**Definição de um modelo de referência de dados
educacionais para a descoberta de conhecimento**

Vanessa Araujo Borges

Tese de Doutorado do Programa de Pós-Graduação em Ciências de
Computação e Matemática Computacional (PPG-CCMC)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Vanessa Araujo Borges

Definição de um modelo de referência de dados educacionais para a descoberta de conhecimento

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutora em Ciências – Ciências de Computação e Matemática Computacional. *EXEMPLAR DE DEFESA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Ellen Francine Barbosa

USP – São Carlos
Agosto de 2017

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

A732d Araujo Borges, Vanessa
 Definição de um modelo de referência de dados
educacionais para a descoberta de conhecimento /
Vanessa Araujo Borges; orientadora Ellen Francine
Barbosa. -- São Carlos, 2017.
 184 p.

 Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2017.

 1. Learning Analytics. 2. Academic Analytics. 3.
Mineração de Dados Educacionais. 4. Sistemas
Educacionais. 5. Business Intelligence. I. Francine
Barbosa, Ellen, orient. II. Título.

Vanessa Araujo Borges

Definition of an educational data reference model for
knowledge discovery

Doctoral dissertation submitted to the Instituto de
Ciências Matemáticas e de Computação – ICMC-
USP, in partial fulfillment of the requirements for the
degree of the Doctorate Program in Computer Science
and Computational Mathematics. *EXAMINATION
BOARD PRESENTATION COPY*

Concentration Area: Computer Science and
Computational Mathematics

Advisor: Profa. Dra. Ellen Francine Barbosa

USP – São Carlos
August 2017

AGRADECIMENTOS

Agradeço a Deus, por me conceder força e determinação para buscar por mais este objetivo.

À Ellen Francine Barbosa, professora e orientadora, pela confiança, palavras de incentivo e, principalmente, pela amizade. Seu apoio constante e disponibilidade irrestrita fizeram toda a diferença.

Aos professores da pós-graduação do ICMC por toda a contribuição para a minha formação e aos funcionários pelo constante auxílio. Em especial, aos professores Renata Pontin de Mattos Fortes, Gustavo Enrique de Almeida Prado Alves Batista pelas valiosas ideias, dicas e sugestões que foram essenciais para a condução deste trabalho.

Aos meus pais, Cida e Borges e irmãs, Clarissa e Sylvia Carolina, pela compreensão, carinho e amor incondicional. A presença de vocês tornam minha caminhada sempre mais suave.

Ao meu esposo Bruno por todo o amor, carinho, paciência e incentivo. Sua presença em todos os momentos e apoio constante, foram fundamentais para a concretização deste trabalho.

Aos grandes amigos que conheci em São Carlos: Everton, Chris, Vinícios, Thainara, Mel, Victor, André, Bárbara Nadai, Igor, Bárbara Rauber, Sandro e Marília, por todo o apoio e constante incentivo.

Aos amigos do LabES, em especial: André Abe, Gambi, Draylson, Gustavo, Joice, Anderson e Maria Lydia que tornaram o ambiente de trabalho sempre descontraído e muito agradável.

Às funcionárias do setor de Pós-Graduação ICMC/USP, Ana Paula, Ana Carolina e Glaucia, pela atenção e apoio.

Ao CNPq pelo apoio financeiro.

Ao Senac MS, representado por Giulliano Cassio de Souza Rosa, Michelle Annita Seibert Kist e Vitor Melo pela compreensão e incentivo nessa fase final do doutorado.

Por fim, deixo minha sincera gratidão a todas as pessoas que contribuíram, direta ou indiretamente, para a concretização deste trabalho.

RESUMO

BORGES, V. A. **Definição de um modelo de referência de dados educacionais para a descoberta de conhecimento.** 2017. 184 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Sistemas educacionais possuem diversas funcionalidades capazes de apoiar a interação entre alunos e professores de maneira dinâmica, síncrona e assíncrona. Uma das formas de monitorar a eficácia do processo educacional é por meio da utilização dos dados armazenados nesses sistemas como fonte de informação. Pesquisas em *Learning Analytics*, *Academic Analytics* e Mineração de Dados Educacionais, buscam explorar os dados de sistemas educacionais utilizando processamento analítico e técnicas de mineração de dados. No entanto, há uma série de fatores que dificultam a gestão eficiente do processo educacional a partir dos dados de sistemas educacionais. A transformação de dados provenientes de diferentes tipos de sistemas educacionais, como Sistemas de Gestão de Aprendizagem e Sistemas Acadêmicos, é uma tarefa complexa devido à natureza heterogênea dos dados. Dados provenientes desses sistemas podem ser analisados considerando diferentes *stakeholders*, sob várias perspectivas e níveis de granularidade. Neste cenário, um modelo de referência para a descoberta de conhecimento a partir de dados de sistemas educacionais, denominado Modelo de Referência de Dados Educacionais (EDRM), foi desenvolvido neste trabalho. O EDRM é um modelo dimensional no formato *star schema*, estruturado em um *Data Warehouse*, projetado para ser uma fonte única de dados integrados e correlacionados voltada à tomada de decisão. Assim, é possível armazenar dados de diversas fontes, combiná-los e, por fim, realizar análises que levem as instituições a desenvolver uma melhor compreensão, rastrear tendências e descobrir lacunas e ineficiências acerca do processo educacional. Neste trabalho, o EDRM foi validado por meio de um estudo de caso, utilizando bases de dados reais coletadas de diferentes sistemas educacionais. Os resultados mostram que o EDRM é eficiente em tarefas com diferentes objetivos, utilizando processamento analítico e mineração de dados.

Palavras-chave:

Learning Analytics, *Academic Analytics*, *Educational Data Mining*, *Business Intelligence*, Sistemas Educacionais

ABSTRACT

BORGES, V. A. **Definition of an educational data reference model for knowledge discovery.** 2017. 184 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2017.

Educational systems support dynamic, synchronous and asynchronous interaction between students and educators. Researches in Learning Analytics, Academic Analytics and Educational Data Mining explore data from educational systems for knowledge discovery through analytical processing, statistical analysis and data mining. However, there are some factors that hinder an efficient management of the educational process. The transformation of data from different kinds of educational system, as Learning Management Systems and Student Information Systems, can be even more difficult due to data heterogeneity. Data from these systems can be analyzed considering different stakeholders, under different perspectives and under different granularities. Motivated by this scenario, in this work we propose Modelo de Referência de Dados Educacionais (EDRM), a reference data model for knowledge discovery in data from educational systems. EDRM is an analytical model structured under a Data Warehouse architecture following a multidimensional data model. EDRM is projected for being an resource of integrated and correlated data focused in decision taking in the educational process. EDRM was developed considering a deep analysis of data and functionalities from different educational systems. In this sense, data from different kinds of systems and sources can be used unified, integrated and consistently. This allows institutions to better comprehend their data, as well as discover patterns, gaps and inefficiencies about their educational process. In this work, EDRM was validated in a case study using real-world databases from different educational systems. The results indicate that EDRM is efficient in tasks with different objectives, using Learning Analytics and Educational Data Mining techniques, and analyzing different perspectives.

Keywords:

Learning Analytics, Academic Analytics, Educational Data Mining, Business Intelligence, Educational Systems

LISTA DE ILUSTRAÇÕES

Figura 1 – Página principal de um curso no <i>Moodle</i>	34
Figura 2 – Página principal e algumas ferramentas do <i>TidiaAe</i>	37
Figura 3 – Página inicial e área restrita a participantes de um curso oferecido na plataforma <i>Edx</i>	38
Figura 4 – Modelo de referência para o processo de <i>Learning Analytics</i>	46
Figura 5 – Etapa do processo de mineração de dados	61
Figura 6 – Elementos básicos de um <i>Data Warehouse</i>	83
Figura 7 – Modelo <i>star schema</i> proposto por Moscoso-Zea, Andres-Sampedro e Luján-Mora (2016).	90
Figura 8 – Modelos <i>star schema</i> propostos por Dimokas <i>et al.</i> (2008).	91
Figura 9 – Modelo dimensional <i>Snowflake Schema</i> proposto por Valdiviezo-Díaz <i>et</i> <i>al.</i> (2015).	92
Figura 10 – Processo de caracterização dos dados do EDRM.	96
Figura 11 – Modelo <i>Star Schema</i> do EDRM.	110
Figura 12 – Modelo <i>Star Schema</i> completo do EDRM	119
Figura 13 – Exemplo de granularidade da perspectiva curso.	121
Figura 14 – Matriz de <i>stakeholders</i> por perspectiva do EDRM.	123
Figura 15 – Quantidade de usuários por modalidades da instituições de ensino analisadas.	130
Figura 16 – Visão geral do estudo de caso seguindo as etapas do processo de <i>Business</i> <i>Intelligence</i>	134
Figura 17 – Modelo <i>star schema</i> do EDRM para as instituições I1, I2 e I3 conside- rando seus <i>stakeholders</i> e níveis de granularidade.	136
Figura 18 – Processamento de ETL para cada fonte de dados.	137
Figura 19 – Mapeamento analítico realizado a partir dos dados da instituição I1.	138
Figura 20 – Mapeamento analítico realizado a partir dos dados das instituições I2 e I3.	139
Figura 21 – <i>Dashboard</i> desenvolvido para a instituição I1.	140
Figura 22 – <i>Dashboard</i> desenvolvido para a instituição I2.	140
Figura 23 – Árvore de decisão para a disciplina <i>Interação Humano-Computador</i>	152
Figura 24 – Árvore de decisão para a disciplina Introdução à Engenharia de Com- putação.	153

Figura 25 – Árvore de decisão para a predição de evasão.	154
Figura 26 – Análise de silhueta para agrupamento de alunos.	156
Figura 27 – Resultado do agrupamento de alunos considerando K=6.	156
Figura 28 – Média de notas dos alunos <i>vs</i> quantidade de disciplinas matriculadas e cursadas com aproveitamento.	157
Figura 29 – Análise de silhueta para agrupamento de disciplinas.	158
Figura 30 – Resultado do agrupamento de disciplinas considerando K=5.	159
Figura 31 – Média de notas nas disciplinas <i>vs</i> quantidade de alunos matriculados e aprovados.	159

LISTA DE QUADROS

Quadro 1 – Características de <i>Learning Analytics</i> e <i>Academic Analytics</i>	45
Quadro 2 – Índice de maturidade de <i>Academic Analytics</i>	51
Quadro 3 – Mineração de dados educacionais: áreas de aplicação	72
Quadro 4 – Diferença entre ambientes operacionais e informacionais.	80
Quadro 5 – Principais funcionalidades de sistemas educacionais.	99
Quadro 6 – Atributos definidos a partir de pesquisas nas área de Indicadores de Desempenho, Mineração de Dados Educacionais, <i>Business Intelligence</i> , <i>Data Warehouse</i> e modelo dimensional de dados educacionais.	104
Quadro 8 – Análises de medidas e indicadores de desempenho educacional.	107
Quadro 9 – Dimensões em conformidade e comuns do modelo <i>star schema</i> do EDRM.	113
Quadro 12 – Ferramentas utilizadas no processo de <i>Business Intelligence</i>	132
Quadro 14 – Sumário dos atributos da instância do EDRM utilizado nas aplicações preditivas e descritivas.	143
Quadro 15 – Exemplo de matriz atributo-valor de A1.	144
Quadro 16 – Exemplo de matriz atributo-valor de A3.	144
Quadro 17 – Exemplo de matriz atributo-valor de A4.	144
Quadro 18 – Exemplo de matriz atributo-valor de A2.	145
Quadro 19 – Parâmetros testados nos algoritmos de classificação.	145

LISTA DE TABELAS

Tabela 1 – Resultados da predição de aprovação nas disciplinas da aplicação A1. .	149
Tabela 2 – Resultados da predição de evasão da aplicação A2.	153

LISTA DE ABREVIATURAS E SIGLAS

AA	<i>Academic Analytics</i>
AIH	<i>Adaptive and Intelligent Hypermedia</i>
BI	<i>Business Intelligence</i>
CMS	<i>Course Management System</i>
DM	<i>Data Marts</i>
DW	<i>Data Warehouse</i>
EDRM	<i>Educational Data Reference Model</i>
IA	<i>Institutional Analytics</i>
IES	Instituições de Ensino Superior
KPI	<i>Key Performance Indicator</i>
LA	<i>Learning Analytics</i>
LMS	<i>Learning Management Systems</i>
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
MDW	Mineração de Dados na Web
MOOC	<i>Massive Open Online Courses</i>
MT	Mineração de Textos
OBIEE	<i>Oracle Business Intelligence Enterprise Edition</i>
OLAP	<i>On-line Analytical Processing</i>
OLTP	<i>On-line Transaction Processing</i>
PDI	<i>Pentaho Data Integrator</i>
RDBMS	<i>Relational Database Management System</i>
SAD	Sistema de Apoio à Decisão
SGBD	Sistema Gerenciador de Banco de Dados
SIS	<i>Student Information Systems</i>
SK	<i>Surrogate Key</i>
SNA	<i>Social Network Analysis</i>
TI	Tecnologia da Informação
TM	<i>Text Mining</i>

UAB	Universidade Aberta do Brasil
WUM	<i>Web Usage Mining</i>
XML	<i>Extensible Markup Language</i>

SUMÁRIO

1	INTRODUÇÃO	23
1.1	Contexto e motivação	23
1.2	Objetivos	29
1.3	Organização do trabalho	30
2	SISTEMAS EDUCACIONAIS: VISÃO GERAL	31
2.1	Considerações iniciais	31
2.2	Tipos de sistemas educacionais	31
2.2.1	Moodle	34
2.2.2	Sakai e TidiaAe	36
2.2.3	Edx	38
2.3	Características de sistemas educacionais	39
2.4	Considerações finais	41
3	ANÁLISE DE DADOS EDUCACIONAIS	43
3.1	Considerações iniciais	43
3.2	Analytics	44
3.2.1	Learning Analytics	44
3.2.2	Academic Analytics	50
3.2.3	Indicadores de desempenho educacional	52
3.3	Mineração de dados educacionais	59
3.3.1	Mineração de dados	59
3.3.2	Métodos para mineração de dados educacionais	65
3.3.3	Aplicações de mineração de dados educacionais	71
3.4	Considerações finais	75
4	PROCESSO DECISÓRIO ORIENTADO A DADOS	77
4.1	Considerações iniciais	77
4.2	Ambientes operacionais e informacionais	78
4.3	Data warehouse	79
4.3.1	Características dos dados	81

4.3.2	Arquitetura de data warehouse	82
4.3.3	Modelagem dimensional	83
4.4	Representação lógica do modelo dimensional	86
4.5	Business intelligence	88
4.6	Aplicações de data warehouse em sistemas educacionais	89
4.7	Considerações finais	93
5	MODELO DE REFERÊNCIA DE DADOS EDUCACIONAIS	95
5.1	Considerações iniciais	95
5.2	Caracterização dos dados educacionais	96
5.2.1	Dados e funcionalidade de sistemas educacionais	97
5.2.2	Dados para análise educacional a partir de indicadores de desempe- nho e mineração de dados educacionais	100
5.2.3	Sumarização de atributos do EDRM	103
5.3	Modelo dimensional star schema do EDRM	109
5.4	Definição dos stakeholders	122
5.5	Decisões iniciais para a instanciação do modelo	123
5.6	Considerações finais	124
6	ESTUDO DE CASO: INSTANCIÇÃO DO MODELO DE REFE- RÊNCIA DE DADOS EDUCACIONAIS	127
6.1	Considerações iniciais	127
6.2	Planejamento e definições do estudo de caso	128
6.2.1	Características dos dados	129
6.2.2	Análises planejadas	130
6.2.3	Ferramentas utilizadas	131
6.2.4	Segurança da informação	133
6.3	Processamento e análise dos dados	134
6.3.1	Etapa 1: instanciação do EDRM e processo de ETL	135
6.3.2	Etapa 2: análises e dashboards	137
6.3.2.1	<i>Etapa 2a: análises baseadas no processamento analítico</i>	<i>137</i>
6.3.2.2	<i>Etapa 2b: análises a partir de tarefas de mineração de dados</i>	<i>141</i>
6.3.3	Análise do estudo de caso	160
6.4	Considerações finais	162
7	CONCLUSÕES	163
7.1	Visão geral da pesquisa realizada	163
7.2	Contribuições de pesquisa	165

7.3	Dificuldades e limitações	167
7.4	Produção científica	168
7.5	Trabalhos futuros	168
REFERÊNCIAS		171

INTRODUÇÃO

1.1 Contexto e motivação

Ao longo dos últimos anos, o processo educacional vem passando por uma significativa transformação, incorporando tecnologias que proporcionam maior interação, colaboração e comunicação em ambientes cada vez mais personalizados e adaptáveis às necessidades de alunos e professores (BIENKOWSKI; FENG; MEANS, 2012; ROMERO; VENTURA, 2010). Instituições de ensino vêm ampliando sua capacidade física oferecendo além dos tradicionais cursos presenciais, cursos híbridos (*blended learning*) ou, até mesmo, totalmente a distância (EaD).

Inúmeras universidades, como *Massachusetts Institute of Technology*, *Stanford* e *Berkeley* vêm promovendo cursos *online* gratuitos e de qualidade por meio do *Open CourseWare Consortium* (OpenCourseWare, 1999). No Brasil, embora existam várias iniciativas relacionadas à educação híbrida ou a distância, a popularização de cursos *online* para o ensino superior iniciou em 2005 com criação da Universidade Aberta do Brasil (UAB)(MEC, 2005). Tal programa foi uma iniciativa do Ministério da Educação em parceria com empresas estatais, com o objetivo de democratizar o ensino superior no Brasil.

Sistemas educacionais estão cada vez mais populares devido à evolução da Web aliada à evolução dos sistemas computacionais. Esses sistemas, atualmente, possuem diversas funcionalidades capazes de apoiar a interação entre alunos e professores de maneira dinâmica, síncrona e assíncrona. Dentre diversos tipos de sistemas educacionais, destacam-se o *Moodle* (MOODLE, 2006), *Sakai* (COUNTERMAN *et al.*, 2004), *TidiaAe* (TIDIA, 2006), entre outros. Em uma perspectiva relacionada, sistemas educacionais utilizados em

massa, também conhecidos por *Massive Open Online Courses* (MOOC), como *Edx* (MIT and Harvard University, 2013), *Coursera* (NG; KOLLER, 2012) e *Udacity* (THRUN; STAVENS; SOKOLSKY, 2011), vêm sendo utilizados por diversas instituições de ensino com o objetivo de promover cursos gratuitos a qualquer pessoa, em qualquer lugar e a qualquer tempo (KHALIL; EBNER, 2016). Tais sistemas buscam auxiliar a automatização de tarefas comuns relativas ao ensino por possibilitarem, entre outras funcionalidades, a condução de atividades colaborativas, a comunicação por meio de ferramentas síncronas e assíncronas associadas, além de oferecer apoio à disponibilização do material didático (BARBOSA; MALDONADO, 2011).

Em contraste à crescente disponibilização de ambientes *online* de apoio ao processo de ensino e aprendizagem, é possível observar a dificuldade tanto na gestão desses sistemas educacionais como no seu monitoramento a partir de metas institucionais. Martin, Sauvageot e Tchatchoua (2011) afirmam que a escassez de análises regulares para a avaliação do desempenho institucional é o maior obstáculo para o progresso no setor. Se o objetivo educacional é desempenhar um papel importante na globalização, é de extrema importância a realização de melhorias em suas práticas de gestão, operação e desempenho.

Outra dificuldade refere-se à alta taxa de evasão em cursos de diferentes modalidades. No Brasil, por exemplo, a evasão em cursos superiores em instituições públicas de ensino atingiu, em 2014, a taxa de 49%, de acordo com os dados do Censo de Educação Superior 2015¹. Esse também é um problema recorrente, em sistemas educacionais utilizados em massa ou MOOCs devido ao grande volume de alunos com perfis heterogêneos em busca de novos conhecimentos. Pesquisas que investigam formas de aumentar a taxa de conclusão buscam monitorar seus sistemas educacionais por meio de indicadores de desempenho para otimizar o processo educacional e disponibilizar indicadores acadêmicos para que os alunos possam acompanhar sua evolução (DEDE; HO; MITROS, 2016).

Segundo Arroyo *et al.* (2016), as instituições vêm realizando análises de dados em busca de aprimorar seu processo educacional para promover melhores experiências, conter custos e gerenciar recursos adequadamente. Os autores enfatizam que a única forma de manter pessoas, processos e cultura educacional em desenvolvimento é afastando métodos de baixa eficácia e adotando práticas que melhor se adaptam às necessidades dos alunos.

Sob essa perspectiva, para promover melhorias para sucesso acadêmico, instituições de ensino estão fazendo uso de recursos de Tecnologia da Informação (TI) para gerar e otimizar análises por meio de ferramentas e algoritmos. A utilização de TI em educação não é algo trivial, pois tanto a tecnologia como o ensino e aprendizagem vêm passando por

¹ Website MEC: <http://portal.mec.gov.br/component/tags/tag/32044-censo-da-educacao-superior>

significativas transformações ao longo do tempo. A crescente utilização de TI, principalmente explorando dados educacionais, é evidenciada em pesquisas como o relatório anual *Top 10 Issues – EDUCAUSE Review* (GRAJEK, 2017), que classifica as dez principais tendências em assuntos de TI que são estratégicas para a educação com foco em ensino superior. De acordo com a pesquisa, nos anos de 2016 e 2017, algumas questões importantes associadas aos dados educacionais foram enfatizadas entre as dez principais (GRAJEK, 2016; GRAJEK, 2017). Em quinto lugar em 2016 e terceiro lugar em 2017, a pesquisa aponta a importância da **Gestão de Dados Institucionais e Governança** com foco na melhoria dessa gestão por meio de padronização dos dados, integrações, proteção e governança. Em sétimo lugar em 2016 e sexto lugar em 2017, **Tomada de Decisão a partir dos Dados** é evidenciada com o objetivo de desenvolver métodos eficazes de *Business Intelligence*, relatórios e análises para garantir informações relevantes à tomada de decisão que possam ser consultadas por administradores/gestores, professores e alunos.

Nesse contexto, Martin, Sauvageot e Tchatchoua (2011) afirmam que a gestão de sistemas educacionais pode ter diversos propósitos, dentre eles: (1) determinar o estado atual de um sistema educacional; (2) acompanhar o seu desenvolvimento e progresso ao longo do tempo; (3) monitorar o andamento de uma estratégia ou o progresso de um planejamento que foi implementado; e (4) gerenciar a instituição como um todo. Isso explicita que, embora pesquisadores, universidades e empresas estejam empenhados em desenvolver plataformas *online* e materiais educacionais de qualidade com o objetivo popularizar o ensino, ainda existe um grande desafio tecnológico e computacional para tornar a experiência de aprendizagem única para cada aluno.

Um instrumento que pode ser utilizado para promover a melhoria durante o processo de ensino e aprendizagem é o conhecimento que pode ser extraído da grande massa de dados gerados e facilmente acessíveis por sistemas educacionais (DEDE; HO; MITROS, 2016; BROOKS; THAYER, 2016). Instituições de ensino vêm gerando grandes volumes de dados que poderiam ser utilizados de forma estratégica, tanto para realizar melhorias no processo educacional, como para obter vantagens competitivas por meio de uma visão rápida e estratégica para mitigar riscos e descobrir novas oportunidades de melhoria.

Os dados dos sistemas educacionais podem ser provenientes de diversas funcionalidades, projetadas para possibilitar a gestão do ensino e aprendizagem a partir de diferentes modalidades educacionais (presencial, híbrida ou a distância). Dados sobre alunos, disciplinas e professores, bem como dados relacionados à interação destes atores entre si e com os sistemas educacionais, são armazenados em quantidades cada vez maiores. Com essa grande quantidade de dados armazenados, há um vasto conhecimento do processo

educacional que pode ser explorado para entendimento e otimização do mesmo por meio de dados analíticos dimensionais em *Data Warehouse* e Mineração de Dados Educacionais.

Converter dados brutos em informações, capazes de gerar conhecimento potencialmente útil, trata-se uma atividade não trivial que pode gerar impactos durante o processo educacional. Essa descoberta de conhecimento potencialmente útil em sistemas educacionais tem sido amplamente investigada com o objetivo de auxiliar professores, administradores e pesquisadores a realizarem uma melhor gestão destes sistemas e, conseqüentemente, promover melhorias no processo educacional.

Sob essa perspectiva, pesquisadores vêm buscando formas de otimizar a atuação de diferentes tipos de *stakeholders* por meio de análise de grandes volumes de dados. Dentre as linhas de pesquisa relevantes para essa finalidade destacam-se as baseadas em *Analytics: Learning Analytics* (LA), *Academic Analytics* (AA) ou *Institutional Analytics* (IA) e Mineração de Dados Educacionais (MDE) (ARROWAY *et al.*, 2016; CAMPBELL; DEBLOIS; OBLINGER, 2007; MOSCOSO-ZEA; ANDRES-SAMPEDRO; LUJÁN-MORA, 2016).

Em linhas gerais, *Analytics* refere-se ao uso de dados, análise estatística, e modelos explicativos e preditivos para a obtenção de percepções para realizar ações a partir de questões complexas (YANOSKY; ARROWAY, 2015). LA e AA são aplicações de *Analytics*, cada uma abordando um grande domínio estratégico. Enquanto LA busca melhorar o sucesso dos alunos, AA tem por objetivo alavancar as práticas de gestão da instituição como um todo.

A MDE, por sua vez, é uma linha de pesquisa explorada tanto no domínio de LA como de AA. Pesquisas em MDE buscam utilizar métodos de Mineração de Dados para realizar a descoberta de padrões potencialmente úteis a partir de dados educacionais. Desse modo, é possível combinar grandes volumes de dados com técnicas estatísticas e modelos gerados por algoritmos de aprendizado de máquina para prover apoio a tomada de decisão orientada a dados (CAMPBELL; OBLINGER, 2007).

Tanto LA como AA utilizam dados provenientes de sistemas de ensino e aprendizagem para realização de predições, intervenções e adaptações durante o processo educacional. Assim, o processo de tomada de decisão a partir de dados educacionais vai muito além da geração dos tradicionais relatórios. Trata-se de apoio a decisão a partir de dados brutos, com o objetivo de possibilitar melhorias no ensino, aprendizagem e sucesso acadêmico (CAMPBELL; OBLINGER, 2007; COOPER; ADAM, 2012).

Para sintetizar a notação referente às linhas de pesquisa utilizadas neste trabalho, foi definido o termo Análise Educacional para abranger as grandes áreas de pesquisa em

análise de dados educacionais: *Learning Analytics*, *Academic Analytics* e Mineração de Dados Educacionais.

Para a utilização de Análise Educacional, é importante diferenciar o tipo de sistemas educacionais de acordo com seus dados (COOPER; ADAM, 2012). Sistemas de Gestão de Aprendizagem (*Learning Management Systems* – LMS) são sistemas educacionais que possuem dados que podem ser analisados em tempo real. Esse tipo de sistema possui basicamente um conjunto de ferramentas de colaboração, comunicação (síncronas e assíncronas associadas), controle de acesso, apoio a avaliação e acompanhamento do aluno, armazenamento de material didático, entre outras. Já os Sistemas de Gestão do Aluno ou Sistema Acadêmico (*Student Information Systems* – SIS) possuem dados com informações históricas que podem ser carregados em intervalos de tempo maiores. Esse tipo de sistema é utilizado para apoiar o processo administrativo da instituição de ensino. Suas principais funcionalidades estão associadas a gestão de *stakeholders*, notas dos alunos, controle de frequência, ementa do curso, entre outras.

De modo geral, as duas categorias podem ser utilizadas individualmente ou em conjunto. Quando SIS é utilizado individualmente, as análises geradas são referentes ao processo de gestão administrativa da instituição e análises de *stakeholders* (comuns no contexto presencial). Ao combinar SIS com LMS, é possível agregar ao modelo educacionais interações de colaboração e comunicação (comuns em contextos à distância ou híbridos).

Os pesquisadores Campbell e Oblinger (2007) relatam que informações em tempo real podem gerar modelos analíticos mais interessantes para o sucesso acadêmico. No entanto, análises de dados em tempo real são complexas quando consideradas as características de:

- **Armazenamento:** dados em tempo real geram maior volume, ou seja, requerem maior espaço de armazenamento. Um exemplo está nos dados de acesso de LMS quando comparados a dados de matrícula em um SIS. Os acessos podem ocorrer diariamente, ou até mesmo em intervalos menores, enquanto as matrículas podem ocorrer uma vez por curso por aluno.
- **Granularidade:** está relacionada ao nível de detalhamento, ou seja, é possível gerar análises considerando dados diários ou sua sumarização por ano, por exemplo. A granularidade é decidida de forma equilibrada considerando os dados que a instituição de ensino possui, as análises que são desejáveis e a possibilidade de armazenar maior ou menor volume de dados.

Nesse sentido, o armazenamento e utilização de dados históricos precisam ser estruturados e bem definidos para sua efetiva utilização considerando múltiplas fontes de dados, diferentes perspectivas correlacionadas e níveis de granularidade.

Segundo Rudy, Miranda e Suryani (2014), sistemas educacionais geram grandes volumes de dados operacionais diariamente que precisam ser acessados a qualquer momento para que seja possível incorporar ao processo educacional práticas de *Analytics*. No entanto, analisar grandes volumes de dados heterogêneos é um processo complexo que demanda conhecimento acerca da estrutura dos dados operacionais, considerando múltiplas fontes, para que seja possível obter informações potencialmente úteis. Dentre os vários desafios em transformar dados em informação, está a transformação de dados operacionais em dados analíticos de qualidade, ou seja, integrados e consistentes. Além disso, a estruturação desses dados é um fator essencial para sua utilização efetiva.

Pesquisas em *Business Intelligence*, que abordam técnicas e tecnologias para a transformação grandes volumes de dados em informação, vêm sendo amplamente exploradas com o objetivo de auxiliar o processo decisório orientado a dados no contexto educacional. A definição de modelos de dados analíticos, estruturadas e armazenadas em repositórios informacionais ou *Data Warehouse* de forma integrada e consistente vêm gerando grandes benefícios às instituições por possibilitarem acesso eficiente e correlacionado aos dados para fins de apoio a decisão.

Diversas pesquisas vêm sendo conduzidas nesse sentido como, Pasyeka e Pasyeka (2016), Moscoso-Zea, Andres-Sampedro e Luján-Mora (2016), Rudy, Miranda e Suryani (2014), Nebic e Mahnič (2010), Wai e Aung (2009) e Dimokas *et al.* (2008). No entanto, essas pesquisas abordam questões específicas considerando análises pontuais, sem a utilização efetiva do modelo, principalmente com relação a utilização de diferentes perspectivas educacionais correlacionadas contendo vários níveis de granularidade.

Uma vez que o processo decisório tem papel fundamental no contexto educacional, considerando-se a carência de modelos de dados de referência para apoiar o processo de descoberta de conhecimento potencialmente útil, a proposição de um modelo de referência mostra-se bastante relevante. Visando a reutilização de estruturas de dados complexas e heterogêneas de sistemas educacionais, neste trabalho propõe-se um Modelo de Referência de Dados Educacionais (*Educational Data Reference Model – EDRM*) desenvolvido para apoiar o processo decisório orientado a dados no contexto educacional por meio de dados analíticos explorados a partir de pesquisas em *Learning Analytics*, *Academic Analytics* e Mineração de Dados Educacionais. A ideia é que o modelo seja capaz de reduzir a complexidades de análises e extração de padrões potencialmente úteis à tomada de decisão no contexto educacional.

1.2 Objetivos

Diante do cenário apresentado na seção anterior, em linhas gerais, o presente trabalho consiste em investigar dados de sistemas educacionais para a proposição de um modelo de dados analítico de referência que facilite a descoberta de conhecimento a partir de evidências por parte de instituições educacionais. Mais especificamente, a definição de um Modelo de Referência de Dados Educacionais (*Educational Data Reference Model – EDRM*), um modelo dimensional definido para dar subsídio à tomada de decisão orientada a dados por meio de processamento analíticos e Mineração de Dados Educacionais. Com isso, pretende-se que o modelo possa reduzir a complexidades de análises e extração de padrões potencialmente úteis à tomada de decisão no contexto educacional a fim de:

1. Possibilitar análises considerando diferentes *stakeholders* (alunos, professores, tutores, monitores, gestores, etc);
2. Possibilitar análises sob diferentes níveis de granularidade (instituição, departamento, curso, disciplina, turma, etc);
3. Possibilitar análises que integre diferentes fontes de dados de uma mesma instituição (LMS e SIS);
4. Possibilitar análises em nível de usuário, instituição e inter-institucionais;
5. Possibilitar análises *ad hoc* a partir de um planejamento estratégico previamente estabelecido;
6. Possibilitar análise que cruze dados de perspectivas distintas (acesso, comunicação, notas, etc);
7. Possibilitar que o modelo possa ser analisados independente de plataforma analítica (Pentaho ² e OBIEE ³, por exemplo).
8. Possibilitar a utilização de dados históricos e de tempo real como fonte de informação;
9. Possibilitar a realização de inferências utilizando Mineração de Dados Educacionais como parte do processo decisório.

Para avaliar a viabilidade do modelo foi realizado um estudo de casos relatados no Capítulo 6. De modo geral esses casos consideram três instituições de ensino superior

² Pentaho website: <http://community.pentaho.com/>

³ OBIEE website: <http://www.oracle.com/technetwork/middleware/bi/index.html>

nacionais (duas públicas e uma privada), com modalidades de ensino presencial, híbrida e a distância que utilizam os dois tipos de sistemas educacionais investigados, SIS e LMS.

1.3 Organização do trabalho

Neste capítulo foram apresentados o contexto no qual este trabalho se insere, as motivações para a sua realização e os principais objetivos relacionados.

No Capítulo 2 é apresentada uma visão geral a respeito de dos principais tipos de sistemas educacionais utilizados com foco na experiência de aprendizagem do aluno e atuação dos demais *stakeholders*.

No Capítulo 3 é apresentado o processo de Mineração de Dados e suas respectivas etapas. Em seguida é apresentada uma visão geral acerca da área de Mineração de Dados Educacionais, considerando seus principais métodos e aplicações.

No Capítulo 4 são apresentadas técnicas de modelagem de analítica de dados e armazenamento em repositórios informacionais denominados *Data Warehouse* para fins de apoio a decisão baseada em evidências.

No Capítulo 5 é apresentado o Modelo de Referência de Dados Educacionais (*Educational Data Reference Model* – EDRM) propostos no escopo deste trabalho, resultante da investigação de dados reais de sistemas educacionais e pesquisas em *Learning Analytics*, *Academic Analytics* e Mineração de Dados Educacionais.

No Capítulo 6 é descrito um estudo de casos que consideram diferentes análises realizadas com dados de sistemas educacionais de três instituições de ensino.

Por fim, no Capítulo 7 são caracterizados os principais resultado, as contribuições, bem como as perspectivas de trabalhos futuros.

SISTEMAS EDUCACIONAIS: VISÃO GERAL

2.1 Considerações iniciais

Conforme discutido anteriormente, no contexto deste trabalho, é definido um Modelo de Referência de Dados Educacionais (*Educational Data Reference Model* – EDRM). O modelo de dados proposto foi projetado para integrar dados oriundos de diferentes tipos sistemas educacionais, como *Learning Management Systems* (LMS) e *Student Information Systems* (SIS). Em linhas gerais, esses dados foram estruturados em um modelo dimensional para a sua representação de forma analítica com o objetivo de possibilitar a tomada de decisão orientada a dados durante o processo educacional.

Neste capítulo são apresentados os principais tipos de sistemas educacionais existentes, fornecendo uma visão geral de suas principais funcionalidades. Essa revisão forneceu subsídios à definição do modelo proposto. São explorados aspectos relativos à utilização dos dados considerando diferentes funcionalidades que podem ser traduzidas em perspectivas e níveis de granularidade. Na Seção 2.2 são caracterizados os principais tipos de sistemas educacionais com foco em ações realizadas pelo aluno e seu contexto de aprendizagem. Na Seção 2.3 são descritas características de sistemas educacionais considerando as funcionalidades importantes à sua nova geração.

2.2 Tipos de sistemas educacionais

De modo geral, sistemas educacionais têm por objetivo apoiar o processo de ensino e aprendizagem fornecendo um conjunto de ferramentas quem buscam auxiliar a disponibilização de material didático, a avaliação e acompanhamento dos aprendizes,

atividades de comunicação síncrona e assíncrona, entre outras (EBERSPACHER *et al.*, 1999; REYES; TCHOUNIKINE, 2005). Tais sistemas fornecem apoio tecnológico e computacional, para que instituições de ensino ofereçam cursos em diferentes modalidades (presencial, híbrida ou a distância), independentemente da quantidade de alunos e de sua localização geográfica.

De fato, sistemas educacionais vêm ampliando o alcance das salas de aula incluindo um conjunto integrado de ferramentas para apoiar a experiência de aprendizagem em qualquer lugar, a qualquer tempo e hora. Esses sistemas influenciam a maneira como as informações e as ideias são criadas e compartilhadas, influenciando também como alunos e professores constroem o conhecimento durante o processo educacional (DAVIS; CARMEAN; WAGNER, 2009). Contudo, grande parte desses sistemas não possuem funcionalidades analíticas eficientes para apoiar o monitoramento e a gestão do processo educacional.

Sob essa perspectiva, a gestão de instituições de ensino pode ser realizada a partir de um conjunto de sistemas que possuem diversas finalidades como: gerenciador de recursos corporativos (*Enterprise Resource Planning* – ERP), gerenciador de relacionamento com o cliente (*Customer Relationship Management* – CRM), sistema de informação do aluno ou sistema acadêmico (*Student Information Systems* – SIS), sistema de gestão de aprendizagem (*Learning Management Systems* – LMS), entre outros. No entanto, LMS e SIS são os principais sistemas educacionais utilizados quando se deseja analisar ações realizadas pelo aluno no contexto o qual ele está inserido (modalidades presencial, híbrida ou a distância). Lang e Pirani (2014) enfatizam que LMS e SIS são sistemas universais que, nos últimos anos, fazem parte da maioria das experiências de aprendizagem das instituições.

Outros tipos de sistemas educacionais, como sistemas tutores inteligentes, sistemas hipermídia inteligentes e adaptativos, sistemas de avaliação e questionários, entre outros, também são voltados à experiência de aprendizagem do aluno. Contudo, esses sistemas ainda não possuem ampla utilização por parte de instituições de ensino. Além disso, alguns desses sistemas podem ser encontrados como módulos ou funcionalidades disponibilizadas por LMS como, por exemplo, questionário e avaliações.

Sob essa perspectiva, para a definição do EDRM proposto neste trabalho, buscou-se enfatizar análises em sistemas acadêmicos (SIS) e sistemas de gestão de aprendizagem (LMS) com o objetivo de criar um modelo de dados analítico, para a tomada de decisão, com foco na experiência de aprendizagem do aluno e atuação dos demais *stakeholders* nesse contexto.

Sistemas acadêmicos

Sistemas acadêmicos (*Student Information Systems – SIS*), são sistemas educacionais cuja finalidade é gerenciar informações do aluno, incluindo o processo de admissão, registro de cursos e notas, auditoria, habitação, entre outros (LANG; PIRANI, 2014). Além disso, é capaz de fornecer auto-serviço para alunos (como matrícula e gestão financeira) e acesso a catálogos de cursos, horários de aulas, notas, *feedback*, e assim por diante.

Esse tipo de sistema possui dados com informações históricas que podem ser analisados em intervalos de tempo maiores. Adicionalmente, SIS podem ser utilizados para apoiar o processo administrativo da instituição. Suas principais funcionalidades estão associadas a gestão de *stakeholders*, notas dos alunos, controle de frequência, ementa do curso, entre outras.

Vale ressaltar que SIS podem ser considerados como uma das principais fontes de dados de gestão educacional, pois possuem dentre diversas informações cadastrais, informações de matrícula, alocação de docentes, notas intermediárias e finais.

Sistemas de gestão de aprendizagem

Sistemas de gestão de aprendizagem (*Learning Management Systems – LMS*) também conhecidos como sistema de gestão de curso (*Course Management System – CMS*), são sistemas computacionais que possibilitam a automatização das tarefas relativas ao ensino, oferecendo apoio à disponibilização de material didático, à condução de atividades colaborativas e à avaliação e acompanhamento dos aprendizes (LANG; PIRANI, 2014). Tais sistemas registram os dados de todas as atividades dos alunos envolvidos, como a leitura, a escrita, os questionários, as tarefas em tempo real e interações com seus pares.

Nota-se que neste capítulo não foram detalhados sistemas do tipo SIS. Em geral, instituições de grande porte desenvolvem seu próprio sistema SIS como, a USP, por exemplo, conta com o Júpter¹ para a graduação, e o Janus² para a pós graduação; a UFMS com o SISCAD³; a UFSCar com o SIGA⁴, entre outros. No caso de instituições de menor porte, é comum a personalização de algumas funcionalidades de LMS mais robustos, como o Moodle⁵, para possibilitar parte da gestão tipicamente realizadas por SIS.

Sob essa perspectiva, em geral, SIS são desenvolvidos e mantidos pela própria instituição de ensino utilizadora, não havendo um consenso de SIS amplamente utilizado por

¹ USP Júpter website: <https://uspdigital.usp.br/jupiterweb/>

² USP Janus website: <https://uspdigital.usp.br/janus/>

³ UFMS SISCAD website: <https://siscad.ufms.br/>

⁴ UFSCar SIGA Website: <https://sistemas.ufscar.br/siga/>

⁵ Moodle website: <https://moodle.com/>

instituições educacionais. Por outro lado, os LMS são amplamente utilizados, principalmente os disponíveis como *software* livre e, em alguns casos, são otimizados para instituições que não possuem seu próprio sistema acadêmico (SIS).

Na próxima seção, são apresentados alguns sistemas educacionais existentes, do tipo LMS, a fim de fornecer uma visão geral a respeito das pesquisas conduzidas na área. Tais sistemas foram escolhidos pois, neste trabalho de doutorado, buscou-se investigar fontes de dados educacionais potencialmente úteis e que possuam ampla utilização por parte as instituições de ensino.

2.2.1 Moodle

O Moodle (*Modular Object-Oriented Dynamic Learning Environment*) (COMMUNITY, 2016; PROJECT, 2017; MOODLEDOCS, 2006; DOUGIAMAS; TAYLOR, 2003), ilustrado na Figura 1, é um sistema de gestão de aprendizagem (LMS) cujo desenvolvimento teve início em 1999. Suas principais funcionalidades são (PROJECT, 2017; COMMUNITY, 2016):

Figura 1 – Página principal de um curso no Moodle.

Fonte: Elaborada pelo autor.

- Gerenciamento do site: o site é gerenciado pelo usuário administrador, definido na instalação.
- Gerenciamento de usuários: cada usuário requer apenas uma conta para todo servidor. Desse modo é possível que o aluno participe de vários cursos utilizando a mesma conta.

- Gerenciamento dos cursos: o tutor de um curso possui controle sobre toda a configuração do mesmo. Atividades como fóruns, notícias, *quizzes*, recursos, exercícios, avaliações, *chats* e *workshops*, também são programadas pelo tutor. A maior parte das áreas de texto (fontes, recursos, fóruns, notícias) pode ser editada usando um editor HTML incorporado.
- Relatórios de acompanhamento e progresso do aluno: relatórios de acessos, notas, conclusão de atividades e conclusão do curso.
- Módulos de interação entre os usuários: permitem a troca de informações entre os usuários do *Moodle*. O sucesso na comunicação interativa ocorre especificamente por meio do conhecimento dos usuários e da manipulação adequada dos recursos oferecidos. Os seguintes módulos estão disponíveis no sistema:
 - Módulo de tarefas: auxilia o professor a conduzir tarefas e avaliar os alunos. O *feedback* do professor é adicionado na página da tarefa para cada aluno e uma notificação é enviada por *e-mail*.
 - Módulo de *chat*: único módulo do sistema que permite interação síncrona por texto entre os alunos e o professor.
 - Módulo de enquete: similar a uma pesquisa, pode ser usado para votação ou para obter *feedback* de cada estudante.
 - Módulo de fórum: os diferentes tipos de fóruns são disponibilizados exclusivamente pelos professores e são utilizados para a comunicação assíncrona entre os participantes ou, ainda, para a publicação de notícias dos cursos. Podem ser criados quantos fóruns de discussão forem necessários.
 - Módulo *quiz*: permite a publicação e correções automáticas de questões em diversos formatos, tais como respostas curtas, verdadeiro ou falso, relacionamento, aleatórias, numéricas, textos descritivos e com imagens integradas.
 - Módulo de arquivo: serve para carregar arquivos para o servidor de um modo prático e simples. Apoia a exibição de qualquer conteúdo eletrônico tais como documentos, apresentações, animações, vídeos, sons, entre outros.
 - Módulo *workshop*: permite compartilhamento e trabalho conjunto dos alunos em documentos onde o professor pode gerenciar e avaliar o trabalho. Os professores podem prover exemplos de documentos para exercícios práticos.
 - Módulo glossário: permite a criação de um dicionário com definições dos termos usados nos conteúdos dos cursos.

O sistema *Moodle* está disponível como um *software* livre (*Open Source*), sob as condições da *GNU Public License* (GPL) (Free Software Foundation, 2007). Esse sistema tem sido amplamente utilizado no contexto educacional e, por isso, possui diversas pesquisas que buscam extrair conhecimento potencialmente útil a partir dos seus dados (SAEL; BEHJA, 2013; MAZZA; BOTTURI, 2007; ROMERO; VENTURA; GARCIA, 2008; ROMERO *et al.*, 2013). O conhecimento, obtido seja por modelos descritivos ou por modelos preditivos, é utilizado para adaptar o sistema educacional a diferentes contextos de ensino e aprendizagem, ou até mesmo como apoio à tomada de decisão por parte de diferentes *stakeholders*, tais como professores, tutores, projetistas, entre outros.

2.2.2 Sakai e TidiaAe

O *Sakai* (SAKAI, 2014; FARMER; DOLPHIN, 2005; COUNTERMAN *et al.*, 2004) é um sistema desenvolvido por meio da colaboração: *University of Michigan*, *Stanford University*, *Indiana University* e *Massachusetts Institute of Technology*, que iniciaram o desenvolvimento deste sistema para substituir os sistemas desenvolvidos por cada uma delas separadamente. O projeto teve como objetivo desenvolver e distribuir um novo sistema de aprendizagem colaborativo.

O projeto *Sakai* é composto por *softwares* educacionais pré-integrados, ou seja, um grupo de ferramentas *Open Source* a fim de produzir um sistema de aprendizagem colaborativo. Além disso, o *Sakai* é considerado um *framework* descrito sobre o padrão IMS (IMS Global Learning Consortium, 2006; KOPER; OLIVIER; ANDERSON, 2003) e as normas da OKI OSID (OKI, 2008) para prover a integração entre suas camadas e assegurar a portabilidade das ferramentas. Suas principais funcionalidades são (COUNTERMAN *et al.*, 2004): (1) apoio à avaliação e a atividades colaborativas (fórum, grupos de discussão, *chat*); (2) compartilhamento de arquivos; (3) lista de *e-mails*; (4) apoio às atividades gerenciais (gráficos, estatísticas de grupos, classes e estudantes); (5) sistemas de alerta em tempo real. Algumas dessas funcionalidades estão disponíveis para serem acessadas por meio de dispositivos móveis (SAKAI, 2014).

O projeto *Ae* (Aprendizado Eletrônico) dentro do programa *Tidia* (Tecnologia da Informação para o Desenvolvimento da Internet Avançada). O *Tidia* foi proposto com o objetivo de auxiliar atividades de aprendizado eletrônico (TidiaAe Portal, 2008). As funcionalidades da plataforma do *TidiaAe* são voltadas para alunos e professores visando dar apoio ao conteúdo aprendido em aula presencial ou à distância, como mostra a Figura 2 (TIDIA, 2006). Essa plataforma foi desenvolvida para promover a geração de conhecimento, a formação de recursos humanos e a produção científica, tecnológica e de produtos na área de tecnologia da informação aplicada à educação, explorando as

características da Internet avançada apoiada pelo uso de redes de alta velocidade. Nesse sistema, cada pessoa, pode criar um grupo de atividade personalizado que pode conter recursos como: integração de vídeos do Youtube, edição de textos colaborativos e, no caso dos professores da USP, integração com os sistemas de graduação e pós-graduação, o *Júpiter* e o *Janus* (MARTINS, 2009).

Figura 2 – Página principal e algumas ferramentas do *TidiaAe*.

Fonte: Elaborada pelo autor.

O desenvolvimento e implantação do sistema de apoio à aprendizagem eletrônica do projeto *TidiaAe* é desenvolvido em parceria com o projeto *Sakai*, utilizando o núcleo básico do *Sakai* para desenvolver sua plataforma e suas ferramentas colaborativas. Sua arquitetura foi desenvolvida baseada em componentes a fim de facilitar sua construção, implementação, manutenção e evolução, permitindo que novas funcionalidades sejam acrescentadas ao longo do tempo. O processo de gerência de aprendizagem do *TidiaAe* é realizado pelo seguinte conjunto de ferramentas: gerenciador de contexto de aprendizagem, gerenciador de usuário, gerenciador de ações de aprendizagem e gerenciador de ferramentas.

Além das ferramentas de gerência, foram desenvolvidas algumas ferramentas básicas de sistemas de ensino e aprendizagem, tais como: (1) *whiteboard*: ferramenta responsável pela captura de informações do professor durante a aula de forma síncrona, por meio de uma lousa eletrônica; tais informações são posteriormente disponibilizadas aos alunos locais ou remotos, utilizando *tablet PC* ou computadores; (2) *chat*; (3) comunicador instantâneo; (4) laboratório remoto – ferramenta que engloba todo o ciclo de vida de um experimento,

que vai desde a sua criação até a sua efetiva realização pelo usuário; (5) hipertexto; (6) correio eletrônico/*e-mail*; e (7) fórum de discussão.

Ressalta-se, por fim, que o projeto *TidiaAe* envolve pesquisa e desenvolvimento. Nesse sentido, os produtos tecnológicos resultantes foram disponibilizados para serem utilizados por universidades públicas e privadas e por empresas interessadas em treinamento eletrônico, sendo liberados sob a licença GPL (TIDIA, 2006).

2.2.3 Edx

*Edx*⁶ é uma plataforma criada a partir de uma parceria entre a *Harvard University* e o *Massachusetts Institute of Technology* (MIT). A plataforma *Edx*, ilustrada na Figura 3, é uma iniciativa sem fins lucrativos que foi planejada com o objetivo de oferecer aprendizagem *online* para multidões ao redor do mundo.

Figura 3 – Página inicial e área restrita a participantes de um curso oferecido na plataforma *Edx*.

Fonte: Elaborada pelo autor.

Considerando a busca pela educação em massa, a plataforma *Edx* também é caracterizada como uma plataforma robusta capaz de apoiar cursos no formato MOOC (*Massive Open Online Courses*)(GRÜNEWALD *et al.*, 2013). De modo geral, os MOOCs são sistemas educacionais que seguem um fluxo contínuo de aprendizagem, ou seja, possuem uma sequência bem definida de conteúdos educacionais, vídeos, avaliações e atividades interativas.

⁶ *Edx* website: <https://www.edx.org>

O projeto *Edx*, além de buscar oferecer conhecimento de qualidade de forma acessível, também tem como objetivo entender melhor o processo de ensino e aprendizagem. Nesse sentido, a plataforma tem sido avaliada como uma grande ferramenta de pesquisa em aprendizagem.

Vale ressaltar que o *Edx* é um sistema distribuído no forma de *software* livre (*Open Source*) e vem recebendo, nos últimos anos, inúmeras contribuições que estão sendo desenvolvidas em forma de componentes⁷. Sua estrutura foi desenvolvida para participação em massa e, por isso, possui uma enorme quantidade de dados potencialmente úteis à extração de padrões relevantes ao contexto educacional.

2.3 Características de sistemas educacionais

Instituições de ensino buscam cada vez mais explorar o potencial de grandes volumes dados produzidos em seus sistemas educacionais com o objetivo de enfrentar os desafios estratégicos atuais. No entanto, uma das questões que impossibilitam a gestão eficiente do processo educacional por meio de sistemas educacionais tradicionais, como LMS e SIS, é a ausência de padronização dos dados principalmente quando múltiplas fontes precisam ser incorporadas e tratadas de forma correlacionada em um único sistema de apoio à decisão.

Análise Educacional (*Learning Analytics*, *Academic Analytics* e Mineração de Dados Educacionais), descrita no Capítulo 4, refere-se a importantes áreas de pesquisa que pode ser utilizada para identificar maneiras de reduzir custos institucionais ou antecipar informações para a realização de intervenções em tempo real. Desse modo, mudanças significativas devem afetar a utilização de sistemas educacionais para considerar a evoluções em integrações, personalização, análise, entre outros.

Em uma pesquisa realizada por Brown, Dehoney e Millichap (2015), os pesquisadores identificaram que a próxima geração de sistemas educacionais deverá ser capaz de:

1. **Integração e interoperabilidade:** o sistema deverá permitir que diferentes tipos de ferramentas possam ser integradas por meio de extensões ou *plugins*. Isso significa que as próximas gerações de sistemas educacionais devem garantir o intercâmbio de informação em um formato unificado e interoperável. Essa integração deve possibilitar, ainda, dados estruturados que possam ser utilizados para análises.
2. **Personalização:** a personalização é amplamente dependente da interoperabilidade. Mecanismos de interoperabilidade, como os padrões de dados, são os principais

⁷ Repositórios *Edx*: <http://code.edx.org/>

fatores para promover a personalização e melhorias na experiência do usuário. Um ecossistema de aprendizagem, que permite aos alunos e professores atuarem na personalização de seu ambiente, pode ser uma poderosa ferramenta para auxiliar o processo de ensino e aprendizagem.

3. **Análise, recomendação e avaliação da aprendizagem:** grande parte dos sistemas educacionais possuem recursos de *Analytics* e relatórios gerenciais. Esses recursos poderiam ser desacoplados dos sistemas educacionais nativos, aprimorados e facilmente adaptados a qualquer sistema educacional. Essa integração poderia ser feita de forma transparente para os usuários finais. Desse modo, as análises poderiam ser visualizados em *dashboards* dentro do sistema ou utilizando especificações de interoperabilidade do sistema educacional. Para isso, dois aspectos são fundamentais:
 - Ampliação do escopo dos dados: dados demográficos, biográficos e de classificação; atividades no curso (engajamento, tempo de resolução de problemas); artefatos do aluno (*posts* em fórum de discussão). Todos esses tipos de dados devem ser incorporados à abordagem de Análise Educacional adotada. A padronização dos dados são novamente a chave para essa ampliação.
 - Análise de aprendizagem para todas as partes interessadas. A próxima geração de Análise Educacional deve abordar pelo menos os níveis de capacitação dos alunos, aperfeiçoamento instrucional contínuo e supervisão/gestão institucional.
4. **Colaboração:** a tecnologia deve fornecer a alunos e professores novas oportunidades para construir, colaborativamente, caminhos únicos para alcançar os objetivos de aprendizagem. Alunos não devem ser limitados a formar colaborações apenas com seus pares em um curso, ao contrário, eles devem ser motivados a organizar colaborações interinstitucionais, descobrir novos conteúdos e participar de outras comunidades de aprendizagem.
5. **Acessibilidade e *design* universal:** sistemas educacionais podem ajudar instituições a apoiar a aprendizagem de pessoas com deficiência, adotando uma abordagem de *design* universal no contexto de sistemas de aprendizagem digital. Entende-se por *design* universal produtos e sistemas desenvolvidos para serem utilizados por todas as pessoas, na maior extensão possível, sem a necessidade de adaptação ou *design* especializado.

Considerando as características apontadas, o foco deste trabalho, por meio do modelo de referência proposto, é fornecer apoio às três primeiras, a depender do modelo de dados transacional/operacional do sistema educacional. Para isso, o EDRM foi desenvolvido

considerando a integração de diferentes fontes de dados em um único repositório. Além disso, o modelo é capaz de abranger diferentes níveis de granularidade a serem analisados a partir de diferentes perspectivas possibilitando, assim, a personalização e a análise de dados a partir de perspectivas distintas porém correlacionadas.

2.4 Considerações finais

Neste capítulo foram apresentados os principais tipos de sistemas educacionais com foco no aluno, ou seja, sistemas quem visam apoiar o processo de ensino e aprendizagem. Em seguida foi apresentada uma síntese a respeito de alguns LMS, destacando suas principais características e funcionalidades.

Ênfase foi dada em sistemas do tipo LMS devido a sua ampla utilização por instituições de ensino e sua disponibilização como *software* livre. Sistemas educacionais do tipo SIS foram abordados, porém, com menor ênfase em suas características, pois não há uma plataforma amplamente utilizada por instituições.

Um aspecto relevante a ser considerado refere-se às limitações dos sistemas educacionais atuais que impedem a utilização eficiente desses sistemas na gestão e monitoração do processo educacional. Desse modo, sistemas educacionais atuais não são capazes de atuar efetivamente como mecanismo de apoio ao processo de ensino e aprendizagem por meio de integração e interoperabilidade, personalização, análise, entre outros. Alguns desses problemas podem ser minimizados por meio da análise dos dados fornecidos por esses sistemas. Nesse sentido, dados provenientes de sistemas educacionais podem ser considerados grandes e ricas fontes de informação que podem ser utilizadas para a compreensão dos processos de aprendizagem de uma heterogênea população de alunos.

Nesse contexto, a maioria dos LMS descritos possuem diversas funcionalidades semelhantes (ferramentas síncronas e assíncronas, apoio a atividades colaborativas, disponibilização de conteúdo educacional, entre outras). Contudo, a estruturação de dados heterogêneos, provenientes de diferentes fontes de dados não é uma atividade trivial.

Assim, é relevante a definição do Modelo de Referência de Dados Educacionais, projetado como um repositório central a partir de uma arquitetura padrão, capaz de possibilitar acesso rápido a informações de diferentes perspectivas. Desse modo, dados estruturados e integrados de forma consistente, passam a ser importantes instrumentos para auxiliar o processo de tomada de decisão no contexto educacional.

O EDRM foi desenvolvido como um modelo de referência para dados provenientes de sistemas educacionais com o objetivo prover uma estrutura de dados analítica unificada,

integrada e consistente capaz de possibilitar análise correlacionadas de dados de diferentes sistemas educacionais. Seu principal propósito consiste na estruturação dos dados para a disponibilização de informações estratégicas, de forma integrada e consistente, para fins de apoio à decisão durante o processo educacional.

No próximo capítulo será apresentada uma síntese de Análise Educacional. Na Seção 3.2, são definidas as linhas de pesquisa em *Learning Analytics* e *Academic Analytics*. Essas linhas de pesquisa vêm se consolidando como as principais formas de potencializar o processo de tomada de decisão a partir de dados obtidos de sistemas educacionais. Na Seção 3.3, são definidos os principais conceitos de Mineração de Dados Educacionais. São apresentados os métodos mais utilizados e suas principais áreas de aplicação (modelos de conhecimento, comportamento e experiência do usuário, entre outros).

ANÁLISE DE DADOS EDUCACIONAIS

3.1 Considerações iniciais

Realizar a descoberta de conhecimento a partir dos dados educacionais deixou de ser apenas um grande interesse passando a ser prioridade para a maioria das instituições de ensino (BROOKS; THAYER, 2016; ARROWAY *et al.*, 2016). Segundo Arroway *et al.* (2016), os principais fatores que motivam instituições a investir em Análise Educacional são: a alta retenção de alunos, o sucesso acadêmico e a redução do tempo de conclusão. Reduzir os custos e otimizar os recursos institucionais são considerados fatores importantes, porém secundários.

O processo de Análise Educacional envolve a coleta de dados, a organização da informação (algumas vezes considerando múltiplas fontes e formatos de dados) e a análise e manipulação dos dados, de modo que o resultado possa otimizar oportunidades de aprendizagem e resultados educacionais. Considerando a necessidade de investigar técnicas e tecnologias que auxiliem a descoberta de conhecimento a partir de dados educacionais, mais especificamente a investigação e especificação de um modelo de dados de referência (EDRM), vários temas e áreas de pesquisa são consideradas como embasamento teórico deste trabalho, como *Learning Analytics*, *Academic Analytics*, Indicadores de Desempenho e Mineração de Dados Educacionais.

Neste capítulo é apresentada uma revisão dos principais temas de pesquisa nos quais este trabalho está fundamentado. Na Seção 3.2, são definidas as linhas de pesquisa *Learning Analytics* e *Academic Analytics* e formas de definir estratégias de análise a partir de Indicadores de Desempenho. Na Seção 3.3 são sumarizados os conceitos de

Mineração de Dados e seus principais métodos utilizados no contexto de Mineração de Dados Educacionais (predição, agrupamento, mineração de relações, entre outras).

3.2 Analytics

A análise de grandes volumes de dados, também conhecida por *Analytics*, possui um papel fundamental no contexto educacional. Técnicas e tecnologias vêm sendo desenvolvidas com o objetivo de explorar dados de sistemas educacionais a fim de obter informações relevantes ao processo decisório.

Segundo Bichsel (2012), *Analytics* corresponde ao uso de dados, análise estatística e modelos explicativos e preditivos para obtenção de percepções permitindo, ainda, a atuação em questões complexas. Os autores enfatizam que *Analytics* refere-se ao planejamento estratégico e está diretamente envolvida na tomada de decisão.

De acordo com Cooper e Adam (2012), *Analytics* permite realizar avaliação de ações passadas, para estimar o potencial de ações futuras, de modo a possibilitar uma melhor tomada de decisão para a adoção de estratégias mais eficazes tanto para a instituição, como para o indivíduo.

Siemens *et al.* (2011) relatam que, com o rápido desenvolvimento de métodos e ferramentas para realizar análises de grandes volumes de dados (*Big Data*), é possível combinar novos processos de gestão e métricas nas instituições. Quando aplicados ao setor educacional, *Analytics* se divide em duas grandes linhas de pesquisa: *Academic Analytics* (AA) e *Learning Analytics* (LA), conforme ilustrado no Quadro 1.

Em linhas gerais, AA se concentra na maneira pela qual a análise é alavancada para melhorar as práticas de negócios das instituições de ensino, enquanto LA busca melhorar o sucesso dos alunos e seus resultados de aprendizagem (BROOKS; THAYER, 2016). Segundo Arroway *et al.* (2016), ambas compartilham características semelhantes como, requerer dados de qualidade, infraestrutura técnica, *stakeholders* envolvidos e necessidade de utilização efetiva de usuários chave.

A seguir são detalhadas as áreas de LA e AA e sumarizados conceitos de indicadores de desempenho educacional.

3.2.1 Learning Analytics

Segundo Siemens *et al.* (2011), *Learning Analytics* pode ser definido como o processo de coletar, medir e analisar dados de alunos e seus contextos para possibilitar a compreensão e otimização do ambiente de aprendizagem. Seu principal objetivo é investigar

Quadro 1 – Características de *Learning Analytics* e *Academic Analytics*

Tipo de <i>Analytics</i>	Nível	Quem se beneficia?
<i>Learning Analytics</i>	Nível de usuário: análise individual do desempenho em relação aos objetivos de aprendizagem, recursos de aprendizagem e hábitos de estudo de outros colegas.	Aluno, professor, auxiliar/tutor
	Nível de curso: redes sociais, desenvolvimento conceitual, linguística, entre outros.	
	Departamental: modelagem preditiva, padrões de falha/sucesso, entre outros	Administrador, financeiro, vendas e comercialização
<i>Academic Analytics</i>	Institucional: perfil do aluno, desempenho do corpo docente, qualidade do curso e alocação de recursos	Aluno, professor, tutor e demais apoios ao aluno
	Regional: comparações entre processos	Administrador, governo, instituição
	Nacional e Internacional	Governo, organizações, instituição

Fonte: Adaptada de Siemens e Long (2011).

e propor melhorias para o sucesso do aluno. Seus principais atores são: alunos, professores e auxiliares/tutores. Shum (2012) introduziu o conceito de LA para fornecer mecanismos para a identificação de padrões e comportamentos, tanto em nível individual como de grupo. Esse conceito é baseado em três níveis de LA:

1. **Nível micro:** permite realizar análises em nível de **usuário** individual, considerando o contexto que esse usuário está inserido (turma, por exemplo). Nesse nível, análises baseadas em similaridades para possíveis intervenções podem ser realizadas. Análises mais complexas de dados de acesso, colaboração e comunicação também podem ser exploradas, sempre com foco no indivíduo (aluno, professor, etc). O principal benefício desse nível é possibilitar a identificação de alunos em situação de risco e fornecer intervenções, fornecer aos alunos uma visão de seus hábitos de aprendizagem e recomendar melhorias.
2. **Nível meso:** permite realizar análises em nível **institucional**. Nesse nível, são utilizados métodos analíticos de apoio a decisão e realizadas análises preditivas, com foco em obtenção de melhorias administrativas e organizacionais. O principal benefício desse nível é promover melhorias no processo de tomada de decisões administrativas e alocação de recursos organizacionais, possibilitar transparência para criar um melhor entendimento dos sucessos e desafios institucionais, possibilitar análise de impacto a partir de diferentes variáveis proporcionando uma tomada de decisão mais abrangente, aumento da produtividade organizacional por meio de informações atualizadas e respostas rápidas aos problemas.

3. **Nível macro:** permite o compartilhamento de dados, a fim de possibilitar análises **interinstitucionais**. Nesse nível, é comum realizar *benchmarking*, ou seja, comparações entre práticas, serviços e produtos educacionais em diferentes escalas geográficas (internacional, nacional, estadual, regional, etc). O principal benefício desse nível é possibilitar novas definições e atualizações de modelos acadêmicos e abordagens pedagógicas.

Siemens e Long (2011) fazem uma distinção entre AA e LA, conforme descrito no Quadro 1, que pode ser traduzida nos níveis de LA propostos por Shum (2012). De modo geral, AA equivale ao níveis **macro** e **meso** de Shum (2012) e LA equivale a análises de nível **micro**. Na visão da Siemens e Long (2011), LA se concentra explicitamente no processo de aprendizagem em busca do sucesso do aluno.

Já Barneveld, Arnold e Campbell (2012) definem LA como análises com foco no aluno e seu comportamento durante a aprendizagem, coletando dados do gerenciamento de cursos e informações do aluno, a fim de gerenciar o sucesso do aluno.

Chatti *et al.* (2014), com o objetivo de descrever os requisitos de uma plataforma eficaz de LA, definiram um modelo de referência para LA baseado em quatro dimensões, conforme Figura 4. As quatro dimensões (contextos, *stakeholders*, métodos e objetivos), foram definidas a partir de publicações em LA e áreas correlatas.

Figura 4 – Modelo de Referência para o Processo de *Learning Analytics*.

Fonte: Adaptada de Chatti *et al.* (2014).

Contexto

Essa dimensão tem por objetivo definir **quais tipos de dados o sistema educacional armazena, gerencia e utiliza para análises**. Para isso, LA utiliza como principais fontes de dados sistemas acadêmicos (SIS) e/ou sistema de gestão de aprendizagem (LMS). Conforme detalhado no Capítulo 2, esses sistemas possuem grandes volumes de dados heterogêneos e com inúmeras características inerentes ao contexto educacional que podem ser exploradas para apoiar o processo de tomada de decisão.

Stakeholders

Essa dimensão tem por objetivo identificar **quem é o alvo da análise**. Em linhas gerais, LA pode ser orientada para diversos *stakeholders* como alunos, professores, tutores, instituições educacionais, administradores, tomadores de decisões e pesquisadores. Cada *stakeholder* possui características importantes e tipos de análises distintas. Um exemplo são as análises para professores que buscam melhorar sua eficiência em práticas de ensino.

Instituições de ensino podem fazer o uso de LA para apoiar a tomada de decisão, melhorar métricas de sucesso acadêmico, identificar riscos e oportunidades de melhoria, ajustar o planejamento financeiro e determinar necessidades de contratação. Para que LA seja utilizada de forma efetiva é necessário, ainda, considerar questões éticas principalmente associadas a privacidade. Para isso, é preciso determinar limites de análises e preservar informações confidenciais dos *stakeholders*.

Objetivos

Essa dimensão busca determinar **por que são coletados e analisados os dados de sistemas educacionais** considerando, por exemplo, os seguintes objetivos (BARNEVELD; ARNOLD; CAMPBELL, 2012):

- **Monitoramento e análise:** busca rastrear as atividades dos alunos e gerar relatórios para apoiar a tomada de decisões de professores, instituição, entre outros. O monitoramento também pode estar associado ao projeto instrucional, por meio da avaliação do processo de aprendizagem, para a obtenção da melhoria contínua. Adicionalmente, busca-se analisar como o progresso do aluno ocorre, com o objetivo de ajudar professores a realizar adequações no projeto instrucional.
- **Predição e intervenção:** tem por objetivo identificar padrões para realizar predições acerca do desempenho do aluno a partir de dados educacionais. Os modelos preditivos podem ser usados para gerar iniciativas de intervenção em relação aos

alunos. Uma análise efetiva de predição do desempenho pode apoiar professores ou instituições em intervenção, sugerindo ações que devem ser tomadas para melhorar o desempenho do aluno.

- **Acompanhamento e aconselhamento:** tem por objetivo orientar alunos a iniciar ou continuar módulos de aprendizagem, bem como instruções de áreas ou matérias específicas em um curso. O acompanhamento vai além da tutoria e se concentra no apoio ao aluno durante todo o processo, ou seja, durante o tempo em que tanto o tutor/mentor quanto o aluno fazem parte da mesma organização. Como parte desse apoio, os tutores fornecem orientação no planejamento de carreira, supervisionam a realização de metas e ajudam a preparar novos desafios.
- **Avaliação e *feedback*:** tem por objetivo apoiar a avaliação ou auto-avaliação para melhorar a eficiência e eficácia do processo de aprendizagem. Desse modo, é possível realizar um *feedback* inteligente, com informações personalizadas geradas a partir dos dados educacionais.
- **Adaptação:** tem por objetivo mostrar aos alunos o que fazer em seguida, ou seja, a organização das atividades, recursos de aprendizagem e atividades instrucionais de acordo com as necessidades individuais do aluno.
- **Personalização e recomendação:** tem por objetivo apoiar os alunos na tomada de decisões a respeito de sua própria aprendizagem e, continuamente, personalizar seu ambiente de aprendizagem (LMS). A personalização sugere mudanças no modelo de aprendizagem passando de conhecimento-empurrado (*knowledge-push*) para conhecimento-puxado (*knowledge-pull*). No modelo de aprendizagem baseado no conhecimento-empurrado, o fluxo de informação é dirigido pela instituição/professor. No modelo de aprendizagem baseado em conhecimento-puxado, o aluno buscar o conhecimento por meio da navegação no sistema educacional. No entanto, uma preocupação associada à abordagem conhecimento-puxado é a possibilidade de sobrecarga de informações. Sob essa perspectiva, torna-se essencial a utilização de sistemas de recomendação para apoiar a aprendizagem auto-dirigida por conhecimento-puxado. O objetivo do LA nesse caso, é ajudar o aluno a decidir o que fazer em seguida, recomendando conhecimentos por meio de recursos de aprendizagem ou pessoas, a partir de atividades e navegações similares.

Esses objetivos mencionados são muitas vezes difíceis de mensurar e, por isso, precisam de um conjunto de **indicadores de desempenho** também conhecidos como KPI (*Key Performance Indicator*), bem como suas **métricas** e **medidas**. Segundo Martin,

Sauvageot e Tchatchoua (2011), os indicadores de desempenho podem atuar como o meio pelo qual tomadores de decisão e administradores podem definir precisamente os resultados que esperam de suas estratégias e planos.

Chatti *et al.* (2014) salientam que o grande desafio nessa dimensão é definir simultaneamente **objetivo**, acompanhados de seus respectivos indicadores, métrica e medidas antes de iniciar análises em LA. Frente a esse desafio, na próxima seção, é apresentada uma descrição detalhada de indicadores de desempenho educacional com as definições de suas principais características e formas de utilização.

Métodos

O objetivo dessa dimensão é definir **como o sistema realiza a análise dos dados coletados**. Em LA é possível aplicar diferentes técnicas para detectar padrões potencialmente úteis por meio de diferentes métodos, como (BARNEVELD; ARNOLD; CAMPBELL, 2012):

- **Estatística:** a maioria dos LMS possuem ferramentas de relatórios que oferecem as estatísticas básicas da interação do aluno com outros *stakeholders* ou com o próprio sistema educacional. Exemplos desses relatórios são as estatísticas de uso (como o tempo *online*), número de visitas por página, número total de visitas, distribuição de visitas ao longo do tempo, frequência de postagens, percentual de material acessado, entre outros.
- **Visualização da informação:** análises estatísticas em forma de tabelas nem sempre são de simples compreensão. Representar visualmente os resultados de forma que sua interpretação seja simples e intuitiva é um grande desafio. Segundo Mazza (2009), devido à nossa capacidade de interpretação visual, muitas vezes a representação visual é mais eficaz quando comparada a textos, tabelas, entre outras. Reconhecendo o poder das representações visuais, os relatórios tradicionais, cada vez mais, estão sendo substituídos por *dashboard* que mostram representações gráficas de diferentes tipos de indicadores.
- **Mineração de dados:** também conhecida como descoberta de conhecimento em base de dados (*Knowledge Discovery in Databases* – KDD), é a descoberta de padrões potencialmente úteis a partir de bancos de dados, textos, imagens, Web, entre outros. Em geral, diferentes métodos de Mineração de Dados (MD) podem ser utilizados no contexto educacional. Na Seção 3.3, é detalhada a utilização de MD para melhorar o processo de ensino e aprendizado, também conhecida como Mineração de Dados Educacionais (MDE).

Em síntese, o que se pretende evidenciar é a abrangência de pesquisas em LA utilizando dados educacionais heterogêneos, em diferentes formatos e para atender inúmeros objetivos e *stakeholders*.

3.2.2 Academic Analytics

Academic Analytics (AA), também conhecida por *Institutional Analytics* (IA), tem por objetivo promover melhorias em processos organizacionais, fluxos de trabalho, alocação de recursos e medição institucional por meio da utilização dos dados acadêmicos e institucionais. AA é semelhante à análise de negócios pois busca realizar melhorias no processo organizacional. Seus principais atores são administradores, gestores, instituição, governo, organizações, entre outros (BROOKS; THAYER, 2016; SIEMENS, 2013).

Modelo de maturidade de Academic Analytics

Considerando a crescente busca por *Analytics* e a necessidade de medir sua performance institucional, Brooks e Thayer (2016) definiram um modelo de maturidade para avaliar a capacidade de análise das instituições. Esse modelo possui 32 itens avaliados que classificam as seguintes dimensões:

1. **Investimento e recursos:** avalia a capacidade da institucional de investir em recursos especializados, profissionais qualificados, treinamento de usuários chaves, entre outros.
2. **Cultura e tomada de decisão:** avalia a eficiência institucional, principalmente de seus líderes, de tomar decisão baseada em dados.
3. **Eficiência dos dados:** avalia se a instituição possui dados de qualidade, ou seja, dados que viabilizam a comparações entre áreas de uma mesma instituição ou entre instituições.
4. **Infraestrutura técnica:** avalia se a instituição possui *software*, principalmente para armazenamento, gestão e análise dos dados.
5. **Envolvimento de instituições de pesquisa com a qualidade institucional:** avalia se existe uma comunicação eficaz entre o departamento de tecnologia e a área de pesquisa institucional.
6. **Políticas:** avalia se são implementadas práticas de segurança da informação e políticas de privilégios de acesso ao dados institucionais e individuais.

As dimensões do modelo de maturidade são avaliados em uma escala, com atributos que variam de 1 (ausente ou *ad hoc*) a 5 (otimizado), conforme descrito no Quadro 2.

Quadro 2 – Índice de maturidade de *Academic Analytics*.

Índice de Maturidade	Significado
1 – Ausente ou <i>ad hoc</i>	Atualmente, não temos essa capacidade, ou tratamos de forma improvisada e irregular.
2 – Repetitivo	Temos uma capacidade estabelecida, mas as nossas práticas são na sua maioria informais.
3 – Definido	Temos uma capacidade padronizada e temos procedimentos documentados e / ou responsabilidades relacionadas a ele.
4 – Gerenciado	Gerenciamos essa capacidade para obter resultados previsíveis com base em indicadores de desempenho medidos de forma confiável.
5 – Otimizado	Além de medir o desempenho, reavaliamos regularmente a forma como oferecemos essa capacidade, a fim de melhorar as práticas e gerenciar riscos.

Fonte: Adaptada de Brooks e Thayer (2016).

Um dos aspectos principais da aplicação de um índice de maturidade para avaliar o programa analítico de uma instituição é que ele possa ajudar a identificar as áreas gerais e específicas que podem ser direcionadas para melhoria.

A partir da utilização do modelo de maturidade por meio de exames individuais, Brooks e Thayer (2016) identificaram que, em geral, a capacidade analítica das instituições de ensino nos últimos anos ainda é inferior ao desejável estabelecido pelo modelo de maturidade. Observou-se ainda que o nível médio de maturidade analítica das instituições situa-se entre **3 – Definido** e **4 – Gerenciado**, ou seja, as instituições possuem um nível moderado de maturidade caracterizado por formatos padronizados, processos documentados e responsabilidades claramente delineadas. Desse modo, para alcançar o próximo nível de maturidade, é necessário que as instituições sejam capazes de gerenciar a capacidade de obtenção de resultados previsíveis a partir da medição de indicadores de desempenho (*Key Performance Indicator* – KPI).

Em síntese, tanto *Learning Analytics* como *Academic Analytics* são linhas de pesquisa emergentes que podem ser complementares quando se deseja abranger vários níveis de granularidade (usuário, curso, departamento, instituição, região, país) e diferentes *stakeholders*. Entretanto, uma limitação comum em pesquisa em *Learning Analytics* e *Academic Analytics* está na ausência de metodologia para definir modelos analíticos consistentes. Desse modo, pesquisas que definem processos de referência e níveis de maturidade vêm sendo importantes nesse sentido. Contudo, ainda há limitações quando busca-se analisar o contexto educacional de forma estratégica utilizando dados analíticos

para auxiliar o processo de tomada de decisão. Os indicadores de desempenho educacional inserem-se nesse contexto.

3.2.3 Indicadores de desempenho educacional

Indicadores de desempenho representam um conjunto de medidas centradas no desempenho organizacional que são utilizadas por tomadores de decisão e administradores com o objetivo de analisar precisamente os resultados esperados de suas estratégias e planos. Parmenter (2015), define indicadores de desempenho como um instrumento de gestão essencial às atividades de monitoramento e avaliação das organizações, assim como seus projetos, programas e políticas, pois permitem acompanhar o alcance das metas, identificar avanços, melhorias de qualidade, correção de problemas, necessidades de mudança.

É importante destacar que indicadores de desempenho, métricas e medidas são conceitos distintos (ECKERSON, 2007). Indicadores de desempenho devem refletir direcionadores ou *drivers* de valor estratégico. São geralmente compostos por índices, faixa de valores, frequências, comparativos, entre outros. Métricas representam medidas de uma atividade tomadas ao longo de um período, que servem de subsídio aos indicadores. Um exemplo de métrica é a *quantidade de alunos por ano*, enquanto que um indicador de desempenho pode ser representado pelo *percentual de novos alunos que ingressaram por vestibular por ano*.

Indicadores de desempenho sempre devem estar associados a um objetivo ou meta estratégica previamente estabelecidas. Países em desenvolvimento possuem políticas e metas educacionais como parte essencial do seu planejamento estratégico (MARTIN; SAUVAGEOT; TCHATCHOUA, 2011).

Entidades gestoras e governos estão exigindo cada vez mais a transparência em processos administrativos e educacionais de instituições de ensino. Essas exigências possuem como ponto de partida um planejamento estratégico, com desdobramentos em demonstração de desempenho por meio de indicadores. Desse modo, a definição de sistemas de indicadores tornou-se um instrumento de gestão indispensável para instituições de ensino (MARTIN; SAUVAGEOT; TCHATCHOUA, 2011).

Indicadores de desempenho educacional são desenvolvidos para atribuir valor estatístico à qualidade do ensino, considerando não somente o desempenho dos alunos, mas também, o contexto econômico e social em que as instituições estão inseridas.

Os indicadores são úteis principalmente no monitoramento dos sistemas educacionais, considerando o acesso, a permanência e a aprendizagem dos alunos. Dessa forma,

contribuem para a criação de políticas públicas voltadas à melhoria da qualidade da educação e dos serviços oferecidos à sociedade por meio das instituições (INEP, 2015).

Em linhas gerais, um conjunto de indicadores deve revelar a existência de problemas e facilitar a sua medição. Isso acontece devido à necessidade de realizar cálculos de indicadores a partir de dados brutos utilizando análises estatísticas como porcentagens, taxas, proporção, índices entre outros.

Para realizar análises abrangentes, recomenda-se que um sistema de indicadores inclua alguns dados brutos em grande escala, como o *total da população estudantil por nível ou disciplina*, que pode servir de base para o cálculo de indicadores tais como *taxas de crescimento ou porcentagens de desagregação da população estudantil em diversas categorias* (PARMENTER, 2015).

Para definir um indicador, é necessária uma visão geral simples e precisa de todo o sistema educacional. É necessário, também, apresentar dados relativos a vários anos uma vez que determinados aspectos de um sistema podem ser observados apenas ao longo de um determinado período de tempo.

Um sistema de indicadores permite a criação de transparência, possibilitando escolhas objetivas e eficiência na alocação e uso eficiente de recursos. Desse modo, é possível medir o desempenho e controlar a qualidade dos serviços educacionais podendo ter diversos propósitos no contexto educacional. Martin, Sauvageot e Tchatchoua (2011), identificou três formas de utilização de sistemas de indicadores:

- **Sistemas de indicadores para informação pública:** tem por objetivo informar o público em geral sobre o estado do sistema. Possui informações sobre tendência da população acadêmica (novas matrículas, gênero, etc), formação, diplomas e certificados, recursos humanos (pesquisadores, docentes, etc), recursos financeiros e até mesmo performance em pesquisa científica.
- **Sistemas de indicadores para monitoramento de políticas:** tem por objetivo monitorar o andamento de uma estratégia ou o progresso de um planejamento que foi implementado. Possui informações que podem ser utilizadas para monitorar a implementação de uma política ou plano de ensino, fornecer informações confiáveis sobre o desenvolvimento de setores e permitir que as instituições avaliem o seu próprio desempenho (excelência em pesquisa, ensino e aprendizagem, entre outros).
- **Sistemas de indicadores para a gestão:** tem por objetivo contribuir para o desenvolvimento estratégico e fornecer informações básicas para a tomada de decisão. Possui informações essenciais para gestores e não ao público em geral. Pode ser

correlacionado com questões sócio-econômicos como, por exemplo, adaptar o ensino superior às necessidades do mercado de trabalho.

A análise de grupos de indicadores permite explicar fatores como relacionamentos de causa-efeito em sistemas educacionais, com o objetivo de preservar a transparência. No entanto, essas interpretações são bastante complicadas e, por isso, é importante que o grupo selecionado de indicadores permita uma variedade de pontos de vista.

Sob essa perspectiva, as principais características de um bom indicador de desempenho são (PARMENTER, 2015): (1) relevância; (2) capacidade de sumarizar a informação sem que ela seja distorcida; (3) natureza estruturada e multifacetada, permitindo a sua ligação com outros indicadores, resultando em análises transversais do sistema; (4) precisão e comparabilidade; e (5) confiabilidade e exatidão.

Essas características têm por objetivo proporcionar um meio de (PARMENTER, 2015): (1) medir um trabalho a ser feito a fim de atingir um determinado objetivo; (2) identificar situações problemáticas ou incertas; (3) responder a preocupações políticas e a questões que levaram à sua escolha como indicador; e (4) comparar seu valor atual com o valor de referência, valor padrão, ou seu próprio valor em diferentes períodos observados.

É importante destacar que definir um conjunto de indicadores considerando tais características não é uma atividade trivial. Entretanto, essa é a única forma de fornecer um instrumento de gestão eficiente para os decisores e um meio de compreender a situação atual do processo educacional.

No Brasil, os indicadores de desempenho de Instituições de Ensino Superior (IES) são medidos a partir de três indicadores de qualidade: (1) desempenho do aluno – o conceito obtido a partir dos resultados do ENADE; (2) cursos superiores – Conceito Preliminar de Curso (CPC); e (3) instituições de educação superior – Índice Geral de Cursos Avaliados da Instituição (ICG) (ALMEIDA *et al.*, 2016). Segundo INEP (2015), com esses três indicadores é possível garantir o desenvolvimento de políticas públicas, informações e orientações para a sociedade, bem como o desenvolvimento do Plano de Desenvolvimento Institucional (PDI) e revisão do planejamento de IES.

Uma importante pesquisa realizada pelo INEP refere-se ao Censo da Educação Superior. Essa pesquisa busca coletar informações relevantes cerca das IES com o objetivo de oferecer à comunidade acadêmica e à sociedade em geral informações detalhadas sobre a situação e as grandes tendências do setor. O Censo da Educação Superior reúne informações sobre as instituições de ensino superior, seus cursos de graduação presencial ou a distância, cursos sequenciais, vagas oferecidas, inscrições, matrículas, ingressantes e concluintes e

informações sobre docentes nas diferentes formas de organização acadêmica e categoria administrativa.

Nesse contexto, todas as instituições e cursos participantes do Censo da Educação Superior são registrados no MEC para que seja possível a coleta de dados para compor as estatísticas básicas. Esses dados são disponibilizados pelo InepData¹ em forma de Sinopse Estatística e Microdados. Uma síntese dos indicadores de desempenho definidos para o Censo da Educação Superior é descritas a seguir (INEP, 2015; INEP, 2016):

- Número de instituições de educação superior, por organização acadêmica (federal, estadual, municipal) e localização (capital e interior);
- Número de cursos, matrículas, concluintes, vagas oferecidas, candidatos inscritos e ingressos em cursos de graduação - presenciais e a distância, por organização acadêmica e grau acadêmico;
- Total de matrículas em cursos de graduação - presenciais e a distância, por sexo, organização acadêmica e grau acadêmico;
- Total de concluintes em cursos de graduação presenciais e a distância, por sexo, organização acadêmica e grau acadêmico;
- Total de ingressos por processo seletivo e outras formas nos cursos de graduação presenciais e a distância, por organização acadêmica e grau acadêmico;
- Número de ingressos por processo seletivo nos cursos de graduação presenciais e a distância;
- Número de ingressos por outras formas nos cursos de graduação presenciais e a distância;
- Matrículas nos cursos de graduação presenciais e a distância, por faixa etária;
- Matrículas de alunos portadores de necessidades especiais nos cursos de graduação presenciais e a distância, por tipo de necessidade especial;
- Matrículas nos cursos de graduação presenciais e a distância, por raça/etnia;
- Total de matrículas (alunos matriculados, matrícula trancada, matrícula desvinculada do curso, transferência para outro curso, falecido) nos cursos de graduação presenciais e a distância, por nacionalidade e procedência do aluno;
- Dados gerais dos cursos de graduação presenciais e a distância, por categoria administrativa das IES;

¹ Website InepData: <http://portal.inep.gov.br/inep-data>

- Número total de docentes (em exercício e afastados), por organização acadêmica e grau de formação;
- Número total de docentes (em exercício e afastados), por organização acadêmica e sexo;
- Número total de docentes em exercício, por organização acadêmica e regime de trabalho;
- Número total de docentes em exercício, por organização acadêmica e grau de formação;
- Número total de docentes em exercício, por organização acadêmica e sexo;
- Número total de funcionários técnico-administrativos (em exercício e afastados), por organização acadêmica e grau de formação;
- Número total de funcionários técnico-administrativos (em exercício e afastados), por organização acadêmica e sexo;
- Relação matrículas dos cursos de graduação presenciais/função docente em exercício, por organização acadêmica;
- Número de cursos de graduação presenciais, por organização acadêmica e localização (capital e interior);
- Número de cursos de graduação presenciais, por organização acadêmica e categoria administrativa das IES;
- Número total de ingressos nos cursos de graduação presenciais, por organização acadêmica e todas as formas de ingressos
- Número de vagas oferecidas, candidatos inscritos e ingressos por vestibular e outros processos seletivos, por organização acadêmica;
- Número total de ingressos em cursos de graduação presenciais, por organização acadêmica , grau acadêmico (bacharelado, licenciatura, tecnólogo e não aplicável) e sexo;
- Relação candidatos inscritos/vaga oferecida e ingressos/vagas oferecidas, nos cursos de graduação presenciais, por organização acadêmica;
- Matrículas em cursos de graduação presenciais, por organização acadêmica e localização (capital e interior);
- Matrículas em cursos de graduação presenciais, por organização acadêmica e categoria administrativa das IES;
- Matrículas em cursos de graduação presenciais, por organização acadêmica e turno;
- Matrículas em cursos de graduação presenciais, por organização acadêmica e sexo;

- Matrículas total em cursos de graduação presenciais, por sexo, organização acadêmica e grau acadêmico (bacharelado, licenciatura, tecnólogo e não aplicável);
- Número de concluintes em cursos de graduação presenciais, por organização acadêmica e localização (capital e interior);
- Número de concluintes dos cursos de graduação presenciais, por organização acadêmica e categoria administrativa das IES;
- Número de concluintes em cursos de graduação presenciais, por organização acadêmica e turno;
- Número de concluintes em cursos de graduação presenciais, por organização acadêmica e sexo;
- Número de concluintes total em cursos de graduação presenciais, por organização acadêmica, grau acadêmico (bacharelado, licenciatura e tecnólogo) e sexo;
- Número de concluintes, número de cursos e matrículas nos cursos de graduação presenciais e a distância;
- Número de vagas oferecidas, candidatos inscritos, relação candidatos inscritos/vaga oferecida e ingressos por processos seletivos, nos cursos de graduação presenciais e a distância;
- Número de polos, ingressos totais, ingressos por processos seletivos, matrículas e concluintes, nos cursos de graduação a distância, por organização acadêmica;
- Número de cursos de graduação a distância, por organização acadêmica e categoria administrativa das IES;
- Número de vagas oferecidas, candidatos inscritos e ingressos por processos seletivos, nos cursos de graduação a distância, por organização acadêmica;
- Matrículas dos cursos de graduação a distância, por organização acadêmica e categoria administrativa das IES;
- Número de concluintes dos cursos de graduação a distância, por organização acadêmica e categoria administrativa das IES;
- Número de cursos e matrículas nos cursos sequenciais presenciais e a distância;
- Número de vagas oferecidas, candidatos inscritos, relação candidatos inscritos/vaga oferecida e ingressos por vestibular e outros processos seletivos, nos cursos sequenciais presenciais e a distância;

- Número de cursos, vagas oferecidas, candidatos inscritos, ingressos por vestibular e outros processos seletivos, matrículas e concluintes, nos cursos sequenciais de formação específica presenciais, por organização acadêmica;
- Número de polos, ingressos totais, ingressos por vestibular e outros processos seletivos, matrículas e concluintes, nos cursos sequenciais de formação específica a distância, por organização acadêmica;
- Número total de docentes (em exercício e afastados), por grau de formação e regime de trabalho;
- Número total de funcionários técnico-administrativos, por grau de formação;
- Números de concluintes, cursos, matrículas nos cursos de graduação presenciais e a distância e docentes;
- Número de vagas oferecidas, candidatos inscritos, relação candidatos inscritos/vagas oferecidas e ingressos por vestibular e outros processos seletivos, nos cursos de graduação presenciais e a distância.

Para a proposição do EDRM, alguns desse indicadores de desempenho disponibilizados pelo INEP foram considerados para a derivação de atributos importantes ao processo decisório. Um exemplo está no indicador *número de concluintes total ou taxa de aprovação* que se refere a uma habilidade acadêmica do aluno. Esse indicador possui várias informações que podem ser utilizadas na composição do modelo, tais como: **aluno**, **nota**, **conceito**, **tempo** e **curso**.

Basicamente, a **nota final do aluno** e/ou **conceito** são as informações que podem ser consideradas para realizar o cálculo da taxa de aprovação, ou seja, são atributos de **notas**.

Tempo pode ser uma variabilidade desse indicador, por exemplo, *taxa de aprovação por ano ou por semestre*. Nesse caso **ano** e **semestre** são atributos de **tempo**.

Curso pode ser uma variabilidade desse indicador, por exemplo, *taxa de aprovação por curso, por departamento ou por instituição*. Nesse caso **curso**, **departamento** e **instituição** são atributos de **curso**.

Por fim, outras variabilidades podem ocorrer considerando atributos com características diferentes, porém correlacionados, como *taxa de aprovação no curso X, por modalidade Y, por gênero e por ano*. Nesse caso, um único indicador possui as perspectivas de **tempo** (ano), **curso** (nome/identificador do curso e modalidade) e **aluno** (nome/identificador do aluno e gênero).

Na próxima seção é descrita a Mineração de Dados Educacionais (MDE), uma importante linha pesquisa em Análise Educacional que está na intersecção de *Learning Analytics* e *Academic Analytics*.

3.3 Mineração de dados educacionais

A Mineração de Dados Educacionais (MDE) tem como objetivo converter os dados brutos provenientes de sistemas educacionais em informações úteis que podem causar algum impacto no processo educacional. Tal conversão é possível por meio da aplicação do processo de Mineração de Dados (MD) sobre os dados educacionais (PEÑA-AYALA, 2014; ROMERO; VENTURA, 2007).

O processo de MDE pode ser entendido como uma particularidade dos processos de MD. Em linhas gerais, o processo pode ser composto por uma mesma sequência de etapas que compõem a MD. Entretanto, as particularidades do domínio educacional trazem algumas diferenças à MDE.

Romero e Ventura (2010) destacam que os dados provenientes de sistemas educacionais possuem características únicas quando comparadas a outros domínios. Suas diferenças estão relacionadas aos **objetivos**, ao **tipo de dados** e aos **métodos utilizados**.

O **objetivo** da utilização do processo de MD em cada aplicação é diferente. Por exemplo, em MDE há várias iniciativas de pesquisas e ferramentas que buscam a melhoria do processo de aprendizagem, seja analisando a aprendizagem do aluno ou, ainda, observando diretamente o processo educacional. Esses objetivos são, por vezes, difíceis de quantificar e exigem seu próprio conjunto especial de técnicas e métricas.

Em sistemas educacionais, existem muitos **tipos diferentes de dados** disponíveis, como dados sobre alunos, disciplinas e professores, dados resultantes interações, entre outros. Esses dados são específicos para a área educacional e, portanto, têm informações e relações com outros dados considerando vários níveis de granularidade.

Em relação aos **métodos utilizados**, em sistemas educacionais busca-se resolver problemas com características específicas que requerem um tratamento diferente. Embora a maioria dos métodos tradicionais aplicados nas diferentes etapas de MD possam ser aplicados diretamente, existem problemas em que tais métodos precisam ser adaptados.

3.3.1 Mineração de dados

A Mineração de Dados (MD) é um processo de análise e descoberta de dados que, sob limitações aceitáveis de eficiência computacional, produz uma determinada enumeração

de padrões a respeito dos dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a). Esse processo pode ser automático ou semi-automático e os padrões descobertos devem ser novos, representativos, não triviais e resultar em alguma vantagem (WITTEN; FRANK; HALL, 2011).

Etapas da mineração de dados

Para efetuar o processo de MD é necessário, inicialmente, identificar os problemas que se deseja resolver e, com isso, definir o objetivo do processo. Além disso, é importante verificar quais os dados estão disponíveis ou, ainda, a possibilidade de obter os dados necessários para alcançar o objetivo.

A fim de atingir esses objetivos, o processo de MD é dividido em etapas a serem cumpridas. Existem, na literatura, diversas abordagens para a divisão dessas etapas. Fayyad, Piatetsky-Shapiro e Smyth (1996b) dividem o processo em nove etapas. Já para Weiss e Indurkha (1998), o processo de MD é dividido em quatro etapas. Em uma revisão mais recente, entretanto, Rezende *et al.* (2003) consideram a divisão do processo em cinco grandes etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento. A Identificação do Problema pode ser vista como uma etapa anterior ao processo computacional de MD, que se refere ao conhecimento do domínio a ser explorado. Da mesma forma a Utilização de Conhecimento é uma etapa posterior ao processo computacional de MD, que refere-se à utilização do conhecimento obtido, conforme ilustrado na Figura 5.

Neste trabalho, optou-se por adotar a definição proposta por Rezende *et al.* (2003) por considerar que o processo de MD, em cinco etapas, possui uma sequência de etapas concisa e bem definida.

1 – Identificação do problema

Nessa etapa, delimita-se o escopo do problema com o qual irá trabalhar, definindo o objetivo ao aplicar do processo de MD. De modo geral, o analista define a base os dados com a qual irá trabalhar, o que se espera obter com a análise dos dados e como o resultado da análise pode ser utilizado no contexto de sua organização (REZENDE *et al.*, 2003).

Conforme Rezende *et al.* (2003), durante a identificação dos problemas é relevante: (1) estabelecer os critérios de desempenho; (2) estabelecer a relação entre simplicidade e precisão do conhecimento extraído; e (3) definir se o conhecimento resultante deve ser compreensível a seres humanos ou se um modelo tipo caixa preta é suficiente.

Figura 5 – Etapa do processo de mineração de dados.

Fonte: Rezende *et al.* (2003).

As decisões tomadas nesse ponto devem guiar as próximas etapas e poderão ter reflexo no desempenho da aplicação. Nesse sentido, com o auxílio de especialistas no domínio, uma atividade imprescindível a ser realizada nessa fase é um estudo sobre do domínio do problema para se adquirir um conhecimento inicial, visando o auxílio na tomada de decisão nas etapas subsequentes do processo (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996a).

2 – Pré-processamento

A etapa de pré-processamento refere-se à aplicação de várias técnicas para a captação, organização, tratamento e preparo dos dados. Nessa etapa, busca-se melhorar a qualidade dos dados efetuando ajustes e formatando esses dados para que possam servir como entrada a métodos de extração do conhecimento.

O pré-processamento pode ser considerado um processo semi-automático, pois entende-se que essa etapa depende da capacidade do analista de dados em identificar os problemas presentes nos dados e, posteriormente, selecionar e utilizar os métodos mais adequados para solucionar cada um dos problemas.

No trabalho de Batista (2003), as tarefas realizadas por métodos de pré-processamento foram classificadas em dois grupos: fortemente dependente de domínio e fracamente dependente de domínio.

As tarefas **fortemente dependentes de domínio** são aquelas que podem ser efetivamente realizadas com o uso de conhecimento específico ao domínio. Métodos automáticos podem ser utilizados para realizar uma tarefa fortemente dependente de domínio. Como principais tarefas fortemente dependente de domínio pode-se citar: identificação de inconsistência, identificação de poluição, verificação de integridade, identificação de atributos duplicados e redundantes, atributos com valores *defaults*.

As tarefas **fracamente dependente de domínio** são aquelas que podem ser realizadas por métodos que extraem dos próprios dados as informações necessárias para tratar o problema de pré-processamento de dados. Vale ressaltar que embora essa tarefa ainda dependa de um conhecimento de domínio, pois é necessário selecionar um método correto para solucionar o problema, é possível utilizar métodos mais automáticos do que aqueles aplicados em tarefas fortemente dependentes de domínio. Dentre as principais tarefas fracamente dependentes de domínio estão: tratamento de valores desconhecidos, identificação e descrição de valores extremos, tratamento de conjuntos de dados com classes desbalanceadas, seleção de atributos e construção de atributos.

De uma maneira geral, as atividades de pré-processamento podem ser agrupadas em sub-fases: extração e integração, transformação, limpeza, seleção e redução dos dados (REZENDE *et al.*, 2003; HAN; KAMBER; PEI, 2006).

- **Extração e integração:** os dados disponíveis podem estar em diferentes fontes, como arquivos texto, arquivos estruturados (XML entre outros), banco de dados relacional ou orientados a documento (não relacional), entre outros. Desse modo, é necessário a obtenção desses dados e sua unificação com o objetivo de formar uma única fonte de dados que possa ser utilizada como entrada para algoritmos de extração de padrões.
- **Transformação:** nessa fase o objetivo é transformar a representação dos dados a fim de superar quaisquer limitações existentes nos algoritmos para a realização da extração de padrões. De modo geral, a decisão de quais transformações são necessárias depende do algoritmo a ser utilizado. Algumas transformações comuns que podem ser aplicadas aos dados são: (1) resumo, por exemplo, quando os dados sobre acessos do aluno ao LMS são agrupados para formar resumos diários; (2) transformação de tipo, por exemplo, quando um atributo do tipo data é transformado em outro tipo para que o algoritmo de Extração de Padrões possa utilizá-lo adequadamente; e (3) normalização de atributos contínuos, que consiste em transformar os valores dos atributos, de seus valores originais para um conjunto de valores específico (por exemplo $[-1,1]$ ou $[0,1]$). Esse tipo de transformação é importante principalmente para

métodos que utilizam medidas de distância entre atributos, como por exemplo KNN (K-vizinhos mais próximos) ou em algoritmos de classificação como redes neurais.

- **Limpeza:** os dados disponíveis podem apresentar problemas provenientes do processo de coleta de dados. Esses problemas podem ser desde erros de digitação, até erro na leitura dos dados pelos sensores. Considerando que esses dados serão utilizados na extração de padrões como apoio à tomada de decisão, é extremamente importante que os dados sejam de qualidade. A limpeza dos dados pode ser realizada utilizando o conhecimento de domínio, como encontrar registros com valores inválidos em algum atributo, granularidade incorreta ou exemplos errôneos. Pode-se também efetuar algumas limpezas independente de domínio como, tomar decisões estratégicas de tratamento de atributos incompletos, remoção de ruídos e tratamento de conjunto de exemplos não balanceados (BATISTA *et al.*, 2000).
- **Seleção e redução dos dados:** devido às restrições de espaço em memória ou tempo de processamento, o número de exemplos e de atributos disponíveis para análise pode inviabilizar a utilização de algoritmos de Extração de Padrões. Como solução a esse problema, é possível utilizar métodos de redução dos dados antes de iniciar a busca pelos padrões. A redução de dados pode ser feita de três formas (WEISS; INDURKHYA, 1998): (1) redução do número de exemplos; (2) redução do número de atributos; e (3) redução do número de valores de um atributo.

Em sistemas educacionais a dificuldade intrínseca ao pré-processamento é ampliada devido à natureza heterogênea dos dados educacionais considerando: (1) **diferentes fontes de dados:** dados podem ser provenientes de arquivos de log (documentos textuais), bastante utilizado em Mineração de Uso, banco de dados, entre outros; (2) **diferentes atores:** alunos, professores, tutores, projetistas; e (3) **diferentes níveis de granularidade:** nível de evento ou ação, nível de atividades, nível do aluno, nível da disciplina, nível do curso, entre outros. Dessa forma, pode-se considerar que o pré-processamento dos dados é uma atividade vital para o sucesso de aplicações de MDE.

3 – Extração de padrões

Delimitado o problema e com os dados devidamente representados, o processo de MD avança, então, para a fase de extração de padrões. As tarefas a serem realizadas nessa etapa são definidas de acordo com o objetivo final do processo de extração de conhecimento. Assim, pode-se resumir as principais atividades de extração de padrões em duas grandes tarefas: **preditivas** e **descritivas**.

Atividades **preditivas** são aquelas que produzem um modelo descrito pelo conjunto de dados para prever o valor de uma ou mais variáveis de interesse (rótulos). Atividades **descritivas**, por sua vez, produzem conhecimento novo baseado nos dados por meio de padrões que podem ser interpretados por humanos (KANTARDZIC, 2003).

Para as atividades **preditivas**, de maneira geral, são utilizados algoritmos de máquina supervisionados. Esses algoritmos, conhecidos como indutores, exigem um conjunto de exemplos de treinamento para os quais o atributo classe tenha valor conhecido (MITCHELL, 1997; MONARD; BARANAUSKAS, 2003). Essa classe de algoritmos possui duas subclasses principais: os algoritmos de **classificação** e os algoritmos de **regressão**. **Classificação** é referente ao processo em que o atributo classe tem valor categórico, enquanto **regressão** representa a predição de variáveis com valores contínuos. A aplicação de tarefas preditivas em MD tem como principal aplicação a categorização automática.

Já atividades **descritivas** utilizam predominantemente algoritmos não-supervisionados, os quais extraem padrões de dados não-rotulados. As principais tarefas desse ramo são a obtenção de regras de associação, o agrupamento e a sumarização de dados.

As **regras de associação** são relações lógicas inferidas entre dados correlacionados de uma ou mais bases de dados, analisados conjuntamente (AGRAWAL; SRIKANT, 1994). Em outras palavras, as regras de associação indicam relações entre dois conjuntos disjuntos de atributos L e R , representando tais relações na forma $L \rightarrow R$, na qual L é conhecido como antecedente e R como o conseqüente da regra.

O **agrupamento de dados**, também conhecido como *clustering*, visa agrupar dados de acordo com alguma medida de similaridade entre eles. Basicamente, esse processo tem por objetivo formar grupos nos dados de maneira que dados dentro de um mesmo grupo sejam os mais similares possível, enquanto objetos de grupos diferentes apresentem máxima dissimilaridade. Como resultado do processo, pode-se obter agrupamentos *flat*, nos quais os grupos são isolados, ou hierárquicos, nos quais os grupos apresentam alguma estrutura hierárquica.

Por fim, a **sumarização** é uma atividade que visa obter uma descrição compacta para um conjunto de dados.

4 – Pós-processamento e utilização do conhecimento

Uma vez obtidos os padrões dos dados, esses devem ser analisados e interpretados. Deve-se avaliar, junto a um especialista, aspectos do conhecimento extraído de acordo com o contexto do problema, como a representatividade do conhecimento obtido, a novidade contida nos resultados e a maneira que o conhecimento deverá ser utilizado.

Em relação à validade do conhecimento, deve-se verificar se os padrões obtidos condizem com a realidade e são aplicáveis ao que se propôs no objetivo inicial. Por exemplo, em aplicações que envolvam tarefas preditivas, pode-se avaliar os modelos com medidas relativas à precisão na predição de dados não vistos, como taxa de erro, precisão e cobertura. A avaliação de modelos descritivos é um pouco mais complexa, uma vez que o critério de avaliação varia de acordo com o objetivo do processo. Dessa forma, algumas vezes podem não existir medidas objetivas para avaliar o resultado do processo, ficando a avaliação a cargo de um especialista do domínio (avaliação subjetiva).

Outro aspecto importante refere-se à compreensibilidade do conhecimento. Os algoritmos de extração de padrões podem gerar uma quantidade de padrões muito grande, dificultando a compreensão por parte do usuário e fazendo necessária a aplicação de mecanismos que forneçam aos usuários apenas os padrões mais interessantes (SILBERSCHATZ; TUZHILIN, 1995).

Uma tarefa que tem um importante papel no auxílio ao especialista nessa etapa do processo, mas que também pode ser utilizada ao longo das demais etapas, é a **visualização** dos dados. Card, Mackinlay e Schneiderman (1999) definem essa tarefa como uma representação visual de dados interativa baseada no computador para ampliar a cognição. Diversos tipos de gráficos e diagramas podem ser utilizados ao longo de todo o processo pelo analista, facilitando a compreensão dos resultados das diversas etapas e auxiliando a tomada de decisão. A análise visual pode, por exemplo, indicar falhas no pré-processamento dos dados e a necessidade de refazê-lo.

Com o correto desenvolvimento das etapas anteriores, pode-se garantir que o conhecimento, ao final do processo, é válido e útil, podendo ser aplicado no apoio à tomada de decisão para os aspectos estabelecidos nos objetivos do processo (REZENDE *et al.*, 2003).

3.3.2 Métodos para mineração de dados educacionais

A Mineração de Dados Educacionais (MDE) é uma área de pesquisa emergente que desenvolve métodos e aplica técnicas da estatística, aprendizagem de máquina e mineração de dados com o objetivo de analisar os dados coletados durante todo o processo de ensino e aprendizagem (ROMERO; VENTURA, 2010; LIÑÁN; PÉREZ, 2015; DUTT; ISMAIL; HERAWAN, 2017).

Em MDE, busca-se converter os dados brutos provenientes de ambientes de apoio ao ensino e aprendizagem em informações úteis que podem gerar algum impacto no contexto educacional. Esse processo não difere de outras áreas de aplicação, por seguir os

mesmos passos do processo de MD (REZENDE *et al.*, 2003): identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento.

Considerando a etapa de extração de padrões, várias tarefas podem ser utilizadas no contexto educacional, como: predição (classificação e regressão), agrupamento, mineração de relações (regras de associação), detecção de *outlier*, mineração de textos, análise de redes sociais, destilação para facilitar a decisão humana e descoberta com modelos.

Alguns desses métodos são tradicionais da área de MD e outros são abordagens que envolvem combinações de métodos de MD, estatística, visualização e modelos educacionais. Cada um desses métodos será sintetizado a seguir.

Predição

Conforme descrito durante na etapa de Extração de Padrões (Seção 3.3.1), para atividades de predição são utilizados principalmente algoritmos supervisionados, ou seja, algoritmos que exigem um conjunto de dados para treinamento.

A predição busca produzir um modelo que possibilite realizar inferências de uma única característica (variável de interesse ou rótulo) a partir de um conjunto de dados (variáveis preditoras) (KANTARDZIC, 2003). Desse modo, um modelo preditivo pode ser entendido como uma função $f(x, \alpha) \approx k$, na qual x é um conjunto de variáveis preditoras, α são parâmetros desconhecidos e k é uma variável predita. Assim, deseja-se estimar o valor de k a partir da descoberta de α utilizando x .

Em Mineração de Dados Educacionais, a predição pode auxiliar na detecção de fatores comportamentais do aluno a partir de dados gerados em interações em fóruns de discussão, avaliações práticas, entre outros. Romero *et al.* (2011) descrevem quatro tipos de problemas principais que utilizam algoritmos de classificação e regressão para realizar predições a partir de dados provenientes de ambientes educacionais.

- **Sucesso acadêmico:** realiza a predição a respeito do desempenho de alunos ao longo do curso (graduação ou pós-graduação, por exemplo) ou dos possíveis desistentes (DEKKER; PECHENIZKIY; VLEESHOUWERS, 2009; GRÜNEWALD *et al.*, 2013; NGHE; JANECEK; HADDAWY, 2007; SUPERBY; VANDAMME; MESKENS, 2006).
- **Resultados específicos de cursos:** realizar predições para antecipar resultados de alunos em um curso específico. Com isso, busca-se predizer se o aluno será aprovado ou reprovado no curso, o índice de desistência, ou a pontuação que será adquirida ao

final do curso entre outros (BRESFELEAN *et al.*, 2008; HÄMÄLÄINEN *et al.*, 2004; KOTSIANTIS; PIERRAKEAS; PINTELAS, 2003; MINAEI-BIDGOLI *et al.*, 2003).

- **Êxito na próxima tarefa:** realizar predições para verificar se o aluno terá êxito na próxima tarefa. Com isso, é possível selecionar a próxima questão de acordo com o nível de conhecimento atual do aluno (DESMARAIS; PU, 2005; JONSSON *et al.*, 2005).
- **Habilidades, hábitos e motivação:** realizar predições do nível de motivação ou compromisso, experiência na utilização do ambiente educacional ou estratégia de intervenção por meio de recomendação (LEE, 2001; COCEA; WEIBELZAHN, 2006; HURLEY; WEIBELZAHN, 2007).

É importante destacar que no contexto de MDE são utilizados basicamente dois tipos de técnicas de predição: a classificação e a regressão. Segundo (ROMERO *et al.*, 2011; SHAHIRI; HUSAIN; RASHID, 2015), na classificação, os algoritmos mais utilizados são *Decision tree* e *Support Vector Machine*. Em relação à regressão, os algoritmos mais populares são *Neural Networks*, *Linear Regression* e *Support Vector Machine* para regressão.

Agrupamento de dados

O agrupamento de dados, ou *clustering*, é uma atividade descritiva que utiliza principalmente algoritmos não supervisionados, conforme definido na etapa de Extração de Padrões (Seção 3.3.1).

Um grupo (*cluster*) é um conjunto de entidades semelhantes sendo que entidades pertencentes a diferentes grupos não são semelhantes. Um grupo pode ser definido também como uma aglomeração de pontos no espaço tal que a distância entre quaisquer dois pontos no grupo é menos do que a distância entre qualquer ponto no grupo e qualquer ponto fora deste (EVERITT, 1974). Considerando o agrupamento de dados, seu principal objetivo é dividir um conjunto de dados em grupos que foram definidos de acordo com a sua similaridade.

No contexto educacional, Romero *et al.* (2011) definiram os problemas de agrupamento em três principais grupos a partir de uma revisão da literatura:

- **Análise de agrupamento de conteúdo educacional:** pesquisas que propuseram utilizar técnicas que agrupam materiais educacionais a partir de ferramentas baseadas em ontologias e web semântica (TANE; SCHMITZ; STUMME, 2004). Outro assunto investigado refere-se à criação de ferramentas de busca inteligente, que agrupam material com base na sua semelhança semântica (DRIGAS; VRETTAROS, 2004).

- **Agrupamento de alunos por comportamento:** pesquisas que propõem ferramentas Web para agrupar alunos com base diferentes critérios. No trabalho de Christodouloupoulos e Papanikolaou (2007), foi desenvolvida uma ferramenta para que professores possam selecionar o método de agrupamento de forma manual ou automática e, em seguida, definir o número de grupos a serem criados. Os modelos *fuzzy* são considerados nas pesquisas de Tian *et al.* (2008), a qual apresentou um sistema de educacional baseado em um conjunto *fuzzy* como algoritmo para agrupar os alunos em classes homogêneas, conforme a personalidade e as estratégias de aprendizagem, de modo a proporcionar aprendizagem personalizada.
- **Análise de agrupamento como uma ferramenta para melhorar ambientes educacionais:** diversos pesquisadores sugerem que a interação do grupo é um elemento fundamental no processo de ensino e aprendizagem (MAULL; SALDIVAR; SUMNER, 2010; FU; OFOGHLU, 2008). No entanto, a formação de grupos com alto nível de interação vem sendo difícil em ambientes virtuais por conta das limitações de comunicação, das diferenças de aprendizagem e do contexto social. Nesse contexto, alguns autores propõem técnicas utilizando agrupamento para promover melhorias nos ambientes educacionais.

Mineração de relações

Em Mineração de Relações o objetivo é descobrir relações entre variáveis em um conjunto de dados e codificá-las em regras para que possam ser utilizadas posteriormente. Em MD essas relações lógicas inferidas entre dados correlacionados são chamadas de regras de associação.

Existem diferentes tipos de mineração de relações como (BIENKOWSKI; FENG; MEANS, 2012):

- **Regras de associação:** qualquer relacionamento entre variáveis.

A mineração de regras de associação é uma das mais importantes técnicas de MD. Seu objetivo é derivar regras de conhecimento, referindo-se a relacionamentos entre variáveis em um conjunto de dados (AGRAWAL; IMIELIŃSKI; SWAMI, 1993).

A regra de associação pode ser formalizada como: seja $I = \{I_1, I_2, \dots, I_m\}$ um conjunto de atributos com m itens e seja $T = \{t_1, t_2, \dots, t_n\}$ uma base de dados com n transações. A regra tem a seguinte forma: **se X então Y** ($X \rightarrow Y$) é válido em um conjunto de transações T onde $X \cap Y = \emptyset$ e $X \subset I$ e $Y \subset I$.

Um exemplo de regra de associação em MDE seria associar a interação do aluno em fóruns de discussão com o seu desempenho.

Existem diversas medidas que podem ser utilizadas para guiar a extração de regras de associação. As principais e mais básicas são o suporte e a confiança. O suporte pode ser descrito como a probabilidade de que uma transação qualquer satisfaça tanto X quanto Y , ao passo que a confiança é a probabilidade de que uma transação satisfaça Y , dado que ela satisfaz X .

Desse modo, dado um conjunto de transações, o problema de mineração de regras de associação pode ser visto como o processo de encontrar as regras significativas, ou seja, aquelas que apresentem o suporte e a confiança iguais ou maiores do que os valores mínimos determinados pelo usuário.

- **Mineração de padrões sequenciais:** identificar associações temporais entre variáveis.

Trata-se de uma forma mais restritiva de mineração de regras de associação em que a ordem dos itens acessados é considerada. A mineração de padrões sequenciais tenta descobrir padrões a partir da presença de um conjunto de itens ordenados dentro de um período (PARTHASARATHY *et al.*, 1999).

Esse método pode ser aplicado em vários domínios tanto no meio acadêmico como na indústria. Um exemplo está na área da saúde, a qual pode utilizar dados temporais para prever o aparecimento da doença a partir de uma sequência de sintomas. Na indústria financeira, pode ser utilizada para prever o risco de investimento com base numa sequência de eventos do mercado financeiro (PARTHASARATHY *et al.*, 1999).

No contexto de MDE, a mineração de sequências pode ser utilizada para inúmeras finalidades. Pahl e Donnellan (2002) realizaram uma pesquisa que busca descobrir e comparar padrões de comportamento esperados com o objetivo de descrever um caminho de aprendizagem ideal a partir de dados temporais. Já na pesquisa de Ha, Bae e Park (2000), os autores investigaram qual é a melhor forma de organizar o ambiente educacional oferecendo sugestões para os alunos que possuem características de navegação semelhantes.

Detecção de outlier

A detecção de *outlier*, também definida por diversos autores como detecção de novidade, detecção de anomalias, detecção de ruído, detecção de desvio ou mineração de exceções, refere-se à identificação daquele elemento que parece desviar-se acentuadamente de outros membros da amostra em que ele ocorre (HODGE; AUSTIN, 2004). Em outras

palavras, a detecção de *outlier* é um tipo de análise de dados que visa determinar e informar pontos de dados significativamente diferentes dos demais dados.

No contexto educacional, a detecção de *outlier* pode ser utilizada, por exemplo, para auxiliar na detecção de alunos que não se adaptaram ao processo de aprendizagem (UENO, 2004); e para detectar regularidades e desvios nas ações entre aluno e professor (MUEHLENBROCK, 2005).

Mineração de textos

Mineração de textos (MT) refere-se ao processo de extração de informação útil (conhecimento) em documentos de textos não-estruturados (FELDMAN; SANGER, 2006). No contexto educacional, MT tem sido utilizada para analisar o conteúdo de fóruns de discussão, chats, páginas da Web, documentos, entre outros (ROMERO; VENTURA, 2013). Com exemplo de pesquisas relacionadas pode-se citar: agrupamento de documentos de acordo com seus temas e semelhanças e fornecendo resumos (HAMMOUDA; KAMEL, 2007); encontrar e organizar o material usando a informação semântica (TANE; SCHMITZ; STUMME, 2004); avaliar o progresso da conversa para observar se há contribuição para o tema (DRINGUS; ELLIS, 2005); detectar o foco da conversa e de discussões encadeadas e classificar temas (KIM *et al.*, 2006).

Análise de redes sociais

Análise de redes sociais (*Social Network Analysis* – SNA) é uma área de pesquisa que investiga um conjunto de técnicas utilizadas para estudar a troca de recursos entre os atores (indivíduos, grupos ou organizações). Um desses recursos é a informação, em que padrões regulares de troca de informações são as redes sociais (grafo), com atores (nós do grafo) e as relações de troca de informações (arestas do grafo que conectam os nós) (HAYTHORNTHTWAITE, 1996).

Em ambientes educacionais, a SNA pode ser usada para apoiar atividades coletivas que ocorrem em grupos de aprendizagem com base em um modelo relacional de mensagens trocadas entre os participantes de um fórum de discussão (REYES; TCHOUNIKINE, 2005).

Destilação para o facilitar decisões humanas

Destilação para o Facilitar Decisões Humanas é uma área cujo principal objetivo é apresentar dados complexos de forma inteligível utilizando visualização, sumarização e interfaces interativas, de modo a facilitar a decisão humana e, assim, oferecer apoio em

tomadas de decisão. Essa abordagem se sobrepõe à análise de dados visuais permitindo que os seres humanos identifiquem rapidamente as características dos dados. Uma forma de realizar pesquisas nessa área é por meio de processamento analítico dos dados educacionais utilizando *Business Intelligence*.

Em ambientes educacionais a ideia é obter estatísticas descritivas de dados educacionais para obter dados globais, resumos e relatórios sobre o comportamento do aluno. Além disso, também é possível promover a visualização dessas informações por meio de técnicas gráficas que ajudem a facilitar a compreensão e exploração desses dados (ROMERO; VENTURA, 2013).

O uso da destilação de dados também pode ser utilizado para categorizar as ações dos estudantes. Desse modo, é possível auxiliar o desenvolvimento de um modelo de predição mais robusto (BAKER; ISOTANI; CARVALHO, 2011).

Descoberta com modelos

Descoberta com modelos é uma técnica que envolve a utilização de um modelo validado de um fenômeno (como predição, agrupamento, ou ainda manualmente por meio da engenharia do conhecimento) como um componente de uma análise mais aprofundada (ROMERO; VENTURA, 2013; BIENKOWSKI; FENG; MEANS, 2012). Um exemplo está nas pesquisas de Jeong e Biswas (2008) que utilizam a abordagem para a construção de modelos para a análise de comportamento dos alunos em diferentes etapas do processo de aprendizagem.

3.3.3 Aplicações de mineração de dados educacionais

A MDE é uma área de atuação que busca responder questões cada vez mais complexas (ROMERO; VENTURA, 2007; ROMERO; VENTURA, 2013). Analisando problemas comuns encontrados na literatura, Romero e Ventura (2010) agruparam esses problemas em formas de aplicação, sendo que cada modelo busca responder perguntas específicas, conforme descrito no Quadro 3.

Nesta seção, são apresentadas algumas das áreas de aplicação de MDE encontradas na literatura (ROMERO; VENTURA, 2010; ROMERO; VENTURA, 2013): modelo de conhecimento, comportamento, experiência, perfil e domínio do usuário.

Modelagem do conhecimento do usuário

De modo geral, os modelos do usuário são utilizados para personalizar e adaptar o comportamento de sistemas computacionais às necessidades específicas dos usuários. Desse

Quadro 3 – Mineração de dados educacionais: áreas de aplicação.

Aplicações	Questões
Conhecimento do usuário	Qual conteúdo o aluno conhece?
Comportamento do usuário	O que os padrões de comportamento dos alunos significam para a sua aprendizagem?
Experiência do usuário	Os alunos estão satisfeitos com a sua experiência?
Perfis do usuário	Em quais grupos o aluno agrupado?
Domínio usuário	Qual é o nível correto em que se deve dividir tópicos em módulos e como esses módulos devem ser dispostos?

Fonte: Adaptada de Romero e Ventura (2010).

modo, a partir das informações contidas no modelo do usuário, é possível que os sistemas façam intervenções específicas com o objetivo de **dizer a coisa certa no momento certo** (FISCHER, 2001).

Inferir o que o usuário sabe, ou seja, modelar o conhecimento do usuário, requer a análise de todos os dados que representam as interações entre os alunos e o ambiente educacional. Tais dados representam informações armazenadas sobre as características do aluno (conhecimento, motivação, atitudes, personalidade) e questões sociais (BIENKOWSKI; FENG; MEANS, 2012; COSTA *et al.*, 2013). Vale ressaltar que as inferências podem ser feitas por um modelo computacional preditivo ou por um professor observando os dados do aluno em um painel.

A modelagem do conhecimento do usuário tem sido amplamente utilizada na construção de sistemas hipermídia adaptativos, sistemas de recomendação, sistemas tutores inteligentes, sistemas especialistas, entre outros.

Baker *et al.* (2008) propuseram um método para rastrear conhecimento utilizando uma abordagem de aprendizagem de máquina para estimar a probabilidade de um aluno acertar sem saber ou errar sabendo.

Bienkowski, Feng e Means (2012) salientam que diversas empresas utilizam a modelagem de conhecimento do usuário em *software* comercial. Uma vertente de modelagem está em determinar se o aluno está habilitado ou não a mudar para o novo conteúdo de aprendizagem. Existem ainda empresas que verificam quando os alunos estão realizando atividades com resultados acima da média. Nesse caso, eles recebem sugestões para passar para o próximo conteúdo; caso contrário, eles são encorajados a consolidar as competências atuais e trabalhar em pré-requisitos.

Como um exemplo do uso de modelagem de conhecimento do usuário, um *software* educacional pode coletar dados como quantos minutos são gastos em uma unidade, quantas sugestões foram utilizadas, e erros comuns. Os dados para um aluno podem, então, ser comparados com um modelo construído a partir de um grande número de estudantes.

Modelando o comportamento do usuário

Modelar o comportamento do usuário pode ser feito a partir do mesmo tipos de dados utilizados para prever seu conhecimento (Seção 3.3.3). Adicionalmente, pode-se utilizar medidas como: o tempo que o aluno gastou *online*, seu comparecimento e seus atrasos. Eventualmente, pode-se considerar também o nível de conhecimento do aluno a partir de resultados de avaliações.

Baker *et al.* (2008) utilizam MDE para detectar comportamentos inadequados como, por exemplo, verificar se o aluno está “trapaceando” (do inglês *Gaming the System*) ao interagir com o *software* educacional.

No trabalho de Macfadyen e Dawson (2010), os autores utilizaram um modelo de regressão que incorpora variáveis que descrevem fatores comportamentais do aluno, como, por exemplo, o número total de mensagens publicadas em fórum de discussão, o número total de mensagens enviadas via correio eletrônico e o número total de avaliações concluídas para realizar previsões a respeito da nota.

Vale destacar que a modelagem de comportamento do usuário também é utilizada em empresas de *software*. Um exemplo está na empresa Zynga² (jogos sociais), que tenta prever o que os usuários querem e o que vão fazer em seguida, para descobrir como fazer jogos mais divertidos e conseqüentemente obter usuários mais motivados e comprometidos (BIENKOWSKI; FENG; MEANS, 2012).

Modelando a experiência do usuário

No modelo de experiência do usuário busca-se verificar a satisfação do usuário por meio de dados coletados. Esse modelo tem sido amplamente utilizado por empresas de comércio eletrônico e, com menor ênfase, em ambientes de apoio ao ensino e aprendizagem (BIENKOWSKI; FENG; MEANS, 2012).

De modo geral, as empresas utilizam avaliações em diferentes etapas do fluxo de contratação para medir o grau de satisfação do usuário e a taxa de retenção. A partir dos dados coletados também é possível obter as informações necessárias com o objetivo de redesenhar a experiência do usuário de forma a aumentar a taxa de retenção.

No contexto educacional, os dados são coletados por meio de questionários de acompanhamento, e também por meio de suas escolhas, comportamentos e desempenho. Com essas informações torna-se possível desenvolver o modelo de experiência do aluno

² Zynga: <http://zynga.com>

para melhorar a aprendizagem e reduzir o abandono, principalmente em cursos totalmente à distância.

Nas pesquisas de Dawson, Heathcote e Poole (2010), os autores analisam os dados de diversos sistemas de informação de uma instituição de ensino (com todo o ciclo de vida do aluno) para combinar esses dados e, assim, obter maior detalhamento acerca dos alunos. A partir dos dados coletados foi desenvolvido um modelo de experiência detalhado do usuário, a fim de obter uma melhor compreensão dos alunos e, conseqüentemente, melhorar a experiência de aprendizagem na instituição de ensino.

Perfis do usuário

Um perfil do usuário é uma coleção de dados pessoais que descrevem suas características essenciais. A construção dos perfis de alunos ou grupos pode ser feita utilizando mineração de dados e algoritmos de aprendizagem de máquina, com o objetivo de diferenciar alunos ou grupos a partir de suas preferências, interesses, objetivos de aprendizagem, dados pessoais, dados demográficos, aspectos de aprendizagem, entre outras características.

Considerando a pesquisa de Kardan e Conati (2011), os autores propuseram um *framework* de modelagem de usuário que utiliza registros de interações em ambientes educacionais para identificar os diferentes tipos de alunos. Essa informação é utilizada para classificar novos alunos para, posteriormente, oferecer apoio às interações quando são detectados comportamentos prejudiciais à aprendizagem. Classificação também pode ser utilizada para a criação de grupos de estudos de alunos ou outras atividades colaborativas.

Empresas de jogos frequentemente identificam o perfil dos usuários para aumentar sua motivação e envolvimento durante o jogo. Os jogadores são classificados segundo suas características como: jogadores casuais, jogadores de final de semana, jogadores sociais, gastadores, entre outros. Essa classificação, de modo geral, é importante para proporcionar uma recomendações mais assertivas principalmente em grupos com hábitos de compra (BIENKOWSKI; FENG; MEANS, 2012).

Modelagem de domínio

O modelo de domínio é frequentemente utilizado para representar os conceitos chaves que representam um determinado assunto ou área, ou seja, um domínio. No modelo também são identificados os relacionamentos entre os conceitos chaves ou unidades de estudos (BRUSILOVSKY; MILLÁN, 2007).

Pesquisas de modelagem de domínio em MDE investigam como a aprendizagem pode ser afetada a partir de alterações na forma como os assuntos (tópicos) são divididos

ou sequenciados. Além disso, é importante considerar o nível de granularidade em que os assuntos são apresentados, podendo ser mais específicos ou mais gerais.

Em ambientes educacionais o modelo de domínio tem demonstrado sua efetividade por proporcionar melhorias de aprendizagem ao apresentar sequências de conceitos relacionados em ordens pré-definidas a partir dos dados coletados no ambiente. Um exemplo está nas pesquisas de Martin *et al.* (2011), em que os autores descrevem um estudo que consiste no uso da curva de aprendizagem (medidas de desempenho relacionadas às oportunidades de aprendizagem) em três contextos distintos. Desse modo, foi possível demonstrar como as curvas de aprendizagem podem ser utilizadas para conduzir mudanças em ambientes educacionais personalizados (sistemas tutores inteligentes).

Considerando a modelagem de domínio, Martin *et al.* (2011) demonstram que ao traçar curvas de aprendizagem para vários subconjuntos do modelo de domínio é possível comparar as curvas resultantes para tentar determinar quais os componentes do conhecimento são de fato efetivos. Adicionalmente, foi feita uma análise da grande quantidade de dados dos alunos disponível para efetuar ajustes do modelo de domínio do sistema.

3.4 Considerações finais

Neste capítulo foram apresentadas as linhas de pesquisa *Learning Analytics* (LA) e *Academic Analytics* (AA) que investigam a descoberta de conhecimento a partir da análise de dados de sistemas educacionais. Em geral, LA explora formas de melhorar o sucesso do aluno. Seus dados são analisados em nível micro (individual do usuário) e meso (institucional). Por outro lado, pesquisas em AA buscam propor melhorias nas práticas de gestão institucional. A análise de seus dados pode chegar ao nível macro (interinstitucional).

Para potencializar a utilização de LA e AA, uma síntese de Indicadores de Desempenho Educacional foi apresentada. A definição de indicadores de desempenho visa a definição metas e objetivos institucionais que devem ser monitoradas ao longo do processo educacional.

Neste capítulo também foi definida a área de Mineração de Dados Educacionais (MDE), que se refere a um importante linha de pesquisa que está na intersecção de LA e AA. Modelos do comportamento, conhecimento, experiência, domínio e perfil do usuário foram traçados para melhorar a compreensão sobre as principais tarefas que podem ser realizadas em contextos distintos. Considerando que as linhas de pesquisa citadas, cada

uma com as suas peculiaridades, atuam na descoberta de conhecimento a partir de dados educacionais, neste trabalho, essas três áreas foram denominada Análise Educacional.

Sob essa perspectiva, para a proposição do EDRM, Análise Educacional foi amplamente investigadas, principalmente na etapa de caracterização dos dados que definem os atributos relevantes à definição do modelo. Nessa etapa foram realizadas análises a partir de pesquisas em sistemas educacionais, Análise Educacional e apoio à decisão orientado a dados (*Data Warehouse*, *Business Intelligence* e modelagem dimensional) para a definição de atributos, medidas e métricas necessários à definição das tabelas fato e dimensões do modelo dimensional proposto neste trabalho.

No próximo capítulo é apresentada uma síntese das técnicas utilizadas para a definição do EDRM. Essas técnicas são importantes para a estruturação e armazenamento de dados analíticos considerando diferentes perspectivas, *stakeholders* e níveis de granularidade para fins de apoio à decisão orientada a dados.

PROCESSO DECISÓRIO ORIENTADO A DADOS

4.1 Considerações iniciais

Nos últimos anos técnicas e tecnologias para apoiar o processo de tomada de decisão orientada a dados vêm sendo amplamente exploradas no contexto educacional (LANG; PIRANI, 2016; DEDE; HO; MITROS, 2016). Segundo pesquisas realizadas para compor o relatório de mercado de aplicações empresariais 2015, do Centro de Análise e Pesquisa da EDUCAUSE (ECAR)(LANG; PIRANI, 2016), há uma grande tendência na utilização de *Business Intelligence* (BI) e *Data Warehouse* (DW) por instituições de ensino superior com o objetivo de tornar os dados educacionais acessíveis para diversos propósitos. Em 2015, 9 em cada 10 instituições (88%) utilizaram relatórios de BI, comparados a 8 em 10 (80%) em 2013. Da mesma forma, 8 em 10 instituições (79%) forneceram um sistema de DW em 2015, em comparação com 7 em cada 10 (71%) em 2013. Isso evidencia que, embora criar e gerenciar sistemas de apoio à decisão não seja uma atividade trivial, seus benefícios potenciais são de extrema relevância para a gestão do processo educacional.

Conforme observado por Zorrilla *et al.* (2005), embora a maioria dos sistemas educacionais sejam utilizados para publicar ou acessar materiais do curso, eles não fornecem aos professores ferramentas necessárias para que possam rastrear e avaliar todas as ações realizadas por seus alunos e, assim, promover melhorias na eficácia do curso e da aprendizagem.

Sob essa perspectiva, embora existam grandes volumes de dados disponibilizados por sistemas educacionais que poderiam ser utilizados para melhorias dos processos

institucionais, esses dados são complexos e heterogêneos, demandando grande esforço das instituições no projeto e desenvolvimento de uma estrutura de dados analítica para apoio à decisão. Nesse sentido, a definição de um modelo de referência de dados educacionais, torna-se relevante para iniciar um processo decisório estratégico orientada a dados.

O modelo de referência proposto (EDRM), foi desenvolvido segundo um modelo dimensional de dados utilizando a estrutura de um DW. Assim, o EDRM é capaz de fornecer subsídios à utilização de dados provenientes de diversas fontes, com o objetivo de abranger análises segundo as principais perspectivas educacionais utilizadas por instituições de ensino. Além disso, o modelo possibilita análises considerando diferentes *stakeholders* e níveis de granularidade.

Neste capítulo são apresentadas as principais técnica investigadas para a definição de modelos de dados analíticos para a apoio à decisão. Os tópicos aqui abordados são importantes para a definição do EDRM como uma estrutura analítica de dados integrados e consistentes.

Na Seção 4.2 são apresentadas as principais diferenças entre ambientes operacionais e ambientes informacionais. Na Seção 4.3 é apresentado o conceito de *Data Warehouse*, considerando suas características e arquitetura. Além disso, é definida a modelagem dimensional, seus elementos, hierarquias e granularidades. Na Seção 4.4 é apresentado o processamento analítico OLAP (*On-line Analytical Processing*). Esse processamento representa uma das principais formas de realizar análises e consulta de dados em um *Data Warehouse*. Na Seção 4.5, é discutido o uso de vários elementos e técnicas descritos anteriormente em uma arquitetura de *Business Intelligence*. Por fim, na Seção 4.6, são descritas algumas pesquisas que abordam a utilização de modelos dimensionais e *Data Warehouse* utilizados no contexto educacional.

4.2 Ambientes operacionais e informacionais

Sistemas educacionais, de modo geral, são sistemas que processam transações (OLTP – *On-line Transaction Processing*) armazenando seus dados em banco de dados relacional, ou seja, entidades, relacionamentos e restrições. Esse tipo de armazenamento pode ser considerado como uma **representação em nível operacional** (BARBIERI, 2011).

Outra forma de representar os dados é por meio de um processamento analítico (OLAP – *On-line Analytical Processing*), utilizado para representar **níveis informacionais ou estratégicos**. A representação nesse nível permite estruturar os dados para que possam

ser realizadas análises da informação sob inúmeras perspectivas ou dimensões, ou seja, uma estrutura de dados dimensional.

Desse modo, ambientes de dados de **suporte à decisão ou informacionais**, são essencialmente diferentes de ambientes **operacionais ou transacionais**. Segundo Inmon (2005), essa diferenciação é caracterizada por razões como:

- Os dados que atendem necessidades operacionais são dados fisicamente diferentes daqueles que atendem a necessidades informacionais ou analíticas.
- A tecnologia de suporte para processamento operacional é fundamentalmente diferente da tecnologia usada para apoiar necessidades informacionais ou analíticas.
- A comunidade de usuários de dados operacionais é diferente daquela servida por dados informativos ou analíticos.
- As características de processamento para o ambiente operacional e informacional são fundamentalmente diferentes.

Por essas razões, a melhor forma de construir sistemas é separando o processamento operacional do informacional ou analítico. No Quadro 4, são destacadas as principais diferenças desses ambientes.

A definição de um ambiente *Data Warehousing* tem início na transformação de dados operacionais em dados informacionais. Desse modo, é possível separar o processo transacional do analítico com o objetivo de possibilitar análises utilizando diversas fontes de dado simultaneamente. Isso significa que ao definir estruturas de processamento distintas, o modelo analítico torna-se flexível para sofrer alterações sem influenciar as estruturas de dados originais dos sistemas educacionais. Na próxima seção são apresentados os conceitos de *Data Warehouse*, suas características, arquitetura, formas de modelagem e elementos.

4.3 Data warehouse

Ambientes de *Data Warehousing* têm por objetivo consolidar dados de interesse, a partir de sistemas de informação heterogêneos, em uma única base de dados denominada de *Data Warehouse* (DW) ou armazém de dados. Esses ambientes são capazes de garantir a eficiência e a flexibilidade necessária à recuperação de informações importantes ao processo decisório.

Segundo Inmon, Strauss e Neushloss (2010), DW é uma base de processamento informacional que atua a partir de uma coleção de dados, com o objetivo de apoiar decisões

Quadro 4 – Diferença entre ambientes operacionais e informacionais.

Características	Ambiente Operacional	Ambiente Informacional
Tipos de operações mais frequentes	Atualização, inserção, consulta e remoção	Realiza consulta/leitura
Características da transação	Pequena e simples, acessam poucos registros por vez	Longas e complexas, acessam muitos registros por vez e realizam várias rotinas e junções
Tipos de usuários	Administradores e projetistas de sistema, usuários finais	Usuários definidos para apoio à decisão: administradores, gestores, entre outros
Volume de transação	Relativamente alto	Relativamente baixo
Volume de dados	<i>Megabytes a Gigabytes</i>	<i>Gigabytes a Terabytes</i>
Granularidade dos dados	detalhado	agregado / detalhado
Conteúdo	Valores correntes/atuais	Valores sumarizados, calculados, integrados de diversas fontes de dados
Organização dos dados	Por aplicação / sistema de informação	Por assunto
Natureza dos dados	Dinâmica	Estática até a atualização programada de tempos em tempos
Projeto de banco de dados	Normalizados para suporte às propriedades de atomicidade, consistência, isolamento e durabilidade (ACID)	Dimensional, para suprir necessidades de suporte à decisão
Utilização	Relacional, própria para computação transacional	Dimensional, própria para atividades analíticas
Tempo de resposta	Otimizado para faixas abaixo de 1 segundo	Análises complexas e tempos de resposta maiores

Fonte: Ponniah (2001), Inmon (2005).

estratégicas e gerenciais. Trata-se de um repositório de dados integrados, desenvolvido para apoiar os processos de tomada de decisão, garantir a obtenção eficiente de informações estratégicas, além de manter os dados gerenciais de alta qualidade.

Em geral, o acesso às informações integradas de diferentes fontes de dados é realizado em duas etapas (CIFERRI, 2002): (1) a informação de cada fonte de dados é extraída previamente, traduzida, filtrada e integrada a informações relevantes de outras fontes de dados e finalmente armazenada em um DW; e (2) as consultas, quando realizadas, são executadas diretamente no DW, sem a necessidade de acesso direto às fontes de dados de origem.

Dessa forma, as informações integradas tornam-se disponíveis para consultas ou análises por meio de sistemas de apoio à decisão. Um exemplo é construir uma aplicação de *Data Warehousing* para que *stakeholders* de uma instituição de ensino superior possam realizar análises a partir da extração de dados de um ou mais sistemas educacionais, como SIS e LMS. Os usuários desses sistemas poderiam realizar análises integradas de informações de gestão institucional e gestão do aluno. Esses dados poderiam, ainda, ser utilizados de

forma contínua por meio de análises a partir de sistemas de *Analytics*, mineração de dados e *softwares*.

Para facilitar o consumo dos dados, pode-se utilizar a ideia de *Data Marts* (DM). Segundo Kimball e Ross (2013), *Data Mart* é subconjunto de dados de um DW referente a um assunto em especial ou diferentes níveis de sumarização. Um DM é um DW orientado por assunto ou área organizacional, ou seja, é uma versão reduzida de um DW concentrados por departamentos específicos.

Em suma, um DW pode ser considerado o núcleo do ambiente de *Data Warehousing*, criado a partir da necessidade de separação entre os ambientes operacional e informacional. Cada um desses ambientes possui características próprias de dados, descritas na seção a seguir.

4.3.1 Características dos dados

Os dados presentes em um DW são caracterizados por serem orientados por assunto, integrados, históricos (variável no tempo) e não voláteis (INMON, 2005; INMON; STRAUSS; NEUSHLOSS, 2010). Essas características são importantes como forma de apoio ao processo gerencial e tomada de decisão.

Armazenar dados **orientados por assunto** significa armazená-los de acordo com os temas de negócio de maior interesse tais como usuário e curso. Essa abordagem é baseada em uma estrutura de dados desnormalizados, que possui somente dados relevantes à tomada de decisão.

Para realizar a tomada de decisão, é necessário reunir todos os dados relevantes de diversas fontes ou sistemas de informação. A **integração dos dados** presentes no DW é a base para uma análise organizacional precisa e consistente. No processo de extração, somente um subconjunto de dados transacionais, necessários à análise estratégica, são transformados e armazenados no DW.

Um DW, devido à sua própria natureza, deve conter dados **históricos**, ou seja, dados relevantes a algum período de tempo, em contraste com ambientes transacionais cujos dados são relativos aos momentos acessados. Cada estrutura de dados no DW deve conter um elemento de tempo dependendo de sua granularidade (dia, mês, ano, entre outros). A natureza **variável no tempo** dos dados em um DW permite: (1) análise do passado; (2) relacionar informações com o presente; e (3) realizar previsões para o futuro. Em geral, dados históricos possuem volumes de dados muito superiores aos ambientes transacionais, representando uma complexidade adicional à administração dos dados.

A característica de **não volatilidade** refere-se ao fato de que o conteúdo do DW permanecerá estável ao longo do tempo. Isso significa que apenas dois tipos de transações são efetuadas: (1) transações de manutenção: para realizar a carga dos dados de forma consistente; e (2) transações de leitura: para a realização de consultas de apoio à decisão. Para a tomada de decisão informada por dados de sistemas educacionais, essas características são essenciais devido à diversidade de fontes de dados distintas, que podem ser analisadas de forma correlacionada, bem como a natureza heterogênea desses dados. Adicionalmente, os dados históricos de um DW permitem que aplicações de *Learning Analytics*, *Academic Analytics* e Mineração de Dados Educacionais possam ser realizadas de forma efetiva.

Por fim, em um DW tradicional a manutenção dos dados, ou seja, a carga de dados, é a única operação de escrita realizada. Em geral, a carga ocorre em períodos noturnos, para que não haja grandes processamentos de leitura de dados simultaneamente. No entanto, é importante destacar que é possível gerar análises de DW considerando dados extraídos de forma *online*, ou seja, carga de dados em tempo real conforme pesquisa de Kakish e Kraft (2012).

4.3.2 Arquitetura de data warehouse

A composição de um ambiente de *Data Warehousing* é realizada a partir de quatro componentes: fontes de dados transacionais, Extração, Transformação e Carga, área de apresentação dos dados e área de acesso aos dados, conforme ilustrado na Figura 6.

A primeira etapa tem por objetivo combinar diferentes fontes de dados em um único repositório. Por exemplo, em aplicações educacionais pode-se combinar dados de SIS e LMS. Também podem ser utilizados dados complementares e recursos externos, como dados públicos, disponíveis no Portal Brasileiro de Dados Abertos¹, entre outros. Em geral, essas fontes de dados são transacionais ou operacionais e estão em uma estrutura de dados relacional.

A segunda etapa refere-se ao processo de Extração Transformação e Carga (*Extract Transform Load* – ETL). O processo de ETL inicia-se na compreensão dos dados do sistema de origem para a realização da **extração dos dados**. O processo **transformação** é semelhante à etapa de pré-processamento de Mineração de Dados, descrita na Seção 3.3.1. Por fim, é necessário organizar esses dados e realizar a **carga** em um repositório central.

A terceira etapa considera que os dados devem ser organizados em uma estrutura própria para responder consultas com base em processos de negócio. A principal estrutura

¹ Website do Portal Brasileiro de Dados Abertos: <http://dados.gov.br/>

Figura 6 – Elementos básicos de um *Data Warehouse*.

Fonte: Adaptada de Kimball e Ross (2013).

de dados para auxiliar na tomada de decisão é o modelo dimensional, que possui dois principais formatos: *Star Schema* e *Snowflake Schema*. Esses modelos são detalhados a seguir.

A quarta etapa refere-se a área de acesso aos dados. Nessa etapa, ferramentas de *Business Intelligence* podem ser utilizadas para fazer análises e gerar relatórios, bem como modelos de Mineração de Dados de dados podem ser utilizados.

4.3.3 Modelagem dimensional

Em um ambiente de *Data Warehousing*, as análises efetuadas para apoiar o processo de tomada de decisão, em geral, representam requisições dimensionais aos dados de um DW, cujo objetivo é possibilitar a visualização desses dados considerando diferentes perspectivas.

De acordo com Kimball e Ross (2013), modelagem dimensional corresponde a uma técnica de projeto de banco de dados, desenvolvida para dar suporte a consultas de usuários finais em um DW. Os autores afirmam que a modelagem dimensional é amplamente aceita como a principal técnica para a apresentação de dados analíticos, pois trata de dois importantes aspectos: fornecer dados que sejam compreensíveis aos usuários finais e proporcionar desempenho em consultas.

Adicionalmente, a modelagem dimensional garante que os usuários possam entender facilmente os dados, além de permitir que o *software* navegue e forneça resultados de forma rápida e eficiente. Um dos motivos que torna o modelo dimensional eficiente é a redundância dos dados inerentes a um modelo desnormalizado.

Embora modelos dimensionais sejam muitas vezes instanciados em sistemas de gerenciamento de banco de dados relacional, eles são bastante diferentes dos modelos de terceira forma normal (3NF) que buscam remover redundâncias de dados. Ambos, 3NF e modelos dimensionais, podem ser representados em diagrama entidade e relacionamento porque consistem em tabelas relacionais. Desse modo, a principal diferença entre 3NF e modelos dimensionais é o grau de normalização.

Dado a simplicidade da estrutura do modelo dimensional, há casos em que sua representação se assemelha diretamente o negócio modelado. Isso acontece devido ao número reduzido de tabelas, quando comparado ao modelo transacional, e a utilização de descritores com características do negócio. Essa simplicidade do modelo dimensional também gera benefícios em desempenho. Otimizadores de banco de dados processam *schemas* simples (com menos junções) de forma mais eficiente (KIMBALL; ROSS, 2013).

Por fim, modelos dimensionais são facilmente adaptáveis a mudanças, principalmente mudanças do negócio ou comportamento do usuário. Isso significa que quando os dados já estão estruturados, o modelo dimensional não precisa ser ajustado caso o usuário necessite de novas maneiras para analisar o seu negócio e gerar conhecimento.

No contexto educacional, a modelagem dimensional fornece diversas vantagens quando deseja-se analisar perspectivas educacionais totalmente distintas, porém correlacionadas como a performance do aluno (notas) associadas à quantidade de interações por meio de ferramentas de comunicação síncronas e assíncronas. Essa vantagens são detalhadas no Capítulo 6.

Elementos de um Modelo Dimensional

Um modelo dimensional possui um conjunto de medidas numéricas, também chamadas de **fatos**, que são os objetivos das análises, e um conjunto de **dimensões**, que determinam o contexto das medidas. As medidas são valores numéricos que podem ser definidos em função das suas dimensões. Um exemplo são as notas obtidas por alunos em uma disciplina, a qual a nota é a medida e disciplina e aluno são dimensões.

As medidas numéricas relevantes ao negócio podem ser: **aditivas**, **semi-aditivas** ou **não aditivas**. Uma medida é **aditiva** quando é possível ser somada por meio de todas as suas dimensões. Um exemplo é a medida numérica “quantidade de acessos ao LMS”, ou

seja, pode-se somar todos os acessos ao LMS para gerar análises por aluno, curso, entre outros. Nesse caso, acesso é a medida aditiva e aluno e curso são as dimensões.

Por outro lado, a medida **semi-aditiva** pode ser somada por meio de algumas de suas dimensões. Em dimensões em que a medida não pode ser somada, significa que essa soma não possui significado algum.

As medidas **não aditivas** simplesmente não podem ser somadas, ou seja, aquelas que não podem ser sumarizadas associadas às dimensões. Percentuais são exemplos de valores armazenados nas medidas que não permitem sumarização. Por exemplo, não faz sentido somar o percentual de ausências de um aluno.

Com relação às dimensões, cada uma delas são definidas para determinar o contexto de um assunto referente a um fato de negócio. Possuem uma ou mais hierarquias naturais, além de atributos descritivos que representam características de um fato.

Existem duas principais abordagens utilizadas para realizar a representação do modelo dimensional, como (KIMBALL; ROSS, 2013):

- **Star Schema**: nesse modelo as tabelas de dimensões são diretamente relacionadas a uma tabela central (fato). É um modelo simples e eficiente. Não há preocupação com economia de espaço de armazenamento e nem preocupação com normalização. Respeita-se o preceito de informação rápida. Contudo, esse modelo apresenta maior redundância dos dados, consequentemente maior volume, devido à desnormalização.
- **Snowflake Schema**: nesse modelo as dimensões são normalizadas e as hierarquias são separadas, reduzindo assim redundâncias, o que resulta em agilidade na manutenção. Por outro lado, as consultas envolvem mais tabelas, podendo haver queda de desempenho.

A utilização do modelo *Star Schema* foi considerada para a definição do EDRM por sua simplicidade em termos de modelagem. A definição do modelo *Star Schema* permite maior facilidade de manutenção e expansão por possuir menos relacionamentos e maior desnormalização. Adicionalmente, esse modelo apresenta melhor desempenho nas consultas para a obtenção da informação de forma rápida e eficiente.

Hierarquias e Granularidades

Em um DW os dados são organizados segundo diferentes níveis de granularidade ou detalhamento. A granularidade refere-se ao nível de sumarização dos elementos e de

detalhes disponíveis, ou seja, quanto mais detalhado menor a granularidade. Sua definição é um dos principais fatores a ser considerado na modelagem de dados.

A vantagem de se ter um nível de granularidade baixo é o maior detalhamento nas análises. Por outro lado, tem-se como ponto negativo o maior volume de dados, o que pode reduzir a performance computacional. Quando se possui um nível de granularidade baixo, o tamanho do DW é muito grande, porém praticamente qualquer consulta pode ser respondida.

No caso da granularidade com nível maior, o volume de dados é menor e, consequentemente, a performance em consultas é superior. Em contrapartida, com menos detalhamento nos dados, pode não ser possível responder a todas as perguntas necessárias. Nesse sentido, a definição clara e coerente da granularidade é fundamental para o desenvolvimento de um modelo. Um exemplo é realizar consultas para a análise financeira de uma instituição de ensino cujo objetivo é consolidar os valores de matrículas realizadas por departamento ou por curso anualmente, semestralmente ou até mesmo diariamente. Esse tipo de consulta permite aos usuários realizar análises em nível agregado (departamento, por exemplo) para obter uma visão geral das questões financeiras associadas a matrículas realizadas/efetivadas, podendo percorrer hierarquias de agregação até o menor nível de granularidade (curso, por exemplo) para obter dados detalhados e específicos quando necessário.

Ressalta-se que a definição de níveis de granularidade detalhados (menor granularidade) ou agregados (maior granularidade) pode responder a análises de diversos tipos, seja planejada ou *ad hoc*.

Para o EDRM, foi considerado o menor nível de granularidade possível devido à necessidade de se considerar um projeto de DW de referência, capaz de prover estrutura para qualquer sistema educacional do tipo SIS e LMS. Contudo, em caso de muitos níveis, Inmon (2005) recomenda que dois (ou mais) níveis granulares de detalhamento, na maioria dos casos, são eficientes e atendem boa parte das consultas. Desse modo, cabe ao instanciador do EDRM definir a eficiência de seu modelo selecionando dois ou mais níveis de interesse.

4.4 Representação lógica do modelo dimensional

O principal propósito de um ambiente de *Data Warehousing* consiste na disponibilização de informações consistentes e integradas a usuários de sistemas de apoio à decisão. Dentre as principais ferramentas de análise e consulta de dados em um DW, pode-se citar ferramentas OLAP e ferramentas de Mineração de Dados.

Segundo Codd, Codd e Salley (1993), OLAP (*On-line Analytical Processing*) foi introduzido para definir a categoria de processamento analítico a partir de bancos de dados históricos voltados aos processos de gerência e tomada de decisão.

Ferramentas OLAP são caracterizadas por permitir análises sofisticadas utilizando visões dimensionais complexas por meio de diferentes operações analíticas. Essas operações definem formas de navegar em uma estrutura de dados logicamente definida no formato de um cubo, ou seja, uma representação para apresentar visões dos dados no espaço de múltiplas dimensões ou multidimensional.

Ao realizar análises de um DW a partir de um cubo OLAP é possível navegar por dados por meio de operações de *slice* (fatias) e *dice* (cubos), percorrer hierarquias (níveis de granularidade) por meio de operações de *drill-down*, *roll-up*, entre outras. Essas operações permitem que sejam realizadas agregações, sumarizações e consolidações, além de possibilitar a visualização dos dados sob diferentes perspectivas. Adicionalmente, existem várias abordagens para os processos OLAP, como (MOSLEY *et al.*, 2010):

- **MOLAP (*Multidimensional On-line Analytical Processing*)**: ou OLAP multidimensional é implementado por um banco de dados multidimensional especializado. Essas estruturas especializadas implementam operações analíticas diretamente sobre os dados.
- **ROLAP (*Relational On-line Analytical Processing*)**: ou OLAP relacional é implementado por meio do armazenamento de dados em SGBD relacional. Cria visões multidimensionais dinâmicas, mas pelo fato de não serem pré-processadas utilizam-se de consultas complexas e com tempo de resposta significativo quando comparados ao MOLAP.
- **HOLAP (*Hybrid On-line Analytical Processing*)**: realiza combinações de ROLAP e MOLAP. Essas implementações variam de acordo com o projetista.

O EDRM foi desenvolvido para utilizar a abordagem ROLAP. Essa decisão foi tomada por considerar que, em geral, os sistemas educacionais já utilizam dados em SGBD relacional e essa estrutura poderia ser aproveitada para o processamento analítico. Contudo, há limitações na utilização de SGBD de forma analítica que podem ser minimizadas com ferramentas ROLAP adequadas, capazes de otimizar consultas complexas e tempo de resposta.

Por fim, é importante destacar que, a representação de cubos OLAP, em geral, é realizada por meio da definição de um metadado modelado para traduzir possíveis operações lógicas em modelos dimensionais ou multidimensionais.

4.5 Business intelligence

O processo de *Business Intelligence* nada mais é do que a utilização de técnicas e ferramentas como, modelagem dimensional, DW, OLAP entre outros, de forma ordenada e consistente para transformar dados brutos em informações potencialmente úteis.

Em 1958, Luhn (1958) citou pela primeira vez o termo *Business Intelligence* (BI) e desde então esse conceito vêm sendo evoluído. BI tornou-se um termo que abrange muitas tecnologias como *Analytics*, Mineração de Dados (MD), *Data Warehouse* (DW), ferramentas de análise de negócio, entre outras.

Loshin (2012) define BI como processos, tecnologias e ferramentas necessárias para transformar dados em informação, informação em conhecimento e conhecimento em planejamento para impulsionar ações de negócios. Na perspectiva de Evelson e Nicolson (2008), BI pode ser definido como um conjunto de metodologias, processos, arquiteturas e tecnologias que transformam dados brutos em informações significativas e úteis.

De acordo com essas definições, BI permite que os usuários de negócios tomem decisões direcionadas a partir dos dados e em tempo real. Essas definições permitem compreender BI como um conjunto de tecnologias que envolvam a transformação de grandes volumes de dados em informação.

A arquitetura básica de BI utiliza vários elementos e técnicas para a transformação de dados em informação. Em geral, a arquitetura pode ser dividida em três principais componentes:

- **Extração, transformação e carga (*Extract Transform Load* – ETL):** processo responsável por extrair dados de sistemas transacionais, efetuar transformações (assim como é feito no pré-processamento de Mineração de Dados) e, por último, armazená-las em um repositório que facilite o acesso às informações.
- **DW (*Data Warehouse*):** repositório de dados analítico que disponibiliza dados estruturados para apoio à decisão.
- **Camada de apresentação:** camada que pode ser composta por diversas técnicas para auxiliar a apresentação da informação armazenada no DW. Os dados disponíveis nessa camada podem ser consumidos por técnicas e ferramentas como relatórios, processamento OLAP, *dashboard*, entre outros.

Vale destacar que não existe uma solução única de BI, ou seja, cada organização possui um cenário particular que busca responder questões específicas. No contexto educa-

cional, a utilização de BI em pesquisas voltadas a Análise Educacional buscam utilizar um conjunto de soluções para auxiliar no processo de tomada de decisão.

Na próxima seção, algumas pesquisas recentes que utilizam modelo dimensional disponibilizados em estruturas de *Data Warehouse* são descritas com o objetivo de ilustrar a aplicabilidade do modelo no contexto educacional.

4.6 Aplicações de data warehouse em sistemas educacionais

Iniciativas de sistemas de apoio à decisão estão, cada vez, unificando instituições a partir de definições estratégicas fundamentais centradas na análise de dados de sistemas educacionais. A utilização de DW para a integração de dados de várias fontes vem possibilitando a estruturação da informação de forma consistente, de modo a aumentar a acessibilidade, visibilidade e utilidade dessas informações.

Sob essa perspectiva, pesquisas que utilizam modelagem dimensional de dados disponibilizados em estruturas de *Data Warehouse* no contexto educacional como, Moscoso-Zea, Andres-Sampedro e Luján-Mora (2016), Pasyeka e Pasyeka (2016), Rudy, Miranda e Suryani (2014), Nebic e Mahnič (2010), Wai e Aung (2009) e Dimokas *et al.* (2008) vêm sendo amplamente exploradas com o objetivo de consolidar dados de interesse considerando informações heterogêneas, em um único repositório de dados.

Dentre essas pesquisas, os modelos dimensionais em estrutura de DW que possuem um maior conjunto de informações com foco no aluno e seu contexto de aprendizagem são os modelos propostos por Moscoso-Zea, Andres-Sampedro e Luján-Mora (2016), Valdiviezo-Díaz *et al.* (2015) e Dimokas *et al.* (2008).

Na pesquisa de Moscoso-Zea, Andres-Sampedro e Luján-Mora (2016) foram realizadas entrevistas com os *stakeholders* para definir os principais processos educacionais que deveriam estar mapeados em um DW. O resultado das entrevistas permitiu a compreensão do cenário e a definição dos dados que seriam extraídos do modelo operacional em um processo de ETL. Um modelo *star schema* foi desenvolvido para realizar análises acerca do conhecimento do alunos, conforme ilustrado na Figura 7.

Nessa pesquisa, os autores não descreveram os indicadores ou objetivos de análises, bem como não há detalhamento dos atributos definidos de cada uma das dimensões. Contudo, foi possível perceber que todas as dimensões possuem as características do aluno, como gênero, localização, carreira, estado civil, aluno, docente, tempo e ensino médio.

Figura 7 – Modelo *star schema* proposto por Moscoso-Zea, Andres-Sampedro e Luján-Mora (2016).

Fonte: Adaptada de Moscoso-Zea, Andres-Sampedro e Luján-Mora (2016).

Outro trabalho relevante nessa linha é o proposto por Dimokas *et al.* (2008), que possui dois modelos no formato *star schema* desenvolvidos para uma universidade. Esses modelos foram projetados com o objetivo de realizar análises departamentais, conforme ilustrado na Figura 8. Segundo os autores, as análises em nível de departamento foram desenvolvidas para que a instituição pudesse medir o progresso anual de cada aluno individualmente e enviar tais informações para entidades governamentais como, por exemplo, o ministério da educação.

Desse modo, a estrutura do primeiro modelo, ilustrado na Figura 8a, possui três dimensões (**aluno**, **curso** e **tempo**) e uma tabela de fato (**notas**). A dimensão **aluno**, possui informação do quinquênio (registro de data), ano de matrícula, idade e nome do currículo. A dimensão **curso**, possui nome do currículo, semestre, tipo, nome do curso e código do curso. Por fim, nas tabelas de fato são armazenadas as quantidades de exames, a média da nota final e o curso.

A segunda estrutura, ilustrada na Figura 8b, possui duas dimensões (**graduado** e **tempo**) e uma tabela de fato (**notas**). A dimensão **graduado**, possui informação do quinquênio (registro de data), ano de matrícula, idade e nome do currículo. A tabela de fato possui o número de graduados e a média da nota final.

A partir da definição dos dois modelos e carga dos dados no DW, as análises foram realizadas por meio de operações OLAP. Essas análises consideraram: (1) média de notas por aluno em um intervalo de tempo; (2) média de alunos graduados em um intervalo de tempo; e (3) quantidade de graduados em um período. Como análise estatística, foi considerado o coeficiente de correlação de *Pearson* para avaliar o grau de correlação entre a notado do aluno e a duração do curso.

Figura 8 – Modelos *star schema* propostos por Dimokas *et al.* (2008). (a) *Star Schema* de notas. (b) *Star Schema* de aluno graduado.Fonte: Dimokas *et al.* (2008).

Na pesquisa realizada por Valdiviezo-Díaz *et al.* (2015), foi desenvolvido um modelo dimensional para a realização do processo de BI a partir dos dados do LMS *Moodle*. O estudo buscou responder basicamente duas questões: (1) É possível definir um modelo de tutoria *online* que pode adaptável ao perfil do aluno? (modelo descritivo); e (2) É possível prever o sucesso de um processo de tutoria *online* para um curso e um aluno? (modelo preditivo). Para isso, foi realizado um estudo de caso considerando um modelo de ensino baseado em competências para a modalidade a distância. A partir dessas questões foram definidos os indicadores de desempenho taxa de sucesso do aluno e comunicação de alunos e professores na plataforma, baseados em mensagens, *chat*, e outros. Por fim, foi desenvolvido o modelo dimensional, no formato *Snowflake Schema*, conforme ilustrado na Figura 9.

Esse modelo possui quatro perspectivas: **usuário**, **curso** (matrícula), **chat** e **fórum de discussão**. Suas principais medidas utilizadas nos experimentos são: **curso**, **módulo**, **ações**, a **taxa de sucesso** (baseada em estado do aluno aprovado ou reprovado), **número de recursos**, **número de comentários do professor para o aluno**, **número de mensagens recebidas**, **número de mensagens lidas**, **número de mensagens de chat**, **número de participantes no chat**, **número de mensagens no fórum de discussão**. Por fim, foram realizadas análises probabilísticas (*Bayesian Network*) utilizando a taxa de performance do modelo dimensional definido.

Figura 9 – Modelo dimensional *Snowflake Schema* proposto por Valdiviezo-Díaz *et al.* (2015).

Fonte: Valdiviezo-Díaz *et al.* (2015).

Cabe destacar que nessa pesquisa os autores não descrevem detalhadamente todos os atributos do modelo. Além disso, as análises realizadas foram pontualmente criadas para responder as duas questões de pesquisas derivadas em dois indicadores de desempenho.

Dentro desse contexto, é importante observar que as pesquisas que utilizam modelo dimensionais como fonte de dados analítica possuem um grande potencial como sistema de apoio à decisão. No entanto, no contexto educacional, essas pesquisas abordam questões específicas considerando análises pontuais, sem a utilização efetiva do modelo, principalmente em relação à utilização de diferentes perspectivas correlacionadas.

Desse modo, é relevante a investigação de modelos de dados capazes de possibilitar uma implementação consistente de dados analíticos considerando diferentes perspectivas, *stakeholders* e níveis de granularidades.

4.7 Considerações finais

Neste capítulo foi apresentada uma síntese das principais estruturas utilizadas para a realização do processo decisório orientada a dados. Esses conceitos serviram de base para a definição do EDRM desenvolvido no escopo deste trabalho.

Em geral, a utilização do *Data Warehouse* é recomendada para sistemas analíticos por possibilitar melhorias de eficiência e flexibilidade, importantes para a recuperação da informação durante o processo decisório. No contexto educacional, as características de um DW são decisivas para resolver problemas inerentes à heterogeneidade dos dados educacionais considerando suas diversas fontes de dados, perspectivas e níveis de granularidade.

Em virtude dessas características, o EDRM foi desenvolvido a partir das definições de modelagem dimensional no formato *Star Schema*, para dar suporte a consultas em um DW. Com isso, tem-se uma estrutura de dados passível de processamento OLAP com seus dados armazenados de forma dimensional em um SGBD relacional (ROLAP), para que instituições de ensino possam utilizar o mesmo tipo de infraestrutura de dados do sistema educacional para armazenar o modelo analítico. Além disso, essa mesma estrutura é adequada a extração de padrões por tarefas de mineração de dados devido a disponibilidade de dados históricos padronizados, integrados e consistentes. Assim, foi possível estruturar dados complexos e heterogêneos em um modelo consistente, flexível e unificado.

No próximo capítulo é apresentado o modelo de referência proposto no escopo deste trabalho de doutorado, considerando o mapeamento dos atributos de interesse e a estruturação do modelo a partir de um repositório informacional de dados ou *Data Warehouse*, que possui um conjunto de esquemas dimensionais conectados por dimensões comuns compatíveis.

MODELO DE REFERÊNCIA DE DADOS EDUCACIONAIS

5.1 Considerações iniciais

Neste capítulo, é apresentado o Modelo de Referência de Dados Educacionais (*Educational Data Reference Model* – EDRM) proposto neste trabalho. Esse modelo é resultante de pesquisas em Análise Educacional (*Learning Analytics*, *Academic Analytics* e Mineração de Dados Educacionais), tendo sido definido para dar subsídio à tomada de decisão orientada a dados por meio de processamento analítico e Mineração de Dados.

O EDRM foi desenvolvido a partir de análises de dados e funcionalidades de sistemas educacionais, para que dados de diferentes tipos de sistemas, mas com características semelhantes, possam ser utilizados de forma unificada, integrada e consistente. Sua arquitetura é estruturada segundo um *Data Warehouse* seguindo um modelo dimensional de dados, projetado para ser uma fonte única de dados integrados e correlacionados, voltada à tomada de decisão. Desse modo, a partir do EDRM, é possível armazenar dados de diferentes fontes, combiná-los e, por fim, realizar análises para que instituições possam desenvolver uma melhor compreensão, rastrear tendências e descobrir lacunas e ineficiências acerca do processo educacional.

Na Seção 5.2 são definidos os atributos relevantes ao EDRM a partir de pesquisas que exploram funcionalidades e características dos dados de sistemas educacionais. Para isso, são exploradas formas de realizar análises nos dados a partir da definição de indicadores de desempenho e sua operacionalização por meio de processamento analítico (*Business Intelligence*, *Data Warehouse* e modelo dimensional de dados) e tarefas de Mineração de

Dados Educacionais. Na Seção 5.3 é detalhado o EDRM considerando suas tabelas fato com medias numéricas e dimensões com atributos que descrevem as medidas. Na Seção 5.4 são definidos alguns dos principais *stakeholders* no contexto educacional. Por fim, na Seção 5.5 são abordados itens que precisam ser considerados por instituições de ensino antes de iniciar a instanciação do EDRM.

5.2 Caracterização dos dados educacionais

Conforme discutido anteriormente, o EDRM é um modelo dimensional no formato *star schema*, composto por tabelas fato que possuem atributos para representar medias numéricas, como a *nota final* do aluno na disciplina, conectadas a dimensões, que possuem atributos para representar o detalhamento das medidas, como a *situação/status da nota* do aluno que pode possuir o valor ‘*aprovado*’ ou ‘*reprovado*’. Basicamente, medida é a associação de uma grandeza numérica a uma característica ou frequência em que o evento ocorre, enquanto métrica é um conjunto de medidas tomadas ao longo de um período.

Sob essa perspectiva, para a definição de atributos relevantes, que possibilitem análises informacionais e analíticas para apoio a decisão baseada em evidências, foram realizadas revisões da literatura a partir de pesquisas de sistemas educacionais, análise educacional e apoio a decisão orientada a dados utilizando *data warehouse*, *business intelligence* e modelagem dimensional de dados. Entende-se por atributos relevantes ao EDRM aqueles que: (1) fazem parte dos dados de sistemas educacionais amplamente utilizados por instituições de ensino; (2) possam ser analisados ou medidos; e (3) possuam algum significado importante ao processo decisório no contexto educacional. Na Figura 10, os passos para a definição dos atributos base do EDRM são ilustrados.

Figura 10 – Processo de caracterização dos dados do EDRM.

Fonte: Elaborada pelo autor.

Inicialmente foram exploradas pesquisas que descrevem funcionalidades de sistemas educacionais. Essas funcionalidades foram organizadas e categorizadas para abranger conjuntos de dados do mesmo assunto para que os atributos de cada categoria seja refinado para compor dimensões e tabelas fato do EDRM.

Em seguida, foram exploradas pesquisas em indicador de desempenho e mineração de dados educacionais. Indicador de desempenho foram considerados por representar um conjunto de medidas e métricas centradas no desempenho, que são utilizadas por tomadores de decisão para a realização de análises precisas acerca do processo educacional. Pesquisas que exploram indicador de desempenho são importantes para a condução da gestão estratégica e monitoramento institucional. Pesquisas em mineração de dados educacionais aplicam técnicas de estatística, aprendizagem de máquina e mineração de dados para analisar dados durante o processo educacional. Desse modo, conjuntos de atributos relevantes ao ensino e aprendizagem são explorados com o objetivo de extrair novos padrões de dados e realizar inferências.

Por fim, foram analisadas pesquisas que abordam a utilização de *business intelligence*, *data warehouse* e modelos dimensionais. Essas pesquisas relatam a utilização de modelagem de dados de diferentes formas para a realização processamento analítico e execução de tarefas de mineração de dados em busca de conhecimento potencialmente útil acerca do processo educacional.

A partir dos resultados dos passos descritos, foram obtidas categorias e subcategorias que possuem conjuntos de atributos separados por assunto para a composição das dimensões e tabelas fato do EDRM. A seguir, são detalhadas as investigações realizadas para a obtenção de atributos relevantes à definição do EDRM.

5.2.1 Dados e funcionalidade de sistemas educacionais

Conforme descrito no Capítulo 2, sistemas de gestão de aprendizagem (*Learning Management Systems* – LMS) e sistemas acadêmicos (*Student Information Systems* – SIS) são sistemas educacionais que se destacam por possuírem dados voltados ao aluno e seu processo de aprendizagem, sendo esse o foco deste trabalho. Sob essa perspectiva, sistemas educacionais dos tipos SIS e LMS constituem elementos fundamentais quando se deseja realizar análises acerca do processo educacional, principalmente considerando experiências de aprendizagem do aluno.

De modo geral, sistemas do tipo LMS possuem dados que podem ser analisados em tempo real, devido às suas características de interatividade entre usuários e formas de comunicação para a realização de atividades colaborativas. A comunicação pode ser

realizada de forma síncrona e assíncrona por meio de ferramentas como *e-mail*, *chat*, fórum de discussão, entre outras (BROWN; DEHONEY; MILLICHAP, 2015). Outras funcionalidades comuns em LMS estão relacionadas à gestão e controle de acesso, ferramentas de apoio à avaliação e acompanhamento do aluno e armazenamento do material didático.

Sistemas do tipo SIS, por sua vez, possuem dados históricos provenientes dos processos administrativos institucionais. Por isso, dados provenientes de SIS são menos voláteis quando comparados aos dados oriundos de LMS. Suas principais funcionalidades são a gestão de *stakeholders*, notas, controle de frequência, ementa do curso, perfil dos docentes, entre outras (BROWN; DEHONEY; MILLICHAP, 2015).

Sistemas educacionais do mesmo tipo possuem várias funcionalidades comuns que resultam em dados com estruturas distintas, porém com informações semelhantes. Em virtude dessas características, a partir de revisões da literatura, trabalhos que descrevem sistemas educacionais amplamente utilizados foram investigados. O objetivo dessa investigação foi determinar as funcionalidades comuns aos mesmos, para identificar dados potenciais ao processo analítico e, assim, compor o EDRM, seja com atributos descritivos ou medidas. Nessa análise, foram consideradas pesquisas que explicam e detalham características e funcionalidades de diversos sistemas. Dentre essas pesquisas podemos citar: *Moodle* (MARCHENKO, 2016; NETO; CASTRO, 2015; IGLESIAS-PRADAS; AZCARATE; AGUDO-PEREGRINA, 2015; VALDIVIEZO-DÍAZ *et al.*, 2015), *TidiaAe* (TIDIA, 2006), *Edx* (MARCHENKO, 2016; PARDOS; KAO, 2015; MIT and Harvard University, 2013) e *Sakai* (LI; WANG, 2010; FARMER; DOLPHIN, 2005; COUNTERMAN *et al.*, 2004).

Uma síntese das principais funcionalidades voltadas ao aluno e seu contexto de aprendizagem pode ser observada na Tabela 5. Nessa tabela, cada uma das funcionalidades possui uma categoria, definida como forma de identificação de um conjunto de dados referentes a um mesmo assunto. Além disso, há uma breve descrição com as características da funcionalidade. Por fim, descreve-se o tipo de sistema educacional predominante para cada funcionalidade. Por exemplo, a funcionalidade **Gestão institucional** lida com dados de **matrícula** e **calendário acadêmico** (itens da coluna descrição) que são itens que possuem características da instituição de ensino, ou seja, da categoria **Institucional** e são predominantemente informações de sistemas acadêmicos e de gestão de aprendizagem.

Esse conjunto de funcionalidades foi categorizado com as principais funcionalidades de sistemas educacionais com foco na experiência de aprendizagem do aluno. Cada uma

Quadro 5 – Principais funcionalidades de sistemas educacionais.

Funcionalidade	Categoria	Descrição	Sistema Pre-dominante
Gestão institucional	Institucional	Possui dados com as características da instituição; forma como a instituição organiza seu processo educacional (cursos, disciplinas, turmas, entre outros); calendário de eventos, calendário acadêmico (dias úteis e dias letivos); informações de ingresso do aluno, formas de matrícula, vagas ocupadas e disponíveis; dados biográficos, ementa; planos pedagógicos e estruturas gerais de ensino e aprendizagem.	SIS e LMS
Realizar comunicação e colaboração	Comunicação e colaboração	Possibilita a interação entre usuários (troca de informação); formação de grupos e compartilhamento de suas produções e conhecimentos. Utilizam ferramentas para: (1) comunicação em tempo real por meio de ferramentas como <i>chat</i> e webconferência e correio eletrônico; (2) comunicação assíncrona por meio de fórum de discussão e ferramenta de <i>e-mail</i> /mensagem; e (3) ferramentas colaborativas como Wikis e e-livros.	LMS
Gestão de <i>stakeholders</i>	Usuários	Gestão de perfil dos <i>stakeholders</i> como aluno, docente, pesquisador, tutor, gestor, administrador, coordenador, entre outros.	SIS e LMS
Gestão de Recursos Educacionais	Recursos Educacionais	Materiais de ensino, aprendizagem e pesquisa (conteúdo para cursos, módulos, livros, vídeos, testes, <i>software</i> e quaisquer outras ferramentas, materiais ou técnicas usadas que permitam o acesso ao conhecimento).	LMS ou sistemas específicos como repositórios de objetos de aprendizagem
Controle de acessos	Segurança	Acesso livre para qualquer visitante, acesso privado para usuários com permissão. O permissionamento é controlado por módulos de autenticação e autorização. Todos os acessos monitorados com atributos de segurança como, data, hora, IP, entre outros.	SIS e LMS
	Acessos	Possibilita a consulta de informações por meio de relatórios como: o número de visitas (total, por página, ao longo do tempo); frequência de postagens, percentual de material lido, entre outros. Esses relatórios podem ser exportados em diversos formatos de arquivos.	
Avaliações e acompanhamento	Habilidade acadêmica	Questionários para avaliações; controle de notas; controle de frequência; controle de <i>feedback</i> .	SIS e LMS
Gestão documentos eletrônicos	Armazenamento	Armazenamento de arquivos em diferentes formatos e estruturas.	LMS

das categorias definidas será refinada para a definição de atributos do modelo dimensional proposto no escopo deste trabalho.

Ressalta-se que as funcionalidades, quando correlacionadas e disponíveis em um modelo analítico consistente e integrado, possuem informações abrangentes e estratégicas para o processo decisório no contexto educacional. Para refinar a caracterização dos dados, na próxima seção são descritas as pesquisas em exploração análise educacional. Ênfase foi dada em pesquisas que abrangem indicador de desempenho educacionais e mineração de dados educacionais.

5.2.2 Dados para análise educacional a partir de indicadores de desempenho e mineração de dados educacionais

Uma vez determinadas as funcionalidades principais dos sistemas educacionais, é necessário determinar quais pesquisas possuem informações importantes à composição do EDRM de maneira a possibilitar o apoio a todas elas. Nesse sentido, foram identificadas pesquisas nas áreas de indicador de desempenho educacional para realizar *Learning Analytics* e *Academic Analytics* e análises de Mineração de Dados Educacionais.

A partir dessas pesquisas, foi possível extrair um conjunto de atributos relacionados às funcionalidades previamente definidas. Considerando esses trabalhos, foram analisadas as principais referências relatadas em busca de atributos relevantes, de maneira que o EDRM possa dar apoio ao processamento analítico e à execução de tarefas de Mineração de Dados Educacionais.

Uma forma de identificar dados que possuem alto potencial à realização de análises é por meio de **indicadores de desempenho educacional**. Segundo Rhys, David e Tim (2010), os esforços em pesquisas voltadas à definição de indicador de desempenho educacional ocorrem em três unidades funcionais: (1) ensino e aprendizagem; (2) pesquisa; e (3) serviços. No entanto, dessas três áreas, a mais investigada refere-se aos ensino e aprendizagem.

Assim, a partir de indicadores criados e utilizados para o Censo da Educação Superior¹ do INEP (2015) e trabalhos identificados na literatura, atributos foram adaptados para serem incorporados ao modelo. Os indicadores de desempenho medidos pelo INEP (2016) estão sumarizados no Capítulo 3 deste trabalho. Exemplos de medidas presentes nos indicadores que são relevantes ao EDRM: número de instituições, total de matrículas (por curso), número de docentes em exercício, relação de matrículas por curso/modalidade, número de alunos concluintes, entre outras. Todas essas medidas podem ser complementadas por agrupadores, como: por tempo (ano, semestre, entre outros), por gênero, por modalidade (presencial, híbrida e a distância), por estado civil, por cor/raça entre outros.

Martin, Sauvageot e Tchatchoua (2011) também realizaram pesquisas que identificam indicador de desempenho para o ensino superior. Esse estudo foi desenvolvido para ser utilizado como guia para a definição de indicador de desempenho, com o objetivo de ajudar instituições a realizarem a gestão e a comunicação de seus resultados. Dentre os casos relatados estão indicadores de desempenho da União Europeia, África do Sul, indicadores definidos pela UNESCO, entre outros. Exemplos de indicador de desempenho relatados nesse estudo são: taxa de matrícula por faixa etária, taxa de alunos que se formaram no

¹ Website Censo da Educação Superior: <http://inep.gov.br/censo-da-educacao-superior>

tempo previsto, proporção de graduados por faixa etária e percentual de abandono por faixa etária.

Cabe destacar que indicadores de desempenho resumem uma quantidade considerável de dados com o objetivo de apresentar uma visão geral de uma situação que está sendo analisada. Contudo, poucas são as instituições que utilizam efetivamente indicadores associados à estratégia institucional, sendo grande parte das análises realizadas de forma exploratória ou *ad hoc*. Por isso, poucos são os estudos que definem indicadores de desempenho e realizam a análises dos dados institucionais baseadas nesses indicadores como forma de orientar a execução de metas e objetivos estratégicos.

Em continuidade à caracterização dos atributos, na sequência são apresentadas pesquisas que utilizam **Mineração de Dados Educacionais** (MDE) para gerar conhecimento potencialmente útil durante o processo decisório. Em MDE os autores Dutt, Ismail e Herawan (2017), Kumar, Singh e Handa (2017), Peña-Ayala (2014), Romero e Ventura (2010), Romero e Ventura (2007) e Baker e Yacef (2009) conduziram diversas pesquisas que consolidam o estado arte de MDE nos últimos dez anos. Nesse sentido, os trabalhos que se destacam são os realizados por Moscoso-Zea, Andres-Sampedro e Luján-Mora (2016), Neto e Castro (2015), Valdiviezo-Díaz *et al.* (2015), Yu e Jo (2014), Sael e Behja (2013), Dimokas *et al.* (2008) e Romero, Ventura e Garcia (2008). Todos os atributos relevantes ao EDRM estão descritos no Quadro 6. Sob essa perspectiva, embora diversas pesquisas tenham sido analisadas, para exemplificar a identificação dos atributos a partir da análises de trabalhos relacionados em MDE, quatro trabalhos são sumarizadas a seguir.

No trabalho de Romero, Ventura e Garcia (2008), foi desenvolvido um estudo de caso a partir dos dados do LMS *Moodle* com o objetivo de realizar mineração de uso. Desse modo, a partir do modelo de dados relacional do LMS, na etapa de pré-processamento, foi definida uma tabela com a sumarização de todos os dados necessários à extração de padrões. Os atributos sumarizados nessa pesquisa são: identificador do curso, quantidade de atividades concluídas, quantidade de atividades aprovados/reprovados (notas em atividades), quantidade de mensagens de *chat* enviadas, quantidade de *posts* enviados ao fórum de discussão, quantidade de *posts* lidos do fórum de discussão, tempo total gasto nas atividades e nota final da atividade do aluno.

Já no trabalho de Neto e Castro (2015), foram realizadas análises para descobrir regras de associação com o objetivo de identificar situações de abandono do curso. Os atributos utilizados nesse experimento são: quantidade de acessos, quantidade de *posts* em fórum de discussão, quantidade de acessos do aluno ao LMS, quantidade de dias entre o início do curso e o primeiro acesso, quantidade de *posts* respondidos por outros alunos, quantidade de *posts* realizados, quantidade de atividades entregues, média de tempo entre a data final

do exercício e sua entrega, quantidade de mensagens enviadas ao professor e quantidade de mensagens respondidas pelo professor.

Valdiviezo-Díaz *et al.* (2015) também estabeleceram um conjunto de atributos para realizar um processo de *Business Intelligence* a partir de dados da modalidade a distância de cursos no LMS *Moodle*. Os principais atributos relatados no estudo são: curso, módulo, ações, taxa de sucesso (baseada em estado do aluno aprovado ou reprovado), número de recursos educacionais, número de comentários do professor para o aluno, número de mensagens recebidas, número de mensagens lidas, número de mensagens de *chat*, número de participantes no *chat* e número de mensagens no fórum de discussão.

No trabalho de Yu e Jo (2014), foi conduzida uma análise de desempenho de um grupo de alunos a partir de análises de regressão linear para determinar quais fatores podem influenciar o desempenho do aluno. Os atributos considerados nas análises são: a frequência de login no LMS, tempo total de estudo no LMS, regularidade do intervalo de aprendizado no LMS, número de *downloads*, quantidade de interações com outros alunos e quantidade de interações com o instrutor.

Por fim, a última etapa de análise de atributos foi realizada a partir de pesquisas que abordam a utilização de **Modelos de Dados de Apoio à Decisão** no contexto educacional. Nessa linha, com os trabalhos de Moscoso-Zea, Andres-Sampedro e Luján-Mora (2016), Pasyeka e Pasyeka (2016), Rudy, Miranda e Suryani (2014), Muntean *et al.* (2011), Nebic e Mahnič (2010), Wai e Aung (2009) e Dimokas *et al.* (2008) foram analisados trabalhos que utilizam *Business Intelligence*, *Data Warehouse* e modelos dimensionais para a estruturação de dados analíticos no contexto educacional. Para exemplificar essa etapa, dois importantes trabalhos são sumarizados.

No trabalho de Nebic e Mahnič (2010), os autores buscaram analisar os dados do LMS *Moodle* que foi personalizado para atender as demandas institucionais. Nessa pesquisa os autores realizaram análises a partir das seguintes métricas: (1) quantidade de novos cursos abertos por instituição e por tempo; (2) quantidade novos usuários inscritos por instituição e por tempo; (3) quantidade de atividades por tipo e por tempo; e (4) correlação de atividade com número de alunos, resultados de aprendizagem e notas ao longo do tempo. Para isso, foi desenvolvido um modelo *star schema* com quatro dimensões (**usuário**, **tempo**, **data** e **curso**) e uma tabela fato (**atividades do usuário**). A tabela fato possui atributos que descrevem ações realizadas no curso, como: **nome da ação**, **endereço IP**, **código de erro**, **atividade**, entre outros. Na dimensão **usuário** são listados os atributos: **login**, **primeiro** e **último nome do usuário**, **senha**, **identificador do usuário**, **e-mail**, **filiação**, **identificador de remoção**, **confirmação do usuário** e **identificador de aceite da política de privacidade**. A dimensão **curso** possui informações que compõem uma hierarquia como: **instituição**, **faculdade**, **departamento**

e curso. Além disso os atributos **nome**, **visibilidade** e **data de criação** para cada item da hierarquia também são descritos. Por fim, a dimensão **data** e a dimensão **hora** possuem atributos com informações de tempo (**dia**, **trimestre**, **mês**, **ano**, etc). Ressalta-se que embora o trabalho não detalhe todos os atributos utilizados, dados de ferramentas de comunicação (fórum de discussão) e atividades (*quiz*) também foram analisadas.

Já Muntean *et al.* (2011) estabeleceram um modelo *star schema* com quatro dimensões (**ferramenta**, **tempo**, **curso** e **pessoa**) e três tabelas fato (**utilização**, **atividade** e **notas**) para analisar dados sob o ponto de vista da utilização do sistema. Os principais objetivos são: investigar a frequência de acessos ao sistema educacional por parte dos usuários, identificar quais são os recursos educacionais mais utilizados e, por fim, identificar quais são os docentes mais ativos no sistema durante o processo educacional. Nesse contexto, o modelo dimensional definido possui a tabela fato **utilização** com os atributos: **número de alunos cadastrados**, **número de alunos ativos**, **identificador do curso** e **duração**. A tabela fato **atividade** possui os atributos: **tipo de ação** (**visualização**, **adição**, **atualização**, **remoção**, etc), **identificador da ferramenta utilizada**, **identificador do usuário** e **identificador do curso**. Por fim, a tabela fato **notas** possui os atributos: **perfil do usuário**, **identificador do curso**, **média de nota** e **se a nota foi satisfatória** (**sim** ou **não**). A dimensão **ferramenta** possui informações de ferramentas como *chat*, *fórum de discussão*, *blog*, *questionário*, *pesquisa*, entre outras. A dimensão **curso** é composta pelo **identificador do curso**, **nome do curso**, **quantidade de créditos**, **categoria**, **programa do curso**, **currículo**, entre outras. A dimensão **pessoa** possui as características do usuário, como: **identificador do usuário**, **perfil** (por exemplo, **administrador**, **coordenador**, **tutor** e **aluno**), **nome**, **gênero** e **data de nascimento**. Por fim, a dimensão **tempo** possui quatro níveis: **dia**, **mês**, **semestre** e **ano**.

5.2.3 Sumarização de atributos do EDRM

Considerando as análises realizadas para a definição dos atributos relevantes ao EDRM, no Quadro 6 são categorizados os principais assuntos levantados nas pesquisas exploradas. A coluna **categoria** foi definida a partir das macro funcionalidades dos sistemas educacionais. Essas categorias foram desmembradas em **subcategorias** para a representação de dados que possuem a mesma natureza. Para cada subcategoria há uma breve descrição, um conjunto de **atributos** e um conjunto de **medidas e métricas** associadas.

Quadro 6 – Atributos definidos a partir de pesquisas nas área de Indicadores de Desempenho, Mineração de Dados Educacionais, *Business Intelligence*, *Data Warehouse* e modelo dimensional de dados educacionais.

Categoria	Subcategoria	Descrição	Atributos	Medidas e Métricas
Usuário	Características Principais	Dados que possuem as principais características dos usuários	Nome e sobrenome, tipo (aluno, professor, tutor, gestor, entre outros) data de nascimento, país, cidade e estado de origem, tipo de usuário (aluno, professor, tutor, gestor, coordenador, entre outros), instituição, modalidade (presencial, híbrida ou a distância), data de início e fim na instituição status do usuário (ativo, cancelado, jubilado, transferido, entre outros), documento de identificação, gênero.	Quantidade de alunos, professores tutores, gestores, coordenadores, entre outros. Quantidade de usuários por modalidade, por instituição de origem, por gênero, por localidade, entre outros.
	Perfil Socioeconômicos	Dados que representam as características da individuais do aluno usuário, principalmente o aluno.	Data e local de nascimento, endereço residencial, cor/raça, estado civil, ocupação do aluno, cidade, estado, país, portador de deficiência, tipo de deficiência, ano inicial e final do ensino fundamental, ano inicial e final do ensino médio, formação do aluno (escolaridade), classificação do ensino médio do aluno (pública, estadual, municipal ou privada) ocupação do responsável, formação do responsável, renda familiar/domiciliar, quantidade de moradores na residência, quantidade de dependentes.	Quantidade de alunos por gênero, por cor/raça, portador de deficiência, por localização (país, cidade, estado), estrangeiros, por faixa de renda familiar, entre outros.
Habilidade Acadêmica	Notas	Dados que descrevem o desempenho do aluno em avaliações e atividades ao longo da vida acadêmica	Média da nota final, notas em atividades, classificação ou <i>ranking</i> , horas ou créditos completos, estágios.	Quantidade de alunos com notas postadas, classificação dos x melhores alunos, quantidade de alunos concluintes, média de notas por aluno por turma, disciplina, curso.
Institucional	Matrícula	Dados que descrevem a inscrição/matricula em cursos ou disciplinas.	Data de início da matrícula na disciplina, data de término da matrícula na disciplina, data de início da matrícula no curso, data de término da matrícula no curso, forma de ingresso, status da matrícula.	Quantidade de matrículas por modalidade, por forma de ingresso, por status.

Categoria	Subcategoria	Descrição	Atributos	Medidas e Métricas
	Curso	Dados com todas as características institucionais de implementação de cursos em seu departamento, execução de turmas, entre outros.	Instituição, departamento, curso, grau acadêmico do curso (bacharelado, licenciatura, tecnológico, outros), disciplina, turma, disciplina pré-requisito, formato da disciplina, modalidade (presencial, híbrida ou a distância), data de criação do curso, data de fechamento do curso, data de criação da disciplina, data de fechamento da disciplina, data de criação da turma, data de fechamento da turma, carga horária da turma, carga horária da disciplina.	Quantidade de instituições, quantidade de carga horária executada por usuário, quantidade de departamentos, quantidade de cursos, quantidade de turmas, quantidade de disciplinas, carga horária total da turma, carga horária total da disciplina.
	Pesquisa	Dados com características de pesquisas de <i>benchmarking</i> interinstitucionais.	Instituição, produções bibliográficas, nacionalidade do aluno, carga horária realizada.	Quantidade de produções bibliográficas, quantidade de alunos, quantidade de cursos, quantidade de disciplinas, quantidade de docentes estrangeiros, quantidade de alunos estrangeiros, crescimento institucional, posição nos <i>rankings</i> mundiais.
Comunicação e Colaboração	Comunicação	Dados que descrevem ações realizadas por <i>stakeholders</i> de forma interativas. Acessos, postagens e leituras em <i>chat</i> , fórum de discussão, <i>quiz</i> e mensagens.	Descrição da mensagem, ação realizada (leitura ou envio), data de envio da mensagem, data de leitura da mensagem.	Quantidade de comunicações enviadas ou lidas, quantidade de <i>post</i> em fórum, quantidade de mensagens enviadas por <i>chat</i> , quantidade de alunos que se comunicaram com professores ou tutores, quantidade de comunicações enviadas ao professor, quantidade de <i>e-mails</i> enviados ou lidos, quantidade de <i>quizzes</i> concluídos.
Segurança e Acessos	Acessos	Dados que possuem os rastros de acessos dos usuários.	Origem, controle de acesso, tempo <i>online</i> .	Quantidade de acessos por página, por módulo, por usuário, por IP, por componente. Quantidade de acessos por instituição, departamento, setor, curso, disciplina e turma.
Administrativo e Financeiro	Administrativo e Financeiro	Dados com as características financeiras do ponto de vista do aluno.	Montante de auxílio financeiro, valor pago, valor do investimento, pendências financeiras.	Quantidade de alunos bolsistas, quantidade de alunos pagantes, quantidade de alunos com pendências financeiras.
Recursos Educacionais	Recursos Educacionais	Materiais educacionais disponibilizados aos alunos	Materiais didáticos, multimídia, entre outros.	Quantidade de materiais disponibilizados, quantidade de materiais acessados.

Diante dos atributos, medidas e métricas definidas, cada uma das categorias e subcategorias foram analisadas considerando os indicadores de desempenho educacionais levantados na literatura. Todas as subcategorias definidas anteriormente foram consideradas e estão descritas no Quadro 7. A subcategoria **TM** (Tempo) foi adicionada devido à necessidade de orientação no tempo de modelos dimensionais de dados. Isso significa que qualquer análise realizada estará associada a algum período do tempo em diferentes níveis de granularidade. Desse modo, cada indicador pode pertencer a mais de uma subcategoria simultaneamente representando suas possíveis correlações que, para o modelo dimensional do EDRM, são diferentes perspectivas e níveis de granularidade que podem ser explorados.

Quadro 7 – Subcategorias definidas para avaliar possíveis correlações de perspectivas de indicadores de desempenho.

Sigla	Subcategoria
US	Usuário – características principais considerando os diferentes tipos: aluno, professor, coordenador, gestor, entre outros.
DS	Usuário – dados socioeconômicos
NT	Notas
CR	Curso – instituição, departamento, setor, curso, disciplina, turma, entre outros.
MT	Matrícula
CM	Comunicação
AC	Acesso
PS	Pesquisa
RE	Recursos Educacionais
AF	Administrativo e Financeiro
TM	Tempo – ano, semestre, trimestre, mês, entre outros.

Considerando essas subcategorias, no Quadro 8 são descritos os indicadores de desempenho correlacionados com essas subcategorias. Um exemplo é a subcategoria *matrícula* que pode ser monitorada por meio de indicadores como: (i) *taxa de matrícula por gênero*, onde *gênero* é um atributo da subcategoria *usuário*; (ii) *taxa de matrícula por idade e por ano*, onde *idade* é um atributo da subcategoria *usuário* e *ano* da subcategoria *tempo*; (iii) *taxa de matrícula por semestre e por nível socioeconômico*, onde *semestre* é um atributo da subcategoria *tempo* e *nível sócio econômico* da subcategoria *dados socioeconômicos*; e (vi) *taxa de matrícula por instituição e por ano*, onde *instituição* é um atributo do grupo *curso* e *ano* do grupo *tempo*.

Outro exemplo está na subcategoria *notas*, quando se deseja realizar análises preditivas para o indicador *taxa de aprovação/reprovação*. Esse indicador pode ser analisado da seguinte forma: (i) *predição do aluno ser aprovado na disciplina X* considerando como atributo as *notas do aluno* em disciplinas semelhantes já cursadas; e (ii) *predição do aluno ser aprovado no curso*, ou seja, passar em todas disciplinas. Nesse caso, para realizar essa análise pode-se utilizar atributos de *acessos* e *notas dos alunos*.

Quadro 8 – Análises de medidas e indicadores de desempenho educacional.

Categoria	Exemplos de indicadores e métricas	(1) US	(2) NT	(3) CR	(4) CM	(5) AC	(6) MT	(7) DS	(8) PS	(9) AF	(10) RE	(11) TM
Habilidade Acadêmica	Taxa de aprovação / reprovação	x	x	x								x
	Taxa de desistência real	x	x	x		x						x
	Taxa de possível desistência	x	x	x		x						x
	Taxa de evasão real	x	x	x		x						x
	Taxa de possível evasão	x	x	x		x		x				x
	Tempo médio de titulação do aluno concluinte	x	x	x			x					x
	Média de notas finais	x	x	x								x
Segurança e Relatórios (Controle de Acessos)	Média de frequência de acesso	x		x		x						x
	Média de horas que o aluno acessou o sistema	x		x		x					x	x
	Percentual de material acessado	x		x		x						x
	Quantidade de visitas por página	x		x		x						x
	Quantidade de alunos logados por turno	x		x		x						x
	Percentual de alunos que acessam via dispositivos móveis	x		x		x						x
Comunicação	Média de tempo <i>online</i>	x		x	x	x						x
	Percentual de participação	x		x	x	x						x
	Frequência de postagens	x		x	x	x						x
Institucional (Matrícula)	Taxa de matrícula no curso	x		x			x					x
	Percentual de trancamento de matrícula	x		x			x					x
	Percentual de rematrículas	x		x			x					x
Perfil do Usuário	Percentual de alunos provenientes de ensino público e de bairros de baixa renda a serem evadidos	x	x	x			x	x				x
Institucional (Administrativo)	Taxa de crescimento institucional (carga horária realizada, quantidade de produções bibliográficas, quantidade de alunos, quantidade de disciplinas, etc)	x		x		x	x	x	x		x	x
	Quantidade de usuários por perfil (docente, aluno, coordenador, etc)	x										x
	Média de horas aula	x		x								x
	Taxa de ocupação de sala	x		x								x
	Tamanho médio das turmas	x		x								x
	Número de projetos de iniciação científica	x		x								x

A partir das pesquisas realizadas foi possível definir os atributos importantes ao modelo dimensional do EDRM considerando pesquisas que definem indicadores de desempenho para nortear análises de metas institucionais, bem como realizam processamento analítico e análises preditivas para a descoberta de conhecimento potencialmente úteis no contexto educacional. Na próxima seção é apresentada a especificação do EDRM segundo um modelo dimensional do formato *star schema*. São detalhadas as tabelas fato e dimensões definidas considerando as categorias, subcategorias e atributos definidos.

5.3 Modelo dimensional star schema do EDRM

O EDRM proposto neste trabalho foi desenvolvido para ser um modelo de referência centrado no aluno, cujo objetivo é fornecer subsídio ao processo de apoio a decisão orientada a dados por meio de análises relevantes ao processo educacional. Sua estrutura foi definida como uma rede de esquemas dimensionais entrelaçados por dimensões comuns compatíveis, considerando os diversos assuntos modelados. O EDRM é um modelo dimensional no formato *star schema* que possui um conjunto de tabelas fato, também conhecidas como constelação de fatos, em que várias tabelas fato podem possuir um conjunto de dimensões exclusivas e/ou comuns. Em geral, tabelas fato armazenam as medidas numéricas relevantes ao negócio, enquanto que as dimensões possuem o detalhamento do assunto em vários níveis de granularidade.

O modelo dimensional do EDRM remete a seis esquemas no formato *star schema*, conforme ilustrado na Figura 11. Isso significa que o modelo dimensional do EDRM é composto por seis *Data Marts* (DM) com as seguintes perspectivas:

- (1) **Institucional (usuário):** perspectiva desenvolvida para a realização de análises institucionais e administrativas a partir do cruzamento de dados dos usuários para realizar análises institucionais considerando característica de matrícula e dados socioeconômicos dos usuários. Ao realizar análises a partir dessa perspectiva, são explorados alunos matriculados que podem possuir dados socioeconômicos.
- (2) **Institucional (curso):** perspectiva desenvolvida para a realização de análises institucionais e administrativas a partir do cruzamento de dados institucionais ao longo do tempo. Esse tipo de perspectiva visa garantir uma visão estratégica da instituição, possibilitando, ainda, análises interinstitucionais. Ao realizar análises a partir dessa perspectiva, são exploradas características institucionais para todos os usuários.
- (3) **Atividades:** perspectiva desenvolvida para possibilitar análises a partir de atividades desenvolvidas por alunos em cursos, disciplinas e/ou turmas. Ao realizar análises a

Figura 11 – Modelo *Star Schema* do EDRM.

Fonte: Elaborada pelo autor.

partir dessa perspectiva, são explorados apenas usuários que realizaram atividades a partir do sistema educacional.

- (4) **Notas:** perspectiva desenvolvida para dar subsídio a análises referentes a habilidades acadêmicas, ou seja, demonstrar o desempenho do aluno ao longo da vida acadêmica. Ao realizar análises a partir dessa perspectiva, são explorados apenas usuários que

possuem notas, ou seja, alunos que são ou foram ativos em algum momento da vida acadêmica.

- (5) **Recursos educacionais:** perspectiva que possui as principais características dos recursos educacionais, bem como suas informações de acesso e utilizações. Desse modo é possível realizar análises de disponibilidade e utilização eficiente dos recursos. Ao realizar análises a partir dessa perspectiva, são explorados apenas usuários que utilizam ou já utilizaram recursos educacionais.
- (6) **Estatística de acesso:** perspectiva que possui dados com os rastros de acesso dos usuários. Essa perspectiva pode ser utilizada para a extração de padrões de uso de sistemas educacionais, também conhecido como mineração de uso. Nessa perspectiva é possível também realizar análises generalistas, ou seja, da instituição como um todo devido à disponibilidade dos dados em seu menor nível de granularidade. Assim, essa perspectiva invariavelmente deverá possuir o maior volume de dados do modelo, devido à sua concentração de dados históricos de ações realizadas nos sistemas educacionais analisados. Ao realizar análises a partir dessa perspectiva, são explorados apenas usuários que utilizaram o sistema educacional ao menos uma vez.
- (7) **Comunicação:** perspectiva que possui dados de atividades interativas como comunicação e colaboração realizadas por *stakeholders*. Em geral, as funcionalidades de comunicação são as mais utilizadas e, por isso, foram mapeadas nessa perspectiva. Ao realizar análises a partir dessa perspectiva, são explorados apenas usuários que utilizaram pelo menos uma vez as ferramentas de comunicação como *chat*, *fórum*, *e-mail*, entre outros. Nesse caso, é comum encontrar os perfis de alunos e docentes/tutores.

Considerando a definição do modelo dimensional de dados, essas perspectivas possuem um conjunto de medidas numéricas relevantes ao negócio. Desse modo, cada uma dessas medidas é armazenada em tabelas fato para que possam ser exploradas. Como forma de otimização do modelo e compatibilidade de dados comuns em diferentes perspectivas, foram definidas cinco dimensões em conformidade (*conformed dimensions*). Entende-se por dimensões em conformidade aquelas que possuem um mesmo sentido semântico para os diversos esquemas dimensionais de diferentes DM que se cruzam (INMON, 2005; BARBIERI, 2011). O EDRM foi desenvolvido utilizando esse tipo de dimensão com o objetivo de garantir que as informações sejam compatíveis em diferentes perspectivas. As dimensões em conformidade definidas neste trabalho são:

- **Usuário:** possui as principais características de usuários de diferentes perfis (aluno, professor, gestor, entre outros).

- **Instituição:** possui as características institucionais, como descrição da instituição, seu nome, categoria administrativo, entre outras.
- **Curso:** possui as características abaixo da hierarquia de instituição, como descrição do departamento, do curso, da disciplina, entre outras.
- **Tempo:** possui características que descrevem o tempo em diversos níveis de granularidade.
- **Acesso:** possui informação com as características de acesso ao sistema educacional.

Por fim, foram mapeadas dimensões exclusivas, ou seja, dimensões que possuem atributos referentes a uma única perspectiva ou tabela fato:

- **Mensagem:** possui características provenientes de ferramentas de comunicação dos sistemas educacionais.
- **Socioeconômico:** possui características importantes associadas aos indicadores populacionais, sociais e econômicos das instituições analisadas.
- **Matrícula:** possui informações utilizadas na matrícula do aluno.
- **Atividade:** possui informações acerca das atividades realizadas pelos alunos em cursos, disciplinas e/ou turmas.
- **Nota:** possui as informações com as características da nota, por exemplo, o docente que aferiu a nota, entre outras.

Em geral, tanto as dimensões comuns como as em conformidade são características descritivas das medidas definidas nas tabelas fato. Isso significa que, por exemplo, ao realizar análises exploratórias a respeito da *média de notas dos alunos em uma determinada disciplina por ano por docente*, a tabela fato **nota** é utilizada para retornar o valor quantitativo das notas dos alunos (operação aritmética de média), os dados dos docentes são retornados a partir da dimensão **usuário**, na dimensão **curso** são recuperadas as disciplinas e na dimensão **tempo** está o período de tempo analisado, ou seja, o ano.

Na Figura 12 está detalhado o modelo *star schema* completo do EDRM. Nesse modelo estão ilustradas todas as **tabelas fato**, **dimensões em conformidade** e **dimensões comuns** e seus relacionamentos. Todos os atributos definidos no modelo estão listados no Quadro 9.

Como boa prática de projetos de *Data Warehouse*, Kimball e Ross (2013) recomendam a realização de junções entre dimensões e fatos por meio de chaves substitutas denominadas *Surrogate Key* (SK). De modo geral, SK são números inteiros atribuídos a cada entidade modelada, que são utilizados como substitutos de uma chave natural do modelos relacionais (processamento operacional). Isso significa que todas as junções entre dimensões e fatos são baseadas em SK ao invés de chaves naturais, cabendo à lógica de extração dos dados fazer essa correspondência. Por isso, esses conceitos são utilizados no EDRM para estabelecer o relacionamento de tabelas fatos e dimensões.

Os atributos do modelo estão padronizados da seguinte forma: (1) chaves primárias de dimensões e chaves estrangeiras de tabelas fato possuem o sufixo *sk*; (2) chaves primárias de tabelas fato possuem o prefixo *pk*; (3) atributos numéricos possuem o prefixo *num*; (4) atributos descritivos/textuais possuem o prefixo *des*; e (5) atributos que são chaves naturais/identificadores do modelo transacional possuem o prefixo *idt*. Por fim, dimensões comuns e em conformidade possuem o prefixo *D* e tabelas fato possuem o prefixo *F*.

Quadro 9 – Dimensões em conformidade e comuns do modelo *star schema* do EDRM.

Relação	Atributos	Domínio	Descrição
D.TEMPO (conformed)	tempo_sk	BIGINT	Primary Key
	dt_data	DATETIME	Todas as datas em um intervalo de anos
	num_ano	INTEGER	Número que identifica o ano
	num_mes	INTEGER	Número que identifica o mês - intervalo de 1 a 12
	num_dia_mes	INTEGER	Número que identifica o dia do mês - intervalo de 1 a 31
	num_dia_ano	INTEGER	Número que identifica o dia do ano - intervalo de 1 a 365
	num_dia_semana	INTEGER	Dia da semana - intervalo de 1 a 7
	des_dia_semana	VARCHAR	Descrição do dia da semana - segunda-feira, terça-feira, quarta-feira, quinta-feira, sexta-feira, sábado e domingo
	num_semana_ano	INTEGER	Número que identifica a semana no ano
	num_trimestre	INTEGER	Número que identifica o trimestre - intervalo de 1 a 3
	des_trimestre	VARCHAR	Descrição do trimestre - TRIM1, TRIM2 e TRIM3
	num_semestre	INTEGER	Número que identifica o semestre - intervalo de 1 a 6
	des_semestre	VARCHAR	Descrição do semestre - SEM1, SEM2, SEM3, SEM4, SEM5, SEM6
D.USUARIO (conformed)	usuario_sk	BIGINT	Primary Key
	idt_usuario	BIGINT	Identificador do usuário
	des_nome	VARCHAR	Nome do usuário

Relação	Atributos	Domínio	Descrição
	fl_status_usuario	VARCHAR	Status do usuário - ativo, cancelado, em exercício, afastado
	dt_nascimento	DATETIME	Data de nascimento do usuário
	des_instituicao	VARCHAR	Descrição da instituição do usuário
	des_perfil_usuario	VARCHAR	Descrição do perfil do usuário - aluno, docente, monitor, tutor, gerente, docente avançado e administrador, nc
	des_cidade_usuario	VARCHAR	Nome da cidade do usuário
	des_estado_usuario	VARCHAR	Nome do estado do usuário
	des_pais_usuario	VARCHAR	Nome do país do usuário
	fl_genero	VARCHAR	Descrição do Gênero - Masculino (M), Feminino (F), nc
	dt_inicio	DATETIME	Data de início do usuário na instituição
	dt_fim	DATETIME	Data de fim do usuário na instituição
D_INSTITUICAO (conformed)	instituicao_sk	BIGINT	Primary Key
	idt_instituicao	BIGINT	Identificador da instituição
	des_sigla_instituicao	VARCHAR	Sigla da instituição
	des_nome_instituicao	VARCHAR	Nome completo da instituição
	des_instituicao_categoria_administrativa	VARCHAR	Categoria da instituição - federal, municipal, privada
	des_pais_instituicao	VARCHAR	Descrição do país da instituição
	des_cidade_instituicao	VARCHAR	Descrição da cidade da instituição
	des_estado_instituicao	VARCHAR	Descrição do estado da instituição
	des_instituicao_modalidade	VARCHAR	Modalidade da instituição - presencial, híbrida, a distância, nc
	dt_criacao_instituicao	DATETIME	Data de criação da instituição
D_CURSO (conformed)	curso_sk	BIGINT	Primary Key
	idt_curso	BIGINT	Identificador do curso
	idt_departamento	BIGINT	Identificador do departamento
	idt_disciplina	BIGINT	Identificador da disciplina
	idt_disciplina_pre_requisito	BIGINT	Identificador da disciplina pré requisito
	idt_turma	BIGINT	Identificador da turma
	des_departamento	VARCHAR	Nome do departamento
	des_sigla_departamento	VARCHAR	Sigla do departamento
	des_nome_curso	VARCHAR	Nome do curso
	des_disciplina	VARCHAR	Nome da disciplina
	des_disciplina_modalidade	VARCHAR	Modalidade da disciplina - presencial, híbrida, a distância, nc
	des_turma	VARCHAR	Nome da turma
	des_turma_modalidade	VARCHAR	Modalidade da turma - presencial, híbrida, a distância, nc
	des_curso_grau_academico	VARCHAR	Grau acadêmico do curso - bacharelado, licenciatura, tecnólogo
	num_disciplina_carga_horaria	DECIMAL	Carga horária da disciplina
	num_turma_carga_horaria	DECIMAL	Carga horária da turma
	dt_criacao_curso	DATETIME	Data de criação do curso
	dt_inicio_curso	DATETIME	Data de início do curso
	dt_fim_curso	DATETIME	Data de conclusão do curso
	dt_criacao_disciplina	DATETIME	Data de criação da disciplina

Relação	Atributos	Domínio	Descrição
	dt_inicio_disciplina	DATETIME	Data de início da disciplina
	dt_fim_disciplina	DATETIME	Data de conclusão da disciplina
	dt_criacao_turma	DATETIME	Data de criação da turma
	dt_inicio_turma	DATETIME	Data de início da turma
	dt_fim_turma	DATETIME	Data de conclusão da turma
D.SOCIOECONOMICO	socioeconomico_sk	BIGINT	Primary Key
	idt_usuario	BIGINT	Identificador do usuário
	des_endereco_residencial	VARCHAR	Endereço residencial
	des_raca	VARCHAR	Descrição da cor/raça - branca, parda, amarela, indígena, nc
	des_estado_civil	VARCHAR	Estado civil
	des_ocupacao	VARCHAR	Nome da profissão/ocupação do usuário
	des_escolaridade	VARCHAR	Nível de escolaridade do usuário - doutorado, mestrado, curso superior completo, curso superior incompleto, ensino médio completo, ensino médio incompleto, ensino fundamental completo, ensino fundamental incompleto, nc
	des_resp_ocupacao	VARCHAR	Nome da profissão/ocupação do responsável
	des_resp_escolaridade	VARCHAR	Nível de escolaridade do responsável - doutorado, mestrado, curso superior completo, curso superior incompleto, ensino médio completo, ensino médio incompleto, ensino fundamental completo, ensino fundamental incompleto, nc
	num_renda_familiar	DECIMAL	Valor da renda familiar do usuário
	num_integrates_residencia	INTEGER	Quantidade de integrantes na residência do usuário
	des_cidade_natal	VARCHAR	Cidade de nascimento do usuário
	des_estado_natal	VARCHAR	Estado de nascimento do usuário
	des_pais_natal	VARCHAR	País de nascimento do usuário
	dt_nascimento	DATETIME	Data de nascimento do usuário
	fl_genero	VARCHAR	Descrição do Gênero - Masculino (M), Feminino (F), nc
	fl_portador_deficiencia	VARCHAR	Se o usuário é portador de deficiência - S (sim), N (não), nc
	des_tipo_deficiencia	VARCHAR	Descrição do tipo de deficiência do usuário
	des_documento_identificacao	VARCHAR	Código do documento de identificação do usuário
D.ACESSO	acesso_sk	BIGINT	Primary Key
	idt_acesso	BIGINT	Identificador do acesso
	des_ip	VARCHAR	Protocolo de Internet (Internet Protocol - IP)
	des_acao	VARCHAR	Ação realizada pelo usuário - atualização, visualização, criação, login, logout, nc
	des_modulo	VARCHAR	Módulo do sistema educacional que a ação foi realizada

Relação	Atributos	Domínio	Descrição
	des_url	VARCHAR	URL do sistema educacional acessada
	des_nome_evento	VARCHAR	Nome do evento realizado
	des_origem	VARCHAR	Origem da ação - web, mobile, nc
	des_componente	VARCHAR	Nome do componente acessado
D_MENSAGEM	mensagem_sk	BIGINT	Primary Key
	idt_mensagem	BIGINT	Identificador da mensagem de <i>e-mail</i>
	idt_chat	BIGINT	Identificador da mensagem de <i>chat</i>
	idt_forum	BIGINT	Identificador da mensagem no fórum de discussão
	idt_curso	BIGINT	Identificador do curso
	idt_disciplina	BIGINT	Identificador da disciplina
	idt_turma	BIGINT	Identificador da turma
	des_tipo_mensagem	VARCHAR	Tipo da ação da mensagem - lida, enviada, nc
	des_mensagem_curta	VARCHAR	Mensagem completa
	des_mensagem_longa	VARCHAR	Mensagem reduzida
D_MATRICULA	matricula_sk	BIGINT	Primary Key
	idt_matricula	BIGINT	Identificador da matrícula
	idt_usuario	BIGINT	Identificador do usuário
	idt_curso	BIGINT	Identificador do curso
	idt_disciplina	BIGINT	Identificador da disciplina
	dt_curso_matricula	DATETIME	Data da matrícula do usuário no curso
	dt_disciplina_matricula	DATETIME	Data da matrícula do usuário na disciplina
	des_tipo_ingresso	VARCHAR	Forma de ingresso - processo seletivo, portador de diploma, transferência, nc
	fl_matricula_disciplina_status	VARCHAR	Status da matrícula na disciplina - realizada, cancelada, em andamento, nc
	fl_matricula_curso_status	VARCHAR	Status da matrícula no curso - realizada, cancelada, em andamento, nc
D_NOTA	nota_sk	BIGINT	Primary Key
	idt_nota	BIGINT	Identificador da nota
	idt_curso	BIGINT	Identificador do curso
	idt_disciplina	BIGINT	Identificador da disciplina
	idt_turma	BIGINT	Identificador da turma
	fl_status_nota	VARCHAR	Status da nota - AP (aprovado), RN (reprovado), nc
D_ATIVIDADE	atividade_sk	BIGINT	Primary Key
	idt_curso	BIGINT	Identificador do curso
	idt_disciplina	BIGINT	Identificador da disciplina
	idt_turma	BIGINT	Identificador da turma
	idt_atividade	BIGINT	Identificador da atividade
	des_atividade	VARCHAR	Descrição da atividade
F_ATIVIDADE	atividade_pk	BIGINT	Primary Key
	instituicao_sk	BIGINT	Surrogate Key
	curso_sk	BIGINT	Surrogate Key
	atividade_sk	BIGINT	Surrogate Key
	usuario_sk	BIGINT	Surrogate Key

Relação	Atributos	Domínio	Descrição
	tempo_sk	BIGINT	Surrogate Key
	num_qtd_atividades	INTEGER	Quantidade de atividade
	num_nota_final	DECIMAL	Nota final da atividade
	num_avg_tempo_atividade	DECIMAL	Média de tempo da atividade
	num_tempo_atividades	INTEGER	Tempo total da atividade
	num_qtd_usuarios	INTEGER	Quantidade de usuários que realizam atividades
	num_qtd_cursos	INTEGER	Quantidade de cursos que realizam a atividade
	num_qtd_disciplinas	INTEGER	Quantidade de disciplinas que realizam a atividade
	num_qtd_turmas	INTEGER	Quantidade de turmas que realizam a atividade
	num_qtd_tentativas_resposta	INTEGER	Quantidade de tentativas de resposta da atividade
	dt_atividade	DATETIME	Data de realização da atividade
F_- INSTITUCIONAL _USUARIO	institucional_usuario_pk	BIGINT	Primary Key
	usuario_sk	BIGINT	Surrogate Key
	instituicao_sk	BIGINT	Surrogate Key
	curso_sk	BIGINT	Surrogate Key
	tempo_sk	BIGINT	Surrogate Key
	matricula_sk	BIGINT	Surrogate Key
	socioeconomico_sk	BIGINT	Surrogate Key
	num_qtd_frequencia	INTEGER	Quantidade de frequências do aluno
	num_qtd_ausencia	INTEGER	Quantidade de ausências do aluno
	num_renda_familiar	DECIMAL	Valor da renda familiar do usuário
	num_integrantes_residencia	DECIMAL	Quantidade de integrantes na residência do usuário
	num_qtd_usuarios	INTEGER	Quantidade de usuários
	num_qtd_matriculas_curso	INTEGER	Quantidade de matrículas realizadas no curso
	num_qtd_matriculas_disciplina	INTEGER	Quantidade de matrículas realizadas na disciplina
	dt_matricula	DATETIME	Data da matrícula
F_- INSTITUCIONAL _CURSO	institucional_curso_pk	BIGINT	Primary Key
	curso_sk	BIGINT	Surrogate Key
	tempo_sk	BIGINT	Surrogate Key
	usuario_sk	BIGINT	Surrogate Key
	num_disciplina_carga_horaria	DECIMAL	Carga horária da disciplina
	num_curso_carga_horaria	DECIMAL	Carga horária do curso
	num_qtd_cursos	INTEGER	Quantidade de cursos
	num_qtd_disciplinas	INTEGER	Quantidade de disciplinas
	num_qtd_turmas	INTEGER	Quantidade de turmas
	num_qtd_instituicao	INTEGER	Quantidade de instituições
	dt_institucional_curso	DATETIME	Data de criação da turma ou disciplina ou curso
F_ESTADISTICA _ACESSO	estatistica_acesso_pk	BIGINT	Primary Key
	usuario_sk	BIGINT	Surrogate Key
	instituicao_sk	BIGINT	Surrogate Key
	curso_sk	BIGINT	Surrogate Key
	tempo_sk	BIGINT	Surrogate Key

Relação	Atributos	Domínio	Descrição
	acesso_sk	BIGINT	Surrogate Key
	num_qtd_acessos	INTEGER	Quantidade de acessos realizados ao sistema educacional
	num_qtd_cursos	INTEGER	Quantidade de cursos acessados
	num_qtd_usuarios	INTEGER	Quantidade de usuários que acessaram o sistema educacional
	num_qtd_disciplinas	INTEGER	Quantidade de disciplinas acessadas
	num_qtd_ips	INTEGER	Quantidade de IPs distintos que acessaram o sistema educacional
	dt_acesso	DATETIME	Data de acesso
F_- COMUNICACAO	comunicacao_pk	BIGINT	Primary Key
	usuario_sk	BIGINT	Surrogate Key
	curso_sk	BIGINT	Surrogate Key
	tempo_sk	BIGINT	Surrogate Key
	mensagem_sk	BIGINT	Surrogate Key
	num_qtd_comunicacao	INTEGER	Quantidade de comunicações realizadas por alunos indiferente da ferramenta
	num_qtd_usuarios	INTEGER	Quantidade de usuários que utilizaram ferramentas de comunicação
	num_qtd_cursos	INTEGER	Quantidade de cursos que os usuários se comunicaram
	num_qtd_disciplinas	INTEGER	Quantidade de disciplinas que os usuários se comunicaram
	num_qtd_msg_forum	INTEGER	Quantidade de mensagens enviadas via fórum de discussão
	num_qtd_msg_chat	INTEGER	Quantidade de mensagens enviadas via <i>chat</i>
	num_qtd_msg_email	INTEGER	Quantidade de mensagens enviadas via <i>e-mail</i>
	num_qtd_topicos_forum	INTEGER	Quantidade de tópicos de discussão de fórum
	dt_comunicacao	DATETIME	Data que a comunicação foi realizada
F_NOTA	nota_pk	BIGINT	Primary Key
	usuario_sk	BIGINT	Surrogate Key
	instituicao_sk	BIGINT	Surrogate Key
	curso_sk	BIGINT	Surrogate Key
	tempo_sk	BIGINT	Surrogate Key
	nota_sk	BIGINT	Surrogate Key
	num_qtd_notas	INTEGER	Quantidade de notas publicadas
	num_qtd_usuarios	INTEGER	Quantidade de usuários que utilizaram que possuem notas publicadas
	num_qtd_cursos	INTEGER	Quantidade de cursos que os usuários se possuem notas publicadas
	num_qtd_disciplinas	INTEGER	Quantidade de disciplinas que os usuários possuem notas publicadas
	num_qtd_turmas	INTEGER	Quantidade de turmas que os usuários possuem notas publicadas
	num_qtd_instituicoes	INTEGER	Quantidade de instituições que os usuários possuem notas publicadas
	dt_nota	DATETIME	Data de de publicação da nota

Figura 12 – Modelo *Star Schema* completo do EDRM

Fonte: Elaborada pelo autor.

É importante salientar que cada *Data Mart* pode ser explorado individualmente ou em conjunto. Isso significa que é possível associar análises de **notas** do aluno com a **estatística de acesso** ou com sua característica de **comunicação**, por exemplo. Essa correlação entre perspectivas é possível devido ao conjunto de *Data Marts* definidos no EDRM. Esses *Data Marts*, quando correlacionados por meio de um cubo virtual, possibilitam análises de dados

entre perspectivas distintas. Desse modo, análises abrangentes, considerando diferentes perspectivas correlacionadas tornam-se viáveis e, com isso, possibilitam a correlação de informações potenciais durante o processo decisório.

Adicionalmente, considerando os níveis de *Learning Analytics* propostos por Shum (2012), descritos no Capítulo 3, o EDRM foi desenvolvido para possibilitar sua abrangência segundo os níveis: (1) **micro**: permite realizar análises em nível de **usuário** individual, considerando o contexto que esse usuário está inserido; (2) **meso**: permite realizar análises em nível **institucional** para a obtenção de melhorias administrativas e organizacionais; e (3) **macro**: permite o compartilhamento de dados, a fim de possibilitar análises **interinstitucionais**.

Sob essa perspectiva, análises no EDRM podem ser realizadas segundo os três níveis. O primeiro nível (micro) refere-se à exploração dos dados em nível de usuário (aluno, professor, tutor, gestor, entre outros). O segundo nível (meso) é realizado por meio da instanciamento do EDRM para cada uma das instituições de ensino considerando apenas análises individuais por instituição. Por fim, o terceiro nível (macro) refere-se à instanciamento de apenas um EDRM para diferentes instituições de ensino, ou seja, apenas um EDRM para todas as instituições que buscam análises comparativas de práticas, serviços e produtos.

Outra decisão importante a ser tomada quanto à instanciamento do EDRM é o nível de granularidade a ser adotado. A granularidade, conforme definido no Capítulo 4, refere-se ao nível de sumarização dos elementos considerando seus detalhes disponíveis; quanto mais detalhado um elemento menor será sua granularidade. Sua definição é um dos principais fatores a ser considerado na modelagem de dados dimensional. Basicamente, a vantagem de se ter um nível de granularidade baixo é o maior detalhamento nas análises. Por outro lado, tem-se como ponto negativo o maior volume de dados, o que pode reduzir a performance computacional. No caso da granularidade com nível maior, o volume de dados é menor e, conseqüentemente, a performance em consultas é superior. Em contrapartida, com menos detalhamento nos dados, pode não ser possível responder a todas as perguntas necessárias. Nesse sentido, a definição clara e coerente da granularidade é fundamental para o desenvolvimento de um modelo.

Segundo Inmon (2005), o projeto de um DW capaz de considerar qualquer entidade arquitetônica deve ser definido em seu menor nível de granularidade. Em caso de muitos níveis, Inmon (2005) recomenda que dois (ou mais) níveis granulares de detalhamento, na maioria dos casos, são eficientes e atendem boa parte das consultas.

Dado que o EDRM foi desenvolvido como modelo de referência, seu nível de granularidade considerado foi o menor nível. Desse modo, cabe ao instanciador do EDRM

definir a eficiência de seu modelo selecionando dois ou mais níveis de interesse considerando, ainda, a quantidade de espaço disponível para armazenar mais ou menos dados (respectivamente menor ou maior granularidade). Para o EDRM, foram definidos níveis de granularidade para as dimensões curso, tempo e usuário. Na dimensão curso, análises podem ser realizadas em níveis crescentes de quantidade de dados, conforme ilustrado na Figura 13: departamento, curso, disciplina e turma.

Figura 13 – Exemplo de granularidade da perspectiva curso.

Fonte: Elaborada pelo autor.

Para a dimensão tempo foram considerados os níveis: ano, semestre, trimestre, mês, semana e dia. Por fim, o usuário pode estar localizado em instituições que atuam em diversas localizações. Desse modo o usuário pode estar situado em um país, estado, cidade, instituição. Esse tipo de granularidade é bastante adequada em análises de nível **macro**, ou seja, em comparações interinstitucionais ou comparações de práticas em diferentes escalas geográficas.

É importante observar que níveis de granularidade podem ser omitidos, ou seja, caso a instituição possua apenas parte da hierarquia como departamento, disciplina e turma, essa hierarquia também será válida.

Na próxima seção são definidos os principais *stakeholders* que podem estar presentes na dimensão Usuário, possibilitando, assim, importantes análises a partir de diferentes *stakeholders*.

5.4 Definição dos stakeholders

A análise educacional pode ser orientada para diversos *stakeholders*. Entende-se por *stakeholders* as partes interessadas no negócio, ou seja, **quem são os alvos das análises**. No contexto educacional, os *stakeholders* são as partes interessadas no processo educacional como alunos, professores, tutores, instituições educacionais, administradores, tomadores de decisões, pesquisadores, entre outros. Para o processo de tomada de decisão baseada em dados, cada *stakeholder* pode possuir tipos de análises distintas compatível com o seu perfil e comportamento. Isso significa que, por exemplo, pode-se realizar análises centradas no aluno, em busca do sucesso acadêmico, ou centrada no professor, para fins de gestão estratégica institucional. Cada uma dessas análises é centrada em diferentes *stakeholders* com diferentes perfis e comportamentos.

Dutt, Ismail e Herawan (2017) realizaram uma revisão sistemática em Mineração de Dados Educacionais e identificaram três principais tipos de *stakeholders*: **aluno**, **professor** e **sistema educacional**. Nessa pesquisa, os autores afirmam que as interações entre esses três atores geram grandes volumes de dados que podem possuir um valor de informação inestimável para a instituição. Já Atif *et al.* (2013) identificaram outros três *stakeholders*, **instituição departamento** e **alunos**.

Para Clow (2012), os principais *stakeholders* no contexto educacional são: (1) **aluno**: qualquer pessoa envolvida na aprendizagem; (2) **professor**: qualquer pessoa envolvida diretamente em facilitar a aprendizagem – inclui assistentes de ensino, professores associados, professores adjuntos, professores, pessoal acadêmico e colegas; (3) **gestor**: qualquer pessoa responsável pela organização ou administração de professores – inclui a administração de nível departamental e institucional (por exemplo, gerentes, administradores, chefes de departamento, diretores executivos, presidentes, reitores, vice-reitores e reitores); e (4) **formulador de políticas**: qualquer pessoa responsável pela definição da política, seja em nível local, regional, estadual, nacional ou transnacional / intergovernamental, incluindo os financiadores.

Diante da diversidade de *stakeholders* envolvidos no processo educacional, para a composição do EDRM foi realizada a análise matricial que define quais *stakeholders* podem estar diretamente associados às perspectivas definidas no modelo, conforme ilustrado na Figura 14. Nessa matriz, as linhas possuem as perspectivas educacionais, ou seja, atividade primárias, enquanto as colunas possuem os principais *stakeholders*. Essa definição foi baseada nas pesquisas que definiram os atributos base do EDRM. Conforme observado por Clow (2012), *stakeholders* distintos podem exercer o mesmo papel no contexto educacional,

ou seja, professores, tutores e monitores podem realizar as mesmas atividades, acessar e produzir os mesmo tipos de dados.

Figura 14 – Matriz de *stakeholders* por perspectiva do EDRM.

		STAKEHOLDERS						
PERSPECTIVA		Aluno	Professor	Tutor	Monitor	Coordenador	Gestor	Administrador
Sistemas Educacionais	Institucional		X	X	X	X	X	X
	Comunicação	X	X	X	X	X	X	X
	Estatísticas de acesso	X	X	X	X	X	X	X
	Notas	X					X	X
	Recursos Educacionais	X	X	X	X	X	X	X
	Pesquisa	X	X			X	X	X
	Financeiro	X	X	X	X	X	X	X
	Comercial	X					X	X
	Biblioteca	X	X	X	X	X	X	X
	Recursos Humanos		X	X	X	X	X	X
Outros Sistemas (financeiro, CRM, etc)	Outros	...	...	...	...	...	...	...

Fonte: Elaborada pelo autor.

É importante observar que a definição de *stakeholders* detalhada na matriz representa apenas uma sugestão de estrutura que poderá ser adotada na dimensão **usuários** para a definição dos *stakeholders*. Contudo, em geral, ambientes educacionais possuem padronizações desses perfis definidos por instituição. Essa padronização é válida no contexto do EDRM e foi explicitada no estudo de caso relatado no Capítulo 6.

Após a definição do modelo, para que seja viável sua utilização é necessário que decisões iniciais sejam tomadas antes de sua instanciação. As principais decisões são relatadas a seguir.

5.5 Decisões iniciais para a instanciação do modelo

A implementação de iniciativas em análise educacional exige que sejam tomadas várias decisões iniciais à sua realização. Essas decisões são determinantes para que o modelo possa ser utilizado de forma eficaz, ou seja, com foco em resultados para atingir objetivos e metas institucionais previamente definidas. Desse modo, antes de começar um projeto de análise educacional para a instanciação do EDRM é necessário:

1. Entender o cenário atual da instituição;

- Identificar os problemas institucionais que podem ser resolvidos por meio de análises;
2. Entender o cenário atual dos dados;
 - Identificar quais são os tipos de sistemas que serão analisados: SIS, LMS, financeiro, biblioteca, entre outros.
 - Avaliar políticas da instituição e definir políticas de segurança da informação associadas;
 - Avaliar quais são os principais *stakeholders* desses sistemas: aluno, docente, administrativo organizacional ou institucional.
 - Avaliar se as fontes de dados a serem exploradas possuem dados suficientes para o tipo de análise desejada.
 3. Avaliar que tipo de análises serão realizadas: *ad hoc*, baseada em indicadores de desempenho e/ou análises preditivas;
 4. Entender quais análises, relatórios e gráficos devem compor o *dashboard* para possibilitar análise que sejam abrangentes e relevantes ao processo decisório;
 5. Avaliar a aderência dos dados institucionais ao EDRM;
 - Identificar se os dados dos sistemas educacionais a serem analisados possuem atributos aderentes à estrutura do EDRM;
 - Definir os *Data Marts* do EDRM que serão instanciados para o conjunto de dados identificados. Se necessário, realizar incrementos à instância definida.
 6. Definir uma boa arquitetura tecnológica considerando a viabilidade financeira da instituição e a necessidade de um profissional capacitado para a definição da infraestrutura considerando a instanciação do modelo, bem como sua atualização.

5.6 Considerações finais

Neste capítulo foi detalhada a estrutura dimensional definida para a composição do EDRM. Inicialmente foram estabelecidos os atributos, medidas e métricas derivados a partir de pesquisas que abordam apoio a decisão orientada a dados utilizando Indicadores de Desempenho, Mineração de Dados Educacionais, modelos de dados dimensionais, *Data Warehouse* e processos de *Business Intelligence*. Esse conjunto de atributos foram categorizados e passaram a representar atributos de tabelas fato e dimensões, ou seja,

atributos quantitativos e qualitativos. A partir dessas definições foi possível obter um modelo dimensional no formato *star schema* capaz de abranger diversas perspectivas educacionais de forma correlacionada considerando diferentes *stakeholders* e níveis de granularidade.

Vale destacar que a definição de um modelo de dados analítico não é uma atividade trivial, principalmente considerando a heterogeneidade dos dados provenientes de sistemas educacionais e a possibilidade de utilização de diferentes fontes de dados (SIS e/ou LMS). Contudo, uma vez que os dados referentes ao domínio foram definidos e estruturados, essa complexidade passa a ser considerada apenas no processo de ETL sendo diretamente proporcional ao número de fontes de dados a serem analisadas, ou seja, quanto maior a quantidade de fontes de dados maior a complexidade do processo.

Com relação à eficiência do modelo, O EDRM será excepcionalmente eficiente quando os dados possuírem características como: grande volume de dados de sistemas educacionais heterogêneo e/ou de alta dimensionalidade (muitos atributos ou colunas) e a instituição possuir mapeamento de indicadores de desempenho que possam nortear as análises necessárias.

Ressalta-se que o modelo aqui descrito é uma proposição de um modelo de referência definido por meio da pesquisa de dados provenientes de sistemas educacionais para o desenvolvimento de *Data Warehouse* no formato *star schema*. Nesse sentido, o modelo não representa a única solução, mas sim uma solução abrangente que pode ser utilizada por qualquer instituição educacional.

No próximo capítulo é relatado um estudo de caso que utiliza a estrutura do EDRM para realizar análises a partir de indicadores de desempenho ou de forma *ad hoc* a partir de dados de três instituições de ensino.

ESTUDO DE CASO: INSTANCIÇÃO DO MODELO DE REFERÊNCIA DE DADOS EDUCACIONAIS

6.1 Considerações iniciais

Segundo Travasso, Gurov e Amaral (2002), o processo experimental vem se tornando o centro do processo científico o qual é capaz de explorar fatores críticos e possibilitar a geração de novos fenômenos para que teorias possam ser formuladas e corrigidas. Uma das formas de experimentação é por meio de estudos de caso. Estudos de caso são experimentos conduzidos com o propósito de investigar uma entidade ou um fenômeno dentro de um espaço de tempo específico. O estudo de caso é considerado um tipo de análise qualitativa desenvolvida para gerar *insights* exploratórios.

Sob essa perspectiva, neste capítulo é apresentado um estudo de caso cujo objetivo é analisar a viabilidade do Modelo de Referência de Dados Educacionais (*Educational Data Reference Model* – EDRM) a partir de dados gerados em contextos educacionais reais. Desse modo, deseja-se avaliar a eficácia do EDRM por meio da geração de análises considerando contextos educacionais variáveis como: diferentes perspectivas correlacionadas, níveis de granularidade e *stakeholders*. O estudo de caso realizado aborda dados referentes às principais funcionalidades de sistemas educacionais SIS e LMS utilizados por instituições de ensino superior, públicas e privadas, que atuam a partir das modalidades presencial, híbrida e a distância.

Na Seção 6.2 são detalhadas as características do estudo de caso considerando as particularidades de cada uma das instituições analisadas e as ferramentas utilizadas para a condução do mesmo. Na Seção 6.3 são detalhadas as três formas de processamentos que realizadas neste estudo considerando as etapas do processo de *Business Intelligence*. Por fim, na Seção 6.2.4 são discutidos aspectos de segurança da informação e confidencialidade dos dados, aplicados às instituições participantes deste estudo de caso.

6.2 Planejamento e definições do estudo de caso

O estudo de caso foi realizado a partir de dados cedidos por três instituições de ensino superior. Para manter a confidencialidade das instituições analisadas, neste trabalho elas são denominadas I1, I2 e I3. Dentre essas instituições, I1 e I2 são públicas e I3 é privada.

Além disso, os dados disponibilizados pelas instituições são referentes a diferentes modalidades sendo, I1 segundo as modalidades presencial, híbrida e a distância e I2 e I3 segundo a modalidade híbrida. O presente estudo de caso consiste na criação de instâncias do EDRM a partir de dados reais das instituições de ensino superior I1, I2 e I3, com o objetivo de:

- Realizar análise de dados provenientes de fontes de dados/provedores distintas;
- Realizar análises correlacionadas de perspectivas distintas;
- Realizar análises considerando diferentes *stakeholders*;
- Realizar análises em nível detalhado ou agregado considerando diversos níveis de granularidade dos dados;
- Realizar inferência por meio de tarefas de mineração de dados como parte do processo decisório;
- Utilizar as informações de forma fácil e acessíveis;
- Utilizar informações consistentes;
- Utilizar uma base de dados segura e consistente para a tomada de decisão baseada em evidências.

6.2.1 Características dos dados

No escopo deste trabalho, foram utilizados dados provenientes do LMS *Moodle* (provenientes das instituições I1, I2 e I3) em três versões distintas e um sistema acadêmico (SIS) desenvolvido pela própria instituição de ensino (I1).

A estrutura de dados do *Moodle* é constituída por aproximadamente 203 tabelas que armazenam dados de gestão de aprendizagem considerando dados provenientes de ferramentas de comunicação e colaboração. As diferentes versões analisadas do *Moodle* significam pequenas diferenças de atributos em cada modelo de dados relacional. Desse modo, poucos foram os ajustes necessários ao processamento dos dados a partir dessas diferentes versões. Quanto ao SIS-I1 (sistema educacional do tipo SIS desenvolvido e utilizado pela instituição I1), sua estrutura de dados possui aproximadamente 106 tabelas que representam todos os dados necessários à gestão institucional de I1.

As três instituições participantes deste estudo de caso possuem dados gerados por seus sistemas educacionais LMS e/ou SIS, conforme descrito no Quadro 10. Um fator determinante para a caracterização das perspectivas a serem analisadas, bem como o volume dos dados utilizado é a modalidade de ensino praticada por cada uma das instituições. Neste estudo de caso, três modalidades estiveram presentes: presencial, híbrida e a distância.

Em síntese, a modalidade presencial refere-se ao ensino tradicional em sala de aula. Em geral, seus dados são relativos a cursos presenciais cujos dados são armazenados em sistemas acadêmicos, sendo constituídos, predominantemente, por dados administrativos relativos à gestão acadêmica institucional.

A modalidade híbrida, por sua vez, tem por objetivo complementar a modalidade presencial por meio da utilização de recursos computacionais que possibilitem, principalmente, a comunicação, a colaboração e o compartilhamento de recursos educacionais durante o processo de ensino e aprendizagem. Os dados provenientes dessa modalidade podem possuir características de gestão e, ao mesmo tempo, de colaboração e comunicação entre *stakeholders*.

Por fim, a modalidade a distância utiliza diversas ferramentas para possibilitar um processo completo de ensino e aprendizagem apoiado por tecnologia da informação e comunicação. Em geral, essa modalidade utiliza um sistema de gestão de aprendizagem, opcionalmente associado a um sistema acadêmico, para fazer toda a parte de colaboração, comunicação e gestão do processo educacional.

Quadro 10 – Sistemas educacionais analisados.

Instituição	LMS	SIS	Modalidade	Categoria administrativa
I1	Moodle versão 1.9.8	SIS-I1	Presencial, híbrida e a distância	Pública
I2	Moodle versão 2.6	–	Híbrida	Pública
I3	Moodle versão 2.7	–	Híbrida	Privada

A abrangência deste estudo de caso é evidenciada pela quantidade e diversidade de dados de usuários de diferentes modalidades de ensino adotadas por cada uma das instituições. Os números que refletem essa afirmação podem ser observados na Figura 15.

Figura 15 – Quantidade de usuários por modalidades das instituições de ensino analisadas.

Fonte: Elaborada pelo autor.

Outra característica importante da amostra de dados deste estudo de caso refere-se ao seu volume físico e intervalo de tempo associados ao conjunto de dados. No Quadro 11, para cada uma das instituições de ensino são exibidas a quantidade de tuplas, a quantidade de dados, em unidade de medida de informação *Megabyte* (MB), e o intervalo de tempo, medido em anos, em que os dados foram gerados.

Quadro 11 – Volume físico e tempo de utilização do SIS e LMS de cada uma das instituições de ensino superior.

Sistema educacional	Quantidade de tuplas	Volume físico (MB)	Intervalo de tempo (ano)
I1 (LMS e SIS)	67.581.512	73.487,52	1996 - 2016
I2 (LMS)	106.420.111	34.214,66	2012 - 2015
I3 (LMS)	335.128	166,38	2015 - 2016

6.2.2 Análises planejadas

De modo geral, as instituições analisadas não possuem gestão e monitoramento baseada em dados de forma automatizada para o acompanhamento do processo educacional

como um todo. Sua gestão ocorre a partir de outras ferramentas e relatórios de seus sistemas. O *Moodle*, por exemplo, possui relatórios de acessos (*log*), notas e participação em atividades. Entretanto, grande parte dos sistemas educacionais, inclusive o *Moodle*, não permitem análises dos dados de forma *ad hoc*, além de não possuir relatórios baseados em indicadores de desempenho ou métricas mais complexas considerando conjuntos de agrupadores, como a *quantidade de alunos evadidos por idade, por gênero, por ano*, sendo *gênero, idade, ano* os agrupadores dessa análise.

Com relação ao cenário atual dos dados das instituições, conforme descrito anteriormente, todas as instituições analisadas possuem seus dados armazenados em SGBD relacional. Quanto ao volume de dados armazenados nos modelos relacionais, para as três instituições os dados são suficientes para o processamento analítico, porém, somente a instituição I1 possui dados históricos suficientes para gerar resultados relevantes em tarefas de mineração de dados.

Por fim, parte das análises realizadas foram baseadas em medidas e métricas pré-definidas e todas elas compõem os *dashboards* desenvolvidos. Para a instituição I1 foram realizadas as seguintes análises: (1) quantidade de cursos por departamento; (2) carga horária de disciplina por ano (por docente, por departamento); (3) quantidade de cursos e disciplinas (por ano, por modalidade); (4) quantidade de usuários por tipo; (5) quantidade de alunos por departamento; e (6) quantidade de alunos que utilizam ferramentas de comunicação. Já as análises de mineração de dados foram realizadas a partir para a descoberta de padrões com o objetivo de: (1) realizar previsões acerca do sucesso do aluno em disciplinas; (2) realizar previsões de evasão do aluno no curso; (3) realizar análise de grupos de perfis de alunos; e (4) realizar análise de grupos de disciplinas.

Considerando as instituições I2 e I3 foram realizadas as seguintes análises: (1) listar os dez primeiros cursos em que os usuários mais utilizaram ferramentas de comunicação (*chat* e fórum de discussão); (2) listar os dez primeiros cursos os quais os alunos obtiveram a maior média de notas; (3) quantidade de tipos de usuários por ano; (4) quantidade de matrículas realizadas por ano e por curso; e (5) média de ausências dos alunos por curso, por disciplina, por semestre.

6.2.3 Ferramentas utilizadas

Segundo as especificações do EDRM, definidas no Capítulo 5, os dados referentes ao domínio educacional são padronizados considerando diferentes perspectivas que correlacionam dados de funcionalidades comuns de sistemas educacionais com foco na experiência de aprendizagem do aluno e atuação dos demais *stakeholders*. Desse modo, a

partir do EDRM usuários interagem com os dados integrados, limpos e estruturados, por meio de ferramentas de análise e consulta de dados em busca de informações acerca do processo educacional. Tais ferramentas devem possibilitar a utilização desses dados, de forma simples e intuitiva, por meio de interfaces de navegação e visualização. Em especial, as ferramentas devem permitir que informações relevantes acerca do processo educacional sejam representadas ao longo do tempo, permitindo que análises possam ser realizadas para a detecção de tendências, análises comparativas, previsões, monitorações, entre outras.

Para iniciar o processo de ETL foi necessário avaliar os dados informacionais de origem dos sistemas educacionais. As fontes de dados analisadas estavam disponíveis no formato de dois SGBDs *Open Source* distintos: *MySQL*¹ e *PostgreSQL*². Em seguida, considerando o processo de ETL realizado resultando em uma instância de EDRM por instituição, outras ferramentas foram utilizadas para a análise dos dados, como: *Pentaho Community*³ (ferramenta *open source*) e *Oracle OBIEE*⁴ (ferramenta empresarial proprietária). Essas ferramentas foram escolhidas para possibilitar uma avaliação do EDRM considerando sua instanciação e utilização a partir de ferramentas distintas.

Nesse contexto, a instância do EDRM para cada instituição de ensino foi desenvolvida separadamente utilizando a ferramenta de ETL *Pentaho Data Integrator* (PDI) para a construção dos modelos. Após a definição de cada instância do modelo de referência, foram consideradas para a exploração dos dados ferramentas gratuitas, disponibilizadas como *software* livre, bem como ferramentas proprietárias amplamente utilizadas em contextos corporativos. As análises realizadas, foram definidas nesse estudo de caso conforme definido no Quadro 12, com a utilização de ferramentas proprietárias, para transformar dados provenientes da instância do EDRM em informação potencialmente útil, como se deseja validar.

Quadro 12 – Ferramentas utilizadas no processo de *Business Intelligence*.

Tipo de <i>software</i>	Processamento analítico (OLAP)	Análises e <i>dashboards</i>
Open Source	<i>Pentaho Schema Workbench</i>	<i>Pentaho BI Server</i>
Proprietário	<i>Oracle Administration Tool</i> (versão de teste)	OBIEE – <i>Oracle Business Intelligence Enterprise Edition</i> (versão de teste)

A partir dessa diferenciação, é possível constatar a utilização do EDRM independente de ferramenta de *Business Intelligence*. Dessa forma, para realizar aplicações reais

¹ Website *MySQL*: <https://www.mysql.com/>

² Website *PostgreSQL*: <https://www.postgresql.org/>

³ Website *Pentaho Community*: <http://community.pentaho.com/>

⁴ Website *Oracle BI*: <https://oracle.com/solutions/business-analytics/business-intelligence/>

do EDRM proposto no escopo deste estudo de caso, a ferramenta necessária à realização de atividades de exploração dos dados pode ser escolhida pela instituição considerando sua disponibilidade financeira para aquisição de *software* e a necessidade de profissionais com o conhecimento necessário à exploração dos dados a partir dessas ferramentas.

Por fim, para realizar experimentos de descoberta de padrões a partir de Mineração de Dados, foram utilizadas bibliotecas de desenvolvimento em *Python* na criação de modelos preditivos e descritivos. Em particular, foram utilizadas as bibliotecas *Scikit-learn*⁵, para a geração de modelos por meio de algoritmos de aprendizado de máquina, e *Matplotlib*⁶, para a geração de gráficos e visualização dos resultados.

6.2.4 Segurança da informação

Como todos os outros ativos que uma instituição mantém atualmente, os ativos de informação são extremamente valiosos. Nesse sentido, para garantir a confidencialidade dos dados das instituições de ensino foi firmado um acordo por escrito de compartilhamento dos dados das instituições participantes deste estudo de caso.

Esse acordo define principalmente: (1) os dados não podem ser publicados ou redistribuídos a pessoas não autorizadas neste acordo; (2) os dados pessoais devem ser anonimizados; (3) devem ser utilizadas todas as medidas necessárias para a proteção e segurança dos dados; e (4) os dados somente podem ser utilizados para a finalidade de pesquisa científica considerando o escopo dessa tese de doutorado.

Sob essa perspectiva, todos os dados pessoais de usuários dos sistemas foram anonimizados, ou seja, substituídos por dados aleatórios com o objetivo de não permitir a identificação de usuários e seus dados pessoais.

Adicionalmente, os dados disponibilizados foram utilizados localmente em um único computador e, após anonimizados, processados por meio do *cluster* Euler, disponibilizado pelo ICMC-USP para pesquisa realizada com o uso dos recursos computacionais do Centro de Ciências Matemáticas Aplicadas à Indústria (CeMEAI)⁷, financiado pela FAPESP.

A utilização do processamento por meio do *cluster* Euler foi importante devido o grande volume de dados, disponibilizados pelas instituições de ensino, que demandaram alto processamento computacional durante o processo de ETL, considerando a menor granularidade possível da hierarquia dos dados.

⁵ Website Scikit-learn: <http://scikit-learn.org/stable/>

⁶ Website Matplotlib: <https://matplotlib.org/>

⁷ Website CeMEAI: <http://www.cemeai.icmc.usp.br>

Vale destacar que é possível realizar esse processamento em computadores com configurações usuais (processador I5, 8Gb de memória, entre outros), o que significaria maior tempo de processamento dos dados a depender de seu volume.

6.3 Processamento e análise dos dados

O experimento conduzido foi realizado a partir de três processamentos de *Business Intelligence* (BI), sendo um para cada instituição analisada. Todos esses processos seguem a mesma sequência de passos inerentes ao processo de BI, conforme ilustrado na Figura 16.

Figura 16 – Visão geral do estudo de caso seguindo as etapas do processo de *Business Intelligence*.

Fonte: Elaborada pelo autor.

A primeira etapa do processo de BI aqui discutido refere-se à transformação de dados operacionais em dados informacionais. Nessa etapa, são utilizadas uma ou mais fontes de dados, no caso os sistemas educacionais com dados operacionais ou OLTP (SIS e/ou LMS) de cada instituição analisada. Nessa etapa, os dados operacionais são transformados

em dados informacionais por meio de um processo de Extração, Transformação e Carga (*Extract Transform Load* – ETL) em que esses dados são transformados na estrutura dimensional e armazenados no *Data Warehouse* segundo uma instância definida do EDRM.

A segunda etapa refere-se à realização de análises dos dados a partir do *Data Warehouse* definido. Nessa etapa, os dados são utilizados para a realização de análises que, em geral, têm como objetivo avaliar métricas e indicadores de desempenho previamente estabelecidos. Em alguns casos, entretanto, tais indicadores não são determinados, levando à realização de análises *ad hoc*. Essa etapa foi dividida em duas etapas com o objetivo de fornecer detalhes sobre as análises realizadas. Na etapa 2(a), as análises ocorrem a partir da criação de metadados necessários à tradução do modelo dimensional em cubos OLAP. Desse modo, os dados podem ser explorados por meio do processamento analítico realizado a partir dos cubos OLAP desenvolvidos. Já na etapa 2(b) os dados históricos armazenados no *Data Warehouse* são utilizados para a descoberta de padrões por meio de tarefas de mineração de dados.

Nas seções a seguir, as decisões tomadas em cada uma das etapas para o processamento e análise dos dados desse estudo de caso são descritas.

6.3.1 Etapa 1: instancição do EDRM e processo de ETL

O processo de ETL foi realizado transformando os dados dos diferentes sistemas educacionais de origem na estrutura estabelecida para o EDRM. O EDRM proposto neste trabalho é genérico e possui uma estrutura que visa atender às mais diferentes análises no contexto educacional. Dessa forma, de acordo com a disponibilidade de dados utilizados e armazenados pelas instituições, bem como dos objetivos das análises a serem realizadas, apenas um subconjunto da estrutura do EDRM foi necessário para atender à demanda de cada uma das instituições analisadas. Como exemplo de variações na instância do EDRM pode-se citar instituições que aplicam somente a modalidade presencial. Em geral, essas instituições provavelmente não possuem dados relativos à comunicação armazenados em seus SGBD. Nesse sentido, para a realização de análises de dados de instituições desse tipo, o EDRM pode ser instanciado sem efetuar a carga de dados no *Data Mart* de comunicação.

Para o estudo de caso aqui realizado, após a observação dos dados armazenados pelas instituições, gerou-se um subconjunto do EDRM para o armazenamento dos seus dados. O modelo *star schema* do EDRM instanciado, bem como seus *stakeholders* e níveis de granularidade adotados, estão ilustrados na Figura 17. Nesse modelo, estão ilustrados cinco *Data Marts* nota, comunicação, institucional_curso, institucional_usuario e estatística_acesso.

Figura 17 – Modelo *star schema* do EDRM para as instituições I1, I2 e I3 considerando seus *stakeholders* e níveis de granularidade.

Fonte: Elaborada pelo autor.

Uma vez definido o modelo a ser utilizado, passou-se à transformação dos dados e carga nas tabelas do DW. Para isso, um mapeamento entre as tabelas dos modelos de dados de origem e as tabelas de dimensões e fatos do EDRM foi estabelecido. Esse mapeamento foi realizado por meio de processos de ETL, utilizando a ferramenta PDI da *Pentaho*. Um exemplo do mapeamento do processo de ETL a partir de uma fonte de dados está ilustrado na Figura 18. De modo geral, esse processo executa um conjunto de subprocessos que realizam a transformação dos dados e sua carga em cada uma das dimensões e tabela fato que fazem parte do *data mart* nota. Ao final desse processo, o *data mart* nota estará totalmente criado com suas entidades, relacionamento e restrições no formato *star schema* definido considerando, ainda, todos os dados do sistema educacional de origem padronizados, integrados e consistentes.

Uma decisão a ser tomada no processo de ETL é sobre o tratamento da ausência de dados. Mesmo considerando um subconjunto do EDRM no modelo instanciado, ainda é possível que alguns dos atributos do modelo não estejam disponíveis em todas as fontes de dados. O tratamento desses casos deve ser padronizado, a fim de evitar incorreções nas análises realizadas.

Figura 18 – Processamento de ETL para cada fonte de dados.

Fonte: Elaborada pelo autor.

Nesse estudo de caso, por exemplo, dados relativos à data de nascimento de aluno não estão disponíveis no modelo relacional do LMS *Moodle*. Nesses casos, assim como em outros casos de dados ausentes ocorridos nesse estudo de caso, optou-se por preencher os valores ausentes com o valor *nc*, significando *não consta*.

Uma vez criado o modelo dimensional a partir da instanciação do EDRM e da respectiva carga dos dados, os dados encontram-se dispostos em um *Data Warehouse* segundo uma estrutura dimensional no formato *star schema*. A partir desse ponto, é possível realizar análises por meio de processamento analítico e mineração de dados.

6.3.2 Etapa 2: análises e dashboards

Para realizar análise e consulta em um DW pode-se utilizar processamento analítico (*On-line Analytical Processing* – OLAP) ou realizar descoberta de padrões a partir de modelos preditivos e descritivos de mineração de dados. Nas seções a seguir, análises baseadas em processamento analítico e mineração de dados, desenvolvidos neste estudo de caso, são apresentadas.

6.3.2.1 Etapa 2a: análises baseadas no processamento analítico

Em um ambiente de *Data Warehousing* as análises para apoiar o processo de tomada de decisão são realizadas a partir de requisições dimensionais aos dados de um *Data Warehouse* (DW). Para isso, ferramentas de processamento analítico (*On-line Analytical Processing* – OLAP) são adequadas por permitirem análises complexas e agregadas. Neste

De forma análoga no processamento analítico apresentado na Figura 20, são definidos os cubos com suas hierarquias, agregações e demais operações a serem realizadas. Ao final, a ferramenta produz um metadado que representa a tradução do mapeamento lógico realizado para o modelo físico. Por fim, esse mapeamento resulta em um metadado que, na próxima etapa do processo, será consumido pela ferramenta *Pentaho BI Server* para que possam ser realizadas as operações analíticas nos dados.

Figura 20 – Mapeamento analítico realizado a partir dos dados das instituições I2 e I3.

Fonte: Elaborada pelo autor.

A partir da estrutura analítica definida, os dados das instituições analisadas podem ser consumidos por ferramentas de relatórios de *dashboards*. Neste estudo de caso, foram desenvolvidos *dashboards* com o objetivo de mostrar graficamente informações que podem ser exploradas de forma correlacionada.

As análises foram baseadas nas métricas previamente definidas na Seção 5.5, que descreve as decisões iniciais a serem tomadas. Desse modo, dando sequência ao processo de *Business Intelligence*, foram desenvolvidos *dashboards* para as instituições analisadas. Para a instituição I1, foi definido um *dashboard* por meio da *suite* de ferramentas OBIEE, conforme ilustrado na Figura 21. Já para as instituições I2 e I3 foi utilizada a *Pentaho BI Server*.

Figura 21 – Dashboard desenvolvido para a instituição I1.

Fonte: Elaborada pelo autor.

Figura 22 – Dashboard desenvolvido para a instituição I2.

Fonte: Elaborada pelo autor.

Cabe destacar que foram realizadas análises definidas a partir de tópicos bastante discutidos em trabalhos que abordam a aplicação de *Learning Analytics* e *Academic Analytics* a partir de dados de sistemas educacionais. Adicionalmente, as análises realizadas podem ser replicadas a qualquer instituição de ensino que possua sistemas educacionais que produzem dados com atributos compatíveis aos definidos no EDRM.

Com relação à adequação do estudo de caso aos níveis de *Learning Analytics*, foram desenvolvidos modelos capazes de abranger os níveis micro (nível de usuário) e meso

(nível institucional). Contudo, considerando a definição do EDRM, e a pesquisa realizada neste estudo de caso, é possível constatar a viabilidade do modelo em nível macro de *Learning Analytics*. Para realizar essa instanciação em nível macro, basta utilizar uma única instância do EDRM para todas as instituições de ensino. Para isso é necessário adequar o processo de ETL de modo a garantir que os dados das instituições sempre pertençam a mesmo nível de granularidade. Essa instanciação do modelo em nível macro é possível devido à dimensão em conformidade *d_instituição* e à estrutura do modelo baseada em chaves substitutas ou *Surrogate Key*. A condução de um estudo de caso em nível macro, considerando os mesmos dados das instituições I1, I2 e I3, está prevista como um trabalho futuro a ser realizado.

6.3.2.2 Etapa 2b: análises a partir de tarefas de mineração de dados

Para complementar a análise dos dados, a partir dos dados da instituição I1, foram realizados testes no EDRM com o objetivo de extrair novos padrões de dados para realizar inferências utilizando algoritmos de mineração de dados e aprendizado de máquina. Em aplicações de mineração de dados educacionais, os dados utilizados são relativos aos diferentes *stakeholders* do processo educacional, como aqueles armazenados em um repositório integrado e consistente segundo especificações definidas para EDRM, desenvolvido no escopo deste trabalho.

Preparação do experimento

Neste estudo de caso, a fim de validar a completude do modelo proposto, foram realizadas quatro aplicações de mineração de dados utilizando algoritmos de aprendizado de máquina a partir dos dados armazenados no EDRM instanciado por I1, conforme descrito no Quadro 13.

No Quadro 13 são definidas as quatro aplicações realizadas. Para as aplicações A1 e A2 foram construídos modelos preditivos utilizando dados de alunos que já cursaram disciplinas nos cursos analisados para a predição do comportamento (classe) de novos alunos. Nessas aplicações foram utilizados algoritmos de aprendizado de máquina supervisionados considerados estado da arte para mineração de dados educacionais preditivo (SHAHIRI; HUSAIN; RASHID, 2015): Árvores de Decisão, Redes Neurais, Regressão Logística, Naïve Bayes, K-NN (K Vizinhos Mais Próximos) e SVM (Máquinas de Vetor e Suporte).

As aplicações A3 e A4 são baseadas nas notas dos alunos em disciplinas, com o objetivo de encontrar grupos concisos de alunos e disciplinas que permitam identificar padrões de comportamento recorrentes. Para essas aplicações foi utilizado o algoritmo *K-Means*, um dos mais tradicionais algoritmos de agrupamento de dados. A escolha pelo

algoritmo K-Means deve-se à simplicidade e interpretabilidade direta do modelo gerado (MACQUEEN, 1967).

Quadro 13 – Aplicações de mineração de dados educacionais utilizando o EDRM.

Aplicação	Análise	Tipo de Tarefas	Algoritmos
A1	Uma aplicação para predição de sucesso em disciplinas	Preditiva	Árvores de Decisão, Redes Neurais, Regressão Logística, Naïve Bayes, KNN e SVM
A2	Uma aplicação para predição de evasão		
A3	Uma aplicação para análise de grupos de perfis de alunos	Descritiva	<i>K-Means</i>
A4	Uma aplicação para a análise de grupos de disciplinas		

Nesse experimento foram utilizados dados de alunos de quatro cursos da área de Computação da instituição I1: Ciência da Computação, Engenharia de Computação, Sistemas de Informação e Tecnologia em Análise e Desenvolvimento de Sistemas. Esses cursos foram escolhidos devido ao projeto de pesquisa definido para a execução desses experimentos ter sido conduzido com o auxílio de professores e gestores de cursos dessas áreas. Dessa forma, a validação dos resultados foi intuitiva por representar o contexto real do dia-a-dia desses professores/pesquisadores, ou seja, cursos e disciplinas que esses professores atuam ou já atuaram seja como docente e/ou coordenador.

A seleção dos dados no EDRM instanciado foi realizada por meio da redução dos dados e composição de um modelo de exploração. Desse modo, os dados de exploração possuem informações relativas a alunos matriculados no período entre 2005 e 2013. Esse período foi considerado para a obtenção de dados com maior completude de valores nos atributos relevantes às análises realizadas. Em geral, dados de instituições podem estar incompletos em alguns períodos devido a evoluções que ocorrem ao longo do tempo na instituição. Essas mudanças institucionais podem resultar em mudanças nos dados a partir de incrementos realizados no sistema educacional ou até mesmo implementações de novas funcionalidades. Contudo, é importante observar que essa análise pode ser realizada com a instância total do EDRM. Isso implica apenas em maior tempo de processamento devido ao maior volume de dados analisados.

Um sumário de atributos utilizados nessas tarefas pode ser encontrado no Quadro 14. Foram utilizados dados provenientes do modelo dimensional, mais especificamente das dimensões *d_curso*, *d_usuario*, *d_matrícula* e *d_nota*, bem como da tabela de consolidação, fato *f_nota*.

Os atributos foram utilizados diretamente a partir do EDRM ou por meio da agregação de um de seus atributos para a construção dos modelos preditivos ou descritivos.

Por exemplo, para as aplicações A1, A2, A3 e A4 o atributo `idt_disciplina` foi obtido diretamente da tabela `d_curso` do EDRM. Por outro lado, para essas mesmas aplicações o atributo calculado `num_avg_nota` que significa a média de notas finais do aluno, foi calculado a partir do atributo `num_nota_final` disponível na tabela `f_nota`.

Quadro 14 – Sumário dos atributos da instância do EDRM utilizado nas aplicações preditivas e descritivas.

Tabela	Atributo de origem	Atributos calculados	Aplicações			
			A1	A2	A3	A4
d_curso	idt_disciplina		x	x	x	x
	des_disciplina		x	x	x	x
d_matricula	des_tipo_ingresso			x		
f_nota	num_nota_final	num_avg_nota (média da nota final)	x	x	x	x
d_nota	fl_status_nota	num_percentual_aprovacao_curso (percentual de aprovação no curso)		x	x	
	fl_status_nota	num_percentual_aprovacao_disciplina (percentual de aprovação na disciplina)				x
f_institucional_usuario	num_qtd_ausencias	num_avg_ausencias (média de ausências)		x		
	dt_matricula	num_ano_matricula (ano)		x		
	dt_matricula	num_dia_matricula (dia)		x		
	dt_matricula	num_semestre_matricula (semestre)		x		

Os atributos utilizados nas aplicações relatadas foram organizados em tabelas atributo-valor de forma a torná-las manipuláveis por algoritmos de aprendizado de máquina. Nesse Quadro, as linhas correspondem aos exemplos e as colunas aos atributos que os descrevem. Cada célula interna da tabela atributo-valor é preenchida com uma medida que associe o exemplo ao atributo. A última coluna representa o atributo de interesse (atributo classe), o qual se deseja prever em aplicações preditivas. Em aplicações descritivas, não há atributo classe.

Na aplicação A1, a matriz atributo-valor foi gerada **para cada disciplina**, ou seja, cada matriz gerada possui **todas as disciplinas de todos os cursos exceto a que será analisada**. Essa matriz, exemplificada no Quadro 15, possui nas linhas apenas alunos que já cursaram a disciplina que esta sendo analisada. Cada célula interna foi preenchida com a nota final do aluno na disciplina correspondente. Além disso, para cada aluno, somente foram preenchidas as notas das disciplinas que ele havia cursado previamente à disciplina analisada. Caso o aluno não tenha cursado uma determinada disciplina, aquela posição foi preenchida com o valor -1. Por fim, a coluna classe foi preenchida com os valores “Aprovado” ou “Reprovado”, de acordo com o aproveitamento do aluno na disciplina analisada.

Por sua vez, na aplicação A3, uma matriz atributo-valor foi gerada **para cada curso**, ou seja, cada matriz possui **todas as disciplinas de cada curso**. Nessa matriz, exemplificada no Quadro 16, as linhas representam os alunos daquele curso, enquanto as

Quadro 15 – Exemplo de matriz atributo-valor de A1.

	<i>disciplina₁</i>	<i>disciplina₂</i>	...	<i>disciplina_M</i>	<i>Classe_disciplina_analisada</i>
<i>aluno₁</i>	<i>nota₁₁</i>	<i>nota₁₂</i>	...	<i>nota_{1M}</i>	<i>Aprovado</i>
<i>aluno₂</i>	<i>nota₂₁</i>	<i>nota₂₂</i>	...	<i>nota_{2M}</i>	<i>Reprovado</i>
⋮	⋮	⋮	⋱	⋮	⋮
<i>aluno_N</i>	<i>nota_{N1}</i>	<i>nota_{N2}</i>	...	<i>nota_{NM}</i>	<i>Aprovado</i>

colunas representavam as disciplinas cursadas por alunos daquele curso. Caso o aluno não tenha cursado uma determinada disciplina, aquela posição foi preenchida com o valor -1.

Quadro 16 – Exemplo de matriz atributo-valor de A3.

	<i>disciplina₁</i>	<i>disciplina₂</i>	...	<i>disciplina_M</i>
<i>aluno₁</i>	<i>nota₁₁</i>	<i>nota₁₂</i>	...	<i>nota_{1M}</i>
<i>aluno₂</i>	<i>nota₂₁</i>	<i>nota₂₂</i>	...	<i>nota_{2M}</i>
⋮	⋮	⋮	⋱	⋮
<i>aluno_N</i>	<i>nota_{N1}</i>	<i>nota_{N2}</i>	...	<i>nota_{NM}</i>

Na aplicação A4, exemplificada no Quadro 17, as matrizes atributo-valor por curso são matrizes transpostas das matrizes utilizadas na aplicação A3, com as linhas correspondendo às disciplinas e as colunas aos alunos.

Quadro 17 – Exemplo de matriz atributo-valor de A4.

	<i>aluno₁</i>	<i>aluno₂</i>	...	<i>aluno_M</i>
<i>disciplina₁</i>	<i>nota₁₁</i>	<i>nota₁₂</i>	...	<i>nota_{1M}</i>
<i>disciplina₂</i>	<i>nota₂₁</i>	<i>nota₂₂</i>	...	<i>nota_{2M}</i>
⋮	⋮	⋮	⋱	⋮
<i>disciplina_N</i>	<i>nota_{N1}</i>	<i>nota_{N2}</i>	...	<i>nota_{NM}</i>

Por fim, na Aplicação A2, exemplificada no Quadro 18, a matriz atributo-valor foi composta por **todos os alunos nas linhas e nas colunas algumas métricas extraídas do aproveitamento do aluno nas disciplinas por ele cursadas durante o curso**: média de notas obtidas; média de faltas; carga horária total; e percentual de aproveitamento.

Os atributos aqui utilizados foram normalizados no intervalo $[0,1]$ a fim de evitar discrepâncias de escala entre os valores dos diferentes atributos. Por exemplo, o atributo correspondente à média das notas tem uma escala muito menor do que a carga horária total – enquanto o primeiro tem valor máximo 10, o segundo tem valores em torno de centenas. A coluna classe foi preenchida com os valores “Evadiu” ou “Não Evadiu”, de acordo com o status do aluno no curso.

Para as aplicações A1 e A2, os modelos de classificação foram avaliados por meio de validação cruzada em k pastas (*k-fold cross validation*) (STONE, 1974). O objetivo

Quadro 18 – Exemplo de matriz atributo-valor de A2.

	<i>media_notas</i>	<i>media_faltas</i>	<i>carga_horaria</i>	<i>per_aproveitamento</i>	<i>Classe_aluno</i>
<i>aluno</i> ₁	<i>mn</i> ₁	<i>mf</i> ₁	<i>ch</i> ₁	<i>pa</i> ₁	Evadiu
<i>aluno</i> ₂	<i>mn</i> ₂	<i>mf</i> ₂	<i>ch</i> ₂	<i>pa</i> ₂	Não Evadiu
⋮	⋮	⋮	⋮	⋮	⋮
<i>aluno</i> _{<i>N</i>}	<i>mn</i> _{<i>N</i>}	<i>mf</i> _{<i>N</i>}	<i>ch</i> _{<i>N</i>}	<i>pa</i> _{<i>N</i>}	Não evadiu

desse procedimento é medir, com maior precisão estatística, o comportamento do modelo construído com os dados que se possui quando defrontado com novos dados.

Nessa avaliação, o conjunto de dados é dividido em k subconjuntos e são realizados k procedimentos de treinamento e teste dos classificadores. Em cada procedimento, um subconjunto (*fold*) é selecionado como conjunto de teste, enquanto os outros $k - 1$ subconjuntos são selecionados para o conjunto de treinamento. A cada repetição, o valor da medida de avaliação do classificador é armazenado. Ao final, computa-se a média dessas medidas de avaliação obtidas ao longo das k repetições.

Nos experimentos aqui conduzidos, utilizou-se k igual a 3. A escolha desse valor de k deu-se para a obtenção de um número razoável de exemplos dentro de cada uma das k pastas para algumas disciplinas em que o número de alunos cursantes era reduzido. Como medida de avaliação, adotou-se a acurácia preditiva dos classificadores comparados.

Para os algoritmos parametrizados, os valores desses parâmetros foram ajustados em cada um dos conjuntos de dados utilizados, de maneira a otimizar os resultados obtidos. Os valores testados para os principais parâmetros desses algoritmos estão especificados no Quadro 19. Parâmetros não especificados indicam que foi adotado o valor padrão do algoritmo na biblioteca utilizada.

Quadro 19 – Parâmetros testados nos algoritmos de classificação.

Técnica	Algoritmo	Parâmetros	Significado	Valores Testados
SVM	SVC (Support Vector Classifier)	Kernel	Função utilizada para separação das classes.	Radial Basis Function (RBF) e Linear
		C	Penalidade aplicada por erro de classificação.	Aleatoriamente de uma função exponencial com escala 100.
		gamma	Influência de uma instância na classificação.	Aleatoriamente de uma função exponencial com escala 0,1.
Redes Neurais	MLP (Multi Layer Perceptron)	solver	Algoritmo utilizado para atualização dos pesos da rede a cada iteração.	SGD (stochastic gradient descent)
		learning_rate	Taxa de aprendizado da rede neural.	Constante

Técnica	Algoritmo	Parâmetros	Significado	Valores Testados
		momentum	Taxa de atualização dos pesos no algoritmo SGD.	Aleatoriamente de uma função exponencial com escala 0,1.
		alpha	Termo de regularização da rede neural.	Aleatoriamente de uma função exponencial com escala 0,001.
		activation	Função de ativação dos neurônios da rede neural.	Logística
		learning_rate_init	Valor inicial da taxa de aprendizado.	Aleatoriamente de uma função exponencial com escala 0,01.
		hidden_layer_sizes	Tamanho das camadas ocultas da rede neural.	(200,50,20), (100,20), (100,50), (200,20), (200,50), (50,10), (200,), (100,), (50,)
		max_iter	Máximo de iterações para convergência do modelo.	500
Árvores de Decisão	C4.5	criterion	Critério para divisão dos nós na árvore.	Medida Gini
		max_depth	Profundidade máxima da árvore.	[3,10]
Vizinhos Mais Próximos	KNN	metric	Medida utilizada para cálculo da distância entre as instâncias.	Distância Euclideana.
		n_neighbors	Número de vizinhos mais próximos na atribuição da classe.	[1,15]
Agrupamento de dados	K-Means	n_clusters	Número de clusters a ser encontrado	[2, 30]
		metric	Medida utilizada para cálculo da distância entre as instâncias.	Distância Euclideana.

Nas seções a seguir, os resultados de cada uma das aplicações são analisados. Inicialmente, são apresentados os resultados que envolvem aplicações preditivas A1 e A2. Em seguida, são reportados os resultados de aplicações descritivas, A3 e A4.

Aplicação A1: predição do sucesso em disciplinas

O resultado da aplicação A1, cujo objetivo é a construção de modelos preditivos de aprovação ou reprovação em disciplinas, podem ser observados na Tabela 1. Nessa tabela, estão expressas as porcentagens de acurácia preditiva média obtida por cada um dos classificadores ao longo do processo de validação cruzada, seguido pelo valor do desvio

padrão obtido. Nas duas últimas colunas, são apresentados o percentual de alunos na classe majoritária (tipicamente a classe **aprovado**) e o total de alunos que compuseram a base de dados para aquela disciplina, ou seja, o número de alunos que cursaram aquela disciplina.

Com o resultado dessa primeira aplicação foi possível observar que, para a grande maioria das disciplinas, os modelos preditivos construídos obtiveram bom desempenho na predição de aprovação dos alunos cursantes. Além disso, todos os algoritmos construíram modelos que superaram a classe majoritária nos conjuntos de dados.

Em particular, o classificador SVM foi o que obteve melhor desempenho ao longo das disciplinas, seguido pelas redes neurais MLP. Observa-se ainda que, mesmo em disciplinas com alto índice de reprovação, como *Fundamentos de Teoria da Computação*, *Algoritmos e Programação I* e *Algoritmos e Programação II*, os algoritmos de classificação obtiveram alto índice de acurácia, indicando a eficácia dos mesmos na caracterização das classes.

Apesar de apresentarem melhor desempenho, os modelos aprendidos pelas SVMs e MLPs não são diretamente interpretáveis por humanos, sendo classificados como “modelos caixa preta” (MITCHELL *et al.*, 2000). Nesse tipo, o modelo aprendido é armazenado por meio de matrizes de números ou fórmulas matemáticas que, a olho humano, nada representam. Para aplicações em que os administradores precisem apenas de um indicador apontando quando um aluno está em risco de reprovação ou não, algoritmos como SVM e MLP são adequados.

Porém, para problemas em que é necessário entender como a decisão foi tomada pelo algoritmo, explicitando o conjunto de fatores que levam o aluno à aprovação ou reprovação em uma determinada disciplina, algoritmos como as Árvore de Decisão são mais indicados. Uma árvore de decisão é uma estrutura *top-down* que pode ser interpretada como um conjunto de regras do tipo “SE ... ENTÃO ...”. A partir dessas regras, gestores podem identificar quais outras disciplinas influenciam na aprovação ou reprovação dos alunos.

Exemplos de árvores de decisão obtidas para as disciplinas de *Interação Humano-Computador*, tipicamente ofertada como disciplina obrigatória no 8º período do curso de Sistemas de Informação e como disciplina optativa aos demais cursos, e *Introdução à Engenharia de Computação*, ofertada apenas no 1º semestre do curso de Engenharia de Computação, podem ser observados nas Figuras 23 e 24. Na primeira, os alunos são classificados como reprovados por meio das regras:

- “SE nota em *Computação e Sociedade* é maior que 1,4 E nota em *Administração de Sistemas I* é maior que 0,4 E nota em *Qualidade de Software* é menor ou igual a 3,2 ENTÃO Classe = Reprovado”.

- “SE nota em *Computação e Sociedade* é menor ou igual a 1,4 E nota em *Gerência de Projetos* é menor ou igual a 5,0 E nota em *Inteligência Artificial* é menor ou igual a 4,9 ENTÃO Classe = Reprovado”;

Tabela 1 – Resultados da predição de aprovação nas disciplinas da aplicação A1.

[illegible]

Tabela 1 – Resultados da predição de aprovação nas disciplinas da aplicação A1.

Disciplina	Árv. Dec.	SVM	KNN	Reg. Log.	N. Bayes	R. Neurais	Maj.	Acad.
Est. de Dados e Prog.	80,5 ± 5,4	80,8 ± 3,6	77,8 ± 1,8	77,0 ± 6,5	73,6 ± 4,7	76,1 ± 4,2	70,7	406
Ferramenta Cad	82,0 ± 6,5	87,2 ± 3,9	82,7 ± 3,6	86,5 ± 3,5	84,9 ± 5,4	85,0 ± 5,9	63,2	133
Fund. de Redes de Comp.	88,4 ± 5,6	91,0 ± 1,8	85,2 ± 3,2	87,1 ± 3,9	81,3 ± 2,3	90,3 ± 2,7	52,3	155
Fund. de Tec. da Inf.	82,9 ± 2,5	87,8 ± 1,2	84,5 ± 1,0	84,8 ± 3,0	82,8 ± 2,8	86,6 ± 2,7	52,0	691
Fund. de Teo. da Comp.	87,6 ± 1,4	88,5 ± 1,2	88,2 ± 2,1	87,3 ± 4,0	84,9 ± 4,7	88,5 ± 3,8	57,4	1144
Gerência de Projetos	69,2 ± 3,7	82,9 ± 7,4	72,5 ± 5,8	65,9 ± 2,9	72,3 ± 13,3	80,4 ± 9,3	73,9	88
Gov. de Tec. da Inf.	89,8 ± 2,4	94,8 ± 4,0	88,5 ± 3,1	90,3 ± 2,8	88,4 ± 1,7	90,4 ± 3,2	73,7	156
Gov. de Tec. da Inf. II	71,6 ± 1,5	69,5 ± 6,7	79,1 ± 6,0	63,8 ± 7,0	79,1 ± 6,0	75,3 ± 7,6	52,8	53
Impl. e Exp. Algorítmica	84,2 ± 8,1	90,3 ± 0,4	90,3 ± 0,4	74,2 ± 12,3	90,3 ± 0,4	90,3 ± 0,4	90,3	31
Inf. Apl. à Administração	89,1 ± 4,4	90,9 ± 5,1	89,1 ± 3,8	86,4 ± 2,2	86,4 ± 0,2	91,9 ± 4,4	86,4	110
Inteligência Artificial	82,7 ± 1,0	83,7 ± 2,9	82,8 ± 5,1	80,5 ± 1,3	83,4 ± 2,1	83,7 ± 2,6	68,7	307
Int. Humano-computador	82,5 ± 1,4	91,8 ± 4,4	86,0 ± 2,9	87,7 ± 2,9	86,5 ± 2,2	90,1 ± 5,8	82,5	171
Int. à Bioinformática	79,9 ± 7,5	94,9 ± 3,6	97,4 ± 3,6	92,5 ± 0,3	90,1 ± 3,1	90,1 ± 3,1	90,0	40
Int. à Criptografia Comp.	72,2 ± 7,9	86,1 ± 3,9	86,1 ± 3,9	97,2 ± 3,9	75,0 ± 11,8	91,7 ± 6,8	66,7	36
Int. a Eng. da Computação	88,7 ± 2,4	90,9 ± 4,6	90,5 ± 5,8	86,8 ± 3,5	74,2 ± 4,1	88,2 ± 2,4	72,4	221
Int. a Sistemas Digitais	92,1 ± 1,2	93,7 ± 0,8	92,1 ± 0,6	92,7 ± 0,5	85,5 ± 4,6	92,9 ± 0,4	54,7	1567
Laboratório de Hardware	83,0 ± 3,3	83,0 ± 3,3	82,4 ± 2,1	77,4 ± 2,9	80,5 ± 1,2	83,0 ± 1,4	76,1	159
Ling. de Montagem	81,0 ± 2,0	81,8 ± 5,8	79,7 ± 5,5	78,4 ± 6,1	76,6 ± 8,7	82,8 ± 2,8	63,8	389
Ling. de Programação O. O.	90,0 ± 0,7	93,3 ± 2,2	89,3 ± 2,9	91,5 ± 1,4	87,8 ± 7,3	91,3 ± 1,5	57,1	401
Ling. Formais e Autômatos	82,4 ± 1,1	82,0 ± 0,7	80,5 ± 4,9	80,8 ± 2,5	72,8 ± 3,8	83,9 ± 3,5	73,2	261
Métodos Numéricos	71,0 ± 4,9	76,5 ± 6,2	72,8 ± 3,8	69,1 ± 2,3	71,0 ± 1,8	71,0 ± 1,8	70,9	55
Mét. Quant. Apl. à Adm. I	86,9 ± 2,5	86,9 ± 2,5	86,9 ± 2,5	82,4 ± 1,3	77,1 ± 0,6	87,6 ± 3,4	77,1	153
Mét. Quant. Apl. à Adm. II	83,2 ± 1,7	84,5 ± 0,1	85,2 ± 2,3	82,0 ± 2,3	77,4 ± 0,7	81,9 ± 2,4	77,4	155
Microcontroladores e Apl.	93,0 ± 5,8	95,4 ± 3,3	91,1 ± 6,3	90,8 ± 6,5	81,9 ± 6,1	93,2 ± 5,4	59,1	44

Tabela 1 – Resultados da predição de aprovação nas disciplinas da aplicação A1.

Disciplina	Árv. Dec.	SVM	KNN	Reg. Log.	N. Bayes	R. Neurais	Maj.	Acad.
Operações de Sist. Inf.	82,8 ± 1,0	82,8 ± 1,0	82,8 ± 1,0	82,8 ± 1,0	82,8 ± 1,0	82,8 ± 1,0	82,8	116
Prob. e Estatística	68,7 ± 4,2	73,5 ± 4,7	74,5 ± 4,1	68,7 ± 2,1	57,2 ± 3,0	73,5 ± 3,9	54,3	208
Prog. de Computadores I	70,5 ± 0,9	70,5 ± 0,9	70,5 ± 0,9	70,5 ± 0,9	69,9 ± 0,9	68,7 ± 1,9	67,1	943
Prog. Linear	57,6 ± 2,3	72,6 ± 2,0	75,4 ± 5,8	66,2 ± 13,0	57,3 ± 5,6	63,4 ± 13,5	57,6	33
Prog. Para a Web	90,7 ± 2,6	94,2 ± 1,4	87,6 ± 2,0	92,6 ± 2,3	90,4 ± 2,1	94,5 ± 2,3	77,3	313
Qualidade de Software	86,3 ± 3,4	88,1 ± 1,7	88,1 ± 1,7	82,1 ± 2,5	87,5 ± 2,9	89,9 ± 0,8	89,3	168
Redes de Computadores	79,4 ± 0,4	79,9 ± 1,7	78,2 ± 3,7	81,4 ± 3,7	78,2 ± 3,2	82,4 ± 3,9	53,6	403
Redes Sem Fio	68,1 ± 4,5	81,7 ± 4,2	71,2 ± 10,8	65,1 ± 2,6	74,2 ± 4,3	75,8 ± 1,4	75,8	66
Segurança de Redes	64,4 ± 8,6	79,4 ± 4,0	71,9 ± 2,4	66,5 ± 7,8	69,3 ± 5,7	71,9 ± 2,4	71,8	39
Seg. e Aud. de Sistemas	79,3 ± 2,4	87,1 ± 3,2	80,7 ± 4,1	78,0 ± 7,3	81,3 ± 8,7	76,8 ± 1,8	76,8	155
Sist. de Apoio à Decisão	52,9 ± 21,1	68,6 ± 9,3	66,1 ± 11,4	59,7 ± 6,6	48,3 ± 11,0	64,3 ± 3,6	57,8	45
Sistemas Distribuídos	89,3 ± 3,5	91,4 ± 1,8	92,1 ± 6,7	88,5 ± 2,8	86,4 ± 2,8	88,5 ± 4,2	87,1	140
Sistemas Operacionais	84,3 ± 0,4	91,7 ± 1,2	85,6 ± 2,7	87,6 ± 2,5	83,5 ± 4,2	86,6 ± 3,5	56,1	606
Sociedade e Ambiente	88,4 ± 3,1	90,1 ± 2,0	92,7 ± 6,0	87,7 ± 5,2	84,3 ± 2,5	91,0 ± 2,9	81,0	121
Teo. dos Grafos e seus Alg.	79,9 ± 8,4	81,9 ± 4,2	85,9 ± 2,7	79,9 ± 3,1	79,9 ± 4,2	84,6 ± 4,7	65,8	94
Tóp. em Bancos de Dados	75,6 ± 7,1	77,4 ± 7,7	73,2 ± 7,4	66,9 ± 2,6	74,7 ± 11,4	78,8 ± 5,2	72,3	92
Tóp. em Sist. de Informação I	75,1 ± 5,3	80,4 ± 5,0	75,9 ± 13,6	71,6 ± 6,0	68,5 ± 7,4	71,6 ± 8,2	57,6	79
Tóp. em Sist. de Informação II	80,0 ± 7,3	83,7 ± 3,1	83,6 ± 7,6	78,6 ± 8,5	76,0 ± 1,5	87,5 ± 3,2	63,3	40
Tóp. em Sist. de Informação III	77,5 ± 12,6	87,4 ± 3,9	79,9 ± 4,1	82,4 ± 3,9	70,0 ± 6,4	84,8 ± 6,5	60,0	149

Figura 23 – Árvore de decisão para a disciplina *Interação Humano-Computador*.

Fonte: Elaborada pelo autor.

Em uma análise da árvore de decisão apresentada no contexto dos cursos da instituição analisada, é possível perceber que outras disciplinas da área de Engenharia de Software, com conteúdo programático diretamente relacionado com a disciplina *Interação Humano-Computador* estão envolvidas, como *Gerência de Projetos* e *Qualidade de Software*. Com isso, alunos reprovados em *Interação Humano-Computador* e com notas baixas nessas outras disciplinas podem ter dificuldades com os conteúdos dessas disciplinas.

Outras disciplinas elencadas na árvore de decisão, *Inteligência Artificial* e *Computação e Sociedade*, são ofertadas no semestre imediatamente anterior ao da oferta dessa disciplina. Isso levanta a hipótese de que o comportamento do aluno pode ter correlação com o aproveitamento ou não nessa disciplina. Cabe destacar que muitos dos alunos nessa fase final estão realizando estágio obrigatório em empresas, o que pode diminuir o tempo de dedicação às disciplinas e, conseqüentemente, seu aproveitamento.

Já para a disciplina de *Introdução à Engenharia de Computação*, as regras extraídas para a reprovação são:

- “SE nota em *Fundamentos de Teoria da Computação* é menor ou igual a 1,7 E nota em *Linguagem de Montagem* é menor ou igual a 0,75 ENTÃO Classe = Reprovado”
- “SE nota em *Fundamentos de Teoria da Computação* é maior que 1,7 E nota em *Introdução a Sistemas Digitais* é menor ou igual a 0,5 (nota 0 ou não cursou a disciplina, caso em que tem nota -1 na matriz) E nota em *Sociedade e Ambiente* é maior que -0,5 (nota -1 indica que não cursou a disciplina) ENTÃO Classe = Reprovado”

Figura 24 – Árvore de decisão para a disciplina Introdução à Engenharia de Computação.

Fonte: Elaborada pelo autor.

No caso de *Introdução à Engenharia de Computação*, por ser uma disciplina oferecida nos períodos iniciais do curso, todas as disciplinas envolvidas na árvore de decisão são, também, do primeiro ou segundo semestre. Nesse caso, o modelo nos permite perceber que o aproveitamento dessa disciplina está diretamente relacionado ao aproveitamento do aluno no curso como um todo. Segundo professores da instituição I1, de modo geral, em períodos iniciais desses cursos, o aluno possui dificuldades devido à transição do ensino médio para o ensino superior. Nesse sentido, alunos com baixo aproveitamento em disciplinas iniciais são mais propensos a evadir dos cursos.

Aplicação A2: predição de evasão

Os resultados da aplicação A2 podem ser observados na Tabela 2. De maneira análoga aos resultados da aplicação A1, os modelos mostraram-se eficientes na predição de evasão, conseguindo obter resultados acima da classe majoritária (não evasão).

Tabela 2 – Resultados da predição de evasão da aplicação A2.

Algoritmo	Acurácia média
Árvore de Decisão	78,66 ± 7,15
SVM	76,48 ± 9,64
KNN	77,41 ± 7,35
Reg. Log.	76,24 ± 7,56
Naïve Bayes	78,41 ± 6,34
MLP	75,82 ± 8,64
Classe Majoritária	65,06
# Alunos	1205

Na aplicação A2, o desempenho de todos os algoritmos de classificação foi bastante similar. O melhor desempenho, em média, foi obtido pelo algoritmo de árvore de decisão. O modelo construído por esse algoritmo pode ser observado na Figura 25.

Figura 25 – Árvore de decisão para a predição de evasão.

Fonte: Elaborada pelo autor.

Nas regras da árvore de decisão da aplicação A2, uma vez que os atributos são normalizados, a interpretação de seus valores não é diretamente relacionada ao contexto do problema. Dessa forma, sabe-se, apenas, que valores próximos a 0 estão próximos dos menores valores para aquele atributo, enquanto valores próximos a 1 são valores mais próximos dos maiores valores. Assim, a interpretação das regras obtidas pela árvore para a caracterização da evasão é a seguinte:

- “SE a carga horária total cursada com aproveitamento pelo aluno (*totalTime*) é muito pequena (menor ou igual a 0,016 em valores normalizados) ENTÃO Classe = Evadiu”
- “SE a carga horária total cursada com aproveitamento pelo aluno (*totalTime*) é pequena (menor ou igual a 0,1011 mas maior que 0,016 em valores normalizados) E a porcentagem de aprovação (*approvalPercent*) é maior que 35% ENTÃO Classe = “Evadiu”

Em ambas as regras, é possível perceber que um fator comum para a evasão dos alunos é estarem no começo do curso. Esse é um fenômeno recorrente em cursos das áreas de Ciências Exatas, em particular na Computação. Geralmente, os semestres iniciais representam um desafio maior aos alunos, quer seja por dificuldades na transição do ensino médio para o ensino superior, quer seja porque as disciplinas desses semestres trazem maiores novidades ao aluno, com uma curva de aprendizado maior. Uma vez não conseguindo sucesso nos semestres iniciais, os alunos tendem a abandonar os cursos.

Aplicação A3: análise de grupos de perfis de alunos

Nessa aplicação, o objetivo foi encontrar grupos de alunos em cada curso analisado de acordo com as notas obtidas pelos alunos nas disciplinas. Para cada aluno de cada grupo, algumas medidas foram analisadas, como **nota média**, **quantidade de disciplinas cursadas** e **quantidade de disciplinas aprovadas**.

Nessa seção são apresentados e discutidos apenas os resultados obtidos para um dos cursos, Sistemas de Informação, contando com 570 alunos. Nessa aplicação, assim como na aplicação A4, optou-se por separar a análise por cursos para possibilitar análises concisas com foco no contexto de cada curso evitando, assim, ruídos desnecessários aos quais os algoritmos de agrupamento são bastante sensíveis. A escolha por apresentar, aqui, os resultados do curso de Sistemas de Informação deve-se ao fato de esse curso possuir a maior quantidade de alunos dentre todos os cursos analisados. Os resultados para os demais cursos são bastante similares, possibilitando análise análoga à do curso aqui reportado.

Inicialmente, para uma análise de grupos obtidos pelo algoritmo K-Means, é necessário determinar, a priori, a quantidade de *clusters* a serem encontrados. Para isso, o parâmetro k do algoritmo foi variado no intervalo $[2, 30]$. Para cada variação, foi medido o valor de silhueta (ROUSSEEUW, 1987) médio, o qual reflete quão bem alocado cada exemplo está dentro de seu *cluster*, baseando-se na distância dos exemplos ao centro do seu *cluster* e ao centro dos demais *clusters*. O valor de silhueta varia no intervalo $[-1, 1]$ e, quanto mais próximo de 1, melhor é o agrupamento. Para essa aplicação, o melhor valor médio de silhueta foi para $k = 6$, com valor 0,4388. O valor de silhueta para cada exemplo pode ser observado na Figura 26.

Para possibilitar a visualização do agrupamento obtido, foi aplicada a técnica de *Principal Component Analysis* (PCA) (PEARSON, 1901), transformando os dados para duas dimensões. O resultado pode ser observado na Figura 27. É possível observar alguns grupos de alunos bem definidos, como os *clusters* 0 e 2, enquanto outro *clusters* são um pouco mais dispersos.

A fim de entender o comportamento dos alunos em cada grupo, primeiramente mediu-se, para alunos com matrícula realizada em pelo menos cinco disciplinas ao longo do curso, as médias de notas e de aprovação. O motivo de limitar a matrícula em pelo menos cinco disciplinas é considerar apenas dados mais representativos e menos propensos a ruídos. Os resultados podem ser observados na Figura 28. Nessa Figura, tem-se, na parte superior, a média de notas obtidas por cada aluno, de cada *cluster*, nas disciplinas por ele cursada. Na parte inferior, tem-se a quantidade de disciplinas cursadas pelos alunos nas

Figura 26 – Análise de silhueta para agrupamento de alunos.

Fonte: Elaborada pelo autor.

Figura 27 – Resultado do agrupamento de alunos considerando K=6.

Fonte: Elaborada pelo autor.

colunas transparentes e a quantidade de disciplinas que o aluno foi aprovado nas colunas opacas.

De maneira geral, conforme ilustrado nas Figuras 26 e 28, é possível perceber que o *cluster* mais característico é o *cluster* 0, que possui alunos com média mais baixa (em

Figura 28 – Média de notas dos alunos *vs* quantidade de disciplinas matriculadas e cursadas com aproveitamento.

Fonte: Elaborada pelo autor.

geral, abaixo da média de notas mínima de 6,0), com baixo aproveitamento de disciplinas e poucas disciplinas diferentes em que se matriculou. Esse é um cenário clássico de alunos mais propensos à evasão, uma vez que os alunos ficam retidos em semestres e disciplinas iniciais, desmotivando-se a continuar nos cursos.

Por outro lado, tem-se nos *clusters* 1, 4 e 5 alunos que possuem, em geral, um bom desempenho. Nesses casos os alunos alcançaram média superior à mínima para sua aprovação. Alunos que se encaixam nesses grupos têm maiores chances de sucesso acadêmico e de conclusão do curso.

Com esses resultados, gestores podem promover ações direcionadas aos alunos em busca de seu sucesso acadêmico. Por exemplo, os gestores podem realizar ações de acompanhamento e priorização dos alunos alocados no *cluster* 0 a fim de aumentar sua retenção nos cursos. Aos alunos dos *clusters* 1, 4 e 5, os gestores podem promover ações de incentivo à pesquisa e ao desenvolvimento pessoal. Esses alunos podem ser estimulados à participação em tarefas de pesquisa e extensão, como forma de preparação para o mercado de trabalho e iniciação de projetos de pós-graduação. Adicionalmente, com a utilização de outros *data marts* presentes no EDRM proposto neste trabalho, pode-se realizar análise de perspectivas associadas como fatores sócio-econômicos e formação educacional prévia do

aluno. É importante observar que não foi possível realizar esta análise como parte deste estudo de caso uma vez que os dados necessários não estavam disponíveis no conjunto de dados disponibilizado pelas instituições.

Aplicação A4: análises de grupos de disciplinas

Por fim, na aplicação A4 o objetivo foi encontrar grupos de disciplinas de cada curso de acordo com as notas obtidas por alunos nas mesmas. Para cada disciplina em cada grupo encontrado, analisou-se, também, a nota média da disciplina, a quantidade de alunos que a cursaram e a quantidade de alunos aprovados.

Assim como na aplicação A3, foi realizada uma análise de silhueta para determinação do valor de k a ser utilizado pelo algoritmo K-Means. O melhor valor encontrado foi de $k = 5$, com silhueta igual a 0,4074. O valor e silhueta de cada exemplo nessa aplicação pode ser observado na Figura 29.

Figura 29 – Análise de silhueta para agrupamento de disciplinas.

Fonte: Elaborada pelo autor.

O agrupamento correspondente, utilizando aplicação de PCA para redução a duas dimensões, pode ser observado na Figura 30. O comportamento das disciplinas também permite identificar alguns grupos bastante concisos, em particular os grupos 1, 3 e 4.

Figura 30 – Resultado do agrupamento de disciplinas considerando $K=5$.

Fonte: Elaborada pelo autor.

Para entender o comportamento apresentado, mediu-se as disciplinas com mais de 30 alunos cursantes que possuem as médias de nota e aprovação dos alunos cursantes. Os resultados estão ilustrados na Figura 31.

Figura 31 – Média de notas nas disciplinas *vs* quantidade de alunos matriculados e aprovados.

Fonte: Elaborada pelo autor.

Desse modo, é possível observar que as disciplinas alocadas nos *clusters* 1 e 4 possuem as maiores quantidades de alunos matriculados, mas baixas taxas de aprovação e baixas médias. Nesses grupos encontram-se disciplinas como Introdução a Sistemas Digitais, Fundamentos de Teoria da Computação, Algoritmos e Programação I e Fundamentos de Tecnologia da Informação. Além disso, todas essas disciplinas possuem média total inferior a 4,5 e taxa de aprovação abaixo de 40%. Ressalta-se que tais disciplinas estão alocadas no primeiro semestre do curso, o que explica as altas taxas de matrícula. Ainda, os resultados explicitam que no primeiro semestre existem altas taxas de reprovação, indicando grande dificuldade dos alunos nessa etapa.

Os *clusters* 0 e 2 representam disciplinas presentes na trilha comum dos cursos, considerando-se a participação regular e as notas médias, enquanto o *cluster* 3 caracteriza disciplinas dos períodos finais dos cursos, com baixa participação porém notas médias superiores, devido à maturidade adquirida pelos alunos. As disciplinas com média igual a 10.0 e 100% de aprovação são *Atividades Complementares*, *Estágio Obrigatório*, *Projeto Final I* e *Projeto Final II* que, conforme mencionado anteriormente, estão presentes no roteiro final dos cursos e possuem um comportamento binário de aprovação ou reprovação apenas.

Em conjunto com os resultados obtidos na aplicação A3, os resultados da aplicação A4 podem ajudar gestores a observar períodos dos cursos em que os alunos necessitam de maior acompanhamento. Em particular, fica claro que os semestres iniciais do curso são aqueles que apresentam maior dificuldade aos alunos. Para esses, projetos de monitoria e acompanhamento extra podem ser indicados, proporcionando maior auxílio aos alunos. Acompanhamento e auxílio em disciplinas intermediárias, como aquelas alocadas nos *clusters* 0 e 2, também pode ser essencial para estimular os alunos a buscarem uma boa formação e sucesso em suas carreiras.

6.3.3 Análise do estudo de caso

Durante a condução do estudo de caso, as fontes de dados originais adotadas pelos sistemas educacionais das diferentes instituições analisadas possuíam características e estruturas diferentes. Por meio do processo de ETL resultando no mapeamento dos dados para a estrutura do EDRM, foi possível a padronização dos dados heterogêneos em uma fonte única de dados integrados e consistentes. Dessa forma, facilita-se a análise integrada dos dados, bem como sua exploração considerando diferentes hierarquias/granularidades e *stakeholders*.

A heterogeneidade dos dados das instituições analisadas neste estudo de caso pode ser observada a partir das características das instituições, considerando o nível de utilização dos sistemas educacionais (maior quantidade de dados históricos), bem como a abrangência de suas funcionalidades (maior diversidade de atributos). Isso significa que, quanto mais funcionalidades forem exploradas por parte dos *stakeholders* e maior tempo de utilização dos sistemas educacionais, mais efetiva será a descoberta de conhecimento.

Neste estudo de caso, as análises foram realizadas guiadas por medidas e métricas pré-definidas. Um exemplo de análise baseada em medidas pré-definidas está na correlação de curso (`d_curso`), usuário (`d_usuario`), tempo (`d_tempo`) do *data mart* `f_institucional_usuario` para estabelecer a **média de ausências dos alunos por curso, por disciplina e por semestre**. As medidas e métricas utilizadas neste estudo de caso foram descritas na Seção 6.2.2. Entretanto, outras análises exploratórias *ad hoc*, sem um objetivo prévio, também são permitidas pelo EDRM, uma vez que a navegação pelos dados em diferentes perspectivas é facilitada pelo modelo aqui proposto. Essa característica permite a utilização do EDRM considerando diferentes níveis de maturidade analítica da instituição de ensino analisada.

Considerando a diversidade de modalidades de ensino (presencial, híbrida ou a distância) adotadas por cada instituição, também foi possível observar implicações relevantes acerca da utilização do modelo, principalmente no que tange aos tipos de dados armazenados. Por exemplo, a instituição I1, a qual atua nas modalidades a distância, híbrida e presencial, tinha necessidades específicas quando comparada às demais instituições analisadas, que atuam somente na modalidade híbrida. A modalidade a distância, quando comparada à modalidade híbrida, armazena mais dados, por exemplo, de comunicação entre os *stakeholders*. Além disso, cursos a distância, por exigirem a interação constante dos *stakeholders* por meio dos sistemas, tendem a gerar maior volume de dados relativos a acesso e utilização dos sistemas.

Neste estudo de caso, foram realizadas análises nos níveis micro (usuário) e meso (instituição) de *Learning Analytics*. Um exemplo de nível micro está na análise de **alunos que utilizam ferramentas de comunicação**. Já em nível meso, pode-se citar a **carga horária total da disciplina por ano (por docente, por departamento)**. Ressalta-se, entretanto, que o EDRM é compatível com análises em nível macro por meio de dimensão `d_instituição`. Essa dimensão permite o armazenamento dos dados de diferentes instituições em uma única instância do EDRM, de modo a possibilitar análises interinstitucionais.

Nas aplicações de mineração de dados reportadas, modelos preditivos e descritivos foram gerados a partir de dados armazenados no EDRM, utilizando dados relativos a **curso, matrícula, nota e usuário**. As análises descritivas de mineração de dados possibilitam aos gestores o melhor entendimento do comportamento dos *stakeholders* de sua instituição. Já

as análises preditivas fornecem um importante ferramental para acompanhamento desses *stakeholders*, possibilitando a antecipação de eventuais situações individuais.

Uma vez que o EDRM fornece uma fonte de dados integrada e com diferentes visões dos dados por meio de seus *data marts*, a construção de modelos de mineração de dados é simples quando comparada a construção de modelos a partir das fontes originais dos dados. Além disso, a replicação das análises relatadas em outros cenários é facilitada, uma vez que o EDRM, como um modelo de referência, representa uma fonte de dados padronizada com suporte a cursos presenciais, híbridos e a distância.

6.4 Considerações finais

Neste capítulo, foram realizadas aplicações práticas do EDRM a partir de dados reais de sistemas educacionais de três instituições de ensino superior (I1, I2 e I3). Inicialmente, foi realizada uma análise dos dados operacionais de cada sistema educacional e definida uma instância do EDRM adequada ao conjunto de dados das instituições analisadas. A partir dessa análise, foi estabelecido um subconjunto de *Data Marts* do EDRM adequados a cada instituição para, em seguida, realizar a instanciamento do modelo passando pelas etapas do processo de *Business Intelligence* (BI).

O processo de BI realizado utiliza a instância definida do EDRM com os dados de cada instituição analisada. Considerando o *Data Warehouse* gerado para cada instituição a partir do EDRM, os dados dessas instituições foram explorados por meio de processamento analítico e modelos de mineração de dados.

Buscando sintetizar a pesquisa conduzida, no próximo capítulo é apresentada uma visão geral da pesquisa realizada, as principais contribuições deste trabalho, as publicações realizadas e esperadas, bem como as linhas de pesquisas futuras.

CONCLUSÕES

Neste capítulo, é apresentada uma visão geral do trabalho realizado, descrevendo suas principais contribuições em função das linhas de pesquisa investigadas: *Learning Analytics*, *Academic Analytics*, *Business Intelligence*, Mineração de Dados Educacionais e *Data Warehousing*. São, também, apresentadas as principais dificuldades e limitações encontradas durante a realização deste trabalho, bem como as perspectivas futuras para a continuidade da pesquisa.

7.1 Visão geral da pesquisa realizada

Inúmeras pesquisas vêm sendo conduzidas com o objetivo de utilizar dados de sistemas educacionais para promover melhorias e avanços no contexto educacional. O estudo e análise dos dados presentes nesses sistemas relativos aos diferentes níveis de granularidade e *stakeholders* envolvidos no processo educacional visam o estabelecimento de mecanismos de apoio ao processo decisório.

Dentre as pesquisas que utilizam dados de sistemas educacionais, linhas de pesquisa que abordam temas em Análise Educacional, como *Learning Analytics*, *Academic Analytics* e Mineração de Dados Educacionais, estão sendo investigadas. Essas pesquisas buscam investigar formas de realizar análises a partir de dados potencialmente úteis. Assim, a partir desses dados é possível promover avanços nos resultados da aprendizagem por meio de apoio à decisão informada por dados. Um exemplo de decisão a partir dos dados é a identificação de padrões recorrentes no comportamento de alunos e professores em disciplinas; definições e monitoramento estratégias educacionais; e o acompanhamento do progresso dos alunos em busca do seu sucesso acadêmico dentre outras possibilidades.

Uma das principais dificuldades de aplicações de Análise Educacional consiste na análise e identificação dos dados necessários ao processo decisório no contexto educacional. Adicionalmente, diferentes ambientes educacionais utilizam diferentes modelos de dados, o que dificulta a identificação e manipulação dos dados de interesse.

Desse modo, neste trabalho foi proposto um modelo de referência de dados educacionais, denominado EDRM (*Educational Data Reference Model*), para a descoberta de conhecimento a partir de dados de sistemas educacionais. Tal modelo visa minimizar problemas relativos à definição e utilização de dados analíticos para apoio ao processo decisório. O EDRM foi desenvolvido a partir de análises de dados e funcionalidades de sistemas educacionais a fim de que dados provenientes de diferentes tipos e sistemas possam ser utilizados de forma unificada, integrada e consistente.

A arquitetura do EDRM é estruturada em um *Data Warehouse* (DW) seguindo um modelo dimensional de dados, projetado para ser uma fonte única de dados integrados e correlacionados voltada à tomada de decisão. A carga de dados para o EDRM deve ser realizada a partir do mapeamento e transformação dos dados dos sistemas educacionais.

É importante observar que, embora a construção da infraestrutura de um DW não seja algo trivial, essa construção faz-se necessária apenas uma vez. Uma vez que o DW é construído corretamente, ele fornece à instituição uma base flexível e reutilizável. Os dados granulares encontrados em um DW são a chave para a reutilização, pois podem ser usados por diversos *stakeholders* de maneiras diferentes.

Entre as vantagens obtidas com a utilização do EDRM está a possibilidade de análise dos dados educacionais a partir de diferentes visões. Além disso, outro benefício relevante é a flexibilidade em utilizar uma mesma fonte de dados estruturada de forma consistente para prover análises históricas de toda a instituição.

Destaca-se, ainda, a simplicidade do modelo para acomodar necessidades futuras. Mudança de regras, leis e necessidades do negócio são cenários de mudança de requisitos de informação inevitáveis. Em um DW a instituição pode rapidamente responder a essas mudanças com ajustes no processo de ETL e adequações em suas análises.

Adicionalmente, dados analíticos, disponíveis em ambiente de *Data Warehousing*, também podem ser utilizados como apoio à extração de padrões em mineração de dados. O grande volume de dados históricos que possuem seu detalhamento em diferentes níveis de granularidade são excelentes fontes de dados limpos, integrados e estruturados passíveis de extração de padrões.

A fim de avaliar a estrutura definida para o EDRM, também foi realizado um estudo de caso que explorou os dados de três instituições de ensino superior por meio da

instanciação de um subconjunto de *data marts* do EDRM realizada considerando as etapas do processo de *Business Intelligence*. Ao final do processo de *Business Intelligence*, foram realizadas análises a partir de processamento analítico e mineração de dados considerando diferentes atores e em diferentes níveis de granularidade. Ressalta-se que as aplicações de mineração de dados realizadas permitiram a descoberta de padrões interessantes acerca dos dados, possibilitando o maior entendimento dos dados contidos nesses sistemas, bem como a possibilidade de predizer, com um alto grau de precisão, o comportamento futuro de alunos. Na próxima seção são sumarizadas as principais contribuições e limitações de pesquisa.

7.2 Contribuições de pesquisa

A principal contribuição deste trabalho refere-se à proposição do EDRM (*Educational Data Reference Model*), um modelo de referência de dados educacional. O EDRM é um modelo de dados analítico de referência capaz de fornecer subsídios à realização do processo decisório baseado em evidências a partir de dados de sistemas educacionais. Considerando as investigações realizadas neste trabalho, tendo em vista o estado da arte acerca do tema abordado, não existe na literatura um modelo analítico de referência para o domínio educacional com a abrangência do EDRM em termos diversidade de atributos categorizados que definem diferentes perspectivas educacionais. Essas perspectivas podem ser analisadas individualmente ou de forma correlacionada considerando, ainda, dados que descrevem as ações de diferentes *stakeholders* e podem estar estruturados em níveis de granularidade.

Como contribuição associadas às principais linhas de pesquisa investigadas para a definição do EDRM destacam-se:

- ***Learning Analytics***: o EDRM foi definido para possibilitar que instituições realizem análises a partir de diferentes níveis de *Learning Analytics*: (1) micro – em nível de usuário, considerando o contexto que o mesmo está inserido; (2) meso – em nível institucional para a obtenção de melhorias administrativas e organizacionais; e (3) macro – em nível interinstitucional, possibilitando comparações entre práticas e serviços educacionais. O estudo de caso relatado considerou análises em nível micro e meso. Contudo, as análises realizadas também evidenciaram a viabilidade do modelo em nível macro. Ainda assim, um estudo de caso considerando o nível macro deve ser conduzido em curto prazo, como parte dos trabalhos futuros associados a esta pesquisa.

- **Academic Analytics:** a estrutura do EDRM permite que instituições com diferentes níveis de maturidade realizem a descoberta de conhecimento a partir de uma instância do mesmo. Dessa maneira, mesmo as instituições que não possuem uma cultura de tomada de decisão baseada em evidências são capazes de explorar as informações de sistemas educacionais disponibilizadas por meio de análises *ad hoc* a partir da estrutura do EDRM, com o objetivo de gerar *insights* exploratórios acerca do processo educacional. Em contrapartida, instituições que realizam sua gestão a partir de metas e objetivos estratégico bem definidos, podem obter de forma eficiente respostas analíticas do processo educacional por meio da estrutura do EDRM.
- **Mineração de Dados Educacionais:** o EDRM possui uma estrutura de dados orientada por assunto para armazenar dados históricos. A estrutura do EDRM possibilita o armazenamento de dados integrados de forma consistente para a realização de análises, comparações e redução dos dados considerando diferentes perspectivas educacionais. Assim, a estrutura do EDRM provê uma padronização de dados relevante à descoberta de padrões a partir de tarefas de mineração de dados.

Outra contribuição deste trabalho a ser destacada é a abrangência da estrutura definida para o EDRM. As perspectivas e atributos do EDRM foram identificadas a partir de uma revisão bibliográfica de pesquisas que exploram análises de dados de sistemas educacionais, indicadores de desempenho e diversos outros modelos de dados desenvolvidos para apoio à decisão. Dessa forma, foi possível obter um amplo conjunto de características, medidas e métricas capazes de abranger diversas perspectivas e níveis de granularidades existentes no contexto educacional. Esse amplo conjunto de atributos, distribuídos em esquemas dimensionais entrelaçados por dimensões comuns, permite maior diversidade de análises relevantes ao processo educacional. Ainda, o EDRM possui todas as vantagens inerentes à arquitetura de um DW, tais como:

- Arquitetura padrão e previsível;
- Estrutura dimensional capaz de garantir alta performance durante a análise dos dados;
- Possibilidade de separar a estrutura de sua representação, ou seja, os dados podem ser reorganizados de diversas formas sem que haja necessidade de modificações em sua estrutura;
- Possibilidade de visualização de diversas maneiras como, gráficos, relatórios e matrizes;

- Flexibilidade para realizar análises atemporais a partir de dados históricos, ou seja, análises do passado, presente e previsões futuras; e
- Acesso rápido a informação a diversas perspectivas correlacionadas com recursos de aumentar ou diminuir o nível de detalhamento.

Ressalta-se, também, que dado a simplicidade da estrutura do EDRM, é possível definir instâncias a partir de diferentes contextos institucionais, contendo apenas os dados necessários as análises adequadas à instituição utilizadora. Além disso, devido à simplicidade da estrutura dimensional definida para o EDRM, caso alguma instituição necessite realizar análises a partir de dados não contemplados, é possível realizar sua expansão ou modificação. Dessa forma, mesmo o EDRM tendo sido desenvolvido voltado a sistemas educacionais que possuem dados voltados ao ensino superior, é possível sua modificação para lidar com contextos de outros níveis de ensino.

Por fim, uma contribuição decorrente da proposição do EDRM e das aplicações desenvolvidas no estudo de caso realizado é o mapeamento do sistema educacional *Moodle* para o EDRM. Ao realizar validações do EDRM a partir do estudo de casos, foram efetuadas todas as etapas de um processo de *Business Intelligence* considerando dados de três versões distintas do sistema educacional *Moodle*. Dessa forma, foi desenvolvido um projeto que realiza a extração dos dados informacionais do sistema *Moodle*, sua limpeza e demais transformações, bem com a carga desses dados em uma instância do EDRM para cada instituição. A partir da instância do EDRM torna-se possível a utilização dos modelos desenvolvidos para a realização do processamento analítico e da descoberta de padrões a partir de tarefas de mineração de dados.

7.3 Dificuldades e limitações

Uma das dificuldades enfrentadas durante a condução deste trabalho foi a necessidade de investigação e estudo de diversas linhas de pesquisa necessárias tanto para o estudo teórico como prático. De fato, a investigação foi realizada de forma multidisciplinar considerando linhas de pesquisa complementares necessárias à realização os objetivos deste trabalho: engenharia de software com ênfase em informática na educação (*Learning Analytics* e *Academic Analytics*), inteligência artificial (Mineração de Dados Educacionais) e banco de dados (modelagem dimensional e *Data Warehouse*).

Com relação à limitação presente na condução deste trabalho, pode-se citar sua validação restrita ao contexto educacional nacional principalmente voltada ao ensino superior. Essa limitação deve-se ao fato das bases de dados analisadas serem estritamente

de instituições de ensino superior, que podem possuir atributos de interesse específicos quando comparados ao ensino médio ou profissionalizante. Assim, diferenças na estruturas da instituição considerando seus cursos, corpo docente, formas de avaliação (nota ou competência), entre outras, podem ser investigadas. Contudo, as investigações realizadas por meio de revisões bibliográficas da literatura apontam para a viabilidade deste modelo para os demais níveis de ensino. Um estudo acerca dessa limitação considerando o EDM como modelo base que a ser adaptado ou estendido para essa finalidade deve ser realizado como trabalho futuro descrito em maiores detalhes a seguir.

7.4 Produção científica

Como principais publicações resultantes das atividades realizadas durante este trabalho de doutorado, destacam-se:

- BORGES, V. A.; NOGUEIRA, B. M.; BARBOSA, E. F. A multidimensional data model for the analysis of learning management systems under different perspectives. In: **2016 IEEE Frontiers in Education Conference (FIE)**, p. 1-8, 2016.
- BORGES, V. A.; NOGUEIRA, B. M.; BARBOSA, E. F. et al. Uma análise exploratória de tópicos de pesquisa emergentes em informática na educação. **2015 Revista Brasileira de Informática na Educação (RBIE)**, SBC, v. 23, n. 2, p. 85-96, 2015.

Em decorrência dos resultados alcançados está sendo elaborado um artigo completo a ser submetido para o periódico *Computers & Education*¹.

- BORGES, V. A.; NOGUEIRA, B. M.; BARBOSA, E. F. Towards an educational data reference models. In: **2017 Computers & Education**, 2017.

7.5 Trabalhos futuros

Dando continuidade às atividades conduzidas durante este trabalho de doutorado, pode-se destacar como principais perspectivas futuras de pesquisa as seguintes linhas de atuação:

- Investigar da aplicabilidade do modelo proposto para os níveis de ensino fundamental, médio e profissionalizante.

¹ Website C&E: <https://www.journals.elsevier.com/computers-and-education/>

- Instanciar o EDRM para um estudo de caso considerando o nível macro de *Learning Analytics* para explorar a abrangência do modelo proposto quanto a realização de análises interinstitucionais.
- Investigar em profundidade de aspectos sócio econômicos em nível nacional. Uma forma realizar essa pesquisa é por meio da definição de incrementos ao modelo para a exploração da dados públicos abertos disponibilizados que podem ser obtidos, por exemplo, por meio do Portal Brasileiro de Dados Abertos².
- Realizar pesquisas que investigam as implicações da utilização do modelo sob o ponto de vista de alunos e professores. O EDRM é um modelo de dados que foi desenvolvido para abranger o maior número de análises relevantes ao processo decisório no contexto educacional. Contudo, para explorar melhor a utilização do EDRM, pode-se investigar a abrangência do modelo a partir de análises realizadas e validadas por alunos e professores.
- Investigar a linha de pesquisa PLE (*Personal Learning Environment*), para a utilização de dados coletados de diversas fontes de dados considerando canais formais (como sistemas educacionais) e canais informais (como mídias sociais). (CHATTI *et al.*, 2012).

² Website do Portal Brasileiro de Dados Abertos: dados.gov.br/

REFERÊNCIAS

AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: **Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: ACM, 1993. p. 207–216. ISBN 0-89791-592-5. Disponível em: <<http://doi.acm.org/10.1145/170035.170072>>. Citado na página 68.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: **VLDB'94: Proceedings of the XX International Conference on Very Large Data Bases**. San Francisco, CA, EUA: Morgan Kaufmann Publishers Inc., 1994. p. 487–499. ISBN 1-55860-153-8. Citado na página 64.

ALMEIDA, M. C. A.; SANTOS, L. L.; ALMEIDA, W. M.; FREITAS, W. A.; VEIGA, C. P.; DUCLOS, L. C. Integração de indicadores para avaliação do desempenho de uma instituição de ensino superior. **Revista ESPACIOS**, v. 37, 2016. Disponível em: <<http://www.revistaespacios.com/a16v37n07/163707e1.html>>. Citado na página 54.

ARROWAY, P.; MORGAN, G.; O'KEEFE, M.; YANOSKY, C. R. **Learning Analytics in Higher Education**. 2016. Citado nas páginas 24, 26, 43 e 44.

ATIF, A.; RICHARDS, D.; BILGIN, A.; MARRONE, M. Learning analytics in higher education: a summary of tools and approaches. p. 68–72, 2013. Citado na página 122.

BAKER, R.; CORBETT, A.; ROLL, I.; KOEDINGER, K. Developing a generalizable detector of when students game the system. **User Modeling and User-Adapted Interaction**, Springer Netherlands, v. 18, n. 3, p. 287–314, 2008. ISSN 0924-1868. Disponível em: <<http://dx.doi.org/10.1007/s11257-007-9045-6>>. Citado nas páginas 72 e 73.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, Sociedade Brasileira de Computação, Porto Alegre, RS, Brasil, v. 19, n. 2, p. 3–13, 2011. Citado na página 71.

BAKER, R.; YACEF, K. The state of educational data mining in 2009: A review and future visions. **Journal of Educational Data Mining**, v. 1, n. 1, p. 3–17, 2009. Citado na página 101.

BARBIERI, C. **BI2 – Business Intelligence: Modelagem & Qualidade**. [S.l.]: Elsevier, 2011. ISBN 9788535247244. Citado nas páginas 78 e 111.

BARBOSA, E. F.; MALDONADO, J. C. Collaborative development of educational modules: a need for lifelong learning. **E-infrastructures and technologies for lifelong learning: next generation environments**, Information Science Reference, p. 175–211, 2011. Citado na página 24.

BARNEVELD, A. V.; ARNOLD, K. E.; CAMPBELL, J. P. Analytics in higher education: Establishing a common language. **EDUCAUSE learning initiative**, v. 1, n. 1, p. 1–11, 2012. Citado nas páginas 46, 47 e 49.

BATISTA, G. E.; CARVALHO, C.P.L.F., A.; MONARD, M. C. Applying one-sided selection to unbalanced datasets. In: **MICAI 2000: Advances in Artificial Intelligence**. Springer Berlin Heidelberg, 2000, (Lecture Notes in Computer Science, v. 1793). p. 315–325. ISBN 978-3-540-67354-5. Disponível em: <http://dx.doi.org/10.1007/10720076_29>. Citado na página 63.

BATISTA, G. H. **Pré-processamento de dados em aprendizado de máquina supervisionado**. Tese (Doutorado) — ICMC-USP, São Paulo, mar 2003. Citado na página 61.

BICHSEL, J. **Analytics in higher education: Benefits, barriers, progress, and recommendations**. 2012. Citado na página 44.

BIENKOWSKI, M.; FENG, M.; MEANS, B. Enhancing teaching and learning through educational data mining and learning analytics: An issue brief. **Washington, DC: SRI International**, p. 1–77, 2012. Citado nas páginas 23, 68, 71, 72, 73 e 74.

BRESFELEAN, V. P.; BRESFELEAN, M.; GHISOIU, N.; COMES, C. A. Determining students academic failure profile founded on data mining methods. In: **ITI 2008 - 30th International Conference on Information Technology Interfaces**. [S.l.: s.n.], 2008. p. 317–322. ISSN 1330-1012. Citado na página 67.

BROOKS, D. C.; THAYER, T.-L. B. **Institutional Analytics in Higher Education**. 2016. Citado nas páginas 25, 43, 44, 50 e 51.

BROWN, M.; DEHONEY, J.; MILLICHAP, N. The next generation digital learning environment. **A Report on Research. ELI Paper. Louisville, CO: Educause**, April 2015. Citado nas páginas 39 e 98.

BRUSILOVSKY, P.; MILLÁN, E. The adaptive web. In: BRUSILOVSKY, P.; KOBSA, A.; NEJDL, W. (Ed.). Berlin, Heidelberg: Springer-Verlag, 2007. cap. User Models for Adaptive Hypermedia and Adaptive Educational Systems, p. 3–53. ISBN 978-3-540-72078-2. Disponível em: <<http://dl.acm.org/citation.cfm?id=1768197.1768199>>. Citado na página 74.

CAMPBELL, J.; DEBLOIS, P.; OBLINGER, D. Academic analytics: A new tool for a new era. **EDUCAUSE Review**, v. 4, July/August 2007. Citado na página 26.

CAMPBELL, J.; OBLINGER, D. Academic analytics. **EDUCAUSE Quarterly**, p. 1–20, october 2007. Citado nas páginas 26 e 27.

CARD, S.; MACKINLAY, J.; SCHNEIDERMAN, B. Information visualization. In: CARD, S.; MACKINLAY, J.; SCHNEIDERMAN, B. (Ed.). **Readings in Information Visualization: Using Vision to Think**. 1. ed. [S.l.]: Morgan Kaufmann Publishers, 1999. cap. 1, p. 1–34. Citado na página 65.

- CHATTI, M. A.; DYCKHOFF, A. L.; SCHROEDER, U.; THÜS, H. A reference model for learning analytics. **International Journal of Technology Enhanced Learning**, Inderscience Publishers, v. 4, n. 5-6, p. 318–331, 2012. Citado na página 169.
- CHATTI, M. A.; LUKAROV, V.; THUS, H.; MUSLIM, A.; YOUSEF, A. M. F.; WAHID, U.; GREVEN, C.; CHAKRABARTI, A.; SCHROEDER, U. Learning analytics: Challenges and future research directions. **elead**, v. 10, n. 1, 2014. ISSN 1860 – 7470. Disponível em: <<http://nbn-resolving.de/urn:nbn:de:0009-5-40350>>. Citado nas páginas 46 e 49.
- CHRISTODOULOPOULOS, C.; PAPANIKOLAOU, K. A group formation tool in an e-learning context. In: **Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on**. [S.l.: s.n.], 2007. p. 117–123. ISSN 1082-3409. Citado na página 68.
- CIFERRI, C. D. A. **Distribuição dos Dados em Ambientes de Data Warehousing: O Sistema WebD2W e Algoritmos Voltados à Fragmentação Horizontal dos Dados**. Tese (Doutorado), 2002. Citado na página 80.
- CLOW, D. The learning analytics cycle: closing the loop effectively. In: **ACM. Proceedings of the 2nd international conference on learning analytics and knowledge**. [S.l.], 2012. p. 134–138. Citado na página 122.
- COCEA, M.; WEIBELZAHN, S. Can log files analysis estimate learners level of motivation? In: _____. **LWA 2006: Lernen - Wissensentdeckung - Adaptivitat, 14th Workshop on Adaptivity and User Modeling in Interactive Systems (ABIS 2006)**. [S.l.]: University of Hildesheim, Institute of Computer Science, 2006. (Hildesheimer Informatik-Berichte, 1/2006), p. 32–35. Citado na página 67.
- CODD, E. F.; CODD, S. B.; SALLEY, C. T. Providing olap (on-line analytical processing) to user-analysts: An it mandate. **Codd and Date**, Codd & Date, Inc., v. 32, 1993. Citado na página 87.
- COMMUNITY, M. **Moodle Org**. 2016. Disponível em: <<https://moodle.org/>>. Citado na página 34.
- COOPER, ADAM. A brief history of analytics. **CETIS Analytics Series**, Citeseer, v. 1, n. 9, p. 1–21, 2012. Citado nas páginas 26, 27 e 44.
- COSTA, E.; BAKER, R. S.; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. **Jornada de Atualização em Informatica na Educação**, v. 1, n. 1, p. 1–29, 2013. Citado na página 72.
- COUNTERMAN, C.; GOLDEN, G.; GOLLUB, R.; NORTON, M.; SEVERANCE, C.; SPEELMON, L. **Technical Report Sakai Project**. 2004. Disponível em: <<http://sakaiproject.org/>>. Citado nas páginas 23, 36 e 98.
- DAVIS, B.; CARMEAN, C.; WAGNER, E. D. The evolution of the lms: From management to learning. **Santa Rosa, CA: e-Learning Guild**, 2009. Citado na página 32.

- DAWSON, S.; HEATHCOTE, L.; POOLE, G. Harnessing ict potential: The adoption and analysis of ict systems for enhancing the student learning experience. **International Journal of Educational Management**, Emerald Group Publishing Limited, v. 24, n. 2, p. 116–128, 2010. Citado na página 74.
- DEDE, C.; HO, A.; MITROS, P. Big data analysis in higher education: promises and pitfalls. **EDUCAUSE**, v. 51, p. 22–34, 2016. Disponível em: <<http://er.educause.edu/articles/2016/8/big-data-analysis-in-higher-education-promises-and-pitfalls>>. Citado nas páginas 24, 25 e 77.
- DEKKER, G.; PECHENIZKIY, M.; VLEESHOUWERS, J. Predicting students drop out: A case study. **EDM**, v. 9, p. 41–50, 2009. Citado na página 66.
- DESMARAIS, M. C.; PU, X. A bayesian student model without hidden nodes and its comparison with item response theory. **International Journal of Artificial Intelligence in Education**, IOS Press, v. 15, n. 4, p. 291–323, 2005. Citado na página 67.
- DIMOKAS, N.; MITTAS, N.; NANOPOULOS, A.; ANGELIS, L. A prototype system for educational data warehousing and mining. In: **Panhellenic Conference on Informatics**. [S.l.: s.n.], 2008. p. 199–203. Citado nas páginas 11, 28, 89, 90, 91, 101 e 102.
- DOUGIAMAS, M.; TAYLOR, P. C. **World Conference on Educational Multimedia, Hypermedia and Telecommunications**, abril 2003. Disponível em: <<http://dougiamas.com/writing/edmedia2003/>>. Citado na página 34.
- DRIGAS, A.; VRETTAROS, J. An intelligent tool for building e-learning content-material using natural language in digital libraries. **WSEAS Transactions on Information Science and Applications**, v. 1, n. 5, p. 1197–1205, 2004. Citado na página 67.
- DRINGUS, L. P.; ELLIS, T. Using data mining as a strategy for assessing asynchronous discussion forums. **Computers & Education**, n. 1, p. 141–160, 2005. ISSN 0360-1315. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0360131504000788>>. Citado na página 70.
- DUTT, A.; ISMAIL, M. A.; HERAWAN, T. A systematic review on educational data mining. **IEEE Access**, IEEE, 2017. Citado nas páginas 65, 101 e 122.
- EBERSPACHER, H. F.; VASCONCELOS, C. D.; JAMUR, J. H.; ELEUTERIO, M. A. M. Eureka: Um ambiente de aprendizagem cooperativa baseado na Web para educação a distância. In: **SBIE'99: X Simpósio Brasileiro de Informática na Educação**. Curitiba/PR, Brasil: [s.n.], 1999. p. 341–348. Citado na página 32.
- ECKERSON, W. Ten characteristics of a good KPI. **The Data Warehousing Institute**, 2007. Citado na página 52.
- EVELSON, B.; NICOLSON, N. Topic overview: Business intelligence. Forrester, november 2008. Disponível em: <<https://www.forrester.com/report/Topic+Overview+Business+Intelligence/-/E-RES39218>>. Citado na página 88.

EVERITT, B. **Cluster analysis**. [S.l.]: Heinemann Educational [for] the Social Science Research Council, 1974. (Reviews of current research). ISBN 9780435822989. Citado na página 67.

FARMER, J.; DOLPHIN, I. Sakai: elearning and more. **EUNIS 2005-Leadership and Strategy in a Cyber-Infrastructure World**, 2005. Citado nas páginas 36 e 98.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Knowledge discovery and data mining: Towards a unifying framework. In: **Proceedings of the Second International Conference on Knowledge Discovery and Data Mining**. AAAI Press, 1996. (KDD'96), p. 82–88. Disponível em: <<http://dl.acm.org/citation.cfm?id=3001460.3001477>>. Citado nas páginas 60 e 61.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Advances in knowledge discovery and data mining. In: FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Ed.). Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996. cap. From Data Mining to Knowledge Discovery: An Overview, p. 1–34. ISBN 0-262-56097-6. Disponível em: <<http://dl.acm.org/citation.cfm?id=257938.257942>>. Citado na página 60.

FELDMAN, R.; SANGER, J. **The text mining handbook: advanced approaches in analyzing unstructured data**. [S.l.]: Cambridge University Press, 2006. Citado na página 70.

FISCHER, G. User modeling in human-computer interaction. **User Modeling and User-Adapted Interaction**, Kluwer Academic Publishers, v. 11, n. 1-2, p. 65–86, 2001. ISSN 0924-1868. Citado na página 72.

Free Software Foundation. **GNU General Public License**. 2007. Disponível em: <<http://www.gnu.org/copyleft/gpl.html>>. Citado na página 36.

FU, H.; OFOGLU, M. A conceptual subspace clustering algorithm in e-learning. In: **Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on**. [S.l.: s.n.], 2008. v. 3, p. 1983–1988. ISSN 1738-9445. Citado na página 68.

GRAJEK, S. Top 10 it issues, 2016: Divest, reinvest, and differentiate. jan 2016. Disponível em: <<http://er.educause.edu/articles/2016/1/top-10-it-issues-2016>>. Citado na página 25.

_____. Top 10 it issues, 2017: Foundations for student success. jan 2017. Disponível em: <<http://er.educause.edu/articles/2017/1/top-10-it-issues-2017-foundations-for-student-success>>. Citado na página 25.

GRÜNEWALD, F.; MEINEL, C.; TOTSCHNIG, M.; WILLEMS, C. Designing moocs for the support of multiple learning styles. In: **Scaling up Learning for Sustained Impact**. [S.l.]: Springer, 2013. p. 371–382. Citado nas páginas 38 e 66.

- HA, S. H.; BAE, S. M.; PARK, S.-C. Web mining for distance education. In: **Management of Innovation and Technology, 2000. ICMIT 2000. Proceedings of the 2000 IEEE International Conference on**. [S.l.: s.n.], 2000. p. 715–719. Citado na página 69.
- HÄMÄLÄINEN, W.; SUHONEN, J.; SUTINEN, E.; TOIVONEN, H. Data mining in personalizing distance education courses. In: **Proceedings of the 21st ICDE World Conference on Open Learning and Distance Education**. [S.l.: s.n.], 2004. p. 18–21. Citado na página 67.
- HAMMOUDA haled; KAMEL, M. Data mining in e-learning. In: PIERRE, S. (Ed.). **E-Learning Networked Environments and Architectures**. Springer London, 2007, (Advanced Information and Knowledge Processing). p. 374–404. ISBN 978-1-84628-351-2. Disponível em: <http://dx.doi.org/10.1007/978-1-84628-758-9_13>. Citado na página 70.
- HAN, J.; KAMBER, M.; PEI, J. **Data mining: concepts and techniques**. [S.l.]: Morgan kaufmann, 2006. Citado na página 62.
- HAYTHORNTHWAITE, C. Social network analysis: An approach and technique for the study of information exchange. **Library & Information Science Research**, v. 18, n. 4, p. 323–342, 1996. ISSN 0740-8188. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0740818896900031>>. Citado na página 70.
- HODGE, V.; AUSTIN, J. A survey of outlier detection methodologies. **Artificial Intelligence Review**, Kluwer Academic Publishers, v. 22, n. 2, p. 85–126, 2004. ISSN 0269-2821. Disponível em: <<http://dx.doi.org/10.1007/s10462-004-4304-y>>. Citado na página 69.
- HURLEY, T.; WEIBELZAHN, S. Eliciting adaptation knowledge from on-line tutors to increase motivation. In: CONATI, C.; MCCOY, K.; PALIOURAS, G. (Ed.). **User Modeling 2007**. Springer Berlin Heidelberg, 2007, (Lecture Notes in Computer Science, v. 4511). p. 370–374. ISBN 978-3-540-73077-4. Disponível em: <http://dx.doi.org/10.1007/978-3-540-73078-1_47>. Citado na página 67.
- IGLESIAS-PRADAS, S.; AZCARATE, C. R. de; AGUDO-PEREGRINA, A. F. Assessing the suitability of student interactions from moodle data logs as predictors of cross-curricular competencies. **Computers in Human Behavior**, v. 47, p. 81 – 89, 2015. ISSN 0747-5632. Learning Analytics, Educational Data Mining and data-driven Educational Decision Making. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0747563214005238>>. Citado na página 98.
- IMS Global Learning Consortium. **IMS Project**. 2006. Disponível em: <<http://www.imsproject.org/>>. Citado na página 36.
- INEP. Indicadores educacionais. oct 2015. Disponível em: <<http://portal.inep.gov.br/indicadores-educacionais>>. Citado nas páginas 53, 54, 55 e 100.
- _____. Sinopse estatística de educação superior 2014. oct 2016. Disponível em: <<http://portal.inep.gov.br/web/guest/sinopses-estatisticas-da-educacao-superior>>. Citado nas páginas 55 e 100.

- INMON, W. H. **Building the data warehouse**. 3. ed. [S.l.]: John Wiley & sons, 2005. ISBN 0-471-08130-2. Citado nas páginas 79, 80, 81, 86, 111 e 120.
- INMON, W. H.; STRAUSS, D.; NEUSHLOSS, G. **DW 2.0: The architecture for the next generation of data warehousing**. [S.l.]: Morgan Kaufmann, 2010. ISBN 978-0-12-374319-0. Citado nas páginas 79 e 81.
- JEONG, H.; BISWAS, G. Mining student behavior models in learning-by-teaching environments. In: **The 1st International Conference on Educational Data Mining – EDM**. [S.l.: s.n.], 2008. p. 127–136. Citado na página 71.
- JONSSON, A.; JOHNS, J.; MEHRANIAN, H.; ARROYO, I.; WOOLF, B.; BARTO, A.; FISHER, D.; MAHADEVAN, S. Evaluating the feasibility of learning student models from data. In: **AAAI Workshop on Educational Data Mining**. [S.l.: s.n.], 2005. p. 1–6. Citado na página 67.
- KAKISH, K.; KRAFT, T. A. Etl evolution for real-time data warehousing. In: **Proceedings of the Conference on Information Systems Applied Research ISSN**. [S.l.: s.n.], 2012. v. 2167, p. 1508. Citado na página 82.
- KANTARDZIC, M. **Data Mining: Concepts, Models, Methods, and Algorithms**. Piscataway, NJ, EUA: John Wiley & Sons, 2003. ISBN 0471228524. Citado nas páginas 64 e 66.
- KARDAN, S.; CONATI, C. A framework for capturing distinguishing user interaction behaviors in novel interfaces. **The 4th International Conference on Educational Data Mining – EDM**, v. 2011, p. 4th, 2011. Citado na página 74.
- KHALIL, M.; EBNER, M. Learning analytics in moocs: Can data improve students retention and learning? In: ASSOCIATION FOR THE ADVANCEMENT OF COMPUTING IN EDUCATION (AACE). **EdMedia: World Conference on Educational Media and Technology**. [S.l.], 2016. p. 581–588. Citado na página 24.
- KIM, J.; CHERN, G.; FENG, D.; SHAW, E.; HOVY, E. Mining and assessing discussions on the web through speech act analysis. In: **Proceedings of the Workshop on Web Content Mining with Human Language Technologies at the 5th International Semantic Web Conference**. [S.l.: s.n.], 2006. Citado na página 70.
- KIMBALL, R.; ROSS, M. **The data warehouse toolkit: The definitive guide to dimensional modeling**. [S.l.]: John Wiley & Sons, 2013. Citado nas páginas 81, 83, 84, 85 e 113.
- KOPER, R.; OLIVIER, B.; ANDERSON, T. IMS Learning Design Information Model. **IMS Global Learning Consortium**, 2003. Citado na página 36.
- KOTSIANTIS, S. B.; PIERRAKEAS, C.; PINTELAS, P. E. Preventing student dropout in distance learning using machine learning techniques. In: SPRINGER. **Knowledge-Based Intelligent Information and Engineering Systems**. [S.l.], 2003. p. 267–274. Citado na página 67.

- KUMAR, M.; SINGH, A.; HANDA, D. Literature survey on educational dropout prediction. **International Journal of Education and Management Engineering**, 2017. ISSN 2305-3623. Disponível em: <<http://www.mecspress.org/ijeme>>. Citado na página 101.
- LANG, L.; PIRANI, J. A. The learning management system evolution. cds spotlight report. research bulletin. **EDUCAUSE**, ERIC, 2014. Citado nas páginas 32 e 33.
- _____. **The 2015 Enterprise Application Market: Business Intelligence Reporting and Data Warehouse Systems**. 2016. Citado na página 77.
- LEE, M.-G. Profiling students' adaptation styles in web-based learning. **Computers & Education**, v. 36, n. 2, p. 121–132, 2001. ISSN 0360-1315. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0360131500000464>>. Citado na página 67.
- LI, J.; WANG, X. Study the use of sakai in distance education. In: IEEE. **Educational and Information Technology (ICEIT), 2010 International Conference on**. [S.l.], 2010. v. 3, p. V3–167. Citado na página 98.
- LIÑÁN, L. C.; PÉREZ, Á. A. J. Educational data mining and learning analytics: differences, similarities, and time evolution. **International Journal of Educational Technology in Higher Education**, Springer, v. 12, n. 3, p. 98–112, 2015. Citado na página 65.
- LOSHIN, D. **Business intelligence: the savvy manager's guide**. [S.l.]: Newnes, 2012. Citado na página 88.
- LUHN, H. P. The automatic creation of literature abstracts. **IBM Journal os Research and Development**, v. 2, n. 2, p. 159–165, 1958. Citado na página 88.
- MACFADYEN, L. P.; DAWSON, S. Mining LMS data to develop an early warning system for educators: A proof of concept. **Computers & Education**, n. 2, p. 588–599, 2010. ISSN 0360-1315. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0360131509002486>>. Citado na página 73.
- MACQUEEN, J. B. Some methods for classification and analysis of multivariate observations. In: CAM, L. M. L.; NEYMAN, J. (Ed.). **Proceedings of the V Berkeley Symposium on Mathematical Statistics and Probability**. [S.l.]: University of California Press, 1967. v. 1, p. 281–297. Citado na página 142.
- MARCHENKO, O. Functional comparison of open edx and moodle platforms. **Processo de educação moderna em ambiente aberto**, n. 1, p. 171–178, 2016. Citado na página 98.
- MARTIN, B.; MITROVIC, A.; KOEDINGER, K.; MATHAN, S. Evaluating and improving adaptive educational systems with learning curves. **User Modeling and User-Adapted Interaction**, Springer Netherlands, v. 21, n. 3, p. 249–283, 2011. ISSN 0924-1868. Disponível em: <<http://dx.doi.org/10.1007/s11257-010-9084-2>>. Citado na página 75.
- MARTIN, M.; SAUVAGEOT, C.; TCHATCHOUA, B. Constructing an indicator system or scorecard for higher education. a practical guide. **International Institute for Educational Planning. Paris. UNESCO**, 2011. Citado nas páginas 24, 25, 49, 52, 53 e 100.

- MARTINS, R. **Projeto complementa a educação formal de forma interativa**. 2009. Disponível em: <<http://www.e-educador.com/index.php/mundo-high-tech-mainmenu-99/5325-tidia>>. Citado na página 37.
- MAULL, K. E.; SALDIVAR, M. G.; SUMNER, T. Observing online curriculum planning behavior of teachers. In: **EDM**. [S.l.: s.n.], 2010. p. 303–304. Citado na página 68.
- MAZZA, R. **Introduction to information visualization**. [S.l.]: Springer Science & Business Media, 2009. Citado na página 49.
- MAZZA, R.; BOTTURI, L. Monitoring an online course with the gismo tool: A case study. **Journal of Interactive Learning Research**, AACE, Chesapeake, VA, v. 18, n. 2, p. 251–265, April 2007. ISSN 1093-023X. Disponível em: <<http://www.editlib.org/p/22972>>. Citado na página 36.
- MEC. **UAB – Universidade Aberta do Brasil, Ministério da Educação**. 2005. Disponível em: <<http://www.capes.gov.br/uab>>. Citado na página 23.
- MINAEI-BIDGOLI, B.; KASHY, D.; KORTMEYER, G.; PUNCH, W. Predicting student performance: an application of data mining methods with an educational web-based system. In: **Frontiers in Education, 2003. FIE 2003 33rd Annual**. [S.l.: s.n.], 2003. v. 1, p. T2A–13–18 Vol.1. ISSN 0190-5848. Citado na página 67.
- MIT and Harvard University. **Edx – Online Education**. 2013. Disponível em: <<https://www.edx.org/school/microsoft>>. Citado nas páginas 24 e 98.
- MITCHELL, J.; COSTELLO, S.; MITCHELL, J.; IMAGE, E. *et al.* International e-vet market research report: A report on international market research for australian vet online products and services. John Mitchell & Associates, 2000. Citado na página 147.
- MITCHELL, T. M. **Machine Learning**. [S.l.]: McGraw-Hill Higher Education, 1997. ISBN 0070428077. Citado na página 64.
- MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE, S. O. (Ed.). **Sistemas Inteligentes: Fundamentos e Aplicacoes**. 1. ed. [S.l.]: Manole, 2003. cap. 4, p. 89–114. Citado na página 64.
- MOODLE, D. G. **Moodle – A Free, Open Source Course Management System for Online Learning**. 2006. Disponível em: <<http://moodle.org/>>. Citado na página 23.
- MOODLEDOCS. **Moodle Docs – The documentation for Moodle that everyone may contribute**. 2006. Disponível em: <<http://docs.moodle.org/>>. Citado na página 34.
- MOSCOSO-ZEA, O.; ANDRES-SAMPEDRO; LUJÁN-MORA, S. Datawarehouse design for educational data mining. In: **2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET)**. [S.l.: s.n.], 2016. p. 1–6. Citado nas páginas 11, 26, 28, 89, 90, 101 e 102.

MOSLEY, M.; BRACKETT, M. H.; EARLEY, S.; HENDERSON, D. **DAMA guide to the data management body of knowledge**. [S.l.]: Technics Publications, 2010. Citado na página 87.

MUEHLENBROCK, M. Automatic action analysis in an interactive learning environment. In: **The 12th international conference on artificial intelligence in education, AIED**. [S.l.: s.n.], 2005. p. 73–80. Citado na página 70.

MUNTEAN, M.; BOLOGA, A.; BOLOGA, R.; FLOREA, A. Business intelligence systems in support of university strategy. **Recent Researches in Educational Technologies**, p. 118–123, 2011. Citado nas páginas 102 e 103.

NEBIC, Z.; MAHNIČ, V. Data warehouse for an e-learning platform. In: WORLD SCIENTIFIC AND ENGINEERING ACADEMY AND SOCIETY (WSEAS). **Proceedings of the 14th WSEAS international conference on Computers: part of the 14th WSEAS CSCC multiconference-Volume II**. [S.l.], 2010. p. 415–420. Citado nas páginas 28, 89 e 102.

NETO, F. A. A.; CASTRO, A. Elicited and mined rules for dropout prevention in online courses. In: **2015 IEEE Frontiers in Education Conference (FIE)**. [S.l.: s.n.], 2015. p. 1–7. Citado nas páginas 98 e 101.

NG, A.; KOLLER, D. **Coursera**. 2012. Disponível em: <<https://www.coursera.org/>>. Citado na página 24.

NGHE, N. T.; JANECEK, P.; HADDAWY, P. A comparative analysis of techniques for predicting academic performance. In: **Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE '07. 37th Annual**. [S.l.: s.n.], 2007. p. T2G–7–T2G–12. ISSN 0190-5848. Citado na página 66.

OKI. **Open Knowledge Initiative – Accelerated interoperability through simplified integration**. 2008. Disponível em: <<http://www.okiproject.org/>>. Citado na página 36.

OpenCourseWare. **OWC – OpenCourseWare**. 1999. Disponível em: <<http://www.ocwconsortium.org/>>. Citado na página 23.

PAHL, C.; DONNELLAN, D. Data mining technology for the evaluation of web-based teaching and learning systems. 7th Int. Conference on E-Learning in Business, Government and Higher Education, 2002. Citado na página 69.

PARDOS, Z. A.; KAO, K. moocrp: An open-source analytics platform. In: **Proceedings of the Second (2015) ACM Conference on Learning Scale**. New York, NY, USA: ACM, 2015. p. 103–110. ISBN 978-1-4503-3411-2. Disponível em: <<http://doi.acm.org/10.1145/2724660.2724683>>. Citado na página 98.

PARMENTER, D. **Key Performance Indicators: Developing, Implementing, and Using Winning KPIs**. Wiley, 2015. ISBN 9781118925102. Disponível em: <<https://books.google.com.br/books?id=bKkxBwAAQBAJ>>. Citado nas páginas 52, 53 e 54.

- PARTHASARATHY, S.; ZAKI, M. J.; OGIHARA, M.; DWARKADAS, S. Incremental and interactive sequence mining. In: **Proceedings of the Eighth International Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 1999. p. 251–258. ISBN 1-58113-146-1. Disponível em: <<http://doi.acm.org/10.1145/319950.320010>>. Citado na página 69.
- PASYEKA, N.; PASYEKA, M. Construction of multidimensional data warehouse for processing students' knowledge evaluation in universities. In: **2016 13th International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET)**. [S.l.: s.n.], 2016. p. 822–824. Citado nas páginas 28, 89 e 102.
- PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, v. 2, n. 6, p. 559–572, 1901. Citado na página 155.
- PEÑA-AYALA, A. Educational data mining: A survey and a data mining-based analysis of recent works. **Expert Systems with Applications**, v. 41, n. 4, p. 1432 – 1462, 2014. ISSN 0957-4174. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0957417413006635>>. Citado nas páginas 59 e 101.
- PONNIAH, P. **Data Warehousing Fundamentals**. [S.l.]: John Wiley & Sons, USA, 2001. Citado na página 80.
- PROJECT, M. **Moodle**. 2017. Disponível em: <<https://moodle.com/>>. Citado na página 34.
- REYES, P.; TCHOUNIKINE, P. Mining learning groups' activities in forum-type tools. In: **Proceedings of th 2005 Conference on Computer Support for Collaborative Learning: Learning 2005: The Next 10 Years!** International Society of the Learning Sciences, 2005. p. 509–513. ISBN 0-8058-5782-6. Disponível em: <<http://dl.acm.org/citation.cfm?id=1149293.1149360>>. Citado nas páginas 32 e 70.
- REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; PAULA, M. F. Mineração de dados. In: REZENDE, S. O. (Ed.). **Sistemas Inteligentes: Fundamentos e Aplicacoes**. 1. ed. [S.l.]: Manole, 2003. cap. 12, p. 307–335. Citado nas páginas 60, 61, 62, 65 e 66.
- RHYS, R.-J.; DAVID, H.; TIM, C. Educational performance indicators. **Management of Sustainable Development**, v. 2, 2010. Citado na página 100.
- ROMERO, C.; ESPEJO, P. G.; ZAFRA, A.; ROMERO, J. R.; VENTURA, S. Web usage mining for predicting final marks of students that use moodle courses. **Computer Applications in Engineering Education**, Wiley Online Library, n. 1, p. 135–146, 2013. Citado na página 36.
- ROMERO, C.; VENTURA, S. Educational data mining: A survey from 1995 to 2005. **Expert Systems with Applications**, v. 33, n. 1, p. 135–146, 2007. ISSN 0957-4174. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0957417406001266>>. Citado nas páginas 59, 71 e 101.

_____. Educational data mining: a review of the state of the art. **Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on**, IEEE, v. 40, n. 6, p. 601–618, 2010. ISSN 1094-6977. Disponível em: <<http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5524021>>. Citado nas páginas 23, 59, 65, 71, 72 e 101.

_____. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, Wiley Online Library, v. 3, n. 1, p. 12–27, 2013. Citado nas páginas 70 e 71.

ROMERO, C.; VENTURA, S.; GARCIA, E. Data mining in course management systems: Moodle case study and tutorial. **Computers & Education**, v. 51, n. 1, p. 368–384, 2008. ISSN 0360-1315. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S0360131507000590>>. Citado nas páginas 36 e 101.

ROMERO, C.; VENTURA, S.; PECHENIZKIY, M.; BAKER, R. S. **Handbook of educational data mining**. [S.l.]: Taylor & Francis US, 2011. Citado nas páginas 66 e 67.

ROUSSEEUW, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, Elsevier Science Publishers B. V., v. 20, n. 1, p. 53–65, 1987. ISSN 0377-0427. Citado na página 155.

RUDY; MIRANDA, E.; SURYANI, E. Implementation of datawarehouse, datamining and dashboard for higher education. **Journal of Theoretical & Applied Information Technology**, v. 64, n. 3, 2014. Citado nas páginas 28, 89 e 102.

SAEL, A. M. N.; BEHJA, H. Web usage mining data preprocessing and multi level analysis on moodle. In: **Computer Systems and Applications (AICCSA), 2013 ACS International Conference on**. [s.n.], 2013. p. 1–7. ISSN 2161-5322. Disponível em: <<http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6616427>>. Citado nas páginas 36 e 101.

SAKAI. **Sakai Project**. 2014. Disponível em: <<http://sakaiproject.org/>>. Citado na página 36.

SHAHIRI, A. M.; HUSAIN, W.; RASHID, N. A. A review on predicting student's performance using data mining techniques. **Procedia Computer Science**, v. 72, p. 414 – 422, 2015. ISSN 1877-0509. Disponível em: <<http://www.sciencedirect.com/science/article/pii/S1877050915036182>>. Citado nas páginas 67 e 141.

SHUM, S. J. B. Learning analytics. **UNESCO policy brief**, 2012. Citado nas páginas 45, 46 e 120.

SIEMENS, G. Learning analytics: The emergence of a discipline. **American Behavioral Scientist**, SAGE Publications, p. 0002764213498851, 2013. Citado na página 50.

SIEMENS, G.; GASEVIC, D.; HAYTHORNTHWAITE, C.; DAWSON, S.; SHUM, S. B.; FERGUSON, R.; DUVAL, E.; VERBERT, K.; BAKER, R. Open learning analytics: an integrated & modularized platform. **CETIS Analytics Series**, 2011. Citado na página 44.

- SIEMENS, G.; LONG, P. Penetrating the fog: Analytics in learning and education. **EDUCAUSE review**, ERIC, v. 46, n. 5, p. 30, 2011. Citado nas páginas 45 e 46.
- SILBERSCHATZ, A.; TUZHILIN, A. On subjective measures of interestingness in knowledge discovery. In: **KDD '95: Proceedings of I International Conference on Knowledge Discovery and Data Mining**. [s.n.], 1995. p. 275–281. Disponível em: <citeseer.ist.psu.edu/silberschatz95subjective.html>. Citado na página 65.
- STONE, M. Cross-Validatory Choice and Assessment of Statistical Predictions. **Journal of the Royal Statistical Society. Series B (Methodological)**, Blackwell Publishing for the Royal Statistical Society, v. 36, n. 2, p. 111–147, 1974. Citado na página 144.
- SUPERBY, J.; VANDAMME, J.; MESKENS, N. Determination of factors influencing the achievement of the first-year university students using data mining methods. In: **Workshop on Educational Data Mining**. [S.l.: s.n.], 2006. p. 37–44. Citado na página 66.
- TANE, J.; SCHMITZ, C.; STUMME, G. Semantic resource management for the web: An e-learning application. In: **Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters**. New York, NY, USA: ACM, 2004. p. 1–10. ISBN 1-58113-912-8. Disponível em: <<http://doi.acm.org/10.1145/1013367.1013369>>. Citado nas páginas 67 e 70.
- THRUN, S.; STAVENS, D.; SOKOLSKY, M. **Udacity**. 2011. Disponível em: <<https://www.udacity.com/>>. Citado na página 24.
- TIAN, F.; WANG, S.; ZHENG, C.; ZHENG, Q. Research on e-learner personality grouping based on fuzzy clustering analysis. In: **Computer Supported Cooperative Work in Design, 2008. CSCWD 2008. 12th International Conference on**. [S.l.: s.n.], 2008. p. 1035–1040. Citado na página 68.
- TIDIA. Tecnologia da informação para o desenvolvimento da internet avançada. 2006. Disponível em: <<http://www.tidia.fapesp.br/portal>>. Citado nas páginas 23, 36, 38 e 98.
- TidiaAe Portal. **Ae Aprendizado Eletrônico – Portal**. 2008. Disponível em: <<http://www.tidia-ae.usp.br/portal>>. Citado na página 36.
- TRAVASSO, G. H.; GUROV, D.; AMARAL, E. A. G. **Introdução à Engenharia de Software**. [S.l.], 2002. Citado na página 127.
- UENO, M. Online outlier detection system for learning time data in e-learning and its evaluation. In: **International conference on computers and advanced technology in education**. [S.l.: s.n.], 2004. p. 248–253. Citado na página 70.
- VALDIVIEZO-DÍAZ, P.; CORDERO, J.; REÁTEGUI, R.; AGUILAR, J. A business intelligence model for online tutoring process. In: **2015 IEEE Frontiers in Education Conference (FIE)**. [S.l.: s.n.], 2015. p. 1–9. Citado nas páginas 11, 89, 91, 92, 98, 101 e 102.

WAI, T. T.; AUNG, S. S. Metadata based student data extraction from universities data warehouse. In: **2009 International Conference on Signal Processing Systems**. [S.l.: s.n.], 2009. p. 670–673. Citado nas páginas 28, 89 e 102.

WEISS, S. M.; INDURKHYA, N. **Predictive data mining: a practical guide**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1998. ISBN 1-55860-403-0. Citado nas páginas 60 e 63.

WITTEN, I.; FRANK, E.; HALL, M. **Data Mining: Practical machine learning tools and techniques**. [S.l.]: Morgan Kaufmann, 2011. Citado na página 60.

YANOSKY, R.; ARROWAY, P. **The Analytics Landscape in Higher Education**. 2015. Citado na página 26.

YU, T.; JO, I.-H. Educational technology approach toward learning analytics: Relationship between student online behavior and learning performance in higher education. In: ACM. **Proceedings of the Fourth International Conference on Learning Analytics and Knowledge**. [S.l.], 2014. p. 269–270. Citado nas páginas 101 e 102.

ZORRILLA, M. E.; MENASALVAS, E.; MARIN, D.; MORA, E.; SEGOVIA, J. Web usage mining project for improving web-based learning sites. In: SPRINGER. **International Conference on Computer Aided Systems Theory**. [S.l.], 2005. p. 205–210. Citado na página 77.