Caracterização de alunos em ambientes de
ensino online: estendendo o uso da DAMICORE
para minerar dados educacionais

Luis Fernando de Souza Moro

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Luis Fernando de Souza Moro

Caracterização de alunos em ambientes de ensino online: estendendo o uso da DAMICORE para minerar dados educacionais

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional *VERSÃO REVISADA*.

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. Seiji Isotani

USP – São Carlos
Junho de 2015

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

d867c de Souza Moro, Luis Fernando
Caracterização de alunos em ambientes de ensino
online: estendendo o uso da DAMICORE para minerar
dados educacionais / Luis Fernando de Souza Moro;
orientador Seiji Isotani. -- São Carlos, 2015.
117 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Ciências de Computação e Matemática
Computacional) -- Instituto de Ciências Matemáticas
e de Computação, Universidade de São Paulo, 2015.

1. Mineração de Dados Educacionais. 2. Seleção de
Atributos para Mineração de Dados. I. Isotani,
Seiji, orient. II. Título.

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____

Luis Fernando de Souza Moro

Characterization of students in online learning
environments: extending the use of dAMICORE to
educational data Mining.

Master dissertation submitted to the Instituto de
Ciências Matemáticas e de Computação - ICMC-
USP, in partial fulfillment of the requirements for the
degree of the Master Program in Computer Science
and Computational Mathematics. *FINAL VERSION*.

Área de Concentração: Computer Science and
Computational Mathematics

Advisor: Prof. Dr. Seiji Isotani

USP – São Carlos
June 2015

Agradecimentos

À Vanessa Kiraly, com admiração e gratidão por sua compreensão, carinho, presença e incansável apoio ao longo do período de elaboração deste trabalho.

A Renata Moreira Grass por toda a sua ajuda e dedicação para a conclusão desse trabalho.

Aos meus amigos que tanto me apoiaram durante essa caminhada, em especial, ao Daniel Tozadore pelos momentos de alegria.

À minha família, Heitor Luiz Rodrigues Moro, Sandra Aparecida de Souza Moro, Marcio Augusto de Souza Moro e Ana Carolina de Souza Moro, pelo apoio e suporte em todos os momentos.

Ao meu orientador Seiji Isotani, pelo apoio e orientação dedicados a mim para que esse trabalho fosse concluído com êxito.

Aos meus colegas do CAEd, em especial para Carla Rodriguez, Aparecida Maria Zem Lopes e Helena Macedo Reis, que sempre estiveram ao meu lado apoiando e auxiliando o desenvolvimento desse trabalho.

Resumo

Com a popularização do uso de recursos tecnológicos na educação, uma enorme quantidade de dados, relacionados às interações entre alunos e esses recursos, é armazenada. Analisar esses dados, visando caracterizar os alunos, é tarefa muito importante, uma vez que os resultados dessa análise podem auxiliar professores no processo de ensino e aprendizagem. Entretanto, devido ao fato de as ferramentas utilizadas para essa caracterização serem complexas e pouco intuitivas, os profissionais da área de ensino acabam por não utilizá-las, inviabilizando a implementação de tais ferramentas em ambientes educacionais. Dentro desse contexto, a dissertação de mestrado aqui apresentada teve como objetivo analisar os dados provenientes de um sistema tutor inteligente, o MathTutor, que disponibiliza exercícios específicos de matemática, para identificar padrões de comportamento dos alunos que interagiram com esse sistema durante um determinado período. Essa análise foi realizada por meio de um processo de Mineração de Dados Educacionais (EDM), utilizando a ferramenta DAMICORE, com o intuito de possibilitar que fossem geradas, de forma rápida e eficaz, informações úteis à caracterização dos alunos. Durante a realização dessa análise, seguiram-se algumas fases do processo de descobrimento de conhecimento em bases de dados, “seleção”, “pré-processamento”, “mineração dos dados” e “avaliação e interpretação”. Na fase de “mineração de dados”, foi utilizada a ferramenta DAMICORE, que encontrou padrões que foram estudados na fase de “avaliação e interpretação”. A partir dessa análise foram encontrados padrões comportamentais dos alunos, por exemplo, alunos do sexo masculino apresentam rendimento superior ou inferior ao de alunas do sexo feminino e quais alunos terão um bom ou mau rendimento nas etapas finais do processo de ensino. Como principal resultado temos que uma das hipóteses criadas, “Alunos que obtiveram bom desempenho no “pós-teste imediato” apresentaram dois dos três seguintes comportamentos: poucas interações na “intervenção”, baixo tempo interagindo com o sistema na “intervenção” e poucos *misconceptions* no “pré-teste”, teve sua acurácia comprovada dentre os dados utilizados nessa pesquisa. Assim, por meio desta pesquisa concluiu-se que a utilização da DAMICORE em contexto educacional pode auxiliar o professor a inferir o desempenho dos seus alunos oferecendo a ele a oportunidade de realizar as intervenções pedagógicas que auxiliem alunos com possíveis dificuldades e apresente novos desafios para aqueles com facilidade no tema estudado

Palavras-Chave: Mineração de Dados Educacionais, Atributos.

Abstract

With the popularization of the use of technological resources in education, a huge amount of data, related to the interactions between students and these resources, is stored. Analyzing this data, due to characterize the students, is an important task, since the results of this analysis can help teachers on teaching and learning process. However, due to the fact that the tools used to this characterization are complex and non-intuitive, the educational professionals do not use it, invalidating the implementation of such tools at educational environments. Within this context, this master's dissertation aimed analyzing the convenient data from an educational web system named MathTutor, which offers specific math exercises to identify behavioral patterns of students who interacted with this system during some period. This analysis was performed by a process known as Educational Data Mining, using the tool named DAMICORE, in order to enable quickly and effectively the construction of helpful information to the characterization of the students. During the course of this analysis, some phases of the process of knowledge discovery in databases were followed: "selection", "preprocessing", "data mining" and "evaluation and interpretation". In "data mining" phase, the tool DAMICORE was used to find behavioral patterns of students which were studied at the "evaluation and interpretation" phase. From this analysis, behavioral patterns of students were found, for example, male students have higher or lower yield against the female students and which students are going to have a good or bad yield on the final steps of the educational process. As the main result we have one of the made assumptions, "Students who get good performance in the "immediate posttest" showed two of the following behaviors: few interactions in the "intervention", low time interacting with the system in the "intervention" and few misconceptions in "pretest", has proven its accuracy among the data used in this dissertation. Thus, through this research, it was concluded that the use of DAMICORE at educational context can help teacher to infer the performance of their students offering him the opportunity to perform the pedagogical interventions that help students who faces difficulties and show new challenges for those who have facilities in the subject studied.

Keywords: Educational Data Mining, Attributes

Sumário

Agradecimentos	7
Resumo	8
Abstract.....	9
Sumário.....	10
Lista de Figuras	12
Lista de Tabelas	14
1. Introdução	16
1.1. Problemática e Justificativa.....	17
1.2. Questões de Pesquisa.....	19
1.3. Objetivos	20
1.4. Organização.....	20
2. Fundamentação Teórica.....	22
2.1. Descoberta de Conhecimento em Bases de Dados.....	22
2.1.1. Seleção	23
2.1.2. Pré-processamento de dados	24
2.1.3. Transformação	24
2.1.4. Mineração de Dados	25
2.1.5. Interpretação e Avaliação	25
2.2. A área de Pesquisa em Mineração de Dados Educacionais	26
2.2.1. Técnicas utilizadas em Mineração de Dados Educacionais.....	27
3. Trabalhos Relacionados.....	32
4. Escopo do Estudo	40
4.1. Contextos: escolas, alunos e o experimento	40
4.1.1. MathTutor	40
4.1.2. O experimento.....	43
4.1.3. A etapa de “intervenção”	45
Assim, todos os alunos tiveram suas interações armazenadas em arquivos de log, que foram transformados em planilhas (.....	47
4.2. A estrutura dos Dados Utilizados ou a Planilha de <i>log</i>	47
4.3. DAMICORE.....	49
4.3.1. Técnicas utilizadas pela DAMICORE	50
5. Seleção de Dados Educacionais e sua Relevância para a Caracterização de Alunos.....	53

5.1.	Mapeamento Sistemático	54
5.1.1.	Definição das questões de pesquisa do Mapeamento	54
5.1.2.	Condução das atividades do Mapeamento Sistemático	55
5.2.	Análise dos Estudos Primários	57
5.3.	Síntese dos estudos	64
5.4.	Ameaças à Validade e Considerações Finais	65
6.	Desenvolvimento e Discussão	67
6.1.	Fase 1: seleção dos dados	68
6.2.	Fase 2: Pré-processamento dos Dados	71
6.3.	Fase 3: mineração dos Dados	71
6.4.	Fase 4: interpretação e avaliação dos dados	73
6.5.	Resultados Obtidos	93
7.	Conclusão	100
	Referencias	103
	ANEXO A – Recorte da planilha de log gerada pelo MathTutor	108
	APÊNDICE A – Tabela de classificação dos artigos	117

Lista de Figuras

Figura 1 - Mineração de Dados adaptado de Silva, 2004.....	22
Figura 2 - Fases do Descobrimento de Conhecimento em Bases de Dados (adaptado de Fayyad, Piatetsky-Shapiro E Smyth, 1996).....	23
Figura 3 - MathTutor: interface <i>Software</i> de Auxílio ao Módulo Ensino de Números Decimais	41
Figura 4 - Math Tutor: <i>drag and drop</i> Sistema de Auxílio ao Ensino de Números Decimais.	42
Figura 5 - MathTutor: exercício <i>Software</i> de Auxílio ao Ensino de Números Decimais.....	43
Figura 6 - Etapas do Experimento (Adaptado De McLaren et al., 2012).....	44
Figura 7 - Passos de um Mapeamento Sistemático (adaptado de Petersen et al. (2008))	54
Figura 8 - Quantidade dos Objetivos encontrados nos estudos	61
Figura 9 - Quantidade dos Atributos encontrados nos estudos	62
Figura 10 - Quantidade dos Atributos encontrados nos estudos que possuem o Objetivo de Caracterizar os Alunos.....	63
Figura 11 - Fases do KDD personalizadas	67
Figura 12 - Árvore da intervenção gerada pela DAMICORE	72
Figura 13 - As nove árvores geradas pela DAMICORE	73
Figura 14 - Árvore da intervenção com porcentagens de desempenho do "pós-teste imediato"	74
Figura 15 - Intervalo de cores utilizadas para classificação das árvores	78
Figura 16 - Árvore da intervenção classificada de acordo com o desempenho no “pós-teste imediato” e com “ramos” selecionados	79
Figura 17 - Árvores classificadas de acordo com os atributos daTabela 9.....	82
Figura 18 - Árvores com “ramos” selecionados.....	84
Figura 19 - Exemplo de “ramos” classificados - escola Escola A, tipo de Intervenção Adaptive.....	90
Figura 20 - Comparação entre árvores classificadas	91
Figura 21 - Gráficos do índice de confirmação das hipóteses sobre o comportamento dos alunos na escola Escola A em cada um dos tipos de intervenção	94
Figura 22 - Gráficos do índice de confirmação das hipóteses sobre o comportamento dos alunos na escola Escola B em cada um dos tipos de intervenção.....	94
Figura 23 - Gráficos do índice de confirmação das hipóteses sobre o comportamento dos alunos na escola Escola C em cada um dos tipos de intervenção.....	95

Figura 24 - Comportamentos dos alunos das três escolas de acordo com o tipo de intervenção	96
Figura 25 – Árvores de tempo e interação na intervenção “PS” nas três escolas.....	97
Figura 26 - Gráfico Comportamentos geral.....	98

Lista de Tabelas

Tabela 1 - Técnicas de EDM	28
Tabela 2 - Alunos por escola e tipo de intervenção.....	46
Tabela 3 - Atributos da planilha de log	47
Tabela 4 - <i>String</i> de busca em inglês.....	55
Tabela 5 - Quantidade de estudos retornados por cada base de dados eletrônica, total de estudos candidatos e seleção final.	56
Tabela 6 - Classes dos Atributos e suas respectivas Descrições	57
Tabela 7 - Classes definidas para os Objetivos e suas respectivas descrições	59
Tabela 8 - Atributos selecionados para a fase de “mineração de dados”	69
Tabela 9 - Atributos selecionados para interpretação do conhecimento	69
Tabela 10 - Atributos mais utilizados nas pesquisas de EDM	70
Tabela 11 – Fração dos dados da planilha de desempenho dos alunos nos testes.....	75
Tabela 12 - Exemplo de cálculo de valores para classificação dos alunos.....	77
Tabela 13 - Recorte da planilha com as informações dos alunos para cada um dos atributos utilizados na fase de interpretação e avaliação	81
Tabela 14 - Cálculo da fórmula para classificar o “ramo” 1 da Figura 16.....	86
Tabela 15 - Cálculo da fórmula para a cor roxa na classificação do “ramo” 2 da Figura 16...	87
Tabela 16 - Cálculo da fórmula para a cor verde na classificação do “ramo” 2 da Figura 16.	88
Tabela 17 - Cálculo da fórmula para classificação do “ramo” 3 da Figura 16.....	89
Tabela 18 - Hipóteses sobre os comportamentos apresentados pelos alunos.....	92

1. Introdução

Aprimorar produtos e serviços e fornecer aos seus consumidores melhores alternativas são os principais objetivos da maioria das grandes empresas. Assim, visando entender o comportamento e necessidades desses consumidores e encontrar alternativas mais inteligentes para oferecer produtos/serviços personalizados, empresas como Google¹, Microsoft², HTC³ e Nokia⁴ armazenam, por meio de seus sistemas digitais, alguns dados referentes aos seus usuários (Lemos A. e Josgrilberg, 2009).

Entretanto, para que os dados armazenados revelem algum conhecimento relevante sobre os usuários e para que possam ser úteis para realizar qualquer tipo de melhoria nos produtos/serviços oferecidos, é importante que os dados armazenados sejam analisados adequadamente. Com uma análise adequada é possível obter resultados valiosos, gerando conhecimento relevante sobre, por exemplo: quais os produtos mais vendidos, em quais dias da semana o site é mais acessado, qual o horário de maior movimento, quais alunos possuem mais dificuldade em determinado assunto/disciplina, etc. (Costa et al., 2012). No entanto, essa análise não é trivial. Para que, de um conjunto de dados armazenados sejam extraídas as informações relevantes, é necessária a aplicação de técnicas de Mineração de Dados (DM) que, com o apoio de ferramentas específicas, possibilitam a análise e favorecem a compreensão de uma grande quantidade de dados em um tempo hábil.

No contexto da educação, a criação e adoção de diversas tecnologias que apoiam a educação à distância (EAD) tem viabilizado o armazenamento digital de grandes quantidades de dados relacionados ao processo de ensino-aprendizagem (Kampff, 2009). Como exemplo, podem ser citados os Ambientes Virtuais de Aprendizagem (AVA) (Kumr, Rajpakala, Ragade, 1998) e os Sistemas Tutores Inteligentes (ITS, do termo em inglês: *Intelligent Tutoring Systems*) (Chakraborty et al., 2010). Essas tecnologias trouxeram diversas funcionalidades que apoiam tanto o ensino presencial quanto o semi-presencial e à distância, permitindo ao professor propor atividades extraclasse aos seus alunos por meio de ferramentas como: i) fóruns de discussão, nos quais é possível debater os assuntos discutidos em aula; ii) espaços de compartilhamento de conteúdo para troca de informações e materiais; e iii) murais digitais para disponibilizar lembretes, enunciados e notas, bem como

¹ <http://www.google.com/intl/pt-BR/about/company/>

² <http://www.microsoft.com/pt-br/about/default.aspx>

³ <http://www.htc.com/pt/about/>

⁴ <http://br.networks.nokia.com/no-brasil/sobre-nos>

recebimento de resoluções de exercícios. Nessas ferramentas são armazenadas as informações sobre as interações entre alunos-alunos, alunos-professor e o sistema.

Com essa evolução e com o uso em larga escala dessas tecnologias no ambiente de ensino, fica evidente o acúmulo de grandes quantidades de dados referentes ao contexto educacional. Por exemplo, só no Brasil, em 2011, foram registrados mais de 3,5 milhões de estudantes na modalidade EAD (CIED, 2011). Dessa forma, analisar e entender esses dados educacionais pode ser muito importante para auxiliar professores a conhecerem melhor seus alunos e criarem práticas pedagógicas mais efetivas; para alunos entenderem como podem aumentar seu desempenho, e ainda, para desenvolvedores de tecnologias educacionais que podem utilizar esse conhecimento com o objetivo de criar ambientes mais inteligentes e personalizados.

Conforme apresentado, a busca de informações sobre o comportamento e desempenho dos alunos é um campo interessante e crescente de investigação. Neste contexto, surgiu a área de pesquisa conhecida como Mineração de Dados Educacionais (do inglês *Educational Data Mining* – EDM), a qual trabalha com o desenvolvimento de métodos específicos para explorar dados armazenados em sistemas educacionais. Assim, se tornam mais eficazes tarefas como a de entender como os alunos aprendem quando interagem com sistemas educacionais *Web*, se os alunos estão entediados ou até mesmo se os alunos estão burlando o sistema (Baker; Isotani e Carvalho 2011).

No entanto, encontrar os dados mais adequados para descobrir informações relevantes sobre os alunos ainda é uma tarefa muito difícil e importante para a EDM. Diversos pesquisadores têm realizado esforços nessa área, visando descobrir como maximizar a EDM (Gottardo; Kaestner e Noronha; 2012, Rigo et al., 2012, Manhães et al. 2011, etc.) utilizando os atributos mais relevantes e dessa forma prever determinados comportamentos dos alunos estudados.

Na próxima seção são apresentadas a problemática, que envolve algumas das dificuldades encontradas pelos professores que utilizam as tecnologias digitais no processo de ensino e aprendizagem, e a justificativa para a realização dessa pesquisa.

1.1.Problemática e Justificativa

Diante da crescente inclusão de recursos tecnológicos na educação, recursos estes que vão desde simples softwares educativos até os mais elaborados AVAs e ITS, que ampliam o

acesso e as formas de interação, além de gerar e armazenar uma enorme quantidade de dados, a busca por informações relevantes para melhorar e apoiar o processo de ensino-aprendizagem torna-se uma tarefa extremamente necessária e importante (Baker, Isotani e Carvalho, 2011; Gotardo, Cereda e Junior, 2013).

Analisando tais dados é possível reconhecer o comportamento dos alunos e descobrir, por exemplo, porque determinados alunos abandonaram o curso (Martíns, Lopes e Raabe, 2012), quando os alunos terão um mau desempenho (Gotardo, Kaestner, Noronha, 2012) e qual o seu estilo cognitivo (Flores et al., 2013).

No entanto, como os professores encontram dificuldades em interpretar as informações obtidas a partir desse grande volume de dados, cria-se um obstáculo em identificar em tempo hábil os comportamentos dos alunos e suas necessidades na realização de atividades educacionais. Deste modo, devido à necessidade de identificar tais comportamentos a tempo de o professor promover melhorias na metodologia do ensino e de analisar possíveis dificuldades e/ou facilidades na aprendizagem de seus alunos, a busca de informações sobre o comportamento e o desempenho dos alunos que interagem com um sistema educacional é atualmente um campo interessante e crescente de investigação.

É o caso do experimento realizado por McLaren et al. (2012) que envolveu três escolas do ensino médio e fundamental dos EUA, no ano de 2012. Durante esse experimento os alunos interagiram com um sistema tutor inteligente, chamado MathTutor⁵, e tiveram suas interações armazenadas em banco de dados e arquivos de *log*. Para dar valor a essas informações, transformando-as em conhecimento, foi necessário realizar uma análise sobre os dados coletados a fim de identificar as características que definiam algum tipo de padrão comportamental. Uma vez que essa análise não é trivial, é importante que metodologias, técnicas e ferramentas específicas sejam utilizadas, de forma que resultados de fácil interpretação possam ser gerados e sejam úteis para os envolvidos no processo de ensino e aprendizagem.

Além disso, encontrar informações relevantes que maximizem o processo de Mineração de Dados (do inglês *Data Mining* - DM) não é tarefa trivial. Rigo et al., (2012), afirma que em muitos trabalhos, a seleção de dados para a DM não visou maximizar o processo, mas sim a facilidade de acesso a esses dados. Dessa forma, são de grande importância para a EDM os trabalhos que objetivam buscar e compreender quais os melhores dados a serem utilizados na DM.

⁵ <https://adapterrex.cs.cmu.edu/>

Inserido nesse contexto, esse trabalho se justifica e tem como objetivo analisar os dados provenientes da interação dos alunos com um sistema tutor inteligente, o MathTutor, que disponibiliza exercícios específicos de matemática em páginas *Web*, avaliando quais os atributos que melhor auxiliem, a partir da utilização de técnicas de mineração de dados, no reconhecimento de padrões de comportamento que auxiliem no processo de ensino e aprendizagem.

Para atingir esse objetivo, técnicas da EDM foram empregadas visando gerar resultados úteis e de fácil interpretação para auxiliar o professor a caracterizar os alunos. Além disso, buscou-se identificar quais dados são relevantes durante o processo de mineração de dados de modo a maximizar a eficácia dos resultados obtidos durante o processo de mineração dos dados. A combinação de tais procedimentos visou oferecer aos professores e pesquisadores da área, informações relevantes sobre os tipos de comportamentos apresentados pelos alunos quando estes interagem com um sistema tutor inteligente.

Na seção seguinte são apresentadas as questões de pesquisa elaboradas com o intuito de guiar o estudo aqui proposto.

1.2. Questões de Pesquisa

Para orientar o desenvolvimento do presente trabalho, as seguintes questões de pesquisa foram elaboradas:

- (Questão 1) Quais informações (dados) são relevantes para entender as diferenças e semelhanças entre os comportamentos manifestados por alunos do ensino fundamental quando interagem com um sistema tutor inteligente?** Atualmente existem diversos ambientes digitais de apoio ao ensino e aprendizagem, nos quais alunos e professores interagem de forma mais simples e rápida. Na maioria dos casos, essas interações são armazenadas para que, posteriormente, esses dados possam ser estudados e assim, trazer algum tipo de informação para os envolvidos. Porém, entender quais são as informações mais relevantes e qual a melhor forma de armazená-las é de extrema importância para que o estudo desses dados seja realizado com eficiência.
- (Questão 2) Quais comportamentos manifestados por alunos do ensino fundamental podem ser identificados ao analisar dados registrados em um sistema tutor inteligente (MathTutor) durante o processo de**

aprendizagem de Decimais? Ao analisar dados dos alunos que interagiram com um sistema tutor inteligente de um curso de matemática, é importante identificar um método para determinar qual o comportamento do aluno e como ele pode afetar a sua performance ao final de determinado percurso de aprendizagem. Dessa forma, é possível oferecer ao professor indícios de quais alunos necessitam (ou necessitarão) de uma intervenção pedagógica mais personalizada.

A partir das questões discutidas nesta subseção, apresentam-se na seção seguinte os objetivos desta pesquisa.

1.3. Objetivos

O objetivo desta pesquisa é analisar os dados provenientes da interação dos alunos com um sistema tutor inteligente, o MathTutor, que disponibiliza exercícios específicos de matemática em páginas *Web*, utilizando técnicas de mineração de dados, para reconhecer padrões de comportamento que possam auxiliar no processo de ensino e aprendizagem.

A partir desse objetivo, definem-se dois objetivos específicos:

O1 → Encontrar os atributos mais importantes para maximizar a fase de “mineração dos dados” e, posteriormente, minerar os dados selecionados, utilizando técnicas de agrupamento.

O2 → Elaborar e validar hipóteses que infiram comportamentos adotados pelos alunos ao final do percurso de aprendizagem, com base nos dados coletados no início do percurso.

Na próxima seção é apresentada a organização desta dissertação com uma breve descrição sobre cada capítulo.

1.4. Organização

Esta dissertação está organizada da seguinte forma: no capítulo 1 foi introduzida a contextualização desta pesquisa, no qual se discorreu mais especificamente sobre a aplicação da mineração de dados em ambientes educacionais, sua problemática atual e as questões e objetivos desta pesquisa.

O capítulo 2 trata da fundamentação teórica da pesquisa, com a apresentação dos métodos e técnicas conhecidos para realizar a EDM. No capítulo 3 são tratados alguns trabalhos relacionados com o tema dessa pesquisa. No capítulo 4 é definido o escopo da pesquisa, a partir dos dados utilizados e da ferramenta utilizada para trabalhar esses dados. No capítulo 5 são exibidos como foram selecionados os dados utilizados nesta pesquisa, bem como um mapeamento sistemático desenvolvido para apoiar esta fase, no âmbito de maximizar a EDM. No capítulo 6 são exibidos os resultados alcançados em cada uma das fases realizadas durante todo o processo de EDM. Por fim, no capítulo 7 são apresentadas as conclusões, as contribuições da pesquisa e a sugestão de trabalhos futuros que poderão ser realizados para melhorar os resultados já obtidos.

2. Fundamentação Teórica

Neste capítulo são apresentadas as técnicas que embasaram este estudo, oferecendo, dessa forma, um referencial teórico sólido para que este trabalho fosse realizado com sucesso. O capítulo está subdividido em duas seções: na seção 2.1 é apresentada a técnica de descoberta de conhecimento em bases de dados e na seção 2.2 é discutida a área de pesquisa de mineração de dados educacionais.

2.1. Descoberta de Conhecimento em Bases de Dados

A Mineração de dados (DM) é um processo utilizado para extrair informações úteis de grandes volumes de dados, por meio da combinação de técnicas estatísticas, de banco de dados e de inteligência artificial (Figura 1). Cabe a ela descobrir as correlações, padrões frequentes, associações ou estruturas casuais entre as informações armazenadas nos repositórios de dados. O resultado de todo esse processo é chamado de “conhecimento”. (Abdullaha et al., 2011).

Figura 1 - Mineração de Dados adaptado de Silva, 2004.

Segundo Baykal (2006), a DM é uma das fases do processo de Descoberta de Conhecimentos em Bases de Dados (do inglês, *knowledge-discovery in databases* - KDD).

Essas fases visam primeiramente selecionar os dados (“seleção”) mais significativos para descobrir algum conhecimento desejado. Em seguida, realiza-se a fase de “pré-processamento” com o objetivo de adequar dados que estejam inconsistentes, além de reduzir

o conjunto de dados a serem minerados. Na sequência, a fase de “transformação” faz o tratamento e formatação dos dados de maneira que possam ser utilizados em algoritmos e ferramenta de DM. Finalmente, esses dados são minerados (“mineração de dados”) gerando padrões que necessitam ser estudados e interpretados (“interpretação e avaliação”) para que o novo conhecimento seja explicitamente identificado. As fases para a descoberta do conhecimento estão representadas na Figura 2, adaptada de Fayyad, Piatetsky-Shapiro e Smyth et al. (1996), e serão detalhadas a seguir.

Figura 2 - Fases do Descobrimento de Conhecimento em Bases de Dados (adaptado de Fayyad, Piatetsky-Shapiro E Smyth, 1996).

2.1.1. Seleção

A fase de “seleção” de dados é a primeira no processo de descobrimento de informação e possui impacto significativo sobre a qualidade do resultado final, uma vez que nessa fase é escolhido o conjunto de dados contendo todas as possíveis “variáveis” (também chamadas de características ou atributos) e “registros” (também chamados de casos ou observações) que farão parte da análise (Han e Kamber, 2006). Normalmente a escolha dos

dados fica a critério de um especialista de domínio, ou seja, alguém que possui conhecimentos significativos sobre o assunto em questão.

O processo de “seleção” é bastante complexo, uma vez que os dados podem vir de uma série de fontes diferentes (*datawarehouses*, planilhas, sistemas legados e sensores) e podem possuir os mais diversos formatos.

2.1.2. Pré-processamento de dados

O “pré-processamento” de dados consiste em avaliar e possivelmente alterar os dados, visando tornar a fase de “mineração de dados” mais eficiente. Assim, antes de aplicar o algoritmo de mineração de dados, diversas tarefas de pré-processamento devem ser realizadas. De acordo com Koutri (2004) as três tarefas mais importantes do “pré-processamento” são:

- Limpeza de dados: remove itens irrelevantes e entradas de registro desnecessárias para o processo de mineração;
- Transformação dos dados e enriquecimento: calcula novos atributos baseados nos já existentes, convertendo atributos numéricos em atributos nominais, fornecendo significado para referências contidas no *log*, etc.;
- Redução dos dados: realiza a redução da dimensionalidade dos dados, ou seja, retira aqueles atributos que não são importantes para a DM.

2.1.3. Transformação

A “transformação” dos dados é a fase da extração do conhecimento que antecede a fase de “mineração de dados”. Após serem selecionados, limpos e processados, os dados precisam ser armazenados e formatados adequadamente para que os algoritmos possam ser aplicados. Essa fase é muito importante pois sem ela a maioria das ferramentas de DM não teria a capacidade de trabalhar com os dados disponíveis.

Um bom exemplo disso são as grandes corporações nas quais é comum encontrar computadores executando diferentes sistemas operacionais e diferentes sistemas gerenciadores de bancos de dados (SGDB). Os dados desejados podem estar dispersos e, portanto, pode ser necessário o agrupamento dos mesmos em um repositório único ou marcados de alguma forma para que, posteriormente, sejam minerados adequadamente. Com essa grande diversidade, é muito importante que os dados sejam formatados para o padrão específico da ferramenta de DM (Han e Kamber, 2006).

2.1.4. Mineração de Dados

Existem diversas definições que tentam explicar o conceito de mineração de dados. Cabena et al. (1998) definem DM como a somatória de técnicas de diferentes áreas, como conhecimento de máquinas, reconhecimento de padrões, estatísticas, banco de dados e visualização, sendo assim um campo interdisciplinar que visa extrair informações de grandes bases de dados. Han e Kamber (2006, p. 8) afirmam que "A mineração de dados é um processo essencial no qual são aplicados métodos inteligentes para extrair padrões dos dados".

Com isso, pode-se notar que apesar das diferentes definições, em geral, a DM visa estudar grandes quantidades de dados utilizando técnicas de diversas áreas para encontrar relações entre dados e que, após algumas observações, produzam novos conhecimentos.

2.1.5. Interpretação e Avaliação

A última fase do processo de descoberta de conhecimento é a “interpretação e avaliação” das informações obtidas na fase de DM. Por meio de análises e interpretações dos resultados gerados, procura-se criar artefatos que facilitarão o entendimento dos resultados. Tais artefatos podem vir na forma de gráficos, tabelas, árvores, etc. Dessa forma, esta fase tem como principal objetivo gerar algum tipo de conhecimento “novo” para o usuário final. Esse conhecimento pode então ser utilizado de forma a aperfeiçoar processos, melhorar sistemas, encontrar padrões, entre outros. Com os avanços nos estudos da área de descoberta de conhecimento em bases de dados, algumas particularidades foram notadas em algumas áreas como a área da educação. Segundo Baker (2010), uma dessas particularidades, por exemplo, se baseia na diferença entre os dados comumente utilizados na DM e os dados utilizados na EDM, uma vez que os dados educacionais seguem uma hierarquia multi-nível e não possuem independência entre si.

Assim, para gerar conhecimento de qualidade no âmbito educacional, foi criada a área de pesquisa em mineração de dados educacionais (*Educational Data Mining* - EDM). Na seção 2.2.1 serão abordadas mais especificamente as particularidades encontradas no âmbito educacional.

2.2. A área de Pesquisa em Mineração de Dados Educacionais

A Mineração de Dados Educacionais (EDM) é definida por Baker e Yacef (2009, p. 4) como sendo “[...] uma das áreas de pesquisa emergentes e interdisciplinares no contexto educacional [...]”. Ela pode, ainda, ser definida como a aplicação de técnicas de mineração de dados para dados educacionais (Romero et al., 2008). A principal preocupação da EDM é desenvolver métodos que possam explorar as informações extraídas de ambientes educacionais para entender melhor os estudantes e os contextos nos quais eles aprendem (Abdullaha et al., 2011).

A EDM surgiu oficialmente no ano de 2005, em Pittsburgh, EUA, onde foi organizado o primeiro *workshop* sobre o tema (*Workshop on Educational Data Mining*) como parte do *20th National Conference on Artificial Intelligence* (AAAI 2005) (Costa et al. 2012). Antes disso, a EDM era debatida em conferências sobre Inteligência Artificial na Educação (AIED) e Sistemas de Tutoria Inteligentes (ITS).

Para analisar a efetividade das pesquisas e publicações em EDM, Romero e Ventura (2007) realizaram um estudo no qual mostram uma lista completa de trabalhos sobre os assuntos publicados entre 1995 e 2005. Para determinar a influência desses trabalhos é calculado o número de citações que cada artigo recebeu por meio do Google Scholar (GS) que, como Bartneck e Hu (2009) notaram, apesar de possuir imperfeições, é considerado como a fonte mais completa de citações na atualidade. De acordo com essa contagem, Romero e Ventura (2007) afirmam que esses artigos publicados entre 1995 e 2005 têm sido altamente influentes tanto na área de EDM quanto em áreas afins (Ha, Bae e Park, 2000; Hamalainen, Suhonen, Sutinen e Toivonen, 2004; Sanjeev e Zytchow, 1995; Tang e McCalla, 2002; Zai'ane e Luo, 2001).

Com essa alta influência, resultados de diversas pesquisas foram publicados e utilizados por outros pesquisadores. Dentre os fatores que afetaram a área de EDM estão: (1) a abertura de repositórios de dados educacionais públicos, como o *Datashop PSLC* (Koedinger Cunningham e Leber, 2008), que fez com que grandes quantidades de dados de uma variedade de ambientes de aprendizagem estivessem disponíveis *online*, gratuitamente, para qualquer pesquisador; e (2) a maior disseminação e utilização de ambientes gerenciadores de cursos como o Moodle⁶, WEBCAT⁷ e EDx⁸, que armazenam diversas

⁶ <http://moodle.com/>

⁷ <http://web-cat.org/>

⁸ <https://www.edx.org/>

informações sobre os alunos e suas interações com o sistema em grandes bancos de dados (Baker e Yacef, 2009).

Atualmente, muitos professores utilizam esses ambientes gerenciadores de cursos tanto para tarefas simples, como enviar resultados das avaliações e avisos aos alunos, quanto para tarefas complexas, como a aplicação de questionários, discussão em fóruns, desenvolvimento de atividades colaborativas em *wikis*, entre outros. Porém, somente um grupo reduzido de docentes utiliza as informações armazenadas por essas ferramentas para gerar algum tipo de conhecimento.

Na próxima seção são apresentadas algumas técnicas utilizadas para minerar dados educacionais e gerar conhecimento.

2.2.1. Técnicas utilizadas em Mineração de Dados Educacionais

A mineração de dados é um recurso aplicado com sucesso em *e-commerce* já há algum tempo (Srivastava, 2000). Pode-se perceber esses avanços ao simplesmente navegar pela *Web* e notar que, ao realizar qualquer tipo de busca por algum conteúdo, diversas propagandas relacionadas a essa busca são lançadas em nossa tela. Em outras áreas, como na educação (*e-learning*), resultados promissores estão sendo obtidos. Segundo Romero e Ventura (2007), as principais diferenças entre as duas áreas de aplicação são:

- Domínio: no *e-commerce*, o objetivo é orientar as compras dos clientes. Já em *e-learning* o objetivo é orientar a aprendizagem dos alunos;
- Dados: no comércio eletrônico, os dados utilizados são geralmente *logs* de acesso ao servidor. Em *e-learning*, os dados são mais complexos, pois tratam de informações sobre as interações do aluno com o sistema, ou seja, trata de comportamentos humanos, o que pode estar atrelado a uma grande gama de fatores;
- Objetivo: no *e-commerce* tem-se um único objetivo: aumentar o lucro da empresa, e essa é uma variável que pode ser medida de maneira relativamente simples. No *e-learning* tem-se como objetivo melhorar o processo de ensino e aprendizagem de alunos e professores, algo que é muito subjetivo e difícil de ser medido; e,
- Técnicas: ao trabalhar com ensino, aprendizado e comportamento, depara-se com dados específicos que necessitam ser analisados com mais cautela, pois representam interações de alunos com um sistema educacional, e não somente quantidades e medidas, como no *e-commerce*.

Além disso, como afirma Baker (2010), para a avaliação da utilização de determinado sistema educacional pelos alunos por meio da mineração de dados educacionais, pode ser interessante considerar simultaneamente os dados referentes a diferentes níveis de interação dos alunos: interação do aluno com o teclado, com a resposta, com a sessão, com o nível do aluno, com o nível de sala de aula, e com o nível da escola. Além disso, os dados são bastantes dependentes uns dos outros, o que implica na utilização de novas técnicas estatísticas ou pelo menos a adequação das já utilizadas na DM.

Portanto, ainda que os resultados da aplicação da DM no âmbito do *e-commerce* e de outras importantes áreas possam ser considerados relevantes para os contextos educativos, é necessário considerar o desenvolvimento de novas técnicas ou a adaptação das técnicas já existentes. Essas técnicas são apresentadas na Tabela 1, conforme sua categorização nas subáreas de EDM, seguindo o que consta na taxonomia proposta por Baker e Yacef (2009):

Tabela 1 - Técnicas de EDM

TÉCNICA	MÉTODOS
1. Predição	Classificação Regressão
2. Agrupamento	
3. Mineração de Relações	Mineração de Regras de Associação Mineração de Correlações Mineração de Padrões Sequenciais Mineração de Causas
4. Destilação de dados para facilitar decisões humanas	
5. Descobertas com modelos	

A seguir, cada uma das técnicas é apresentada detalhadamente.

2.2.1.1. Predição

Na predição, o objetivo principal é desenvolver modelos que consigam retirar uma perspectiva específica dos dados. Assim, desenvolver esses modelos, de forma que encontrem aspectos específicos nos dados (variáveis preditivas) por meio de análises e fusões de diversos

aspectos encontrados nos dados (variáveis preditoras) é a principal tarefa dessa técnica. Para se conseguir utilizar a predição é necessário que boa parte dos dados seja codificada manualmente para que, assim, a correta identificação de uma ou mais variáveis ocorra (Costa et al., 2012).

Baker; Isotani e Carvalho (2011, p. 5), afirmam que existem duas vantagens em utilizar a predição:

[...] Primeiro, métodos de predição são utilizados para estudar quais aspectos de um modelo são importantes para predição, dando informação sobre o construto sendo examinado (exemplos de constructo modelado incluem curvas de aprendizagem e representações de tipos variados de comportamento). Segundo, os métodos de predição auxiliam a predizer o valor das variáveis utilizadas em um modelo.

Na EDM, normalmente são utilizados dois tipos de técnicas de predição, a classificação e a regressão. A classificação é uma técnica que utiliza a variável preditiva (valor será predito baseado nos valores de outros atributos) binária ou categórica. Os algoritmos mais comuns utilizados na classificação são: árvores de decisão, regressão logística e regressão *step*. Já quando a variável é numérica, são utilizados os algoritmos de regressão. Dentre os mais utilizados estão: regressão linear, redes neurais e máquinas de suporte vetorial (Baker; Isotani e Carvalho, 2011).

Tanto na classificação quanto na regressão, as variáveis preditoras (atributos usados para fazer a predição da variável preditiva) podem assumir valores categóricos (a presença de asas em uma determinada espécie: SIM ou NÃO) ou numéricos (quantidade de insetos em uma determinada espécie). Porém, os métodos utilizados terão diferentes eficiências dependendo do comportamento das variáveis preditoras (independentes) (Costa et al. 2012).

2.2.1.2. Agrupamento

De acordo com (Costa et al. , 2012, p. 11), o objetivo principal de um algoritmo de agrupamento é

“[...] achar dados que se agrupam naturalmente, classificando-os em diferentes grupos e/ou categorias. Esses grupos e categorias não são conhecidos inicialmente. Através de técnicas de agrupamento os grupos/categorias são automaticamente identificados através da manipulação

das características dos dados. É possível criar esses grupos/categorias utilizando diferentes unidades de análise”.

Fica, então, como principal tarefa para esse tipo de técnica, trabalhar com dados que não possuem nenhum tipo de classificação ou que são pouco conhecidos.

2.2.1.3. Mineração de Relações

Na mineração de relações, a principal ideia é encontrar relações entre as muitas variáveis de um banco de dados. Isto comumente é feito descobrindo quais as variáveis estão conectadas mais fortemente. Existem quatro tipos de mineração de relações: regras de associação, correlações, padrões sequenciais e causas (Costa et al., 2012).

2.2.1.4. Destilação de dados para facilitar decisões humanas

Essa abordagem tem como objetivo exibir os dados de forma mais simples e de fácil compreensão auxiliando, assim, a tomada de decisões. Sua principal técnica é a “visualização da informação”. Porém, os métodos utilizados na EDM são diferentes dos comumente usados na área de “visualização de informação” (HersHKovitz e Nachmias 2008; Kay et al., 2006). Algumas dessas particularidades são destacadas por Baker (2010):

- A organização dos dados é feita através da estrutura do material de aprendizagem e da estrutura do contexto de aprendizagem; e,
- Os dados são refinados para facilitar decisões humanas em dois pontos principais: identificação, no qual os dados são apresentados de forma que humanos possam identificar os padrões mais facilmente, e classificação, no qual a destilação de dados pode ser usada também para apoiar modelos de predição. Neste caso, parte dos dados é exibida para ser rotulada por humanos.

2.2.1.5. Descobertas com modelos

A “descoberta com modelos” inicia-se através de um modelo inferido por um método de predição, de agrupamento ou ainda, manualmente. A única condição é que esse modelo possa ser validado de alguma forma. Após isso, o modelo descoberto é utilizado como um componente em outra análise (Costa et al. 2012).

A “descoberta com modelos” tornou-se um método bastante popular em pesquisas de EDM, apoiando diversas análises como a identificação de *outliers*⁹ através de previsões do modelo, o exame de quais variáveis melhor prevê a construção de um modelo, etc.

O primeiro ponto antes de uma descoberta com análise do modelo, é ter certeza que o modelo adotado é válido tanto em termos globais quanto para a situação específica em que será empregado. O ideal é realizar uma validação cruzada, na qual o modelo é repetidamente treinado em uma parte dos dados e testado em uma parte diferente (Hershkovitz et al., 2013).

Algumas vantagens da utilização da “descoberta com modelos” são apontadas por Hershkovitz et al. (2013). Primeiramente, essa técnica permite que a investigação científica use definições que podem ser discutidas e inspecionadas, enquanto os modelos utilizados desta forma são publicados e discutidos entre os pesquisadores, mesmo após a técnica já ter sido utilizada. Em segundo lugar, a descoberta com os modelos deixa trilhas de dados claros que podem ser inspecionados mais tarde, o que facilita a compreensão do método utilizado para os que não estão familiarizados com a técnica. Em terceiro lugar, algumas construções que são difíceis para humanos rotularem à mão (como o conhecimento latente do estudante) podem ser modeladas e depois estudadas. Em quarto lugar, a descoberta com modelos permite que as análises ocorram em escalas o que de outra maneira seria inviável.

Nesse trabalho a técnica utilizada foi a de agrupamento. Através da análise dos dados selecionados foram gerados agrupamentos para os indivíduos (alunos), de acordo com suas características e, dessa forma, alguns algoritmos foram utilizados para avaliar essas características e gerar esses agrupamentos de forma eficiente. A aplicação dessa técnica será apresentada com mais detalhes na seção 4.3.1.

No capítulo seguinte serão apresentados alguns trabalhos relacionados à EDM nos quais são descritas formas de caracterização dos alunos que interagiram com sistemas computacionais educacionais.

⁹ Indivíduos que apresentam um grande afastamento dos demais da série

3. **Trabalhos Relacionados**

A EDM tem se consolidado como uma potencial linha de pesquisa para a melhoria da qualidade do ensino. No Brasil, ainda que haja muitos esforços dos pesquisadores para expandir e enriquecer esta área, ela é ainda pouco explorada. No entanto, com a crescente expansão de cursos à distância e de outras tecnologias de suporte ao ensino e aprendizagem, já é grande o volume de dados educacionais que, explorados corretamente e com qualidade, podem gerar informações valiosas às instituições de ensino. Assim, a tendência é que a EDM ganhe cada vez mais relevância em ambientes educacionais e, deste modo, enriqueça a qualidade do ensino em ambientes EAD e em escolas, de forma a ajudar o Brasil a ocupar lugar de destaque no cenário educacional mundial (Costa et al. 2012).

Em termos internacionais, países como Estados Unidos, Canadá e Espanha lideram o número de pesquisas em EDM, uma vez que cada vez mais professores, escolas e universidades têm utilizado *softwares* educacionais. Além disso, nos Estados Unidos, por exemplo, estão sendo disponibilizados aos pesquisadores bancos de dados públicos, permitindo que se obtenha com facilidade dados relevantes para futuras análises a partir de técnicas vindas da área de EDM, além de reduzir tempo e custos envolvidos com pesquisas nesta área (Baker; Isotani e Carvalho 2011).

Deste modo, diversos estudos têm utilizado a EDM em conjunto com *softwares* educacionais para caracterizar os alunos de acordo com dados de suas interações com sistemas *Web* educacionais. Assim, é possível apontar aos professores quando comportamentos inadequados ocorrem durante as interações, quais as potenciais facilidades e dificuldades dos alunos em diferentes assuntos e, portanto, oferecer a possibilidade de um suporte adequado para atender às necessidades de cada aluno.

Outra vertente que tem como suporte a EDM são os Sistemas Tutores Inteligentes (do inglês *Intelligent Tutoring Systems* - ITS), caracterizados por apresentarem comportamentos inteligentes que oferecem suporte e *feedback* aos alunos durante a resolução de exercícios em determinado ambiente de ensino computacional. Assim, de acordo com determinados comportamentos apresentados pelos alunos durante a realização de atividades educacionais, os ITS visam classificar os alunos e dessa forma oferecer a cada um o suporte necessário para que se possa obter o máximo de conhecimento durante tais atividades. Esse suporte aos alunos provém de um tutor que, animado ou não, dá dicas e conselhos aos alunos durante as atividades educacionais, fazendo com que o aluno trabalhe mais a fundo os temas nos quais

possui maiores dificuldades de aprendizagem, além de evitar que ele se distraia com assuntos fora de sua tarefa (*off task behavior*).

Na pesquisa de Baker et al. (2006), os autores verificaram que os alunos que interagem com sistemas tutores inteligentes tendem a apresentar um comportamento conhecido como “*gaming the system*”, no qual as ferramentas oferecidas pelo sistema, como dicas para resolução de exercícios, são utilizadas pelos alunos com o intuito de responder mais rapidamente a todos os exercícios ao invés de serem utilizadas de forma a agregar conhecimento. Um exemplo de “*gaming the system*” se mostra quando alunos pedem ajuda ao sistema diversas vezes para tentar obter a resposta final de um exercício sem ter tentado resolvê-lo de fato. Para contornar esse tipo de comportamento e garantir que o aluno obtenha aprendizado durante suas atividades, introduziu-se um tutor que oferece exercícios focados exatamente no assunto que determinado aluno não priorizou durante sua interação com o sistema, além de expressar visão negativa sobre alunos que apresentam o comportamento “*gaming the system*”. Para isso, foi criado um agente animado chamado "Scooter, o Tutor", de forma a incorporar e apresentar aos alunos, por meios visuais, uma gama mais ampla de emoções.

A projeção do Scooter objetivou reduzir o incentivo ao comportamento “*gaming the system*” e fortalecer assuntos nos quais os alunos adotaram tal comportamento, oferecendo uma nova oportunidade de aprendizado com qualidade. Além do mais, buscou-se afetar minimamente alunos que não adotavam tais comportamentos prejudiciais.

Para verificar a influência do agente tutor animado no ambiente computacional, Baker et al. (2006) realizaram um experimento utilizando um ITS aplicado ao ensino de matemática do ensino médio em duas escolas de cinco classes no subúrbio de Pittsburgh, cujos alunos já haviam interagido por meses com tal sistema tutor de modo a estarem familiarizados com ele. No experimento, foram apresentadas a cada aluno duas versões do sistema: uma com o Scooter (condição experimental) e outra sem (controle), sendo que cada uma das versões abordava assuntos didáticos diferentes. Dentre os 51 alunos que participaram da condição experimental, 12 eram isentos de pré ou pós-teste e, dentre os mesmos que participaram da condição controle, 17 eram isentos de pré ou pós-teste.

Como resultados desse experimento, a presença do Scooter na condição experimental foi associada a uma considerável, embora apenas marginalmente significativa, redução da frequência do comportamento “*gaming the system*” observado: 33% dos alunos na condição controle apresentaram o comportamento “*gaming the system*” contra 18% de alunos que

interagiram com o ambiente computacional na presença do Scooter. No entanto, verificou-se que esses 18% não burlavam menos o sistema do que os 33% na condição controle.

Além disso, foi visto que os alunos efetivamente aprendiam quando lhes eram passados exercícios suplementares, mas ainda assim esses alunos não deixavam de burlar o sistema. Do mesmo modo, a expressão de desagrado apresentada pelo Scooter não contribuiu para uma mais efetiva aprendizagem tão pouco para reduzir o comportamento "*gaming the system*".

É possível concluir, portanto, que, ainda que pouca, houve redução de "*gaming the system*", fazendo surgir a hipótese de que o fato de burlar o sistema resultar em exercícios complementares passados pelo Scooter tenha feito os alunos evitarem adotar esse tipo de comportamento.

Outro trabalho que utilizou a caracterização de alunos em tutores inteligentes foi realizado por Baker et al. (2004). Este trabalho apresenta o quanto os alunos são prejudicados em termos educacionais quando adotam comportamentos que os distraem das tarefas de aprendizagem ou que os permitem burlar o sistema de alguma forma. Para isso, realizou-se um experimento em duas escolas nos subúrbios de uma cidade de médio porte no Nordeste dos Estados Unidos com um conjunto de cinco salas de aula do ensino médio com alunos de faixa etária entre 12 a 14 anos. Durante a condução desse experimento, os alunos assistiam a uma aula teórica em um dia e no dia seguinte eram submetidos a atividades guiadas por ITS e relacionadas à disciplina de matemática.

Analisando os dados, verificou-se que, em geral, o tutor foi bem sucedido. Os resultados mostram que, no "pré-teste", os alunos obtiveram 40% de êxito em solucionar exercícios corretamente, enquanto que, no pós-teste, apresentaram 71% de êxito, possibilitando que esta melhora significativa seja associada ao fato de que os alunos realizaram com menor frequência atividades fora de sua tarefa educacional.

Carmona et al. (2010), em sua pesquisa, analisou dados provenientes de interações de alunos com o Moodle para apresentar resultados aos professores em formas de regras, de modo a permitir que tais regras auxiliassem os professores na tomada de decisões sobre as atividades complementares do curso.

Para obtenção desses resultados, foram avaliados cinco cursos com alunos que apresentavam maior uso das atividades e recursos de um sistema AVA com o qual interagiram, resultando em um total de 293 alunos. A partir de informações provenientes da interação destes alunos com o sistema, foram avaliados os seguintes atributos: identificação do curso; número de trabalhos concluídos; número de atribuições nas quais o aluno passou;

número de atribuições nas quais o aluno falhou; número de *quizzes* em que passou; número de *quizzes* em que falhou; número de mensagens enviadas pelo bate-papo; número de mensagens enviadas para o professor; número de mensagens enviadas para o fórum; número de mensagens lidas no fórum e média do aluno.

A avaliação desses atributos teve por objetivo permitir que os pesquisadores conseguissem analisar uma possível relação entre o uso de atividades complementares oferecidas em um curso e as notas finais obtidas pelos alunos. Deste modo, essa pesquisa busca auxiliar o professor nas tarefas de ensino criando regras para o uso ou não de atividades de apoio de forma a aprimorar o ensino e aprendizado dos alunos dentro da sala de aula. Tais regras foram criadas a partir da utilização de dois algoritmos evolucionários, SDIGA e MESDIF.

De forma a validar os resultados encontrados na pesquisa a partir da utilização dos algoritmos evolucionários, Carmona et al. (2010), os comparou com resultados encontrados a partir do uso de algoritmos tradicionais (Apriori-SD e CN2-SD). Tal comparação permitiu concluir que o uso de algoritmos evolucionários permitiu a descoberta de um conjunto reduzido de regras, tornando-as de fácil interpretação do professor, garantindo mais qualidade de ensino e suporte aos alunos.

Já em uma pesquisa realizada por Kampff (2009), buscou-se estudar mais a fundo ambientes de educação à distância (EAD) de modo a criar melhorias e alternativas para se evitar que alunos reprovem e/ou abandonem o curso. Kampff (2009) acredita que muitos professores não estão preparados de forma adequada para realizar as atividades de ensino e aprendizagem e, mesmo os que estão preparados, encontram dificuldade para realizar o acompanhamento de seus alunos. Para facilitar esse acompanhamento, foram criados e avaliados alertas automatizados para informar os professores sobre determinados comportamentos relevantes.

Esse tipo de comunicação teve por objetivo facilitar o entendimento do professor sobre seus alunos sem que o mesmo necessite avaliar uma grande quantidade de dados em busca de algum padrão. Além disso, como se tratam de alertas objetivos que não exigem grandes esforços para a compreensão de qual é a situação do aluno, são trazidas ao professor novas e úteis informações que podem otimizar o tempo gasto para entender de forma mais completa seus alunos e, dessa forma, realizar intervenções no processo de ensino e aprendizagem.

Para validar se realmente seria possível gerar tais alertas e se a intervenção dos professores auxiliados por esse tipo de comunicação seria de fato funcional, realizou-se um

experimento na Universidade Luterana do Brasil no qual foram avaliados dados de 161 alunos que utilizavam um AVA em um curso de EAD.

Para avaliação desses dados, aplicou-se o processo de KDD e, com isso, foi possível gerar indicadores sobre o comportamento desses alunos em relação à reprovação ou evasão do curso. Como próximo passo à obtenção das regras de classificação, 230 alunos foram observados durante um semestre letivo. Ao reunir, inicialmente, os dados obtidos, cada aluno foi representado por 158 atributos, todos validados, possibilitando o trabalho com uma tabela bastante completa. Na fase de preparação de dados, novos atributos foram criados a partir dos existentes, de forma a tornar as informações mais ricas, contextualizadas e/ou resumidas. A tabela final, enriquecida, alcançou o número de 230 atributos. Para a DM, foram selecionados 87 atributos, pois alguns foram resumidos ou agrupados em novos atributos gerados durante a fase de preparação dos dados.

A observação desses alunos gerava alertas para os professores que, então, realizavam intervenções durante os processos de ensino. Os resultados alcançados com esses alertas comprovaram a melhoria do índice de aprovação de alunos, bem como a redução do índice de evasão na disciplina acompanhada.

Esses resultados nos permitem concluir que essa pesquisa traz interessantes contribuições para a melhoria da qualidade de educação à distância no Brasil, uma vez que ela possibilita identificar, por meio da análise de dados oriundos de um AVA, quais os comportamentos comumente apresentados quando alunos tendem a reprovar ou abandonar o curso, além de disponibilizar essa informação ao professor por meio de alertas, permitindo que ele possa intervir de modo a reduzir o nível de evasão do curso.

Além disso, nesta pesquisa, pode ser visto claramente que os atributos utilizados na EDM podem sofrer diversas mudanças em prol de encontrar a melhor forma de se trabalhar com eles, visando maximizar os resultados da mineração. Com o aumento do uso de *softwares* de apoio pedagógico, bem como a utilização de DB com informações de estudantes, novos desafios aos profissionais de Informática em Educação vêm surgindo. Assim, surge, por exemplo, o desafio de como manipular com qualidade grandes volumes de informações sobre os estudantes armazenadas nesses repositórios.

Conforme visto nos trabalhos anteriores, é de grande importância a definição dos dados que serão utilizados na DM, ainda que seja tarefa complexa definir quais são os atributos mais relevantes para esta fase, segundo afirmam Gottardo, Kaestner, Noronha (2012): "os autores utilizaram um conjunto restrito de atributos, o qual não reflete todos os aspectos de interação, atividades e tarefas normalmente desenvolvidas por um estudante em

um Ambiente Virtual de Aprendizagem (AVA)". Com isso pode-se perceber que definir quais são os atributos mais relevantes para a DM é uma tarefa muito complicada.

O que é muito comum de se encontrar nos trabalhos relacionados à EDM são pesquisas por vezes limitadas a um escopo de avaliação restrito aos conjuntos de atributos tratados, que na maioria dos casos são selecionados a partir de critérios que privilegiam aspectos técnicos, tais como a disponibilidade de atributos em bases de dados ou o funcionamento de determinado algoritmo. Ao se realizar uma análise mais abrangente que envolva equipes multidisciplinares, aspectos adicionais podem ser destacados como relevantes, tais como as interações com colegas e docentes ou as especificidades de materiais instrucionais, ferramentas de apoio utilizadas nas aulas e metodologias de ensino [Rigo et al., 2012].

A comunidade de Informática na Educação tem realizado pesquisas que trabalham a identificação de atributos importantes de grandes massas de dados educacionais. Atributos esses com maior potencial para auxiliar a análise de perfis de estudantes, tornando possível compreender e interpretar comportamentos dos alunos e quais os atributos estão ligados a esses comportamento [Rigo et al., 2012].

No trabalho de Manhães et al. (2011), a dificuldade na escolha desses atributos é retratada. Além de não se ter um bom conhecimento sobre quais os melhores atributos devem ser utilizados nesse trabalho, também foram encontradas outras dificuldades, como a limitação de acesso aos dados e pouca informação sobre os dados coletados.

Para a definição e estruturação do modelo de processo de previsão de desempenho, Gottardo, Kaestner, Noronha (2012) propõe atributos agrupados em três diferentes dimensões, apresentadas a seguir:

1 - "Perfil geral de uso do AVA: nesta dimensão o objetivo é identificar dados que representem aspectos de planejamento, organização e gestão do tempo do estudante para a realização do curso. Para isso definiram-se indicadores gerais de quantidade e tempo médio de acessos aos recursos do AVA. Foram incluídos também atributos que representem atividades rotineiras e regulares dos acessos dos aprendizes..."

2 - "Segunda Dimensão: interação Estudante-Estudante: com esta dimensão pretende-se verificar se os estudantes interagem entre si usando as ferramentas disponíveis, como fóruns, *chats*, envio ou recebimento de mensagens. Espera-se, com estes atributos, identificar a existência de colaboração e cooperação entre estudantes. Fato esse denominado por Schrire (2006) como “aprendendo com os outros” (em inglês “*learning with others*”)..."

3 - "Terceira Dimensão: interação Estudante-Professor: nesta dimensão o objetivo é averiguar como professores ou tutores interagem com estudantes no contexto do AVA...".

Portanto, são perceptíveis, de acordo com os trabalhos esboçados acima, os esforços realizados na área de EDM que têm como objetivo comum melhorar a qualidade e eficácia de ambientes educacionais para que os alunos possam receber maior suporte durante seu processo de aprendizagem. Isso significa, por exemplo, que professores podem avaliar informações a partir da EDM e então classificar alunos de acordo com diversas características apresentadas durante a interação com os ambientes de ensino. No entanto, como já visto no trabalho de Garcia *et al.* (2011), as ferramentas de DM não foram projetadas especificamente para fins educacionais e/ou pedagógicos, pois não apresentam uma interface simples que possibilitaria a familiarização dos professores com tais ferramentas.

Além disso, a carga horária excessiva dos professores impede que eles consigam avaliar seus alunos com maior atenção e até mesmo realizar cursos para o aprendizado da utilização das ferramentas de EDM. Pesquisas indicam que a maioria dos professores desenvolve problemas de saúde pelas difíceis condições de trabalho a que são submetidos em seu dia a dia (Oliveira et al., 2013).

Em Reis et al. (2005, p. 1480), os autores afirmam que:

“[...] a categoria docente é uma das mais expostas a ambientes conflituosos e de alta exigência de trabalho, tais como tarefas extraclasse, reuniões e atividades adicionais” e que “77,8% dos professores submetidos a essas condições desenvolvem distúrbios psíquicos menores (DPM, distúrbios mentais comuns, tais como depressão, ansiedade, distúrbios somatoformes e neurastenia”

Já em Carlotto e Palazzo (2006, p.1017) afirmam que “[...] no atual modelo, muitas são as atribuições impostas ao professor, aparte de seu interesse e, muitas vezes, de sua carga horária”. Isso significa que, além das atividades em sala de aula, os professores devem cumprir tarefas administrativas, planejar, reciclar-se, investigar, orientar alunos e atender aos pais. Outro agravante de tal situação são os baixos salários que fazem com que os professores acabem por se responsabilizar com um maior número de turmas para lecionar, de modo que possam receber um salário mais adequado aos seus gastos mensais (Domingues, Toschi e Oliveira 2000).

Todas essas informações tornam evidente a falta de tempo adequado dos professores em avaliar seus alunos, identificar suas facilidades e dificuldades e ainda possíveis comportamentos inadequados. Tal avaliação proporcionaria um ensino personalizado, auxiliando no aprimoramento do aprendizado desses alunos.

Dentro deste contexto, buscou-se neste trabalho encontrar/gerar um processo de KDD capaz de caracterizar os alunos junto a um conjunto de atributos que maximizasse tal caracterização, trazendo resultados satisfatórios para todos os envolvidos no processo. A partir da conclusão dessa tarefa, espera-se, no futuro, gerar um processo que seja intuitivo, rápido e de simples utilização pelos professores, de forma a auxiliá-los a conhecer melhor seus alunos sem que precisem aumentar de forma significativa os esforços já realizados. Para isso, é importante que cada uma das fases do KDD seja muito bem analisada e realizada da forma mais automática possível.

Com a redução da interação dos professores com as fases do KDD, fazendo com que os mesmos apenas interajam nas fases imprescindíveis, e assim, que sejam gerados resultados que facilitem seu trabalho no dia a dia, mais professores utilizariam as ferramentas de EDM e, dessa forma, mais dados e hipóteses seriam gerados e validados contribuindo para a melhoria dos processos de ensino e aprendizagem e para maior qualidade das interações entre alunos e professores.

Neste trabalho, as fases do KDD foram estudadas e avaliadas, buscando-se automatizar ao máximo cada uma delas, gerando ainda dados adequados para a tarefa de DM. No capítulo seguinte serão apresentados os dados utilizados, a forma com que eles foram armazenados e também a ferramenta selecionada para realizar a DM.

4. Escopo do Estudo

Neste capítulo é apresentado o contexto em que esta pesquisa foi desenvolvida. Na seção 4.1 é apresentado o contexto no qual foram gerados os dados utilizados nesta pesquisa, na seção 4.1.1 é apresentado o sistema *Web MathTutor*, na seção 4.1.2 é apresentado o experimento do qual os dados foram utilizados nessa pesquisa. Já em 4.1.3 a etapa de intervenção do experimento é tratada com mais detalhes, pois foi de extrema importância para essa pesquisa, na seção 4.2 é detalhada a estrutura da planilha de *log* armazenada pelo sistema tutor inteligente MathTutor, e na seção 4.3 a ferramenta de DM DAMICORE e seu funcionamento são apresentados.

4.1. Contextos: escolas, alunos e o experimento

Para a realização do estudo proposto, foram utilizados, em um primeiro momento, os dados de um experimento desenvolvido com 330 alunos que cursavam a disciplina de matemática, nos ensinos fundamental e médio, de 3 escolas norte americanas (Escola A, Escola B e Escola C) no ano de 2012. Durante esse experimento, os alunos interagiram com o sistema tutor inteligente conhecido como MathTutor, trabalhando com exemplos errôneos em números decimais. Em um segundo momento, os dados resultantes da interação dos alunos com o MathTutor foram selecionados e tratados para serem “minerados” por meio de uma ferramenta de DM, a DAMICORE (seção 4.3).

4.1.1. MathTutor

O MathTutor é um sistema tutor inteligente de acesso aberto (gratuito e livre para qualquer um que deseja utilizá-lo) para o ensino e aprendizagem de matemática. O sistema oferece orientação detalhada, interativa e passo a passo para a resolução de problemas, além de relatórios detalhados do desempenho dos estudantes para toda a comunidade escolar (professores, pais, etc.) (Aleven; McLaren e Sewall, 2009).

Ao utilizar o sistema, os alunos têm a oportunidade de realizar diversos tipos de exercícios, com a opção de solicitar dicas quando estiver com dificuldades na sua resolução. A Figura 3 mostra um exemplo dessa interface, na qual é possível observar, na parte superior da interface do sistema, uma lista numérica (lista numérica, fundo branco e fonte verde) com todos os exercícios oferecidos naquele módulo.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32

Jorge has 3 cups of different sizes. The first cup can hold 0.47 L, the second can hold 0.5 L, and the third can hold 0.313 L. Jorge's friend asks him to pick the smallest cup. Which cup should Jorge choose?

Jorge said this incorrect answer: I should pick the 0.5 L cup. I found this by ordering the decimals from smallest to largest.

0	.	5	
0	.	4	7
0	.	3	1 3

Which answer best explains what Jorge did wrong?
He thinks decimals are the same as whole numbers and ____.

- ☐ longer decimals are smaller
- ☒ shorter decimals are smaller
- ☐ shorter decimals are larger
- ☐ longer decimals are smaller than zero

Here is a table showing the size of each cup. Rearrange the rows so that the numbers go from smallest to largest, top to bottom.

0	.	5	
0	.	4	7
0	.	3	1 3

Ok

Message Window

That's correct.

← Previous Next →

Figura 3 - MathTutor: interface *Software* de Auxílio ao Módulo Ensino de Números Decimais

Ainda conforme a Figura 3, ao interagir com o sistema, respondendo corretamente a primeira parte da questão selecionada (quadro vermelho – questão; quadro amarelo – resposta), outra pergunta referente ao mesmo contexto é exibida (quadro branco, à esquerda) e no canto inferior direito encontra-se um quadro cuja função é exibir as mensagens de orientação do sistema para o aluno.

Além da opção de selecionar a resposta correta dentro de uma lista, o sistema ainda oferece a opção de *drag and drop* (arrastar e soltar), como pode ser visto na Figura 4.

Figura 4 - Math Tutor: *drag and drop* Sistema de Auxílio ao Ensino de Números Decimais

Outro exemplo de exercício é apresentado Figura 4, no qual o aluno precisa escolher, entre as três alternativas propostas, a melhor resposta de acordo com o enunciado do exercício.

Figura 5 - MathTutor: exercício *Software* de Auxílio ao Ensino de Números Decimais

Com os diversos tipos de exercícios e um *feedback* para cada uma de suas interações, a ideia é que os alunos se sintam motivados a resolverem os exercícios propostos. Como já dito anteriormente, o MathTutor gera diversos relatórios sobre as interações dos alunos com o sistema. Para analisar esses relatórios foi utilizada a ferramenta DAMICORE que será descrita na seção 4.3.

4.1.2. O experimento

A realização do experimento consistiu em seis etapas: “pré-teste”, “questionário demográfico e de experiência matemática” (Q1), os “problemas da intervenção, questionário sobre a experiência do aluno com os materiais” (Q2), um “pós-teste imediato” e um “teste de retenção” (McLaren et al., 2012), todas elas realizadas com os alunos que já haviam interagido durante determinado período de tempo com o sistema tutor inteligente MathTutor..

Figura 6 - Etapas do Experimento (Adaptado De McLaren et al., 2012)

Na etapa “pré-teste”, buscou-se classificar os alunos de acordo com o seu conhecimento prévio. Assim, esse teste ocorreu sem nenhuma intervenção do professor, de forma que o aluno utilizou apenas o conhecimento que possui, sem ter assistido a aulas ou ter passado por quaisquer interações com o professor sobre o tema do teste.

Na “Q1” os alunos foram indagados com perguntas relacionadas às suas características como idade, sexo, conhecimento em matemática, conhecimento em números decimais, conhecimentos em informática, etc., visando, dessa forma, conhecer e caracterizar os alunos que participaram do experimento.

Na etapa “intervenção”, foram apresentados aos alunos um conjunto de exercícios guiados pelo sistema MathTutor no qual, ao resolver os problemas, os alunos trabalharam os conceitos da disciplina e tiveram a possibilidade de adquirir o conhecimento necessário sobre aquele tema.

Na “Q2” buscou-se por informações sobre a experiência do aluno com o sistema e com o conteúdo apresentado, por exemplo, quanto tempo o aluno interage com o computador, sua opinião sobre os exercícios, se faria mais exercícios, sua opinião sobre a apresentação do sistema, se o sistema era de fácil entendimento, etc.

A etapa “pós-teste imediato” teve como objetivo descobrir o quanto os alunos aprenderam na intervenção. Através da resolução de exercícios, assim como no “pré-teste”, buscou-se classificar os alunos, mas agora após a “intervenção”, verificando quais alunos aprenderam algo e o quanto eles aprenderam.

A etapa “teste de retenção” teve como objetivo verificar a quantidade de informações que os alunos conseguem assimilar num longo período, após a “intervenção”. Assim, esse

teste foi realizado com uma diferença maior de tempo em relação ao tempo dos outros testes e intervenções. Com o seu resultado, pôde-se verificar o que os alunos realmente aprenderam, e não apenas o que decoraram para resolver os exercícios no “pós-teste” (Isotani et al., 2011 e McLaren et al., 2012).

4.1.3. A etapa de “intervenção”

Durante a etapa “intervenção”, três grupos de alunos foram formados aleatoriamente e cada um deles recebeu um tipo diferente de intervenção. As intervenções foram baseadas nos seguintes fundamentos, como descrito em Gogvadze et al. (2011) e Isotani et al. (2011):

- (PS) Resolução de problema parcialmente apoiado: nesta condição, são apresentados ao aluno problemas para que o mesmo resolva. Além disso, existe um indicativo que exibe respostas incorretas em vermelho e respostas corretas em verde. Caso o aluno erre um dos problemas, assim que o erro é constatado, um botão é exibido na tela com o rótulo "Mostre-me a resposta". Desse ponto em diante, o aluno pode selecionar o botão para ver a solução correta para o problema (a combinação de *feedback* verde/vermelho e o botão "Mostre-me a resposta" é a razão desta condição ser chamada de resolução de problemas "parcialmente apoiada", em vez de simplesmente “resolução de problemas”, no qual não seria fornecido nenhum *feedback* ou ajuda).
- (ErrEx) Exemplos errôneos: nesta condição foram apresentados aos alunos materiais com o apoio de exemplos errôneos. Ou seja, eles recebem um *word problem* (quando um exercício matemático tem suas informações apresentadas como texto ao invés de utilizar notações matemáticas) e, para auxiliá-los a resolver este problema, são apresentados dois exemplos: um correto e um incorreto. Em seguida, os alunos respondem a duas perguntas, uma para explicar porque o exemplo incorreto é errado e outra para dar conselhos a um colega sobre como resolver corretamente o problema.
- (Adp) Exemplos Errôneos adaptados: funciona como o ErrEx, porém os temas dos exercícios são selecionados de acordo com o desempenho do aluno no “pré-teste”, visando sanar as dúvidas dos temas em que foram identificados mais erros.

Em cada escola, uma quantidade diferente de alunos realizou os testes, conforme pode ser visto na Tabela 2, que exibe os alunos por escola e por tipo de intervenção.

Tabela 2 - Alunos por escola e tipo de intervenção

Escola\Intervenção	Adaptative	ErrEx	ErrEx	P S	Total
Escola A	47	47	47	141	
Escola B	23	23	24	70	
Escola C	39	40	40	119	
Total	109	110	111	330	

Assim, todos os alunos tiveram suas interações armazenadas em arquivos de log, que foram transformados em planilhas (ANEXO A – Recorte da planilha de log gerada pelo MathTutor). Nas planilhas, cada registro (linha da planilha) armazena a ação do aluno em um determinado momento, por exemplo, quando ele seleciona uma resposta, finaliza um exercício, pede auxílio ao sistema, etc. Essa planilha será detalhada na próxima seção.

4.2. A estrutura dos Dados Utilizados ou a Planilha de log

A planilha de *log* consiste em uma tabela em que cada interação do aluno com o MathTutor (apresentado na seção 4.1.1) é armazenada. Para armazenar essas interações de forma eficiente, diversas informações são utilizadas. Nesta pesquisa, a planilha obtida através do experimento citado anteriormente possui 40 atributos, ou seja, 40 itens que são utilizados para detalhar a ação que o aluno tomou e todo o contexto que envolveu a sua decisão em um determinado momento. A planilha contém informações como “duração”, “tipo da ação”, “tipo do exercício com que está interagindo”, “saída do sistema”, etc. Esses atributos e suas respectivas descrições podem ser vistos na Tabela 3, bem como as diferentes formas de armazenamento.

Tabela 3 - Atributos da planilha de log

Nome do atributo	Descrição	Tipo
<i>Row</i>	Linha	<i>Number</i>
<i>Sample Name</i>	Nome da amostra	<i>String</i>
<i>Anon Student Id</i>	Identificação do estudante	<i>String</i>
<i>Session Id</i>	Identificação da sessão	<i>String</i>
<i>Time</i>	Data e hora	<i>Date</i>
<i>Time Zone</i>	Fuso horário	<i>String</i>
<i>Duration (sec)</i>	Duração em segundos	<i>Number</i>
<i>Student Response Type</i>	Tipo da resposta do estudante	<i>String</i>

<i>Student Response Subtype</i>	Subtipo da resposta do estudante	<i>String</i>
<i>Tutor Response Type</i>	Tipo da resposta do tutor	<i>String</i>
<i>Tutor Response Subtype</i>	Subtipo da resposta do tutor	<i>String</i>
<i>Level(Assignment)</i>	Tipo e ordem dos exercícios	<i>String</i>
<i>Level(ProblemSet)</i>	Nome do exercício	<i>String</i>
<i>Problem Name</i>	Nome, número e tipo do exercício	<i>String</i>
<i>Problem View</i>	Quantidade de vezes que viu aquele exercício	<i>Number</i>
<i>Problem Start Time</i>	Data e horário do início da resolução do exercício	<i>Date</i>
<i>Step Name</i>	Nome da etapa	<i>String</i>
<i>Attempt At Step</i>	Número da tentativa na etapa do exercício	<i>Number</i>
<i>Outcome</i>	Saída	<i>String</i>
<i>Selection</i>	Nome e número do exercício + etapa	<i>String</i>
<i>Action</i>	Ação	<i>String</i>
<i>Input</i>	Entrada	<i>String</i>
<i>Feedback Text</i>	<i>Misconception</i> ¹⁰	<i>String</i>
<i>Feedback Classification</i>	Classificação do <i>feedback</i>	-
<i>Help Level</i>	Nível de ajuda	-
<i>Total Num Hints</i>	Número total de dicas	<i>Number</i>
<i>Condition Name</i>	Tipo e ordem dos exercícios	
<i>Condition Type</i>	Tipo da condição	<i>String</i>
<i>KC(Single-KC)</i>	-	-

¹⁰ Falso conceito criado pelo aluno

<i>KC Category(Single-KC)</i>	-	-
<i>School</i>	Escola	<i>String</i>
<i>Class</i>	Classe	<i>String</i>
<i>CF(calculated_offset)</i>	-	-
<i>CF(question)</i>	-	-
<i>CF(server_adjusted_event_time)</i>	-	-
<i>CF(server_name)</i>	-	-
<i>CF(server_time_millis)</i>	-	-
<i>CF(step_id)</i>	-	-
<i>CF(tool_event_time)</i>	-	-
<i>CF(tutor_event_time)</i>	-	-

Essa planilha de *log* é recebida em formato de texto (txt) e cada interação do aluno é armazenada em uma linha diferente, na qual as informações são separadas por uma tabulação. Cada vez que o aluno realiza alguma ação no sistema, como selecionar uma resposta, o sistema armazena diversos dados sobre essa ação: identificador do aluno (“*Anon Student Id*”), tipo do exercício (“*Problem Name*”), tempo gasto para realizar aquela ação (“*Duration (sec)*”), ação realizada (“*Action*”), etc.

A seguir será apresentada a DAMICORE ferramenta de DM utilizada para buscar padrões nos dados já apresentados. Está ferramenta que trabalha repositórios de dados e já foi utilizada em diversos tipos de dados, vem obtendo bons resultados e agora será utilizada para trabalhar com dados educacionais.

4.3. DAMICORE

A proposta da ferramenta DAMICORE é trabalhar com a combinação de algoritmos já utilizada com sucesso em outros campos da ciência: a *Normalized Compression Distance* (NCD), da Teoria da Informação (Cilibrasi e Vitanyi, 2005); *Neighbor Joining* (NJ), da

filogenética (Felsenstein, 2003) e *FastNewman* (FN), de Redes Complexas (McLaren et al., 2008). Com a utilização dessas três técnicas, NCD, NJ e FN, um minerador de dados inteligente conhecido como DAMICORE (*data mining of code repository*) foi gerado (Carvalho, 2012).

A DAMICORE é uma ferramenta de DM que possui utilização bastante simples. Para utilizá-la basta colocar em um repositório arquivos do mesmo tipo, cujo conteúdo são os dados que necessitam ser minerados. É necessário que os indivíduos que serão tratados sejam separados cada um em um arquivo diferente. Com isso, é aplicada a técnica NCD para o cálculo da distância, para o agrupamento e para a compressão dos dados analisados. Após isso, a técnica NJ constrói uma rede filogenética e por fim, a técnica FN identifica e forma os *clusters* dos dados de entrada (Sanches, Cardoso e Delbem, 2011).

Uma das ideias iniciais deste trabalho era a de realizar a comparação entre o desempenho da DAMICORE com outras ferramentas bastante utilizadas para minerar dados, como a Weka e Rapidminer. No entanto, além do curto tempo para realização dessa comparação, algumas dificuldades no gerenciamento dos dados foram encontradas. Assim, por exemplo, ao se trabalhar com a Weka, a maior dificuldade surgiu em fases anteriores à fase de DM, nas quais era necessário pré-processar e transformar os dados para um formato que a ferramenta fosse capaz de interpretá-los e, uma vez se tratando de grande quantidade de dados, isso implicou em maior complexidade para essa tarefa.

Deste modo, optou-se pela utilização da DAMICORE nesta pesquisa devido à facilidade e simplicidade encontrados em seu uso e gerenciamento, o que não é comumente encontrado em outras ferramentas de DM ou EDM, conforme explicitado acima.

A seguir serão descritas as técnicas utilizadas na DAMICORE.

4.3.1. Técnicas utilizadas pela DAMICORE

A NCD (Cilibrasi e Vitanyi, 2005) pode ser usada para vários tipos de dados. De fato, é uma métrica que pode encontrar relações entre os dados, mesmo quando não se conhecem as características que podem ser usadas para representar adequadamente o objeto analisado. Como descrito por Sanches, Cardoso e Delbem (2011, p.76):

“[...] a ideia básica de seu funcionamento é que, dado dois objetos “a” e “b”, os mesmos são considerados próximos se “a” pode ser significativamente

comprimido usando a informação de “b”, e vice-versa, ou seja, se duas entidades são semelhantes é mais fácil (ou mais sucinto) descrever a segunda entidade apenas se referindo às suas diferenças em relação a primeira.”

A segunda técnica utilizada na DAMICORE é a NJ, que vem da filogenética, campo da biologia que investiga as relações entre as espécies (ou outros objetos). Uma estrutura de árvore (às vezes chamada de uma rede) é utilizada para descrever as relações. Os tipos de dados utilizados podem ser tanto discretos (como a ausência de asas em uma determinada espécie) quanto contínuos (como o tamanho da espécie). Os métodos mais recentes para a reconstrução filogenética, em geral, buscam por árvores de maior probabilidade. No entanto, o método antigo do *Neighbor Join* é um dos algoritmos mais simples e eficientes para a filogenia, que constrói filogenias competitivas como as produzidas pelos métodos mais confiáveis (Paiva et al., 2011).

Além disso, Sanches, Cardoso e Delbem p.76 (2011), afirma que:

“[...] o NJ não depende da evolução dos modelos biológicos ou outro tipo de conhecimento prévio do domínio do problema para funcionar bem, ou seja, exige apenas uma estimativa da matriz da distância entre os objetos. Esses aspectos são os principais motivos que explicam porque o NJ é usado atualmente em laboratórios de biologia. Em outras palavras, o NJ possivelmente apresenta o melhor desempenho entre custo e benefício para lidar com uma diversidade relativamente grande dos tipos de dados. Dessa maneira, é um algoritmo interessante para o desenvolvimento de uma ferramenta para a identificação de semelhanças hierárquicas entre os códigos de software. A saída do algoritmo NJ é uma árvore de relações, assim, outro método é necessário para extrair os clusters potenciais (subtipos em biologia) escondidos na topologia da árvore.”

Assim, para dar continuidade à DM e tomar as corretas decisões sobre as redes (árvores, florestas ou grafos) relevantes, algoritmos foram desenvolvidos no campo das Redes Complexas (Newman, 2010). Entre eles, o Fast Newman (FN) obteve sucesso para vários contextos de rede em grande escala.

Uma vez encontrada uma filogenia, deve-se verificar se existem subtipos, por exemplo, sub-árvores da árvore filogenética com elevado grau de independência. Daqui para frente são utilizadas as Redes Complexas para descobrir *clusters* na rede (árvore) gerada pelo

NJ. A FN pode lidar com redes de grande escala para encontrar as principais comunidades representativas de um sistema. Assim, este algoritmo se torna interessante devido a dois aspectos: i) a sua escalabilidade o torna adequado para a manipulação de muitas funções/grãos, ii) ela funciona corretamente sobre os dados organizados em um modelo de rede, que permita a sua aplicação, ou seja, as árvores filogenéticas (Sanches, Cardoso e Delbem, 2011).

O processamento de toda a tecnologia proposta, combinando NCD, NJ e FN para mineração de dados organizados em um repositório de códigos é chamado de DAMICORE.

Porém, como mostrado na seção 2.1, para minerar os dados relevantes, algumas etapas devem ser seguidas, sendo a primeira delas a fase de “seleção”, na qual serão escolhidos quais dados deverão ser utilizados nas fases seguintes do KDD. Para dar suporte a essa fase foi realizado um mapeamento sistemático, descrito na seção 5.1, visando encontrar na literatura embasamento para retirar, manter e criar novos atributos para serem investigados.

5. Seleção de Dados Educacionais e sua Relevância para a Caracterização de Alunos

Em diversos trabalhos sobre mineração de dados, pesquisadores vêm avaliando diferentes métodos de seleção de atributos com o objetivo de identificar quais deles irão proporcionar, de forma mais completa e minuciosa, conhecimento sobre os dados analisados sem que haja perda de detalhes que enriqueçam a avaliação das informações. John, Kohavi e Peger (1994), por exemplo, propuseram a metodologia *wrapper*, que consiste na avaliação de conjuntos de atributos usando um algoritmo que funciona como uma caixa preta para encontrar os melhores subconjuntos de atributos. Já Carlos et. al realizaram a seleção dos atributos a partir do método conhecido como “qui quadrado”, o qual avalia quantitativamente a relação entre os atributos utilizados e o resultado de um experimento.

No presente trabalho, a fase de “seleção” dos dados foi feita sob uma abordagem diferente à dos trabalhos supracitados. Essa abordagem permitiu investigar os trabalhos disponíveis nas bases de dados mais relevantes na área de EDM com o objetivo de descobrir quais são os atributos mais utilizados para a mineração de dados educacionais para que, dessa forma, sejam escolhidos atributos que forneçam resultados expressivos a esta pesquisa.

Como citado por Kravvaris, Kermanidis e Thanou, 2012; Baker, Gowda e Corbett, 2011, materiais de apoio para realizar a fase de “seleção” dos dados são imprescindíveis, uma vez que esses materiais dão suporte para uma avaliação completa das pesquisas da área desejada. Para isso, neste trabalho optou-se por realizar um estudo de mapeamento sistemático uma vez que ele é projetado com o intuito de fornecer uma visão geral de uma área de investigação, determinando se existem evidências documentadas em pesquisas sobre determinado tema, além de identificar a quantidade dessas evidências (Kitchenham e Charters, 2007). Neste trabalho, a área de investigação é a de mineração de dados educacionais e o tema pesquisado está relacionado aos atributos utilizados com maior frequência nessa área, ou seja, atributos com maior potencial dentro da EDM para que se possa gerar grande margem de conhecimento sobre alunos (prever comportamentos, identificar alunos que burlam o sistema, etc.).

Com o resultado do mapeamento sistemático, foi possível obter um ponto inicial sobre quais atributos utilizar na pesquisa proposta, bem como avaliar a importância dos atributos identificados baseando-se em estudos de outros pesquisadores com dados e objetivos similares. Dessa forma, a fase de “seleção” se tornou mais simples e eficiente, proporcionando resultados de forma mais rápida e com maior precisão.

Nas seções seguintes, cada uma das etapas do mapeamento sistemático é apresentada com mais detalhes, tal como quais foram as ações tomadas e o modo de condução do estudo. Na seção 5.1 é apresentada a metodologia empregada para a realização do mapeamento; na seção 5.2 é apresentada a análise feita e os resultados obtidos, na seção 5.3 é apresentada uma síntese dos estudos encontrados; e na seção 5.4 são apresentadas as considerações finais e as ameaças à validade desse mapeamento, bem como as considerações finais quanto à relevância dos dados.

5.1. Mapeamento Sistemático

O objetivo deste mapeamento foi descobrir quais são os atributos mais utilizados pelos pesquisadores na mineração de dados educacionais. Este estudo baseou-se nas diretrizes propostas por Petersen et al. (2008) e foi conduzido de acordo com os cinco passos a seguir: (i) definição das questões de pesquisa, (ii) realização da pesquisa de estudos primários relevantes, (iii) triagem dos documentos, (iv) *keywording* dos resumos, e (v) extração de dados e mapeamento. Na Figura 7 pode-se observar os passos propostos por Petersen et al. (2008).

Figura 7 - Passos de um Mapeamento Sistemático (adaptado de Petersen et al. (2008))

5.1.1. Definição das questões de pesquisa do Mapeamento

Para atender à necessidade de identificar a relevância dos dados educacionais e selecionar aqueles mais adequados para caracterizar alunos em ambientes de aprendizagem, foram definidas três questões a serem respondidas pelo mapeamento sistemático:

- **QPM01** – Qual é o principal foco dos trabalhos referentes à mineração de dados educacionais?

- **QPM02** – Quais são os atributos mais utilizados em trabalhos referentes à mineração de dados educacionais?
- **QPM03** – No caso referente à “Caracterização dos Alunos”, quais são os atributos mais utilizados em trabalhos referentes à mineração de dados educacionais?

5.1.2. Condução das atividades do Mapeamento Sistemático

Esta etapa é compreendida pela execução da *string* (frase) de pesquisa nas bibliotecas digitais. A fim de obter um maior número de estudos primários, as seguintes bases foram consultadas, pois, de acordo com Dybå et al. (2007), estas bibliotecas estão entre as mais relevantes para a área de Computação:

- **ACM Digital Library** (portal.acm.org);
- **IEEEExplore** (ieeexplore.ieee.org);
- **Scopus** (<http://www.scopus.com>);
- **Elsevier - via Science Direct** (www.sciencedirect.com); e,
- **Springer Link - via Science Direct** (www.sciencedirect.com).

Por meio de uma *string* específica definida em conjunto com especialistas na área de EDM, buscou-se por artigos que se enquadravam no escopo deste trabalho, chamados de estudos primários (Petersen et al. (2008)). Para encontrar esses estudos, disponíveis nas bibliotecas selecionadas, foi utilizada uma *string* de busca, na língua inglesa, constituída de três partes e formatada de acordo com as regras de cada biblioteca digital, conforme apresentado na coluna “Palavra-chave” da Tabela 4. A primeira parte “*attribute*” trata dos atributos; a segunda parte “*educational data mining*” conecta tais atributos com a EDM e a última parte “*student data*” visa encontrar artigos que trabalhem informações referentes aos alunos.

Tabela 4 - *String* de busca em inglês.

Palavra-chave	Termos alternativos e/ou sinônimos
<i>Attribute</i>	<i>Property OR Feature OR Features OR "Labeled Data" AND</i>
<i>Educational Data Mining</i>	<i>EDM AND</i>
<i>Student data</i>	<i>"Learning Data"</i>

Na Tabela 4, o operador *boolean* OR, que surge na coluna “Termos alternativos e/ou sinônimos”, conecta os sinônimos da *string* de pesquisa, enquanto o operador *AND* conecta as

partes. Nenhum campo específico (filtro) foi utilizado para limitar as buscas nas bibliotecas digitais selecionadas. Os estudos selecionados foram analisados com o intuito de identificar aqueles relevantes ao contexto da pesquisa. Essa análise consistiu na aplicação de critérios de inclusão e exclusão previamente definidos. Os critérios de inclusão elaborados foram os seguintes:

- Caso vários artigos mostrem estudos similares, somente o mais recente será incluído;
- Se houver várias versões do mesmo artigo, e.g. uma resumida e outra completa, a mais completa será incluída e,
- Artigos que descrevem quais atributos foram utilizados durante as tarefas de mineração dos dados.

Já os critérios de exclusão permitem maior precisão na eliminação de estudos considerados irrelevantes ao contexto da pesquisa em andamento. Por essa razão, durante a análise dos estudos retornados, todos aqueles que se enquadraram em ao menos um dos critérios de exclusão abaixo foram descartados. Os critérios são:

- Artigos que não estejam relacionados com mineração dos dados educacionais;
- Estudos que não estejam na língua portuguesa ou inglesa;
- Artigos que não apresentem nenhuma informação sobre os atributos utilizados nas tarefas de mineração dos dados e,
- Relatórios técnicos, documentos que estejam disponíveis na forma de resumos ou apresentações/*slides* e também estudos secundários (ou seja, revisões e mapeamentos sistemáticos da literatura).

O mapeamento sistemático foi realizado no período do dia 20/09/2013 ao dia 02/05/2013. Primeiramente, foram encontrados 215 estudos primários, apresentados na Tabela 5.

Tabela 5 - Quantidade de estudos retornados por cada base de dados eletrônica, total de estudos candidatos e seleção final.

Base de dados	Quantidade
ACM Digital Library	36
Scopus	84
Science Direct	38
IEEEExplore	57
Total	215
Candidatos	163
Seleção final	32

A revisão desses trabalhos foi realizada aos pares de maneira simultânea. Na primeira etapa fez-se a leitura dos títulos, resumos e palavras-chave de todos os artigos, criando uma lista de estudos selecionados por revisor. Na segunda etapa discutiu-se sobre quais estudos deveriam ser selecionados e, ao final, o conjunto inicial de 215 estudos foi reduzido a 163 estudos que realmente estavam relacionados com o tópico investigado. Após esse passo, foi realizada novamente a leitura dos resumos, introdução e conclusão dos 163 estudos selecionados aplicando os critérios de inclusão e exclusão, o que resultou em um subconjunto de 32 estudos. Desses 32 estudos, 2 foram excluídos por não exibirem de forma clara e completa quais foram os atributos (i.e. dados) utilizados durante o trabalho. Em seguida, foi realizada a leitura integral dos estudos selecionados a fim de encontrar e catalogar os atributos e objetivos descritos nos mesmos. Esta etapa será discutida detalhadamente na próxima seção.

5.2. Análise dos Estudos Primários

Esta seção apresenta uma visão geral dos atributos de EDM encontrados nos 30 estudos selecionados, bem como seus objetivos de pesquisa. Para melhor organizar os objetivos dos trabalhos e os atributos utilizados para minerar os dados, foram criadas classes (categorias) que permitiram agrupar informações semelhantes, visto que, devido à falta de uma padronização de termos, essas informações podem ter sido citadas em artigos distintos por diferentes maneiras, como “Tempo”, “Tempo Médio”, “Tempo Final”, “Duração”, etc.

Adicionalmente, também foram criadas classes que permitiram agrupar informações que estão relacionadas umas com as outras. Assim, por exemplo, atributos como “Idade”, “Nome”, “Sexo”, “Religião”, entre outros, foram agrupados na classe “Dados Pessoais”. Da mesma forma, “Tempo total em minutos”, “Média de tempo para responder a uma pergunta”, “Duração”, “Tempo inicial”, “Tempo final”, etc. foram agrupados na classe “Tempo”. A Tabela 6 apresenta as classes criadas para os atributos.

Tabela 6 - Classes dos Atributos e suas respectivas Descrições

Classe Atributo	Descrição
Sessão	Informações referentes ao acesso do aluno a um determinado sistema.
Dados pessoais	Informações referentes aos dados pessoais do aluno.

Nota	Informações referentes ao desempenho do aluno.
Tempo	Informações referentes ao tempo.
Chat	Informações referentes às interações do aluno no <i>chat</i> .
Wiki	Informações referentes às interações do aluno com a <i>Wiki</i> .
Fórum	Informações referentes às interações do aluno com o Fórum.
Tarefa	Informações referentes às tarefas que o aluno realizou.
E-mail	Informações referentes às interações do aluno no <i>e-mail</i> .
Tentativa	Informações referentes às tentativas do aluno em determinado exercício, questionário, etc.
Dica	Informações referentes às dicas que os alunos receberam.
SVN	Informações referentes às interações no SVN.
Curso	Informações referentes ao curso.
Motivação	Informações referentes à motivação do estudante.
Turma	Informações referentes à turma na qual o estudante está cursando a disciplina.
Escola	Informações sobre escolas na qual o aluno estuda ou já estudou.
Mensagem Professor	Informações referentes às mensagens trocadas com o professor.
Off-task behavior	Informações referentes aos comportamentos fora das tarefas de ensino.
Gaming the system	Informações referentes aos alunos que burlaram o sistema.

O mesmo procedimento foi realizado com os “objetivos” dos artigos. Dessa forma, após estudar cada artigo foram criadas classes de objetivos definidas na Tabela 7.

Tabela 7 - Classes definidas para os Objetivos e suas respectivas descrições

Classe Objetivo	Descrição
Definição dos Atributos	Definir quais são os melhores atributos para utilizar na DM.
Definição da Interface	Definir qual a melhor interface para professores, alunos e todos os envolvidos em ferramentas de auxílio aos processos de ensino e aprendizagem.
Predizer ajuda que o aluno necessita	Predizer quando e qual a forma de auxílio o aluno necessita em determinada fase do ensino e aprendizagem.
Exibir indicadores de ensino dos alunos	Exibir indicadores que demonstrem o desempenho do aluno para professores, alunos e todos envolvidos no ensino e aprendizagem.
Caracterizar os alunos	Caracterizar os alunos de forma a criar grupos com características similares ou diferentes.
Prever quais alunos irão cometer falhas	Predizer quais alunos irão falhar em determinada tarefa.
Demonstrar qualidade de algoritmo	Avaliar o desempenho de determinado algoritmo.
Auxiliar definições variáveis em cursos	Encontrar evidência para se criar modelos para cursos (Número de páginas na <i>Web</i> , quantidade de exercícios, etc.).
Exibir sugestões para o ensino dos alunos	Exibe sugestões práticas de como o professor deve agir com determinados alunos.
Como desenvolver novos métodos para a EDM	Como desenvolver novos métodos para se obter sucesso ao minerar dados educacionais.
Classificação de técnicas de	Avalia as técnicas de ensino utilizadas pelos professores

ensino	nos processos de ensino.
Desenvolver processos para a criação de ITS	Busca encontrar a melhor forma de se criar um ITS.
Avaliação de métodos colaborativos	Avalia quais os melhores métodos para se trabalhar colaborativamente.
Definição do melhor momento para se classificar os alunos	Estudo para descobrir o melhor momento para se caracterizar os alunos.

Com base nessas classes, os artigos foram estudados de modo a responder as questões definidas anteriormente na seção 5.1.1. Os resultados encontrados são apresentados a seguir na forma de gráficos.

QPM01 – Qual é o principal foco dos trabalhos referentes à mineração de dados educacionais?

Analisando o gráfico da Figura 8, que exhibe os objetivos encontrados nos estudos (em alguns estudos foi encontrado mais que um objetivo) e a quantidade de vezes que cada um desses objetivos é mencionado nos trabalhos selecionados, é possível perceber que o objetivo da maioria dos trabalhos (13 dentre os selecionados) da área de EDM busca “Caracterizar os alunos”, ou seja, criar classificações para os alunos de forma a agrupá-los sob determinada característica.

Além disso, foram também identificados outros objetivos mais frequentes: “Exibir indicadores de ensino dos alunos” (3 trabalhos), que tratam de exibir indicadores (classificações de características) que demonstrem a todos os envolvidos nas etapas de ensino e aprendizagem (alunos, professores, pais, etc.), o desempenho dos alunos ao trabalharem nas atividades propostas dentro e fora da sala de aula; “Demonstrar qualidade de algoritmo” (4 trabalhos), por meio da avaliação da qualidade de um algoritmo ao trabalhar com a EDM e “Definição dos Atributos” (3 trabalhos), que objetiva descobrir qual o melhor conjunto de atributos para realizar a fase de “mineração dos dados”.

A Figura 8 a seguir relaciona todos os objetivos com a frequência com que foram identificados.

Figura 8 - Quantidade dos Objetivos encontrados nos estudos

QPM02 – Quais são os atributos mais utilizados em trabalhos referentes à mineração de dados educacionais?

Em relação aos atributos mais utilizados na EDM, foram encontradas 19 categorias (“Sessão”, “Dado pessoal”, “Nota, Tempo”, “Chat”, “Wiki”, “Fórum”, “Tarefa”, “E-mail”, “Tentativa”, “Dica”, “SVN”, “Curso”, “Motivação”, “Turma”, “Escola”, “Mensagem Professor”, “Off-task behavior” e “Gaming the system”). O gráfico da Figura 9 apresenta a frequência com que essas categorias foram encontradas nos trabalhos mapeados e é possível observar que seis categorias aparecem mais frequentemente nos trabalhos: “Tarefa” (72 ocorrências), “Tempo” (62 ocorrências), “Fórum” (47 ocorrências), “Nota” (43 ocorrências), “Curso” (33 ocorrências) e “Dados Pessoais” (23 ocorrências).

Figura 9 - Quantidade dos Atributos encontrados nos estudos

As categorias que aparecem com um número maior de ocorrências no gráfico da Figura 9, respondem à QPM1 e representam os atributos mais utilizados em trabalhos de EDM. No caso, a categoria "Tarefa" retrata toda e qualquer atividade realizada pelo aluno, podendo ser questionários, provas, exercícios, entre outros. "Tempo" envolve todos os atributos que tenham alguma relação direta com o tempo que o aluno dedicou em determinada ocasião, podendo incluir a realização de uma tarefa, a resposta a uma questão, o tempo logado no sistema, etc. A categoria "Fórum" retrata as interações dos alunos dentro do fórum, tais como envio, recebimento e leitura de mensagens. "Nota" se refere às informações armazenadas relacionadas a qualquer tipo de nota que o aluno obteve como resultado da avaliação do seu desempenho em determinado exercício, atividade, disciplina, etc. "Curso" traz informações referentes ao curso em questão, como nome do curso, duração, etc. Por fim, a categoria "Dados Pessoais" faz referência a todos os dados que envolvem informações sobre os alunos, como nome, idade, sexo, etc. Além dessas categorias, o gráfico da Figura 9 apresenta ainda outros atributos que foram identificados, porém em um menor número de trabalhos. "Sessão", "Chat", "Wiki", "E-mail", "Tentativa", "Dica", "SVN", "Motivação", "Turma", "Escola", "Mensagem Professor", "Off-task behavior", "Gaming the system".

QPM03 – No caso referente à "Caracterização dos Alunos", quais são os atributos mais utilizados em trabalhos referentes à mineração de dados educacionais?

Para responder à QPM3, foram avaliados apenas os trabalhos que tinham como objetivo "Caracterizar os Alunos". O resultado pode ser visto no gráfico da Figura 10, na qual

os atributos (“Sessão”, “Dado pessoal”, “Nota”, “Tempo”, “Chat”, “Wiki”, “Fórum”, “Tarefa”, “E-mail”, “Tentativa”, “Dica”, “SVN”, “Curso”, “Motivação”, “Turma”, “Escola”, “Mensagem Professor”, “Off-task behavior”, “Gaming the system”) são exibidos de acordo com o número de vezes que foram utilizados nos estudos. Dentre esses estão os que representam as características mais utilizadas nesse tipo de pesquisa: “Tempo” (45 ocorrências), “Tarefa” (37 ocorrências), “Curso” (18 ocorrências), “Nota” (15 ocorrências), “Dados Pessoais” (11 ocorrências) e “Fórum” (10 ocorrências).

Figura 10 - Quantidade dos Atributos encontrados nos estudos que possuem o Objetivo de Caracterizar os Alunos

Além desses atributos, o gráfico da Figura 10 apresenta ainda outros atributos (“Sessão”, “Chat”, “Wiki”, “E-mail”, “Tentativa”, “Dica”, “SVN”, “Motivação”, “Turma”, “Escola”, “Mensagem Professor”, “Off-task behavior”, “Gaming the system”) que foram identificados com menos frequência nos estudos mapeados.

O APÊNDICE A – Tabela de classificação dos artigos, traz uma tabela com os campos: “Nome do artigo”, “Objetivo” e os “Atributos”, que sintetiza os dados dos 32 artigos utilizados nessa pesquisa e que, posteriormente, tiveram seus dados utilizados para auxiliar nas fases de “mineração de dados” e na “interpretação e avaliação do conhecimento”.

5.3. Síntese dos estudos

Dentre os 30 estudos selecionados, 13 deles tinham como objetivo caracterizar os alunos. Essa caracterização consiste em classificá-los e, possivelmente, agrupá-los de acordo com algumas de suas características comuns (apresentação de alguma ferramenta, modelos para prever comportamentos, avaliação de algum algoritmo, etc.). Esses estudos buscam caracterizar os alunos de diferentes formas e por meio de diversas técnicas e algoritmos. Além disso, a maioria deles possui outros objetivos (elaborar técnicas de visualização dos dados, descobrir melhores dados para serem minerados, etc.) que, ao serem concluídos, auxiliam na caracterização desses alunos.

Em alguns trabalhos são apresentadas as ferramentas utilizadas para a caracterização dos alunos. Em Martín et al. (2008) é apresentada a ferramenta *Sigma*, que tem como principal objetivo fornecer informações sobre o comportamento dos alunos ao realizar atividades de aprendizagem. Em Romero, Ventura e Garcia (2007), as ferramentas *Weka* e *Keel* são utilizadas para minerar dados educacionais, exibindo de forma teórica e prática como gerar conhecimento sobre os alunos que trabalham em Ambientes Virtuais de Aprendizagem. Em Wixon et al. (2012) é apresentado o primeiro detector automático de comportamento WTF (*Without Thinking Fastidiously*) e em Zorrilla, García e Álvarez (2010) é trabalhado um *data warehouse*, que busca gerar e armazenar modelos de mineração de dados para caracterizar os alunos.

Outros estudos, como os realizados por Baker, Gowda e Corbett (2011) trabalham com modelos para prever comportamentos dos alunos, que podem ser usados tanto para compreender as condições de aprendizagem quanto para dirigir intervenções destinadas a aumentar a robustez de aprendizagem dos alunos. Nessa mesma linha, Drazdilova et al. (2011) buscam auxiliar os professores, oferecendo a eles suporte para avaliar os dados armazenados sobre os alunos que trabalharam em um AVA. Também Romero et al. (2009) buscam auxiliar o professor com informações sobre os alunos e o uso dos recursos educacionais baseados na *Web*, gerando resultados que informam sobre os benefícios e/ou malefícios do uso desses recursos.

Alguns trabalhos também investigam processos anteriores à mineração de dados, como Kazanidis et al. (2012), que buscam avaliar a relação entre o desempenho dos alunos e as características do curso. Nesse trabalho, verificam-se quais são os dados mais importantes para avaliar os alunos e maximizar a tarefa de mineração dos dados. Já Kravvaris, Kermanidis e Thanou (2012) procuram avaliar o sucesso ou o fracasso dos alunos que participam de um

determinado curso e analisar os atributos do conjunto de dados armazenados com as informações das interações do aluno no AVA, a fim de obter informações úteis sobre o curso. Em uma vertente semelhante, o trabalho de Hershkovitz et al. (2011) busca encontrar informações sobre a melhor estrutura para cursos com suporte *Web*, por meio de algoritmos de agrupamento, para posteriormente identificar associações entre essas informações e determinadas características do curso, como o tamanho do curso e a forma como as informações de apoio ao alunos serão disponibilizadas para eles.

Dos três últimos trabalhos relacionados com a caracterização dos alunos, dois apresentam e avaliam algoritmos utilizados para a EDM. Em Carmona et al. (2010) são apresentados dois algoritmos evolutivos, que trabalham com o objetivo de caracterizar subgrupos de estudantes cujas notas finais são significativamente diferentes das notas da maioria dos alunos. E Zafra e Ventura (2012) apresentam um algoritmo genético, utilizado para agrupar os alunos e realizar uma comparação de seu desempenho com outros algoritmos, para que, dessa forma, sua acurácia fosse comprovada. O último trabalho, Drazdilova et al. (2009) tem uma proposta diferente da dos trabalhos supracitados, uma vez que ele busca, por meio da análise dos registros armazenados da utilização dos alunos em um AVA, gerar uma rede que conecte os alunos através de relações definidas pelas atividades em que eles trabalharam. Através dessa rede é possível verificar, de forma detalhada, o nível de intensidade das relações entre os alunos e os diversos tipos de atividades que eles realizaram.

5.4. Ameaças à Validade e Considerações Finais

A fim de garantir uma seleção imparcial, as questões de pesquisa e os critérios de inclusão e exclusão de artigos foram definidos previamente. No entanto, é necessário considerar a ameaça em relação à avaliação da qualidade dos estudos incluídos nesta pesquisa, já que foram selecionados sem qualquer atribuição de pontuação.

Outra possível ameaça identificada é a possibilidade de alguns artigos relevantes não terem sido incluídos devido à utilização de um conjunto limitado de bibliotecas digitais. Além disso, o sistema de classificação e categorias elaboradas representa também uma ameaça à validade, pois, como mostrado por Pretorius e Budgen (2008), a melhor maneira de se classificar e categorizar os resultados só é obtida ao final da seleção.

De qualquer forma, o mapeamento realizado foi útil para identificar os atributos com grande potencial para pesquisas na área de EDM. Especificamente para esta pesquisa, foram identificados determinados atributos como sendo de grande relevância e, portanto, foram

utilizados para encontrar informações valiosas sobre os alunos estudados. Os atributos escolhidos para essa pesquisa estão descritos na seção 6.1.

Além disso, vale ressaltar que este trabalho pode auxiliar pesquisadores da área educacional na identificação dos atributos mais relevantes para condução de trabalhos relacionados

No capítulo seguinte serão exibidas as fases da EDM realizadas para a geração de conhecimento a partir dos dados selecionados.

6. Desenvolvimento e Discussão

Após a conclusão do mapeamento sistemático, foram identificados os atributos com grande potencial para pesquisas na área de EDM. A partir dessa informação, conforme apresentado na seção 2.1, foi possível selecionar com maior acurácia quais eram os dados mais relevantes para realizar cada uma das fases do KDD, de forma a trazer conhecimento sobre os alunos e seus respectivos comportamentos e desempenhos ao interagirem com o MathTutor.

Na fase de “seleção”, optou-se por utilizar os dados referentes à etapa de “intervenção” do experimento por se tratar de uma etapa inicial na qual o aluno interage com o sistema tutor inteligente pela primeira vez, além de poder fornecer informações de como o aluno irá interagir em atividades subsequentes. Assim, para dar início ao processo, as fases do KDD (Figura 2) foram personalizadas a fim de torná-las adequadas a esta pesquisa.

No total foram definidas quatro fases para gerar o conhecimento desejado (i.e. caracterizar o aluno de acordo com seu desempenho): Fase 1 – “seleção”; Fase 2 “pré-processamento”; Fase 3 “mineração dados”; e Fase 4 “interpretação e avaliação”, as quais podem ser observadas na Figura 11.

Figura 11 - Fases do KDD personalizadas

Na Figura 11, os dados brutos da intervenção foram selecionados (Fase 1) e pré-processados (Fase 2) para que pudessem ser inseridos na ferramenta de mineração de dados

DAMICORE. Como discutido anteriormente, esses dados são referentes às interações entre os alunos e o MathTutor, e foram obtidos durante a etapa de “intervenção” do experimento, apresentada na seção 4.1.3.

Após a inserção dos dados na DAMICORE (Fase 3), foram geradas árvores que, através dos algoritmos da ferramenta, agruparam os alunos. Por fim, foi realizada a fase de interpretação e avaliação dos dados (Fase 4), classificando-os com base em alguns atributos. Tal classificação teve a finalidade de inferir, a partir da análise dos dados da etapa “intervenção” do experimento, qual será o comportamento apresentado pelo aluno na etapa “pós-teste”. As quatro fases KDD são descritas com mais detalhes nas seções seguintes.

6.1. Fase 1: seleção dos dados

A primeira fase do KDD é a seleção dos dados. Para apoiar esta fase foram utilizados os resultados do mapeamento sistemático apresentados no capítulo 5, além de discussões com o especialista do sistema (responsável pelo experimento com exemplos errôneos no MathTutor) e alguns testes realizados na ferramenta escolhida. Esse procedimento permitiu descobrir e compreender quais atributos possuem maior potencial para obter os melhores resultados na mineração de dados, isto é, obter informações que resultem em conhecimento relevante sobre o comportamento dos alunos. Assim, a partir dos atributos disponíveis na Tabela 3, foram criados dois subconjuntos de atributos, apresentados nas Tabela 8 e Tabela 9.

O primeiro dos subconjuntos (Tabela 8) foi utilizado para potencializar a Fase 3 de “mineração dos dados”, uma vez que apresenta os atributos mais relevantes para essa fase. Para selecionar esses atributos algumas características foram utilizadas como critérios de exclusão:

- Repetição: quando outro campo já contém os dados do campo em questão;
- Vazio: quando uma grande quantidade dos dados referentes a este campo não possui nenhum valor; e,
- Representatividade: quando o campo não traz nenhuma informação importante para detectar os comportamentos estudados.

Excluir os campos de acordo com os critérios acima descritos torna a mineração dos dados mais eficiente, uma vez que dados irrelevantes para os objetivos de pesquisa deixaram de ser utilizados, evitando retrabalhos.

Tabela 8 - Atributos selecionados para a fase de “mineração de dados”

Atributo	Descrição
Atributo	Descrição
<i>Anon Student Id</i>	Código do Estudante
<i>Duration (sec)</i>	Duração da Interação
<i>Level(Assignment)</i>	Tipo e ordem dos exercícios
<i>Problem Name</i>	Nome do exercício
<i>Problem View</i>	Número de vezes que resolveu esse exercício
<i>Problem Start Time</i>	Horário de início de resolução do exercício
<i>Step Name</i>	Nome do passo em que está
<i>Attempt At Step</i>	Número de tentativas
<i>Outcome</i>	Saída do exercício (" <i>Correct</i> ", " <i>Incorrect</i> " ou "vazio")
<i>Selection</i>	Nome do problema

O segundo subconjunto de atributos (Tabela 9), resultante da discussão com o especialista do experimento e também da análise do resultado do mapeamento sistemático e de alguns testes realizados para descobrir quais atributos agrupavam os alunos de forma mais adequada, foi utilizado para viabilizar a Fase 4 de “interpretação e a avaliação dos dados”.

Tabela 9 - Atributos selecionados para interpretação do conhecimento

Atributo	Descrição
Pós-teste	Porcentagem de acerto do aluno na fase “pós-teste imediato”
Tempo	Tempo gasto pelo aluno interagindo com o sistema
Interações	Quantidade de interações com o sistema realizadas pelo aluno
Misconception	Quantidade de <i>misconceptions</i> que aluno cometeu no “pré-teste”
Idade	Idade do aluno
Série	Série do aluno
Sexo	Sexo do aluno

Notas	O quanto o aluno se acha competente de modo geral
Conhecimento em decimais	O quanto o aluno se acha competente ao trabalhar com números decimais
Desempenho escolar	Como o aluno avalia o seu desempenho escolar

Dessa forma, foram selecionados todos os atributos e os respectivos dados a serem tratados na fase de “mineração de dados” (Fase 3) e na fase de “interpretação e avaliação dos dados” (Fase 4).

Como pode ser visto na Tabela 10, dentre os atributos mais utilizados em trabalhos de EDM, apenas “Fórum” não foi tratado no Mapeamento Sistemático, uma vez que a ferramenta utilizada (MathTutor) é um ITS e portanto não oferece suporte para esse tipo de ambiente.

Tabela 10 - Atributos mais utilizados nas pesquisas de EDM

Atributos mais citados nas pesquisas de EDM	Atributos similares utilizados nessa pesquisa
Tempo	Duração da Interação, Tempo
Tarefa	Tipo e Ordem dos Exercícios, Nome do Problema, Nome do Exercício, Número de vezes que resolveu esse Exercício,
Curso	Série, Tipo e Ordem dos Exercícios
Nota	Pós Teste, Conhecimento em Informática, Conhecimento em Decimais, Desempenho Escolar
Dado Pessoal	Código do Estudante, Idade, Sexo
Sessão	Horário de Início da resolução do Exercício
Tentativa	Número de vezes que resolveu esse Exercício
Fórum	

Contudo, antes do tratamento dos dados nas fases supracitadas, foi necessário prepará-los, conforme procedimentos da fase de “pré-processamento dos dados” (Fase 2), descritos a seguir.

6.2. Fase 2: Pré-processamento dos Dados

Esta fase do KDD teve como objetivo preparar os dados para que pudessem ser inseridos na ferramenta de DM. Para isso, foram aplicadas na prática as alterações determinadas na fase de seleção, de forma que os atributos selecionados (Tabela 8) permaneceram na planilha de *log* (seção 4.2) e aqueles julgados como dispensáveis foram retirados. Esse procedimento retrata o processo de limpeza e redução dos dados, no qual foram realizadas formatações dos dados, como a troca de vírgula (“,”) por ponto (“.”) e a padronização das datas para o formato “dd/mm/yyyy”. Após isso, os dados foram organizados para que pudessem ser inseridos na ferramenta DAMICORE. Tal organização se deu pela separação dos alunos de acordo com o tipo de intervenção (seção 4.1.3) pela qual cada um passou, gerando, dessa forma, três agrupamentos por escola (ErrEx, Adaptative e PS).

Buscando otimizar o trabalho, procedimentos que automatizam todas as tarefas desta fase foram desenvolvidos na linguagem de programação Java. Em um primeiro momento, a partir da seleção de alguns parâmetros de entrada (local das planilhas de *log* e local de saída dos arquivos prontos para fase de DM), as planilhas de *log* foram trabalhadas para gerar uma planilha final contendo apenas os atributos selecionados na fase de “seleção” (seção 2.1.1). Em seguida, todo o conteúdo foi filtrado realizando as alterações de acerto de pontuação e data. Por fim, os alunos foram separados de acordo com a intervenção que receberam e as planilhas referentes a cada um dos alunos foram armazenadas em seus respectivos diretórios (ErrEx, Adaptative e PS).

Tendo em vista que a DAMICORE não necessita de um tipo específico de arquivos para realizar a mineração de dados (arquivos de texto, planilhas ou imagens), ou seja, qualquer conjunto de arquivos pode ser utilizado, desde que cada indivíduo esteja representado em um arquivo diferente e que todos os arquivos sejam de um mesmo tipo, a fase de DM pode ser realizada sem maiores dificuldades dentro de cada um dos diretórios criados, possibilitando que se seguisse para a Fase 3, “mineração de dados” sem a necessidade da fase de “transformação”.

6.3. Fase 3: mineração dos Dados

Após as fases de “seleção” (Fase 1) e “pré-processamento” (Fase 2), deu-se início à fase de “mineração dos dados” (Fase 3), na qual os dados já selecionados foram inseridos na DAMICORE e separados por cada escola participante (Escola B, Escola A e Steel Valley) e

pelo tipo de intervenção (ErrEx, Adaptative e PS), conforme apresentado na seção 2. Dessa forma foram geradas as “árvores da intervenção”, que agrupam os alunos de acordo com suas características conforme mostrado na Figura 12. A estrutura de cada árvore gerada é composta por “folhas”, as quais representam cada um dos alunos por meio de seus respectivos códigos de identificação, e “ramos” (conjunto de folhas), que representam os agrupamentos dos alunos criados pela DAMICORE.

Figura 12 - Árvore da intervenção gerada pela DAMICORE

Ao final da fase “mineração de dados” (Fase 3), nove árvores foram geradas para classificar os alunos das escolas Escola A, Escola B e Escola C para cada tipo de intervenção (Adaptative, ErrEx e PS) como mostra a Figura 13. Cada um desses grupos, representados por uma escola e um tipo de intervenção, é denominado “turma”.

A partir da análise das nove árvores da intervenção geradas pela DAMICORE, foi possível encontrar evidências sobre possíveis comportamentos apresentados pelos alunos quando interagem com o sistema tutor inteligente. Com isso, objetivou-se descobrir quais foram os atributos utilizados pela DAMICORE para criação dos agrupamentos verificando os elementos que tornam dois ou mais indivíduos similares (agrupados em um mesmo ramo), além dos elementos que tornam dois ou mais indivíduos diferentes (agrupados em ramos da árvore fortemente separados entre si). Essas análises são apresentadas na seção 6.4, na qual é descrita a Fase 4 “interpretação e avaliação dos dados”.

Figura 13 - As nove árvores geradas pela DAMICORE

6.4. Fase 4: interpretação e avaliação dos dados

Para realizar a última fase do KDD, na qual os dados são interpretados e avaliados de forma a caracterizar os alunos, foi necessário propor métodos para analisar as árvores geradas pela DAMICORE (Figura 13), visando encontrar “ramos” (agrupamento de alunos) que trouxessem informações relevantes sobre os comportamentos dos alunos. Deste modo, primeiramente verificou-se o desempenho dos alunos conforme apresentado na Tabela 11, na

qual se pode observar uma fração da planilha de dados em que é armazenado o percentual de desempenho dos alunos de uma determinada escola em cada um dos testes do experimento.

Dessa planilha foram escolhidos para avaliar os resultados da fase de DM os dados referentes à coluna “pós-teste imediato”, que são dados provenientes de uma etapa posterior à etapa de “intervenção” e foram utilizados para verificar se em cada um dos ramos das árvores da intervenção geradas pela DAMICORE havia alunos com um desempenho semelhante ao obtido no “pós-teste imediato”. Isso provaria que a DAMICORE é capaz de prever padrões comportamentais dos alunos, como prever quais alunos irão apresentar bom rendimento no “pós-teste imediato”, e quais alunos irão apresentar mau desempenho no “pós teste imediato”.

Deste modo, conforme ilustrado na Figura 14, procurou-se descobrir se o percentual de desempenho obtido pelos alunos no “pós-teste imediato” pode se relacionar com o agrupamento dos alunos representado por cada ramo da árvore gerada a partir dos dados da “intervenção”.

Figura 14 - Árvore da intervenção com porcentagens de desempenho do "pós-teste imediato"

Assim, os dados do “pós-teste imediato” foram utilizados da seguinte forma: o desempenho (em porcentagem) de cada aluno no “pós-teste imediato”, assim como sua posição na árvore gerada pela DAMICORE foram avaliados, como pode ser observado nos dois ramos circutados em vermelho na Figura 14. Após a avaliação de todos os alunos da forma supracitada, foi possível relacionar o desempenho do aluno no “pós-teste imediato” com sua posição (“folha”) da árvore.

Em seguida, foram avaliados os “ramos”, visando identificar se neles existia algum padrão, ou seja, se a maioria dos alunos agrupados em determinado ramo apresentou o mesmo comportamento no “pós-teste imediato”. Caso fosse observado algum padrão, seria possível afirmar que a DAMICORE estava inferindo algum comportamento dos alunos.

Tabela 11 – Fração dos dados da planilha de desempenho dos alunos nos testes

Aluno	Pré- Teste	Pós- Teste imediato	Teste de Retenção
Stu_3b44dad3f5f02569842648314f3fff58	24%	26%	10%
Stu_093094291041497d02bf245519cfe6d2	58%	78%	82%
Stu_8f7e95d9997f7d100b66fafcca94f2c1	56%	56%	40%
Stu_0ac04da4a49722789f4ccfc44238d6b4	34%	26%	0%
Stu_e479d00f3fefe397fc3a5b6a891d619d	54%	66%	74%
Stu_0872de89ae081f549bf5fc7bbebb8a15	42%	30%	38%
Stu_c1824597c7b72e54956748744ab7a909	34%	30%	44%
Stu_27d8ce0717bcd7ec3cfe826978efd35	34%	28%	36%
Stu_6e016b7e345c75d5638469f09453655f	32%	68%	68%
Stu_65eafccdb2546a5e5eec9ae008dd2341	36%	38%	48%
Stu_89d032f6c405f9dae7c55a80a21dedd8	38%	46%	52%
Stu_2d09adda68ea56de9eacaf9dad558a44	30%	48%	58%
Stu_042bd140fd520923457dfe78db5d2824	40%	68%	76%
Stu_6752849051cf8c1a04ea56ffa6eb2258	30%	22%	44%
Stu_ce9f23cf1143f0b0ccc752b26203d777	38%	36%	36%
Stu_dc36d04de7311cae9d038586ff3d5a87	52%	54%	62%
Stu_274c7089040344e8412c918736596ca5	46%	32%	46%
Stu_4e2767ecf231c6b6a205c6951e1fb469	60%	76%	82%

Para obter uma avaliação mais precisa e mais sistemática dos agrupamentos gerados pela DAMICORE (“ramos”) e sua relação com o desempenho do aluno no “pós-teste imediato”, alguns intervalos foram criados para classificar o desempenho obtido pelos alunos nessa etapa do experimento. Para isso, calcularam-se a média e o desvio padrão para cada uma das turmas com relação ao desempenho obtido por cada aluno no “pós-teste imediato”. A partir desse cálculo, foram criados os intervalos para classificar os alunos, sendo eles:

“Abaixo do Desvio Padrão”, “Valor Intermediário Abaixo da Média”, “Média”, “Valor Intermediário Acima da Média” e “Acima do Desvio padrão”.

Avaliando os “ramos” das árvores geradas pela DAMICORE e verificando, de acordo com os intervalos criados, quais alunos se encontravam nesses “ramos”, buscou-se por padrões mais específicos que revelassem, de forma bastante simplificada, o comportamento dos alunos.

Na Tabela 12 é possível observar em “*Anon Study ID*” os códigos referentes a cada um dos alunos e suas respectivas taxas de acerto em “Pós-teste imediato”. A seguir, tem-se o cálculo da média e do desvio padrão em relação ao desempenho desse recorte de alunos obtido no “pós-teste imediato”. Ao final da tabela, são exibidos os seis intervalos criados, seus respectivos valores e os cálculos a partir da média e do desvio padrão utilizados para se chegar a esses valores.

Tabela 12 - Exemplo de cálculo de valores para classificação dos alunos

	Anon Study ID	Pós-Teste Imediato		
	Stu_2569842648314f3fff58	26%		
	Stu_497d02bf245519cfe6d2	78%		
	Stu_7d100b66fafcca94f2c1	56%		
	Stu_22789f4ccfc44238d6b4	26%		
	Stu_e397fc3a5b6a891d619d	66%		
	Stu_1f549bf5fc7bbebb8a15	30%		
	Stu_2e54956748744ab7a909	30%		
	Stu_db7ec3cfe826978efd35	28%		
	Stu_75d5638469f09453655f	68%		
	Média	42%		
	Desvio Padrão	18%		
Abaixo do Desvio Padrão	Valor Intermediário Abaixo da Média	Média	Valor Intermediário Acima da Média	Acima do Desvio Padrão
24%	36%	42%	51%	60%
Média - Desvio Padrão	Abaixo do Desvio Padrão + (Média - Abaixo do Desvio Padrão) / 2	Média	(Média + Acima do Desvio Padrão) / 2 - Média	Média + Desvio Padrão

Com base nesses intervalos, as “folhas” das árvores geradas pela DAMICORE foram coloridas para auxiliar a visualização dos desempenhos obtidos no “pós-teste imediato” por alunos que foram agrupados em um mesmo “ramo”, de modo que se torne mais simples e intuitiva a observação da possível relação entre os “ramos” das árvores geradas com os dados da “intervenção” e o desempenho obtido no “pós-teste imediato” por cada aluno ali agrupado. Na Figura 15 estão as cores que identificam cada um dos intervalos.

Figura 15 - Intervalo de cores utilizadas para classificação das árvores

De acordo com os intervalos da Figura 15, a cor azul representa alto, grande ou muito e a cor vermelha representa baixo, pequeno ou pouco, seguindo o mesmo padrão para as cores intermediárias.

O agrupamento de cores em determinados pontos das árvores demonstra que a DAMICORE agrupou os alunos com comportamentos semelhantes. O primeiro comportamento estudado foi o desempenho do aluno no "pós-teste imediato", e, se o resultado obtido consistir em “ramos” contendo alunos coloridos com cores iguais ou próximas (de acordo com a Figura 15), pode-se afirmar que alunos com desempenho similar foram agrupados em um mesmo "ramo".

Para os atributos que não possuem forma numérica explícita, “idade”, “série” e “sexo”, o padrão estabelecido foi: a cor azul representa alunos com mais idade e pertencentes a séries mais altas e a cor vermelha representa alunos com menos idade e pertencentes a séries mais baixas. Já para o atributo “sexo”, para tornar mais intuitivo, a cor azul foi estabelecida

Para conseguir interpretar os dados das árvores, foi avaliado o segundo subconjunto dos atributos selecionados, apresentados na Tabela 9. O objetivo de tal avaliação foi analisar cada um dos atributos inseridos na Tabela 9 da mesma forma que foram avaliados os dados do “pós-teste imediato”. Para isso a planilha de *log* foi consultada e as informações obtidas na etapa de “intervenção” referentes a cada um dos alunos foram coletadas (a Tabela 13 representa uma fração dessa planilha) e utilizadas para colorir cada uma das folhas em suas respectivas árvores, fornecendo uma árvore capaz de relatar como ocorreram os agrupamentos dos alunos. Por exemplo, após colorir cada uma das folhas utilizando o atributo “tempo”, como pode ser visto na Figura 17, pode-se afirmar que determinado “ramo” agrupa alunos que gastaram muito tempo para realizar as atividades na etapa da “intervenção” (folhas coloridas de azul, roxo ou verde), ou que gastaram pouco tempo (folhas coloridas de vermelho, laranja ou amarelo) e com isso concluir se em determinado ramo existe um comportamento comum, como baixo tempo, alto tempo, etc.

Dessa forma, o restante dos atributos também foi utilizado para colorir as árvores. Assim, conforme mostrado na Figura 17, o atributo “interações”, por exemplo, coloriu as folhas de azul, roxo ou verde quando os alunos realizaram muitas interações, enquanto coloriu de vermelho, laranja ou amarelo quando poucas interações foram realizadas. Outro exemplo se dá de acordo com o “sexo” dos alunos, em que as folhas que representam os alunos do sexo masculino foram coloridas de azul e as que representam alunos do sexo feminino foram coloridas de vermelho.

Para todas as árvores que foram coloridas a partir da análise individual dos atributos (Figura 17), buscou-se analisar se havia semelhança entre o agrupamento observado nas árvores coloridas de acordo com os dados do “pós-teste imediato” com as árvores coloridas de acordo com cada atributo apresentado na Tabela 9. Dentre os resultados obtidos, verificou-se que é possível identificar relação entre os “ramos” coloridos a partir dos dados do “pós-teste imediato” e os “ramos” coloridos a partir dos valores dos atributos encontrados na Tabela 9. Por exemplo, foram encontrados alunos com poucas interações em um “ramo” no qual estão agrupados os alunos com bom desempenho no “pós-teste imediato”, assim como foram encontrados alunos do sexo feminino nos “ramos” de alunos com um bom desempenho no “pós-teste imediato”.

Tabela 13 - Recorte da planilha com as informações dos alunos para cada um dos atributos utilizados na fase de interpretação e avaliação

Aluno	Pós-Teste	Tempo	Interações	Idade	Sexo	Serie	Conhecimento em Matematica	Conhecimento em Números Decimais	Conhecimento em Computadores
Stu_23457dfe78db5d	40%	7.239	1185	11	male	6	neutral	neutral	strongly agree
Stu_4727eb2d62ddb7	42%	4.550	714	11	female	6	agree	neutral	agree
Stu_789f4ccfc44238	58%	7.551	2592	12	female	6	neutral	neutral	agree
Stu_83d89bde2ed557	34%	3.303	651	12	male	6	strongly agree	strongly agree	agree
Stu_078a154a21c27d	46%	4.558	684	11	female	6	strongly agree	agree	strongly agree
Stu_b646094180c7ec	34%	2.735	705	11	female	6	strongly agree	neutral	strongly agree
Stu_fbce3a0eccec0f	30%	4.338	796	11	male	6	agree	agree	strongly agree
Stu_01eae1ce75008f	24%	3.802	801	11	female	6	agree	neutral	strongly agree
Stu_de9eacaf9dad55	60%	6.767	1164	11	male	6	agree	disagree	strongly agree
Stu_fdb6dd2649cee6	36%	3.640	646	11	female	6	agree	neutral	agree
Stu_42a32a887b7acb	30%	4.017	741	12	male	6	agree	agree	strongly agree
Stu_69842648314f3f	32%	6.284	1725	11	male	6	disagree	neutral	strongly agree
Stu_b261fa5078486f	38%	3.516	611	11	female	6	strongly agree	strongly agree	strongly agree
Stu_c2c94483a3c23a	56%	4.685	787	11	female	6	neutral	neutral	strongly agree
Stu_7c2142908a1013	34%	3.998	965	11	male	6	strongly disagree	strongly disagree	strongly disagree
Stu_8e5fe28ab8ab25	52%	5.631	1002	11	female	6	strongly agree	agree	strongly agree

Steel Valey - ADP

Figura 17 - Árvore classificadas de acordo com os atributos da Tabela 9

Essa comparação foi feita “ramo” a “ramo” para todas as árvores, utilizando cada um dos atributos. O objetivo foi avaliar se em cada um dos “ramos”, identificados pela análise do desempenho do aluno no “pós-teste imediato”, havia um comportamento comum apresentado pelos alunos pertencentes a um mesmo “ramo”.

Dentre os resultados obtidos, verificou-se que é possível identificar relação entre os “ramos” coloridos a partir dos dados do “pós-teste imediato” e os “ramos” coloridos a partir dos valores dos atributos encontrados na Tabela 9. Por exemplo, foram encontrados alunos com poucas interações em um ramo no qual estão agrupados os alunos com bom desempenho no “pós-teste imediato”, assim como foram encontrados alunos do sexo masculino nos “ramos” de alunos com um mau desempenho no “pós-teste imediato”.

Deste modo, os “ramos” circulados na Figura 18 podem ser avaliados em busca das relações exemplificadas acima, que, quando identificadas e analisadas, podem criar um entendimento para a forma com que a DAMICORE realizou os agrupamentos e permitir afimar que o resultado gerado (Figura 18) tem potencial para inferir se alguma característica ou comportamento influencia no desempenho do aluno.

Para inferir essa influência, foi necessário classificar e avaliar todos os “ramos” de forma a identificar padrões que se repetissem dentre as turmas estudadas, trazendo assim, informações relevantes para o entendimento dos critérios utilizados pela DAMICORE para agrupar alunos em “ramos” quando realiza a criação das árvores.

A próxima tarefa realizada consistiu em definir em qual intervalo (dentre os descritos na Figura 15) se encontram os “ramos” selecionados nas árvores coloridas de acordo com os atributos da Tabela 9, bem como definir aqueles que não pertencem a nenhuma faixa, de modo a criar uma forma sistemática para avaliar os “ramos”. Para isso uma fórmula foi criada, pela qual se classificam os “ramos” que possuem uma maioria de indivíduos com comportamento semelhante apresentado no “pós-teste imediato”.

Assim essa fórmula foi elaborada e calibrada para respeitar os ramos identificados nas árvores coloridas com o desempenho dos alunos no “pós-teste imediato” de modo que as árvores coloridas a partir dos comportamentos dos alunos perante os outros atributos poderão ter seus “ramos” identificados a partir da fórmula já testada e calibrada, o que tornará o processo sistemático.

Figura 18 - Árvores com “ramos” selecionados

Essa fórmula consiste no seguinte cálculo:

- Verifica-se no “ramo” (R) de interesse qual cor é a maioria;
 - Se houver empate, a métrica é aplicada para ambas as cores e, a que obtiver o maior resultado é a escolhida;
 - Se houver novo empate, a cor mais próxima do azul conforme intervalos apresentados na Figura 15 é a escolhida.
- Multiplica-se o número de “folhas” que possuem a cor que é maioria por 2 (esse será o peso inicial (Pi));
- Para cada “folha” que possua a cor vizinha direta conforme Figura 15 da cor que é maioria ou para as duas cores mais próximas apenas nos casos das cores vermelha e azul, soma-se 1 (V1 e V2);
- Para o restante dos elementos, que não são possuem as cores citadas anteriormente, subtrai-se 1 (V3, V4 e V5);
- Divide-se o resultado pelo número de elementos do “ramo”;
- O valor resultante é o que representa o “ramo”.

Ou seja:

$$R = (P_i + V_1 + V_2 - V_3 - V_4 - V_5) / N$$

Em que:

R = Ramo de interesse a ser classificado

P_i = Número de “folhas” com a cor que é maioria em um “ramo”, multiplicado por 2;

V_n = Número de “folhas” com a cor vizinha à cor identificada como majoritária em P_i, sendo 1 e 2 as duas cores mais próximos à cor identificada em P_i e 3, 4 e 5 as mais distantes;

N = Número de “folhas” no “ramo”.

Por exemplo, a árvore exibida na Figura 16 retrata o desempenho dos alunos na etapa da “intervenção”, enquanto as “folhas” coloridas (de acordo com os intervalos estabelecidos na Figura 15) representam os desempenhos que os alunos obtiveram no “pós-teste imediato”. Assim, os “ramos” identificados com o número 1 e 2 (roxo e azul) representam os alunos que

obtiveram alto desempenho no “pós-teste imediato” (com graus diferentes entre si) e o “ramo” com o número 3 (laranja) representa os alunos que obtiveram baixo desempenho no “pós-teste imediato”. Embora seja simples de se notar quais são as cores predominantes em cada um dos “ramos” supracitados, o mesmo não acontece para todos os outros “ramos” aqui estudados e por isso a necessidade da implementação da fórmula desenvolvida.

Para verificar se o valor resultante da fórmula indicava realmente determinado comportamento, foram realizados os cálculos para todos os ramos de cada turma e, em seguida, foi calculado o desvio padrão. Assim, para determinado ramo, se o resultado calculado pela fórmula ficasse acima do desvio padrão, esse ramo era considerado apto a ser colorido com a cor majoritária (i.e. alunos pertencentes àquele ramo se enquadravam em determinado comportamento).

Desta forma, aplicando a fórmula para o ramo 1 (R) da Figura 16, encontram-se os seguintes resultados (Tabela 14):

Tabela 14 - Cálculo da fórmula para classificar o “ramo” 1 da Figura 16

Cor	Quantidade de “folhas”	Operação	Total
Roxa (Maioria)	4	multiplica por 2	8
Verde	4	soma 4	4
Azul	1	soma 1	1
Amarela	1	subtrai 1	1
Laranja	1	subtrai 1	1
Vermelha	0	subtrai 0	0
Número de “folhas” no “ramo”	11	Soma	11
		Final	$11/8 = 1,375$

Conforme demonstrado pela Tabela 14, é possível observar que o “ramo” 1 (R) da Figura 16 possui 5 “folhas” azuis, 2 roxas, 1 verde, nenhuma amarela, laranja ou vermelha. Assim, o resultado final obtido aplicando a fórmula em questão é dado por:

Pi → Azul, cor majoritária dentre as “folhas” do Ramo 1: $5 \times 2 = 10$

V1 → Roxa, cor vizinha à Pi: $2 \times 1 = 2$

V2 → Verde, cor vizinha à Pi: $1 \times 1 = 1$

V3 → Amarela, cor distante à Pi: $0 \times (-1) = 0$

V4 → Laranja, cor distante à Pi: $0 \times (-1) = 0$

V5 → Vermelha, cor distante à Pi: $0 \times (-1) = 0$

Conforme a fórmula desenvolvida, a soma dos resultados de Pi a V2 seguida da subtração de V3 a V5 nos leva ao valor 13, que, quando dividido pelo número N de “folhas” no “ramo” (8), nos traz o valor resultante para o “ramo” estudado, sendo $R = 1,625$. Ao comparar esse resultado com o desvio padrão encontrado para essa turma (0,76), nota-se que o valor R é maior que o desvio padrão e, portanto, é possível colorir o “ramo” 1 de azul (cor que colore a maioria das folhas nesse “ramo”) significando que o “ramo” 1 pode ser classificado por agrupar alunos que apresentaram muito bom desempenho no “pós-teste imediato”.

A fórmula desenvolvida para classificação dos “ramos” foi aplicada para classificar o “ramo” 2 da Figura 16, que possui 1 “folha” azul, 4 roxas, 4 verdes, 1 amarela, 1 laranja e nenhuma vermelha. Como neste caso há empate entre as cores majoritárias (4 roxas e 4 verdes), é necessário aplicar a fórmula separadamente para cada uma das duas cores e verificar quem possui o maior resultado.

Na Tabela 15, é apresentado o cálculo feito para a cor roxa:

Tabela 15 - Cálculo da fórmula para a cor roxa na classificação do “ramo” 2 da Figura 16

Cor	Quantidade de “folhas”	Operação	Total
Roxa (Maioria)	4	multiplica por 2	8
Verde	4	soma 4	4
Azul	1	soma 1	1
Amarela	1	subtrai 1	1
Laranja	1	subtrai 1	1
Vermelha	0	subtrai 0	0

Número de “folhas” no “ramo”	11	Soma	11
		Final	11/8 = 1,375

Na Tabela 16, é apresentado o cálculo feito para a cor verde:

Tabela 16 - Cálculo da fórmula para a cor verde na classificação do “ramo” 2 da Figura 16

Cor	Quantidade de “folhas”	Operação	Total
Verde (Maioria)	4	multiplica por 2	8
Roxa	4	soma 2	4
Amarela	1	soma 1	1
Azul	1	subtrai 1	1
Laranja	1	subtrai 1	1
Vermelha	0	subtrai 0	0
Número de “folhas” no “ramo”	11	Soma	11
		Final	11/8 = 1,375

Ao refazer os cálculos, o resultado final obtido também foi de 1,375, mostrando que também houve empate no cálculo da métrica para as duas cores. Conforme definido durante a criação da fórmula, em caso de segundo empate entre as cores majoritárias, a cor mais próxima do azul é a cor selecionada e, portanto, para a classificação do “ramo” 2 da Figura 16, a cor mais próxima do azul dentre o verde e o roxo é a cor roxa. Após definição de qual cor é a majoritária no “ramo” 2, sabendo que seu $R = 1,375$ é maior que o desvio padrão calculado para essa turma (0,76) o “ramo” é então colorido de roxo, significando que o “ramo 2” da Figura 16 pode ser classificado por agrupar alunos que apresentam bom desempenho no “pós-teste imediato”.

Por fim, para classificar o “ramo” 3 da Figura 16, a aplicação da fórmula desenvolvida é mostrada abaixo, conforme Tabela 17:

Tabela 17 - Cálculo da fórmula para classificação do “ramo” 3 da Figura 16

Cor	Quantidade de “folhas”	Operação	Total
Laranja (Maioria)	3	multiplica por 2	6
Vermelha	2	soma 2	2
Amarela	2	soma 2	2
Verde	0	subtrai 0	0
Roxa	0	subtrai 0	0
Azul	0	subtrai 0	0
Número de “folhas” no “ramo”	7	Soma	10
		Final	10/8 = 1,25

Como demonstrado na Tabela 17, o valor R para o “ramo” 3 é de 1,25 e, por se apresentar maior que o desvio padrão (0,76) calculado para a turma, o “ramo” 3 da Figura 16 é então colorido de laranja, significando que pode ser classificado por agrupar alunos que apresentam baixo desempenho no “pós-teste imediato”.

Conforme explicitado acima, os “ramos” de todas as árvores geradas durante a fase de “mineração de dados” pela DAMICORE foram coloridos com as respectivas cores que representam o seu comportamento. Para “ramos” que não possuíam um comportamento predominante, manteve-se a cor preta como coloração. Na Figura 19, são exibidas as árvores geradas pela DAMICORE de acordo com o desempenho dos alunos da escola Escola A no “pós-teste imediato” que foram coloridas conforme o desempenho apresentado pelos alunos em cada atributo exibido na Tabela 9, e, em seguida, seus “ramos” foram coloridos após a aplicação da fórmula para cada um deles.

Figura 19 - Exemplo de “ramos” classificados - escola Escola A, tipo de Intervenção Adaptive

Como exemplificado na Figura 19, um mesmo “ramo” foi classificado para cada atributo exibido na Tabela 9 por diferentes cores, significando que um “ramo” pode apresentar um comportamento para determinada característica dos alunos ali agrupados, o que indica que, a partir dos dados da “intervenção”, foram identificados determinados comportamentos no “pós-teste imediato”, a partir dos quais buscou-se encontrar padrões em características relevantes desses mesmos alunos.

A partir da representação conforme Figura 19, foi possível verificar como cada um dos atributos analisados ou a combinação de alguns deles influenciou no desempenho dos alunos. Um exemplo está ilustrado na Figura 20, na qual a árvore da esquerda contempla os alunos agrupados de acordo com o desempenho no “pós-teste imediato” enquanto a árvore da direita contempla os alunos agrupados de acordo com o “tempo” que dedicaram à interação com o sistema.

Figura 20 - Comparação entre árvores classificadas

Analisando a Figura 20, pode-se notar que os alunos que apresentaram desempenho acima da média (Figura 20, árvore esquerda, ramo 1) interagiram menos tempo com o sistema (Figura 20, árvore direita, ramo 1) e os alunos com desempenho abaixo da média (Figura 20, árvore esquerda, ramo 2) interagiram mais tempo com o sistema (Figura 20, árvore direita, ramo 2).

Buscando por padrões semelhantes ao exemplificado no parágrafo anterior, avaliaram-se “ramos” das árvores coloridas de acordo com o desempenho dos alunos apresentado no “pós-teste imediato” visando encontrar semelhanças dentre os comportamentos dos alunos para cada um dos ramos coloridos nas árvores do “pós-teste imediato”. Com isso, poderá ser inferido qual o comportamento de alunos que obtiveram bom ou mau desempenho no “pós-teste imediato”. Por exemplo, foi analisado para alunos que apresentaram desempenho acima da média no “pós-teste imediato”, qual foi seu número de interações com o sistema, quanto tempo foi dedicado a cada interação, número de *misconceptions* no “pós-teste imediato”, qual seu sexo, idade e série e como o aluno considerava seu desempenho em matemática ou computação.

Deste modo, a busca por esses padrões revelou algumas características frequentes que influenciam o desempenho do aluno, como o número de interações com o sistema, o tempo dedicado a cada interação e o número de *misconceptions* no desempenho do aluno no "pós-teste imediato". Além disso, certa tendência relacionada ao sexo do aluno também foi notada.

Com base na observação desses padrões, algumas hipóteses foram criadas a fim de entender e explicar o agrupamento dos alunos realizado pela DAMICORE. Essas hipóteses foram elaboradas testando os quatro atributos que mostraram ser mais influentes na determinação do comportamento dos alunos no "pós-teste imediato" e, assim, no seu agrupamento em “ramos”. Os quatro atributos são: tempo, número de interações, *misconceptions* e sexo.

Com isso, após diversas discussões com o especialista do sistema e a avaliação desses padrões encontrados, as hipóteses foram criadas (Tabela 18). Essas hipóteses têm como objetivo entender os “ramos” gerados pela DAMICORE e em paralelo identificar o comportamento dos alunos com alguma prova estatística que a sustente.

Tabela 18 - Hipóteses sobre os comportamentos apresentados pelos alunos

Hipótese	Comportamento
C1 - Ótimo Desempenho	Alunos que obtiveram bom desempenho no pós-teste, apresentaram todos os seguintes comportamentos: poucas interações e baixo tempo interagindo com o sistema na intervenção e poucos <i>misconceptions</i> no pré-teste.
C2 - Péssimo Desempenho	Alunos que obtiveram mau desempenho no pós-teste apresentaram muitas interações e alto tempo interagindo com o sistema na intervenção, e

muitos *misconceptions* no pré-teste.

C3 - Gênero Dependente	Alunos que obtiveram bom desempenho no pós-teste são do sexo feminino ou alunos que obtiveram mau desempenho são do sexo masculino.
C4 – Bom Desempenho	Alunos que obtiveram bom desempenho no pós-teste apresentaram dois dos três seguintes comportamentos: poucas interações na intervenção, baixo tempo interagindo com o sistema na intervenção e poucos <i>misconceptions</i> no pré-teste.
C5 - Mau Desempenho	Alunos que obtiveram mau desempenho no pós-teste apresentaram dois dos três seguintes comportamentos: muitas interações na intervenção, alto tempo interagindo com o sistema na intervenção e muitos <i>misconceptions</i> no pré-teste.

Após a formulação das hipóteses supracitadas, fez-se necessário verificar a acurácia de cada uma delas e comprovar se de fato há relação entre os comportamentos apresentados na “intervenção” e os resultados no “pós-teste imediato”. Para isso, foi analisado o desempenho dos alunos dentro de cada escola de forma geral e de cada intervenção separadamente. Os resultados obtidos são apresentados na seção 6.5.

6.5.Resultados Obtidos

Para verificar se as hipóteses realmente se confirmavam e com que frequência essa confirmação ocorria, foram gerados gráficos que evidenciam a acurácia de cada uma delas a partir dos dados apurados por turma e pelo tipo de “intervenção” pelas quais os alunos passaram.

Os gráficos exibidos na

Figura 21, por exemplo, mostram a acurácia das hipóteses de acordo com o tipo de comportamento apresentado pelos alunos da escola Escola A nos três tipos de intervenção.

Figura 21 - Gráficos do índice de confirmação das hipóteses sobre o comportamento dos alunos na escola Escola A em cada um dos tipos de intervenção

Com base nos gráficos da

Figura 21, é possível concluir quais hipóteses tiveram sua acurácia comprovada na escola Escola A. A hipótese “Ótimo Desempenho” (C1) variou de acordo com cada uma das intervenções e obteve acurácia de 100% somente na intervenção do tipo “PS”. Já a hipótese “Péssimo Desempenho” (C2) obteve acurácia de 100% apenas na intervenção do tipo “ErrEx”, não se comprovando (0%) nas demais intervenções, “PS” e “Adp”. Já a hipótese “Gênero Dependente” (C3) obteve 67% de acurácia apenas na intervenção “Adp”, não se comprovando nas demais, “ErEx e P S”. Entretanto, as hipóteses “Bom Desempenho” (C4) e “Mau Desempenho” (C5) obtiveram acurácia de 100% em todos os tipos de intervenção, com exceção à hipótese “Mau Desempenho” (C5) que não foi comprovada na intervenção de tipo “PS” (0%).

Já para a escola Escola B, conforme apresentado nos gráficos da Figura 22, foi possível observar que a hipótese “Ótimo Desempenho” (C1) obteve acurácia de 100% apenas na intervenção “PS”, não se comprovando nas intervenções “Adp” e “ErrEx” (0%). Já com relação à hipótese “Péssimo Desempenho” (C2), o resultado referente às intervenções “ErrEx” e “Adp” foi de acurácia de 100%, enquanto que na intervenção “PS” o índice ficou em 0%.

Huston

Figura 22 - Gráficos do índice de confirmação das hipóteses sobre o comportamento dos alunos na escola Escola B em cada um dos tipos de intervenção

A hipótese “Gênero Dependente” (C3) obteve acurácia de 100% apenas na intervenção “Adp”, não se comprovando nas demais “ErrEx” e “PS” (0%). Por fim, as hipóteses “Bom Desempenho” (C4) e “Mau Desempenho” (C5) apresentaram acurácia de 100% nas intervenções “Adp” e “ErrEx”, enquanto que na intervenção “PS” apenas a hipótese “Bom Desempenho” (C4) obteve acurácia de 100%.

Na escola Escola C, conforme apresentado na Figura 23, é possível observar que a hipótese “Ótimo Desempenho” (C1) obteve acurácia de 100% em todas as intervenções.

Figura 23 - Gráficos do índice de confirmação das hipóteses sobre o comportamento dos alunos na escola Escola C em cada um dos tipos de intervenção

Com relação à hipótese “Péssimo Desempenho” (C2), a única intervenção com acurácia de 100% foi a “ErrEx”. A hipótese “Gênero Dependente” (C3) apresentou acurácia de 100% nas intervenções “Adp” e “ErrEx”. Com relação à hipótese “Bom Desempenho” (C4) obteve-se acurácia de 100% nas três intervenções realizadas. A hipótese “Mau Desempenho” (C5) obteve acurácia de 100% apenas nas intervenções “Adp” e “ErrEx”.

Ao compararmos os índices de acurácia das hipóteses nas três escolas, nota-se que a hipótese “Bom Desempenho” (C4) obteve acurácia de 100% em todas as escolas e em todas as intervenções, mostrando-se muito promissora para os resultados desta pesquisa. Além disso, na escola Steel Valley, todas as hipóteses foram comprovadas para o tipo de intervenção “ErrEx”, o que também pode ser considerado um resultado bastante interessante. É possível, portanto, analisar os gráficos gerados e verificar, para cada escola, qual o tipo de intervenção permitiu que os alunos obtivessem melhor desempenho no processo de ensino e aprendizagem.

Na Figura 24 são apresentados os gráficos resultantes do cálculo da média dos índices de acurácia obtidos em cada tipo de intervenção nas três escolas.

Figura 24 - Comportamentos dos alunos das três escolas de acordo com o tipo de intervenção

Analisando os gráficos da Figura 24, verifica-se que todas as hipóteses obtiveram determinado índice de acurácia nas intervenções “Adp” e “ErrEx”, enquanto que na intervenção “PS” apenas as hipóteses “Ótimo Desempenho” (C1) e “Bom Desempenho” (C4) tiveram sua acurácia comprovada. É possível ainda verificar que a hipótese “Bom Desempenho” (C4) obteve 100% de acurácia em todas as intervenções.

Visto que as intervenções “Adp” e “ErrEx” utilizam uma teoria de ensino e a “PS” uma outra (seção 4.1.3), essas diferenças na previsão do comportamento dos alunos podem estar relacionadas a essas diferenças teóricas. Por exemplo, como demonstrado nesta pesquisa, os alunos que participaram da intervenção “PS” dedicaram pouco tempo e realizaram poucas interações com o sistema, conforme pode ser visto pelas cores vermelha, laranja e amarela (em geral, desempenho abaixo da média) que classificam os ramos das árvores da Figura 25.

Dentro do grupo desses alunos, aqueles que obtiveram mau desempenho, juntamente com o fato de terem dedicado pouco tempo e interagido em número reduzido de vezes com o sistema, receberam pouco *feedback* do sistema tutor inteligente MathTutor, que possui ferramentas para guiar os alunos durante seu aprendizado. Assim, provavelmente esses alunos não receberam os conceitos importantes que o sistema deveria passar para eles.

Essa mesma situação não é vista nas intervenções “Adp” e “ErrEx”, uma vez que as hipóteses “Péssimo Desempenho” (C2) e “Mau Desempenho” (C4) obtiveram maior acurácia, demonstrando que os alunos interagiram mais com o sistema e, deste modo, puderam obter mais do seu potencial.

Figura 25 – Árvores de tempo e interação na intervenção “PS” nas três escolas

Em síntese, é apresentado na Figura 26, um gráfico que resume o índice geral de acurácia alcançado pelas hipóteses C1, C2, C3, C4 e C5, considerando o comportamento dos alunos das três escolas estudadas durante os três tipos de intervenção.

Figura 26 - Gráfico Comportamentos geral

O gráfico da Figura 26 evidencia que, além da hipótese “Bom Desempenho” (C4), que obteve um índice de acurácia geral de 100%, a hipótese “Mau Desempenho” (C5) também apresentou um índice elevado de acurácia geral (67%) e 100% de acurácia para alunos nas intervenções “Adp” e “ErrEx”. Portanto, ambas as hipóteses C4 e C5 são relevantes para avaliar os resultados do estudo proposto. Outra hipótese que apresentou índice de acurácia razoável foi a hipótese “Ótimo Desempenho” (C1), que obteve acurácia geral de 61% e acurácia de 100% na intervenção “PS”.

É importante destacar que, para a hipótese “Gênero Dependente” (C3), nenhum tipo de análise sobre a divergência entre número de alunos do sexo masculino e feminino foi realizada, podendo ter influenciado os resultados encontrados para essa hipótese.

No mais, os resultados alcançados possibilitaram constatar que, no contexto no qual este estudo foi realizado, alguns atributos, como o tempo que o aluno dedica à interação com o sistema, a quantidade de interações realizadas e a quantidade de *misconceptions* que o aluno possui em relação ao assunto que lhe é proposto, têm um papel importante na identificação dos comportamentos dos alunos.

Esses resultados puderam revelar formas mais adequadas para identificar e classificar agrupamentos em cada um dos tipos de intervenção. Por exemplo, conforme pode ser visto na Figura 24, para as intervenções “Adp” e “ErrEx”, as hipóteses que obtiveram alto índice de acurácia foram: “Bom Desempenho” (C4) e “Mal Desempenho” (C5). Além disso, para a

intervenção “PS” as hipóteses que obtiveram os maiores índices de acurácia foram: “Ótimo Desempenho” (C1) e “Bom Desempenho” (C4), significando que, com esses resultados, professores podem descobrir quais alunos terão dificuldades e quais terão facilidades na utilização de um determinado tipo de material de apoio para o processo de ensino e aprendizagem.

Finalmente, é importante salientar que apesar dos resultados relevantes alcançados nesta pesquisa, é necessária uma amostragem maior de dados para que a acurácia das hipóteses possa ser comprovada e as conclusões possam ser estendidas para outros contextos, domínios e ambientes de ensino.

7. Conclusão

Neste trabalho foram identificados, por meio de técnicas de EDM, padrões relevantes de comportamentos de alunos que interagiram com o sistema tutor inteligente, que disponibiliza exercícios específicos de matemática, o MathTutor. Para alcançar os objetivos definidos, foram executadas as fases do KDD, trabalhando com dados de um experimento realizado com alunos que interagiram com três versões do mesmo material no MathTutor e a ferramenta de mineração de dados (DAMICORE). De forma a retomar os objetivos desta pesquisa e as contribuições realizadas transcrevemos as questões de pesquisa presentes definidas na seção 1.2.

Q1 → Quais informações (dados) são relevantes para entender as diferenças e semelhanças entre os comportamentos manifestados por alunos do ensino fundamental e médio quando interagem com um sistema tutor inteligente?

Em um primeiro momento foi trabalhado um processo de “mapeamento sistemático” que auxiliou na seleção dos atributos para a EDM. Com o resultado desse mapeamento, e o auxílio de um especialista no domínio dos dados armazenados do experimento, foi possível responder esta questão de pesquisa. No que diz respeito aos dados que registram o comportamento dos alunos quando interagem com o sistema e permitem a realização de agrupamentos, os seguintes atributos foram selecionados como relevantes para realizar a fase de DM: "Código do Estudante", "Duração da Interação", "Tipo e ordem dos exercícios", "Nome do exercício", "Número de vezes que resolveu esse exercício", "Horário de início de resolução do exercício", "Nome do passo que está", "Número de tentativas", "Saída do exercício ("Correct", "Incorrect" ou vazio)", "Nome do problema", "Ação que o aluno efetuou". Com os atributos relevantes selecionados as demais fases do KDD foram realizadas para que conhecimentos sobre o comportamento dos alunos fossem gerados. Dessa forma, foram analisados os resultados gerados pela DAMICORE e, com isso, algumas hipóteses foram elaboradas, buscando prever qual seria o desempenho que o aluno teria nos testes finais, de acordo com o seu desempenho na fase de intervenção do experimento. As hipóteses geradas foram submetidas a testes de validação que incluíram a análise de artefatos (árvores da intervenção) e a geração de gráficos que exibem o índice de acurácia de cada uma das hipóteses. Os testes de validação das hipóteses possibilitaram responder a segunda questão de pesquisa desse trabalho.

Q2 → Quais comportamentos, manifestados por alunos do ensino fundamental e médio podem ser identificados ao analisar dados registrados em um sistema tutor inteligente no início de um curso de matemática?

Os índices de acurácia validaram algumas das hipóteses elaboradas e permitiram prever os seguintes comportamentos: Alunos que obtiveram bom desempenho no pós-teste apresentaram dois dos três comportamentos a seguir: (i) poucas interações na intervenção; (ii) baixo tempo interagindo com o sistema na intervenção; e (iii) poucos *misconceptions* no pré-teste. Foi possível prever também que alunos que obtiveram mau desempenho no pós-teste apresentaram dois dos três comportamentos a seguir: (i) muitas interações na intervenção; (ii) alto tempo interagindo com o sistema na intervenção; e (iii) muitos *misconceptions* no pré-teste. É possível concluir que com os procedimentos e técnicas adotados nessa pesquisa foi possível identificar, com algum grau de precisão, os alunos que teriam um bom ou um mau desempenho ao final do experimento. Valendo-se desses resultados podem-se encontrar indícios de que o professor pode utilizar o comportamento inicial do aluno com o sistema para intervir no processo de ensino e aprendizagem com certa antecedência, podendo auxiliar os alunos que possuem mais dificuldade ou propor novos desafios para aqueles com mais facilidade na execução das tarefas propostas.

Parte dos resultados desta pesquisa foram publicados em duas oportunidades (Moro et al., 2013) e (Moro et al., 2014). Na primeira, foram descritos os procedimentos que seriam seguidos para a realização desta pesquisa, exibindo a metodologia que seria utilizada bem como os resultados que eram visados. Na segunda, foram apresentados alguns dos resultados alcançados e a forma com que esses resultados foram obtidos.

Apesar dos resultados positivos alcançados, ainda é necessário validar as hipóteses com um grupo maior de alunos, bem como verificar se os resultados alcançados conseguem realmente auxiliar os professores (usuários leigos aos conceitos de DM) a identificarem o comportamento de seus alunos de forma mais efetiva. Para conseguir essas validações, novos experimentos, com outros grupos de alunos e em outros contextos educativos, necessitam ser realizados. Dessa forma, está previsto como trabalho futuro o desenvolvimento de um projeto de pesquisa no qual o mesmo experimento será realizado em outro contexto educativo. O objetivo desse novo projeto será o de acompanhar o desempenho dos alunos e professores envolvidos para comparar se os comportamentos manifestados pelos alunos nessa pesquisa se repetem, validando com maior precisão, os resultados encontrados, e também avaliar se os resultados gerados realmente podem auxiliar os professores na tarefa de melhor conhecer seus alunos, suas dificuldades e facilidades. É previsto, por exemplo, que nesse novo experimento

os professores participantes recebam as informações referentes aos comportamentos dos alunos a cada etapa, visando auxiliá-lo a identificar as necessidades manifestadas pelos alunos. Com essas informações à disposição, o professor poderá, durante o processo de ensino e aprendizagem auxiliar aqueles alunos com maiores dificuldades e gerar novos desafios para aqueles com maiores facilidades no tema.

Além disso, devem ser implementadas melhorias no mapeamento sistemático. A *string* de busca necessita ser desenvolvida visando encontrar um maior número de trabalhos, incluindo trabalhos importantes que não puderam ser tratados nesta pesquisa e também trabalhos provenientes de bases nacionais. Assim, com o mapeamento sistemático mais bem desenvolvido, a etapa de seleção de atributos deve ser revista e validada de acordo com os novos dados encontrados, em busca da possibilidade de melhores resultados serem alcançados.

Em relação aos recursos de apoio ao professor, considera-se interessante o desenvolvimento de uma ferramenta que apresente a ele, de forma rápida e objetiva, resultados que o auxiliem a interagir com a sua turma. Para isso, deverão ser implementadas na ferramenta de EDM, técnicas de Interação Humano-Computador que aprimorem a forma de apresentação das funcionalidades para o professor tirar o máximo proveito de sua utilização.

Outra linha com potencial para ampliar os resultados alcançados com a pesquisa aqui apresentada seria estudar as intervenções realizadas com os alunos. Um estudo dessas intervenções pode trazer resultados importantes, mostrando aos professores e pesquisadores da área, por exemplo, qual o tipo de intervenção pode ser mais adequado para determinada situação de ensino e aprendizagem.

Referencias

- ABDULLAHA, Z.; HERAWANB, T.; AHMADB, N. e DERISC, M. M. Mining significant association rules from educational data using critical relative support approach. *Porcoecdeida - Social and Behavioral Science* 28, p. 97–101, 2011.
- ALEVEN, V.; MCLAREN, B.; e SEWALL, J. Scaling up programming by demonstration for intelligent tutoring systems development: An open-access website for middle school mathematics learning. *IEEE Transactions on Learning Technologies*. V. 2(2), p. 64-78, 2009.
- ASSOCIATION FOR THE ADVANCEMENT OF ARTIFICIAL INTELLIGENCE (AAAI). Disponível em: <<http://www.aaai.org/Conferences/AAAI/aaai05.php>>. Acesso em 21 de Jan. 2013.
- BAKER, R. S. J. D. Data Mining for Education. *International Encyclopedia of Education*. E. 3, p. 7-8, 2010.
- BAKER, R. S. J. D. Data mining for education. *International encyclopedia of education*, 7, 112-118, 2010.
- BAKER, R. S. J. D. e YACEF, K. The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*. v. 1(1), p. 3-17, 2009.
- BAKER, R. S. J. D.; CORBETT A. T.; KOEDINGER K. R. e WAGNER A. Z. Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". *CHI*. 2004.
- BAKER, R. S. J. D.; GOWDA, S. M. e CORBETT, A. T. Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key. *Proceedings of the 4th International Conference on Educational Data Mining*, p. 179–188, 2011.
- BAKER, R. S. J. D.; ISOTANI, S. e CARVALHO, A. M. J. B. Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19(2), 2011.
- BAKER, R.S.J.D. in press. Data Mining For Education. In *International Encyclopedia of Education* (3rd edition), B. MCGAW, PETERSON, P., BAKER Ed. Elsevier, Oxford, UK.
- BAKER, R.S.J.D.; CORBETT, A.T.; KOEDINGER, K.R.; EVENSON, S.E.; ROLL, I.; WAGNER, A.Z.; NAIM, M.; RASPAT, J.; BAKER, D.J. e BECK, J. Adapting to When Students Game an Intelligent Tutoring System. In *Proceedings of the International Conference on Intelligent Tutoring Systems*. p. 392-401, 2006.
- BARTNECK, C. e HU, J. Scientometric Analysis Of The CHI Proceedings. *Proceedings of the Conference on Human Factors in Computing Systems*. p. 699-708, 2009.
- BAYKAL, A. Veri madenciligi uygulama alanlari, D.U. Ziya Gokalp Egitim Fakultesi Dergisi, 7, p. 95-107, 2006.
- BECK, J.E. e MOSTOW, J. How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. In *Proceedings of the International Conference on Intelligent Tutoring Systems*. p. 353-362. 2008.
- CABENA, P.; HADJINIAN, P.; STADLER, R.; JAAPVERHEES, e ZANASI, A. *Discovering Data Mining: From Concept to Implementation*. Prentice Hall, 1998.
- CARLOTTO, M. S. e PALAZZO, L. S. Síndrome de burnout e fatores associados: um estudo epidemiológico com professores. *Cadernos de Saúde Pública*, Rio de Janeiro, v. 22, n. 5, p. 1017-26, 2006.
- CARMONA, C. J.; GONZÁLEZ, P.; JESUS, M. J.; ROMERO, C. e VENTURA, S. Evolutionary algorithms for subgroup discovery applied to e-learning data. *IEEE*

- EDUCON Education Engineering – The Future of Global Learning Engineering Education. 2010.
- CARVALHO, H. L. Método de análise para a coordenação dos processos de produção sob a ótica de redes de inovação colaborativas apoiado por agente inteligente evolutivo. Tese de doutorado apresentada para a Escola de Engenharia de São Carlos. 2012.
- CHAKRABORTY, S.; ROY, D.; BHOWMICK K. P. e BASU A. An authoring system for developing Intelligent Tutoring System. Students' Technology Symposium (TechSym), IEEE, 2010.
- CILIBRASI, A. R. e VITANYI A. P. Clustering by compression. IEEE Trans. Information Theory, v. 51(4), p. 1523-1545, 2005.
- CONGRESSO INTERNACIONAL ABED DE EAD (CIED), 2011. Disponível em: <http://www.abed.org.br/congresso2012/anais/ANAIS_18CIAED_poster.pdf>. Acesso em 21 de Jan. 2013.
- COSTA, E.; BAKER R. S. J. d.; AMORIM L.; MAGALHÃES J. e MARINHO T. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. Jornada de Atualização em Informática na Educação, v. 1, p. 1-29, 2012.
- DOMINGUES, J.J.; TOSCHI, N. S. e OLIVEIRA, J. F. A reforma do Ensino Médio: a nova formulação curricular e a realidade da escola pública. Revista Educação & Sociedade n.º 70. Campinas, ano XXI, pp. 63-99, 2000.
- DRAZDILOVA, P.; SLANINOVA, K.; MARTINOVIC, J.; OBADI, G. e SNASEL, V. Creation of Students' Activities from Learning Management System and their Analysis. Conference on Computational Aspects of Social Networks, p. 155, 2009
- EDX <<https://www.edx.org/>> Acesso em 2 de Mar. 2014.
- FAYYAD, U.; PIATETSKY-SHAPIO, G. e SMYTH, P. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996.
- FELSENSTEIN J. Inferring phylogenies. Sinauer Associates. 2003.
- FLORES, M. S.; FROZZA, R.; BAGATINI, D. e CRUZ, M. E. J. K. Uso de rede neural artificial para determinar o estilo cognitivo do aprendiz. Anais do 24º Simpósio Brasileiro de Informática na Educação (SBIE). 2013.
- GARCÍA, E.; ROMERO, C.; VENTURA, S. e CASTRO, C. A collaborative educational association rule mining tool Internet and Higher Education 14, p. 77–88, 2011.
- GOGUADZE, G.; SOSNOVSKY, S.; ISOTANI, S. e MCLAREN, B.M. Evaluating a bayesian student model of decimal misconceptions. In the Proceedings of the 4th International Conference on Educational Data Mining. 2011.
- GOOGLE. <<http://www.google.com/intl/pt-BR/about/company/>> Acesso em 2 de Mar. 2014.
- GOTARDO, R.; CEREDA, P. R. M. e JUNIOR, E. R. H. Predição do Desempenho do Aluno usando Sistemas de Recomendação e Acoplamento de Classificadores. Anais do SBIE. 2013.
- GOTTARDO, E.; KAESTNER, C. E NORONHA, R.V. Avaliação de Desempenho de Estudantes em Cursos de Educação a Distância Utilizando Mineração de Dados. Anais do XXXII Congresso da Sociedade Brasileira de Computação, 2012.
- GOTTARDO, E.; KAESTNER, C. e NORONHA, R. V. Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais. Anais do 23º Simpósio Brasileiro de Informática na Educação. 2012.
- HA, S.; BAE, S.; e PARK, S. Web mining for distance education. In IEEE international conference on management of innovation and technology. p. 715–719, 2000.

- HAMALAINEN, W.; SUHONEN, J.; SUTINEN, E.; e TOIVONEN, H. Data mining in personalizing distance education courses. In World conference on open learning and distance education. 2004.
- HAN, J. e KAMBER, M. Data Mining Concepts and Techniques, Morgan Kauffmann Publishers, Second Edition. 2006.
- HERSHKOVITZ, A. e NACHMIAS, R. Developing a log-based motivation measuring tool. EDM, p. 226–233, 2008.
- HERSHKOVITZ, A.; AZRAN, R.; HARDOF-JAFFE, S. e NACHMIAS, R. Types of online hierarchical repository structures. Internet and Higher Education 14. P. 107–112, 2011.
- HERSHKOVITZ, A.; BAKER, R. S. J.; GOBERT, J.; WIXON, M. e PEDRO, M. S. Discovery With Models: A Case Study on Carelessness in Computer-Based Science Inquiry. 2013, Disponível em: <<http://abs.sagepub.com/content/57/10/1480>>. Acesso em 1 de Jan. 2013.
- HTC. <<http://www.htc.com/pt/about/>> Acesso em 2 de Mar. 2014.
- ISOTANI, S.; ADAMS, D.; MAYER, R.E.; DURKIN, K.; RITTLE-JOHNSON, B. e MCLAREN, B.M.: Can erroneous examples help middle-school students learn decimals? In: Kloos, EC-TEL 2011. V. 6964, p. 181–195, 2011.
- JOHN, G.H., KOHAVI, R., e PFLEGER, K. Irrelevant features and the subset selection problem, Machine Learning: Proceedings of the Eleventh International Conference, California: Morgan Kaufmann, p. 121-129, 1994.
- KAMPFF, A. J. C. Mineração de Dados Educacionais para Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio a Prática Docente. Tese de Doutorado apresentada à Universidade Federal do Rio Grande do Sul, 2009.
- KAY, J.; MAISONNEUVE, N.; YACEF, K. e REIMANN, P. The Big Five and Visualisations of Team Work Activity. In Proceedings of Intelligent Tutoring Systems. p. 197-206, 2006.
- KAZANIDIS, I.; VALSAMIDIS, S.; LOUKAS, A.; KONTOGIANNIS, S. e KARAKOS A. Measuring and mining LMS data. 16th Panhellenic Conference on Informatics. p. 296, 2012.
- KITCHENHAM, B. e CHARTERS, S. Guidelines for performing Systematic Literature Reviews in Software Engineering. EBSE Technical Report. 2007.
- KOEDINGER, K.R.; CUNNINGHAM, K.; A.; S. e LEBER, B. An open repository and analysis tools for fine-grained, longitudinal learner data. In Proceedings of the 1st International Conference on Educational Data Mining. p. 157-166, 2008.
- KOUTRI, M.; AVOURIS, N.; e DASKALAKI, S. A survey on web usage mining techniques for web-based adaptive hypermedia systems. 2004.
- KRAVVARIS, D.; KERMANIDIS, K. L. e THANOU, E. Success Is Hidden in the Students' Data. Artificial Intelligence Applications and Innovations IFIP Advances in Information and Communication Technology. v. 382, p. 401-410, 2012.
- KUMR, A.; RAJPAKALA, R. K.; RAGADE, J. P. e WONG*. The Virtual Learning Environment System. FIE Conference. 1998.
- LEMOES, A. e JOSGRILBERG, F. Comunicação e mobilidade aspectos socioculturais das tecnologias móveis de comunicação no Brasil. E. 1. Bahia: EDUFBA, p. 12, 2009.
- MANHÃES, L. M. B.; CRUZ, S. M. S.; COSTA, R. J. M.; ZAVALETA J. E ZIMBRÃO G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. Anais do XXII SBIE - XVII WIE. 2011.
- MARTÍN, M.; ÁLVAREZ, A.; FERNÁNDEZ-CASTRO, I. e URRETAVIZCAYA, M. Generating Teacher Adapted Suggestions for Improving Distance Educational Systems with SIgMa. Eighth IEEE International Conference on Advanced Learning Technologies. 2008.

- MARTINS, L. C.; LOPES, D. A. e RAABE, A. Um Assistente de Predição de Evasão aplicado a uma disciplina Introdutória do curso de Ciência da Computação. Anais do SBIE. 2013.
- MATHTUTOR. <<https://adapterrex.cs.cmu.edu/>> Acesso em 2 de Mar. 2014.
- MCLAREN, B.M.; ADAMS, D.; DURKIN, K.; GOGUADZE, G.; MAYER, R.E.; RITTLE-JOHNSON, B.; SOSNOVSKY, S.; ISOTANI, S. e van VELSEN, M. To err is human, to explain and correct is divine: A study of interactive erroneous examples with middle school math students. In: Ravenscroft, EC-TEL. v. 7563, p. 222–235, 2012.
- MCLAREN, B.M.; LIM, S.J. e KOEDINGER, K.R. When and how often should worked examples be given to students? New results and a summary of the current state of research. Proceedings of the 30th Annual Conference of the Cognitive Science Society. p. 2176–2181, 2008.
- MICROSOFT. <<http://www.microsoft.com/pt-br/about/default.aspx>> Acesso em 2 de Mar. 2014.
- MOODLE <<http://moodle.com/>> Acesso em 2 de Mar. 2014.
- MORO, L. F. S.; LOPES, A.Z.; DELBEM, A. C. B.; ISOTANI, S. Os Desafios para Minerar Dados Educacionais de Forma Rápida e Intuitiva: o Caso da DAMICORE e a Caracterização de Alunos em Ambientes de eLearning. In: DesafIE! - II Workshop de Desafios da Computação Aplicada à Educação, 2013, Maceió. II Workshop de Desafios da Computação Aplicada à Educação, v. 1, p. 1367-1376, 2013.
- MORO, L. F. S.; RODRIGUES, C. L.; ANDRADE F. R. H.; DELBEM, A. C. B.; ISOTANI, S. Caracterização de Alunos em Ambientes de Ensino Online: Estendendo o Uso da DAMICORE para Minerar Dados Educacionais. III Congresso Brasileiro de Informática na Educação, Workshops, WMDAVA, 2014.
- NEWMAN, M. Networks: An Introduction. Oxford University Press. 2010.
- NOKIA. <<http://br.networks.nokia.com/no-brasil/sobre-nos>> Acesso em 2 de Mar. 2014.
- OLIVEIRA, S. A. C.; SOBRINHO, C. L. N.; GAMA, R. N. e SILVA, V. N. Estudo da Prevalência dos Distúrbios Psíquicos Menores (dpm), no Bairro George Américo, Feira de Santana, Bahia. Seminário de Iniciação Científica da UEFS, 2013.
- PAIVA, J. G. S. FLORIAN-CRUZ, L. PEDRINI H.; TELLES, G. P. e MINGHIM, R. IMPROVED Similarity Trees and their Application to Visual Data Classification. IEEE Transactions on Visualization and Computer Graphics. v. 17(12), 2011.
- PETERSEN, K.; FELDT, R.; MUJTABA, S. e MATTSSON, M. Systematic Mapping Studies in Software Engineering. Proceedings of the 12th international conference on Evaluation and Assessment in Software Engineering. p. 68-77, 2008.
- PRETORIUS, R. e BUDGEN, D. A Mapping Study on Empirical Evidence Related to the Models and Forms used in the UML. In 2nd ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), ACM, p. 342–344. 2008.
- REIS, E.J.; CARVALHO, F.M.; DE ARAÚJO, T.M.; PORTO, L.A. e SILVANY NETO, A.M. Work and psychological distress among public school teachers in Vitória da Conquista, Bahia State, Brazil. Cadernos de Saúde Pública, 21(5), p. 1480-1490. 2005.
- RIGO, S. J.; CAZELLA, S. C.; E CAMBRUZZI, W. Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In Anais do Workshop de Desafios da Computação Aplicada à Educação, p. 168–177, 2012.

- ROMERO, C. e VENTURA S. Expert Systems with Applications. v. 33, p. 135–146, 2007.
- ROMERO, C.; GONZALEZ, P.; VENTURA, S.; JESUS, M. J. e HERRERA, F. Evolutionary algorithms for subgroup discovery in e-learning: A practical application using Moodle data. Expert Systems with Applications 36, p. 1632–1644, 2009.
- ROMERO, C.; VENTURA S.; ESPEJO P. G. e HERVÁS C. Data mining algorithms to classify students. Proceedings of Educational Data Mining. p. 20-21, 2008.
- ROMERO, C.; VENTURA, S. e GARCIA, E. Data mining in course management systems: Moodle case study and tutorial. Computers & Education. v. 51(1), p. 368–384, 2007.
- SANCHES, A. CARDOSO, J. M. P. e DELBEM, A. C. B. Identifying Merge-Beneficial Software Kernels for Hardware Implementation. International Conference on Reconfigurable Computing and FPGAs. p. 74-79, 2011.
- SILVA, M. P. S. Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka, 2004. Disponível em: <<http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf>>. Acesso em 01/12/2013.
- SRIVASTAVA, J.; COOLEY, R.; DESHPANDE, M. e TAN, P. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, v. 1(2), p. 12–23, 2000.
- TANG, T.; e MCCALLA, G. Smart recommendation for an evolving e-learning system. International Journal on E-Learning, 4(1), p. 105–129. 2005.
- WEB-CAT <<http://web-cat.org/>> Acesso em 2 de Mar. 2014.
- WIKIPÉDIA, 2013. Disponível em: <<http://stats.wikimedia.org/PT/TablesWikipediaPT.htm>>. Acesso em 2 de Mar. 2014.
- WIXON, M.; BAKER, R. S. J.; GOBERT, J.; OCUMPAUGH, J. e BACHMANN, M. WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously. UMAP'12 Proceedings of the 20th international conference on User Modeling, Adaptation, and Personalization. 2012.
- ZAFRA, A.; VENTURA, S. Multi-instance genetic programming for predicting student performance in web based educational environments. Applied Soft Computing 12. p. 2693–2706, 2012.
- ZAÏANE, O. e LUO, J. Web usage mining for a better web-based learning environment. Proceedings of Conference on Advanced Technology for Education. Banff, Alberta, p. 60-64, 2001.
- ZORRILLA, M.; GARCÍA, D. e ÁLVAREZ, E. A Decision Support System to improve e-Learning Environments. EDBT/ICDT Workshops, 2010.

ANEXO A – Recorte da planilha de log gerada pelo MathTutor

104

Titulo	Row	Sample Name	Anon Student Id	Session Id	Time
Aluno 1	1	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:27:17.0
Aluno 2	2	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:28:09.0
Aluno 3	3	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:28:24.0
Aluno 4	4	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:29:01.0
Aluno 5	5	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:30:12.0
Aluno 6	6	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:30:42.0
Aluno 7	7	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:30:42.0
Aluno 8	8	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:30:44.0
Aluno 9	9	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:30:46.0
Aluno 10	10	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:30:59.0
Aluno 11	11	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:31:04.0
Aluno 12	12	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:31:06.0
Aluno 13	13	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:31:10.0
Aluno 14	14	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:31:10.0
Aluno 15	15	Escola A Jr High School	Stu_002769db730fa4d61ed6f962b38eab6f	S9f4f1a821af40c9def9836470f845b1e	2012-10-03 12:31:12.0

Titulo	Time Zone	Duration (sec)	Student Response Type	Student Response Subtype	Tutor Response Type	Tutor Response Subtype	105 Level(Assignment)
Aluno 1	UTC	0	ATTEMPT				ErrEx-CBA
Aluno 2	UTC	52	ATTEMPT		RESULT		ErrEx-CBA
Aluno 3	UTC	15	ATTEMPT		RESULT		ErrEx-CBA
Aluno 4	UTC	37	ATTEMPT		RESULT		ErrEx-CBA
Aluno 5	UTC	71	ATTEMPT		RESULT		ErrEx-CBA
Aluno 6	UTC	15	ATTEMPT		RESULT		ErrEx-CBA
Aluno 7	UTC	15	ATTEMPT				ErrEx-CBA
Aluno 8	UTC	2	ATTEMPT		RESULT		ErrEx-CBA
Aluno 9	UTC	0	ATTEMPT				ErrEx-CBA
Aluno 10	UTC	13	ATTEMPT		RESULT		ErrEx-CBA
Aluno 11	UTC	5	ATTEMPT		RESULT		ErrEx-CBA
Aluno 12	UTC	2	ATTEMPT		RESULT		ErrEx-CBA
Aluno 13	UTC	2	ATTEMPT		RESULT		ErrEx-CBA
Aluno 14	UTC	2	ATTEMPT		RESULT		ErrEx-CBA
Aluno 15	UTC	0	ATTEMPT				ErrEx-CBA

Titulo	Tutor Response Subtype	Level (Assignment)	Level (ProblemSet)	Problem Name	Problem View	Problem Start Time	Step Name
Aluno 1		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-0-TST-C	1	2012-10-03 12:27:17.0	
Aluno 2		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-0-TST-C	1	2012-10-03 12:27:17.0	Regz_AddDecimals1_C UpdateTextField
Aluno 3		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-0-TST-C	1	2012-10-03 12:27:17.0	Regz_SubtrDecimals1_C UpdateTextField
Aluno 4		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-0-TST-C	1	2012-10-03 12:27:17.0	Regz_SubtrDecimals2_C UpdateTextField
Aluno 5		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-0-TST-C	1	2012-10-03 12:27:17.0	Regz_AddDecimals2_C UpdateTextField
Aluno 6		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-0-TST-C	1	2012-10-03 12:27:17.0	Regz_AddDecimals3_C UpdateTextField
Aluno 7		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-0-TST-C	1	2012-10-03 12:27:17.0	
Aluno 8		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-0-TST-C	1	2012-10-03 12:27:17.0	done ButtonPressed
Aluno 9		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-1-TST-C	1	2012-10-03 12:30:46.0	
Aluno 10		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-1-TST-C	1	2012-10-03 12:30:46.0	Pegz_AddDecimals1_C UpdateTextField
Aluno 11		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-1-TST-C	1	2012-10-03 12:30:46.0	Pegz_AddDecimals1_C UpdateTextField
Aluno 12		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-1-TST-C	1	2012-10-03 12:30:46.0	Pegz_AddDecimals1_C UpdateTextField
Aluno 13		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-1-TST-C	1	2012-10-03 12:30:46.0	Pegz_AddDecimals1_C UpdateTextField
Aluno 14		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-1-TST-C	1	2012-10-03 12:30:46.0	done ButtonPressed
Aluno 15		ErrEx-CBA	AdaptErrEx_2.0-14	AdaptErrEx_2.0-14-2-TST-C	1	2012-10-03 12:31:12.0	

Titulo	Attempt At Step	Outcome	Selection	Action	Input 107
Aluno 1			NtpDate	NtpTimeCheck	2012-10-03 08:27:17.450 -0400
Aluno 2	1	Correct	Regz_AddDecimals1_C	UpdateTextField	1.6
Aluno 3	1	InCorrect	Regz_SubtrDecimals1_C	UpdateTextField	0.80
Aluno 4	1	Correct	Regz_SubtrDecimals2_C	UpdateTextField	0.497
Aluno 5	1	Correct	Regz_AddDecimals2_C	UpdateTextField	0.606
Aluno 6	1	Correct	Regz_AddDecimals3_C	UpdateTextField	6.01
Aluno 7			likert_Regz_AddDecimals3_C	UpdateLikert	5: Very sure
Aluno 8	1	Correct	done	ButtonPressed	-1
Aluno 9			NtpDate	NtpTimeCheck	2012-10-03 08:30:46.338 -0400
Aluno 10	1	InCorrect	Pegz_AddDecimals1_Ahundredths	UpdateTextField	0
Aluno 11	2	InCorrect	Pegz_AddDecimals1_Atenths	UpdateTextField	1
Aluno 12	3	InCorrect	Pegz_AddDecimals1_Aones	UpdateTextField	3
Aluno 13	4	Correct	Pegz_AddDecimals1_C	UpdateTextField	231.000
Aluno 14	1	Correct	done	ButtonPressed	-1
Aluno 15			NtpDate	NtpTimeCheck	2012-10-03 08:31:12.697 -0400

108					
Titulo	Feedback Text	Feedback Classification	Help Level	Total Num Hints	Condition Name
Aluno 1					ErrEx-CBA
Aluno 2					ErrEx-CBA
Aluno 3	negz;regz				ErrEx-CBA
Aluno 4					ErrEx-CBA
Aluno 5					ErrEx-CBA
Aluno 6					ErrEx-CBA
Aluno 7					ErrEx-CBA
Aluno 8					ErrEx-CBA
Aluno 9					ErrEx-CBA
Aluno 10					ErrEx-CBA
Aluno 11					ErrEx-CBA
Aluno 12					ErrEx-CBA
Aluno 13					ErrEx-CBA
Aluno 14					ErrEx-CBA
Aluno 15					ErrEx-CBA

Titulo	Condition Type	KC(Single-KC)	KC Category(Single-KC)	School	Class	CF(calculated_offset)	CF(question) 109
Aluno 1	Condition			Escola A Jr-High	Escola A_Fall_2012	68	
Aluno 2	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 3	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 4	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 5	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 6	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 7	Condition			Escola A Jr-High	Escola A_Fall_2012		How sure are you that you solved this problem correctly?
Aluno 8	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 9	Condition			Escola A Jr-High	Escola A_Fall_2012	79	
Aluno 10	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 11	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 12	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 13	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 14	Condition	Single-KC		Escola A Jr-High	Escola A_Fall_2012		
Aluno 15	Condition			Escola A Jr-High	Escola A_Fall_2012	84	

Titulo	CF (server_adjusted_event_time)	CF (server_name)	CF (server_time_millis)	CF (step_id)	CF (tool_event_time)	110 CF (tutor_event_time)
Aluno 1	2012/10/03 12:27:17.465 UTC	3.north-america.pool.ntp.org	1,34927E+12		2012/10/03 12:27:17.533 UTC	
Aluno 2	2012/10/03 12:28:09.616 UTC			1	2012/10/03 12:28:09.590 UTC	2012/10/03 12:28:09.684 UTC
Aluno 3	2012/10/03 12:28:24.623 UTC			3	2012/10/03 12:28:24.598 UTC	2012/10/03 12:28:24.691 UTC
Aluno 4	2012/10/03 12:29:01.314 UTC			5	2012/10/03 12:29:01.289 UTC	2012/10/03 12:29:01.382 UTC
Aluno 5	2012/10/03 12:30:12.840 UTC			7	2012/10/03 12:30:12.815 UTC	2012/10/03 12:30:12.908 UTC
Aluno 6	2012/10/03 12:30:42.824 UTC			9	2012/10/03 12:30:42.829 UTC	2012/10/03 12:30:42.892 UTC
Aluno 7	2012/10/03 12:30:42.792 UTC				2012/10/03 12:30:42.860 UTC	
Aluno 8	2012/10/03 12:30:44.56 UTC			11	2012/10/03 12:30:44.30 UTC	2012/10/03 12:30:44.124 UTC
Aluno 9	2012/10/03 12:30:46.354 UTC	3.north-america.pool.ntp.org	1,34927E+12		2012/10/03 12:30:46.433 UTC	
Aluno 10	2012/10/03 12:30:59.99 UTC			59	2012/10/03 12:30:59.116 UTC	2012/10/03 12:30:59.178 UTC
Aluno 11	2012/10/03 12:31:04.824 UTC			59	2012/10/03 12:31:04.841 UTC	2012/10/03 12:31:04.903 UTC
Aluno 12	2012/10/03 12:31:06.774 UTC			59	2012/10/03 12:31:06.775 UTC	2012/10/03 12:31:06.853 UTC
Aluno 13	2012/10/03 12:31:10.237 UTC			59	2012/10/03 12:31:10.254 UTC	2012/10/03 12:31:10.316 UTC
Aluno 14	2012/10/03 12:31:10.253 UTC			67	2012/10/03 12:31:10.301 UTC	2012/10/03 12:31:10.332 UTC
Aluno 15		3.north-america.pool.ntp.org	1,34927E+12		2012/10/03 12:31:12.797 UTC	

APÊNDICE A – Tabela de classificação dos artigos

Id do artigo	Título do Artigo	Classe Objetivo	Atributos
19	<i>A Decision Support System to improve e-Learning Environments</i>	Definição dos Atributos, Definição da Interface	Sessão, Dado pessoal, Sessão, Nota, Tempo, Dado pessoal, Tempo, Chat, Curso, Wiki, Fórum, Tarefa, Tarefa, Fórum, Tarefa, Fórum, Tarefa, E-mail, Tarefa, E-mail, Tentativa, Tarefa, Tarefa, Tarefa, Nota, Tarefa
33	<i>Addressing the Testing Challenge with a Web-Based E-Assessment System that Tutors as it Assesses</i>	Predizer ajuda que o aluno necessita	Tarefa, Tarefa, Tarefa, Tarefa, Dica, Dica, Dica, Dica, Dica, Dica, Tentativa, Tentativa, Tarefa, Tempo, Tempo
3	<i>Clustering and Sequential Pattern Mining of Online Collaborative Learning Data</i>	Exibir indicadores de ensino dos alunos	Tarefa, Tarefa, Tarefa, Tarefa, Wiki, Wiki, Wiki, Wiki, Wiki, Wiki, SVN, SVN, SVN, SVN
4	<i>The Design of eLeTK – Software System for Enhancing On-Line Educational Environments</i>	Exibir indicadores de ensino dos alunos	Tarefa, Tarefa, Tentativa, Tarefa, Tempo, Tempo
7	<i>Evolutionary algorithms for subgroup discovery</i>	Definição dos Atributos	Curso, Tarefa, Tarefa, Tarefa, Tarefa, Tarefa, Tarefa, Chat, E-

	<i>applied to e-learning data</i>		mail, Fórum, Mensagem Professor, Nota
8	<i>Predicting School Failure and Dropout by Using Data Mining Techniques</i>	Prever quais alunos irão cometer falhas	Nota, Nota, Nota, Nota, Nota, Nota, Motivação, Nota, Dado pessoal, Nota, Dado pessoal, Dado pessoal, Turma
16	<i>Predicting NDUM Student's Academic Performance Using Data Mining Techniques</i>	Definição dos Atributos	Dado pessoal, Dado pessoal, Dado pessoal, Dado pessoal, Escola, Dado pessoal, Escola, Nota, Nota, Nota, Escola, Escola, Motivação, Nota, Turma
20	<i>Subgroup discovery in an e-learning usage study based on Moodle</i>	Demonstrar qualidade de algoritmo	Curso, Tarefa, Tarefa, Tarefa, Tarefa, Tarefa, Chat, Mensagem Professor, Fórum, Fórum
21	<i>Measuring and mining LMS data</i>	Caracterizar os alunos, Definição dos Atributos	Curso, Curso, Curso, Sessão, Sessão, Tempo, Sessão, Sessão, Curso, Curso
44	<i>Generating Teacher Adapted Suggestions for Improving</i>	Caracterizar os alunos, Exibir sugestões para o ensino dos alunos	Tarefa, Nota, Nota, Tempo, Tempo, Tarefa

	<i>Distance Educational Systems with SIgMa</i>		
48	<i>Research and Design of Applying Association Rule in Course Management System</i>	Exibir sugestões para o ensino dos alunos	Tempo, Nota, Tentativa, Tentativa, Nota, Chat, Tarefa, Fórum, Fórum, Curso, Curso, Tempo, Nota, Tentativa, Fórum, Fórum, Curso, Curso, Tempo, Tempo, Nota
49	<i>Non-linear correlation techniques in educational data mining</i>	Como desenvolver novos métodos para a EDM	Escola, Tarefa, Curso, Sessão, Sessão, Tempo, Escola, Nota
50	<i>Analysis of E-learning in Data Mining – A Dreamed Vision for Empowering Rural Students in India</i>	Caracterizar os alunos	Sessão, Tempo, Nota, Nota, Tarefa, Tentativa, Chat, Fórum
55	<i>Creation of Students' Activities from Learning Management System and their Analysis</i>	Caracterizar os alunos	Curso, Tempo, Escola, Dado pessoal, Tarefa, Tarefa
3	<i>Data mining in course management systems Moodle case study and tutorial</i>	Caracterizar os alunos	Curso, Tarefa, Tarefa, Tarefa, Tarefa, Chat, Mensagem Professor, Fórum, Fórum, Tempo, Tempo, Nota

4	<i>Multi-instance genetic programming for predicting student performance in Web based educational environments</i>	Caracterizar os alunos	Tarefa, Tempo, Fórum, Fórum, Tempo, Tarefa, Tarefa, Tarefa
11	<i>Evolutionary algorithms for subgroup discovery in e-learning A practical application using Moodle data</i>	Caracterizar os alunos, Classificação de técnicas de ensino	Curso, Tarefa, Tarefa, Tarefa, Tarefa, Tarefa, Tarefa, Chat, Mensagem Professor, Fórum, Fórum, Nota
1	<i>A decision support system to improve e-learning</i>	Caracterizar os alunos	Sessão, Dado pessoal, Dica, Dado pessoal, Tempo, Dado

	<i>environments</i>		pessoal, Tempo, Chat, Curso, Wiki, Fórum, Tarefa, Fórum, Tarefa, Fórum, Tarefa, E-mail, Tarefa, E-mail, Tarefa, Tarefa, Tarefa, Tarefa, Nota, Nota, Nota
2	<i>A data mining approach to guide students through the enrollment process based on academic performance</i>	Desenvolver processos para a criação de ITS	Curso, Tentativa, Nota, Tarefa, Tarefa, Curso, Curso, Nota
3	<i>Application of machine-learning techniques to analyse student interactions and improve the collaboration process</i>	Avaliação de métodos colaborativos	Fórum, Fórum, Fórum, Fórum, Fórum, Fórum, Fórum, Fórum, Fórum, Fórum
4	<i>Automatically detecting a student's preparation for future learning Help use is key</i>	Caracterizar os alunos	Dica, Tempo, Tempo, Tempo, Tarefa, Off-task behavior, Tempo, Gaming the system, Tarefa, Tarefa
5	<i>Mining rare association rules from e-learning data</i>	Demonstrar qualidade de algoritmo	Curso, Tentativa, Tarefa, Tarefa, Tarefa, Fórum, Fórum, Tempo, Tempo, Tempo, Nota
6	<i>Mirroring of Knowledge Practices based on User-defined Patterns</i>	Exibir indicadores de ensino dos alunos	Tarefa, Dado pessoal, Dado pessoal, Dado pessoal, Motivação,

			Curso, Curso, Curso, Tempo, Curso
7	<i>Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data</i>	Prever quais alunos irão cometer falhas	Nota, Nota, Nota, Nota, Nota, Nota, Nota, Motivação, Nota, Dado pessoal, Dado pessoal, Turma, Dado pessoal, Nota, Turma, Dado pessoal, Tempo, Nota
8	<i>Predicting students final performance from participation in on Line discussion Forums</i>	Demonstrar qualidade de algoritmo	Fórum, Fórum, Fórum, Fórum, Fórum, Tempo, Nota, Fórum, Fórum
9	<i>Recommender system In collaborative learning environment using an influence diagram</i>	Avaliação de métodos colaborativos	Fórum, Fórum, Fórum, Fórum, Fórum, Fórum, Fórum, Fórum
10	<i>Success Is Hidden in the Students' Data</i>	Caracterizar os alunos, Definição dos Atributos, Definição do melhor momento para se classificar os alunos, Demonstrar qualidade de algoritmo	Dado pessoal, Dado pessoal, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Tempo, Nota
11	<i>Types of online hierarchical repository structures</i>	Caracterizar os alunos	Curso, Curso, Curso, Curso, Curso, Curso, Curso, Curso

12	<i>WTF Detecting students who are conducting inquiry without thinking fastidiously</i>	Caracterizar os alunos	Tarefa, Tempo, Tempo, Tarefa, Tentativa, Tempo, Tempo, Tempo, Tempo, Tempo, Tentativa, Tempo, Tempo, Tempo, Tempo, Tempo