UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

ERNANI GOTTARDO

ESTIMATIVA DE DESEMPENHO ACADÊMICO DE ESTUDANTES EM UM AVA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS

DISSERTAÇÃO

CURITIBA
2012

ERNANI GOTTARDO

ESTIMATIVA DE DESEMPENHO ACADÊMICO DE ESTUDANTES EM UM AVA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS.

Dissertação apresentada ao Programa de Pós-graduação em Computação Aplicada da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do grau de “Mestre em Computação Aplicada”. Área de Concentração: Engenharia de Sistemas Computacionais. Linha de Pesquisa: Sistemas de Informação.

Orientador: Prof. Dr. Robinson Vida Noronha

Co-Orientador: Prof. Dr. Celso Antonio Alves Kaestner

CURITIBA
2012

Dados Internacionais de Catalogação na Publicação

G686 Gottardo, Ernani

Estimativa de desempenho acadêmico de estudantes em um AVA utilizando técnicas de mineração de dados / Ernani Gottardo. — 2012.

84 f. : il. ; 30 cm

Orientador: RobinsonVida Noronha.

Co-orientador: Celso Antonio Alves Kaestner.

Dissertação (Mestradado) – Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Computação Aplicada. Área de concentração: Engenharia de Sistemas de Computacionais. Linha de Pesquisa: Sistemas de Informação, Curitiba, 2012.

Bibliografia: p. 79-84.

1. Mineração de dados (Computação). 2. Ensino a distância. 3. Aprendizagem. 4. Software educacional. 5. Internet na educação. 6. Computação – Dissertações. I. Noronha, Robinson, orient. II. Kaestner, Celso Antonio Alves, co-orient. III. Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Computação Aplicada. III. Título.

CDD (22. ed.) 004

Biblioteca Central da UTFPR, Campus Curitiba

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Câmpus Curitiba
Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Informática – DAINF
Programa de Pós-Graduação em Computação Aplicada - PPGCA
Mestrado Profissional

Título da Dissertação

"Estimativa de Desempenho Acadêmico de Estudantes em um AVA Utilizando Técnicas de Mineração de Dados"

por

Ernani Gottardo

Esta dissertação foi apresentada como requisito parcial à obtenção do grau de MESTRE EM COMPUTAÇÃO APLICADA - Área de Concentração: Engenharia de Sistemas Computacionais, pelo PPGCA - Programa de Pós-Graduação em Computação Aplicada - Mestrado Profissional – da Universidade Tecnológica Federal do Paraná - UTFPR - Câmpus Curitiba, às 9:30 horas do dia 04 de dezembro de 2012. O trabalho foi aprovado pela Banca Examinadora, composta pelos professores:

Prof. **Robinson Vida Noronha, Dr.**
presidente - (UTFPR - CT)

Prof. **Clovis Torres Fernandes, Dr.**
(ITA)

Prof. **Andrey Ricardo Pimentel, Dr.**
(UFPR)

Prof. **João Alberto Fabro, Dr.**
(UTFPR-CT)

Prof. **Hilton José Silva de Azevedo, Dr.**
(UTFPR-CT)

Departamento Acadêmico de Informática - DAINF
Programa de Pós-Graduação em Computação Aplicada - PPGCA
Câmpus Curitiba
Bloco B - ramal: 3310-4644 • Fax: 3310-4646

A Folha de Aprovação assinada encontra-se na Coordenação do Programa.

AGRADECIMENTOS

- A Deus por ter me dado forças e saúde para chegar até aqui e me protegido nas inúmeras viagens de Erechim-RS a Curitiba-PR durante o Mestrado.

- A minha esposa Elenice Gottardo e minha filha Bruna Gottardo que me incentivaram e apoiaram durante as longas horas de estudo e viagens.

- Ao meu orientador, Prof. Robinson Vida Noronha, pela disponibilidade, colaboração, dedicação e paciência com que me conduziu durante a realização deste trabalho. Por tudo isso, posso afirmar que foi um privilégio tê-lo como meu orientador.

- Ao meu co-orientador, Prof. Celso A. A. Kaestner, pelas valiosas colaborações prestadas para a realização deste trabalho.

- Ao Prof. João A. Fabro, coordenador do PPGCA, pela costumeira colaboração e celeridade no atendimento às demandas solicitadas à coordenação do programa.

- Aos professores João A. Fabro, Gustavo A. G. Lugo, Gilson Y. Sato, Hilton J. S. Azevedo pelos pertinentes apontamentos e sugestões realizados na apresentação do projeto e nos seminários de acompanhamento desta pesquisa.

- Aos professores João A. Fabro, Jean M. Simão, Tania M. Centeno, Gilda M. Friedlaender, Luiz Nacamura Jr., Paulo C. Stadzisz, Murilo V.G. Silva e Celso A. A. Kaestner pelos relevantes conhecimentos transmitidos em suas disciplinas.

- Aos professores Clóvis Torres Fernandes, Andrey Ricardo Pimentel, Hilton J. S. Azevedo, João A. Fabro e Robinson Vida Noronha pela participação na banca de defesa e pelos apontamentos realizados que contribuíram para a melhoria deste trabalho.

RESUMO

GOTTARDO, Ernani. ESTIMATIVA DE DESEMPENHO ACADÊMICO DE ESTUDANTES EM UM AVA UTILIZANDO TÉCNICAS DE MINERAÇÃO DE DADOS. 84 f. Dissertação – Programa de Pós Graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba, 2012.

Alguns ambientes educacionais têm incorporado *softwares* que são utilizados como apoio ou, em alguns casos, como condição básica para a disponibilização de cursos. Neste cenário, destacam-se os Ambientes Virtuais de Aprendizagem (AVA) usados para apoiar o desenvolvimento de cursos presenciais, semipresenciais e a distância. Os AVA caracterizam-se por armazenar um grande volume de dados. Contudo, esses ambientes carecem de ferramentas que permitam extrair informações úteis para o desenvolvimento de processos de acompanhamento eficiente dos estudantes. Diante disso, esta pesquisa investiga como os dados armazenados em um AVA poderiam ser processados para geração de informações relacionadas a estimativas de desempenho acadêmico futuro de estudantes. Para obter essas informações, primeiramente fez-se necessário a seleção de um conjunto de atributos para representar estudantes em um curso a distância (EAD) utilizando um AVA. O conjunto de atributos foi escolhido considerando-se três dimensões, selecionadas a partir da análise de referências teóricas da literatura sobre cursos EAD: perfil de uso do AVA, interação estudante-estudante e interação bidirecional estudante-professor. Aplicando-se técnicas de mineração de dados sobre o conjunto de atributos selecionados, foi possível então a obter estimativas sobre o desempenho futuro de estudantes. Essas estimativas poderiam apoiar o desenvolvimento de processos de acompanhamento efetivo dos estudantes, atividade de fundamental importância em cursos EAD. Neste trabalho, um estudo com sete experimentos foram realizados e apresentam diferentes cenários em que as estimativas sobre o desempenho podem ser obtidas. Os resultados desses experimentos apontam para a viabilidade desta proposta, tendo em vista os índices promissores de acurácia obtidos na classificação de estudantes quanto ao seu desempenho final nos cursos.

Palavras-chaves: Mineração de Dados Educacionais, Educação a Distância, Ambientes Virtuais de Aprendizagem, Estimativa de Desempenho de Estudantes.

ABSTRACT

GOTTARDO, Ernani. ACADEMIC PERFORMANCE PREDICTION OF STUDENTS IN A LMS USING DATA MINING TECHNIQUES. 84 f. Dissertação – Programa de Pós Graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba, 2012.

Some educational environments have incorporated software to support or, in some cases, as a basic condition to the availability of courses. In this scenario, stand out Learning Management Systems (LMS) used to support the development of classroom, blended or distance courses. Learning Management System are characterized by storing a large volume of data. However, these environments lack tools to extract useful information for the development of efficient processes for monitoring students'. Thus, this research investigates how data stored in a LMS could be processed to generate information regarding estimates of students' future academic performance. To obtain this information, first became necessary to select a set of attributes to represent students in an online course using a LMS. This set of attributes was chosen considering three dimensions, selected through the analysis of theoretical bases about online courses: LMS use profile, student-student interaction and bidirectional student-teacher interaction. Applying data mining techniques on the set of selected attributes, it was possible to obtain estimates of students' future performance. These estimates can support the development of effective processes for monitoring students, activity of fundamental importance in distance learning. In this research, a study with seven experiments were conducted and present different scenarios where estimates of performance can be obtained. The results of these experiments indicate the viability of this proposal, given the promising accuracy rates obtained in the classification of students regarding their final performance in courses.

Keywords: Educacional Data Mining, Distance Learning Education, Virtual Learning Environments, Students' Performance Prediction.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AVA: Ambiente Virtual de Aprendizagem

EAD: Ensino a Distância

EDM: *Educational Data Mining*

CMC: *Combination Of Multiple Classifiers*

ITS: *Intelligent Tutoring System*

SQL: *Structured Query Language*

SUMÁRIO

1. INTRODUÇÃO

O uso de ambientes computacionais de apoio a atividade educacional propicia o registro de dados sobre o processo de interação de estudantes que utilizam esses ambientes (Holliman e Scanlon, 2006; Macfadyen e Dawson, 2010; Romero e Ventura, 2010; Romero-Zaldivar et al., 2012).

Merceron e Yacef (2005) destacam que os dados armazenados são bastante diversificados. Eles variam desde registros de acesso e interações diversas com o sistema até dados com significados semânticos como respostas a testes ou participações em fóruns e *chats*.

Ambientes computacionais de apoio à atividade educacional apresentam alguns recursos comuns, voltados principalmente para o apoio ao desenvolvimento de atividades pedagógicas que poderiam ser úteis a professores e estudantes (Macfadyen e Dawson, 2010; Romero-Zaldivar et al., 2012). Dentre esses recursos, podem-se destacar ferramentas de interação, como fóruns de discussão e *chats* ou salas de bate-papo, que podem auxiliar na realização de atividades de aprendizado ativo. De acordo com Hrastinski (2009), em um ambiente de aprendizado ativo espera-se que o estudante seja ator principal do processo de aprendizado e o nível de interação e participação deve estar diretamente ligado ao desenvolvimento do conhecimento.

Além disso, estudantes utilizando um Ambiente Virtual de Aprendizagem (AVA) podem dispor de uma gama de opções para interagir e trabalhar de maneira colaborativa com colegas e professores. Nesse sentido, Nistor e Neubauer (2010) lembram que ambientes de aprendizagem *online* não devem ser apenas imagens de cursos tradicionais, devendo ter uma construção didática que considere as vantagens e desvantagens dessa tecnologia.

Apesar do número de informações que um típico AVA armazena em seu banco de dados sobre alunos e cursos, muitas dessas informações não são processadas de maneira adequada. Considera-se um processamento adequado aquele capaz de disponibilizar informações úteis (e.g. estudantes com probabilidade de reprovação ou evasão, descoberta de grupos de estudantes com características semelhantes) para apoiar professores na tomada de decisão ou realização de

prognósticos sobre os estudantes.

Essa lacuna é abordada nos trabalhos de Rabbany, Takaffoli e Zaïane (2011), Zorrilla et al., (2005) e Kampff (2009). Nesses trabalhos, os autores destacam a necessidade e importância do acompanhamento e da análise detalhada de todas as atividades desenvolvidas pelos estudantes com o objetivo de identificar necessidades específicas e oferecer a eles um auxílio personalizado.

Algumas plataformas de apoio ao ensino disponibilizam ferramentas de relatório sobre atividades desenvolvidas por estudantes. Entretanto, é difícil para professores extrair delas informações úteis que apóiem um acompanhamento efetivo dos estudantes (Romero, Ventura e Garcia, 2008; Macfadyen e Dawson, 2010; Romero-Zaldivar et al., 2012). Logo, instrutores geralmente podem ter apenas uma visão geral dos conteúdos que os estudantes acessaram ou das discussões em que esses se envolveram.

A dificuldade em obter rinformações importantes à práxis escolar pode ser considerada um elemento desmotivador ao uso do AVA pelo professor. Para esse professor hipotético, algumas perguntas precisariam ser respondidas, tais como:

- Como acompanhar efetivamente um estudante?
- Como verificar se os estudantes estão interagindo entre si?
- É possível identificar o quanto o estudante está apreendendo?
- Como identificar o estudante que está desmotivado e prestes a abandonar o curso?

Diante disso, um questionamento maior se impõe. Seria possível desenvolver um modelo de processo que pudesse auxiliar o professor a responder algumas dessas indagações?

As perguntas apresentadas definem possíveis caminhos e linhas de pesquisas. Vislumbra-se que os Ambientes Virtuais de Aprendizagem deixem de ser repositórios de dados sobre os estudantes que realizam um curso e tornem-se ferramentas poderosas de apoio ao professor.

Nesse cenário, os AVA poderiam agregar funcionalidades para auxiliar o professor na tomada de decisões ou para intervir diretamente no processo de aprendizagem do estudante. Poder-se-ia esperar que automatizar, mesmo que parcialmente, a tarefa de acompanhar estudantes nos AVA, se conseguida, reduziria a quantidade de trabalho do professor.

Um processo de automatização do acompanhamento dos estudantes poderia

considerar o que foi apresentado por Baker (2010). Em seu trabalho, Baker destaca que é fundamental considerar todos os possíveis tipos de interação do estudante no AVA. Como exemplo dessas interações, pode-se listar o seguinte:

- Número de acessos
- Frequência de acesso
- Participação em fóruns e *chats*
- Interação com professores e outros estudantes e
- Acesso aos materiais e atividades propostas

A coleta dessas informações pode ser feita de maneira não intrusiva, pois os AVA armazenam e disponibilizam essas informações em bancos de dados ou outros formatos de arquivos, como arquivos texto, por exemplo.

As limitações dos AVA na disponibilização de informações podem levar a dificuldades no desenvolvimento de atividades de acompanhamento dos estudantes que permitam monitorar o processo de aprendizagem nesses ambientes (Romero, Ventura e Garcia, 2008; Romero-Zaldivar et al., 2012). Esse fato é relevante tendo em vista que, segundo Moore e Kearsley (2007), a qualidade de um curso tem relação direta com a qualidade do processo de acompanhamento realizado.

Levando-se em conta as limitações dos AVA em fornecer informações ou ferramentas adequadas que permitam aos professores acompanhar os estudantes que realizam um curso utilizando esses ambientes, apresenta-se a seguir o problema que orienta esta pesquisa:

Como as informações disponíveis em uma base de dados de um AVA podem ser utilizadas para gerar inferências sobre o desempenho acadêmico de aprendizes?

No contexto deste trabalho, o termo “desempenho” ou “desempenho acadêmico” refere-se ao resultado final em termos de nota ou conceito obtido pelos estudantes em uma disciplina ou curso.

Buscando respostas ao problema que norteia o desenvolvimento desta pesquisa, no presente trabalho investigaram-se inicialmente referências teóricas da literatura para fundamentar a seleção de atributos que pudessem representar estudantes em um AVA. A partir dessa investigação, três dimensões foram

consideradas fundamentais para representação dos estudantes neste projeto de pesquisa: perfil de uso do AVA, interação estudante-estudante e interação bidirecional estudante-professor.

Baseando-se nas três dimensões apresentadas, um conjunto de atributos possíveis de serem obtidos a partir de uma base de dados de um AVA foi selecionado. Um estudo com sete experimentos foi então realizado com o objetivo de investigar cenários de uso e também possíveis limitações do conjunto de atributos selecionado.

Nesses experimentos, os algoritmos RandomForest e MultilayerPercepetron (Witten et al., 2011) foram aplicadas sobre o conjunto de atributos selecionado. Como resultado da aplicação desses algoritmos, estimativas sobre o desempenho acadêmico dos estudantes puderam ser obtidas. Espera-se que essas estimativas possam auxiliar a acompanhar o desempenho de estudantes em um curso realizado em um AVA.

1.1. OBJETIVOS

A seguir será apresentado o objetivo geral que motiva a realização deste trabalho, bem como os objetivos específicos que contribuem para a realização do objetivo geral.

1.1.1. OBJETIVO GERAL

Este trabalho tem como objetivo geral investigar a possibilidade de geração de inferências relativas ao desempenho de estudantes por meio de técnicas de mineração de dados utilizando atributos disponíveis em uma base de dados de AVA.

1.1.2. OBJETIVOS ESPECÍFICOS

Buscando-se atingir o objetivo geral desta pesquisa, destacam-se os seguintes objetivos específicos:

- Analisar a base de dados de um AVA para identificar quais informações poderiam ser utilizadas para representar um estudante em um curso EAD.
- Gerar inferências que indiquem estimativas de desempenho acadêmico futuro de estudantes em um curso EAD, utilizando técnicas de mineração de dados.
- Avaliar que atributos são relevantes para a geração de inferências relativas ao desempenho acadêmico de estudantes, eliminando possíveis atributos irrelevantes.
- Realizar um estudo de caso com experimentos que permitam avaliar a precisão das inferências sobre o desempenho acadêmico.
- Analisar a precisão das inferências especificamente em relação a correta identificação dos estudantes com desempenho inferior.

1.2. RELEVÂNCIA DA PESQUISA

Um fator importante para a efetividade de experiências de ensino é a capacidade de professores em monitorar o processo de aprendizagem e tomar decisões com base em eventos observados (Moore e Kearsley, 2007; Romero-Zaldivar et al., 2012). Por exemplo, segundo Ricarte e Falci Junior, 2011, a informação de que um recurso disponibilizado não foi acessado pelos estudantes, conforme esperado, poderia indicar que: I) o professor não deixou claro que o conteúdo deveria ser estudado, ou seja, ofereceu instruções inadequadas; II) o conteúdo não está sendo encontrado pelos estudantes, indicando um possível problema de visibilidade. Poder-se-ia ainda incluir um terceiro item à lista apresentada acima: os estudantes, deliberadamente, não acessaram o AVA ou os recursos disponibilizados nele, sendo necessário investigar fatores, principalmente

externos ao AVA, que levaram a isso.

O processo de monitoramento e avaliação constitui um diferencial qualitativo entre cursos em um ambiente EAD (Moore e Kearsley, 2007). Um professor que monitora os eventos que ocorrem em um AVA teria uma condição privilegiada para tomar decisões. Entretanto, atualmente este cenário hipotético ainda está muito longe da realidade nas instituições de ensino (Romero-Zaldivar et al., 2012). Eleutério e Bortolozzi (2004) complementam que em cenários que envolvem turmas com grande quantidade de estudantes, comuns em cursos EAD, essa tarefa é ainda mais desafiadora aos professores.

Diante dessa situação, disponibilizar informações envolvendo estimativas sobre o desempenho futuro de estudantes poderia ser útil para a tarefa de monitoramento e avaliação. A partir dessas estimativas, um professor poderia identificar estudantes com risco maior de reprovação e implementar ações pedagógicas específicas, como, por exemplo, a realização de atividades de revisão de conteúdo. As inferências sobre o desempenho futuro poderiam ainda auxiliar o professor na personalização do ensino ou focar as atenções em problemas específicos de determinado grupo de estudante.

Para essa tarefa de monitorar e avaliar, considerar indicadores relacionados com a interação, tais como estudante-professor, professor-estudante, estudante-estudante, colaboração e *feedback* é essencial, segundo Borba et al. (2008) e Pereira (2010). Entretanto, uma das principais limitações encontradas nos trabalhos correlatos (Ibrahim e Rusli, 2007; Minaei-Bidgoli et al., 2003; Márquez-Vera, Romero e Ventura, 2011) refere-se ao fato de que esses indicadores não são considerados nesses trabalhos.

1.3. ESTRUTURA DO TRABALHO

O Capítulo 2 é dedicado à revisão bibliográfica, destacando-se a apresentação e análise de trabalhos correlatos. Nesse capítulo, apresenta-se ainda uma revisão conceitual envolvendo técnicas computacionais ligadas à área de mineração de dados utilizadas na realização da pesquisa.

O Capítulo 3 apresenta as referências teóricas que nortearam a escolha das três dimensões para representação de aprendizes em um AVA. Ainda neste capítulo, são descritos atributos, selecionados com base nas três dimensões apresentadas, que servem como base para a obtenção de inferências sobre o desempenho dos estudantes.

No Capítulo 4 são apresentados os procedimentos utilizados para seleção e tratamento de dados a serem utilizados para obtenção de inferências relativas ao desempenho de estudantes. Nesse capítulo, também são apresentados os resultados de um estudo que envolve sete experimentos realizados. Cada experimento investiga diferentes cenários em que o conjunto de dados proposto para representar aprendizes em um AVA poderia ser utilizado para obtenção de prognósticos sobre o desempenho desses aprendizes.

A análise dos resultados do estudo com os experimentos realizados é apresentada no Capítulo 5. A análise é conduzida por meio de síntese e discussão dos resultados mais expressivos obtidos nos experimentos. Apresenta-se ainda nesse capítulo comparações dos resultados entre os experimentos.

Finalizando este documento, o Capítulo 6 apresenta as considerações finais onde são discutidos os resultados e as possíveis contribuições deste trabalho e suas limitações. Ainda nesse capítulo são apontadas algumas perspectivas de continuidade da pesquisa em trabalhos futuros.

2. REVISÃO BIBLIOGRÁFICA

Este capítulo está organizado em duas partes. Na primeira parte, é apresentada uma revisão conceitual de técnicas computacionais relacionadas com a área de mineração de dados aplicáveis ao contexto desta pesquisa. Na sequência, são destacados trabalhos e relatos de pesquisas referentes ao processo de busca de informações em bases de dados educacionais.

2.1. TÉCNICAS COMPUTACIONAIS

O desenvolvimento e utilização de sistemas informatizados (e.g. AVA), nas mais diversas áreas, tem permitindo a captura e armazenamento de vastas quantidades de dados. Romero, Ventura e García (2008) observam que na área educacional a quantidade de dados armazenada também vem crescendo. Para ilustrar esse cenário, cita-se o uso de AVA que registram todas as atividades realizadas nesses ambientes, tais como acessos e leituras, escritas em fóruns, respostas a questionários e comunicações diversas entre os participantes. Isso faz com que um grande volume de dados seja criado, tornando difícil a análise manual dos mesmos.

Com o objetivo de gerar informações a partir de grandes bancos de dados pode-se utilizar um processo conhecido como KDD (*Knowledge Discovery in Databases*) ou, em Português, Descoberta de Conhecimento em Bancos de Dados (Fayyad et al., 1996). O processo KDD foi utilizado como base para o desenvolvimento desta pesquisa, tendo em vista que esse processo define uma sequência clara de etapas e procedimentos para guiar a atividade de descoberta de padrões em banco de dados (Han e Kamber, 2006).

Fayyad et al., (1996) definem KDD como processo não trivial de identificação de padrões em dados que sejam válidos, inéditos, potencialmente úteis e compreensíveis. O processo KDD pode ser dividido em diversas etapas, reforçando que se trata de um processo interativo (com muitas decisões tomadas pelo usuário

do KDD) e iterativo, conforme mostrado na Figura 1.

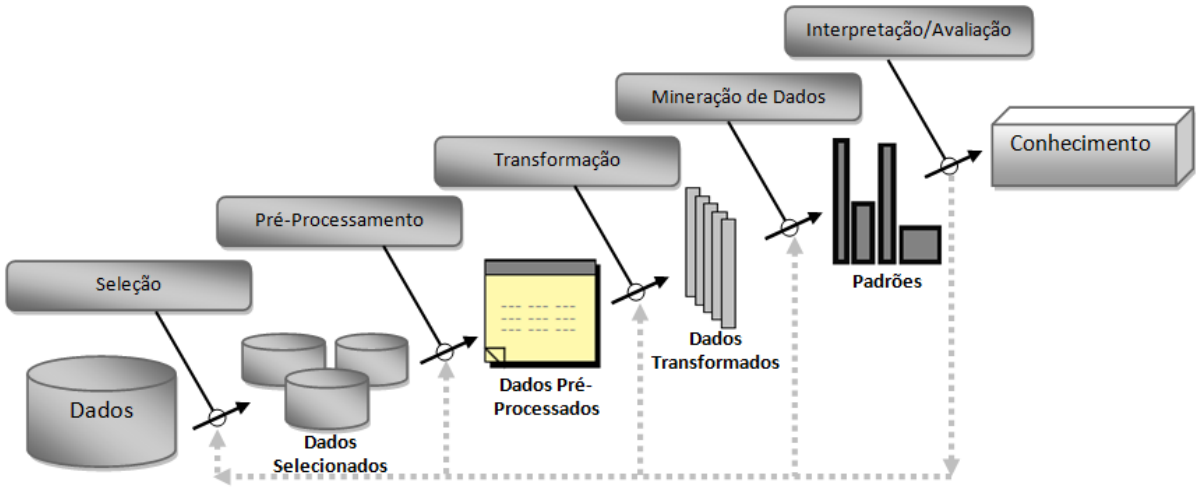

Figura 1 - Etapas da Descoberta de Conhecimento (adaptado de Fayyad et al., 1996)

Fayyad et al. (1996) e Han e Kamber (2006) observam que a fase de mineração de dados é frequentemente tratada como sinônimo de KDD. No escopo deste trabalho, mineração de dados será tratada como uma das etapas do processo KDD. Entretanto, conforme observam Han e Kamber (2006) essa etapa é essencial, pois tem o potencial de descobrir padrões ocultos e permitir a sua avaliação.

Han e Kamber (2006) definem mineração de dados como a extração ou "mineração" de conhecimento a partir de grandes volumes de dados. Para Fayyad, Piatesky-Shapiro e Smyth (1996) mineração de dados é a etapa no processo KDD que consiste na aplicação de algoritmos de descoberta que, considerando limitações computacionais aceitáveis, produzem uma enumeração particular de padrões (ou modelos) a partir de um conjunto de dados.

Técnicas de mineração de dados têm se desenvolvido muito nos últimos anos tendo sua aplicação atingido um número grande de áreas, como vendas, marketing, serviços financeiros (Fayyad et al., 1996; Witten et al., 2011) e, mais recentemente, a área educacional (Romero et al. 2008a; Baker, 2010; Zorrilla et al., 2005).

Especificamente com relação a ambientes educacionais, Zorrilla et al. (2005) apontam que a aplicação de técnicas de mineração de dados pode ser orientada para diferentes atores (e.g. educadores, professores, gestores e estudantes), cada qual com seu ponto de vista, como se pode ver na Figura 2. Educadores, professores e gestores são responsáveis por projetar, planejar, construir e manter sistemas educacionais. Estudantes usam e interagem com esses sistemas. Baseado

nas informações disponíveis sobre cursos, dados de uso e interação de estudantes, diferentes técnicas de mineração de dados podem ser aplicadas para descobrir conhecimentos úteis que ajudam a melhorar o processo de aprendizagem. O conhecimento descoberto pode ser usado não somente por gestores, professores e educadores, mas também pelos próprios usuários (estudantes).

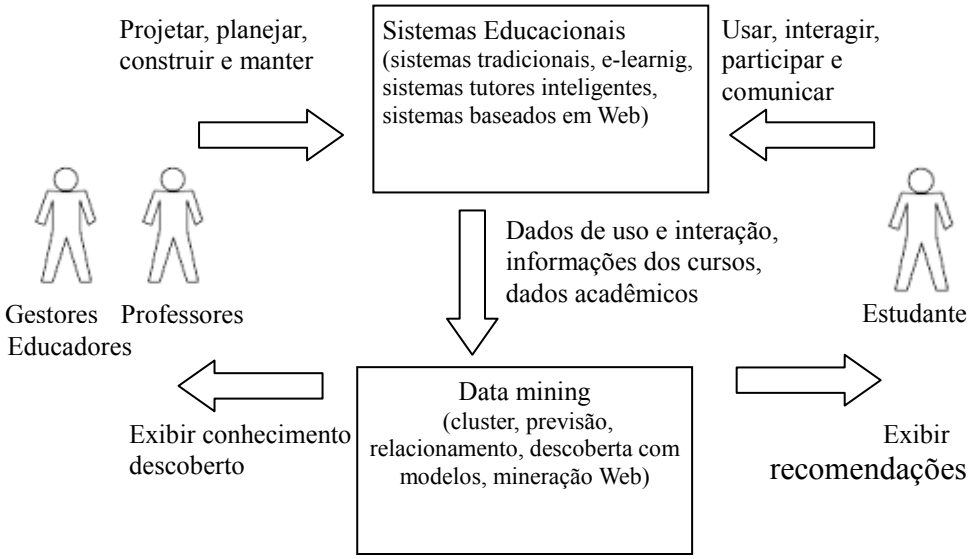

Figura 2 - O ciclo de aplicação de Mineração de Dados em sistemas educacionais (adaptado de Romero e Ventura, 2007).

Romero e Ventura (2007) destacam que diversas técnicas de mineração de dados têm sido utilizadas no contexto educacional. Dentre as técnicas mais utilizadas, esses autores destacam as que se enquadram no grupo “Previsão”.

Nesse grupo enquadram-se as técnicas que têm por objetivo desenvolver modelos que possam inferir um aspecto particular dos dados (variável a ser prevista) por meio de alguma combinação de outros aspectos destes dados (variáveis preditoras). Em linhas gerais, existem três tipos de técnicas de previsão: classificação, regressão e estimativa de densidade. Para cada técnica de previsão, as variáveis de entrada podem representar categorias ou valores contínuos, sendo que o grau de efetividade dos métodos é fortemente dependente do tipo de variáveis de entrada utilizadas.

Neste trabalho, a variável a ser prevista é o resultado ou nota final dos estudantes e as variáveis preditoras são os atributos propostos no Quadro 1 (pág. 37). A técnica de previsão a ser utilizada neste trabalho é a Classificação, dado que se busca distribuir os estudantes em classes categóricas de acordo com a nota

obtida no curso. Romero et al. (2008b) salientam que existem diversos métodos de classificação, dentre os quais destacam-se:

- Classificação Estatística - Procedimento em que itens individuais são agrupados em classes distintas baseados em inferências estatísticas. Essa técnica também é conhecida como Classificação Probabilística. Alguns exemplos de algoritmos desse grupo são: *linear discriminant analysis, least mean square quadratic, kernel e k nearest neighbors* (Witten et al., 2011).
- Árvore de Decisão - É um conjunto organizado em uma estrutura hierárquica em que uma instância de dados é classificada seguindo um caminho que satisfaça as condições, iniciando-se pela raiz e chegando até uma folha da árvore. Ressalta-se que uma árvore de decisão pode facilmente ser convertida para um conjunto de regras de classificação. C4.5, CART e RandomForest são exemplos de algoritmos desse grupo (Han e Kamber, 2006).
- Regras de Classificação - Neste grupo, enquadram-se técnicas de mineração de dados em que regras SE-ENTÃO são produzidas a partir da análise do conjunto de dados. Exemplos de algoritmos dessa classe são o AprioriC e CN2 (Han e Kamber, 2006).
- Redes Neurais – Trata-se de um paradigma computacional em que o modelo baseia-se no funcionamento de estruturas corticais do cérebro humano. Essa técnica consiste na interconexão de elementos processadores chamados nós ou neurônios que trabalham em conjunto para produzir uma função de saída (e.g. classificação de um estudante). *Multilayer Perceptron e Radial Basis Function Network* (Witten et al., 2011) são alguns exemplos de modelos desse grupo.

É importante observar que os métodos de Classificação Estatística e Rede Neurais são métodos de classificação que apresentam boas taxas de acertos, porém geram resultados difíceis de serem entendidos por pessoas (Romero et al., 2008b). Essa conclusão deve-se ao fato de que são técnicas do tipo "*black-box*" (caixa-preta), que normalmente não demonstram a estrutura do padrão da solução encontrada.

Neste trabalho, dois algoritmos disponíveis na ferramenta Weka (Witten et al., 2011) que implementam os métodos de classificação conhecidos como *Random Forests*, baseado em árvore de decisão, e *Multilayer Perceptron*, que utiliza redes

neurais, serão utilizados. No Weka, esses algoritmos estão disponíveis com o nome RandomForest e MultilayerPerceptron, respectivamente. Desta maneira, na sequência deste trabalho, os termos “algoritmo RandomForest” e “algoritmo MultilayerPerceptron” serão utilizados como referências à implementação desses dois métodos de classificação na ferramenta Weka.

A escolha desses algoritmos baseia-se no fato de que eles implementam técnicas de classificação que apresentam bons resultados e cujo uso tem sido frequente em aplicações relacionadas com a área de mineração de dados (Anil et al.,1996; Witten et al., 2011). Além disso, esses algoritmos ou técnicas de classificação apresentaram resultados promissores quando aplicados ao contexto educacional, como nos trabalhos de Manhães et al. (2011), Márquez-Vera et al. (2011) e Ibrahim e Rusli (2007).

O algoritmo RandomForest enquadra-se na categoria conhecida como “*ensemble methods”.* Essa técnica consiste em combinar um conjunto arbitrário de modelos de classificação com o objetivo de melhorar os resultados obtidos. No caso do algoritmo RandomForest, cada modelo do conjunto é construído utilizando uma árvore de decisão, formando assim uma “floresta”. Durante a classificação, cada árvore do conjunto “vota” e a classe escolhida é a que obtiver maior número de votos (Han e Kamber, 2006).

O algoritmo MultilayerPerceptron implementa a técnica conhecida como Redes Neurais Artificiais Diretas (*Feed-forward Artificial Neural Network)*. Inspiradas em redes neurais biológicas, as redes neurais artificiais são compostas de um conjunto conectado de unidades de processamento (neurônios) onde cada conexão tem um peso associado (Witten et al., 2011). Segundo Anil et al. (1996), redes neurais artificiais podem ser entendidas como um grafo direcionado em que os neurônios artificiais são os nós e as arestas direcionadas com peso são as conexões de entrada e saída desses neurônios. Em redes neurais artificiais diretas (*feed-forward*) o grafo formado não inclui laços de realimentação.

Anil et al. (1996) observam que a técnica de rede neural artificial direta utilizada com freqüência em aplicações que envolvem classificação é a Multilayer Perceptron. Neste tipo de rede, os neurônios são organizados em um número arbitrário de camadas (*layers*), possuindo apenas conexões direcionais entre as camadas. O treinamento da rede neural Multilayer Perceptron é feito usando a técnica conhecida como *backpropagation* (Witten et al., 2011).

Segundo Han e Kamber (2006), a técnica *backpropagation* treina a rede neural artificial processando iterativamente o conjunto de dados de treinamento e comparando o resultado da previsão para cada instância com o valor real conhecido. Os pesos das conexões entre os neurônios são então ajustados para diminuir o erro observado. Os valores do referido ajuste são retropropagados da camada de saída para as camadas intermediárias da rede, daí a origem do termo *backpropagation*.

A avaliação do desempenho de um modelo de classificação geralmente envolve a análise da habilidade de previsão ou correta separação das classes. Uma técnica muito utilizada para essa atividade é conhecida como "*confusion matrix*" ou matriz de confusão (Witten et al., 2011). Utilizando-se essa técnica, os resultados da classificação são apresentados como uma matriz bidimensional, com uma linha e coluna para cada classe. Cada elemento da matriz mostra o número de instâncias correta ou incorretamente classificadas considerando-se o conjunto de testes utilizado.

Em uma matriz de confusão com duas classes (e.g. "Sim" e "Não") quatro possíveis resultados podem ser representados: verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN). Verdadeiro positivo e verdadeiro negativo representam as classificações corretas. Um falso positivo ocorre quando o resultado é incorretamente previsto na classe "Sim" (ou positivo), quando na verdade é "Não" (negativo). Um falso negativo ocorre quando o resultado é incorretamente previsto como negativo (classe "Não"), quando na verdade é positivo

A Tabela 1 ilustra a representação de uma matriz de confusão e os resultados possíveis, considerando-se um modelo com duas classes ("Sim" e "Não"). Nessa tabela, os elementos na diagonal principal da matriz, destacados com fundo cinza, mostram o número de classificações corretas para cada classe. Os demais elementos representam os diferentes tipos de erro. É importante ressaltar que essa técnica pode ser generalizada para modelos com qualquer número de classes.

Tabela 1 - Matriz de confusão com diferentes resultados para previsão de duas classes.

		Classes Previstas pelo Modelo de Classificação	
		Sim	Não
Classes Reais / Conhecidas	Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Adaptado de Witten et al., 2011, p. 162.

A partir de uma matriz de confusão, pode-se obter um conjunto de medidas para avaliar o desempenho de um modelo de classificação. Uma dessas medidas é a acurácia, que mede a taxa de acerto global, ou seja, o número de classificações

corretas dividido pelo número total de instâncias dos dados a serem classificados. A acurácia pode ser definida pela expressão: acurácia = (VP + VN)/(VP + VN + FP + FN) (Witten et al., 2011). Essa medida também foi utilizada em trabalhos correlatos consultados (Romero et al., 2008a; Minaei-Bidgoli et al., 2003; Ibrahim e Rusli, 2007; Manhães et al., 2011).

Além da acurácia, Han e Kamber (2006) destacam que existem outras medidas que podem ser utilizadas para avaliação do desempenho de modelos de classificação. A Tabela 2 apresenta algumas dessas medidas juntamente com a descrição e a fórmula que pode ser utilizada para o cálculo. Essas medidas são relevantes para o presente trabalho, pois permitem avaliar o desempenho do processo de classificação individualmente para cada classe.

Tabela 2 - Medidas de desempenho de classificadores.

Medida	Descrição	Fórmula
Precisão	Proporção de instâncias que realmente pertencem à classe "x" entre todas as instâncias que foram classificadas como classe "x".	$VP_x / (VP_x + FP_x)$
Taxa de Verdadeiro Positivo	Proporção de instâncias que foram classificadas como classe "x", entre todas as instâncias que realmente pertencem à classe "x", ou seja, que percentual da classe foi capturado. É equivalente a medida "Recobertura" (*Recall).*	$VP_x / (VP_x + FN_x)$
Taxa de Falso Positivo	Proporção de instâncias que foram classificadas como classe "x", mas que pertencem a uma classe diferente, entre todas as instâncias que não são da classe "x"	$FP_x / ((VP + VN + FP + FN) – (VP_x + FN_x))$

Fonte: Adaptado de Witten et al. (2011)

A acurácia global é apresenta na ferramenta Weka e em trabalhos correlatos como uma taxa percentual, enquanto as demais medidas de desempenho são normalmente exibidas como taxas unitárias. Neste trabalho, esse padrão será adotado para apresentação da acurácia e das demais medidas de desempenho.

Técnicas de mineração de dados que utilizam algoritmos de classificação requerem que a base de dados seja separada em duas partes: conjunto de treinamento e conjunto de teste. Usando esta abordagem, um modelo de classificação é construído utilizando-se o conjunto de treinamento. Na sequência, esse modelo é aplicado para classificar as instâncias do conjunto de teste. Dessa maneira, é possível obter as medidas de desempenho do classificador (e.g. acurácia, precisão) avaliando-se os resultados obtidos no conjunto de teste.

Witten et al. (2011) destacam que os critérios utilizados para separação dos conjuntos de treinamento e teste têm uma grande influência no resultado final do processo. Os referidos autores citam alguns métodos possíveis, destacando o método "*K-fold Cross-Validation*" como uma das técnicas que apresentam os melhores resultados.

Essa técnica consiste inicialmente em definir um número fixo (K) de partições dos dados. Suponha-se, como exemplo, que sejam definidas três partições (K). Neste caso, inicialmente os dados são divididos aleatoriamente em três partições aproximadamente iguais. Na sequência, dois dos conjuntos são usados para treinamento e o terceiro é usado para teste. O processo é repetido três vezes para que, no final, todos os casos tenham sido usados exatamente uma vez para teste. As taxas de desempenho final do modelo são obtidas calculando-se a média das taxas de cada execução. Witten et al. (2011), baseando-se em estudos realizados, sugerem a adoção do número dez como valor padrão para o número de partições dos dados (K).

2.2. BUSCA DE INFORMAÇÕES EM BASES DE DADOS EDUCACIONAIS

O trabalho de Minaei-Bidgoli et al. (2003) demonstra a aplicação de uma técnica chamada de *Combination of Multiple Classifiers* - CMC. Essa técnica consiste em aplicar diversos algoritmos de classificação no conjunto de dados objeto de estudo, seguidos por um "voto". A classe que obtiver o maior número de "votos" será a escolhida. O estudo utilizando essa técnica baseou-se em dados de utilização de um AVA, tais como resposta de questões, trabalhos de casa, número de acessos, tempo de acesso, entre outros. O objetivo principal do trabalho consistiu em classificar estudantes cursando a disciplina "*Introductory Physics*" de acordo com a previsão de nota final. Os autores relatam que a técnica CMC apresentou taxa média de acerto de 70,9% na classificação de estudantes em um experimento considerando 3 classes.

Uma comparação entre a acurácia de métodos de aprendizagem de máquina Redes Neurais, Árvores de Decisão e Regressão Linear foi realizada por Ibrahim e

Rusli (2007). Eles desenvolveram um estudo buscando prever a aprovação ou reprovação em um curso na área de tecnologia da informação que utilizou um ITS - *Intelligent Tutoring System*. Nesse estudo foram utilizadas as seguintes informações: conhecimento prévio em programação e tecnologia da informação, situação financeira da família, tipo de escola anterior (internato ou não). O resultado do estudo mostrou que as três técnicas utilizadas apresentaram resultados promissores, atingindo percentual de precisão nas previsões acima de 80%.

Com o objetivo de identificar estudantes que apresentam risco de evasão do curso de graduação de Engenharia Civil da Universidade Federal do Rio de Janeiro - UFRJ, Manhães et al. (2011) realizaram um estudo para identificar quais algoritmos são mais adequados para mineração de dados educacionais. Os autores utilizaram dez algoritmos de classificação e concluíram que a acurácia média obtida entre eles era semelhante, variando entre 75 a 80%. Além da acurácia, os tipos de erros de classificação falso positivo e falso negativo foram analisados no estudo. Considerou-se no estudo que uma elevada taxa de erro para falso positivo (aluno classificado como sem risco de evasão e que na realidade evade) não é adequada para a solução do problema. Levando-se em conta a taxa de falso positivo, os algoritmos MultilayerPerceptron e RandomForest (Witten et al., 2011) apresentaram os melhores resultados dentre os dez algoritmos testados nesse estudo.

Conforme observam Márquez-Vera et al. (2011), grande parte das pesquisas relacionadas com previsão de desempenho são realizadas considerando o caso específico da educação a distância no ensino superior. Diante disso, o trabalho dos autores referenciados diferencia-se pela utilização de dez algoritmos de mineração de dados para realizar inferências sobre o desempenho de 670 estudantes do ensino médio da cidade de Zacatecas, México. Como particularidade, os dados utilizados nos experimentos foram obtidos totalmente por meio de formulários de pesquisas aplicados aos estudantes. As taxas de acurácia reportadas ficaram acima de 90%, apresentado assim os maiores valores entre os trabalhos correlatos citados nesta pesquisa.

O trabalho de Maia et al. (2010) destaca-se dentre os trabalhos que tratam a previsão de desempenho pela abordagem utilizada nesta pesquisa. Nesse trabalho, as inferências sobre desempenho futuro de estudantes em disciplinas de um curso de graduação são realizadas a partir das notas obtidas em disciplinas já cursadas. Os autores apresentaram um modelo de representação de aprendizes baseado na

teoria dos grafos. Nesse modelo, os alunos e as disciplinas foram modelados como nós e a relação entre eles como arestas de um grafo. A partir de relações de similaridade obtidas por meio do grafo, modelos de crescimento baseados em redes complexas foram capazes de prever a evolução do grafo e, consequentemente, as notas obtidas pelos alunos. Os autores reportaram que, do ponto de vista das disciplinas, existe uma grande variação nos valores médios dos erros observados, variando de 3,6% a 100%. Entretanto, os autores concluem que um erro médio elevado para uma disciplina poderia indicar o seguinte: I) ela não possui grande relação com as outras disciplinas do currículo; ou II) a avaliação apresenta algum grau de discrepância com os resultados obtidos em outras disciplinas.

Com o objetivo de fornecer *feedback* para autores de materiais e tutores a respeito do uso dos materiais disponíveis, Ricarte e Falci Junior (2011) desenvolveram um estudo utilizando algoritmos de agrupamento. Com a aplicação dessa técnica em cursos utilizando um AVA na Universidade Estadual de Campinas –UNICAMP, os autores buscaram identificar, a partir de padrões de uso do AVA, grupos de estudantes com comportamento similar. Como resultado do trabalho, foram gerados relatórios a professores e alunos identificando grupos e padrões de acesso aos recursos. Entretanto, os autores reforçam a importância de aprofundar a análise realizada com o objetivo de relacionar grupos e padrões de uso com o desempenho obtido nos cursos.

Merceron e Yacef (2005) desenvolveram um estudo demonstrando como o uso da técnica Regras de Associação (Witten et al., 2011) pode ajudar a descobrir conhecimentos pedagógicos relevantes extraídos do banco de dados do ITS Logic-ITA utilizado na Universidade de Sydney. Buscou-se descobrir quais são os erros que ocorrem com freqüência durante a resolução dos exercícios. Estes resultados puderam ser utilizados por professores para revisão de materiais do curso ou para enfatizar detalhes durante a explicação dos conceitos aos estudantes.

Além disso, algumas pesquisas que têm usado métodos de Mineração de Dados Educacionais – EDM (*Educational Data Mining*) para inferir a relação existente entre o estado emocional de estudantes e o desempenho apresentado por esses estudantes (Mcquiggan et al., 2008). Técnicas de EDM também foram usadas para analisar o impacto do atributo não intelectual autodisciplina no desenvolvimento da aprendizagem (Gong et al., 2009) e para detectar se um estudante está insatisfeito ou frustrado (D'Mello et al. 2008) .

Identificar quando um estudante está tentando manipular o sistema (em inglês "*gaming the system*") foi o objetivo do trabalho de Baker et al. (2006). Um exemplo de tentativa de manipular o sistema é o fato de um estudante solicitar repetidas vezes ajuda ao ambiente computacional antes mesmo de tentar resolver a questão.

Levando-se em consideração a natureza social do aprendizado, pesquisas têm focado no estudo de como estudantes participam e interagem em um ambiente de aprendizagem virtual (Li e Huang, 2008; Macfadyen e Dawson, 2010; Rabbany et al., 2011).

Outros trabalhos têm utilizado técnicas conhecidas como Análise de Rede Social (Wassermann e Faust, 1994) para investigar o relacionamento entre os participantes em um curso a distância. Destaca-se nesta abordagem de pesquisa o trabalho de Rabbany et al. (2011) que investigaram a importância da análise de redes sociais, enfocando a utilização da técnica chamada de "mineração de comunidades" (em inglês "*community mining"*) para o estudo e descoberta de estruturas relevantes em fóruns de discussão. O referido trabalho envolveu o desenvolvimento de um *software* que permite a visualização dos participantes em um fórum de discussão, suas interações, destacando diferentes papéis assumidos pelos estudantes, tais como líderes ou periféricos.

Nesta mesma linha, o trabalho de Gottardo e Noronha (2012) investiga as possibilidades geradas pela utilização de técnicas de análise de redes sociais com o objetivo de obter padrões de interação em fóruns de discussão que poderiam ter influência na aprendizagem. No referido trabalho, algumas medidas de centralidade e agrupamento foram utilizadas para a geração e visualização de grafos que demonstram padrões de interação de alunos e professores nos fóruns de discussão.

Conforme destacado anteriormente, os trabalhos que utilizam a abordagem de análise de redes sociais demonstram a viabilidade de utilização dessa técnica para obtenção de informações sobre o desenvolvimento de "comunidades de aprendizagem" em cursos a distância. Todavia, estas pesquisas não têm investigado empiricamente a possível influência direta dos aspectos de interação no desempenho acadêmico dos estudantes.

Segundo Cobo et al. (2011), os fóruns de discussão representam um dos recursos mais comuns em ambientes de aprendizagem *online*. Dessa maneira, as atividades dos estudantes em fóruns de discussão tornaram-se fontes importantes de informações que podem facilitar as tarefas de monitoramento durante o curso.

Eleutério e Bortolozzi (2004) destacam que os fóruns de discussão representam um recurso indispensável à formação de comunidades de aprendizagem a distância. Esses autores propuseram o desenvolvimento do sistema AMANDA que consiste em um ambiente para mediação de discussões. O objetivo principal desse *software* é aumentar a interatividade e reduzir o trabalho de professores e tutores no acompanhamento dessas discussões. Além de mecanismos de mediação, o referido sistema possui mecanismos de avaliação dos participantes que considera indicadores de participação dos estudantes na discussão.

Nesta mesma linha, algumas pesquisas têm investigado alternativas para simplificar e facilitar o processo de análise de conteúdos de fóruns de discussão (Dringus e Ellis, 2005; Azevedo et al., 2010). Técnicas de mineração de dados são utilizadas nesses trabalhos como ferramentas para obter indicadores quantitativos e qualitativos da participação de cada estudante. Como exemplo de indicadores quantitativos, pode-se citar os seguintes: número de postagens, tempo entre as postagens e tamanho das postagens. Um indicador qualitativo utilizado foi o número de postagens que estão relacionadas ao tema do fórum.

Cobo et al. (2011) apresenta uma proposta de uso de séries temporais e algoritmos de agrupamento hierárquico para identificar diferentes padrões de comportamento adotados por estudantes em fóruns de discussão. Um algoritmo de agrupamento hierárquico foi utilizado para agrupar estudantes com comportamento similar. Neste trabalho, séries temporais foram utilizadas para representar as atividades (e.g. escrita e leitura) desenvolvidas ao longo do tempo.

Com o objetivo de dar suporte ao processo de obter de indicadores qualitativos a partir de mensagens de textos disponíveis em fóruns de discussão, Azevedo et al. (2011) propõem a utilização de técnicas de mineração de texto, utilizando grafos. Nesse caso, os vértices do grafo representam os conceitos mais relevantes do texto minerado. As arestas ligando os vértices do grafo são construídas considerando-se a proximidade entre as palavras dentro do texto. O principal indicador qualitativo apresentado no trabalho foi chamado de relevância temática. O objetivo desse indicador é identificar se as contribuições textuais produzidas pelos discentes são relevantes para o debate. Os autores realizaram um estudo de caso com 403 mensagens em cinco fóruns de discussão para comparar os resultados da técnica de mineração de texto utilizando grafos com avaliações

feitas por cinco professores. Como resultado, os autores reportaram que o grau de similaridade entre a média da análise automática feita pela técnica proposta e a dos cinco professores ficou entre 70,3% e 96,4%.

Brooks et al. (2006) observam que o processo de desenvolvimento de AVA e ITS têm ocorrido de maneira amplamente independente. Os referidos autores investigam as possibilidades geradas pela convergência dessas duas áreas. O trabalho avalia como os AVA poderiam ser construídos incluindo técnicas utilizadas no desenvolvimento de ITS. Por outro lado, no trabalho desses autores são apresentadas alternativas para tornar os ITS mais abertos e reusáveis.

Bittencourt e Costa (2011) destacam que os ambientes clássicos de educação a distância (AVA) apresentam limitações relativas à falta de controle e avaliação personalizada, além de deficiências na adaptação às características dos usuários. Diante dessas limitações, esses autores investigam soluções e metodologias para o desenvolvimento de ambientes educacionais adaptativos e semânticos. A principal característica desses ambientes refere-se à preocupação com a automatização, interoperabilidade e reúso entre aplicações. Com isso, busca-se prover adaptação aos usuários e integração com ferramentas externas.

Trabalhos na área de mineração de dados educacionais envolvem o desenvolvimento de ferramentas para extrair informações a partir de ambientes educacionais (Kotsiantis, 2011; Azevedo et al., 2011, Eleutério e Bortolozzi, 2005). Entretanto, existem algumas dificuldades apontadas por Romero e Ventura (2007) para que professores possam usar adequadamente essas ferramentas. Dentre essas dificuldades, destaca-se o fato de que elas não são totalmente automatizadas ou são muito complexas. Os modelos e ferramentas propostos normalmente requerem a intervenção parcial de professores e gestores tanto para descobrir conhecimento quanto para analisar os resultados obtidos com ela. Como exemplo, o autor cita casos em que usuários são responsáveis pela seleção de tabelas e escolha de algoritmos de mineração e definição ou ajustes de seus parâmetros.

Considerando-se os trabalhos de pesquisa apresentados nesta seção, pode-se constatar os esforços realizados pela comunidade científica para disponibilizar informações para apoiar o processo de ensino. Contudo, mesmo com os resultados promissores relatados nas pesquisas apresentadas, é consenso entre os autores citados que mais avanços são possíveis e necessários para desenvolver artefatos tecnológicos que apóiem as atividades educacionais. Nesse sentido, Romero e

Ventura (2010), Baker (2010) e Baker et al. (2011) reforçam que a aplicação de técnicas de mineração de dados para descoberta de informações em base de dados educacionais é recente e que avanços são altamente relevantes.

Nos trabalhos destacados anteriormente, diferentes abordagens foram utilizadas para se obter informações relacionadas com ambientes educacionais. Dentre essas abordagens, pode-se destacar as seguintes: I) estimativas de desempenho acadêmico ou evasão, II) descoberta de grupos de estudantes com características similares, III) análise de redes sociais, IV) avaliação quantitativa e qualitativa de participações em fóruns de discussão e V) organização do conteúdo de fóruns de discussão para facilitar a análise.

A proposta apresentada neste trabalho de pesquisa enquadra-se na primeira abordagem destacada acima; ou seja, objetiva-se disponibilizar informações contendo estimativas de desempenho acadêmico futuro de estudantes utilizando um AVA. Os trabalhos consultados que se enquadram nessa abordagem enfatizam a comparação de diferentes algoritmos ou técnicas de aprendizagem de máquina, utilizando um conjunto restrito de atributos. Esses conjuntos de atributos não refletem todos os aspectos de interação, atividades e tarefas normalmente desenvolvidas por um estudante em um AVA. Entretanto, foram considerados suficientes por seus autores para alcançar o objetivo dos trabalhos que era a comparação de técnicas de mineração de dados. Como o foco do presente trabalho é a geração de inferências sobre o desempenho acadêmico, trabalhou-se com um conjunto de atributos que pudesse representar as principais atividades desenvolvidas pelos estudantes em um AVA.

Além disso, diferentemente de alguns trabalhos correlatos, no presente trabalho foi utilizado um AVA com código fonte e banco de dados aberto. Essas condições poderiam facilitar o desenvolvimento de futuras pesquisas que busquem ampliar ou melhorar os resultados obtidos neste trabalho.

3. REPRESENTAÇÃO DE ESTUDANTES EM UM AVA

Para a obtenção de inferências relativas ao desempenho acadêmico futuro de

estudantes em um curso EAD utilizando em AVA é necessário primeiramente escolher um conjunto de atributos que representem adequadamente esses estudantes. Para a realização desta pesquisa, essa escolha foi realizada considerando as referências teóricas detalhadas na Seção 3.1.

3.1. REFERENCIAL TEÓRICO

Uma referência fundamental utilizada para definir que atributos selecionar para representar estudantes em um curso EAD foi a “Teoria de Interação em Educação a Distância” (Moore, 1989). Essa teoria destaca três tipos de interação relevantes:

- Entre o estudante e o conteúdo ou objeto de estudo
- Entre o estudante e o especialista que elaborou o material em questão ou algum outro especialista atuando como instrutor
- Entre o estudante e outros estudantes, sozinho ou em grupo, com ou sem a presença em tempo real de um instrutor.

Corroborando a teoria de Moore (1989), Romero-Zaldivar et al. (2012) complementam que, de um ponto de vista abstrato, uma parte significativa das atividades desenvolvidas em qualquer experiência de aprendizagem *online* são baseadas em interações. Adicionalmente, aspectos relacionados com interação entre estudante-estudante e interação bidirecional estudante-professor são considerados também por vários autores como fundamentais em um processo de aprendizado *online* (Dringus e Ellis, 2005; Holliman e Scanlon, 2006; Li e Huang, 2008; Rabbany et al., 2011; Schrire, 2006).

Holliman e Scanlon (2006) destacam que a interação também pode acontecer de maneira passiva em um curso a distância. Essa situação é caracterizada pelos estudantes que lêem ativamente as contribuições de outros estudantes sem necessariamente escrever em fórum de discussão, por exemplo.

Além disso, outra referência importante para escolha dos atributos para representar estudantes em um curso EAD é o trabalho de Hrastinski (2008). Hrastinski realizou uma revisão bibliográfica, considerando as principais pesquisas

que abordam a importância da participar e interagir em ambientes de aprendizado *online*. Nesse estudo, o autor identificou seis níveis, em ordem crescente de complexidade, em que a participação de um estudante foi contextualizada em 36 pesquisas consultadas. A Tabela 3 resume os seis níveis de participação de um estudante propostos por Hratinski (2008). Essa tabela mostra também o número e percentual de pesquisas desenvolvidas em cada um dos níveis. Os autores das pesquisas consultadas salientam que as pesquisas que abordam os níveis 5 e 6 da Tabela 3 normalmente utilizam questionários ou entrevistas com os estudantes. A partir desses questionários ou entrevistas são obtidos indicadores com quantidade percebida de escrita ou se os estudantes sentem-se parte de um diálogo.

Tabela 3 - Níveis de concepção da participação de estudantes em ambientes de aprendizagem *online*.

Nível	Concepção	Número de Pesquisas	Percentual de Pesquisas
1	Participação como quantidade de acesso ao ambiente de aprendizado *online*	1	3
2	Participação como quantidade de escrita	10	28
3	Participação como qualidade da escrita	9	25
4	Participação como quantidade de escrita e leitura	2	6
5	Participação como quantidade real e percebida de escrita	2	6
6	Participação como fazer parte de um diálogo	13	33
Total		36	100

Fonte: Hrastinski, 2008, p. 1757.

3.2. PROPOSTA DE MODELO DE INFERÊNCIA DE DESEMPENHO DE ESTUDANTES EM UM AVA

Considerando-se as referências teóricas destacadas acima, escolheram-se três dimensões para representar os estudantes em um AVA: perfil de uso do AVA, interação estudante-estudante e interação bidirecional estudante-professor. A Figura 3 apresenta essas três dimensões, detalhadas a seguir:

- **Perfil de Uso do AVA -** Nesta dimensão, mapeiam-se dados que indiquem aspectos de planejamento, organização e gestão do tempo do aluno para a realização do curso. Para isso utilizaram-se indicadores gerais de quantidade e tempo médio de acessos aos recursos do AVA e de existência de rotina ou

regularidade desses acessos. Levar em conta esses indicadores é de fundamental importância, segundo Baker (2010).

- **Interação Estudante-Estudante -** Nesta dimensão, busca-se avaliar se os estudantes interagem entre si, usando as ferramentas disponíveis, como fóruns ou *chats,* para o envio e recebimento de mensagens. Segundo Schrire (2006), essa informação poderia indicar a existência de colaboração ou cooperação entre estudantes, caracterizando o processo definido por como "aprendendo com os outros".
- **Interação Bidirecional Estudante-Professor -** Nesta dimensão, investiga-se como professores ou tutores interagem com estudantes no contexto do curso. Este tipo de informação tem sua importância destacada por Holliman e Scanlon (2006). Eles ressaltam que professores ou tutores tem um papel fundamental no sentido de facilitar e incentivar a colaboração entre estudantes e também corrigir possíveis desvios para manter o foco no tema proposto.

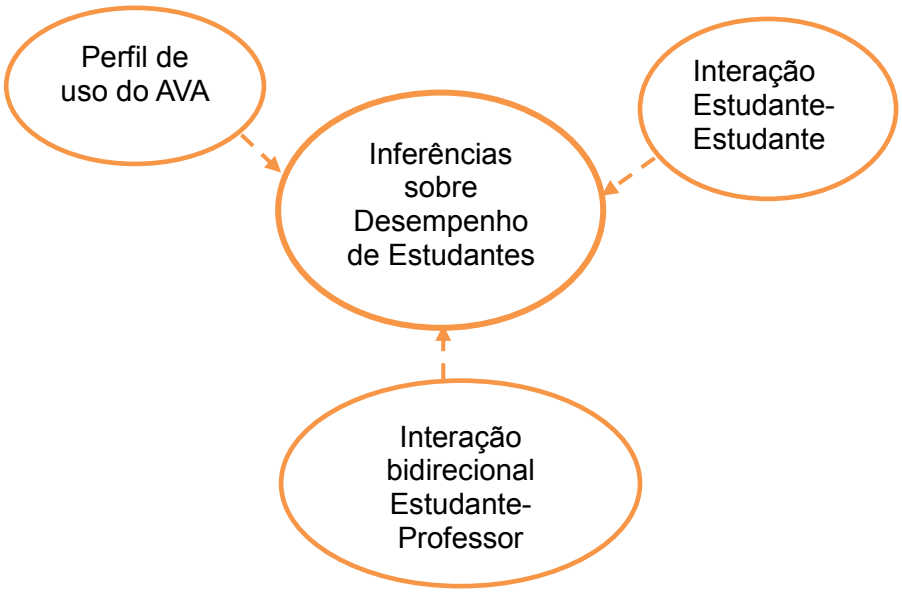

Figura 3 – Dimensões do Modelo de Inferência de Desempenho de Estudantes.

Considerando-se as três dimensões apresentadas, selecionou-se um conjunto de atributos que poderiam ser associados a cada dimensão e que poderiam ser extraídos da base de dados de um AVA. O Quadro 1 apresenta esses atributos e sua relação com as três dimensões apresentadas. O primeiro atributo "id_estudante" representa o código único de identificação do estudante no AVA. Esse atributo não

está relacionado com nenhuma das três dimensões e foi incluído para possibilitar a identificação e busca de informações adicionais sobre o estudante, caso necessário.

Dimensão	Atributo	Representação
	Código identificador do estudante no AVA	id_estudante
Perfil de Uso do AVA	Número total de acesso ao AVA	nr_logins
	Número total de postagens realizadas em fóruns	nr_tot_posts
	Número de postagens de outros participantes lidas em fóruns	nr_posts_lidos
	Número total de revisões em postagens anteriores realizadas em fóruns	nr_tot_rev
	Número de sessões de chat que o estudante participou	nr_chats
	Número de mensagens enviadas ao chat	nr_msg_chat
	Número de questões respondidas	nr_questoes_resp
	Número de questões respondidas corretamente	nr_questoes_acert
	Tempo médio decorrido entre os diversos acessos ao sistema	tempo_decorrido
	Número de dias transcorridos entre o início do curso e o primeiro acesso do estudante no AVA	tempo_decor_prim_login
Interação Estudante-Estudante	Número total de respostas postadas em fóruns referindo-se a postagens de outros estudantes	nr_posts_env
	Número de postagens em fóruns de outros estudantes que fazem referência a postagem do estudante	nr_posts_rec
	Número de mensagens recebidas de outros estudantes durante a realização do curso	nr_msg_rec_est
	Número de mensagens enviadas a outros estudantes durante a realização do curso	nr_msg_env_est_est
Interação bidirecional Estudante-Professor	Número de postagens do estudante que tiveram respostas feitas por professores ou tutores do curso	resp_est_prof
	Número de postagens de professores ou tutores que tiveram respostas feitas pelo estudante	resp_prof_est
	Número de mensagens enviadas ao professor/tutor durante a realização do curso	nr_msg_env_est_prof
	Número de mensagens recebidas do professor/tutor durante a realização do curso	nr_msg_rec_prof
Objetivo da Previsão	Resultado final obtido pelo estudante no curso. Representa classe objetivo de classificação	resultado_final

Quadro 1 - Atributos selecionados para representação de estudantes em um AVA

O processo de seleção dos atributos para compor cada uma das classes foi realizado analisando-se as funcionalidades e as tabelas do banco de dados disponíveis em uma instalação padrão do AVA Moodle (Moodle, 2012). Para as dimensões que envolvem interação, avaliou-se que atividades disponíveis em uma instalação padrão desse AVA permitem interação. Observou-se que se enquadram

nesse requisito as atividades conhecidas como fóruns de discussão e mensagens. A partir disso, as tabelas do banco de dados que armazenam dados dessas atividades foram analisadas para a seleção dos atributos que poderiam ser utilizados. A seleção de atributos para compor a dimensão perfil de uso do AVA foi feita de maneira semelhante, porém sem levar em consideração atividades que envolvam interação com outros participantes. Para a dimensão Perfil de Uso do AVA, analisou-se também uma tabela de registros de acesso que o Moodle armazena sobre ações gerais realizadas no ambiente.

Em uma instalação padrão, o banco de dados do AVA Moodle tem aproximadamente 250 tabelas de onde certamente seria possível extrair outros atributos, além daqueles selecionados nesse trabalho. Porém, com os 18 atributos selecionados e ilustrados no Quadro 1 pretende-se atingir o primeiro objetivo definido nesta pesquisa. Esse objetivo envolve identificar quais informações disponíveis uma base de dados de um AVA poderiam ser utilizadas para representação de estudantes nesses ambientes.

Deve-se destacar que, com as três dimensões e o conjunto de atributos apresentados, não se pretende definir um método de avaliação da aprendizagem. Apenas investiga-se a possibilidade de obter estimativas de desempenho acadêmico que possam auxiliar o professor a identificar comportamentos e perfis dos estudantes.

4. SELEÇÃO E TRATAMENTO DE DADOS

Este capítulo é dedicado inicialmente a apresentação do procedimento utilizado para selecionar dados experimentais baseados nos atributos propostos para representação dos estudantes em um AVA. Utilizando o conjunto de dados selecionados e aplicando-se técnicas de mineração de dados, inferências sobre o desempenho acadêmico dos estudantes puderam finalmente ser obtidas. Essas inferências são apresentadas na sequência desse capítulo em um estudo desenvolvido com sete experimentos que exploram diferentes cenários e possibilidades de obtenção dessas inferências.

A aplicação dos algoritmos RandomForest e MultilayerPerceptron para a realização de inferências sobre o desempenho de estudantes foi feita utilizando-se uma base de dados do ambiente Moodle. Essa base de dados contém informações de estudantes matriculados em disciplinas de cursos de especialização *lato-sensu* ofertados totalmente a distância. Visando manter a confidencialidade das informações utilizadas, será preservada a identificação dos estudantes e também da disciplina objeto desse estudo.

Como o foco do trabalho refere-se à estimativa do resultado ou nota final, não foram considerados estudantes desistentes, tendo em vista a indisponibilidade do resultado final nesses casos. Além disso, a evasão tem relação com fatores específicos como, por exemplo, ambiente sócio-econômico e público-alvo de cursos, tendo sido foco exclusivo de trabalhos como de Manhães et al. (2011) e Kampff (2009).

De posse do banco de dados de estudantes, a primeira atividade realizada consistiu em identificar disciplinas candidatas a servirem de base para o estudo. Os critérios considerados nessa escolha foram os seguintes:

- Maior número de alunos que concluíram a disciplina
- Maior número de ofertas da disciplina para turmas diferentes
- Disponibilidade do resultado de avaliações parciais e finais no banco de dados do AVA
- Maior número de recursos do AVA utilizados (fóruns, *chats*, questionários,

etc).

Baseando-se nos critérios elencados acima, escolheu-se uma disciplina com duas turmas já encerradas, totalizando uma população de 140 estudantes concluintes. Foram então desenvolvidos procedimentos para extração do conjunto de atributos selecionados para representar os estudantes, ilustrados no Quadro 1, a partir do banco de dados do Moodle.

Com o objetivo de destacar algumas características do conjunto de dados extraídos, apresenta-se na sequência duas tabelas. A Tabela 4 apresenta a distribuição dos dados de cada um dos atributos do conjunto de dados dos 140 estudantes objetos desse estudo. A Tabela 5 demonstra a correlação existente entre o atributo “Resultado_Final” e os demais atributos propostos para representação dos estudantes.

Tabela 4 - Distribuição dos dados dos estudantes selecionados.

Atributo	Menor Valor	Maior Valor	Média	Desvio Padrão
NR_LOGINS	155	1339	444,26	211,72
TEMPO_DECOR_PRIM_LOGIN	0	25	1,32	2,92
TEMPO_DECORRIDO	0,45	4,69	1,73	0,77
NR_MSG_ENV_EST_EST	0	5	0,25	0,74
NR_MSG_ENV_EST_PROF	0	3	0,19	0,59
NR_MSG_REC_PROF	0	22	1,09	3,39
NR_MSG_REC_EST	0	9	0,25	1,27
NR_TOT_POSTS	25	189	57,44	22,10
NR_POSTS_LIDOS	103	1420	344,63	219,44
NR_TOT_REV	0	58	8,54	9,15
NR_POSTS_ENV	0	78	10,70	16,31
NR_POSTS_REC	0	39	10,70	6,65
RESP_EST_PROF	0	58	6,51	6,79
RESP_PROF_EST	0	19	2,03	3,29
NR_QUESTOES_RESP	8	14	11,85	1,75
NR_QUEST_ACERT	6	14	10,84	1,89
NR_CHATS	0	31	7,90	6,58
NR_MSG_CHAT	0	1393	257,45	303,47
RESULTADO_FINAL	67	97	82,27	5,13

Para viabilizar a utilização de algoritmos que requerem que a classe seja um valor discreto e também para facilitar a interpretação dos resultados, foi realizado procedimento de discretização do atributo “Resultado_Final”, descrito a seguir.

Optou-se por dividir os estudantes em três classes chamadas de “A”, “B” e “C” de acordo com a nota obtida no curso. Na classe “A” e “C” serão alocados os

estudantes com as notas mais altas e mais baixas, respectivamente. Os demais estudantes farão parte da classe “B”.

Tabela 5 - Correlação entre o atributo "Resultado_Final" e os demais atributos.

Atributo	Correlação com o Atributo “Resultado_Final”
NR_LOGINS	0,2723
TEMPO_DECOR_PRIM_LOGIN	-0,0020
TEMPO_DECORRIDO	-0,2444
NR_MSG_ENV_EST_EST	-0,1466
NR_MSG_ENV_EST_PROF	-0,1513
NR_MSG_REC_PROF	-0,3298
NR_MSG_REC_EST	-0,1771
NR_TOT_POSTS	0,2247
NR_POSTS_LIDOS	0,3602
NR_TOT_REV	0,2926
NR_POSTS_ENV	0,0611
NR_POSTS_REC	0,1316
RESP_EST_PROF	0,0048
RESP_PROF_EST	-0,0640
NR_QUESTOES_RESP	-0,0334
NR_QUEST_ACERT	0,1174
NR_CHATS	0,5001
NR_MSG_CHAT	0,4686

Analisando-se as notas dos estudantes das turmas selecionadas para o presente trabalho, verificou-se que a distribuição dessas notas aproximam-se de uma curva normal. Esse fato pode ser observado através da Figura 4 que apresenta um histograma da distribuição das notas dos estudantes selecionados, gerado pela ferramenta Weka. No eixo horizontal desse gráfico são apresentadas as notas dos estudantes, divididas em nove intervalos iguais, com a média geral 82 ao centro. O eixo vertical desse gráfico representa o número de estudantes em cada um dos intervalos de distribuição das notas.

Tomando-se como base uma distribuição normal, inicialmente, na classe “B” foram alocados os estudantes cujas notas situem-se em aproximadamente 80% da área central da curva normal. Os demais estudantes foram distribuídos nas classes “C” e “A”, representando, respectivamente, aproximadamente 10% da área inferior e superior da curva normal.

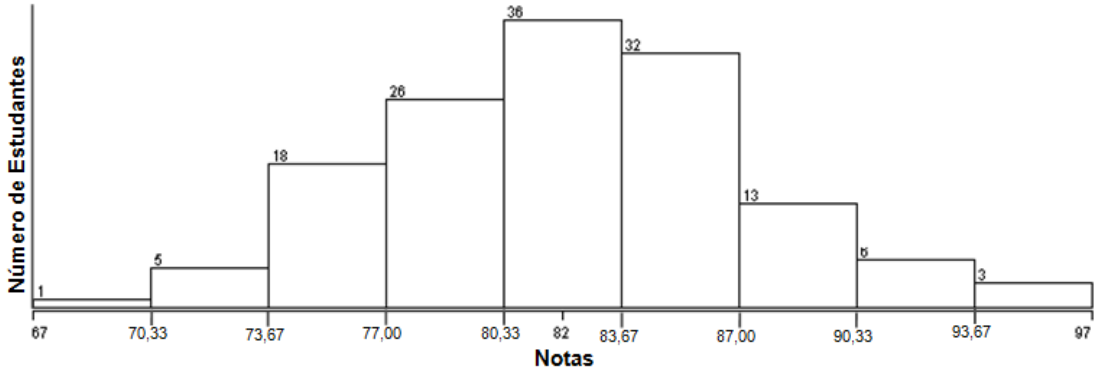

Figura 4 - Histograma de distribuição das notas dos estudantes

A Figura 5 apresenta graficamente uma curva normal com a distribuição das classes levando-se em consideração o método descrito acima. A região central da Figura 5 (destacada em cinza) compreende aproximadamente 80% da área total e representa a classe “B”. Nas extremidades da curva (cor branca) localizam-se os estudantes das classes “C” e “A”. Os números apresentados no eixo horizontal representam as notas que delimitam as classes, com o valor médio global das notas (82) ao centro. O eixo vertical representa o número de estudantes.

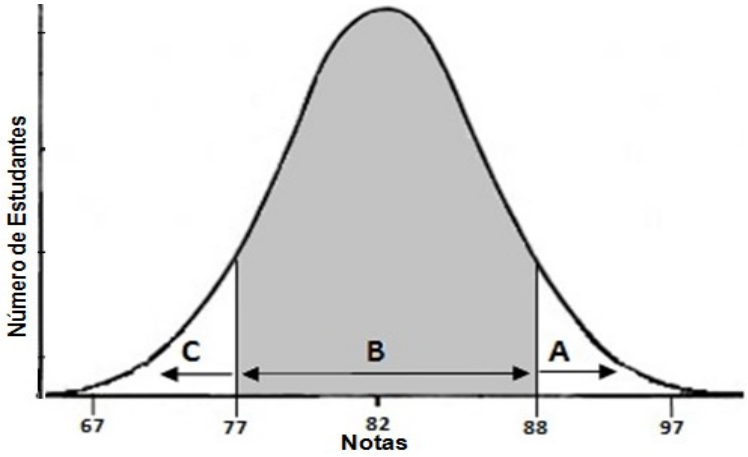

Figura 5 - Visualização da distribuição das classes

Na Tabela 6, as características individuais das três classes são destacadas. Nessa tabela são apresentados o título e a descrição das classes, bem como o número de estudantes e o intervalo de notas em cada uma das três classes criadas.

Tabela 6 - Distribuição das classes obtidas pelo do processo de discretização

Título da Classe	Descrição	Número de Estudantes	Intervalo de Notas
A	Alunos com desempenho superior	16	Maior que 88
B	Alunos com desempenho intermediário	109	Entre 77 e 88, inclusive
C	Alunos com desempenho inferior	15	Menor que 77

Ao analisar os valores da Tabela 6, pode-se observar que o número de estudantes na classe “A” e “C” é diferente. Esse fato poderia ser interpretado como uma inconsistência do processo de discretização, já que foi definido o mesmo percentual (10%) para as duas classes. Entretanto, esta situação ocorre para evitar que estudantes com a mesma nota sejam alocados em classes diferentes. Como exemplo, na classe “C” inicialmente 14 (10% dos 140) estudantes foram alocados, incluindo-se nesse número dois estudantes com a nota 76. Porém, existem três estudantes com a nota 76 na base de dados. Diante disso, mais um estudante foi incluído na classe “C”, totalizando assim 15, conforme apresentado na Tabela 6.

No procedimento de discretização apresentado acima, o valor de dois parâmetros importantes foram escolhidos pelo autor: o número de classes e o método para distribuição dos estudantes nas classes. Devido à relevância e potencial impacto desses parâmetros nos resultados do processo de classificação, serão destacados a seguir elementos que motivaram essas escolhas.

Nos trabalhos correlatos apresentados no Capítulo 2, observou-se que o número de classes utilizadas é variado e não existe uma discussão específica relativa a essa questão. Como exemplo dessa variação, observa-se no trabalho de Romero et al. (2008a) a utilização de 4 classes, enquanto em Minaei-Bidgoli et al. (2003) foram realizados experimentos com 2, 3 e 9 classes.

Desta maneira, optou-se por utilizar três classes neste trabalho. Com esse número é possível manter a facilidade da interpretação dos resultados das técnicas de classificação. Entende-se que a dificuldade de interpretação dos resultados poderia aumentar proporcionalmente ao aumento do número de classes. Isso porque, quanto maior o número de classes, maior a semelhança entre elas, podendo-se chegar a situações em que mínimas variações nos padrões de uso e interação coloquem estudantes em classes diferentes. Por outro lado, a definição de apenas duas classes agruparia na mesma classe estudantes com maiores

diferenças nos padrões de uso e interação. Isso potencialmente prejudicaria a identificação de grupos específicos de estudantes que poderiam interessar a um professor.

Além do número de classes, o número de estudantes em cada classe é outro parâmetro relevante. Conforme descrito anteriormente, optou-se por agrupar em cada uma das classes “A” e “C” aproximadamente 10% dos estudantes com as maiores e menores notas, respectivamente. A identificação desses dois grupos específicos de estudantes poderia ser considerada importante para um professor interessado em investigar possíveis fatores que contribuem para a um bom ou mau desempenho.

Considerando a distribuição dos dados experimentais apresentada, poder-se-ia concluir que quanto maior o percentual de estudantes nas classes A e C, maior o número de estudantes com desempenho intermediário juntar-se-ão a elas. Isso poderia dificultar a identificação de possíveis grupos com padrões diferenciados de uso e interação com o AVA. Por outro lado, utilizar classes com número de instâncias desbalanceadas pode influenciar o desempenho dos algoritmos de classificação. Diante disso, na sequência deste capítulo, serão apresentados experimentos que avaliam o impacto da alteração na distribuição do número de estudantes nas classes e também alternativas para tratar a questão do desbalanceamento entre o número de instâncias das classes.

É importante ressaltar que existem alternativas ao processo de discretização apresentado acima. O trabalho de Gottardo et al. (2012) demonstra a utilização de um algoritmo para automatização do procedimento de discretização. No referido trabalho foi aplicado o algoritmo de discretização não supervisionado disponível na ferramenta Weka (Witten et al., 2011). Este algoritmo utiliza a técnica conhecida como *equal-width*, que divide o intervalo de valores possíveis em subintervalos de mesmo tamanho. No entanto, a distribuição dos estudantes nas classes utilizando-se esse método é altamente dependente de possíveis notas com valores muito baixos ou altos. Por exemplo, um único estudante com uma nota extremamente baixa altera completamente a distribuição das classes, chegando-se a situações em que apenas um estudante compõe a classe.

4.1. ESTUDO REALIZADO

A geração de inferências ou estimativas relativas ao desempenho acadêmico de estudantes tem relação direta com os objetivos desta pesquisa apresentados na Seção 1.1. Buscando-se atingir esses objetivos, um estudo com sete experimentos foram realizados utilizando-se o conjunto de dados descrito no início deste capítulo. Nesses experimentos, procurou-se identificar diferentes cenários em que o conjunto de atributos propostos para representação de estudantes pode ser utilizado para a geração de estimativa de desempenho acadêmico.

Para o desenvolvimento deste estudo e seus experimentos foram utilizados os algoritmos de classificação RandomForest e MultilayerPerceptron, apresentados no Capítulo 2, e disponíveis na ferramenta Weka. Salienta-se que em todos os experimentos realizados foi utilizado o método "*K-fold Cross-Validation*" como técnica para a estratificação da base de dados em conjunto de treinamento e teste. Foi adotado 10 como valor padrão para o número de partições dos dados (K), conforme sugerem Witten et al. (2011).

Na sequência, serão apresentadas as características e os resultados dos sete experimentos utilizando-se os dados objetos desse estudo.

4.1.1. EXPERIMENTO 1 – CONJUNTO DE DADOS ORIGINAL

Nesse experimento, foi considerado o conjunto de atributos proposto para representação de estudantes sem nenhuma transformação ou pré-processamento. Desta maneira, foram utilizados os atributos apresentados no Quadro 1 (pág. 37), apenas considerando a aplicação do processo de discretização do atributo "Resultado_Final", conforme apresentado na Tabela 6.

O algoritmo RandomForest obteve uma taxa de acurácia de 74,3%, ou seja, 104 dos 140 estudantes foram classificados corretamente. A Tabela 7 demonstra a matriz de confusão desse experimento. Para facilitar a interpretação, a diagonal

principal da matriz de confusão, que apresenta as classificações corretas, foi destacada com fundo cinza. Esse destaque foi utilizado nas demais tabelas desse capítulo que apresentam outras matrizes de confusão.

Tabela 7 –Matriz de Confusão do algoritmo RandomForest no Experimento 1.

		Classes Previstas		
		A	B	C
Classes Corretas	A	2	14	0
	B	2	100	7
	C	0	13	2

Analisando-se os dados apresentados na Tabela 7, percebe-se que em ambas as classes “A” e “C” apenas dois estudantes foram corretamente classificados.

Na Tabela 8, são destacadas três medidas de desempenho do algoritmo RandomForest, permitindo a visualização de maiores detalhes dos resultados obtidos.

Tabela 8 -Medidas de Desempenho algoritmo RandomForest no Experimento 1.

Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,125	0,016	0,5
B	0,917	0,871	0,787
C	0,133	0,056	0,222
Média	0,743	0,686	0,694

O algoritmo MultilayerPerceptron obteve neste experimento uma taxa de acurácia de 80,7%, classificando corretamente 113 dos 140 estudantes. Para visualizar mais detalhes dessa classificação a Tabela 9 demonstra a matriz de confusão e a Tabela 10 apresenta algumas medidas de desempenho adicionais desse classificador.

Tabela 9 –Matriz de Confusão do algoritmo MultilayerPerceptron no Experimento 1.

		Classes Previstas		
		A	B	C
Classes Corretas	A	11	5	0
	B	7	95	7
	C	0	8	7

Tabela 10 -Medidas de Desempenho algoritmo MultilayerPerceptron no Experimento 1.

Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,688	0,056	0,611
B	0,872	0,419	0,88
C	0,467	0,056	0,5
Média	0,807	0,339	0,808

Analisando-se a matriz de confusão apresentada na Tabela 9, percebe-se que 11 dos 16 estudantes da classe “A” foram corretamente classificados, representando 68,7%. Já para a classe “C”, 7 dos 15 estudantes foram corretamente classificados, ou seja, 46,6%.

4.1.2. EXPERIMENTO 2 – CONJUNTO DE DADOS DISCRETIZADOS

Neste experimento foi utilizada uma versão do conjunto de dados experimental com todos os atributos transformados para valores discretos. Essa alternativa baseia-se no fato de que alguns algoritmos de mineração de dados, mesmo os capazes de trabalhar com valores contínuos, podem apresentar melhores resultados e maior velocidade de execução quando se utilizam valores discretos (Witten et al., 2011).

Desta maneira, foi executado processo de discretização supervisionado dos atributos da base de dados original utilizando-se o algoritmo padrão da ferramenta Weka, que é baseado em Fayyad e Irani (1993). Os resultados desse experimento são apresentados abaixo.

Utilizando os dados discretizados, 108 dos 140 estudantes (77,1%) foram classificados corretamente pelo o algoritmo RandomForest. A Tabela 11 com a matriz de confusão e a Tabela 12 com outras medidas de desempenho fornecem informações adicionais sobre o resultado desse experimento.

A aplicação do algoritmo MultilayerPercepetron neste experimento resultou em valores muito próximos aos obtidos pelo algoritmo RandomForest para a matriz de confusão e as demais medidas de desempenhos consideradas.

Tabela 11 –Matriz de Confusão do algoritmo RandomForest no Experimento 2.

		Classes Previstas		
		A	B	C
Classes Corretas	A	0	15	0
	B	0	108	1
	C	0	16	0

Tabela 12 -Medidas de Desempenho algoritmo RandomForest no Experimento 2.

Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0	0	0
B	0,991	1	0,777
C	0	0,008	0
Média	0,771	0,779	0,605

Um aspecto importante a ser destacado a partir da observação da matriz de confusão apresentada na Tabela 11 refere-se ao fato que nenhum dos estudantes das classes “A” e “C” foram corretamente classificados.

4.1.3. EXPERIMENTO 3 – SELEÇÃO DE ATRIBUTOS

Neste experimento utilizou-se uma base de dados reduzida considerando a técnica de seleção de atributos baseados em sua relevância para a estimativa de desempenho de um estudante.

Witten et al. (2011) destacam que muitos algoritmos são projetados para identificar os atributos mais apropriados a serem utilizados em suas decisões. Na prática, entretanto, os autores apontam que a inclusão de atributos irrelevantes em um conjunto de dados frequentemente tem um impacto negativo em algoritmos de aprendizagem de máquina. Desta maneira, com este experimento pretende-se investigar a possível presença e o impacto de atributos irrelevantes no conjunto de dados proposto.

Neste processo foi utilizado o algoritmo de seleção de atributos supervisionado padrão disponível na ferramenta Weka, baseado em Hall (1998). A execução desse algoritmo sobre o conjunto completo de atributos apresentados no Quadro 1 (pág. 37) resultou na identificação dos seguintes atributos: NR_MSG_REC_PROF, NR_QUEST_ACERT, NR_CHATS e NR_MSG_CHAT. Considerando-se apenas essa lista de atributos foram executados os algoritmos de classificação utilizados neste estudo, cujos resultados são descritos abaixo.

O algoritmo RandomForest obteve acurácia de 73,6%, classificando corretamente 103 dos 140 estudantes. A matriz de confusão destacada na Tabela 13 e as medidas demonstradas na Tabela 14 fornecem informações adicionais sobre os resultados desse experimento.

Tabela 13 –Matriz de Confusão do algoritmo RandomForest no Experimento 3.

		Classes Previstas		
		A	B	C
Classes Corretas	A	7	9	0
	B	8	94	7
	C	0	13	2

Analisando-se os dados apresentados na Tabela 13, percebe-se que na classe “A” foram classificados corretamente 7 dos 16 estudantes. Na classe “C” apenas 2 dos 15 estudantes foram corretamente classificados.

O algoritmo MultilayerPerceptron apresentou taxa de acurácia de 75%, ou seja, 105 dos 140 estudantes foram corretamente classificados. As Tabelas 15 e 16 destacadas na sequência apresentam, respectivamente, a matriz de confusão e algumas medidas adicionais que detalham os resultados do algoritmo MultilayerPerceptron neste experimento.

Tabela 14 -Medidas de Desempenho algoritmo RandomForest no Experimento 3.

Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,438	0,065	0,467
B	0,862	0,71	0,81
C	0,133	0,056	0,222
Média	0,736	0,566	0,708

Tabela 15 –Matriz de Confusão do algoritmo MultlayerPerceptron no Experimento 3.

		Classes Previstas		
		A	B	C
Classes Corretas	A	2	14	0
	B	5	100	4
	C	0	12	3

Tabela 16 -Medidas de Desempenho algoritmo MultilayerPerceptron no Experimento 3.

Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,125	0,04	0,286
B	0,917	0,839	0,794
C	0,2	0,032	0,429
Média	0,75	0,661	0,696

A matriz de confusão destacada na Tabela 15 demonstra que as classificações corretas nas classes “A” e “C” são pequenas. Na classe “A” 2 estudantes foram classificados corretamente, e na classe “C” este valor foi 3.

4.1.4. EXPERIMENTO 4 – BALANCEAMENTO DE CLASSES

O objetivo principal desse quarto experimento refere-se ao tratamento de um

problema conhecido como classificação de dados desbalanceados. Este problema ocorre quando o número de instâncias de uma classe é muito menor que o número de instâncias de outra classe (Gu et al., 2008).

Este problema torna-se mais relevante quando a classe ou classes com número reduzido de instâncias é aquela cuja classificação correta representa o maior interesse no contexto de utilização do processo de classificação.

Analisando-se a distribuição das classes apresentadas na Tabela 6, percebe-se claramente que essa situação fica caracterizada nos dados utilizados neste trabalho. Neste caso, a classe “B”, que agrupa estudantes com desempenho intermediário, concentra 109 das 140 instâncias dos dados experimentais.

Entretanto, considera-se de fundamental relevância a correta classificação de estudantes pertencentes à classe minoritária “C”, que agrupa os estudantes com desempenho inferior. Essa importância explica-se tendo em vista que os estudantes dessa classe teriam a maior probabilidade de reprovação. Diante disso, a correta identificação desse grupo de estudantes poderia servir de base para ações que busquem evitar reprovações ou melhorar o desempenho destes alunos.

Tradicionalmente, os algoritmos de classificação têm sido desenvolvidos com o objetivo de maximizar a taxa de acurácia global, que é independente da distribuição individual de cada classe. Conforme observam Han e Kamber (2006), as técnicas de classificação assumem que a distribuição das classes é balanceada e que os custos dos erros são iguais entre as diferentes classes. Esta situação pode ser observada analisando-se os resultados dos experimentos 1, 2 e 3. As tabelas com a matriz de confusão desses experimentos mostram que o percentual de estudantes corretamente classificados para as classes minoritárias “A” e “C” é significativamente menor em relação à classe majoritária “B”.

Uma alternativa para tratar essa questão consiste em aplicar técnicas de amostragem ou balanceamento de classes durante a fase de pré-processamento. Segundo Márquez-Vera, et al. (2011), uma técnica amplamente utilizada em aplicações de mineração de dados e que está disponível no Weka como um filtro de dados supervisionado é conhecida como SMOTE (*Synthetic Minority Over-sampling Technique*) (Chawla et al., 2002).

Essa técnica ajusta a frequência relativa entre classes majoritárias e minoritárias nos dados. Em linhas gerais, a técnica SMOTE consiste em introduzir, sinteticamente, instâncias de classes minoritárias, considerando a técnica de

agrupamento *k-nearest-neighbor (*Witten et al., 2011).

É possível definir dois parâmetros principais para a criação das instâncias sintéticas: percentual de sobreamostragem e número de vizinhos. Neste experimento a técnica SMOTE foi aplicada aos dados originais utilizando-se o valor 150 para o parâmetro “percentual de sobreamostragem”, de forma a aumentar significativamente (150%) o número de instâncias da classe “C”. Para o parâmetro “número de vizinhos” foi utilizado o valor 5, padrão sugerido pela ferramenta Weka.

Após a aplicação dessa técnica sobre o conjunto de dados original, mesmo utilizado no Experimento 1, o número de instâncias das classes ficou distribuído conforme apresentado na Tabela 17. Pode-se observar que a classe “C” que tinha originalmente 15 estudantes (ver Tabela 6) passou a ter 37, ou seja, um aumento de 150%. As demais classes permaneceram inalteradas. Dessa maneira, o conjunto de dados resultante, após a aplicação da técnica SMOTE, conta com um total de 162 instâncias.

Tabela 17 - Distribuição das classes após a aplicação da técnica SMOTE.

Título da Classe	Descrição	Número de Estudantes
A	Alunos com desempenho superior	16
B	Alunos com desempenho intermediário	109
C	Alunos com desempenho inferior	37

Tomando-se como base os dados transformados pela técnica SMOTE, foram aplicados os dois algoritmos utilizados neste trabalho. Os resultados obtidos são apresentados a seguir.

O algoritmo RandomForest alcançou 129 (79,6%) do total de 140 estudantes classificados corretamente. A matriz de confusão destacada na Tabela 18 e as medidas demonstradas na Tabela 19 fornecem informações adicionais sobre os resultados desse experimento.

Analisando-se o resultado da matriz de confusão apresentada na Tabela 18 pode-se observar que 29 estudantes foram classificados corretamente na Classe “C”. A ferramenta Weka permite analisar o resultado da classificação individualmente para cada instância dos dados do conjunto de testes. Através da análise do resultado da classificação individual desses 29 estudantes, observou-se que 9 são estudantes que fazem parte do conjunto de dados original da classe “C”, ou seja,

não são estudantes inseridos sinteticamente pela técnica SMOTE. Isso significa que 9 dos 15 estudantes “reais” da classe “C” foram classificados corretamente. No Experimento 1, o algoritmo RandomForest classificou corretamente apenas 2 estudantes da classe “C”, conforme observar na Tabela 7.

Tabela 18 –Matriz de Confusão do algoritmo RandomForest no Experimento 4.

		Classes Previstas		
		A	B	C
Classes Corretas	A	2	14	0
	B	2	98	9
	C	0	8	29

Tabela 19 -Medidas de Desempenho algoritmo RandomForest no Experimento 4.

Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,125	0,014	0,5
B	0,899	0,415	0,817
C	0,784	0,072	0,763
Média	0,796	0,297	0,773

Neste experimento, considerando-se o algoritmo MultilayerPerceptron, 122 (75,3%) dos 140 estudantes foram corretamente classificados. Resultados adicionais desse experimento podem ser encontrados na Tabela 20 que apresenta a matriz de confusão e na Tabela 21 que demonstra outras medidas de desempenho do algoritmo MultilayerPerceptron.

Analisando-se o resultado da matriz de confusão apresentada na Tabela 20 pode-se observar que 26 estudantes foram classificados corretamente na Classe “C”. Através da análise do resultado da classificação individual desses 26 estudantes na ferramenta Weka, pode-se observar que 10 são estudantes que fazem parte do conjunto de dados original da classe “C”, ou seja, não são estudantes inseridos sinteticamente pela técnica SMOTE. Isso significa que 10 dos 15 estudantes “reais” da classe “C” foram classificados corretamente. No Experimento 1, o algoritmo

MultilayerPerceptron classificou corretamente 7 estudantes da classe “C”, conforme observa-se na Tabela 9.

Tabela 20 –Matriz de Confusão do algoritmo MultlayerPerceptron no Experimento 4.

		Classes Previstas		
		A	B	C
Classes Corretas	A	6	10	0
	B	8	90	11
	C	0	11	26

Tabela 21 -Medidas de Desempenho algoritmo MultilayerPerceptron no Experimento 4.

Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,375	0,055	0,429
B	0,826	0,396	0,811
C	0,703	0,088	0,703
Média	0,753	0,292	0,748

4.1.5. EXPERIMENTO 5 – AVALIAÇÃO DE SÉRIES TEMPORAIS

Conforme apresentado na Revisão Bibliográfica, alguns trabalhos investigaram a aplicação de técnicas de mineração de dados com o objetivo de realizar inferências ou projeções a respeito do desempenho de estudantes. Entretanto, de maneira geral, são analisados os resultados considerando dados envolvendo todo o período de realização do curso, que só podem ser obtidos ao final do mesmo.

Diante disso, o experimento apresentado a seguir pretende investigar os resultados e a viabilidade de obter inferências sobre estimativas de desempenho durante a realização do curso. Nesse experimento, utilizaram-se dados do curso em análise, coletados em três intervalos de tempo distintos. Desta maneira, coletaram-se três diferentes conjuntos de dados, cada um contendo informações considerando

a data de início do curso até a data final de cada período em questão.

As aulas do curso utilizado neste estudo foram realizadas em um período de um ano. Diante disso, o primeiro período contém dados dos 120 primeiros dias de aulas do curso. O segundo e terceiro período envolvem, respectivamente, os primeiros 240 e 360 dias de aulas. Para a finalização do curso, os estudantes dispõem de 120 dias adicionais para a realização de um trabalho de conclusão. Este período final, não foi considerado neste experimento, pois não foram verificadas a disponibilização de atividades (e.g. fóruns, *chats*, questionários) no AVA para os estudantes nesse período.

Os conjuntos de dados considerados neste experimento são baseados nos atributos propostos no Quadro 1 (pág. 37), com a aplicação da técnica balanceamento das classes SMOTE apresentada no experimento 4. Essa escolha deve-se ao fato de que essa técnica trata o problema do desbalanceamento de classes presente nos dados experimentais. Da mesma forma que nos experimentos anteriores, foram utilizados os algoritmos RandomForest e MultilayerPerceptron neste experimento.

Utilizando o conjunto de dados do período 1 (120 dias), o algoritmo RandomForest conseguiu classificar corretamente 116 (71,6%) dos 140 estudantes. Na Tabela 22 é apresentada a matriz de confusão com os resultados do algoritmo RandomForest com os dados do período 1.

Tabela 22 –Matriz de Confusão do algoritmo RandomForest no período 1 do Experimento 5.

		Classes Previstas		
		A	B	C
Classes Corretas	A	3	13	0
	B	8	89	12
	C	1	12	24

Neste mesmo conjunto de dados (período 1), o algoritmo MultilayerPerceptron obteve acurácia de 71,6%, ou seja, 116 dos 140 estudantes foram classificados corretamente. A Tabela 23 apresenta a matriz de confusão com o resultado da aplicação do algoritmo MultilayerPerceptron sobre os dados do período 1 desse experimento.

Tabela 23 –Matriz de Confusão do algoritmo MultilayerPerceptron no período 1 do Experimento 5.

		Classes Previstas		
		A	B	C
Classes Corretas	A	3	11	2
	B	11	89	9
	C	1	12	24

A Tabela 24 apresenta outras medidas de desempenho obtidas pelos dois algoritmos utilizando o conjunto de dados do período 1 desse experimento.

Tabela 24 - Medidas de desempenho dos algoritmos no período 1 do Experimento 5.

	RadomForest			MultilayerPerceptron		
Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,188	0,062	0,25	0,188	0,082	0,2
B	0,817	0,472	0,781	0,817	0,434	0,795
C	0,649	0,096	0,667	0,649	0,088	0,686
Média	0,716	0,345	0,702	0,716	0,32	0,711

Tomando-se como base o conjunto de dados do período 2 (240 dias), o algoritmo RandomForest obteve os seguintes resultados: 122 (75,3%) dos 140 estudantes classificados corretamente. A matriz de confusão apresentada na Tabela 25 demonstra o desempenho do algoritmo RandomForest com os dados do período 2.

Neste mesmo conjunto de dados (período 2), o algoritmo MultilayerPerceptron alcançou 115 (71%) do total de 140 estudantes classificados corretamente. Resultados adicionais desse experimento podem ser encontrados na Tabela 26 que apresenta a matriz de confusão do algoritmo MultilayerPerceptron.

Na Tabela 27 são destacadas três medidas adicionais que podem ser utilizadas para avaliar o desempenho dos dois algoritmos utilizando o conjunto de dados do segundo período desse experimento.

Tabela 25 –Matriz de Confusão do algoritmo RandomForest no período 2 do Experimento 5.

		Classes Previstas		
		A	B	C
Classes Corretas	A	3	13	0
	B	5	95	9
	C	0	13	24

Tabela 26 –Matriz de Confusão do algoritmo MultilayerPerceptron no período 2 do Experimento 5.

		Classes Previstas		
		A	B	C
Classes Corretas	A	6	10	0
	B	13	86	10
	C	1	13	23

Tabela 27 - Medidas de desempenho dos algoritmos no período 2 do Experimento 5.

	RadomForest			MultilayerPerceptron		
Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,188	0,034	0,375	0,375	0,096	0,3
B	0,872	0,491	0,785	0,789	0,434	0,789
C	0,649	0,072	0,727	0,622	0,08	0,697
Média	0,753	0,35	0,731	0,71	0,32	0,72

Por fim, considerando-se o conjunto de dados do período 3 (360 dias), o algoritmo RandomForest classificou corretamente 129 (79,6%) do total de 140 estudantes. Na Tabela 28 a matriz de confusão com os resultados desse algoritmo é apresentada.

Nesse mesmo conjunto de dados (período 3), o algoritmo MultilayerPerceptron alcançou 120 (74,1%) dos 140 estudantes classificados corretamente. Mais detalhes do resultado desse algoritmo encontram-se na matriz de confusão da Tabela 29.

Tabela 28 –Matriz de Confusão do algoritmo RandomForest no terceiro período do Experimento 5.

		Classes Previstas		
		A	B	C
Classes Corretas	A	3	13	0
	B	5	101	3
	C	0	12	25

Tabela 29 –Matriz de Confusão do algoritmo MultilayerPerceptron no terceiro período do Experimento 5.

		Classes Previstas		
		A	B	C
Classes Corretas	A	6	10	0
	B	12	88	9
	C	0	11	26

Com o objetivo de fornecer detalhes adicionais sobre o desempenho dos dois algoritmos utilizando o conjunto de dados do terceiro período desse experimento apresenta-se a Tabela 30.

Tabela 30 - Medidas de desempenho dos algoritmos no terceiro período do Experimento 5.

	RadomForest			MultilayerPerceptron		
Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,188	0,034	0,188	0,375	0,082	0,333
B	0,927	0,472	0,802	0,807	0,396	0,807
C	0,676	0,024	0,893	0,703	0,072	0,743
Média	0,796	0,326	0,78	0,741	0,291	0,746

Uma particularidade a ser destacada neste experimento refere-se ao caso de um estudante que pertence a classe “C” e foi incorretamente classificado na classe “A”. Esta situação pode ser observada na matriz de confusão apresentada nas

Tabelas 22, 23 e 26.

4.1.6. EXPERIMENTO 6 – DISTRIBUIÇÃO DOS ESTUDANTES NAS CLASSES

No processo de discretização do atributo “Resultado_Final”, descrito no início deste capítulo, optou-se por agrupar aproximadamente 10% dos estudantes com as maiores e menores notas, respectivamente, nas classes “A” e “C”. Na descrição do processo de discretização foram destacados elementos que motivaram essa escolha, salientando-se que esse parâmetro poderia influenciar os resultados do processo de classificação.

Dessa maneira, a realização deste experimento tem como objetivo avaliar qual o impacto que mudanças no percentual de distribuição dos estudantes nas classes têm nos resultados do processo de classificação.

Para a realização desse experimento, foi utilizado o conjunto de dados original com os atributos descritos no Quadro 1 (pág. 37). Apenas os valores do atributo “Resultado_Final” foi discretizado, conforme procedimento descrito no início deste capítulo.

Nesse experimento, foram realizadas duas simulações de distribuição dos estudantes nas classes. Na primeira simulação, foi utilizado percentual de 5% dos estudantes para cada uma das classes “A” e “C” e os 90% restantes na classe “B”. A Tabela 31 mostra como ficou a distribuição das classes dos 140 estudantes utilizando-se os percentuais dessa primeira simulação. Nessa tabela, são apresentados o título e a descrição das classes, bem como o número de estudantes e o intervalo de notas em cada uma das três classes.

Utilizando o conjunto de dados com a distribuição das classes apresentado na Tabela 31, o algoritmo RandomForest classificou corretamente 126 (90%) dos 140 estudantes. Na Tabela 32, apresenta-se a matriz de confusão com o resultado obtido pelo algoritmo RandomForest com distribuição de 5% nas classes “A” e “C”.

O algoritmo MultilayerPerceptron, nesse mesmo conjunto de dados (5%), classificou corretamente 116 (82,9%) dos 140 estudantes. Mais detalhes desse experimento podem ser visualizados na Tabela 33 que apresenta a matriz de

confusão com os resultados do algoritmo MultilayerPerceptron.

Tabela 31 - Distribuição dos estudantes com 5% para cada uma das classe “A” e “C”.

Título da Classe	Descrição	Número de Estudantes	Intervalo de Notas
A	Alunos com desempenho superior	7	Maior que 91
B	Alunos com desempenho intermediário	126	Entre 75 e 91
C	Alunos com desempenho inferior	7	Menor que 75

Tabela 32 – Matriz de confusão do algoritmo RandomForest com distribuição de 5% nas classes “A” e “C”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	0	7	0
	B	0	126	0
	C	0	7	0

Tabela 33 –Matriz de Confusão do algoritmo MultilayerPerceptron com distribuição de 5% nas classes “A” e “C”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	1	6	0
	B	8	114	4
	C	0	6	1

A Tabela 34 resume três medidas de desempenho dos algoritmos que demonstram os resultados obtidos no processo de classificação individualizado por classe nessa primeira simulação (5%).

Para a segunda simulação foi utilizado percentual de 15% dos estudantes para cada uma das classes “A” e “C” e os 70% restantes na classe “B”. A Tabela 35 mostra como ficou a distribuição das classes dos 140 estudantes utilizando-se os percentuais dessa segunda simulação. Nessa tabela, são apresentados o título e a descrição das classes, bem como o número de estudantes e o intervalo de notas em

cada uma das três classes.

Tabela 34 - Medidas de desempenho dos algoritmos com distribuição de 5% nas classes “A” e “C”.

	RadomForest			MultilayerPerceptron		
Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0	0	0	0,143	0,06	0,111
B	1	1	0,9	0,905	0,857	0,905
C	0	0	0	0,143	0,03	0,2
Média	0,9	0,9	0,81	0,829	0,776	0,83

Tabela 35 - Distribuição dos estudantes com 15% para cada uma das classe “A” e “C”.

Título da Classe	Descrição	Número de Estudantes	Intervalo de Notas
A	Alunos com desempenho superior	22	Maior que 87
B	Alunos com desempenho intermediário	94	Entre 78 e 87
C	Alunos com desempenho inferior	24	Menor que 78

Utilizando o conjunto de dados com a distribuição das classes apresentado na Tabela 35, o algoritmo RandomForest classificou corretamente 94 (67,1%) dos 140 estudantes. A matriz de confusão com os resultados do algoritmo RandomForest é mostrada na Tabela 36.

Tabela 36 – Matriz de confusão do algoritmo RandomForest com distribuição de 15% nas classes “A” e “C”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	5	17	0
	B	3	80	11
	C	0	15	9

O algoritmo MultilayerPerceptron, nesse mesmo conjunto de dados (15%), classificou 87 (62,1%) dos 140 estudantes corretamente. Mais detalhes desse

experimento podem ser visualizados na Tabela 37 que apresenta a matriz de confusão com os resultados desse algoritmo.

Tabela 37 –Matriz de Confusão do algoritmo MultilayerPerceptron com distribuição de 15% nas classes “A” e “C”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	8	14	0
	B	15	69	10
	C	0	14	10

Outras três medidas que indicam o desempenho do processo de classificação, individualizado por classes com distribuição de 15% nas classes “A” e “C”, são apresentadas na Tabela 38.

Tabela 38 - Medidas de desempenho dos algoritmos com distribuição de 15% nas classes “A” e “C”.

	RadomForest			MultilayerPerceptron		
Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,227	0,025	0,625	0,364	0,127	0,348
B	0,851	0,696	0,714	0,734	0,609	0,711
C	0,375	0,095	0,45	0,417	0,086	0,5
Média	0,671	0,487	0,655	0,621	0,443	0,618

4.1.7. EXPERIMENTO 7 – AVALIAÇÃO INDIVIDUAL DAS DIMENSÕES

Para a seleção dos atributos que foram utilizados para representar os estudantes em um AVA apresentados no Quadro 1 (pág. 37) foram consideradas três dimensões: perfil de uso do AVA, interação estudante-estudante e interação bidirecional estudante-professor.

Com exceção do experimento 3, os demais experimentos apresentados até aqui foram realizados utilizando-se um conjunto de atributos completos, incluindo as

três dimensões apresentadas acima. Esse experimento foi desenvolvido com o objetivo de avaliar as inferências sobre o desempenho de estudantes que podem ser obtidas tomando-se cada uma das dimensões isoladamente.

Dessa maneira, nesse experimento foram utilizados três conjuntos de dados, cada um contendo apenas os atributos específicos de cada dimensão. Os três conjuntos de dados contém o atributo “Resultado_Final”, pois este representa a classe objetivo do processo de classificação. Além disso, os valores desse atributo foram discretizados, conforme procedimento descrito no início desse capítulo. Assim como nos demais experimentos, utilizaram-se os algoritmos RandomForest e MultilayerPerceptron.

O primeiro conjunto de dados contém os dez atributos da dimensão “Perfil de Uso do AVA.Utilizando o conjunto de dados original com os atributos dessa, o algoritmo RandomForest classificou corretamente 108 (77,1%) dos 140 estudantes. Na Tabela 39 apresenta-se a matriz de confusão com os resultados do algoritmo RandomForest.

Tabela 39 – Matriz de confusão do algoritmo RandomForest para a dimensão “Perfil de Uso do AVA”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	4	12	0
	B	3	100	6
	C	0	11	4

O algoritmo MultilayerPerceptron, nesse mesmo conjunto de dados (dimensão “Perfil de uso do AVA”), classificou corretamente 105 (75%) dos 140 estudantes. Mais detalhes desse experimento podem ser visualizados na Tabela 40 que apresenta a matriz de confusão com os resultados do algoritmo MultilayerPerceptron.

Para destacar mais detalhes sobre os resultados obtidos pelos dois algoritmos utilizados a Tabela 41 é apresentada. Nessa tabela são apresentadas medidas de desempenho dos algoritmos para cada uma das três classes considerando-se a dimensão “Perfil de Uso do AVA”.

Tabela 40 –Matriz de Confusão do algoritmo MultilayerPerceptron para a dimensão “Perfil de Uso do AVA”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	3	13	0
	B	7	98	4
	C	0	11	4

Tabela 41 - Medidas de Desempenho para a dimensão “Perfil de Uso do AVA”.

	RadomForest			MultilayerPerceptron		
Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,25	0,024	0,571	0,188	0,056	0,3
B	0,917	0,742	0,813	0,899	0,774	0,803
C	0,267	0,048	0,4	0,267	0,032	0,5
Média	0,771	0,586	0,741	0,75	0,613	0,713

O segundo conjunto de dados deste experimento é composto pelos 4 atributos da dimensão “Interação Estudante-Estudante. Utilizando esse segundo conjunto de dados, o algoritmo RandomForest alcançou 96 (68,6%) dos 140 estudantes classificados corretamente. A matriz de confusão destacada na Tabela 42 apresenta os resultados detalhados obtidos pelo algoritmo RandomForest.

Tabela 42 –Matriz de Confusão do algoritmo RandomForest para a dimensão “Interação Estudante-Estudante”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	3	13	0
	B	9	90	10
	C	0	12	3

O algoritmo MultilayerPerceptron, utilizando o conjunto de dados da dimensão “Interação Estudante-Estudante”, classificou corretamente 103 (73,6%) dos 140

estudantes. Resultados adicionais desse experimento podem ser encontrados na Tabela 43 que apresenta a matriz de confusão do algoritmo MultilayerPerceptron.

Tabela 43 –Matriz de Confusão do algoritmo MultlayerPerceptron para a dimensão “Interação Estudante-Estudante”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	0	16	0
	B	3	103	3
	C	0	105	0

Para destacar mais detalhes sobre os resultados obtidos pelos dois algoritmos utilizados a Tabela 44 é apresentada. Nessa tabela são apresentadas medidas de desempenho dos algoritmos para cada uma das três classes considerando-se a dimensão “Interação Estudante-Estudante”.

Tabela 44 - Medidas de Desempenho para a dimensão “Interação Estudante-Estudante”.

	RadomForest			MultilayerPerceptron		
Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,188	0,073	0,25	0	0,024	0
B	0,826	0,806	0,783	0,945	1	0,769
C	0,2	0,08	0,231	0	0,024	0
Média	0,686	0,645	0,663	0,736	0,784	0,598

Para o terceiro conjunto de dados deste experimento foram utilizados os 4 atributos da dimensão “Interação bidirecional Estudante-Professor”. Utilizando-se esse terceiro conjunto de dados, o algoritmo RandomForest obteve uma taxa de acurácia global de 65%, ou seja, 91 dos 140 estudantes foram classificados corretamente. A Tabela 45 demonstra a matriz de confusão desse experimento.

O algoritmo MultilayerPerceptron obteve neste experimento uma taxa de acerto de 76,4%, classificando corretamente 107 dos 140 estudantes. Para visualizar mais detalhes dessa classificação a Tabela 46 demonstra a matriz de confusão desse classificador.

Tabela 45 –Matriz de Confusão do algoritmo RandomForest para a dimensão “Interação bidirecional Estudante-Professor”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	2	13	1
	B	9	88	12
	C	1	13	1

Tabela 46 –Matriz de Confusão do algoritmo MultilayerPerceptron para a dimensão “Interação bidirecional Estudante-Professor”.

		Classes Previstas		
		A	B	C
Classes Corretas	A	0	16	0
	B	0	107	2
	C	0	15	0

A Tabela 47 apresenta algumas medidas de desempenho adicionais dos dois algoritmos com os dados da dimensão “Interação bidirecional Estudante-Professor”, permitindo a visualização de maiores detalhes dos resultados obtidos.

Tabela 47 - Medidas de Desempenho para a dimensão “Interação bidirecional Estudante-Professor”.

	RadomForest			MultilayerPerceptron		
Classe	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão	Taxa Verdadeiro Positivo	Taxa Falso Positivo	Precisão
A	0,125	0,081	0,167	0	0	0
B	0,807	0,839	0,772	0,982	1	0,775
C	0,067	0,104	0,071	0	0,016	0
Média	0,65	0,673	0,628	0,764	0,78	0,604

5. ANÁLISE DOS RESULTADOS

A análise dos resultados do estudo será feita comparando-se inicialmente os resultados dos experimentos 1, 2, 3 e 4, pois esses experimentos utilizam conjuntos de dados que envolvem o período completo de realização do curso. Na sequência, serão destacadas também algumas considerações sobre os resultados dos experimentos 5, 6 e 7, já que nesses casos são utilizados conjuntos de dados com diferentes características (períodos de tempo, distribuição das classes e dimensões).

5.1. ANÁLISE DOS RESULTADOS: EXPERIMENTOS 1 A 4

A Tabela 48 apresenta um resumo com os resultados dos experimentos 1, 2, 3 e 4 executados sobre o conjunto de dados contendo o período completo de realização do curso. Nela, o percentual de acurácia global médio e o desvio padrão de 100 execuções são apresentados para cada algoritmo e experimento. Para a estratificação da base de dados em conjuntos de teste e treinamento foi usado o método *"K-fold Cross-Validation*", apresentado no Capítulo 2, com o parâmetro K igual a 10.

Tabela 48 - Acurácia média e desvio padrão em 100 execuções dos classificadores utilizados nos experimentos 1, 2, 3 e 4.

Classificador	Experimento 1	Experimento 2	Experimento 3	Experimento 4
RandomForest	77,4 ± 7,78	77,2 ± 2,99	72,7 ± 9,92	**78,4 ± 8,48**
MultilayerPerceptron	**80,1 ± 8,88**	77,2 ± 2,99	76,9 ± 8,07	77,1 ± 9,83

Avaliando-se os resultados descritos na Tabela 48, observa-se que, em números absolutos, no experimento 1 que utiliza um conjunto de atributos completo e sem a aplicação de nenhuma técnica de transformação, o algoritmo MultilayerPerceptron obteve 80,1% de acurácia. Esse foi o melhor resultado em termos de acurácia em todos os experimentos do estudo realizado. Por sua vez, o algoritmo RandomForest alcançou a melhor taxa de acurácia (78,4%) no experimento 4, utilizando uma base de dados com as classes balanceadas pela

técnica SMOTE.

Uma observação pode ser feita em relação ao experimento 3 que considerou apenas o conjunto dos 4 atributos mais relevantes apresentados na Seção 4.1.3. Conforme apresentado na quarta coluna da Tabela 48, as taxas de acurácia de 72,7% e 76,9% obtidas, respectivamente, pelos algoritmos RandomForest e MultilayerPerceptron são as menores, em números absolutos, dentre os quatro experimentos.

Com o objetivo de testar a significância estatística dos resultados obtidos, utilizou-se a técnica de teste estatístico conhecida como “T-pareado” - *pair-wise T-Test* (Witten et al., 2011) com nível de significância de 5%. Para a realização desse teste, utilizou-se o ambiente *Weka Experiment Environment* – WEE, disponível na ferramenta Weka.

A partir do resultado do teste “T-pareado”, observou-se que não existe diferença, considerando o nível de significância de 5%, entre os resultados dos quatro experimentos apresentados na Tabela 48.

É importante destacar que as medidas de desempenho “Taxa de Verdadeiro Positivo”, “Taxa de Falso Positivo” e “Precisão” dos classificadores variaram significativamente entre as diferentes classes em todos os experimentos do estudo. Esta constatação pode ser feita avaliando-se os resultados das medidas de desempenho dos algoritmos apresentadas nos experimentos 1, 2, 3 e 4.

A Tabela 49 resume as taxas de verdadeiro positivo obtidas para cada classe nesses quatro experimentos, permitindo a comparação das diferenças entre os resultados de cada classe. Nessa tabela, observa-se que as taxas de verdadeiro positivo para classe “B” ficaram entre 99% (experimento 2) e 83% (experimento 4), sendo sempre maiores em relação às outras classes. As classes “C” e “A” apresentaram índices menores, chegando a 0% no experimento 2.

De acordo com o que foi mencionado na Seção 4.1.4, identificar corretamente os estudantes da classe “C” poderia ser relevante para um professor que deseja desenvolver estratégias pedagógicas para diminuir a reprovação, por exemplo. Com esse objetivo o experimento 4 foi realizado aplicando-se a técnica SMOTE ao conjunto de dados original. Os resultados podem ser observados na última coluna da Tabela 49. No experimento 4, o algoritmo RandomForest alcançou taxa de verdadeiro positivo de 78% e o MultilayerPerceptron 70% para a classe “C”. O valor dessa medida mais próximo a esses foi 47%, obtido no experimento 1 com o

algoritmo MultilayerPerceptron. A Figura 6 demonstra graficamente as taxas de verdadeiro positivo obtidas para a Classe “C” nos quatro experimentos realizados, demonstrando que no experimento 4 as taxas foram superiores.

Tabela 49 – Taxas de verdadeiro positivo obtidas para cada classe nos experimentos 1 a 4.

Classe	Classificador	Taxa Verdadeiro Positivo Experimento 1	Taxa Verdadeiro Positivo Experimento 2	Taxa Verdadeiro Positivo Experimento 3	Taxa Verdadeiro Positivo Experimento 4
A	RandomForest	13%	0%	44%	13%
	MultilayerPerceptron	69%	0%	13%	38%
B	RandomForest	92%	99%	86%	90%
	MultilayerPerceptron	87%	99%	92%	83%
C	RandomForest	13%	0%	13%	78%
	MultilayerPerceptron	47%	0%	20%	70%

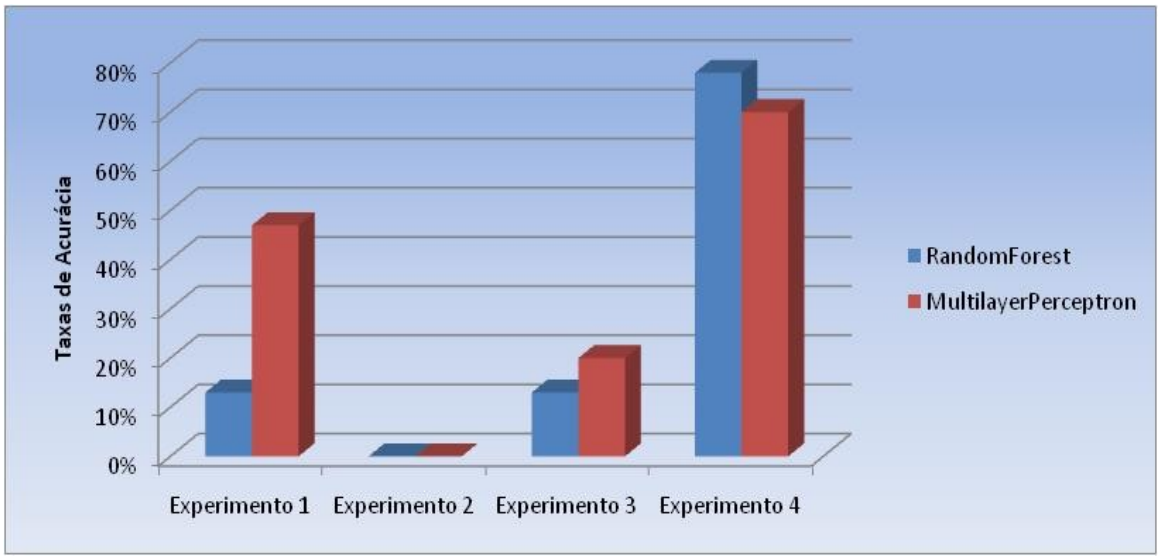

Figura 6 - Taxas de verdadeiro positivo obtidas para a Classe "C" nos experimentos 1 a 4

5.2. ANÁLISE DOS RESULTADOS: EXPERIMENTOS 5 a 7

No experimento 5 são analisadas as possibilidades de realização de inferências sobre o desempenho em diferentes etapas do curso. A Figura 7

apresenta graficamente os resultados obtidos nesse experimento. Nessa figura são apresentados os percentuais de acurácia dos algoritmos nos três conjuntos de dados representando os diferentes períodos desse experimento. Para fins de comparação, foram incluídos no gráfico da Figura 7 os valores obtidos no experimento 4, que considera os dados do período completo do curso.

Analisando-se os dados apresentados na Figura 7, podem-se observar as diferenças entre os resultados obtidos nos quatro períodos de realização do curso, incluindo o período completo. Para o algoritmo RandomForest, a diferença máxima é de 8% entre o período 1 e o final do curso. No caso do algoritmo MultilayerPerceptron a maior diferença foi 4,3%, sendo observada entre o período 2 e o final do curso.

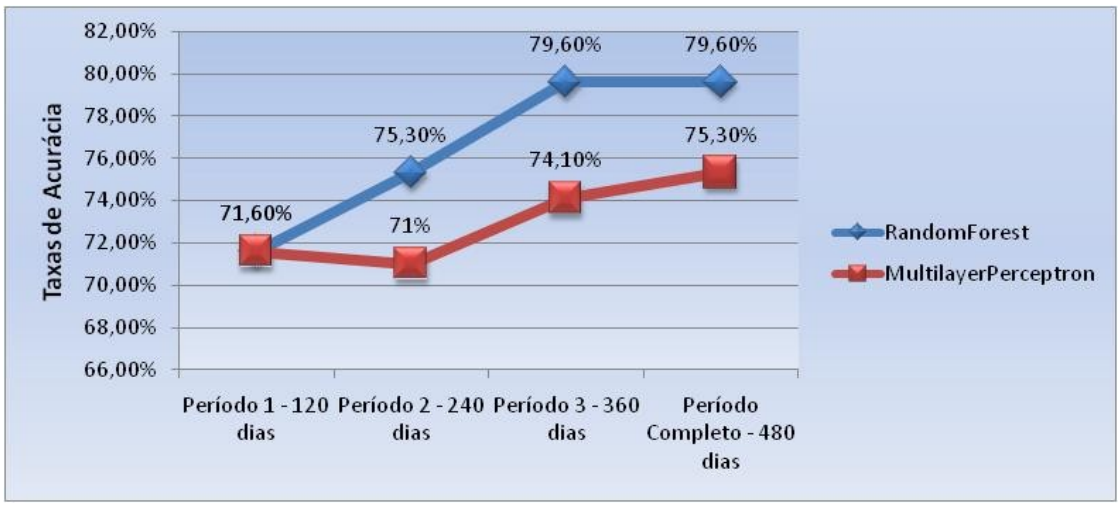

Figura 7 - Taxas de acurácia obtidas nos diferentes períodos do experimento 5.

A distribuição do número de estudantes em cada um das três classes utilizadas neste trabalho potencialmente poderia influenciar os resultados dos classificadores, conforme discutido no início do Capítulo 4. No experimento 6, foram simulados dois cenários distintos de distribuição dos estudantes nas classes. Na primeira simulação utilizou-se 5% dos estudantes em cada uma das classes “A” e “C”. Na segunda simulação alterou-se este percentual para 15%.

Os resultados do experimento 6 demonstraram que utilizando o percentual de 5% para as classes “A” e “C” as taxas de acurácia global foram superiores: 90% para o algoritmo RandomForest e 82,9% para o algoritmo MultilayerPerceptron. Com o percentual de 15% para as classes “A” e “C” a acurácia global diminuiu para 67,1% e 62,1% para os algoritmos RandomForest e MultilayerPercetron, respectivamente.

Além da acurácia global, as medidas de desempenho “Taxa de Verdadeiro Positivo”, “Taxa de Falso Positivo” e “Precisão” dos classificadores para cada classe foram diferentes nas duas simulações, conforme se pode observar na Tabela 50. Nessa tabela as taxas unitárias de precisão são apresentadas de maneira individual para cada classe e algoritmo nas duas simulações de distribuição dos estudantes nas classes.

Analisando-se os dados da Tabela 50, pode-se observar que a precisão das inferências na simulação com 15% dos estudantes para as classes “A” e “C” é melhor distribuída entre as três classes quando comparadas com a simulação considerando 5%. Na simulação considerando 5% dos estudantes para as classes “A” e “C”, a precisão dessas classes foi significativamente menor, chegando a zero para essas classes no caso do algoritmo RandomForest.

Tabela 50 - Taxas unitárias de precisão por classe do Experimento 6.

Distribuição Percentual de cada classe “A” e “C”	Classe	Precisão Algoritmo RandomForest	Precisão Algoritmo MultilayerPerceptron
5%	A	0	0,111
	B	0,9	0,905
	C	0	0,2
15%	A	0,625	0,348
	B	0,714	0,711
	C	0,45	0,5

Para selecionar os atributos que fizeram parte do conjunto de dados utilizado para representar os estudantes em um AVA, três dimensões foram consideradas: perfil de uso do AVA, interação estudante-estudante e interação bidirecional estudante-professor. Diferentemente dos demais, o experimento 7 foi realizado considerando cada uma das três dimensões individualmente. Na Tabela 51 são apresentadas as taxas de acurácia obtidas no experimento 7 para cada dimensão tomada individualmente.

As informações apresentadas na Tabela 51 permitem observar que, mesmo considerando apenas os atributos de cada classe individualmente, foi possível obter taxas de acurácia entre 65% e 77,1%. Para efeito de comparação, no experimento 1, utilizando todos os atributos das três dimensões, as taxas de acurácia ficaram entre 74,3% e 80,7%.

Tabela 51 - Taxas de acurácia de cada dimensão obtidas no experimento 7.

Dimensão	Acurácia RandomForest	Acurácia MultilayerPerceptron
Perfil de Uso do AVA	77,1%	75%
Interação Estudante-Estudante	68,6%	73,6%
Interação bidirecional Estudante-Professor	65%	76,4%

5.3. IMPLICAÇÕES

Em busca de respostas ao problema de pesquisa apresentado, os resultados dos sete experimentos do estudo realizado demonstram algumas possibilidades de utilização dos atributos ilustrados no Quadro 1 (pág. 37) para representar estudantes em um AVA e gerar inferências sobre o desempenho acadêmico desses estudantes.

Além da gerar inferências, um objetivo definido neste trabalho foi verificar a precisão dessas inferências. Nos experimentos realizados neste estudo foi possível avaliar também o impacto da aplicação de técnicas como, discretização, balanceamento de classes e relevância de atributos, na precisão das inferências sobre o desempenho dos estudantes. Nesses experimentos, avaliou-se também a viabilidade de obtenção e a precisão das inferências sobre o desempenho de estudantes em três intervalos de tempo distintos, durante a realização do curso.

Analisando-se os resultados dos sete experimentos realizados, percebe-se que as taxas de acurácia global variaram entre 72% e 80%, podendo ser consideradas expressivas no contexto educacional. Essa afirmação baseia-se no resultado de Hämäläinen e Vinni (2011) que, utilizando pesquisas envolvendo estimativas de desempenho, realizaram um estudo para verificar as taxas de acurácia obtidas nessas pesquisas. Os referidos autores relatam que, devido às características particulares dos estudos, existem variações entre os resultados observados. Porém, destacam que a média da acurácia foi de 72% entre as pesquisas consideradas.

Tendo em vista que a inclusão de atributos irrelevantes poderia impactar negativamente o processo de obter as inferências, avaliar a relevância do conjunto de atributos selecionado e eliminar atributos irrelevantes constitui um objetivo deste trabalho. A técnica utilizada no experimento 3 demonstrou que taxas de acurácia

entre 72,7% e 76,9% podem ser obtidas considerando apenas os quatro atributos mais relevantes entre os 18 atributos ilustrados no Quadro 1 (pág. 37): NR_ MSG_REC_PROF, NR_QUEST_ACERT, NR_CHATS e NR_MSG_CHAT. Entretanto, as taxas de acurácia dos experimentos que utilizam todos os atributos ilustrados no Quadro1 (pág. 36) são maiores, em números absolutos. Isso indica a viabilidade de utilizar todos os atributos selecionados, sem a necessidade de eliminação de atributos com menor relevância.

Avaliando-se os resultados do estudo com os sete experimentos e a análise dos resultados apresentada neste capítulo, pode-se concluir que o Experimento 4 foi o que teve os melhores resultados. Mesmo não sendo o experimento que obteve as melhores taxas de acurácia global, esse experimento destaca-se pelo fato de obter melhores taxas de classificação para a classe “C”. Conforme destacado anteriormente, pode-se considerar que isso é de fundamental importância, pois identificar corretamente os estudantes da classe “C” poderia auxiliar professores na realização de ações pedagógicas no sentido de evitar reprovação ou melhorar o desempenho.

É importante ressaltar também que os resultados obtidos pelos algoritmos RandomForest e MultilyerPerceptron são semelhantes em todos os cenários investigados nos sete experimentos do estudo realizado. Diante disso, o algoritmo RandomForest poderia ser considerado mais adequado no contexto tratado neste trabalho, pois apresenta um tempo de processamento menor. Além disso, diferentemente do algoritmo MultilayerPerceptron, que é do tipo *black-box*, a estrutura do padrão da solução encontrada pelo algoritmo RandomForest pode ser facilmente entendida por pessoas através da geração de uma árvore de decisão.

6. CONCLUSÃO E TRABALHO FUTUROS

Acompanhar de maneira detalhada e individual estudantes em cursos EAD tem se mostrado um desafio para profissionais e instituições que atuam nesta modalidade de ensino. É consenso entre os autores citados neste trabalho que a implementação de processos efetivos de acompanhamento dos estudantes tem relação direta com a qualidade dos cursos. Diante disso, esforços têm sido feitos pela comunidade científica no desenvolvimento de soluções tecnológicas que forneçam informações relevantes para auxiliar a gestão do processo de ensino desses cursos.

Neste trabalho investigou-se como os dados armazenados por um AVA poderiam ser transformados em informações potencialmente úteis para apoiar o acompanhamento de estudantes em cursos EAD. As informações geradas foram inferências envolvendo estimativas de desempenho acadêmico futuro de estudantes.

Para a geração das referidas inferências, inicialmente foi necessário analisar quais informações disponíveis em uma base de dados de um AVA poderiam ser utilizadas para representar aprendizes realizando um curso EAD. Esse objetivo foi alcançado por meio do uso de três dimensões, a saber: perfil de uso do AVA, interação estudante-estudante e interação bidirecional estudante-professor. As três dimensões mencionadas serviram como base para a seleção de um conjunto de atributos, apresentados no Quadro 1 (pág. 37), que foram utilizados como referência para a geração das inferências sobre o desempenho acadêmico dos estudantes.

Os resultados obtidos com a aplicação de técnicas de mineração de dados sobre o conjunto de atributos selecionados demonstram que é possível obter inferências relativas ao desempenho dos estudantes com taxas de acurácia global variando entre 72% e 80%. Espera-se que a disponibilização dessas inferências a professores ou gestores possam ser úteis para o desenvolvimento de ações de monitoramento ou para o desenvolvimento de estratégias pedagógicas que busquem auxiliar os estudantes a melhorar o desempenho no curso.

Entretanto, apenas a taxa de acurácia global pode ser insuficiente para avaliar a qualidade do modelo de classificação, principalmente quando o número de instâncias das classes é desbalanceado, como no caso desse estudo. A utilização

da técnica de balanceamento SMOTE, utilizada no quarto experimento, demonstrou ser eficaz para a correta classificação de classes minoritárias. Nesse experimento, foi possível atingir taxas de acurácia entre 70 e 78% para a classe minoritária “C”, que agrupa os estudantes com desempenho inferior. A correta identificação desse grupo de estudantes poderia ser útil para o desenvolvimento de ações pedagógicas que busquem diminuir as reprovações. Como exemplo, uma ação pedagógica poderia consistir em incentivar os estudantes da classe “C” a realizar atividades de revisão ou recuperação de conteúdos.

Um aspecto importante pode ser destacado a partir do resultado do experimento 3. Conforme definição apresentada na seção 4.1.3, esse experimento utiliza um conjunto reduzido de atributos, identificado por técnica de seleção de atributos mais relevantes. As taxas de acurácia de 72,6% e 76,9% obtidas, respectivamente, pelos algoritmos RandomForest e MultilayerPercpetron são estatisticamente equivalentes às taxas obtidas nos experimentos que utilizam o conjunto completo de atributos. Trabalhar com um conjunto reduzido de atributos, além de diminuir a complexidade do processamento dos dados, poderia facilitar a replicação deste estudo utilizando AVA que não dispõe de todos os atributos utilizados neste trabalho.

Por outro lado, salienta-se que com a utilização do conjunto de atributos completos (experimentos 1, 2 e 4) apresentados no Quadro 1 (pág. 37), as taxas de acurácia não foram prejudicadas, ficando entre 77,2% e 80,1%. Isso indica que os dois algoritmos utilizados foram eficientes na identificação de possíveis atributos irrelevantes que poderiam degradar o seu desempenho. Esse fato pode indicar a viabilidade da utilização de um conjunto amplo de atributos para representação de estudantes.

Os resultados obtidos no experimento 5 indicam a viabilidade de se realizar inferências relativas ao desempenho de estudantes, obtendo-se taxas de acurácia próximas a 72%, mesmo em etapas iniciais de realização do curso. Estas informações poderiam ser úteis para o desenvolvimento de ações envolvendo estudantes da turma em andamento e não apenas de turmas futuras. Essas inferências, disponibilizadas em períodos iniciais do curso, poderiam ainda auxiliar professores no acompanhamento individual de estudantes e no desenvolvimento de estratégias pedagógicas que possam melhorar o desempenho.

A distribuição do número de estudantes em cada classe teve impacto nas

taxas de acurácia global e individual de cada classe, conforme se pode observar analisando os resultados do sexto experimento. Essa constatação aponta para a importância da definição desse parâmetro, que deve estar relacionado com o objetivo a ser alcançado com o processo de classificação. Nesse trabalho, considerou-se relevante a identificação de padrões que poderiam levar a desempenhos diferenciados, principalmente o desempenho inferior (classe C). Dessa maneira, as classes “A” e “C” foram definidas com uma distribuição de 10% dos estudantes em cada uma, fazendo com que as classes fiquem desbalanceadas. Para tratar o problema do desbalanceamento, uma alternativa que apresentou resultados promissores foi apresentada no quarto experimento.

A seleção dos atributos para representar os estudantes em um curso a distância utilizando um AVA foi baseada em três dimensões. Analisando-se os resultados do sétimo experimento, pode-se confirmar que as três dimensões são relevantes para a geração de estimativas sobre o desempenho acadêmico de estudantes. Os resultados desse experimento mostraram que, mesmo considerando-se as dimensões individualmente, foi possível obter taxas de acurácia entre 65% e 77,1%. Essas taxas ficam próximas às obtidas nos demais experimentos, que ficaram entre 72% e 80%.

Levando-se em consideração os resultados apresentados e discutidos até aqui, pode-se concluir que a principal contribuição deste trabalho foi demonstrar a viabilidade e apresentar os procedimentos que podem ser utilizados para obter inferências relativas ao desempenho de estudantes em um curso EAD. Considerando-se os resultados do estudo realizado, pode-se concluir que para a obtenção dessas inferências as seguintes etapas poderiam ser seguidas:

1. Extração do conjunto de atributos apresentados no Quadro 1 (pág. 37)
2. Discretização do atributo “Resultado_Final” para facilitar o processamento e entendimento dos resultados.
3. Aplicação da técnica SMOTE para balancear as classes, caso forem desbalanceadas, dando ênfase às classes que se tem especial interesse (neste trabalho enfatizou-se a classe “C”).
4. Aplicação do algoritmo RandomForest utilizando a base balanceada pela técnica SMOTE para classificação dos estudantes.
5. Análise dos resultados da classificação utilizando a matriz de confusão e medidas como, precisão e taxas de verdadeiro e falso positivo.

6. Disponibilização dos resultados a professores, tutores ou gestores de cursos.

Considerando-se os resultados obtidos neste trabalho e também suas limitações, vislumbram-se algumas perspectivas de investigações e desenvolvimentos futuros que serão apresentados a seguir.

Para verificar a generalização do conjunto de atributos proposto nesse trabalho, futuras pesquisas poderiam ser desenvolvidas considerando a aplicação do estudo apresentado neste trabalho em diferentes cursos a distância de outras instituições.

Adicionalmente, considera-se relevante realizar estudos para avaliar os resultados da aplicação de técnicas conhecidas como Classificação Baseada em Custo (*"cost-sensitive classification"*) para diferenciar o custo dos erros de classificação para cada classe. A investigação de cenários com diferentes números de classes poderá ainda ser tratada em futuras pesquisas, com o objetivo de avaliar os impactos de mudanças neste parâmetro no processo de classificação dos estudantes.

A partir das estimativas de desempenho acadêmico de estudantes apresentadas, novos esforços poderiam ser feitos na tentativa de dotar os AVA de recursos inteligentes, tais como monitorar e adaptar conteúdo ou monitorar estudantes na expectativa de identificar estudantes com risco de desempenho insatisfatório.

O conjunto de atributos utilizados nesta pesquisa envolve apenas indicadores quantitativos e baseia-se apenas em dados possíveis de serem obtidos diretamente da base de dados de um AVA. Essa limitação poderá ser tratada em futuros trabalhos que avaliem a ampliação do conjunto de atributos propostos, considerando principalmente a inclusão de características que representem aspectos qualitativos dos estudantes. Como exemplo, poder-se-ia utilizar técnicas de mineração de texto para obter indicadores que verifiquem se as postagens de um estudante em um fórum de discussão são relacionadas ou não com tema desse fórum. Nessa linha, poderia ainda ser investigada a contribuição da inclusão de atributos obtidos de fontes externas aos AVA, tais como indicadores socioeconômicos, resultados de obtidos em outros cursos ou níveis de ensino anterior.

Outra possibilidade de investigação futura refere-se a verificar a relação entre medidas possíveis de serem obtidas por meio de técnicas conhecidas como Análise

de Redes Sociais, tais como *closeness centrality, betweenness centrality, clustering* (Wassermann e Faust, 1994) e o desempenho apresentado pelos estudantes em um curso a distância. A utilização dessa técnica poderá contribuir para o aprofundar a análise da importância das interações ocorridas entre os estudantes em um curso a distância.

Investigar detalhes sobre os casos de estudantes em que o processo de classificação falhou completamente, ou seja, um estudante da classe “A” que foi classificado erroneamente como “C” e vice-versa. Essa investigação poderá apontar a necessidade de inclusão de novos atributos à lista apresentada neste trabalho para aprimorar a representação de estudantes em um AVA e melhorar o processo de classificação.

REFERÊNCIAS

ANIL, J.K.; MAO, J.; MOHIUNDDIN, K.M. Artificial Neural Networks: A Tutorial. **IEEE Computer Society**, vol. 29, n. 3, p. 31-44, 1996.

AZEVEDO, B.F.T.; REATEGUI, E.B; BEHAR, P.A. Qualitative Analysis of Discussion Forums. **In IADIS International Conference on e-Learning, Freiburg, Alemanha,** p. 251-258, 2010.

AZEVEDO, B.F.T.; BEHAR, P.A.; REATEGUI, E.B. Análise das mensagens de fóruns de discussão através de um software para mineração de textos. **Anais do XXII SBIE-XVII WIE**, p. 20-29, 2011.

BAKER, R.S.J.D., Data Mining for Education. **International Encyclopedia of Education** (3rd edition), p. 112-118. Elsevier: Oxford, UK, 2010.

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.; 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.

BAKER, R.S.J.D.; ISOTANI, S.; CARVALHO, A.M.J.B.D. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, Vol. 19, No. 2. p. 2-13, 2011.

BITTENCOURT, I.I.; COSTA, E. Modelos e Ferramentas para a Construção de Sistemas Educacionais Adaptativos e Semânticos. **Revista Brasileira de Informática na Educação**, Vol. 19, No. 1. p. 85-98, 2011.

BORBA, M.C.D; MALHEIROS, A.P.S.D; ZULATTO, R.B.A. **Educação a Distância *online*.** 2.ed. Belo Horizonte: Autêntica, 2008.

BROOKS, C. A.; GREER, J. E.; MELIS, E.; ULLRICH, C. Combining ITS and eLearning Technologies: Opportunities and Challenges. In Ikeda, M., Ashley, K. D., and Chan, T.-W., Intelligent Tutoring Systems, Vol. 4053, **Lecture Notes in Computer Science,** p. 278–287. Springer, 2006.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. **Journal of Artificial Intelligence**, No. 16, 321–357, 2002.

COBO, G.; GARCÍA-SOLÓRZANO, D.; SANTAMARÍA, E.; MORÁN, J.A.; MELENCHÓN, J.; MONZO, C. Modeling Students Activity in Online Discussion Forums: a Strategy Based on Time Series and Agglomerative Hierarchical Clustering. **In Proceedings of the Fourth International Conference on Educational Data Mining**, p. 253-257, 2011.

D'MELLO, S.K.; CRAIG, S.D.; WITHERSPOON, A.W.; MCDANIEL, B.T.; GRAESSER, A.C. Automatic Detection of Learner's Affect from Conversational Cues. **User Modeling and User-Adapted Interaction**, p.45-80, 2008.

DRINGUS, L.P.; ELLIS, T. Using Data Mining as a Strategy for Assessing Asynchronous Discussion Forums. **Computers & Education, n. 45**, p.141-160, 2005.

ELEUTÉRIO, M.A.; BORTOLOZZI, F. AMANDA: An ITS for Mediating Asynchronous Group Discussions. **Lecture Notes in Computer Science - Intelligent Tutoring Systems** , vol. 3220, p. 102-115, 2004.

FAYYAD, U.M.; IRANI, K. B. Multi-interval discretization of continuous valued attributes for classification learning. In: **Thirteenth International Joint Conference on Articial Intelligence**, p.1022-1027, 1993.

FAYYAD, U.M.; PIATESKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. Advances in Knowledge Discovery and Data Mining, **AAAI Press**, 1996.

FAYYAD, U.M.; PIATESKY-SHAPIRO, G.; SMYTH, P. The KDD Process for Extracting Useful Knowledge from Volumes of Data. **Comunications of ACM**, vol.39, no.11, p.27-34, 1996.

GONG, Y.; RAI, D.; BECK, J.; HEFFERNAN, N. Does Self-Discipline Impact Students' Knowledge and Learning? **In Proceedings of the 2nd International Conference on Educational Data mining**, p.61-70, 2009.

GOTTARDO, E.; KAESTNER, C.; 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.; NORONHA, R.V. Social Networks Applied to Distance Education Courses: Analysis of Interaction in Discussion Forums. **In Proceedings of 18th Brazilian Symposium on Multimedia and the Web**, 2012.

GU, Q.; CAI, Z.; ZHU, L.; HUANG, B. Data Mining on Imbalanced Data Sets. **In Proceedings of International Conference on Advanced Computer Theory and Enginnering**, p.1020-1024, 2008.

HALL, M. A. **Correlation-based Feature Subset Selection for Machine Learning**. Hamilton, New Zealand, 1998.

HÄMÄLÄINEN, W.; VINNI, M. Classifiers for Educational Data Mining. In: Romero et al. **Handbook of Educational Data Mining**. Flórida, CRC Press, p. 57-71, 2011.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**.California, 2.ed. Morgan Kaufmann, 2006.

HOLLIMAN, R.; SCANLON, E. Investigating Cooperation and Collaboration in Near Synchronous Computer Mediated Conferences. **Computers & Education, n. 46**, p.322-335, 2006 .

HRASTINSKI, S. What is Online Learner Participation? A Literature Review. **Computers & Education, n. 51**, p.1755-1765, 2008.

HRASTINSKI, S. A Theory of Online Learning as Online Participation. **Computers & Education, n. 52**, p.78-82, 2009 .

IBRAHIM, Z.; RUSLI, D. Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression**, In Proceedings of the 21º Annual SAS Malaysia Forum**, Kuala Lumpur, Malaysia, p. 1–6, 2007.

KAMPFF, A.J.C. **Mineração de Dados Educacionais para a Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente**. 2009, 186p, Tese (Doutorado em Informática na Educação) – Programa de Pós-Graduação em Informática na Educação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009.

KOTSIANTIS, S.B. Use of Machine Learning Techniques for Educational Purposes: A Decision Support System for Forecasting Student's Grades". **Springer Science+Business Media B.V.** DOI 10.1007/s10462-011-9234-x, 2011.

LI, Y.; HUANG, R. Analyzing Peer Interaction in Computer-Supported Collaborative Learning: Model, Method and Tool. **Lecture Notes in Computer Science (LNCS)**. n. 5169, p.125-136, 2008.

MACFADYEN, L.P.; DAWSON, S. Mining LMS Data to Develop an “Early Warning System” for Educators: A Proof of Concept. **Computers & Education, n. 54**, p.588-599, 2010.

MAIA, R.F.; SPINA, E.M.; SHIMIZU, S.S. Sistema de Previsão de Desempenho de Alunos para Auxilio a Aprendizagem e Avaliação de Disciplinas. **Anais do XXI SBIE-XVI WIE**, 2010.

MANHÃES, L.M.B.; CRUZ, S.M.S.; COSTA, R.J.M.; ZAVALETA, J.; ZIMBRÃO, G. Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados. **Anais do XXII SBIE-XVII WIE**, p. 150-159 , 2011.

MÁRQUEZ-VERA, C.; ROMERO, C.; VENTURA, S. Predicting School Failure Using Data Mining. **In Proceedings of 4ª International Conference on Educational Data Mining**, p. 271-275, 2011.

MCQUIGGAN, S.; MOTT, B.; LESTER, J. Modeling Self-Efficacy in Intelligent Tutoring Systems: An Inductive Approach**. User Modeling and User-Adapted Interaction,** v.18, p. 81-123, 2008.

MERCERON, A.; YACEF, K. Educational data mining: a case study**. Artificial Intelligence in Education. Proceedings of the 12th International Conference on Artificial Intelligence**, p.467-474, 2005.

MINAEI-BIDGOLI, B.; KASHY, A.D.; KORTEMEYER, G.; PUNCH, F.W. Predicting Student Performance: An application of data mining methods with the educational web-based system**. Proceedings of International Conference in Frontiers of Education**, p. 13-18, 2003.

MOODLE, disponível em <http://moodle.org>. Acesso em 24/03/2012.

MOORE, M. G. Three Types of Interaction. **The American Journal of Distance Education**, Vol. 3, No. 2, p. 1–6, 1989.

MOORE, M. G.; KEARSLEY, G. **Educação a distância: uma visão integrada**. São Paulo: Thomson Learning, 2007.

NISTOR, N.; NEUBAUER, K. From Participation to dropout: Quantitative Participation patterns in online University Courses. **Computers & Education, n. 55**, p.663-672, 2010.

PEREIRA, J.M**.** Educação Superior a Distância, Tecnologias de Informação e Comunicação e Inclusão Social no Brasil. **Revista de Economia Política de las Tecnologias de la Información y Comunicación**, vol. XII, n.2, Mayo-Agosto, 2010.

RABBANY, R.K.; TAKAFFOLI M.; ZAÏANE, O.R. Analyzing Participation of Students in Online Courses Using Social Network Analysis Techniques. **In Proceedings of the Fourth International Conference on Educational Data Mining**, p. 22-30, 2011.

RICARTE, I. L. M.; FALCI JUNIOR, G. R. A Methodology for Mining Data from Computer-Supported Learning Environments. **Informática na Educação: teoria & prática**, Porto Alegre, v. 14, n. 2, p. 83-94, 2011.

ROMERO, C.; VENTURA, S. Educational Data mining: A Survey from 1995 to 2005. **Expert Systems with Applications**, v.33, p.125-146, 2007.

ROMERO, C.; VENTURA, S.; GARCÍA, E. Data mining in course management systems: Moodle case study and tutorial. **Computers & Education**, n. 51, p.368-384, 2008a.

ROMERO, C.; VENTURA, S.; ESPEJO, G.P.; HERVÁS, C. Data mining Algorithms to Classify Students. **In Proceedings of the 1st International Conference on Educational Data mining**, p.8-17, 2008b.

ROMERO, C.; VENTURA, S. Educational Data Mining: A Review of the State of Art. **IEEE Transactions on Systems , Man, and Cybernetics – Part C: Application and Reviews**. Vol. 40, n.6., p. 601-618, 2010.

ROMERO-ZALDIVAR, V.A.; PARDO, A.; BURGOS, D.; KLOOS, C.D. Monitoring Student Progress Using Virtual Appliances: A Case Study. **Computers & Education, n. 58**, p.1058-1067, 2012.

SCHRIRE, S. Knowledge building in asynchronous discussion groups: going beyond quantitative analysis. **Computers & Education**, Vol. 46, n.1, p.49–70, 2006.

WASSERMANN, S.; FAUST, K. **Social Network Analysis: Methods and Application**. Cambridge University Press, Cambridge, 1994.

WITTEN, I.H.; FRANK E.; HALL, M.A. **Data mining: Practical machine learning tools and techniques**. San Francisco: Morgan Kaufmann, 3 ed., 2011.

ZORRILLA, M. E.; MENASALVAS, E.; MARIN, D.; MORA, E.; SEGOVIA, J. Web usage mining project for improving web-based learning sites**. In Web mining workshop**, Cataluña, p. 1-22, 2005.