UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

FRANCISCO DA CONCEIÇÃO SILVA

FERRAMENTA PARA VISUALIZAÇÃO DE DIAGNÓSTICO DE BAIXO DESEMPENHO GERADO A PARTIR DO MÉTODO DE CLASSIFICAÇÃO NO PROCESSO DE MINERAÇÃO DE DADOS, COM BASE NAS INTERAÇÕES EM FÓRUNS DE DISCUSSÃO

DISSERTAÇÃO DE MESTRADO

SÃO LUIS

2015

FRANCISCO DA CONCEIÇÃO SILVA

FERRAMENTA PARA VISUALIZAÇÃO DE DIAGNÓSTICO DE BAIXO DESEMPENHO GERADO A PARTIR DO MÉTODO DE CLASSIFICAÇÃO NO PROCESSO DE MINERAÇÃO DE DADOS, COM BASE NAS INTERAÇÕES EM FÓRUNS DE DISCUSSÃO

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia de Computação e Sistemas da Universidade Estadual do Maranhão, como parte dos requisitos para a obtenção do título de Mestre em Engenharia de Computação e Sistemas. Área de Concentração: Tecnologia da Informação. Linha de Pesquisa: Informática na Educação.

Orientador: Prof. Dr. Luis Carlos Costa Fonseca
Coorientador: Prof. MSc. Josenildo Costa da Silva
Coorientador: Prof. Ms. Reinaldo de Jesus da Silva

SÃO LUIS
2015

Silva, Francisco da Conceição.

Ferramenta para visualização de diagnóstico de baixo desempenho gerado a partir do método de classificação no processo de mineração de dados, com base nas interações em fóruns de discussão. /Francisco da Conceição Silva – São Luis, 2015.

82 f.

Dissertação (Mestrado) – Curso de Pós-graduação em Engenharia de Computação e Sistemas (PECS). Universidade Estadual do Maranhão, 2015.

Orientador: Prof. Dr. Luis Carlos Costa Fonseca

1. Baixo desempenho. 2. Mineração de dados. 3. Classificação. 4. Fórum. 5. AVA. I. Título

CDU: 37.091.31:004.891.3

UNIVERSIDADE ESTADUAL DO MARANHÃO

CENTRO DE CIÊNCIAS TECNOLÓGICAS

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO E SISTEMAS

FERRAMENTA PARA VISUALIZAÇÃO DE DIAGNÓSTICO DE BAIXO DESEMPENHO GERADO A PARTIR DO MÉTODO DE CLASSIFICAÇÃO NO PROCESSO DE MINERAÇÃO DE DADOS, COM BASE NAS INTERAÇÕES EM FÓRUNS DE DISCUSSÃO

FRANCISCO DA CONCEIÇÃO SILVA

Aprovada em: 23 / 10 / 2015.

BANCA EXAMINADORA:

Prof. Dr. Luis Carlos Costa Fonseca
(Presidente/Orientador)
Departamento de Engenharia da Computação
Universidade Estadual do Maranhão – UEMA

Prof. Dr. Angelo Rodrigo Bianchini
Departamento de Educação II – Campus Bacanga
Universidade Federal do Maranhão – UFMA

Prof. Ms. Reinaldo de Jesus da Silva
Departamento de Engenharia da Computação
Universidade Estadual do Maranhão – UEMA

Prof. MSc. Josenildo Costa da Silva
Departamento Acadêmico de Informática
Instituto Federal de Educação, Ciência e Tecnologia do Maranhão – IFMA

Dedico esta pesquisa

Aos meus pais,
Antonia Gomes da Conceição e
Raimundo Pereira da Silva.

À minha esposa Leydiane e
minha filha Mariana.

A meus irmãos, sobrinhos,
amigos, colegas de curso, professores
e funcionários da UEMA.

AGRADECIMENTOS

Agradeço a Deus, por ter me dado força nos momentos difíceis dessa jornada.

Ao meu orientador, professor Dr. Luis Carlos Costa Fonseca, pela orientação nesta pesquisa e pela amizade. Agradeço pelas oportunidades que me proporcionou para que eu pudesse ampliar meus conhecimentos.

Aos meus coorientadores, professor MSc. Josenildo Costa da Silva e professor Ms. Reinaldo de Jesus da Silva, pela amizade e momentos de aprendizado, que foram valiosos e muito contribuíram para a realização desta pesquisa.

Ao professor Dr. Angelo Rodrigo Bianchini, pela participação nesta banca examinadora e por suas honrosas contribuições a esta pesquisa.

Aos meus familiares pelo apoio em todos os momentos da minha vida.

À minha esposa Leydiane e minha filha Mariana, que com carinho, delicadeza e compreensão, me motivam a seguir em frente sempre.

Aos colegas de curso, pelos bons momentos passados juntos.

E a todos que, direta ou indiretamente, contribuíram para que eu chegasse até aqui.

O gênio é composto por 2% de talento e de 98% de perseverante aplicação.

Ludwig van Beethoven

RESUMO

As ferramentas colaborativas e de comunicação tem sido usadas largamente nos contextos educacionais e os Ambientes Virtuais de Aprendizagem (AVAs), que são uma modalidade de Ensino a Distância (EAD), estão cada vez mais sendo inseridos em universidades, escolas e empresas. Essa comunicação ocorre de diversas formas, tais como *chats*, fóruns de discussão, *wikis*, dentre outras. Os fóruns, especialmente, consistem em espaços para discussões e trocas de ideias sobre assuntos definidos por seus participantes, possibilitando uma experiência favorável ao processo de aprendizagem. Na EAD existe um problema recorrente e muito desafiador, que é a evasão de alunos, cujas taxas de desistência são altas e preocupantes. Neste sentido, esta pesquisa apresenta o desenvolvimento de um modelo preditivo de baixo desempenho em um AVA, a partir das interações de alunos em fóruns de discussão. O objetivo foi realizar a previsão de baixo desempenho de alunos, que é considerado um forte indício para evasão, gerando relatórios que auxilie as partes interessadas na tomada de decisão. Para isso, foram realizados experimentos com conjuntos de dados distintos, onde a Mineração de Dados (MD) foi aplicada através de cinco algoritmos de classificação, sendo comparado o desempenho de cada um, a fim de que um modelo com melhor desempenho fosse obtido. Para a visualização dos resultados obtidos no processo de MD foi desenvolvida uma ferramenta com o objetivo de melhor apresentar os resultados obtidos às partes interessadas, sendo um auxílio na tomada de decisão.

Palavras-Chave: Baixo desempenho. Mineração de dados. Classificação. AVA. Fórum.

ABSTRACT

Abstract. The tools colaborative and of communication has been used broadly in the education contexts and the Virtual Learning Environments (VLEs), that are a modality of Distance Education (DE), they are more and more being inserted in universities, schools and companies. That communication happens in several ways, such as chats, discussion forums, wikis, among others. The forums, especially, consist of spaces for discussions and changes of ideas on defined subjects for their participants, making possible a favorable experience to the learning process. In EAD an recurrent and very challenging problem exists, that is the students' dropout, whose dropout rates are high and preoccupying. In this sense, this research presents the development of a model preditivo of low acting in an AVA, starting from the students' interactions in discussion forums. The objective was to accomplish the forecast of low acting of students, that considered a strong indicator evasion, generating reports that it aids the interested parts in the socket of decision. For that, experiments were accomplished with groups of different data, where the Data Mining (DM) was applied through five classification algorithms, being compared the acting of each one, so that a model with better acting was obtained. For the visualization of the results obtained in the process of DM a tool was developed with the best objective to present the results obtained to the interested parts, being an aid in the socket of decision.

Key words: Underperforming. Data mining. Classification. VLE. Forum.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ARFF: Atribute-Relation File Format (Formato de Arquivo Atributo-Relação)

APM: Aprovado por Média

AVA: Ambiente Virtual de Aprendizagem

BD: Banco de Dados

CSV: Comma Separeted Value (Valores Separados por Vírgula)

EAD: Educação a Distância

GPL: General Public License (Licença Pública Geral)

IES: Instituições de Ensino Superior

KDD*: Knowledge Discovery Database (Descoberta de Conhecimento em Banco de Dados)*

MD: Mineração de Dados

MDE: Mineração de Dados Educacionais

MEC: Ministério da Educação e Cultura

MOODLE: Modular Object-Oriented Dynamic Learning Environment (Objeto Orientado para Ambiente Dinâmico de Aprendizagem)

SO: Sistemas Operacionais

TBD: Tendência a Baixo Desempenho

TP: Técnicas de Programação

USA: United States American (Estados Unidos da América)

VLEs: Virtual Learning Environments (Ambientes Virtuais de Aprendizagem)

WEKA: Waikato Environment for Knowledge Analyis (Ambiente de Análise de Conhecimento Waikato)

SUMÁRIO

1 INTRODUÇÃO

O desenvolvimento tecnológico, a partir da segunda metade do século XX, impulsionou e está transformando a maneira de ensinar e de aprender. Nos últimos anos os Ambientes Virtuais de Aprendizagem (AVAs) estão sendo cada vez mais utilizados no âmbito acadêmico e corporativo como uma opção tecnológica para atender esta demanda educacional (PEREIRA, 2007, p.4).

Ferramentas de comunicação e colaboração estão sendo largamente usadas em contextos escolares, e como resultado, AVAs estão sendo mais e mais utilizados para adicionar tecnologia *web* em seus cursos e para suplementar os tradicionais cursos face-a-face (COLE; FOSTER, 2007). Essa comunicação pode acontecer de diversas formas, como por exemplo, *chats*, fóruns de discussão, *wikis*, dentre outras. Em relação aos fóruns, eles consistem em espaços para discussões e trocas de ideias sobre assuntos definidos por seus participantes e podem permitir uma experiência de aprendizagem favorável ao processo pedagógico (ABAWAJY, 2012).

O desenvolvimento tradicional de cursos de aprendizagem *online* é uma atividade laboriosa em que o desenvolvedor (geralmente o professor) tem de escolher os conteúdos que serão apresentados, decidir a estrutura dos conteúdos e determinar os elementos dos conteúdos mais apropriados para cada tipo de potenciais usuários do curso. Dada a complexidade dessas decisões, um projeto curto é dificilmente possível, mesmo quando é feito cuidadosamente. Além disso, na maioria dos casos, é necessário evoluir e possibilitar modificações no conteúdo dos cursos, estrutura e navegação baseados nas informações de uso dos alunos, tudo isso para proporcionar aos alunos condições de permanência e avanço em um curso.

No AVA, o professor precisa ter uma postura de mediação, incentivando os alunos a participarem, acompanhando seus desempenhos, dando retorno e orientação durante toda a caminhada (MOORE; KEARSLEY, 2007; ROMERO-ZALDIVAR *et al.*, 2012). Apenas reagir à demanda dos alunos, apresentando respostas às suas indagações, empobrece o papel do professor (ALVES; NOVA, 2003); por isso, ele deve ter postura proativa e mediadora para gerir sua sala de aula virtual de forma a integrar seus alunos em uma comunidade virtual de aprendizagem, que estabeleça trocas significativas e onde cada um esteja engajado

nos estudos. Para que a mediação do professor possa acontecer, alguns fatores são essenciais para contribuir no desenvolvimento da autonomia dos alunos (PALLOF; PRATT, 2004):

- estrutura de curso bem planejada;
- ferramentas de diálogo;
- atendimentos individuais;
- técnicas de mediação:
    - humanização;
    - participação;
    - estilo da mensagem e;
    - *feedback*.

Ao aluno é necessário que mantenha regularidade no acesso, com atenção às orientações do professor mediador, realizando as atividades nos prazos estabelecidos. Para Moore e Kearsley (2007), o aluno deve participar ativamente a fim de construir seu conhecimento de forma mais elaborada, não dependendo do contato face-a-face com o professor. No entanto, conforme os autores, quando o aluno perde o prazo de entrega das atividades, torna-se difícil seguir e, normalmente, desiste do curso.

Nesse sentido, quando há um ambiente que proporcione as condições adequadas para o processo de ensino e aprendizagem (TINTO, 2000), onde há a ação proativa do professor mediador e a participação ativa do aluno, estas relações podem possibilitar maior dinâmica nesse processo, com resultados bem promissores.

Uma característica comum em ambientes computacionais utilizados no contexto educacional refere-se a sua capacidade de coletar e armazenar uma grande quantidade de dados sobre os usuários (ROMERO et al., 2008). Esses dados são bastante amplos, variando desde registros de acesso, interações diversas com o sistema até dados com ricos significados semânticos tais como as mensagens em fóruns e *chats* (ABAWAJY 2012; ROMERO-ZALDIVAR et al., 2012).

É importante ressaltar que a simples criação de vastas bases de dados torna-se pouco útil sem a disponibilização de ferramentas adequadas para sua análise e interpretação de forma automática. Algumas plataformas de apoio ao ensino disponibilizam ferramentas simples de relatório que permitem a extração de

algumas informações sobre atividades desenvolvidas por alunos. Entretanto, é difícil para professores realizarem uma análise de alto nível de modelos e padrões dessas informações, visto que é difícil fazer isso manualmente em grandes bases de dados. O professor poderia estar motivado a fazer uma análise dessas, por exemplo, pelos seguintes questionamentos: qual o padrão para identificar um aluno com baixo desempenho? Como prever alunos desmotivados ou prestes a abandonar o curso?

Na EAD existe um problema recorrente e muito desafiador, que é a evasão de alunos e, conforme Obbadi e Jurberg (2005), as altas taxas de desistência estão preocupando as instituições que oferecem esta modalidade de ensino. Gibson (1998) apresenta alguns fatores que explicam os motivos de baixo desempenho e, consequentemente, abandono do curso por parte dos alunos:

- fatores do aluno (atributos de motivação e persistência, autoconfiança acadêmica, etc);
- fatores situacionais (apoio da família e empregador, mudanças em circunstâncias da vida pessoal);
- fatores do sistema educacional (qualidade, dificuldades com o material instrucional, suporte institucional);

Para Tinto (2000), os alunos têm mais chances de aprender e persistir quando se encontram em ambientes que:

- possuem altas expectativas para a sua aprendizagem e apresentam isso de forma clara e consistente;
- fornecem apoio acadêmico e social para suas necessidades, essenciais para a promoção da retenção e da aprendizagem;
- provêm *feedback* frequente sobre a sua aprendizagem; também diz respeito a sistemas de alertas precoces que avisem as instituições sobre alunos que precisam de assistência que faça diferença;
- oportunizam o envolvimento com outros alunos e professores em aprendizagens significativas, em comunidades de aprendizagem que favoreçam o sentimento de pertença e engajamento.

Os dados gerados nas interações entre professores e alunos, dos alunos entre si e deles com os recursos disponibilizados em um AVA são volumosos, pouco explorados e reuni-los é uma atividade complexa e exaustiva. Dentre esses dados, os fóruns de discussão fornecem informações sobre as interações nas discussões

propostas em um curso ou disciplina, onde eles podem se expressar a respeito dos temas tratados, bem como responder a postagens de outros alunos, gerando, assim, uma comunicação mais dinâmica (ABAWAJY, 2012).

Uma forma de fazer a identificação de relações relevantes em grandes bases de dados, como a de um AVA, é por meio da Mineração de Dados (MD), que busca explorar e analisar esses dados para identificar regras, padrões ou desvios. A MD é um processo para extração de conhecimento que estão implícitos, são previamente desconhecidos e são muito úteis para um contexto de estudo.

Tendo como base o contexto educacional, existe um campo de pesquisa em MD chamado Mineração de Dados Educacionais (MDE), que consiste de métodos e ferramentas de análise de dados para observar o comportamento dos alunos para auxiliar o professor a detectar possíveis erros, deficiências e melhorias. É uma abordagem indutiva muito interessante que cria modelos para descobrir automaticamente relações ocultas presentes nos dados dos alunos que podem ser utilizadas na melhoria da aprendizagem (ROMERO et al., 2008).

A utilização de MDE pode viabilizar melhores condições para que o professor tenha êxito na mediação pedagógica a seus alunos e pode ser aplicada a dados gerados em ambientes virtuais para encontrar correlações entre os dados disponíveis, que, conforme Garcia *et. al*, (2007), visam:

- otimizar os conteúdos em um portal educacional por meio da descoberta dos conteúdos que mais interessam aos usuários;
- extrair padrões úteis para ajudar educadores e desenvolvedores de materiais a avaliar e interpretar as atividades de um curso *on-line*, as formas como são executadas e seus resultados;
- guiar automaticamente as atividades dos alunos, gerando e recomendando materiais;
- personalizar o ensino virtual com base na agregação de perfis de usuários e ontologias de domínio;
- etc;

O grande volume de dados possibilitou, em paralelo ao avanço das técnicas de MD, um estudo mais acurado do fenômeno da evasão. O uso de técnicas de MDE possibilita identificar padrões de acesso que apontam se a realização de

atividades e interações dos alunos os levam a obter êxito (ou não) e pode contribuir para reduzir os índices de evasão e reprovação.

Quando se utiliza um AVA, não basta disponibilizar conteúdos e abrir espaços de discussão, é preciso acompanhar o percurso do aluno, mediando o processo de aprendizagem. A realização de MD em AVAs tem por finalidade a melhoria do processo de aprendizagem, a fim de que se torne uma estratégia promissora para atender aos objetivos propostos para a aprendizagem dos alunos.

1.1 Motivação e Objetivos

A questão norteadora desta pesquisa reside na inquietação decorrente do seguinte cenário: os altos índices de evasão e reprovação são preocupantes para o avanço da EAD. De fato, para Romero et al. (2012), o fracasso escolar é conhecido como "o problema das mil causas", dado que uma série de fatores do tipo pessoal, acadêmico (quanto maior a escolaridade, menor a evasão), físico, familiar, social, dentre outros, podem influenciar no baixo desempenho do aluno, ocasionando fracasso ou abandono de curso.

Segundo Favero (2006, p.153):

> [...] foi estudada a evasão que ocorre em cursos na modalidade a distância. O estudo realizado permitiu verificar que esse problema é uma realidade e quase todas as instituições que oferecem cursos a distância, senão todas, enfrentam esse problema.

Para este autor, evasão é considerada como a desistência do curso pelo aluno, independentemente da quantidade de participações efetuadas. É um fenômeno comum na EAD e os motivos devem ser pesquisados, buscando verificar os cursos em que a evasão ocorre com maior frequência.

Diante desse cenário, quais fatores influenciam, direta ou indiretamente, na evasão e reprovação de alunos e como construir um modelo computacional que se proponha fazer esse diagnóstico? No intuito de responder a essas questões e possibilitar a identificação de indicadores de baixo desempenho, a presente pesquisa apresenta um modelo preditivo para diagnóstico de baixo desempenho de alunos em AVAs, a partir de suas interações em fóruns de discussão, para servir como ponto de partida aos interessados diretos (professores, tutores, gestores, etc.) na tomada de decisão. Busca-se gerar um modelo que correlacione certas variáveis dessas interações, no sentido de prever desistências ou reprovações.

Para chegar a este objetivo, os seguintes objetivos específicos foram propostos:

- Realizar um estudo bibliográfico para ampliar a compreensão sobre os fenômenos evasão e reprovação em AVAs;
- Investigar e definir as técnicas de MD a serem utilizadas na elaboração do modelo preditivo proposto;
- Conhecer a estrutura da base de dados do MOODLE para preparação dos dados a serem utilizados;
- Realizar experimentos para obtenção de um modelo preditivo de MD para fazer o diagnóstico de tendências de baixo desempenho, permitindo analisar o efeito das interações de alunos em fóruns de discussão no seu desempenho acadêmico.
- Implementar uma ferramenta para visualização de diagnóstico gerado por MD.

A ferramenta proposta para visualização de diagnóstico se configura como uma das motivações para o desenvolvimento desta pesquisa e visa facilitar a compreensão das tendências de baixo desempenho geradas por MD em AVAs, a partir das interações de alunos em fóruns de discussão, servindo como ponto de partida na tomada de decisão.

Esta ferramenta utiliza modelos gerados por algoritmos de árvore de decisão e os dados utilizados na MD para proporcionar às partes interessadas condições de verificar em quais tendências os alunos estão inseridos, bem como em quais regras eles aparecem.

Outra motivação para o desenvolvimento desta pesquisa, que é decorrente da implementação da ferramenta, diz respeito à possibilidade de tomadas de decisão a partir dos resultados gerados pela ferramenta, no sentido de favorecer ao aluno condições de permanência e êxito em um curso.

Por fim, a pesquisa foi motivada pela possibilidade de viabilizar a disponibilidade da ferramenta proposta como um módulo em um AVA, para que o próprio professor possa realizar todo o processo de descoberta de conhecimento referente aos dados de seus alunos, bem como poder visualizar e analisar os resultados obtidos nesse processo, para que possa intervir pedagogicamente, possibilitando melhorias na aprendizagem dos alunos.

1.2 Estrutura da pesquisa

Os capítulos a seguir farão uma abordagem sobre a temática desta pesquisa e, a começar desta introdução, onde nela podemos fazer uma leitura inicial, que proporciona uma visão geral de todo a pesquisa, os demais capítulos estão organizados como segue.

O **capítulo 2** é destinado à revisão de literatura, onde é feita uma abordagem sobre o fenômeno chamado evasão, a fim de que seja ampliada a compreensão desse fenômeno, para que seja possível proporcionar as condições para enfrentar esse problema na educação à distância. É abordada a MDE, as principais técnicas e ferramentas que podem ser utilizadas para esta finalidade. É feita uma descrição sobre o MOODLE, apresentando a estrutura de sua base de dados e outros detalhes desta ferramenta. É feita uma abordagem sobre as interações de alunos em fóruns de discussão. Por fim, nas considerações finais temos a definição das técnicas que foram utilizadas para a geração das regras de classificação nesta pesquisa.

No **capítulo 3** são apresentados os trabalhos relacionados ao tema desta pesquisa, a fim de consolidar as discussões realizadas aqui.

No **capítulo 4** é feita a apresentação dos procedimentos metodológicos para os experimentos de MD, onde é apresentada a arquitetura do modelo preditivo proposto nesta pesquisa e feita sua descrição. São descritas as etapas de seleção, pré-processamento e transformações de dados, bem como definidos os atributos para a tabela de sumarização, para que sejam aplicadas as técnicas de MD adotadas nesta pesquisa, a fim de que os objetivos propostos aqui sejam alcançados. São definidas as técnicas de MD e algumas métricas de confiabilidade que serão aplicadas ao modelo computacional gerado, para analisar se o diagnóstico obtido neste estudo é confiável e aplicável. Finalizando o capítulo, é apresentada a arquitetura da ferramenta proposta para a visualização de diagnóstico gerado por MD.

O **capítulo 5** *é destinado à análise dos resultados obtidos nos experimentos* realizados, onde são analisadas algumas taxas de desempenho, definida a técnica de classificação mais adequada em cada contexto de dados e apresentados indicadores das tendências trabalhadas na pesquisa; é apresentada ainda a ferramenta proposta e algumas comparações entre indicadores obtidos.

No **capítulo 6**, são descritas as considerações finais, que ampliam as discussões referentes a esta pesquisa, a respeito de sua relevância no contexto em que é aplicado, bem como faz a abertura para trabalhos futuros.

Por fim, as referências que nortearam o desenvolvimento desta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma abordagem a respeito dos tópicos referentes a esta pesquisa, mostrando uma visão geral sobre pesquisas em cada assunto tratado.

2.1 Evasão na EAD

A EAD tem sido muito difundida no Brasil, sendo que a cada ano há uma ampliação nas ofertas de cursos e de matrícula de alunos nesta modalidade. Em contrapartida, um problema que se mostra recorrente e desafia a EAD diz respeito à evasão de alunos. As altas taxas de desistência estão deixando perplexas as instituições que oferecem esta modalidade de ensino (OBBADI; JURBERG, 2005). Conforme essas autoras:

> embora os números e índices de evasão em cursos a distância sejam bastante díspares e não coincidentes, todos concordam que os números de desistência dos alunos que frequentam cursos nesta modalidade são maiores do que os índices relatados pelas instituições tradicionais em cursos presenciais (OBBADI; JURBERG, 2005, p.49).

As autoras conduziram um estudo sobre a problemática da evasão na EAD e apresentaram alguns índices que evidenciam o desafio da evasão nos cursos nesta modalidade, conforme visto na **tabela 1**.

Tabela 1: Índices que evidenciam o desafio da evasão (Adaptada de OBBADI; JURBERG, 2005).

	Europa	USA e países asiáticos	Brasil
Índice (%)	de 20 a 30	50	62 (IES[1])
			21 (certificados pelo MEC[2])

A literatura internacional e nacional tem gerado muitas pesquisas sobre a evasão, sendo que muitas delas tratam desse fenômeno relacionado ao ensino superior (ABBAD et al., 2010).

A **tabela 2** mostra a distribuição dos índices de evasão no Brasil em um estudo mais recente, entre 2010 e 2013, segundo o censo da EAD (ABED, 2013), onde no ano de 2013 foi registrado um aumento significativo nos índices de evasão

[1] IES – Instituições de Ensino Superior
[2] MEC – Ministério da Educação e Cultura

em todos os tipos de cursos apresentados, sendo que a média simples de cursos regulamentados totalmente a distância foi de 19% e de semipresenciais foi de 14,6%.

Tabela 2: Índices de evasão registrados no período 2010-2013 no Censo EAD-BR (ABED, 2013, p.32).

Tipo de cursos	2010	2011	2012	2013
Autorizados pelo MEC	18,6%	20,5%	11,74%	16,94%
Livres não corporativos	22,3%	23,6%	10,05%	17,08%
Livres corporativos	7,6%	20%	3%	14,62%
Disciplinas EAD	—	17,6%	3,10%	10,49%

Frente a esse cenário preocupante, percebe-se que esses números de evasão na EAD são alarmantes e exigem um esforço efetivo no sentido de entender e explicar as suas causas de forma a propiciar ações corretivas e preventivas em relação à evasão.

As supostas causas de evasão na EAD são apontadas por Bruno *et. al* (2010) como:

> O insuficiente domínio técnico do uso do computador (principalmente da internet), falta da tradicional relação face a face entre professores e acadêmicos, dificuldade de expor ideias numa comunicação escrita a distância e a falta de um agrupamento de pessoas numa instituição física (BRUNO *et. al.*, 2010, p.10).

Outras possíveis causas de evasão são mencionadas ainda por Abbad *et. al.* (2010), conforme segue: falta de tempo, situação financeira, falta de adaptação ao sistema de cursos a distância, falta de dedicação, frustração das expectativas, ausência de integração entre colegas, falta de habilidade para administrar o tempo de estudo, dentre outras.

No intuito de melhor compreender o fenômeno da evasão no ensino, Tinto (2000) propôs um modelo teórico para o estudo da permanência e da evasão em cursos de graduação, onde supõe que a permanência é função do compromisso do aluno para concluir o curso, do comportamento do aluno com obrigações externas

ao ambiente acadêmico, da formação escolar anterior, da inteligência acadêmica (intelectual) e da integração social do aluno (pessoal). Esse autor ressalta que diferenças pessoais (demográficas) são menos importantes na determinação da evasão.

Nesse sentido, Abbad et al (2010) conduz a discussão de Tinto (2000) para o contexto da EAD, ao admitir que:

> a persistência em cursos a distância está associada a: conhecimento prévio sobre conteúdos semelhantes aos abordados pelo curso; motivação pessoal; necessidade e capacidade de balancear família e carreira; independência; autodisciplina (menor evasão); nível de escolaridade (quanto maior a escolaridade, menor a evasão) (ABBAD et al., 2010, p.295).

Ainda conforme as autoras, a influência de idade e sexo sobre evasão em curso a distância são inconsistentes e não conclusivas, de modo que tal influência não é adequada para o diagnóstico de evasão.

Conforme dados disponibilizados pelo MEC (BRASIL, 2005), o baixo desempenho, associado à vulnerabilidade econômica e social dos alunos, interfere nos indicadores de aprovação e conclusão dos níveis de ensino, bem como, nas taxas de abandono. Esse baixo desempenho ocorre, essencialmente, em semestres iniciais de um curso (BRUNO, 2011). Isso implica dizer que as disciplinas oferecidas nestes semestres iniciais são um fator que podem definir a situação do aluno, no sentido da continuidade ou não no curso. Nesta pesquisa foram utilizados dados oriundos de fóruns de discussão de disciplinas do segundo módulo, que corresponde à parte inicial do curso, para identificar indicadores de baixo desempenho dos alunos.

A identificação de indicadores de baixo desempenho em cursos a distância é importante para que seja possível proporcionar as condições necessárias que reduza ou elimine a evasão. Para isso, são necessários métodos e ferramentas de análise de dados a fim de observar o comportamento dos alunos para auxiliar as partes interessadas na tomada de decisão. Nesse sentido, a MDE é uma abordagem muito interessante que pode ser usada para descobrir automaticamente relações ocultas presentes nas informações sobre os alunos, em especial, as interações realizadas em fórum de discussão, que podem ser utilizadas na melhoria da aprendizagem (ROMERO et al., 2008).

2.2 Mineração de Dados Educacionais

O enorme crescimento no uso de AVAs gerou a necessidade de novas abordagens para analisar comportamentos de aprendizagem dos alunos que sejam adequados para esses ambientes. Tais abordagens têm permitido a captura automática e registro de informações sobre interações de alunos, fornecendo uma rica fonte de dados sobre comportamentos de aprendizagem (ROMERO et al., 2011). Dentre essas abordagens, destaca-se aqui a MDE, cujo objetivo é usar conjuntos de dados educacionais em larga escala para entender melhor a aprendizagem e fornecer informações sobre os processos de aprendizagem (ROMERO et al., 2011).

O desenvolvimento de métodos de MD para gerenciar e interpretar dados de interações de alunos em um ambiente *web* visa capturar e modelar os padrões de comportamento e perfis de usuários interagindo nesse ambiente (MOBASHER, 2004; ROMERO et al., 2011).

A capacidade humana para analisar grandes conjuntos de dados e encontrar relações significativas entre eles é muito limitada, sendo assim, difícil se apropriar de tais relações que não são previamente conhecidas, embora sejam potencialmente úteis. Desta forma, a MD pode ser um auxílio importante para encontrar informações que valem "ouro" (DEOGUN et al., 1997).

A MD utiliza técnicas eficientes para a descoberta de conhecimento (*knowledge discovery database* - KDD) (LIN; CERCONE, 1997) e é uma área que pode contribuir na descoberta automática de conhecimento que é potencialmente útil, por meio de algoritmos de aprendizado de máquina. As ferramentas de MD processam esses dados de forma a buscar correlações importantes entre eles e tem sido objeto de estudos interdisciplinares, tendo se desenvolvido muito nos últimos anos. A aplicação de MD abrange um número grande de áreas, como Inteligência Artificial, Estatísticas, Banco de Dados (FAYYAD et al., 1996; WITTEN et al., 2011) e, mais recentemente, a área educacional (ROMERO et al. 2008).

No campo educacional, minerar grandes bases de dados pode ajudar a localizar quais características e/ou comportamentos que contribuem para o êxito (ou não) na realização de um curso via web. Em um ambiente baseado na web é possível utilizar MD para analisar as interações em fóruns de discussão para predizer se um aluno tem tendência a baixo desempenho acadêmico, que pode levá-

lo a evadir ou ser reprovado em um curso; predizer fatores que podem afetar as reações de alunos, dada uma estratégia pedagógica específica, dentre outras (ROMERO et al., 2013).

2.2.1 Etapas do processo de descoberta de conhecimento

No processo de KDD é importante a realização de etapas anteriores à MD, a fim de que, quando nesta fase, a MD possa realizar a descoberta de padrões de comportamento potencialmente úteis. Para isso, Fayyad et al. (1996) propôs as etapas descritas na **figura 1**.

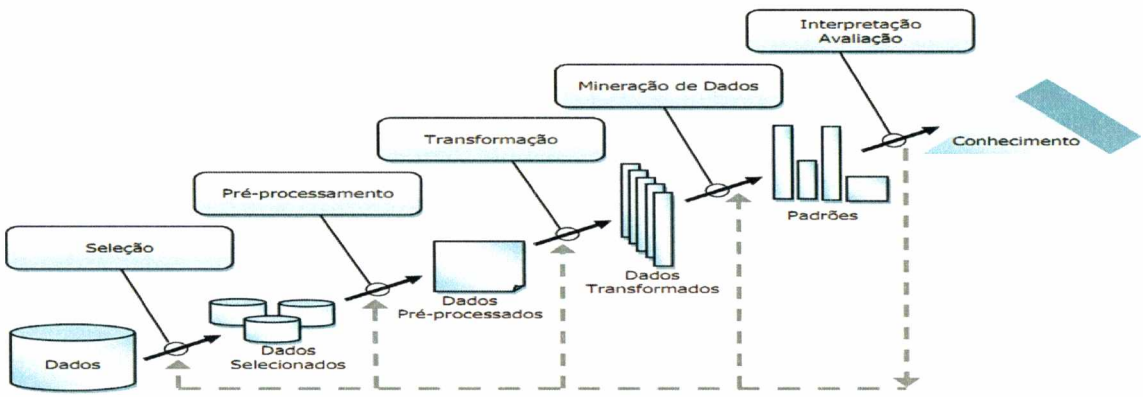


Figura 1: Etapas para Descoberta de Conhecimento (adaptada de Fayyad et al., 1996).

As etapas do processo de descoberta do conhecimento anteriores à MD demandam um grande esforço referente à obtenção e ajustes dos dados e, quando bem realizado, proporcionam um bom desempenho da etapa posterior, a MD (que representa 20% do esforço para a descoberta de conhecimento), pois é nela que relações implícitas e potencialmente úteis são extraídas (ZHANG et al., 2003).

A **seleção** consiste em recuperar os dados de uma fonte de dados para que os mesmos sejam submetidos às etapas posteriores no processo de KDD. Esses dados normalmente não estão em uma forma que possam ser diretamente analisados usando estatística ou técnicas de MD. Desta forma, uma etapa importante anterior à análise envolve o **pré-processamento**, que consiste na preparação adequada dos dados coletados. Esta etapa envolve um múltiplo estágio e processos largamente contextualizados (ROMERO et al., 2011). O pré-processamento deve excluir dados desnecessários sem empobrecer os dados e pode envolver a remoção de registros irrelevantes, substituição de registros ausentes e eliminação de periféricos (ROMERO et al., 2011).

Variáveis de dados individuais podem precisar ser **transformadas** para permitir análises mais significativas. Isto pode envolver, por exemplo, mapeamentos de dados para valores numéricos para permitir comparações (ROMERO et al., 2011); pode acontecer ainda de uma variável com muitas classes ser transformada para apresentar apenas algumas dessas classes, conforme critérios estabelecidos. Esta última situação está relacionada com a seleção de características, que é o processo de escolha de variáveis de dados para serem usadas na análise e eliminação de variáveis irrelevantes ou redundantes do conjunto de dados (ROMERO et al., 2011).

Para HAN et al. (2012), o processo de MD é realizado através de grupos de tarefas, tais como descoberta direta de conhecimento e descoberta indireta de conhecimento.

- *descoberta direta de conhecimento (aprendizado supervisionado)* – neste grupo, a supervisão no aprendizado vem dos exemplos marcados no conjunto de dados de treinamento; consiste em associar, a partir de classes determinadas, cada registro de dados a uma delas; lida com valores numéricos e contínuos, estimando um valor; no contexto educacional, poderia se estimar o resultado final do aluno em uma disciplina a partir de suas interações em fóruns de discussão, por exemplo;
- *descoberta indireta do conhecimento (aprendizado não-supervisionado)*: busca encontrar relações a partir do cruzamento entre todos os dados disponíveis, no intuito de localizar padrões, relações novas e úteis; por meio da *clusterização* (agrupamento), busca-se distribuir os variados dados analisados (heterogêneos) em grupos com características similares (mais homogêneas); a extração de regras de associação também se insere neste contexto não supervisionado.

A seguir, são apresentadas algumas das metodologias utilizadas tanto para aprendizado supervisionado quanto não supervisionado. É feita uma abordagem das tarefas, métodos e ferramentas existentes, para a definição do aparato mais apropriado para um determinado contexto de estudo. Ressalta-se que para esta pesquisa foi considerada a descoberta direta de conhecimento, com ênfase na classificação de alunos, a partir de suas interações em fóruns de discussão.

2.2.2 Classificação

A classificação é uma tarefa de aprendizado supervisionado, cujo método de análise de dados visa extrair modelos que descrevem classes de dados importantes. Esses modelos são chamados de classificadores e preveem rótulos categóricos de classe (HAN et al., 2012). Visa predizer um comportamento futuro, baseado em vários valores de um conjunto de dados.

Em um ambiente baseado na web é possível utilizar MD, através de classificação, para predizer, por exemplo, as tendências de evasão e reprovação de alunos em cursos por meio de árvores de decisão, predizer quais fatores afetam as reações de alunos, favoráveis ou não, a uma estratégia pedagógica específica, dentre outras. (ROMERO et al., 2013).

A seguir, são apresentados alguns métodos de classificação que podem ser utilizados para a descoberta automática de conhecimento.

2.2.2.1 Árvore de Decisão

É um modelo de representação de conhecimento que se adapta bem a tarefas de classificação, com boa visualização das características que impactam em cada classe. Sua representação pode ser gráfica ou textual, podendo ser traduzida em regras do tipo: SE <condição> ENTÃO <classificação>. Cada caminho da raiz até a folha corresponde a uma regra da forma T_{i1} ^ ... ^ T_{ij} -> $(C = c)$, onde c é o valor da classe na folha e cada T_{i1} é um teste booleano com valor no atributo de A_{ij} (ROMERO et al., 2011). Novos nós são colocados na árvore conforme sejam mais relacionadas com a raiz, de forma encadeada e campos não determinantes são desprezados na construção da árvore.

As técnicas baseadas em árvore de decisão são bastante adequadas no contexto educacional, pois elas geram um resultado que é mais compreensível e fácil de interpretar, sendo um grande auxílio na tomada de decisão. De fato, analisar os resultados a partir de uma árvore de decisão facilita em muito a tomada de decisão, ainda mais com a extração de regras de classificação possibilitada pela árvore.

Um algoritmo de classificação busca encontrar relacionamentos entre os atributos e uma ou mais classes e no caso de árvore de decisão, por exemplo, da árvore gerada pode-se extrair regras de classificação. Estas regras podem ser

utilizadas posteriormente para predizer a classe de um novo registro. A ordem de apresentação das regras estabelece uma lista de decisão, a ser aplicada em sequência. A primeira regra na lista tem maior prioridade para predizer a classe e quando um registro é classificado, nenhuma outra regra posterior será aplicada sobre ele.

Em uma regra gerada por algoritmo de árvore de decisão, o nó raiz sempre estará presente na regra. Um exemplo de uma regra de classificação, obtida a partir de uma árvore de decisão, é dado a seguir:

total_participacao_em_todos_os_foruns <= 1.849889 and
nota_media_em_foruns <= 4.75 and
total_posts_em_todos_os_foruns <= 0.120737: TBD (324.0/161.0)

Neste exemplo, o atributo **total_participacao_em_todos_os_foruns** é o atributo raiz da regra. Ao lado direito de cada regra há as informações referentes à classe da regra, como sua identificação (neste exemplo, a classe TBD) e dois parâmetros: o primeiro se refere ao número de casos que se enquadram na regra e o segundo indica o número de casos que não se enquadraram na regra.

O conjunto de regras geradas possibilita classificar corretamente todos os exemplos utilizados, o que em bases de dados volumosos nem sempre é possível (ROMERO et al., 2011).

Para a criação da árvore, é utilizada a métrica da entropia, que é uma medida definida na teoria da informação[3] e que busca definir a pureza de um conjunto de instâncias, necessitando de um conjunto de instâncias positivas e negativas. A entropia de um conjunto de dados S é definida como:

$$Entropy(S) = \sum_{i=1}^{v} -p_i \, log_2 p_i$$

Onde p_i é a proporção de S pertencente a i.

Outra medida utilizada é o ganho de informação (*infomation gain*), que é a redução esperada na entropia, causada pelo particionamento das instâncias de acordo com o atributo de predição, isto é, o quanto se espera que a entropia seja reduzida quando se sabe o valor do atributo A. Esta medida é definida como:

[3] Teoria da Informação é um ramo da matemática que estuda quantificação da informação, evolvendo dois aspectos cruciais: a eficiência da representação da informação gerada pela fonte e a taxa de transmissão à qual é possível enviar a informação com confiabilidade através de um canal ruidoso. Disponível em: http://users.isr.ist.utl.pt/~vab/FTELE/cap1.pdf.

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Onde, Values(A) é o subconjunto de todos os possíveis valores para o atributo A e S_v é o subconjunto de S em que o atributo A tem valor v.

Conforme Romero et al. (2011), para a construção de uma árvore de decisão, o algoritmo age recursivamente na fase de aprendizagem, realizando subdivisões dos dados até que as folhas sejam classes puras ou que exista um critério de parada especificado, como o número de casos enquadrado. Sua estrutura em árvore facilita a compreensão humana e, em um contexto educacional, essa característica se torna essencial para proporcionar a usuários como professores, gestores, dentre outros, melhores condições de analisarem os resultados obtidos com esse modelo.

A implementação mais conhecida de árvore de decisão diz respeito ao algoritmo ID3 (*Indutive Decision Tree*), sendo ele a base para implementações de outros algoritmos de árvore de decisão, como o C4.5 (QUINLAN, 1993) (J48 na ferramenta Weka). Algumas das principais vantagens do C4.5 em relação ao ID3 são apresentadas a seguir.

O C4.5 trabalha com variáveis discretas e contínuas (o ID3 só trabalha com dados discretos), através de um processo de discretização interno. Ele realiza a poda da árvore, que consiste em aumentar a capacidade de generalização da árvore de decisão (QUINLAN, 1993), evitando, assim, que ocorra o *overfit* (sobreajuste de um conjunto de dados). O *overfit* é influenciado principalmente pela pouca quantidade de instâncias nos dados e quando há ruído nos dados. Para Mitchell (1997), em um espaço de hipóteses **H** (árvores possíveis do C4.5, por exemplo), uma hipótese **h** ∈ **H** superestima (*overfit*) um conjunto de dados de treinamento (D) se existe alguma outra hipótese **h'** ∈ **H** tal que, **h** possui um erro menor que **h'** em D, mas **h'** possui um erro menor no conjunto de dados completo.

Outra vantagem do C4.5 é que, ao invés de usar o ganho de informação, ele usa a razão do ganho, definida como:

$$\text{GainRatio}(S,A) = \frac{\text{Gain}(S,A)}{\text{SplitInfo}(S,A)}$$

$$\text{SplitInfo}(S,A) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} x log_2\left(\frac{|D_j|}{|D|}\right)$$

Onde S é a entropia do conjunto de treinamento D e D_j é o um subconjunto de D em que o atributo A está presente.

A razão do ganho evita que atributos com muitos valores sejam beneficiados na seleção de atributos, o que possibilita a geração de árvores mais curtas e menos complexas *(*QUINLAN*, 1993)*.

Outras implementações de árvore de decisão são, por exemplo, os algoritmos CART (*SimpleCart* no Weka) e BFTree.

O algoritmo CART (*Classification and Regression Tree*) foi proposto por Breiman et al. (1984) e consiste em uma técnica não-paramétrica para induzir tanto árvores de classificação (se o atributo é categórico) quanto árvores de regressão (se o atributo é contínuo). Este algoritmo gera árvores sempre binárias, que são percorridas da sua raiz até as folhas através de respostas a perguntas do tipo "sim" ou "não", utilizando técnica de pesquisa exaustiva para definir os limiares utilizados nos nodos para dividir os atributos contínuos. A expansão da árvore é feita realizando pós-poda (diferentemente de outros algoritmos do tipo) por meio da redução do fator custo-complexidade. O processo de pós-poda tem o objetivo te tornar a árvore mais eficiente, pois a deixa mais simples, precisa e com ótima capacidade de generalização (BREIMAN et al., 1984).

O algoritmo BFTree foi proposto por Shi (2007) para indução de árvores de decisão binárias, tendo como base a heurística *best-first*, para construção do primeiro melhor classificador através de divisão binária para atributos numéricos e nominais. Para a criação da árvore de decisão, o algoritmo considera o atributo com maior ganho de informação.

2.2.2.2 Métodos Bayesianos

São métodos baseados em estatística que classificam uma instância em determinada classe a partir da probabilidade desta instância pertencer a esta classe (ROMERO et al., 2011). Eles podem prever probabilidades de associação de classe, como a probabilidade de que uma determinada tupla pertença a uma classe particular e tem como base o teorema de *Bayes*.

Dado um conjunto de instâncias de treinamento e um conhecimento *a priori*, o teorema de *Bayes* pode ser aplicado para definir a hipótese mais provável e é definido como:

$P(h|D) = (P(D|h)P(h)) / P(D)$

Onde:

- P(h) é a probabilidade da hipótese ser verdadeira (*priori* da hipótese);
- P(D) é a probabilidade do conjunto de dados D ser observado;
- P(D|h) é a probabilidade do conjunto de dados D ser observado dado que h é verdadeira;
- P(h|D) é a probabilidade de h ser verdadeira dado o conjunto de dados D (hipótese a *posteriori*)

Sabe-se que D é constante para todas as hipóteses, pois é o conjunto de treinamento, logo, P(D) pode ser retirada do cálculo, ficando:

$P(h|D) = P(D|h)P(h)$

Se a probabilidade de todas as hipóteses é a mesma (equiprovável), ela é retirada do cálculo, ficando:

$P(h|D) = P(D|h)$ => Verossimilhança

A verossimilhança é uma função da probabilidade condicional, onde a verossimilhança (L) de um conjunto de parâmetros (θ) dado alguma observação (x) é igual a probabilidade daquela observação ter ocorrido dados os valores daqueles parâmetros. Pode-se empregar esta função com a ideia de que o(s) valores(es) de θ que maximizam a probabilidade dos dados observados (x) seria um estimador de θ (CRAMER, 1986).

Dada uma amostra desconhecida X, com valores de seus atributos iguais a $x_1, x_2, ..., x_n$ e sabendo que existem m classes possíveis $C_1, C_2, ..., C_m$, calcula-se m probabilidades $P(C_i \mid X)$, i = 1, 2, ..., m. Cada um dos valores $P(C_i \mid X)$ representa a probabilidade de que a amostra X pertença a uma classe C_i específica, considerando que se conhece os valores dos atributos de X. $P(C_i|X)$ é chamada probabilidade de C_i condicionada a X ou probabilidade de que ocorra a classe Ci, dado que se conhece os valores dos atributos de X (ROMERO et al., 2011).

Dentre as implementações de métodos *bayesianos*, duas delas são amplamente utilizadas, e são os algoritmos *Naive Bayes* e *Bayes Net*.

O algoritmo *Naive Bayes* é uma técnica comparável em desempenho com classificadores que usam árvores de decisão e apresenta precisão alta e boa

escalabilidade (HAN et al., 2012). Ele é conhecido como classificador ingênuo porque o efeito do valor de um atributo sobre uma determinada classe é calculado de forma independente do restante dos outros atributos, ou seja, ele considera apenas probabilidades independentes, ou independência condicional de classe, tornando sensivelmente mais simples as comparações envolvidas.

As redes *bayesianas* (*BayesNet* no Weka) representam bem o conhecimento incerto através de dependências estatísticas, que são apresentadas visualmente como uma estrutura de grafo (ROMERO et al., 2011). Uma rede *bayesiana* é uma forma de representar o conhecimento de um domínio onde não se tem certeza de todas as variáveis presentes. Através da probabilidade pode-se responder, com níveis de certeza, a questões formuladas com base em evidências de uma situação.

2.2.2.3 Redes Neurais Artificiais

Estas técnicas remontam a 1943 com os trabalhos de McCulloch e Pitts, mas somente em 1953, Nathaniel Rochester simulou a primeira rede neural. É uma técnica baseada no funcionamento do cérebro humano, no tocante à forma como os neurônios reagem e propagem estímulos na rede neural (HAYKIN, 1999). Há três tipos básicos de redes neurais: *perceptron*, função de base radial e mapas auto organizáveis.

Em especial, a rede *perceptron* apresenta aprendizagem por retropropagação (*backpropagation*). A representação mais simples de um neurônio é dada conforme a **figura 2**.

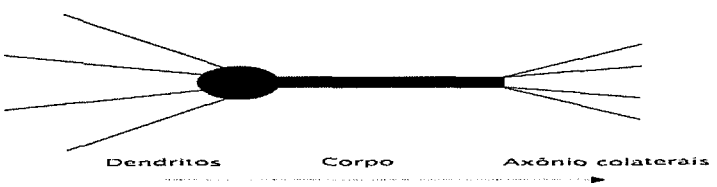


Figura 2. Modelo de um neurônio.

Em geral, em uma rede neural pode-se ter I nós de entrada, H nós intermediários e O nós de saída, conforme ilustra a **figura 3**:

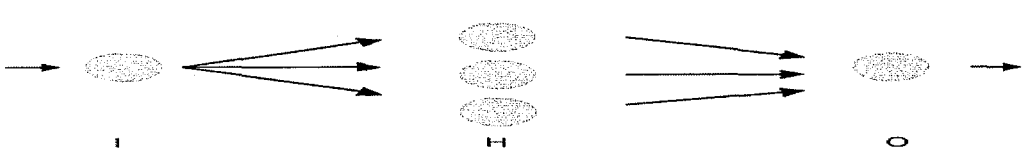


Figura 3. Rede Neural.

Na fase de treinamento, isto é, para um conjunto conhecido de valores de entrada e saída, ajustam-se os pesos de forma que o erro seja aceitável.

Seja $k\delta$ o erro do k-ésimo elemento de saída, o erro total deve então ser reduzido a cada iteração.

$$E_p = \frac{1}{2}\sum_{k=1}^{m} \delta_k^2$$

Isto é conseguido atualizando-se os pesos tanto da camada intermediária como da camada de saída (HAYKIN, 1999).

Abaixo segue o esquema de aprendizado por retropropagação de erros:

1. Aplicar o vetor X em todos os elementos da camada de entrada
2. Calcular os valores de propagação da camada intermediária
3. Calcular os valores de saída da camada intermediária
4. Calcular os valores de propagação da camada de saída
5. Calcular as respectivas saídas
6. Calcular os erros para cada elemento de saída
7. Calcular os erros para cada elemento intermediário
8. Atualizar os pesos da camada de saída
9. Atualizar os pesos da camada intermediária
10. Calcular o erro total
11. Se o erro for aceitável, encerrar. Caso contrário, passar para outro conjunto de treinamento retornando ao passo 1.

Uma rede neural constituída por um único neurônio é denominada de *perceptron* de camada única, equivalente à regressão logística univariada. Após a fase de treinamento, a rede está pronta para predizer a categoria de um novo vetor X (HAYKIN, 1999).

2.2.2.4 Algoritmos Genéticos

Algoritmos genéticos representam uma das poucas ideias que causaram uma repercussão semelhante ao conceito de seleção natural, proposto por Charles Darwin. Eles inspiram-se no processo de evolução natural e seus métodos de busca e otimização procuram resolver problemas encontrados no mundo real.

Conforme Mitchell (1997, pg. 249): "algoritmos genéticos tem sido aplicados com sucesso em uma variedade de tarefas de aprendizagem e para outros

problemas de otimização" (**tradução nossa**). A utilização desses algoritmos na robótica tem contribuído na aprendizagem de regras para controlar robôs; na área de inteligência artificial, são utilizados para aperfeiçoar a topologia e parâmetros de aprendizagem em uma rede neural artificial (MITCHELL, 1997).

Algoritmos genéticos abordam o seguinte problema: pesquisar de forma não determinística, com base na aleatoriedade e probabilidade, um espaço de hipóteses candidatas para identificar a melhor hipótese, que é definida como a que otimiza uma medida numérica predefinida para o problema à mão, chamada de "hipótese de aptidão".

Para Mitchell (1997, p.251), a função de seleção probabilística de uma hipótese no espaço de soluções (população) é dada como segue:

$$\Pr(h_i) = \frac{Fitness(h_i)}{\sum_{j=1}^{p} Fitness(h_j)}$$

De um modo geral, algoritmos genéticos apresentam as seguintes características (MITCHELL, 1997):

- Espaço de hipóteses: conjunto de indivíduos, onde cada indivíduo representa uma solução possível; tal solução é a medida de desempenho correspondente à função de aptidão do indivíduo, determinando quão apto ele é na solução de um problema;
- Operações genéticas: determinam a geração de população sucessora, através de um conjunto de operações que recombinam e transformam membros selecionados da população atual; essas operações correspondem à versão idealizada de operações genéticas encontradas na evolução biológica; as principais operações genéticas são: seleção, *crossover* (cruzamento) e mutação.
- Função de aptidão (*fitness*): define o critério de classificação de hipóteses potenciais e para selecionar probabilisticamente a inclusão na população da próxima geração;

Em muitos sistemas naturais, organismos individuais aprendem a se adaptar significativamente durante sua vida. Em se tratando de algoritmos genéticos, esse processo de adaptação ocorre de forma análoga aos processos biológicos.

Nesse sentido, sua adaptação ocorre através de buscas por soluções para problemas de otimização, de forma análoga ao processo de evolução natural. Na

maioria das vezes, eles convergem para um bom resultado, no entanto, seu desempenho nem sempre é ótimo. Para maximizar esse processo, a parametrização variável pode ter desempenho melhor em alguns problemas.

Existem diversas áreas que podem se beneficiar de algoritmos genéticos, como: Programação genética, Gerenciamento de redes, Problemas de otimização complexos, Aprendizagem de máquina, dentre outras.

2.2.3 Regras de Associação

São algoritmos que têm por objetivo encontrar relações entre dados que ocorrem com determinada frequência e que possam ser utilizados para identificar padrões de comportamentos.

Seu formalismo é dado como segue:

$\Gamma = \{i_1; i_2; i_m\}$: conjunto de m itens distintos e D uma base de dados formada por um conjunto de transações, onde cada T = conjunto de itens, tal que $T \subseteq \Gamma$ (AGRAWAL el al., 1993);

Uma regra de associação é igual:

$A \Rightarrow B$; onde $A \subset \Gamma$, $B \subset \Gamma$, $|A| > 0$, $|B| > 0$ e $A \cap B = \varnothing$

A = antecedente, B = consequente

O total de itens de um conjunto é chamado de tamanho do conjunto.

Parâmetros:

- confiança: é o resultado obtido pelo número de vezes em que A e B aparecem em uma mesma transação em relação ao total de vezes que A aparece no conjunto de transações;
- suporte: indica o número de ocorrências da regra A => B no conjunto das transações.

Tais parâmetros devem ser ajustados no processo de MD por interferirem diretamente na quantidade e qualidade das regras geradas; em grandes volumes de dados, é preciso definir o número mínimo de casos em que as regras se aplicam (suporte), além de um valor mínimo de confiança.

Na MDE, regras de associação podem ser interessantes para identificar padrões de navegação.

2.2.4 Clusterização

É uma técnica que busca descobrir conhecimento de forma indireta, a partir da identificação de grupos de dados com características semelhantes, identificar agrupamentos de dados que podem ser classificados em uma classe comum descoberta através de algoritmos com base estatística, procurando estabelecer os elementos centroides de cada *cluster* (BERRY; LINOFF, 1997).

A realização desta técnica é feita a partir de critérios apropriados para o particionamento de conjuntos de dados em subconjuntos (MARAVALLE et al., 1997), tais como:

- homogeneidade: é medida no interior de um mesmo *cluster*, verificando o grau de similaridade;
- separação: medida do quão diversos entre si são os elementos de *clusters.*

Na educação, a clusterização pode ser utilizada para a formação de grupos de trabalho.

2.2.5 Ferramentas para MD

Existe uma variedade de ferramentas de MD e frameworks que são de uso geral e específico, sendo que algumas dessas ferramentas são comerciais, como DBMiner, SPSS, *Clementine*, DB2 *Intelligent Miner*, e outras são de domínio público, como o **Weka** e RapidMiner (ROMERO et al., 2008).

O **Weka** é um software desenvolvido pelo grupo de aprendizado de máquina da Universidade de Waikato e tem por objetivo fornecer uma ampla coleção de algoritmos de aprendizado de máquina que resolvam problemas de MD. Os algoritmos podem ser aplicados diretamente sobre um conjunto de dados ou o programa pode ser executado a partir de um código Java.

Ele contém ferramentas para dados pré-processados com suporte a tarefas de classificação, regressão, agrupamento, regras de associação e visualização e, dentre alguns dos formatos de arquivos que ele utiliza, temos os formatos CSV (*Comma Separeted Value*) e ARFF (*Atribute-Relation File Format*), que correspondem a um arquivo de texto ASCII para descrever uma lista de instâncias compartilhando um conjunto de atributos, do tipo chave-valor.

O formato CSV tem em sua primeira linha a definição dos atributos, separados por vírgula e nas linhas seguintes, a definição das instâncias, cujos atributos são separados por vírugla.

O formato ARFF tem em sua primeira linha o nome da relação, identificado por **@relation** e possui duas seções:

- a seção de cabeçalho (**@attribute**) => contém informações referentes à definição dos atributos do conjunto de dados a ser considerado;
- a seção de dados (**@data**) => contém os dados da base de dados, separados por vírgula.

O **RapidMiner** foi desenvolvido pela Unidade de Inteligência Artificial da Universidade de Dortmund (Alemanha) em 2002, tendo sua primeira versão de código-livre e aberto, sob a licença GPL (Licença Pública Geral – *General Public License*) lançada em 2004. Ela cobre uma ampla faixa de tarefas de MD, oferecendo diversos algoritmos que trabalham com combinações de operadores, demandando que o seu utilizador inclua aqueles necessários à tarefa que pretende realizar (KAMPFF, 2009).

2.3 Moodle

O Moodle (*Modular Object-Oriented Development Learning Environment*) é um sistema gerenciador de cursos de aprendizagem de código aberto utilizado como auxílio aos educadores para criar efetivas comunidades de aprendizagem online (MOODLE ORG, 2015; COLE; FOSTER, 2007). É um sistema de distribuição livre, sob licença *open-source*, e em todo o mundo diferentes instituições, desde escolas primárias ate universidades, o utilizam como apoio no processo de ensino e aprendizagem, tanto em cursos a distância, como presenciais. A arquitetura do Moodle, sua implementação, interoperabilidade e intencionalidade são consideradas boas e tem uma comunidade muito forte, em 75 línguas; está presente em mais de 220 países e em mais de 20.300 sites (Al-AJLAN, 2008).

O Moodle é um projeto modular que facilita a criação de novos cursos, adicionando conteúdo que envolverá aprendizagem. Sua elaboração busca dar suporte a um estilo de aprendizagem chamado pedagogia construtivista social[4]

[4] Estilo cujo princípio é acreditar que os alunos aprendem melhor quando interagem com o material de ensino, constrói novos materiais para outros e interagem com outros alunos sobre o material (RICE, 2006).

(ROMERO et al., 2008). O Moodle não força a utilização desse estilo de aprendizagem, mas este é o tipo de comportamento que ele permite que o professor tenha.

Conforme Romero et al. (2008), o Moodle pode apresentar a flexibilidade modular descrita na **tabela 3**.

Tabela 3: Flexibilidade modular do MOODLE (Adaptada de ROMERO et al., 2008, p.4) (grifos nossos).

Cinco tipos de material de curso estático:	Seis tipos de material de curso interativo	Cinco tipos de atividades para interação entre alunos
• uma página de texto; • uma página *web*; • um *link* para alguma coisa na *web*; • visualizar um dos diretórios do curso; • um rótulo que mostra algum texto ou imagem.	• atribuições; • seleção; • diário; • lição; • perguntas; • exames.	• *chat*; • **fórum**; • glossário; • *wiki;* • seminários.

O Moodle mantém em sua base de dados *logs* detalhados de todas as atividades que os alunos desenvolvem (RICE, 2006), constando a trajetória dos materiais que os alunos acessaram. Além disso, ele registra os *cliques* dos alunos para fins de navegação e tem um sistema de visualização dos *logs*; os filtros desse sistema podem mostrar os *logs* por: curso, participante, dia e atividade. Através dos *logs* os professores podem determinar quais alunos estão ativos nos cursos, o que estão fazendo e quais deles estão fazendo. Desta forma, podem obter relatórios completos das atividades de um aluno, ou todos os alunos de uma atividade específica, podendo ser exibidas as atividades de diferentes dias ou horas (ROMERO et al., 2008).

A base de dados do Moodle registra todas as interações referentes ao uso do ambiente, à medida que o usuário o utiliza (ROMERO et al., 2008). As interações referentes à visualização dos módulos e realização dos exercícios, por exemplo, são armazenadas em tabelas correspondentes (grupos de tabelas que guardam informações sobre os módulos ativos e grupos de tabelas que guardam informações referentes aos questionários), dado a estrutura modular do curso.

O MOODLE possui um banco de dados extenso e com muitas tabelas, demandando muito estudo para melhor compreendê-lo. Ele registra as informações em um banco de dados que contém mais de 200 tabelas e o mesmo pode ser ampliado, conforme novas funcionalidades sejam acrescidas.

As tabelas seguem um padrão de organização em grupos, sendo que o nome de cada tabela sugere a que grupo pertence, conforme o modelo: mdl_(nome do grupo). Como exemplo, segue alguns nomes de tabelas agrupadas: **mdl_assignment, mdl_assignment_submissions, mdl_assignment_upgrade, mdl_course, mdl_course_categories, mdl_course_completions.**

2.4 Interações em fóruns de discussão

Os fóruns são uma poderosa ferramenta de comunicação e consistem em espaços de discussões e trocas de ideias a respeito de temas propostos por seus participantes. Através deles, cada participante pode tecer comentários sobre os temas discutidos, possibilitando, assim, o entendimento mútuo.

É uma ferramenta assíncrona, isto é, independente de tempo, que permite a conversa de todos com todos, cada qual a seu tempo, possibilitando a criação de um ambiente centrado na interação online. Sendo um espaço aberto para alunos e professores, os fóruns devem ser utilizados como estratégia de comunicação e diálogo entre tais atores, fazendo com que eles se movimentem na busca de entendimento e produção do saber (SCHERER, 2009).

Pesquisas recentes indicam que, quando bem concebidos, os fóruns motivam e melhoram a experiência de aprendizagem dos participantes, favorecendo o processo pedagógico e possibilitando ao aluno lograr êxito em cursos a distância (ABAWAJY, 2012). Conforme o autor, o objetivo final de fóruns de discussão assíncrono é criar um ambiente de aprendizagem online para atingir altos níveis de aprendizagem e para isso, aponta algumas das principais características que podem influenciar e diferenciar os vários tipos de fóruns de discussão:

- **grau de interação** - o sucesso ou fracasso de uma discussão online assíncrona está relacionado com a qualidade das interações predominantemente aluno-aluno e a profundidade da aprendizagem que ocorre nas discussões;
- **requisitos de participação** - participação em fórum de discussão fornece ao aluno oportunidades para aprendizagem ativa através da

leitura de outros comentários, postar suas próprias questões e fornecer resposta a outras postagens;

- **volume e frequência de postagens** - discussões assíncronas fornecem ao aluno tempo de reflexão e permite aos alunos compartilhar suas próprias perspectivas e analisar ponto de vista dos outros; a quantidade de mensagens geradas em discussões é um fator muito importante para tornar um fórum de discussão atraente e gerenciável; quando o número de mensagens aumenta, o fórum pode causar problema ao aluno na identificação de conteúdo relevante, digerir e fornecer resposta, dado que tem de fazer a triagem de postagens muitas vezes irrelevantes e desordenadas.
- **atividade de discussão** - está relacionada com a concepção do próprio fórum, no sentido de definir se a participação do aluno é obrigatória ou não, se há atribuição de nota, bem como os temas de discussão;
- **resposta** - a característica comum em fóruns de discussão é que os alunos postam mensagens em um local permanente, onde podem ser lidas e comentadas; além do mais, as postagens dos professores/instrutores, na forma de resposta são importantes para motivar os alunos a contribuírem na discussão; no entanto, respostas pouco frequentes, inexistentes, atrasadas, irrelevantes ou negativas, podem render discussões inúteis, impedindo a aprendizagem eficaz; os alunos podem deixar de contribuir se eles não recebem nenhuma resposta imediata ou comentários de outros participantes.

A discussão é normalmente considerada uma poderosa ferramenta para desenvolver habilidades pedagógicas, como o pensamento crítico, colaboração e reflexão. Os fóruns de discussão oferecem muitas vantagens pedagógicas, como incentivo à reflexão, análise e pensamento de ordem superior.

2.5 Considerações finais

Para esta pesquisa foi utilizado o Moodle como o AVA onde os dados foram coletados, porque, além das características descritas na **seção 2.3** deste capítulo, o Brasil é o terceiro país que mais o utiliza (MOODLE ORG, 2015). E ainda, por ele ser o AVA da instituição de ensino onde os dados foram coletados.

Dentre as tarefas apresentadas neste capítulo, esta pesquisa utilizou a tarefa de classificação, uma vez que ela se insere no contexto dos experimentos realizados. A criação dos modelos gerados por MD, através da ferramenta Weka, foi feita utilizando:

- três algoritmos baseados em árvores de decisão: **J48**, **BFTree** e **SimpleCart**;
- dois algoritmos baseados em estatística: **BayesNet** e **NaiveBayes**.

Para a ferramenta proposta à apresentação de diagnóstico gerado por MD, foram utilizadas as regras de classificação geradas pelas técnicas baseadas em **árvore de decisão**, a fim de auxiliar os interessados diretos desse ambiente (professores, tutores, gestores, etc.) na tomada de decisão.

Ressalta-se que esta pesquisa focou no uso de métodos baseados em árvore de decisão. Os métodos baseadas em estatística foram utilizadas a fim de que fosse possível fazer a comparação entre técnicas diferentes, para que os resultados obtidos pudessem ser mais bem consolidados. Devido a isso, não foram explorados em maior profundidade os métodos baseados em estatística.

3 TRABALHOS RELACIONADOS

Na literatura encontramos algumas iniciativas de MD em AVAs, sendo algumas delas no ambiente MOODLE, que visam descobrir padrões de comportamento a partir das interações dos alunos nesses ambientes. Dentre elas, as que mais se aproximam da metodologia proposta nesta pesquisa são as que seguem.

Kampff (2009) descreve uma pesquisa em que propõe uma arquitetura para sistemas de alertas, com alertas pré-definidos e outros gerados a partir de MD. Essa arquitetura identifica, por meio de MD gerados pela interação no AVA NetAula[5], comportamentos e características de alunos com risco de evasão ou reprovação e, então, alerta o professor, por meio de agrupamentos de alunos (grupos maiores que um (01) aluno) com características similares, para que o mesmo possa estabelecer comunicação personalizada e contextualizada com esses alunos. Desta forma, auxilia a atuação pedagógica do professor no acompanhamento das situações de aprendizagem. Os grupos identificados são dinâmicos, sendo gerados no momento oportuno para conscientizar o professor e os alertas são gerados nos resultados da MD, no formato de regras de classificação.

A avaliação da arquitetura apontou, conforme a autora, que as intervenções realizadas pelo professor, a partir dos alertas, direcionadas a grupos que compartilhavam necessidades específicas, contribuíram para a melhoria dos índices de aprovação e para redução dos índices de evasão dos alunos na disciplina acompanhada.

A pesquisa de Kampff (2009) faz uso da tarefa de classificação a partir da técnica *RuleLeaning* e, na geração de regras de decisão, utiliza a técnica *DecisionTree*, ambas implementadas na ferramenta RapidMiner, mas não há a comparação do desempenho das duas técnicas, nem mesmo com outras técnicas.

Marques (2014) propôs uma metodologia para MD educacionais em nove passos, baseada no estudo de Fayyad et al. (1996). O objetivo é identificar padrões de acesso dos alunos que evadem cursos na modalidade de EAD, por meio da MDE, gerando regras que caracterizam o perfil de acesso desses alunos. Busca tornar possível prever desistências por meio da análise de características de acesso

[5] É um AVA desenvolvido pela fábrica de software do setor de TI da Universidade Luterana do Brasil – ULBRA, para atender as necessidades da instituição, integrado ao sistema acadêmico. Disponível em: http://www.ulbra.br/ead. Acesso em 15/12/2014.

e características sociais do aluno no AVA MOODLE do SENAI-PB, possibilitando sugerir soluções para o problema dado, e em um tempo hábil para evitar a reprovação do aluno.

Para o estudo, Marques (2014) utilizou MD para classificação de exemplos, a fim de induzir um atributo presente nos dados. O trabalho focou na predição de uma variável categórica e binária, com o atributo "status do aluno", podendo ser "aprovado" ou "reprovado", por evasão ou não cumprimento das atividades propostas. Para isso, foi utilizado: a ferramenta RapidMiner®, escolhida pelo fato de ter interface gráfica intuitiva e ser fácil de usar, além de possibilitar o desenvolvimento incremental; o algoritmo C45/J48 (para o algoritmo J48, foi necessário instalar o pacote de expansão do WEKA), referente a árvore de decisão;

A autora procurou mapear o perfil de acesso dos alunos que desistem dos cursos oferecidos na EAD e pretende, futuramente, identificar esses perfis principalmente em 25% iniciais das aulas do curso, evidenciando automaticamente os resultados encontrados para os responsáveis educacionais.

O trabalho de Marques (2014) propõem uma metodologia para MD, mas os resultados obtidos da aplicação dessa metodologia são visualizados na própria interface do RapidMiner. Em relação a ferramentas de MD como esta, Romero et al. (2013, p.138) argumenta que elas não são especificamente projetadas para propósitos pedagógicos/educacionais e é muito complicado para um educador usá-las.

Santana (2014) propôs a aplicação de técnicas de classificação em um conjunto de dados de alunos de um curso para obter resultados como forma de apoio à tomada de decisão. Para isso, comparou algumas técnicas de classificação na interação Perfil de Uso do AVA, tendo como variável alvo o desempenho do aluno. Os resultados que foram considerados satisfatórios pelo autor foram gerados pela aplicação da técnica de árvore de decisão J48, que alcançou taxa de 74% de precisão. O trabalho de Santana (2014) faz uma discussão muito pertinente quanto à utilização e comparação de diversas técnicas de MD para avaliação do perfil de uso do AVA. No entanto, ele não propõe nenhuma estratégia para visualização dos resultados obtidos, no sentido de apresentar em maiores detalhes as regras de classificação geradas pelas árvores de decisão do algoritmo J48, bem como, não faz correlações entre as variáveis envolvidas no processo de MD.

Romero et al. (2013) desenvolveu uma ferramenta de MD específica para o Moodle, para que educadores, que normalmente não são experientes em MD, possam fazer uso dela, tendo em vista que tal ferramenta fornece algoritmos e parâmetros padrões para facilitar a utilização por parte desses usuários. Esta ferramenta tem o objetivo de predizer notas finais de alunos em cursos nesse ambiente e para isso, compara o desempenho de diversas técnicas de MD a partir de dados de alunos.

Para avaliar o desempenho e utilidade dos diferentes algoritmos de classificação utilizados em seu trabalho, alguns experimentos foram realizados usando todos os dados disponíveis (numéricos) e filtrados (categóricos, filtro por linha, filtro por coluna) de 438 alunos de 7 cursos de engenharia no Moodle da Universidade de Cordova, no sentido de se obter maior precisão nos resultados.

Para Romero et al. (2013), os resultados apontaram que os modelos obtidos usando dados categóricos são mais compreensíveis que aqueles que usam dados numéricos porque facilitam ao professor interpretar esses intervalos e magnitudes precisas.

Finalmente, os autores recomendam o uso de algoritmos de árvores de decisão, indução de regras e regras *fuzzy*, disponíveis na ferramenta, porque elas são modelos que favorecem resultados compreensíveis, permitem fazer uma interpretação do modelo obtido e podem ser usados para a tomada de decisão.

Os resultados obtidos pela ferramenta proposta por Romero et al. (2013) são gerados em arquivos no formato textual, o que ainda pode causar maior dificuldade na leitura e interpretação desses resultados.

Nesse sentido, a pesquisa proposta aqui busca ampliar as contribuições dos trabalhos correlatos, uma vez que, além do modelo preditivo, resultante da comparação de desempenho de algumas técnicas de MD, também apresenta uma ferramenta para visualização de diagnóstico. Apresenta ainda a comparação de atributos obtidos no processo de MD para proporcionar um nível maior de compreensão às partes interessadas. O diagnóstico apresentado na ferramenta é obtido a partir dos resultados dos experimentos realizados no Weka, permitindo visualizar mais detalhada e facilmente as tendências geradas pelo modelo, tais como: tendência a baixo desempenho e tendência à aprovação.

4 PROCEDIMENTOS METODOLÓGICOS PARA OS EXPERIMENTOS DE MD E DESENVOLVIMENTO DA FERRAMENTA PROPOSTA

A seguir são apresentados os procedimentos metodológicos para os experimentos realizados nesta pesquisa, visando gerar um modelo preditivo para diagnóstico de baixo desempenho a partir da utilização de MD. São apresentados ainda os procedimentos adotados para o desenvolvimento da ferramenta proposta.

4.1 Metodologia de pesquisa

O método científico é um procedimento cujo objetivo é conhecer, interpretar e intervir na realidade, tendo como diretriz problemas formulados que sustentam regras e ações adequadas à constituição do conhecimento, tendo por finalidade informar, descrever ou persuadir um fato (TARTUCE, 2006).

Nesta pesquisa faz-se uso de métodos indutivos, que é um processo cuja abstração é considerada insuficiente para propiciar um conhecimento completo do universo, pois entende que o conhecimento é fundamentado exclusivamente na experiência, sem levar em consideração princípios preestabelecidos, que é característico do método dedutivo (método que é baseado em regras e evidências, análise, síntese e enumeração matemáticas, partindo do geral para o específico) (MARCONI; LAKATOS, 2003). No método indutivo, parte-se do específico para o geral e apenas por meio da observação é possível formular uma hipótese explicativa da causa do fenômeno, ou seja, chega-se a conclusões que são apenas prováveis (MARCONI; LAKATOS, 2003).

A experimentação é um processo de validação de técnicas, abordagens, ferramentas e teses, e na computação, esse processo deve ser parte fundamental na criação de modelos (BASILI et al., 1986). O experimento é uma etapa de pesquisa científica no qual o pesquisador manipula e controla uma ou mais variáveis independentes e observa a variação nas variáveis dependentes (KERLINGER; LEE, 1964), com o objetivo de capturar a relação entre as causas e seus respectivos efeitos. Desta forma, o experimento é um método que investiga relações causais entre as variáveis para comprovar ou refutar pressupostos teóricos.

Tichy (1998) apresenta alguns benefícios da experimentação:

- pode conduzir a novos conhecimentos úteis e não esperados e, dessa forma, abrir novas áreas de investigação;

- pode acelerar o processo de eliminação de abordagens infrutíferas, orientando a engenharia e a teoria às direções certas.

Quanto à abordagem, a pesquisa é de caráter quantitativo, que é baseada na medida (normalmente numérica) de poucas variáveis objetivas, na ênfase em comparação de resultados e no uso intensivo de técnicas estatísticas.

A realização de experimentos permite testar teorias, explorar fatores críticos e trazem à tona novas questões, de modo que teorias possam ser formuladas ou corrigidas.

Desta forma, as etapas abrangidas nesta pesquisa são:

- Revisão de literatura:
- Coleta de dados;
- Experimentação: preparação dos dados coletados e aplicação de técnicas de MD para gerar o modelo de classificação através da ferramenta Weka;
- Análise dos resultados obtidos e apresentação da ferramenta proposta para visualização de diagnóstico gerado por MD.

O desenvolvimento da ferramenta para visualização de diagnóstico visa facilitar a compreensão por parte de usuários que não tenham domínio no uso de ferramentas como o Weka. Ela visa possibilitar várias estatísticas a respeito da situação dos alunos em um curso.

4.2 Arquitetura do modelo preditivo

A proposta metodológica desta pesquisa teve foco na utilização de dados oriundos da base de dados de um curso técnico em um AVA/Moodle de uma instituição de ensino que oferece cursos a distância. O objetivo foi diagnosticar alunos com tendência a baixo desempenho, aplicando técnicas de MD em conjuntos de dados contendo informações sobre interações realizadas por esses alunos em fóruns de discussão.

A **figura 4** mostra a arquitetura do modelo preditivo proposto nesta pesquisa, adaptada das etapas definidas por Fayyad et al. (1996), e as **seções 4.3, 4.4 e 4. 5** fazem sua descrição. Ressalta-se que a referida arquitetura trata das etapas necessárias à geração do modelo preditivo, e não de um modelo para implementação computacional. A proposta de implementação é apresentada na **seção 4. 7**.

Em Romero et al., (2011, p. 33), a transformação de dados é considerada como pertencente ao pré-processamento, e nesta pesquisa optou-se pela realização das transformações necessárias nos dados também nesta etapa, bem como a definição dos dados originais e, a partir dos dados originais, foram gerados os dados filtrados. Após o pré-processamento, foi feita a utilização de **algoritmos de MD** para induzir modelos para identificar, a partir das interações realizadas nos fóruns de discussão, se um aluno tem tendência a baixo desempenho ou aprovação, além de identificar atributos que estão relacionados à evasão.

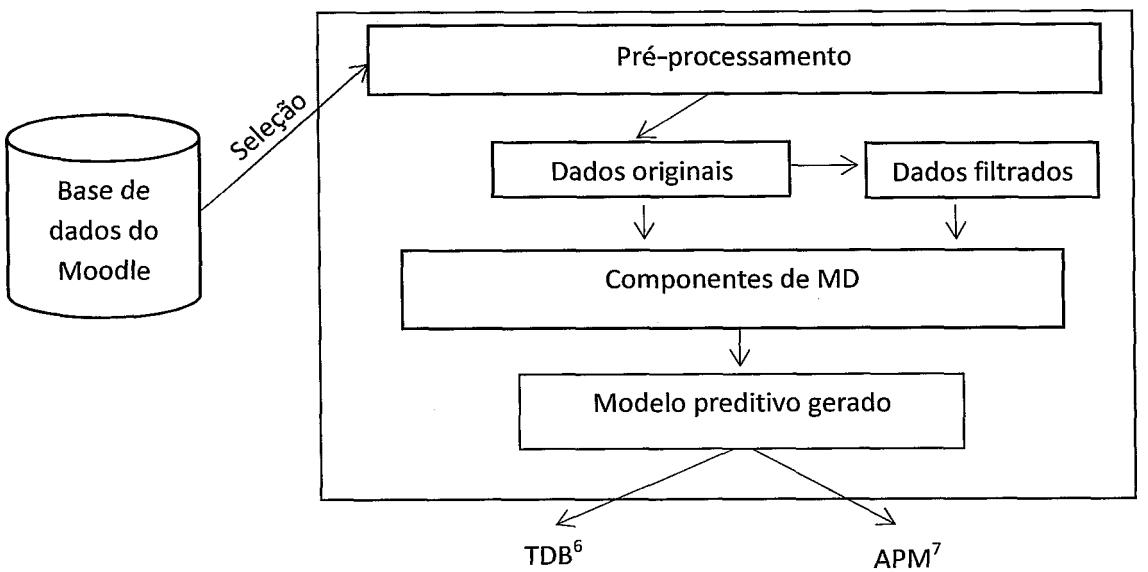


Figura 4. Arquitetura do modelo preditivo para prever baixo desempenho (adaptada de Fayyad et al., 1996)

4.3 Seleção de dados

A seleção de dados ocorreu a partir da escolha do curso de Técnico em Informática, composto de 25 disciplinas e com um total de 1171 alunos. No entanto, para cada disciplina tratada aqui, foram considerados apenas aqueles alunos que participaram em algum momento da disciplina. Dentre essas disciplinas, foram selecionadas três delas: **Sistemas Operacionais (SO)**, **Técnicas de Programação (TP)** e **Banco de Dados (BD)**, oferecidas no segundo módulo. Estas disciplinas foram escolhidas porque apresentaram de 02 (dois) a 03 (três) fóruns de discussão, quantitativo considerado adequado para o contexto desta pesquisa, bem como por

[6] TBD – Tendência a Baixo Desempenho
[7] APM – Aprovado Por Média

apresentarem a situação do aluno ao final da disciplina (**resultado**). A disciplina de SO teve **3 fóruns** de discussão, as de TP e BD, **2 fóruns** cada uma.

Os dados foram obtidos pela seleção realizada a partir das tabelas e atributos da base de dados do Moodle que registram as interações dos alunos nos fóruns de discussão, bem como no próprio AVA da instituição de ensino. O motivo da utilização de disciplinas para o processo de experimentação desta pesquisa diz respeito ao fato de que, no curso oferecido, quando o aluno reprova em alguma disciplina, automaticamente o mesmo está reprovado no curso.

4.4 Pré-processamento e definição dos conjuntos de dados

Para a criação da tabela de sumarização, referente ao conjunto de dados originais, novos atributos foram adicionados ou transformados, enquanto que outros foram desconsiderados por falta de relevância no contexto em estudo, para que fossem criados conjuntos de dados mais representativos e que pudessem possibilitar melhores resultados nos experimentos.

Tabela 4: Atributos da tabela de sumarização.

Atributo	Descrição
nota_media_em_foruns	nota média do aluno nos fóruns
total_participacao_em_todos_os_foruns	total de fóruns que o aluno participou
total_posts_em_todos_os_foruns	total de postagens em todos os fóruns
media_posts_por_forum	média de postagens por fórum
Resultado	resultado final do aluno na disciplina (atributo classe) [Aprovado Por Média, Aprovado Por Final, Reprovado Por Média, Reprovado Por Final]

A seleção dos atributos de cada tupla foi realizada manualmente e considerou o quantitativo de participação em fóruns e as postagens nas discussões, conforme sugerido por Abawajy (2012), por representarem fortes indicadores de interação nesta atividade, sendo que a tabela de sumarização ficou com 05 (cinco) atributos, conforme mostra a **tabela 4**. Outro critério adotado para a seleção de atributos está relacionado com fato de que foram consideradas somente disciplinas que já tinham sido encerradas, o que justifica a presença de atributos referentes à nota média (em fóruns e por postagens em fóruns). Para que o modelo seja

aplicado em uma disciplina em andamento, faz-se necessário uma nova configuração dos atributos.

O atributo **resultado** apresenta 04 (quatro) classes, como visto na **tabela 4**. Estas classes foram transformadas em apenas duas, sendo que foi mantida a classe Aprovado Por Média (doravante chamada de **APM)**, enquanto que as demais foram transformadas na classe Tendência a Baixo Desempenho (doravante chamada de **TBD**). Este procedimento foi adotado porque a aprovação por média é o que se espera de um aluno em uma situação ótima, entretanto, quando isso não acontece, significa que alguns fatores influenciaram para que o mesmo tivesse desempenho abaixo do esperado. Desta forma, as três classes foram consideradas uma só por elas representarem os alunos nesta condição.

A partir dos dados coletados, foram criados dois conjuntos de dados para cada disciplina em estudo, onde foi feita uma limpeza de dados através de filtro por linha. O primeiro conjunto de dados considerou apenas os alunos que participaram em algum momento das atividades propostas na disciplina e o segundo conjunto de dados foi formado apenas com os alunos que participaram da atividade fórum, com participação em pelo menos um fórum de discussão.

Com os dados pré-processados, foi necessário transformá-los nos formatos adequados aos algoritmos e ferramentas de MD utilizados, tais como em arquivos CSV ou em outros formatos específicos.

4.4.1 Definição dos conjuntos de dados da disciplina de SO

Para a disciplina de **SO**, a distribuição das classes referentes ao resultado do primeiro conjunto de dados, antes e depois da transformação aplicada no atributo **resultado**, é apresentada conforme a **tabela 5**.

Desta forma, o primeiro conjunto de dados foi formado por: **502 instâncias, 5 atributos e 2 classes**, sendo definido como o conjunto de dados originais. O segundo conjunto de dados da disciplina de **SO** foi obtido a partir do primeiro através de um filtro por linha, que consistiu em manter apenas os alunos que participaram de pelo menos um fórum da disciplina, o que resultou em apenas 353 alunos, sendo mantidos os 5 atributos do conjunto de dados anterior, bem como as duas classes referentes ao resultado, cuja distribuição é apresentada na **tabela 6**.

Tabela 5: Distribuição das classes do primeiro conjunto de dados da disciplina de SO, referente ao resultado.

Classe não transformada			Classe transformada		
Resultado	Total de Alunos	Percentual	**Resultado**	Total de Alunos	Percentual
Aprovado Por Média	**288**	**57,37%**	APM	**288**	**57,37%**
Aprovado Por Final	**148**	**29,48%**	TBD	**214**	**42,63%**
Reprovado Por Média	**63**	**12,55%**			
Reprovado Por Final	**3**	**0,6%**			
TOTAL	**502**	**100%**	TOTAL	**502**	**100%**

Tabela 6. Distribuição das classes do segundo conjunto de dados da disciplina de SO, referente ao resultado.

Resultado	Total de Alunos	Percentual
APM	**247**	**69,97%**
TBD	**106**	**30,03%**
TOTAL	**353**	**100%**

4.4.2 Definição dos conjuntos de dados da disciplina de TP

Para a disciplina de **TP**, a distribuição das classes referentes ao resultado do primeiro conjunto de dados, antes e depois da transformação aplicada no atributo **resultado**, é apresentada conforme a **tabela 7**.

Tabela 7: Distribuição das classes do primeiro conjunto de dados da disciplina de TP, referente ao resultado.

Classe não transformada			Classe transformada		
Resultado	Total de Alunos	Percentual	**Resultado**	Total de Alunos	Percentual
Aprovado Por Média	**240**	**49,79%**	APM	**240**	**49,79%**
Aprovado Por Final	**190**	**39,42%**	TBD	**242**	**50,21%**
Reprovado Por Média	**45**	**9,34%**			
Reprovado Por Final	**7**	**1,45%**			
TOTAL	**482**	**100%**	TOTAL	**482**	**100%**

Desta forma, o primeiro conjunto de dados foi formado por: **482 instâncias, 5 atributos e 2 classes**, sendo definido como o conjunto de dados originais. O segundo conjunto de dados da disciplina de **TP** foi obtido a partir do primeiro através de um filtro por linha, que consistiu em manter apenas os alunos que participaram de pelo menos um fórum da disciplina, o que resultou em apenas 241 alunos, sendo mantidos os 5 atributos do conjunto de dados anterior, bem como as duas classes referentes ao resultado, cuja distribuição é mostrada na **tabela 8**.

Tabela 8. Distribuição das classes do segundo conjunto de dados da disciplina de TP, referente ao resultado.

Resultado	Total de Alunos	Percentual
APM	**167**	**69,29%**
TBD	**74**	**30,71%**
TOTAL	**241**	**100%**

4.4.3 Definição dos conjuntos de dados da disciplina de BD

Para a disciplina de **BD**, a distribuição das classes referentes ao resultado do primeiro conjunto de dados, antes e depois da transformação no atributo **resultado**, é apresentada conforme a **tabela 9**.

Desta forma, o primeiro conjunto de dados foi formado por: **459 instâncias, 5 atributos e 2 classes**, sendo definido como o conjunto de dados originais.

Tabela 9: Distribuição das classes do primeiro conjunto de dados da disciplina de BD, referente ao resultado.

Classe não transformada			Classe transformada		
Resultado	Total de Alunos	Percentual	**Resultado**	Total de Alunos	Percentual
Aprovado Por Média	**217**	**47,28%**	APM	**217**	**47,28%**
Aprovado Por Final	**207**	**45,10%**	TBD	**242**	**52,72%**
Reprovado Por Média	**31**	**6,77%**			
Reprovado Por Final	**4**	**0,87%**			
TOTAL	**459**	**100%**	TOTAL	**459**	**100%**

O segundo conjunto de dados da disciplina de **BD** foi obtido a partir do primeiro através de um filtro por linha, que consistiu em manter apenas os alunos que participaram de pelo menos um fórum da disciplina, o que resultou em apenas 250 alunos, sendo mantidos os 5 atributos do conjunto de dados anterior, bem como as duas classes referentes ao resultado, cuja distribuição é apresentada na **tabela 10**.

Tabela 10. Distribuição das classes do segundo conjunto de dados da disciplina de BD, referente ao resultado.

Resultado	Total de Alunos	Percentual
APM	**174**	**69,6%**
TBD	**76**	**30,4%**
TOTAL	**250**	**100%**

Em conjuntos de dados como de um AVA é comum o desbalanceamento de dados, que ocorre quando o número de instâncias de uma classe é muito maior ou menor que o de outra, ou de outras. O problema de dados desbalanceados é que muitos algoritmos de classificação tendem a dar mais atenção às classes de maior frequência (classes majoritárias) na fase de treinamento, pois eles buscam maximizar a taxa de exatidão total do modelo, que é independente da distribuição das classes, ao passo que na etapa de teste tendem a ter baixa sensibilidade às classes minoritárias (ROMERO et al.; 2012, 2013).

Uma característica que é comum nos conjuntos de dados originais criados a partir da transformação das classes referentes ao resultado em todas as disciplinas é que as classes tiveram uma distribuição mais uniforme, ou seja, estão mais balanceadas. Ao observar as **tabelas 5, 7 e 9** (dados originais), onde especifica **a classe não transformada** em cada disciplina, fica evidente uma grande diferença entre os percentuais de alunos que foram aprovados (soma das aprovações por média e por final) e os que foram reprovados (soma das reprovações por média e por final), sendo que a aprovação prevaleceu acima de **86%**, enquanto que o percentual de reprovação ficou abaixo de **14%**. Com a transformação, percebe-se a ocorrência de um balanceamento entre as classes, onde o percentual de aprovação em cada disciplina oscilou entre **47%** e **57%**, enquanto que o percentual de reprovação oscilou entre **42%** e próximo de **53%**.

Em ralação aos dados filtrados, percebe-se que houve um desbalanceamento entre as classes em todas as disciplinas, sendo que a classe APM apresentou índice acima de **69%** e a classe TBD ficou abaixo de **31%**.

4.5 Mineração de dados

Para a realização dos experimentos foi utilizada a ferramenta de MD Weka, sendo que foram utilizadas as cinco técnicas de classificação apresentadas nesta pesquisa: **J48, BFTree, SimpleCart** (os três algoritmos baseados em árvore de decisão); ***Naive Bayes*** **e** ***Bayes Net*** (os dois algoritmos baseados em estatística). Optou-se pelo uso de árvore de decisão pelo fato dessa técnica proporcionar melhor compreensão por parte do público alvo a que se direciona esta pesquisa (ROMERO et al., 2013), uma vez que os resultados obtidos com outras técnicas de MD não são tão bem compreensíveis a um professor, por exemplo. Em relação ao uso de técnicas estatísticas, elas foram utilizadas para que fosse possível fazer a comparação entre técnicas diferentes, a fim de se ter um resultado mais consistente.

Foi aplicada ainda uma técnica de validação de dados, chamada **Validação Cruzada de Dez Partições** (*cross validation 10-folds*), que consiste em dividir os dados em dez partições aleatórias, onde são retiradas nove dessas partições para serem utilizadas no conjunto de treinamento e uma partição para o conjunto de testes (WITTEN et al., 2011). A validação cruzada é uma técnica bastante adequada quando se trabalha com conjuntos de dados pequenos e que não se pode fazer a divisão destes conjuntos em treinamento e teste, como os que são utilizados nesta pesquisa. Desta forma, a validação cruzada busca tentar amenizar o problema de pouca disponibilidade de dados.

Estudos relatados sugerem a adoção do número 10 como valor padrão para o número de partições dos dados (WITTEN et al., 2011), por permitir a previsão do comportamento da rede no futuro por meio da avaliação da exatidão da classificação obtida. Essa técnica é um recurso que possibilita teste e validação, através de parâmetros de validade e confiabilidade. Dessa primeira iteração é obtida a primeira precisão do modelo. Em seguida, para cada algoritmo de classificação aplicado, mais nove iterações percorrem todas as possibilidades de escolha. A precisão final do classificador é calculada, considerando a média das precisões das dez iterações.

Para a avaliação do desempenho dos algoritmos foram utilizadas as métricas *Precision* (percentual de amostras positivas classificadas corretamente

sobre o total de amostras classificadas como positivas), *Recall* (percentual de amostras positivas classificadas corretamente sobre o total de amostras positivas) e *F-measure* (média ponderada de *Precision e Recall*).

Para cada disciplina tratada nesta pesquisa, foram realizados 2 experimentos, um com dados originais e outro com dados filtrados, num total de 6 experimentos, com 5 execuções cada um (considerando que cada experimento faz uso de 5 métodos de MD), ou seja, 30 execuções no total.

O objetivo dos experimentos foi verificar se um modelo preditivo com maior precisão para o diagnóstico de tendência de baixo desempenho seria obtido usando dados filtrados ou dados originais. Buscou-se ainda, verificar a partir do segundo conjunto de dados em cada disciplina, quais indicativos das interações em fóruns de discussão influenciaram no resultado final do aluno na respectiva disciplina. Por fim, foi feita a comparação entre os atributos mais relevantes em cada disciplina para verificar se algum deles esteve presente em todas as disciplinas em estudo, o que certamente caracteriza esse(s) atributo(s) como sendo muito representativo(s) quanto à influencia causada no resultado final do aluno.

4.6 Ferramenta para visualização de diagnóstico utilizada nesta pesquisa

A ferramenta utilizada nesta pesquisa está sendo desenvolvida na linguagem de programação PHP 6.0, com servidor APACHE 2.2 e banco de dados MySql 5.0. Ela possui três módulos (ver arquitetura na **figura 5**) e, até o momento, dois deles estão sendo implementados: **Pré-processamento de dados** e **Visualizador de Diagnóstico**. Nesta pesquisa é apresentado o módulo **Visualizador de Diagnóstico**, da ferramenta proposta.

O módulo de Pré-processamento controla as rotinas de acesso e consulta ao banco de dados para obter os dados necessários à MD; este módulo está descrito na **seção 4.3** desta pesquisa. O módulo de MD é responsável por todo o processo de MD; nesta pesquisa, a MD é realizada a partir da ferramenta Weka (ver **seção 4.4**), no entanto, está sendo viabilizada a possibilidade de implementação deste módulo da ferramenta proposta. O módulo visualizador de diagnóstico é voltado à apresentação de diagnóstico das tendências trabalhadas; maiores detalhes são descritos na **seção 5.3**, **capítulo 5**, referente à análise dos resultados a partir da ferramenta proposta.

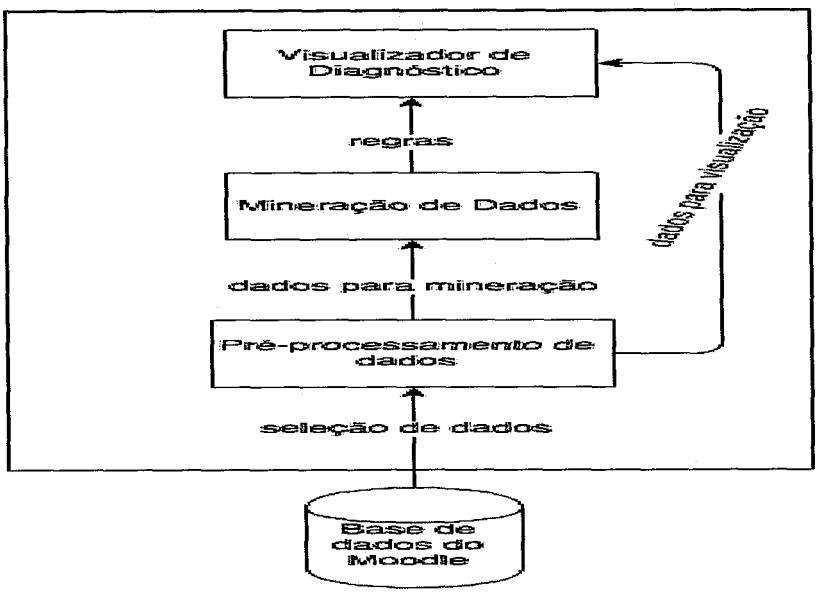


Figura 5. Arquitetura da ferramenta para visualização de diagnóstico

Definida as tabelas e os atributos a serem utilizados, é realiza tanto as consultas na base de dados do Moodle, como a formatação do resultado em um formato de arquivo reconhecido pelo Weka (no caso, o arquivo .csv). Esses resultados são gravados em arquivos, que são os conjuntos de dados contendo os atributos definidos na **tabela 4** e as instâncias, que representam os alunos. Ressalta-se que alguns dados foram obtidos a partir do próprio AVA, através de seus recursos de exportação.

Após a MD, as regras geradas pelos modelos baseados em árvore de decisão foram adaptadas em regras de classificação para uso na ferramenta proposta e, juntamente com os dados de alunos utilizados na MD, foram carregados na ferramenta para apresentar o diagnóstico. A adaptação da árvore de decisão em regras de classificação foi feita manualmente e gravada em arquivo no formato .txt, para ser utilizado pela ferramenta proposta.

A **figura 6** apresenta a tela inicial do módulo **Visualizador de Diagnóstico** da ferramenta proposta nesta pesquisa, onde são carregados os dados para MD e as regras de classificação que foram geradas no *Weka* e adaptadas para uso na ferramenta.

A ferramenta tem o propósito de possibilitar o diagnóstico para novos alunos e poderá ser usada por gestores para visualizar as tendências de alunos ao longo de suas participações no curso.

Figura 6. Tela inicial da aplicação web para auxiliar o diagnóstico de evasão

Ela representa uma ampliação da análise que se faz a partir do Weka, pois apresenta de forma mais detalhada quantitativos de alunos em cada tendência trabalha, e na regra de classificação específica em que eles se inserem. Além do mais, possibilita ampliar a compreensão aos interessados diretos (professores, tutores, gestores, etc.) na tomada de decisão. A ideia é integrá-la como um módulo do Moodle para facilitar a análise da situação dos alunos a partir do diagnóstico apresentado no próprio AVA e já está sendo estudada esta possibilidade.

5 ANÁLISE DOS RESULTADOS

Este capítulo apresenta a análise dos resultados obtidos nos experimentos realizados com os conjuntos de dados e as técnicas de MD definidos em capítulos anteriores, onde, para cada disciplina, é feita a comparação das taxas de desempenho dos algoritmos nos conjuntos de dados usados, é definida a técnica de MD que obteve melhor desempenho e são apresentados os principais indicadores que podem influenciar nas tendências tratadas nesta pesquisa.

Foi realizada a comparação entre os indicadores utilizados em cada processo de MD nas disciplinas em estudo, onde são apresentados índices e algumas relações geradas por eles entre as disciplinas.

Os resultados obtidos nos experimentos de MD desta pesquisa também foram analisados a partir da ferramenta proposta para a visualização de diagnóstico como fins de teste e análise do seu funcionamento. Como se trata apenas de teste, é apresentado somente uma análise, referente a dados originais da disciplina de **TP**. Através desta ferramenta é apresentado o diagnóstico gerado por MD, a fim de que a visualização de dados seja mais bem compreendida. Esta aplicação tem como entrada os dados obtidos no experimento, bem como as regras geradas na MD.

5.1 Análise das taxas de desempenho, definição da técnica mais adequada para classificação e principais indicadores

5.1.1 Disciplina de SO

A **tabela 11** mostra algumas taxas referentes ao desempenho dos classificadores na disciplina de **SO**. Os valores em negrito na **tabela 11** representam a média ponderada das medidas e permitem identificar mais facilmente o classificador com melhor desempenho.

No experimento com dados originais, as taxas de *Precision* e *Recall* dos algoritmos nas classes trabalhadas se mantiveram com um bom equilíbrio, com valores entre **0,62** e abaixo de **0,81**, enquanto que o *Recall* de todos os algoritmos para a classe TBD, referente a dados filtrados, ficou abaixo de **0,40**, e a medida de *Precision* teve valores entre **0,51** e **0,67**. O processo de filtragem piorou o *Recall* de TBD e isso pode ser atribuído, também, ao desbalanceamento que há entre as classes APM e TBD. Além disso, a precisão do classificador fica menor quando há menos dados disponíveis (ROMERO et al., 2013).

Tabela 11. Desempenho dos classificadores nos experimentos da disciplina de SO.

		dados originais			dados filtrados			
Métodos	Algoritmo	*Precision*	*Recall*	*F-Measure*	*Precision*	*Recall*	*F-Measure*	Classe
Árvores de Decisão	J48	0,717	0,626	0,668	0,667	0,264	0,378	TBD
		0,746	0,816	0,779	0,749	0,943	0,835	APM
		0,733	**0,735**	**0,732**	**0,724**	**0,739**	**0,698**	
	BFTree	0,654	0,645	0,649	0,600	0,283	0,385	TBD
		0,739	0,747	0,743	0,749	0,919	0,825	APM
		0,703	**0,703**	**0,703**	**0,704**	**0,728**	**0,693**	
	SimpleCart	0,667	0,654	0,660	0,650	0,245	0,356	TBD
		0,747	0,757	0,752	0,744	0,943	0,832	APM
		0,713	**0,713**	**0,713**	**0,716**	**0,734**	**0,689**	
Bayesianos	Naive Bayes	0,696	0,621	0,657	0,561	0,302	0,393	TBD
		0,740	0,799	0,768	0,750	0,899	0,818	APM
		0,721	**0,723**	**0,721**	**0,693**	**0,720**	**0,690**	
	BayesNet	0,680	0,636	0,657	0,514	0,340	0,409	TBD
		0,742	0,778	0,759	0,753	0,862	0,804	APM
		0,715	**0,717**	**0,716**	**0,681**	**0,705**	**0,685**	

O primeiro experimento usou dados originais e os melhores índices de desempenho no experimento foram registrados pelo algoritmo J48 (com médias *Precision* = 0,733, *Recall* = 0,735 e *F-Measure* = 0,732), seguido pelo algoritmo Naive Bayes (com médias *Precision* = 0,721, *Recall* = 0,723 e *F-Measure* = 0,721). Os índices obtidos mostram que para dados originais, as duas técnicas tiveram desempenho muito semelhantes, com uma leve vantagem das técnicas baseadas em árvore de decisão em relação às bayesianas. Considerando que as classes APM e TBD neste experimento estão praticamente balanceadas, pode-se dizer que o desempenho apresentado por estes algoritmos indica que as instâncias foram classificadas de forma mais acertada.

No segundo experimento, usando dados filtrados por linha, percebe-se um maior distanciamento das técnicas baseadas em árvore de decisão em relação aos métodos bayesianos. Os melhores índices de desempenho foram registrados pelo algoritmo J48 (com médias *Precision* = 0,724, *Recall* = 0,739 e *F-Measure* = 0,698), seguidos pelo desempenho apresentado pelo algoritmo Naive Bayes (com médias *Precision* = 0,693, *Recall* = 0,720 e *F-Measure* = 0,690). No entanto, conforme mencionado anteriormente, existe um desbalanceamento entre as classes APM, com muito mais instâncias, e TBD, em menor quantidade, o que pode significar que o bom desempenho desses algoritmos tenha sido influenciado pela classe majoritária (ROMERO et al., 2013), ou seja, os algoritmos podem ter classificado

corretamente muito mais instancias da classe majoritária, em detrimento da minoritária.

Outra informação essencial para a análise dos experimentos realizados refere-se à matriz de confusão (GOLDSCHIMIDT; PASSOS, 2005). Ela é uma característica que trata da qualidade da classificação realizada, mostrando quais foram os casos em que o algoritmo se confundiu na classificação. A diagonal principal da matriz apresenta as instâncias classificadas corretamente para cada classe e corresponde à precisão do classificador. As **tabelas 12 e 13** mostram a matriz de confusão do algoritmo J48 usando dados originais e filtrados, respectivamente.

Tabela 12. Matriz de confusão do algoritmo J48 no experimento com dados *originais*, disciplina de SO.

	TBD	APM
TBD	**134**	**80**
APM	**53**	**235**

Ao verificar a matriz de confusão gerada pelo algoritmo J48 no experimento com dados originais (**tabela 12**) percebe-se que 80 instâncias foram classificadas erroneamente como sendo da classe **APM**, sendo que eram da classe **TBD**. Da mesma forma, 53 instâncias foram classificadas erroneamente como sendo de **TBD**, quando, na realidade, eram de **APM**.

Tabela 13. Matriz de confusão do algoritmo J48 no experimento com dados filtrados por linha, disciplina de SO.

	TBD	APM
TBD	**28**	**78**
APM	**14**	**233**

A **tabela 13** apresenta a matriz de confusão gerada pelo algoritmo J48 em dados filtrados por linha, onde se ver que a grande maioria das instâncias de **TBD** foram classificadas erroneamente como sendo de **APM**, correspondendo a 78 instâncias e apenas 28 foram classificadas corretamente; por outro lado, apenas 14 instâncias foram classificadas como sendo de **TBD**, quando, na verdade, eram de **APM**.

A matriz de confusão mostra se o algoritmo está errando muito mais em uma classe do que na outra, principalmente quando se trata de dados desbalanceados, como é o caso do conjunto de dados filtrados por linha (**tabela 13**), que apresentou

um desbalanceamento de suas classes e isso refletiu no grande número de instâncias classificadas erroneamente como **APM**, quando eram **TBD**. Além disso, a precisão do classificador é menor se há poucos dados (ROMERO et al., 2013).

As técnicas baseadas em árvores de decisão obtiveram índices significativos de precisão nos conjuntos de dados utilizados nesta pesquisa, sendo que o algoritmo J48 foi o que obteve melhor desempenho, apesar de uma pequena diferença com o algoritmo Naive Bayes. Através das regras de classificação geradas por ele é possível apontar quais fatores são mais indicativos para diagnosticar alunos com tendência a baixo desempenho ou aprovação.

No primeiro experimento o algoritmo J48 gerou uma árvore de decisão composta por 8 regras de classificação (ver **figura 7**) e elas indicam que quando a **nota media em fóruns** foi maior que 4.67, o aluno foi aprovado por média sem depender de outras variáveis. Abaixo dessa média, o algoritmo identificou, por exemplo, que quando não houve participação em fórum (**total_foruns <= 0**, correspondendo a mais de 20% dos casos), ocorreu a classe TBD.

```
nota_media_em_foruns <= 4.67
|   total_foruns <= 0: TBD (149.0/41.0)
|   total_foruns > 0
|   |   total_foruns <= 2
|   |   |   nota_media_em_foruns <= 4.17
|   |   |   |   nota_media_em_foruns <= 2.75: APM (19.0/6.0)
|   |   |   |   nota_media_em_foruns > 2.75
|   |   |   |   |   total_posts_em_todos_os_foruns <= 1: TBD (27.0/7.0)
|   |   |   |   |   total_posts_em_todos_os_foruns > 1
|   |   |   |   |   |   nota_media_em_foruns <= 3.08: TBD (3.0/1.0)
|   |   |   |   |   |   nota_media_em_foruns > 3.08: APM (6.0)
|   |   |   nota_media_em_foruns > 4.17: TBD (5.0)
|   |   total_foruns > 2: TBD (4.0)
nota_media_em_foruns > 4.67: APM (289.0/69.0)
```

Figura 7. Árvore de decisão gerada pelo algoritmo J48 com dados originais, disciplina de SO.

Conforme tratado por Abawajy (2012), os fóruns podem motivar e melhorar a experiência de aprendizagem dos participantes, favorecendo o processo pedagógico, além de possibilitar ao aluno lograr êxito em cursos a distância. Quando isso não ocorre, pode acontecer da não participação em fórum produzir o efeito contrário ao que diz este autor, como foi o caso apontado pelo algoritmo nesse experimento.

As variáveis **total_foruns** e **total_posts_em_todos_foruns** foram aspectos importantes para predizer o resultado final na disciplina, quando a nota média em fóruns foi menor ou igual a 4.67.

Em relação ao segundo experimento, o algoritmo J48 gerou uma árvore de decisão composta por 5 regras (ver **figura 8**) e elas indicam que a **nota média em fóruns** e **quantidade de postagens nos fóruns**, foram fatores importantes para definir o resultado do aluno na disciplina.

```
nota_media_em_foruns <= 5.33
|   total_foruns <= 2
|   |   nota_media_em_foruns <= 2.75: APM (19.0/6.0)
|   |   nota_media_em_foruns > 2.75
|   |   |   total_posts_em_todos_os_foruns <= 1: TBD (27.0/7.0)
|   |   |   total_posts_em_todos_os_foruns > 1: APM (25.0/10.0)
|   total_foruns > 2: TBD (7.0)
nota_media_em_foruns > 5.33: APM (275.0/63.0)
```

Figura 8. Árvore de decisão gerada pelo algoritmo J48 com dados filtrados por linha, disciplina de SO.

Uma observação exclusiva para o contexto desses experimentos é que o algoritmo considerou a participação em mais de dois fóruns, em alguns casos, como sendo um indício de baixo desempenho, tanto que quando o total de fóruns por aluno foi superior a 2, esses alunos foram definidos como sendo da classe TBD. Considerando que a disciplina de SO teve 3 fóruns de discussão, desconsiderando o resultado final dos alunos e observando a participação, fica evidente que apenas dois deles foram significativos para alguns alunos.

5.1.2 Disciplina de TP

A **tabela 14** mostra algumas taxas referentes ao desempenho dos classificadores na disciplina de **TP**. Os valores em negrito na **tabela 14** representam a média ponderada das medidas e permitem identificar mais facilmente o classificador com melhor desempenho.

Em dados originais, as taxas de *Recall* dos algoritmos nas classes trabalhadas oscilaram bastante, tendo valores entre **0,59** e **0,80**, enquanto que os valores para *Precision* ficaram entre **0,66** e **0,76**. O intervalo de valores para *Recall* foi bem maior que o de *Precision*, partindo desde uma classificação ruim (menor que **0,65**), em alguns momentos, até uma considerada boa (acima de **0,75**). De modo geral, em cada classe tratada, as taxas de *Precision* ficaram bem próximas, da

mesma forma que as taxas de *Recall*, ou seja, os algoritmos conseguiram classificar de forma equilibrada as instâncias.

Tabela 14. Desempenho dos classificadores nos experimentos da disciplina de TP.

		dados originais			dados filtrados			
Método	Algoritmo	*Precision*	*Recall*	*F-Measure*	*Precision*	*Recall*	*F-Measure*	Classe
Árvore de Decisão	J48	0,687	0,789	0,735	0,675	0,365	0,474	TBD
		0,750	0,638	0,689	0,766	0,922	0,837	APM
		0,718	**0,714**	**0,712**	**0,738**	**0,751**	**0,725**	
	BFTree	0,672	0,806	0,733	0,639	0,311	0,418	TBD
		0,755	0,604	0,671	0,751	0,922	0,828	APM
		0,714	**0,705**	**0,702**	**0,717**	**0,734**	**0,702**	
	SimpleCart	0,664	0,802	0,727	0,581	0,338	0,427	TBD
		0,747	0,592	0,660	0,753	0,892	0,816	APM
		0,706	**0,697**	**0,694**	**0,700**	**0,722**	**0,697**	
Bayesianos	Naive Bayes	0,684	0,744	0,713	0,478	0,297	0,367	TBD
		0,717	0,654	0,684	0,733	0,856	0,790	APM
		0,701	**0,699**	**0,699**	**0,655**	**0,685**	**0,660**	
	BayesNet	0,693	0,690	0,692	0,500	0,216	0,302	TBD
		0,689	0,692	0,690	0,722	0,904	0,803	APM
		0,691	**0,691**	**0,691**	**0,654**	**0,693**	**0,649**	

Usando dados filtrados, percebe-se que os valores de *Precision* e *Recall* em cada classe tiveram variação de valores bem maior que com dados originais. As taxas de *Precision* tiveram valores entre **0,47** e **0,76**, enquanto que as de *Recall* tiveram valores entre **0,21** (muito ruim) e **0,92** (muito bom). Isto demonstra o quanto a classificação é afetada pelo desbalanceamento de dados e também quando há poucas instâncias para classificar (ROMERO et al., 2013), no sentido de favorecer a classe majoritária (ver **tabela 8**). Pode-se ver isso a partir das taxas de *Recall* para a classe APM, que tiveram valores entre e **0,85** e **0,92**, enquanto que, para a classe TBD, os valores de *Recall* oscilaram entre **0,21** e **0,36**.

Nos dois experimentos realizados para esta disciplina, as técnicas baseadas em árvore de decisão obtiveram os melhores desempenhos. Em dados originais, o algoritmo J48 foi o melhor deles (Precision = 0,718, Recall = 0,714 e F-Measure = 0,712), seguido pelo algoritmo BFTree (Precision = 0,714, Recall = 0,705 e F-Measure = 0,702). Em dados filtrados, novamente os dois algoritmos foram melhores: J48 (*Precision* = 0,738, *Recall* = 0,751 e *F-Measure* = 0,725) e BFTree (Precision = 0,717, Recall = 0,734 e F-Measure = 0,702).

As **tabelas 15 e 16** mostram a matriz de confusão do algoritmo J48 usando dados originais e filtrados, respectivamente.

Tabela 15. Matriz de confusão do algoritmo J48 no experimento com dados originais, disciplina de TP.

	TBD	APM
TBD	**191**	**51**
APM	**87**	**153**

Ao verificar a matriz de confusão gerada pelo algoritmo J48 no experimento com dados originais (**tabela 15**) percebe-se que 51 instâncias foram classificadas erroneamente como sendo de **APM**, sendo que eram de **TBD**. Da mesma forma, 87 instâncias foram classificadas erroneamente como sendo de **TBD**, quando, na realidade, eram de **APM**.

Tabela 16. Matriz de confusão do algoritmo J48 no experimento com dados filtrados por linha, disciplina de TP.

	TBD	APM
TBD	**27**	**47**
APM	**13**	**154**

A **tabela 16** apresenta a matriz de confusão gerada pelo algoritmo J48 em dados filtrados por linha, onde se ver que um pouco mais da metade das instâncias de **TBD** foram classificadas erroneamente como sendo de **APM**, correspondendo a 47 instâncias (67%) e apenas 27 foram classificadas corretamente (36%); por outro lado, apenas 13 instâncias (~7,8%) foram classificadas como sendo de **TBD**, quando, na verdade, eram de **APM**.

```
nota_media_em_foruns <= 7.38
|   nota_media_em_foruns <= 3.63: TBD (258.0/77.0)
|   nota_media_em_foruns > 3.63
|   |   total_participacao_em_todos_os_foruns <= 1: APM (21.0/4.0)
|   |   total_participacao_em_todos_os_foruns > 1: TBD (32.0/11.0)
nota_media_em_foruns > 7.38: APM (171.0/36.0)
```

Figura 9. Árvore de decisão gerada pelo algoritmo J48 com dados originais, disciplina de TP.

No primeiro experimento o algoritmo J48 gerou uma árvore de decisão composta por 4 regras de classificação (ver **figura 9**), utilizando apenas as variáveis **nota_media_em_foruns** e **total_participacao_em_todos_os_foruns**. Estras regras indicam que quando a **nota media em fóruns** foi maior que 7.38, o aluno foi aprovado por média sem depender de outras variáveis. Abaixo dessa média, o

algoritmo identificou que quando a média final é menor ou igual a 3.63, os alunos nesta condição estão na classe TBD. Acima dessa média e tendo participado de mais de um fórum, o algoritmo também classificou os alunos nesta condição como sendo da classe TBD. Esta disciplina ofereceu 2 fóruns de discussão, no entanto, o algoritmo considerou relevante apenas a participação em um deles.

```
nota_media_em_foruns <= 8.38
|   total_posts_em_todos_os_foruns <= 2
|   |   nota_media_em_foruns <= 3.63
|   |   |   total_posts_em_todos_os_foruns <= 1: TBD (11.0/2.0)
|   |   |   total_posts_em_todos_os_foruns > 1
|   |   |   |   nota_media_em_foruns <= 2.5: TBD (2.0)
|   |   |   |   nota_media_em_foruns > 2.5: APM (2.0)
|   |   nota_media_em_foruns > 3.63
|   |   |   total_posts_em_todos_os_foruns <= 1: APM (21.0/4.0)
|   |   |   total_posts_em_todos_os_foruns > 1
|   |   |   |   nota_media_em_foruns <= 4.75: TBD (4.0)
|   |   |   |   nota_media_em_foruns > 4.75: APM (58.0/20.0)
|   total_posts_em_todos_os_foruns > 2: TBD (20.0/5.0)
nota_media_em_foruns > 8.38: APM (123.0/20.0)
```

Figura 10. Árvore de decisão gerada pelo algoritmo J48 com dados filtrados por linha, disciplina de TP.

Em relação ao segundo experimento, o algoritmo J48 gerou uma árvore de decisão composta por 8 regras (ver **figura 10**), desta vez utilizando as variáveis **nota_media_em_foruns** e **total_de_posts_em_todos_os_foruns**. Quando a nota média foi maior que 8.38, todos os alunos nesta condição foram considerados como sendo da classe APM. Abaixo desta nota, uma série de relações entre essas duas variáveis foram estabelecidas para definir a situação final do aluno. Alunos que postaram acima de duas mensagens foram considerados como sendo da classe TBD. Para os que postaram até duas mensagens, ainda foi considerada a nota média em fóruns para a definição de qual classe o aluno estaria inserido.

5.1.3 Disciplina de BD

A **tabela 17** mostra algumas taxas referentes ao desempenho dos classificadores na disciplina de **BD**. Os valores em negrito na **tabela 13** representam a média ponderada das medidas e permitem identificar mais facilmente o classificador com melhor desempenho.

Em dados originais, as taxas de *Precision e Recall* dos algoritmos nas classes trabalhadas se mantiveram com um bom equilíbrio, com valores entre **0,69** e abaixo de **0,81**. Em relação a dados filtrados, percebe-se que o *Recall* de todos os algoritmos para a classe TBD ficou abaixo de **0,30**, enquanto que esta mesma medida, para a classe APM, teve índices acima de **0,86**. A medida de *Precision* teve

valores entre **0,47** e **0,65** para a classe TBD e valores acima de **0,74** para a classe APM. Novamente, percebe-se que o processo de filtragem e desbalanceamento contribuiu para que o desempenho da classe TBD fosse baixo.

Tabela 17. Desempenho dos classificadores nos experimentos da disciplina de BD.

		dados originais			dados filtrados			
Método	Algoritmo	*Precision*	*Recall*	*F-Measure*	*Precision*	*Recall*	*F-Measure*	Classe
Árvore de Decisão	J48	0,749	0,814	0,780	0,566	0,395	0,465	TBD
		0,770	0,696	0,731	0,766	0,868	0,814	APM
		0,759	**0,758**	**0,757**	**0,706**	**0,724**	**0,708**	
	BFTree	0,749	0,814	0,780	0,549	0,368	0,441	TBD
		0,770	0,696	0,731	0,759	0,868	0,810	APM
		0,759	**0,758**	**0,757**	**0,695**	**0,716**	**0,698**	
	SimpleCart	0,749	0,814	0,780	0,523	0,303	0,383	TBD
		0,770	0,696	0,731	0,743	0,879	0,805	APM
		0,759	**0,758**	**0,757**	**0,676**	**0,704**	**0,677**	
Bayesianos	Naive Bayes	0,779	0,715	0,746	0,565	0,342	0,426	TBD
		0,709	0,774	0,740	0,755	0,885	0,815	APM
		0,746	**0,743**	**0,743**	**0,697**	**0,720**	**0,697**	
	BayesNet	0,768	0,752	0,760	0,574	0,408	0,477	TBD
		0,730	0,747	0,738	0,770	0,868	0,816	APM
		0,750	**0,749**	**0,750**	**0,711**	**0,728**	**0,713**	

O experimento usando dados originais apontou que as técnicas baseadas em árvore de decisão foram ligeiramente melhores que as *bayesianas*, sendo que os três algoritmos de árvore de decisão tiveram exatamente os mesmos índices de desempenho (com médias *Precision* = 0,759, *Recall* = 0,758 e *F-Measure* = 0,757), seguido pelo algoritmo *Naive Bayes* (com médias *Precision* = 0,750, *Recall* = 0,749 e *F-Measure* = 0,750). Os índices obtidos mostram que para dados originais, as duas técnicas tiveram desempenho muito semelhantes. No segundo experimento, usando dados filtrados por linha, percebe-se que o algoritmo *Bayes Net* obteve o melhor índice de desempenho (com médias *Precision* = 0,711, *Recall* = 0,738 e *F-Measure* = 0,713), seguido pelo algoritmo J48 (com médias *Precision* = 0,706, *Recall* = 0,724 e *F-Measure* = 0,708).

Apesar de o algoritmo *Bayes Net* ter o melhor desempenho em dados filtrados, optou-se por fazer a análise dos resultados gerados pelo algoritmo J48, visto que a diferença de percentual foi muito pequena entre os dois e também por se tratar de uma técnica baseada em árvore de decisão.

As **tabelas 18 e 19** mostram a matriz de confusão do algoritmo J48 usando dados originais e filtrados, respectivamente.

Tabela 18. Matriz de confusão do algoritmo J48 no experimento com dados originais, disciplina de BD.

	TBD	APM
TBD	**197**	**45**
APM	**66**	**151**

Ao verificar a matriz de confusão gerada pelo algoritmo J48 no experimento com dados originais (**tabela 18**) percebe-se que 45 instâncias foram classificadas erroneamente como sendo de **APM**, sendo que eram de **TBD**. Da mesma forma, 66 instâncias foram classificadas erroneamente como sendo de **TBD**, quando, na realidade, eram de **APM**.

Tabela 19. Matriz de confusão do algoritmo J48 no experimento com dados filtrados por linha, disciplina de BD.

	TBD	APM
TBD	**30**	**46**
APM	**23**	**151**

A **tabela 19** apresenta a matriz de confusão gerada pelo algoritmo J48 em dados filtrados por linha, onde se ver que 46 instâncias de **TBD** (60%) foram classificadas erroneamente como sendo de **APM**, e apenas 30 foram classificadas corretamente (40%). Por outro lado, apenas 23 instâncias (15%) foram classificadas como sendo de **TBD**, quando, na verdade, eram de **APM**.

As técnicas baseadas em árvores de decisão obtiveram índices significativos de desempenho nos conjuntos de dados utilizados nos experimentos da disciplina de BD, sendo que, em geral, o algoritmo J48 foi o que obteve melhor desempenho. A partir das regras de classificação geradas por ele é possível apontar quais fatores são mais indicativos para diagnosticar alunos com tendência a baixo desempenho.

```
nota_media_em_foruns <= 7: TBD (264.0/65.0)
nota_media_em_foruns > 7: APM (195.0/43.0)
```

Figura 11. Árvore de decisão gerada pelo algoritmo J48 com dados originais, disciplina de BD.

No primeiro experimento o algoritmo J48 gerou uma árvore de decisão composta apenas por 2 regras de classificação (ver **figura 11**), onde o atributo **nota_media_em_foruns** foi o único fator determinante para indicar a situação final

do aluno. Para isso, quando a nota media em fóruns foi menor ou igual a 7, o aluno foi identificado como pertencente à classe TBD, significando que, aproximadamente, um pouco mais da metade do total de alunos (57%) se encontram nesta condição.

```
nota_media_em_foruns <= 7: TBD (55.0/22.0)
nota_media_em_foruns > 7: APM (195.0/43.0)
```

Figura 12. Árvore de decisão gerada pelo algoritmo J48 com dados filtrados por linha, disciplina de BD.

No segundo experimento o algoritmo J48 gerou uma árvore de decisão também composta por apenas 2 regras (ver **figura 12**), com apenas um atributo (**nota_media_em_foruns)** determinando a situação do aluno e estas regras indicam que quando a **nota média em fóruns** foi menor ou igual a 7, o aluno foi identificado como pertencente à classe TBD, significando que, aproximadamente, 22% do total de alunos se encontram nesta condição.

A **tabela 20** sintetiza o desempenho do algoritmo J48 nos experimentos realizados, onde é apresentada a medida de *F-mesuare*, que foi utilizada na definição do algoritmo de melhor desempenho.

Tabela 20. Síntese do desempenho do algoritmo J48 nos experimentos realizados

			Valores de *F-measure*		
Dados	Experimentos	Classe	*F-measure*	Valor mínimo	Valor máximo
Originais	Experimento 1 – SO	APM	0,779	0,668	0,780
	Experimento 2 – TP	APM	0,689		
	Experimento 3 – BD	APM	0,731		
	Experimento 1 – SO	TBD	0,668		
	Experimento 2 – TP	TBD	0,735		
	Experimento 3 – BD	TBD	0,780		
Filtrados	Experimento 1 – SO	APM	0,835	0,378	0,837
	Experimento 2 – TP	APM	0,837		
	Experimento 3 – BD	APM	0,814		
	Experimento 1 – SO	TBD	0,378		
	Experimento 2 – TP	TBD	0,474		
	Experimento 3 – BD	TBD	0,465		

Ao observar a **tabela 20**, é possível verificar que quando foram usados dados originais, os modelos gerados pelo algoritmo J48 tiveram os melhores desempenhos. Embora o valor máximo de *F-measure* nas classes em dados

filtrados tenha sido **0,837**, o valor mínimo foi muito ruim, ficando em **0,378**. Em relação aos valores dessa mesma medida para as classes em dados originais, eles ficaram entre **0,668 e 0,780**, que é um intervalo de desempenho considerado bom, ainda que o valor máximo em dados originais tenha sido menor que o valor máximo em dados filtrados.

Uma possível causa para a pouca eficiência de alguns algoritmos nos conjuntos de dados utilizados, como o *Naive Bayes*, pode estar relacionada ao fato de que alguns algoritmos tendem a selecionar atributos apropriadamente e ignorar aqueles que são redundantes ou irrelevantes, enquanto que outros algoritmos não fazem isso (MIHAESCU; BURDESCU, 2006). Outro fator possível está relacionado com o baixo número de atributos. Por fim, pode estar relacionado à natureza e implementação do próprio algoritmo, que pode ser mais apropriado para usar dados numéricos ou categóricos (ROMERO et al., 2013).

5.2 Comparação de indicadores

Nesta seção são apresentados os indicadores que foram considerados no processo de MD como sendo aqueles com maior relevância na tarefa de classificação (ver **tabela 21**), em cada disciplina analisada. A identificação desses indicadores é um fator de grande importância para auxiliar as partes interessadas na tomada de decisão, tanto em relação á própria disciplina, como também para um curso.

Tabela 21. Indicadores utilizados nas regras de classificação geradas pelo algoritmo J48 nas disciplinas analisadas.

Disciplina	Indicadores (atributos)
SO	nota_media_em_foruns, total_foruns total_posts_em_todos_os_foruns
BD	nota_media_em_foruns
TP	nota_media_em_foruns total_participacao_em_todos_os_foruns total_posts_em_todos_os_foruns

À exceção da variável **total_participação_em_todos_os_foruns**, que ocorreu apenas nas regras de classificação geradas a partir de dados originais, na disciplina de TP, as demais variáveis estiveram presentes tanto nas regras geradas usando dados originais, quanto filtrados.

Ao observar a **tabela 21**, e considerando o contexto dos experimentos realizados, percebe-se que a variável **nota_media_em_foruns** foi considerada em 100% dos experimentos para a geração das regras de classificação e pode ser vista como o principal indicador para determinar a situação final de um aluno em uma disciplina, com base nas interações realizadas nos fóruns de discussão das disciplinas trabalhadas. Isso pode ser observado pelo fato de que em todas as regras geradas, esta variável sempre apareceu como o nó raiz da árvore, que é considerado pelo classificador como sendo o atributo que tem grande impacto na classificação dos casos dados.

Pode-se perceber ainda, que a variável **total_posts_em_todos_os_foruns** também foi de grande importância para 2/3 (dois terços) das disciplinas (SO e TP), pois ela representa a quantidade de mensagens postadas pelos alunos nos fóruns em que participaram.

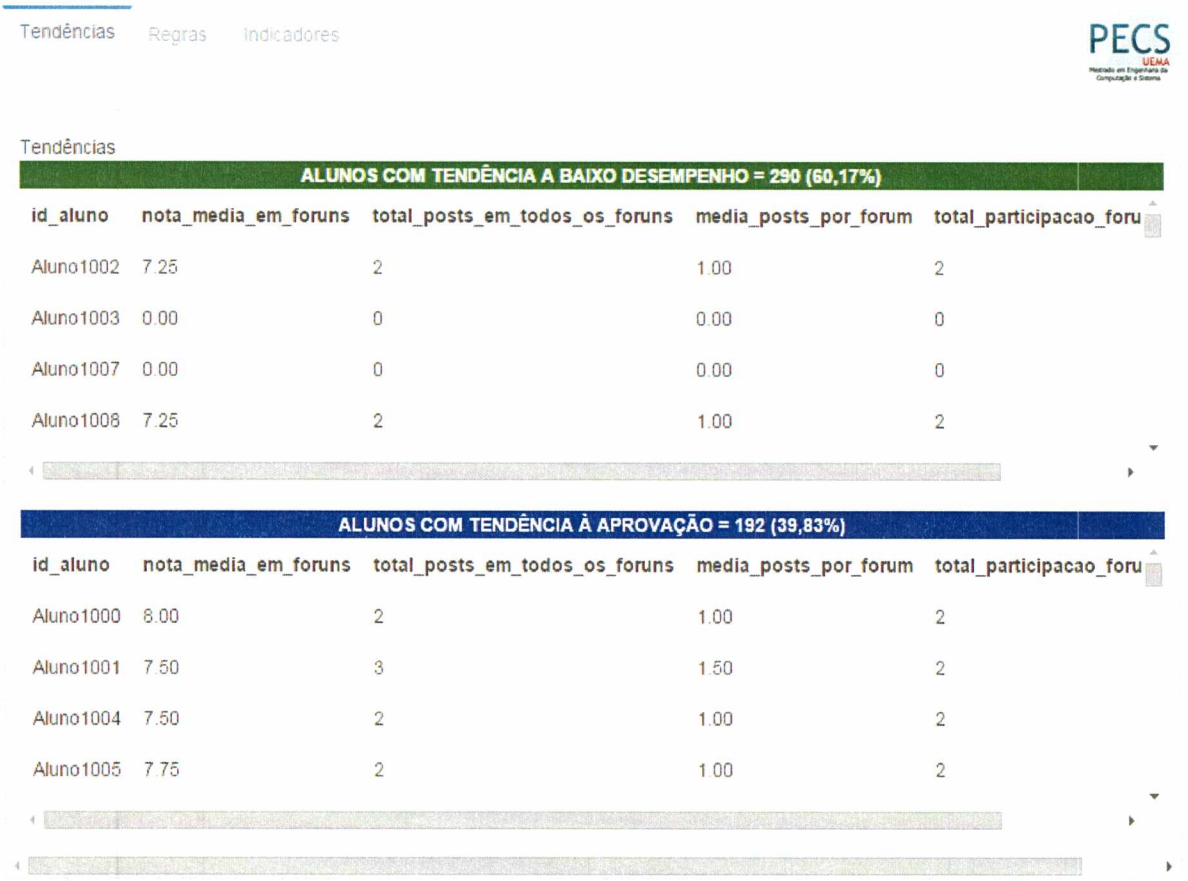
Algumas comparações podem ser feitas entre os indicadores apresentados na **tabela 14**, nas diferentes disciplinas.

Na disciplina de BD, apenas foi considerada a variável **nota_media_em_foruns**, tanto em dados originais, como filtrados, sendo que os alunos que tiveram nota abaixo de 7 foram definidos como sendo da classe TBD e os demais, da classe APM. Na disciplina de TP, com dados filtrados, mesmo a nota média em fórum tendo sido considerada a partir de **4.75**, o algoritmo considerou que, se o aluno postou **1** mensagem, ele foi definido como sendo da classe APM. Situação semelhante ocorre na disciplina de SO, sendo que a nota média considerada foi ainda menor, entre **3.08** e **4.17**, e a quantidade de postagens igual a **1** mensagem, onde o aluno também foi considerado como sendo da classe APM.

Em compensação, em alguns casos, também na disciplina de TP, se o aluno teve nota média em fóruns acima de **3.63** e até **4.75**, mesmo tendo postado mais de **1** mensagem, o mesmo foi considerado como sendo da classe TBD. Isso pode significar que, muitas postagens podem não causar efeito prático em relação ao resultado final do aluno. A quantidade de postagens em fóruns é discutida por Abawajy (2012), onde conforme o autor, o aumento no número de mensagens pode causar problema relacionado a conteúdo, uma vez que dificulta digerir e fornecer resposta, dado que é preciso fazer a triagem de postagens muitas vezes irrelevantes e desordenadas.

5.3 Diagnóstico através da ferramenta desenvolvida para esta pesquisa

Na **figura 13** são exibidas as tendências geradas na árvore de decisão referente a dados originais da disciplina de **TP,** juntamente com o total de alunos que fazem parte de cada uma dessas tendências. Através desta tela é possível ver quais alunos se inserem em cada tendência. Percebe-se que o percentual de alunos com tendência à aprovação na disciplina em estudo foi de 39,83%, enquanto que o percentual de alunos com tendência a baixo desempenho correspondeu a 60,17%.


Tendências Regras Indicadores

PECS UEMA

Tendências

ALUNOS COM TENDÊNCIA A BAIXO DESEMPENHO = 290 (60,17%)

id_aluno	nota_media_em_foruns	total_posts_em_todos_os_foruns	media_posts_por_forum	total_participacao_foru
Aluno1002	7.25	2	1.00	2
Aluno1003	0.00	0	0.00	0
Aluno1007	0.00	0	0.00	0
Aluno1008	7.25	2	1.00	2

ALUNOS COM TENDÊNCIA À APROVAÇÃO = 192 (39,83%)

id_aluno	nota_media_em_foruns	total_posts_em_todos_os_foruns	media_posts_por_forum	total_participacao_foru
Aluno1000	8.00	2	1.00	2
Aluno1001	7.50	3	1.50	2
Aluno1004	7.50	2	1.00	2
Aluno1005	7.75	2	1.00	2

Figura 13. Tela com as tendências trabalhadas na pesquisa.

Conforme tratado por Abawajy (2012), os fóruns podem motivar e melhorar a experiência de aprendizagem dos participantes, favorecendo o processo pedagógico, além de possibilitar ao aluno lograr êxito em cursos a distância. Apesar disso, percebe-se que menos da metade dos alunos nesta disciplina obtiveram êxito no final (aproximadamente 40%).

A **figura 14** mostra a tela onde são exibidas as regras geradas na árvore de decisão, bem como o total de alunos que fazem parte de cada regra e, como visto anteriomente, o algoritmo gerou uma árvore de decisão composta por 4 regras.

Através das regras de classificação geradas é possível apontar quais fatores são mais indicativos para diagnosticar alunos com tendência a baixo desempenho. Elas indicam que a **nota média em fóruns** e a **participação em fóruns de discussão** foram fatores importantes para definir o resultado do aluno na disciplina.

Cada regra gerou um quantitativo de alunos e ao observá-las mais a fundo, vê-se que aqueles cuja nota média em fóruns foi maior que 7.38 foram aprovados sem depender de outras condições, perfazendo um total de 171 alunos (**regra 4**). Os alunos que obtiveram nota média em fóruns entre 3.63 e 7.38 e participaram de mais de 1 fórum foram classificados como com tendência a baixo desempenho, num total de aproximadamente 32 (**regra 3**).

Figura 14. Tela com as regras de classificação adaptadas a partir do algoritmo J48.

A partir da análise das regras é possível observar a distribuição das classes em cada regra, e como pode-se ver, a classe APM está distribuída em duas das 4 regras obtidas (**regras 2 e 4**), sendo que a classe TBD está presenta nas **regras 1 e 3**.

Esta visualização de dados possibilitará uma melhor análise dos resultados obtidos a partir da MD realizada no experimento proposto. Outras telas estão em

fase de implementação para apresentar mais detalhes do diagnóstico gerado e não foram apresentadas neste momento.

6 CONSIDERAÇÕES FINAIS

Esta pesquisa investigou como os dados armazenados em um AVA podem ser transformados em informações potencialmente úteis para apoiar o acompanhamento de alunos em cursos EAD.

Nesta pesquisa foi feito um levantamento do estado da arte referente a um dos principais problemas da EAD, que é a evasão. Foram apresentados índices tanto externos como internos ao Brasil, onde se observou que esses índices são altos e preocupantes. No Brasil, o percentual de evasão é de aproximadamente 20% (ABED, 2013) e as causas para ocorrência deste fenômeno são as mais diversas, dentre as quais, as apresentadas por Abbad et al. (2010) e Bruno et al. (2010). Foi visto ainda que o baixo desempenho é um dos fatores que interferem negativamente nas taxas de evasão, essencialmente nos semestres iniciais de um curso (BRUNO, 2011).

No intuito de proporcionar as condições necessárias para reduzir ou eliminar a evasão na EAD, se torna essencial a identificação de indicadores de baixo desempenho em cursos a distância. Desta forma, foi visto que a utilização de métodos e ferramentas de análise de dados é importante para observar o comportamento dos alunos, servindo como auxilio às partes interessadas na tomada de decisão. Nesse sentido, a MD foi considerada como sendo uma abordagem bastante adequada para o contexto educacional, visto que ela pode ser utilizada para descobrir automaticamente relações ocultas presentes nas informações sobre os alunos, em especial, as interações realizadas em fórum de discussão, e, assim, contribuir na melhoria da aprendizagem (ROMERO et al., 2008).

Foi visto que os fóruns representam a principal ferramenta de comunicação em um AVA e, conforme Abawajy (2012), o objetivo de um fórum de discussão é proporcionar um ambiente de aprendizagem online para se obter altos níveis de aprendizagem.

Foram apresentadas algumas tarefas de MD, dentre as quais, destacou-se a tarefa de classificação para o propósito desta pesquisa, onde foram apresentadas as técnicas utilizadas para a realização dos experimentos tratados nos **capítulos 4 e 5**.

Através dos experimentos realizados buscou-se obter um modelo preditivo com desempenho bom o suficiente para que fosse capaz de predizer quando um

aluno apresenta características tendenciosas a baixo desempenho a partir de suas interações em fóruns de discussão.

As técnicas baseadas em árvores de decisão são recomendadas no contexto educacional, uma vez que elas geram um resultado mais compreensível e fácil de interpretar ao usuário que as utilizar para a tomada de decisão (ROMERO et al., 2013) e nos experimentos desta pesquisa observou-se que elas obtiveram os melhores desempenhos. Em **83%** dos experimentos realizados, o algoritmo J48 teve o melhor desempenho, sendo que a média de *Recall* nesses experimentos, que é uma boa medida para determinar o desempenho real de um classificador, foi de **73,96%**, e a média de *F-measure* foi de **72,48%**, sendo esta uma medida determinante para a escolha do classificador mais adequado. Desta forma, as técnicas baseadas em árvore de decisão são as mais indicadas, dentre as testadas, para a geração de um diagnóstico mais preciso das tendências tratadas nesta pesquisa.

Ao se observar as taxas de *F-measure*, tanto em dados originais, quanto em dados filtrados, percebe-se que os modelos obtidos a partir de dados originais, no contexto dos experimentos realizados aqui, tiveram melhor desempenho.

Foi feita ainda comparação entre os indicadores identificados pelo algoritmo J48 nas diferentes disciplinas trabalhadas, para possibilitar melhor compreensão para auxiliar as partes interessadas na tomada de decisão.

Tão importante quanto a MD é a visualização dos resultados gerados para as partes interessadas, uma vez que, nem sempre, tais usuários tem um conhecimento a fundo da tecnologia e do que representam esses resultados quando os mesmos são obtidos a partir de ferramentas como o Weka. Desta forma, esta pesquisa apresentou, ainda, uma ferramenta para visualização de dados, para proporcionar maior facilidade na compreensão e análise dos resultados gerados pelos algoritmos classificadores utilizados.

Como visto nos resultados, a ferramenta proporciona uma análise mais detalhada da MD gerada pelo Weka, uma vez que ela faz listagens específicas de alunos, tais como: por tendência à aprovação, tendência a baixo desempenho, bem como, para cada regra gerada pelo modelo, faz a listagem de alunos em cada uma delas, apresentando percentuais correspondentes.

O foco desta pesquisa foi especificamente nas interações em fóruns de discussão, que tem sua relevância em uma análise de interações em AVAs, uma vez

que representa a principal ferramenta de comunicação nesses ambientes. Como meta futura, pretende-se reunir dados de diversos recursos do Moodle para uma análise de caráter mais amplo, bem como abranger mais cursos e disciplinas. Pretende-se ainda verificar a possibilidade de extração de regras de classificação de técnicas diferentes de **árvore de decisão e de extração de regras**.

A ferramenta apresentada nesta pesquisa ainda não está finalizada, apesar das funções que possui. Como trabalhos futuros, pretende-se adicionar novas funcionalidades a ela e, para isso, seguem alguns desafios:

- finalizar a implementação dos módulos pré-processamento de dados e visualizador de diagnóstico;
- implementar o módulo de MD;
- mostrar a situação individual de cada aluno de forma mais detalhada;
- implementar a exibição dos indicadores de baixo desempenho e;
- integrá-la como um módulo do Moodle;

Através da MD em AVAs é possível verificar a relação entre uma abordagem pedagógica e o aprendizado do aluno, a fim de que o professor avalie se sua abordagem realmente está ajudando ou não o aluno a ter um bom desempenho nas atividades propostas.

REFERÊNCIAS

ABAWAJY, J. Analysis of Asynchronous Online Discussion Forums for Collaborative Learning. International Journal of Education and Learning, v. 1, n. 2, p. 2012.

ABBAD, Gardênia da Silva; ZERBINI, Thaís; SOUZA, Daniela Borges Lima de. **Panorama das pesquisas em educação a distância no Brasil**. In: Estudos de Psicologia, 15(3), setembro-dezembro/2010. Disponível em: <http://www.scielo.br/epsic >. Acesso em: 15 out. 2014.

AGRAWAL, R.; IMIELINSKI, T.; SRIKANT R. **Missing Association Rules between Sets of Items in Large Databases**. In: Proc. of the ACM SIGMOD Intl. Conf. on Management of Data, US, Washington, 1993, 207-216.

AL-AJLAN Ajlan, ZEDANs Husein. Why Moodle, in International Workshop on Future Trends of Distributed Computing System. IEEE 2008.

ALVES, Lynn; NOVA, Cristiane. **Educação a distância**: Limites e Possibilidades. Capítulos do Livro: ALVES, Lynn; NOVA, Cristiane. Educação a Distância: Uma nova concepção de aprendizado e interatividade. São Paulo: Futura, 2003.

ASSOCIAÇÃO BRASILEIRA DE EDUCAÇÃO A DISTÂNCIA. **Censo EAD.BR. Relatório Analítico da aprendizagem a distância no Brasil 2013**. Disponível em: http://www.abed.org.br/censoead2013/CENSO_EAD_2013_PORTUGUES.pdf. Acesso em 24 abr 2015.

BASILI VR, Selby RW, Hutchens DH. **Experimentation in software engineering. Software Engineering**, IEEE Trans- actions on. 1986; 7:733-43.

BRASIL, MEC, SEB. **PLANO DE QUALIDADE PARA A EDUCAÇÃO BÁSICA**: Diagnóstico e ações para elevar o nível de qualidade do ensino nas escolas brasileiras, Brasília, março de 2005.

BERRY, Michael J. A.; LINOFF, Gordon. **Data Mining Techniques:** For Marketing,Sales, and Customer Support. US: Jonh Wiley & Sons, 1997.

BRUNO, Maria, F. **Causas da evasão em curso de graduação a distância em administração em uma universidade pública federal**; Rev. Teoria e Prática da Educação, v14, n.3p. 43-56; Set-Dez. 2011. Disponível em: <http:periódicos.uem.br/ojs/index.php/TeorPratEduc/article/view/18487 >. Acesso em: 21 nov. 2015.

BRUNO, G. J. (et all). **Evasão na educação a distância**: um estudo sobre a evasão em uma instituição de ensino superior, 2010. Disponível em: <http://www.abed.org.br/congresso2010/cd/252010220450.pdf>. Acesso em: 23 out. 2014.

BREIMAN, L., FRIEDMAN, J. H., Olshen, R. A., & Stone, C. J. **Classification and Regression Trees**. Wadsworth. 1984.

COLE, J., FOSTER,H. **Using Moodle: Teaching with the Popular Open Source Course Management System**. Second Edition. O'Reilly Community Press: Printed in the United States of America, November, 2007.

CRAMER, J. S. **Econometric Applications of Maximum Likelihood Methods**. Cambridge University Press. Cambridge. 1986

DEOGUN, Jitender S. et al. **Data Mining: Trend in Research and Development**. In: LIN, T. Y; CERCONE, N.Rough Sets and Data Mining: Analysis for Imprecise Data. Springer, 1997.

FAVERO, R. V. M.; FRANCO, S. R. Um estudo sobre a permanência e a evasão na Educação a Distância. **RENOTE - Revista Novas Tecnologias na Educação**, [S.1.], v.4, n.2, 2006.

FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery:** An Overview. Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

GARCIA, Enrique et al. **Drawbacks and solutions of applying association rule mining in learning management systems**. International Workshop on Applying Data Mining in e-learning (ADML'07), in Second European Conference on Technology Enhanced Learning (EC-TEL07). Crete, Greece, September, 2007. Disponível em: <http://wwwis.win.tue.nl/~tcalders/pubs/GARCIAADML2007.pdf>. Acessado em: 24/08/2015.

GIBSON. Chere Campbell. **The distance learner's academic self-concept**. Capítulo de Livro: GIBSON, Chere Campbell. Distance Learners in higher education: Institutional responses for quality outcomes. Madson, WI: Atwood Plublishing, 1998.

GOLDSCHIMIDT, R.; PASSOS, E. **Data mining**: Um guia prático. Rio de Janeiro: Campus, 2005.

HAN, Jiawei; KAMBER, MICHELINE; PEI, Jian. **Data Mining: concepts and techniques**. 3nd. ed. San Francisco: Morgan Kaufmann Publishers / Elsevier, 2012.

HAYKIN, S. **Redes neurais**: Princípios e prática. Tradução Engel, P. M. Bookman, 1999.

KAMPFF, A. J. C. **Mineração de Dados Educacionais para Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente**. 2009. Tese (Doutorado) – PPGIE/UFRGS.

KERLINGER F. N.; Lee H. B. **Foundations of behavioral research**: Educational and psychological inquiry. New York: Holt, Rinehart and Winston. 1964.

LIN, T. Y; CERCONE, N. **Roug Sets and Data Mining: Analysis for Imprecise Data**. springer, 1997.

MARCONI, Maria de Andrade; LAKATOS, Eva Maria. **Fundamentos de Metodologia Científica**. 5ª Ed. São Paulo: Atlas, 2003.

MARQUES, J.L. de Q. **Mineração de dados educacionais**: um estudo de caso utilizando o ambiente virtual do SENAI. Campina Grande - Paraíba. 2014.

MARAVALLE, M.; SIMEONE, B.; NALDINI, R. **Clustering on trees**. Computational Statistics & Data Analysis, v. 24, n. 2, 1997, p. 217-234.

MIHAESCU, C.; BURDESCU, D. **Testing attribute selection algorithms for classification performance on real data**, In: Proceedings of the IEEE Conference Intelligent Systems, London, UK, 2006, pp 581–586.

MITCHELL, T. M.. **Machine Learning**. McGraw-Hill, New York, 1997.

MOBASHER, B., **Web Usage Mining and Personalization, in Practical Handbook of Internet Computing**, M.P. Singh, Editor. CRC Press, Boca Raton, FL, 2004, pp. 1–35.

MOODLE Org. Disponível em: <https://moodle.org/>. Acesso em: 13 jan. 2015.

MOORE, M., KEARSLEY, G. **Educação a distância**: uma visão integrada. São Paulo: Thompson Pioneira, 2007.

OBBADI, M.; JURBERG, C. **Educação a distância**: algumas reflexões sobre a desistência. Tecnologia Educacional. Ano 33, n. 167/169, out. /04; p.47-58, jun. 2005.

PALLOF, Rena M. PRATT, Keith. **O aluno virtual:** um guia para trabalhar com estudantes online. Porto Alegre: Artmed, 2004.

PEREIRA, A. T. C. **Ambientes Virtuais de Aprendizagem – Em Diferentes Contextos**. Rio de Janeiro: Editora Ciência Moderna Ltda., 2007.

QUINLAN, J. R. **C4.5**: Programs for Machine Learning. Morgan Kaufmann, 1993. Publishers, San Mateo, CA.

Rice, W.H. (2006) **Moodle E-learning Course Development**. A complete guide to succeful learning using Moodle. Packt publishing.

ROMERO, C., VENTURA, S., G, ENRIQUE. (2008). Data mining in course management systems: Moodle case study and tutorial. **Computers & Education**, 51(1): 368–384.

_____, VENTURA, S., PECHENIZKIY, M., BAKER, R.S.J.d. **Handbook of Educational Data Mining,** Ed. C R C, 2011, 503p.

_____, VERA, C., M., SOTO, S., V. Predicción del Fracaso Escolar mediante Técnicas de Minería de Datos. **IEEE-RITA Vol. 7, Núm. 3, Nov. 2012**.

_____. ESPEJO, P., G, P. G., ZAFRA . A. (2013). **Web usage mining for predicting final marks of students that use Moodle courses**. Disponível em: <https://www.deepdyve.com/lp/wiley/web-usage-mining-for-predicting-final-marks-of-students-that-use-moodle-courses-P23vC7HY6h/1>. Acesso em 23 de abril de 2015.

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.

SANTANA, Leandro, C. (2014). Avaliação do Perfil de Uso no Ambiente Moodle Utilizando Técnicas de Mineração de Dados. **Revista Brasileira de Informática na Educação. Pág. 269.**

SCHERER, Suely. **Educação bimodal: habitantes, visitantes ou transeuntes?** In: VALENTE, J. A. e BUSTAMANTE, S. B. V. Educação a Distância: prática e formação do profissional reflexivo. São Paulo: Avercamp, 2009. 259 p.

SHI, H. **Best-first decision tree learning**. Master's thesis, University of Waikato, Hamilton, NZ. 2007.

TARTUCE, T. J. A. **Métodos de pesquisa**. Fortaleza: UNICE – Ensino Superior, 2006. Apostila.

TICHY, W. F. **Should computer scientists experiment more? Computer**. 1998; 31(5):32-40.

TINTO, V. **Taking Student Retention Seriously**: Rethinking the First Year of College. NACADA Journal, 2000. Disponível em: http://www.fyecd2009.qut.edu.au/resources/SPE_VincentTinto_5Feb09.pdf. Acessado em: 23/06/2015.

WITTEN, I.H; Frank, E; Hall, M.A. **Data Mining**: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann, 3 ed, 2011.

ZHANG, Schichao; ZHANG, Chengqi; YANG, Qiang. **Data Preparation for Data Minig**. In: Applied Artificial Intelligence, 17, p.375-381, 2003.