UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS
MESTRADO PROFISSIONAL EM ENGENHARIA DA COMPUTAÇÃO E SISTEMAS

LUCIANA DA CONCEIÇÃO FERREIRA MENDES

ANALISE E PREDIÇÃO DE DESEMPENHO DE ALUNOS DA GRADUAÇÃO A PARTIR DA INSERÇÃO DE MATERIAL DE APOIO EM AMBIENTES VIRTUAIS DE APRENDIZAGEM

São Luís
2019

LUCIANA DA CONCEIÇÃO FERREIRA MENDES

ANALISE E PREDIÇÃO DE DESEMPENHO DE ALUNOS DA GRADUAÇÃO A PARTIR DA INSERÇÃO DE MATERIAL DE APOIO EM AMBIENTES VIRTUAIS DE APRENDIZAGEM

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia da Computação e Sistemas como requisito parcial para obtenção do título de mestre em Engenharia de Computação e Sistemas

Orientador: Prof. Dr. Reinaldo de Jesus da Silva.

São Luís

2019

Mendes, Luciana da Conceição Ferreira.
Análise e predição de desempenho de alunos da graduação a partir da inserção de material de apoio em ambientes virtuais de aprendizagem / Luciana da Conceição Ferreira Mendes. – São Luís, 2019.
53f.

Dissertação (Mestrado) – Programa de Pós-Graduação em Engenharia de Computação e Sistemas, Universidade Estadual do Maranhão, 2019.

Orientador: Prof. Dr. Reinaldo de Jesus da Silva.

1. Desempenho. 2. Mineração de dados. 3. Diagnóstico. 4. AVA. I. Título.

CDU 004.891.3

LUCIANA DA CONCEIÇÃO FERREIRA MENDES

ANALISE E PREDIÇÃO DE DESEMPENHO DE ALUNOS DA GRADUAÇÃO A PARTIR DA INSERÇÃO DE MATERIAL DE APOIO EM AMBIENTES VIRTUAIS DE APRENDIZAGEM

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia da Computação e Sistemas como requisito parcial para obtenção do título de mestre em Engenharia de Computação e Sistemas

Aprovada em: 28 / 02 /2019

BANCA EXAMINADORA

Prof. Dr. Reinado de Jesus da Silva (Orientador)
Universidade Estadual do Maranhão (UEMA)

Prof. Dr. Luís Carlos da Costa Fonseca (Coorientador)
Universidade Estadual do Maranhão (UEMA)

Prof. Dr. Marco Antônio Nogueira Gomes
Universidade Estadual do Maranhão (UEMA)

Prof. Dra. Aurea Celeste Costa Ribeiro
Universidade Estadual do Maranhão (UEMA)

AGRADECIMENTOS

A Deus, pela vida e oportunidade que tem dado a mim.

Aos meus pais pelo cuidado e incentivo nesta caminhada até a realização dos meus sonhos.

Aos meus professores que me ajudaram a chegar até aqui, meu orientador professor Doutor Reinaldo de Jesus da Silva e meu coorientador professor Doutor Luis Carlos Costa Fonseca.

Ao meu noivo Luis Fernandes da Silva Farias Junior, pelo apoio e suporte nos momentos que pensei em desistir.

A minha amiga Melkyanne Mendes Silva pelo companheirismo durante todo o mestrado e à minha amiga Paula Sousa da Costa que me obrigou a terminar a dissertação antes que chegue um concurso.

E a todos aqueles que, de alguma forma, contribuíram para a elaboração desta dissertação.

“A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê”.

Arthur Schopenhauer.

RESUMO

Os ambientes virtuais de aprendizagem (AVA) são cada dia mais utilizados, pois possibilitam que a educação chegue mais longe e para mais pessoas. Sua utilização não se limita apenas a educação a distância, mas também auxilia na educação presencial, agindo como uma ferramenta a mais no processo de aprendizagem. No entanto, gerenciar o processo de aprendizagem nos AVAs com qualidade de integração e acompanhamento exige cada vez mais do professor que precisa utilizar materiais atualizados e ao mesmo tempo acompanhar e medir o aprendizado dos alunos através de avaliações. Neste sentido, este trabalho apresenta um modelo preditivo de mineração de dados em um AVA, a partir da inserção de material didático por parte dos professores no ensino presencial. O objetivo foi verificar o impacto que a inserção de material tem no desempenho de alunos do ensino presencial que utilizam AVA como extensão da sala de aula. Para isso, foram realizados experimentos com conjuntos de dados distintos de turmas onde havia sido inserido material de apoio e turmas onde não havia sido inserido material de apoio. Em seguida foram aplicadas as técnicas de mineração de dados com melhor desempenho em casos de classificação de dados, Redes Bayesianas e Árvore de decisão. Posteriormente foi comparado o desempenho de cada técnica de mineração, a fim de que um modelo preditivo fosse obtido. Dessa forma, foi possível demonstrar através da análise dos dados gerados por professores e alunos, que a inserção de material didático no AVA, contribui para o melhor desempenho dos alunos da graduação presencial que usam o AVA como extensão da sala de aula.

Palavras-chave: Desempenho. Mineração de dados. Diagnóstico. AVA.

ABSTRACT

Virtual learning environments (VLEs) are increasingly being used as they enable education to reach more and more people. Its use is not limited only to distance education, but also assists in face-to-face education as a more advanced form of learning. However, managing the learning process in VLEs with quality integration and monitoring requires more and more of the teacher who needs to use updated materials and at the same time, monitor and measure students' learning through assessments. In this regard, this work presents a predictive model of data mining in a VLE, from the insertion of didactic material by the teachers in the classroom teaching. The objective was to verify the impact that the non-insertion of material has on the performance of classroom students using VLE as an extension of the classroom. Toward this, experiments were carried out with different data sets, where data mining techniques Bayesian networks and Decision Tree were applied, and the performance of each one was compared in order to obtain a more accurate model. As falows, it was possible to demonstrate, through the analysis of the data generated, that the insertion of didactic material in the VLE contributes to the better performance of the undergraduate students using VLE as an extension of the classroom.

Key words: Low performance. Data Mining. Diagnosis VLE Support material.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AVA – Ambiente Virtual de Aprendizagem

EAD – Educação a Distância

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

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

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

SUMÁRIO

1 INTRODUÇÃO

As interações que os alunos têm entre si, com os professores e com recursos educacionais são indicadores valiosos da eficácia de uma experiência de aprendizagem. O crescente uso das tecnologias de informação e comunicação permite que essas interações sejam registradas, para que as técnicas de análise ou de mineração sejam usadas para obter uma compreensão mais profunda do processo de aprendizagem e propor melhorias. Mas com a crescente variedade de ferramentas que estão sendo usadas, o monitoramento do progresso dos alunos está se tornando um desafio (ROMERO; VENTURA; GARCIA, 2008).

Um fator importante que contribui para a eficácia de uma experiência de aprendizagem é a capacidade dos professores de monitorar o conjunto processo de aprendizagem e potencialmente atuar com base nos eventos observados. Na situação ideal, um professor que monitora todos os eventos que acontecem em um ambiente de aprendizagem teria uma posição privilegiada para ajustar os parâmetros disponíveis para melhorar a experiência geral para os alunos. Mas este cenário hipotético ainda está muito longe da realidade nas instituições educacionais atuais e, pior ainda, há várias forças afastando-se desse objetivo, como por exemplo, o fato de não existir uma ferramenta de diálogo para sala de aula ou o mau planejamento da estrutura do curso (ROMERO-ZALDIVAR *et al*., 2012).

O uso de ambientes computacionais no contexto educacional coleta e armazena uma grande quantidade de dados sobre dos discentes. Esses dados são bastante amplos, e variam desde interações diversas com o sistema, registros de acesso até dados ricos em significado e relevância como as mensagens trocadas em fóruns e chats.

É interessante destacar que a simples criação de grandes bases de dados torna-se inútil sem a disponibilização de recursos apropriados para sua análise e interpretação de forma aperfeiçoada. Algumas plataformas que apoiam o ensino oferecem ferramentas simples de relatórios que possibilitam a obtenção de informações sobre as atividades desempenhadas pelos discentes.

As técnicas de mineração de dados educacionais (MDE) permitem que os padrões e aspectos sobre o contexto em que os alunos estão se tornem evidentes, assim como suas dificuldades. Dessa forma, o uso de técnicas de MDE torna possível

o apoio a atividades dos professores a partir da produção de informações adicionais que não estariam disponíveis apenas pela observação direta do professor em sala de aula (RIGO *et al.*, 2014).

Ainda segundo Rigo *et al.* (2014), a não adoção de recursos de recursos mediação digital, faz com que muitas vezes uma visão equivocada do cenário seja obtida, fazendo com que alguns recursos, como diversificação de práticas pedagógicas, melhoria do processo de ensino aprendizagem, facilidade de acesso a material e recursos não sejam aplicados para melhoria do contexto.

Dessa forma, a motivação para a realização dessa pesquisa deu-se pela possibilidade de apoiar gestores e professores no acompanhamento do desempenho dos alunos, pois a ausência desse pode acarretar problemas como retenção e evasão, que consequentemente podem gerar perdas não só para os estudantes, mas também para as Universidades e mesmo para o país.

Sendo assim, este trabalho foi dividido em duas etapas: Na primeira, procurou-se responder aos objetivos geral e específicos, mediante pesquisa descritiva e explicativa, resultando em um modelo preditivo do sistema. Na segunda etapa, analisou-se os resultados obtidos com a análise do modelo.

Como embasamento teórico, há dois pontos principais: AVA e Mineração de dados. No que tange o AVA, pretende-se demonstrar sua importância como extensão da sala de aula, mesmo para os cursos de graduação presencial, apresentando o impacto que ele pode exercer no desempenho do aluno quando usado como ferramenta auxiliar de aprendizado.

No que se refere a mineração de dados, faz-se uso da mesma para geração do modelo preditivo a ser analisado, pois essa técnica permite que uma coleção de dados que incialmente não contém informação alguma, possam ser transformados em conhecimento e consequentemente usados como ferramenta de transformação.

Ao longo dos anos, pesquisas têm sido desenvolvidas na área de mineração de dados, como Kampff *et al*. (2014) que buscam identificar perfis de alunos com risco de evasão ou reprovação e Guércio *et al.* (2014) que propõem a criação de um modelo que auxilie o professor na análise do desempenho dos alunos no decorrer da oferta de determinada disciplina, no intuito de melhorar o desempenho dos estudantes, entre outras que utilizando métodos e técnicas de mineração que auxiliem nos processos de ensino e de aprendizagem dentro do AVA.

A utilização de uma mineração de dados, integrado ao AVA para medir os impacto negativo que a ausência de acompanhamento do desempenho dos alunos e a consequente não atuação do professor diante dos eventos, o que pode acarretar não só para os estudantes e também para as Universidades, ainda é um obstáculo a ser ultrapassado, visto que se tem uma infinidade de material didático disponibilizados por instituições públicas e privadas, prontos para serem utilizados pelos seus professores e alunos, mas que não sabem ou não têm ferramentas adequadas para avaliar os impactos da utilização de AVAs.

Com base nos pressupostos evidenciados, considera-se relevante desenvolver um modelo preditivo para medir os impactos da utilização do material didático no AVA. Uma das vantagens dos modelos de mineração é a possibilidade de geração de padrões que permitem a comparação entre os resultados obtidos. Nesse aspecto, tem-se a pretensão de oferecer aos usuários não apenas a análise de desempenho, mas também a possibilidade de apoio na tomada de decisões.

Nessa perspectiva, a problemática que norteou esta dissertação foi entender: “de que forma a inserção de material de apoio em ambientes virtuais por parte do professor, pode impactar no desempenho dos alunos”. Partindo desse pressuposto, abordar-se-á na próxima seção o objetivo geral.

1.2 Objetivo geral

Desenvolver um modelo preditivo para medir os impactos no desempenho dos alunos a partir da inserção de material didático em ambientes virtuais de aprendizagem.

1.3 Objetivos específicos

- Realizar estudo bibliográfico para ampliar a compreensão sobre fatores que influenciam no desempenho de alunos em AVAs.
- Analisar e definir as técnicas de MD a serem usadas no modelo preditivo.
- Entender descrever o funcionamento da ferramenta Weka e utilizar a mesma para gerar o modelo preditivo a ser usado neste trabalho.

- Realizar um estudo de caso com experimentos para diagnosticar o baixo desempenho de alunos da graduação, de forma que seja possível analisar o efeito da não inserção de material de apoio no AVA.

1.4 Metodologia

Quanto ao ponto de vista da natureza será uma pesquisa aplicada, em virtude de gerar conhecimentos para aplicação prática dirigida à solução do problema da não inserção de material didático ou de apoio ao aprendizado nos AVAs.

Do ponto de vista da forma de abordagem do problema a pesquisa será qualitativa, visto que, considera que há uma relação dinâmica entre o mundo real e o sujeito, de forma a selecionar casos específicos para que sejam realizadas as observações, criando padrões e classificações.

Quantos aos procedimentos técnicos adotar-se-á como pesquisa bibliografia e que a busca de fontes pertinentes ao tema em questão, ou seja, material já publicado, constituído basicamente de livros, artigos de periódicos e, de informações disponibilizadas na internet.

Ainda, nesse contexto, já sob o ponto de vista dos seus objetivos, corresponde a uma pesquisa de cunho exploratório cujo desenvolvimento baseia-se em um estudo de caso com emprego da ferramenta Weka (*Waikato Environment for Knowledge Analyis*) para a MD, com dados provenientes do Sistema de Gerenciamento de Gestão Integrado da Universidade Estadual do Maranhão (SigUema).

1.5 Apresentação do trabalho

Este trabalho está dividido em sete capítulos, dispostos da seguinte forma: o primeiro capítulo apresenta a importância do uso de AVAs na educação, bem como a importância da mineração de dados na pesquisa em questão. O objetivo geral, que permitiu delimitar o estudo e facilitando a sua orientação; os objetivos específicos, permitindo e determinando as etapas que devem ser cumpridas para se alcançar o objetivo geral; e elementos teóricos que clarificam as ideias da pesquisa. O segundo capítulo define AVA descrevendo suas funcionalidades. No terceiro capitulo são

apresentados os trabalhos relacionados da dissertação; o quarto capítulo apresenta a metodologia adotada na pesquisa; já o quinto capítulo o modelo preditivo gerado; sexto capítulo exibe a pesquisa realizada, assim como os seus resultados e discussões. E, por fim, no sétimo, as conclusões e trabalhos futuros da pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta uma abordagem a respeito dos tópicos referentes a este trabalho, mostrando o que há de pesquisas sobre ambientes virtuais de aprendizagem e mineração de dados.

2.1 Ambiente virtual de aprendizagem

Com o desenvolvimento e crescimento da Web, houve grande investimento na construção de plataformas, que possibilitassem transpor a sala de aula para o meio virtual. Dessa forma, surgem os AVAs espelhados em práticas tradicionais e com os seguintes fins: oferta de curso a distância e uso de mediação tecnológica para diminuir os encontros presenciais (FELCHER; PINTO; FERREIRA, 2017).

Kalinke (2014, p.74), define AVA como: “novos espaços destinados à aprendizagem e nos quais ela pode ser favorecida. São espaços com características próprias e que permitem novas formas e encaminhamentos aos processos de ensino e aprendizagem”.

Ainda, segundo o autor, qualquer ambiente computacional utilizado para o ensino e aprendizagem é um AVA, desde que utilize a Internet para disponibilização de conteúdo, troca de informações, interação entre usuários e outras possibilidades que sejam positivas para finalidades educacionais.

Dessa forma, observa-se que o AVA é uma ferramenta com a qual o docente tem a possibilidade de disponibilizar recursos aos discentes, como textos, exercícios, links de vídeo-aulas, cronogramas entre outros. Mas em alguns casos os AVAs são usados apenas como repositor de material didático das mais diversas fontes.

É importante que haja interações em AVAs, pois a colaboração entre professor e aluno ajuda a desenvolver soluções para possíveis problemas cognitivos implícitos na interação e na comunicação que aluno e professor teriam pessoalmente.

Sendo assim, a função que o professor desempenha no contexto educacional é de extrema importância, pois ele é o mediador entre os alunos e os objetos de conhecimento. O professor também tem papel de organizador do ambiente de aprendizagem, pois mesmo que os alunos sejam estimulados a buscarem por

material de apoio, a intervenção do professor nesse processo é fundamental na avaliação da qualidade do conteúdo.

Uma das vantagens de ambientes virtuais em relação aos tradicionais é a maior exposição de material de apoio, o que não seria possível em sala de aula, já que o custo com materiais é sempre um empecilho, limitando as escolhas do professor (FELCHER; PINTO; FERREIRA, 2017).

Dentre as ferramentas comumente encontradas em um ambiente virtual de aprendizagem (AVA), podem-se citar: fóruns, envio de mensagens instantâneas, web-conferência, chat, correio eletrônico, quadro branco compartilhado, navegação *web* compartilhada, wiki, compartilhamento de arquivos e aplicativos, agenda, gestão de turmas, gestão de grupos, gestão de usuários, sistema de avaliação, questionários e banco de questões. Não sendo necessário que um ambiente virtual tenha todos essas características para ser considerado um AVA.

Um dos ambientes virtuais de aprendizagem mais difundido é o Moodle, dessa forma neste trabalho falaremos um pouco dele, além do ambiente a ser usado como estudo de caso que é o turma virtual.

2.1.1 Moodle

Moodle (*Modular Object Oriented Distance Learning*) é uma plataforma de aprendizagem projetada para oferecer educadores, administradores e aprendizes com um único sistema robusto, seguro e integrado para criar ambientes de aprendizagem personalizados (MOODLE ORG, 2018).

O moodle é um software livre de apoio à aprendizagem, pode ser instalado em várias plataformas que consigam executar a linguagem php tais como Unix, Linux, Windows. MAC OS. Como base de dados podem ser utilizados MySQL, PostgreSQL, Oracle, Access, Interbase ou ODBC (MOODLE ORG, 2018).

Seu desenvolvimento é de forma colaborativa por uma comunidade virtual, a qual reúne programadores, designers, administradores, professores e usuários do mundo inteiro e está disponível em diversos idiomas.

A plataforma vem sendo utilizada não só como ambiente de suporte à educação a distância, mas também como apoio a cursos presenciais, formação de grupos de estudo, treinamento de professores.

Algumas funcionalidades como, links para as atividades disponibilizados para cada curso, o conteúdo teórico de cada tópico do curso, o calendário que evidencia as datas de entrega de atividades e marca datas dos próximos eventos, últimas notícias, chat e fórum são disponibilizadas pela plataforma.

Outra parte importante das funcionalidades são os cadastros de administradores, professores, monitores e alunos. A plataforma Moodle traz como diferencial outras funcionalidades que abrangem desde o cadastro de professores, cursos, currículos, turmas, alunos, até a geração de relatórios (MOODLE ORG, 2018).

Para os professores é possível trabalhar com documentos padronizados, e como recursos de revisão, impressão e envio das tarefas via e-mail ou postagem na plataforma, sem restrições que este material seja arquivo de texto, fotografias, gráficos, diagramas, áudio, vídeo e, podendo ainda, acompanhar a evolução da aprendizagem dos seus alunos fornecendo-lhe, a cada etapa do curso suporte fazendo comentários e emitindo orientações que os ajudarão num melhor proveito das tarefas, podendo ainda, contar com o recurso de elaborar avaliações e atividades (MOODLE ORG, 2018).

O Moodle mantém em sua base de dados logs detalhados de todas as atividades que os alunos desenvolvem, constando a trajetória dos materiais que os alunos acessaram (RICE, 2006),

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; VENTURA; GARCIA, 2008).

2.1.2 Turma virtual

A Turma Virtual é uma ferramenta de ensino complementar colocada à disposição dos docentes e discentes. Ela é um espaço construído para ajudar no aprendizado dos discentes, criando uma extensão da sala de aula no Sistema Integrado de Gestão Acadêmica da Uema (SigUema). Criada e desenvolvida pela Universidade Federal do Rio Grande do Norte, a ferramenta encontra-se nos Portais

do Docente e do Discente, permitindo o intercâmbio virtual de informações entre discentes e docentes de uma turma.

Na turma virtual, o docente pode cadastrar o plano de curso, o cronograma de aulas, as avaliações, referências bibliográficas, conteúdo programático, criar fórum e chat para a turma, lançar a frequência dos discentes, gerenciar grupos, imprimir diário de turma e lista de presença, cadastrar materiais para disponibilizar para os discentes, cadastrar atividades e questionários para que os discentes resolvam visualizar estatísticas de notas, alunos e acessos etc. Além dessas e de outras funcionalidades, o docente também pode, na Turma Virtual, efetuar o lançamento de notas ou conceitos e fechar a turma, como observa-se na Figura 1.

Os discentes, de modo similar aos docentes, também têm acesso às turmas virtuais, podendo acessar todo o material disponibilizado pelos docentes, além de visualizar suas notas, acessar seu histórico, imprimir comprovantes de vínculo e matrícula, solicitar trancamento de curso, fazer sua matricula online, contatar o professor entre outras funcionalidades.

Figura 1- Funcionalidades gerais da turma virtual na visão docente

Fonte: Turma Virtual, SigUema.

A base de dados do Turma Virtual registra todas as interações referentes ao uso do ambiente através de logs detalhados à medida que o usuário o utiliza. As interações referentes à visualização das disciplinas, notas e envio dos exercícios, por

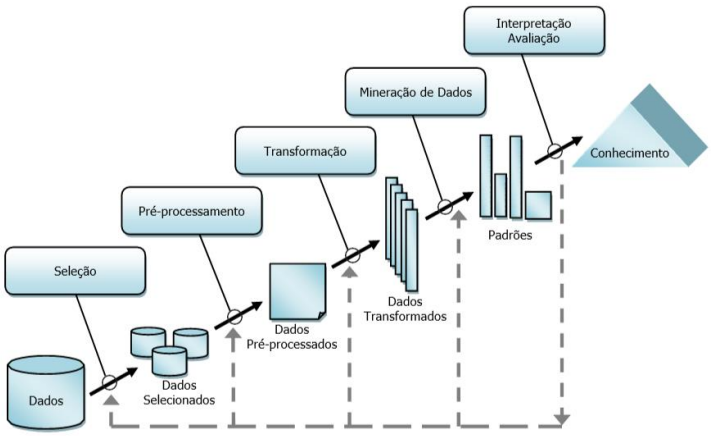
exemplo, são armazenadas em que tabelas separadas, também chamadas de tabelas de log.

2.2 Mineração de dados e descoberta de conhecimento

Com a globalização, o uso de sistemas informatizados tornou-se praticamente obrigatório. A cada dia cresce a quantidade de dados de diversas áreas, gerados e armazenados em bases digitais. Bases de comércio, saúde, educação, ciência entre outras são muitos valorizadas, dessa forma, muitos esforços tem sido empreendido para analisá-las, pois muitos padrões novos e relevantes podem ser identificados a partir de dados contidos nessas bases.

Sendo assim é praticamente impossível para o ser humano analisar os enormes conjuntos de informações e encontrar padrões significativos neles, que não eram previamente conhecidos e que são potencialmente uteis. Dessa forma minerar dados pode ajudar encontrar informações que farão toda a diferença. A MD pode ser definida como uma etapa principal de um processo mais abrangente conhecido como descoberta de conhecimento em bases de dados ou em inglês *Knowledge Discovery in Databases* (KDD). Em KDD verifica-se ainda a inclusão de mais duas grandes etapas (Figura 2), são elas: pré-processamento de dados que nada mais é que preparação de dados, abrangendo mecanismos para captação, organização e tratamento dos dados e pós-processamento dos resultados obtidos na MD.

Figura 2 - Etapas para descoberta de conhecimento.



Fonte: Kampff *et al*. (2014).

Sendo assim, observa-se que MD possui uma definição abrangente, na qual KDD é descrito como um processo geral de descoberta de conhecimento

composto pelas três grandes etapas mencionadas. Os padrões mencionados devem ser novos, compreensíveis e úteis, ou seja, deverão trazer algum benefício novo que possa ser compreendido rapidamente pelo usuário para uma possível tomada de decisão.

Entretanto, há uma ausência de consenso entre os autores sobre uma definição para o termo MD, dificultando o estabelecimento de uma definição única. Há autores que consideram *Data Mining* como sinônimo de KDD, referindo-se a ambas como uma disciplina que objetiva a extração automática de padrões uteis e implícitos de grandes coleções de dados, mas neste trabalho não faremos a distinção entre os termos (DIAS, 2008).

2.3 Mineração de dados educacionais

A Mineração de dados educacional (MDE) pode ser definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais e de acordo com Baker *et al*. (2011 *apud* COUTINHO; KAESTNER; NORONHA, 2014, p. 1) tem o seguinte objetivo.

> A área de Mineração de Dados Educacionais tem como objetivo a aplicação de técnicas computacionais para o tratamento das grandes massas de dados geradas em Ambientes Virtuais de Aprendizagem (AVA). A EDM tem como base proporcionar a descoberta de conhecimentos que sejam relevantes, únicos e válidos, bem como: a identificação de padrões entre os alunos; a análise preditiva de desempenho; e a identificação de perfis, de forma a auxiliar a gestão qualitativa da EAD.

Sendo assim, é possível compreender de forma mais eficaz e adequada os educandos, como eles aprendem, o papel do contexto na qual a aprendizagem ocorre, além de outros fatores que influenciam a aprendizagem. Por exemplo, é possível identificar em que situação um tipo de abordagem instrucional (aprendizagem individual ou colaborativa, por exemplo) proporciona melhores benefícios educacionais ao educando. Também é possível verificar se o educando está desmotivado ou confuso e, assim, personalizar o ambiente e os métodos de ensino para oferecer melhores condições de aprendizagem (BAKER; ISOTANI; CARVALHO, 2011).

A área de Mineração de Dados Educacionais (MDE) procura desenvolver ou adaptar métodos e algoritmos de mineração existentes, de tal forma que os mesmos sirvam para compreender melhor os dados em contextos educacionais que são produzidos principalmente por discentes. Os métodos de MDE visam entender melhor o estudante no seu processo de aprendizagem, analisando-se sua interação com os ambientes de aprendizagem (DIAS *et al*., 2014).

Ainda segundo Baker, Isotani e Carvalho (2011), muitos métodos utilizados em MDE, são originalmente da área de MD. Dessa forma, muitos métodos necessitam de alguma modificação para serem usados em MDE por conta da necessidade de se considerar a hierarquia em vários níveis da informação. Além do que existe uma falta de independência estatística nos tipos de dados encontrados ao coletar informações em ambientes educacionais.

Dessa forma, surge a necessidade de adequação dos algoritmos de MD existentes para lidar com especificidades inerentes aos dados educacionais, como a hierarquia dos dados. Por outro lado, há uma necessidade da criação de ambientes que possam coletar melhor esses dados, para que o processo de MD seja facilitado para cada um dos atores envolvidos, principalmente o professor.

2.3.1 Técnicas de mineração de dados educacionais

Neste tópico são apresentadas as técnicas de MDE utilizadas para a análise de aplicações educacionais segundo a taxonomia de Baker, Isotani e Carvalho (2011), as quais são:

- Predição (Classificação e Regressão);
- Agrupamento;
- Mineração de Relações (Mineração de Regras de Associação, Mineração de Correlações, Mineração de Padrões Sequenciais e Mineração de Causas;
- Destilação de dados para facilitar decisões humanas;
- Descobertas com modelos.

Dos métodos destacados na taxonomia acima, serão destacados: Predição, Agrupamento e Mineração de Relações, por conta da relevância para este trabalho. Os demais serão vistos de maneira mais sucinta

2.3.1.1 Predição

Na tarefa de predição, a meta é desenvolver modelos que façam inferência sobre aspectos específicos dos dados (variáveis preditivas) por meio da análise e associação dos diversos aspectos encontrados nos dados (variáveis preditoras). Um modelo preditivo pode ser entendido como uma função f(X;b)˜Y, onde X é um conjunto de variáveis preditoras, b são parâmetros desconhecidos e Y é a variável preditiva Y. Em outras palavras, desejasse estimar o valor de Y por meio da descoberta de b utilizando-se X. No processo de predição, é fundamental que boa parte dos dados sejam rotulados manualmente, ou seja, a aprendizagem do modelo ocorrerá de forma supervisionada e dar-se-á utilizando um conjunto de treinamento com valores previamente conhecidos de Y (COSTA *et al.*, 2013).

De acordo com Baker, Isotani e Carvalho (2011) há dois benefícios relacionados à utilização da predição em EDM. Primeiro, os métodos de predição podem ser utilizados para estudar quais aspectos de um modelo são importantes para predição. Esta estratégia é frequentemente utilizada em pesquisas que tentam, de forma direta, predizer os benefícios educacionais de determinadas técnicas e ferramentas para um conjunto de estudantes, isso sem considerar os fatores intermediários. Segundo os métodos de predição auxiliam a predizer o valor das variáveis utilizadas em um modelo. O intuito de utilizar essa abordagem é verificar quais dados são mais importantes para o modelo, pois analisar todos os dados de um grande banco de dados para gerar um modelo é inviável, do ponto de vista financeiro e de tempo.

Sendo assim, o modelo pode ser construído utilizando parte dos dados e então ser aplicado para modelar dados mais extensos. Esse tipo de técnica pode auxiliar no desenvolvimento e uso de atividades instrucionais, pois consegue-se estimar os benefícios educacionais antes mesmo da atividade ser aplicada aos alunos (BAKER*;* ISOTANI; CARVALHO, 2011).

Em EDM, dois tipos de técnicas são usadas com mais frequência: classificação e regressão. Na classificação a variável preditiva é binária ou categórica e na regressão a variável preditiva é contínua. Em ambos os casos, as variáveis preditoras podem ser categóricas ou contínuas (COSTA *et al.*, 2013).

Na classificação, os algoritmos mais utilizados são árvores de decisão e máquina de vetores de suporte. Em relação à regressão, os algoritmos mais populares são regressão linear, redes neurais e máquinas de vetores de suporte para regressão.

Na sequência são apresentados alguns algoritmos de predição, descrevendo o método de indução e classificação de árvore de decisão e o de redes bayesianas, além do modelo de regressão, apresentado através da abordagem de regressão linear.

2.3.1.2 Árvore de decisão

Árvores de decisão são modelos estatísticos que utilizam treinamento supervisionado para classificação e predição dos dados. Dessa forma, no conjunto de treinamento as variáveis preditivas Y são conhecidas. Uma árvore de decisão possui uma estrutura de árvore, onde cada nó interno (não-folha), pode ser entendido como um atributo de teste, e cada nó-folha (nó-terminal) possui um rótulo de classe. O nó de nível mais elevado numa árvore de decisão é chamado de nó-raiz (COSTA *et al.*, 2013).

A árvore de decisão classifica a instancia de acordo com o caminho que satisfaz as condições desde o nó raiz até o nó folha, baseando-se nos parâmetros do modelo aprendido. Ao final de todo processo, a instancia é rotulada de acordo com o nó folha (QUINLAN, 1993).

O algoritmo para a construção de uma árvore de decisão age de maneira recursiva na fase de aprendizagem, subdividindo os dados até que as folhas sejam classes puras ou que exista um critério de parada especificado, como o número de casos enquadrados. A representação da árvore de decisão pode ser gráfica ou textual, podendo ser traduzida em regras do tipo SE <condição> ENTÃO <classificação> (KAMPFF *et al*., 2014).

A implementação mais clássica de árvore de decisão refere-se ao algoritmo ID3 (*Indutive Decision Tree*), sendo o mesmo a base para implementações de outros algoritmos de árvore de decisão, como o C4.5 (QUILAN, 1993) (J48 na ferramenta Weka), que é uma versão otimizada do ID3. Algumas das principais vantagens do C4.5 em relação ao ID3 são apresentadas são: o C4.5 trabalha com variáveis discretas e contínuas (o ID3 só trabalha com dados discretos), através de um

processo de discretização interna. Ele realiza a poda da árvore, que consiste em aumentar a capacidade de generalização da árvore de decisão, evitando, assim, que ocorra o overfit (superestima de um conjunto de dados). O overfit é influenciado principalmente pela pouca quantidade de instâncias nos dados e quando há ruído nos dados (SILVA *et al.*, 2015).

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

O algoritmo CART (*Classification and Regression Trees*) consiste de uma técnica não paramétrica que induz tanto árvores de classificação quanto árvores de regressão, dependendo se o atributo é nominal (classificação) ou contínuo (regressão). 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. A utilização da pós-poda no CART é vista pelos autores supracitados como extremamente eficiente, pois, segundo eles, produz árvores mais simples, precisas e com ótima capacidade de generalização (SILVA *et al.*, 2015).

O algoritmo BFTree foi proposto por Shi (2007) para indução de árvores de decisão binárias, baseado na 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.

2.3.1.3 Redes bayesianas

Os métodos probabilísticos bayesianos baseados no Teorema de Bayes, são utilizados para resolverem problemas que envolvem tarefas de predição de classificação, casos em que as informações disponíveis são imprecisas ou incompletas.

A noção fundamental da Estatística Bayesiana é a Probabilidade Condicional, definida por probabilidades de associação de classe, como a probabilidade de que uma determinada tupla pertença a uma classe particular P(H|E)

no qual H é a hipótese e E é a evidência. Para computar a probabilidade de uma hipótese H, é necessário levar em consideração o valor da evidência E. Quando não existir evidências, tem-se a probabilidade incondicional P(H).

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:

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

Onde:

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

O cálculo é feito a partir da equação abaixo, dada por: $P(H \mid E)= P(H \cap E) /P(E)$, onde o numerador é a probabilidade de H e E ocorrerem simultaneamente e o denominador é a probabilidade de ocorrer E.

A formulação do teorema de Bayes envolve estas probabilidades. A Equação seguinte apresenta o teorema formulado por Bayes: $P(H \mid E)=P(E \mid H)P(H) /P(E)$.

A estatística bayesiana pode ser usada para classificação de uma forma um tanto simples, sendo chamada de classificadores Bayesianos, cujo objetivo é a descrição e identificação de classes e também a previsão de classes de objetos que não foram classificados (DIAS *et al.*, 2014).

O classificador bayesiano mais simples é conhecido como Naïve Bayes e considera a hipótese de que todas as variáveis são independentes. Ele apresenta bom desempenho em vários domínios e é considerado robusto à presença de ruídos e atributos irrelevantes, contudo, a suposição de independência total entre os atributos pode torna-se um pouco rígida, tornando o classificador incapaz de tratar problemas de classificação reais onde as variáveis geralmente apresentam interdependência.

As Redes Bayesianas são modelos probabilísticos representados por grafos acíclicos e direcionados, mostrando as relações de causalidade entre as variáveis de um problema. As Redes Bayesianas possuem uma parte gráfica que é qualitativa, e uma parte quantitativa que são as tabelas com a distribuição de probabilidades das variáveis (DIAS *et al.*, 2014).

A construção do modelo probabilístico de redes bayesianas envolve duas etapas: a primeira é a construção do grafo e a segunda é a avaliação dos valores de probabilidades nas tabelas associadas a cada nodo, chamadas de Tabelas de Probabilidade Condicional (TPC); essas etapas podem ser realizadas utilizando o conhecimento de especialistas do domínio, utilizando bases de dados, ou pela junção das duas abordagens.

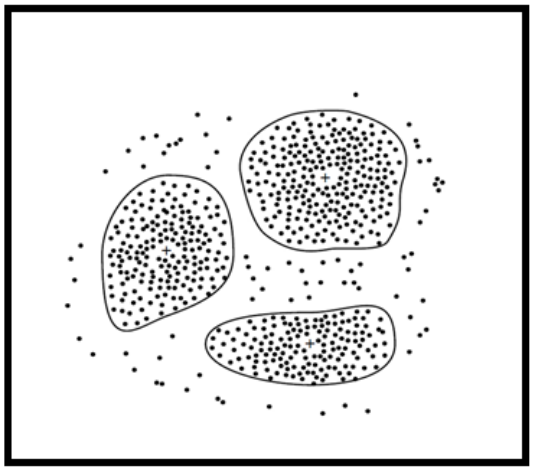
2.3.1.4 Agrupamento

A técnica de agrupamento objetiva identificar e aproximar os registros similares. Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém diferentes dos outros registros nos demais agrupamentos. Esta técnica não necessita que os registros sejam previamente categorizados (aprendizado não-supervisionado). Além disso, ela não tem a pretensão de classificar, estimar ou predizer o valor de uma variável, ela apenas identifica os grupos de dados similares, conforme mostra a Figura 3. Exemplos:

- Segmentação de mercado para um nicho de produtos;
- Separação de comportamentos suspeitos para auditoria

As aplicações das tarefas de agrupamento são as mais variadas possíveis: pesquisa de mercado, reconhecimento de padrões, processamento de imagens, análise de dados, segmentação de mercado, taxonomia de plantas e animais, pesquisas geográficas, classificação de documentos da Web, detecção de comportamentos atípicos (fraudes), entre outras. Normalmente a tarefa de agrupamento é combinada com outras tarefas, além de serem usadas na fase de preparação dos dados (CAMILO; SILVA, 2009).

Figura 3 - Registro agrupado em três clusters.



Fonte: Camilo e Silva (2009)

2.3.1.5 Mineração de relações

Em mineração de relações, a meta é descobrir se existe relações entre variáveis em bancos de dados. Esta tarefa pode envolver a tentativa de aprender quais variáveis são mais importantes, ou mesmo envolver as relações entre quaisquer variáveis presentes nos dados. Para identificar essas relações, existem quatro tipos de mineração: regras de associação, correlações, sequências e causas fortemente associadas com uma variável específica, previamente conhecida.

Na mineração de regras de associação, procura-se gerar/identificar regras do tipo se-então (if-then) que permitam associar o valor observado de uma variável ao valor de uma outra variável. Dessa forma, caso uma condição seja verdadeira (por exemplo, variável Y possui valor 1) e uma regra associe essa condição ao valor de uma outra variável X, então podemos inferir o valor desta variável X. Por exemplo, ao analisar um conjunto de dados seria possível identificar uma regra que faz a associação entre a variável “objetivo do educando”, uma variável binária que pode ter os valores alcançado ou não alcançado, e uma outra variável binária “pedir ajuda ao professor” que pode ter os valores sim ou não. Neste contexto, se o educando tem como objetivo aprender geometria, mas está com dificuldade, isto é, a variável meta do educando tem valor não alcançado, então é provável que ele peça ajuda do professor, ou seja, a variável pedir ajuda ao professor tem valor positivo (DIAS *et al.*, 2014).

Em mineração de correlações, o objetivo é achar correlações lineares (positivas ou negativas) entre variáveis. Por exemplo, ao analisar um conjunto de dados, seria possível identificar a existência de uma correlação positiva entre uma variável que indica a quantidade de tempo que um educando passa externalizando comportamentos que não estão relacionados as tarefas passadas pelo professor (por exemplo, conversas paralelas, brincadeiras e outras perturbações que ocorrem em sala de aula) e a nota que este educando recebe na próxima prova (DIAS *et al.*, 2014).

Em mineração de sequências, a meta principal é achar a associação temporal entre eventos e o impacto destes eventos no valor de uma variável. Neste caso, é possível determinar qual trajetória de atos e ações de um educando pode, eventualmente, levar a uma aprendizagem efetiva. Dessa forma, é possível criar um conjunto de atividades instrucionais que podem melhorar a qualidade do ensino fazendo com que os educandos externalizem ações que vão ajudá-los a construir seu conhecimento e desenvolver as habilidades necessárias para trabalhar com o conteúdo apresentado pelo professor.

Em mineração de causas, desenvolve-se algoritmos para verificar se um evento causa outro evento por meio da análise dos padrões de covariância. Por exemplo, se considerarmos o exemplo anterior onde um educando externaliza comportamentos inadequados que não contribuem para resolver a tarefa dada pelo professor, o educando, em muitos casos, recebe uma nota ruim na prova final. Nesta situação, o comportamento do educando pode ser a causa do mesmo não aprender e, assim, resulta em uma performance ruim na prova. Contudo, pode ser que o educando externalize tal comportamento inadequado devido à dificuldade em aprender, e portanto, a causa da performance ruim na prova não é o comportamento em si, mas sim a dificuldade de aprendizagem do educando. Analisando o padrão de covariância, a mineração de causa pode auxiliar na inferência de qual evento foi a causa do outro (DIAS *et al.*, 2014).

2.3.1.6 Destilação de dados para facilitar decisões humanas

Na área de destilação de dados para facilitar decisões humanas, são realizadas pesquisas que tem como objetivo apresentar dados complexos de forma a facilitar sua compreensão e expor suas características mais importantes.

Os métodos dessa subárea da MDE facilitam a visualização da informação contida nos dados educacionais coletados por softwares educacionais. Estes métodos limpam os dados para auxiliar as pessoas na identificação de padrões. Em diversas ocasiões, esses padrões são previamente conhecidos, mas são difíceis de serem visualizados e/ou descritos formalmente.

2.3.1.7 Descoberta com modelos

Em descoberta com modelos, parte-se de um modelo gerado por um método de predição, tal como classificação, ou por um método de agrupamento, ou ainda manualmente, por meio de engenharia de conhecimento. Em seguida, esse modelo é utilizado como componente, ou ponto de partida, em outra análise com técnicas de predição ou mineração de relações (COSTA *et al.*, 2013).

2.3.2 Ferramentas de mineração de dados educacional

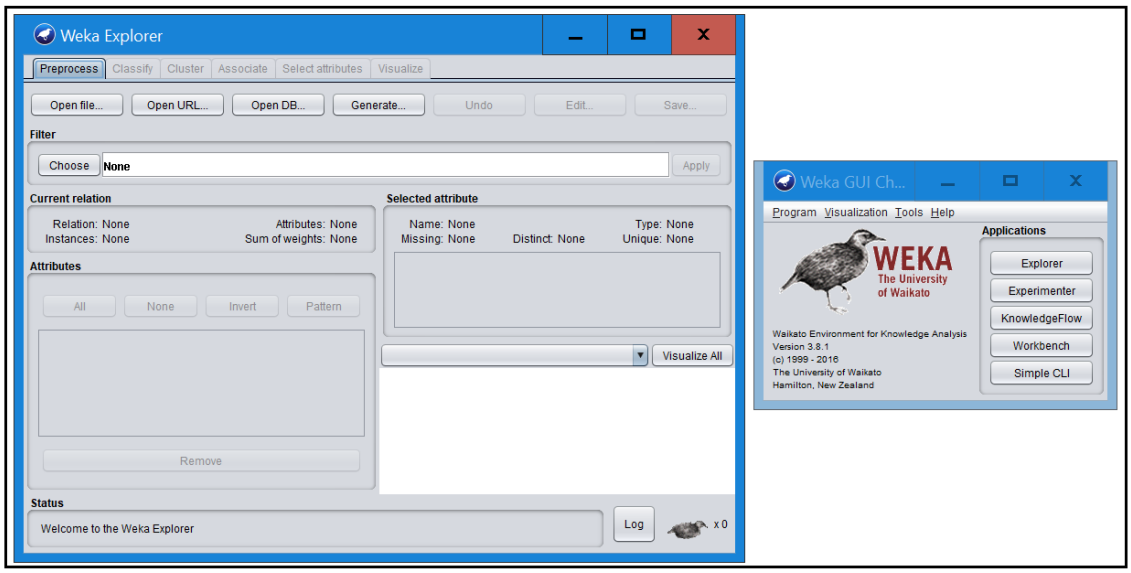
Existem diversas ferramentas de mineração, comerciais e acadêmicas, disponíveis que proveem algoritmos de mineração, algoritmos de pré-processamento, técnicas de visualização, entre outros, como: DBMiner, Clementine, IBM Intelligent Miner, Weka. Apesar dos esforços da comunidade de MDE em propor e construir ferramentas de mineração que levem em conta as particularidades da mineração no contexto educacional, duas dessas ferramentas são muito utilizadas na literatura: Weka e Rapidminer. Mas por motivo de aplicabilidade neste estudo, apenas a ferramenta Weka será mostrada mais detalhadamente.

2.3.2.1 Weka

Weka é uma coleção de algoritmos de aprendizagem de máquina e ferramentas de pré-processamento. É uma ferramenta de código aberto e foi desenvolvido na Universidade de Waikato na Nova Zelândia. Weka possui uma variedade de algoritmos de aprendizagem, que incluem ferramentas de pré-processamento. Além disso, oferece suporte a todo processo de mineração, que inclui suporte a preparação dos dados de entrada, avaliação estatística da aprendizagem,

visualização dos dados de entrada e os resultados (HALL *et al*., 2009). Todas as funcionalidades disponíveis podem ser acessadas através de uma interface comum, apresentada na Figura

Figura 4 - Interface gráfica inicial do Weka e a Interface gráfica Explorer



Fonte: Mendes (2018).

A primeira opção da ferramenta é o Explorer. Esta é a opção mais simples para se utilizar. Ela oferece uma interface que possibilita ao usuário acessar funcionalidades oferecidas pelo Weka e que podem ser acessadas por meio da interface gráfica e suas opções. O usuário pode, por exemplo, escolher ler um arquivo ARFF e construir uma modelo, utilizando algum dos algoritmos disponíveis. A interface possibilita utilizar quaisquer um dos algoritmos implementados pela ferramenta, apresentando dicas e os resultados de cada aprendizagem realizada com a base de dados escolhida (HALL *et al*., 2009).

Outra opção disponibilizada na interface gráfica é o *KnowledgeFlow*. Esta opção oferece uma interface gráfica que permite ao usuário construir um fluxo para o processamento dos seus dados.

O Weka oferece alguns algoritmos incrementais e podem ser usados para processar um conjunto de dados muito grande. Essa interface permite que o usuário escolha entre caixas que representam esses algoritmos, arraste-os e estabeleça a configuração desejada. Isso permite que o usuário construa um fluxo para o processamento do conjunto de dados escolhido por meio da conexão desses

componentes. Esses componentes podem estar representando as fontes de dados, ferramentas de pré-processamento, algoritmos de aprendizagem, métodos de avaliação e visualização (HALL *et al*., 2009).

A terceira opção disponibilizada pelo Weka é o *Experimenter*. Essa opção oferece uma interface gráfica que possibilita aos usuários um auxílio em uma questão prática: descobrir quais métodos e parâmetros funcionam melhor para um determinado problema. O usuário pode fazer isso de forma interativa ao aplicar algum algoritmo em sua base de dados. Entretanto, essa interface permite ao usuário automatizar esse processo, tornando mais fácil executar diferentes algoritmos e filtros com diferentes parâmetros (HALL *et al*., 2009).

A última opção oferece a funcionalidade mais básica da ferramenta, onde o Weka pode ser acessado utilizando linhas de comando. Essa opção possibilita o acesso a todas as funcionalidades do sistema. Além disso, o Weka oferece a API Java, que permite a construção de aplicações que utilizem todas as funcionalidades disponibilizadas pela ferramenta. A API também oferece suporte a construção do arquivo ARFF que é o formato específico aceito pelos algoritmos implementados pela ferramenta (HALL *et al*., 2009).

A escolha da ferramenta se deu não só pelas funcionalidades anteriormente citadas ou pelo fato de ser uma ferramenta gratuita, mas principalmente por ser ela a mais usada nos trabalhos utilizados como base para a pesquisa em questão.

3 TRABALHOS RELACIONADOS

Na literatura encontram-se algumas iniciativas de MD em AVA's, sendo, que visam analisar o desempenho de alunos e consequentemente descobrir padrões de comportamento a partir das interações dos alunos nesses ambientes. Dentre elas, as que mais se aproximam da metodologia proposta neste trabalho são as que seguem.

Silva *et al.* (2015) apresentam o desenvolvimento de um modelo preditivo de MD em um AVA, a partir das interações de alunos em fóruns de discussão. O objetivo era realizar o diagnóstico 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 o autor realizou experimentos com conjuntos de dados distintos, levando em consideração as seguintes características: total de fóruns que o aluno participou, total de postagens em todos os fóruns, média de postagens por fórum, resultado final do aluno na disciplina (atributo classe), aprovado por média, aprovado por final, reprovado por média, reprovado por final. A técnica de MD foi aplicada através de cinco algoritmos de classificação: J48, BFTree, SimpleCart, Bayesianos e BayesNet, sendo comparado o desempenho de cada um, a fim de que um modelo com melhor desempenho fosse obtido. O algoritmo J48 alcançou melhor desempenho (73,96%) sendo a mais indicada, dentre as testadas, para a geração de um diagnóstico mais preciso das tendências tratadas no trabalho.

Através de técnicas de MD, Kampff *et al.* (2014) buscam identificar perfis de alunos com risco de evasão ou reprovação, visando à geração de alertas para sensibilizar o professor sobre possíveis problemas. Segundo os autores, nos resultados obtidos, há evidências de que o índice de evasão nas turmas com os alertas tenha sido significativamente inferior ao índice de evasão observado na amostra de dados históricos.

Guércio *et al.* (2014) propõem a criação de um modelo que auxilie o professor na análise do comportamento dos alunos no decorrer da oferta de determinada disciplina, no intuito de melhorar o desempenho dos estudantes.

Para tanto, foram criados três conjuntos de treinamento, divididos por tempo de acordo com o período de duração das disciplinas. No caso, os conjuntos

foram divididos nos períodos de 6, 12 e 18 semanas após o início do curso, para avaliar o comportamento dos alunos durante esses intervalos de tempo.

Para cada um dos períodos de tempo descritos, foram analisados o total de acessos a plataforma e a interação ocorrida nos fóruns. A análise de acessos de alunos foi realizada selecionando o total de acessos de alunos na disciplina assim como o total de acessos a fóruns, recursos e atividades. Para a análise da interação entre alunos e tutores, foram selecionadas a quantidade de postagens criados por cada aluno e o total dessas postagens que foram respondidas por alunos e tutores. O quadro 1 descreve de maneira simplificada as dimensões e os atributos utilizados para avaliar o desempenho dos alunos.

Segundo os autores, a acurácia média encontrada foi de 73% na classificação do desempenho e foi possível observar a possibilidade de transformar os dados armazenados na base de dados da plataforma Moodle em conhecimento, gerando regras muito úteis para o apoio a tomada de decisões.

Marques (2014) propôs uma metodologia para MDE 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 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, a autora utilizou o método de MD denominado predição, por meio da classificação de exemplos, que induz 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.

Marques (2014) 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.

Segundo a autora, durante o estudo, foi possível observar a relação existente entre alguns atributos, demonstrando a importância da interação constante entre aluno e professor (MARQUES, 2014).

Santana, Maciel e Rodrigues (2014) tiveram como objetivo realizar a avaliação da dimensão perfil de uso no ambiente Moodle. Foram utilizados dados de um curso ofertado na modalidade semipresencial extraídos do banco de dados do AVA Moodle. Foram utilizados sete algoritmos para analisar o desempenho do perfil, onde o J48 obteve o melhor desempenho, alcançando 74% de acurácia.

Brito *et al.* (2014) propõe a utilização de técnicas de MD para tentar predizer o desempenho dos alunos no primeiro período do curso de Ciência da Computação da UFPB, através das suas notas de ingresso no vestibular.

O estudo buscou encontrar relação entre a nota de ingresso de estudantes e o seu desempenho nas disciplinas de Cálculo Diferencial e Integral I, Física Aplicada à Computação I, Cálculo Vetorial e Geometria Analítica do primeiro período do curso de Ciência da Computação da UFPB. Através da ferramenta *Weka*, obteve-se precisão superior a 70%, utilizando um conjunto de três atributos de entrada: Média Geral, Média de Matemática e a Média de Física obtidas no processo seletivo para entrada na UFPB. O autor aponta que se tem conhecimento da existência de outras variáveis que podem influenciar o desempenho do aluno no primeiro período do curso, porém afirma que estas são muitas vezes subjetivas e difíceis de serem recuperadas, como motivação do aluno no curso, taxa de aprovação da turma, situação socioeconômica, entre outras.

Nesse sentido, o trabalho aqui proposto busca ampliar as contribuições dos trabalhos correlatos, uma vez que, insere no contexto uma variável nunca antes utilizada na análise do desempenho de discentes, buscando assim, elucidar o problema proposto.

A análise a ser apresentada será obtida a partir dos resultados dos experimentos realizados no Weka, permitindo visualizar as tendências a baixo desempenho e mesmo à reprovação.

O diferencial deste trabalho é análise de duas varáveis, ou seja, professor e aluno. Além da abordagem ser do ponto de vista do professor analisando as atividades desenvolvidas por ele, não olhando só pelo ponto de vista das atividades de responsabilidade do aluno. O quadro 1 mostra a comparação entre os trabalhos relacionados e o trabalho aqui proposto.

Quadro 3- Comparação entre os trabalhos relacionados

AUTORES	TÉCNICA DE MINERAÇÃO DE DADOS	CARACTERÍSTICAS ANALISADAS	FONTE DE DADOS
Kampff *et al*. (2014)	Algoritmos RuleLearner e DecisionTree da ferramenta de mineração RapidMiner	Indicadores de baixa frequência de acesso aos fóruns, baixo rendimento nas atividades e baixo acesso ao AVA.	AVA e informações baseadas na mineração de dados educacionais
Silva *et al*. (2015)	Árvore de decisão Bayesianos	Nota média do aluno nos fóruns, total de fóruns que o aluno participou, total de postagens em todos os fóruns, média de postagens por fórum, resultado final do aluno na disciplina (atributo classe) [Aprovado Por Média, Aprovado Por Final, Reprovado Por Média, Reprovado Por Final.	Moodle
Marques (2014)	Árvore de decisão (J48)	Dados dos alunos, interação com o AVA, pontuação em atividades avaliativas, cumprimento de forma satisfatória das atividades ou situações problema, etc. Indicadores gerais de quantidade de acessos aos recursos do AVA.	AVA Senai
Guércio *et al*. (2014)	Random Tree, Random Forest e J48	Total de acessos a atividades, total de acessos a fóruns, total de acessos a recursos, nota final na disciplina, postagens realizadas, postagens realizadas respondidas pelos alunos, pelos tutores e professores.	Moodle (UFJF)
Brito *et al*. (2014)	SMO	Nota dos alunos em duas disciplinas entre os anos de 2006 e 2013.	Amostra de alunos presenciais UFPB
Santana, Maciel e Rodrigues (2014)	Árvore de Decisão (J48)	Número total de acesso ao fórum, número total de interações com as vídeo-aulas, número total de interações com o material da disciplina (Caderno), número total de interações com as apresentações em Slides. Tempo médio de acesso no ambiente	Moodle
Mendes (2018)	Árvores de Decisão (j48) e Redes Bayesianas	Nota média do aluno nas avaliações, quantidade de frequência do aluno, quantidade downloads de arquivos postado pelo docente, quantidade de material postado pelo docente, total de postagem do professor, resultado final do aluno na disciplina.	Turma Virtual

Fonte: Mendes (2018).

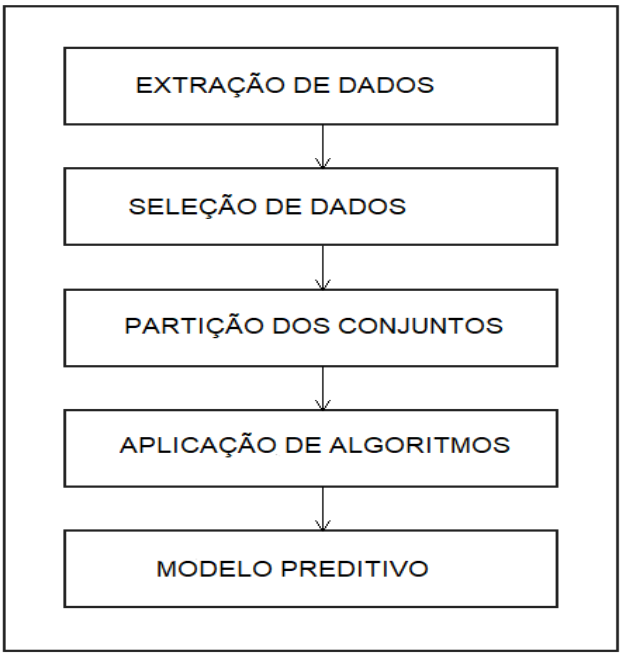
Assim como o trabalho desenvolvido nessa pesquisa, os trabalhos citados analisam o desempenho de alunos e consequentemente buscam estabelecer padrões de comportamento a partir das interações dos alunos nesses ambientes e consequentemente ajudar professores e gestores educacionais na detecção desses problemas e no apoio aos alunos, no sentido de escolher as estratégias de aprendizagem que mais podem gerar resultados positivos para os mesmos.

4 ARQUITETURA DO MODELO PREDITIVO

A seguir é descrita a arquitetura utilizada para geração do modelo preditivo do trabalho em questão.

A figura 5 mostra a arquitetura do modelo computacional proposto neste trabalho baseados em Silva *et al*. (2015), e os parágrafos seguintes fazem sua descrição:

Figura 5 - Arquitetura do modelo preditivo para diagnóstico de desempenho



Fonte: Adaptado de Silva *et al*. (2015).

- **Extração de dados:** Nessa fase os dados serão extraídos da base do SigUema. Os dados selecionados para extração foram os que se encaixavam no domínio da pesquisa, ou seja, dados de alunos e professores cujas turmas havia material didático inserido e dados da mesma turma ofertada em semestre anterior sem a inserção de material didático.
- **Seleção dos dados:** A fase de pré-processamento constitui-se das etapas de seleção de dados e aplicação de filtros na arquitetura proposta. Em seguida, foi feita a utilização de algoritmos de MD para identificar, a partir das postagens de material de apoio na turma virtual,

se um aluno tem tendência a baixo desempenho ou aprovação. Os dados serão obtidos pela seleção realizada a partir das tabelas e atributos da base de dados do turma virtual que registram as inserções de material de apoio. Para a criação da tabela de sumarização, referente ao conjunto de dados originais, novos atributos serão adicionados ou transformados, enquanto que outros serão desconsiderados por falta de relevância no contexto em estudo, para que sejam criados conjuntos de dados mais representativos e que possibilitem melhores resultados nos experimentos.

- **Partição dos conjuntos:** A partir dos dados coletados, serão criados dois conjuntos de dados para cada disciplina em estudo, onde o primeiro considerará apenas dados dos alunos que estavam em turmas onde foi postado em algum momento material de apoio; o segundo conjunto de dados será formado com dados filtrados por linha, ou seja, será feita uma limpeza de dados, cujo objetivo foi manter apenas os alunos que acessaram pelo menos uma vez o material postado.
- **Aplicação dos algoritmos:** Nessa etapa serão aplicados os algoritmos de classificação descritos na sessão 4.2.1.1.1 e na sessão 4.2.1.1.2, a saber: J48 e redes bayesianas, por terem sido considerados os mais eficazes para o tipo de problema proposto neste trabalho.

Em seguida será 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. Dessa forma, a primeira iteração será obtida a primeira precisão do modelo. Em seguida, para cada algoritmo de classificação aplicado, mais nove iterações percorrem todas as possibilidades de escolha, que resulta em mais nove valores de precisão. A precisão final do classificador é calculada, considerando a média das precisões das dez iterações (SILVA *et al*., 2015).

E finalmente, para a avaliação do desempenho dos algoritmos serão 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 que representa uma boa métrica para avaliar qual algoritmo utilizar) (SILVA *et al*., 2015).

- **Modelo preditivo:** Após todas as etapas descritas anteriormente neste capítulo deve ser possível nesta etapa a geração de um modelo preditivo onde será viável estabelecer uma relação entre o baixo desempenho dos alunos com a não postagem de material na turma virtual, ou mesmo o descarte da não postagem de material na turma virtual como fator que influencia o baixo desempenho dos alunos.

5 PROCEDIMENTOS METODOLÓGICOS PARA OS EXPERIMENTOS DE MINERAÇÃO DE DADOS

A seguir são apresentados os procedimentos metodológicos para os experimentos que foram realizados neste trabalho, visando gerar um modelo preditivo para diagnóstico de baixo desempenho a partir da utilização de MD.

A proposta metodológica deste trabalho tem foco na utilização de dados oriundos da base de dados dos cursos de graduação da Universidade Estadual do Maranhão, para diagnosticar o perfil de alunos com mau desempenho, por meio das diversas interações realizadas por eles na turma virtual.

O primeiro conjunto de dados, com 450 alunos dos cursos de Engenharia Agronômica, Engenharia de Pesca, Veterinária e Zootecnia, por serem cursos com disciplinas afins distribuídos em 10 turmas: piscicultura, topografia, química analítica, genética e solos, sendo duas turmas de cada disciplina. Dessa forma, foi obtido pela seleção realizada a partir das tabelas e atributos da base de dados do Turma Virtual que registram as interações dos alunos no download de arquivos postados pelos docentes. Dessa forma, o quadro 2 de sumarização apresenta os atributos selecionados**.**

Quadro 4- Atributos de Sumarização

Atributo	**Descrição**
nota_media_em_aval	Nota média do aluno nas avaliações
qtd_frequencia	Quantidade de frequência do aluno
qtd_down_turma	Quantidade downloads de arquivos postado pelo docente
total_post_prof	total de postagem do professor
resultado	Resultado final do aluno na disciplina (atributo classe)

Fonte: Elaborado pela autora (2019).

A distribuição das classes referentes ao resultado do primeiro conjunto de dados está assim apresentada: Aprovado Por Média = 250 (55,55%), Aprovado Por Final = 98 (21,17%), Reprovado Por Média = 80 (17,77%) e Reprovado Por Final = 22 (0,4%). Estas classes foram transformadas em apenas duas, sendo que foi mantida a classe Aprovado Por Média (anteriormente chamada de A.M), enquanto que as demais foram transformadas na classe Baixo Desempenho (anteriormente chamada

de B.D). Com isso, uma nova distribuição pode ser observada, conforme mostrado na Tabela 1.

Tabela 1- Distribuição das classes

Resultado	Total de Alunos	Percentual
A.M	250	55,55%
B.D	200	44,44%
TOTAL	**450**	**100%**

Fonte: Elaborada pela autora (2019).

O procedimento foi adotado porque a aprovação por média é o que se espera de um aluno em uma situação muito boa, mas quando isso não ocorre, 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 e porque a análise de desempenho é o que se busca neste trabalho.

O segundo conjunto de dados foi obtido a partir do primeiro através de um filtro por linha, que consistiu em manter apenas os alunos que estavam matriculados em turmas onde houve postagem de material de apoio, o que resultou em 273 alunos nessa condição, sendo mantidos os 5 atributos do conjunto de dados anterior e a distribuição das classes referentes ao resultado ficou assim: A.M = 273 (60,66%) e B.D = 177 (39,33%), conforme se observa na Tabela 2

Tabela 2- Distribuição de dados por linha

Resultado	Total de Alunos	Percentual
A.M	273	60,66%
B.D	177	39,33%
TOTAL	**450**	**100%**

Fonte: Elaborada pela autora (2019).

Para a mineração de dados deste trabalho fez-se uso da ferramenta Weka (*Waikato Environment for Knowledge Analyis*) que é uma coleção de algoritmos de aprendizagem de máquina e ferramentas de pré-processamento. É uma ferramenta de código aberto e foi desenvolvido na Universidade de Waikato na Nova Zelândia. Weka possui uma variedade de algoritmos de aprendizagem, que incluem ferramentas de pré-processamento. Além disso, oferece suporte a todo processo de mineração,

que inclui suporte a preparação dos dados de entrada, avaliação estatística da aprendizagem, visualização dos dados de entrada e os resultados. Os modelos gerados a partir de árvores de decisão utilizaram o algoritmo J48, e os baseados em estatística utilizaram o algoritmo NaiveBayes.

Em seguida 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. Dessa forma, a primeira iteração será obtida a primeira precisão do modelo. Em seguida, para cada algoritmo de classificação aplicado, mais nove iterações percorrem todas as possibilidades de escolha, que resulta em mais nove valores de precisão.

A precisão final do classificador é calculada, considerando a média das precisões das dez iterações. E finalmente, para a avaliação do desempenho dos algoritmos serão 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 que representa uma boa métrica para avaliar qual algoritmo utilizar) (SILVA *et al.*, 2015).

6 ANÁLISE DOS RESULTADOS

Os dados foram obtidos pela seleção realizada a partir das tabelas e atributos da base de dados do Turma Virtual que registram as interações entre professores e alunos. 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.

Para os dados obtidos das turmas de piscicultura, o experimento inicial utilizou dados originais e os melhores índices de desempenho no experimente foram registrados pelo algoritmo J48, seguido pelo algoritmo de redes. Observou-se (Tabela 3) que para dados originais, as duas técnicas tiveram desempenho muito semelhantes, com uma pequena vantagem do algoritmo J48 sobre o NaiveBayes.

Tabela 3- Distribuição de dados do primeiro conjunto de dados Piscicultura

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,723	0,731	0,722
Bayesianos	NaiveBayes	0,720	0,721	0,722

Fonte: Elaborada pela autora (2019).

No segundo experimento, usando dados filtrados por linha, foi perceptível o melhor desempenho do algoritmo J48, seguidos pelo desempenho apresentado pelo algoritmo NaiveBayes (Tabela 4).

Tabela 4- Distribuição de dados do segundo conjunto de dados Piscicultura

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,710	0,720	0,719
Bayesianos	NaiveBayes	0,722	0,721	0,720

Fonte: Elaborada pela autora (2019).

Para os dados obtidos das turmas de Topografia, o experimento inicial utilizou dados originais e os melhores índices de desempenho no experimente foram registrados pelo NaiveBayes, seguido pelo algoritmo J48. Observou-se (Tabela 5) que

para dados originais, as duas técnicas tiveram desempenho muito semelhantes, com uma pequena vantagem do algoritmo NaiveBayes sobre o J48.

Tabela 5- Distribuição de dados do primeiro conjunto de dados Topografia

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,728	0,732	0,731
Bayesianos	NaiveBayes	0,719	0,721	0,720

Fonte: Elaborada pela autora (2019).

No segundo experimento, usando dados filtrados por linha, o melhor desempenho foi do algoritmo J48, seguidos pelo desempenho apresentado pelo algoritmo NaiveBayes (Tabela 6).

Tabela 6- Distribuição de dados do segundo conjunto de dados Topografia

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,733	0,721	0,687
Bayesianos	NaiveBayes	0,693	0,702	0,679

Fonte: Elaborada pela autora (2019).

Para os dados obtidos das turmas de Química Analítica, o experimento inicial utilizou dados originais e os melhores índices de desempenho no experimente foram registrados pelo algoritmo J48, seguido pelo algoritmo de NaiveBayes. Observou-se (Tabela 7) que para dados originais, as duas técnicas tiveram desempenho muito semelhantes, com uma pequena vantagem do algoritmo J48 sobre o NaiveBayes.

Tabela 7- Distribuição de dados do segundo conjunto de dados Química

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,729	0,728	0,729
Bayesianos	NaiveBayes	0,719	0,721	0,725

Fonte: Elaborada pela autora (2019).

No segundo experimento, usando dados filtrados por linha, o melhor desempenho foi do algoritmo J48, seguidos pelo desempenho apresentado pelo algoritmo NaiveBayes (Tabela 8).

Tabela 8 - Distribuição de dados do segundo conjunto de dados Química

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,729	**0,735**	**0,727**
Bayesianos	NaiveBayes	0,720	0,721	0,725

Fonte: Elaborada pela autora (2019).

Para os dados obtidos das turmas de Genética, o experimento inicial utilizou dados originais e os melhores índices de desempenho no experimento também foram registrados pelo algoritmo J48, seguido pelo algoritmo de NaiveBayes. Observou-se (Tabela 9) que para dados originais, as duas técnicas tiveram desempenho muito semelhantes, com uma pequena vantagem do algoritmo J48 sobre o NaiveBayes.

Tabela 9- Distribuição de dados do primeiro conjunto de dados Genética

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,729	0,731	0,733
Bayesianos	NaiveBayes	0,718	0,720	0,722

Fonte: Elaborada pela autora (2019).

No segundo experimento, usando dados filtrados por linha, o melhor desempenho também foi do algoritmo J48, seguidos pelo desempenho apresentado pelo algoritmo NaiveBayes (Tabela 10).

Tabela 10- Distribuição de dados do segundo conjunto de dados Genética

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,729	0,730	0,729
Bayesianos	NaiveBayes	0,728	0,721	0,721

Fonte: Elaborada pela autora (2019).

Para os dados obtidos das turmas da disciplina de Solos o experimento inicial utilizou dados originais e os melhores índices de desempenho no experimento novamente foram registrados pelo algoritmo J48, seguido pelo algoritmo de NaiveBayes. Observou-se (Tabela 11) que para dados originais, as duas técnicas tiveram desempenho muito semelhantes, com uma pequena vantagem do algoritmo J48 sobre o NaiveBayes.

Tabela 11- Distribuição de dados do primeiro conjunto de dados Solos

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,730	0,728	0,729
Bayesianos	NaiveBayes	0,719	0,721	0,725

Fonte: Elaborada pela autora (2019).

No segundo experimento, usando dados filtrados por linha, o melhor desempenho foi do algoritmo J48, seguidos pelo desempenho apresentado pelo algoritmo NaiveBayes (Tabela 12).

Tabela 12- Distribuição de dados do segundo conjunto de dados Solos

Método	Algoritmo	Precision	Recall	F-Measure
Árvores de Decisão	J48	0,732	0,730	0,730
Bayesianos	NaiveBayes	0,725	0,727	0,725

Fonte: Elaborada pela autora (2019).

A taxa de acerto dos algoritmos ficou acima de 70%, de forma que o maior índice de precisão no experimento foi registrado pelo algoritmo J48 (73,8092%), seguido pelo algoritmo NaiveBayes (73,1236%). Os resultados obtidos (Tabela 13), mostram que para dados originais, as duas técnicas tiveram desempenho muito semelhantes, com uma leve vantagem da técnica J48 em relação às baseadas em redes bayesianas.

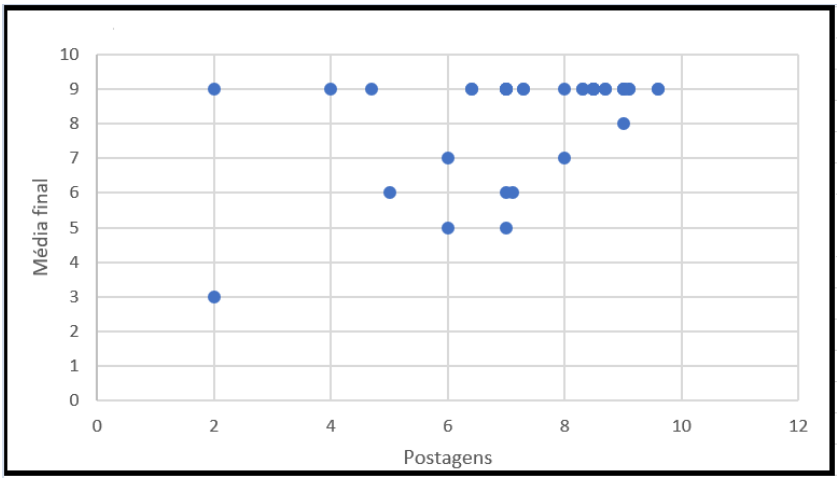
Tabela 13- Precisão de acerto em cada conjunto de dados (% de acerto)

Método	Algoritmo	Originais	Filtrados por linha
Árvores de Decisão	J48	73,8092	**73,5611**
Bayesianos	NaiveBayes	73,1236	72,6998

Fonte: Elaborada pela autora (2019).

No segundo experimento, usando dados filtrados por linha, mantiveram-se os resultados em relação ao experimento anterior, visto que os melhores desempenhos foram obtidos pelo algoritmo J48 (todas acima de 73%), sendo que as técnicas de baseadas em redes bayesianas ficaram abaixo de 72%. Dessa forma, a maior taxa de acerto no experimento foi registrada pelo algoritmo J48, que ficou abaixo dos 73%.

Figura 6 - gráfico de dispersão da quantidade de postagens



Fonte: Elaborado pela autora (2019).

Dessa forma, na figura 6 é exibida a tendência de desempenho dos alunos que foi gerada pela árvore de decisão, sobre a quantidade de material de apoio postado e as médias finais dos mesmos. É possível observar que as turmas com maior quantidade de postagens de material têm tendência a melhor desempenho. Nota-se que a partir de 6 postagens as médias dos alunos melhoram, enquanto as turmas com menos de 6 postagens tem tendência a baixo desempenho.

7 CONSIDERAÇÕES FINAIS

Desenvolver um modelo preditivo para medir os impactos no desempenho dos alunos a partir da inserção de material didático em ambientes virtuais de aprendizagem. Nesse sentido, ao fazer esse estudo, esperava-se que pudesse ficar notória a importância da postagem de material didáticos em AVAs.

Nesse contexto, foi feito uso do mecanismo de mineração de dados educacional através do uso de algoritmos de classificação para que fosse possível encontrar padrões e analisa-los sobre a ótica não só do desempenho do aluno, mas também da participação do professor.

Sendo assim, foi feita a revisão da literatura para que houvesse embasamento teórico do estudo. Em seguida deu-se a coleta de dados da base do SigUema, de forma que foram extraídas as turmas onde havia material didático inserido pelo professor e turmas anteriores onde não haviam esses dados. A preparação dos dados coletados para aplicação de técnicas de MD aconteceu em seguida a fim de obter um modelo de classificação utilizando a ferramenta Weka e aplicando-se os algoritmos de classificação.

A taxa de acerto dos algoritmos ficou acima de 70%, de forma que o maior índice de precisão no experimento foi registrado pelo algoritmo J48 (73,8092%), seguido pelo algoritmo NaiveBayes (73,1236%). Os resultados obtidos mostram que para dados originais, as duas técnicas tiveram desempenho muito semelhantes, com uma leve vantagem da técnica J48 em relação às baseadas em redes bayesianas.

Dessa forma, através da análise dos dados obtidos, foi possível alcançar o objetivo proposto, ou seja, foi observado que os alunos com melhor desempenho foram aqueles que estavam inseridos em turmas onde havia sido postado material didático em detrimento daqueles que estavam inseridos em turma onde não foi postado material.

Já com relação aos objetivos específicos, mostrou-se que as técnicas de mineração de dados baseados em algoritmos de classificação (mais precisamente o NaiveBayes e o J48), são eficazes para geração do modelo preditivo no que se refere a busca por padrões, permitindo assim que uma coleção de dados que incialmente não continha informação alguma, pudesse ser transformada em conhecimento e consequentemente usada como ferramenta de transformação.

Para trabalhos futuros, objetiva-se tornar a interpretação e compreensão dos resultados do modelo obtido de mais fácil entendimento. Para tanto, apresentam-se alguns desafios: criação de uma ferramenta para mostrar o diagnóstico ao público interessado para cada um dos resultados trabalhados neste estudo; apresentar a situação individual de cada aluno e integrar essa ferramenta como um módulo do SigUema. Importante informar ainda que esta ferramenta já está em desenvolvimento e que utiliza as regras de classificação geradas por modelos baseados em árvores de decisão. O foco deste trabalho foi especificamente na postagem de material de apoio na turma virtual por parte dos docentes.

Sendo assim, por meio da mineração de dados educacional é possível analisar 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 em sala de aula.

REFERÊNCIAS

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o Brasil. **Brazilian Journal of Computers in Education**, v. 19, n. 2, p. 3, 2011.

BRITO, D. M. *et al*. Predição de desempenho de alunos do primeiro período baseado nas notas de ingresso utilizando métodos de aprendizagem de máquina. **Brazilian Symposium on Computers in Education** (Simpósio Brasileiro de Informática na Educação-SBIE), v. 25, n. 1, 2014.

CAMILO, C. O.; SILVA, J. C. **Mineração de dados:** conceitos, tarefas, métodos e ferramentas. Goiânia: Universidade Federal de Goiás (UFC), 2009. p. 1-29.

COSTA, E. *et al*. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. **Jornada de Atualização em Informática na Educação**, v. 1, n. 1, p. 1-29, 2013.

COUTINHO, E.; KAESTNER, C. A. A.; NORONHA, R. V. Estimativa de desempenho acadêmico de estudantes: análise da aplicação de técnicas de mineração de dados em cursos a distância. **Revista Brasileira de Informática na Educação**, v. 22, n. 1, 2014.

DIAS, M. M. *et al*. **Mineração de dados educacionais**: relato de experiência no ambiente virtual LABSQL. [S.l.:s.n], 2014.

DIAS, M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. **Acta Scientiarum. Technology**, v. 24, p. 1715-1725, 2008.

FELCHER, C. D. O.; PINTO, A. C. M.; FERREIRA, A. L. A. O uso do facebook como ambiente virtual de aprendizagem para o ensino dos números racionais. **Revista Paranaense de Educação Matemática**, v. 6, n. 10, 2017.

GUÉRCIO, H. *et al*. Análise do Desempenho Estudantil na Educação a Distância Aplicando Técnicas de Mineração de Dados. In: WORKSHOPS DO CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. 2014. **Anais…** [S.l:s.n], 2014. p. 641.

HALL, M. *et al.* The weka data mining software: an update. **SIGKDD Explor. Newsl.**, v. 11, n. 1, p. 10-18, 2009.

KALINKE, M. A. **Tecnologias no ensino**: a linguagem matemática na web. Curitiba: CRV, 2014.

KAMPFF, A. J. C. *et al*. Identificação de perfis de evasão e mau desempenho para geração de alertas num contexto de educação a distância. **RELATEC: Revista Latinoamericana de Tecnología Educativa**, Madri, v. 13, n. 2 p. 61-76, 2014.

MARQUES, J. L. Q. **Mineração de dados educacionais**: um estudo de caso utilizando o ambiente virtual do SENAI. [S.l.:s.n], 2014.

MENDES, Luciana. Análise e predição de desempenho de discentes a partir da inserção de material de apoio em AVAs. In: CONCONFERÊNCIA INTERNACIONAL SOBRE INFORMÁTICA NA EDUCAÇÃO, 11, 2018, Brasília. Anais […]. Disponível em: http://www.tise.cl/Volumen14/TISE2018/336.pdf. Acesso em 12 de dezembro. 2018.

MOODLE Org. Disponível em: <https://moodle.org/>. Acesso em: 13 dez. 2018.

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

RICE, W. H. **Moodle E-learning Course Development:** a complete guide to succeful learning using Moodle. [S.l.]: Packt publishing., 2006.

RIGO, S. J. *et al*. Aplicações de Mineração de Dados Educacionais e Learning Analytics com foco na evasão escolar: oportunidades e desafios. **Revista Brasileira de Informática na Educação**, v. 22, n. 1, 2014.

ROMERO, C.; VENTURA, S.; GARCÍA, E.. Data mining in course management systems: Moodle case study and tutorial. **Computers & Education**, v. 51, n. 1, p. 368-384, 2008.

ROMERO-ZALDIVAR, V. *et al.* Monitoring student progress using virtual appliances: A case study. **Computers & Education**, v. 58, n. 4, p. 1058-1067, 2012.

SANTANA, L. C.; MACIEL, A. M.; RODRIGUES, R. L. Avaliação do perfil de uso no ambiente moodle utilizando técnicas de mineração de dados. **Brazilian Symposium on Computers in Education** (Simpósio Brasileiro de Informática na Educação-SBIE), v. 25, n. 1, p. 269, 2014.

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

SILVA, F. *et al*. Um modelo preditivo para diagnóstico de evasão baseado nas interações de alunos em fóruns de discussão. In: **Brazilian Symposium on Computers in Education** (Simpósio Brasileiro de Informática na Educação-SBIE), 2015. p. 1187.