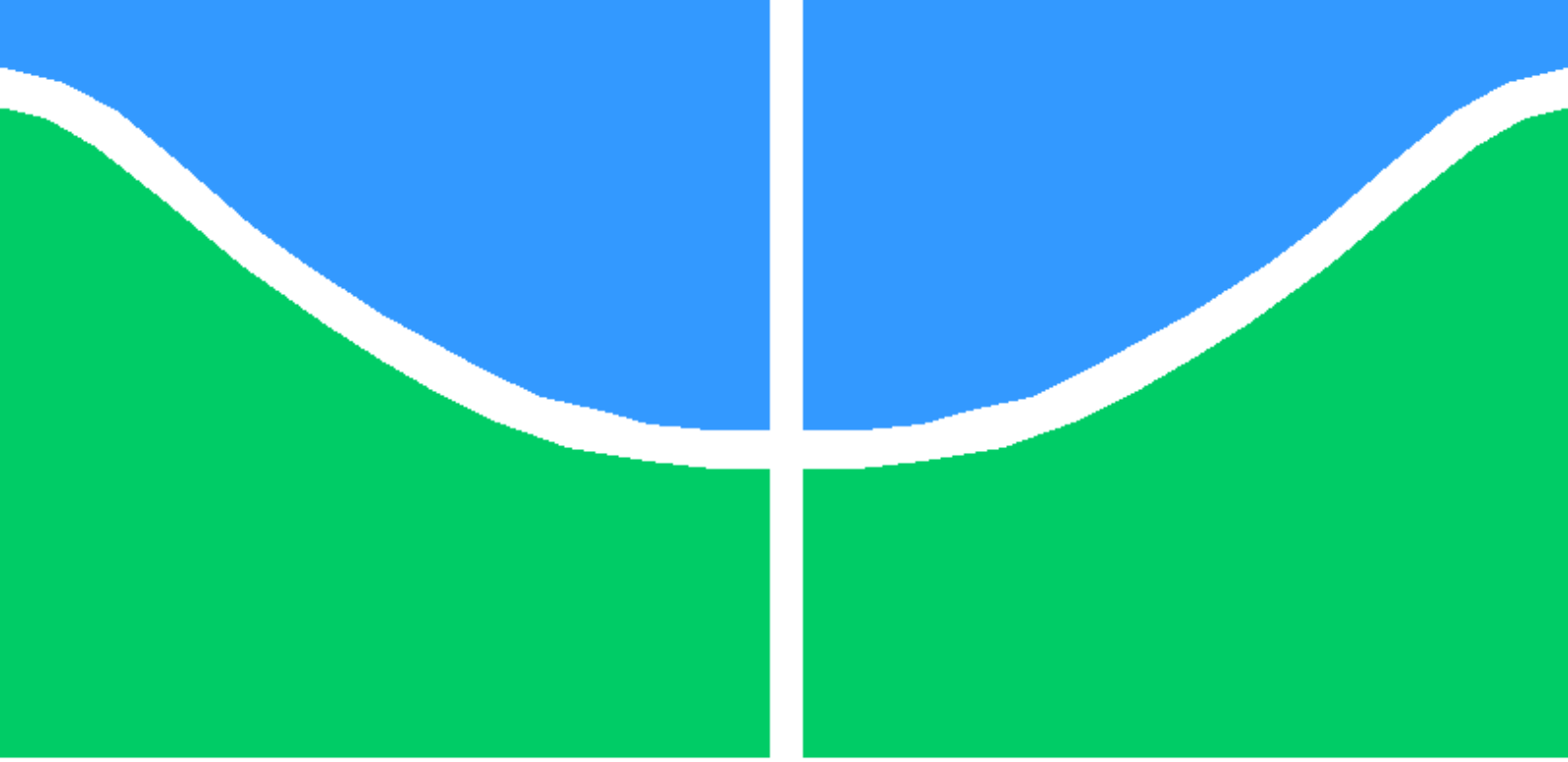


ANÁLISE DE LOGS DE INTERAÇÃO EM AMBIENTE EDUCACIONAL CORPORATIVO VIA MINERAÇÃO DE DADOS EDUCACIONAIS

VINÍCIUS COUTINHO GUIMARÃES COELHO

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

ANÁLISE DE LOGS DE INTERAÇÃO EM AMBIENTE EDUCACIONAL CORPORATIVO VIA MINERAÇÃO DE DADOS EDUCACIONAIS

VINÍCIUS COUTINHO GUIMARÃES COELHO

Orientador: PROF. DR. DANIEL GUERREIRO E SILVA, ENE/UNB

Coorientador: PROF. DR. JOÃO PAULO C. LUSTOSA DA COSTA, ENE/UNB

DISSERTAÇÃO DE MESTRADO EM ENGENHARIA ELÉTRICA

PUBLICAÇÃO PPGENE.DM - 662/2017
BRASÍLIA-DF, 26 DE JUNHO DE 2017.

FICHA CATALOGRÁFICA
VINÍCIUS COUTINHO GUIMARÃES COELHO
Análise de Logs de Interação em Ambiente Educacional Corporativo via Mineração de Dados Educacionais
2017xv, 75p., 201x297 mm
(ENE/FT/UnB, Mestre, Engenharia Elétrica, 2017)
Dissertação de Mestrado - Universidade de Brasília
Faculdade de Tecnologia - Departamento de Engenharia Elétrica

REFERÊNCIA BIBLIOGRÁFICA

VINÍCIUS COUTINHO GUIMARÃES COELHO (2017) Análise de Logs de Interação em Ambiente Educacional Corporativo via Mineração de Dados Educacionais. Dissertação de Mestrado em Engenharia Elétrica, Publicação **662/2017**, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 75p.

CESSÃO DE DIREITOS

AUTOR: VINÍCIUS COUTINHO GUIMARÃES COELHO
TÍTULO: Análise de Logs de Interação em Ambiente Educacional Corporativo via Mineração de Dados Educacionais.
GRAU: Mestre ANO: 2017

VINÍCIUS COUTINHO GUIMARÃES COELHO
QMSW 5 Lt. 8 Bloco 4 - Brasília/DF

UNIVERSIDADE DE BRASÍLIA
FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA

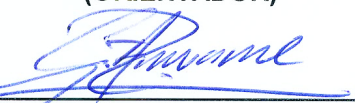
ANÁLISE DE LOGS DE INTERAÇÃO EM AMBIENTE EDUCACIONAL CORPORATIVO VIA MINERAÇÃO DE DADOS EDUCACIONAIS

VINÍCIUS COUTINHO GUIMARÃES COELHO

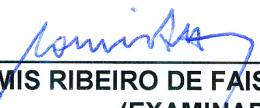
DISSERTAÇÃO DE MESTRADO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA ELÉTRICA DA FACULDADE DE TECNOLOGIA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE.

APROVADA POR:

DANIEL GUERREIRO E SILVA, Dr., ENE/UNB
(ORIENTADOR)



GEORGES DANIEL AMVAME NZE, Dr., ENE/UNB
(EXAMINADOR INTERNO)



ROMIS RIBEIRO DE FAISSOL ATTUX, Dr., UNICAMP
(EXAMINADOR INTERNO)

Brasília, 26 de Junho de 2017.

Agradecimentos

Agradeço a Deus pois sem ele nada tem sentido nessa vida.

À minha família, em especial minha mãe Célia Guimarães que sempre me apoiou e acreditou nas horas que realmente precisei. Tudo é graças a você.

Agradeço a meu orientador Professor Daniel Guerreiro e Silva, pela paciência e apoio incondicional para a condução deste trabalho e pelas preciosas orientações em momentos críticos.

Ao meu coorientador Professor João Paulo C. Lustosa da Costa, pelo desafio lançado para realização deste mestrado e pelo apoio incondicional.

Ao grande amigo Daniel A. da Silva, por acreditar e proporcionar tantas oportunidades e pelo suporte e auxílio na condução desse projeto de mestrado.

Ao Professor Rafael Timóteo, pelo apoio junto aos projetos de pesquisa que participei no decorrer dessa jornada.

Agradeço à ENAP - Escola Nacional de Administração Pública (TED 83/2016) e ao projeto MP/DIPLA (TED 05/2016) pelo apoio à pesquisa no decorrer deste projeto de mestrado.

Aos amigos do projeto DIPLA, pela convivência diária agradável e por todo companheirismo. Aos amigos Vitor Lopes e Alessandro Mendes pelo apoio com a dissertação e questões de Aprendizado de Máquina.

À minha esposa Tatiana, pela cumplicidade, paciência, apoio, cobrança nas horas que foi necessário e amor acima de tudo. À nossa ”Gaviota” pelo companheirismo.

A meu querido filho Pedro Lucas, pela paciência, vibração e amor incondicional. Somos capazes!

RESUMO

A Mineração de Dados Educacionais (MDE) (do inglês, *Educational Data Mining*) tem sido uma ferramenta crucial para a melhora da Educação a Distância (EAD), permitindo, por exemplo, a identificação de características de participantes, a análise preditiva de desempenho bem como o reconhecimento dos tipos e padrões de aprendizado. A literatura científica apresenta uma vasta quantidade de trabalhos relacionados a ambientes educacionais de Instituições de Ensino Superior. Entretanto, tais ambientes possuem um modelo pedagógico com características específicas comuns a cursos de graduação e pós-graduação. Neste trabalho de mestrado, é proposto um modelo de aplicação de técnicas de EDM para um Ambiente Virtual de Aprendizagem (AVA) corporativo, de âmbito governamental. Foram gerados dados referentes aos logs de interação de cerca de 70 mil alunos em 45 turmas de 7 cursos na modalidade a distância da Escola Nacional de Administração Pública (Enap), entre 2015 e 2016. Por meio de técnicas de classificação usando árvores de decisão, verifica-se o relacionamento entre as interações realizadas pelos alunos ao longo do curso e as notas finais obtidas. Foi utilizada uma metodologia de agrupamento dos dados de interação divididos em semanas, com o intuito de viabilizar possíveis intervenções antes do término dos cursos. Foi possível concluir que o modelo proposto alcançou bons resultados quando comparados à literatura específica e que foi capaz de gerar indicadores relacionados aos perfis de interação dos alunos, que são passíveis de utilização para o combate às taxas de evasão e reprovação, nos cursos a distância ofertados por uma instituição corporativa governamental de ensino.

Palavras chaves: Mineração de Dados Educacionais, Aprendizado de Máquina, Classificação Supervisionada, Educação Corporativa.

ABSTRACT

Educational Data Mining has been a crucial tool for the improvement of Distance Education, allowing, for example, the identification of characteristics of participants, predictive performance analysis as well as the recognition of learning types and patterns. The scientific literature shows a vast amount of work related to educational environments of Higher Education Institutions. However, such environments have a pedagogical model with specific characteristics common to undergraduate and postgraduate courses. In this master's work, a model of application of EDM techniques for a corporate Virtual Learning Environment (VLE) is proposed, , of governmental scope. Data were generated for interaction logs of about 70 thousand students in 45 classes of 7 courses in the distance modality of the National School of Public Administration (Enap) between 2015 and 2016. Through classification techniques using decision trees, relationship between the interactions carried out by the students along the course and the final grades obtained is verified. A methodology was used to group the interaction data divided into weeks, in order to enable possible interventions before the end of the courses. It was possible to conclude that the proposed model achieved good results when compared to the specific literature and it was able to generate indicators related to the students interaction profiles, which can be used to combat dropout and failure rates in distance courses offered in governmental educational institution.

Keywords: *Educational Data Mining, Machine Learning, Supervised Classification, Corporate Education.*

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE TERMOS E SIGLAS

AM	Aprendizado de Máquina (AM) (do inglês, *Machine Learning*)
ARFF	*Attribute Relation File Format* (ARFF)
AVA	Ambientes Virtuais de Aprendizagem)
CGEAD	Coordenação Geral de Educação a Distância (CGEAD)
CRISP-DM	do inglês *Cross Industry Standard Process for Data Mining*
EAD	Educação a distância
IES	Instituições de Ensino Superior
ITS	Sistemas de Tutoria Inteligente (ITS) (do inglês, *Intelligent Tutor System*)
KDD	Descoberta de Conhecimentos em Bases de Dados (do inglês, *Knowledge Discovery in Databases – KDD*)
MD	Mineração de Dados (MD) (do inglês, *Data Mining*)
MDE	Mineração de Dados Educacionais (do inglês, *Educational Data Mining)*
TIC	Tecnologias da Informação e Comunicação

Capítulo 1

INTRODUÇÃO

A crescente utilização da Educação a Distância (EAD) é destaque em todas as suas áreas de aplicação devido à sua grande amplitude no atendimento das demandas por aprendizagem. A partir desse crescimento, a utilização dessa metodologia tem ganhado destaque na educação corporativa e vem sendo adotada em programas de qualificação e formação profissional, bem como em instituições de ensino superior na oferta de cursos de graduação e pós graduação.

Neste contexto, diversas instituições na Europa, Canadá e também no Brasil adotaram essa modalidade como ferramenta para o desenvolvimento e capacitação dos servidores públicos. No Brasil, a Escola Nacional de Administração Pública (Enap) tem como missão o desenvolvimento de competências dos servidores públicos para aumentar a capacidade de governo na gestão de políticas públicas. No entanto, a utilização da EAD está relacionada a um processo de ensino e aprendizagem mediado diretamente por tecnologias. Isto significa que é necessário que os atores envolvidos nesse processo possuam uma alfabetização tecnológica básica, para que possa haver interação com os ambientes de estudo [Abbad 2007].

Os ambientes educacionais utilizados na EAD, denominados Ambientes Virtuais de Aprendizagem (AVA), registram em suas bases de dados todas as interações realizadas pelos alunos no decorrer dos cursos. Esses registros, devido ao nível de detalhe, geram imensas massas de dados que são humanamente impossíveis de serem processadas. Porém, existem técnicas computacionais que auxiliam com essas atividades para o processamento de grandes massas de dados em busca de conhecimentos que podem contribuir com a melhoria da EAD.

A área de Mineração de Dados Educacionais (MDE) tem como objetivo principal a aplicação de técnicas computacionais para o tratamento das grandes massas de dados geradas em AVA. A MDE 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 de forma quantitativa e qualitativa, a melhoria na oferta de cursos utilizando a EAD [Baker et al. 2011b].

Entre as atividades presentes na MDE, uma das mais utilizadas é a classificação supervi-

sionada de padrões, que se caracteriza por organizar objetos em classes pré-definidas. Trata-se de uma abordagem sistemática para construção de modelos de classificação a partir de conjuntos de dados. Existem diversas técnicas que podem ser utilizadas, e.g. classificadores baseados em árvores de decisão, classificadores baseados em regras, redes neurais artificiais, máquinas de vetores suporte e classificadores bayesianos[Tan et al. 2009].

Em ambientes educacionais, a predição de desempenho dos alunos possui dois contextos distintos para sua aplicação: 1) o estudo da influência dos atributos de um modelo específico para a previsão de uma classe e 2) previsão de um resultado para uma classe alvo de saída de acordo com os atributos preditores utilizados. É possível, neste sentido, direcionar técnicas de classificação para a análise e previsão de desempenho dos alunos, possibilitando a identificação de padrões que podem ser monitorados, como indicadores de intervenção para a melhoria da EAD [Baker et al. 2010].

1.1 MOTIVAÇÃO

O presente trabalho teve seu início a partir do projeto ”Educação mediada por tecnologias”, – TED[1] firmado entre a UnB (Universidade de Brasília) e a Enap–, com a participação do autor junto à equipe de tecnologia, voltada para a pesquisa e implementação de soluções inovadoras relacionadas ao cenário da EAD em instituições corporativas de cunho governamental.

Diferentemente das Instituições de Ensino Superior (IES), os cursos ofertados em ambientes corporativos geralmente são de curta duração e estão focados unicamente nos conteúdos e objetos educacionais que são utilizados a partir de uma plataforma tecnológica. Nesse contexto, são geradas imensas massas de dados relacionadas à interação dos alunos com o ambiente dos cursos através do registro de *logs* em tabelas que são armazenadas por um AVA, como por exemplo um dos mais utilizados, o Moodle.

No Brasil, segundo estudos realizados pela Associação Brasileira de Educação a Distância (ABED) em 2015/2016, os índices de evasão estão em torno de 40% nas instituições que oferecem cursos totalmente a distância[ABED 2015]. Nesse contexto, conforme o estudo apresentado por [Baker et al. 2011a], a área de Mineração de Dados Educacionais (MDE) demonstra possibilidades promissoras para a exploração dos dados provenientes de ambientes educacionais, através de técnicas de aprendizado de máquina, possibilitando o desenvolvimento de métodos que viabilizem a compreensão de forma mais eficaz e adequada de como os alunos aprendem e quais fatores estão relacionados a esse aprendizado.

No cenário da MDE, destacam-se os trabalhos realizados por [Baker and Yacef 2009] [Romero 2010] e [Peña-Ayala 2014], que, cronologicamente, apresentam uma revisão do estado da arte sobre a utilização de técnicas de MDE não somente para o combate à evasão mas

[1]Termo de Descentralização - Convênios entre órgãos

também, para questões como a modelagem dos estudantes, suporte pedagógico e descobertas científicas, entre outras.

Entretanto, conforme pode ser observado nos trabalhos citados anteriormente, bem como em [Bresfelean 2007, Bunkar et al. 2012, Hoe et al. 2013, Mishra et al. 2014, Guleria et al. 2014, Jindal and Borah 2015], tais trabalhos estão altamente concentrados em dados provenientes de cenários relacionados a um tipo específico de instituição, as Instituições de Ensino Superior (IES). A metodologia para oferta de cursos em EAD nessas instituições possui características únicas, inerentes ao tipo de curso que é ofertado, no caso, cursos de graduação. Essas características, que são comuns em alguns estudos relacionados aos dados de IES, muitas vezes não estão presentes em dados provenientes de outros tipos de instituições, como aquelas de educação corporativa governamental, i.e. a Enap.

Dentro deste contexto específico — educação a distância no ambiente corporativo governamental — a principal motivação para este trabalho concentra-se no estudo da aplicação de técnicas de MDE, utilizando dados oriundos de uma instituição focada na EAD corporativa. Nesta instituição, os cursos ofertados possuem características específicas, como a sua modalidade de oferta, com tutoria e sem tutoria, bem como a característica relacionada ao tempo de duração dos cursos, em torno de 30 dias distribuídos em quatro semanas de duração. Os cursos sem tutoria agrupam, nesta instituição, a maior quantidade de alunos que constituem os cursos com maior representatividade.

1.2 OBJETIVOS

Partindo da motivação apresentada anteriormente, o objetivo principal desse trabalho pode ser sintetizado da seguinte forma:

Estudar e analisar uma proposta de modelo de Mineração de Dados que possibilite a descoberta de conhecimentos relacionados à interação dos alunos com o AVA, utilizando dados históricos da oferta de cursos de uma instituição focada na EAD corporativa de cunho governamental, ou seja, educação ao longo da vida (do inglês *lifelong learning*).

Para alcançar o objetivo proposto, foram delimitados alguns objetivos específicos, conforme listado a seguir:

- Analisar e propor uma metodologia para extração das informações relacionadas à interação dos alunos com o AVA Moodle;

- Criar bases de dados que apresentem as interações dos alunos separadas em intervalos semanais, contemplando 7 dias de interação, desde a primeira até a terceira semana de realização dos cursos. Essas bases devem considerar a forma de composição dos dados onde em uma das bases os dados representarão as interações de cada semana de forma isolada, ou seja, os dados somente da semana que passou sem considerar as semanas anteriores. Na outra base, serão considerados os dados de forma incremental, onde, ao final da semana,

serão considerados os dados da semana atual e também os dados das semanas anteriores;

- Analisar e comparar os resultados a partir de técnicas de classificação supervisionada com árvores de decisão utilizando o algoritmo C4.5;

- Estudar e comparar qual composição das bases de dados em semanas separadas ou semanas incrementais, é mais promissora em relação aos resultados e

- Estudar e comparar os resultados obtidos sob a ótica da melhor composição da base de dados a partir da técnica de rebalanceamento de classes.

Através dos objetivos expostos, foram realizados os experimentos apresentados no Capítulo 4 com o intuito de responder às seguintes questões:

1. Qual é a melhor abordagem em relação à composição dos *datasets* para o estudo de caso proposto?
2. A técnica de balanceamento de classes (*RESAMPLE*) pode ser considerada para melhoria dos resultados no estudo de caso proposto?
3. O modelo proposto alcançou um bom desempenho para os padrões da literatura?
4. É possível gerar indicadores de interação que auxiliem o combate à evasão e reprovação a partir do modelo proposto?

1.3 PUBLICAÇÕES DO AUTOR

No decorrer da realização deste trabalho de mestrado, o autor buscou a publicação de artigos científicos para embasamento da pesquisa proposta. Inicialmente, foi publicado um artigo visando o estudo dos registros em sistemas de comunicação relacionados à barreiras na utilização da EAD. Em seguida, foi publicado um artigo que refere-se diretamente ao trabalho proposto na seção de experimentos.

[Coelho et al. 2015] Coelho, V. C. G., Costa, J. P. C. L. d., Souza, D. d. C. R. d., Canedo, E. D., Silva, D. G. e., and Sousa Júnior, R. T. d. (2015). **Mineração de dados educacionais para identificação de barreiras na utilização da educação a distância**. In 21º Congresso Internacional ABED de Educação a Distância. ABED.

[Coelho et al. 2016] Coelho, V. C. G., da Costa, J. P. C. L., da Silva, D. A., de Sousa Júnior, R. T., de Mendonça, F. L., and Silva, D. G. (2016). **Mineração de dados educacionais no ensino a distância governamental**. In Conferências Ibero-Americanas WWW/Internet e Computação Aplicada 2016, pages 1–10. CIAWI.

1.4 ORGANIZAÇÃO DO TRABALHO

O restante desse trabalho esta distribuído da seguinte forma: no Capítulo 2, serão apresentados os conceitos teóricos que foram utilizados para embasamento das pesquisas, iniciando pela questão da tecnologia e a educação na geração de dados, que são utilizados em processos de descoberta de conhecimentos através de metodologias de Mineração de Dados por meio de técnicas consagradas de Aprendizado de Máquina. Neste capítulo, também serão abordadas as principais questões envolvendo a Mineração de Dados Educacionais, com um detalhamento sobre suas possibilidades e aplicações.

O Capítulo 3 dedica-se à apresentação dos trabalhos que estão relacionados à mesma linha de pesquisa abordada com a presente dissertação. Os trabalhos selecionados abrangem a aplicação de técnicas de MDE através de algoritmos de classificação supervisionada e árvores de decisão, além de questões sobre o tratamento de classes desbalanceadas e análise de *logs* de interação em ambientes educacionais.

Em seguida, no Capítulo 4, é apresentado o estudo de caso que foi conduzido a partir da utilização de uma base de dados com registros de interações dos alunos, nos cursos oferecidos entre 2015 e 2016 por uma instituição focada na EAD corporativa governamental, a Enap. Nesse capítulo, é utilizado um *framework* específico para a condução de projetos de Mineração de Dados, que possui fases distintas desde o entendimento do negócio até a implementação do modelo de mineração.

No Capítulo 5 serão apresentadas as discussões em relação aos resultados gerais do projeto de mineração. Inicialmente serão discutidos os resultados sobre a performance do algoritmo de classificação em relação à composição dos datasets. Em seguida, será apresentado o resultado alcançado através da utilização da técnica de balanceamento de classes, quando considerado o dataset que obteve a melhor performance na etapa anterior.

Por fim, o Capítulo 6 apresenta as conclusões para os estudos apresentados nos capítulos anteriores e para os experimentos propostos, bem como as possibilidades para a sequência dessa dissertação na seção de trabalhos futuros.

Capítulo 2

REFERENCIAL TEÓRICO

Este capítulo apresenta o embasamento teórico utilizado para o desenvolvimento desse projeto de pesquisa. Serão apresentadas as questões relacionadas à descoberta de conhecimentos em bases de dados de ambientes educacionais, abordando assuntos como a utilização da tecnologia em ambientes educacionais, a aplicação de técnicas de mineração de dados via algoritmos de aprendizado de máquina e a Mineração de Dados Educacionais.

2.1 EAD e TECNOLOGIA

A Educação a Distância pode ser definida como um processo de ensino e aprendizagem mediado por tecnologias, em ambientes separados por espaço e tempo, onde as tecnologias interativas evidenciam a base para o processo de educação através da interação e interlocução entre todos os atores envolvidos [Moran 2002].

O uso da EAD está baseado na utilização de Tecnologias da Informação e Comunicação (TIC), que requerem que seus usuários possuam uma devida alfabetização tecnológica, como, por exemplo, o manuseio de editores de texto, planilhas, e-mail, participação em chats e utilização de buscas na internet [Almeida et al. 2013].

No modelo tradicional de educação, as informações são registradas em papel ou sistemas básicos de secretaria acadêmica onde geralmente são armazenados dados relacionados à frequência dos alunos, informações pedagógicas do curso ou matéria, além dos objetivos curriculares e alguns poucos dados individualizados dos alunos. No entanto, a educação baseada na tecnologia e na web, a EAD, possui muito mais informação disponível, tendo em vista que os ambientes tecnológicos educacionais podem gravar todas as informações sobre as ações e interações dos alunos em arquivos de *log* e em sistemas de bancos de dados. A mineração desses dados pode construir modelos analíticos que permitem descobrir padrões interessantes e tendências em informações relacionadas aos alunos, cursos e conteúdos [Romero et al. 2008].

Conforme estudos sobre fatores que afetam o desempenho de estudantes, [Peña-Ayala 2014, Baker and Yacef 2009, Romero 2010], percebe-se que algumas barreiras ainda precisam ser vencidas para que a EAD seja realmente acessível e para que sejam levadas em consideração as limitações e dificuldades dos usuários, os quais podem apresentar brechas para o processo inclusivo da aprendizagem. A falta de uma alfabetização tecnológica pode gerar um impacto negativo nos indivíduos que têm seus primeiros contatos com a EAD e também para os que já são usuários dessa modalidade, mas que ainda possuem dificuldades de adaptação tecnológica.

A necessidade do uso de TIC pode gerar a evasão de alunos que possuam um grande potencial de aprendizagem, mas com limitações relacionadas aos conhecimentos tecnológicos. Essa situação pode excluí-los das oportunidades e vantagens que o modelo de ensino e aprendizagem da EAD disponibiliza [Albertin and Brauer 2012]. Esse modelo de aprendizagem está relacionado à criação de um ambiente que seja propício para promover a colaboração e a interação dos atores envolvidos no processo de aprendizagem, que são denominados Ambientes Virtuais de Aprendizagem [Pereira et al. 2007].

A utilização de um AVA está diretamente relacionada à geração de grandes massas de dados devido à quantidade de interações e dados que são armazenados. Com essa grande quantidade de informações armazenadas, que não são passíveis de análise por seres humanos, há possibilidades de aplicação de técnicas computacionais para a descoberta de conhecimentos relevantes acerca do comportamento dos alunos e também sobre os conteúdos dos cursos que podem auxiliar para a melhoria da EAD e para a redução de taxas de reprovação e evasão [Baker and Yacef 2009].

2.2 DESCOBERTA DE CONHECIMENTOS

A utilização de uma grande diversidade de sistemas computacionais, aliada à necessidade de armazenamento e tratamento da imensa quantidade de dados gerados, faz parte da realidade em diversas áreas de atuação, como, por exemplo, bancos, instituições financeiras, governos, educação, ciência, entre outros. A análise e extração de conhecimentos nessas imensas massas de dados, que são geradas através da utilização de TIC, torna-se humanamente impossível sem o auxílio de técnicas computacionais. Contudo, os dados brutos, ou seja, os dados operacionais que são provenientes de processos transacionais, são de pouca contribuição, sem o devido tratamento, para o processo de tomada de decisão [Barbieri 2011].

Para que os dados possam ser devidamente utilizados como insumos de relevância no processo decisório, é necessário que seja realizada uma transformação em sua forma e conteúdo. Os dados transacionais devem ser transformados em informação e disponibilizados em um ambiente adequado de coleta, armazenamento e publicação. Essas informações possibilitam que as instituições possam utilizar técnicas para descoberta de conhecimen-

tos, gerando insumos informacionais estratégicos de acordo com o domínio de aplicação [Tan et al. 2009].

Como pode ser observado na Figura. 2.1, a entrada inicial é o dado bruto onde são realizadas intervenções até que possa ser gerada a experiência, ou seja, o conhecimento válido para o domínio estudado. O processo de transformação do dado bruto até a geração de uma experiência, aplicado ao contexto desse trabalho, é também conhecido como Descoberta de Conhecimentos em Bases de Dados (do inglês, *Knowledge Discovery in Databases - KDD)*.

Figura 2.1: Evolução dos dados até a experiência

2.2.1 Processo de KDD

Analisando a perspectiva do conhecimento a ser extraído, [Fayyad et al. 1996b] definem o processo de KDD como:

> *"Um processo não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados."*

Nessa definição, a *não trivialidade* do processo está relacionada à dificuldade na percepção e interpretação de forma adequada dos inúmeros fatos observados, bem como a dificuldade de utilizar de forma dinâmica as interpretações disponibilizadas, com o intuito de decidir quais ações podem ser aplicadas a cada caso em particular.

O fato de ser *interativo* refere-se à necessidade incontestável da participação do homem para o controle do processo através da utilização de recursos computacionais direcionados para a análise e interpretação dos fatos observados e os resultados obtidos no decorrer do processo. A presença do homem se dá em dois papéis: (1) o analista de dados ou cientista de dados, com o perfil relacionado ao entendimento e domínio do processo e (2) o especialista de domínio, que possui conhecimentos específicos no âmbito da aplicação na qual se insere o problema a ser resolvido [Goldschmidt and Bezerra 2015].

Durante o processo de KDD, pode ser necessário que hajam refinamentos sucessivos para encontrar os resultados mais adequados e satisfatórios ao domínio, ou seja, podem haver *iterações* integrais ou parciais até que sejam alcançados objetivos realmente representativos.

Um dos principais objetivos descritos pelo KDD trata da identificação de padrões que sejam compreensíveis, ou seja, de fácil entendimento, de forma clara e concisa. Os co-

nhecimentos devem ser válidos, verdadeiros e adequados ao domínio em análise. Devem ser novos, acrescentando conhecimentos desconhecidos ou agregando padrões previamente existentes. Por fim, os padrões devem ser úteis e proporcionar novos benefícios [Fayyad et al. 1996b].

As etapas do processo de KDD proposto por [Fayyad et al. 1996b] podem ser visualizadas na Figura 2.2. Cada etapa está relacionada ao desenvolvimento de atividades específicas onde a entrada é o dado bruto e a saída final é o conhecimento.

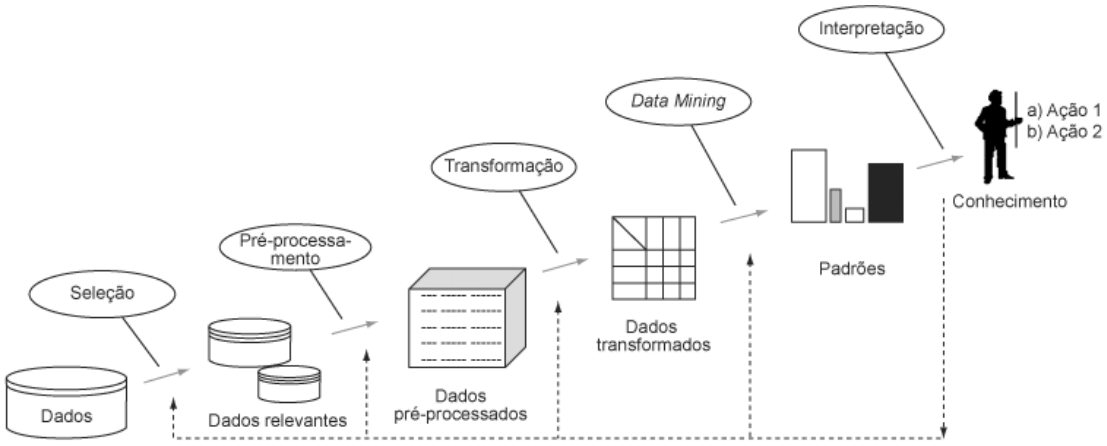


Figura 2.2: Etapas do processo de KDD
Adaptado de [Fayyad et al. 1996b]

Na figura, é possível visualizar que o início do processo se dá a partir da Seleção da base de dados que será utilizada, com o intuito de possibilitar o entendimento sobre o domínio da aplicação e também a seleção do conhecimento prévio relevante visando identificar o objetivo para o KDD. A fase de Pré-processamento trata da limpeza dos dados através da realização de operações básicas para remoção de ruídos (*outliers*) e para a definição de estratégias para o tratamento de informações ausentes (*missing values*). A próxima etapa realiza as Transformações necessárias a partir da redução da dimensão dos dados e da quantidade efetiva de variáveis. Com os dados transformados e tratados, é possível a realização da Mineração dos Dados, que aborda o planejamento e seleção dos algoritmos a serem aplicados, de acordo com o objetivo definido, em busca de padrões que sejam passíveis de análise e possam ser utilizados como fonte de informações em processos de tomada de decisão [Baker et al. 2010].

Em contrapartida, [Brachman and Anand 1996] defendem uma outra perspectiva, onde o KDD é um processo e que não se limita à descoberta de padrões, mas que está relacionado, entre outras, à negociação com os ”donos dos dados”, manipulação e grande interação com os dados, definindo o processo de KDD da seguinte forma:

> *”O processo de KDD consiste de uma sequência de interações complexas, que se estende sobre um determinado período de tempo, entre um analista de dados e uma coleção de dados, possivelmente auxiliado por um conjunto heterogêneo de ferramentas computacionais.”*

Essa definição considera o fato de que o analista de dados deve sempre estar presente e com alto nível de relacionamento com todas as etapas envolvidas no processo. Essa interação deve possibilitar que o analista de dados formule hipóteses relacionadas ao universo de dados de forma global e empírica, direcionando onde intensificar a exploração para geração de indicadores e informações úteis e válidas [Goldschmidt and Bezerra 2015].

2.3 MINERAÇÃO DE DADOS

A Mineração de Dados (MD) (do inglês, *Data Mining*) é o processo de extrair padrões ocultos e previamente desconhecidos de dados brutos, com a intenção de transformar essas grandes quantidades de dados em informações úteis. Pode ser definida como um processo operacional para descoberta de conhecimentos em grandes massas de dados, de forma automática ou semiautomática, para a identificação de padrões em dados que possibilitem conhecimentos relevantes, únicos e válidos. Os padrões descobertos devem possuir valores significativos e devem levar a alguma vantagem, geralmente de natureza econômica ou estratégica. Os dados utilizados em ambientes de MD estão inevitavelmente presentes em quantidades substanciais [Witten and Frank 2005].

Vale ressaltar que a MD é apenas uma das etapas do processo de KDD, conforme apresentado na Figura 2.2. Neste contexto, conhecimento significa relacionamento e padrões entre elementos de dados, presentes na MD e utilizados como insumos de um estágio para descoberta dentro do processo de KDD [Adriaans and Zantinge 1996].

A identificação de padrões trata do conhecimento representado levando em consideração normas sintáticas na utilização de algum tipo de linguagem formal, que seja passível de interpretação por seres humano. Um exemplo de representação do conhecimento seria uma linguagem baseada em equações onde operadores matemáticos são utilizados para relacionar variáveis, e.g. $A = bX + C$ [Goldschmidt and Bezerra 2015].

Padrões podem ser classificados em dois tipos básicos, preditivos e descritivos. **Padrões Preditivos** possuem a característica de tentar resolver um problema específico prevendo valores de um ou mais atributos em função de um outro atributo ou uma classe alvo. Esse padrões podem ser avaliados pelo julgamento de quão efetivos eles são na predição de algum fato futuro baseado em atributos e classes. **Padrões Descritivos** têm como objetivo central a apresentação de informações que sejam interessantes ao especialista de domínio. Possuem uma dificuldade mais acentuada de avaliação, em virtude de sua real contribuição estar relacionada ao fato de esses padrões sugerirem ações que sejam úteis para o especialista de domínio e na observação de quão efetivas essas ações se apliquem ao contexto da aplicação [Hand et al. 2001].

Uma outra visão sobre MD se dá como a aplicação de algoritmos específicos para extração de padrões de dados. Isso demonstra que a ênfase está na aplicação de algoritmos, ao contrário dos próprios algoritmos. Neste sentido, é possível definir a relação entre

o aprendizado de máquina e mineração de dados da seguinte forma: a mineração de dados é um processo, durante o qual os algoritmos de aprendizado de máquina são utilizados como ferramentas para extrair padrões potencialmente valiosos dentro de grandes conjuntos de dados [Fayyad et al. 1996a].

Trata-se de um campo multidisciplinar que teve suas origens a partir de tecnologias de bancos de dados, aprendizado de máquina, inteligência artificial e estatística entre outras áreas. É um campo onde os elementos estatísticos são utilizados nas funções como classificação, *clustering*, regressão e associação. No entanto, a mineração de dados engloba uma variedade de tarefas que não são de natureza estatística. Por exemplo, a preparação dos dados, a inspeção e limpeza que são de grande importância e, quando combinados, podem ser responsáveis por mais de 60% de todo o tempo de um projeto de MD [Tan et al. 2009].

Outra definição, segundo [Alpaydin 2014], é que a aplicação de métodos de aprendizado de máquina para grandes bancos de dados é chamada de Mineração de Dados. A analogia ao termo é que um grande volume de terra e matéria-prima extraído de uma mina, quando processado, leva a uma pequena quantidade de material muito precioso. Da mesma forma, na Mineração de Dados, um grande volume de dados é processado para construir um modelo simples, porém de uso valioso.

A MD está diretamente relacionada com o Aprendizado de Máquina, pois, na MD, o objetivo de suas atividades está relacionado à aplicação de algoritmos específicos para extração de padrões em bases de dados. Conforme ressaltado na seção anterior, a ênfase da MD está na aplicação e utilização de algoritmos de Aprendizado de Máquina como ferramentas para descobrir padrões que sejam potencialmente valiosos para o processo de KDD [Fayyad et al. 1996b].

2.4 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina (AM) (do inglês, *Machine Learning*) pode ser definido como um método de análise que automatiza o desenvolvimento de modelos analíticos usando algoritmos que aprendem interativamente a partir de dados. O aprendizado de máquina permite que os computadores encontrem *insights* ocultos sem serem explicitamente programados para procurar algo específico. O aspecto iterativo do aprendizado de máquina é importante porque, conforme os modelos são expostos a novos dados, eles são capazes de se adaptar de forma independente, ou seja, aprendem com os cálculos anteriores para produzir decisões e resultados confiáveis e reproduzíveis [Sammut and Webb 2011].

Em um contexto prático, o aprendizado pode ser caracterizado a partir do momento em que um determinado comportamento é alterado, baseado em acontecimentos, para gerar um melhor desempenho futuro. É possível testar o aprendizado através da observação do comportamento, comparando-o com o comportamento passado [Witten and Frank 2005]. A visão sobre o aprendizado relaciona-se à utilização de técnicas que possibilitem a evolução

dos resultados alcançados de acordo com o nível de interação com o ambiente aplicado. Um modelo de aprendizado denota mudanças que sejam adaptativas no sentido de permitir que esse modelo realize a mesma tarefa, ou tarefas extraídas da mesma população, de forma mais eficiente e mais eficaz a cada ciclo de interação [Simon 1983]. Outra visão, porém com a mesma aplicação, pode ser observada em [Witten et al. 2016], onde o aprendizado é caracterizado quando ocorre uma mudança de comportamento de maneira a proporcionar um melhor funcionamento no futuro.

O conceito de AM pode ser sintetizado como a capacidade de um programa de computador aprender com a experiência (E) relacionada a alguma classe de tarefas (T), baseada em uma medida de desempenho (P). Dessa forma, o desempenho em tarefas (T), quando medido por (P), melhora com a experiência em (E) [Mitchell et al. 1997].

Dessa forma, tanto a tarefa a ser realizada quanto a medida de desempenho são dependentes e, muitas vezes específicas do problema em análise. Embora a experiência de aprendizado também seja dependente do problema, ela pode ser classificada segundo diferentes paradigmas. Os três principais são: Aprendizado Supervisionado (do inglês *supervised learning*), Aprendizado Não-Supervisionado (do inglês *unsupervised learning*) e Aprendizado por Reforço (do inglês *reinforcement learning*).

2.4.1 Aprendizado supervisionado

O Aprendizado Supervisionado é uma técnica de AM para deduzir uma função de dados de treinamento onde esses dados consistem em pares de objetos de entrada (tipicamente vetores) e saídas desejadas. A saída da função pode ser um valor contínuo (no problema de regressão), ou pode prever um rótulo de classe do objeto de entrada (no problema de classificação). A tarefa do aprendizado supervisionado é prever o valor da função para qualquer objeto de entrada válido, depois de ter visto um número de exemplos de treinamento. Para conseguir isso, o aprendizado supervisionado tem de generalizar a partir dos dados apresentados para situações não vistas de uma forma "razoável" [Mitchell et al. 1997].

Os modelos de Aprendizado Supervisionado são caracterizados pela capacidade de construir modelos que "aprendem" a partir de observações existentes, replicando esse aprendizado na previsão de observações futuras, prevendo os resultados que sejam de interesse. Os algoritmos utilizados possuem características relacionadas à capacidade de generalização com base em regularidades constatadas a partir de uma determinada base de treinamento, ou seja, utilizam um conhecimento prévio do domínio para orientar a generalização de situações futuras [Luger 2013].

Conforme informado anteriormente, no paradigma de Aprendizado Supervisionado existem duas atividades principais, a Classificação e a Regressão, que estão relacionadas aos tipos de dados utilizados. Tanto a Regressão como a Classificação são problemas de aprendizagem supervisionados onde há uma entrada x com uma saída y resultando em uma tarefa

que é aprender o modelo de mapeamento da entrada para a saída. A abordagem no AM é de assumir um modelo definido até um conjunto de parâmetros da seguinte forma:

$$y = g(x|\theta) \tag{2.1}$$

Desta forma, pode ser visualizado em 2.1 a questão dos tipos de dados para cada atividade onde $g(\cdot)$ é o modelo e θ são seus parâmetros. Dessa forma, y é um número ou um código de classe. O objeto $g(\cdot)$ é a função de regressão ou, na classificação, é a função discriminante que separa as instâncias de classes diferentes. O programa de AM otimiza os parâmetros em θ, de forma que o erro de aproximação seja minimizado, ou seja, que as estimativas sejam tão próximas quanto possível dos valores corretos referenciados no conjunto de treinamento [Alpaydin 2014].

Nas tarefas de classificação, os atributos do conjunto de dados são divididos em dois tipos, os atributos preditivos e o atributo alvo. Os atributos preditivos registram as características (do inglês *features*) ou seja, os atributos que possuem os dados que serão os "influenciadores" para que se classifique em um atributo alvo. O atributo alvo é a característica a qual deseja-se predizer de acordo com as características dos atributos preditivos. Conforme já apresentado, em atividades de classificação, o atributo alvo é categórico com rótulos que representem as classes. A tarefa de classificação consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram. Trata-se de uma abordagem sistemática para construção de modelos de classificação a partir de conjuntos de dados pré-existentes. Existem diversas técnicas que podem ser utilizadas, como os classificadores baseados em árvores de decisão, classificadores baseados em regras, redes neurais artificiais, máquinas de vetores de suporte e classificadores bayesianos [Tan et al. 2009].

Entre as tarefas de classificação, a técnica baseada na construção de árvores de decisão se destaca quando é necessária a identificação de padrões descritivos e preditivos. A árvore de decisão é um modelo de classificação estruturado em forma hierárquica, que é fácil de entender mesmo por usuários leigos e pode ser eficientemente induzido a partir de dados. Trata-se de um modelo de representação de conhecimentos onde cada nó interno representa uma decisão sobre um atributo que determina como os dados estão particionados pelos seus nós filhos. Alguns dos principais métodos de classificação em MDE são baseados na construção de árvores de decisão [Wu et al. 2008].

Árvores de Decisão

Em uma definição generalista, uma árvore de decisão pode ser definida como uma estrutura de dados hierárquica que implementa a estratégia de divisão e conquista. É um método eficiente, que pode ser usado tanto para a classificação quanto para a regressão. Os algo-

ritmos constroem uma árvore a partir de uma dada amostra de treinamento devidamente rotulada com as classes de interesses [Alpaydin 2014].

A estratégia de dividir e conquistar ocorre quando um problema complexo é decomposto em subproblemas mais simples e recursivamente a mesma estratégia é aplicada a cada subproblema. A capacidade de discriminação de uma Árvore de Decisão advém das características de divisão do espaço definido pelos atributos em subespaços e da associação de uma classe a cada subespaço [Frank and Hall 2001]. Em geral, a construção de uma árvore de decisão é realizada de acordo com alguma abordagem recursiva de particionamento do conjunto de dados [Sammut and Webb 2011].

Uma árvore de decisão tem como entrada dados que descrevem um conjunto de propriedades para produzir, por exemplo, uma decisão *booleana* sim ou não. Funções com uma gama maior de classes também podem ser representadas, mas, por simplicidade, aqui se considera o caso booleano. Cada nó interno na árvore corresponde a um teste do valor de uma das propriedades, e os ramos do nó são rotulados com os possíveis valores do teste. Cada nó folha na árvore especifica o valor booleano a ser retornado se essa folha for atingida [Duda et al. 2012].

A figura 2.3 apresenta um modelo clássico de árvore de decisão.

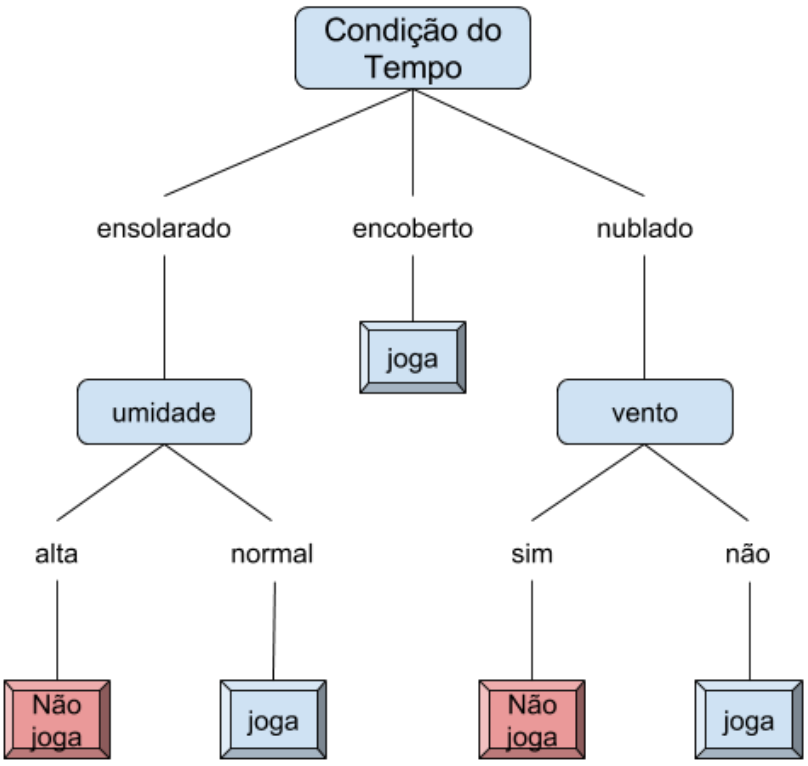


Figura 2.3: Exemplo de árvore de decisão
adaptado de [Quinlan 1986]

Inicialmente, a raiz da árvore considera todo o conjunto de dados com exemplos misturados das várias classes presentes. A partir daí, um predicado, denominado como ponto de

separação, é escolhido como sendo a condição que melhor separa ou discrimina as classes. Um predicado envolve exatamente um dos atributos preditores para o problema em questão, induzindo uma divisão do conjunto de dados em dois ou mais subconjuntos disjuntos, cada um dos quais associado a um nó filho. Cada novo nó abrange, portanto, um subconjunto do conjunto de dados global que é recursivamente separado até que o subconjunto associado a cada nó folha consista, inteira ou predominantemente, de registros de uma mesma classe [Quinlan 1986].

Quando árvores de decisão são construídas, muitas das arestas ou sub-árvores podem refletir ruídos ou erros. Isso pode ocorrer devido ao problema conhecido como sobre-ajuste (do inglês *overfitting*), que significa um aprendizado muito específico do conjunto de treinamento, não permitindo ao modelo generalizar bem. Para detectar e excluir essas arestas e sub-árvores, são utilizados métodos de poda (do inglês *pruning*) da árvore, cujo objetivo é melhorar a taxa de acerto do modelo para novos exemplos que não tenham sido utilizados no conjunto de treinamento [Li et al. 2001].

Algoritmo C4.5

Um dos principais algoritmos de indução de árvores de decisão é o algoritmo C4.5, que representa uma significativa evolução do algoritmo ID3. Esse algoritmo possui a capacidade de lidar tanto com atributos categóricos (ordinais ou não-ordinais) como com atributos contínuos. Para lidar com atributos contínuos, o algoritmo define um limiar e então divide os exemplos de forma binária: aqueles cujo valor do atributo é maior que o limiar e aqueles cujo valor do atributo é menor ou igual ao limiar [Quinlan 1993].

O C4.5 permite que os valores desconhecidos para um determinado atributo (em inglês *missing values*) sejam representados com um sinal de '?', de forma que o algoritmo trate esses valores de forma especial, não utilizando-os nos cálculos de ganho e entropia. Utiliza a medida de razão de ganho (*Gain Ratio*) para selecionar o atributo que melhor divide os exemplos. Essa medida se mostrou superior ao ganho de informação (*info gain*), gerando árvores mais precisas e menos complexas [Quinlan 1993].

Para calcular o índice da Razão de Ganho é necessário encontrar o valor da Entropia que, segundo a Teoria da Informação [Cover and Thomas 2012], mede a qualidade do dado em relação aos atributos a partir da expressão matemática apresentada na Equação 2.2.

$$Entropia(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus} \tag{2.2}$$

Nessa fórmula, o cálculo da Entropia do conjunto de dados representado por S leva em consideração a proporção de exemplos positivos $p_{\oplus}$ e a proporção de exemplos negativos $p_{\ominus}$.

Em seguida é calculado o ganho de informação que utiliza a expressão apresentada na

Equação 2.3.

$$Ganho(S, A) \equiv Entropia(S) \; - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{2.3}$$

Para casos onde o conjunto de dados é separado em partições para validação, o valor de $SplitInfo$ representa a informação potencial gerada, dividindo o conjunto de dados de treinamento D em v partições, correspondendo a v resultados no atributo A conforme 2.4.

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{|D_j|}{|D|}) \tag{2.4}$$

O critério de razão de ganho (*Ganho*) seleciona atributos de acordo com a razão entre seu ganho e seu conteúdo de informação intrínseca, ou seja, a quantidade de informação contida na resposta à pergunta: "Qual é o valor desse atributo?" O critério de razão de ganho, portanto, tenta medir com que eficiência um atributo fornece informações sobre a classificação correta de um exemplo. $Ganho(S, A)$ = redução esperada da Entropia devido à classificação de A onde:

$$Ganho(S, A) \equiv Entropia(S) \; - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{2.5}$$

$$GainRatio(D, S) = \frac{Ganho(D, S)}{H(\frac{|D_1|}{|D|}, ... \frac{|D_s|}{|D|})} \tag{2.6}$$

Considerando a abordagem baseada na Teoria da Informação [Cover and Thomas 2012] apresentada nas Equações, o algoritmo C4.5 utiliza a característica de busca de baixo para cima, transformando em nós folha aqueles ramos que não apresentam nenhum ganho significativo ou seja, produzindo árvores de decisão a partir de uma abordagem recursiva [Quinlan 1996]. Isso pode ser observado no Algoritmo 2.1.

A estratégia de indução da árvore no algoritmo C4.5 pode ser classificada como "gulosa", pois executa sempre o melhor passo avaliado localmente, sem se preocupar se este passo, junto à sequência completa de passos, vai produzir a melhor solução ao final. Como informado anteriormente, utiliza a técnica de "Dividir para conquistar" onde, partindo da raiz, criam-se sub-árvores até chegar nas folhas, o que implica em uma divisão hierárquica em múltiplos subproblemas de decisão, que tendem a ser mais simples que o problema original [Quinlan 1993].

Métodos de poda

Existem diversas formas de realizar a poda em uma árvore de decisão, que devem ser classificadas como pré-poda ou pós-poda [Quinlan 1986].

Algoritmo 2.1 C4.5

Require: Um conjunto de dados com atributos em D
  if D é "puro" OU outros critérios de parada foram cumpridos **then**
    terminar
  end if
  for all atributo $a \in D$ **do**
    Calcule os critérios da teoria da informação se dividirmos em a
  end for
  a_{best} = Melhor atributo de acordo com os critérios calculados
  $Tree_v$ = Criar um nó de decisão que teste a_{best} na raiz
  D_v = Sub-conjuntos de dados induzidos de D com base em a_{best}
  for all D_v **do**
    $Tree_v$ = C4.5(D_v)
    Anexe $Tree_v$ ao ramo correspondente da Árvore
  end for
  return $Tree$

O método pré-poda é realizado durante o processo de construção da árvore, onde simplesmente para-se de dividir o conjunto de elementos e se transforma o nó corrente em um nó folha da árvore. O ganho de informação, por exemplo, pode ser utilizado como critério de poda. Caso todas as divisões possíveis utilizando um atributo A gerem ganhos menores que um valor pré-estabelecido, então esse nó vira folha, representando a classe mais frequente no conjunto de exemplos.

Já o pós-poda é realizado após a construção da árvore de decisão, removendo ramos completos, onde tudo que está abaixo de um nó interno é excluído e esse nó é transformado em folha, representando a classe mais frequente no ramo. Para cada nó interno da árvore, o algoritmo calcula a taxa de erro caso a sub-árvore abaixo desse nó seja podada. Em seguida, é calculada a taxa de erro caso não haja a poda. Se a diferença entre as duas grandezas for menor que um valor pré-estabelecido, a árvore é podada. Caso contrário, não ocorre a poda.

Empregar critérios de parada severos incorre no risco de criar árvores de decisão pequenas e sub-equipadas. Por outro lado, o uso de critérios de parada mais flexíveis tende a gerar grandes árvores de decisão sobre-ajustadas ao conjunto de treinamento. Métodos de poda sugeridos originalmente em [Breiman et al. 1984] foram desenvolvidos para resolver este dilema. De acordo com esta metodologia, um critério de parada flexível permite que a árvore de decisão sobreponha o conjunto de treinamento. Em seguida, a árvore sobre-ajustada é cortada para trás em uma árvore menor, removendo sub-ramos que não estão contribuindo para a precisão de generalização. Foi demonstrado em vários estudos que empregar métodos de poda pode melhorar o desempenho de generalização de uma árvore de decisão, especialmente em domínios ruidosos [Rokach and Maimon 2014].

O parâmetro denominado fator de confiança (do inglês *confidence factor*) é usado em árvores de decisão como fator de poda. Com um fator de confiança maior, menos poda na árvore é realizada, tendendo ao sobre-ajuste dos exemplos de treinamento. Com um fator

de confiança mais baixo, mais poda é realizada, resultando em uma árvore menor e mais generalizada [Quinlan 1993].

Existem outras formas de avaliação como o Valor Preditivo Positivo (VPP) e Valor Preditivo Negativo (VPN) que são altamente suscetíveis em situações de desbalanceamento de classes, podendo facilmente induzir a uma conclusão errada sobre o desempenho dos sistemas [Kohavi and Provost 1998].

2.4.2 Aprendizado não-supervisionado

Diferentemente dos algoritmos de aprendizado supervisionado, que assumem a existência de um ”professor” ou de uma medida de adequação para classificação de exemplos de treinamento, os algoritmos de aprendizado não-supervisionado eliminam a existência dessa medida de referência e requerem que o próprio algoritmo de aprendizado avalie os conceitos envolvidos, por meio de observação e descoberta [Duda et al. 2012].

O aprendizado não-supervisionado não possui a vantagem de um ambiente de treinamento com casos para calibração de um modelo de classificação; em vez disso, os algoritmos não supervisionados propõem hipóteses para explicar as observações. Os algoritmos avaliam as hipóteses usando critérios como simplicidade, generalidade e performance, para testar hipóteses por meio de experimentos que os próprios algoritmos concebem em sua abordagem computacional [Luger 2013].

No aprendizado não-supervisionado, o algoritmo k-means é conhecido por ser um dos mais utilizados em tarefas com essas características, pois é um algoritmo simples e que proporciona resultados efetivos em diversas aplicações [Wu et al. 2008].

Algoritmo k-means

Proposto por [MacQueen et al. 1967], o algoritmo de Análise de Agrupamento k-means é um dos mais conhecidos e utilizados, além de ser o que possui o maior número de variações. Trata-se de um método de agrupamento simples e efetivo onde é possível, por exemplo, comprovar o processo de minimização da distância quadrática total de cada ponto de um grupo, em relação ao centroide de referência. Após a estabilização das interações, cada ponto estará atribuído ao centroide mais próximo e consequentemente ocorre um efeito generalizado de minimização da distância quadrática total de todos os pontos aos seus centros. Contudo, não existem garantias de que o método encontre essa generalização, sendo necessário reiniciá-lo diversas vezes com diferentes pontos de partida (centroides), escolhendo o melhor resultado com a menor distância quadrática total [Witten and Frank 2005].

Um exemplo de interação com o k-means, utilizando um grupo de números aleatórios, pode ser observado na Figura 2.4. A figura demonstra, de forma sintetizada, um exemplo utilizando a lógica do algoritmo k-means para formar dois agrupamentos, considerando a

população composta pelos elementos em $\{2, 6, 9, 1, 5, 4, 8\}$. Inicialmente, foram escolhidos como sementes os dois primeiros elementos e, como critério para definir o valor do centroide após a união, foi usada a média. Ao final, os elementos $c1$ e $c2$ apresentam os valores dos centroides de cada um dos agrupamentos após a adição de um novo elemento.

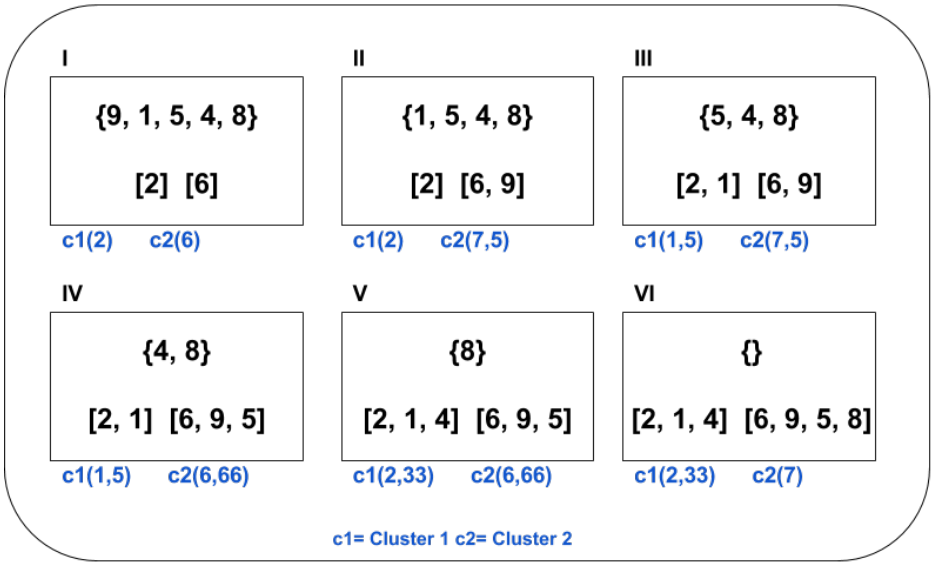


Figura 2.4: Exemplo de execução do algoritmo *k-means*
adaptado de [Prass et al. 2004]

O algoritmo apresenta características de escalabilidade e confiança de uma forma geral, mas possui algumas limitações. Exige que as variáveis sejam numéricas ou binárias onde as aplicações frequentemente envolvem dados categorizados. Para esses casos, uma alternativa é converter os dados categorizados em valores numéricos. Outra limitação está relacionada com a sensibilidade do algoritmo para tratamento de valores discrepantes (*outliers*). Um único objeto com valor muito extremo pode modificar substancialmente a distribuição dos dados [Huang 1997].

2.4.3 Aprendizado por reforço

O terceiro paradigma da aprendizagem, baseado em reforço, aborda a questão de como um agente autônomo, que detecta e age em seu ambiente, pode aprender a escolher as melhores ações para atingir seus objetivos. Este problema muito genérico cobre tarefas como aprender a controlar um robô móvel, aprender a otimizar operações em fábricas e aprender a jogar jogos de tabuleiro. Cada vez que o agente executa uma ação em seu ambiente, um instrutor pode fornecer uma recompensa ou penalidade para indicar a conveniência do estado resultante. Por exemplo, ao treinar um agente para jogar um jogo o instrutor pode fornecer uma recompensa positiva quando o jogo é ganho, negativa quando perder e a recompensa zero em todos os outros estados. A tarefa do agente é aprender com essa recompensa indireta, atrasada, para escolher sequências de ações que produzam a maior recompensa de forma acumulativa [Mitchell et al. 1997].

No aprendizado por reforço, o processo de ajuste dos parâmetros é feito pela interação contínua com o ambiente para minimizar (ou maximizar) um determinado índice de desempenho. Assim, não há um supervisor indicando a saída esperada a cada estímulo fornecido como entrada, mas sim uma espécie de "crítico" que atribui uma nota para a resposta da máquina de aprendizado ao estímulo, com o objetivo de alcançar o nível máximo de sucesso no seu funcionamento com base em um índice estabelecido [Kaelbling et al. 1996].

Na aprendizagem por reforço, o agente aprende com uma série de reforços - recompensas ou punições. Por exemplo, a falta de gorjeta no final da viagem dá ao agente de táxi uma indicação de que fez algo errado. O ponto para uma vitória no final de um jogo de xadrez diz ao agente que fez algo certo. Cabe ao agente decidir quais das ações anteriores ao reforço foram as mais responsáveis pelo caminho correto [Russell et al. 1995].

2.4.4 Medidas de avaliação e desempenho

Quando um modelo de AM é utilizado, é necessário que hajam parâmetros que permitam testar a confiabilidade e a performance em relação aos resultados obtidos. Existem diversas metodologias para validações e testes, segundo [Monard and Baranauskas 2003, Witten et al. 2016]: validação cruzada (do inglês *cross validation*), *holdout*, amostragem aleatória, entre outras.

Na validação cruzada, os exemplos são aleatoriamente divididos em r partições mutuamente exclusivas (*folds*) de tamanho aproximadamente igual de exemplos. Os exemplos nos $(r-1)$ são usados para treinamento e a hipótese induzida é testada no remanescente. Este processo é repetido r vezes, cada vez considerando um diferente para teste. O desempenho na validação cruzada é a média dos desempenhos calculados em cada um dos r.

Na metodologia de *holdout*, dividem-se os exemplos em uma porcentagem fixa, geralmente em $70/30$, ou seja, 70% dos dados para treinamento e 30% para testes.

Com a amostragem aleatória as hipóteses são induzidas a partir de cada conjunto de treinamento onde o desempenho final é calculado como a média dos desempenhos de todas as hipóteses induzidas e calculadas em conjuntos de teste independentes extraídos aleatoriamente.

Para a avaliação dos resultados, são utilizadas métricas como: matriz de confusão, índice de correção e incorreção de instâncias mineradas, estatística kappa, erro médio absoluto, erro relativo médio, precisão, F-measure, dentre outros.

Em problemas multiclasse (duas ou mais classes), o resultado em um conjunto de teste é frequentemente exibido como uma matriz de confusão bidimensional, com uma linha e coluna para cada classe. Cada elemento da matriz mostra o número de exemplos de teste para os quais a classe real é a linha e a classe prevista é a coluna. Bons resultados correspondem a grandes números na diagonal principal e pequenos, idealmente zero, valores fora da diagonal [Witten et al. 2016]

A matriz de confusão na Figura 2.5(a) disponibiliza métricas relevantes para medir o desempenho de um algoritmo. Essas métricas baseiam-se em quatro possíveis resultados relacionados à assertividade da previsão de classes. Os Positivos Verdadeiros (do inglês *True Positive*) (TP), que são os itens classificados como verdadeiros e realmente são verdadeiros, Verdadeiros Negativos (do inglês *True Negative*) (TN), que são os itens classificados como negativos e que são realmente negativos, os Falsos Positivos (do inglês *False Positive*) (FP), que são os itens classificados como positivos e que são falsos e os itens classificados como Falsos Negativos (do inglês *False Negative*) (FN), que são os itens classificados como negativo mas na verdade são positivos [Monard and Baranauskas 2003].

Classe Predita / Classe Verdadeira	Aprovados (X)	Reprovados (Y)	Evadidos (Z)
Aprovados (X)	26	4	3
Reprovados (Y)	12	43	4
Evadidos (Z)	4	54	33

(a) Matriz de Confusão

Aprovados (X)	26	TP
	76	TN
	65	FN
	16	FP

(b) Análise da Classe Aprovados (X)

Figura 2.5: Exemplo de matriz de confusão para análise de previsão de resultados em EAD

Na Figura 2.5(b) são apresentados com legendas os índices de classificação isolados para a classe Aprovados (X). Analisando a estrutura da matriz de confusão, é possível identificar que o item destacado com a cor azul representa os itens classificados como **TP**. Na coluna de Classe Predita em (X), estão marcados com a cor laranja os itens classificados como **FP**. Analisando a diagonal partindo de (X), estão destacados os itens em laranja que representam os itens classificados como **TN** e, por final, em amarelo estão os ítens classificados como **FN**.

2.5 MINERAÇÃO DE DADOS EDUCACIONAIS

A área de pesquisa relacionada à Mineração de Dados Educacionais foi definida por [Baker and Yacef 2009] da seguinte forma:

> "*uma disciplina emergente, preocupada com o desenvolvimento de métodos para exploração dos tipos de dados únicos provenientes dos ambientes educacionais e como utilizar esses métodos para entender melhor os alunos e as características de como eles aprendem.*"

Em 2009, foi lançado o primeiro volume da Revista de Mineração de Dados Educacionais *(Journal of Educational Data Mining*[1]*)*, publicado pela recém formada Sociedade Internacional de Mineração de Dados Educacionais (*International Society of Educational Data Mining*[2]). Neste volume, o trabalho de [Baker and Yacef 2009] apresenta uma revisão do estado da arte e uma visão sobre as tendências futuras através de uma análise de trabalhos relacionados às técnicas de mineração de dados em ambientes educacionais.

No Brasil, o trabalho de [Baker et al. 2011b] registrou as possibilidades e oportunidades de aplicação dos conceitos de MDE. Este trabalho pode ser considerado como um grande empurrão para a aplicação de técnicas de *Data Mining* em ambientes educacionais, em especial na oferta da educação em modalidade à distância.

Conforme observado no trabalho de [Costa et al. 2013], a MDE pode ser definida como uma área emergente que procura desenvolver, aplicar e adaptar métodos de Descoberta de Conhecimentos em Bases de Dados (KDD) com o intuito de identificar modelos de conhecimentos a partir das grandes bases de dados que são geradas pelos ambientes educacionais utilizados na EAD.

A literatura relacionada à MDE aborda a aplicação de técnicas, entre elas, a classificação, a regressão e o agrupamento de dados, para o tratamento da grande quantidade de dados que são gerados a partir da utilização da EAD, ou seja, os dados gerados pelos usuários ao acessarem ambientes educacionais e ao utilizarem as ferramentas e os meios de interação aplicados [Romero 2010, Baker et al. 2011a].

A MDE é um campo multidisciplinar que explora os diferentes tipos de dados provenientes de ambientes educacionais, sendo que o principal objetivo é a análise destes dados para a resolução de problemas relacionados, os quais envolvem diferentes grupos de usuários ou participantes que possivelmente enxergam as informações educacionais de uma forma singular [Romero 2010].

Conforme observado no trabalho realizado por [Romero and Ventura 2013], a MDE reúne diversas áreas de pesquisa em suas aplicações, conforme pode ser observado na Figura 2.6. Essas áreas se relacionam em torno das atividades envolvidas com MDE.

[1] Journal of Educational Data Mining, Article 1, Vol 1, No 1

[2] International Society of Educational Data Mining - http://www.educationaldatamining.org/

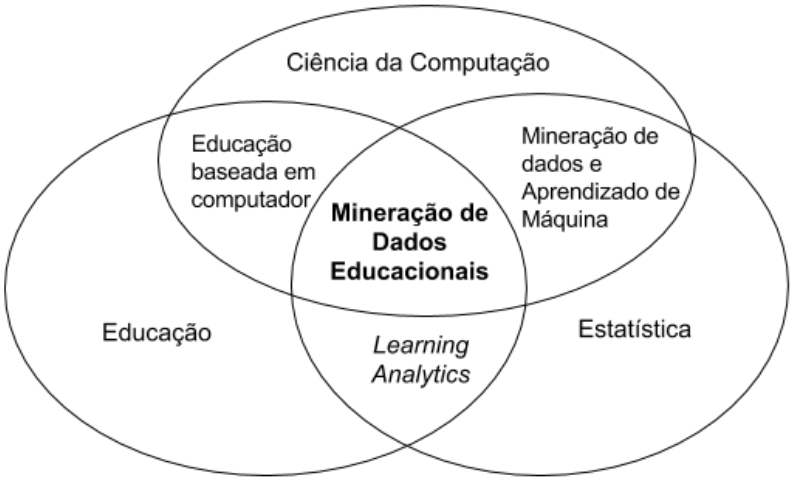


Figura 2.6: Áreas envolvidas com a MDE
Adaptado de [Romero and Ventura 2013]

A MDE está relacionada à Estatística, *Learning Analytics*, Educação, Educação baseada em computadores, Ciência da Computação, além de Mineração de dados e Aprendizado de Máquina.

2.5.1 Modelos em MDE

Conforme apresentado anteriormente na Seção de Mineração de Dados, na MDE existem dois tipos de modelos básicos que norteiam os projetos de acordo com os objetivos definidos para seu sucesso: os modelos preditivos e os modelos descritivos [Fayyad et al. 1996b].

Nos modelos preditivos, as tarefas objetivam predizer o valor de um determinado atributo (variável) baseado nos valores de outros atributos. O atributo a ser predito é comumente conhecido como a variável preditiva, dependente ou alvo, enquanto os atributos usados para fazer a predição são conhecidos como as variáveis preditoras, independentes ou explicativas. Um exemplo de aplicação em MDE é a criação de modelos preditores para auxiliar na previsão do desempenho dos alunos no combate à evasão [Fayyad et al. 1996c].

Os modelos descritivos são caracterizados por tarefas utilizadas para descrever os padrões e tendências revelados pelos dados. A descrição geralmente oferece uma possível interpretação para os resultados obtidos. A tarefa de descrição é muito utilizada em conjunto com as técnicas de análise exploratória de dados, para comprovar a influência de certas variáveis no resultado obtido. Tarefas Descritivas procuram encontrar padrões (correlações, tendências, grupos, trajetórias e anomalias) que descrevam os dados. Um exemplo de aplicação em MDE é a possibilidade de análise do nível de influência de uma determinada característica (*feature*) dos alunos, descrevendo como acontecem os relacionamentos para alcançar uma determinada classe de nota ao final do curso [Fayyad et al. 1996c].

2.5.2 Tarefas em MDE

A partir dos Modelos apresentados anteriormente, a MDE possui diversas tarefas para aplicação de algoritmos para descoberta de modelos de mineração provenientes de dados educacionais. Essas tarefas foram classificadas nos estudos realizados por [Baker et al. 2010, Baker et al. 2011a] que propuseram uma taxonomia, conforme segue:

- Predição
  - Classificação
  - Regressão
- Agrupamento (*Cluster*)
- Mineração de Relações
  - Mineração de Regras de Associação
  - Mineração de Correlações
  - Mineração de Padrões Sequenciais
  - Mineração de Causas
- Destilação de Dados para facilitar discussões humanas
- Descoberta com Modelos

Classificação

Conforme informado anteriormente, na MDE existem dois tipos de técnicas de predição que são mais utilizadas: a Classificação e a Regressão. Ambas são utilizadas na análise preditiva porém, nas tarefas de regressão os valores são numéricos ou contínuos enquanto na classificação o atributo alvo é caracterizado por ser uma classe nominal.

A Figura 2.7 representa o funcionamento de um modelo classificador, que tem como entrada um conjunto de treinamento, que consiste de um conjunto de amostras (ou instâncias) de dados onde a classe já é conhecida. A partir desse conjunto de dados, um processo de aprendizado supervisionado induz um modelo classificador que, em seguida, é testado junto a um conjunto de testes, que consiste de um conjunto de amostras cujas classes são ocultas/-desconhecidas e precisam ser preditas a partir do modelo de treinamento [Costa et al. 2013].

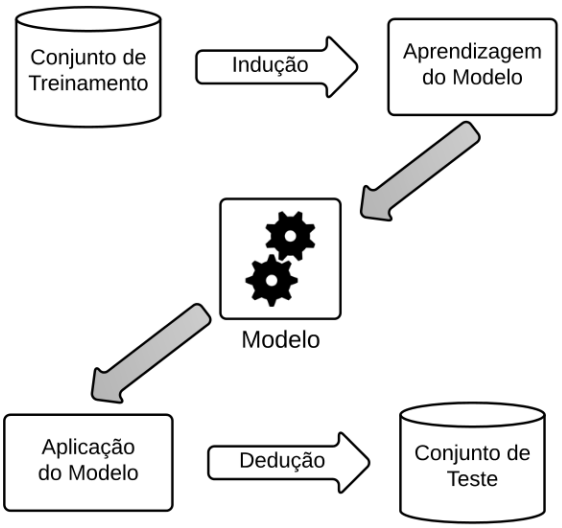


Figura 2.7: Elementos de um modelo classificador
Adaptado de [Costa et al. 2013]

Reforçando o conceito já mencionado anteriormente, vale ressaltar que, entre as técnicas de classificação, destacam-se as Árvores de Decisão, Redes Neurais (do inglês *Neural Network*), Máquina de Vetor-Suporte (do inglês *support vector machine*) entre outras.

Regressão

O objetivo da análise de regressão é determinar os valores de parâmetros para uma função que fazem com que a função se ajuste melhor a um conjunto de observações de dados fornecida. A Equação 2.6 expressa essas relações em símbolos, caracterizando a regressão como o processo de estimar o valor de um alvo contínuo (y) como uma função (F) de um ou mais preditores $(x1, x2, ..., xn)$, um conjunto de parâmetros $(\theta1, \theta2, ..., \theta n)$ e uma medida do erro (e) [Fayyad et al. 1996a].

$$Y = F(x, \theta) + e \tag{2.7}$$

A regressão ajuda a identificar o comportamento de uma variável quando outras variáveis são alteradas no processo. Em suma, quando a intenção é atribuir objetos a diferentes categorias, então usamos algoritmos de classificação e quando queremos predizer valores futuros, então usamos algoritmos de regressão. Existem diversas formas de executar tarefas de regressão, entre elas a Regressão Linear, Regressão Não-Linear e Regressão Multi Variada [Tan et al. 2009].

Agrupamento (Cluster)

Em agrupamento em *clusters*, a atividade principal trata na busca por dados que se agrupem naturalmente, classificando-os em diferentes grupos e/ou categorias de acordo com características presentes. Estes grupos e categorias não são conhecidos inicialmente e através das técnicas de agrupamento, os grupos/categorias são automaticamente identificados através da manipulação das características presentes nos dados. É possível criar esses grupos/categorias utilizando diferentes unidades de análise, por exemplo é possível achar grupos de

escolas com o objetivo de investigar as diferenças e similaridades entre elas, achar grupos de alunos para investigar as diferenças e similaridades entre eles ou até grupos de ações para investigar os padrões de comportamento dos alunos [Romero 2010].

Regras de Associação

A tarefa de Regras de associação, que também pode ser denominada como Descoberta de Associações, consiste em encontrar subconjuntos de itens que ocorrem de forma simultânea e frequente em uma fração mínima e previamente estabelecida do conjunto de dados. Pode ser formalmente definida como a tarefa de busca por regras de associação frequentes e válidas em um conjunto de dados, a partir da especificação dos parâmetros de suporte e confiança mínimos [Agrawal et al. 1993].

Correlações

Em mineração de correlações, a meta 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 negativa entre uma variável que indica a quantidade de tempo que um aluno passa externalizando comportamentos que não estão relacionados as tarefas passadas pelo professor (e.g. conversas paralelas, brincadeiras e outras perturbações que ocorrem em sala de aula) e a nota que este aluno recebe na próxima prova [Baker et al. 2011a].

Padrões Sequenciais

Em mineração de sequências, o objetivo 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 aluno 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 alunos externalizem ações que vão ajudá-los a construir seu conhecimento e desenvolver as habilidades necessárias para trabalhar com o conteúdo apresentado pelo professor [Baker et al. 2010].

Mineração de Causas

Em mineração de causas, desenvolvem-se algoritmos e técnicas para verificar se um evento causa outro evento através da analise dos padrões de covariância. Por exemplo, se considerarmos o exemplo onde um aluno externaliza comportamentos inadequados que não contribuem para resolver a tarefa dada pelo professor, o aluno, em muitos casos, recebe uma nota ruim na prova final. O comportamento do aluno pode ser a causa dele não aprender e, assim, resultando em uma performance ruim na prova. Contudo, pode ser que o aluno externalize tal comportamento inadequado devido a 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 aluno. Analisando o padrão de covariância, a mineração de causa pode inferir qual evento foi a causa do outro [Baker et al. 2011a].

Destilação de Dados

A área de Destilação de Dados visa facilitar decisões humanas realizando pesquisas que tem como objetivo apresentar dados complexos de forma a facilitar sua compreensão e expor suas características mais importantes. Através da destilação, é possível que os dados sejam utilizados para inferir aspectos e tomar decisões que, anteriormente, não poderiam ser tomadas e nem automatizadas apenas com o uso dos métodos de MDE. Os métodos dessa sub-área da MDE facilitam a visualização da informação contida nos dados educacionais que são coletados por softwares específicos. Tais métodos purificam os dados para auxiliar a identificação de padrões. Diferente de outras técnicas, os padrões são previamente conhecidos, mas são difíceis de serem visualizados ou descritos formalmente. O uso da destilação de dados também é muito útil para categorizar as ações dos estudantes, o que possibilita o desenvolvimento de um modelo de predição mais robusto [Baker et al. 2009].

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. Um exemplo clássico é a utilização de técnicas de clusterização para compor modelos de classificação onde os agrupamentos gerados pela clusterização tornam-se atributos preditores para atividades de classificação [de Souza Mendes et al. 2014].

2.5.3 Métodos e aplicações

Na MDE, existem diversos métodos que podem ser utilizados de acordo com os objetivos previstos para a mineração e também sobre as características dos dados. Os principais métodos abordam Árvores de decisão, Classificadores Baseados em Regras, Classificadores Bayesianos, Classificadores K-NN (vizinho mais próximo), Redes Neurais Artificiais, entre outros [Baker et al. 2010].

As possibilidades de aplicação das técnicas de MDE são bastante abrangentes, sendo, segundo [Baker and Yacef 2009], as principais aplicações:

Modelagem do estudante, que se refere ao estudo dos diferentes modelos cognitivos relacionados aos alunos como: emoções, cognição, conhecimento de domínio, estratégias de aprendizagem, realizações, características, preferências e habilidades de aprendizagem, avaliação e estado afetivo. Nessa abordagem, o objetivo principal é a representação das características do usuário, adaptando as experiências de ensino às necessidades específicas de aprendizagem [Peña-Ayala 2014].

Identificar as diferenças existentes entre os estudantes possibilita o acompanhamento do aprendizado de forma personalizada. Um exemplo seria a modelagem das características dos alunos em Sistemas de Tutoria Inteligente (ITS) (do inglês, *Intelligent Tutor System*).

Através de técnicas de MDE, é possível modelar atributos do estudante para detectar comportamentos inadequados, verificando se o estudante está "trapaceando com o sistema". Por exemplo, o estudante pede diversas dicas ao STI somente para descobrir a resposta de um determinado problema [Romero and Ventura 2013].

Modelagem do domínio, que abrange uma área importante na utilização de técnicas de MDE para a descoberta de modelos que representem a estrutura de um domínio, ou seja, um modelo que reúna características presentes em uma determinada aplicação. Por meio da combinação de arcabouços da modelagem de psicometria com algoritmos de espaço de busca, alguns trabalhos têm conseguido desenvolver abordagens automáticas de descoberta que, a partir de dados, conseguem identificar as características presentes na estrutura dos dados analisados [Costa et al. 2013].

Suporte pedagógico, que estuda a descoberta de conhecimentos tanto em softwares de apoio à aprendizagem quanto em outros domínios, como a aprendizagem colaborativa, a modelos pedagógicos que sejam mais eficientes para grupos específicos de estudantes. Trata-se de uma tarefa com maior complexidade, devido à necessidade de análise em domínios específicos que consideram as particularidades dos alunos relacionadas às particularidades dos modelos pedagógicos em questão [Baker and Yacef 2009].

Descoberta científica está focada na exploração e confirmação de teorias científicas educacionais, proporcionando uma melhor compreensão dos fatores que impactam no processo de aprendizagem, procurando desenvolver melhores sistemas de apoio ao ensino e à aprendizagem [Costa et al. 2013].

Capítulo 3

TRABALHOS RELACIONADOS

Este capítulo apresenta o levantamento de artigos científicos, trabalhos de mestrado e doutorado que contribuíram no processo de pesquisa para realização dos experimentos propostos no Capítulo 4.

3.1 MINERAÇÃO DE DADOS EDUCACIONAIS

Conforme apresentado no Capítulo 2, a Mineração de Dados Educacionais é uma área de pesquisa relativamente nova que, a partir de 2008, inicia-se com trabalhos científicos relacionados à descoberta de conhecimentos em ambientes educacionais. Antes de sua existência, os autores utilizavam eventos relacionados à Inteligência Artificial para publicação de pesquisas no tema. Um trabalho relevante publicado antes da existência da área específica de pesquisa foi realizado por [Romero and Ventura 2007], onde é apresentado um levantamento do estado da arte entre os anos de 1995 a 2005.

Na primeira edição da Revista de Mineração de Dados Educacionais, publicada, no ano de 2009, [Baker and Yacef 2009] apresentaram outra revisão do estado da arte em MDE, apresentando visões futuras sobre a utilização dessas técnicas.

No Brasil, a publicação realizada por [Baker et al. 2011a] apresenta de forma elucidativa as possibilidades e técnicas para aplicação de MDE no cenário da educação brasileira. Nesse artigo, os autores apresentam uma proposta de taxonomia com as possíveis aplicações de técnicas clássicas da Mineração de Dados, quando consideradas as particularidades dos cenários e tipos de dados envolvidos com a EAD. Na MDE, as técnicas passíveis de aplicação têm sido frequentemente utilizadas para fornecer suporte e mensagens de *feedback* a professores, recomendações a estudantes, identificação de grupos de estudantes com características comuns e para previsão de desempenho ou risco de evasão.

O artigo de [Romero and Ventura 2007] apresenta um levantamento do estado da arte em relação à pesquisa MDE. Este trabalho apresentou o grande aumento do interesse na pesquisa relacionada com a aplicação de tais técnicas. Os autores registram a evolução dos sistemas

utilizados para a educação, desde os modelos tradicionais baseados em salas de aula até a evolução da utilização de sistemas de tutoria automática baseada em inteligência artificial e também como a aplicação das técnicas de DM podem gerar conhecimentos e auxílio no processo educacional. Os autores propõem um ciclo interativo de atividades de DM em ambientes educacionais, conforme Figura 3.1.

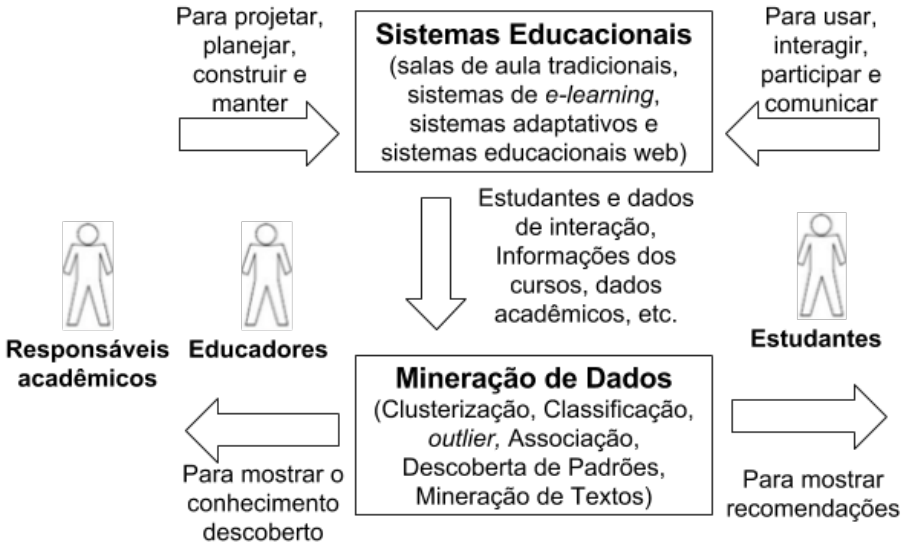


Figura 3.1: Ciclo de técnicas de DM
Adaptado de [Romero and Ventura 2007]

Pode-se observar que as técnicas de DM propostas para utilização em ambientes educacionais são a clusterização, a classificação, a identificação de *outlier*, as regras de associação e a mineração de textos.

Através do trabalho realizado por [Peña-Ayala 2014], é possível identificar que o estudo da aplicação de técnicas de MDE está altamente concentrado em cenários relacionados a um tipo específico de instituição de ensino, as IES. A metodologia de oferta em EAD das IES está focada, naturalmente, nos cursos que tais instituições oferecem, graduação, pós-graduação e especialização. Esses trabalhos possuem características específicas em relação à metodologia em que o ensino é ofertado, tais como: informações pré-acadêmicas dos alunos, duração dos cursos, informações sobre indicadores econômicos e variáveis relacionadas a outras atividades das instituições.

No trabalho realizado por [Baruque et al. 2007], foi proposta a criação de um *Data Mart*[1], que consistiu em um processo de engenharia reversa da base de dados do Moodle para possibilitar uma melhor compreensão dos relacionamentos existentes entre as entidades. Este tipo de solução viabiliza o isolamento de dados direcionados para análises específicas, tendo como base a construção de esquemas em estrela para registrar fatos que podem ser analisados de acordo com dimensões específicas. Porém, esse trabalho fornece somente uma

[1] *Data Mart* é um repositório de dados projetado para atender uma determinada área de conhecimento. Um sub-conjunto de dados de um repositório (Data Warehouse) [Kimball and Ross 2011].

visão sobre as possibilidades de análises em cenários de *Data Marts*, ao contrário do exposto nesta dissertação, que utiliza uma consulta direta na base relacional para criar uma tabela com todos os registros de interação ocorridos em um grupo de cursos específico do AVA Moodle.

Pode-se observar no trabalho de Romero [Romero et al. 2008], uma proposta de criação de uma tabela de sumarização dos dados referentes aos registros das atividades no AVA. Porém, diferentemente dessa abordagem proposta em 2008, o experimento realizado nesta dissertação de mestrado avaliou inicialmente quais ações de cada um dos módulos estão presentes nos cursos analisados. Após essa análise, foram elencados quais módulos e atividades deviam compor a base de dados a ser utilizada no projeto de mineração. Esse artigo serve de base para a proposta dos experimentos realizados e apresentados no Capítulo 4.

3.2 ALGORITMOS DE CLASSIFICAÇÃO NA MDE

Em relação às tarefas mais utilizadas para mineração de dados provenientes de aplicações educacionais, o estudo de [Peña-Ayala 2014], que analisou 242 trabalhos entre 2010 a 2013, apontou que a classificação foi o tipo de tarefa mais considerado nos estudos, com 42,15% dos trabalhos, seguida por agrupamento (26,86%), regressão (15,29%) e regras de associação (6,61%). As demais tarefas juntas atingiram 9,19% dos estudos. Em [Márquez-Vera et al. 2016] também são descritos registros onde as tarefas de classificação são as mais utilizados para análise e previsão de resultados e identificação das causas de evasão em ambientes de EAD.

Entre as técnicas de classificação, existem duas possibilidades de utilização de algoritmo,s que podem ser classificados como *White-Box*(caixa branca) e *Black-Box*(caixa preta). Essas abordagens de aplicação estão diretamente relacionadas a como acontece a saída dos dados para interpretação dos algoritmos. A seguir, serão apresentadas as particularidades de cada uma dessas abordagens.

Um modelo classificador apropriado para um ambiente educacional deve ser preciso e compreensível para que os instrutores e administradores de cursos possam usá-lo para a tomada de decisões [Romero et al. 2013a].

Os algoritmos baseados em técnicas ”caixa branca” (do inglês *white box*) fornecem modelos que podem ser facilmente compreendidos por seres humanos e usados diretamente no processo de tomada de decisão. A utilização de modelos baseados nesse tipo de abordagem pode ser observado em diversos trabalhos na literatura relacionada à MDE, como em [Romero et al. 2013b] e [Marquez-Vera et al. 2013]. Tais algoritmos atuam na geração de modelos que fornecem uma explicação para os resultados das atividades de classificação, servindo como suporte ao processo de tomada de decisão.

Em contextos educacionais focados no entendimento dos padrões extraídos, os mode-

los *white box*, como as árvores de decisão, são preferíveis aos modelos de caixa preta (do inglês *black box*), como as redes neurais, que são mais precisos porém, menos compreensíveis. Técnicas de visualização também são muito úteis para mostrar resultados de uma maneira que seja mais fácil de interpretar. Por exemplo, é melhor mostrar apenas um subconjunto de regras de associação em formato gráfico em vez de mostrar todas as regras descobertas (normalmente centenas ou milhares) em um formato de texto tradicional [Romero and Ventura 2013].

A utilização de modelos *black-box* geralmente possibilitam melhores resultados em relação à quantidade de acertos para a análise preditiva em ambiente educacionais. Porém, a utilização desses modelos, como por exemplo, Redes Neurais Artificiais e Árvores Aleatórias, não fornece uma explicação para o resultado da classificação, não sendo tipicamente utilizados diretamente para a tomada de decisão [Marquez-Vera et al. 2013]

3.2.1 Árvores de decisão

Em comparação com outras técnicas, os algoritmos de árvore de decisão são mais poderosos para analisar a relação entre variáveis independentes e variáveis dependentes devido ao esquema de busca em árvore [Barros et al. 2012].

A árvore de decisão pode ser considerada com a técnica de classificação supervisionada mais amplamente aplicada em ambientes de dados educacionais. As etapas de aprendizagem e classificação da indução da árvore de decisão são simples e rápidas, com possibilidade de aplicação a qualquer domínio. Estudos como [Lakshmi et al. 2013] e [Adhatrao et al. 2013] apresentam pesquisas comparativas de análise e desempenho entre os principais algoritmos de árvores de decisão: ID3, CART e C4.5.

No trabalho realizado por [Lin et al. 2013], os autores utilizam técnicas de árvores de decisão com o objetivo de desenvolver um sistema personalizado de aprendizagem baseada em criatividade. Esse sistema busca fornecer caminhos de aprendizagem personalizados, para otimizar o desempenho da criatividade em ambientes de EAD. Os experimentos realizados apontaram resultados onde a utilização do caminho de aprendizado, sugerido por uma árvore de decisão para os alunos, caracteriza uma probabilidade de 90% em obter uma pontuação de criatividade acima da média. Tais resultados sugerem que a técnica empregada pode fornecer insumos relevantes na aprendizagem adaptativa relacionada à criatividade.

No trabalho realizado por [Kabakchieva 2013], foi utilizado o *framework* CRISP-DM para aplicação de algoritmos de classificação, entre eles árvores de decisão, em dados de uma universidade na Bulgária. O estudo reúne dados relativos a características pessoais e pré-universitárias para previsão da performance dos alunos, utilizando metodologias de *holdout* e *cross validation*. Os resultados alcançados, quando analisados à luz das taxas de predição, não registraram valores representativos em sua fase inicial, onde as taxas variaram entre 52-67%. Esses resultados demonstram como um projeto de MDE pode evoluir a partir

de estudos iniciais.

A necessidade de uma instituição educacional obter conhecimento prévio sobre alunos matriculados para prever seu desempenho em futuros acadêmicos é tratado em [Adhatrao et al. 2013]. Esse trabalho analisou um conjunto de dados contendo informações sobre os alunos, como gênero, notas obtidas nos exames, notas e classificação nos exames de admissão, além de resultados no primeiro ano dos alunos. Foram utilizados os algoritmos de classificação com árvores de decisão ID3 e C4.5, prevendo o desempenho geral e individual dos alunos recém admitidos em exames futuros.

A qualidade na educação é tratada como um fator relevante para o aluno selecionar um instituição de ensino. No trabalho realizado por [Guleria et al. 2014], através da utilização de técnicas de árvores de decisão, os autores identificaram um atributo específico com grande influência na classificação do desempenho dos alunos, possibilitando a identificação do perfil do aluno com possibilidades de falha nos exames finais.

3.2.2 Tratamento de classes desbalanceadas

Quando considerados dados oriundos de aplicações educacionais, os conjuntos de dados exibem distribuições de classes onde quase todos os casos são atribuídos a uma classe e muito menos casos a uma classe menor, que geralmente é a classe mais interessante. Um classificador induzido de um conjunto de dados desbalanceado tem, tipicamente, uma baixa taxa de erro para a classe majoritária e uma taxa de erro inaceitável para a classe minoritária [Kotsiantis and Pintelas 2003].

Apesar de a evasão ser um problema nas instituições de ensino que utilizam a EAD, o número de casos ainda é, em geral, menor em relação ao número de alunos não evadidos. Nesse sentido, o problema é caracterizado pelo desbalanceamento das classes presentes em bases de dados direcionadas para análise preditiva e descritiva de desempenho. A existência de classes desbalanceadas faz com que os algoritmos de aprendizagem tendam a ignorar as classes menos frequentes (classes minoritárias) e só considerar as mais frequentes (classes majoritárias). Como resultado, o classificador não é capaz de classificar corretamente as instâncias de dados que correspondem a classes menos frequentes [Marquez-Vera et al. 2013].

Para tratar o problema do desbalanceamento, [Thai-Nghe et al. 2009] utilizaram técnicas de amostragem e de aprendizado sensível ao custo. Os resultados demonstram que o rebalanceamento de classes possibilita a melhora nos resultados quando comparados às bases desiguais. Os experimentos para essa conclusão abordaram a utilização de árvores de decisão, Redes Bayesianas e Máquinas de Vetor-Suporte.

3.3 ANÁLISE DE *LOGS* DE INTERAÇÃO

Os sistemas de computadores possibilitam o registro de todas as ações realizadas durante sua utilização. A técnica de análise de *logs* de interação consiste na exploração dos registros das interações (cliques de mouse para navegação) dos usuários em um sistema. Em ambientes educacionais de EAD, o clique do mouse representa uma interação com um determinado objeto de aprendizagem [Romero and Ventura 2007].

No trabalho realizado por [Gottardo et al. 2014], os autores utilizaram variáveis relacionadas ao nível de interação com o AVA, considerando a interação estudante-estudante e interação bidirecional estudante-professor para análise preditiva do desempenho dos alunos. Foram utilizados os modelos de classificação *Random Forest* e Redes Neurais com *Multilayer Perceptron*. Porém, nesse artigo, os autores utilizaram somente algoritmos classificados como *BlackBox* (caixa preta) ou seja, mesmo com a performance considerável dos algoritmos, não é possível extrair informações sobre quais características dos alunos estão relacionadas a cada classificação realizada. Com o intuito de suprir esse gargalo, no modelo proposto nessa dissertação, buscamos analisar as iterações dos alunos utilizando um algoritmo *WhiteBox* visando o entendimento sobre quais atributos influenciam no desempenho dos estudantes. Vale ressaltar que esse trabalho realizado em 2014 serve como base para a produção dessa presente dissertação de mestrado.

A análise de logs de interação pode proporcionar diversas vantagens para os atores envolvidos com a EAD. Através da aplicação de técnicas de MDE em dados provenientes de Ambientes Virtuais de Aprendizagem, como o Moodle, surgem soluções que podem ser utilizadas, por exemplo, para que instrutores possam visualizar dados de interação dos alunos de forma global, identificando comportamentos atípicos que possam ser analisados de forma mais aprofundada. Outra aplicação é a identificação de grupos com comportamentos semelhantes que, através do classificador, pode gerar regras para avaliar se existe alguma relação entre as características dos grupos classificados e os atributos indutores [Romero et al. 2008].

Existe uma relação direta entre a quantidade de dados a serem estudados com o tipo de análise a ser realizada. No trabalho realizado por [Romero and Ventura 2013], é apresentado um modelo de análise de *logs* de interação no Moodle onde a quantidade de dados está relacionada com a granularidade envolvida em cada tipo de análise. A Figura 3.2 demonstra como acontece essa relação, de acordo com a dimensão a ser considerada.

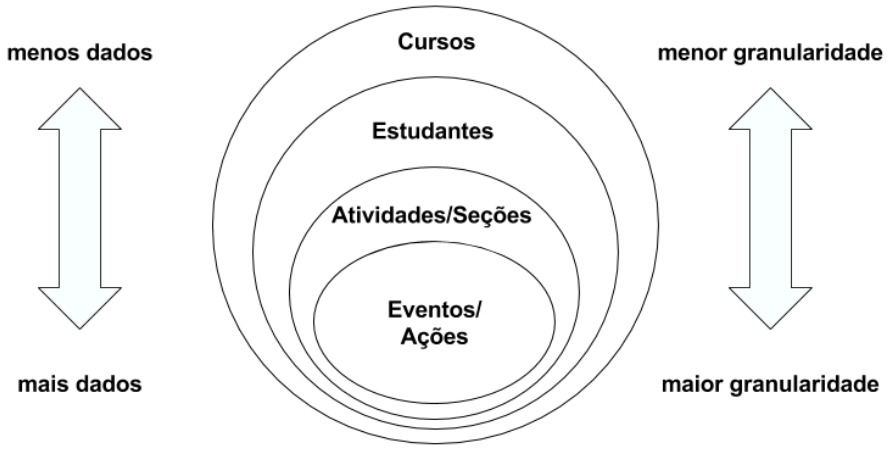


Figura 3.2: Relação granularidade x quantidade
Adaptado de [Romero and Ventura 2013]

Quanto mais detalhada for a informação a ser analisada, maior será a quantidade de informações presente nas bases de dados. Por exemplo, na Figura 3.2, uma análise envolvendo somente os Cursos contará com informações relacionadas ao tipo de curso, carga horária, nível de dificuldade, entre outros. Já uma análise envolvendo Eventos/Ações possuirá informações sobre quais eventos e ações foram realizadas por um estudante que realizou algum curso, baseado em atividades e seções de uso.

Capítulo 4

ESTUDO APLICADO

Este capítulo apresenta as características da metodologia aplicada nos experimentos realizados a partir da base de dados disponibilizada pela instituição de ensino. Os experimentos seguiram metodologias relacionadas às técnicas de classificação elencadas através da pesquisa apresentada no Capítulo 3, para avaliação e comparação dos resultados obtidos de acordo com o cenário avaliado.

4.1 METODOLOGIA

O método de pesquisa aplicado foi o da pesquisa quantitativa de caráter empírico, a partir da coleta de dados na busca de resultados que possam ser quantificados [Gil 2010]. O principal objetivo dessa pesquisa é testar hipóteses que tratam da relação entre causa e efeito das interações em ambientes virtuais de aprendizagem utilizados em cursos de curta duração na educação corporativa.

Os experimentos tiveram como base de análise a duração dos cursos que são ofertados pela instituição de ensino. Os cursos possuem uma característica relevante relacionada ao período de duração, com aproximadamente 40 horas, geralmente ofertados em um espaço de tempo de 30 dias, ou seja, 4 semanas de duração.

A análise dos dados de acordo com o andamento semanal das interações realizadas no AVA foi o foco principal para a proposição do modelo utilizado. Sendo assim, ao final da primeira semana de realização dos cursos, o modelo proposto possibilita a análise das interações ocorridas nesse intervalo de tempo, classificando os alunos de acordo com o desempenho ao final do curso.

Nesse sentido, a análise das interações ocorre em três momentos distintos, primeiro ao final de sete dias após o início dos cursos, denominada no modelo como **S1** (Semana 1), o segundo ao final dos 14 dias, denominado **S2** (Semana 2) e por fim, ao final de 21 dias de realização do curso, denominado **S3** (Semana 3). Esse modelo tem como objetivo a avaliação de qual composição de *dataset*, analisando duas composições diferentes em relação

ao espaço de tempo, possibilita melhores resultados em relação às taxas de TP e FP.

Para validação do modelo proposto, foram realizados dois experimentos, utilizando a mesma base de dados com estruturas diferenciadas em relação à composição dos *datasets* utilizados. No primeiro experimento (DS1), são consideradas as semanas de interação de forma isolada, ou seja, os dados das interações de cada uma das semanas são analisadas de forma independente. Em outras palavras, são analisados somente os dados da primeira semana, em seguida são analisados isoladamente os dados da segunda semana e assim sucessivamente. No segundo experimento (DS2), estão sendo consideradas as semanas de interação de forma incremental, ou seja, ao final da primeira semana, é realizada uma primeira análise, ao final da segunda semana são analisadas, na mesma base de dados, as interações da primeira e da segunda semana e, por fim, na terceira semana, são analisados os dados das interações realizadas na primeira, segunda e terceira semana de forma unificada. A quarta semana de interação não foi considerada tendo em vista que ao final dessa semana o curso será finalizado, inviabilizando possíveis intervenções.

A utilização de um framework em projetos de MD incentiva a aplicação de práticas já consagradas em projetos de sucesso, oferecendo às organizações uma estrutura necessária para obtenção de resultados melhores e mais rápidos [Shearer 2000]. Os experimentos seguiram as fases propostas em um *framework* específico, conforme apresentado na Subseção 4.1.1. Este *framework* auxiliou na condução organizada e bem documentada de diversas atividades na execução dos experimentos.

4.1.1 *Framework* CRISP-DM

O *framework* CRISP-DM (do inglês, *Cross Industry Standard Process for Data Mining*) propõe a organização dos projetos de mineração de dados em seis fases: (A) Entendimento do negócio, (B) Entendimento dos dados, (C) Pré-processamento dos dados, (D) Modelagem, (E) Avaliação e (F) Implementação, conforme apresentado na Figura. 4.1. Nesta representação, as setas internas indicam as dependências mais importantes e frequentes entre as fases. O círculo exterior simboliza a natureza cíclica da mineração de dados, onde as lições aprendidas durante todo o processo podem desencadear novas questões para as fases do projeto [Wirth and Hipp 2000].

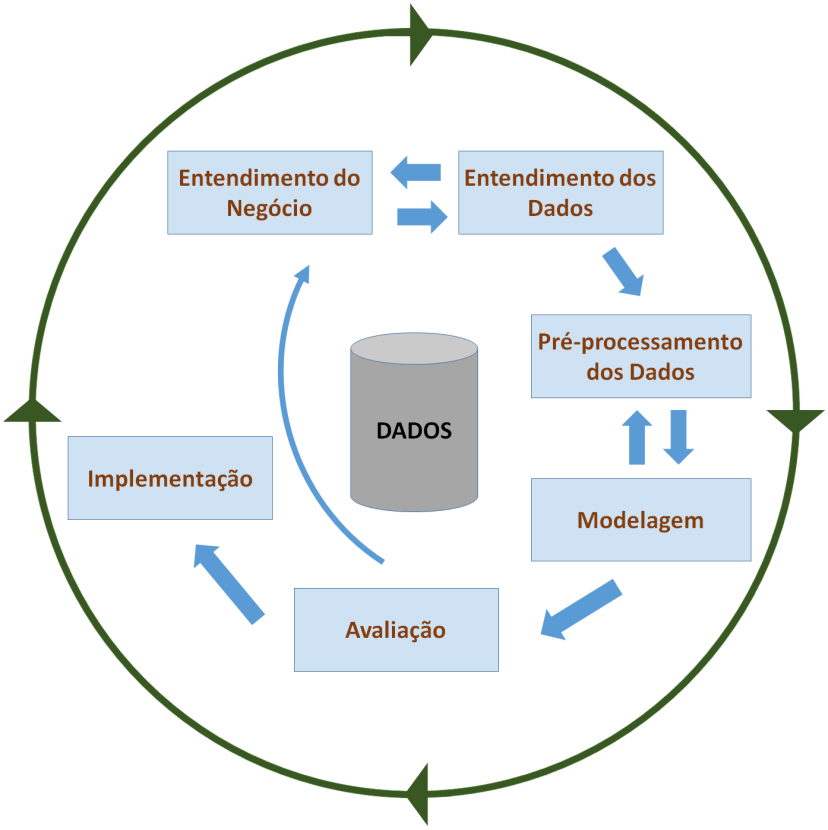


Figura 4.1: Etapas do *framework* CRISP-DM
adaptado de [Wirth and Hipp 2000]

Na fase (A) de entendimento do negócio, devem ser identificados os objetivos e as metas para a mineração de dados, gerando um plano para o projeto. A fase de entendimento dos dados (B) aborda a coleta inicial dos dados e também serve para a familiarização dos envolvidos com os tipos de dados específicos do projeto. Na fase (C), o pré-processamento dos dados é realizado para a construção do conjunto de dados a ser utilizado no modelo para as atividades de mineração. Esse é um ponto crucial, onde é comum a necessidade de melhor entendimento e familiarização com os dados, retornando para as etapas anteriores. A fase (D) define o modelo que será utilizado para a mineração dos dados, o que, em termos práticos, envolve escolher as atividades específicas de mineração. A fase (E) avalia os resultados da fase anterior levando em consideração as metas de precisão e confiabilidade. Esta etapa avalia o grau de satisfação que o modelo proposto proporciona para o projeto de mineração, visando verificar se existe algum motivo para que o modelo não seja eficiente. Por fim, a fase (F) é referente à implementação dos resultados obtidos através das atividades de mineração. Nesta fase são determinadas as estratégias e o planejamento para o monitoramento efetivo dos resultados diretamente aplicados ao negócio do projeto de mineração.

4.2 ENTENDIMENTO DO NEGÓCIO

Esta é a fase inicial, onde devem ser identificados os objetivos e as metas para a mineração de dados, gerando um plano para o projeto. O principal objetivo desta fase foi compreender, a partir de uma perspectiva de negócio, quais seriam as possibilidades e os fatores que possam influenciar no resultado do projeto de mineração de dados[Wirth and Hipp 2000].

Neste contexto, a Coordenação Geral de Educação a Distância (CGEAD) é responsável pela oferta dos cursos a distância na Enap. O catálogo de cursos ofertados é direcionado para o aperfeiçoamento e a formação dos servidores públicos no Brasil. São utilizadas duas metodologias para oferta dos cursos a distância, os cursos sem tutoria ou auto-instrucional e os cursos com tutoria, ou seja, instrucional.

A modalidade sem tutoria, ou auto-instrucional, trata dos cursos onde os alunos utilizam o AVA para acessar os conteúdos teóricos, o material de apoio para estudo, além de exercícios de fixação e exercícios avaliativos para classificação dos alunos como aprovados, reprovados ou evadidos. A modalidade com tutoria ou instrucional trata dos cursos que possuem o acompanhamento de tutores durante a sua realização. A principal diferença entre esses dois cursos são os módulos do AVA que são utilizados: nos cursos sem tutoria, são utilizados somente módulos estáticos, que possuem características de interação somente do aluno com o AVA. Os cursos com tutoria possuem características de interação entre alunos, tutores e o AVA, sendo que os tutores são responsáveis pela condução de atividades específicas que incluem a interação entre os alunos e os tutores, entre os alunos e o AVA, bem como a interação de alunos entre si.

4.2.1 Objetivos e metas

Os cursos ofertados possuem geralmente 30 dias de duração ou 4 semanas. O projeto de mineração deve considerar que as análises devem ser passíveis de realização entre as semanas de realização dos cursos, ou seja, ao final da primeira semana de realização, devem ser analisados os dados desta semana para geração de indicadores que possibilitem uma análise descritiva de como se comportam os alunos conforma a propensão a uma determinada nota final. Para as análises das semanas seguintes, é necessário validar qual é o modelo que possibilita um maior índice de assertividade e com os melhores índices de *True Positive* (TP) e *False Positive* (FP).

Até o início da realização desse estudo de caso (janeiro de 2016), não havia na Enap indicadores que possibilitassem o acompanhamento relacionado às interações dos alunos com o AVA durante a realização dos cursos. Tais indicadores poderiam possibilitar a intervenção, por parte da CGEAD, no combate aos índices de evasão e reprovação.

Portanto, o plano para o projeto de mineração foi definido com o objetivo de gerar in-

dicadores semanais relacionados à interação dos alunos com o AVA, com informações que possibilitem a realização de ações relacionadas ao combate da evasão e da reprovação nos cursos de maior quantidade de alunos, no caso, os cursos auto-instrucionais (sem tutoria).

Os indicadores devem possibilitar a geração de informações em intervalos semanais, a partir do início dos cursos. Tais informações devem possuir características que descrevam os perfis dos alunos de acordo com a nota final obtida nos cursos, possibilitando a compreensão, por parte da CGEAD, de como se comportam esses alunos em relação à interação deles com os objetos de aprendizagem e a nota final.

Em síntese, a Enap atua na oferta de cursos a distância para capacitação de servidores públicos no Brasil utilizando um AVA baseado no software Moodle. O plano para o projeto de mineração trata da extração de informações de logs do Moodle que possibilitem a análise de interações dos alunos com os módulos presentes nos cursos para que sejam aplicadas técnicas de MDE, especificamente a classificação através da geração de árvores de decisão, possibilitando uma análise descritiva e preditiva do desempenho dos alunos, quando consideradas as interações em intervalos semanais.

Nesse sentido, é necessário verificar qual é o melhor modelo de análise das interações quando consideradas as semanas dos cursos. Foram identificadas duas formas para essa análise, a primeira que considera o modelo com as semanas de forma isolada, ou seja, ao final de cada semana, são analisados os dados isolados da semana que passou. Ao final da primeira semana (S1), é realizada uma primeira verificação, na segunda semana (S2) são analisados os dados somente das interações ocorridas entre 7 e 14 dias de realização dos cursos e, por fim, o mesmo procedimento na terceira semana (S3). A segunda visão para validação deve considerar as semanas de forma incremental, ou seja, os dados das semanas serão acumulados para as análises posteriores. Ao final da primeira semana, são analisados os dados da mesma forma que o primeiro modelo (S1), na segunda semana serão analisados os dados da semana 1 mais os dados da semana 2 (S1_S2). Na terceira semana, serão analisados conjuntamente os dados da semana 1, semana 2 e semana 3 (S1_S2_S3).

4.3 ENTENDIMENTO DOS DADOS

Esta fase aborda a coleta inicial dos dados e também serve para a familiarização dos pesquisadores envolvidos com os tipos de dados específicos do projeto [Wirth and Hipp 2000]. Nessa fase, são definidas questões relevantes como:

- quais atributos das bases de dados possuem características promissoras,
- quais desses atributos parecem irrelevantes e podem ser descartados,
- se existem dados suficientes para gerar indicadores generalizáveis ou que possibilitem predições precisas,

- como serão tratados os casos de valores omissos.

A Figura 4.2 apresenta as características de como os dados são tratados e armazenados para a oferta dos cursos. Os dados cadastrais e as informações sobre os detalhes de caráter administrativo são armazenados no ambiente da Secretaria Virtual, que utiliza um software especificamente desenvolvido para a Enap, denominado *WebCef*. Os dados dos cursos como os conteúdos, os exercícios e todas as informações referentes às interações e registros acadêmicos são armazenados no AVA, que é baseado no ambiente do *software* Moodle[1] com o sistema gerenciador de banco de dados *PostgreSql*[2].

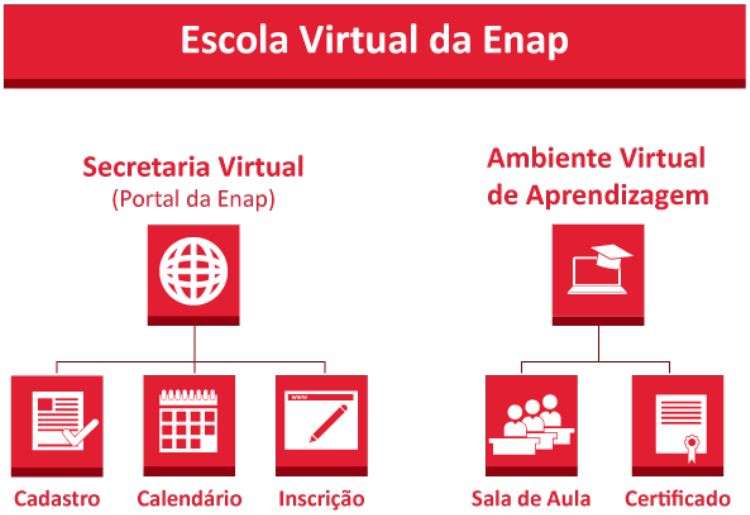


Figura 4.2: Estrutura para armazenamento dos dados
Retirado de *enapvirtual.enap.gov.br*

A utilização do AVA está condicionada à seleção de quais módulos estarão disponíveis para a realização de um curso. A utilização dos módulos representa quais objetos de aprendizagem serão utilizados para a oferta de um determinado curso. Cada módulo possui um tipo de característica de interação que o aluno pode realizar quando utilizar o AVA. A Figura 4.3 apresenta um exemplo de estrutura de conteúdos e seus respectivos módulos que foram utilizados para a criação de um curso no AVA.

Os objetos destacados na Figura 4.3 representam os módulos utilizados em um curso. Esses módulos possuem características diversas como, por exemplo, a apresentação de conteúdos, a disponibilização de materiais de referência, exercícios para reflexão e avaliação, entre outros. Quando o aluno acessa o ambiente do curso, cada um desses módulos possui características específicas de acordo com sua finalidade.

[1]Disponível em https://moodle.org
[2]Disponível em https://www.postgresql.org

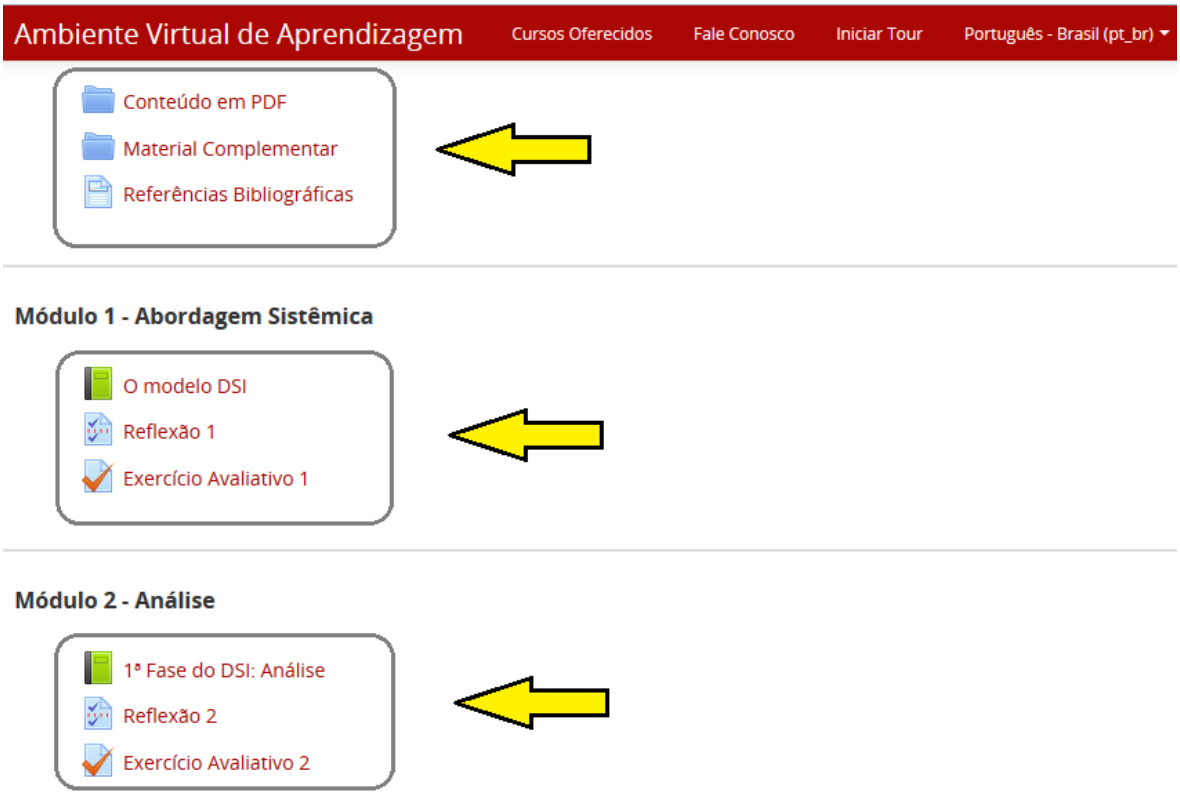


Figura 4.3: Exemplo de estrutura dos cursos no AVA

4.3.1 Definição dos dados

A base de dados do AVA Moodle foi definida como a fonte de informações para a criação e composição dos *datasets*, com os atributos e classes a serem verificadas através das atividades de classificação. Esta base de dados possui aproximadamente 361 tabelas, sendo responsável por todos os registros relacionados à utilização do AVA para as ofertas de cursos realizadas entre o ano de 2015 e 2016.

Entre as tabelas do sistema Moodle, foi identificada a tabela nomeada como *mdl_logstores_standard_log*, que é responsável por armazenar todas as informações dos acessos e interações realizadas, ou seja, os *logs* de utilização, dentro do AVA. A Tabela 4.1 apresenta os atributos que compõem essa estrutura e como são armazenados.

Tabela 4.1: Campos e características - Tabela de *log* do Moodle

Nome do Atributo	**Tipo**	**Descrição**
id	bigserial	Campo de id da tabela de log
eventname	character varying	Registra os tipos de eventos
component	character varying	Registro do componente acessado
action	character varying	Registra o tipo de ação realizada no componente
target	character varying	Tabela alvo do evento
objecttable	character varying	Tabela alvo do registro
objectid	int	Identificador do objeto que foi acionado
crud	character varying	Registra o tipo de ação realizada copy, read, update ou delete
edulevel	int	Registro do componente de tabela respossável
contextid	int	Contexto de realização da ação
contextlevel	int	Nível do contexto da ação
contextinstanceid	int	Identificado da instância acionada
userid	int	Identificador do aluno
courseid	int	Identificador do curso
relateduserid	int	Identifica usuários assumindo papel para possíveis alterações administrativas
anonymous	int	Registra se o usuário se logou no sistema ou se foi uma ação externa anônima
other	text	Registro geral
timecreated	int	Registra o momento da ação
origin	character varying	Local que originou o registro
ip	character varying	Identificador do IP utilizado para o registro
realuserid	int	Registro de userid quando for acessado por outro usuário

Após a identificação da tabela que possui os registros mais promissores, foram identificados os atributos que seriam capazes de caracterizar as interações dos alunos com o AVA. Dentre os atributos presentes na tabela de *log*, foram elencados somente os que possuem informações que estejam relacionadas à interação dos alunos com os módulos do AVA. A Tabela 4.2 apresenta quais foram os selecionados para criação dos *datasets*.

Tabela 4.2: Atributos selecionados para composição do *Dataset*

Atributo	Tipo	Descrição
component	character varying	Registro do componente acessado
action	character varying	Registra o tipo de ação realizada no componente
target	character varying	Registra a característica da ação realizada
userid	int	Identificador do aluno
courseid	int	Identificador do curso
timecreated	int	Registra data e hora da evento

Através da estruturação dos atributos apresentados na Tabela 4.2, foi possível identificar informações completas sobre quais tipos de interações ocorreram por cada aluno nos cursos ofertados.

Em seguida, para garantir a geração de indicadores que possibilitem predições válidas, foram elencados os cursos com características similares em relação às interações e os objetos educacionais utilizados e com o maior número de alunos inscritos. Nesse caso, foram selecionados os cursos com maior representatividade para a instituição, agrupando dados de aproximadamente 71 mil alunos em 45 turmas de 7 cursos realizados entre os anos de 2015 e 2016, conforme descrito na Tabela 4.3.

Tabela 4.3: Definição dos cursos para composição do *Dataset*

Cursos Selecionados	Qtd. Turmas	Qtd. Alunos
A Previdência Social dos Servidores Públicos Regime Próprio e Complementar	7	10.461
Atendimento ao Cidadão	8	11.505
Ética e Serviço Público	8	14.180
Formação de Pregoeiros	5	6.994
Gestão da Informação e Documentação Conceitos Básicos em Gestão Documental	4	8.647
Introdução à Gestão de Processos	6	8.130
Orçamento Público Conceitos Básicos	7	11.162
TOTAL	**45**	**71.079**

4.4 PREPARAÇÃO DOS DADOS

Nesta fase, o pré-processamento dos dados é realizado para a construção do conjunto de dados, denominado *dataset*, a ser utilizado no modelo definido para mineração [Wirth and Hipp 2000].

A seleção dos dados compreende a identificação de quais informações, dentre as bases de dados existentes, devem ser efetivamente consideradas durante o projeto de mineração. Em um contexto geral, os dados que são utilizados para análises encontram-se organizados em

bases transacionais que, por características nativas, sofrem constantes atualizações ao longo do tempo. Considera-se uma boa prática que os dados sejam copiados para um ambiente dedicado às atividades do projeto de mineração para que não haja interferência nas rotinas dessas bases transacionais. No caso da aplicação em questão, foi realizada uma cópia íntegra do banco de dados e isolado um ambiente no gerenciador de bases de dados Postgresql.

Após a seleção dos cursos alvo para a atividade de mineração, o passo seguinte abordou a seleção de quais atributos da tabela de *logs* serão utilizados para identificar as interações dos alunos com o AVA. Dentre os atributos presentes na tabela de *log* no sistema Moodle, foram selecionados os que registram informações sobre que tipo de interação ocorreu em determinado momento, com base nos objetos de aprendizagem (módulos) que são utilizados em comum para esses cursos selecionados. Essa seleção se deu pois, os cursos na modalidade sem tutoria possuem em sua grande maioria os mesmos módulos e ações que são passíveis de interação junto ao AVA. A preparação do *dataset* para a realização das próximas atividades visou a implementação dos algoritmos na fase de modelagem, tendo como base os atributos apresentados na Seção anterior, Tabela 4.2.

Os atributos selecionados possuem funções específicas, sendo que os campos *component*, *action* e *target* são responsáveis pelos registros das interações dos alunos com os objetos educacionais presentes nos cursos. Com os dados recentes nesses atributos é possível extrair informações relacionadas à quantidade de interações que cada aluno realizou com cada objeto educacional dos cursos realizados. Com esses atributos, será possível criar uma tupla (linha) com a seguinte estrutura:

- *userid, courseid, component, action, target, timecreated*.

Através dessa estrutura de atributos, foi realizada a consulta relacionada aos *logs* das interações realizadas, conforme pode ser observado, por exemplo, na Tabela 4.4.

Tabela 4.4: Extração de interações

userid	**courseid**	**component**	**action**	**target**	**timecreated**
125xxx	176	mod_book	view	chapter	03/08/2015 10:23:45
125xxx	176	mod_book	view	course_module	03/08/2015 10:23:49
125xxx	176	mod_glossary	view	course_module	03/08/2015 10:25:32
125xxx	176	mod_glossary	view	entry	03:08/2015 10:32:42
125xxx	176	mod_folder	view	course_module	04/08/2015 09:23:54

A partir dos dados provenientes da extração das informações com a estrutura apresentada na Tabela 4.4, foi realizada a integralização da quantidade de interações realizadas pelos alunos durante o intervalo de 30 dias, separada em quatro partes que representam as semanas de realização dos cursos.

Após a integralização dos dados, foi definido o conjunto de características que participarão da composição dos *datasets* com as interações separadas por intervalos de semanas. A

Tabela 4.5 apresenta as *features* (características) que foram consideradas na composição dos *datasets*.

Tabela 4.5: Composição e características do *Dataset*

Nome do atributo	Descrição
primeiro_acesso	Quantidade de dias para o primeiro acesso dos alunos ao AVA
book_view	Quantidade de acessos ao módulo de conteúdos, livro.
quiz_view_attempt	Quantidade de acessos à revisão de tentativas no módulo de exercícios
quiz_view_course	Quantidade de visualizações ao módulo de exercícios do curso
folder_view	Quantidade de acessos ao ambiente de conteúdos de apoio do curso, biblioteca
page_view	Quantidade de acessos às páginas externas de conteúdos de apoio
questionnaire_view	Quantidade de visualizações às atividades pontuadas do curso
questionnaire_submitted	Quantidade de atividades pontuadas submetidas para avaliação
glossary_view	Quantidade de visualizações ao módulo de glossário
glossary_view_entry	Quantidade de visualizações às entradas de glossário disponíveis entre os conteúdos
pontuação_final	Pontuação final (de 0 a 100 pontos) obtida pelos alunos ao final do curso

Todos os atributos listados na Tabela 4.5 são numéricos e inteiros, possuindo valores entre $0...n$. Nessa fase, realizou-se a atividade de discretização do atributo alvo nota_final, que foi separado em categorias específicas de acordo com as notas dos alunos. Resultou-se assim em um novo atributo **nota final**.

A discretização é uma técnica essencial em projetos relacionados a descoberta de conhecimento e tarefas de mineração de dados. O objetivo principal é transformar um conjunto de atributos contínuos em discretos, associando valores categóricos a intervalos e assim transformando dados quantitativos em dados qualitativos [Garcia et al. 2013]. Nesse sentido, o atributo nota_final foi discretizado em três classes distintas, conforme segue:

- Classe **EVA** (Evadido), para as notas com 0 pontos obtidos, agrupando os alunos que não realizaram nenhum tipo de atividade pontuada, caracterizando o abandono do curso;
- Classe **REP** (Reprovado), para as notas com valores entre 1 e 59 pontos obtidos, agrupando os alunos que realizaram pelo menos uma atividade, obtendo pontos, porém, não alcançando a quantidade mínima para aprovação e;
- Classe **APV** (Aprovado), para notas com valores entre 60 e 100 pontos obtidos, agrupando os alunos que realizaram atividades e alcançaram a nota mínima para aprovação no curso.

A preparação do atributo **nota final** através da atividade de discretização foi necessária tendo em vista que o objetivo desse projeto de mineração, conforme apresentado na Seção 4.2, envolve a aplicação de técnicas de classificação através do algoritmo J48[3].

Com a distribuição dos alunos entre as classes do novo atributo **nota final**, foi possível visualizar a distribuição da quantidade de alunos em cada uma delas, conforme a Figura

[3]Na ferramenta WEKA, o algoritmo C4.5 [Quinlan 1993] recebe o nome de J48

4.4. É possível observar que a quantidade de alunos aprovados (APV) é superior à soma das outras duas classes (REP e EVA). Essa observação se faz necessária, tendo em vista que é perceptível nesse momento que as classes para esse *dataset* estão desbalanceadas. Na Seção de Modelagem, será apresentada uma proposta para tratamento e análise em relação ao desbalanceamento das classes.

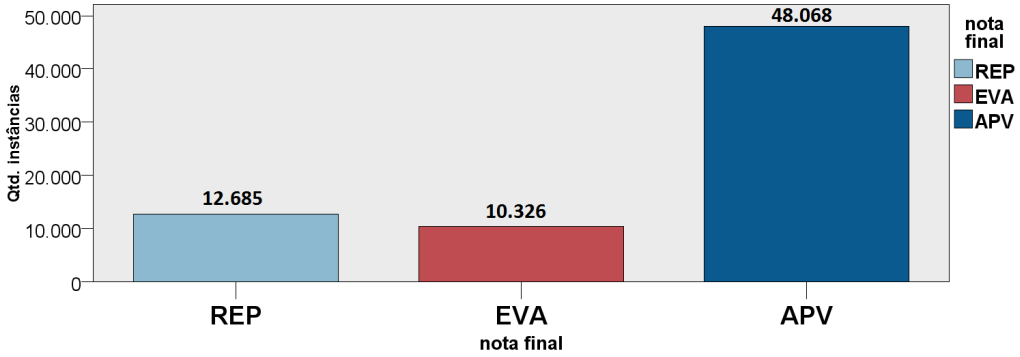


Figura 4.4: Distribuição das classes de notas após discretização

Durante a realização dos experimentos propostos neste Capítulo, foi utilizada uma técnica direcionada para o tratamento de classes desbalanceadas para comparação dos resultados na execução dos algoritmos. Conforme apresentado no Capítulo 2 (Tratamento de classes desbalanceadas), é comum em ambientes educacionais que as classes estejam presentes em quantidades desproporcionais.

4.4.1 Estatísticas descritivas

Os atributos possuem valores específicos, onde os valores de mínimos e máximos podem ser observados na Tabela 4.6.

Tabela 4.6: Estatísticas descritivas - Mínimos e Máximos

Variável	**Mínimo**			**Máximo**		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**
primeiro_acesso	0	0	0	7	7	7
book_view	1	0	0	3.105	5.824	8.262
quiz_view_attempt	0	0	0	2.183	1.160	2.272
quiz_view_course	0	0	0	3.540	806	1.638
folder_view	0	0	0	590	208	208
page_view	0	0	0	342	288	487
questionnaire_view	0	0	0	118	328	188
questionnaire_submitted	0	0	0	59	52	59
glossary_view	0	0	0	896	638	364
glossary_view_entry	0	0	0	1.179	663	405

Analisando os valores, pode-se verificar que os atributos relacionados a conteúdos como *book_view*, *glossary* e *quiz* possuem valores máximos elevados quando comparados com os

outros atributos. Essa questão ocorre devido a utilização de um AVA em ambiente EAD onde os conteúdos possuem uma carga mais elevada de interação quando comparados com os outros módulos. Outra informação presente é a existência de valores mínimos com zero, o que caracteriza que houve alunos que não interagiram com algum dos módulos analisados durante a realização dos cursos.

A Tabela 4.7 apresenta os valores médios de cada atributo bem como o respectivo valor do desvio padrão de cada um desses atributos.

Tabela 4.7: Estatísticas descritivas - Médias e Desv. Padrão

Variável	**Média**			**Desv. Padrão**		
	S1	**S2**	**S3**	**S1**	**S2**	**S3**
primeiro_acesso	1,48	1,48	1,48	2,14	2,14	2,14
book_view	78,56	27,86	23,04	138,95	84,20	90,74
quiz_view_attempt	10,10	5,94	6,39	31,81	19,18	23,17
quiz_view_course	11,99	6,78	6,92	31,85	18,51	19,98
folder_view	4,74	1,14	0,87	9,57	4,26	3,85
page_view	4,15	0,94	0,65	9,08	4,09	3,83
questionnaire_view	0,81	0,85	1,18	3,32	3,41	3,49
questionnaire_submitted	0,30	0,44	0,75	1,59	1,64	2,01
glossary_view	1,66	0,17	0,12	8,92	4,02	2,28
glossary_view_entry	1,22	0,40	0,29	10,77	5,91	4,21

Analisando os valores médios apresentados, é possível verificar que algumas interações registraram valores médios abaixo de 1, o que significa que houve casos de baixa interação com determinados módulos do AVA. Para o atributo book_view, os valores de desvio padrão foram elevados nas três semanas, o que pode caracterizar uma grande variabilidade entre os perfis de acesso a esse módulo.

Após a análise estatística dos dados, foi possível identificar a existência de valores discrepantes entre os atributos selecionados. Tal fato pode caracterizar a existência de ruído, ou seja, *outliers* (valores discrepantes), que indicam a possibilidade de afirmações baseadas em valores fora de contexto. Porém, quando os dados analisados são provenientes de ambientes educacionais, os valores discrepantes normalmente são observações verdadeiras pois existem alunos excepcionais que têm sucesso com pouco esforço ou que falham contra todas as expectativas [Hämäläinen and Vinni 2010].

Como atividade final dessa etapa, os dados foram separados em arquivos específicos para utilização na ferramenta WEKA, no formato *Attribute Relation File Format* (ARFF) para leitura dos dados [Hall et al. 2009].

```
@relation 'SEM_TUTORIA_1_SEMANA'

@attribute primeiro_acesso numeric
@attribute S1_course_view numeric
@attribute S1_folder_view numeric
@attribute S1_book_view numeric
@attribute S1_book_print numeric
@attribute S1_glossary_view numeric
@attribute S1_questionnaire_view numeric
@attribute S1_questionnaire_submit numeric
@attribute S1_quiz_view numeric
@attribute S1_quiz_attempt numeric
@attribute 'nota final' {EVA,APR,REP}

@data
0,20,0,0,0,0,0,0,0,0,EVA
0,297,0,572,0,0,0,0,55,0,EVA
0,70,0,224,0,0,0,0,0,0,APR
6,12,0,8,0,0,0,0,2,0,APR
1,16,0,100,0,0,0,0,0,0,APR
0,30,0,95,0,0,0,0,0,0,EVA
1,52,0,351,0,0,0,0,0,0,APR
```

Figura 4.5: Estrutura dos dados - ARFF

A sintaxe dos arquivos ARFF é composta por uma estrutura inicial com os nomes e os respectivos tipos de cada um dos atributos. Em seguida, são registrados em cada linha os valores para cada um dos atributos, onde cada linha representa um aluno. A Figura 4.5 é um exemplo da composição do ARFF referente à primeira semana (S1), onde os dados das interações de cada aluno estão presentes após a marcação @*data*.

Com os dados, passamos para as atividades de modelagem para o projeto de mineração, onde serão definidas as características dos algoritmos e também a metodologia para validação do modelo proposto.

4.5 MODELAGEM

A fase de modelagem é onde são definidos os modelos de mineração de dados que serão utilizados. Em suma, esta fase envolve as atividades específicas para definição de quais algoritmos serão aplicados em busca dos resultados registrados durante o entendimento do negócio [Wirth and Hipp 2000]. Para esse projeto foi definida a atividade de classificação supervisionada com árvores de decisão, para uma análise preditiva de desempenho e descritiva sobre as características de interação.

Conforme exposto na Seção 4.1, os dados foram separados em dois *datasets* que possuem características distintas em relação a sua composição. A Tabela 4.8 apresenta como foram compostos os dados para a criação de cada um dos datasets.

Tabela 4.8: Composição dos *Datasets*

Dataset	Composição	Características
DS1	S1, S2 e S3	Dados de cada semana isolados, analisados de forma separada
DS2	S1, S1_S2 e S1_S2_S3	Dados das semanas agrupados, analisados de forma conjunta

A Figura 4.6 apresenta de forma gráfica a sequência de ações previstas para o modelo proposto. Inicialmente foram coletados e armazenados os dados referentes à quantidade de interações dos alunos com cada módulo dos cursos e suas respectivas notas já discretizadas em classes. Em seguida, foram separados os *datasets* para realização dos experimentos (DS1 e DS2), que utilizaram a técnica de *Cross Validation*, ou validação cruzada, em 10 partições. A validação utilizou o algoritmo J48 configurado com o parâmetro de quantidade mínima de instâncias $M = 300$ e o fator de poda em $C = 0,25$ para avaliar qual *dataset* possui melhores resultados. Vale ressaltar que o valor de M seguiu orientações da instituição e buscou agrupar cerca de 0,5% do total da massa de dados para esse agrupamento. Em seguida, com o *dataset* selecionado, foi utilizada uma técnica para o rebalanceamento das classes e realização novamente do *cross validation*, baseado nos mesmos parâmetros do Resultado 1 (R1). Por fim, foram comparados os resultados em R1 e R2 para definição de qual modelo detém os melhores resultados.

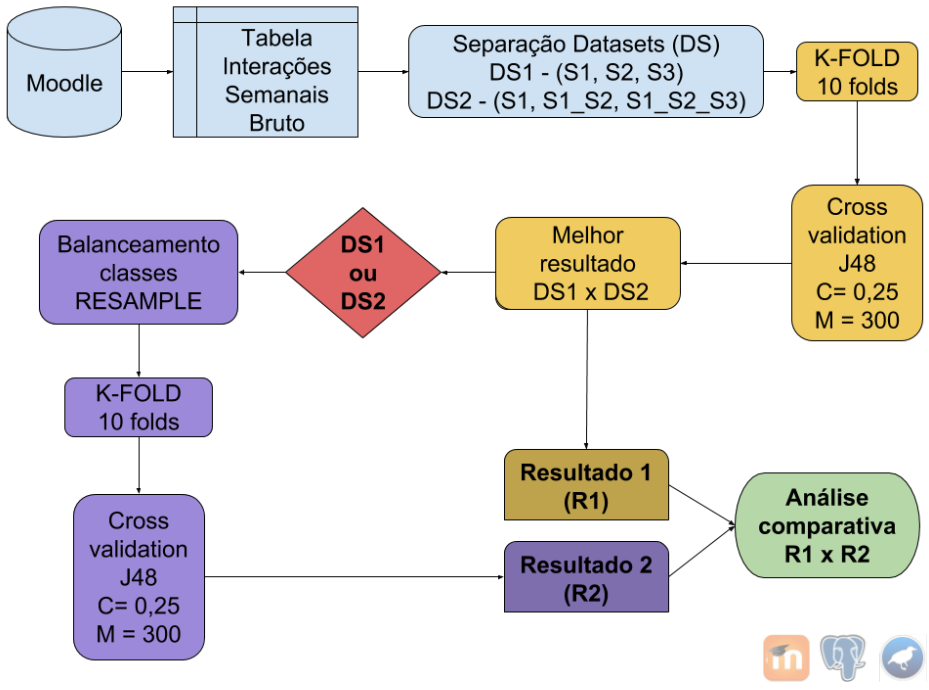


Figura 4.6: Arquitetura proposta para os experimentos

Para a atividade de balanceamento das classes, foi utilizado um filtro específico da ferramenta WEKA, denominado *RESAMPLE*. Esse filtro produz uma subamostra aleatória de um conjunto de dados usando amostragem com ou sem reposição. Pode ser utilizado para criar uma nova amostra com os dados distribuídos de maneira mais uniforme entre as

classes e/ou aumentar ou diminuir o tamanho da amostra [Frank et al. 2005].

Na ferramenta WEKA, o algoritmo *RESAMPLE* possui alguns parâmetros de configuração que permitem ajustar o balanceamento das classes. No experimento realizado foram configurados alguns parâmetros, conforme apresentado a seguir.

- $biasToUniformClass = 1$ — significa que o balanceamento das classes será uniforme, mantendo a mesma quantidade de instâncias;
- $noReplacement = False$ — não serão feitas substituições nem alterações na quantidade de instâncias e;
- $sampleSizePercent = 100\%$ — configura o percentual de instâncias do *dataset* original que resultarão no *dataset* balanceado.

4.5.1 Metodologia de validação

A utilização da técnica de *10-fold cross validation* proporciona a realização de uma validação cruzada do *dataset* levando em consideração todos os dados, ou seja, em algum momento o dado é considerado ou para a etapa de treinamento, ou para a etapa de teste. Os resultados apresentam a média de instâncias classificadas corretamente e também o índice de TP e FP para cada uma das classe avaliadas [Witten et al. 2016].

O *dataset* DS1 reúne os dados para a análise de semanas de forma isolada e o DS2 reúne os dados das semanas de forma incremental. Essa separação tem como objetivo a comparação de qual composição apresenta os melhores resultados. Em seguida, foi aplicada uma técnica para balanceamento das classes. Como pode ser observado na Seção 4.4, Figura 4.4, as classes de notas possuem valores discrepantes quando comparadas as classes REP e EVA em relação à classe APV.

4.6 AVALIAÇÃO

Esta etapa avalia o grau em que o modelo atende aos objetivos de negócios e/ou determina se existe alguma razão para que este modelo seja deficiente. Além das descobertas que estão necessariamente relacionados aos objetivos originais do negócio, outros achados também podem revelar desafios, informações ou sugestões adicionais, para direcionamentos futuros [Wirth and Hipp 2000].

Resultados DS1

A Figura 4.7 apresenta os resultados obtidos para o DS1 a partir da técnica de 10 *fold cross validation*. Os valores considerados para a avaliação compreendem o percentual correto, que significa o total geral de alunos classificados corretamente, e a taxa de TP e FP que representa a acurácia do modelo onde as instâncias classificadas como TP foram previstas corretamente e as classificadas como FP significam a rotulação de classes diferentes das que deveriam ter sido aplicadas.

S1								
Dataset	Fold	% Correto	TP			FP		
			EVA	APV	REP	EVA	APV	REP
S1	1	67,9516	0,185	0,969	0,000	0,035	0,887	0,000
	2	68,0501	0,215	0,958	0,012	0,040	0,868	0,004
	3	68,2372	0,184	0,970	0,000	0,034	0,891	0,000
	4	67,7828	0,192	0,958	0,000	0,046	0,882	0,000
	5	67,7406	0,187	0,962	0,000	0,040	0,887	0,000
	6	69,0068	0,236	0,959	0,000	0,044	0,862	0,000
	7	68,036	0,216	0,958	0,000	0,046	0,868	0,000
	8	67,417	0,224	0,957	0,000	0,050	0,858	0,000
	9	68,4581	0,164	0,971	0,000	0,032	0,903	0,000
	10	68,2004	0,238	0,965	0,000	0,044	0,854	0,000
MÉDIA		**68,088**	**0,204**	**0,963**	**0,001**	**0,041**	**0,876**	**0,000**
DESV. PADRÃO		**0,43498173**	**0,025**	**0,006**	**0,004**	**0,006**	**0,016**	**0,001**

(a) DS1 - Semana 1

S2								
Dataset	Fold	% Correto	TP			FP		
			EVA	APV	REP	EVA	APV	REP
S2	1	67,206	0,000	1,000	0,000	0,000	1,000	0,000
	2	67,5577	0,000	1,000	0,000	0,000	1,000	0,000
	3	67,6702	0,000	1,000	0,000	0,000	1,000	0,000
	4	67,9094	0,000	1,000	0,000	0,000	1,000	0,000
	5	67,5014	0,000	1,000	0,000	0,000	1,000	0,000
	6	68,4581	0,000	1,000	0,000	0,000	1,000	0,000
	7	67,7406	0,000	1,000	0,000	0,000	1,000	0,000
	8	66,99949	0,000	1,000	0,000	0,000	1,000	0,000
	9	67,0782	0,000	1,000	0,000	0,000	1,000	0,000
	10	67,1451	0,000	1,000	0,000	0,000	1,000	0,000
MÉDIA		**67,527**	**0,000**	**1,000**	**0,000**	**0,000**	**1,000**	**0,000**
DESV. PADRÃO		**0,44838989**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**

(b) DS1 - Semana 2

S3								
Dataset	Fold	% Correto	TP			FP		
			EVA	APV	REP	EVA	APV	REP
S3	1	67,206	0,000	1,000	0,000	0,000	1,000	0,000
	2	67,5577	0,000	1,000	0,000	0,000	1,000	0,000
	3	67,6702	0,000	1,000	0,000	0,000	1,000	0,000
	4	67,9094	0,000	1,000	0,000	0,000	1,000	0,000
	5	67,5014	0,000	1,000	0,000	0,000	1,000	0,000
	6	68,4581	0,000	1,000	0,000	0,000	1,000	0,000
	7	67,7406	0,000	1,000	0,000	0,000	1,000	0,000
	8	66,9949	0,000	1,000	0,000	0,000	1,000	0,000
	9	68,0782	0,000	1,000	0,000	0,000	1,000	0,000
	10	67,1451	0,000	1,000	0,000	0,000	1,000	0,000
MÉDIA		**67,614**	**0,000**	**1,000**	**0,000**	**0,000**	**1,000**	**0,000**
DESV. PADRÃO		**0,44949594**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**	**0,000**

(c) DS1 - Semana 3

Figura 4.7: Resultados dos folds - Dataset 1 (DS1)

Os dados apresentados na Figura 4.7 foram sintetizados para facilitar a leitura dos valores médios obtidos através da técnica de 10-folds. A Tabela 4.9 apresenta os valores obtidos para o percentual de itens corretos e taxas de TP e FP para cada classe.

Tabela 4.9: Síntese resultados DS1

Semana (S)	% Correto	TP			FP		
		EVA	APV	REP	EVA	APV	REP
S1	68,088	0,204	0,963	0,001	0,041	0,876	0,000
S2	67,527	0,000	1,000	0,000	0,000	1,000	0,000
S3	67,614	0,000	1,000	0,000	0,000	1,000	0,000

Quando analisadas as taxas de TP e FP de cada uma das classes, é possível identificar que os dados de S2 e S3 apresentaram valores fixos para as classes em todas as interações dos *folds*, $0,000$, $1,000$ e $0,000$ respectivamente. A Figura 4.7(b), que representa os dados de S2 e a Figura 4.7(c), que representa os dados de S3, demonstram que para essas semanas, o modelo DS1 não obteve capacidade de generalização, classificando todas as instâncias com a classe APV. Esse fato será discutido posteriormente na seção de análise dos resultados.

Resultados DS2

A Figura 4.8 apresenta os resultados obtidos a partir do *Dataset* 2 (DS2) que é composto pelos dados incrementais das semanas de interação.

S1								
Dataset	Fold	% Correto	TP			FP		
			EVA	APV	REP	EVA	APV	REP
S1	1	67,9516	0,185	0,969	0,000	0,035	0,887	0,000
	2	68,0501	0,215	0,958	0,012	0,040	0,868	0,004
	3	68,2372	0,184	0,970	0,000	0,034	0,891	0,000
	4	67,7828	0,192	0,958	0,000	0,046	0,882	0,000
	5	67,7406	0,187	0,962	0,000	0,040	0,887	0,000
	6	69,0068	0,236	0,959	0,000	0,044	0,862	0,000
	7	68,036	0,216	0,958	0,000	0,046	0,868	0,000
	8	67,417	0,224	0,957	0,000	0,050	0,858	0,000
	9	68,4581	0,164	0,971	0,000	0,032	0,903	0,000
	10	68,2004	0,238	0,965	0,000	0,044	0,854	0,000
MÉDIA		**68,088**	**0,204**	**0,963**	**0,001**	**0,041**	**0,876**	**0,000**
DESV. PADRÃO		**0,43498173**	**0,025**	**0,006**	**0,004**	**0,006**	**0,016**	**0,001**

(a) DS2 - Semana 1

S1_S2								
Dataset	Fold	% Correto	TP			FP		
			EVA	APV	REP	EVA	APV	REP
S1_S2	1	73,4947	0,696	0,892	0,160	0,088	0,507	0,029
	2	73,1851	0,722	0,895	0,125	0,085	0,533	0,027
	3	73,1148	0,668	0,904	0,139	0,082	0,547	0,026
	4	73,1851	0,666	0,902	0,139	0,082	0,549	0,027
	5	72,8475	0,716	0,897	0,094	0,089	0,542	0,024
	6	73,9026	0,702	0,903	0,124	0,080	0,534	0,029
	7	73,1851	0,666	0,900	0,141	0,082	0,548	0,026
	8	72,6787	0,685	0,898	0,133	0,088	0,520	0,024
	9	73,0585	0,712	0,885	0,144	0,090	0,524	0,030
	10	72,8437	0,727	0,899	0,106	0,095	0,529	0,021
MÉDIA		**73,150**	**0,696**	**0,898**	**0,131**	**0,086**	**0,533**	**0,026**
DESV. PADRÃO		**0,35055195**	**0,024**	**0,006**	**0,019**	**0,005**	**0,014**	**0,003**

(b) DS2 - Semana 2

S1_S2_S3								
Dataset	Fold	% Correto	TP			FP		
			EVA	APV	REP	EVA	APV	REP
S1_S2_S3	1	81,9077	0,842	0,912	0,440	0,039	0,266	0,074
	2	81,6967	0,847	0,913	0,431	0,035	0,290	0,072
	3	82,0625	0,837	0,932	0,398	0,034	0,316	0,059
	4	81,556	0,860	0,908	0,430	0,040	0,287	0,071
	5	81,8092	0,865	0,920	0,390	0,041	0,303	0,059
	6	82,7237	0,842	0,923	0,440	0,034	0,285	0,065
	7	81,7952	0,838	0,917	0,421	0,037	0,292	0,068
	8	81,8374	0,852	0,914	0,440	0,041	0,277	0,068
	9	82,4001	0,843	0,908	0,479	0,036	0,263	0,074
	10	82,0881	0,852	0,926	0,414	0,039	0,293	0,061
MÉDIA		**81,873**	**0,845**	**0,916**	**0,431**	**0,038**	**0,289**	**0,068**
DESV. PADRÃO		**0,34903031**	**0,009**	**0,008**	**0,025**	**0,003**	**0,016**	**0,006**

(c) DS2 -Semana 3

Figura 4.8: Resultados dos folds - Dataset 2 (DS2)

Os resultados obtidos no DS2 estão sintetizados na Tabela 4.10, onde é possível verificar que o percentual médio de itens classificados corretamente foi de aproximadamente 68% na semana 1 (S1), 73% para a semana 2 (S1_S2) e 81% para a semana 3 (S1_S2_S3). Como os dados de S1 não variaram do DS1 para o DS2, essa semana não foi considerada nesse momento. Analisando os valores isolados de S1_S2, a classe APV obteve valores de 89% para TP e 53% para FP, REP com 13% de TP e 2% de FP e a classe EVA com 69% de TP e 8% de FP. Os dados de S1_S2_S3 apresentaram os melhores percentuais de TP e FP onde foram classificadas para a classe APV com 91% de TP e 28% de FP, para a classe REP foram 43% de TP com 6% de FP, e por fim, a classe EVA com 84% de TP com 3% de FP.

Tabela 4.10: Síntese resultados DS2

Semana (S)	**% Correto**	**TP**			**FP**		
		EVA	**APV**	**REP**	**EVA**	**APV**	**REP**
S1	68,088	0,204	0,963	0,001	0,041	0,876	0,000
S1_S2	73,150	0,696	0,898	0,121	0,086	0,533	0,026
S1_S2_S3	81,873	0,845	0,916	0,431	0,038	0,289	0,068

Os resultados obtidos com o DS2 registraram um aumento considerável para a taxa de assertividade, representado pelos itens classificados como TP e diminuiu para os índices de erro que foram classificados como FP. A comparação entre esses resultados será melhor debatida no Capítulo de Discussão de Resultados.

Resultados com rebalanceamento de classes - (Rs)

Conforme proposto na fase de Modelagem, após a seleção de qual das técnicas gerou o melhor resultado, no caso a modelagem DS2, foi utilizada uma técnica para o balanceamento das classes a partir do algoritmo *RESAMPLE* disponível na ferramenta WEKA. Esse algoritmo consiste em produzir uma subamostra aleatória de um conjunto de dados, baseada em amostragem com e sem substituição, para igualar a quantidade de instâncias entre as classes, gerando um *dataset* balanceado. A definição de um *dataset* balanceado consiste na existência de uma quantidade similar de instâncias para as classes presentes em uma base a ser analisada.

A Figura 4.9 apresenta os resultados obtidos através da técnica de *cross validation* utilizando como referências os dados presentes em DS2 devidamente balanceados.

S1									
Dataset	Fold	% Correto	TP			FP		
			EVA	APV	REP	EVA	APV	REP
S1	1	49,7468	0,827	0,429	0,470	0,287	0,150	0,255
	2	49,3528	0,802	0,423	0,509	0,270	0,152	0,275
	3	49,0011	0,820	0,425	0,476	0,300	0,148	0,250
	4	49,8734	0,836	0,433	0,482	0,291	0,148	0,248
	5	48,4102	0,854	0,413	0,444	0,302	0,137	0,260
	6	49,7046	0,819	0,439	0,462	0,281	0,157	0,257
	7	48,762	0,825	0,417	0,482	0,295	0,139	0,262
	8	48,9308	0,860	0,416	0,460	0,315	0,127	0,245
	9	48,2977	0,843	0,412	0,463	0,310	0,137	0,252
	10	50,3166	0,843	0,454	0,417	0,300	0,161	0,230
MÉDIA		**49,240**	**0,833**	**0,426**	**0,467**	**0,295**	**0,146**	**0,253**
DESV. PADRÃO		**0,66694527**	**0,018**	**0,013**	**0,025**	**0,013**	**0,010**	**0,012**

(a) Rs - Semana 1

S2								
Dataset	Fold	% Correto	TP			FP		
			EVA	APV	REP	EVA	APV	REP
S2	1	45,3292	0,894	0,398	0,275	0,413	0,097	0,201
	2	46,511	0,897	0,433	0,236	0,405	0,129	0,178
	3	45,484	0,898	0,417	0,254	0,412	0,105	0,192
	4	46,3281	0,881	0,429	0,264	0,417	0,120	0,171
	5	45,9623	0,880	0,422	0,248	0,418	0,128	0,174
	6	44,8227	0,868	0,404	0,281	0,407	0,103	0,206
	7	44,0349	0,881	0,385	0,289	0,416	0,103	0,208
	8	46,1452	0,879	0,424	0,264	0,417	0,115	0,177
	9	46,117	0,873	0,403	0,260	0,407	0,119	0,203
	10	46,7145	0,891	0,438	0,240	0,416	0,126	0,166
MÉDIA		**45,745**	**0,884**	**0,415**	**0,261**	**0,413**	**0,115**	**0,188**
DESV. PADRÃO		**0,83190177**	**0,010**	**0,017**	**0,017**	**0,005**	**0,012**	**0,016**

(b) Rs - Semana 2

S3								
Dataset	Fold	% Correto	TP			FP		
			EVA	APV	REP	EVA	APV	REP
S3	1	44,4851	0,902	0,367	0,340	0,476	0,035	0,171
	2	45,695	0,912	0,398	0,314	0,473	0,041	0,152
	3	43,6832	0,917	0,373	0,306	0,474	0,044	0,173
	4	43,2893	0,903	0,363	0,327	0,481	0,050	0,168
	5	45,076	0,922	0,374	0,348	0,478	0,037	0,158
	6	45,09	0,914	0,382	0,345	0,467	0,039	0,165
	7	45,1463	0,928	0,377	0,346	0,471	0,041	0,162
	8	45,3433	0,905	0,384	0,344	0,479	0,042	0,152
	9	44,9071	0,898	0,380	0,345	0,467	0,033	0,171
	10	46,1095	0,918	0,397	0,335	0,417	0,045	0,149
MÉDIA		**45,083**	**0,913**	**0,379**	**0,342**	**0,474**	**0,041**	**0,164**
DESV. PADRÃO		**0,86044236**	**0,010**	**0,011**	**0,015**	**0,019**	**0,005**	**0,009**

(c) Rs - Semana 3

Figura 4.9: Resultados dos folds utilizando RESAMPLE

A síntese dos resultados obtidos através do balanceamento das classes, é apresentado na Tabela 4.11.

Tabela 4.11: Síntese resultado RESAMPLE (Rs)

Semana (S)	**% Correto**	**TP**			**FP**		
		EVA	**APV**	**REP**	**EVA**	**APV**	**REP**
Rs - S1	49,240	0,833	0,426	0,467	0,295	0,146	0,253
Rs - S2	67,701	0,864	0,588	0,579	0,148	0,138	0,246
Rs - S3	75,626	0,855	0,746	0,739	0,040	0,100	0,214

O rebalanceamento das classes proporcionou um ajuste ao modelo, porém, é possível observar que os resultados obtidos em S2 e S3 não superaram os índices alcançados com o

dataset DS2. Para S1 o balanceamento de classes proporcionou uma melhoria considerável para as classes EVA e REP, porém, para a classe APV os resultados de TP decaíram consideravelmente com uma melhoria para os itens classificados como FP. Esses resultados e a comparação entre eles serão analisados no próximo Capítulo, Discussão dos Resultados.

Capítulo 5

DISCUSSÃO DOS RESULTADOS

Nesse capítulo, serão discutidos os resultados obtidos em relação à composição dos datasets, para avaliar qual dos modelos propostos possibilitou melhores resultados em relação aos índices globais de assertividade e também as taxas de TP e FP. Em seguida, serão discutidos os resultados em relação à técnica de balanceamento das classes para os dados do DS2 e comparados aos resultados originais. Ao final, serão discutidos os resultados obtidos com a indução das árvores de decisão para o conjunto de dados de DS2, rebalanceados com a técnica *RESAMPLE*.

Conforme a proposta de modelo apresentada no Capítulo 4 e os objetivos apresentados no Capítulo 1, esse projeto de mineração de dados partiu do estudo e análise de um modelo desenvolvido com os dados históricos das interações dos alunos com um AVA, na EAD corporativa governamental, para descoberta de conhecimentos para o suporte ao combate à reprovação e evasão.

Nesse sentido, foram realizadas atividades para validar uma melhor composição de *dataset* considerando a granularidade temporal (semanal) dos dados. Para realização dos experimentos, foram criados dois *datasets* de referência, ambos compostos por dados das interações dos alunos com o AVA, separados por semanas.

O *dataset* 1 (DS1) foi criado com os dados das interações das três semanas de realização dos cursos, para análise de forma isolada, ou seja, ao final de cada semana são analisadas as interações dos alunos utilizando o algoritmo C4.5. O segundo *dataset* (DS2) é composto pelos dados das interações semanais, porém, de forma acumulada até o final de cada uma das três semanas propostas para análises e o *dataset* RS que possui os dados de DS2 com as classes balanceadas.

5.1 COMPOSIÇÃO DOS *DATASETS*

A partir das referências apresentadas, foram realizados os experimentos utilizando dados relativos à oferta de cursos em EAD para mais de 70 mil alunos da Enap durante 2015 e

2016. O objetivo inicial dos experimentos foi a comparação em relação ao comportamento dos dados de acordo com a composição utilizada. Para o DS2, que é composto pelos dados das semanas de forma agrupada, os resultados para o índice total de assertividade e de TP e FP registrados foram consideravelmente superiores, conforme destacado na Tabela 5.1

Tabela 5.1: Comparação dos resultados obtidos em DS1 e DS2

Semana (S)	**% Correto**	**TP**			**FP**		
		EVA	**APV**	**REP**	**EVA**	**APV**	**REP**
DS1 - S2	67,527	0,000	1,000	0,000	0,000	1,000	0,000
DS2 - S2	**73,150**	**0,696**	**0,898**	**0,131**	**0,086**	**0,533**	**0,026**
DS1 - S3	67,614	0,000	1,000	0,000	0,000	1,000	0,000
DS2 - S3	**81,873**	**0,845**	**0,916**	**0,431**	**0,038**	**0,289**	**0,068**

O dataset DS2, que é composto pelos dados das semanas de forma incremental, obteve melhores resultados quando comparado ao DS1. A semana 1 (S1) não foi considerada nessa comparação pois os valores em ambos cenários são idênticos. Quando analisados os dados de DS2 em relação à semana 2 e semana 3, é possível identificar que o modelo, ao contrário do acontecido em DS1, obteve uma maior capacidade de generalização, porém, as taxas alcançadas para TP e FP ainda apresentam discrepâncias que merecem atenção.

Em síntese, os dados presentes no DS2 apresentaram resultados mais confiáveis quando comparados ao DS1. O fato de as classes EVA e REP apresentarem valores nulos para TP e FP caracteriza que o modelo com essas estrutura dos dados não é passível de utilização. Outro fator importante que pode ser observado é que a classe APV apresenta altos níveis de TP nas três semanas, com 96% na S1, e 100% para S2 e S3. Porém, esses altos índices de TP estão acompanhados de altos valores de FP, com 87% na S1 e 100% na S2 e S3. Essa situação caracteriza que o modelo está ”chutando”as respostas, classificando todos os resultados como APV.

Outro fator passível de análise são os resultados de TP e FP obtidos com o DS2. A classe EVA apresentou índices consideráveis de TP, com 69% , e um valor de FP que pode ser considerado aceitável, com 8%. Com isso, é possível afirmar que o modelo DS2 teve a capacidade de acertar a classificação de alunos evadidos com 69% de classificações corretas e 8% de classificações incorretas. Para a classe APV, os valores melhoraram com a utilização do DS2, porém, ainda apresentam índices críticos de TP e FP que não caracterizam uma boa classificação, com 89% e 53% respectivamente. Por fim, a classe REP apresentou baixos níveis de TP, somente 13% e 2% de FP, o que não caracteriza um bom desempenho para essa classe. Analisando o modelo DS2 à vista de S2, é possível identificar que a classe APV, devido aos índices de TP e FP, compromete os demais resultados, inviabilizando a possibilidade de aplicação e utilização desse modelo em ambientes de produção.

Para os dados até a terceira semana (S3) do *dataset* DS2, os valores obtidos para as três classes foram consideravelmente superiores à semana anterior (S2) e também quando comparados aos resultados obtidos na mesma semana do *dataset* DS1. A classe EVA apresentou

índices consideráveis para TP, com 84% aliados a uma taxa de somente 3% de FP. Os valores para APV e REP também apresentaram grande evolução. A classe APV apresentou índice de 91% para TP e 28% para FP, registrando um decréscimo considerável nas taxas de FP, decaindo de 53% para 28%. Porém, ainda é possível afirmar que o modelo está deficiente na classificação para a classe APV pois, por exemplo, para cada 100 alunos, o modelo está acertando a classificação de 53 como APV e está indicando como aprovados outros 28 alunos que, na verdade, foram reprovados ou evadiram. A classe REP apresentou evolução nos índices de TP, subindo o índice para 43% com 6% de FP, caracterizando uma deficiência no modelo para essa classe.

Os resultados analisados até esse ponto demonstraram que o *dataset* DS2, quando consideradas todas as classes, apresentou o valor médio global superior para as taxas de TP aliado a um valor médio global inferior para as taxas de FP, obteve o melhor desempenho na análise das interações.

5.2 BALANCEAMENTO DE CLASSES

Os resultados obtidos com o *dataset* composto pelos dados com as classes balanceadas (RS) registraram índices de total de acertos inferiores ao *dataset* DS2 em todas as três análises semanais propostas pelo modelo. Entre as classes analisadas, as taxas de TP para a classe APV decaíram consideravelmente em todas as semanas e as taxas de FP aumentaram, caracterizando uma maior complexidade em classificar alunos aprovados quando a quantidade de alunos presentes nas outras classes é igual. Para as classes EVA e REP, os resultados melhoraram significativamente com um grande aumento nas taxas de TP, porém, as taxas de FP também aumentaram consideravelmente. A Tabela 5.2 apresenta em destaque os resultados obtidos através de RS.

Tabela 5.2: Comparação dos resultados de *RESAMPLE* (RS) e DS2

Semana (S)	**% Correto**	**TP**			**FP**		
		EVA	**APV**	**REP**	**EVA**	**APV**	**REP**
DS2 - S1	68,088	0,204	0,963	0,001	0,041	0,876	0,000
RS - S1	**49,240**	**0,833**	**0,426**	**0,467**	**0,295**	**0,146**	**0,253**
DS2 - S2	73,150	0,696	0,898	0,131	0,086	0,533	0,026
RS - S2	**67,701**	**0,864**	**0,588**	**0,579**	**0,148**	**0,138**	**0,246**
DS2 - S3	81,873	0,845	0,916	0,431	0,038	0,289	0,068
RS - S3	**75,626**	**0,855**	**0,746**	**0,739**	**0,040**	**0,100**	**0,214**

Analisando os resultados apresentados na Tabela 5.2, é possível confirmar que a utilização de técnicas para o tratamento de dados com classes desbalanceadas proporcionam um acréscimo considerável na qualidade dos resultados obtidos com a classificação supervisionada via árvores de decisão.

Quando comparadas as taxas de TP e FP presentes em DS2 e RS, é possível identificar que o decréscimo dos índices para a classe APV se deu paralelamente ao aumento das taxas para as classes EVA e REP. Tal fato leva em consideração que os alunos classificados anteriormente como APV, devido a quantidade minoritária de instâncias presentes nas classes REP e EVA, foram reclassificados quando utilizados dados de maior representatividade em relação à distribuição entre as classes.

Em síntese, os resultados obtidos em RS apresentaram os melhores índices de acertos e erros, com uma taxa média global (todas as classes) de TP em torno de 68% e 17% para FP.

5.3 ANÁLISE DO MODELO

Após a validação de qual composição de dados proporciona melhores resultados e do balanceamento das classes dessa composição, foi induzido um modelo de árvore de decisão para cada uma das semanas analisadas. Uma das vantagem de modelos baseados em árvores é a facilidade na sua interpretação.

Os modelos em árvores de decisão possibilitam o entendimento das características descritivas dos padrões de interações que, quando combinados, levam a uma determinada classe. Esses padrões descrevem como os alunos, quando generalizados pelo modelo proposto, se comportam em relação à utilização dos objetos de aprendizagem disponíveis no AVA.

Semana 1

A partir dos dados de S1 em RS, semana 1 dos dados rebalanceados, o modelo inferiu a árvore de decisão apresentada na Figura 5.1.

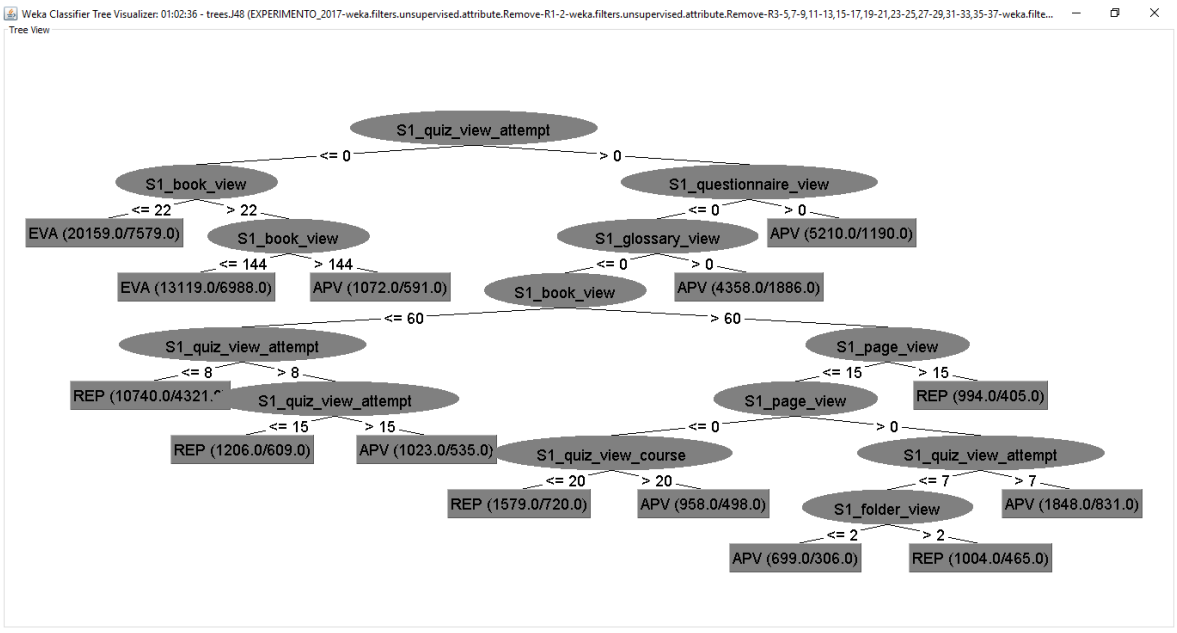


Figura 5.1: Árvore de decisão gerada a partir de RS para a primeira semana

A leitura de uma árvore de decisão pode ser realizada através de conjuntos de regras do tipo se, $entao$ $(IF, THEN)$, que apresentam algumas características relacionadas aos perfis dos alunos de acordo com a nota final obtida. O nó inicial de uma árvore caracteriza o atributo com as características mais influenciadoras do modelo. No caso de S1, o atributo caracterizado como nó inicial está relacionado às interações realizadas com o módulo de exercícios de fixação (quiz_view_attempt). Analisando o próximo nível, podemos identificar que a separação da árvore partiu de valores (≤ 0) e (> 0). Analisando os dados, pode-se induzir regras que servem como indicadores, por exemplo:

- **se** ($S1\_quiz\_view\_attempt > 0$) **e** ($S1\_questionnaire\_view > 0$) $\rightarrow$ **APV**;
- **se** ($S1\_quiz\_view\_attempt \leq 0$) **e** ($S1\_book\_view \leq 22$) $\rightarrow$ **EVA**;
- **se** ($S1\_quiz\_view\_attempt \leq 0$) **e** ($S1\_book\_view > 22$) **e** ($S1\_book\_view > 144$) $\rightarrow$ **APV**
- **se** ($S1\_quiz\_view\_attempt \leq 0$) **e** ($S1\_book\_view > 22$) **e** ($S1\_book\_view \leq 144$) $\rightarrow$ **EVA**

Conforme pode ser observado nas regras geradas a partir da árvore induzida para a semana 1, não houve uma generalização para a classe APV no primeiro nível da árvore. Para essa semana, o modelo inferiu a importância de duas variáveis para poder generalizar os alunos classificados como APV a partir dos atributos $S1\_quiz\_view\_attempt$ e $S1\_questionnaire\_view$.

Tendo por base os dados apresentados, uma forma de utilização dos resultados pelos Coordenadores de curso da Enap seria considerar, por exemplo, que ao final da primeira semana os alunos não realizarem nenhuma interação com o módulo de atividades e menos de 22 interações com o módulo de conteúdos, eles estão tendenciosos à evasão e que caso essa quantidade de interações alcance 144, eles poderão reverter essa tendência para uma aprovação. Como exemplo de possível intervenção, poderiam ser enviadas mensagens automáticas para os alunos que obtiveram essas características ao final da primeira semana, incentivando que eles utilizem melhor os conteúdos de referência disponíveis nos cursos, onde tal comunicação pode trazer resultados interessantes no combate à evasão, conforme pode ser observado em [Almeida et al. 2016].

Outro fator relevante refere-se a questão dos atributos selecionados como nós em uma árvore de decisão. Somente os atributos com capacidade de generalização são utilizados para a construção de árvores com o algoritmo C4.5. Conforme apresentado no Capítulo 1, os atributos são avaliados de acordo com o valor da informação, calculado através do ganho de informação (*Information Gain - InfoGain*). Considerando esses fatores, uma informação valiosa extraída desse modelo é que entre os atributos analisados, os que estão presentes na árvore apresentada na Figura 5.1 destacam-se como foco de atenção. Nesse sentido, pode-se identificar que, entre os onze atributos de entrada, para os dados da primeira semana no *dataset* RS, seis possuem maior representatividade:

- $S1\_quiz\_view\_attempt$
- $S1\_questionnaire\_view$
- $S1\_book\_view$
- $S1\_page\_view$
- $S1\_folder\_view$
- $S1\_glossary\_view$

A partir desses resultados, seria possível, por exemplo, a Enap considerar essas informações para a construção de novos cursos, repassando esses dados aos conteudistas como orientação para valorização dos objetos de aprendizagem que representam maiores contribuições para o sucesso dos alunos ao final dos cursos. Essa mesma interpretação pode ser replicada aos resultados alcançados em todas as semanas analisadas.

Semana 2

A árvore apresentada na Figura 5.2 tem como nó inicial a característica relacionada às tentativas realizadas nos exercícios avaliativos dos cursos, ou seja, quantas vezes os alunos submeteram as atividades, ao final da segunda semana de realização dos cursos. Essa interação é representada na árvore pelo atributo ($questionnaire\_submitted$).

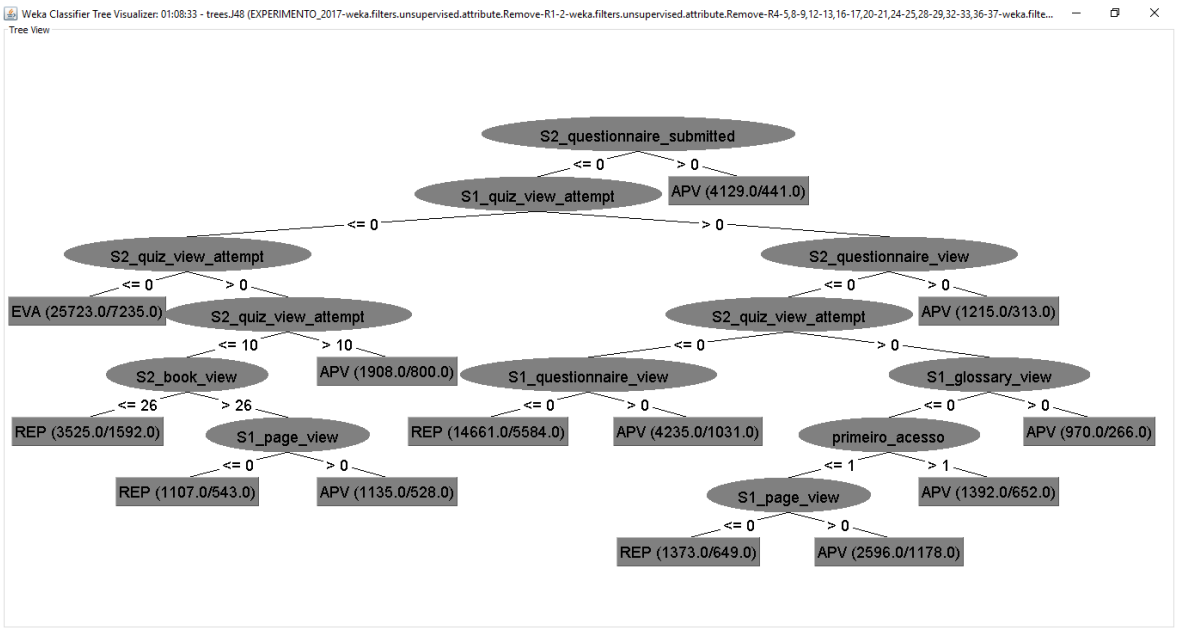


Figura 5.2: Árvore de decisão gerada para semana 2 - S2

A partir desse modelo é possível extrair algumas regras que foram induzidas, representando indicadores para a segunda semana, por exemplo:

- **se** ($S2\_questionnaire\_submitted > 0$) $\rightarrow$ **APV**;
- **se** ($S2\_questionnaire\_submitted \leq 0$) **e** ($S1\_quiz\_view\_attempt \leq 0$) **e** ($S2\_quiz\_view\_attempt \leq 0$) $\rightarrow$ **EVA**;
- **se** ($S2\_questionnaire\_submitted \leq 0$) **e** ($S1\_quiz\_view\_attempt \leq 0$) **e** ($S2\_quiz\_view\_attempt > 0$) **e** ($S2\_quiz\_view\_attempt \leq 10$) **e** ($S2\_book\_view \leq 26$) $\rightarrow$ **REP**;

Quando analisados os atributos presentes na segunda semana, é possível identificar que o modelo, diferente da semana 1, utilizou o atributo $primeiro\_acesso$ como nó de decisão para as classes APV e REP.

Semana 3

Como os dados analisados para inferência das árvores de decisão consideraram os dados acumulados das semanas de interações, os resultados obtidos ao final da terceira semana apresentam os atributos de todas as três semanas em sua estrutura, conforme apresentado na Figura 5.3

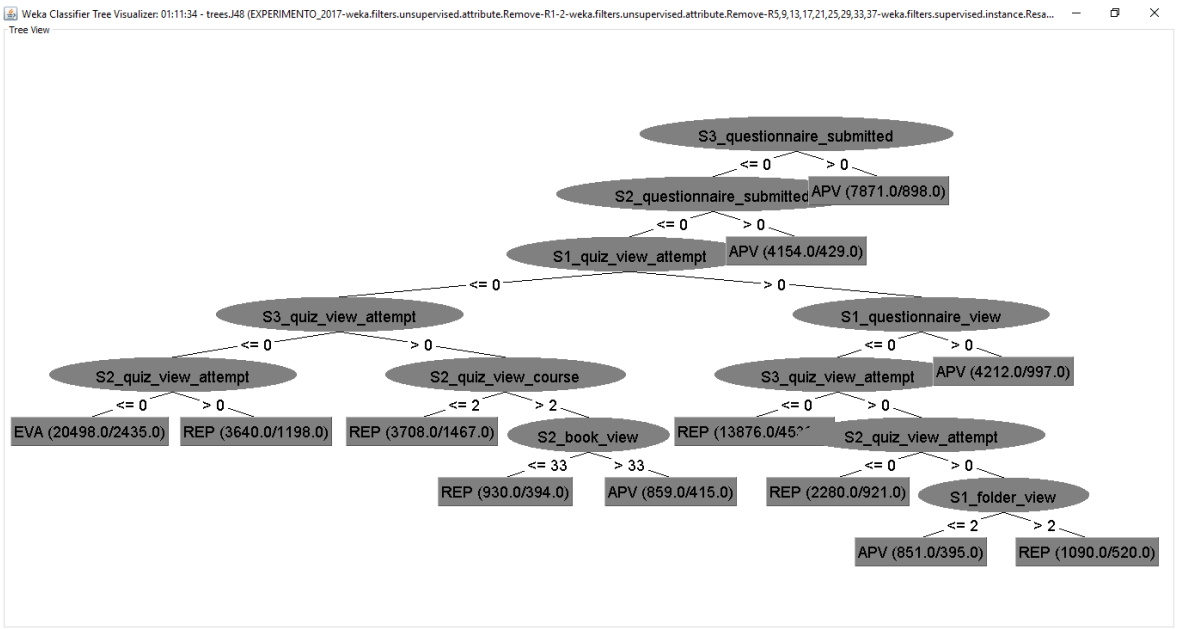


Figura 5.3: Árvore de decisão gerada para semana 3 - S3

Na árvore induzida pelos dados da terceira semana, o atributo de partida para os nós de decisão foi o ($S3\_questionnaire\_submitted$) que representa as interações com o módulo de exercícios pontuados. Percebe-se que a referência para esse indicador é o nível de iteração superior ou igual a zero, tendo em vista que não existem números negativos (menor que 0) nesse *dataset*, o que pode caracterizar que nos cursos analisados, o fato do aluno ter alguma interação com esses atributos ao final da terceira semana caracteriza um perfil de sucesso. Algumas regras foram extraídas, entre elas:

- **se** ($S3\_questionnaire\_submitted > 0$) $\rightarrow$ **APV**;

- **se** ($S3\_questionnaire\_submitted \leq 0$) **e** ($S2\_questionnaire\_submitted \leq 0$) **e** ($S1\_quiz\_view\_attempt \leq 0$) **e** ($S3\_quiz\_view\_attempt \leq 0$) **e** ($S2\_quiz\_view\_attempt \leq 0$) $\rightarrow$ **EVA**;

- **se** ($S3\_questionnaire\_submitted \leq 0$) **e** ($S2\_questionnaire\_submitted \leq 0$) **e** ($S1\_quiz\_view\_attempt \leq 0$) **e** ($S3\_quiz\_view\_attempt \leq 0$) **e** ($S2\_quiz\_view\_attempt > 0$) $\rightarrow$ **REP**;

Nesse momento é possível identificar que o perfil dos alunos das classes REP e EVA possuem características bem parecidas, somente com o atributo $S2\_quiz\_view\_attempt$ delimitando a divisão entre os alunos. Outro fato que pode ser observado nas árvores induzidas para as três semanas é que os nós principais de partida estão relacionados aos módulos de exercícios e que a interação com esses módulos foi a característica principal para classificação dos alunos aprovados. Já para os alunos evadidos e reprovados, além da interação com os módulos de exercícios o módulo livro também influenciou em todas as semanas analisadas.

Outra observação interessante é que os atributos relacionados aos exercícios, *quiz* e *questionnaire*, estão presentes nos primeiros níveis da hierarquia das árvores induzidas em todas as semanas e que os atributos relacionados aos conteúdos como *book*, *page* e *folder*, estão nos níveis mais baixos das árvores. Esse indicadores podem, por exemplo, caracterizar um curso que não explore tanto os conteúdos utilizados nos módulos para a resposta dos exercícios. Essas informações podem ser utilizadas pelos coordenadores de cursos da Enap como insumo para demandar a melhoria ou criação de cursos em seu catálogo, visando uma melhoria qualitativa no processo de utilização dos objetos educacionais.

5.3.1 Síntese dos resultados

Em síntese, é possível afirmar, com base nos resultados obtidos, que o perfil dos alunos que são aprovados ao final dos cursos com tutoria ofertados pela Enap, é caracterizado pela interação em todas as semanas com os módulos de exercícios e questionários pontuados. Pode-se afirmar também que, na primeira semana de realização dos cursos, diferentemente das outras semanas, os atributos relacionados aos conteúdos dos cursos como *book* e *glossary* estão presentes nos perfis de ambas as classes.

Em relação à geração de indicadores para auxiliar o combate dos índices de reprovação e evasão, é possível propor indicadores de alerta para riscos de insucesso nos cursos para cada uma das semanas, através de relatórios direcionados para os coordenadores de cursos na Enap. Um exemplo de indicadores poderia se dar da seguinte forma:

- Indicador de evasão (S1) - Listagem de alunos com acessos entre $[22...144]$ cliques

(interações) no módulo de conteúdos (*book*) e que ainda não tenham realizado nenhuma interação com o módulo de exercícios;

- Indicador de evasão (S2) - Listagem de alunos que não interagiram com os módulos de exercícios;
- Indicador de reprovação (S2) - Listagem de alunos que não interagiram com os módulos de exercícios nem com o módulo de conteúdo externo (*page*) e que tenham acessado os conteúdos (*book*) menos que 26 vezes;
- Indicador de reprovação (S3) - Listagem de alunos que não interagiram com o módulo de exercício avaliativo na primeira e segunda semana juntamente com 2 ou menos interações com o módulo de exercício de fixação na segunda semana e menos de 33 acessos ao módulo de conteúdos.

Outra informação valiosa está relacionada ao perfil de comportamento dos alunos de acordo com a classe da nota final obtida. Com base nos resultados é possível identificar que os alunos classificados como EVA não interagem com os módulos de exercícios (*quiz* e *questionnaire*) em nenhuma das três semanas e que, especificamente para a classe REP, essa ausência de interação está relacionada a uma baixa interação com o módulo de conteúdos (*book*). Pode-se identificar também que os alunos aprovados possuem característica relacionadas principalmente com a interação com os módulos de exercícios, porém, sem relacionamento com os outros módulos. Tal comportamento pode caracterizar alunos que procuram ”burlar” os sistema, indo diretamente para os questionários sem navegar nos conteúdos para reciclagem e fixação.

Em vista dos resultados alcançados, foi possível comprovar que a utilização de um modelo de classificação supervisionada através de árvores de decisão é eficiente para geração de indicadores relacionados às interações dos alunos com os objetos de aprendizagem presentes em um AVA, presente na oferta da EAD em cenários corporativos governamentais.

Capítulo 6

CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho de mestrado consistiu no estudo de técnicas de Mineração de Dados Educacionais, especificamente a classificação supervisionada com árvores de decisão, para análise de uma proposta de modelo preditivo de desempenho utilizando dados provenientes de um Ambiente Virtual de Aprendizagem de uma instituição de ensino corporativo com atuação em âmbito governamental.

Para construção do modelo proposto, foram extraídos dados relacionados às interações dos alunos com um AVA, que representam a quantidade de cliques em cada um dos módulos presentes nos cursos ofertados, entre os anos de 2015 e 2016 pela Escola Nacional de Administração Pública. As conclusões serão apresentadas com o intuito de responder algumas das questões abordadas na Seção 1.3 além de outros pontos cercados por novas indagações surgidas ao longo dos experimentos.

Realizaram-se três experimentos para validação das composições de dados utilizando o algoritmo de classificação supervisionada C4.5. Os resultados apresentaram que a composição de dados que utilizou as interações semanais de forma consecutiva (DS2) unida à da técnica de balanceamento *RESAMPLE* (RS) obteve o melhor desempenho, quando consideradas as taxas médias globais de itens classificados como TP e FP que alcançaram 68% e 17% respectivamente. Vale destacar que essa taxa média global considera a média de TP como a taxa de acerto e da média de FP para a taxa de erro, considerando as taxas alcançadas para as três classes de forma agrupada.

Um fator que pôde ser confirmado é que a performance do modelo aumenta de acordo com a quantidade de dados presentes como referência. Ou seja, ao final da primeira semana de interações, os dados não possuem tantas características para uma classificação eficiente. Contudo, quando analisados os dados da segunda e terceira semana, o modelo melhorou a capacidade de generalização devido a proporção de informações presentes nos dados analisados.

Ao avaliar os resultados obtidos ao final de cada uma das semanas propostas, nota-se as taxas de classificação correta para as duas primeiras semanas, 49,24% e 67,70% respectivamente, não se encaixam nos padrões da literatura como um bom resultado para projetos de MDE [Wu et al. 2008]. Porém, para os dados da terceira semana a taxa de classificação correta foi de 75,62%, caracterizando um resultado com valores aceitáveis. Neste sentido é possível concluir que a composição de dados proposta é passível de implementação para uma análise preditiva de desempenho ao final da terceira semana de realização dos cursos sem tutoria na Enap.

Em síntese, a composição de dados incremental, aliada à técnica de balanceamento propostas no modelo utilizado, possibilita a geração de indicadores que podem auxiliar o combate à reprovação e evasão. Pode-se acompanhar os níveis de interação dos alunos no decorrer da realização de novas ofertas de cursos que possuam as mesmas características dos que foram utilizados como base para criação dos *datasets*, através de um monitoramento semanal.

Nesse sentido, é possível concluir que a aplicação de técnicas de MDE em ambientes de educação corporativa para a oferta de cursos de curta duração, utilizando modelos de classificação supervisionada com árvores de decisão, pode gerar indicadores promissores para análise preditiva de desempenho e disponibilizar informações descritivas sobre os padrões de interação dos alunos.

A principal contribuição deste trabalho para a área de Mineração de Dados Educacionais está no estudo da técnica de Aprendizado Supervisionado utilizando o algoritmo C4.5 para a geração de indicadores, presentes nas regras induzidas por árvores de decisão, que podem auxiliar a Enap no combate às taxas de evasão e reprovação em cursos sem tutoria.

Outra contribuição importante foi a utilização de dados de uma instituição de ensino corporativo, fora do contexto de grande parte dos trabalhos em MDE que estão centralizados na análise de dados oriundos de Instituições de Ensino Superior. Essa contribuição comprova que as técnicas já consagradas na aplicação de MDE via classificação supervisionada, também são passíveis de aplicação em dados gerados em ambiente de EAD corporativa.

6.1 TRABALHOS FUTUROS

Como perspectivas de trabalhos futuros, são levantadas as seguintes indagações:

- Construir de um *Data Warehouse* específico para análise de dados relacionados à interação dos alunos com o intuito de centralizar essas informações em um ambiente apropriado para armazenamento de dados históricos e que facilite possíveis consultas posteriores;
- Unir as informações de interação dos alunos ao tipo de conteúdo que está sendo uti-

lizado nos módulos, de forma a pontuar a interação dos alunos de acordo com a importância de cada um dos módulos que ele interage;

- Utilizar o modelo proposto considerando cursos de outras instituições de EAD corporativa;
- Estudar o modelo proposto para análise de dados de uma Instituição de Ensino Superior a fim de verificar se as interações, mesmo em cursos de longa duração, podem gerar indicadores que auxiliem no processo de tomada de decisão;
- Integralizar as funcionalidades de ETL e Aprendizado de Máquina em um módulo compatível com o AVA Moodle para análise de interações em períodos específicos de tempo.
- Aplicar outras técnicas de classificação, por exemplo o algoritmo C5.0 (que ainda é de formato proprietário), para validar o modelo proposto bem como outras técnicas de Aprendizado de Máquina, como por exemplo o aprendizado não-supervisionado.

REFERÊNCIAS BIBLIOGRÁFICAS

[Abbad 2007] Abbad, G. d. S. (2007). Educação a distância: o estado da arte e o futuro necessário. *Revista do Serviço Público*, 58(3):351–374.

[ABED 2015] ABED (2015). br 2010. relatório analítico da aprendizagem da educação a distância no brasil. associação brasileira de educação a distância.

[Adhatrao et al. 2013] Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., and Honrao, V. (2013). Predicting students' performance using id3 and c4. 5 classification algorithms. *arXiv preprint arXiv:1310.2071*.

[Adriaans and Zantinge 1996] Adriaans, P. and Zantinge, D. (1996). Data mining. harlow. *England: Addison Wesley*.

[Agrawal et al. 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM.

[Albertin and Brauer 2012] Albertin, A. L. and Brauer, M. (2012). Resistência à educação a distância na educação corporativa. *Revista de Administração Pública*, 46(5):1367–1389.

[Almeida et al. 2016] Almeida, L. R. d., da Costa, J. P. C. L., Sousa Júnior, R. T. d., Freitas, E. P., Canedo, E. D., Prettz, J., Zacarias, E., and Galdo, G. D. (2016). Motivating attendee's participation in distance learning via an automatic messaging plugin for the moodle platform. In *Frontiers in Education Conference (FIE)*. IEEE.

[Almeida et al. 2013] Almeida, O. C. d. S. d., Abbad, G., Meneses, P. P. M., and Zerbini, T. (2013). Evasão em cursos a distância: fatores influenciadores. *Revista Brasileira de Orientação Profissional*, 14(1):19–33.

[Alpaydin 2014] Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

[Baker et al. 2010] Baker, R. et al. (2010). Data mining for education. *International encyclopedia of education*, 7(3):112–118.

[Baker et al. 2011a] Baker, R., Isotani, S., and Carvalho, A. (2011a). Mineraçao de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, 19(02):03.

[Baker et al. 2009] Baker, R. S., de Carvalho, A., Raspat, J., Aleven, V., Corbett, A. T., and Koedinger, K. R. (2009). Educational software features that encourage and discourage "gaming the system". In *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, pages 475–482.

[Baker and Yacef 2009] Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*, 1(1):3–17.

[Baker et al. 2011b] Baker, R. S. J., Isotani, S., and de Carvalho, A. M. J. B. (2011b). Mineração de dados educacionais: oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19(2).

[Barbieri 2011] Barbieri, C. (2011). *BI2: business intelligence: modelagem e qualidade*. Campus.

[Barros et al. 2012] Barros, R. C., Basgalupp, M. P., De Carvalho, A. C., and Freitas, A. A. (2012). A survey of evolutionary algorithms for decision-tree induction. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(3):291–312.

[Baruque et al. 2007] Baruque, C. B., Amaral, M. A., Barcellos, A., da Silva Freitas, J. C., and Longo, C. J. (2007). Analysing users' access logs in moodle to improve e learning. In *Proceedings of the 2007 Euro American conference on Telematics and information systems*, page 72. ACM.

[Brachman and Anand 1996] Brachman, R. J. and Anand, T. (1996). The process of knowledge discovery in databases. In *Advances in knowledge discovery and data mining*, pages 37–57. American Association for Artificial Intelligence.

[Breiman et al. 1984] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. wadsworth & brooks. *Monterey, CA*.

[Bresfelean 2007] Bresfelean, V. P. (2007). Analysis and predictions on students' behavior using decision trees in weka environment. In *Proceedings of the Information Technology Interfaces (ITI)*, pages 25–28. IEEE.

[Bunkar et al. 2012] Bunkar, K., Singh, U. K., Pandya, B., and Bunkar, R. (2012). Data mining: Prediction for performance improvement of graduate students using classification. In *Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on*, pages 1–5. IEEE.

[Coelho et al. 2015] Coelho, V. C. G., Costa, J. P. C. L. d., Souza, D. d. C. R. d., Canedo, E. D., Silva, D. G. e., and Sousa Júnior, R. T. d. (2015). Mineração de dados educacionais para identificação de barreiras na utilização da educação a distância. In *21º Congresso Internacional ABED de Educação a Distância*. ABED.

[Coelho et al. 2016] Coelho, V. C. G., da Costa, J. P. C., da Silva, D. A., de Sousa Júnior, R. T., de Mendonça, F. L., and Silva, D. G. (2016). Mineração de dados educacionais no ensino a distância governamental. In *Conferências Ibero-Americanas WWW/Internet e Computação Aplicada 2016*, pages 1–10. CIAWI.

[Costa et al. 2013] Costa, E., Baker, R. S., Amorim, L., Magalhães, J., and Marinho, T. (2013). Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29.

[Cover and Thomas 2012] Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[de Souza Mendes et al. 2014] de Souza Mendes, A., de Sousa Junior, R. T., Martins, V. A., and de Deus, F. E. G. (2014). Application of data mining techniques in the characterization of internal personnel turnover. In *Information Systems and Technologies (CISTI), 2014 9th Iberian Conference on*, pages 1–6. IEEE.

[Duda et al. 2012] Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

[Fayyad et al. 1996a] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996a). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

[Fayyad et al. 1996b] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996b). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

[Fayyad et al. 1996c] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. (1996c). *Advances in knowledge discovery and data mining*, volume 21. AAAI press Menlo Park.

[Frank and Hall 2001] Frank, E. and Hall, M. (2001). A simple approach to ordinal classification. In *European Conference on Machine Learning*, pages 145–156. Springer.

[Frank et al. 2005] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., and Trigg, L. (2005). Weka. *Data Mining and Knowledge Discovery Handbook*, pages 1305–1314.

[Garcia et al. 2013] Garcia, S., Luengo, J., Sáez, J. A., Lopez, V., and Herrera, F. (2013). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750.

[Gil 2010] Gil, A. C. (2010). Métodos e técnicas de pesquisa social. In *Métodos e técnicas de pesquisa social*. Atlas.

[Goldschmidt and Bezerra 2015] Goldschmidt, R. and Bezerra, E. (2015). *Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações*. Elsevier Brasil.

[Gottardo et al. 2014] Gottardo, E., Kaestner, C. A. A., and Noronha, R. V. (2014). Estimativa de desempenho acadêmico de estudantes: Análise da aplicaçao de técnicas de mineraçao de dados em cursos a distância. *Revista Brasileira de Informática na Educação*, 22(01):45.

[Guleria et al. 2014] Guleria, P., Thakur, N., and Sood, M. (2014). Predicting student performance using decision tree classifiers and information gain. In *I.C. on Parallel, Distributed and Grid Computing*, pages 126–129. IEEE.

[Hall et al. 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

[Hämäläinen and Vinni 2010] Hämäläinen, W. and Vinni, M. (2010). Classifiers for educational data mining. *Handbook of educational data mining*, pages 57–74.

[Hand et al. 2001] Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of data mining*. MIT press.

[Hoe et al. 2013] Hoe, A. C. K., Ahmad, M. S., Hooi, T. C., Shanmugam, M., Gunasekaran, S. S., Cob, Z. C., and Ramasamy, A. (2013). Analyzing students records to identify patterns of students' performance. In *Research and Innovation in Information Systems (ICRIIS), 2013 International Conference on*, pages 544–547. IEEE.

[Huang 1997] Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*, pages 21–34. Citeseer.

[Jindal and Borah 2015] Jindal, R. and Borah, M. D. (2015). Predictive analytics in a higher education context. *IT Professional*, 17(4):24–33.

[Kabakchieva 2013] Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, 13(1):61–72.

[Kaelbling et al. 1996] Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

[Kimball and Ross 2011] Kimball, R. and Ross, M. (2011). *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons.

[Kohavi and Provost 1998] Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine Learning*, 30(2-3):271–274.

[Kotsiantis and Pintelas 2003] Kotsiantis, S. and Pintelas, P. (2003). Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, 1(1):46–55.

[Lakshmi et al. 2013] Lakshmi, T. M., Martin, A., Begum, R. M., and Venkatesan, V. P. (2013). An analysis on performance of decision tree algorithms using student's qualitative data. *International Journal of Modern Education and Computer Science*, 5(5):18.

[Li et al. 2001] Li, W., Han, J., and Pei, J. (2001). Cmar: Accurate and efficient classification based on multiple class-association rules. In *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, pages 369–376. IEEE.

[Lin et al. 2013] Lin, C. F., Yeh, Y.-c., Hung, Y. H., and Chang, R. I. (2013). Data mining for providing a personalized learning path in creativity: An application of decision trees. *Computers & Education*, 68:199–210.

[Luger 2013] Luger, G. F. (2013). *Inteligência Artificial - Tradução Daniel Vieira*. Pearson Education do Brasil, 6 edition.

[MacQueen et al. 1967] MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

[Márquez-Vera et al. 2016] Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., and Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124.

[Marquez-Vera et al. 2013] Marquez-Vera, C., Morales, C. R., and Soto, S. V. (2013). Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 8(1):7–14.

[Mishra et al. 2014] Mishra, T., Kumar, D., and Gupta, S. (2014). Mining students' data for prediction performance. In *Advanced Computing & Communication Technologies (ACCT), 2014 Fourth International Conference on*, pages 255–262. IEEE.

[Mitchell et al. 1997] Mitchell, T. M. et al. (1997). Machine learning.

[Monard and Baranauskas 2003] Monard, M. C. and Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. *Sistemas Inteligentes-Fundamentos e Aplicações*, 1(1).

[Moran 2002] Moran, J. M. (2002). O que é educação a distância. *São Paulo*.

[Peña-Ayala 2014] Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, 41(4):1432–1462.

[Pereira et al. 2007] Pereira, A. T. C., Schmitt, V., and Dias, M. (2007). Ambientes virtuais de aprendizagem. *AVA-Ambientes Virtuais de Aprendizagem em Diferentes Contextos. Rio de Janeiro: Editora Ciência Moderna Ltda*, page 23.

[Prass et al. 2004] Prass, F. S. et al. (2004). Estudo comparativo entre algoritmos de análise de agrupamentos em data mining.

[Quinlan 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

[Quinlan 1993] Quinlan, J. R. (1993). C4. 5: Programming for machine learning. *Morgan Kauffmann*, 38.

[Quinlan 1996] Quinlan, J. R. (1996). Improved use of continuous attributes in c4. 5. *Journal of artificial intelligence research*, 4:77–90.

[Rokach and Maimon 2014] Rokach, L. and Maimon, O. (2014). *Data mining with decision trees: theory and applications*. World scientific.

[Romero et al. 2013a] Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., and Ventura, S. (2013a). Web usage mining for predicting final marks of students that use moodle courses. *Computer Applications in Engineering Education*, 21(1):135–146.

[Romero et al. 2013b] Romero, C., Olmo, J. L., and Ventura, S. (2013b). A meta-learning approach for recommending a subset of white-box classification algorithms for moodle datasets. In *Educational Data Mining 2013*.

[Romero and Ventura 2007] Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146.

[Romero and Ventura 2013] Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

[Romero et al. 2008] Romero, C., Ventura, S., and García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*, 51(1):368–384.

[Romero 2010] Romero, Cristóbal ; Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

[Russell et al. 1995] Russell, S., Norvig, P., and Intelligence, A. (1995). A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25:27.

[Sammut and Webb 2011] Sammut, C. and Webb, G. I. (2011). *Encyclopedia of machine learning*. Springer Science & Business Media.

[Shearer 2000] Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

[Simon 1983] Simon, H. A. (1983). Why should machines learn? In *Machine learning*, pages 25–37. Springer.

[Tan et al. 2009] Tan, P.-N., Steinbach, M., and Kumar, V. (2009). *Introdução ao datamining: mineração de dados*. Ciência Moderna.

[Thai-Nghe et al. 2009] Thai-Nghe, N., Busche, A., and Schmidt-Thieme, L. (2009). Improving academic performance prediction by dealing with class imbalance. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*, pages 878–883. IEEE.

[Wirth and Hipp 2000] Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39. Citeseer.

[Witten and Frank 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Witten et al. 2016] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Wu et al. 2008] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., et al. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37.