UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

JOSÉ GONÇALVES DE OLIVEIRA JÚNIOR

IDENTIFICAÇÃO DE PADRÕES PARA A ANÁLISE DA EVASÃO EM CURSOS DE GRADUAÇÃO USANDO MINERAÇÃO DE DADOS EDUCACIONAIS

DISSERTAÇÃO

Curitiba
2015

JOSÉ GONÇALVES DE OLIVEIRA JÚNIOR

IDENTIFICAÇÃO DE PADRÕES PARA A ANÁLISE DA EVASÃO EM CURSOS DE GRADUAÇÃO USANDO MINERAÇÃO DE DADOS EDUCACIONAIS

Dissertação submetida ao Programa de Pós-Graduação em Computação Aplicada da Universidade Tecnológica Federal do Paraná como requisito parcial para a obtenção do título de Mestre em Computação Aplicada.

Área de concentração: *Engenharia de Sistemas Computacionais*

Orientador: Prof. Dr. Celso Antônio Alves Kaestner
Coorientador: Prof. Dr. Robinson Vida Noronha

Curitiba
2015

Dados Internacionais de Catalogação na Publicação

O48i 2015

Oliveira Júnior, José Gonçalves de
Identificação de padrões para a análise da evasão em cursos de graduação usando mineração de dados educacionais / José Gonçalves de Oliveira Júnior.-- 2015.
86 f. : il. ; 30 cm

Texto em português, com resumo em inglês
Dissertação (Mestrado) - Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Computação Aplicada, Curitiba, 2015
Bibliografia: f. 69-72

1. Universidade Tecnológica Federal do Paraná. 2. Mineração de dados (Computação). 3. Evasão escolar. 4. Computação - Dissertações. I. Kaestner, Celso Antônio Alves, orient. II. Noronha, Robinson Vida, coorient. III. Universidade Tecnológica Federal do Paraná - Programa de Pós-graduação em Computação Aplicada. IV. Título.

CDD: Ed. 22 -- 621.39

Biblioteca Central da UTFPR, Câmpus Curitiba

ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO Nº 38

Aos 08 dias do mês de agosto de 2015 realizou-se na sala B-205 a sessão pública de Defesa da Dissertação de Mestrado intitulada "Identificação de Padrões para a Análise da Evasão em Cursos de Graduação usando Mineração de Dados", apresentada pelo aluno **José Gonçalves de Oliveira Júnior** como requisito parcial para a obtenção do título de Mestre em Computação Aplicada, na área de concentração “Engenharia de Sistemas Computacionais”, linha de pesquisa “Sistemas Inteligentes e Lógica”.

Constituição da Banca Examinadora:

Prof. Dr. Celso Antonio Alves Kaestner, UTFPR - CT (Presidente) _____

Prof. Dr. Robinson Vida Noronha, UTFPR- CT _____

Prof. Dr. Leandro Augusto da Silva, UP - Mackenzie _____

Prof. Dr. Laudelino Cordeiro Bastos, UTFPR- CT_____

Prof. Dr. Hilton José Silva de Azevedo, UTFPR- CT_____

Em conformidade com os regulamentos do Programa de Pós-Graduação em Computação aplicada e da Universidade Tecnológica Federal do Paraná, o trabalho apresentado foi considerado _____ (aprovado/reprovado) pela banca examinadora. No caso de aprovação, a mesma está condicionada ao cumprimento integral das exigências da banca examinadora, registradas no verso desta ata, da entrega da versão final da dissertação em conformidade com as normas da UTFPR e da entrega da documentação necessária à elaboração do diploma, em até _____ dias desta data.

Ciente (assinatura do aluno): _____

(para uso da coordenação)

A Coordenação do PPGCA/UTFPR declara que foram cumpridos todos os requisitos exigidos pelo programa para a obtenção do título de Mestre.

Curitiba PR, _____/_____/_____ _____

"A Ata de Defesa original está arquivada na Secretaria do PPGCA".

Agradecimentos

A Deus por ter me dado saúde e inspiração para a realização deste gratificante trabalho.

À minha família, a qual amo muito, pelo carinho, paciência e incentivo.

Ao meu orientador, prof. Celso Antônio Alves Kaestner, pela disponibilidade, colaboração, dedicação e paciência com que me conduziu durante a realização deste trabalho. Foi um privilégio tê-lo como meu orientador.

Ao meu coorientador, prof. Robinson Vida Noronha, pelo incentivo e valiosas contribuições para este trabalho.

Ao prof. Marco Aurelio Wehrmeister, coordenador do PPGCA, pela colaboração e auxílio no atendimento às demandas solicitadas à coordenação do programa.

Aos professores Hilton José Silva de Azevedo, Laudelino Cordeiro Bastos e Adolfo Gustavo Serra Seca Neto pelos pertinentes apontamentos e sugestões realizados na apresentação do projeto e nos seminários de acompanhamento desta pesquisa.

Aos professores Cesar Augusto Tacla, Myriam Regattieri de Biase da Silva Delgado, Murilo Vicente Gonçalves da Silva, Nadia Puchalski Kozievitch, Laudelino Cordeiro Bastos, Gustavo Alberto Gimenez Lugo, Maria Cláudia Figueiredo Pereira Emer, Robinson Vida Noronha e Celso Antônio Alves Kaestner pelos relevantes conhecimentos transmitidos em suas disciplinas.

Aos professores Leandro Augusto da Silva, Hilton José Silva de Azevedo, Robinson Vida Noronha, Laudelino Cordeiro Bastos e Celso Antônio Alves Kaestner pela participação na banca de defesa e pelos apontamentos realizados que contribuíram para a melhoria deste trabalho.

Ao prof Maurício Alves Mendes, Pró-Reitor de Graduação da UTFPR, pelo apoio incondicional à realização desta pesquisa.

Ao Ivantuil Lapuente Garrido, Diretor de Gestão de Tecnologia da Informação da UTFPR, pelo apoio na aplicação deste projeto para as necessidades da instituição.

À Rosane Beatriz Zanetti Putz pelo incentivo e apoio para a realização deste trabalho.

Resumo

OLIVEIRA JÚNIOR, José Gonçalves. Identificação de Padrões para a Análise da Evasão em Cursos de Graduação Usando Mineração de Dados Educacionais. 2015. 86 f. Dissertação - Programa de Pós-graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

A mineração de dados educacionais é uma área recente de pesquisa que está ganhando popularidade por causa de seus potenciais para as instituições de ensino. Um dos desafios dessas instituições é a redução da evasão escolar. A evasão no ensino superior é um fenômeno em crescimento e tornou-se foco de preocupação para pesquisadores de diferentes áreas. Entretanto, as características da evasão ainda são pouco estudadas e há carência de informações e modelos de identificação dos seus motivos. Esta pesquisa propõe uma abordagem computacional para a identificação de padrões a serem utilizados na análise da evasão de estudantes em cursos presenciais de graduação, a fim de auxiliar os tomadores de decisão das instituições de ensino. Propõe-se um método para seleção dos melhores atributos para tarefa de classificação, que considera as classes "haverá evasão" e "não haverá evasão", baseado na seleção e criação de atributos. Os experimentos foram realizados com dados de alunos da Universidade Tecnológica Federal do Paraná, consolidados em um *Data Warehouse*, que permitiu investigar a evasão entre os anos de 1980 e 2014. Nesta pesquisa são abordados os problemas mais comuns que ocorrem na mineração de dados educacionais, como a seleção do subconjunto de atributos, dados desbalanceados, valores discrepantes e sobreajuste. Os resultados experimentais apresentam os atributos mais relevantes a previsão da evasão, indicando a contribuição da criação de atributos na tarefa de mineração de dados, permitindo com estas inferências apoiar a tomada de decisão pelos gestores educacionais situados nos níveis estratégico, tático e operacional.

Palavras-chave: Mineração de Dados Educacionais. Criação de Atributos. Seleção de Subconjunto de Atributos.

Abstract

OLIVEIRA JÚNIOR, José Gonçalves. Pattern Identification for Dropout Analysis in Undergraduate Courses using Educational Data Mining. 2015. 86 f. Dissertação - Programa de Pós-graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

Educational data mining is a recent research area that is gaining popularity because of their potential for educational institutions. One of the challenges of these institutions is to reduce the course dropout. The dropout in higher education is a phenomenon in growth and has become the focus of concern for researchers from different areas. However, the avoidance features are poorly studied and there is a lack of information and identification of models of their motives. This research proposes a computational approach for identifying patterns to be used in the analysis of dropout students in undergraduate classroom courses, in order to assist decision-makers in educational institutions. The proposed method selects the best attributes for classification task, in which the classes "dropout" and "non-dropout" are considered, based on the feature subset selection and feature creation. The experiments were conducted with the undergraduate students' data at the Federal University of Technology - Paraná, consolidated in a Data Warehouse, that allowed the dropout investigation between the years 1980 and 2014. In this research are discussed the most common problems that occur in educational data mining, such as feature subset selection, unbalanced data, outliers and overfitting. The experimental results show the most relevant attributes to dropout prediction, indicating the contribution of the feature creation in the data mining task, allowing with these inferences to support the decision-making by educational managers located in strategic, tactical and operational levels.

Keywords: Educational Data Mining. Feature Creation. Feature Subset Selection.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Lista de Abreviações

DM	*Data Mining*
DW	*Data Warehouse*
EDM	*Educational Data Mining*
ENEM	Exame Nacional do Ensino Médio
FSS	*Feature Subset Selection*
IES	Instituições de Ensino Superior
KDD	*Knowledge Discovery in Databases*
MEC	Ministério da Educação
ML	*Machine Learning*
MLP	*Multilayer Perceptron*
OLAP	*Online Analytical Processing*
SISU	Sistema de Seleção Unificada
SMO	*Sequential Minimal Optimization*
SMOTE	*Synthetic Minority Oversampling Technique*
SVM	*Support Vector Machines*
UTFPR	Universidade Tecnológica Federal do Paraná
WEE	*WEKA Experiment Environment*
WEKA	*Waikato Environment for Knowledge Analysis*

Capítulo 1

Introdução

Detectar antecipadamente quais estudantes não terão êxito na conclusão do curso tem sido um grande desafio para a comunidade acadêmica e pesquisadores da área de educação. Essa possível detecção antecipada poderia fornecer informações que permitissem a tomada de decisões pelos gestores educacionais (e.g. coordenadores de curso, diretores de ensino, entre outros) para reverter esta situação. Na busca por um dispositivo ou mecanismo inteligente que seja capaz de realizar essa detecção antecipada, alguns pesquisadores da área de Informática em Educação têm empregado técnicas computacionais de mineração de dados. Nesse contexto, bases de dados acadêmicas (e.g. Sistemas de Controle Acadêmico e Ambientes Virtuais de Aprendizagem) têm sido investigadas por meio de algoritmos de mineração de dados [Baker et al., 2011], [Gottardo et al., 2014], [Rigo et al., 2012] e [Borges et al., 2015].

Os estudos relativos à evasão iniciaram com referenciais e teorias que buscam explicar a evasão e a retenção. Apesar de não haver um conceito homogêneo, a partir de 1970, autores como Tinto [Tinto, 1975] passaram a abordar o modelo de integração do estudante, destacando que a decisão de se evadir é tomada em função da falta de integração com o ambiente acadêmico e social da instituição, sendo esta integração influenciada pelas características individuais, pelas expectativas para a carreira ou curso e pelas intenções e compromissos assumidos antes do início do curso.

Um dos primeiros trabalhos a sistematizar o problema da evasão no Brasil foi realizado a partir de uma comissão nacional, instituída pelo Ministério da Educação (MEC). A Comissão Especial para o Estudo da Evasão nas Universidades Brasileiras [SESu/MEC, 1996] surgiu dentro de um contexto de discussão de avaliação institucional, definido pelos indicadores do Programa de Avaliação Institucional das Universidades Brasileiras (PAIUB), realizado por diferentes instituições de ensino, especificamente as públicas. Antes desse trabalho, os estudos realizados enfatizavam apenas levantamentos estatísticos e estudos de casos de forma fragmentada, realizados por iniciativa do MEC e de universidades públicas. Entretanto, tais estudos não desenvolveram a problemática de forma a criar políticas institucionais, avaliações, ações administrativas e pedagógicas, ou seja, acompanhamentos necessários para minimizar o problema. Assim, essa iniciativa foi um primeiro esforço conjunto de diferentes Instituições de Ensino Superior (IES) públicas para organizar de forma sistemática um estudo que procurou definir uma única metodologia, objetivando identificar causas e possíveis soluções para o problema. Os objetivos finais dessa Comissão foram esclarecer o conceito de evasão, analisar as taxas e as causas desse fenômeno e uniformizar uma metodologia a ser empregada pelas instituições.

Nos estudos da Comissão Especial para Estudo da Evasão [SESu/MEC, 1996], encontram-se também pesquisas sobre o desempenho de universidades europeias e norte-americanas numa série histórica de 1960 a 1986. Nessas pesquisas, os melhores rendimentos do sistema universitário são apresentados pela Finlândia, Alemanha, Holanda e Suíça enquanto que os piores resultados se verificam nos Estados Unidos, Áustria, França e Espanha. De acordo com a investigação, nos Estados Unidos as taxas de evasão nos últimos 30 anos estão em torno de 50%. Número semelhante encontra-se na França onde as taxas, em 1980, eram de 60 a 70% em algumas Universidades. Na Áustria, por sua vez, aponta-se uma taxa de evasão de 43%, sendo que apenas 13% dos estudantes concluíram seus cursos nos prazos previstos.

Para auxiliar a análise da evasão, uma alternativa de busca de informação muito promissora é o uso de "descoberta de conhecimento em bases de dados" e uso de "técnicas de mineração de dados na educação", chamada Mineração de Dados Educacionais (*Educational Data Mining* - EDM) [Márquez-Vera et al., 2013b].

EDM é um campo que explora a estatística, a aprendizagem de máquina (*Machine Learning* - ML) e os algoritmos de mineração de dados (*Data Mining* - DM) sobre os diferentes tipos de dados da área de ensino. Seu objetivo principal é analisar estes tipos de dados, a fim de resolver questões em pesquisa educacional [Sachin and Vijay, 2012].

A previsão da evasão escolar pode ser associada à tarefa de mineração de dados chamada classificação, que tem como objetivo a associação de uma classe a cada elemento considerado, a partir de um conjunto de propriedades (ou atributos previsores) inerentes ao mesmo elemento. No caso da tarefa em questão, cada elemento corresponde a um aluno, e as classes consideradas são "haverá evasão" e "não haverá evasão".

No contexto de mineração de dados, a criação de atributos consiste em criar novos atributos a partir de outros existentes, de modo que informações importantes sejam capturadas em um conjunto de dados mais eficazmente. Neste caso busca-se encontrar elementos de simples interpretação que possam ser usados como indicativos de uma potencial situação de evasão.

A seleção de atributos é uma técnica de mineração de dados aplicada para reduzir a dimensionalidade dos dados, facilitando a aplicação de algoritmos de mineração. A redução da dimensionalidade produz uma representação mais compacta, mais facilmente interpretável do conceito alvo, focalizando a atenção do usuário sobre as variáveis mais relevantes [Witten et al., 2011]. Neste trabalho a seleção de atributos foi empregada para mostrar que os atributos criados são efetivamente relevantes para a previsão da evasão.

Em resumo, este trabalho propõe a criação de um método de seleção dos melhores atributos para classificação, a fim de identificar padrões para a análise da evasão escolar em alunos de cursos presenciais de graduação, a partir de informações acadêmicas, organizadas em um *Data Warehouse*, aplicando-se os algoritmos de mineração de dados para gerar inferências[1] com a finalidade de auxiliar os tomadores de decisão no ambiente educacional. Aplica-se neste estudo uma abordagem computacional, mesmo sabendo-se da limitação da quantidade de atributos que podem ser capturados dos sistemas de gestão acadêmica. Outros trabalhos sobre evasão utilizam uma abordagem ligada à psicologia e/ou à educação, que estão fora do escopo desta pesquisa.

[1] Inferência estatística é o processo pelo qual é possível tirar conclusões acerca da população usando informação de uma amostra, tendo como questão central saber usar os dados da amostra para obter conclusões acerca da população [Morais, 2005].

1.1 Contextualização

Com o desenvolvimento da Internet na década de 1970 e a adoção em larga escala da *World Wide Web* desde os anos 1990 aumentaram a velocidade de geração e coleta de dados de negócio em ritmo exponencial [Chen et al., 2012]. Para atender a essa demanda há a necessidade de uma nova geração de teorias e ferramentas para ajudar os seres humanos na extração de informações úteis (conhecimento) dos volumes de rápido crescimento de dados digitais computacionais.

Essas teorias e ferramentas são o assunto do campo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD). KDD é processo não trivial de identificação de padrões em dados que sejam válidos, inéditos, potencialmente úteis e compreensíveis [Fayyad et al., 1996]. DM é a etapa no processo KDD que consiste na aplicação de algoritmos de descoberta de conhecimento que, considerando limitações computacionais aceitáveis, produzem uma enumeração particular de padrões (ou modelos) a partir de um conjunto de dados [Fayyad et al., 1996].

De acordo com Fayyad [Fayyad et al., 1996], KDD refere-se ao processo global de descoberta de conhecimento útil a partir de dados e DM refere-se a uma determinada etapa neste processo. Muitas vezes, no entanto, se emprega o termo "Mineração de Dados" no contexto amplo, isto é, como sinônimo de KDD. A mineração de dados é a aplicação de algoritmos específicos para extrair padrões de dados. O processo KDD pode envolver significativa iteração e pode conter *loops* (processos de *feedback*) entre quaisquer dois passos. O fluxo básico está ilustrado na Figura 1.1.

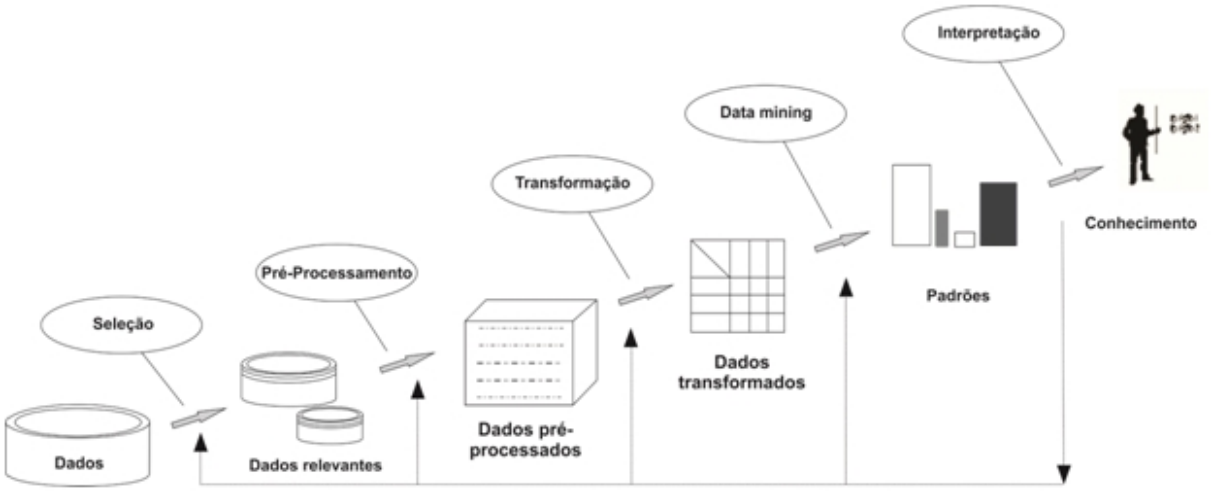


Figura 1.1: Etapas do processo de KDD [Fayyad et al., 1996] (adaptado)

KDD é uma área ativa de pesquisa e continua a evoluir, a partir da interseção de campos de pesquisa, como a aprendizagem de máquina, reconhecimento de padrões, bases de dados, estatística, inteligência artificial, aquisição de conhecimento para sistemas especialistas, visualização de dados e computação de alto desempenho. O objetivo é extrair conhecimento de alto nível a partir de dados de baixo nível no contexto de grandes conjuntos de dados [Fayyad et al., 1996].

Uma área correlata ao KDD é o *Data Warehousing*, que se refere à coleta e limpeza de dados transacionais para torná-los disponíveis para a análise *on-line* e apoio à decisão. *Data*

Warehousing ajuda a definir o cenário para KDD em dois aspectos importantes: (1) limpeza de dados e (2) acesso aos dados [Fayyad et al., 1996].

O foco deste trabalho é gerar inferências, obtidas a partir de um processo de KDD, com a finalidade de auxiliar os tomadores de decisão no ambiente educacional. As decisões dentro de uma organização podem ser classificadas quanto à atividade administrativa a que ela pertence, segundo três níveis [Kendall and Kendall, 2010], conforme indicado na Figura 1.2.

- Nível estratégico - o propósito deste nível é desenvolver estratégias para que a organização seja capaz de atingir seus objetivos. As atividades deste nível não possuem um período com ciclo uniforme, podendo ser irregulares, ainda que alguns planos estratégicos se façam dentro de planejamentos anuais ou em períodos pré-estabelecidos (e.g. Plano de Desenvolvimento Institucional). Nas IESs o nível estratégico da área de ensino é representado, normalmente, por uma pró-reitoria de graduação, que tem a responsabilidade de planejar, coordenar e supervisionar a execução de atividades do ensino.

- Nível tático - as decisões nesse nível são normalmente relacionadas com o controle administrativo e são utilizadas para decidir sobre as operações de controle, formular novas regras de decisão que irão ser aplicadas por parte do pessoal de operação e designação de recursos. Neste nível são necessárias informações sobre o funcionamento planejado (normas, expectativas), variações a partir de um funcionamento planejado, a explicação destas variações e a análise das possibilidades de decisão no curso das ações. Nas IESs o nível tático da área de ensino pode ser representado, por exemplo, pelas diretorias de ensino de câmpus, responsáveis por coordenar e supervisionar a execução das atividades do ensino.

- Nível operacional - a decisão nesse nível é um processo pelo qual se assegura que as atividades operacionais sejam bem desenvolvidas. O controle operacional utiliza procedimentos e regras de decisões pré-estabelecidas. Uma grande parte destas decisões são programadas e os procedimentos a serem seguidos são geralmente muito estáveis. Nas IESs o nível operacional da área de ensino é representado pelas coordenações de curso.

Figura 1.2: Modelo da pirâmide organizacional [Kendall and Kendall, 2010] (adaptado)

Pretende-se assim apoiar a tomada de decisão nas áreas de ensino nos três níveis organizacionais, com especial atenção ao nível operacional, mo qual encontram-se as coordenações de curso.

1.2 Motivação

A evasão escolar faz parte dos debates e reflexões do dia-a-dia da educação brasileira e ocupa espaço de relevância no cenário das políticas de educação pública. Sobre esse assunto existem diversos trabalhos recentes nas áreas de estatística [Silva Filho et al., 2007], educação [Aparecida et al., 2011], psicologia [Bardagi and Hutz, 2014] e computação [Manhães et al., 2011], entre outros.

A evasão gera a cada ano enormes custos financeiros. Segundo o pesquisador Oscar Hipólito, do Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia[2], as perdas financeiras com evasão no ensino superior alcançou em 2009 a marca de 9 bilhões de reais. Em outro estudo[3] realizado pela UNB, Universidade de Brasília, mostra-se que mais de 6,3 mil alunos deixaram de concluir o curso em 2014, gerando um custo financeiro de 95 milhões de reais.

As IESs, de maneira geral, têm desenvolvido iniciativas para tratar o problema da evasão. A Universidade Tecnológica Federal do Paraná (UTFPR) instituiu, em 2014, a Comissão de Evasão e Retenção (portaria 873[4], de 22 de maio de 2014), com a finalidade de realizar uma análise dos índices de evasão e retenção nos cursos de graduação. Essa comissão foi mais um motivador para a presente pesquisa.

Nos trabalhos preliminares da comissão supracitada demandou-se a criação de um *Data Warehouse*, para permitir uma melhor análise dos dados acadêmicos da instituição e auxiliar no diagnóstico do problema. Esse *Data Warehouse* contempla registros acadêmicos de alunos da UTFPR, conforme abordagem detalhada em [Oliveira Júnior et al., 2015].

Outro motivador para esta pesquisa é o estudo da criação e seleção de atributos, permitindo a obtenção de indicadores de fácil interpretação que forneçam aos gestores educacionais um prognóstico da possibilidade de evasão.

Aplicar-se-á neste trabalho uma abordagem computacional para o problema, utilizando-se mineração de dados educacionais, um assunto ainda pouco explorado pela literatura acadêmica.

1.3 Objetivos

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

[2]http://g1.globo.com/educacao/noticia/2011/02/pais-perde-r-9-bilhoes-com-evasao-no-ensino-superior-diz-pesquisador.html

[3]http://www.andifes.org.br/evasoes-na-unb-causam-prejuizo-de-r-95-milhoes-por-ano/

[4]http://www.utfpr.edu.br/estrutura-universitaria/diretorias-de-gestao/diretoria-de-gestao-de-pessoas/portarias-do-reitor/portarias-2014/portaria-0873_2014-designa-comissao-responsavel-pelos-indices-de-evasao-e-retencao-nos-cursos-de-graduacao

1.3.1 Objetivo Geral

Este trabalho tem como objetivo geral investigar a identificação de padrões para auxiliar a tomada de decisão de gestores educacionais com a finalidade de analisar a evasão de estudantes em cursos presenciais de graduação, utilizando as técnicas de mineração de dados "seleção de subconjunto de atributos" e "criação de atributos", a partir de informações disponíveis em bases de dados acadêmicas, para uma melhor representação computacional da evasão.

1.3.2 Objetivos Específicos

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

- Propor uma abordagem computacional para a extração de padrões com a finalidade de analisar a evasão de estudantes em cursos presenciais de graduação.
- Criar um *Data Warehouse* com dados acadêmicos agregados dos alunos de cursos de graduação, com a finalidade de facilitar a aplicação dos algoritmos de mineração e a análise dos dados.
- Gerar inferências para indicar os fatores de possível abandono dos alunos de cursos de graduação, utilizando técnicas de mineração de dados.
- Criar novos indicadores com o intuito de auxiliar na previsão[5] da evasão escolar.
- Analisar a performance dos indicadores criados.
- Realizar experimentos para verificar a aplicabilidade da proposta em situações reais.

1.4 Estrutura do Trabalho

O Capítulo 2 é dedicado ao referencial teórico, apresentando os conceitos envolvendo técnicas computacionais ligadas à área de mineração de dados utilizadas na realização da pesquisa. Destacam-se ainda a apresentação e a análise de trabalhos correlatos.

O Capítulo 3 apresenta um estudo preliminar da evasão e a metodologia de pesquisa utilizada neste trabalho. É apresentado o método proposto para a seleção dos melhores atributos para a classificação, utilizado para realizar as inferências com a finalidade de auxiliar os tomadores de decisão no ambiente educacional.

O Capítulo 4 apresenta os experimentos realizados para a escolha dos melhores atributos a serem utilizados na mineração de dados. São apresentadas também a análise dos resultados obtidos a partir dos experimentos realizados.

Finalizando este documento, o Capítulo 5 apresenta as considerações finais nas quais são apresentadas as conclusões e as possíveis contribuições deste trabalho e suas limitações. Ainda nesse capítulo são apontadas algumas possibilidades de continuidade da pesquisa em trabalhos futuros.

[5] A palavra previsão sugere que se quer ver algo antes que ele exista. Alguns autores preferem a palavra predição e outros utilizam o termo projeção [Morettin and Toloi, 2006]. Neste texto utiliza-se consistentemente a palavra previsão, com o sentido indicado acima.

Capítulo 2

Referencial Teórico

Neste capítulo são apresentados os principais conceitos de mineração de dados utilizados neste trabalho e são analisados diversos trabalhos relacionados ao tema da mineração de dados educacionais, com foco na identificação de padrões para a análise da evasão escolar.

2.1 Fundamentação Teórica

Em mineração de dados educacionais alguns problemas se destacam, que são: os dados desbalanceados, os valores discrepantes, o sobreajuste e a seleção de atributos.

O problema com dados desbalanceados acontece porque os algoritmos de aprendizagem tendem 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 correspondentes a classes pouco representadas [Márquez-Vera et al., 2013b]. Esse problema ocorre na aplicação dos algoritmos de mineração para a análise da evasão. Como exemplo, pode-se citar que nos primeiros semestres cursados a maioria dos alunos encontram-se ativos (regulares, trancados, em mobilidade acadêmica ou sub judice). Uma maneira de resolver esse problema é agir durante o pré-processamento dos dados através da realização de uma amostragem ou balanceamento de distribuição das classes [Márquez-Vera et al., 2013b]. Uma abordagem amplamente utilizada no balanceamento das classes é a aplicação do algoritmo SMOTE (*Synthetic Minority Oversampling Technique*) [Chawla et al., 2002]. Esse algoritmo ajusta a frequência relativa entre classes majoritárias e minoritárias, introduzindo sinteticamente instâncias de classes minoritárias, usando a abordagem K-nn (k vizinhos mais próximos) [Witten et al., 2011].

O sobreajuste (*overfitting*) é um problema importante relacionado com a acurácia[1]. Sobreajuste significa que o modelo foi muito ajustado aos dados de treinamento. O modelo resultante é tão especializado que não se pode generalizar dados futuros [Hämäläinen and Vinni, 2011]. No domínio da educação, sobreajuste é um problema crítico, porque há muitos atributos disponíveis para a construção de um modelo complexo, mas há somente poucos dados para aprendê-lo com precisão [Hämäläinen and Vinni, 2011].

Outro problema comum em EDM são os valores discrepantes (*outliers*), que se referem a pontos de dados que se afastam significativamente da maioria, os quais não se encaixam no mesmo modelo dos demais. Valores discrepantes podem ocorrer devido ao ruído,

[1] Acurácia é uma medida de avaliação do desempenho de um modelo de classificação, que mede a taxa de acerto global, ou seja, o número de classificações corretas dividido pelo número total de instâncias a serem classificadas.

mas em dados educacionais são normalmente observações verdadeiras. Há sempre os alunos excepcionais, que têm sucesso com pouco esforço ou falham contra todas as expectativas [Hämäläinen and Vinni, 2011]. Uma maneira de se identificar os valores discrepantes pode ser feita com a aplicação do cálculo da amplitude interquartil (*Inter Quartile Range* - IQR), utilizada em estatística descritiva como uma medida de dispersão, definida pela diferença entre o $1^o(Q1)$ e o $3^o(Q3)$ quartil. Os limites superiores e inferiores são calculados conforme expressão apresentada abaixo, e os valores fora destes limites são considerados discrepantes:

$$LimiteInferior = max\{min(dados); Q_1 - outlier\_factor(Q_3 - Q_1)\} \tag{2.1}$$

$$LimiteSuperior = min\{max(dados); Q_3 + outlier\_factor(Q_3 - Q_1)\} \tag{2.2}$$

2.1.1 *Data Warehousing*

Um DW é uma coleção de dados orientada por assuntos, integrada, variante no tempo, não volátil, que tem por objetivo dar suporte aos processos de tomada de decisão [Inmon, 2005]. As quatro palavras-chave orientada por assunto, integrada, variante no tempo, e não volátil distinguem DW de outros sistemas de repositório de dados, tais como: sistemas de bancos de dados relacionais, sistemas de processamento de transações e sistemas de arquivos [Han et al., 2011].

Os DWs generalizam e consolidam dados no espaço multidimensional, e a sua construção envolve a limpeza de dados, integração e transformação de dados, e pode ser visto como um passo importante para o pré-processamento na mineração de dados, além de fornecer ferramentas para o processamento analítico *on-line* (*OnLine Analytical Processing* - OLAP), para análise interativa de dados multidimensionais de granularidades variadas, facilitando a mineração de dados [Han et al., 2011].

Os DWs muitas vezes adotam uma arquitetura de três camadas, conforme apresentado na Figura 2.1 [Han et al., 2011]:

- Na camada inferior encontra-se o servidor de banco de dados, que é quase sempre um sistema de banco de dados relacional. Ferramentas de *back-end* e utilitários são usados para alimentar os dados oriundos de bancos de dados operacionais ou outras fontes externas.

- A camada intermediária é composta pelo servidor OLAP, que normalmente é implementado usando um modelo OLAP relacional (ROLAP), i.e. um Sistema de Gerenciamento de Banco de Dados Relacionais (SGBDR) estendido que mapeia as operações em dados multidimensionais para operações relacionais padrão; ou, um modelo multidimensional (MOLAP), i.e. um servidor de propósito específico que implementa diretamente dados e operações multidimensionais.

- A camada superior é composta pelo cliente *front-end*, que contém ferramentas de consulta e de informação, ferramentas de análise e/ou ferramentas de mineração de dados (e.g. análise de tendências, previsão, etc.).

Sistemas de DW usam ferramentas de *back-end* e utilitários para popular e atualizar seus dados (Figura 2.1). Estas ferramentas e utilitários incluem as seguintes funções [Han et al., 2011]:

- extração de dados, que normalmente reúne dados de múltiplas fontes;

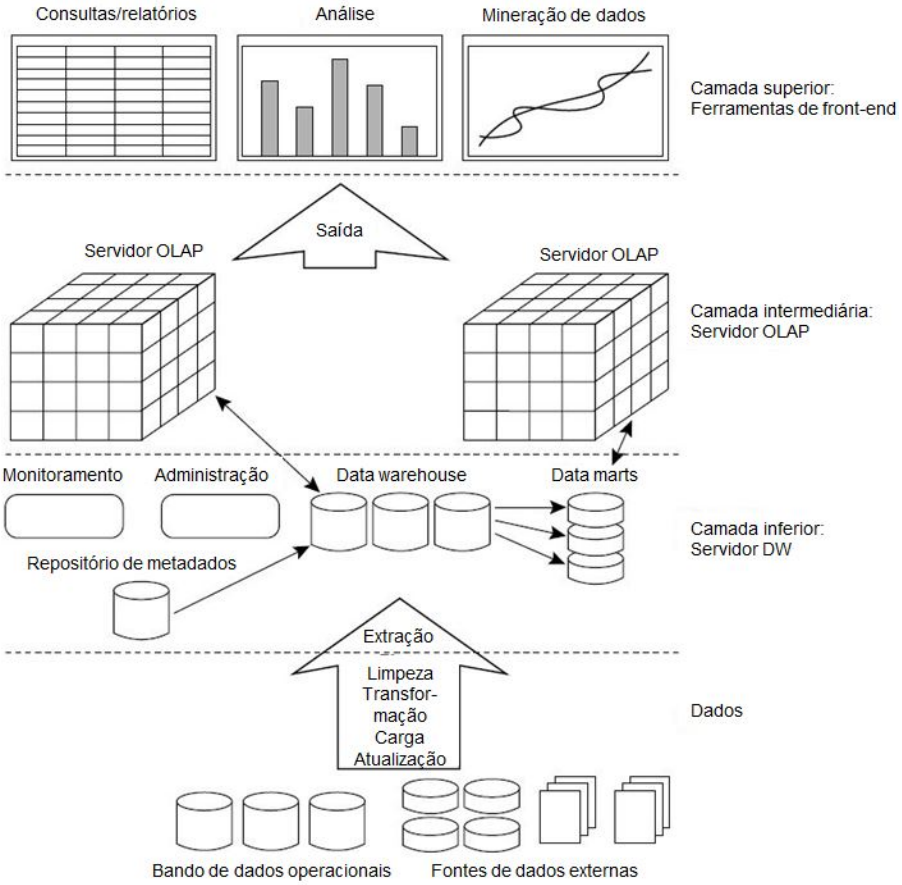


Figura 2.1: Arquitetura de um DW em três camadas [Han et al., 2011] (adaptado)

- limpeza de dados, que detecta erros nos dados para retificá-los quando possível;
- transformação dos dados, que converte os dados para o formato do DW;
- carga, a operação que classifica, sumariza, consolida, computa visualizações, verifica a integridade, e constrói índices e partições; e
- atualização, que propaga as atualizações das fontes de dados para o DW.

A limpeza e transformação de dados são passos importantes para melhorar a qualidade dos dados e, posteriormente, os resultados de mineração de dados. É evidente que a presença de um armazém de dados é um precursor muito útil para exploração de dados, e se este não estiver disponível, muitos dos passos envolvidos no armazenamento de dados terão que ser realizados para preparar os dados para a mineração [Witten et al., 2011].

Uma das tecnologias muitas vezes discutidas no contexto de DW é o processamento de sistemas de gerenciamento de banco de dados multidimensionais, às vezes chamado de processamento *On-Line Analytical Processing* (OLAP) [Inmon, 2005]. Sistemas de gerenciamento de banco de dados multidimensionais fornecem um sistema de informação com a estrutura que permite que uma organização tenha acesso muito flexível aos dados, para separá-los de várias maneiras, e explorar de forma dinâmica a relação entre os dados resumidos e detalhados [Inmon, 2005].

A modelagem multidimensional, uma técnica de projeto lógico normalmente usada para construção de DW, é definida sobre dois pilares: tabelas fato e tabelas dimensão. Tabela fato é a tabela primária em um modelo dimensional, onde as medidas de desempenho numéricas do negócio são armazenadas [Kimball and Ross, 2011]. As tabelas de dimensão são complementos integrais para uma tabela fato. As tabelas dimensão contêm os descritores textuais do negócio. Em um modelo dimensional bem concebido, as tabelas de dimensão têm muitas colunas ou atributos [Kimball and Ross, 2011].

No ambiente de SGBDR uma tabela fato é construída com um registro para cada medição discreta. Esta tabela fato é rodeada por um conjunto de tabelas de dimensão, descrevendo precisamente o que é conhecido no âmbito de cada registro da medição. Este modelo é chamado de *Star schema*, ou modelo Estrela [Kimball and Ross, 2011]. As tabelas de dimensão muitas vezes representam relações hierárquicas no negócio. Assim, a informação descritiva hierárquica é armazenada de forma redundante, mas isto é feito com o espírito de facilidade de uso e desempenho da consulta. Este modelo é chamado *Snowflake schema*, ou modelo Flocos de Neve [Kimball and Ross, 2011].

2.1.2 Métodos de Classificação

A ideia de que a mente humana organiza o seu conhecimento usando o processo natural de classificação é muito frequente. Classificação é o processo de associação de objetos específicos (instâncias) em um conjunto de categorias (classes ou conceitos), com base nas respectivas propriedades do objeto. A classificação é um procedimento em que os itens individuais são colocados em grupos com base na informação oriunda de características inerentes aos itens e com base em um conjunto de treinamento previamente rotulado [Gorunescu, 2011].

A classificação é computacionalmente obtida com o uso de algoritmos de classificação, ou simplesmente classificadores. São empregados nesta pesquisa os seguintes classificadores: árvore de decisão [Romero et al., 2008], classificador baseado em regras [Han et al., 2011], vizinho mais próximo [Hämäläinen and Vinni, 2011], redes neurais [Hämäläinen and Vinni, 2011], máquina de vetores de suporte [Hämäläinen and Vinni, 2011] e *ensemble methods* [Chau and Phung, 2013], pois estes são os principais classificadores encontrados nos trabalhos de EDM.

Uma árvore de decisão é um conjunto de condições organizadas em uma estrutura hierárquica. É um modelo preditivo em que um exemplo é classificado seguindo o caminho de condições satisfeitas a partir da raiz da árvore até atingir uma folha, que vai corresponder a um rótulo de classe [Romero et al., 2008]. As árvores de decisão têm muitas vantagens: elas são simples, fáceis de entender e podem lidar com variáveis de diferentes tipos (variáveis numéricas ou categóricas) [Hämäläinen and Vinni, 2011]. Quando uma árvore de decisão é construída, muitos dos ramos refletem anomalias nos dados de treinamento devido ao ruído ou desvios. Métodos de poda de árvores resolvem este problema de sobreajuste dos dados [Han et al., 2011]. As árvores de decisão são consideradas modelos de fácil compreensão, porque um processo de raciocínio pode ser dado para cada conclusão, exceto se a árvore obtida é muito grande (uma série de nós e folhas) [Romero et al., 2008].

Nos classificadores baseados em regras o modelo aprendido é representado como um conjunto de regras do tipo "*if-then*" (se...então...). As regras são uma boa maneira de representar informação ou pedaços de conhecimento [Han et al., 2011]. Um exemplo desse classificador é o algoritmo JRip, que implementa uma aprendizagem de regra proposicional (*Repeated In-*

cremental Pruning to Produce Error Reduction - RIPPER, ou Poda Incremental Repetida para Produzir Redução de Erro) [Cohen, 1995]. O algoritmo basicamente divide-se em duas fases: a primeira gera um conjunto de regras para a comparação e a segunda otimiza o conjunto de regras iniciais para diminuir erros e tornar o processo mais seletivo, sendo esses passos repetidos inúmeras vezes.

O classificador K-nn, ou K vizinhos mais próximos, é uma técnica baseada em aprendizagem por analogia, ou seja, comparando uma determinada tupla teste com tuplas de treinamento que são semelhantes. As tuplas de treinamento são descritas por n atributos. Cada tupla representa um ponto em um espaço n-dimensional. Desta forma, todas as tuplas de formação são armazenadas num espaço padrão de n dimensões. Quando uma dada tupla é desconhecida, um classificador k-vizinho mais próximo procura o espaço padrão para as tuplas de treinamento k que estão mais próximas da tupla desconhecida. Estas tuplas de treinamento k são os k "vizinhos mais próximos" da tupla desconhecida [Han et al., 2011].

As Redes Neurais Artificiais formam um paradigma também conhecido como redes de processamento paralelo distribuído. Tratam-se de elementos de processamento interconectados chamados de nós ou neurônios que trabalham em conjunto para produzir uma função de saída. As Redes Neurais Feed-Forward (RNFF) são o tipo de redes neurais mais utilizado, que possuem as seguintes camadas de nós: uma para os nós de entrada, uma para nós de saída e pelo menos uma camada de nós ocultos. Em cada camada oculta os nós são conectados aos nós da camada anterior e da camada seguinte, e as arestas estão associadas com pesos individuais. O modelo mais comum contém apenas uma camada oculta. RNFFs também podem representar qualquer tipo (não linear) de limites de classe [Hämäläinen and Vinni, 2011]. A principal desvantagem é que as RNFFs precisam de uma grande quantidade de dados para o treinamento, muito mais do que os conjuntos típicos de dados educacionais contém. Elas são muito sensíveis ao sobreajuste, e o problema é ainda mais crítico com conjuntos de treinamento pequenos. O modelo de rede neural é do tipo "*black box*" (caixa preta) e é difícil de se entender as explicações para os seus resultados [Hämäläinen and Vinni, 2011].

O classificador "Máquina de Vetores de Suporte" (*Support Vector Machines* - SVM) é um método definido inicialmente para dados com separação linear. O objetivo é encontrar o hiperplano de maior margem que separa as classes. No caso de dados que não sejam linearmente separáveis utiliza-se o "Truque do Kernel" (*Kernel Trick*) que mapeia o espaço de atributos original em um espaço de maior dimensão, onde se espera que os dados sejam linearmente separáveis. No entanto, as SVMs têm restrições similares às das redes neurais: os dados devem ser numéricos contínuos (ou quantificados); o modelo não é facilmente interpretável, e a seleção dos parâmetros adequados (especialmente a função de *kernel*) pode ser difícil [Hämäläinen and Vinni, 2011].

Os métodos de classificação denominados "*ensemble methods*" (métodos de conjunto de classificadores) compõe um modelo composto, formado por uma combinação de classificadores. Os classificadores individuais votam e uma previsão de rótulo da classe é devolvida pelo conjunto baseado na coleta de votos. *Ensemble methods* tendem a ser mais precisos do que os classificadores componentes. No algoritmo *Random Forest*, por exemplo, cada modelo do conjunto é construído utilizando uma árvore de decisão, formando assim uma "floresta". Durante a classificação, cada árvore do conjunto "vota" e a classe escolhida é a que obtiver maior número de votos [Han et al., 2011].

Métricas de Avaliação dos Algoritmos de Classificação

A avaliação do desempenho de um modelo de classificação geralmente envolve a análise da habilidade de previsão ou a correta separação das classes. Uma estrutura muito utilizada para essa atividade é conhecida como matriz de confusão [Han et al., 2011]. Em uma matriz de confusão os resultados da classificação são apresentados como uma matriz bidimensional, com uma linha e coluna para cada classe, conforme mostrado no Quadro 2.1. Cada elemento da matriz mostra o número de instâncias corretas ou incorretamente classificadas considerando-se o conjunto de testes utilizado [Fayyad et al., 1996].

A partir de uma matriz de confusão pode-se obter um conjunto de medidas para avaliar o desempenho de um modelo de classificação. Uma dessas medidas é a acurácia, que mede a taxa de acerto global, ou seja, o número de classificações corretas dividido pelo número total de instâncias dos dados a serem classificados. A fórmula da acurácia é definida por:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.3}$$

conforme descrição das variáveis no Quadro 2.1.

	Classe c1 prevista	**Classe c2 prevista**
Classe c1 verdadeira	Taxa de verdadeiro positivo (VP)	Taxa de falso negativo (FN)
Classe c2 verdadeira	Taxa de falso positivo (FP)	Taxa de verdadeiro negativo (VN)

Quadro 2.1: Matriz de confusão para classes binárias
Fonte: Witten et al. [Witten et al., 2011] (adaptado)

A acurácia da classificação em conjunto *r* é medida pela taxa de classificação, que define a proporção de registros corretamente classificados no conjunto *r* [Hämäläinen and Vinni, 2011].

Além da acurácia, Han et al. [Han et al., 2011] destacam outras métricas de avaliação que podem ser utilizadas no desempenho de modelos de classificação, conforme detalhado no Quadro 2.2.

Medida	Fórmula
Acurácia, taxa de reconhecimento	$\frac{VP+VN}{P+N}$
Taxa de erro, taxa de classificação incorreta	$\frac{FP+FN}{P+N}$
Sensibilidade, taxa de verdadeiro positivo, recobrimento	$\frac{VP}{P}$
Especificidade, taxa de verdadeiro negativo	$\frac{VN}{N}$
Precisão	$\frac{VP}{VP+FP}$
F, F_1, *F-score*, média harmônica de precisão, recobrimento	$\frac{2\times precisao}{precisao+recobrimento}$
F_β, onde β é um número real não negativo	$\frac{(1+\beta^2)\times precisao\times recobrimento}{(\beta^2\times precisao)+recobrimento}$
Média Geométrica. Indica o equilíbrio entre os desempenhos de classificação nas classes majoritárias e minoritárias	$MG = \sqrt{VP\times VN}$

Quadro 2.2: Medidas de avaliação de algoritmos de classificação
Fonte: Han et al. [Han et al., 2011] e Márquez-Vera et al. [Márquez-Vera et al., 2013a] (adaptado)

2.1.3 Seleção de Atributos

Um dos pontos chave na aplicação dos algoritmos de aprendizagem é ter dados suficientes para o aprendizado. É fácil verificar que o número de dados de aprendizagem cresce exponencialmente com a dimensão. Este aumento exponencial é a primeira consequência da chamada "Maldição da Dimensionalidade" (*Curse of Dimensionality*). De modo mais geral, a maldição da dimensionalidade é a expressão de todos os fenômenos que aparecem em dados com alta dimensão, e que têm na maioria das vezes consequências indesejáveis sobre o comportamento e desempenho dos algoritmos de aprendizagem [Verleysen and François, 2005], como por exemplo a degradação da acurácia. A redução da dimensionalidade, uma técnica de redução de dados, visa obter uma representação reduzida dos dados, minimizando a perda de conteúdo de informação [Han et al., 2011]. Um princípio lógico aplicado para a redução da dimensionalidade é regido pela Navalha de Occam (*Occam's Razor*), na qual dadas duas explicações para o mesmo fenômeno, deve-se sempre escolher a mais simples [Blumer et al., 1987].

A redução da dimensionalidade dos dados, eliminando atributos inadequados, melhora o desempenho dos algoritmos de aprendizagem [Witten et al., 2011]. Mais importante ainda, a redução de dimensionalidade produz uma representação mais compacta, mais facilmente interpretável do conceito alvo, focando a atenção do usuário sobre as variáveis mais relevantes [Witten et al., 2011].

Algoritmos de aprendizado de máquina, incluindo algoritmos de árvore de decisão, tais como ID3 (Iterative Dichotomiser 3) [Quinlan, 1986], C4.5 (sucessor do ID3) [Quinlan, 1993], e CART (Classification And Regression Tree) [Breiman et al., 1984], e algoritmos baseados em instância, tal como IBL (Instance-Based Learning)[Aha et al., 1991], são

conhecidos por degradar seu desempenho diante de muitos atributos que não são necessários para prever a saída desejada [Kohavi and John, 1997]. Algoritmos como Naïve Bayes são robustos em relação a atributos irrelevantes (ou seja, o seu desempenho degrada muito lentamente quando atributos mais irrelevantes são adicionados), mas, o seu desempenho pode degradar rapidamente se atributos correlacionados são adicionados, mesmo que os atributos sejam relevantes [Kohavi and John, 1997].

O problema da seleção de subconjunto de atributos (*Feature Subset Selection* - FSS) é encontrar um subconjunto de atributos originais de um conjunto de dados de tal forma que um algoritmo de indução, que é executado nos dados contendo apenas esses atributos, gere um classificador com a maior acurácia possível. FSS é uma tarefa de mineração que tem por objetivo reduzir a dimensionalidade dos dados, onde são detectados e removidos atributos irrelevantes, fracamente relevantes ou redundantes [Han et al., 2011]. A seleção do subconjunto de atributos possui duas abordagens principais: *filter* e *wrapper*.

A abordagem *filter*, mostrada na Figura 2.2, seleciona os atributos usando uma etapa de pré-processamento. A principal desvantagem dessa abordagem é que ela ignora totalmente os efeitos do subconjunto de atributos selecionados no desempenho do algoritmo de indução [Kohavi and John, 1997]. Em contrapartida, a abordagem *filter* possui um menor custo computacional, quando comparado com a abordagem *wrapper*.

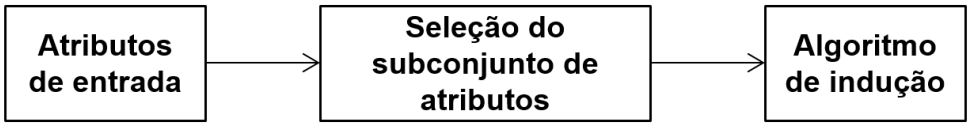


Figura 2.2: Abordagem *filter* para a seleção de um subconjunto de atributos [Kohavi and John, 1997] (adaptado)

Na abordagem *wrapper* [John et al., 1994], o algoritmo de seleção do subconjunto de atributos existe como um invólucro em torno do algoritmo de indução. O algoritmo de seleção do subconjunto de atributos realiza uma pesquisa para um bom subconjunto usando o próprio algoritmo de indução, como parte da função de avaliação de subconjuntos de atributos.

A ideia por trás da abordagem *wrapper*, mostrada na Figura 2.3, é simples. O algoritmo de indução é executado no conjunto de dados, geralmente dividido em conjuntos de treinamento e validação. O subconjunto de atributos com a avaliação mais alta é escolhido como o último conjunto no qual se deve executar o algoritmo de indução. A avaliação é feita usando validação cruzada com fator n. A validação cruzada é um procedimento de teste experimental largamente utilizado na avaliação de algoritmos de classificação. A ideia é dividir um conjunto de dados em K subconjuntos disjuntos de tamanhos aproximadamente iguais. Em seguida, executa-se K experimentos, onde, em cada experimento, o subconjunto de ordem K é removido. O sistema é treinado com o restante dos dados, e em seguida, o sistema de formação é testado no subconjunto mantido fora. No final desta validação cruzada K vezes, cada exemplo foi utilizado num teste definido uma única vez. Este procedimento tem a vantagem de que todos os conjuntos de teste são independentes [Witten et al., 2011]. O classificador resultante é, então, avaliado em um conjunto de teste independente que não foi usado durante a pesquisa.

Para se percorrer o espaço de busca, cada estado representa um subconjunto de atributos. Para n atributos, existem n *bits* em cada estado, e cada *bit* indica se um recurso está presente (1) ou ausente (0). Os operadores determinam a conectividade entre os estados, e utilizam-se os operadores que adicionam ou excluem um único atributo de um estado. A Figura 2.4 mostra

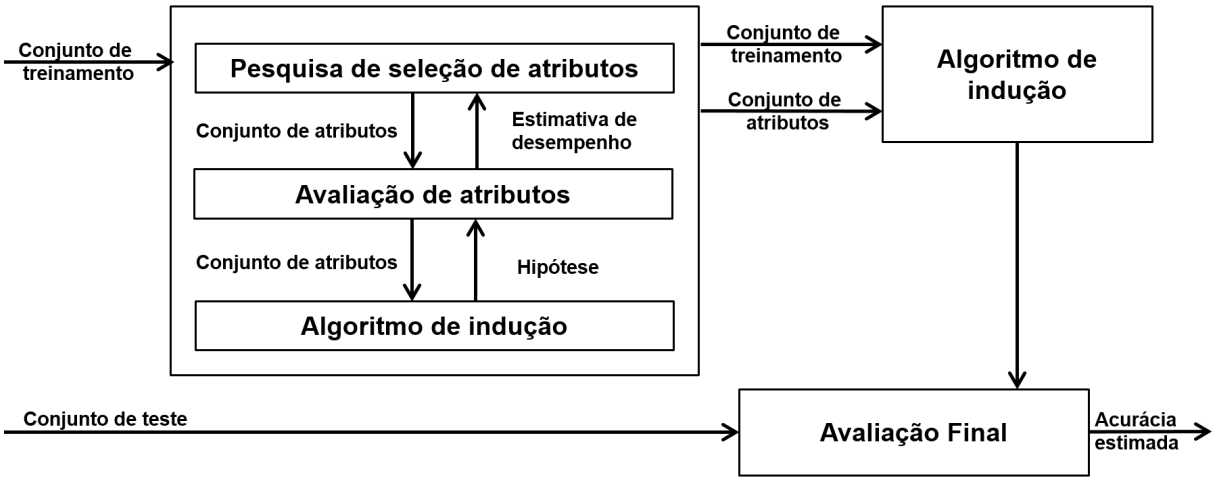


Figura 2.3: Abordagem *wrapper* para a seleção de um subconjunto de atributos [Kohavi and John, 1997] (adaptado)

o espaço de estados e operadores para um problema com quatro atributos. O tamanho do espaço de busca para n atributos é $O(2^n)$, sendo impraticável pesquisar todo o espaço de forma exaustiva, a menos que n seja pequeno.

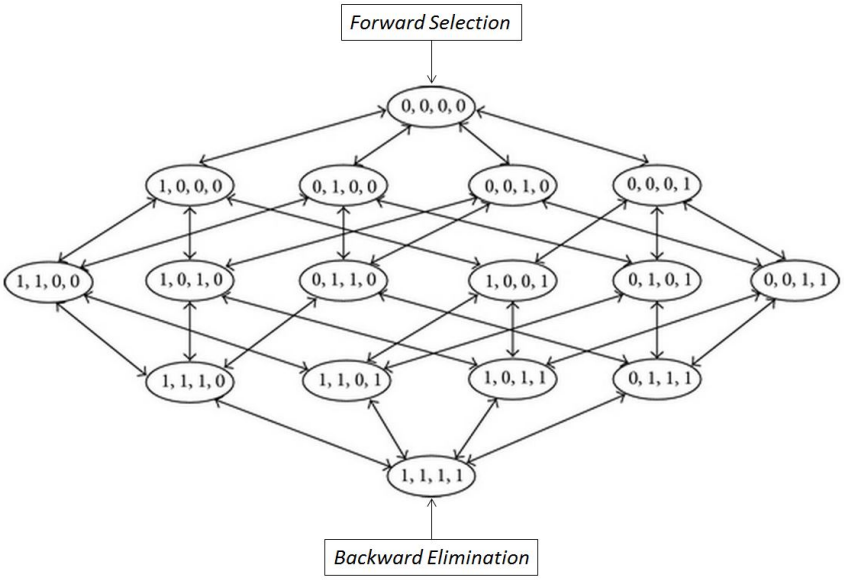


Figura 2.4: Espaço de busca para a seleção de subconjunto de atributos, em um reticulado com quatro atributos [Kohavi and John, 1997]

Existem diferentes estratégias para explorar o reticulado (*lattice*), dado o espaço de busca e o operador de transição de estado [Kohavi and John, 1997]:

- Busca de subida de encosta: a partir do melhor nó atual são consideradas todas as possíveis expansões e são retidos os nós com avaliação máxima;
- Busca em feixe local: semelhante à busca de subida de encosta, mas, somente os melhores k nós são retidos;

- *Best first search*: seleciona os nós mais promissores que foram gerados, mas, ainda não expandidos.

Como descrito anteriormente, o espaço de estados é comumente organizado de tal forma que cada nó representa um subconjunto de atributos e cada operador representa a adição ou supressão de um atributo. O principal problema com esta organização é que a busca deve expandir (i.e. gerar sucessores) cada nó no caminho do subconjunto de atributos inicial para o melhor subconjunto de atributos. O Quadro 2.3 apresenta o resumo da instanciação do problema de busca.

Estado	Vetor booleano, um *bit* por atributo
Estado inicial	O conjunto de atributos vazio (0,0,0 ..., 0)
Heurística	Validação cruzada com fator 5 repetido múltiplas vezes com uma pequena penalidade (0.1%) para cada atributo
Algoritmo de busca	Busca de subida de encosta ou busca em feixe local
Objetivo	Encontrar o estado (i.e. o subconjunto de atributos) que maximiza o escore de avaliação

Quadro 2.3: Instanciação do problema de busca
Fonte: Kohavi e John [Kohavi and John, 1997] (adaptado)

Para percorrer o espaço de busca com o objetivo de selecionar o subconjunto de atributos são utilizados dois algoritmos: *backward elimination* e *forward selection*. O algoritmo *backward elimination* inicia com o conjunto completo de atributos e gulosamente remove os atributos que degradam a acurácia do algoritmo de indução. O algoritmo *forward selection* inicia com o conjunto de atributos vazio e gulosamente adiciona atributos [John et al., 1994].

Para otimizar a busca, Kohavi e John [Kohavi and John, 1997] propõem os operadores compostos (*compound operators*), que são operadores criados dinamicamente após o conjunto padrão dos nós filhos, pois há mais informação na avaliação dos nós filhos do que apenas a identificação do nó com a avaliação máxima. Operadores compostos combinam operadores que levam aos melhores filhos em um único operador dinâmico.

2.1.4 Criação de Atributos

A criação de novos atributos poderia capturar informações importantes em um conjunto de dados de forma mais eficiente do que os atributos originais. A criação de novos atributos a partir de outros existentes, produz, em alguns casos, a definição de novos conceitos. Por exemplo, um corpo em movimento pode ter dois atributos medidos: distância percorrida e quantidade de tempo gasto nesse movimento. A redução desses dois atributos em um único atributo fornece o conceito de velocidade. Em outro exemplo, os atributos peso e altura de diversas pessoas podem ser reduzidos para um único atributo. Essa redução produziu o conceito de Índice de Massa Corpórea - IMC.

A redução de atributos com a produção de novos atributos, em alguns casos, pode ser representada por meio de uma expressão ou equação matemática. Por exemplo, a velocidade (V) é a redução dos atributos distância (D) e tempo (T) gasto por um corpo e é expressa pela relação $V = \frac{D}{T}$. O IMC expressa a relação entre o peso (P) de uma pessoa e sua altura (A) e é calculado pela expressão $IMC = \frac{P}{(A*A)}$.

Neste trabalho investiga-se quais novos indicadores podem ser obtidos de um Sistema de Controle Acadêmico e como esses novos indicadores podem estar relacionados com a evasão escolar.

2.2 Trabalhos Correlatos

Os trabalhos correlatos elencados nessa seção dizem respeito somente às abordagens computacionais para análise do problema da evasão, de forma que outras abordagens para o problema da evasão não são analisadas.

O trabalho de Kotsiantis et al. [Kotsiantis et al., 2003] apresenta uma série de experimentos com dados fornecidos pelos cursos de informática da Hellenic Open University, com o objetivo de identificar o algoritmo de aprendizado mais adequado para a previsão do abandono de curso. A comparação de seis algoritmos de classificação mostrou que o algoritmo Naïve Bayes foi o mais adequado. Os resultados indicam uma acurácia de 63%, baseado somente em dados demográficos, e uma acurácia de 83% antes da metade do período acadêmico.

O trabalho de Dekker et al. [Dekker et al., 2009] apresenta os resultados de um estudo de caso de mineração de dados educacionais com a finalidade de prever a evasão de estudantes do curso de Engenharia Elétrica, da Universidade de Eindhoven, após o primeiro semestre de seus estudos. Os resultados experimentais mostraram que classificadores bastantes simples e intuitivos (e.g. árvores de decisão) dão um resultado útil, com acurácia entre 75 e 80%.

O trabalho de Antunes [Antunes, 2010] exemplifica como diferentes técnicas de mineração de dados (classificação e regras de associação) podem ser combinadas para antecipar o fracasso dos alunos. Os resultados experimentais, aplicados em um conjunto de dados de alunos de graduação do Instituto Superior Técnico de Lisboa, matriculados na disciplina de Fundamentos de Programação, mostram que a precisão desses novos métodos é muito promissora quando comparada com os classificadores treinados com conjuntos de dados menores.

Manhães et al. [Manhães et al., 2012] comparam seis algoritmos de classificação e apresentam uma abordagem quantitativa, aplicados em uma base de dados de informações acadêmicas dos alunos de graduação da UFRJ (Universidade Federal do Rio de Janeiro), com o objetivo de identificar precocemente alunos em risco de evasão. Os melhores resultados foram obtidos com o algoritmo Naïve Bayes, obtendo acurácia de 80%.

O trabalho de Gottardo et al. [Gottardo et al., 2012] aborda técnicas de mineração de dados educacionais utilizadas para geração de inferências sobre o desempenho de estudantes a partir de dados coletados em séries temporais. O objetivo principal foi investigar a viabilidade da obtenção destas informações em etapas iniciais de realização do curso, para apoiar a tomada de ações. Os resultados obtidos demonstraram que é possível obter inferências com acurácia próxima a 75%, utilizando os algoritmos *RandomForest* e *MultilayerPerceptron* (MLP), mesmo em períodos iniciais do curso.

O trabalho de Miranda et al. [Miranda et al., 2014] propõe um modelo de DW, *dashboard* e uso de técnicas de mineração de dados para uma ferramenta analítica aplicada em IESs. O resultado obtido foi um modelo de DW, mineração de dados e ferramenta analítica para essas instituições, com o intuito de melhorar o desempenho e ajudar no processo de tomada de decisão.

Márquez-Vera et al. [Márquez-Vera et al., 2013b] propõem a aplicação de técnicas de mineração de dados para prever insucesso escolar e abandono, em um estudo de caso com dados

de 670 estudantes do ensino médio de Zacatecas, México. Os autores apresentam um método de previsão do insucesso escolar, mostrado na Figura 2.5, composto pelas seguintes etapas:

- Coleta de dados. Consiste em reunir todas as informações disponíveis sobre os alunos. Para fazer isso, o conjunto de fatores que podem afetar o desempenho dos alunos deve ser identificado e recolhido a partir de diferentes fontes de dados disponíveis.
- Pré-processamento. O conjunto de dados é preparado para a aplicação das técnicas de mineração de dados. Para fazer isso, métodos tradicionais de pré-processamento tais como: limpeza de dados, transformação de variáveis e particionamento de dados são aplicados. Outras técnicas como a seleção de atributos e o rebalanceamento dos dados também podem ser aplicadas a fim de resolver os problemas de alta dimensionalidade e dados desbalanceados.
- Mineração de dados. Os algoritmos de DM são aplicados para predizer o fracasso do aluno como um problema de classificação. Para fazer esta tarefa é proposta a utilização de algoritmos de classificação com base em regras e árvores de decisão. Além disso, uma abordagem de classificação sensível ao custo também é utilizada. Finalmente, diferentes algoritmos podem ser executados, avaliados e comparados, a fim de determinar quais obtêm os melhores resultados.
- Interpretação. Os modelos obtidos são analisados para detectar o insucesso do estudante.

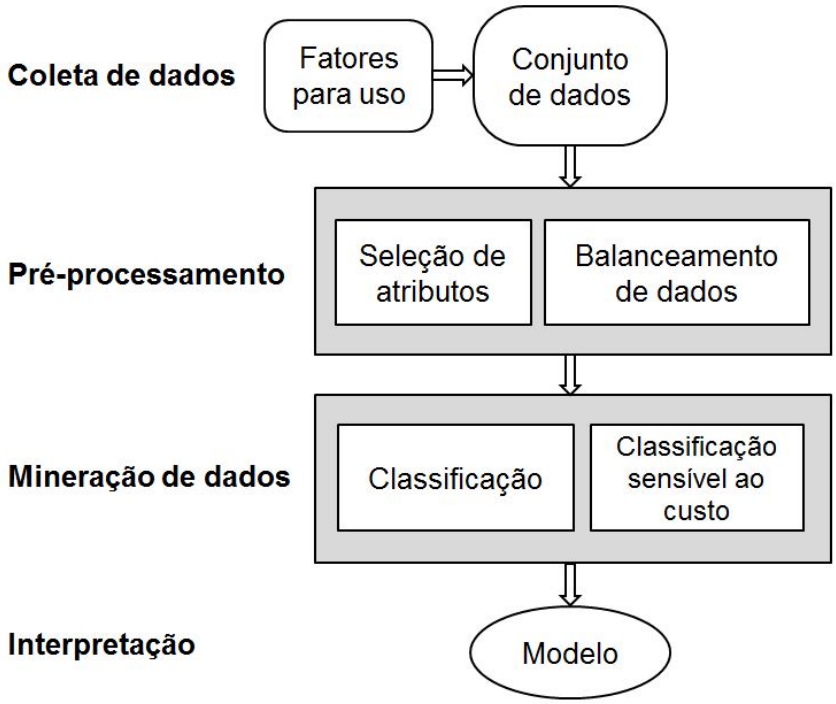


Figura 2.5: Método de previsão do insucesso escolar [Márquez-Vera et al., 2013b] (adaptado)

As acurácias obtidas nos experimentos utilizando os 15 melhores atributos variaram de 88% a 97% nos dez classificadores utilizados. Os autores concluem que algoritmos de classificação podem ser utilizados com sucesso, a fim de prever um desempenho acadêmico do aluno. Os experimentos demonstraram a utilidade das técnicas de seleção de atributos, onde o

conjunto de atributos foi reduzido de 77 para 15 atributos, obtendo-se menos regras e condições, sem perder o desempenho na classificação.

O trabalho de Chau e Phung [Chau and Phung, 2013] apresenta uma abordagem com um esquema de reamostragem híbrida dos dados e aplicação do algoritmo *RandomForest* para a tarefa de classificação de dados educacionais desbalanceados, baseados no desempenho do aluno. Os experimentos comprovaram a eficácia para a previsão da situação final do estudante no curso. A proposta de classificação de dados educacionais desbalanceados, conforme mostrado na Figura 2.6, possui três etapas principais:

- Pré-processamento dos dados: esta etapa é responsável pelo tratamento dos dados faltantes e transformação de dados.
- Reequilíbrio com um esquema híbrido de *oversampling* e *undersampling*: trabalha-se com os dados desbalanceados através de um reequilíbrio nos conjuntos de dados, antes do processo de construção do classificador.
- Processo de construção do classificador com *RandomForest*.

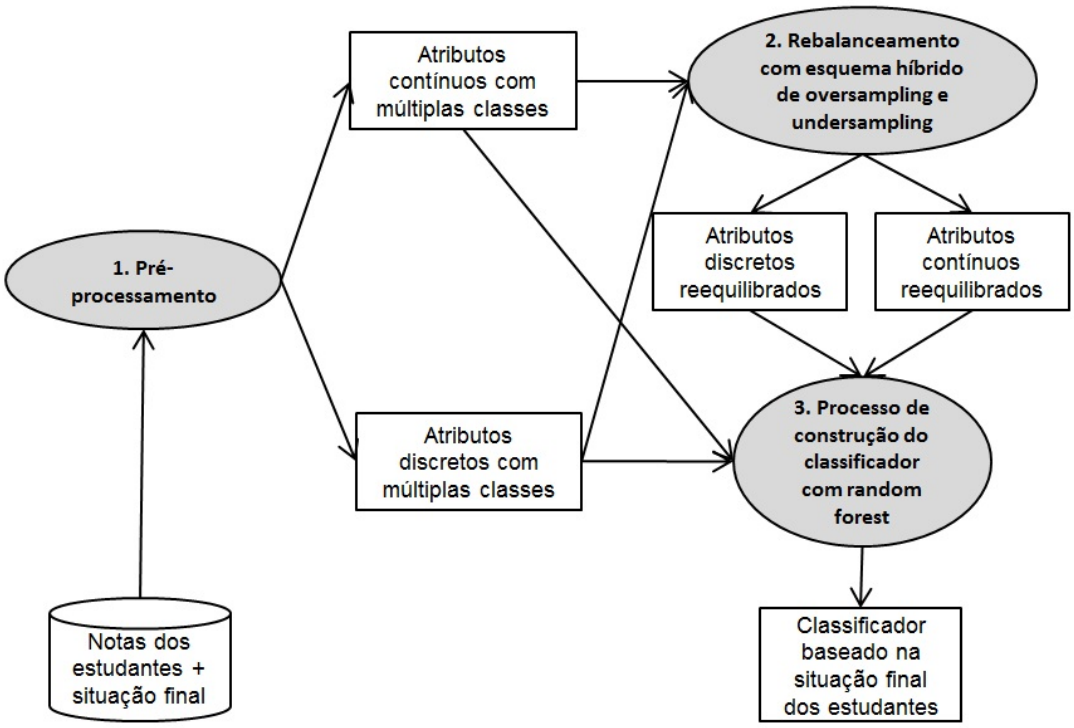


Figura 2.6: Abordagem para classificação de dados educacionais desbalanceados [Chau and Phung, 2013] (adaptado)

Os resultados dos experimentos apresentaram acurácias entre 71% a 94% no conjunto de atributos após o balanceamento das classes. A abordagem proposta tem provado ser eficaz para a previsão do estado final do aluno, sendo utilizada em um sistema de apoio à decisão.

O método de predição da evasão, proposto por Márquez-Vera et al. [Márquez-Vera et al., 2013b], a abordagem para a classificação de dados educacionais desbalanceados, proposto por Chau e Phung [Chau and Phung, 2013], juntamente com o processo de descoberta de conhecimento em bases de dados de Fayyad [Fayyad et al., 1996], são utilizados como referência para o método proposto de seleção dos melhores atributos para classificação.

Capítulo 3

Mineração de Dados Educacionais: Análise da Evasão

Neste capítulo são apresentados o escopo da pesquisa e um estudo preliminar da evasão. Apresenta-se também os atributos extraídos para a tarefa de mineração de dados e a metodologia de pesquisa utilizada neste trabalho. E por fim, é apresentado o método proposto para a seleção dos melhores atributos para a classificação, utilizado para realizar as inferências, com a finalidade de auxiliar os tomadores de decisão no ambiente educacional.

3.1 Definição do Escopo da Pesquisa

O objeto de estudo desta pesquisa para a execução dos experimentos são os dados acadêmicos dos alunos de cursos presenciais graduação da UTFPR. Até 2014 essa instituição não possuía nenhum curso de graduação à distância. Dentre os cursos de graduação, foram selecionados somente os cursos com oferta semestral, que em 2015, 1º semestre, representavam 97% do total de cursos de graduação ofertados, conforme mostrado na Figura 3.1.

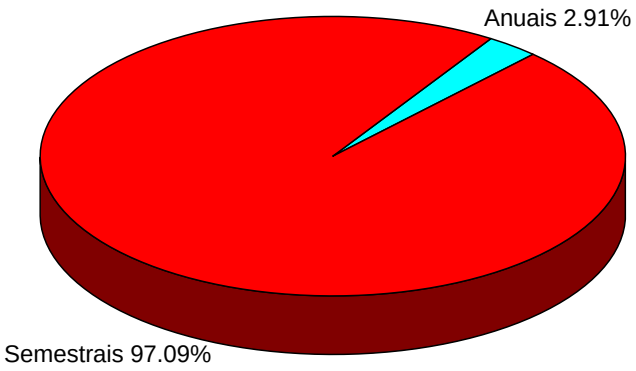
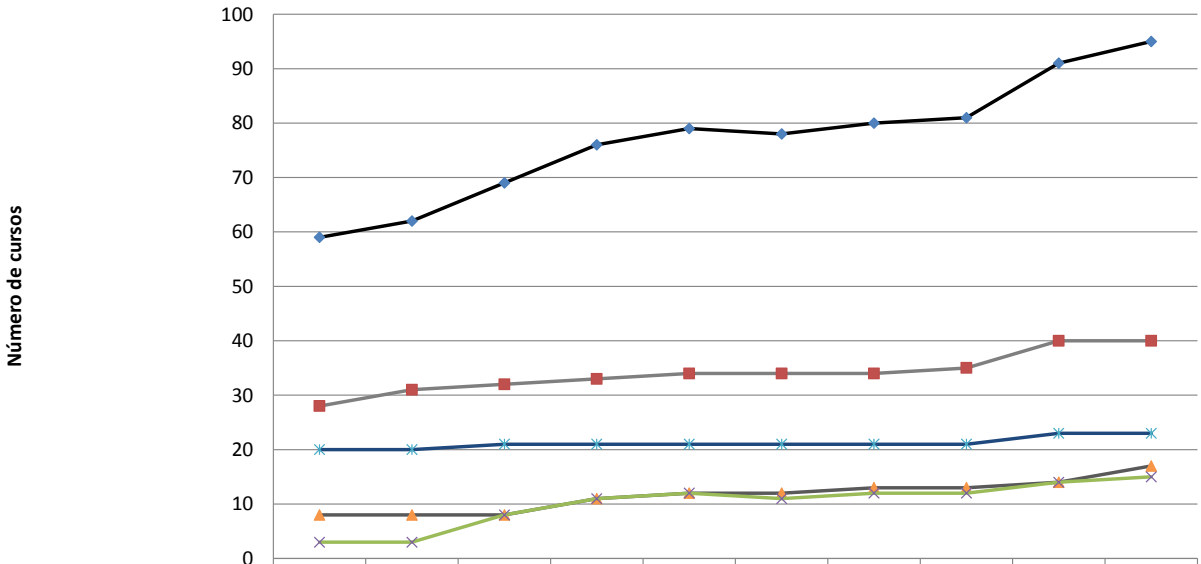


Figura 3.1: Periodicidade dos cursos de graduação ofertados em 2015/1
Fonte: Sistema Acadêmico da UTFPR

A Figura 3.2 apresenta a quantidade de cursos semestrais de graduação da UTFPR com oferta entre os anos de 2010 e 2014, mostrando a crescente oferta de cursos de graduação nesse período.



	2010/1	2010/2	2011/1	2011/2	2012/1	2012/2	2013/1	2013/2	2014/1	2014/2
Total	59	62	69	76	79	78	80	81	91	95
Bacharelado (engenharia)	28	31	32	33	34	34	34	35	40	40
Bacharelado (outros)	8	8	8	11	12	12	13	13	14	17
Licenciatura	3	3	8	11	12	11	12	12	14	15
Tecnologia	20	20	21	21	21	21	21	21	23	23

Figura 3.2: Total de cursos semestrais ofertados entre 2010 e 2014
Fonte: Sistema Acadêmico da UTFPR

3.2 Estudo Preliminar da Evasão

Conforme observado por Mendes Braga et al. [Mendes Braga et al., 2003], a evasão é mais intensa nos períodos iniciais dos cursos. Sendo assim, procurou-se investigar em que período dos cursos a evasão acumulada atingisse 80% (percentual baseado no Princípio de Pareto). Essa informação foi investigada nos cursos semestrais com 6, 8 e 10 períodos, com dados extraídos do Sistema Acadêmico da UTFPR. As Figuras 3.3, 3.4 e 3.5 exibem os diagramas de Pareto dos desistentes por período em cursos de graduação com 6, 8 e 10 períodos.

Os resultados mostram que a desistência até o 3º período ocorreu com 80,28% dos alunos em cursos com 6 períodos, 81,68% dos alunos em cursos com 8 períodos e 81,36% dos alunos em cursos com 10 períodos. Ou seja, independente da duração do curso (6, 8 ou 10 períodos), 80% da evasão acontece até o 3º período do curso.

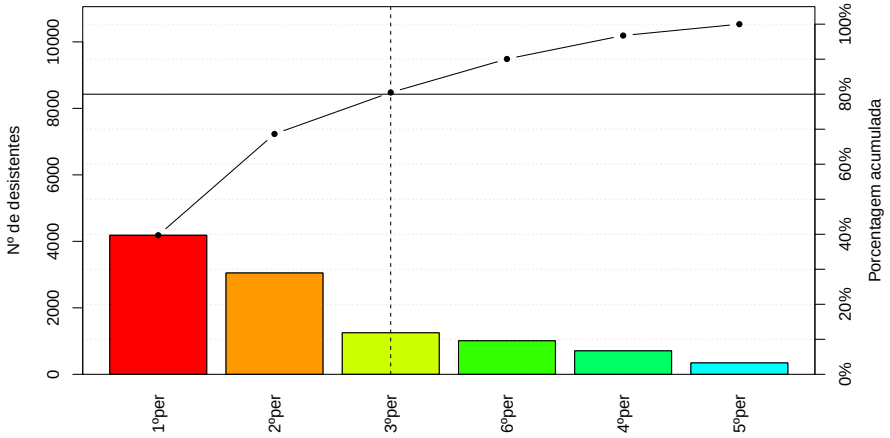


Figura 3.3: Desistentes por período em cursos de graduação com 6 períodos (1981 a 2014)

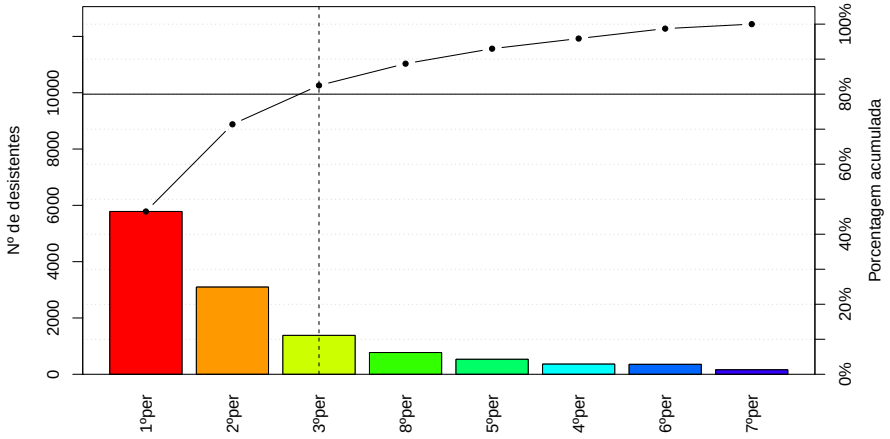


Figura 3.4: Desistentes por período em cursos de graduação com 8 períodos (1999 a 2014)

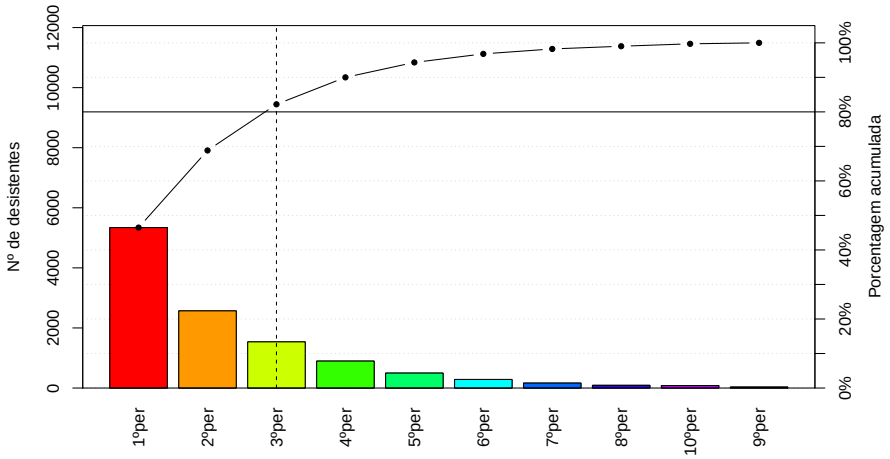


Figura 3.5: Desistentes por período em cursos de graduação com 10 períodos (1978 a 2014)

Para se ter a medida da evasão em uma instituição é necessária a definição da taxa de evasão, definindo-se os tipos de evasão. A Comissão Especial para Estudo da Evasão [SESu/MEC, 1996] caracterizou a evasão da seguinte forma:

- evasão de curso: quando o estudante desliga-se do curso superior em situações diversas tais como: abandono (deixa de matricular-se), desistência (oficial), transferência ou reopção (mudança de curso), exclusão por norma institucional;
- evasão da instituição: quando o estudante desliga-se da instituição na qual está matriculado;
- evasão do sistema: quanto o estudante abandona de forma definitiva ou temporária o ensino superior.

Na UTFPR, a Comissão de Evasão e Retenção caracterizou a evasão da seguinte forma:

- evasão de curso: quando o estudante desliga-se do curso superior;
- evasão do câmpus: quando o estudante desliga-se do câmpus em que está matriculado;
- evasão da instituição: quando o estudante desliga-se da instituição na qual está matriculado.

Essa comissão da UTFPR definiu, seguindo as bases conceituais sobre evasão e retenção discutidas nos seminários promovidos pela SETEC - Secretaria de Educação Profissional e Tecnológica, do MEC, e pela ANDIFES - Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior, a seguinte equação para o cálculo da evasão semestral:

$$E_i^n = A_i^n - C_i^n - (A_{i+1}^n - I_{i+1}^n) \tag{3.1}$$

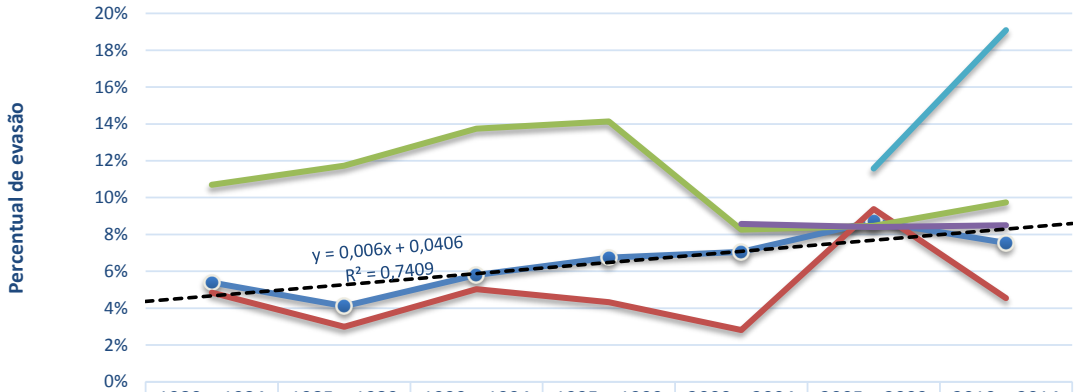
Em que:

$i \rightarrow$ semestre em análise;
$n \rightarrow$ ano em análise;
$E_i^n \rightarrow$ número de alunos evadidos no semestre i do ano n;
$A_i^n \rightarrow$ número de alunos com matrícula ativa;
$C_i^n \rightarrow$ número de alunos concluintes;
$I_i^n \rightarrow$ número de alunos ingressantes.

Assim, a taxa de evasão semestral é expressa por:

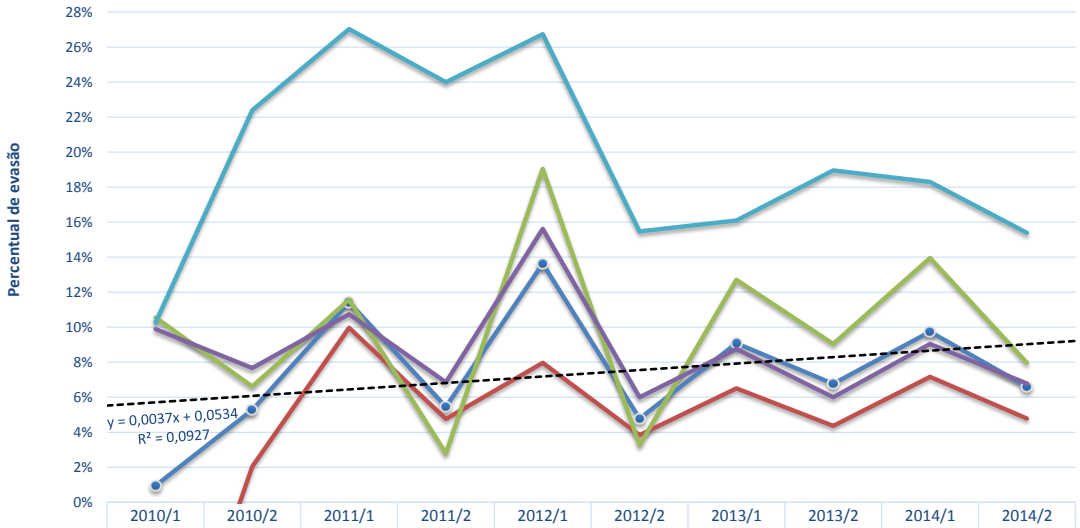
$$\%E_i^n = \frac{E_i^n}{A_i^n} \times 100 \tag{3.2}$$

Com a aplicação da equação da taxa de evasão semestral nos dados históricos da UTFPR, retirando-se os alunos que mudaram de curso, permitiu-se gerar os gráficos da taxa de evasão por quinquênio (de 1980 a 2014), conforme mostrado na Figura 3.6, e da taxa de evasão semestral de 2010 a 2014, conforme indicado na Figura 3.7. Nessas duas figuras a linha pontilhada representa a linha de tendência, utilizando a regressão linear com os valores de evasão de todos os cursos. Interpretando o coeficiente de correlação (R^2) das duas figuras pode-se observar a taxa de evasão semestral por quinquênio apresenta uma correlação forte e a taxa de evasão semestral de 2010 a 2014 possui uma correlação bem fraca.



	1980 a 1984	1985 a 1989	1990 a 1994	1995 a 1999	2000 a 2004	2005 a 2009	2010 a 2014
Todos	5,40%	4,11%	5,79%	6,74%	7,06%	8,71%	7,54%
Engenharias	4,84%	2,99%	5,04%	4,33%	2,81%	9,38%	4,55%
Tecnologias	10,70%	11,74%	13,75%	14,13%	8,26%	8,47%	9,74%
Demais Bacharelados					8,57%	8,40%	8,51%
Licenciaturas						11,58%	19,09%

Figura 3.6: Taxa de evasão semestral por quinquênio, de 1980 a 2014



	2010/1	2010/2	2011/1	2011/2	2012/1	2012/2	2013/1	2013/2	2014/1	2014/2
Todos	0,95%	5,27%	11,43%	5,45%	13,63%	4,78%	9,08%	6,77%	9,74%	6,60%
Engenharias	-14,01%	2,04%	9,97%	4,76%	7,97%	3,84%	6,51%	4,37%	7,17%	4,79%
Tecnologias	10,55%	6,61%	11,57%	2,79%	19,05%	3,26%	12,71%	9,05%	13,96%	7,97%
Demais Bacharelados	9,89%	7,67%	10,74%	6,86%	15,62%	6,00%	8,76%	6,00%	9,05%	6,78%
Licenciaturas	10,26%	22,39%	27,03%	24,00%	26,74%	15,47%	16,10%	18,96%	18,30%	15,39%

Figura 3.7: Taxa de evasão semestral de 2010 a 2014

3.3 Modelagem do *Data Warehouse*

O DW é uma ferramenta que auxilia enormemente o pré-processamento para a mineração de dados. O DW proposto neste trabalho é um *Virtual Warehouse*, isto é, um conjunto de visões sobre bancos de dados operacionais onde apenas algumas das possíveis visões são materializadas [Han et al., 2011]. Foi escolhida esta abordagem pois permite uma flexibilidade na

criação do DW. As tabelas fato utilizadas no DW utilizam como fonte de dados, além de tabelas transacionais, duas séries históricas que foram inferidas com os dados de alunos da UTFPR.

3.3.1 Arquitetura do *Data Warehouse*

Nos trabalhos preliminares da comissão de Evasão e Retenção da UTFPR (ver Seção 1.2) demandou-se a criação de um *Data Warehouse*, para permitir uma melhor análise dos dados acadêmicos da instituição e auxiliar no diagnóstico do problema.

A arquitetura do DW está baseada em três camadas: dados, negócio e apresentação, e o seu detalhamento encontra-se em Oliveira Júnior et al. [Oliveira Júnior et al., 2015].

3.3.2 Séries Históricas

Uma série histórica, também denominada série temporal, é uma sequência de dados obtidos em intervalos regulares de tempo durante um período específico. Um DW torna-se efetivo na avaliação de dados com granularidade temporal. Sendo assim, para permitir a análise de dados de anos anteriores foram criadas duas séries históricas.

A primeira série histórica representa o vínculo do aluno com o curso em cada período letivo. Essa série histórica contém os principais dados acadêmicos do aluno em cada período letivo vinculado ao curso, como: situação do aluno (regular, trancado, etc.), coeficiente de rendimento, número de disciplinas matriculadas, número de disciplinas reprovadas, entre outros atributos. Foi utilizada a granularidade do período letivo (semestral, anual) por ser uma medida que representa a evolução do aluno na matriz curricular do curso. A série histórica foi inferida entre os anos de 1979 e 2014, baseada no histórico escolar dos alunos, contendo dados de 105.117 discentes de cursos de graduação e técnico, em um total de 839.223 tuplas.

A segunda séria histórica criada representa as disciplinas/turma cursadas pelos alunos (foram excluídas as disciplinas de crédito consignado) em cada período letivo, contendo informações como: média e desvio padrão das notas, número de alunos matriculados, número de alunos reprovados por nota, número de alunos reprovados por frequência, carga horária semanal da disciplina, entre outros atributos. A série histórica foi inferida entre os anos de 1979 e 2014, baseada no histórico escolar dos alunos, contendo 430.857 tuplas de 13.110 disciplinas.

3.4 Extração de Atributos para a Mineração de Dados

A extração dos dados compreendeu uma pesquisa (entre as diversas fontes existentes na organização) sobre a disponibilidade, precisão, validade e consistência dos dados. Para essa etapa, a implementação do DW foi um grande facilitador, permitindo automatizar as atividades de extração, transformação e carga dos dados, além de melhorar a qualidade e confiabilidade dos mesmos, onde foi necessário a retificação de informações.

Os dados utilizados neste trabalho, conforme autorização da Pró-Reitoria de Graduação e Educação Profissional - PROGRAD, tiveram o foco nos resultados dos cursos, não permitindo a identificação de qualquer aluno, mantendo a confidencialidade bem como a privacidade de seus conteúdos.

Em função da indisponibilidade de obtenção de alguns atributos (e.g. dados de auxílio estudantil), entende-se que o conjunto de atributos final formado não é o ideal, mas, já permite realizar muitas inferências.

Para a tarefa de mineração de dados foram utilizados os atributos descritos no Quadro 3.1, extraídos de dados acadêmicos dos alunos da UTFPR, dados do ENEM[1] e dados de um questionário socioeconômico aplicado aos alunos no início do curso.

A definição de classes investigativas do problema indica se um aluno se evadiu (sim ou não). Os alunos que não se evadiram são representados pelos alunos ativos: regulares, trancados, em mobilidade acadêmica, sub judice ou alunos já formados. Os alunos que evadiram do curso são representados pelos alunos: desistentes, que mudaram ou transferiram de curso, falecidos ou jubilados.

As seções a seguir apresentam o detalhamento dos atributos já existentes na base de dados de registros acadêmicos e os atributos criados.

3.4.1 Atributos já Existentes

A relação abaixo apresenta os atributos já existentes na base de dados, baseados em informações já utilizadas pela instituição. A numeração dos atributos indicados na relação abaixo segue a codificação utilizada no Quadro 3.1.

Atributo nº 01 - Grau: indica qual o grau que o curso confere ao aluno, que são: bacharelados (engenharia), demais bacharelados, tecnologia ou licenciatura.

Atributo nº 02 - Gênero: indica do gênero do aluno, masculino ou feminino, constante no seu registro civil.

Atributo nº 03 - Estado civil: indica o estado civil do aluno, que pode ser: solteiro, divorciado, casado, separado judicialmente ou viúvo.

Atributo nº 04 - Tipo de escola anterior: indica se o aluno é oriundo de escola pública ou privada.

Atributo nº 07 - Tipo de cota: indica em qual política de cotas o aluno foi selecionado no processo seletivo. Os tipos de cota utilizados entre 2008 e 2012 são: não cotista e cotista (escola pública). A partir de 2013[2] são utilizados os seguintes tipos de cota:
- não cotista,
- categoria I (oriundo de família com renda igual ou inferior a 1,5 salários-mínimos per capita e se enquadra no grupo PPI - Pretos, Pardos e Indígenas),
- categoria II (oriundo de família com renda igual ou inferior a 1,5 salários-mínimos per capita e não se enquadra no grupo PPI),
- categoria III (oriundo de família independente de renda e PPI) e
- categoria IV (oriundo de família independente de renda e não se enquadra no grupo PPI)

Atributo nº 09 - Idade do aluno: indica a idade do aluno ao ingressar no curso.

Atributo nº 14 - Coeficiente de rendimento: é o índice de rendimento acadêmico, e leva em consideração tanto a média final quanto a carga horária da disciplina. O coeficiente de

[1] Exame Nacional do Ensino Médio, utilizado como forma de ingresso para os cursos de graduação na UTFPR desde 2010, através do SISU (Sistema de Seleção Unificada)

[2] Ano em que a UTFPR passou a adotar integralmente a Lei de Cotas - Lei 12.711/2012

Nº	Atributo	Tipo	Atributo criado
01	grau (engenharia, bacharelado, tecnologia ou licenciatura)	Categórico	
02	genero (masculino ou feminino)	Categórico	
03	estado_civil	Categórico	
04	tipo_escola_anterior (pública ou privada)	Categórico	
05	reentrada_mesmo_curso (sim/não)	Categórico	Sim
06	mudou_de_curso (sim/não)	Categórico	Sim
07	tipo_cota	Categórico	
08	previsao_evasao_dificuldade (sim/não)	Categórico	Sim
09	idade_inicio_curso	Numérico	
10	total_semestres_trancados	Numérico	Sim
11	emprestimos_biblioteca_por_semestre	Numérico	Sim
12	regressao_coeficiente	Numérico	Sim
13	percentual_frequencia	Numérico	Sim
14	coeficiente_rendimento	Numérico	
15	percentual_aprov	Numérico	Sim
16	nota_final_enem	Numérico	
17	nota_linguagem	Numérico	
18	nota_humanas	Numérico	
19	nota_natureza	Numérico	
20	nota_matematica	Numérico	
21	nota_redacao	Numérico	
22	micro_regiao_origem (mesma do câmpus ou outra)	Categórico	Sim
23	meso_regiao_origem (mesma do câmpus ou outra)	Categórico	Sim
24	regiao_origem (mesma do câmpus ou outra)	Categórico	Sim
25	socio_renda_familiar	Categórico	
26	socio_mora_com	Categórico	
27	socio_reside_em	Categórico	
28	socio_trabalho	Categórico	
29	socio_necessidade_trabalhar	Categórico	
30	socio_part_economica_na_familia	Categórico	
31	socio_escolaridade_pai	Categórico	
32	socio_escolaridade_mae	Categórico	
33	socio_nr_membros_familia	Categórico	
34	socio_fez_cursinho	Categórico	
35	socio_motivo_escolha_curso	Categórico	
36	evasao (sim/não) [atributo alvo]	Categórico	

Quadro 3.1: Atributos utilizados nos experimentos

rendimento é calculado pela seguinte equação[3]:

$$CR = \frac{\sum(NF \times CH)}{10 \times \sum CH} \tag{3.3}$$

Em que:
CR = coeficiente de rendimento do aluno;
NF = nota final na disciplina/unidade curricular, expressa de 0,0 a 10,0;
CH = carga horária total da disciplina/unidade curricular.

Atributo nº 16 a 21 - Notas no ENEM: refere-se às notas obtidas no ENEM, no ano anterior ao ingresso do aluno no curso.

Atributo nº 26 a 35 - Dados socioeconômicos: são atributos extraídos de um questionário (*on-line*) que os alunos preenchem no momento do ingresso no curso.

Atributo nº 36 - Evasão: é o atributo alvo da classificação. Indica se o aluno se evadiu (sim ou não).

3.4.2 Criação de Atributos

Considerando informações existentes na base de dados de alunos da UTFPR e utilizando medidas estatísticas para a sua definição, este trabalho propõe a criação dos seguintes atributos, com o objetivo de auxiliar no processo de mineração de dados.

Dificuldade Média das Disciplinas Cursadas pelo Aluno

O primeiro atributo criado utiliza a dificuldade de uma disciplina/turma cursada pelos alunos, definido pela relação inversa do percentual de aprovação dos alunos na disciplina/turma. A equação da dificuldade da disciplina/turma cursada pelos alunos em um período letivo é representada por:

$$Dif(d) = \log_2 \left(\frac{Ap(d) + Rep(d)}{Ap(d)} \right) \tag{3.4}$$

Em que:
$Dif(d) \rightarrow$ dificuldade de aprovação da disciplina/turma (d) em um período letivo;
$Ap(d) \rightarrow$ número de alunos aprovados na disciplina/turma;
$Rep(d) \rightarrow$ número de alunos reprovados na disciplina/turma.
Observação: Foi utilizado o $\log_2$ com a finalidade de atenuar e normalizar os valores.

A partir daí é possível computar o atributo denominado de "dificuldade média das disciplinas cursadas pelo aluno". Esse atributo agrega um componente coletivo (percentual dos alunos aprovados na disciplina/turma) no desempenho individual do aluno. A ideia é criar um atributo que possa auxiliar na previsão da evasão escolar, além de contribuir com os demais

[3]http://www.utfpr.edu.br/estrutura-universitaria/pro-reitorias/prograd/legislacao/utfpr-1/bacharelado/034-15-regulamento-da-organizacao-didatico-pedagogica-correcao-04-2015

atributos na tarefa de mineração de dados. O cálculo desse atributo é feito com a seguinte equação:

$$DM(a) = \frac{\sum_{i=1}^{n} Dif(D_n) - \sum_{j=1}^{m} Dif(D_m)}{n+m} \tag{3.5}$$

Em que:

$DM(a) \rightarrow$ dificuldade média das disciplinas cursadas pelo aluno (a);
$n \rightarrow$ total de disciplinas que o aluno obteve aprovação;
$m \rightarrow$ total de disciplinas que o aluno reprovou;
$D_n \rightarrow$ disciplina que o aluno obteve aprovação;
$D_m \rightarrow$ disciplina que o aluno reprovou.

Na investigação de quais seriam os valores aceitáveis de dificuldade média das disciplinas cursadas pelos alunos, foi aplicado o cálculo em uma amostra composta de 16.766 alunos de cursos semestrais de graduação formados na UTFPR entre os anos de 1983 e 2014, que ingressaram no curso no 1º período. Foram selecionados os alunos formados por serem a referência de desempenho acadêmico de sucesso. Foram selecionadas na amostra as disciplinas da matriz curricular do curso, composta de disciplinas obrigatórias, optativas e equivalentes, sendo excluídas as disciplinas de enriquecimento curricular, crédito consignado, estágio curricular obrigatório e atividades complementares.

As Figuras C.1, C.2 e C.3, do Apêndice C, mostram respectivamente o histograma, a densidade e o *boxplot* da média das disciplinas cursadas até o 3º período pelos alunos formados.

Segue abaixo o resumo dos valores estatísticos do atributo de dificuldade média de disciplinas cursadas pelos alunos formados (até o 3º período):

Mínimo	-0,7500	Máximo	1,5500	Mediana	0,2200
1º Quartil	0,0800	3º Quartil	0,3600	Média	0,2220

Para a amostra selecionada o valor da variância interquartil foi de 0,28. Desta forma, o intervalo para exclusão dos valores discrepantes, utilizando $1,5 \times IQR$, são os valores situados fora do intervalo [-0.37 .. 0.80], resultando em 2,73% de valores discrepantes na amostra selecionada. Ou seja, os alunos que neste atributo estiverem fora desse intervalo poderão ser considerados "em risco de evasão".

Regressão Linear do Coeficiente de Rendimento

O objetivo desse atributo é verificar qual a tendência (aumento, diminuição ou estabilidade) do coeficiente de rendimento para o próximo semestre a ser cursado pelo aluno. Para isso, é calculado o coeficiente de rendimento das disciplinas cursadas pelo aluno em cada semestre letivo. Com os coeficientes de cada semestre é calculado o coeficiente de regressão (β_1) da equação de regressão linear simples, conforme a expressão:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{3.6}$$

Em que:

$Y_i \rightarrow$ valor observado para a variável dependente Y no i-ésimo nível da variável independente X.

$\beta_0 \rightarrow$ constante de regressão. Representa o intercepto da reta com o eixo Y.

$\beta_1 \rightarrow$ coeficiente de regressão. Representa a variação de Y em função da variação de uma unidade da variável X.

$X_i \rightarrow$ i-ésimo nível da variável independente X $(i = 1,2,...,n)$.

$\varepsilon_i \rightarrow$ é o erro que está associado à distância entre o valor observado Y_i e o correspondente ponto na curva, do modelo proposto, para o mesmo nível i de X.

Valores positivos de β_1 indicam a tendência de aumento do coeficiente de rendimento no próximo semestre e valores negativos indicam a tendência de diminuição do coeficiente de rendimento no semestre seguinte. A ideia desse atributo surgiu nas reuniões da Comissão de Evasão e Retenção da UTFPR.

A Figura 3.8 exemplifica a evolução do coeficiente de rendimento de um aluno, em cada semestre, utilizando a regressão linear simples. A reta em azul representa a linha de tendência do coeficiente de rendimento, chamado de coeficiente de rendimento parcial.

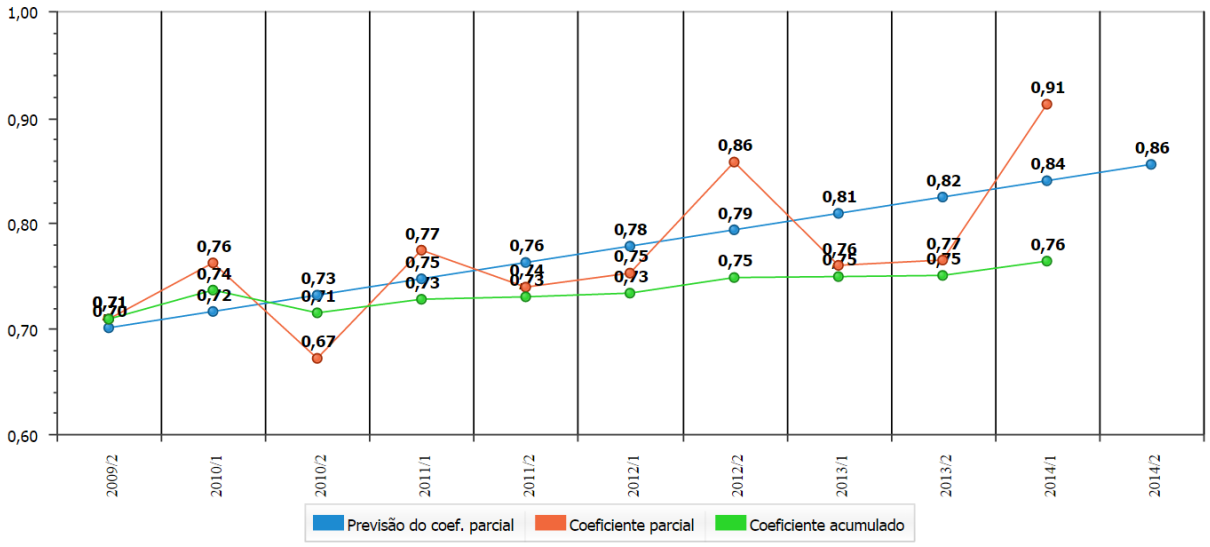


Figura 3.8: Evolução do coeficiente de rendimento de um aluno

Enriquecimento de Dados

Os seguintes atributos foram incorporados à base, sendo obtidos de forma simples, a partir de informações disponíveis na base de dados. São eles:

- Reentrada no mesmo curso: indica se o aluno está reiniciando o mesmo curso, no mesmo câmpus. Isso acontece principalmente no 2º semestre, quando os alunos utilizam a mesma nota do ENEM para reingressar no mesmo curso, tendo como benefício a melhoria do coeficiente de rendimento, pois são convalidadas apenas as disciplinas aprovadas.

- Mudança de curso: indica se o aluno é oriundo de outro curso de graduação da instituição.

- Total de semestres trancados: indica a quantidade de semestres em que o aluno esteve com a matrícula trancada.
- Empréstimos na biblioteca: indica a média de empréstimos de livros na biblioteca por semestre cursado. Essa informação foi extraída do Sistema Integrado de Bibliotecas Pergamum[4], utilizado pela UTFPR desde 2002.
- Percentual de frequência: indica o percentual de frequência das disciplinas cursadas. Essa informação foi extraída dos diários de classe.
- Percentual de aprovação: indica o percentual de aprovação das disciplinas cursadas. Essa informação foi extraída do histórico escolar dos alunos.
- Microrregião, mesorregião e região de origem dos calouros: estes três atributos indicam se o aluno é oriundo da mesma microrregião, mesorregião ou região (IBGE - Instituto Brasileiro de Geografia e Estatística) do câmpus. Essa informação é importante para verificar alguma influência que possa ocorrer com alunos oriundos de outras localidades diferentes do município do câmpus.

3.5 Metodologia

Este trabalho consiste de uma pesquisa exploratória de natureza aplicada [Gerhardt and Silveira, 2009], pois objetiva estudar métodos para a identificação de padrões para auxiliar a tomada de decisão de gestores educacionais com a finalidade de analisar a evasão de estudantes em cursos presenciais de graduação.

Como método científico adotou-se o método dedutivo [Gerhardt and Silveira, 2009], uma vez que com base em um conhecimento técnico e científico já formalmente conhecido é possível desenvolver e avaliar uma solução computacional que ofereça suporte de maneira consistente com tais conhecimentos (e/ou premissas).

Trata-se de uma pesquisa quantitativa, uma vez que a abordagem adotada para análise do método proposto ocorrerá por meio dos resultados mensuráveis obtidos com os experimentos realizados.

Em relação aos procedimentos técnicos, foram realizados levantamentos bibliográficos que fundamentaram o desenvolvimento do método proposto de seleção dos melhores atributos para classificação. Os experimentos foram realizados subsequentemente neste método para prova de conceito e análise do método propriamente dito.

3.5.1 Método para Seleção dos Melhores Atributos

Como visto na Seção 2.1.3, a redução da dimensionalidade melhora o desempenho dos algoritmos de classificação. Sendo assim, este trabalho propõe um método de seleção dos melhores atributos a serem utilizados na mineração de dados, indicado no diagrama da Figura 3.9. Esse método baseia-se nos trabalhos de Márquez-Vera et al. [Márquez-Vera et al., 2013b], Chau e Phung [Chau and Phung, 2013], juntamente com o processo de descoberta de conhecimento em bases de dados de Fayyad [Fayyad et al., 1996].

[4] http://www.pergamum.pucpr.br/redepergamum/pergamum_index.php

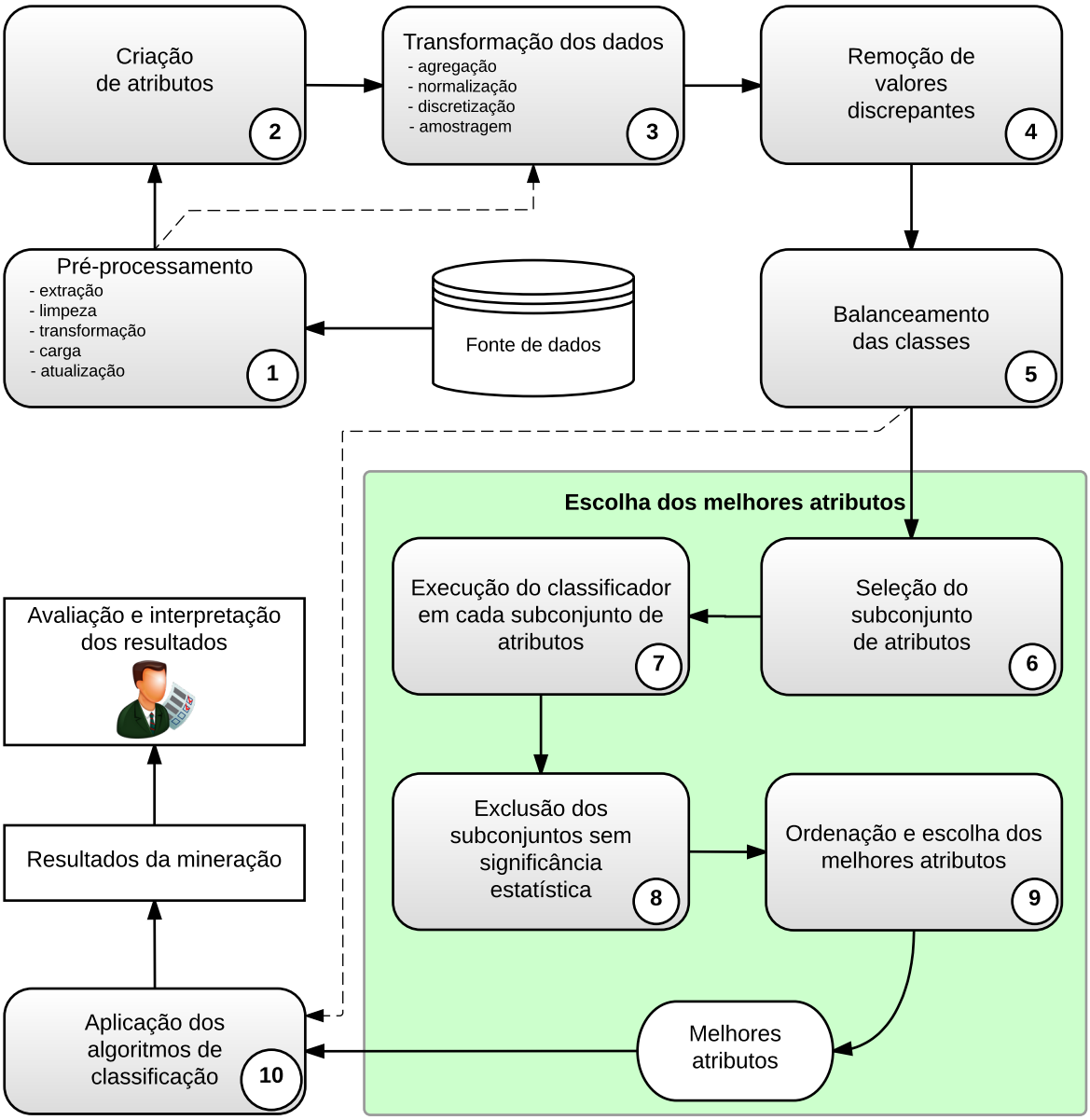


Figura 3.9: Método de seleção dos melhores atributos para classificação

A revisão bibliográfica do Capítulo 2 serviu de embasamento teórico para o método proposto, referenciando o detalhamento das técnicas e algoritmos citados. O método é composto de dez etapas que são indicadas a seguir. Os números mostrados na Figura 3.9 referem-se aos indicados, entre parênteses, nas subseções.

(1) Pré-processamento

Nesta etapa são realizadas as atividades de extração, limpeza, transformação, carga e atualização dos dados, conforme os procedimentos tradicionais empregados em mineração de dados [Fayyad et al., 1996]. Para a extração dos atributos sugere-se criar pelo menos três *da-*

tasets, contendo amostras distintas, para permitir um resultado menos suscetível ao sobreajuste (*overfitting*).

(2) Criação de Atributos

A criação de novos atributos pode capturar informações importantes em um conjunto de forma mais eficiente do que os atributos originais. Este trabalho propõe a criação de novos atributos, detalhados na Seção 3.4.2, considerando informações existentes na base de dados e utilizando medidas estatísticas para a sua definição. O objetivo dos novos atributos é criar índices quantitativos que sejam simples e fáceis de interpretar, e que sirvam como "sinais de alerta" para os gestores educacionais, permitindo a tomada de ações a tempo de evitar a evasão. Esta etapa é facultativa, conforme indicado na Figura 3.9 com a linha tracejada entre a etapa 1 e a etapa 3, apesar de altamente recomendada.

(3) Transformação dos Dados

Nesta etapa são realizadas as tarefas de agregação, normalização, discretização e amostragem dos dados, também seguindo os procedimentos tradicionais empregados em mineração de dados, como descrito em Han et al. [Han et al., 2011].

(4) Remoção dos Valores Discrepantes

Nesta etapa é verificada a necessidade de remoção de valores discrepantes (*outliers*). Como visto na Seção 2.1, pode-se utilizar nesta etapa o cálculo da amplitude interquartil. Os limites superiores e inferiores são calculados e os valores fora destes limites são considerados valores discrepantes.

(5) Balanceamento das Classes

Apesar da evasão ser um problema nas instituições de ensino, o número de casos de evasão ainda é, em geral, menor em relação ao número de alunos não evadidos. Sendo assim, o problema se caracteriza pelo desbalanceamento das classes. Este problema faz com que os algoritmos de aprendizagem tendam a ignorar as classes menos frequentes (classes minoritárias) e só considerar nas mais frequentes (classes majoritárias). Como resultado, o classificador não é capaz de classificar corretamente as instâncias de dados correspondentes a classes pouco representadas [Márquez-Vera et al., 2013b].

Uma abordagem amplamente utilizada no balanceamento das classes é a aplicação do algoritmo SMOTE (*Synthetic Minority Oversampling Technique*) [Chawla et al., 2002]. Esse algoritmo, empregado neste trabalho, ajusta a frequência relativa entre classes majoritárias e minoritárias, introduzindo sinteticamente instâncias de classes minoritárias, considerando a técnica de agrupamento K-nn [Witten et al., 2011].

(6) Seleção do Subconjunto de Atributos

A seleção do subconjunto de atributos é um método de redução da dimensionalidade quando são detectados e removidos atributos irrelevantes, fracamente relevantes ou redundantes, conforme detalhado na Seção 2.1.3. Nesta etapa são aplicados algoritmos com a abordagem

filter (que independe do classificador a ser utilizado) e algoritmos com a abordagem *wrapper*, que utilizam um classificador na seleção dos atributos.

O Quadro 3.2 apresenta uma sugestão de algoritmos de seleção de atributos a serem utilizados nessa etapa, todos disponíveis na ferramenta WEKA (*Waikato Environment for Knowledge Analysis*). É importante salientar que o algoritmo com a abordagem *wrapper* requer o uso de um classificador para realizar a seleção de atributos. Os algoritmos constantes no Quadro 3.2 estão descritos em detalhes em Witten et al. [Witten et al., 2011].

Abordagem	**Algoritmo**	**Método de busca**
Filter	CfsSubsetEval	BestFirst e GeneticSearch
	ChiSquaredAttributeEval	Ranking
	GainRatioAttributeEval	Ranking
	InfoGainAttributeEval	Ranking
	OneRAttributeEval	Ranking
	ReliefFAttributeEval	Ranking
	SymmetricalUncertAttributeEval	Ranking
Wrapper	WrapperSubsetEval	BestFirst e GeneticSearch

Quadro 3.2: Sugestão de algoritmos de seleção de atributos
Fonte: Witten et al. [Witten et al., 2011] (adaptado)

(7) Execução do Classificador em Cada Subconjunto de Atributos

Depois de selecionados os subconjuntos de atributos, os classificadores devem ser avaliados quanto ao desempenho, utilizando-se como medida a acurácia. Nos experimentos realizados, os algoritmos foram executados dez vezes nos subconjuntos selecionados, usando a técnica de validação cruzada (fator $n = 10$).

Para esta etapa foi utilizado nos experimentos o ambiente WEKA Experiment Environment - WEE [Hall et al., 2009], aplicando-se o meta classificador FilteredClassifier. Nesse meta classificador foi aplicado o filtro "attribute.Remove" (para a seleção do subconjunto de atributos) e posteriormente foi aplicado o classificador. Para os subconjuntos selecionados pelos algoritmos com a abordagem *filter* foram selecionados apenas os dez melhores atributos ranqueados.

(8) Exclusão dos Subconjuntos sem Significância Estatística

O objetivo dessa etapa é descartar o subconjunto de atributos cuja acurácia seja muito inferior à melhor acurácia obtida com o classificador no experimento.

Para avaliar a significância estatística dos resultados obtidos, utiliza-se a técnica de teste estatístico conhecida como "T-pareado" (*pairwise T-test*) [Witten et al., 2011], com nível de significância de 5%.

A partir do resultado do teste "T-pareado", considerando o nível de significância de 5%, são desprezados os subconjuntos de atributos selecionados em que a acurácia não obteve significância estatística, quando comparados com a melhor acurácia obtida (denominada *test*

base no WEE). Caso todos os subconjuntos selecionados pela abordagem *filter* não obtenham significância estatística, deve-se selecionar o subconjunto com a melhor acurácia, para assim permitir realizar o desempate na próxima etapa.

(9) Ordenação e Escolha dos Melhores Atributos

Para se obter os melhores atributos, após o descarte dos subconjuntos sem significância estatística, este trabalho propõe o seguinte procedimento:

1. Ordena-se de forma decrescente a frequência em que o atributo foi selecionado pelos algoritmos WrapperSubsetEval e CfsSubsetEval, que não utilizam o método de busca *Ranking*;

2. Ordena-se de forma crescente pela posição média em que o atributo foi classificado pelos algoritmos que utilizaram o método de busca Ranking.

3. Selecionam-se os n melhores atributos. Sugere-se o valor de $n = 10$.

(10) Aplicação dos Algoritmos de Classificação

Nesta etapa são aplicados os algoritmos de classificação no *dataset* com os melhores atributos selecionados. Caso os melhores atributos já estejam selecionados e haja a necessidade de criação de outro *dataset*, utiliza-se o fluxo alternativo indicado na Figura 3.9 com a linha tracejada entre a etapa 5 e a etapa 10. Para a mineração de dados educacionais é recomendado o uso de algoritmos de classificação do tipo "caixa branca", que geram modelos de fácil interpretação e podem ser usados diretamente para a tomada de decisão [Márquez-Vera et al., 2013b].

Após a aplicação dos algoritmos de classificação o resultado é disponibilizado para a sua avaliação e interpretação.

Para validar o método proposto de seleção e escolha dos melhores atributos para a mineração foram realizados experimentos com seis *datasets*, composto de dados acadêmicos de alunos da UTFPR, que são detalhados a seguir.

Capítulo 4

Experimentos

Este capítulo apresenta os experimentos realizados para a escolha dos melhores atributos a serem utilizados na mineração de dados. São apresentadas também a análise dos resultados do estudo com os experimentos.

4.1 Experimentos Realizados

Os primeiros experimentos realizados neste trabalho geraram um artigo apresentado no I Workshop de Mineração de Dados em Ambientes Virtuais de Ensino/Aprendizagem, do 3º Congresso Brasileiro de Informática na Educação (CBIE 2014), enquadrado no tópico de interesse: "Mineração de Dados Educacionais no Ensino Superior". O artigo intitulado "Correlação da Evasão de Cursos com o Empréstimo de Livros em Biblioteca" [Oliveira Júnior et al., 2014] permitiu realizar os primeiros experimentos para investigação dos temas propostos. Os experimentos foram realizados com 3.605 alunos da UTFPR indicaram haver indícios da correlação entre o empréstimo de livros em biblioteca com a permanência do aluno no curso. Foram aplicados cinco algoritmos de classificação e o melhor desempenho foi obtido com o classificador de árvore de decisão J48, com acurácia próxima a 80%.

No prosseguimento da pesquisa foi criado, e colocado em produção na UTFPR, um DW com a finalidade de auxiliar a análise dos dados sobre a evasão e na extração de atributos para a mineração de dados. Essa tarefa originou o artigo apresentado no II Workshop de Mineração de Dados em Ambientes Virtuais de Ensino/Aprendizagem, do 4º Congresso Brasileiro de Informática na Educação (CBIE 2015), intitulado "Uma Abordagem de *Data Warehouse* Educacional para Apoio à Tomada de Decisão" [Oliveira Júnior et al., 2015], enquadrado no tópico de interesse "Bancos de Dados Educacionais". O artigo apresenta uma abordagem baseada no uso de visões materializadas para construção do *Data Warehouse* e no uso de *table functions* para agregação das regras de negócio, tendo como sub produto uma ferramenta OLAP para apoio à tomada de decisão para gestores educacionais (coordenadores de curso, diretores de ensino, entre outros) intitulado "Relatórios Analíticos de Gestão". Esse software foi registrada junto ao INPI (Instituto Nacional da Propriedade Industrial) sob o nº BR 512015001572-9. A aplicação da modelagem multidimensional permitiu aos gestores educacionais a obtenção de informações em diversos níveis, auxiliando na tarefa de tomada de decisão.

Com o DW implementado, foram realizados novos experimentos de mineração de dados para a aplicação do método proposto de seleção dos melhores atributos para classificação.

Foi utilizado nos experimentos o ambiente de mineração de dados WEKA, reconhecido como um sistema de referência em mineração de dados e aprendizado de máquina [Hall et al., 2009].

Para a criação dos *datasets* no formato ARFF[1] (*Attribute-Relation File Format*), os atributos que possuíam valores ausentes foram substituídos pelo caractere "?", e nos atributos numéricos com casas decimais, a vírgula foi substituída por ponto, conforme definido pela ferramenta WEKA.

4.1.1 Criação de Atributos

Para a realização dos experimentos, foram criados 11 atributos (conforme indicado no Quadro 3.1), que estão detalhados na Seção 3.4.2. A criação desses atributos foi fruto de uma investigação dos registros existentes na base de dados que poderiam compor estes novos atributos, e que pudessem ser úteis na análise da evasão.

Com os atributos criados, somados aos atributos já existentes, foram gerados três *datasets*, em que foram utilizados os registros de alunos avaliados durante seis semestres letivos, conforme detalhado na Tabela 4.1. Foi escolhida a janela de tempo de seis semestres por considerar um período suficiente para analisar a evasão. Pode-se escolher, conforme a necessidade, uma janela de tempo diferente da utilizada neste trabalho.

Tabela 4.1: *Datasets* criados para os experimentos com os dados de todos os alunos da UTFPR

Dataset	**Ingressantes**	**Ano/Semestre fim**	**Total de alunos**	***Outlier***	**% de *outlier***	**Nº de atributos**
Dataset1	2010/1	2012/2	2565	130	5,07%	36
Dataset2	2011/1	2013/2	2847	153	5,37%	36
Dataset3	2012/1	2014/2	3438	113	3,29%	36

4.1.2 Normalização e Remoção dos Valores Discrepantes

Nos dados utilizados foram normalizados os valores de notas do ENEM para o intervalo [0.00,1.00]. Para a criação dos *datasets* foram removidos os valores discrepantes do atributo idade, utilizando a variação interquartil com $outlier\_factor = 3$.

As Figuras D.1, D.2 e D.3, no Apêndice D, mostram o diagrama dos atributos do *dataset1*, *dataset2* e *dataset3* antes do balanceamento das classes, com os valores discrepantes removidos.

4.1.3 Balanceamento das Classes

Para o balanceamento das classes foi aplicado o algoritmo SMOTE [Chawla et al., 2002], com os percentuais de instâncias sintéticas inseridas, indicados na Tabela 4.2.

Com a aplicação do algoritmo SMOTE, muitas das instâncias sintéticas criadas ficam com mais casas decimais que o atributo utiliza. Por exemplo, o atributo coeficiente de rendimento é definido com 2 casas decimais. Após a aplicação do algoritmo SMOTE o atributo ficou

[1] http://www.cs.waikato.ac.nz/ml/weka/arff.html

Tabela 4.2: Distribuição de classes do atributo alvo

Dataset	Nº de registros	Não Evadidos	Evadidos	% de evadidos	% de instâncias inseridas
Dataset1	2435	1498	937	38,48%	59,87%
Dataset2	2694	1626	1068	39,64%	52,24%
Dataset3	3325	1885	1440	43,31%	30,90%

com mais de 2 casas decimais, influenciando os resultados da classificação. Para contornar esse problema aplicou-se, após o balanceamento das classes, o filtro de atributo não supervisionado NumericCleaner, disponível na ferramenta WEKA, para reduzir a quantidade de casas decimais de alguns atributos.

4.1.4 Escolha dos Melhores Atributos

Os *datasets* utilizados nos experimentos possuem 35 atributos previsores. Para a escolha dos melhores atributos utilizou-se o método que está detalhado na Seção 3.5.1.

Seleção do Subconjunto de Atributos

O primeiro passo é a aplicação dos algoritmos de seleção de atributos nos *datasets*. Os algoritmos de seleção de atributos utilizados nos experimentos foram os seguintes:
- Abordagem *filter*: CfsSubsetEval, ChiSquaredAttributeEval, GainRatioAttributeEval, InfoGainAttributeEval, OneRAttributeEval, ReliefFAttributeEval e SymmetricalUncertAttributeEval;
- Abordagem *wrapper*: WrapperSubsetEval, utilizando os classificadores de árvore de decisão (J48), baseado em regras (JRip), redes neurais artificiais (MultilayerPerceptron), máquina de vetores de suporte (SVM, com a implementação SMO - *Sequential Minimal Optimization*), métodos de conjunto de classificadores (*RandomForest*) e o classificador K vizinhos mais próximos (IBk). Os algoritmos supracitados, disponíveis na ferramenta WEKA, estão descritos em detalhes em Witten et al. [Witten et al., 2011].

Execução do Classificador em Cada Subconjunto de Atributos

Depois de selecionados os subconjuntos de atributos, os classificadores foram executados dez vezes nesses subconjuntos, usando a técnica de validação cruzada (fator $n = 10$), ou seja, o classificador foi executado 100 vezes em cada subconjunto de atributos. Para essa etapa foi utilizado o ambiente WEE [Hall et al., 2009], utilizando o meta classificador FilteredClassifier. Com esse meta classificador foi aplicado o filtro "attribute.Remove" (para a seleção do subconjunto de atributos) e posteriormente foi aplicado o classificador. Para os subconjuntos selecionados pelos algoritmos com a abordagem *filter* foram selecionados apenas os 10 melhores atributos ranqueados.

As Tabelas A.1 a A.6, no Apêndice A, apresentam a acurácia e o seu desvio padrão da aplicação dos classificadores nos subconjuntos de atributos selecionados, indicando se houve degradação da significância estatística.

Exclusão dos Subconjuntos sem Significância Estatística

Com o objetivo de avaliar a significância estatística dos resultados obtidos, utilizou-se a técnica de teste estatístico conhecida como "T-pareado" (*pairwise T-test*) [Witten et al., 2011], com nível de significância de 5%.

A partir do resultado do teste "T-pareado", considerando o nível de significância de 5%, foram desprezados os atributos selecionados em que a acurácia não obteve significância estatística quando comparados com a melhor acurácia obtida (denominada *test base* no WEE).

Ordenação e Escolha dos Melhores Atributos

Para se obter os melhores atributos foi utilizado o seguinte procedimento:

1. Ordenou-se de forma decrescente a frequência em que o atributo foi selecionado pelos algoritmos WrapperSubsetEval e CfsSubsetEval, que não utilizam o método de busca Ranking;

2. Ordenou-se de forma crescente pela posição média em que o atributo foi classificado pelos algoritmos que utilizam o método de busca Ranking.

3. Selecionou-se os 10 melhores atributos.

As Tabelas B.1 a B.6, no Apêndice B, exibem os resultados da escolha dos melhores atributos.

4.1.5 Seleção de Atributos pelo Algoritmo "OneRAttributeEval"

Um dos algoritmos com a abordagem *filter* utilizado na etapa de seleção do subconjunto de atributos foi o algoritmo OneRAttributeEval, que faz o ranqueamento dos atributos aplicando-se o classificador OneR. O algoritmo OneR baseia-se no pressuposto de que "frequentemente um único atributo é suficiente para determinar a classe". O atributo escolhido por esse algoritmo é aquele que possuir menor erro de predição, discretizando-se os atributos numéricos [Holte, 1993]. Mesmo sabendo-se que com um único atributo pode não ser suficiente para se realizar uma boa classificação, o algoritmo OneRAttributeEval é muito útil para avaliar a influência dos atributos criados, avaliando cada atributo individualmente.

A Tabela 4.3 mostra os cinco melhores atributos ranqueados pelo algoritmo OneRAttributeEval, aplicados no *dataset1*, *dataset2* e *dataset3*.

A seguir são apresentados experimentos aplicados nos três níveis organizacionais.

4.1.6 Experimento com a Base Completa

O intuito desse experimento é trazer informação da mineração de dados para os gestores educacionais situados no nível estratégico, como por exemplo, uma pró-reitoria de graduação. Nesse experimento foram aplicados os algoritmos de classificação para todos os alunos de cursos semestrais de graduação da instituição, que ingressaram pelo SISU no 1º semestre de 2012 e foram avaliados até o 2º semestre de 2014 (6 semestres letivos). Foram utilizados os atributos selecionados dentre os dez melhores em cada classificador, conforme Apêndice B. Esse conjunto de registros foi nomeado de *dataset4*.

Tabela 4.3: Cinco melhores atributos ranqueados pelo algoritmo OneRAttributeEval

Dataset1			
Nº	Atributo	Acurácia	Atributo criado
12	regrecao_coeficiente	90,5478%	sim
13	percentual_frequencia	86,5731%	sim
32	socio_esc_mae	82,7321%	
8	previsao_evasao_dificuldade	81,0287%	sim
34	socio_fez_cursinho	80,5611%	
Dataset2			
Nº	Atributo	Acurácia	Atributo criado
12	regrecao_coeficiente	90,7749%	sim
13	percentual_frequencia	85,4859%	sim
8	previsao_evasao_dificuldade	79,2743%	sim
15	percentual_aprov	78,9360%	sim
14	coeficiente_rendimento	77,3063%	
Dataset3			
Nº	Atributo	Acurácia	Atributo criado
12	regrecao_coeficiente	91,1671%	sim
13	percentual_frequencia	83,3687%	sim
15	percentual_aprov	83,2626%	sim
14	coeficiente_rendimento	82,6790%	
8	previsao_evasao_dificuldade	81,1936%	sim

Na mineração de dados educacionais é importante saber os motivos que deram origem à classificação dos registros. Sendo assim, foram utilizados neste experimento somente algoritmos de classificação do tipo "caixa branca", que, neste caso, são os classificadores de árvore de decisão (J48) e baseado em regras (JRip).

Na aplicação do algoritmo J48 foram modificados os seguintes parâmetros:

- *minNumObj = 50*, para efetuar a poda da árvore. Este parâmetro representa o número mínimo de instâncias por folha. Assim, o tamanho da árvore de decisão foi reduzido de 203 para 27 elementos, conforme indicado na Figura 4.1.

- *binarySplits = true*, para gerar uma árvore binária, que facilita a interpretação pelo usuário final.

O algoritmo de árvore de decisão J48 obteve a acurácia de 85,78%, classificando corretamente 3.234 dos 3.770 registros. A Tabela 4.4 mostra a matriz de confusão desse experimento.

Tabela 4.4: Matriz de confusão gerada com o algoritmo J48 no *dataset4*

		Classes previstas	
		evasão=não	**evasão=sim**
Classes corretas	**evasão=não**	1677	208
	evasão=sim	328	1557

Na Tabela 4.5 são destacadas três medidas de desempenho do algoritmo J48, permitindo a visualização de mais detalhes dos resultados obtidos.

Tabela 4.5: Medidas de desempenho do algoritmo J48 no *dataset4*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
evasão=não	0,890	0,174	0,836
evasão=sim	0,826	0,110	0,882
Média	0,858	0,142	0,859

A Figura 4.1 mostra a árvore de decisão gerada pelo algoritmo J48 no *dataset4*. O atributo "dificuldade média das disciplinas cursadas pelo aluno" foi selecionado como o nó raiz, indicando a contribuição desse atributo criado na tarefa de classificação.

O mesmo experimento foi realizado aplicando-se o algoritmo baseado em regras JRip no *dataset4*. Esse algoritmo obteve a acurácia de 87,69%. A Tabela 4.6 mostra a matriz de confusão desse experimento.

Tabela 4.6: Matriz de confusão gerada com o algoritmo JRip no *dataset4*

		Classes previstas	
		evasão=não	**evasão=sim**
Classes corretas	**evasão=não**	1706	179
	evasão=sim	285	1600

A Tabela 4.7 apresenta três medidas de desempenho do algoritmo JRip, permitindo a visualização de mais detalhes dos resultados obtidos.

Tabela 4.7: Medidas de desempenho do algoritmo JRip no *dataset4*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
evasão=não	0,905	0,151	0,857
evasão=sim	0,849	0,095	0,899
Média	0,877	0,123	0,878

A Tabela 4.8 apresenta as 12 regras geradas pelo algoritmo de classificação JRip no *dataset4*.

Tabela 4.8: Conjunto de regras geradas com o classificador JRip no *dataset4*

Nº da regra	**Regra**				**Total de instâncias**	**Classificadas incorretamente**	**Acurácia**
1	(previsao_evasao_dificuldade = Sim)	(percentual_aprov <= 0.14)	=>	evasao=Sim	727	10	98,6%
2	(previsao_evasao_dificuldade = Sim)	(emp_biblio_por_semestre <= 1)	=>	evasao=Sim	420	35	91,7%
3	(previsao_evasao_dificuldade = Sim)	(grau = Licenciatura)					
		(nota_linguagem >= 0.56)	=>	evasao=Sim	88	7	92,0%
4	(coeficiente_rendimento <= 0.49)		=>	evasao=Sim	70	6	91,4%
5	(previsao_evasao_dificuldade = Sim)	(grau = Tecnologia)	=>	evasao=Sim	63	13	79,4%
6	(regrecao_coeficiente <= -0.12)	(regrecao_coeficiente <= -0.17)					
		(percentual_aprov <= 0.7)	=>	evasao=Sim	123	17	86,2%
7	(previsao_evasao_dificuldade = Sim)	(grau = Bacharelado_(outros))	=>	evasao=Sim	41	14	65,9%
8	(percentual_aprov <= 0.68)	(emp_biblio_por_semestre <= 0)	=>	evasao=Sim	13	1	92,3%
9	(percentual_aprov <= 0.71)	(previsao_evasao_dificuldade = Não)	=>	evasao=Sim	16	1	93,8%
10	(regrecao_coeficiente <= -0.07)	(percentual_frequencia >= 0.76)					
	(coeficiente_rendimento >= 0.63)	(coeficiente_rendimento <= 0.71)					
		(nota_final_enem <= 0.65)	=>	evasao=Sim	18	3	83,3%
11	(previsao_evasao_dificuldade = Sim)	(cota = Cotista_(2008-2012))					
	(emp_biblio_por_semestre >= 8)	(emp_biblio_por_semestre <= 15)	=>	evasao=Sim	25	6	76,0%
12			=>	evasao=Não	2166	351	83,8%
Total					**3770**	**464**	**87,7%**

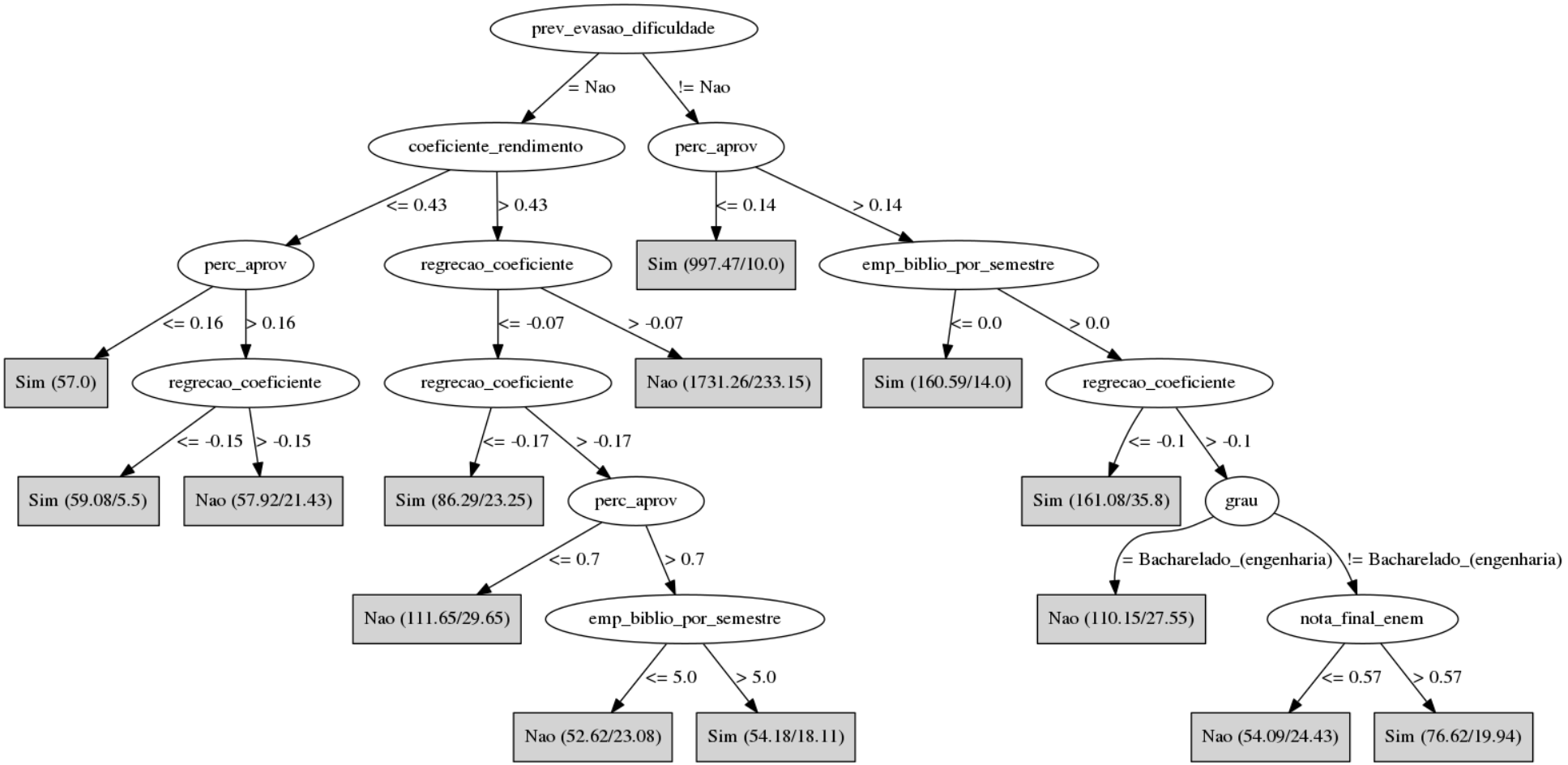


Figura 4.1: Árvore de decisão gerada com o classificador J48 no *dataset4*

4.1.7 Experimento Restrito a um Câmpus

O intuito desse experimento é trazer informação da mineração de dados para os gestores educacionais situados no nível tático da instituição, como por exemplo, os diretores de ensino. Para esse experimento, foi escolhido o câmpus Curitiba por ser o câmpus que teve o maior número de ingressantes em 2012. Foram selecionados os registros dos alunos que ingressaram pelo SISU no 1º semestre de 2012, e foram avaliados até o 2º semestre de 2014 (6 semestres letivos).

O atributo evasão desse novo *dataset* possui 576 alunos não evadidos e 390 alunos evadidos. Foi nomeado esse novo conjunto de registros de *dataset5*.

Foram utilizados os mesmos atributos do experimento anterior, aplicados nos classificadores de árvore de decisão (J48) e baseado em regras (JRip).

Para o algoritmo J48 foram alterados os seguintes parâmetros:

- *minNumObj = 10*, para efetuar a poda da árvore. Assim, o tamanho da árvore de decisão foi reduzido de 35 para 23 elementos, conforme indicado na Figura 4.2.
- *binarySplits = true*, para gerar uma árvore binária, que facilita a interpretação pelo usuário final.

O algoritmo de árvore de decisão J48 obteve a acurácia de 87,06%, classificando corretamente 841 dos 966 registros. A Tabela 4.9 mostra a matriz de confusão desse experimento.

Tabela 4.9: Matriz de confusão gerada com o algoritmo J48 no *dataset5*

		Classes previstas	
		evasão=não	**evasão=sim**
Classes corretas	**evasão=não**	529	47
	evasão=sim	78	312

Na Tabela 4.10 são destacadas três medidas de desempenho do algoritmo J48, permitindo a visualização de mais detalhes dos resultados obtidos.

Tabela 4.10: Medidas de desempenho do algoritmo J48 no *dataset5*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
evasão=não	0,918	0,200	0,871
evasão=sim	0,800	0,082	0,869
Média	0,871	0,152	0,871

A Figura 4.2 mostra a árvore de decisão gerada pelo algoritmo J48 no *dataset5*. O atributo "dificuldade média das disciplinas cursadas pelo aluno" foi novamente selecionado como o nó raiz, indicando a contribuição desse atributo criado na tarefa de classificação.

O mesmo experimento foi realizado aplicando-se o algoritmo baseado em regras JRip no *dataset5*. Esse algoritmo obteve a acurácia de 86,54%. A Tabela 4.11 mostra a matriz de confusão desse experimento.

A Tabela 4.12 apresenta três medidas de desempenho do algoritmo JRip, permitindo a visualização de mais detalhes dos resultados obtidos.

Tabela 4.11: Matriz de confusão gerada com o algoritmo JRip no *dataset5*

		Classes previstas	
		evasão=não	**evasão=sim**
Classes corretas	**evasão=não**	515	61
	evasão=sim	69	321

Tabela 4.12: Medidas de desempenho do algoritmo JRip no *dataset5*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
evasão=não	0,894	0,177	0,882
evasão=sim	0,823	0,106	0,840
Média	0,865	0,148	0,865

A Tabela 4.13 apresenta as 6 regras geradas pelo algoritmo de classificação JRip no *dataset5*.

Tabela 4.13: Conjunto de regras geradas com o classificador JRip no *dataset5*

Nº da regra	**Regra**				**Total de instâncias**	**Classificadas incorretamente**	**Acurácia**
1	(previsao_evasao_dificuldade = Sim)		=>	evasao=Sim	313	42	86,6%
2	(coeficiente_rendimento <= 0.55)	(coeficiente_rendimento <= 0.43)	=>	evasao=Sim	44	8	81,8%
3	(regrecao_coeficiente <= -0.17)	(perc_aprov <= 0.69)	=>	evasao=Sim	22	2	90,9%
4	(emp_biblio_poo_semestre <= 0)	coeficiente_rendimento <= 0.73)	=>	evasao=Sim	14	3	78,6%
5	(socio_reside_em = Pensão)	(coeficiente_rendimento <= 0.7)	=>	evasao=Sim	6	1	83,3%
6			=>	evasao=Não	567	74	86,9%
Total					**966**	**130**	**86,5%**

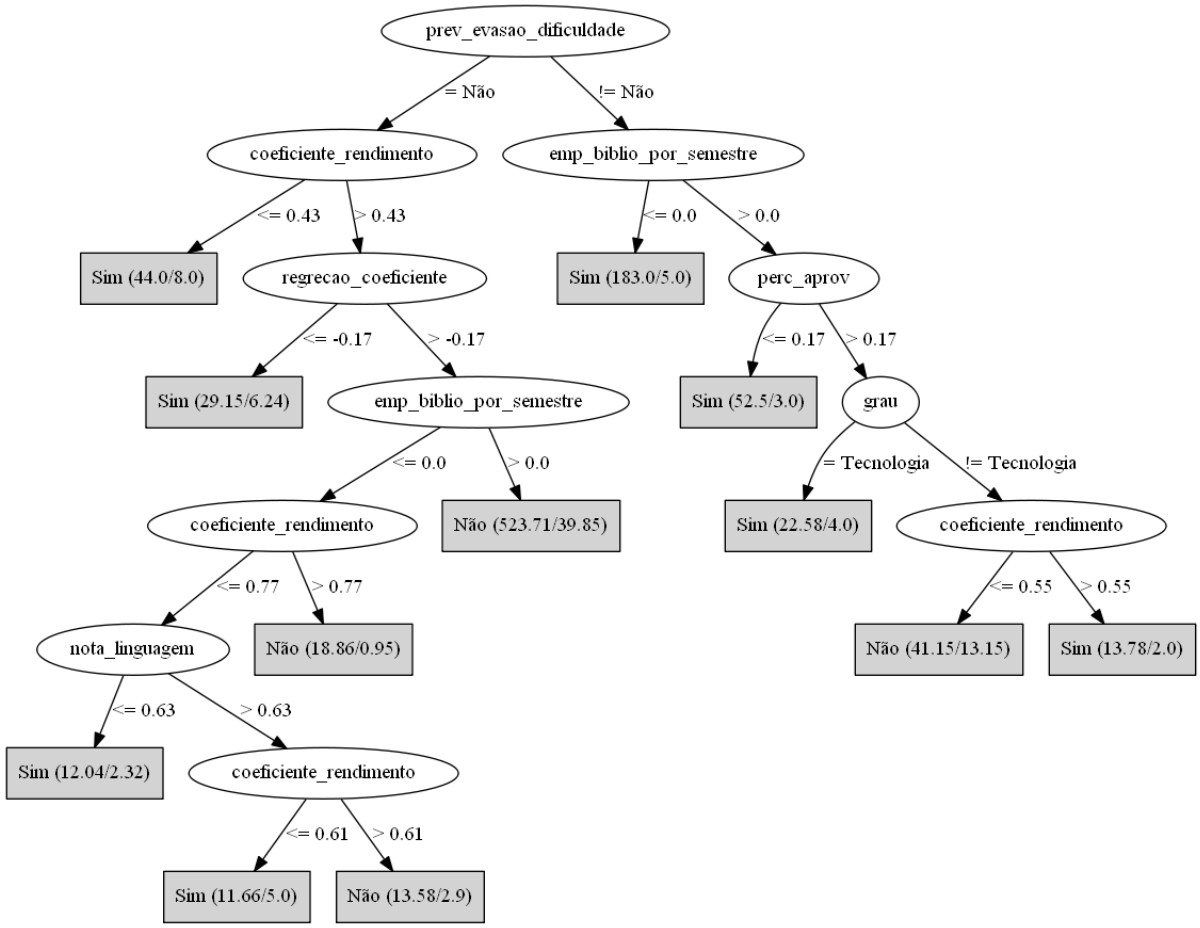


Figura 4.2: Árvore de decisão gerada com o classificador J48 no *dataset5*

4.1.8 Experimento Restrito a um Curso

O intuito desse experimento é trazer informação da mineração de dados para os gestores educacionais situados no nível operacional da instituição, como por exemplo, os coordenadores de curso. Foi escolhido o curso de Engenharia Mecânica, do câmpus Curitiba, por ser o curso que teve o maior número de ingressantes em 2012. Desse curso, foram selecionados os registros dos alunos que ingressaram pelo SISU, no 1º semestre de 2012, e foram avaliados até o 2º semestre de 2014 (6 semestres letivos).

O atributo evasão desse novo *dataset* possui 73 alunos não evadidos e 17 alunos evadidos. Foi nomeado esse novo conjunto de registros de *dataset6*.

Foram utilizados os mesmos atributos do experimento anterior, aplicado nos classificadores de árvore de decisão (J48) e baseado em regras (JRip).

O algoritmo de árvore de decisão J48 obteve a acurácia de 91,1111%, classificando corretamente 82 dos 90 registros. A Tabela 4.14 mostra a matriz de confusão desse experimento.

Na Tabela 4.15 são destacadas três medidas de desempenho do algoritmo J48, permitindo a visualização de mais detalhes dos resultados obtidos.

A Figura 4.3 mostra a árvore de decisão gerada pelo algoritmo J48 no *dataset6*.

Tabela 4.14: Matriz de confusão gerada com o algoritmo J48 no *dataset6*

		Classes previstas	
		evasão=não	**evasão=sim**
Classes corretas	**evasão=não**	64	4
	evasão=sim	4	13

Tabela 4.15: Medidas de desempenho do algoritmo J48 no *dataset6*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
evasão=não	0,945	0,235	0,945
evasão=sim	0,765	0,055	0,765
Média	0,911	0,201	0,911

O mesmo experimento foi realizado aplicando-se o algoritmo baseado em regras JRip no *dataset6*. Esse algoritmo obteve a acurácia de 85,5556%. A Tabela 4.16 mostra a matriz de confusão desse experimento.

Tabela 4.16: Matriz de confusão gerada com o algoritmo JRip no *dataset6*

		Classes previstas	
		evasão=não	**evasão=sim**
Classes corretas	**evasão=não**	68	5
	evasão=sim	8	9

A Tabela 4.17 apresenta três medidas de desempenho do algoritmo JRip, permitindo a visualização de mais detalhes dos resultados obtidos.

A Tabela 4.18 apresenta as 3 regras geradas pelo algoritmo de classificação JRip no *dataset6*.

De forma análoga aos dois experimentos anteriores, onde foram utilizados alunos de todos os câmpus e de um câmpus específico, tanto a árvore de decisão gerada pelo algoritmo J48 (Figura 4.3), quanto as regras geradas pelo algoritmo JRip (Tabela 4.18) apresentam aos gestores educacionais uma ferramenta para auxílio na análise da evasão. Como os resultados foram extraídos do conjunto de alunos de um curso, essa informação pode ser útil aos gestores educacionais situados no nível operacional da instituição, neste caso, um coordenar de curso.

4.2 Análise dos Resultados

A análise dos resultados do estudo será feita inicialmente com os dados levantados sobre a evasão, antes da aplicação dos algoritmos de mineração. Na sequência, é analisado o método proposto de seleção dos melhores atributos para classificação. Concluindo, são analisados os resultados da mineração com todos os alunos da instituição, com alunos de um câmpus e com alunos de um curso, abrangendo assim os três níveis organizacionais para a tomada de decisão.

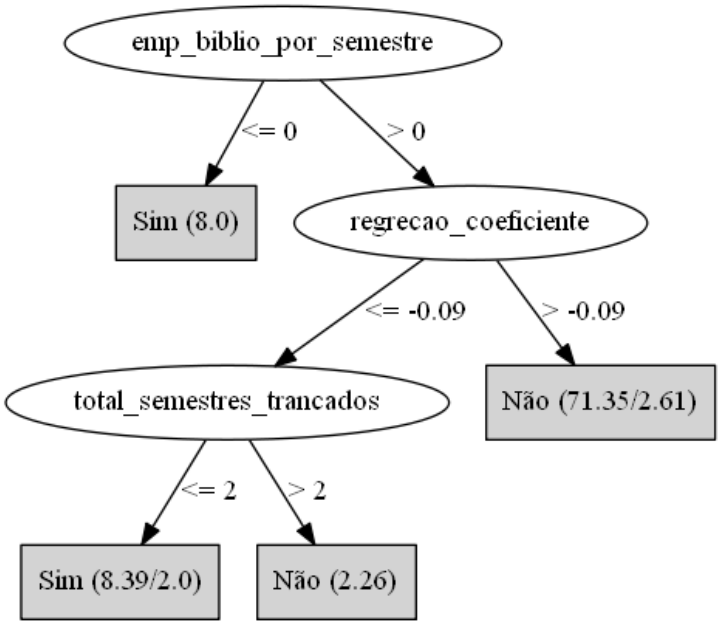


Figura 4.3: Árvore de decisão gerada com o classificador J48 no *dataset6*

Tabela 4.17: Medidas de desempenho do algoritmo JRip no *dataset6*

Classe	**Taxa de verdadeiro positivo**	**Taxa de falso positivo**	**Precisão**
evasão=não	0,932	0,471	0,895
evasão=sim	0,529	0,068	0,643
Média	0,856	0,395	0,847

Tabela 4.18: Conjunto de regras geradas com o classificador JRip no *dataset6*

Nº da regra	**Regra**				**Total de instâncias**	**Classificadas incorretamente**	**Acurácia**
1	(emp_biblio_por_semestre <= 0)		=>	evasao=Sim	8	0	100,0%
2	(regrecao_coeficiente <= -0.09)		=>	evasao=Sim	10	4	60,0%
3			=>	evasao=Não	72	3	95,8%
Total					**90**	**7**	**92,2%**

4.2.1 Análise Preliminar da Evasão

Antes da aplicação dos algoritmos de mineração, foi investigado em que período (da matriz curricular do curso) os alunos se evadem. Foi verificado que aproximadamente 80% da evasão de curso ocorre até o 3º período, independente se o curso possui duração de 6, 8 ou 10 períodos em sua matriz curricular. Essa informação permite aos gestores educacionais concentrar os esforços no acompanhamento dos alunos principalmente até o 3º período, no qual se concentram 80% das evasões.

Analisando os gráficos da taxa de evasão semestral por quinquênio (Figura 3.6) e da taxa de evasão semestral de 2010 a 2014 (Figura 3.7), observa-se que a evasão na instituição é um fenômeno em crescimento, indicando uma necessidade de intervenção pelos gestores educacionais. Na Figura 3.7, a taxa de evasão dos cursos de engenharia no 1º semestre de 2010

apresenta o valor -14,01%. Esse valor negativo foi ocasionado pois, em 2010, houve uma migração dos alunos dos antigos cursos de engenharia industrial para os cursos de engenharia sem a denominação "industrial". Esse fato influenciou o valor da taxa de evasão de apenas 0,95% no 1º semestre de 2010. Fazendo uma simulação, se o valor da taxa de evasão para o 1º semestre de 2010 fosse de 7,71% a linha de tendência da evasão semestral de 2010 a 2014 tenderia à estabilidade, com coeficiente de regressão próximo a zero.

Os cursos de Licenciatura são os que apresentam as maiores taxas de evasão. Esses cursos tiveram em 2014 uma taxa de evasão semestral duas vezes maior que a taxa de evasão semestral da UTFPR, como pode-se verificar na Figura 3.7. As altas taxas de evasão nas Licenciaturas ocorrem também nas demais IESs há bastante tempo [SESu/MEC, 1996], ocasionando como prejuízo a falta de professores para o ensino médio [Ristoff, 1999].

4.2.2 Análise do Método de Seleção dos Melhores Atributos

A seleção dos melhores atributos para a mineração, aplicada na análise da evasão, demonstrou que os melhores atributos variam de acordo com a amostra selecionada, que no estudo abrangeu alunos com ingresso em três anos distintos, avaliados durante 6 semestres letivos. Isso ressalta a importância da seleção dos melhores atributos utilizando o método proposto.

A abordagem de seleção de subconjunto de atributos *wrapper* obteve a melhor acurácia com os classificadores J48, JRip, MLP, SMO, IBk e *RandomForest*, com acurácias variando entre 83 e 94%, conforme indicado nas Tabelas do Apêndice A.

Na aplicação dos seis classificados supracitados no *dataset3* somente o algoritmo com abordagem *wrapper* obteve significância estatística.

Na aplicação dos algoritmos MLP e *RandomForest* no *dataset1*, *dataset2* e *dataset3*, nenhum dos subconjuntos de atributos selecionados com a abordagem *filter* obteve significância estatística.

De acordo com os resultados apresentados nas Tabelas do Apêndice B, dos dez melhores atributos classificados, pelo menos cinco deles são atributos criados, auxiliando assim a tarefa de análise da evasão. O atributo com maior destaque foi o atributo "regressão linear do coeficiente de rendimento", que ficou classificado nos seis algoritmos de classificação em 1º, 2º ou 3º lugar. O atributo "dificuldade média das disciplinas cursadas pelo aluno" ficou em 2º lugar nos dois classificadores tipo "caixa branca" (J48 e JRip), que são os classificadores recomendados quando se quer entender os motivos da classificação. O atributo coeficiente de rendimento, que mede o desempenho individual do aluno, foi selecionado entre os dez melhores atributos nos classificadores, com exceção do classificador IBk. Analisando a Tabela B.6, verifica-se que o classificador IBk selecionou apenas sete atributos, sendo esse o classificador que selecionou a menor quantidade de atributos nos experimentos.

Avaliando-se os atributos criados na aplicação do algoritmo de classificação OneR, na Tabela 4.3, verifica-se que dos 11 atributos criados três deles se destacam: "regressão linear do coeficiente de rendimento", "percentual de frequência" e "dificuldade média das disciplinas cursadas pelo aluno", que foram ranqueados entre os cinco melhores atributos no *dataset1*, *dataset2* e *dataset3*, avaliados pelo algoritmo OneR. O atributo "regressão linear do coeficiente de rendimento" apresentou nos três *datasets* acurácia superior a 90%, podendo ser considerada expressiva no contexto educacional.

4.2.3 Análise do Experimento com a Base Completa

No experimento com todos os alunos de graduação da instituição que ingressaram pelo SISU, representado pelo *dataset4*, pode-se verificar que o algoritmo JRip obteve acurácia de 87,69% e o algoritmo J48 obteve acurácia de 85,78%, podendo ser consideradas expressivas no contexto educacional.

No algoritmo J48 foi alterado o parâmetro *minNumObj=50* para realizar a poda da árvore, reduzindo o tamanho da árvore de 203 para 27 elementos, tornando-se assim um modelo de mais fácil interpretação.

Analisando as matrizes de confusão geradas pelos algoritmos J48 (Tabela 4.4) e JRip (Tabela 4.6) verifica-se que a classificação gerou mais falsos positivos que falsos negativos. Para a análise da evasão ter mais falsos positivos (prever o aluno como não evadido, mas, ele se evadiu) é um problema. Uma forma de minimizar isso é aplicar a classificação sensível ao custo, atribuindo um peso maior para os falsos positivos.

Nesse experimento a árvore de decisão gerada pelo algoritmo J48 apresentou como nó raiz o atributo "dificuldade média das disciplinas cursadas pelo aluno", indicando novamente a contribuição da criação de atributos na tarefa de classificação. Nesse experimento o algoritmo JRip gerou como resultado um conjunto de 12 regras, que são visualizadas na Tabela 4.8.

Como os resultados desse experimento foram extraídos do conjunto de alunos de todos os câmpus, essa informação pode ser útil aos gestores educacionais situados no nível estratégico da instituição, como por exemplo, uma pró-reitoria de graduação, podendo também ser estendido aos gestores educacionais dos níveis tático e operacional.

4.2.4 Análise do Experimento Restrito a um Câmpus

No experimento com todos os alunos de graduação de um câmpus, que ingressaram pelo SISU, representado pelo *dataset5*, pode-se verificar que o algoritmo JRip obteve acurácia de 86,54% e o algoritmo J48 obteve acurácia de 87,06%, podendo ser consideradas expressivas no contexto educacional.

No algoritmo J48 foi alterado o parâmetro *minNumObj=10* para realizar a poda da árvore, reduzindo o tamanho da árvore de 49 para 23 elementos, tornando-se assim um modelo de mais fácil interpretação.

Analisando as matrizes de confusão geradas pelos algoritmos J48 (Tabela 4.9) e JRip (Tabela 4.11) verifica-se que a classificação gerou mais falsos positivos que falsos negativos. Novamente, para a análise da evasão ter mais falsos positivos (prever o aluno como não evadido, mas, ele se evadiu) é um problema. Uma forma de minimizar isso é aplicar a classificação sensível ao custo.

Nesse experimento a árvore de decisão gerada pelo algoritmo J48 apresentou como nó raiz novamente o atributo "dificuldade média das disciplinas cursadas pelo aluno", indicando a sua importância na tarefa de classificação. Nesse experimento o algoritmo JRip gerou como resultado um conjunto de seis regras, que são visualizadas na Tabela 4.13.

Como os resultados desse experimento foram extraídos do conjunto de alunos de um câmpus, essa informação pode ser útil aos gestores situados no nível tático da instituição, como por exemplo, uma diretoria de ensino, podendo também ser estendido aos gestores educacionais do nível operacional.

4.2.5 Análise do Experimento Restrito a um Curso

No experimento com os alunos de um curso, que ingressaram pelo SISU, representado pelo *dataset6*, pode-se verificar que o algoritmo JRip obteve acurácia de 85,55% e o algoritmo J48 obteve acurácia de 91,11%, podendo ser consideradas expressivas no contexto educacional.

No algoritmo J48 não houve a necessidade de poda da árvore, já que a mesma possui apenas sete elementos.

Nesse experimento a árvore de decisão gerada pelo algoritmo J48 apresentou como nó raiz o atributo "média de empréstimo na biblioteca por semestre", indicando a importância da criação de atributos na tarefa de classificação. Nesse experimento o algoritmo JRip gerou como resultado um conjunto de somente três regras, que são visualizadas na Tabela 4.18.

Como os resultados desse experimento foram extraídos do conjunto de alunos de um curso, essa informação pode ser útil aos gestores educacionais situados no nível operacional da instituição, como por exemplo, os coordenadores de curso.

Esse experimento apenas mostra um exemplo de mineração de dados de um curso. É importante ressaltar que a mineração de dados de cursos dificilmente serão similares, devendo ser realizadas individualmente para cada curso.

4.2.6 Resumo dos Experimentos

Tanto as árvores de decisão (Figuras 4.1, 4.2 e 4.3), geradas pelo algoritmo J48, quanto o conjunto de regras (Tabelas 4.8, 4.13 e 4.18), geradas pelo algoritmo JRip, apresentam aos gestores educacionais uma ferramenta muito útil para o auxílio na identificação dos motivos da evasão, auxiliando assim o processo de tomada de decisão.

Nos dois conjuntos de regras geradas pelo algoritmo JRip (Tabelas 4.8, 4.13), os atributos "dificuldade média das disciplinas cursadas pelo aluno', "regressão linear do coeficiente de rendimento" e "média de empréstimos na biblioteca" constam nas regras, reforçando novamente a importância da criação de atributos.

O atributo "dificuldade média das disciplinas cursadas pelo aluno" revelou-se uma boa medida de prognóstico de desempenho do aluno, tendo como diferencial um componente coletivo em sua avaliação individual.

O atributo "regressão linear do coeficiente de rendimento", que indica a tendência do desempenho do aluno para o semestre subsequente, também se revelou um bom indicador para ser usado na previsão da evasão.

Com os resultados supracitados pode-se concluir que a criação de atributos contribuiu efetivamente para a tarefa de mineração de dados.

Capítulo 5

Conclusões e Trabalhos Futuros

Este capítulo apresenta as conclusões desta pesquisa, elenca as contribuições e indica possíveis trabalhos futuros.

5.1 Conclusões

Neste trabalho investigou-se como os dados armazenados em sistemas de registros acadêmicos poderiam ser transformados em informações potencialmente úteis para apoiar a tomada de decisão com a finalidade de reduzir a evasão, com o uso de técnicas de mineração de dados.

Conforme verificado neste trabalho, a evasão ainda é um fenômeno em crescimento na UTFPR, justificando a necessidade de se gerar inferências para a identificação de padrões para a sua análise.

Este trabalho apresentou um método de seleção dos melhores atributos para classificação, aplicado na previsão da evasão escolar, utilizando a criação de atributos e a seleção dos melhores atributos previsores. Esse método contribuiu para o processo de identificação de padrões a serem utilizados na previsão da evasão no ensino superior.

A arquitetura de *Data Warehouse* [Oliveira Júnior et al., 2015] implementada auxiliou no processo de mineração de dados, principalmente por realizar as tarefas de limpeza, extração, transformação e carga dos dados, passos importantes na tarefa de mineração de dados, além de permitir a geração de relatórios analíticos dos dados sobre a evasão, disponibilizados aos gestores educacionais da UTFPR.

O algoritmo de seleção de atributos que apresentou os melhores resultados para a acurácia foi o WrapperSubsetEval, que utiliza a abordagem *wrapper*, empregando os classificadores de árvore de decisão (J48), baseado em regras (JRip), máquina de vetores de suporte (SVM, com a implementação SMO), redes neurais artificiais (MultilayerPerceptron), métodos de conjunto de classificadores (*RandomForest*) e o classificador K vizinhos mais próximos (IBk). Este resultado é consistente com o indicado em Hall e Holmes [Hall and Holmes, 2003], em que a abordagem *wrapper* também aparece com os melhores resultados.

Entre os algoritmos de classificação do tipo "caixa branca" as melhores acurácias foram obtidas com o classificador baseado em regras JRip. As redes neurais artificiais, com o algoritmo Multilayer Perceptron, apresentaram as melhores acurácias entre todos os classificadores utilizados nos experimentos.

Na escolha dos melhores atributos para a tarefa de mineração, pelo menos metade dos dez melhores atributos foram atributos criados, indicando a sua contribuição na tarefa de classificação. A criação do atributo "dificuldade média das disciplinas cursadas pelo aluno" melhorou a acurácia dos algoritmos de classificação, tendo como diferencial agregar um componente coletivo ao desempenho individual do aluno. Outro atributo criado que auxilia na predição da evasão é o atributo "regressão linear do coeficiente de rendimento", que indica a tendência do desempenho do aluno para o semestre subsequente.

Os algoritmos de classificação do tipo "caixa branca", como os algoritmos de árvore de decisão (J48) e baseado em regras (JRip), revelaram-se mais adequados na mineração de dados educacionais, pois permitem ao usuário entender os motivos do resultado da mineração.

Na análise preliminar da evasão, foi constatado que 80% das evasões concentram-se até o 3º período do curso, independente do total de períodos do curso (6, 8 ou 10 períodos). Isto reduz o escopo a ser analisado e permite concentrar os esforços em medidas de diminuição da evasão.

A abordagem computacional proposta neste trabalho possui vantagens, como a de se gerar inferências sobre todo o conjunto de alunos matriculados na instituição, mas possui também desvantagens, apresentando apenas os sintomas e assim ocultando as reais causas da evasão. Uma maneira de melhorar os resultados da mineração de dados pode ser com a coleta de outras informações dos alunos, como por exemplo, através de questionários específicos, uso de ambientes virtuais de aprendizagem, entre outros, tendo-se o cuidado de armazenar essas informações ao longo do tempo.

Com a abordagem computacional proposta para a análise da evasão, permite-se identificar padrões que podem auxiliar a tomada de decisão dos gestores educacionais, utilizando-se indicadores e/ou conjunto de regras que permitem avaliar a possibilidade da evasão de cada aluno, abrangendo do nível estratégico (pró-reitorias de graduação) até o nível operacional (coordenações de curso), com o uso de técnicas de mineração de dados.

5.2 Trabalhos Futuros

Analisar a evasão apenas utilizando as inferências geradas pela mineração de dados pode ser insuficiente do ponto de vista computacional. Entende-se como importante unir os resultados de mineração de dados com análises multidimensionais em um *Data Warehouse* [Oliveira Júnior et al., 2015]. Dessa forma pretende-se trabalhar na inclusão de mais assuntos no *Data Warehouse* e a criação *Dashboards* para os principais perfis de gestores educacionais, criando alertas que possibilitem realizar a intervenção junto ao aluno em tempo oportuno.

Na parte de mineração de dados a pesquisa também pode ser ampliada. Pretende-se aplicar o método proposto em outras bases de dados e também avaliar a aplicação de classificação sensível ao custo, onde erros de classificação de diferentes tipos (instância "não haverá evasão" classificada como "haverá evasão" e vice-versa) tenham a si associados custos diferentes.

É importante, também, ampliar o escopo e tratar a evasão em conjunto o problema da retenção de alunos, que faz com que os graduandos demorem um tempo superior ao previsto para a conclusão do curso.

Sugere-se ainda realizar uma comparação multidisciplinar com as áreas da psicologia e educação, aprofundando o estudo, minimizando assim as deficiências de uma abordagem apenas computacional, que foi o foco deste trabalho.

Neste sentido, este trabalho pode ser considerado como um ponto de partida, constituindo-se em uma ferramenta que fornece dados quantitativos para uma análise mais ampla do problema da evasão em cursos superiores de graduação.

Referências Bibliográficas

[Aha et al., 1991] Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, 6(1):37–66.

[Antunes, 2010] Antunes, C. (2010). *Anticipating student s failure as soon as possible*. Chapman & Hall/CRC Press, New York, NY.

[Aparecida et al., 2011] Aparecida, C., Baggi, S., and Lopez, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior*, 16(2):355–374.

[Baker et al., 2011] Baker, R., Isotani, S., and Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19(02):03.

[Bardagi and Hutz, 2014] Bardagi, M. and Hutz, C. S. (2014). Evasão universitária e serviços de apoio ao estudante: uma breve revisão da literatura brasileira. *Revista Psicologia. Revista da Faculdade de Ciências Humanas e da Saúde. ISSN 1413-4063*, 14(2):279–301.

[Blumer et al., 1987] Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. (1987). Occam's razor. *Information processing letters*, 24(6):377–380.

[Borges et al., 2015] Borges, V. A., Nogueira, B. M., and Barbosa, E. F. (2015). Uma análise exploratória de tópicos de pesquisa emergentes em informática na educação. *Revista Brasileira de Informática na Educação*, 23(01):85.

[Breiman et al., 1984] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

[Chau and Phung, 2013] Chau, V. T. N. and Phung, N. H. (2013). Imbalanced educational data classification: An effective approach with resampling and random forest. In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference on*, pages 135–140. IEEE.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

[Chen et al., 2012] Chen, H., Chiang, R. H., and Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4):1165–1188.

[Cohen, 1995] Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning, Lake Tahoe, California*.

[Dekker et al., 2009] Dekker, G. W., Pechenizkiy, M., and Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. In *Educational Data Mining 2009*, pages 41–50, Córdoba, Espanha.

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

[Gerhardt and Silveira, 2009] Gerhardt, T. E. and Silveira, D. T. (2009). *Métodos de pesquisa*. Editora da UFRGS.

[Gorunescu, 2011] Gorunescu, F. (2011). *Data Mining: Concepts, models and techniques*, volume 12. Springer.

[Gottardo et al., 2012] Gottardo, E., Kaestner, C., and Noronha, R. V. (2012). Previsão de desempenho de estudantes em cursos ead utilizando mineração de dados: uma estratégia baseada em séries temporais. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 23.

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

[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.

[Hall and Holmes, 2003] Hall, M. A. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering, IEEE Transactions on*, 15(6):1437–1447.

[Hämäläinen and Vinni, 2011] Hämäläinen, W. and Vinni, M. (2011). Classifiers for educational data mining. *Handbook of Educational Data Mining, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series*, pages 57–71.

[Han et al., 2011] Han, J., Kamber, M., and Pei, J. (2011). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

[Holte, 1993] Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1):63–90.

[Inmon, 2005] Inmon, W. H. (2005). *Building the Data Warehouse, 4rd Edition*. Wiley India Pvt. Limited, 4rd edition.

[John et al., 1994] John, G. H., Kohavi, R., Pfleger, K., et al. (1994). Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the Eleventh International Conference*, pages 121–129.

[Kendall and Kendall, 2010] Kendall, K. E. and Kendall, J. E. (2010). *Systems Analysis and Design*. Prentice Hall Press, Upper Saddle River, NJ, USA, 8th edition.

[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 John, 1997] Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324.

[Kotsiantis et al., 2003] Kotsiantis, S. B., Pierrakeas, C., and Pintelas, P. E. (2003). Preventing student dropout in distance learning using machine learning techniques. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 267–274. Springer.

[Manhães et al., 2012] Manhães, L. M. B., CRUZ, S., Costa, R. J. M., Zavaleta, J., and Zimbrão, G. (2012). Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. *Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo*.

[Manhães et al., 2011] Manhães, L. M. B., Cruz, S. d., Costa, R. J. M., Zavaleta, J., and Zimbrão, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. *Anais do XXII Simpósio Brasileiro de Informática na Educação (SBIE) - XVII Workshop de Informática na Escola (WIE), Aracaju-SE*.

[Márquez-Vera et al., 2013a] Márquez-Vera, C., Cano, A., Romero, C., and Ventura, S. (2013a). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied intelligence*, 38(3):315–330.

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

[Mendes Braga et al., 2003] Mendes Braga, M., Peixoto, M. D. C. L., and Bogutchi, T. F. (2003). A evasão no ensino superior brasileiro: O caso da ufmg. *Avaliação*, 8(3):161–189.

[Miranda et al., 2014] Miranda, E., Suryani, E., et al. (2014). Implementation of datawarehouse, datamining and dashboard for higher education. *Journal of Theoretical & Applied Information Technology*, 64(3).

[Morais, 2005] Morais, C. (2005). *Escalas de medida, estatística descritiva e inferência estatística*. Instituto Politécnico de Bragança, Escola Superior de Educação.

[Morettin and Toloi, 2006] Morettin, P. A. and Toloi, C. (2006). *Análise de séries temporais*. Blucher.

[Oliveira Júnior et al., 2015] Oliveira Júnior, J. G., Bastos, L., and Kaestner, C. (2015). Uma abordagem de data warehouse educacional para apoio à tomada de decisão. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 4.

[Oliveira Júnior et al., 2014] Oliveira Júnior, J. G., Noronha, R. V., and Kaestner, C. (2014). Correlação da evasão de cursos com o empréstimo de livros em biblioteca. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 3.

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

[Quinlan, 1993] Quinlan, J. R. (1993). *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann.

[Rigo et al., 2012] Rigo, S. J., Cazella, S. C., and Cambruzzi, W. (2012). Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In *Anais do Workshop de Desafios da Computação Aplicada à Educação*, pages 168–177.

[Ristoff, 1999] Ristoff, D. I. (1999). Considerações sobre evasão. *Ristoff, DI Universidade em foco: reflexões sobre a educação superior. Florianópolis: Insular*, pages 119–129.

[Romero et al., 2008] Romero, C., Ventura, S., Espejo, P. G., and Hervás, C. (2008). Data mining algorithms to classify students. In *The First International Conference on Educational Data Mining*, pages 8–17.

[Sachin and Vijay, 2012] Sachin, R. B. and Vijay, M. S. (2012). A survey and future vision of data mining in educational field. In *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on*, pages 96–100. IEEE.

[SESu/MEC, 1996] SESu/MEC (1996). Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. Technical report, ANDIFES/ABRUEM/SESu/MEC.

[Silva Filho et al., 2007] Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., and Lobo, M. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, 37(132):641–659.

[Tinto, 1975] Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, pages 89–125.

[Verleysen and François, 2005] Verleysen, M. and François, D. (2005). The curse of dimensionality in data mining and time series prediction. In *Computational Intelligence and Bioinspired Systems*, pages 758–770. Springer.

[Witten et al., 2011] Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

Apêndice A

Acurácia Obtida com os Classificadores

Tabela A.1: Acurácia e seu desvio padrão obtidos com o classificador J48 nos subconjuntos de atributos

Algoritmo de seleção	**Método de busca**	***Dataset1***	***Dataset2***	***Dataset3***
ChiSquaredAttributeEval	Ranking	83,59 ± 2,04	83,72 ± 2,02	85,71 ± 1,97 *
GainRatioAttributeEval	Ranking	83,67 ± 2,11	83,65 ± 1,98	85,64 ± 2,03 *
InfoGainAttributeEval	Ranking	83,59 ± 2,04	83,72 ± 2,02	85,71 ± 1,97 *
SymmetricalUncert	Ranking	83,59 ± 2,04	83,72 ± 2,02	85,75 ± 2,03 *
OneRAttributeEval	Ranking	83,74 ± 1,94	83,24 ± 2,07	85,71 ± 1,97 *
ReliefFAttributeEval	Ranking	83,17 ± 1,91	81,99 ± 2,05 *	84,52 ± 2,00 *
WrapperSubsetEval	BestFirst	83,94 ± 2,02	83,81 ± 2,04	87,15 ± 1,73
WrapperSubsetEval	GeneticSearch	84,60 ± 2,59	81,36 ± 2,02 *	87,26 ± 1,79
CfsSubsetEval	BestFirst	83,86 ± 2,08	81,82 ± 2,09 *	83,94 ± 2,02 *
CfsSubsetEval	GeneticSearch	83,65 ± 2,10	83,75 ± 1,98	83,88 ± 2,01 *

* degradação da significância estatística

Tabela A.2: Acurácia e seu desvio padrão obtidos com o classificador JRip nos subconjuntos de atributos

Algoritmos de seleção	**Método de busca**	***Dataset1***	***Dataset2***	***Dataset3***
ChiSquaredAttributeEval	Ranking	89,17 ± 5,25	88,15 ± 5,49	88,15 ± 3,63 *
GainRatioAttributeEval	Ranking	89,89 ± 5,22	88,39 ± 5,40	88,28 ± 3,72 *
InfoGainAttributeEval	Ranking	89,17 ± 5,25	88,15 ± 5,49	88,15 ± 3,63 *
OneRAttributeEval	Ranking	89,55 ± 5,44	88,60 ± 5,39	88,15 ± 3,63 *
ReliefFAttributeEval	Ranking	86,82 ± 3,74 *	85,13 ± 3,83 *	84,59 ± 2,17 *
SymmetricalUncert	Ranking	89,17 ± 5,25	88,15 ± 5,49	88,28 ± 3,72 *
WrapperSubsetEval	BestFirst	90,49 ± 5,29	89,08 ± 5,49	89,27 ± 3,91
WrapperSubsetEval	GeneticSearch	90,29 ± 5,11	88,98 ± 6,02	89,50 ± 3,60
CfsSubsetEval	BestFirst	88,17 ± 3,90	85,09 ± 3,95 *	83,99 ± 1,95 *
CfsSubsetEval	GeneticSearch	89,59 ± 5,46	87,44 ± 4,25	84,80 ± 2,01 *

* degradação da significância estatística

Tabela A.3: Acurácia e seu desvio padrão obtidos com o classificador SMO nos subconjuntos de atributos

Algoritmos de seleção	**Método de busca**	***Dataset1***	***Dataset2***	***Dataset3***
ChiSquaredAttributeEval	Ranking	87,47 ± 1,81 *	83,20 ± 1,92 *	81,91 ± 1,90 *
GainRatioAttributeEval	Ranking	87,05 ± 1,83 *	82,96 ± 1,94 *	81,89 ± 1,91 *
InfoGainAttributeEval	Ranking	87,47 ± 1,81 *	83,20 ± 1,92 *	81,91 ± 1,90 *
OneRAttributeEval	Ranking	88,83 ± 1,62	83,54 ± 1,95 *	81,91 ± 1,90 *
ReliefFAttributeEval	Ranking	86,89 ± 1,84 *	81,59 ± 1,87 *	81,19 ± 1,88 *
SymmetricalUncert	Ranking	87,47 ± 1,81 *	83,20 ± 1,92 *	81,89 ± 1,91 *
WrapperSubsetEval	BestFirst	87,58 ± 2,05 *	86,77 ± 1,81	83,36 ± 2,22
WrapperSubsetEval	GeneticSearch	89,26 ± 1,60	86,83 ± 1,73	82,87 ± 1,87
CfsSubsetEval	BestFirst	83,13 ± 2,01 *	80,43 ± 1,99 *	81,19 ± 1,87 *
CfsSubsetEval	GeneticSearch	88,32 ± 1,63 *	82,50 ± 1,99 *	81,19 ± 1,87 *

* degradação da significância estatística

Tabela A.4: Acurácia e seu desvio padrão obtidos com o classificador MLP nos subconjuntos de atributos

Algoritmos de seleção de atributos	**Método de busca**	***Dataset1***	***Dataset2***	***Dataset3***
ChiSquaredAttributeEval	Ranking	91,54 ± 1,64 *	88,78 ± 1,96 *	90,44 ± 1,59 *
GainRatioAttributeEval	Ranking	93,03 ± 1,47 *	90,75 ± 1,64 *	90,67 ± 1,42 *
InfoGainAttributeEval	Ranking	91,54 ± 1,64 *	88,78 ± 1,96 *	90,44 ± 1,59 *
OneRAttributeEval	Ranking	90,70 ± 1,72 *	87,80 ± 1,80 *	90,44 ± 1,59 *
ReliefFAttributeEval	Ranking	89,07 ± 1,86 *	86,03 ± 1,73 *	81,79 ± 1,83 *
SymmetricalUncert	Ranking	91,54 ± 1,64 *	88,78 ± 1,96 *	90,67 ± 1,42 *
WrapperSubsetEval	BestFirst	94,75 ± 1,26	92,55 ± 1,42	92,73 ± 1,28
WrapperSubsetEval	GeneticSearch	93,24 ± 1,32 *	90,69 ± 1,61 *	92,46 ± 1,29
CfsSubsetEval	BestFirst	90,82 ± 1,48 *	89,28 ± 1,77 *	82,94 ± 1,87 *
CfsSubsetEval	GeneticSearch	92,54 ± 1,51 *	84,73 ± 2,81 *	86,45 ± 1,83 *

* degradação da significância estatística

Tabela A.5: Acurácia e seu desvio padrão obtidos com o classificador *RandomForest* nos subconjuntos de atributos

Algoritmo de seleção	**Método de busca**	***Dataset1***	***Dataset2***	***Dataset3***
ChiSquaredAttributeEval	Ranking	76,19 ± 2,17 *	83,34 ± 1,92 *	85,70 ± 1,66 *
GainRatioAttributeEval	Ranking	80,96 ± 1,77 *	84,89 ± 2,05 *	85,71 ± 1,78 *
InfoGainAttributeEval	Ranking	76,19 ± 2,17 *	83,34 ± 1,92 *	85,70 ± 1,66 *
OneRAttributeEval	Ranking	69,96 ± 2,18 *	82,20 ± 1,90 *	85,70 ± 1,66 *
ReliefFAttributeEval	Ranking	70,67 ± 2,21 *	80,41 ± 1,90 *	82,84 ± 1,96 *
SymmetricalUncert	Ranking	76,19 ± 2,17 *	83,34 ± 1,92 *	85,71 ± 1,78 *
WrapperSubsetEval	BestFirst	87,40 ± 1,60	88,13 ± 1,78	84,47 ± 2,03 *
WrapperSubsetEval	GeneticSearch	87,35 ± 1,52	87,31 ± 1,92	87,30 ± 1,79
CfsSubsetEval	BestFirst	84,65 ± 1,91 *	81,92 ± 1,98 *	80,62 ± 1,69 *
CfsSubsetEval	GeneticSearch	79,30 ± 1,93 *	84,15 ± 1,79 *	83,76 ± 1,88 *

* degradação da significância estatística

Tabela A.6: Acurácia e seu desvio padrão obtidos com o classificador IBk nos subconjuntos de atributos

Algoritmo de seleção	**Método de busca**	***Dataset1***	***Dataset2***	***Dataset3***
ChiSquaredAttributeEval	Ranking	78,05 ± 2,17 *	73,87 ± 2,42 *	71,77 ± 2,41 *
GainRatioAttributeEval	Ranking	82,04 ± 2,22 *	66,69 ± 4,99 *	75,52 ± 2,10 *
InfoGainAttributeEval	Ranking	78,05 ± 2,17 *	73,87 ± 2,42 *	71,77 ± 2,41 *
OneRAttributeEval	Ranking	79,46 ± 2,37 *	73,68 ± 2,85 *	71,77 ± 2,41 *
ReliefFAttributeEval	Ranking	79,88 ± 2,40 *	77,07 ± 1,79 *	74,13 ± 2,11 *
SymmetricalUncert	Ranking	78,05 ± 2,17 *	73,87 ± 2,42 *	75,52 ± 2,10 *
WrapperSubsetEval	BestFirst	86,76 ± 2,05 *	86,06 ± 1,79	90,28 ± 1,47
WrapperSubsetEval	GeneticSearch	84,90 ± 2,25 *	85,76 ± 1,85	88,36 ± 1,61 *
CfsSubsetEval	BestFirst	90,32 ± 1,35	80,65 ± 2,12 *	70,46 ± 2,49 *
CfsSubsetEval	GeneticSearch	82,66 ± 2,03 *	72,57 ± 3,34 *	75,67 ± 2,92 *

* degradação da significância estatística

Apêndice B

Classificação dos Melhores Atributos

Tabela B.1: Classificação dos melhores atributos utilizando o classificador J48

Classificação	Nº	Atributo	Atributo criado	Nº de vezes selecionado**	Posição média*
1º	12	regressao_coeficiente	Sim	8	8
2º	8	previsao_evasao_dificuldade	Sim	7	2
3º	14	coeficiente_rendimento		6	4
4º	15	percentual_aprov	Sim	5	3
5º	10	total_semestres_trancados	Sim	5	20
6º	11	emprestimos_biblioteca_por_semestre	Sim	3	8
7º	27	socio_reside_em		3	13
8º	1	grau		3	15
9º	32	socio_escolaridade_mae		2	7
10º	29	socio_necessidade_trabalhar		2	11

* posição média nos algoritmos de seleção de atributos por ranqueamento
** frequência que o atributo foi selecionado nos algoritmos Wrapper e Cfs

Tabela B.2: Classificação dos melhores atributos utilizando o classificador JRip

Classificação	Nº	Atributo	Atributo criado	Nº de vezes selecionado**	Posição média*
1º	12	regressao_coeficiente	Sim	9	6
2º	8	previsao_evasao_dificuldade	Sim	8	2
3º	10	total_semestres_trancados	Sim	8	20
4º	15	percentual_aprov	Sim	6	3
5º	14	coeficiente_rendimento		5	3
6º	13	percentual_frequencia	Sim	5	4
7º	7	cota		5	22
8º	27	socio_reside_em		4	13
9º	1	grau		4	17
10º	5	reentrada_mesmo_curso	Sim	4	33

* posição média nos algoritmos de seleção de atributos por ranqueamento
** frequência que o atributo foi selecionado nos algoritmos Wrapper e Cfs

Tabela B.3: Classificação dos melhores atributos utilizando o classificador SMO

Classificação	**Nº**	**Atributo**	**Atributo criado**	**Nº de vezes selecionado****	**Posição média***
1º	12	regressao_coeficiente	Sim	5	1
2º	13	percentual_frequencia	Sim	5	2
3º	10	total_semestres_trancados	Sim	5	28
4º	31	socio_esc_pai		3	6
5º	2	genero		3	23
6º	16	nota_final_enem		3	29
7º	32	socio_esc_mae		2	3
8º	8	previsao_evasao_dificuldade	Sim	2	4
9º	15	percentual_aprov	Sim	2	7
10º	14	coeficiente_rendimento		2	8

* posição média nos algoritmos de seleção de atributos por ranqueamento

** frequência que o atributo foi selecionado nos algoritmos Wrapper e Cfs

Tabela B.4: Classificação dos melhores atributos utilizando o classificador MLP

Classificação	**Nº**	**Atributo**	**Atributo criado**	**Nº de vezes selecionado****	**Posição média***
1º	10	total_semestres_trancados	Sim	4	-
2º	12	regressao_coeficiente	Sim	4	-
3º	13	percentual_frequencia	Sim	4	-
4º	14	coeficiente_rendimento		2	-
5º	16	nota_final_enem		2	-
6º	17	nota_linguagem		2	-
7º	19	nota_natureza		2	-
8º	3	estado_civil		1	-
9º	5	reentrada_mesmo_curso	Sim	1	-
10º	6	mudou_de_curso	Sim	1	-

* posição média nos algoritmos de seleção de atributos por ranqueamento

** frequência que o atributo foi selecionado nos algoritmos Wrapper e Cfs

Tabela B.5: Classificação dos melhores atributos utilizando o classificador *Random Forest*

Classificação	**Nº**	**Atributo**	**Atributo criado**	**Nº de vezes selecionado****	**Posição média***
1º	15	percentual_aprov	Sim	5	-
2º	14	coeficiente_rendimento		5	-
3º	13	percentual_frequencia	Sim	5	-
4º	12	regressao_coeficiente	Sim	5	-
5º	10	total_semestres_trancados	Sim	5	-
6º	8	previsao_evasao_dificuldade	Sim	5	-
7º	1	grau		5	-
8º	22	micro_regiao_origem	Sim	4	-
9º	19	nota_natureza		3	-
10º	16	nota_final_enem		3	-

* posição média nos algoritmos de seleção de atributos por ranqueamento

** frequência que o atributo foi selecionado nos algoritmos Wrapper e Cfs

Tabela B.6: Classificação dos melhores atributos utilizando o classificador IBk

Classificação	**Nº**	**Atributo**	**Atributo criado**	**Nº de vezes selecionado****	**Posição média***
1º	10	total_semestres_trancados	Sim	4	-
2º	13	percentual_frequencia	Sim	3	-
3º	12	regressao_coeficiente	Sim	2	-
4º	8	previsao_evasao_dificuldade	Sim	2	-
5º	1	grau		1	-
6º	23	meso_regiao_origem	Sim	1	-
7º	24	regiao_origem	Sim	1	-

* posição média nos algoritmos de seleção de atributos por ranqueamento

** frequência que o atributo foi selecionado nos algoritmos Wrapper e Cfs

Apêndice C

Gráficos do Atributo "Dificuldade Média das Disciplinas Cursadas"

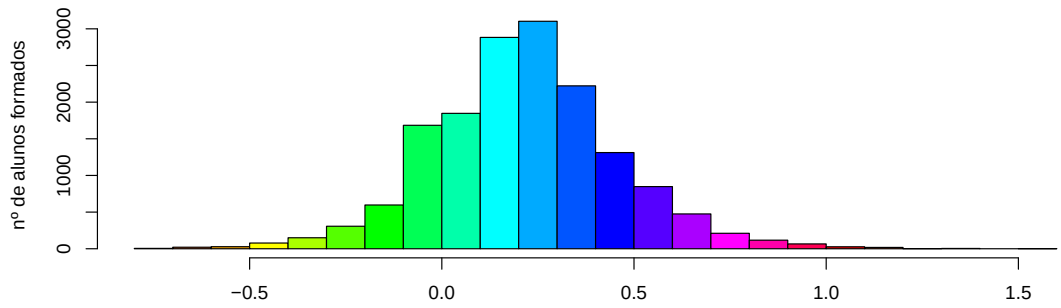


Figura C.1: Histograma da dificuldade média das disciplinas cursadas até o 3º período dos alunos formados entre 1983 e 2014

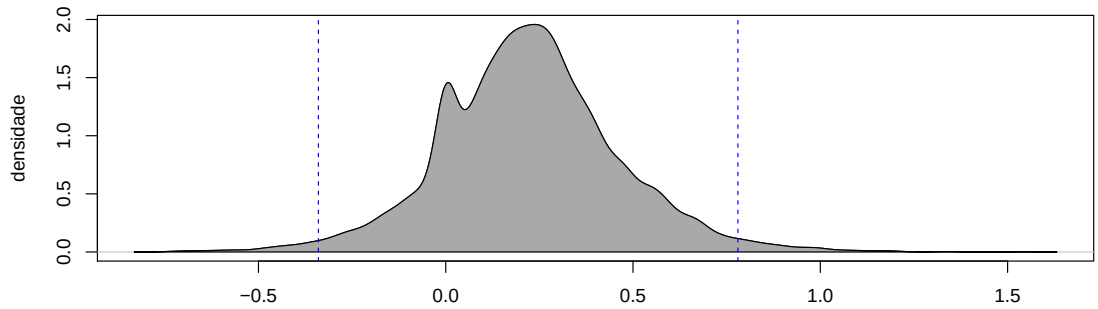


Figura C.2: Densidade da dificuldade média das disciplinas cursadas até o 3º período dos alunos formados entre 1983 e 2014

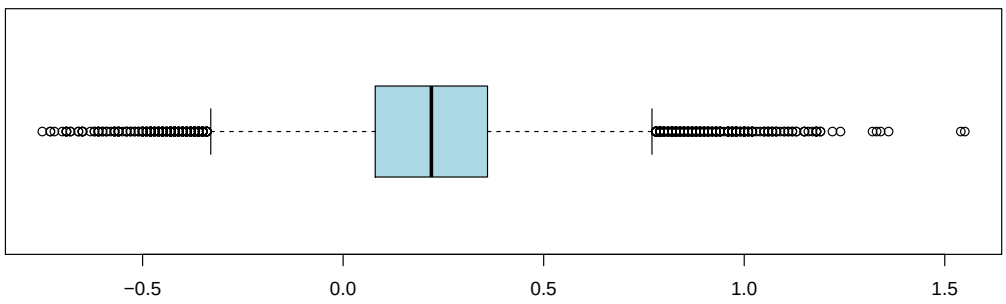


Figura C.3: Boxplot da dificuldade média das disciplinas cursadas até o 3º período dos alunos formados entre 1983 e 2014

Apêndice D

Diagramas dos Atributos Antes do Balanceamento das Classes

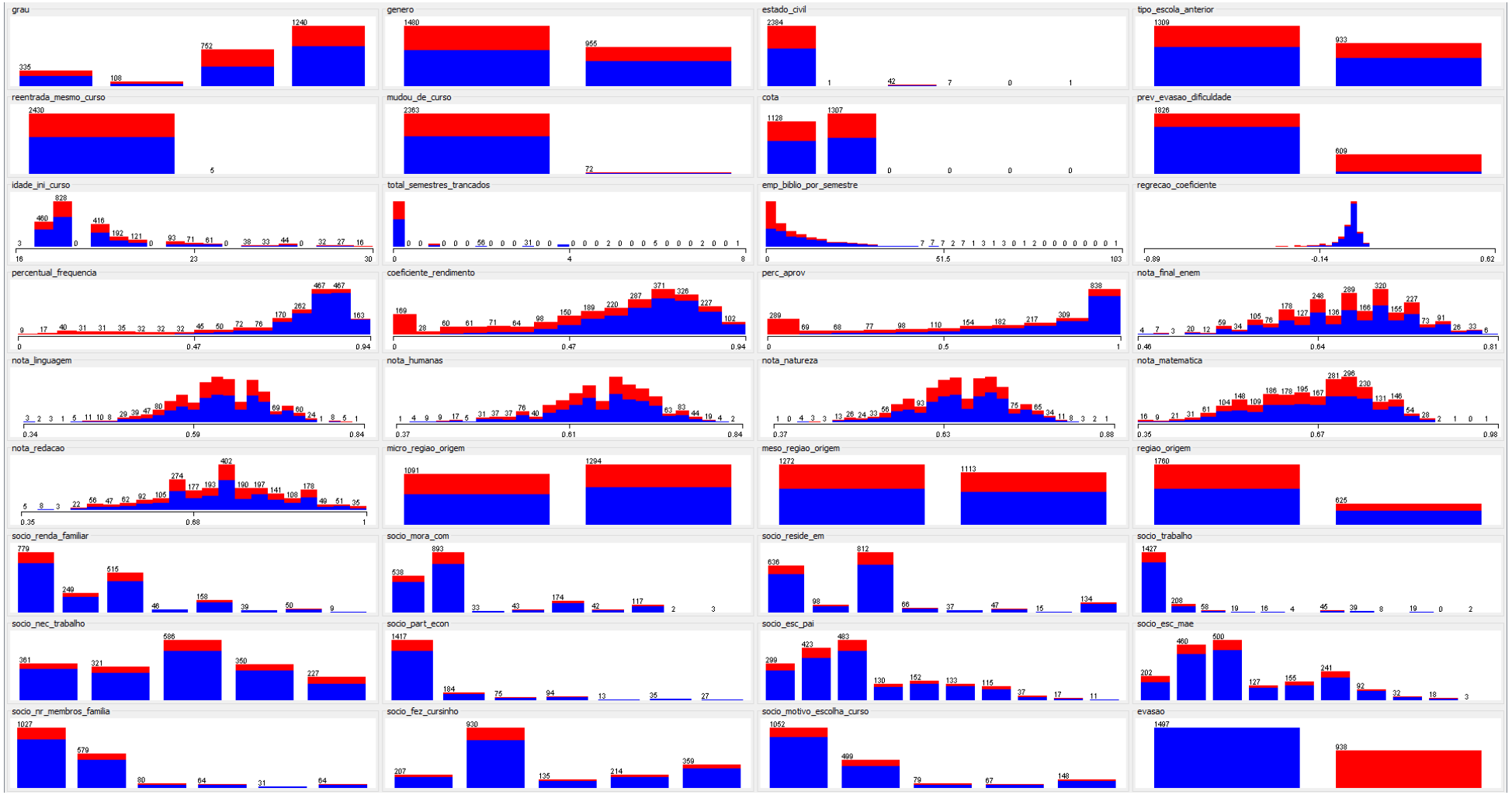


Figura D.1: Diagrama dos atributos do *dataset1* antes do balanceamento das classes

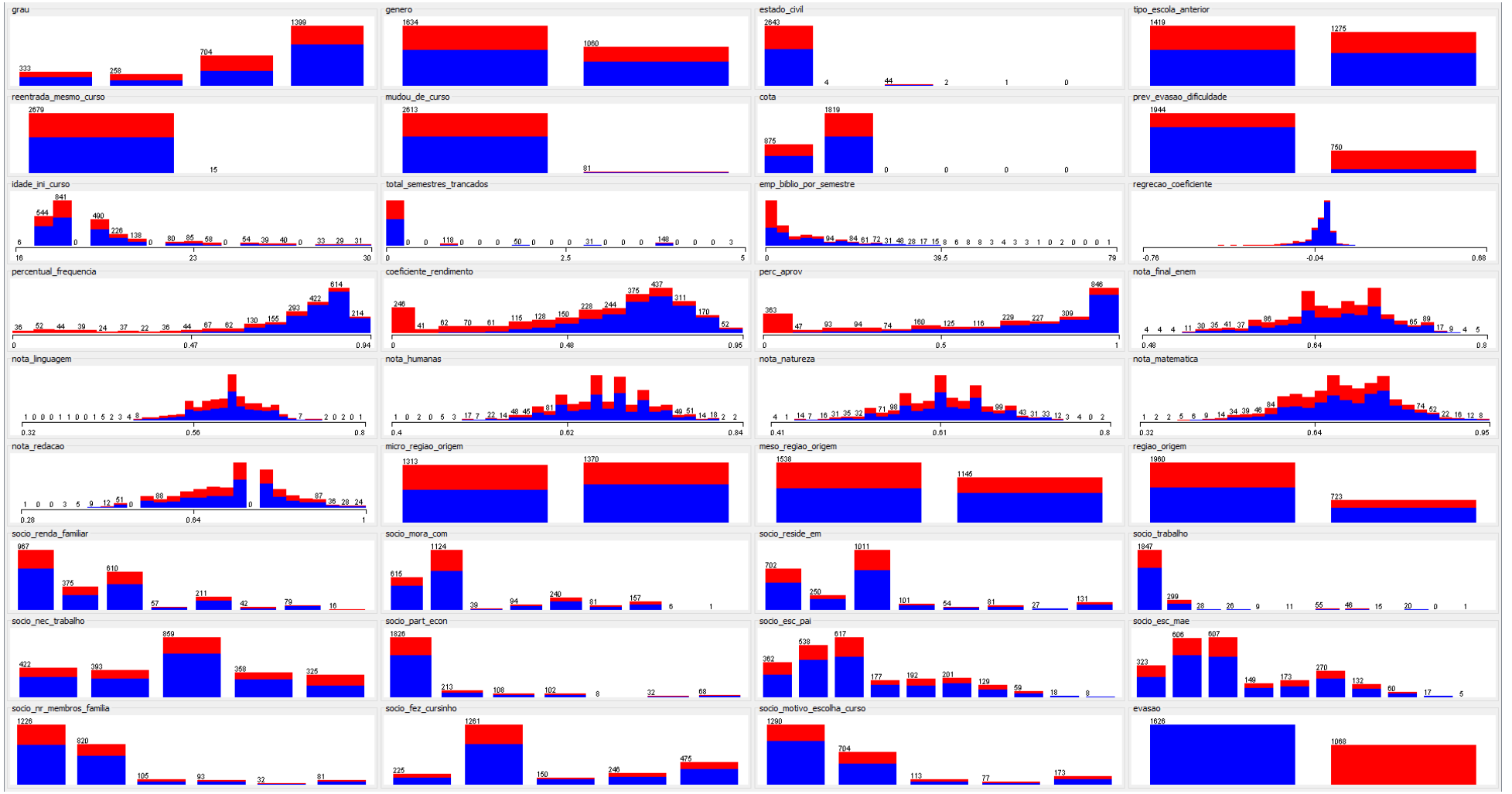


Figura D.2: Diagrama dos atributos do *dataset2* antes do balanceamento das classes

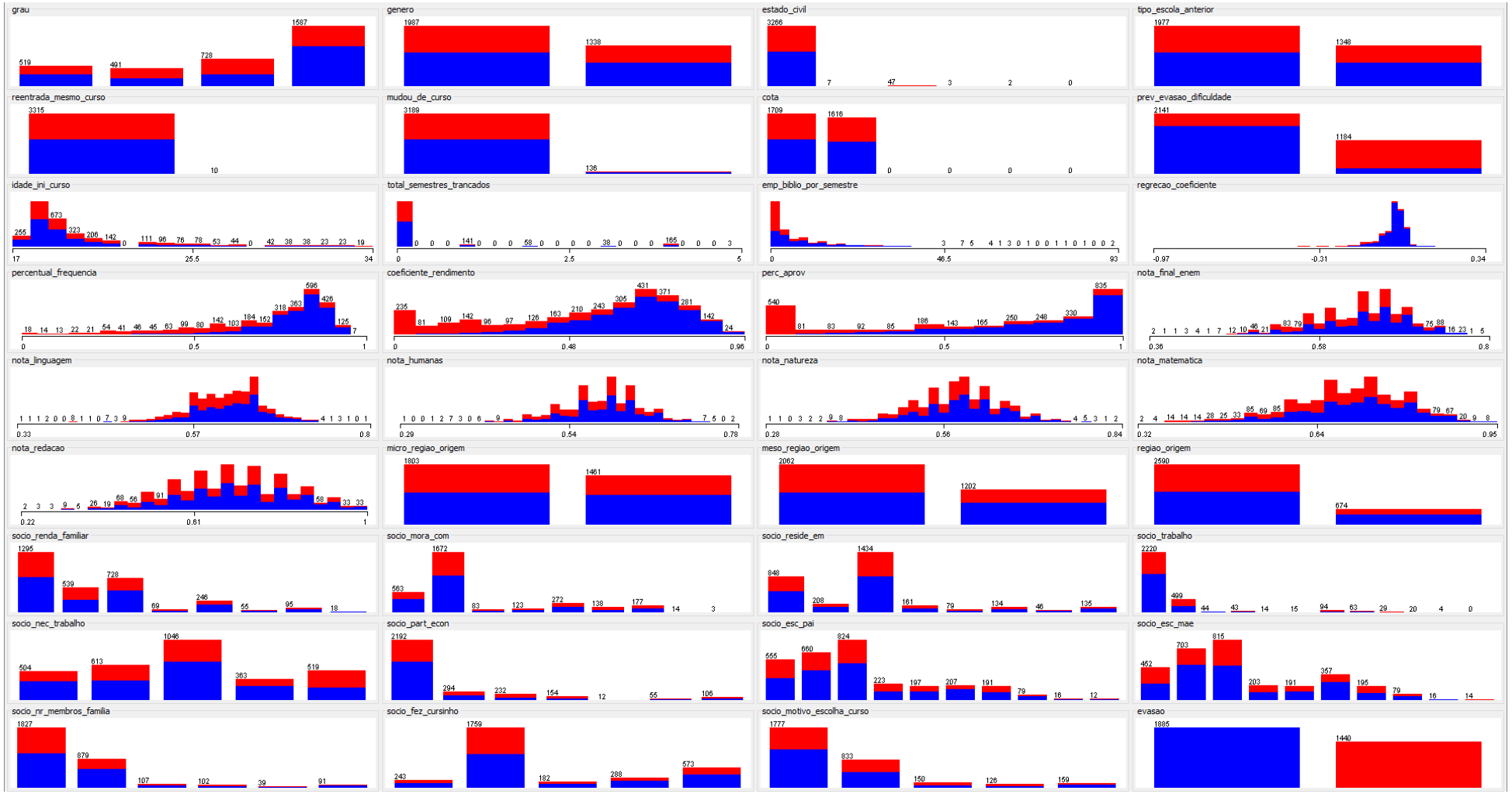


Figura D.3: Diagrama dos atributos do *dataset3* antes do balanceamento das classes