



PREDIÇÃO DO DESEMPENHO ACADÊMICO DE GRADUANDOS UTILIZANDO MINERAÇÃO DE DADOS EDUCACIONAIS

Laci Mary Barbosa Manhães

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutor em Engenharia de Sistemas e Computação.

Orientador(es): Geraldo Zimbrão da Silva
Sérgio Manuel Serra da Cruz

Rio de Janeiro
Fevereiro de 2015

PREDIÇÃO DO DESEMPENHO ACADÊMICO DE GRADUANDOS UTILIZANDO
MINERAÇÃO DE DADOS EDUCACIONAIS

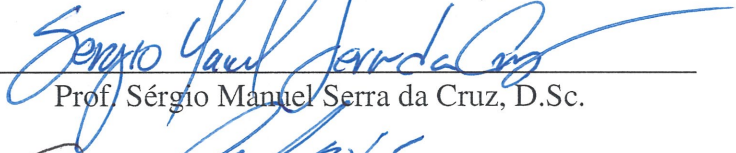
Laci Mary Barbosa Manhães

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ
COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS
REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTOR EM
CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:



Prof. Geraldo Zimbrão da Silva, D.Sc.



Prof. Sérgio Manuel Serra da Cruz, D.Sc.



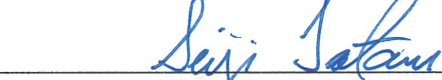
Prof. Geraldo Bonorino Xexéo, D.Sc.



Prof. Nelson Francisco Favilla Ebecken, D.Sc.



Prof. Josefino Cabral Melo Lima, Ph.D.



Prof. Seiji Isotani, Ph.D.

RIO DE JANEIRO, RJ - BRASIL

FEVEREIRO DE 2015

Manhães, Laci Mary Barbosa

Predição do Desempenho Acadêmico de Graduandos
Utilizando Mineração de Dados Educacionais / Laci Mary
Barbosa Manhães. – Rio de Janeiro: UFRJ/COPPE, 2015.

XVII, 140 p.: il.; 29,7 cm.

Orientadores: Geraldo Zimbrão da Silva

Sérgio Manuel Serra da Cruz

Tese (doutorado) – UFRJ/ COPPE/ Programa de
Engenharia de Sistemas e Computação, 2015.

Referências Bibliográficas: p. 130-139.

1. Mineração de dados Educacionais. 2. Arquitetura. 3.
Desempenho Acadêmico. 4. Experimentos. I. Silva,
Geraldo Zimbrão da *et al.* II. Universidade Federal do Rio
de Janeiro, COPPE, Programa de Engenharia de Sistemas
e Computação. III. Título.

Dedico este trabalho ao meu pai da terra e ao meu Pai do Céu.

AGRADECIMENTOS

Agradeço a Deus, Ele quis esta tese muito mais do que eu. Agradeço ao meu pai que sempre e incondicionalmente me deu suporte para eu ser o que sou e chegar aonde cheguei. A minha família e amigos pela torcida e por suportar minha ausência.

Aos professores Geraldo Zimbrão e Sérgio Serra pela orientação e incentivo ao longo destes anos. A todos os professores que ao longo dos anos foram me ajudando a construir o conhecimento, especialmente aqueles que eu encontrei na UFRJ.

Agradeço ao professor Erickson Almendra Rocha, ex-diretor da Escola Politécnica da UFRJ, e a Roberto Vieira, diretor da DRE/UFRJ, pelo auxílio em informações relevantes na execução deste trabalho.

Ao amigo Sérgio Serra por ter sido um anjo bom que Deus colocou no meu caminho.

Aos amigos Macário Costa, Jorge Zavaleta e muitos outros que torceram por mim, eu agradeço a amizade, companheirismo e espírito de equipe.

Treino a minha mente na aquisição do conhecimento para contemplar a
Verdadeira Sabedoria.

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutor em Ciências (D.Sc.)

PREDIÇÃO DO DESEMPENHO ACADÊMICO DE GRADUANDOS UTILIZANDO MINERAÇÃO DE DADOS EDUCACIONAIS

Laci Mary Barbosa Manhães

Fevereiro/2015

Orientadores: Geraldo Zimbrão da Silva

Sérgio Manuel Serra da Cruz

Programa: Engenharia de Sistemas e Computação

Este trabalho apresenta uma proposta de arquitetura baseada em Mineração de Dados Educacionais (EDM) para predição do desempenho acadêmico de graduandos. O objetivo deste trabalho é fornecer aos gestores educacionais das universidades públicas brasileiras, não especialista em EDM, uma abordagem que oferece informações úteis sobre o desempenho acadêmico dos graduandos e predizer os que estão em risco de abandonar o sistema de ensino. A arquitetura EDM WAVE engloba todo o processo de descoberta de conhecimento em dados (pré-processamento, mineração de dados e pós-processamento). A arquitetura e os modelos propostos foram testados através de estudos experimentais que utilizaram dados do mundo real de graduandos da Universidade Federal do Rio de Janeiro (UFRJ), durante um período de 16 anos.

Nossa abordagem é uma das primeiras que utiliza apenas dados acadêmicos que variam no tempo, armazenados no sistema de gestão acadêmica, nenhum dado social ou econômico é considerado nas análises. Os resultados experimentais mostram que a arquitetura proposta é capaz de predizer o desempenho acadêmico dos graduandos a cada semestre letivo com precisão em torno de 80%. Além da predição, também foi possível identificar as principais variáveis que distinguem os estudantes que obtêm sucesso ou não na conclusão do curso de graduação.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Doctor of Science (D.Sc.)

PREDICTING ACADEMIC PERFORMANCE OF UNDERGRADUATE STUDENTS
USING EDUCATIONAL DATA MINING

Laci Mary Barbosa Manhães

February/2015

Advisors: Geraldo Zimbrão da Silva

Sérgio Manuel Serra da Cruz

Department: Computer Science and Engineering

This thesis presents an architecture based on Educational Data Mining (EDM) for the prediction of academic performance of undergraduate students. The objective of this work is to provide educational managers of Brazilian public universities, non-specialist in EDM, an approach that offers useful information about the academic performance of the students and predicts those students that are at risk of leaving the education system. The architecture encompasses the process of Knowledge Discovery from Data (pre-processing, data mining and post-processing). The EDM WAVE architecture and the data models were developed and tested through experimental studies using real-world data of students from Federal University of Rio de Janeiro (UFRJ), for a period of 16 years.

Our approach is one of the first to use only time-varying academic data, stored in the academic management system, no social or economic data is considered in the analyzes. The experimental results show that the architecture is able to predict the academic performance of the students every semester our results present 80% of corrected. In addition to the prediction, it was also possible to identify the main features that distinguish students who succeed or not in the completion of the undergraduate degree course.

Sumário

1	Introdução	1
1.1	Definição do Problema	2
1.2	Objetivos da Tese	5
1.3	Contribuição.....	6
1.4	Metodologia da Pesquisa	8
1.5	Organização da Tese	10
2	Caracterização do Problema e Trabalhos Correlatos	11
2.1	Definição dos Termos Utilizados.....	11
2.2	Análise e Contextualização do Problema	14
2.2.1	Abordagem do Problema sob a Perspectiva do Estudante.....	16
2.2.2	Abordagem do Problema sob a Perspectiva da Instituição.....	19
2.2.3	Abordagem do Problema sob a Perspectiva da Sociedade e do País.....	21
2.3	Contextualização da Trabalho de Tese.....	24
2.4	Trabalhos Correlatos em Diversos Níveis de Aplicação.....	26
2.5	Trabalhos Correlatos em EDM	27
2.5.1	Trabalhos Direcionados a Identificar os Atributos Relevantes para a Caracterização dos Estudantes	29
2.5.2	Trabalhos Direcionados a Identificar e Comparar o Desempenho dos Algoritmos.....	30
2.5.3	Trabalhos Relacionados Utilizando Métodos Estatísticos e/ou Outras Análises.....	33
2.6	Repositórios de Base de Dados Educacionais	34
2.7	Conclusões	35
3	WAVE: Uma Arquitetura Apoiada em EDM para IFES	39
3.1	Descoberta de Conhecimento em Dados	39
3.1.1	Bases de Dados	41
3.1.2	Pré-processamento dos Dados	41
3.1.3	Mineração de Dados e EDM	42
3.1.4	Pós-processamento dos Dados	43
3.2	Funcionalidades da Mineração de Dados	43

3.3	Processo de Construção de um Modelo	46
3.4	Arquitetura	48
3.4.1	Camada de Dados	49
3.4.2	Camada de Aplicação	50
3.4.3	Camada de Apresentação	54
3.4.4	Funcionalidades da Arquitetura	56
3.5	Conclusões	57
4	Experimentos, Testes e Avaliação Crítica	59
4.1	Contextualização do Problema na Graduação da UFRJ	59
4.2	Base de Dados da UFRJ	60
4.2.1	Descrição dos Atributos Originais Extraídos da Base de Dados do SIGA62	
4.3	Definição do Modelo de Dados dos Graduandos	65
4.4	Definição dos Algoritmos Utilizados nos Experimentos	68
4.5	Particionamento da Base de Dados	70
4.6	Ferramentas para o Processo ETL	70
4.7	Ferramentas de Mineração de Dados	71
4.7.1	Weka Explorer (WE)	71
4.7.2	Weka Experiment Environment (WEE)	72
4.7.3	Descrição dos Arquivos Utilizados pelo Weka	73
4.8	Estudo de Caso 01: Avaliação de 12 Algoritmos Classificadores Utilizando Dados do Curso de Engenharia Civil e suas Ênfases	76
4.8.1	Descrição dos Algoritmos Utilizados nos Experimentos	76
4.8.2	Descrição da Base de Dados dos Experimentos	77
4.8.3	Descrição dos Experimentos e Avaliação dos Resultados	79
4.9	Estudo de Caso 02: Uma Abordagem Quantitativa dos Fatores que Influenciam o Desempenho Acadêmico dos Estudantes de Graduação da UFRJ	86
4.9.1	Descrição dos Algoritmos Utilizados no Experimentos	87
4.9.2	Descrição da Base de Dados Utilizada nos Experimentos	87
4.9.3	Descrição dos Experimentos e Avaliação dos Resultados	88
4.10	Estudo de Caso 03: Curso de Engenharia Civil e suas Ênfases Predição do Desempenho Acadêmico até o Quinto Semestre Letivo	98
4.10.1	Descrição dos algoritmos Utilizados nos Experimentos	99
4.10.2	Descrição da Base de Dados Utilizada nos Experimentos	99
4.10.3	Descrição dos Experimentos e Avaliação dos Resultados	99

4.11	Estudo de Caso 04: Estudantes do Curso de Engenharia Produção e suas Ênfases Predição do Desempenho Acadêmico até o Quinto Semestre Letivo	105
4.11.1	Descrição dos Algoritmos Utilizados nos Experimentos	105
4.11.2	Descrição da Base de Dados.....	105
4.11.3	Descrição dos Experimentos e Avaliação dos Resultados.....	105
4.12	Estudo de Caso 05: Estudantes do Curso de Engenharia Mecânica e suas Ênfases Predição do Desempenho Acadêmico até o Quinto Semestre Letivo	108
4.12.1	Descrição dos Algoritmos Utilizados nos Experimentos	108
4.12.2	Descrição da Base de Dados.....	108
4.12.3	Descrição dos Experimentos e Avaliação dos Resultados.....	109
4.13	Estudo de Caso 06: Avaliação da Arquitetura EDM WAVE	111
4.13.1	Descrição dos Algoritmos Utilizados nos Experimentos	112
4.13.2	Descrição da Base de Dados.....	112
4.13.3	Modelo de Dados dos Estudantes.....	112
4.13.4	Descrição do Experimento e Avaliação dos Resultados	114
4.14	Estudo de Caso 07: Análise de 6 cursos de graduação da UFRJ	117
4.14.1	Descrição dos Algoritmos Utilizados nos Experimentos	117
4.14.2	Descrição da Base de Dados.....	118
4.14.3	Definição do Modelo de Dados dos Estudantes	119
4.14.4	Descrição dos Experimentos e Avaliação dos Resultados.....	119
4.15	Geração dos Modelos para Visualização da Mineração de Dados.....	123
4.16	Conclusão.....	124
5	Capítulo: Conclusões.....	126
5.1	Trabalhos Futuros	128
6	Referências Bibliográficas.....	130
7	Apêndice.....	140

Lista de Tabelas

Tabela 2.1: Estudos e abordagens dos trabalhos relacionados.	37
Tabela 3.1: <i>Layout</i> do relatório com a predição dos classificadores para n estudantes..	55
Tabela 4.1: Lista de atributos originais da base de dados do SIGA.	62
Tabela 4.2: Exemplo de dados acadêmicos de um estudante de graduação obtidos a partir do SIGA.	64
Tabela 4.3: Modelo de dados dos estudantes de graduação.	67
Tabela 4.4: Identificação e breve descrição dos classificadores.....	69
Tabela 4.5: Estrutura do arquivo Weka (.arff) para o conjunto de treinamento.....	73
Tabela 4.6: Estrutura do arquivo Weka (.arff) do conjunto de teste.	74
Tabela 4.7: Estrutura do arquivo Weka (.arff) com o resultado da predição da classe... 75	
Tabela 4.8: Modelo de dados dos estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando dados do primeiro semestre letivo. 78	
Tabela 4.9: Análise da importância dos atributos para classificação segundo o método de Ganho da Informação 78	
Tabela 4.10: Resultados dos classificadores para a predição de duas classes de estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando validação cruzada com 10 conjuntos para dados do primeiro semestre letivo. 79	
Tabela 4.11: Resultados dos classificadores para a predição de duas classes de estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando seleção randômica do conjunto de treinamento e teste para dados de estudantes utilizando dados do primeiro semestre letivo. 80	
Tabela 4.12: Resultados dos classificadores para a predição de duas classes de estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando validação cruzada com 10 conjuntos para dados do primeiro semestre letivo. 81	
Tabela 4.13: Resultados dos classificadores para a predição de duas classes de estudantes especificando o conjunto de treinamento e teste para estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando dados do primeiro semestre letivo. 82	

Tabela 4.14: Acurácia, média das acurácias e desvio padrão dos Experimentos.	83
Tabela 4.15: Quantidade de estudantes distribuídos nas três classes por ano de ingresso.	87
Tabela 4.16: Análise do desempenho dos classificadores segundo critérios quantitativos para estudantes ingressaram 2003-1.	89
Tabela 4.17: Análise do desempenho do classificador <i>Naive Bayes</i> para estudantes ingressaram 2003-1, 2003-2, 2004-1 e 2004-2.	95
Tabela 4.18: Arquivo com os atributos do Modelo de Dados dos Estudantes para predição do desempenho acadêmico no segundo semestre letivo.	100
Tabela 4.19: Resultados dos classificadores para a predição do segundo semestre letivo dos estudantes da Engenharia Civil ano ingresso 2003-1.	101
Tabela 4.20: Arquivo com os atributos do Modelo de Dados dos Estudantes para predição do desempenho acadêmico no terceiro semestre letivo.	101
Tabela 4.21: Resultados dos classificadores para a predição do terceiro semestre letivo dos estudantes da Engenharia Civil ano ingresso 2003-1.	102
Tabela 4.22: Arquivo com os atributos do modelo de dados dos estudantes para predição do desempenho acadêmico no quarto semestre letivo.	102
Tabela 4.23: Resultados dos classificadores para a predição do quarto semestre letivo dos estudantes da Engenharia Civil ano ingresso 2003-1.	103
Tabela 4.24: Arquivo com os atributos do Modelo de Dados dos Estudantes para predição do desempenho acadêmico no quinto semestre letivo.	104
Tabela 4.25: Resultados dos classificadores para a predição do quinto semestre letivo dos estudantes da Engenharia Civil ano ingresso 2003-1.	104
Tabela 4.26: Resultados dos classificadores para a predição do segundo semestre letivo dos estudantes da Engenharia Produção ano ingresso 2005-1.	106
Tabela 4.27: Resultados dos classificadores para a predição do terceiro semestre letivo dos estudantes da Engenharia Produção ano ingresso 2005-1.	107
Tabela 4.28: Resultados dos classificadores para a predição do quarto semestre letivo dos estudantes da Engenharia Produção ano ingresso 2005-1.	107
Tabela 4.29: Resultados dos classificadores para a predição do quinto semestre letivo dos estudantes da Engenharia Produção ano ingresso 2005-1.	108
Tabela 4.30: Resultados dos classificadores para a predição do segundo semestre letivo dos estudantes da Engenharia Mecânica ano ingresso 2007-1	109
Tabela 4.31: Resultados dos classificadores para a predição do terceiro semestre letivo	

dos estudantes da Engenharia Mecânica ano ingresso 2007-1.	110
Tabela 4.32: Resultados dos classificadores para a predição do quarto semestre letivo dos estudantes da Engenharia Mecânica ano ingresso 2007-1.	110
Tabela 4.33: Resultados dos classificadores para a predição do quinto semestre letivo dos estudantes da Engenharia Mecânica ano ingresso 2007-1.	111
Tabela 4.34: Quantidade de estudantes distribuídos nas duas classes por ano de ingresso.	112
Tabela 4.35: Modelo de dados dos estudantes de graduação predição para o segundo semestre letivo.	113
Tabela 4.36: Taxas de acerto e erro dos classificadores, VP, FN, VN, FP e MC para o curso de Engenharia Civil.	115
Tabela 4.37: Taxas de acerto e erro dos classificadores, VP, FN, VN, FP e MC para o curso de Engenharia Mecânica.	115
Tabela 4.38: Taxas de acerto e erro dos classificadores, VP, FN, VN, FP e MC para o curso de Engenharia de Produção.	116
Tabela 4.39: Número de estudantes no conjunto de treinamento distribuídos em duas classes.	118
Tabela 4.40: Número de estudantes para os conjuntos de testes para cada curso de graduação e por ano/semestre de ingresso.	118
Tabela 4.41: Porcentagem de estudantes em cada classe nos conjuntos de teste.	119
Tabela 4.42: Valores do <i>Kappa</i>	121
Tabela 4.43: Média e desvio padrão das taxas de acerto para a classe não-progresso.	122
Tabela 4.44: <i>Layout</i> do relatório com a predição de um grupo de n estudantes graduação.	124
Tabela 7.1: Cursos de graduação em Engenharia da Escola Politécnica UFRJ.	140

Lista de Figuras

Figura 3.1: Processo de Descoberta de Conhecimento em Dados (KDD) utilizando EDM e as fases de desenvolvimento metodológico adotada nesta tese - adaptação (HAN, KAMBER, 2006).....	40
Figura 3.2: Síntese do Descoberta de Conhecimento em Dados.	41
Figura 3.3: Esquema de modelo preditivo de dados.....	46
Figura 3.4: Esquema de construção, análise e validação do modelo preditivo de dados.	47
Figura 3.5: Arquitetura EDM WAVE baseada em três camadas.	49
Figura 3.6: Repositório de Conhecimento da arquitetura EDM WAVE.....	53
Figura 3.7: Esquema de execução de um algoritmo na arquitetura EDM WAVE.....	54
Figura 4.1: O gráfico ilustra as acurácias dos classificadores obtidos nos Experimentos.	84
Figura 4.2: Da esquerda para direita temos os gráficos: (a) número de disciplinas cursadas; (b) número de disciplinas aprovadas; (c) número de disciplinas RM; e (d) número de disciplinas RFM.....	91
Figura 4.3: De cima para baixo temos os gráficos que apresentam média das disciplinas aprovadas, CR do período e CRA dos estudantes: (a) Cancelados, (b) AFP e (c) Concluintes.....	92
Figura 4.4: Da esquerda para a direita, temos os gráficos que mostram o número de disciplinas cursadas para os estudantes: (a) cancelados, (b) AFP e (c) concluintes. Os gráficos que apresentam o número de disciplinas aprovadas para: (d) cancelados, (e) ativos e (f) concluintes. Os gráficos com o número de disciplinas RM para estudantes: (g) cancelados, (h) ativos e (i) concluintes. Os gráficos com o número de disciplinas RFM para estudantes: (j) cancelados, (k) ativos e (l) concluintes.....	96
Figura 4.5: Da esquerda para a direita, temos os gráficos que mostram a média das disciplinas aprovadas para os estudantes: (a) cancelados, (b) AFP e (c) concluintes. Os gráficos que apresentam o CR para: (d) cancelados, (e) ativos e (f) concluintes. Os gráficos com o CRA para estudantes: (g) cancelados, (h) ativos e (i) concluintes.....	97
Figura 4.6: Da esquerda para direita temos os gráficos: (a) Exemplos corretamente classificados (acurácia) por curso de graduação. (b) Taxa de acerto da classe não-	

progresso por curso.....	116
Figura 4.7: Porcentagem de exemplos corretamente classificados pelos algoritmos	
<i>Naive Bayes</i>	120
Figura 4.8: Porcentagem de exemplos corretamente classificados (taxa de acerto) pelo	
algoritmo Naive Bayes para a classe não-progresso.	121
Figura 4.9: Porcentagem de exemplos corretamente classificados (taxa de acerto) pelo	
algoritmo Naive Bayes para a classe progresso.....	121

Lista de Termos e Abreviações

ANDIFES - Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior

AVA - Ambiente Virtual de Aprendizagem

EAD - Educação a distância

EDM - Mineração de Dados Educacionais (*Educational Data Mining - EDM*)

ENADE - Exame Nacional de Desempenho de Estudantes

ENEM - Exame Nacional do Ensino Médio

ETL - Extração Transformação Carga (*Extract Transform and Load - ETL*)

GUI - Graphical User Interface

IA - Inteligência Artificial

IFES - Instituições Federais de Ensino Superior

KDD - Descoberta de Conhecimento em Dados (*Knowledge Discovery from Data - KDD*)

KMR - Repositório de Conhecimento (*Knowledge Management Repository*)

MD - Mineração de Dados

OCDE - Organização para a Cooperação e Desenvolvimento Econômico

OECD - Organisation for Economic Cooperation and Development

PISA - *Programme for International Student Assessment*

REUNI - Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais

SAE - Secretaria de Assuntos Estratégicos da Presidência da República

SGA – Sistema de Gestão Acadêmica

SIGA - Sistema de Gestão Acadêmica da UFRJ

SISU - Sistema de Seleção Unificada

STEM – Ciências, Tecnologia, Engenharia e Matemática (acrônimo dos termos *Science Technology, Engineering, and Mathematics*)

UFRJ - Universidade Federal do Rio de Janeiro

1 Introdução

O Brasil está se tornando um país cada vez mais competitivo no cenário internacional, as empresas brasileiras requerem profissionais com competências e habilidades específicas. A própria Secretaria de Assuntos Estratégicos da Presidência da República (SAE) investiga se os brasileiros estão preparados para atender as novas exigências do mercado de trabalho (SAE, 2013a). A produtividade do Brasil é aproximadamente cinco vezes menor que a dos Estados Unidos e, segundo a Organização para a Cooperação e Desenvolvimento Econômico (OCDE), apenas 11% dos brasileiros na faixa etária entre 25 a 64 anos possuem graduação. Nos países ricos, essa taxa é de 31% (OECD, 2012, SAE, 2013a). A escassez de mão de obra qualificada entre os brasileiros abre perspectivas para que o governo e as empresas busquem mão de obra estrangeira para ocupar as vagas ociosas. O governo admite que a educação dos brasileiros deveria ser a melhor solução para o problema, no entanto, deixa claro que é um caminho demorado. Para solucionar o problema da falta de mão de obra, o governo incentiva a imigração de profissionais e espera que os brasileiros busquem mais formação para garantir seu espaço no mercado de trabalho (SAE, 2013b). O governo relata que há necessidade de seis milhões de profissionais estrangeiros nos próximos anos para suprir a falta de pessoal especializado nas áreas de engenharia e saúde (SAE, 2013c). As discussões em torno do assunto prosseguem envolvendo o governo, sindicatos, órgãos de classe e a sociedade civil organizada. Embora as universidades não tenham se posicionado claramente a respeito do assunto, elas estão diretamente relacionadas ao problema.

O déficit nacional de mão de obra especializada é uma questão multifacetada, ele é parcialmente decorrente de problemas que ocorrem em diversas universidades: elevadas taxas de evasão e retenção. Atualmente, as universidades oferecem a cada ano um crescente número de vagas, no entanto, o número de formados reduz a cada ano (INEP, 2012a). A ocupação de uma vaga em uma universidade pública seguido do abandono tornou-se um problema generalizado, independente da instituição, gerando perdas pessoais, sociais e financeiras (SOARES, 2000, 2006, 2009, MEC, 2007, ANDIFES, 2008, 2012, TIGRINHO, 2008, FORGRAD, 2012, INEP, 2012b, LIMA JR, 2012).

De um modo geral, as universidades podem atrair maior ou menor número de estudantes em função de vários fatores, por exemplo, sua localização geográfica, forma de ingresso, número de vagas ofertadas, formas de ingresso, qualidade dos cursos, adequação dos cursos ao contexto socioeconômico da região. Da mesma forma, os cursos de graduação são atrativos quando oferecem boas perspectivas de entrada no mercado de trabalho, grade curricular atualizada e adequada à formação profissional do estudante.

1.1 Definição do Problema

As Instituições Federais de Ensino Superior (IFES) são importantes atores no cenário nacional, elas atuam nos segmentos de ensino, pesquisa e extensão e anualmente recebem milhares de estudantes e devolvem para a sociedade um número de profissionais aquém do esperado (INEP, 2012a). Por exemplo, a Universidade Federal do Rio de Janeiro (UFRJ) é a terceira universidade em oferta de vagas (3.669) no SISU (MEC, 2014). Considerada uma das maiores IFES do Brasil, possui em torno de 36 mil estudantes de graduação e mais de 100 cursos de graduação (UFRJ, 2014).

A UFRJ participa do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (REUNI) instituído pelo Decreto nº 6.096, de 24 de abril de 2007 (MEC, 2007). O REUNI mostra que o governo e as IFES estão preocupados e mobilizados para mitigar o problema da evasão. O documento relata que “*os índices de evasão de estudantes nos cursos de graduação atingem, em alguns casos, níveis alarmantes*”. O governo propõe como meta a elevação gradual da taxa de conclusão média dos cursos de graduação presenciais para noventa por cento.

O insucesso na conclusão da graduação acontece quando o estudante para de frequentar o curso ou não consegue obter rendimento acadêmico suficiente para fazer jus à diplomação. Neste último caso, algumas universidades aplicam o jubramento, que consiste no processo de forçar o estudante a deixar o curso após um determinado tempo, mesmo que ele não tenha completado os estudos.

A complexidade que envolve este antigo problema atinge indistintamente as IFES e vem sendo continuamente examinado pelo governo (MEC, 1997). Este definiu que a evasão pode ser entendida em três eixos:

- (1) *Evasão de curso* - quando o estudante desliga-se do curso de graduação em situações diversas: abandono (deixa de se matricular), desistência (oficial),

transferência (mudança de curso) ou exclusão por norma institucional;

- (2) *Evasão da instituição* - quando o estudante desliga-se da instituição na qual está matriculado;
- (3) *Evasão do sistema* - quando o estudante abandona de forma definitiva ou temporária o ensino superior.

No contexto desta tese, focaremos nossa pesquisa apenas no segmento graduação, adotou-se o primeiro eixo considerado pelo governo federal.

Embora o problema da evasão aconteça em diversas universidades, os índices de evasão podem variar entre as instituições e entre os cursos de graduação (MEC, 1997, INEP, 2009, 2011, 2012a). Relacionar as causas da evasão e retenção, identificar os fatores que mais influenciam as baixas taxas de diplomação e elevada evasão e atribuir uma ordem de importância a estes fatores é um trabalho multidisciplinar complexo e fortemente ligado ao contexto da IFES. No entanto, a evasão, quando o estudante se desliga do curso em situações diversas, não é um ato repentino, podem-se identificar os sinais que antecedem a saída definitiva do estudante (BARDAGI, 2007, 2009).

Dentre as diversas unidades organizacionais (Institutos, Escolas e Faculdades) da UFRJ existe uma percepção do problema da evasão. Alguns estudos já avaliaram o problema. Dentre eles destacamos BARROSO e FALCÃO (2004) que explica as condições que motivam a evasão sob três agrupamentos:

- (1) *Econômica* – o estudante apresenta impossibilidade de permanecer no curso por questões socioeconômicas;
- (2) *Vocacional* – o estudante não se identificou com o curso;
- (3) *Institucional* – abandono por fracasso nas disciplinas iniciais, deficiência prévia de conteúdos anteriores, inadequação aos métodos de estudo, dificuldades de relacionamento com colegas ou com membros da instituição.

Os gestores acadêmicos discutem o problema e buscam soluções (UFRJ, 2009), no entanto, o acompanhamento do desempenho dos estudantes de graduação é feito de maneira empírica, subjetiva e não sistemática, pois depende primordialmente da experiência acadêmica e do envolvimento de um pequeno grupo de docentes. Geralmente, estes docentes desempenham inúmeras atividades além das ligadas a sala de aula e aos laboratórios de pesquisa, portanto é difícil acompanhar e reconhecer as necessidades individuais de cada estudante. Torna-se necessário investigar e propor mecanismos que automatizem e viabilizem o acompanhamento do desempenho acadêmico dos estudantes e identifique aqueles que estão em situação de risco de

evasão.

Nesta tese, caracterizaremos o estudante em risco de evasão e investigaremos se este estudante apresenta desempenho acadêmico que o difere do estudante que será capaz de concluir o curso. Com base nesta premissa, é possível identificar quais são os estudantes que estão mais propensos ou em risco de evasão do curso de graduação. Portanto, será possível desenvolver estratégias computacionais que podem prever quais os estudantes que estão em risco de evasão e quais os estudantes que poderão obter a diplomação. Caracterizar quantitativamente estes tipos de estudantes se constitui um dos grandes desafios desta pesquisa. Tendo em vista que a área da mineração de dados possui recursos e técnicas que podem ser utilizados para resolver problemas de predição em diversas áreas e, atualmente, a mineração de dados está sendo ampla e consistentemente utilizada para resolver problemas envolvendo dados educacionais. Torna-se, portanto, interessante investigar a aplicação da mineração de dados no contexto do problema proposto nesta tese.

Via de regra, os sistemas de gerenciamento acadêmicos utilizados pelas IFES armazenam grandes volumes de dados sobre os estudantes. No entanto, converter estes dados em informações úteis passa pela realização de diversas etapas, este processo é conhecido como Descoberta de Conhecimento em Dados (*Knowledge Discovery from Data - KDD*) (HAN, KAMBER, 2006). A mineração de dados é a parte central deste processo. Depois de selecionadas, as bases de dados passam por um rigoroso processo de pré-processamento de dados, a fim obter dados de qualidade para serem utilizados no processo analítico de mineração de dados (HAN, KAMBER, 2006). Existem poderosas ferramentas de mineração de dados disponíveis no mercado, no entanto, executar todo o processo de KDD requer conhecimento especializado além de tempo. Estes fatores tornam a utilização da mineração de dados mais complexa para gestores acadêmicos não especialistas em bancos de dados e análise de dados. O uso de técnicas de mineração de dados dentro do contexto educacional é chamado de *Mineração de Dados Educacionais (Educational Data Mining - EDM)*, definido recentemente como um campo que explora estatística, aprendizado de máquina e algoritmos de mineração de dados aplicados a diferentes tipos de dados educacionais (ROMERO, VENTURA, 2010).

A mineração de dados depende fortemente da qualidade e da quantidade de dados disponíveis para executar os algoritmos. Uma das limitações está em encontrar uma fonte segura e selecionar o maior número possível de dados (atributos) relevantes para

investigar o problema. O Sistema de Gestão Acadêmica (SIGA) (UFRJ, 2014) é uma fonte segura de dados e possui as mesmas informações acadêmicas para todos os estudantes de graduação da UFRJ. O SIGA é um sistema de informação legado que armazena dados acadêmicos sobre todos os estudantes da UFRJ. O sistema identifica cada estudante por um número de matrícula e armazena o ano e semestre de ingresso, identificação do curso de graduação, disciplinas cursadas a cada período letivo, notas, coeficiente de rendimento no período, coeficiente de rendimento acumulado, situação no período, situação no curso entre outras informações acadêmicas.

Na condução desta pesquisa, consideram-se a seguinte hipótese:

É possível prever o desempenho acadêmico de estudantes de diversos cursos de graduação utilizando técnicas de mineração de dados educacionais aplicadas à base de dados extraídas do Sistema de Gestão Acadêmica das IFES?

1.2 Objetivos da Tese

Esta tese tem como objetivo geral identificar o desempenho acadêmico dos estudantes de graduação da UFRJ utilizando técnicas de mineração de dados educacionais, contribuindo para que os gestores acadêmicos identifiquem os estudantes que apresentam risco de evasão do sistema. A UFRJ recebe anualmente centenas de estudantes e carece de mecanismos automáticos que facilitem este tipo de avaliação. Atualmente, não existe um procedimento administrativo formal de identificação de estudantes nesta situação. Além disso, o grande número de estudantes que se matriculam a cada semestre inviabiliza procedimentos de identificação por meios não automatizados. Portanto, é necessário atender a uma demanda latente e investigar os novos métodos necessários para desenvolver mecanismos que automatizem e viabilizem a detecção precoce de estudantes em risco evasão e, conseqüentemente, contribuir com redução dos índices de evasão nos cursos de graduação.

Para atender ao objetivo geral da tese foi criada uma arquitetura modular em camadas que incorpora técnicas de mineração de dados educacionais para prever o desempenho de estudantes ao término de cada semestre letivo e apontar aqueles que apresentam risco de evasão.

Dentro desse contexto, definem-se os seguintes objetivos específicos:

- (1) Caracterizar e apresentar o problema da evasão e retenção, esta classe de problemas é compartilhada por diversas IFES, em especial na UFRJ;

- (2) Identificar os diferentes tipos de desempenho acadêmico que os estudantes podem apresentar durante o curso de graduação;
- (3) Propor uma metodologia apoiada por técnicas de mineração de dados aplicados a base de dados de informações acadêmicas, de modo a identificar as características (variáveis) mais significativas para distinguir o desempenho acadêmico dos graduandos em diversos cursos de uma IFES, tendo a UFRJ como cenário principal;
- (4) Projetar e desenvolver uma arquitetura computacional que seja capaz de ser acoplada ao sistema legado das IFES e apoiar os gestores acadêmicos a identificar o desempenho acadêmico do estudante e emitir alertas aos gestores sobre aqueles estudantes em possível risco de evasão. A arquitetura deverá contemplar generalizações para utilizar em diversos cursos de graduação da UFRJ e ser adaptável para utilização em outras IFES;
- (5) Realizar experimentos quantitativos e discutir e apresentar os resultados desta pesquisa.

1.3 Contribuição

O presente trabalho visa contribuir para a redução da evasão no ensino superior brasileiro. Este trabalho concebe, projeta, desenvolve e testa uma nova abordagem. Esta nova abordagem é composta pela arquitetura denominada EDM WAVE (MANHÃES *et al.*, 2014b, 2014c, 2014d). Esta arquitetura baseada em três camadas é fundamentada nos conceitos da mineração de dados educacionais enunciados por (BAKER, YACEF, 2009, ROMERO, VENTURA, 2010, BAKER *et al.*, 2011, MÁRQUEZ-VERA, 2013). A arquitetura EDM WAVE permite automatizar a predição do desempenho acadêmico dos estudantes a cada semestre letivo. A arquitetura EDM WAVE se propõe a trabalhar acoplada ao sistema legado de gestão acadêmica das IES, os SGA já existentes podem incorporar as funcionalidades da arquitetura EDM WAVE, contribuindo assim para redução nos custos de desenvolvimento de novos SGA.

A arquitetura EDM WAVE foi concebida para auxiliar os gestores acadêmicos das IFES na tomada de decisão e evitar (reduzir) os altos índices de evasão e a permanência dos estudantes além do tempo previsto para conclusão dos cursos de graduação. Os gestores acadêmicos das IFES podem acompanhar sistematicamente o desempenho acadêmico dos estudantes de graduação, identificando aqueles que estão com

difficultades de cumprir com as exigências acadêmicas. O acompanhamento do desempenho acadêmico permitirá que os gestores identifiquem estudantes em risco de evasão. Deste modo, podem-se planejar ações que mitiguem o processo de evasão ou mesmo diminuam os índices de evasão do curso e da universidade. O acompanhando sistemático dos estudantes permite identificar precocemente em qual semestre letivo ocorre declínio no desempenho, identificar as possíveis causas do declínio e evitar que essas causas perdurem no tempo. Além de permitir que a universidade não utilize apenas dados estatísticos para a análise do problema.

A pesquisa foi realizada utilizando dados reais dos estudantes extraídos diretamente do Sistema SIGA (UFRJ, 2014). A arquitetura EDM WAVE é uma das primeiras a utilizar somente variáveis com dados de estudante que variam com o tempo (*time-varying student data*). Nesta tese foram estabelecidos novos modelos de dados de estudantes que pudessem ser comuns a todos os cursos de graduação da UFRJ. Foram propostos modelos de dados focados em um número reduzido de variáveis que descrevem dados acadêmicos de estudantes em diversos períodos letivos do curso de graduação. Os modelos de dados dos graduandos foram construídos a partir do resultado de várias investigações e experimentação, como descritos nos 7 estudos de casos apresentados nesta tese.

Os estudos de casos analisados nesta tese foram utilizados para avaliar vários algoritmos classificadores, foram testados 12 algoritmos classificadores. Os critérios para comparar e avaliar os algoritmos classificadores foram: acurácia, taxa de erro e acerto para cada classe analisada, matriz de confusão, tempo de execução, valor do Kappa e interpretação do modelo. O algoritmo classificador *NaïveBayes* apresentou os melhores resultados gerais e um modelo de classificação mais interpretável. Os resultados apresentados por este algoritmo permitiram o desenvolvimento de análises quantitativas e representações gráficas com maior valor informativo.

Esta tese contribuiu para o aperfeiçoamento da mineração de dados educacionais (EDM), pois desenvolvemos uma arquitetura voltada para atender um dos tópicos de estudo em aberto da EDM que é a predição do desempenho acadêmico de estudantes em cursos de graduação. Ampliamos o universo de pesquisa de um número limitado de cursos, como mostrado nos trabalhos relacionados, para um grande número de cursos de graduação, como demonstramos a partir da utilização do cenário real investigado na UFRJ. Os estudos de casos apresentados nesta tese foram elaborados de modo que possam ser reproduzidos por qualquer equipe que pretenda utilizar EDM em outras

universidades de ensino presencial ou EAD. Portanto, esta tese serve como base para que as pesquisas em EDM possam avançar e alcançar novas metas.

Em termos de produtos gerados pela tese, delineamos os requisitos de uma arquitetura que pode ser acoplada aos sistemas de Gestão Acadêmica existentes na IFES, avaliando-a no âmbito de cursos de graduação da UFRJ. Além disso, em termos de publicações produzimos ao longo da pesquisa 7 artigos (MANHÃES *et al.*, 2011, 2012, 2014a, 2014b, 2014c, 2014d, 2015) e 1 artigo submetido a um periódico (*journal*) ainda em fase de avaliação no momento da defesa da tese.

1.4 Metodologia da Pesquisa

A metodologia que norteia as ações para elaborar as estratégias voltadas para diagnosticar o desempenho acadêmico dos discentes em cursos de graduação e o desenvolvimento de uma arquitetura EDM WAVE para apoiar os gestores acadêmicos na tomada de decisão está apoiada em diversas atividades descritas a seguir.

O desenvolvimento do estudo da tese passou por diversos estágios, foram seguidas as diretrizes do processo de Descoberta de Conhecimento em Dados (*Knowledge Discovery from Data - KDD*) (HAN, KAMBER, 2006). Este processo será descrito na seção 3.1.

Em primeiro lugar, iniciou-se pelo entendimento do insucesso universitário dentro de diversas perspectivas: dos estudantes, da IES, do governo e sociedade. Identificamos o problema da evasão e retenção de discente em cursos de graduação nas universidades nacionais e internacionais, particularmente, dentro do contexto da UFRJ. Investigou-se as soluções computacionais já utilizadas para solucionar o problema abordado. Nesta fase, formulamos os objetivos do projeto, identificamos os requisitos do problema e os dados necessários para aplicar a mineração de dados educacionais.

Na fase do entendimento dos dados, os dados foram coletados e explorados. Nesta fase inicial do processo de KDD, a base de dados utilizada foi extraída do sistema de gestão acadêmico da UFRJ. Verificou-se se os dados são adequados e relevantes.

Utilizou-se o processo de Extração Transformação e Carga (*Extract Transform and Load - ETL*) na fase de seleção e transformação dos dados do KDD. Os dados originais coletados do SIGA foram explorados e transformados. Identificou-se que os dados originais eram relevantes, no entanto, o formato não era adequado para utilizar técnicas de mineração de dados. Esta fase de preparação dos dados envolveu uma seleção

cuidadosa de dados e transformação em novos atributos relevantes e no formato adequado para serem usados em aplicações de mineração de dados. Nesta fase também foram selecionados variáveis de classe para serem utilizadas pelos algoritmos classificadores, cujos valores foram definidos em função dos novos dados derivados dos dados originais do SIGA.

Após a fase de seleção e transformação do KDD, os novos dados estão prontos para a fase de mineração de dados e aplicação dos algoritmos classificadores para a criação de um modelo preditivo de dados. Nesta tese serão apresentados, no capítulo 4, diversos estudos de casos com os seguintes objetivos: (1) avaliar os novos modelos de dados dos estudantes e sua adequação na predição do desempenho acadêmico dos estudantes a cada semestre letivo; (2) avaliar e produzir um comparativo entre 12 algoritmos classificadores aplicados aos novos modelos de dados dos estudantes e identificar quais algoritmos são mais adequados ao contexto do problema dos estudantes da UFRJ. Na seção 3.2 serão descritos os diversos critérios para avaliar os algoritmos utilizados e na seção 4.4 serão descritos os algoritmos classificadores utilizados. Por fim, reavaliar os modelos obtidos através da aplicação de novas entradas de graduandos.

A ferramenta chamada *Waikato Environment for Knowledge Analysis (Weka)* é uma um software *open source* baseado em código Java que implementa diversos algoritmos de aprendizado de máquina (WITTEN, FRANK, 2005). A ferramenta Weka proporciona diversos mecanismos de análise para aplicações de mineração de dados.

Por fim uma arquitetura EDM WAVE de três camadas foi concebida para agregar novos valores e funcionalidades ao Sistema de Gestão Acadêmica (SGA) das universidades públicas federais brasileiras. Na seção 3.4 foram descritos detalhes da arquitetura. A arquitetura acrescenta funcionalidades aos sistemas de gestão acadêmicos legados a partir da perspectiva de avaliar sistematicamente o desempenho dos estudantes ou prever aqueles que estão em risco de abandonar o curso de graduação. Além de caracterizar os diferentes desempenhos de estudantes dos cursos de graduação;

A metodologia adotada nesta tese contempla diversas fases de estudo, sendo responsável pelo desenvolvimento de uma estrutura para prever o desempenho acadêmico dos estudantes de graduação da UFRJ. Os resultados obtidos em cada fase do KDD foram discutidos, assim como os resultados obtidos na investigação dos estudos de casos.

1.5 Organização da Tese

Este trabalho está organizado da seguinte forma: No capítulo 2 contextualiza-se o problema dentro do campo da mineração de dados educacionais (EDM) e apresentam-se trabalhos correlatos. No capítulo 3 são discutidos detalhes da proposta de solução para o problema abordado. Apresentam-se detalhes da arquitetura, suas funcionalidades e interações. No capítulo 4 descrevem-se sete estudos de casos, nestes estudos de casos foram realizados diversos experimentos e testes para validar a proposta de tese. Diversas métricas foram utilizadas para avaliar os modelos de dados dos estudantes e o conjunto de algoritmos classificadores utilizados em cada experimento. No capítulo 5 apresentem-se a conclusão do estudo, limitações e considerações sobre trabalhos futuros.

2 Caracterização do Problema e Trabalhos Correlatos

A educação é um dos fundamentos para o desenvolvimento do ser humano. Não se pode pensar em um cidadão consciente, participante e atuante desvinculado de uma formação educacional de qualidade. Embora não esteja no escopo desta tese tratar em profundidade os diversos aspectos relacionados à educação. No entanto, trataremos neste capítulo de um assunto fortemente relacionado à educação: diplomação, retenção e evasão nos cursos de graduação das IFES. Veremos que o insucesso escolar é um assunto abordado por diversos segmentos acadêmicos e pelo governo federal (MEC, 1997, INEP, 2009, 2011, 2012a, OECD, 2012).

Este capítulo apresenta considerações sobre o problema, define os termos relacionados, discute trabalhos correlatos e suas limitações sobre o levantamento dos dados acadêmicos em IFES.

2.1 *Definição dos Termos Utilizados*

No Brasil, existem diversos termos utilizados para categorizar o sucesso ou insucesso dos discentes. De uma maneira geral, o sucesso dos discentes nas IFES está relacionado à obtenção do diploma de graduação (diplomação). Outros termos são utilizados para descrever o insucesso escolar, os mais empregados são: evasão, abandono e retenção (atraso). A seguir, veremos como os termos ligados ao sucesso e insucesso escolar são empregados e definidos em diversos contextos.

A diplomação certifica a conclusão do curso de graduação e confere o grau de formado ao discente. Para a conclusão do curso, o discente deve cumprir todas as exigências curriculares do curso de graduação, além de outras exigências legais do MEC, por exemplo, o ENADE. Para o MEC (1997), “*As preocupações maiores de qualquer instituição de ensino superior, em especial quando públicas, devem ser a de bem qualificar seus estudantes e a de garantir bons resultados em termos de número de diplomados que libera a cada ano para o exercício profissional.*”

Por outro lado, os termos utilizados para caracterizar ou descrever o insucesso

escolar não possuem uma definição precisa. O sítio do movimento “Todos pela Educação” (TPE, 2008) disponibiliza um glossário, onde são mostradas as definições para evasão e abandono:

“Evasão: condição do aluno que, matriculado em determinada série, em determinado ano letivo, não se matricula na escola no ano seguinte, independentemente de sua condição de rendimento escolar ter sido de aprovado ou de reprovado” e

“Abandono: condição do aluno que deixa de frequentar a escola durante o andamento de determinado ano letivo.”

Neste sítio não está claro se as definições acima são para todos os níveis escolares ou específicos para ensino fundamental e médio.

Em outro documento do MEC, especialmente criado para analisar as taxas de diplomação, retenção e evasão dos cursos de graduação das universidades públicas brasileiras (MEC, 1997), encontramos definições mais precisas sobre os termos utilizados no ensino superior brasileiro. Esse documento foi criado pela Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras (MEC, 1997) e mostra a preocupação do governo com as altas taxas de evasão nas IFES. As discussões apresentadas no documento refletem sobre a complexidade do problema no ensino superior brasileiro e definem que a evasão pode ser entendida em três eixos:

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

A evasão de curso trata, portanto, dos estudantes que por qualquer razão desligam-se do curso de graduação no qual estão matriculados, sendo que o abandono do curso é uma situação dentro do termo mais geral chamado de evasão de curso.

A definição de retenção de estudantes apresentada pelo MEC (1997) descreve a situação em que o estudante, ainda matriculado na instituição, permanece nos cursos além do tempo máximo de integralização curricular. A semântica da palavra retenção pode induzir a um erro de interpretação, no Brasil ela não significa uma ideia oposta à evasão. Pelo contrário, a retenção pode ser entendida como um atraso e significa que o

discente não está progredindo no curso para obter a diplomação. Em LOBO (2011) encontramos a seguinte observação: “nos EUA a retenção é uma política de combate à evasão, antievasão.”.

A retenção escolar acarreta perdas pessoais para o estudante, para IFES, para sociedade e em última análise para o país (DIAS, 2009, VASCONCELOS, 2011). Para a IFES é um investimento de recursos sem retorno porque a instituição alocou uma vaga, mas não teve êxito na diplomação. A sociedade padece pela falta de profissional qualificado e pelo mau uso dos recursos públicos. O estudante na situação de retenção é um forte candidato a ser um estudante evadido. No entanto, apesar da importância, não há muitos estudos sobre o assunto, segundo DIAS (2009) este é um fenômeno mais fácil de ser estudado porque o estudante retido ainda permanece vinculado à instituição, sendo mais fácil encontrá-lo. Por outro lado, o estudante evadido quebra o vínculo perde-se o contato.

O problema da evasão estudantil é mais discutido do que a retenção, segundo DIAS (2009) a evasão é mais facilmente percebida pelos gestores acadêmicos porque há uma ausência concreta de estudantes, o número de matrículas diminui e vagas ficam ociosas. Por outro lado, na retenção o estudante permanece na instituição, a vaga ainda está ocupada, mas o papel da IFES de formar o cidadão para a sociedade e diploma-lo para o mercado de trabalho não está sendo cumprido. Outro fator que diferencia a evasão da retenção está no fato que a evasão é mais acentuada no início do curso de graduação, SILVA FILHO (2007) afirma que a taxa de evasão no primeiro ano de curso é duas a três vezes maior do que nos anos seguintes. A retenção excessiva é mais perceptível quando o estudante está adiantado no curso, mas está fora do prazo para diplomação (DIAS, 2009). As causas da retenção podem ser as mesmas da evasão acrescido do alto índice de reprovação, dificuldades financeiras momentâneas ou envolvimento em outras atividades, os fatores motivadores ainda não estão definidos, também depende da política de jubramento de cada IFES (VASCONCELOS, 2011).

Nesta tese, também utilizaremos a expressão *insucesso* para referenciar, de um modo geral, a evasão e/ou retenção escolar. E a palavra *sucesso* como referência a trajetória acadêmica do discente que recebe a diplomação ao término da formação no curso de graduação.

Outras definições do MEC (1997) serão utilizadas ao longo do nosso trabalho:

- *Ano/período-base* - corresponde ao ano e semestre de ingresso do estudante na universidade;

- *Ingressante* - estudante que ingressou em um curso de graduação em um ano/período-base, independentemente da forma de ingresso;
- *Diplomado* (ou concluinte) - estudante que concluiu o curso de graduação dentro do prazo máximo de integralização curricular fixado pelas normas da IFES, contado a partir do ano/período-base de ingresso;
- *Retido* - estudante que, apesar de esgotado o prazo máximo de integralização curricular fixado pelas normas da IFES, ainda não concluiu o curso, mantendo-se matriculado na universidade;
- *Evadido* - estudante que deixou o curso de graduação sem obter a diplomação.

2.2 *Análise e Contextualização do Problema*

O aprendizado formal dentro do âmbito escolar constitui importante instrumento de desenvolvimento humano, tanto no campo pessoal, intelectual, profissional, social e financeiro. O diploma universitário é mais que um instrumento, ele sintetiza o objetivo de muito brasileiros que, conseqüentemente, almejam ascensão pessoal, profissional e financeira na sociedade. O discente, uma vez matriculado em um curso universitário, depara-se com inúmeros desafios e obstáculos para cumprir as exigências acadêmicas. Isto leva muitos estudantes a interromper de forma temporária ou definitiva o curso de graduação no qual estão matriculados.

O insucesso dos discentes na obtenção da formação nos cursos de graduação das IFES brasileiras representa um problema complexo e atingem inúmeras instituições distribuídas em todo o espaço geográfico brasileiro. Embora sendo um problema generalizado, suas causas e, conseqüentemente, sua solução ou mitigação ou prevenção dependem fundamentalmente do contexto onde ele ocorre. No entanto, segundo LOBO (2011) há poucos estudos sobre a evasão no Brasil, quando ocorrem são desenvolvidos por pequenos grupos de docentes de forma não sistemática e limitados a apenas um curso de graduação, enquanto em outros países ocorrem estudos em maior número e são conduzidos de forma sistemática e mais ampla.

Relatamos, nesta seção, os principais estudos que investigaram os aspectos relacionados ao problema da evasão e/ou retenção em diversas universidades do Brasil. Os estudos apresentam uma análise das causas dentro do contexto universitário da IFES, mostram dados estatísticos que comprovam a gravidade do problema e avaliam as

consequências para os estudantes, para universidade e sociedade (RAMALHO FILHO, 2008). No entanto, observamos que os trabalhos estão focados no levantamento de causas, poucos mencionam ações de prevenção ou possíveis reversão do quadro.

A totalidade dos artigos analisados relata que a evasão universitária pode ocorrer em função de diversos motivos (BARROSO, FALCÃO, 2004, SOARES, 2000, 2006, 2009, SILVA FILHO, 2007, ANDRIOLA, 2009, DIAS, 2010, LOBO, 2011), sua origem e motivação pode estar na pessoa do estudante e/ou na instituição de ensino. No entanto, as consequências da não obtenção do diploma de conclusão do curso de graduação afetam o estudante, a instituição e o país. O modelo desenvolvido por TINTO (1975, 1993, 2006) utilizado nas universidades americanas, sugere que a evasão ocorre pela falta de integração do estudante com o ambiente acadêmico e/ou social da universidade. O modelo de TINTO não se aplica completamente a realidade das universidades brasileiras como afirma ANDRIOLA (2009) porque os cursos e os fatores sociais são diferentes da realidade de outros países.

No trabalho desenvolvido por VELOSO e ALMEIDA (2001) na Universidade Federal de Mato Grosso – UFMT, os autores fizeram um trabalho de pesquisa para identificar o problema da evasão em 14 cursos de graduação daquela universidade. Eles mencionaram que durante o curso de graduação, entrada do estudante na instituição, e o momento de sua saída, esperada formatura, uma série de fatores ocorrem que atuam positivamente ou negativamente, influenciando a continuidade do estudo, a evasão como interrupção do processo educacional ocorre por diversos motivos relacionados ao próprio estudante e relacionados à instituição. Os autores destacam que a universidade não tem um plano de procurar o estudante evadido e verificar as causas sob o seu ponto de vista, o que seria o mais adequado. Portanto, a análise das causas é parcial porque só considera dados de um questionário respondido pelos coordenadores de curso. O trabalho não considerou qualquer informação do estudante.

Na pesquisa de (SILVA FILHO, 2007), o autor destaca que:

“A evasão estudantil no ensino superior é um problema internacional que afeta o resultado dos sistemas educacionais. As perdas de estudantes que iniciam mas não terminam seus cursos são desperdícios sociais, acadêmicos e econômicos. No setor público, são recursos públicos investidos sem o devido retorno. No setor privado, é uma importante perda de receitas. Em ambos os casos, a evasão é uma fonte de ociosidade de professores, funcionários, equipamentos e espaço físico.”

Muitas são as visões sobre o problema, para o Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia e Lobo & Associados Consultoria (LOBO, 2011), a evasão é um problema de gestão acadêmica. Em concordância com os demais trabalhos relacionados, eles mostram que há vários tipos de evasão. No entanto, ressalta a importância de desenvolver análises e pesquisas sobre o assunto, visto que há uma escassez de trabalhos cientificamente conduzidos no Brasil. Faltam iniciativas, realização de pesquisas e estudos sistemáticos sobre a evasão dentro do nosso sistema de ensino superior que permitisse indicar com precisão quais são as melhores práticas para combatê-la, e não somente se espelhando em estudos realizados em outros países.

A evasão compromete a estrutura acadêmica, pois originalmente foi previsto atender a um quantitativo maior de estudantes (CAMPELLO, LINS, 2008, SOARES, 2009).

Alguns autores apontam a ineficiência das IFES para tratar o problema, salientando o distanciamento e a falta de interlocução entre os discentes, professores e gestores acadêmicos, por exemplo: os coordenadores de curso. Além disso, não há instrumentos públicos e ferramenta computacionais que permitam aos gestores acadêmicos acompanhar sistematicamente o desempenho acadêmico dos estudantes matriculados nos cursos, falta um projeto de acompanhamento e apoio psicopedagógico (VELOSO, ALMEIDA, 2001, SILVA FILHO, 2007).

O Censo da Educação Superior mostra a cada ano uma visão geral da situação do ensino superior no Brasil (INEP, 2009, 2011, 2012a). No entanto, o acesso a informações mais detalhadas sobre o insucesso dos discentes nas IFES é muito restrito. O estudo detalhado para o entendimento do problema da evasão e retenção dentro das IFES carece de dados. Destacamos a falta de definição de quais dados são relevantes, fontes confiáveis e acesso por parte dos pesquisadores.

Entendemos que o problema a ser abordados nesta tese se desdobra e pode ser abordado sob três diferentes perspectivas: estudante, instituição e sociedade (país).

2.2.1 Abordagem do Problema sob a Perspectiva do Estudante

Muitos estudantes que evadem das IFES não são devidamente acompanhados ou instruídos para fazer um relato dos motivos da evasão. No entanto, alguns segmentos da UFRJ fizeram estudos estatísticos para acompanhar o problema, dentro das Engenharias destacamos SOARES (2000, 2006, 2009) e SARAIVA e MASSON (2003). No Instituto

de Física da UFRJ um trabalho realizado por BARROSO e FALCÃO (2004) com os estudantes de graduação verificou-se que há três fatores importantes associados à evasão:

- (1) *Econômica* - impossibilidade de manutenção do vínculo por questões socioeconômicas;
- (2) *Vocacional* - percepção de uma escolha de curso inadequada aos interesses do discente;
- (3) *Institucional* – abandono por inadequação (fracasso nas disciplinas iniciais, deficiência prévia de conteúdos acadêmicos, inadequação aos métodos de estudo) ou dificuldades de relacionamento (com os colegas e com os membros da instituição).

De fato, estes itens são constantes entre as causas de evasão e retenção de estudantes. Muitos estudantes encontram dificuldades de permanecer no curso por questões financeiras, item (1), mesmo sendo cursos gratuitos, existem altos custos relacionados à aquisição de material didático, transporte, moradia e alimentação. Por outro lado, muitos estudantes são atraídos precocemente pelo mercado de trabalho, devido à crescente demanda por mão-de-obra, surgem propostas de emprego para os estudantes antes do término do curso agravando ainda mais a evasão e a retenção, pois em muitos casos os estudantes envolvidos com outras atividades não conseguem se dedicar de modo satisfatório ao curso.

No item (2) relacionam-se estudantes que insatisfeitos com o curso de graduação que estão cursando podem pedir transferência para outros cursos de graduação na mesma instituição, ou podem reiniciar um curso em outra universidade ou abandonam o sistema de ensino superior. De qualquer forma, a opção errada de curso gera prejuízos para o estudante e para a instituição.

No item (3) está bastante relacionado às dificuldades de obter média para aprovação nas disciplinas.

Outro estudo realizado por SOUZA (2008) identificou os mesmos problemas que pesam na decisão de abandonar o curso:

- (1) Falta de condição financeira para se manter no curso;
- (2) Incompatibilidade de horários do curso, trabalho e sustento próprio ou da família;
- (3) Escolha da graduação feita sem a análise adequada da futura atividade profissional ou do mercado de trabalho.

Em SILVA (2011) analisou o problema da evasão e retenção no curso de graduação em Ciências Econômicas da Universidade Federal de Pernambuco. Este trabalho estatístico relacionou a retenção como precursor da evasão. A pesquisa revelou que os estudantes que irão prolongar a permanência destacam os seguintes fatores: estão insatisfeitos com o curso, ocupação com o trabalho remunerado, os estudantes consideram elevado o grau de dificuldade do curso, o conhecimento básico exigido é deficiente e as reprovações em disciplinas.

O trabalho de pesquisa de SOUZA *et al.* (2012) baseou-se nas informações do sítio da Capes, o objetivo era fazer um levantamento estatístico da quantidade de dissertações e teses defendidas no período de 2000 a 2011 no Brasil sobre o tema evasão no ensino superior. Foram encontrados 32 trabalhos, sendo 28 dissertações de mestrados e 4 teses de doutorado no período considerado. Segundo os autores 64% das pesquisas visam compreender os fatores que levam o estudante à evasão numa determinada IES; 6% analisam historicamente o processo da evasão; 6% analisam a relação entre os indicadores de satisfação dos estudantes com relação à determinada IES e a evasão; 12% estudaram o perfil do estudante que evade; 3% analisaram quais cursos apresentam o maior índice de evasão; 9% desenvolvem e analisam propostas de trabalho relacionadas à tecnologia com a intenção de diminuir os índices de reprovação e de evasão. Neste estudo, evidenciaram-se alguns fatores: falta de condições financeiras, influência familiar, falta de vocação para a profissão, repetência, dificuldades na aprendizagem, decepção com a qualidade do curso, localização da IES, idade do estudante (a taxa de evasão é maior entre os estudantes mais velhos), projeto pedagógico deficiente, indisposição com professores e/ou colegas, infraestrutura precária, entre outros motivos.

Outro trabalho estatístico realizado por MELLO (2012) na Universidade Federal de Pelotas – UFPel, estuda os fatores que contribuem para evasão no curso superior de Administração no período de 2009 a 2011. Os autores fizeram um levantamento das conclusões obtidas em trabalhos nacionais de outros autores. Em essência as justificativas apresentadas pelos estudantes evadidos foram: compromisso com o trabalho remunerado; impossibilidade de conciliar horários de trabalho e aulas; expectativas diferentes com relação ao curso; dificuldades de se adaptar a didática e/ou relacionamento com o professor; dificuldade de aprendizagem dos conteúdos das disciplinas devido ao ensino médio deficiente; dificuldade de adaptação ao processo educacional da universidade que é diferente do ensino médio, a universidade exige

autonomia nos estudos, cobra a pesquisa e não simples reprodução de textos; reprovação em disciplina(s); várias reprovações na mesma disciplina; dificuldade de realizar atividades extraclasse; falta de ajuda extraclasse nos conteúdos das disciplinas.

Em resumo, observa-se que a evasão não é causada por apenas uma única e forte razão, mas por uma série de pequenos problemas que vão desmotivando o estudante. Estes casos traçam um diagnóstico de estudantes que abandonam o curso de graduação sem aviso ou explicação, ou justificam a saída devido a dificuldades financeiras, demonstrando desinteresse em permanecer estudando no curso (LOBO, 2011).

2.2.2 Abordagem do Problema sob a Perspectiva da Instituição

A Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (ANDIFES) e a Rede Universia realizaram seminário na Universidade Federal do Amapá - Macapá com o tema "Evasão e Retenção Discente nas IFES" (UFPE, 2008). Neste evento, representantes do governo e das universidades discutiram o tema amplamente. O representante da Secretaria de Ensino Superior do Ministério da Educação apresentou considerações sobre as "Políticas de Combate à Evasão". O professor Nelson Cardoso Amaral da Universidade Federal de Goiás discursou sobre "Evasão e permanência nas IFES" apontando diversas causas (AMARAL, 2008).

A evasão e retenção ocorrem em quase todas as instituições de ensino brasileiras públicas e particulares. Nas IFES o insucesso é um assunto importante e que merece atenção (CAMPELLO, LINS, 2008), a gestão acadêmica possui a responsabilidade de adequar à utilização dos recursos humanos e patrimoniais, necessários para atender ao número de vagas oferecidas nas formas de ingresso da graduação (SOARES, 2009). Os seguintes itens fazem parte das inúmeras responsabilidades atribuídas a IFES para conduzir os cursos de graduação, destacadas por SOARES (2009):

- (1) Oferta de vagas;
- (2) Preenchimento da totalidade das vagas;
- (3) Redução da evasão e da retenção dos estudantes durante o curso de graduação;
- (4) Responsabilidade na excelência da formação técnica, humana, ética e moral do estudante;
- (5) Corpo docente qualificado e comprometido;
- (6) Pessoal técnico-administrativo qualificado e comprometido;

- (7) Infraestrutura adequada;
- (8) Sistemas informatizados que permitam uma gestão eficiente;
- (9) Currículos e projetos pedagógicos atualizados e compatíveis com as demandas do mercado de trabalho;
- (10) Adequada forma de seleção e a admissão dos estudantes.

O amplo estudo estatístico de SOARES (2009) sobre o assunto destacou-se dois problemas das IFES em geral, e bem observados na UFRJ: (i) preenchimento das vagas ociosas nos cursos menos concorridos; (ii) evasão e retenção, mesmo nos cursos mais concorridos. Portanto, o aumento do número de vagas não é o fundamental e sim promover medidas para minimizar os dois problemas citados acima.

Outro ponto de discussão é a forma de ingresso nas IFES, baixas notas nos processos de ingresso (vestibular, ENEM ou convênios), cotas e opção por cursos onde a relação candidato/vaga é baixa. Não encontramos estudos que relacione a forma de ingresso na IFES com a evasão, mas sabe-se (conforme relatado no item 2.2.1) que uma das causas da evasão e retenção, principalmente nos primeiros semestres letivos, é devido à falta de condições de atender as exigências acadêmicas do curso. Em alguns casos, a reprovação nos primeiros períodos deve-se a deficiências oriundas do ensino médio, nos cursos de Engenharia estes estudantes têm dificuldades de acompanhar as disciplinas de Cálculo e Física (SOARES, 2009).

Muitas IFES reconhecem que possuem altas taxas de evasão e de retenção em alguns de seus cursos de graduação. No entanto, expressam dificuldades para determinar uma política adequada para reduzir estes índices (CAMPELLO, LINS, 2008). Embora as causas da evasão e retenção sejam questões difíceis de definir, as consequências e seus efeitos são bem perceptíveis.

Em CHRISPIM e WERNECK (2003), o curso de Engenharia de Produção da Universidade Federal de Juiz de Fora – UFJF foi analisado. Um dos argumentos para o insucesso dos estudantes está na fase inicial do curso, às disciplinas básicas iniciais geram desinteresse e desmotivação, promovendo altos índices de evasão. A sugestão para o problema é a (re)adequação da grade curricular, colocando disciplinas específicas logo no primeiro semestre letivo, promovendo a motivação dos discentes e, consequentemente, reduzindo a evasão.

SOARES (2000, 2006, 2009) realizou um estudo na Escola Politécnica da UFRJ, o autor apresentou análise da evasão e retenção nos cursos de Engenharia, uma das causas da evasão deve-se a conflitos vocacionais quando o estudante identifica que sua escolha

pelo curso não corresponde as suas expectativas ou não foi sua primeira opção. No período em que foi feito o estudo a forma de ingresso era por meio de vestibular, os estudantes de segunda e terceira opção eram mais propensos a evasão e retenção. Ele destacou que a escolha precoce por uma especialização dentro da Engenharia propicia a evasão dos estudantes imaturos. Por outro lado, os cursos que oferecem a unificação dos módulos básicos das Engenharias retardando a escolha da especialidade propiciam a evasão dos estudantes que já estão firmes da sua escolha.

Estudantes em situação de retenção são potenciais para evasão (DIAS, 2009, LODER, 2011). Estes estudantes também representam investimentos sem retorno porque ocupam vagas, são contabilizados nos custos e investimentos da instituição, mas não se formam. Logo, haverá o constrangimento de serem jubilados pela universidade por ultrapassar o período de integralização do curso.

2.2.3 Abordagem do Problema sob a Perspectiva da Sociedade e do País

Embora o problema se desdobre em tantos custos para a sociedade, parece estar sendo minimizado ou não se sabe a quem cobrar providências. As afirmativas de CAMPELLO e LINS (2008) mostram isto:

*“O que se constata é que poucas pesquisas têm sido realizadas no sentido de elucidar as razões que levam estudantes em todo o Brasil a **abandonar** um curso de graduação ou mesmo a **postergar** a data de sua formatura. Faz-se, portanto, urgente que se estabeleça uma sistemática de avaliação que permita diagnosticar esta situação em diferentes regiões do país.”* e

“A retenção também apresenta seus impactos negativos, ao não permitir que profissionais de nível superior venham a atuar nas suas respectivas áreas do conhecimento no prazo inicialmente previsto.”

A falta de formação de profissionais qualificados no mercado para enfrentar os desafios do mundo contemporâneo acarreta problemas para economia do país, sem pessoal capacitado para desenvolver novas tecnologias setores como as empresas, indústrias, agricultura e prestação de serviços tornam-se menos produtivos e menos competitivos (SILVA FILHO, 2007, CAMPELLO, LINS, 2008, LOBO, 2011).

Segundo LOBO (2011), estudar o fenômeno da evasão deveria ser uma política

governamental voltada para garantir a qualidade acadêmica e uso dos recursos (públicos e privados). Exemplos de iniciativas de discutir o problema e apontar soluções são encontrados nos Estados Unidos, há uma década 14 mil sítios disponibilizavam dados, diretrizes e resultados dos programas do governo e das universidades sobre evasão.

O governo federal criou um Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais - REUNI (MEC, 2007) que estabeleceu como uma de suas metas a redução das taxas de evasão dos cursos de graduação das universidades públicas. O documento também apresenta diversas considerações econômicas e sociais relacionadas ao problema da evasão e retenção no ensino superior brasileiro.

O portal da ANDIFES (2012) mostra algumas considerações importantes discutidas pelo MEC com as instituições públicas e privadas. Um dos temas abordados foi os altos índices de evasão no ensino superior brasileiro, a fim de evitar que vagas financiadas pelo recurso público fiquem ociosas. As principais causas apontadas foram à decepção do estudante com o curso escolhido e a falta de condições financeiras ou acadêmicas para acompanhar o ritmo das aulas. A Universidade de Brasília (UnB) se propôs a apoiar o MEC e implantar o REUNI. O estudo de BRITO (2013) mostra o processo de implantação do REUNI na UnB, o levantamento das taxas de evasão mostra como média geral dos cursos de graduação no período de 2002 a 2006 em torno de 35,5%, em casos específicos, alcança o teto de 74,1%. O estudo também ressalta a falta de pesquisas no Brasil sobre as causas da evasão e os fatores que influenciam as taxas de diplomação na educação superior. O estudo mostrou as seguintes médias para cada tipo de saída, no período 2002 a 2010: 31,47% são desligados da UnB, em razão de problemas com o rendimento; 30,95% são desligados da instituição, em razão de abandono; 22,50% são desligados da entidade por diversos outros motivos; 15,08% são desagregados da universidade por desligamento voluntário. Com exceção do desligamento por rendimento, não há muito detalhes sobre os outros motivos da evasão entre os estudantes.

Algumas questões relacionadas à evasão passam pela responsabilidade dos governos na adoção de certas políticas públicas e pela responsabilidade das IFES, como sugere (LOBO, 2011):

- (1) *Baixa qualidade da educação básica brasileira* – deficiência de conteúdos básicos interfere no entendimento de conteúdos mais complexos;
- (2) *Baixa eficiência e formação no ensino médio* – o estudante sai do ensino

médio sem autonomia e competências, não há uma mudança de mentalidade para perceber que está se capacitando para ser um profissional;

- (3) *Limitação das políticas de financiamento ao estudante* – inclusive para estudantes das IFES, muitos deixam de estudar por falta de condições financeiras para se manter no curso;
- (4) *Escolha precoce da especialização profissional* - a estrutura do ensino brasileiro permite um excesso de especializações, mas de 200 tipos de especialização para a graduação em Engenharia. Muitas das especializações atendem a um mercado de trabalho muito restrito;
- (5) *Dificuldade de mobilidade estudantil* – dificuldades de transferências de curso entre as universidades brasileiras e estrangeiras. O Brasil ainda não unificou os currículos com as instituições estrangeiras, como estão fazendo os países desenvolvidos;
- (6) *Exigências para autorização e reconhecimento de curso* – muitos estudantes entram em um curso sem o conhecimento se ele é reconhecido pelo MEC, muitas vezes seu tempo e dinheiro são desperdiçados ou seu diploma não é considerado válido;
- (7) *Falta de empenho para combater a evasão* – muitas IFES não possuem políticas de combate a evasão, pois consideram que o estudante não permaneceu no curso porque é academicamente fraco ou não se dedicou o suficiente, portanto, não é um problema da IFES;
- (8) *Escolha dos docentes* – muitos docentes não têm formação pedagógica adequada, eles se sentem acomodados devido a estabilidade, desvalorização da função de docência, cobrança de desempenho com valorização exclusiva da produção científica.

Além disso, um menor número de graduandos sendo formados compromete o ingresso nos cursos de mestrado e doutorado e, conseqüentemente, a produção científica no Brasil. Isto ocorre, principalmente, nos cursos de pós-graduação *stricto sensu* das áreas tecnológicas (SILVA FILHO, 2007).

Como ressalta SOARES (2009) muitos pontos têm que ser discutidos para que os índices de evasão e retenção desejados pelo governo possam ser alcançados. Somente o aumento do número de vagas nas universidades não garante a diplomação de novos profissionais para o mercado de trabalho, enfatiza-se a necessidade manter os estudantes até a sua completa formação e diplomação, combatendo os altos índices de evasão e

retenção, principalmente nas habilitações de menor demanda.

Existe um custo para formar um estudante em uma universidade pública. Independentemente se a vaga está ocupada, as universidades precisam manter a infraestrutura, laboratórios, material de ensino, bibliotecas e pagamento de professores e funcionários. Os dados do Censo da Educação Superior mostram ampliação no número de vagas no ensino superior, mas a taxa de graduação alcança índices muito abaixo do esperado pelo programa REUNI (MEC, 2007, INEP, 2009, 2011, 2012a, SOUZA *et al.*, 2012).

O assunto motiva discussões entre representantes das IFES, governo e sociedade. Algumas tomadas de decisão devem ser à luz do maior número de informações e reflexões sobre os fatores que influenciam a evasão e retenção e suas consequências.

2.3 Contextualização da Trabalho de Tese

O nosso trabalho foi desenvolvido dentro do âmbito da UFRJ, onde há um forte empenho dos gestores acadêmicos para reduzir as causas e as consequências da evasão e retenção nos cursos de graduação. O problema das altas taxas de evasão foi amplamente discutido por vários coordenadores durante o workshop “Pensando na Graduação” (UFRJ, 2009).

Durante o processo de desenvolvimento deste estudo foram feitas várias entrevistas com alguns coordenadores de cursos de graduação e com o diretor da Escola de Engenharia da UFRJ - Escola Politécnica. O objetivo das entrevistas foi discutir detalhes do problema e identificar quais são as expectativas dos gestores com relação a uma ferramenta computacional para auxiliá-los na identificação dos alunos em risco de evasão. O entendimento da complexidade do problema na instituição foi de vital importância para definirmos uma proposta de solução.

De fato, verificamos que os problemas e as considerações apresentados nos itens anteriores são complexos e demandam um estudo de várias equipes multidisciplinares. Além disso, as limitações de acesso as bases de dados demandaram tempo de espera longo e restringiram parte do nosso trabalho. Por exemplo, não tivemos acesso a bases de dados com informações socioeconômicas dos estudantes e não havia um procedimento formal na universidade de entrar em contato com o estudante e registrar quais os motivos que levou a abandonar as suas atividades acadêmicas. Portanto, as condições socioeconômicas no momento do ingresso do estudante e na evasão não serão

analisadas dentro do contexto deste trabalho de tese.

O sistema de gestão acadêmica da UFRJ (SIGA) é um sistema legado que armazena grandes quantidades de dados acadêmicos, mas não possui funcionalidades para processar sistematicamente informações dos estudantes na perspectiva de avaliar o desempenho e prever aqueles que estão em risco de evasão (UFRJ, 2014). A base de dados do SIGA possui uma grande quantidade de dados dos estudantes de graduação. Estes dados podem ser utilizados em análises de dados utilizando EDM. A mineração de dados educacionais EDM aplicados aos dados acadêmicos dos estudantes de graduação possibilitaria a construção de uma abordagem capaz de identificar (predizer) quais estudantes estão potencialmente sujeitos a evasão/retenção nos cursos de graduação. Para auxiliar num diagnóstico mais abrangente e correto, a arquitetura EDM WAVE permite prever o desempenho acadêmico dos estudantes ao longo de todos os semestres letivos, permitindo que gestores acadêmicos possam tomar decisões e reverter o número de evasões nos cursos de graduação.

Recentemente, SOUZA *et al.* (2012) afirma que poucos são os trabalhos científicos no Brasil que propõem a utilização da tecnologia da informação para auxiliar na gestão do problema nas IFES. Enfim, a complexidade do tema evidencia a necessidade dos gestores acadêmicos buscarem ferramentas computacionais que os auxiliem a traçar estratégias que promovam a permanência dos estudantes no ensino superior. Portanto, identificar precocemente os potenciais estudantes sujeitos a evasão e/ou retenção é uma das primeiras necessidades que os gestores acadêmicos das IFES precisam para direcionar medidas de mitigação do problema.

Um ponto importante na identificação precoce dos estudantes sujeitos a evasão e/ou retenção é o tempo que o estudante leva para tomar a decisão, há alguns indícios que podem ser observados como, por exemplo, abandono de disciplinas, notas baixas, menor frequência, trancamento do período, entre outros. Nos trabalhos de SOUZA (2008) e MELLO (2012), foram realizadas entrevistas com os estudantes evadidos. Eles apontaram que tinham consciência do motivo que os levou a abandonar o curso, ou seja, todos foram claros em afirmar que sua decisão não foi repentina ou abrupta.

Portanto, o nosso estudo identificou os seguintes questionamentos:

- (1) Há uma maneira de identificar potenciais estudantes sujeitos a evasão ou retenção ao longo do curso?
- (2) Há formas de identificar diferenças entre estudantes que conseguem concluir o curso, daqueles que permanecem matriculados fora do prazo de conclusão,

e por fim, daqueles que abandonam os cursos de graduação?

- (3) Sistemas informatizados baseados em EDM podem ajudar na gestão acadêmica, em particular na predição de estudantes em risco de evasão e retenção?
- (4) Estudantes de diferentes cursos da UFRJ possuem o mesmo comportamento se considerarmos as três condições: concluintes, ativos fora do prazo e evadidos?

Por fim, os gestores acadêmicos da UFRJ podem elencar outros questionamentos além dos relacionados acima. A base de dados do SIGA possui diversas informações acadêmicas sobre os estudantes da UFRJ. Esta tese considera que a investigação dessas bases de dados utilizando técnicas de mineração de dados educacionais poderá permitir a identificação precoce do desempenho acadêmico dos estudantes de graduação da UFRJ, contribuindo para que os gestores acadêmicos identifiquem os estudantes que apresentam risco de evasão do sistema ou retenção. A utilização da arquitetura EDM WAVE acoplada ao sistema de gestão acadêmica SIGA permitirá que gestores acadêmicos não familiarizados com as tecnologias de mineração de dados educacionais possam acompanhar o desempenho acadêmico dos graduandos e detectar aqueles com problemas em cumprir com as exigências curriculares do curso de graduação.

2.4 Trabalhos Correlatos em Diversos Níveis de Aplicação

Os trabalhos correlatos apresentam o problema do insucesso dos discentes em diversos contextos e níveis de abrangência. A maioria deles utiliza uma perspectiva de análise estatística para analisar os dados. Poucos trabalhos propõem a utilização de alguma tecnologia da computação como parte da solução do problema. No entanto, antes de apresentar os trabalhos correlatos que utilizam EDM, distinguiremos três diferentes níveis de abordagem encontrados na literatura:

- (1) Na *disciplina ou conjunto de disciplinas em um período de tempo específico* - neste caso, trata-se da análise das reprovações e do abandono que podem ocorrer em uma ou mais disciplinas específicas. Esta análise limita-se a um contexto muito reduzido, pois a grade curricular de um curso de graduação é composta por diversas disciplinas e eixos temáticos. Os principais autores que investigaram esta linha são (HAMALAINEN *et al.*, 2004, MINAEI-BIDGOLI *et al.*, 2006);

- (2) *No curso de graduação* - trata-se da evasão e/ou retenção que podem ocorrer em um curso de graduação de uma IES. Os principais autores que investigaram esta linha são (BARROSO, FALCÃO, 2004, SOARES, 2000, 2006, 2009);
- (3) *Na IES ou em todas as IES do país* – trata-se do estudo macro do problema a nível institucional ou nacional. Os principais autores que investigaram esta linha são (HERZOG, 2005, SILVA FILHO, 2007, INEP, 2009).

Esta tese apresenta análises nos níveis (2) e (3). No capítulo 4, apresentaremos diversos experimentos sobre diferentes cursos de graduação e uma análise quantitativa identificando as diferentes características dos graduandos da UFRJ.

2.5 Trabalhos Correlatos em EDM

A mineração de dados é um importante componente da Descoberta de Conhecimento em Dados (*Knowledge Discovery from Data - KDD*) (HAN, KAMBER, 2006), estes conceitos estão relacionados à descoberta de informações potencialmente úteis a partir de grandes quantidades de dados. A mineração de dados tem sido aplicada em um grande número de domínios do conhecimento. Nos últimos anos, tem havido um interesse crescente no uso de mineração de dados para investigar questões científicas no âmbito da educação, denominada mineração de dados educacionais (*Educational Data Mining – EDM*). Configura-se como uma área de pesquisa que investiga o desenvolvimento de métodos para fazer descobertas de conhecimento utilizando dados oriundos de contextos educativos (BAKER, YACEF, 2009, BAKER, 2010, ROMERO, VENTURA, 2010, PAIVA, 2014).

A utilização da mineração de dados dentro do contexto educacional motivou a formação de uma comunidade de pesquisa (BAKER, YACEF, 2009, BAKER *et al.*, 2011). O periódico publicado pela comunidade acadêmica mostra várias publicações e discute diversos temas de pesquisa em *Educational Data Mining* (EDM), entre eles o estudo sobre os fatores associados ou relacionados ao insucesso dos estudantes nos cursos de graduação. Como em outras áreas de pesquisa, a comunidade EDM se esforça para criar repositórios de dados disponibilizados pelas universidades a fim de ter dados consistentes para desenvolver aplicações na área de educação. Em BAKER e YACEF (2009), os autores comentam a escassez de contribuições explorando EDM na América do Sul.

Um artigo do tipo *survey* escrito por ROMERO e VENTURA (2010) descreve os mais relevantes documentos relacionados à EDM. O artigo introduz os principais conceitos relacionados à mineração de dados aplicada ao cenário educacional. Os conceitos relacionados à EDM não são muito diferentes da mineração de dados em outros contextos. A mineração de dados depende de diversas áreas de pesquisa, as quais contribuem para aprimorá-la, tais como, estatística, inteligência artificial, visualização, aprendizado de máquina e banco de dados (HAN, KAMBER, 2006). Em mineração de dados diversas técnicas podem ser utilizadas para solucionar problemas em diversos contextos (HAN, KAMBER, 2006, WITTEN, FRANK, 2005, WITTEN *et al.*, 2011). Portanto, os autores supracitados defenderam a definição da EDM como uma linha de pesquisa, dentro da grande área da mineração de dados, direcionada a utilizar técnicas de mineração de dados aplicados a diferentes tipos de dados educacionais.

O artigo de ROMERO e VENTURA (2010) enfatiza a necessidade de se projetarem e criarem novas ferramentas apoiadas em EDM que facilitem a integração entre o conhecimento educacional, dados de ambientes educacionais, e mineração de dados. Particularmente, um dos temas mencionados é a necessidade de aplicar EDM na predição do desempenho dos estudantes. **De forma apropriada, os autores definem a predição como o ato de estimar o valor desconhecido de uma variável que descreve o estudante.**

Em 2011, BAKER *et al.* (2011) publicaram o excelente artigo “Mineração de Dados Educacionais: Oportunidades para o Brasil” publicado na Revista Brasileira de Informática na Educação. Na ocasião os autores mencionaram que a maior concentração das pesquisas está sendo realizada em instituições estrangeiras e faltam trabalhos relacionados no Brasil. No final de 2011, nós iniciamos a publicação de uma série de trabalhos relacionados ao estudo da predição do desempenho de graduandos utilizando EDM (MANHÃES *et al.*, 2011, 2012, 2014a, 2014b, 2014c, 2014d, 2015). Nossa pesquisa e trabalhos utilizando EDM são um dos primeiros no Brasil a abordar uma recente área de pesquisa em mineração de dados.

Basicamente, os trabalhos utilizando EDM seguem duas linhas de pesquisas principais, dividindo-se em: (i) identificar atributos relevantes para caracterizar estudantes ou (ii) identificar e comparar o desempenho de algoritmos classificadores. A primeira linha está relacionada em utilizar a EDM para identificar, em base de dados educacionais, os atributos mais relevantes para caracterizar os grupos de estudantes. A segunda está relacionada ao estudo e seleção dos algoritmos mais apropriados para

solução do problema, estimando o desempenho quando aplicados a dados educacionais.

2.5.1 Trabalhos Direcionados a Identificar os Atributos Relevantes para a Caracterização dos Estudantes

Nesta subseção, descrevemos os artigos mais relevantes na análise dos atributos e identificação das características dos estudantes.

Em HAMALAINEN *et al.* (2004), duas disciplinas de programação oferecidas na modalidade *online* em um curso de Ciência da Computação foram analisadas. Estas disciplinas apresentavam altos índices de reprovações e abandono. Os autores utilizaram regras de associação e modelos probabilísticos para identificar as características mais importantes para predizer os resultados finais nas duas disciplinas.

A Universidade *Michigan State* desenvolveu um sistema educacional online denominado LON-CAPA (*Learning Online Network with Computer-Assisted Personalized Approach*), os autores utilizaram regras de associação para extrair padrões de dados, a fim de identificar atributos e valores que caracterizam os grupos de estudantes. A seguir, verificaram se existe associação entre os atributos dos estudantes e a situação de aprovação ou reprovação na disciplina de Física oferecida neste ambiente (MINAEI-BIDGOLI *et al.*, 2004a, 2004b, 2006).

Em (CAMPELLO, LINS, 2008), os autores estudaram a evasão e retenção de estudantes no curso de graduação em Engenharia de Produção da Universidade Federal de Pernambuco (UFPE). Os autores utilizaram dados socioeconômicos, avaliações do vestibular, histórico escolar, entre outros. Eles analisaram um conjunto de 280 estudantes que ingressaram no curso entre 2000 a 2006, foram considerados 136 (48,6%) estudantes evadidos/retidos. Os autores mencionaram a utilização de mineração de dados, o trabalho foi conduzido no sentido de identificar os agrupamentos (*clusters*) de estudantes, segundo os autores cada agrupamento mostrou evidências das causas de evasão/retenção.

A dissertação de SOUZA (2008) buscou através da utilização de regras de associação e árvore de decisão demonstrar que há um padrão que descreve os estudantes que evadiram dos cursos de engenharia da Universidade Federal Fluminense (UFF). A autora utilizou dados acadêmicos dos estudantes da UFF, o banco de dados utilizado foi Oracle e a ferramenta de mineração foi *Oracle Data Miner*. Esta ferramenta de mineração ainda não é muito explorada no meio acadêmico e é restrita em termos de

documentação. A autora trabalhou com o conjunto de estudantes cancelados e as disciplinas de maior reprovação. Segundo a pesquisa, a autora mostrou as disciplinas mais relacionadas a evasão dos cursos de graduação em engenharia da UFF.

A tese de KAMPFF (2009) desenvolvida na Universidade Federal do Rio Grande do Sul (UFRGS) aplica técnicas de mineração de dados aos dados de estudantes gerados pela interação em um Ambiente Virtual de Aprendizagem (AVA), o objetivo era identificar comportamentos e características de estudantes com risco de abandono ou reprovação, o trabalho utilizou a extração de regras. O experimento utilizou dados de uma mesma disciplina on-line, coletados durante várias edições, totalizando 1564 estudantes. A tese não focou na apresentação dos resultados obtidos pelos algoritmos, seu objetivo era descrever um sistema de alerta para os professores da disciplina.

2.5.2 Trabalhos Direcionados a Identificar e Comparar o Desempenho dos Algoritmos

Em MINAEI-BIDGOLI e PUNCH (2003) utilizaram uma combinação de algoritmos classificadores com o objetivo de prever as notas finais de uma disciplina do curso on-line LON-CAPA. O experimento utilizou os resultados de 12 exercícios de casa realizados por 261 estudantes inscritos na disciplina online de Física. Os autores demonstraram que a combinação de múltiplos classificadores obteve maior acurácia (86,8%) em comparação com as acurácias medidas individualmente, o experimento utilizou validação cruzada com 10 conjuntos (*10-fold cross validation*).

Os autores KOTSIANTIS *et al.* (2003) apresentaram um estudo comparativo entre seis algoritmos de aprendizado de máquina, objetivando encontrar o mais apropriado para prever o abandono dos estudantes na disciplina “Introdução à Informática” de um curso na modalidade EAD na universidade *Hellenic Open University*, na Grécia. Neste trabalho, os autores analisaram 350 estudantes utilizando dados pessoais, participação nas atividades e notas nas avaliações. As seguintes técnicas de aprendizagem de máquina foram utilizadas: árvore de decisão, redes neurais artificiais, classificador *Naive Bayes*, aprendizagem baseada em instâncias, análise de regressão logística e *Support Vector Machine* (SVM). A conclusão do experimento mostrou que não havia diferença estatística entre os algoritmos estudados, mas o algoritmo classificador *Naive Bayes* foi ligeiramente melhor comparando com os demais.

O artigo de HAMALAINEN e VINNI (2006) apresentou um trabalho realizado na

University of Joensuu, Finlândia. Os autores analisaram dados de duas disciplinas ministradas no programa de Educação à Distância (EAD) do curso de Ciência da Computação. Os dados foram coletados durante dois anos letivos, sendo analisados 125 estudantes da disciplina de “Programação I” e 88 estudantes da disciplina de “Programação II”. A base de dados constituiu-se de atributos correspondentes a resultados de exercícios e notas finais. Foram comparados 5 classificadores: regressão linear, *Support Vector Machine* (SVM), *Naive Bayes*, redes bayesianas (*Bayesian networks*) e *Bayesian multinets*. O experimento utilizou validação cruzada de 10 conjuntos. Os autores concluíram que o classificador *Naive Bayes* foi melhor para prever potenciais estudantes desistentes nas disciplinas.

SUPERBY *et al.* (2006) analisaram o desempenho acadêmico dos estudantes das universidades belgas de língua francesa (*Belgian French Speaking Universities*). A pesquisa constituía em identificar no primeiro ano acadêmico estudantes em risco de falhar nos estudos ou abandonar, estabeleceram-se três níveis de risco: baixo, médio e alto. Foram coletados dados de 533 estudantes entre os anos 2003 a 2004. A análise para classificar os estudantes nos três grupos de risco foi feita utilizando os algoritmos: redes neurais (*neural network*), *random forests* e árvore de decisão (*decision tree*). O artigo utilizou questionários respondidos pelos estudantes das várias universidades envolvidas no estudo, os autores especularam que os resultados não foram satisfatórios porque a análise foi realizada com vários estudantes de diferentes universidades.

O trabalho de RUSLI *et al.* (2008) descreveu a utilização de três algoritmos preditivos para analisar o desempenho dos estudantes na *Faculty of Information Technology and Quantitative Sciences* na *Universiti Teknologi MARA* na Malásia. Os algoritmos utilizados foram: regressão logística, redes neurais artificiais e neuro-fuzzy. Os resultados mostraram que o sistema neuro-fuzzy foi que apresentou melhor resultado. O estudo foi feito com 393 estudantes e as variáveis utilizadas estão diretamente ligadas às formas de ingresso na universidade.

O artigo de GARCIA *et al.* (2009) descreve uma ferramenta de mineração de dados colaborativa baseada em regras de associação para auxiliar professores de cursos online a compartilhar e avaliar as informações descobertas. Os dados utilizados são *logs* de cursos online, os algoritmos de regras de associação são utilizados para descobrir relações entre os dados, os professores avaliam estas regras e utilizam um sistema de pontuação segundo a importância das mesmas. Mais detalhes sobre a ferramenta de mineração de dados KEEL (*Knowledge Extraction based on Evaluatory Learning*)

(ALCALÁ-FDEZ *et al.*, 2009). A ferramenta proposta pelos autores é similar a ferramenta de mineração de dados Weka (HALL *et al.*, 2009, BOUCKAERT *et al.*, 2010).

Um trabalho mais abrangente foi realizado por DEKKER *et al.* (2009) no departamento de Engenharia Elétrica da *Eindhoven University of Technology* da Holanda. Os autores aplicaram técnicas de mineração de dados para identificar estudantes que abandonaram ou reprovaram no primeiro ano de graduação do curso de Engenharia Elétrica (DEKKER *et al.*, 2009). Eles utilizaram dados pré-universitários dos estudantes com atributos informando: tipo de curso feito anteriormente, notas obtidas nas disciplinas Ciências, Matemática e outras. Os dados posteriores ao ingresso dos estudantes foram às notas obtidas em três exames parciais. Este trabalho foi feito com dados de 648 estudantes entre os anos de 2000 e 2009. Os autores testaram os algoritmos: árvore de decisão, classificadores bayesianos, regras de associação. Segundo os pesquisadores o classificador baseado em árvore de decisão obteve melhor resultado.

Em sua tese de doutorado HUANG (2011) utilizou um conjunto de modelos matemáticos (estáticos e mineração de dados) para predição do desempenho acadêmico dos estudantes na disciplina “*Engineering Dynamics*” na *Utah State University*, Estados Unidos. Nos quatro semestres de 2008 a 2011 foram coletados dados de 323 estudantes. As variáveis utilizadas foram às notas das disciplinas que são pré-requisitos a esta disciplina e as notas dos exames parciais da disciplina. Quatro técnicas de modelagem matemática foram utilizadas: regressão linear múltipla, rede neural, redes RBF e SVM. O autor mostrou que as quatro técnicas modelagem matemática apresentaram uma média de acurácia na predição superior a 80%. O autor relata que dados do conhecimento prévio dos estudantes influenciaram significativamente na acurácia do modelo de predição do desempenho acadêmico dos estudantes. O resultado obtido é uma predição da nota na disciplina. O autor utilizou os pacotes de software comerciais SPSS 18 e o MATLAB para executar os algoritmos dos modelos matemáticos investigados.

O trabalho de ZAFRA *et al.* (2011) propõe uma nova representação baseado em *Multiple Instance Learning* (MIL) para melhorar a eficiência em predizer o desempenho acadêmico de estudantes em Ambiente Virtual de Aprendizagem (AVA) em comparação com a utilização dos algoritmos clássicos. Os autores mostraram que para os dados de estudantes que utilizaram o sistema AVA da *Cordoba University*, Espanha, o sistema

MIL foi mais apropriado.

O trabalho desenvolvido por TONTINI *et al.* (2011) em uma IES da rede privada brasileira utilizou como base de dados às respostas de um questionário aplicado aos estudantes dos cursos de graduação. O questionário pedia uma avaliação da infraestrutura da instituição, qualidade do curso, expectativa profissional dos estudantes com relação à futura profissão e informações socioeconômicas. O autor considerou as respostas de 300 estudantes evadidos entre 2009/1 e 2009/2, e mais 300 estudantes escolhidos aleatoriamente que permaneceram matriculados na instituição. A análise das respostas para identificar estudantes em risco de evasão seguiu os procedimentos: método estatístico, análise de agrupamentos (*clustering*) e redes neurais artificiais RBF.

CHEEWAPRAKOBKIT (2013) investigou os fatores que afetam o desempenho acadêmico, ele analisou dados acadêmicos e pessoais dos estudantes de um programa internacional. O conjunto de dados analisados possui 1600 registros com 22 atributos entre os anos de 2001 a 2011 na universidade da Tailândia. Os autores utilizaram dois classificadores e utilizaram a validação cruzada com 10 conjuntos para conduzir o experimento. Os resultados experimentais mostraram que o classificador árvore de decisão obteve acurácia de 85% e o classificador redes neurais obteve acurácia de 84%.

2.5.3 Trabalhos Relacionados Utilizando Métodos Estatísticos e/ou Outras Análises

Conforme enunciado anteriormente, a deficiência em manter o discente na universidade é um problema antigo e motivo de preocupação em muitas universidades. Um estudo feito na Escócia no departamento de Matemática da *Napier University* mostrou que mais de um quarto dos novos estudantes que entram nos cursos de graduação são reprovados ou abandonam (JOHNSTON, 1997). Este trabalho mostrou vários aspectos do problema. Particularmente, destacou as dificuldades de obtenção de dados adicionais sobre os motivos que levam os estudantes a não conseguirem progresso, lá também não há um acompanhamento após a saída do estudante. O trabalho envolveu a criação de questionários e realizou entrevistas com estudantes, professores, coordenadores de curso e outros membros da universidade. Durante as entrevistas, houve desacordo por ambas as partes sobre as causas do *insucesso*, visto que os questionários foram feitos sob as perspectivas da universidade e não dos estudantes. Os dados foram coletados em 1994 e 1995. A partir dos questionários foi feito um

levantamento estatístico, as conclusões da pesquisa revelam que a grande maioria dos estudantes que abandonam ou tem problemas estão concentrados no primeiro ano acadêmico. A pesquisa sugere que os problemas não acadêmicos podem contribuir mais para o fracasso do estudante do que problemas acadêmicos e a gama dos problemas não acadêmicos é ampla e complexa. Além disso, a percepção pessoal dos membros das universidades sobre o grau de influência exercido por esses problemas não foi consensual. O artigo menciona que a predição incorreta da situação final do estudante (falso positivo e falso negativo) afeta a orientação tanto de professores quanto de estudantes. Existem grupos de estudantes que apresentaram bons resultados acadêmicos, mas não retornam no segundo ano do curso. O estudo sugere ações da instituição para reverter o quadro principalmente aplicando-as ao primeiro ano acadêmico.

Os trabalhos de MOORE (1995) e DAVIES (1997) coletaram dados através de entrevistas e identificaram um grande número de fatores que influenciam no desempenho dos estudantes. MOORE (1995) mostra que o insucesso dos estudantes é influenciado por um amplo número de fatores. Entre eles estão a antipatia pelo curso escolhido ou consideram o curso inadequado por razões pessoais e acadêmicas. DAVIES (1997) comparou as expectativas dos estudantes que eram bem sucedidos com os não bem sucedidos, ele concluiu que nenhum dos grupos pode ser diferenciado em termos de sua aparente motivação e da importância do curso para suas vidas. No entanto, os grupos se diferenciam em sua relativa satisfação com vários aspectos da universidade. O grupo de estudantes com problemas mostrou insatisfação com a qualidade do ensino e suporte as atividades das aulas. Problemas pessoais e financeiros foram menos relevantes. A conclusão de DAVIES (1997) mostrou que os estudantes com problemas falham pela falta de qualidade e assistência nos estudos nas salas de aula. Este trabalho ressalta que é responsabilidade da instituição e dos gestores acadêmicos, que coordenam os cursos, melhorar a interação entre a instituição e os estudantes para que estes obtenham êxito.

2.6 Repositórios de Base de Dados Educacionais

A utilização de técnicas de mineração de dados depende, essencialmente, da qualidade da base de dados disponíveis. Em (CASTRO *et al.*, 2007, BAKER, YACEF, 2009), os autores mencionam a necessidade de criar repositórios de dados educacionais para que os experimentos possam ser reproduzidos por outros pesquisadores. Alguns

repositórios públicos de dados educacionais são encontrados na Universidade de Pittsburgh, no *Pittsburgh Science of Learning Center (PSLC)* (PSLC, 2010). O grupo “*Education Group at the World Bank*” (EWB) possui vários dados estatísticos sobre educação, mas nenhuma base sobre evasão escolar na graduação (EWB, 2009). A *Organisation for Economic Cooperation and Development - OECD*, ligado ao *Programme for International Student Assessment – PISA* possui várias bases de dados com informações socioeconômicas de estudantes e instituições o público alvo são estudantes entre 15 e 16 anos de 60 países diferentes.

2.7 Conclusões

O estudo sobre o insucesso dos estudantes na graduação é tratado em diversas universidades em todo o mundo. Governo, sociedade, IES e IFES mostram grande preocupação com o assunto, pois os índices de evasão e/ou retenção são muito altos. Os custos pessoais e financeiros são difíceis de calcular.

O sucesso do discente é quantitativamente definido como a diplomação ou formação no curso de graduação. Por outro lado, vários termos são utilizados para definir o insucesso, sendo evasão e retenção os adotados pelo MEC, e utilizados neste trabalho.

Vários trabalhos investigaram o problema em diversas universidades brasileiras, destacamos que o problema pode ser analisado sob três principais perspectivas: estudante, instituição e sociedade (país).

O insucesso dos estudantes também é tratado dentro de diversos contextos e níveis de abrangência: (i) na disciplina ou conjunto de disciplinas em um período de tempo específico; (ii) no curso de graduação; e (iii) na IES ou em todas as IES do país.

Dentre os trabalhos avaliados, verifica-se que são aplicados a pequenos contextos e apresentam algumas limitações, como: (i) consideram um pequeno número de disciplinas; (ii) observam dados que refletem pequenos intervalos de tempo; e (iii) aplicam as técnicas de mineração de dados a um número reduzido de estudantes.

Entre os trabalhos relacionados neste capítulo, apenas as pesquisas apresentadas por DEKKER *et al.* (2009) e por CAMPELLO e LINS (2008) na IFES brasileira UFPE, avaliaram curso de graduação em Engenharia.

Os dados são essenciais para análise e definição do problema que acontece dentro do contexto das universidades, portanto há diversos procedimentos para coletar dados e definir o problema. As fontes de dados podem ser questionários, entrevistas, base de

dados pessoais e acadêmicos.

As comunidades acadêmicas estimulam a criação de repositórios de dados sobre educação. No entanto, nenhum dos trabalhos mencionados em (CASTRO *et al.*, 2007, BAKER, YACEF, 2009) e disponibilizam bases de dados com informações sobre insucesso dos estudantes na graduação. Como relatado em (BAKER, YACEF, 2009, BAKER *et al.*, 2011) a utilização de técnicas de mineração de dados aplicada a educação ainda é um assunto muito recente, há ainda dúvidas de quais dados (atributos) devem ser utilizados e quais técnicas de mineração de dados são mais adequadas.

A EDM é uma parte da mineração de dados que utiliza dados educacionais, estudos mais recentes utilizam EDM para tratar problemas que envolvem dados educacionais e estudos anteriores utilizam métodos estatísticos ou outras formas de análise dos dados.

Os autores ROMERO e VENTURA (2010) apresentaram os pontos em aberto e que precisam ser explorados em EDM: (i) desenvolvimento de ferramentas de EDM para educadores e gestores acadêmicos que não são peritos em mineração de dados; (ii) as operações de pré-processamento das informações, facilidades de configurações dos algoritmos e interpretação dos resultados dos algoritmos estão a parte do interesse dos educadores, por isso a necessidade de criação de ferramentas mais genéricas, configuráveis e de simples manipulação; (iii) não há ferramentas de EDM que possam ser reutilizadas em qualquer sistema educacional, em especial no contexto das IFES brasileiras; e (iv) não há uma padronização para entrada de dados e resultado dos modelos obtidos, após as fases de pré-processamento, mineração de dados e pós-processamento dos dados educacionais. Os autores também enfatizam a necessidade de criar ferramentas de mineração de dados que integre o domínio do conhecimento educacional a utilização das técnicas de mineração de dados.

A adoção de mineração de dados educacionais para a predição da situação acadêmica é um campo de investigação ainda não consolidado, necessita de investigações mais profundas e complementares tanto na definição dos atributos a serem utilizados quanto nas técnicas de mineração de dados empregadas (CASTRO *et al.*, 2007, BAKER, YACEF, 2009, DEKKER *et al.*, 2009, BAKER *et al.*, 2011). Os autores, em linhas gerais, indicam pontos que precisam ser pesquisados para aprimorar a utilização da mineração de dados na identificação de estudantes com risco de evasão nos cursos de graduação. Os principais pontos são: (i) transformação dos dados (os dados colhidos nem sempre são diretamente tratados pelos algoritmos); (ii) identificar os atributos mais relevantes; (iii) identificar os algoritmos mais adequados e (iv) aplicar

os algoritmos para identificar outros grupos de estudantes. Os itens (i, ii e iii) estão dentro da fase de pré-processamento de dados e o item (iv) está dentro da fase de mineração de dados do KDD (HAN, KAMBER, 2006).

A Tabela 2.1 mostra um comparativo entre a nossa abordagem e os trabalhos relacionados, identificando o alvo de estudo de cada trabalho. A maior parte dos trabalhos analisa o problema do insucesso dos alunos em um curso de graduação (coluna 5) e utilizam métodos quantitativos e qualitativos para analisar os dados (coluna 6). Estes trabalhos não mencionam a utilização de recursos computacionais para analisar os dados. Os autores que utilizaram MD, IA e EDM como métodos de análise dos dados (coluna 6) demonstraram a utilização de recursos computacionais. A (coluna 2) apresenta os trabalhos que estudaram os atributos mais relevantes para caracterizar o problema. A (coluna 3) apresenta os trabalhos mais focados na avaliação dos algoritmos e técnicas de DM. A (coluna 4) apresenta os estudos que analisaram o desempenho de estudantes em disciplinas específicas. A (coluna 5) identifica trabalhos que investigaram o problema em um curso de graduação ou em cursos de uma área da engenharia.

Tabela 2.1: Estudos e abordagens dos trabalhos relacionados.

Trabalhos Relacionados	Estudos dos atributos relevantes	Estudos dos algoritmos	Análise do desempenho em disciplinas	Análise do desempenho em curso de graduação	Método de análise dos dados
MOORE, 1995				x	quantitativo e qualitativo
DAVIES, 1997				x	quantitativo e qualitativo
JOHNSTON, 1997				x	quantitativo e qualitativo
SOARES, 2000, 2006, 2009				x	quantitativo e qualitativo
VELOSO, ALMEIDA, 2001				x	estatístico e qualitativo
CHRISPIM e WERNECK, 2003				x	qualitativo
KOTSIANTIS <i>et al.</i> , 2003		x	x		IA
MINAEI-BIDGOLI, PUNCH, 2003		x	x		MD
SARAIVA, MASSON, 2003				x	estatístico e qualitativo
BARROSO, FALCÃO, 2004				x	quantitativo e qualitativo
HAMALAINEN <i>et al.</i> , 2004	x	x	x		MD
MINAEI-BIDGOLI <i>et al.</i> , 2004a, 2004b, 2006	x	x	x		MD

HERZOG, 2005		x		x	MD
HAMALAINEN, VINNI, 2006		x	x		MD
SUPERBY <i>et al.</i> , 2006		x		x	MD
SILVA FILHO, 2007				x	quantitativo e qualitativo
CAMPELLO, LINS, 2008	x			x	MD
RUSLI <i>et al.</i> , 2008		x		x	MD
SOUZA, 2008	x			x	MD
ANDRIOLA, 2009				x	quantitativo e qualitativo
DEKKER <i>et al.</i> , 2009		x		x	EDM
GARCIA <i>et al.</i> , 2009		x	x		EDM
KAMPFF, 2009	x		x		EDM
SOARES, 2009				x	quantitativo e qualitativo
DIAS, 2010				x	quantitativo e qualitativo
HUANG, 2011		x	x		modelos matemáticos e MD
LOBO, 2011				x	quantitativo e qualitativo
LODER, 2011				x	qualitativo
SILVA, 2011				x	quantitativo e análise multicritério
TONTINI <i>et al.</i> , 2011		x		x	quantitativo e qualitativo e MD
ZAFRA <i>et al.</i> , 2011		x	x		EDM
MELLO, 2012				x	qualitativo
CHEEWAPRAKOBKIT, 2013		x		x	EDM
MANHÃES <i>et al.</i> , (2011, 2012, 2014a, 2014b, 2014c, 2014d, 2015)	x	x		x	EDM

Nossa proposta e abordagem excederam todos os demais trabalhos relacionados em volume de dados, complexidade e abrangência de cursos contemplados. Nossa pesquisa abrangeu um estudo detalhado identificando os melhores atributos, avaliação de 12 algoritmos classificadores, analisamos diversos cursos de graduação e apresentamos análises comparativas sobre todos os cursos de graduação da UFRJ. Nos próximos capítulos apresentaremos a arquitetura com base nas fases do KDD e utilizando mineração de dados educacionais (EDM).

3 WAVE: Uma Arquitetura Apoiada em EDM para IFES

Neste capítulo, apresentamos a arquitetura EDM WAVE baseada nos requisitos da EDM enunciados por (BAKER, YACEF, 2009, ROMERO, VENTURA, 2010) e descrevemos os fundamentos teóricos e práticos utilizados para sua concepção. A arquitetura EDM WAVE foi concebida para auxiliar a gestão acadêmica das IFES e permitirá automatizar a predição do desempenho acadêmico dos estudantes a cada semestre letivo. Essa abordagem difere dos trabalhos relacionados não só no que se refere aos métodos computacionais utilizados como também na capacidade de acompanhar o desempenho dos estudantes ao longo de cada semestre letivo.

A proposta foi desenvolvida em sintonia com a linha de pesquisas denominada Descoberta de Conhecimento em Dados (*Knowledge Discovery from Data* - KDD). HAN e KAMBER (2006) definem a mineração de dados como parte do processo de KDD. No entanto, a mineração de dados, especialmente tratada por EDM por utilizar dados gerados a partir de um contexto educacional, também pode ser tratada dentro do KDD sem a necessidade de adaptações relevantes.

Este capítulo está organizado da seguinte forma. Na seção 3.1 serão apresentadas as principais características do processo de KDD e a definição de mineração de dados. Na seção 3.2, discutimos as funcionalidades da mineração de dados. Em seguida, na seção 3.3, descrevemos o processo de construção de um modelo. Posteriormente, na seção 3.4, apresentamos detalhes da arquitetura EDM WAVE. Na seção 3.5 apresentamos as conclusões do capítulo.

3.1 *Descoberta de Conhecimento em Dados*

Nos últimos anos a aquisição de dados não tem sido problema, pois a tecnologia envolvendo a captura dos dados e armazenamento evoluiu consistentemente. As bases de dados institucionais tornaram-se gigantescas. Entretanto, os dados coletados precisam ser analisados de modo que a informação implícita nestas bases de dados possa ser interpretada tornando-se conhecimento para tomada de decisão. Em seu livro,

HAN e KAMBER (2006) apresentaram o processo de Descoberta de Conhecimento em Dados (*Knowledge Discovery from Data - KDD*), que consiste em uma sequência de passos iterativos, eles distinguem a mineração de dados como parte do processo, como mostra a figura 3.1.

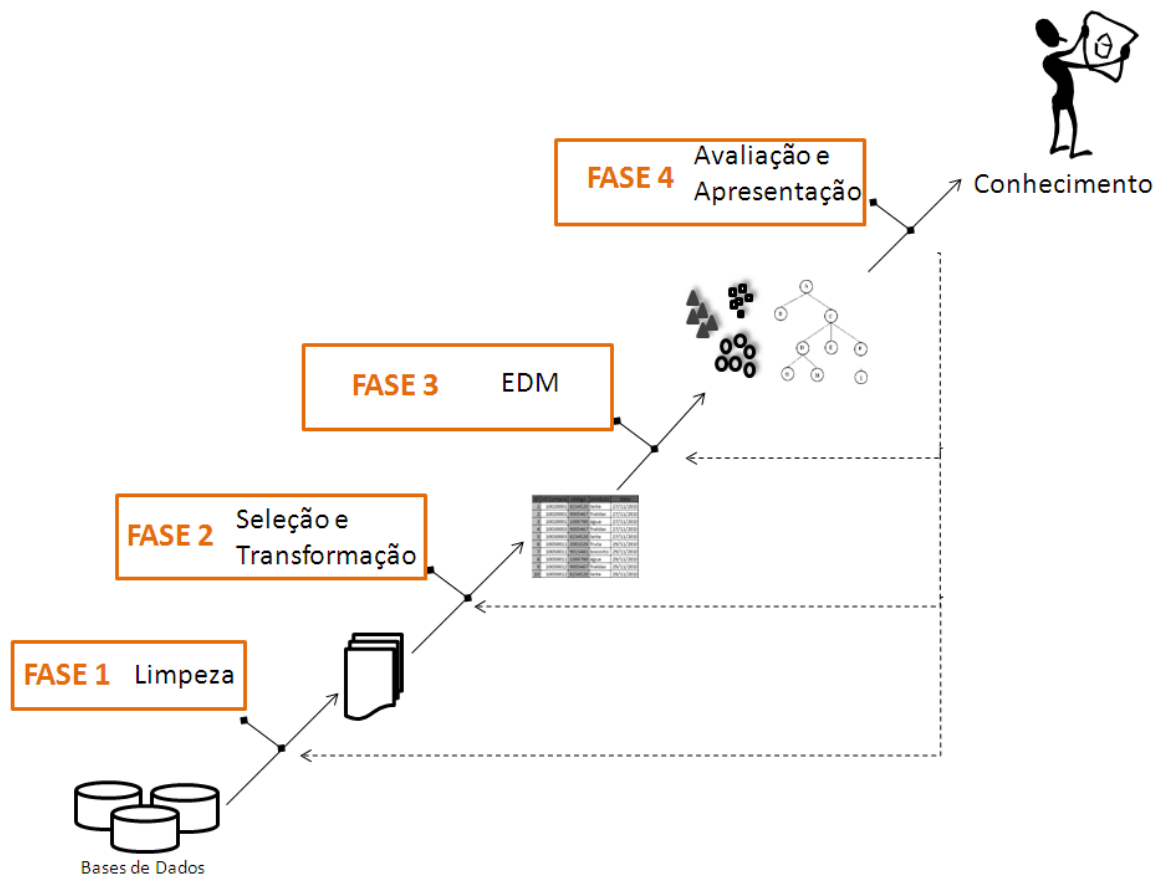


Figura 3.1: Processo de Descoberta de Conhecimento em Dados (KDD) utilizando EDM e as fases de desenvolvimento metodológico adotada nesta tese - adaptação (HAN, KAMBER, 2006).

O termo mineração de dados muitas vezes é referenciado como KDD (HAN, KAMBER, 2006, OLSON, DELEN, 2008). Figura 3.2 sintetiza o processo de descoberta de conhecimento. Observamos que os dados coletados nas bases de dados são transformados na fase de pré-processamento dos dados. A seguir, são aplicadas técnicas de mineração de dados, consideramos a utilização do termo EDM para especificar a utilização da mineração de dados no âmbito do tratamento de dados educacionais. O resultado da aplicação da mineração de dados também é transformado para gerar a visualização da informação. As descrições de cada componente são

mostradas a seguir:

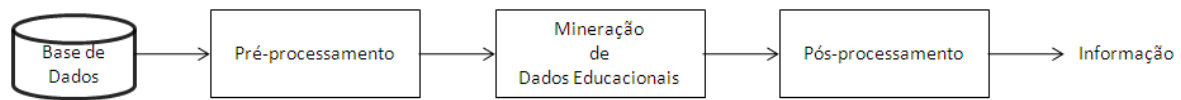


Figura 3.2: Síntese do Descoberta de Conhecimento em Dados.

3.1.1 Bases de Dados

As bases de dados utilizadas no processo de KDD podem ser geradas a partir de diferentes formas de estruturação e armazenamento de dados: banco de dados relacional, planilhas eletrônicas, *data warehouse*, arquivos de log (*log files*), *data stream*, dados da web, arquivos de dados simples (*flat files*), arquivos RDF, entre outros. Independente da fonte, os dados devem passar pelo processo de pré-processamento para serem transformados para o formato adequado para aplicar as técnicas de mineração de dados.

3.1.2 Pré-processamento dos Dados

A etapa de pré-processamento de dados engloba as fases de limpeza, seleção e preparação dos dados até obter o melhor conjunto de dados, HAN e KAMBER (2006) descrevem as seguintes fases:

- (1) *Limpeza dos dados* - remoção de dados incorretos ou inconsistentes para melhorar a qualidade dos dados;
- (2) *Seleção e Transformação dos dados* – quando dados relevantes para análise são obtidos das bases de dados. Os dados selecionados devem ser transformados e consolidados no formato apropriado para etapa de mineração.

A fase de pré-processamento dos dados são atividades iterativas que são realizadas até chegar ao conjunto de atributos relevantes para aplicar as técnicas de mineração de dados.

3.1.3 Mineração de Dados e EDM

A mineração de dados possui aplicação em diversos segmentos tais como biomedicina, engenharia, negócio, educação e outras áreas que envolvem descoberta de conhecimento em dados. Portanto, ela necessita de recursos de diversos campos de estudos para permitir sua maior utilização, tornando-se, então, intercessão de diversas áreas: banco de dados, aprendizado de máquina, estatística, reconhecimento de padrões, recuperação da informação, inteligência artificial e visualização da informação. Segundo a perspectiva de banco de dados apresentada por HAN e KAMBER (2006), mineração de dados é um processo de descoberta de conhecimento interessante escondido em bases de dados ou outro repositório de informação. Em WITTEN *et al.* (2011) mineração de dados é definida como o processo de descoberta de padrões nos dados. Segundo CARVALHO (2005) mineração de dados é o “uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano.”. Neste caso, não há grandes discordâncias entre os estudiosos sobre a definição do termo mineração de dados.

A definição de EDM apresentada por (BAKER, YACEF, 2009, ROMERO, VENTURA, 2010) diz que EDM é um campo de pesquisa que está relacionada ao desenvolvimento de métodos da mineração de dados para explorar tipos de dados educacionais e usar estes métodos para o melhor entendimento do estudante e o meio educacional em que ele está inserido. Segundo (BAKER, YACEF, 2009, BAKER *et al.*, 2011), as principais subáreas de pesquisa em EDM são:

- (1) Predição (*prediction*)
 - classificação (*classification*)
 - regressão (*regression*)
 - estimação de densidade (*density estimation*)
- (2) Agrupamento (*clustering*)
- (3) Mineração de relações (*relationship mining*)
 - mineração de regras de associação (*association rule mining*)
 - mineração de correlações (*correlation mining*)
 - mineração de padrões sequenciais (*sequential pattern mining*)
 - mineração de causas (*causal mining*)
- (4) Destilação de dados para facilitar decisões humanas (*distillation of data for*

human judgment)

- (5) Descobertas com modelos (*discovery with models*)

O estudo desenvolvido nesta tese teve como base a subárea da EDM chamada de predição, em particular a classificação. Na seção 3.2 mostraremos mais detalhes sobre as funcionalidades inerentes a predição de dados em EDM.

3.1.4 Pós-processamento dos Dados

Na etapa de pós-processamento, destacamos as seguintes operações (HAN, KAMBER, 2006):

- (1) *Avaliação de padrões* – identificação e interpretação de padrões interessantes, ou seja, aquisição de alguma informação relevante obtida a partir da análise dos dados na fase de mineração de dados;
- (2) *Apresentação dos resultados* - conjunto de atividades relacionadas a representação e visualização dos resultados da mineração de dados.

3.2 Funcionalidades da Mineração de Dados

Mineração de dados é a principal parte da análise de dados dentro do processo de descoberta de conhecimento. Para isso, diversas técnicas podem ser empregadas para realizar a mineração de dados, estas técnicas devem ser escolhidas de acordo com o tipo de dado disponível e com o conhecimento que se deseja obter a partir destes dados. Após a análise das características dos dados é que se pode determinar qual(is) as mais adequadas para explorar padrões de informação escondidos.

HAN e KAMBER (2006) descrevem algumas funcionalidades da mineração de dados. Por exemplo, que tipo de padrão deve ser encontrado quando se utiliza determinada técnica de mineração de dados. Em geral, as técnicas de mineração de dados podem ser classificadas em duas categorias: descritiva e preditiva. A categoria descritiva tem por objetivo analisar os dados, descrever suas características e apresentar propriedades interessantes gerais dos dados. A categoria preditiva tem por objetivo analisar os dados, a fim de construir um ou um conjunto de modelos, e tentar fazer inferências sobre os mesmos de modo que o sistema possa fazer predições ou prever o comportamento de novos conjuntos de dados. Os modelos descrevem aspectos específicos dos dados, portanto, necessita-se de uma quantidade de exemplos (registros)

que devem possuir um conjunto de características (atributos), que descrevam corretamente grupos ou classes distintas.

O conceito de classificação dentro da mineração de dados pode ser definido como um processo de encontrar um modelo ou função que descreve e distingue classes de dados com o propósito de utilizar o modelo encontrado para prever a classe de um novo elemento cuja identificação da classe é desconhecida (HAN, KAMBER, 2006, WITTEN *et al.*, 2011). O modelo gerado baseia-se na análise de um conjunto de treinamento com os rótulos das classes bem definidos e conhecidos. A classificação utiliza algoritmos supervisionados para inferir (prever) o grupo ou classe dos novos exemplos (registros). O algoritmo precisa de um conjunto de dados, na qual os exemplos (registros) possuem classes conhecidas, para aprender a identificar quais valores de atributos são importantes para definir ou caracterizar exemplos de cada classe. Há diversos algoritmos classificadores e diversas formas de representar o conhecimento. A escolha dos algoritmos para aprendizado do modelo depende das diversas características encontradas nos dados de entrada. A qualidade e a quantidade de dados influenciam diretamente no aprendizado do modelo. Neste caso, a qualidade representa o quanto o conjunto de entrada é significativo para descrever a classe a ser aprendida e a quantidade representa um número adequado de exemplos na base para treinamento e teste do modelo aprendido.

Os algoritmos classificadores podem fazer inferências com base em dados anteriores com o objetivo de fornecer previsões ou mostrar tendências. A inferência realizada pelo algoritmo classificador baseia-se nos valores dos atributos que compõem a base de dados. O algoritmo constrói um modelo baseado nas características que mais se aproximam da descrição de uma determinada classe. Quando um novo elemento não rotulado é testado pelo algoritmo, ele compara os valores dos atributos do novo elemento aos valores utilizados na construção do modelo que definiu cada classe. Desta forma, um novo rótulo da classe é definido para este novo elemento. As variáveis preditivas (*predicted variables*) são utilizadas pelos algoritmos para inferir a que classe pertence o novo exemplo (registro) (HAN, KAMBER, 2006). Outro termo também é muito utilizado para identificar a classe dos exemplos, chama-se atributo de classe.

Gregory Piatetsky-Shapiro (KDNUGETS, 2014) faz distinção entre o termo classificação e predição. “Por exemplo, o algoritmo *árvore de decisão* aplicado a dados existentes, com classes conhecidas, formam um modelo de classificação. Quando se aplica este modelo a novos dados cuja classe não é conhecida, obtém-se a predição da

classe. O pressuposto é que os novos dados vêm de uma distribuição semelhante utilizada para construir a árvore de decisão. Em muitos casos, isso é uma suposição correta e é por isso que se pode usar a árvore de decisão para a construção de um modelo preditivo”. Para Gregory Piatetsky-Shapiro a diferença entre classificação e predição é uma questão de definição. “A classificação é utilizada para dados existentes, por exemplo, grupo de pacientes com base em seus dados médicos conhecidos e resultado do tratamento, eu chamaria isso de uma classificação. Se utilizar um modelo de classificação para prever o resultado do tratamento para um novo paciente seria uma predição.”.

A classificação pode utilizar diferentes técnicas, indo desde técnicas mais simples de classificação até as mais complexas (NUGENT, CUNNINGHAM, 2004, HAN, KAMBER, 2006). Os algoritmos que geram modelos mais simples de serem interpretados são: árvores de decisão e regras de indução. Os que utilizam redes neurais e “*Support Vector Machines*” (SVM) não são transparentes na forma como os dados são classificados.

Os algoritmos podem responder diferentemente de acordo com a qualidade e o tipo dos dados de entrada, portanto, é necessário testar os algoritmos e utilizar algumas métricas para avaliar se o resultado da predição é satisfatório. HAN e KAMBER (2006) apresentam alguns critérios de comparação para algoritmos classificadores. A lista a seguir apresenta critérios para avaliar o desempenho dos algoritmos classificadores:

- (1) *Acurácia* – é a precisão de um classificador, dado um determinado conjunto de teste obtém a porcentagem dos exemplos (tuplas) que estão corretamente classificadas pelo classificador. Em outras palavras, mostra o quanto o modelo foi preciso para acertar os dados não rotulados do conjunto de teste;
- (2) *Taxa de erro ou acerto* – significa o quanto o modelo acertou ou errou na predição dos exemplos de cada classe analisada;
- (3) *Matriz de confusão* – é um recurso muito útil para análise do resultado do classificador, pois mostrar o quantitativo para as diferentes classes investigadas;
- (4) *Kappa* – utiliza-se a medida estatística *Kappa* para medir o número de respostas concordantes, ou seja, no número de casos cujo resultado é o mesmo entre o previsto e o observado em um conjunto de dados. O coeficiente *Kappa* é calculado levando-se em consideração todas as classes e é útil para mensurar o grau de concordância ou qualidade do classificador. O

valor estatístico próximo de 0 (zero) representa resultado de classificação ruim e quanto mais próximo de 1 (um) indica resultado excelente ou maior grau de concordância entre as observações (WITTEN *et al.*, 2011).

As bases de dados dos graduandos da UFRJ (disponíveis no sistema SIGA) são constantemente atualizadas e identificam a situação acadêmica do estudante a cada semestre letivo. Estas bases também possuem a identificação dos estudantes que chegaram à conclusão do curso de graduação, evadiram do curso ou estão em situação de pendência. Portanto, a bases de dados disponíveis possuem a identificação das possíveis classes de desempenho acadêmico dos estudantes que serão estudadas nesta tese. Portanto, neste trabalho, vamos adotar a definição de classificação como um processo de encontrar um modelo ou função que descreve e distingue classes de dados ou conceitos com o propósito de utilizar o modelo encontrado para predizer a classe de novos elementos cuja identificação de classe é desconhecida. Os modelos encontrados utilizando classificadores aplicados às bases de dados do SIGA ajudarão na predição da classe dos novos estudantes sendo, portanto, será possível predizer a situação acadêmica dos graduandos a cada semestre letivo identificando os estudantes mais propensos a abandonar o curso de graduação.

3.3 *Processo de Construção de um Modelo*

De uma maneira em geral, um modelo pode ser entendido como uma abstração, uma caixa preta, que faz previsões sobre o futuro baseado em informações do passado ou do presente (THEARLING, 2010). Como mostra a figura abaixo:

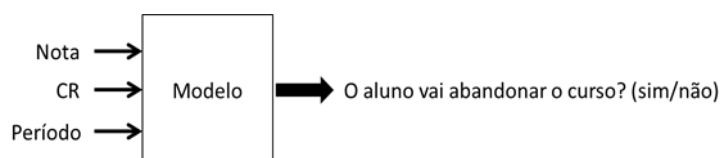


Figura 3.3: Esquema de modelo preditivo de dados.

A Figura 3.3 mostra um modelo onde diversas informações dos estudantes são as entradas de dados, a partir dos dados de entrada o modelo pode inferir sobre estes dados e produzir como saída uma predição. A precisão da resposta do modelo, entre outras coisas, depende da qualidade e da quantidade de dados disponíveis. Neste caso, isto se aplica tanto na criação do modelo quanto na sua validação ou utilização. Então, um

modelo é construído em função das informações disponíveis contidas na base de dados e da seleção da técnica de mineração de dados adequada.

A geração do modelo depende dos dados da base e dos algoritmos utilizados para aprendizagem do modelo. O modelo criado precisa ser validado, ou seja, deve-se verificar se ele atende a condições impostas para solucionar o problema. Vários modelos podem solucionar o problema, mas alguns podem apresentar melhor desempenho.

O processo de construção de um modelo passa por duas fases distintas e interativas: a primeira, chamada de descritiva, requer um conjunto de dados de treinamento que serão utilizados pelo algoritmo classificador para construir o modelo descritivo dos dados. A segunda fase, chamada de preditiva, testa o modelo gerado na primeira fase utilizando novos dados, conjuntos de teste. Uma pessoa analisa os resultados, e verifica se o modelo atende ao propósito. O modelo é validado para ser usado com novos dados e obter a predição. Se o modelo não atender, as seguintes ações devem ser realizadas:

- (1) Executar o algoritmo novamente modificando parâmetros ou utilizando novos dados;
- (2) Utilizar outro algoritmo a fim de encontrar o modelo preditivo desejado.

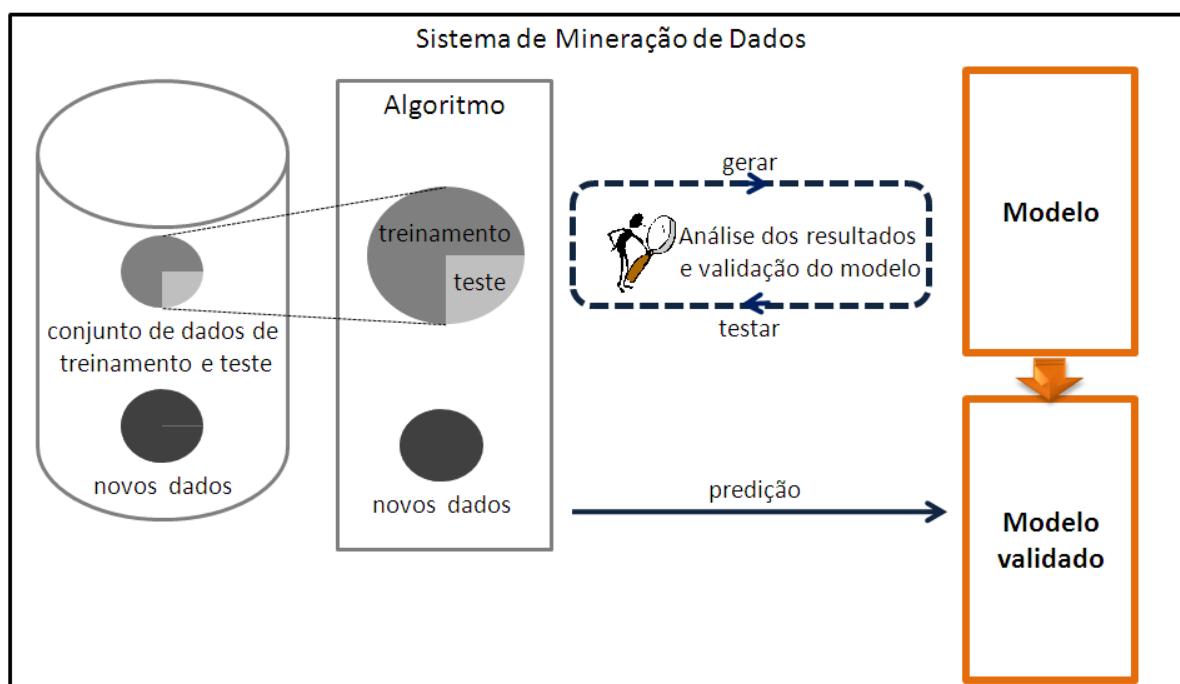


Figura 3.4: Esquema de construção, análise e validação do modelo preditivo de dados.

Este esquema para construção do modelo pode ser acompanhado na Figura 3.4. O processo de construção de um modelo preditivo para um determinado contexto precisa ser feito apenas uma vez podendo ser usado várias vezes para novos dados.

Portanto, o modelo obtido pode ser apresentado de diversas formas dependendo do algoritmo empregado. Os modelos podem ser descritos utilizando regras do tipo *se então* (*if then*), árvores de decisão, modelos estatísticos, redes neurais, entre outros.

3.4 Arquitetura

Nesta seção apresentaremos a arquitetura chamada EDM WAVE, ela foi concebida para inferir o desempenho acadêmico dos estudantes (MANHÃES *et al.*, 2011, 2012, 2014a, 2014b, 2014c, 2014d, 2015). Ela foi projetada com bases nos requisitos da EDM (BAKER, YACEF, 2009, ROMERO, VENTURA, 2010, BAKER *et al.*, 2011). O propósito da arquitetura é identificar e prever o desempenho acadêmico dos graduandos periodicamente. Utilizando apenas dados acadêmicos que variam com o tempo, permite que gestores acadêmicos, não especialistas em EDM, identifiquem estudantes em risco de evasão do sistema de ensino universitário.

Esta arquitetura foi concebida para agregar novos valores e funcionalidades ao Sistema de Gestão Acadêmica (SGA) das universidades públicas federais brasileiras. Atualmente, esses sistemas são softwares legados de difícil manutenção e que armazenam grandes quantidades de dados acadêmicos, mas faltam as funcionalidades para realizar um tratamento sistemático de informações dos estudantes a partir da perspectiva de avaliar o desempenho do estudante ou prever aqueles que estão em risco de abandono escolar.

A arquitetura proposta tem como objetivo ampliar os SGA, devido a sua natureza e sob a ótica do desenvolvimento de software a utilização de uma arquitetura adjacente e complementar ao sistema legado é menos arriscada e mais econômica do que a implementação de um novo sistema. Portanto, a arquitetura EDM WAVE apresentada na Figura 3.5, é uma abordagem mais adequada para adicionar novas funcionalidades analíticas, mantendo os sistemas existentes.

O modelo de arquitetura adotado para a implementação da EDM WAVE é o sistema multicamadas (*multi-tiers*). Particularmente, a arquitetura EDM WAVE foi concebida como uma arquitetura de três camadas: camada de dados, camada de aplicação e uma camada de apresentação.

- (1) A *camada de dados* consiste por bases de dados fornecidas pelo SGA;
- (2) A *camada de aplicação* gerencia as principais funcionalidades da arquitetura e as regras de processamento de dados. A camada de aplicação é constituída por três componentes: ETL, EDM e Repositório de Conhecimento (*Knowledge Management Repository - KMR*);
- (3) A *camada de apresentação* é o nível mais alto da arquitetura. Ela é responsável por tratar da interação com o usuário. O gestor educacional pode acessar o sistema diretamente usando a interface gráfica do usuário. (*Graphical User Interface - GUI*).

A Figura 3.5 ilustra detalhes da arquitetura EDM WAVE e as três camadas apresentadas acima.

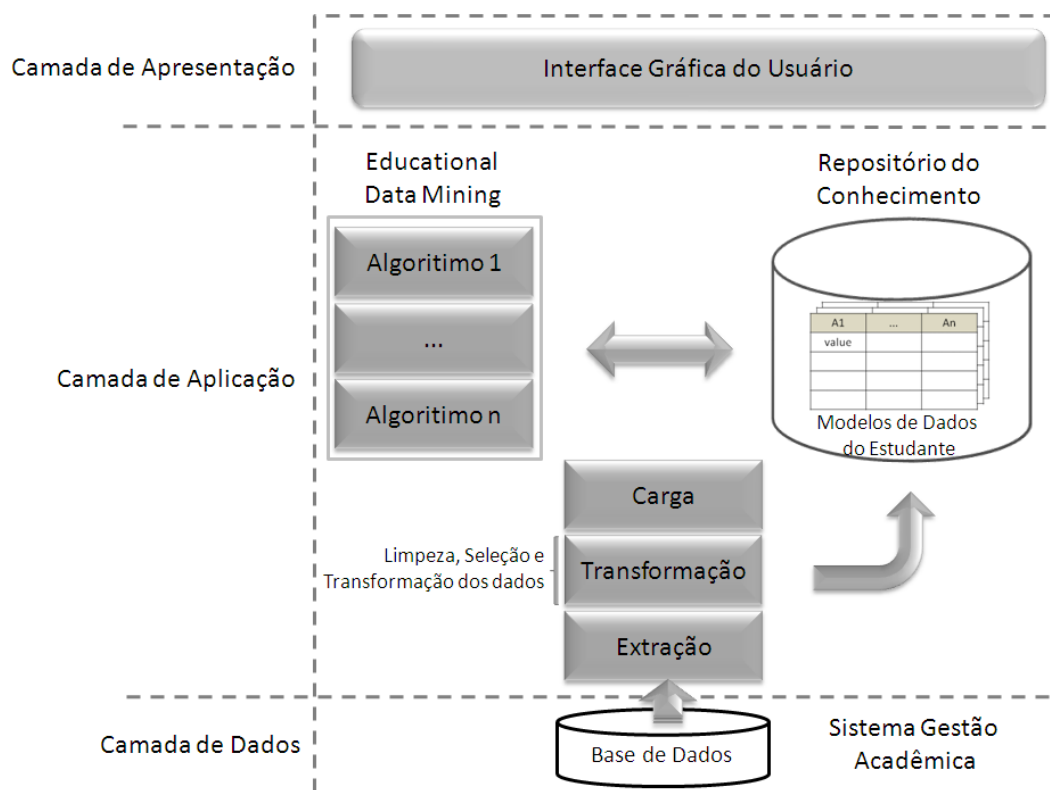


Figura 3.5: Arquitetura EDM WAVE baseada em três camadas.

3.4.1 Camada de Dados

A camada de dados consiste na coleção de bases de dados oriundas do Sistema de Gestão Acadêmica da universidade. O sistema de gestão acadêmica da UFRJ (SIGA) é um sistema legado que armazena grandes quantidades de dados acadêmicos

estruturados em um banco de dados relacional implementado no *Sistema Gerenciados de Banco de Dados MS-SQL-Server versão 2005*. As bases de dados deste sistema armazenam um expressivo volume de tabelas, as bases são confiáveis e armazenam informações acadêmicas de todos os estudantes da graduação de todos os cursos da UFRJ.

Está fora do escopo desta tese apresentar e discutir as principais características do *schema* relacional da base de dados do Sistema SIGA da UFRJ. Assim se garante a generalidade da proposta da arquitetura EDM WAVE.

As bases de dados do SIGA identificam os dados de cada estudante por um número de matrícula. Cada estudante possui diversas informações acadêmicas, entre elas o ano e semestre de ingresso no curso de graduação, a identificação do curso de graduação e os dados relacionados departamento que oferece o curso, disciplinas cursadas a cada período letivo, notas, coeficiente de rendimento no período, coeficiente de rendimento acumulado, situação no período, situação no curso entre outras informações acadêmicas. Os dados acadêmicos dos estudantes são periodicamente atualizados no sistema SIGA.

3.4.2 Camada de Aplicação

A camada de aplicação garante o isolamento dos dados, independência de aplicação, isola o atual sistema de problemas de desempenho e implementa as principais funcionalidades da arquitetura. Ela é constituída por três componentes: ETL, EDM e Repositório de Conhecimento (*Knowledge Management Repository - KMR*). As descrições de cada componente e as interações são mostradas a seguir:

3.4.2.1 Extração, Transformação e Carga

O processo de Extração, Transformação e Carga, mais conhecido pela sigla ETL (*Extract Transform and Load*) (KIMBALL, CASERTA, 2004), é mais comumente associado aos sistemas de *Data Warehouse* e *Data Mart*. No entanto, a utilização deste processo na arquitetura EDM WAVE facilitou consideravelmente a realização das primeiras fases do pré-processamento de dados previstos no KDD para serem realizadas antes da aplicação de EDM. A fase de pré-processamento de dados consome muito tempo, esta atividade demorada foi melhorada utilizando o processo de ETL.

Este processo utiliza ferramentas de software para realizar diversas atividades.

Primeiro, realiza-se a extração de dados de diversas fontes e sistemas. Os dados extraídos podem passar por seleção, limpeza e transformação conforme regras de negócios e por fim a carga dos dados para um determinado sistema da organização.

- (1) *Extração* - o primeiro componente ETL é responsável por carregar e preparar os registros para os próximos componentes. Esta primeira parte estabelece a ligação com a fonte do Sistema de Gestão Acadêmica (SIGA) ou com outros arquivos para extrair os dados dos estudantes. A extração converte os dados da base de dados para um determinado formato, de modo que possam ser utilizados pelos softwares que realizam a fase de transformação dos dados;
- (2) *Transformação* – a parte de transformação dos dados oriundos da base de dados do SIGA necessitou que se realizassem três etapas:
 - a. *Seleção* - extrair descritores das bases de dados dos sistemas acadêmicos. Apenas determinados atributos foram selecionados do conjunto armazenado na base de dados. A seção 4.2.1 descreve detalhes sobre os atributos originais extraídos da base de dados do SIGA;
 - b. *Limpeza* – limpar e transformar os descritores em dados adequados para a fase de mineração de dados. Os dados oriundos da base de dados possuíam valores inconsistentes e dados faltando. Nesta etapa foram realizados todos os acertos possíveis para que o maior número de registros pudesse ser aproveitado para constituir os arquivos a serem utilizados na fase de EDM;
 - c. *Transformação dos dados* - aplicar regras de transformação aos dados extraídos e criar novos atributos para armazenar os novos valores calculados. A seguir, agrupar e transformar os dados em arquivos apropriados para serem utilizados pelos algoritmos classificadores;
- (3) *Carga* - a fase de carregamento dos dados é usada para formatar e carregar os dados produzidos pelas etapas anteriores para serem usadas pelo componente de Repositório de Conhecimento. Nesta fase, os arquivos são gerados em formato para serem lidos pelos algoritmos classificadores utilizados na arquitetura. A reposição ou acréscimo de novos dados constituem opções de projeto e dependem de novas atualizações dos dados acadêmicos dos estudantes durante o decorrer dos semestres letivos.

O processo de ETL tem passos repetidos para executar extração de dados, seleção de dados, limpeza de dados e transformação de dados.

3.4.2.2 Repositório de Conhecimento (*Knowledge Management Repository* - KMR)

A camada de aplicação contempla também um Repositório de Conhecimento. Este repositório é constituído de uma coleção de dados sobre os estudantes e seus cursos. Cada conjunto de dados é definido como um modelo de dados dos estudantes. Portanto, um modelo compreende as características dos estudantes (atributos) relevantes para obter a predição do desempenho acadêmico.

Existem diversos modelos de dados dos estudantes, isto se faz necessário devido a algumas variações entre os cursos de graduação das IFES, exigindo que se tenha um modelo de dados dos estudantes para cada curso de graduação. Além disso, para fazer a predição a cada semestre letivo no decorrer do curso, é necessário dispor de conjunto de atributos diferentes. Cada modelo de dados dos estudantes apresenta características diferentes, que variam de acordo com o número de semestres concluídos pelo estudante. Estes dados são conhecimentos sobre o desempenho acadêmico dos estudantes obtidos em semestres anteriores, tendo como base os dados acadêmicos das disciplinas cursadas. Estes dados fornecidos para a camada de mineração de dados são conjuntos de treinamento importantes para os algoritmos classificadores determinarem com maior precisão o desempenho acadêmico dos novos estudantes. Detalhes sobre os modelos de dados utilizados na arquitetura serão apresentados na seção 4.3.

A arquitetura permite que o *repositório de conhecimento* possa ser atualizado com novas informações sobre o rendimento acadêmico dos novos estudantes, tornando-o mais robusto e consistente. O repositório foi criado para armazenar dados dos estudantes que serão utilizados posteriormente como conjuntos de treinamento pelos algoritmos classificadores do módulo EDM. No caso da UFRJ, estes dados devem ser extraídos do SIGA, utilizando a arquitetura EDM WAVE, ao final de cada semestre, após os lançamentos das notas dos exames finais. Além desses dados, o repositório armazena o resultado após a execução dos algoritmos no módulo EDM. Os resultados seguem o mesmo formato do modelo de dados dos estudantes com um campo adicional que identifica o valor da predição. Detalhes serão discutidos nos estudos de casos apresentados no capítulo 4.

A Figura 3.6 ilustra a parte da arquitetura que envolve o repositório e os arquivos

contendo os dados dos estudantes (modelos de dados dos estudantes), que serão utilizados pelo EDM.

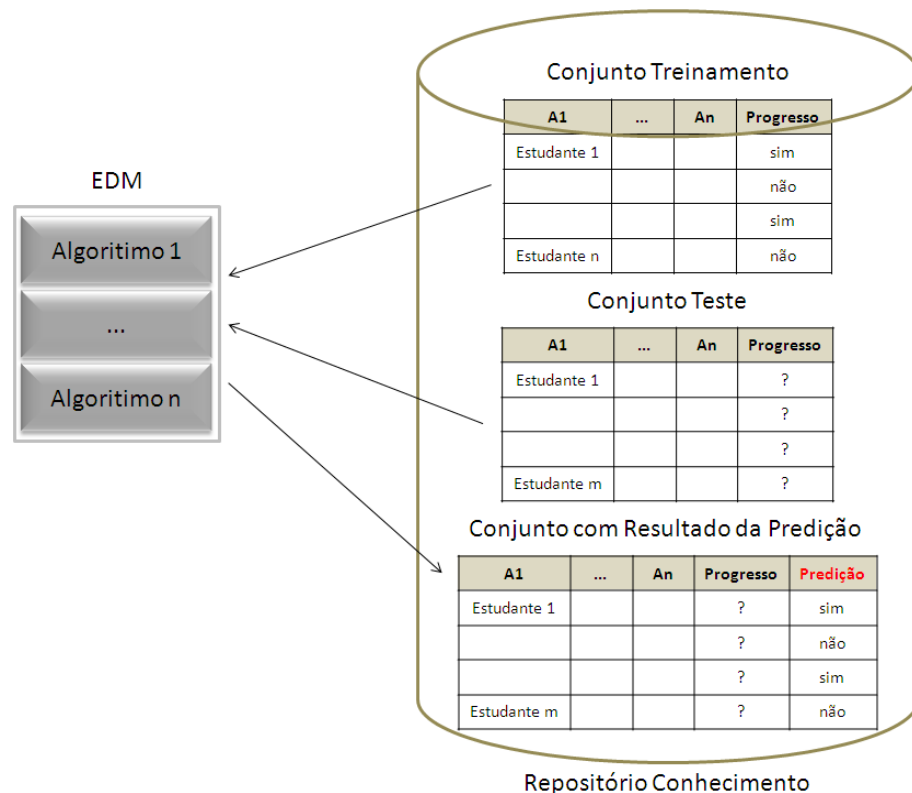


Figura 3.6: Repositório de Conhecimento da arquitetura EDM WAVE.

3.4.2.3 Mineração de Dados Educacionais (EDM)

A camada da aplicação possui o componente EDM. Este componente define o conjunto de algoritmos classificadores utilizados na arquitetura EDM WAVE. Este componente da arquitetura é ativado e recebe os conjuntos de dados, objetos de análise, estudantes cujo desempenho para o próximo semestre letivo deverá ser predito. O repositório fornece os arquivos de dados dos estudantes (treinamento e teste) conforme a Figura 3.6 para executar a predição. Cada algoritmo que executa a predição utilizando os conjuntos de treinamento e teste retorna um terceiro arquivo contendo o resultado da predição no formato de uma coluna de dados, ou seja, é acrescentado o atributo “predição” para identificar o valor inferido pelo algoritmo classificador. A Figura 3.7 mostra o esquema de execução do componente EDM da arquitetura.

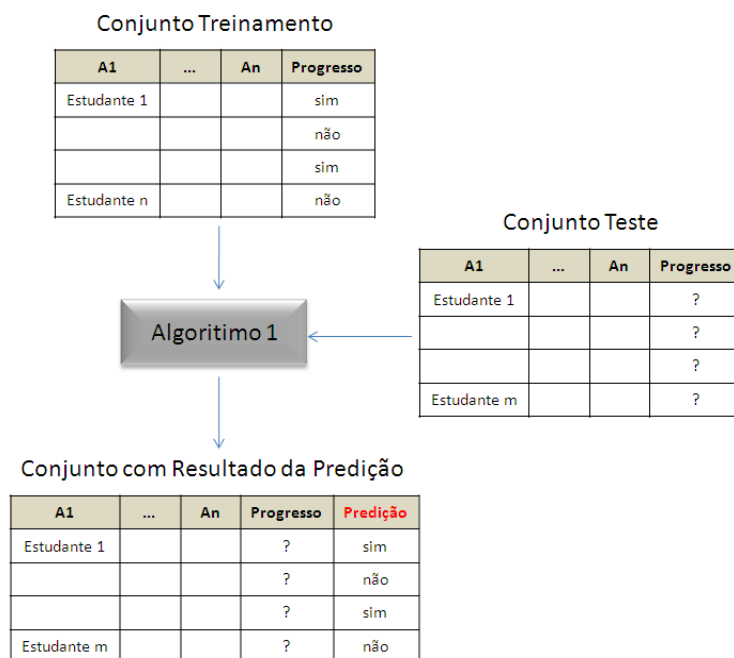


Figura 3.7: Esquema de execução de um algoritmo na arquitetura EDM WAVE.

3.4.3 Camada de Apresentação

Na arquitetura, a camada de apresentação tem papel importante na interação com usuário final, neste caso o gestor. Por exemplo, o gestor educacional pode interagir com a arquitetura de modo a executar as atividades previstas até obter o resultado individual de todos os algoritmos classificadores predizendo o desempenho do estudante no próximo semestre letivo.

A camada de apresentação exibe sob o formato de tabela a identificação do estudante e a predição do desempenho obtido por cada algoritmo classificador. Os detalhes do funcionamento dos algoritmos são transparentes para o gestor acadêmico, apenas o desempenho esperado de cada estudante é exibido. Desta forma, podem-se avaliar quantitativamente quais dos cinco algoritmos empregados chegaram ao mesmo resultado, o maior número de resultados coincidentes indicará a provável predição final.

3.4.3.1 Funcionalidades da camada de apresentação

A arquitetura oferece uma interface de comunicação com os usuários. O diagrama de caso de uso apresentado na Figura 3.8 descreve a sequência típica dos principais eventos-chave que representam as solicitações e comunicação dos usuários com a arquitetura. As principais comunicações exteriores são as seguintes:

- (1) O gestor acadêmico solicita a predição do desempenho acadêmico de um grupo de estudantes para o próximo semestre letivo;
- (2) A interface solicita que o usuário informe curso, turma e período, para que possa identificar no sistema acadêmico o grupo de estudantes para fazer a predição;
- (3) Opcionalmente, caso o sistema acadêmico não consiga ser acessado. Pode-se solicitar que o usuário preencha os dados dos estudantes, a partir destes dados informados pode-se iniciar a predição do desempenho para estes estudantes.

3.4.3.2 Modelo do relatório

A arquitetura EDM WAVE possui um conjunto de algoritmos classificadores. Cada algoritmo retorna uma predição para cada estudante do curso de um determinado período. Esta composição de algoritmos classificadores é utilizada para identificar o número de classificadores com a mesma predição, conferindo maior confiabilidade e reforçando os resultados globais. A Tabela 3.1 mostra um exemplo da disposição de um relatório que pode ser analisado pelo gestor educacional. O relatório mostra individualmente, em linhas, os resultados de cada estudante. As colunas mostram os resultados de predição de cada classificador. Valor "1" indica o progresso e "0" nenhum progresso. A última coluna (?) mostra o resultado da composição. Valor "1" é usado quando a maioria dos classificadores possui a mesma predição atribuindo progresso para o estudante. Por outro lado, o valor "0" é atribuído quando a maioria dos classificadores indica nenhum progresso. No entanto, o gestor educacional tem autonomia para interpretar os resultados. No exemplo de *layout* apresentado na Tabela 3.1 os seguintes algoritmos foram utilizados: *Naïve Bayes* (NB), *Multilayer Perceptron* (MLP), *Support Vector Machine com kernel polinomial* (SVM1) e *kernel RBF* (SVM2) e tabela de decisão (TD).

Tabela 3.1: *Layout* do relatório com a predição dos classificadores para n estudantes.

Estudante ID	NB	MLP	SVM1	SVM2	DT	?
Estudante 1	0	0	0	0	1	0
...
Estudante n	1	1	1	1	1	1

3.4.4 Funcionalidades da Arquitetura

As três camadas da arquitetura interagem entre si e com atores externos e com o sistema legado. O ator externo, por exemplo, o gestor acadêmico, é responsável por enviar a mensagem inicial que inicia a interação entre os componentes da arquitetura. No entanto, a sequência global do comportamento iniciados pode ser difícil de acompanhar. Os itens abaixo representam essa sequência de uma forma simples e lógica e por ordem temporal.

- (1) Receber a solicitação para iniciar a predição de um grupo de estudantes;
- (2) Solicitar que o usuário identifique o grupo de estudantes por (curso/turma/período);
- (3) Solicitar a camada de dados informações acadêmicas do grupo de estudantes;
- (4) Transformar (processar dados dos estudantes) de modo a gerar um arquivo (teste) no formato adequado para serem utilizados pelos algoritmos classificadores;
- (5) Solicitar ao repositório de conhecimento os arquivos de treinamento;
- (6) Encaminhar ao componente EDM os arquivos treinamento e o arquivo teste;
- (7) Receber os arquivos resultantes de cada classificador;
- (8) Mostrar resultado da predição.

A sequência permite mostrar utilizando um alto nível de abstração às dependências entre os componentes. Pelo diagrama ilustrado na Figura 3.8 existe a interface com o usuário, com o sistema acadêmico que são as bases de dados, uma interface com a implementação dos classificadores, por exemplo, os utilizados pela ferramenta Weka (HALL *et al.*, 2009, BOUCKAERT *et al.*, 2010). E com os demais componentes da arquitetura.

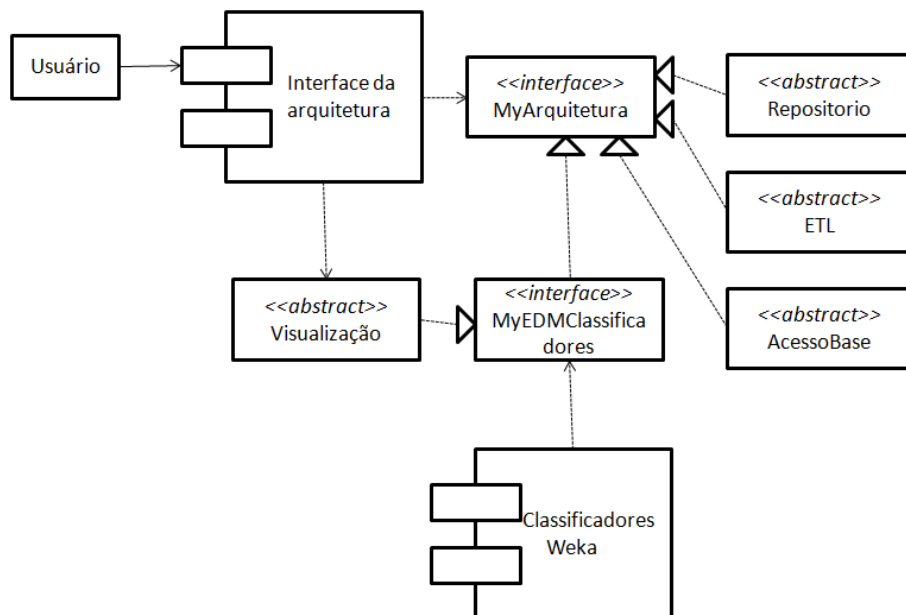


Figura 3.8: Diagrama de componentes (simplificado) da arquitetura EDM WAVE.

3.5 Conclusões

De uma maneira em geral, o processo de descoberta de conhecimento abrange diversas fases desde a preparação dos dados, passando pela mineração de dados até a interpretação dos resultados. A mineração de dados compreende uma etapa importante deste processo. A mineração de dados pode ser utilizada em um amplo leque de atividades para análise dos dados, pois apresenta diversos algoritmos aplicáveis a diferentes tipos de dados e também adequados a lidar com diversas informações disponíveis nestes dados. As formas de apresentação os resultados obtidos através das análises podem variar de acordo com o algoritmo utilizado, sendo o analista responsável por determinar qual algoritmo melhor se aplica ao contexto dos dados que estão sendo analisados.

O problema abordado nesta tese, embora ocorra em várias universidades, possui causas que estão intrínsecas ao contexto onde ele ocorre. O modelo de arquitetura proposto foi construído com bases em dados de estudantes que já passaram pela universidade, no entanto, todo o processo de escolha dos dados, adequação aos algoritmos classificadores, e funcionalidades aqui propostas tornam a arquitetura EDM WAVE bastante adaptável às futuras extensões.

A arquitetura é modular, em camadas e pode ser facilmente adaptável para acompanhar as mudanças que podem ocorrer ao longo do tempo nos requisitos do sistema, é muito provável que aconteçam mudanças de comportamento humano ou das

leis que regem a entrada de estudos nas IFES.

A arquitetura proposta neste trabalho, por ser genérica, poderá ser utilizada em outras IFES, espera-se com um número reduzido de adaptações.

4 Experimentos, Testes e Avaliação

Crítica

A avaliação da arquitetura EDM WAVE nesta tese foi realizada tendo como base dados reais de estudantes dos cursos de graduação da Universidade Federal do Rio de Janeiro (UFRJ), coletados por um período de 16 anos. A UFRJ é uma das maiores universidades públicas do Brasil com mais de 100 cursos de graduação que cobrem todas as áreas das Ciências e cerca de 50.000 estudantes de graduação.

Neste capítulo serão apresentadas as bases de dados e os diversos experimentos realizados para testar e validar a hipótese da tese. Este capítulo está organizado da seguinte forma. Na Seção 4.1 serão apresentadas as principais ideias que contextualizam o problema do insucesso dos graduandos na UFRJ. Na seção 4.2 descrevemos as bases de dados da UFRJ utilizadas neste estudo. Em seguida, na seção 4.3, definimos o Modelo de Dados dos Graduandos. Posteriormente, na seção 4.4, apresentamos os algoritmos classificadores utilizados nos experimentos. Na seção 4.5 definimos as formas de particionar as bases de dados. Na seção 4.6 definimos o processo de ETL utilizado. Na seção 4.7 definimos as ferramentas de mineração de dados. As seções 4.8 a 4.14 descrevemos os estudos de casos. Na seção 4.15 descrevemos a visualização dos resultados. Por fim, apresentamos a conclusão do capítulo.

4.1 Contextualização do Problema na Graduação da UFRJ

De um modo em geral, o problema da evasão universitária atinge diversos cursos de graduação. Embora ainda não haja estudos recentes sobre a evasão nos cursos em Ciências, Tecnologia, Engenharia e Matemática (no inglês o termo abreviado é STEM). Eles são alvo de muita preocupação entre os diretores responsáveis por estas áreas na UFRJ.

Em diversos segmentos da universidade, o problema da evasão vem sendo discutido. Particularmente, a Escola Politécnica da UFRJ, que oferece cursos de graduação para formação de engenheiros em diversas áreas, tem mostrado profundo interesse sobre este tema. Apesar da excelência na formação dos engenheiros e da concorrência por uma

vaga em seus cursos o problema da evasão preocupa a direção da Escola. Alguns estudos quantitativos mostraram o percentual de evasão dos estudantes que ingressaram nos períodos de 1990 a 2000, em alguns cursos as taxas de evasão variam entre 24,7% a 52,9% (SARAIVA, MASSON, 2003). As evasões ocorrem ao longo do curso e com maior frequência nos quatro primeiros semestres letivos. A concentração da evasão no início do curso também foi verificada em outras universidades como relatado por DEKKER *et al.* (2009) e JOHNSTON (1997).

4.2 Base de Dados da UFRJ

A UFRJ utiliza o Sistema de Gestão Acadêmico (SIGA) para manter informações acadêmicas dos estudantes e disciplinas. Conforme salientado, o sistema ainda não oferece recursos de gestão acadêmica, para os diretores, gestores e coordenadores de curso, portanto, ao final de cada semestre os gestores acadêmicos precisam planejar o cronograma para o próximo semestre, oferecendo disciplinas de acordo com o número de estudantes estimados. Esta atividade é uma tarefa complexa devido ao número irregular de estudantes que provavelmente estarão frequentando o curso no próximo semestre. Devido à grande quantidade de graduandos e baixa efetividade dos serviços de assessoria acadêmica, é difícil identificar quais estudantes estão em risco de evasão ou mesmo os já evadidos.

Os cursos de graduação da UFRJ são divididos em semestres letivos. A base de dados utilizada nesta tese é compartilhada com o SIGA, foram selecionados dados acadêmicos dos estudantes que ingressaram nos dois semestres letivos dentre os anos de 1994 até 2010.

Os dados utilizados nesta tese foram obtidos através da direção da Escola Politécnica que solicitou ao DRE (Divisão de Registro de Estudantes) da UFRJ e aos responsáveis pelo SIGA a colaboração com esta pesquisa. Em dezembro de 2010, os responsáveis forneceram as bases de dados contendo informações acadêmicas de todos os estudantes da UFRJ no período de 1994 a 2010. A base de dados recebida não possui identificação do estudante.

Os dados armazenados no SIGA não ofereciam qualquer tipo de padronização para serem utilizados diretamente pelos algoritmos classificadores no processo de EDM. Além disso, existiam inúmeros problemas de inconsistência nos dados da base, isto gerou muitas dificuldades no processo de análise dos dados.

- (1) Não há identificação em qual semestre o estudante concluiu o curso;
- (2) Nos casos de abandono definitivo, não há identificação de quando o estudante parou de frequentar o curso de graduação;
- (3) No caso de trancamento por alguns períodos, não existe identificação de quais períodos o estudante permaneceu com a matrícula trancada. Em períodos posteriores, os dados do estudante reaparecem indicando que ele voltou a frequentar o curso;
- (4) Quando não reaparece a informação, considera-se que ocorreu um abandono definitivo do curso de graduação;
- (5) Enorme quantidade de dados repetidos e dados inconsistentes. Por exemplo, estudantes com uma ou várias disciplinas com diferentes conceitos (notas) e/ou situação da disciplina (AP, RFM, RM e RF) no mesmo período;
- (6) Muitos registros com valores de CR e CRA acima do valor máximo;
- (7) Vários valores, em torno de 30, para o atributo situação de matrícula no período, e não há documentação de referência do significado dos termos utilizados;
- (8) Falta de documentação da equivalência de disciplinas. Por exemplo, muitos estudantes apresentam aprovação em disciplinas que não pertencem à grade do curso. No entanto, não há identificação de quais destas disciplinas são equivalentes às disciplinas da grade curricular do curso;
- (9) Estudantes com inscrição de ingresso no curso em um determinado semestre/período, mas que não fizeram as disciplinas iniciais neste período de ingresso.

A atividade de pré-processamento foi desenvolvida utilizando o processo ETL (descrito anteriormente). Mesmo utilizando este recurso, a fase de pré-processamento consumiu grande parte do tempo do desenvolvimento da tese. Toda informação a respeito dos dados era obtida através de conversas e entrevista com os administradores do sistema acadêmico. O sistema SIGA não possuía documentação, portanto, o trabalho de entendimento dos dados foi através da manipulação direta desses dados. Nesta fase, verificamos que era necessário transformar a maioria dos dados para obter um modelo de dados que atendesse aos objetivos da tese. Neste caso, o novo modelo de dados não deveria ser influenciado pelos problemas encontrados nos dados e no sistema da UFRJ, listados acima. Além disso, o maior número possível de registros da base de dados original do SIGA deveria ser utilizado. Novas bases de dados tiveram de ser criadas

para armazenar os novos dados gerados durante o processo de pré-processamento de dados.

4.2.1 Descrição dos Atributos Originais Extraídos da Base de Dados do SIGA

A Tabela 4.1 ilustra os principais conjuntos de atributos originais da base de dados do sistema acadêmico SIGA extraídos para este estudo.

Tabela 4.1: Lista de atributos originais da base de dados do SIGA.

Nº	Atributos do SIGA	Descrição
1	codCursoAtual	Código do curso onde o estudante está atualmente matriculado
2	nomeCursoAtual	Nome do curso onde o estudante está atualmente matriculado
3	codUnidade	Código da Escola, Instituto ou Faculdade onde o curso é oferecido na universidade
4	nomeUnidade	Nome da Escola, Instituto ou Faculdade onde o curso é oferecido na universidade
5	codCentro	Código do centro da UFRJ
6	nomeCentro	Nome do centro da UFRJ
7	formaIngresso	Forma de ingresso do estudante na universidade (vestibular, transferências e outros)
8	segmentacaoIngresso	Ano e semestre letivo que o estudante ingressou na universidade
9	situacaoMatriculaAtual	Situação atual da matrícula do estudante
10	CRA	O CRA (Coeficiente de Rendimento Acumulado) é a média ponderada pelo número de créditos das notas das disciplinas já cursadas durante todo o curso. O SIGA calcula o CR e o CRA do estudante
11	periodoDisciplina	Ano/semestre letivo da disciplina cursada pelo estudante
12	CR	O CR (Coeficiente de Rendimento) é a média ponderada pelo número de créditos das notas obtidas nas disciplinas cursadas em um determinado período do curso
13	situacaoMatriculaNoPeriodo	Situação da matrícula do estudante no período letivo
14	codDisciplina	Código da disciplina cursada
15	nomeDisciplina	Nome da disciplina cursada
16	creditos	Número de créditos da disciplina
17	conceito	Valor numérico (nota) atribuído à disciplina cursada
18	nomeCurto	Situação atribuída à disciplina cursada: AP (Aprovado), RFM (Reprovado por Falta e Média), RM (Reprovado por Média) e RF (Reprovado por Falta)

O atributo original do SIGA “*situacaoMatriculaAtual*” (atributo Nº 9 da Tabela 4.1) descreve a situação do estudante no curso de graduação até o momento da extração dos dados em novembro de 2010. Existem vários valores para este atributo que identifica a situação de matrícula do estudante no curso de graduação (*Aband Def*, *Abandono*, *Aluno*

em Int, Ativa, Canc a Pedido, Canc Conc Int, Canc Dec Judic, Canc Faltou Matricula, Canc Opcao Curso, Canc out mot, Canc Sancao Dis, Conclusao, Especial, Expulsao, Jubilamento, Mat. Desat., Mobilidade, Morte, Rem Aut, Rem Automatica, Rem Ex-offic, Rem Isenc Vest, Rem p/ Tranf, Rematricula, Trancada para o doutorado sanduiche, Trancamento, Trancamento Automatico, Trancamento Solicitado, Transferencia e Ult Prazo Tranc).

O atributo original do SIGA “*situacaoMatriculaNoPeriodo*” (atributo N° 13 da Tabela 4.1) descreve a situação da matrícula do estudante em cada período cursado, possui vários valores diferentes: (*AbandDef, Abandono, Aluno em Int, Ativa, Canc a Pedido, Canc Conc Int, Canc Dec Judic, Canc Faltou Matricula, Canc Opcao Curso, Canc out mot, Conclusao, Especial, Expulsao, Jubilamento, Mobilidade, Morte, Rem Aut, Rem Automatica, Rem Ex-offic, Rem Isenc Vest, Rem p/ Tranf, Rematricula, Trancada para o doutorado sanduiche, Trancamento, Trancamento Automatico, Trancamento Solicitado, Transferencia, Ult Prazo Tranc*).

A Tabela 4.2 ilustra um exemplo dos dados acadêmicos originais obtidos do SIGA para um graduando da UFRJ, a figura mostra todas as disciplinas cursadas ao longo dos períodos letivos, cada linha se refere aos dados de uma disciplina cursada.

Tabela 4.2: Exemplo de dados acadêmicos de um estudante de graduação obtidos a partir do SIGA.

codCurso	nomeCurs	codUnid	nomeUnic	codCentr	nomeCen	formaIngr	segmenta	situacaoM	cracumula	periodoDi	craPeriod	situacaoM	codDiscip	nomeDisc	creditos	conceito	nomeCurt
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-1		70 Ativa	EEH210	Engenhari	2	70 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-1		70 Ativa	FIS111	Fisica Expi	1	59 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-1		70 Ativa	IQG111	Química E	4	64 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-1		70 Ativa	FIT112	Física I - A	4	72 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-1		70 Ativa	MAC118	Calculo Di	6	71 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-1		70 Ativa	MAB124	Programa	3	76 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-1		70 Ativa	EEL200	Introduca	2	70 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-2		70 Ativa	MAE125	Algebra Li	4	72 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-2		70 Ativa	FIT122	Física II - A	4	74 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-2		70 Ativa	IQG112	Química E	2	65 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-2		70 Ativa	MAC128	Calculo Di	4	74 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-2		70 Ativa	EEG206	Expressao	5	55 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-2		70 Ativa	FIS121	Física Expi	1	64 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1994-2		70 Ativa	MAB224	Programa	3	91 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-1		69 Ativa	MAC238	Calculo Di	4	73 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-1		69 Ativa	FIN231	Física Expi	1	62 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-1		69 Ativa	EEG207	Expressao	5	64 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-1		69 Ativa	MAB231	Calculo Ni	4	80 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-1		69 Ativa	EEA212	Mecanica	4	28 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-1		69 Ativa	FIM230	Física III -	4	89 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-2		69 Ativa	EEL201	Instrumer	4	74 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-2		69 Ativa	EEL206	Historia d	2	80 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-2		69 Ativa	FIN241	Física Expi	1	61 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-2		69 Ativa	FIM240	Física IV -	4	68 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-2		69 Ativa	EEL201	Probabilic	4	60 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-2		69 Ativa	MAC243	Calculo IV	4	63 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-2		69 Ativa	FCF245	Filosofia c	2	100 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1995-2		69 Ativa	EEA212	Mecanica	4	72 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1996-2		66 Ativa	EEL351	Elettronica	6	67 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1996-2		66 Ativa	EEL354	Teoria Ele	5	50 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1996-2		66 Ativa	EEL355	Circuitos l	5	55 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1996-2		66 Ativa	EEL353	Sistemas l	4	76 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1996-2		66 Ativa	EEL352	Circuitos l	5	31 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1997-1		66 Ativa	EEL364	Teoria Ele	4	55 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1997-1		66 Ativa	EEL352	Circuitos l	5	69 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1997-1		66 Ativa	EEL363	Controle l	5	60 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1997-1		66 Ativa	EEL321	Organizac	4	77 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1997-2		64 Ativa	EEL365	Sistemas l	5	58 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1997-2		64 Ativa	EEL361	Elettronica	6	12 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1997-2		64 Ativa	EEL362	Circuitos l	5	55 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1997-2		64 Ativa	EEL474	Comunica	4	67 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1998-1		63 Ativa	EEL361	Elettronica	6	70 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1998-1		63 Ativa	EEL472	Sintese M	4	38 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1998-1		63 Ativa	EEL475	Organizac	4	67 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1998-2		62 Ativa	EEL471	Elettronica	6	50 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1998-2		62 Ativa	EEL312	Economia	4	61 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1998-2		62 Ativa	EEL484	Comunica	4	22 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1999-1		60 Ativa	EEL485	Software l	4	20 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1999-1		60 Ativa	EEL483	Controle l	4	71 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1999-1		60 Ativa	EEL523	Microcom	4	62 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1999-1		60 Ativa	EEL613	Redes Nei	4	0 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1999-2		58 Ativa	EEL485	Software l	4	58 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1999-2		58 Ativa	EEL740	Comunica	5	33 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1999-2		58 Ativa	EEL615	Elettronica	5	15 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 1999-2		58 Ativa	EEL473	Controle l	5	44 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2000-1		58 Ativa	EEL873	Engenhari	4	50 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2000-1		58 Ativa	EEL770	Sistemas l	5	62 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2001-1		57 Ativa	EEL191	Engenhari	4	70 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2001-1		57 Ativa	EEL878	Redes de	4	51 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2001-1		57 Ativa	EEL871	Banco de l	4	0 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2001-1		57 Ativa	EEL879	Redes de	4	72 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2002-1		57 Rematrã	EEE387	Conversac	5	55 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2002-1		57 Rematrã	EEL856	Sist.de Co	4	49 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2002-1		57 Rematrã	EEL710	Instrum	5	83 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2002-1		57 Rematrã	EEL760	Controle l	5	5 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2002-2		57 Ativa	EEE387	Conversac	5	0 RFM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2002-2		57 Ativa	EEL615	Elettronica	5	72 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2002-2		57 Ativa	EEL856	Sist.de Co	4	0 RFM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2002-2		57 Ativa	EEL472	Sintese M	4	50 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2002-2		57 Ativa	EEL760	Controle l	5	0 RFM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2003-1		57 Rematrã	EEL740	Comunica	5	43 RM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2003-2		57 Rematrã	EEL740	Comunica	5	80 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2003-2		57 Rematrã	EEL473	Controle l	5	75 AP	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2003-2		57 Rematrã	EEL760	Controle l	5	0 RFM	
3,6E+09	Engenhari	36010000	Escola Pol	36000000	Centro de Vestibul	1994-1		Conclusão		57 2005-1		57 Rematrã	EEL670	Linguager	5	58 AP	

4.3 Definição do Modelo de Dados dos Graduandos

Os dados acadêmicos obtidos do SIGA não puderam ser utilizados diretamente pela ferramenta de mineração de dados. Devido ao formato não adequado e a falta de padronização e documentação. Portanto, foram criados novos modelos de dados dos estudantes para serem armazenados no repositório de conhecimento da arquitetura e posteriormente aplicados aos algoritmos classificadores no processo de EDM. Neste trabalho, os modelos de dados dos estudantes foram construídos a partir do resultado de várias investigações e experimentação, o conjunto de atributos obtidos a partir destas análises constitui o modelo de dados dos graduandos.

A Tabela 4.3 ilustra a descrição dos atributos criados e utilizados nos modelos de dados dos graduandos, estes atributos foram criados a partir dos atributos originais do SIGA (Tabela 4.1).

O atributo “*Id estudante*” (atributo N° 1 da Tabela 4.3) foi utilizado para identificar todos os dados relacionados a um estudante específico.

Foram criados prefixos para facilitar a identificação de alguns atributos ao longo dos semestres letivos. Os atributos com prefixos (*01S*, *02S*, ..., *13S*) indicam o semestre cursado. Por exemplo, (*01S*) indica que o atributo mantém dados do primeiro semestre cursado, (*02S*) indica que o atributo mantém dados do segundo semestre e assim sucessivamente. Para fins da pesquisa, considerou-se analisar até o 13° semestre a partir do semestre de ingresso no curso.

Os atributos “(*01S*, *02S*, ..., *13S*)_SitPeriodo” (atributo N° 7 da Tabela 4.3) foram criados a partir dos dados do atributo original do SIGA “*situacaoMatriculaNoPeriodo*” (atributo N° 13 da Tabela 4.1). Este atributo é utilizado para identificar as situações acadêmicas dos estudantes em cada período letivo. Dos diversos valores encontrados no SIGA, originaram-se três valores distintos para identificar a situação do estudante no período cursado: (*APROVADO*, *PAROU* e *ATIVO*). O valor (*APROVADO*) foi atribuído a todos os estudantes que obtiveram pelo menos uma disciplina aprovada no período. O valor (*ATIVO*) foi atribuído aos estudantes que não obtiveram aprovação em alguma disciplina no período letivo, mas estavam regularmente matriculados. O valor (*PAROU*) foi atribuído a todos os estudantes que não possuíam qualquer informação acadêmica no período.

Os atributos “(*01S*, *02S*, ..., *13S*)_CRPeriodo” (atributo N° 8 da Tabela 4.3) foram utilizados para manter o coeficiente de rendimento (CR) obtido em cada período

cursado.

Os atributos “(01S, 02S, ..., 13S)_NoDisc” (atributo N° 9 da Tabela 4.3) foram criados para armazenar o número de disciplinas em que o estudante se matriculou no período letivo. Por exemplo: O atributo “01S_NoDisc” registra que o estudante cursou 7 disciplinas no primeiro semestre letivo e “02S_NoDisc” registra que o estudante cursou 3 disciplinas no segundo semestre letivo e assim sucessivamente.

Os atributos “(01S, 02S, ..., 13S)_NoAP” (atributo N° 10 da Tabela 4.3) foram criados para manter o número de disciplinas nas quais o estudante obteve aprovação em cada semestre letivo cursado.

Os atributos “(01S, 02S, ..., 13S)_MediaAP” (atributo N° 11 da Tabela 4.3) foram criados para manter a média aritmética obtida nas disciplinas aprovadas em cada período letivo. Este cálculo é diferente do CR do período porque só considera as disciplinas aprovadas.

Os atributos “(01S, 02S, ..., 13S)_NoRFM” (atributo N° 12 da Tabela 4.3) foram criados para manter o número de disciplinas reprovadas por falta e/ou média em cada período letivo.

Os atributos “(01S, 02S, ..., 13S)_NoRM” (atributo N° 13 da Tabela 4.3) foram criados para manter o número de disciplinas reprovadas por média em cada período letivo.

Os cursos de graduação da UFRJ oferecem uma grade curricular com disciplinas dispostas em períodos semestrais. As disciplinas oferecidas no primeiro período são normalmente feitas por todos os novos estudantes. As demais disciplinas da grade curricular podem ter algum pré-requisito ou exigências, mas elas podem ser cursadas em diferentes períodos por cada estudante. No sistema de créditos não existe o conceito de turma de estudantes que cumprem a mesma grade do início ao fim do curso de graduação. No modelo de dados dos estudantes proposto nesta tese levou-se em consideração dados específicos das disciplinas do primeiro semestre. Foram identificadas 6 a 7 disciplinas introdutórias nas grades curriculares dos cursos de graduação, portanto, foram utilizados os prefixos (1D, 2D, ..., 7D) para identificar cada disciplina da grade do primeiro semestre letivo. Na Tabela 4.3 identificam-se três conjuntos de atributos que armazenam valores das disciplinas do primeiro semestre. Os atributos “(1D, 2D, ..., 7D)_Disciplina” foram criados para armazenar a identificação das disciplinas do primeiro período da grade curricular do curso de graduação. Os atributos “(1D, 2D, ..., 7D)_Conceito” foram criados para armazenar as notas (valor

numérico) obtidas nas disciplinas da grade curricular do primeiro semestre letivo. Os atributos “(1D, 2D, ..., 7D)_SitDisciplina” foram criados para armazenar a situação da disciplina de primeiro período (AP, RM, RFM).

O atributo de classe é utilizado para auxiliar o algoritmo classificador a prever a classe dos novos registros. Pode-se definir um dos vários atributos como atributo de classe, a escolha deve ser com base na regra que define o desempenho acadêmico dos estudantes em um determinado período do curso.

Tabela 4.3: Modelo de dados dos estudantes de graduação.

Nº	Atributos	Descrição	Valor	Tipo
1	IdEstudante	Identificador do estudante	Código Id	String
2	IdIngresso	Identificador do ano e período em que o estudante ingressou na universidade	Código Id	String
3	IdCurso	Identificador do curso de graduação no qual o estudante está matriculado	Código Id	String
4	IdUnidadeCurso	Identificador da Escola, Instituto ou Faculdade onde o curso é oferecido na universidade	Código Id	String
5	CRA	Coeficiente Rendimento Acadêmico Acumulado (CRA) é a média de aproveitamento das disciplinas cursadas durante todo o curso de graduação	{0 to n}	Numérico
6	(01S, 02S, ..., 13S)_Periodo	Período letivo identificado por (ano-semester)	Código Id	String
7	(01S, 02S, ..., 13S)_SitPeriodo	Mantém a situação da matrícula do estudante no período	{APROVADO, PAROU, ATIVO}	String
8	(01S, 02S, ..., 13S)_CRPeriodo	Mantém o coeficiente de rendimento do período cursado	{0 to n}	Numérico
9	(01S, 02S, ..., 13S)_NoDisc	Mantém o número de disciplinas cursadas em cada período letivo	{0 to n}	Numérico
10	(01S, 02S, ..., 13S)_NoAP	Mantém o número de disciplinas aprovadas em cada período letivo	{0 to n}	Numérico
11	(01S, 02S, ..., 13S)_MediaAP	Mantém a média aritmética obtida nas disciplinas aprovadas em cada período letivo	{0 to 100}	Numérico
12	(01S, 02S, ..., 13S)_NoRFM	Mantém o número de disciplinas reprovadas por falta e/ou média em cada período letivo	{0 to n}	Numérico
13	(01S, 02S, ..., 13S)_NoRM	Mantém o número de disciplinas reprovadas por média em cada período letivo	{0 to n}	Numérico

Nº	Atributos	Descrição	Valor	Tipo
14	(1D, 2D, ...,7D)_Disciplina	Identifica as disciplinas do primeiro semestre da grade curricular do curso de graduação	Código disciplina	<i>String</i>
15	(1D, 2D, ...,7D)_Conceito	Armazena as notas (valor numérico) obtidas nas disciplinas da grade curricular do primeiro semestre letivo	{0 to 100}	Numérico
16	(1D, 2D, ...,7D)_SitDisciplina	Mantém a situação na disciplina do primeiro semestre letivo: AP (Aprovado), RFM (Reprovado por Falta e Média), RM (Reprovado por Média) e RF (Reprovado por Falta)	{AP, RM, RFM}	<i>String</i>
17	{atributo de classe}	Atributo de classe é utilizado pelo algoritmo classificador para inferir o valor da classe dos exemplos.		

4.4 Definição dos Algoritmos Utilizados nos Experimentos

O sucesso da mineração de dados resulta no emprego de diversos algoritmos originalmente criados para aprendizado de máquina. Estes algoritmos foram desenvolvidos e baseados em métodos simples ou mais complexos, os trabalhos de (CARVALHO, 2005, HAN, KAMBER, 2006, KOTSIANTIS *et al.*, 2007, WU *et al.*, 2008, WITTEN *et al.*, 2011) detalham os métodos empregados.

Os algoritmos classificadores quando aplicados a uma base de dados descrevem modelos de classes ou conceitos (HAN, KAMBER, 2006). Os modelos obtidos são utilizados para identificar novos exemplos cuja classe é desconhecida. Para criar um modelo, é necessário treinar os algoritmos classificadores utilizando exemplos corretamente rotulados em classes conhecidas a priori. Este processo denomina-se aprendizagem supervisionada. Neste caso, o algoritmo classificador constrói o modelo (aprendizagem) a partir do conjunto de treinamento composto por amostras (exemplos) com a identificação da classe as quais elas pertencem. A aprendizagem é verificada quando o conjunto de teste é aplicado ao modelo aprendido. Existem várias formas de representar ou descrever um modelo, resultado dos algoritmos classificadores, as mais comuns são através de regras, árvores, tabelas de decisão, redes neurais, métodos estatísticos, entre outros.

A importância da identificação dos algoritmos mais adequados para prever o desempenho acadêmico dos estudantes foi discutida na seção 2.5.2. No entanto, nesta tese foi feita uma ampla análise utilizando diversos algoritmos classificadores utilizados

em aplicações de mineração de dados tradicionais. Nos estudos de casos apresentados a seguir, são comparados os desempenhos dos algoritmos classificadores utilizando a nova base de dados gerada, modelo de dados dos graduandos. A comparação do desempenho dos algoritmos aplicados ao domínio do problema serve para justificar a escolha do algoritmo mais adequado para EDM. A escolha dos algoritmos utilizados nos experimentos deve-se a larga utilização dos mesmos em diversos contextos (WITTEN, FRANK, 2005, WU *et al.*, 2008, WITTEN *et al.*, 2011). A Tabela 4.4 mostra o nome que identifica o algoritmo classificador e uma breve descrição dos métodos empregados pelos algoritmos. Os detalhes sobre os métodos utilizados por cada algoritmo podem ser obtidos nas referências (WITTEN, FRANK, 2005, WU *et al.*, 2008, WITTEN *et al.*, 2011).

Tabela 4.4: Identificação e breve descrição dos classificadores.

Identificação	Descrição
AdaBoost	<i>"Adaptive Boosting"</i> . O <i>Boosting algorithm</i> é um <i>machine learning ensemble meta-algorithm</i> .
BayesNet	<i>"Bayesian network"</i> . Classificador baseado em um tipo de modelo estatístico (<i>probabilistic directed acyclic graphical model</i>)
DecisionTable	Tabela de decisão simples (Decision table model)
J48	Árvore de decisão (decision tree), implementação do C4.5
JRip	Aprendizado baseado em regras (rule-based learner), implementação do RIPPER
MultilayerPerceptron	Rede neural artificial baseado no (Perceptron-based)
Naive Bayes	Classificador probabilístico simples baseado na aplicação do teorema de Bayes
OneR	Árvore de decisão baseado no modelo (One-level decision tree)
RandomForest	Randomized decision tree
SimpleLogistic	Modelos lineares de regressão logística (<i>Logistic regression model</i>)
SVM with Poly Kernel (SVM1) e SVM with RBF Kernel (SVM2)	Máquina de Vetor de Suporte (<i>Support Vector Machine - SVM</i>)

O problema apresentado nesta tese está situado dentro do contexto de aplicação da EDM, portanto, a investigação de diversos algoritmos de aprendizado de máquina serve para justificar quais são os mais adequados para serem utilizados em aplicações EDM e também que investigam o desempenho acadêmico dos graduandos. Na seção de experimentos, comparam-se o desempenho dos diversos algoritmos classificadores, listados abaixo, utilizando a nova base de dados gerada neste estudo: modelo de dados dos graduandos. Lista dos nomes dos algoritmos classificadores implementados na

ferramenta Weka:

- (1) *AdaBoost* (AD)
- (2) *BayesNet* (BN)
- (3) *DecisionTable* (DT)
- (4) *J48* (J48)
- (5) *JRip* (JR)
- (6) *MultilayerPerceptron* (MP)
- (7) *NaiveBayes* (NB)
- (8) *OneR* (OR)
- (9) *RandomForest* (RF)
- (10) *SimpleLogistic* (SL)
- (11) *SVM com PolyKernel* (SVM1)
- (12) *SVM com RBFKernel* (SVM2)

4.5 *Particionamento da Base de Dados*

Existem vários métodos de divisão da base de dados para obter os subconjuntos de treinamento e teste para serem utilizados pelos algoritmos classificadores, um dos métodos mais empregados é a validação cruzada (*k-fold cross-validation*). Este método divide a base de dados em k conjuntos (HAN, KAMBER, 2006). A forma mais comum é utilizar a divisão em 10 conjuntos. Outra forma de particionamento a base de dados é utilizar a divisão em duas partes, conjunto de treinamento e conjunto de teste.

Nos experimentos foram utilizadas a validação cruzada com 10 conjuntos e o particionamento utilizando conjuntos de treinamento e teste.

4.6 *Ferramentas para o Processo ETL*

O processo de ETL foi realizado em nosso estudo utilizando os softwares da Microsoft Access e Excel. O Excel é uma ferramenta simples e possui recursos para extração, transformação e carregamento (ETL). Os dois programas foram adequados para carregar o volume de dados da base de dados disponibilizada pelo SIGA. As funcionalidades dos programas permitiram a manipulação dos atributos. O Excel disponibiliza diversas funções matemáticas e estatísticas, possui diversos recursos de visualização dos dados e permite gerar arquivos em formato adequado para serem lidos

pelos algoritmos de mineração de dados.

4.7 Ferramentas de Mineração de Dados

As ferramentas de mineração de dados dispõem de recursos de análise de dados e implementam diversos algoritmos utilizados na mineração de dados. A ferramenta de mineração de dados Weka (HALL *et al.*, 2009, BOUCKAERT *et al.*, 2010) disponibiliza vários algoritmos classificadores. Ela foi utilizada neste trabalho devido a: facilidade de aquisição, o software está disponível para *download* na página do desenvolvedor sem custo de utilização, a ferramenta dispõe de várias versões de algoritmos empregados na mineração de dados e disponibilidade recursos estatísticos para comparar o desempenho dos algoritmos. Seus algoritmos são implementados na linguagem de programação Java e podem ser utilizados (instanciados) por outros sistemas.

A ferramenta Weka disponibiliza dois ambientes para realizar os experimentos: *Weka Explorer* (WE) e *Weka Experiment Environment* (WEE) (SCUSE, REUTEMANN, 2008). Cada ambiente oferece diversas formas de selecionar as bases de dados, particionamento das mesmas e aplicação de diversos algoritmos.

4.7.1 Weka Explorer (WE)

O *Weka Explorer* (WE) é um ambiente da ferramenta Weka que permite a seleção e execução de um algoritmo classificador por vez (BOUCKAERT *et al.*, 2010). Este ambiente oferece quatro opções de particionamento da base de dados: (i) *use training set*, (ii) *supplied test set*, (iii) *cross-validation* e (iv) *percentage split*.

A opção *supplied test set* permite, diretamente, especificar o conjunto de teste separado do conjunto de treinamento. Esta opção da ferramenta permite que o conjunto de treinamento e o conjunto de teste possam ser trabalhados distintamente. O conjunto de treinamento deve possuir um atributo de classe, que identifica a classe de cada exemplo (registro) do conjunto. A partir deste conjunto de treinamento, os algoritmos fazem a aprendizagem do modelo. O conjunto de teste não precisa de um valor para a classe dos exemplos, pode-se utilizar um ponto de interrogação (?) para indicar valor não informado. O resultado da predição é mostrado pela ferramenta através de um terceiro arquivo com os dados do conjunto de teste e mais um campo (atributo) chamado “*predicted*” que informa o valor predito para cada registro. A importância da

opção *supplied test set* da ferramenta é a facilidade de selecionar os arquivos que compõem o conjunto de treinamento e teste. Por exemplo, pode-se utilizar como conjunto de teste dados antigos para se verificar a precisão dos algoritmos ou utilizar conjunto de teste com dados de novos estudantes e fazer a predição do desempenho acadêmico (ROMERO, VENTURA, 2010). A possibilidade de definir os exemplos que compõem os dois conjuntos pode ser usada para fazer análises quantitativas do desempenho do algoritmo classificador. Este ambiente favorece a realização de experimentos mais próximos das funcionalidades propostas na arquitetura EDM WAVE.

Este ambiente da ferramenta oferece vários recursos para analisar os algoritmos. No entanto, cada algoritmo deve ser executado individualmente. A comparação entre os diferentes algoritmos deve ser feita utilizando outros recursos externos a este ambiente.

4.7.2 Weka Experiment Environment (WEE)

A ferramenta Weka disponibiliza o ambiente *Weka Experiment Environment* (WEE), ele é apropriado para realizar comparações entre o desempenho de vários algoritmos de mineração de dados (SCUSE, REUTEMANN, 2008, BOUCKAERT *et al.*, 2010). O WEE permite selecionar um ou mais algoritmos disponíveis na ferramenta e analisar os resultados de modo a identificar se um classificador é, estatisticamente, melhor do que os demais. Por exemplo, cada algoritmo é executado n vezes e seu desempenho final é a média das n execuções. O ambiente pode ser configurado estabelecendo um número de execuções, no entanto, o número padrão para cada algoritmo selecionado é de 10 execuções.

O WEE oferece três opções de divisão da base de dados: (i) *Cross-validation*, (ii) *Train/Test Percentage Split (data randomized)* e (iii) *Train/Test Percentage Split (order preserved)*. Utilizando o padrão de particionar a base de dados em 10 conjuntos (*10-fold cross-validation*) significa que um classificador é executado 100 vezes para os conjuntos de treinamento e teste.

Este ambiente avalia a acurácia dos algoritmos. Um dos algoritmos avaliados é escolhido como base de comparação (*baseline*), a partir dos resultados obtidos por cada classificador são assinalados aqueles algoritmos que possuem diferença estatística entre os seus resultados e o resultado do algoritmo *baseline*. O padrão de configuração para esta comparação é *pair-wise T-Test* com significância de 5%.

4.7.3 Descrição dos Arquivos Utilizados pelo Weka

A ferramenta Weka utiliza diversos tipos de arquivos para identificar os dados. Nesta tese foram utilizados arquivos de texto com a extensão (.arff). Deve-se utilizar um arquivo para cada modelo de dados de estudantes que se deseja fazer a predição, cada modelo será explicado a seguir nas seções que descrevem os estudos de casos. No entanto, um modelo de dados pode ser utilizado por diversos algoritmos.

Apresentaremos a seguir, o formato utilizado para o arquivo referente ao conjunto de treinamento, conjunto de teste e o arquivo que mostra o resultado obtido após a aplicação do algoritmo de mineração de dados.

A Tabela 4.5 ilustra um exemplo de conjunto de treinamento, o arquivo é composto pela descrição do modelo de dados utilizado, das linhas 3 a 24, e pelos dados, a partir da linha 27. O exemplo utilizado mostra o modelo de dados para predição do progresso do estudante no segundo semestre letivo, tendo como base dados do primeiro semestre letivo. O atributo de classe (Tabela 4.5 linha 24) com a informação da situação acadêmica do estudante no próximo semestre, neste caso, o segundo semestre. Neste exemplo, utilizaram-se três classes distintas para descrever o desempenho acadêmico do estudante.

Tabela 4.5: Estrutura do arquivo Weka (.arff) para o conjunto de treinamento.

1	@relation Treinamento_EC_1S
2	
3	@attribute 01S_CRPeriodo numeric
4	@attribute 01S_SitPeriodo {APROV,ATIVA,PAROU}
5	@attribute 01S_NoDisc numeric
6	@attribute 01S_NoAP numeric
7	@attribute 01S_MediaAP numeric
8	@attribute 01S_NoRFM numeric
9	@attribute 01S_NoRM numeric
10	@attribute 1D_Conceito numeric
11	@attribute 1D_SitDisciplina {AP,RM,RFM}
12	@attribute 2D_Conceito numeric
13	@attribute 2D_SitDisciplina {AP,RM,RFM}
14	@attribute 3D_Conceito numeric
15	@attribute 3D_SitDisciplina {AP,RM,RFM}
16	@attribute 4D_Conceito numeric
17	@attribute 4D_SitDisciplina {AP,RM,RFM}
18	@attribute 5D_Conceito numeric
19	@attribute 5D_SitDisciplina {AP,RM,RFM}
20	@attribute 6D_Conceito numeric
21	@attribute 6D_SitDisciplina {AP,RM,RFM}
22	@attribute 7D_Conceito numeric
23	@attribute 7D_SitDisciplina {AP,RM,RFM}
24	@attribute 02S_SitPeriodo {APROV,PAROU,ATIVA}
25	@data
26	
27	84,APROV,7,7,82.857143,0,0,95,AP,85,AP,76,AP,80,AP,80,AP,73,AP,91,AP,APROV

31	67,APROV,7,7,68.571,0,0,82,AP,71,AP,74,AP,79,AP,55,AP,50,AP,69,AP,?
32	82,APROV,7,7,83.286,0,0,89,AP,87,AP,93,AP,72,AP,82,AP,76,AP,84,AP,?
33	63,APROV,7,7,67.714,0,0,86,AP,74,AP,79,AP,50,AP,63,AP,67,AP,55,AP,?
34	53,APROV,7,6,64.167,0,1,81,AP,71,AP,80,AP,50,AP,53,AP,21,RM,50,AP,?
35	56,APROV,7,6,58.667,0,1,65,AP,52,AP,59,AP,72,AP,53,AP,42,RM,51,AP,?
36	65,APROV,7,7,63.857,0,0,83,AP,54,AP,68,AP,54,AP,54,AP,50,AP,84,AP,?
37	58,APROV,7,7,62.286,0,0,86,AP,71,AP,65,AP,50,AP,54,AP,57,AP,53,AP,?
38	45,APROV,7,4,67.250,0,3,73,AP,38,RM,83,AP,22,RM,53,AP,11,RM,60,AP,?
39	41,APROV,7,3,63.667,0,4,73,AP,30,RM,58,AP,60,AP,35,RM,5,RM,40,RM,?
40	64,APROV,7,7,66.857,0,0,86,AP,75,AP,67,AP,57,AP,50,AP,70,AP,63,AP,?

A Tabela 4.7 ilustra o arquivo resultado da predição para o conjunto de teste. Observa-se que este arquivo gerado pela ferramenta Weka utiliza os mesmos dados do conjunto de teste e acrescenta um novo campo (atributo) que contém a predição da classe para cada exemplo do conjunto de teste. Este valor gerado pelo algoritmo classificador é destacado na linha 25.

Tabela 4.7: Estrutura do arquivo Weka (.arff) com o resultado da predição da classe.

1	@relation Treinamento_EC_1S_predicted
2	
3	@attribute 01S_CRPeriodo numeric
4	@attribute 01S_SitPeriodo {APROV,ATIVA,PAROU}
5	@attribute 01S_NoDisc numeric
6	@attribute 01S_NoAP numeric
7	@attribute 01S_MediaAP numeric
8	@attribute 01S_NoRFM numeric
9	@attribute 01S_NoRM numeric
10	@attribute 1D_Conceito numeric
11	@attribute 1D_SitDisciplina {AP,RM,RFM}
12	@attribute 2D_Conceito numeric
13	@attribute 2D_SitDisciplina {AP,RM,RFM}
14	@attribute 3D_Conceito numeric
15	@attribute 3D_SitDisciplina {AP,RM,RFM}
16	@attribute 4D_Conceito numeric
17	@attribute 4D_SitDisciplina {AP,RM,RFM}
18	@attribute 5D_Conceito numeric
19	@attribute 5D_SitDisciplina {AP,RM,RFM}
20	@attribute 6D_Conceito numeric
21	@attribute 6D_SitDisciplina {AP,RM,RFM}
22	@attribute 7D_Conceito numeric
23	@attribute 7D_SitDisciplina {AP,RM,RFM}
24	@attribute 'prediction margin' numeric
25	@attribute 'predicted 02S_SitPeriodo' {APROV,PAROU,ATIVA}
26	@attribute 02S_SitPeriodo {APROV,PAROU,ATIVA}
27	
28	@data
29	50,APROV,7,4,63.5,1,2,79,AP,32,RM,59,AP,0,RFM,28,RM,66,AP,50,AP,0.657855,APROV,?
30	65,APROV,7,6,75.667,0,1,78,AP,84,AP,93,AP,75,AP,64,AP,35,RM,?,?,1,APROV,?
31	68,APROV,7,5,75,0,2,80,AP,46,RM,75,AP,90,AP,53,AP,47,RM,77,AP,1,APROV,?
32	64,APROV,7,7,67.571,0,0,77,AP,76,AP,75,AP,51,AP,55,AP,77,AP,62,AP,1,APROV,?
33	67,APROV,7,7,68.571,0,0,82,AP,71,AP,74,AP,79,AP,55,AP,50,AP,69,AP,1,APROV,?
34	82,APROV,7,7,83.286,0,0,89,AP,87,AP,93,AP,72,AP,82,AP,76,AP,84,AP,1,APROV,?
35	63,APROV,7,7,67.714,0,0,86,AP,74,AP,79,AP,50,AP,63,AP,67,AP,55,AP,1,APROV,?
36	53,APROV,7,6,64.167,0,1,81,AP,71,AP,80,AP,50,AP,53,AP,21,RM,50,AP,1,APROV,?

37	56,APROV,7,6,58.667,0,1,65,AP,52,AP,59,AP,72,AP,53,AP,42,RM,51,AP,1,APROV,?
38	65,APROV,7,7,63.857,0,0,83,AP,54,AP,68,AP,54,AP,54,AP,50,AP,84,AP,1,APROV,?
39	58,APROV,7,7,62.286,0,0,86,AP,71,AP,65,AP,50,AP,54,AP,57,AP,53,AP,1,APROV,?
40	45,APROV,7,4,67.25,0,3,73,AP,38,RM,83,AP,22,RM,53,AP,11,RM,60,AP,0.998598,APROV,?

4.8 Estudo de Caso 01: Avaliação de 12 Algoritmos Classificadores

Utilizando Dados do Curso de Engenharia Civil e suas Ênfases

Neste estudo, realizaram-se quatro experimentos com o objetivo de comparar o desempenho de doze algoritmos de mineração de dados utilizando diferentes formas de particionar a base de dados. A análise dos resultados serve para identificar se as formas de particionar a base de dados, utilizando: (i) o método de validação cruzada com 10 conjuntos (*10 folds cross-validation*) e (ii) seleção dos conjuntos de treinamento e teste, não interferem nos resultados dos algoritmos. Os algoritmos foram analisados levando-se em consideração duas formas diferentes de dividir a base de dados e utilizando duas classes de estudantes.

Resumidamente, este estudo de caso tem os seguintes objetivos:

- (1) Comparar o desempenho de 12 algoritmos de mineração de dados considerando a acurácia dos classificadores;
- (2) Verificar se a forma de particionar a base de dados interfere no desempenho dos algoritmos;
- (3) Considerar a base de dados dividida em duas classes: não-concluintes e concluintes;
- (4) Identificar os atributos mais relevantes para tratar o problema da evasão nos cursos de graduação.

Parte deste trabalho foi publicada e apresentada no Simpósio Brasileiro de Informática na Educação (SBIE2011). A apresentação foi feita na seção técnica intitulada *Uso de Tecnologias de IA na Educação*, o título do artigo: *Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados*. (MANHÃES *et al.*, 2011).

4.8.1 Descrição dos Algoritmos Utilizados nos Experimentos

Foram comparados os 12 algoritmos disponíveis no Weka com seus valores de configuração padrão: AdaBoost (AD), BayesNet (BN), DecisionTable (DT), J48 (J48),

JRip (JR), MultilayerPerceptron (MP), NaiveBayes (NB), OneR (OR), RandomForest (RF), SimpleLogistic (SL), SVM com PolyKernel (SVM1) e SVM com RBF Kernel (SVM2).

O algoritmo OneR foi escolhido como base de referência (*baseline*), a escolha do algoritmo *baseline* é experimental, optou-se pelo OneR por ser um classificador muito simples, por utilizar um método de classificação de custo reduzido e obter uma acurácia alta (CARVALHO, 2005, HAN, KAMBER, 2006, WITTEN *et al.*, 2011).

Neste estudo de caso, somente a acurácia (Ac.) dos classificadores foi utilizada como parâmetro de comparação entre os algoritmos.

4.8.2 Descrição da Base de Dados dos Experimentos

A base de dados utilizada neste experimento foi extraída do sistema acadêmico SIGA da UFRJ, ela contém informações sobre os estudantes de graduação que ingressaram na Escola Politécnica no curso de Engenharia Civil e suas cinco ênfases. Foram selecionadas as informações acadêmicas referentes ao primeiro semestre letivo de 887 estudantes que ingressaram no período de 1994 a 2005.

A base de dados dos estudantes foi dividida em duas classes distintas e bem definida. A primeira classe identificada como (*não-concluinte*) composta por 324 estudantes que não concluíram o curso por iniciativa própria (abandono ou trancamento de matrícula); ou por imposição da universidade (reprovação por nota, ultrapassar o prazo para conclusão do curso e sanção disciplinar). A segunda classe (*concluinte*) composta por 563 estudantes que concluíram todos os requisitos para aprovação e conclusão do curso de graduação.

Identificaram-se as disciplinas mais cursadas relativas ao primeiro semestre, a saber: Introdução a Engenharia Civil (EEC200), Engenharia e Meio Ambiente (EEH210), Programação de Computadores I (MAB124), Cálculo Diferencial e Integral I (MAC118) e Química (IQG111). Todos os 887 estudantes selecionados da base de dados frequentaram estas disciplinas.

4.8.2.1 Modelo de Dados dos Estudantes

Um dos objetivos deste estudo de caso foi identificar os atributos da base de dados mais relevantes para compor os modelos de dados dos estudantes. A tabela a seguir

ilustra a lista de atributos avaliados neste estudo de caso.

Tabela 4.8: Modelo de dados dos estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando dados do primeiro semestre letivo.

Atributo	Descrição	Valor	Tipo
01S_CR	CR do primeiro período	{0 to 100}	Numérico
EEC200_Conceito	Nota da disciplina	{0 to 100}	Numérico
EEC200_SitDisciplina	Situação da disciplina	{AP,RM,RFM}	String
EEH210_Conceito	Nota da disciplina	{0 to 100}	Numérico
EEH210_SitDisciplina	Situação da disciplina	{AP,RM,RFM}	String
MAB124_Conceito	Nota da disciplina	{0 to 100}	Numérico
MAB124_SitDisciplina	Situação da disciplina	{AP,RM,RFM}	String
MAC118_Conceito	Nota da disciplina	{0 to 100}	Numérico
MAC118_SitDisciplina	Situação da disciplina	{AP,RM,RFM}	String
IQG111_Conceito	Nota da disciplina	{0 to 100}	Numérico
IQG111_SitDisciplina	Situação da disciplina	{AP,RM,RFM}	String
STATUS*	Atributo de classe	{Concluinte,Não-Concluinte}	String

O ganho de informação é um método utilizado para avaliar o quanto um atributo influencia o critério de classificação do algoritmo (HAN, 2005). A tabela mostra os atributos seguindo a ordem do mais importante para o menos importante, como segue: nota na disciplina Cálculo Diferencial e Integral I (MAC118); 01S_CR; situação na disciplina Cálculo Diferencial e Integral I (MAC118); notas das disciplinas: Química (IQG111), Engenharia e Meio Ambiente (EEH210), Programação de Computadores I (MAB124), Introdução a Engenharia Civil (EEC200); situação das disciplinas: Introdução a Engenharia Civil (EEC200), Química (IQG111), Programação de Computadores I (MAB124), Engenharia e Meio Ambiente (EEH210).

Tabela 4.9: Análise da importância dos atributos para classificação segundo o método de Ganho da Informação

=== Run information ===
Evaluator: weka.attributeSelection.InfoGainAttributeEval
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation: EC_887EstudantesTreinamentoTestOrderPreserved
Instances: 887
Attributes: 12
01S_CR
EEC200_Conceito
EEC200_SitDisciplina
EEH210_Conceito
EEH210_SitDisciplina
MAB124_Conceito
MAB124_SitDisciplina
MAC118_Conceito
MAC118_SitDisciplina

IQG111_Conceito		
IQG111_SitDisciplina		
STATUS		
Evaluation mode: 10-fold cross-validation		
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===		
average merit	average rank	attribute
0.358 +- 0.009	1.2 +- 0.4	8 MAC118_Conceito
0.353 +- 0.009	1.8 +- 0.4	1 01S_CR
0.303 +- 0.008	3 +- 0	9 MAC118_SitDisciplina
0.225 +- 0.009	4.2 +- 0.4	10 IQG111_Conceito
0.209 +- 0.007	4.9 +- 0.54	4 EEH210_Conceito
0.198 +- 0.006	5.9 +- 0.3	6 MAB124_Conceito
0.184 +- 0.004	7.1 +- 0.3	2 EEC200_Conceito
0.179 +- 0.004	8.1 +- 0.3	3 EEC200_SitDisciplina
0.174 +- 0.006	8.8 +- 0.6	11 IQG111_SitDisciplina
0.147 +- 0.004	10.3 +- 0.46	7 MAB124_SitDisciplina
0.143 +- 0.007	10.7 +- 0.46	5 EEH210_SitDisciplina

Os resultados apresentados na Tabela 4.9 mostram que o conceito (nota) em Cálculo Diferencial e Integral I (MAC118) e o CR do primeiro período, considerando apenas os dados do primeiro semestre letivo, são os atributos mais importantes para a correta classificação do desempenho acadêmico dos estudantes do curso de Engenharia Civil da UFRJ.

4.8.3 Descrição dos Experimentos e Avaliação dos Resultados

Os experimentos a seguir mostram que foram utilizados diversos períodos disponíveis entre 1994 a 2010 da base de dados original do SIGA para formar os conjuntos de treinamento e teste. A escolha de diversos períodos para compor os experimentos permite minimizar que influências externas que tenham ocorrido em algum período de tempo venham afetar as nossas análises.

4.8.3.1 Experimento 1

Este experimento foi executado no ambiente WEE da ferramenta Weka (SCUSE, REUTEMANN, 2008, BOUCKAERT *et al.*, 2010), foram selecionados doze algoritmos classificadores. A base de dados foi dividida em 10 conjuntos utilizando o método de validação cruzada (*10 folds cross-validation*). Os algoritmos, aplicados a base de dados, foram executados 10 vezes, valor padrão de configuração do ambiente.

Tabela 4.10: Resultados dos classificadores para a predição de duas classes de estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando validação cruzada com 10 conjuntos para dados do primeiro semestre letivo.

Algoritmo	Acurácia
OR	84.98
JR	82.38
AB	81.86
NB	81.14
BN	81.02
RF	80.88
J48	80.86
SL	80.81
DT	80.57
SVM2	79.12
MP	79.00
SVM1	77.13

A ferramenta calculou a média das acurácias obtidas em cada rodada dos classificadores. A ferramenta mostrou que todos os algoritmos avaliados neste experimento eram significativamente diferentes (*Confidence: 0.05 (two tailed)*) com relação ao *baseline* OneR (WITTEN, FRANK, 2005). Observamos que o algoritmo OneR apresenta o maior valor e o SVM1 o menor, os demais algoritmos não apresentam valores significativos diferentes entre si.

4.8.3.2 Experimento 2

Este experimento foi realizado no ambiente WEE da ferramenta Weka (SCUSE, REUTEMANN, 2008, BOUCKAERT *et al.*, 2010). A configuração do ambiente foi modificada para contemplar outra forma de particionar a base de dados, *Train/Test Percentage Split (data randomized)* é uma opção da ferramenta que utiliza um processo randômico para selecionar os exemplos e dividir a base em dois conjuntos distintos treinamento e teste. O padrão da ferramenta seleciona 66% dos registros para o conjunto de treinamento e 34% para o conjunto de teste. O ambiente WEE utiliza a mesma base de dados e executa todos os algoritmos selecionados 10 vezes, como no experimento 1, o ambiente pode ser configurado para alterar o número de execuções, o padrão é 10 execuções para cada algoritmo selecionado.

O objetivo deste experimento é verificar se a alteração na forma de seleção da base de dados utilizando conjunto de treinamento e teste afeta a porcentagem de acerto dos classificadores.

Tabela 4.11: Resultados dos classificadores para a predição de duas classes de estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando seleção randômica do conjunto de treinamento e teste para dados de

estudantes utilizando dados do primeiro semestre letivo.

Algoritmo	Acurácia
OR	84.69
AB	82.70
NB	81.94
SL	81.64
JR	81.61
DT	81.41
BN	81.35
J48	81.21
RF	80.68
SVM2	79.32
SVM1	78.89
MP	78.53

A ferramenta calculou a média das acurácias obtidas nas 10 execuções dos classificadores. Observou-se que os resultados foram bem próximos dos obtidos no primeiro experimento.

Comparando os resultados dos experimentos 1 e 2, verificam-se que as duas formas de dividir a base de dados não interferiram no percentual de acerto dos algoritmos.

4.8.3.3 Experimento 3

O terceiro experimento foi realizado no ambiente WE da ferramenta Weka, neste ambiente é necessário executar cada um dos 12 algoritmos individualmente. Neste experimento a base de dados utilizada contempla 887 estudantes do curso de Engenharia Civil, esta base foi dividida em 10 conjuntos (*10-fold cross-validation*). A tabela mostra a acurácia (Ac.), verdadeiro positivo (VP), falso negativo (FN), verdadeiro negativo (VN), falso positivo (FP) e matriz de confusão (MC).

Tabela 4.12: Resultados dos classificadores para a predição de duas classes de estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando validação cruzada com 10 conjuntos para dados do primeiro semestre letivo.

Algoritmo	Ac.	VP	FN	VN	FP	MC
OR	84.89	0.72	0.28	0.92	0.08	234 90 44 519
JR	82.86	0.72	0.28	0.89	0.11	233 91 61 502
AB	82.19	0.6	0.4	0.95	0.05	193 131 27 536
RF	81.29	0.73	0.27	0.86	0.14	235 89 77 486
DT	81.17	0.63	0.37	0.91	0.09	205 119 48 515
J48	81.17	0.68	0.32	0.89	0.11	220 104 63 500

SL	80.95	0.63	0.37	0.91	0.09	205 119 50 513
NB	80.95	0.65	0.35	0.9	0.1	210 114 55 508
BN	80.38	0.62	0.38	0.91	0.09	202 122 52 511
SVM2	79.03	0.6	0.4	0.9	0.1	194 130 56 507
MP	78.69	0.61	0.39	0.89	0.11	197 127 62 501
SVM1	76.66	0.59	0.41	0.87	0.13	191 133 74 489

4.8.3.4 Experimento 4

O quarto experimento foi realizado no ambiente WE da ferramenta Weka. Neste experimento a base de dados foi particionada em dois conjuntos: treinamento e teste. O ambiente WE oferece a opção *Supplied test set* que permite especificar os conjuntos de treinamento e teste. Este experimento, em particular, ilustra como a arquitetura EDM WAVE realiza a predição do desempenho acadêmico dos estudantes.

Neste experimento, o conjunto de treinamento é composto 599 registros (68%), 234 da classe (não-concluente) (39%) e 365 da classe (concluente) (61%). O conjunto de teste composto por 288 registros (32%), 90 estudantes da classe (não-concluente) (31%) e 198 da classe (concluente) (69%). *Não-concluente* corresponde aos alunos da base que não concluíram o curso de graduação e a classe *concluente* corresponde aos alunos que concluíram o curso (formados).

No conjunto de teste um ponto de interrogação foi colocado no lugar do valor a ser retornado como resultado da predição. Os resultados referentes ao desempenho dos classificadores não são diretamente obtidos da ferramenta necessitando de cálculos adicionais.

A Tabela 4.13 apresenta a acurácia (Ac.), verdadeiro positivo (VP), falso negativo (FN), verdadeiro negativo (VN) e falso positivo (FP) e matriz de confusão (MC), considerando-se o conjunto de teste da base de dados dos estudantes da Engenharia Civil.

Tabela 4.13: Resultados dos classificadores para a predição de duas classes de estudantes especificando o conjunto de treinamento e teste para estudantes do curso de Engenharia Civil que ingressaram no período de 1994 a 2005 utilizando dados do primeiro semestre letivo.

Algoritmo	Ac.	VP	FN	VN	FP	MC
SL	82.29	0.64	0.36	0.94	0.06	70 40 11 167
NB	81.25	0.65	0.35	0.91	0.09	72 38 16 162
BN	80.56	0.65	0.35	0.90	0.10	71 39 17 161
J48	80.21	0.64	0.36	0.90	0.10	70 40 17 161
JR	80.21	0.74	0.26	0.84	0.16	81 29 28 150
OR	79.86	0.65	0.35	0.89	0.11	71 39 19 159
SVM2	79.86	0.56	0.44	0.94	0.06	62 48 10 168
RF	79.17	0.76	0.24	0.81	0.19	84 26 34 144
SVM1	78.82	0.48	0.52	0.98	0.02	53 57 4 174
AB	76.39	0.68	0.32	0.81	0.19	75 35 33 145
MP	75.69	0.70	0.30	0.79	0.21	77 33 37 141
DT	75.00	0.66	0.34	0.80	0.20	73 37 35 143

4.8.3.5 Discussão dos Resultados dos Experimentos

Estes experimentos são importantes porque mostram o desempenho dos algoritmos quando se mantém a mesma base de dados e se modifica a forma de particionar esta base de dados.

A Tabela 4.14 ilustra os resultados das acurácias dos 12 algoritmos e o desvio padrão obtidos em cada experimento. As acurácias obtidas estão muito próximas, portanto, os quatro experimentos mostram que as formas de particionar a base de dados não interferem na acurácia dos algoritmos.

Tabela 4.14: Acurácia, média das acurácias e desvio padrão dos Experimentos.

Algoritmo	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Média	DP
DT	80.57	81.41	81.17	75.00	79.54	3.05
AB	81.86	82.70	82.19	76.39	80.79	2.95
OR	84.98	84.69	84.89	79.86	83.61	2.50
MP	79.00	78.53	78.69	75.69	77.98	1.54
JR	82.38	81.61	82.86	80.21	81.77	1.16
SVM1	77.13	78.89	76.66	78.82	77.88	1.15
RF	80.88	80.68	81.29	79.17	80.51	0.93
SL	80.81	81.64	80.95	82.29	81.42	0.68
J48	80.86	81.21	81.17	80.21	80.86	0.46
BN	81.02	81.35	80.38	80.56	80.83	0.44
NB	81.14	81.94	80.95	81.25	81.32	0.43
SVM2	79.12	79.32	79.03	79.86	79.33	0.37

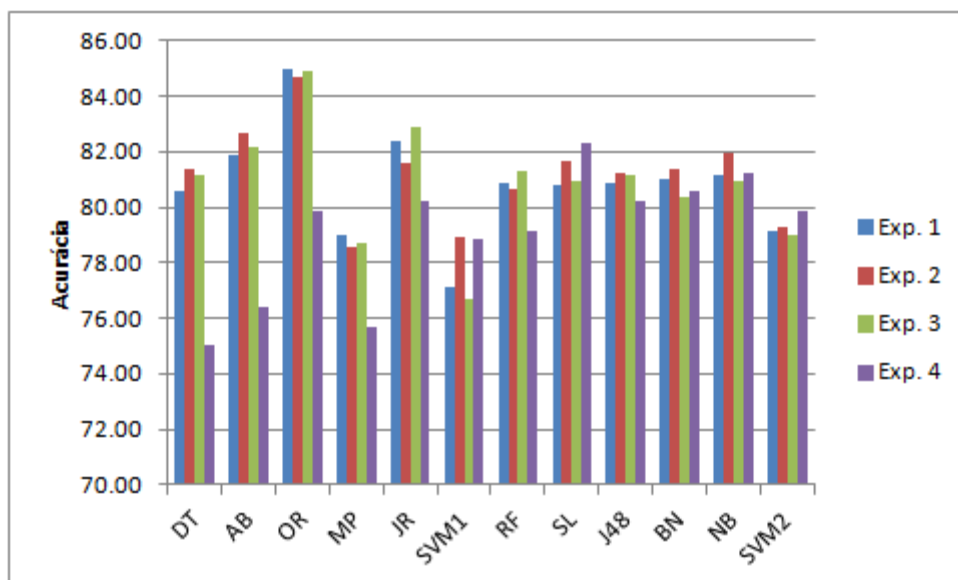


Figura 4.1: O gráfico ilustra as acurácias dos classificadores obtidos nos Experimentos.

Observando a tabela e o gráfico acima, o algoritmo OR possui uma acurácia superior a 84.0 para os três primeiros experimentos, no experimento 4 a acurácia obtida é menor, no entanto, aproximadamente 80%. Os algoritmos DT, AB, JR e RF possuem acurácia acima de 80% para os três primeiros experimentos, mas a acurácia do experimento 4 é bem inferior com relação aos três primeiros experimentos. Os algoritmos MP, SVM1 e SVM2 possuem acurácias inferior a 80.0 para todos os experimentos. Os algoritmos SL, J48, BN e NB são os que apresentam maior homogeneidade entre as acurácias dos quatros experimentos e acima de 80%. Entre estes algoritmos, observamos que o algoritmo NB é que possui acurácias mais altas em todos os quatro experimentos e que possui menor desvio padrão entre as acurácias.

Os experimentos 3 e 4 efetuados no ambiente WE da ferramenta Weka permite utilizar outras métricas para avaliar os algoritmos classificadores. Além da acurácia, um algoritmo pode diferir do outro nos valores das taxas de acerto e erro na classificação dos exemplos positivos e negativos. A taxa de erro ou falso negativo (FN) mostra que o algoritmo possui uma elevada possibilidade de classificar um estudante como concluinte quando de fato o estudante não concluíra o curso. Neste caso, considera-se um erro grave do algoritmo. Por outro lado, a taxa de erro ou falso positivo (FP) é um erro brando do classificador porque atribui a um estudante concluinte uma classificação errada como não concluinte. Neste caso, é menos grave do que classificar um estudante com risco de evasão como sem risco. O erro do algoritmo em classificar um estudante

no grupo de risco de evasão sem de fato ocorrer à evasão, falso positivo, é considerado um erro brando, menos grave. Observam-se que os algoritmos classificadores apresentaram taxa de erro bastante diferenciadas. No Experimento 3, os algoritmos que possuem maior taxa de erro grave são: SVM1 (41%), AB e SVM2 (40%), MP (39%), BN (38%), DT e SL (37%), os algoritmos NB, J48, JR e OR possuem erro grave menor que 35%. No Experimento 4 os algoritmos tiveram erro grave ainda maior, SVM1 (52%) e SVM2 (44%) os demais algoritmos abaixo de 36%.

Os experimentos 3 e 4 mostram que o classificador (RF) apresentou melhor resultado para a predição da classe (não-concluente) taxa de acerto (VP) superior a 73%. O classificador SVM1 apresentou o pior resultado (48%) para a taxa de acerto (VP) para a classe (não-concluente).

Outra análise feita para identificar as causas dos erros dos classificados foi realizada diretamente na base de dados. Observou-se que os dados (características) de alguns estudantes não seguiam o padrão das classes a qual eles pertenciam. Por exemplo, estudantes com rendimento acadêmico abaixo da média para o curso de graduação, concluíram o curso. Outro grupo de estudantes que possui comportamento fora do padrão da classe de estudantes que evade, são os que possuem rendimento acadêmico elevado, mas não completaram o curso. Estes casos comprometem a predição feita pelos classificadores refletindo aumento na taxa de erro. No entanto, a remoção destes exemplos não é aconselhável porque a base perderia seu reflexo da realidade.

De um modo geral, os desempenhos obtidos pelos algoritmos de mineração de dados dos mais simples aos mais sofisticados foram semelhantes. No entanto, a acurácia dos classificadores e a taxa de erro são fortemente influenciadas pelas características dos dados, isto é, estudantes que evadem do curso mesmo com rendimento acadêmico alto e estudantes que concluem o curso com rendimento acadêmico abaixo da média, esses casos estão fora do padrão das classes aprendidas pelos algoritmos classificadores.

Embora, a acurácia não seja uma das formas mais adequadas para avaliar os algoritmos. No entanto, as acurácias são superiores a 75% para todos os classificadores investigados. Os resultados preliminares destes 4 algoritmos mostram que os atributos utilizados no modelo de dados dos estudantes são importantes e relevantes para serem considerados na predição do desempenho acadêmico dos estudantes. Estes resultados mostram que é possível fazer a predição de estudantes com risco de evasão com um número reduzido de atributos. Verificou-se que o atributo mais importante para esta base de dados é a *nota na disciplina de Cálculo Diferencial e Integral I* e o coeficiente

de rendimento do primeiro semestre letivo.

4.9 Estudo de Caso 02: Uma Abordagem Quantitativa dos Fatores que Influenciam o Desempenho Acadêmico dos Estudantes de Graduação da UFRJ

A predição do desempenho acadêmico dos estudantes de graduação passa por uma importante etapa que é identificar quais são os fatores que caracterizam os estudantes ao longo do curso de graduação. Na base de dados fornecida pelo SIGA identificamos três classes de estudantes como descritas a seguir: (i) *cancelados* - estudantes que interromperam o curso em algum período antes da formatura (evasão); (ii) *ativos fora do prazo* - estudantes que permaneceram matriculados além do prazo médio para conclusão do curso; e (iii) *concluintes* - estudantes que concluíram o curso de graduação.

Portanto, este estudo de caso tem os seguintes objetivos:

- (1) Identificar as três classes de estudantes;
- (2) Verificar quais algoritmos de mineração de dados são melhores na predição de três classes de exemplos de estudantes;
- (3) Identificar os principais fatores que distinguem as três classes de estudantes que frequentam os cursos de graduação da UFRJ;
- (4) Identificar qual algoritmo possui melhor resultado e que possa ser interpretável e convertido em dados gráficos;
- (5) Apresentar uma análise quantitativa, a fim de identificar, apresentar e quantificar as variáveis que representam os principais fatores que influenciam a conclusão, a evasão e permanência além do tempo médio para concluir o curso de graduação.

Parte deste trabalho foi publicado em dois artigos: o primeiro apresentado no VIII Simpósio Brasileiro de Sistema de Informação (SBSI 2012), o título do artigo: *Identificação dos Fatores que Influenciam a Evasão em Cursos de Graduação Através de Sistemas Baseados em Mineração de Dados: uma Abordagem Quantitativa* (MANHÃES *et al.*, 2012). O segundo artigo publicado no 6th International Conference on Computer Supported Education - CSEDU 2014, o título do artigo: *Identifying the Factors Related to Differents Undergraduate Students' Academic Performance Using*

4.9.1 Descrição dos Algoritmos Utilizados no Experimentos

Foram comparados os 12 algoritmos disponíveis no Weka com seus valores de configuração padrão: AdaBoost (AD), BayesNet (BN), DecisionTable (DT), J48 (J48), JRip (JR), MultilayerPerceptron (MP), NaiveBayes (NB), OneR (OR), RandomForest (RF), SimpleLogistic (SL), SVM com PolyKernel (SVM1) e SVM com RBF Kernel (SVM2).

4.9.2 Descrição da Base de Dados Utilizada nos Experimentos

A base de dados utilizada neste estudo de caso é compartilhada com o SIGA. Neste estudo de caso, foram selecionados dados acadêmicos dos estudantes que ingressaram nos dois semestres letivos dos anos de 2003 e 2004. A base contempla estudantes de 250 cursos de graduação oferecidos por 28 unidades da UFRJ. Escolheram-se os estudantes que ingressaram nos anos de 2003 e 2004 porque estes já dispõem de situação acadêmica definida, previamente registrada no sistema acadêmico através do atributo "*situacaoMatriculaAtual*" (Tabela 4.1).

A Tabela ilustra os quatro conjuntos de estudantes subdivididos nas três situações estabelecidas neste trabalho, as situações finais foram obtidas depois de 12 semestres letivos a partir do ano de ingresso (início do curso).

Tabela 4.15: Quantidade de estudantes distribuídos nas três classes por ano de ingresso.

Ano de ingresso	Cancelado	Ativo FP	Concluente	Total
2003-1	1448 (0.38)	365 (0.10)	1995 (0.52)	3808
2003-2	1204 (0.40)	342 (0.11)	1494 (0.49)	3040
2004-1	1733 (0.41)	605 (0.14)	1900 (0.45)	4238
2004-2	1255 (0.40)	616 (0.20)	1280 (0.41)	3151

4.9.2.1 Modelo de Dados dos Estudantes

O modelo de dados dos estudantes utilizado neste estudo de caso é apresentado na (Tabela 4.3). Todos os atributos foram considerados até o 12º semestre letivo a partir do ano/semestre de ingresso nos cursos.

Para realizar o primeiro objetivo proposto neste estudo de caso, as três classes de estudantes foram identificadas a partir da redução e adaptação do atributo original

“*situacaoMatriculaAtual*” do SIGA. O atributo “*Status*” foi criado como atributo de classe, o atributo possui três valores que descrevem a situação final da matrícula do estudante: *cancelado*, *ativa fora do prazo (AFP)* e *concluinte*. O termo “*cancelado*” foi atribuído a todos os estudantes com matrícula cancelada por: (i) iniciativa do estudante: cancelamento ou trancamento da matrícula; e (ii) iniciativa da instituição: matrícula cancelada por abandono do curso, não cumprimento das exigências curriculares e outros. O termo “*ativa fora do prazo*” (AFP) foi atribuído a todos os estudantes que tinham matrícula ativa e ultrapassaram o prazo médio para conclusão do curso, até o 12º semestre a partir do ano de ingresso no curso de graduação. Por fim, o termo “*concluinte*” foi atribuído a todos os estudantes que cumpriram com todos os requisitos da grade curricular do curso e foram diplomados.

4.9.3 Descrição dos Experimentos e Avaliação dos Resultados

4.9.3.1 Experimento 1

A abordagem neste experimento está voltada para atingir os seguintes objetivos propostos para este estudo de caso: (2) verificar quais algoritmos de mineração de dados são melhores na predição de três classes de exemplos de estudantes; e (4) identificar qual algoritmo possui melhor resultado e que possa ser interpretável e convertido em dados gráficos.

Nesta seção, apresentamos os resultados obtidos após a aplicação dos algoritmos classificadores sobre a base de dados do sistema SIGA. Ressalta-se que nem todos os algoritmos implementados na ferramenta dão suporte a análise de três diferentes classes de exemplos. Portanto, a construção de modelos mais complexos envolve a escolha dos algoritmos que suportam análise multiclasse. Além disso, os algoritmos mais sofisticados demandam mais tempo para construir os modelos e os mais simples perdem um pouco na precisão dos modelos, mas ganham na flexibilidade e interpretabilidade (SUMATHI, SIVANANDAM, 2006).

Neste estudo de caso optou-se por utilizar o método de validação cruzada com o número de conjuntos igual a 10 ($k=10$), devido ao grande número de exemplos disponíveis na base de dados.

A Tabela 4.16 mostra o tempo de execução em segundos para a construção do modelo. Observou-se que alguns classificadores levaram um tempo considerável para a construção dos modelos para três classes de estudantes, a construção do modelo para

duas classes foi consideravelmente menor. A construção do modelo do algoritmo Multilayer Perceptron (MP) levou mais de 12 horas de processamento. A tabela mostra o número de instancias da base de dados corretamente classificadas (acurácia) e o número de instancias incorretamente classificadas. As menores taxas de acerto dos classificadores ficaram em torno de 80%. A tabela mostra a matriz de confusão para as três classes, As taxas de acerto (VP) e as taxas de erro (FP) para as três classes. E o valor do *Kappa* para cada um dos modelos gerados pelos classificadores.

Tabela 4.16: Análise do desempenho dos classificadores segundo critérios quantitativos para estudantes ingressaram 2003-1.

Crítérios	OR	BN	NB	MP	SL	SVM1
Tempo exec.(s)	0.09	0.36	0.22	5626.39	687.71	154.53
Corretamente Classificada	3005 (78.91%)	3000 (78.78%)	3036 (79.73%)	3147 (82.64%)	3332 (87.50%)	3328 (87.39%)
Incorretamente Classificada	803 (21.09%)	808 (21.22%)	772 (20.27%)	661 (17.36%)	476 (12.50%)	480 (12.61%)
Kappa	0.61	0.64	0.66	0.69	0.78	0.78
MC	1066 51 331 105 49 211 63 42 1890	1070 164 214 27 257 81 47 275 1673	1081 148 219 27 242 96 39 243 1713	1081 148 219 27 242 96 39 243 1713	1254 62 132 76 179 110 39 57 1899	1249 64 135 77 184 104 51 49 1895
VP Cancelado	0.74	0.74	0.74	0.83	0.87	0.86
VP Ativo	0.13	0.70	0.66	0.40	0.49	0.50
VP Concluente	0.95	0.84	0.86	0.90	0.95	0.95
FP Cancelado	0.07	0.03	0.03	0.10	0.05	0.05
FP Ativo	0.03	0.13	0.11	0.03	0.04	0.03
FP Concluente	0.19	0.16	0.17	0.19	0.13	0.13
Crítérios	SVM2	AB	DT	JR	J48	RF
Tempo exec.	456.39	0.22	11.67	5.2	1.31	0.62
Corretamente Classificada	3239 (85.06%)	2972 (78.05%)	3152 (82.77%)	3217 (84.48%)	3152 (82.77%)	3258 (85.56%)
Incorretamente Classificada	569 (14.94%)	836 (21.95%)	656 (17.23%)	591 (15.52%)	656 (17.23%)	550 (14.44%)
Kappa	0.73	0.59	0.69	0.73	0.69	0.74
Matriz de confusão	1157 42 249 49 136 180 27 22 1946	1153 0 95 193 0 172 176 0 1819	1148 43 257 64 128 173 63 56 1876	1176 71 201 44 177 144 51 80 1864	1172 68 208 67 126 172 84 57 1854	1211 57 180 63 146 156 44 50 1901
VP Cancelado	0.80	0.80	0.79	0.81	0.81	0.84
VP Ativo	0.37	0.00	0.35	0.49	0.35	0.40
VP Concluente	0.98	0.91	0.94	0.93	0.93	0.95
FP Cancelado	0.03	0.16	0.05	0.04	0.06	0.05
FP Ativo	0.02	0.00	0.03	0.04	0.04	0.03
FP Concluente	0.23	0.26	0.24	0.19	0.21	0.19

Observamos que todos os classificadores apresentaram valores consideravelmente satisfatórios com relação à taxa de acerto geral do classificador (acurácia), em torno de 78% ou superior. A taxa de acerto da classe *Cancelados* possui o menor valor (0.74) para os classificadores (ON), (BN) e (NB). A taxa de acerto para a classe *Ativo* foi melhor entre os classificadores (BN) e (NB), o algoritmo (OR) obteve um valor baixo

0.13. Os algoritmos mais sofisticados perdem na questão do tempo para construção dos modelos e na complexidade dos modelos que são construídos. Portanto, dos algoritmos avaliados o (NB) mostrou resultados aceitáveis e ofereceu maior facilidade na conversão dos modelos resultante para valores a serem convertidos na criação de gráficos.

4.9.3.2 Experimento 2: Abordagem Quantitativa

Os modelos gerados pelos algoritmos de mineração de dados apresentam um determinado nível de interpretabilidade, isso significa o quanto o modelo gerado pode ser compreendido ou interpretado pelo humano.

O *Naive Bayes* não sendo o que apresentou melhor acurácia (79.73%), mas o seu modelo de representar o conhecimento aprendido é mais facilmente interpretado e transformado em gráficos.

Os gráficos da Figura 4.2 mostram no eixo abscissas (x) os semestres letivos cursados a partir de 2003-1 a 2008-2 (12 semestres letivos a partir da data de ingresso em 2003-1). No eixo das ordenadas (y) o número de disciplinas cursadas. Todos os gráficos contemplam as três classes de estudantes (*Cancelado*, *Ativa fora do prazo (AFP)* e *Concluinte*), conforme mostram as legendas, a cor (vermelho) para os *Cancelados*, amarelo para *Ativa fora do prazo (AFP)* e verde para *Concluintes*.

A Figura 4.2 mostra os seguintes gráficos: (a) ilustra o número de disciplinas cursadas. O gráfico (b) apresenta o número de disciplinas aprovadas em cada semestre letivo. O gráfico (c) mostra o número de disciplinas onde os estudantes foram reprovados por média (RM). O gráfico (d) ilustra o número de disciplinas onde os estudantes foram reprovados por falta e média (RFM).

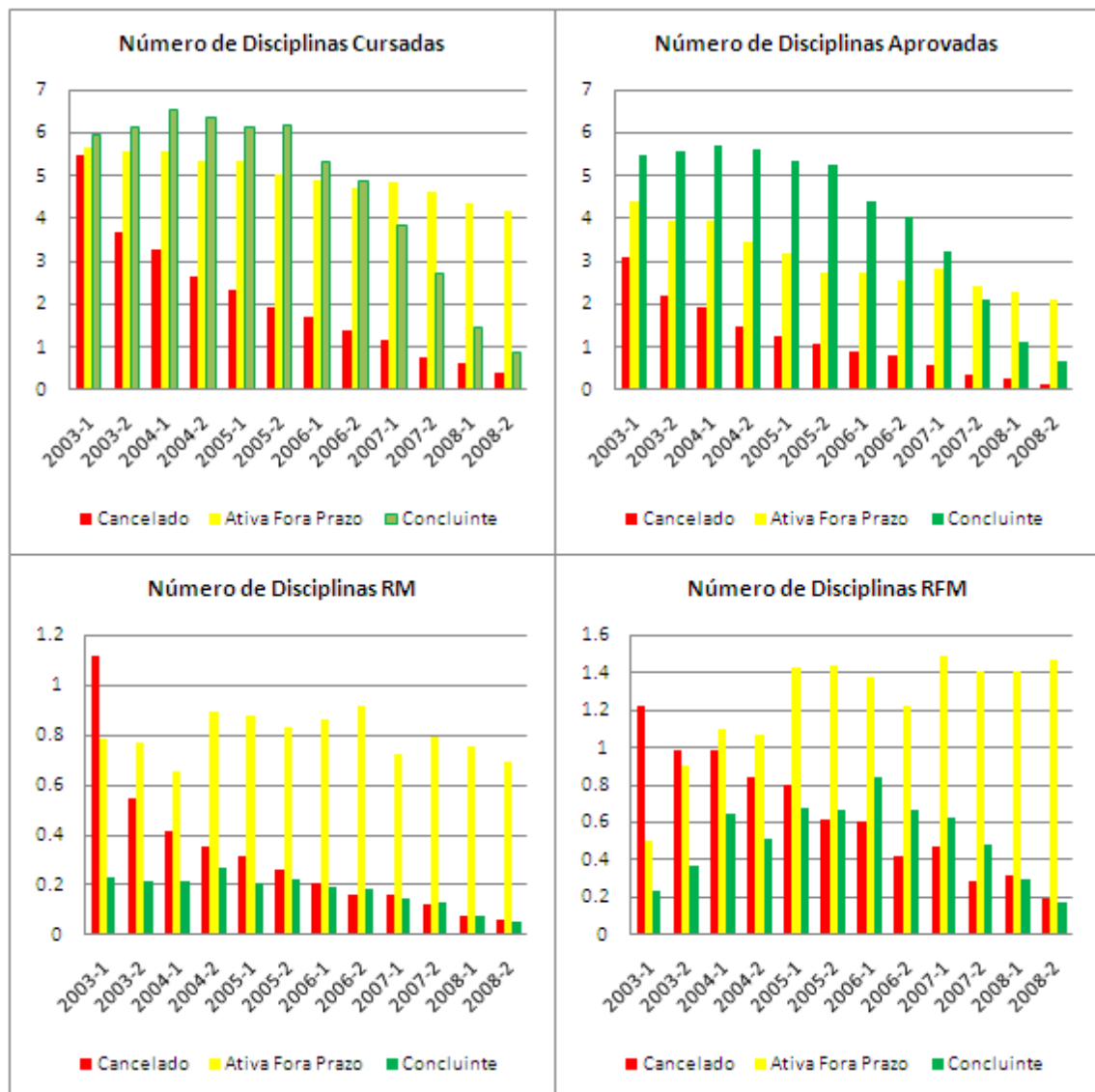


Figura 4.2: Da esquerda para direita temos os gráficos: (a) número de disciplinas cursadas; (b) número de disciplinas aprovadas; (c) número de disciplinas RM; e (d) número de disciplinas RFM.

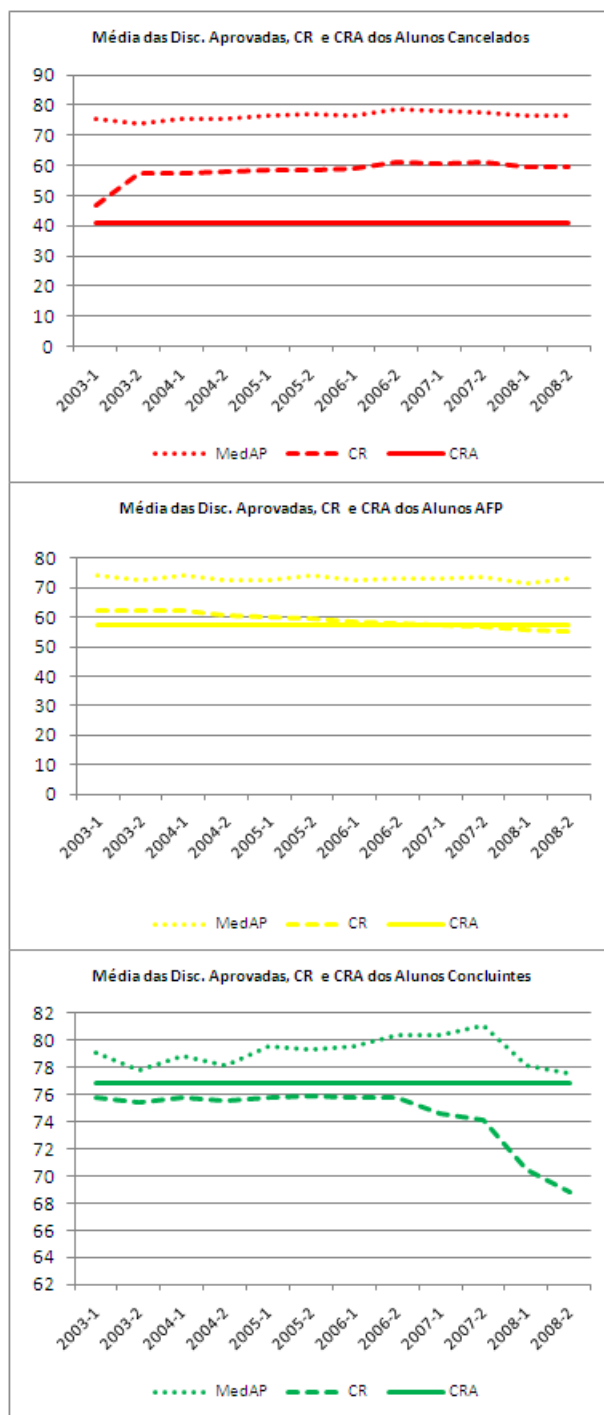


Figura 4.3: De cima para baixo temos os gráficos que apresentam média das disciplinas aprovadas, CR do período e CRA dos estudantes: (a) Cancelados, (b) AFP e (c) Concluintes.

Para os estudantes da classe *Concluinte* possuem valores de média das disciplinas aprovadas, CR período e CRA elevado acima de 76.0. A média das disciplinas aprovadas e o alto número de disciplinas inscritas distinguem este grupo dos demais dois grupos de estudantes.

A seguir foram destacadas as informações mais relevantes sobre os fatores que caracterizam as três classes de estudantes analisadas neste estudo de caso.

Observa-se que os estudantes com matrícula cancelada possuem as seguintes características:

- (1) O número de disciplinas cursadas reduz ao longo dos períodos letivos (Gráfico 4.2a);
- (2) O número de disciplinas aprovadas reduz a cada período (Gráfico 4.2b), acompanhando o número de disciplinas cursadas (inscritas) (Gráfico 4.2a);
- (3) Possuem pelo menos uma disciplina RM no primeiro período (Gráfico 4.2c);
- (4) Possuem pelo menos uma disciplina RFM no primeiro período (Gráfico 4.2d);
- (5) A média das disciplinas aprovadas está entre 70 a 80, como as demais classes de estudantes (Gráfico 4.3a), mas o número de disciplinas cursadas é menor a cada semestre (Gráfico 4.2a). Isso indica que o aluno se dedica a poucas disciplinas e consegue nota alta neste número reduzido de disciplinas aprovadas (Gráfico 4.2b);
- (6) No primeiro ano letivo o CR é o menor se comparado com as demais classes (Gráfico 4.3a);
- (7) A média do CR aumenta ao longo do curso (Gráfico 4.3a), mas para um número reduzido de disciplinas matriculadas (Gráfico 4.3a);
- (8) A média do CRA para esta classe é em torno de 41, o menor valor entre as três classes analisadas (Gráfico 4.3a).

A partir das análises quantitativas, observa-se que os estudantes com matrículas ativas fora do prazo (AFP) são estudantes que apresentam um comportamento regular ao longo de todo o curso. Apesar de se matricular em 5 ou 6 disciplinas por semestre, possuem em torno de duas disciplinas reprovadas por nota ou por abandono. Isto compromete o CR e o CRA distinguindo dos demais graduandos.

Com relação ao subconjunto de estudantes que mantiveram suas matrículas ativas até o ano/período de 2009-1, eles apresentaram as seguintes características:

- (1) Matriculam-se em torno de 5 disciplinas por semestre letivo, número superior a classe Cancelados e inferior a classe Concluinte (Gráfico 4.2a);
- (2) O número de disciplinas aprovadas está em torno de 2 a 4, se destacam dos cancelados, que possuem um número menor, e dos concluintes, que possuem um número maior de disciplinas aprovadas, como mostra o (Gráfico 4.2a);

- (3) Possuem, ao longo do curso, em torno de uma disciplina RM. O (Gráfico 4.2c) destaca a diferença entre as classes distintas de graduandos;
- (4) Possuem, ao longo do curso, uma ou mais disciplinas RFM (Gráfico 4.2a);
- (5) Os (Gráficos 4.2c e d) ilustram que a classe AFP possui pelo menos duas disciplinas RM ou RFM, ao longo de todo o período de matrícula no curso de graduação;
- (6) A média das disciplinas aprovadas é alta como as demais classes (Gráfico 4.3b);
- (7) As médias do CR e CRA são bem próximas em torno de 60, e se destacam das demais classes como mostra o gráfico (Gráfico 4.3b).

Os estudantes que concluíram o curso possuíam uma regularidade de comportamento durante todo o curso. Destacamos: o número elevado de disciplinas cursadas e a média alta de notas nas disciplinas. Observa-se que, o número de reprovações aumenta ao final do curso, provavelmente em função do estágio curricular ou outra atividade. Com relação aos estudantes que concluíram o curso, verifica-se que estes apresentam as seguintes características:

- (1) Mantêm um número alto de disciplinas cursadas, igual ou superior a seis, diminuindo nos últimos períodos do curso (Gráfico 4.2a);
- (2) Possuem um alto índice de aprovações nas disciplinas, superior as demais classes (Gráfico 4.2b);
- (3) O número de disciplinas RM destes estudantes é próximo de zero durante todo curso (Gráfico 4.2c);
- (4) O número de disciplinas RFM aumenta nos últimos semestres do curso (Gráfico 4.2d);
- (5) A média das disciplinas aprovadas é alta superior a 78 (Gráfico 4.3c);
- (6) A média do CR é elevada em torno de 76 (Gráfico 4.3c), no entanto nos últimos semestres o CR diminui acompanhando um aumento nas disciplinas reprovada RFM (Gráfico 4.2d);
- (7) A média do CRA para este grupo de estudantes é de 76.90 (Gráfico 4.3c).

4.9.3.3 Experimento 3: Abordagem Quantitativa

Procedimentos análogos utilizando o algoritmo *Naive Bayes* foram realizados para os subconjuntos de estudantes que ingressaram 2003-2, 2004-1 e 2004-2. Os resultados

apresentaram valores bem próximos dos obtidos para base de 2003-1.

Tabela 4.17: Análise do desempenho do classificador *Naive Bayes* para estudantes ingressaram 2003-1, 2003-2, 2004-1 e 2004-2.

Crítérios	2003-1	2003-2	2004-1	2004-2
Tempo exec. (s)	0.20	0.12	0.20	0.13
Corretamente Classificada	3031 (79.60%)	2412 (79.34%)	3440 (81.17%)	2564 (81.37%)
Incorretamente Classificada	777 (20.40%)	628 (20.66%)	798 (18.83%)	587 (18.63%)
<i>Kappa</i>	0.65	0.66	0.70	0.71
Matriz de confusão	1080 148 220 27 241 97 40 245 1710	893 120 191 16 236 90 39 172 1283	1367 151 215 23 423 159 41 209 1650	972 104 179 9 461 146 31 118 1131
VP Cancelado	0.75	0.74	0.79	0.78
VP Ativo	0.66	0.69	0.70	0.75
VP Concluinte	0.86	0.86	0.87	0.88
FP Cancelado	0.03	0.03	0.03	0.02
FP Ativo	0.11	0.11	0.10	0.09
FP Concluinte	0.18	0.18	0.16	0.17

A seguir serão apresentados gráficos comparando os quatro semestres de ingresso: 2003-1, 2003-2, 2004-1 e 2004-2.

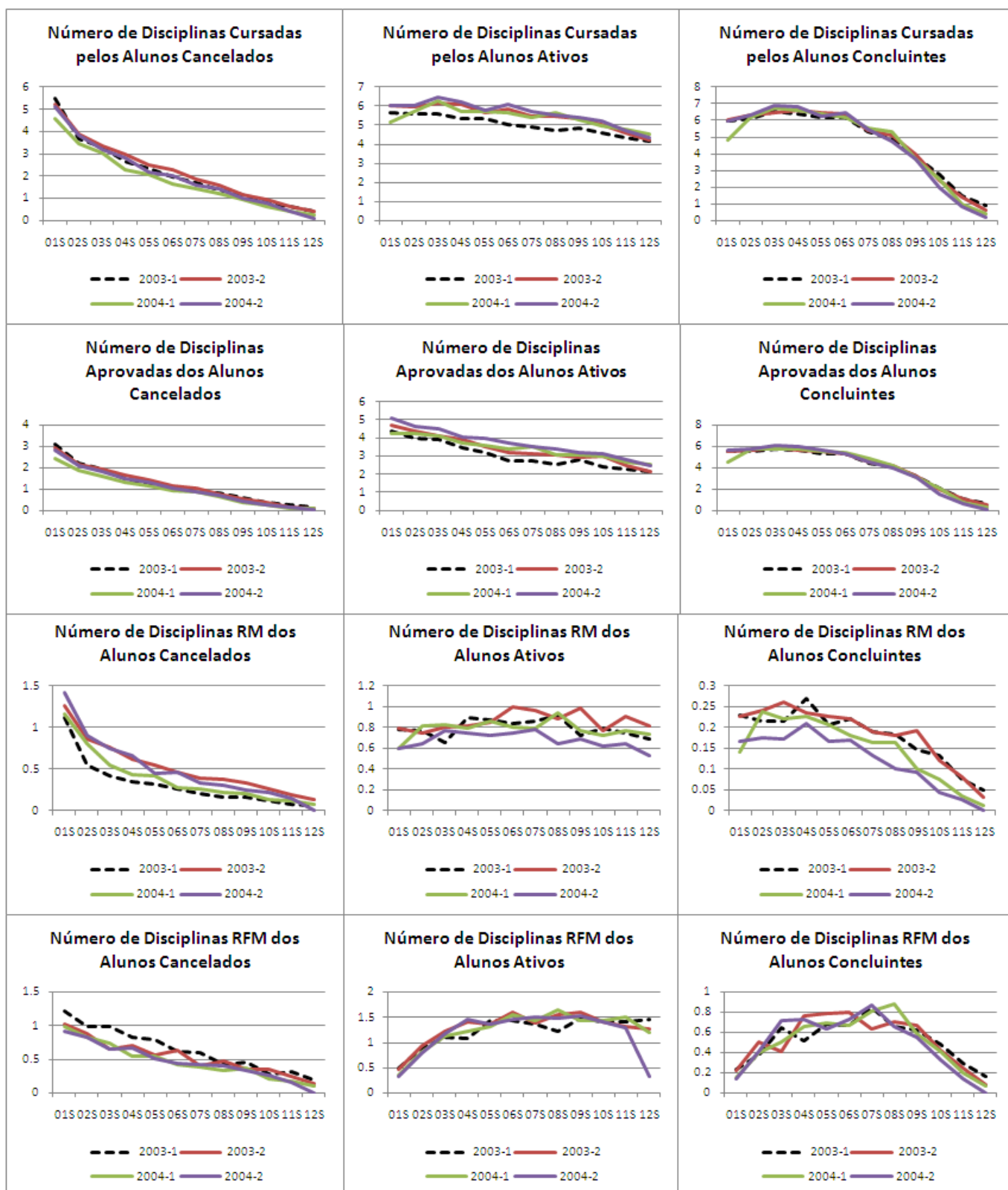


Figura 4.4: Da esquerda para a direita, temos os gráficos que mostram o número de disciplinas cursadas para os estudantes: (a) cancelados, (b) AFP e (c) concluintes. Os gráficos que apresentam o número de disciplinas aprovadas para: (d) cancelados, (e)

ativos e (f) concluintes. Os gráficos com o número de disciplinas RM para estudantes: (g) cancelados, (h) ativos e (i) concluintes. Os gráficos com o número de disciplinas RFM para estudantes: (j) cancelados, (k) ativos e (l) concluintes.

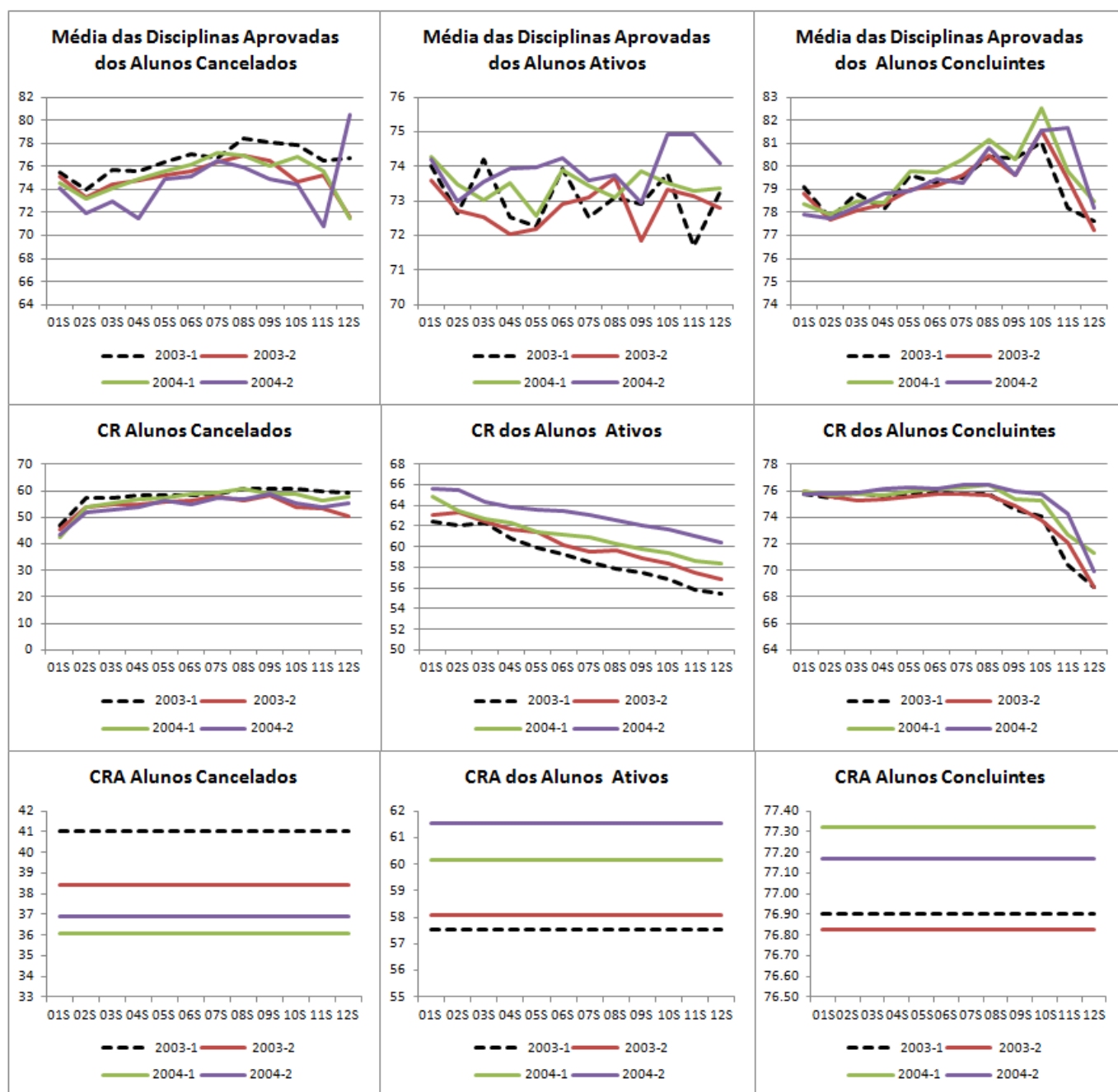


Figura 4.5: Da esquerda para a direita, temos os gráficos que mostram a média das disciplinas aprovadas para os estudantes: (a) cancelados, (b) AFP e (c) concluintes. Os gráficos que apresentam o CR para: (d) cancelados, (e) ativos e (f) concluintes. Os gráficos com o CRA para estudantes: (g) cancelados, (h) ativos e (i) concluintes.

Observando as Figuras 4.4 e 4.5 reconhecemos que não há diferenças acentuadas

para os 4 conjuntos de dados de graduandos que ingressaram nos períodos de 2003-1, 2003-2, 2004-1 e 2004-2. As três classes distintas de alunos (cancelados, ativos fora do prazo (AFP) e concluintes) apresentam as mesmas características detalhadas no Experimento 2 (4.9.3.2.).

4.9.3.4 Discussão dos resultados

Como descrito anteriormente, a base de dados analisada contempla um número significativo de exemplos para cada subconjunto de estudantes (Tabela 4.15).

Os algoritmos analisados demonstraram taxa de acerto geral (acurácia) superior a 78% (Tabela 4.16 e Tabela 4.17). Embora, o classificador *Naive Bayes* não tenha apresentado a melhor acurácia quando comparado aos demais algoritmos, seu rendimento global atende aos objetivos do trabalho que consiste em apresentar uma análise quantitativa dos principais fatores relacionados a evasão, permanência dos estudantes além do prazo médio para conclusão do curso e conclusão. Os modelos gerados pelo algoritmo *Naive Bayes* possibilitaram uma análise numérica e geração de gráficos facilitando a interpretação dos dados, tornando mais facilmente adaptado ao processo de visualização da informação (HAN, KAMBER, 2006, WU *et al.*, 2008).

A análise quantitativa utilizando dados do algoritmo *Naive Bayes* mostrou informações importantes sobre o comportamento ao longo do curso das três classes distintas de estudantes estabelecidas a partir dos dados do SIGA: Cancelados, Ativo fora do prazo (AFP) e Concluintes.

Por fim, a mesma análise quantitativa foi feita para estudantes que ingressaram nos anos de 2003-2, 2004-1 e 2004-2. Pode-se comprovar que as mesmas características encontradas entre os estudantes que ingressaram em 2003-1 também persistiram para os estudantes que ingressaram na UFRJ nos semestres subsequentes.

4.10 Estudo de Caso 03: Curso de Engenharia Civil e suas Ênfases

Predição do Desempenho Acadêmico até o Quinto Semestre Letivo

Neste estudo de caso, temos como objetivo fazer a predição e identificar o desempenho acadêmico dos estudantes no segundo, terceiro, quarto e quinto semestres letivos do curso de graduação a partir do semestre de ingresso no curso. Para cada semestre analisado consideram-se um modelo de dados diferente, portanto diferentes

atributos da base de dados devem ser considerados tendo em vista o semestre letivo que se deseja fazer a predição do desempenho do estudante.

4.10.1 Descrição dos algoritmos Utilizados nos Experimentos

Os algoritmos classificadores utilizados são: *NaiveBayes* (NB), *MultilayerPerceptron* (MP), *Support Vector Machines* usando *polynomial kernel* (SVM1) e *RBF kernels* (SVM2), e *DecisionTable* (DT).

4.10.2 Descrição da Base de Dados Utilizada nos Experimentos

4.10.2.1 Conjunto de Treinamento

O conjunto de treinamento foi composto por estudantes que ingressaram no primeiro semestre letivo dos seguintes anos: 1994-1, 1996-1, 1998-1, 2000-1, 2002-1, 2004-1, 2006-1 e 2008-1. O conjunto de treinamento reuniu 1066 registros de estudantes dos cursos da Engenharia Civil e suas ênfases.

4.10.2.2 Conjunto de Teste

O conjunto de teste é formado por dados de 73 estudantes que ingressaram no ano/semestre de 2003-1 do curso de Engenharia Civil e suas ênfases da Escola Politécnica da UFRJ.

4.10.3 Descrição dos Experimentos e Avaliação dos Resultados

Neste experimento foram consideradas três classificações de progresso nos semestres. A regra de progresso estabeleceu: (i) (APROVADO) indica que o estudante obteve aprovação em pelo menos uma disciplina no semestre letivo; (ii) (PAROU) indica que o estudante parou, ou seja não se matriculou em nenhuma disciplina no semestre letivo; e (iii) (REPROVADO) indica estudantes que não registraram progresso no semestre, nenhuma aprovação nas disciplinas e reprovação (RFM) e/ou (RM) nas disciplinas cursadas no semestre letivo. Para os 73 estudantes do curso de Engenharia Civil analisados no conjunto de teste, identificamos as seguintes distinção no primeiro

semestre letivo conforme a regra de progresso (APROVADOS=60, REPROVADOS = 13 e PAROU=0). Esta informação está armazenada no atributo “01S_SitPeriodo {APROVADO,REPROVADO,PAROU}” que registra para cada estudante a situação de progresso no primeiro semestre letivo.

A base de dados foi particionada utilizando um conjunto de treinamento e conjunto de teste.

4.10.3.1 Predição para o segundo semestre letivo

A Tabela 4.18 mostra o modelo de dados dos estudantes para predição do desempenho acadêmico no segundo semestre letivo. O mesmo modelo de dados dos estudantes é utilizado pelos conjuntos de treinamento e teste. O atributo de classe especificado na linha 24 (@attribute 02S_SitPeriodo {APROVADO,PAROU,REPROVADO}) é utilizado pelos algoritmos classificadores para predizer os valores dos exemplos no conjunto de teste.

Tabela 4.18: Arquivo com os atributos do Modelo de Dados dos Estudantes para predição do desempenho acadêmico no segundo semestre letivo.

```

1 @relation ModeloDadosEstudantes_Previsao_2S
2
3 @attribute 01S_craPeriodo numeric
4 @attribute 01S_SitPeriodo {APROVADO,REPROVADO,PAROU}
5 @attribute 01S_NoDisc numeric
6 @attribute 01S_NoAP numeric
7 @attribute 01S_MediaAP numeric
8 @attribute 01S_NoRFM numeric
9 @attribute 01S_NoRM numeric
10 @attribute 1D_Conceito numeric
11 @attribute 1D_SitDisciplina {AP,RM,RFM}
12 @attribute 2D_Conceito numeric
13 @attribute 2D_SitDisciplina {AP,RM,RFM}
14 @attribute 3D_Conceito numeric
15 @attribute 3D_SitDisciplina {AP,RM,RFM}
16 @attribute 4D_Conceito numeric
17 @attribute 4D_SitDisciplina {AP,RM,RFM}
18 @attribute 5D_Conceito numeric
19 @attribute 5D_SitDisciplina {AP,RM,RFM}
20 @attribute 6D_Conceito numeric
21 @attribute 6D_SitDisciplina {AP,RM,RFM}
22 @attribute 7D_Conceito numeric
23 @attribute 7D_SitDisciplina {AP,RM,RFM}
24 @attribute 02S_SitPeriodo {APROVADO,REPROVADO,PAROU}
25
26 @data

```

A Tabela 4.19 mostra a acurácia e a matriz de confusão obtidas para os algoritmos

empregados neste experimento. Estes resultados mostram que a acurácia dos algoritmos foi acima de 83% índice elevado considerando a avaliação dos algoritmos para três classes diferentes. A matriz de confusão mostra que os algoritmos acertaram na predição da classe (APROVADO) e (PAROU), mas não acertaram nenhum caso da classe (REPROVADO). Analisando o conjunto de teste, dos 13 graduandos da classe REPROVADO no primeiro semestre, no segundo semestre 7 estão na classe PAROU, 4 continuam na classe REPROVADO e 2 passaram para classe APROVADO.

Tabela 4.19: Resultados dos classificadores para a predição do segundo semestre letivo dos estudantes da Engenharia Civil ano ingresso 2003-1.

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	61	83.6 %		61	83.6 %		64	87.7 %		64	87.7 %		64	87.7 %	
ERROU	12	16.4 %		12	16.4 %		9	12.3 %		9	12.3 %		9	12.3 %	
Matriz Confusão															
APROV	54	2	3	55	2	2	57	2	0	57	2	0	58	1	0
PAROU	3	7	0	2	6	2	3	7	0	3	7	0	4	6	0
REPROVADO	0	4	0	0	4	0	0	4	0	0	4	0	0	4	0

4.10.3.2 Predição para o terceiro semestre letivo

A Tabela 4.20 mostra o modelo de dados dos estudantes para predição do desempenho acadêmico no terceiro semestre letivo. O mesmo modelo de dados dos estudantes é utilizado pelos conjuntos de treinamento e teste. O atributo de classe especificado na linha 24 (*@attribute 03S_SitPeriodo {APROVADO,PAROU,REPROVADO}*) é utilizado pelos algoritmos classificadores para predizer os valores dos exemplos no conjunto de teste.

Tabela 4.20: Arquivo com os atributos do Modelo de Dados dos Estudantes para predição do desempenho acadêmico no terceiro semestre letivo.

1	@relation ModeloDadosEstudantes_Previsao_3S
2	
3	@attribute 02S_craPeriodo numeric
4	@attribute 02S_SitPeriodo {APROVADO,PAROU,REPROVADO}
5	@attribute 02S_NoDisc numeric
6	@attribute 02S_NoAP numeric
7	@attribute 02S_MediaAP numeric
8	@attribute 02S_NoRFM numeric
9	@attribute 02S_NoRM numeric
10	@attribute 1D_Conceito numeric
11	@attribute 1D_SitDisciplina {AP,RM,RFM}
12	@attribute 2D_Conceito numeric
13	@attribute 2D_SitDisciplina {AP,RM,RFM}
14	@attribute 3D_Conceito numeric
15	@attribute 3D_SitDisciplina {AP,RM,RFM}

```

16 @attribute 4D_Conceito numeric
17 @attribute 4D_SitDisciplina {AP,RM,RFM}
18 @attribute 5D_Conceito numeric
19 @attribute 5D_SitDisciplina {AP,RM,RFM}
20 @attribute 6D_Conceito numeric
21 @attribute 6D_SitDisciplina {AP,RM,RFM}
22 @attribute 7D_Conceito numeric
23 @attribute 7D_SitDisciplina {AP,RM,RFM}
24 @attribute 03S_SitPeriodo {APROVADO,PAROU,REPROVADO}
25
26 @data

```

A Tabela 4.21 mostra os resultados dos classificadores para a predição das três classes considerando a acurácia dos classificadores e a matriz de confusão. Pelos resultados apresentados na tabela a acurácia está acima de 82% para predição das três classes de estudantes para todos os classificadores utilizados no experimento. A matriz de confusão mostra que o número de acertos para as classes (APROVADO) e (PAROU) são elevados, mas os classificadores não conseguiram identificar exemplos da classe (REPROVADO).

Tabela 4.21: Resultados dos classificadores para a predição do terceiro semestre letivo dos estudantes da Engenharia Civil ano ingresso 2003-1.

	(NB)	(MP)	(SVM1)	(SVM2)	(DT)
ACERTOU	60 82.2 %	60 82.2 %	63 86.3 %	64 87.7 %	64 87.7 %
ERROU	13 17.8 %	13 17.8 %	10 13.7 %	9 12.3 %	9 12.3 %
Matriz Confusão					
APROVADO	49 2 3	49 2 3	53 1 0	53 1 0	53 1 0
PAROU	2 11 0	2 11 0	2 10 1	2 11 0	2 11 0
REPROVADO	2 4 0	2 4 0	4 2 0	4 2 0	4 2 0

4.10.3.3 Predição para o quarto semestre letivo

A Tabela 4.22 mostra o modelo de dados dos estudantes para predição do desempenho acadêmico no quarto semestre letivo. O mesmo modelo de dados dos estudantes é utilizado pelos conjuntos de treinamento e teste. O atributo de classe especificado na linha 24 (*@attribute 04S_SitPeriodo {APROVADO,PAROU,REPROVADO}*) é utilizado pelos algoritmos classificadores para prever os valores dos exemplos no conjunto de teste.

Tabela 4.22: Arquivo com os atributos do modelo de dados dos estudantes para predição do desempenho acadêmico no quarto semestre letivo.

```

1  @relation ModeloDadosEstudantes_Previsao_4S
2
3  @attribute 03S_craPeriodo numeric
4  @attribute 03S_SitPeriodo {APROVADO,PAROU,REPROVADO}
5  @attribute 03S_NoDisc numeric
6  @attribute 03S_NoAP numeric
7  @attribute 03S_MediaAP numeric
8  @attribute 03S_NoRFM numeric
9  @attribute 03S_NoRM numeric
10 @attribute 1D_Conceito numeric
11 @attribute 1D_SitDisciplina {AP,RM,RFM}
12 @attribute 2D_Conceito numeric
13 @attribute 2D_SitDisciplina {AP,RM,RFM}
14 @attribute 3D_Conceito numeric
15 @attribute 3D_SitDisciplina {AP,RM,RFM}
16 @attribute 4D_Conceito numeric
17 @attribute 4D_SitDisciplina {AP,RM,RFM}
18 @attribute 5D_Conceito numeric
19 @attribute 5D_SitDisciplina {AP,RM,RFM}
20 @attribute 6D_Conceito numeric
21 @attribute 6D_SitDisciplina {AP,RM,RFM}
22 @attribute 7D_Conceito numeric
23 @attribute 7D_SitDisciplina {AP,RM,RFM}
24 @attribute 04S_SitPeriodo {APROVADO,PAROU,REPROVADO}
25
26 @data

```

Observamos pela Tabela 4.23 que a acurácia dos classificadores ficou mais precisa acima dos 89% comparando com os experimentos que avaliam o segundo e terceiro semestres letivos dos estudantes que ingressaram em 2003-1. O algoritmo *Naive Bayes* conseguiu prever a classe (REPROVADO).

Tabela 4.23: Resultados dos classificadores para a predição do quarto semestre letivo dos estudantes da Engenharia Civil ano ingresso 2003-1.

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	67	91.8 %		65	89.0 %		67	91.8 %		70	95.9 %		67	91.8 %	
ERROU	6	8.2 %		8	11.0 %		6	8.2 %		3	4.1 %		6	8.2 %	
Matriz Confusão															
APROVADO	50	1	4	50	2	3	53	1	1	55	0	0	53	0	2
PAROU	0	15	1	0	15	1	1	14	1	1	15	0	2	14	0
REPROVADO	0	0	2	2	0	0	1	1	0	1	1	0	2	0	0

4.10.3.4 Predição para o quinto semestre letivo

A Tabela 4.24 mostra o modelo de dados dos estudantes para predição do desempenho acadêmico no quinto semestre letivo. O mesmo modelo de dados dos estudantes é utilizado pelos conjuntos de treinamento e teste. O atributo de classe especificado na linha 24 (*@attribute*

05S_SitPeriodo{APROVADO,PAROU,REPROVADO}) é utilizado pelos algoritmos classificadores para prever os valores dos exemplos no conjunto de teste.

Tabela 4.24: Arquivo com os atributos do Modelo de Dados dos Estudantes para predição do desempenho acadêmico no quinto semestre letivo.

1	@relation ModeloDadosEstudantes_Previsao_5S
2	
3	@attribute 04S_craPeriodo numeric
4	@attribute 04S_SitPeriodo {APROVADO,PAROU,REPROVADO}
5	@attribute 04S_NoDisc numeric
6	@attribute 04S_NoAP numeric
7	@attribute 04S_MediaAP numeric
8	@attribute 04S_NoRFM numeric
9	@attribute 04S_NoRM numeric
10	@attribute 1D_Conceito numeric
11	@attribute 1D_SitDisciplina {AP,RM,RFM}
12	@attribute 2D_Conceito numeric
13	@attribute 2D_SitDisciplina {AP,RM,RFM}
14	@attribute 3D_Conceito numeric
15	@attribute 3D_SitDisciplina {AP,RM,RFM}
16	@attribute 4D_Conceito numeric
17	@attribute 4D_SitDisciplina {AP,RM,RFM}
18	@attribute 5D_Conceito numeric
19	@attribute 5D_SitDisciplina {AP,RM,RFM}
20	@attribute 6D_Conceito numeric
21	@attribute 6D_SitDisciplina {AP,RM,RFM}
22	@attribute 7D_Conceito numeric
23	@attribute 7D_SitDisciplina {AP,RM,RFM}
24	@attribute 05S_SitPeriodo {APROVADO,PAROU,REPROVADO}
25	
26	@data

A Tabela 4.25 mostra os resultados dos classificadores para a predição das três classes considerando a acurácia dos classificadores e a matriz de confusão. Observamos pela tabela que a acurácia dos classificadores ficou acima dos 87%. Isto mostra que quanto mais o estudante avança nos semestres letivos maior é a precisão dos algoritmos para identificar o desempenho dos estudantes. Este conjunto de teste oferece apenas um exemplo de estudante na situação (REPROVADO) no quinto semestre e os algoritmos *Naive Bayes* e *Multilayer Perceptron* conseguiram identificá-lo corretamente.

Tabela 4.25: Resultados dos classificadores para a predição do quinto semestre letivo dos estudantes da Engenharia Civil ano ingresso 2003-1.

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	64	87.7 %		69	94.5 %		71	97.3 %		71	97.3 %		69	94.5 %	
ERROU	9	12.3 %		4	5.5 %		2	2.7 %		2	2.7 %		4	5.5 %	
Matriz Confusão															
APROVADO	47	1	7	54	0	1	55	0	0	55	0	0	53	0	2
PAROU	1	16	0	3	14	0	1	16	0	1	16	0	1	16	0
REPROVADO	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0

4.11 Estudo de Caso 04: Estudantes do Curso de Engenharia Produção e suas Ênfases Predição do Desempenho Acadêmico até o Quinto Semestre Letivo

Neste estudo de caso, temos como objetivo fazer a predição para identificar o desempenho acadêmico dos estudantes no segundo, terceiro, quarto e quinto semestres letivos do curso de graduação em engenharia Produção e suas ênfases. Para cada semestre analisado considera-se um modelo de dados diferente, portanto diferentes atributos da base de dados devem ser considerados.

4.11.1 Descrição dos Algoritmos Utilizados nos Experimentos

Os algoritmos classificadores utilizados são: *NaiveBayes* (NB), *MultilayerPerceptron* (MP), *Support Vector Machines* usando polynomial kernel (SVM1) e *RBF kernels* (SVM2), e *DecisionTable* (DT).

4.11.2 Descrição da Base de Dados

4.11.2.1 Conjunto de Treinamento

O conjunto de treinamento foi composto por estudantes que ingressaram no primeiro semestre letivo dos seguintes anos: 1994-1, 1996-1, 1998-1, 2000-1, 2002-1, 2004-1, 2006-1 e 2008-1. O conjunto de treinamento possui 681 registros de estudantes dos cursos de Engenharia Produção e suas ênfases.

4.11.2.2 Conjunto de Teste

O conjunto de teste é formado por dados de 48 estudantes que ingressaram no ano/semestre de 2005-1 do curso de Engenharia Produção e suas ênfases da Escola Politécnica da UFRJ.

4.11.3 Descrição dos Experimentos e Avaliação dos Resultados

Neste experimento foram consideradas três classificações de progresso nos

semestres. A regra de progresso estabeleceu: (i) (APROVADO) indica que o estudante obteve aprovação em pelo menos uma disciplina no semestre letivo; (ii) (PAROU) indica que o estudante parou, ou seja não se matriculou em nenhuma disciplina no semestre letivo; e (iii) (REPROVADO) indica estudantes que não registraram progresso no semestre, nenhuma aprovação nas disciplinas e reprovação (RFM) e/ou (RM) nas disciplinas cursadas no semestre letivo. Para os 48 estudantes do curso de Engenharia Produção analisados no conjunto de teste, identificamos as seguintes distinção no primeiro semestre letivo conforme a regra de progresso (APROVADOS=46, REPROVADOS = 1 e PAROU=1). Esta informação está armazenada no atributo “*01S_SitPeriodo {APROVADO,REPROVADO,PAROU}*” que registra para cada estudante a situação de progresso no primeiro semestre letivo.

O método utilização para particionar a base de dados foi utilizando um conjunto de treinamento e conjunto de testes.

4.11.3.1 Predição para o segundo semestre letivo

O modelo de dados dos estudantes é o mesmo utilizado para os demais cursos de graduação para predição do segundo semestre letivo (seção 4.10.3.1 Predição para o segundo semestre letivo).

A Tabela 4.26 mostra o resultado dos classificadores: o número de acertos e erros, acurácia (porcentagem de acertos e erros) e matriz confusão. A acurácia obtida pelos algoritmos classificadores é bastante alta acima de 89%. Este fato deve-se a grande maioria dos dados pertencerem a classe (APROVADO).

Tabela 4.26: Resultados dos classificadores para a predição do segundo semestre letivo dos estudantes da Engenharia Produção ano ingresso 2005-1.

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	43	89.6 %		43	89.6 %		47	97.9 %		47	97.9 %		47	97.9 %	
ERROU	5	10.4 %		5	10.4 %		1	2.1 %		1	2.1 %		1	2.1 %	
Matriz Confusão															
APROVADO	42	3	1	41	5	0	46	0	0	46	0	0	46	0	0
PAROU	1	1	0	0	2	0	1	1	0	1	1	0	1	1	0
REPROVADO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4.11.3.2 Predição para o terceiro semestre letivo

O modelo de dados dos estudantes é o mesmo utilizado para os demais cursos de

graduação para predição do terceiro semestre letivo (seção 4.10.3.2 Predição para o terceiro semestre letivo).

A Tabela 4.27 mostra o resultado dos classificadores: o número de acertos e erros, acurácia (porcentagem de acertos e erros) e matriz confusão. A acurácia é alta superior a 87%, aumentou o número de estudantes da classe (PAROU) e (REPROVADO). Nenhum classificador acertou a classe (REPROVADO).

Tabela 4.27: Resultados dos classificadores para a predição do terceiro semestre letivo dos estudantes da Engenharia Produção ano ingresso 2005-1.

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	43	89.6 %		43	89.6 %		42	87.5 %		42	87.5 %		42	87.5 %	
ERROU	5	10.4 %		5	10.4 %		6	12.5 %		6	12.5 %		6	12.5 %	
Matriz Confusão															
APROVADO	40	0	0	40	0	0	40	0	0	40	0	0	40	0	0
PAROU	2	3	1	3	3	0	4	2	0	4	2	0	4	2	0
REPROVADO	1	1	0	2	0	0	2	0	0	2	0	0	2	0	0

4.11.3.3 Predição para o quarto semestre letivo

O modelo de dados dos estudantes é o mesmo utilizado para os demais cursos de graduação para predição do quarto semestre letivo (seção 4.10.3.3 Predição para o quarto semestre letivo).

A Tabela 4.28 mostra o resultado dos classificadores: o número de acertos e erros, acurácia (porcentagem de acertos e erros) e matriz confusão. Os resultados mostram que a acurácia chegou a 100% para a maioria dos algoritmos. Neste conjunto de dados não observou-se exemplos da classe (REPROVADO).

Tabela 4.28: Resultados dos classificadores para a predição do quarto semestre letivo dos estudantes da Engenharia Produção ano ingresso 2005-1.

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	46	95.8 %		48	100.0 %		48	100.0 %		48	100.0 %		48	100.0 %	
ERROU	2	4.2 %		0	0.0 %		0	0.0 %		0	0.0 %		0	0.0 %	
Matriz Confusão															
APROVADO	40	0	0	40	0	0	40	0	0	40	0	0	40	0	0
PAROU	0	6	2	0	8	0	0	8	0	0	8	0	0	8	0
REPROVADO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4.11.3.4 Predição para o quinto semestre letivo

O modelo de dados dos estudantes é o mesmo utilizado para os demais cursos de

graduação para predição do quinto semestre letivo (seção 4.10.3.4 Predição para o quinto semestre letivo).

A tabela mostra o resultado dos classificadores: o número de acertos e erros, acurácia (porcentagem de acertos e erros) e matriz confusão. Acurácia acima de 95% e a taxa de acerto para a classe (PAROU) foi de 100%.

Tabela 4.29: Resultados dos classificadores para a predição do quinto semestre letivo dos estudantes da Engenharia Produção ano ingresso 2005-1.

	(NB)	(MP)	(SVM1)	(SVM2)	(DT)
ACERTOU	46 95.8 %	47 97.9 %	46 95.8 %	46 95.8 %	46 95.8 %
ERROU	2 4.2 %	1 2.1 %	2 4.2 %	2 4.2 %	2 4.2 %
Matriz Confusão					
APROVADO	40 2 0	41 1 0	40 2 0	40 2 0	40 2 0
PAROU	0 6 0	0 6 0	0 6 0	0 6 0	0 6 0
REPROVADO	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0

4.12 Estudo de Caso 05: Estudantes do Curso de Engenharia Mecânica e suas Ênfases Predição do Desempenho Acadêmico até o Quinto Semestre Letivo

Neste estudo de caso, temos como objetivo fazer a predição para identificar o desempenho acadêmico dos estudantes no segundo, terceiro, quarto e quinto semestres letivos do curso de graduação em Engenharia Mecânica e suas ênfases. Para cada semestre analisado considera-se um modelo de dados diferente, portanto diferentes atributos da base de dados devem ser considerados.

4.12.1 Descrição dos Algoritmos Utilizados nos Experimentos

Os algoritmos classificadores utilizados são: *NaiveBayes* (NB), *MultilayerPerceptron* (MP), *Support Vector Machines usando polynomial kernel* (SVM1) e *RBf kernels* (SVM2), e *DecisionTable* (DT).

4.12.2 Descrição da Base de Dados

4.12.2.1 Conjunto de Treinamento

O conjunto de treinamento foi composto por estudantes que ingressaram no primeiro semestre letivo dos seguintes anos: 1994-1, 1996-1, 1998-1, 2000-1, 2002-1, 2004-1, 2006-1 e 2008-1. O conjunto de treinamento possui dados de 846 estudantes dos cursos de Engenharia Mecânica e suas ênfases.

4.12.2.2 Conjunto de Teste

O conjunto de teste é formado por 69 estudantes que ingressaram no ano e semestre de 2007-1 nos cursos de Engenharia Mecânica e suas ênfases.

4.12.3 Descrição dos Experimentos e Avaliação dos Resultados

Neste experimento foram consideradas três classificações de progresso nos semestres: (i) (APROVADO) indica que o estudante obteve aprovação em pelo menos uma disciplina no semestre letivo; (ii) (PAROU) indica que o estudante não cursou nenhuma disciplina no semestre letivo; e (iii) (REPROVADO) indica estudantes que não foram aprovados em nenhuma disciplina e foram reprovados RFM e/ou RM nas disciplinas cursadas no semestre letivo.

A base de dados foi particionada utilizando um conjunto de treinamento e conjunto de teste.

4.12.3.1 Predição para o segundo semestre letivo

O modelo de dados dos estudantes é o mesmo utilizado para os demais cursos de graduação para predição do segundo semestre letivo (seção 4.10.3.1 Predição para o segundo semestre letivo).

A Tabela 4.30 mostra o resultado dos classificadores: o número de acertos e erros, acurácia (porcentagem de acertos e erros) e matriz confusão. A acurácia acima de 84%, a taxa de acerto da classe (PAROU) está em torno de 50%. O algoritmo *Naive Bayes* identificou um exemplo da classe (REPROVADO).

Tabela 4.30: Resultados dos classificadores para a predição do segundo semestre letivo dos estudantes da Engenharia Mecânica ano ingresso 2007-1

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	59	85.5 %		58	84.1 %		62	89.9 %		62	89.9 %		62	89.9 %	
ERROU	10	14.5 %		11	15.9 %		7	10.1 %		7	10.1 %		7	10.1 %	
Matriz Confusão															
APROVADO	54	0	4	56	0	2	58	0	0	58	0	0	58	0	0
PAROU	1	4	3	4	2	2	4	4	0	4	4	0	4	4	0
REPROVADO	2	0	1	3	0	0	3	0	0	3	0	0	3	0	0

4.12.3.2 Predição para o terceiro semestre letivo

O modelo de dados dos estudantes é o mesmo utilizado para os demais cursos de graduação para predição do terceiro semestre letivo (seção 4.10.3.2 Predição para o terceiro semestre letivo).

A Tabela 4.31 mostra o resultado dos classificadores: o número de acertos e erros, acurácia (porcentagem de acertos e erros) e matriz confusão. O menor valor de acurácia entre os classificadores foi de 81%. A taxa de acerto para a classe (PAROU) foi superior a 54% e nenhum dos algoritmos identificou a classe (REPROVADO).

Tabela 4.31: Resultados dos classificadores para a predição do terceiro semestre letivo dos estudantes da Engenharia Mecânica ano ingresso 2007-1.

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	56	81.2 %		58	84.1 %		61	88.4 %		61	88.4 %		61	88.4 %	
ERROU	13	18.8 %		11	15.9 %		8	11.6 %		8	11.6 %		8	11.6 %	
Matriz Confusão															
APROVADO	49	3	4	52	3	1	53	3	0	53	3	0	54	2	0
PAROU	3	7	1	3	6	2	3	8	0	3	8	0	4	7	0
REPROVADO	2	0	0	2	0	0	2	0	0	2	0	0	2	0	0

4.12.3.3 Predição para o quarto semestre letivo

O modelo de dados dos estudantes é o mesmo utilizado para os demais cursos de graduação para predição do quarto semestre letivo (seção 4.10.3.3 Predição para o quarto semestre letivo).

A Tabela 4.32 mostra o resultado dos classificadores: o número de acertos e erros, acurácia (porcentagem de acertos e erros) e matriz confusão. O menor valor de acurácia entre os classificadores foi de 88%, a menor taxa de acerto da classe (PAROU) foi igual a 83%, e os algoritmos *Naive Bayes* e *Multilayer Perceptron* identificaram exemplos da classe (REPROVADO).

Tabela 4.32: Resultados dos classificadores para a predição do quarto semestre letivo dos estudantes da Engenharia Mecânica ano ingresso 2007-1.

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	61	88.4 %		63	91.3 %		66	95.7 %		66	95.7 %		67	97.1 %	
ERROU	8	11.6 %		6	8.7 %		3	4.3 %		3	4.3 %		2	2.9 %	
Matriz Confusão															
APROVADO	49	0	5	50	1	3	54	0	0	54	0	0	54	0	0
PAROU	2	10	0	1	11	0	0	12	0	0	12	0	0	12	0
REPROVADO	1	0	2	1	0	2	2	1	0	2	1	0	2	0	1

4.12.3.4 Predição para o quinto semestre letivo

O modelo de dados dos estudantes é o mesmo utilizado para os demais cursos de graduação para predição do quinto semestre letivo (seção 4.10.3.4 Predição para o quinto semestre letivo).

A Tabela 4.33 mostra o resultado dos classificadores: o número de acertos e erros, acurácia (porcentagem de acertos e erros) e matriz confusão. A menor acurácia foi superior a 87%, a taxa de acerto para a classe (PAROU) foi de 100% para a maioria dos classificadores e apenas o *Naive Bayes* identificou exemplos da classe (REPROVADO).

Tabela 4.33: Resultados dos classificadores para a predição do quinto semestre letivo dos estudantes da Engenharia Mecânica ano ingresso 2007-1.

	(NB)			(MP)			(SVM1)			(SVM2)			(DT)		
ACERTOU	60	87.0 %		60	87.0 %		63	91.3 %		64	92.8 %		64	92.8 %	
ERROU	9	13.0 %		9	13.0 %		6	8.7 %		5	7.2 %		5	7.25 %	
Matriz Confusão															
APROVADO	48	3	5	50	4	2	54	2	0	54	2	0	54	2	0
PAROU	0	10	0	0	10	0	0	9	1	0	10	0	0	10	0
REPROVADO	1	0	2	2	1	0	3	0	0	3	0	0	2	1	0

4.13 Estudo de Caso 06: Avaliação da Arquitetura EDM WAVE

Nesta seção, estamos focados para avaliar as técnicas EDM usados na arquitetura EDM WAVE. A seguir apresentamos experimentos sobre três tradicionais cursos de graduação de engenharia: Civil, Mecânica e de Produção.

O objetivo deste estudo de caso é investigar a aplicação da arquitetura EDM WAVE e sua aplicabilidade em lidar com condições reais de uso, auxiliando gestores acadêmicos, não especialistas em EDM, a identificar estudantes em risco de evasão do sistema de ensino universitário. Para isso, investigasse os melhores algoritmos, associado a modelo de dados dos estudantes e a quantidade de classes que podem ser investigada para fazer a predição do desempenho acadêmico dos estudantes a cada semestre letivo.

Comparando com os estudos de casos anteriores (seção 4.10, 4.11 e 4.12) nos quais três classes de estudantes foram analisadas para fazer a predição para o próximo semestre letivo. Este experimento utiliza duas classes para observar os resultados.

Parte deste trabalho foi publicado em *ACM Symposium on Applied Computing*

(SAC) *Track on Intelligent, Interactive and Innovative Learning environments (ACM SAC2014)*. O título do artigo: *WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM* (MANHÃES *et al.*, 2014b).

4.13.1 Descrição dos Algoritmos Utilizados nos Experimentos

Os algoritmos de mineração empregados neste estudo de caso foram: *NaiveBayes* (NB), *MultilayerPerceptron* (MP), *Support Vector Machines* usando *polynomial kernel* (SVM1) e RBF kernels (SVM2), e *DecisionTable* (DT).

4.13.2 Descrição da Base de Dados

Nós selecionamos os registros, oferecido pelos administradores acadêmicos, sobre o progresso acadêmico dos estudantes individuais em cada semestre, no período de 1994 a 2010 o progresso acadêmico denota conclusão bem sucedida das exigências acadêmicas.

Em nossos experimentos, considerou-se o conjunto de dados dos estudantes do primeiro semestre para obter a predição para o semestre seguinte (segundo semestre letivo do ano). Reunimos os oito conjuntos (1995-1, 1997-1, ..., 2009-1) em um único conjunto de dados para cada curso. Este conjunto com dados de todos os estudantes de um mesmo curso de graduação que ingressaram na universidade em diferentes anos (1995-1, 1997-1, ..., 2009-1) formam a base de dados utilizada neste estudo de caso. Embora a estrutura do conjunto de dados (atributos) é o mesmo, os registros sobre o estudante são diferentes entre os cursos de engenharia.

Tabela 4.34: Quantidade de estudantes distribuídos nas duas classes por ano de ingresso.

Curso	Ano de ingresso	Não-progresso	Progresso	Total
EC	(1995-1, 1997-1, ..., 2009-1)	106 (20%)	416 (80%)	522
EM	(1995-1, 1997-1, ..., 2009-1)	100 (21%)	383 (79%)	483
EP	(1995-1, 1997-1, ..., 2009-1)	35 (10%)	319 (90%)	354

4.13.3 Modelo de Dados dos Estudantes

Em nossos experimentos, foram utilizados os seguintes atributos para compor o modelo de dados dos estudantes:

Tabela 4.35: Modelo de dados dos estudantes de graduação predição para o segundo semestre letivo.

Nº	Atributos	Descrição	Valor	Tipo
1	Id estudante	Identificador do estudante		String
2	Ano Ingresso	Ano e período em que o estudante ingressou na universidade		String
3	Id Curso	Identifica o curso de graduação		String
4	01S_Periodo	Primeiro período letivo identificado por (ano-semester)		String
5	01S_SitPeriodo	Situação da matrícula do estudante no primeiro semestre letivo	{não-progresso, progresso}	String
6	01S_CR_Periodo	Armazena o coeficiente de rendimento no primeiro semestre letivo	{0 to n}	Numérico
7	01S_NoDisc	Armazena o número de disciplinas cursadas no primeiro semestre letivo	{0 to n}	Numérico
8	01S_NoAP	Armazena o número de disciplinas aprovadas no primeiro semestre letivo	{0 to n}	Numérico
9	01S_MediaAP	Armazena a média aritmética obtida nas disciplinas aprovadas no primeiro semestre letivo	{0 to 100}	Numérico
10	01S_NoRFM	Armazena o número de disciplinas reprovadas por falta e/ou média no primeiro semestre letivo	{0 to n}	Numérico
11	01S_NoRM	Armazena o número de disciplinas reprovadas por média no primeiro semestre letivo	{0 to n}	Numérico
12	(1D, 2D, ...,7D)_Disciplina	Identifica as disciplinas do primeiro semestre da grade curricular do curso de graduação	id disciplina	String
13	(1D, 2D, ...,7D)_Conceito	Armazena as notas (valor numérico) obtidas nas disciplinas da grade curricular do primeiro semestre letivo	{0 to 100}	Numérico
14	(1D, 2D, ...,7D)_SistDisciplina	Armazena a situação na disciplina do primeiro semestre letivo: AP (Aprovado), RFM (Reprovado por Falta ou Média), RM (Reprovado por Média)	{AP, RM, RFM}	String
15	02S_SitPeriodo	O atributo de classe é utilizado pelo algoritmo classificador para inferir o valor da classe dos exemplos para o segundo semestre letivo	{não-progresso, progresso} {?,?}	String

Este estudo de caso considerou o atributo (*02S_SitPeriodo*), utilizado para armazenar o valor da situação do estudante no segundo semestre letivo, como atributo

de classe para predição deste valor para novos exemplos. Nos experimentos apresentados neste estudo de caso, foram considerados duas classes de valores para os atributos que identificam a situação do estudante no semestre letivo (*01S_SitPeriodo* e *02S_SitPeriodo*). O valor da primeira classe é atribuído quando o estudante não obteve progresso do semestre (não-progresso-NP). A segunda classe refere-se a um estudante que teve progresso no semestre (progresso-P).

Na arquitetura EDM WAVE, o gestor acadêmico pode definir (configurar) uma regra de progresso segundo seus critérios. Por exemplo, o estudante tem um progresso positivo quando seu desempenho em um determinado semestre está acima de um padrão mínimo estabelecido pelo gestor acadêmico. O gestor acadêmico pode utilizar alguns dos seguintes critérios para compor a regra de progresso no semestre acadêmico: número de disciplinas aprovadas (um ou duas, no mínimo), número de créditos, valor mínimo de CR.

No primeiro semestre letivo um novo estudante (calouro) é automaticamente matriculado em disciplinas iniciais da grade do curso de graduação (por volta dos 6 ou 7 disciplinas). Três atributos foram usados para armazenar os valores para cada uma dessas disciplinas iniciais: Identificação da disciplina, nota da disciplina e situação na disciplina (AP, RFM e RM) como mostra a Tabela 4.35.

4.13.4 Descrição do Experimento e Avaliação dos Resultados

Neste estudo de caso, foi verificado o comportamento dos algoritmos de mineração de dados utilizando o método de validação cruzada com 10 conjuntos para particionar a base de dados utilizada. Foram analisadas duas classes para prever o progresso do estudante no segundo semestre letivo.

Considerou-se a seguinte regra de progresso para atribuir um valor para os atributos (*01S_SitPeriodo*) e o atributo de classe (*02S_SitPeriodo*). O valor (progresso-P) atribuído a todos os graduandos que terminaram o semestre com no mínimo uma disciplina aprovada e o valor (não-progresso-NP) para os estudantes que não obtiveram aprovação em nenhuma disciplina, ou foram reprovados por (RFM) ou (RM), ou pararam o curso.

4.13.4.1 Experimento 1:

A tabela mostra os resultados dos algoritmos para o curso de Engenharia Civil.

Tabela 4.36: Taxas de acerto e erro dos classificadores, VP, FN, VN, FP e MC para o curso de Engenharia Civil.

		NB	MLP	SVM1	SVM2	DT
Exemplos corretamente classificados		467 (89.46%)	462 (88.51%)	478 (91.57%)	476 (91.19%)	479 (91.76%)
Exemplos incorretamente classificados		55 (10.54%)	60 (11.49%)	44 (8.43%)	46 (8.81%)	43 (8.24%)
VP		0.79	0.68	0.65	0.61	0.70
FN		0.21	0.32	0.35	0.39	0.30
VN		0.92	0.94	0.98	0.99	0.97
FP		0.08	0.06	0.02	0.01	0.03
MC	P	383 33	390 26	409 7	411 5	405 11
	NP	22 84	34 72	37 69	41 65	32 74

4.13.4.2 Experimento 2:

A tabela mostra os resultados dos algoritmos para o curso de Engenharia Mecânica.

Tabela 4.37: Taxas de acerto e erro dos classificadores, VP, FN, VN, FP e MC para o curso de Engenharia Mecânica.

		NB	MLP	SVM1	SVM2	DT
Exemplos corretamente classificados		421 (87.16%)	424 (87.78%)	434 (89.86%)	427 (88.41%)	436 (90.27%)
Exemplos incorretamente classificados		62 (12.84%)	59 (12.22%)	49 (10.14%)	56 (11.59%)	47 (9.73%)
VP		0.67	0.65	0.55	0.46	0.55
FN		0.33	0.35	0.45	0.54	0.45
VN		0.92	0.94	0.99	1.00	1.00
FP		0.08	0.06	0.01	0.01	0.01
MC	P	354 29	359 24	379 4	381 2	381 2
	NP	33 67	35 65	45 55	54 46	45 55

4.13.4.3 Experimento 3:

A tabela mostra os resultados dos algoritmos para o curso de engenharia de produção.

Tabela 4.38: Taxas de acerto e erro dos classificadores, VP, FN, VN, FP e MC para o curso de Engenharia de Produção.

		NB	MLP	SVM1	SVM2	DT
Exemplos corretamente classificados		332 (93.79%)	330 (93.22%)	338 (95.48%)	341 (96.33%)	340 (96.05%)
Exemplos incorretamente classificados		22 (6.21%)	24 (6.78%)	16 (4.52%)	13 (3.67%)	14 (3.95%)
VP		0.83	0.63	0.63	0.63	0.69
FN		0.17	0.37	0.37	0.37	0.31
VN		0.95	0.97	0.99	1.00	0.99
FP		0.05	0.03	0.01	0.00	0.01
MC	P	303 16	308 11	316 3	319 0	316 3
	NP	6 29	13 22	13 22	13 22	11 24

4.13.4.4 Discussão dos resultados

Neste estudo de caso, focamos no estudo dos algoritmos tendo em vista duas classes de estudantes. O principal objetivo é identificar os classificadores com uma taxa de acerto da classe não-progresso mais elevada. Pois, o estudante sem aprovação nas disciplinas do próximo semestre ou segundo outro critério especificado na regra de progresso, indica uma alta probabilidade de abandono do sistema de ensino.

A Figura 4.6 mostra os gráficos com as principais variáveis quantitativas que comparam o resultado dos classificadores para cada curso de graduação avaliado neste estudo de caso. O gráfico da (Figura 4.6a) mostra a porcentagem de acerto dos classificadores (acurácia), observamos que a taxa é elevada acima de 87% para todos três cursos avaliados. O curso de Engenharia de Produção mostrou melhor resultado. o algoritmo *Naive Bayes* obteve a melhor taxa de acerto para a classe (não-progresso).

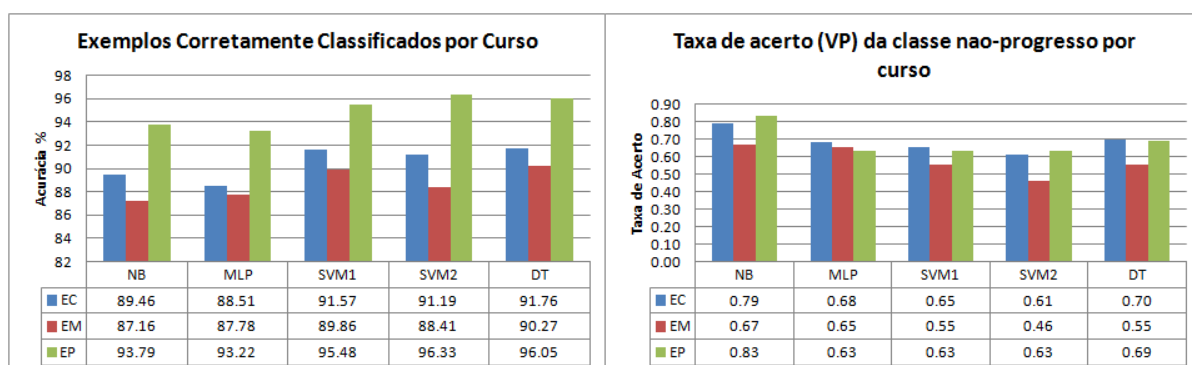


Figura 4.6: Da esquerda para direita temos os gráficos: (a) Exemplos corretamente classificados (acurácia) por curso de graduação. (b) Taxa de acerto da classe não-progresso por curso.

4.14 Estudo de Caso 07: Análise de 6 cursos de graduação da UFRJ

Neste estudo de caso foram investigados seis cursos de graduação da UFRJ: Direito, Farmácia, Física, Engenharia Civil, Engenharia Mecânica e Engenharia de Produção.

Os cursos foram escolhidos porque pertencem a departamentos distintos da universidade, o perfil dos estudantes que ingressam nestes cursos são diferenciados. Os cursos possuem diversos números de entradas de estudantes por ano/semestre. Os cursos possuem taxas de evasão e práticas pedagógicas diferenciadas entre si.

Nos estudos de casos anteriores, verificamos que o modelo de dados dos estudantes propostos e os algoritmos selecionados e testados atendem as demandas da predição do progresso dos estudantes para um grande número de cursos de graduação da UFRJ. No entanto, na avaliação dos diversos algoritmos de mineração de dados, verificamos que o algoritmo *Naive Bayes* mostrou um conjunto de características adequadas para utilizá-lo na arquitetura EDM WAVE.

Portanto, neste estudo de caso, apresentamos uma análise mais aprofundada da aplicação do algoritmo *Naive Bayes*. Este algoritmo classificador apresentou a maior taxa de acerto para prever os estudantes que não terão progresso no próximo semestre letivo, classe positiva (não-progresso). Além disso, o algoritmo *Naive Bayes* apresenta um modelo de predição mais interpretável. Desta forma, o resultado de sua predição pode ser facilmente convertido em gráficos possibilitando análise quantitativa ou outras formas de representar o conhecimento para os usuários da arquitetura EDM WAVE.

Este trabalho foi aceito como artigo para ser publicado e apresentado na conferência intitulada *20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2014)*. A apresentação foi feita na parte da conferência especificada como *Data Mining for Educational Assessment and Feedback workshop (ASSESS 2014)*. O título do artigo: *Evaluating Performance and Dropouts of Undergraduates using Educational Data Mining*. Title of the presentation: *Evaluating Performance and Dropouts of Undergraduates using Educational Data Mining* (MANHÃES et al., 2014c).

4.14.1 Descrição dos Algoritmos Utilizados nos Experimentos

Estabeleceu-se a adoção do algoritmo *Naive Bayes* (NB) para compor a arquitetura EDM WAVE.

4.14.2 Descrição da Base de Dados

Neste estudo de caso a base de dados foi extraída do SIGA da UFRJ. Foram avaliados os estudantes com ano de ingresso entre 1994 a 2010. Os cursos de graduação do tipo STEM estudados foram: Engenharia Civil (EC), Engenharia Mecânica (EM), Engenharia de Produção (EP), Farmácia (FAR), Física (FIS) e Direito (DIR).

4.14.2.1 Conjunto de Treinamento

Neste experimento, para cada curso de graduação, o conjunto de treinamento foi composto por estudantes que ingressaram no primeiro semestre letivo dos seguintes anos: 1994-1, 1996-1, 1998-1, 2000-1, 2002-1, 2004-1, 2006-1 e 2008-1. A Tabela 4.39 mostra para cada curso de graduação o número de estudantes em cada classe analisada.

Tabela 4.39: Número de estudantes no conjunto de treinamento distribuídos em duas classes.

	EC	EM	EP	FAR	FIS	DIR
Não-progresso	81 (17%)	58 (14%)	25 (8%)	63 (10%)	326 (52%)	335 (16%)
Progresso	408 (83%)	358 (86%)	290 (92%)	548 (90%)	297 (48%)	1785 (84%)
Total	489	416	315	611	623	2120

4.14.2.2 Conjunto de Teste

Foram selecionados os dados dos estudantes que ingressaram no primeiro semestre letivo dos seguintes anos: 1995-1, 1997-1, 1999-1, 2001-1, 2003-1, 2005-1, 2007-1 e 2009-1. Considerando os seis cursos analisados, foram construídos 48 conjuntos de testes. A Tabela 4.40 mostra o número de estudantes em cada conjunto de teste, identificam-se o curso de graduação e o ano/semestre de ingresso.

Tabela 4.40: Número de estudantes para os conjuntos de testes para cada curso de graduação e por ano/semestre de ingresso.

Ano/Sem	EC	EM	EP	FAR	FIS	DIR
1995-1	53	60	34	72	84	266
1997-1	68	50	41	70	87	285
1999-1	63	47	38	69	55	285
2001-1	49	45	51	71	34	272
2003-1	70	60	42	78	75	264
2005-1	71	65	41	74	61	273
2007-1	67	66	43	73	74	253
2009-1	60	61	39	96	37	262

A Tabela 4.41 mostra a porcentagem de estudantes distribuídos nas duas classes;

esses valores têm por base os conjuntos de teste. Por exemplo, 19% dos estudantes de Engenharia Civil pertencem à classe não-progresso. Se compararmos Tabela 4.39 e Tabela 4.41, a porcentagem de estudantes por classe é semelhante para os conjuntos de treinamento e teste em todos os 6 cursos de graduação.

Tabela 4.41: Porcentagem de estudantes em cada classe nos conjuntos de teste.

Classe	EC	EM	EP	FAR	FIS	DIR
Não-progresso	0.19	0.19	0.08	0.14	0.51	0.16
Progresso	0.81	0.81	0.92	0.86	0.49	0.84

4.14.3 Definição do Modelo de Dados dos Estudantes

O modelo de dados utilizado neste estudo de caso consiste nos atributos apresentados na Tabela 4.35. O mesmo modelo de dados dos estudantes foi definido para o conjunto de treinamento e teste.

4.14.4 Descrição dos Experimentos e Avaliação dos Resultados

Este estudo de caso avalia o algoritmo *Naive Bayes* usado na arquitetura EDM WAVE aplicado a dados do primeiro semestre letivo para obter a predição do progresso do graduando no segundo semestre letivo.

Esse modelo de dados dos estudantes para o conjunto de treinamento possui o atributo de classe (*02S_SitPeriodo*) com o valor do desempenho dos graduandos no segundo semestre letivo (progresso ou não-progresso).

Neste estudo de caso, foram considerados dois valores para a situação no período letivo. A regra de progresso pode ser configurada de acordo com algum critério estabelecido pelo gestor acadêmico. Neste estudo de caso, considerou-se a seguinte regra de progresso para atribuir um valor para a situação do período: o valor (não-progresso) foi atribuído ao estudante que não obteve nenhum progresso no semestre, nenhuma disciplina aprovada, reprovações por (RFM) ou (RM), ou pararam o curso no meio do semestre letivo. Esses são os estudantes com maior probabilidade de abandonar o curso de graduação. O segundo valor (progresso) foi atribuído quando o estudante obteve progresso no semestre letivo, adotou-se no mínimo uma disciplina aprovada (AP).

As figuras a seguir apresentam gráficos e tabelas que auxiliam na análise quantitativa do desempenho do algoritmo *Naive Bayes* para os 48 conjuntos de testes

utilizados neste estudo de caso. Neste contexto, várias métricas de classificação podem ser usadas para indicar o desempenho do classificador. Na sequência das figuras, temos: Na Figura 4.7, a porcentagem de exemplos corretamente classificados pelo algoritmo (acurácia) *Naive Bayes*. As medidas estatísticas calculadas a partir da matriz de confusão: taxa de acerto da classe positiva (verdadeiro positivo - VP) Figura 4.8 e a taxa de acerto da classe negativa (verdadeiro negativo – VN) (Figura 4.9). A Tabela 4.42 apresenta o valor *Kappa* (Kappa de Cohen) para cada um dos 48 conjuntos de teste. O *Kappa* é uma medida estatística utilizada para mensurar a qualidade do classificador indicando o nível de concordância, avalia o número de respostas concordantes além do que seria esperado ao acaso. Quando a medida ficar próximo do 0 (zero) significa uma maior discordância das informações, quanto mais próximo de 1 (um) indica uma maior ligação e concordância (COHEN, 1960, WITTEN *et al.*, 2011).

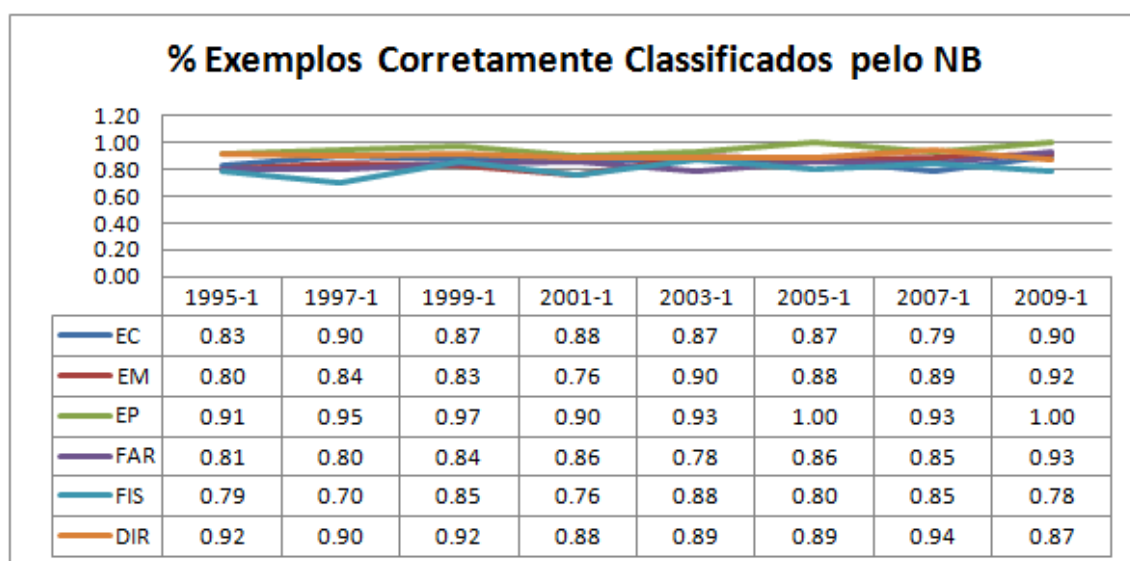


Figura 4.7: Porcentagem de exemplos corretamente classificados pelos algoritmos *Naive Bayes*.

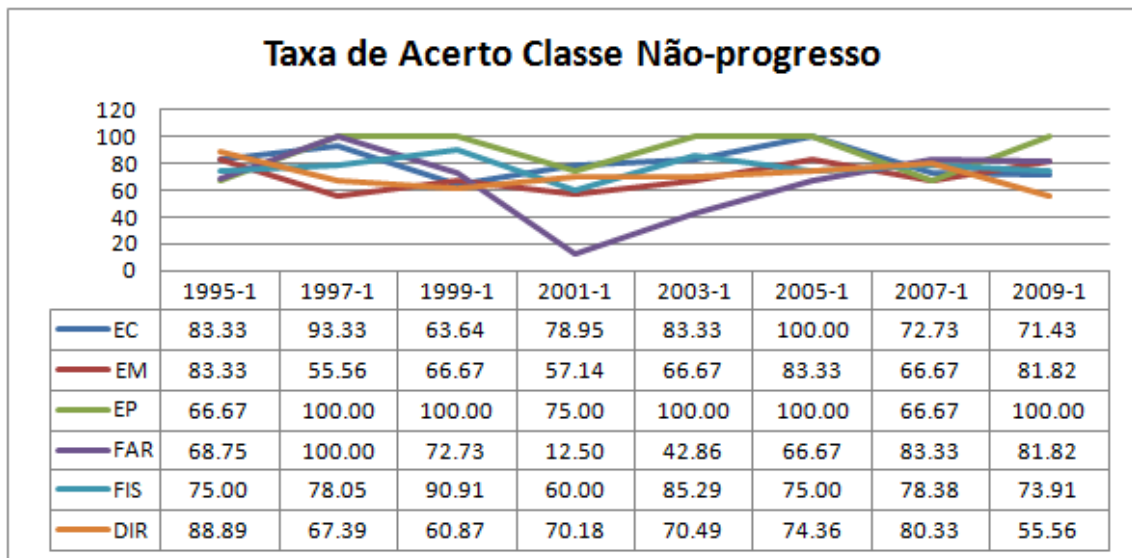


Figura 4.8: Porcentagem de exemplos corretamente classificados (taxa de acerto) pelo algoritmo Naive Bayes para a classe não-progresso.

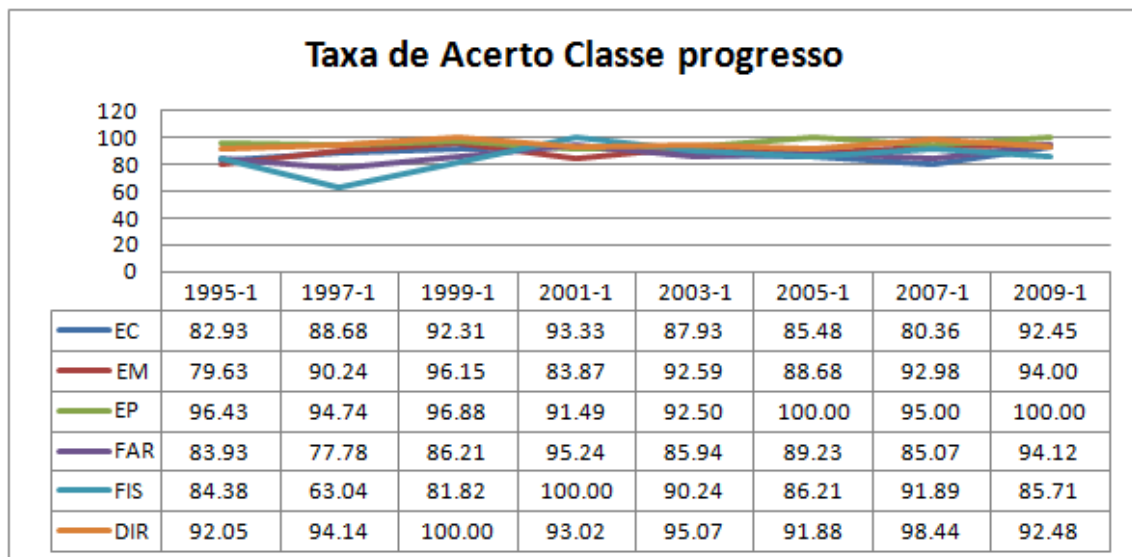


Figura 4.9: Porcentagem de exemplos corretamente classificados (taxa de acerto) pelo algoritmo Naive Bayes para a classe progresso.

Tabela 4.42: Valores do *Kappa*

Ano/Sem	EC	EM	EP	FAR	FIS	DIR
1995-1	0.58	0.36	0.68	0.48	0.57	0.64
1997-1	0.73	0.46	0.72	0.41	0.41	0.62
1999-1	0.56	0.65	0.91	0.50	0.71	0.49
2001-1	0.74	0.42	0.50	0.10	0.55	0.64
2003-1	0.61	0.52	0.54	0.28	0.76	0.69
2005-1	0.60	0.64	1.00	0.47	0.61	0.60
2007-1	0.41	0.57	0.53	0.41	0.70	0.83
2009-1	0.57	0.73	1.00	0.68	0.56	0.47

4.14.4.1 Discussão dos resultados

Nesta seção, avaliamos os resultados experimentais do algoritmo *Naive Bayes* usado na arquitetura. Por sua vez, definiu a classe positiva, não-progresso e classe negativa como progresso. Nos conjuntos de dados, a maioria dos estudantes pertence à classe progresso. No entanto, nosso interesse é medir o desempenho do classificador quando se prevê a classe não-progresso.

A Figura 4.7 apresenta a porcentagem dos exemplos corretamente classificados considerando as duas classes (acurácia) do classificador para cada conjunto de teste dos seis cursos de graduação avaliados neste estudo de caso. O gráfico mostra de forma mais clara que a acurácia do classificador está em torno de 80%.

A Figura 4.8 e Figura 4.9 apresentam o desempenho do algoritmo *Naive Bayes* usando medidas calculadas a partir da matriz de confusão. A Figura 4.8 apresenta as taxas de acerto da classe positiva (não-progresso), o gráfico auxilia a perceber que para a maioria dos conjuntos analisados a taxa de acerto da classe não-progresso está acima de 60%. A Tabela 4.43 apresenta uma análise estatística (média e desvio padrão) dos resultados das taxas de acerto da classe não-progresso para os seis cursos de graduação.

Tabela 4.43: Média e desvio padrão das taxas de acerto para a classe não-progresso.

	EC	EM	EP	FAR	FIS	DIR
Média	80.84	70.15	88.54	66.08	77.07	71.01
Desvio Padrão	11.89	11.35	16.02	27.14	9.04	10.51

Observando a Tabela 4.43 o curso de Farmácia apresenta a maior variação de valores encontrados pelo classificador.

Identificamos que a maior parte dos conjuntos de estudantes analisados neste estudo de caso apresenta taxa de acerto do classificador para a classe negativa (progresso) superior a 80%. Portanto, a arquitetura EDM WAVE baseada no algoritmo *Naive Bayes* e o modelo de dados de estudantes proposto neste estudo de caso apresenta bons resultados para identificar os estudantes com melhor desempenho.

Comparando os resultados para as taxas de acerto da classe não-progresso para a classe progresso obtidos para estes conjuntos de dados, podemos observar que não há grandes diferenças, considerando que o número de exemplos da classe não-progresso é menor do que a classe progresso e há alguns estudantes que abandonam o curso sem uma razão previsível. Neste caso, o erro do classificador pode ser minimizado.

A Tabela 4.42 apresenta valores *Kappa* para cada conjunto de dados. *Kappa* é outra

medida para avaliar o desempenho do classificador. De acordo com a Tabela 4.42, todos os valores *Kappa* estão acima de 0. Indicando que o classificador é adequado.

4.15 Geração dos Modelos para Visualização da Mineração de Dados

A *Visual data mining* é uma área importante que precisa unir as diversas características resultantes dos algoritmos de mineração de dados e representá-las visualmente com significado. A *visual data mining* não é o tópico de foco desta tese. No entanto, elaboramos uma forma de apresentar o resultado dos algoritmos na predição para cada graduando que auxilie o gestor acadêmico visualizar informações para tomar decisões e evitar os índices de evasão universitária.

Em nossa arquitetura utilizamos uma composição de algoritmos classificadores. Tomando como exemplo as duas classes (não-progresso e progresso), pode-se atribuir a cada saída do classificador os seguintes valores: (0) quando a predição é não-progresso e (1) quando a predição for progresso no próximo semestre letivo. A composição consiste em apresentar uma coleção de classificadores com seus respectivos resultados e apresentar uma coluna que assinale qual classe a maioria dos classificadores atribuiu ao graduando. Este recurso de apresentar uma composição de algoritmos e levar em consideração a predição da maioria torna mais confiável e reforçar os resultados individuais de cada classificador.

A Tabela 4.44 mostra um exemplo de relatório que pode ser analisado pelo gestor educacional. O relatório mostra individualmente, para cada estudante de graduação, a predição de cada classificador. As linhas são utilizadas para especificar cada graduando e as colunas a predição de cada classificador. Valor "0" é utilizado quando a predição é (não-progresso) e o valor "1" indica (progresso). A última coluna (Predição do desempenho?) apresenta o resultado da composição de classificadores. Valor "0" é atribuído quando a maioria dos classificadores obteve como resultado da predição a classe (não-progresso) e o valor "1" quando a maioria dos algoritmos obteve como resultado da predição a classe (progresso). No entanto, o gestor educacional tem autonomia para interpretar os resultados.

Tabela 4.44: *Layout* do relatório com a predição de um grupo de n estudantes graduação.

Estudante ID	Algoritmo 1	...	Algoritmo n	Predição do desempenho?
Estudante 1	0	0	1	0
...
Estudante n	1	1	1	1

4.16 Conclusão

Este capítulo teve como objetivo apresentar estudos de casos que pudessem comprovar a funcionalidade da arquitetura EDM WAVE. Através dos estudos de casos vários objetivos foram analisados. No entanto, dois objetivos principais podem ser destacados. Primeiro, a investigação dos modelos de dados mais adequados e que pudessem ser utilizados pelo maior número de cursos de graduação. Segundo, avaliar o maior número de algoritmos classificadores que atendessem de forma satisfatória a predição do desempenho acadêmico dos graduandos a cada semestre letivo. Identificar quais desses algoritmos apresentam modelos de classificação que possam ser mais facilmente convertidos em representações gráficas. Através dessas representações pode-se melhor interpretar os resultados da predição do desempenho acadêmico e, por consequência, destacar as principais características que distinguem o desempenho acadêmico dos estudantes de graduação.

Observou-se que vários estudos mencionaram o período de maior ocorrência das evasões, citam que elas ocorrem no início do curso de graduação (JOHNSTON, 1997, SARAIVA, MASSON, 2003, BARROSO, FALCÃO, 2004, DEKKER *et al.*, 2009). O abandono de curso não é uma decisão imediata do estudante, ocorrem indícios que podem ser identificados, quando um estudante tem progresso acadêmico insatisfatório, isto indica que o estudante está enfrentando dificuldades. O primeiro ano é crítico, pela adaptação do estudante a instituição. Neste estudo, observamos alguns fenômenos que ocorrem com os graduandos da UFRJ (MANHÃES *et al.*, 2012, 2014a), principalmente no primeiro ano letivo, identificamos mudança significativa entre o desempenho acadêmico dos estudantes que concluíram os cursos de graduação com relação àqueles que evadiram ou ficaram ativos fora do prazo de conclusão do curso.

Os modelos de dados encontrados e definidos para fazer a predição de cada semestre letivo mostraram-se bastante adequados. A estrutura dos dados foi a mesma utilizada em vários experimentos descritos ao longo dos estudos de casos apresentados neste capítulo.

Portanto, os conjuntos de atributos usados para prever o desempenho em cada semestre acadêmico é eficiente, conciso e pode ser facilmente extraído do SGA da universidade.

Nesta tese, nós avaliamos um total de 12 algoritmos de mineração de dados públicos e disponíveis na biblioteca da ferramenta de mineração de dados Weka (HALL *et al.*, 2009, BOUCKAERT *et al.*, 2010). Diversos detalhes sobre o desempenho de todos os algoritmos avaliados foram apresentados nos estudos de casos e nos artigos publicados (MANHÃES *et al.*, 2011, 2012, 2014b, 2014c, 2014d, 2015). Entre os algoritmos classificadores avaliados nos estudos de casos, o algoritmo *Naive Bayes* apresentou melhor resultado geral. Este algoritmo apresenta um modelo interpretável e seus resultados numéricos podem ser facilmente convertidos em gráficos, ele obteve uma precisão na classificação (acurácia) em torno de 80% quando aplicado a base de dados de informações acadêmicas dos estudantes. O modelo de classificação apresentado pelo algoritmo *Naive Bayes* foi utilizado para ilustrar uma abordagem quantitativa, na qual três desempenhos acadêmicos distintos foram investigados.

Este capítulo apresentou estudos de casos que investigaram dois ou três diferentes desempenhos, ou classes distintas de exemplos, enquanto todos os outros trabalhos relacionados trataram apenas duas classes de dados.

Também tratamos neste capítulo da apresentação dos resultados da predição para cada estudante para os gestores acadêmicos. Uma composição de algoritmos foi utilizada para atribuir o valor da predição final do desempenho dos estudantes, todos algoritmos que foram utilizados para fazer a predição apresentam seus resultados, portanto, a predição final do desempenho dos estudantes recebe seu valor em função do número de classificadores com o mesmo resultado de predição.

5 Capítulo: Conclusões

Elevadas taxas de abandono dos cursos de graduação têm muitas consequências indesejadas, não só para os estudantes, mas também para a sociedade e para as IFES. Este é um problema complexo, e motiva vários estudos em diversos campos interdisciplinares. A UFRJ possui diversos cursos de graduação e seus graduandos diferem em muitos aspectos, incluindo o tipo de instrução e conhecimento prévio obtidos nos ensinamentos fundamental e médio; fatores socioeconômicos e motivações para obtenção de um diploma universitário.

A proposta desta tese foi apresentar uma arquitetura baseada em EDM para auxiliar gestores acadêmicos a prever o desempenho acadêmico dos estudantes de graduação e identificar aqueles que estão em risco de evadir do sistema de ensino. As maiores dificuldades encontradas para a realização deste trabalho foram obter o acesso aos dados dos estudantes de graduação e na possibilidade de alteração do sistema SIGA.

A primeira dificuldade diz respeito aos dados sobre todos os estudantes de graduação da UFRJ, ela foi contornada a partir da utilização dos dados acadêmicos dos armazenados no Sistema de Gestão Acadêmica da UFRJ. O acesso aos dados foi permitido pelo diretor da Escola Politécnica da UFRJ que solicitou que o diretor da DRE/UFRJ disponibilizasse as bases de dados do SIGA.

A segunda dificuldade foi o entendimento dos dados do SIGA, sem documentação que descrevessem os atributos e seus respectivos valores a etapa de pré-processamento dos dados demandou muito tempo de trabalho.

A arquitetura EDM WAVE não foi completamente implementada, ela é uma proposta que apresenta como vantagem usar apenas os dados dos graduandos armazenados no sistema de gestão acadêmica, sem a necessidade de utilizar dados (sociais e econômicos). Nossa proposta é uma das únicas que utiliza somente variáveis com dados de estudantes que variam com o tempo (*time-varying student data*), ou seja, as previsões executadas pelos algoritmos classificadores são realizadas com dados acadêmicos atualizados semestralmente. Por exemplo, utilizam-se dados do semestre corrente para fazer a previsão para o próximo semestre letivo. Portanto, a nossa abordagem é altamente focada na identificação das desistências iminentes e dar retorno (*feedback*) para os estudantes e gestores educacionais logo após o final do primeiro

semestre letivo ou ao final de cada semestre letivo.

Portanto, a proposta da arquitetura EDM WAVE é de trabalhar acoplada aos sistemas legados de gestão acadêmica. Esta solução envolve menores custos do que o desenvolvimento de novos SGA com a incorporação das funcionalidades da arquitetura EDM WAVE.

Os benefícios diretos da aplicação da arquitetura EDM WAVE baseada em EDM, neste contexto são: (i) identificar ao longo do curso, e principalmente nos períodos iniciais, os estudantes mais propensos a evadir e aqueles com possibilidade de permanecerem matriculados além do prazo médio para conclusão do curso; (ii) permitir que a universidade não utilize apenas dados estatísticos na análise do problema da evasão; (iii) identificar os fatores de sucesso e insucesso específicos para cada curso e relacionar estes fatores ao currículo do curso.

Os resultados obtidos nos estudos de casos avaliados nesta tese mostram que a arquitetura EDM WAVE fornece suporte para os gestores educacionais fazerem o monitoramento do progresso dos graduandos a cada semestre letivo e identificar aqueles que apresentam maiores riscos de abandonar o sistema educacional.

Os estudos de casos analisados nesta tese foram utilizados para avaliar vários algoritmos classificadores. Dentre os algoritmos avaliados o algoritmo classificador *Naïve Bayes* apresentou bons resultados gerais. Este algoritmo apresentou um modelo de classificação mais interpretável e facilmente convertido em gráficos. Os resultados apresentados por este algoritmo permitiram o desenvolvimento de análises quantitativas e representações gráficas com maior valor informativo.

Em termos de publicações esta tese produziu 7 trabalhos:

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., et. al, 2011, *Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados*. Anais do Simpósio Brasileiro de Informática na Educação (XXII SBIE-XVII WIE), Vol. 1. No. 1, 150-159, 2011.

MANHÃES, L. M. B., CRUZ, S.M.S., ZIMBRÃO, G., et. al, 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 (SBSI 2012) - Trilhas Técnicas, pp. 468-479.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2014a. *The Impact of High Dropout Rates in a Large Public Federal Brazilian University: A Quantitative Approach Using Educational Data Mining*. In: CSEDU, 2014, Barcelona, Spain, 6th International Conference on Computer Supported Education, 2014, 124-129.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2014b, *WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM*. In: Symposium of Applied Computing (SAC 2014), Gyeongju, Korea.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2014c, *Evaluating Performance and Dropouts of Undergraduates using Educational Data Mining*. The 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2014), Data Mining for Educational Assessment and Feedback workshop (ASSESS 2014).

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2014d, Investigating Withdraw of STEM Courses in a Brazilian University with EDM. 2nd Symposium on knowledge Discovery, Mining and Learning (KDMiLe 2014).

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2015, *Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree*. Programs. In: Symposium of Applied Computing (SAC'15), April 13–17, 2015, Salamanca, Spain. <http://dx.doi.org/10.1145/2695664.2695918>.

5.1 Trabalhos Futuros

Como trabalhos futuros, consideramos aplicar procedimentos semelhantes para outras IFES, verificando se os resultados até agora observados se repetem para outras instituições de ensino. Além dos cursos de graduação tradicionais, pretendemos investigar a utilização da arquitetura EDM WAVE para cursos de graduação oferecidos na modalidade *on-line* ou EAD.

Investigar mais profundamente a identificação dos três desempenhos dos graduandos nos semestres letivos. Os estudos de casos apresentados nas seções 4.10, 4.11 e 4.12 investigaram três desempenhos durante o semestre: (i) (APROVADO) indica que o estudante obteve aprovação em pelo menos uma disciplina no semestre letivo; (ii) (PAROU) indica que o estudante parou, ou seja não se matriculou em nenhuma disciplina no semestre letivo; e (iii) (REPROVADO) indica estudantes que não registraram progresso no semestre, nenhuma aprovação nas disciplinas e reprovação (RFM) e/ou (RM) nas disciplinas cursadas no semestre letivo. Consideramos que mais estudos devem ser feitos para identificar e prever semestralmente o desempenho de graduandos, a fim de relacionar o desempenho semestral com a situação final no curso: (a) *cancelados* - estudantes que interromperam o curso em algum período antes da formatura (evasão); (b) *ativos fora do prazo* - estudantes que permaneceram matriculados além do prazo médio para conclusão do curso; e (c) *concluintes* - estudantes que concluíram o curso de graduação.

Aprimorar a automatização da fase de pré-processamento dos dados. Esta fase detém procedimentos que precisam da intervenção humana, principalmente porque os dados coletados da base de dados do SGA da UFRJ precisam passar por um processo de limpeza e verificação, por exemplo, dados incompletos (falta de valores para alguns atributos), dados inconsistentes (valores errados como CR=900), registros duplicados entre outros problemas.

Explorar técnicas existentes ou desenvolver novas técnicas de análises de dados permitindo que mais características do problema do desempenho acadêmico possam ser investigadas.

Aprimorar a visualização dos resultados e investir na interatividade do usuário com a arquitetura é uma tarefa a ser perseguida. Certamente os gestores acadêmicos precisam de mais respostas e novas perguntas surgiram sobre o desempenho acadêmico dos graduandos. Os graduandos também precisam ser contemplados com módulos de visualização e alertas sobre seu próprio desempenho acadêmico. A partir deste estudo podem-se criar processos automatizados de construção de roteiros de estudos e orientações de conteúdos personalizados e individualizados.

6 Referências Bibliográficas

ALCALÁ-FDEZ, J., SÁNCHEZ, L., GARCÍA, S., *et al.*, 2009, *KEEL: a software tool to assess evolutionary algorithms for data mining problems*. Soft Computing, 13(3), 307-318.

AMARAL, N. C., 2008, *Evasão e Permanência nas IFES*. Universidade Federal do Amapá (UNIFAP), Macapá, AP. Disponível em: http://www.andifes.org.br/wp-content/files_flutter/13625988891_-_Nelson_Cardoso_Amaral_-_UFG.pdf. Acesso em: 16 ago. 2013.

ANDIFES, 2008, *Evasão e Retenção Discente nas IFES*. Universidade Federal do Amapá (UNIFAP), Macapá, AP. Disponível em: <http://www.andifes.org.br/?cat=583>. Acesso em: 27 maio 2013.

ANDIFES, 2012, *MEC e universidades estudam planos para combater evasão*. Disponível em: <http://www.andifes.org.br/?p=12901.13/02/2012>. Acesso em: 23 ago. 2013.

ANDRIOLA, W., 2009, *Fatores Associados à Evasão Discente na Universidade Federal do Ceará (UFC) de acordo com as Opiniões de Docentes e de Coordenadores de Cursos*. REICE: Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio em Educación, 7(4), 342-355.

ASTIN, A. W., 1994, *Minorities in American higher education: Recent trends, current prospects and recommendations*. San Francisco: Jossey-Bass.

BAKER, R.S.J.d., YACEF, K., 2009, *The State of Educational Data Mining in 2009: A Review and Future Visions*. Journal of Educational Data Mining (JEDM), Volume 1, Issue 1, October 2009, 3-17.

BAKER, R.S.J.d., 2010, *Data Mining for Education*. In McGaw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), vol. 7, 112-118. Oxford, UK: Elsevier.

BAKER, R.S.J.d., ISOTANI, S., CARVALHO, A., 2011, *Mineração de Dados Educacionais: Oportunidades para o Brasil*. Revista Brasileira de Informática na Educação, 19(02), 3-13. Disponível em: <http://dx.doi.org/10.5753/RBIE.2011.19.02.03>. Acesso em: 30 set. 2011.

BARDAGI, M. P., 2007, *Evasão e comportamento vocacional de universitários. Estudos sobre o desenvolvimento de carreira na graduação*. Tese de Doutorado. Porto Alegre, RS: UFRGS – Programa de Pós-Graduação em Psicologia.

BARDAGI, M. P., HUTZ, C.S., 2009, *Não havia outra saída: percepções de estudantes evadidos sobre o abandono do curso superior*. Psico-USF (Impr.), vol.14, n.1, 95-105. ISSN 1413-8271. Disponível em: <http://dx.doi.org/10.1590/S1413-82712009000100010>. Acesso em: 30 set. 2014.

- BARROSO, M. F., FALCÃO, E. B. M., 2004, *Evasão Universitária: O Caso do Instituto de Física da UFRJ*. IX Encontro Nacional de Pesquisa em Ensino de Física, 2004.
- BOUCKAERT, R., EIBE, F., HALL, M., *et al.*, 2010, WEKA Manual for Version 3-6-4. 2010. Disponível em:
<http://ufpr.dl.sourceforge.net/project/weka/documentation/3.6.x/WekaManual-3-6-4.pdf>
 . Acesso em: 15 abr. 2011.
- BRITO, M.I.L., 2013. *Implementação do REUNI na UnB (2008 – 2011): limites na ampliação de vagas e redução da evasão*. Dissertação de Mestrado Profissional em Educação, Universidade de Brasília, Brasília.
- CAMPELLO, A. V. C., LINS, L. N., 2008, *Metodologia de Análise e Tratamento da Evasão e Retenção em Cursos de Graduação de Instituições Federais de Ensino Superior*. XXVIII Encontro Nacional de Engenharia de Produção. Rio de Janeiro, RJ, Brasil.
- CARVALHO, L. A. V., 2005, *Datamining – A mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração*. Rio de Janeiro: Editora Ciência Moderna Ltda.
- CASTRO, F., VELLIDO, A., NEBOT, À., *et al.*, 2007, Applying Data Mining Techniques to e-Learning Problems, *Studies in Computational Intelligence (SCI)*. 62, 183–221, Springer-Verlag Berlin Heidelberg. Disponível em:
<http://sci2s.ugr.es/keel/pdf/specific/capitulo/ApplyingDataMiningTechniques.pdf> .
 Acesso em: 10 abr. 2009.
- CHEEWAPRAKOBKIT, P., 2013, *Study of Factors Analysis Affecting Academic Achievement of Undergraduate Students in International Program*. In: Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS). Vol I, Hong Kong, March 13-15, 2013.
- CHRISPIM, E. M., WERNECK, R. F., 2003, Contexto e prática em Engenharia de Produção: estudo de caso de uma organização como fonte de conhecimento. XXIII ENEGEP. Ouro Preto: ABEPRO.
- COHEN, J. 1960. *A coefficient of agreement for nominal scales*. Educational and Psychological Measurement, 20 (1), 37-46.
- DAVIES, P., 1997, *Within our control?: Improving retention rates in FE*. FEDA. 1997
- DEKKER G., PECHENIZKIY M., VLEESHOUWERS J., 2009, *Predicting Students Drop Out: A Case Study*. In Proceedings of the International Conference on Educational Data Mining, Cordoba, Spain, T. BARNES, M. DESMARAIS, C. ROMERO and S. VENTURA Eds., 41-50, 2009. . Disponível em:
<http://www.educationaldatamining.org/EDM2009/uploads/proceedings/dekker.pdf> .
 Acesso em: 10 abr. 2009.
- DIAS, A.F.M., CERQUEIRA, G.S., LINS, L.N., 2009, *Fatores Determinantes da Retenção Estudantil em um Curso de Graduação em Engenharia de Produção*. In: COBENGE 2009 – XXXVII Congresso Brasileiro de Ensino de Engenharia. Recife –

PE, 2009.

DIAS, E. C. M., THEÓPHILO, C. R., LOPES, M. A. S., 2010, *Evasão no ensino superior: estudo dos fatores causadores da evasão no curso de Ciências Contábeis da Universidade Estadual de Montes Claros – UNIMONTES – MG*. In: Anais do Congresso USP de Iniciação Científica em Contabilidade. São Paulo: Êxito Editora, 2010.

EWB, 2009, Education Group at the World Bank. Disponível em: www.worldbank.org/education/tertiary. Acesso em: 17 abr. 2009.

FAYYAD, U. M., PIATETSKY-SHAPIRO, G., SMYTH, P., UTHURUSAMY, R., 1996, *Advances in knowledge discovery and data mining*. Edition 1st edition, March Publisher MIT Press, 1996.

FORGRAD, 2012, Encontro Regional do Fórum Brasileiro de Pró-reitores de Graduação (FORGRAD Nordeste 2012). Disponível em: <http://www.ufal.edu.br/noticias/2012/12/evasao-e-retencao-nas-universidades-problemas-discutidos-no-forgrad-2012>. Acesso em: 27 maio 2013.

GARCIA, E., ROMERO, C., VENTURA, S., GEA, M., DE CASTRO, C., 2009, *Collaborative Data Mining Tool for Education*. International Working Group on Educational Data Mining.

GOSMAN, E. J., DANDRIDGE, B. A., NETTLES, M. T., THOENY, A. R., 1983, *Predicting student progression: The influence of race and other student and institutional characteristics on college student performance*. Research in Higher Education, 18, 209-236.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, *et al.*, 2009, The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1. 10-18, 2009.

HAMALAINEN, W., LAINE, T.H., SUTINEN, E., 2004, *Data mining in personalizing distance education courses*. In world conference on open learning and distance education, Hong Kong, 1–11. 2004.

HAMALAINEN, W., VINNI, M., 2006, *Comparison of machine learning methods for intelligent tutoring systems*. in Proc. Int. Conf. Intell. Tutoring Syst., Taiwan, 2006, 525-534.

HAN, J., 2005, *Feature selection based on rough set and information entropy*. Granular Computing, 2005 IEEE International Conference on , vol.1, 153-158, 2005.

HAN, J., KAMBER, M., 2006, *Data Mining Concepts and Techniques*. Morgan Kauffmann Publishers, Second Edition, 2006.

HERZOG, S., 2005, *Measuring determinants of student Return vs. Dropout/Stopout vs. Transfer: a First-to-Second Year Analysis of New Freshman*. Research in Higher Education. v.46, n.8, December 2005. p.883-928. DOI: 10.1007/s11162-005-6933-7. Springer Netherlands.

HUANG, S., 2011, *Predictive Modeling and Analysis of Student Academic Performance*

in an Engineering Dynamics Course. Ph.D. Thesis dissertation, Utah State University, Logan, Utah, USA.

INEP, 2009, *Resumo Técnico do Censo da Educação Superior 2009*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: <http://portal.inep.gov.br>. Acesso em: 10 out. 2011.

INEP, 2011, *Resumo Técnico do Censo da Educação Superior 2011*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: <http://portal.inep.gov.br/web/censo-da-educacao-superior/resumos-tecnicos>. Acesso em: 10 out. 2012.

INEP, 2012a, *Resumo Técnico do Censo da Educação Superior 2012*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: <http://portal.inep.gov.br/web/censo-da-educacao-superior/resumos-tecnicos>. Acesso em: 15 dez. 2014.

INEP, 2012b, *Investimentos Públicos em Educação*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: <http://portal.inep.gov.br/estatisticas-gastoseducacao>. Acesso em: 10 out. 2012.

JOHNSTON, V., 1997, *Why do first year students fail to progress to their second year? An academic staff perspective*. Department of Mathematics. Napier University. Paper presented at the British Educational Research Association Annual Conference. September 11-14, 1997: University of York. Disponível em: <http://www.leeds.ac.uk/educol/documents/000000453.htm>. Acesso em: 10 abr. 2010.

KAMPFF, A. J. C., 2009, *Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente*. Tese de doutorado em Informática na Educação. Universidade Federal do Rio Grande do Sul.

KDNUGGETS, 2014, Data Mining Community Top Resource. Disponível em: <http://www.kdnuggets.com/faq/classification-vs-prediction.html>. Acesso em: 16 set. 2014.

KIMBALL, R., CASERTA, J., 2004, *The data warehouse ETL toolkit*. John Wiley & Sons.

KOTSIANTIS, S., PIERRAKEAS, C., PINTELAS, P., 2003, *Preventing student dropout in distance learning using machine learning techniques*. KES, eds. V. Palade, R. Howlett & L. Jain, Springer, 2003. volume 2774 of Lecture Notes in Computer Science, pp. 267–274. 1087-6545 online DOI: 10.1080/08839510490442058.

KOTSIANTIS, S. B., ZAHARAKIS, I. D., AND PINTELAS, P. E., 2007, *Supervised machine learning: A review of classification techniques*. 3-24.

LAVRAC, N., DZEROSKI, S., 1994, *Inductive Logic Programming: Techniques and applications*. Ellis Horwood, New York. 1994.

LIMA JR, P., OSTERMANN, F., REZENDE, F., 2012, *Análise dos condicionantes sociais da evasão e retenção em cursos de graduação em Física à luz da sociologia de Bourdieu*. Revista Brasileira de Pesquisa em Educação em Ciências, Vol. 12, No 1,

2012. Disponível em: <http://revistas.if.usp.br/rbpec/article/view/248>. Acesso em: 27 maio 2013.

LOBO, M. B. D. C. M., 2011, *Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções*. Instituto Lobo/Lobo & Associados Consultoria.

LODER, L.L., NAKAO, O.S., 2011, Evasão e Retenção em Cursos de Engenharia. Sessão Dirigida. XXXIX Congresso Brasileiro de Educação em Engenharia (COBENGE 2011). Blumenau – SC.

LYKOURENTZOU, I. GIANNOUKOS, I., NIKOLOPOULOS, V., *et al.*, 2009, *Dropout prediction in e-learning courses through the combination of machine learning techniques*. Computers & Education, Volume 53, Issue 3, (November, 2009), 950-965.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., *et. al*, 2011, *Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados*. Anais do Simpósio Brasileiro de Informática na Educação (XXII SBIE-XVII WIE), Vol. 1. No. 1, 150-159, 2011.

MANHÃES, L. M. B., CRUZ, S.M.S., ZIMBRÃO, G., *et. al*, 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 (SBSI 2012) - Trilhas Técnicas, pp. 468-479.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2014a, *The Impact of High Dropout Rates in a Large Public Federal Brazilian University: A Quantitative Approach Using Educational Data Mining*. In: CSEDU, 2014, Barcelona, Spain, 6th International Conference on Computer Supported Education, 2014, 124-129.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2014b, *WAVE: an Architecture for Predicting Dropout in Undergraduate Courses using EDM*. In: Symposium of Applied Computing (SAC 2014), Gyeongju, Korea.

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2014c, *Evaluating Performance and Dropouts of Undergraduates using Educational Data Mining*. The 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2014), Data Mining for Educational Assessment and Feedback workshop (ASSESS 2014).

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2014d, *Investigating Withdraw of STEM Courses in a Brazilian University with EDM*. 2nd Symposium on knowledge Discovery, Mining and Learning (KDMiLe 2014).

MANHÃES, L.M.B., CRUZ, S.M.S., ZIMBRÃO, G., 2015, *Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree*. Programs. In: Symposium of Applied Computing (SAC'15), April 13–17, 2015, Salamanca, Spain. <http://dx.doi.org/10.1145/2695664.2695918>.

MÁRQUEZ-VERA, C., CANO, A., ROMERO, C., VENTURA, S., 2013, *Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data*. Applied Intelligence, 1-16.

MEC, 1997, *Diplomação, Retenção e Evasão nos Cursos de Graduação em Instituições de Ensino Superior Públicas*. Ministério da Educação e Cultura. Disponível em: http://www.udesc.br/arquivos/id_submenu/102/diplomacao.pdf. Acesso em: 27 maio 2013.

MEC, 2007, *Diretrizes Gerais do Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais – REUNI*. Ministério da Educação e Cultura. Disponível em: <http://portal.mec.gov.br/sesu/arquivos/pdf/diretrizesreuni.pdf>. Acesso em: 17 fev. 2011.

MEC, 2014. *Passa de 1,1 milhão o número de candidatos inscritos no Sisu até as 19 horas do último dia*. Disponível em: http://portal.mec.gov.br/index.php?option=com_content&view=article&id=20504:passa-de-11-milhao-o-numero-de-candidatos-inscritos-no-sisu-ate-as-19-horas-do-ultimo-dia&catid=410&Itemid=86. Acesso em: 20 nov. 2014.

MELLO, S.P. T., SANTOS, E. G., 2012, *Diagnóstico e alternativas de contenção da evasão no curso de administração em uma universidade pública no sul do Brasil*. Revista Gestão Universitária na América Latina - GUAL, v. 5, p. 67-80, 2012.

MENDES, E. F., VIEIRA, M. T. P., 2009, *Kira: Uma Ferramenta Instrucional para Apoiar a Aplicação do Processo de Mineração de Dados*. XX Simpósio Brasileiro de Informática na Educação.

MINAEI-BIDGOLI, B., PUNCH, W.F., 2003, *Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System*. In: Cantu, P.E., et al. (eds.): Genetic and Evolutionary Computation Conference (GECCO 2003). 2252–2263.

MINAEI-BIDGOLI, B., KORTEMEYER, G., PUNCH, W., 2004a, *Association analysis for an online education system*. Information Reuse and Integration (IRI 2004). Proceedings of the 2004 IEEE International Conference. (November, 2004), 504,509, 8-10.

MINAEI-BIDGOLI, B., TAN, P. N., PUNCH, W. F., 2004b, *Mining interesting contrast rules for a web-based educational system*. In Machine Learning and Applications (ICMLA, 2004). Proceedings. 2004 International Conference on (pp. 320-327). IEEE.

MINAEI-BIDGOLI, B., TAN, P., KORTEMEYER G., PUNCH, W.F., 2006, *Association analysis for a web-based educational system*. Data Mining in E-Learning. WitPress. Southampton, Boston, 2006.

MOORE, R., 1995, *Retention rates research project - final report*, Sheffield Hallam University.

NUGENT, C., CUNNINGHAM, P., 2004, *A Case-Based Explanation System for 'Black-Box' Systems*. Trinity College Dublin, Department of Computer Science TCD-CS-2004-20, pp10, Dublin, Ireland. Disponível em: <https://www.cs.tcd.ie/publications/tech-reports/reports.04/TCD-CS-2004-20.pdf>. Acesso em: 10 abr. 2010.

OECD. 2012. *Brazil, in Education at a Glance 2012: OECD Indicators*. OECD Publishing. Disponível em: <http://dx.doi.org/10.1787/eag-2012-42-en>. Acesso em: 10

abr. 2013.

OLSON, D. L., DELEN, D., 2008, *Advanced data mining techniques*. Springer.

PAIVA, R., BITTENCOURT, I.I., SILVA, A.P., ISOTANI, S., JAQUES, P., 2014. *A Systematic Approach for Providing Personalized Pedagogical Recommendations Based on Educational Data Mining*. In: International Conference on Intelligent Tutoring Systems, 2014, Honolulu. Lecture Notes in Computer Science, 2014. p. 362-367.

PAL, S., 2012, *Mining educational data to reduce dropout rates of engineering students*. International Journal of Information Engineering and Electronic Business (IJIEEB), 4(2), 1.

PANG-NING, T., STEINBACH, M., KUMAR, V., 2005, *Introduction to Data Mining*. Addison-Wesley. 2005.

PASCARELLA, E. T., TERENZINI, P. T., 1991, *How college affects students*. San Francisco: Jossey-Bass.

PSLC, 2010. Pittsburgh Science of Learning Center –PSLC. Disponível em: <http://www.learnlab.org/technologies/datashop/> . Acesso em: 10 abr. 2010.

RAMALHO FILHO, R., 2008, *Redução da evasão discente: diagnóstico e metas propostos pelas universidades federais ao Programa REUNI (análise preliminar)*. Universidade Federal do Amapá (UNIFAP), Macapá/AP. Disponível em: http://www.andifes.org.br/wp-content/files_flutter/13625996794_-_Rodrigo_Ramalho_-_SESu-DEDES-MEC.pdf . Acesso em: 26 ago. 2013.

ROMERO, C., VENTURA, S., 2010, *Educational Data Mining: A Review of the State of the Art, Systems, Man, and Cybernetics, Part C: Applications and Reviews*. IEEE Transactions on, vol.40, no.6, 601-618, 2010. doi: 10.1109/TSMCC.2010.2053532.

ROMERO, C., VENTURA, S., 2013, *Data Mining in Education*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, In Press. Volume 3, Issue 1, 12-27, 2013.

RUSLI, N.M., IBRAHIM, Z., JANOR, R. M., 2008, *Predicting students' academic achievement: Comparison between logistic regression, artificial neural network, and Neuro-fuzzy*. Information Technology, 2008. ITSIM 2008. International Symposium on , vol.1, no., pp.1,6, 26-28 Aug. 2008 doi: 10.1109/ITSIM.2008.4631535.

SAE, 2013a. *Qualificação da mão de obra brasileira, uma nova urgência (Brasil Econômico, em 27.02.2012)*. Secretaria de Assuntos Estratégicos da Presidência da República. Disponível em: <http://www.sae.gov.br/site/?p=14995>. Acesso em: 27 maio 2013.

SAE, 2013b, *Brasil vai simplificar visto de trabalho para estrangeiros (Folha de São Paulo, em 17.05.2013)*. Secretaria de Assuntos Estratégicos da Presidência da República. Disponível em: <http://www.sae.gov.br/site/?p=16481>. Acesso em: 27 maio 2013.

SAE, 2013c, *Brasil precisa de 6 milhões de profissionais estrangeiros, diz SAE (BBC*

Brasil, em 22.04.2013). Disponível em: <http://www.sae.gov.br/site/?p=15768>. Acesso em: 27 maio 2013.

SARAIVA, S. , MASSON. M., 2003, *Evasão e Permanência em uma Instituição de Tradição: um estudo sobre o processo de evasão de estudantes em cursos de Engenharia na Escola Politécnica da UFRJ*. Relatório de Pesquisa, 2003.

SCUSE, D., REUTEMANN, P., 2008, *WEKA Experimenter Tutorial for Version 3-5-8*.

SHMUELI, G., PATEL, N. R., BRUCE, P. C., 2007, *Data mining in excel: Lecture notes and cases*.

SILVA FILHO, R.L.L., MOTEJUNAS, P.R., HIPÓLITO, O., LOBO, M. B. C. M., 2007, *A Evasão no Ensino Superior Brasileiro*. Cadernos de Pesquisa. v. 37, n. 132, p. 641-659, set./dez. São Paulo: Fundação Carlos Chagas, 2007. Disponível em: <http://dx.doi.org/10.1590/S0100-15742007000300007>. Acesso em: 10 abr. 2010.

SILVA, M.V. A., FERREIRA, M. O., 2011, *Estudo sobre evasão e retenção no curso de graduação em ciências econômicas do CAA/UFPE. Considerando uma abordagem microeconômica*. XIX Conic, III Coniti e VII Joic. Universidade Federal de Pernambuco. Recife, PE.

SILVEIRA, I., 2010, Editorial. Revista Brasileira de Informática na Educação, 18(3), 01. 2010. Disponível em: <http://www.br-ie.org/pub/index.php/rbie/article/view/1281/1129>. Acesso em: 10 abr. 2010.

SOARES, I. S., 2000, *UFRJ - A Engenharia de Produção - Opção no Vestibular, Evasão, Reprovação e Novo Vestibular*. In: VI Encontro de Ensino de Engenharia- Anais, UFRJ/UFJF, Itaipava, RJ.

SOARES, I. S. 2006. *Evasão, retenção e orientação acadêmica: UFRJ - Engenharia de Produção – Estudo de Caso*. In: Anais do XXXIV COBENGE - Congresso Brasileiro de Ensino de Engenharia. Ed. Universidade de Passo Fundo, Passo Fundo, RS.

SOARES, I. S., 2009. *UFRJ - Escola Politécnica - Vestibular 1993-2009 - Revisão Histórica - Vagas, Evasão e Retenção*. In: COBENGE 2009 - XXXVII Congresso Brasileiro de Educação em Engenharia. Recife, PE.

SOUZA, S.L., 2008, *Evasão no ensino superior: um estudo utilizando a mineração de dados como ferramenta de gestão do conhecimento em um banco de dados referente à graduação de engenharia*. Dissertação de Mestrado, COPPE/UFRJ, Engenharia Civil, Rio de Janeiro, RJ, Brasil.

SOUZA, C.T., PETRÓ, C.S., GESSINGER, R.M., 2012, II CLABES - Conferencia Latino Americana sobre el Abandono em la Educación Superior. PUCRS - Porto Alegre, RS.

SUMATHI, S., SIVANANDAM, S.N., 2006, *Introduction to Data Mining and its Applications*. Springer-Verlag, Berlin Heidelberg, 2006.

SUPERBY, J.F., VANDAMME, J-P., MESKENS, N., 2006. *Determination of factors influencing the achievement of the first-year university students using data mining*

methods. In Proceedings International Conference Intelligent Tutoring System of the Workshop on Educational Data Mining, Taiwan, 2006, pp. 1-8.

TERENZINI, P. T., PASCARELLA, E. T., 1977, *Voluntary freshman attrition and patterns of social and academic integration in a university: A test of a conceptual model*. Research in Higher Education, 6, 25-43.

THEARLING, K., 2010, *An Introduction to Data Mining*. Disponível em: http://www.thearling.com/dmintro/dmintro_2.htm. Acesso em: 02 dez. 2010.

TIGRINHO, L. M. V., 2008, *Evasão Escolar nas Instituições de Ensino Superior*. Disponível em: <http://www2.cartaconsulta.com.br/espacodocoordenador/?p=93> . Acesso em: 27 maio 2013.

TINTO, V., 1975. *Dropout from higher education: a theoretical synthesis of recent research*. Review of Educational Research, New York, n. 45, p. 89-125, 1975.

TINTO, V., 1993, *Leaving college: Rethinking the causes and cures of student attrition*. Chicago: University of Chicago Press.

TINTO, V., 2006, *Research and practice of student retention: what next?* Journal of College Student Retention: Research, Theory and Practice, 8(1), 1-19.

TONTINI, G., WALTER, S.A., SILVANA, A., 2011, XI Colóquio Internacional sobre Gestão Universitária na América do Sul. Florianópolis, SC.

TPE, 2008, Glossário. TODOS PELA EDUCAÇÃO. Disponível em: <http://www.todospelaeducacao.org.br/busca/?busca=glossario&buscar=>. Acesso em: 29 set. 2013.

UFPE, 2008, Workshop Evasão e Retenção Discente nas IFES. Universidade Federal de Pernambuco. Macapá. Disponível em: http://www.ufpe.br/agencia/index.php?option=com_content&view=article&id=33274:a&catid=19&Itemid=72. Acesso em: 16 ago. 2013.

UFRJ, 2009. Workshop Pensando na graduação, Mesa Redonda com o tema: *Reprovações e abandono. O que pode ser feito?* Rio de Janeiro: UFRJ, 2009.

UFRJ, 2014. Universidade Federal do Rio de Janeiro. Disponível em: <http://www.ufrj.br/>. Acesso em: 20 nov. 2014.

VAPNIK, V. N., 1995, *The nature of Statistical learning theory*. Springer-Verlag, New York, 1995.

VASCONCELOS, A.L.F.S., SILVA, M.N., 2011, XI Colóquio Internacional sobre Gestão Universitária na América do Sul. Florianópolis, SC.

VELOSO, T. C. M. A., ALMEIDA, E. P., 2001, *Evasão nos Cursos de Graduação da Universidade Federal de Mato Grosso, campus universitário de Cuiabá: Um Processo de Exclusão*.

WITTEN, I.H., FRANK, E., 2005, *Data Mining: Practical machine learning tools and*

techniques. 2nd edition Morgan Kaufmann, San Francisco.

WITTEN, I. H., EIBE, F., HALL, A. M., 2011, Data mining: practical machine learning tools and techniques. 3rd The Morgan Kaufmann series in data management systems ed., 2011. ISBN 978-0-12-374856-0.

WU, X., KUMAR, V., ROSS, Q.J., GHOSH, J., et. al, 2008, *Top 10 algorithms in data mining*. Journal of Knowledge and Information Systems. Springer London. vol. 14, Issue 1, 1-37.

ZAFRA, A., ROMERO, C., VENTURA, S., 2011, Multiple instance learning for classifying students in learning management systems, 2011.

7 Apêndice

Os estudantes da Escola politécnica foram tratados separadamente por cursos com suas ênfases. A Tabela 7.1 lista os cursos de graduação oferecidos pela Escola Politécnica da UFRJ de 1994 a 2010.

Tabela 7.1: Cursos de graduação em Engenharia da Escola Politécnica UFRJ.

Curso de Graduação em Engenharia
Engenharia (Básico)
Engenharia (Ciclo Básico)
Engenharia Civil
Eng Civil - Ênfase em Construção Civil
Eng Civil - Ênfase em Estruturas
Eng Civil - Ênfase em Geotecnia
Eng Civil - Ênfase em Mecânica dos Solos
Eng Civil - Ênfase em Obras Hidraul e San
Eng Civil - Ênfase em Recursos Hídricos
Eng Civil - Ênfase em Transportes
Eng Civil - Ênfase em Engenharia dos Transportes
Eng Civil - Ênf em Construção Civil
Eng Civil - Ênf em Estruturas
Eng Civil - Ênf em Geotecnia
Eng Civil - Ênf em Mecânica dos Solos
Eng Civil - Ênf em Recursos Hídricos
Eng Civil - Ênf em Transportes
Eng de Produção - Área: Engenharia Econômica
Eng de Produção - Área: Gerência de Produção
Engenharia de Produção
Engenharia Ambiental
Engenharia de Computação e Informação
Engenharia Eletrônica e de Computação
Engenharia de Controle e Automação
Engenharia de Petróleo
Engenharia de Materiais
Engenharia Elétrica
Engenharia Elétrica: Ênfase em Eletrônica
Engenharia Elétrica: Ênfase em Eletrotécnica
Engenharia Mecânica: Ênf. em Mecânica
Engenharia Mecânica
Engenharia Metalúrgica
Engenharia Naval
Engenharia Naval e Oceânica
Engenharia Nuclear