



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Utilizando Mineração de Dados para Análise de gênero nos cursos de Computação na UnB

Gustavo Carlos Couto
Marília Alves da Nóbrega Alberto Dantas

Monografia apresentada como requisito parcial
para conclusão do Curso de Computação — Licenciatura

Orientador
Prof. Dr. Jan Mendonça Correa

Brasília
2014

Universidade de Brasília — UnB
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Curso de Computação — Licenciatura

Coordenador: Prof. Dr. Wilson Henrique Veneziano

Banca examinadora composta por:

Prof. Dr. Jan Mendonça Correa (Orientador) — CIC/UnB
Prof.^a Dr.^a Maristela Terto de Holanda — CIC/UnB
Prof.^a Dr.^a Maria Emilia Machado Telles Walter — CIC/UnB

CIP — Catalogação Internacional na Publicação

Couto, Gustavo Carlos.

Utilizando Mineração de Dados para Análise de gênero nos cursos de Computação na UnB / Gustavo Carlos Couto, Marília Alves da Nóbrega Alberto Dantas. Brasília : UnB, 2014.

80 p. : il. ; 29,5 cm.

Monografia (Graduação) — Universidade de Brasília, Brasília, 2014.

1. mineração, 2. dados, 3. meninas, 4. computação

CDU 004.4

Endereço: Universidade de Brasília
Campus Universitário Darcy Ribeiro — Asa Norte
CEP 70910-900
Brasília-DF — Brasil



Universidade de Brasília

Instituto de Ciências Exatas
Departamento de Ciência da Computação

Utilizando Mineração de Dados para Análise de gênero nos cursos de Computação na UnB

Gustavo Carlos Couto
Marília Alves da Nóbrega Alberto Dantas

Monografia apresentada como requisito parcial
para conclusão do Curso de Computação — Licenciatura

Prof. Dr. Jan Mendonça Correa (Orientador)
CIC/UnB

Prof.^a Dr.^a Maristela Terto de Holanda Prof.^a Dr.^a Maria Emília Machado Telles Walter
CIC/UnB CIC/UnB

Prof. Dr. Wilson Henrique Veneziano
Coordenador do Curso de Computação — Licenciatura

Brasília, 14 de Julho de 2014

Dedicatória

Gustavo: Dedico este trabalho aos meus pais, pelo apoio, atenção e carinho nas horas mais difíceis. Às minhas irmãs pelo companheirismo e descontração. À Thaís por estar sempre ao meu lado. Aos professores e colegas que compartilharam o seu conhecimento e possibilitaram mais esta conquista.

Marília: Dedico este trabalho primeiramente a Deus, que me deu fé para continuar nessa jornada, mesmo nos momentos mais difíceis. Aos meus pais e meu irmão, que além do suporte emocional, me ajudaram com ideias e opiniões, não deixando que fugisse do escopo do trabalho. Aos meus avós, tios e primos, que me encorajaram e acreditaram em mim em todas as horas. Ao meu amigo e sócio André, por entender a falta de tempo e pouca participação na empresa nessa última fase de escrita do projeto, e ainda me encorajar e acreditar na minha capacidade. A todos os meus amigos e amigas, que mesmo longe me ajudam, seja com carinho ou conversas honestas. E por fim a todos os professores do curso, que compartilharam todo conhecimento possível, desde o primeiro semestre de estudos.

Agradecimentos

Gostaríamos de agradecer primeiramente ao professor Jan por sua orientação, suporte e guia nesse desafio, mesmo quando não tínhamos nenhum objetivo definido. Agradecemos também à professora Maria Emilia que nos introduziu ao projeto "Meninas da Computação" e nos ajudou com todas as suas ideias e opiniões, nos fazendo manter o foco. À professora Maristela por ter nos recebido no projeto "Meninas da Computação" e nos dado a oportunidade de participar da coleta de dados utilizada neste trabalho, além de ter nos guiado na utilização das informações recebidas. À Maria Inez, do DGP, que nos cedeu todos os dados necessários e todo o suporte que precisávamos, além de ter contribuído com o ponto de vista estatístico. À professora Aletéia, que é a coordenadora do projeto "Meninas na Computação", nos ajudando com informações sobre ele e dando todo o suporte necessário. À professora Maria de Fátima, ao professor Wilson e ao professor Luciano, por ajudarem com sugestões de pesquisa, quando estávamos no processo inicial do projeto.

Resumo

A Mineração de Dados é uma área relativamente nova da Ciência da Computação, que visa o processo de descoberta de modelos, resumos e valores derivados de uma determinada coleção de dados. Este trabalho tem como objetivo extrair padrões encontrados nos dados dos alunos dos cursos de Ciência da Computação, Licenciatura em Computação e Engenharia da Computação da Universidade de Brasília desde sua criação, e de algumas alunas do Ensino Fundamental e Médio do Distrito Federal, de forma a identificar tendências e criar hipóteses a respeito da diferença de gênero nesses cursos, utilizando análises estatísticas e Mineração de Dados. Nesse projeto serão descritas algumas das técnicas de Mineração de Dados, principalmente as aplicadas nesse projeto, assim como os aplicativos utilizados. E por fim, serão descritos os resultados encontrados a partir das análises realizadas e sua importância para a UnB e seus gestores.

Palavras-chave: mineração, dados, meninas, computação

Abstract

Data Mining is a relatively new area of the Computer Science, and aims the discovery process of models, extracts and values derived from a certain collection of data. This research objective is to extract patterns found in the data of students from Computer Science, Teaching degree in Computer and Computer Engineering at the University of Brasilia, and from some Elementary and High School female students from the Federal District, in order to identify tendencies and create hypotheses regarding the genre difference in those courses, using statistical analysis and Data Mining. In this project some techniques of Data Mining are described, especially those applied in the project, as well as applications that were used. And finally, will describe the results found based on the analysis that were performed and its importance to the UnB and its managers.

Keywords: mining, data, girls, computer science

Sumário

1	Introdução	1
1.1	Motivação	1
1.2	Objetivos Gerais	2
1.3	Objetivos Específicos	2
1.4	Estrutura do trabalho	2
2	Revisão Teórica	4
2.1	Mulheres e Computação	4
2.2	Sistema de Banco de Dados	5
2.3	Mineração de Dados	5
2.3.1	Classificação	11
2.3.2	Rregressão	12
2.3.3	Detecção de Anomalias	13
2.3.4	Clusterização	14
2.3.5	Regras de Associação	15
2.3.6	Padrões Sequenciais	15
2.4	Weka, ferramenta de aprendizagem de máquina	16
3	Estudo de Caso 1: Dados dos formulários do Projeto Meninas na Computação	19
3.1	Coleta de Dados	19
3.2	Formatação e Limpeza dos Dados	21
3.3	Método de Análise: Usando os dados na plataforma Weka	23
3.3.1	Convertendo para ARFF	23
3.3.2	Carregando os dados e fazendo a análise inicial	26
3.3.3	Utilizando o algoritmo de classificação e analisando a saída	26
3.4	Resultados da Mineração de Dados	32
4	Estudo de Caso 2: Cursos de Grduação de Computação da UnB	34
4.1	Introdução e Análise Estatística	34
4.2	Formatação dos Dados	39
4.3	Análise e Resultados	41
5	Conclusão	52
5.1	Trabalhos Futuros	53
A	Querys utilizadas no Estudo de Caso 1	55

B Saída da árvore de decisões com dois perfis do Estudo de Caso 1	62
C <i>Querys</i> utilizadas no Estudo de Caso 2	67
Referências	79

Listas de Figuras

1.1	Logomarca do projeto Meninas da Computação [10]	2
2.1	A evolução da tecnologia de sistemas de banco de dados [8].	7
2.2	As raízes de Mineração de Dados [7].	8
2.3	Mineração de dados como um passo no processo de descoberta [8]	8
2.4	Geração e utilização de Armazém de Dados [8].	10
2.5	Estágios da construção de um modelo de classificação (concessionária) [7]. .	12
2.6	Anomalia nos dados [7].	13
2.7	Exemplo de uma Clusterização bem sucedida. [7].	15
2.8	Esquema cerveja-fralda. [7].	15
2.9	Tela inicial do Weka.	18
3.1	Questionário aplicado as meninas durante as Semanas Nacionais de Ciência e Tecnologia.	20
3.2	Formato dos dados recebidos pelo CESPE.	21
3.3	Mapeamento de colunas da planilha com atributos da tabela.	22
3.4	Tela inicial do <i>Explorer</i>	24
3.5	Formatos: 1. Banco de Dados, 2. CSV e 3. ARFF.	25
3.6	Tela inicial com os dados carregados.	26
3.7	Tela para análise visual dos dados. Vermelho: quem não sabe banco de dados; Azul: quem sabe banco de dados	27
3.8	Lista de classificadores.	27
3.9	Parte 1 da saída da árvore de decisões.	29
3.10	Parte 2 da saída da árvore de decisões.	29
3.11	Árvore de decisões.	30
3.12	Parte 1 da saída da árvore de decisões com os dois perfis.	31
4.1	Ingressantes por ano e sexo no curso de Ciência da Computação	36
4.2	Formandos por ano e sexo no curso de Ciência da Computação	36
4.3	Desligados por ano e sexo no curso de Ciência da Computação	36
4.4	Motivo do desligamento no curso de Ciência da Computação	36
4.5	Ingressantes por ano e sexo no curso de Licenciatura em Computação . . .	37
4.6	Ingressantes por ano e sexo no curso de Licenciatura em Computação . . .	37
4.7	Desligados por ano e sexo no curso de Licenciatura em Computação . . .	37
4.8	Motivo do desligamento no curso de Licenciatura em Computação	37
4.9	Ingressantes por ano e sexo no curso de Engenharia da Computação	38
4.10	Desligados por ano e sexo no curso de Engenharia da Computação	38
4.11	Motivo do desligamento no curso de Engenharia da Computação	38

4.12	Recorte da planilha com os dados dos alunos de cursos de Computação da UnB.	39
4.13	Recorte da planilha com os dados das menções dos alunos de cursos de Computação da UnB.	39
4.14	Recorte da planilha com os dados das disciplinas e departamentos existentes na UnB.	40
4.15	Gráfico com o comparativo de média de notas por gênero e curso.	41
4.17	Gráfico com o comparativo de média de notas por gênero em disciplinas.	42
4.16	Gráfico com o comparativo de média de notas por gênero ao longo do tempo.	42
4.18	Gráfico com o comparativo de média de notas de Cálculo 1 por gênero em cada curso.	43
4.19	Gráfico com o comparativo de média de notas de Física 1 por gênero em cada curso.	43
4.20	Gráfico com o comparativo de média de notas de Computação Básica por gênero em cada curso.	44
4.21	Gráfico com a porcentagem de desistência por quantidade de semestres em Ciência da Computação.	44
4.22	Gráfico com a porcentagem de desistência por quantidade de semestres em Computação (Licenciatura).	45
4.23	Gráfico com a porcentagem de desistência por quantidade de semestres em Engenharia da Computação.	45
4.24	Rede bayesiana gerada a partir da classificação do sexo dos alunos.	46
4.25	Gráfico pizza com a relação cotas/sexo geradas pela rede bayesiana.	47
4.26	Gráfico pizza com a relação tipo de escola/sexo geradas pela rede bayesiana.	48
4.27	Gráfico pizza com a relação raça/sexo geradas pela rede bayesiana.	48
4.28	Gráfico pizza com a relação forma de ingresso/sexo geradas pela rede bayesiana.	49
4.29	Gráfico pizza com a relação forma de saída/sexo geradas pela rede bayesiana.	50
4.30	Gráfico pizza com a relação curso/sexo geradas pela rede bayesiana.	50
4.31	Gráfico pizza com a relação grau/gênero geradas pela rede bayesiana.	51
4.32	Gráfico pizza com a relação idade/gênero geradas pela rede bayesiana.	51
B.1	Parte 1 da saída da árvore de decisões com os dois perfis.	63
B.2	Parte 2 da saída da árvore de decisões com os dois perfis.	64
B.3	Parte 3 da saída da árvore de decisões com os dois perfis.	65
B.4	Parte 4 da saída da árvore de decisões com os dois perfis.	66

Capítulo 1

Introdução

1.1 Motivação

A Universidade de Brasília (UnB) foi criada há mais de 50 anos, com uma quantidade de alunos que a cada semestre aumenta consideravelmente. Consequentemente, necessita de um local para armazenar todas essas informações. O Centro de Processamento de Dados (CPD), existente desde 1991, é onde são armazenados e processados esses dados. Mas essas não são as únicas atribuições do CPD, nem as mais importantes, porém dentre seus princípios estão: Confiabilidade da informação e Segurança da informação; características importantes para manutenção e conservação desses dados [5]. Um dos setores em especial é a Gerência de Estratégia de Dados, que tem como competência, proporcionar a disponibilidade, confiabilidade, integridade e guarda dos bancos de dados sob custódia do Centro [5].

A análise dessas informações contidas nos bancos de dados do CPD, são muito importantes para a UnB, auxiliando os gestores em suas decisões, sejam elas no âmbito do ensino, da pesquisa ou da extensão.

Foi verificado desde 2010 entre as professoras do Departamento de Ciência da Computação (CIC), uma divergência entre a quantidade de alunos e alunas dos cursos relacionados à Computação. Então foi criado o projeto Meninas da Computação, com o objetivo de fornecer informação de qualidade às jovens em processo de escolha do curso de graduação de nível superior para prosseguimento nos estudos [10]. A Figura 1.1 mostra a logomarca do projeto. No ano de 2013, com o apoio do CNPq, o mesmo grupo de docentes criou o projeto Meninas na Engenharia, e com isso também foram realizadas pesquisas com essas jovens a respeito do conhecimento sobre computação, gerando então um banco de dados. Porém não foi realizado nenhum estudo de mineração de dados de gênero com os dados obtidos no projeto nem na UnB.

O foco desse trabalho é explorar técnicas de mineração de dados, no domínio da UnB, mais precisamente nos cursos relacionados a Computação, abordando o tema de gênero, e então obtendo informações relevantes a respeito do assunto, auxiliando os gestores, sejam ele da UnB ou dos departamentos, ou até mesmo do projeto mencionado acima a tomarem decisões importantes.



Figura 1.1: Logomarca do projeto Meninas da Computação [10]

1.2 Objetivos Gerais

O objetivo geral deste trabalho é extrair padrões nos dados providos pela UnB e pelos projetos Meninas da Computação e Meninas da Engenharia, que possam identificar tendências e hipóteses a respeito das diferenças de gênero nos cursos de Computação vinculados ao Departamento de Ciência da Computação.

1.3 Objetivos Específicos

Os objetivos específicos são justamente os passos do processo de Mineração de Dados: Limpeza dos Dados, Integração de dados, Seleção de dados, Transformação dos dados, Mineração de Dados, Avaliação de Padrões e Apresentação do conhecimento.

- Coleta de Dados através da aplicação de questionários a alunas do Ensino Fundamental e Médio;
- Limpeza e Integração de Dados obtidos nos formulários de pesquisa do projeto e dos dados obtidos na UnB UnB;
- Mineração de Dados e análise estatística dos dados após a limpeza, que será realizado separadamente para cada estudo de caso;
- Avaliação de padrões e Apresentação do conhecimento mais relevante ao projeto, de acordo com cada caso.

1.4 Estrutura do trabalho

Este trabalho é composto dos seguintes capítulos:

- Capítulo 2: Referencial Teórico. Nesse capítulo será explicado sobre Mineração de Dados em geral e seus métodos de mineração, assim como sobre as Mulheres na Computação e Sistemas de Banco de Dados;
- Capítulo 3: Estudo de Caso 1: Dados dos formulários do Projeto Meninas na Computação. Nesse capítulo serão realizados todos os passos referentes a mineração nos dados recebidos neste estudo de caso;

- Capítulo 4: Estudo de Caso 2: Cursos de Graduação de Computação da UnB. Nesse capítulo serão realizados análises estatísticas sobre os dados recebidos neste estudo de caso, assim como a mineração deles;
- Capítulo 5: Conclusão e Trabalhos Futuros.

Capítulo 2

Revisão Teórica

2.1 Mulheres e Computação

Nos anos iniciais da criação da Computação, a participação das mulheres era muito efetiva. Foram estas, que criaram a maioria dos programas na Segunda Guerra Mundial, e ocupavam posições de responsabilidade e influência na indústria computacional da época [1].

O número de mulheres na Computação começou a decair desde meados dos anos 1980 nos Estados Unidos e Reino Unido, mesmo com o aumento da participação destas em outras áreas de Ciência e Tecnologia. Isso causa [2] uma perda para as mulheres, no sentido da redução de melhores oportunidades de trabalho, mas também uma perda para a Ciência da Computação, que está possuindo menos perspectiva feminina, um fato que pode ter consequências negativas na sociedade. E [14] algumas pesquisas mostraram que estudantes mulheres tiveram menos experiência em computadores do que seus colegas homens no começo do curso e para a maioria das mulheres, Ciência da Computação não era sua primeira escolha.

E, ainda nesse âmbito de ensino superior, de acordo com pesquisas realizadas, apenas 4% das meninas americanas que estão finalizando ensino médio pretendem fazer Ciência da Computação. Além disso, 31% dos estudantes que estão fazendo este mesmo curso nos Estados Unidos, são mulheres. No nível de pós-graduação, esse número é ainda menor, 16% do número de diplomas de Doutorado entregues, foram a mulheres.

No cenário do Brasil, especificamente em Brasília , desde 2010 um grupo de docentes tem estudado este tema através de projetos apoiados pela UnB e CNPq. Estes projetos foram iniciados para investigar a hipótese dos cursos na área de computação e engenharia não serem a primeira escolha das estudantes recém egressadas do ensino médio, hipótese essa levantada a partir de interpretação de dados levantados pelo Centro de Seleção e de Promoção de Eventos (CESPE/UnB) que apontam ser de no máximo 10% a quantidade de alunos do gênero feminino que ingressam nas vagas destinadas aos cursos de Ciência da Computação, Licenciatura em Computação e Engenharia da Computação da Universidade de Brasília.

A partir da identificação desta diferença no percentual de alunos e alunas nos cursos de computação da UnB, o projeto elaborou um questionário, que será abordado mais detalhadamente no Capítulo 3.2, com 14 perguntas objetivas acerca do tema “percepção sobre computação”, para tentar levantar alguns motivos relacionados ao baixo número de

meninas nos cursos de computação, juntamente com este questionário foram construídas atividades de divulgação das diversas áreas da tecnologia e computação voltadas para o universo feminino na tentativa de iniciar um diálogo mais efetivo com essas alunas do ensino médio.

A primeira aplicação dos questionários e o início das atividades de divulgação ocorreram na Semana Nacional de Ciência e Tecnologia (SNCT) de 2011 em Brasília, o que se repetiu nos anos de 2012 e 2013.

Além de uma investigação sobre a diferença de gêneros já dentro dos cursos supracitados na UnB, este trabalho também fará uma análise sobre os dados resultantes da aplicação deste questionário ao longo destes três anos com o intuito de verificar e qualificar se as atividades propostas para atender aos objetivos e cumprir as metas descritas na formalização do projeto estão sendo alcançadas, não abstendo-se também de sugerir ajustes e apontar redirecionamentos mais eficazes para o maior sucesso do projeto, ou seja, que se obtenha sucesso em mostrar que computação também é coisa de menina.

2.2 Sistema de Banco de Dados

Segundo Ham e Kramber (2012) em tradução livre do inglês: Um sistema de banco de dados, também chamado de sistema de gerenciamento de banco de dados (SGBD), consiste em uma coleção de dados interrelacionados, conhecidos como base de dados, e um conjunto de softwares que gerenciam e acessam os dados.

Atualmente existem diversos SGBDs, onde destacam-se os sistemas baseados em orientação a objetos e os sistemas de banco de dados relacionais, para o desenvolvimento deste trabalho vamos focar nos SGBDs do tipo entidade-relacionamento por questões de estruturação do banco disponibilizado pela UnB.

Ham e Kramber (2012) definem um SGBD relacional como: “Uma base de dados relacional é um conjunto de tabelas, onde um nome único é atribuído a cada tabela. Cada tabela é constituída por um conjunto de colunas de atributos (colunas ou campos) e, geralmente, armazena um grande conjunto de tuplas (registros ou linhas). Cada tupla em uma tabela relacional representa um objeto identificado por uma chave única e descrito por um conjunto de valores de atributo.”

A importante implicação em usarmos este tipo de SGBD, é a possibilidade de implementarmos algoritmos de buscas utilizando linguagens de busca relacional como por exemplo SQL, desta forma o primeiro passo da mineração de dados, que consiste na coleta dos dados, pode ser simplificado e automatizado de forma mais intuitiva, tornando mais simples as etapas subsequentes, limpeza e integração dos dados.

2.3 Mineração de Dados

A mineração de dados é uma área relativamente nova da Ciência da Computação, também conhecida como o processo de Descoberta de Conhecimento em Banco de Dados (KDD). Esta área nasceu de uma evolução natural da tecnologia da informação [8].

A partir dos anos 70, foram criados Sistemas de Gerenciamento de Bancos de Dados (SGBD), que contribuiram para a maior facilidade nas operações realizadas em Banco de Dados, como a busca por instâncias específicas por exemplo. Esta contribuição foi

implementada através de algumas técnicas, ferramentas e linguagens como: Modelagem de Dados, Sistemas de Bancos de Dados Relacionais, Linguagens de Query (SQL), interfaces de usuário, dentre outros elementos.

Com as inovações alcançadas nos anos 70, após a metade dos anos 80 e início dos anos 90, surgiram Bancos de Dados mais avançados, que juntamente com a propagação da internet levaram a multiplicação da quantidade de dados, ou seja, geraram e tornaram disponíveis uma enorme quantidade de dados. A facilidade alcançada com estas inovações juntamente com a quantidade de dados despertou interesses na descoberta de padrões, começou-se a querer extrair conhecimento destes dados. Assim tornou-se necessária uma análise de dados avançada, onde é possível inserir estudos mais aprofundados no campo da Mineração de Dados. Na Figura 2.1 temos uma visão geral dessa evolução.

Em geral, estudos em mineração de dados são uma combinação da Estatística, Inteligência Artificial e pesquisa de banco de dados [11]. A Figura 2.2 mostra essa relação entre esses conceitos. Mas, como a Estatística e a Inteligência Artificial foram introduzidas nessa área?

A Estatística é bastante utilizada com a mineração de dados, uma vez que algumas das principais ferramentas implementadas pela mineração de dados utilizam técnicas da estatística, como as de vizualização e descritivas de dados que são usadas no que é chamado de análise descritiva dos dados, que são um conjunto de técnicas usadas para identificar a relação existentes entre diferentes variáveis.

Já a Inteligência Artificial usa algoritmos que buscam a solução otimizada, o que possibilita desenvolver uma mineração de dados baseada no modelo de raciocínio humano e possibilita a aprendizagem de máquina, elevando a participação dos sistemas computacionais de simples entidades passivas de processamento a entidades com poder de decisão, levando a mineração a possibilidade de responder perguntas mais complexas em diversos banco de dados.

Existem também outras áreas que utilizam da mineração de dados, dentre elas estão: Economia/Administração, que possuem uma grande quantidade de dados e diferentes banco de dados desde dados web e *e-commerce* até dados financeiros e transações bancárias, que precisam de análise para tomadas de decisões; Saúde, que também possuem vários e diferentes bancos de dados no domínio da área médica e farmacêutica; Pesquisa científica, que possui vários e enormes bancos de dados inexplorados em diferentes áreas que não podem ser explorados em meios tradicionais [7].

Então, o que é Mineração de Dados? É o processo de descoberta de vários modelos, resumos e valores derivados de uma determinada coleção de dados [9]. E também, qual é o objetivo de Mineração de Dados? É dar sentido a grandes quantidades de dados, em sua maioria sem supervisão [3].

O processo de mineração de dados é feito através de uma sequencia interativa de passos [8], que são resumidos pela Figura 1.3, são eles:

1. Limpeza dos dados: remoção de dados inconsistentes e interferências;
2. Integração de dados: várias fontes de dados podem ser combinadas;
3. Seleção de dados: os dados relevantes para a análise são filtrados do banco de dados;
4. Transformação dos dados: os dados são transformados e consolidados na forma apropriada para mineração, através de operações de junção ou de filtragem;

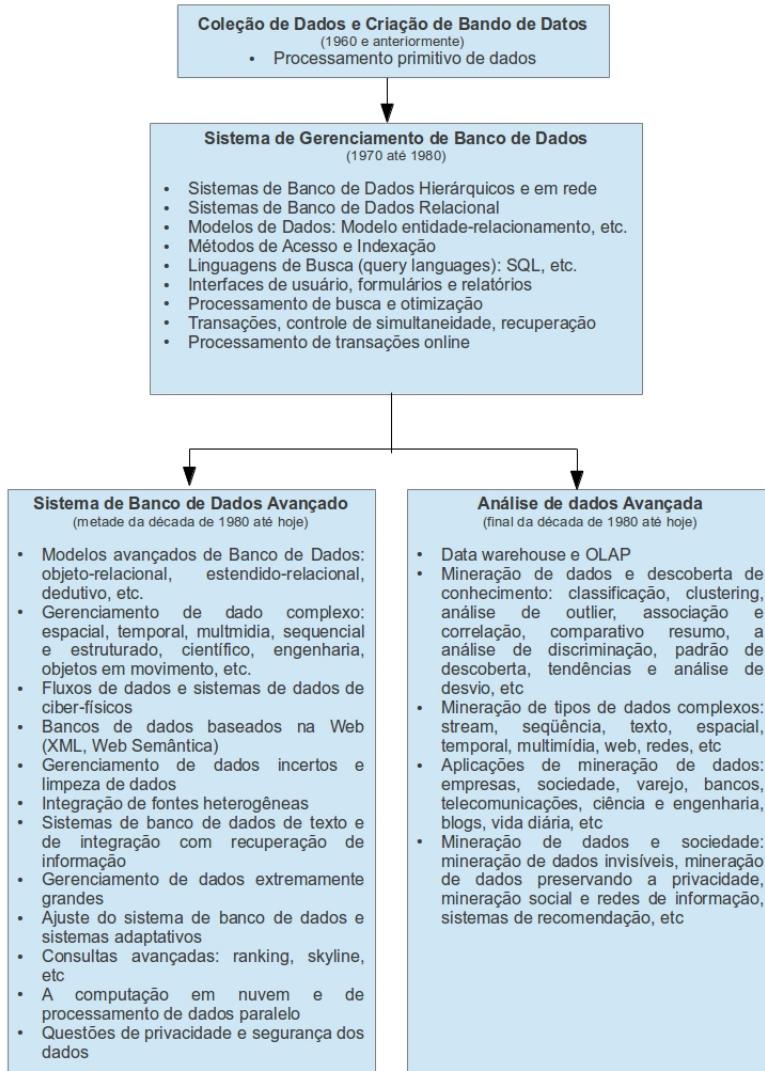


Figura 2.1: A evolução da tecnologia de sistemas de banco de dados [8].

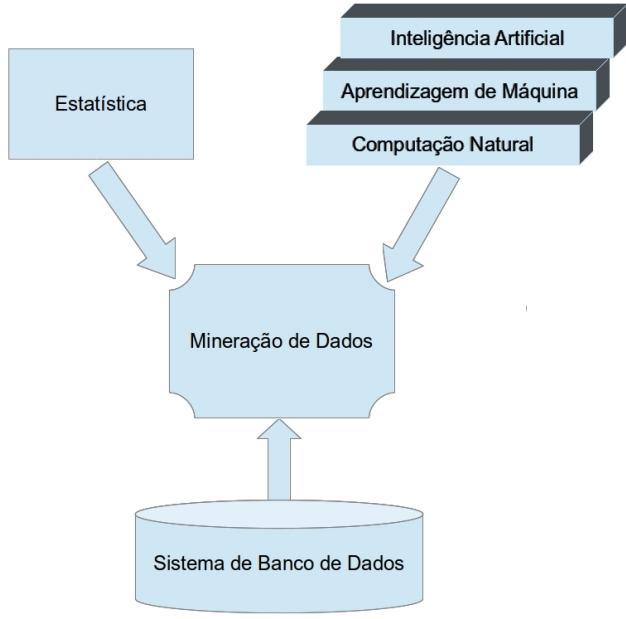


Figura 2.2: As raízes de Mineração de Dados [7].

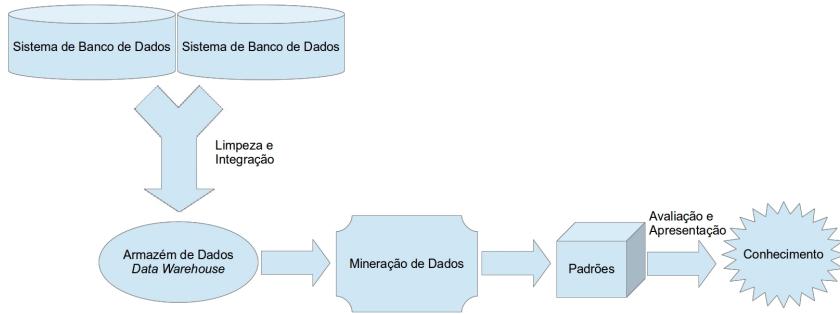


Figura 2.3: Mineração de dados como um passo no processo de descoberta [8]

5. Mineração de dados: um processo essencial onde métodos inteligentes são aplicados para extrair padrões de dados;
6. Avaliação de Padrões: verificação da usabilidade dos padrões extraídos;
7. Apresentação do conhecimento: visualização e as técnicas de representação do conhecimento são utilizadas para apresentar o conhecimento minerado para usuários.

Porém antes de iniciar o processo descrito no passo a passo, é necessário que se tenha em mãos os dados, ou seja, é preciso que se tenha acesso à algum sistema previamente populado com os dados que se deseja minerar, ou que tenha ocorrido algum processo de coleta de dados como por exemplo o que chamamos de *Web Mining* ou Mineração da Web. No caso de trabalharmos com *Web Mining*, podem existir diversos tipos de dados que despertam um grande interesse na realização da mineração, como os listados [13]:

- Conteúdo: Conteúdo de páginas web, em sua maior parte texto e imagens;

- Estrutura: Conteúdo que descreve a organização do conteúdo, ou seja, além do conteúdo e imagens, também a estrutura HTML ou XML;
- Uso: Dados que descrevem padrões de uso de páginas Web como endereços IP, referências de página e data e hora de acesso;
- Perfil de Usuário: Dados que fornecem informações demográficas dos usuários da página. Incluindo dados de registro e informações do perfil do usuário.

Com os dados em mãos, é realizado o primeiro passo do processo KDD, Limpeza dos dados. A importância desse passo é dada pela necessidade de manusear dados provisórios/-coletados, que podem vir incompletos, com interferência e/ou com erros com o intuito de melhorar a qualidade dos dados [6]. Interferências na mineração são as informações desnecessárias que se encontram nos dados, um exemplo mais comum seriam propagandas, barras de menu, etc. encontradas em páginas Web [16].

Gorunescu (2011) mostra formas de limpeza de cada um desses problemas. No caso de interferências, são utilizadas várias técnicas de filtragem para remover ou reduzir o efeito da interferência. No caso de valores que diferem significativamente da média de valores, esses dados podem ser removidos ou ser utilizados parâmetros (estatística) que não são tão sensíveis a esses valores extremos. O caso mais frequente é de valores incompletos, nessa situação podem ser eliminados os objetos que contém essa falta de dados, ou então realizada uma estimativa, sua substituição, ou até mesmo, caso seja possível, ignorá-los.

Os dois primeiros passos, limpeza e integração de dados, aparecem na Figura 1.3 juntos, pois quando múltiplas fontes de dados precisam ser integradas (armazém de dados ou sistemas de informação da web) a necessidade de limpeza aumenta significativamente. E isso acontece porque normalmente essas fontes possuem dados redundantes e é necessária a consolidação das diferentes formas de dados e de informações duplicadas antes de integrá-las [12].

Armazém de Dados, ou *Data Warehouse*, é uma repositório de informações coletada de várias fontes, armazenados em um esquema unificado, e normalmente encontram-se em um único local. Então para criar um armazém de dados é necessária ser feita a limpeza dos dados, para então serem integrados em um único local [8]. A Figura 1.4 mostra como funciona a geração e utilização deles.

Após o problema da qualidade dos dados ser resolvido, é necessário realizar um pré-processamento, equivalente a Seleção e Transformação, no qual consiste nos seguintes procedimentos [7]:

- Junção, que consiste na combinação de dois ou mais atributos (ou objetos) em apenas um, para reduzir assim sua quantidade e obter dados mais estáveis, com menos variação.
- Amostragem, que é o principal método de seleção de dados, representando o processo de extrair uma amostra que representa todo o grupo de dados. Porém, é preciso ter cuidado na hora de determinar o tamanho da amostra, pois é importante balancear a efetividade do processo de mineração de dados (obtido reduzindo a quantidade de dados sendo processada) e a perda significante de informações devido a baixa quantidade de dados. Esse problema é resolvido através da Estatística, que possui várias técnicas específicas, utilizadas dependendo do problema a ser resolvido.

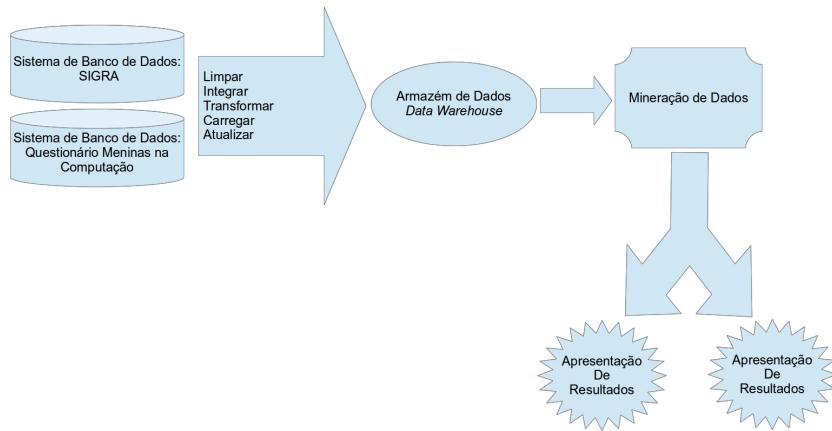


Figura 2.4: Geração e utilização de Armazém de Dados [8].

- Redução dimensional, que é realizado quando acontece um aumento do tamanho e propagação dos dados, e consequentemente o processamento de mais dados vai ser difícil, porque vai precisar de mais memória, que significa redução na velocidade de computação. Portanto, é obtido uma redução dimensional, que reduz a quantidade de tempo e memória necessária para o processamento dos dados, uma melhor visualização, eliminação de características irrelevantes e uma possível redução de interferência.
- Seleção de características, que é utilizada para eliminar características irrelevantes e redundantes, que podem causar confusão, utilizando métodos específicos.
- Criação de características, que é o processo de criação de novos (artificiais) atributos, que podem capturar melhor informações importantes nos dados do que os originais.
- Discretização e binarização, que é a transição de dados contínuos para dados discretos/categóricos (p. ex. mudança de valores reais para valores inteiros), e conversão de múltiplos valores em valores binários (p. ex. converter uma imagem de 256 cores em uma imagem preto e branco).
- Transformação de atributos, que é o princípio de conversão de atributos antigos em novos utilizando certa transformação (p. ex. funções matemáticas e normalização), e transformação que melhora o processo de mineração.

Após esse processamento dos dados é realizada a mineração dos dados ou definição da pesquisa, mas Gorunescu (2011) ainda adiciona que o passo de processamento dos dados pode ser repetido sempre que necessário.

Na mineração de dados, são utilizadas duas técnicas de aprendizagem, Aprendizagem Supervisionada e Aprendizagem não Supervisionada [8]. A primeira delas, é uma função de predição, ou seja, ao receber um objeto, conjunto de valores, será retornado um valor previsto que seguirá a tendência dos dados recebidos. O processo de classificação (método de predição) utiliza essa técnica, por exemplo, em estudos médicos, estão interessados em entender a progressão de algumas doenças, que são influenciadas por certos fatores de risco. Já a segunda técnica, ao contrário de Aprendizagem Supervisionada, o modelo

é adaptado a observações e não há uma saída deduzida, apenas objetos de entrada. O processo de clusterização (método descritivo) utiliza essa técnica, por exemplo, em uma empresa estão interessados em identificar clientes com o mesmo comportamento em relação a compra de certos tipos de produtos, e também a identificação de exceções nos dados que possam ser considerados identificação de fraude [7]. Então, no geral as funcionalidades da Mineração de Dados podem ser classificadas em duas categorias, descritiva e preditiva [8].

Todos esses passos serão explicados mais detalhadamente nos próximos capítulos, onde utilizaremos da mineração de dados para resolução dos problemas encontrados.

Então, como foi dito anteriormente nesta seção, a Mineração de Dados pode ser dividida em duas métodos: Métodos Preditivos e Métodos Descritivos; onde o primeiro deles utiliza de variáveis existentes para prever valores futuros (não existentes ainda) de outras variáveis, por exemplo, classificação, regressão, detecção de anomalias, etc.; e o segundo revela padrões nos dados, facilmente interpretados pelo usuário, por exemplo, clusterização, regras de associação, padrões sequenciais, etc. [7].

2.3.1 Classificação

O processo de Classificação [7] pode ser considerado também pelo processo de Taxonomia, que a princípio apareceu como ciência de classificar organismos vivos, mas depois transformou-se em uma ciência de classificação em geral. Portanto, a taxonomia (classificação) é o processo de colocar um objeto (conceito) específico em um conjunto de categorias, baseado em suas respectivas propriedades.

O processo de classificação é baseado em quatro componentes fundamentais [7]:

- Classe, que é a variável categórica que representa o rótulo colocado em objetos logo após sua classificação, por exemplo, lealdade de um cliente, classificação das estrelas (galáxia), classe de um terremoto (furacão), etc.
- Preditores, que são as variáveis independentes do modelo, representadas pelas características (atributos) dos dados a serem classificados e baseadas em qual classificação será feita, por exemplo, consumo de álcool e cigarro, pressão do sangue, frequência da compra, estado civil, características das imagens (de um satélite), registros geológicos específicos, direção e velocidade do vento, temporada, localização da ocorrência de um fenômeno, etc.
- Conjunto de dados de treinamento consiste no conjunto de dados contendo valores dos dois componentes anteriores, e é utilizado para “treinar” o modelo, para que reconheça a classe apropriada, baseado nos preditores disponíveis. Exemplos deste componente são: grupos de pacientes que tiveram infartos, grupos de clientes de um supermercado (investigado por pesquisas internas), bancos de dados contendo imagens de um monitoramento telescópico e acompanhamento de objetos astronômicos, bancos de dados em pesquisas de terremotos e bancos de dados em pesquisas de furacões.
- Conjunto de dados de teste contendo os novos dados que serão classificados pelo modelo (classificador) construído acima, e a precisão da classificação (performance do modelo) que, desta maneira, pode ser avaliada.

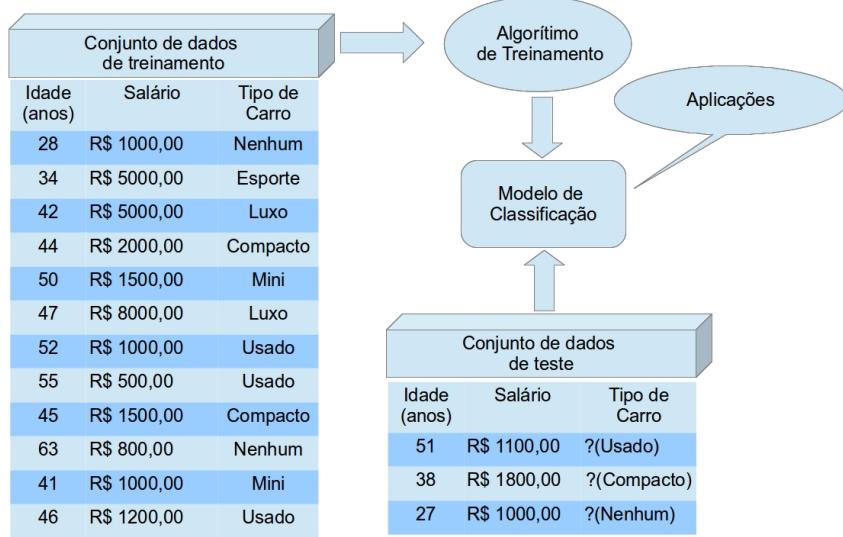


Figura 2.5: Estágios da construção de um modelo de classificação (concessionária) [7].

A Figura 2.5 [7] ilustra o processo de classificação, onde podemos visualizar os estágios de construção de um modelo de classificação de tipos de carros, que podem ser comprados por diferentes tipos de pessoas. No primeiro momento desta construção é selecionando um conjunto de dados de treinamento, que neste caso é uma tabela contendo a idade, o salário e o tipo de carro de um comprador específico. Após a seleção dos dados é aplicado um algoritmo de treinamento que gera um modelo de classificação relacionando os parâmetros descritos no conjunto de treinamento. Com o modelo de classificação disponível são realizados testes com um conjunto de dados para esta finalidade, onde é feito a classificação, neste caso relacionar cada perfil de comprador a um tipo de carro, e verificado o grau de certeza gerado pelo modelo. De acordo com o grau de certeza adquirido é necessário que sejam realizados ajustes no modelo de classificação até que seja possível realizar a classificação com um grau de certeza ótimo para a situação.

2.3.2 Regressão

A análise regressiva ou simplesmente regressão teve origem no final do século XIX com o trabalho do famoso geneticista Francis Galton que inovou com a noção de “Regressão para a média” onde dizia que a partir de duas medições dependentes, aos valores estimados para a segunda medição estão perto dos valores da média registrados na primeira medição. Levando para a área da estatística, este conceito de “Regressão para a média” implica a aplicação de um modelo matemático que estabelece, através de uma equação de regressão, a relação entre os valores de variáveis pré-definidas ou independentes e os valores de variáveis de saída ou simplesmente variáveis dependentes.

A análise regressiva visa levantar os seguintes pontos:

- Determinar a relação quantitativa entre duas variáveis;
- Prever os valores de uma variável de acordo com os valores de outra variável, determinar os efeitos das variáveis preditoras nas variáveis de resposta.

A partir deste método estatístico podemos desenvolver diversas aplicações na área de mineração de dados, como por exemplo [7]:

- No comércio, podemos prever a quantidade de venda de um novo produto baseado na quantidade de recursos gastos com a publicidade;
- Na área de meteorologia, prevendo a velocidade e a direção dos ventos em função da temperatura, umidade, pressão do ar, etc.;
- Na medicina, mensurando o efeito do peso de nascimento dos pais no peso de nascimento dos seus filhos;
- Na bolsa de valores, prevendo as tendências no mercado de ações baseado em análises de séries temporais.

2.3.3 Detecção de Anomalias

A Detecção de Anomalias/Divergências, como o nome sugere, lida com a descoberta de significantes divergências do “comportamento normal” [7]. A Figura 2.6 ilustra sugestivamente a existência de anomalias nos dados, onde podemos observar as estrelas de quatro e cinco pontas assim como o ouriço representando instâncias, dispostas em um gráfico onde o eixo x representa um atributo e o eixo y outro. Nessa disposição as estrelas de quatro pontas parecem mais próximasumas das outras, representando assim uma semelhança maior entre si, diferente da estrela de cinco pontas e do ouriço, que podem ser anomalias.

Já Han e Kramber (2012) dão um exemplo real de utilização dessa técnica. Um auditor de transações de uma empresa de cartões de crédito, para proteger seus clientes de fraude em seus serviços, monitoram as transações mais discrepantes dos casos típicos. Por exemplo, se o valor de uma compra é muito maior do que o normal para o dono do cartão, e se a compra ocorre em um local longe da cidade de residência do mesmo, então a compra é suspeita. É necessário detectar essas transações assim que elas ocorrem e entrar em contato com o dono do cartão de crédito para verificação. Essa é uma prática comum em várias empresas de cartão de crédito.

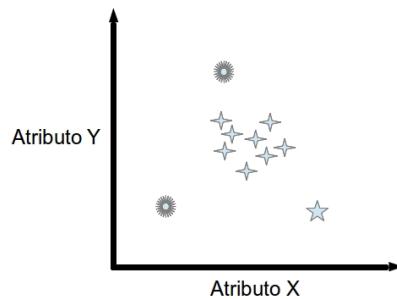


Figura 2.6: Anomalia nos dados [7].

2.3.4 Clusterização

Ainda segundo Gorunescu (2011), clusterização significa o método de dividir um conjunto de dados (registros/ tuplas/ vetores/ instâncias/ objetos/ amostras) em vários grupos (clusters), baseado em uma métrica de similaridade predeterminada. Portanto, podemos considerar o método de clusterização como um processo de “ classificação ” de objetos similares em subconjuntos dos quais os elementos possuem características em comum. É importante ressaltar também que além do termo Clusterização dos dados (clusterização), existem vários termos com significados similares, incluindo análise de clusters, classificação automática, taxonomia numérica, análise tipológica, etc. Também é importante não confundir o processo de Classificação (2.3.1) com o processo de Clusterização, pois no primeiro estamos lidando com um objeto que recebe um rótulo, pertencendo a uma classe em particular, enquanto no segundo um conjunto de objetos inteiro é particionado em subgrupos bem definidos. Um exemplo do segundo processo na vida real é: em um supermercado, diferentes tipos de produtos são colocados em departamentos separados (queijo, produtos de carne, utensílios, etc.), pessoas que se reúnem em grupos (clusters) em uma reunião baseada em afinidades, e divisão de animais ou plantas em grupos bem definidos (espécie, gênero, etc.).

Dado um conjunto de objetos, cada um deles é caracterizado por um conjunto de atributos, e sendo fornecida uma medida de similaridade, que ao dividi-los em grupos (clusters), tomar cuidado com objetos pertencentes a um mesmo cluster mais similares a um do que ao outro; e objetos em clusters diferentes são menos similares a um do que ao outro [7].

Esse processo de Clusterização vai ser bem sucedido se a similaridade intra-cluster e a não similaridade inter-clusters sejam ambas maximizadas. A Figura 2.7 mostra exatamente um processo de Clusterização bem sucedido [7].

Então, para realizar essa medida entre dois objetos, são utilizadas métricas, cada uma delas de acordo com a natureza dos dados e com o objetivo proposto. Gorunescu (2011) cita alguns dos medidores mais populares são:

- Minkowski;
- Tanimoto;
- Pearson’s r;
- Mahalanobis.

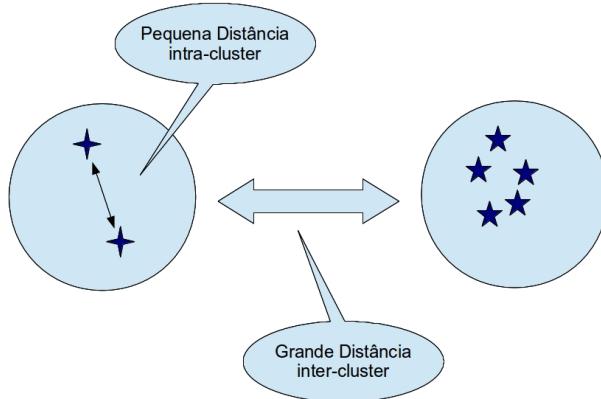


Figura 2.7: Exemplo de uma Clusterização bem sucedida. [7].

2.3.5 Regras de Associação

Para Gorunescu (2011), pela Descoberta de Regras de Associação nós entendemos o processo de identificação de regras de dependência entre diferentes grupos de fenômenos. Por exemplo, vamos supor que temos uma coleção de conjuntos, cada um contendo um número de objetos. Nossa objetivo é encontrar regras que conectem (associem) esses objetos e então, baseado nessas regras, poder prever a ocorrência de um novo item. A Figura 2.8 mostra o famoso exemplo da combinação Cerveja e Fralda, onde foi realizado um trabalho de mineração de dados nos registros de um supermercado a fim de descobrir comportamentos comuns aos clientes e desenvolver conhecimentos para auxiliar os executivos do supermercado a implementarem novas políticas para a ampliação do mercado. Como um dos resultados da mineração, foi apresentado a relação entre a compra de cerveja e fralda, ou seja, homens que compravam fraldas também compravam cerveja. A partir da identificação dessa relação, que em um primeiro momento pudesse parecer absurda, os executivos implementaram um novo mapa de distribuição dos produtos pelo supermercado, colocando prateleiras de cerveja também ao lado do setor de fraldas.

2.3.6 Padrões Sequenciais

Em várias aplicações como: biologia computacional, acesso Web, análise de conexões (logins) quando utilizando sistemas, os dados estão normalmente no formato de sequências. De uma forma mais sucinta, a questão deste contexto é a seguinte: dada uma sequencia de eventos discretos com um padrão, por exemplo « ... ABACDACEBABC

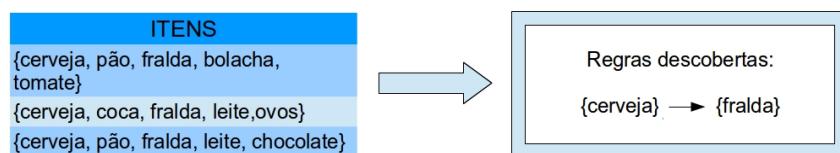


Figura 2.8: Esquema cerveja-fralda. [7].

... », e processando estes nós desejamos descobrir padrões que são frequentemente repetidos (no exemplo acima, A seguido de B, A seguido de C, etc.). Dada uma sequencia na seguinte forma: “ Tempo 1 (Temperatura = 28° C) -> Tempo 2 (Humidade = 67 por cento, Pressão = 765mm/Hg) ”, consiste de itens (atributo/valor) e/ou grupos de itens, que temos que descobrir padrões, a ocorrência de eventos nesses padrões. Gorunescu (2011) cita alguns exemplos da vida real, onde podemos utilizar esta técnica.

- Análise de grandes banco de dados nos quais são registradas sequencias de dados a respeito de transações comerciais variadas de um supermercado (por exemplo, o identificador do cliente - quando utilizando cartão de crédito, a data na qual foi realizada a transação, os produtos comprados - utilizando a tecnologia de código de barras, etc.), para agilizar a venda;
- Na Medicina, quando diagnosticando uma doença, os sintomas registrados são analisados em tempo real para a descoberta de padrões sequencias de uma doença específica, como por exemplo: “Nos primeiros três dias tiveram dores de cabeça e tosses desagradáveis, seguidos de outros dois dias com febre alta de 38-39° C, etc.”;
- Na Meteorologia - de forma geral - descobrindo padrões da mudança global de clima (aquecimento global, por exemplo), ou particularmente, descobrindo a ocorrência de furacões e tsunamis momentâneos, baseados em sequencia de eventos anteriores.

2.4 Weka, ferramenta de aprendizagem de máquina

Segundo Witten e Frank (2005), nenhum esquema de aprendizagem de máquina é completamente apropriado para todos os problemas de mineração de dados. Esse aprendiz universal existiria apenas em um mundo idealista, porque conjuntos de dados reais variam, e para obter modelos precisos a lógica do algoritmo de aprendizagem deve coincidir com a estrutura do domínio.

O Weka possui uma coleção de algoritmos de aprendizagem de máquina e várias ferramentas de pré-processamento de dados. Um de seus objetivos como ferramenta é de possibilitar o teste rápido de métodos existentes em diferentes conjuntos de dados de forma flexível. Esse programa disponibiliza um grande suporte para todo o processo de experimentação da mineração dos dados, desde a preparação dos dados de entrada e avaliação estatística de esquemas de aprendizagens até a visualização dos resultados da aprendizagem. Portanto, essa ferramenta é diversa e de fácil compreensão, pois possui uma interface simples e possibilita os usuários a compararem diferentes métodos de aprendizagem e selecionar o que melhor resolve seu problema.

O aplicativo Weka foi desenvolvido na Universidade de Waikato na Nova Zelândia, e essa palavra é uma sigla que significa *Waikato Environment for Knowledge Analysis*, que em sua tradução literal significa Ambiente de Análise de Conhecimento Waikato. Porém, pelo fato da palavra *weka* também representar um pássaro que não voa, encontrado apenas nas ilhas da Nova Zelândia, algumas pessoas de fora da universidade desconhecem o significado original. A linguagem de programação utilizada para desenvolver essa ferramenta foi Java, e ele foi distribuído sob os termos da GNU, *General Public License*, que traduzido literalmente significa Licença Pública Geral. Com isso, possibilita do sistema

ser executado em quase qualquer plataforma, mas tem sido testado nos principais sistemas operacionais utilizados hoje em dia (Linux, Windos e Macintosh).

Uma das grandes vantagens do uso dessa ferramenta é que ela inclui vários métodos para quase todos os problemas de mineração de dados citados na sessão 2.3, sendo eles: regressão, classificação, clusterização e regras de associação. Além disso, ela possui várias formas de visualizar os dados e ferramentas de pré-processamento, permitindo ao usuário a experiência de conhecer melhor os dados em questão. Todos os algoritmos utilizados no Weka requerem dados de entrada na forma de tabela relacional, portanto o programa aceita arquivos de variados formatos, inclusive conexões com diferentes bancos de dados através de conexões JDBC (*Java Database Connectivity*), e após isso os converte para o formato padrão utilizado pelos algoritmos (ARFF).

Existem várias formas de utilizar o aplicativo, uma delas é aplicar um dos métodos de aprendizagem a um grupo de dados e após isso, analisar a informação de saída, de forma a possuir mais informações sobre esses dados. Uma segunda forma é a de poder utilizar modelos aprendidos anteriormente para prever informações futuras. E por último, a possibilidade de aplicar vários algoritmos aos mesmo dados e comparar seus resultados de forma a escolher o que melhor deles para o problema. Esses métodos de aprendizagem são chamados de classificadores, e o Weka permite selecioná-los de forma interativa através de um menu. A maioria desses classificadores possuem parâmetros que podem ser ajustados, dependendo de seu tipo.

De um modo geral, a forma mais fácil de utilizar o Weka é através de uma interface gráfica chamada *Explorer*, e com ela podemos ter acesso a várias facilidades apenas utilizando o menu de seleção e o preenchimento de formulários. O exemplo mais simples de utilização dessa ferramenta é a leitura de um conjunto de dados, e a geração de uma árvore de decisões a partir dele. Uma das vantagens dessa interface é que ela impõe ao usuário a trabalhar de forma correta, através de escolhas em menus, impossibilitando erros ao deixar cinza as opções indisponíveis, e as que ficaram disponíveis na forma de um formulário. Além disso, possui dicas úteis que aparecem ao passar o mouse por cima de cada item, explicando suas funções. Mas apesar disso, o Weka pode dar uma sensação de facilidade que pode fazer o usuário se enganar, apesar de possuir valores padrões para garantir o menor esforço, se ele não estiver sabendo o que faz, não irá conseguir entender o que os resultados significam.

Além do *Explorer*, existem outras duas interfaces gráficas no Weka, a *Knowledge Flow*, que permite ao usuário criar configurações para processamento de dados em streaming, e a *Experimenter*, que possibilita a comparação de técnicas de aprendizagem de forma mais avançada. Porém nessa pesquisa utilizaremos apenas a interface gráfica *Explorer*.

A Figura 2.9 mostra a tela inicial do Weka, com as interfaces disponíveis. O programa também possui uma documentação online que possui a lista completa de algoritmos disponíveis e está sempre atualizada. Nessa documentação mostra também como é possível inserir um algoritmo de aprendizagem próprio.

A plataforma está disponível para download através do endereço eletrônico <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>, tanto no instalador específico de um sistema operacional, ou um arquivo Java jar executável.

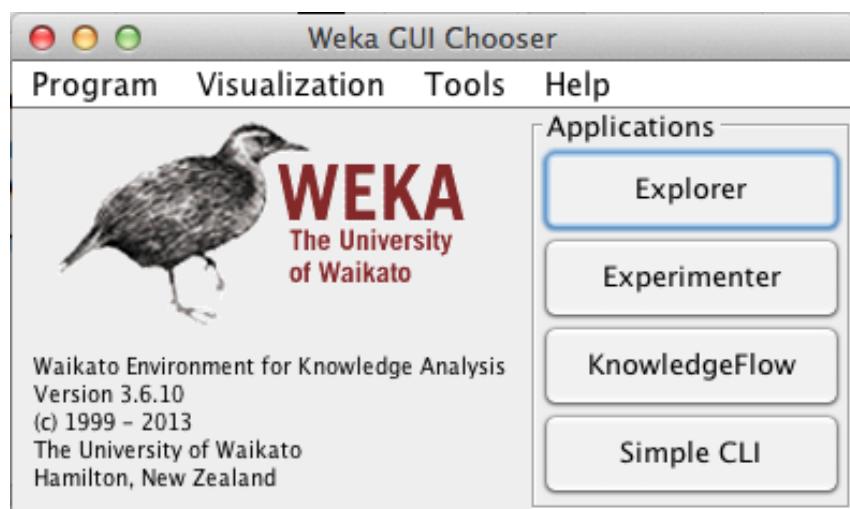


Figura 2.9: Tela inicial do Weka.

Capítulo 3

Estudo de Caso 1: Dados dos formulários do Projeto Meninas na Computação

Com o objetivo de realizar uma mineração de dados mais interativa, utilizamos a sequência de passos sugeridos por Han e Kramber (2012) citados na introdução deste trabalho, utilizando então, diversas ferramentas para facilitar este processo.

3.1 Coleta de Dados

Durante os anos de 2011 a 2013, nas feiras de Ciência e Tecnologia realizadas em Brasília, o projeto Meninas na Computação e Meninas na Engenharia, com o apoio de alunas e professoras do Departamento de Ciência da Computação, divulgou a área para meninas do Ensino Médio através de posters, banners, jogos, conversas. Além disso, foi aplicado um questionário de percepção sobre computação a essas mesmas pessoas, conforme mostrado na Figura 3.1.

No questionário em questão haviam questões de identificação como: Sexo, Série Escolar, qual área de estudos pretende prestar vestibular (Exatas, Biológicas ou Humanas) se está pensando em fazer um curso superior para Computação; Questões gerais de percepção sobre computação como: locais que usa o computador, atividades realizadas no computador; E questões específicas, de sim, não ou talvez, como: Trabalhar em computação dá prestígio? Quem trabalha em computação ganha bem? Entre outras. Esse questionário será mais detalhado posteriormente.

Os questionários foram impressos pelo CESPE (Centro de Seleção e Promoção de Eventos da Universidade de Brasília), no formato de preenchimento similar ao de concursos, e após aplicados foram enviados de volta ao Centro para processamento dos dados. Após isso, os resultados foram devolvidos em forma de planilha para os professores responsáveis pelo projeto. Portanto, a coleta de dados do nosso projeto foi realizada de forma indireta pelos participantes do Projeto de Pesquisa Meninas da Computação, e processados pelo CESPE.



Preencha os **círculos** completamente e com nitidez, utilizando caneta esferográfica de tinta preta.

Questionário: PERCEPÇÃO sobre Computação

2012

SEXO:

- Feminino Masculino

SÉRIE ESCOLAR:

- Ensino Fundamental 2º ano - Ensino Médio/Técnico Supletivo
 1º ano - Ensino Médio/Técnico 3º ano - Ensino Médio/Técnico Ensino Superior

PARA QUAL ÁREA VOCÊ PRETENDE FAZER UM CURSO SUPERIOR?

- Exatas Biológicas e Saúde Humanas

VOCÊ ESTÁ PENSANDO EM FAZER UM CURSO SUPERIOR PARA COMPUTAÇÃO?

- Sim Não Não sei ainda

PERCEPÇÃO SOBRE COMPUTAÇÃO

1. Marque todos os locais em que você usa o computador:

- | | |
|---|---|
| <input type="radio"/> em casa | <input type="radio"/> no trabalho |
| <input type="radio"/> em casa de parentes | <input type="radio"/> em lan-house |
| <input type="radio"/> em casa de amigos | <input type="radio"/> biblioteca |
| <input type="radio"/> na escola | <input type="radio"/> centros de inclusão digital |

2. Marque todas as atividades que você realiza ou realizou no computador:

- | | |
|---|--|
| <input type="radio"/> edição de texto (word, outros) | <input type="radio"/> e-mail |
| <input type="radio"/> edição de imagem (outros) | <input type="radio"/> jogos |
| <input type="radio"/> planilha (excel, outros) | <input type="radio"/> desenvolvimento de páginas |
| <input type="radio"/> banco de dados (access, outros) | <input type="radio"/> programação |
| <input type="radio"/> acesso a internet (pesquisas sobre conteúdos, notícias, outros) | <input type="radio"/> outros. |
| <input type="radio"/> redes sociais (facebook, orkut, etc) | |

NAS QUESTÕES DE 3 A 14, MARQUE APENAS UMA OPÇÃO.

	SIM	NÃO	TALVEZ
3. Um curso superior de computação só ensina a usar software?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4. Um curso superior de computação usa pouca matemática?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5. A maioria dos alunos de computação é do sexo masculino?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6. É preciso saber usar computadores para fazer um curso superior de computação?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7. É preciso fazer curso superior de computação para trabalhar na área?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8. A sua família gostaria que você fizesse vestibular para computação?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
9. É difícil conseguir emprego em computação depois de formado?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10. Quem trabalha com computação tem poucas horas de lazer?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
11. Trabalhar em computação permite que você exerça sua criatividade?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
12. Trabalhar em computação dá prestígio?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
13. Quem trabalha em computação ganha bem?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
14. Trabalhar em computação permite atuar em outras áreas diferentes?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figura 3.1: Questionário aplicado as meninas durante as Semanas Nacionais de Ciência e Tecnologia.

Sexo	Série Escolar	Área pretende fazer Curso Superior							Curso Superior para Computação						
		1 = Exatas , 2 = Biológicas e Saúde e 3 = Humanas / " . " = Sem Marcação / " * " = Dupla Marcação							1 = SIM , 2 = NÃO e 3 = Não sei ainda / " . " = Sem Marcação / " * " = Dupla Marcação						
F	6							3							3
F	2							2							3
F	2							1							3
F	1							3							2
F	1							3							2
F	6							3							2
F	6							3							2
F	4							2							1
F	4							2							2
F	2							2							3

Questão 01	Questão 02	Questão 03	Questão 04	Questão 05	Questão 06	Questão 07	Questão 08	Questão 09	Questão 10	Questão 11	Questão 12	Questão 13	Questão 14
1 = Com marcação	1 = Com marcação	Questão 03	Questão 04	Questão 05	Questão 06	Questão 07	Questão 08	Questão 09	Questão 10	Questão 11	Questão 12	Questão 13	Questão 14
0 = Sem marcação	0 = Sem marcação												
1 = SIM , 2 = NÃO e 3 = TALVEZ / " . " = Sem Marcação / " * " = Dupla Marcação													
11011100	1110111000	2	3	1	1	2	3	3	2	1	1	3	1
11110000	11101100101	2	3	1	1	1	1	2	3	1	1	1	1
11110000	10000110001	2	3	1	1	1	3	2	3	1	1	1	1
11100000	110011101011	2	2	1	1	3	2	2	3	1	2	3	1
10100000	110011101011	2	2	1	1	3	2	2	3	1	2	3	1
10011100	1111111000	2	2	1	1	1	3	2	2	1	1	1	1
10000110	10001110001	2	2	1	1	1	3	3	2	2	1	1	1
11111111	111011101011	3	1	1	1	2	3	1	1	1	2	1	1
11100010	11000011010	2	2	2	2	1	1	3	2	1	2	1	1
10100000	11001110001	2	3	1	2	1	2	2	3	1	3	3	1

Figura 3.2: Formato dos dados recebidos pelo CESPE.

3.2 Formatação e Limpeza dos Dados

Foram entregues três planilhas, cada uma delas representando o seu ano de resposta correspondente. Os dados estavam distribuídos em colunas, onde cada uma delas representava uma questão. O questionário possuía dois tipos de questões: as que permitiam apenas uma resposta, como por exemplo: Para qual área você pretende fazer um curso superior?; e as que permitiam várias respostas, como por exemplo: Marque todos os locais em que você usa o computador. Para o primeiro tipo, cada resposta recebe uma numeração de acordo com a ordem que aparece no questionário, e a planilha recebida possui esses números de acordo com a escolha do respondente. Já no segundo tipo de questão, em uma única coluna, é mostrado um número binário para cada possível resposta, onde o número "1" representa que aquela resposta foi marcada, e o número "0" aquela que não foi marcada, porém por ser um campo numérico, na planilha não aparecerá as primeiras respostas não marcadas, por exemplo, caso uma resposta seja "00100", na planilha aparecerá apenas "100". A figura 3.2 mostra um recorte de uma das planilhas recebidas.

Para realizar uma limpeza nos dados, foi criado um banco de dados utilizando o SGBD (Sistema de Gerenciamento de Banco de Dados) MySQL, selecionado por ser rápido, grátis e fácil de instalar. Dentro dele foi criada uma tabela, onde cada atributo desta representa uma coluna na planilha. Para realizar a importação dos dados da planilha para a tabela, foi utilizado o aplicativo Sequel Pro, que é um gerenciador de Bancos de Dados MySQL para usuários do sistema operacional Mac, este foi escolhido por ser grátis, de fácil manipulação e ter um algoritmo de importação de dados eficiente.

Foi realizada a conversão da planilha para o formato CSV (Valores separados por vírgula), e após isso, foi iniciada a importação pela plataforma selecionada. Tendo como primeiro passo, a seleção do arquivo a ser importado, e em seguida, o mapeamento das colunas do arquivo com os atributos da tabela e suas tipagens, conforme demonstrado na Figura 3.3. No mapeamento, o ano foi o único dado inserido manualmente, de acordo com a planilha selecionada. Já no caso das colunas de várias respostas, foi escolhido temporariamente o primeiro atributo como armazenador do dado.

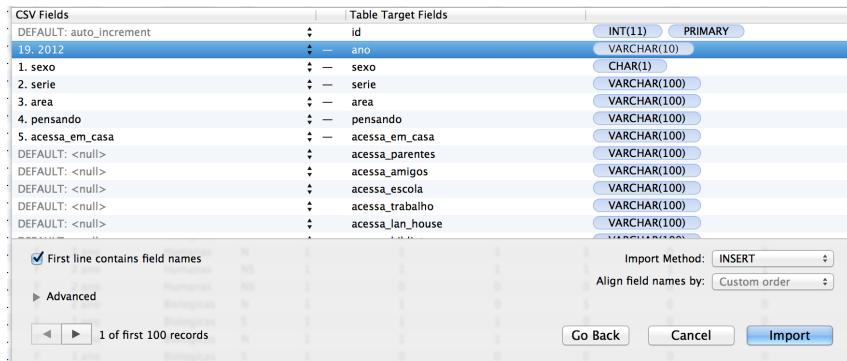


Figura 3.3: Mapeamento de colunas da planilha com atributos da tabela.

Na criação da tabela no banco de dados, foram modelados atributos para realizar a separação das respostas selecionadas nas questões um e dois. Na questão um foram modelados oito atributos relacionados ao acesso a computadores. As respostas possíveis foram: em casa, em casa de parentes, em casa de amigos, na escola, no trabalho, em lan-house, biblioteca e em centros de inclusão digital. No caso da questão dois, foram modelados onze atributos relacionados às atividades realizadas no computador. As respostas possíveis foram: edição de texto, edição de imagem, planilha, banco de dados, acesso a internet, redes sociais, email, jogos, desenvolvimento de páginas, programação e outros.

Para manter o banco de dados consistente, foram realizadas algumas *queries* para manter a padronização de cada atributo. Por exemplo, na coluna de sexo, as respostas possíveis eram: M para Masculino e F para Feminino. Então, tudo que fosse diferente dessas duas opções, deveriam ser considerados nulos.

No atributo de série, foi necessário fazer a tradução dos dados de acordo com a ordem apresentada no questionário, que foi a ordem apresentada pelo CESPE. Todas as respostas com o número um, deveriam ser substituídas por Ensino Fundamental, de número dois por 1º ano, de número três por 2º ano, de número quatro por 3º ano, de número cinco por Supletivo e de número seis por Ensino Superior, todos os outros deveriam ser considerados nulos.

Assim como o campo de série, o campo de área também precisou ser traduzido. Essa questão perguntava qual área o respondente estava pretendendo fazer o curso superior, e as respostas número um deveriam ser traduzidas por Exatas, de número dois por Biológicas e de número três por Humanas, todas as outras respostas deveriam ser traduzidas por nulo.

A última questão da primeira parte, perguntava aos respondentes se eles estavam pensando em fazer um curso superior para computação, e as respostas possíveis eram: Sim, Não e Não sei ainda. Foi necessário fazer a tradução para essas respostas, onde um representava Sim, dois representava Não e 3 representava Não Sei, todas as respostas diferentes dessas deveriam representar nulo.

Como foi explicado anteriormente, os dados advindos da planilha para a coluna da questão um foi gerado no formato numérico, portanto para realizar a atualização dos novos atributos foi necessário gerar uma *query* específica para cada atributo. Primeiro foi necessário verificar se o campo era nulo, caso verdadeiro, o retorno seria zero, pois o respondente não marcou essa resposta. Após isso, foi verificado se o campo em questão é

menor do que o valor mínimo possível para ele ser verdadeiro, pois se for menor, o campo não foi marcado, então o retorno é zero também. Caso seja maior, há a possibilidade do respondente ter marcado essa resposta, então o campo é atualizado com o número correspondente a ordem da resposta, podendo ele ser zero ou um.

Da mesma forma que foi realizado na questão um, foi necessária realizar uma atualização dos campos correspondentes a questão dois, que também possibilitava a escolha de mais de uma resposta no questionário. A *query* utilizada nesses campos é um pouco maior, pelo fato da questão dois possuir mais opções de resposta, resultando em uma string maior.

As questões seguintes, de três a quatorze possuem as mesmas opções de resposta, sendo possível escolher apenas uma delas. Por serem questões de percepção, as respostas são: Sim, Não e Talvez. O comando utilizado é similar entre essas questões, substituindo as opções de número um pela letra S, de número dois pela letra N, de número três pela letra T, e o que for diferente disso por nulo.

Finalmente, após toda a transformação dos dados, foi realizada a seleção, onde houve a separação entre os respondentes que estão pensando em fazer computação e os outros, para ser realizada uma mineração de dados separada desses dois perfis. Foram utilizadas duas *queries* para isso.

3.3 Método de Análise: Usando os dados na plataforma Weka

Após a formatação e limpeza dos dados realizada na seção anterior, utilizamos a interface gráfica *Explorer* da ferramenta Weka, conforme explicado na seção 2.4, para realizar a análise desses dados. Na Figura 3.4 podemos visualizar a tela inicial da ferramenta em questão. Nessa interface gráfica podemos visualizar seis abas na parte superior da tela, onde após a seleção dos dados, podemos encontrar seis painéis diferentes, que correspondem às várias tarefas de mineração de dados que o Weka suporta.

De forma geral, os passos de utilização dessa interface gráfica e geração de uma árvore de decisões, é necessário preparar os dados, iniciar o *Explorer*, e carregá-los na plataforma. Com isso, é então selecionado um método de construção dessa árvore, para então esta ser gerada e poder realizar a interpretação da saída. Caso o resultado não seja satisfatório, pode ser construída outra árvore com outro algoritmo ou método, para realizar a comparação dos resultados obtidos, visualizando graficamente esses modelos e os próprios conjuntos de dados, verificando qualquer erro que pode ter ocorrido.

3.3.1 Convertendo para ARFF

Normalmente os dados estão em forma de planilhas ou bancos de dados, porém o método de armazenamento nativo do Weka é o formato ARFF. O programa pode realizar essa conversão automaticamente, porém para deixar o carregamento dos dados mais rápido, é necessário ser realizada uma conversão manual desses dados. Como os dados em planilha podem ser mais facilmente convertidos para ARFF, foi necessário exportar os dados selecionados para o formato CSV. Esse formato consiste de uma lista de instâncias e os valores de atributos para cada instância são separados por vírgula, ou seja, no caso

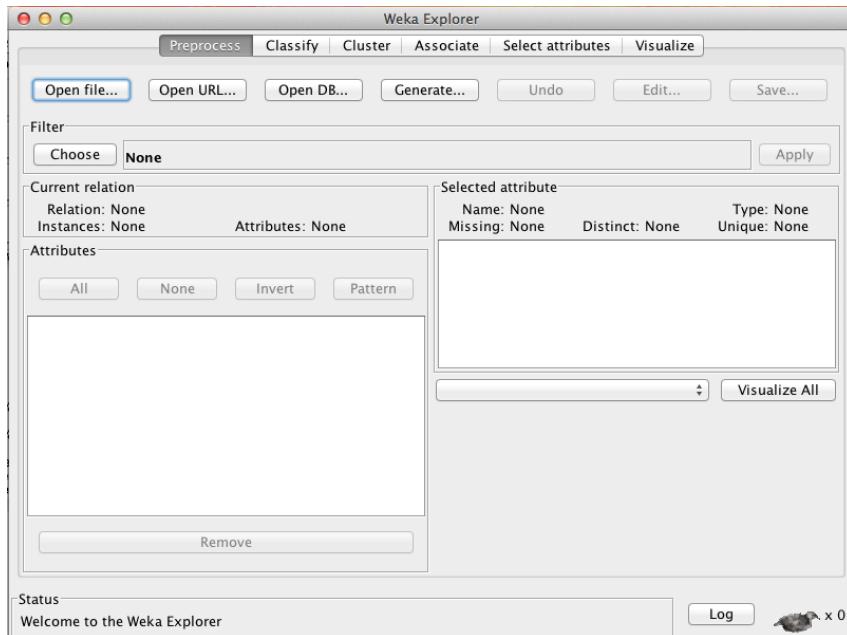


Figura 3.4: Tela inicial do *Explorer*.

específico cada linha representa os dados de um aluno e cada atributo é separado por vírgulas. A maioria das planilhas e dos programas de bancos de dados permitem ao usuário exportar os dados para este formato, como uma lista de registros separando os items por vírgula. Após isso, para realizar a conversão, é necessário abrir o arquivo em um editor de texto, adicionar a tag @relation seguido de um nome para o conjunto de dados, a tag @attribute seguido da informação do atributo e por fim a tag @data em uma linha, e nas linhas seguintes os dados separados por vírgula, então é necessário salvar o arquivo como um texto, com a extensão arff. Na Figura 3.5 podemos ver a distinção entre os formatos e o resultado da conversão manual.

ano	serie	sabe_banco_dados	sabe_usar_internet
2011	1 ano	N	S
2011	1 ano	S	S
2011	1 ano	N	S
2011	1 ano	S	S
2011	1 ano	N	S
2011	1 ano	N	S
2011	1 ano	N	S
2011	1 ano	N	S
2011	1 ano	N	S
2011	1 ano	S	S
2011	1 ano	N	S
2011	3 ano	N	S

1. Banco de Dados

2. CSV

```
@relation questionario

@attribute ano {2011, 2012, 2013}
@attribute serie {Fundamental, "1 ano", "2 ano", "3 ano", Supletivo}
@attribute sabe_banco_dados {S, N}
@attribute sabe_usar_internet {S, N}
@attribute sabe_programar {S, N}
@attribute sabe_basico {S, N}
@attribute q4_sup_pouca_mat {S, N, T}
@attribute q7_prec_curs_sup_p_trab {S, N, T}
@attribute q10_trab_pouco_lazer {S, N, T}
@attribute q11_trab_comp_criatividade {S, N, T}
@attribute q12_trab_comp_prestigio {S, N, T}
@attribute q13_trab_comp_ganha_bem {S, N, T}
@attribute q14_trab_comp_atuar_outras_areas {S, N, T}

@data
2011,"1 ano","N","S","S","S","S","N","T","S","S","T"
2011,"1 ano","S","S","S","S","N","S","T","S","S","S"
2011,"1 ano","N","S","N","S","T","S","T","S","N","T"
2011,"1 ano","S","S","S","S","T","S","T","S","S","T"
2011,"1 ano","N","S","N","N","S","T","S","T","S,"
```

3. ARFF

Figura 3.5: Formatos: 1. Banco de Dados, 2. CSV e 3. ARFF.

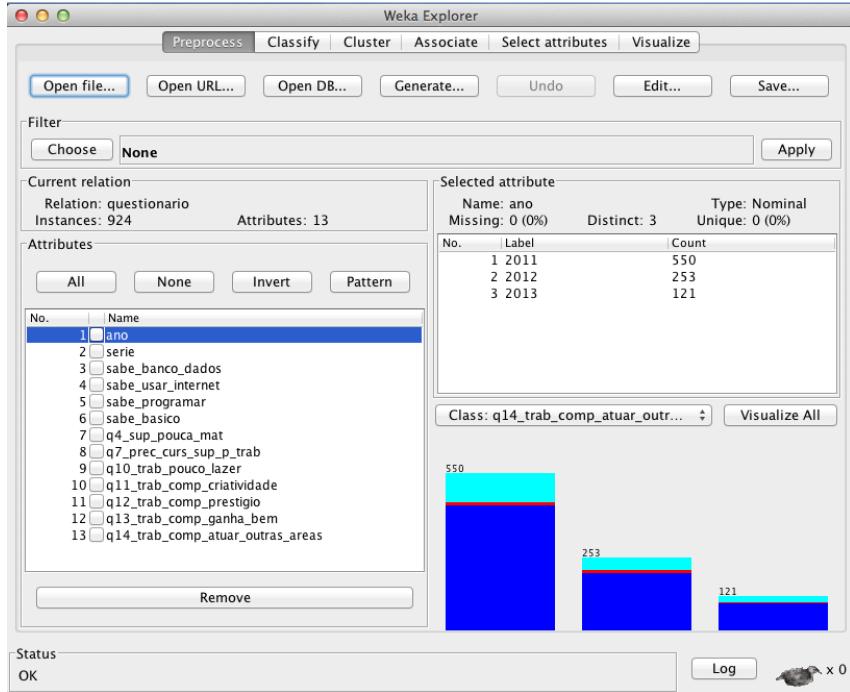


Figura 3.6: Tela inicial com os dados carregados.

3.3.2 Carregando os dados e fazendo a análise inicial

Após entrar no Weka e selecionar a interface *Explorer*, foi selecionado o arquivo de importação através do botão *open file* encontrado no canto superior esquerdo da tela. Depois que o arquivo for carregado, será exibido na tela a quantidade de atributos e instâncias, além de exibir em uma área reservada os respectivos atributos. A Figura 3.6 mostra os dados do questionário filtrados pelo perfil de quem estava pensando em prestar vestibular para algum curso de computação.

No quadro inferior direito dessa tela, é possível selecionar um classificador, onde a partir deste, pode ser feita uma análise comparativa. Por exemplo, selecionando o atributo *sabe-banco-dados* como classificador, onde a cor vermelha representa quem não sabe banco de dados e a cor azul para quem sabe, e clicando no botão para visualizar o comparativo de todos os atributos, conseguimos chegar a algumas conclusões apenas observando essa tela. Na Figura 3.7 temos a imagem dessa tela de visualização.

3.3.3 Utilizando o algoritmo de classificação e analisando a saída

Com todos os dados carregados na interface, e a análise inicial feita, é necessário fazer a mineração de dados, e para isso, algoritmos de classificação foram utilizados. Então, como primeiro passo, é preciso clicar no botão *Choose* no canto superior esquerdo para escolher qual tipo será utilizado. Ao realizar essa ação, foi aberta uma tela com várias pastas, onde existem vários tipos de algoritmos. Nesse caso, foi utilizado o algoritmo de árvore de decisões J48, que é encontrado dentro da pasta *trees*. A Figura 3.8 mostra essa lista de classificadores.

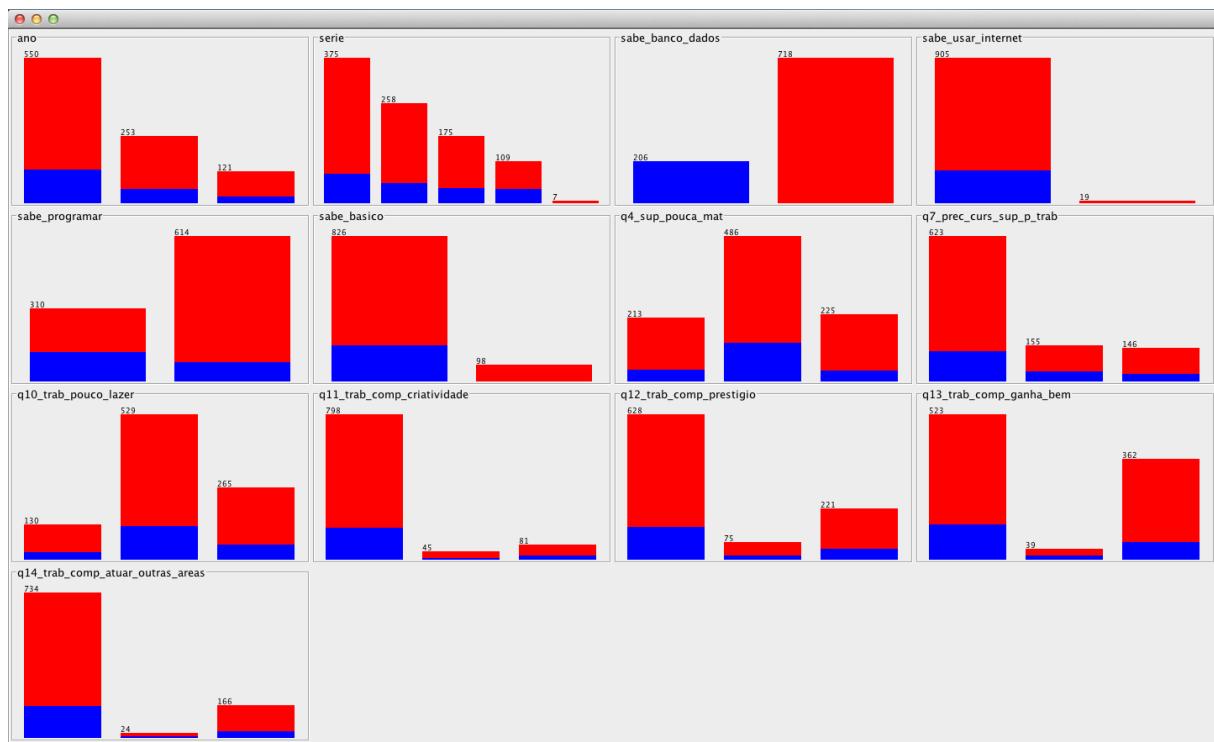


Figura 3.7: Tela para análise visual dos dados. Vermelho: quem não sabe banco de dados; Azul: quem sabe banco de dados

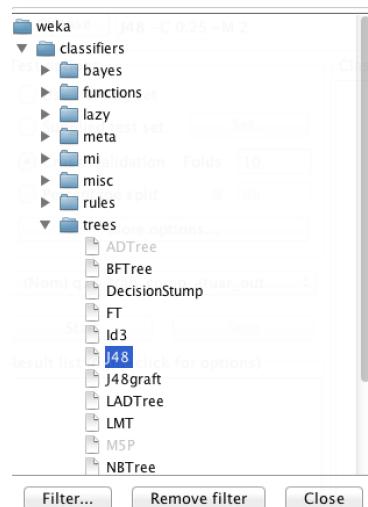


Figura 3.8: Lista de classificadores.

Após isso, é necessário escolher o atributo classificador e então apertar o botão *Start*. Porém, dependendo do tamanho dos dados, do classificador escolhido e do algoritmo, essa execução pode demorar algum tempo. E enquanto o algoritmo está executando, o passarinho weka fica se movimentando até a saída ser mostrada no painel principal. Só que nesse caso, como temos poucos dados, a saída é imediata.

As Figuras 3.9 e 3.10 mostram a saída completa da mineração, utilizando o atributo *sabe-programacao* como classificador. No início do arquivo de saída, é mostrado um resumo do conjunto de dados utilizado, que nesse caso foi o *10-fold cross-validation*. Segundo Witten e Frank (2005), na *cross-validation* os dados são divididos em dez partições aproximadamente iguais e em cada turno, uma é utilizada para teste e o restante para treinamento e ao final de cada um deles é calculado a taxa de erro. Portanto, o procedimento de aprendizagem é executado dez vezes em diferentes conjuntos de dados de treinamento, consequentemente dez estimativas de erros, gerando um valor médio para esse no resultado final. Esse é o tipo de teste padrão da ferramenta, mas é possível selecionar outros tipos de teste no painel chamado *Test options*, acima do local onde foi escolhido o atributo classificador e abaixo do campo onde o algoritmo classificador é selecionado. Nesse caso, o primeiro nó da árvore, ou raiz, é o atributo *sabe-banco-dados*, onde este é dividido nas respostas possíveis (Sim ou Não), seguido do atributo *q11-trab-comp-criatividade*, referente a questão onze do questionário, e dependendo da resposta deste atributo (Sim, Não ou Talvez) são utilizados diferentes atributos, que transformam em uma árvore de altura 5. Quando ao final da linha na qual encontra-se o atributo, existir o símbolo referente a dois pontos, significa que este representa uma folha da árvore, seguido do número de instâncias que utilizaram este mesmo caminho para chegar a esta folha. Este número é decimal pelo fato do algoritmo utilizar instâncias fracionais pra tentar descobrir valores ausentes. As instâncias incorretamente classificadas aparecem separadas por uma barra, por exemplo, na última linha da árvore temos: *sabe-banco-dados = N: N (718.0/186.0)*, significando que das 718 instâncias chegaram a esta folha, 186 foram classificadas incorretamente. Abaixo da árvore, encontramos o número de folhas e o número total de nós, ou seja, o tamanho da árvore.

Na próxima parte da saída, Figura 3.10, encontramos a estimativa da predição da performance da árvore que foi obtido utilizando determinado tipo de teste. Nesse caso, quase 30% das instâncias foram classificadas incorretamente na *cross-validation*. No final do arquivo, temos a matriz de confusão, e pode ser observado que 63 instâncias da classe *Sim* foram atribuídos a classe *Não*, e 208 da classe *Não* foram atribuídos a classe *Sim*.

Além do erro de classificação, é possível visualizar também a estatística *Kappa*, o erro médio absoluto e a raiz do erro médio quadrado das estimativas encontradas. A estatística *Kappa*, segundo Witten e Frank (2005), representa a porcentagem de sucesso de predição a partir das instâncias corretamente classificadas na matriz de confusão. A raiz do erro médio quadrado é a raiz quadrada da perda média quadrática. Já o erro médio absoluto apesar de ser calculado de forma similar, utiliza o valor absoluto ao invés da diferença quadrática.

Ao final, na Figura 3.11 temos a árvore de forma visual, que pode ser gerada a partir da lista de resultados no canto inferior esquerdo da tela, basta clicar com o botão direito e selecionar *Visualize tree*.

Para realizar um análise comparativa entre as pessoas entrevistadas que estavam pensando em fazer computação e as que não estavam, geramos outro grupo de dados incluindo

```

==== Run information ====
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: questionario-weka.filters.unsupervised.attribute.Remove-R6-weka.filters.unsupervised.attribute.Remove-R4
Instances: 924
Attributes: 11
ano
serie
sabe_banco_dados
sabe_programar
q4_sup_pouca_mat
q7_prec_curs_sup_p_trab
q10_trab_pouco_lazer
q11_trab_comp_criatividade
q12_trab_comp_prestigio
q13_trab_comp_ganha_bem
q14_trab_comp_atuar_outras_areas
Test mode:10-fold cross-validation
==== Classifier model (full training set) ====
J48 pruned tree
-----
sabe_banco_dados = S
| q11_trab_comp_criatividade = S
| | q14_trab_comp_atuar_outras_areas = S: S (145.0/49.0)
| | q14_trab_comp_atuar_outras_areas = N: N (4.0/1.0)
| | q14_trab_comp_atuar_outras_areas = T
| | | serie = Fundamental: S (13.0/4.0)
| | | serie = 1 ano
| | | | q7_prec_curs_sup_p_trab = S: N (2.0)
| | | | q7_prec_curs_sup_p_trab = N: N (0.0)
| | | | q7_prec_curs_sup_p_trab = T: S (3.0/1.0)
| | | serie = 2 anos: N (3.0)
| | | serie = 3 anos: S (5.0/2.0)
| | | serie = Supletivo: S (0.0)
| | q11_trab_comp_criatividade = N
| | | q7_prec_curs_sup_p_trab = S: N (3.0)
| | | q7_prec_curs_sup_p_trab = N: S (6.0/1.0)
| | | q7_prec_curs_sup_p_trab = T: S (1.0)
| | q11_trab_comp_criatividade = T
| | | ano = 2011
| | | | q12_trab_comp_prestigio = S: S (5.0/2.0)
| | | | q12_trab_comp_prestigio = N: S (2.0)
| | | | q12_trab_comp_prestigio = T: N (5.0)
| | | ano = 2012: N (7.0)
| | | ano = 2013: S (2.0)
sabe_banco_dados = N: N (718.0/186.0)

Number of Leaves : 18
Size of the tree : 26

```

Figura 3.9: Parte 1 da saída da árvore de decisões.

```

Time taken to build model: 0 seconds
==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances      653          70.671 %
Incorrectly Classified Instances   271          29.329 %
Kappa statistic                      0.2561
Mean absolute error                  0.4036
Root mean squared error              0.4625
Relative absolute error              90.493 %
Root relative squared error         97.9556 %
Total Number of Instances           924

==== Detailed Accuracy By Class ====
      TP Rate    FP Rate    Precision    Recall    F-Measure    ROC Area    Class
          0.329     0.103     0.618     0.329     0.429     0.575     S
          0.897     0.671     0.726     0.897     0.803     0.575     N
Weighted Avg.      0.707     0.48      0.69      0.707     0.677     0.575

==== Confusion Matrix ====
      a     b  <-- classified as
102   208 |   a = S
       63   551 |   b = N

```

Figura 3.10: Parte 2 da saída da árvore de decisões.

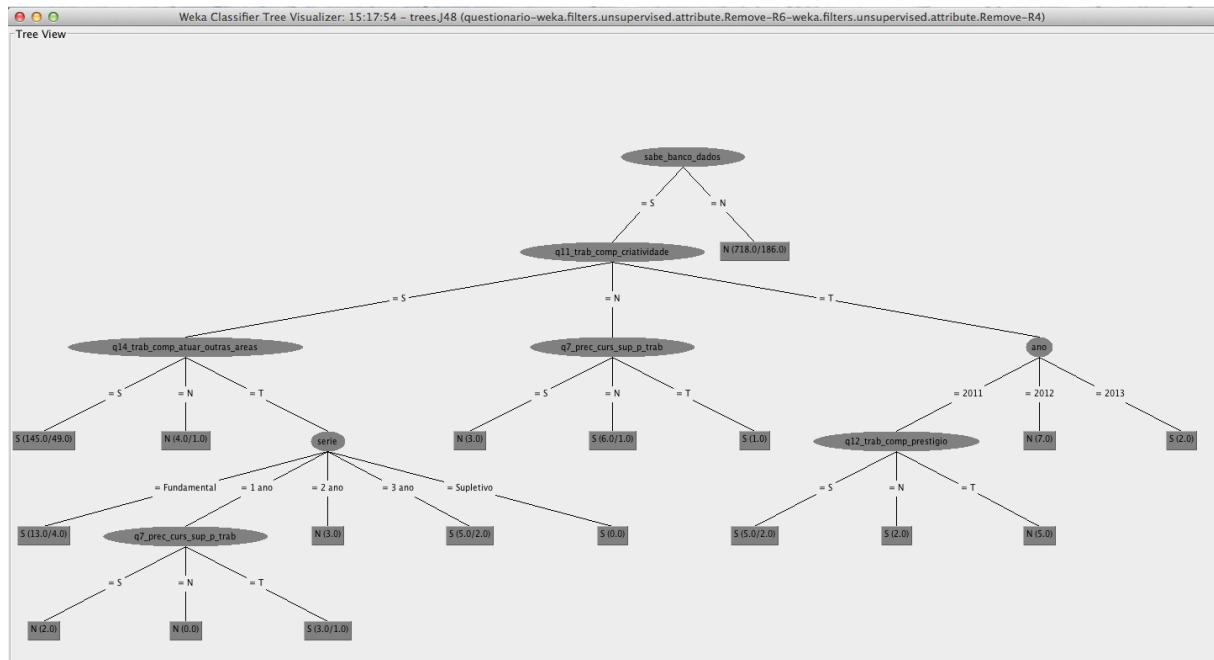


Figura 3.11: Árvore de decisões.

os dois perfis e adicionando o atributo *pensando* como diferenciador entre eles. Em seguida, utilizando o atributo *pensando* como classificador, foi executado o algoritmo J48 nos novos dados. Apesar de possuir apenas 43% das instâncias classificadas corretamente, foi possível realizar algumas análises nesses dados. A Figura 3.12 mostra um trecho da árvore de saída gerada.

```

== Run information ==
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: questionario-weka.filters.unsupervised.attribute.Remove-R5,7-weka.filters.unsupervised.attribute.Remove-R7
Instances: 2539
Attributes: 11
ano
serie
pensando
sabe_banco_dados
sabe_programar
q4_sup_pouca_mat
q10_trab_pouco_lazer
q11_trab_comp_criatividade
q12_trab_comp_prestigio
q13_trab_comp_ganha_bem
q14_trab_comp_atuar_outras_areas
Test mode:10-fold cross-validation
== Classifier model (full training set) ==
J48 pruned tree
-----
serie = Fundamental
| sabe_banco_dados = S
| q11_trab_comp_criatividade = S
| ano = 2011: S (53.0/12.0)
| ano = 2012
|   sabe_programar = S
|     q4_sup_pouca_mat = S: NS (3.0/1.0)
|     q4_sup_pouca_mat = N: S (13.0/2.0)
|     q4_sup_pouca_mat = T: NS (6.0)
|   sabe_programar = N
|     q12_trab_comp_prestigio = S
|       q14_trab_comp_atuar_outras_areas = S
|         q13_trab_comp_ganha_bem = S: S (3.0/1.0)
|         q13_trab_comp_ganha_bem = N: NS (0.0)
|         q13_trab_comp_ganha_bem = T: NS (7.0/3.0)
|       q14_trab_comp_atuar_outras_areas = N: N (1.0)
|       q14_trab_comp_atuar_outras_areas = T: N (2.0)
|     q12_trab_comp_prestigio = N: S (1.0)
|     q12_trab_comp_prestigio = T: NS (3.0)
|   ano = 2013
|     q10_trab_pouco_lazer = S: N (1.0)
|     q10_trab_pouco_lazer = N
|       q14_trab_comp_atuar_outras_areas = S: NS (6.0/2.0)
|       q14_trab_comp_atuar_outras_areas = N: NS (0.0)
|       q14_trab_comp_atuar_outras_areas = T: S (2.0)
|     q10_trab_pouco_lazer = T: S (4.0)
|   q11_trab_comp_criatividade = N: NS (5.0/2.0)
|   q11_trab_comp_criatividade = T: S (11.0/1.0)
sabe_banco_dados = N
q13_trab_comp_ganha_bem = S
ano = 2011: S (155.0/64.0)
ano = 2012
| q4_sup_pouca_mat = S
|   sabe_programar = S: S (7.0/2.0)
|   sabe_programar = N: NS (13.0/5.0)
| q4_sup_pouca_mat = N
|   sabe_programar = S: NS (20.0/8.0)
|   sabe_programar = N: S (41.0/14.0)
| q4_sup_pouca_mat = T: NS (31.0/12.0)
ano = 2013
| q4_sup_pouca_mat = S
|   q14_trab_comp_atuar_outras_areas = S: NS (6.0/3.0)
|   q14_trab_comp_atuar_outras_areas = N: S (0.0)
|   q14_trab_comp_atuar_outras_areas = T: S (2.0)
| q4_sup_pouca_mat = N
|   q10_trab_pouco_lazer = S: N (2.0)
|   q10_trab_pouco_lazer = N
|     sabe_programar = S: S (4.0/1.0)
|     sabe_programar = N: NS (10.0/5.0)
|   q10_trab_pouco_lazer = T: NS (6.0/3.0)
| q4_sup_pouca_mat = T
|   q14_trab_comp_atuar_outras_areas = S: S (7.0/1.0)
|   q14_trab_comp_atuar_outras_areas = N: S (0.0)
|   q14_trab_comp_atuar_outras_areas = T: NS (4.0/1.0)
q13_trab_comp_ganha_bem = N
sabe_programar = S: NS (5.0)
sabe_programar = N
| ano = 2011
|   q12_trab_comp_prestigio = S: S (13.0/5.0)
|   q12_trab_comp_prestigio = N
|     q4_sup_pouca_mat = S: NS (2.0)
|     q4_sup_pouca_mat = N: S (2.0/1.0)
|     q4_sup_pouca_mat = T: NS (0.0)
|   q12_trab_comp_prestigio = T: NS (5.0/2.0)
ano = 2012: NS (5.0/1.0)
ano = 2013
| q4_sup_pouca_mat = S: NS (2.0)
| q4_sup_pouca_mat = N: N (5.0)
| q4_sup_pouca_mat = T: N (0.0)

```

Figura 3.12: Parte 1 da saída da árvore de decisões com os dois perfis.

3.4 Resultados da Mineração de Dados

A partir da análise inicial, na Figura 3.7, é possível perceber no terceiro gráfico (sabe_banco_dados) que enquanto 206 respondentes, ou 22% do total, já utilizaram banco de dados, 718 pessoas, ou 78% do total, nunca realizou esse tipo de atividade.

Já no segundo gráfico (serie), comparando com os dados de série escolar dos respondentes, onde visualizamos respectivamente os dados de ensino fundamental, primeiro, segundo e terceiro ano, e por último supletivo. Nesse caso fica claro visualmente que, sem contar com o Supletivo, a proporção de pessoas que já utilizaram banco de dados aumenta de acordo com o nível escolar, principalmente no terceiro ano. Com isso, foi possível fazer uma nova busca no banco de dados, e trazer dados mais precisos, onde temos uma proporção menor do que a geral para o ensino Fundamental e primeiro ano, de 20% e 21% respectivamente, porém no segundo e terceiro ano, temos 23% e 34% respectivamente.

Além disso, podemos visualizar no quinto gráfico (sabe_programar), que apesar de menos pessoas saberem programação, existe uma maior quantidade de respondentes que sabem bancos de dados e programação do que quem não sabe nenhum dos dois. Apesar de ficar claro no gráfico, existe uma confirmação ao buscar dados precisos no banco de dados. Dos 310 respondentes que sabem programar, 124 (40%) também sabem banco de dados. E daqueles 614 que afirmam não saber programar, apenas 82 (13%) afirmam ter conhecimento sobre bancos de dados.

Ao analisarmos a árvore da Figura 3.9, confirmamos o que encontramos na análise inicial e destacamos os seguintes pontos:

- Foi alcançada uma predição de 718 instâncias que são classificadas nos atributos “não saber banco de dados” e “não saber programação”, apesar da classificação incorreta de 186 delas.
- Foi alcançada a predição de 145 instâncias que são classificadas nos atributos “saber banco de dados” e responderam sim na questão onze (Computação utiliza criatividade) e na quatorze (Computação é interdisciplinar). Essas 145 instâncias foram classificadas corretamente no atributo de “saber programação”, apesar de 49 outras instâncias terem sido incorretamente classificadas.

A partir desses pontos, os dados sugerem que no universo abordado nos questionários, quem não sabe banco de dados, também não sabe programar. Além de sugerir também que quem sabe banco de dados, reconhece a interdisciplinaridade e o uso da criatividade na área da computação, também sabem programar.

Analisando a árvore gerada utilizando como classificadores os perfis “pensando em fazer computação” e “não pensando em fazer computação”, representada na figura de 3.12 destacamos os seguintes pontos, divididos pela série cursada pelas alunas no momento em que responderam os questionários.

Alunas do ensino fundamental:

- Que acreditam que trabalhar na área da computação oferece uma boa remuneração, pensam em prestar vestibular para a área da computação. 219 questionários (22%) caem neste caso. A predição alcançada neste atributo foi de 155 (71%) corretas e 64 (29%) erradas.

- Que sabem programar, acreditam no prestígio da profissão e que se faz necessário o uso da criatividade, pensam prestar o vestibular para área da computação. 72 questionários (7%) caem neste caso. A predição alcançada neste atributo foi de 51 (71%) corretas e 21 (29%) erradas.
- Que sabem programar e acham que o curso requer muita matemática, pensam em cursar a área da computação. 15 questionários (2%) caem neste caso. A predição alcançada foi de 13 (87%) corretas e 2 (13%) erradas.

Alunas do primeiro ano do ensino médio:

- Que sabem programar e acreditam no prestígio na área, pensam em prestar vestibular para área da computação. 176 questionários (24%) caem neste caso. A predição alcançada foi de 124 (70%) corretas e 52 (30%) erradas.
- Alunas do primeiro ano que não sabem programar, acreditam que a área é interdisciplinar e sabem que o curso utiliza muita matemática, pensam em prestar vestibular para a área. 94 questionários (13%) caem neste caso. A predição alcançada foi de 65 corretas (69%) e 29 incorretas (31%).

Alunas do segundo ano do ensino médio:

- Que sabem banco de dados e acreditam no prestígio da profissão querem fazer computação. 61 questionários (10%) caem neste caso. A predição alcançada foi de 40 corretas (65%) e 21 incorretas (35%).

Alunas do terceiro ano do ensino médio:

- Que sabem programar e sabem banco de dados, além de reconhecer o prestígio da profissão e a boa remuneração, pensam em seguir a área da computação. 37 questionários (7%) caem neste caso. A predição alcançada foi de 26 corretas (70%) e 11 incorretas (30%).
- Que não sabem banco de dados e não acreditam na interdisciplinaridade, não querem seguir a área da computação. 46 questionários caem neste caso. A predição alcançada foi de 30 corretas (65%) e 16 incorretas (35%).

A partir da análise desses pontos, que foram os que apresentaram maior destaque e relevância na predição. Os dados indicam que saber programar desde o ensino fundamental é de grande importância para as alunas seguirem na área da computação. O fator de saber programar juntamente com saber banco de dados, também conseguem manter o interesse na computação pelas alunas ao longo das séries do ensino médio.

A divulgação da informação sobre a área de atuação, prestígio da profissão e a boa remuneração, podem influenciar na escolha em qual área prestar vestibular. Meninas quem possuem um conhecimento maior sobre o mercado na área de computação, tendem a preferir esta área no vestibular.

Um fato importante que deve ser destacado também, é o fato de que meninas que apreciam a Matemática, possuem tendências a escolherem a área da computação como profissão.

Capítulo 4

Estudo de Caso 2: Cursos de Graduação de Computação da UnB

4.1 Introdução e Análise Estatística

Após a análise dos resultados obtidos no Capítulo 3, Estudo de Caso 1, e realizar a interpretação da percepção que as mulheres do ensino fundamental e médio possuem sobre a ciência da computação , como previsto no início deste trabalho, expandimos a pesquisa a fim de investigar as diferenças de gênero nos curso da área da computação da Universidade de Brasília, ou seja, investigamos os cursos de Ciência da Computação, Licenciatura em Computação e Engenharia da Computação.

Para esta investigação obtivemos os dados relativos aos alunos e alunas destes cursos a partir do ano de 1983, ressaltando que primeiramente foi criado o curso de Ciência da Computação nesta data, logo após em 1997 foi criado o curso de Licenciatura em Computação e somente em 2008 foi criado o curso de Engenharia da Computação, consequentemente os dados referente a cada curso começam a ser analisados cronologicamente a partir da data de sua criação. Estes dados foram extraídos do Sistema de Informação Acadêmica de Graduação – SIGRA da unb assim como os gráficos inseridos nesta seção.

Antes de iniciarmos o processo de mineração de dados para investigar a diferença de gênero nos cursos de computação da UnB, é interessante destacarmos algumas informações estatísticas que em uma primeira análise já evidenciam alguma diferença de gênero nesses cursos, como: Número de ingressantes de acordo com o gênero em cada curso, Número de formandos e desligados por motivos adversos, Comparação do Índice de Rendimento Acadêmico – IRA. Em todos os gráficos ingressantes por ano e sexo (Figuras 4.1, 4.5 e 4.9) há uma queda nos resultados pelo fato do ano de 2014 estar incompleto.

Na Figura 4.1 podemos visualizar o gráfico do número de ingressantes por ano e sexo no curso de Ciência da Computação a partir do ano de sua criação, assim conseguimos visualizar facilmente que o número de alunos do gênero masculino sempre foi maior que o número de alunas, porém, seguindo a mesma linha de análise, na Figura 4.2 visualizamos o gráfico com o número de formandos por ano e sexo no mesmo curso, e nele observamos que nos primeiros anos a quantidade de formandos e formandas era mais próximo, destacando que no ano de 1992 formaram mais mulheres que homens, já na Figura 4.3 visualizamos o gráfico por ano e sexo dos alunos desligados.

Além dos gráficos de linha, mostramos também nas Figuras 4.4, 4.8 e 4.11 os gráficos de pizza, onde visualizamos o motivo do desligamento, podendo analisar em qual curso o desligamento por não cumprir o estado de condição tem maior incidências por exemplo.

Observamos que os motivos do desligamento entre os alunos e alunas do mesmo curso mantém praticamente o mesmo padrão, porém é importante destacar que no curso de engenharia os alunos do gênero feminino não foram desligados por abandono, já 11% dos alunos masculinos foram desligados por este motivo, e no curso de Ciência da Computação destacamos que os alunos do gênero feminino não foram desligados por reprovarem três vezes na mesma matéria, mas 6% dos alunos masculinos foram desligados por esse motivo.

Expandido a comparação dos motivos de desligamento entre os cursos, destacamos os seguintes pontos que podem servir de base para uma investigação mais detalhada na fase de mineração de dados e talvez sejam relacionados a outros atributos de cada curso.

- O desligamento por não cumprir condição é maior no curso de Licenciatura, principalmente se compararmos os alunos do gênero feminino, onde a diferença é ainda maior.
- O desligamento voluntário é maior no curso de Ciência da Computação, chegando a ser o dobro em relação ao mesmo motivo nos outros cursos.
- O desligamento por abandono no curso de Engenharia é muito inferior se comparado aos outros cursos.
- O desligamento por novo vestibular no curso de Engenharia é muito superior se comparado aos outros cursos.
- O desligamento por transferência praticamente existe somente no curso de Ciencia da Computação.

Podemos observar os gráficos individuais para os cursos de Licenciatura e Engenharia nas Figuras 4.5, 4.6, 4.7, 4.9 e 4.10. Outra constatação importante que requer uma investigação profunda nas próximas seções é o fato de o número de desligamentos dos alunos do gênero masculino do curso de Engenharia da Computação está em pleno crescimento, não ocorrendo as baixas e altas observadas nos outros cursos.

Ingressantes por Ano e Sexo Ciência da Computação

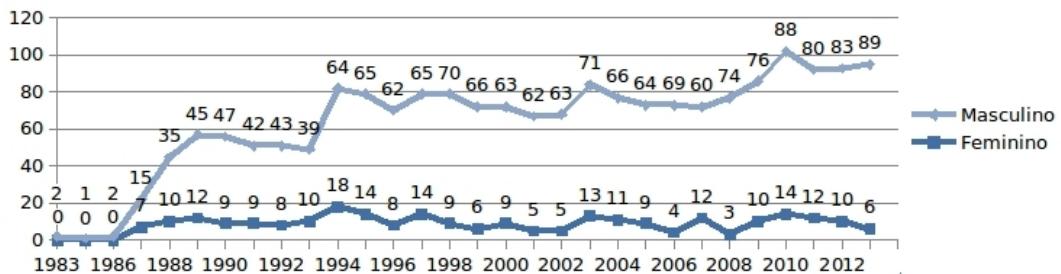


Figura 4.1: Ingressantes por ano e sexo no curso de Ciência da Computação

Formados por Ano e Sexo Ciência da Computação

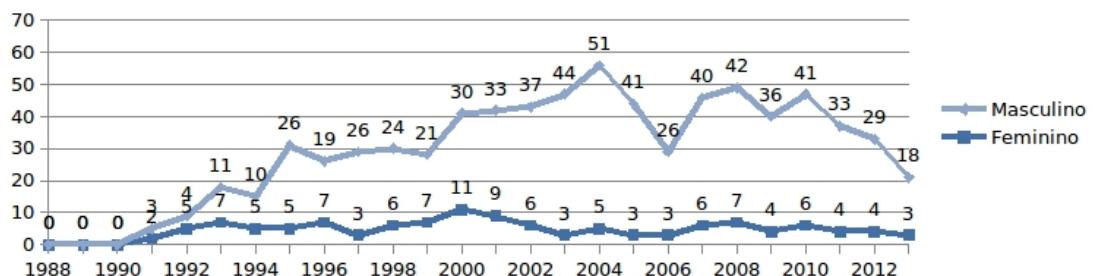


Figura 4.2: Formandos por ano e sexo no curso de Ciência da Computação

Desligados por Ano e Sexo Ciência da Computação

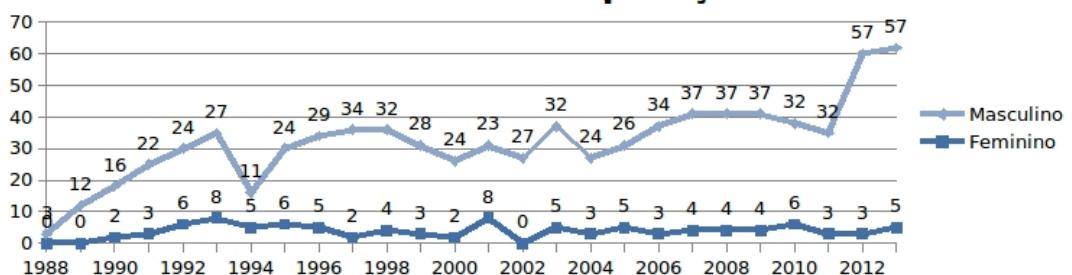
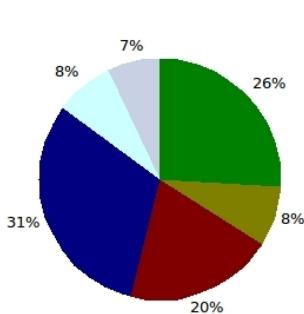


Figura 4.3: Desligados por ano e sexo no curso de Ciência da Computação

Desligamentos Ciência da Computação Feminino



Desligamentos Ciência da Computação Masculino



Figura 4.4: Motivo do desligamento no curso de Ciência da Computação

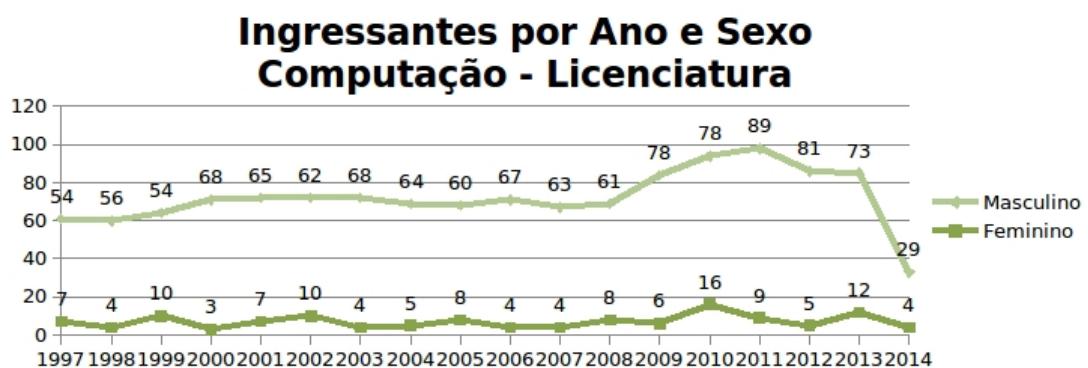


Figura 4.5: Ingressantes por ano e sexo no curso de Licenciatura em Computação

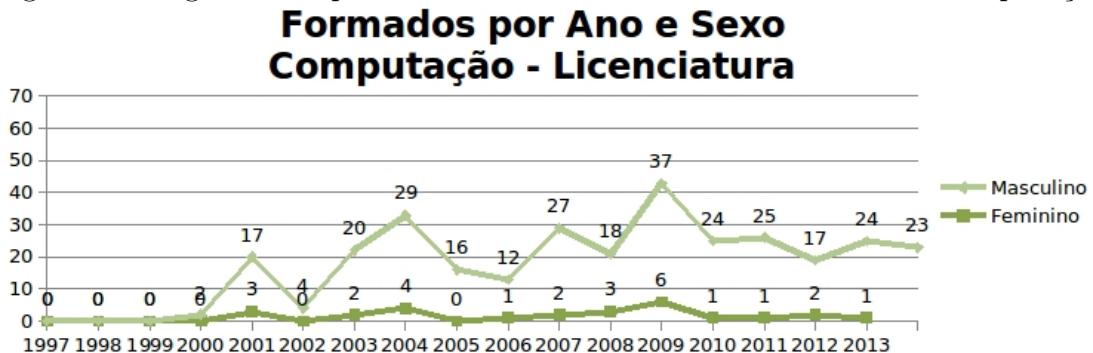


Figura 4.6: Ingressantes por ano e sexo no curso de Licenciatura em Computação

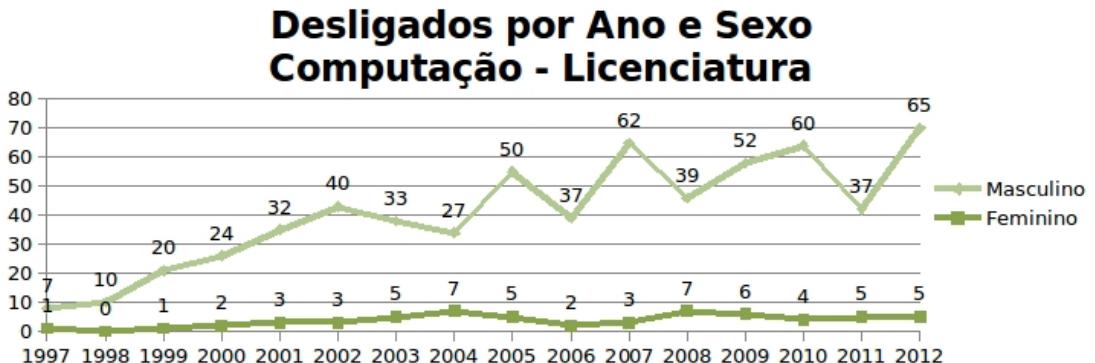


Figura 4.7: Desligados por ano e sexo no curso de Licenciatura em Computação



Figura 4.8: Motivo do desligamento no curso de Licenciatura em Computação

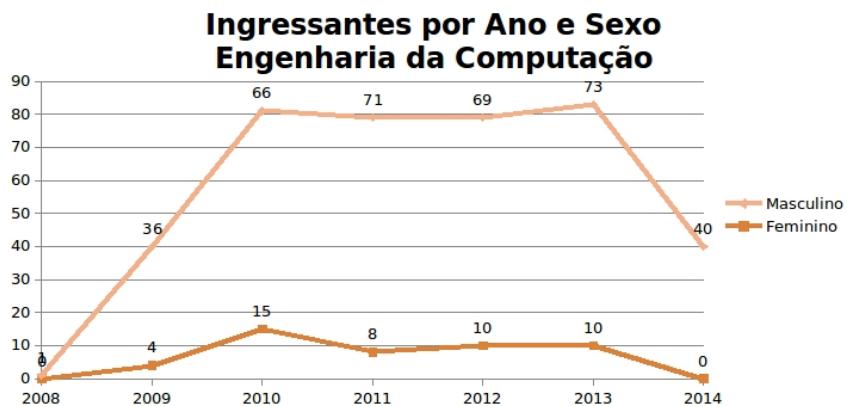


Figura 4.9: Ingressantes por ano e sexo no curso de Engenharia da Computação

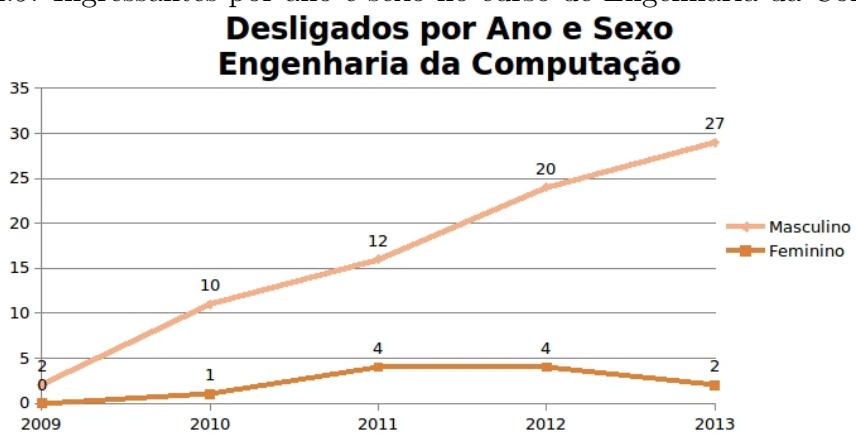


Figura 4.10: Desligados por ano e sexo no curso de Engenharia da Computação



Figura 4.11: Motivo do desligamento no curso de Engenharia da Computação

A fim de investigar mais profundamente esta relação e a diferença de gêneros nestes cursos e adentrar além da análise dos quantitativos de alunos ingressantes, formados e desligados, aplicamos a mineração de dados e apresentamos os resultados e análises nas seções que se seguem.

4.2 Formatação dos Dados

Foram entregues três diferentes tipos de planilha, um deles com dados dos alunos, outro com os dados de menções dos alunos desde 2000 até 2013 e o último com dados das disciplinas e seus departamentos. O primeiro tipo de planilha é composto de três abas: a primeira delas contém os dados em questão; a segunda possui um dicionário para os atributos, por exemplo, a coluna V2 significa Sexo do Aluno; e a terceira possui o dicionário dos valores dos dados, por exemplo, na coluna de Sexo do Aluno, existem duas possibilidades de valores, F ou M, e seus significados seriam Feminino ou Masculino, respectivamente. O segundo tipo de planilha foi um pouco similar ao primeiro, porém com dados referentes a menções dos alunos de cursos de computação e dividido em vários arquivos, sendo duas planilhas por ano letivo desde 2000 até 2013 referente a cada semestre. E o terceiro tipo de planilha possui colunas auto-explicativas, não necessitando de dicionário de dados, e dividido em duas abas, a primeira com dados do departamento e a segunda com dados das disciplinas relacionadas a estes departamentos. As Figuras 4.12, 4.13 e 4.14 mostram recortes dos tipos de planilhas recebidas.

+ Computacao feminino 30 jan 2014	Dicionário	Tabelas																																																																																																																																																																																
<table border="1"> <thead> <tr> <th>MatricAluno</th><th>V2</th><th>V3</th><th>V4</th><th>V5</th><th>V6</th><th>V7</th><th>V9</th><th>V10</th><th>V13</th><th>V14</th><th>V15</th><th>V17</th><th>V18</th><th>V19</th><th>V20</th></tr> </thead> <tbody> <tr><td>198700991</td><td>F</td><td>2</td><td>22</td><td>17-May-69</td><td>22</td><td>RJ</td><td>0</td><td>52455335100</td><td>1672170 - DF</td><td>0</td><td>8700991</td><td>0</td><td>0</td><td>RIO DE JANEIRO</td><td>12194</td></tr> <tr><td>198707448</td><td>F</td><td>2</td><td>22</td><td>12-Mar-69</td><td>22</td><td>SP</td><td>0</td><td>37661973120</td><td>1080909/SEP/DF</td><td>0</td><td>8707448</td><td>0</td><td>0</td><td>SAO PAULO</td><td>9947</td></tr> <tr><td>198711178</td><td>F</td><td>2</td><td>22</td><td>12-Oct-68</td><td>22</td><td>SP</td><td>0</td><td></td><td>0 768531 SSP DF</td><td>0</td><td>8711178</td><td>0</td><td>0</td><td></td><td>22485</td></tr> <tr><td>198711283</td><td>F</td><td>2</td><td>22</td><td>9-Aug-68</td><td>22</td><td>GO</td><td>0</td><td></td><td>0 2456045 SEP GO</td><td>0</td><td>8711283</td><td>0</td><td>0</td><td></td><td>22493</td></tr> <tr><td>198711291</td><td>F</td><td>2</td><td>22</td><td>2-Feb-71</td><td>22</td><td>RS</td><td>0</td><td></td><td>0 1054518 SSP DF</td><td>0</td><td>8711291</td><td>0</td><td>0</td><td></td><td>35403</td></tr> <tr><td>198735620</td><td>F</td><td>2</td><td>142</td><td>2-Jun-65</td><td>142</td><td></td><td>0</td><td></td><td>0 V120923-I RNE</td><td>0</td><td>8735620</td><td>0</td><td>0</td><td></td><td>23238</td></tr> <tr><td>198752541</td><td>F</td><td>2</td><td>22</td><td>28-Jul-66</td><td>22</td><td>DF</td><td>0</td><td></td><td>0 805184 SSP DF</td><td>0</td><td>8752541</td><td>0</td><td>0</td><td></td><td>23570</td></tr> <tr><td>198801274</td><td>F</td><td>2</td><td>22</td><td>12-Feb-70</td><td>22</td><td>RJ</td><td>0</td><td>52341615104</td><td>1130498</td><td>0</td><td>8801274</td><td>0</td><td>0</td><td>NITEROI</td><td>12165</td></tr> <tr><td>198801355</td><td>F</td><td>2</td><td>22</td><td>24-Dec-70</td><td>22</td><td>DF</td><td>0</td><td>52462153187</td><td>1133014/DF</td><td>0</td><td>8801355</td><td>0</td><td>0</td><td></td><td>75374</td></tr> <tr><td>198811156</td><td>F</td><td>2</td><td>22</td><td>10-Aug-71</td><td>22</td><td>PE</td><td>0</td><td></td><td>0 1182897 SSP DF</td><td>0</td><td>8811156</td><td>0</td><td>0</td><td></td><td>32577</td></tr> </tbody> </table>	MatricAluno	V2	V3	V4	V5	V6	V7	V9	V10	V13	V14	V15	V17	V18	V19	V20	198700991	F	2	22	17-May-69	22	RJ	0	52455335100	1672170 - DF	0	8700991	0	0	RIO DE JANEIRO	12194	198707448	F	2	22	12-Mar-69	22	SP	0	37661973120	1080909/SEP/DF	0	8707448	0	0	SAO PAULO	9947	198711178	F	2	22	12-Oct-68	22	SP	0		0 768531 SSP DF	0	8711178	0	0		22485	198711283	F	2	22	9-Aug-68	22	GO	0		0 2456045 SEP GO	0	8711283	0	0		22493	198711291	F	2	22	2-Feb-71	22	RS	0		0 1054518 SSP DF	0	8711291	0	0		35403	198735620	F	2	142	2-Jun-65	142		0		0 V120923-I RNE	0	8735620	0	0		23238	198752541	F	2	22	28-Jul-66	22	DF	0		0 805184 SSP DF	0	8752541	0	0		23570	198801274	F	2	22	12-Feb-70	22	RJ	0	52341615104	1130498	0	8801274	0	0	NITEROI	12165	198801355	F	2	22	24-Dec-70	22	DF	0	52462153187	1133014/DF	0	8801355	0	0		75374	198811156	F	2	22	10-Aug-71	22	PE	0		0 1182897 SSP DF	0	8811156	0	0		32577		
MatricAluno	V2	V3	V4	V5	V6	V7	V9	V10	V13	V14	V15	V17	V18	V19	V20																																																																																																																																																																			
198700991	F	2	22	17-May-69	22	RJ	0	52455335100	1672170 - DF	0	8700991	0	0	RIO DE JANEIRO	12194																																																																																																																																																																			
198707448	F	2	22	12-Mar-69	22	SP	0	37661973120	1080909/SEP/DF	0	8707448	0	0	SAO PAULO	9947																																																																																																																																																																			
198711178	F	2	22	12-Oct-68	22	SP	0		0 768531 SSP DF	0	8711178	0	0		22485																																																																																																																																																																			
198711283	F	2	22	9-Aug-68	22	GO	0		0 2456045 SEP GO	0	8711283	0	0		22493																																																																																																																																																																			
198711291	F	2	22	2-Feb-71	22	RS	0		0 1054518 SSP DF	0	8711291	0	0		35403																																																																																																																																																																			
198735620	F	2	142	2-Jun-65	142		0		0 V120923-I RNE	0	8735620	0	0		23238																																																																																																																																																																			
198752541	F	2	22	28-Jul-66	22	DF	0		0 805184 SSP DF	0	8752541	0	0		23570																																																																																																																																																																			
198801274	F	2	22	12-Feb-70	22	RJ	0	52341615104	1130498	0	8801274	0	0	NITEROI	12165																																																																																																																																																																			
198801355	F	2	22	24-Dec-70	22	DF	0	52462153187	1133014/DF	0	8801355	0	0		75374																																																																																																																																																																			
198811156	F	2	22	10-Aug-71	22	PE	0		0 1182897 SSP DF	0	8811156	0	0		32577																																																																																																																																																																			

Figura 4.12: Recorte da planilha com os dados dos alunos de cursos de Computação da UnB.

MatricAluno	V2_h	V3_h	CodDisc	V5_h	V6_h	V7_h	V8_h	V2
199203265	2000	1	116475	B	6	SS	5	M
199203265	2000	1	130010		1	CC	0	M
199318127	2000	1	116530	A	4	MM	0	M
199318127	2000	1	117200	A	4	MM	10	M
199318127	2000	1	116360	A	4	MM	4	M
199318127	2000	1	116581	A	4	MS	20	M
199318127	2000	1	175307	B	2	MS	17	M
199331565	2000	1	113051	A	6	MM	15	M
199403124	2000	1	116530	A	4	MM	0	F
199403124	2000	1	185035	K	4	SS	0	F
199403124	2000	1	116629	E	4	SS	0	F
199403124	2000	1	130010		1	CC	0	F
199403124	2000	1	181021	C	4	MS	16	F

Figura 4.13: Recorte da planilha com os dados das menções dos alunos de cursos de Computação da UnB.

	DEPARTAMENTO	DISCIPLINA
COD_DISCIPLINA	COD_DEPARTAMENTO	NOM_DISCIPLINA
186091	40	ADMINISTRAÇÃO DE SERVIÇOS NO SETOR PÚBLICO
180203	40	ADMINISTRAÇÃO FINANCEIRA E ORÇAMENTÁRIA
186112	40	ADMINISTRACAO PUBLICA COMPARADA
113107	113	ALGEBRA 1
117145	113	ALGEBRA 3
113123	113	ALGEBRA LINEAR
113611	113	ALGEBRA PARA ENSINO 1 E 2
113204	113	ANALISE 1
113212	113	ANALISE 2
115894	115	ANALISE DE DADOS CATEGORIZADOS
115177	115	ANALISE MULTIVARIADA 1
115185	115	ANALISE MULTIVARIADA 2
115231	115	APLICACOES DA ESTATISTICA 1

Figura 4.14: Recorte da planilha com os dados das disciplinas e departamentos existentes na UnB.

Assim como na primeiro Estudo de Caso (seção 3.2), foi utilizado o SGBD (Sistema Gerencial de Bancos de Dados) MySQL e o aplicativo Sequel Pro para a manipulação desses dados. Então, foi utilizado o mesmo procedimento utilizado no capítulo anterior para importação dos dados das planilhas para o banco de dados: conversão para CSV e mapeamento das colunas com atributos. Porém, não precisou ser criada nenhuma coluna adicional na geração das tabelas do banco de dados.

Após o mapeamento, foi necessário realizar a tradução dos dados. Os primeiros atributos convertidos foram aqueles que só possuíam uma possível tradução pelo fato dos universo apenas incluir estas opções, sendo eles: nível, nível da opção, nível do curso, duração, duração do curso, código da faculdade, código do departamento e a forma do curso. Os três primeiros atributos são todos referentes ao nível do curso que o aluno estava ou está cursando, por exemplo, Graduação ou Pós-graduação. Os campos relacionados a duração são auto-explicativos, pois mostram a duração do curso, essa pode ser normal (plena), ou pode ser uma duração mais curta. Apesar dos códigos da faculdade e do departamento também serem auto-explicativos, a forma do curso não é. Neste último atributo existe a classificação do curso como presencial ou a distância.

Os próximos atributos a serem convertidos foram aqueles que possuíam apenas duas opções, sendo eles: pessoa com deficiência, prioridade da opção, aluno é registrado, código do grau, turno do curso e tipo de escola. Os campos de pessoa com deficiência e se o aluno é registrado só possuem sim ou não como resposta. No caso do campo de prioridade da opção, os possíveis valores são principal ou secundária, significando a prioridade do curso escolhida na hora de prestar o vestibular. E por último os campos código do grau, turno do curso e tipo de escola contém a informação se o curso é Bacharelado ou Licenciatura, se o curso é Diurno ou Noturno e se o aluno é advindo de escola Pública ou Privada, respectivamente.

Por existir um campo relacionado a nacionalidade do aluno e outro ao seu país de nascimento, foi necessária realizar a tradução de ambos os campos além da limpeza daqueles que não possuíam essa informação.

No caso da data de nascimento, o formato advindo da planilha foi 07-Aug-89 (07 de Agosto de 1989), portanto foi necessário converter para o formato de data do banco de dados (1989-08-07) para poder realizar qualquer tipo de classificação com esses dados posteriormente.

Além dos atributos listados anteriormente, também temos o campo de cotas, no qual diz em qual grupo de cotas aquele aluno se encaixa, raça, forma de ingresso, forma de ingresso na opção, forma de saída, forma de saída da opção e classificação por idade.

No caso do terceiro tipo de planilha não foi necessário fazer nenhum tipo de limpeza nos dados inseridos, portanto após a limpeza das planilhas anteriores foi realizada apenas a seleção dos dados. Para isso, foi necessário fazer o casamento das quatro tabelas criadas (alunos, menções, cursos, departamentos).

4.3 Análise e Resultados

Conforme pode ser visualizado na *Listing C.6*, localizada no Apêndice C, foi adicionado um campo chamado nota, onde foi colocado a menor nota possível daquela menção para serem realizados cálculos matemáticos acerca deles, substituindo a menção SS pelo número 9, MS por 7, MM por 5, MI por 3, II por 1 e todos os outros por 0. A partir disso, foi realizada uma análise, agrupando os dados por sexo e curso, e calculando a média dessas notas. Ao final, foram gerados dois gráficos, um com a média geral dos cursos, e outro com a média ao longo do tempo. As Figuras 4.15 e 4.16 mostram esses gráficos.

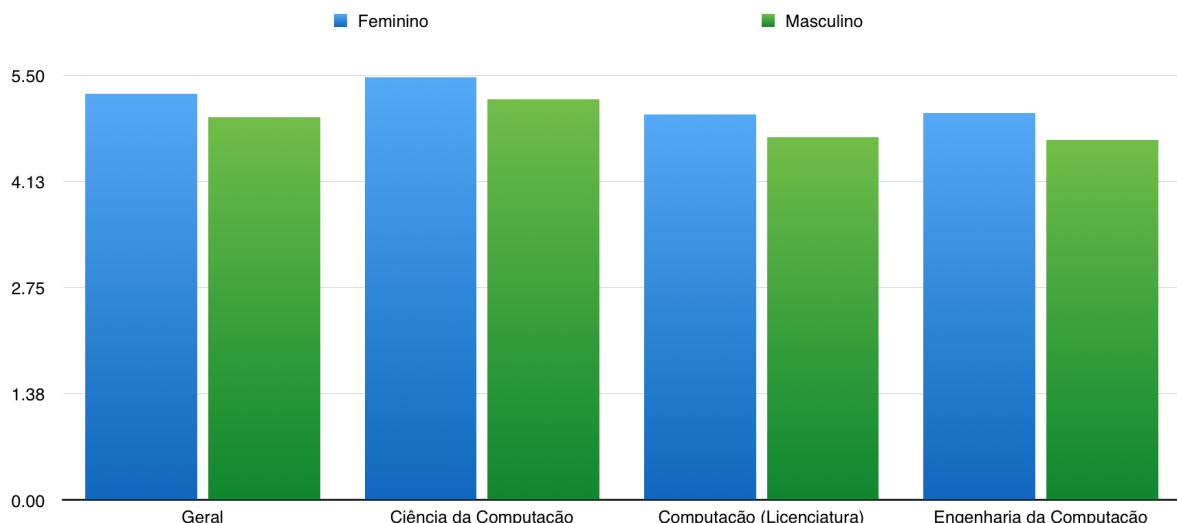


Figura 4.15: Gráfico com o comparativo de média de notas por gênero e curso.

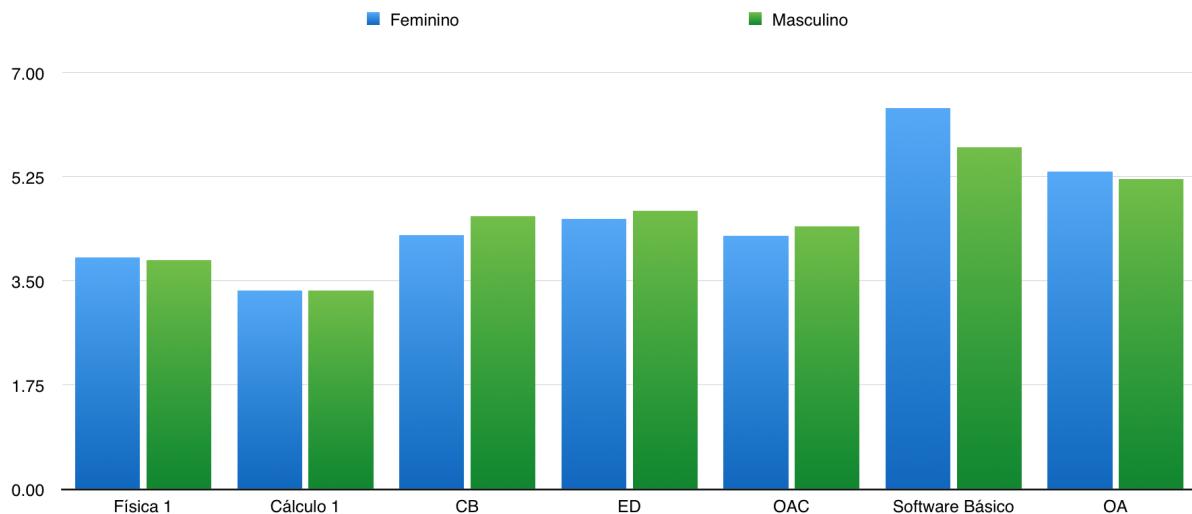


Figura 4.17: Gráfico com o comparativo de média de notas por gênero em disciplinas.

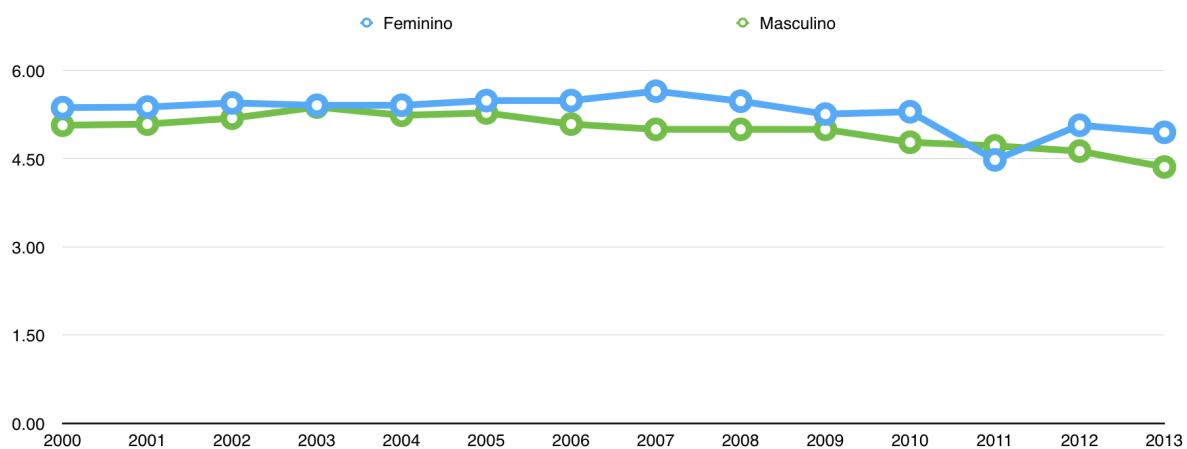


Figura 4.16: Gráfico com o comparativo de média de notas por gênero ao longo do tempo.

Foram também selecionadas as disciplinas em que os alunos possuem mais dificuldades (Física 1, Cálculo 1, Computação Básica, Estrutura de Dados, Organização e Arquitetura de Computadores, Software Básico e Organização de Arquivos) e feito um gráfico comparativo por gênero, conforme a Figura 4.17.

Com essa informação, foi possível gerar um gráfico comparativo entre os cursos com as disciplinas de menor nota, no caso Cálculo 1, Física 1 e Computação Básica. As Figuras 4.18, 4.19 e 4.20, respectivamente, mostram essa comparação.

Por último, subtraindo ano e semestre de saída por seu ano e semestre de entrada, foi descoberto o tempo de permanência, em semestres, dos alunos que saíram da UnB por qualquer forma diferente de formatura. Com isso, pode ser identificado a quantidade de períodos que os alunos normalmente deixam o curso, seja de forma voluntária ou forçada. As Figuras 4.21, 4.22 e 4.23 mostram gráficos com essa informação.

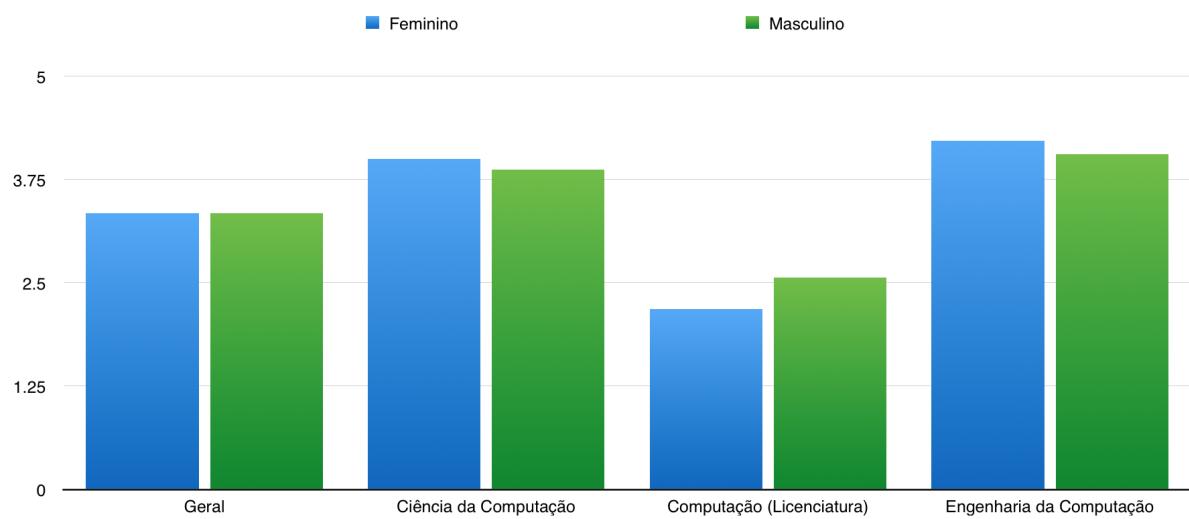


Figura 4.18: Gráfico com o comparativo de média de notas de Cálculo 1 por gênero em cada curso.

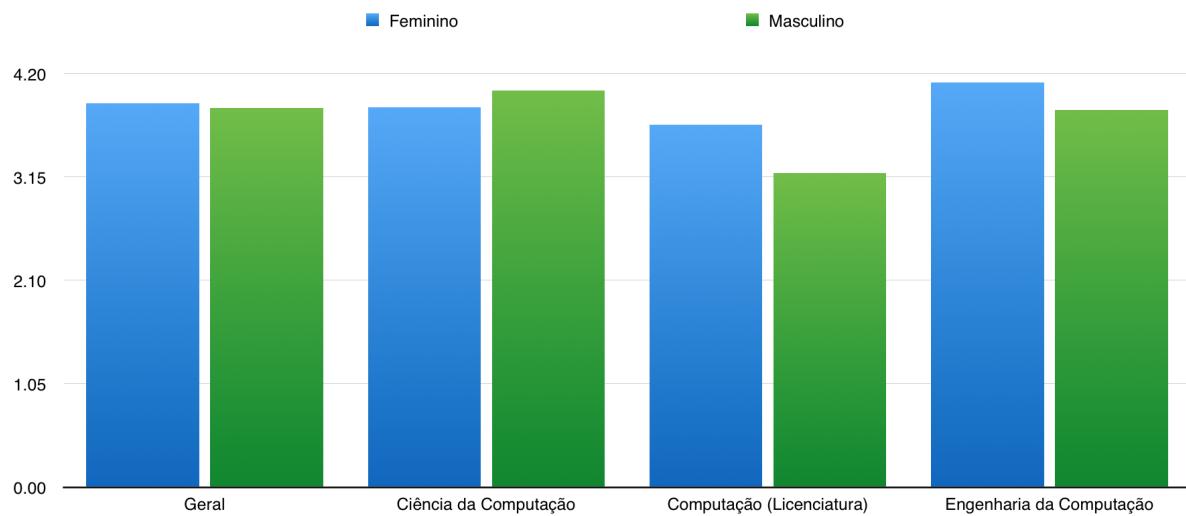


Figura 4.19: Gráfico com o comparativo de média de notas de Física 1 por gênero em cada curso.

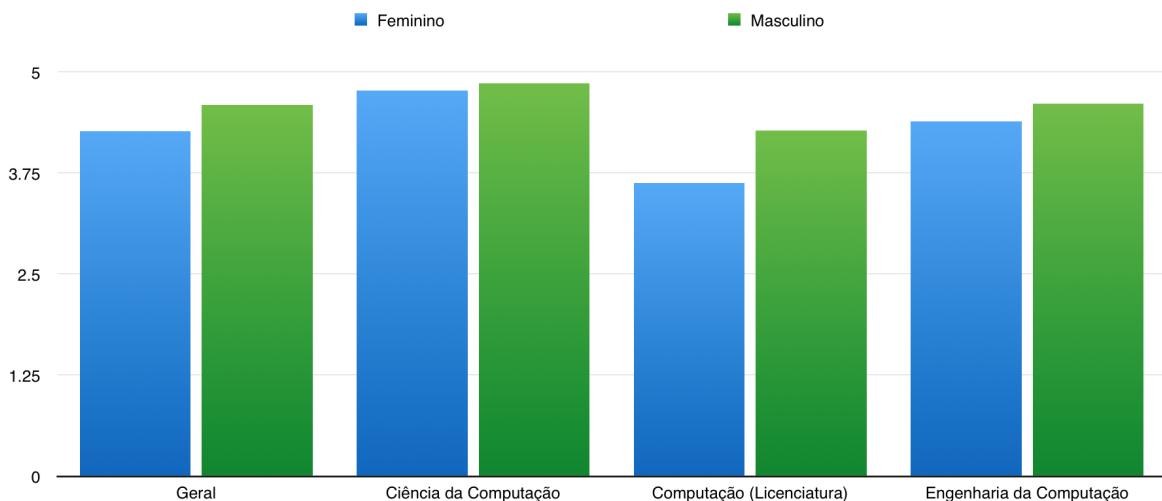


Figura 4.20: Gráfico com o comparativo de média de notas de Computação Básica por gênero em cada curso.

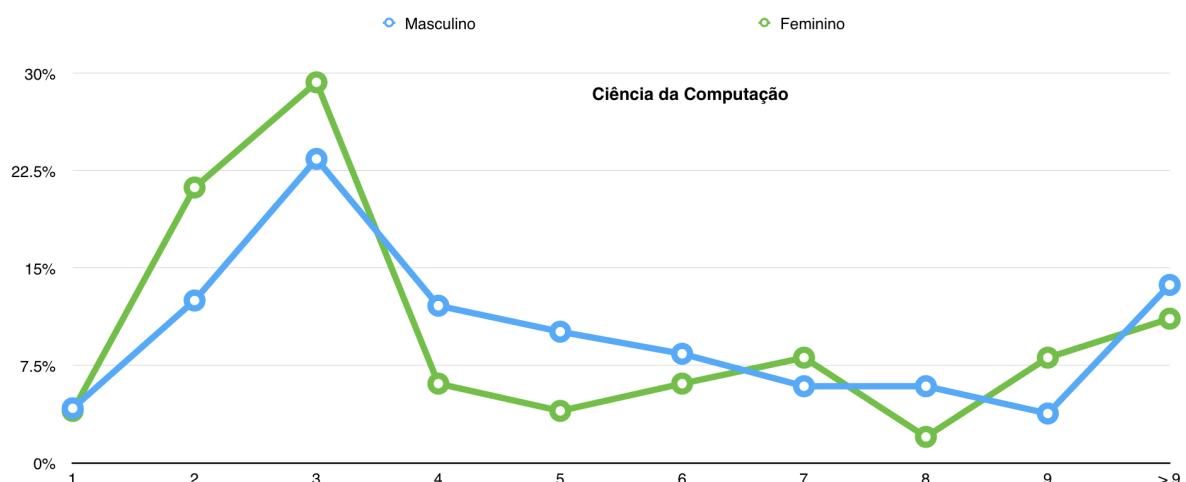


Figura 4.21: Gráfico com a porcentagem de desistência por quantidade de semestres em Ciência da Computação.

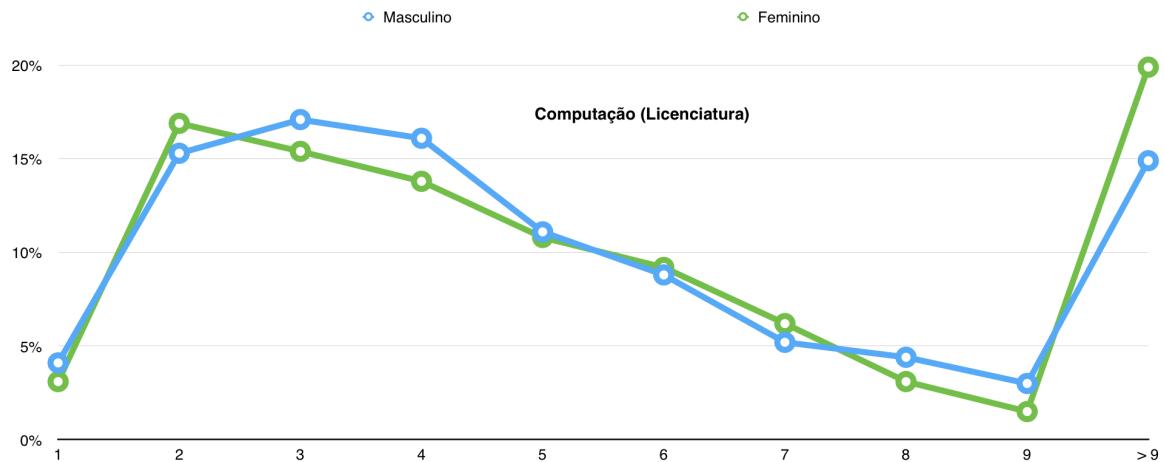


Figura 4.22: Gráfico com a porcentagem de desistência por quantidade de semestres em Computação (Licenciatura).

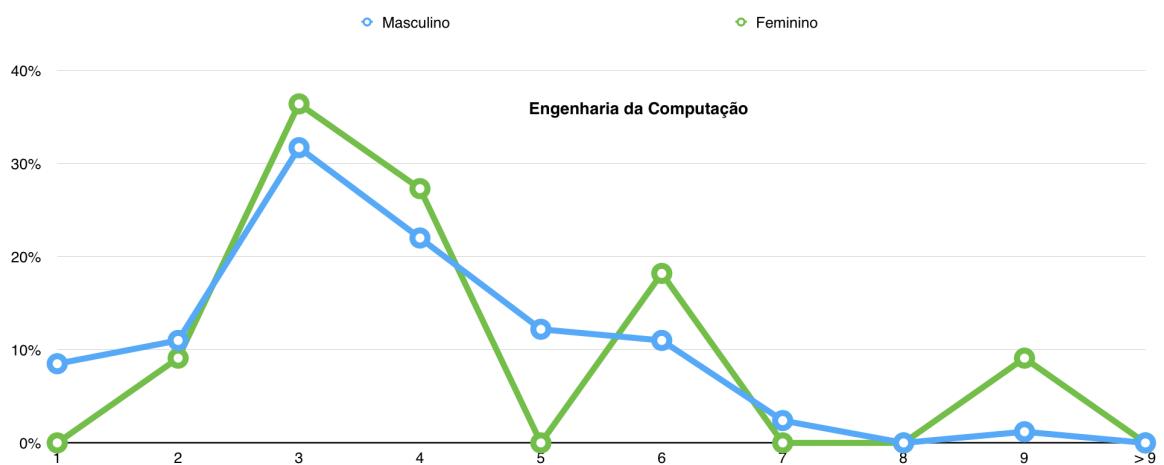


Figura 4.23: Gráfico com a porcentagem de desistência por quantidade de semestres em Engenharia da Computação.

Após essa análise inicial, foi realizado o procedimento de mineração de dados com os dados do alunos, fazendo a conversão de CSV para ARFF, para carregar os dados no programa Weka utilizando o algoritmo Bayesiano BayesNet. Segundo Witten e Frank (2005), a regra de Bayes calcula a probabilidade de um evento A dado um evento B, que seria a probabilidade da interseção entre os dois eventos dividido pela probabilidade do evento B. Portanto, utilizando o atributo sexo como classe, ou evento B, chegamos a rede bayesiana da Figura 4.24.

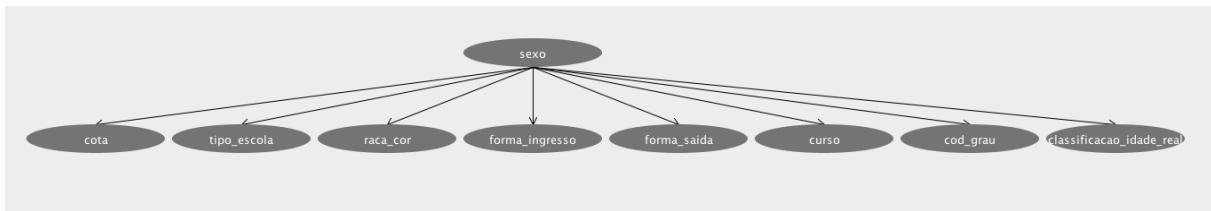


Figura 4.24: Rede bayesiana gerada a partir da classificação do sexo dos alunos.

A partir da figura com o gráfico 4.15, pode ser visualizado que as alunas de cursos da computação, independentemente de curso, possuem a média de notas ligeiramente maiores que a dos homens. Além disso, há uma superioridade em médias por parte dos alunos de Ciência da Computação, tanto do gênero feminino quanto do masculino, em relação ao dos outros cursos.

No gráfico comparativo das médias ao longo dos anos (Figura 4.16), é possível notar que em 2011 houve uma queda significativa na média das alunas, tornando a média dos alunos, que não houve nenhum tipo de mudança brusca, superior a delas. Porém, em 2012 as notas voltaram a uma proporção semelhante a do ano 2010. No geral, é possível visualizar uma diminuição da média dos alunos, tendo em 2013 a menor média de notas dos alunos, com exceção de 2011.

Na Figura 4.17 temos o comparativo de médias das disciplinas em que os alunos normalmente tem mais dificuldade. É possível visualizar que apesar das dificuldades, os alunos tem menos dificuldade nas disciplinas de Software Básico e Organização de Arquivos, e em ambas, as mulheres tem uma média de notas melhores que a dos homens, principalmente em Software Básico. Nas disciplinas restantes, a diferença de gênero é bem pequena, mas Computação Básica, Estrutura de Dados e Organização e Arquitetura de Computadores, os homens tem notas ligeiramente melhores que as mulheres.

Com essa informação, ao analisar as disciplinas individualmente por curso, pode perceber que os alunos de Computação (Licenciatura) possuem uma dificuldade muito grande com a disciplina de Cálculo 1, principalmente as mulheres, enquanto que nos cursos de Ciência e Engenharia da Computação, apesar da nota ainda ser baixa, é maior do que a média geral e a nota das alunas é um pouco maior que a dos homens. No caso de Física 1, que é obrigatória para os cursos de Ciência e Engenharia da Computação, possuem padrões diferenciados, pois enquanto os alunos do primeiro curso possuem notas melhores que das mulheres e acima da média, as alunas do curso de Engenharia possuem notas melhores que a dos homens e acima da média. No caso de Computação (Licenciatura), no qual a disciplina de Física 1 não é obrigatória, as mulheres possuem uma média significativamente maior que a dos homens, mas ambas abaixo da média geral, principalmente os alunos. Por último, a disciplina de Computação Básica, todos os cursos possuem o mesmo padrão da média geral, no qual os homens possuem notas melhores que as mulheres, a única peculiaridade é a nota significativamente baixa das alunas do curso Noturno de Computação.

Esses resultados, juntamente dos gráficos sobre desligamento dos alunos da seção 4.1, facilitam a interpretação dos gráficos relativos a porcentagem do período de desligamento dos alunos por curso. Na Figura 4.21, é possível visualizar que há um grande número de alunas desligadas no segundo e terceiro período de curso, com 21.2% e 29.3% respecti-

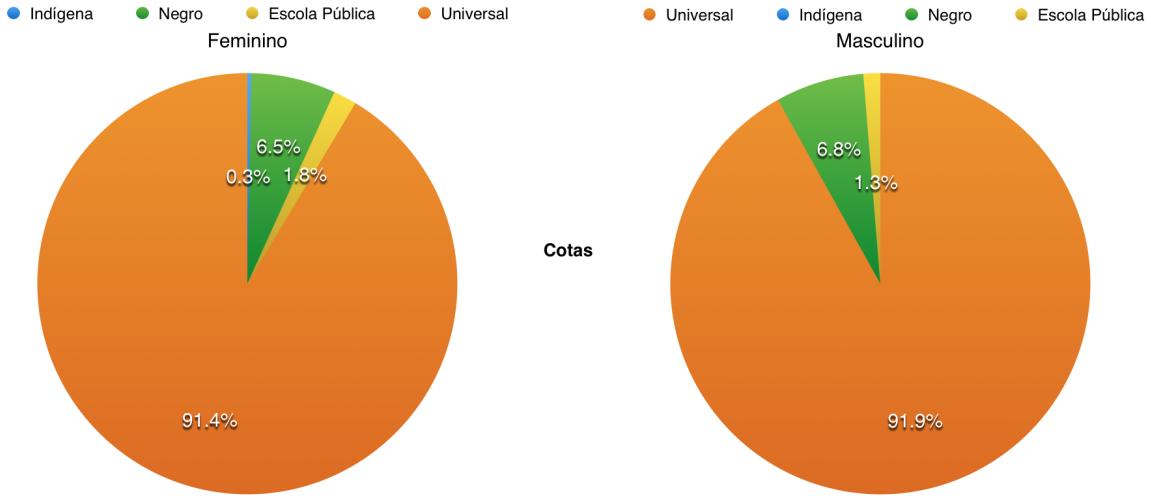


Figura 4.25: Gráfico pizza com a relação cotas/sexo geradas pela rede bayesiana.

vamente, porém no quarto período essa porcentagem cai para 6.1%. No caso dos alunos o período com maior desligamento é o terceiro período, com 23.4%, diminuindo gradativamente ao longo dos próximos períodos. No caso de Licenciatura em Computação, o padrão de desligamento entre homens e mulheres é bem semelhante, assim como a porcentagem entre os semestres. O maior valor em um semestre das mulheres da Licenciatura é de 16.9% no segundo semestre e 17.1% no terceiro semestre no caso dos homens. O curso de Engenharia da Computação, apesar de ter sido criado recentemente, também possui desligamentos, e assim como o curso de Licenciatura, o perfil de homens e mulheres é bem semelhante, possuindo 36.4% das alunas desligadas no terceiro semestre e 31.7% dos alunos desligados neste mesmo período.

Ao interpretar os dados da rede bayesiana mostrada na figura 4.24, algumas informações podem ser retiradas. A primeira delas é a proporção por cotas, onde é possível perceber nos gráficos 4.25, que foram gerados a partir dos dados da rede, a proporção de cotas Indígenas para as mulheres é de 0.3%, o que é baixo, mas ainda maior do que a dos homens que não possuem nenhum aluno oriundo dessa cota. Além disso, é possível perceber que existe uma proporção maior de cotas para escolas públicas dentre as alunas do que dos alunos.

A comparação do atributo sexo pelo tipo de escola, visualizada na figura 4.26, mostra que, apesar da proporção de alunos de escola privada é maior do que a de alunos de escola pública, existe uma diferença por gênero. A proporção de alunos do sexo masculino advindos de escola pública, que é 11.5%, é maior do que a das alunas, com 10.2%.

A Figura 4.27 mostra a comparação bayesiana entre sexo e raça/cor. Com ela, pode ser verificado que, apesar de não possuir nenhuma cota Indígena para os homens, 0.1% deles são dessa raça, e apesar dos homens possuírem uma proporção maior de cotas para Negros, as mulheres possuem uma proporção maior de negras no curso, com 3.3%, contra 2.9% dos homens. Além dos negros, as mulheres surpassam também na proporção de alunos de cor Amarela, com uma proporção de 1.5%, contra 1.1% dos homens.

Comparando a forma de ingresso com o gênero dos alunos (Figura 4.28), é possível visualizar algumas peculiaridades. Enquanto que os homens possuem uma proporção

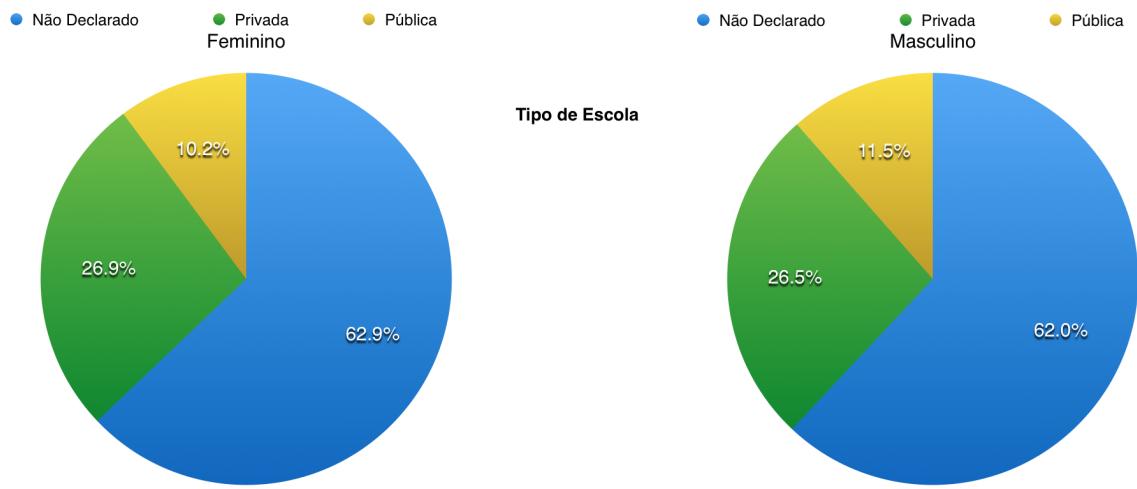


Figura 4.26: Gráfico pizza com a relação tipo de escola/sexo geradas pela rede bayesiana.

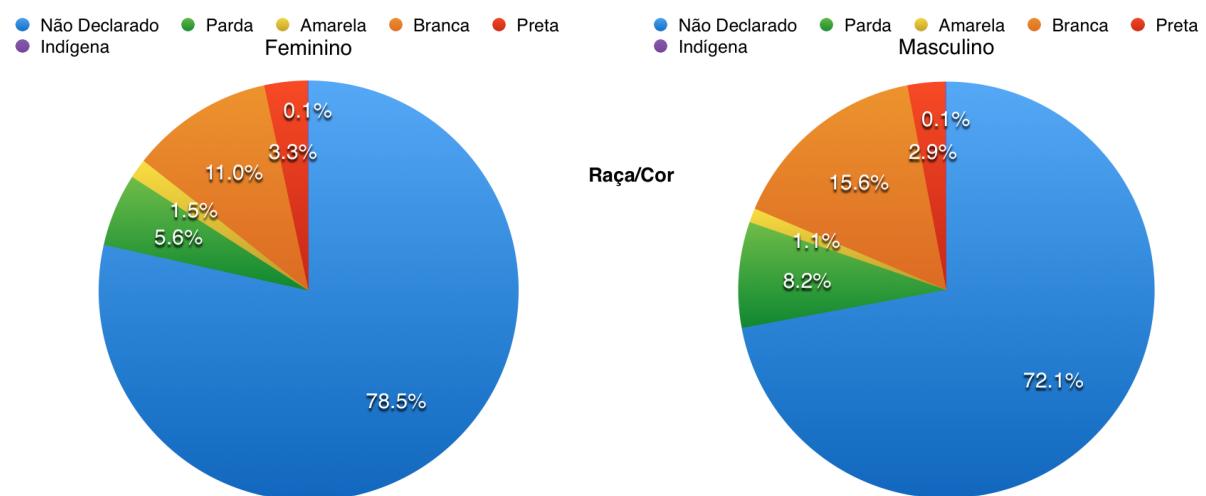


Figura 4.27: Gráfico pizza com a relação raça/sexo geradas pela rede bayesiana.

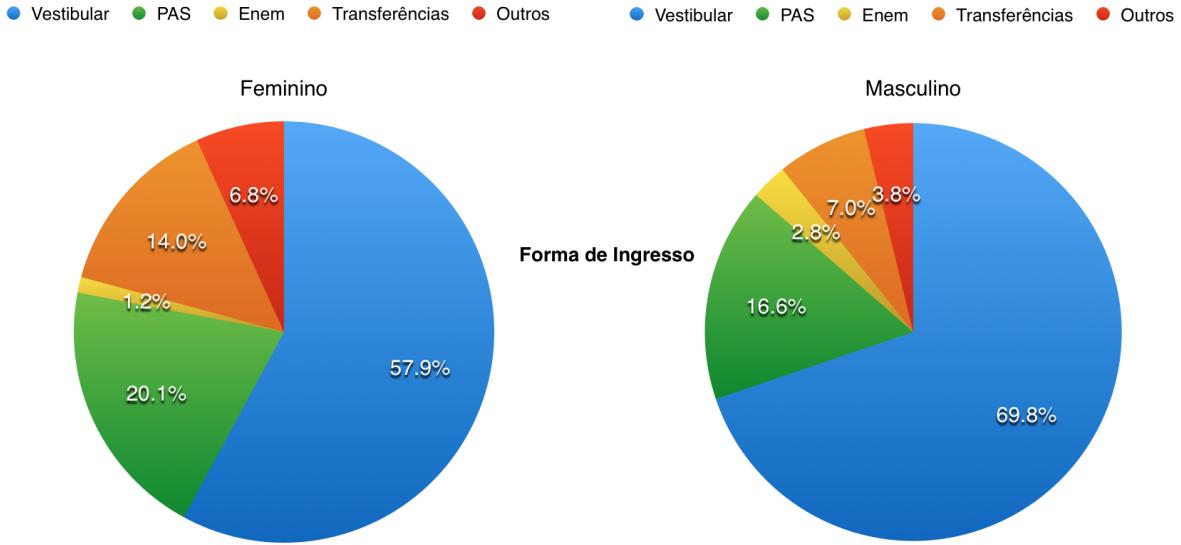


Figura 4.28: Gráfico pizza com a relação forma de ingresso/sexo geradas pela rede bayesiana.

maior de alunos advindos pelo vestibular e Enem do que as mulheres, 69.8% contra 57.9% e 2.8% contra 1.2% respectivamente, elas possuem uma maior proporção maior pelo PAS e por transferências, com 20.1% contra 16.6% e 14% contra 7% respectivamente.

Os gráficos de proporção de Forma de Saída por gênero, representado na Figura 4.29, mostra que a proporção de mulheres formadas (33.3%) é maior do que a dos homens (28.3%). Além disso, a proporção de desligamento por não cumprir condição é menor, com 14.3%, enquanto a dos alunos é de 20.3%. É importante ressaltar que esse gráfico difere das Figuras 4.4, 4.8 e 4.11, pelo fato de ser uma rede bayesiana, ou seja, uma predição, e não estar dividido por cursos.

Na Figura 4.30, podem ser visualizados os gráficos referentes a comparação bayesiana entre curso e gênero. É possível perceber uma proporção maior de mulheres no curso de Ciência da Computação (60.1%) do que nos outros. Apesar de possuir também uma proporção maior no mesmo curso, essa proporção, comparada com as mulheres, é menor (53.7%), porém nos outros cursos, a proporção dos homens é maior.

Os gráficos com a comparação por grau e gênero, Figura 4.31, confirmam o encontrado na Figura 4.30. É possível perceber que existe uma proporção muito maior de mulheres no curso de Bacharelado do que em Licenciatura (71%), assim como os homens (64.7%), porém essa proporção é maior no gráfico das alunas do que na dos alunos.

Por fim, temos os gráficos da Figura 4.32, referentes a comparação entre a idade de entrada na universidade e o gênero. No caso das mulheres, existe uma proporção muito grande de mulheres até os 24 anos de idade, que somados chegam a uma proporção de 88.1% de todas as alunas. Apesar dos homens também possuirem uma proporção semelhante, existe uma proporção menor entre os homens até 18 anos (35.8%) em relação às mulheres com a mesma faixa etária (43.8%). Além disso, existe uma proporção maior de homens na faixa etária de 30 a 34 anos (9.2%) do que das alunas (6%).

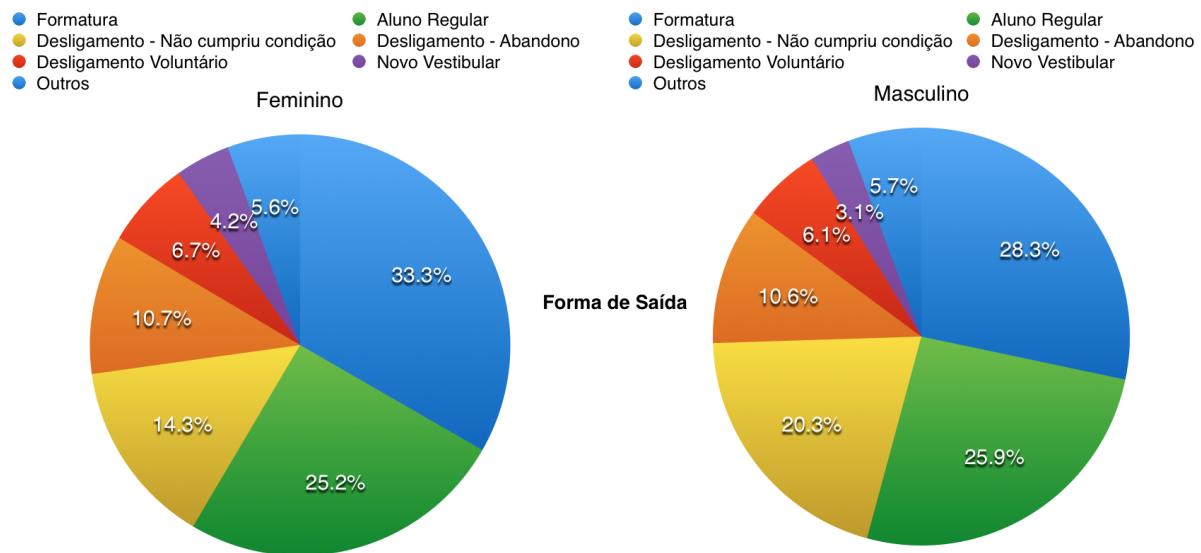


Figura 4.29: Gráfico pizza com a relação forma de saída/sexo geradas pela rede bayesiana.

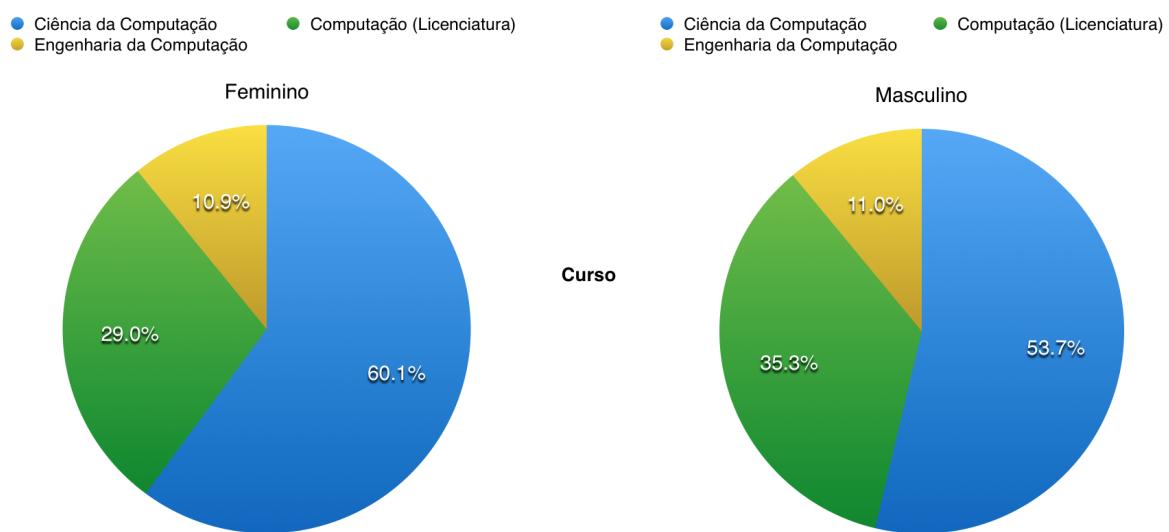


Figura 4.30: Gráfico pizza com a relação curso/sexo geradas pela rede bayesiana.

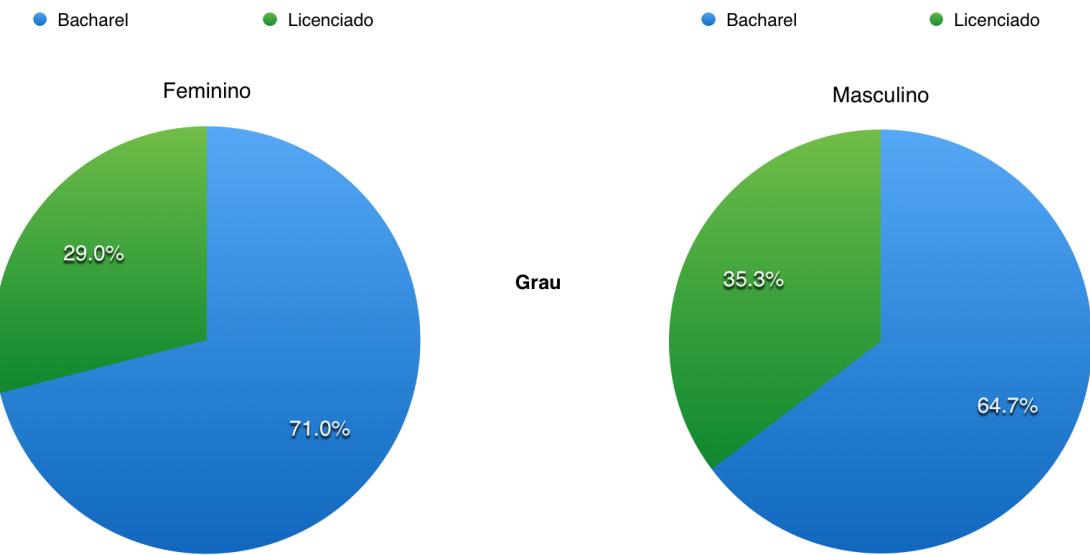


Figura 4.31: Gráfico pizza com a relação grau/gênero geradas pela rede bayesiana.

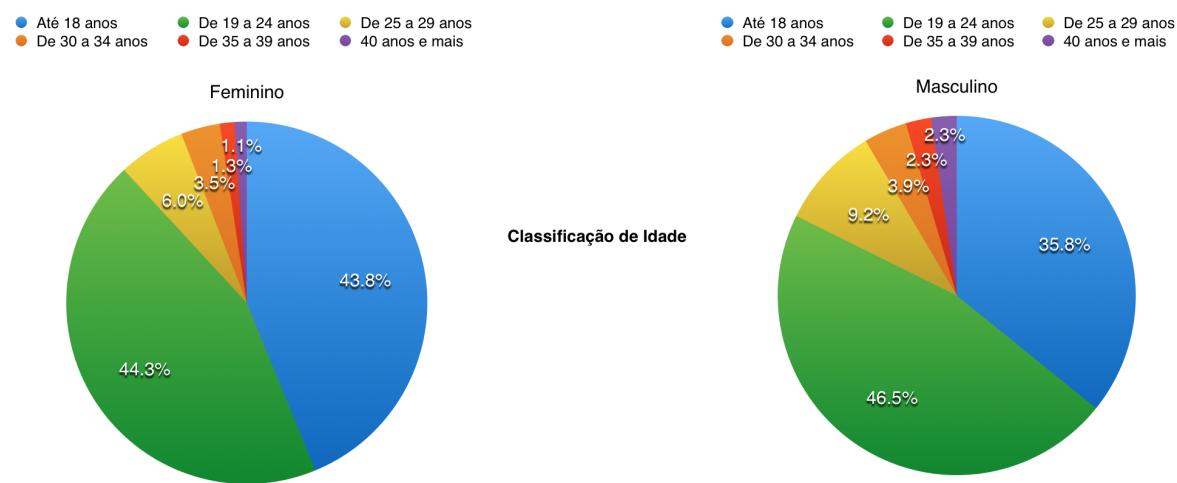


Figura 4.32: Gráfico pizza com a relação idade/gênero geradas pela rede bayesiana.

Capítulo 5

Conclusão

Neste trabalho de graduação, buscou-se encontrar padrões de gênero em dados referentes a estudantes do Ensino Fundamental e Médio, coletados em planilhas, e Superior, que foram extraídos do banco de dados Sigra da Universidade de Brasília, utilizando ferramentas de Mineração de Dados.

Primeiramente foi necessário solicitar a leitura dos dados coletados pelo Cespe - UnB, e após isso, conforme os passos sugeridos no capítulo 1, foi realizada a limpeza, integração, seleção e transformação dessa informação. Com os dados no formato correto, foi possível realizar a análise inicial, para então minerar os dados, e ao final ser feita a avaliação. Esse processo pode ser repetido até serem encontrados padrões que podem ser utilizados, e ao final a apresentação do conhecimento adquirido.

No Estudo de Caso 1, capítulo 3, foi verificado que, apesar do número baixo de meninas que sabem banco de dados, esse conhecimento vai crescendo de acordo com seu nível de conhecimento. Através da árvore de decisões é sugerido que as mulheres que não sabem banco de dados, não sabem programação também. Além disso, sugere que quem sabe banco de dados, acredita que na computação é utilizada a criatividade e acredita que esse curso é interdisciplinar, sabe programar. Portanto, é sugerido que o conhecimento de banco de dados e programação estão interligados.

Na mineração de dados relacionada a pergunta de estar pensando em fazer computação foram geradas algumas previsões. É sugerido que as meninas do Ensino Fundamental que acreditam que o profissional em computação é bem remunerado, pensam em fazer o curso de computação. No caso do primeiro ano do Ensino Médio, é sugerido que aquelas que sabem programar e acreditam no prestígio da profissão, pensam em fazer computação. Para o segundo ano, é previsto que aquelas que sabem banco de dados e acreditam no prestígio da Computação, pensam em fazer o curso. E por último, no caso do terceiro ano, a previsão de querer fazer o curso de Computação é daquelas que sabem programar e trabalhar com banco de dados, além de acreditar no prestígio da profissão e na boa remuneração.

No Estudo de Caso 2, capítulo 4, foram extraídos dados do SIGRA da Universidade de Brasília referentes aos alunos de cursos de Computação da universidade em questão. Porém, juntamente com esses dados, foram entregues algumas informações estatísticas sobre eles. Realizando uma análise inicial sobre essas estatísticas, foi possível verificar que os cursos de Licenciatura e Engenharia da Computação possuem padrões semelhantes entre homens e mulheres nos motivos de saída da universidade, porém no curso de Ciência da

Computação a maior causa de desligamento das alunas deste curso é Abandono, enquanto no caso dos alunos é o fato de não cumprir condição.

A partir dos dados recebidos, também foi realizada a média dos alunos nas principais disciplinas cursadas por eles, e com isso verificou-se que as disciplinas de Física 1, Cálculo 1 e Computação Básica, fazem parte do grupo de disciplinas com médias mais baixas pelos alunos. Analisando-as individualmente foi descoberto que o curso de Licenciatura possui uma média extremamente baixa em Cálculo 1 comparada com a média dos outros cursos, principalmente as mulheres. Além disso, foi possível visualizar que a média das alunas de Ciência da Computação na disciplina de Física 1 é menor que a dos homens, e nos outros cursos essa média das mulheres é significativamente maior que a dos homens. No caso de Computação Básica, o padrão é similar em todos os cursos, com exceção das mulheres da Licenciatura, que possuem uma média muito abaixo da média.

Com os dados de ano de entrada e ano de saída dos alunos, foi possível verificar padrões de número de semestres cursados por alunos que são desligados da UnB sem colar grau por curso. O curso de Engenharia da Computação não foi possível verificar padrões pelo fato dele ser recente e não possuir muitos desligamento, mas ainda sim, assim como o curso de Licenciatura e dos homens da Ciência da Computação, os períodos com mais desligados são do segundo ao quarto, com esses valores crescendo e decrescendo proporcionalmente. No caso das mulheres da Ciência da Computação é um pouco diferente, existe um crescimento do segundo para o terceiro semestre, mas no quarto semestre há uma queda brusca de aproximadamente 23 pontos percentuais, o que leva a crer que o terceiro semestre é um período decisivo para essas alunas.

Por fim, foi realizada uma classificação bayesiana nos dados dos alunos por gênero. Essa classificação gerou resultados relacionados ao tipo de escola, onde homens possuem uma proporção maior que cursou em escolas públicas, enquanto que comparada com os alunos, as alunas possuem uma proporção maior que cursou em escolas privadas. Essa proporção é diferenciada também na classificação por raça, onde as mulheres possuem uma variedade racial maior que a dos homens. No caso da forma de ingresso, a proporção das alunas advindas do Programa Seriado e de Trasferências é maior que a dos alunos, que possuem uma proporção maior em Vestibular e Enem. Já na forma de saída, há uma proporção maior nas formaturas, e uma proporção menor no desligamento por não cumprir condição das mulheres em relação aos homens. O curso com maior proporção de mulheres é o de Ciência da Computação. E por fim, existe uma concentração maior de homens com idade entre 19 e 24 anos, enquanto as mulheres possuem uma proporção quase igual de alunas com menos de 19 anos e entre 19 e 24 anos, que somadas chega a quase 90% do total de mulheres.

5.1 Trabalhos Futuros

Como sugestão para trabalhos futuros, seria a comparação das reprovações dos alunos em disciplinas, com o período do currículo no qual elas se encontram, assim como a comparação do período de saída dos alunos com as disciplinas cursadas por estes. É possível também realizar a influência de determinada turma em disciplinas nas quais a cursada é pré-requisito, por exemplo, os alunos de determinada turma de Cálculo 1, conseguiu ter um resultado significativo em Cálculo 2?

Outra sugestão seria realizar uma pesquisa para saber porque os alunos de Licenciatura possuem uma dificuldade tão grande em disciplinas da Matemática, e o que fazer para mudar essa realidade. Tentar explicar o que causa esse acúmulo de desistências no segundo e terceiro período das alunas de Ciência da Computação.

No caso das alunas de Ensino Fundamental e Médio, reformular o questionário aplicado de forma a poder extrair mais informações, e após isso, tentar encontrar a diferença em perfis de diferentes anos, para avaliar a evolução desses alunos. Estudar a parte educacional de computação no Ensino Fundamental e Médio de forma a realizar algum tipo de divulgação de informações mais específicas sobre computação mais efetivamente.

Apêndice A

Querys utilizadas no Estudo de Caso 1

Listing A.1: Padronização da coluna de sexo

```
UPDATE dados SET sexo = NULL WHERE sexo NOT IN ( 'F' , 'M' );
```

Listing A.2: Tradução do atributo Série

```
UPDATE dados SET serie = 'Fundamental' WHERE serie = '1'  
UPDATE dados SET serie = '1_ano' WHERE serie = '2'  
UPDATE dados SET serie = '2_ano' WHERE serie = '3'  
UPDATE dados SET serie = '3_ano' WHERE serie = '4'  
UPDATE dados SET serie = 'Supletivo' WHERE serie = '5';  
UPDATE dados SET serie = 'Superior' WHERE serie = '6';  
UPDATE dados SET serie = NULL WHERE serie NOT IN ( 'Fundamental' ,  
'1_ano' , '2_ano' , '3_ano' , 'Supletivo' , 'Superior' )
```

Listing A.3: Substituição do campo da Área

```
UPDATE dados SET area = 'Exatas' WHERE area = '1'  
UPDATE dados SET area = 'Biologicas' WHERE area = '2'  
UPDATE dados SET area = 'Humanas' WHERE area = '3'  
UPDATE dados SET area = NULL WHERE area NOT IN ( 'Exatas' ,  
'Biologicas' , 'Humanas' )
```

Listing A.4: Tradução do campo pensando

```
UPDATE dados SET pensando = 'S' WHERE pensando = '1'  
UPDATE dados SET pensando = 'N' WHERE pensando = '2'  
UPDATE dados SET pensando = 'NS' WHERE pensando = '3'  
UPDATE dados SET pensando = NULL WHERE pensando NOT IN ( 'S' ,  
'N' , 'NS' )
```

Listing A.5: Distribuição das respostas da questão um

```
UPDATE dados SET acessa_centro_inclusao =  
IF(acessa_em_casa IS NULL, 0, IF(acessa_em_casa < 1, 0,  
LEFT(RIGHT(acessa_em_casa, 1), 1)))  
UPDATE dados SET acessa_biblioteca =
```

```

    IF(acessa_em_casa IS NULL, 0, IF(acessa_em_casa < 10, 0,
                                     LEFT(RIGHT(acessa_em_casa, 2), 1)))
UPDATE dados SET acessa_lan_house = IF(acessa_em_casa IS NULL, 0,
                                         IF(acessa_em_casa < 100, 0, LEFT(RIGHT(acessa_em_casa, 3), 1)))
UPDATE dados SET acessa_trabalho = IF(acessa_em_casa IS NULL, 0,
                                         IF(acessa_em_casa < 1000, 0, LEFT(RIGHT(acessa_em_casa, 4), 1)))
UPDATE dados SET acessa_escola = IF(acessa_em_casa IS NULL, 0,
                                         IF(acessa_em_casa < 10000, 0, LEFT(RIGHT(acessa_em_casa, 5), 1)))
UPDATE dados SET acessa_amigos =
    IF(acessa_em_casa IS NULL, 0, IF(acessa_em_casa < 100000, 0,
                                      LEFT(RIGHT(acessa_em_casa, 6), 1)))
UPDATE dados SET acessa_parentes =
    IF(acessa_em_casa IS NULL, 0, IF(acessa_em_casa < 1000000, 0,
                                      LEFT(RIGHT(acessa_em_casa, 7), 1)))
UPDATE dados SET acessa_em_casa =
    IF(acessa_em_casa IS NULL, 0, IF(acessa_em_casa < 10000000, 0,
                                      LEFT(RIGHT(acessa_em_casa, 8), 1)))

```

Listing A.6: Distribuição das respostas da questão dois

```

UPDATE dados SET sabe_outros =
    IF(sabe_edicao_texto IS NULL, 0, IF(sabe_edicao_texto < 1, 0,
                                         LEFT(RIGHT(sabe_edicao_texto, 1), 1)))
UPDATE dados SET sabe_programacao =
    IF(sabe_edicao_texto IS NULL, 0, IF(sabe_edicao_texto < 10, 0,
                                         LEFT(RIGHT(sabe_edicao_texto, 2), 1)))
UPDATE dados SET sabe_dev_pag =
    IF(sabe_edicao_texto IS NULL, 0, IF(sabe_edicao_texto < 100, 0,
                                         LEFT(RIGHT(sabe_edicao_texto, 3), 1)))
UPDATE dados SET sabe_jogos =
    IF(sabe_edicao_texto IS NULL, 0, IF(sabe_edicao_texto < 1000, 0,
                                         LEFT(RIGHT(sabe_edicao_texto, 4), 1)))
UPDATE dados SET sabe_email =
    IF(sabe_edicao_texto IS NULL, 0, IF(sabe_edicao_texto < 10000, 0,
                                         LEFT(RIGHT(sabe_edicao_texto, 5), 1)))
UPDATE dados SET sabe_redes_sociais =
    IF(sabe_edicao_texto IS NULL, 0,
        IF(sabe_edicao_texto < 100000, 0,
            LEFT(RIGHT(sabe_edicao_texto, 6), 1)))
UPDATE dados SET sabe_acesso_internet =
    IF(sabe_edicao_texto IS NULL, 0,
        IF(sabe_edicao_texto < 1000000, 0,
            LEFT(RIGHT(sabe_edicao_texto, 7), 1)))
UPDATE dados SET sabe_banco_dados =
    IF(sabe_edicao_texto IS NULL, 0,
        IF(sabe_edicao_texto < 10000000, 0,
            LEFT(RIGHT(sabe_edicao_texto, 8), 1)))

```

```

UPDATE dados SET sabe_planilha =
    IF(sabe_edicao_texto IS NULL, 0,
        IF(sabe_edicao_texto < 100000000, 0,
            LEFT(RIGHT(sabe_edicao_texto, 9), 1)))
UPDATE dados SET sabe_edicao_imagem =
    IF(sabe_edicao_texto IS NULL, 0,
        IF(sabe_edicao_texto < 100000000, 0,
            LEFT(RIGHT(sabe_edicao_texto, 10), 1)))
UPDATE dados SET sabe_edicao_texto =
    IF(sabe_edicao_texto IS NULL, 0,
        IF(sabe_edicao_texto < 1000000000, 0,
            LEFT(RIGHT(sabe_edicao_texto, 11), 1)))

```

Listing A.7: Query de atualização das questões de três a quatorze

— Questao 3

```

UPDATE dados SET q3_sup_ens_soft = 'S'
    WHERE q3_sup_ens_soft = '1'
UPDATE dados SET q3_sup_ens_soft = 'N'
    WHERE q3_sup_ens_soft = '2'
UPDATE dados SET q3_sup_ens_soft = 'T'
    WHERE q3_sup_ens_soft = '3'
UPDATE dados SET q3_sup_ens_soft = NULL WHERE q3_sup_ens_soft
    NOT IN ('S', 'N', 'T')

```

— Questao 4

```

UPDATE dados SET q4_sup_pouca_mat = 'S'
    WHERE q4_sup_pouca_mat = '1'
UPDATE dados SET q4_sup_pouca_mat = 'N'
    WHERE q4_sup_pouca_mat = '2'
UPDATE dados SET q4_sup_pouca_mat = 'T'
    WHERE q4_sup_pouca_mat = '3'
UPDATE dados SET q4_sup_pouca_mat = NULL WHERE q4_sup_pouca_mat
    NOT IN ('S', 'N', 'T')

```

— Questao 5

```

UPDATE dados SET q5_maior_comp_masc = 'S'
    WHERE q5_maior_comp_masc = '1'
UPDATE dados SET q5_maior_comp_masc = 'N'
    WHERE q5_maior_comp_masc = '2'
UPDATE dados SET q5_maior_comp_masc = 'T'
    WHERE q5_maior_comp_masc = '3'
UPDATE dados SET q5_maior_comp_masc = NULL WHERE
    q5_maior_comp_masc NOT IN ('S', 'N', 'T')

```

— Questao 6

```

UPDATE dados SET q6_sab_usar_comp_p_cursar = 'S'

```

```

WHERE q6_sab_usar_comp_p_cursar = '1'
UPDATE dados SET q6_sab_usar_comp_p_cursar = 'N'
WHERE q6_sab_usar_comp_p_cursar = '2'
UPDATE dados SET q6_sab_usar_comp_p_cursar = 'T'
WHERE q6_sab_usar_comp_p_cursar = '3'
UPDATE dados SET q6_sab_usar_comp_p_cursar = NULL WHERE
q6_sab_usar_comp_p_cursar NOT IN ('S', 'N', 'T')

```

-- Questao 7

```

UPDATE dados SET q7_prec_curs_sup_p_trab = 'S'
WHERE q7_prec_curs_sup_p_trab = '1'
UPDATE dados SET q7_prec_curs_sup_p_trab = 'N'
WHERE q7_prec_curs_sup_p_trab = '2'
UPDATE dados SET q7_prec_curs_sup_p_trab = 'T'
WHERE q7_prec_curs_sup_p_trab = '3'
UPDATE dados SET q7_prec_curs_sup_p_trab = NULL WHERE
q7_prec_curs_sup_p_trab NOT IN ('S', 'N', 'T')

```

-- Questao 8

```

UPDATE dados SET q8_fam_gost_vest_comp = 'S'
WHERE q8_fam_gost_vest_comp = '1'
UPDATE dados SET q8_fam_gost_vest_comp = 'N'
WHERE q8_fam_gost_vest_comp = '2'
UPDATE dados SET q8_fam_gost_vest_comp = 'T'
WHERE q8_fam_gost_vest_comp = '3'
UPDATE dados SET q8_fam_gost_vest_comp = NULL WHERE
q8_fam_gost_vest_comp NOT IN ('S', 'N', 'T')

```

-- Questao 9

```

UPDATE dados SET q9_dif_empr_dps_formado = 'S'
WHERE q9_dif_empr_dps_formado = '1'
UPDATE dados SET q9_dif_empr_dps_formado = 'N'
WHERE q9_dif_empr_dps_formado = '2'
UPDATE dados SET q9_dif_empr_dps_formado = 'T'
WHERE q9_dif_empr_dps_formado = '3'
UPDATE dados SET q9_dif_empr_dps_formado = NULL WHERE
q9_dif_empr_dps_formado NOT IN ('S', 'N', 'T')

```

-- Questao 10

```

UPDATE dados SET q10_trab_pouco_lazer = 'S'
WHERE q10_trab_pouco_lazer = '1'
UPDATE dados SET q10_trab_pouco_lazer = 'N'
WHERE q10_trab_pouco_lazer = '2'
UPDATE dados SET q10_trab_pouco_lazer = 'T'
WHERE q10_trab_pouco_lazer = '3'
UPDATE dados SET q10_trab_pouco_lazer = NULL WHERE

```

```
q10_trab_pouco_lazer NOT IN ( 'S' , 'N' , 'T' )
```

— Questao 11

```
UPDATE dados SET q11_trab_comp_criatividade = 'S'  
                      WHERE q11_trab_comp_criatividade = '1'  
UPDATE dados SET q11_trab_comp_criatividade = 'N'  
                      WHERE q11_trab_comp_criatividade = '2'  
UPDATE dados SET q11_trab_comp_criatividade = 'T'  
                      WHERE q11_trab_comp_criatividade = '3'  
UPDATE dados SET q11_trab_comp_criatividade = NULL WHERE  
q11_trab_comp_criatividade NOT IN ( 'S' , 'N' , 'T' )
```

— Questao 12

```
UPDATE dados SET q12_trab_comp_prestigio = 'S'  
                      WHERE q12_trab_comp_prestigio = '1'  
UPDATE dados SET q12_trab_comp_prestigio = 'N'  
                      WHERE q12_trab_comp_prestigio = '2'  
UPDATE dados SET q12_trab_comp_prestigio = 'T'  
                      WHERE q12_trab_comp_prestigio = '3'  
UPDATE dados SET q12_trab_comp_prestigio = NULL WHERE  
q12_trab_comp_prestigio NOT IN ( 'S' , 'N' , 'T' )
```

— Questao 13

```
UPDATE dados SET q13_trab_comp_ganha_bem = 'S'  
                      WHERE q13_trab_comp_ganha_bem = '1'  
UPDATE dados SET q13_trab_comp_ganha_bem = 'N'  
                      WHERE q13_trab_comp_ganha_bem = '2'  
UPDATE dados SET q13_trab_comp_ganha_bem = 'T'  
                      WHERE q13_trab_comp_ganha_bem = '3'  
UPDATE dados SET q13_trab_comp_ganha_bem = NULL WHERE  
q13_trab_comp_ganha_bem NOT IN ( 'S' , 'N' , 'T' )
```

— Questao 14

```
UPDATE dados SET q14_trab_comp_atuar_outras_areas = 'S'  
                      WHERE q14_trab_comp_atuar_outras_areas = '1'  
UPDATE dados SET q14_trab_comp_atuar_outras_areas = 'N'  
                      WHERE q14_trab_comp_atuar_outras_areas = '2'  
UPDATE dados SET q14_trab_comp_atuar_outras_areas = 'T'  
                      WHERE q14_trab_comp_atuar_outras_areas = '3'  
UPDATE dados SET q14_trab_comp_atuar_outras_areas = NULL WHERE  
q14_trab_comp_atuar_outras_areas NOT IN ( 'S' , 'N' , 'T' )
```

Listing A.8: Query de seleção dos dados separados por perfil

— Respondentes pensando em fazer curso de Computacao

```
SELECT ano ,  
serie ,
```

```

IF( sabe_banco_dados = '1' , 'S' , 'N') AS
'sabe_banco_dados',
IF( sabe_acesso_internet = '1' , 'S' , IF(
    sabe_redes_sociais = '1' , 'S' , IF(
        sabe_email = '1' , 'S' , 'N' ))) AS
'sabe_usar_internet',
IF( sabe_dev_pag = '1' , 'S' , IF(
sabe_programacao = '1' , 'S' , 'N')) AS
'sabe_programar',
IF( sabe_edicao_texto = '1' , 'S' , IF(
    sabe_edicao_imagem = '1' , 'S' , IF(
        sabe_planilha = '1' , 'S' , 'N' ))) AS
'sabe_basico',
q4_sup_pouca_mat,
q7_prec_curs_sup_p_trab,
q10_trab_pouco_lazer,
q11_trab_comp_criatividade,
q12_trab_comp_prestigio,
q13_trab_comp_ganha_bem,
q14_trab_comp_atuar_outras_areas
FROM dados d
WHERE sexo = 'F'
    AND serie NOT IN ( 'Superior' )
    AND pensando = 'S'
    AND serie IS NOT NULL
    AND sabe_banco_dados IS NOT NULL
    AND 'q4_sup_pouca_mat' IS NOT NULL
    AND 'q7_prec_curs_sup_p_trab' IS NOT NULL
    AND 'q10_trab_pouco_lazer' IS NOT NULL
    AND 'q11_trab_comp_criatividade' IS NOT NULL
    AND 'q12_trab_comp_prestigio' IS NOT NULL
    AND 'q13_trab_comp_ganha_bem' IS NOT NULL
    AND 'q14_trab_comp_atuar_outras_areas' IS NOT NULL

```

— Respondentes que nao estao pensando em fazer curso de
— Computacao ou nao sabem

```

SELECT ano ,
    serie ,
    IF( sabe_banco_dados = '1' , 'S' , 'N') AS
'sabe_banco_dados',
    IF( sabe_acesso_internet = '1' , 'S' , IF(
        sabe_redes_sociais = '1' , 'S' , IF(
            sabe_email = '1' , 'S' , 'N' ))) AS
'sabe_usar_internet',
    IF( sabe_dev_pag = '1' , 'S' , IF(
sabe_programacao = '1' , 'S' , 'N')) AS

```

```

'sabe_programar',
IF( sabe_edicao_texto = '1', 'S', IF(
    sabe_edicao_imagem = '1', 'S', IF(
        sabe_planilha = '1', 'S', 'N' ))) AS
'sabe_basico',
q4_sup_pouca_mat,
q7_prec_curs_sup_p_trab,
q10_trab_pouco_lazer,
q11_trab_comp_criatividade,
q12_trab_comp_prestigio,
q13_trab_comp_ganha_bem,
q14_trab_comp_atuar_outras_areas
FROM dados d
WHERE sexo = 'F'
    AND serie NOT IN ('Superior')
    AND pensando != 'S'
    AND serie IS NOT NULL
    AND sabe_banco_dados IS NOT NULL
    AND 'q4_sup_pouca_mat' IS NOT NULL
    AND 'q7_prec_curs_sup_p_trab' IS NOT NULL
    AND 'q10_trab_pouco_lazer' IS NOT NULL
    AND 'q11_trab_comp_criatividade' IS NOT NULL
    AND 'q12_trab_comp_prestigio' IS NOT NULL
    AND 'q13_trab_comp_ganha_bem' IS NOT NULL
    AND 'q14_trab_comp_atuar_outras_areas' IS NOT NULL

```

Apêndice B

Saída da árvore de decisões com dois perfis do Estudo de Caso 1

```

==== Run information ====
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: questionario-weka.filters.unsupervised.attribute.Remove-R5,7-weka.filters.unsupervised.attribute.Remove-R7
Instances: 2539
Attributes: 11
ano
serie
pensando
sabe_banco_dados
sabe_programar
q4_sup_pouca_mat
q10_trab_pouco_lazer
q11_trab_comp_criatividade
q12_trab_comp_prestigio
q13_trab_comp_ganha_bem
q14_trab_comp_atuar_outras_areas
Test mode:10-fold cross-validation

==== Classifier model (full training set) ====
J48 pruned tree
-----
serie = Fundamental
| sabe_banco_dados = S
| | q11_trab_comp_criatividade = S
| | | ano = 2011: S (53.0/12.0)
| | | ano = 2012
| | | | sabe_programar = S
| | | | | q4_sup_pouca_mat = S: NS (3.0/1.0)
| | | | | q4_sup_pouca_mat = N: S (13.0/2.0)
| | | | | q4_sup_pouca_mat = T: NS (6.0)
| | | | sabe_programar = N
| | | | | q12_trab_comp_prestigio = S
| | | | | | q14_trab_comp_atuar_outras_areas = S
| | | | | | | q13_trab_comp_ganha_bem = S: S (3.0/1.0)
| | | | | | | q13_trab_comp_ganha_bem = N: NS (0.0)
| | | | | | | q13_trab_comp_ganha_bem = T: NS (7.0/3.0)
| | | | | | q14_trab_comp_atuar_outras_areas = N: N (1.0)
| | | | | | q14_trab_comp_atuar_outras_areas = T: N (2.0)
| | | | | q12_trab_comp_prestigio = N: S (1.0)
| | | | | q12_trab_comp_prestigio = T: NS (3.0)
| | | ano = 2013
| | | | q10_trab_pouco_lazer = S: N (1.0)
| | | | q10_trab_pouco_lazer = N
| | | | | q14_trab_comp_atuar_outras_areas = S: NS (6.0/2.0)
| | | | | q14_trab_comp_atuar_outras_areas = N: NS (0.0)
| | | | | q14_trab_comp_atuar_outras_areas = T: S (2.0)
| | | | | q10_trab_pouco_lazer = T: S (4.0)
| | | | q11_trab_comp_criatividade = N: NS (5.0/2.0)
| | | | q11_trab_comp_criatividade = T: S (11.0/1.0)
| | sabe_banco_dados = N
| | | q13_trab_comp_ganha_bem = S
| | | | ano = 2011: S (155.0/64.0)
| | | | ano = 2012
| | | | | q4_sup_pouca_mat = S
| | | | | | sabe_programar = S: S (7.0/2.0)
| | | | | | sabe_programar = N: NS (13.0/5.0)
| | | | | q4_sup_pouca_mat = N
| | | | | | sabe_programar = S: NS (20.0/8.0)
| | | | | | sabe_programar = N: S (41.0/14.0)
| | | | | q4_sup_pouca_mat = T: NS (31.0/12.0)
| | | | ano = 2013
| | | | | q4_sup_pouca_mat = S
| | | | | | q14_trab_comp_atuar_outras_areas = S: NS (6.0/3.0)
| | | | | | q14_trab_comp_atuar_outras_areas = N: S (0.0)
| | | | | | q14_trab_comp_atuar_outras_areas = T: S (2.0)
| | | | | q4_sup_pouca_mat = N
| | | | | | q10_trab_pouco_lazer = S: N (2.0)
| | | | | | q10_trab_pouco_lazer = N
| | | | | | | sabe_programar = S: S (4.0/1.0)
| | | | | | | sabe_programar = N: NS (10.0/5.0)
| | | | | | q10_trab_pouco_lazer = T: NS (6.0/3.0)
| | | | | q4_sup_pouca_mat = T
| | | | | | q14_trab_comp_atuar_outras_areas = S: S (7.0/1.0)
| | | | | | q14_trab_comp_atuar_outras_areas = N: S (0.0)
| | | | | | q14_trab_comp_atuar_outras_areas = T: NS (4.0/1.0)
| | | q13_trab_comp_ganha_bem = N
| | | | sabe_programar = S: NS (5.0)
| | | | sabe_programar = N
| | | | | ano = 2011
| | | | | | q12_trab_comp_prestigio = S: S (13.0/5.0)
| | | | | | q12_trab_comp_prestigio = N
| | | | | | | q4_sup_pouca_mat = S: NS (2.0)
| | | | | | | q4_sup_pouca_mat = N: S (2.0/1.0)
| | | | | | | q4_sup_pouca_mat = T: NS (0.0)
| | | | | | q12_trab_comp_prestigio = T: NS (5.0/2.0)
| | | | ano = 2012: NS (5.0/1.0)
| | | | ano = 2013
| | | | | q4_sup_pouca_mat = S: NS (2.0)
| | | | | q4_sup_pouca_mat = N: N (5.0)
| | | | | q4_sup_pouca_mat = T: N (0.0)

```

Figura B.1: Parte 1 da saída da árvore de decisões com os dois perfis.

```

| q13_trab_comp_ganha_bem = T
| q12_trab_comp_prestigio = S
|   q11_trab_comp_criatividade = S
|     sabe_programar = S: S (51.0/21.0)
|     sabe_programar = N
|       q4_sup_pouca_mat = S
|         ano = 2011: S (12.0/2.0)
|         ano = 2012: S (3.0/1.0)
|         ano = 2013: NS (6.0/2.0)
|       q4_sup_pouca_mat = N: NS (53.0/25.0)
|       q4_sup_pouca_mat = T: NS (34.0/12.0)
|   q11_trab_comp_criatividade = N
|     sabe_programar = S: NS (4.0/1.0)
|     sabe_programar = N: S (8.0/3.0)
|   q11_trab_comp_criatividade = T: NS (9.0/2.0)
| q12_trab_comp_prestigio = N
|   q14_trab_comp_atuar_outras_areas = S
|     q10_trab_pouco_lazer = S: S (2.0)
|     q10_trab_pouco_lazer = N: S (10.0/4.0)
|     q10_trab_pouco_lazer = T: N (12.0/5.0)
|   q14_trab_comp_atuar_outras_areas = N: NS (6.0/2.0)
|   q14_trab_comp_atuar_outras_areas = T: NS (3.0/1.0)
| q12_trab_comp_prestigio = T
|   q11_trab_comp_criatividade = S
|     q14_trab_comp_atuar_outras_areas = S
|       q4_sup_pouca_mat = S
|         q10_trab_pouco_lazer = S: N (1.0)
|         q10_trab_pouco_lazer = N: S (12.0/5.0)
|         q10_trab_pouco_lazer = T: NS (6.0/3.0)
|       q4_sup_pouca_mat = N
|         q10_trab_pouco_lazer = S: S (3.0/1.0)
|         q10_trab_pouco_lazer = N
|           ano = 2011: N (4.0/2.0)
|           ano = 2012: N (4.0/1.0)
|           ano = 2013: S (5.0/2.0)
|         q4_sup_pouca_mat = T: NS (26.0/10.0)
|       q14_trab_comp_atuar_outras_areas = N: N (1.0)
|     q14_trab_comp_atuar_outras_areas = T: NS (35.0/16.0)
|   q11_trab_comp_criatividade = N: NS (3.0/1.0)
|   q11_trab_comp_criatividade = T
|     q14_trab_comp_atuar_outras_areas = S: N (11.0/4.0)
|     q14_trab_comp_atuar_outras_areas = N: NS (0.0)
|     q14_trab_comp_atuar_outras_areas = T: NS (12.0/3.0)
serie = 1 ano
| q12_trab_comp_prestigio = S
|   sabe_programar = S: S (124.0/52.0)
|   sabe_programar = N
|     q14_trab_comp_atuar_outras_areas = S
|       ano = 2011
|         q4_sup_pouca_mat = S: S (13.0/6.0)
|         q4_sup_pouca_mat = N: S (65.0/29.0)
|         q4_sup_pouca_mat = T: NS (46.0/23.0)
|       ano = 2012
|         q10_trab_pouco_lazer = S: S (6.0/2.0)
|         q10_trab_pouco_lazer = N
|           q4_sup_pouca_mat = S: S (12.0/5.0)
|           q4_sup_pouca_mat = N: NS (17.0/3.0)
|           q4_sup_pouca_mat = T: S (15.0/6.0)
|         q10_trab_pouco_lazer = T
|           sabe_banco_dados = S: S (3.0/1.0)
|           sabe_banco_dados = N
|             q4_sup_pouca_mat = S: N (1.0)
|             q4_sup_pouca_mat = N: N (7.0/3.0)
|             q4_sup_pouca_mat = T: NS (3.0/1.0)
|           ano = 2013: NS (50.0/20.0)
|         q14_trab_comp_atuar_outras_areas = N: NS (4.0/2.0)
|       q14_trab_comp_atuar_outras_areas = T
|         q4_sup_pouca_mat = S
|           q10_trab_pouco_lazer = S: S (2.0)
|           q10_trab_pouco_lazer = N: S (3.0/1.0)
|           q10_trab_pouco_lazer = T: N (3.0/1.0)
|         q4_sup_pouca_mat = N
|           sabe_banco_dados = S: S (3.0/1.0)
|           sabe_banco_dados = N
|             ano = 2011: NS (18.0/8.0)
|             ano = 2012
|               q10_trab_pouco_lazer = S: S (1.0)
|               q10_trab_pouco_lazer = N: NS (2.0)
|               q10_trab_pouco_lazer = T: N (3.0)
|             ano = 2013
|               q13_trab_comp_ganha_bem = S: NS (2.0)
|               q13_trab_comp_ganha_bem = N: NS (0.0)
|               q13_trab_comp_ganha_bem = T: N (3.0/1.0)
|             q4_sup_pouca_mat = T
|               q13_trab_comp_ganha_bem = S: S (5.0/1.0)
|               q13_trab_comp_ganha_bem = N: S (0.0)
|               q13_trab_comp_ganha_bem = T: NS (7.0/2.0)
|             q12_trab_comp_prestigio = N: N (38.0/20.0)
|             q12_trab_comp_prestigio = T
|               q14_trab_comp_atuar_outras_areas = S
|                 sabe_banco_dados = S
|                   ano = 2011: S (5.0)
|                   ano = 2012
|                     sabe_programar = S: NS (2.0)
|                     sabe_programar = N: S (4.0/1.0)
|                   ano = 2013
|                     sabe_programar = S: S (2.0/1.0)
|                     sabe_programar = N: NS (2.0)

```

Figura B.2: Parte 2 da saída da árvore de decisões com os dois perfis.

```

    | sabe_banco_dados = N: NS (117.0/54.0)
q14_trab_comp_atuar_outras_areas = N: N (5.0/1.0)
q14_trab_comp_atuar_outras_areas = T
    | q10_trab_pouco_lazer = S
        | q13_trab_comp_ganha_bem = S: S (2.0)
        | q13_trab_comp_ganha_bem = N: S (1.0)
        | q13_trab_comp_ganha_bem = T: NS (4.0)
q10_trab_pouco_lazer = N
    | ano = 2011: NS (16.0/4.0)
    | ano = 2012: N (5.0/2.0)
    | ano = 2013
        | q4_sup_pouca_mat = S: N (3.0/1.0)
        | q4_sup_pouca_mat = N: S (0.0)
        | q4_sup_pouca_mat = T: S (3.0/1.0)
q10_trab_pouco_lazer = T
    | ano = 2011
        | sabe_programar = S: NS (3.0/1.0)
        | sabe_programar = N
            | q4_sup_pouca_mat = S: S (2.0/1.0)
            | q4_sup_pouca_mat = N: S (4.0)
            | q4_sup_pouca_mat = T
                | q11_trab_comp_criatividade = S: N (6.0/1.0)
                | q11_trab_comp_criatividade = N: N (0.0)
                | q11_trab_comp_criatividade = T: NS (3.0/1.0)
    | ano = 2012: N (13.0/4.0)
    | ano = 2013
        | sabe_programar = S: N (2.0)
        | sabe_programar = N: NS (3.0/1.0)
serie = 2 ano
q14_trab_comp_atuar_outras_areas = S
q4_sup_pouca_mat = S
    | q12_trab_comp_prestigio = S
        | ano = 2011: S (20.0/7.0)
        | ano = 2012: N (12.0/6.0)
        | ano = 2013: S (5.0/1.0)
    | q12_trab_comp_prestigio = N: N (5.0/1.0)
    | q12_trab_comp_prestigio = T: S (17.0/6.0)
q4_sup_pouca_mat = N
    | q12_trab_comp_prestigio = S
    | sabe_banco_dados = S: (40.0/21.0)
    | sabe_banco_dados = N
        | q13_trab_comp_ganha_bem = S
            | q10_trab_pouco_lazer = S: S (4.0/1.0)
            | q10_trab_pouco_lazer = N
                | q11_trab_comp_criatividade = S: NS (58.0/34.0)
                | q11_trab_comp_criatividade = N: NS (0.0)
                | q11_trab_comp_criatividade = T: S (2.0/1.0)
            | q10_trab_pouco_lazer = T
                | ano = 2011: S (11.0/0.3)
                | ano = 2012: NS (5.0/1.0)
                | ano = 2013: S (3.0/2.0)
        | q13_trab_comp_ganha_bem = N: S (3.0/1.0)
        | q13_trab_comp_ganha_bem = T
            | ano = 2011
                | q10_trab_pouco_lazer = S: S (0.0)
                | q10_trab_pouco_lazer = N
                    | sabe_programar = S: NS (6.0/3.0)
                    | sabe_programar = N: S (9.0/3.0)
                    | q10_trab_pouco_lazer = T: N (8.0/3.0)
            | ano = 2012
                | sabe_programar = S: S (5.0/2.0)
                | sabe_programar = N: N (8.0/4.0)
            | ano = 2013: N (10.0/5.0)
    | q12_trab_comp_prestigio = N
        | ano = 2011: S (3.0/1.0)
        | ano = 2012: S (1.0)
        | ano = 2013: NS (3.0/1.0)
    | q12_trab_comp_prestigio = T
        | q11_trab_comp_criatividade = S
            | sabe_banco_dados = S
                | sabe_programar = S: NS (6.0/1.0)
                | sabe_programar = N: N (6.0/2.0)
            | sabe_banco_dados = N
                | q10_trab_pouco_lazer = S: S (2.0/1.0)
                | q10_trab_pouco_lazer = N
                    | q13_trab_comp_ganha_bem = S: NS (6.0/2.0)
                    | q13_trab_comp_ganha_bem = N: N (1.0)
                    | q13_trab_comp_ganha_bem = T: N (15.0/7.0)
                | q10_trab_pouco_lazer = T
                    | ano = 2011
                        | sabe_programar = S: S (3.0/1.0)
                        | sabe_programar = N: NS (13.0/5.0)
                    | ano = 2012
                        | sabe_programar = S: NS (4.0)
                        | sabe_programar = N: S (6.0/2.0)
                    | ano = 2013: S (2.0/1.0)
            | q11_trab_comp_criatividade = N: NS (0.0)
            | q11_trab_comp_criatividade = T: N (9.0/4.0)
q4_sup_pouca_mat = T: NS (94.0/50.0)
q14_trab_comp_atuar_outras_areas = N
    | q13_trab_comp_ganha_bem = S: N (4.0/2.0)
    | q13_trab_comp_ganha_bem = N: S (2.0)
    | q13_trab_comp_ganha_bem = T: N (4.0/1.0)
q14_trab_comp_atuar_outras_areas = T
    | q11_trab_comp_criatividade = S: NS (83.0/40.0)
    | q11_trab_comp_criatividade = N: S (6.0/1.0)

```

Figura B.3: Parte 3 da saída da árvore de decisões com os dois perfis.

```

q11_trab_comp_criatividade = T
| q10_trab_pouco_lazer = S: NS (3.0/1.0)
| q10_trab_pouco_lazer = N
|   ano = 2011: NS (4.0/1.0)
|   ano = 2012: N (2.0/1.0)
|   ano = 2013: NS (0.0)
| q10_trab_pouco_lazer = T: N (11.0/3.0)
serie = 3 ano
sabe_programar = S
| q12_trab_comp_prestigio = S
| q13_trab_comp_ganha_bem = S
|   sabe_banco_dados = S: S (26.0/11.0)
|   sabe_banco_dados = N: NS (35.0/20.0)
| q13_trab_comp_ganha_bem = N: S (1.0)
| q13_trab_comp_ganha_bem = T
|   q4_sup_pouca_mat = S: S (2.0)
|   q4_sup_pouca_mat = N: N (9.0/4.0)
|   q4_sup_pouca_mat = T: NS (12.0/5.0)
q12_trab_comp_prestigio = M: S (7.0/3.0)
q12_trab_comp_prestigio = T
|   ano = 2011: S (7.0/3.0)
|   ano = 2012: N (14.0/7.0)
|   ano = 2013: N (5.0/1.0)
sabe_programar = N
ano = 2011
| q14_trab_comp_atuar_outras_areas = S
| q10_trab_pouco_lazer = S: NS (10.0/3.0)
| q10_trab_pouco_lazer = N: NS (95.0/48.0)
| q10_trab_pouco_lazer = T
|   q4_sup_pouca_mat = S: N (4.0/2.0)
|   q4_sup_pouca_mat = N
|     q13_trab_comp_ganha_bem = S: NS (10.0/4.0)
|     q13_trab_comp_ganha_bem = N: N (0.0)
|     q13_trab_comp_ganha_bem = T: N (15.0/5.0)
|   q4_sup_pouca_mat = T
|     sabe_banco_dados = S: S (2.0/1.0)
|     sabe_banco_dados = N: NS (6.0/2.0)
q14_trab_comp_atuar_outras_areas = N: N (5.0/2.0)
q14_trab_comp_atuar_outras_areas = T
| q11_trab_comp_criatividade = S
|   sabe_banco_dados = S: NS (4.0/1.0)
|   sabe_banco_dados = N: N (30.0/16.0)
| q11_trab_comp_criatividade = N: NS (1.0)
| q11_trab_comp_criatividade = T: NS (3.0/1.0)
ano = 2012
| q14_trab_comp_atuar_outras_areas = S
| q13_trab_comp_ganha_bem = S: N (66.0/29.0)
| q13_trab_comp_ganha_bem = N: NS (2.0)
| q13_trab_comp_ganha_bem = T
|   sabe_banco_dados = S: NS (7.0/2.0)
|   sabe_banco_dados = N: N (39.0/16.0)
q14_trab_comp_atuar_outras_areas = N: S (3.0/1.0)
q14_trab_comp_atuar_outras_areas = T
| q13_trab_comp_ganha_bem = S
|   q4_sup_pouca_mat = S: NS (0.0)
|   q4_sup_pouca_mat = N: NS (7.0/3.0)
|   q4_sup_pouca_mat = T: N (2.0)
|   q13_trab_comp_ganha_bem = N: N (1.0)
|   q13_trab_comp_ganha_bem = T: N (15.0/2.0)
ano = 2013
| q12_trab_comp_prestigio = S
| q13_trab_comp_ganha_bem = S
|   q4_sup_pouca_mat = S: NS (1.0)
|   q4_sup_pouca_mat = N
|     sabe_banco_dados = S: NS (3.0/1.0)
|     sabe_banco_dados = N: N (20.0/6.0)
|     q4_sup_pouca_mat = T: S (4.0/2.0)
|     q13_trab_comp_ganha_bem = N: NS (2.0)
|     q13_trab_comp_ganha_bem = T: NS (12.0/3.0)
|   q12_trab_comp_prestigio = N: N (9.0/6.0)
|   q12_trab_comp_prestigio = T: NS (2.0)
serie = Supletivo
q12_trab_comp_prestigio = S (10.0/4.0)
q12_trab_comp_prestigio = N: S (1.0)
q12_trab_comp_prestigio = T: NS (2.0)

Number of Leaves : 219
Size of the tree : 342

Time taken to build model: 0.01 seconds
== Stratified cross-validation ==
== Summary ==
Correctly Classified Instances      1118      43.718 %
Incorrectly Classified Instances    1429      56.282 %
Kappa statistic                      0.1224
Mean absolute error                  0.4085
Root mean squared error              0.4911
Relative absolute error              93.8371 %
Root relative squared error          105.2512 %
Total Number of Instances           2539

== Detailed Accuracy By Class ==


|               | TP    | Rate | FP    | Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---------------|-------|------|-------|------|-----------|--------|-----------|----------|-------|
| Weighted Avg. | 0.437 |      | 0.318 |      | 0.432     | 0.437  | 0.431     | 0.58     |       |
|               |       |      |       |      |           |        |           |          |       |


== Confusion Matrix ==


|     |     |     | <-- classified as |
|-----|-----|-----|-------------------|
| a   | b   | c   |                   |
| 469 | 89  | 366 | a = S             |
| 173 | 154 | 282 | b = N             |
| 360 | 159 | 487 | c = NS            |


```

Figura B.4: Parte 4 da saída da árvore de decisões com os dois perfis.

Apêndice C

Querys utilizadas no Estudo de Caso 2

Listing C.1: Tradução dos atributos com apenas um possível valor

— Níveis dos cursos

```
UPDATE dados_cpd SET nivel = 'Graduacao' WHERE nivel = '2';
UPDATE dados_cpd SET nivel_opcao = 'Graduacao'
                     WHERE nivel_opcao = '2';
UPDATE dados_cpd SET nivel_curso = 'Graduacao'
                     WHERE nivel_curso = '2';
```

— Duração dos cursos

```
UPDATE dados_cpd SET duracao = 'Duracao_Plena'
                     WHERE duracao = '2';
UPDATE dados_cpd SET duracao_curso = 'Duracao_Plena'
                     WHERE duracao_curso = '2';
```

— Faculdade na qual se encontra os cursos

```
UPDATE dados_cpd
       SET cod_faculdade = 'Instituto_de_Ciencias_Exatas'
                     WHERE cod_faculdade = '11';
```

— Departamento dos cursos

```
UPDATE dados_cpd
       SET cod_departamento = 'Departamento_de_Ciencia_da_Computacao'
                     WHERE cod_departamento = '116';
```

— Forma do curso (Presencial ou a Distância)

```
UPDATE dados_cpd SET forma_curso = 'Presencial'
                     WHERE forma_curso = '0';
```

Listing C.2: Atributos com dois possíveis valores

— Pessoa com deficiência?

```
UPDATE dados_cpd SET pessoa_deficiencia = 'S'
                     WHERE pessoa_deficiencia = '1';
UPDATE dados_cpd SET pessoa_deficiencia = 'N'
```

```

WHERE pessoa_deficiencia = '0';

-- Prioridade da opcao ao prestar Vestibular
UPDATE dados_cpd SET prioridade_opcao = 'Principal'
WHERE prioridade_opcao = '1';
UPDATE dados_cpd SET prioridade_opcao = 'Secundaria'
WHERE prioridade_opcao = '2';

-- Aluno registrado?
UPDATE dados_cpd SET aluno registrado = 'N'
WHERE aluno registrado = '0';
UPDATE dados_cpd SET aluno registrado = 'S'
WHERE aluno registrado = '1';

-- Grau do curso do aluno
UPDATE dados_cpd SET cod_grau = 'Bacharel' WHERE cod_grau = '1';
UPDATE dados_cpd SET cod_grau = 'Licenciado'
WHERE cod_grau = '2';

-- Turno do curso
UPDATE dados_cpd SET turno_curso = 'Diurno'
WHERE turno_curso = '1';
UPDATE dados_cpd SET turno_curso = 'Noturno'
WHERE turno_curso = '2';

-- Tipo da Escola (Publica ou Privada)
UPDATE dados_cpd SET tipo_escola = 'Nao_Declarado'
WHERE tipo_escola = '0';
UPDATE dados_cpd SET tipo_escola = 'Nao_Declarado'
WHERE tipo_escola = '';
UPDATE dados_cpd SET tipo_escola = 'Publica'
WHERE tipo_escola = '1';
UPDATE dados_cpd SET tipo_escola = 'Privada'
WHERE tipo_escola = '2';

```

Listing C.3: Tradução dos campos de nacionalidade e país de nascimento

```

-- Nacionalidade dos alunos
UPDATE dados_cpd SET nacionalidade = 'Africa do Sul'
WHERE nacionalidade = '2';
UPDATE dados_cpd SET nacionalidade = 'Angola'
WHERE nacionalidade = '8';
UPDATE dados_cpd SET nacionalidade = 'Argentina'
WHERE nacionalidade = '11';
UPDATE dados_cpd SET nacionalidade = 'Birmania'
WHERE nacionalidade = '19';
UPDATE dados_cpd SET nacionalidade = 'Bolivia'

```

```

                WHERE nacionalidade = '20';
UPDATE dados_cpd SET nacionalidade = 'Brasil'
                WHERE nacionalidade = '22';
UPDATE dados_cpd SET nacionalidade = 'Cabo_Verde'
                WHERE nacionalidade = '26';
UPDATE dados_cpd SET nacionalidade = 'Camaroes'
                WHERE nacionalidade = '27';
UPDATE dados_cpd SET nacionalidade = 'Chile'
                WHERE nacionalidade = '31';
UPDATE dados_cpd SET nacionalidade = 'China'
                WHERE nacionalidade = '32';
UPDATE dados_cpd SET nacionalidade = 'Colombia'
                WHERE nacionalidade = '37';
UPDATE dados_cpd SET nacionalidade = 'Congo'
                WHERE nacionalidade = '38';
UPDATE dados_cpd SET nacionalidade = 'Coreia_do_Sul'
                WHERE nacionalidade = '40';
UPDATE dados_cpd SET nacionalidade = 'Costa_do_Marfim'
                WHERE nacionalidade = '41';
UPDATE dados_cpd SET nacionalidade = 'Costa_Rica'
                WHERE nacionalidade = '42';
UPDATE dados_cpd SET nacionalidade = 'Equador'
                WHERE nacionalidade = '49';
UPDATE dados_cpd SET nacionalidade = 'Espanha'
                WHERE nacionalidade = '50';
UPDATE dados_cpd SET nacionalidade = 'Estados Unidos'
                WHERE nacionalidade = '51';
UPDATE dados_cpd SET nacionalidade = 'Franca'
                WHERE nacionalidade = '56';
UPDATE dados_cpd SET nacionalidade = 'Guine'
                WHERE nacionalidade = '64';
UPDATE dados_cpd SET nacionalidade = 'Guine-Bissau'
                WHERE nacionalidade = '65';
UPDATE dados_cpd SET nacionalidade = 'Guine-Equatorial'
                WHERE nacionalidade = '66';
UPDATE dados_cpd SET nacionalidade = 'Haiti'
                WHERE nacionalidade = '67';
UPDATE dados_cpd SET nacionalidade = 'Honduras'
                WHERE nacionalidade = '69';
UPDATE dados_cpd SET nacionalidade = 'India'
                WHERE nacionalidade = '76';
UPDATE dados_cpd SET nacionalidade = 'Ira'
                WHERE nacionalidade = '78';
UPDATE dados_cpd SET nacionalidade = 'Italia'
                WHERE nacionalidade = '83';
UPDATE dados_cpd SET nacionalidade = 'Malasia'

```

```

WHERE nacionalidade = '96';
UPDATE dados_cpd SET nacionalidade = 'Mexico'
WHERE nacionalidade = '104';
UPDATE dados_cpd SET nacionalidade = 'Mocambique'
WHERE nacionalidade = '105';
UPDATE dados_cpd SET nacionalidade = 'Nigeria'
WHERE nacionalidade = '112';
UPDATE dados_cpd SET nacionalidade = 'Paquistao'
WHERE nacionalidade = '118';
UPDATE dados_cpd SET nacionalidade = 'Paraguai'
WHERE nacionalidade = '119';
UPDATE dados_cpd SET nacionalidade = 'Peru'
WHERE nacionalidade = '120';
UPDATE dados_cpd SET nacionalidade = 'Polonia'
WHERE nacionalidade = '121';
UPDATE dados_cpd SET nacionalidade = 'Sao_Tome_e_Principe'
WHERE nacionalidade = '132';
UPDATE dados_cpd SET nacionalidade = 'Senegal'
WHERE nacionalidade = '133';
UPDATE dados_cpd SET nacionalidade = 'Suecia'
WHERE nacionalidade = '141';
UPDATE dados_cpd SET nacionalidade = 'Suica'
WHERE nacionalidade = '142';
UPDATE dados_cpd SET nacionalidade = 'Suriname'
WHERE nacionalidade = '143';
UPDATE dados_cpd SET nacionalidade = 'Togo'
WHERE nacionalidade = '147';
UPDATE dados_cpd SET nacionalidade = 'Trinidad_e_Tobago'
WHERE nacionalidade = '149';
UPDATE dados_cpd SET nacionalidade = 'Uruguai'
WHERE nacionalidade = '155';
UPDATE dados_cpd SET nacionalidade = 'Zimbabwe'
WHERE nacionalidade = '174';
UPDATE dados_cpd SET nacionalidade = 'Republica_Tcheca'
WHERE nacionalidade = '214';
UPDATE dados_cpd SET nacionalidade = 'Nao_Declarado'
WHERE nacionalidade = '';

```

-- País de nascimento dos alunos

```

UPDATE dados_cpd SET pais_nascimento = 'Africa_do_Sul'
WHERE pais_nascimento = '2';
UPDATE dados_cpd SET pais_nascimento = 'Angola'
WHERE pais_nascimento = '8';
UPDATE dados_cpd SET pais_nascimento = 'Argentina'
WHERE pais_nascimento = '11';
UPDATE dados_cpd SET pais_nascimento = 'Belgica'

```

```

WHERE pais_nascimento = '18';
UPDATE dados_cpd SET pais_nascimento = 'Birmania'
WHERE pais_nascimento = '19';
UPDATE dados_cpd SET pais_nascimento = 'Bolivia'
WHERE pais_nascimento = '20';
UPDATE dados_cpd SET pais_nascimento = 'Brasil'
WHERE pais_nascimento = '22';
UPDATE dados_cpd SET pais_nascimento = 'Cabo_Verde'
WHERE pais_nascimento = '26';
UPDATE dados_cpd SET pais_nascimento = 'Camaroes'
WHERE pais_nascimento = '27';
UPDATE dados_cpd SET pais_nascimento = 'Chile'
WHERE pais_nascimento = '31';
UPDATE dados_cpd SET pais_nascimento = 'China'
WHERE pais_nascimento = '32';
UPDATE dados_cpd SET pais_nascimento = 'Colombia'
WHERE pais_nascimento = '37';
UPDATE dados_cpd SET pais_nascimento = 'Congo'
WHERE pais_nascimento = '38';
UPDATE dados_cpd SET pais_nascimento = 'Coreia_do_Sul'
WHERE pais_nascimento = '40';
UPDATE dados_cpd SET pais_nascimento = 'Costa_do_Marfim'
WHERE pais_nascimento = '41';
UPDATE dados_cpd SET pais_nascimento = 'Costa_Rica'
WHERE pais_nascimento = '42';
UPDATE dados_cpd SET pais_nascimento = 'Equador'
WHERE pais_nascimento = '49';
UPDATE dados_cpd SET pais_nascimento = 'Espanha'
WHERE pais_nascimento = '50';
UPDATE dados_cpd SET pais_nascimento = 'Estados Unidos'
WHERE pais_nascimento = '51';
UPDATE dados_cpd SET pais_nascimento = 'Franca'
WHERE pais_nascimento = '56';
UPDATE dados_cpd SET pais_nascimento = 'Guine'
WHERE pais_nascimento = '64';
UPDATE dados_cpd SET pais_nascimento = 'Guine-Bissau'
WHERE pais_nascimento = '65';
UPDATE dados_cpd SET pais_nascimento = 'Guine-Equatorial'
WHERE pais_nascimento = '66';
UPDATE dados_cpd SET pais_nascimento = 'Haiti'
WHERE pais_nascimento = '67';
UPDATE dados_cpd SET pais_nascimento = 'Honduras'
WHERE pais_nascimento = '69';
UPDATE dados_cpd SET pais_nascimento = 'India'
WHERE pais_nascimento = '76';
UPDATE dados_cpd SET pais_nascimento = 'Ira'

```

```

                WHERE pais_nascimento = '78';
UPDATE dados_cpd SET pais_nascimento = 'Italia',
                WHERE pais_nascimento = '83';
UPDATE dados_cpd SET pais_nascimento = 'Japao',
                WHERE pais_nascimento = '86';
UPDATE dados_cpd SET pais_nascimento = 'Malasia',
                WHERE pais_nascimento = '96';
UPDATE dados_cpd SET pais_nascimento = 'Mexico',
                WHERE pais_nascimento = '104';
UPDATE dados_cpd SET pais_nascimento = 'Mocambique',
                WHERE pais_nascimento = '105';
UPDATE dados_cpd SET pais_nascimento = 'Nigeria',
                WHERE pais_nascimento = '112';
UPDATE dados_cpd SET pais_nascimento = 'Paquistao',
                WHERE pais_nascimento = '118';
UPDATE dados_cpd SET pais_nascimento = 'Paraguai',
                WHERE pais_nascimento = '119';
UPDATE dados_cpd SET pais_nascimento = 'Peru',
                WHERE pais_nascimento = '120';
UPDATE dados_cpd SET pais_nascimento = 'Polonia',
                WHERE pais_nascimento = '121';
UPDATE dados_cpd SET pais_nascimento = 'Portugal',
                WHERE pais_nascimento = '122';
UPDATE dados_cpd SET pais_nascimento = 'Sao_Tome_e_Principe',
                WHERE pais_nascimento = '132';
UPDATE dados_cpd SET pais_nascimento = 'Senegal',
                WHERE pais_nascimento = '133';
UPDATE dados_cpd SET pais_nascimento = 'Suecia',
                WHERE pais_nascimento = '141';
UPDATE dados_cpd SET pais_nascimento = 'Suica',
                WHERE pais_nascimento = '142';
UPDATE dados_cpd SET pais_nascimento = 'Suriname',
                WHERE pais_nascimento = '143';
UPDATE dados_cpd SET pais_nascimento = 'Togo',
                WHERE pais_nascimento = '147';
UPDATE dados_cpd SET pais_nascimento = 'Trinidad_e_Tobago',
                WHERE pais_nascimento = '149';
UPDATE dados_cpd SET pais_nascimento = 'Uruguai',
                WHERE pais_nascimento = '155';
UPDATE dados_cpd SET pais_nascimento = 'Inglaterra',
                WHERE pais_nascimento = '162';
UPDATE dados_cpd SET pais_nascimento = 'Zimbabue',
                WHERE pais_nascimento = '174';
UPDATE dados_cpd SET pais_nascimento = 'Republica_Tcheca',
                WHERE pais_nascimento = '214';
UPDATE dados_cpd SET pais_nascimento = 'Nao_Declarado'

```

```
WHERE pais_nascimento = '';
```

Listing C.4: *Query* de atualização da data de nascimento

```
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Jan' , '01');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Feb' , '02');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Mar' , '03');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Apr' , '04');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'May' , '05');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Jun' , '06');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Jul' , '07');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Aug' , '08');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Sep' , '09');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Oct' , '10');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Nov' , '11');
UPDATE dados_cpd
    SET data_nascimento = REPLACE(data_nascimento , 'Dec' , '12');
UPDATE dados_cpd SET data_nascimento =
    CONCAT('19', RIGHT(data_nascimento, 2), '-',
           LEFT(RIGHT(data_nascimento, 5), 2), '-',
           IF(CHAR_LENGTH(data_nascimento) = 8, LEFT(data_nascimento, 2),
              CONCAT('0', LEFT(data_nascimento, 1))));
```

Listing C.5: Código de tradução dos campos restantes

```
-- Tipos de cotas de entrada na UnB
UPDATE dados_cpd SET cota = 'Universal' WHERE cota = '0';
UPDATE dados_cpd SET cota = 'Negro' WHERE cota = '1';
UPDATE dados_cpd SET cota = 'Indigena' WHERE cota = '2';
UPDATE dados_cpd SET cota = 'Escola_Publica_Baixa_Renda__PPI'
                                         WHERE cota = '3';
UPDATE dados_cpd
    SET cota = 'Escola_Publica_Baixa_Renda__Nao_PPI'
                                         WHERE cota = '4';
UPDATE dados_cpd SET cota = 'Escola_Publica_Alta_Renda__PPI'
                                         WHERE cota = '5';
UPDATE dados_cpd SET cota = 'Escola_Publica_Alta_Renda__Nao_PPI'
```

WHERE cota = '6';

— Raca do aluno

UPDATE dados_cpd **SET** raca_cor = 'Nao_Declarado'
WHERE raca_cor = '0';
UPDATE dados_cpd **SET** raca_cor = 'Branca' **WHERE** raca_cor = '1';
UPDATE dados_cpd **SET** raca_cor = 'Preta' **WHERE** raca_cor = '2';
UPDATE dados_cpd **SET** raca_cor = 'Parda' **WHERE** raca_cor = '3';
UPDATE dados_cpd **SET** raca_cor = 'Amarela' **WHERE** raca_cor = '4';
UPDATE dados_cpd **SET** raca_cor = 'Indigena' **WHERE** raca_cor = '5';
UPDATE dados_cpd **SET** raca_cor = 'Nao_dispoe_da_informacao'
WHERE raca_cor = '6';
UPDATE dados_cpd **SET** raca_cor = 'Nao_cadastrada'
WHERE raca_cor = '7';

— Forma de Ingresso do aluno na UnB

UPDATE dados_cpd **SET** forma_ingresso = 'Vestibular'
WHERE forma_ingresso = '1';
UPDATE dados_cpd **SET** forma_ingresso = 'Transferencia_Obrigatoria'
WHERE forma_ingresso = '2';
UPDATE dados_cpd **SET** forma_ingresso = 'Transferencia_Facultativa'
WHERE forma_ingresso = '3';
UPDATE dados_cpd
 SET forma_ingresso = 'Portador_Diploma_Curso_Superior'
 WHERE forma_ingresso = '4';
UPDATE dados_cpd **SET** forma_ingresso = 'Acordo_Cultural--PEC'
 WHERE forma_ingresso = '5';
UPDATE dados_cpd **SET** forma_ingresso = 'Convenio--Int'
 WHERE forma_ingresso = '6';
UPDATE dados_cpd **SET** forma_ingresso = 'Matricula_Cortesia'
 WHERE forma_ingresso = '7';
UPDATE dados_cpd
 SET forma_ingresso = 'PAS--Programa_de_Avaliacao_Seriada'
 WHERE forma_ingresso = '17';
UPDATE dados_cpd **SET** forma_ingresso = 'Convenio--Andifes'
 WHERE forma_ingresso = '20';
UPDATE dados_cpd **SET** forma_ingresso = 'PEC-G_Peppfol'
 WHERE forma_ingresso = '24';
UPDATE dados_cpd **SET** forma_ingresso = 'Enem'
 WHERE forma_ingresso = '27';
UPDATE dados_cpd **SET** forma_ingresso = 'Dupla_Habilitacao'
 WHERE forma_ingresso = '50';

— Forma de saida do aluno da UnB

UPDATE dados_cpd **SET** forma_saida = 'Aluno_Regular'
WHERE forma_saida = '0';

```

UPDATE dados_cpd SET forma_saida = 'Formatura'
                                         WHERE forma_saida = '1';
UPDATE dados_cpd
                                         SET forma_saida = 'Desligamento_/_Rendimento_Academico'
                                         WHERE forma_saida = '2';
UPDATE dados_cpd SET forma_saida = 'Desligamento_/_Jubilamento'
                                         WHERE forma_saida = '3';
UPDATE dados_cpd
                                         SET forma_saida = 'Desligamento_/_Falta_Documentacao'
                                         WHERE forma_saida = '4';
UPDATE dados_cpd
                                         SET forma_saida = 'Desligamento_/_Forca_de_Convenio'
                                         WHERE forma_saida = '5';
UPDATE dados_cpd SET forma_saida = 'Transferencia'
                                         WHERE forma_saida = '6';
UPDATE dados_cpd SET forma_saida = 'Deligamento_Voluntario'
                                         WHERE forma_saida = '7';
UPDATE dados_cpd SET forma_saida = 'Falecimento'
                                         WHERE forma_saida = '9';
UPDATE dados_cpd SET forma_saida = 'Deligamento_Decisao_Judicial'
                                         WHERE forma_saida = '12';
UPDATE dados_cpd SET forma_saida = 'Deligamento_/_Abandono'
                                         WHERE forma_saida = '16';
UPDATE dados_cpd
                                         SET forma_saida = 'Deligamento_/_Nao_cumpriu_condicao'
                                         WHERE forma_saida = '17';
UPDATE dados_cpd SET forma_saida =
                                         'Reprovou_3_vezes_na_mesma_disciplina_/_obrigatoria'
                                         WHERE forma_saida = '20';
UPDATE dados_cpd SET forma_saida = 'Novo_Vestibular'
                                         WHERE forma_saida = '21';
UPDATE dados_cpd
                                         SET forma_saida = 'Vestibular_/_para_outra_Habilitacao'
                                         WHERE forma_saida = '50';
UPDATE dados_cpd SET forma_saida = 'Anulacao_de_Registro'
                                         WHERE forma_saida = '55';

```

— Forma de Ingresso na opcao escolhida

```

UPDATE dados_cpd SET forma_ingresso_opcao = 'Vestibular'
                                         WHERE forma_ingresso_opcao = '1';
UPDATE dados_cpd
                                         SET forma_ingresso_opcao = 'Transferencia_Obrigatoria'
                                         WHERE forma_ingresso_opcao = '2';
UPDATE dados_cpd
                                         SET forma_ingresso_opcao = 'Transferencia_Facultativa'
                                         WHERE forma_ingresso_opcao = '3';

```

```

UPDATE dados_cpd
    SET forma_ingresso_opcao = 'Portador_Diploma_Curso_Superior'
        WHERE forma_ingresso_opcao = '4';
UPDATE dados_cpd
    SET forma_ingresso_opcao = 'Acordo_Cultural--PEC'
        WHERE forma_ingresso_opcao = '5';
UPDATE dados_cpd SET forma_ingresso_opcao = 'Convenio--Int'
        WHERE forma_ingresso_opcao = '6';
UPDATE dados_cpd SET forma_ingresso_opcao = 'Matricula_Cortesia'
        WHERE forma_ingresso_opcao = '7';
UPDATE dados_cpd SET forma_ingresso_opcao = 'Duplo_Curso'
        WHERE forma_ingresso_opcao = '15';
UPDATE dados_cpd
    SET forma_ingresso_opcao = 'PAS--Programa_de_Avaliacao_Seriada'
        WHERE forma_ingresso_opcao = '17';
UPDATE dados_cpd SET forma_ingresso_opcao = 'Convenio--Andifes'
        WHERE forma_ingresso_opcao = '20';
UPDATE dados_cpd SET forma_ingresso_opcao = 'PEC-G_Peppfol'
        WHERE forma_ingresso_opcao = '24';
UPDATE dados_cpd SET forma_ingresso_opcao = 'Enem'
        WHERE forma_ingresso_opcao = '27';
UPDATE dados_cpd SET forma_ingresso_opcao = 'Dupla_Habilitacao'
        WHERE forma_ingresso_opcao = '50';
UPDATE dados_cpd
    SET forma_ingresso_opcao = 'Mudanca_de_Habilitacao'
        WHERE forma_ingresso_opcao = '51';
UPDATE dados_cpd SET forma_ingresso_opcao = 'Mudanca_de_Curso'
        WHERE forma_ingresso_opcao = '52';

-- Forma de saida da opcao escolhida
UPDATE dados_cpd SET forma_saida_opcao = 'Aluno_Regular'
        WHERE forma_saida_opcao = '0';
UPDATE dados_cpd SET forma_saida_opcao = 'Formatura'
        WHERE forma_saida_opcao = '1';
UPDATE dados_cpd
    SET forma_saida_opcao = 'Desligamento--Rendimento_Academico'
        WHERE forma_saida_opcao = '2';
UPDATE dados_cpd
    SET forma_saida_opcao = 'Desligamento--Jubilamento'
        WHERE forma_saida_opcao = '3';
UPDATE dados_cpd
    SET forma_saida_opcao = 'Desligamento--Falta_Documentacao'
        WHERE forma_saida_opcao = '4';
UPDATE dados_cpd
    SET forma_saida_opcao = 'Desligamento--Forca_de_Convenio'
        WHERE forma_saida_opcao = '5';

```

```

UPDATE dados_cpd SET forma_saida_opcao = 'Transferencia'
    WHERE forma_saida_opcao = '6';
UPDATE dados_cpd SET forma_saida_opcao = 'Deligamento_Voluntario'
    WHERE forma_saida_opcao = '7';
UPDATE dados_cpd SET forma_saida_opcao = 'Falecimento'
    WHERE forma_saida_opcao = '9';
UPDATE dados_cpd
    SET forma_saida_opcao = 'Deligamento_Decisao_Judicial'
        WHERE forma_saida_opcao = '12';
UPDATE dados_cpd SET forma_saida_opcao = 'Deligamento_Abandono'
    WHERE forma_saida_opcao = '16';
UPDATE dados_cpd
    SET forma_saida_opcao = 'Deligamento_Nao_cumpriu_condicao'
        WHERE forma_saida_opcao = '17';
UPDATE dados_cpd SET forma_saida_opcao =
    'Reprovou_3_vezes_na_mesma_disciplina_obrigatoria'
        WHERE forma_saida_opcao = '20';
UPDATE dados_cpd SET forma_saida_opcao = 'Novo_Vestibular'
    WHERE forma_saida_opcao = '21';
UPDATE dados_cpd
    SET forma_saida_opcao = 'Vestibular_para_outra_Habilitacao'
        WHERE forma_saida_opcao = '50';
UPDATE dados_cpd SET forma_saida_opcao = 'Mudanca_de_Curso'
    WHERE forma_saida_opcao = '52';
UPDATE dados_cpd SET forma_saida_opcao = 'Anulacao_de_Registro'
    WHERE forma_saida_opcao = '55';

```

— Classificacao dos alunos por idade

```

UPDATE dados_cpd SET classificacao_idade = 'Ate_18_anos'
    WHERE classificacao_idade = '1';
UPDATE dados_cpd SET classificacao_idade = 'De_19_a_24_anos'
    WHERE classificacao_idade = '2';
UPDATE dados_cpd SET classificacao_idade = 'De_25_a_29_anos'
    WHERE classificacao_idade = '3';
UPDATE dados_cpd SET classificacao_idade = 'De_30_a_34_anos'
    WHERE classificacao_idade = '4';
UPDATE dados_cpd SET classificacao_idade = 'De_35_a_39_anos'
    WHERE classificacao_idade = '5';
UPDATE dados_cpd SET classificacao_idade = 'De_40_a_44_anos'
    WHERE classificacao_idade = '6';
UPDATE dados_cpd SET classificacao_idade = 'De_45_a_49_anos'
    WHERE classificacao_idade = '7';
UPDATE dados_cpd SET classificacao_idade = 'De_50_a_54_anos'
    WHERE classificacao_idade = '8';
UPDATE dados_cpd SET classificacao_idade = '55_anos_e_mais'
    WHERE classificacao_idade = '9';

```

Listing C.6: Seleção dos dados de menções dos alunos

```

SELECT ano ,
        semestre ,
        c . nome _ disciplina ,
        d . ‘ sigla ‘ ,
        IF( mencao = ‘ CC ’ , ‘ A ’ , turma) AS
        turma ,
        creditos ,
        mencao ,
        IF( mencao = ‘ SS ’ , 9 , IF( mencao = ‘ MS ’ , 7 ,
                                         IF( mencao = ‘ MM ’ , 5 ,
                                         IF( mencao = ‘ MI ’ , 3 ,
                                         IF( mencao = ‘ II ’ , 1 , 0 )))) ) AS
        nota ,
        dcp . sexo ,
        cota ,
        tipo _ escola ,
        raca _ cor ,
        forma _ ingresso ,
        forma _ saida ,
        IF( ira = 0 , ira , ira / POW(10 , CHAR_LENGTH( ira ) - 1)) AS
        ira ,
        IF( codigo _ curso = 370 , ‘ Ciencia _ da _ Computacao ’ ,
        IF( codigo _ curso = 906 , ‘ Computacao ’ ,
                                         ‘ Engenharia _ da _ Computacao ’)) AS
        curso ,
        ano _ ingresso - YEAR( data _ nascimento ) AS
        idade
FROM      dados _ cpd _ disciplinas dcp
INNER JOIN dados _ cpd dc
            ON dcp . matricula = dc . matricula
INNER JOIN cursos c
            ON dcp . codigo _ disciplina = c . codigo _ disciplina
INNER JOIN departamentos d
            ON c . ‘ departamento ‘ = d . ‘ codigo _ departamento ‘
WHERE      mencao NOT IN ( ‘ AP ’ , ‘ DP ’ );

```

Referências

- [1] Janet Abbate. *Recoding Gender: Women's Changing Participation in Computing*. MIT Press, 2012. [4](#)
- [2] Sapna Cheryan, Victoria Plaut, Paul Davies, and Claude Steele. Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, 97(6):1045–1060, 2009. [4](#)
- [3] Krzysztof Cios, Witold Pedrycz, Roman Swiniarski, and Lucasz Kurgan. *Data Mining: A Knowledge Discovery Approach*. Springer, 2007. [6](#)
- [4] Joel Cooper and Kimberlee Weaver. *Gender and Computers: Understanding the Digital Divide*. Psychology Press, 2003.
- [5] Centro de Processamento de Dados / UnB. <http://www.cpd.unb.br>, acessado em 13/10/2013. [1](#)
- [6] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37, 1996. [9](#)
- [7] Florin Gorunescu. *Data Mining: Concepts, Models and Techniques*, volume 12. Springer, 1 edition, 2011. [vii, 6, 8, 9, 11, 12, 13, 14, 15](#)
- [8] Jiawei Han, Micheline Kamber, and Jian Pei. *Data mining : concepts and techniques*. Morgan Kaufmann, 225 Wyman Street, Waltham, MA 02451, USA, 3 edition, 2012. [vii, 5, 6, 7, 8, 9, 10, 11](#)
- [9] Mehmed Kantardzic. *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley and Sons, Inc., 2 edition, 2011. [6](#)
- [10] CIC/UnB. Meninas na Computação. <http://meninas.cic.unb.br>, acessado em 13/10/2013. [vii, 1, 2](#)
- [11] Daryl Pregibon. Data mining. *Statistical Computing and Graphics Newsletter*, 7(3):8, December 1996. [6](#)
- [12] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, 23(4):3–13, 2000. [9](#)
- [13] Jaideep Srivasta, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1:12–23, Janeiro 2000. [8](#)

- [14] Dorian Stoilescu and Douglas McDougall. Gender digital divide and challenges in undergraduate computer science programs. *Canadian Journal of Education*, 34(1):308–333, 2011. 4
- [15] Ian Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 2005.
- [16] Lan Yi, Bing Liu, and Xiaoli Li. Eliminating noisy information in web pages for data mining. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 296–305. ACM, Agosto 2003. 9