KLEYTON FAZOLIN

TRATAMENTO TEMPORAL EM MINERAÇÃO DE DADOS EDUCACIONAIS PARA FIDELIZAÇÃO DE ESTUDANTES

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

Área de concentração: *Engenharia de Sistemas Computacionais*

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

Curitiba PR
2017

Agradecimentos

À Deus por ter me proporcionar tudo o que tenho e que sou até hoje. ELE tem me acompanhado, guiado, dado inspiração e saúde ao longo de todos os desafios de minha vida.

À minha esposa e super parceira Patricia, que esteve sempre ao meu lado nos bons e maus momentos. Seu apoio nos momentos de dificuldades e incentivo foi o que me motivou a finalizar mais esta etapa de nossas vidas. Te amo.

Aos meus filhos Ana Beatriz e Gabriel, que sempre me recebiam com muito amor e carinho mesmo tendo que ceder vários momentos de brincadeiras para que eu pudesse concluir.

Ao meu orientador Prof. Dr. Celso Antônio Alves Kaestner, pelo constante apoio, paciência e disposição na condução de todo o mestrado.

Ao meu coorientador Prof. Dr. Robinson Vida Noronha, pelas valiosas contribuições.

Ao Instituto Adventista de Tecnologia – IATec, pelo apoio no qual sem ele, não seria viável o início e conclusão deste mestrado.

Aos demais professores do Departamento Acadêmico de Informática da UTFPR e a todos que direta e indiretamente contribuíram para a realização deste sonho.

Resumo

O tratamento temporal tem se revelado importante em muitos problemas de mineração de dados em que a base de dados é formada por dados coletados historicamente [Romero e Ventura 2007]. Um exemplo desta situação ocorre em instituições de ensino, onde os dados históricos dos alunos - tais como o desempenho escolar e a situação financeira - vem sendo adquiridos paulatinamente ao longo do tempo [Romero e Ventura 2007]. Este trabalho apresenta uma proposta de criação de atributos temporais com o objetivo de auxiliar a previsão da evasão de alunos de Ensino Fundamental em escolas particulares, tratada como um problema de classificação. A fidelização e retenção de alunos em instituições de ensino se tornou um dos maiores desafios para a área de gestão destas instituições [Lin 2012]. Uma solução promissora para alcançar esse objetivo é o uso da mineração de dados educacionais, para a identificação de padrões que auxiliem nas tomadas de decisões. Para a realização dos experimentos, os dados de 15.753 alunos da Rede Educacional Adventista – uma das maiores redes educacionais do mundo [“Educação Adventista” 2016] – foram extraídos e utilizados. Após a aplicação dos algoritmos de classificação, verificou-se que o classificador baseado em instâncias KNN conseguiu a melhor acurácia antes do uso dos novos atributos temporais criados, porém o melhor algoritmo para efetuar previsão da evasão no contexto desta pesquisa foi a Árvore de Decisão J4.8, pois permite a interpretação dos fatores que levaram ao resultado final. Os resultados mostram que a abordagem é viável, tendo-se obtido uma acurácia de até 96,57% utilizando o algoritmo J48 e um aumento de 14,39% na acurácia do classificador KNN com o uso dos atributos temporais.

Palavras-chave: Mineração de Dados Educacionais, Criação de Atributos Temporais, Mineração de dados para Fidelização de Alunos.

Abstract

The creation of temporal attributes has proved important in many data mining problems in that the database is formed by data collected historically [Romero e Ventura 2007]. An example of this situation occurs in educational institutions, where the historical data of students – such as school performance and financial situation – has been gradually acquired over time [Romero e Ventura 2007]. This paper presents a proposal for the creation of temporal attributes with the purpose of helping to predict the avoidance of elementary school students in private schools, treated as a classification problem. The loyalty and retention of students in educational institutions has become one of the greatest challenges for the management area of these institutions [Lin 2012]. A promising solution to achieve this goal is the use of educational data mining to identify patterns that aid in decision making. For the experiments, the data of 15,753 students of the Adventist Educational Network – one of the largest educational networks in the world ["Educação Adventista" 2016]– were employed. After the application of the classification algorithms, it was verified that the instance-based KNN classifier obtained the best accuracy before the use of the time attributes created, but the best algorithm to predict the avoidance in the context of this research was the Decision Tree J4.8 algorithm, because it allows the interpretation of the factors that led to the final result. The results show that the approach is feasible, obtaining an accuracy of up to 97.87% in the experiments performed and a gain of up to 14.39% in the accuracy when using the KNN with temporal attributes.

Keywords: Educational Data Mining, Feature Creation, Student Loyalty and Retention.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

BDT	Banco de Dados Temporal
DM	Data Mining
EDM	Educational Data Mining
FN	Falso Negativo
FP	Falso Positivo
GPA	Ponto médio de baixo grau
IBGE	Instituto Brasileiro de Geografia e Estatística
KDD	Knowledge Discovery in Databases
KNN	K Nearest-Neighbors
N	Negativo
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
P	Positivo
SGBD	Sistema de Gerenciamento de Banco de Dados
SSE	Sistema de Secretaria Escolar
SVM	Support Vector Machines
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
WEKA	Waikato Environment for Knowledge Analysis

Capítulo 1

Introdução

Com o avanço da tecnologia uma enorme quantidade de dados vêm sendo coletada e armazenada das mais diversas maneiras em uma escala cada vez maior [Fayyad et al. 1996]. Em muitas aplicações estes dados são coletados ao longo do tempo, que vão sendo sucessivamente incrementados formando uma base de dados temporal.

Os dados se constituem em valiosas fontes de informações que podem ser analisadas quando da ocorrência de eventos específicos, ou em conjuntos de eventos relacionados a particularidades temporais. Pode se observar que estas análises podem ser úteis para abstrair informações implícitas a partir dos dados originais, e também para predizer o comportamento futuro de um processo anteriormente monitorado.

Devido a essa grande quantidade de dados, especialistas de diversas áreas estão tendo dificuldades para provar ou confirmar hipóteses através de planilhas de dados ou relatórios operacionais. A partir do crescimento do volume de dados armazenados, gera-se a necessidade urgente de técnicas e ferramentas que transformem esses dados em informação útil da forma mais rápida e automática possível. A área de Mineração de Dados (*Data Mining* - MD) vem para ajudar a identificar e desenvolver técnicas e métodos para converter dado em informação [Fayyad et al. 1996].

Assim como em várias outras áreas, a área educacional também passa por este tipo de problema. Com essa gigantesca gama de dados sobre os alunos e seu histórico, tentar identificar uma informação útil é uma tarefa extremamente difícil e complexa [Marquez-Vera et al. 2013]. Uma solução que tem dado resultados satisfatórios é o uso da técnica chamada Mineração de Dados Educacionais (Educational Data Mining - EDM) [Romero e Ventura 2007]. Esta nova área tem como foco concentrar-se no desenvolvimento de métodos cujo o principal objetivo é entender melhor os estudantes e o ambiente educacional que o cerca [Romero e Ventura 2010].

A grande maioria das pesquisas que aplicam o EDM estão relacionadas a problemas de estudantes no contexto de ensino superior em cursos presenciais [Marquez-Vera et al. 2013] e/ou ensino a distância [Lykourentzou et al. 2009]. Em contrapartida, pouquíssimas pesquisas foram aplicadas no contexto do ensino médio e fundamental. As poucas pesquisas ligadas a este contexto de ensino tratam apenas de aplicar métodos estatísticos, sem o uso das técnicas de mineração de dados (Data Mining – DM) [Parker 1999].

Um dos pontos que pode ser abordado na EDM é o problema da fidelização e retenção de alunos. Neste trabalho, o termo “retenção” será utilizado para indicar a permanência do aluno na instituição de ensino. A retenção de alunos em escolas particulares tornou-se um dos problemas mais desafiadores para os tomadores de decisões em instituições de ensino privado [Lin 2012]. Com o aumento da concorrência e o alto custo operacional, têm-se estudado maneiras inteligentes e efetivas para se entender o que leva um estudante de ensino médio ou fundamental a permanecer na mesma escola até completar sua carreira acadêmica inicial [Dwayne D. Gremler 1999].

Muitas vezes as instituições concentram seus esforços e recursos visando prioritariamente o recrutamento de novos alunos (pois o recrutamento é essencial para o aumento de alunos matriculados), porém em vários casos não se tem um plano de ação para reter os alunos já matriculados, impactando diretamente na receita da instituição [Fike e Fike 2008]. Na área educacional, a retenção de alunos matriculados é tão ou mais importante do que se montar estratégias para atrair e matricular novos alunos [Kotler e Armstrong 2012].

Com base nas pesquisas do ensino superior, a retenção de alunos afeta diretamente a taxa de formandos. Faculdades com taxas maiores de retenção de calouros tendem a ter no período de quatro anos uma taxa superior de graduados comparado a instituições com alto índice de abandono nos anos iniciais. Nos Estados Unidos a média de retenção é de aproximadamente 55% dos alunos, porém em algumas instituições a taxa pode chegar a menos de 20% nos casos de cursos envolvendo pós-graduação [Druzdzel e Glymour 1994]. Um dos primeiros estudos relacionado a área de retenção, datado no ano de 1982, foi feito nos Estados Unidos por Tinto [Tinto 1982] onde se mostra que a taxa de abandono naquele país se manteve constante entre 45% a 52% nos últimos 100 anos (excluindo o período da segunda guerra mundial).

Olhando pela perspectiva sociológica do problema, Tinto [Tinto 1975], [Tinto e Bean 1988] propõe um modelo de integração de alunos onde se explora as interações conjuntas entre sistemas acadêmicos e sociais para determinar se um aluno irá persistir na instituição. Em sua opinião, a retenção dos alunos está positivamente relacionada ao compromisso acadêmico/social do estudante na graduação. Este compromisso é diretamente dependente de atributos pré-universitários e sociais como por exemplo: status socioeconômico, habilidades acadêmicas, raça e gênero [Tinto 1975].

A fidelização de estudantes pode ser associada à tarefa da mineração de dados chamada classificação, no qual o principal objetivo é a associação de uma classe a cada elemento a partir de um conjunto de atributos pertinentes ao mesmo sujeito. Neste trabalho, a tarefa em questão nos mostra que cada elemento corresponde a um aluno e

as classes consideradas correspondem a “Rematriculou” ou “Não Rematriculou” na etapa seguinte do curso.

Problemas como o citado se caracterizam por possuírem bases de dados temporais, onde o fator tempo aparece como elemento que condiciona a evolução dos dados. Desta forma busca-se neste trabalho encontrar formas de incorporar esta questão temporal ao processo de Mineração de Dados, de modo a encontrar uma solução para a tarefa alvo, melhorando a acurácia dos algoritmos de mineração utilizados, auxiliando assim nas tomadas de decisões pertinentes a fidelização. Neste contexto a incorporação do tratamento temporal no processo da Mineração de Dados busca encontrar elementos de simples interpretação que possam ser utilizados como indicadores de uma potencial situação de evasão/fidelização, uma vez que alguns algoritmos nos mostram o como chegaram ao resultado da classe.

No contexto da mineração de dados, a criação de atributos visa criar novos atributos baseados em outros já existentes de modo que informações importantes sejam extraídas em um conjunto de dados. Buscou-se criar novos atributos que representassem de forma adequada e simplificada a evolução temporal de elementos originais da base de dados, proporcionando assim sua simples interpretação.

Como indicado anteriormente, para validar a proposta será utilizada como aplicação o estudo da retenção de estudantes focado na mineração de dados em bases educacionais. Usando estes dados será proposto um método de criação de atributos para classificação que visa identificar padrões para análise da fidelização de estudantes em escolas privadas do ensino fundamental, a partir de informações socioeconômicas e financeiras, aplicando algoritmos de classificação para gerar indicadores com a finalidade de auxiliar os gestores educacionais e financeiros nas tomadas de decisão.

1.1 Contextualização

O processo de descoberta de conhecimento em banco de dados (Knowledge Discovery in Databases – KDD) segundo Fayyad “é o processo não trivial de identificação de padrões, a partir de dados, que sejam válidos, novos, potencialmente úteis e compreensíveis” [Fayyad et al. 1996]. O problema básico tratado pelo KDD é o mapeamento dos dados brutos convertendo-os em informações usuais e úteis [Tan et al. 2006]. Seguindo a ideia de Fayyad, as informações obtidas devem ser novas, compreensivas e úteis. Ao final do processo o modelo gerado necessita trazer algum benefício novo, que possa ser compreendido de maneira rápida para uma possível tomada de decisão [Witten et al. 2011].

Contudo a definição dos termos KDD e DM ainda não é um consenso na

literatura. Wang [Wang 2005] e Han [Han et al. 2011] consideram que ambos são sinônimos. Porém para Cios [Cios et al. 2007] e o próprio Fayyad [Fayyad et al. 1996] dizem que o processo como um todo é o KDD e que a DM é apenas uma atividade dentro do processo. Todos concordam que o processo de mineração deve ser dividido em algumas etapas como ilustrado na Figura 1.

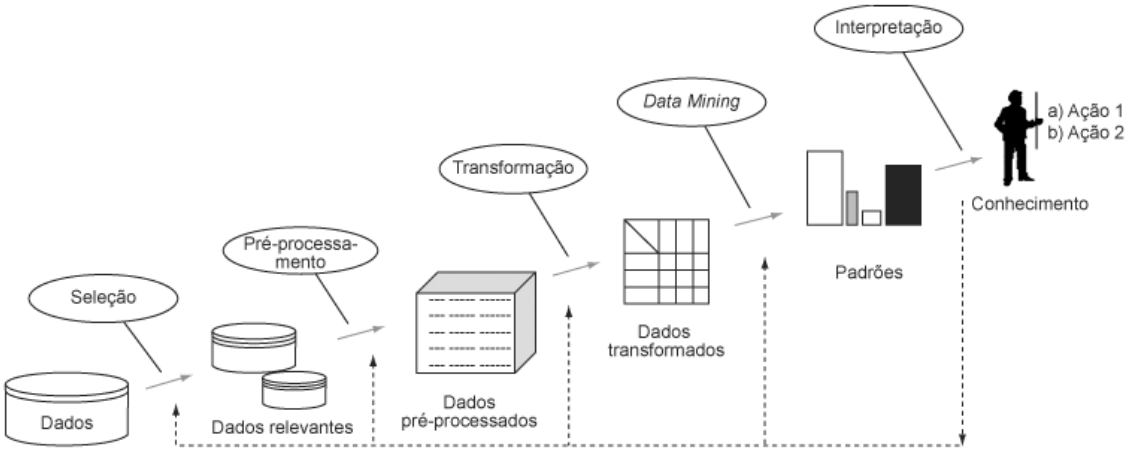

Figura 1: Fases do processo de um KDD [Fayyad et al. 1996].

Como listado na Figura 1, temos as etapas de Seleção, Pré-processamento, Transformação, Mineração de Dados para se chegar em um novo conhecimento.

Uma ferramenta muito útil no processo de KDD é o *Data Warehousing*. O *Data Warehousing* trabalha com a coleta e limpeza dos dados de bancos transacionais (OLTP) convertendo-os para uma estrutura analítica (OLAP) que permite uma melhor compreensão dos dados [Fayyad et al. 1996]. Se um *Data Warehouse* (a definição será detalhada na seção 2.4) não estiver presente, muitas etapas que envolvem o *Data Warehousing* terão que ser realizadas na preparação da mineração de dados [Witten et al. 2011].

1.2 Motivação

A fidelização de estudantes faz parte dos debates e ponderações do dia-a-dia da educação brasileira e mundial, ocupando espaço de relevância no cenário das políticas educacionais em instituições públicas e privadas. Sobre este assunto é encontrado diversos trabalhos nas áreas de marketing [Fabio Bergamo 2011], educação [Tinto e Bean 1988], psicologia [Tinto 1982] e computação [Eckert e Suénaga 2015].

Somente nos Estados Unidos, as receitas perdidas para cada 10 alunos que abandonam os cursos ainda no primeiro semestre de cada ano podem chegar a $326.811 (trezentos e vinte e seis mil oitocentos e onze dólares) [Tillman e Burns 2000]. A taxa média de retenção de alunos em 2010 nos Estados Unidos para o ensino superior foi de 77,1% [NCHEMS 2010]. Porém, este problema não se limita apenas a instituições

norte-americanas, sendo encontrado problemas semelhantes em outros países. No Brasil a taxa de evasão chega a 51% para o ensino superior gerando um grande desafio aos gestores educacionais, levando em conta que este fator pode comprometer a vida financeira de uma instituição privada [Fabio Bergamo 2011].

A motivação para esta pesquisa é o estudo e criação de novos atributos visando gerar indicadores de fácil interpretação para fornecer aos gestores da área educacional/financeira um prenúncio da possibilidade de fidelização/evasão. Neste trabalho será aplicado uma abordagem computacional para o problema, utilizando EDM, pois é um assunto ainda pouco explorado pela literatura acadêmica.

1.3 Objetivo geral

O objetivo geral deste trabalho é gerar indicadores de padrões através da técnica de criação de atributos em mineração de dados, a partir das informações disponíveis nas bases de dados acadêmicas, usando a tarefa de classificação para gerar uma representação computacional sobre a fidelização, com o intuito de auxiliar na tomada de decisão de gestores educacionais/financeiros focando na fidelização de alunos do ensino fundamental em instituições privadas.

1.4 Objetivos específicos

Para atingir o objetivo geral, são considerados os seguintes objetivos específicos:

1. Criar um *Data Warehouse* com dados acadêmicos para ajudar na análise dos dados e posteriormente aplicação dos algoritmos de mineração.

2. Desenvolver um modelo para a extração de padrões com a finalidade de analisar a retenção de estudantes do ensino fundamental.

3. Criar atributos e indicadores baseados na evolução histórica dos dados de pagamentos das mensalidades dos alunos com o intuito de maximizar a assertividade na predição da retenção escolar.

4. Testar qual a performance em relação a acurácia dos algoritmos de classificação no problema alvo e qual a melhora proporcionada pela adição dos atributos temporais.

Além disto, considerando as grandes necessidades dos gestores da área educacional, os conceitos envolvidos na dissertação e seu resultado, podem ajudar no

entendimento dos assuntos relacionados a retenção, auxiliando nos processos de tomada de decisões, contribuindo assim para as pesquisas na área de EDM.

1.5 Estrutura da Dissertação

O Capítulo 2 apresenta a revisão bibliográfica dos principais conceitos na área de EDM com foco em aspectos computacionais dos conceitos de discretização de dados, criação de atributos, balanceamento de dados, *data warehousing*, atributos temporais, algoritmos de classificação e métricas para avaliação de performance para os classificadores.

O Capítulo 3 apresenta os métodos e materiais propostos neste trabalho. Neste capítulo é apresentado um estudo preliminar sobre a fidelização de alunos e por sequencia o método proposta para a criação de atributos temporais utilizado para realizar inferências com a finalidade de auxiliar os gestores na tomada de decisões.

O Capítulo 4 apresenta as diversas implementações e experimentos realizados para a execução dos algoritmos de classificação. Também é apresentada a análise dos resultados obtidos através dos experimentos realizados.

Finalmente, no capítulo 5 são apresentadas as conclusões finais e sugestões que vislumbram aspectos que podem ser abordados em trabalhos futuros. Ainda nesse capítulo são apontadas possíveis contribuições e também algumas limitações deste trabalho.

Capítulo 2
Fundamentação Teórica

Este capítulo apresenta a revisão bibliográfica dos conceitos de mineração de dados, *data warehousing*, discretização de dados, desbalanceamento de dados, seleção de atributos, classificação e finalmente revisão da literatura.

2.1 Bases Temporais

Os Sistemas de Gerenciamento de Banco de Dados – SGBD atuais, possuem somente a possibilidade de armazenamento de um instante de tempo nos registros. Sendo assim, podemos definir que os SGBD nos permitem armazenar apenas informações do estado presente [Serra e Zárate 2015].

Para se armazenar informações temporais, é necessário criar atributos adicionais para cada entidade. Este gerenciamento normalmente fica a cargo do administrador desta base de dados, pois os SGBD comerciais, não possuem uma implementação nativa com esta finalidade [Serra e Zárate 2015]. Porém, há uma necessidade explícita de muitas aplicações de armazenarem não apenas informações do estado corrente de um registro, mas o seu estado passado, presente e futuro. Visando atender essa necessidade, foi criado o Banco de Dados Temporais – BDT, onde sua modelagem é feita para representar as informações com base em elementos temporais [Dean e Mcdermott 1987].

2.1.1 Conceitos de Tempo

A área de BDT tem como principal objetivo especificar os aspectos relacionados a modelagem, recuperação e armazenamento dos dados temporais. Os rótulos temporais relacionados aos registros das tabelas, são rotulados por seu estado: “presente”, ‘passado”, “futuro”, fazendo com que os BDTs se diferencie dos bancos de dados normais do mercado [Edelweiss 2001].

Os rótulos temporais são divididos em: tipos de dados temporais, tipo de tempo e expressões temporais [Edelweiss 2001].

Tipos de Dados Temporais

Os tipos de dados temporais podem ser classificados em “Instante”, “Período” e “Intervalo” onde o Instante é apenas o momento em que ocorre um determinado evento, o Período é o intervalo decorrido entre dois Instantes de tempos conhecidos e Intervalo é o tempo decorrido entre dois Instantes, porém não é conhecido quando começa ou quando termina.

Tipos de Tempo

Os tipos de tempo são classificados como: “Tempo de Transação”, “Tempo de Validade” e “Tempo Definido pelo Usuário” onde o Tempo de Transação é responsável por representar o momento em que é feita a alteração na base de dados, o Tempo de Validade visa representar o tempo em que os dados terão validade e por último o Tempo do Usuário onde o usuário manipula esse tempo através da aplicação.

Expressões Temporais

Segundo [Dean e Mcdermott 1987], as expressões temporais são construções sintáticas utilizadas para *queries* de consulta para determinar um valor temporal que podem ser determinados por um instante, período ou intervalo de tempo.

Quando se busca extrair conhecimentos em um BDT o uso de elementos de dados temporais se torna útil [Dean e Mcdermott 1987].

2.2 Data Warehousing

A integração de diferentes fontes de dados apresenta alguns desafios. Diferentes “departamentos” utilizarão diferentes estilos e formas de inserção e manutenção dos registros, diferentes convenções, períodos de tempo, graus de agregação, chaves primárias e principalmente, terão diferentes tipos de erro. Os dados devem ser agrupados, integrados e limpos. A ideia de se manter todas essas características em um banco de dados é conhecida como *data warehousing*. Os *data warehouses* promovem um único e consistente ponto de acesso aos dados, transcendendo a visão de “departamento” [Witten et al. 2011].

Os *data warehouses* são um passo importante para o pré-processamento na etapa de mineração de dados pois concentram e transformam dados de uma base OLTP para uma base OLAP que é uma ferramenta muito prática para fazer a análise de dados multidimensionais com suas granularidades diferentes, o que facilita a generalização dos dados e a mineração de dados [Han et al. 2011].

Um *data warehouse* é definido como uma “coleção de dados orientado à assunto, integrados, não voláteis, variáveis em relação ao tempo” [Inmon 2005].

Os dados são organizados por assuntos, permitindo assim a flexibilidade requerida para uma análise gerencial dos dados, ao passo que possibilita a estruturação de acordo com as áreas de atuação e objetivos estratégicos da organização [Tan et al. 2006].

A integração dos dados pode ser proveniente de bases localizadas em diversos sistemas operacionais possibilitando uma visão unificada e consistente de todo o cenário. Para isso, o dado é extraído de sua base original e traduzido para um formato adequado para as análises estratégicas e posteriormente tratado para se tentar eliminar as inconsistências, incompatibilidades e ausência de conteúdo essencial [Inmon 2005].

Existe uma importante característica em relação as aplicações comuns: não ser volátil. Isso indica que o *data warehouse* mantém um histórico dos dados que normalmente seriam alterados ou excluídos. Sendo assim, com a retenção do dado, consultas históricas sobre informações armazenadas em longos períodos de tempo podem ser feitas, permitindo uma análise de tendências a partir do estudo de elementos complexos de dados [Inmon 2005].

A variação temporal se deve ao fato dos dados serem armazenados conforme sua inserção ao longo de uma linha de tempo, de forma que consultas podem ser feitas para recuperar o estado de uma informação em um determinado instante de tempo. Neste sentido, o dado é constantemente inserido para se manter um registro histórico das operações ao longo de todo processo [Inmon 2005].

Um *data warehouse* pode possuir três tipos de arquitetura: uma, duas ou três camadas. O tipo mais utilizado é o de três camadas pois possui uma maior flexibilidade permitindo que as informações fiquem armazenadas em camadas distintas, como visto na Figura 2 [Han et al. 2011].

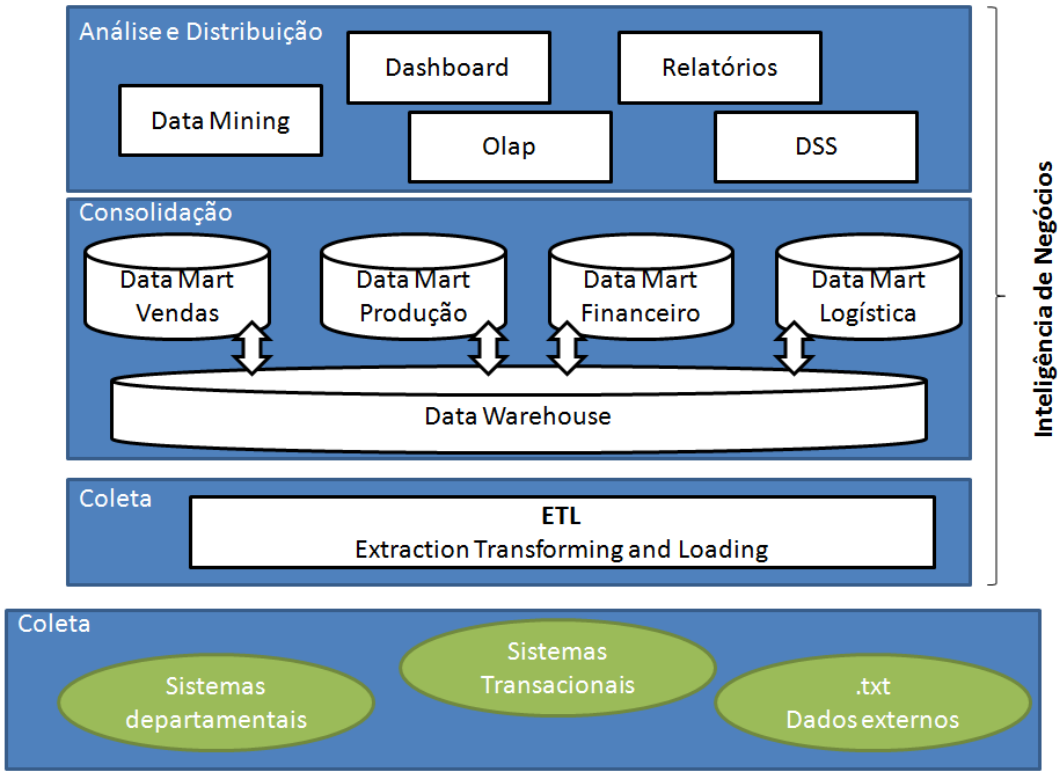

Figura 2: Estrutura de *Data Warehouse* dividida em três camadas [Devmedia 2013].

Na primeira camada podem estar as interfaces que mostram informações diretamente ao usuário final, onde geralmente estão disponibilizadas em formato de gráficos ou outros elementos que facilitem a compreensão das informações.

A segunda camada é onde se concentra os servidores de dados OLAP, que tem a função de promover um acesso eficiente e veloz aos dados compartilhados.

Na terceira e última camada ficam as fontes de dados. Normalmente esta fonte vem de um banco de dados relacional (OLTP).

2.3 Mineração de Dados

As técnicas de mineração de dados podem ser aplicadas em diversas áreas do conhecimento, dentre elas na Educacional que é o objetivo deste estudo. Levando-se em consideração os conceitos elaborados por Fayyad, neste estudo considera-se MD como uma etapa do KDD, e sua principal característica é a aplicação dos algoritmos aos dados pré-processados com o objetivo de gerar indicadores aos gestores ou analistas para apoio à tomada de decisão nos diferentes níveis, sejam eles estratégicos, táticos ou operacionais.

Segundo [Cios et al. 2007] as etapas da DM são:

1. Escolha da tarefa de DM: uma combinação de tarefas deve ser escolhida dentre vários tipos de tarefas possíveis sendo elas a regressão, agrupamento, classificação e associação. Como cada tarefa possui sua própria característica, é necessário um pré-conhecimento de como cada uma funciona para sua melhor aplicação. Neste trabalho será utilizada a tarefa de classificação.

2. Escolha do algoritmo de DM: de acordo com a tarefa escolhida, um ou mais algoritmos serão selecionados e aplicados nos dados, utilizando-se os modelos e parâmetros mais apropriados para o respectivo problema.

3. Aplicação da DM: busca por padrões em meio aos dados visando um interesse particular ou identificação de comportamentos antes não reconhecidos através das análises tradicionais.

De acordo com [Fayyad et al. 1996], as duas principais metas que podem ser alcançadas através da DM são a Previsão e a Descrição. A Previsão faz uso de variáveis existentes no banco de dados para prever valores desconhecidos ou futuros. A Descrição é voltada para a busca de padrões descrevendo os dados e a subsequente apresentação para a interpretação do usuário. A relativa ênfase entre previsão e descrição varia de acordo com o sistema de mineração de dados utilizado. Estes objetivos são conseguidos através da aplicação dos algoritmos.

2.4 Classificação

O processo de classificação é a tarefa de atribuir uma classe ou categoria predefinida aos objetos de uma determinada entrada. Essa classificação leva em consideração os atributos e características deste objeto gerando um modelo que permite obter a classe de saída [Tan et al. 2006]. A Figura 3 mostra este procedimento.

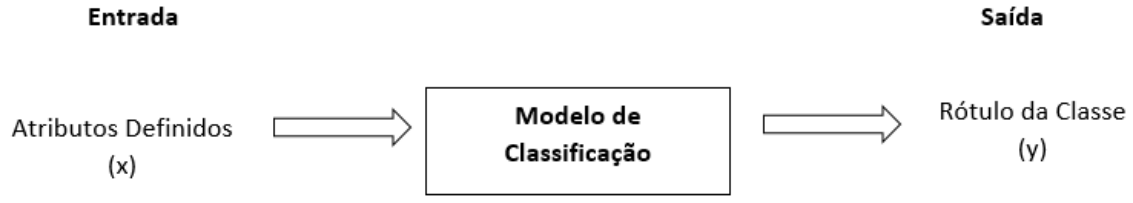

Figura 3: Tarefa de classificação mapeando um conjunto de entradas do atributo (x) em Rótulo da Classe (y) [Tan et al. 2006]

Na grande maioria as técnicas de classificação são separadas em dois tipos de modelos: Descritivos e Preditivos. Um modelo não precisa necessariamente ser de

somente um tipo, podendo assim ser de ambos [Tan et al. 2006].

Modelos Descritivos são modelos que podem servir como uma ferramenta de explicação do que distingue diferentes objetos de suas respectivas classes [Witten et al. 2011]. Um exemplo seria explicar qual a definição de um animal mamífero ou réptil. Supondo que no conjunto de objetos temos os atributos "tem nascimento" e "tem sangue quente" e as classes de saída fossem "espécie". Neste caso, a explicação para a saída da espécie seria que todos os mamíferos possuem nascimento e tem sangue quente.

Modelos Preditivos são modelos que podem ser usados para classificar uma classe de saída desconhecida [Han et al. 2011]. Considerando que temos os mesmos atributos "tem nascimento" e "tem sangue", mas não possuímos a classe a quem pertence. O modelo irá indicar uma classe de saída "mamífero" ou "réptil".

Em termos gerais nas tarefas de classificação separa-se o conjunto de dados em dois subconjuntos: treinamento e teste [Witten et al. 2011].

No subconjunto treinamento cada objeto de registro possui "n" atributos sendo que um deles é a classe de saída. Este conjunto é usado para construir um modelo de classificação que posteriormente será aplicado ao conjunto de teste [Tan et al. 2006].

O conjunto de teste é utilizado para validar a consistência do modelo em predizer qual é a classe de saída que deveria ser classificada uma determinada instância, indicando a acurácia do modelo gerado [Tan et al. 2006].

Segundo Nandeshwar [Nandeshwar et al. 2011] em vários estudos foi notado uma melhora significativa na performance e assertividade dos modelos voltados a previsão da retenção de estudantes quando utilizado técnicas de discretização dos dados, seleção de atributos, balanceamento de dados e validação cruzada de algoritmos (*Cross-Validation*). Valores discrepantes (*outliers*) e sobre ajustados (*overfitting*), também devem ser tratados.

O objetivo da detecção de anomalias é encontrar objetos que são muito diferentes dos outros objetos. Esses objetos discrepantes são chamados de *outliers* pois, pegando um gráfico de dispersão de dados, eles ficam muito "distantes" de outros elementos de dados [Tan et al. 2006]. Esses valores discrepantes podem surgir por serem um ruído, fazendo com que alguns algoritmos de classificação tenham problemas em gerar um modelo correto [Witten et al. 2011].

Existe um problema característico na mineração de dados que é chamado de *overfitting*. O *overfitting* ocorre quando o modelo está muito ajustado em relação ao conjunto de treinamento [Han et al. 2011]. Isso pode fazer com que o modelo gerado

tenha dificuldades de generalizar o problema, fazendo com que dados futuros não sejam classificados corretamente [Hämäläinen e Vinni 2011]. Este fenômeno pode ser evitado ou minimizado separando um conjunto adicional para validação [Lykourentzou et al. 2009].

Existem várias técnicas de classificação, porém para esta pesquisa serão considerados os seguintes classificadores: Árvore de Decisão, regressão logística, redes neurais, máquina de suporte de vetores, vizinhos mais próximos e redes bayesianas, baseando-se nos modelos mais recorrentes no contexto de retenção de estudantes.

2.4.1 Árvore de Decisão

Árvore de Decisão (*Decision Tree*) trabalha como um fluxograma em formato de árvore onde cada nó (que não seja folha) aponta para um teste que deve ser feito sobre um determinado atributo. Cada ligação entre os nós, representa possíveis valores de teste do nó pai e as folhas da classe de saída a qual a instância testada pertence [Goebel e Gruenwald 1999]. Na Figura 4, vemos um exemplo.

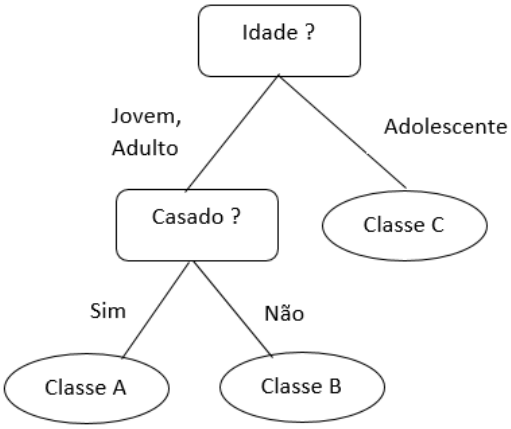

Figura 4: Exemplo de uma Árvore de Decisão [Han et al. 2011]

Depois de criado o modelo da árvore, para se classificar uma nova instancia é necessário seguir o caminho criado a partir do nó raiz até chegar ao nível mais inferior (nó folha) que indicará qual o seu rótulo de saída (no caso da Figura 4 Classe “A” ou B). A grande vantagem está na trajetória percorrida até o nó folha, que representa uma regra e/ou um conjunto de regras, facilitando a interpretação do modelo para o usuário final [Romero e Ventura 2010].

Os principais algoritmos de Árvore de Decisão baseiam-se no C4.5 [Quinlan 1993] criado por J. Ross Quinlan e baseado no algoritmo ID3 (*Iterative Dichotomiser*)

do início da década de 80. Algoritmos derivados incluem versões como o C5.0 [RuleQuest [S.d.]] e o J48, uma versão desenvolvida em Java.

2.4.2 Redes Neurais

O estudo das Redes Neurais Artificiais (*Artificial Neural Networks*) é inspirado na tentativa de simular o sistema biológico neural. O cérebro humano consiste primariamente de células nervosas chamadas neurônios que estão ligados a outros neurônios através dos chamados axônios. Estes axônios são usados para transmitir impulsos nervosos de um neurônio a outro quando os mesmos são estimulados [Tan et al. 2006]. A Rede Neural Artificial é uma analogia à esta estrutura do cérebro, e consiste de um conjunto de nós interligados entre si em várias camadas.

A ideia base é que se tenha as informações de entrada e saída conectadas, onde cada ligação possui um peso associado [Jain et al. 1996]; pesos adequados devem representar adequadamente o mapeamento entre as entradas e as saídas apresentadas. Com a evolução deste paradigma criou-se uma estrutura mais complexa chamada Redes Artificiais de Múltiplas Camadas (*Multilayer Artificial Neural Network*) onde se adicionou uma camada intermediária (escondida) fazendo a ligação entre as camadas de entrada e saída. Para se chegar nos pesos ideais, durante o processo de aprendizado (treinamento), a rede pode ajustar seus pesos para tentar classificar de uma maneira mais assertiva um objeto. As principais desvantagens deste modelo é quase sempre se necessita de um longo período para seu treinamento, alguns ajustes para os parâmetros de entrada são difíceis, e principalmente a dificuldade de se obter uma interpretação da relação existente entre os dados de entrada e suas respectivas saídas [Chen e Du 2009]. Na Figura 5 vemos um exemplo de uma rede *Multilayer* com quatro neurônios de entrada, três na camada “oculta” e dois na camada de saída.

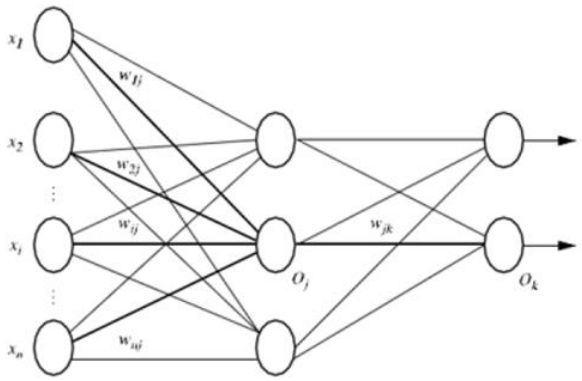

Figura 5: Rede Neural [Han et al. 2011]

2.4.3 Máquinas de Vetores de Suporte

O método “Máquinas de Vetores de Suporte” (*Support Vector Machines* - SVM) é uma técnica supervisionada utilizada preferencialmente para classificações onde é possível a separação linear das classes. O hiperplano que separa as classes é encontrado a partir de determinados “vetores de suporte”, obtidos do conjunto de treinamento de forma a definir a maior margem possível de separação entre as classes. Estes vetores de suporte são geralmente obtidos pelo uso de uma técnica de otimização. O SVM oferece uma maneira de trabalhar com objetos que não são linearmente separáveis transformando as coordenadas de um espaço “x” para um novo espaço “p(x)” de modo que uma margem linearmente separável possa ser aplicada na transformação do espaço [Tan et al. 2006].

O SVM é bastante aplicado onde há problemas com grandes dimensões de dados e se esses dados forem esparsos, porém não é fácil a interpretação de como se chegou ao resultado. Encontrar os melhores parâmetros para se propor um modelo pode ser complexo [Han et al. 2011].

Duas grandes vantagens dos SVM é que pode ser gerado um modelo com elevada generalização por meio da maximização de margem e o aprendizado eficiente de problemas não-lineares através da função de *kernel* [Yu e Kim 2012].

2.4.4 K Vizinhos Mais Próximos

K Vizinhos Mais Próximos (*Nearest-Neighbor* - KNN) é uma estratégia de classificação chamada preguiçosa (*lazy learning*). Previamente não é construído um modelo classificador, mas quando se deseja classificar um novo objeto são usadas as instâncias já rotuladas para classificar novas entradas. Para isso é necessário se obter uma função que calcule a distância entre as instâncias. Sendo assim, ao chegar um novo objeto não classificado, calculam-se as distâncias entre este novo objeto e as instâncias da base, de forma a obter as K mais próximas (ou K vizinhos mais próximos). A classe atribuída ao novo objeto é, em geral, a mais frequente entre as classes do K vizinhos correspondentes [Yang 1994].

Um fator muito importante para esta técnica é a normalização dos valores, pois sem este procedimento poderá haver uma discrepância muito grande entre os valores considerados dos atributos, comprometendo o processo de classificação [Tan et al. 2006].

2.4.5 Näive Bayes

Näive Bayes é um classificador criado baseado em uma técnica probabilística

que emprega o teorema de *Bayes*. Este teorema diz que é possível encontrar a probabilidade de um evento ocorrer através da probabilidade condicional de um outro evento ter ocorrido. Ele é um modelo bem simples e muito utilizado como ponto de partida na comparação de resultados, partindo do princípio que todos os atributos são independentes uns dos outros [Zhang 2004]. Apesar do próprio nome já sugerir *Näive* (Ingênuo), pois segue a premissa de independência entre os preditores, esta abordagem é aplicada com sucesso em vários problemas do mundo real [Zhang 2004].

2.4.6 Discussão

Como visto nesta seção, existem vários métodos de classificação que seguem paradigmas bastante diversos. Fazer a escolha de um algoritmo não é uma tarefa trivial. Sendo assim, W. Hämäläinen criou uma tabela onde mostra os principais métodos de classificação de acordo com oito critérios. Segundo os autores, esses critérios muitas vezes são relevantes no contexto de classificação de dados educacionais [Hämäläinen e Vinni 2011]. Na Tabela 1 vemos essa comparação.

Tabela 1: Comparação de diferentes paradigmas de classificação. O sinal "+" indica que o método oferece suporte à propriedade, "-" que não oferece suporte [Hämäläinen e Vinni 2011].

	Árvore de decisão	Naive Bayes	Bayesianos em Geral	Redes Neurais	K-Vizinhos	SVM
Fronteiras não lineares	+	+	+	+	+	+
Acuracia em pequenos conjuntos de dados	-	+	+/-	-	-	+
Trabalha com dados incompletos	-	+	+	+	+	-
Suporta variáveis mistas	+	+	+	-	+	-
Interpretação natural	+	+	+	-	+	-
Raciocínio eficiente	+	+	+	+	-	+
Aprendizado eficiente	+/-	+	-	-	+/-	+
Atualização eficiente	-	+	+	+	+	-

2.4.7 Métricas para Avaliação da Performance dos Classificadores

Perguntas do tipo "como posso avaliar a performance do meu modelo? " Ou "como eu posso ter uma estimativa confiável? " Ou mesmo "como comparar o desempenho entre os modelos? ", são pontos que devem ser considerados em uma tarefa de MD.

Para um problema de classificação é natural mensurar a performance do classificador baseado na taxa de erro [Witten et al. 2011]. O classificador prediz a classe de cada instancia: se a mesma está correta a mesma é contada como sucesso e em caso contrário como um erro. A taxa de erro é apenas a proporção de erros

cometidos ao se classificar o conjunto de treinamento, e este valor é usado para mensurar a performance do classificador [Witten et al. 2011]. Similarmente a proporção de acertos indica a acurácia (*accuracy*) do modelo, como indicado adiante.

Estimar o erro ajuda os algoritmos de aprendizado a criar um modelo de seleção, ou seja, encontrar um modelo onde a complexidade não seja suscetível a *overffiting* [Tan et al. 2006].

No caso do método de *holdout*, a base de dados original, já com as classes de saída rotuladas, é particionado em dois conjuntos distintos chamados de treinamento e teste. Um modelo de classificação é então gerado a partir do conjunto de treinamento e sua performance é avaliada no conjunto de teste. A proporção dos dados reservados para o treinamento e teste fica a cargo unicamente do analista, embora a proporção 2/3 – 1/3 seja a mais usada [Tan et al. 2006]. Na Figura 6 vemos uma amostra para teste e treinamento.

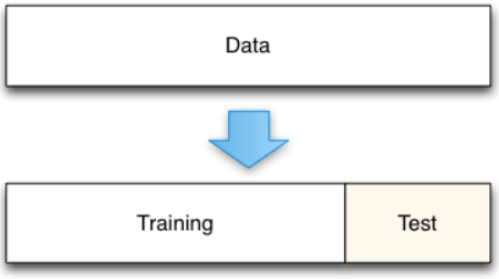

Figura 6: *Holdout* para o conjunto de teste [Roe 2012]

A validação cruzada (*cross-validation*) é utilizada para se estimar a distribuição do erro real do modelo gerado pelo algoritmo de classificação. Quando o modelo é construído a partir dos dados de treinamento um valor único do erro é uma estimativa muito otimista das taxas de erro, considerando a capacidade do modelo conseguir classificar corretamente os dados futuros. A ideia para se construir um modelo adequado é aplicá-lo a novos dados que nunca forem utilizados anteriormente, esperando que o modelo seja generalista o suficiente para conseguir classificar corretamente estas novas instâncias [Witten et al. 2011].

No contexto da retenção de estudantes, a grande maioria das pesquisas utiliza o método *k-fold cross-validation*. Essa é uma simples e intuitiva maneira de se estimar a distribuição estatística do erro. Na abordagem de *k-fold cross-validation* utiliza-se o particionamento da base de dados em “k” partições de mesmo tamanho. Durante cada execução, uma das partições é escolhida para o teste, enquanto o resto deles é usado para o treinamento. Isto é repetido “k” vezes até que cada partição seja utilizada como teste. Assim obtém-se uma distribuição a partir dos “k” valores obtidos, e o erro estimado é a média do erro calculado em cada uma das partições de teste [Tan et al. 2006]. Na Figura 7, vemos um exemplo desta situação para k= 5.

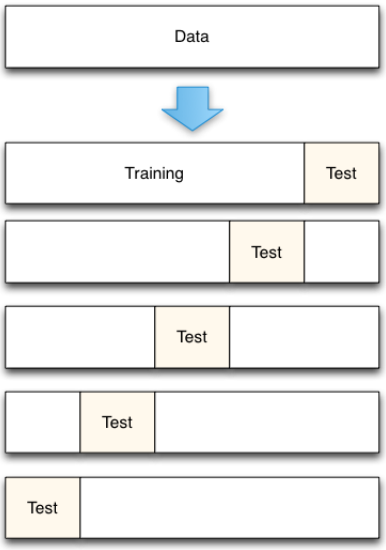

Figura 7: *K-fold cross-validation*, sendo “k” com valor 5 [Roe 2012]

A acurácia (*accuracy*) mede o grau de acerto obtido pelo classificador. Usar os dados de treinamento para calcular a acurácia poderia gerar uma estimativa muito otimista e não realista de um determinado modelo gerado pelo algoritmo de aprendizado. Portanto, a acurácia sempre é medida pelo conjunto de teste que consiste de instancias com classes rotuladas (de saída) que não foram utilizadas pelo modelo de treinamento. A acurácia de um classificador é baseada na porcentagem do conjunto de teste que foram classificados corretamente pelo classificador [Han et al. 2011]. A formula da acurácia é:

$$\frac{VP + VN}{VP + VN + FP + FN}$$

Onde positivo (P) é quando o classificador classifica a amostra que pertence à classe, e negativo (N) é quando classifica como não pertencente à classe. Sendo assim o verdadeiro positivo (VP) se refere a instâncias positivas que foram rotuladas corretamente pelo classificador, enquanto o verdadeiro negativo (VN) são os negativos que foram rotulados corretamente pelo classificador. Os falsos positivos (FP) são os positivos que foram rotulados de maneira incorreta pelo classificador e os falsos negativos (FN) referem-se aos negativos rotulados pelo classificador como positivos [Tan et al. 2006].

A matriz de confusão (*confusion matrix*) é uma tabela muito utilizada para se visualizar estes elementos [Witten et al. 2011]. Na matriz de confusão os resultados da classificação são indicados como uma matriz bidimensional onde cada valor da matriz corresponde ao número de instancias que foram classificadas de forma errada ou correta utilizando-se do conjunto de teste [Fayyad et al. 1996]. Supondo que em certo problema se tenham duas classes de saída, a matriz de confusão ficaria como indicado à Tabela 2 a seguir.

Tabela 2: Matriz de confusão para instancias positivas e negativas [Han et al. 2011]

		Classes Previstas	
		Classe 1	Classe 2
Classes Verdadeiras	Classe 1	Taxa de Verdadeiro Positivo **(VP)**	Taxa de Falso Negativo **(FN)**
	Classe 2	Taxa de Falso Positivo **(FP)**	Taxa de Verdadeiro Negativo **(VN)**

2.4.8 Criação de Atributos

A criação de novos atributos ajuda a abstrair informações importantes em um conjunto de dados de forma mais eficiente do que quando utilizado os atributos em sua forma original [Anzanello et al. 2011].

Com a criação de novos atributos baseados em outros já existentes, pode-se em alguns casos criar um novo conceito: por exemplo, o atributo peso e altura de uma determinada pessoa pode ser resumido para somente um único atributo "Índice de Massa Corpórea" – IMC, pela conhecida fórmula peso/(altura)^2. Outro exemplo é o vislumbre de uma série de pagamentos realizados por um determinado cliente em algumas lojas: apesar de existirem diferentes tipos de pagamentos como cartão de crédito, débito, dinheiro, etc., avaliando se o período de pagamentos pode-se criar um atributo de "inadimplente".

Este trabalho visa investigar e propor novos atributos temporais extraídos da base de dados obtida do Sistema Acadêmico de uma Instituição de Ensino, que possam melhorar a performance dos algoritmos de classificação e seu potencial para auxiliar a previsão da fidelização de alunos.

2.5 Preparação dos dados

2.5.1 Discretização de dados

Em termos gerais os dados coletados e armazenados estão em três formatos: nominal, discreto e/ou contínuo [Han et al. 2011].

Dados discretos e contínuos são tipos de dados ordinais, ou seja, possuem uma ordem entre os valores, enquanto os valores nominais não possuem qualquer forma de ordenação entre eles. Valores discretos são intervalos de um espectro contínuo de valores. Enquanto o número de valores contínuos de um atributo pode ser infinito, os números de valores discretos são muitas vezes menores ou finitos [Witten et al. 2011]. Ambos os tipos de valores dependem do atributo, e podem fazer uma grande diferença no aprendizado de algoritmos de classificação como árvores ou regras por exemplo [Liu et al. 2002].

O principal objetivo da discretização de dados é a redução do número de valores de atributos contínuos através de agrupamentos. Além disso, existem algoritmos que só trabalham sobre dados discretizados, como por exemplo o *Naïve Bayes* [Cios et al. 2007].

2.5.2 Balanceamento de Dados

O desbalanceamento pode ser descrito como o número de casos que uma determinada classe supera significativamente o de outras [Tan et al. 2006]. Chamamos as classes que tem a maior quantidade de classes de “maioritária” e a que possui o menor número de classes de “minoritária”. Nestes casos, os algoritmos de classificação tendem a dar prioridade as classes maioritárias [Nandeshwar et al. 2011].

O problema do desbalanceamento é recorrente em aprendizado de máquinas e aplicações de mineração de dados como visto em muitas tarefas de predição no mundo real. Contudo as técnicas e conceitos para o balanceamento de dados dependem em muito do modelo construído e são motivo de muitas pesquisas realizadas na atualidade. Uma ampla variedade de técnicas de balanceamento de dados tem sido aplicada em conjuntos de dados nas mais diversificadas áreas, como diagnósticos médicos [Su et al. 2006] [Li et al. 2010], classificadores para bases de dados em marketing [Duman et al. 2012], e classificadores para falhas em solda [Liao 2008] entre outros.

Muitas pesquisas têm focado no problema das classes desbalanceadas, tentando promover uma melhor acurácia na predição baseando-se nas classes minoritárias [Drummond e Holte 2003]. Alguns desses métodos são usados no pré-

processamento, onde se busca balancear os dados antes da construção do modelo, enquanto outros desenvolvem algoritmos preditivos que inserem um peso maior dependendo da representação das classes. O pré-processamento mostra-se como a abordagem mais direta e representa a maior promessa em superar os problemas de desbalanceamento de classes [Jo e Japkowicz 2004]. Esta abordagem usa vários métodos para eleger aleatoriamente uma amostragem maior da classe minoritária ou uma amostragem menor para a classe maioritária ou mesmo uma combinação entre as duas [Thammasiri et al. 2014]. A amostragem maior (*over-sampling*) serve para popular uma nova amostra com as classes balanceadas inserindo dados para o conjunto de teste utilizando as classes minoritárias. O *over-sampling* visa equilibrar as populações das classes através da criação de novas amostras da classe minoritária por seleção aleatória, e de sua adição ao conjunto de treinamento. Por outro lado, a amostragem insuficiente (*under-sampling*) visa equilibrar as populações de classes através da remoção de dados da classe maioritária, até que as representações das classes serem igualadas [Liao 2008]. Mesmo que não haja nenhuma evidência comprovada de superioridade da técnica de *under-sampling*, pesquisas mostram que esta técnica leva uma ligeira vantagem em relação a *over-sampling* [Drummond e Holte 2003].

Portanto, as técnicas de balanceamento de dados são conhecidas por promover melhores resultados em relação ao conjunto de dados original [Thammasiri et al. 2014].

Foram encontradas diversas pesquisas voltadas para otimizar as técnicas de balanceamento de conjunto de dados, melhorando o desenvolvimento de modelos preditivos. Estas técnicas são em sua grande maioria, derivada da abordagem de *over/under-sampling* como a técnica de *SMOTE* [Chawla et al. 2011], *Cluster-Based sampling* [Jo e Japkowicz 2004], *Adaptive Synthetic Sampling algorithms* [Chawla et al. 2011], entre outros.

2.5.3 Seleção de Atributos

Muitos fatores afetam o sucesso dos algoritmos de mineração de dados. A qualidade dos dados é um desses fatores. Se a informação é irrelevante ou redundante, se os dados possuem muitos ruídos ou são pouco confiáveis, o processo de descoberta de conhecimento durante o treinamento torna-se difícil e em alguns casos inviável [Hall e Holmes 2003]. A seleção de dados é o processo de identificar e remover o máximo possível de informações redundantes ou irrelevantes. Os algoritmos de aprendizado diferem no que se refere à ênfase dada aos atributos selecionados [Hall e Holmes 2003].

A maioria dos algoritmos de aprendizado são projetados para aprender com os atributos mais apropriados e estes são usados para fazer suas decisões. Por isso, é

necessário a seleção de um pequeno número de características altamente preditivas, afim de evitar o *overfitting* para os dados de treinamento. Independentemente se foi selecionado um atributo ou ignorado, a seleção de atributos antes da etapa de aprendizado pode ser benéfico [Witten et al. 2011].

A redução da dimensionalidade dos dados reduz o tamanho de possibilidades e permite os algoritmos operar de forma mais rápida e eficaz, minimizando a perda de conteúdo, tempo e informações [Han et al. 2011]. Remover atributos relevantes ou manter atributos irrelevantes, pode ser a principal causa no mau funcionamento de um algoritmo empregado, podendo resultar em descobertas de padrões com pouca qualidade. Em contrapartida adicionar um volume de informações irrelevantes ou redundantes pode diminuir em muito a qualidade da mineração [Han et al. 2011].

A redução do conjunto de dados removendo os atributos irrelevantes ou redundantes chama-se subconjunto da seleção de atributos (*Attribute Subset Selection*) [Han et al. 2011]. A melhor maneira para selecionar atributos relevantes é separá-los manualmente, baseado em um profundo entendimento do problema a ser estudado definindo quais atributos realmente são importantes [Witten et al. 2011]. Porém dependendo do tamanho da base dados e de sua dimensionalidade, fica praticamente inviável este tipo de análise, e para isso são aplicados métodos automáticos [Witten et al. 2011]. Estes métodos são basicamente separados em três categorias: *embedded*, *filter* e *wrapper* [Tan et al. 2006].

Os métodos *embedded* ocorrem naturalmente como parte dos algoritmos de mineração de dados, mais especificamente, durante as operações destes algoritmos, onde o próprio algoritmo decide qual atributo usar e qual ignorar. Algoritmos de construção de Árvores de Decisão utilizam esta abordagem [Tan et al. 2006].

A abordagem *filter* é chamada desta forma porque os conjuntos de atributos são filtrados antes do aprendizado começar gerando subconjuntos mais promissores [Witten et al. 2011]. Se esses recursos são selecionados antes do algoritmo de mineração de dados ser executado, esta abordagem fica independente da tarefa de mineração de dados [Tan et al. 2006]. Na Figura 8, vemos como o *filter* trabalha.

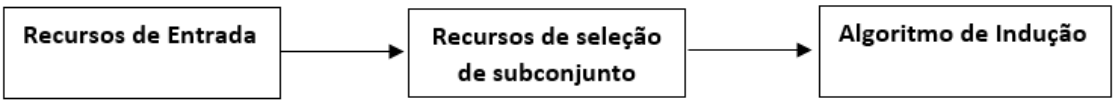

Figura 8: Abordagem *filter* onde os recursos são filtrados independentemente do algoritmo de indução [Kohavi e John 1997].

Porém há uma desvantagem neste tipo de abordagem. Neste caso são

ignorados totalmente os efeitos do subconjunto de atributos selecionado no desempenho do algoritmo de indução [Kohavi e John 1997].

Já na abordagem *wrapper* o algoritmo de aprendizado é "envelopado" em todo o processo recursivo de seleção [Witten et al. 2011]. O método de *wrapper* utiliza o algoritmo de mineração de dados de destino como uma caixa preta, para encontrar o melhor subconjunto de atributos [Tan et al. 2006].

O grande desafio na seleção de subconjuntos é fazer com que o algoritmo de aprendizado se concentre no subconjunto mais relevante, ignorando os outros atributos. Na abordagem *wrapper* o algoritmo de seleção de subconjuntos realiza uma pesquisa para encontrar um bom subconjunto de atributos, usando o algoritmo de indução em si como parte da função avaliadora [Kohavi e John 1997]. O mecanismo por trás da abordagem *wrapper* é relativamente simples (ver Figura 9): (1) o algoritmo de indução é considerado como uma caixa preta; (2) este algoritmo é executado sobre um conjunto de dados, geralmente dividido em conjuntos de treinamento e validação com diferentes conjuntos de recursos retirados dos dados; (3) o subconjunto com a avaliação mais alta é escolhido como conjunto final, onde o algoritmo de indução fará a execução; (4) por fim, o resultado do classificador é então avaliado sobre um conjunto de testes independente do utilizado durante o treinamento [Kohavi e John 1997].

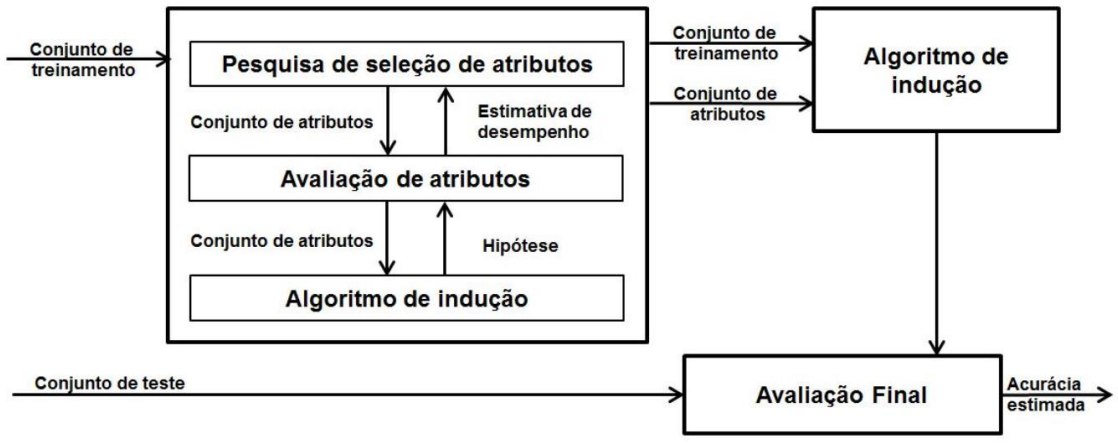

Figura 9: Abordagem *wrapper* para seleção de subconjuntos onde o algoritmo de indução é utilizado como uma caixa preta pelo algo*ritmo de seleção de subconjunto* [Kohavi e John 1997].

2.6 Trabalhos Relacionados

Os trabalhos relacionados nessa seção tratam apenas das abordagens computacionais da análise do problema da fidelização em EDM. Vários autores já

investigaram o problema da evasão/fidelização usando EDM, sendo que a maioria deles o considera como um problema de classificação. Juntamente com a EDM, o assunto da mineração de dados temporal e criação de atributos também foram explorados em alguns trabalhos.

O problema da descoberta na associação temporal dos dados foi introduzido por Agrawal [Agrawal et al. 1993]. Seguiram-se sucessivos refinamentos, generalizações e melhorias [Agrawal et al. 1993, 1996; Parker 1999]. Entre estes, podem-se encontrar algoritmos melhorados para a descoberta de conjuntos de itens frequentes, classificadores, regras de associação generalizadas e quantitativas, e novas medidas para outros tipos de dados, para as mais diversificadas áreas.

Os trabalhos anteriores sobre mineração de dados que incluem aspectos temporais geralmente estão relacionados à análise da sequência de eventos e criação de novos atributos [Agrawal et al. 1996; Bettini et al. 1996]. Normalmente o objetivo é descobrir regularidades na ocorrência de certos eventos e relação temporais entre os diferentes eventos. Pode-se citar em particular o trabalho de Mannila [Mannila et al. 1995], onde o problema de reconhecer episódios frequentes em uma sequência de eventos é discutido. Neste trabalho mostra-se que um episódio é definido como uma coleção de eventos que ocorrem durante intervalos de tempo em um tamanho específico. Por outra lado Srikant [Srikant e Agrawal 1996] analisam o problema de descobrir padrões sequenciais em bases de dados transacionais, gerando novos atributos. A solução adotada consiste em criar uma série sequencial para cada cliente e procurar padrões frequentes em cada sequência.

Pode-se citar ainda Bettini [Bettini et al. 1996], que trabalhou com padrões mais complexos do que nos casos citados acima. Neste caso, são tratadas as distâncias temporais com múltiplas granularidades.

Todos esses trabalhos têm em comum o objetivo de descobrir padrões em seus períodos ou intervalos de tempo através de classificadores ou regras de associação. Eles apresentam algoritmos para encontrar de forma eficiente o que eles chamaram de "classificação cíclica" devido a existência de determinados intervalos de tempo e, portanto, exibem variações cíclicas regulares ao longo do tempo.

Na pesquisa de Ramaswamy [Ramaswamy et al. 1998] os autores estudam como os classificadores e as regras de associação que variam ao longo do tempo, generalizando o trabalho de [Özden et al. 1998]. Eles apresentam a noção de álgebra de calendário para descrever fenômenos temporais de interesse para os usuários e os algoritmos presentes para descobrir "regras e classificadores baseado em calendário", ou seja, classificadores e regras de associação que seguem os padrões temporais estabelecidos nas expressões de calendário fornecidas pelo usuário.

Já voltado a área de EDM, Pittmann [Pittman 2008] realizou em sua tese de doutorado uma pesquisa em que comparou os algoritmos de Árvores de Decisão, redes neurais, regressão logística e classificadores baseados em *Bayes* em estudos que tinham como foco a retenção de alunos. Neste trabalho ele apresenta um ranking com os atributos mais relevantes no contexto de nível superior dentro da Universidade New Southeastern University nos Estados Unidos, sendo alguns deles: a média da nota no ensino médio, a nota no teste de admissão no vestibular, o gênero e a etnia. Foram criados também novos atributos baseados nos já existentes. Levando-se em conta a área da curva de ROC, o melhor algoritmo preditor foi a regressão logística, segundo Pittman. Seu modelo obteve um valor de acurácia de 78% a 81% para a não retenção de um estudante, quando aplicado a uma base com 21.136 instâncias de dados educacionais.

Delen [Delen 2010] cita que para se melhorar a retenção dos alunos deve-se começar com uma minuciosa compreensão das razões por trás dos conflitos nos quais os estudantes estão envolvidos. Tal compreensão é a base para predizer com precisão alunos em risco e adequadamente tomar ações para retê-los. Para isso, ele criou um modelo utilizando validação cruzada em 10 vezes aplicando nos algoritmos SVM, Redes Neurais, Árvore de Decisão e Regressão Logística, em uma base com 23.000 estudantes obtida de uma universidade pública localizada no centro-oeste dos Estados Unidos. O melhor resultado para classificar a classe “Yes”, indicando que o estudante permaneceria na Instituição foi obtido com a Árvore de Decisão com acurácia de 87,23%, acompanhado de perto pela Rede Neural com acurácia de 87,16%.

Baseado em estudos recentes, Lin [Lin 2012] demonstra em seu trabalho que a maior parte dos alunos desistentes são calouros do primeiro ano da graduação. Para tentar mensurar isso, o autor e a equipe de retenção de alunos de um campus da Biola University localizada nos Estados Unidos, coletaram os dados de uma turma da disciplina de Inteligência Artificial para criarem um modelo preditivo onde seria possível identificar os alunos calouros do primeiro ano que são mais suscetíveis a se beneficiar do programa de retenção.

Na primeira etapa, foram gerados novos atributos e incorporados aos atributos existentes, convertendo-se todos os valores dos atributos em numéricos. Utilizando o WEKA [Hall et al. 2009], aplicaram-se os dados a quatorze algoritmos de Árvore de Decisão, nove de Regras, quatro de KNN, sete de Redes Neurais e SVM e cinco baseados em *Bayes* e Redes Bayesianas. O Algoritmo de Árvore de Decisão Alternativo (ADT) obteve a melhor precisão com um índice de 84% de acerto e segundo o estudo, sem sinal de *overfitting*. O modelo foi criado utilizando-se um conjunto de 1.000 novos alunos, sendo que desses cerca de 250 são candidatos a abandonar o curso (supondo uma taxa de 25%). O modelo produziu uma lista de

potenciais alunos evasores com cerca de 37 alunos, sendo que 31 deles realmente abandonaram o curso.

Uma tabela comparativa entre os algoritmos com resultados mais relevantes é mostrada na Tabela 3.

Tabela 3: Precisão encontrados com os principais modelos preditivos [Lin 2012]

	1x data set		2x data set			3x data set		
	Precision	Recall	Precision	Recall	Over-fitting	Precision	Recall	Over-fitting
ADT Tree	83.9%	12.3%	84.0%	12.4%	Unlikely	49.5%	17.6%	Unlikely
NB Tree	77.9%	07.9%	56.4%	08.9%	Unlikely	40.8%	27.8%	Unlikely
CART	73.8%	05.1%	40.4%	49.0%	Likely	44.9%	88.7%	Likely
J48 graft	70.3%	09.6%	44.4%	35.1%	Likely	43.3%	69.8%	Likely
J48	68.8%	09.9%	43.6%	35.2%	Likely	42.7%	69.8%	Likely

Em um trabalho que utiliza redes neurais, florestas aleatórias e Árvore de Decisão, Superby [Superby e Vandamme 2006] classificou em baixo, médio ou alto o risco de abandono dos novos alunos que ingressavam na Universidade da Bélgica no período acadêmico entre 2003 e 2004, a partir de uma amostra de 533 estudantes. Os autores descobriram que o histórico escolar e o fator sócio-familiar formam os melhores preditores como métrica do índice de risco. Os resultados indicaram que a Árvore de Decisão obteve uma taxa de acerto de 40,63%, as florestas aleatórias uma taxa de 51,78% e as redes neurais uma taxa de 51,88%.

Marquez [Marquez-Vera et al. 2013] realizou um trabalho propondo aplicar somente técnicas de classificação que fossem de caixa branca a uma base com dados de 670 estudantes do ensino fundamental em Zacatecas, no México. Primeiro foram utilizados todos os atributos disponíveis, em seguida selecionaram os melhores atributos e por fim foi efetuado o balanceamento de dados e utilizado um custo de classificação sensível às classes. Com base na média de 10 execuções, a Tabela 4 mostra a porcentagem de classificações corretas para todos os atributos levando em conta as classes de saída “Passou” e “Falhou”, e as porcentagens da Acurácia (Ac) e Média Geométrica (GM). Baseado nas mesmas atribuições da Tabela 4, a Tabela 5 mostra os resultados para os dados balanceados, sendo possível visualizar uma sensível melhora considerando a quantidade de acertos na classificação

Tabela 4: Resultados da classificação utilizando todos os atributos [Marquez-Vera et al. 2013]

Algoritmo	Passou	Falhou	Ac	GM
JRip	97,0	**81,7**	95,7	**89,0**
NNge	98,0	76,7	96,1	86,7
OneR	98,9	41,7	93,7	64,2
Prism	99,2	44,2	94,7	66,2
Ridor	95,6	68,3	93,1	80,8
ADTree	**99,2**	78,3	**97,3**	88,1
J48	97,7	55,5	93,9	73,6
RandomTree	98,0	63,3	94,9	78,8
REPTree	97,9	60,0	94,5	76,6
SimpleCart	98,0	65,0	95,1	79,8

Tabela 5: Resultados da classificação utilizando os dados Balanceados [Marquez-Vera et al. 2013]

Algoritmo	Passou	Falhou	Ac	GM
JRip	97,7	**65,0**	94,8	**78,8**
NNge	98,7	78,3	96,9	87,1
OneR	88,8	88,3	88,8	88,3
Prism	99,8	37,1	94,7	59,0
Ridor	97,9	70,0	95,4	81,4
ADTree	**98,2**	86,7	**97,2**	92,1
J48	96,7	75,0	94,8	84,8
RandomTree	96,1	68,3	93,6	79,6
REPTree	96,5	75,0	94,6	84,6
SimpleCart	96,4	76,7	94,6	85,5

Yu [Yu et al. 2007] realizou uma pesquisa um pouco diferente das tradicionais, pois incluiu variáveis não muito usadas nesse tipo de pesquisa, como por exemplo: dados demográficos, indicadores de desempenho acadêmicos de pré-universidade e se o estudante morava no campus ou fora dele. Nessa pesquisa, utilizando Árvore de Decisão, descobriu-se que o atributo “morar ou não dentro do campus” é muito relevante no contexto da base utilizada. Para esta pesquisa foram extraídos 10.000 registros do *Data Warehouse* da Arizona State University, no período de 2002 a 2006, e a taxa de acerto obtida foi de 73,6%.

Para analisar a diminuição no número de graduados no ano de 2010 na Savannah State University no Estados Unidos, Lauría [Lauría et al. 2012] utilizou fontes de dados relacionadas aos históricos do aluno e dados relacionados ao curso escolhido. Para fazer o balanceamento dos dados os autores utilizaram a técnica de *oversampling* e aplicaram os classificadores em árvore para fazer a predição da evasão e comparar com outras técnicas. Nesta comparação incluíram a regressão logística, o

SVM e o algoritmo de Árvore de Decisão C4.5. Os resultados sobre uma base de 27.276 registros mostram que a regressão logística e o SVM possuem uma melhor acurácia em relação à Árvore de Decisão, considerando a tentativa de detectar um possível risco de não-retenção de um estudante.

Eckert [Eckert e Suénaga 2015] mostra em seu estudo que as principais variáveis que resultam na saída de um estudante do curso de Engenharia da Informática da Universidad Gastón Dachary na Argentina, usando dados do período entre 2000 à 2009. Os atributos mais importantes foram – em ordem de importância – as notas obtidas no ensino médio, a nota tirada em Matemática e o período entre a saída do ensino médio e o ingresso na faculdade. Neste estudo ainda se verificou que os estudantes de engenharia possuem um alto índice de desistência.

Marquez [Marquez-Vera et al. 2013] argumenta que a quantidade de dados armazenadas dos alunos dificulta a descoberta de informações úteis, mas com o uso da EDM pode-se chegar a conclusões promissoras. Em contrapartida, há um problema relacionado a EDM, pois a grande maioria das pesquisas realizadas focam somente no fracasso/abandono de estudantes aplicado especificamente ao ensino superior ou ensino a distância. Pouquíssimas pesquisas estão focadas no ensino médio ou fundamental, e as que são encontradas baseiam-se em métodos estatísticos e não na mineração de dados.

Thammasiri [Thammasiri et al. 2014] confirma a conjectura de Marquez quando diz que há muitos anos os métodos estatísticos tradicionais têm sido utilizados para prever a correlação entre os fatores de desgaste e o comportamento no âmbito acadêmico que causam a retenção. Para os autores as principais diferenças entre o uso das técnicas estatísticas e a mineração de dados são [Thammasiri et al. 2014]:

- A mineração de dados provê várias informações adicionais, pois o seu processo é mais amplo e algumas etapas já incluem os métodos estatísticos;

- Técnicas estatísticas são usadas frequentemente para encontrar dados similares, sendo que na Mineração de Dados é utilizada uma abordagem mais direta que trabalha com os dados classificados;

- A Mineração trabalha muito bem com grandes quantidades de dados (milhões e bilhões de registros), já a estatística normalmente não funciona muito bem com grandes bases de dados e com conjuntos de alta dimensionalidade.

- A Mineração de Dados consegue produzir previsões mais precisas e de maior utilidade.

Segundo Nandeshwar [Nandeshwar et al. 2011] há uma desconexão entre as

teorias propostas para explicar a evasão de alunos e os dados disponíveis para se apoiar essas teorias. Baseado na discussão com os administradores da Universidade de Kent nos Estados Unidos, os autores afirmam que, de maneira informal, não se pode chegar a uma conclusão sobre qual é o principal fator que influencia na permanência do estudante. Na maioria dos casos, a situação financeira é considerada o item mais importante para a evasão, mas isto não é um consenso. Porém, quando se analisam os registros, não se tem uma base quantitativa sólida que apoie esta crença. Na realidade pode-se notar que muitas universidades tentam melhorar a retenção trabalhando com vários programas de fidelização como atrair os alunos com os maiores índices de desempenho no vestibular e ensino médio, investir nos estudantes que possuem melhores notas no primeiro ano e, principalmente, investir na qualificação dos professores; no entanto não há um modelo que embase estas escolhas.

Quando se fala em EDM trabalha-se com grandes volumes de dados que podem conter muitas variáveis preditivas [Romero e Ventura 2010]. Com isso encontram-se alguns problemas, principalmente lidando com a ausência de dados e padrões complexos não lineares. Apesar de existirem várias técnicas de aprendizado de máquina que tratam estas questões, tem-se mostrado na literatura que a superioridade entre as diferentes técnicas varia entre os diferentes contextos institucionais. Dependendo dos dados e da formulação do problema qualquer técnica de mineração pode ser superior a outra. Esta falta de consenso pede uma abordagem experimental para se tentar identificar a técnica mais adequada para um problema de predição [Romero et al. 2010].

A maioria dos estudos sobre retenção de alunos se concentram em uma única variável ou um único conjunto de variáveis, ou mesmo no fator já bem estabelecido de ponto médio de baixo grau (GPA) que explica somente uma pequena porcentagem de variação na retenção [Pittman 2008]. Com base nessas pesquisas tem-se notado a necessidade de modelos sofisticados que possam levar em conta as múltiplas variáveis, que conduzem a uma contribuição para predição da variável de relevância no que diz respeito ao desgaste do aluno com a instituição, bem como a necessidade da retenção do mesmo [Pittman 2008].

Capítulo 3

Metodologia Empregada

Este capítulo apresenta a metodologia e os materiais utilizados para o desenvolvimento do modelo proposto.

3.1 Metodologia

Este trabalho consiste de uma pesquisa exploratória de natureza aplicada [Gerhardt et al. 2009], uma vez que objetiva estudar métodos para identificação de padrões que auxiliem na tomada de decisão de gestores educacionais com a finalidade de analisar a fidelização de estudantes em instituições privadas.

Esta pesquisa é quantitativa, pois baseado na abordagem proposta, a análise ocorrerá por meio de resultados passíveis de serem mensurados através dos experimentos realizados.

Baseado no conhecimento técnico e científico visto no Capítulo 2, foi possível desenvolver e avaliar uma solução computacional para o problema da identificação de padrões na EDM objetivando a retenção de estudantes, sendo assim o método utilizado classifica-se como dedutivo [Gerhardt et al. 2009].

Foram realizados levantamentos bibliográficos que fundamentam o desenvolvimento do método proposto da criação de novos atributos para classificação. Esse procedimento técnico proporcionou os fundamentos para a realização dos experimentos baseados nos métodos para prova de conceito e análise do método deste trabalho.

3.2 Método

O método proposto contém 5 (cinco) etapas como apresentado à Figura 10. A primeira etapa é a Análise dos Dados, onde se tenta compreender os dados coletados da base de dados da Rede Educacional Adventista. A segunda etapa é a Catalogação dos Dados onde é realizada toda a tarefa de pré-processamento. A terceira etapa é a Criação de Novos Atributos que ocorre logo após o pré-processamento. A quarta etapa é a Aplicação da Mineração de Dados e a etapa final é a Validação dos Resultados e as Conclusões.

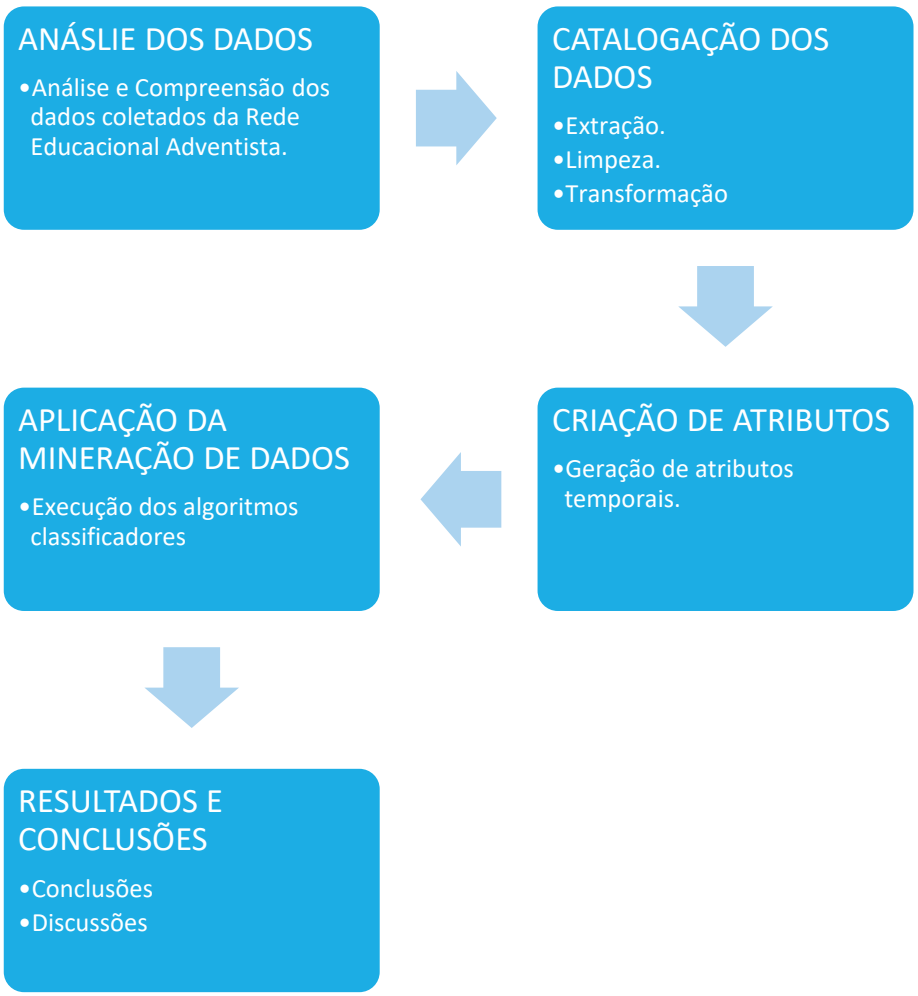

Figura 10: Método proposto dividido cinco etapas, autoria própria (2017).

3.2.1 Análise dos Dados

Diversas informações sobre alunos e responsáveis financeiros estão geralmente disponíveis em um sistema acadêmico. Para a análise dos dados coletados da Rede Educacional Adventista, foi necessário seguir as seguintes etapas:

- Entender a estrutura da base de dados sobre os alunos.
- Encontrar os dados e atributos relacionados aos discentes e seus responsáveis.
- Extrair os dados pertinentes ao contexto de fidelização.

Com os objetivos definidos, é necessário identificar quais os dados disponíveis estão associados às metas pré-definidas.

3.2.2 Catalogação dos Dados e Criação de Atributos

Após a determinação das metas e informações necessárias para alcançá-las, é necessário seguir para o próximo passo, onde prepara-se esses dados adaptando-os em formatos apropriados para as ferramentas de mineração, eliminando ruídos e inconsistências que possam estar presentes. No processo de reorganização dos dados que serão utilizados, geralmente é gerado uma nova estrutura (*data warehouse*) para facilitar e agilizar o acesso aos registros. Este processo também está ligado a preparação dos dados. Para desempenhar esse passo, o especialista em *software* e o especialista em mineração de dados devem trabalhar juntos, sendo que o primeiro possui o conhecimento dos dados extraídos do sistema e o segundo possui o conhecimento de como melhor formatá-lo para aplicar nas técnicas de mineração de dados.

A preparação dos dados envolve algumas etapas como: verificação da consistência dos dados, limpeza e reorganização, quando necessário a normalização e a discretização. A consistência dos dados é validada removendo os ruídos originados normalmente por mal-uso do sistema ou mesmo falha na gravação gerando dados incorretos ou incompletos. A limpeza dos dados é feita removendo dados que não são necessários. A reorganização envolve o processo de adequação do formato dos dados originais para um formato mais fácil de se aplicar na mineração de dados.

Para a implementação e execução deste processo, foi necessário dividir as tarefas em algumas etapas como:

- Pré-processamento das informações;
- Separação das variáveis e atributos mais relevantes;
- Criação de novos atributos baseados nos atributos existentes;
- Balanceamento dos dados.

3.2.3 Aplicação da Mineração de Dados

Para alcançar os objetivos propostos é necessário a aplicação de algumas técnicas de mineração de dados. Conforme já mencionado anteriormente, o problema da fidelização de estudante no contexto da EDM é tratado aqui como um problema de classificação, considerando as classes “Sim” para o caso da fidelização e “Não” para o caso do abandono ou não permanência do estudante na Instituição.

Para a execução deste processo, foram necessários dividir as tarefas nas seguintes etapas:

- Identificação dos algoritmos mais relevantes e com melhor performance para o modelo proposto;

- Aplicação dos algoritmos escolhidos sobre os dados pré-processados para a geração de modelos.

3.2.4 Resultados e Conclusões

Por último é verificado se os resultados obtidos estão alinhados e se são consistentes com as metas estabelecidas. Sendo assim, para um trabalho voltado para classificação, deve-se realizar a análise do modelo e informações adquiridas e se o mesmo obteve o resultado satisfatório utilizando-se uma métrica adequada. Em um eventual resultado insatisfatório, é necessário determinar quais etapas anteriores produziram uma saída inadequada, retornando até este ponto e refazendo os subsequentes.

A fase de resultados e conclusões mostra as informações que foram obtidas nas fases anteriores, os processamentos executados, o resultado desta execução gerando as conclusões, sugestões e trabalhos futuros.

Capítulo 4

Análise da Fidelização com Mineração de Dados Educacionais

Este capítulo apresenta a análise dos dados e a proposta da criação de atributos temporais objeto deste trabalho.

4.1 Análise dos Dados e Definição do Escopo

O conjunto de dados utilizado neste experimento foi extraído da base de dados de uma das maiores redes educacionais do mundo, chamada Rede Educacional Adventista, que está presente em 165 países, representada por 7.783 instituições da educação infantil ao ensino superior, com aproximadamente 90 mil professores e 1,8 milhões de alunos. Na América do Sul existem 888 instituições com 277 mil alunos, e desses 176 mil são residentes no Brasil [Educação Adventista 2016].

Devido à falta de pesquisas baseadas no ensino primário [Marquez-Vera et al. 2013] e a disponibilização dos dados referentes a este grau de ensino pela Rede Educacional Adventista, este trabalho tem como escopo alunos do Ensino Fundamental (1º ao 9º ano). O período letivo escolhido foi entre o ano de 2000 a 2014, visando observar o comportamento acadêmico para todos os alunos durante os anos encontrados neste contexto educacional. Esta mesma base de dados possui informações desde 1998, porém nos anos anteriores não haviam algumas informações utilizadas neste trabalho, pois foram implementadas no sistema acadêmico a partir do ano 2000. Inicialmente o ano de 2015 foi adicionado, porém, foi notado uma discrepância em relação aos outros anos, de forma que o mesmo teve de ser descartado.

Neste trabalho foi utilizado uma amostra composta dos registros de 15.738 alunos localizados na região Sul do Brasil, que inclui os estados do Paraná, Santa Catarina e Rio Grande do Sul.

Para melhor compreender o processo de fidelização dos alunos, antes de se aplicar as técnicas de mineração de dados utilizando os algoritmos de classificação em busca de regras, é importante traçar perfis dos alunos e responsáveis financeiros.

4.1.1 Perfil do Aluno

Para uma primeira análise, os alunos foram “categorizados” por gênero,

sendo encontrada a distribuição de 53% dos alunos do sexo masculino e 47% do sexo feminino como visto na Figura 11. Considerando a diferença de apenas 6% de superioridade na quantidade dos alunos do sexo masculino, pode-se considerar que há um equilíbrio em relação ao gênero.

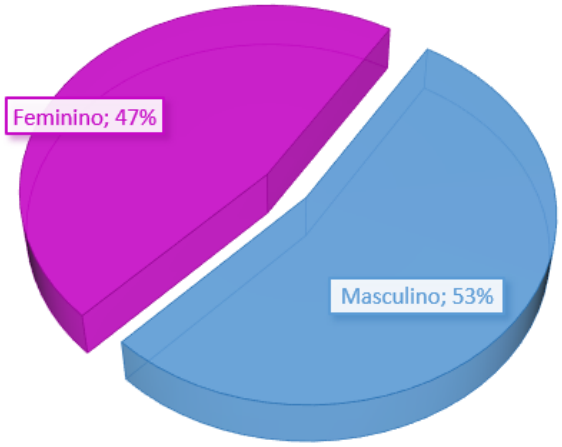

Figura 11: Alunos por gênero, autoria própria (2017)

A informação de faixa etária dos alunos não foi considerada neste trabalho, pois o mesmo é voltado para a educação do ensino fundamental. Analisando os dados, constatou-se que a faixa etária varia única e exclusivamente em relação à série, sendo, portanto, criada uma nova variável chamada MediaNaIdade. Saber se o aluno está na idade ou não para a série correspondente pode ser um indicador útil para compor o perfil do aluno, por isso foi analisado se a idade do aluno está na média da idade dos outros alunos da sala, conforme visto na Figura 12.

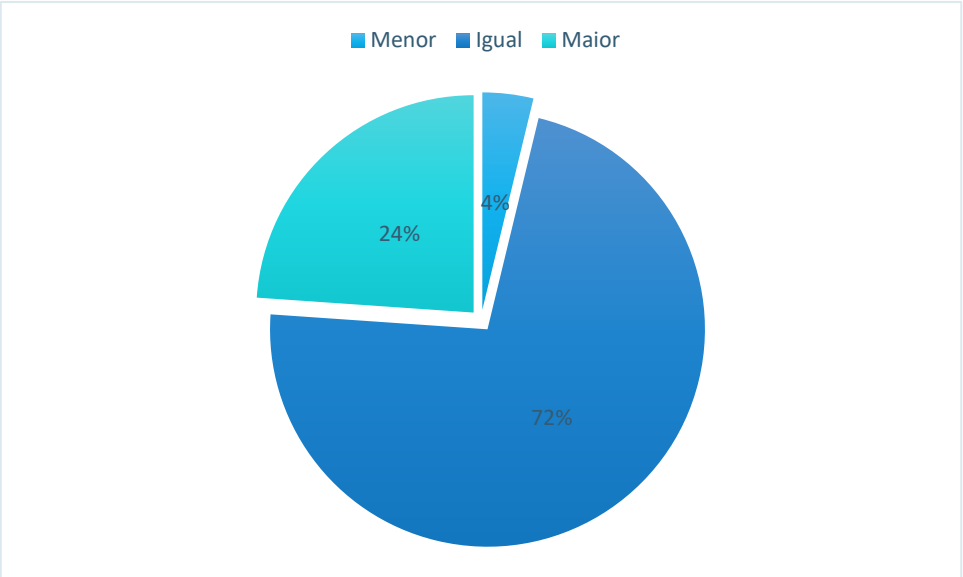

Figura 12: Idade do aluno comparada a idade dos outros alunos da mesma série, autoria própria (2017)

No Ensino Fundamental da Rede Educacional Adventista existem somente

dois turnos de aulas: Manhã e Tarde. Ao traçar o perfil do aluno, pode ser útil a identificação da maior densidade de alunos por turno. Na Figura 13 vemos que ambos os turnos são preenchidos de forma praticamente igual.

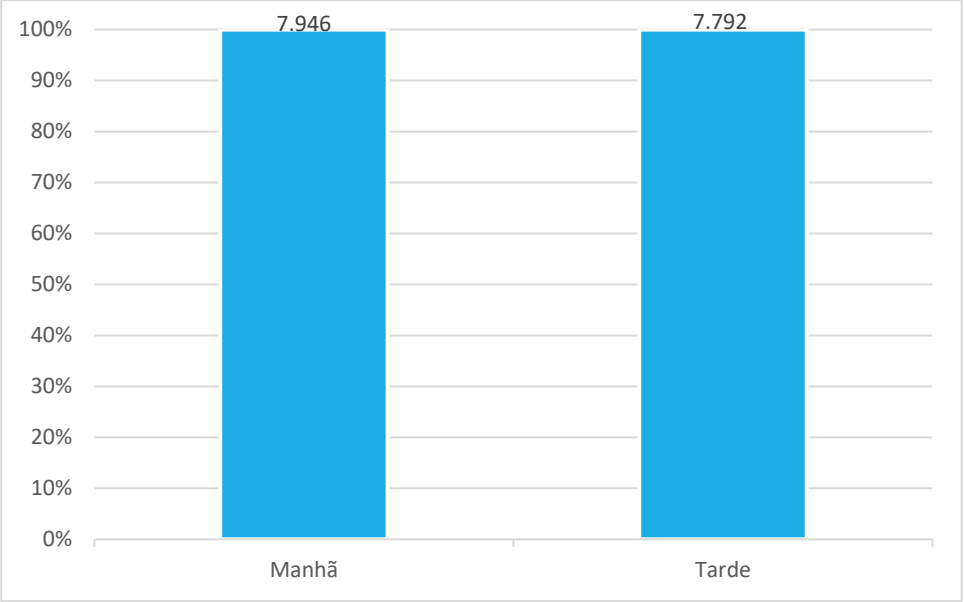

Figura 13: Turno de aula dos alunos, autoria própria (2017)

Baseado na análise dos dados da Rede Educacional Adventista pode-se concluir que o perfil típico dos alunos é bastante homogêneo: cerca de 50% dos alunos são do sexo masculino, normalmente um aluno está na mesma média de idade que os outros alunos da sua série e o mesmo pode estudar tanto de manhã como de tarde.

4.1.2 Perfil do Responsável Financeiro

O responsável financeiro é a pessoa que, perante a instituição de ensino, é o responsável pelo pagamento das mensalidades, não sendo obrigatoriamente o responsável legal. Um exemplo disso é um aluno cujo os pais são separados e um tio arca com as mensalidades. Com esta visão a primeira análise sobre os responsáveis teve como objetivo verificar se os responsáveis financeiros são os responsáveis legais. Na Figura 14 observa-se que na maior parte dos casos o responsável financeiro e responsável legal são a mesma pessoa.

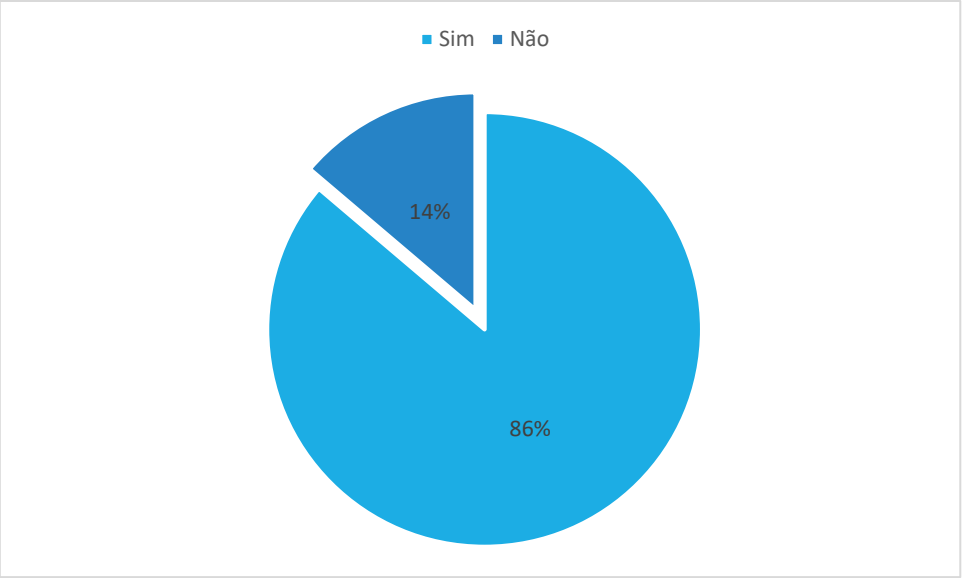

Figura 14: Porcentagem onde o responsável financeiro é responsável legal, autoria própria (2017)

Outra categoria criada foi a distribuição do estado civil dos responsáveis financeiros dos alunos, pois na base de dados original ocorriam valores possíveis como: “Desconhecido", "Solteiro", "Viúvo", "Separado", "Desquitado", "Divorciado", "Maritalmente", "União Estável" e "Casado", que foram reagrupados segundo a Tabela 6. Após a discretização foi verificado que a quantidade de responsáveis “casado” é predominante aos outros estados civis, como indicado à Figura 15.

Tabela 6: Conversão dos valores do estado civil dos responsáveis financeiros para valores discretizados, autoria própria (2017).

Valores Originais	**Valores Reagrupados**
Desconhecido	Solteiro
Solteiro	
Viúvo	
Separado	Separado
Desquitado	
Divorciado	
Maritalmente	Casado
União Estável	
Casado	

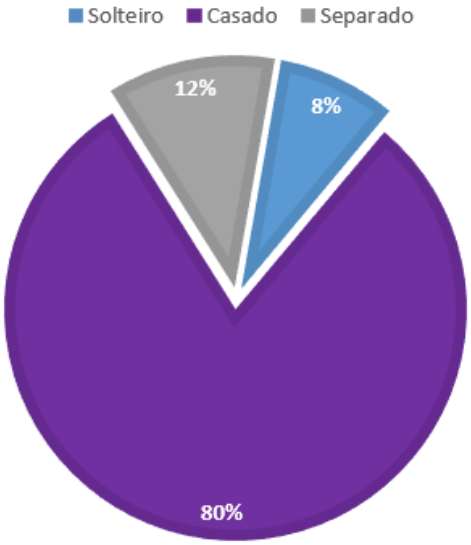

Figura 15: Distribuição dos responsáveis financeiros pelo estado civil, autoria própria (2017)

A categoria de faixa etária dos responsáveis financeiros (Figura 16) mostra que a maioria dos responsáveis se enquadram, segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), na faixa etária entre 31 a 50 anos, havendo poucos jovens ou idosos [IBGE 2010].

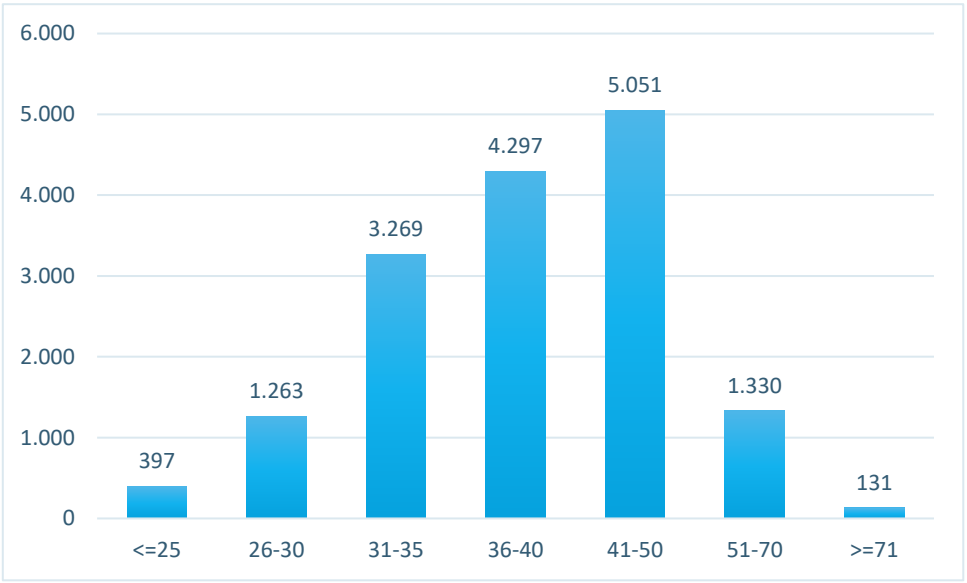

Figura 16: Distribuição dos responsáveis financeiros por faixa etária, autoria própria (2017)

O grau de escolaridade dos responsáveis financeiros também foram utilizados. Verificou-se que a maioria dos responsáveis possuem um grau de instrução acima de Superior Completo, ou seja, possuem algum tipo de Pós-Graduação (stricto-senso ou lato-sensu), seguido por responsáveis com Superior Completo (Figura 17).

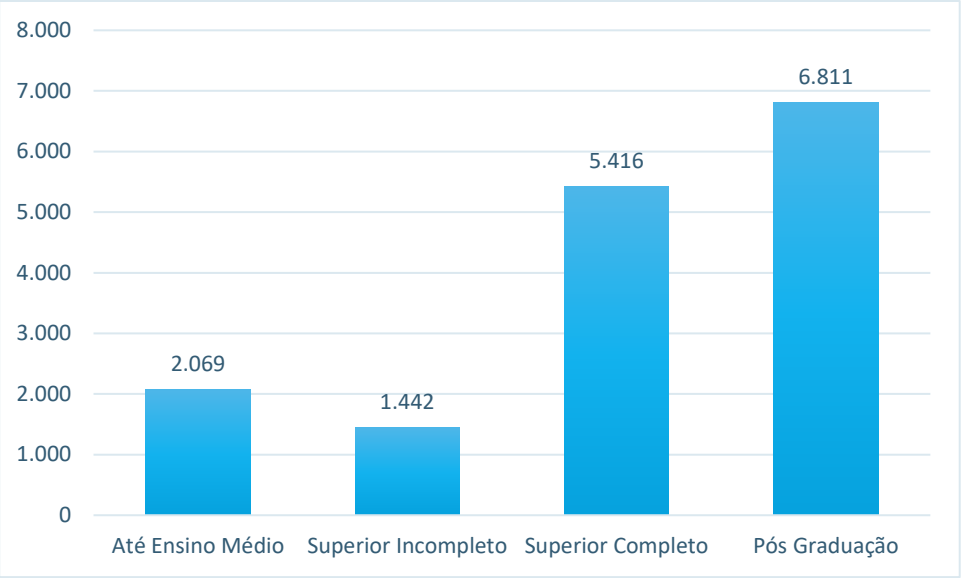

Figura 17: Distribuição dos responsáveis financeiros por grau de escolaridade, autoria própria (2017)

Baseado nos dados analisados, pode-se chegar ao seguinte perfil típico dos responsáveis financeiros: o responsável financeiro normalmente é o responsável legal, é casado, está na faixa etária de adultos e possui ao menos grau superior.

4.1.3 Análise da Fidelização/Evasão

Em média 15,43% dos alunos do Ensino Fundamental não efetuam a rematrícula para o ano posterior, ou seja, mudam para outra escola particular/pública ou mesmo abandonam os estudos, como visto na Figura 18.

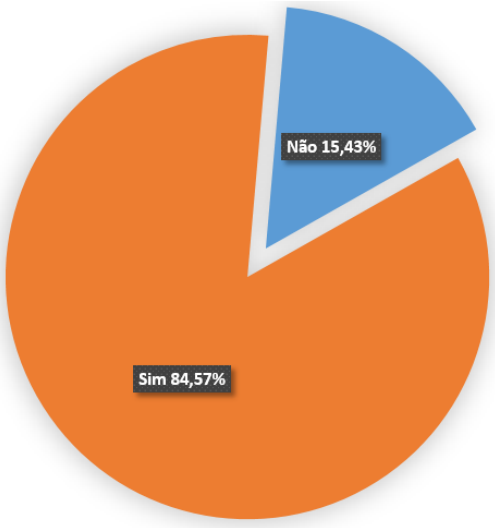

Figura 18: Média geral de alunos que efetuaram a matrícula no ano letivo seguinte, autoria própria (2017)

Numa análise mais aprofundada, separando o índice de retenção por série, percebe-se que com exceção do 5º ano não existe uma diferença relevante entre a retenção relacionada a série do aluno por ano, como visto no gráfico da Figura 19.

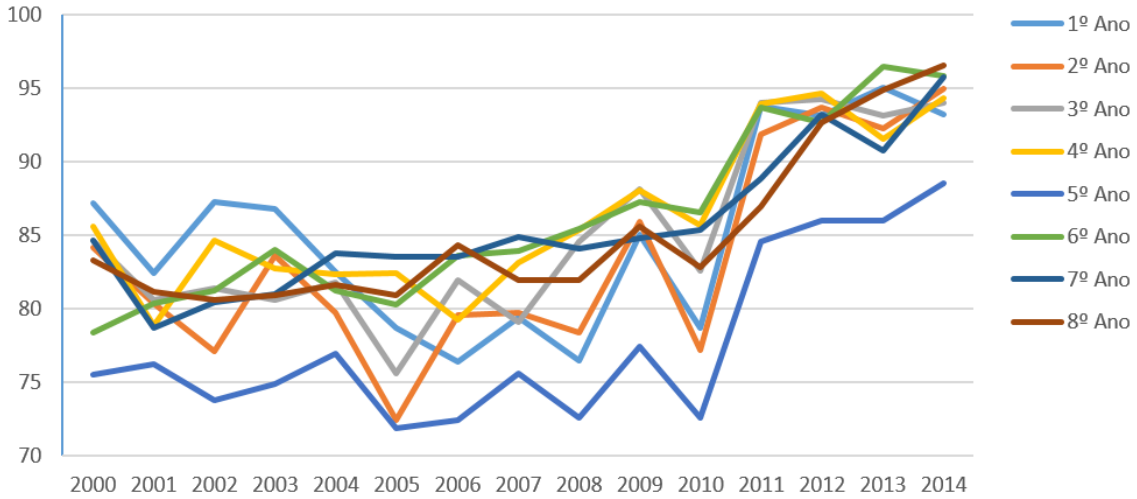

Figura 19: Porcentagem de retenção para a série do 1º ao 8º Ano do Ensino Fundamental entre os anos de 2000 a 2014, autoria própria (2017)

Apesar do Ensino Fundamental possuir séries de 1º ao 9º ano, foram retiradas da base as informações referentes à última série (9º ano), pois o mesmo não possui informações de retenção, pois na sequência não existem mais séries do ensino fundamental.

4.2 Tratamento e Separação dos Dados

4.2.1 Extração de Atributos

Segundo Delen [Delen 2010] as métricas mais utilizadas na separação dos dados em EDM para fidelização e evasão envolvem encontrar atributos relacionados ao contexto social como: estado civil, gênero, grau de instrução e faixa etária. Esses atributos ajudam a encontrar padrões de comportamento esperados em um ambiente educacional.

Para o contexto social, foram encontradas no sistema acadêmico as tabelas apresentadas à Figura 20, obtidas a partir do Sistema de Secretaria Escolar (SSE) utilizado pela Rede Educacional Adventista.

Outras variáveis comumente utilizadas em EDM são os atributos relacionados ao desempenho acadêmico, como notas e controle de presença. Estes atributos tendem a indicar o interesse e integração do aluno em relação a escola, professores e colegas de sala [Tinto 1982].

Como este trabalho trata de instituições privadas, variáveis relacionadas a questões financeiras também são muito pertinentes. Comumente os atributos relacionados ao pagamento das mensalidades e fornecimento de descontos ou bolsas são utilizados para se tentar encontrar algum tipo de padrão [Lauría et al. 2012]. Esses atributos foram encontrados nas Tabelas “Carnês” como visto na Figura 21.

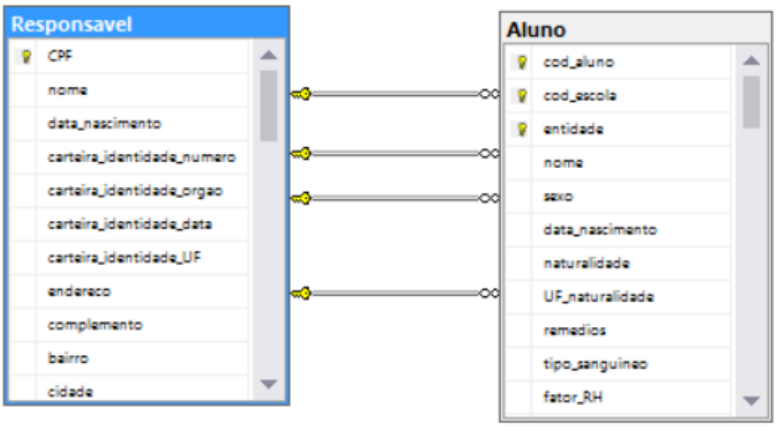

Figura 20: Tabelas que possuem informações dos responsáveis legais/financeiros e alunos no sistema SSE, autoria própria (2017)

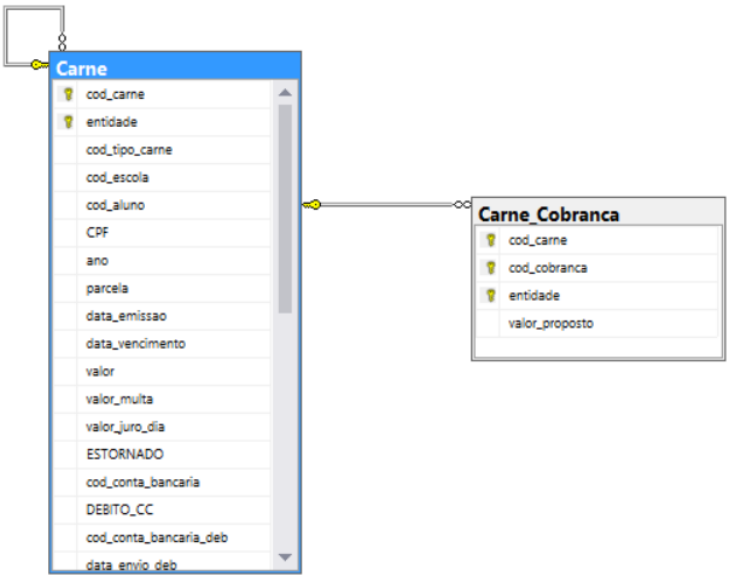

Figura 21: Tabelas que possuem informações dos dados de pagamento no sistema SSE, autoria própria (2017)

A seguir são relacionados todos os atributos extraídos ou gerados através dos dados originais citados acima. Na Tabela 7 pode-se vislumbrar as categorizações e seus possíveis valores destes atributos.

- Atributo **Dependentes_Matriculados**: apresenta a quantidade de alunos que estão vinculados com o mesmo responsável.

- Atributo **TipoAluno**: algumas escolas Adventistas possuem internato; este campo mostra se o aluno é de internato ou externato.

- Atributo **MediaNaIdade**: faz a comparação da idade do aluno com os outros alunos da mesma série.

- Atributo **Genero**: indica o gênero do aluno.

- Atributo **EhAdventista**: neste contexto específico a instituição educacional possui um cunho religioso, portanto se mostrou necessário indicar se o aluno é de uma determinada religião, no caso Adventista do Sétimo Dia.

- Atributo **EstadoCivil**: indica o Estado Civil do responsável financeiro.

- Atributo **GrauInstrucaoResponsavel**: indica o grau de instrução do responsável financeiro.

- Atributo **FaixaEtaria**: indica a faixa etária do responsável financeiro.

- Atributo **TemIrmao**: diferentemente do atributo "Dependentes_Matriculados", este atributo visa mostrar somente se o mesmo possui irmão ou é filho único.

- Atributo **VeioDeForaDoEstado**: indica se o aluno nasceu no mesmo estado em que pertence a escola.

- Atributo **Curso**: indica a série que o aluno frequenta.

- Atributo **AlunoQuantoTempo**: indica a quantos anos o aluno está matriculado na mesma escola.

- Atributo **QtdeAlunos**: indica quantos alunos estão matriculados na mesma sala.

- Atributo **Turno**: indica o turno em que o aluno frequenta.

- Atributo **OcorrenciasBoas**: indica quantas ocorrências de elogio ou bom comportamento o aluno recebeu no ano letivo.

- Atributo **OcorrenciasMaterial**: indica quantas ocorrências de falta de material escolar o aluno recebeu no ano letivo.

- Atributo **OcorrenciasRuins**: indica quantas ocorrências de mal comportamento o

aluno recebeu no ano letivo.

- Atributo **JaReprovou**: indica se o aluno já reprovou alguma vez, nesta escola.

- Atributo **MediaMatematica**: indica a média na disciplina de Matemática que o aluno obteve durante o ano letivo; esta média foi convertida em conceito: (0 – Muito Baixo) <=4, (1 - Baixo) de 4 a 5.9, (2 - Suficiente) de 6 a 6.4, (3 - Regular) de 6.5 a 7.4, (4 - Bom) de 7.5 a 8.4, (5 – Muito Bom) de 8.5 a 9.4 e (6 - Excelente) de 9.5 a 10.

- Atributo **FaltasMatematica**: indica as faltas que o aluno obteve na disciplina de matemática durante o ano letivo.

- Atributo **MediaReligiao**: indica a média na disciplina de religião que o aluno obteve durante o ano letivo; esta média foi convertida em conceito: (0 – Muito Baixo) <=4, (1 - Baixo) de 4 a 5.9, (2 - Suficiente) de 6 a 6.4, (3 - Regular) de 6.5 a 7.4, (4 - Bom) de 7.5 a 8.4, (5 – Muito Bom) de 8.5 a 9.4 e (6 - Excelente) de 9.5 a 10.

- Atributo **FaltasReligiao**: indica as faltas que o aluno obteve na disciplina de religião durante o ano letivo.

- Atributo **MediaBiologia**: indica a média na disciplina de biologia que o aluno obteve durante o ano letivo; esta média foi convertida em conceito: (0 – Muito Baixo) <=4, (1 - Baixo) de 4 a 5.9, (2 - Suficiente) de 6 a 6.4, (3 - Regular) de 6.5 a 7.4, (4 - Bom) de 7.5 a 8.4, (5 – Muito Bom) de 8.5 a 9.4 e (6 - Excelente) de 9.5 a 10.

- Atributo **FaltasBiologia**: indica as faltas que o aluno obteve na disciplina de biologia durante o ano letivo.

- Atributo **MediaHistoria**: indica a média na disciplina de história que o aluno obteve durante o ano letivo; esta média foi convertida em conceito: (0 – Muito Baixo) <=4, (1 - Baixo) de 4 a 5.9, (2 - Suficiente) de 6 a 6.4, (3 - Regular) de 6.5 a 7.4, (4 - Bom) de 7.5 a 8.4, (5 – Muito Bom) de 8.5 a 9.4 e (6 - Excelente) de 9.5 a 10.

- Atributo **FaltasHistoria**: indica as faltas que o aluno obteve na disciplina de História durante o ano letivo.

- Atributo **MediaGeografia**: indica a média na disciplina de Geografia que o aluno obteve durante o ano letivo; esta média foi convertida em conceito: (0 – Muito Baixo) <=4, (1 - Baixo) de 4 a 5.9, (2 - Suficiente) de 6 a 6.4, (3 - Regular) de 6.5

a 7.4, (4 - Bom) de 7.5 a 8.4, (5 – Muito Bom) de 8.5 a 9.4 e (6 - Excelente) de 9.5 a 10.

- Atributo **FaltasGeografia**: indica as faltas que o aluno obteve na disciplina de geografia durante o ano letivo.

- Atributo **MediaIngles**: indica a média na disciplina de inglês que o aluno obteve durante o ano letivo; esta média foi convertida em conceito: (0 – Muito Baixo) <=4, (1 - Baixo) de 4 a 5.9, (2 - Suficiente) de 6 a 6.4, (3 - Regular) de 6.5 a 7.4, (4 - Bom) de 7.5 a 8.4, (5 – Muito Bom) de 8.5 a 9.4 e (6 - Excelente) de 9.5 a 10.

- Atributo **FaltasIngles**: indica as faltas que o aluno obteve na disciplina de inglês durante o ano letivo.

- Atributo **MediaPortugues**: indica a média na disciplina de português que o aluno obteve durante o ano letivo; esta média foi convertida em conceito: (0 – Muito Baixo) <=4, (1 - Baixo) de 4 a 5.9, (2 - Suficiente) de 6 a 6.4, (3 - Regular) de 6.5 a 7.4, (4 - Bom) de 7.5 a 8.4, (5 – Muito Bom) de 8.5 a 9.4 e (6 - Excelente) de 9.5 a 10.

- Atributo **FaltasPortugues**: indica as faltas que o aluno obteve na disciplina de português durante o ano letivo.

- Atributo **MediaEdFisica**: indica a média na disciplina de educação física que o aluno obteve durante o ano letivo; esta média foi convertida em conceito: (0 – Muito Baixo) <=4, (1 - Baixo) de 4 a 5.9, (2 - Suficiente) de 6 a 6.4, (3 - Regular) de 6.5 a 7.4, (4 - Bom) de 7.5 a 8.4, (5 – Muito Bom) de 8.5 a 9.4 e (6 - Excelente) de 9.5 a 10.

- Atributo **FaltasEdFisica**: indica as faltas que o aluno obteve na disciplina de Educação Física durante o ano letivo.

- Atributo **ResponsavelLegalEhFinanceiro**: indica se o responsável legal e financeiro é o mesmo para o aluno.

- Atributo **MesInicio**: dependendo de quando foi feita a matrícula de um aluno, a mensalidade pode iniciar em meses diferentes, como por exemplo um aluno que ingressa em junho, ele somente começará a pagar as mensalidades em julho.

- Atributo **BolsaEDescontosNormalizado**: indica a somatória total de bolsas e/ou descontos que o aluno possui.

- Atributo **CobrouMatricula**: indica se foi cobrada alguma taxa de matrícula no

ingresso do aluno.

- Atributo **QtdeParcelas**: indica em quantas vezes foram parceladas as mensalidades do aluno.

- Atributo **Rematriculou**: indica se ao final do ano letivo o aluno efetuou a matrícula para o próximo ano; no problema considerado este atributo corresponde ao atributo classe.

Tabela 7: Atributos extraídos da base de dados SSE, autoria própria (2017)

Dimensão	Atributo	Valores Possíveis
Social	Dependentes_Matriculados	Número inteiro
	TipoAluno	[Regular-Externato, Interno]
	MediaNaIdade	[Igual,Menor,Maior]
	Genero	[M,F]
	EhAdventista	[0,1]
	EstadoCivil	[Solteiro, Casado, Separado]
	GrauInstrucaoResponsavel	[Acima-de-Graduação,Até-Ensino-Medio,Superior-Completo,Superior-Incompleto]
	FaixaEtaria	[< 25, 26-30, 31-35, 36-40, 41-50, 51-70, >=71]
	TemIrmao	[0,1]
	VeioDeForaDoEstado	[0,1]
Acadêmica	Curso	[1º Ano, ..., 8º Ano]
	AlunoQuantoTempo	Número inteiro
	QtdeAlunos	Número inteiro
	Turno	[M,T,N]
	OcorrenciasBoas	Número inteiro
	OcorrenciasMaterial	Número inteiro
	OcorrenciasRuins	Número inteiro
	JaReprovou	[0,1]
	MediaMatematica	[0,1,2,3,4,5,6]
	FaltasMatematica	Número inteiro
	MediaReligiao	[0,1,2,3,4,5,6]
	FaltasReligiao	Número inteiro
	MediaBiologia	[0,1,2,3,4,5,6]
	FaltasBiologia	Número inteiro

	MediaHistoria	[0,1,2,3,4,5,6]
	FaltasHistoria	Número inteiro
	MediaGeografia	[0,1,2,3,4,5,6]
	FaltasGeografica	Número inteiro
	MediaIngles	[0,1,2,3,4,5,6]
	FaltasIngles	Número inteiro
	MediaPortugues	[0,1,2,3,4,5,6]
	FaltasPortugues	Número inteiro
	MediaEdFisica	[0,1,2,3,4,5,6]
	FaltasEdFisica	Número inteiro
Financeira	ResponsavelLegalEhFinanceiro	[0,1]
	MesInicioParcela	[1,..,12]
	BolsaEDescontosNormalizado	[0 a 5, 6 a 25, 26 a 50, 51 a 75, 76 a 100]
	CobrouMatricula	[0,1]
	QtdeParcelas	Número inteiro
Objetivo da Previsão	Rematriculou	[Sim, Não]

4.2.2 Criação de Novos Atributos

A criação de novos atributos pode ajudar no reconhecimento de informações importantes em um conjunto de dados de forma mais eficiente do que os atributos originais. Em geral, esta criação se fundamenta em uma combinação dos atributos existentes [Liu e Motoda 1998].

Neste trabalho foram criados atributos temporais, aqui denominados “Janela Temporal”, para auxiliar na melhoria do desempenho na classificação relativa à fidelização escolar. O procedimento proposto tende a ser genérico, podendo ser aplicado em outras situações em que um determinado atributo tenha valores formando uma sequência temporal. No contexto deste trabalho o atributo selecionado foi o histórico de pagamento da mensalidade escolar por sua característica temporal.

Considerando que um ano letivo possui um período de 12 meses e cada mês possui uma mensalidade a ser paga, foram classificadas para cada mês três possibilidades referentes ao pagamento baseado na variação temporal:

- Valor “**D**” (pagamento em dia) refere-se ao pagamento em dia, ou seja, o pagamento da mensalidade ocorreu antes ou no dia do vencimento.

- Valor “**A**” (pagamento feito com atraso), refere-se ao pagamento realizado em um

ou mais dias após a data de vencimento da fatura.

- Valor "**N**" (não pago) refere-se à mensalidade que não foi paga até a data da coleta dos dados, excluindo-se as mensalidades que até a data não estavam vencidas.

Na Tabela 8 apresenta-se um exemplo da aplicação para ilustrar o conceito proposto de janela temporal. Nesta ilustração supõe-se que nos meses de Janeiro e Fevereiro os pagamentos foram feitos em dia, no mês de Março o pagamento ocorreu com atraso, em Abril o pagamento foi feito em dia, em Maio com atraso, em Junho não foi feito o pagamento, e assim sucessivamente até Dezembro.

Tabela 8: Exemplo da "Janela Temporal" relativa ao atributo pagamento de mensalidade, autoria própria (2017)

Jan.	Fev.	Mar.	Abr.	Mai.	Jun.	Jul.	Ago.	Set.	Out.	Nov.	Dez.
D	D	A	D	A	D	N	D	D	N	A	A

O próximo passo é selecionar qual o período de tempo em que a "Janela Temporal" será agrupada. Como a separação escolhida foi mensal, é necessário definir a periodicidade deste agrupamento para que se possa criar os novos valores. A periodicidade pode ser agrupada a cada mês, como visto na Tabela 8, ou se pode efetuar o agrupamento por bimestre, trimestre ou quadrimestre, dependendo do problema a ser tratado.

A Tabela 9 mostra um exemplo de como ficaria um agrupamento por trimestre, levando em conta a mesma instância vista na Tabela 8.

Tabela 9: Agrupamento da sequência não sobreposta a cada 3 meses, autoria própria (2017)

Jan.	Fev.	Mar.
D	D	A
↓		
DDA		

Abr.	Mai.	Jun.
D	A	D
↓		
DAD		

Jul.	Ago.	Set.
N	D	D
↓		
NDD		

Out.	Nov.	Dez.
N	A	A
↓		
NAA		

A geração deste novo atributo através do agrupamento ajuda a encontrar comportamentos que não poderiam ser vistos nos dados originais. Independente da escolha da periodicidade, agora é possível identificar se existe alguma relação entre fidelização e a ordem em que são pagas as mensalidades. Neste exemplo é possível identificar que o responsável financeiro efetuou o pagamento corretamente nos primeiros meses, e após ocorreram atrasos ou até mesmo a não efetivação do pagamento. Este procedimento é importante para o processo de fidelização, pois é possível gerar alertas em situações como a deste caso.

Os agrupamentos temporais adotados neste trabalho foram: por bimestre (de

dois em dois meses), trimestre (a cada três meses), e com sobreposição de mês (fazendo o agrupamento a cada 3 meses sequencialmente, com sobreposição da janela temporal).

Para ilustrar, no caso dos atributos bimestrais – que foram portanto subdivididos a cada dois meses – são gerados seis novos atributos conforme indicado nas Tabelas 10 e 11.

Tabela 10: Exemplo de uma instância gerada pela "Janela Temporal" por bimestre não sobreposta, autoria própria (2017).

Jan.	Fev.	Mar.	Abr.	Mai.	Jun.	Jul.	Ago.	Set.	Out.	Nov.	Dez.
D	D	A	D	A	D	N	D	D	N	A	A
↓		↓		↓		↓		↓		↓	
DD		AD		AD		ND		DN		AA	

Tabela 11: Atributos gerados por bimestre, não sobrepostos, autoria própria (2017).

Dimensão	Atributo	Valores Possíveis
Janela Temporal	Primeiro	Combinação de [[A,D,N] | [A,D,N]]
	Segundo	Combinação de [[A,D,N] | [A,D,N]]
	Terceiro	Combinação de [[A,D,N] | [A,D,N]]
	Quarto	Combinação de [[A,D,N] | [A,D,N]]
	Quinto	Combinação de [[A,D,N] | [A,D,N]]
	Sexto	Combinação de [[A,D,N] | [A,D,N]]
Objetivo da Previsão	Rematriculou	[Sim, Não]

Para os atributos trimestrais, que separam a série histórica a cada três meses, são gerados quatro novos atributos como os exemplificados nas Tabelas 9 e 12.

Tabela 12: Atributos gerados por trimestre não sobreposta, autoria própria (2017).

Dimensão	Atributo	Valores Possíveis
Janela Temporal	Primeiro	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Segundo	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Terceiro	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Quarto	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
Objetivo da Previsão	Rematriculou	[Sim, Não]

No caso dos atributos com sobreposição, foram subdivididos de acordo com a periodicidade (de três meses no exemplo) e sobrepostos a cada mês, gerando dez novos atributos novos exemplificados nas Tabelas 13 e 14.

Tabela 13: Agrupamento da sequência sobreposta de mês, autoria própria (2017)

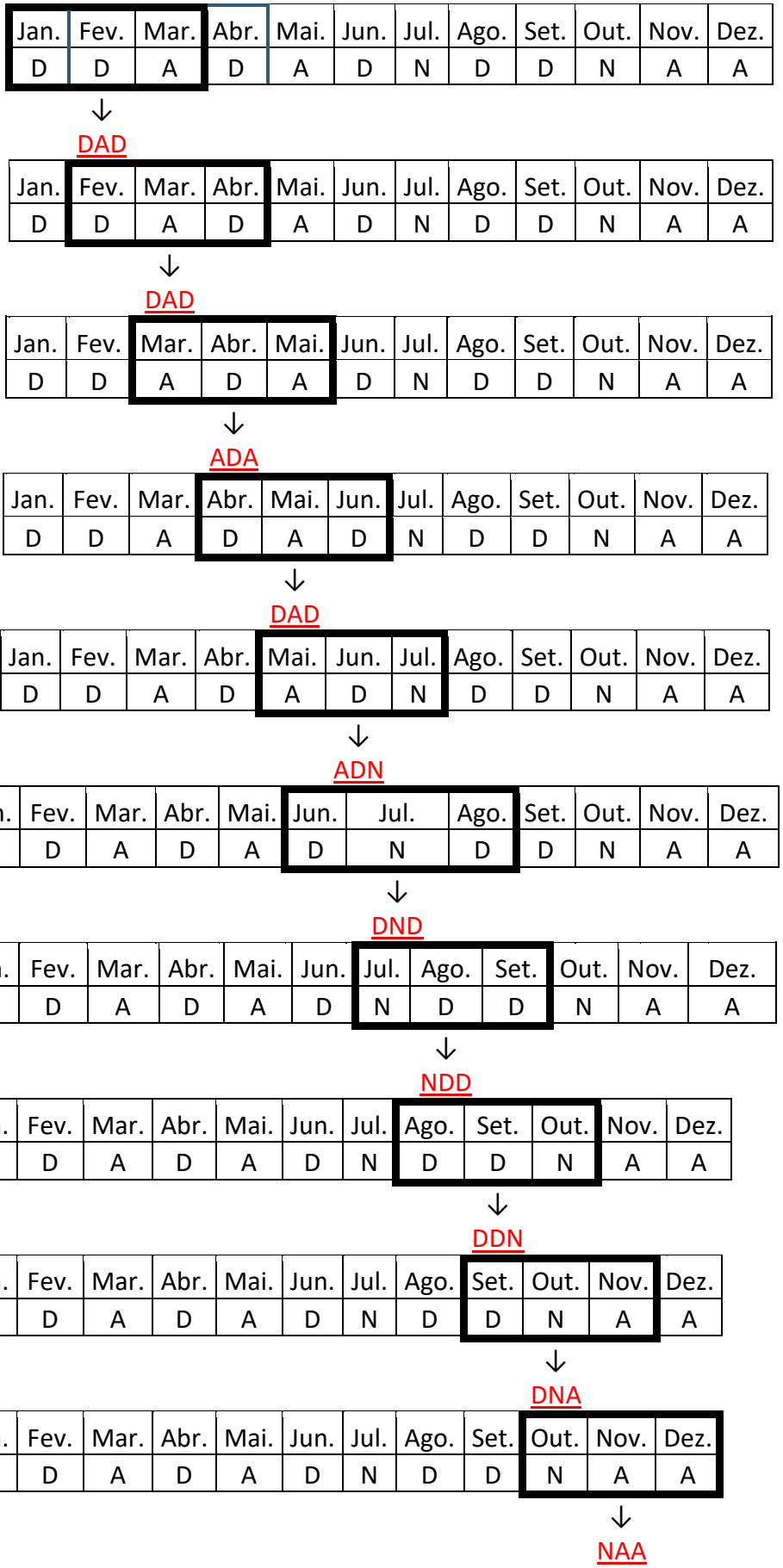

Tabela 14: *Dataset* gerado por trimestre, autoria própria (2017).

Dimensão	Atributo	Valores Possíveis
Janela Temporal	Primeiro	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Segundo	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Terceiro	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Quarto	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Quinto	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Sexto	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Sétimo	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Oitavo	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Novo	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
	Decimo	Combinação de [[A,D,N] | [A,D,N] | [A,D,N]]
Objetivo da Previsão	Rematriculou	[Sim, Não]

4.3 Conclusão

Os atributos propostos neste trabalho e denominados “Janela Temporal” visam identificar um comportamento baseado no período de sequência onde ocorrem os eventos. Estes atributos foram criados com o objetivo de auxiliar os algoritmos de classificação a melhorarem a sua acurácia. Além disto também contribuem para a compreensão pelo usuário da relação entre a fidelização dos alunos e o elemento temporal referente aos pagamentos realizados. Estes objetivos são verificados e comprovados por meio dos procedimentos empíricos descritos no próximo Capítulo.

.

Capítulo 5

Experimentos

Esta seção apresenta os experimentos realizados a partir da implementação proposta, incluindo a descrição dos dados, os experimentos realizados e a análise dos resultados, que visam comprovar que o uso dos atributos temporais criados melhoram a performance dos algoritmos de classificação.

5.1 Experimentos Realizados

Após a análise dos dados foram criados cinco conjuntos de dados (*datasets*). O primeiro *dataset* foi gerado contemplando todos os atributos indicados na seção 4.1.1. O segundo, terceiro e quarto *datasets* foram gerados com dados exclusivamente obtidos pelo procedimento "Janela Temporal", visando tentar identificar qual a melhor forma de agrupamento (a que obtém a melhor acurácia), de acordo com a periodicidade de Bimestre, Trimestre e Período Sobreposto, conforme o descrito nas seções de 5.1.3 a 5.1.6. O quinto e último *dataset* foi criado com base na junção de todos os atributos pré-selecionados e os atributos com a acurácia mais alta na "Janela Temporal", e os experimentos correspondentes são descritos à seção 5.1.7.

Todos os experimentos aqui realizados foram executados utilizando-se o ambiente de mineração de dados Waikato Environment for Knowledge Analysis (WEKA), reconhecido por ser referência em mineração de dados [Hall et al. 2009]. O WEKA foi desenvolvido pela Universidade de Waikato – Nova Zelândia, tendo como principal objetivo prover métodos de análise em conjuntos novos de dados, de uma forma rápida e flexível [WEKA 2017]. Além de uma interface gráfica, o WEKA possui um ambiente para execução em linha de comando e modelagem via *Workflow*, assim como também aceita vários plug-ins que podem nos ajudar nas tarefas de mineração de dados.

5.1.1 Discretização e Normalização dos Dados

Alguns algoritmos são muito influenciados dependendo do tipo e formatação dos dados. Sendo assim os atributos MediaNaIdade, Estado Civil, GrauInstrucaoResponsavel, FaixaEtaria, MediaMatematica, MediaReligiao, MediaBiologia, MediaHistoria, MediaGeografia, MediaIngles, MediaPortugues e MediaEdFisica e BolsaEDescontosNormalizado passaram pelo processo de

discretização, visando a redução do número de valores contínuos para que alguns algoritmos, tais como a Árvore de Decisão, possam ser beneficiados. A limitação e rotulação dos valores possíveis para os atributos citados são vistos na Tabela 7.

5.1.2 Balanceamento dos Dados

Devido ao desbalanceamento dos dados que ocorre nos *datasets*, já que a grande maioria dos alunos na base pertence à classe "Sim" para rematrícula, foi necessário aplicar um algoritmo para balancear os registros. O algoritmo utilizado foi o SMOTE [Chawla et al. 2011], disponível na ferramenta WEKA e de fácil execução neste ambiente.

Inicialmente os dados desbalanceados estavam distribuídos da forma apresentada à Figura 22, sendo que a classe "Sim" para a rematrícula detinha uma superioridade numérica de 5.662 instâncias a mais em relação a classe "Não". A Figura 23 mostra os dados já balanceados após a aplicação da função SMOTE, mostrando que a superioridade numérica da classe "Sim" foi reduzida para 624 instâncias em relação a classe "Não".

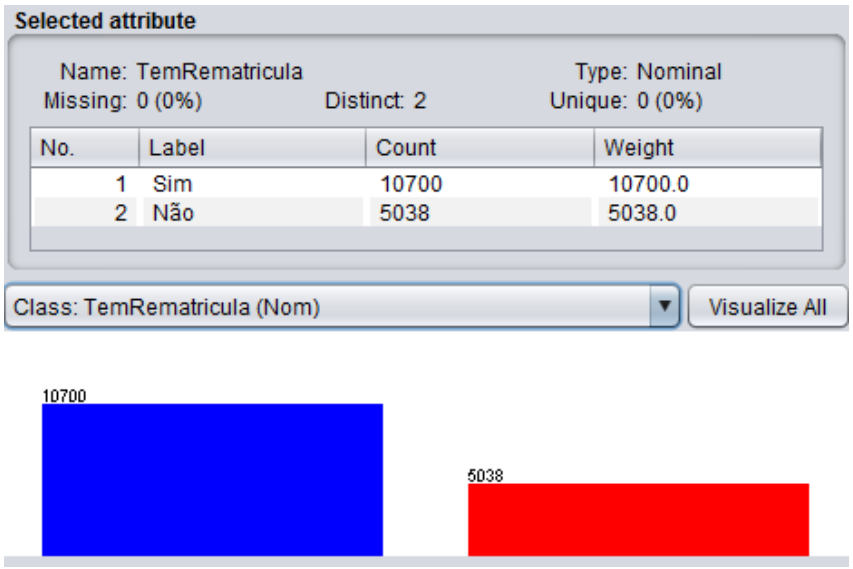

Figura 22: Dados desbalanceados do I experimento, autoria própria (2017).

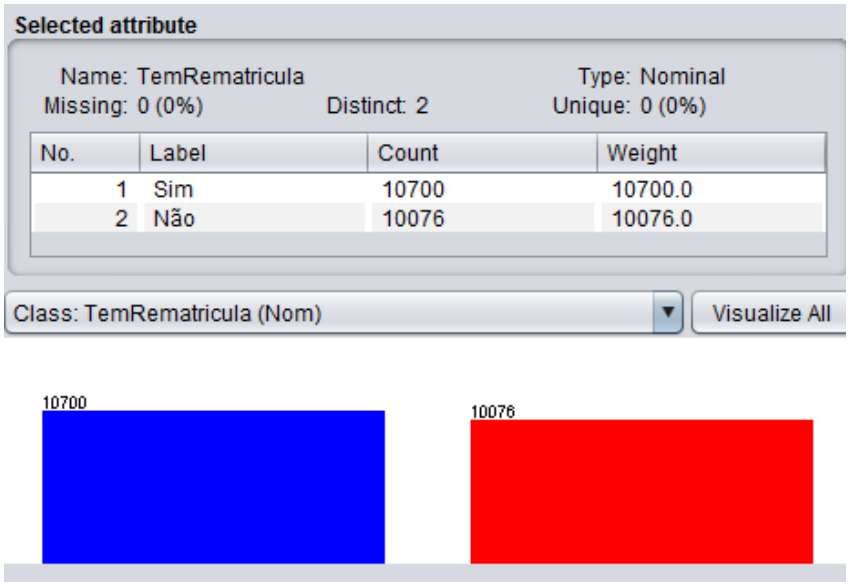

Figura 23: Dados balanceados do I experimento, autoria própria (2017).

5.1.3 Primeiro Experimento: base com os atributos originais

Na primeira abordagem, utilizaram-se os 40 atributos indicados na Tabela 7. Os dados para este *dataset* foram gerados através de consultas realizadas previamente realizadas em uma base de dados Microsoft SQL Server (base de dados do sistema SSE), convertida em um *data warehouse* no qual a estrutura da tabela se encontra no Apêndice C, e exportada para o formato ".ARFF" utilizado pela ferramenta WEKA. Uma amostra deste arquivo pode ser vista no Apêndice D.

Visando encontrar indicadores para auxiliar os gestores educacionais na tarefa de prever a fidelização/evasão e aferir o desempenho dos novos atributos gerados, neste experimento inicial não foram adicionados os atributos criados a partir do procedimento "Janela Temporal".

Este primeiro *dataset* engloba todos os dados dos alunos relacionados a situação socioeconômica e acadêmicas como notas, faltas, grau de escolaridade dos responsáveis financeiros, etc. entre os anos de 2000 a 2014, como mencionado no capítulo 3.2.

Apesar da importância de saber os caminhos e motivos pelo qual se alcançou o resultado, possível por meio dos algoritmos "caixa branca", foram também incluídos nos experimentos os algoritmos de "caixa preta", que não possuem explicitamente a rastreabilidade necessária para se determinar como chegou ao resultado. Estes últimos foram empregados devido à grande quantidade de pesquisas relacionadas a EDM que os utilizaram, e pela necessidade de se comparar a acurácia obtidas nos experimentos pelos diversos algoritmos. Todos os experimentos aqui executados utilizaram o procedimento de validação cruzada de 10 "partições", conforme recomendado por

[Han et al. 2011].

1) *Naïve-Bayes*

Como a classificação do algoritmo *Naïve-Bayes* é por probabilidade e os atributos são considerados independentes, normalmente primeiro executa-se este algoritmo para que se possa ter uma referência inicial que será utilizado na comparação com os demais algoritmos [Zhang 2004].

O algoritmo *Naïve-Bayes* obteve uma acurácia de 81,09%. A Tabela 15 indica o resultado da classificação e a Tabela 16 a matriz de confusão, onde mostra quantitativamente o número de instâncias classificadas da maneira correta e da maneira errada. Por este algoritmo ser considerado "caixa preta", não é possível apresentar uma explicação do processo decisório que levou até o resultado.

Tabela 15: Classificações Corretas e Incorretas para o algoritmo *Naïve-Bayes* no Primeiro experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	16.848	**81,09%**
Classificação Incorreta	3.928	18,91%
Total de Instâncias	20.776	100,00%

Tabela 16: Matriz de confusão para o algoritmo *Naïve-Bayes* no Primeiro experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**8.741**	1.959
Não Rematriculou	1.969	**8.107**

2) SVM

O algoritmo SVM obteve a acurácia de 86,99%. A Tabela 17 mostra a matriz de confusão. Este algoritmo também é considerado "caixa preta", de forma que não é possível apresentar uma explicação do processo decisório que levou até o resultado.

Tabela 17: Matriz de confusão para o algoritmo SVM no Primeiro experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**9.150**	1.560
Não Rematriculou	1.144	**8.923**

3) KNN

O algoritmo KNN foi executado através da função “IBk” implementada no ambiente WEKA. O KNN também é considerado um algoritmo “caixa preta”, não permitindo rastrear o resultado, apenas indicando exemplos semelhantes. Na Tabela 18 são mostradas as classificações corretas e na Tabela 19 a matriz de confusão. A acurácia alcançada foi de 79,25%.

Tabela 18: Classificações Corretas e Incorretas para o algoritmo KNN no Primeiro experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	16.465	**79,25%**
Classificação Incorreta	4.311	20,75%
Total de Instâncias	20.776	100,00%

Tabela 19: Matriz de confusão para o algoritmo KNN no Primeiro experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**8.342**	2.358
Não Rematriculou	1.953	**8.123**

4) Árvore de Decisão

O algoritmo Árvore de Decisão é considerado um algoritmo de “caixa branca”, pois o mesmo permite a compreensão do resultado. Para tanto basta seguir a sequência de testes indicados nos nós da árvore gerada. Neste experimento foi utilizado o algoritmo J48 no ambiente WEKA. Na Tabela 20 é mostrado as classificações corretas e na Tabela 21 a matriz de confusão. A acurácia alcançada foi de 84,92%.

Tabela 20: Classificações Corretas e Incorretas para o algoritmo J48 no Primeiro experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	17.643	**84,92%**
Classificação Incorreta	3.133	15,08%
Total de Instâncias	20.776	100,00%

Tabela 21: Matriz de confusão para o algoritmo J48 no Primeiro experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**9.059**	1.641
Não Rematriculou	1.492	**8.584**

Para este classificador a ferramenta WEKA disponibiliza um gráfico em formato de árvore, mostrando todos os nós e seus ramos até chegar no resultado. Devido a dimensionalidade dos atributos e quantidade de nós gerados pela árvore a apresentação da mesma é feita apenas em formato texto no Apêndice A.

5.1.4 Segundo Experimento: inclusão e uso único dos atributos temporais

Visando identificar a periodicidade do procedimento temporal com melhor performance em relação ao problema proposto, o segundo experimento utilizou exclusivamente os atributos gerados pela "Janela Temporal" e suas respectivas classes de saída ("Sim" e "Não" para Rematrícula). O intuito foi de permitir, após aplicados os classificadores, a realização de uma comparação de acurácia entre as opções consideradas, selecionando a opção de maior acurácia para inseri-la ao *dataset* com os demais atributos já existentes. Baseado nessas informações o segundo *dataset* foi gerado através da periodicidade Trimestral como vistos na Tabela 11. Aplicou-se também o balanceamento dos dados através da função SMOTE.

1) *Naïve-Bayes*

O algoritmo *Naïve-Bayes* obteve uma acurácia de 94,70%. A Tabela 22 indica o resultado da classificação e a Tabela 23 a matriz de confusão.

Tabela 22: Classificações Corretas e Incorretas para o algoritmo *Naïve-Bayes* no Segundo experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	20.451	**94,70%**
Classificação Incorreta	1.144	5,30%
Total de Instâncias	21.595	100,00%

Tabela 23: Matriz de confusão para o algoritmo *Naïve-Bayes* no Segundo experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**11.221**	136
Não Rematriculou	1008	**9230**

2) KNN

O algoritmo KNN foi executado através da função "IBk" implementada no ambiente WEKA. Na Tabela 24 são mostradas as classificações corretas e na Tabela 25 a matriz de confusão. A acurácia alcançada foi de 79,25%.

Tabela 24: Classificações Corretas e Incorretas para o algoritmo KNN no Segundo experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	16.465	**79,25%**
Classificação Incorreta	4.311	20,75%
Total de Instâncias	21.595	100,00%

Tabela 25: Matriz de confusão para o algoritmo KNN no Segundo experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**11.169**	188
Não Rematriculou	734	**9504**

3) Árvore de Decisão

Neste experimento foi utilizado o classificador J48 no ambiente WEKA. Na Tabela 26 são mostradas as classificações corretas e na Tabela 27 a matriz de confusão. A acurácia alcançada foi de 96,04%.

Tabela 26: Classificações Corretas e Incorretas para o algoritmo J48 no Segundo experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	20.740	**95,04%**
Classificação Incorreta	855	3,96%
Total de Instâncias	21.595	100,00%

Tabela 27: Matriz de confusão para o algoritmo J48 no Segundo experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**11.209**	148
Não Rematriculou	707	**9.531**

5.1.5 Terceiro Experimento: base temporal acrescida de atributos temporais trimestrais

Para o terceiro experimento a periodicidade trimestral foi aplicada, gerando quatro novos atributos exemplificados na Tabela 9. Baseado nessas informações o terceiro *dataset* foi gerado como apresentado à Tabela 12. Aplicou-se também o balanceamento dos dados através da função SMOTE.

1) *Naïve-Bayes*

O algoritmo *Naïve-Bayes* obteve uma acurácia de 94,83%. A Tabela 28 indica o resultado da classificação e a Tabela 29 a matriz de confusão.

Tabela 28: Classificações Corretas e Incorretas para o algoritmo *Naïve-Bayes* no Terceiro experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	19.702	**94,83%**
Classificação Incorreta	1.074	5,17%
Total de Instâncias	20.776	100,00%

Tabela 29: Matriz de confusão para o algoritmo *Naïve-Bayes* no Terceiro experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**10.579**	121
Não Rematriculou	953	**9.123**

2) KNN

O algoritmo KNN foi executado através da função "IBk" do ambiente WEKA. Na Tabela 30 é mostrado as classificações corretas e na Tabela 31 a matriz de confusão. A acurácia alcançada foi de 95,80%.

Tabela 30: Classificações Corretas e Incorretas para o algoritmo KNN no Terceiro experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	19.903	**95,80%**
Classificação Incorreta	873	4,20%
Total de Instâncias	20.776	100,00%

Tabela 31: Matriz de confusão para o algoritmo KNN no Terceiro experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**10.520**	180
Não Rematriculou	693	**9.383**

3) Árvore de Decisão

Neste experimento foi utilizado o classificador J48 no ambiente WEKA. Na

Tabela 32 é mostrado as classificações corretas e na Tabela 33 a matriz de confusão. A acurácia alcançada foi de 96,80%.

Tabela 32: Classificações Corretas e Incorretas para o algoritmo J48 no Terceiro experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	20.111	**96,80%**
Classificação Incorreta	665	3,20%
Total de Instâncias	20.776	100,00%

Tabela 33: Matriz de confusão para o algoritmo J48 no Terceiro experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**10.556**	144
Não Rematriculou	739	**9.337**

5.1.6 Quarto Experimento: atributos temporais com sobreposição de meses

Para o Quarto experimento a periodicidade foi subdividida sobrepondo um mês ao outro com o intervalo de três meses, gerando dez novos atributos indicados à Tabela 13. Baseado nessas informações o quarto *dataset* foi gerado como visto à Tabela 14. Aplicou-se também o balanceamento dos dados através da função SMOTE.

1) *Naïve-Bayes*

O algoritmo *Naïve-Bayes* obteve uma acurácia de 94,62%. A Tabela 34 indica o resultado da classificação e a Tabela 35 a matriz de confusão.

Tabela 34: Classificações Corretas e Incorretas para o algoritmo *Naïve-Bayes* no Quarto experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	20.434	**94,62%**
Classificação Incorreta	1.161	5,38%
Total de Instâncias	21.595	100,00%

Tabela 35: Matriz de confusão para o algoritmo *Naïve-Bayes* no Quarto experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**11.228**	129
Não Rematriculou	1032	**9.206**

2) KNN

O algoritmo KNN obteve uma acurácia de 95,82%. Na Tabela 36 são mostradas as classificações corretas e na Tabela 37 a matriz de confusão.

Tabela 36: Classificações Corretas e Incorretas para o algoritmo KNN no Quarto experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	20.693	**95,82%**
Classificação Incorreta	902	4,18%
Total de Instâncias	21.595	100,00%

Tabela 37: Matriz de confusão para o algoritmo KNN no Quarto experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**11.167**	190
Não Rematriculou	712	**9.526**

3) Árvore de Decisão

O classificador J48 obteve uma acurácia de 95,89%. Na Tabela 38 são mostradas as classificações corretas e na Tabela 39 a matriz de confusão.

Tabela 38: Classificações Corretas e Incorretas para o algoritmo J48 no Quarto experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	20.707	**95,89%**
Classificação Incorreta	665	4,11%
Total de Instâncias	20.776	100,00%

Tabela 39: Matriz de confusão para o algoritmo J48 no Quarto experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**11.199**	158
Não Rematriculou	730	**9.508**

5.1.7 Quinto Experimento: com os atributos originais adicionados aos temporais trimestrais

Após a execução dos experimentos da "Janela Temporal", verificou-se no terceiro experimento que a periodicidade trimestral obteve a melhor acurácia, principalmente para Árvore de Decisão (ver Tabela 40). Sendo assim acrescentou-se aos atributos originais do primeiro experimento os quatro novos atributos gerados pelo procedimento "Janela Temporal" com periodicidade trimestral ao *dataset*.

Tabela 40: Comparativo entre o Segundo, Terceiro e Quarto experimentos, autoria própria (2017).

Experimentos	Acurácia (%)		
	KNN	Naïve-Bayes	Árvore Decisão
Segundo (3 Meses)	95,80	**94,83**	**96,80**
Terceiro (2 Meses)	79,25	94,70	95,04
Quarto (Sobreposto)	**95,82**	94,62	95,89

Com os *datasets* assim preparados, se faz necessário a comparação dos resultados de classificação com e sem os atributos temporais (ver seção 4.4.3), visando aferir a melhoria na performance da classificação que é devida à introdução dos atributos temporais gerados.

1) ***Naïve-Bayes***

O algoritmo *Naïve-Bayes* obteve uma acurácia de 83%. A Tabela 41 indica-se o resultado da classificação e a Tabela 42 a matriz de confusão.

Tabela 41: Classificações Corretas e Incorretas para o algoritmo *Naïve-Bayes* no Quinto experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	17.244	**83%**
Classificação Incorreta	3.532	17%
Total de Instâncias	20.776	100%

Tabela 42: Matriz de confusão para o algoritmo *Naïve-Bayes* no Quinto experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**8.832**	1.868
Não Rematriculou	1.664	**8.412**

2) KNN

O algoritmo KNN obteve nesta base uma acurácia de 93,64%. Na Tabela 43 são apresentadas as classificações corretas e na Tabela 44 a matriz de confusão.

Tabela 43: Classificações Corretas e Incorretas para o algoritmo KNN no Quinto experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	19.454	**93,64%**
Classificação Incorreta	1.322	6,36%
Total de Instâncias	20.776	100,00%

Tabela 44: Matriz de confusão para o algoritmo KNN no Quinto I experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**9.907**	793
Não Rematriculou	529	**9.547**

3) Árvore de Decisão

O algoritmo Árvore de Decisão alcançou uma acurácia de 96,57%. Na Tabela 45 são mostradas as classificações corretas e na Tabela 46 a matriz de confusão.

Tabela 45: Classificações Corretas e Incorretas para o algoritmo J48 no Quinto experimento, autoria própria (2017).

	Instâncias	Percentual
Classificação Correta	20.063	**96,57%**
Classificação Incorreta	713	15,08%
Total de Instâncias	20.776	100,00%

Tabela 46: Matriz de confusão para o algoritmo J48 no Quinto experimento, autoria própria (2017).

	Rematriculou	Não Rematriculou
Rematriculou	**10.572**	128
Não Rematriculou	585	**9.491**

Recorda-se que o algoritmo Árvore de Decisão é “caixa branca”, isto é, permite que se obtenha uma explanação de como o resultado foi obtido, o que é importante no problema considerado. Assim como no Primeiro experimento, devido ao tamanho e complexidade da árvore criada (406 nós), a mesma é apresentada integralmente em formato texto no Apêndice B.

5.2 Interpretação dos Resultados

No decorrer do desenvolvimento dos experimentos observou-se que existem algumas variáveis como “Quantidade de Alunos em Sala de Aula”, “Gênero” e “Quantidade de Ocorrências Ruins” que geram bons indicadores na predição da fidelização e evasão em instituições de ensino. Observa-se ainda que não é possível encontrar um atributo específico que se destacasse em relação à predição de evasão desejada antes de algum período de registro da interação entre o aluno e/ou seu responsável e a Instituição.

No primeiro experimento, que utiliza todos os atributos originais sem a “Janela Temporal”, é possível verificar que apesar da quantidade de atributos, a acurácia para todos os classificadores se manteve similar, com uma taxa mínima de 79%. Este resultado indica a potencial qualidade dos dados relacionados, sendo uma acurácia consideravelmente elevada comparada a outros trabalhos do contexto educacional. Apesar da complexidade da Árvore de Decisão obtida neste experimento, detectou-se que pode ser viável a identificação da evasão por meio dos atributos selecionados e trabalhados, permitindo a tomada de decisões que visem minimizar o problema.

No segundo, terceiro e quarto experimentos comparou-se o desempenho entre as periodicidades possíveis no procedimento “Janela Temporal” para se identificar o período com maior relevância em relação à tarefa em questão. Por meio da comparação entre bimestre e trimestre, a acurácia obtida com a periodicidade trimestre obteve um melhor desempenho. Em seguida a periodicidade trimestral foi utilizada para verificar se a sobreposição produziria um melhor desempenho, como indicado à Figura 24. A utilização da periodicidade trimestral sem sobreposição obteve a melhor acurácia.

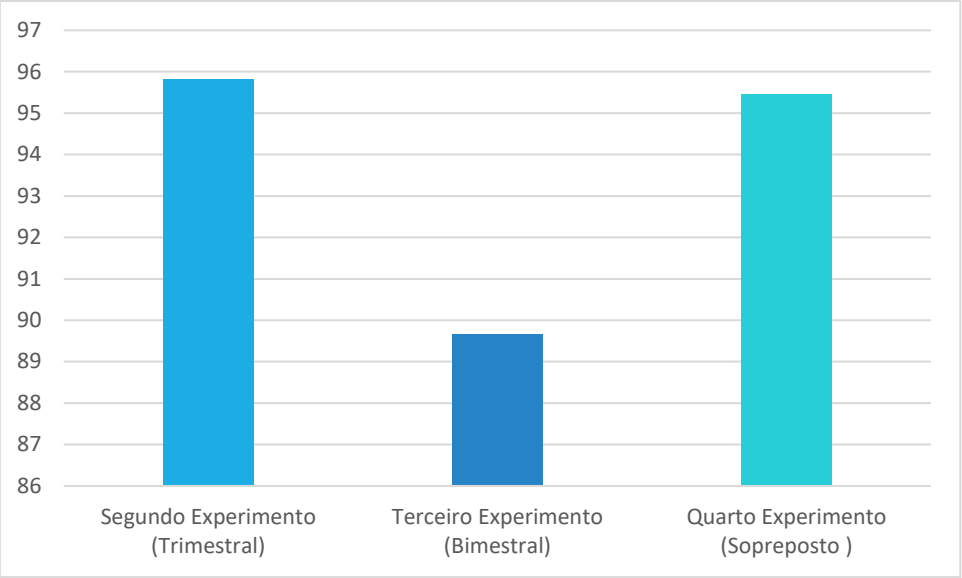

Figura 24: Comparativo da acurácia entre o Segundo, Terceiro e Quarto experimentos, autoria própria (2017).

No quinto experimento foi possível observar que adicionando-se à base original os atributos gerados pelo procedimento "Janela Temporal" com periodicidade trimestral é possível obter um ganho de até 14,39% na acurácia, dependendo do classificador (ver Tabela 47). Todos os algoritmos obtiveram um ganho na acurácia, sendo que o resultado do algoritmo de Árvore de Decisão teve o valor mais elevado de acurácia: 96,57%.

Outro detalhe importante é identificar o motivo pelo qual está ocorrendo a evasão. Apesar do tamanho e complexidade da árvore gerada, apresentada no Apêndice B, é possível identificar alguns comportamentos que predizem a fidelização ou a evasão. Percorrendo a árvore gerada detectou-se que quando se usa as variáveis temporais, somente é possível identificar uma possível evasão a partir do terceiro trimestre. Se houver um comportamento de três atrasos consecutivos no pagamento da mensalidade, na maioria dos casos (85%) a tendência é a não efetivação da rematrícula para o próximo ano na instituição. Segundo a árvore gerada, assim que ocorrem os três primeiros atrasos é necessário olhar para o primeiro trimestre: se neste trimestre já ocorriam atrasos então o aluno tende a permanecer na instituição, no caso contrário deve ocorrer a evasão do aluno.

Tabela 47: Comparativo entre os experimentos Primeiro, Quinto, autoria própria (2017).

Experimentos	Acurácia (%)		
	Primeiro	Quinto	Melhora
KNN	79,25%	93,64%	14,39%
Naïve Bayes	81,09%	83,00%	1,91%
Árvore de Decisão	84,92%	96,57%	11,65%

Outra constatação obtida por meio da análise da Árvore de Decisão gerada pelo quinto experimento foi a relação entre a quantidade de faltas do aluno na disciplina de Biologia e a fidelização. Os alunos que obtiveram uma maior presença nas aulas tendem a efetuar a rematrícula e os alunos que possuem uma menor presença tendem a não efetuar a rematrícula para o próximo ano.

Uma asserção que foi inicialmente indicada pelos gestores educacionais como de importância, e que não foi confirmada, é o valor de desconto e bolsas dados aos alunos. Segundo a árvore gerada, o índice de fidelização entre os alunos que obtiveram um percentual maior de bolsa ou desconto foi o mesmo dos alunos que não efetuaram a rematrícula para o próximo ano, indicando assim, que as bolsas e descontos não estão diretamente relacionadas à evasão.

Em vários nós da árvore gerada o atributo relacionado a quantidade de alunos em sala de aula para uma mesma turma indicou que, quanto menor a quantidade de alunos, maior a probabilidade de permanência do mesmo na Instituição.

O atributo que indica se o responsável financeiro é o mesmo que o responsável legal (ver Tabela 4) mostrou, na grande maioria dos casos, que se estes responsáveis não são os mesmos, o aluno tende a não efetuar a rematrícula para o ano seguinte.

5.3 Conclusão

Como visto neste capítulo a criação de novos atributos pode beneficiar os classificadores, ajudando-os a obterem uma acurácia superior em relação a aplicação em bases de dados utilizadas somente com os atributos originais. A inclusão de atributos que consideram variação temporal também auxilia na previsão do comportamento de evasão, como visto nos experimentos.

Os resultados obtidos também mostram que os experimentos realizados geraram uma importante contribuição para a área de EDM, pois o processo de agrupamento das informações através do procedimento “Janela Temporal” mostrou-se eficaz em relação ao aumento da acurácia em todos os algoritmos de classificação, e em especial para a Árvore de Decisão. Este resultado é particularmente importante pois o algoritmo é o mais adequado para a tarefa do tipo “caixa branca”, permitindo a análise das causas da evasão.

Capítulo 6

Conclusões e Perspectivas

Este trabalho buscou encontrar formas de melhorar a efetividade de algoritmos de Mineração de Dados em bases de dados que se caracterizam pela sua construção incremental, isto é, que são alimentadas periodicamente com os mesmos elementos de dados, formando sequências de valores para cada atributo.

Para tanto, este trabalho propôs um método para geração de atributos baseados em “Janelas Temporais”, que objetivam melhor o desempenho obtido pelos algoritmos aplicados à tarefa de mineração de dados.

A proposta da criação do método foi avaliada no contexto de EDM, tratando-se como aplicação o problema da evasão de alunos em instituições de ensino privadas para o Ensino Fundamental. Constata-se porém que o procedimento pode ser aplicado em outras áreas que possuam uma similar variação temporal dos dados.

Com o aumento da concorrência e o alto custo operacional, encontrar maneiras inteligentes e efetivas para se entender os fatores que levam um estudante de ensino fundamental a permanecer em uma mesma instituição educacional particular por vários anos, tem sido um grande desafio para os gestores educacionais e financeiros. Dados capazes de oferecer um resultado apropriado ao gestor podem ser encontrados dentro de sistemas educacionais, porém o tempo e a complexidade envolvida na extração e análise dos dados, torna esta opção algo impraticável pois na maioria dos casos há uma grande quantidade de cursos, alunos e responsáveis financeiros envolvidos neste contexto. Há poucos sistemas educacionais que disponibilizam ferramentas apropriadas para a realização deste tipo de análise.

Visando encontrar soluções para o problema, este trabalho apresentou uma revisão bibliográfica na área da Mineração de Dados Educacionais, com foco em aspectos computacionais relacionados ao *data warehousing*, à discretização e ao balanceamento de dados, à seleção de melhores atributos e criação de atributos temporais, ao uso de algoritmos de classificação e métricas para avaliação de sua performance.

Nos trabalhos relacionados verificou-se que esta é uma área de pesquisa recente, muito ativa mas pouco desenvolvida aqui no Brasil, sendo um dos motivos que levaram a escolha do tema em questão. Verificou-se também que a grande maioria das pesquisas em EDM são realizadas com o foco no ensino superior ou ensino a distância

deixando uma lacuna aberta para o ensino médio e ensino fundamental. Em muitos casos, chegou-se à conclusão que isto ocorre devido a maior facilidade de acesso aos dados educacionais de ensino superior, uma vez que estando em uma instituição de nível superior a própria instituição libera seus dados para se fazer a pesquisa. Contudo para as instituições de ensino médio ou fundamental, é necessário solicitar os dados a gestores que nem sempre estão dispostos a compartilha-los.

Outro ponto identificado nos trabalhos relacionados é o fato de que várias técnicas de classificação vêm sendo utilizadas, sendo que na grande maioria dos trabalhos são exploradas mais de uma técnica. O classificador mais citado é a Árvore de Decisão, sendo que em 80% dos artigos aparecem o mesmo sendo usado. A escolha da Árvore de Decisão é resultado da sua transparência do modelo gerado, pois permite identificar os atributos utilizados e as regras geradas, facilitando o trabalho do gestor em identificar os fatores relevantes para atacar o problema.

A análise do tipo de desempenho dos alunos é tratada principalmente com os temas de abandono e desempenho no curso/disciplina, apesar de existirem diferentes interpretações do desempenho e abandono. As classes de evasão e abandono, tendem a ser binárias, indicando se o aluno evadiu ou não, assim como feito neste trabalho, já o desempenho pode apresentar uma diversificação maior de opções, classificando de duas a “n” classes distintas.

Em relação aos conjuntos de atributos preditores, há uma tendência de iniciar o processo com todos os atributos disponíveis e durante a elaboração do modelo adicionar ou remover outros atributos conforme o refinamento utilizado. Os preditores mais utilizados estão relacionados a dados socioeconômicos, notas e faltas.

Baseado na metodologia desenvolvida, foi possível examinar e delimitar etapas distintas do processo necessário para a efetivação da mineração de dados sobre os dados em ambiente educacional. Foi possível organizar o fluxo de trabalho realizado e a identificar os problemas que acontecem em cada etapa, permitindo assim sua resolução no momento em que os mesmos ocorriam. A metodologia adotada não evita completamente problemas particulares que ocorrem quando há inconsistências nas informações disponíveis, porém ficou evidente que onde os dados se revelaram insuficientes, incompletos ou inadequados, a aplicação da função ou processo adequado para sua correção influenciaram positivamente para que gerassem resultados significativos no final do processo, auxiliando efetivamente os gestores educacionais. Um exemplo desta situação foi a aplicação da técnica de balanceamento de dados que através da função SMOTE, onde foram geradas instancias sintéticas para que equilibrasse a quantidade de classes “Sim” e “Não” relacionadas ao processo de fidelização.

A criação de atributos temporais na forma de janelas é o ponto de originalidade do trabalho. Os atributos temporais propostos, constituídos por valores sucessivos agrupados, formando janelas nas opções com e sem sobreposição, contribuíram de forma efetiva para a melhoria da performance dos classificadores empregados na tarefa alvo de mineração de dados, como anteriormente descrito. Este procedimento foi de grande influência no processo de identificação de padrões e na aplicação de medidas preventivas para assegurar a fidelização de estudantes no ensino médio e/ou fundamental.

Entre todos os algoritmos de classificação utilizados, o que obteve a melhor acurácia foi o classificador de Árvore de Decisão J48, que após a adição dos atributos gerados atingiu uma acurácia de 96,57%. Apesar de ser considerado um algoritmo de "caixa-branca", a árvore obtida é de grande tamanho, de forma que a análise e compreensão do problema se tornou um pouco complexa. Outros algoritmos aqui aplicados também obtiveram uma melhora em suas acurácias, como o SVM, o *Naïve-Bayes* e o KNN.

A geração do *Data Warehouse* (ver apêndice C) foi fundamental para se alcançar o êxito nesta pesquisa, pois com ele se tornou mais simples a extração dos dados para os formatos exigidos pela ferramenta WEKA, e a formatação dos dados para aplicação nos algoritmos de classificação utilizados neste trabalho.

Com a abordagem computacional proposta neste trabalho foi possível a identificação de padrões comportamentais de alunos e responsáveis financeiros, o que permite aos gestores financeiros e educacionais utilizarem medidas capazes de auxiliar e tratar de forma adequada o problema de fidelização, tendo este trabalho contribuído à área de EDM. A aplicação do modelo gerado irá ajudar a Instituição na previsão da evasão de estudantes, auxiliando no programa de fidelização e retenção de alunos.

Como trabalhos futuros envolvendo a aplicação da metodologia proposta, sugere-se a aplicação dos procedimentos a uma base de dados de menor dimensão ou com o auxílio de algoritmos de seleção de atributos, diminuindo sua quantidade de elementos envolvidos e permitindo assim a geração de árvores menos complexas. Esta abordagem seria adequada para o uso em condições mais controladas e favoráveis à geração de indicadores relacionados a fidelização.

Em relação a EDM esta pesquisa pode ser ampliada aplicando o modelo proposto a outras instituições, como as de nível superior e/ou de educação a distância, ou mesmo aplicando outros classificadores não abordados neste trabalho para a avaliação dos resultados.

Em relação aos atributos temporais criados e seu uso em contextos

educacionais, é possível considerar sua aplicação a outro atributo que não esteja relacionado à variação financeira, visando aferir se o mesmo obtém a mesma melhoria obtida em relação a acurácia.

Também é possível a aplicação dos procedimentos relacionados ao conceito de “Janela Temporal” a outros contextos que não sejam educacionais. Em particular, seria interessante avaliar o uso dos atributos propostos em contextos eminentemente financeiros, como no caso de séries temporais de valores monetários.

O resultado resultante da aplicação alvo deste trabalho pode ser considerado um ponto de partida para trabalhos que visam fornecer informações relevantes a fidelização e abandono de alunos em instituições públicas e privadas, com o intuito de auxiliar gestores educacionais e financeiros a tomarem melhores decisões para evitar este problema.

Referências Bibliográficas

Agrawal, R.; Imieliński, T.; Swami, A. (1993). Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*. , SIGMOD '93. ACM. http://doi.acm.org/10.1145/170035.170072, [acessado em Jul 18].

Agrawal, R.; Mannila, H.; Srikant, R.; Toivonen, H.; Verkamo, A. I. (1996). Advances in Knowledge Discovery and Data Mining pg 307-328. In: Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R.[Eds.]. . Menlo Park, CA, USA: American Association for Artificial Intelligence. p. 307–328.

Anzanello, M. J.; Fogliatto, F. S.; Rossini, K. (2011). Data mining-based method for identifying discriminant attributes in sensory profiling. *Food Quality and Preference*, v. 22, n. 1, p. 139–148.

Bettini, C.; Wang, X. S.; Jajodia, S. (1996). Testing Complex Temporal Relationships Involving Multiple Granularities and Its Application to Data Mining (Extended Abstract). In *Proceedings of the Fifteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. , PODS '96. ACM. http://doi.acm.org/10.1145/237661.237680, [acessado em Jul 18].

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. (9 jun 2011). SMOTE: Synthetic Minority Over-sampling Technique. *arXiv:1106.1813 [cs]*,

Chen, W.-S.; Du, Y.-K. (mar 2009). Using Neural Networks and Data Mining Techniques for the Financial Distress Prediction Model. *Expert Syst. Appl.*, v. 36, n. 2, p. 4075–4086.

Cios, K. J.; Pedrycz, W.; Swiniarski, R. W.; Kurgan, L. A. (2007). *Data Mining: A Knowledge Discovery Approach*. 2007 edition ed. Springer.

Dean, T. L.; Mcdermott, D. V. (1987). Temporal data base management. *Artificial Intelligence*, v. 32, n. 1, p. 1–55.

Delen, D. (nov 2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, v. 49, n. 4, p. 498–506.

Devmedia (2013). Como definir um Data Warehouse na prática usando Transact SQL. http://www.devmedia.com.br/como-definir-um-data-warehouse-na-pratica-usando-transact-sql/32633, [acessado em Mai 17].

Drummond, C.; Holte, R. C. (2003). C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling.

Druzdzel, M. J.; Glymour, C. (1994). Application of the TETRAD II program to the study of student retention in U.S. colleges. http://www.pitt.edu/~druzdzel/abstracts/kdd94.html, [acessado em Nov 12].

Duman, E.; Ekinci, Y.; Tanrıverdi, A. (jan 2012). Comparing alternative classifiers for database marketing: The case of imbalanced datasets. *Expert Systems with Applications*, v. 39, n. 1, p. 48–53.

Dwayne D. Gremler, S. W. B. (1999). The loyalty ripple effect: Appreciating the full value of customers. *International Journal of Service Industry Management*, v. 10, n. 3, p. 271–293.

Eckert, K. B.; Suénaga, R. (2015). Analysis of Attrition-Retention of College Students Using Classification Technique in Data Mining. *Formación universitaria*, v. 8, n. 5, p. 03-12.

Edelweiss, N. (2001). Bancos de Dados Temporais: Teoria e Prática. https://www.researchgate.net/publication/237299628_Bancos_de_Dados_Temporais_Teoria_e_Pratica, [acessado em Jul 18].

Educação Adventista (2016). http://www.educacaoadventista.org.br/conheca-mais/no-mundo/, [acessado em Jun 13].

Fabio Bergamo, A. C. G. (2011). Modelo de lealdade e retenção de alunos para instituições do ensino superior: um estudo teórico com base no marketing de relacionamento. *Brazilian Business Review*, v. 8, n. 2, p. 43–67.

Fayyad, U. M.; Gregory Piatetsky-Shapiro; Padhraic Smyth (1996). Advances in Knowledge Discovery and Data Mining. In: Usama M. Fayyad; Gregory Piatetsky-Shapiro; Padhraic Smyth; Ramasamy Uthurusamy[Eds.]. . Menlo Park, CA, USA: American Association for Artificial Intelligence. p. 1–34.

Fike, D. S.; Fike, R. (10 jan 2008). Predictors of First-Year Student Retention in the Community College. *Community College Review*, v. 36, n. 2, p. 68–88.

Gerhardt, T. E.; Silveira, D. T.; Neis, I. A.; Abreu, S. P. De; Rodrigues, R. S. (2009). *Métodos de pesquisa*. Ed. da UFRGS.

Goebel, M.; Gruenwald, L. (jun 1999). A Survey of Data Mining and Knowledge Discovery Software Tools. *SIGKDD Explor. Newsl.*, v. 1, n. 1, p. 20–33.

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

Hall, M.; Frank, E.; Holmes, G.; et al. (nov 2009). The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.*, v. 11, n. 1, p. 10–18.

Hämäläinen, W.; Vinni, M. (2011). *Classifiers for educational data mining*.

Han, J.; Kamber, M.; Pei, J. (2011). *Data Mining: Concepts and Techniques, Third Edition*. 3 edition ed. Burlington, MA: Morgan Kaufmann.

IBGE (2010). Pirâmide etária. http://vamoscontar.ibge.gov.br/atividades/ensino-fundamental-6-ao-9/49-piramide-etaria.html, [acessado em Mai 22].

Inmon, W. H. (2005). *Building the Data Warehouse*. 4 edition ed. Indianapolis, Ind: Wiley.

Jain, A. K.; Mao, J.; Mohiuddin, K. M. (mar 1996). Artificial neural networks: a tutorial. *Computer*, v. 29, n. 3, p. 31–44.

Jo, T.; Japkowicz, N. (jun 2004). Class Imbalances Versus Small Disjuncts. *SIGKDD Explor. Newsl.*, v. 6, n. 1, p. 40–49.

Kohavi, R.; John, G. H. (dez 1997). Wrappers for feature subset selection. *Artificial Intelligence*, Relevance. v. 97, n. 1–2, p. 273–324.

Kotler, P.; Armstrong, G. (2012). *Princípios de marketing*. 12 edition ed. Pearson.

Lauría, E. J. M.; Baron, J. D.; Devireddy, M.; Sundararaju, V.; Jayaprakash, S. M. (2012). Mining Academic Data to Improve College Student Retention: An Open Source Perspective. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*. , LAK '12. ACM. http://doi.acm.org/10.1145/2330601.2330637, [acessado em Dez 3].

Li, D.-C.; Liu, C.-W.; Hu, S. C. (1 mai 2010). A learning method for the class imbalance problem with medical data sets. *Computers in Biology and Medicine*, v. 40, n. 5, p. 509–518.

Liao, T. W. (out 2008). Classification of weld flaws with imbalanced class data. *Expert Systems with Applications*, v. 35, n. 3, p. 1041–1052.

Lin, S.-H. (abr 2012). Data Mining for Student Retention Management. *J. Comput. Sci. Coll.*, v. 27, n. 4, p. 92–99.

Liu, H.; Hussain, F.; Tan, C. L.; Dash, M. (out 2002). Discretization: An Enabling Technique. *Data Mining and Knowledge Discovery*, v. 6, n. 4, p. 393–423.

Liu, H.; Motoda, H. (1998). *Feature Extraction, Construction and Selection: A Data Mining Perspective*. Norwell, MA, USA: Kluwer Academic Publishers.

Lykourentzou, I.; Giannoukos, I.; Nikolopoulos, V.; Mpardis, G.; Loumos, V. (nov 2009). Dropout prediction in e-learning courses through the combination of machine learning techniques. *Computers & Education*, v. 53, n. 3, p. 950–965.

Mannila, H.; Toivonen, H.; Verkamo, A. I. (1995). Discovering Frequent Episodes in Sequences Extended Abstract. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. , KDD'95. AAAI Press. http://dl.acm.org/citation.cfm?id=3001335.3001370, [acessado em Jul 18].

Marquez-Vera, C.; Morales, C. R.; Soto, S. V. (fev 2013). Predicting School Failure and Dropout by Using Data Mining Techniques. *Tecnologias del Aprendizaje, IEEE Revista Iberoamericana de*, v. 8, n. 1, p. 7–14.

Nandeshwar, A.; Menzies, T.; Nelson, A. (nov 2011). Learning patterns of university student retention. *Expert Systems with Applications*, v. 38, n. 12, p. 14984–14996.

NCHEMS (2010). HigherEdInfo.org: Retention Rates - First-Time College Freshmen Returning Their Second Year. http://www.higheredinfo.org/dbrowser/index.php?measure=92, [acessado em Nov 13].

Özden, B.; Ramaswamy, S.; Silberschatz, A. (1998). Cyclic Association Rules. In *Proceedings of the Fourteenth International Conference on Data Engineering*. , ICDE '98. IEEE Computer Society. http://dl.acm.org/citation.cfm?id=645483.656222, [acessado em Jul 18].

Parker, A. (1999). A Study of Variables that Predict Dropout from Distance Education. *International Journal of Educational Technology*, v. 1, n. 2, p. 1–10.

Pittman, K. (2008). Comparison of data mining techniques used to predict student retention. Nova Southeastern University.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Ramaswamy, S.; Mahajan, S.; Silberschatz, A. (1998). On the Discovery of Interesting Patterns in Association Rules. In *Proceedings of the 24rd International Conference on Very Large Data Bases*. , VLDB '98. Morgan Kaufmann Publishers Inc. http://dl.acm.org/citation.cfm?id=645924.671170, [acessado em Jul 18].

Roe, S. F. (2012). Accurately Measuring Model Prediction Error. http://scott.fortmann-roe.com/docs/MeasuringError.html, [acessado em Nov 22].

Romero, C.; Ventura, S. (jul 2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, v. 33, n. 1, p. 135–146.

Romero, C.; Ventura, S. (nov 2010). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, v. 40, n. 6, p. 601–618.

Romero, C.; Ventura, S.; Pechenizkiy, M. (2010). *Handbook of Educational Data Mining*. 1 edition ed. Boca Raton: CRC Press.

RuleQuest ([S.d.]). Information on See5/C5.0. https://www.rulequest.com/see5-info.html, [acessado em Nov 8].

Serra, A. P.; Zárate, L. E. (2015). Characterization of time series for analyzing of the evolution of time series clusters. *Expert Systems With Applications*, v. 42, n. 1, p. 596–611.

Srikant, R.; Agrawal, R. (25 mar 1996). Mining sequential patterns: Generalizations and performance improvements. In *Advances in Database Technology — EDBT '96*. , Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. https://link.springer.com/chapter/10.1007/BFb0014140, [acessado em Jul 18].

Su, C.-T.; Chen, L.-S.; Yih, Y. (out 2006). Knowledge acquisition through information granulation for imbalanced data. *Expert Systems with Applications*, v. 31, n. 3, p. 531–541.

Superby, J. F.; Vandamme, J. P. (2006). Determination of factors influencing the achievement of the first-year university students using data mining methods. http://libra.msra.cn/Publication/11731349/determination-of-factors-influencing-the-achievement-of-the-first-year-university-students-using, [acessado em Nov 13].

Tan, P.-N.; Steinbach, M.; Kumar, V. (2006). *Introduction to Data Mining*. 1. ed. Addison Wesley: .

Thammasiri, D.; Delen, D.; Meesad, P.; Kasap, N. (1 fev 2014). A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, v. 41, n. 2, p. 321–330.

Tillman, C.; Burns, P. (2000). Presentation on First Year Experience. . http://www.valdosta.edu/cgtillma/powerpoint.ppt, [acessado em Nov 13].

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, v. 45, n. 1, p. 89–125.

Tinto, V. (1982). The Limits of Theory and Practice in Student Attrition. *The Journal of Higher Education*, v. 53, n. 6.

Tinto, V.; Bean, J. P. (1988). Review of Leaving College: Rethinking the Causes and Cures of Student Attrition. *The Journal of Higher Education*, v. 59, n. 6, p. 708–711.

Wang, J. (2005). *Encyclopedia of Data Warehousing and Mining*. Pck edition ed. Idea Group Publishing.

WEKA (2017). Weka 3 - Data Mining with Open Source Machine Learning Software in Java. http://www.cs.waikato.ac.nz/ml/weka/, [acessado em Fev 5].

Witten, I. H.; Frank, E.; Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques, Third Edition*. 3 edition ed. Morgan Kaufmann.

Yang, Y. (1994). Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. , SIGIR '94. Springer-Verlag New York, Inc. http://dl.acm.org/citation.cfm?id=188490.188496, [acessado em Nov 8].

Yu, C. H.; DiGangi, S. A.; Jannasch-Pennell, A.; Lo, W.; Kaprolet, C. (2007). *A Data-Mining Approach to Differentiate Predictors of Retention*.

Yu, H.; Kim, S. (2012). SVM Tutorial — Classification, Regression and Ranking. In: Rozenberg, G.; Bäck, T.; Kok, J. N.[Eds.]. . *Handbook of Natural Computing*. Springer Berlin Heidelberg. p. 479–506.

Zhang, H. (2004). The optimality of naive Bayes. *AA*, v. 1, n. 2, p. 3.

Apêndice A

Árvore de Decisão do Primeiro Experimento

```
FaltasBiologia <= 3.975091
|   jaReprovou = 0
|   |   OcorrenciasRuins <= 0.983655
|   |   |   AlunoQuantoTempo <= 0
|   |   |   |   Curso = Fund-1º-Ano
|   |   |   |   |   TemIrmao = 0
|   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   |   Ano = 2000
|   |   |   |   |   |   |   |   MediaMatematica = 0: Não (32.0/3.0)
|   |   |   |   |   |   |   |   MediaMatematica = 1: Sim (3.0)
|   |   |   |   |   |   |   |   MediaMatematica = 2: Não (0.0)
|   |   |   |   |   |   |   |   MediaMatematica = 3: Não (0.0)
|   |   |   |   |   |   |   |   MediaMatematica = 4: Não (0.0)
|   |   |   |   |   |   |   |   MediaMatematica = 5: Não (0.0)
|   |   |   |   |   |   |   |   MediaMatematica = 6: Não (0.0)
|   |   |   |   |   |   |   Ano = 2001
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.813304: Não (24.0/4.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.813304: Sim (8.0/1.0)
|   |   |   |   |   |   |   Ano = 2002
|   |   |   |   |   |   |   |   MediaNaIdade = Igual: Não (7.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Menor: Não (2.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Maior: Sim (6.0)
|   |   |   |   |   |   |   Ano = 2003
|   |   |   |   |   |   |   |   MediaNaIdade = Igual
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Sim (3.0/1.0)
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053: Não (10.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Menor: Não (0.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Maior: Sim (7.0)
|   |   |   |   |   |   |   Ano = 2004
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.033909: Sim (12.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.033909: Não (33.0/4.0)
|   |   |   |   |   |   |   Ano = 2005
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0: Sim (10.0/3.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 0: Não (25.0/4.0)
|   |   |   |   |   |   |   Ano = 2006
|   |   |   |   |   |   |   |   MediaNaIdade = Igual: Não (28.0/3.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Menor: Não (0.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Maior
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 26.47682: Sim (7.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 26.47682: Não (4.0)
|   |   |   |   |   |   |   Ano = 2007
|   |   |   |   |   |   |   |   EstadoCivil = Solteiro: Sim (4.0)
|   |   |   |   |   |   |   |   EstadoCivil = Casado
|   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75: Sim (1.0)
|   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 76-a-100: Não (0.0)
|   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5: Não (13.0/2.0)
|   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50
|   |   |   |   |   |   |   |   |   |   Genero = M: Não (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   Genero = F: Sim (2.0)
|   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25
|   |   |   |   |   |   |   |   |   |   Turno = M: Não (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   Turno = T: Sim (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   Turno = N: Não (0.0)
|   |   |   |   |   |   |   |   EstadoCivil = Separado: Não (3.0/1.0)
|   |   |   |   |   |   |   Ano = 2008
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação: Não
(49.0/7.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio
|   |   |   |   |   |   |   |   |   CobrouMatricula = 0: Não (3.0/1.0)
|   |   |   |   |   |   |   |   |   CobrouMatricula = 1: Sim (2.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo
|   |   |   |   |   |   |   |   |   EhAdventista = Sim: Sim (2.0)
|   |   |   |   |   |   |   |   |   EhAdventista = Não
|   |   |   |   |   |   |   |   |   |   Genero = M: Não (11.0/1.0)
|   |   |   |   |   |   |   |   |   |   Genero = F: Sim (5.0/2.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Sim
(5.0/1.0)
```

```
|   |   |   |   |   |   |   |   Ano = 2009
|   |   |   |   |   |   |   |   |   MediaNaIdade = Igual: Não (46.0/8.0)
|   |   |   |   |   |   |   |   |   MediaNaIdade = Menor: Não (0.0)
|   |   |   |   |   |   |   |   |   MediaNaIdade = Maior: Sim (4.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2010: Não (32.0/2.0)
|   |   |   |   |   |   |   |   Ano = 2011
|   |   |   |   |   |   |   |   |   VeioDeForaDoEstado = Sim: Não (3.0/1.0)
|   |   |   |   |   |   |   |   |   VeioDeForaDoEstado = Não: Sim (27.0/2.0)
|   |   |   |   |   |   |   |   Ano = 2012: Sim (45.0/12.0)
|   |   |   |   |   |   |   |   Ano = 2013: Sim (36.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2014: Sim (38.0/4.0)
|   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (56.0/13.0)
|   |   |   |   |   |   |   TemIrmao = 1: Sim (146.0/28.0)
|   |   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   |   QtdeAlunos <= 4
|   |   |   |   |   |   |   |   QtdeAlunos <= 1.435747: Sim (2.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 1.435747: Não (7.0/1.0)
|   |   |   |   |   |   |   QtdeAlunos > 4: Não (156.0/2.0)
|   |   |   |   |   |   Curso = Fund-3º-Ano: Não (69.0/5.0)
|   |   |   |   |   |   Curso = Fund-4º-Ano: Não (84.0/6.0)
|   |   |   |   |   |   Curso = Fund-5º-Ano
|   |   |   |   |   |   |   TemIrmao = 0: Não (51.0/9.0)
|   |   |   |   |   |   |   TemIrmao = 1
|   |   |   |   |   |   |   |   Ano = 2000: Não (1.0)
|   |   |   |   |   |   |   |   Ano = 2001: Não (1.0)
|   |   |   |   |   |   |   |   Ano = 2002: Sim (1.0)
|   |   |   |   |   |   |   |   Ano = 2003: Sim (1.0)
|   |   |   |   |   |   |   |   Ano = 2004: Sim (5.0)
|   |   |   |   |   |   |   |   Ano = 2005: Sim (2.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2006: Não (1.0)
|   |   |   |   |   |   |   |   Ano = 2007: Não (3.0)
|   |   |   |   |   |   |   |   Ano = 2008: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2009: Sim (1.0)
|   |   |   |   |   |   |   |   Ano = 2010: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2011: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2012: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2013: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2014: Sim (0.0)
|   |   |   |   |   |   Curso = Fund-6º-Ano
|   |   |   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   |   |   Ano = 2000: Não (4.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2001: Sim (7.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2002: Sim (7.0)
|   |   |   |   |   |   |   |   Ano = 2003: Sim (13.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2004: Sim (7.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2005: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2006: Não (4.0)
|   |   |   |   |   |   |   |   Ano = 2007: Não (2.0)
|   |   |   |   |   |   |   |   Ano = 2008: Não (3.0)
|   |   |   |   |   |   |   |   Ano = 2009: Sim (1.0)
|   |   |   |   |   |   |   |   Ano = 2010: Não (1.0)
|   |   |   |   |   |   |   |   Ano = 2011: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2012: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2013: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2014: Sim (0.0)
|   |   |   |   |   |   |   CobrouMatricula = 1: Não (81.0/5.0)
|   |   |   |   |   |   Curso = Fund-7º-Ano
|   |   |   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.893844: Não (13.0/1.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.893844
|   |   |   |   |   |   |   |   |   EstadoCivil = Solteiro: Sim (0.0)
|   |   |   |   |   |   |   |   |   EstadoCivil = Casado
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 34.496416: Não (3.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 34.496416: Sim (8.0/2.0)
|   |   |   |   |   |   |   |   |   EstadoCivil = Separado: Sim (4.0)
|   |   |   |   |   |   |   CobrouMatricula = 1: Não (21.0/1.0)
|   |   |   |   |   |   Curso = Fund-8º-Ano
|   |   |   |   |   |   |   QtdeAlunos <= 35.490684: Não (21.0)
|   |   |   |   |   |   |   QtdeAlunos > 35.490684
|   |   |   |   |   |   |   |   CobrouMatricula = 0: Sim (7.0)
|   |   |   |   |   |   |   |   CobrouMatricula = 1
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Sim (5.0/1.0)
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053: Não (14.0/2.0)
|   |   |   |   AlunoQuantoTempo > 0
|   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75
|   |   |   |   |   |   Curso = Fund-1º-Ano
|   |   |   |   |   |   |   Turno = M
|   |   |   |   |   |   |   |   FaixaEtaria = <25: Não (2.0)
```

```
|   |   |   |   |   |   |   FaixaEtaria = >=71: Não (0.0)
|   |   |   |   |   |   |   FaixaEtaria = 36-40: Não (6.0/1.0)
|   |   |   |   |   |   |   FaixaEtaria = 51-70: Sim (1.0)
|   |   |   |   |   |   |   FaixaEtaria = 41-50: Sim (4.0)
|   |   |   |   |   |   |   FaixaEtaria = 31-35
|   |   |   |   |   |   |   |   Genero = M: Não (2.0)
|   |   |   |   |   |   |   |   Genero = F: Sim (2.0)
|   |   |   |   |   |   |   FaixaEtaria = 26-30
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.496705: Sim (5.0/1.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.496705: Não (6.0)
|   |   |   |   |   |   Turno = T
|   |   |   |   |   |   |   EstadoCivil = Solteiro: Sim (6.0/1.0)
|   |   |   |   |   |   |   EstadoCivil = Casado: Sim (70.0/21.0)
|   |   |   |   |   |   |   EstadoCivil = Separado: Não (5.0/1.0)
|   |   |   |   |   |   Turno = N: Sim (0.0)
|   |   |   |   |   Curso = Fund-2º-Ano: Não (63.0/3.0)
|   |   |   |   |   Curso = Fund-3º-Ano
|   |   |   |   |   |   QtdeAlunos <= 11: Sim (2.0)
|   |   |   |   |   |   QtdeAlunos > 11: Não (41.0/1.0)
|   |   |   |   |   Curso = Fund-4º-Ano: Não (38.0/9.0)
|   |   |   |   |   Curso = Fund-5º-Ano: Não (52.0/15.0)
|   |   |   |   |   Curso = Fund-6º-Ano
|   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não (13.0)
|   |   |   |   |   |   Dependentes_Matriculados > 0.493053
|   |   |   |   |   |   |   CobrouMatricula = 0: Sim (11.0/1.0)
|   |   |   |   |   |   |   CobrouMatricula = 1
|   |   |   |   |   |   |   |   QtdeAlunos <= 30
|   |   |   |   |   |   |   |   |   MediaNaIdade = Igual: Sim (4.0/1.0)
|   |   |   |   |   |   |   |   |   MediaNaIdade = Menor: Sim (0.0)
|   |   |   |   |   |   |   |   |   MediaNaIdade = Maior: Não (2.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 30: Não (18.0)
|   |   |   |   |   Curso = Fund-7º-Ano
|   |   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não (3.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053
|   |   |   |   |   |   |   |   AlunoQuantoTempo <= 5.499975: Sim (15.0)
|   |   |   |   |   |   |   |   AlunoQuantoTempo > 5.499975: Não (3.0)
|   |   |   |   |   |   CobrouMatricula = 1: Não (17.0/2.0)
|   |   |   |   |   Curso = Fund-8º-Ano
|   |   |   |   |   |   Ano = 2000: Sim (3.0)
|   |   |   |   |   |   Ano = 2001
|   |   |   |   |   |   |   AlunoQuantoTempo <= 2.497997: Sim (2.0)
|   |   |   |   |   |   |   AlunoQuantoTempo > 2.497997: Não (3.0)
|   |   |   |   |   |   Ano = 2002
|   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não (6.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053: Sim (3.0)
|   |   |   |   |   |   Ano = 2003: Sim (5.0)
|   |   |   |   |   |   Ano = 2004
|   |   |   |   |   |   |   AlunoQuantoTempo <= 4: Não (6.0/2.0)
|   |   |   |   |   |   |   AlunoQuantoTempo > 4: Sim (3.0)
|   |   |   |   |   |   Ano = 2005: Não (5.0)
|   |   |   |   |   |   Ano = 2006: Não (3.0)
|   |   |   |   |   |   Ano = 2007: Não (4.0/1.0)
|   |   |   |   |   |   Ano = 2008: Não (2.0)
|   |   |   |   |   |   Ano = 2009: Não (0.0)
|   |   |   |   |   |   Ano = 2010: Não (0.0)
|   |   |   |   |   |   Ano = 2011: Não (1.0)
|   |   |   |   |   |   Ano = 2012: Não (0.0)
|   |   |   |   |   |   Ano = 2013: Não (1.0)
|   |   |   |   |   |   Ano = 2014: Não (0.0)
|   |   |   |   BolsaEDescontosNormalizado = 76-a-100
|   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   Curso = Fund-1º-Ano: Sim (4.0/1.0)
|   |   |   |   |   |   Curso = Fund-2º-Ano: Não (11.0)
|   |   |   |   |   |   Curso = Fund-3º-Ano: Não (7.0/1.0)
|   |   |   |   |   |   Curso = Fund-4º-Ano: Não (3.0)
|   |   |   |   |   |   Curso = Fund-5º-Ano: Não (7.0/1.0)
|   |   |   |   |   |   Curso = Fund-6º-Ano: Sim (7.0/1.0)
|   |   |   |   |   |   Curso = Fund-7º-Ano: Não (10.0/4.0)
|   |   |   |   |   |   Curso = Fund-8º-Ano
|   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não (6.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053: Sim (8.0)
|   |   |   |   |   CobrouMatricula = 1: Não (95.0/9.0)
|   |   |   |   BolsaEDescontosNormalizado = 0-a-5
|   |   |   |   |   Curso = Fund-1º-Ano
|   |   |   |   |   |   Ano = 2000: Não (6.0/1.0)
|   |   |   |   |   |   Ano = 2001
|   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não (26.0/1.0)
```

```
|   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053
|   |   |   |   |   |   |   |   Turno = M: Sim (4.0)
|   |   |   |   |   |   |   |   Turno = T: Não (2.0)
|   |   |   |   |   |   |   |   Turno = N: Sim (0.0)
|   |   |   |   |   |   Ano = 2002
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação: Não (16.0/1.0)
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Não (2.0)
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo
|   |   |   |   |   |   |   |   QtdeAlunos <= 27: Sim (4.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 27: Não (2.0)
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Não (0.0)
|   |   |   |   |   |   Ano = 2003: Não (28.0/4.0)
|   |   |   |   |   |   Ano = 2004
|   |   |   |   |   |   |   Dependentes_Matriculados <= 0.307848: Sim (4.0/1.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 0.307848: Não (59.0/3.0)
|   |   |   |   |   |   Ano = 2005: Não (55.0/4.0)
|   |   |   |   |   |   Ano = 2006
|   |   |   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação
|   |   |   |   |   |   |   |   |   FaixaEtaria = <25: Sim (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = >=71: Sim (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 36-40: Sim (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 51-70: Sim (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 41-50: Sim (2.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 31-35: Não (2.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 26-30: Sim (0.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Não (2.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo: Sim (2.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Sim (0.0)
|   |   |   |   |   |   |   CobrouMatricula = 1: Não (26.0/1.0)
|   |   |   |   |   |   Ano = 2007
|   |   |   |   |   |   |   CobrouMatricula = 0: Não (7.0)
|   |   |   |   |   |   |   CobrouMatricula = 1
|   |   |   |   |   |   |   |   QtdeAlunos <= 19.463006: Sim (7.0/1.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 19.463006: Não (15.0/3.0)
|   |   |   |   |   |   Ano = 2008
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação: Não (13.0/1.0)
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Não (3.0)
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo
|   |   |   |   |   |   |   |   Turno = M: Sim (3.0)
|   |   |   |   |   |   |   |   Turno = T
|   |   |   |   |   |   |   |   |   EstadoCivil = Solteiro: Sim (4.0/1.0)
|   |   |   |   |   |   |   |   |   EstadoCivil = Casado: Não (8.0/2.0)
|   |   |   |   |   |   |   |   |   EstadoCivil = Separado: Não (0.0)
|   |   |   |   |   |   |   |   Turno = N: Sim (0.0)
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Não (1.0)
|   |   |   |   |   |   Ano = 2009: Não (20.0)
|   |   |   |   |   |   Ano = 2010: Não (41.0/3.0)
|   |   |   |   |   |   Ano = 2011
|   |   |   |   |   |   |   Dependentes_Matriculados <= 1.54644
|   |   |   |   |   |   |   |   Turno = M: Não (2.0)
|   |   |   |   |   |   |   |   Turno = T
|   |   |   |   |   |   |   |   |   FaixaEtaria = <25: Não (3.0/1.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = >=71: Não (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 36-40: Não (1.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 51-70: Sim (1.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 41-50: Não (4.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 31-35
|   |   |   |   |   |   |   |   |   |   Genero = M: Não (2.0)
|   |   |   |   |   |   |   |   |   |   Genero = F: Sim (4.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 26-30: Sim (2.0)
|   |   |   |   |   |   |   |   Turno = N: Não (0.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 1.54644: Sim (8.0)
|   |   |   |   |   |   Ano = 2012
|   |   |   |   |   |   |   MediaNaIdade = Igual
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação: Não
(8.0/1.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Sim
(2.0/1.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo: Sim
(8.0/2.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Não (2.0)
|   |   |   |   |   |   |   MediaNaIdade = Menor: Sim (0.0)
|   |   |   |   |   |   |   MediaNaIdade = Maior: Sim (7.0/1.0)
|   |   |   |   |   |   Ano = 2013
|   |   |   |   |   |   |   QtdeAlunos <= 9.275872: Não (7.0)
|   |   |   |   |   |   |   QtdeAlunos > 9.275872: Sim (35.0/9.0)
|   |   |   |   |   |   Ano = 2014: Sim (26.0/2.0)
```

```
|   |   |   |   |   Curso = Fund-2º-Ano: Não (589.0)
|   |   |   |   |   Curso = Fund-3º-Ano: Não (623.0/1.0)
|   |   |   |   |   Curso = Fund-4º-Ano: Não (599.0/9.0)
|   |   |   |   |   Curso = Fund-5º-Ano
|   |   |   |   |   |   Dependentes_Matriculados <= 0.96179: Não (145.0)
|   |   |   |   |   |   Dependentes_Matriculados > 0.96179
|   |   |   |   |   |   |   Ano = 2000: Sim (10.0)
|   |   |   |   |   |   |   Ano = 2001
|   |   |   |   |   |   |   |   QtdeAlunos <= 19.463006: Não (4.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 19.463006: Sim (10.0/1.0)
|   |   |   |   |   |   |   Ano = 2002
|   |   |   |   |   |   |   |   QtdeAlunos <= 34: Não (4.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 34: Sim (7.0)
|   |   |   |   |   |   |   Ano = 2003
|   |   |   |   |   |   |   |   QtdeAlunos <= 24.772817: Não (14.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 24.772817: Sim (5.0/1.0)
|   |   |   |   |   |   |   Ano = 2004: Não (50.0/1.0)
|   |   |   |   |   |   |   Ano = 2005
|   |   |   |   |   |   |   |   QtdeAlunos <= 28.48277: Não (37.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 28.48277
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 31: Sim (3.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 31: Não (14.0/2.0)
|   |   |   |   |   |   |   Ano = 2006: Não (45.0)
|   |   |   |   |   |   |   Ano = 2007: Não (30.0/1.0)
|   |   |   |   |   |   |   Ano = 2008: Não (15.0)
|   |   |   |   |   |   |   Ano = 2009: Não (20.0)
|   |   |   |   |   |   |   Ano = 2010: Não (32.0)
|   |   |   |   |   |   |   Ano = 2011: Não (12.0)
|   |   |   |   |   |   |   Ano = 2012: Não (14.0)
|   |   |   |   |   |   |   Ano = 2013: Não (11.0)
|   |   |   |   |   |   |   Ano = 2014: Não (2.0)
|   |   |   |   |   Curso = Fund-6º-Ano: Não (598.0/18.0)
|   |   |   |   |   Curso = Fund-7º-Ano: Não (617.0/19.0)
|   |   |   |   |   Curso = Fund-8º-Ano: Não (710.0/24.0)
|   |   |   |   BolsaEDescontosNormalizado = 26-a-50
|   |   |   |   |   Curso = Fund-1º-Ano
|   |   |   |   |   |   TemIrmao = 0
|   |   |   |   |   |   |   Ano = 2000: Não (6.0/1.0)
|   |   |   |   |   |   |   Ano = 2001: Sim (4.0/1.0)
|   |   |   |   |   |   |   Ano = 2002: Não (3.0/1.0)
|   |   |   |   |   |   |   Ano = 2003: Sim (5.0/1.0)
|   |   |   |   |   |   |   Ano = 2004
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Sim (4.0/1.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053: Não (11.0)
|   |   |   |   |   |   |   Ano = 2005
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação
|   |   |   |   |   |   |   |   |   FaixaEtaria = <25: Sim (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = >=71: Sim (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 36-40: Não (2.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 51-70: Sim (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 41-50: Sim (1.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 31-35: Sim (3.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 26-30: Não (2.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Não (0.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo: Sim (2.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Não (3.0)
|   |   |   |   |   |   |   Ano = 2006: Não (14.0/3.0)
|   |   |   |   |   |   |   Ano = 2007
|   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 15.250064
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.103474: Sim (3.0)
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.103474: Não (3.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 15.250064: Não (21.0)
|   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (3.0/1.0)
|   |   |   |   |   |   |   Ano = 2008: Não (13.0/3.0)
|   |   |   |   |   |   |   Ano = 2009
|   |   |   |   |   |   |   |   MediaNaIdade = Igual
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Sim (5.0/1.0)
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053: Não (8.0/1.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Menor: Não (0.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Maior: Sim (2.0)
|   |   |   |   |   |   |   Ano = 2010: Não (8.0/1.0)
|   |   |   |   |   |   |   Ano = 2011: Sim (7.0/2.0)
|   |   |   |   |   |   |   Ano = 2012: Sim (8.0/1.0)
|   |   |   |   |   |   |   Ano = 2013: Sim (7.0)
|   |   |   |   |   |   |   Ano = 2014: Sim (5.0)
|   |   |   |   |   |   TemIrmao = 1
|   |   |   |   |   |   |   VeioDeForaDoEstado = Sim: Não (6.0/2.0)
```

```
|   |   |   |   |   |   |   VeioDeForaDoEstado = Não: Sim (60.0/15.0)
|   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   QtdeAlunos <= 4.778116
|   |   |   |   |   |   |   QtdeAlunos <= 2.399861: Sim (4.0)
|   |   |   |   |   |   |   QtdeAlunos > 2.399861: Não (2.0)
|   |   |   |   |   |   QtdeAlunos > 4.778116: Não (110.0/2.0)
|   |   |   |   |   Curso = Fund-3º-Ano
|   |   |   |   |   |   QtdeAlunos <= 7.370504: Sim (5.0/1.0)
|   |   |   |   |   |   QtdeAlunos > 7.370504: Não (129.0/2.0)
|   |   |   |   |   Curso = Fund-4º-Ano: Não (94.0/8.0)
|   |   |   |   |   Curso = Fund-5º-Ano: Não (99.0/22.0)
|   |   |   |   |   Curso = Fund-6º-Ano
|   |   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não (7.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053
|   |   |   |   |   |   |   |   Ano = 2000: Sim (1.0)
|   |   |   |   |   |   |   |   Ano = 2001: Sim (5.0)
|   |   |   |   |   |   |   |   Ano = 2002: Sim (2.0)
|   |   |   |   |   |   |   |   Ano = 2003
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 42: Não (4.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 42: Sim (7.0)
|   |   |   |   |   |   |   |   Ano = 2004: Sim (3.0)
|   |   |   |   |   |   |   |   Ano = 2005: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2006: Não (6.0)
|   |   |   |   |   |   |   |   Ano = 2007: Não (1.0)
|   |   |   |   |   |   |   |   Ano = 2008: Não (2.0)
|   |   |   |   |   |   |   |   Ano = 2009: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2010: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2011: Não (2.0)
|   |   |   |   |   |   |   |   Ano = 2012: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2013: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2014: Sim (0.0)
|   |   |   |   |   |   CobrouMatricula = 1
|   |   |   |   |   |   |   QtdeAlunos <= 31.472983: Sim (7.0/3.0)
|   |   |   |   |   |   |   QtdeAlunos > 31.472983: Não (54.0)
|   |   |   |   |   Curso = Fund-7º-Ano
|   |   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não (9.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053
|   |   |   |   |   |   |   |   Ano = 2000: Sim (2.0)
|   |   |   |   |   |   |   |   Ano = 2001: Sim (4.0)
|   |   |   |   |   |   |   |   Ano = 2002: Sim (4.0)
|   |   |   |   |   |   |   |   Ano = 2003: Sim (5.0)
|   |   |   |   |   |   |   |   Ano = 2004
|   |   |   |   |   |   |   |   |   Genero = M: Não (5.0)
|   |   |   |   |   |   |   |   |   Genero = F: Sim (3.0)
|   |   |   |   |   |   |   |   Ano = 2005: Não (1.0)
|   |   |   |   |   |   |   |   Ano = 2006: Não (1.0)
|   |   |   |   |   |   |   |   Ano = 2007: Não (2.0)
|   |   |   |   |   |   |   |   Ano = 2008: Não (2.0)
|   |   |   |   |   |   |   |   Ano = 2009: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2010: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2011: Não (1.0)
|   |   |   |   |   |   |   |   Ano = 2012: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2013: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2014: Sim (0.0)
|   |   |   |   |   |   CobrouMatricula = 1
|   |   |   |   |   |   |   Turno = M: Não (43.0/1.0)
|   |   |   |   |   |   |   Turno = T
|   |   |   |   |   |   |   |   EhAdventista = Sim: Sim (3.0)
|   |   |   |   |   |   |   |   EhAdventista = Não: Não (21.0/4.0)
|   |   |   |   |   |   |   Turno = N: Não (0.0)
|   |   |   |   |   Curso = Fund-8º-Ano: Não (89.0/16.0)
|   |   |   |   BolsaEDescontosNormalizado = 6-a-25
|   |   |   |   |   MediaMatematica = 0
|   |   |   |   |   |   Curso = Fund-1º-Ano
|   |   |   |   |   |   |   AlunoQuantoTempo <= 0.978623: Não (27.0)
|   |   |   |   |   |   |   AlunoQuantoTempo > 0.978623
|   |   |   |   |   |   |   |   Ano = 2000: Não (7.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2001
|   |   |   |   |   |   |   |   |   TemIrmao = 0
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.558958: Não (3.0)
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.558958: Sim (2.0)
|   |   |   |   |   |   |   |   |   TemIrmao = 1: Sim (5.0)
|   |   |   |   |   |   |   |   Ano = 2002: Sim (8.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2003: Sim (13.0/2.0)
|   |   |   |   |   |   |   |   Ano = 2004
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.322116
```

```
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 29: Sim (12.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 29
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Sim
(4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053
|   |   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 1.498619: Não (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 1.498619: Sim
(3.0/1.0)
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.322116: Não (6.0)
|   |   |   |   |   |   |   |   Ano = 2005
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.444793
|   |   |   |   |   |   |   |   |   |   Genero = M: Não (2.0)
|   |   |   |   |   |   |   |   |   |   Genero = F: Sim (4.0)
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.444793
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 21.49637
|   |   |   |   |   |   |   |   |   |   |   EhAdventista = Sim: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   EhAdventista = Não
|   |   |   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 18.481811: Não (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   QtdeAlunos > 18.481811: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 21.49637: Não (53.0/2.0)
|   |   |   |   |   |   |   |   Ano = 2006
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 28.037931
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.167175: Sim (5.0)
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.167175
|   |   |   |   |   |   |   |   |   |   |   Genero = M
|   |   |   |   |   |   |   |   |   |   |   |   TemIrmao = 0: Não (16.0/5.0)
|   |   |   |   |   |   |   |   |   |   |   |   TemIrmao = 1: Sim (8.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   Genero = F: Sim (15.0/5.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 28.037931: Não (14.0)
|   |   |   |   |   |   |   |   Ano = 2007
|   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação: Sim
(10.0)
|   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Não
(4.0)
|   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 20.496219: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 20.496219: Não (3.0)
|   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Não
(1.0)
|   |   |   |   |   |   |   |   Ano = 2008
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 1.901794: Não (23.0/5.0)
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 1.901794
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.496705: Sim
(16.0/3.0)
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.496705
|   |   |   |   |   |   |   |   |   |   |   Genero = M: Sim (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   Genero = F: Não (2.0)
|   |   |   |   |   |   |   |   Ano = 2009
|   |   |   |   |   |   |   |   |   FaixaEtaria = <25: Sim (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = >=71: Sim (0.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 36-40: Não (6.0/2.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 51-70: Não (1.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 41-50: Sim (4.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 31-35
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não (2.0)
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053: Sim (2.0)
|   |   |   |   |   |   |   |   |   FaixaEtaria = 26-30: Sim (2.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2010: Não (38.0/9.0)
|   |   |   |   |   |   |   |   Ano = 2011: Sim (32.0/3.0)
|   |   |   |   |   |   |   |   Ano = 2012
|   |   |   |   |   |   |   |   |   TemIrmao = 0
|   |   |   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 31.872911: Não (22.0/7.0)
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos > 31.872911: Sim (4.0)
|   |   |   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (2.0)
|   |   |   |   |   |   |   |   |   TemIrmao = 1: Sim (12.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2013: Sim (37.0/2.0)
|   |   |   |   |   |   |   |   Ano = 2014: Sim (33.0/2.0)
|   |   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   |   QtdeAlunos <= 2
|   |   |   |   |   |   |   |   Ano = 2000: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2001: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2002: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2003: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2004: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2005: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2006: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2007: Sim (0.0)
```

```
|   |   |   |   |   |   |   |   Ano = 2008: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2009
|   |   |   |   |   |   |   |   |   CobrouMatricula = 0: Não (3.0/1.0)
|   |   |   |   |   |   |   |   |   CobrouMatricula = 1: Sim (8.0)
|   |   |   |   |   |   |   |   Ano = 2010: Não (3.0)
|   |   |   |   |   |   |   |   Ano = 2011: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2012: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2013: Sim (0.0)
|   |   |   |   |   |   |   |   Ano = 2014: Sim (0.0)
|   |   |   |   |   |   |   QtdeAlunos > 2
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.042435
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 27.483253: Não (16.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 27.483253
|   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação:
Sim (5.0)
|   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Sim
(0.0)
|   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo: Sim
(2.0)
|   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto:
Não (2.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.042435: Não (130.0/1.0)
|   |   |   |   |   |   Curso = Fund-3º-Ano: Não (142.0/5.0)
|   |   |   |   |   |   Curso = Fund-4º-Ano: Não (115.0/22.0)
|   |   |   |   |   |   Curso = Fund-5º-Ano
|   |   |   |   |   |   |   QtdeAlunos <= 24.442421: Não (36.0)
|   |   |   |   |   |   |   QtdeAlunos > 24.442421
|   |   |   |   |   |   |   |   Ano = 2000: Sim (3.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2001
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.826575: Não (4.0)
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.826575: Sim (3.0)
|   |   |   |   |   |   |   |   Ano = 2002: Sim (8.0)
|   |   |   |   |   |   |   |   Ano = 2003
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 37: Sim (7.0/1.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 37: Não (2.0)
|   |   |   |   |   |   |   |   Ano = 2004
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 29: Sim (5.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 29
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 40.46632: Não (7.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 40.46632
|   |   |   |   |   |   |   |   |   |   |   Genero = M: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   Genero = F
|   |   |   |   |   |   |   |   |   |   |   |   TemIrmao = 0: Não (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   TemIrmao = 1: Sim (2.0)
|   |   |   |   |   |   |   |   Ano = 2005
|   |   |   |   |   |   |   |   |   EhAdventista = Sim: Sim (4.0/1.0)
|   |   |   |   |   |   |   |   |   EhAdventista = Não
|   |   |   |   |   |   |   |   |   |   TemIrmao = 0: Não (23.0/1.0)
|   |   |   |   |   |   |   |   |   |   TemIrmao = 1
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.496705: Não
(3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.496705: Sim
(2.0)
|   |   |   |   |   |   |   |   Ano = 2006: Não (2.0)
|   |   |   |   |   |   |   |   Ano = 2007
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 3.491928: Não (4.0)
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 3.491928: Sim (2.0)
|   |   |   |   |   |   |   |   Ano = 2008: Não (8.0/1.0)
|   |   |   |   |   |   |   |   Ano = 2009: Não (12.0)
|   |   |   |   |   |   |   |   Ano = 2010: Não (5.0)
|   |   |   |   |   |   |   |   Ano = 2011: Não (6.0)
|   |   |   |   |   |   |   |   Ano = 2012: Não (0.0)
|   |   |   |   |   |   |   |   Ano = 2013: Não (4.0)
|   |   |   |   |   |   |   |   Ano = 2014: Não (1.0)
|   |   |   |   |   |   Curso = Fund-6º-Ano
|   |   |   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não (9.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 2.497997: Sim (12.0)
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 2.497997
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 43
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-
Graduação: Não (9.0)
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio:
Não (2.0)
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo
|   |   |   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 32: Não (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   QtdeAlunos > 32: Sim (4.0)
```

```
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-
Incompleto: Não (0.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 43: Sim (5.0)
|   |   |   |   |   |   |   CobrouMatricula = 1
|   |   |   |   |   |   |   |   QtdeAlunos <= 31.453894
|   |   |   |   |   |   |   |   |   Turno = M
|   |   |   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim: Não
(25.0/3.0)
|   |   |   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (3.0/1.0)
|   |   |   |   |   |   |   |   |   Turno = T: Sim (7.0/2.0)
|   |   |   |   |   |   |   |   |   Turno = N: Não (0.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 31.453894: Não (54.0/1.0)
|   |   |   |   |   |   Curso = Fund-7º-Ano
|   |   |   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   |   |   QtdeAlunos <= 30.496032: Não (16.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 30.496032
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.676323: Não (5.0)
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.676323: Sim (21.0/2.0)
|   |   |   |   |   |   |   CobrouMatricula = 1: Não (75.0/6.0)
|   |   |   |   |   |   Curso = Fund-8º-Ano
|   |   |   |   |   |   |   CobrouMatricula = 0
|   |   |   |   |   |   |   |   Turno = M
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 33.494815: Não (9.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 33.494815
|   |   |   |   |   |   |   |   |   |   MediaNaIdade = Igual
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Não
(5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.493053: Sim
(19.0/3.0)
|   |   |   |   |   |   |   |   |   |   MediaNaIdade = Menor: Não (3.0)
|   |   |   |   |   |   |   |   |   |   MediaNaIdade = Maior: Sim (7.0)
|   |   |   |   |   |   |   |   Turno = T: Não (6.0)
|   |   |   |   |   |   |   |   Turno = N: Não (0.0)
|   |   |   |   |   |   |   CobrouMatricula = 1: Não (59.0/7.0)
|   |   |   |   |   MediaMatematica = 1: Não (106.0)
|   |   |   |   |   MediaMatematica = 2: Não (2.0)
|   |   |   |   |   MediaMatematica = 3: Não (2.0)
|   |   |   |   |   MediaMatematica = 4: Não (0.0)
|   |   |   |   |   MediaMatematica = 5: Não (0.0)
|   |   |   |   |   MediaMatematica = 6: Não (0.0)
|   |   OcorrenciasRuins > 0.983655
|   |   |   Dependentes_Matriculados <= 0.020812
|   |   |   |   FaltasMatematica <= 0.452528: Sim (124.0/11.0)
|   |   |   |   FaltasMatematica > 0.452528
|   |   |   |   |   MediaMatematica = 0: Não (9.0/1.0)
|   |   |   |   |   MediaMatematica = 1: Sim (5.0)
|   |   |   |   |   MediaMatematica = 2: Não (0.0)
|   |   |   |   |   MediaMatematica = 3: Não (0.0)
|   |   |   |   |   MediaMatematica = 4: Não (0.0)
|   |   |   |   |   MediaMatematica = 5: Não (0.0)
|   |   |   |   |   MediaMatematica = 6: Não (0.0)
|   |   |   Dependentes_Matriculados > 0.020812
|   |   |   |   QtdeAlunos <= 33
|   |   |   |   |   OcorrenciasMaterial <= 0.034403
|   |   |   |   |   |   Curso = Fund-1º-Ano
|   |   |   |   |   |   |   Turno = M
|   |   |   |   |   |   |   |   QtdeAlunos <= 20.563962: Não (3.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 20.563962: Sim (36.0/4.0)
|   |   |   |   |   |   |   Turno = T: Sim (48.0)
|   |   |   |   |   |   |   Turno = N: Sim (0.0)
|   |   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   |   MediaNaIdade = Igual: Não (21.0)
|   |   |   |   |   |   |   MediaNaIdade = Menor: Não (0.0)
|   |   |   |   |   |   |   MediaNaIdade = Maior: Sim (3.0/1.0)
|   |   |   |   |   |   Curso = Fund-3º-Ano: Não (12.0/1.0)
|   |   |   |   |   |   Curso = Fund-4º-Ano
|   |   |   |   |   |   |   AlunoQuantoTempo <= 3: Não (7.0)
|   |   |   |   |   |   |   AlunoQuantoTempo > 3: Sim (2.0)
|   |   |   |   |   |   Curso = Fund-5º-Ano: Não (6.0/1.0)
|   |   |   |   |   |   Curso = Fund-6º-Ano
|   |   |   |   |   |   |   FaltasMatematica <= 0.452528
|   |   |   |   |   |   |   |   AlunoQuantoTempo <= 0.84153
|   |   |   |   |   |   |   |   |   TemIrmao = 0: Não (7.0)
|   |   |   |   |   |   |   |   |   TemIrmao = 1: Sim (3.0/1.0)
|   |   |   |   |   |   |   |   AlunoQuantoTempo > 0.84153
|   |   |   |   |   |   |   |   |   OcorrenciasRuins <= 9: Sim (14.0)
|   |   |   |   |   |   |   |   |   OcorrenciasRuins > 9: Não (3.0/1.0)
|   |   |   |   |   |   |   FaltasMatematica > 0.452528: Não (5.0)
```

```
|   |   |   |   |   |   Curso = Fund-7º-Ano
|   |   |   |   |   |   |   QtdeAlunos <= 30.496032: Sim (21.0/5.0)
|   |   |   |   |   |   |   QtdeAlunos > 30.496032: Não (7.0)
|   |   |   |   |   |   Curso = Fund-8º-Ano
|   |   |   |   |   |   |   Ano = 2000: Sim (0.0)
|   |   |   |   |   |   |   Ano = 2001: Sim (0.0)
|   |   |   |   |   |   |   Ano = 2002: Sim (0.0)
|   |   |   |   |   |   |   Ano = 2003: Sim (0.0)
|   |   |   |   |   |   |   Ano = 2004: Sim (0.0)
|   |   |   |   |   |   |   Ano = 2005: Sim (11.0)
|   |   |   |   |   |   |   Ano = 2006: Sim (1.0)
|   |   |   |   |   |   |   Ano = 2007: Não (3.0/1.0)
|   |   |   |   |   |   |   Ano = 2008: Sim (0.0)
|   |   |   |   |   |   |   Ano = 2009: Não (2.0)
|   |   |   |   |   |   |   Ano = 2010: Não (4.0)
|   |   |   |   |   |   |   Ano = 2011: Sim (0.0)
|   |   |   |   |   |   |   Ano = 2012: Sim (0.0)
|   |   |   |   |   |   |   Ano = 2013: Sim (0.0)
|   |   |   |   |   |   |   Ano = 2014: Sim (0.0)
|   |   |   |   |   |   OcorrenciasMaterial > 0.034403: Não (55.0/7.0)
|   |   |   |   |   QtdeAlunos > 33
|   |   |   |   |   |   Ano = 2000: Sim (2.0/1.0)
|   |   |   |   |   |   Ano = 2001: Não (1.0)
|   |   |   |   |   |   Ano = 2002: Não (0.0)
|   |   |   |   |   |   Ano = 2003: Sim (1.0)
|   |   |   |   |   |   Ano = 2004
|   |   |   |   |   |   |   Turno = M: Não (47.0)
|   |   |   |   |   |   |   Turno = T
|   |   |   |   |   |   |   |   OcorrenciasMaterial <= 0.476286: Sim (4.0)
|   |   |   |   |   |   |   |   OcorrenciasMaterial > 0.476286: Não (2.0)
|   |   |   |   |   |   |   Turno = N: Não (0.0)
|   |   |   |   |   |   Ano = 2005
|   |   |   |   |   |   |   CobrouMatricula = 0: Não (6.0)
|   |   |   |   |   |   |   CobrouMatricula = 1
|   |   |   |   |   |   |   |   FaltasMatematica <= 0.452528
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 42: Sim (26.0/4.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 42: Não (2.0)
|   |   |   |   |   |   |   |   FaltasMatematica > 0.452528: Não (2.0)
|   |   |   |   |   |   Ano = 2006: Não (10.0/1.0)
|   |   |   |   |   |   Ano = 2007
|   |   |   |   |   |   |   MediaMatematica = 0
|   |   |   |   |   |   |   |   OcorrenciasRuins <= 6: Não (10.0)
|   |   |   |   |   |   |   |   OcorrenciasRuins > 6
|   |   |   |   |   |   |   |   |   OcorrenciasMaterial <= 1: Não (8.0/2.0)
|   |   |   |   |   |   |   |   |   OcorrenciasMaterial > 1: Sim (4.0)
|   |   |   |   |   |   |   MediaMatematica = 1: Sim (2.0)
|   |   |   |   |   |   |   MediaMatematica = 2: Não (0.0)
|   |   |   |   |   |   |   MediaMatematica = 3: Não (0.0)
|   |   |   |   |   |   |   MediaMatematica = 4: Não (0.0)
|   |   |   |   |   |   |   MediaMatematica = 5: Não (0.0)
|   |   |   |   |   |   |   MediaMatematica = 6: Não (0.0)
|   |   |   |   |   |   Ano = 2008: Não (23.0/1.0)
|   |   |   |   |   |   Ano = 2009: Não (30.0/3.0)
|   |   |   |   |   |   Ano = 2010: Não (24.0/1.0)
|   |   |   |   |   |   Ano = 2011: Não (18.0)
|   |   |   |   |   |   Ano = 2012: Não (41.0)
|   |   |   |   |   |   Ano = 2013: Não (22.0/1.0)
|   |   |   |   |   |   Ano = 2014
|   |   |   |   |   |   |   FaltasMatematica <= 2.491941: Não (19.0)
|   |   |   |   |   |   |   FaltasMatematica > 2.491941: Sim (3.0)
|   |   jaReprovou = 1
|   |   |   QtdeAlunos <= 21.49637
|   |   |   |   MediaNaIdade = Igual: Não (7.0/2.0)
|   |   |   |   MediaNaIdade = Menor: Sim (1.0)
|   |   |   |   MediaNaIdade = Maior: Sim (4.0)
|   |   |   QtdeAlunos > 21.49637: Sim (55.0/2.0)
FaltasBiologia > 3.975091
|   QtdeAlunos <= 22.878625
|   |   Turno = M
|   |   |   Curso = Fund-1º-Ano: Sim (4.0)
|   |   |   Curso = Fund-2º-Ano
|   |   |   |   QtdeAlunos <= 22: Sim (70.0/6.0)
|   |   |   |   QtdeAlunos > 22: Não (4.0)
|   |   |   Curso = Fund-3º-Ano
|   |   |   |   Dependentes_Matriculados <= 0.493053
|   |   |   |   |   FaltasMatematica <= 7.230786: Não (5.0/1.0)
|   |   |   |   |   FaltasMatematica > 7.230786: Sim (4.0)
|   |   |   |   Dependentes_Matriculados > 0.493053: Sim (59.0/5.0)
```

```
|   |   |   Curso = Fund-4º-Ano
|   |   |   |   FaltasMatematica <= 5.887197
|   |   |   |   |   MediaNaIdade = Igual: Não (6.0)
|   |   |   |   |   MediaNaIdade = Menor: Não (0.0)
|   |   |   |   |   MediaNaIdade = Maior: Sim (4.0/1.0)
|   |   |   |   FaltasMatematica > 5.887197
|   |   |   |   |   QtdeAlunos <= 14.490086
|   |   |   |   |   |   MediaMatematica = 0: Sim (0.0)
|   |   |   |   |   |   MediaMatematica = 1: Sim (0.0)
|   |   |   |   |   |   MediaMatematica = 2: Sim (0.0)
|   |   |   |   |   |   MediaMatematica = 3: Sim (4.0/1.0)
|   |   |   |   |   |   MediaMatematica = 4: Não (4.0)
|   |   |   |   |   |   MediaMatematica = 5: Sim (4.0/1.0)
|   |   |   |   |   |   MediaMatematica = 6: Sim (0.0)
|   |   |   |   |   QtdeAlunos > 14.490086: Sim (109.0/10.0)
|   |   |   Curso = Fund-5º-Ano
|   |   |   |   QtdeAlunos <= 21.876182: Não (624.0/25.0)
|   |   |   |   QtdeAlunos > 21.876182
|   |   |   |   |   CobrouMatricula = 0: Sim (13.0)
|   |   |   |   |   CobrouMatricula = 1: Não (41.0/3.0)
|   |   |   Curso = Fund-6º-Ano: Sim (14.0/1.0)
|   |   |   Curso = Fund-7º-Ano: Sim (6.0)
|   |   |   Curso = Fund-8º-Ano: Sim (5.0/1.0)
|   |   Turno = T
|   |   |   OcorrenciasRuins <= 0.486073
|   |   |   |   Ano = 2000
|   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75: Não (4.0)
|   |   |   |   |   BolsaEDescontosNormalizado = 76-a-100: Sim (0.0)
|   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5: Sim (12.0)
|   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50: Sim (0.0)
|   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25
|   |   |   |   |   |   AlunoQuantoTempo <= 0.5424: Não (2.0)
|   |   |   |   |   |   AlunoQuantoTempo > 0.5424: Sim (4.0/1.0)
|   |   |   |   Ano = 2001: Sim (45.0/1.0)
|   |   |   |   Ano = 2002: Sim (24.0/1.0)
|   |   |   |   Ano = 2003: Sim (37.0)
|   |   |   |   Ano = 2004: Sim (53.0/7.0)
|   |   |   |   Ano = 2005: Sim (36.0/9.0)
|   |   |   |   Ano = 2006
|   |   |   |   |   EhAdventista = Sim: Sim (13.0)
|   |   |   |   |   EhAdventista = Não
|   |   |   |   |   |   Curso = Fund-1º-Ano: Sim (0.0)
|   |   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   |   Dependentes_Matriculados <= 1.496705
|   |   |   |   |   |   |   |   AlunoQuantoTempo <= 1.498619: Sim (10.0/3.0)
|   |   |   |   |   |   |   |   AlunoQuantoTempo > 1.498619: Não (8.0/2.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 1.496705: Não (14.0/1.0)
|   |   |   |   |   |   Curso = Fund-3º-Ano: Sim (5.0/1.0)
|   |   |   |   |   |   Curso = Fund-4º-Ano: Sim (12.0)
|   |   |   |   |   |   Curso = Fund-5º-Ano
|   |   |   |   |   |   |   CobrouMatricula = 0: Não (3.0/1.0)
|   |   |   |   |   |   |   CobrouMatricula = 1: Sim (6.0)
|   |   |   |   |   |   Curso = Fund-6º-Ano: Sim (0.0)
|   |   |   |   |   |   Curso = Fund-7º-Ano: Sim (0.0)
|   |   |   |   |   |   Curso = Fund-8º-Ano: Sim (0.0)
|   |   |   |   Ano = 2007
|   |   |   |   |   MediaNaIdade = Igual: Sim (42.0/4.0)
|   |   |   |   |   MediaNaIdade = Menor: Não (3.0/1.0)
|   |   |   |   |   MediaNaIdade = Maior
|   |   |   |   |   |   EhAdventista = Sim: Sim (2.0)
|   |   |   |   |   |   EhAdventista = Não: Não (8.0/2.0)
|   |   |   |   Ano = 2008
|   |   |   |   |   FaltasMatematica <= 6.42345: Não (5.0)
|   |   |   |   |   FaltasMatematica > 6.42345: Sim (55.0/11.0)
|   |   |   |   Ano = 2009
|   |   |   |   |   FaltasMatematica <= 7.230786: Não (3.0)
|   |   |   |   |   FaltasMatematica > 7.230786: Sim (16.0/2.0)
|   |   |   |   Ano = 2010
|   |   |   |   |   AlunoQuantoTempo <= 3.491928: Não (14.0/1.0)
|   |   |   |   |   AlunoQuantoTempo > 3.491928: Sim (5.0)
|   |   |   |   Ano = 2011: Sim (30.0)
|   |   |   |   Ano = 2012: Sim (12.0)
|   |   |   |   Ano = 2013: Sim (23.0)
|   |   |   |   Ano = 2014: Sim (21.0)
|   |   |   OcorrenciasRuins > 0.486073: Sim (92.0)
|   |   Turno = N: Sim (0.0)
|   QtdeAlunos > 22.878625
|   |   OcorrenciasRuins <= 0.875155
```

```
|   |   |   Ano = 2000
|   |   |   |   CobrouMatricula = 0
|   |   |   |   |   QtdeAlunos <= 30.496032
|   |   |   |   |   |   Turno = M
|   |   |   |   |   |   |   FaltasMatematica <= 10.477524
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.955338
|   |   |   |   |   |   |   |   |   QtdeAlunos <= 26.149397
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 25.750165: Não (18.0/2.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 25.750165: Sim (11.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 26.149397: Não (43.0/2.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.955338: Sim (14.0/2.0)
|   |   |   |   |   |   |   FaltasMatematica > 10.477524: Sim (9.0)
|   |   |   |   |   |   Turno = T: Sim (110.0/14.0)
|   |   |   |   |   |   Turno = N: Sim (0.0)
|   |   |   |   |   QtdeAlunos > 30.496032: Sim (128.0/4.0)
|   |   |   |   CobrouMatricula = 1: Não (16.0)
|   |   |   Ano = 2001: Sim (221.0/4.0)
|   |   |   Ano = 2002: Sim (231.0/3.0)
|   |   |   Ano = 2003: Sim (281.0)
|   |   |   Ano = 2004
|   |   |   |   QtdeAlunos <= 25.814838
|   |   |   |   |   FaltasMatematica <= 6.42345: Não (4.0)
|   |   |   |   |   FaltasMatematica > 6.42345: Sim (47.0/4.0)
|   |   |   |   QtdeAlunos > 25.814838: Sim (337.0/7.0)
|   |   |   Ano = 2005
|   |   |   |   QtdeAlunos <= 24.982336
|   |   |   |   |   FaltasMatematica <= 9.461704
|   |   |   |   |   |   Curso = Fund-1º-Ano: Não (0.0)
|   |   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   |   EhAdventista = Sim: Sim (2.0)
|   |   |   |   |   |   |   EhAdventista = Não
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação
|   |   |   |   |   |   |   |   |   TemIrmao = 0: Não (12.0)
|   |   |   |   |   |   |   |   |   TemIrmao = 1: Sim (2.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Não (0.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo: Sim
(2.0/1.0)
|   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Sim (1.0)
|   |   |   |   |   |   Curso = Fund-3º-Ano: Sim (4.0/2.0)
|   |   |   |   |   |   Curso = Fund-4º-Ano
|   |   |   |   |   |   |   EhAdventista = Sim: Sim (3.0)
|   |   |   |   |   |   |   EhAdventista = Não: Não (2.0)
|   |   |   |   |   |   Curso = Fund-5º-Ano: Não (45.0/1.0)
|   |   |   |   |   |   Curso = Fund-6º-Ano: Não (0.0)
|   |   |   |   |   |   Curso = Fund-7º-Ano: Não (0.0)
|   |   |   |   |   |   Curso = Fund-8º-Ano: Não (0.0)
|   |   |   |   |   FaltasMatematica > 9.461704: Sim (11.0/1.0)
|   |   |   |   QtdeAlunos > 24.982336
|   |   |   |   |   EhAdventista = Sim
|   |   |   |   |   |   VeioDeForaDoEstado = Sim
|   |   |   |   |   |   |   Genero = M: Não (5.0)
|   |   |   |   |   |   |   Genero = F: Sim (2.0)
|   |   |   |   |   |   VeioDeForaDoEstado = Não: Sim (70.0/6.0)
|   |   |   |   |   EhAdventista = Não
|   |   |   |   |   |   MediaMatematica = 0
|   |   |   |   |   |   |   Turno = M: Não (11.0)
|   |   |   |   |   |   |   Turno = T: Sim (5.0/2.0)
|   |   |   |   |   |   |   Turno = N: Não (0.0)
|   |   |   |   |   |   MediaMatematica = 1
|   |   |   |   |   |   |   QtdeAlunos <= 27.7477: Sim (12.0)
|   |   |   |   |   |   |   QtdeAlunos > 27.7477: Não (23.0/7.0)
|   |   |   |   |   |   MediaMatematica = 2
|   |   |   |   |   |   |   QtdeAlunos <= 34.496416: Não (3.0)
|   |   |   |   |   |   |   QtdeAlunos > 34.496416: Sim (4.0)
|   |   |   |   |   |   MediaMatematica = 3
|   |   |   |   |   |   |   Curso = Fund-1º-Ano: Não (0.0)
|   |   |   |   |   |   |   Curso = Fund-2º-Ano: Não (0.0)
|   |   |   |   |   |   |   Curso = Fund-3º-Ano
|   |   |   |   |   |   |   |   Genero = M: Não (2.0)
|   |   |   |   |   |   |   |   Genero = F: Sim (3.0/1.0)
|   |   |   |   |   |   |   Curso = Fund-4º-Ano: Sim (8.0/1.0)
|   |   |   |   |   |   |   Curso = Fund-5º-Ano: Não (4.0)
|   |   |   |   |   |   |   Curso = Fund-6º-Ano: Sim (3.0/1.0)
|   |   |   |   |   |   |   Curso = Fund-7º-Ano: Não (2.0)
|   |   |   |   |   |   |   Curso = Fund-8º-Ano: Não (1.0)
|   |   |   |   |   |   MediaMatematica = 4
|   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   |   |   EstadoCivil = Solteiro: Sim (2.0)
```

```
|   |   |   |   |   |   |   |   EstadoCivil = Casado
|   |   |   |   |   |   |   |   |   MediaNaIdade = Igual
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = <25: Não (0.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = >=71: Não (0.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = 36-40
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-
Graduação: Sim (1.0)
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio:
Não (6.0)
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo
|   |   |   |   |   |   |   |   |   |   |   |   Genero = M: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   Genero = F: Não (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-
Incompleto: Não (0.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = 51-70: Não (1.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = 41-50: Não (12.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = 31-35
|   |   |   |   |   |   |   |   |   |   |   FaltasMatematica <= 8.239094: Sim (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   FaltasMatematica > 8.239094: Não (2.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = 26-30: Sim (1.0)
|   |   |   |   |   |   |   |   |   MediaNaIdade = Menor: Sim (2.0)
|   |   |   |   |   |   |   |   |   MediaNaIdade = Maior
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 37.443807: Sim (9.0/1.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 37.443807: Não (2.0)
|   |   |   |   |   |   |   |   EstadoCivil = Separado: Sim (5.0/1.0)
|   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (3.0)
|   |   |   |   |   |   MediaMatematica = 5
|   |   |   |   |   |   |   TemIrmao = 0
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.299728
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 0.498392: Não (3.0/1.0)
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 0.498392: Sim (19.0/1.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.299728
|   |   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim: Não (5.0)
|   |   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (2.0)
|   |   |   |   |   |   |   TemIrmao = 1: Sim (20.0)
|   |   |   |   |   |   MediaMatematica = 6
|   |   |   |   |   |   |   AlunoQuantoTempo <= 6.198975: Sim (17.0/1.0)
|   |   |   |   |   |   |   AlunoQuantoTempo > 6.198975: Não (2.0)
|   |   |   Ano = 2006
|   |   |   |   Dependentes_Matriculados <= 0.07495: Sim (37.0/1.0)
|   |   |   |   Dependentes_Matriculados > 0.07495
|   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   EhAdventista = Sim
|   |   |   |   |   |   |   VeioDeForaDoEstado = Sim
|   |   |   |   |   |   |   |   QtdeAlunos <= 32.470402: Sim (7.0/1.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 32.470402: Não (3.0)
|   |   |   |   |   |   |   VeioDeForaDoEstado = Não: Sim (71.0/12.0)
|   |   |   |   |   |   EhAdventista = Não
|   |   |   |   |   |   |   Dependentes_Matriculados <= 0.993764: Não (20.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 0.993764
|   |   |   |   |   |   |   |   Turno = M
|   |   |   |   |   |   |   |   |   Curso = Fund-1º-Ano: Não (0.0)
|   |   |   |   |   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 31.823007
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 27.850481: Não (10.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos > 27.850481: Sim (10.0/1.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 31.823007: Não (7.0)
|   |   |   |   |   |   |   |   |   Curso = Fund-3º-Ano
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.062006
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 31.19249: Sim (11.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos > 31.19249: Não (4.0)
|   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.062006: Não
(10.0/1.0)
|   |   |   |   |   |   |   |   |   Curso = Fund-4º-Ano
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75: Sim (1.0)
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 76-a-100: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5: Sim (3.0)
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-
Graduação: Não (3.0)
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio:
Não (0.0)
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo:
Sim (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-
Incompleto: Não (0.0)
|   |   |   |   |   |   |   |   |   Curso = Fund-5º-Ano
```

```
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75: Não (0.0)
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 76-a-100: Não (0.0)
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 29.488976: Não (11.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos > 29.488976
|   |   |   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 38: Sim (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   QtdeAlunos > 38
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 0: Não (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 1: Não (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 2: Não (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 3: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 4: Não (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 5: Não (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 6: Não (0.0)
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50: Sim
(6.0/2.0)
|   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25: Não
(37.0/3.0)
|   |   |   |   |   |   |   |   |   Curso = Fund-6º-Ano
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 0: Não (5.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 1: Não (2.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 2: Sim (1.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 3: Não (4.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 4: Sim (4.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 5: Sim (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 6: Sim (3.0/1.0)
|   |   |   |   |   |   |   |   |   Curso = Fund-7º-Ano
|   |   |   |   |   |   |   |   |   |   TemIrmao = 0: Sim (7.0/1.0)
|   |   |   |   |   |   |   |   |   |   TemIrmao = 1: Não (3.0)
|   |   |   |   |   |   |   |   |   Curso = Fund-8º-Ano: Sim (15.0/3.0)
|   |   |   |   |   |   |   |   Turno = T
|   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Solteiro: Sim (7.0)
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Casado
|   |   |   |   |   |   |   |   |   |   |   FaixaEtaria = <25: Não (4.0)
|   |   |   |   |   |   |   |   |   |   |   FaixaEtaria = >=71: Não (0.0)
|   |   |   |   |   |   |   |   |   |   |   FaixaEtaria = 36-40
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75: Não
(3.0)
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 76-a-100:
Não (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5: Não
(5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 0: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 1: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 2: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 3: Não (1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 4
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <=
1.811647: Não (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados >
1.811647: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 5: Sim (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 6: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25: Não
(17.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   FaixaEtaria = 51-70: Sim (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   FaixaEtaria = 41-50
|   |   |   |   |   |   |   |   |   |   |   |   Curso = Fund-1º-Ano: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   |   |   |   |   |   |   |   Genero = M: Não (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   Genero = F: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   Curso = Fund-3º-Ano: Sim (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   Curso = Fund-4º-Ano: Não (10.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   |   Curso = Fund-5º-Ano
|   |   |   |   |   |   |   |   |   |   |   |   |   TemIrmao = 0: Não (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   TemIrmao = 1: Sim (4.0)
|   |   |   |   |   |   |   |   |   |   |   |   Curso = Fund-6º-Ano: Não (1.0)
|   |   |   |   |   |   |   |   |   |   |   |   Curso = Fund-7º-Ano: Sim (1.0)
|   |   |   |   |   |   |   |   |   |   |   |   Curso = Fund-8º-Ano: Sim (1.0)
|   |   |   |   |   |   |   |   |   |   |   FaixaEtaria = 31-35
|   |   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.241878: Sim
(9.0)
|   |   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.241878: Não
(4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   FaixaEtaria = 26-30: Sim (7.0/1.0)
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Separado
```

```
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.496705
|   |   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 0.498392: Não
(4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 0.498392: Sim (7.0)
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.496705: Não
(3.0)
|   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Sim
(20.0/3.0)
|   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo
|   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 2.497997: Sim (44.0/4.0)
|   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 2.497997
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 29.488976: Sim (6.0)
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos > 29.488976: Não (9.0/3.0)
|   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Sim
(12.0/3.0)
|   |   |   |   |   |   |   |   Turno = N: Sim (0.0)
|   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (111.0/20.0)
|   |   |   Ano = 2007
|   |   |   |   EhAdventista = Sim: Sim (105.0/7.0)
|   |   |   |   EhAdventista = Não
|   |   |   |   |   Curso = Fund-1º-Ano: Sim (0.0)
|   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   MediaMatematica = 0
|   |   |   |   |   |   |   FaixaEtaria = <25: Não (0.0)
|   |   |   |   |   |   |   FaixaEtaria = >=71: Não (0.0)
|   |   |   |   |   |   |   FaixaEtaria = 36-40: Sim (2.0)
|   |   |   |   |   |   |   FaixaEtaria = 51-70: Não (1.0)
|   |   |   |   |   |   |   FaixaEtaria = 41-50: Não (6.0/1.0)
|   |   |   |   |   |   |   FaixaEtaria = 31-35: Sim (2.0)
|   |   |   |   |   |   |   FaixaEtaria = 26-30: Não (1.0)
|   |   |   |   |   |   MediaMatematica = 1: Sim (41.0/8.0)
|   |   |   |   |   |   MediaMatematica = 2: Sim (0.0)
|   |   |   |   |   |   MediaMatematica = 3: Sim (0.0)
|   |   |   |   |   |   MediaMatematica = 4: Sim (0.0)
|   |   |   |   |   |   MediaMatematica = 5: Não (1.0)
|   |   |   |   |   |   MediaMatematica = 6: Não (2.0)
|   |   |   |   |   Curso = Fund-3º-Ano
|   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Sim (20.0/1.0)
|   |   |   |   |   |   Dependentes_Matriculados > 0.493053
|   |   |   |   |   |   |   FaltasMatematica <= 8.239094
|   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim: Não (23.0/7.0)
|   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (3.0)
|   |   |   |   |   |   |   FaltasMatematica > 8.239094
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 2.110319: Sim (42.0/8.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 2.110319: Não (2.0)
|   |   |   |   |   Curso = Fund-4º-Ano
|   |   |   |   |   |   MediaNaIdade = Igual
|   |   |   |   |   |   |   VeioDeForaDoEstado = Sim: Não (2.0)
|   |   |   |   |   |   |   VeioDeForaDoEstado = Não
|   |   |   |   |   |   |   |   QtdeAlunos <= 33.033441
|   |   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 32.118848
|   |   |   |   |   |   |   |   |   |   |   EstadoCivil = Solteiro: Não (1.0)
|   |   |   |   |   |   |   |   |   |   |   EstadoCivil = Casado: Sim (20.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   EstadoCivil = Separado: Não (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 32.118848
|   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 1.498619: Não (15.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 1.498619: Sim (5.0/1.0)
|   |   |   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (13.0/1.0)
|   |   |   |   |   |   |   |   QtdeAlunos > 33.033441: Não (4.0)
|   |   |   |   |   |   MediaNaIdade = Menor: Não (1.0)
|   |   |   |   |   |   MediaNaIdade = Maior: Sim (15.0/2.0)
|   |   |   |   |   Curso = Fund-5º-Ano
|   |   |   |   |   |   MediaNaIdade = Igual
|   |   |   |   |   |   |   QtdeAlunos <= 39.58777
|   |   |   |   |   |   |   |   QtdeAlunos <= 34.089423
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 3.491928
|   |   |   |   |   |   |   |   |   |   TemIrmao = 0
|   |   |   |   |   |   |   |   |   |   |   Genero = M
|   |   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-
Graduação: Não (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-
Medio: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-
Completo: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-
Incompleto: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   Genero = F: Não (10.0/1.0)
```

```
|   |   |   |   |   |   |   |   |   |   TemIrmao = 1: Sim (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 3.491928: Sim (7.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 34.089423: Não (32.0/3.0)
|   |   |   |   |   |   |   |   |   QtdeAlunos > 39.58777: Sim (6.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Menor: Não (2.0)
|   |   |   |   |   |   |   |   MediaNaIdade = Maior: Não (5.0)
|   |   |   |   |   |   |   Curso = Fund-6º-Ano
|   |   |   |   |   |   |   |   MediaMatematica = 0: Não (3.0)
|   |   |   |   |   |   |   |   MediaMatematica = 1: Não (10.0/1.0)
|   |   |   |   |   |   |   |   MediaMatematica = 2
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 1.498619: Não (3.0/1.0)
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 1.498619: Sim (3.0)
|   |   |   |   |   |   |   |   MediaMatematica = 3
|   |   |   |   |   |   |   |   |   FaltasMatematica <= 8.239094: Sim (8.0/1.0)
|   |   |   |   |   |   |   |   |   FaltasMatematica > 8.239094
|   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação: Não (9.0)
|   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Não (0.0)
|   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo: Sim
(3.0/1.0)
|   |   |   |   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Não (0.0)
|   |   |   |   |   |   |   |   MediaMatematica = 4
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 4.309012: Sim (15.0/3.0)
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 4.309012
|   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 37.6111: Não (6.0/1.0)
|   |   |   |   |   |   |   |   |   |   QtdeAlunos > 37.6111: Sim (2.0)
|   |   |   |   |   |   |   |   MediaMatematica = 5
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 4.476244: Sim (13.0)
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 4.476244: Não (3.0/1.0)
|   |   |   |   |   |   |   |   MediaMatematica = 6: Não (5.0)
|   |   |   |   |   |   |   Curso = Fund-7º-Ano
|   |   |   |   |   |   |   |   FaltasMatematica <= 13.383775: Sim (31.0/3.0)
|   |   |   |   |   |   |   |   FaltasMatematica > 13.383775: Não (10.0/2.0)
|   |   |   |   |   |   |   Curso = Fund-8º-Ano: Sim (36.0/5.0)
|   |   |   |   |   Ano = 2008
|   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação
|   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   |   |   TemIrmao = 0
|   |   |   |   |   |   |   |   |   Genero = M
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Solteiro
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos <= 31.472983: Não (3.0)
|   |   |   |   |   |   |   |   |   |   |   QtdeAlunos > 31.472983: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Casado
|   |   |   |   |   |   |   |   |   |   |   MediaNaIdade = Igual: Não (53.0/11.0)
|   |   |   |   |   |   |   |   |   |   |   MediaNaIdade = Menor: Não (2.0)
|   |   |   |   |   |   |   |   |   |   |   MediaNaIdade = Maior
|   |   |   |   |   |   |   |   |   |   |   |   FaltasMatematica <= 8.239094: Sim (7.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   |   FaltasMatematica > 8.239094: Não (7.0/2.0)
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Separado: Sim (9.0/2.0)
|   |   |   |   |   |   |   |   |   Genero = F
|   |   |   |   |   |   |   |   |   |   Curso = Fund-1º-Ano: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 0.358468: Sim (3.0)
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 0.358468: Não (17.0/2.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-3º-Ano: Sim (10.0/3.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-4º-Ano: Sim (17.0/2.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-5º-Ano
|   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 76-a-100: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50: Sim (5.0)
|   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25: Não (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-6º-Ano: Não (5.0/2.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-7º-Ano: Sim (4.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-8º-Ano: Sim (8.0/2.0)
|   |   |   |   |   |   |   |   |   TemIrmao = 1
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 0: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 1: Não (1.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 2: Sim (5.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 3
|   |   |   |   |   |   |   |   |   |   |   MediaNaIdade = Igual
|   |   |   |   |   |   |   |   |   |   |   |   FaltasMatematica <= 10.477524: Não (5.0)
|   |   |   |   |   |   |   |   |   |   |   |   FaltasMatematica > 10.477524: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   MediaNaIdade = Menor: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   MediaNaIdade = Maior: Sim (4.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 4
|   |   |   |   |   |   |   |   |   |   |   VeioDeForaDoEstado = Sim: Não (2.0)
|   |   |   |   |   |   |   |   |   |   |   VeioDeForaDoEstado = Não
|   |   |   |   |   |   |   |   |   |   |   |   CobrouMatricula = 0: Não (3.0/1.0)
```

```
|   |   |   |   |   |   |   |   |   CobrouMatricula = 1: Sim (20.0/3.0)
|   |   |   |   |   |   |   |   MediaMatematica = 5
|   |   |   |   |   |   |   |   |   VeioDeForaDoEstado = Sim: Não (3.0/1.0)
|   |   |   |   |   |   |   |   |   VeioDeForaDoEstado = Não: Sim (17.0/1.0)
|   |   |   |   |   |   |   |   MediaMatematica = 6
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 2.497997: Sim (2.0)
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 2.497997: Não (3.0/1.0)
|   |   |   |   |   ResponsavelLegalEhFinanceiro = Não
|   |   |   |   |   |   EhAdventista = Sim
|   |   |   |   |   |   |   MediaMatematica = 0: Sim (0.0)
|   |   |   |   |   |   |   MediaMatematica = 1: Sim (0.0)
|   |   |   |   |   |   |   MediaMatematica = 2: Sim (0.0)
|   |   |   |   |   |   |   MediaMatematica = 3: Sim (3.0/1.0)
|   |   |   |   |   |   |   MediaMatematica = 4: Sim (2.0)
|   |   |   |   |   |   |   MediaMatematica = 5: Não (3.0)
|   |   |   |   |   |   |   MediaMatematica = 6: Sim (0.0)
|   |   |   |   |   |   EhAdventista = Não: Sim (29.0/1.0)
|   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Sim (34.0/1.0)
|   |   |   |   GrauInstrucaoResponsavel = Superior-Completo
|   |   |   |   |   MediaNaIdade = Igual
|   |   |   |   |   |   EhAdventista = Sim: Sim (28.0/1.0)
|   |   |   |   |   |   EhAdventista = Não
|   |   |   |   |   |   |   EstadoCivil = Solteiro: Sim (2.0/1.0)
|   |   |   |   |   |   |   EstadoCivil = Casado
|   |   |   |   |   |   |   |   FaltasMatematica <= 12.208869
|   |   |   |   |   |   |   |   |   VeioDeForaDoEstado = Sim: Sim (5.0)
|   |   |   |   |   |   |   |   |   VeioDeForaDoEstado = Não
|   |   |   |   |   |   |   |   |   |   Curso = Fund-1º-Ano: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-2º-Ano: Sim (5.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-3º-Ano
|   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 0.498392: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 0.498392: Não (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-4º-Ano
|   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 0.498392: Não (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 0.498392: Sim (12.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-5º-Ano
|   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 0: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 1: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 2: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 3
|   |   |   |   |   |   |   |   |   |   |   |   Genero = M: Não (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   Genero = F: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 4: Sim (6.0/2.0)
|   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 5
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75: Não
(0.0)
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 76-a-100:
Não (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5: Sim
(2.0)
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50: Não
(1.0)
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25: Não
(4.0)
|   |   |   |   |   |   |   |   |   |   |   MediaMatematica = 6: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-6º-Ano
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.496705: Sim
(6.0)
|   |   |   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.496705: Não
(4.0/1.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-7º-Ano
|   |   |   |   |   |   |   |   |   |   |   FaltasMatematica <= 9.141806: Sim (4.0)
|   |   |   |   |   |   |   |   |   |   |   FaltasMatematica > 9.141806: Não (10.0/1.0)
|   |   |   |   |   |   |   |   |   |   Curso = Fund-8º-Ano: Sim (5.0/1.0)
|   |   |   |   |   |   |   |   FaltasMatematica > 12.208869: Não (6.0)
|   |   |   |   |   |   |   EstadoCivil = Separado: Sim (12.0)
|   |   |   |   |   MediaNaIdade = Menor: Sim (7.0)
|   |   |   |   |   MediaNaIdade = Maior: Sim (45.0/2.0)
|   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto
|   |   |   |   |   QtdeAlunos <= 31.472983: Não (16.0/6.0)
|   |   |   |   |   QtdeAlunos > 31.472983: Sim (30.0/3.0)
|   |   |   Ano = 2009
|   |   |   |   EhAdventista = Sim: Sim (48.0/8.0)
|   |   |   |   EhAdventista = Não
|   |   |   |   |   MediaMatematica = 0: Não (6.0)
|   |   |   |   |   MediaMatematica = 1: Não (11.0/2.0)
|   |   |   |   |   MediaMatematica = 2: Não (10.0)
|   |   |   |   |   MediaMatematica = 3
```

```
|   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Sim (4.0)
|   |   |   |   |   |   Dependentes_Matriculados > 0.493053
|   |   |   |   |   |   |   QtdeAlunos <= 38
|   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.577953: Não (15.0/1.0)
|   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.577953: Sim (4.0/1.0)
|   |   |   |   |   |   |   QtdeAlunos > 38: Sim (4.0)
|   |   |   |   |   MediaMatematica = 4
|   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   |   AlunoQuantoTempo <= 3.224889: Sim (30.0/7.0)
|   |   |   |   |   |   |   AlunoQuantoTempo > 3.224889: Não (9.0/2.0)
|   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Sim (3.0)
|   |   |   |   |   MediaMatematica = 5
|   |   |   |   |   |   VeioDeForaDoEstado = Sim: Sim (2.0)
|   |   |   |   |   |   VeioDeForaDoEstado = Não
|   |   |   |   |   |   |   FaixaEtaria = <25: Não (0.0)
|   |   |   |   |   |   |   FaixaEtaria = >=71: Não (0.0)
|   |   |   |   |   |   |   FaixaEtaria = 36-40
|   |   |   |   |   |   |   |   Genero = M: Não (8.0)
|   |   |   |   |   |   |   |   Genero = F
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 2.497997: Sim (3.0)
|   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 2.497997: Não (2.0)
|   |   |   |   |   |   |   FaixaEtaria = 51-70: Sim (2.0/1.0)
|   |   |   |   |   |   |   FaixaEtaria = 41-50: Sim (11.0/3.0)
|   |   |   |   |   |   |   FaixaEtaria = 31-35: Não (8.0/3.0)
|   |   |   |   |   |   |   FaixaEtaria = 26-30: Sim (4.0/1.0)
|   |   |   |   |   MediaMatematica = 6
|   |   |   |   |   |   FaixaEtaria = <25: Sim (1.0)
|   |   |   |   |   |   FaixaEtaria = >=71: Sim (0.0)
|   |   |   |   |   |   FaixaEtaria = 36-40: Sim (6.0/2.0)
|   |   |   |   |   |   FaixaEtaria = 51-70: Não (1.0)
|   |   |   |   |   |   FaixaEtaria = 41-50: Não (4.0)
|   |   |   |   |   |   FaixaEtaria = 31-35: Sim (7.0/2.0)
|   |   |   |   |   |   FaixaEtaria = 26-30: Sim (2.0)
|   |   |   Ano = 2010
|   |   |   |   TemIrmao = 0
|   |   |   |   |   Curso = Fund-1º-Ano: Não (0.0)
|   |   |   |   |   Curso = Fund-2º-Ano
|   |   |   |   |   |   QtdeAlunos <= 31.472983: Não (39.0/5.0)
|   |   |   |   |   |   QtdeAlunos > 31.472983: Sim (7.0/2.0)
|   |   |   |   |   Curso = Fund-3º-Ano: Não (38.0/13.0)
|   |   |   |   |   Curso = Fund-4º-Ano
|   |   |   |   |   |   FaltasMatematica <= 8.239094
|   |   |   |   |   |   |   FaltasMatematica <= 7.230786: Sim (5.0)
|   |   |   |   |   |   |   FaltasMatematica > 7.230786: Não (6.0/1.0)
|   |   |   |   |   |   FaltasMatematica > 8.239094: Sim (7.0)
|   |   |   |   |   Curso = Fund-5º-Ano: Não (38.0/11.0)
|   |   |   |   |   Curso = Fund-6º-Ano
|   |   |   |   |   |   QtdeAlunos <= 32.818545: Não (23.0/2.0)
|   |   |   |   |   |   QtdeAlunos > 32.818545: Sim (9.0/2.0)
|   |   |   |   |   Curso = Fund-7º-Ano: Não (12.0)
|   |   |   |   |   Curso = Fund-8º-Ano: Sim (4.0/1.0)
|   |   |   |   TemIrmao = 1
|   |   |   |   |   MediaMatematica = 0: Sim (0.0)
|   |   |   |   |   MediaMatematica = 1: Não (3.0/1.0)
|   |   |   |   |   MediaMatematica = 2: Não (4.0)
|   |   |   |   |   MediaMatematica = 3
|   |   |   |   |   |   EstadoCivil = Solteiro: Não (1.0)
|   |   |   |   |   |   EstadoCivil = Casado: Sim (8.0/3.0)
|   |   |   |   |   |   EstadoCivil = Separado: Não (2.0)
|   |   |   |   |   MediaMatematica = 4
|   |   |   |   |   |   FaltasMatematica <= 7.230786: Não (4.0/1.0)
|   |   |   |   |   |   FaltasMatematica > 7.230786: Sim (9.0)
|   |   |   |   |   MediaMatematica = 5
|   |   |   |   |   |   Dependentes_Matriculados <= 2.481693: Sim (14.0/1.0)
|   |   |   |   |   |   Dependentes_Matriculados > 2.481693: Não (3.0/1.0)
|   |   |   |   |   MediaMatematica = 6
|   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   |   EhAdventista = Sim: Sim (4.0)
|   |   |   |   |   |   |   EhAdventista = Não
|   |   |   |   |   |   |   |   VeioDeForaDoEstado = Sim: Não (2.0)
|   |   |   |   |   |   |   |   VeioDeForaDoEstado = Não
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados <= 1.496705: Sim (3.0)
|   |   |   |   |   |   |   |   |   Dependentes_Matriculados > 1.496705
|   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo <= 3.491928: Não (4.0/1.0)
|   |   |   |   |   |   |   |   |   |   AlunoQuantoTempo > 3.491928: Sim (2.0)
|   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não: Não (2.0)
|   |   |   Ano = 2011: Sim (342.0/5.0)
|   |   |   Ano = 2012
```

```
|   |   |   |   QtdeAlunos <= 24.993476
|   |   |   |   |   Turno = M
|   |   |   |   |   |   FaltasMatematica <= 7.230786: Não (27.0/3.0)
|   |   |   |   |   |   FaltasMatematica > 7.230786
|   |   |   |   |   |   |   VeioDeForaDoEstado = Sim: Não (6.0/1.0)
|   |   |   |   |   |   |   VeioDeForaDoEstado = Não: Sim (15.0/1.0)
|   |   |   |   |   Turno = T: Sim (32.0/1.0)
|   |   |   |   |   Turno = N: Sim (0.0)
|   |   |   |   QtdeAlunos > 24.993476: Sim (362.0/2.0)
|   |   |   Ano = 2013
|   |   |   |   QtdeAlunos <= 25.718616
|   |   |   |   |   Turno = M
|   |   |   |   |   |   Curso = Fund-1º-Ano: Não (0.0)
|   |   |   |   |   |   Curso = Fund-2º-Ano: Sim (7.0)
|   |   |   |   |   |   Curso = Fund-3º-Ano: Sim (6.0)
|   |   |   |   |   |   Curso = Fund-4º-Ano: Sim (3.0)
|   |   |   |   |   |   Curso = Fund-5º-Ano
|   |   |   |   |   |   |   MediaNaIdade = Igual: Não (34.0/4.0)
|   |   |   |   |   |   |   MediaNaIdade = Menor: Não (0.0)
|   |   |   |   |   |   |   MediaNaIdade = Maior: Sim (2.0)
|   |   |   |   |   |   Curso = Fund-6º-Ano: Não (0.0)
|   |   |   |   |   |   Curso = Fund-7º-Ano: Sim (1.0)
|   |   |   |   |   |   Curso = Fund-8º-Ano: Sim (2.0)
|   |   |   |   |   Turno = T: Sim (51.0)
|   |   |   |   |   Turno = N: Sim (0.0)
|   |   |   |   QtdeAlunos > 25.718616: Sim (281.0)
|   |   |   Ano = 2014: Sim (453.0/2.0)
|   |   OcorrenciasRuins > 0.875155
|   |   |   FaltasBiologia <= 5.946199
|   |   |   |   MediaMatematica = 0
|   |   |   |   |   jaReprovou = 0
|   |   |   |   |   |   Dependentes_Matriculados <= 0.493053: Sim (5.0)
|   |   |   |   |   |   Dependentes_Matriculados > 0.493053: Não (56.0/2.0)
|   |   |   |   |   jaReprovou = 1: Sim (13.0/1.0)
|   |   |   |   MediaMatematica = 1: Sim (121.0/9.0)
|   |   |   |   MediaMatematica = 2: Sim (35.0)
|   |   |   |   MediaMatematica = 3: Sim (46.0)
|   |   |   |   MediaMatematica = 4: Sim (3.0)
|   |   |   |   MediaMatematica = 5: Sim (0.0)
|   |   |   |   MediaMatematica = 6: Sim (0.0)
|   |   |   FaltasBiologia > 5.946199
|   |   |   |   MediaHistoria = 0
|   |   |   |   |   jaReprovou = 0
|   |   |   |   |   |   Curso = Fund-1º-Ano: Sim (0.0)
|   |   |   |   |   |   Curso = Fund-2º-Ano: Sim (4.0)
|   |   |   |   |   |   Curso = Fund-3º-Ano: Sim (0.0)
|   |   |   |   |   |   Curso = Fund-4º-Ano: Sim (0.0)
|   |   |   |   |   |   Curso = Fund-5º-Ano: Sim (0.0)
|   |   |   |   |   |   Curso = Fund-6º-Ano: Não (12.0)
|   |   |   |   |   |   Curso = Fund-7º-Ano
|   |   |   |   |   |   |   FaltasBiologia <= 10.59199: Não (10.0/2.0)
|   |   |   |   |   |   |   FaltasBiologia > 10.59199: Sim (8.0/1.0)
|   |   |   |   |   |   Curso = Fund-8º-Ano: Sim (16.0/4.0)
|   |   |   |   |   jaReprovou = 1: Sim (22.0)
|   |   |   |   MediaHistoria = 1: Sim (326.0/10.0)
|   |   |   |   MediaHistoria = 2: Sim (325.0)
|   |   |   |   MediaHistoria = 3: Sim (812.0)
|   |   |   |   MediaHistoria = 4: Sim (912.0)
|   |   |   |   MediaHistoria = 5: Sim (707.0)
|   |   |   |   MediaHistoria = 6: Sim (152.0)
```

Apêndice B

Árvore de Decisão do Quinto Experimento

```
FaltasBiologia <= 3.975091
|   Terceiro = A|A|A
|   |   Primeiro = A|A|A: Sim (644.0/49.0)
|   |   Primeiro = A|A|D: Sim (59.0/2.0)
|   |   Primeiro = A|A|N: Não (5.0/1.0)
|   |   Primeiro = A|D|A: Sim (44.0)
|   |   Primeiro = A|D|D
|   |   |   QtdeAlunos <= 26.51853
|   |   |   |   Genero = M
|   |   |   |   |   OcorrenciasRuins <= 0.486073: Não (9.0/2.0)
|   |   |   |   |   OcorrenciasRuins > 0.486073: Sim (2.0)
|   |   |   |   Genero = F: Sim (4.0)
|   |   |   QtdeAlunos > 26.51853: Sim (24.0)
|   |   Primeiro = A|D|N: Sim (0.0)
|   |   Primeiro = A|N|A: Sim (0.0)
|   |   Primeiro = A|N|D: Sim (0.0)
|   |   Primeiro = A|N|N: Não (1.0)
|   |   Primeiro = D|A|A: Sim (89.0/7.0)
|   |   Primeiro = D|A|D: Sim (13.0/1.0)
|   |   Primeiro = D|A|N: Não (1.0)
|   |   Primeiro = D|D|A: Sim (21.0)
|   |   Primeiro = D|D|D: Sim (28.0/2.0)
|   |   Primeiro = D|D|N: Sim (0.0)
|   |   Primeiro = D|N|A: Não (1.0)
|   |   Primeiro = D|N|D: Sim (0.0)
|   |   Primeiro = D|N|N: Sim (0.0)
|   |   Primeiro = N|A|A: Não (10.0)
|   |   Primeiro = N|A|D: Sim (0.0)
|   |   Primeiro = N|A|N: Sim (0.0)
|   |   Primeiro = N|D|A: Não (1.0)
|   |   Primeiro = N|D|D: Não (1.0)
|   |   Primeiro = N|D|N: Sim (0.0)
|   |   Primeiro = N|N|A: Não (2.0)
|   |   Primeiro = N|N|D: Sim (0.0)
|   |   Primeiro = N|N|N: Não (5.0)
|   Terceiro = A|A|D: Sim (145.0/12.0)
|   Terceiro = A|A|N: Não (34.0/1.0)
|   Terceiro = A|D|A: Sim (99.0/7.0)
|   Terceiro = A|D|D: Sim (112.0/3.0)
|   Terceiro = A|D|N: Não (6.0)
|   Terceiro = A|N|A: Sim (2.0/1.0)
|   Terceiro = A|N|D: Não (0.0)
|   Terceiro = A|N|N
|   |   jaReprovou = 0: Não (153.0/2.0)
|   |   jaReprovou = 1: Sim (2.0)
|   Terceiro = D|A|A: Sim (114.0/5.0)
|   Terceiro = D|A|D: Sim (90.0/4.0)
|   Terceiro = D|A|N: Não (4.0)
|   Terceiro = D|D|A: Sim (86.0/5.0)
|   Terceiro = D|D|D
```

```
|   |   MediaMatematica = 0: Sim (419.0/18.0)
|   |   MediaMatematica = 1
|   |   |   TemIrmao = 0: Sim (7.0)
|   |   |   TemIrmao = 1
|   |   |   |   CobrouMatricula = 0: Sim (3.0)
|   |   |   |   CobrouMatricula = 1: Não (4.0)
|   |   MediaMatematica = 2: Sim (0.0)
|   |   MediaMatematica = 3: Não (3.0/1.0)
|   |   MediaMatematica = 4: Sim (0.0)
|   |   MediaMatematica = 5: Sim (0.0)
|   |   MediaMatematica = 6: Sim (0.0)
|   Terceiro = D|D|N: Não (16.0)
|   Terceiro = D|N|A: Não (0.0)
|   Terceiro = D|N|D: Não (0.0)
|   Terceiro = D|N|N: Não (48.0)
|   Terceiro = N|A|A: Não (4.0)
|   Terceiro = N|A|D: Não (0.0)
|   Terceiro = N|A|N: Não (11.0)
|   Terceiro = N|D|A: Não (1.0)
|   Terceiro = N|D|D: Não (1.0)
|   Terceiro = N|D|N: Não (17.0)
|   Terceiro = N|N|A: Não (16.0/1.0)
|   Terceiro = N|N|D: Não (9.0)
|   Terceiro = N|N|N: Não (7587.0/18.0)
FaltasBiologia > 3.975091
|   QtdeAlunos <= 22.878625
|   |   Turno = M
|   |   |   QtdeAlunos <= 17.952809: Não (338.0/34.0)
|   |   |   QtdeAlunos > 17.952809
|   |   |   |   OcorrenciasRuins <= 0.486073
|   |   |   |   |   Segundo = A|A|A
|   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Sim
|   |   |   |   |   |   |   TemIrmao = 0
|   |   |   |   |   |   |   |   Terceiro = A|A|A
|   |   |   |   |   |   |   |   |   EstadoCivil = Solteiro
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 0: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 1: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 2: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 3: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 4: Não (9.0/3.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 5: Sim (3.0)
|   |   |   |   |   |   |   |   |   |   MediaMatematica = 6: Não (1.0)
|   |   |   |   |   |   |   |   |   EstadoCivil = Casado
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = <25: Sim (2.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = >=71: Não (1.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = 36-40: Não (48.0/6.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = 51-70: Não (7.0/2.0)
|   |   |   |   |   |   |   |   |   |   FaixaEtaria = 41-50
|   |   |   |   |   |   |   |   |   |   |   FaltasMatematica <= 6.42345: Sim (9.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   FaltasMatematica > 6.42345
|   |   |   |   |   |   |   |   |   |   |   |   Genero = M: Não (35.0/5.0)
|   |   |   |   |   |   |   |   |   |   |   |   Genero = F
|   |   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75:
Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 76-a-
100: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5:
Não (3.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50:
Não (2.0)
```

```
|   |   |   |   |   |   |   |   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25:
Sim (5.0)
|   |   |   |   |   |   |   |   |   |   |   FaixaEtaria = 31-35: Sim (13.0/4.0)
|   |   |   |   |   |   |   |   |   |   |   FaixaEtaria = 26-30: Sim (9.0/3.0)
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Separado: Sim (12.0/2.0)
|   |   |   |   |   |   |   |   |   Terceiro = A|A|D: Não (14.0)
|   |   |   |   |   |   |   |   |   Terceiro = A|A|N: Não (3.0)
|   |   |   |   |   |   |   |   |   Terceiro = A|D|A
|   |   |   |   |   |   |   |   |   |   FaltasMatematica <= 7.230786: Não (10.0/2.0)
|   |   |   |   |   |   |   |   |   |   FaltasMatematica > 7.230786: Sim (4.0)
|   |   |   |   |   |   |   |   |   Terceiro = A|D|D
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Solteiro: Não (0.0)
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Casado: Não (3.0)
|   |   |   |   |   |   |   |   |   |   EstadoCivil = Separado: Sim (2.0)
|   |   |   |   |   |   |   |   |   Terceiro = A|D|N: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = A|N|A: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = A|N|D: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = A|N|N: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = D|A|A
|   |   |   |   |   |   |   |   |   |   MediaNaIdade = Igual: Sim (5.0)
|   |   |   |   |   |   |   |   |   |   MediaNaIdade = Menor: Sim (0.0)
|   |   |   |   |   |   |   |   |   |   MediaNaIdade = Maior: Não (2.0)
|   |   |   |   |   |   |   |   |   Terceiro = D|A|D: Não (3.0)
|   |   |   |   |   |   |   |   |   Terceiro = D|A|N: Não (1.0)
|   |   |   |   |   |   |   |   |   Terceiro = D|D|A: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = D|D|D: Sim (3.0/1.0)
|   |   |   |   |   |   |   |   |   Terceiro = D|D|N: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = D|N|A: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = D|N|D: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = D|N|N: Não (1.0)
|   |   |   |   |   |   |   |   |   Terceiro = N|A|A: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = N|A|D: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = N|A|N: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = N|D|A: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = N|D|D: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = N|D|N: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = N|N|A: Não (0.0)
|   |   |   |   |   |   |   |   |   Terceiro = N|N|D: Não (2.0)
|   |   |   |   |   |   |   |   |   Terceiro = N|N|N: Não (8.0)
|   |   |   |   |   |   |   |   TemIrmao = 1
|   |   |   |   |   |   |   |   |   MediaNaIdade = Igual: Sim (18.0/2.0)
|   |   |   |   |   |   |   |   |   MediaNaIdade = Menor: Sim (0.0)
|   |   |   |   |   |   |   |   |   MediaNaIdade = Maior: Não (5.0/1.0)
|   |   |   |   |   |   |   ResponsavelLegalEhFinanceiro = Não
|   |   |   |   |   |   |   |   AlunoQuantoTempo <= 3: Sim (15.0)
|   |   |   |   |   |   |   |   AlunoQuantoTempo > 3: Não (2.0)
|   |   |   |   |   Segundo = A|A|D
|   |   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75: Não (3.0/1.0)
|   |   |   |   |   |   BolsaEDescontosNormalizado = 76-a-100: Sim (0.0)
|   |   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5: Sim (6.0)
|   |   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50
|   |   |   |   |   |   |   FaltasMatematica <= 8.239094: Não (5.0)
|   |   |   |   |   |   |   FaltasMatematica > 8.239094: Sim (5.0/1.0)
|   |   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25: Sim (12.0/3.0)
|   |   |   |   |   Segundo = A|A|N: Não (3.0)
|   |   |   |   |   Segundo = A|D|A: Sim (23.0/5.0)
|   |   |   |   |   Segundo = A|D|D
|   |   |   |   |   |   QtdeAlunos <= 18.920535: Não (11.0/1.0)
|   |   |   |   |   |   QtdeAlunos > 18.920535
|   |   |   |   |   |   |   BolsaEDescontosNormalizado = 51-a-75: Sim (0.0)
```

```
|   |   |   |   |   |   |   BolsaEDescontosNormalizado = 76-a-100: Sim (0.0)
|   |   |   |   |   |   |   BolsaEDescontosNormalizado = 0-a-5: Sim (11.0/2.0)
|   |   |   |   |   |   |   BolsaEDescontosNormalizado = 26-a-50: Não (2.0)
|   |   |   |   |   |   |   BolsaEDescontosNormalizado = 6-a-25: Sim (4.0)
|   |   |   |   |   Segundo = A|D|N: Não (2.0)
|   |   |   |   |   Segundo = A|N|A: Não (1.0)
|   |   |   |   |   Segundo = A|N|D: Não (3.0)
|   |   |   |   |   Segundo = A|N|N: Não (3.0)
|   |   |   |   |   Segundo = D|A|A
|   |   |   |   |   |   FaixaEtaria = <25: Sim (0.0)
|   |   |   |   |   |   FaixaEtaria = >=71: Sim (0.0)
|   |   |   |   |   |   FaixaEtaria = 36-40: Não (3.0/1.0)
|   |   |   |   |   |   FaixaEtaria = 51-70: Não (3.0)
|   |   |   |   |   |   FaixaEtaria = 41-50: Sim (7.0)
|   |   |   |   |   |   FaixaEtaria = 31-35
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = acima-de-Graduação: Não (2.0)
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = Até-Ensino-Medio: Sim (0.0)
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Completo: Sim (2.0)
|   |   |   |   |   |   |   GrauInstrucaoResponsavel = Superior-Incompleto: Sim (0.0)
|   |   |   |   |   |   FaixaEtaria = 26-30: Não (2.0)
|   |   |   |   |   Segundo = D|A|D
|   |   |   |   |   |   MediaMatematica = 0: Sim (0.0)
|   |   |   |   |   |   MediaMatematica = 1: Sim (3.0)
|   |   |   |   |   |   MediaMatematica = 2: Não (1.0)
|   |   |   |   |   |   MediaMatematica = 3: Não (3.0)
|   |   |   |   |   |   MediaMatematica = 4: Sim (5.0)
|   |   |   |   |   |   MediaMatematica = 5
|   |   |   |   |   |   |   EhAdventista = Sim: Sim (2.0)
|   |   |   |   |   |   |   EhAdventista = Não: Não (3.0)
|   |   |   |   |   |   MediaMatematica = 6: Sim (3.0)
|   |   |   |   |   Segundo = D|A|N: Não (0.0)
|   |   |   |   |   Segundo = D|D|A
|   |   |   |   |   |   AlunoQuantoTempo <= 3.491928: Sim (7.0/1.0)
|   |   |   |   |   |   AlunoQuantoTempo > 3.491928: Não (3.0)
|   |   |   |   |   Segundo = D|D|D
|   |   |   |   |   |   AlunoQuantoTempo <= 3.200947
|   |   |   |   |   |   |   FaltasMatematica <= 6.42345
|   |   |   |   |   |   |   |   AlunoQuantoTempo <= 1.498619: Sim (3.0)
|   |   |   |   |   |   |   |   AlunoQuantoTempo > 1.498619: Não (4.0)
|   |   |   |   |   |   |   FaltasMatematica > 6.42345: Sim (29.0/1.0)
|   |   |   |   |   |   AlunoQuantoTempo > 3.200947
|   |   |   |   |   |   |   QtdeAlunos <= 18.481811: Sim (2.0)
|   |   |   |   |   |   |   QtdeAlunos > 18.481811: Não (24.0/7.0)
|   |   |   |   |   Segundo = D|D|N: Não (1.0)
|   |   |   |   |   Segundo = D|N|A: Não (1.0)
|   |   |   |   |   Segundo = D|N|D: Não (1.0)
|   |   |   |   |   Segundo = D|N|N: Não (4.0)
|   |   |   |   |   Segundo = N|A|A: Não (5.0)
|   |   |   |   |   Segundo = N|A|D: Não (2.0)
|   |   |   |   |   Segundo = N|A|N: Não (0.0)
|   |   |   |   |   Segundo = N|D|A: Não (0.0)
|   |   |   |   |   Segundo = N|D|D: Não (2.0)
|   |   |   |   |   Segundo = N|D|N: Sim (2.0/1.0)
|   |   |   |   |   Segundo = N|N|A: Não (9.0/1.0)
|   |   |   |   |   Segundo = N|N|D: Não (11.0)
|   |   |   |   |   Segundo = N|N|N: Não (103.0)
|   |   |   |   OcorrenciasRuins > 0.486073: Sim (34.0)
|   |   |   Turno = T
|   |   |   |   Primeiro = A|A|A: Sim (245.0/17.0)
|   |   |   |   Primeiro = A|A|D: Sim (39.0/3.0)
```

```
|   |   |   Primeiro = A|A|N: Não (6.0/1.0)
|   |   |   Primeiro = A|D|A: Sim (39.0/6.0)
|   |   |   Primeiro = A|D|D
|   |   |   |   ResponsavelLegalEhFinanceiro = Sim: Sim (36.0/2.0)
|   |   |   |   ResponsavelLegalEhFinanceiro = Não
|   |   |   |   |   Genero = M: Sim (4.0)
|   |   |   |   |   Genero = F: Não (2.0)
|   |   |   Primeiro = A|D|N: Não (2.0)
|   |   |   Primeiro = A|N|A: Não (2.0)
|   |   |   Primeiro = A|N|D: Não (1.0)
|   |   |   Primeiro = A|N|N: Não (6.0/1.0)
|   |   |   Primeiro = D|A|A: Sim (42.0/1.0)
|   |   |   Primeiro = D|A|D: Sim (31.0/1.0)
|   |   |   Primeiro = D|A|N: Não (2.0)
|   |   |   Primeiro = D|D|A: Sim (26.0/1.0)
|   |   |   Primeiro = D|D|D
|   |   |   |   VeioDeForaDoEstado = Sim: Não (3.0/1.0)
|   |   |   |   VeioDeForaDoEstado = Não: Sim (88.0/1.0)
|   |   |   Primeiro = D|D|N: Sim (0.0)
|   |   |   Primeiro = D|N|A: Sim (0.0)
|   |   |   Primeiro = D|N|D: Sim (0.0)
|   |   |   Primeiro = D|N|N: Sim (0.0)
|   |   |   Primeiro = N|A|A: Não (6.0)
|   |   |   Primeiro = N|A|D: Sim (0.0)
|   |   |   Primeiro = N|A|N: Não (3.0)
|   |   |   Primeiro = N|D|A: Não (1.0)
|   |   |   Primeiro = N|D|D: Não (2.0)
|   |   |   Primeiro = N|D|N: Não (1.0)
|   |   |   Primeiro = N|N|A: Não (7.0)
|   |   |   Primeiro = N|N|D: Não (9.0/1.0)
|   |   |   Primeiro = N|N|N: Não (16.0/1.0)
|   |   Turno = N: Sim (0.0)
|   QtdeAlunos > 22.878625
|   |   Terceiro = A|A|A
|   |   |   OcorrenciasRuins <= 0.486073
|   |   |   |   Primeiro = A|A|A
|   |   |   |   |   Segundo = A|A|A: Sim (1387.0/147.0)
|   |   |   |   |   Segundo = A|A|D: Sim (70.0/5.0)
|   |   |   |   |   Segundo = A|A|N: Sim (0.0)
|   |   |   |   |   Segundo = A|D|A: Sim (38.0)
|   |   |   |   |   Segundo = A|D|D: Sim (12.0/1.0)
|   |   |   |   |   Segundo = A|D|N: Sim (0.0)
|   |   |   |   |   Segundo = A|N|A: Sim (0.0)
|   |   |   |   |   Segundo = A|N|D: Sim (0.0)
|   |   |   |   |   Segundo = A|N|N: Sim (0.0)
|   |   |   |   |   Segundo = D|A|A
|   |   |   |   |   |   AlunoQuantoTempo <= 4.070425: Sim (55.0)
|   |   |   |   |   |   AlunoQuantoTempo > 4.070425
|   |   |   |   |   |   |   AlunoQuantoTempo <= 4.682037: Não (2.0)
|   |   |   |   |   |   |   AlunoQuantoTempo > 4.682037: Sim (10.0)
|   |   |   |   |   Segundo = D|A|D: Sim (11.0/2.0)
|   |   |   |   |   Segundo = D|A|N: Sim (0.0)
|   |   |   |   |   Segundo = D|D|A: Sim (8.0/2.0)
|   |   |   |   |   Segundo = D|D|D: Não (7.0/1.0)
|   |   |   |   |   Segundo = D|D|N: Sim (0.0)
|   |   |   |   |   Segundo = D|N|A: Sim (0.0)
|   |   |   |   |   Segundo = D|N|D: Sim (0.0)
|   |   |   |   |   Segundo = D|N|N: Sim (0.0)
|   |   |   |   |   Segundo = N|A|A: Não (3.0)
|   |   |   |   |   Segundo = N|A|D: Sim (0.0)
```

```
|   |   |   |   |   Segundo = N|A|N: Não (1.0)
|   |   |   |   |   Segundo = N|D|A: Sim (0.0)
|   |   |   |   |   Segundo = N|D|D: Sim (0.0)
|   |   |   |   |   Segundo = N|D|N: Sim (0.0)
|   |   |   |   |   Segundo = N|N|A: Sim (0.0)
|   |   |   |   |   Segundo = N|N|D: Sim (0.0)
|   |   |   |   |   Segundo = N|N|N: Não (12.0)
|   |   |   |   Primeiro = A|A|D: Sim (120.0/8.0)
|   |   |   |   Primeiro = A|A|N: Não (9.0/1.0)
|   |   |   |   Primeiro = A|D|A: Sim (116.0/3.0)
|   |   |   |   Primeiro = A|D|D: Sim (51.0/2.0)
|   |   |   |   Primeiro = A|D|N: Sim (0.0)
|   |   |   |   Primeiro = A|N|A: Não (7.0)
|   |   |   |   Primeiro = A|N|D: Não (3.0)
|   |   |   |   Primeiro = A|N|N: Não (6.0)
|   |   |   |   Primeiro = D|A|A: Sim (207.0/17.0)
|   |   |   |   Primeiro = D|A|D: Sim (48.0/6.0)
|   |   |   |   Primeiro = D|A|N: Não (1.0)
|   |   |   |   Primeiro = D|D|A: Sim (47.0/3.0)
|   |   |   |   Primeiro = D|D|D
|   |   |   |   |   Dependentes_Matriculados <= 0.955338
|   |   |   |   |   |   Turno = M: Não (7.0)
|   |   |   |   |   |   Turno = T: Sim (5.0)
|   |   |   |   |   |   Turno = N: Não (0.0)
|   |   |   |   |   Dependentes_Matriculados > 0.955338
|   |   |   |   |   |   Dependentes_Matriculados <= 1.257579: Sim (36.0)
|   |   |   |   |   |   Dependentes_Matriculados > 1.257579
|   |   |   |   |   |   |   Dependentes_Matriculados <= 1.962212: Não (2.0)
|   |   |   |   |   |   |   Dependentes_Matriculados > 1.962212: Sim (8.0)
|   |   |   |   Primeiro = D|D|N: Sim (0.0)
|   |   |   |   Primeiro = D|N|A: Não (3.0)
|   |   |   |   Primeiro = D|N|D: Sim (0.0)
|   |   |   |   Primeiro = D|N|N: Não (1.0)
|   |   |   |   Primeiro = N|A|A: Não (22.0)
|   |   |   |   Primeiro = N|A|D: Não (2.0)
|   |   |   |   Primeiro = N|A|N: Não (2.0)
|   |   |   |   Primeiro = N|D|A: Não (3.0)
|   |   |   |   Primeiro = N|D|D: Sim (0.0)
|   |   |   |   Primeiro = N|D|N: Não (1.0)
|   |   |   |   Primeiro = N|N|A: Não (20.0)
|   |   |   |   Primeiro = N|N|D: Não (6.0)
|   |   |   |   Primeiro = N|N|N: Não (22.0)
|   |   |   OcorrenciasRuins > 0.486073: Sim (1570.0/2.0)
|   |   Terceiro = A|A|D
|   |   |   QtdeAlunos <= 27.987203
|   |   |   |   Turno = M
|   |   |   |   |   FaltasMatematica <= 8.239094
|   |   |   |   |   |   OcorrenciasRuins <= 1
|   |   |   |   |   |   |   Quarto = A|A|A: Não (14.0)
|   |   |   |   |   |   |   Quarto = A|A|D: Sim (3.0/1.0)
|   |   |   |   |   |   |   Quarto = A|A|N: Não (0.0)
|   |   |   |   |   |   |   Quarto = A|D|A: Não (0.0)
|   |   |   |   |   |   |   Quarto = A|D|D: Sim (1.0)
|   |   |   |   |   |   |   Quarto = A|D|N: Não (0.0)
|   |   |   |   |   |   |   Quarto = A|N|A: Não (0.0)
|   |   |   |   |   |   |   Quarto = A|N|D: Não (0.0)
|   |   |   |   |   |   |   Quarto = A|N|N: Não (0.0)
|   |   |   |   |   |   |   Quarto = D|A|A: Sim (4.0)
|   |   |   |   |   |   |   Quarto = D|A|D: Não (2.0)
|   |   |   |   |   |   |   Quarto = D|A|N: Não (0.0)
```

```
|   |   |   |   |   |   |   Quarto = D|D|A: Não (0.0)
|   |   |   |   |   |   |   Quarto = D|D|D: Sim (1.0)
|   |   |   |   |   |   |   Quarto = D|D|N: Não (0.0)
|   |   |   |   |   |   |   Quarto = D|N|A: Não (0.0)
|   |   |   |   |   |   |   Quarto = D|N|D: Não (0.0)
|   |   |   |   |   |   |   Quarto = D|N|N: Não (0.0)
|   |   |   |   |   |   |   Quarto = N|A|A: Não (0.0)
|   |   |   |   |   |   |   Quarto = N|A|D: Não (0.0)
|   |   |   |   |   |   |   Quarto = N|A|N: Não (0.0)
|   |   |   |   |   |   |   Quarto = N|D|A: Não (0.0)
|   |   |   |   |   |   |   Quarto = N|D|D: Não (0.0)
|   |   |   |   |   |   |   Quarto = N|D|N: Não (0.0)
|   |   |   |   |   |   |   Quarto = N|N|A: Não (0.0)
|   |   |   |   |   |   |   Quarto = N|N|D: Não (0.0)
|   |   |   |   |   |   |   Quarto = N|N|N: Não (0.0)
|   |   |   |   |   |   |   OcorrenciasRuins > 1: Sim (5.0)
|   |   |   |   |   |   |   FaltasMatematica > 8.239094: Sim (22.0/2.0)
|   |   |   |   |   Turno = T: Sim (78.0/3.0)
|   |   |   |   |   Turno = N: Sim (0.0)
|   |   |   |   QtdeAlunos > 27.987203: Sim (461.0/10.0)
|   |   |   Terceiro = A|A|N: Não (24.0/4.0)
|   |   |   Terceiro = A|D|A: Sim (382.0/9.0)
|   |   |   Terceiro = A|D|D: Sim (409.0/7.0)
|   |   |   Terceiro = A|D|N: Sim (0.0)
|   |   |   Terceiro = A|N|A: Não (13.0)
|   |   |   Terceiro = A|N|D: Não (4.0)
|   |   |   Terceiro = A|N|N: Não (38.0/5.0)
|   |   |   Terceiro = D|A|A: Sim (451.0/23.0)
|   |   |   Terceiro = D|A|D: Sim (409.0/13.0)
|   |   |   Terceiro = D|A|N: Não (1.0)
|   |   |   Terceiro = D|D|A: Sim (334.0/19.0)
|   |   |   Terceiro = D|D|D: Sim (1861.0/65.0)
|   |   |   Terceiro = D|D|N: Não (1.0)
|   |   |   Terceiro = D|N|A: Não (2.0)
|   |   |   Terceiro = D|N|D: Não (1.0)
|   |   |   Terceiro = D|N|N: Não (13.0)
|   |   |   Terceiro = N|A|A: Não (14.0/2.0)
|   |   |   Terceiro = N|A|D: Não (6.0)
|   |   |   Terceiro = N|A|N: Não (5.0)
|   |   |   Terceiro = N|D|A: Sim (0.0)
|   |   |   Terceiro = N|D|D: Não (4.0)
|   |   |   Terceiro = N|D|N: Não (7.0)
|   |   |   Terceiro = N|N|A
|   |   |   |   VeioDeForaDoEstado = Sim: Sim (3.0/1.0)
|   |   |   |   VeioDeForaDoEstado = Não: Não (13.0)
|   |   |   Terceiro = N|N|D: Não (6.0)
|   |   |   Terceiro = N|N|N
|   |   |   |   OcorrenciasRuins <= 0.875155: Não (577.0/34.0)
|   |   |   |   OcorrenciasRuins > 0.875155
|   |   |   |   |   MediaMatematica = 0
|   |   |   |   |   |   jaReprovou = 0: Não (55.0)
|   |   |   |   |   |   jaReprovou = 1: Sim (3.0)
|   |   |   |   |   MediaMatematica = 1
|   |   |   |   |   |   OcorrenciasMaterial <= 1.689269: Não (17.0/4.0)
|   |   |   |   |   |   OcorrenciasMaterial > 1.689269: Sim (5.0)
|   |   |   |   |   MediaMatematica = 2: Sim (2.0)
|   |   |   |   |   MediaMatematica = 3: Sim (10.0)
|   |   |   |   |   MediaMatematica = 4: Sim (10.0)
|   |   |   |   |   MediaMatematica = 5: Sim (8.0)
|   |   |   |   |   MediaMatematica = 6: Não (0.0)
```

Apêndice C

Estrutura do *Data Warehouse*

Aluno

Column Name	Data Type	Allow Nulls
AlunoId	varchar(150)	☑
Curso	varchar(50)	☑
Dependentes_Matriculados	int	☑
ResponsavelLegalEhFinanc...	varchar(10)	☑
MesInicioParcela	int	☑
TipoAluno	varchar(70)	☑
AlunoQuantoTempo	int	☑
QtdeAlunos	int	☑
Turno	varchar(2)	☑
idadeMedia	int	☑
DataNascimento	date	☑
MediaNaIdade	varchar(50)	☑
Genero	varchar(2)	☑
EhAdventista	varchar(50)	☑
EstadoCivil	varchar(50)	☑
GrauInstrucaoResponsavel	varchar(50)	☑
FaixaEtaria	varchar(50)	☑
Ano	int	☑
OcorrenciasBoas	int	☑
OcorrenciasMaterial	int	☑
OcorrenciasRuins	int	☑
BolsaEDescontos	money	☑
BolsaEDescontosNormaliza...	varchar(50)	☑
TemIrmao	bit	☑
VeioDeForaDoEstado	varchar(50)	☑
CarnesAtraso	int	☑
CarnesEmDia	int	☑
CarnesEmAberto	int	☑
CobrouMatricula	bit	☑
QtdeParcelas	int	☑
jaReprovou	bit	☑
MediaMatematica	money	☑
FaltasMatematica	int	☑
MediaReligiao	money	☑
FaltasReligiao	int	☑
MediaBiologia	money	☑
FaltasBiologia	int	☑
MediaHistoria	money	☑
FaltasHistoria	int	☑
MediaGeografia	money	☑
FaltasGeografia	int	☑
MediaIngles	money	☑
FaltasIngles	int	☑
MediaPortugues	money	☑
FaltasPortugues	int	☑
MediaEdFisica	money	☑
FaltasEdFisica	int	☑
Primeiro	varchar(10)	☑
Segundo	varchar(10)	☑
Terceiro	varchar(10)	☑
Quarto	varchar(10)	☑
Quinto	varchar(10)	☑
Sexto	varchar(10)	☑
Setimo	varchar(10)	☑
Oitavo	varchar(10)	☑
Nono	varchar(10)	☑
Decimo	varchar(10)	☑
TemRematricula	varchar(10)	☑
		☐

Apêndice D

Trecho do arquivo em formato ARFF

```
@relation Completo

@attribute Curso {Fund-1º-Ano,Fund-2º-Ano,Fund-3º-Ano,Fund-4º-Ano,Fund-5º-Ano,Fund-6º-Ano,Fund-7º-Ano,Fund-8º-Ano}
@attribute Dependentes_Matriculados numeric
@attribute ResponsavelLegalEhFinanceiro {Sim,Não}
@attribute MesInicioParcela numeric
@attribute TipoAluno {Regular-Externato,Interno}
@attribute AlunoQuantoTempo numeric
@attribute QtdeAlunos numeric
@attribute Turno {M,T,N}
@attribute MediaNaIdade {Igual,Menor,Maior}
@attribute Genero {M,F}
@attribute EhAdventista {Sim,Não}
@attribute EstadoCivil {Solteiro,Casado,Separado}
@attribute GrauInstrucaoResponsavel {acima-de-Graduação,Até-Ensino-Medio,Superior-Completo,Superior-Incompleto}
@attribute FaixaEtaria {<25,>=71,36-40,51-70,41-50,31-35,26-30}
@attribute Ano {2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014}
@attribute OcorrenciasBoas numeric
@attribute OcorrenciasMaterial numeric
@attribute OcorrenciasRuins numeric
@attribute BolsaEDescontosNormalizado {51-a-75,76-a-100,0-a-5,26-a-50,6-a-25}
@attribute TemIrmao {0,1}
@attribute VeioDeForaDoEstado {Sim,Não}
@attribute CobrouMatricula {0,1}
@attribute QtdeParcelas numeric
@attribute jaReprovou {0,1}
@attribute MediaMatematica {0,1,2,3,4,5,6}
@attribute FaltasMatematica numeric
@attribute MediaReligiao {0,1,2,3,4,5,6}
@attribute FaltasReligiao numeric
@attribute MediaBiologia {0,1,2,3,4,5,6}
@attribute FaltasBiologia numeric
@attribute MediaHistoria {0,1,2,3,4,5,6}
@attribute FaltasHistoria numeric
@attribute MediaGeografia {0,1,2,3,4,5,6}
@attribute FaltasGeografia numeric
@attribute MediaIngles {0,1,2,3,4,5,6}
@attribute FaltasIngles numeric
@attribute MediaPortugues {0,1,2,3,4,5,6}
@attribute FaltasPortugues numeric
@attribute MediaEdFisica {0,1,2,3,4,5,6}
@attribute FaltasEdFisica numeric
@attribute Primeiro {A|A|A,A|A|D,A|A|N,A|D|A,A|D|D,A|D|N,A|N|A,A|N|D,A|N|N,D|A|A,D|A|D,D|A|N,D|D|A,D|D|D,D|D|N,D|N|A,D|N|D,D|N|N,N|A|A,N|A|D,N|A|N,N|D|A,N|D|D,N|D|N,N|N|A,N|N|D,N|N|N}
@attribute Segundo {A|A|A,A|A|D,A|A|N,A|D|A,A|D|D,A|D|N,A|N|A,A|N|D,A|N|N,D|A|A,D|A|D,D|A|N,D|D|A,D|D|D,D|D|N,D|N|A,D|N|D,D|N|N,N|A|A,N|A|D,N|A|N,N|D|A,N|D|D,N|D|N,N|N|A,N|N|D,N|N|N}
```

```
@attribute Terceiro
{A|A|A,A|A|D,A|A|N,A|D|A,A|D|D,A|D|N,A|N|A,A|N|D,A|N|N,D|A|A,D|A|D,D|A|N,D|D|A,D|D|D,D|D
|N,D|N|A,D|N|D,D|N|N,N|A|A,N|A|D,N|A|N,N|D|A,N|D|D,N|D|N,N|N|A,N|N|D,N|N|N}
@attribute Quarto
{A|A|A,A|A|D,A|A|N,A|D|A,A|D|D,A|D|N,A|N|A,A|N|D,A|N|N,D|A|A,D|A|D,D|A|N,D|D|A,D|D|D,D|D
|N,D|N|A,D|N|D,D|N|N,N|A|A,N|A|D,N|A|N,N|D|A,N|D|D,N|D|N,N|N|A,N|N|D,N|N|N}
@attribute TemRematricula {Sim,Não}

@data
Fund-5º-Ano,1,Sim,1,Regular-Externato,3,39,M,Igual,M,Não,Casado,Superior-Completo,41-
50,2001,0,0,0,6-a-
25,0,Sim,0,12,0,4,7,4,7,4,7,4,7,4,7,4,7,4,7,0,0,D|D|A,D|A|D,D|D|D,D|A|A,Sim
Fund-6º-Ano,1,Sim,1,Regular-Externato,4,30,M,Igual,M,Não,Casado,Superior-Completo,41-
50,2002,0,0,3,6-a-
25,0,Sim,0,12,0,5,15,5,15,5,15,5,15,5,15,5,15,5,15,0,0,A|D|A,A|D|A,D|D|D,A|D|A,Sim
```