INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO

“Sistema de Mineração de Dados para Apoiar a Tomada de Decisão em uma Instituição de Ensino Superior: o problema da evasão escolar no IFTM”

Eduardo de Oliveira Araujo

Dissertação de Mestrado

Mestrado em Assessoria de Administração

Porto – 2018

INSTITUTO SUPERIOR DE CONTABILIDADE E ADMINISTRAÇÃO DO PORTO
INSTITUTO POLITÉCNICO DO PORTO

INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO

“Sistema de Mineração de Dados para Apoiar a Tomada de Decisão em uma Instituição de Ensino Superior: o problema da evasão escolar no IFTM”

Eduardo de Oliveira Araujo

Dissertação de mestrado
apresentada ao Instituto Superior de Contabilidade e Administração do Porto para obtenção do grau de Mestre em Assessoria de Administração, sob orientação da Profª. Doutora Ana Azevedo

Porto – 2018

INSTITUTO SUPERIOR DE CONTABILIDADE E ADMINISTRAÇÃO DO PORTO
INSTITUTO POLITÉCNICO DO PORTO

Resumo

Este trabalho visa buscar soluções para a tomada de decisão a respeito da problemática da evasão escolar no âmbito dos cursos de graduação do Instituto Federal do Triângulo Mineiro (IFTM) no período de 2012 a 2016. O objetivo principal deste estudo teve como foco analisar se é possível descobrir a causa da evasão escolar através da mineração de dados na base de dados do Sistema de Controle Acadêmico do IFTM. Para isso, utilizamos a metodologia CRISP-DM, que permite implementar o processo *Knowledge Discovery in Databases (KDD)* para descoberta de conhecimento em base de dados e o programa WEKA (*Waikato Environment for Knowledge Analysis*) para fazer a mineração de dados e encontrar tendências e padrões. Os resultados deste trabalho foram adquiridos através da metodologia investigação-ação, que serviu de suporte à investigação. Conforme identificado durante a etapa de avaliação dos modelos, a maioria dos casos de evasão ocorre quando a frequência das aulas é baixa e a quantidade de disciplinas reprovadas por infrequência é alta, ou seja, a variável que totaliza a frequência está plenamente correlacionada com o sucesso do aluno em ser aprovado nas disciplinas. Dessa forma, o perfil do aluno com maiores e prováveis chances de abandono do curso é o aluno faltoso com alta reprovação por infrequência. De acordo com os resultados obtidos, comprova-se que o investigador promoveu alterações no ambiente organizacional ao introduzir, pela primeira vez, a participação direta das TICs nesta problemática e ao contribuir com um conjunto de ações contendo propostas de melhorias para a gestão administrativa reduzir a evasão escolar.

Palavras-chave: Tecnologia da Informação - Administração; Mineração de dados; Evasão escolar; Processo decisório.

Abstract

This work aims to seek solutions for decision-making regarding the problem of school evasion under the graduation courses of the Instituto Federal do Triângulo Mineiro (IFTM) in the period from 2012 to 2016. The main objective of this study was to analyze whether it is possible to discover the cause of school evasion through data mining in the database of the IFTM academic control system. For this, we use the CRISP-DM methodology, which enables the Knowledge Discovery in Databases (KDD) process to discover the knowledge base and the WEKA program (Waikato Environment for Knowledge Analysis) to do the data mining and find trends and patterns. The results of this work were acquired through the action-research methodology, which served as support for research. The evidence during the assessment phase of the models, most of the cases of evasion occur when the frequency is low and the number of disciplines reproved by infrequency is high, that is, a variable that totals the frequency is fully correlated with the success of the student to be approved in the disciplines. Thus, the profile of the student with the highest and most probable chances of taking the course is the missing student with high reproach for infrequency. According to the results obtained, it is verified that the researcher promoted changes in the organizational environment by introducing, for the first time, the direct participation of ICTs in this problem and by contributing with a set of actions containing proposals for improvements to the administrative management to reduce school dropout.

Key words: Information Technology - Administration; Data mining; School dropout; Decision-making process.

Agradecimentos

Agradeço a Deus pelo dom da vida, pelos desafios e oportunidades que possibilitam o meu crescimento pessoal e profissional, por estar sempre presente nos caminhos que percorro.

Aos meus pais, José e Odília, pelas minha primeira e principal educação que recebi na minha vida, meus mestres eternos.

Aos meus irmãos e irmãs, pelo apoio incondicional e por acreditarem nos meus sonhos.

À minha noiva Poliana, pela compreensão nos momentos ausentes, pelo companheirismo, carinho e amor sempre presentes.

Ao amigo e coordenador da equipe de desenvolvimento de software do IFTM, João Batista, pelos momentos de consultoria e ajuda com o entendimento do Sistema de Controle Acadêmico.

A toda minha família, meu refúgio e abrigo nos momentos mais difíceis da minha vida.

À orientadora Professora Doutora Ana Azevedo pelos ensinamentos, paciência, dedicação e colaboração fundamental para a conclusão do meu trabalho e obtenção do grau de Mestre em Assessoria de Administração.

Aos amigos e colegas de mestrado, pela amizade, por caminharmos juntos nessa jornada, pelo companheirismo, pelas experiências compartilhadas.

A todos os professores de Assessoria de Administração pelos ensinamentos, pela dedicação ao seu trabalho, pelo afeto e carinho a nós alunos.

A todas aquelas pessoas que torceram por mim, de forma direta ou indireta, mas que de alguma forma contribuíram para a concretização de mais um sonho.

“Educar é semear com sabedoria e colher com paciência.”
Augusto Cury

Lista de Abreviaturas

ADS - Análise e Desenvolvimento de Sistemas

BD - Banco de Dados

CC - Coordenação de Curso

CRA - Módulo Controle de Registro Acadêmico

CRCA - Coordenação de Registro e Controle Acadêmico

CRISP-DM (*CRoss Industry Standard Process for Data Mining*) - Processo Padrão Inter-Indústrias para Mineração de Dados

DM (*Data Mining*) - Mineração de Dados

DTIC - Diretoria de Tecnologia da Informação e Comunicação

IA - Investigação-Ação

IFTM - Instituto Federal do Triangulo Mineiro

EaD - Educação a Distância

ENADE - Exame Nacional de Desempenho dos Estudantes

ERP (*Enterprise Resource Planning*) - Planejamento dos Recursos da Empresa

GC - Módulo Gestor de Curso

KDD (*Knowledge Discovery in Databases*) - Descoberta de Conhecimento em Base de Dados

MAC - Módulo Acadêmico

PDI - Plano de desenvolvimento Institucional

PDTIC - Plano Diretor de Tecnologia da Informação e Comunicação

PPC - Projeto Pedagógico do Curso

PPI - Projeto Pedagógico Institucional

PROEN - Pró-reitora de Ensino

PROEXT - Pró-reitora de Extensão

SCA - Sistema de Controle Acadêmico

SEMMA (*Sample, Explore, Modify, Model, Assess*) - Amostragem, Exploração, Modificação, Modelação, Avaliação

SGBD - Sistema Gerenciador de Banco de Dados

SISTEC - Sistema Nacional de Informações da Educação Profissional e Tecnológica

SQL (*Structured Query Language*) - Linguagem de Consulta Estruturada

TI - Tecnologia da Informação

TIC - Tecnologia da Informação e Comunicação

VIRTUAL IF - Ambiente de intranet dos serviços virtuais do IFTM

WEKA - *Waikato Environment for Knowledge Analysis*

Índice Geral

Índice de Tabelas

Índice de Figuras

Introdução

Existe uma enorme quantidade de dados sendo produzida diariamente pelas empresas e instituições nas suas transações de negócios e nos seus processos administrativos. Esses dados podem ser gravados e recuperados nos formatos de textos, imagens, áudios, vídeos ou dados estruturados. De acordo com Hurwitz, Nugent, Halper, & Kaufman (2015), um dos grandes desafios de toda empresa, independentemente de seu tamanho ou do setor em que atua, tem sido a administração e análise de seus dados. A utilização de ferramentas adequadas para analisar essa grande quantidade de dados é importante para descobrir novas informações e produzir novos conhecimentos, que não estavam acessíveis. A mineração de dados pode ser considerada uma dessas ferramentas.

Mineração de dados é "o uso de técnicas automáticas de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que, devido ao volume de dados, não seriam facilmente descobertas a olho nu pelo ser humano" (Carvalho, 2005, p. 3). Dentro dos conceitos gerais e de forma muito similar e complementar, "*Data mining*, ou mineração de dados, é o processo de descoberta de padrões e tendências existentes em repositórios de dados. Esse processo visa basicamente à análise de grandes quantidades de dados com o objetivo principal de descoberta de conhecimento" (Pinheiro, 2008, p. 97).

O Instituto Federal do Triângulo Mineiro (IFTM) possui um armazém de dados com as suas informações armazenadas de forma bem estruturadas e gerenciadas por um sistema proprietário ERP (*Enterprise Resource Planning* - Planejamento dos Recursos da Empresa), daqui em diante referido como ERP-IFTM.

O ERP-IFTM reúne todos os softwares que integram os dados e processos da Reitoria e dos seus *campi*, inclusive os dados do Sistema de Controle Acadêmico (SCA), essencial para a pesquisa deste trabalho. Porém, com exceção dos Indicadores do IFTM, não existem ferramentas específicas para extrair informações que auxiliem a tomada de decisão.

Os Indicadores do IFTM, acessíveis a todos públicos, interno e externo, pelo endereço virtual *indicadores.iftm.edu.br*, extrai do ERP-IFTM informações tais como quantitativos e percentuais de: a) técnicos administrativos: agrupados por sexo, faixa etária, cargo, classe da carreira, titulação e campus; b) docentes: agrupados por tipo (efetivo ou temporário), titulação, jornada de trabalho e campus; c) acadêmicos (alunos): agrupados por nível de educação, campus e movimento, que é a situação que o aluno se encontra como matriculado, formado ou desistente.

Pelo fato de fornecer análises somente descritivas e não preditivas, é perceptível que as informações extraídas dos indicadores não são suficientes para auxiliar a tomada de decisão no nível gerencial, de modo a estimular a permanência de alunos em todas as vagas ofertadas, ou seja, evitar problemas de evasão.

O Relatório de Gestão 2016 do Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro relata que:

> Parte dos problemas da evasão e da retenção podem ser explicados pelas dificuldades apresentadas pelos estudantes ingressantes, em geral oriundos da Educação Básica, os quais manifestam deficiências em aspectos (habilidades, competências) considerados essenciais para o sucesso escolar, assim como por problemas de outras naturezas, em geral decorrentes da desigualdade social, presente na realidade brasileira. (IFTM, 2017, p. 23)

Ainda de acordo com esse relatório de gestão e com o intuito de estimular a permanência dos alunos e evitar a evasão, o IFTM tem implantado diversos projetos e medidas institucionais tais como:

- concessão de bolsas acadêmicas;
- assistência e auxílio estudantil;
- monitorias para reduzir deficiências de aprendizagem;
- atualização de projetos pedagógicos;
- realização de eventos para a formação acadêmica em geral;
- acompanhamento pedagógico;
- projetos institucionais para resolver os problemas da evasão e da retenção escolar;
- projetos institucionais para compreensão das principais características do mercado, das demandas e dos arranjos produtivos.

Acresce que, de acordo com o Informativo IFTM em Ação (IFTM, 2013)[1], durante a realização do evento "III Fórum Internacional sobre Educação Profissional e Evasão Escolar", ocorrido em Belo Horizonte - Brasil, onde estavam presentes os representantes da Pró-reitora de Ensino (PROEN) e dos *Campi* Uberaba e Uberlândia do IFTM, identificou-se que a evasão escolar é um problema que atinge praticamente todas as Instituições Superiores de Ensino do Brasil, da América Latina e da Europa do Sul. Verificou-se também que são necessárias iniciativas estratégicas que promovam a permanência e o sucesso escolar dos estudantes. Nesse sentido, o Plano de Desenvolvimento Institucional (PDI), que constitui em um documento norteador de ações para o planejamento e desenvolvimento institucional, desenvolveu o seu Projeto Pedagógico Institucional (PPI) para o período 2014 a 2018. Entre as diversas metas e ações que regulam este plano acadêmico está o compromisso de promover a formação integral de seus educandos, investindo recursos em ensino, pesquisa e extensão. De acordo com o Plano de Desenvolvimento Institucional (IFTM, 2014), incluso nos

[1] O Informativo IFTM em Ação é um periódico que tem como foco divulgar as atividades desenvolvidas no IFTM relacionadas a ensino, pesquisa, extensão, projetos, eventos, gestão e relação com a sociedade.

objetivos estratégicos, dentro da perspectiva do aluno, destaca-se: “Reduzir as taxas de evasão e retenção de alunos”, conforme tabela 1.

Tabela 1. PDI: Objetivo estratégico nº 4 da perspectiva do aluno.

Objetivo 4 – Reduzir as taxas de evasão e retenção de alunos.				
Meta 1:	Reduzir o nível de evasão para 15% até 2018.			
Indicador:	Índice de evasão escolar			
Responsável:	Pró-reitora de Ensino		**Tipo:** Desdobrável	
Ano 2014	**Ano 2015**	**Ano 2016**	**Ano 2017**	**Ano 2018**
30%	25%	20%	17,5%	15%

Fonte: Plano de Desenvolvimento Institucional, IFTM (2014, p. 30).

A tabela 1 estabelece o índice percentual a ser alcançado com a redução das taxas de evasão escolar para todos os *campi* nos anos de 2014 até 2018. As ações de combate à evasão até aqui desenvolvidas, tanto as relacionadas à dimensão individual do estudante quanto as relativas à dimensão institucional, não estão sendo eficazes para atingir esta meta. O último relatório de gestão disponível, referente ao ano 2017, indica que apenas um terço dos *campi* está conseguindo reduzir o índice de evasão escolar de forma satisfatória.

Portanto, pode-se afirmar que este trabalho com a mineração de dados é muito importante para determinar um novo conhecimento sobre este aspecto.

De acordo com as atribuições da área de Tecnologia da Informação e Comunicação (IFTM, 2017), dentre as diversas competências da Diretoria de Tecnologia da Informação e Comunicação do IFTM (DTIC) estão: identificar novas necessidades do IFTM quanto à Tecnologia da Informação e Comunicação (TIC) e planejar o desenvolvimento de projetos para o atendimento dessas necessidades em consonância com o Plano Diretor de Tecnologia da Informação e Comunicação (PDTIC); planejar e manter, em conjunto com as áreas correlatas, o PDTIC, em consonância com o PDI.

Dessa forma, a participação das TICs na solução do problema supracitado é requerida. Principalmente, porque ainda não há nenhuma pesquisa sobre a evasão escolar no IFTM que envolva uma metodologia que possibilite a implantação do processo para realizar a mineração de dados.

Sendo assim, acreditamos que o estudo das bases de dados que coletam informações dos alunos, desde o seu ingresso e durante toda a sua vida acadêmica, através de recursos tecnológicos e da mineração de dados tem um papel fundamental na investigação do problema da evasão escolar e pode auxiliar a administração do IFTM nos processos decisórios que promovem o sucesso escolar dos estudantes.

Depois de atendido o pedido de autorização solicitado aos dirigentes do IFTM para procedermos ao estudo na base de dados do SCA do Virtual IF com o compromisso de utilizar as informações obtidas apenas para fins de pesquisa de mestrado, buscamos realizar uma pesquisa qualitativa a fim de obter respostas à problemática da evasão escolar.

O objetivo geral desta pesquisa é definir um conjunto de procedimentos que poderão auxiliar a tomada de decisão e como a mineração de dados vai apoiar a gestão administrativa com relação ao problema da evasão escolar nos cursos superiores e presenciais do IFTM.

A partir do objetivo mais abrangente, delimitamos como objetivos específicos:

1. Entender o funcionamento do SCA, indicando as ações que alimentam a base de dados.
2. Definir as movimentações na base de dados que indicam a ocorrência de evasão.
3. Apresentar o entendimento dos dados dos módulos do SCA, avaliando a sua qualidade e indicando os seus aspectos mais relevantes.
4. Preparar os dados do SCA para serem utilizados com os algoritmos de mineração de dados.
5. Criar modelos de classificação e segmentação a partir dos dados coletados.
6. Avaliar os modelos e identificar as barreiras que dificultam o sucesso escolar do aluno e discutir os resultados.
7. Propor estratégias para amenizar a ocorrência de desistência do curso por parte dos alunos do IFTM.

Definimos a metodologia de pesquisa Investigação-Ação (IA) para o desenvolvimento deste projeto pelas suas características, por se tratar de uma metodologia de investigação desenvolvida através da ação, que permite uma análise crítica com a finalidade de investigar e dar respostas ao problema da evasão escolar, que consiste no ato de um aluno abandonar ou ser desligado do seu curso. Inicialmente, serão relacionados os sistemas informatizados do ERP-IFTM que envolvem a vida acadêmica do aluno, com o intuito de descobrir quais ferramentas computacionais armazenam e recuperam dados relevantes destes alunos no SCA. Estes dados serão investigados aplicando-se a metodologia de mineração de dados CRISP-DM na base de dados real do SCA. Por fim, este trabalho será concluído com a análise, a discussão dos resultados, onde se apresentam ações que poderão auxiliar os gestores da instituição na tomada de decisões quanto à problemática da evasão escolar.

O presente estudo foi estruturado em quatro capítulos, descrevendo-se a seguir a estrutura para o resto da dissertação:

O capítulo 1 expõe, através da revisão da literatura, os fundamentos teóricos que fazem parte da mineração de dados e que são necessários para a realização desta pesquisa: definições, metodologias, aplicações e escolha da ferramenta de mineração de dados.

No capítulo 2 apresentam-se o estudo empírico que aborda a IA como metodologia de investigação e o IFTM dentro do contexto da investigação.

No capítulo 3 é implementada a metodologia CRISP-DM com o desenvolvimento das suas seis fases no SCA do IFTM para realizar o processo de mineração de dados em busca do perfil do aluno que evade do seu curso. Na implementação, fase 6 dessa metodologia, é apresentada uma lista contendo um conjunto de ações, que foi construída pelo pesquisador após a análise, a avaliação dos modelos e as interpretações dos resultados, compondo-se de 18 propostas de melhorias para a gestão administrativa no ensino do IFTM.

No capítulo 4 a discussão dos resultados obtidos através da pesquisa que foi realizada.

Por fim, apresenta-se a Conclusão.

Capítulo I – Revisão da Literatura

1 Mineração de Dados

A correta manipulação do grande volume de dados diário que é produzido pelas empresas e instituições é muito importante para que não se perca nenhuma informação significativa, que pode auxiliar os processos administrativos e a tomada de decisões. É neste contexto que surge a necessidade de se investir no trabalho com a utilização das ferramentas de mineração de dados, que podem auxiliar o descobrimento de uma relação entre os dados analisados e produzir uma nova informação.

1.1 Definição de DM e KDD

A Mineração de Dados (Data Mining – DM) é uma parte integral da Descoberta do Conhecimento em Banco de Dados (*Knowledge Discovery in Databases*, KDD), que são dados brutos tratados e transformados em informações úteis. O termo KDD é definido por (Fayyad, Piatetsky-Shapiro, & Smyth, 1996) como sendo todo o processo de extração não trivial de descoberta de conhecimento novo e útil de um determinado conjunto de dados, ou seja, é o processo que busca a descoberta de padrões úteis nos dados armazenados na empresa, que ainda não haviam sido revelados, por meio de técnicas e ferramentas de exploração e análise, que trará algum benefício para a empresa. Essas cinco fases citadas pelos autores serão explicadas no próximo item "Metodologias de Mineração de Dados".

Para Fayyad et al. (1996), o KDD tem 5 (cinco) fases, conforme definido na figura 1. De acordo com os autores, o processo de KDD é interativo quando permite a interferência e controle do fluxo das atividades por parte do usuário e iterativo por seguir uma sucessão de ações de forma sequencial e correlata.

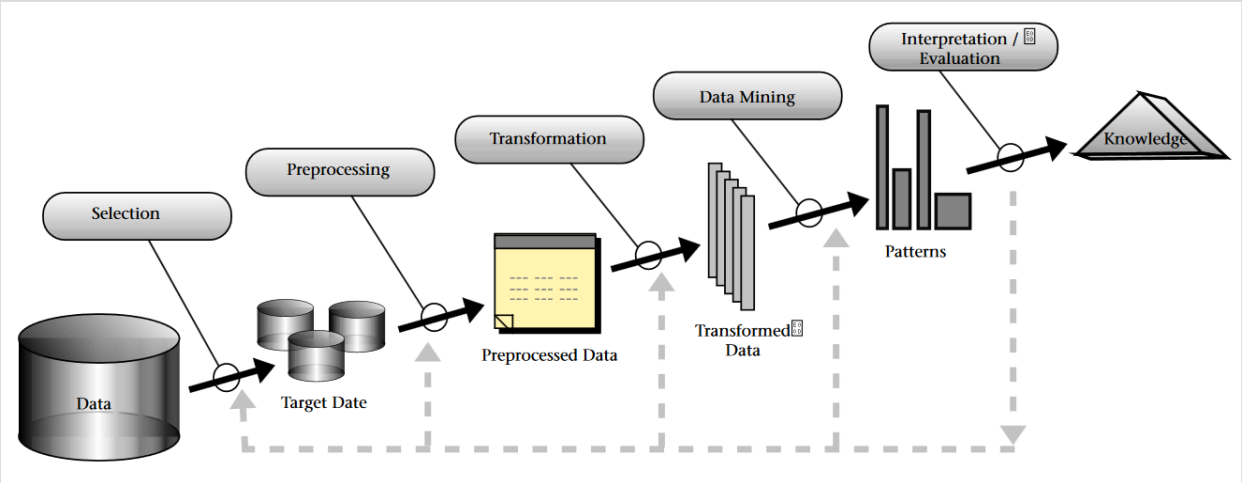

Figura 1. Visão geral das etapas que compõem o processo de KDD.

Fonte: Fayyad et al. (1996).

As cinco fases do KDD são assim definidas:

1. Seleção dos dados (*Selection*):

Na primeira fase do processo de KDD são selecionados e agrupados os conjuntos de dados que são alvo da investigação. Esse agrupamento de dados contém todos os atributos e registros que serão analisados no processo de descoberta de conhecimento, o que torna essa etapa um passo importantíssimo para a qualidade do resultado que será alcançado no final do processo.

2. Pré-processamento e Limpeza dos Dados (*Preprocessing*):

Nessa fase são eliminados os dados inconsistentes e os que são discrepantes, que representam algum erro de observação no conjunto que está sendo avaliado, ou seja, são excluídos os dados que, além de não contribuírem para a investigação, podem atrapalhar todo o processo. Desta forma, neste estágio ocorre a redução da quantidade de atributos que foram selecionados na primeira etapa com o objetivo de favorecer a execução e desempenho dos algoritmos de mineração.

Ao encontrar atributos com dados ausentes pode-se: completá-los através de técnicas de imputação (previsão) ou pela média aritmética do atributo ou apagar todo o registro que contém o atributo vazio.

Se houver necessidade, é possível criar um novo atributo, que será derivado da relação que se estabelece com outros atributos, por exemplo: a idade do aluno pode ser encontrada a partir da data do seu nascimento.

3. Transformação dos Dados (*Transformation*):

Depois de passarem pelas duas primeiras fases do processo de KDD, ou seja, logo após os dados serem selecionados, limpos e pré-processados, o próximo passo do ciclo é transformar e armazenar os atributos em um conjunto de dados apropriado para a utilização dos algoritmos de mineração de dados.

4. Mineração de Dados (*Data Mining*):

Dando continuidade às fases do processo de KDD, tem-se a seguir a mineração de dados propriamente dita, que se inicia com a definição dos algoritmos que serão aplicados para procurar e descobrir novos padrões e regras nos dados. Nesta etapa é necessário indicar os melhores métodos e técnicas especializadas para realizar a mineração de dados que, supostamente, terão mais sucesso para alcançarmos o objetivo do processo de KDD. As principais tarefas de mineração de dados são as seguintes:

a) classificação – por meio da análise de um conjunto de registros disponibilizado, tem o objetivo de “aprender” a reconhecer a qual classe pertence um novo registro.

b) segmentação – é todo o conjunto de dados subdividido em conjuntos menores, conhecidos também como *clusters*, com a intenção de que sejam o mais heterogêneo possível

entre si e possibilitar, a partir desse ponto, pressupor algum resultado ao determinar padrões ou criar novos agrupamentos para análise.

c) associação – tem por objetivo determinar quais variáveis estão relacionadas, ou seja, que ocorrem simultaneamente no mesmo evento. Manifesta-se pela condição "SE variável X, ENTÃO variável Y", segue um exemplo hipotético: um aluno que se matricula na disciplina X, em N% de vezes, também se matricula na disciplina Y.

5. Interpretação e Avaliação dos Resultados (*Interpretation / Evaluation*):

Na última etapa do processo de KDD é necessário interpretar e avaliar o conhecimento descoberto por meio da mineração de dados. É preciso saber se a indagação inicial foi respondida com os resultados obtidos, ou seja, se o objetivo final foi alcançado de forma satisfatória. Em caso negativo, pode-se voltar a qualquer uma das etapas anteriores, escolher um novo algoritmo de mineração de dados ou, se for necessário, alterar o conjunto de dados inicial.

Com o intuito de diferenciar os termos Mineração de Dados e KDD, ainda conforme Fayyad et al. (1996), Mineração de Dados é apenas uma das etapas do KDD na qual se utilizam algoritmos específicos para extração de modelos de dados, e KDD, por sua vez, é o processo completo que busca a extração de conhecimento dos dados.

Segundo Amaral (2016) a mineração de dados é conceituada como sendo um conjunto de processos que tem por objetivo explorar e analisar os dados de históricos de eventos anteriores, através de algoritmos capazes de identificar padrões e associações, com o objetivo de transformar dados em informação e conhecimento de uma forma mais eficaz que outra técnica seria capaz de produzir.

De acordo com Pinheiro (2008, p. 97), "os processos de mineração de dados focam na aplicação de técnicas estatísticas e de inteligência artificial para análise interativa de dados, visando à identificação de padrões de comportamento, tendências e predição".

Carvalho (2005, p. 13), define 5 (cinco) fases para a mineração de dados na empresa (figura 2):

1. Identificação de um problema ou definição de um objetivo a ser alcançado;
2. Descoberta de novas relações por técnicas de mineração de dados;
3. Análise humana das novas relações descobertas;
4. Uso racional das novas relações descobertas;
5. Avaliação dos resultados.

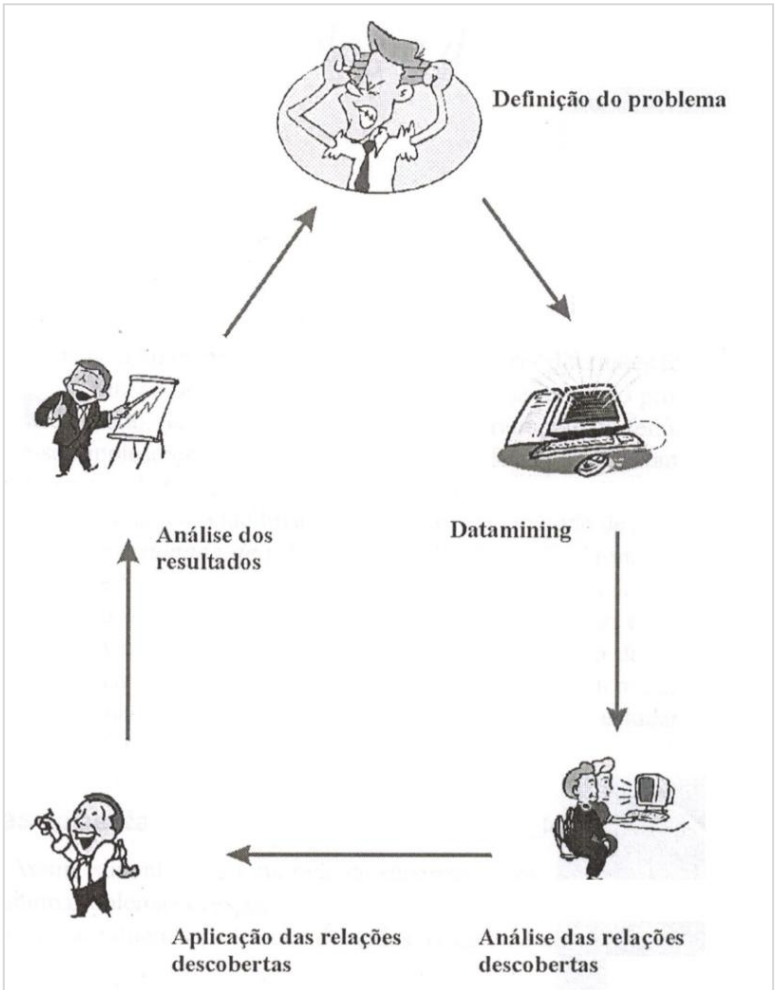

Figura 2. As fases para a mineração de dados na empresa.

Fonte: Carvalho (2005).

Pinheiro (2008), define 6 (seis) etapas do processo de mineração de dados (figura 3):

1. Entendimento do negócio: na primeira etapa é necessário entender o problema e buscar compreender os objetivos do projeto e suas necessidades. É preciso obter uma definição clara do que se está investigando para determinar um plano preliminar das ações a serem executadas.

2. Entendimento dos dados: esta etapa está relacionada com a extração dos dados para se criar uma nova base de dados que será objeto de investigação.

3. Preparação de dados: nessa etapa é feito um pré-processamento onde ocorre a limpeza dos dados, identificando anomalias como dados ausentes ou inconsistentes.

4. Modelagem: são selecionados os dados que podem influenciar nos resultados do modelo a ser construído. Nessa etapa os dados são reduzidos ao se eliminarem as informações confusas e variáveis indiferentes ao modelo.

5. Avaliação: na quinta etapa identifica-se a melhor técnica a ser aplicada e a abordagem para a aplicação dos modelos. Os objetivos do descobrimento de conhecimento

podem ser de verificação, que trabalha com hipóteses formuladas pelos usuários, ou de descoberta, na qual o sistema procura descobrir novos padrões de forma autônoma.

6. Resultados: na interpretação dos resultados, sexta e última etapa, são apresentadas as descobertas obtidas, discutida a melhor forma de utilizá-las na tomada de decisões, definidas as vantagens e desvantagens do modelo e reavaliação de todo o processo.

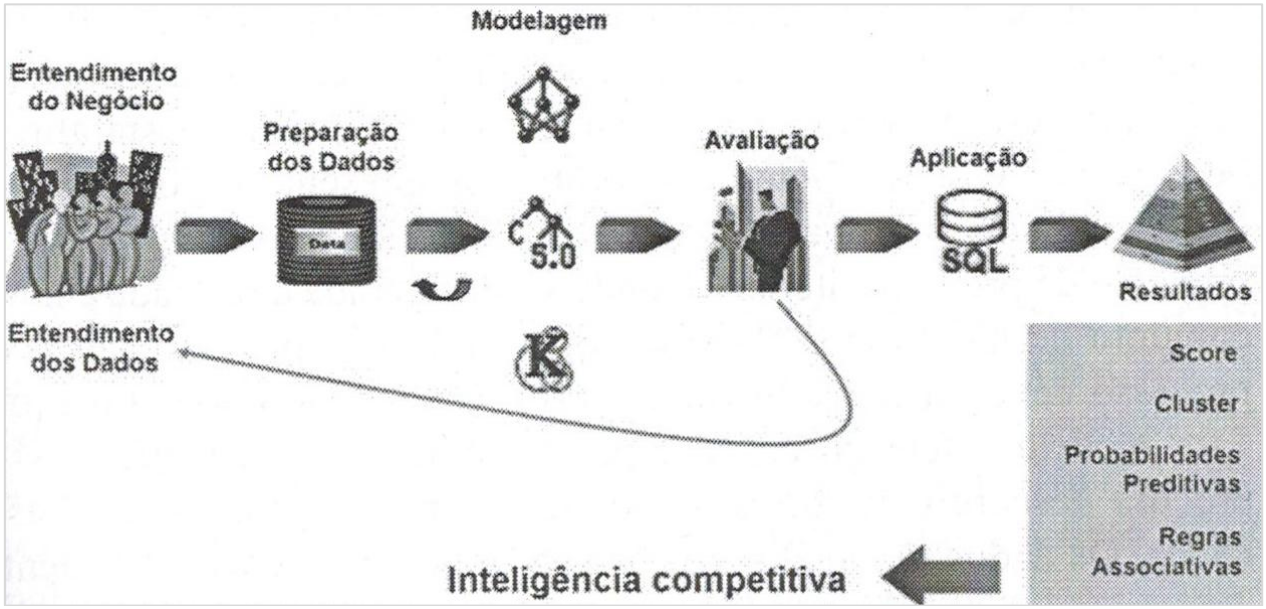

Figura 3. Etapas do processo de mineração de dados.

Fonte: Pinheiro (2008).

1.2 Metodologias de mineração de dados

As metodologias de mineração de dados têm o propósito de obter conhecimento útil para agilizar, melhorar a qualidade e a eficiência no processo de tomada de decisões estratégicas por parte dos gestores.

Sendo assim, os resultados obtidos através da investigação nos dados para reconhecer novos padrões e tendências auxiliará o processo decisório das empresas ao transformar dados em informações valiosas.

De acordo com uma enquete realizada por Piatetsky (2014), apesar do crescente aumento, em 2014, de pessoas que utilizam sua própria metodologia (27,5%, antes 19% em 2007), a metodologia CRISP-DM continua sendo a mais usual para projetos de análise, mineração de dados e ciência dos dados, com 43% de participação dos 200 respondentes, mantendo essencialmente a mesma porcentagem desde 2007 (42%), conforme figura 4. Vemos ainda que o número de pessoas que utilizam a sua própria metodologia cresceu de 19% para 27,5% entre 2007 e 2014. A terceira metodologia mais utilizada é a SEMMA.

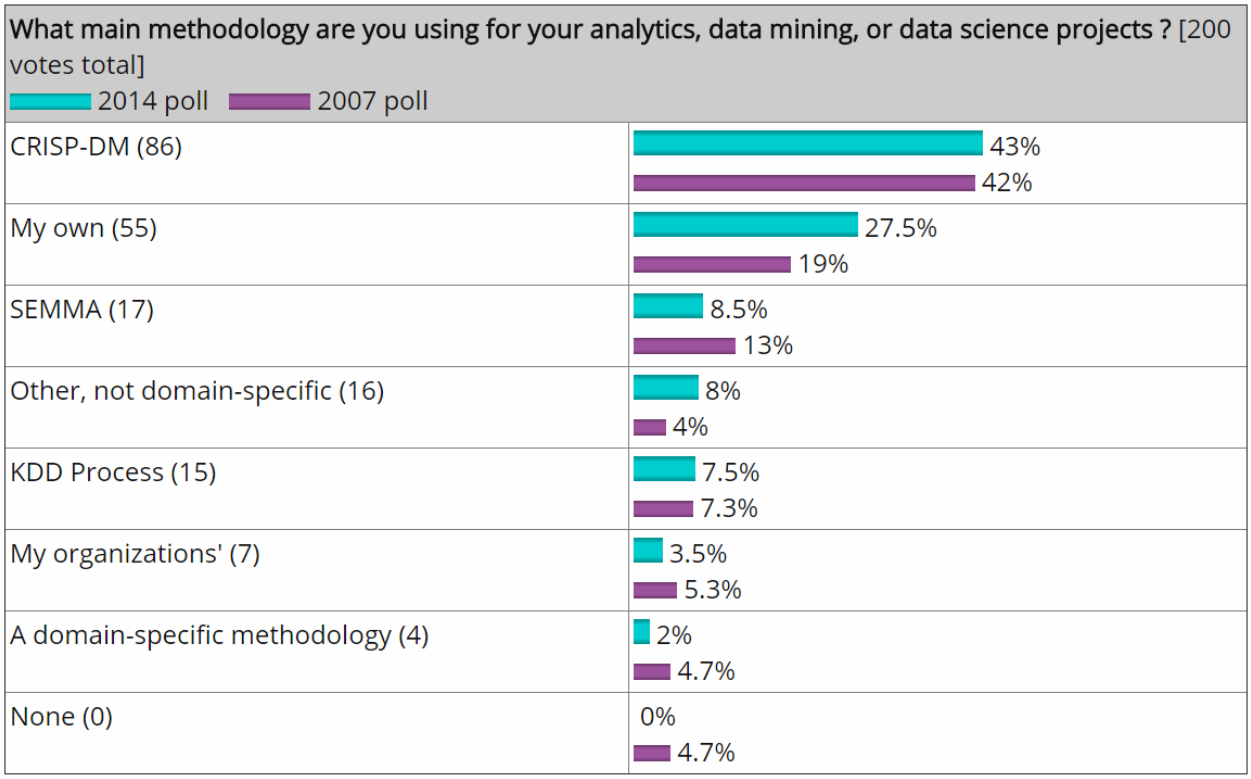

Figura 4. Pesquisa: Qual metodologia principal você está usando para seus projetos de análise, de mineração de dados ou de ciência dos dados?

Fonte: Piatetsky (2014).

Vejamos de seguida com mais detalhes duas das metodologias mais populares, a saber CRISP-DM e SEMMA.

1.2.1 Metodologia CRISP-DM

De acordo com Chapman, Julian, Randy, Thomas, Thomas, Colin & Rüdiger (2000) a metodologia CRISP-DM foi desenvolvida por um consórcio formado por NCR Systems Engineering Copenhagen, DaimlerChrysler AG, SPSS Inc. e OHRA Verzekeringen en Bank Groep B.V em 1996, que não visava fins lucrativos. Segundo os autores, a metodologia CRISP-DM, que é utilizada para o desenvolvimento de projetos de mineração de dados, é formada por um conjunto de fases e processos padrão, flexíveis e independentes da área e das ferramentas utilizadas, ou seja, foi concebida para funcionar em qualquer tipo de negócio e aceitar a aplicação de várias técnicas, porém de forma estruturada e sistemática.

Ainda de acordo com Chapman, et al. (2000), o ciclo da metodologia CRISP-DM compreende seis fases, que interagem entre si, apresentando ciclos e retornos. Por não ser linear, torna-se mais flexível. Todavia, cada fase necessita dos resultados obtidos em cada etapa antecedente, conforme constata-se na figura 5.

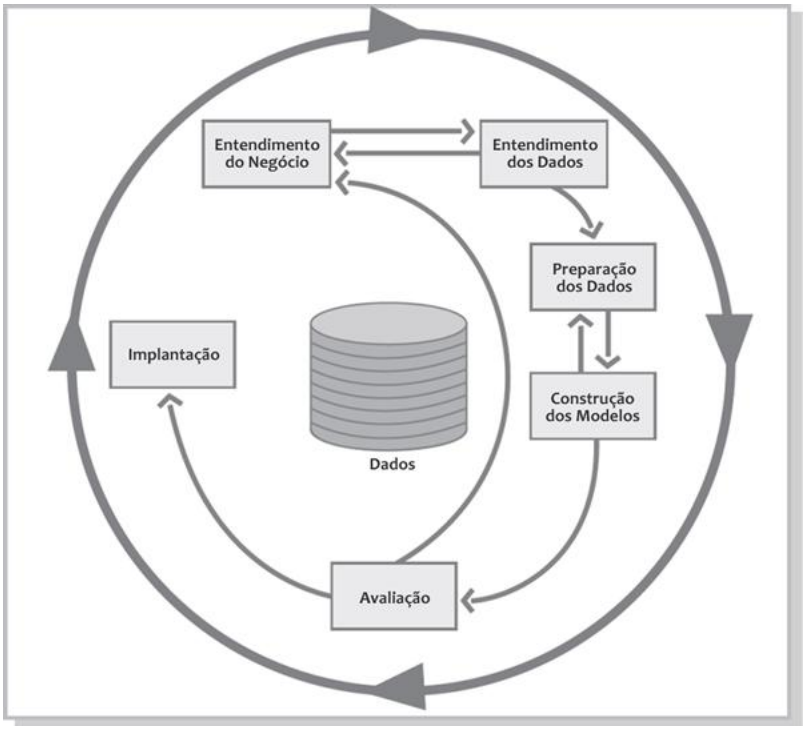

Figura 5. Fases do modelo de referência CRISP-DM.

Fonte: adaptada de Chapman, et al. (2000).

Na figura 5, a natureza cíclica do processo de mineração de dados é representada pelo círculo de setas externo. As dependências mais frequentes e importantes são indicadas pelas setas internas. A seta de saída de uma fase indica o início da próxima.

Explicamos a seguir, de forma sumária, cada uma das fases da metodologia CRISP-DM.

1. Entendimento do negócio (*Business Understanding*):

Na primeira fase do modelo CRISP-DM o foco está em identificar e assimilar o problema da empresa que precisa ser resolvido. É preciso compreender o negócio de acordo com seus objetivos e perspectivas e, por consequência, definir quais são as suas necessidades.

Nessa etapa, faz-se o levantamento de questões e detectam-se quais são os fatores mais significativos e importantes para a mineração de dados, que poderão auxiliar as etapas posteriores. Esse levantamento tem suma importância, pois, poderá influenciar ou até mesmo modificar os resultados finais.

2. Entendimento dos dados (*Data Understanding*):

Após entender o negócio e definir os objetivos, é imprescindível conhecer os dados e identificar quais são mais relevantes para a solução do problema em questão. É necessário verificar e organizar todos os dados disponíveis e que são indispensáveis para decifrar o problema que está sendo investigado. Pode haver necessidade de voltar à fase 1.

Nessa etapa, relatam-se como os dados foram adquiridos e descrevem-se as informações relevantes, como o seu formato e o seu conjunto de valores, de forma a identificar e compreender a informação contida neles que pode ser primordial para o estudo.

3. Preparação dos dados (*Data Preparation*):

Neste terceiro momento é realizado um conjunto de atividades de inspeção e preparação dos dados com o objetivo de se obterem os dados finais com os quais será criado e validado o modelo. Dessa forma, podem-se executar ações para obter dados mais limpos, como: filtrar, combinar e preencher valores vazios.

Nessa etapa, escolhem-se os atributos dos dados que foram selecionados e, em seguida, organizam-se de forma integrada em uma visão única para iniciar a análise.

4. Construção dos modelos (*Modeling*):

Na quarta etapa são selecionadas e aplicadas as técnicas de mineração de dados mais adequadas aos objetivos que foram especificados na fase de entendimento do negócio. Pode haver necessidade de voltar à fase 3.

Nessa etapa, aplicam-se os algoritmos de mineração de dados, que sejam capazes de produzir resultados mais satisfatórios sobre o conjunto final de dados que foram organizados na fase de preparação dos dados. Por consequência, possibilita-se a resolução da questão identificada na fase de entendimento do negócio.

5. Avaliação (*Evaluation*):

Nessa etapa são realizados os testes e a avaliação de desempenho dos modelos obtidos na etapa anterior, sendo necessário verificar se as necessidades identificadas no entendimento do negócio foram atendidas, assim como, se os objetivos do negócio foram alcançados. Pode haver necessidade de voltar à fase 1.

No caso desta dissertação, espera-se avaliar modelos com qualidade, confiabilidade e eficácia suficientes para identificar os motivos que levaram os alunos a abandonarem seus cursos.

6. Implantação (*Deployment*):

A última fase é a implantação, onde se desenvolvem e distribuem os resultados obtidos, sendo necessário que todos os envolvidos conheçam os resultados. Esses resultados possibilitaram a criação de um conjunto de ações para ser implantado dentro da instituição.

1.2.2 Metodologia SEMMA

A metodologia SEMMA foi desenvolvida pela empresa SAS, que atua no mercado em softwares e serviços de *business analytics*. O acrônimo SEMMA significa *Sample*, *Explore*, *Modify*, *Model*, *Assess*, e se refere ao processo de realização de um projeto de mineração de

dados, que evidencia, sobretudo, as características da implementação das técnicas e do processo, conforme se constata na figura 6.

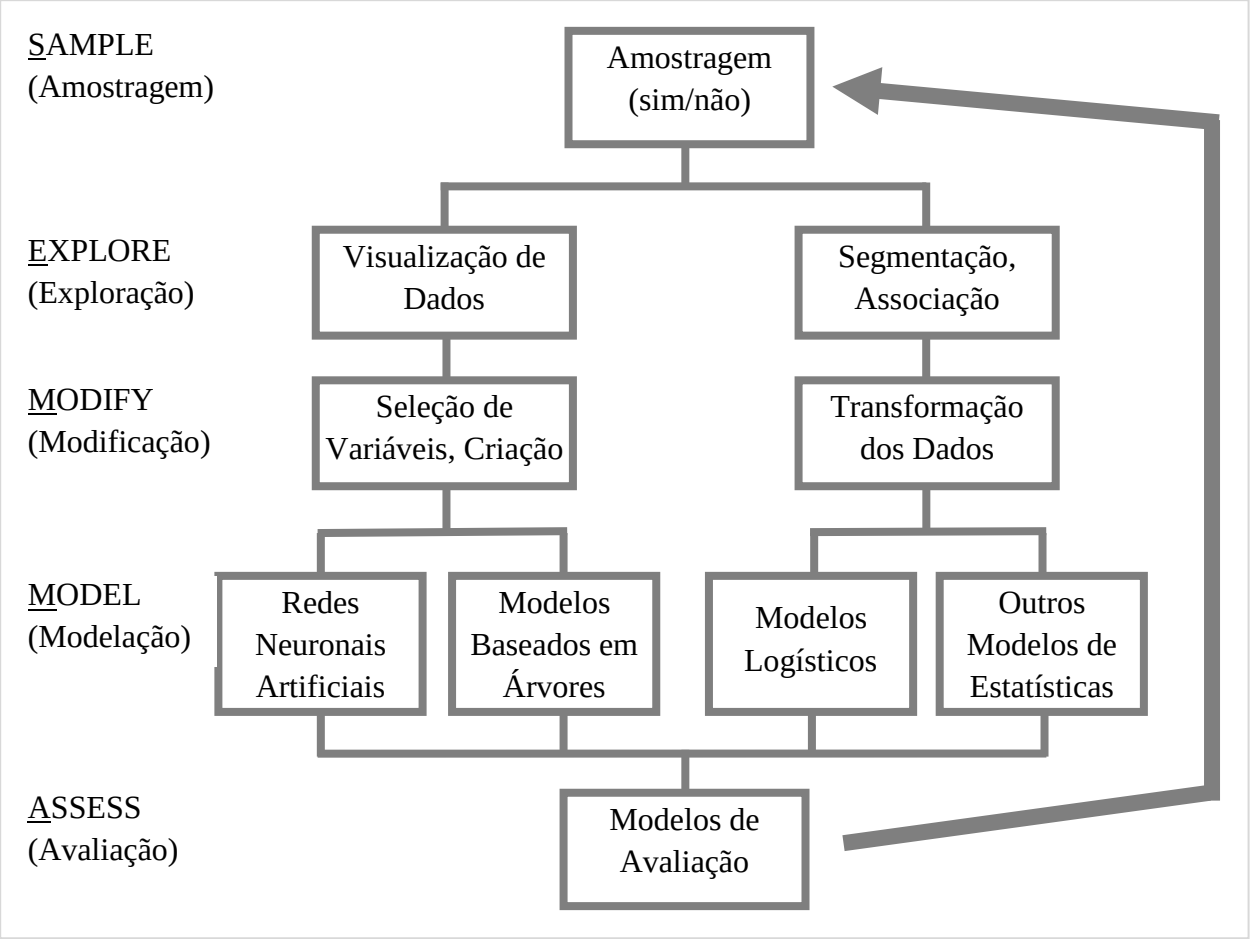

Figura 6. Fases da metodologia SEMMA.

Fonte: adaptada de Olson e Delen (2008) *apud* Nogueira (2014, pp. 7-8).

Explicamos a seguir, de forma sumária, cada uma das fases da metodologia SEMMA.

1. Amostragem (*Sample*):

A primeira fase da SEMMA consiste na amostragem dos dados, extraindo uma parcela dos dados que represente o volume total e que contenha informações significativas para manipulação rápida. Para aumentar a precisão da análise dos modelos, recomenda-se criar partições de dados para treino e validação dos mesmos.

2. Exploração (*Explore*):

A segunda fase é a exploração. Esta fase consiste na exploração dos dados, visual ou numérica, através da pesquisa de padrões, tendências e anomalias não previstas, com o objetivo de adquirir uma melhor compreensão e entendimento do conjunto de dados recolhidos na primeira fase.

3. Modificação (*Modify*):

Na fase modificação são realizadas alterações nos dados através da criação, seleção e transformação das variáveis para melhorar a construção do modelo com o objetivo de adquirir novas informações. Nesta fase pode ocorrer a inclusão de novas variáveis e a eliminação de valores omissos ou a sua substituição pela média. Também pode ser necessário eliminar algumas variáveis, mantendo-se apenas as mais significativas.

4. Modelação (*Model*):

A quarta fase consiste na modelação dos dados. Através de uma combinação de variáveis obtém-se um modelo, que representa padrões nos dados e prevê de forma eficaz e confiável o resultado desejado. As técnicas de modelação em mineração de dados incluem redes neuronais, modelos baseados em árvores de decisão, modelos logísticos e outros métodos estatísticos definidos pelos analistas. Cada uma delas tem os seus pontos fortes, devendo-se escolher a melhor ou mais apropriada para situações específicas através de métodos de otimização e de testes estatísticos significativos.

5. Avaliação (*Assess*):

A última fase do SEMMA é a avaliação. Esta fase consiste na avaliação dos dados através do melhor modelo obtido, verificando-se a utilidade e a confiabilidade dos resultados alcançados a partir do processo de mineração de dados.

1.3 Aplicações de mineração de dados

As técnicas de mineração de dados podem ser aplicadas para a descoberta de conhecimento em várias áreas, tais como: educacional, financeira, médica, industrial, biológica, comercial entre muitas outras, para a resolução de problemas de diversas índoles. Apresentamos a seguir alguns exemplos de problemas referentes a cada uma das áreas:

Financeira: Em se tratando de empréstimos bancários, como classificar novos clientes como prováveis adimplentes ou inadimplentes?

Médica: Como a mineração de dados pode apoiar o diagnóstico da doença de Alzheimer?

Industrial: Qual é a previsão do tempo de vida de uma máquina do setor de produção?

Biológica: Onde podem ser feitas novas ligações entre os átomos para aumentar a estabilidade da estrutura da proteína?

Comercial: Qual a probabilidade de um determinado grupo de clientes comprarem os produtos de uma determinada oferta?

Educacional: Como saber no início do ano letivo quais são os alunos com maiores e prováveis chances de abandono do curso e quais são os principais motivos para este acontecimento?

1.4 A escolha da ferramenta de mineração de dados

Com relação às modalidades de softwares, pode-se agrupar as ferramentas disponíveis para realizar a mineração de dados em dois grupos: as comerciais, que são softwares proprietários de grandes empresas e possuem custos elevados para a aquisição, e as de distribuição gratuita, cuja utilização não implica o pagamento de licenças de uso.

Exemplos de ferramentas comerciais:

IBM SPSS Statistics[2] – Software da empresa IBM - International Business Machines para apoio a tomada de decisão que inclui: aplicação analítica, mineração de dados, mineração de texto e estatística.

Oracle Data Mining (ODM)[3] – É uma ferramenta para a Mineração de Dados desenvolvida pela empresa Oracle para o uso em seu banco de dados Oracle.

Exemplos de ferramentas gratuitas:

O Projeto R (ou simplesmente "R") [4] – É um software de estatística que contém diversos pacotes com diversas funções estatísticas, matemáticas e econométricas.

WEKA (*Waikato Environment for Knowledge Analysis*)[5] – Uma das melhores ferramentas livre, que disponibiliza diversos algoritmos para as tarefas de mineração. Fornece as funcionalidades para pré-processamento, classificação, regressão, agrupamento, regras de associação e visualização.

[2] Disponível em https://www.ibm.com/br-pt/marketplace/spss-statistics, acessado em Outubro de 2018.
[3] Disponível em https://www.oracle.com/technetwork/database/options/advanced-analytics/odm/overview/index.html, acesso em Outubro de 2018.
[4] Disponível em https://www.r-project.org/, acessado em Outubro de 2018.
[5] Disponível em https://www.cs.waikato.ac.nz/ml/index.html, acessado em Outubro de 2018.

Capítulo II – Estudo Empírico

2 A Investigação

Apresenta-se, a seguir, o estudo empírico, utilizando-se da metodologia de investigação-ação.

2.1 A IA como metodologia de investigação

A Investigação-Ação (IA) é "um termo genérico para qualquer processo que siga um ciclo no qual se aprimora a prática pela oscilação sistemática entre agir no campo da prática e investigar a respeito dela" (Tripp, 2005, pp. 445-446). Obtendo-se desse processo um aperfeiçoamento de sua prática e da sua própria investigação ao executar as seguintes ações de forma cíclica: planejar, implementar, descrever e avaliar. Contudo, o autor complementa que "aplicações e desenvolvimentos diferentes do ciclo básico da investigação-ação exigirão ações diferentes em cada fase e começarão em diferentes lugares" (Tripp, 2005, p. 447).

De acordo com Tripp (2005), o processo básico de investigação-ação possui diversos desenvolvimentos, seja por ter se iniciado em diferentes localidades, épocas ou áreas do conhecimento. Dentre esses vários tipos de investigação-ação, o autor destaca a pesquisa-ação. Segundo o autor, não é fácil conceituar a pesquisa-ação, pois ela é um processo comum, habitual e amplamente aplicado em várias áreas, sendo desenvolvida e aperfeiçoada de inúmeras formas.

A pesquisa-ação é descrita como sendo "independente", "não-reativa" e "objetiva" e com a finalidade de "desenvolver o conhecimento e a compreensão como parte da prática (...) surgiu da necessidade de superar a lacuna entre teoria e prática" (Engel, 2000, p. 182).

A escolha dessa metodologia de pesquisa se deve ao fato de que ela "pode ser aplicada em qualquer ambiente de interação social que se caracterize por um problema, no qual estão envolvidos pessoas, tarefas e procedimentos" (Engel, 2000, p. 183).

Nesse trabalho essa metodologia será utilizada para aplicar o processo de mineração de dados através da metodologia CRISP-DM (*CRoss Industry Standard Process for Data Mining*), objetivando a elaboração e o desenvolvimento de um fluxo de trabalho para extrair conhecimentos úteis, que possam auxiliar os gestores do IFTM na tomada de decisão quanto à problemática da evasão escolar por parte dos seus alunos. A escolha da utilização da CRISP-DM se deve ao fato dela ter uma relação próxima com os modelos do processo de KDD e ser independente de ferramenta, sendo que o mesmo processo pode ser aplicado para analisar dados dos mais variados negócios. É relevante destacarmos que, apesar de sua proximidade com o KDD, a CRISP-DM dá uma ênfase maior ao problema de negócio.

2.2 O ciclo da IA

São definidas cinco fases de forma continua em um processo cíclico. O ciclo estrutural da investigação-ação é ilustrado, a seguir, na figura 7:

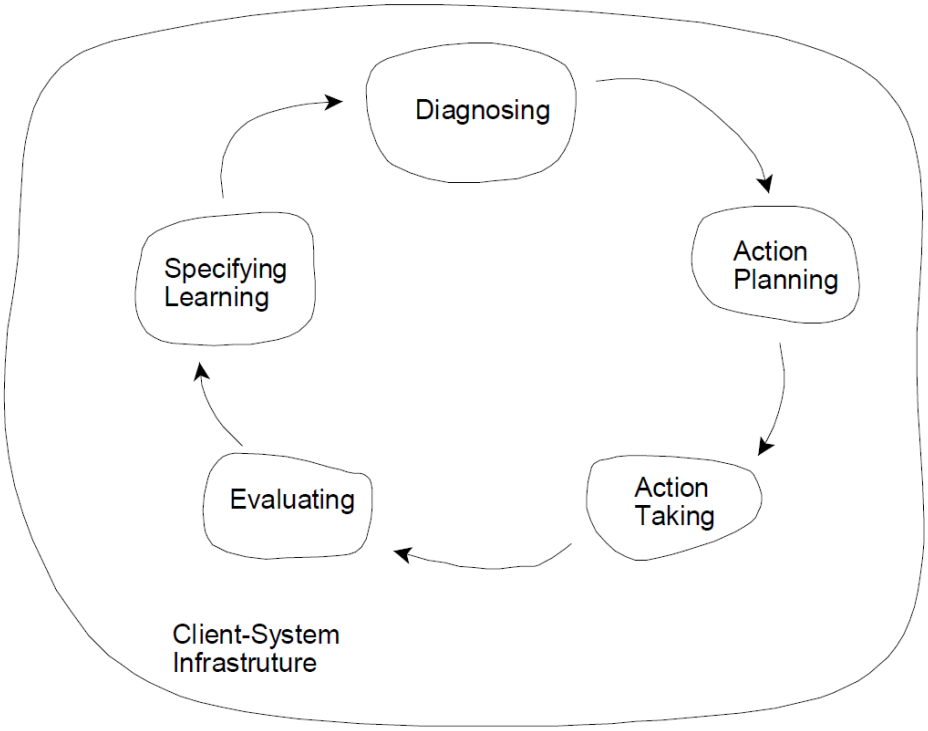

Figura 7. Ciclo da Investigação-Ação

Fonte: Baskerville (1999, p. 14).

Observa-se na figura 7 que a Infraestrutura do Sistema do Cliente (*Client-System Infrastruture*) contém as fases do processo cíclico da investigação e determina como será o ambiente da pesquisa. Neste ambiente existe a colaboração mútua entre os investigadores e a organização. Além disso, ele contém a abrangência e os limites do domínio da pesquisa, assim como a sua amplitude na disseminação do aprendizado adquirido.

Cada uma das cinco fases do ciclo da IA é exemplificada e explicada de forma sumária, a seguir. Explicam-se também as atividades levadas a cabo pelo investigador em cada uma dessas fases.

1. Diagnóstico (*Diagnosing*):

Na primeira fase do ciclo da IA identifica-se um problema, ou seja, define-se uma situação organizacional que, por parte dos investigadores, é algo que o intriga, por isso, impulsiona um desejo de mudança da organização. É o reconhecimento de que algo pode ser melhorado na gestão administrativa da instituição com relação à dificuldade de reduzir o elevado índice de evasão dos alunos.

No caso desta investigação, depois de identificado o problema, afirma-se que nessa proposta de trabalho há um alto grau de importância e relevância para o IFTM, e que a sua

viabilidade se dará através da pesquisa realizada na base de dados da instituição.

A pesquisa preliminar realizada nos primeiros capítulos deste trabalho, juntamente com a revisão da literatura, tem como propósito entender o que realmente está ocorrendo na instituição com relação à situação problemática e comprovar que as ações já implantadas até aqui não estão sendo suficientes para reduzir de forma eficaz a taxa de evasão de alunos. Conforme a pesquisa preliminar comprova, o percentual de evasão está acima do esperado pelo Plano de Desenvolvimento Institucional.

2. Plano de Ação (*Action Planning*):

Para provocar uma mudança na atual situação do elevado índice de evasão dos alunos, identificada no item anterior, é necessário criar, nesta fase, um planejamento em busca de soluções tangíveis, ou seja, desenvolver um plano de ação para reverter essa situação problemática.

A criação de uma lista de ações, que poderão ser implantadas dentro da instituição para a tomada de decisões pela gestão administrativa, tem como objetivo propor mudanças organizacionais para resolver ou diminuir o problema.

Dessa forma, no caso desta dissertação, a abordagem para provocar essa mudança organizacional será iniciada com um processo de investigação contextualizado, a fim de planejar com maior eficiência as ações que serão executadas com a mineração de dados. Assim, a lista de ações que foram planeadas durante essa investigação, foram as seguintes:

- reuniões com os profissionais da área de TI, que formam a equipe de desenvolvedores responsáveis pela construção e manutenção do ERP-IFTM e do SCA, para um melhor entendimento destas ferramentas administrativas.

- utilização das ferramentas de mineração de dados para descobrir o perfil do aluno evadido.

- apresentação das principais causas da evasão, a partir dos resultados obtidos com o processo de mineração de dados.

3. Ação (*Action Taking*):

Nesta fase, ocorre a implementação do plano de ação, ou seja, é posto em prática. Os investigadores trabalham em colaboração mútua com os profissionais da instituição, que estão envolvidos com o processo de investigação, ambos com um papel ativo na intervenção. Dentre as várias formas de intervenção estratégica que podem ser adotadas, tem-se a direta, onde as mudanças no meio são conduzidas, diretamente, pelos investigadores e sua pesquisa, e a indireta, onde as mudanças são solicitadas, indiretamente, pelos investigadores através da sua pesquisa.

No contexto desta investigação, foram necessárias quatro reuniões com a equipe de TI

para ter uma maior compreensão do ERP-IFTM e do SCA. Na primeira reunião, foi realizado um levantamento de todas as ferramentas do ERP-IFTM que estão diretamente relacionadas com o SCA para um melhor entendimento do negócio. Na segunda reunião, foram relacionadas todas as tabelas pertinentes à questão de investigação, para entender como estão organizados os dados no Banco de Dados (BD) do ERP-IFTM. Na terceira reunião, foram enumeradas as colunas (campos) de cada uma das tabelas listadas na reunião anterior, que contêm as informações indispensáveis para a coleta de dados inicial. Na quarta e última reunião, ficaram reconhecidos os valores armazenados em cada uma das colunas selecionada, a fim de, preparar os dados para a execução dos algoritmos de mineração de dados. Em seguida, foi implementada a metodologia CRISP-DM que permitiu a aplicação do processo de KDD para a descoberta de conhecimento na base de dados do ERP-IFTM. Devido à importância da implementação da metodologia CRISP-DM, esta será descrita de forma mais detalhada no capítulo 3. Como resultados, foram sugeridas diversas ações a implementar pelos gestores responsáveis, de modo a diminuir a taxa de evasão dos alunos.

Os principais motivos que levaram a escolha da utilização do Weka para realizar este trabalho de mineração de dados foram: a ferramenta gratuita; a interface gráfica é simples e amigável, o que facilita o seu uso e tornar as etapas da mineração de dados mais intuitivas; é uma ferramenta poderosa com longa data de desenvolvimento e aprimoramento, que está sempre sendo atualizada com novas funcionalidades e variados algoritmos capazes de executar diversas tarefas com o intuito de auxiliar a investigação dos dados.

4. Avaliação (*Evaluating*):

Depois de concluídas as ações, inicia-se o momento de se avaliar os resultados, ou seja, analisar e interpretar os resultados obtidos na fase anterior. Nesta fase, segue-se o monitoramento das ações e a avaliação de sua eficácia.

Foi apresentado um conjunto de ações à gestão administrativa para auxiliar os gestores na tomada de decisões quanto à problemática da evasão escolar, que foram aprovadas de forma positiva.

5. Aprendizagem Específica (*Specifying Learning*):

Na fase de aprendizagem específica, se as ações da fase anterior não obtiveram eficácia comprovada, ou seja, não alcançarem resultados predominantes positivos, os investigadores, a partir das reflexões e discussões dos resultados adquiridos, podem aperfeiçoar a sua pesquisa e impulsionar o início de um novo ciclo de IA.

De maneira satisfatória, os resultados obtidos com as ações dessa pesquisa foram eficazes, ou seja, a busca pelo conhecimento na base de dados do ERP-IFTM, apoiada por essa investigação, foi bem-sucedida. O próximo passo é a comunicação dos resultados, que

pode ser conseguida com a publicação dos frutos dessa experiência, através de um artigo na Revista Inova, que é uma revista especializada e hospedada na Plataforma de Publicações do IFTM.

2.3 O contexto da investigação

A missão do IFTM é "Ofertar a educação profissional e tecnológica por meio do ensino, pesquisa e extensão, promovendo o desenvolvimento na perspectiva de uma sociedade inclusiva e democrática".

É de senso comum que uma instituição de ensino nunca conseguirá ofertar e promover a educação se os seus alunos evadirem dos cursos que estão matriculados.

Dessa forma, o presente estudo tem uma enorme relevância, pois encontrar os motivos pelos quais os alunos não estão concluindo os seus estudos tem uma importância fundamental para descobrir e apresentar ações que diminuam o problema da evasão escolar e com isso, contribuir para a melhoria na qualidade da vida acadêmica e promover a permanência do aluno na instituição até a sua formação.

Enquanto instituição da Rede Federal de Educação Profissional, Científica e Tecnológica, o IFTM oferece cursos técnicos de nível médio e de graduação (tecnologia, licenciaturas, bacharelados) distribuídos em seus 7 (sete) *campi* (Ituiutaba; Paracatu; Patos de Minas; Patrocínio; Uberaba; Uberlândia e Uberlândia Centro) e seus 2 (dois) *campi* avançados (Campina Verde e Uberaba Parque Tecnológico), nas modalidades presencial e a distância (EaD). O IFTM também oferece cursos de pós-graduação *lato sensu* (Especialização) e *stricto sensu* (Mestrado).

De acordo com o Relatório de Gestão do ano de 2012 (IFTM, 2013), constata-se um número alto no total geral de alunos evadidos no IFTM nesse mesmo ano, perfazendo um índice aproximado de 16% do total geral dos alunos matriculados, conforme é ilustrado na tabela 2:

Tabela 2. Quantitativos de alunos evadidos / matriculados por campus do IFTM em 2012.

Campus	Evadidos	Matriculados	Índice %
Uberaba	619	3.226	19,19
Ituiutaba	82	2.160	3,80
Patrocínio	35	782	4,48
Uberlândia	68	1.203	5,65
Uberlândia Centro	111	600	18,50
Paracatu	1.157	4.907	23,58
Total Geral	**2.072**	**12.878**	**16,09**

Fonte: adaptada do SISTEC *apud* IFTM (2013, pp. 217-230).

A tabela 2 não apresenta os dados dos *campi* Uberaba Parque Tecnológico, Patos de Minas e Campina Verde porque os mesmos ainda não haviam sido implementados. O Plano Estratégico de Ações de Permanência e Êxito dos Estudantes do IFTM visa acompanhar, entre outros indicadores institucionais, o da evasão dos alunos, sendo parte de um conjunto de ações que objetivam amenizar o problema da retenção e da evasão no âmbito do IFTM. Esse plano foi realizado pela PROEN através de uma pesquisa que utilizou questionários com os alunos evadidos e possui dentre os seus objetivos específicos identificar para cada curso do IFTM no ano de 2012: o número de alunos evadidos e retidos, os motivos e as estratégias de intervenção para diminuir a evasão.

De acordo com o diagnóstico qualitativo do Plano Estratégico de Ações de Permanência e Êxito dos Estudantes do (IFTM, 2016, pp. 13-14), as principais causas da evasão são:

Relacionadas à dimensão individual do estudante:

a) dificuldades na formação escolar anterior;
b) não adaptação à vida acadêmica;
c) problemas financeiros do estudante ou da família;
d) estudo paralelo em outra instituição;
e) incompatibilidade com o horário de trabalho;
f) distância entre sua moradia e a instituição;
g) problemas familiares;
h) o curso não atendeu às expectativas dos alunos;
i) dificuldades em conciliar o trabalho com os estudos;
j) indisponibilidade de tempo para estudar fora da instituição;
k) falta de transporte adequado para chegar a instituição.

Relativas à dimensão institucional:

a) falta de estrutura do campus em relação a ambiente e ferramentas;
b) desconhecimento do mercado de trabalho;
c) retenções em disciplinas ou estágio.

Dentre as diversas propostas de intervenção apresentadas pelo plano estratégico, destacam-se:

a) oferecer aulas de reforço;
b) criar cursos de nivelamento online para alunos com déficit de conteúdos básicos;
c) criar a figura do professor-tutor de turma;
d) criar grupos de estudo através de monitorias e sobre tendências do mercado;
e) criar uma comissão de orientação profissional;

f) criar novas ferramentas no SCA para os gestores acompanharem as notas e/ou faltas dos alunos.

Ainda de acordo com o plano estratégico, concluiu-se que "se faz necessário conhecer e avaliar a complexidade de fatores sociais, econômicos, culturais e acadêmicos que intervêm na vida acadêmica referente à formação dos estudantes, uma vez que são tais fatores que levam ao êxito ou à desistência do curso" (IFTM, 2016, p. 16).

Todavia, apesar dos esforços e das ações executadas para evitar ou diminuir a evasão, constata-se que a meta não foi cumprida de forma satisfatória para todos os *campi*, conforme revelado no Relatório de Gestão 2016 do IFTM e no Relatório de Gestão 2017 do IFTM, ilustrados na tabela 3:

Tabela 3. Índice de evasão escolar por campus do IFTM nos anos de 2016 e 2017.

Campus	2016			2017		
	Meta	Indicador	Cumpriu?	Meta	Indicador	Cumpriu?
Uberaba	13,00%	16,65%	Parcialmente	15,00%	19,10%	Parcialmente
Ituiutaba	20,00%	28,04%	Parcialmente	17,50%	25,43%	Parcialmente
Patrocínio	10,00%	25,42%	Parcialmente	20,00%	20,29%	Parcialmente
Uberlândia	20,00%	5,84%	Sim	17,50%	14,47%	Sim
Uberlândia Centro	30,00%	20,27%	Sim	25,00%	17,68%	Sim
Paracatu	20,00%	22,80%	Parcialmente	15,00%	16,35%	Parcialmente

Fonte: adaptada do Relatório de Gestão 2016 (IFTM, 2017, pp. 336-337) e do Relatório de Gestão 2017 (IFTM, 2018, pp. 336-337).

A tabela 3 apresenta que apenas 2, do total de 6 *campi*, conseguiram alcançar a meta proposta para os anos de 2016 e 2017 na redução do índice da taxa de evasão escolar de alunos.

Ao compararmos as metas da tabela 3 com a tabela 1, que são as metas definidas originalmente no PDI, 20% (2016) e 17,5% (2017), constata-se que o índice foi ajustado para cada campus. Dessa forma, ainda que um valor de redução da taxa de evasão esteja acima do objetivado inicialmente no PDI, poderá apontar uma meta cumprida, como é caso do Campus Uberlândia Centro com as novas metas de 30,00% e 25,00% e aferições de 20,27% e 17,68% para os anos de 2016 e 2017, respectivamente.

Isto posto, pode-se afirmar que as ações de combate à evasão até aqui desenvolvidas não são eficazes na sua totalidade.

Muitas ações e projetos implantados no IFTM com o intuito de evitar a evasão, conforme exposto no Relatório de Gestão 2016 (IFTM, 2017), como por exemplo, a concessão de bolsas, a assistência e o auxílio estudantil, podem ter um custo financeiro alto e não resolve o problema causador da evasão escolar que, por enquanto, é desconhecido.

A mineração de dados pode ser usada para descobrir mais detalhes sobre a evasão escolar e possivelmente a sua causa.

Para que o trabalho não se torne exaustivo e copiando o mesmo intervalo de tempo do PDI, serão coletados, exclusivamente dos cursos superiores, dados no período de 5 anos. Sendo assim, a pesquisa englobará, exclusivamente, os últimos 5 anos completados: 2012 a 2016. Dessa forma, estarão inclusos os 3 (três) anos concluídos do PDI vigente e os últimos 2 (dois) anos do PDI anterior.

Ao incluir o ano de 2012 na pesquisa, possibilita-se estabelecer um comparativo com os resultados obtidos com o Plano Estratégico de Ações de Permanência e Êxito dos Estudantes do IFTM.

No que concerne à delimitação do problema e à generalidade dos resultados, os motivos que levam o aluno à desistência de um determinado curso talvez não sejam os mesmos de outro aluno que está matriculado em um curso diferente do analisado, assim como, fatores de localização e estrutura física podem diferenciar a conclusão da pesquisa de um campus localizado em cidade distinta da que faz parte da pesquisa.

Os dados da investigação serão coletados, exclusivamente, do sistema ERP do IFTM, mais especificamente da base de dados do SCA. Não serão exploradas as bases de dados fora da Intranet do instituto, como da Internet, secretarias de educação e redes sociais.

Acreditamos que o conhecimento alcançado auxiliará a alta gestão do IFTM na tomada de decisões ao criar uma expectativa de mudança de processos na instituição quanto à problemática da evasão escolar.

A presente proposta de dissertação limita-se única e exclusivamente a apresentar os resultados obtidos pela investigação através da mineração de dados, não se comprometendo em resolver por definitivo os problemas da evasão escolar no IFTM.

Se encontrarmos soluções para o problema proposto, elas serão apresentadas de forma organizada e em uma linguagem clara para que todos possam ter acesso, porém, cabe apenas às autoridades competentes pela administração e gestão escolar tomar a decisão de aplicá-las.

Em conformidade com os objetivos de alcançar resultados mais claros e apresentar os aspectos mais relevantes para o entendimento do negócio e dos dados, inicia-se a abordagem do nosso estudo com uma visão geral do IFTM, suas ferramentas informáticas e as características do funcionamento do Sistema de Controle Acadêmico (SCA) e dos seus módulos, que são agrupamentos de ferramentas e recursos.

É fundamental entendermos o funcionamento do SCA para investigarmos o problema de evasão escolar no IFTM e encontrarmos possíveis soluções para evitá-lo.

O sucesso dessa investigação será obtido quando o processo de mineração de dados no SCA descobrir as características dos alunos que evadem do IFTM.

2.3.1 O IFTM

O IFTM é uma instituição de ensino preocupada com o sucesso pessoal e a formação profissional de seus alunos. Por isso, não mede esforços ao assumir o compromisso de oferecer uma educação profissional e tecnológica de qualidade. Além de manter um excelente capital humano na organização, que inclui uma equipe de professores e pedagogos qualificados e uma equipe administrativa competente, o IFTM possui ferramentas que auxiliam a administração, o controle e a gestão das suas atividades educativas e cotidianas. Dentre essas ferramentas, estão as de gestão tecnológica, onde se destacam o Virtual IF e o ERP-IFTM.

2.3.2 O Virtual IF e o ERP-IFTM

O Virtual IF é o ambiente de intranet do IFTM que tem por objetivo concentrar de forma interligada todos os serviços virtuais da instituição. Com o propósito de ser um local de trabalho colaborativo, o Virtual IF agrupa todos os softwares proprietários e institucionais em um sistema ERP que auxilia a administração e o gerenciamento das atividades desenvolvidas tanto no plano administrativo quanto no acadêmico.

O IFTM mantém uma equipe de desenvolvedores no setor de Tecnologia e Informação (TI) da instituição, denominado “Fábrica de Software”, formada por: 10 (dez) Analistas de TI, 7 (sete) Técnicos de TI e 1 (um) Programador Visual, com o objetivo de ampliar a quantidade e melhorar a qualidade e eficiência dos softwares que compõem o ERP-IFTM, dessa forma, mantendo-o em constante desenvolvimento e aprimoramento.

2.3.3 O Sistema de Controle Acadêmico do Virtual IF

Junto aos vários sistemas de softwares que integram o ERP-IFTM está o Sistema de Controle Acadêmico (SCA), contendo informações cadastrais da vida acadêmica do aluno, como: forma de ingresso, cursos matriculados, disciplinas cursadas, grade horária, controle de frequências, registro de notas e seu aproveitamento.

2.3.3.1 Atribuições dos agentes envolvidos no funcionamento do SCA

De acordo com as Orientações Gerais Quanto ao Funcionamento do SCA (IFTM, 2013), dentre as várias orientações gerais que determinam as atribuições específicas e exigidas dos agentes envolvidos no funcionamento do SCA, destacam-se as que envolvem alteração e/ou inserção de novos registros na base de dados:

Coordenação de Registro e Controle Acadêmico (CRCA):

1. Cadastrar o Projeto Pedagógico de Curso (PPC);
2. Cadastrar os componentes curriculares da Matriz Curricular, constantes no PPC;
3. Vincular o curso à oferta da Matriz Curricular;
4. Inserir as datas referentes ao início e término do período letivo;
5. Informar os dados relativos à parametrização para o funcionamento dos "blocos" específicos de frequência e de notas no diário eletrônico.
6. Executar a totalização dos diários e o fechamento do período letivo dos estudantes;
7. Executar os processos de matrícula e de renovação de matrícula (rematrícula).

Coordenação de Curso (CC):

1. Cadastrar a oferta das Unidades Curriculares para os períodos letivos;
2. Inserir os horários de aulas;
3. Homologar os diários eletrônicos de cada um dos componentes curriculares;
4. Executar os processos de ajustes de matrículas dos estudantes.

Professor:

1. Cadastrar o Plano de Ensino;
2. Realizar lançamentos de frequência e notas dos alunos;
3. Inserir atividades diárias desenvolvidas nas salas de aula;
4. Efetuar o "encerramento" dos blocos de frequência, notas e atividades.

2.3.3.2 Recursos / ferramentas e permissões de acesso aos módulos do SCA

A tabela 4, ilustra todos os módulos do SCA e a disponibilidade para cada um dos agentes / usuários envolvidos, inclusive o aluno:

Tabela 4. Virtual IF: Acesso aos módulos do SCA.

Módulos	Disponibilidade			
	Aluno	Professor	CC	CRCA
Aluno	sim	não	não	não
Assistência Estudantil	sim	não	não	sim
Banco de Estágio, Emprego e Currículo	sim	não	não	não
Coordenação de Registro e Controle Acadêmico	não	não	não	sim
Disco Virtual Acadêmico	sim	sim	sim	não
Eventos	sim	sim	sim	sim
Gestor de Curso	não	não	sim	não
Mural de Recados	sim	sim	sim	não
Professor	não	sim	sim	não
Serviço de Agendamento de recursos	não	sim	sim	não

Observa-se na tabela 4 que a Coordenação de Curso tem acesso a todos os recursos do Professor. Esse acontecimento se deve ao fato de que, normalmente, o responsável pela

coordenação de cursos também é um professor, que ministra algumas disciplinas. Observa-se também que alguns professores e/ou gestores podem ter acesso aos recursos dos alunos, pois os mesmos podem ser ex-alunos ou alunos de cursos de idioma ou mestrado.

2.3.3.3 Módulo Aluno

A utilização do Módulo do Aluno é exclusiva ao aluno e possui diversos recursos / ferramentas para auxiliá-lo na sua vida acadêmica, permitindo-o realizar as seguintes ações:

- solicitar a carteirinha de estudante;
- visualizar suas notas e frequências;
- imprimir grade horária, declaração de matrícula e boletins anteriores;
- acompanhar a situação da solicitação da carteirinha de estudante;
- consultar o cadastro dos responsáveis no "Pais e Filhos", que é um sistema disponível no endereço https://portalif.iftm.edu.br/, que tem como objetivo permitir que os responsáveis possam acompanhar a vida escolar de seus filhos.

Portal do Aluno IFTM – é um aplicativo disponível para dispositivos móveis, como *smartphones* e *tablets*, que utilizam o sistema operacional Android. Apesar de não estar contido dentro do Módulo do Aluno, é uma ferramenta que obtêm dados e informações do ERP-IFTM para possibilitar que aluno visualize suas notas, frequências, arquivos do disco virtual e mensagens do mural de recados.

2.3.3.4 Módulo Assistência Estudantil

O Programa de Assistência Estudantil, ofertado a todos os estudantes dos cursos regulares presenciais do IFTM e sob a responsabilidade da Pró-reitora de Extensão (PROEXT), entre outras finalidades, busca garantir a permanência dos estudantes na instituição até a sua formação.

O benefício oferecido pelo programa é dividido em: Assistência Estudantil, que possui apoio financeiro para garantia de sua permanência nos estudos; e Auxílio Estudantil, com o apoio financeiro ou não, para atenção à saúde biopsicossocial, concessão de alojamento e participação em atividades/eventos acadêmicos.

O Módulo Assistência Estudantil disponibiliza os seguintes recursos / ferramentas, para:

Alunos: inscrever-se e acompanhar o resultado dos editais da assistência estudantil; imprimir o termo de compromisso se aceito, senão solicitar recurso da decisão.

CRCA: cadastrar e publicar os editais da assistência estudantil; realizar o processo seletivo: validar e analisar as inscrições, acompanhar resultados parciais e finais, gerenciar

convocações, termo de compromisso e desligamentos do programa, gerenciar recursos financeiros e emitir relatórios de gestão administrativa.

2.3.3.5 Módulo Banco de Estágio, Emprego e Currículo

O Módulo Banco de Estágio, Emprego e Currículo está disponível a todos os alunos e ex-alunos com o objetivo de motivá-los e incentivá-los a obter sucesso na sua formação acadêmica e profissional.

Essa ferramenta disponibiliza aos alunos os seguintes recursos: consultar suas informações de cadastro pessoal no sistema acadêmico e realizar a manutenção (consulta e alteração) das informações de contato, como e-mail, telefones e correspondência; incluir o currículo para participar de processos seletivos de estágio e emprego e consultar as vagas que foram disponibilizadas por empresas parceiras.

2.3.3.6 Módulo Controle de Registro Acadêmico (CRA)

O Módulo Controle de Registro Acadêmico (CRA) está disponível ao coordenador de registro e controle acadêmico e permite o acesso somente às informações de seu campus.

O CRA disponibiliza aos coordenadores os seguintes recursos / ferramentas, referente a:

Cursos: cadastro de Projeto Pedagógico do Curso (PPC), matriz curricular e oferta da matriz.

Alunos: controle do Cadastro do Aluno: dados pessoais, documentos, endereço, contatos e cursos, edição de matrículas, impressão e/ou tradução do Histórico Escolar, cadastro do Exame Nacional de Desempenho dos Estudantes (ENADE), impressão das carteirinhas de estudante e exclusão de frequência e notas.

Professores: controle do Cadastro do Professor: dados pessoais, documentos, endereço e contatos.

Disciplinas: homologação das disciplinas que são ofertas dentro de uma matriz curricular, fechamento e rematrículas.

Relatórios: emissão de relatórios de gestão e controle administrativos relativos aos alunos, professores e cursos.

2.3.3.7 Módulo Disco Virtual Acadêmico

O Disco Virtual Acadêmico é o espaço virtual destinado aos professores para compartilhar materiais das aulas com os seus alunos. São disponibilizados arquivos em

diversos formatos, como documentos, apresentações e planilhas, organizados por disciplina e curso matriculado.

Esse módulo oferece os seguintes recursos / ferramentas, para:

Professor: gerenciar a inclusão de novos materiais e/ou a exclusão de materiais antigos.

Aluno: explorar os conteúdos disponibilizados pelo professor, inclusive baixá-los.

2.3.3.8 Módulo Eventos

O Módulo Eventos tem o objetivo de acompanhar e gerenciar os eventos acadêmicos do IFTM, como seminários, feiras e simpósios.

Esse módulo oferece os seguintes recursos / ferramentas, para:

Aluno: obter informação sobre os eventos acadêmicos, acompanhar o andamento das suas inscrições e, ao final, imprimir os seus certificados de participação.

Professor: inscrever e consultar as inscrições dos seus alunos orientados.

2.3.3.9 Módulo Gestor de Curso (GC)

O Módulo Gestor de Curso (GC) está disponível ao coordenador de curso e permite o acesso somente às informações dos cursos que coordena.

O GC disponibiliza aos coordenadores de curso os seguintes recursos / ferramentas, referente à:

Cursos: alocação das ofertas de currículo ao inserir as disciplinas para o curso oferecido, cadastro das disciplinas com as suas respectivas cargas horária, criação da matriz curricular que pertence a cada curso e oferta (disponibilização) da matriz curricular.

Alunos: realiza ações relacionadas às matrículas: ajustes, movimentação por turma e emissão da declaração de matrícula.

Disciplinas: homologação das disciplinas que são ofertadas dentro de uma matriz curricular.

Site: criação e publicação automática das páginas do sítio web (www.iftm.edu.br) que divulga os cursos oferecidos em cada campus do IFTM.

Relatórios: emissão de relatórios de gestão e controle administrativos relativos aos alunos e professores.

2.3.3.10 Módulo Mural de Recados

O Módulo Mural de Recados facilita a comunicação acadêmica, permitindo a troca de mensagens, entre:

- Aluno com os seus professores e o seu coordenador de curso.
- Professor com seus alunos.
- Coordenação de Curso com os professores e os alunos dos cursos que é gestor.

2.3.3.11 Módulo Professor

A utilização do Módulo do Professor é exclusiva ao professor e possui diversos recursos / ferramentas para auxiliá-lo na sua vida acadêmica, principalmente, no que diz respeito ao planejamento e acompanhamento das aulas ministradas, permitindo-o realizar as seguintes ações:

- editar o plano de ensino com seu respectivo cronograma para cada disciplina que ministra aulas;
- lançar as frequências dos alunos;
- editar as atividades avaliativas das disciplinas que ministra e lançar as respectivas pontuações dos alunos, incluídos também os estudos autônomos e a recuperação;
- emitir relatórios de controle relativos aos alunos;
- acessar o resultado da Comissão Própria de Avaliação, que objetiva a melhoria das atividades pedagógicas, com perguntas respondidas pelos alunos e relacionadas à sua pontualidade, assiduidade, cordialidade, dinamismo e criatividade ao ministrar as aulas, entre outros critérios de avaliação.

2.3.3.12 Módulo Serviço de Agendamento de Recursos

O Serviço de Agendamento de Recursos permite o controle de disponibilização e reserva de recursos como: auditório, laboratórios, sala de reuniões e equipamentos, por exemplo, *data show*. A Coordenação de Curso executa o cadastro dos agendamentos e realiza o atendimento dos recursos solicitados pelos professores.

Capítulo III – Implementação da Metodologia CRISP-DM

3 A Metodologia CRISP-DM

Este capítulo compreende as seis fases do ciclo da metodologia de mineração de dados CRISP-DM. Durante a sua implementação foi respeitado o fluxo dos ciclos sem a necessidade de retorno a uma etapa antecedente, pelo motivo de que as informações obtidas em cada etapa foram suficientes e significativas para se alcançarem resultados pertinentes e relevantes quanto à questão investigada.

3.1 Fase 1 – Entendimento do negócio

A primeira fase compreende o negócio ao identificar o IFTM dentro da perspectiva do problema da evasão escolar, a fim de obter resultados que permitam alcançar os objetivos deste trabalho.

3.1.1 Definição de evasão

As movimentações na base de dados de matrícula do aluno que indicam a ocorrência de evasão são as seguintes:

1. Não Matriculado – o aluno não renovou a matrícula.

2. Cancelado – o aluno efetuou a matrícula, porém, antes das aulas iniciarem, ele solicitou o seu cancelamento.

3. Desistência – ocorre quando o aluno manifesta não ter mais interesse em continuar com os seus estudos.

4. Desvinculado do curso – a secretaria desligou o aluno do curso por motivos previstos nas resoluções e regulamento do curso, como, por exemplo, o abandono de curso.

Dessa forma, delimitam-se as etapas da aplicação das técnicas de mineração de dados somente aos dados e às movimentações que envolvem essas ocorrências.

3.2 Fase 2 – Entendimento dos dados

Inicialmente, é introduzido o banco de dados do ERP-IFTM com a finalidade de descrever as tabelas do Módulo Acadêmico e como os dados foram obtidos para criar o arquivo “alunos.arff”. Em seguida, será utilizado o programa Weka para explorar os dados em detalhe, avaliar a sua qualidade, indicar todos os seus aspectos relevantes e identificar problemas de anomalias, se existirem. Por fim, os dados serão preparados de forma que estejam mais limpos e prontos para a execução dos algoritmos de mineração de dados.

3.2.1 O banco de dados do ERP-IFTM / Módulo Acadêmico

O Banco de Dados do ERP-IFTM contém 52 *schemas*, que possibilitam agrupamentos físicos de dados no *PostgreSQL*, que é um SGBD (Sistema Gerenciador de Banco de Dados) objeto-relacional de código aberto.

O *schema* frequentemente usado em todos os módulos do Virtual IF é o *Public*, ou público, que contém as tabelas comuns a todos os módulos do ERP-IFTM, das quais foram utilizadas 8, as pertinentes aos dados que serão coletados do MAC (Módulo Acadêmico). A tabela 5 descreve as tabelas de dados.

Tabela 5. Tabelas do Banco de Dados do ERP-IFTM - *Schema Public*.

Tabela	Descrição
cm_campus	Cadastro de todos os *campi* do IFTM.
cm_cidade	Cadastro de cidades de todos os estados do Brasil.
cm_estado_civil	Indica o estado civil do aluno, se solteiro, casado etc.
cm_nacionalidade	A nacionalidade do aluno.
cm_necessidade_especial	Indica se o aluno possui alguma deficiência física ou mental, que exija alguma necessidade especial.
cm_pessoa	Contém os dados pessoais do aluno, como nome, CPF etc. Será utilizada apenas para interligar as tabelas.
cm_pessoa_fisica	Contém outros dados pessoais do aluno, como data de nascimento, etnia, cidade onde nasceu etc.
cm_unidade	Cadastrado das unidades dos *campi* do IFTM.

Fonte: ERP-IFTM.

O *schema* mais importante para a realização dessa investigação é o MAC, que contém ao todo 150 tabelas, das quais foram utilizadas 13, as mais relevantes e indispensáveis. Na tabela 6 são descritas as tabelas de dados.

Tabela 6. Tabelas do Banco de Dados do Módulo Acadêmico.

Tabela	Descrição
aluno	Cadastro do aluno quando ingressa no IFTM.
aluno_escolaridade	O nível de escolaridade do aluno, ensino fundamental, médio, técnico, superior etc.
aluno_ing	Registra os dados de matrícula em cada curso que o aluno estudar.
area_conhecimento	A área de conhecimento do curso, como Ciências Agrárias, Gestão e Negócios, Informação e Comunicação etc.
grade_horaria	Utilizada para saber a duração das aulas em minutos.
mat_cur_ofe	Matriz curricular que é ofertada ao aluno. Contém todas as disciplinas do curso do aluno.
mat_cur_ofe_turno	O período do dia que é oferecido o curso, como matutino, vespertino, noturno etc.
movimento	Contém os tipos de movimentações que podem ocorrer com o aluno como matriculado, formado, desistência etc.
periodo_letivo	O período letivo do curso, se anual, semestral etc.
ppc	Projeto pedagógico de curso, onde estão registrados todos os cursos do IFTM.
ppc_modalidade	Informa a modalidade do curso: presencial ou EaD.
ppc_nivel_categoria_forma	O nível, a categoria ou a forma do curso, como bacharelado, tecnólogo e licenciatura.
tipo_periodo	Indica como está dividido o período letivo do curso, se 1

Tabela	Descrição
	semestre, 2 semestres etc.

Fonte: ERP-IFTM.

3.2.2 Coleta de dados inicial

Com o objetivo de determinar a causa da evasão escolar, foi feito o levantamento das informações no banco de dados do ERP-IFTM, que estão alinhados com o objetivo da pesquisa.

Foi criado um script SQL (*Structered Query Language* - Linguagem de Consulta Estruturada) relacionando todas as tabelas dos *schemas Public* e MAC, citadas anteriormente, para extrair os dados que serão minerados.

Com o propósito de antecipar problemas de qualidade dos dados e apresentar suas soluções, alguns dados extraídos foram corrigidos durante a própria execução dos comandos SQL, como por exemplo:

- A escolaridade do aluno com o valor "ensino médio - supletivo" e "*ensimo* médio", nesses casos com erro de ortografia, foram corrigidos para "ensino médio";
- A cidade onde o aluno nasceu com o valor "São *Apulo*" para "São Paulo" e "*Uebraba*" para "Uberaba", ambos com erros ortográficos;
- Foi alterada a data de nascimento informada errada no cadastro, por exemplo, nascimento em "29/03/2019" alterado para "29/03/1989";
- Valores de horas armazenados no formato decimal foram convertidos para minutos antes de serem somados e, em seguida, convertidos para horas novamente, por exemplo: a soma das faltas em horas 10,5 + 10,5 não poderia exibir 21 horas, o correto são 20 horas e 10 minutos;

Alguns dados extraídos são resultados de cálculos realizados durante a própria execução dos comandos SQL, como por exemplo:

- Calcular a idade que o aluno tinha quando ingressou no curso é a diferença em anos entre as datas de ingresso e de nascimento;
- Calcular a soma das faltas e frequências de todas as disciplinas do aluno;
- Calcular a soma das notas e das cargas horárias de todas as disciplinas do aluno;
- Calcular a quantidade de disciplinas que o aluno foi: aprovado, reprovado, dispensado etc.

3.2.3 O arquivo alunos.arff

O Formato de Arquivo de Atributo-Relação (ARFF) é um método para carga de dados no WEKA, que possibilita definir os tipos de dados que serão carregados e os seus valores.

Compõe-se de duas seções: o cabeçalho, que contêm um nome para a base de dados, uma lista das variáveis e seus tipos de dados, que podem ser: nominal, numérico, *string* (valores de texto arbitrário) e data; os dados, onde cada linha representa uma instância na base dados, com os valores das variáveis separados por vírgulas. Os valores faltosos (*missing values*) são representados pelo caractere "?".

O arquivo "alunos.arff" possui 19 variáveis nominais, 20 numéricas e a variável classe, totalizando 40 atributos, que são detalhados a seguir.

3.2.4 O significado das variáveis

Variáveis nominais:

- ds_sexo – indica o sexo do aluno, "m" para o sexo masculino e "f" para o feminino;
- estado_civil – estado civil do aluno;
- escolaridade – nível de escolaridade do aluno;
- nacionalidade – nacionalidade do aluno;
- ds_naturalidade_cidade – cidade que o aluno nasceu;
- ds_naturalidade_estado – estado onde o aluno nasceu;
- etnia – etnia do aluno;
- necessidade_especial – indica se o aluno é possui alguma deficiência física ou mental;
- bl_dilacao – indica se o aluno obteve o benefício de prorrogar o prazo de conclusão de seus estudos;
- bl_ensino_medio_publico – indica se o aluno estudou no ensino médio público, que equivale à última fase da educação básica, ou seja, a continuidade da educação fundamental;
- bl_aluno_especial – indica se o aluno necessita de educação especial;
- campus – campus de matrícula do aluno;
- cidade_campus – cidade do campus de matrícula do aluno;
- curso – curso que o aluno se matriculou;
- area_conhecimento – agrupa os cursos do IFTM em áreas de conhecimento;
- ppc_nivel_categoria_forma – indica o nível, a categoria ou a forma de estudos de cursos superiores;
- ds_nome_turno – período do dia que o aluno estuda;
- periodo_letivo – indica como está dividido o período letivo do curso;
- ano_ingresso – ano de ingresso do aluno.

Variáveis numéricas:

- id – representa a identificação de cada matrícula do aluno nos cursos oferecidos pelo IFTM. Nesse caso, poderá haver mais de um "id" para o mesmo aluno se ele estiver matriculado em mais de um curso na instituição. Dessa forma, está sendo investigada a existência ou não da evasão em todas as matrículas dos alunos.
- idade_ingresso – idade do aluno no seu ano de ingresso, ou seja, quando se matriculou no curso;
- no_carga_horaria_minima – carga horária mínima que o aluno deverá estudar para concluir o seu curso;
- no_carga_horaria_maxima – carga horária máxima que será permitida ao aluno estudar até concluir o seu curso;
- duracao_aula_no_minutos – duração em minutos das aulas dos alunos;
- soma_no_total_nota – soma das notas de todas as disciplinas que o aluno já foi avaliado no seu curso;
- soma_total_faltas_horas – soma de todas as faltas que o aluno obteve nas aulas do seu curso, calculadas em horas;
- soma_no_total_faltas – soma de todas as faltas que o aluno obteve nas aulas do seu curso, calculadas em quantidade de aulas;
- soma_carga_horaria – soma da carga horária das aulas que o aluno assistiu. Apresenta 2.422 valores distintos;
- soma_total_frequencia_horas – soma de todas as frequências que o aluno obteve, ou seja, a carga horária de aulas que assistiu durante o seu curso, calculadas em horas;
- soma_no_total_frequencia – soma de todas as frequências que o aluno obteve nas aulas do seu curso, calculadas em quantidade de aulas;
- qtd_disciplina_enriquecimento_curricular – quantidade de disciplinas que o aluno estudou, além das disciplinas obrigatórias e optativas do seu curso, que enriqueceram o seu currículo escolar;
- qtd_disciplina_dependencia – quantidade de disciplinas que o aluno ficou de dependência, ou seja, não conseguiu aprovação durante o período acadêmico regular;
- qtd_disciplina_aprovado – quantidade de disciplinas em que o aluno foi aprovado;
- qtd_disciplina_recuperacao – quantidade de disciplinas em que o aluno ficou para recuperação;

- qtd_disciplina_reprovado – quantidade de disciplinas em que o aluno foi reprovado;
- qtd_disciplina_reprovado_em_dependencias – quantidade de disciplinas em que o aluno foi reprovado em dependências, ou seja, não conseguiu ser aprovado mesmo tendo uma segunda oportunidade para recuperar seus estudos e obter a aprovação;
- qtd_disciplina_reprovado_por_infrequencia – quantidade de disciplinas em que o aluno foi reprovado por infrequência, ou seja, não conseguiu ser aprovado porque obteve uma quantidade de faltas acima da permitida;
- qtd_disciplina_dispensado – quantidade de disciplinas em que o aluno foi dispensado, o que é muito comum quando o aluno realizou um curso anterior e que continha uma ou mais disciplinas do curso atual;
- qtd_disciplina_transferido – quantidade de disciplinas transferidas do aluno.

3.2.5 Variável objetivo - classe

De acordo com as informações fornecidas pelo Weka:

Tipo de variável (*Type*): neste caso, a variável é nominal, ou seja, não numérica.

Valores em falta (*Missing*), ou seja, alunos para os quais não existe informação sobre a evasão: neste caso, a variável tem 0% de valores em falta, que indica que o valor da classe é conhecido para todos os alunos.

Valores distintos (*Distinct*): ou seja, quais valores a variável pode possuir, neste caso, existem somente dois valores possíveis, que indicam a situação do aluno quanto à evasão escolar: “true” para os alunos que evadiram dos seus cursos e “false” para os que não evadiram. Tem-se o quantitativo de 2.056 alunos que evadiram contra 2.081 que não evadiram.

Valores únicos (*Unique*): ou seja, que aparece uma única vez, nesse caso há 0% de valores únicos, o que é interessante porque uma variável nominal deve ter poucos únicos, pela razão de facilitar a descoberta de padrões, isto é, redundância, que são as repetições de dados.

3.2.6 Variáveis nominais

O arquivo “alunos.arff” possui 19 variáveis nominais, que são analisadas a seguir: ds_sexo, estado_civil, escolaridade, nacionalidade, ds_naturalidade_cidade, ds_naturalidade_estado, etnia, necessidade_especial, bl_dilacao, bl_ensino_medio_publico, bl_aluno_especial, campus, cidade_campus, curso, area_conhecimento, ppc_nivel_categoria_forma, ds_nome_turno, periodo_letivo e ano_ingresso.

Variável ds_sexo

Essa variável é nominal, tem 2 valores distintos: “m” para o sexo masculino e “f” para o feminino. Tem-se mais pessoas do sexo masculino do que do sexo feminino, ao todo são 2.773 alunos e 1.364 alunas.

Verifica-se que essa variável possui valores bons para uma variável nominal, sendo 0% de valores em falta e 0% de únicos, além disso, não possui nenhuma anomalia, como por exemplo, um valor diferente de “m” ou “f”.

Variável estado_civil

Apresenta 6 valores distintos para o estado civil do aluno, a saber: solteiro (3.476), casado (476), união estável (68), divorciado (62), separado judicialmente (13) e viúvo (2). A grande maioria, 3.476 alunos, possui o estado civil “solteiro”, as demais 5 situações juntas totalizam 621 alunos.

Essa variável possui apenas 1% (40 alunos) em falta, ou seja, que não se sabe o estado civil, por esse motivo será aplicado o filtro “*ReplaceMissingValues*” na etapa de preparação dos dados, que substituirá os valores em falta pela moda. Apresenta 0% de valores únicos, o que é bom para variáveis do tipo nominal, além disso, não se verifica outro qualquer tipo de anomalias.

Variável escolaridade

Essa variável contém 8 valores distintos para o nível de escolaridade do aluno, temos: ensino médio (358), ensino fundamental (311), técnico (7), superior bacharelado (6), magistério (4), superior licenciatura (4), superior tecnólogo (3) e ensino superior (2).

A variável possui 0% de valores únicos, o que é bom, porém, existe um grande número de alunos que não se tem a informação do nível de escolaridade: 3.442 (83%), o que não é bom. Sendo assim, essa variável será eliminada.

Variável nacionalidade

Apresenta 3 valores distintos para a nacionalidade do aluno, a saber: brasileiro nato (3.931), brasileiro naturalizado (8) e estrangeiro (4). Constata-se que, quase em sua totalidade, os alunos são brasileiros natos.

Há 5% (194 alunos) em falta, ou seja, que não se sabe nacionalidade, sendo assim, será aplicado o filtro “*ReplaceMissingValues*” na etapa de preparação dos dados, que substituirá os valores em falta pela moda. Tem-se 0% de valores únicos e, além disso, não se apresenta nenhum outro tipo de anomalias.

Variável ds_naturalidade_cidade

Essa variável possui apenas 7 (0,17%) de valores em falta, o que não é tão ruim, porém, contém 526 valores distintos, que são muitos por se tratar de uma variável nominal e

são 308 (7%) de valores únicos, o que é muito e também não é bom. Sendo assim, essa variável será eliminada.

Variável ds_naturalidade_estado

Constata-se que a maior parte dos alunos nasceu no estado de Minas Gerais (3.264), seguido de São Paulo (417) e depois de Goiás (140), o que não é uma surpresa, pois, o primeiro é o estado onde se localiza fisicamente o IFTM e os últimos são estados vizinhos, os demais estados possuem valores abaixo de 50 cada.

Apresentam 26 valores distintos, o que indica que em apenas um dos estados brasileiros não ocorreu o nascimento dos alunos do IFTM, que é o estado do Acre. São 8 (0,19%) os valores em falta, ou seja, que não se sabe em que estado o aluno nasceu, por isso, será aplicado o filtro "*ReplaceMissingValues*", que substituirá os valores em falta pela moda. Além disso, não se apresenta nenhum outro tipo de anomalias.

Variável etnia

Apresenta 6 valores distintos para a etnia do aluno, a saber: branca (1.577), parda (1.166), não informada (858), que não se deve confundir com valores faltosos, negra (295), amarela (38) e indígena (10). A maior parte é das etnias branca e parda, que somados são 2.743 alunos, contra 343 das demais que foram informadas.

Essa variável possui 5% (193 alunos) em falta, ou seja, que não se sabe a etnia, por esse motivo será aplicado o filtro "*ReplaceMissingValues*" na etapa de preparação dos dados, que substituirá os valores em falta das variáveis nominais pela moda. Apresenta 0% de valores únicos, além disso, não se apresenta nenhum outro tipo de anomalias.

Variável necessidade_especial

Essa variável contém 12 valores distintos, que são: nenhuma (4.106), deficiência auditiva (8), deficiência física (8), deficiência mental (3), mobilidade reduzida (3), deficiência visual (2), portador de baixa visão (2), nanismo (1), ostomia (1), deformidade congênita ou adquirida (1), portador de ocorrência visual simultânea (1) e mobilidade reduzida permanente ou temporária (1). A grande maioria não possui nenhuma necessidade especial (4.106 alunos), contra as demais somadas (31), que representa menos de 1%.

Apesar de apresentar 0% de valores em falta (nenhum aluno), a maioria (99%) não possui nenhuma necessidade especial. Além disso, entre os que possuem alguma necessidade especial, têm-se 5 valores únicos e outros quase únicos (2 a 3 alunos), ou seja, há pouca redundância, o que não é bom para a mineração de dados. Sendo assim, não é interessante investigar essa variável, por isso, será eliminada.

Variável bl_dilacao

Essa variável possui apenas duas possibilidades de valores: “não” quando o aluno não obteve adiamento do prazo para conclusão dos seus estudos e “sim” que obteve. Do total de alunos (4.137), consta que apenas 1 recebeu a prorrogação do prazo.

Apesar de essa variável apresentar 0% de valores em falta, possui 1 valor único, onde se pode ter apenas 2 valores, indicando que não é interessante investigá-la na mineração de dados, portanto, será eliminada.

Variável bl_ensino_medio_publico

Essa variável possui apenas duas possibilidades de valores: “não” quando o aluno não estudou no ensino médio público, que representa a maioria (3.073) e “sim” quando cursou o ensino médio em escola pública, sendo a minoria (1.064), apenas 25% do total de alunos.

Consta 0% de valores em falta e 0% de valores únicos, além disso, não se verifica qualquer outro tipo de anomalias.

Variável bl_aluno_especial

Essa variável possui apenas duas possibilidades de valores: “não” quando a educação do aluno não demanda atenção especial e “sim” quando requer. Todos os alunos (4.137) não necessitam de educação especial.

Apesar de essa variável apresentar 0% de valores em falta e 0% de valores únicos, possui apenas um valor distinto, que a tornar dispensável para a mineração de dados, já que não há nenhuma redundância, portanto, será eliminada.

Variável campus

No período delimitado da pesquisa (2012 a 2016), 7 *campi* distintos tiveram alunos matriculados nos cursos superiores (tecnologia, licenciaturas e bacharelados), a saber: Campus Uberlândia Centro (970), Campus Uberaba (943), Campus Avançado Uberaba Parque Tecnológico (517), Campus Patrocínio (516), Campus Uberlândia (486), Campus Ituiutaba (414) e Campus Paracatu (291). Constata-se que 2 *campi* do IFTM não estão inclusos por não terem oferecido curso superior até o último ano dessa pesquisa (2016), são eles: Campus Avançado Campina Verde e Campus Patos de Minas.

A variável apresenta 0% de valores em falta e 0% de valores únicos, além disso, não se apresenta nenhum tipo de anomalias.

Variável cidade_campus

Dos 7 *campi* que fazem parte dessa pesquisa, são 5 os valores distintos, o seja, 5 cidades onde os alunos estudam: Uberaba (1.460), Uberlândia (1.456), Patrocínio (516), Ituiutaba (414) e Paracatu (291). O maior número de alunos estuda nas cidades de Uberaba e

Uberlândia (2.916), que foram as duas primeiras a fazerem parte da história da criação do IFTM.

Apresenta 0% de valores em falta e 0% de valores únicos, além disso, não consta nenhum tipo de anomalias.

Variável curso

Dos cursos superiores com alunos matriculados, existem 9 cursos com o quantitativo entre 203 a 297 alunos, 8 entre 185 a 102 e 10 entre 12 a 99.

Verifica-se 0% de valores em falta e 0% de valores únicos, além disso, não se verifica qualquer tipo de anomalias.

Variável area_conhecimento

Possui 13 valores distintos, ou seja, são 13 as áreas de conhecimento que os alunos estudam, a saber: Informação e Comunicação (841), Ciências Agrárias (667), Gestão e Negócios (633), Ciência da Computação (408), Informática (374), Produção Alimentícia (250), Ciências Biológicas (168), Ciências e Tecnologia de Alimentos (161), Engenharia / Tecnologia / Gestão (155), Controle e Processos Industriais (102), Ciências Exatas e da Terra (99), Engenharia Elétrica (68) e Ciências Sociais Aplicadas (62). Salienta-se que nem todos os campi possuem todas as áreas e cursos que a pesquisa abrange.

Verifica-se 4% (149) de valores em falta, ou seja, que não se sabe a área de conhecimento do curso do aluno, sendo assim, será aplicado o filtro "*ReplaceMissingValues*". Tem-se 0% de valores únicos, além disso, não se verifica qualquer outro tipo de anomalias.

Variável ppc_nivel_categoria_forma

A variável apresenta 3 valores distintos para o nível, a categoria ou a forma de estudos de cursos superiores, a saber: tecnológico (2.522), bacharelado (989) e licenciatura (626). Destaca-se a forma de curso tecnológico pelo fato do IFTM oferecer vários cursos tecnológicos deste a sua criação.

Apresenta 0% de valores em falta e 0% de valores únicos, além disso, não se verifica qualquer tipo de anomalias.

Variável ds_nome_turno

Apresenta 5 valores distintos, que são: noturno (2.260), diurno (921), matutino (629), multiperiódico (222) e vespertino (97). Observa-se uma maior concentração de alunos no período noturno, o que pode supor que a maioria trabalha durante o dia.

Apresenta 0,19% (8) de valores em falta, ou seja, que não se sabe o período do dia que o aluno estuda, sendo assim, será aplicado o filtro "*ReplaceMissingValues*". Tem-se 0% de valores únicos, além disso, não se verifica qualquer outro tipo de anomalias.

Variável periodo_letivo

Contém 2 valores distintos, dessa forma, a apuração das notas e das frequências dos alunos podem ocorrer nos seguintes períodos: 1 semestre, que representa a maioria (3.036) e 2 semestres, que representa a minoria (1.089), apenas 26% do total de alunos.

Apresenta 0,29% (12) de valores em falta, ou seja, não se sabe como o período letivo do curso do aluno, sendo assim, será aplicado o filtro "*ReplaceMissingValues*". Tem-se 0% de valores únicos e, além disso, não se apresenta nenhum outro tipo de anomalias.

Variável ano_ingresso

Apresenta 5 valores distintos para o ano que os alunos iniciaram o seu curso, na seguinte ordem cronológica: 2012 (763), 2013 (680), 2014 (789), 2015 (790) e 2016 (1.115). Conforme citado no item "2.2 Delimitação do Tema", foram selecionados os alunos que se matricularam nos últimos 5 anos completados: 2012 a 2016. À exceção do ano de 2013, observa-se um aumento no número de alunos ingressantes a cada ano, principalmente, em 2016, quando ocorreu um crescimento de 41% em relação ao ano anterior.

Consta 0% de valores em falta e 0% de valores únicos, além disso, não se verifica qualquer outro tipo de anomalias.

3.2.7 Variáveis numéricas

O arquivo "alunos.arff" possui 20 variáveis numéricas, que são analisadas a seguir: id, idade_ingresso, no_carga_horaria_minima, no_carga_horaria_maxima, duracao_aula_no_minutos, soma_no_total_nota, soma_total_faltas_horas, soma_no_total_faltas, soma_carga_horaria, soma_total_frequencia_horas, soma_no_total_frequencia, qtd_disciplina_enriquecimento_curricular, qtd_disciplina_dependencia, qtd_disciplina_aprovado, qtd_disciplina_recuperacao, qtd_disciplina_reprovado, qtd_disciplina_reprovado_em_dependencias, qtd_disciplina_reprovado_por_infrequencia, qtd_disciplina_dispensado e qtd_disciplina_transferido.

Variável id

A variável tem 100% (4.137) de valores distintos e 100% de valores únicos (4.137), ou seja, não possui nenhuma redundância. Essa é uma situação que indica desinteresse para a mineração de dados, portanto, a variável será eliminada.

Variável idade_ingresso

Essa variável apresenta 50 valores distintos, o que não é relevante para variáveis numéricas. As medidas estatísticas mais comuns para a idade de ingresso são: mínimo de 16

anos, máximo de 70, média de 23.766 e desvio padrão de 7.812. Mostra que o aluno mais novo ingressou aos 16 anos e o mais velho aos 70.

Verifica-se 0% de valores em falta. Além disso, existem 0,15% (6) valores únicos, o que não é relevante no caso de variáveis numéricas. Não se verifica qualquer outro tipo de anomalias, como por exemplo, um valor muito baixo ou até mesmo negativo para a idade, que indicaria que não se trata de um aluno de curso superior ou até mesmo o absurdo de que ele ainda "não teria nascido".

Variável no_carga_horaria_minima

Essa variável possui 28 valores distintos. As medidas estatísticas mais comuns para a carga horária mínima são: mínima de 3.03 horas, máximo de 4.384, média de 1.832,627 e desvio padrão de 841,931.

A variável possui 0% de valores em falta e 0% de valores únicos. O valor mínimo (3.03) é muito baixo e pode indicar um erro de informação, porém a média ainda permaneceu razoável (2.832). Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável no_carga_horaria_maxima

Essa variável possui 5 valores distintos. As medidas estatísticas mais comuns para a carga horária máxima são: mínima de 0 hora, máximo de 4.384, média de 772.972 e desvio padrão de 1.625,27.

A variável possui 54% (2.233) de valores em falta, ou seja, mais da metade não se sabe qual a carga horária máxima permitida ao aluno estudar. Tem-se 0% de valores únicos. O valor mínimo (0), certamente, indica um erro de informação, além disso, existe uma quantidade enorme de valores em falta, portanto, a variável será eliminada.

Variável duracao_aula_no_minutos

Essa variável apresenta 4 valores distintos. As medidas estatísticas mais comuns para a duração das aulas são: mínimo de 40 minutos, máximo de 50, média de 45,02 e desvio padrão de 3,613. Mostra que o aluno que tem a aula mais curta estuda 40 e aquele com a aula mais longa 50 minutos.

Apresenta apenas 0,19% (8) em falta, ou seja, que não se sabe a duração em minutos das aulas do aluno, por esse motivo será aplicado o filtro "*ReplaceMissingValues*" na etapa de preparação dos dados, que substituirá os valores em falta de variáveis numéricas pela média. Tem-se 0% de valores únicos e, além disso, não se apresenta nenhum outro tipo de anomalias.

Variável soma_no_total_nota

Essa variável apresenta 3.248 valores distintos, como já dito antes, não é relevante para variáveis numéricas. As medidas estatísticas mais comuns para a soma total das notas são: mínimo de 0 pontos, máximo de 7.208,51, média de 1.201,819 e desvio padrão de

1.355,072. Os valores muito baixos podem indicar que o aluno está no início do seu curso e os muito altos que já esteja no final ou até mesmo concluído.

Verifica-se 0% de valores em falta e 77% (3.180) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável soma_total_faltas_horas

Essa variável apresenta 2.313 valores distintos. As medidas estatísticas mais comuns para as faltas dos alunos são: mínimo de 0 hora, máximo de 2.428,25, média de 226,147 e desvio padrão de 230,366. Têm-se alunos com pouca ou nenhuma falta, assim como, alunos com muitas faltas.

A variável possui 0% de valores em falta e 38% (1.579) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável soma_no_total_faltas

Essa variável apresenta 881 valores distintos. As medidas estatísticas mais comuns para a quantidade de faltas dos alunos são: mínimo de 0 falta, máximo de 3.147, média de 259,362 e desvio padrão de 291,507. Têm-se alunos com pouca ou nenhuma falta, assim como, alunos com muitas faltas.

A variável possui 0% de valores em falta e 7% (274) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável soma_carga_horaria

Essa variável apresenta 2.422 valores distintos. As medidas estatísticas mais comuns para carga horária das aulas dos alunos são: mínimo de 0 hora, máximo de 6.065,8, média de 1.328,011 e desvio padrão de 1.157,479. Os valores muito baixos podem indicar que o aluno está no início do seu curso e os muito altos que já esteja no final ou até mesmo concluído.

Verifica-se 0% de valores em falta e 50% (2.062) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável soma_total_frequencia_horas

Essa variável apresenta 2.244 valores distintos. As medidas estatísticas mais comuns para as frequências dos alunos são: mínimo de 0 hora, máximo de 1.694,93, média de 323,405 e desvio padrão de 378,739. Os valores muito baixos podem indicar que o aluno está no início do seu curso e os muito altos que já esteja no final ou até mesmo concluído.

A variável possui 0% de valores em falta e 51% (2.107) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável soma_no_total_frequencia

Essa variável apresenta 1.992 valores distintos. As medidas estatísticas mais comuns para as frequências dos alunos são: mínimo de 0 aula, máximo de 5.091, média de 998,347 e

desvio padrão de 1.023,298. Os valores muito baixos podem indicar que o aluno está no início do seu curso e os muito altos que já esteja no final ou até mesmo concluído.

A variável possui 0% de valores em falta e 26% (1.065) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável qtd_disciplina_enriquecimento_curricular

Essa variável apresenta 3 valores distintos. As medidas estatísticas mais comuns para a quantidade de disciplinas de enriquecimento curricular são: mínimo de 0 disciplina, máximo de 2, média de 0,007 e desvio padrão de 0,105. Os valores muito baixos, tanto para mínimo e máximo, de uma forma geral, indicam que os alunos não têm muito interesse por esse tipo de disciplina.

A variável possui 0% de valores em falta e 0% de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável qtd_disciplina_dependencia

Essa variável apresenta 27 valores distintos. As medidas estatísticas mais comuns para a quantidade de disciplinas de dependência são: mínimo de 0 disciplina, máximo de 28, média de 1,492 e desvio padrão de 3,097. Têm-se alunos com pouca ou nenhuma disciplina de dependência, assim como, alunos com muitas.

A variável possui 0% de valores em falta e 0% de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável qtd_disciplina_aprovado

Essa variável apresenta 78 valores distintos. As medidas estatísticas mais comuns para a quantidade de disciplinas aprovadas são: mínimo de 0 disciplina, máximo de 79, média de 13,567 e desvio padrão de 16,424. Têm-se alunos com pouca ou nenhuma disciplina aprovada, talvez por estar no início do seu curso ou ter várias disciplinas reprovadas.

A variável possui 0% de valores em falta e 0,15% (6) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável qtd_disciplina_recuperacao

Essa variável apresenta 4 valores distintos. As medidas estatísticas mais comuns para a quantidade de disciplinas de recuperação são: mínimo de 0 disciplina, máximo de 3, média de 0,026 e desvio padrão de 0,19. Pode-se constatar que, mesmo com o máximo de 4, é baixa a quantidade de disciplina que os alunos ficaram de recuperação.

A variável possui 0% de valores em falta e 0% de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável qtd_disciplina_reprovado

Essa variável apresenta 24 valores distintos. As medidas estatísticas mais comuns para a quantidade de disciplinas reprovadas são: mínimo de 0 disciplina, máximo de 25, média de 1,851 e desvio padrão de 3,122. Têm-se alunos com pouca ou nenhuma disciplina reprovada, talvez por estar no início do seu curso ou ter várias disciplinas aprovadas. Os baixos valores para a média e o desvio padrão indicam que são poucos os alunos com o valor máximo de 25 disciplinas reprovadas.

A variável possui 0% de valores em falta e 0,02% (1) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável qtd_disciplina_reprovado_em_dependencias

Essa variável apresenta 2 valores distintos. As medidas estatísticas mais comuns para a quantidade de disciplinas reprovadas em dependências são: mínimo de 0 disciplina, máximo de 1, média de 0 e desvio padrão de 0,016.

Comparando-se com a variável qtd_disciplina_dependencia, citada anteriormente, onde se tem o valor máximo de 28 disciplinas que o aluno obteve dependência, contra o máximo de 1 disciplina reprovada em dependência, pode-se afirmar que, são poucos os alunos que são reprovados nas disciplinas de recuperação.

A variável possui 0% de valores em falta e 0,02% (1) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável qtd_disciplina_reprovado_por_infrequencia

Essa variável apresenta 38 valores distintos. As medidas estatísticas mais comuns para a quantidade de disciplinas reprovadas por infrequência são: mínimo de 0 disciplina, máximo de 49, média de 3,502 e desvio padrão de 4,894. Os baixos valores para a média e o desvio padrão em relação ao valor máximo, indicam que não são tantos os alunos que são reprovados em muitas disciplinas por infrequência.

A variável possui 0% de valores em falta e 0,10% (4) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável qtd_disciplina_dispensado

Essa variável apresenta 33 valores distintos. As medidas estatísticas mais comuns para a quantidade de disciplinas dispensadas são: mínimo de 0 disciplina, máximo de 49, média de 1,053 e desvio padrão de 3,732. Os baixos valores para a média e o desvio padrão em relação ao valor máximo, indicam que são poucos os alunos que são dispensados de disciplinas do curso.

A variável possui 0% de valores em falta e 0,10% (4) de valores únicos. Além disso, não se apresenta nenhum outro tipo de anomalia.

Variável qtd_disciplina_transferido

Essa variável apresenta 1 valor distinto. As medidas estatísticas mais comuns para a quantidade de disciplinas transferidas são: mínimo de 0 disciplina, máximo de 0, média de 0 e desvio padrão de 0, ou seja, nenhum aluno possui disciplina transferida.

Apesar de essa variável apresentar 0% de valores em falta e 0% de valores únicos, possui apenas um valor distinto, que a tornar dispensável para a mineração de dados, já que não há nenhuma redundância, portanto, será eliminada.

3.2.8 Variáveis correlacionadas

O Weka possui uma ferramenta de visualização que nos permite confirmar se existe correlação entre as variáveis numéricas. Ao utilizar a ferramenta são criadas imagens, onde se verifica se o diagrama de dispersão que foi criado é uma reta. Assim sendo, indica a correlação entre as variáveis.

Se existir correlações entre duas variáveis e essa correlação for igual a 1, significa que uma dessa variáveis é função linear da outra variável. Sendo assim, uma delas deverá ser eliminada, pois se tem a impressão de utilizar uma mesma variável duas vezes, dessa forma não é útil para a mineração de dados.

Após verificar os diagramas de dispersão gerados pelo Weka, observando todas as variáveis duas a duas, foram identificadas apenas duas possíveis correlações, indicadas a seguir.

A primeira correlação ocorre entre as variáveis soma_total_frequencia_horas e soma_no_total_frequencia, que correspondem à totalização das frequências, em horas e em quantidade de aulas, respectivamente. Porém, o diagrama de dispersão diverge de uma reta, conforme a figura 8.

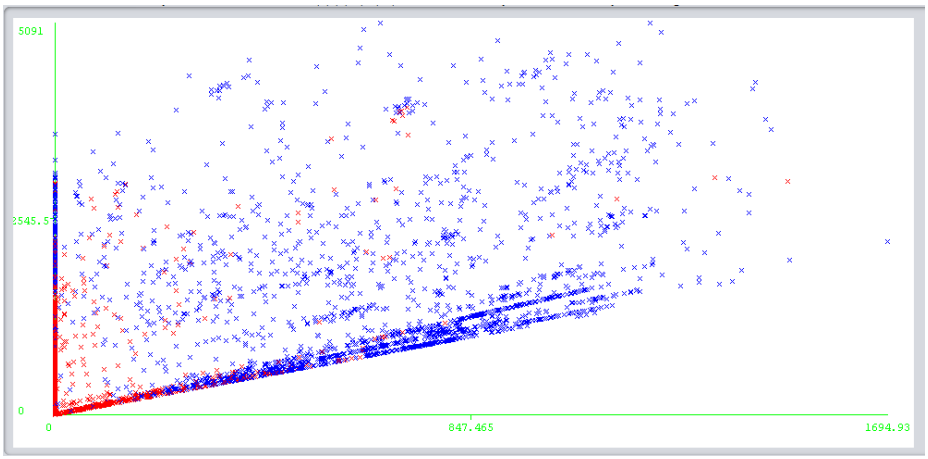

Figura 8. Diagrama de dispersão: soma_total_frequencia_horas x soma_no_total_frequencia.

A segunda correlação ocorre entre as variáveis soma_total_faltas_horas e soma_no_total_faltas, que correspondem à totalização das faltas, em horas e em quantidade

de aulas, respectivamente. Porém, o diagrama de dispersão diverge de uma reta, conforme a figura 9.

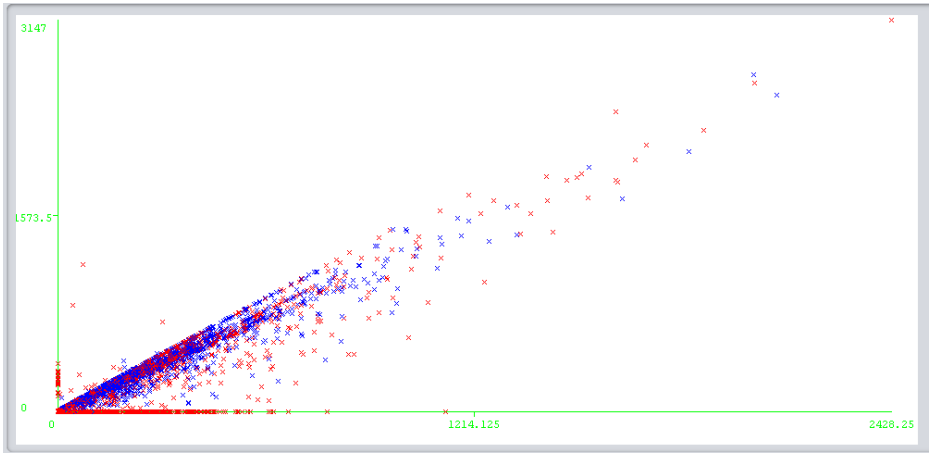

Figura 9. Diagrama de dispersão: soma_total_faltas_horas x soma_no_total_faltas.

Acredita-se que as correlações não foram detectadas em ambos os diagramas de dispersão por causa da ausência de lançamento de valores em um dos pares das variáveis no sistema.

Nesse sentido, conclui-se que não há correlações entre as variáveis do arquivo "alunos.arff", sendo assim, nenhuma variável a mais será eliminada.

3.3 Fase 3 – Preparação dos dados

Após a coleta, análise e exploração dos dados, avança-se para a fase de preparação dos dados, onde houve transformações nas variáveis para obter-se um conjunto de dados adequado para se executar os algoritmos de mineração de dados.

3.3.1.1 Remover as variáveis

Elimina-se as 8 variáveis que apresentaram anomalias na etapa de entendimento dos dados, são elas: escolaridade, ds_naturalidade_cidade, necessidade_especial, bl_dilacao, bl_aluno_especial, id, no_carga_horaria_maxima e qtd_disciplina_transferido. Dessa forma, fica-se com 32 variáveis.

3.3.1.2 Discretizar as variáveis

Os algoritmos habituais de classificação, como por exemplo os algoritmos de classificação JRip, que gera um modelo de classificação obtido por regras e o J48, que gera um modelo de classificação obtido por árvore, não se adéquam bem com valores de grandes amplitudes, inclusive, alguns deles trabalham somente com variáveis não numéricas. Sendo assim, foi necessário distribuir os dados das variáveis com muita amplitude em intervalos, ou

seja, discretizar para possibilitar sua utilização em determinados algoritmos de mineração de dados.

Foram discretizadas as seguintes variáveis: no_carga_horaria_minima, soma_no_total_nota, soma_total_faltas_horas, soma_no_total_faltas, soma_carga_horaria, soma_total_frequencia_horas e soma_no_total_frequencia.

3.3.1.3 Normalizar as variáveis

O agrupamento *k-means*, que é um método de *clustering*, ou seja, uma técnica que realiza agrupamentos de forma automática, foi utilizado com o objetivo de particionar os dados em grupos por aproximação, isto é, pela semelhança de acordo com os próprios dados.

Para possibilitar a utilização do algoritmo de mineração de dados de segmentação *k-means* nas próximas etapas, normalizam-se todas as variáveis numéricas para que apresentem valores entre 0 e 1.

3.4 Fase 4 - Construção dos modelos

Depois de realizada a preparação dos dados, apresenta-se a continuidade ao processo de mineração de dados com a construção dos modelos. Nesta fase, foram escolhidos e ajustados os parâmetros e aplicados os algoritmos de mineração de dados mais adequados ao propósito da nossa investigação, a fim de obter modelos eficientes.

Inicialmente, foram aplicados os algoritmos de classificação de regras JRip, modelo disponível no Anexo I, e de árvore J48, modelo disponível no Anexo II. Em seguida, utiliza-se o algoritmo de segmentação *k-means*.

3.4.1 Modelo de classificação de regras - JRip

Após a execução do algoritmo de classificação JRip, foi gerado um modelo de classificação obtido por regras. O modelo completo encontra-se no Anexo I. A seguir, tem-se um resumo das principais informações obtidas e, logo após, as regras concebidas:

As instâncias classificadas corretamente foram 3.882 (93,8361%) contra apenas 255 (6,1639%) de instâncias classificadas incorretamente.

Com relação ao desempenho dos classificadores de acordo com a métrica da Estatística de Kappa, o nível de concordância de classificação, ou seja, da coerência dos dados classificados é de: 0,8767, que representa uma ótima qualidade por estar próximo a 1.

O modelo gerado nos fornece uma quantidade de 10 regras, descritas e interpretadas a seguir:

1. (soma_total_frequencia_horas = '(-inf-29.94]') and (qtd_disciplina_aprovado <= 11) => evadido=True (1521.0/31.0)

Para um determinado aluno, se o valor da soma total da sua frequência, medida em horas, estiver abaixo ou igual a 29,94 horas e a quantidade de disciplinas aprovadas for menor ou igual a 11, então o aluno evade do seu curso. A regra acerta 1.521 vezes e erra apenas 31.

2. (qtd_disciplina_aprovado <= 6) and (soma_no_total_frequencia = '(-inf-243.5]') => evadido=True (224.0/20.0)

Para um determinado aluno, se a quantidade de disciplinas aprovadas for menor ou igual a 6 e a soma do número total de suas frequências estiver abaixo ou igual a 243 presenças, então o aluno evade do seu curso. A regra acerta 224 vezes e erra 20.

3. (soma_total_frequencia_horas = '(-inf-29.94]') and (qtd_disciplina_aprovado <= 21) => evadido=True (123.0/18.0)

Para um determinado aluno, se o valor da soma total da sua frequência, medida em horas, estiver abaixo ou igual a 29,94 horas e a quantidade de disciplinas aprovadas for menor ou igual a 21, então o aluno evade do seu curso. A regra acerta 123 vezes e erra 18.

4. (qtd_disciplina_aprovado <= 6) and (periodo_letivo = 1 semestre) and (soma_no_total_frequencia = '(243.5-389.5]') => evadido=True (56.0/4.0)

Para um determinado aluno, se a quantidade de disciplinas aprovadas for menor ou igual 6 e o período letivo for igual a 1 semestre e a soma do número total de suas frequências estiver entre 243 a 389 presenças, então o aluno evade do seu curso. A regra acerta 56 vezes e erra 4.

5. (soma_total_frequencia_horas = '(29.94-193.14]') and (qtd_disciplina_aprovado <= 13) and (periodo_letivo = 1 semestre) and (qtd_disciplina_reprovado >= 2) => evadido=True (19.0/1.0)

Para um determinado aluno, se o valor da soma total da sua frequência, medida em horas, estiver entre 29,94 a 193,14 horas e a quantidade de disciplinas aprovadas for menor ou igual a 13 e o período letivo for igual a 1 semestre e a quantidade de disciplinas reprovadas for maior ou igual 2, então o aluno evade do seu curso. A regra acerta 19 vezes e erra apenas 1.

6. (qtd_disciplina_aprovado <= 6) and (soma_total_frequencia_horas = '(29.94-193.14]') and (soma_no_total_faltas = '(228.5-inf)') => evadido=True (7.0/0.0)

Para um determinado aluno, se a quantidade de disciplinas aprovadas for menor ou igual 6 e o valor da soma total da sua frequência, medida em horas, estiver entre 29,94 a 193,14 horas e a soma do número total de suas faltas estiver acima de 228 faltas, então o aluno evade do seu curso. A regra acerta 7 vezes e não erra nenhuma.

7. (qtd_disciplina_reprovado_por_infrequencia >= 5) and (soma_carga_horaria = '(565.21-757]') and (ds_nome_turno = noturno) and (idade_ingresso <= 39) => evadido=True (15.0/0.0)

Para um determinado aluno, se a quantidade de disciplinas reprovadas por infrequência for maior ou igual 5 e a soma da carga horária estiver entre 565,21 a 757 horas e o turno for igual a noturno e a idade de ingresso for menor ou igual a 39 anos, então o aluno evade do seu curso. A regra acerta 15 vezes e não erra nenhuma.

8. (qtd_disciplina_reprovado_por_infrequencia >= 5) and (soma_total_frequencia_horas = '(-inf-29.94]') and (soma_no_total_faltas = '(228.5-inf)') => evadido=True (15.0/5.0)

Para um determinado aluno, se a quantidade de disciplinas reprovadas por infrequência for maior ou igual 5 e o valor da soma total da sua frequência, medida em horas, for inferior ou igual a 29,94 horas e a soma do número total de suas faltas estiver acima de 228 faltas, então o aluno evade do seu curso. A regra acerta 15 vezes e erra 5.

9. (qtd_disciplina_reprovado_por_infrequencia >= 8) and (curso = licenciatura em quimica) and (qtd_disciplina_dependencia <= 9) => evadido=True (5.0/0.0)

Para um determinado aluno, se a quantidade de disciplinas reprovadas por infrequência for maior ou igual a 8 e o curso for Licenciatura em Química e a quantidade de disciplinas em dependência for menor ou igual 9, então o aluno evade do seu curso. A regra acerta 5 vezes e não erra nenhuma.

10. => evadido=False (2152.0/150.0)

Indica que em todos os casos que não satisfazem nenhuma das regras anteriores, o aluno não evadiu, sendo que esta regra acerta 2.152 e erra 150 vezes.

3.4.2 Modelo de classificação de árvore - J48

Após a execução do algoritmo de classificação J48, utilizando-se os valores padrões dos parâmetros, foi gerado um modelo de classificação obtido por árvore. No entanto, por ter gerado uma árvore muito grande, dificultando a leitura, ajusta-se o parâmetro *confidenceFactor* para 0.01 e obtém-se uma árvore de tamanho 15 contendo 10 folhas, conforme figura 10.

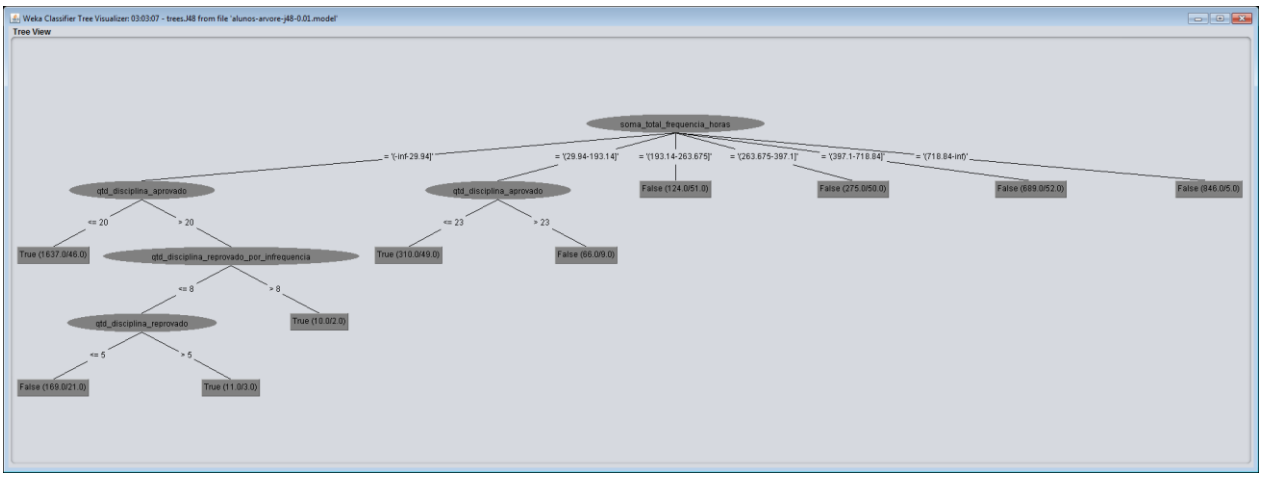

Figura 10. Visualização da árvore do algoritmo J48.

O modelo completo que foi produzido encontra-se no Anexo II. A seguir, tem-se um resumo das principais informações obtidas e, logo após, a árvore concebida:

As instâncias classificadas corretamente foram 3.833 (92,6517%) contra apenas 304 (7,3483%) de instâncias classificadas incorretamente.

O desempenho dos classificadores de acordo com a métrica da Estatística de Kappa é de 0,853, o que representa uma ótima qualidade por estar próximo a 1.

A seguir, realiza-se a leitura do modelo gerado pelo algoritmo J48, onde cada elipse representa um nó da árvore, que contém uma condição numa das variáveis. Percorrem-se todos os caminhos possíveis desde o primeiro nó, representado pela primeira elipse, ou seja, a raiz da árvore, até alcançarmos uma das folhas, representadas por retângulos, que contém um dos valores da variável objetivo: *True* ou *False*, dessa forma, descobrimos em que condições ocorreu a evasão do aluno.

Assim, procede-se à leitura de todas as 10 folhas da árvore:

1. soma_total_frequencia_horas = '(-inf-29.94]'
   | qtd_disciplina_aprovado <= 20: True (1637.0/46.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, for menor ou igual a 29,94 horas e a quantidade de disciplinas aprovadas for menor ou igual a 20, então o aluno evade do seu curso. A regra acerta com 1.637 alunos e erra 46 vezes.

2. soma_total_frequencia_horas = '(-inf-29.94]'
   | qtd_disciplina_aprovado > 20
   | | qtd_disciplina_reprovado_por_infrequencia <= 8
   | | | qtd_disciplina_reprovado <= 5: False (169.0/21.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, for menor ou igual a 29,94 horas e a quantidade de disciplinas aprovadas for maior do que 20 e a quantidade de disciplinas reprovadas por infrequência for menor ou igual a 8 e a quantidade

de disciplinas reprovadas for menor ou igual a 5, então o aluno não evade do seu curso. A regra acerta com 169 alunos e erra 21 vezes.

3. soma_total_frequencia_horas = '(-inf-29.94]'
    | qtd_disciplina_aprovado > 20
    | | qtd_disciplina_reprovado_por_infrequencia <= 8
    | | | qtd_disciplina_reprovado > 5: True (11.0/3.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, for menor ou igual a 29,94 horas e a quantidade de disciplinas aprovadas for maior do que 20 e a quantidade de disciplina reprovadas por infrequência for menor ou igual a 8 e a quantidade de disciplinas reprovadas for maior do que 5, então o aluno evade do seu curso. A regra acerta com 11 alunos e erra 3 vezes.

4. soma_total_frequencia_horas = '(-inf-29.94]'
    | qtd_disciplina_aprovado > 20
    | | qtd_disciplina_reprovado_por_infrequencia > 8: True (10.0/2.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, for menor ou igual a 29,94 horas e a quantidade de disciplinas aprovadas for maior do que 20 e a quantidade de disciplina reprovadas por infrequência for maior do que 8, então o aluno evade do seu curso. A regra acerta com 10 alunos e erra 2 vezes.

5. soma_total_frequencia_horas = '(29.94-193.14]'
    | qtd_disciplina_aprovado <= 23: True (310.0/49.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, estiver entre 29,94 a 193,14 horas e a quantidade de disciplinas aprovadas for menor ou igual a 23, então o aluno evade do seu curso. A regra acerta com 310 alunos e erra 49 vezes.

6. soma_total_frequencia_horas = '(29.94-193.14]'
    | qtd_disciplina_aprovado > 23: False (66.0/9.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, estiver entre 29,94 a 193,14 horas e a quantidade de disciplinas aprovadas for maior do que 23, então o aluno não evade do seu curso. A regra acerta com 66 alunos e erra 9 vezes.

7. soma_total_frequencia_horas = '(193.14-263.675]': False (124.0/51.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, estiver entre 193,14 a 263,675 horas, então o aluno não evade do seu curso. A regra acerta com 124 alunos e erra 51 vezes.

8. soma_total_frequencia_horas = '(263.675-397.1]': False (275.0/50.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, estiver entre 263,675 a 397,1 horas, então o aluno não evade do seu curso. A regra acerta com 275 alunos e erra 50 vezes.

9. soma_total_frequencia_horas = '(397.1-718.84]': False (689.0/52.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, estiver entre 397,1 a 718,84 horas, então o aluno não evade do seu curso. A regra acerta com 689 alunos e erra 52 vezes.

10. soma_total_frequencia_horas = '(718.84-inf)': False (846.0/5.0)

Se o valor da soma total da frequência de um determinado aluno, medida em horas, for maior do que 718,84 horas, então o aluno não evade do seu curso. A regra acerta com 846 alunos e erra apenas 5 vezes.

3.4.3 Modelo de segmentação - K-means

O algoritmo de segmentação *k-means* foi executado diversas vezes, ajustando os seus parâmetros para obter um resultado mais satisfatório aos objetivos pré-estabelecidos e melhorar a precisão ou facilitar a compreensão do conhecimento extraído. Como resultado, obtém-se um modelo de segmentos, que posteriormente foi utilizado para classificar.

A classificação é realizada de forma automática pelo algoritmo, sem nenhuma pré-classificação existente, ou seja, sem a necessidade da supervisão humana. Porém, o *k-means* não determina automaticamente o número de segmentos a serem gerados, por isso, foi necessário experimentar alguns valores até obtermos um modelo com um número menor de erros (*squared errors*). Também foi necessário informar atributos (variáveis) a serem ignorados.

Assim, procedem-se as experimentações:

1. Utilizando os valores padrões e ignorando somente o atributo classe, obtivemos:

   Number of iterations: 5

   Within cluster sum of squared errors: 26018.865277321376

Observa-se um *squared errors* muito grande, com valor acima de 26.000.

2. Utilizando os valores padrões e ignorando 21 atributos:

Observa-se que, conforme eram ignoradas as variáveis nominais, o erro diminuía. Assim como foi constatado nos modelos de classificação JRip e J48, as variáveis nominais não tem muita relevância para o resultado final, sendo assim, foram ignoradas a classe e as 14 variáveis nominais: ds_sexo, estado_civil, nacionalidade, ds_naturalidade_estado, etnia, bl_ensino_medio_publico, campus, cidade_campus, curso, area_conhecimento, ppc_nivel_categoria_forma, ds_nome_turno, periodo_letivo e ano_ingresso. Houve uma

redução no erro, aproximadamente, de 26.018 para 1.543, ou seja, uma melhora bastante significativa.

Em seguida, ignora-se 6 variáveis numéricas, que também não apresentaram grandes relevâncias durante as experimentações, porém com as suas eliminações houve uma redução no erro, aproximadamente, de 1.543 para 476, ou seja, uma diminuição considerável. As variáveis ignoradas foram: idade_ingresso, no_carga_horaria_minima, duracao_aula_no_minutos, soma_carga_horaria, soma_no_total_frequencia e soma_no_total_faltas. Essas duas últimas variáveis, apesar de não terem sido eliminadas na etapa anterior, ao se verificar as variáveis correlacionadas, serão ignoradas agora por não apresentarem alguma importância significativa. Ao ignorá-las, houve uma redução no erro, aproximadamente, de 581 para 476.

Em resumo, permaneceram as variáveis que informam os valores de frequência, nota e disciplinas cursadas pelo aluno. Obtêm-se os seguintes resultados:

Number of iterations: 19

Within cluster sum of squared errors: 476.5255685122146

3. Alterando os valores padrões:

Altera-se o número de segmentos gerados de 2 para 3, ou seja, muda-se o valor de *numClusters* = 3. Obtêm-se os seguintes resultados:

Number of iterations: 20

Within cluster sum of squared errors: 465.58358962226

Ao criar mais um segmento, reduziu-se pouco os erros, aproximadamente, de 476 para 465. Esse novo segmento foi gerado com apenas 17 alunos, o que não se apresentou como algo interessante para a investigação e não trouxe nenhuma melhoria para os resultados esperados. Sendo assim, define-se por ficar com apenas 2 segmentos.

O modelo completo que foi produzido encontra-se no Anexo III. Os segmentos que foram obtidos (*cluster centroids*) serão avaliados na próxima etapa do processo de mineração de dados.

3.5 Fase 5 - Avaliação

Nesta fase, avalia-se a qualidade e eficácia dos modelos, que foram produzidos na etapa anterior, ao verificar se os resultados obtidos alcançam o objetivo do negócio, ou seja, se são capazes de auxiliar na investigação do problema da evasão escolar no IFTM.

3.5.1 Qualidade dos modelos de classificação

Os modelos de classificação são avaliados com um conjunto de medidas como, por exemplo, os casos classificados corretamente e incorretamente, Estatística de Kappa, erros absolutos e relativos etc.

Os valores comparativos para os dois modelos, regras, que foi gerado na execução do algoritmo JRip e árvore, que foi gerado na execução do algoritmo J48, ambos na etapa anterior, são apresentados na tabela 7.

Tabela 7. Comparativo entre os modelos de regras e árvore.

Medida	Regras		Árvore	
Correctly Classified Instances	3882	93.8361 %	3833	92.6517 %
Incorrectly Classified Instances	255	6.1639 %	304	7.3483 %
Kappa statistic		0.8767		0.853
Mean absolute error		0.1011		0.1186
Root mean squared error		0.2354		0.2468
Relative absolute error		20.2292 %		23.7219 %
Root relative squared error		47.0753 %		49.3641 %
Total Number of Instances		4137		4137

O modelo de classificação de regras JRip possui o índice de confiança muito bom. Os alunos evadidos classificados corretamente somam 93,8361% e os incorretos apenas 6,1639%.

O modelo de classificação de árvore J48 também possui o índice de confiança muito bom, porém, um pouco abaixo do JRip. Os alunos evadidos classificados corretamente somam 92,6517% e os incorretos apenas 7,3483%.

Além disso, os erros absolutos 0,1011 e relativos 20,2292%, também são menores no caso das regras, pois, no caso da árvore têm-se 0,1186 e 23,7219%, respectivamente.

A estatística *kappa* é mais próxima de 1 nas regras, onde tem-se o valor 0,8767 contra 0,853 na árvore, o que indica o modelo de regras como sendo o mais estável.

3.5.2 Precisão detalhada por classes

Verifica-se a precisão detalhada por classes de ambos os modelos, a fim de constatar se todos os valores, exceto os falsos positivos (*FP Rate*), estão o mais próximo possível de 1, pois, quanto mais perto desse valor, melhor a qualidade dos modelos. Vejamos, a seguir, os valores para os dois modelos:

Modelo de regras - JRip:

```
=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure  MCC    ROC Area PRC Area Class
               0,947   0,071   0,931     0,947  0,939      0,877  0,949    0,928    False
               0,929   0,053   0,946     0,929  0,937      0,877  0,949    0,941    True
Weighted Avg.  0,938   0,062   0,938     0,938  0,938      0,877  0,949    0,934
```

Modelo de árvore - J48:

```
=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure  MCC    ROC Area PRC Area Class
               0,947   0,094   0,910     0,947  0,928      0,854  0,960    0,956    False
               0,906   0,053   0,944     0,906  0,925      0,854  0,960    0,948    True
Weighted Avg.  0,927   0,074   0,927     0,927  0,926      0,854  0,960    0,952
```

Classe não evadidos:

Verifica-se que para ambos os modelos os valores das classes são próximos, porém, há um pequeno aumento nos valores da classe dos não evadidos, exceto em *Precision* e *PRC Area*.

Classe evadidos:

Pode-se afirmar que a classe dos evadidos apresenta valores um pouco maiores no modelo de regras, exceto para *PRC Area*. Constata-se também que essa classe foi a que obteve uma melhor precisão em ambos os modelos, o que o torna melhor.

3.5.3 Matriz confusão

A matriz confusão exibe a distribuição dos registros em termos de suas classes atuais e de suas classes previstas. Isso indica a qualidade do modelo atual. A seguir, analisa-se a matriz confusão de ambos os modelos de classificação.

Modelo de regras - JRip:

No modelo de regras JRip, mostrado na tabela 8, as classes False e True são previstas. As classificações corretas são indicadas em negrito.

Tabela 8. Matriz confusão do modelo de regras gerado pelo algoritmo JRip.

	False (previsto)	True (previsto)	Total
False	**1.971**	110	2.081
True	145	**1.911**	2.056
Total	2.116	2.021	4.137

A matriz de confusão pode ser lida, na horizontal e na vertical.

Lendo a tabela na horizontal

Existem 2.081 alunos classificados na classe False, ou seja, o aluno não evadiu:

- 1.971 desses alunos estão corretamente classificados como não evadidos.
- 110 desses alunos estão incorretamente classificados como não evadidos.

Há 2.056 alunos na classe True, ou seja, o aluno evadiu:

- 145 desses alunos estão incorretamente classificados como evadidos.
- 1.911 desses alunos estão corretamente classificados como evadidos.

Lendo a tabela na vertical

Existem 2.116 alunos classificados na classe False, ou seja, o aluno não evadiu:

- 1.971 desses alunos estão corretamente classificados como não evadidos.
- 145 desses alunos estão incorretamente classificados como não evadidos.

Existem 2.021 alunos classificados na classe True, ou seja, o aluno evadiu:

- 110 desses alunos estão incorretamente classificados como evadidos.
- 1.911 desses alunos estão corretamente classificados como evadidos.

Modelo de árvore - J48:

No modelo de árvore J48, mostrado na tabela 9, as classes False e True são previstas. As classificações corretas são indicadas em negrito.

Tabela 9. Matriz confusão do modelo de regras gerado pelo algoritmo J48.

	False (previsto)	True (previsto)	Total
False	**1.971**	110	2.081
True	194	**1.862**	2.056
Total	2.165	1.972	4.137

Lendo a tabela na horizontal

Existem 2.081 alunos classificados na classe False, ou seja, o aluno não evadiu:

- 1.971 desses alunos estão corretamente classificados como não evadidos.
- 110 desses alunos estão incorretamente classificados como não evadidos.

Há 2.056 alunos na classe True, ou seja, o aluno evadiu:

- 194 desses alunos estão incorretamente classificados como evadidos.
- 1.862 desses alunos estão corretamente classificados como evadidos.

Lendo a tabela na vertical

Existem 2.165 alunos classificados na classe False, ou seja, o aluno não evadiu:

- 1.971 desses alunos estão corretamente classificados como não evadidos.
- 194 desses alunos estão incorretamente classificados como não evadidos.

Existem 1.972 alunos classificados na classe True, ou seja, o aluno evadiu:

- 110 desses alunos estão incorretamente classificados como evadidos.
- 1.862 desses alunos estão corretamente classificados como evadidos.

A matriz de confusão, em ambos os modelos, é calculada pela função de pesquisa classificação. Para avaliarmos a qualidade dos modelos, verifica-se a distribuição das instâncias em termos de suas classes atuais e de suas classes previstas.

Nesse caso, é mais importante que os verdadeiros negativos estejam mais próximos o possível de zero, do que os falsos negativos, ou seja, é mais relevante encontrarmos menos erros ao definir as regras pelas quais um aluno evade de seu curso, do que constatar menos erros ao defini-las para o aluno que não evade.

Observando-se a tabela 8, o modelo erra 110 vezes ao definir as regras pelas quais os alunos evadem dos seus cursos e erra 145 vezes ao fazê-lo para os que não evadem.

Observando-se a tabela 9, o modelo também erra 110 vezes ao definir as regras pelas quais os alunos evadem dos seus cursos e erra 194 vezes ao fazê-lo para os que não evadem.

A quantidade de erros ao analisar a evasão é a mesma para ambos os modelos, sendo assim, por cometer menos erros ao analisar a não evasão, pode-se afirmar que modelo de regras JRip é mais confiável.

Com relação à geração de falsos positivos (110 alunos classificados incorretamente como evadidos) e de falsos negativos (145 alunos classificados incorretamente como não evadidos) no caso do modelo de classificação de regras JRip, por representarem, respectivamente, apenas 5,29% e 7,05% da amostra de 4.137 alunos investigados, considera-se uma margem de erro aceitável.

3.5.4 O que nos dizem os modelos

Analisando ambos os modelos se verificou que:

- As variáveis que informam os valores de frequência, nota e disciplinas cursadas pelo aluno são mais significantes para determinar a evasão, apesar do modelo de regras incluírem outras variáveis. Observa-se que a variável soma_total_frequencia_horas está presente em quase todas as regras apresentadas para o JRip e foi evidenciada no J48.
- A soma total das frequências do aluno, medida em horas, indica um dos principais problemas porque o nó da árvore é formado por uma condição nessa variável, porém, só existem alunos evadidos quando a soma das suas frequências está abaixo de 193,14 horas. Sendo assim, alunos com frequências acima desse valor não apresentam evasão. Essa condição também se verifica no modelo de regras, onde está variável aparece em 5 das 10 regras.
- Não basta apenas verificar a frequência das aulas, medida em horas, para determinar se um aluno evadiu. No caso da árvore, essa variável está sempre relacionada com a quantidade de disciplinas aprovadas quando ocorre uma evasão. Por exemplo, a maior parte das evasões (1.637 casos identificados com 46 falhas) ocorre quando a frequência está abaixo de 29,94 horas e a quantidade de disciplinas aprovadas é menor ou igual a 20. Para o modelo de regras, a maior parte das evasões (1.521 casos identificados com 31 falhas) ocorre quando as frequências são inferiores a 29,94 horas e a quantidade de disciplinas aprovadas é menor ou igual a 11. Enfim, para ambos os modelos, a maioria

dos casos de evasão ocorrem quando se têm baixas frequências e poucas disciplinas aprovadas.

- No caso do modelo de regras, que contém regras que envolvem as frequências das aulas, medida em horas, quando está entre 29,94 a 193,14 horas, há apenas uma regra que não se relaciona com a quantidade de disciplinas aprovadas, para se relacionar com a quantidade de disciplinas reprovadas por infrequência, maior ou igual a 5, e a soma total de faltas, acima de 228,5. Sendo assim, o aluno evade se tiver baixa frequência e, consequentemente, diversas disciplinas reprovadas por infrequência e bastante faltas.
- Um grande número de disciplinas aprovadas, por exemplo, mais do que 20, não é garantia de que o aluno não irá evadir, pois, foram identificados no modelo de árvore que alunos nessa condição se evadiram porque a sua quantidade de disciplinas reprovadas por infrequência foi maior do que 8.
- Houve algumas poucas evasões para as situações onde a quantidade de disciplinas reprovadas por infrequência foi igual ou menor que 8, ao mesmo tempo que a quantidade de disciplinas reprovadas por nota foi maior do 5, de acordo com o modelo de árvore.
- No modelo de regras existem 3 regras que determinam 100% de evasão, todas relacionadas com a baixa frequência e/ou reprovação por infrequência:
    1. Se a quantidade de disciplinas aprovadas for menor ou igual 6, ou seja, baixa aprovação, e o valor da soma total da sua frequência, medida em horas, estiver entre 29,94 a 193,14 horas e a soma do número total de suas faltas estiver acima de 228 faltas, ou seja, baixa frequência: 7 alunos evadidos.
    2. Se a quantidade de disciplinas reprovadas por infrequência for maior ou igual 5 e a soma da carga horária estiver entre 565,21 a 757 horas e o turno for igual a noturno e a idade de ingresso for menor ou igual a 39 anos: 15 alunos evadidos. Essa regra é mais complexa, além da reprovação por infrequência, a evasão ocorre com alunos mais jovens que estudam no período noturno e a carga horária do curso está entre 565,21 a 757 horas.
    3. Se a quantidade de disciplinas reprovadas por infrequência for maior ou igual a 8 e o curso for Licenciatura em Química e a quantidade de disciplinas em dependência for menor ou igual 9: 5 alunos evadidos. Essa regra é mais específica, além de muita reprovação por infrequência, a evasão ocorre somente com os alunos do curso de Licenciatura em Química e que tenha no máximo 9 disciplinas dependentes.

3.5.5 Qualidade do modelo de segmentação

Inicialmente, identificam-se as características do centroide de cada um dos elementos.

O algoritmo de segmentação *k-means*, que foi aplicado na etapa anterior, nos retornou o resultado exibido na tabela 10:

Tabela 10. Segmentos obtidos pelo algoritmo *k-means*.

Attribute	Cluster#		
	Full Data	0	1
	(4137.0)	(2495.0)	(1642.0)
soma_no_total_nota	0.1667	0.0519	0.3412
soma_total_faltas_horas	0.0931	0.0903	0.0974
soma_total_frequencia_horas	0.1908	0.0604	0.389
qtd_disciplina_enriquecimento_curricular	0.0034	0.0004	0.0079
qtd_disciplina_dependencia	0.0533	0.0381	0.0763
qtd_disciplina_aprovado	0.1717	0.0435	0.3665
qtd_disciplina_recuperacao	0.0085	0.0056	0.013
qtd_disciplina_reprovado	0.074	0.0454	0.1176
qtd_disciplina_reprovado_em_dependencias	0.0002	0	0.0006
qtd_disciplina_reprovado_por_infrequencia	0.0715	0.0868	0.0482
qtd_disciplina_dispensado	0.0215	0.0167	0.0288

Da mesma forma que ocorreu na avaliação dos modelos de classificação, no modelo de segmentação *k-means* a variável que totaliza a frequência está plenamente correlacionada com o sucesso do aluno em ser aprovado nas disciplinas, sendo assim, um valor baixo para essa variável pode indicar uma grande possibilidade de evasão.

Verifica-se que no centroide do 1º segmento a **soma total das frequências é a mais baixa** (0.0604), possui menores quantidades de disciplinas em dependência (0.0381), recuperação (0.0056), reprovação (0.0454) e reprovação por dependências (0), entretanto, tem **maior reprovação por infrequência** (0.0868) e menores notas (0.0519), consequentemente, há uma **maior quantidade de disciplinas reprovadas** (0.0435).

No centroide do 2º seguimento, ao contrário, a **soma total das frequências é a mais alta** (0.389) e, apesar de ter maiores quantidades de disciplinas em dependência (0.0763), recuperação (0.013), reprovação (0.1176) e reprovação por dependências (0.0006), contudo, tem **menor reprovação por infrequência** (0.0482) e maiores notas (0.3412), consequentemente, há uma **maior quantidade de disciplinas aprovadas** (0.3665).

Destaca-se a seguinte constatação: quanto maior a frequência, menor a reprovação por infrequência e maior a quantidade de disciplinas aprovadas. O contrário também é verdadeiro, quanto menor a frequência, maior a reprovação por infrequência e menor a quantidade de disciplinas aprovadas.

Em seguida, verifica-se se algum dos segmentos tem mais tendência para evadir do curso. Para isso, acrescenta-se mais um atributo ao conjunto de dados para indicar qual o segmento a que pertence cada aluno.

No atributo "*ignoreAttributeIndices*" informam-se os índices das 21 variáveis ignoradas e mantêm-se com 2 segmentos, conforme teste realizado na etapa anterior ao criar o modelo.

Verifica-se, conforme o gráfico da figura 11, a existência de mais uma variável, "cluster", correspondendo a exatamente cada um dos 2 segmentos definidos anteriormente, o segmento 1, com 2.495 alunos, está definido na cor mais escura (azul) e o segmento 2, com 1.642 alunos, na cor mais clara (vermelha).

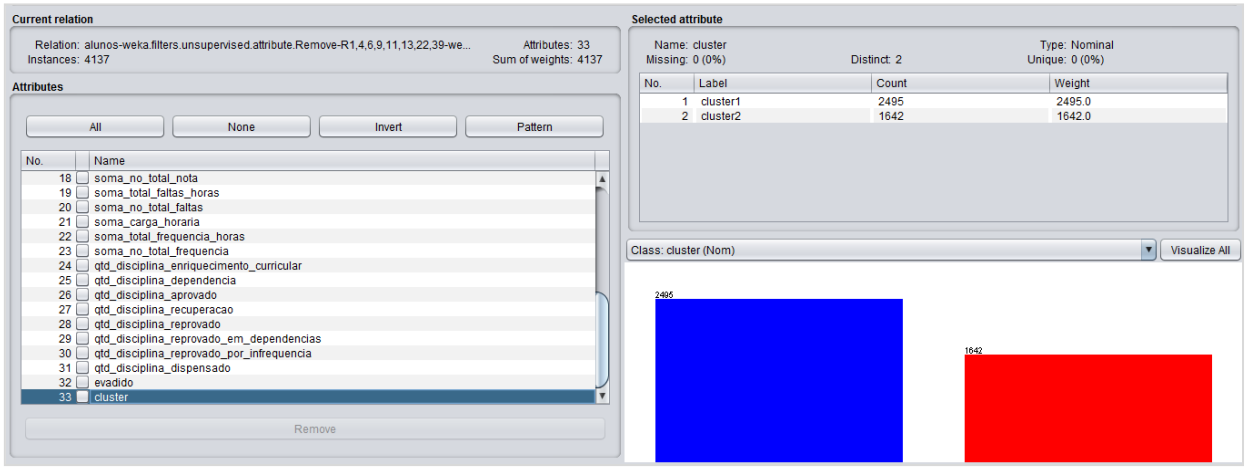

Figura 11. Novo atributo "*cluster*" com 2 segmentos.

Confirma-se que os vários elementos de cada um dos segmentos têm as mesmas características dos centroides. Vejamos, a seguir, essa relação com as variáveis em destaque soma_total_frequencia_horas e qtd_disciplina_reprovado_por_infrequencia:

Soma total da frequência:

Verificamos, a seguir, a variável que informa a soma total da frequência, conforme gráfico da figura 12:

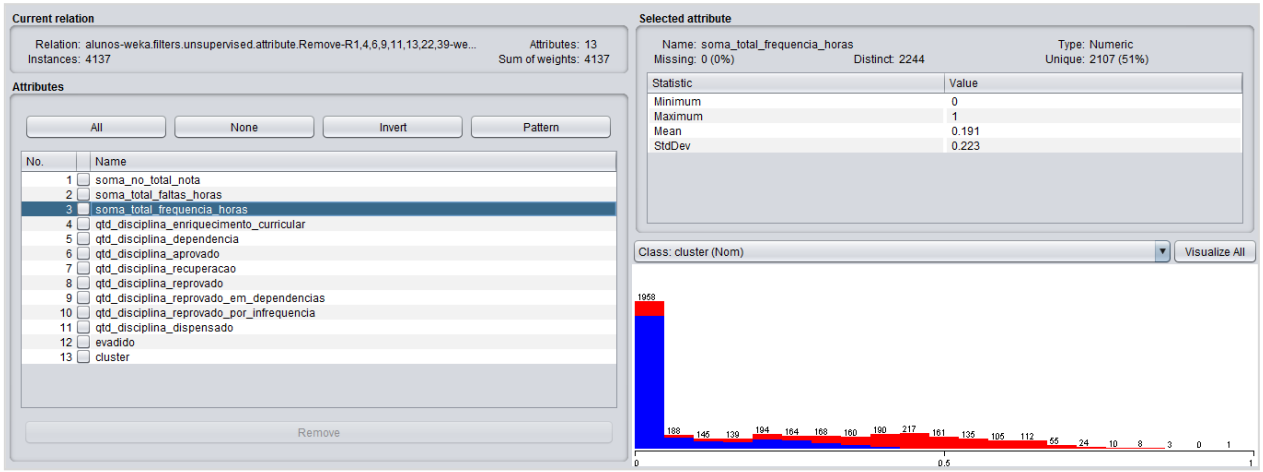

Figura 12. Atributo soma_total_frequencia_horas com a classe "cluster".

Com relação à variável que informa a soma total das frequências do aluno, observada no gráfico da figura 12, a proporção dos mais escuros (azuis), que são os alunos faltosos, é mais recorrente nos valores próximos de zero.

Quantidade de disciplinas reprovadas por infrequência:

Verificamos, a seguir, a variável que informa a quantidade de disciplinas reprovadas por infrequência, conforme gráfico da figura 13:

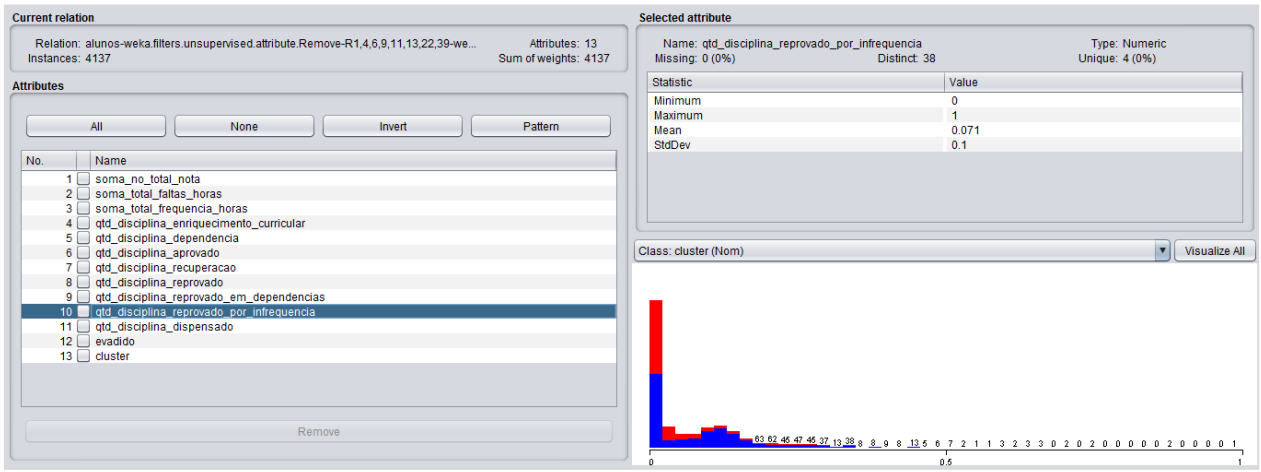

Figura 13. Atributo qtd_disciplina_reprovado_por_infrequencia com a classe "cluster".

Com relação à variável que informa a quantidade de disciplinas reprovadas por infrequência, observada no gráfico da figura 13, a proporção dos mais claros (vermelhos), que são os alunos assíduos, é mais recorrente nos valores próximos de zero.

Sendo assim, nomeiam-se os segmentos da variável *cluster* da seguinte forma:

- Segmento 1: alunos faltosos com alta reprovação por infrequência.
- Segmento 2: alunos assíduos com baixa reprovação por infrequência.

Em seguida, analisa-se se algum destes segmentos tem mais tendência para evadir, verificando o atributo "evadido", conforme gráfico da figura 14:

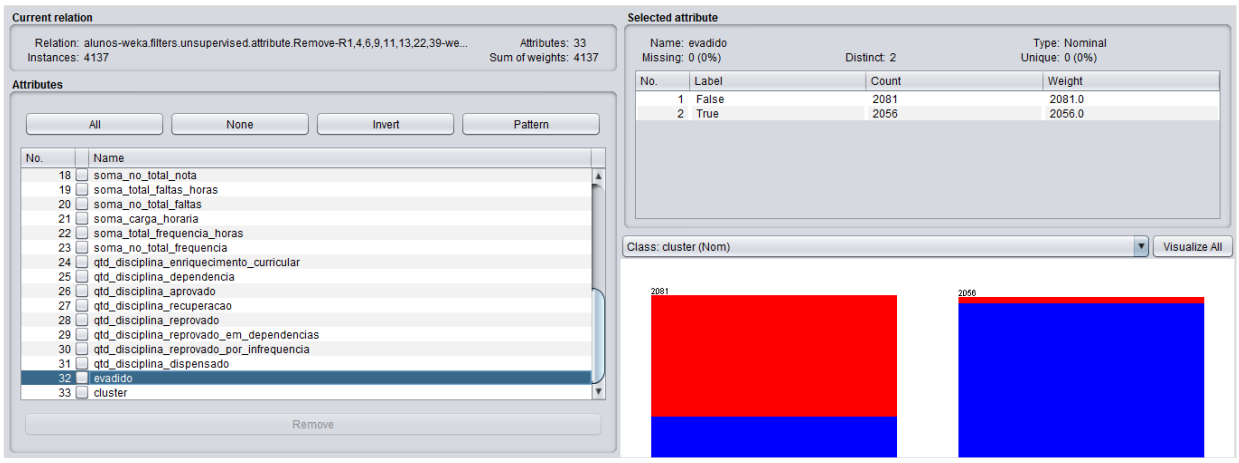

Figura 14. O atributo evadido com as características dos centroides.

Constata-se, conforme gráfico da figura 14, que a proporção dos escuros (azuis) é superior nos evadidos (valor da variável "*True*"), por isso, pode-se afirmar que os alunos faltosos com alta reprovação por infrequência têm mais tendência de evadir-se.

Observa-se também que a proporção dos claros (vermelhos) é baixa nos evadidos, ou seja, são pouquíssimos alunos assíduos e com baixa reprovação por infrequência que tem a tendência de evadir-se, por motivos que não envolva a frequência.

Ainda conforme o gráfico da figura 14, chama-se a atenção para os escuros (azuis) do segmento 1. Apesar de estarem no grupo dos "não evadidos", eles pertencem ao *cluster* 1, ou seja, dos alunos faltosos com alta reprovação por infrequência. Sendo assim, pode-se pré-determinar que estes sejam os casos mais prováveis de evasões futuras.

Com relação à geração de falsos positivos (110 alunos classificados incorretamente como evadidos) e de falsos negativos (145 alunos classificados incorretamente como não evadidos) no caso do modelo de classificação de regras JRip, por representarem, respectivamente, apenas 5,29% e 7,05% da amostra de 4.137 alunos investigados, considera-se uma margem de erro aceitável.

3.6 Fase 6 – Implementação

Após a avaliação de desempenho dos modelos obtidos, prossegue-se a última fase da metodologia CRISP-DM com o desenvolvimento e a distribuição dos resultados alcançados através das técnicas de mineração de dados.

Apresenta-se, a seguir, um conjunto de ações para ser implantado dentro da instituição, que poderá auxiliar os gestores na tomada de decisões quanto à problemática da evasão escolar:

1. Acompanhar a assiduidade dos alunos: o preenchimento e o acompanhamento dos diários de classe, como o próprio nome diz, deve ser "diário", sendo assim, deve-se incentivar o uso do Diário Eletrônico, ferramenta desenvolvida e disponibilizada pela Fábrica de Software do IFTM. Nessa condição, se o aluno faltou "hoje" e o professor registrou sua infrequência de forma online, imediatamente a Secretaria do Curso tomará conhecimento e poderá tomar as providências necessárias. Estabelecer ações diferenciadas para: primeira falta, falta consecutiva ou proximidade com o limite de faltas aceitável, ou seja, intervir proporcionalmente de acordo com a quantidade de faltas.
2. Avaliar individualmente cada caso de infrequência quando necessário. Dessa forma, os motivos que levam a ausência de um aluno não devem ser generalizados. Nos casos de maior gravidade, por exemplo, quando se aproxima da reprovação por

infrequência, considerar a necessidade de visitar o aluno em sua residência para uma conversa pessoal, evidenciando o quanto é importante para a instituição a sua presença nas salas de aula e o seu futuro na sociedade.

3. Criar uma rede de contatos e de confiança, por exemplo: solicitar a um colega de classe entrar em contato com o aluno ausente, ao invés de uma pessoa que trabalha na Secretaria do Curso, que possivelmente é estranha a ele, dessa forma, a preocupação da escola se torna mais pessoal.

4. Divulgar para a família do aluno, educadores e sociedade, nas mídias e redes sociais, os projetos que envolvem as ações que visam o sucesso dos estudantes, como o Plano Estratégico de Ações de Permanência e Êxito dos Estudantes do IFTM, com a finalidade de sensibilizar e conscientizar a todos e incluí-los na responsabilidade de se combater o problema da evasão escolar.

5. Utilizar os meios de comunicação seja por cartas, e-mails, mensagens no celular ou pelas redes sociais, para promover os pontos fortes da instituição, incentivar a assiduidade do aluno e divulgar a importância das frequências nas aulas para alcançar o seu sucesso pessoal e profissional.

6. Investir tanto na qualidade de ensino, quanto na qualidade do clima escolar, ou seja, aperfeiçoar, continuamente, os processos de atendimento e de convivência entre os alunos, as famílias e os colaboradores, fortalecendo assim o vínculo com a instituição.

7. Favorecer a permanência do aluno com problemas de socialização, disciplina e dificuldades de aprendizagem.

8. Definir estratégias para tornar o ambiente escolar mais atrativo, agradável e próximo das necessidades curriculares dos alunos, despertando o seu interesse pelo processo ensino / aprendizagem, oferecendo respostas para as suas ansiedades e dúvidas. Espaços de convivência e lazer aliados ao ambiente de aprendizagem podem estimular a permanência do aluno. Por conseguinte, deve-se acompanhar e avaliar se essas ações tiveram algum efeito redutivo nas infrequências.

9. Criar projetos que envolvam a escola, o aluno, a família e a comunidade, por exemplo, nas áreas de esporte, cultura e conscientização ambiental, de forma que estimule a socialização, o envolvimento e a permanência do aluno no ambiente escolar, valorize o relacionamento, o trabalho em grupo e a cooperação, de forma que proporcione uma integração melhor e mais ampla na relação entre o aluno e os educadores. Desse jeito, o espaço de ensino se torna um local para o exercício da cidadania, onde todos são mais participativos e responsáveis pela educação.

10. Criar um programa de deslocamento para auxiliar o aluno que mora longe e não possui transporte adequado para se locomover até a instituição, por exemplo: beneficiar esses alunos com um cartão de passe escolar ou criar uma campanha de carona solitária.
11. Oferecer formas alternativas de disponibilizar os conteúdos das aulas presenciais, seja em tutoriais, apostilas, livros interativos ou vídeos. Dessa forma, os alunos que não compareceram à aula por motivo de choque de horário com o trabalho, poderão acompanhar o progresso da sua turma à distância. Pode-se criar outra opção de avaliação para esses casos.
12. Todavia, fazer um levantamento sobre o aluno que falta de forma excessiva e verificar se é possível alterar a modalidade do seu curso, oferecendo-o à distância.
13. Viabilizar a criação de ambientes virtuais com o objetivo de reforço das aulas presenciais, pois, se bem planejados, serão de grande valia tanto para os alunos faltosos, quanto para os que possuem certa dificuldade de aprendizado.
14. Manter uma equipe capacitada de desenvolvedores de conteúdos digitais de alta interatividade e ludicidade. Os recursos tecnológicos podem ter um papel importante ao aproximar o aluno do universo digital deixando os conteúdos das aulas mais atrativos e despertando o seu interesse pela disciplina. Enfim, melhorar a qualidade do ensino.
15. Estabelecer níveis de gravidade no problema da evasão, por exemplo: o aluno que evadiu do IFTM para continuar os seus estudos em outra instituição de ensino, independente do motivo, não deixa de ser um problema de evasão, porém, é menos crítico do que aquele aluno que evadiu e não ingressou em nenhum outro curso. Assim, devem-se priorizar soluções para os problemas mais graves de evasão. Outro exemplo é o caso do aluno que mudou de cidade, que é menos crítico do que aquele aluno que evadiu por não ter mais interesse em trabalhar na área do curso. Para esse último caso, poderá ser aplicado um teste de aptidão profissional no momento do seu ingresso, de forma que possa receber orientações para escolher um curso que tenha uma maior afinidade.
16. Criar e manter projetos distintos para as seguintes situações: aluno evadido e em processo de evasão, pois se acredita que as ações para resgatar um aluno evadido são diferentes e mais complexas do que impedir que o aluno que está em processo de evasão abandone os seus estudos.
17. Planejar o acolhimento do aluno que foi resgatado da evasão, no que se refere a: recuperar o conteúdo das aulas perdidas, acompanhar o ritmo de aprendizado da turma

e definir as estratégias para se evitar uma nova evasão.

18. Acompanhar e avaliar diariamente o aluno em processo de evasão e/ou resgatado da evasão, conduzir a sua aprendizagem, o seu desenvolvimento e, principalmente, a sua assiduidade.

Capítulo IV – Discussão

4 Discussão dos Resultados

Conforme se constatou durante a etapa da coleta de dados inicial, a pesquisa realizada neste trabalho utilizou dados que necessitam de uma maior atenção relativamente à sua qualidade. Isto posto, recomenda-se que as informações erradas ou faltantes sejam corrigidas diretamente na base de dados e que rotinas de programação de validação sejam implementadas para que não ocorram mais inserções erradas.

4.1 Conexões com o Plano Estratégico de Ações de Permanência e Êxito dos Estudantes do IFTM

Prossegue-se a discussão, estabelecendo conexões entre os resultados dessa pesquisa e o Plano Estratégico de Ações de Permanência e Êxito dos Estudantes do IFTM, citado no capítulo 2, que foi criado para proporcionar um conjunto de ações que garantam o êxito dos discentes, ou seja, que amenize o problema da evasão escolar. De acordo com o Plano Estratégico (IFTM, 2016), o total de 167 alunos evadidos dos cursos de Tecnologia em Análise e Desenvolvimento de Sistemas e Técnico em Eletrônica Concomitante ao Ensino Médio, dos três primeiros períodos de integralização, responderam ao questionário proposto. Ao mesmo tempo em que mapeava as causas que levaram os alunos do IFTM a evadirem, esse Plano propunha ações para a redução do índice de evasão. Dentre os momentos que foram abrangidos, temos: implantação e acompanhamento das ações propostas. Por consequência, foram planejadas estratégias de intervenção: com indicação das ações de acordo com as causas identificadas, prazos e responsáveis para minimizar os indicadores de retenção e evasão.

Conforme se identificou durante a etapa de avaliação dos modelos, a maioria dos casos de evasão ocorre quando a frequência das aulas é baixa e a quantidade de disciplinas reprovadas por infrequência é alta. Portanto, a variável que totaliza a frequência está plenamente correlacionada com o sucesso do aluno em ser aprovado nas disciplinas, sendo assim, um valor baixo para essa variável determina um aumento na quantidade de disciplinas reprovadas por infrequência e, consequentemente, pode indicar uma grande possibilidade de evasão futura.

A partir do resultado desta pesquisa, onde afirma-se que os alunos faltosos com alta reprovação por infrequência têm mais tendência de evadir-se, é possível inferir que, de acordo com o Plano Estratégico, as principais causas da evasão são:

Relacionadas à dimensão individual do estudante:

- Estudo paralelo em outra instituição, que poderá dificultar a assiduidade no curso do IFTM;

- Incompatibilidade com o horário de trabalho, que poderá impedir o aluno de frequentar a aula no horário certo;
- Distância entre sua moradia e a instituição, que poderá ocasionar atrasos e dificuldades para estar presente no curso todos os dias;
- Dificuldades em conciliar o trabalho com os estudos, onde o compromisso com o trabalho poderá se priorizado em detrimento dos estudos;
- Falta de transporte adequado para chegar à instituição, que poderá dificultar o aluno a chegar no local das aulas, acumulando infrequências constantes.

Relativas à dimensão institucional:

- Retenções em disciplinas ou estágio, que poderá impedir o aluno de seguir o planejamento normal do curso, ficando com o seu rendimento escolar prejudicado.

4.2 Perfil do aluno com maiores e prováveis chances de abandono do curso

De acordo com o resultado desta pesquisa, pode-se afirmar que evadirá aquele aluno faltoso com alta reprovação por infrequência.

Ao estabelecer uma conexão do resultado desta pesquisa com o Plano Estratégico de Ações de Permanência e Êxito dos Estudantes do IFTM, pode-se deduzir que se evadirá aquele aluno que se enquadra em uma ou mais das seguintes situações:

- estudar ao mesmo tempo em outra instituição;
- trabalhar em horários coincidentes com as aulas;
- fracassar na conciliação da sua vida profissional com a acadêmica;
- morar distante do local de estudo e não possuir transporte adequado para se deslocar;
- por fim, não avançar com os estudos por ficar retido em disciplinas ou estágio.

Conclusão e Trabalho Futuro

5 Conclusão e Trabalho Futuro

Depois de finalizado o processo de mineração de dados, neste capítulo conclui-se o trabalho com uma síntese dos resultados obtidos, assim como suas limitações, e as linhas de trabalho futuro, tendo como pretensão a continuidade do trabalho iniciado com esta pesquisa.

5.1 Conclusão

Neste estudo específico, o problema de negócio em questão é o elevado número de alunos que abandonaram os cursos de graduação do IFTM no período de 2012 a 2016.

Inicialmente, foi realizado um estudo da revisão de literatura, que foi enriquecedor por proporcionar um conhecimento mais amplo e definido da metodologia CRISP-DM e, consequentemente, permitir a implementação do processo de KDD para a descoberta de conhecimento em base de dados.

A execução de todas as etapas da mineração de dados possibilitou explorar e analisar de forma eficaz e eficiente os dados do SCA do Virtual IF produzindo resultados satisfatórios quanto à problemática da evasão escolar.

Porém, os dados que foram coletados não são o suficiente para diagnosticar causas de evasão como: dificuldades de aprendizado, falta de interesse pelos estudos, ausência de incentivo por parte dos pais, abandono por motivo de doença crônica, entre outros.

A análise e discussão dos resultados só foi capaz de comprovar apenas a baixa frequência às aulas como causa da evasão. Entretanto, foi possível considerar indicações de outras prováveis causas da evasão. Esse resultado possibilitou recomendar várias ações que poderão ser implementadas e acompanhadas pela equipe de gestão. Na etapa de discussão foram relacionadas algumas dessas ações que podem auxiliar quanto à questão da evasão.

Dentre os diversos problemas encontrados, identificou-se como a principal e mais crítica causa da evasão a infrequência às aulas. Dessa forma, é imprescindível realizar um acompanhamento mais rigoroso quanto ao registro dos diários de classe, assim como, implantar um plano de ações mais efetivo, com metas viáveis e instrumentos para acompanhar as suas ações, que por sua vez, devem estimular a frequência do aluno e diminuir os índices da evasão escolar.

Enfim, as ações de combate à evasão devem ser aplicadas por meio de procedimentos de gestão eficiente e os seus resultados precisam ser acompanhados durante todo o ano letivo, ou seja, combater a evasão escolar exige: planejamento com qualidade e monitoramento diário.

Este trabalho não abrangeu todos os cursos oferecidos pelo IFTM, assim como não englobou todos os anos e também não contém todas as modalidades ofertadas.

5.2 Trabalho Futuro

A conclusão deste trabalho não apresenta o fim, mas sim o ponto de partida para que novos processos de mineração de dados sejam realizados com o banco de dados do SCA do Virtual IF no IFTM.

Durante a execução deste trabalho, foram identificadas diversas oportunidades de trabalho futuro, tais como:

Analisar outros cursos além de graduações, como por exemplo: ensino médio, educação profissional técnica de nível médio, pós-graduações, cursos de extensão etc.

Realizar um trabalho de análise dos dados dos cursos da modalidade EaD e fazer um comparativo com os cursos presenciais, modalidade investigada por esse trabalho.

Fazer pesquisas de investigação na base de dados dos anos anteriores a 2012 e posteriores a 2016. Talvez, criar uma equipe de trabalho que dê continuidade às pesquisas de mineração de dados em todos os anos e em períodos mais curtos, que pode ser muito interesse, por exemplo, para os cursos com o período letivo semestral.

Identificar quais são os cursos com menor índice de evasão e investigar se é possível estabelecer os fatores para essa ocorrência. Se for possível, encontrar soluções para reduzir a evasão nos outros cursos.

Verificar se outros fatores da vida acadêmica do aluno, como o programa de bolsas e assistência estudantil, contribuem para a redução dos índices de evasão.

Experimentar algoritmos de mineração de associação, como por exemplo: *Apriori* e *FP-Grow* para investigar se há relações entre os atributos.

Enfim, explorar novas técnicas de mineração de dados, que abrangem todos os processos que envolvem o aluno, consequentemente, melhorar a informação disponibilizada.

Bibliografia

Abernethy, M. (12 de Maio de 2010). *Mineração de dados com o WEKA, Parte 1: Classificação e armazenamento em cluster*. Acesso em 5 de Novembro de 2016, disponível em IBM: http://www.ibm.com/developerworks/br/opensource/library/os-weka1/

Abernethy, M. (12 de Maio de 2010). *Mineração de dados com o WEKA, Parte 2: Classificação e armazenamento em cluster*. Acesso em 5 de Novembro de 2016, disponível em IBM: http://www.ibm.com/developerworks/br/opensource/library/os-weka2/

Amaral, F. (2016). *Aprenda Mineração de Dados – Teoria e prática.* Rio de Janeiro: Alta Book.

Barbieri, C. (2011). *BI2 – Business Intelligence: modelagem e qualidade.* Rio de Janeiro: Elsevier.

Baskerville, R. L. (Outubro de 1999). Investigating Information Systems With Action Research. *Communications of Association for Information Systems, Vol. II, Article 19*, 1-32.

Carvalho, L. A. (2005). *Datamining – A mineração de Dados no Marketing, Medicina, Economia, Engenharia e Administração.* Rio de Janeiro: Ciência Moderna.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (Agosto de 2000). *CRISP-DM 1.0. Step-by-step data mining guide.* Acesso em 15 de Abril de 2017, disponível em The Modeling Agency: https://www.the-modeling-agency.com/crisp-dm.pdf

Davenport, T. (2014). *Big data no trabalho: derrubando mitos e descobrindo oportunidades.* Rio de Janeiro: Elsevier.

Davenport, T. (2014). *Dados demais!: como desenvolver habilidades analíticas para resolver problemas complexos, reduzir riscos e decidir melhor.* Rio de Janeiro: Elsevier.

Engel, G. I. (2000). Pesquisa-ação. *Educar, n. 16*, p. 181-191.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *From Data Mining to Knowledge Discovery in Databases*. Acesso em 17 de Março de 2017, disponível em Western Science: http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf

Hurwitz, J., Nugent, A., Halper, F., & Kaufman, M. (2015). *Big data para leigos.* Rio de Janeiro: Alta Books.

IFTM. (Outubro de 2013). *Informativo IFTM em Ação - Representantes da PROEN e dos Campus Uberaba e Uberlândia participaram do III Fórum Internacional sobre*

Educação Profissional e Evasão Escolar em Belo Horizonte. Acesso em 9 de Outubro de 2016, disponível em IFTM: http://www.iftm.edu.br/comunicacao/pdf/ano1_out_ed4.pdf

IFTM. (9 de Julho de 2013). *Orientações Gerais Quanto ao Funcionamento do Sistema de Controle Acadêmico (SCA).* Acesso em 28 de Julho de 2017, disponível em IFTM: http://www.iftm.edu.br/ERP/MAC/CRA/arquivos/orientacoes_gerais_proen.pdf

IFTM. (30 de Abril de 2013). *Projeto: Um estudo sobre a evasão nos cursos presenciais do Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro - IFTM*. Acesso em 9 de Outubro de 2016, disponível em IFTM: http://www.iftm.edu.br/proreitorias/ensino/eventos/PDF/eventos/APRESENTACAO_PROJETO_EVASAO.pdf

IFTM. (11 de Abril de 2013). *Relatório de Gestão 2012 do Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro - IFTM*. Acesso em 15 de Abril de 2017, disponível em IFTM: http://www.iftm.edu.br/processo-de-contas/pdf/Relatorio_Gestao_2012.pdf

IFTM. (30 de Outubro de 2014). *Plano de Desenvolvimento Institucional 2014 - 2018.* Acesso em 4 de Março de 2017, disponível em IFTM: http://www.iftm.edu.br/pdi/arquivos/pdi2014_2018.pdf

IFTM. (14 de Abril de 2016). *Plano Estratégico de Ações de Permanência e Êxito dos Estudantes do IFTM*. Acesso em 24 de Fevereiro de 2017, disponível em IFTM: http://www.iftm.edu.br/proreitorias/ensino/permanenciaeexito/plano/documentos/plano-estrategico.pdf

IFTM. (11 de Janeiro de 2017). *Indicadores do IFTM*. Acesso em 1 de Julho de 2017, disponível em IFTM: http://indicadores.iftm.edu.br/

IFTM. (3 de Março de 2017). *Relatório de Gestão 2016 do Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro - IFTM*. Acesso em 4 de Abril de 2017, disponível em IFTM: http://www.iftm.edu.br/processo-de-contas/pdf/Relatorio_Gestao_2016.pdf

IFTM. (11 de Janeiro de 2017). *Tecnologia da Informação e Comunicação*. Acesso em 17 de Março de 2017, disponível em IFTM: http://www.iftm.edu.br/proreitorias/desenvolvimento/tic/

IFTM. (27 de Março de 2018). *Relatório de Gestão 2017 do Instituto Federal de Educação, Ciência e Tecnologia do Triângulo Mineiro*. Acesso em 2 de Abril de 2018, disponível em IFTM: http://www.iftm.edu.br/processo-de-contas/pdf/relato%CC%81rio_de_gesta%CC%83o__-_versa%CC%83o_14_-_com_resolucao_do_consup.pdf

Mayer-Schonberger, V. (2013). *Big data: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana.* Rio de Janeiro: Elsevier.

Nogueira, D. (2014). *Agile Data Mining: Uma metodologia ágil para o desenvolvimento de projetos de data mining.* Dissertação (Dissertação em Engenharia Informática e Computação) - FEUP, Porto.

Piatetsky, G. (Outubro de 2014). *CRISP-DM, ainda é a principal metodologia utilizada para análise, mineração de dados ou projetos de ciência dos dados*. Acesso em 8 de Julho de 2017, disponível em KDnuggets: http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Pinheiro, C. A. (2008). *Inteligência Analítica – Mineração de dados e descoberta de conhecimento.* Rio de Janeiro: Ciência Moderna.

SAS. (17 de Maio de 2018). *Sobre o SAS.* Acesso em 30 de Maio de 2018, disponível em SAS: https://www.sas.com/pt_br/company-information.html

Tan, P.-N., Steinbach, M., & Kumar, V. (2009). *Introdução ao DATAMINING Mineração de dados.* Rio de Janeiro: Ciência Moderna Ltda.

Tripp, D. (set./dez. de 2005). Pesquisa-ação: uma introdução metodológica. *Educação e Pesquisa, v. 31, n. 3*, p. 443-466, Tradução de Lólio Lourenço de Oliveira.

University of Waikato. (11 de Janeiro de 2017). *Downloading and installing Weka*. Acesso em 22 de Abril de 2017, disponível em Weka: http://www.cs.waikato.ac.nz/ml/weka/downloading.html

Anexos

Anexo I – Modelo de Classificação de Regras – JRip

```
=== Run information ===

Scheme:       weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1
Relation:     alunos-weka.filters.unsupervised.attribute.Remove-
R1,4,6,9,11,13,22,39-
weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.supervised.attribute.Discretize-R15,18,19,20,21,22,23-
precision6
Instances:    4137
Attributes:   32
              ds_sexo
              estado_civil
              nacionalidade
              ds_naturalidade_estado
              etnia
              idade_ingresso
              bl_ensino_medio_publico
              campus
              cidade_campus
              curso
              area_conhecimento
              ppc_nivel_categoria_forma
              ds_nome_turno
              periodo_letivo
              no_carga_horaria_minima
              duracao_aula_no_minutos
              ano_ingresso
              soma_no_total_nota
              soma_total_faltas_horas
              soma_no_total_faltas
              soma_carga_horaria
              soma_total_frequencia_horas
              soma_no_total_frequencia
              qtd_disciplina_enriquecimento_curricular
              qtd_disciplina_dependencia
              qtd_disciplina_aprovado
              qtd_disciplina_recuperacao
              qtd_disciplina_reprovado
              qtd_disciplina_reprovado_em_dependencias
              qtd_disciplina_reprovado_por_infrequencia
              qtd_disciplina_dispensado
              evadido
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

JRIP rules:
===========

(soma_total_frequencia_horas = '(-inf-29.94]') and (qtd_disciplina_aprovado
<= 11) => evadido=True (1521.0/31.0)
(qtd_disciplina_aprovado <= 6) and (soma_no_total_frequencia = '(-inf-
243.5]') => evadido=True (224.0/20.0)
(soma_total_frequencia_horas = '(-inf-29.94]') and (qtd_disciplina_aprovado
<= 21) => evadido=True (123.0/18.0)
(qtd_disciplina_aprovado <= 6) and (periodo_letivo = 1 semestre) and
(soma_no_total_frequencia = '(243.5-389.5]') => evadido=True (56.0/4.0)
(soma_total_frequencia_horas = '(29.94-193.14]') and
(qtd_disciplina_aprovado <= 13) and (periodo_letivo = 1 semestre) and
(qtd_disciplina_reprovado >= 2) => evadido=True (19.0/1.0)
```

```
(qtd_disciplina_aprovado <= 6) and (soma_total_frequencia_horas = '(29.94-
193.14]') and (soma_no_total_faltas = '(228.5-inf)') => evadido=True
(7.0/0.0)
(qtd_disciplina_reprovado_por_infrequencia >= 5) and (soma_carga_horaria =
'(565.21-757]') and (ds_nome_turno = noturno) and (idade_ingresso <= 39) =>
evadido=True (15.0/0.0)
(qtd_disciplina_reprovado_por_infrequencia >= 5) and
(soma_total_frequencia_horas = '(-inf-29.94]') and (soma_no_total_faltas =
'(228.5-inf)') => evadido=True (15.0/5.0)
(qtd_disciplina_reprovado_por_infrequencia >= 8) and (curso = licenciatura
em quimica) and (qtd_disciplina_dependencia <= 9) => evadido=True (5.0/0.0)
 => evadido=False (2152.0/150.0)

Number of Rules : 10


Time taken to build model: 0.85 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3882               93.8361 %
Incorrectly Classified Instances       255                6.1639 %
Kappa statistic                          0.8767
Mean absolute error                      0.1011
Root mean squared error                  0.2354
Relative absolute error                 20.2292 %
Root relative squared error             47.0753 %
Total Number of Instances             4137

=== Detailed Accuracy By Class ===

               TP Rate FP Rate Precision Recall F-Measure MCC   ROC Area PRC Area
Class
               0,947   0,071   0,931     0,947  0,939     0,877 0,949    0,928
False
               0,929   0,053   0,946     0,929  0,937     0,877 0,949    0,941
True
Weighted Avg.  0,938   0,062   0,938     0,938  0,938     0,877 0,949    0,934

=== Confusion Matrix ===

    a    b   <-- classified as
 1971  110 |    a = False
  145 1911 |    b = True
```

Anexo II – Modelo de Classificação de Árvore – J48

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.01 -M 2
Relation:                alunos-weka.filters.unsupervised.attribute.Remove-
R1,4,6,9,11,13,22,39-
weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.supervised.attribute.Discretize-R15,18,19,20,21,22,23-
precision6
Instances:    4137
Attributes:   32
              ds_sexo
              estado_civil
              nacionalidade
              ds_naturalidade_estado
              etnia
              idade_ingresso
              bl_ensino_medio_publico
              campus
              cidade_campus
              curso
              area_conhecimento
              ppc_nivel_categoria_forma
              ds_nome_turno
              periodo_letivo
              no_carga_horaria_minima
              duracao_aula_no_minutos
              ano_ingresso
              soma_no_total_nota
              soma_total_faltas_horas
              soma_no_total_faltas
              soma_carga_horaria
              soma_total_frequencia_horas
              soma_no_total_frequencia
              qtd_disciplina_enriquecimento_curricular
              qtd_disciplina_dependencia
              qtd_disciplina_aprovado
              qtd_disciplina_recuperacao
              qtd_disciplina_reprovado
              qtd_disciplina_reprovado_em_dependencias
              qtd_disciplina_reprovado_por_infrequencia
              qtd_disciplina_dispensado
              evadido
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

soma_total_frequencia_horas = '(-inf-29.94]'
|   qtd_disciplina_aprovado <= 20: True (1637.0/46.0)
|   qtd_disciplina_aprovado > 20
|   |   qtd_disciplina_reprovado_por_infrequencia <= 8
|   |   |   qtd_disciplina_reprovado <= 5: False (169.0/21.0)
|   |   |   qtd_disciplina_reprovado > 5: True (11.0/3.0)
|   |   qtd_disciplina_reprovado_por_infrequencia > 8: True (10.0/2.0)
soma_total_frequencia_horas = '(29.94-193.14]'
|   qtd_disciplina_aprovado <= 23: True (310.0/49.0)
|   qtd_disciplina_aprovado > 23: False (66.0/9.0)
soma_total_frequencia_horas = '(193.14-263.675]': False (124.0/51.0)
soma_total_frequencia_horas = '(263.675-397.1]': False (275.0/50.0)
```

```
soma_total_frequencia_horas = '(397.1-718.84]': False (689.0/52.0)
soma_total_frequencia_horas = '(718.84-inf)': False (846.0/5.0)

Number of Leaves  :     10

Size of the tree :      15


Time taken to build model: 0.22 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        3833               92.6517 %
Incorrectly Classified Instances       304                7.3483 %
Kappa statistic                          0.853
Mean absolute error                      0.1186
Root mean squared error                  0.2468
Relative absolute error                 23.7219 %
Root relative squared error             49.3641 %
Total Number of Instances             4137

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area
Class
               0,947    0,094    0,910      0,947   0,928      0,854    0,960     0,956
False
               0,906    0,053    0,944      0,906   0,925      0,854    0,960     0,948
True
Weighted Avg.  0,927    0,074    0,927      0,927   0,926      0,854    0,960     0,952

=== Confusion Matrix ===

    a    b   <-- classified as
 1971  110 |    a = False
  194 1862 |    b = True
```

Anexo III – Modelo de Segmentação – K-means

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -
periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
Relation:     alunos-weka.filters.unsupervised.attribute.Remove-
R1,4,6,9,11,13,22,39-
weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances:    4137
Attributes:   32
              soma_no_total_nota
              soma_total_faltas_horas
              soma_total_frequencia_horas
              qtd_disciplina_enriquecimento_curricular
              qtd_disciplina_dependencia
              qtd_disciplina_aprovado
              qtd_disciplina_recuperacao
              qtd_disciplina_reprovado
              qtd_disciplina_reprovado_em_dependencias
              qtd_disciplina_reprovado_por_infrequencia
              qtd_disciplina_dispensado
Ignored:
              ds_sexo
              estado_civil
              nacionalidade
              ds_naturalidade_estado
              etnia
              idade_ingresso
              bl_ensino_medio_publico
              campus
              cidade_campus
              curso
              area_conhecimento
              ppc_nivel_categoria_forma
              ds_nome_turno
              periodo_letivo
              no_carga_horaria_minima
              duracao_aula_no_minutos
              ano_ingresso
              soma_no_total_faltas
              soma_carga_horaria
              soma_no_total_frequencia
              evadido
Test mode:    evaluate on training data


=== Clustering model (full training set) ===


kMeans
======

Number of iterations: 19
Within cluster sum of squared errors: 476.5255685122146

Initial starting points (random):

Cluster 0: 0.529169,0.120663,0.459282,0,0,0.620253,0,0.08,0,0,0
Cluster 1: 0.39734,0.081746,0.708655,1,0,0.303797,0,0.08,0,0,0.244898
```

```
Missing values globally replaced with mean/mode

Final cluster centroids:
                                                        Cluster#
Attribute                                          Full Data            0
1
                                                (4137.0)             (2495.0)
(1642.0)
=============================================================================
=
soma_no_total_nota                                    0.1667       0.0519
0.3412
soma_total_faltas_horas                               0.0931       0.0903
0.0974
soma_total_frequencia_horas                           0.1908       0.0604
0.389
qtd_disciplina_enriquecimento_curricular              0.0034       0.0004
0.0079
qtd_disciplina_dependencia                            0.0533       0.0381
0.0763
qtd_disciplina_aprovado                               0.1717       0.0435
0.3665
qtd_disciplina_recuperacao                            0.0085       0.0056
0.013
qtd_disciplina_reprovado                               0.074       0.0454
0.1176
qtd_disciplina_reprovado_em_dependencias             0.0002            0
0.0006
qtd_disciplina_reprovado_por_infrequencia             0.0715       0.0868
0.0482
qtd_disciplina_dispensado                             0.0215       0.0167
0.0288



Time taken to build model (full training data) : 0.33 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      2495 ( 60%)
1      1642 ( 40%)
```