UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

GABRIEL CLÁVILA SOARES

SAM - UMA ABORDAGEM ESPECÍFICA DE MINERAÇÃO DE DADOS SOCIOECONÔMICOS DE ALUNOS DO IF AMAZONAS PARA APOIO AO PROCESSO DE CONCESSÃO DE ASSISTÊNCIA ESTUDANTIL

Recife
2020

GABRIEL CLÁVILA SOARES

SAM - UMA ABORDAGEM ESPECÍFICA DE MINERAÇÃO DE DADOS SOCIOECONÔMICOS DE ALUNOS DO IF AMAZONAS PARA APOIO AO PROCESSO DE CONCESSÃO DE ASSISTÊNCIA ESTUDANTIL

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Banco de Dados

Orientador (a): Professor Dr. Fernando da Fonseca de Souza

Recife
2020

Catalogação na fonte
Bibliotecária Mariana de Souza Alves CRB4-2105

S676s Soares, Gabriel Clávila.
SAM - uma abordagem específica de mineração de dados socioeconômicos de alunos do IF Amazonas para apoio ao processo de concessão de assistência estudantil / Gabriel Clávila Soares. – 2020.
150 f.: il., fig., tab.

Orientador: Fernando da Fonseca de Souza.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação. Recife, 2020.
Inclui referências, apêndices e anexos.

1. Banco de Dados. 2. Descoberta do Conhecimento. 3. CRISP-DM. 4. Análise Comparativa de Algoritmos. I. Souza, Fernando da Fonseca de. (orientador) II. Título.

025.04 CDD (22. ed.) UFPE-CCEN 2020-152

Gabriel Clávila Soares

SAM - Uma Abordagem Específica de Mineração de Dados Socioeconômicos de Alunos do IF Amazonas para Apoio ao Processo de Concessão de Assistência estudantil

Dissertação apresentada ao Programa de Pós-Graduação Profissional em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre Profissional em 15 de julho de 2020.

Aprovado em: 15/07/2020.

BANCA EXAMINADORA

Prof. Luciano de Andrade Barbosa
Centro de Informática / UFPE

Profa.Aida Araújo Ferreira
Instituto Federal de Pernambuco

Prof. Fernando da Fonseca de Souza
Centro de Informática / UFPE
(Orientador)

AGRADECIMENTOS

A conclusão deste trabalho não seria possível sem a participação de pessoas fundamentais que fazem parte da minha vida, que por meio de manifestações de apoio, carinho e compreensão, serviram de motivação para concluir esta etapa importante em minha vida.

Assim, expresso toda minha gratidão: Primeiramente a Deus, pelo dom da vida, pela a proteção divina e por sempre ouvir as minhas preces e me guiar para que eu possa alcançar os meus objetivos.

A minha esposa Neylianne e às minhas filhas Serena e Nirvana, que sempre me apoiaram sendo compreensivas durante os meus períodos de ausências ou em relação ao tempo que tive que dedicar para realização desta pesquisa, além de todo carinho e amor que serviram de motivação.

A minha Mãe Maria Lucimar e as minhas tias Maria Madalena e Mirian Marly, pelo amor incondicional, por sempre me apoiarem e incentivarem em toda a minha vida acadêmica e por serem exemplos de força e integridade. Aos demais familiares e amigos, os quais eu não poderia nomear aqui pois são muitos, por todo o incentivo e pelas palavras de motivação.

Aos colegas do IFAM Campus Manaus Zona Leste, em especial dos membros da CGTIC, pelas sugestões, manifestações de apoio, ajuda e compromisso assumidos durante os períodos da minha ausência. Ao IFAM, em especial aos gestores do Campus Manaus Zona Leste por oferecer todo o apoio e as condições necessárias para a minha participação neste Mestrado.

Ao Professor e Orientador Dr. Fernando da Fonseca de Souza, por acreditar nesta pesquisa, pela compreensão, apoio, segurança, atenção e disponibilidade que sempre prestou durante todo o período de orientação deste trabalho. Aos demais professores do CIN-UFPE, por todo o conhecimento compartilhado, bem como os servidores TAE, em especial à Joelma, por todo o atendimento e apoio prestados sempre com muita gentileza e respeito.

Aos colegas de turma de mestrado Marcelo, Paulo e Fagner pelas constantes trocas de experiências e conhecimentos e a boa convivência e em especial ao Carlos, Diego, Eliandro, Gustavo, Lucas, Leonardo, Rogério e Welington, pelos momentos de descontração, estudo e apoio.

RESUMO

Os recursos tecnológicos vêm cada vez mais sendo utilizados pelas organizações, sejam elas públicas ou privadas, para armazenar os seus dados, desde instituições públicas de ensino até grandes corporações financeiras. Com isso, todos os dias surgem novas bases com grandes volumes de dados armazenados. Entretanto, devido à grande quantidade de dados que só aumenta a cada dia, torna-se inviável a realização da análise desses dados de forma manual por pessoas. Visando automatizar este processo de análise de grandes bases de dados e transformá-los em conhecimento útil se faz necessária a aplicação da Mineração de Dados por meio de algoritmos. Atualmente os Institutos Federais de Educação, Ciência e Tecnologia utilizam o Plano Nacional de Assistência Estudantil (PNAES) que objetiva conceder bolsas financeiras de auxílios aos estudantes em situação de vulnerabilidade socioeconômica. Porém, esse processo de seleção para concessão das bolsas de assistência estudantil demanda um certo tempo para a análise socioeconômica dos alunos. Além disso, os dados armazenados não estão sendo explorados para gerar conhecimento útil. Diante disso, surge a questão: Qual processo automatizado que utilize o algoritmo de mineração de dados mais eficiente e eficaz para identificar quais alunos estão aptos a receber a assistência estudantil? Portanto, esta pesquisa objetiva desenvolver uma abordagem específica de Mineração de Dados Socioeconômicos dos alunos do Instituto Federal do Amazonas visando automatizar o processo de concessão da assistência estudantil. Para isso, também será necessário realizar uma análise comparativa dos algoritmos de mineração visando identificar qual é o mais eficiente e eficaz em predizer os alunos aptos a receber a assistência estudantil. Por fim, esta abordagem deve gerar conhecimento útil como evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica do Instituto Federal do Amazonas.

Palavras-chave: Descoberta do Conhecimento. CRISP-DM. Análise Comparativa de Algoritmos.

ABSTRACT

Technological resources are increasingly being used by organizations, whether public or private, to store their data, ranging from public educational institutions to large financial corporations. As a result, new databases appear every day with large volumes of stored data. However, due to the large amount of data that increases every day, it is not feasible to carry out manual analysis on them. In order to automate such a process of analyzing large databases and transforming them into useful knowledge, it is necessary to apply Data Mining through algorithms. Currently, the Federal Institutes of Education, Science and Technology use the National Student Assistance Plan (PNAES) that is aimed at granting financial support to aid students under socioeconomic vulnerability. However, this selection process requires a considerable amount of time for the socioeconomic analysis of students. In addition, the stored data currently is not being explored to generate useful knowledge. Therefore, this question arises: What is automatized process along with the most efficient and effective data mining algorithm to identify which students are able to receive assistance? Therefore, this research is aimed at developing a specific approach for mining socioeconomic data of students enrolled at the Federal Institute of Amazonas to automate the process of granting student's assistance. For this, it is also be necessary to carry out a comparative analysis of data mining algorithms in order to identify which is the most efficient and effective in predicting students able to receive such assistance. Finally, this approach shall also generate useful knowledge such as determining the profile of students under socioeconomic vulnerability at the Federal Institute of Amazonas.

Keywords: Knowledge Discovery. CRISP-DM. Comparative Analysis of Algorithms.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARFF	Attribute Relation File Format
AUC	Area Under Curve
CFS	*Correlation-based Feature Selection*
CSV	Commom Separated Values
CMZL	Campus Manaus Zona Leste
CRISP-DM	Cross Industry Standard Process for Data Mining
DCBD	Descoberta de Conhecimento em Banco de Dados
DM	Data Mining
ENEM	Exame Nacional do Ensino Médio
IF	Instituto Federal de Educação, Ciência e Tecnologia
IFAM	Instituto Federal de Educação, Ciência e Tecnologia do Amazonas
IFES	Institutos Federais de Ensino Superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KDD	Kwoledge Discovery in Databases
KNN	k-Nearest Neighbors
MD	Mineração de Dados
MLP	*MultiLayer Perceptron*
PART	Partial Decision Trees
PNAES	Programa Nacional de Assistência Estudantil
RNA	Redes Neurais Artificiais
ROC	Receiver Operating Characteristics
SAM	Student Assistance Mining
SGBD	Sistema Gerenciador de Banco de Dados
SMO	Sequential Minimal Optimization
SQL	Structured Query Language
SVM	Support Vector Machine
WEKA	Waikato Environment for Knowledge Analysis

SUMÁRIO

1 INTRODUÇÃO

Atualmente, com o constante avanço tecnológico as organizações perceberam que a tecnologia é a melhor forma de manter as suas informações. A partir disso, imensos volumes de dados passaram a ser sistematicamente coletados e armazenados gerando milhões de bases de dados.

Entretanto, devido ao seu grande volume surgem dificuldades no momento de explorar esses dados, pois excedem a capacidade humana e a habilidade técnica na sua interpretação. Segundo Goldschmidt e Passos (2005), existe a necessidade da aplicação de ferramentas computacionais capazes de analisar, relacionar e interpretar esses dados.

Sendo assim, para converter esses dados em informações úteis é necessária a realização de um processo de várias etapas denominado de Descoberta de Conhecimento em Banco de dados (DCBD), em inglês *Knowledge Discovery in Databases* (KDD). O KDD é um processo de várias etapas, utilizado para identificar padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados. Dentre essas etapas está a mineração de dados, formada por um conjunto de ferramentas e técnicas que, por meio da utilização de algoritmos, exploram um conjunto de dados, extraindo ou evidenciando padrões e predizendo conhecimento. Tal conhecimento pode ser apresentado por essas ferramentas de diversas formas: classificação, agrupamentos, hipóteses, regras, árvores de decisão, grafos ou dendrogramas (FAYYAD et al., 1996).

O Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM), mais especificamente o Campus Manaus Zona Leste, considerando o Decreto Nº 7.234, de 19 de julho de 2010 que dispõe sobre o Programa Nacional de Assistência Estudantil – PNAES - seleciona estudantes para receberem benefícios por meio do Programa Socioassistencial Estudantil. Para ser beneficiário desse programa, o aluno deve estar regularmente matriculado e frequentando um curso presencial e que comprove estar em situação de vulnerabilidade socioeconômica, tendo em vista a finalidade do programa de propiciar-lhes condições favoráveis a permanência, êxito e conclusão de seus cursos.

A mineração de dados é uma importante ferramenta que permite a descoberta de padrões em uma determinada base de dados. Por meio da aplicação dessas técnicas, é possível identificar com mais eficiência e eficácia informações relevantes

para serem utilizadas pelo processo decisório de qualquer organização, beneficiando particularmente ações que envolvam gastos públicos.

1.1 MOTIVAÇÃO

Atualmente, a análise dos formulários socioeconômicos preenchidos durante o processo seletivo do programa social assistencial do Instituto Federal do Amazonas (IFAM), mais especificamente o Campus Manaus Zona leste, demanda muito tempo ao setor de Serviço Social, o que acaba influenciando na data de início do pagamento das bolsas assistenciais aos estudantes, uma vez que todo o processo seletivo deve estar finalizado para que os pagamentos possam ser realizados.

Outro fato importante é que toda essa base de dados socioeconômicos dos alunos não está sendo explorada para extração de informações úteis que poderiam auxiliar a direção geral do campus a identificar o perfil do aluno em situação de vulnerabilidade e com isso analisar, por exemplo, se o recurso financeiro aplicado atualmente é suficiente para atender às reais necessidades desses alunos.

Esses dois fatos mencionados acima reforçam a necessidade de um processo específico que auxilie na melhoria de ambas as situações. A utilização do processo de KDD com a aplicação da Mineração de Dados pode ser a melhor opção, haja vista que a sua concepção visava principalmente à descoberta de padrões ou a realizar predições, sendo que esses padrões possam ser expressos de forma a gerar conhecimento útil (WITTEN et. al., 2011).

Nesse contexto, a aplicação de um processo específico de KDD, bem como uma análise comparativa de técnicas de mineração de dados que possa determinar o algoritmo de mineração mais eficiente e eficaz para tanto certamente pode contribuir para dar agilidade à tomada de decisão do Serviço Social. Além disso, poderá permitir a ampliação da capacidade de extrair informações auxiliando a direção geral do campus em seus planejamentos estratégicos voltados para os alunos, relacionados aos aspectos financeiros, culturais e sociais. Tal processo pode, ainda, também ser aplicado a situações análogas em todas as unidades dos Institutos Federais.

1.2 QUESTÕES DA PESQUISA

Diante das motivações apresentadas anteriormente, pode-se perceber que os problemas do processo de seleção da assistência estudantil do IFAM Campus Manaus Zona Leste estão diretamente ligados ao tempo gasto na análise socioeconômica dos alunos; à falta de definição do perfil do aluno em situação de vulnerabilidade socioeconômica; e às possíveis informações (conhecimento útil) que não estão sendo exploradas nessa base de dados. A partir destes problemas, surgem algumas questões de pesquisa que se pretende responder com este trabalho:

- É possível por meio da mineração de dados, evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica?
- Qual é o algoritmo de mineração de dados mais eficiente e eficaz para identificar quais alunos estão aptos a receber a assistência estudantil?
- Quais os possíveis conhecimentos úteis que podem ser gerados a partir de aplicação da mineração de dados sobre essa base de dados?

1.3 DEFINIÇÃO DE HIPÓTESES

Levando em consideração as questões de pesquisa levantadas, as seguintes hipóteses serão investigadas ao longo deste trabalho:

- H1: A mineração de dados torna mais rápido o processo de análise socioeconômica dos alunos melhorando a eficácia dessa análise.
- H2: O emprego de técnicas de mineração de dados permite definir o perfil do aluno em situação de vulnerabilidade socioeconômica a partir dos seus dados socioeconômicos.

1.4 OBJETIVOS

Para responder às questões da pesquisa e comprovar as hipóteses estabelecidas, nesta seção são definidos o objetivo geral e os objetivos específicos deste trabalho.

1.4.1 Objetivo geral

Esta pesquisa tem como objetivo geral desenvolver uma abordagem especifica para aplicação da Mineração de Dados sobre os dados socioeconômicos de alunos que utilize o algoritmo mais eficiente e eficaz neste contexto, visando tornar mais eficiente o processo de análise para concessão da assistência estudantil do Instituto Federal do Amazonas.

1.4.2 Objetivos Específicos

Os objetivos específicos dessa pesquisa são:

- Identificar junto às assistentes sociais os dados mais relevantes para análise socioeconômica do aluno e que irão influenciar no treinamento e teste dos algoritmos de mineração;
- Realizar uma análise comparativa dos algoritmos de mineração de dados de modo a identificar qual(is) obteve(tiveram) os melhores desempenhos; e
- Gerar informações que sirvam de conhecimento útil, evidenciando o perfil do aluno em situação de vulnerabilidade socioeconômica e consequentemente apoiando na tomada de decisão do processo seletivo da assistência estudantil.

1.5 ESTRUTURA DA DISSERTAÇÃO

A dissertação, além deste capítulo, deverá ser estruturada como segue:

- Capítulo 2 apresenta e detalha a metodologia utilizada que guiará os processos e os métodos que serão empregados no desenvolvimento do trabalho;
- Capítulo 3 corresponde à fundamentação conceitual, no qual são apresentados os conceitos que embasam esta pesquisa como o KDD e as suas etapas, a mineração de dados e as suas principais tarefas, técnicas, métricas de avaliação e a metodologia CRISP-DM para a sua aplicação;

- Capítulo 4 apresenta uma análise dos trabalhos relacionados, no qual são apresentadas várias pesquisas que são relacionadas à pesquisa proposta nesta dissertação, bem como os critérios de escolha e análise desses trabalhos e uma análise crítica sobre os mesmos;
- Capítulo 5 corresponde à abordagem específica de Mineração de Dados desenvolvida neste trabalho, mostrando como foram aplicadas as fases de entendimento dos dados, preparação dos dados e modelagem da metodologia CRISP-DM sobre a base de dados do programa socioassistencial;
- Capítulo 6 apresenta os experimentos realizados, no qual após aplicação dos algoritmos de Mineração dos Dados, foram aplicadas as métricas de avaliação visando analisar o desempenho e a eficácia do conhecimento útil gerado pelos algoritmos;
- Capítulo 7 corresponde às conclusões da dissertação e apresenta as suas contribuições, as limitações encontradas durante a realização da pesquisa e os trabalhos futuros que poderão ser realizados a partir dela; e
- Por fim, são listadas as referências bibliográficas utilizadas como base para elaboração deste trabalho, seguidas de anexos e apêndices.

2 METODOLOGIA DE TRABALHO

Em função dos objetivos desta pesquisa apontarem para a utilização de técnicas de mineração de dados, foi considerada a aplicação do processo de KDD, sendo utilizado como metodologia o CRISP-DM, acrônimo para *Cross Industry Standard Process for Data Mining*. Sua escolha deveu-se, principalmente, por se tratar de um modelo de processo proposto especificamente para projetos de Mineração de Dados (SHEARER, 2000). Ele é composto por seis fases: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação (CHAPMAN et al., 2000). Portanto, o processo metodológico de desenvolvimento desta pesquisa está estruturado em três etapas sequencialmente inter-relacionadas, baseadas na metodologia CRISP-DM, conforme mostrado na Figura 1.

Figura 1– Etapas da Metodologia Aplicada Baseada em CRISP-DM

Fonte: Elaborada pelo Autor (2020)

As subseções seguintes descrevem as referidas etapas.

2.1 PRIMEIRA ETAPA

A primeira etapa da metodologia aplicada a este trabalho, correspondente à primeira fase da metodologia CRISP-DM - Entendimento do Negócio - ,é constituída por uma pesquisa bibliográfica, a qual consiste de um levantamento em várias publicações cientificas para obter conceitos fundamentais relacionados à descoberta do conhecimento em banco de dados (KDD) e à mineração de dados (*data mining*) com o foco na aplicação de suas técnicas sobre dados socioeconômicos, assim como na comparação e avaliação da eficácia dessas técnicas. Segundo Marconi e Lakatos (2010), o objetivo da pesquisa bibliográfica é colocar o pesquisador em contato direto com o que foi escrito, falado ou filmado sobre determinado assunto.

Simultaneamente a esse levantamento bibliográfico, foi realizada uma pesquisa de campo de cunho exploratório, por meio de entrevistas com as assistentes sociais do campus para saber como ocorre atualmente o processo de seleção dos alunos para receber as bolsas sócio assistenciais. Essa pesquisa de campo foi seguida de um acompanhamento in loco de toda a realização desse processo de seleção, visando identificar os problemas enfrentados pela assistência social bem como suas causas e efeitos. Ainda de acordo com Marconi e Lakatos (2010), a pesquisa de campo de cunho exploratório é formada por investigações de pesquisas empíricas cujo objetivo é a formulação de questões ou de um problema, tendo três possíveis finalidades: desenvolver hipóteses; aumentar os conhecimentos do pesquisador em relação a um ambiente, fato ou fenômeno, para aumentar a sua precisão ao realizar uma pesquisa futura; ou modificar e esclarecer conceitos.

O terceiro elemento desta etapa é a análise de trabalhos relacionados. Para tanto, foi realizada uma revisão sistemática de literatura visando encontrar tais trabalhos. Assim, foram definidos critérios iniciais a serem utilizados na estratégia de seleção desses trabalhos, de modo a encontrar um melhor alinhamento com o contexto desta pesquisa. Um deles foi o de inclusão de trabalhos que aplicaram o KDD e a mineração de dados, visando predição ou identificação, sobre bases de dados relacionados a alunos. Outro, para atender à contemporaneidade da pesquisa, foi o critério de exclusão de trabalhos publicados antes de 2014.

2.1.1 Critérios para Análise dos Trabalhos Relacionados

O primeiro critério definido foi o objetivo pelo qual a mineração de dados foi utilizada no trabalho analisado, ou seja, se foi utilizada para realizar predição, identificação ou ambas, já que o foco deste estudo é utilizar as duas situações para a resolução das suas questões.

O segundo critério levado em consideração nesta análise foi o tipo de tarefa da mineração de dados definido na pesquisa analisada: classificação, associação, regressão, entre outras, tendo em vista que neste estudo mais de uma tarefa poderá ser utilizada. Isto consequentemente influenciará em outro critério muito importante que é a quantidade de algoritmos aplicados.

O terceiro critério definido foi se a pesquisa analisada levou em consideração ou teve como foco principal da mesma a aplicação do KDD e da mineração sobre os dados socioeconômicos dos alunos e como foi feita essa abordagem. Esse critério é importante devido ao fato de que este estudo é aplicado exclusivamente sobre dados socioeconômicos, já que seu foco está diretamente relacionado ao processo seletivo de bolsas assistenciais estudantis.

O quarto critério é a quantidade de algoritmos aplicados na pesquisa analisada, dado que um dos objetivos deste estudo é realizar uma análise comparativo dos algoritmos em relação ao tipo de tarefa definido.

O quinto critério é a quantidade de métricas de avaliação aplicadas na pesquisa analisada. Este critério é importante já que o principal ponto da análise comparativa dos algoritmos é o desempenho dos mesmos, e para que isso possa ser avaliado se faz necessária a aplicação dessas métricas.

A partir dos critérios de inclusão e exclusão citados neta seção, foi definida a string de busca. Para tanto, as palavras chave foram utilizadas com os operadores lógicos AND e OR, como mostrado a seguir:

String de Busca: ("mineração de dados" OR "descoberto do conhecimento") AND ("predição" OR "identificação") AND ("evasão" OR "desempenho") AND "socioeconômicos".

A máquina de busca utilizada para as pesquisas foi o Google Scholar[1]. Após a aplicação da string de busca foram realizadas diversas triagens para reduzir a quantidade de trabalhos relacionados a serem analisados, visando o seu

---

1 https://scholar.google.com/

alinhamento principalmente com o tema e os objetivos desta pesquisa. O Quadro 1 mostra a quantidade de trabalhos sendo reduzida a cada aplicação de um critério de triagem dos trabalhos.

Quadro 1– Triagem dos Trabalhos Relacionados

Critérios Aplicados	**Quantidade Resultante de Trabalhos**
Aplicação da String de busca no Google Scholar	549
Aplicação do critério de exclusão de trabalhos com mais de 5 anos	387
Excluir trabalhos de conclusão de curso	311
Excluir trabalhos que não trabalham com dados de estudantes	183
Excluir trabalhos que abordam EAD	116
Excluir trabalhos que não abordam processo de seleção	38
Excluir trabalhos que não abordam dados socioeconômicos	6

Fonte: Elaborado pelo Autor (2020)

Conforme mostrado no Quadro 1, após a aplicação dos critérios de avaliação e os métodos de triagem, foram selecionados seis trabalhos que atenderam ao critério inicial de exclusão, ou seja, somente trabalhos de 2014 em diante visando a contemporaneidade da pesquisa, e que também atenderam ao critério inicial de inclusão que trata da aplicação da mineração de dados visando a predição ou identificação sobre bases de dados de alunos, além dos demais critérios visando o alinhamento dos trabalhos com o tema e os objetivos desta pesquisa.

2.2 SEGUNDA ETAPA

A segunda etapa da metodologia aplicada a este trabalho, correspondente à segunda fase da metodologia CRISP-DM - Entendimento dos dados - se inicia com a coleta de dados, obtida nas bases de dados geradas por meio do preenchimento de formulários eletrônicos. Foi utilizada como instrumento a pesquisa documental, pois nessas bases de dados se concentra todos os dados sociais, econômicos e culturais dos estudantes. Lembrando que foram coletados somente os dados dos estudantes do Campus Manaus Zona Leste do IFAM, o qual é o ambiente lócus desta pesquisa. Logo após, foi realizada a terceira fase da metodologia CRISP-DM - Preparação de Dados, na qual serão feitas a limpeza e transformação dos dados, de

modo a permitir selecionar os dados realmente relevantes para a pesquisa e remover as possíveis inconsistências. Finalizando esta segunda etapa, serão selecionas e aplicadas as técnicas de mineração de dados, correspondendo à quarta fase da metodologia CRISP-DM - Modelagem.

2.3 TERCEIRA ETAPA

A terceira e última etapa da metodologia aplicada a este trabalho, correspondente a quinta e sexta fases da metodologia CRISP-DM - Avaliação e Implantação - consiste na realização da análise dos resultados da pesquisa. Nela foi realizada uma análise comparativa dos algoritmos utilizados, por meio da aplicação das métricas de avaliação, buscando identificar quais deles obtiveram um melhor desempenho na resolução dos problemas em questão. Sendo tais resultados baseados na abordagem quantitativa, pois análises, interpretações e inferências serão feitas sobre dados estatísticos, sendo complementada pela abordagem qualitativa que atribui um significado aos resultados frente ao processo seletivo da assistência estudantil.

2.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foi descrita a metodologia utilizada para a realização esta pesquisa, destacando suas etapas, processos, métodos e técnicas. Esta metodologia é constituída por três etapas associadas às fases da metodologia CRISP-DM. O próximo capítulo apresenta a fundamentação conceitual deste trabalho.

3 FUNDAMENTAÇÃO CONCEITUAL

Este capítulo apresenta conceitos, definições e outras informações sobre o *Knowledge Discovery in Databases* (KDD), a mineração de dados e suas principais tarefas e técnicas, bem como as métricas de avaliação de algoritmos de DM e a metodologia CRISP-DM que será utilizada na aplicação da mineração de dados desta pesquisa.

Todas as informações relacionadas a esses assuntos servem de base conceitual para o desenvolvimento deste trabalho.

3.1 O PROCESSO DE DESCOBERTA DO CONHECIMENTO EM BANCO DE DADOS

Devido à constante evolução tecnológica, a velocidade de coletar e armazenar dados vem se tornando muito maior do que a de processar e analisar. Consequentemente, as organizações geram grandes e múltiplas bases de dados sem o conhecimento das informações valiosas que estão deixando de ser analisadas e que poderiam auxiliar na tomada de decisão ou no planejamento estratégico. Um dos fatores que cooperaram para o surgimento de grandes bases de dados ao longo dos últimos anos foi a redução dos custos de armazenamento. Entretanto, o grande volume desses dados acaba excedendo a habilidade técnica e a capacidade humana na sua análise e interpretação (GOLDSCHIMIDT; PASSOS, 2005).

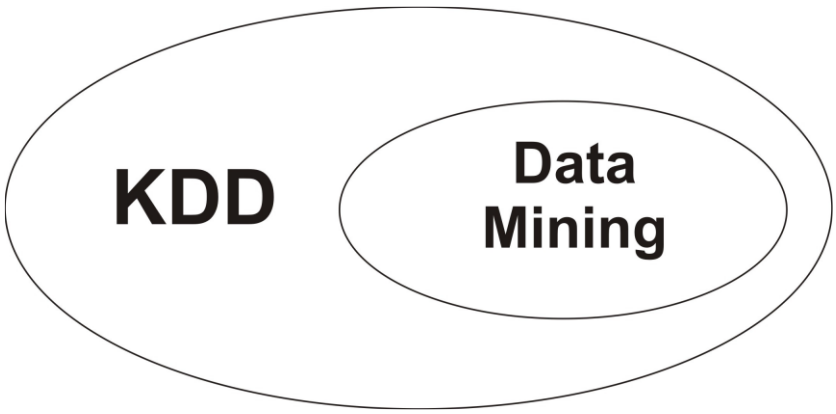
A análise de grandes quantidades de dados pelo ser humano é inviável sem o auxílio de ferramentas computacionais especificas que automatizem esse processo. Neste contexto, torna-se essencial o desenvolvimento de ferramentas que auxiliem os seres humanos, automatizando as tarefas como a análise, a interpretação e o relacionamento dos dados para que possam gerar o conhecimento, auxiliando na tomada de decisão e na definição de estratégias de ação em cada contexto de aplicação (GOLDSCHMIDT; PASSOS, 2005).

A Descoberta de Conhecimento em Banco de Dados (DCBD ou *Knowledge Discovery in Databases* – KDD) foi então desenvolvida para atender a este contexto. O termo KDD foi formalizado por Fayyad, em 1989, para expressar um processo de

várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis, a partir de dados armazenados em grandes volumes (FAYYAD et al., 1996). É um processo não trivial, porque normalmente encontra-se complexidade no decorrer de sua execução. O termo interativo significa que haverá a necessidade de um elemento que seja responsável pelo controle do processo, o qual na maioria das vezes é o usuário, e que este usuário poderá optar pela retomada em qualquer uma das etapas do processo. Já o termo iterativo significa que o processo pode ser repetido diversas vezes de forma integral ou parcial para se chegar a um resultado satisfatório e, a cada repetição, gera um resultado parcial, o qual será usado na próxima repetição, possibilitando assim realizar sucessivos refinamentos (FAYYAD et al., 1996).

Alguns autores consideram sinônimos os termos KDD (*Knowledge Discovery in Databases*) e Mineração de Dados (*Data Mining*) como é o caso de Han et al. (2011) e Wang (2005), porém a maioria dos autores considera a mineração como sendo parte do processo de KDD. Em seu texto de formalização, Fayyad et al. (1996) estabelecem os limites e diferenças de cada área, nas quais o termo KDD refere-se a todo processo de descoberta de conhecimento útil de um banco de dados, enquanto que a mineração de dados está inserida como a etapa principal neste processo. A diferença pode ser visualizada na Figura 2.

Figura 2– Relação entre o KDD e o Data Mining

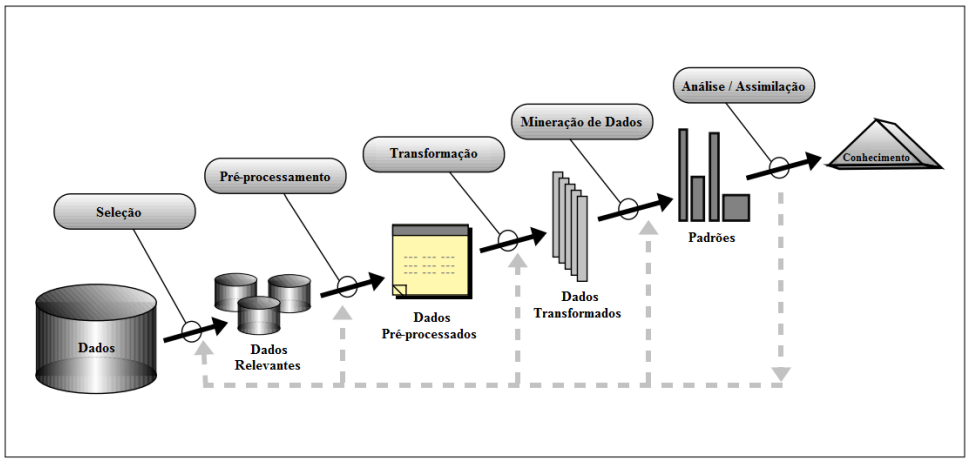


Fonte: Elaborada pelo Autor (2020)

3.1.1 Etapas do Processo de KDD

As etapas definidas por Fayyad et al. (1996) para o processo de descoberta de conhecimento em banco de dados são apresentadas na Figura 3.

Figura 3– Etapas do Processo de KDD



Fonte: FAYYAD et al. (1996)

O processo de KDD é iniciado por meio da compreensão do domínio da aplicação e da definição dos objetivos a serem alcançados. Assim sendo, o usuário define o que deseja conhecer e quais são seus objetivos com o conhecimento a ser adquirido.

A primeira etapa do processo é a seleção dos dados, a qual possui um impacto relevante na qualidade do resultado final, uma vez que é nesta fase que é selecionado o conjunto de dados que contém as possíveis variáveis (atributos) ou amostra de dados sobre os quais será feita a análise para descoberta do conhecimento. Esta etapa pode torna-se complexa, pois dependendo do caso, o conjunto de elementos pode ser proveniente de várias fontes diferentes de dados e possuir diversos formatos, sendo necessário criar uma nova base de dados a partir desses elementos.

A segunda etapa é o pré-processamento dos dados na qual é realizada a limpeza desses dados por meio de várias operações como remoção de dados inválidos e redundantes, a padronização dos valores das variáveis, além do tratamento e remoção dos ruídos e de dados discrepantes ao conjunto (*outliers*).

A terceira etapa visa uma transformação dos dados, a fim de diminuir o número de variáveis envolvidas no processo. Para facilitar a utilização dos dados 111pelas técnicas de mineração de dados é necessário fazer certas adequações no conjunto de dados de acordo com o objetivo da tarefa.

Na quarta etapa ocorre a mineração de dados, a qual consiste em aplicar técnicas e algoritmos de aprendizagem de máquina a fim de extrair informações e padrões sobre grandes bases de dados.

Após a realização da mineração dos dados, inicia-se a quinta e última etapa. As predições ou os padrões enumerados devem ser interpretados e avaliados, ou seja, deve ser realizado um trabalho de pós-processamento, visando verificar o que consiste à nova descoberta e se o objetivo principal foi alcançado de acordo com algum critério definido pelo especialista do domínio da aplicação ou do processo do KDD. Os resultados obtidos da mineração são avaliados quanto a sua qualidade, utilidade e relevância. Assim sendo, a interpretação, compreensão e aplicação desses resultados tornarão o conhecimento adquirido pelo processo de KDD um indicador relevante para a tomada de decisão.

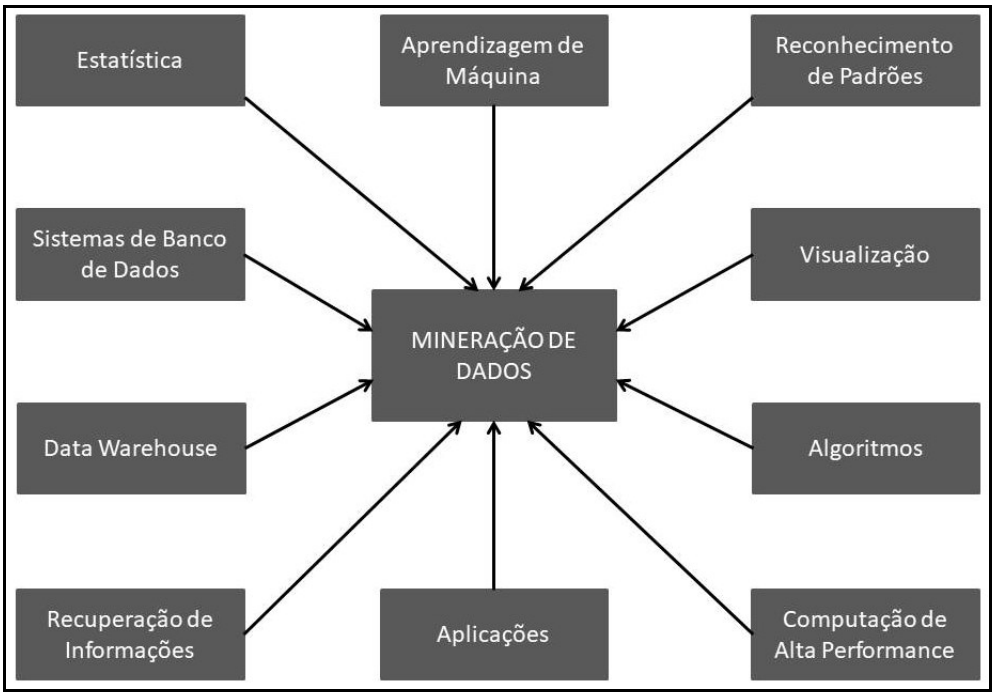
3.2 MINERAÇÃO DE DADOS

A mineração de dados pode ser vista como um resultado natural da evolução da tecnologia da informação. Isso se deve ao constante crescimento nos últimos anos dos sistemas de informações e do grande volume de bases de dados gerados por eles, além da redução do custos com a capacidade de armazenamento, fazendo com que as bases de dados só tendam a continuar crescendo (HAN et al., 2011).

Assim, a mineração de dados (em inglês, Data Mining) é um conjunto de ferramentas e técnicas que, por meio do uso de algoritmos, tem a capacidade de explorar um conjunto de dados, extraindo ou evidenciando padrões e predizendo o conhecimento (FAYYAD et al., 1996). Segundo Tan et al. (2009), a mineração de dados é um processo de descoberta automática de informações em grades volumes de dados. Para Han et al. (2011), é um processo de descoberta de padrões que representam o conhecimento armazenado em grandes bases de dados. De acordo com Witten et al. (2011), a mineração de dados tem como objetivo realizar predições a partir de dados ou transformar os dados em informações na forma de fatos ou padrões.

O processo da mineração de dados tem integrado técnicas de outras áreas do conhecimento como estatística, aprendizado de máquina, banco de dados, reconhecimento de padrões, recuperação da informação, algoritmos, entre outras, tornando-se assim objeto de estudo multidisciplinar. Essa característica multidisciplinar das pesquisas relacionadas à mineração de dados tem ampliado tanto a sua aceitação quanto a sua aplicação (HAN et al., 2011). A Figura 4 mostra as áreas associadas à multidisciplinaridade da mineração de dados.

Figura 4– Multidisciplinaridade da Mineração de Dados



Fonte: Adaptada de Han et al. (2011)

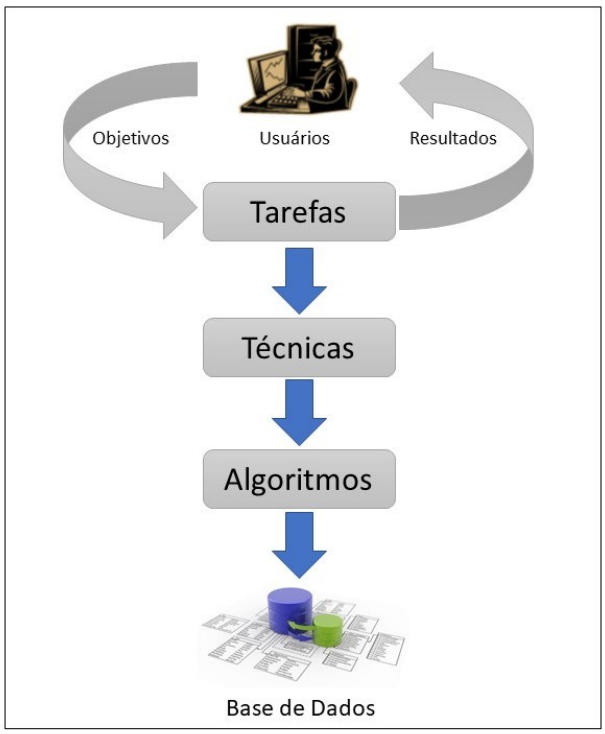
3.3 TAREFAS DA MINERAÇÃO DE DADOS

Na mineração de dados várias tarefas podem ser realizadas de acordo com os objetivos que se pretende alcançar. A tarefa em si consiste em especificar qual tipo de conhecimento que se quer obter a partir dos dados armazenados ou que tipo de regularidades ou padrões que se buscam descobrir, ou ainda quais os tipos de padrões que podem se desviar do habitual (TAN et. al., 2009).

Uma tarefa de mineração de dados pode utilizar diferentes técnicas para alcançar seus objetivos, com isso, geralmente cria-se uma dependência entre elas. Essas técnicas de mineração especificam métodos ou algoritmos visando garantir a

descoberta dos padrões que interessam. Sendo assim, uma técnica pode utilizar diferentes tipos de métodos ou algoritmos de mineração de dados para realizar determinada tarefa. A Figura 5 logo abaixo ilustra a relação entre esses três elementos.

Figura 5– Relação entre Tarefas, Técnicas e Algoritmos de DM



Fonte: Elaborada pelo Autor (2020)

Segundo Peña-Ayala (2014) e Han et al. (2011), as tarefas de mineração de dados de acordo com os objetivos pretendidos podem ser classificadas em preditivas ou descritivas.

As tarefas descritivas têm como objetivo estabelecer correlações, associações e tendências entre os atributos da base de dados analisada, resumindo os relacionamentos entre esses dados e encontrando informações relevantes que até então eram difíceis de serem visualizadas. Essas tarefas muitas vezes possuem uma natureza exploratória sendo frequentemente necessária a utilização de técnicas de pós-processamento para validar e explicar os resultados (TAN et al., 2009).

As tarefas preditivas têm como objetivo permitir estimar valores desconhecidos ou futuros de variáveis de interesse a partir de valores de outras variáveis que são conhecidos e que são independentemente relacionadas (HAND et al., 2001). Geralmente o atributo a ser classificado ou previsto é denominado de

variável alvo ou atributo classe, enquanto os demais atributos utilizados para realizar a predição são denominados de atributos independentes ou variáveis preditoras.

Nesse contexto, de acordo com Peña-Ayala (2014) e Han et al. (2011), as tarefas de classificação e regressão são classificadas como preditivas e as tarefas de associação e agrupamento são classificadas como descritivas. O Quadro 2 exemplifica algumas técnicas que podem ser aplicadas em uma determinada tarefa da mineração de dados.

Quadro 2– Tarefas de DM e algumas das suas Técnicas

Tipos de Tarefas de DM	Tarefas de DM	Técnicas de DM
Preditivas	Classificação	Árvore de Decisão Regras de Classificação Classificação Bayesiana Regressão Logística SVM Redes Neurais
	Regressão	Métodos Estatísticos Redes Neurais
Descritivas	Associação	Regras de Associação Padrões Sequenciais
	Agrupamento	Clusterização

Fonte: Elaborado pelo Autor (2020)

No estudo realizado por Peña-Ayala (2014) foram levantadas quais as tarefas de mineração de dados que foram mais utilizadas sobre bases de dados educacionais, com análise de 242 trabalhos entre os anos de 2010 e 2013, sendo apontada a tarefa de classificação como a mais utilizada nesses trabalhos com 42,15%, seguida pela tarefa de agrupamento (26,86%), a tarefa de regressão (15,29%) e regras de associação (6,61%). Levando em consideração não apenas esse estudo, mas também outras fontes que são citadas neste trabalho no capítulo dos trabalhos relacionados, podem-se constatar que essas são as tarefas da mineração de dados mais utilizadas. Sendo assim, essas quatro tarefas serão descritas a seguir.

3.3.1 Classificação

Uma das tarefas mais conhecidas e utilizadas é a classificação, a qual consiste em descobrir uma função capaz de mapear um conjunto de registros da base de dados em determinados grupos de rótulos categóricos previamente

definidos, denominados de classes. Uma vez definida esta função, ela deve ser aplicada para classificar novos registros que não foram previamente categorizados, podendo assim prever a classe na qual os registros se enquadram (MICHIE et al., 1994).

De acordo com Goldschmidt e Passos (2005), essa tarefa classifica as amostras em classes pré-definidas de acordo com a semelhança dos atributos das amostras da base de dados que já foi analisada. Sendo assim, segundo Han et al. (2011), o conjunto de dados fornecidos é analisado, no qual cada dado já contém o rótulo e com isso indica a qual categoria ele pertence. Dessa forma, o modelo dessa tarefa vai “aprendendo” a classificar novos dados.

Em suma, a tarefa de classificação gera um modelo que se baseia na análise prévia de um conjunto de dados de treinamento para criar o classificador, com isso ela geralmente exige que um objeto ou dado seja comparado com outros dados ou objetos que supostamente pertençam a classes previamente definidas (HAN; KAMBER, 2006).

3.3.2 Associação

A tarefa da descoberta de regras de associação consiste na busca por itens que frequentemente ocorram de forma simultânea em amostras de banco de dados (GOLDSCHMIDT; PASSOS, 2005). É utilizada na descoberta de padrões que identificam características altamente associadas entre os dados, visando encontrar relações entre os atributos. Esses padrões descobertos são normalmente representados por subconjunto de características, mas principalmente na forma de regras de implicação (HAN et al., 2011).

De acordo com Han e Kamber (2006), uma regra de associação é definida da seguinte forma: Se X então Y ou $X \rightarrow Y$, onde X e Y são conjuntos de termos e $X \cap Y = \emptyset$. Diz-se que X é o antecedente da regra, enquanto Y é o seu consequente, onde essa regra pode ter vários atributos tanto no antecedente quanto no consequente.

3.3.3 Regressão

A tarefa de regressão consiste na busca por uma função com capacidade para mapear amostras de base de dados em valores reais. Esta tarefa é bastante similar à tarefa de classificação, sendo diferenciada pelo atributo alvo que é restrito apenas a atributos numéricos (MICHIE et al., 1994).

De acordo com Witten et al. (2011), embora existam outros métodos com esse objetivo, é a metodologia estatística mais frequentemente aplicada para previsão numérica. Além disso, esta tarefa também pode ser utilizada para identificação de tendências de distribuição baseando-se nas informações disponíveis (HAN et al.,2011).

3.3.4 Agrupamento

A tarefa de agrupamento, também chamada de clusterização ou segmentação, tem como objetivo agrupar registros de uma base de dados em subconjuntos ou grupos denominados de clusters, de tal forma que esses registros sejam mais semelhantes entre si, ou seja, compartilhem de propriedades comuns e que os distinguem de outros grupos (clusters) (FAYYAD, 1996).

Dessa forma, quanto maior a semelhança (homogeneidade) dos registros dentro de um grupo e maior as suas diferenças para os outros grupos, melhor é o agrupamento, já que o seu objetivo é aumentar a semelhança intragrupo e reduzir a semelhança intergrupo (GOLDSCHMIDT; PASSOS, 2005).

Esta tarefa se diferencia da classificação porque não necessita que os dados sejam previamente categorizados. Os rótulos podem ser gerados a partir do próprio processo de agrupamento. Após as formações dos clusters, cada conjunto formado pode ser visto como uma classe, a partir da qual as regras podem ser derivadas (HAN et al., 2011).

3.4 TÉCNICAS DE MINERAÇÃO DE DADOS

Após a definição da tarefa a ser utilizada, deve-se então definir a técnica e consequentemente o algoritmo ou os algoritmos que serão aplicados na fase de mineração dos dados. Para a realização de cada tarefa, diversas técnicas podem

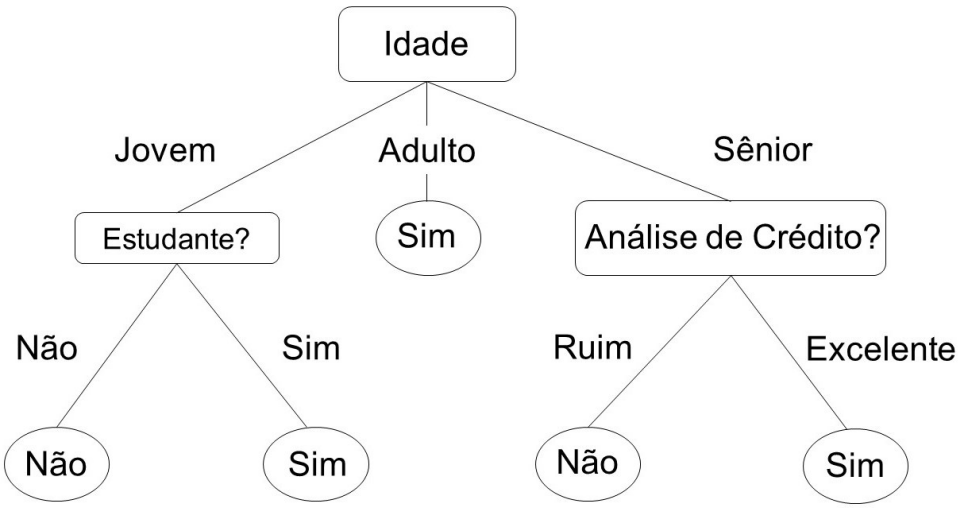
ser aplicadas, testadas, comparadas e até combinadas, podendo cada uma delas apresentar vantagens e desvantagens, sendo que o fator que definirá a escolha de determinada técnica é o problema em questão que deve ser tratado (VIEIRA, 2014).

Assim sendo, é importante ter o conhecimento sobre as técnicas de mineração para poder definir quais poderão ser aplicadas de acordo com o problema em questão. A seguir, são descritas as técnicas de mineração de dados mais utilizadas e citadas na literatura.

3.4.1 Árvore de Decisão

A técnica de árvore de decisão consiste em um fluxograma com estrutura de árvore, no qual cada nó interno representa um teste de um dos atributos, cada ramo representa o resultado desses testes e cada nó folha (nó terminal) representa o rótulo da classe (HAN et al., 2011). O nó mais alto que fica no topo da árvore é chamado de nó raiz (WITTEN et al., 2011). A Figura 6 mostra um exemplo de uma árvore de decisão.

Figura 6– Exemplo de Árvore de Decisão



Fonte: Elaborada pelo Autor (2020)

De acordo com Tan et al. (2009), uma árvore de decisão possui três tipos de nós que são descritos a seguir:

- Um nó raiz (topo) é onde começa a árvore, o qual não possui arestas chegando. Esse nó representa a condição de teste para o atributo que melhor separa as instâncias que possuem características diferentes;
- Nós internos são os nós intermediários, com cada um possuindo uma aresta chegando e duas saindo. Assim, cada ramificação representa um resultado de um teste; e
- Nós folhas (terminais), os quais possuem uma única aresta chegando e nenhuma saindo. Cada um desses nós representa o atributo alvo ou determinado rótulo de uma classe.

As árvores de decisão funcionam da seguinte forma: dado um atributo associado ao rótulo da classe, o qual será exibido nas folhas da árvore, o atributo mais significativo é definido como nós raiz (topo da árvore), de forma recursiva novos atributos são adicionados como nós internos, os quais são realizados testes sobre eles, gerando ramos, sendo que cada ramo representa uma saída desse teste, particionando os nós em mais internos. Assim sendo, um caminho é percorrido da raiz até o nó terminal (folha), o qual detém a predição da classe para uma instância. Dessa forma, as árvores de decisão podem ser representadas por regras de classificação (HAN et al., 2011).

Existem medidas que podem ser utilizadas para selecionar os atributos que melhor dividem as instâncias em classes distintas. A principal medida utilizada em árvores de decisão é chamada de ganho de informação. Segundo Taconeli (2008), em árvores de decisão deve-se buscar árvores com tamanho reduzido e baixa heterogeneidade em seus nós terminais. Geralmente, o método da poda é utilizado para identificar e reduzir os ramos, principalmente aqueles que podem refletir ruídos na árvore, aumentando assim a sua capacidade preditiva (JAMES et al., 2013).

As árvores de decisão são muito utilizadas, pois a sua representação em forma é intuitiva e de fácil assimilação humana. Outros fatores que contribuem para a sua ampla utilização são: a sua aprendizagem é rápida e simples, contribuindo para um bom desempenho e precisão; podem manipular dados multidimensionais; e além disso não requerem nenhum conhecimento do domínio ou configurações de parâmetros, sendo uma técnica apropriada para descobertas de conhecimentos exploratórios (HAN et al., 2011).

3.4.2 Regras de Classificação

Regras de classificação são uma técnica que tem como objetivo encontrar os relacionamentos entre os atributos e a classe alvo, de modo que a regra encontrada possa predizer a qual classe uma nova instância pertence. Regras são uma forma simples e intuitiva para representar informação e conhecimento (HAN et al., 2011). As regras de classificação são regras do tipo “IF.....THEN” e podem ser expressas da seguinte forma (TAN et al., 2009):

IF condição THEN conclusão

Um exemplo da regra é a R1:

R1: IF transporte = ônibus AND renda_familiar <= 998 THEN recebe_auxilio = SIM

A parte “IF” (lado esquerdo) da regra é chamada de antecedente ou precondição da regra. A parte “THEN” (lado direito) da regra é chamada de consequente. Na parte do antecedente, a precondição consiste no teste de um ou mais atributos (por exemplo transporte = ônibus AND renda_familiar <= 998). Já o consequente da regra consiste em predizer a classe, no caso desse exemplo, está sendo predito que o aluno vai receber o auxílio. Dada uma instância, se a condição do antecedente da regra for verdadeira diz-se que a regra foi satisfeita ou que a regra cobre a instância (HAN et al., 2011).

Existem métricas que podem ser aplicadas para avaliar essas regras, geralmente são utilizadas a cobertura e a precisão. A cobertura é o percentual de instâncias que foram cobertas pela regra em relação ao total, isto é, todos os valores das variáveis foram verdadeiros no antecedente da regra. Para a precisão, é analisada primeiramente a cobertura das instâncias e calculado o percentual delas que a regra pode classificar corretamente (HAN et al., 2011).

De acordo com Kampff (2009), a principal diferença entre as regras geradas por algoritmos específicos da técnica de regras de classificação para as regras extraídas de árvores de decisão é a restrição de que toda regra obtida por meio de uma árvore tenha o atributo raiz em sua condição. Outra importante diferença é a ordem das regras que estabelece um fluxo sequencial de decisão, priorizando a predição da classe para a primeira regra. Assim sendo, após uma instância ser classificada nenhuma regra posterior poderá ser aplicada sobre ela.

3.4.3 Classificação Bayesiana

Essa técnica é um modelo probabilístico, baseado em estatística de probabilidade condicional por meio da aplicação do teorema de Bayes (HAN et al., 2011). De acordo com esse teorema, pode-se encontrar a probabilidade de um determinado evento ocorrer, dada a probabilidade de um outro evento que já ocorreu (TAN et al., 2009). Em suma, os classificadores bayesianos também chamados de Naive Bayes, dado que A e B são eventos, calcula a probabilidade de A dado B, conforme mostrar a Equação 3.1:

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)} \tag{3.1}$$

Onde:

- P(A) e P(B) são as probabilidades de A e B ocorrerem independentemente um do outro;
- P(B|A) é a probabilidade de B ocorrer dado que A tenha ocorrido; e
- P(A|B) é a probabilidade condicional, ou seja, é a probabilidade de A ocorrer dado que B já ocorreu.

Segundo Zhang (2004) e Castro e Ferrari (2016), os classificadores bayesianos apresentam uma alta taxa de acurácia e desempenho de processamento quando aplicadas a grandes bases de dados.

Esse classificador é denominado ingênuo (Naive) pois o mesmo assume que os atributos são independentes, mesmo sabendo que isso é importante na maioria dos problemas práticos de classificação. Apesar dessa premissa “ingênua” e simples, segundo Mitchell (1997) essa premissa torna o Naive Bayes eficiente, pois o mesmo lê os dados do conjunto de treinamento uma única vez e calcula todas as probabilidades, além disso, ele pode ser usado de forma incremental ou alterado facilmente com a inclusão de novos dados, uma vez que as probabilidades podem ser recalculadas.

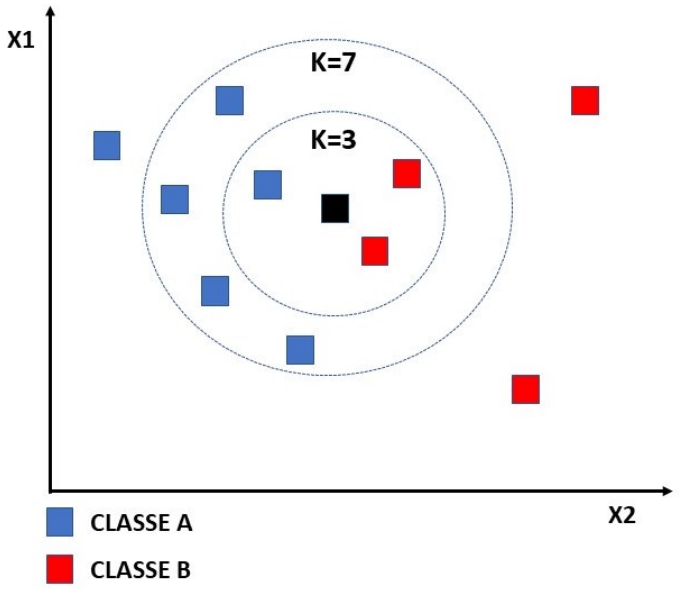
3.4.4 KNN (k-Nearest Neighbors)

O k-*Nearest Neighbors* ou k vizinhos mais próximos (kNN) é uma técnica baseada em instâncias, cujo objetivo é determinar o rótulo de classificação de uma

nova instância baseando-se na sua similaridade em relação às instâncias do conjunto de treinamento. Essa similaridade é medida por meio da utilização de uma função que calcula a distância entre essas instâncias, assim o algoritmo consiste em rotular a nova instância baseado nas k instâncias mais próximas. Assim sendo, a variável k é quem vai representar a quantidade de vizinhos mais próximos que serão utilizados para definir a qual classe a nova instância pertence (KOTSIANTIS, 2007).

De acordo com Silva et al. (2016), na tarefa de regressão o KNN serve para estimar valores de uma função real obtidos por meio do cálculo da média dos vizinhos mais próximos. Já para a tarefa de classificação é verificada a classe mais frequente entre os k vizinhos mais próximos. A Figura 7 mostra o processo de classificação, aplicando a técnica do kNN.

Figura 7– Exemplo da Classificação do KNN para dois Valores de K



Fonte: Adaptada de Han et al. (2011)

O exemplo ilustrado destaca a importância da variável k e como ela influencia na classificação de novas instâncias. Se k=3, a nova instância pertence à classe B, pois tem dois vizinhos mais próximos, ou seja, classe mais frequente. Para k=7, a instância pertenceria à classe A, com quatro vizinhos mais próximos. Nesse contexto, o valor de k geralmente é definido de forma empírica variando de acordo com o problema em questão ou o conjunto de dados, quando a escolha do k ótimo pode ter o seu desempenho (precisão) avaliado na fase de validação cruzada (ZHANG et al., 2017).

Segundo Kotsiantis (2007), existem diversas formas de calcular a distância entre as instâncias e assim mensurar a sua proximidade, porém a mais simples e mais utilizada é a distância euclidiana, a qual é mostrada na Equação 3.2:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.2}$$

A principal desvantagem da técnica do KNN é que ela exige um alto consumo computacional, já que para cada nova instância deve ser calculada a distância para todas as instâncias de treinamento, o que pode ir sendo ampliado, caso o conjunto de dados possua um grande número de variáveis.

3.4.5 Regressão Logística

A regressão logística é uma técnica estatística utilizada para obter um resultado discreto ou categórico binário, ou seja, geralmente ela é dicotômica (BAKER; INVENTADO, 2014). Esta técnica trabalha com um modelo probabilístico, já que o resultado da função, aplicada a cada instância, deve retornar a probabilidade de ela pertencer a uma das classes alvo. Assim sendo, a regressão logística tem a capacidade de estimar a probabilidade de sucesso (pi) e, consequentemente, a de fracasso (1 – pi) de determinados dados avaliados (DELEN, 2011).

Segundo Kleinbaum e Klein (2010), uma das vantagens da regressão logística no processo de classificação de uma variável binária é permitir o uso de conjunto de variáveis independentes, tanto numéricas quanto categóricas. Em suma, esse modelo consiste em relacionar um conjunto de n variáveis independentes V1, V2,.....,Vn a uma variável resultante e dicotômica X, com a variável X podendo assumir o valor 0 ou 1. Por exemplo, o modelo pode permitir estimar a probabilidade da ocorrência de um evento para (Y=1) ou (pi) (HOSMER et al., 2008). As equações 3.3 e 3.4 mostram as funções de probabilidade de sucesso (pi) e fracasso (1-pi):

$$P(Y=1) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)} \tag{3.3}$$

E, consequentemente,

$$P(y=0)=1-p(y=1)=\frac{1}{1+\exp(\beta_0+\beta_1 x_1+\ldots+\beta_n x_n)} \quad (3.4)$$

Onde as variáveis β são os parâmetros do modelo, e a etapa de aprendizagem ou treinamento ocorre por meio da estimação dos coeficientes dessas variáveis, os quais são obtidos pelo método da máxima verossimilhança (MONTGOMERY et al., 2011).

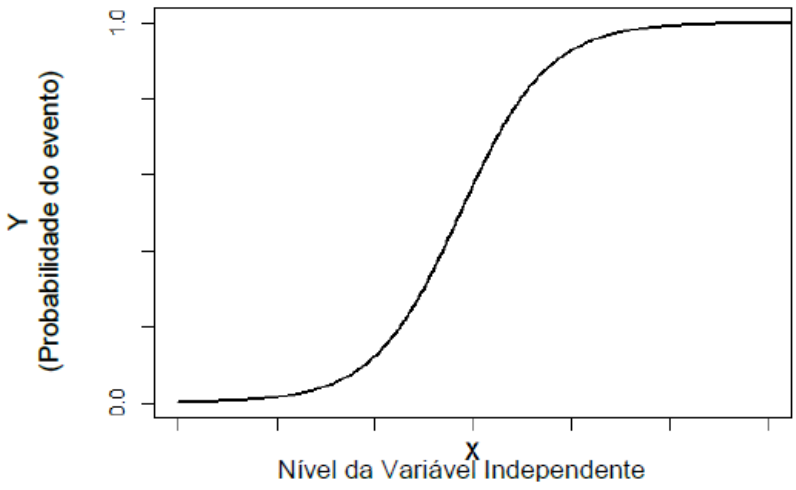
De acordo com Hair et al. (2009), o modelo de regressão logística também é chamado de análise ou função logit, sendo basicamente limitada à previsão de duas classes, embora existam formas alternativas que permitam trabalhar com mais de duas classes. A função logit é expressa pela Equação 3.5:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} \quad (3.5)$$

Onde β0, β1, ... βk, são os coeficientes das variáveis que explicam a ocorrência de um determinado evento, e pi o número de variáveis preditoras. A função logit utiliza a forma específica de curva logística em forma de S para manter o domínio entre 0 e 1 conforme é mostrado na Figura 8.

Dessa forma, para cada ocorrência, o modelo prevê um valor probabilístico entre 0 e 1. Se a probabilidade prevista for maior que 0,50, então estima-se que o resultado seja 1 (sucesso, o evento ocorreu); caso contrário, o resultado é estimado como 0 (fracasso, o evento não ocorreu). Esse limiar pode ser configurado de acordo com a necessidade e o contexto da aplicação da técnica.

Figura 8– Relação Logística entre Variáveis Dependente e Independente



Fonte: Hair et al. (2009)

A probabilidade se aproxima de 0 quando os valores da variável independente forem muito baixos, porém nunca alcançarem tal valor. No mesmo contexto, quanto maior for o valor da variável independente, os valores previstos crescem na curva. Entretanto, a inclinação passa a diminuir, e a probabilidade até se aproxima de 1, mas não chega a exceder esse valor, ou seja, a probabilidade nunca fica abaixo de 0 ou acima de 1 (HAIR et al., 2009).

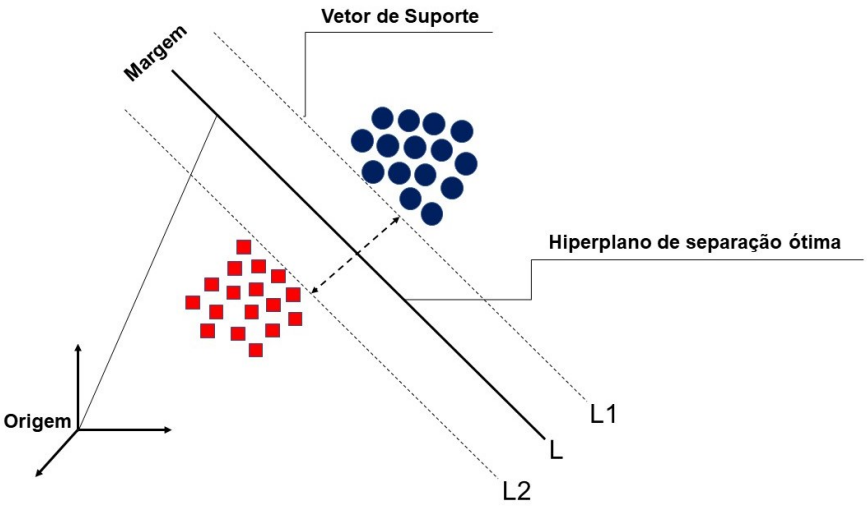
A regressão logística, assim como os algoritmos de árvores de decisão, pode ser uma boa escolha quando se trata de facilitar a interpretação do modelo, já que os coeficientes gerados para cada variável mostram a sua relevância na classificação da classe alvo do estudo (HOSMER; LEMESHOW, 2000).

3.4.6 Support Vector Machines

A *Support Vector Machines* ou Máquinas de Vetores de Suporte (SVM) é uma técnica baseada na teoria de aprendizagem estatística e otimização matemática. Esta técnica utiliza o conceito de hiperplano, o qual é uma generalização de um plano em diferentes dimensões. A partir disso, o algoritmo realiza um mapeamento não linear para transformar os dados de treinamento inicial em uma dimensão de alta dimensionalidade, e nessa nova dimensão ele procura o hiperplano de separação ótimo linear que será utilizado para separar as instâncias de cada classe (HAN et al., 2011).

O algoritmo SVM busca construir um hiperplano (L) com a máxima margem de separação (distância entre as bordas, L1 e L2), ou seja, se baseia nas instâncias das bordas que são chamados de vetores de suporte, conforme mostra a Figura 9.

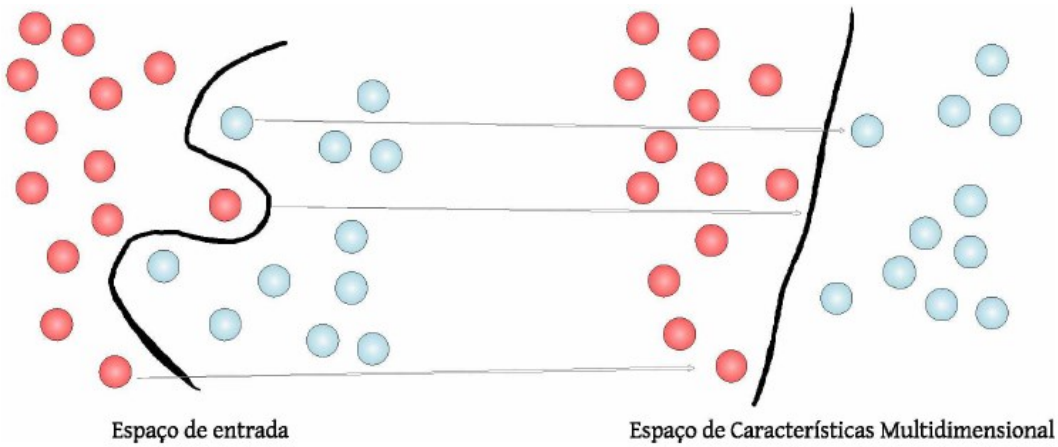
Figura 9– Exemplo do SVM



Fonte: Adaptada de Nascimento et al. (2009)

Quanto maior a margem de separação das classes do classificador, maior será sua capacidade de generalização, permitindo assim, aumentar a capacidade de predição de uma nova instância (MUCHERINO et al., 2009; TAN et al., 2009). O SMV pode ser utilizado tanto para padrões lineares como não-lineares, sendo os lineares aqueles cujo conjunto de dados pode ser separado linearmente por um hiperplano e os não-lineares aqueles nos quais não houve essa possibilidade (LORENA; CARVALHO, 2007). Segundo Hastie et al. (2009), para esses padrões não-linearmente separáveis podem ser utilizadas as representações de kernels que permitem realizar um mapeamento dos dados de entrada para um vetor multidimensional, assim possibilitando que os dados possam ser separados linearmente por um hiperplano, conforme mostra a Figura 10.

Figura 10– Transformação da SVM



Fonte: Adaptada de Hastie et al. (2009)

Nesse contexto, assim como outros algoritmos de mineração de dados, o desempenho do SVM depende do ajuste de parâmetros, principalmente por tratar de solucionar problemas do mundo real (VIANA et al., 2007). Assim sendo, o SVM possui alguns parâmetros que podem ser ajustados, os quais podem ocasionar a perda do desempenho do algoritmo caso os valores desses parâmetros sejam definidos inadequadamente. O principal parâmetro do SVM é a função *kernel*, descrita seguir.

Função Kernel, segundo Haykin (2009), é utilizada pelo SVM para separar dados linearmente ou não-linearmente separáveis. Por meio dessa função, o SVM consegue acesso a espaços complexos de forma simplificada. A definição do *kernel* a princípio é feita de acordo com a natureza do problema e sendo realizada de forma empírica. Nesse contexto, as funções kernel mais utilizadas são: as polinomiais, os RBF (Radial-Basis Function) e as sigmoides, mostradas no Quadro 3.

Quadro 3– Resumo das funções de Kernel

Tipo de Kernel	Função $k(x,x_i)$	Comentários
Polinomial	$(\mathbf{x}^T\mathbf{x}+1)^p$	Dimensão *p* é definida a priori pelo usuário
RBF	$\exp(-\frac{1}{2\sigma^2}\|\|x - x_i\|\|^2)$	A largura σ^2, comum a todos os kernels, é definida a priori pelo usuário
Sigmoidal	$\tanh(\beta_0\mathbf{x}^T\mathbf{x}_i+\beta_1)$	O teorema de Mercer é satisfeito apenas para valores de β_0,β_1.

Fonte: Adaptado de Haykin (2009)

De acordo com Haykin (2009) e Braga et al. (2007), os resultados obtidos pelo SVM são equivalentes e, muitas vezes, melhores comparados a outros algoritmos, como por exemplo as redes neurais artificiais. No entanto, o seu desempenho depende da seleção dos parâmetros da função kernel. Caso isso seja feito de forma inadequada, pode resultar na perda de sua acurácia.

3.4.7 Redes Neurais Artificiais

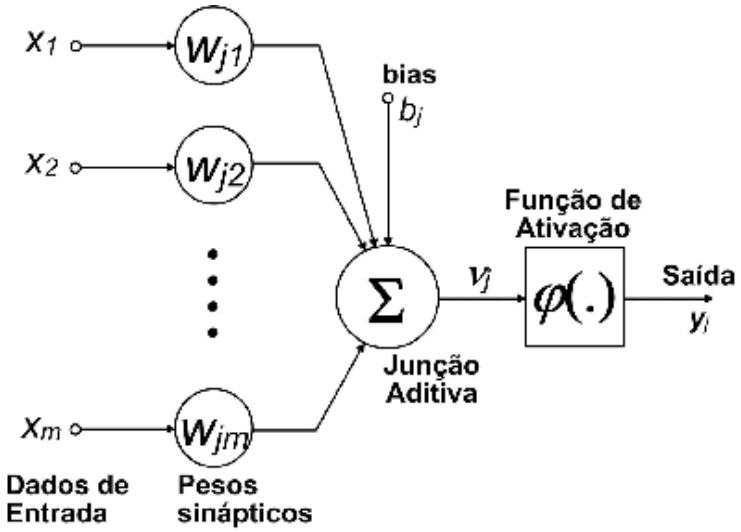
As redes neurais artificiais (RNA) são técnicas que buscam simular o funcionamento dos neurônios do cérebro humano, incluindo algumas características como: aprendizado, realização de novas descobertas, velocidade de

processamento, processamento simultâneo e capacidade de adaptação (HAYKIN, 2009).

As RNA são compostas de múltiplos nós conectados em rede. Assim como a disposição dos neurônios de um cérebro humano, esses nós conectados interagem entre si, o que pode ser comparável às conexões sinápticas (SILVA et al., 2010).

Cada um desses nós pode receber uma entrada e realizar operações, sendo a saída de cada nó obtida por meio do uso de uma função denominada função de ativação, e por fim, essas saídas são transmitidas para o próximo nó (neurônio). Cada conexão entre os nós possui um peso associado, o qual durante o processo de aprendizado pode ser ajustado para ampliar a sua capacidade de classificar corretamente uma instância (HAYKIN, 2009). A Figura 11 mostra a estrutura de um neurônio artificial (nó) que também pode ser chamado de *perception*.

Figura 11– Estrutura de um Neurônio Artificial



Fonte: Adaptada de Han et al. (2011)

Segundo Silva et al. (2010) e Han et al. (2011), e conforme ilustrado na Figura 11, a estrutura de um neurônio artificial é composta pelos seguintes elementos:

- Dados de entrada (Xj) - são dados advindos do meio externo que representam os valores de variáveis de uma aplicação especifica;
- Pesos sinápticos (Wjm) - são os valores que servirão para ponderar os dados de entrada e assim quantificando a relevância de cada um;
- Junção Aditiva (Σ) - é uma junção de soma responsável pela adição dos sinais ponderados de entrada pelos seus respectivos pesos do neurônio;

- Função de ativação - é geralmente não linear e limita a saída do neurônio dentro de um intervalo de valores; e
- Saída (yj) - é valor resultante produzido pelo neurônio.

As arquiteturas das redes neurais definem a forma como os seus neurônios artificiais estarão interligados e consequentemente na forma como as informações serão direcionadas entre eles. Nessas arquiteturas, uma RNA pode ter uma estrutura em camadas dividas em três partes: camada de entrada, camadas intermediárias ou ocultas e camadas de saída (SILVA et al., 2010). Já levando em consideração a constituição de suas camadas e a disposição dos seus neurônios e as suas interligações, as RNA podem ser classificadas em: redes de camada única, redes multicamadas, redes recorrentes ou realimentadas e redes em estrutura reticulada (SILVA et al., 2010).

As redes neurais possuem algumas desvantagens como a necessidade de um longo período de treinamento e de ajustes finos de parâmetros, além dos modelos produzidos por elas serem de difícil interpretação, já que não é possível identificar claramente as relações entre as suas entradas e saídas (HARRISON, 1998). No entanto, essa técnica pode ser utilizada em uma variedade de aplicações, como por exemplo nas tarefas de classificação, regressão e agrupamento. Além disso, ela pode identificar padrões para os quais não foi treinada (HAYKIN, 2009).

3.5 MÉTRICAS DE AVALIAÇÃO

Uma questão importante, na aplicação de modelos classificadores, é avaliar o desempenho desses classificadores durante o processo de predição das classes para novas instâncias de dados. Para isso, diversas métricas foram desenvolvidas e são utilizadas de acordo com os objetivos da mineração ou do contexto no qual os dados estão sendo utilizados. O ideal é saber como o modelo se comporta quando novas instâncias de dados são aplicadas, ou seja, diferente dos dados usados na definição dos seus parâmetros (SILVA et al., 2016).

Nesta seção serão apresentadas as métricas mais utilizadas para avaliar o desempenho dos modelos de classificação com o objetivo de assegurar qual o modelo é o melhor ou mais preciso na solução do problema do estudo em questão.

3.5.1 Matriz de Confusão

A avaliação do desempenho de um modelo de classificação consiste na análise de sua capacidade de classificar ou separar as classes de forma correta. Para isso, um método muito utilizado é denominado de matriz de confusão (HAN; KAMBER, 2006). Em uma matriz de confusão os resultados obtidos pelo classificador são representados por uma matriz bidimensional, sendo uma linha e coluna para cada classe, conforme a Figura 12. Nessa matriz, cada elemento mostra o número de instâncias classificadas correta ou incorretamente considerando o conjunto de testes utilizado (FAYYAD et al., 1996).

Figura 12– Matriz de Confusão

		Classe Verdadeira *(Referência)*	
		Negativo	Positivo
Classe Predita *(Modelo)*	Negativo	**VN**	**FN**
	Positivo	**FP**	**VP**

Fonte: Adaptada de Fawcett (2006)

Onde dado um classificador e instâncias a serem classificadas, há quatro resultados possíveis:

- Verdadeiros positivos (VP) - é o número de instâncias positivas corretamente classificadas como positivas pelo classificador;
- Falsos positivos (FP) - é o número de instâncias negativas incorretamente classificadas como positivas pelo classificador;
- Verdadeiros negativos (VN) - é o número de instâncias negativas corretamente classificadas como negativas pelo classificador; e
- Falsos negativos (FN) - é o número de instâncias positivas incorretamente classificadas como negativas pelo classificador.

Portanto, tendo um classificador e um conjunto de instâncias (ou conjunto de teste), uma matriz de confusão de dimensões de 2 por 2 pode ser construída mostrando as disposições do conjunto de instâncias para as duas classes. A matriz gerada é bastante útil, pois serve de base e possibilita o cálculo de várias outras métricas de avaliação de modelos classificadores (FAWCETT, 2006; SILVA et al., 2016).

Essas métricas que também podem ser utilizadas para avaliar o desempenho de um modelo classificador são descritas nas subseções seguintes e foram baseadas nos estudos de Fawcett (2006), Tan et al. (2009) e Silva et al. (2016).

3.5.2 Acurácia (*Accuracy*)

A acurácia é a métrica da taxa de acerto global de um classificador. Ela é definida pela razão do número de classificações corretas das duas classes (VP + VN) pelo número total de instâncias classificadas (VP+VN+FP+FN). Sendo assim, o melhor caso a ser obtido para a matriz de confusão é o preenchimento somente da diagonal principal, o que resultaria em uma acurácia de 100%. A sua fórmula é mostrada na Equação 3.6:

$$\text{Acurácia} = (VP + VN) / (VP+VN+FP+FN) \quad (3.6)$$

É importante ressaltar que além da acurácia, outras métricas devem ser utilizadas para analisar o desempenho dos algoritmos classificadores, já que um algoritmo pode se diferenciar do outro por meio das taxas de acerto e erro na classificação das instâncias positivas e negativas. Por exemplo, se um algoritmo classificador possuir uma elevada taxa de erro para o falso negativo, ele não é um classificador adequado para ser aplicado. Nesse contexto, outras métricas muito utilizadas na avaliação do desempenho de classificadores são descritas a seguir.

3.5.3 Precisão (*Precision*)

A precisão é definida como a razão entre a quantidade de acertos de verdadeiros positivos (VP) pela quantidade de todos os exemplos classificados como positivos (VP + FP). Assim sendo, a precisão calcula o percentual de frequência que o modelo acerta quando este faz uma previsão positiva. Quanto maior for a precisão, menor será o erro de falsos positivos gerado pelo classificador. A sua fórmula é mostrada na Equação 3.7:

$$\text{Precisão} = VP / (VP + FP) \quad (3.7)$$

3.5.4 Sensibilidade (*Recall*)

Já a sensibilidade, também chamada de *recall*, é definida como a razão entre a quantidade de acertos de verdadeiros positivos (VP) pela soma de verdadeiros positivos com falsos negativos (VP + FN). Ou seja, o recall calcula o percentual de acertos de verdadeiros positivos previstos de forma correta pelo classificador, traduzindo assim, a confiança de que todas as instâncias positivas foram encontradas pelo modelo, no qual um recall alto indica que o classificador produziu poucos resultados positivos classificados como falso negativos. A sua fórmula é mostrada na Equação 3.8:

$$\text{Recall} = VP / (VP + FN) \tag{3.8}$$

Um alto *recall* indica que o classificador produziu poucos exemplos positivos classificados como falso negativos.

É importante ressaltar que construir um modelo que alcance valores altos para ambas as métricas é um desafio para os algoritmos de classificação. Na maioria das vezes, os modelos são desenvolvidos priorizando a otimização de apenas uma dessas métricas (*precision* ou *recall*). Por exemplo, um modelo no qual cada instância de teste seja declarada com sendo positiva terá uma alta precisão, mas um baixo *recall*. Inversamente a isso, um modelo no qual todas as instâncias forem declaradas como sendo positivas terá um alto *recall*, mas uma baixa precisão.

3.5.5 F-Measure

A medida F (do inglês, F-Measure ou F1) é uma medida que calcula a média harmônica ou média ponderada das métricas *Precision* e *Recall*. Quanto mais próximo de 1 o valor da medida, melhor é o algoritmo e quanto mais próximo de 0, pior é o algoritmo. A sua fórmula é mostrada na Equação 3.9 (WITTEN et al., 2011):

$$\text{F-Measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \tag{3.9}$$

Assim sendo, a F-Measure mede a eficiência do classificador levando em consideração o erro nas duas classes, sendo necessário que a classificação correta das duas classes aumente para que o valor da métrica aumente.

3.5.6 Estatística Kappa (*Statistic Kappa*)

A estatística Kappa é uma métrica que avalia o nível de concordância de uma tarefa de classificação. Ela mede o grau de concordância entre as classes preditas e observadas, deduzindo o número esperado de acertos (utilizando uma classificação ao acaso) do número real de acertos do classificador (WITTEN et al.,2011). Segundo Burn e Weir (2011), a estatística Kappa é utilizada para avaliar o grau de concordância além do que seria esperado tão somente pelo acaso, e o seu valor varia de -1 a 1, onde o valor 1 significa que os dois observadores concordam exatamente já que classificaram todas as instâncias da mesma forma; o valor 0 significa que não há nenhuma relação entre as classificações dos dois observadores, acima da concordância de acasos que eram esperados; e, por fim, o valor -1 significa os dois observadores classificaram exatamente o oposto um do outro: enquanto um observador sempre classifica como “SIM”, o outro sempre classifica como “NÃO”. A fórmula para calcular o índice Kappa foi definida por Cohen (1960), sendo apresentada conforme a Equação 3.10:

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e} \tag{3.10}$$

Onde (P_o) é equivalente à proporção de acertos nas classes, ou a acurácia observada, e (P_e) é a probabilidade de concordância esperada (concordância ao acaso) ou precisão esperada (COHEN, 1960). De acordo com Koch e Landis (1977), foram estabelecidos rótulos de acordo com os intervalos dos valores do índice Kappa para descrever a força de concordância associada, conforme mostra o Quadro 4.

Quadro 4– Rótulos da Força de Concordância do Índice Kappa

Estatística Kappa	**Força da Concordância**
< 0.00	Pobre
0.00 a 0.20	Leve
0.21 a 0.40	Justa
0.41 a 0.60	Moderada
0.61 a 0.80	Substancial
0.81 a 1.00	Quase Perfeito

Fonte: Adaptado de Koch e Landis (1977)

3.5.7 Área sob a curva ROC

A curva *Receiver Operation Characteristics* (ROC) é uma outra importante métrica de avaliação do desempenho de classificadores. Segundo Fawcett (2006), muitos trabalhos têm demonstrado a importância da curva ROC na avalição e comparação de classificadores em detrimento da utilização somente da métrica de acurácia, geralmente considerada insuficiente enquanto medida de desempenho de classificadores. A curva ROC utiliza uma abordagem gráfica bidimensional na qual no eixo vertical (eixo y) é plotada a taxa de verdadeiros positivos e no eixo horizontal (eixo x) é plotada a taxa de falsos positivos. Cada curva ROC é um modelo de classificação no qual cada ponto dessa curva é obtido por meio do cálculo das taxas de verdadeiros e falsos positivos a partir da matriz de confusão desse modelo (PRATI et al., 2008). Sendo assim, um modelo de classificação pode ser considerado bom quanto mais próxima do vértice superior esquerdo do gráfico a sua curva ROC estiver.

Uma outra abordagem para avaliar qual modelo possui o melhor desempenho como classificador é o método de calcular Área Sob a Curva ROC, também chamada de AUC (*Area Under Curve*). Devido ao fato de uma curva ROC ser uma representação bidimensional do desempenho do classificador, tornou-se necessário transformá-la em uma simples medida escalar para representar o desempenho e assim poder comparar os classificadores. Sendo a AUC uma porção da área da unidade de quadrado, o seu valor irá sempre estar entre o valor mínimo 0,0 (pior caso) e o valor máximo e melhor medida 1,0 (modelo ideal). No entanto, não deve possuir valores menores que 0,5 para classificadores reais, pois essa é a área que produz a linha diagonal com início no ponto mínimo (0,0) e fim no ponto máximo (1,1) do plano (FAWCETT, 2006).

Quando a curva ROC é construída interligando vários pontos, a área pode ser calculada pela regra do trapézio por meio da Equação 3.11:

$$AUC = \sum_{i=1}^{i=k-1} (FP_{i+1} - FP_i)(TP_i - TP_{i+1})/2 \qquad (3.11)$$

Sendo assim, uma AUC com valor 1,0 representa um classificador perfeito; já uma AUC com valor de 0,5 representa um classificador sem valor. Tanto Mehdi et al. (2011), quanto Devi e Sehgal (2017), utilizam um sistema de pontuação tradicional

para classificar o desempenho de um classificador levando em consideração o valor obtido pela AUC, o qual pode-se ver abaixo:

- 0,90 a 1,0 = Excelente – Caso a AUC obtida pelo classificador esteja nessa faixa, ele é considerado um excelente classificador para a tarefa de classificação;
- 0,80 a 0,90 = Bom – Caso a AUC obtida pelo classificador esteja nessa faixa ele é considerado um bom classificador para a tarefa de classificação;
- 0,70 a 0,80 = Justo – Caso a AUC obtida pelo classificador esteja nessa faixa ele é considerado um classificador justo para a tarefa de classificação, mas não é inferior aos classificadores que obtiveram a AUC igual ou superior a 0,80;
- 0,60 a 0,70 = Pobre – Caso a AUC obtida pelo classificador esteja nessa faixa ele é considerado um classificador pobre e, portanto, não sendo adequado para a tarefa de classificação; e
- 0,50 a 0,60 = Falho – Caso a AUC obtida pelo classificador esteja nessa faixa ele é considerado um péssimo classificador e, portanto, também não é adequado para a tarefa de classificação.

3.5.8 Média Ponderada

Nesta pesquisa também foi utilizada como medida de avaliação dos algoritmos de classificação, a média ponderada referente às métricas: Taxa de Verdadeiros Positivos (VP Rate), Taxa de Falsos Positivos (FP Rate), Precisão, Recall, F-Measure e a Área sob a Curva ROC (AUC). A média ponderada é calculada a partir do resultado da respectiva métrica selecionada para determinada classe, a quantidade de registros referente à respectiva classe e a quantidade total geral de registros pertencentes a todas as classes do problema.

Como exemplo do cálculo da média ponderada, a Tabela 1 mostra uma matriz de confusão e a taxa de verdadeiros positivos obtida por um dado algoritmo que classificou os alunos que poderiam estar aptos ou não-aptos a receber a assistência estudantil do IFAM Campus Manaus Zona Leste.

Tabela 1– Exemplo de uma Matriz de Confusão para Classificação à Assistência Estudantil

	Classificações referente à Classe APTO	Classificações referente à Classe NÃO APTO	Taxa de Verdadeiros Positivos (VP Rate) relativos a cada Classe	Classe ALVO
Amostras pertencentes à Classe de APTO	2700	156	0,945	APTO
Amostras pertencentes à Classe de NÃO APTO	715	60	0,077	NÃO APTO
Média Ponderada calculada para a métrica VP Rate:			**0,760**	

Fonte: Elaborada pelo Autor (2020)

Para o cálculo da média ponderada para a métrica VP Rate, conforme Santana Júnior (2018) foi definida a seguinte Equação 3.12:

$$MP = \frac{(VMClasseA \times QTDClasseA) + (VMClasseB \times QTDClasseB)}{QTDGERAL} \quad (3.12)$$

Onde:

- MP: Indica a Média Ponderada para a métrica VP Rate;

VMClasseA: Indica o valor da métrica VP Rate para a classe APTO;

- QTDClasseA: Indica a quantidade total de registros para a classe APTO;
- VMClasseB: Indica o valor da métrica VP Rate para a classe NÃO APTO;
- QTDClasseB: Indica a quantidade total de registros para a classe NÃO APTO;
- QTDGERAL: Indica a quantidade total de registros pertencentes a todas as classes.

Substituindo pelos valores da matriz de confusão da Tabela 1 tem-se:

$$MP = \frac{(0,945 \times 2856) + (0,077 \times 775)}{3631} = 0,760$$

A mesma fórmula pode ser derivada para as outras métricas alterando-se apenas os valores de VMClasseA e VMClasseB para os respectivos valores da

métrica a ser usada. Nesta pesquisa a média ponderada será considerada junto aos algoritmos testados para analisar o desempenho dos mesmo durante os experimentos.

3.6 SELEÇÃO DE ATRIBUTOS

Segundo Bellman (1961), na mineração de dados, a quantidade de classificadores que devem ser considerados aumenta exponencialmente com a quantidade de atributos do conjunto de dados selecionado, tornando mais difícil para o algoritmo de mineração definir um modelo preciso. Este problema é chamado de maldição da dimensionalidade. De acordo com Liu e Motoda (1998b), várias pesquisas sobre seleção de atributos indicam que uma grande quantidade de atributos irrelevantes pode introduzir ruídos nos dados, ocasionando problemas na aprendizagem do algoritmo e consequentemente erros na classificação. Além disso, tais problemas podem ocasionar dificuldade em extrair informações que sejam realmente relevantes para classificação.

A seleção de atributos é uma técnica de redução de dimensionalidade na qual são identificados e removidos atributos irrelevantes ou redundantes (HAN et al.,2011). Assim sendo, esta técnica é responsável por identificar, selecionar ou até mesmo ordenar por nível de relevâncias os atributos que possuem maior relação ou que estejam fortemente ligados ao atributo classificador, ou seja, a classe à qual pretende-se predizer (LIU; MOTODA, 1998a).

A seleção de atributos é considerada uma das formas de melhorar o desempenho do processo de classificação já que é uma técnica muito útil na redução do custo de processamento e de simplificação dos modelos gerados na classificação (LÔBO, 2015). Segundo Borges (2006), quando a seleção de atributos é aplicada junto à tarefa de classificação, a taxa de erro do classificador é minimizada. Assim sendo, a seleção de atributos torna-se fundamental para a redução do tempo de treinamento do algoritmo e na melhora da acurácia de um modelo de classificação. A técnica de seleção de atributos está dividida em três abordagens: *Filter*, *Wrapper* e *Embedded* (LEE, 2005; SAEYS et al., 2007; FACELI et al., 2011; LAZAR et al., 2012), descritas a seguir.

- Abordagem *Filter* (Filtro): Essa abordagem de seleção de atributos introduz um processo separado, ocorrendo na etapa de pré-

processamento, ou seja, antes da aplicação dos algoritmos de aprendizagem propriamente ditos, no qual o objetivo é filtrar atributos irrelevantes, segundo algum critério, antes que o aprendizado ocorra (Baranauskas, 2001; Blum and Langley, 1997; John et al., 1994). Essa abordagem leva em consideração as características gerais do conjunto de dados original para selecionar os atributos mais relevantes, sendo totalmente independente do algoritmo de aprendizado (indutor), o qual receberá como entrada apenas o subconjunto de atributos que foram selecionados pelo filtro. De acordo com Freitas (1998), o objetivo do filtro é selecionar um subconjunto de atributos que preserva a informação pertinente no conjunto original de atributos.

- Abordagem *Wrapper*: Essa abordagem também ocorre externamente ao algoritmo de aprendizado, porém utiliza o mesmo algoritmo como uma "caixa preta" para analisar, a cada iteração, o subconjunto de atributos em questão (LEE, 2005). Os métodos aplicados na abordagem *Wrapper* geram um subconjunto candidato de atributos a partir do conjunto de treinamento e aplicam o algoritmo de aprendizado sobre este subconjunto. Por fim, utilizam a precisão resultante do algoritmo aplicado para avaliar o subconjunto de atributos em questão (LEE, 2005; PARMEZAN et al., 2012). Esse processo é repetido para cada subconjunto de atributos até que um critério de parada determinado seja satisfeito (DASH, LIU, 1997; KOHAVI, JOHN, 1997). Na abordagem *Wrapper* os atributos são avaliados por meio de estimativas de precisão geradas por algoritmos de aprendizado pré-determinados (FREITAS, 1998).
- Abordagem *Embedded* (Embutida): Nessa abordagem a seleção de atributos é realizada internamente pelo próprio algoritmo de aprendizagem (LEE, 2005). Dessa forma, os métodos aplicados nessa abordagem selecionam o subconjunto de atributos dinamicamente na construção do modelo de classificação, durante a etapa de treinamento, na qual geralmente são específicos para um dado algoritmo de aprendizagem (PARMEZAN et al., 2012).

Segundo Kohavi e John (1997), as abordagens geralmente utilizadas principalmente em modelos de classificação são o *Filter* e o *Wrapper*. As técnicas da abordagem *Wrapper* geralmente possuem melhor capacidade preditiva, pois

avaliam cada subconjunto de atributos utilizando um algoritmo de aprendizagem. Entretanto, requerem várias execuções desse algoritmo, o que eleva o custo computacional em relação à abordagem *Filter* (PAES et al., 2013). De acordo com Lima (2016), outra desvantagem do *Wrapper* é essa dependência de um algoritmo de aprendizagem, pois a sua solução para a seleção de atributos acaba não sendo generalizada. Assim sendo, um subconjunto selecionado poder ser o melhor para um classificador A, mas pode não ser o melhor para um classificador B.

Já a abordagem *Filter* possui algumas vantagens em relação às outras como a execução rápida, a possiblidade de ser aplicada em qualquer tipo de conjunto de dados e a capacidade de reduzir em até 50% a quantidade de atributos dos conjuntos de dados (HALL, 2000). Segundo Parmezan et al. (2012), os métodos da abordagem *Filter* são menos custosos computacionalmente, além de serem baseados somente nas características intrínsecas dos próprios dados de entrada, possibilitando que eles sejam acoplados a qualquer processo de aprendizado. Além disso, de acordo com Goswami e Chakrabarti (2014), a seleção de atributos baseada na abordagem *Filter* é a mais utilizada tanto em pesquisas acadêmicas, quanto na indústria.

Dentre os métodos mais utilizados para avaliar um subconjunto de atributos, um dos que mais se destaca é CFS (*Correlation-based Feature Selection*) (LIU; YU, 2005). O CFS é um método de seleção de atributos supervisionado, baseado na abordagem *Filter* e com avaliação de subconjuntos (HALL, 2000). O algoritmo CFS classifica os subconjuntos gerados com uma função de avaliação heurística baseada em correlação. Esse algoritmo avalia o mérito de um subconjunto de atributos em função da correlação individual de cada atributo com a classe (atributo-classe) e o grau de correlação entre os atributos (atributo-atributo) (HALL, 1999). O subconjunto com o maior mérito encontrado pela heurística será selecionado.

A maior vantagem do CFS é que diferentemente da maioria das técnicas baseadas na abordagem *Filter*, ele não só analisa a relevância do atributo, mas também a redundância, diminuindo assim a dimensionalidade. O CFS não faz nenhuma adequação para trabalhar com dados anômalos, tornando-o sensível ao desbalanceamento de classes.

3.7 BALANCEAMENTO DAS CLASSES

O problema de desbalanceamento de classes, no domínio da classificação, ocorre quando o número de instâncias de uma classe é muito menor que o número de instâncias da outra classe (GU et al., 2008). Nesse contexto, os algoritmos de aprendizagem tendem a ignorar as classes menos frequentes (classes minoritárias) e a valorizar as mais frequentes (classes majoritárias), fazendo com que o classificador não seja capaz de classificar as instâncias de dados correspondentes às classes menos frequentes corretamente (MARQUEZ-VERA et al., 2013). Este problema pode se tornar mais relevante se a classe com o menor número de instâncias é aquela cuja classificação correta pode representar maior interesse para os objetivos do projeto de mineração de dados.

De acordo com Márquez-Vera et al. (2013), os algoritmos de classificação foram desenvolvidos com objetivo de maximizar a taxa de acurácia global, sendo independente da distribuição das classes. Nesse contexto, segundo Han e Kamber (2006), as técnicas de classificação assumem que a distribuição das classes está balanceada e que os custos dos erros são iguais para todas as classes.

Visando reduzir esses problemas, existem duas abordagens para realizar o balanceamento das classes: o *Undersampling* e o *Oversampling* (Liu et al., 2008). O *Undersampling* consiste na redução do número de instâncias, por meio de uma escolha aleatória, somente das classes majoritárias e mantendo a proporção das classes minoritárias. Já o *Oversampling* consiste em replicar, criar sinteticamente ou projetar instâncias, aumentando a proporção das classes minoritárias e mantendo a proporção das classes majoritárias (CHAWLA, 2009).

De acordo com Márquez-Vera et al. (2013), uma técnica da abordagem *Oversampling* que é amplamente utilizada em modelos de classificação é o algoritmo SMOTE *(Synthetic Minority Over-sampling Technique*) (CHAWLA et al., 2002).

O algoritmo SMOTE ajusta a frequência relativa entre classes majoritárias e minoritárias, introduzindo instâncias sintéticas nas classes minoritárias e aumentando as suas proporções. Essas instâncias sintéticas são introduzidas por meio de uma interpolação entre os k vizinhos mais próximos das instâncias minoritárias já existentes (WITTEN et al., 2011). Isso faz com que as instâncias

sejam escolhidas de forma aleatória dependendo da quantidade da amostragem, não sendo simples réplicas dos dados existentes.

As principais vantagens dessa técnica são que ela foi criada para evitar o sobre ajuste (*overfitting*) e, por ser uma técnica de *Oversampling*, impedir que sejam removidas instâncias que podem ser importantes para o processo de aprendizado, o que poderia reduzir significativamente o conjunto de dados utilizado para testes, reduzindo a qualidade do modelo preditivo.

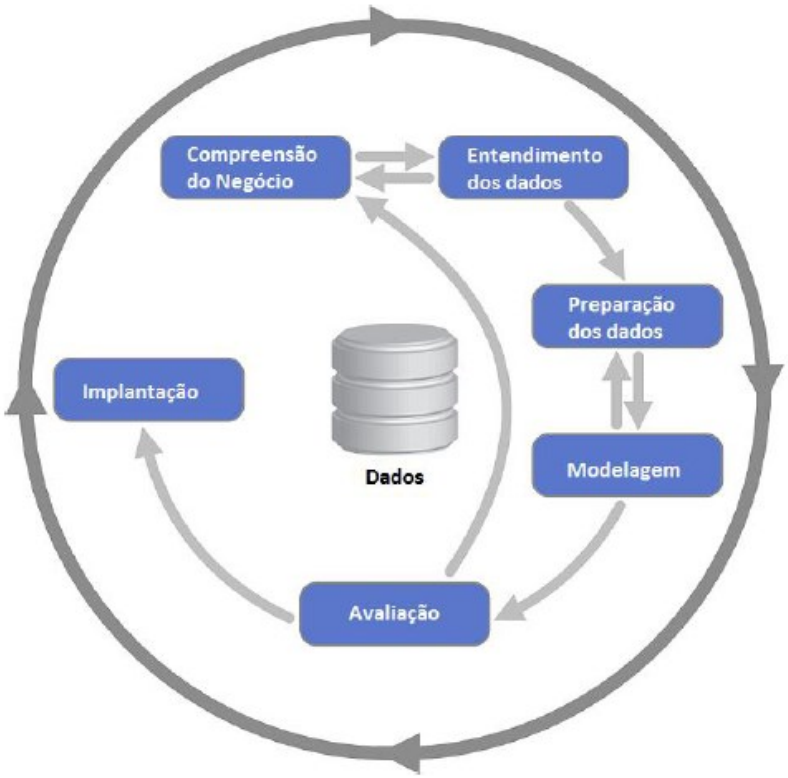
3.8 METODOLOGIA CRISP-DM

Com a crescente utilização da mineração de dados nas mais diversas áreas do conhecimento, surgiu a necessidade de estabelecer normas, padrões e processos que pudessem guiar a implementação de projetos de DM em qualquer área. Diante dessa necessidade, em 1996, um consórcio formado por diversas empresas da área de DM como NCR, DaimlerChrysler AG, SPSS Inc. e OHRA desenvolveram o CRISP-DM (*Cross Industry Standard Process for Data Mining*) (CHAPMAN et al., 2000; SHEARER,2000).

Segundo Chapman et al. (2000), a metodologia CRISP-DM define um processo padronizado para o desenvolvimento de projetos de mineração de dados, possuindo um conjunto de fases, sendo independente da área de aplicação e da tecnologia a ser utilizada no projeto, tornando-se assim bastante versátil quanto ao tipo de negócio ou técnicas aplicadas, mas sempre mantendo uma forma estruturada e sistêmica.

O ciclo da metodologia CRISP-DM é composto por seis fases: compreensão ou entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação (CHAPMAN et al., 2000; SHEARER,2000). A Figura 13 ilustra as fases do ciclo, o qual é bastante flexível já que não é linear, não sendo rígido quanto a sequência de execução das fases, pois permite retornar à fase anterior sempre que necessário. Entretanto, cada fase necessita dos resultados de sua fase anterior (CHAPMAN et al., 2000).

Figura 13– As Fases do Ciclo da Metodologia CRISP-DM



Fonte: Shearer (2000)

Como pode ser visto na Figura 13, as setas internas indicam as dependências mais frequentes e importantes entre as fases, de modo que dependendo do resultado de cada fase ou dos objetivos a serem alcançados pelo projeto de DM, o usuário poderá retornar a uma fase anterior para realizar alguma alteração e seguir com o processo. Já o círculo de setas externas simboliza a natureza cíclica do processo de mineração de dados (CHAPMAN et al., 2000; SHEARER,2000). Detalhando as fases, tem-se:

- Entendimento do Negócio (*Business Understanding*) - É a primeira fase do processo e consiste em compreender os objetivos do projeto de mineração de dados a partir da perspectiva do negócio. Esse conhecimento é convertido em um problema de DM. Além disso, é definido também um plano preliminar do projeto DM para alcançar esses objetivos;
- Entendimento dos Dados (*Data Understanding*) - A segunda fase do processo consiste em conhecer os dados visando identificar quais os mais relevantes para solução do problema em questão. Nesta fase também é possível identificar problemas na qualidade dos dados e detectar subconjuntos de dados e variáveis interessantes que podem ser utilizadas

para formular hipóteses sobre informações ocultas;

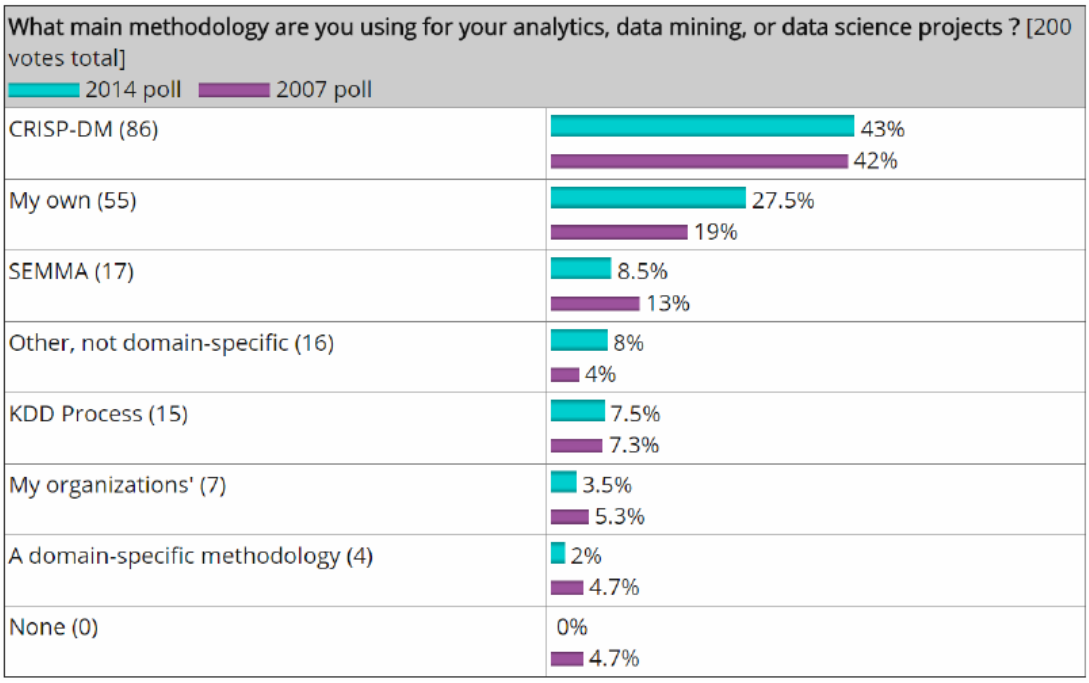
- Preparação dos Dados (*Data Preparation*) - A terceira fase do processo consiste em realizar ações necessárias para construir o conjunto final de dados. Essas ações incluem a seleção de tabelas e variáveis, transformação e limpeza dos dados para que estejam preparados para a aplicação das técnicas de mineração de dados. Estima-se que a fase de preparação de dados geralmente exige entre 50 e 70% do tempo e esforço de um projeto de mineração de dados;
- Modelagem (*Modeling*) - A quarta fase do processo consiste na seleção e aplicação das técnicas de mineração de dados e seus algoritmos de acordo com o problema de mineração definido. Nesta fase são utilizadas as configurações padrão dos parâmetros dos modelos e estes vão sendo ajustados para obtenção de melhores resultados. Em suma, as técnicas ou algoritmos de DM são aplicados visando obter os melhores resultados sobre o conjunto final de dados gerado na fase de preparação dos dados, e assim, possibilitar a resolução do problema definido na fase de entendimento do negócio;
- Avaliação (*Evaluation*): A quinta fase do processo consiste em revisar o modelo construído e avaliar os resultados obtidos, verificando se as necessidades levantadas na fase entendimento do negócio foram atendidas, ou seja, se o objetivo do negócio foi alcançado, de forma que, ao final desta etapa, deve-se definir se os resultados obtidos serão utilizados ou não. O processo permite retornar para a primeira fase, caso a avaliação dos resultados não seja satisfatória ou não alcance os objetivos do negócio; e
- Implantação (Deployment) - A sexta e última fase do processo consiste na organização e apresentação dos resultados e conhecimentos obtidos, de forma que os usuários possam utilizá-los de forma prática e efetiva dentro da organização nos processos de tomada de decisão. Dependendo dos requisitos, esta fase pode gerar desde a produção de relatórios até a repetição do processo de DM, na qual os seus conceitos devem ser apresentados se forma intuitiva aos usuários finais.

Além do CRISP-DM existem outras metodologias aplicadas em projetos de DM, como por exemplo o SEMMA (*Sample, Explore, Modify, Model, Assess*). É outra

metodologia padronizada para projetos de mineração de dados que foi desenvolvida pelo SAS Institute. O SEMMA também é um processo cíclico de cinco fases, no entanto, esse modelo é menos utilizado do que o CRISP-DM, pois a sua aplicação está muito ligada ao software de mineração da SAS (AZEVEDO; SANTOS, 2008).

No estudo realizado por Piatetsky (2014), cerca de 200 participantes que utilizam a mineração de dados responderam a uma enquete que buscava saber qual a metodologia de DM é a mais utilizada. O estudo ainda fez um comparativo entre os anos de 2007 e 2014, por meio do qual foi mostrado que a metodologia CRISP-DM continua sendo a mais usada nos projetos de DM, sendo que em 2007 eram 42% e em 2014 foram 43%. Outro dado relevante mostrado nesse estudo foi o aumento no número de pessoas que utilizam a sua própria metodologia, subindo de 19% em 2007 para 27,5% em 2014. Já a metodologia SEMMA foi a terceira mais utilizada, conforme mostra a Figura 14.

Figura 14- Pesquisa - Qual metodologia principal você está usando para seus projetos de análise, de mineração de dados ou de ciência dos dados?



Fonte: Piatetsky (2014)

Portanto, devido ao fato de ser uma metodologia de grande aceitação e bastante utilizada tanto na área acadêmica como na empresarial, sendo desenvolvida por empresas com foco na área de DM, a metodologia CRISP-DM será utilizada nesta pesquisa para aplicação da mineração de dados sobre a base de

dados do programa de assistência estudantil do IFAM Campus Manaus Zona Leste.

3.9 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo, inicialmente foram explicados os conceitos sobre o KDD e a mineração de dados, sendo destacadas as diferenças entre ambos. Além disso, são descritas as características que levam a mineração de dados a ser cada vez mais utilizada nos dias atuais. Este estudo objetivou descobrir se a mineração de dados pode ser utilizada para solucionar o problema tratado nesta pesquisa. Logo em seguida, são descritas as tarefas de DM mais utilizadas sobre bases de dados educacionais e como cada uma dessas tarefas pode utilizar diferentes técnicas na resolução de um problema. Foram descritas também as principais técnicas de DM. Este estudo objetivou obter o conhecimento sobre as várias tarefas e técnicas para auxiliar na definição das tarefas e técnicas que serão aplicadas nesta pesquisa.

Logo após, são descritas as métricas de avaliação que serão utilizadas para avaliar o desempenho dos algoritmos de DM, já que um dos objetivos desta pesquisa é realizar uma análise comparativa dos algoritmos, além de descrever as técnicas de seleção de atributos e de balanceamento de classes. Por fim foi apresentada a metodologia CRISP-DM, sendo descrita todas as fases dessa metodologia, além de mostrar que ela é a mais utilizada por especialistas em DM e com isso justificando a sua escolha para ser empregada no processo de mineração de dados desta pesquisa.

Tendo adquirido os conhecimentos técnicos inerentes ao objeto de estudo deste trabalho, baseado no que foi exposto neste capítulo, torna-se possível analisar outras pesquisas correlatas ao mesmo contexto deste trabalho. Assim sendo, o próximo capítulo apresenta uma análise de trabalhos relacionados a esta pesquisa encontrados na literatura pesquisada, considerando alguns aspectos como a aplicação da mineração de dados sobre bases de dados de alunos, utilização de dados socioeconômicos e comparação de algoritmos ou aplicação de métricas de avaliação.

4 TRABALHOS RELACIONADOS

Buscando ressaltar a relevância tanto cientifica quanto social, bem como a contemporaneidade da linha de pesquisa aplicada neste estudo, são analisados neste capitulo os trabalhos correlatos encontrados na Revisão Sistemática de Literatura (RSL) descrita no Capítulo 2.

É importante ressaltar que na pesquisa exploratória da RSL não foram encontradas pesquisas nas quais a mineração de dados fosse aplicada sobre dados socioeconômicos de alunos visando à predição da necessidade de receber ou não bolsas de auxílios acadêmicos. Também não foram encontrados trabalhos visando à identificação de perfis de alunos em situação de vulnerabilidade socioeconômica.

Segundo Cordeiro (2017), as pesquisas aplicando a mineração de dados sobre a questão da evasão escolar vêm crescendo nos últimos anos. Foi observado nos trabalhos analisados que a mineração de dados foi aplicada visando a predição do aluno se evadir, analisando os dados acadêmicos de desempenho nas disciplinas, ou visando a identificação do perfil de aluno propenso à evasão.

Este capítulo apresenta uma descrição de cada trabalho encontrado na RSL e apresenta um quadro comparativo desses trabalhos levando em consideração os critérios definidos no Capítulo 2. Por fim, as considerações finais trazem a análise crítica de cada trabalho correlato, assim como as principais contribuições desta pesquisa em comparação com esses trabalhos.

4.1 DESCRIÇÃO DOS TRABALHOS RELACIONADOS

Mineração de Dados Educacionais: Um estudo sobre os dados socioeconômicos na educação na base de dados do INEP (SANTOS, 2019)

Em Santos (2019), o objetivo foi identificar quais os fatores socioeconômicos que influenciam no desempenho dos estudantes que prestam o ENEM, sendo analisados por região, assim como os fatores relacionados à infraestrutura oferecida pelas escolas públicas de ensino médio do estado do Pará que também poderiam exercer essa influência. As bases de dados foram obtidas junto ao INEP, das quais foram selecionados os dados sobre as notas dos candidatos bem como as suas

informações socioeconômicas. Outra base foi a do Censo Escolar, da qual foram obtidos os dados referentes as características estruturais das escolas e formação do corpo docente. Essas bases são referentes somente ao ano de 2016.

A metodologia utilizada consistiu na aplicação do KDD sobre essas bases de dados. Na fase de pré-processamento dos dados foi utilizada a técnica de PCA por meio do software WEKA, a qual permitiu reduzir a quantidade de variáveis sem perder as informações obtidas no conjunto inicial. Já na fase de mineração de dados foi aplicado o algoritmo das Redes Bayesianas, fazendo o uso do software Bayesware Discovery. A autora justificativa a sua escolha dessa técnica já que ela tem como pressuposto o relacionamento das variáveis e suas probabilidades de ocorrência.

Como resultados, a pesquisa evidencia que cada região possui um conjunto de variáveis influentes tendo apenas duas variáveis que se repetem em todas as regiões: o tipo de dependência administrativa escolar e a renda familiar, Nas regiões Norte, Nordeste e Centro-Oeste pode-se observar que fatores a respeito do aluno possuir computador em sua residência e ter acesso à Internet, influência nas suas notas obtidas no exame do Enem.

Mineração em dados do ENEM para a Predição do Desempenho Acadêmico no Âmbito da Rede Federal de Educação Tecnológica (SANTANA JÚNIOR, 2018)

Em Santana Júnior (2018), o objetivo foi aplicar técnicas de mineração de dados para classificação de desempenho de alunos no ENEM em duas classes (baixo-rendimento, bom-rendimento), a partir dos dados socioeconômicos e informações preenchidas pelos alunos no ENEM 2014. Outros objetivos atrelados ao principal são: identificar quais as características dos alunos que mais influenciam na tarefa de classificação do desempenho dos alunos no ENEM e comparar as técnicas de mineração de dados adotadas, avaliando o desempenho de cada uma delas na tarefa de classificação dos alunos.

A base de dados utilizada foi a do INEP, a qual armazena informações sobre alunos, escolas, desempenhos obtidos pelos alunos nas provas objetivas e de redação, além de informações socioeconômicas e culturais relativas aos alunos. . Foram selecionados apenas os dados referentes aos alunos que pertencem à Rede

Federal de Educação Tecnológica e pretendeu-se desenvolver classificadores voltados a este público especificamente.

Foi utilizada a metodologia CRISP-DM que define um processo padronizado para o desenvolvimento de projetos envolvendo mineração de dados e foi definida a Validação Cruzada de 10 folds (*K-fold Cross Validation*) para treinamento de testes dos algoritmos. E, por fim, os algoritmos de classificação aplicados foram JRIP, PART, J48, Random Forest, Naive Bayes e SVM e as métricas de avaliação utilizadas foram a Acurácia, Matriz de Confusão, Curvas ROC e Área Sob a Curva ROC (AUC).

Os algoritmos que obtiveram os melhores resultados junto à tarefa de classificação de desempenho acadêmico dos alunos submetidos ao ENEM foram os modelos Random Forest, SVM e Naive Bayes, porém como são modelos de caixa-preta, foram aplicados modelos de caixa branca (JRIP, PART e J48), dos quais o JRIP obteve os melhores resultados. Nesse contexto, os resultados obtidos pelo modelo Random Forest foram os melhores e mais satisfatórios quando aplicado ao conjunto de dados pré-processados durante os experimentos realizados. Levando em consideração as regras geradas, foi adotado o JRIP apenas como modelo secundário a fim de tentar explicar aos gestores educacionais como as classificações foram feitas.

Mineração de Dados Educacionais nos Resultados do ENEM de 2015 (SIMON; CAZELLA, 2017)

O estudo de Simon e Cazella (2017) teve como objetivo gerar um modelo de predição do indicador de desempenho médio na área de Ciências da Natureza e suas Tecnologias, dos alunos de escolas do ensino médio a partir de dados do ENEM 2015. A base de dados foi obtida do INEP, da qual foram selecionados somente os dados do desempenho obtidos na prova de Ciências da Natureza e suas Tecnologias e os dados socioeconômicos dos alunos que realizaram o ENEM 2015, sendo agrupados por tipo de escolas: privada, federal, estadual e municipal.

A metodologia utilizada consistiu na aplicação das fases do KDD. Na fase da mineração de dados por meio da tarefa de classificação, os autores optaram pela técnica de árvore de decisão, aplicada por meio do algoritmo J48 utilizando o software WEKA. Foi também definida no treinamento e teste do algoritmo a *cross-*

validation (validação cruzada) com o valor do *fold* igual a dez, com a variável a ser predita chamada de Média Escola. E, por fim, as métricas de avaliação utilizadas foram a Acurácia, Matriz de Confusão, Curvas ROC e Área Sob a Curva ROC (AUC).

Entre os resultados obtidos, segundo o algoritmo J48, está a identificação das variáveis independentes mais importantes hierarquicamente para a decisão sobre a classe da variável independente, respectivamente e em ordem decrescente: no primeiro nível, o Tipo Escola; no segundo nível, o Nível Socioeconômico; e no terceiro nível em diante, as variáveis alternaram de posição em função dos valores dos níveis superiores. Sobre o desempenho médio dos alunos em Ciências da Natureza e suas Tecnologias, destaca-se que os grupos que tiveram desempenho acima de 550 pontos ocorrem nas escolas privadas, apenas no nível socioeconômico muito alto; federais, nos níveis muito alto, alto e médio alto; estaduais, apenas no nível muito alto; e nas municipais, apenas no nível médio alto.

Prevendo o Desempenho dos Candidatos do ENEM através de Dados Socioeconômicos (STEARNS et al., 2017)

O estudo de Stearns et. al (2017) teve como objetivo analisar a capacidade de predizer o desempenho de estudantes na disciplina de Matemática no ENEM, baseando-se apenas em seus dados socioeconômicos. Outro objetivo relacionado ao principal foi o de comparar a capacidade de generalização de dois métodos de agrupamento por árvore de decisão para a nota dos estudantes no exame. Os autores justificam a escolha da disciplina de Matemática do ENEM devido à alta variância das notas. A base de dados foi obtida por meio do INEP, sendo referente a dados de alunos que realizaram o ENEM em 2014.

A metodologia utilizada consistiu em aplicar diretamente a mineração de dados, sendo definida a utilização da tarefa de regressão, por meio da qual foi realizada uma análise comparativa de duas técnicas de *boosting* para agregar árvores de decisão: o *Gradient Boosting* e o *AdaBoost*. Como métricas para avaliar a regressão foram utilizadas a *Mean Absolute Percentage Error* (MAPE), *Mean Absolute Error* (MAE) e *R-Squared* (R^2), sendo essas duas últimas combinadas com a *Cross-Validation* de 10 *folds*. A comparação do desempenho dos regressores

foi feita utilizando Paired t-Test sobre as amostras obtidas por meio dos *folds* da validação.

Os melhores resultados foram obtidos com o *Gradient Boosting* com alta relevância estatística. Esses resultados apontaram que os fatores socioeconômicos podem influenciar nas notas dos alunos, sendo representado em um ranking das *features*, podendo ser utilizado como base para entender o viés nas notas.

Prática da Mineração de Dados no Exame Nacional do Ensino Médio (SILVA et al., 2014)

Na pesquisa de Silva et. al. (2014), o objetivo foi identificar fatores socioeconômicos que podem estar relacionados ao desempenho dos alunos na prova do ENEM. A base de dados utilizada foi obtida por meio do INEP. Dela foram selecionados os dados do desempenho das provas e dos questionários socioeconômicos dos alunos relativos ao ENEM 2010, sendo definidos como foco da análise somente dados das capitais da região sudeste: São Paulo, Rio de Janeiro, Belo Horizonte e Vitória.

A metodologia utilizada consistiu na aplicação das fases do KDD, sendo que na fase da mineração de dados foi utilizada a tarefa de associação, aplicando a técnica de regras de associação por meio do algoritmo Apriori, sendo, por fim, os experimentos realizados utilizando o software RapidMiner 5.1. Foram realizadas quatro simulações utilizando valores diferentes para os parâmetros de suporte e confiança que são utilizados para a avaliação das regras no algoritmo Apriori. A própria quantidade de regras que foram geradas em cada simulação também foi analisada.

Os resultados identificaram fatores socioeconômicos que diminuem o desempenho dos alunos, como a renda familiar baixa, a escolaridade dos pais de nível primário e a quantidade alta de pessoas que moram com o estudante.

Predição de Desempenho de Escolas Privadas usando o ENEM como Indicador de Qualidade Escolar (ADEODATO et al., 2014)

Na pesquisa de Adeodato et. al. (2014), o objetivo foi avaliar e predizer a qualidade do ensino médio privado brasileiro a partir de dados do ENEM e do Censo Escolar. A primeira base de dados foi obtida por meio do INEP, sendo referente ao ENEM de 2011 que continha dados socioeconômicos e do desempenho dos alunos

nas provas. A segunda base foi a do censo escolar 2011 que detalha as condições das escolas, como infraestrutura e corpo docente, sendo definido como foco da pesquisa somente dados de escolas privadas.

Foram escolhidas a metodologia CRISP-DM para padronizar o processo de aplicação da mineração de dados e a abordagem Domain-Driven Data Mining (D3M), sendo utilizadas as técnicas de Regressão Logística, Árvore de Decisão e Indução de Regras. A regressão logística produziu um classificador capaz de gerar uma pontuação de propensão ao sucesso da escola, a partir das suas características e daquelas dos seus docentes, discentes e famílias. A árvore de decisão foi utilizada para explicar como seria a sequência decisória ideal na visão de um especialista humano. Já a indução de regras foi utilizada para gerar uma base de regras para explicar as decisões e identificar nichos de alta relevância no domínio.

As métricas utilizadas para avaliar o desempenho do classificador da regressão logística foram a área sob a Curva ROC (AUC_ROC) e a máxima distância do KS (Max_KS) e para medir a qualidade das regras foram utilizadas a cobertura, confiança e o lift. Os resultados mostraram que os principais fatores que influenciam a boa qualidade das escolas estão ligados a situação econômica e financeira, seja de maneira direta (renda familiar) ou indiretamente, ou em aspectos culturais (nível de educação da mãe ou do pai) da família.

Após a análise dos trabalhos relacionados, pode-se notar que existem diversos trabalhos atualmente que aplicam tanto o KDD quanto a mineração de dados sobre dados de estudantes, os quais se diferem quanto ao tipo de ensino abordado no estudo, ou seja, ensino público ou privado, quanto à esfera política, podendo ser municipal, estadual ou federal, e por fim quanto à abrangência geográfica, se o estudo abordou dados no nível nacional, por região, por estado ou tendo como foco uma dada instituição de ensino.

Já de acordo com os critérios que foram definidos para análise, eles também se diferem quanto ao objetivo e consequentemente quanto à tarefa de mineração de dados que foi definida no estudo para alcançar os objetivos do respectivo trabalho. O Quadro 5 mostra um comparativo dos trabalhos relacionados que foram analisados de acordo com os critérios definidos no início deste capítulo.

Quadro 5– Comparativo dos Trabalhos Analisados

Descrição	Objetivo	Tipo de Tarefa	Aborda Dados socioeconômicos?	Quantidade de Algoritmos	Quantidade de Métricas
Santos (2019)	Identificação	Classificação	Sim	Um	Não utilizou
Santana Júnior (2018)	Predição	Classificação	Sim	Seis	Seis
Simon, Augusto; Cazella, Sílvio (2017)	Predição	Classificação	Sim	Um	Seis
Stearns et al. (2017)	Predição	Regressão	Sim	Dois	Três
Silva et al. (2014)	Identificação	Associação	Sim	Um	Não utilizou
Adeodato et. al (2014)	Predição	Regressão	Sim	Três	Duas

Fonte: Elaborado pelo Autor (2020)

4.2 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Como pôde ser visto, todos os trabalhos encontrados na literatura pesquisada que abordaram os dados socioeconômicos dos estudantes tiveram a mesma fonte de dados, ou seja, por meio do INEP, na realização do ENEM em diversos anos diferentes, sendo que poucos trabalhos, menos de 20, ou seja, algo em torno de 3,5% dos 549 trabalhos encontrados pela string de busca realizaram uma análise comparativa de algoritmos aplicados sobre esses dados. Outro fato relevante é que os dados socioeconômicos usualmente foram agrupados ou combinados com outros dados, como os desempenhos de notas ou do censo escolar.

Nesse contexto, um dos estudos mais recentes sobre aplicação do KDD e da mineração de dados relacionados a dados socioeconômicos de alunos foi o de Silva (2019), no qual foi utilizado apenas um algoritmo, o de Redes Bayesianas, na tarefa de classificação, já que seu foco não era comparar o desempenho de algoritmos. Os dados socioeconômicos não foram os únicos a serem utilizados na pesquisa, sendo os mesmos combinados com dados do censo escolar relacionados à infraestrutura das escolas, ou seja, os dados socioeconômicos por si só também não eram o foco único do estudo.

A pesquisa encontrada que teve como um dos seus principais focos realizar uma análise comparativa de algoritmos foi a de Santana Júnior (2018), na qual foram utilizados seis algoritmos da tarefa de classificação, sendo três de caixa-preta

e três de caixa-branca, tendo sido aplicadas diversas métricas para avaliar seus desempenhos. Os algoritmos de caixa-preta tiveram o melhor desempenho, porém os algoritmos de caixa-branca são melhores em expressar o conhecimento de modo que um especialista no domínio da aplicação possa compreender, sendo considerado o melhor algoritmo o JRIP. Apesar desse estudo ser o mais completo e o que mais se aproxima aos objetivos deste trabalho, ele não foca somente nos dados socioeconômicos e acaba combinando esses dados com os de desempenho das notas dos estudantes.

Na pesquisa de Simon e Cazella (2017) foi utilizado apenas um algoritmo, o J48, da tarefa de classificação e, mesmo sendo aplicadas diversas métricas de avaliação do desempenho do algoritmo, não foi possível realizar uma análise comparativa com outros algoritmos já que este não era o foco da pesquisa. Por outro lado, sobre os dados socioeconômicos, não foram descritos quais os aspectos utilizados como parâmetros, sendo que nos experimentos com a aplicação do algoritmo esses dados foram sintetizados em uma única variável denominada nível socioeconômico.

Já no estudo realizado por Stearns et al. (2017) foi realizada uma análise comparativa de duas técnicas de *boosting* da tarefa de regressão para agregar árvores de decisão, sendo aplicadas diversas métricas para avaliar seus desempenhos. Os autores descreveram quais dados socioeconômicos foram utilizados e como resultado produziram um ranking dos dados mais relevantes considerando a quantidade de vezes que um dado foi utilizado para um nó de decisão da árvore.

Outra pesquisa que também utiliza somente um algoritmo foi a de Silva et. al. (2014). Nela foi aplicado o algoritmo Apriori da tarefa de associação, sendo assim, também não foi foco desse estudo realizar comparações de desempenho de algoritmos. No entanto, em relação aos dados socioeconômicos, os autores descreveram não só quais dados foram utilizados, como também quais desses dados foram identificados como influenciadores no desempenho dos alunos no ENEM.

Dentre os trabalhos analisados, o que mais utilizou técnicas diferentes na sua abordagem foi o de Adeodato et. al (2014). Porém, cada uma delas foi utilizada para um determinado contexto e para obter um objetivo diferente. Portanto, esse foi outro estudo que não teve como um de seus objetivos comparações de algoritmos. A

abordagem sobre os dados socioeconômicos, assim como na pesquisa de Silva (2019), foi a de agrupar esses dados com os do censo escolar. Com isso não fica claro quais dados socioeconômicos foram utilizados e nos resultados somente nas regras geradas pela técnica de indução de regras é que alguns desses dados aparecem.

O diferencial desta pesquisa é a utilização de uma quantidade e variedade maior de algoritmos a serem comparados da tarefa de classificação, tendo em vista que uma de suas abordagens é comparativa. Também, uma quantidade e variedade maior de métricas de avaliação será aplicada sobre esses algoritmos. Além disso, a aplicação desses algoritmos tem como foco único dados socioeconômicos dos alunos, deixando de lado dados de desempenho escolar, haja vista que os dados desta pesquisa serão extraídos do processo seletivo da assistência estudantil.

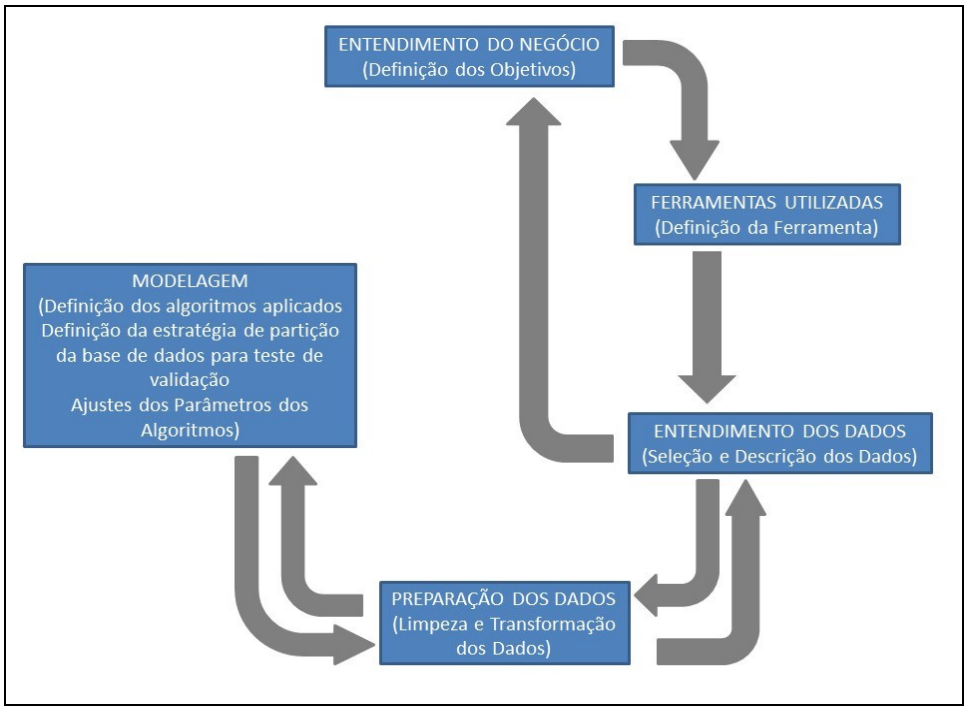
O próximo capítulo descreve a abordagem especifica de Mineração de Dados desenvolvida nesta dissertação, mostrando a sua divisão em etapas e quais temas cada uma dessas etapas aborda.

5 ABORDAGEM PROPOSTA: STUDENT ASSISTANCE MINING (SAM)

Neste capítulo é apresentado o desenvolvimento de uma abordagem específica com o objetivo de aplicar mineração de dados sobre bases de dados socioeconômicos de alunos dos Institutos Federais (IF). Para este estudo, serão considerados os dados do Instituto Federal do Amazonas, mais especificamente o Campus Manaus Zona Leste. Seu objetivo visa apoiar tanto a tomada de decisão na análise socioeconômica do programa de assistência estudantil quanto gerar informações úteis para a alta gestão do campus. Esta abordagem deve ser eficaz, eficiente e especifica de acordo com o cenário encontrado e servir de guia desde a seleção dos dados dos beneficiários até a análise dos resultados.

A metodologia CRISP-DM que está sendo utilizada neste trabalho permite esta personalização, por meio do desenvolvimento dos modelos previstos no seu escopo. As definições feitas em cada etapa da metodologia vão variar de acordo com o cenário, sendo assim, esta personalização pode ser entendida como uma abordagem, ou seja, a aplicação e suas possíveis adaptações de uma metodologia para um determinado caso específico. A Figura 15 mostra as etapas da metodologia CRISP-DM adotadas pela abordagem SAM proposta neste trabalho.

Figura 15– Etapas da Abordagem SAM



Fonte: Elaborada pelo Autor (2020)

Nesse contexto serão utilizados os conceitos gerais e a forma estruturada e sistêmica da metodologia CRISP-DM para guiar o desenvolvimento da abordagem SAM. Tal desenvolvimento leva em consideração a pesquisa realizada por Piatesky (2014), descrita na Seção 3.8, a qual comprovou que o CRISP-DM é a metodologia mais utilizada em projetos de DM. Além disso, o CRISP-DM pode ser considerado como o modelo genérico que servirá de base para a construção dessa abordagem especifica.

Em suma, devido à natureza e aos objetivos desta pesquisa fazerem uso de algoritmos de mineração de dados, visando determinar qual é o mais eficiente e eficaz na predição de alunos aptos a receber a assistência estudantil, foi considerada a aplicação do KDD como o processo global neste estudo, utilizando a metodologia CRISP-DM para alcançar os seus objetivos. Nas seções seguintes do presente capítulo será detalhada cada etapa do CRISP-DM e quais métodos e técnicas foram utilizados em cada uma delas, desde a etapa de entendimento do negócio até a etapa de modelagem.

5.1 ENTENDIMENTO DO NEGÓCIO

A primeira etapa do CRISP-DM visa obter o entendimento sobre os requisitos do projeto sob a perspectiva da área do negócio e transformar esse entendimento em uma definição de um problema de mineração de dados para a área do negócio abordada (WIRTH; HIPP, 2000). Assim sendo, nesta seção será descrito em detalhes o Programa Nacional de Assistência Estudantil – PNAES, quando e com que objetivo ele foi criado, bem como os objetivos do projeto de mineração de dados.

5.1.1 Plano Nacional de Assistência Estudantil – PNAES

O Fórum Nacional de Pró-Reitores de Assuntos Comunitários e Estudantes (FONAPRECE) elaborou uma proposta de Plano de Assistência Estudantil, sendo apresentado à Associação Nacional de Dirigentes das Instituições Federais de Ensino Superior (ANDIFES) em julho de 2007 e encaminhada ao Ministério da Educação (MEC). A Portaria Normativa nº 39, de 12 de dezembro de 2007 instituiu em âmbito nacional o Programa Nacional de Assistência Estudantil – PNAES,

visando atender os estudantes de cursos de graduação presenciais das Instituições Federais de Ensino Superior (IFES), o qual começou a ser implementado a partir do ano 2008 (BRASIL, 2010).

Contudo, está portaria foi alterada em 2010 e o PNAES foi regulamentado por meio do Decreto nº 7.234, de 19 de julho de 2010, sendo considerado um marco histórico para a política de assistência estudantil, por definir suas áreas de ação e ser referencial para os programas e projetos realizados nas IFES do Brasil, além de ser um instrumento jurídico com mais força e que permite uma estabilidade maior ao programa. Nesta nova regulamentação foram definidos os objetivos dos PNAES:

> I – Democratizar as condições de permanência dos jovens na educação superior pública federal;
> II – Minimizar os efeitos das desigualdades sociais e regionais na permanência e conclusão da educação superior;
> III – Reduzir as taxas de retenção e evasão; e
> IV – Contribuir para a promoção da inclusão social pela educação (BRASIL, 2010).

Além disso, são definidas as áreas onde deverão ser desenvolvidas as ações de assistências estudantil do PNAES:

> I - moradia estudantil; II - alimentação; III - transporte; IV - atenção à saúde; V - inclusão digital; VI - cultura; VII - esporte; VIII - creche; IX - apoio pedagógico; e X - acesso, participação e aprendizagem de estudantes com deficiência, transtornos globais do desenvolvimento e altas habilidades e superdotação (BRASIL, 2010).

Essas ações de assistência estudantil têm como objetivo viabilizar a igualdade de oportunidades para os alunos, além de contribuir para a melhoria do desempenho acadêmico dos mesmos e agir, de forma preventiva, nas situações de retenção e evasão que possam ocorrer devido à insuficiência de suas condições financeiras (BRASIL, 2010).

O referido decreto também estabeleceu em seu Art. 5º o público-alvo dessas ações, no qual é descrito que o PNAES deve atender, prioritariamente, estudantes oriundos da rede pública de educação básica ou com renda per capita de até um salário mínimo e meio, sem prejuízo de demais requisitos definidos pelas IFES (BRASIL, 2010). Esse critério de renda é mais restritivo que o anterior da Portaria nº 39, porém há possibilidade de acrescentar outros critérios por parte das instituições.

Desse modo o PNAES, apesar de ser um programa em âmbito federal, sua execução é descentralizada, com cada IFES possuindo autonomia de gestão para aplicar os recursos financeiros disponibilizados de acordo com suas necessidades e especificidades locais. Além disso, o decreto define que caberá às IFES definir os

critérios e a metodologia para selecionar os alunos que serão beneficiados, podendo considerar além da situação de vulnerabilidade socioeconômica, outros critérios condicionantes que podem ser adicionados, tais como: frequência e rendimento escolar. Essa autonomia permitiu que os Instituto Federais ampliassem o PNAES para alunos dos cursos integrados e técnicos.

Por fim, com esse decreto os Institutos Federais de Educação, Ciência e Tecnologia passaram a fazer parte do PNAES e a receberem recursos orçamentários para a assistência estudantil. Entretanto, é definido que as IFES devem adotar mecanismos de monitoramento e avaliação do PNAES, visando obter informações para auxiliar a tomada de decisão, bem como identificar e corrigir problemas e possíveis falhas de alinhamento estratégico. Assim sendo, elas devem planejar e executar análises, tendo como base critérios de eficiência, eficácia e efetividade, com o objetivo de aperfeiçoar a política da assistência estudantil e consequentemente melhorar a qualidade dos gastos públicos (ALMEIDA, 2018).

5.1.2 Definição dos Objetivos do Projeto de Mineração de Dados

O objetivo principal deste projeto de mineração é classificar os alunos participantes do processo de seleção do PNAES do Campus Manaus Zona Leste do IFAM em duas classes (APTO, NÃO APTO), a partir dos dados socioeconômicos preenchidos no questionário eletrônico da primeira fase do processo de seleção. A partir dessa classificação, o modelo produzido poderá predizer se os futuros alunos estarão aptos ou não a receber a assistência estudantil do PNAES, visando tornar mais ágil o processo de análise socioeconômica dos estudantes.

Além disso, existem outros objetivos atrelados ao principal, como a identificação e confirmação dos dados mais relevantes para análise socioeconômica definidos pelos especialistas do negócio, em especifico nesta pesquisa, as assistentes sociais do campus e a comparação dos algoritmos de mineração de dados aplicados, avaliando o desempenho de cada um deles na tarefa de classificação dos alunos participantes do PNAES e identificando qual algoritmo possui o melhor desempenho e eficácia.

Assim sendo, o escopo definido desta pesquisa consiste na predição de alunos aptos ou não a serem beneficiados pela assistência estudantil, os quais participaram do processo de seleção do Programa Nacional de Assistência

Estudantil (PNAES) do Campus Manaus Zona Leste (CMZL) do Instituto Federal de Educação, Ciência e Tecnologia do Amazonas (IFAM).

Assim, para alcançar estes objetivos serão aplicados algoritmos de mineração de dados visando identificar qual deles possui o melhor desempenho e eficácia com o objetivo de tornar mais ágil o processo de análise socioeconômica no processo de seleção do PNAES.

5.2 FERRAMENTAS UTILIZADAS

Atualmente existem dezenas de ferramentas de mineração de dados disponíveis no mercado, desde aquelas fornecidas por grandes empresas de software que são proprietárias como Oracle Data Minin[2], IBM SPSS Modeler[3] e o Microsoft Analysis Services[4], quanto as ferramentas que possuem distribuição gratuita como o KNIME[5], Orange Canvas[6], R[7], TANAGRA[8], WEKA[9] e o RapidMiner[10], sendo que este último também possui versões proprietárias.

Devido a essa quantidade de ferramentas existentes, pode-se encontrar na literatura estudos comparativos delas, principalmente das que apresentam licença do tipo open source ou do tipo freeware, como por exemplo os estudos realizados por Wahbeh et al. (2011), Gomes (2014), Boscarioli et al. (2014) e Viterbo et al. (2016). Esses estudos comparativos deixam claro que não existe uma ferramenta que seja a melhor entre todas, já que mostram que todas as ferramentas possuem vantagens e desvantagens na relação de umas com as outras. De acordo com Gomes (2014), nas conclusões dos seus estudos ela afirma que a definição da ferramenta depende dos objetivos do projeto de mineração de dados e dos conhecimentos técnicos do seu responsável.

A ferramenta escolhida para ser utilizada neste trabalho foi o WEKA (*Waikato Environment for Knowledge Analysis*), devido a algumas de suas características, tais como: a facilidade na aquisição, haja vista, que este software está disponível para download na página do desenvolvedor sem custos de operação; possuir um número

---

2 https://www.oracle.com/database/technologies/advanced-analytics/odm.html
3 https://www.ibm.com/products/spss-modeler
4 https://docs.microsoft.com/pt-br/analysis-services/?view=asallproducts-allversions
5 https://www.knime.com/
6 https://orange.biolab.si/
7 https://www.r-project.org/
8 http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
9 https://www.cs.waikato.ac.nz/ml/weka/
10 https://rapidminer.com/

considerável de algoritmos de mineração de dados para utilização; a possibilidade de alteração dos parâmetros de execução dos algoritmos; disponibilidade de recursos estatísticos para comparar os resultados e desempenho entre os algoritmos testados; e possuir uma ampla comunidade de usuários e ser extensível por meio de uma API Java.

5.2.1 Weka

O software WEKA (*Waikato Environment for Knowledge Analysis*) é uma ferramenta de código aberto e interface amigável, sendo desenvolvido na linguagem de programação JAVA[11] que facilita a sua portabilidade, ou seja, pode ser executado em diversas plataformas como Windows[12], Linux[13] e Mac Os[14]. Por isso, tem sido bastante utilizado no meio acadêmico e em pesquisas na área de mineração de dados, tornando-se um dos mais populares (WITTEN et al., 2011; HALL et al., 2009).

O WEKA foi desenvolvido na Universidade Waikato, na Nova Zelândia, no ano de 1999 sendo a sua licença *General Public License* (GPL), o que significa que é um software de distribuição e difusão livre. Essa ferramenta é formada por um conjunto de algoritmos, desde os mais clássicos aos algoritmos mais atuais, os quais implementam várias técnicas para resolução de tarefas como: pré-processamento, classificação, regras de associação, regressão e agrupamento (WITTEN et al., 2011; HALL et al., 2009).

O WEKA ainda possibilita a inclusão de novos algoritmos, desde que atendam aos requisitos descritos em sua documentação. Segundo Witten et al. (2011), o WEKA fornece suporte para todo o processo de mineração de dados, desde a preparação dos dados de entrada até a avaliação dos sistemas de aprendizado de maneira estatística, contando ainda com uma grande variedade de ferramentas de pré-processamento. Além disso, ele possibilita ser utilizado de diversas formas, em função do mesmo possuir cinco diferentes ambientes implementados que podem ser acessados por meio da sua tela inicial conforme mostra a Figura 16.

11 https://www.oracle.com/java/
12 https://www.microsoft.com/en-us/windows/
13 https://www.linux.org/
14 https://www.apple.com/macos/what-is/

Figura 16– Tela Inicial da Ferramenta Weka

Fonte: https://waikato.github.io/weka-wiki/downloading_weka/ (2020)

De acordo com Witten et al. (2011) e Hall et al. (2009), esses cinco diferentes ambientes são descritos da seguinte forma:

- *Explorer* - Proporciona um ambiente gráfico intuitivo de manipulação de dados para a utilização de diversos algoritmos. Trata-se da interface mais simples de usar, conduzindo o utilizador através de menus e formulários, orientando-o de forma que as aplicações sejam utilizadas na ordem apropriada, além de fornecer sugestões úteis que aparecem na tela auxiliando o usuário;
- *Experimenter* - Permite testar técnicas diferentes de classificação ou regressão, além de valores de parâmetros de modo a compará-las verificando qual funciona melhor para o problema em questão. Apesar de estas operações serem igualmente possíveis tanto no Explorer como no KnowledgeFlow. No Experimenter, no entanto, é possível automatizar esse processo tornando mais fácil a execução dos testes de comparações;
- *KnowledgeFlow* - Permite o desenvolvimento de projetos de mineração de dados num ambiente gráfico com fluxos de informação;
- *Workbench* - Foi implementado nas versões mais atuais do WEKA, e é um ambiente que combina todos os outros em uma única interface; e
- *Simple CLI* - Proporciona uma interface que permite a execução direta de comandos do WEKA. Embora disponibilize todas as funcionalidades, requer um elevado grau de conhecimento dos comandos que poderão ser

utilizados.

Existem dois formatos de arquivos que podem ser utilizados pelo WEKA: o CSV (Commom Separated Values) e o ARFF (Attribute Relation File Format). O formato ARFF é um formato próprio do WEKA, o qual contém informações como a definição do domínio dos atributos, um cabeçalho no qual são declarados os atributos e as instâncias que representam os dados que serão explorados (HALL et al., 2009). É importante mencionar que a própria ferramenta também permite a conversão de arquivos CSV para o formato ARFF. Essa conversão facilita a utilização, pois a ferramenta pode rapidamente ler um único arquivo ARFF no qual todas as instâncias de dados podem estar inseridas e construir modelos preditivos a partir da aplicação de algoritmos que já estão implementados no WEKA.

5.3 ENTENDIMENTO DOS DADOS

A segunda etapa do CRISP-DM tem como objetivo a seleção, a descrição e o entendimento dos dados (SHARMA et al., 2012). Assim sendo, nesta seção será descrita a base de dados de estudantes participantes do processo seletivo do PNAES do Campus Manaus Zona Leste do IFAM, sobre a qual serão aplicados os algoritmos de mineração de dados, desde a forma como a base foi obtida até a seleção dos dados que podem ser os mais relevantes para serem utilizados pelos algoritmos.

5.3.1 Base de Dados

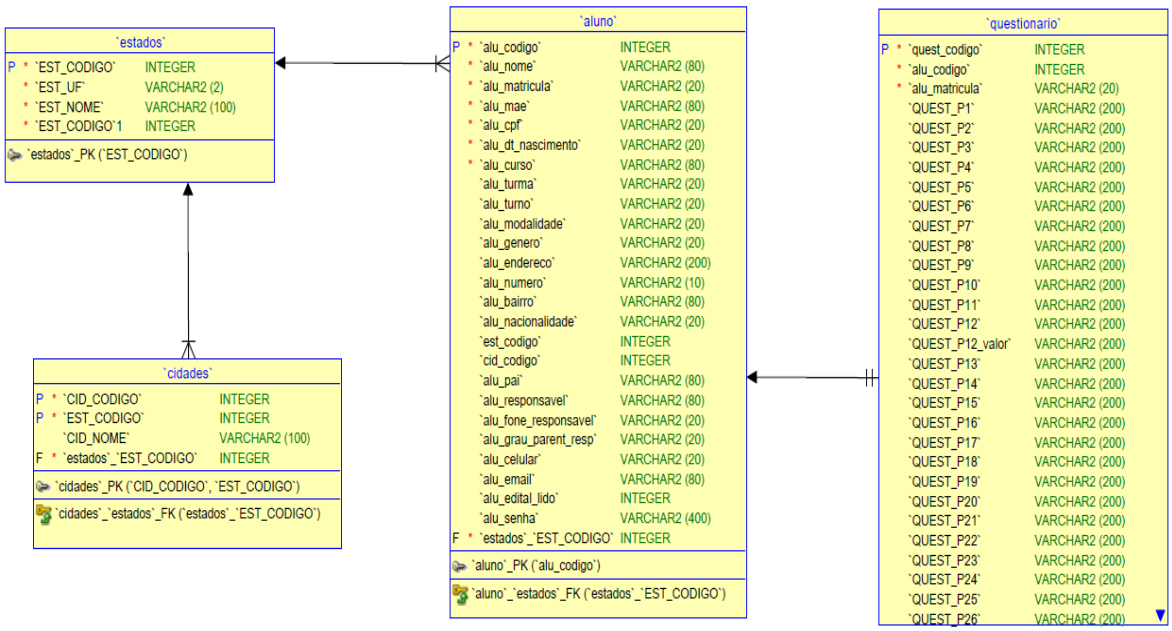
O PNAES foi implementado no Campus Manaus Zona Leste do IFAM a partir do segundo semestre de 2011, sendo inicialmente realizado todo de forma manual, ou seja, os alunos preenchiam formulários a mão e depois as assistentes sociais tinham que ler todas as respostas de todos os questionários. Já nos anos de 2012 e 2013 foi utilizado o software Microsoft Access para automatizar o processo, e entre os anos de 2014 e 2015 foi utilizado o software online *Survey Monkey* também para criação de um questionário online e para armazenamento dos dados.

Somente no ano 2016, após a ampliação da equipe de TI do Campus, foi possível desenvolver uma ferramenta própria e institucional, sendo a mesma utilizada até os dias atuais. Devido às constantes mudanças de software, alterações

de perguntas e respostas dos questionários e mudanças de gestão que podem ter ocasionado perda de informações, levando em consideração que o Campus possui uma ferramenta única e estável sendo utilizada desde 2016, a base de dados selecionada será obtida por essa ferramenta com dados armazenados de 2016 até o segundo semestre de 2019.

O questionário eletrônico aplicado no processo de seleção do PNAES do Campus Zona Leste possui mais de cem (100) perguntas, sendo algumas obrigatórias e outras condicionadas a respostas de outras perguntas. De 2016 até o segundo semestre de 2019 um total de três mil seiscentos e trinta e um (3.631) alunos preencheu este questionário para participar do processo de seleção do PNAES. As perguntas do questionário eletrônico são mostradas de forma organizada no APÊNDICE A e o exemplo de um questionário preenchido e gerado em arquivo com a extensão PDF pelo sistema do questionário é mostrado no ANEXO A. O questionário eletrônico do PNAES é uma ferramenta desenvolvida utilizando a linguagem de programação PHP[15] e os seus dados são armazenados em um SGBD MySQL[16], o esquema do modelo relacional do banco de dados dessa aplicação é mostrado na Figura 17.

Figura 17– Modelo Relacional da Aplicação do Questionário Socioeconômico

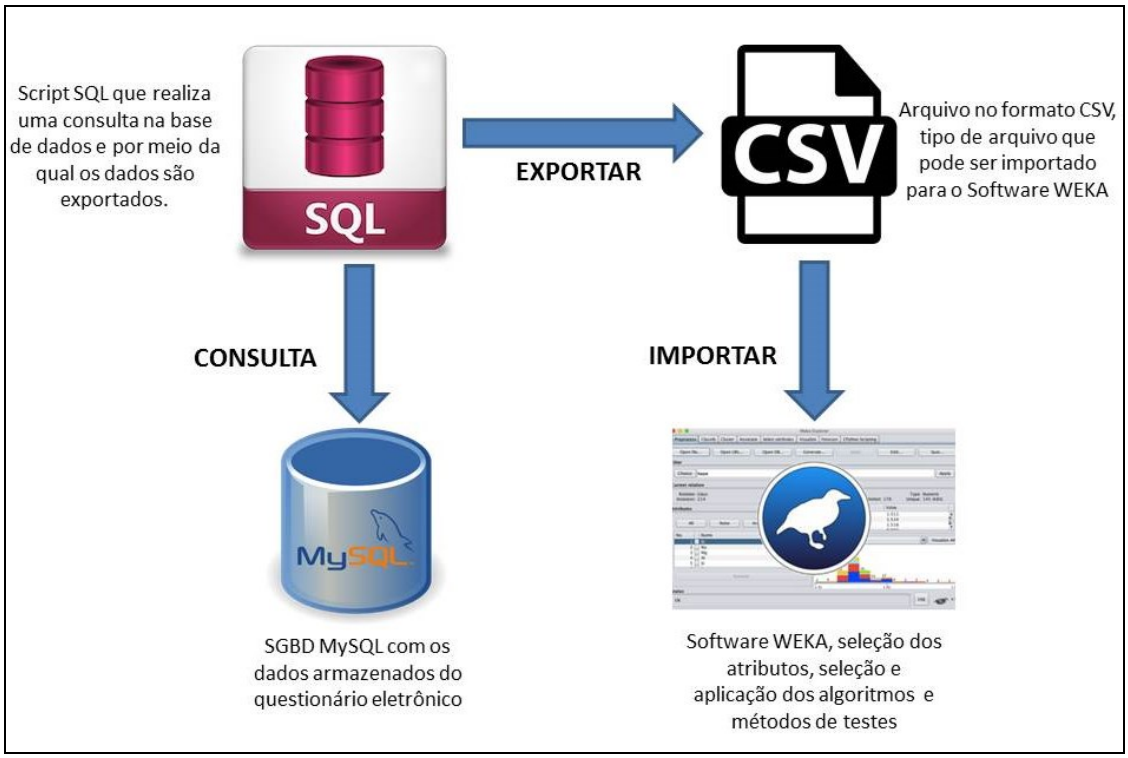


Fonte: Elaborada pelo Autor (2020)

15 https://www.php.net/
16 https://www.mysql.com/

Ao término do prazo estabelecido no edital do PNAES para o preenchimento do questionário eletrônico pelos alunos, é realizada uma consulta SQL sobre a base de dados armazenados no SGBD, cujos dados que resultam da realização dessa consulta são exportados para um arquivo do tipo planilha eletrônica do Microsoft Excel[17] e enviado para o setor de assistência social do Campus. Porém, para a realização desta pesquisa, os dados armazenados no SGBD foram exportados para um arquivo com extensão CSV, para que esses dados pudessem ser importados para o software WEKA, conforme mostra a Figura 18.

Figura 18– Importação dos Dados para o Software WEKA



Fonte: Elaborada pelo Autor (2020)

5.3.2 Seleção de Dados

O PNAES é um programa voltado para a assistência estudantil, portanto, esta é a área de negócio em questão do projeto de mineração de dados desta pesquisa. Assim sendo, as assistentes sociais do Campus Manaus Zona Leste são as especialistas dessa área de negócios. São as assistentes sociais que possuem todo o conhecimento sobre a legislação do PNAES, planejam, organizam e executam o

17 https://www.microsoft.com/pt-br/microsoft-365/excel

processo de seleção dos estudantes para o recebimento da assistência estudantil, atuando e atendendo diretamente pais e alunos.

Nesse contexto, são essas especialistas do negócio que definem quais são os dados mais relevantes para análise socioeconômica dos estudantes, para definir se o estudante está apto ou não a prosseguir no processo de seleção. Vale ressaltar, conforme descrito na Seção 5.1.1, o Decreto do PNAES dá essa autonomia para as IFES definirem os critérios e metodologia de seleção dos beneficiados, porém, ele deixa bem claro que o principal critério é a vulnerabilidade socioeconômica, ou seja, questões financeiras que impactem na renda familiar do estudante.

Assim sendo, foram definidos pelas especialistas 11 atributos a serem selecionados, sendo 10 perguntas do questionário socioeconômico que são de fato relevantes para a análise socioeconômica dos alunos mais o campo GÊNERO. Essas 10 perguntas foram selecionadas porque estão diretamente ligadas à renda ou ao que de certa forma influencia na situação financeira, não só do estudante, mas de toda a sua família. O campo GÊNERO que é dos dados pessoais do aluno também foi selecionado, pois ele será importante para gerar o conhecimento útil relacionado a evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica, ou seja, descobrir se existem mais alunos do gênero feminino, masculino ou LGBT no Campus Manaus Zona Leste nesta situação de vulnerabilidade. As perguntas selecionadas são mostradas no Quadro 6.

Quadro 6– Perguntas Selecionadas para Análise Socioeconômica

Campo no BD	**Descrição do Campo (Pergunta e suas Respectivas Respostas do Questionário Socioeconômico)**	**Tipo de Dado**
ALU_GENERO	GÊNERO: Masculino Feminino LGBT	STRING
QUEST_P05	P05 – VOCÊ TEM FILHOS? Não tenho; Sim, tenho um (1); Sim, tenho dois (2); Sim, tenho Três (3); Sim, mais que Três;	STRING
QUEST_P11	P11 – QUAL SUA SITUAÇÃO ATUAL DE MORADIA? Moro Sozinho(a); Com o pai, a mãe e irmãos; Com o cônjuge; Em casa de familiares; Em casa de amigos;	STRING

	Pensão/Hotel/Pensionato; República; Moradia mantida pela família; Com Filhos; Só com a mãe; Só com o pai; Com irmãos; Casa do estudante paga pelo poder público; Internato IFAM - CAMPUS MANAUS ZONA LESTE; Outras moradias coletivas religiosa, pública, entre outros tipos;	
QUEST_P12	P12 – TIPO DE MORADIA DA SUA FAMÍLIA? Própria; Própria financiada; Alugada; Cedida; Herdada;	STRING
QUEST_P13	P13 – QUAL O PRINCIPAL MEIO DE TRANSPORTE QUE VOCÊ UTILIZA PARA CHEGAR AO IFAM - ZONA LESTE? Transporte próprio (carro ou moto); A pé ou bicicleta; Transporte coletivo; Carona; Mototáxi;	STRING
QUEST_P15	P15 – VOCÊ TRABALHA? Sim, com carteira assinada; Sim, sem carteira assinada; Não;	STRING
QUEST_P19	P19 – QUAL A RENDA MENSAL DO SEU GRUPO FAMILIAR? SEM RENDA; Até MEIO salário mínimo - (522,50); Até UM salário mínimo - (1.045,00); Até UM E MEIO salário mínimo - (1.567,50); Até DOIS salários mínimos - (2.090,00); Até DOIS E MEIO salários mínimos - (2.612,50); Até TRÊS salários mínimos - (3.135,00); Até TRÊS E MEIO salários mínimos - (3.657,50); Até QUATRO salários mínimos - (4.180,00); Até QUATRO E MEIO salários mínimos - (4.702,50); Até CINCO salários mínimos - (5.225,00); Até CINCO E MEIO salários mínimos - (5.747,50); Até SEIS salários mínimos - (6.270,00); Mais que SEIS salários mínimos - (6.792,50);	STRING
QUEST_P20	P20 – QUANTAS PESSOAS, INCLUINDO VOCÊ, VIVEM DA RENDA MENSAL DO SEU GRUPO FAMILIAR INFORMADA ACIMA? Uma (1); Duas (2); Três (3); Quatro (4); Cinco (5); Seis (6); Sete (7); Oito (8); Nove (9); Dez (10) ou mais	STRING
QUEST_P21	P21 – QUAL A SUA PARTICIPAÇÃO NA VIDA ECONÔMICA DO SEU GRUPO FAMILIAR? Sou sustentado pela família ou por outras pessoas;	STRING

	Recebo ajuda financeira da família ou de outras pessoas; Sou responsável apenas pelo meu próprio sustento; Sou responsável pelo meu sustento e contribuo para o sustento da família; Sou responsável principal pelo sustento de minha família;	
QUEST_P22	P22 – QUAL É A ESCOLARIDADE DE SEU PAI? OU DA PESSOA QUE AO CRIOU COMO PAI: Não teve pai ou pessoa que exerceu tal papel na criação; Sem instrução, não alfabetizado; Sem instrução, sabe ler e escrever; Ensino fundamental 1a a 4a – INCOMPLETO; Ensino fundamental 1a a 4a – COMPLETO; Ensino fundamental 5a a 8a – INCOMPLETO; Ensino fundamental 5a a 8a – COMPLETO; Ensino Médio antigo 2o grau – INCOMPLETO; Ensino Médio antigo 2o grau – COMPLETO; Ensino Superior – INCOMPLETO; Ensino Superior – COMPLETO; Especialização, Mestrado ou Doutorado;	STRING
QUEST_P23	P23 – QUAL É A ESCOLARIDADE DE SEU MÃE? OU DA PESSOA QUE AO CRIOU COMO MÃE: Não teve mãe ou pessoa que exerceu tal papel na criação; Sem instrução, não alfabetizado; Sem instrução, sabe ler e escrever; Ensino fundamental 1a a 4a – INCOMPLETO; Ensino fundamental 1a a 4a – COMPLETO; Ensino fundamental 5a a 8a – INCOMPLETO; Ensino fundamental 5a a 8a – COMPLETO; Ensino Médio antigo 2o grau – INCOMPLETO; Ensino Médio antigo 2o grau – COMPLETO; Ensino Superior – INCOMPLETO; Ensino Superior – COMPLETO; Especialização, Mestrado ou Doutorado;	STRING

Fonte: Elaborado pelo Autor (2020)

Conforme descrito na Seção 5.3.1, uma consulta SQL é realizada sobre a base dados da aplicação do questionário socioeconômico por meio da qual os dados são selecionados para gerar o arquivo CSV que é importado para a ferramenta WEKA. Assim sendo, o script dessa consulta SQL já é realizado selecionando somente os dados que foram previamente definidos pelas especialistas do negócio. Essa consulta SQL é mostrada na Figura 19.

Figura 19– Script SQL de Consulta dos Dados Selecionados

```
SELECT aluno.`alu_genero` AS GENERO,
questionario.`QUEST_P5` AS "TEM_FILHOS",
questionario.`QUEST_P11` AS "MORA_COM_QUEM",
questionario.`QUEST_P12` AS "TIPO_MORADIA",
questionario.`QUEST_P13` AS "MEIO_TRASPORTE",
questionario.`QUEST_P15` AS "ALUNO_TRABALHA",
questionario.`QUEST_P19` AS "RENDA_MENSAL_FAMILIAR",
questionario.`QUEST_P20` AS "QUANT_DEPENDE_RENDA",
questionario.`QUEST_P21` AS "PARTICIPA_RENDA",
questionario.`QUEST_P22` AS "ESCOLARIDADE_PAI",
questionario.`QUEST_P23` AS "ESCOLARIDADE_MAE"
FROM `aluno`,`questionario`
WHERE aluno.`alu_codigo` = questionario.`alu_codigo`
```

Fonte: Elaborada pelo Autor (2020)

5.4 PREPARAÇÃO DOS DADOS

A etapa de preparação dos dados da metodologia CRISP-DM também conhecida como pré-processamento dos dados é uma fase crucial no processo de descoberta de conhecimento em bases de dados (CASTRO; FERRARI, 2016). Ela consiste basicamente na limpeza, integração e transformação dos dados. Ainda nessa etapa, é possível realizar a redução tanto dos atributos quanto dos registros, bem como permite a definição de novos atributos, a partir dos já existentes, com o objetivo de aumentar a quantidade de variáveis de entrada para a aplicação dos algoritmos de mineração de dados (GOLDSCHMIDT et al., 2015; CASTRO; FERRARI, 2016).

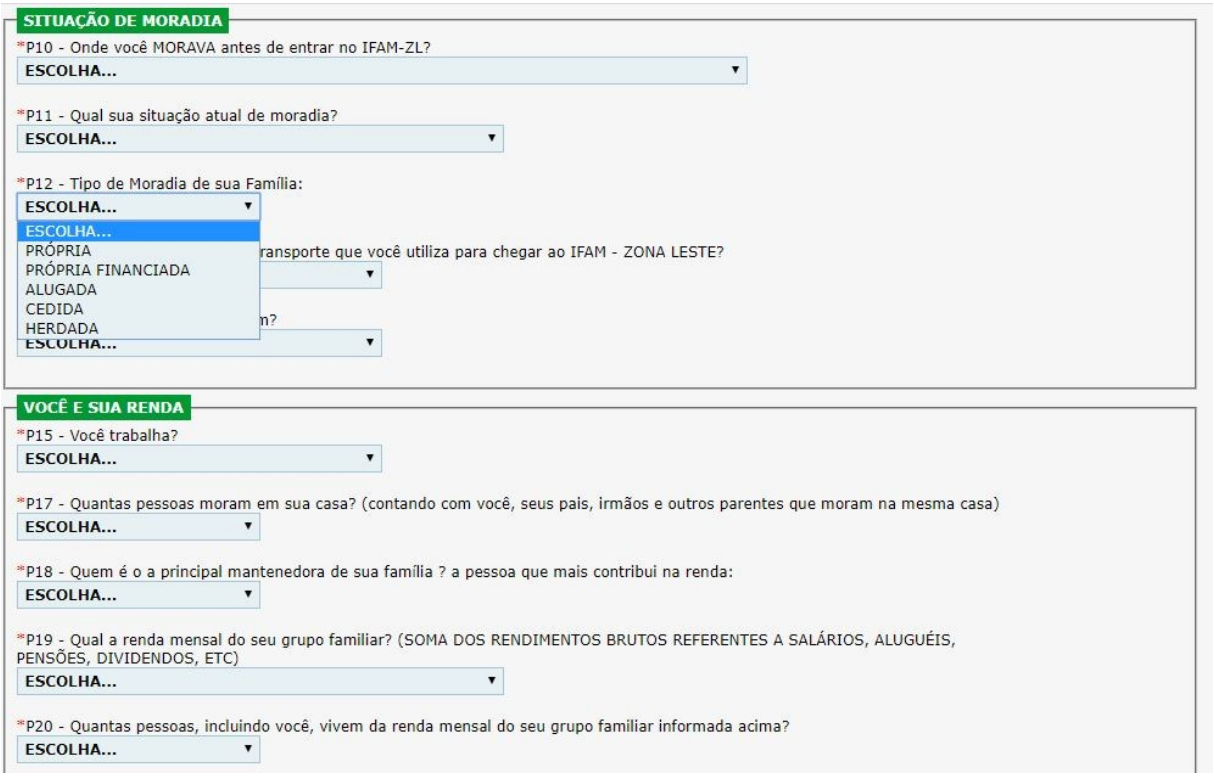
Nesta seção serão descritos como foram realizadas as tarefas da etapa da preparação dos dados: limpeza, integração, redução e transformação dos dados, que foram previamente selecionados da base de dados selecionada.

5.4.1 Limpeza e Integração dos Dados

Devido a algumas características da ferramenta desenvolvida para o preenchimento do questionário eletrônico com os dos dados socioeconômicos dos alunos, alguns problemas que poderiam ser encontrados como valores ausentes, valores ruidosos ou inconsistentes foram evitados.

A primeira característica do questionário eletrônico desenvolvido que evitou a ocorrência desses problemas foi que todas as respostas das perguntas eram campos obrigatórios, impossibilitando assim dados ausentes na base de dados. A segunda característica é que os campos de resposta do questionário eram quase todos do tipo “lista suspensa”, ou seja, já possuíam respostas pré-definidas e não permitiam respostas “abertas” com texto digitado ou mais de uma resposta para uma mesma pergunta, conforme mostra a Figura 20.

Figura 20– Tela do Sistema com os Campos do Questionário



Fonte: Ferramenta do Questionário Eletrônico do IFAM-CMZL (2020)

Assim sendo, como a base de dados não possuía dados ausentes, ruidosos ou inconsistentes, não foi necessária a realização da tarefa de Limpeza dos Dados. Da mesma forma, não houve a necessidade da realização da tarefa da Integração dos Dados, haja visto que toda a base de dados utilizada nesta pesquisa foi obtida de uma única fonte de dados, gerada por meio da ferramenta do questionário eletrônico. Conforme descrito na Seção 5.3.1, apesar de terem existido outras bases de dados anteriormente geradas por outras ferramentas, com o passar dos anos e a constante troca de ferramentas e de servidores responsáveis, algumas bases foram perdidas, sendo levada em consideração para esta pesquisa apenas a base de

dados da ferramenta do questionário eletrônico de 2016 até 2019.

5.4.2 Criação de Novos Atributos

O único atributo novo criado foi o atributo alvo (CLASSE), já que a base de dados não possuía esses dados. O atributo alvo foi criado a partir da listagem divulgada pelo Campus Manaus Zona Leste com o resultado da assistência estudantil contendo o nome e a situação, informando se o pedido foi deferido ou indeferido. Essa listagem é mostrada no ANEXO B.

O atributo foi gerado a partir da comparação do nome do aluno que já estava na base de dados com o nome do aluno que estava na listagem da assistência estudantil. Então, por meio da coluna RESULTADO da listagem foi definido o atributo alvo que foi denominado SITUACAO, no qual havia o valor contendo a informação AUXÍLIO CONCEDIDO que foi convertido em APTO ou o valor INDEFERIDO, convertido em NÃO APTO. Além disso, também foram levados em consideração os períodos que as listagens foram divulgadas por semestre e ano para realizar as comparações que geraram o atributo alvo. Todos esses procedimentos foram realizados por meio da ferramenta Microsoft Excel no arquivo CSV, antes do mesmo ser importado pela ferramenta WEKA.

Apesar de existirem outras ferramentas que automatizem essa tarefa, optou-se pela utilização do Microsoft Excel para que os especialistas da área de negócio e demais interessados na utilização da abordagem que não são da área de tecnologia, pudessem compreender de que forma o novo atributo foi criado, já que o Microsoft Excel é uma ferramenta de uso mais comum a todos.

Após a criação do atributo alvo com as classes APTO e NÃO APTO é possível verificar que a base de dados armazenada entre os anos de 2016 a 2019, contendo 3.631 instâncias de dados, está com as classes desbalanceadas. A base possui 2856 instâncias para classe APTO e somente 775 instâncias para a classe NÃO APTO, ou seja, a classe APTO possui 78,66% do total de instâncias e a classe NÃO APTO apenas 21,34%. Assim, a proporção de instâncias para a classe APTO é quase quatro vezes maior, portanto a base de dados está com as classe desbalanceadas.

5.4.3 Padronização de Atributos

A padronização dos dados é um processo realizado para resolver as diferenças de unidades e escalas dos dados, como a capitalização que é a diferença entre maiúsculas e minúsculas; os caracteres especiais, como acentuação; padronização de formatos, como dados do tipo data; e conversão de unidades, como dados com valores em moeda e unidades de peso e distância (CASTRO; FERRARI, 2016). Esse processo pode reduzir bastante os valores distintos dos atributos, ocasionando melhoras no desempenho dos algoritmos, além de evitar possíveis problemas que caracteres especiais podem ocasionar nos algoritmos. Todo esse processo retirou caracteres especiais como acentos, cedilhas, virgulas e outros símbolos e foi realizado por meio da ferramenta Microsoft Excel.

Assim como ocorreu na tarefa de criação de novos atributos, apesar de haver ferramentas que automatizem a padronização dos dados, optou-se por utilizar o Microsoft Excel para realizar esta tarefa para que os especialistas da área de negócio e demais interessados na utilização da abordagem que não são da área de tecnologia, pudessem compreender de que forma foi realizada a padronização.

O atributo RENDA_MENSAL_FAMILIAR armazena os valores baseados no salário mínimo vigente. Durante o período armazenado na base de dados, entre 2016 a 2019, o salário mínimo teve quatro valores diferentes, conforme mostra o Quadro 7.

Quadro 7– Valores do Salário Mínimo por Períodos

Períodos	**Valores do Salário Mínimo**
2016	R$ 880,00 (oitocentos e oitenta reais)
2017	R$ 937,00 (novecentos e trinta e sete reais)
2018	R$ 954,00 (novecentos e cinquenta e quatro reais)
2019	R$ 998,00 (novecentos e noventa e oito reais)

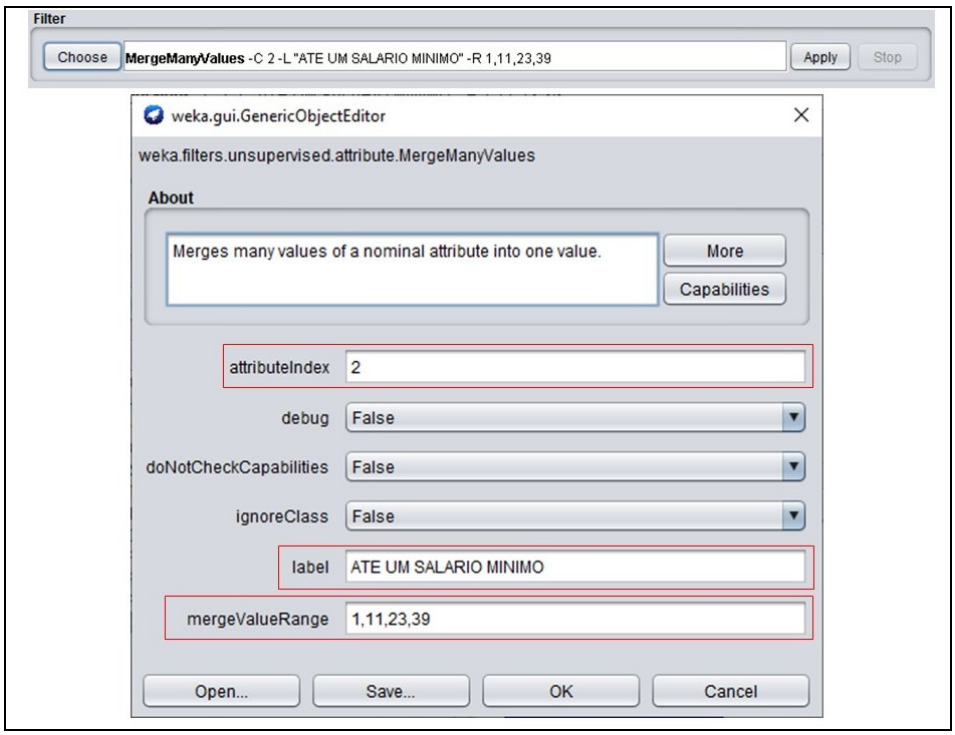
Fonte: Elaborado pelo Autor (2020)

Com isso, o atributo RENDA_MENSAL_FAMILIAR tinha 54 valores distintos diferentes, já que cada uma das 14 possíveis opções de respostas para o atributo ainda tinha 4 variações devido às mudanças dos valores do salário mínimo. Para evitar redundância de dados, os valores em dinheiro foram retirados, por exemplo, o

valor “ATÉ UM SALÁRIO MÍNIMO - (1.045,00)” ficou somente “ATE UM SALARIO MINIMO”.

Esse procedimento foi realizado por meio da aplicação do filtro não supervisionado de atributos da ferramenta WEKA, o *MergeManyValues*. A função deste filtro é realizar a mesclagem de vários valores de um atributo nominal em um único valor, o filtro possui três parâmetros que devem ser configurados: *attributeIndex*, *mergeValueRange* e *label*. O *attributeIndex* é o índice do atributo que vai ser feita a mesclagem dos valores, o *mergeValueRange* é o índice dos valores do atributo que serão mesclados e o *label* é o novo nome que o valor recebe após a mesclagem dos valores. O Filtro e seus parâmetros são mostrados na Figura 21.

Figura 21– Aplicação do Filtro MergeManyValues



Fonte: Elaborada pelo Autor (2020)

Vale ressaltar que o processo de discretização para converter valores numéricos ou contínuos em valores nominais ou categóricos não foi necessário ser realizado, haja visto que quase todos os dados da base de dados original são nominais, com a exceção do atributo QUANT_DEPEND_RENDA, o qual mesmo estando armazenado em forma nominal pode ser interpretado e convertido para valores ordinais, porém essa conversão pode ser realizado automaticamente pelo algoritmo que estiver sendo aplicado.

5.4.4 Seleção de Atributos e Balanceamento das Classes

Apesar de já ter havido uma redução dos dados realizada por meio da seleção dos dados feita pelos especialistas do negócio, neste caso as assistentes sociais, essa seleção inicial foi realizada apenas do ponto de vista do entendimento do negócio.

Visando analisar como a técnica de seleção de atributos iria influenciar no desempenho das técnicas de mineração de dados e diante das informações levantadas na Seção 3.6, devido a sua aceitação, generalidade, simplicidade e a rapidez em sua execução, nesta pesquisa foi definida a abordagem Filter para a seleção de atributos, sendo realizada por meio da aplicação do algoritmo CFS.

Com o objetivo de atender ao maior número possível de alunos, a assistência estudantil do IFAM Campus Manaus Zona Leste tende a ter uma baixa quantidade de indeferimentos de solicitações de bolsas dos auxílios, obviamente sempre se baseando no orçamento disponível para a assistência. Isso faz com que o número de alunos com pedidos deferidos para receber os auxílios seja muito maior do que os que tiveram os pedidos indeferidos, fazendo com que a base de dados fique desbalanceada.

Diante do exposto na Seção 3.7, com a justificativa da necessidade de realizar o balanceamento das classes e a descrição e as vantagens do algoritmo SMOTE, ele foi definido nesta pesquisa como o algoritmo para realizar o balanceamento das classes.

5.5 MODELAGEM

A modelagem é a quarta etapa do CRISP-DM. Nela, primeiramente são definidos os algoritmos de mineração de dados que serão aplicados sobre a base de dados já preparada. Segundo Kohavi et al. (1996), os resultados mostram que não existe uma única técnica ou algoritmo capaz de obter o melhor desempenho em relação aos outros algoritmos em todos os domínios de aplicações. Assim sendo, os algoritmos são selecionados sempre buscando alcançar os objetivos definidos durante a etapa do entendimento do negócio, ou seja, em função da aderência ao problema a ser tratado (WIRTH; HIPP, 2000).

Logo após definir os algoritmos a serem aplicados, é necessário criar um

projeto de testes visando avaliar a qualidade de um modelo, no qual a base de dados pode ser dividida em duas partes (dados de treinamento e dados de teste) ou em três partes (treinamento, validação e teste). Este particionamento pode ser realizado por meio de quatro diferentes métodos: *K-fold Cross Validation*, *Leave-One-Out*, *Holdout* e *Bootstrap*. E, por fim, são feitos ajustes dos parâmetros dos algoritmos para a realização dos testes preliminares a partir dos dados preparados.

Dessa forma, nesta pesquisa o modelo de mineração utilizado foi o preditivo, por meio da tarefa de classificação, utilizando sete diferentes técnicas: Árvore de Decisão, Regras de Classificação, Classificação Bayesiana, KNN, Regressão Logística, Support Vector Machine e Redes Neurais Artificiais, todas descritas na Seção 3.4. Essas técnicas foram escolhidas por serem as mais utilizadas na tarefa de classificação, a qual é a tarefa de mineração que possui mais aderência aos objetivos do entendimento do negócio, além de poderem ser utilizadas em diversos outros contextos de aplicação (WU et al., 2008; WITTEN et al., 2011).

Nesta seção serão descritos os algoritmos que foram aplicados, bem como o projeto de testes para validação dos modelos com o método de particionamento da base de dados utilizado e os parâmetros que foram ajustados.

5.5.1 Algoritmos Selecionados

Foram definidos doze algoritmos divididos entre as sete diferentes técnicas de mineração de dados, conforme mostra o Quadro 8. Essas escolhas foram baseadas em importantes trabalhos que possuíam uma abordagem comparativa de algoritmos de mineração de dados aplicados sobre dados de estudantes e que apontam tais algoritmos como os predominantes em pesquisas na área (BAKER; YACEF, 2009; PEÑA-AYALA et al., 2009; ROMERO;VENTURA, 2010; MANHÃES, 2011 e PEÑA-AYALA, 2014).

Quadro 8– Descrição dos Algoritmos Selecionados

Técnica de Data Mining	**Algoritmo da Ferramenta WEKA**	**Descrição do Algoritmo**
Árvore de Decisão	J48	Este algoritmo é uma implementação do algoritmo C4.5, o qual produz um classificador em forma de árvore de decisão e utiliza a abordagem *top-down*, na qual a

		árvore é construída do topo para a base, a partir da escolha do atributo que melhor separa as amostras. Os nós da árvore podem ser um nó folha, no qual todas os objetos pertencem à mesma classe, ou seja, a classe alvo, ou pode ser um nó de decisão (é o atributo que melhor separa as amostras). Para selecionar esse atributo que melhor divide as amostras, é utilizada a medida de razão de ganho (*Gain Ratio*), devido ao fato de que essa medida se mostrou superior ao ganho de informação (*info Gain*), gerando árvores mais precisas e menos complexas (QUINLAN, 1993). O algoritmo usa esse particionamento recursivamente, de modo que uma vez que o atributo apareceu em um nó, ele não é mais considerado nos seus dependentes. Este particionamento recursivo termina quando todos as amostras para um dado nó pertencem à mesma classe (CASTRO; FERRARI, 2016).
	Random Forest	O algoritmo *Random Forest* cria diversas árvores de decisão de maneira aleatória. Então essas árvores da floresta são combinadas em um único classificador final, o qual irá determinar a classe que uma determinada instância pertence (BREIMAN, 2001). O conjunto de amostras da base de dados é dividido de forma aleatória em diversos subconjuntos, e para cada subconjunto uma árvore é gerada e um classificador é identificado, o qual é utilizado em uma votação para eleger a classe. Cada árvore gerada a partir de um subconjunto vota, e aquela que obtiver o maior número de votos será a classe definida.
Regras de Classificação	PART	O algoritmo PART (*Partial Decision Trees*) se baseia na geração de uma árvore de decisão, e a partir dela extrai um conjunto de regras de classificação (FRANK; WITTEN,1998). Primeiramente ele extrai a primeira regra do caminho com a maior cobertura da árvore, na qual as instâncias cobertas por esta regra são removidas dos dados de treinamento. Então o processo é repetido para gerar a segunda regra, e assim, sucessivamente. Em suma, ele usa abordagem de dividir e conquistar e constrói uma árvore de decisão parcial para cada interação e percorre a “melhor” folha dentro de uma regra (DEVASENA, 2014).
	JRip	Um algoritmo que implementa o princípio de regras proporcionais é o *Repeated Incremental Pruning to Produce Error Reducition* (RIPPER), o qual utiliza a poda incremental repetida visando a redução de erro (COHEN, 1995). É um dos algoritmos mais utilizados na tarefa de classificação (WITTEN et al., 2011). Ele opera em dois estágios: o primeiro gera um conjunto de regras para serem comparadas e o segundo otimiza essas regras visando diminuir os erros de classificação, sendo esses passos repetidos inúmeras vezes.
	OneR	Abreviatura para *One Rule* ou uma regra, isto porque este algoritmo gera uma árvore de decisão de um único nível. Para isto, ele monta uma tabela com as regras mais frequentes e suas predições, na qual é selecionada a regra que apresentar menos erros na predição. Após isso, todas as predições são baseadas nesta regra.
Classificação Bayesiana	Naive Bayes	Este algoritmo é um classificador probabilístico baseado no teorema de Bayes, o qual parte do princípio de que dado um conjunto de dados de treinamento no qual cada

		amostra já possua previamente a sua classe determinada, é possível predizer em qual classe uma nova amostra será classificada (MITCHELL, 1997). Em suma, ele calcula a probabilidade que uma nova amostra tem de pertencer a uma das classes possíveis, ou seja, prediz a classe mais provável da nova amostra (JOHN; LANGLEY, 1995).
	BayesNet	É um algoritmo de aprendizagem baseado nas redes bayesianas. Para estimar as tabelas de probabilidade condicional da rede é utilizado o algoritmo de aprendizado denominado de árvore Naive Bayes aumentada. Esta rede bayesiana de aprendizagem, por sua vez, utiliza outros algoritmos de buscas variadas e métricas de qualidade (WITTEN et al., 2011).
Regressão Logística	SimpleLogistic	É um algoritmo que constrói modelos de regressão logística, ajustando-os utilizando o método *LogitBoost* com funções de regressão simples como base de aprendizado e determinando quantas interações devem ser executadas por meio da validação cruzada (*cross validation*), dando suporte à seleção automática de atributos (WITTEN et al., 2011).
K-NN	IBk	Algoritmo de aprendizado baseado em instâncias que implementam o classificador do K vizinhos mais próximos (KNN). O mesmo utiliza uma função de similaridade baseada na distância euclidiana dos atributos (AHA et al., 1991). Neste algoritmo o valor do parâmetro K pode ser determinado automaticamente por meio da validação cruzada (*cross validation*). As predições de mais de um vizinho podem ser ponderadas de acordo com a distância da instância de teste e duas fórmulas diferentes são implementadas para converter a distância em um peso, ou seja, pode-se atribuir um peso para cada vizinho de K (WITTEN et al., 2011).
Support Vector Machine (SVM)	SMO	O algoritmo SMO (*Sequential Minimal Optimization*) é utilizado para classificação de vetores de suporte (SVM), usando *kernels* polinomiais ou gaussianos (PLATT 1998, KEERTHI et al., 2001). Neste algoritmo, os valores ausentes são substituídos globalmente, os atributos nominais são convertidos em atributos binários e os atributos numéricos são normalizados por padrão (WITTEN; FRANK, 2005).
	LibSVM	Classificador do tipo *wrapper* que permite a implementação de máquinas de vetores de suporte de terceiros no WEKA (WITTEN et al., 2011). Ele fornece acesso à biblioteca LIBSVM (CHANG; LIN, 2011), a qual fornece vários tipos de SVM para classificação e regressão, podendo ser selecionadas várias funções *kernels* como: linear, núcleos polinomiais, de base radial e sigmoides.
Redes Neurais Artificiais	MLP	O algoritmo gera uma rede neural artificial com pelo menos uma camada intermediária e utiliza a retro propagação de erro (*black-propagation*) para a fase de treinamento da rede, durante a qual os pesos da rede são ajustados com base no erro. Todos os nós são sigmoides (exceto para classes numéricas, nesse caso, os nós de saída se tornam unidades lineares sem limite). Os nós que representam os neurônios são ativados por meio de funções logísticas (HAYKIN, 2009).

Fonte: Elaborado pelo Autor (2020)

5.5.2 Particionamento da Base de Dados para Testes de Validação do Modelo

Esta etapa é necessária para aplicar estratégias que permitam avaliar o desempenho do modelo gerado em dados ainda desconhecidos. Nesse contexto, independentemente das métricas de avalição a serem aplicadas para avaliar o desempenho de um modelo, não é adequado avaliá-lo por seu desempenho apresentado no processo de treinamento (indução). Sempre é necessário saber como o modelo se comporta quando aplicado a dados que ainda não conheça, ou seja, que não foram usados no processo de sintonização de seus parâmetros (SILVA et al., 2016). Em suma, a proposta básica é garantir que os modelos sejam treinados com um conjunto de dados e testados por um conjunto distinto do primeiro. Essa separação garante que os modelos possam predizer (classificar) a partir de dados desconhecidos, com o grupo de teste formado por instâncias independentes do processo de construção do modelo. Este processo visa evitar que nos modelos preditivos ocorram um fenômeno bastante conhecido chamado de *overfitting* (em português, sobre ajuste) (CECHINEL; CAMARGO, 2018; SILVA et al., 2016).

O fenômeno do sobre ajuste ocorre quando o modelo preditivo apresenta um bom desempenho na fase de treinamento, porém uma baixa capacidade de generalização para novos conjuntos de dados (CECHINEL; CAMARGO, 2018). Segundo Silva et al. (2016), os modelos que se sobre ajustam às instâncias utilizadas no processo de indução perdem desempenho em relação ao erro de generalização, e se forem avaliados por seu desempenho nas novas instâncias de dados, terão avaliações muito ruins. Nesse contexto, o modelo funciona perfeitamente para a base de dados de treinamento, porém não conseguirá generalizar para novos conjuntos de dados.

Portanto, para evitar esse fenômeno é necessária a aplicação das estratégias de particionamento de dados. Logo abaixo são descritas as estratégias mais utilizadas para particionamento de bases de dados para testes de validação de modelos preditivos:

- *Holdout* - Esta estratégia divide de forma aleatória os registros em dois subconjuntos mutuamente exclusivos, sendo um conjunto para treinamento (indução) do modelo preditivo e um conjunto de teste (GOLDSCHMIDT et al., 2015). Segundo Han et al. (2011), esta divisão de subconjuntos é definida sendo 2/3 para o treinamento e 1/3 para testes;

- *K-Fold Cross-Validation* (Validação Cruzada com K Conjuntos) - Esta estratégia consiste em dividir aleatoriamente a base de dados com N elementos em K subconjuntos disjuntos (folds), os quais devem conter o número aproximado de elementos (N/K). Nesse processo apenas um dos K subconjuntos é utilizado para teste enquanto os outros K – 1 subconjuntos são utilizados para treinamento. Então, o processo é repetido K vezes, no qual todos os folds são utilizados tanto para treinamento quanto para teste (KOHAVI, 1995; GOLDSCHMIDT et al., 2015);
- *Stratifield K-Fold Cross-Validation* (Validação Cruzada com K Estratificada) - Aplicável em problemas de classificação, esta estratégia é bastante similar à validação cruzada com K conjuntos, porém o que a diferencia é que nesta estratégia a proporção das classes na amostragem dos K subconjuntos é considerada (GOLDSCHMIDT et al., 2015). Por exemplo, dada uma base de dados original na qual seus objetos são divididos em duas classe com a proporção de 20% de objetos pertencente à classe 1 e 80% dos objetos pertencente à classe 2, deve-se garantir ao se fazer a separação das bases de treinamento e teste que ambas terão a mesma proporção das classes em cada K subconjunto, ou seja, 20% da classe 1 e 80% da classe 2 (CASTRO; FERRARI, 2016);
- *Leave-One-Out* - Esta estratégia também é uma variação da validação cruzada com K conjuntos, porém a sua principal característica é que em cada K subconjuntos somente haverá um único registro, ou seja, K = N (N é a quantidade de registros). Esta estratégia deve ser utilizada em bases de dados pequenas por ser computacionalmente dispendiosa (CASTRO; FERRARI, 2016; GOLDSCHMIDT et al., 2015); e
- *Bootstrap* - Esta estratégia é um pouco similar ao *Holdout*. Ela divide os dados em dois subconjuntos, sendo um para treinamento e outro para teste (SILVA et al., 2016). Porém, o conjunto de treinamento é criado a partir de N sorteios aleatórios e com reposição do conjunto de dados original (contendo N instâncias). Já o conjunto de testes é formado por instâncias originais que não foram sorteadas para o conjunto de treinamento (GOLDSCHMIDT et al., 2015). De acordo com Han et al. (2011), dado que a escolha das N instâncias permite a reamostragem dos

> mesmos, é bastante provável que algumas instâncias sejam escolhidas mais de uma vez para compor o conjunto de treinamento. Devido a essas características, essa estratégia se torna interessante em contextos nos quais o conjunto de dados original para indução do modelo preditivo é pequeno (SILVA et al., 2016).

Para esta pesquisa foi definida a utilização a estratégia *Holdout*, na qual a base de dados foi dividida em dois subconjuntos, um com 70% das instâncias dos dados para treinamento e outro com 30% das instâncias para testes. Esta estratégia foi definida devido à necessidade da utilização da técnica de Balanceamento das Classes pela abordagem proposta SAM. Dessa forma, as classes serão balanceadas somente no subconjunto de treinamento. Já o subconjunto de testes não será balanceado visando retratar o cenário do mundo real.

5.5.3 Ajustes de Parâmetros dos Algoritmos

A fase de modelagem, além da definição dos algoritmos de mineração a serem aplicados sobre os dados, também envolve a realização de testes preliminares voltados a calibração de parâmetros do(s) algoritmo(s). Assim sendo, as técnicas de modelagem de dados são experimentadas e, em cada uma delas, diversos valores de parâmetros são testados (GOLDSCHMIDT et al., 2015). De acordo com Viana et al. (2007), o desempenho dos algoritmos de classificação é sensível aos ajustes dos parâmetros, principalmente quando se trata de problemas do mundo real. Com o objetivo de aumentar a eficácia dos algoritmos, foram realizados ajustes dos parâmetros de alguns algoritmos utilizados nesta pesquisa.

Diversos testes preliminares com alterações nos parâmetros de todos os 12 (doze) algoritmos foram realizados. Para tanto, foi utilizada a própria ferramenta WEKA, alterando os parâmetros dos algoritmos e realizando o treinamento e testes e verificando se havia aumento no desempenho das métricas de avaliação aplicadas. Apesar de haver outras ferramentas que automatizam esse processo de ajustes, visando encontrar a melhor configuração dos parâmetros, optou-se por WEKA com o objetivo de utilizar uma ferramenta única tanto na realização da modelagem quanto na avaliação dos modelos classificadores.

Alguns algoritmos obtiveram o melhor desempenho sem alteração de seus parâmetros, ou seja, com os seus parâmetros padrão. Já outros algoritmos obtiveram os seus melhores desempenhos com uma determinada configuração de seus parâmetros. Essas determinadas configurações dos parâmetros que fizeram com que os algoritmos obtivessem melhores desempenhos avaliadas pelas métricas de avaliação são mostradas no Quadro 9.

Quadro 9– Parâmetros Ajustados dos Algoritmos

ALGORITMO	**PARÂMETROS AJUSTADOS NO WEKA**	**DESCRIÇÃO**
J48	*Unpruned* = True	Remoção de PODA da Árvore de Decisão
Random Forest	Parâmetros Padrão	Parâmetros Padrão
PART	*Unpruned* = True	O Algoritmo PART também gera uma árvore de decisão, então assim como o J48, foi retirada a PODA na árvore gerada.
JRip	Parâmetros Padrão	Parâmetros Padrão
OneR	Parâmetros Padrão	Parâmetros Padrão
Naive Bayes	Parâmetros Padrão	Parâmetros Padrão
BayesNet	*searchAlgorithm* = TAN -S ENTROPY *scoreType* = ENTROPY	Método **TAN** foi selecionado para pesquisar estruturas de rede bayesianas. Já no tipo de pontuação foi aplicada a **ENTROPY** como medida usada para julgar a qualidade da estrutura da rede.
SimpleLogistic	Parâmetros Padrão	Parâmetros Padrão
IBK (KNN = 1)	Parâmetros Padrão	Parâmetros Padrão
SMO	C = 2.0 Kernel = Puk	O parâmetro de complexidade C, controla a sensibilidade, ou o grau de erros permitidos, pela flexibilidade da linha desenhada ao separar as classes. O valor C = 2.0 torna mais flexível. E o Kernel utilizado foi o PUK com os seus parâmetros padrão.
LibSVM	SVMType = nu-SVC (classification) Cost = 0.0 kernelType = polynominal	O tipo do SVM selecionado foi o nu-SVC para classificação, o custo foi reduzido para 0 (zero) e o tipo de kernel utilizado foi o polinomial.
MLP	Parâmetros Padrão	Parâmetros Padrão

Fonte: Elaborado pelo Autor (2020)

As configurações mostradas no Quadro 9 foram utilizadas no capítulo dos experimentos visando realizar a análise comparativa dos melhores desempenhos de cada algoritmo obtidos por meio da aplicação das métricas de avaliação.

5.6 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foi mostrada a abordagem proposta nesta pesquisa, denominada de *Student Assistance Mining* (SAM), a qual foi guiada pela aplicação da metodologia CRISP-DM. Assim, a primeira etapa, a do entendimento do negócio, permitiu: obter o conhecimento acerca do Plano Nacional de Assistência Estudantil (PNAES); definir os objetivos do projeto da mineração de dados; e adotar a ferramenta WEKA para ser utilizada na realização de treinamento, testes e aplicação dos algoritmos de mineração, bem como na aplicação das métricas de avaliação.

A segunda etapa, a de preparação dos dados, permitiu conhecer a base de dados da assistência estudantil do Campus Manaus Zona Leste e selecionar os dados que foram utilizados na aplicação dos algoritmos, sendo essa seleção baseada no conhecimento dos especialistas do negócio.

Na terceira etapa, a de entendimento dos dados, foram realizadas diversas tarefas como a criação de um novo atributo, o atributo alvo SITUACAO; a padronização dos atributos com a retirada de caracteres especiais como acentos, virgulas, cedilhas, traços, símbolos e a mesclagem dos valores do atributo RENDA_MENSAL_FAMILIAR; foi definida a abordagem e a técnica a ser utilizada para a seleção de atributos, tendo sido selecionada a abordagem *Filter* com a aplicação da técnica CFS. Por fim, foi definida a técnica aplicada para realizar o balanceamento das classes que foi a técnica de *Oversampling* SMOTE.

Já na quarta etapa, a de modelagem, foram definidos os algoritmos que foram aplicados nos experimentos para que os seus desempenhos fossem comparados. Além disso, também foi definida a estratégia de particionamento da base de dados para realizar os testes dos modelos, para os quais foi definida a estratégia *Holdout* onde a base de dados foi dividida em dois subconjuntos um com 70% das instâncias para treinamento e outro com 30% das instâncias para teste. Por fim, foram realizados testes preliminares com os ajustes dos parâmetros dos algoritmos visando obter o melhor desempenho dos mesmos, definindo assim quais algoritmos tiveram os seus parâmetros ajustados e quais permaneceram com as configurações

dos seus parâmetros padrão. O Quadro 10 mostra um resumo com todas as etapas da abordagem SAM e as tarefas presentes em cada uma das etapas.

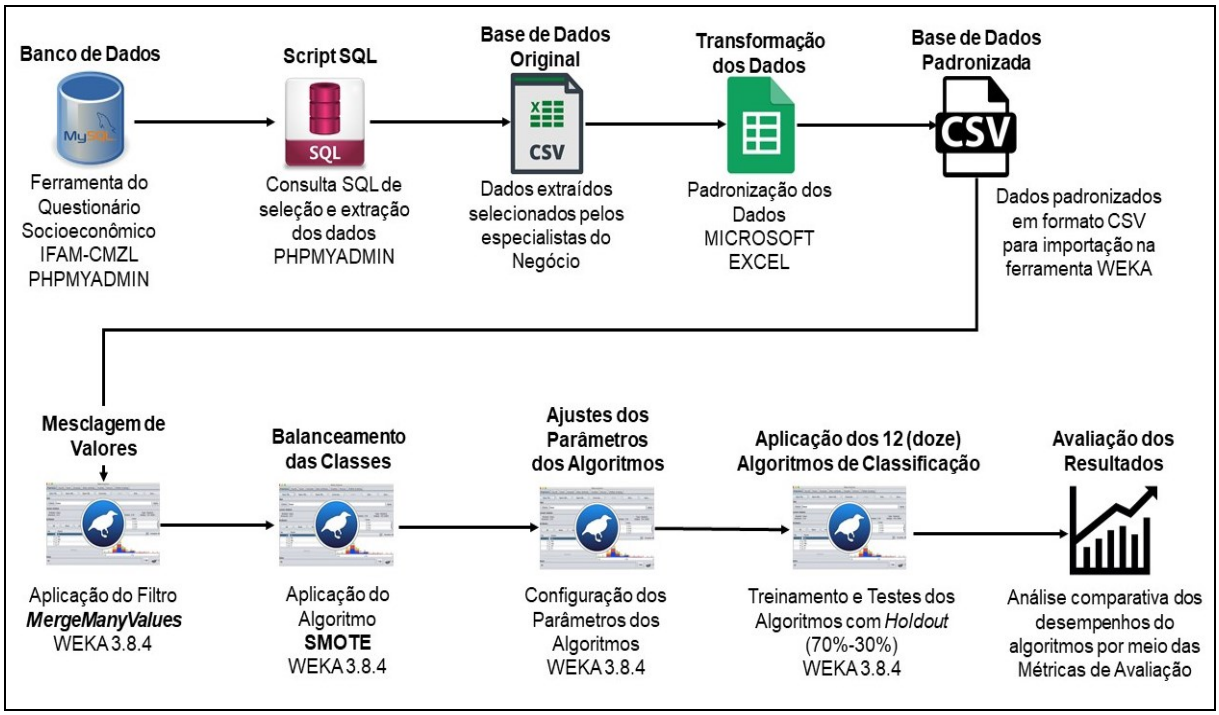
Quadro 10– Resumo da Abordagem SAM

Etapas da Abordagem SAM	Tarefas Realizadas
ENTENDIMENTO DO NEGÓCIO	• Obter o conhecimento acerca do Plano Nacional de Assistência Estudantil (PNAES); e • Definição dos objetivos do projeto da mineração de dados: o objetivo principal deste projeto de mineração é classificar os alunos participantes do processo de seleção do PNAES do Campus Manaus Zona Leste do IFAM em duas classes (APTO, NÃO APTO), a partir dos dados socioeconômicos preenchidos no questionário eletrônico da primeira fase do processo de seleção.
FERRAMENTAS UTILIZADAS	• Definição da ferramenta WEKA para realização dos experimentos com os algoritmos de Mineração de Dados.
ENTENDIMENTO DOS DADOS	• Obter o conhecimento e os dados armazenados na Base de Dados da ferramenta utilizada no processo de seleção do Programa de Assistência Estudantil do Instituto; e • Realização da Seleção dos Dados junto aos especialistas do negócio para definir quais os atributos a serem utilizados no projeto de Mineração de Dados.
PREPARAÇÃO DOS DADOS	• Criação de Novos Atributos, onde o atributo alvo (CLASSE) SITUACAO precisou ser criado; • Realização da Padronização dos Dados, onde atributos que possuíam diversos valores redundantes foram mesclados por meio da aplicação do filtro *MergeManyValues*; • Foi definida a técnica de Seleção de Atributos mais adequada para o projeto para ser aplicada nos experimentos, sendo a Abordagem *Filter* por meio da aplicação do algoritmo CFS; e a Definição da técnica para realizar o Balanceamento das Classes, sendo definido a técnica *Oversampling* por meio da aplicação do algoritmo SMOTE.
MODELAGEM	• Definição dos Algoritmos que foram aplicados nos experimentos, sendo definidos os algoritmos mais utilizados em estudos comparativos junto à tarefa de classificação; • Definição da técnica de Particionamento da Base de Dados para testes de validação, sendo definido o *Holdout* com a proporção de 70%-30%; • Definição dos ajustes a serem realizados nos parâmetros dos algoritmos para obter os melhores desempenhos; e • Determinação do algoritmo mais eficiente e eficaz para ser utilizado neste contexto.

Fonte: Elaborado pelo Autor (2020)

Todas as informações obtidas com as definições das abordagens, métodos e técnicas a serem utilizados permitiram determinar um design genérico para a aplicação dos modelos classificadores nos experimentos desta pesquisa, o qual é mostrado na Figura 22.

Figura 22– Design para Aplicação dos Algoritmos nos Experimentos



Fonte: Elaborada pelo Autor (2020)

No próximo capítulo são descritos os experimentos realizados nesta pesquisa seguindo o design mostrado na Figura 22, definido pela Abordagem *Student Assistance Mining* (SAM) desenvolvida neste capítulo.

6 EXPERIMENTOS E ANÁLISE DE DESEMPENHO DOS ALGORITMOS DE MINERAÇÃO DE DADOS

Neste capítulo são descritos os experimentos realizados nesta pesquisa visando avaliar o desempenho da abordagem proposta SAM e os respectivos resultados alcançados. Para isso, todos os experimentos seguiram o design definido no capítulo anterior, mostrado na Figura 22.

A base de dados utilizada foi a base padronizada obtida após as etapas de entendimento e preparação dos dados do CRISP-DM, aplicadas ao banco de dados gerado a partir da ferramenta do questionário eletrônico do IFAM-CMZL, compreendendo 3.631 instâncias armazenadas no período entre os anos de 2016 e 2019. Além disso, foram definidos, pelos especialistas do negócio, 11 atributos como sendo os mais relevantes para a análise socioeconômica dos alunos. Também foi criado o atributo alvo, totalizando assim 12 atributos. O Quadro 11 mostra esses atributos após a etapa de preparação dos dados.

Quadro 11– Atributos da Base de Dados Padronizada

DESCRIÇÃO	NOME ATRIBUTO	TIPO DO ATRIBUTO	INTERVALO
Gênero do Aluno	GENERO	NOMINAL	[MASCULINO, FEMININO, LGBTT]
O aluno tem filhos?	TEM_FILHOS	NOMINAL	[NAO TENHO, SIM TENHO UM (1), SIM TENHO DOIS (2), SIM TENHO TRES (3), SIM MAIS QUE TRES]
Com quem o aluno mora?	MORA_COM_QUEM	NOMINAL	[MORO SOZINHO(A), COM O PAI A MAE E IRMAOS, COM O CONJUGE, EM CASA DE FAMILIARES, EM CASA DE AMIGOS, PENSAO/HOTEL/PENSIONATO, REPUBLICA, MORADIA MANTIDA PELA FAMILIA, COM FILHOS, SO COM A MAE, SO COM O PAI, COM IRMAOS, CASA DO ESTUDANTE PAGA PELO PODER PUBLICO, INTERNATO IFAM CAMPUS MANAUS ZONA LESTE, OUTRAS MORADIAS COLETIVAS RELIGIOSA, PUBLICA, ENTRE OUTROS TIPOS]
Qual o tipo de moradia?	TIPO_MORADIA	NOMINAL	[PROPRIA, PROPRIA FINANCIADA, ALUGADA, CEDIDA, HERDADA]
Qual o meio de transporte utilizado para ir a aula?	MEIO_TRANSPORTE	NOMINAL	[TRANSPORTE PROPRIO (CARRO OU MOTO), A PE OU BICICLETA, TRANSPORTE COLETIVO, CARONA, MOTOTAXI]
O aluno trabalha?	ALUNO_TRABALHA	NOMINAL	[SIM, NAO]
Qual a renda mensal do grupo familiar?	RENDA_MENSAL_FAMILIAR	NOMINAL	[SEM RENDA, ATE MEIO SALARIO MINIMO, ATE UM SALARIO MINIMO, ATE UM E MEIO SALARIO MINIMO, ATE DOIS SALARIOS MINIMOS, ATE DOIS

			E MEIO SALARIOS MINIMOS, ATE TRES SALARIOS MINIMOS, ATE TRES E MEIO SALARIOS MINIMOS, ATE QUATRO SALARIOS MINIMOS, ATE QUATRO E MEIO SALARIOS MINIMOS; ATE CINCO SALARIOS MINIMOS, ATE CINCO E MEIO SALARIOS MINIMOS, ATE SEIS SALARIOS MINIMOS, MAIS QUE SEIS SALARIOS]
Quantas pessoas dependem dessa renda mensal?	QUANT_DEPENDE_RENDA	NOMINAL	[UMA (1), DUAS (2), TRES (3), QUATRO (4), CINCO (5), SEIS (6), SETE (7), OITO (8), NOVE (9), DEZ (10) OU MAIS]
Qual a participação do aluno na renda do seu grupo familiar?	PARTICIPA_RENDA	NOMINAL	[SOU SUSTENTADO PELA FAMILIA OU POR OUTRAS PESSOAS, RECEBO AJUDA FINANCEIRA DA FAMILIA OU DE OUTRAS PESSOAS, SOU RESPONSAVEL APENAS PELO MEU PROPRIO SUSTENTO, SOU RESPONSAVEL PELO MEU SUSTENTO E CONTRIBUO PARA O SUSTENTO DA FAMILIA, SOU RESPONSÁVEL PRINCIPAL PELO SUSTENTO DE MINHA FAMILIA]
Qual o nível de escolaridade do pai do aluno?	ESCOLARIDADE_PAI	NOMINAL	[NAO TEVE PAI OU PESSOA QUE EXERCEU TAL PAPEL NA CRIAÇAO, SEM INSTRUCAO E NAO ALFABETIZADO, SEM INSTRUCAO E SABE LER E ESCREVER, ENSINO FUNDAMENTAL 1 A 4 INCOMPLETO, ENSINO FUNDAMENTAL 1 A 4 COMPLETO, ENSINO FUNDAMENTAL 5 A 8 INCOMPLETO, ENSINO FUNDAMENTAL 5 A 8 COMPLETO, ENSINO MEDIO ANTIGO 2 GRAU INCOMPLETO, ENSINO MEDIO ANTIGO 2 GRAU COMPLETO, ENSINO SUPERIOR INCOMPLETO, ENSINO SUPERIOR COMPLETO, ESPECIALIZACAO MESTRADO OU DOUTORADO]
Qual o nível de escolaridade da mãe do aluno?	ESCOLARIDADE_MAE	NOMINAL	[NAO TEVE PAI OU PESSOA QUE EXERCEU TAL PAPEL NA CRIAÇAO, SEM INSTRUCAO E NAO ALFABETIZADO, SEM INSTRUCAO E SABE LER E ESCREVER, ENSINO FUNDAMENTAL 1 A 4 INCOMPLETO, ENSINO FUNDAMENTAL 1 A 4 COMPLETO, ENSINO FUNDAMENTAL 5 A 8 INCOMPLETO, ENSINO FUNDAMENTAL 5 A 8 COMPLETO, ENSINO MEDIO ANTIGO 2 GRAU INCOMPLETO, ENSINO MEDIO ANTIGO 2 GRAU COMPLETO, ENSINO SUPERIOR INCOMPLETO, ENSINO SUPERIOR COMPLETO, ESPECIALIZACAO MESTRADO OU DOUTORADO]
Atributo Alvo Criado	SITUACAO	NOMINAL	[APTO, NAO APTO]

Fonte: Elaborado pelo Autor (2020)

A estratégia Holdout foi definida na abordagem SAM para o particionamento da base de dados para o treinamento e testes dos algoritmos, a Tabela mostra como

ficou definido a divisão dos subconjuntos dos dados com a distribuição das instâncias.

Tabela 2– Distribuição de Instâncias de Dados nos Subconjuntos de Treinamento e Teste

Base de Dados	Base de Dados para Treinamento (70%)	Base de Dados para Testes (30%)
3631	**2542**	**1089**

Fonte: Elaborada pelo Autor (2020)

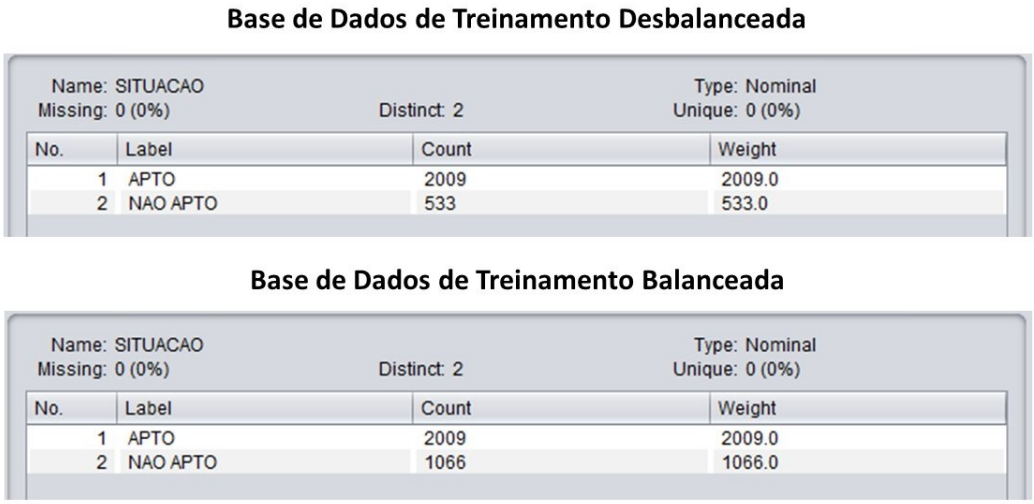
A ferramenta de mineração de dados definida para ser utilizada em todos os experimentos foi o WEKA (*Waikato Enviroment for Knowledge Analysis*), mais especificamente o ambiente Explorer. Os motivos que levaram à escolha dessa ferramenta foram descritos na Seção 5.2. Na ferramenta WEKA, apesar de já ter sido definida a técnica SMOTE na Seção 5.4.4 para realizar o balanceamento das classes, foram realizados testes preliminares com outras técnicas disponíveis no WEKA como *ClassBalancer*, *StratifiedRemoveFolds* e o próprio SMOTE com os valores do parâmetro K = 1, 3 e 5. O SMOTE com o valor de K = 1 foi a técnica de balanceamento de classes que permitiu aos algoritmos aplicados obterem os melhores desempenhos.

Esse algoritmo é utilizado no WEKA como um filtro supervisionado de instâncias. Ele foi aplicado com os parâmetros (-C 0 –K 1 –P 100.0 –S 1), onde C = 0 detecta de forma automática a classe minoritária que será utilizada pelo SMOTE; o valor padrão de K é 5, porém para esta pesquisa o valor utilizado foi k = 1, onde um vizinho mais próximos foi usado para efetuar a interpolação dos atributos para a geração dos exemplos sintéticos. Essa alteração do valor de K ocorreu devido à grande diferença de quantidade de registros de uma classe para outra, quando um valor de K maior poderia gerar novas instâncias sintéticas que poderiam ser classificadas incorretamente. Já o valor P = 100 é para criar 100% de instâncias por meio do SMOTE, e por fim, o S = 1 é o valor utilizado para a geração da amostragem aleatória, gerando assim os novos exemplos sintéticos para a classe minoritária por meio da interpolação entre suas instâncias a fim de obter o balanceamento dos dados.

Conforme descrita na seção 5.5.2 somente a base de dados de treinamento que contêm 2542 instâncias foi balanceada e base de dados de teste com 1089

instâncias de dados não foi balanceada de forma a representar os dados reais. A Figura 23 mostra como as classes estavam distribuídas antes do balanceamento e como elas ficaram após a realização do balanceamento. Na base de dados de treinamento desbalanceada a classe NÃO APTO possuía apenas 26,53% de instâncias em relação à classe APTO, após a realização do balanceamento a classe NÃO APTO passou a ter 53,06% de instâncias em relação à classe APTO. Assim sendo, contando com as instâncias sintéticas geradas, a base de dados de treinamento balanceada passou a possuir 3.075 instâncias que foram utilizadas nos experimentos.

Figura 23– Base de Dados de Treinamento Antes e Após o Balanceamento das Classes



Fonte: Elaborada pelo Autor (2020)

Por fim, para avaliar o desempenho dos algoritmos de mineração foram aplicadas as métricas de avaliação descritas na Seção 3.5: Matriz de Confusão, Acurácia, Precisão, Sensibilidade, Medida F, Estatística Kappa e Área Sob a Curva ROC (AUC). Por meio dos resultados obtidos por essas métricas é que foi conduzida a análise comparativa dos algoritmos em todos os experimentos realizados.

Todos os experimentos desta pesquisa foram realizados em um Notebook com Processador Intel Core I5-8250U 1.80 GHz, 8 Gigabytes de memória RAM, 1 Terabyte de HD e Sistema Operacional Windows 10 Pro 64 bits. Além disso, ressaltamos que todos os casos relacionados ao tempo de execução dos algoritmos foram realizados sob as mesmas condições de uso.

6.1 EXPERIMENTO 1 – APLICAÇÃO DOS ALGORITMOS E ANÁLISE COMPARATIVA DE SEUS DESEMPENHOS

Visando a análise comparativa de seus desempenhos, o primeiro experimento realizado nesta pesquisa consistiu na aplicação dos 12 (doze) algoritmos descritos na Seção 5.5.1: J48, Random Forest, PART, JRip, OneR, NaiveBayes, BayesNet, SimpleLogistic, IBK, SMO, LibSVM e MLP. A referida análise teve como objetivo a identificação do algoritmo mais eficiente e eficaz na predição dos alunos “aptos” ou não a receberem a assistência estudantil do IFAM Campus Manaus Zona Leste.

Como descrito na Seção 5.5.3, nos testes realizados alguns algoritmos obtiveram os seus melhores desempenhos relacionados à eficácia com os seus parâmetros padrão. Já outros precisaram de alguns ajustes nos parâmetros. Para estes nos quais houve ajustes, são mostrados na análise comparativa os seus resultados tanto com os seus parâmetros padrão quanto com os parâmetros ajustados. Vale ressaltar que os algoritmos que foram executados com os parâmetros padrão permaneceram assim, porque após a realização dos ajustes dos seus parâmetros, os desempenhos permaneceram os mesmos ou apresentaram uma diminuição em seus desempenhos relacionados a sua eficácia.

Assim sendo, foram executados 7 (sete) algoritmos com as configurações padrão de seus parâmetros e 5 (cinco) algoritmos foram executados com as configurações padrão e com os ajustes dos seus parâmetros, totalizando 17 diferentes aplicações dos algoritmos. Após a aplicação dos algoritmos o primeiro passo foi analisar a capacidade de predição deles, verificando o quantitativo e os percentuais de classificação corretas e incorretas das instâncias da base de dados de cada um dos algoritmos. A Tabela 3 mostra um *ranking* criado sendo ordenado do melhor desempenho para o pior desempenho na classificação correta das instâncias.

Tabela 3– Ranking dos Algoritmos em Relação à Capacidade de Classificar Corretamente as Instâncias

	Algoritmos	**Quantidade de Instâncias Classificadas Corretamente**	**Percentual de Instâncias Classificadas Corretamente**	**Quantidade de Instâncias Classificadas Incorretamente**	**Percentual de Instâncias Classificadas Incorretamente**
1º	SMO (com ajuste de parâmetros)	932	85,5831%	157	14,4169%
2º	Random Forest	905	83,1038%	184	16,8962%

3º	IBK (K = 1)	891	81,8182%	198	18,1818%
4º	LibSVM (sem ajuste de parâmetros)	868	79,7062%	221	20,2938%
5º	OneR	865	79,4307%	224	20,5693%
6º	SMO (sem ajuste de parâmetros)	859	78,8797%	230	21,1203%
7º	LibSVM (com ajuste de parâmetros)	853	78,3287%	236	21,6713%
8º	J48 (com ajuste de parâmetros)	851	78,1451%	238	21,8549%
9º	PART (com ajuste de parâmetros)	850	78,0533%	239	21,9467%
10º	JRip	843	77,4105%	246	22,5895%
11º	MLP	842	77,3186%	247	22,6814%
12º	J48 (sem ajuste de parâmetros)	838	76,9513%	251	23,0487%
13º	SimpleLogistic	832	76,4004%	257	23,5996%
14º	PART (sem ajuste de parâmetros)	827	75,9412%	262	24,0588%
15º	BayesNet (com ajuste de parâmetros)	813	74,6556%	276	25,3444%
16º	Naive Bayes	765	70,2479%	324	29,7521%
17º	BayesNet (sem ajuste de parâmetros)	763	70,0643%	326	29,9357%

Fonte: Elaborada pelo Autor (2020)

Após serem analisados os desempenhos dos algoritmos em relação à capacidade de classificar corretamente as instâncias da base de dados, conforme mostrado na Tabela 3, o próximo passo deste experimento foi realizar a análise da matriz de confusão dos algoritmos **SMO** com os parâmetros ajustados por ter obtido o melhor desempenho na classificação correta das instâncias e o **LibSVM** com os parâmetros padrão e o **LibSVM** com os seus parâmetros ajustados para mostrar como esse ajuste apesar de parecer ter um desempenho inferior, melhorou muito a distribuição da classificação das instâncias para ambas as classes. Esta análise visou verificar como esses quantitativos e percentuais de classificação corretos e incorretos estavam distribuídos entre as duas classes alvo APTO e NAO APTO.

A Tabela 4 mostra os resultados da matriz de confusão do algoritmo **SMO**, o qual obteve o primeiro lugar no *ranking* de desempenho de classificação correta das instâncias com o percentual de 85,5831%. Além disso, conforme mostrado na Tabela 4, ele também apresentou um bom desempenho na classificação correta das

instâncias para ambas as classes, obtendo 97,58% para a classe APTO e 38,46% para NAO APTO.

Tabela 4– Resultados da Matriz de Confusão do Algoritmo SMO em Quantitativos e Percentuais

	APTO	NAO APTO
APTO	**847**	21
NAO APTO	136	**85**

	APTO	NAO APTO
APTO	**97,58%**	2,42%
NAO APTO	61,54%	**38,46%**

Fonte: Elaborada pelo Autor (2020)

A Tabela 5 mostra os resultados da matriz de confusão do algoritmo **LibSVM**, sendo aplicado com as configurações padrão de seus parâmetros. Esta aplicação do algoritmo apesar de obter o 4º (quarta) posição no ranking em classificar corretamente as instâncias com o percentual de 79,7062%, essa aplicação é a que pior distribuiu a classificação correta das instâncias para ambas as classes. Conforme mostrado na Tabela 5, ele classificou corretamente somente as instâncias da classe APTO, obtendo 100%, enquanto a classe NAO APTO obteve 0% de instâncias classificadas corretamente. O **LibSVM** só obteve um alto percentual de classificação correta das instâncias devido ao fato de a base de teste está desbalanceada, onde a classe APTO possui quase 80% das instâncias de dados. Dessa forma, o algoritmo **LibSVM** mostrou um poder de predição nulo para a classe NAO APTO.

Tabela 5– Resultados da Matriz de Confusão do Algoritmo LibSVM em Quantitativos e Percentuais sem Ajuste de Parâmetros

	APTO	NAO APTO
APTO	**868**	0
NAO APTO	221	**0**

	APTO	NAO APTO
APTO	**100,00%**	0,00%
NAO APTO	100,00%	**0,0%**

Fonte: Elaborada pelo Autor (2020)

Já a Tabela 6 mostra os resultados da matriz de confusão do algoritmo LibSVM sendo aplicado com o ajuste dos parâmetros mostrado na Seção 5.5.3. Esta aplicação do algoritmo apesar de ter um desempenho inferior na classificação correta das instâncias ficando em 7º (sétimo) no ranking em relação a sua aplicação com os parâmetros padrão, realiza a distribuição bem melhor da classificação correta das instâncias para ambas as classes APTO e NÃO APTO.

Conforme mostrado na Tabela 6, ele também apresentou um aumento significativo no desempenho da classificação das instâncias para a classe NAO APTO, com 43,44%, porém uma redução na classificação correta de instâncias para a classe APTO, como 87,21%, em relação à aplicação sem ajuste de parâmetros, houve uma redução de 12,79% para a classe APTO. Dessa forma, o algoritmo LibSVM com os ajustes de parâmetros buscou equilibrar melhor a classificação correta das instâncias para as duas classes ampliando o poder de predição para a classe NAO APTO.

Tabela 6– Resultados da Matriz de Confusão do Algoritmo LibSVM em Quantitativos e Percentuais com Ajuste de Parâmetros

	APTO	NAO APTO
APTO	**757**	111
NAO APTO	125	**96**

	APTO	NAO APTO
APTO	**87,21%**	12,79%
NAO APTO	56,56%	**43,44%**

Fonte: Elaborada pelo Autor (2020)

O último passo deste experimento foi avaliar os algoritmos quanto a sua eficácia e eficiência. Para a avaliação da eficácia são mostrados os resultados obtidos por meio das métricas de avaliação utilizadas nas 17 (dezessete) aplicações dos algoritmos testados neste experimento. A ferramenta WEKA mostra os resultados obtidos de algumas métricas de avaliação para cada classe, além do resultado final consolidado obtido por meio do cálculo da média ponderada para as métricas: TP-Rate, FP-Rate, Precision, Recall, F-Measure e ROCArea (AUC). Sendo esse cálculo da média ponderada realizado conforme descrito na Seção 3.5.8. A Figura 24 mostra um exemplo dos resultados das métricas de avaliação gerados pela ferramenta WEKA.

Figura 24– Resultados das Métricas de Avaliação Gerados pelo WEKA

```
Correctly Classified Instances        2760               76.0121 %
Incorrectly Classified Instances       871               23.9879 %
Kappa statistic                          0.0309
Mean absolute error                      0.3194
Root mean squared error                  0.4245
Relative absolute error                 95.1073 %
Root relative squared error            103.6005 %
Total Number of Instances             3631
Ignored Class Unknown Instances                 1

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,945    0,923    0,791      0,945    0,861      0,039    0,572     0,826     APTO
                 0,077    0,055    0,278      0,077    0,121      0,039    0,572     0,259     NAO APTO
Weighted Avg.    0,760    0,737    0,681      0,760    0,703      0,039    0,572     0,705

=== Confusion Matrix ===

    a    b   <-- classified as
 2700  156 |    a = APTO
  715   60 |    b = NAO APTO
```

Fonte: Elaborada pelo Autor (2020)

A Tabela 7 mostra os resultados consolidados obtidos com as médias ponderadas dos resultados das métricas de avaliação utilizadas.

Tabela 7– Resultados Obtidos pelas Métricas de Avaliação Aplicadas aos Algoritmos

Algoritmo	**Acurácia**	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**Kappa**	**(ROC Area)**
SMO (com ajuste de parâmetros)	85,58%	0,856	0,495	0,850	0,856	0,835	0.4471	0,680
Random Forest	83,10%	0,831	0,455	0,817	0,831	0,821	0.4189	0,783
IBK (K = 1)	81,81%	0,818	0,421	0,811	0,818	0,814	0.4143	0,709
LibSVM (com ajuste de parâmetros)	78,32%	0,783	0,477	0,778	0,783	0,781	0.3139	0,653
PART (com ajuste de parâmetros)	78,05%	0,781	0,471	0,778	0,781	0,779	0.3135	0,690
J48 (com ajuste de parâmetros)	78,14%	0,781	0,474	0,778	0,781	0,780	0.3129	0,688
MLP	77,31%	0,773	0,503	0,767	0,773	0,770	0.2782	0,661
PART (sem ajuste de	75,94%	0,759	0,584	0,740	0,759	0,748	0.1908	0,640

parâmetros)								
BayesNet (com ajuste de parâmetros)	74,65%	0,747	0,570	0,736	0,747	0,741	0.1835	0,636
J48 (sem ajuste de parâmetros)	76,95%	0,770	0,652	0,728	0,770	0,742	0.1413	0,561
SimpleLogistic	76,40%	0,764	0,657	0,723	0,764	0,737	0.1278	0,582
SMO (sem ajuste de parâmetros)	78,87%	0,789	0,718	0,730	0,789	0,736	0.0976	0,535
Naive Bayes	70,24%	0,702	0,558	0,712	0,702	0,707	0.1103	0,560
BayesNet (sem ajuste de parâmetros)	70,06%	0,701	0,589	0,711	0,701	0,706	0.1078	0,560
JRip	77,41%	0,774	0,695	0,719	0,774	0,734	0.1018	0,541
OneR	79,43%	0,794	0,754	0,732	0,794	0,726	0.0597	0,520
LibSVM (sem ajuste de parâmetros)	79,70%	0,797	0,797	?	0,797	?	0	0,500

Fonte: Elaborada pelo Autor (2020)

Conforme destacado na cor laranja na Tabela 7, é possível observar que dentre os 12 (doze) algoritmos testados em suas 17 (dezessete) aplicações diferentes realizadas, os algoritmos, **SMO (com ajuste dos parâmetros)**, ***RandomForest*** e **IBK** foram os que obtiveram os melhores desempenhos em eficácia junto à tarefa de classificar e predizer se um aluno participante do processo de seleção da assistência estudantil do IFAM-CMZL, está APTO ou NÃO APTO a receber as bolsas da assistência. Estas três aplicações de algoritmos obtiveram uma acurácia superior a 80%. Já a aplicação do LibSVM com os parâmetros padrão obteve o pior desempenho, apesar de ter uma acurácia de 79,70%, na métrica da Estatística Kappa obteve o valor 0 (zero), e nas métricas Precision e F-Measure obteve valores nulos representados pelo símbolo "**?**", além de ter o pior desempenho na métrica AUC (ROC Area) com o valor de **0,500**.

Na Tabela 7 também é destacado (em laranja) que o algoritmo SMO (com ajuste dos parâmetros) obteve o melhor desempenho em 7 (sete) das 8 (oito) métricas aplicadas: a maior acurácia com **85,58**%, Taxa de Verdadeiros Positivos (TP-Rate) com **0,856**, *Precision* com **0,850**, *Recall* com **0,856**, F-Measure com **0,835**, Statistic Kappa com **0.4471**. Já a aplicação do algoritmo *RandomForest* obteve o melhor desempenho na métrica AUC (PRC Area) com o valor de **0,783**. E,

por fim, o algoritmo **IBK** que obteve o melhor desempenho na métrica Falsos Positivos (FP-Rate) com **0,421**, foi o algoritmo que obteve a menor média ponderada nas classificações incorretas das instâncias.

Em relação à avaliação da eficiência foi analisado o tempo que cada aplicação dos algoritmos levou para realizar o treinamento e os testes de cada modelo por meio da estratégia *Holdout* (70% - 30%). A Tabela 8 mostra os resultados obtidos com o tempo calculado em segundos para a execução do treinamento e testes dos algoritmos, sendo os mesmos classificados do menor para o maior tempo de execução.

Tabela 8– Classificação da Aplicação dos Algoritmos em Relação ao Tempo de Execução

CLASSIFICAÇÃO	ALGORITMOS	TEMPO DE EXECUÇÃO EM SEGUNDOS (TREINAMENTO E TESTE)
1º	Naive Bayes	**00.51 segundos**
2º	J48 (com ajuste de parâmetros)	**00.58 segundos**
3º	J48 (sem ajuste de parâmetros)	**00.59 segundos**
4º	BayesNet (com ajuste de parâmetros)	**00.60 segundos**
5º	BayesNet (sem ajuste de parâmetros)	**00.69 segundos**
6º	OneR	**00.71 segundos**
7º	PART (sem ajuste de parâmetros)	**00.94 segundos**
8º	IBK (K = 1)	**01.09 segundos**
9º	PART (com ajuste de parâmetros)	**01.32 segundos**
10º	JRip	**01.41 segundos**
11º	SimpleLogistic	**02.23 segundos**
12º	LibSVM (com ajuste de parâmetros)	**02.38 segundos**
13º	LibSVM (sem ajuste de parâmetros)	**02.68 segundos**
14º	Random Forest	**07.29 segundos**
15º	SMO (sem ajuste de parâmetros)	**10.09 segundos**
16º	SMO (com ajuste de parâmetros)	**40.88 segundos**
17º	MLP	**202.21 segundos**

Fonte: Elaborada pelo Autor (2020)

Como pode ser visto na Tabela 8, os algoritmos mais eficientes foram os algoritmos probabilísticos bayesianos ***Naive Bayes*** e ***BayesNet***, juntamente com o algoritmo de árvore de decisão **J48**. Porém, conforme já mostrado nas tabelas 3 e 7

relacionadas à eficácia, esses algoritmos ficaram entre os piores desempenhos. Já os algoritmos que foram menos eficientes com um maior tempo de execução, foram o algoritmo de Máquina de Vetores de Suporte, as duas aplicações do **SMO** com e sem ajuste de parâmetros, o algoritmo ***Random Forest*** e o algoritmo de Redes Neurais Artificiais, o **MLP**. Neste contexto, optou-se por avaliar somente o desempenho da eficiência da aplicação dos três algoritmos que obtiveram os melhores desempenhos em eficácia: o **SMO (com ajuste de parâmetros)**, o ***Random Forest*** e o **IBK**. O desempenho destes três algoritmos está destacado em laranja na Tabela 8, sendo **IBK** o algoritmo mais eficiente e **SMO** o menos eficiente dos três.

Entretanto apesar de o **SMO (com ajuste de parâmetros)** ter obtido o melhor desempenho em 6 (seis) das 8 (oito) métricas de avaliação aplicadas para análise da eficácia, ele teve um tempo de execução muito elevado o qual foi bem superior aos dois outros algoritmos mais eficazes, sendo a penúltima aplicação em relação a eficiência. Levando em consideração que a base de dados utilizada no experimento possui 4406 instâncias de dados, esse aumento no tempo de execução se tornará ainda maior conforme o volume da base de dados for aumentando e consequentemente ampliando esse tempo em relação ao ***Random Forest*** o segundo algoritmo mais eficaz.

Assim sendo, por ser o algoritmo que obteve o segundo melhor desempenho em quase todas as métricas de avaliação e o melhor desempenho na métrica AUC, aplicado sem precisar de ajuste em seus parâmetros, além de ter um tempo de execução para treinamento e testes muito menor em relação ao **SMO (com ajuste de parâmetros)** e relativamente pequeno em relação às aplicações dos outros algoritmos, o algoritmo ***Random Forest*** foi considerado o algoritmo mais eficiente e eficaz junto à tarefa de classificação para predizer se os alunos do IFAM-CMZL estão aptos ou não a receber a assistência estudantil.

6.2 EXPERIMENTO 2 – ANÁLISE DAS REGRAS GERADAS PELOS ALGORITMOS DE REGRAS DE CLASSIFICAÇÃO

O segundo experimento realizado nesta pesquisa consistiu na análise das regras geradas pelos algoritmos de regras de classificação. Conforme foi definido na fase de modelagem da abordagem proposta descrita na Seção 5.5.1, apenas 3 (três)

dos algoritmos de regras de classificação foram escolhidos: **PART**, **JRip** e **OneR**, os quais inclusive foram aplicados no primeiro experimento. Este experimento teve como objetivo evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica, ao verificar quais atributos e seus valores tiveram maior ocorrência para a classe APTO nas regras geradas.

Conforme mostrado no primeiro experimento desta pesquisa, os resultados obtidos por meio das métricas de avaliação mostraram que quase todos os algoritmos de regras de classificação tiveram os desempenhos mais baixos entre os 12 (doze) algoritmos aplicados sobre os dados socioeconômicos dos alunos, com a exceção do algoritmo **PART** com o ajuste do parâmetro (*Unpruned* = *True*), ou seja, sem a poda da árvore, o qual obteve o melhor desempenho conforme mostra a Tabela 9.

Tabela 9– Resultados Obtidos pelas Métricas de Avaliação Aplicadas aos Algoritmos de Regras de Classificação

Algoritmo	Acurácia	TP Rate	FP Rate	Precision	Recall	F-Measure	Kappa	(ROC Area)
PART (com ajuste de parâmetros)	78,05%	0,781	**0,471**	**0,778**	**0,781**	**0,779**	**0.3135**	**0,690**
PART (sem ajuste de parâmetros)	75,94%	0,759	0,584	0,740	0,759	0,748	0.1908	0,640
JRip	77,41%	0,774	0,695	0,719	0,774	0,734	0.1018	0,541
OneR	**79,43%**	**0,794**	0,754	0,732	0,794	0,726	0.0597	0,520

Fonte: Elaborada pelo Autor (2020)

Na Tabela 9 são destacados na cor laranja os desempenhos do algoritmo **PART** com ajuste de parâmetros, mostrando que ele obteve os melhores desempenhos em 6 (seis) das 8 (oito) métricas de avaliação aplicadas. Por este motivo, para este experimento foi considerado somente as regras de classificação geradas pela aplicação do algoritmo **PART** com o ajuste de parâmetros.

No ambiente *Explorer* da ferramenta WEKA, além dos resultados obtidos pelas métricas de avaliação, também são mostradas as regras geradas pelos algoritmos de regras de classificação. No caso do algoritmo **PART** é chamada lista de decisão PART e ao final desta lista é mostrado o número total de regras geradas.

Já ao lado de cada regra é mostrado o número de registros classificados corretamente por ela. Caso haja tanto registro corretos quanto incorretos classificados pela regra, ela é mostrada da seguinte forma (Nº de registros corretos / Nº de registros incorretos), conforme mostra a Figura 25.

Figura 25– Lista de Decisão PART com as Regras Geradas pelo Algoritmo

```
PART decision list
------------------

RENDA_MENSAL_FAMILIAR = MAIS QUE SEIS SALARIOS MINIMOS AND
ESCOLARIDADE_MAE = ENSINO MEDIO COMPLETO: NAO APTO (6.0)

RENDA_MENSAL_FAMILIAR = MAIS QUE SEIS SALARIOS MINIMOS AND
QUANT_DEPENDE_RENDA = QUATRO (4): NAO APTO (5.0)

RENDA_MENSAL_FAMILIAR = MAIS QUE SEIS SALARIOS MINIMOS AND
ESCOLARIDADE_MAE = ENSINO FUNDAMENTAL 5A A 8A INCOMPLETO: NAO APTO (3.0)

RENDA_MENSAL_FAMILIAR = SEM RENDA AND
PARTICIPA_RENDA = SOU RESPONSAVEL PRINCIPAL PELO SUSTENTO DE MINHA FAMILIA: APTO (22.0)

RENDA_MENSAL_FAMILIAR = SEM RENDA AND
GENERO = FEMININO AND
ESCOLARIDADE_MAE = ENSINO FUNDAMENTAL 1A A 4A INCOMPLETO AND
MEIO_TRANSPORTE = TRANSPORTE COLETIVO: APTO (23.0)

ESCOLARIDADE_MAE = NAO TEVE MAE OU PESSOA QUE EXERCEU TAL PAPEL NA CRIACAO AND
TEM_FILHOS = NAO: APTO (7.0)

RENDA_MENSAL_FAMILIAR = ATE CINCO SALARIOS MINIMOS AND
MORA_COM_QUEM = COM O PAI A MAE E IRMAOS AND
QUANT_DEPENDE_RENDA = CINCO (5) AND
ESCOLARIDADE_MAE = ENSINO SUPERIOR COMPLETO: NAO APTO (8.0)

ESCOLARIDADE_MAE = NAO TEVE MAE OU PESSOA QUE EXERCEU TAL PAPEL NA CRIACAO AND
TEM_FILHOS = SIM TENHO TRES (3): APTO (2.0/1.0)

ESCOLARIDADE_MAE = ENSINO FUNDAMENTAL 1A A 4A COMPLETO AND
TEM_FILHOS = SIM TENHO UM (1): APTO (18.0)

ESCOLARIDADE_MAE = ENSINO FUNDAMENTAL 1A A 4A COMPLETO AND
PARTICIPA_RENDA = SOU SUSTENTADO PELA FAMILIA OU POR OUTRAS PESSOAS AND
ESCOLARIDADE_PAI = NAO TEVE PAI OU PESSOA QUE EXERCEU TAL PAPEL NA CRIACAO: APTO (16.0)

Number of Rules  :      980
```

Fonte: Elaborada pelo Autor (2020)

Como o objetivo deste experimento é identificar o perfil do aluno em situação de vulnerabilidade socioeconômica, foram analisadas somente as regras com maior ocorrência para a classe APTO, ou seja, os perfis identificados classificados corretamente para receber a assistência estudantil. Devido ao grande número de regras geradas, sendo 673 regras, para fins de análise foram consideradas apenas as regras que englobaram no mínimo 10 registros e 3 atributos. Pode-se perceber um baixo desempenho do algoritmo com os atributos selecionados para o perfil, no

qual as regras geradas abarcaram poucos registros. Assim sendo, as regras mais relevantes encontradas são descritas a seguir.

1. **RENDA_MENSAL_FAMILIAR = SEM RENDA AND MORA_COM_QUEM = COM O PAI A MAE E IRMAOS AND MEIO_TRANSPORTE = TRANSPORTE COLETIVO AND TIPO_MORADIA = PROPRIA AND ESCOLARIDADE_MAE = ENSINO MEDIO COMPLETO: APTO (15.0)**

A Regra 1 mostra que 15 registros de alunos que não possuem renda familiar, que moram com o pai a mãe e os irmãos, utilizam o transporte coletivo para ir as aulas, possuem casa própria e cuja a mãe possui somente o Ensino Médio completo, foram classificados como “aptos” a receber a assistência estudantil.

2. **RENDA_MENSAL_FAMILIAR = SEM RENDA AND ESCOLARIDADE_MAE = ENSINO FUNDAMENTAL 1A A 4A INCOMPLETO AND MEIO_TRANSPORTE = TRANSPORTE COLETIVO: APTO (19.0)**

A Regra 2 mostra que 19 registros de alunos que não possuem renda familiar, cuja a mãe possui somente o Ensino Fundamental I incompleto e que utilizam o transporte coletivo para ir as aulas, foram classificados como “aptos” a receber a assistência estudantil.

3. **RENDA_MENSAL_FAMILIAR = ATE UM SALARIO MINIMO AND MORA_COM_QUEM = EM CASA DE AMIGOS AND PARTICIPA_RENDA = SOU SUSTENTADO PELA FAMILIA OU POR OUTRAS PESSOAS: APTO (12.0)**

A Regra 3 mostra que 12 registros de alunos que possuem renda familiar de até um salário mínimo, moram em casa de amigos e que são sustentados pelos familiares, foram classificados como “aptos” a receber a assistência estudantil.

4. **RENDA_MENSAL_FAMILIAR = ATE UM SALARIO MINIMO AND ESCOLARIDADE_MAE = SEM INSTRUCAO SABE LER E ESCREVER AND QUANT_DEPENDE_RENDA = QUATRO (4) AND PARTICIPA_RENDA = SOU SUSTENTADO PELA FAMILIA OU POR OUTRAS PESSOAS: APTO (11.0)**

A Regra 4 mostra que 11 registros de alunos que possuem renda mensal familiar de até um salário mínimo, cuja a mãe sabe somente ler e escrever, onde são 4 (quatro) pessoas da família que dependem dessa renda e que são sustentados pelos familiares, foram classificados como “aptos” a receber a assistência estudantil.

5. **MORA_COM_QUEM = COM O CONJUGE AND ESCOLARIDADE_MAE = SEM INSTRUCAO SABE LER E ESCREVER AND ESCOLARIDADE_PAI = SEM INSTRUCAO SABE LER E ESCREVER AND PARTICIPA_RENDA = SOU SUSTENTADO PELA FAMILIA OU POR OUTRAS PESSOAS AND GENERO = FEMININO: APTO (10.0)**

A Regra 5 mostra que 10 registros de alunos o nos quais o aluno mora com o cônjuge, e que tanto a mãe quanto pai não possuem instrução e sabem apenas ler e escrever, onde os mesmos são sustentados pelos familiares e são do gênero feminino, foram classificados como “aptos” a receber a assistência estudantil.

6. **ESCOLARIDADE_PAI = ENSINO FUNDAMENTAL 1A A 4A INCOMPLETO AND QUANT_DEPENDE_RENDA = QUATRO (4) AND RENDA_MENSAL_FAMILIAR = ATE UM SALARIO MINIMO: APTO (19.0)**

A Regra 6 mostra que 19 registros de alunos nos quais o pai do aluno possui somente o Ensino Fundamental I incompleto, existem 4 (quatro) pessoas que dependem da renda familiar e que essa renda mensal familiar é de até um salário mínimo , foram classificados como “aptos” a receber a assistência estudantil.

7. **MEIO_TRANSPORTE = TRANSPORTE COLETIVO AND PARTICIPA_RENDA = SOU SUSTENTADO PELA FAMILIA OU POR OUTRAS PESSOAS AND MORA_COM_QUEM = COM A MAE E IRMAOS AND QUANT_DEPENDE_RENDA = QUATRO (4) AND ESCOLARIDADE_MAE = ENSINO MEDIO COMPLETO AND TIPO_MORADIA = PROPRIA: APTO (12.0)**

A Regra 7 mostra que 12 registros de alunos nos quais o aluno utiliza o transporte coletivo para ir às aulas, onde o mesmo é sustentado pelos familiares, mora somente com a mãe e os irmãos, a quantidade de pessoas que dependem da renda familiar é de 4 (quatro) pessoas, a mãe do aluno possui somente o Ensino Médio completo e que o tipo de moradia da sua família é própria, foram classificados como “aptos” a receber a assistência estudantil.

8. **MORA_COM_QUEM = COM O PAI A MAE E IRMAOS AND ESCOLARIDADE_PAI = ENSINO MEDIO COMPLETO AND ESCOLARIDADE_MAE = ENSINO MEDIO COMPLETO AND TIPO_MORADIA = PROPRIA AND RENDA_MENSAL_FAMILIAR = ATE UM E MEIO SALARIO MINIMO AND GENERO = FEMININO AND QUANT_DEPENDE_RENDA = QUATRO (4): APTO (11.0/5.0)**

A Regra 8 mostra que 11 registros de alunos foram classificados corretamente como “aptos” a receber a assistência estudantil, porém houve 5 registros que foram classificados incorretamente como “não aptos” a receber a

assistência. Esse um conhecimento útil importante gerado, para que seja feita uma reanálise sobre os dados destes 5 alunos que deixaram de receber a assistência. Nesta regra o aluno mora com o pai a mãe e os irmãos, tanto o pai quanto a mãe do aluno possuem o ensino médio completo, o tipo de moradia da família é própria, a renda mensal familiar é de até um salário mínimo e meio, é do gênero feminino e que existem 4 pessoas que dependem da renda familiar.

9. **MORA_COM_QUEM = COM O PAI A MAE E IRMAOS AND ESCOLARIDADE_PAI = ENSINO MEDIO COMPLETO AND ESCOLARIDADE_MAE = ENSINO MEDIO COMPLETO AND TIPO_MORADIA = PROPRIA AND RENDA_MENSAL_FAMILIAR = ATE UM E MEIO SALARIO MINIMO AND GENERO = MASCULINO AND QUANT_DEPENDE_RENDA = QUATRO (4): APTO (17.0/8.0)**

A Regra 9 mostra que 17 registros de alunos foram classificados corretamente como “aptos” a receber a assistência estudantil, porém houve 8 registros que foram classificados incorretamente como “não aptos” a receber a assistência. Assim como a Regra 8, esse é outro conhecimento útil importante gerado, para que seja feita uma reanálise sobre os dados destes 8 alunos que deixaram de receber a assistência. Esta regra define como “apto” a receber a assistência o aluno que mora com o pai a mãe e os irmãos, tanto a mãe quanto o pai do aluno possuem o Ensino Médio completo, o tipo de moradia da família é própria, cuja renda mensal familiar é de até um salário mínimo e meio, é do gênero masculino e existem 4 pessoas que dependem da renda mensal familiar.

10. **MORA_COM_QUEM = COM O PAI A MAE E IRMAOS AND ESCOLARIDADE_PAI = ENSINO MEDIO COMPLETO AND ESCOLARIDADE_MAE = ENSINO MEDIO COMPLETO AND GENERO = FEMININO AND QUANT_DEPENDE_RENDA = QUATRO (4): APTO (11.0/3.0)**

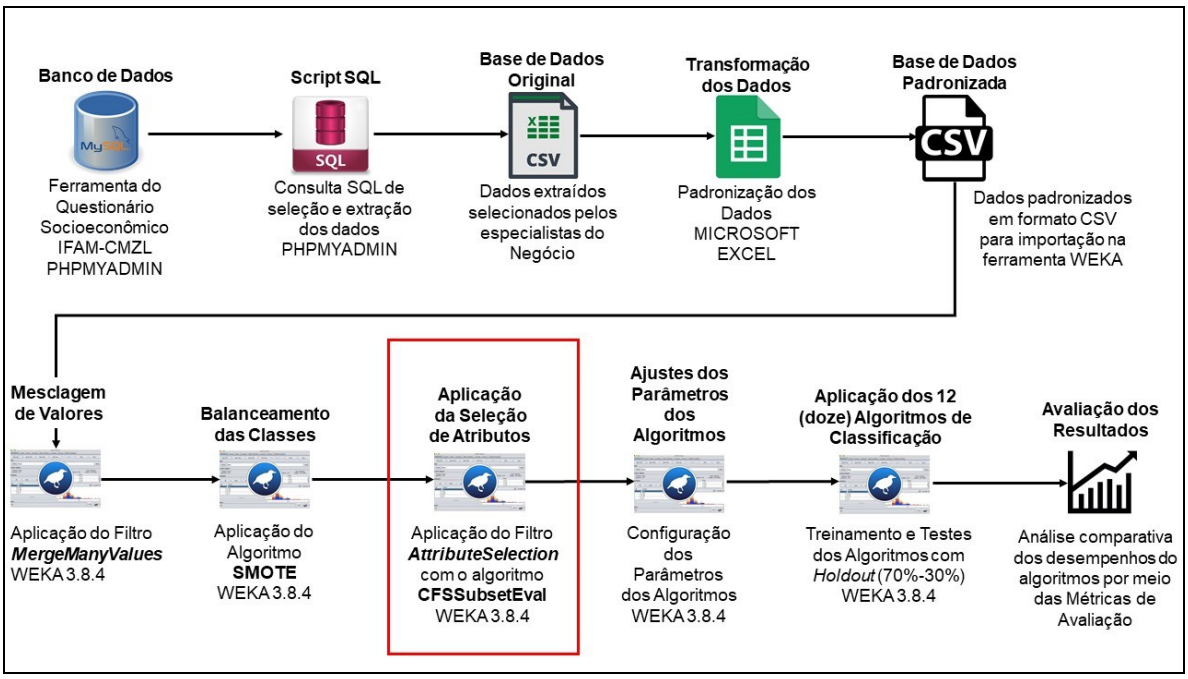
A Regra 10 mostra que 11 registros de alunos foram classificados corretamente como “aptos” a receber a assistência estudantil, porém houve 3 registros que foram classificados incorretamente como “não aptos” a receber a assistência. Assim como as Regra 8 e 9, esse é outro conhecimento útil importante gerado, para que seja feita uma reanálise sobre os dados destes 3 alunos que deixaram de receber a assistência. Esta regra define como “apto” a receber a assistência o aluno que mora com o pai a mãe e os irmãos, tanto a mãe quanto o pai do aluno possuem o Ensino Médio completo, é do gênero feminino e existem 4 pessoas que dependem da renda mensal familiar.

A partir das regras geradas pelo algoritmo **PART**, podemos verificar que algumas informações foram evidenciadas a partir dos padrões encontrados, como serem do GÊNERO FEMININO, a RENDA MENSAL FAMILIAR, a qual na maioria das regras o valor é de ATÉ UM E MEIO SALÁRIO MÍNIMO. Outra informação importante foi a utilização do TRANSPORTE COLETIVO como meio de transporte usado para ir à aula. Também ficou evidenciada pelas regras o baixo nível de escolaridade dos pais dos alunos, os quais possuem no máximo o ENSINO MÉDIO COMPLETO e que os alunos moram com O PAI A MÃE E OS IRMÃOS, ou seja, geralmente tendo 4 PESSOAS que dependem da renda familiar mensal.

6.3 EXPERIMENTO 3 – APLICAÇÃO DOS ALGORITMOS COM A SELEÇÃO DE ATRIBUTOS E ANÁLISE COMPARATIVA DOS SEUS DESEMPENHOS

Este terceiro experimento foi bastante similar ao primeiro, consistindo na aplicação dos 12 (doze) algoritmos já testados e na análise comparativa de seus desempenhos. Porém, o que diferenciou os dois experimentos entre si foi que neste terceiro foi utilizada a técnica de seleção de atributos por meio do algoritmo CFS (*Correlation-based Feature Selection*) da abordagem *Filter* descrita na Seção 3.6 e definida na Seção 5.4.4. Assim sendo, este experimento teve como objetivos analisar como a seleção de atributos influenciou no desempenho dos algoritmos, além de verificar quais os atributos mais relevantes e, com isso, gerar conhecimento útil para os especialistas do negócio. Para realização deste experimento foi necessário inserir no *design* da abordagem SAM esta etapa de seleção de atributos que ocorreu após o balanceamento das classes e antes dos ajustes dos parâmetros dos algoritmos, conforme mostrado na Figura 26.

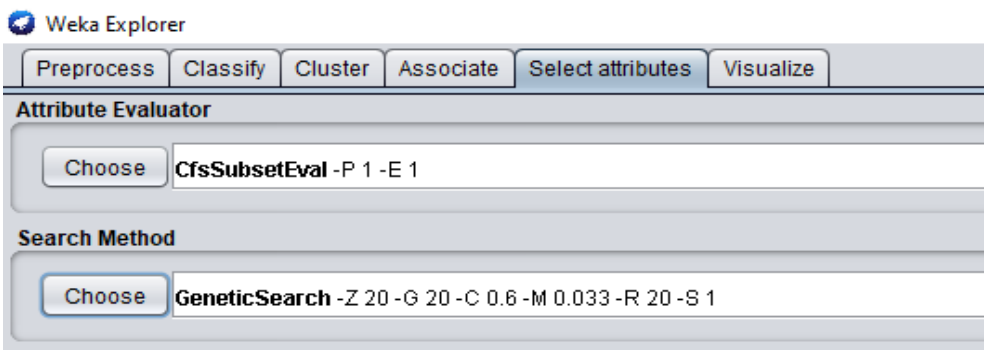
Figura 26– Design para Aplicação dos Algoritmos com a Inclusão da Técnica da Seleção de Atributos



Fonte: Elaborada pelo Autor (2020)

Inicialmente foi realizada a análise da aplicação do algoritmo CFS por meio da área denominada *Select atributes* do ambiente *Explorer* da ferramenta WEKA, conforme mostra a Figura 27. Nesta mesma figura são mostrados dois parâmetros escolhidos para os testes da seleção de atributos: o *Attribute Evaluator* e o *Search Method*. O parâmetro *Attribute Evaluator* é a técnica ou algoritmo que será aplicado e o *Search Method* é o método de busca que esta técnica ou algoritmo irá utilizar na seleção dos atributos. Os métodos de busca aplicados foram o *Best First* e o *Genetic Search*, por serem os mais utilizados em pesquisas relacionadas à seleção de atributos, como em Oliveira Júnior (2015) e Viniski e Guimarães (2017).

Figura 27– Área de Seleção de Atributos e seus Parâmetros no Ambiente Explorer do WEKA



Fonte: Elaborada pelo Autor (2020)

Nessas aplicações, foi utilizada para validação dos modelos encontrados a opção de *Cross Validation*, com 10 partições (f*olds*) seguindo o mesmo método definido em Oliveira Júnior (2015) e Viniski e Guimarães (2017). As imagens com as telas da ferramenta WEKA mostrando essas duas aplicações e os seus resultados estão no APÊNDICE B. Já a Tabela 10 mostra o comparativo dos resultados obtidos pelos dois métodos de busca utilizados, ordenando os atributos pelo percentual do nível de relação com o atributo alvo.

Tabela 10– Comparativo dos Resultados dos Métodos de Busca do Algoritmo de Seleção de Atributos CSF

Attribute Evaluator: ***CSF Subset Evaluator***		
Atributos	Search Method: ***Best First***	Search Method: ***Genetic Search***
TEM_FILHOS	100%	100%
MORA_COM_QUEM	100%	100%
TIPO_MORADIA	70%	70%
MEIO_TRANSPORTE	100%	100%
RENDA_MENSAL_FAMILIAR	100%	100%
QUANT_DEPENDE_RENDA	100%	100%
ESCOLARIDADE_PAI	100%	100%
ESCOLARIDADE_MAE	100%	100%
ALUNO_TRABALHA	0%	0%
GENERO	50%	40%
PARTICIPA_RENDA	90%	90%

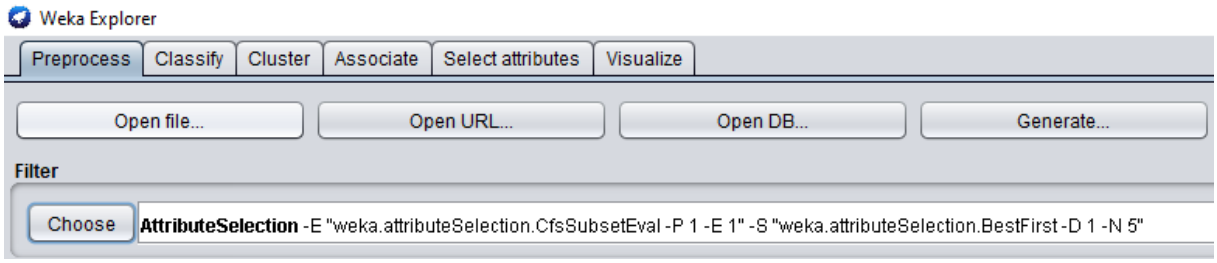
Fonte: Elaborada pelo Autor (2020)

Conforme mostrado na Tabela 10, os dois métodos de busca obtiveram resultados bastante semelhantes, sendo a única diferença uma redução de 10% para o atributo GENERO no método *Genetic Search* em relação ao método *Best First*. Porém, o resultado mais importante foi que ambos obtiveram os mesmos resultados tanto para os atributos mais relevantes como para o atributo menos relevante, os quais foram os atributos GENERO e ALUNO_TRABALHA que obtiveram apenas 50% e 0% respectivamente nos dois métodos de busca.

Assim sendo, para este experimento, os atributos GENERO e ALUNO_TRABALHA foram removidos da base de dados por meio da aplicação do filtro de seleção atributos na área de pré-processamento do ambiente Explorer da

ferramenta WEKA. Foi utilizado o algoritmo CFS com o método de busca *Best First*, conforme mostra a Figura 28.

Figura 28– Aplicação do Filtro de Seleção de Atributos no WEKA



Fonte: Elaborada pelo Autor (2020)

Após a remoção dos atributos GENERO e ALUNO_TRABALHA, o próximo passo deste experimento, assim como no primeiro experimento, foi a execução de sete (7) algoritmos com as configurações padrão de seus parâmetros e cinco (5) algoritmos sendo executados com as configurações padrão e com os ajustes dos seus parâmetros, totalizando 17 diferentes aplicações. Após a aplicação dos algoritmos, foi analisada a capacidade de predição deles, após a seleção de atributos, verificando-se como esta seleção influenciou os percentuais de classificação corretas das instâncias da base de dados de cada um dos algoritmos. A Tabela 11 mostra um *ranking* criado sendo ordenado do maior desempenho para o menor desempenho na classificação correta das instâncias, além de mostrar o comparativo dos percentuais de antes e após a aplicação da seleção de atributos.

Tabela 11– *Ranking* dos Algoritmos e Comparativo dos Percentuais de Classificação Correta das Instâncias Antes e Após a Seleção de Atributos

	Algoritmos	**Percentual de Instâncias Classificadas Corretamente EXPERIMENTO 1**	**Percentual de Instâncias Classificadas Corretamente EXPERIMENTO 3**	**Diferença de Percentual entre os Experimentos**
1º	SMO (com ajuste de parâmetros)	85,5831%	84,9403%	- 0,6428%
2º	Random Forest	83,1038%	82,0018%	- 1,102%
3º	IBK (K = 1)	81,8182%	81,7264%	- 0,0918%
5º	LibSVM (sem ajuste de parâmetros)	79,7062%	79,7062%	0%
4º	OneR	79,4307%	79,4307%	0%
6º	SMO (sem ajuste de parâmetros)	78,8797%	78,8797%	0%
7º	LibSVM (com ajuste de parâmetros)	78,3287%	75,5739%	- 2,7548%

8º	J48 (com ajuste de parâmetros)	78,1451%	77,5941%	- 0,551%
9º	PART (com ajuste de parâmetros)	78,0533%	76,8595%	- 1,1938%
11º	JRip	77,4105%	73,1864%	- 4,2241%
10º	MLP	77,3186%	78,7879%	+ 0,1134%
12º	J48 (sem ajuste de parâmetros)	76,9513%	76,9513%	0%
13º	SimpleLogistic	76,4004%	76,0331%	- 0,3673%
14º	PART (sem ajuste de parâmetros)	75,9412%	77,7778%	+ 1,8366%
15º	BayesNet (com ajuste de parâmetros)	74,6556%	73,4619%	- 1,1937%
16º	Naive Bayes	70,2479%	70,2479%	0%
17º	BayesNet (sem ajuste de parâmetros)	70,0643%	70,2479%	- 0,1863%

Fonte: Elaborada pelo Autor (2020)

Como se pode perceber na Tabela 11, somente as aplicações do algoritmo **MLP** e o **PART com parâmetros padrão** obtiveram aumento no desempenho da classificação correta das instâncias após a aplicação da seleção de atributos. No entanto outras 10 (dez) aplicações de algoritmos tiveram reduções significativas nos seus desempenhos. Da mesma forma, 5 (cinco) aplicações dos algoritmos tiveram o mesmo desempenho do primeiro experimento (sem a aplicação da seleção de atributos).

O último passo deste experimento foi avaliar os algoritmos quanto a sua eficácia e eficiência e como a seleção de atributos influenciou nos resultados. Para a avaliação da eficácia são mostrados os resultados obtidos por meio das métricas de avaliação utilizadas nas 17 (dezessete) aplicações dos algoritmos testados neste experimento, antes e após a realização da seleção de atributos. A Tabela 12 mostra os resultados consolidados obtidos das médias ponderadas das métricas de avaliação utilizadas.

Tabela 12– Resultados Obtidos pelas Métricas de Avaliação Aplicadas aos Algoritmos Antes e Após a Seleção de Atributos

SELEÇÃO DE ATRIBUTOS (SL)	**Acurácia**	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Mensure**	**Statistic Kappa**	**ROC Area**
SMO (Com ajuste de Parâmetros)								
Antes da SL	85,58%	0,856	0,495	0,850	0,856	0,835	0.4471	0,680
Após a SL	84,94%	0,849	**0,484**	0,839	0,849	0,831	0.4401	**0,683**

RANDOM FOREST								
Antes da SL	83,10%	0,831	0,455	0,817	0,831	0,821	0.4189	0,783
Após a SL	82,00%	0,820	**0,440**	0,809	0,820	0,813	0.4056	0,776
IBK (K=1)								
Antes da SL	81,81%	0,818	0,421	0,811	0,818	0,814	0.4143	0,709
Após a SL	81,72%	0,817	0,431	0,809	0,817	0,812	0.4061	**0,726**
LibSVM (Com ajuste de parâmetros)								
Antes da SL	78,32%	0,783	0,477	0,778	0,783	0,781	0.3139	0,653
Após a SL	75,57%	0,756	**0,450**	0,771	0,756	0,762	0.2882	0,653
PART (Com ajuste de parâmetros)								
Antes da SL	78,05%	0,781	0,471	0,778	0,781	0,779	0.3135	0,690
Após a SL	76,85%	0,769	**0,464**	0,773	0,769	0,771	0.2989	0,690
J48 (Com ajuste de parâmetros)								
Antes da SL	78,14%	0,781	0,474	0,778	0,781	0,780	0.3129	0,688
Após a SL	77,59%	0,776	**0,462**	0,777	0,776	0,777	0.3121	**0,720**
MLP								
Antes da SL	77,31%	0,773	0,503	0,767	0,773	0,770	0.2782	0,661
Após a SL	**78,78%**	**0,788**	**0,439**	**0,789**	**0,788**	**0,788**	**0.3476**	**0,688**
PART (Sem ajuste de parâmetros)								
Antes da SL	75,94%	0,759	0,584	0,740	0,759	0,748	0.1908	0,640
Após a SL	**77,77%**	**0,778**	**0,576**	**0,753**	**0,778**	**0,762**	**0.2271**	0,625
BAYESNET (Com ajuste de parâmetros)								
Antes da SL	74,65%	0,747	0,570	0,736	0,747	0,741	0.1835	0,636
Após a SL	73,46%	0,735	0,584	0,727	0,735	0,731	0.1555	0,625
J48 (Sem ajuste de parâmetros)								
Antes da SL	76,95%	0,770	0,652	0,728	0,770	0,742	0.1413	0,561
Após a SL	76,95%	0,770	**0,635**	**0,733**	0,770	**0,745**	**0.1584**	**0,572**
SIMPLELOGISTIC								
Antes da SL	76,40%	0,764	0,657	0,723	0,764	0,737	0.1278	0,582
Após a SL	76,03%	0,760	0,682	0,713	0,760	0,729	0.096	0,573
SMO (Sem ajuste de parâmetros)								
Antes da SL	78,87%	0,789	0,718	0,730	0,789	0,736	0.0976	0,535
Após a SL	78,87%	0,789	0,718	0,730	0,789	0,736	0.0976	0,535
NAIVE BAYES								
Antes da SL	70,24%	0,702	0,558	0,712	0,702	0,707	0.1103	0,560
Após a SL	70,24%	0,702	0,595	0,710	0,702	0,706	0.1045	0,557
BAYESNET (sem ajuste de parâmetros)								
Antes da SL	70,06%	0,701	0,589	0,711	0,701	0,706	0.1078	0,560
Após a SL	**70,24%**	**0,702**	**0,588**	**0,712**	**0,702**	**0,707**	**0.1103**	0,557
JRIP								
Antes da SL	77,41%	0,774	0,695	0,719	0,774	0,734	0.1018	0,541
Após a SL	73,18%	0,732	**0,645**	0,708	0,732	0,719	0.0948	**0,545**
OneR								

Antes da SL	79,43%	0,794	0,754	0,732	0,794	0,726	0.0597	0,520
Após a SL	79,43%	0,794	0,754	0,732	0,794	0,726	0.0597	0,520
LibSVM (Sem ajuste de parâmetros)								
Antes da SL	79,70%	0,797	0,797	?	0,797	?	0	0,500
Após a SL	79,70%	0,797	0,797	?	0,797	?	0	0,500

Fonte: Elaborada pelo Autor (2020)

Assim como houve redução na classificação correta das instâncias, a Tabela 12 mostra que também houve redução no desempenho dos resultados das métricas de avaliação aplicadas na grande maioria dos algoritmos (escritas em letra vermelha). Já as métricas que obtiveram um aumento no desempenho são destacadas na cor laranja. Dessa forma, somente 2 (duas) aplicações dos algoritmos obtiveram aumentos significativos em seus desempenhos nas métricas de avaliação aplicadas, o **MLP** que obteve aumento de desempenho em todas as métricas aplicas e o **PART com os parâmetros padrão** que obteve aumento em 7 (sete) das 8 (oito) métricas. Já as aplicações do ***BayesNet*** e o **J48** ambos com os parâmetros padrão apesar de terem obtido aumento no desempenho em quase todas as métricas de avaliação, foram pequenos aumentos com pouca significância.

Já o restante das aplicações dos algoritmos teve redução nos desempenhos das métricas de avaliação utilizadas, incluindo os três algoritmos mais eficazes antes da aplicação da seleção de atributos o **SMO (com ajuste de parâmetros)**, o ***Random Forest*** e o **IBK** que tiveram redução de desempenho em quase todas as métricas aplicadas, porém mesmo com essas reduções eles permaneceram como sendo os três mais eficazes após a aplicação da seleção de atributos.

Em relação à avaliação da eficiência, com a exclusão dos atributos GENERO e ALUNO_TRABALHA realizada por meio da seleção de atributos, já era esperado que houvesse uma redução no tempo de execução dos algoritmos, uma vez que a retirada de um atributo reduz o esforço computacional no processamento deles. Assim sendo, após a seleção de atributos, foi analisado o tempo que cada aplicação dos algoritmos levou para realizar o treinamento e os testes de cada modelo por meio da estratégia *Holdout* (70% - 30%). A Tabela 13 mostra os resultados obtidos com o tempo calculado em segundos para a execução do treinamento e testes dos algoritmos, sendo os mesmos classificados do menor para o maior tempo de execução

Tabela 13– Classificação da Aplicação dos Algoritmos em Relação ao Tempo de Execução Antes e Após a Seleção de Atributos

	Algoritmos	Tempo de Execução em segundos (Treinamento e Teste)		Diferença nos Tempos de Execução
		Antes da Seleção de Atributos	Após a Seleção de Atributos	
1º	Naive Bayes	**00.51 segundos**	**00.51 segundos**	**0**
2º	J48 (com ajuste de parâmetros)	**00.58 segundos**	**00.47 segundos**	**- 0.11**
3º	J48 (sem ajuste de parâmetros)	**00.59 segundos**	**00.42 segundos**	**- 0.17**
4º	BayesNet (com ajuste de parâmetros)	**00.60 segundos**	**00.56 segundos**	**- 0.04**
5º	BayesNet (sem ajuste de parâmetros)	**00.69 segundos**	**00.62 segundos**	**- 0.07**
6º	OneR	**00.71 segundos**	**00.50 segundos**	**- 0.21**
7º	PART (sem ajuste de parâmetros)	**00.94 segundos**	**00.72 segundos**	**- 0.22**
8º	IBK (K = 1)	**01.09 segundos**	**00.87 segundos**	**- 0.22**
9º	PART (com ajuste de parâmetros)	**01.32 segundos**	**01.40 segundos**	**+ 0.08**
10º	JRip	**01.41 segundos**	**00.79 segundos**	**- 0.62**
11º	SimpleLogistic	**02.23 segundos**	**02.53 segundos**	**+ 0.30**
12º	LibSVM (com ajuste de parâmetros)	**02.38 segundos**	**01.61 segundos**	**- 0.77**
13º	LibSVM (sem ajuste de parâmetros)	**02.68 segundos**	**02.33 segundos**	**- 0.35**
14º	Random Forest	**07.29 segundos**	**05.65 segundos**	**- 1.64**
15º	SMO (sem ajuste de parâmetros)	**10.09 segundos**	**09.45 segundos**	**- 0.64**
16º	SMO (com ajuste de parâmetros)	**40.88 segundos**	**40.04 segundos**	**- 0.84**
17º	MLP	**202.21 segundos**	**184.07 segundos**	**- 18.14**

Fonte: Elaborada pelo Autor (2020)

Como pode ser visto na Tabela 13, apenas duas aplicações dos algoritmos tiveram um aumento no tempo de execução: **PART (com ajuste de parâmetros)** e **SimpleLogistic**. Os demais algoritmos tiveram redução no tempo de execução e melhora na sua eficiência após a realização da seleção de atributos. Entretanto, assim como no Experimento 1 desta pesquisa, optou-se por avaliar somente o desempenho da eficiência da aplicação dos três algoritmos que obtiveram os melhores desempenhos em eficácia: **SMO (com ajuste de parâmetros)**, ***Random Forest*** e o **IBK**. O desempenho destes três algoritmos está destacado em laranja na Tabela 13, com **IBK** permanecendo antes e após a realização da seleção de atributos como o algoritmo mais eficiente e o **SMO** como o menos eficiente dos três.

Assim sendo, este experimento mostrou que apesar do atributo GENERO ter apenas 50% de relevância e o atributo ALUNO_TRABALHA com 0% de relevância em relação ao atributo alvo (CLASSE) conforme a aplicação do algoritmo CFS para

a seleção de atributos, as suas exclusões impactaram de forma negativa no desempenho dos algoritmos, não tendo influência significativa na definição do algoritmo mais eficaz e eficiente junto à tarefa de classificação para predizer quais os alunos que estão aptos ou não a receber a assistência estudantil do IFAM-CMZL. E por fim, este experimento também serviu para destacar que a técnica seleção de atributos não é necessária para a abordagem proposta SAM devido aos atributos utilizados neste experimento, uma vez que já haviam sidos selecionados previamente pelos especialistas no negócio, na etapa de preparação dos dados da abordagem.

6.4 CONSIDERAÇÕES FINAIS DO CAPÍTULO

Neste capítulo foram descritos os três experimentos realizados nesta pesquisa. Primeiramente foi descrita a base de dados utilizada nos experimentos e a realização do balanceamento das classes por meio do algoritmo SMOTE, utilizando o ambiente *Explorer* da ferramenta WEKA.

O primeiro experimento consistiu na aplicação dos 12 (doze) algoritmos utilizando a estratégia *Holdout* (70% - 30%) para partição da base de dados para treinamento e testes definidos na etapa de modelagem da abordagem SAM e a utilização das métricas de avaliação definidas na Seção 3.5. Os 12 (doze) algoritmos foram aplicados, sendo 7 (sete) algoritmos com as configurações padrão de seus parâmetros e 5 (cinco) algoritmos tanto com as configurações padrão quanto com os ajustes dos seus parâmetros visando melhorar os seus desempenhos, totalizando assim 17 aplicações diferentes dos algoritmos. As métricas de avaliação foram aplicadas para realizar a análise comparativa dos desempenhos dos algoritmos e descobrir qual deles é o mais eficaz.

Na análise comparativa da eficácia, primeiro foi analisada a capacidade de classificação correta e incorreta das instâncias de dados. Logo após foi analisada a matriz de confusão do algoritmo SMO (com ajuste de parâmetros) que obteve o melhor desempenho na classificação correta das instâncias, e das aplicações do LibSVM com parâmetros padrão e com ajuste de parâmetros para verificar como esta classificação correta e incorreta estava dividida entre as classes APTO e NAO APTO e, por fim, foram analisados os resultados das métricas de avaliação. Após todas essas análises, o algoritmo **SMO (com ajuste de parâmetros)** foi o que

obteve a maior eficácia. Por fim, foi analisado o tempo de execução para o treinamento e testes dos algoritmos visando identificar qual a algoritmo que foi o mais eficiente. Entretanto, o principal objetivo desta pesquisa é identificar o algoritmo que possa ser considerado o mais eficiente e eficaz. Assim, foi analisado somente o tempo de execução dos três algoritmos com o melhor desempenho na eficácia, sendo o algoritmo **IBK** o mais eficiente e **SMO** o menos eficiente. Assim sendo, como o ***Random Forest*** obteve o melhor desempenho na métrica AUC e o segundo melhor desempenho nas outras 7 (sete) métricas aplicadas, não precisar de alterações em seus parâmetros e ser muito mais eficiente que o **SMO (com ajuste de parâmetros)**, pode-se assim considerar o algoritmo ***Random Forest*** como o mais eficaz e eficiente.

Assim sendo, este experimento além de alcançar um dos objetivos principais desta pesquisa, também responde à sua segunda questão de pesquisa: Qual é o algoritmo de mineração de dados mais eficiente e eficaz para identificar quais alunos estão aptos a receber a assistência estudantil do IFAM-CMZL? A resposta é o algoritmo ***Random Forest*.**

O segundo experimento consistiu na análise das regras geradas pelo algoritmo de regras de classificação que obteve os melhores desempenhos nas métricas de avaliação, sendo a aplicação do algoritmo **PART** com ajuste do parâmetro de remoção da poda o que obteve os melhores desempenhos. Para esta análise foram definidos alguns critérios para selecionar as regras. O primeiro deles foi o de analisar somente regras geradas para a classe APTO, já que o objetivo deste experimento foi evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica. Outros critérios definidos foram: o número de registros que foram classificados pela regra, sendo no mínimo 10, bem como o número de atributos que a regra deveria ter, no mínimo 3. Isso foi necessário para reduzir o número de regras a serem mostradas, haja visto que o algoritmo gerou um total de 673 regras. A partir dos padrões encontrados nas regras geradas, foi possível verificar quais atributos e seus valores tiveram maior ocorrência para a classe APTO, podendo auxiliar a evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica.

Assim sendo, este experimento além de alcançar um dos objetivos principais desta pesquisa, também responde à sua primeira questão da pesquisa: É possível por meio da mineração de dados, evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica? A resposta é que os alunos, em situação de

vulnerabilidade socioeconômica do IFAM Campus Manaus Zona Leste, são do gênero feminino, cuja renda mensal familiar é de até um salário mínimo; utilizam o transporte coletivo para ir às aulas; não trabalham e são dependentes economicamente da família; tanto o pai quanto a mãe possuem no máximo o ensino médio completo; e essas famílias são compostas geralmente por 4 pessoas que dependem diretamente da renda mensal familiar.

O terceiro experimento foi bastante semelhante ao primeiro, com a única diferença que os algoritmos foram utilizados após a aplicação da técnica de seleção de atributos. Esta foi aplicada por meio do filtro *SelectAttribution* usando como avaliador de atributos o algoritmo CFS e o método de busca foi o *BestFirst*. Nesse experimento, os atributos GENERO e ALUNO_TRABALHA foram excluídos da base de dados por terem respectivamente apenas 50% e 0% de relevância em relação à classe alvo. Após a aplicação da técnica, as métricas de avaliação foram empregadas para realizar a análise comparativa e avaliar como a seleção de atributos influenciou nos desempenhos dos algoritmos.

Tanto na análise da classificação correta e incorreta das instâncias dos dados como nos resultados das métricas de avaliação aplicadas, ficou comprovado que houve muito mais redução nos desempenhos dos algoritmos do que aumentos. Somente a aplicação dos algoritmos **MLP** com as configurações padrão teve melhorias significativas nos desempenhos das métricas de avaliação. Assim sendo, a técnica de seleção de atributos, apesar da redução no desempenho das métricas de avaliação, não ocasionou alterações relevantes em relação às aplicações dos três algoritmos mais eficazes **SMO (com ajuste de parâmetros)**, ***Random Forest*** e **IBK**. Já em relação à eficiência quase todos os algoritmos tiveram redução no tempo de execução, melhorando os seus desempenhos. Isso se deve ao fato de que reduzir a quantidade de atributos também reduz o esforço computacional. Somente duas aplicações tiveram um pequeno aumento de tempo de execução.

Contudo, o objetivo do Experimento 3 foi o de analisar se os atributos selecionados pelos especialistas do negócio na etapa de preparação dos dados da abordagem SAM estão corretos e, se mesmo após essa definição dos atributos, seria necessário aplicar uma técnica de seleção de atributos automatizada na abordagem proposta SAM. Conforme comprovado neste terceiro experimento, não há necessidade de aplicar a técnica da seleção de atributos, pois mesmo os atributos apresentando apenas 50% e 0% de relevância para a classe alvo, ainda

são muito importantes para um bom desempenho dos algoritmos e para a análise socioeconômica dos alunos.

Após a realização dos três experimentos, é possível responder à terceira questão da pesquisa: Quais os possíveis conhecimentos úteis que podem ser gerados a partir de aplicação da mineração de dados sobre essa base de dados? Os conhecimentos úteis gerados estão diretamente relacionados ao resultado do processo de concessão da assistência estudantil, ressaltando que há erros na análise socioeconômica realizada manualmente. Conforme mostrado pelos resultados das métricas de avaliação dos algoritmos aplicados nos experimentos 1 e 3, e comprovadas por algumas regras geradas pelo experimento 2, alguns alunos que de acordo com um determinado padrão encontrado estavam APTOS a receber a assistência estudantil por algum motivo foram considerados NÃO APTOS. Esses erros encontrados na análise provavelmente influenciaram negativamente na construção dos modelos e classificadores e, consequentemente, na sua capacidade de predição.

Em relação às hipóteses desta pesquisa, a hipótese H1 afirma que: A mineração de dados torna mais rápido o processo de análise socioeconômica dos alunos melhorando a eficácia dessa análise. Após a realização dos experimentos foi possível comprovar que esta hipótese é verdadeira, haja visto que conforme descrito na motivação desta pesquisa, a análise socioeconômica dos dados dos alunos demanda uma grande quantidade de tempo atualmente. Porém, após a implantação da abordagem SAM e aplicação do algoritmo mais eficiente e eficaz, o *Random Forest*, essa análise socioeconômica pode ser realizada em minutos dependendo da quantidade de alunos inscritos, além de manter a eficácia da análise que foi baseada na base de dados histórica dos últimos 4 anos da assistência estudantil.

Já a hipótese H2 afirma que: O emprego de técnicas de mineração de dados permite definir o perfil do aluno em situação de vulnerabilidade socioeconômica a partir dos seus dados socioeconômicos. Após a realização dos experimentos foi possível comprovar que esta hipótese também é verdadeira, já que por meio das regras geradas pelo algoritmo PART, pode-se identificar quais são os atributos e os valores que mais ocorrem e possuem mais relevância. Esses padrões permitiram evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica.

O próximo capítulo apresenta a conclusão, as contribuições e limitações desta pesquisa bem como as propostas de trabalhos futuros.

7 CONCLUSÕES

Este trabalho desenvolveu uma abordagem específica para Mineração de Dados Socioeconômicos de alunos do Campus Zona Leste do IFAM, baseada na metodologia CRISP-DM, denominada SAM. Para tanto, esta dissertação buscou identificar o algoritmo de mineração de dados mais eficiente e eficaz para definir quais alunos estão aptos a receber a assistência estudantil, além de evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica, apoiando assim a tomada de decisão das assistentes sociais na definição dos alunos ingressantes aptos para o recebimento da assistência estudantil do PNAES. A base de dados original foi obtida por meio dos dados armazenados do programa de assistência estudantil, porém, apenas uma pequena parte foi utilizada no processo de mineração.

Dessa forma, tanto o objetivo geral quanto os objetivos específicos desta pesquisa foram alcançados, na qual foi desenvolvida a abordagem proposta SAM. Para tanto, foram selecionados somente os atributos diretamente relacionados à renda familiar do aluno, ressaltando-se que esses atributos foram selecionados em conjunto com os especialistas do negócio, as assistentes sociais do IFAM-CMZL. Foram definidas duas diferentes classes para a classificação dos alunos e aplicadas as métricas de avaliação que avaliaram o desempenho dos algoritmos e a capacidade de predição dos mesmos junto a essas classes, além do tempo de execução dos algoritmos, sendo o algoritmo ***Random Forest*** definido como o eficiente e eficaz. Além disso, por meio do algoritmo de regras de classificação **PART**, foi evidenciado o perfil do aluno em situação de vulnerabilidade socioeconômica utilizando os padrões encontrados pelo algoritmo.

Ao se alcançar os objetivos do trabalho tornou-se possível responder as questões da pesquisa e comprovar as suas hipóteses. Assim, por meio de uma técnica de mineração de dados, com as regras de classificação **PART**, pode-se evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica. Foi também definido, pelos experimentos realizados, o algoritmo ***Random Forest*** como o mais eficiente e eficaz para identificar os alunos aptos a receber a assistência estudantil. E, por fim, ainda por meio das regras do algoritmo **PART** pode-se gerar conhecimentos úteis, ressaltando-se que pode haver erros na análise socioeconômica realizada atualmente de forma manual, pois alguns alunos que de

acordo com um determinado padrão encontrado estavam APTOS a receber a assistência estudantil, por algum motivo foram considerados NÃO APTOS.

Assim sendo, a utilização da mineração de dados por meio da abordagem SAM obteve um resultado satisfatório uma vez que permitiu identificar o algoritmo mais eficiente e eficaz, além de gerar novas informações, como evidenciar o perfil do aluno em situação de vulnerabilidade socioeconômica e, também, auxiliar na avaliação do conjunto de atributos selecionados, comprovando a sua relevância no desempenho dos algoritmos. SAM ainda mostrou que a técnica de seleção de atributos não trouxe melhorias no desempenho dos algoritmos, causando redução no desempenho de quase todas as aplicações testadas.

7.1 CONTRIBUIÇÕES

As principais contribuições desta pesquisa foram:

- Desenvolvimento de abordagem específica para aplicar algoritmos de mineração de dados sobre bases de dados socioeconômicos de alunos dos Institutos Federais (IF), denominada Student Assistance Mining (SAM);
- Desenvolvimento de uma metodologia de pré-processamento dos dados da assistência estudantil do IFAM-CMZL para que os algoritmos de mineração de dados fossem aplicados sobre uma base de dados padronizados;
- Realização de uma análise comparativa do desempenho dos algoritmos de mineração de dados da tarefa de classificação aplicados à base de dados padronizada, para identificar o algoritmo mais eficiente e eficaz na classificação de alunos que estão aptos a receber a assistência estudantil. Nesse caso, o Random Forest foi o algoritmo que obteve o melhor equilíbrio entre a eficácia e a eficiência nesta análise comparativa;
- A identificação do perfil do aluno em situação de vulnerabilidade socioeconômica por meio das regras geradas pelo algoritmo que obteve o melhor desempenho dentre os algoritmos de regras de classificação, o PART; e

- A avaliação dos atributos definidos pelos especialistas do negócio, para verificar a relevância dos mesmos, por meio de uma análise comparativa de antes e após a aplicação da técnica de seleção de atributos.

7.2 LIMITAÇÕES

As limitações encontradas durante o desenvolvimento desta pesquisa foram:

- O quantitativo de instâncias disponíveis na base de dados para serem utilizadas na análise comparativa do desempenho, conforme descrito na Seção 5.5.2, os dados referentes aos anos de 2011 a 2015, foram perdidos. Isso limitou a utilização de uma amostra maior de instâncias. Essa redução de instâncias limitou a eficácia dos algoritmos, o que consequentemente gerou menos regras do que o esperado para alguns atributos, haja visto que quanto menos instâncias de dados, menos padrões são encontrados durante a classificação;
- O ambiente de aplicação da pesquisa limitado a somente um campus do IFAM, já que o mesmo possui atualmente cerca de outros 15 campi além do CMZL. Isso também influencia na redução de instâncias de dados;
- Direcionamento para dados obtidos de uma ferramenta específica, a qual é a ferramenta atualmente utilizada no Campus Manaus Zona Leste. Uma vez que os outros campi podem utilizar outras ferramentas na realização do processo seletivo da assistência estudantil, isso pode causar uma redução tanto na capacidade do pré-processamento dos dados quanto na abrangência da abordagem proposta; e
- Na terceira e última etapa da metodologia, a última fase da metodologia CRISP-DM, a fase de IMPLANTAÇÃO não pôde ser realizada devido à Pandemia do Covid-19 que iniciou no primeiro semestre de 2020 e coincidiu justamente com a realização desta fase, impedindo, portanto, a sua aplicação.

7.3 TRABALHOS FUTUROS

Considerando a continuidade desta dissertação, a principal sugestão de trabalho futuro é o desenvolvimento e a implantação de um software que implemente toda a abordagem proposta nesta pesquisa, desde a importação dos dados, pré-processamento, aplicação do algoritmo mais eficiente e eficaz, o Random Forest, até a predição dos alunos aptos e não aptos a receberem a assistência estudantil, automatizando todo o processo da abordagem e facilitando a utilização do conhecimento extraído pelos especialistas do negócio.

Outras sugestões de trabalhos futuros são:

- Ampliação do ambiente da pesquisa e da base dados, com a unificação das instâncias dos dados da assistência estudantil de todos os 16 (dezesseis) Campi do IFAM, tornando a abordagem proposta mais abrangente e permitindo assim, definir o perfil do aluno em situação de vulnerabilidade socioeconômica de todo o instituto;
- Aplicação de outras técnicas de mineração de dados da tarefa de classificação com algoritmos meta classificadores, como por exemplo: AdaBoostM1, Bagging e LogitBoost, ou a utilização da combinação de algoritmos de mineração na tarefa de classificação usando diferentes métodos de comitê de classificadores como Stacking e Vote, por exemplo; e por fim, realizar a análise comparativa do desempenho dessas técnicas com as demais utilizadas nesta pesquisa.

REFERÊNCIAS

ADEODATO, P. J. L.; FILHO, M. M. S.; RODRIGUES, R. L. **Predição de desempenho de escolas privadas usando o ENEM como indicador de qualidade escolar**. In: Simpósio Brasileiro de Informática na Educação-SBIE. 2014. p. 891-895. Disponível em: <https://www.br-ie.org/pub/index.php/sbie/article/view/3025>. Acesso em: 17 de agosto de 2019.

AHA, D. W.; KIBLER, D.; ALBERT, M. K. **Instance-based learning algorithms**. Machine learning, v. 6, n. 1, p. 37-66, 1991. Disponível em: < https://link.springer.com/article/10.1007/BF00153759> Acesso em: 02 de abril de 2020.

ALMEIDA, J. de A. **Monitoramento do Programa Nacional de Assistência Estudantil (PNAES): Uma Análise do IFPE Campus Belo Jardim**. 2018. 108 f. Dissertação (Mestrado Profissional em Políticas Públicas) - Departamento de Ciência Política, Universidade Federal de Pernambuco, Recife, 2018. Disponível em: <https://attena.ufpe.br/handle/123456789/32324> Acesso em: 15 de fevereiro de 2020.

AZEVEDO, A. I. R. L.; SANTOS, M. F. **KDD, SEMMA and CRISP-DM: a parallel overview**. In: IADIS European Conference on Data Mining, 2008. p.182–185. Disponível em: <https://recipp.ipp.pt/bitstream/10400.22/136/3/KDD-CRISP-SEMMA.pdf>. Acesso em: 31 de outubro de 2019.

BARANAUSKAS, J. A. **Extração automática de conhecimento por múltiplos indutores**. 2001. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Ribeirão Preto, 2001. doi:10.11606/T.55.2001.tde-08102001-112806. Acesso em: 22 de março de 2020.

BAKER, R. S.; INVENTADO, P. S. Educational Data Mining and Learning Analytics. In: LARUSSON, J. A.; WHITE, B. **Learning Analytics: From Research to Practice**. New York: Springer, 2014. p. 61-75. Disponível em: < https://link.springer.com/chapter/10.1007/978-1-4614-3305-7_4>. Acesso em: 26 de setembro de 2019.

BAKER, R. S.; YACEF, K. **The state of educational data mining in 2009: A review and future visions**. JEDM-Journal of Educational Data Mining, v. 1, n. 1, p. 3-17, 2009. ISSN 2157-2100. Disponível em: < https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8> Acesso em: 30 de março de 2020.

BELLMAN, R. E. **Adaptive control processes: a guided tour**. Princeton, NJ: Princeton University Press, 1961.

BLUM, A. L.; LANGLEY, P. **Selection of relevant features and examples in machine learning**. Artificial Intelligence, 1997, p. 245–271.

BOSCARIOLI, C. et al. **Analyzing HCI Issues in Data Clustering Tools**. In:

International Conference on Human Interface and the Management of Information. London: Springer, Cham, 2014. v. 8521, p. 22-33. Disponível em: < https://doi.org/10.1007/978-3-319-07731-4_3> Acesso em: 19 de fevereiro de 2020.

BORGES, H. B. **Redução de dimensionalidade em bases de dados de expressão gênica**. 2006. 123 f. Dissertação (Mestrado em Informática) - Pontifícia Universidade Católica do Paraná, Curitiba, 2006. Disponível em: < https://www.ppgia.pucpr.br/pt/arquivos/mestrado/dissertacoes/2006/2006_helyane.pdf> Acesso em: 20 de março de 2020.

BRAGA, A. P.; CARVALHO, A.P.L.F.; LUDERMIR, T.B. **Redes neurais artificiais: teoria e aplicações**. 2. ed. Rio de Janeiro: LTC, 2007.

BRASIL. Decreto 7.234, de 19 de julho de 2010. **Dispõe sobre o Programa Nacional de assistência estudantil (PNAES)**. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2010/decreto/d7234.htm>. Acesso em: 15 de fevereiro de 2020.

BREIMAN, L. **Random forests**. Machine learning, Springer, v. 45, n. 1, p. 5–32, 2001. Disponível em: < https://link.springer.com/article/10.1023/A:1010933404324> Acesso em: 01 de abril de 2020.

CASTRO, L. N.; FERRARI, D. G. **Introdução à Mineração de Dados: conceitos básicos, algoritmos e aplicações**. 1. ed. São Paulo: Saraiva, 2016.

CECHINEL, C.; CAMARGO, S. da S. **Mineração de dados educacionais: avaliação e interpretação de modelos de classificação**. In: JAQUES, Patrícia Augustin; SIQUEIRA; Sean; BITTENCOURT, Ig; PIMENTEL, Mariano. (Org.) Metodologia de Pesquisa Científica em Informática na Educação: Abordagem Quantitativa. Porto Alegre: SBC, 2020. (Série Metodologia de Pesquisa em Informática na Educação, v. 2) Disponível em: <https://metodologia.ceie-br.org/livro-2> Acesso em: 03 de abril de 2020.

CHANG, C-C.; LIN, C-J. **LIBSVM: A library for support vector machines**. ACM transactions on intelligent systems and technology (TIST), v. 2, n. 3, p. 1-27, 2011. Disponível em: < https://dl.acm.org/doi/abs/10.1145/1961189.1961199 > Acesso em: 02 de abril de 2020.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C.; WIRTH, R. **CRISP-DM 1.0 Step-by-step data mining guide**. Illinois: SPSS, 2000.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. **Smote: Synthetic minority over-sampling technique**. Journal of Artificial Intelligence Research, v. 16, p. 321–357, 2002. Disponível em: < https://www.jair.org/index.php/jair/article/view/10302> Acesso em: 28 de março de 2020.

CHAWLA, N. V. **Data mining for imbalanced datasets: An overview**. In: Data mining and knowledge discovery handbook. Springer, Boston, MA, 2009. p. 875-886. Disponível em: <https://link.springer.com/chapter/10.1007/978-0-387-09823-4_45> Acesso em: 28 de março de 2020.

COHEN, J. **A coefficient of agreement for nominal scales**. Educational and Psychological Measurement, v. 20, n. 1, p. 37-46, 1960. ISSN 1552-3888. Disponível em:<https://journals.sagepub.com/doi/abs/10.1177/001316446002000104?journalCode=epma> Acesso em: 08 de outubro de 2019.

COHEN, W. W. **Fast effective rule induction**. In: Machine learning proceedings 1995. Morgan Kaufmann, 1995. p. 115-123. Disponível em: < https://www.sciencedirect.com/science/article/pii/B9781558603776500232> Acesso em: 01 de abril de 2020.

CORDEIRO, R. G. **Identificação do comportamento dos estudantes evadidos de cursos técnicos utilizando técnicas de mineração de dados**. 2017. 79 f. Dissertação (Mestrado em Sistemas Aplicados à Engenharia e Gestão) – Campus Centro, Instituto Federal Fluminense, Campos dos Goytacazes, 2017. Disponível em:<http://portal1.iff.edu.br/pesquisa-e-inovacao/pos-graduacao-stricto-sensu/mestrado-profissional-em-sistemas-aplicados-a-engenharia-e-a-gestao/dissertacoes-1/identificacao-do-comportamento-dos-estudantes-evadidos-de-cursos-tecnicos-utilizando-tecnicas-de-mineracao-de-dados>. Acesso em 12 de agosto de 2019.

DASH, M.; LIU, H. **Feature selection for classification**. Intelligent data analysis, v. 1, n. 3, p. 131-156, 1997.

DELEN, D. **Predicting Student Attrition with Data Mining Methods**. Journal of College Student Retention: Research, Theory & Practice, v. 13, n. 1, p.17–35, 2011. ISSN 1521-0251. Disponível em: < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.826.3380&rep=rep1&type=pdf>. Acesso em: 27 de setembro de 2019.

DEVASENA, L. C. **Proficiency Comparison of ZeroR, RIDOR and PART Classifiers for Intelligent Heart Disease Prediction**. International Journal of Advances in Computer Science and Technology (IJACST), v. 3, n. 11, p. 12-18, 2014.

DEVI, J.; SEHGAL, N. **A Technique for Improving Software Quality using Support Vector Machine**, International Journal of Computer Sciences and Engineering, v. 5, n. 6, p.100-105, 2017. Disponível em: < https://www.ijcseonline.org/full_paper_view.php?paper_id=1309>. Acesso em: 15 de outubro de 2019.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F de. **Inteligência Artificial: Uma Abordagem de Aprendizagem de Máquina**, LTC/Grupo Gen, 2011.

FAWCETT, T. **An introduction to ROC analysis. Pattern recognition letters**. v. 27, n. 8, p. 861-874, 2006. ISSN 0167-8655. Disponível em: < https://www.cin.ufpe.br/~fatc/AM/roc.pdf>. Acesso em: 09 de outubro de 2019.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases.** AI Magazine, [S.l.], v. 17, n. 3, p.37-54. 1996. Disponível em: <https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230>. Acesso em: 17 de junho de 2019.

FRANK E.; WITTEN I. H. **Generating Accurate Rule Sets Without Global**

Optimization. In: Fifteenth International Conference on Machine Learning, p. 144-151, 1998. Disponível em: <https://dl.acm.org/doi/10.5555/645527.657305> Acesso em: 01 de abril de 2020.

FREITAS, A. A. **Data mining and knowlwdge discovery with evolutionary algorithms**. Springer-Verlag Berlin Heidelberg New York, 1998.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: Um Guia Prático**. Rio de Janeiro: Elsevier, 2005.

GOLDSCHMIDT, R.; BEZERRA, E.; PASSOS, E. **Data mining: conceitos, técnicas, algoritmos, orientações e aplicações**. 2. ed. Rio de Janeiro-RJ: Elsevier, 2015.

GOMES, T. M. dos S. **Ferramentas open source de Data Mining**. 2014. 147 f. Dissertação (Mestrado em Informática e Sistemas) - Instituto Superior de Engenharia de Coimbra, Instituto Politécnico de Coimbra, Coimbra, 2014. Disponível em: < http://comum.rcaap.pt/handle/10400.26/14084> Acesso em: 18 de fevereiro de 2020.

GOSWAMI, S.; CHAKRABARTI, A. **Feature selection: A practitioner view**. International Journal of Information Technology and Computer Science (IJITCS), v. 6, n. 11, p. 66, 2014. Disponível em: < http://j.mecs-press.net/ijitcs/ijitcs-v6-n11/IJITCS-V6-N11-10.pdf > Acesso em: 25 de março de 2020.

GU, Q.; CAI, Z.; ZHU, L.; HUANG, B. **Data mining on imbalanced data sets**. In: 2008 International Conference on Advanced Computer Theory and Engineering. IEEE, 2008. p. 1020-1024. Disponível em: < https://ieeexplore.ieee.org/abstract/document/4737112> Acesso em: 27 de março de 2020.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise multivariada de dados**. 6. ed. Porto Alegre: Bookman, 2009. ISBN 0-13-032929-0.

HALL, M. A. **Correlation-based Feature Selection for Machine Learning**. 1999. 178 f. Tese (Doutorado) - Department of Computer Science, University of Waikato, Hamilton, 1999.

HALL, M. A. **Correlation-based feature selection for discrete and numeric class machine learning**. In: Proceedings of the Seventeenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000. (ICML '00), p. 359-366. Disponível em: <http://dl.acm.org/citation.cfm?id=645529.657793> Acesso em: 25 de março de 2020.

HALL, M. et al. **The WEKA data mining software: an update**. ACM SIGKDD explorations newsletter, v. 11, n. 1, p. 10-18, 2009. Disponível em: < http://www.cms.waikato.ac.nz/~ml/publications/2009/weka_update.pdf > Acesso em: 20 de fevereiro de 2020.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. 2. ed. San Francisco: Morgan Kaufmann Publishers, 2006.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: Concepts and Techniques**. 3.ed. Waltham: Morgan Kaufmann Publishers Inc., 2011. ISBN 9780123814807.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. MIT press, 2001. ISBN 026208290X.

HARRISON, T. H. **Intranet Data Warehouse**. [S.l.]: Berkeley, 1998.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning:** prediction, inference and data mining. 2nd ed. New York: Springer Verlag, 2009. 745 p.

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed. [S.l.]: New Jersey: Pearson Education Inc., 2009.

HOSMER, D. W.; LEMESHOW, S. **Applied logistic regression**. New Jersey: John Wiley & Sons, Inc., 2000.

HOSMER, D. W.; LEMESHOW, S.; MAY, S. **Applied regression analysis**. New Jersey**:** John Wiley & Sons, Inc., 2008.

JAMES, G.; HASTIE, T.; WITTEN, D.; TIBSHIRANI, R. **An introduction to statistical learning: with applications in R**. New York: Springer, 2013. ISBN 978-1-4614-7138-7.

JOHN, G. H.; KOHAVI, R.; PFLEGER, K. **Irrelevant features and the subset selection problem**. In: Machine Learning Proceedings. Morgan Kaufmann, 1994. p. 121-129. Disponível em: < https://www.sciencedirect.com/science/article/pii/B9781558603356500234> Acesso em: 22 de março de 2020.

JOHN G. H.; LANGLEY P. **Estimating continuous distributions in Bayesian classifiers**. In: Eleventh conference on uncertainty in artificial intelligence. p. 338-345, 1995. Disponível em: < https://arxiv.org/abs/1302.4964> Acesso em: 02 de abril de 2020.

KAMPFF, A. J. C. **Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente**. 2009. 186 f. Tese (Doutorado em Informática na Educação) - Centro Interdisciplinar de Novas Tecnologias na Educação, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2009. Disponível em: < https://lume.ufrgs.br/handle/10183/19032>. Acesso em: 18 de setembro de 2019.

KEERTHI, S. S.; SHEVADE, S. K.; BHATTACHARYYA, C.; MURTHY, K. R. K. **Improvements to Platt's SMO algorithm for SVM classifier design**. Neural Computation, v. 13, n. 3, p. 637–649, 2001. Disponível em: < https://www.mitpressjournals.org/doi/abs/10.1162/089976601300014493> Acesso em: 02 de abril de 2020.

KLEINBAUM, D. G.; KLEIN, M. Analysis of matched data using logistic regression. In: KLEINBAUM, D. G.; KLEIN, M. **Logistic Regression**. New York: Springer, 2010. p.389-428. ISBN 978-1-4419-1741-6. Disponível em: <

https://link.springer.com/chapter/10.1007%2F0-387-21647-2_8>. Acesso em: 28 de setembro de 2019.

KOCH, G. G.; LANDIS, J. R. **The measurement of observer agreement for categorical data**. Biometrics, v. 33, n. 1, p. 159-174, jan. 1977. ISSN 0006-341X. Disponível em: <https://www.jstor.org/stable/2529310?seq=1> Acesso em: 08 de outubro de 2019.

KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI), v. 2, p. 1137-1143, 1995. Disponível em: < https://www.researchgate.net/profile/Ron_Kohavi/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection/links/02e7e51bcc14c5e91c000000.pdf> Acesso em: 03 de abril de 2020.

KOHAVI, R.; JOHN, G. H. **Wrappers for feature subset selection**. Artificial intelligence, v. 97, n. 1-2, p. 273-324, 1997. Disponível em: < https://www.sciencedirect.com/science/article/pii/S000437029700043X> Acesso em: 24 de março de 2020.

KOHAVI, R.; SOMMERFIELD, D.; DOUGHERTY, J. **Data mining using mlc++, a machine learning library in c++**. In: Proceedings of the 8th International Conference on Tools with Artificial Intelligence. Washington, DC, USA: IEEE Computer Society, 1996. (ICTAI '96), p. 234–. ISBN 0-8186-7686-8. Disponível em: <http://dl.acm.org/citation.cfm?id=850949.853584>. Acesso em: 29 de março de 2020.

KOTSIANTIS, S. B. **Supervised machine learning: A review of classification techniques**. Emerging artificial intelligence applications in computer engineering, v. 160, p. 3–24, 2007. Disponível em: < http://www.informatica.si/index.php/informatica/article/viewFile/148/140>. Acesso em: 22 de setembro de 2019.

LAZAR, C. et al. **A survey on filter techniques for feature selection in gene expression microarray analysis**. IEEE/ACM Transactions on Computational Biology and Bioinformatics, v. 9, n. 4, p. 1106-1119, 2012. Disponível em: < https://dl.acm.org/doi/10.1109/TCBB.2012.33 > Acesso em: 22 de março de 2020.

LEE, H. D. **Seleção de atributos importantes para a extração de conhecimento de bases de dados**. 2005. 182 f. Tese (Doutorado em Ciência da Computação e Matemática Aplicada) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2005, Disponível em: < https://www.teses.usp.br/teses/disponiveis/55/55134/tde-22022006-172219/publico/tese_huei.pdf > Acesso em: 21 de março de 2020.

LIMA, R. A. F. de. **Estratégias de seleção de atributos para detecção de anomalias em transações eletrônicas**. 2016. 94 f. Dissertação (Mestrado em Ciência da Computação) — Departamento de Ciência da Computação, Universidade Federal de Minas Gerais, Belo Horizonte, 2016, Disponível em: < https://repositorio.ufmg.br/bitstream/1843/ESBF-ACXKTA/1/rafaelfrancalima.pdf > Acesso em: 25 de março de 2020.

LIU, H.; MOTODA, H. **Feature Extraction, Construction and Selection: A Data Mining Perspective**. Kluwer Academic Publishers, Norwell, MA, USA. ISBN 0792381963. (1998a).

LIU, H.; MOTODA, H. **Feature Selection for Knowledge Discovery and Data Mining**. Kluwer Academic Publishers, Norwell, MA, USA. ISBN 079238198X. (1998b).

LIU, H.; YU, L. **Toward Integrating Feature Selection Algorithms for Classification Vand Clustering**. IEEE Transactions on Knowledge and Data Engineering, v.17, n. 4, p. 491-502, 2005. Disponível em: < https://ieeexplore.ieee.org/abstract/document/1401889> Acesso em: 26 de março de 2020.

LIU, X-Y.; WU, J.; ZHOU, Z-H. **Exploratory undersampling for class-imbalance learning**. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), v. 39, n. 2, p. 539-550, 2008. Disponível em: < https://ieeexplore.ieee.org/abstract/document/4717268> Acesso em: 28 de março de 2020.

LÔBO, M. T. F. **Contribuições ao Problema de Seleção de Atributos Mineração de Dados**. 2015. 119 f. Tese (Doutorado em Computação) – Universidade Federal Fluminense, Niterói, 2015. Disponível em: < http://www.ic.uff.br/PosGraduacao/frontend-tesesdissertacoes/download.php?id=707.pdf&tipo=trabalho>. Acesso em: 20 de março de 2020.

LORENA, A. C.; CARVALHO, A. C. **Uma introdução às support vector machines**. Revista de Informática Teórica e Aplicada, v. 14, n. 2, p. 43–67, 2007. Disponível em: <https://seer.ufrgs.br/rita/article/view/rita_v14_n2_p43-67>. Acesso em: 03 de outubro de 2019.

MAIMON, O.; ROKACH, L. **Data Mining and Knowledge Discovery Handbook**. 2nd. Springer, New York, 2010. ISBN 978-0-387-09822-7. Disponível em: < https://link.springer.com/book/10.1007%2F978-0-387-09823-4 >. Acesso em: 05 de abril de 2020.

MANHÃES, L.M.B.; CRUZ, S.M.S.; COSTA, R.J.M.; ZAVALETA, J.; ZIMBRÃO, G. **Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados**. Anais do XXII SBIE-XVII WIE, p. 150-159, 2011. Disponível em: < https://www.br-ie.org/pub/index.php/sbie/article/view/1585> Acesso em: 31 de março de 2020.

MÁRQUEZ-VERA, C.; MORALES, C. R.; SOTO, S. V. **Predicting school failure and dropout by using data mining techniques**. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, IEEE, v. 8, n. 1, p. 7–14, 2013. Disponível em: < https://ieeexplore.ieee.org/abstract/document/6461622 > Acesso em: 27 de março de 2020.

MEHDI T.; BASHARDOOST N.; AHMADI M. **Kernel smoothing for ROC curve and estimation for thyroid stimulating hormone**, Int J Public Health Res Special Issue, Specia, p.239–242, 2011. Disponível em: <http://journalarticle.ukm.my/3560/>.

Acesso em: 15 de outubro de 2019.

MICHIE, D.; SPIEGELHALTER, D.; TAYLOR, C. **Machine Learning, Neural and Statistical Classifications**. Ellis Horwood, 1994.

MITCHELL, T. M. **Machine Learning**. New York, NY, USA: McGraw-Hill Science/Engineering/Math, 1997. ISBN 0070428077.

MONTGOMERY, D. C.; RUNGER, G. C.; HUBELE, N. F. **Engineering statistics**. 5. Ed. Hoboken: John Wiley & Sons, 2011.

MUCHERINO, A.; PAPAJORGJI, P.J.; PARDALOS, P.M. **Data mining in agriculture.** Gainesville: Springer, 2009.

NASCIMENTO, R. F. F.; ALCÂNTARA, E.; KAMPEL, M.; STECH, J. L.; NOVO, E.; FONSECA, L. M. G. **O algoritmo support vector machines (svm): avaliação da separação ótima de classes em imagens ccd-cbers-2**. In: Simpósio Brasileiro de Sensoriamento Remoto, v. 14, p. 2079–2086, 2009. Disponível em: < http://marte.sid.inpe.br/col/dpi.inpe.br/sbsr%4080/2008/10.20.10.59/doc/2079-2086.pdf>. Acesso em: 02 de outubro de 2019.

OLIVEIRA JÚNIOR, J. G. de. **Identificação de padrões para a análise da evasão em cursos 2015 de graduação usando mineração de dados educacionais**. 2015. 86 f. Dissertação (Mestrado) - Programa de Pós-graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná, Curitiba, 2015. Disponível em: < http://repositorio.utfpr.edu.br:8080/jspui/handle/1/1995> Acesso em: 30 de maio de 2020.

PAES, B. C.; PLASTINO, A.; FREITAS, A. A. **Seleção de atributos aplicada à classificação hierárquica**. In: Symposium on Knowledge Discovery, Mining and Learning-KDMiLe. 2013. Disponível em: < https://homepages.dcc.ufmg.br/~gfrancis/kdmile/papers/KDMiLe.pdf > Acesso em: 23 de março de 2020.

PARMEZAN, A. R. S. et al. **Avaliação de Métodos para Seleção de Atributos Importantes para Aprendizado de Máquina Supervisionado no Processo de Mineração de Dados**. 2012. Relatórios técnicos do laboratório de bioinformática, Universidade estadual do oeste do Paraná, Foz do Iguaçu, 2012. Disponível em: < http://179.106.223.20:8000/portal_labi/papers/Tech_Parmezan_UNIOESTE_2012_Avaliacao.pdf> Acesso em: 24 de março de 2020.

PEÑA-AYALA, A. **Educational data mining: A survey and a data mining-based analysis of recent works**. Expert Systems with Applications**,** v. 41, n. 4, PART 1, p. 1432-1462, 2014. ISSN 0957-4174. Disponível em: < https://www.sciencedirect.com/science/article/pii/S0957417413006635>. Acesso em: 12 de setembro de 2019.

PEÑA-AYALA, A.; DOMÍNGUEZ, R.; MEDEL, J. D. J. **Educational data mining: a sample of review and study case**. World Journal On Educational Technology, v. 1, n. 2, p. 118-139, 2009. ISSN 1309-0348. Disponível em: < http://archives.un-pub.eu/index.php/wjet/article/viewFile/134/55> Acesso em: 30 de março de 2020.

PLATT, J. **Fast training of support vector machines using sequential minimal optimization**. In B. Schölkopf, C. Burges, & A. Smola (Eds.), Advances in kernel methods: Support vector learning. Cambridge, MA: MIT Press. p. 185–209, 1998.

PIATETSKY, G. **CRISP-DM, ainda é a principal metodologia utilizada para análise, mineração de dados ou projetos de ciência dos dados**. 2014. Disponível em KDnuggets: <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>. Acesso em 1 de novembro de 2019.

PRATI, R.; BATISTA, G.; MONARD, M. **Curvas ROC para avaliação de classificadores**. Revista IEEE América Latina, v. 6, n. 2, p. 215-222, 2008. Disponível em: <http://conteudo.icmc.usp.br/pessoas/gbatista/files/ieee_la2008.pdf>. Acesso em: 12 de outubro de 2019.

QUINLAN, J. R. C4.5: **Programs for machine learning**. Morgan Kauffmann, 1993.

ROMERO, C.; VENTURA, S. **Educational data mining: a review of the state of the art**. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), v. 40, n. 6, p. 601-618, 2010. Disponível em: < https://ieeexplore.ieee.org/abstract/document/5524021> Acesso em: 31 de março de 2020.

SANTANA JÚNIOR, W. M. **Mineração de dados do ENEM para a predição do desempenho acadêmico no âmbito da rede federal de educação tecnológica**. 2018. 168 f. Dissertação (Mestrado em Ciência da Computação) – Centro de Informática, Universidade Federal de Pernambuco, Recife, 2018. Disponível em: <https://repositorio.ufpe.br/handle/123456789/29994>. Acesso em: 12 de agosto de 2019.

SANTOS, A. M. T. B. **Mineração de dados educacionais: Um estudo sobre dados socioeconômicos na educação na base de dados do INEP**. 2019. 86 f. Dissertação (Mestrado em Engenharia Elétrica) - Instituto de Tecnologia, Universidade Federal do Pará, Belém, 2019. Disponível em: < http://repositorio.ufpa.br/jspui/handle/2011/11265>. Acesso em: 13 de agosto de 2019.

SAEYS, Y.; INZA, I.; LARRANAGA, P. **A review of feature selection techniques in bioinformatics**. bioinformatics, v. 23, n. 19, p. 2507-2517, 2007.

SHARMA, S.; OSEI-BRYSON, K-M.; KASPER, G. M. **Evaluation of an integrated Knowledge Discovery and Data Mining process model**. Expert Systems with Applications, v. 39, n. 13, p. 11335-11348, 2012. Disponível em: < https://www.sciencedirect.com/science/article/abs/pii/S0957417412002886> Acesso em: 22 de fevereiro de 2020.

SHEARER, C. **The CRISP-DM Model: The New Blueprint for Data Mining**. Journal of Data Warehousing**,** v. 5, n. 4, p. 13-22, 2000. Disponível em: <https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf>. Acesso em: 28 de outubro de 2019.

SILVA, I.; SPATTI, D.; FLAUZINO, R. **Redes Neurais Artificiais para Engenharia e Ciências Aplicadas**. São Paulo: Artliber, 2010. ISBN 9788588098534.

SILVA, L. A.; MORINO, A. H.; SATO, T. M. C. **Prática de Mineração de Dados no Exame Nacional do Ensino Médio**. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação, v. 3, n. 1, p. 651–660, 2014. Disponível em: < https://www.br-ie.org/pub/index.php/wcbie/article/view/3289>. Acesso em: 14 de agosto de 2019.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados Com Aplicações em R**. 1. ed. Rio de Janeiro: Elsevier, 2016.

SIMON, A.; CAZELLA, S. **Mineração de Dados Educacionais nos Resultados do ENEM de 2015**. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação, p. 754-763, 2017. Disponível em: <https://br-ie.org/pub/index.php/wcbie/article/view/7461>. Acesso em: 15 de agosto de 2019.

STEARNS, B.; RANGEL, F.; FIRMINO, F.; RANGEL, F.; OLIVEIRA, J. **Prevendo Desempenho dos Candidatos do ENEM Através de Dados Socioeconômicos**. In: Anais do XXXVI Concurso de Trabalhos de Iniciação Científica da SBC. 2017. Disponível em: <https://sol.sbc.org.br/index.php/ctic/article/view/3244>. Acesso em: 16 de agosto de 2019.

TACONELI, C. A. **Árvores de classificação multivariadas fundamentadas em coeficientes de dissimilaridade e entropia**. 2008. 99 p. Dissertação (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2008. Disponível em: < https://teses.usp.br/teses/disponiveis/11/11134/tde-15102008-082243/pt-br.php>. Acesso em: 16 de setembro de 2019.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao Datamining: mineração de dados**. 1ª Edição. Rio de Janeiro: Ciência Moderna, 2009. ISBN 8573937610.

VIANA, R. et al. **Svm with stochastic parameter selection for bovine leather defect. classification**. In: MERY, D.; RUEDA, L. (Ed.). Advances in Image and Video Technology. [S.l.]: Springer Berlin Heidelberg, 2007, (Lecture Notes in Computer Science, v. 4872). p. 600-612. ISBN 978-3-540-77128-9. Disponível em: < https://link.springer.com/content/pdf/10.1007/978-3-540-77129-6_52.pdf>. Acesso em 04 de outubro de 2019.

VIEIRA, F.D. **Modelos baseados em técnicas de mineração de dados para suporte à certificação racial de ovinos**. 2014. 88 p. Dissertação (Mestrado em Engenharia Agrícola) – Faculdade de Engenharia Agrícola, Universidade Estadual de Campinas, Campinas, 2014. Disponível em: < http://repositorio.unicamp.br/jspui/handle/REPOSIP/257128>. Acesso em: 14 de setembro de 2019.

VINISKI, A. D.; GUIMARÃES, A. M. **Técnicas de seleção de atributos para mineração de dados de alta dimensionalidade gerados por espectroscopia no infravermelho próximo–NIR**. Anais SULCOMP, v. 8, 2017. Disponível em: < http://periodicos.unesc.net/sulcomp/article/view/3142> Acesso em: 30 de maio de

2020.

VITERBO, J. et al. **Avaliação de Ferramentas de Apoio ao Ensino de Técnicas de Mineração de Dados em Cursos de Graduação**. In: WORKSHOP SOBRE EDUCAÇÃO EM COMPUTAÇÃO (WEI), 24., 2016, Porto Alegre. Anais do XXIV Workshop sobre Educação em Computação. Porto Alegre: Sociedade Brasileira de Computação 2016. p. 11-20. Disponível em <DOI: https://doi.org/10.5753/wei.2016.9644>. Acesso em: 19 de fevereiro de 2020.

WAHBEH, A. H. et al. **A comparison study between data mining tools over some classification methods**. International Journal of Advanced Computer Science and Applications, v. 8, n. 2, p. 18-26, 2011. Disponível em: <https://www.researchgate.net/profile/Mohammed_Al-Kabi/publication/251422102_A_Comparison_Study_between_Data_Mining_Tools_over_some_Classification_Methods/links/00b495345aa85dd423000000/A-Comparison-Study-between-Data-Mining-Tools-over-some-Classification-Methods.pdf>. Acesso em: 18 de fevereiro de 2020.

WANG, J. **Encyclopedia of data warehousing and mining**. 2. Ed. Hershey: IGI Global, 2008. ISBN 978-1-60566-010-3.

WIRTH, R.; HIPP, J. **CRISP-DM: Towards a standard process model for data mining**. In: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining. London, UK: Springer-Verlag, 2000. p. 29-39. Disponível em: < http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf> Acesso em: 15 de fevereiro de 2020.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3. ed. Amsterdam: Morgan Kaufmann, 2011. ISBN: 978-0-12-374856-0.

WU, X. et al. **Top 10 algorithms in data mining**. Knowledge and information systems, v. 14, n. 1, p. 1-37, 2008. Disponível em: < https://link.springer.com/article/10.1007/s10115-007-0114-2> Acesso em: 29 de março de 2020.

ZHANG, S.; LI, X.; ZONG, M.; ZHU, X.; CHENG, D. **Learning k for kNN Classification**. ACM Transactions on Intelligent Systems and Technology, v. 8, n. 3, p. 1–19, 12 jan. 2017. Disponível em: < https://dl.acm.org/citation.cfm?id=2990508>. Acesso em: 23 de setembro de 2019.

ZHANG, H. **The optimality of naive Bayes**. In: Proceedings of the 17th FLAIRS conference, AAAI Press, v. 1, n. 2, p. 562-567, 2004. Disponível em: <https://www.aaai.org/Papers/FLAIRS/2004/Flairs04-097.pdf>. Acesso em: 20 de setembro de 2019.

APÊNDICE A – DADOS PESSOAIS E PERGUNTAS DO QUESTIONÁRIO SOCIOECONÔMICO DO IFAM-CMZL

Quadro A.1 – Todos os Dados Pessoais dos Alunos e as Perguntas Necessárias para o Processo de Concessão da Assistência Estudantil do IFAM-CMZL

DADOS PESSOAIS			
Nome	Turma	Bairro	Grau Parentesco Resp.
Matrícula	Turno	Nacionalidade	
Nome da Mãe	Modalidade	Cidade	Fone Celular
Nº CPF	Gênero	Estado	E-mail
Data de Nascimento	Endereço	Nome do Pai	
Curso	Número	Nome do Responsável	
QUESTIONÁRIO SOCIOECONÔMICO – VOCÊ E SUA FAMILIA			
P01 - Como você se considera:		P05 - Você tem filhos?	
P02 - Você indicou indígena, quais línguas você domina?		P06 - Você participa financeiramente no sustento dos filhos?	
P03 - Qual é a sua religião?		P07 - Você paga pensão?	
P04 - Seu estado civil?		P09 - Qual o estado civil de seus pais?	
QUESTIONÁRIO SOCIOECONÔMICO – SITUAÇÃO DE MORADIA			
P10 - Onde você MORAVA antes de entrar no IFAM-ZL?		P13 - Qual o principal meio de transporte que você utiliza para chegar ao IFAM - ZONA LESTE?	
P11 - Qual sua situação atual de moradia?		P14 - Na casa de sua família tem?	
P12 - Tipo de Moradia de sua Família:			
QUESTIONÁRIO SOCIOECONÔMICO – VOCÊ E SUA RENDA			
P15 - Você trabalha?		P19 - Qual a renda mensal do seu grupo familiar?	
P17 - Quantas pessoas moram em sua casa?		P20 - Quantas pessoas, incluindo você, vivem da renda mensal do seu grupo familiar informada acima?	
P18 - Quem é o a principal mantenedora de sua família? A pessoa que mais contribui na renda:		P21 - Qual a sua participação na vida econômica do seu grupo familiar?	
QUESTIONÁRIO SOCIOECONÔMICO – VOCÊ E A ESCOLA			
P22 - Qual é a escolaridade de seu pai? Ou da pessoa que ao criou como pai:		P29 - Você frequentou cursinho pré-vestibular durante pelo menos seis meses?	
P23 - Qual é a escolaridade de sua mãe? Ou da pessoa que oa criou como mãe:		P30 - Quantas vezes prestou vestibular?	
P24 - Em que tipo de escola você cursou o Ensino Fundamental?		P31 - De que forma você entrou nesse curso?	
P25 - Você frequentou escola particular no EF, utilizou bolsa de estudo?		P32 - Em que turno você está frequentando a maior parte das disciplinas?	
P26 - Qual o tipo de Ensino Médio você cursou?		P33 - Qual o seu domínio da língua estrangeira Inglês?	
P27 - Em que tipo de escola você cursou o Ensino Médio?		P34 - Qual o seu domínio da língua estrangeira Francês?	
		P35 - Qual o seu domínio da língua estrangeira Espanhol?	

P28 - No Ensino Médio particular utilizou bolsa de estudo?	P36 - Você se identifica com a escola?
QUESTIONÁRIO SOCIOECONÔMICO – VOCÊ E SUAS ATIVIDADES	
P37 - Qual sua principal fonte de informação de acontecimentos atuais? P38 - Qual desses assuntos lhe desperta mais interesse P39 - Qual é a média de livros que você lê em um ano? P40 - Que tipo de livro você mais lê? P41 - Com que frequência você participa de atividades extraclasse Artística-Culturais? P42 - Com que frequência você participa de atividades extraclasse de Movimento Estudantil?	P43 - Com que frequência você participa de atividades extraclasse de Movimentos Ecológicos? P44 - Com que frequência você participa de atividades extraclasse de Movimentos Religiosos? P45 - Com que frequência você participa de atividades extraclasse de Movimentos Sociais? P46 - Com que frequência você participa de atividades extraclasse Política – Partidárias? P47 - Com que frequência você participa de atividades extraclasse de Sociedades Científicas?
QUESTIONÁRIO SOCIOECONÔMICO – VOCÊ E SUA SAÚDE	
P48 - Em geral, quando você precisa de atendimento médico você procura: P49 - Sua última consulta médica ocorreu: P50 - Com relação a seus cuidados dentários, você: P51 - Qual atividade física você mais pratica? P52 - Com que frequência você pratica essa atividade? P53 - Esta atividade é normalmente encarada por você como? P54 - Caso você não pratique nenhuma atividade física fora do IFAM_CMZL, qual a razão principal? P55 - Você já teve alguma dificuldade significativa ou crise emocional nos últimos meses? P56 - Você já procurou atendimento psico-pedagógico alguma vez em sua vida? P57 - Você já procurou atendimento psicológico alguma vez em sua vida? P58 - Você já procurou atendimento psiquiátrico alguma vez em sua vida? P59 - Alguma vez em sua vida, você já tomou medicação psiquiátrica, mesmo que tenha sido por pouco tempo? P60 - Quanto interfere na sua vida ou no contexto acadêmico, a adaptação a novas situações: cidade, moradia, separação da família, entre outras?	P61 - Quanto interfere na sua vida ou no contexto acadêmico, os relacionamentos familiares? P62 - Quanto interfere na sua vida ou no contexto acadêmico, as relações amorosas/conjugais? P63 - Quanto interfere na sua vida ou no contexto acadêmico, a situação de violência física ou sexual? P64 - Quanto interfere na sua vida ou no contexto acadêmico, o assédio moral? P65 - Quanto interfere na sua vida ou no contexto acadêmico, os conflitos de valores/conflitos religiosos? P66 - Quanto interfere na sua vida ou no contexto acadêmico, a dificuldade de acesso a materiais e meios de estudo livros, computador e outros? P67 - Quanto interfere na sua vida ou no contexto acadêmico, a dificuldades financeiras interfere na sua vida ou no contexto acadêmico? P68 - Quanto interfere na sua vida ou no contexto acadêmico, as dificuldades de aprendizagem? P69 - Quanto interfere na sua vida ou no contexto acadêmico, a falta de disciplina/hábito de estudo? P70 - Quanto interfere na sua vida ou no contexto acadêmico, a carga horária excessiva de trabalho? P71 - Quanto interfere na sua vida ou no

	contexto acadêmico, a carga excessiva de trabalhos acadêmicos?
QUESTIONÁRIO SOCIOECONÔMICO – VOCÊ E O MUNDO DIGITAL	
P72 - Com que idade você acessou pela primeira vez um computador? P73 - Onde ocorreu esse primeiro acesso ao computador? P74 - Qual o domínio que você tem em relação ao microcomputador? P75 - Você considera importante o uso de recursos tecnológicos? P76 - Para quais fins você utiliza as tecnologias? P77 - Você gostaria que sua sala de aula utilizasse recursos tecnológicos para aprendizagem (celular, tablets)? P78 - Você possui: P79 - Você tem um celular? P80 - Seu celular é um smartphone?	P81 - Qual o tipo de tecnologia tem seu telefone? P82 - Qual a frequência você usa o celular? P83 - Quando você anda com o seu celular? P84 - Você costuma utilizar e baixar aplicativos móveis no seu celular? P85 - Possui algum tipo de acesso à internet? P86 - Onde você mais costuma acessar? P87 - Quanta vezes na semana? P88 - Quantas horas por dia? P89 - Entre as opções abaixo, marque o principal motivo que levam você a utilizar a internet:
QUESTIONÁRIO SOCIOECONÔMICO – VOCÊ E O SEU FUTURO	
P90 - Julgue o grau de motivação que o levaram a escolher este Campus P91 - O que você pretende fazer logo após se formar?	P92 - Depois de formado você pretende trabalhar. Imagina-se: P93 - Depois de formado você pretende estudar. Imagina-se: P94 - A escola tem contribuido para o seu projeto de vida?
QUESTIONÁRIO SOCIOECONÔMICO – A ESCOLHA DE BENEFÍCIOS	
P95 - Você já buscou atendimento no serviço social? P96 - Você já foi beneficiário da assistência estudantil? P97 - Qual/Quais Benefício(s)? P98 - Em qual/quais anos você recebeu esse(s) Benefício(s)? P99 - Você necessita receber Benefício Alimentação?	P100 - Você necessita receber Benefício Material Didático? P101 - Você necessita receber Benefício Transporte? P102 - Você necessita receber Benefício Internato? P103 - Você necessita receber Benefício Creche? P104 - Você necessita receber Benefício Moradia?

Fonte: Elaborado pelo Autor (2020)

APÊNDICE B – TELAS DA FERRAMENTA WEKA COM OS RESULTADOS DA APLICAÇÃO DA SELEÇÃO DE ATRIBUTOS

Figura B.1 – Resultado da Aplicação do Algoritmo *CFSSubsetEval* com Método de Busca *BestFirst* na Ferramenta WEKA

21:40:24 - BestFirst + CfsSubsetEval

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     REGISTRO_GERAL_CONSOLIDADO_2016-2019_SOCIOASSISTENCIAL_IFAM_CMZL - PADRONIZADO-weka.filters.supervised.instance.SMOTE-C0-K1-P100.0-S1
Instances:    3075
Attributes:   12
              GENERO
              TEM_FILHOS
              MORA_COM_QUEM
              TIPO_MORADIA
              MEIO_TRANSPORTE
              ALUNO_TRABALHA
              RENDA_MENSAL_FAMILIAR
              QUANT_DEPENDE_RENDA
              PARTICIPA_RENDA
              ESCOLARIDADE_PAI
              ESCOLARIDADE_MAE
              SITUACAO
Evaluation mode:    10-fold cross-validation


=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

number of folds (%)  attribute
          5( 50 %)    1 GENERO
         10(100 %)    2 TEM_FILHOS
         10(100 %)    3 MORA_COM_QUEM
          7( 70 %)    4 TIPO_MORADIA
         10(100 %)    5 MEIO_TRANSPORTE
          0(  0 %)    6 ALUNO_TRABALHA
         10(100 %)    7 RENDA_MENSAL_FAMILIAR
         10(100 %)    8 QUANT_DEPENDE_RENDA
          9( 90 %)    9 PARTICIPA_RENDA
         10(100 %)   10 ESCOLARIDADE_PAI
         10(100 %)   11 ESCOLARIDADE_MAE
```

Fonte: Elaborado pelo Autor (2020)

Figura B.2 – Resultado da Aplicação do Algoritmo *CFSSubsetEval* com Método de Busca *GeneticSearch* na Ferramenta WEKA

21:56:10 - GeneticSearch + CfsSubsetEval

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval -P 1 -E 1
Search:       weka.attributeSelection.GeneticSearch -Z 20 -G 20 -C 0.6 -M 0.033 -R 20 -S 1
Relation:     REGISTRO_GERAL_CONSOLIDADO_2016-2019_SOCIOASSISTENCIAL_IFAM_CMZL - PADRONIZADO-weka.filters.supervised.instance.SMOTE-C0-K1-P100.0-S1
Instances:    3075
Attributes:   12
              GENERO
              TEM_FILHOS
              MORA_COM_QUEM
              TIPO_MORADIA
              MEIO_TRANSPORTE
              ALUNO_TRABALHA
              RENDA_MENSAL_FAMILIAR
              QUANT_DEPENDE_RENDA
              PARTICIPA_RENDA
              ESCOLARIDADE_PAI
              ESCOLARIDADE_MAE
              SITUACAO
Evaluation mode:    10-fold cross-validation


=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

number of folds (%)  attribute
          4( 40 %)    1 GENERO
         10(100 %)    2 TEM_FILHOS
         10(100 %)    3 MORA_COM_QUEM
          7( 70 %)    4 TIPO_MORADIA
         10(100 %)    5 MEIO_TRANSPORTE
          0(  0 %)    6 ALUNO_TRABALHA
         10(100 %)    7 RENDA_MENSAL_FAMILIAR
         10(100 %)    8 QUANT_DEPENDE_RENDA
          9( 90 %)    9 PARTICIPA_RENDA
         10(100 %)   10 ESCOLARIDADE_PAI
         10(100 %)   11 ESCOLARIDADE_MAE
```

Fonte: Elaborado pelo Autor (2020)

ANEXO A – QUESTIONÁRIO SÓCIOECONÔMICO GERADO PELA FERRAMENTA

Figura A.1 – Exemplo do Questionário Socioeconômico Respondido pelo Aluno Gerado pela Ferramenta

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA AMAZONAS

QUESTIONARIO SÓCIOECONÔMICO 2019

ESTUDANTE: [redacted]

CODIGO/PERGUNTA/RESPOTA

P01 - Como você se considera: **PARDA**
P03 - Qual é a sua religião? **SEM RELIGIÃO**
P04 - Seu estado civil? **SOLTEIRO(A)**
P05 - Você tem filhos? **NÃO TENHO**
P09 - Qual o estado civil de seus pais? **SEPARADOS**
P10 - Onde você MORAVA antes de entrar no IFAM-ZL? **EM MANAUS**
P11 - Qual sua situação atual de moradia? **COM A MÃE E IRMÃOS**
P12 - Tipo de Moradia de sua Família: **PRÓPRIA**
P13 - Qual o principal meio de transporte que você utiliza para chegar ao IFAM - ZONA LESTE? **TRANSPORTE COLETIVO**
P14 - Na casa de sua família tem? **BANHEIRO, GELADEIRA, TV E LAVADORA DE ROUPA**
P15 - Você trabalha? **NÃO**
P17 - Quantas pessoas moram em sua casa? **CINCO (5)**
P18 - Quem é o a principal mantenedora de sua família ? a pessoa que mais contribui na renda: **IRMÃO/IRMÃ**
P19 - Qual a renda mensal do seu grupo familiar? **ATÉ UM E MEIO SALÁRIO MINIMO - (1.497,00)**
P20 - Quantas pessoas, incluindo você, vivem da renda mensal do seu grupo familiar informada acima? **CINCO (5)**
P21 - Qual a sua participação na vida econômica do seu grupo familiar? **SOU SUSTENTADO PELA FAMÍLIA OU POR OUTRAS PESSOAS**
P22 - Qual é a escolaridade de seu pai? Ou da pessoa que ao criou como pai: **ENSINO MÉDIO ANTIGO 2O GRAU ? COMPLETO**
P23 - Qual é a escolaridade de sua mãe? Ou da pessoa que oa criou como mãe: **ENSINO MÉDIO ANTIGO 2O GRAU ? COMPLETO**
P24 - Em que tipo de escola você cursou o Ensino Fundamental? **SOMENTE ESCOLA PÚBLICA**
P26 - Qual o tipo de Ensino Médio você cursou? **ENSINO MÉDIO PADRÃO**
P27 - Em que tipo de escola você cursou o Ensino Médio? **SOMENTE ESCOLA PÚBLICA**
P29 - Você frequentou cursinho pré-vestibular durante pelo menos seis meses? **NÃO**
P30 - Quantas vezes prestou vestibular? **DUAS VEZES**
P31 - De que forma você entrou nesse curso? **PROCESSO SELETIVO INSTITUCIONAL AMPLA CONCORRÊNCIA**
P32 - Em que turno você está frequentando a maior parte das disciplinas? **MATUTINO**
P33 - Qual o seu domínio da língua estrangeira Inglês? **REGULAR**
P34 - Qual o seu domínio da língua estrangeira Francês? **REGULAR**
P35 - Qual o seu domínio da língua estrangeira Espanhol? **REGULAR**
P36 - Você se identifica com a escola? **SIM**
Porque? **SIM**
P37 - Qual sua principal fonte de informação de acontecimentos atuais? **INTERNET**
P38 - Qual desses assuntos lhe desperta mais interesse **NOTÍCIAS LOCAIS E/OU NACIONAIS**
P39 - Qual é a média de livros que você lê em um ano? **CINCO**
P40 - Que tipo de livro você mais lê? **LITERATURA (ROMANCE/CONTOS/POEMAS)**
P41 - Com que frequência você participa de atividades extraclasse Artística-Culturais? **OCASIONALMENTE**
P42 - Com que frequência você participa de atividades extraclasse de Movimento Estudantil? **OCASIONALMENTE**
P43 - Com que frequência você participa de atividades extraclasse de Movimentos Ecológicos? **OCASIONALMENTE**
P44 - Com que frequência você participa de atividades extraclasse de Movimentos Religiosos? **OCASIONALMENTE**
P45 - Com que frequência você participa de atividades extraclasse de Movimentos Sociais? **OCASIONALMENTE**
P46 - Com que frequência você participa de atividades extraclasse Política – Partidárias? **OCASIONALMENTE**
P47 - Com que frequência você participa de atividades extraclasse de Sociedades Científicas? **OCASIONALMENTE**
P48 - Em geral, quando você precisa de atendimento médico você procura: **REDE PÚBLICA**
P49 - Sua última consulta médica ocorreu: **NO ÚLTIMO ANO**
P50 - Com relação a seus cuidados dentários, você: **VAI AO DENTISTA SE ESTIVER COM DOR OU OUTRO PROBLEMA**
P51 - Qual atividade física você mais pratica? **CAMINHADA/CORRIDA**
P52 - Com que frequência você pratica essa atividade? **OCASIONALMENTE**
P53 - Esta atividade é normalmente encarada por você como? **ATIVIDADE DE LAZER**
P55 - Você já teve alguma dificuldade significativa ou crise emocional nos últimos meses? **NÃO**
P56 - Você já procurou atendimento psico-pedagógico alguma vez em sua vida? **NÃO**
P57 - Você já procurou atendimento psicológico alguma vez em sua vida? **NÃO**
P58 - Você já procurou atendimento psiquiátrico alguma vez em sua vida? **NÃO**
P59 - Alguma vez em sua vida, você já tomou medicação psiquiátrica, mesmo que tenha sido por pouco tempo? **NÃO, NUNCA TOMEI**
P60 - Quanto interfere na sua vida ou no contexto acadêmico, a adaptação a novas situações: cidade, moradia, separação da família, entre outras? **POUCO**
P61 - Quanto interfere na sua vida ou no contexto acadêmico, os relacionamentos familiares? **POUCO**
P62 - Quanto interfere na sua vida ou no contexto acadêmico, as relações amorosas/conjugais? **POUCO**
P63 - Quanto interfere na sua vida ou no contexto acadêmico, a situação de violência física ou sexual? **POUCO**
P64 - Quanto interfere na sua vida ou no contexto acadêmico, o assédio moral? **POUCO**
P65 - Quanto interfere na sua vida ou no contexto acadêmico, os conflitos de valores/conflitos religiosos? **POUCO**
P66 - Quanto interfere na sua vida ou no contexto acadêmico, a dificuldade de acesso a materiais e meios de estudo livros, computador e outros? **MUITO**
P67 - Quanto interfere na sua vida ou no contexto acadêmico, a dificuldades financeiras interfere na sua vida ou no contexto acadêmico? **MUITO**
P68 - Quanto interfere na sua vida ou no contexto acadêmico, as dificuldades de aprendizagem? **MUITO**
P69 - Quanto interfere na sua vida ou no contexto acadêmico, a falta de disciplina/hábito de estudo? **POUCO**
P70 - Quanto interfere na sua vida ou no contexto acadêmico, a carga horária excessiva de trabalho? **MUITO**
P71 - Quanto interfere na sua vida ou no contexto acadêmico, a carga excessiva de trabalhos acadêmicos? **MUITO**
P72 - Com que idade você acessou pela primeira vez um computador? **14**
P73 - Onde ocorreu esse primeiro acesso ao computador? **LAN HOUSE**
P74 - Qual o domínio que você tem em relação ao microcomputador? **TEM EXPERIÊNCIA**
P75 - Você considera importante o uso de recursos tecnológicos? **IMPORTANTE**
P76 - Para quais fins você MAIS utiliza as tecnologias? **REDES SOCIAIS E COMUNICAÇÃO**
P77 - Você gostaria que sua sala de aula utilizasse recursos tecnológicos para aprendizagem (celular, tablets)? **SIM**
P78 - Você possui: **SOMENTE CELULAR**
P79 - Você tem um celular? **SIM**
P80 - Seu celular é um smartphone? **SIM**
P81 - Qual o tipo de tecnologia tem seu telefone? **ANDROID**
P82 - Qual a frequência você usa o celular? **DE 5 A 7 HORAS**
P83 - Quando você anda com o seu celular? **SEMPRE**
P84 - Você costuma utilizar e baixar aplicativos móveis no seu celular? **NÃO**
P85 - Possui algum tipo de acesso à internet? **SIM**

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA AMAZONAS

QUESTIONARIO SÓCIOECONÔMICO 2019

P86 - Onde você mais costuma acessar? **NA SUA PRÓPRIA CASA**
P87 - Quanta vezes na semana? **TODOS OS DIAS**
P88 - Quantas horas por dia? **TRÊS OU MAIS**
P89 - Entre as opções abaixo, marque o principal motivo que levam você a utilizar a internet: **BATE-PAPO (WHATSAPP)**
P90 - Julgue o grau de motivação que o levaram a escolher este Campus: **FORMAÇÃO PROFISSIONAL VOLTADA PARA O MERCADO DE TRABALHO**
P91 - O que você pretende fazer logo após se formar? **TRABALHAR E CONTINUAR ESTUDANDO**
P92 - Depois de formado você pretende trabalhar. Imagina-se: **TRABALHANDO EM QUALQUER ÁREA QUE TIVER OPORTUNIDADE**
P93 - Depois de formado você pretende estudar. Imagina-se: **INICIANDO UM CURSO DE GRADUAÇÃO**
P94 - A escola tem contribuido para o seu projeto de vida? **SIM**
Porque? **PORQUE TEM ME AJUDADO COM A FORMAÇÃO PROFISSIONAL E PESSOAL.**
P95 - Você já buscou atendimento no serviço social? **NÃO**
P96 - Você já foi beneficiário da assistência estudantil? **NÃO**
P99 - Você necessita receber Benefício Alimentação? **SIM**
P100 - Você necessita receber Benefício Material Didático? **SIM**
P101 - Você necessita receber Benefício Transporte? **SIM**
P102 - Você necessita receber Benefício Internato? **NÃO**
P103 - Você necessita receber Benefício Creche? **NÃO**
P104 - Você necessita receber Benefício Moradia? **NÃO**
P105 - Escreva abaixo uma sugestão de Projeto ou Benefício que NÃO existe no CMZL: **BOLSAS DE ESTÁGIOS REMUNERADOS**

Eu LARISSA LESLER DO NASCIMENTO COSTA declaro para todos os fins que li e aceito todas as condições do Edital Nº 07/2019 e acima estão tabuladas as minhas respostas a esta pesquisa.

Assinatura ESTUDANTE: _____ **Assinatura RESPONSAVEL:** _____

Fonte: Ferramenta do Questionário Eletrônico da Assistência Estudantil do IFAM-CMZL

ANEXO B – LISTAGEM COM RESULTADO DA ASSISTÊNCIA ESTUDANTIL DO IFAM-CMZL

Figura B.1 – Listagem com o Nome dos Aprovados no Processo de Concessão da Assistência Estudantil do IFAM-CMZL

PUBLICAÇÃO DO RESULTADO FINAL DA ASSISTÊNCIA ESTUDANTIL - CMZL 2019 (EM: 26/04/2019)		
QT	NOME	RESULTADO
1	[redacted]	TRANSPORTE
2	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
3	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
4	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
5	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
6	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
7	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
8	[redacted]	TRANSPORTE
9	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
10	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
11	[redacted]	INDEFERIDO
12	[redacted]	INDEFERIDO
13	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
14	[redacted]	MORADIA E MATERIAL DIDÁTICO
15	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
16	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
17	[redacted]	TRANSPORTE
18	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
19	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
20	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
21	[redacted]	INDEFERIDO
22	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
23	[redacted]	INDEFERIDO
24	[redacted]	TRANSPORTE
25	[redacted]	INDEFERIDO
26	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
27	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
28	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
29	[redacted]	MÁXIMO COMPLEMENTAR
30	[redacted]	INDEFERIDO
31	[redacted]	INDEFERIDO
32	[redacted]	INDEFERIDO
33	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
34	[redacted]	INDEFERIDO
35	[redacted]	TRANSPORTE
36	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
37	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
38	[redacted]	INDEFERIDO
39	[redacted]	INDEFERIDO
40	[redacted]	MÁXIMO COMPLEMENTAR
41	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
42	[redacted]	INDEFERIDO
43	[redacted]	TRANSPORTE E MATERIAL DIDÁTICO
44	[redacted]	TRANSPORTE

Fonte: Campus Manaus Zona Leste – IFAM