Universidade Regional do Cariri
Programa de Mestrado Profissional em Matemática em Rede Nacional

Mineração de Dados Educacionais: um estudo sobre a Proficiência em Matemática no Ceará

Herlane Martins Araújo

Juazeiro do Norte
2022

Universidade Regional do Cariri
Programa de Mestrado Profissional em Matemática em Rede Nacional

Herlane Martins Araújo

Mineração de Dados Educacionais: um estudo sobre a Proficiência em Matemática no Ceará

Dissertação de Mestrado apresentada à Universidade Regional do Cariri-URCA como parte dos requisitos para a obtenção do título de Mestre em Matemática.

Orientadora: Prof.ª Dra. Kátia Pires do Nascimento Sacramento
Coorientador: Prof. Dr. Vinícius Pereira Sacramento

Juazeiro do Norte
2022

Catalogação na fonte
Cícero Antônio Gomes Silva – CRB-3 n° /1385

A663

Araújo, Herlane Martins.

Mineração de dados educacionais: Um estudo sobre a proficiência em matemática no Ceará / Herlane Martins Araújo – Juazeiro do Norte-Ce, 2022,

51 f.: il.;30cm.

Dissertação de mestrado - Universidade Regional do Cariri– URCA / Programa de mestrado profissional em matemática em rede nacional - PROFMAT

Orientadora: Profª. Dra. Katia Pires do Nascimento Sacramento

Coorientador: Prof.Dr. Vinicius Pereira Sacramento

1. KDD . 2. Mineração de dados educacionais 3. Árvore de decisões I. Título

CDD:510

Herlane Martins Araújo

Mineração de Dados Educacionais: um estudo sobre a Proficiência em Matemática no Ceará

Dissertação de Mestrado apresentada à Universidade Regional do Cariri-URCA como parte dos requisitos para a obtenção do título de Mestre em Matemática.

Dissertação aprovada. Juazeiro do Norte - CE, 25 de Janeiro de 2022.

Documento assinado digitalmente
gov.br KATIA PIRES NASCIMENTO DO SACRAMENTO
Data: 04/05/2022 16:20:39-0300
Verifique em https://verificador.iti.br

Orientadora
Prof.ª Dr.ª Kátia Pires do Nascimento Sacramento
Universidade Regional do Cariri - URCA

Documento assinado digitalmente
gov.br Vinicius Pereira co Sacramento
Data: 03/05/2022 09:30:48-0300
Verifique em https://verificador.iti.br

Coorientador
Prof. Dr. Vinícius Pereira do Sacramento
Universidade Federal do Cariri - UFCA

Paulo César Cavalcante de Oliveira

Prof. Dr. Paulo César Cavalcante de Oliveira
Universidade Regional do Cariri - URCA

Documento assinado digitalmente
gov.br ROSILDA BENICIO DE SOUZA
Data: 19/04/2022 12:11:49-0300
Verifique em https://verificador.iti.br

Prof.ª Dr.ª Rosilda Benício de Souza
Universidade Federal do Cariri - UFCA

Abstract

This dissertation was a search for *Knowledge Discovery* through the *Knowledge Discovery in Databases - KDD* process and Data Mining, applied to the integration of the database from different sources: the Educational Indicators (EI) formulated by the INEP and data on Mathematics Proficiency in Ceará's municipalities as measured by SPAECE/2019. Through the KDD process, this investigation sought to develop a *Knowledge Model* based on data mining, which can classify and predict the performance of Ceará municipalities in the SPAECE Mathematics Assessment based on the context established by the Educational Indicators. The result was the construction of a Decision Tree model with an accuracy of 85.86%, structured and which points out the relevant EI for the investigated problem.

Keywords: KDD, Educational Data Mining, Decision Tree, SPAECE, Mathematics Proficiency.

Resumo

Esta dissertação foi uma busca de *Descoberta de Conhecimento* através do processo *Knowledge Discovery in Databases - KDD* e da Mineração de Dados, aplicados à integração da base de dados de diferentes fontes: os Indicadores Educacionais-IE formulados pelo INEP e os dados da Proficiência em Matemática dos municípios cearenses aferida pelo SPAECE/2019. Através do processo KDD, esta investigação buscou desenvolver um *Modelo de Conhecimento* baseado na mineração de dados, que possa classificar e prever o desempenho dos municípios cearenses na Avaliação de Matemática do SPAECE com base no contexto estabelecido pelos Indicadores Educacionais. O resultado foi a construção de um modelo de Árvore de Decisões de acurácia de 85,86%, estruturada e que aponta os IE relevantes para o problema investigado.

Palavras-chave: KDD, Mineração de Dados Educacionais, Árvore de decisões, SPAECE, Proficiência em Matemática.

Prefácio

Esta dissertação de mestrado foi submetida à Universidade Regional do Cariri - URCA como requisito parcial para obtenção do grau de Mestre em Matemática.

A dissertação foi desenvolvida no Programa de Mestrado Profissional em Matemática em Rede Nacional - PROFMAT, tendo como orientadora a **Prof.ª Dra. Kátia Pires do Nascimento Sacramento**. O **Prof. Dr. Vinícius Pereira Sacramento** foi coorientador deste trabalho.

Agradecimentos

Agradeço a Deus pela saúde.

À minha mãe Edenia, pelo apoio incondicional.

À toda minha família, especialmente, Tia Edna, Vó Nedina, Paula e Samuel, grandes incentivadores dos meus estudos, e aqueles que fazem torcida por mim lá de cima, meu pai Paulo, Tio Manu e Vô Pequeno.

À minha companhia diária Séfora, pelo incentivo e pelo suporte nos momentos difíceis.

À Maiara e Aristéia, amigas de alma.

Agradeço imensamente à minha orientadora Prof.ª Katia Pires do Nascimento Sacramento por acreditar no meu trabalho, pela paciência, orientação e disposição em me atender.

Ao Prof. Vinícius Pereira Sacramento, pelo ensino e assistência.

Aos colegas Cléber, Robério, José, Risoleta, Tamires, Jeovane, Bárbara, Taty, Luiz, Genilson e Raimundo, pelos momentos maravilhosos em sala de aula.

Ao Prof. Guttenberg Sergistótanes, pela amizade, por ter me ensinado os primeiros passos na pesquisa acadêmica.

Aos Professores Mário de Assis Oliveira, Paulo Cesar Cavalcante de Oliveira, Francisco Valdemiro Braga, Flávio França Cruz e José Tiago Nogueira Cruz, pelo ensino e assistência.

A todos que fazem o PROFMAT/URCA.

"Para Vó Nedina, minha mãe Edenia e Tia Edna, com todo meu amor".

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Durante a década de 90 e nos anos 2000, o Brasil vivenciou um período de implementação de reformas educacionais. Este cenário instigou a necessidade de um monitoramento educacional frente às novidades e à descentralização político-institucional da Educação. O marco da educação brasileira, a Lei de Diretrizes e Bases da Educação Nacional (LDB) de 1996, determinou a elaboração do Plano Nacional de Educação (PNE), política de diretrizes e planejamento no âmbito educacional que exige atuação conjunta das três esferas de governo (Federal, Estadual e Municipal) em regime de colaboração.

O panorama que se formou necessitava de um sistema de monitoramento educacional para acompanhamento sistemático de um conjunto de indicadores educacionais e estatísticos em intervalos regulares de tempo, para fornecer, dessa forma, elementos importantes para o planejamento e execução de ações pró-melhoria da qualidade de ensino [13].

Desta forma, o Brasil constituiu um sistema de monitoramento educacional baseado em três componentes: o Censo Escolar, o Sistema de Avaliação de Educação Básica (SAEB) e o Índice de Desenvolvimento da Educação Básica (IDEB), em um misto de coleta de informações e aplicação de avaliações em larga escala. Realidade que repercutiu também no âmbito dos estados, de modo que desde 1992, o Ceará implementou o Sistema Permanente de Avaliação da Educação Básica do Ceará (SPAECE) como avaliação externa em larga escala.

A proposta de monitoramento da educação encontra suporte no uso de tecnologias da informação no contexto da consolidação da Era Digital no final do século XX. Essa realidade permitiu um grande levantamento de dados sobre o ambiente educacional, e gerou uma demanda de extração de informações úteis a partir de bancos de dados armazenados digitalmente.

A demanda de extração de informação a partir de um grande volume de dados não era apenas no cenário educacional, era uma realidade em diversas áreas, e fomentou o desenvolvimento do campo de pesquisa interdisciplinar para transformação dos dados em informação (conhecimento).

Estruturou-se a teoria de *Knowledge Discovery in Databases - KDD* referente ao processo de descoberta de conhecimento em um banco de dados, auxiliado por

ferramentas e técnicas automatizadas que facilitem a extração e interpretação de informações. Uma das etapas do processo KDD é a Mineração de Dados (MD), embora por vezes seja usada como sinônimo do processo todo, constitui um campo de pesquisa que ganhou espaço na investigação em base de dados educacionais e consolidou uma área de pesquisa à parte, chamada Mineração de Dados Educacionais (MDE) entendida como uma combinação das áreas de Computação, Educação e Estatística [21].

Há a disposição um grande volume de dados provenientes de interações e de registros contínuos de informações sobre os sistemas educacionais. Entendendo que a estrutura da Educação é complexa, além das avaliações externas em larga escala, outros aspectos devem estar envolvidos para aferição da qualidade da educação, como formação e valorização dos professores, estrutura e entorno das escolas, gestão educacional, fatores econômicos e de vulnerabilidade social dos alunos e aspectos socioemocionais advindos das relações entre os sujeitos da escola, dentre outros. E esta é a realidade que os Indicadores Educacionais buscam aferir, pois são capazes de agregar valor analítico e avaliativo às estatísticas levantadas [10] [21].

Com inspiração na aplicabilidade da MDE, esta pesquisa propôs-se a investigar a Proficiência em Matemática alcançada pelos municípios cearenses, a partir de dois bancos de dados: (a) o de Indicadores Educacionais fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e (b) o resultado (por município) da Avaliação de Matemática do SPAECE/2019 referente à 3ª série do Ensino Médio da rede pública de ensino do estado do Ceará.

Através do processo KDD, objetivou-se minerar dados dos referidos bancos, com o propósito de *descoberta de conhecimento*, através da execução da tarefa *preditiva* de classificação de dados. Deste modo, como objetivo geral, esta investigação buscou desenvolver um *Modelo de Conhecimento* baseado na mineração de dados, que possa classificar e prever o desempenho dos municípios cearenses na Avaliação de Matemática do SPAECE com base nos dados de indicadores educacionais.

No tocante a objetivos específicos, esta investigação propôs-se a (1) identificar os principais atributos do conjunto de dados, considerando suas relevâncias e níveis de influência para classificação e predição da proficiência em matemática dos municípios cearenses; (2) criar um fluxo de processamento automatizado que permita a reprodução do modelo para outros anos a partir das novas edições do SPAECE e (3) contribuir no estudo e desenvolvimento de modelos preditivos de aferição de desempenho educacional.

Esta dissertação de mestrado está organizada da seguinte forma:

O Capítulo 2 apresenta os objetivos desta investigação.

O Capítulo 3 apresenta a fundamentação teórica acerca do processo de *Knowledge Discovery in Databases*, do qual a Mineração de Dados é uma etapa, neste ínterim, discorre sobre trabalhos relacionados à área de Mineração de Dados Educacionais. Apresenta também a Base de Dados desta investigação: os Indicadores Educacionais/INEP e a avaliação externa em larga escala do SPAECE.

O Capítulo 4 destaca a metodologia do processo KDD, concentrando-se em explicar a tarefa de mineração de dados de **Classificação**, a técnica da **Árvore de**

decisões e a ferramenta computacional utilizada, o algoritmo J48 implementado pelo *software* WEKA.

O Capítulo 5 apresenta as três principais etapas do processo KDD aplicadas aos bancos de dados selecionados, compondo o estudo de caso que se deseja realizar.

O Capítulo 6 apresenta conclusões sobre os principais resultados obtidos na pesquisa e as contribuições desta dissertação.

Capítulo 2

Objetivos

Esta seção apresenta os objetivos da pesquisa, com o propósito de esclarecer o que esta investigação buscou alcançar.

2.1 Objetivo Geral

Esta pesquisa buscou desenvolver um *Modelo de Conhecimento*, através da aplicação da Mineração de Dados, que possa classificar e prever a Proficiência em Matemática dos municípios cearenses na Avaliação do SPAECE, usando como banco de dados duas fontes distintas: (a) o de Indicadores Educacionais fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) e (b) o resultado por município da Avaliação de Matemática do SPAECE/2019 referente à 3ª série do Ensino Médio da rede pública de ensino do estado do Ceará.

2.2 Objetivos Específicos

- Identificar os principais atributos do conjunto de dados, considerando suas relevâncias e níveis de influência para classificação e predição da proficiência em matemática dos municípios cearenses;

- Criar um fluxo de processamento automatizado que permita a reprodução do modelo para outros anos a partir das novas edições do SPAECE;

- Contribuir para o estudo e desenvolvimento de modelos preditivos de aferição de desempenho educacional.

Capítulo 3

Revisão Bibliográfica

Este capítulo apresenta o embasamento teórico para a realização da descoberta do conhecimento através da Mineração de Dados. Apresenta também, a base de dados da pesquisa que são os Indicadores Educacionais fornecidos pelo INEP e a proficiência em Matemática na avaliação do SPAECE.

3.1 Knowledge Discovery in Databases - KDD

Em 2020 foram gerados 40 trilhões de gigabytes de dados em todo mundo, decorrentes do uso massivo de tecnologias da informação e da expansão do uso da internet, características da Era Digital consolidada no final do século XX. Neste cenário, começou-se a discutir como extrair informações úteis para a sociedade a partir dos bancos de dados armazenados digitalmente.

Em 1989 foi formalizado o termo *Knowledge Discovery in Databases - KDD*, como referência ao amplo conceito de procurar conhecimento a partir de bases de dados. Discutia-se sobre ferramentas e teorias que pudessem auxiliar os seres humanos na extração do conhecimento do volume de dados digitais em crescimento exponencial [12], [8].

Em um mundo que gera dados tão rápido, o método manual de análise e interpretação de informações é lento e em razão de ser realizado por analistas familiarizados com os dados, é caro e subjetivo [8]. Deste modo, buscava-se uma forma de automatizar este trabalho, mesmo que parcialmente.

Goldschmidt e Passos [12], indicam que a ideia de descobrir conhecimento é multidisciplinar e, originam-se de diversas áreas como Estatística, Inteligência Computacional e Aprendizado de Máquina, Reconhecimento de Padrões e Banco de Dados. De tal modo, o interesse no conhecimento que as análises de dados podem gerar, é científico e econômico. É um conhecimento que se conecta com o mundo real através da astronomia, marketing, finanças, saúde, varejo, geologia, educação, etc.

O artigo publicado em 1996 por Fayyad, Shapiro e Smyth é esclarecedor sobre a utilidade do KDD, definindo-o como um campo "preocupado com o desenvolvimento de métodos e técnicas para dar sentido aos dados", e em uma definição

mais precisa apontam que *"KDD é um processo, de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados"* [8]. Para uma melhor compreensão do processo KDD, e para referenciais futuros, é importante esmiuçar as características relatadas a partir de Goldschmidt e Passos [12]:

- Não trivial: se refere à complexidade normalmente presente na execução de processos de KDD;
- Iterativo: repetições do processo de KDD na busca de resultados satisfatórios por meio de refinamentos sucessivos;
- Dados: itens elementares, captados e armazenados por recursos da Tecnologia da Informação;
- Conhecimento útil: a formulação que pode envolver e relacionar dados e informações;
- Padrão válido: conhecimento verdadeiro e adequado ao contexto da aplicação de KDD;
- Padrão novo: deve acrescentar novos conhecimentos aos conhecimentos existentes no contexto da aplicação de KDD.

Dado que o KDD é um processo que busca extrair conhecimento a partir de um banco de dados, como etapa deste processo está "a aplicação de métodos específicos de mineração de dados para descoberta e extração de padrões".

> "Embora muitos usem mineração de dados como sinônimo de KDD, na primeira conferência internacional sobre KDD, realizada na cidade de Montreal, Canadá, em 1995, foi proposto que a terminologia descoberta de conhecimentos em bases de dados se referisse a todo o processo de extração de conhecimentos a partir de dados. Foi proposto também que a terminologia mineração de dados fosse empregada exclusivamente para a etapa de descoberta do processo de KDD" [12].

A Mineração de Dados - MD é uma etapa dentro do processo de KDD, na organização de Goldschmidt e Passos [12] na Figura 3.1, é possível observar que antes da MD, os dados passam por um pré-processamento que compreende as funções de captação, organização e ao tratamento dos dados e após serem minerados, procede-se ao tratamento do conhecimento obtido. Na Figura 3.1 também observa-se a participação humana em todo processo, destacando-se a característica *interativa* do KDD, que controla os recursos computacionais e trabalha na análise e interpretação dos resultados [12].

O alcance de extração de padrões pelo processo KDD pode ser indicado pela expressão *Modelo de Conhecimento* que aponta qualquer abstração de conhecimento,

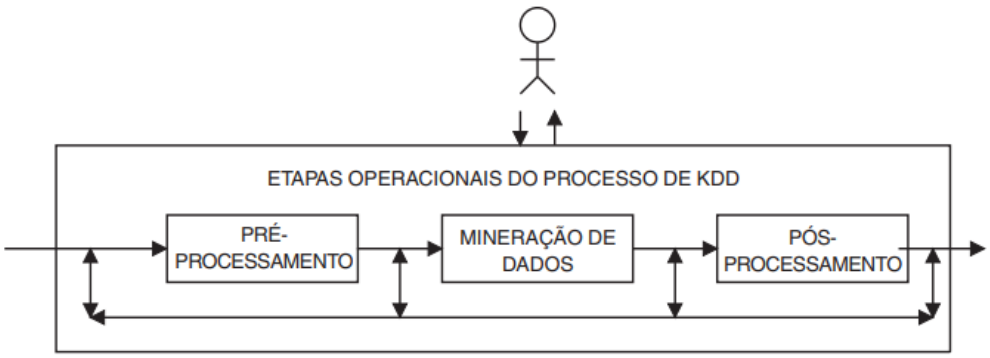

Figura 3.1: Etapas operacionais do Processo KDD
Fonte: Goldschmidt e Passos (2015)

expressa em alguma linguagem, que descreva algum conjunto de dados. E é com base nesse modelo que se pode avaliar o cumprimento das expectativas em relação ao objetivo de aplicação [12].

Os *objetivos da aplicação*, por sua vez, compreendem as características esperadas do modelo de conhecimento a ser produzido ao final do processo. Tais objetivos retratam, portanto, restrições e expectativas dos especialistas no domínio da aplicação acerca do modelo de conhecimento a ser gerado.

O *problema* a ser submetido ao processo de KDD envolve três componentes: o conjunto de dados, o especialista do domínio da aplicação e objetivos da aplicação [12].

3.1.1 Pré-processamento de dados

A etapa operacional de Pré-processamento engloba a captação, organização e o tratamento dos dados, preparando-os para serem minerados [12]. São realizadas as funções:

- **Seleção de Dados**: essa função envolve a identificação dos dados existentes que devem, de fato, ser considerados durante o processo de KDD [11];
- **Limpeza dos Dados**: consiste no tratamento feito sobre os dados selecionados de maneira a garantir a qualidade dos fatos representados. As informações que estiverem inconsistentes, ausentes ou erradas devem ser arrumadas para que não prejudiquem a qualidade dos modelos de conhecimento que serão retirados ao final do processo de KDD [12];
- **Codificação dos Dados**: nessa etapa os dados devem ser codificados de forma Numérica – Categórica, transformando os valores reais em categoria ou intervalos; ou Categórica – Numérica, que representa numericamente valores de atributos categóricos, para que possam ser utilizados como entrada para os algoritmos de Mineração de Dados [11];

- **Enriquecimento dos Dados**: consiste em conseguir de alguma forma mais informação que possa ser agregada aos registros existentes, enriquecendo os dados, para que estes forneçam mais informações para o processo de descoberta de conhecimento [12].

A Mineração de Dados será abordada na sequência, na seção 3.2, para dar destaque aos seus propósitos e execução.

3.1.2 Pós-processamento de dados

Esta é a etapa final do processo KDD, contempla a depuração e/ou síntese dos padrões descobertos. Se propõe a uma organização do conhecimento obtido, simplificação do Modelo de Conhecimento, com uma proposta de melhorar a visualização da descoberta através de gráficos, diagramas ou relatórios demonstrativos. O objetivo é basicamente facilitar interpretação e a avaliação do conhecimento adquirido. As operações a serem realizadas nesta etapa são: Simplificação do Modelo de Conhecimento, Transformações do Modelo de Conhecimento e Organização e Apresentação dos Resultados [12], [8].

3.2 Mineração de Dados

3.2.1 Definição e Histórico

Data Mining ou Mineração de Dados assemelha-se ao processo de extração de minerais preciosos a partir do trabalho em uma mina. A analogia corresponde na prática, ao exercício de minerar bancos de dados para extrair deles algo "precioso", no caso, o conhecimento que será útil para a sociedade. Neste exercício, o homem interage com a tecnologia, e tem como ferramenta de mineração, o uso de algoritmos adequados à obtenção do conhecimento, este, entendido como algo que permite uma tomada de decisão para a agregação de valor [12].

Goldschmidt e Passos [12] definem quatro gerações no histórico da MD, baseados no trabalho de Gregory Piatetsky-Shapiro [19]:

- Primeira geração: surge na década de 1980, baseada em ferramentas orientadas para a pesquisa com foco em tarefas únicas. Essas tarefas incluíam construir um classificador do tipo Indução de regra (árvore de decisão), Rede Neural, descoberta de clusters ou ainda a visualização de dados.

- Segunda geração: surge em 1995 com o desenvolvimento de ferramentas chamadas "suites", estas, eram dirigidas ao fato de que o processo de descoberta do conhecimento requer múltiplos tipos de análise dos dados, permitindo ao usuário realizar diversas tarefas de descoberta (geralmente classificação, clusterização e visualização) e suportavam transformação de dados. Tais ferramentas requerem conhecimento significativo da teoria estatística, exigindo o auxílio de especialistas em análise de dados.

- Terceira geração: nesta geração as interfaces são orientadas para o usuário e procuram esconder a complexidade da Mineração de Dados.

- Quarta geração: compreende o desenvolvimento e a aplicação de técnicas e ferramentas que auxiliem o homem na própria condução do complexo processo de KDD.

Percebe-se que na evolução da MD, surge uma preocupação em simplificar as ferramentas de mineração para que usuários, ainda que não detenham expertise, consigam manipular os ambientes de *software* ou até mesmo algoritmos isolados em busca de tratamento e análise de dados. Foi a evolução dessas ferramentas que permitiu a realização desta pesquisa, tornando possível a aproximação da realidade da MD com o problema a ser submetido ao processo de KDD.

Um conceito importante a ser apresentado, é a medida utilizada na avaliação da qualidade de uma regra, chamada de **acurácia**, também denominada confiança ou precisão da regra [9], [12]. Esta medida costuma ser calculada pelo *software* que processa os dados, fornecendo ao usuário um valor que represente a relevância dos resultados por ele obtidos, permitindo a fluidez da iteração e interação do processo KDD.

Como etapa do processo KDD, a MD corresponde à aplicação de algoritmos capazes de extrair conhecimentos a partir dos dados pré-processados [9]. Esta aplicação faz correspondência com o objetivo que se deseja atingir, e deste modo, remete à **tarefa** da mineração de dados, ou seja, a qual tipo de descoberta que se pretende realizar em uma base de dados. Estas tarefas serão abordadas na Subseção 3.2.2.

3.2.2 Tarefas da Mineração de Dados

As tarefas da MD são basicamente duas: (a) *predição* e (b) *descrição* dos dados. Uma tarefa "consiste na especificação do que se pretende buscar, ou que tipo de regularidade ou padrões interessa encontrar"[24]. Desta forma, "as funcionalidades da MD são usadas para especificar os tipos de informações a serem obtidas nas tarefas de mineração"[9].

A Figura 3.2 apresenta uma estrutura de classificação das tarefas da MD, baseada no trabalho de Han, Kamber e Pei. As tarefas (a) *preditivas* fazem induções a partir dos dados objetivando predições [9]:

- Classificação: consiste em descobrir uma função que mapeie um conjunto de registros em um conjunto de rótulos categóricos predefinidos, denominados classes. Uma vez descoberta, tal função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram [12];

- Regressão: compreende a busca por uma função que mapeie os registros de um banco de dados em valores reais. Esta tarefa é similar à tarefa de classificação, sendo restrita apenas a atributos numéricos [12].

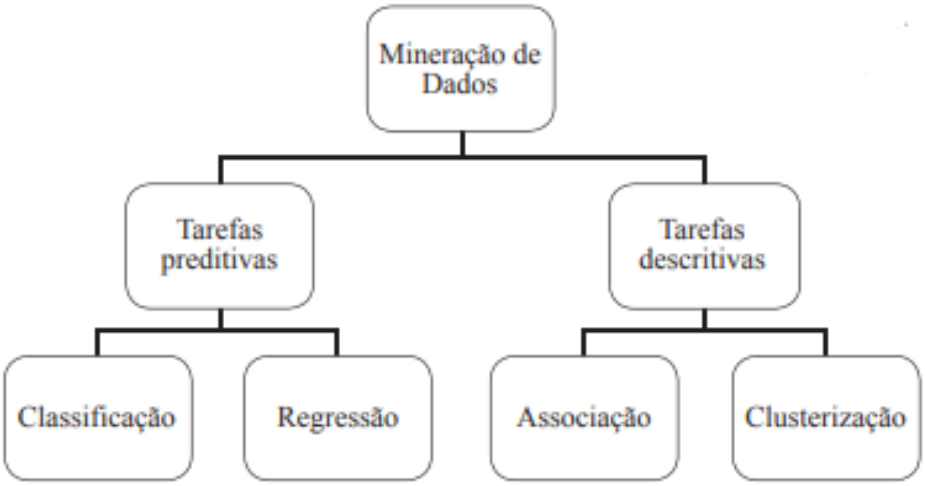

Figura 3.2: Tarefas da Mineração de Dados
Fonte: Vieira e Oliveira [24]

As tarefas (b) *descritivas* caracterizam as propriedades gerais dos dados [9], são elas:

- Associação: corresponde à descoberta de regras de associação que apresentam valores de atributos que ocorrem concomitantemente em uma base de dados [9];

- Clusterização: utilizada para separar os registros de uma base de dados em subconjuntos ou *clusters*, de tal forma que os elementos de um *clusters* compartilhem de propriedades comuns que os distingam de elementos em outros *clusters* [12].

3.3 Estudos sobre Mineração de Dados Educacionais

A MD quando utilizada no contexto educacional recebe a denominação Mineração de Dados Educacionais - MDE, e ganhou destaque a partir de 2008 quando foi realizada a primeira conferência internacional exclusiva para discussão do MDE, em Montreau no Canadá [23].

No Brasil, a MDE é difundida através de eventos e periódicos sobre Informática na Educação e tem como marco importante para o desenvolvimento de pesquisas na área, o artigo de Ryan Baker, Seiji Isotani e Adriana Carvalho, publicado em 2011, intitulado “Mineração de Dados Educacionais: Oportunidades para o Brasil”. A predição da performance dos alunos é uma das aplicações de MDE mais antigas e também mais utilizadas [21].

As pesquisas na área de MDE destacam o crescente número de dados sobre educação coletados pelos sistemas informatizados das escolas e universidades, pelos

Ambientes Virtuais de Aprendizagem - AVA, advindos da expansão dos cursos a distância e pelo uso das tecnologias de informação e comunicação - TIC's na Educação.

Deste modo, os dados oriundos de ambientes educacionais podem ser trabalhados com a finalidade de melhorar as relações de ensino-aprendizagem, o financiamento educacional, a administração escolar e até mesmo nortear as políticas públicas educacionais. Enumeram-se na sequência, leituras na área de MDE relevantes e inspiradoras para a execução desta pesquisa.

1. Nascimento *et al* utilizaram bases de dados educacionais fornecidas pelo INEP para aplicação de técnicas de mineração de dados com a finalidade de melhor explicar indicadores como a evasão e reprovação escolar no ensino fundamental [7];

2. Rogério Silva em sua dissertação apresentou uma solução de mineração de dados em um processo de KDD para predição e estimação do desempenho dos alunos do Ensino Médio dos Institutos Federais usando como base o Exame Nacional do Ensino Médio [23];

3. Bezerra *et al* analisaram a evasão escolar no último ano do ensino fundamental nas escolas públicas estaduais e municipais do estado de Pernambuco. Árvore de Decisão, Indução de Regras e Regressão Logística foram as técnicas para extração de conhecimento aplicadas visando identificar o perfil do aluno evasor e estimar a propensão à evasão [1];

4. Júnior *et al* basearam-se em dados educacionais fornecidos pelo INEP e trabalharam a mineração de dados educacionais para produzirem um estudo descritivo das variáveis explicativas do modelo utilizado para predizer a aprovação e a reprovação escolar [16];

5. Manhães *et al* alcançaram um bom resultado ao utilizarem técnicas de mineração de dados para prever o risco de evasão nos cursos de graduação [18].

3.4 Base de dados

3.4.1 Indicadores Educacionais

O Ministério da Educação, por meio do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) mantém em seu portal, desde 1995, dados dos Indicadores Educacionais - IE coletados a partir do Censo Escolar.

> "Os Indicadores Educacionais atribuem valor estatístico à qualidade do ensino, atendo-se não somente ao desempenho dos alunos, mas também ao contexto econômico e social em que as escolas estão inseridas. Eles consideram informações como o acesso, a permanência e a aprendizagem dos alunos. Pode-se obter esses dados em diferentes granularidades como nível nacional, regional ou nível das escolas"[7].

Para este estudo é utilizada a base de dados a nível de municípios. Os IE considerados para esta pesquisa são:

1. **Taxa de distorção idade-série**: tem como objetivo identificar a proporção de alunos com mais de 2 anos de atraso escolar;
2. **Taxa de rendimento**: o cálculo das taxas de rendimento (aprovação, reprovação e abandono) tem como referência as informações de rendimento e movimento dos alunos coletadas na segunda etapa do Censo Escolar;
3. **Média de horas aulas diárias**: expressa o tempo médio de permanência dos alunos na escola;
4. **Percentual de docentes com curso superior**: identifica o percentual de professores do Ensino Médio, em relação ao total de professores da etapa, que possuem o Ensino Superior completo;
5. **Indicador de nível socioeconômico**: busca situar o conjunto dos alunos atendidos por cada escola em um estrato, definido pela posse de bens domésticos, pela renda e pela contratação de serviços por parte da família dos alunos e pelo nível de escolaridade dos pais [25];
6. **Média de alunos por turma**: permite avaliar o tamanho médio das turmas para diferentes etapas de ensino;
7. **Adequação da formação docente**: tem como objetivo avaliar a adequação da formação docente à disciplina que leciona, segundo as orientações legais [25];
8. **Esforço docente**: objetiva medir o esforço empreendido pelos docentes para o exercício da profissão;
9. **Complexidade da gestão da escola**: o indicador classifica as escolas em níveis de 1 a 6 de acordo com sua complexidade de gestão, níveis elevados indicam maior complexidade. Considerou-se que complexidade de gestão está relacionada ao porte da escola, número de turnos de funcionamento, quantidade e complexidade de modalidades/etapas oferecidas [7];
10. **Indicador de regularidade docente**: avalia a regularidade do corpo docente com base na observação da permanência dos professores nas escolas nos últimos cinco anos [25];

Esse conjunto de indicadores é um instrumento de gestão, que embasa o planejamento e o monitoramento de práticas que buscam expressar a partir de um significado particular, um resultado, uma característica ou o desempenho de uma ação, de um processo ou de um serviço, ou política. Entende-se que os IE podem positivamente assumir uma função diagnóstica de amplitude social e subsidiar a formulação de políticas públicas porém, são limitados quanto ao propósito avaliar a qualidade da educação [25].

3.4.2 Sistema Permanente de Avaliação da Educação Básica do Ceará – SPAECE

Histórico e Políticas Educacionais

Na perspectiva de garantir o direito constitucional à Educação e atingir sua finalidade de pleno desenvolvimento do educando, a gestão brasileira das políticas educacionais buscou estudar e levantar dados do ambiente educacional.

Neste contexto é que a avaliação educacional se desenvolve no Brasil na década de 90, para ajudar a instituir uma política nacional de avaliação que ficou explicitada através da criação do Sistema de Avaliação da Educação Básica (SAEB), o Exame Nacional do Ensino Médio (ENEM) e o Exame Nacional de Curso (ENC), para o ensino superior [15].

O governo do Estado do Ceará financiou a criação do Sistema de Avaliação do Rendimento Escolar em 1992 com a finalidade de realizar uma avaliação experimental para detectar problemas de aprendizagem. Tal aplicação foi realizada de forma censitária com os alunos de todas as escolas da rede estadual do município de Fortaleza, com testes padronizados de Português e Matemática para a 4ª e a 8ª série do 1º grau. Fato que deu destaque ao Ceará como um dos estados pioneiros na criação de um sistema próprio de avaliação da Educação Básica [14], [15].

Em 1996, a publicação da Lei de Diretrizes e Bases da Educação - LDB nº 9394/96 destaca a avaliação educacional como uma área de múltiplas direções que alcança "os sistemas, os cursos, as instruções, os currículos, os programas de ensino, os professores, as dimensões cognitivas e não cognitivas buscando romper com seus próprios limites". Neste cenário, a avaliação do Sistema Permanente de Avaliação da Educação Básica do Ceará - SPAECE se consolidou no ano 2000, sendo oficialmente institucionalizado através da portaria nº 101/00 [15].

Atualmente, o SPAECE tem três focos: a Avaliação da Alfabetização – SPAECE-Alfa (aplicada ao 2º ano do ensino fundamental); a Avaliação do Ensino Fundamental (nos 5º e 9º anos e EJA); e a Avaliação do Ensino Médio (na 3ª série e EJA).

Ao interesse desta pesquisa, a Avaliação do Ensino Médio é realizada, anualmente, envolve todas as escolas da Rede Estadual de ensino e seus anexos, localizadas nos 184 municípios cearenses. Para os alunos da Educação de Jovens e Adultos, oferecida de forma presencial, é realizada também, uma avaliação para os 1º e 2º períodos do Ensino Médio [2].

O Ceará realiza a referida avaliação em parceria com o Centro de Políticas Públicas e Avaliação da Educação da Universidade Federal de Juiz de Fora (CAED/UFJF), tal instituição indica:

> "O conjunto de informações coletadas pelo SPAECE permite diagnosticar a qualidade da educação pública em todo o estado do Ceará, produzindo resultados por aluno, turma, escola, município, credes e estado. Ao mesmo tempo, os resultados têm servido de base para implementação de políticas públicas educacionais e de práticas pedagógicas

> inovadoras nas escolas estaduais e municipais. O SPAECE tornou-se um instrumento essencial na fomentação de debate público e na promoção de ações orientadas para a melhoria e execução da democratização do ensino, garantindo a todos igualdade de acesso e permanência na escola"[2].

Através do que informa o CAED/UFJF, percebe-se que os resultados das avaliações do SPAECE possibilitam a reformulação de políticas públicas voltadas para a melhoria educacional da rede pública de ensino do estado do Ceará [14].

Padrão de Desempenho e Níveis de Proficiência

No que se refere ao que é investigado na avaliação em larga escala do SPAECE, a matriz de referência da prova aplicada aos alunos do Ensino Médio, atualmente, se aproxima dos temas avaliados pelo ENEM. Deste modo, a matriz da prova de Matemática do SPAECE abrange 76 descritores agrupados em quatro domínios: espaço e forma; grandezas e medidas; números, operação e álgebra; tratamento da informação. A Figura 3.3 apresenta a matriz de referência de Matemática para a 3ª série do ensino médio.

Para aferição do desempenho dos alunos, o SPAECE adota a Teoria Clássica dos Testes (TCT) e a Teoria de Resposta ao Item (TRI). Os resultados gerais são descritos pela TCT, que consiste em quantificar o acerto do aluno no teste ou escore bruto, e os mais específicos seguem com a contribuição da TRI [14].

O resultado apresentado pela TRI é uma medida de proficiência que se utiliza do modelo logístico unidimensional de três parâmetros: discriminação, dificuldade e probabilidade de acerto ao acaso. Esse resultado é capaz de posicionar o desempenho do aluno em uma escala de proficiência [14].

A aferição da Proficiência em Matemática se baseia nos Padrões de Desempenho Estudantil, que são categorias definidas a partir de cortes numéricos que agrupam os níveis da Escala de Proficiência, com base nas metas educacionais estabelecidas pelo SPAECE. Esses cortes dão origem a quatro Padrões de Desempenho: até 250 pontos – Muito crítico; 250 a 300 – Crítico; 300 a 350 – Intermediário; e acima de 350 – Adequado, que se utilizam de uma gradação de cores para dar melhor visualização aos resultados, conforme figura 3.4.

De certo, para esta pesquisa é importante fazer convergir os valores dos padrões de desempenho e o nível de proficiência de modo a fornecer um diagnóstico pedagógico sobre a aprendizagem em Matemática. Essa interpretação pode ser compreendida como:

- **Padrão de desempenho - Muito crítico, Nível de proficiência - abaixo de 250**: Os alunos que apresentam este padrão de desempenho revelam ter desenvolvido competências e habilidades que se encontram muito aquém do que seria esperado para o período de escolarização em que se encontram. Portanto, necessitam de uma intervenção focalizada de modo a progredir com sucesso em seu processo de escolarização. Esses alunos são capazes, ao final do 3º ano do ensino médio, apenas, de identificar forma ampliada de uma figura simples

3ª SÉRIE DO ENSINO MÉDIO	
TEMA I. INTERAGINDO COM NÚMEROS E FUNÇÕES	
D16	Estabelecer relações entre representações fracionárias e decimais dos números racionais.
D19	Resolver problema envolvendo juros simples.
D20	Resolver problema envolvendo juros compostos.
D24	Fatorar e simplificar expressões algébricas.
D28	Reconhecer a representação algébrica ou gráfica da função polinomial de 1º grau.
D40	Relacionar as raízes de um polinômio com sua decomposição em fatores do 1º grau.
D42	Resolver situação problema envolvendo o cálculo da probabilidade de um evento.
TEMA II. CONVIVENDO COM A GEOMETRIA	
D49	Resolver problema envolvendo semelhança de figuras planas.
D50	Resolver situação problema aplicando o Teorema de Pitágoras ou as demais relações métricas no triângulo retângulo.
D51	Resolver problema usando as propriedades dos polígonos (soma dos ângulos internos, número de diagonais e cálculo do ângulo interno de polígonos regulares).
D52	Identificar planificações de alguns poliedros e/ou corpos redondos.
D53	Resolver situação problema envolvendo as razões trigonométricas no triângulo retângulo (seno, cosseno, tangente).
D54	Calcular a área de um triângulo pelas coordenadas de seus vértices.
D55	Determinar uma equação da reta a partir de dois pontos dados ou de um ponto e sua inclinação.
D56	Reconhecer, dentre as equações do 2°grau com duas incógnitas, as que representam circunferências.
D57	Identificar a localização de pontos no plano cartesiano.
D58	Interpretar geometricamente os coeficientes da equação de uma reta.
TEMA III. VIVENCIANDO AS MEDIDAS	
D64	Resolver problema utilizando as relações entre diferentes unidades de medidas, de capacidade e de volume.
D65	Calcular o perímetro de figuras planas numa situação problema.
D67	Resolver problema envolvendo o cálculo de área de figuras planas.
D71	Calcular a área da superfície total de prismas, pirâmides, cones, cilindros e esfera.
D72	Calcular o volume de prismas, pirâmides, cilindros e cones em situação-problema.
TEMA IV. TRATAMENTO DA INFORMAÇÃO	
D76	Associar informações apresentadas em listas e/ ou tabelas aos gráficos que as representam, e vice-versa.
D78	Resolver problema envolvendo medidas de tendência central: média, moda ou mediana.

Figura 3.3: Matriz de Referência de Matemática - SPAECE
Fonte: CAED/UFJF [2]

em malha quadriculada; resolver problemas de subtração de números racionais escritos na forma decimal com o mesmo número de casas decimais, localizar informações em gráficos de colunas duplas, resolver problemas envolvendo conversão de kg para g ou relacionando diferentes unidades de medidas de tempo

PADRÃO DE DESEMPENHO	MUITO CRÍTICO	CRÍTICO	INTERMEDIÁRIO	ADEQUADO
NÍVEL DE PROFICIÊNCIA	ABAIXO DE 250	250 A 300	300-350	ACIMA DE 350

Figura 3.4: Nível de proficiência x Padrão de desempenho

(mês/trimestre/ano); resolver problemas que envolvam subtração de números decimais com o mesmo número de casas; localizar dados em tabelas de múltipla entrada.

- **Padrão de desempenho - Crítico, Nível de proficiência - 250 – 300**: Os alunos que apresentam esse padrão de desempenho demonstram ter começado um processo de sistematização e domínio das habilidades consideradas essenciais ao período de escolarização em que se encontram. Contudo, também para esse grupo de alunos, é importante o investimento de esforços para que se possam desenvolver habilidades que envolvam resoluções de problemas com um grau de complexidade um pouco maior. Além das habilidades apresentadas no padrão de desempenho anterior, esses alunos revelam, ao final do 3º ano do ensino médio, ser capazes de localizar números inteiros e números racionais, positivos e negativos, na forma decimal, na reta numérica, reconhecer e aplicar em situações simples o conceito de porcentagem, utilizar o conceito de progressão aritmética (PA); calcular probabilidades simples; identificar a fração como parte de um todo, sem apoio da figura; calcular o valor numérico de uma expressão algébrica, incluindo potenciação.

- **Padrão de desempenho - Intermediário, Nível de proficiência - 300 – 350**: Os alunos que apresentam este padrão de desempenho demonstram ter ampliado o leque de habilidades tanto no que diz respeito à quantidade quanto no que se refere à complexidade dessas habilidades, as quais exigem um maior refinamento do processo cognitivo nelas envolvido. Além das habilidades apresentadas no padrão de desempenho anterior, esses alunos, ao final do 3º ano do ensino médio, são capazes de, por exemplo, calcular o volume dos sólidos a partir da medida de suas arestas; solucionar problemas envolvendo propriedades dos polígonos regulares inscritos (hexágono), para calcular seu perímetro; reconhecer o significado da palavra perímetro; identificar crescimento e decrescimento em um gráfico de função, calcular o resultado de uma divisão em partes proporcionais e identificar o termo seguinte em uma sequência dada (PG); calcular expressões numéricas com números inteiros e decimais positivos e negativos; ler informações fornecidas em gráficos envolvendo regiões do plano cartesiano.

- **Padrão de desempenho - Adequado, Nível de proficiência - Acima de 350**: Os alunos que apresentam este padrão de desempenho revelam ser capazes de

realizar tarefas que exigem um raciocínio algébrico e geométrico mais avançado para resoluções de problemas, além de desenvolverem habilidades que superam aquelas esperadas para o período de escolaridade em que se encontram. Além das habilidades apresentadas no padrão de desempenho anterior, esses alunos revelam ser capazes, ao final do 3º ano do ensino médio, de calcular o volume de um paralelepípedo; efetuar cálculos de divisão com números racionais (forma fracionária e decimal simultaneamente); resolver problemas usando sistemas de equação do primeiro grau ou que recaem em equação do segundo grau; resolver problemas de contagem envolvendo permutação; calcular a probabilidade de um evento, usando princípio multiplicativo para eventos independentes; resolver equações exponenciais simples; resolver problemas envolvendo relações métricas no triângulo retângulo; resolver problemas simples envolvendo funções exponenciais; utilizar a definição de PA e PG para resolver um problema e calcular a área total de uma pirâmide regular.

A partir do conhecimento do processo KDD e do banco de dados, apresenta-se no Capítulo 4 a metodologia de desenvolvimento da pesquisa.

Capítulo 4

Metodologia, Técnica e Ferramenta

Esta é uma pesquisa de *Descoberta de Conhecimento*, condizente com a proposta de busca efetiva por conhecimentos a partir da abstração dos dados existentes. A metodologia se baseou nas etapas operacionais do processo KDD, descritas na Figura 3.1, no Capítulo 3, quais sejam: **Pré-processamento**, **Mineração** e **Pós-processamento dos dados**. A descrição das etapas aplicadas ao banco de dados selecionado estão descritas no Capítulo 5.

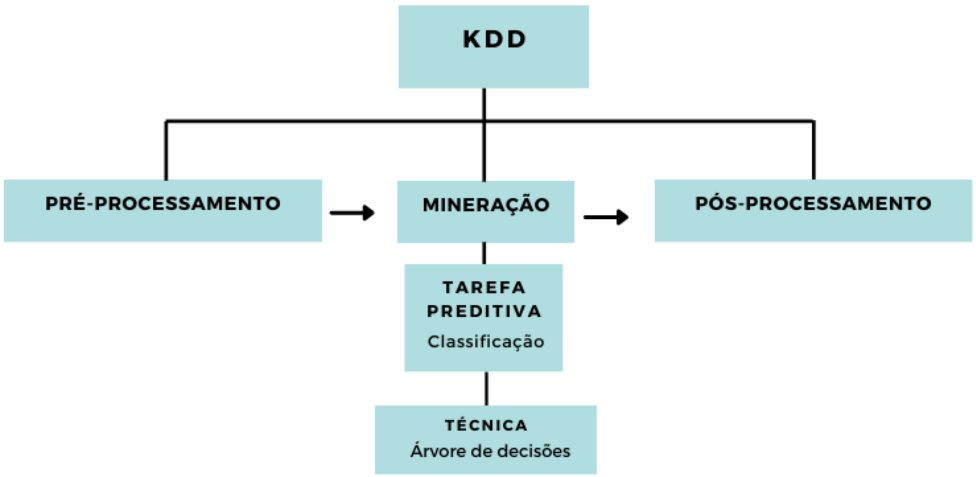

Figura 4.1: Metodologia do processo KDD
Fonte: Elaboração própria

O guia prático de mineração de dados elaborado por Goldschmidt e Passos, indica três componentes envolvidos no processo KDD [12], descritos a seguir no perfil desta pesquisa:

- O problema a ser submetido ao processo de KDD: o conjunto de dados utilizados nesta pesquisa é formado pelos Indicadores Educacionais do INEP e pela Proficiência em Matemática levantada pelo SPAECE, deste modo, definiu-se

como objetivo da aplicação do KDD, a produção de um modelo preditivo para classificar a relação entre os IE's e o desempenho em matemática;

- Os recursos disponíveis para solução do problema mencionado: a **ferramenta** de KDD empregada como recurso computacional para análise de dados foi o *software* livre WEKA.

- Os resultados obtidos a partir da aplicação dos recursos no problema: o modelo de conhecimento desenvolvido nesta pesquisa e o histórico de ações realizadas estão descritas no Capítulo 5.

A Figura 4.1 esquematiza a sequência envolvida no processo KDD, e apresenta as escolhas realizadas na etapa de MD, de modo que para realizar a **tarefa preditiva de Classificação**, optou-se pela **técnica** de Árvore de decisões.

4.1 Tarefa de Classificação

A **Classificação** é uma das tarefas de KDD mais importantes e mais populares, na definição de Ferrari e Castro:

> "Classificar um objeto significa atribuir a ele um rótulo, chamado classe, de acordo com a categoria à qual ele pertence. Para que isso seja possível, um algoritmo de classificação é usado na construção de um modelo de classificação, também chamado de classificador"[9].

Nesta pesquisa, os objetos são os dados de indicadores educacionais dos municípios cearenses e a proposta é classificá-los de acordo com a proficiência em matemática obtida por eles na avaliação do SPAECE, desta forma, as classes correspondem aos níveis de desempenho em matemática: muito crítico, crítico, intermediário e adequado. Esclarece Goldschmidt e Passos, como se dá a relação entre objeto e classe, e como a tarefa de classificação tem um comportamento preditivo:

> "[...]essa tarefa pode ser compreendida como a busca por uma função que permita associar corretamente cada registro X_i de um banco de dados a um único rótulo categórico, Y_j denominado classe. Uma vez identificada, essa função pode ser aplicada a novos registros de forma a prever a classe em que tais registros se enquadram"[12].

Nesta investigação, a descoberta do conhecimento envolvendo a tarefa de classificação buscou a função que relaciona o conjunto de diversos dados de indicadores educacionais com o conjunto que contém os atributos categóricos de desempenho em matemática. Uma vez construída, essa função se mostra como um produto relevante que permite prever a proficiência em matemática dos municípios a partir do mapeamento dos indicadores educacionais.

4.2 Árvore de Decisões

Uma das principais **técnicas** de MD são as que se baseiam na construção de Árvores de Decisões (AD). A construção de uma AD segue alguma abordagem recursiva de particionamento da base de dados [12] e mais:

> "[...] funciona como um fluxograma em forma de árvore, onde cada nó (não folha) indica um teste feito sobre um valor (...). As ligações entre os nós representam os valores possíveis do teste do nó superior, e as folhas indicam a classe (categoria) a qual o registro pertence. Após a árvore de decisão montada, para classificarmos um novo registro, basta seguir o fluxo na árvore (mediante os testes nos nós não-folhas) começando no nó raiz até chegar a uma folha. Pela estrutura que formam, as árvores de decisões podem ser convertidas em Regras de Classificação. O sucesso das árvores de decisão, deve-se ao fato de ser uma técnica extremamente simples, não necessita de parâmetros de configuração e geralmente tem um bom grau de assertividade"[3].

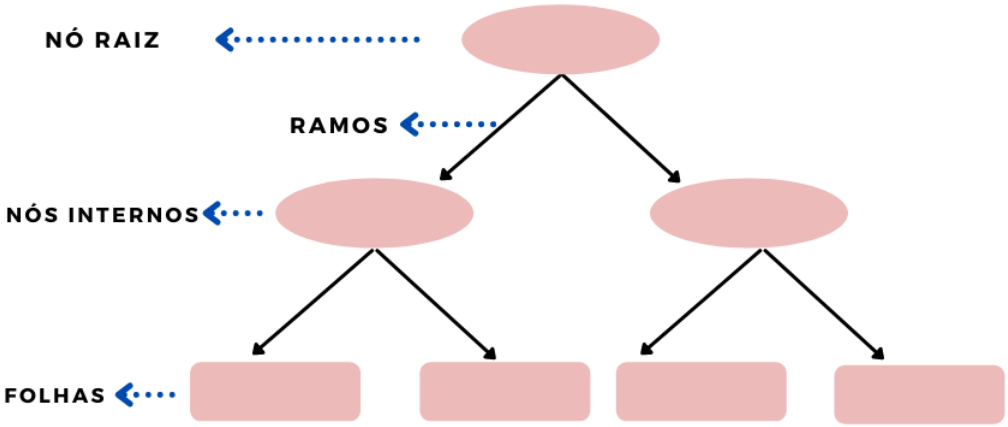

Figura 4.2: Elementos de uma Árvore de Decisões
Fonte: Elaboração própria

Através da Figura 4.2 é possível observar a estrutura de uma AD genérica, que ilustra as terminologias citadas acima, que têm os seguintes conceitos:

- *Nó raiz*: é o atributo que melhor divide o conjunto de dados, é aquele que confere o maior ganho de informação ao modelo [9], [22];
- *Nós internos*: são representadas pelas condições lógicas (SE/ENTÃO) que determinam o caminho dentro da árvore [22];
- *Nós folhas*: estão na parte inferior da árvore, são os rótulos de classe [9], [22].

A AD usa a estratégia dividir para conquistar para resolver um problema de decisão. Deste modo, um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia[17].

> "[...] a estruturação do modelo adota a estratégia "dividir para conquistar", baseando-se no conceito de razão de ganho de informação que identifica por meio da redução de entropia o quanto informativo um atributo é, para então selecionar a separação ótima, ou seja, o quanto espera-se que a entropia se reduza caso um determinado nó seja escolhido para fazer a partição dos dados."[12]

No processo de construção da árvore, há a avaliação dos pontos de separação de cada nó interno da árvore e a identificação de qual o melhor ponto de separação [12]. Essas medidas são calculadas com base nas equações [12], [7]:

- Redução de Entropia - Ganho de informação considerando a partição da base de dados associada ao nó raiz:

$$info(S) = -\sum_{i=1}^{k} \frac{freq(C_i, S)}{|S|} \times log_2 \left(\frac{freq(C_i, S)}{|S|} \right) \tag{4.1}$$

Onde:
- S representa a partição da base de dados;
- *freq*(C_i, S) representa o número de vezes em que a classe C_i acontece em S;
- C_i é a classe: i = 1, 2, 3, ..., k, k=número de classes
- |S| denota o número de casos do conjunto de treinamento;
- k indica o número de classes distintas;
- a unidade da informação é denominada *bits*.

- Redução de Entropia - Cálculo do valor da informação esperada, $info_x$(S), para o atributo X da partição S:

$$info_x(S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times info_x(S_i) \tag{4.2}$$

Onde:
- n é o número de valores possíveis que o atributo pode assumir, ou seja, o número de nós internos;
- S representa a quantidade de ocorrências na partição em análise;
- S_i representa a quantidade de ocorrências de uma classe contida no conjunto S.

E o ganho de informação será dado por:

$$ganho(X) = info(S) - info_x(S) \tag{4.3}$$

Deste modo, através da Equação 4.1 calcula-se a *entropia* do conjunto de dados completo e assim, define-se o nó raiz. E através da Equação 4.2 calcula-se a *entropia* dos demais atributos. Por fim, é calculado através da Equação 4.3 o ganho de informação de cada atributo.

> "De uma forma simplificada, o grau de entropia de um conjunto de atributos expressa o grau de complexidade da informação contida no referido conjunto. Assim, quanto menor a entropia, menor a quantidade de informação codificada em um ou mais atributos. Em contrapartida, quanto maior a entropia de um conjunto de atributos, maior a relevância destes atributos na descrição do conjunto de dados"[12].

O algoritmo construtor da árvore de decisões buscará o atributo com o maior valor da razão de ganho, de modo que quanto menor o valor da entropia, maior a pureza da partição [9].

4.3 WEKA

Em uma melhor definição da **ferramenta** empregada, a suite WEKA (Waikato Enviroment for Knowledge Analysis) é formada por um conjunto de implementações de algoritmos de diversas técnicas de MD. O WEKA está implementado na linguagem Java, podendo ser utilizada em diversos sistemas operacionais. O WEKA é um *software* livre, ou seja, está sob domínio da licença GPL e está disponível em http://www.cs.waikato.ac.nz/ml/weka [6].

A suite WEKA permite que seja realizada a abordagem de *Aprendizado Supervisionado*, quando a descoberta de conhecimento se dá a partir dos dados apresentados na forma de pares ordenados (entrada, saída desejada), atributos estes previamente conhecidos [12].

No WEKA, foi escolhido o J48 como **algoritmo classificador** para processar os dados inseridos e realizar a construção da AD. Este algoritmo calcula a razão de ganho de informação a partir dos atributos e constrói o modelo baseado em árvore de decisões.

Na utilização do algoritmo J48, o WEKA fornece um recurso para treinamento e avaliação do modelo, que é a **Validação Cruzada *k-fold***:

> "[...] consiste em dividir aleatoriamente o conjunto de dados com N elementos em K subconjuntos disjuntos (folds), com aproximadamente o mesmo número de elementos (N/K). Neste processo, cada um dos K subconjuntos é utilizado como conjunto de teste e os (K – 1) demais subconjuntos são reunidos em um conjunto de treinamento. Assim, o processo é repetido K vezes, sendo gerados e avaliados K modelos de conhecimento"[12].

Deste modo, esta investigação baseou-se nas etapas operacionais do processo KDD, buscando realizar na mineração de dados a tarefa *preditiva* de classificação de

dados, com auxílio da ferramenta WEKA, tornando possível a construção da árvore de decisões pelo uso do algoritmo J48, com validação cruzada do tipo *k-fold* para k = 10 folds.

Capítulo 5

Estudo de caso

Este capítulo apresenta as três principais etapas do processo KDD descritas no livro de Goldschmidt e Passos [12], aplicadas ao banco de dados de Indicadores Educacionais do INEP e aos dados da Proficiência em Matemática dos municípios cearenses levantados pelo SPAECE.

5.1 Pré-processamento de dados

Para o estudo do **desempenho em matemática** foi escolhido o resultado dos alunos da 3ª série do Ensino Médio da rede pública de ensino do estado do Ceará na Avaliação de Matemática do SPAECE, disponível no portal da Secretaria de Educação do Ceará - SEDUC, referentes ao ano de 2019, que são os resultados mais atualizados até a data de realização desta pesquisa. Tais dados estão disponíveis em formato quantitativo (referente ao nível de proficiência) e qualitativo (referente ao padrão de desempenho), conforme apresentado na Figura 3.4.

Os dados brutos escolhidos para o estudo do **panorama educacional** dos municípios cearenses, constam nos dados abertos dos Indicadores Educacionais disponibilizados no portal do INEP, referentes ao ano de 2019, condizente com a temporalidade dos dados do SPAECE. Tais informações podem ser acessadas na forma de microdados, organizados em tabelas com dados quantitativos.

A preparação das informações para o processo de mineração começa com a **Limpeza** dos dados para remoção de ruídos e dados inconsistentes. Na sequência, durante a fase de **Seleção de Dados** foram escolhidos 13 indicadores educacionais, de modo que cada indicador consta nos microdados do INEP em uma tabela própria com extensão *xslx*. Neste estudo, o indicador Taxa de Rendimento, que apresenta uma porcentagem distribuída entre as Taxas de Aprovação, Reprovação e Abandono, foi considerado como três indicadores distintos.

Para selecionar os dados relevantes à pesquisa, foi necessário aplicar filtros nas tabelas baixadas. A extração dos dados disponibilizados pelo INEP, teve como critérios de filtragem região - nordeste, unidade federativa - Ceará; localização - total (que inclui escolas localizadas na zona urbana e rural); dependência administrativa - pública (que faz referência aos dados das escolas públicas do município nas esferas

estadual, federal e municipal) e por fim, o valor total do indicador educacional referente às quatro séries do EM.

Na sequência, foi realizada a seleção na tabela onde consta o resultado do SPAECE e os seguintes filtros foram usados: Edição - 2019 - FINAL, com base nesta escolha, foi possível extrair o atributo Indicador de Padrão de Desempenho por município na disciplina de Matemática - que classifica o município dentro de quatro níveis de desempenho do SPAECE: muito crítico, crítico, intermediário e adequado - e também, permitiu extrair o atributo Proficiência Média - que indica a pontuação alcançada pelo município.

A **Integração** das informações provenientes dos dois bancos de dados distintos (INEP e SEDUC), teve como critério de organização os valores referentes aos municípios, desta forma,a tabela final apresenta em sua composição 184 (cento e oitenta e quatro) linhas de dados que é quantidade de municípios cearenses.

Coluna	**INDICADOR EDUCACIONAL**	**Tipo de dado**
1	Taxa de distorção idade-série	Numérico
2	Taxa de aprovação	Numérico
3	Taxa de reprovação	Numérico
4	Taxa de abandono	Numérico
5	Média de horas aulas diárias	Numérico
6	Percentual de docentes com curso superior	Numérico
7	Indicador de nível socioeconômico	Numérico
8	Média de alunos por turma	Numérico
9 a 13	Adequação da Formação Docente Grupos 1 ao 5	Numérico
14 a 19	Esforço docente Níveis 1 a 6	Numérico
20 a 25	Complexidade da gestão da escola Níveis 1 a 6	Numérico
26 a 29	Indicador de Regularidade Docente Níveis 1 a 4: baixa, média-baixa, média-alta e alta	Numérico
30	Proficiência em Matemática da 3ª série do EM no SPAECE/2019 nos quatro níveis: Muito Crítico, Crítico, Intermediário e Avançado.	Nominal

Tabela 5.1: Base de Dados

Neste momento da pesquisa, está criada a **tabela de base de dados estruturados** do estudo, unindo as informações provenientes dos dois bancos de dados, INEP e SPAECE. Tal tabela conta com 184 linhas de dados, e uma linha de cabeçalho, as linhas são chamadas de **objetos ou instâncias**. A tabela de base apresenta 30 colunas com os dados dos indicadores educacionais, que são chamados de **atributos**; destas, 29 colunas apresentam dados numéricos, são os chamados **atributos preditivos**, cujos valores serão analisados para que seja descoberto o modo como eles se relacionam com o **atributo classe**, este por sua vez, figura na 30ª coluna e traz os

dados do desempenho em Matemática alcançado pelos municípios, conforme Tabela 5.1.

A estrutura de dados orienta que a última coluna da tabela deve apresentar um atributo com dados nominais, e esta recebe o nome de atributo classe, ou apenas **classe**, que indica a classe a qual a observação pertence. Convertida a tabela de dados para o formato *csv*, com valores separados por vírgula, na fase de **Transformação**, agora os objetos estão prontos para serem minerados no *software* livre WEKA [26].

5.2 Mineração de Dados

A etapa de Mineração de Dados está vinculada à tarefa que se deseja executar. Para o presente estudo, seguindo a organização de Goldschmidt e Passos [12], foi usada a técnica de Árvore de Decisão objetivando a tarefa preditiva de *classificação*, com apoio do *software* WEKA.

No ambiente WEKA, no menu *Preprocess*, ilustrado na Figura 5.1, é possível observar a lista de atributos, as opções de filtros, um pequeno resumo sobre os atributos, dentre outras funcionalidades. Destaca-se no ambiente *Preprocess* desta pesquisa, o resumo do atributo classe, que é o Desempenho em Matemática de cada município na avaliação do SPAECE, em formato de gráfico de colunas, apontando a contagem de 166 municípios no nível **Crítico**, 2 municípios no nível **Muito Crítico** e apenas 16 municípios no **Intermediário**.

5.2.1 Árvore de Decisões

A tarefa de classificação escolhida para esta pesquisa, com base no perfil do banco de dados, foi a classificação supervisionada baseada em separabilidade (entropia) que foi executada através de algoritmos que geram árvores de decisão. No WEKA Explorer, após o pré-processamento dos dados, no menu *Classify*, optou-se por usar a metodologia de teste *Cross-validation* que usa validação cruzada do tipo *k-fold*.

A árvore de decisões é uma técnica que facilita a interpretação dos resultados, permite um melhor entendimento do fenômeno estudado e permite a identificação de variáveis com maior poder de separação dos dados [1].

Ao lembrar que o processo KDD é interativo envolvendo várias decisões feitas pelo usuário [8]. Neste ínterim, é pertinente a esta etapa da Mineração realizar testes e estudar as saídas no WEKA para podar a árvore em busca do ganho de informação.

Nesta pesquisa, foi usado o algoritmo J48 para criar a árvore de decisões a partir do banco de dados selecionado, deste modo, mantendo-se a configuração padrão do algoritmo, obteve-se um modelo de **acurácia de 85,86%**, exibido na Figura 5.2.

Esta árvore de decisões contém 9 (nove) folhas para classificar as três características do atributo classe: **Muito Crítico**, **Crítico** e **Intermediário** (nenhum município do Ceará está classificado no nível **Adequado**). Observa-se que as folhas

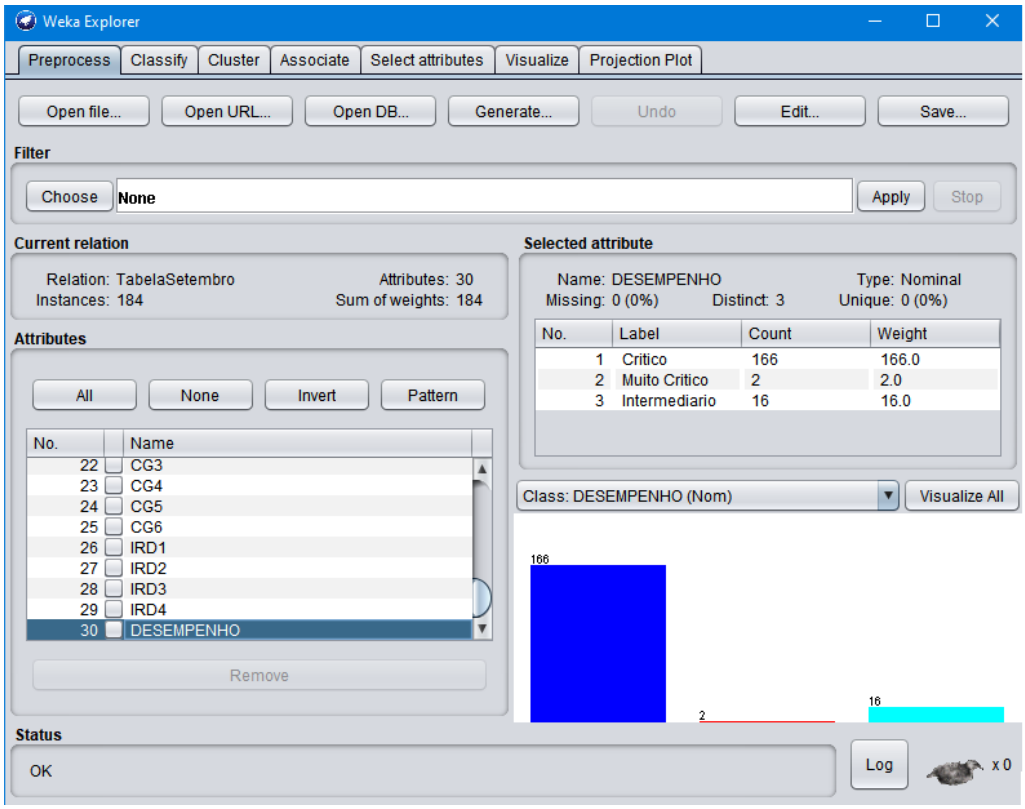

Figura 5.1: Menu Preprocess - WEKA Explorer

contemplam apenas os classificadores **Crítico** e **Intermediário**, indicando que para este modelo eles são o de maior relevância.

Nó raiz

O primeiro resultado que a árvore mostra é o atributo Taxa de distorção idade-série (TAXDIS) no lugar do nó raiz, considerado pelo algoritmo como o atributo que melhor divide o conjunto de dados conforme Figura 5.2. Este resultado evidencia a relação positiva entre o desempenho escolar medido através da Proficiência em Matemática e a defasagem idade-série, já discutida na literatura educacional brasileira.

Portella *et al* [20] indica que alunos em situação de distorção idade-série estão mais propensos ao abandono escolar, ao baixo desempenho e à reprovação. Os autores relatam também, que a defasagem idade-série se relaciona com piores níveis socioeconômicos, de educação familiar/capital cultural e ainda com a cor negra e com o sexo masculino. Por todas essas inferências, é compreensível que o modelo tenha indicado a TAXDIS como o atributo mais relevante, o que evidencia que o modelo gerado a partir dos dados, condiz com a realidade educacional do país, e com as discussões levantas nas pesquisas sobre o tema que destacam a importância dos aspectos sociais para a compreensão dos fenômenos educacionais.

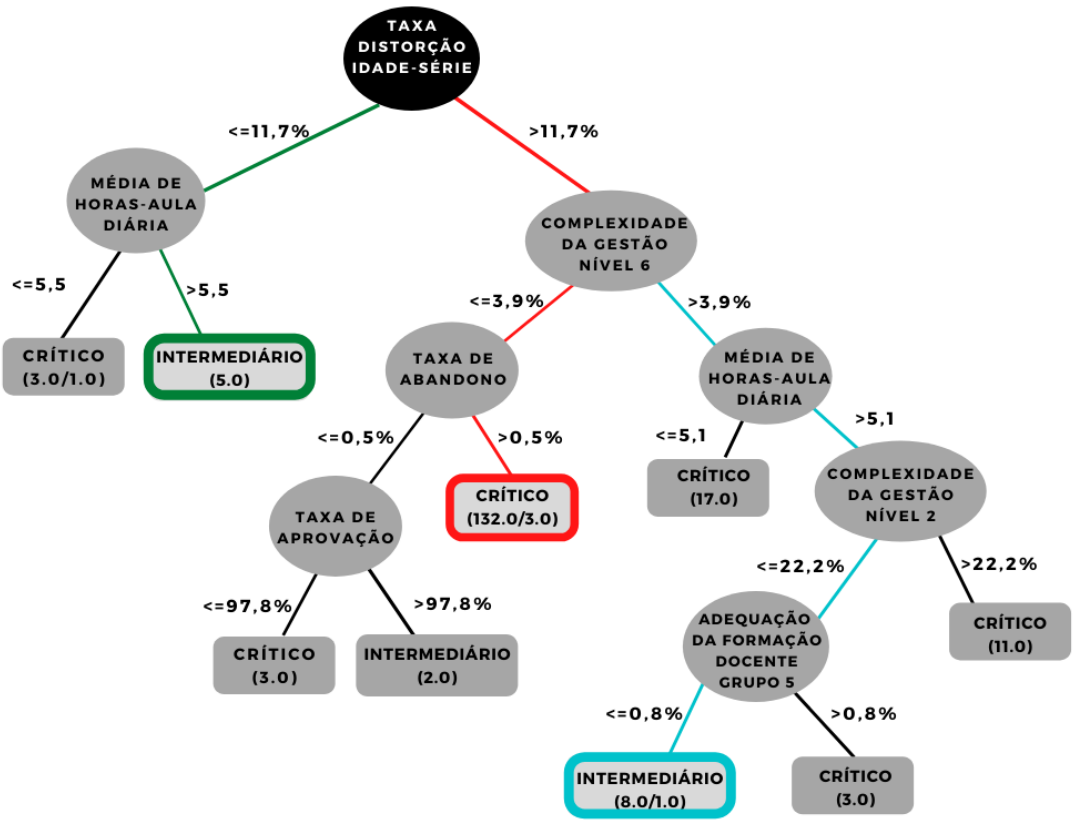

Figura 5.2: Árvore de decisões gerado pelo Algoritmo J48

A realidade da TAXDIS no Ensino Médio no Ceará é uma média de 23,29% de alunos fora da faixa etária referente a sua série e o Ceará apresenta 91 municípios com a TAXDIS acima da média estadual. Em uma inferência simples, sendo a média de alunos por turma do estado de 35,64, a porcentagem de TAXDIS indica que são aproximadamente 8 alunos por turma em situação de defasagem idade-série, como ilustrado na Figura 5.3. Na perspectiva docente, a heterogeneidade de idades dos estudantes em sala de aula transformam as atividades pedagógicas num desafio cotidiano.

Proficiência Nível Intermediário

O estudo dos Nós Folha remete à conjunção de todos os predicados existentes em um determinado caminho, sendo esse, o caminho da árvore que parte do nó raiz e termina no nó folha obedecendo as regras na forma de "se" "então" [12]. Com base nisso, destacam-se no estudo do nó folha que apresenta como valor a Proficiência em Matemática em nível **Intermediário**, dois caminhos apresentados em tons de verde e azul na Figura 5.2, com seu estudo esmiuçado nos tópicos seguintes.

- Nó Folha - Cor: Verde

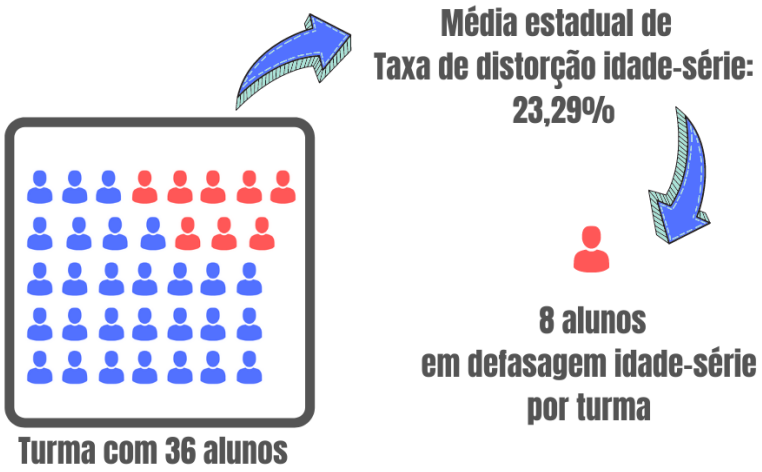

Figura 5.3: Ilustração da sala de aula estadual

Dos 184 municípios cearenses, 16 deles alcançaram o nível **Intermediário** na Proficiência em Matemática no SPAECE 2019. Em uma análise da esquerda para a direita da Árvore de Decisão da Figura 5.2, o caminho destacado de verde classifica 5 municípios no nível **Intermediário** através de dois atributos: a TAXDIS $\leq 11,7\%$ e a Média de horas aulas diárias - MHAD $> 5,5$ horas, conforme ilustrado na Figura 5.4.

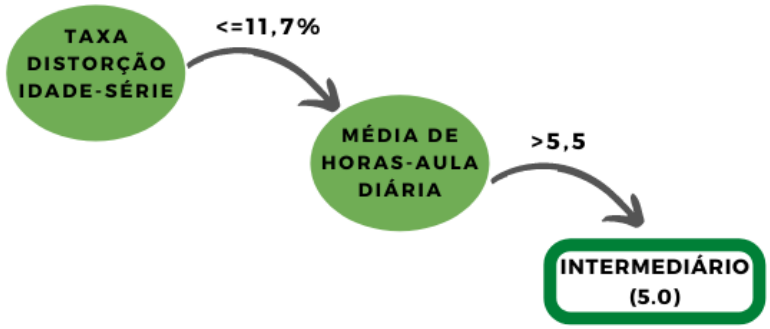

Figura 5.4: Nó folha (Verde) - Desempenho Intermediário

Estes municípios são Pedra Branca, Sobral, Pacujá, Bela Cruz e Cariré, são cidades que apresentam uma TAXDIS de menos da metade da média estadual, que é de 23,29%, e a combinação desta realidade com a quantidade de horas que o aluno permanece em aula no ambiente escolar, quando esta é maior que 5,5 horas por dia, ilustram o cotidiano destes cinco municípios que alcançaram o nível Intermediário de desempenho em Matemática.

Importante destacar que quatro desses municípios têm o INSE - Indicador de Nível Socioeconômico - acima da média estadual, o que os coloca com um público de alunos que tem wi-fi em casa ainda que não possuam computador e que os seus

responsáveis têm o ensino fundamental incompleto ou completo e/ou ensino médio completo. A tradução destes dados para a realidade das escolas indica que estas atendem a um perfil de aluno que tem condição econômica suficiente para permanecer na escola, inclusive em horário integral, sugestivo de um ambiente familiar que incentiva a educação, o que condiz com o panorama apresentado neste galho da árvore.

- Nó Folha - Cor: Azul

O Nó Folha destacado de azul na Árvore de Decisões (Figura 5.2), gerada no WEKA, advém de uma ramificação do Nó raiz e de uma sequência de testes que define os nós internos do galho, de modo que a sequência classifica corretamente a Proficiência em Matemática de desempenho **Intermediário** atingida por 7 municípios, um resultado relevante já que é quase metade da quantidade de cidades que alcançou esse desempenho em todo Estado do Ceará.

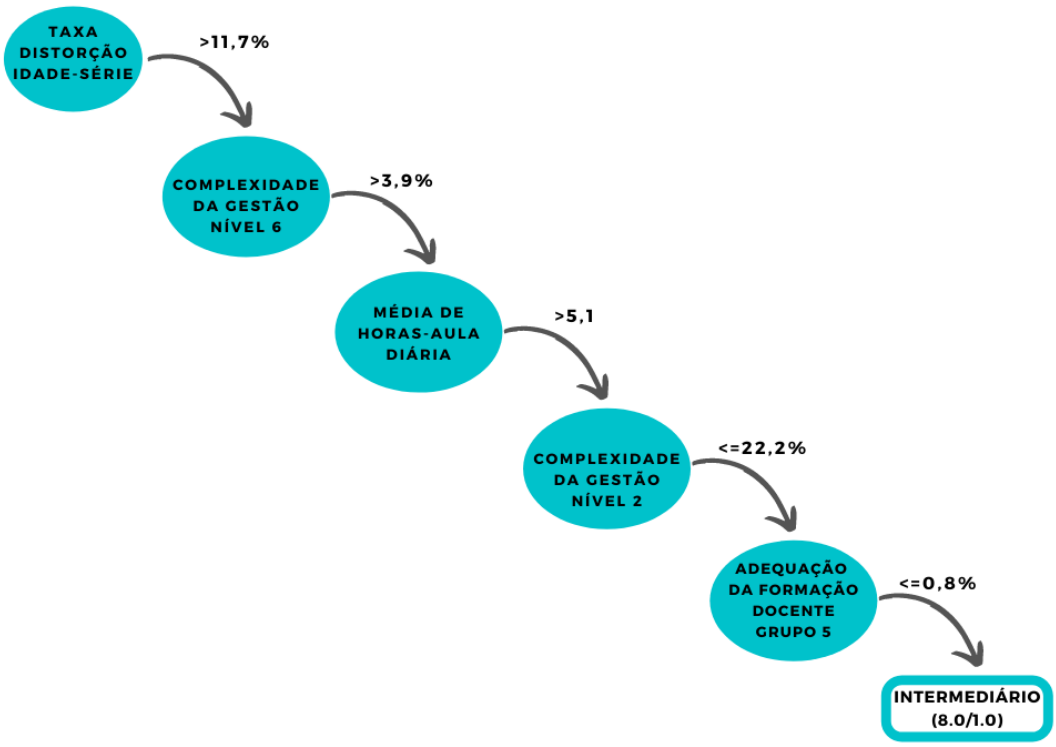

Figura 5.5: Nó folha (Azul) - Desempenho Intermediário

O ramo azul da árvore tem seus nós internos concentrados basicamente em quatro atributos preditivos: Taxa de distorção idade-série, Complexidade da gestão (CG), Média de horas-aula diária e Adequação da formação docente (AFD). De pronto, destaca-se este conjunto de atributos como uma das interações mais ricas

da árvore, por relacionar aspectos das três dimensões que permeiam o ambiente escolar: a realidade do aluno, a realidade do professor e a realidade da escola (gestão e infraestrutura), conforme esquematizado na Figura 5.6.

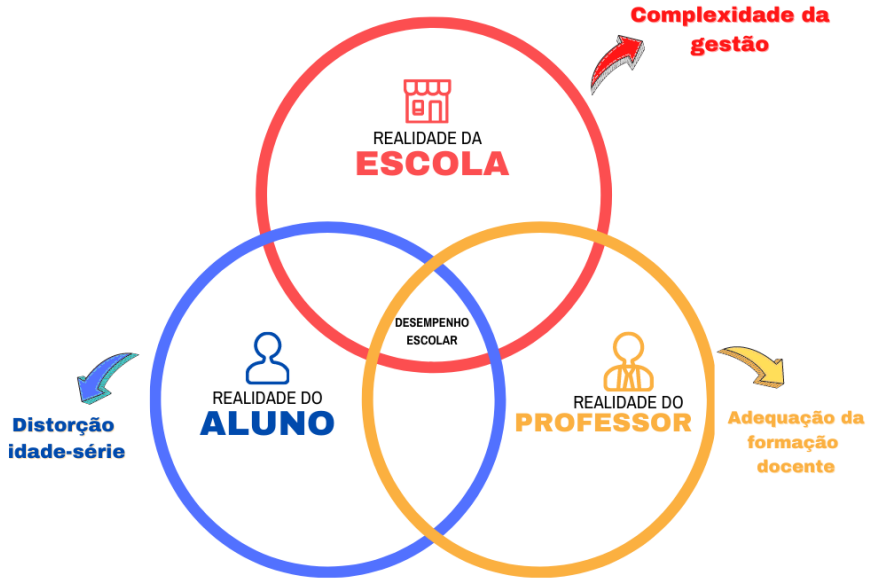

Figura 5.6: Interações escola-professor-aluno no galho azul da árvore

A partir do Nó raiz, o galho destacado de azul, apresenta a realidade educacional quando a TAXDIS é maior que 11,7%, dentre os municípios que alcançaram o desempenho **Intermediário**, essa defasagem idade-série pode chegar até 30,4%.

Na sequência, é apresentado pelo algoritmo, o atributo Complexidade da Gestão - Nível 6 (CG6), que indica o nível de gestão mais complexo da categoria, são municípios que têm em seu cenário, escolas com mais de 500 alunos, operando em 3 turnos, com 4 ou mais etapas de ensino, mostrando que a presença de escolas desse porte, aliada a outros fatores, expõe que a alta complexidade da gestão pode não influenciar de forma negativa na proficiência educacional [5], uma vez que seleciona municípios onde $CG6 > 3,9\%$.

Por fim, antes de alcançar o nó folha, o algoritmo apresenta no modelo o nó interno Adequação da Formação Docente - Grupo 5 (AFD5). O Grupo 5, na classificação do INEP, reporta a docentes que não possuem curso superior completo, e na estrutura da Árvore de Decisões se mostra como um fator negativo para o desempenho na Proficiência em Matemática, uma vez que para o percentual menor ou igual a 0,8%, aponta para o nível intermediário, e já quando a $AFD5 > 0,8\%$ aponta para um desempenho crítico.

No contexto das correlações entre os indicadores educacionais, Costa [4] indica que quanto maior a proporção de docências sem formação específica, maiores são as taxas de abandono e de distorção idade-série, e pior é a taxa de reprovação. Esta

relação entre IE é também observada na estrutura do modelo da AD construída pelo algoritmo J48, e tem reflexos no desempenho em matemática dos estudantes.

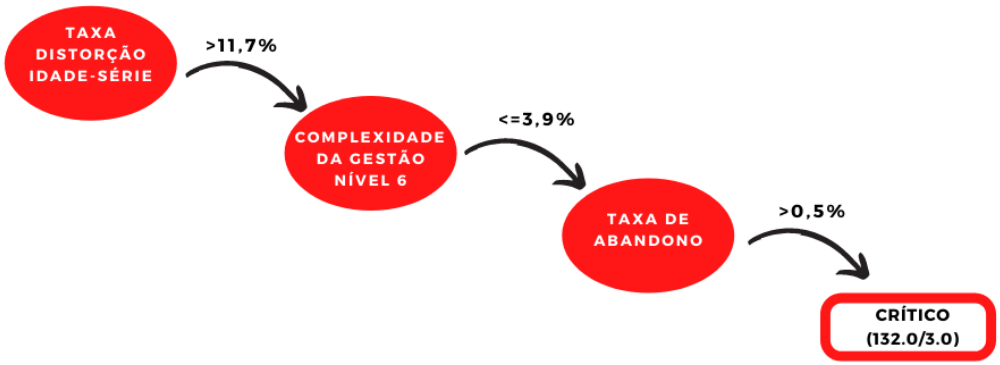

Figura 5.7: Nó folha (Vermelho) - Desempenho Crítico

Proficiência Nível Crítico

O estado do Ceará apresenta 166 municípios classificados no nível **Crítico** de Proficiência em Matemática avaliada através do SPAECE em 2019. Este resultado quando confrontado com os indicadores educacionais do mesmo ano, através da mineração de dados, com ajuda da técnica de árvore de decisões, originou o modelo da Figura 5.2, apresentado anteriormente.

No referido modelo, para explicar e predizer o contexto dos municípios do nível **Crítico** de desempenho em matemática, há um galho que chama a atenção, uma vez que classifica corretamente 129 municípios, um número significativo. Este galho da árvore está destacado de vermelho e será analisado na sequência.

- Nó Folha - Cor: Vermelho

A análise do galho destacado de vermelho na árvore de decisões da Figura 5.2, remete à discussões anteriores em relação aos indicadores educacionais Complexidade da Gestão e Taxa de distorção idade-série. O cenário dos municípios classificados neste galho da árvore, conta com índice de TAXDIS $> 11,7\%$ podendo chegar a até 46,6% e envolve municípios que em sua maioria não apresentam escolas de grande complexidade de gestão para os parâmetros avaliados pelo indicador. Dos 129 municípios selecionados pelo galho vermelho apenas 19 apresentam escolas no nível CG6, para todo restante a CG6 é zero.

Através da Figura 5.7 é possível visualizar o componente, que juntamente com a TAXDIS e a CG6, determina a criticidade do desempenho em matemática dos referidos municípios: é a Taxa de Abandono (TAXABA), quando esta é maior que 0,5%. A TAXABA nas 129 unidades municipais citadas tem uma média de 5,3%, alcançando até o índice de 14,2%. Esse indicador selecionado pelo algoritmo J48, dialoga com a taxa de distorção idade-série, nó raiz da AD, estabelecendo uma

relação causa-consequência discutida na literatura e observada na realidade escolar. Considera-se como um dos motivos para a existência de defasagem idade-série, o abandono escolar, quando o aluno deixa de frequentar a escola por um período [20].

A próxima seção **Pós-processamento de dados** apresenta a etapa do processo KDD em que os dados levantados e discutidos nesta seção foram sistematizados em uma retomada ao problema proposto neste trabalho.

5.3 Pós-processamento de dados

O processo KDD tem a característica de ter etapas correlacionadas e interdependentes, deste modo, a transcrição da árvore de decisões da Figura 5.2 é considerada uma atividade de pós-processamento por ser uma simplificação do Modelo de Conhecimento construído.

Esta etapa final do processo KDD, contempla a depuração e/ou síntese dos padrões descobertos. Se propõe a uma organização do conhecimento obtido, com uma proposta de melhorar a visualização da descoberta através de gráficos, diagramas ou relatórios demonstrativos. O objetivo é basicamente facilitar a interpretação e a avaliação do conhecimento adquirido através da síntese do Modelo de Conhecimento.

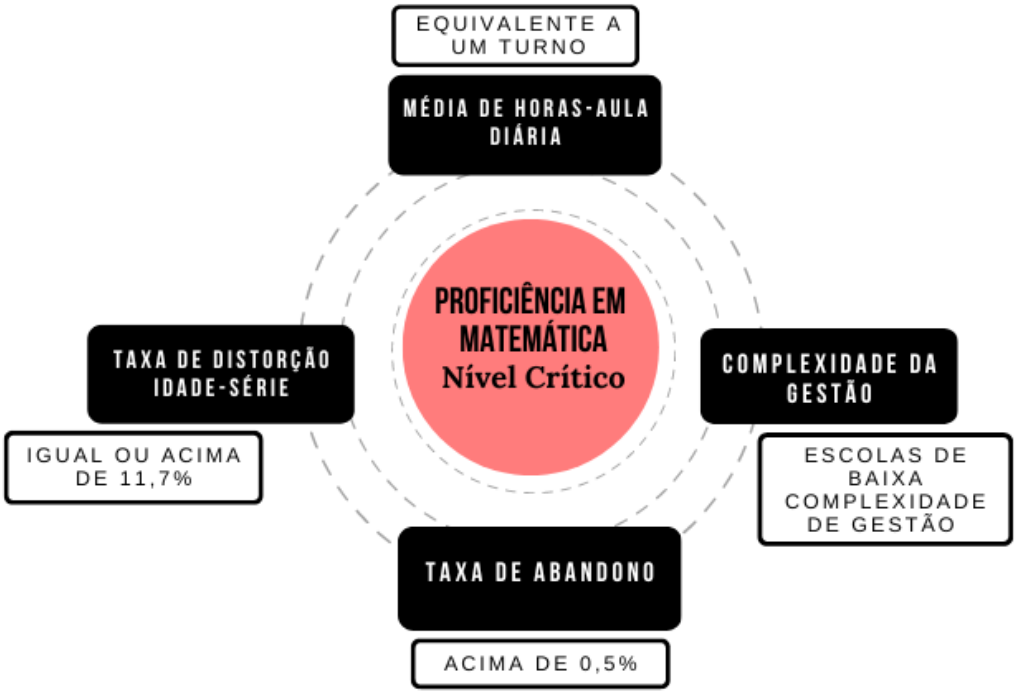

Figura 5.8: Diagrama - Proficiência em Matemática Nível Crítico
Fonte: Elaboração própria

Ao relembrar o início da investigação, para cada município foram relacionados 12 indicadores educacionais, distribuídos em 29 colunas com informações numé-

ricas, chegando a um total de 5.336 objetos numéricos. E na busca de estabelecer uma relação dessas informações com os atributos classe (nominais), o algoritmo J48 utilizou-se dos cálculos de *entropia*, bem como utilizou-se da *validação cruzada* para $k = 10$ *folds*, de modo que dividiu o grande conjunto de dados em 10 subconjuntos com aproximadamente a mesma quantidade de elementos, testou esses dados e realizou esse processo 10 vezes até construir a árvore de decisões da Figura 5.2.

Percebe-se assim, o poder da Mineração de Dados em classificar dentre tantos atributos, os que são de fato relevantes para o problema investigado, e mais, a AD fornece níveis estruturados de importância, sequência e valores dos atributos.

Na intenção de realizar as tarefas de pós-processamento, de destacar padrões e melhorar a visualização da descoberta, construiu-se diagramas que ilustram as principais observações dos dois atributos classe mais relevantes para a árvore de decisões: Nível de Desempenho Crítico e Nível de Desempenho Intermediário.

O diagrama da Figura 5.8, ilustra o contexto que envolve e caracteriza o **desempenho crítico em matemática** nos municípios cearenses. Tal desempenho está "cercado"dentre quatro indicadores, que figuram nos nós da árvore, ou seja, mantêm uma relação entre si, que são a TAXDIS, a MHAD (quando esta é menor que 5,1 horas - um turno), a CG6 e a TAXABA.

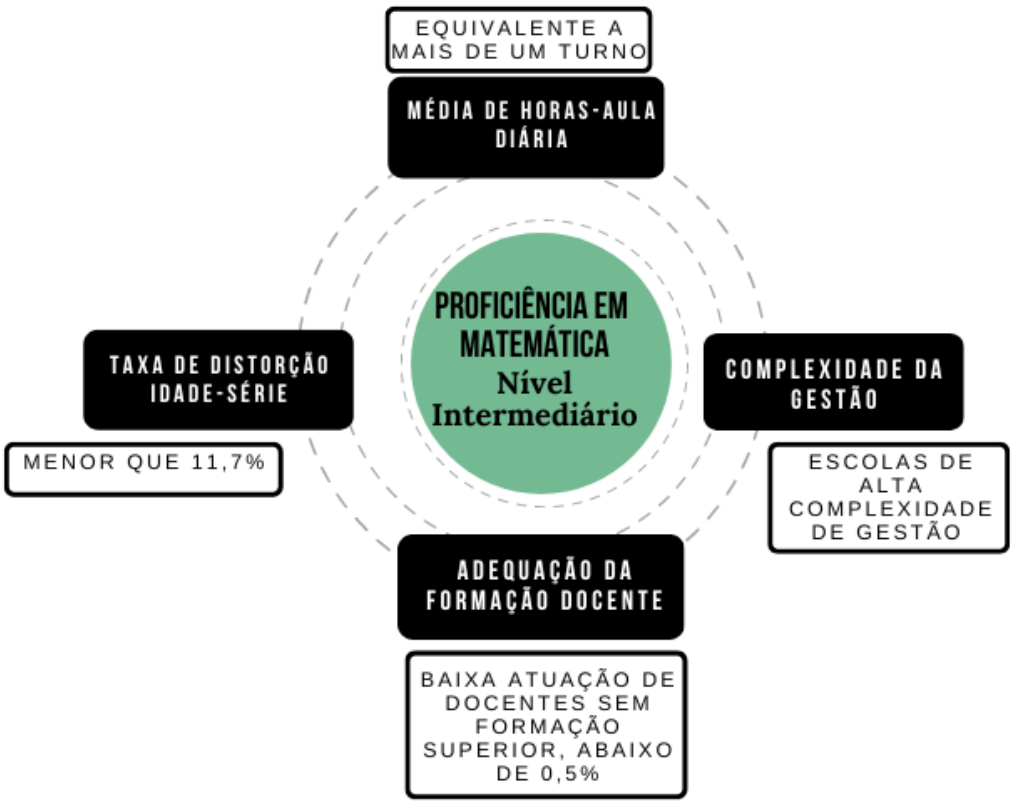

Figura 5.9: Diagrama - Proficiência em Matemática Nível Intermediário
Fonte: Elaboração própria

O diagrama da Figura 5.9, por sua vez, ilustra o contexto que envolve e

caracteriza o **desempenho intermediário em matemática** dos municípios cearenses. O propósito da ilustração é apresentar como o desempenho está "cercado"dentre quatro indicadores (que não são os mesmos da Figura 5.8), componentes dos nós da AD e inter-relacionados, são eles: a TAXDIS, a MHAD (quando acima de 5,1 horas - mais de um turno), a CG6 e AFD5.

Estabelecendo uma comparação entre os dois diagramas, os primeiros indicadores educacionais que separam o atributo crítico do atributo intermediário são a Taxa de distorção idade-série em combinação com a Média de horas-aula diárias. Observa-se que para o nível crítico, a taxa de abandono em consonância com os outros indicadores, é expressiva para o desempenho. E na realidade dos municípios que alcançaram o nível intermediário, a Adequação da formação docente, em combinação com os primeiros IE acima citados, é relevante para o desempenho.

É possível inferir que se a gestão educacional concentrar ações sobre a realidade expressada por cada um desses indicadores destacados nos diagramas, os municípios cearenses podem alcançar um melhor desempenho em matemática na avaliação do SPAECE.

Capítulo 6

Conclusões

Esta dissertação foi uma busca de *Descoberta de Conhecimento* através do processo *Knowledge Discovery in Databases - KDD* e da Mineração de Dados, aplicados aos dados de Indicadores Educacionais e aos dados da Proficiência em Matemática dos municípios cearenses levantados pelo SPAECE.

A pesquisa foi norteada pelas etapas do processo KDD descritas no livro de Goldschmidt e Passos [12]: Pré-processamento, Mineração e Pós-processamento de dados. Com destaque para a MD, pois esta é a etapa que se vincula à tarefa que se deseja executar. Por sua vez, foi realizada a tarefa *preditiva* de classificação de dados, objetivando relacionar as informações dos Indicadores Educacionais com o resultado obtido pelos alunos da 3ª série do Ensino Médio, dos municípios cearenses na avaliação do SPAECE/2019 e a partir daí extrair um *Modelo de Conhecimento* com a aplicação esperada de classificar a relação entre os IE e o desempenho em matemática dos municípios cearenses.

Considera-se que o *objetivo de aplicação* do processo KDD foi alcançado, tendo como produto de *Modelo de Conhecimento* uma Árvore de Decisões de acurácia de 85,86% desenvolvida com base nos dados selecionados. E através da árvore foi possível identificar os IE relevantes para o problema investigado, e mais, estruturados em importância, sequência e valores.

Todo o fluxo do processo, desde a integração das bases de dados de diferentes fontes, foi sistematizado e poderá ser reaplicado a novos dados de outras edições do SPAECE e Censo Escolar,o que otimiza a pesquisa e diminui o esforço técnico [23].

Aponta-se esta investigação como um incentivo aos Estados, gestores, professores e pesquisadores, para que usem da tecnologia em favor da educação, ao passo que a MDE é capaz de contribuir com o desenvolvimento de modelos preditivos de aferição do desempenho educacional.

Referências

[1] Bezerra, C., Scholz, R., Adeodato, P., Lucas, T., e Ataide, I. (2016). Evasão escolar: aplicando mineração de dados para identificar variáveis relevantes. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 27, página 1096.

[2] Caed/UFJF (2022). O sistema | spaece. Disponível em: `https://https:spaece.caedufjf.net/o-sistema/o-spaece/`. Acesso em: 07 de janeiro de 2022.

[3] Camilo, C. O. e Silva, J. C. d. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, 1(1):1–29.

[4] Costa, R., Britto, A., e Waltenberg, F. (2020). Efeitos da formação docente sobre resultados escolares do ensino médio. *Estudos Econômicos (São Paulo)*, 50(3):369–409.

[5] Da costa, F. S. (2021). Influência sa complexidade da gestão escolar no desempenho em língua portuguesa dos estudantes da 1ª série do ensino médio das escolas estaduais do amazonas.

[6] Damasceno, M. (2015). Introdução a mineração de dados utilizando o weka. Disponível em: `https://connepi.ifal.edu.br/ocs/anais/conteudo/anais/files/conferences/1/schedConfs/1/papers/258/public/258-4653-1-PB.pdf`. Acesso: 07 de setembro de 2021.

[7] do Nascimento, R. L. S., da Cruz Junior, G. G., e de Araújo Fagundes, R. A. (2018). Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. *RENOTE*, 16(1).

[8] Fayyad, U., Piatetsky-Shapiro, G., e Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

[9] Ferrari, D. G. e Silva, L. N. D. C. (2017). *Introdução a mineração de dados*. Saraiva Educação SA.

[10] Ferreira Filho, L. N., Vidal, E. M., e Júnior, J. A. d. F. P. (2020). Avaliação em larga escala no ceará e as políticas de accountability–o protagonismo do spaece. *Revista Práxis Educacional, Vitória da Conquista*, 16(43):452–471.

[11] Furlan, M. (2018). Algoritmos e técnicas para a mineração de dados. Disponível em: `https://cepein.femanet.com.br/BDigital/arqTccs/1511420203.pdf`. Acesso em: 07 de janeiro de 2022.

[12] Goldschmidt, R. e Passos, E. (2005). *Data Mining*. Elsevier Brasil.

[13] Gomes, M. B. e Azevedo, L. C. T. d. (2011). A prática do monitoramento da educação no município e na escola. Disponível em: `https://educere.bruc.com.br/CD2011/pdf/5736_3231.pdf`. Acesso em: 18 novembro 2021.

[14] Hippolyto, L. d. Q. (2013). Avaliação dos resultados do 3º ano do ensino médio em matemática no ceará e sua repercussão na prática pedagógica dos professores: um estudo descritivo a partir dos testes do spaece nos anos 2008, 2009 e 2010.

[15] Júnior, A. G. M. e de Farias, M. A. (2016). Spaece: Uma história em sintonia com avaliação educacional do governo federal. *Revista de Humanidades*, 31(2):525–547.

[16] Junior, R. N., do Nascimento, R. L. S., de Araújo Fagundes, R. A., e de Mattos Neto, P. S. G. (2019). Estimação de índices de aprovação e reprovação escolar do ensino médio. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, página 339.

[17] Lorena, A. C., Gama, J., e Faceli, K. (2000). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Grupo Gen-LTC.

[18] Manhães, L. M. B., Da Cruz, S. M. S., Costa, R. J. M., Zavaleta, J., e Zimbrão, G. (2012). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)*, volume 1.

[19] Piatetsky-Shapiro, G. (1999). The data-mining industry coming of age. *IEEE Intelligent Systems and their Applications*, 14(6):32–34.

[20] Portella, A. L., Bussmann, T. B., e Oliveira, A. M. H. d. (2017). A relação de fatores individuais, familiares e escolares com a distorção idade-série no ensino público brasileiro. *Nova economia*, 27:477–509.

[21] Rodrigues, R. L., Ramos, J. L. C., Silva, J. C. S., e Gomes, A. S. (2014). A literatura brasileira sobre mineração de dados educacionais. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 3, página 621.

[22] Santana, F. (2020). Minerando dados | Árvores de decisão (projeto passo a passo). Disponível em: `https://https://minerandodados.com.br/arvores-de-decisao-conceitos-e-aplicacoes/`. Acesso em: 18 novembro 2021.

[23] Silva Filho, R. L. C. (2017). Modelo de análise e predição do desempenho dos alunos dos institutos federais de educação usando o enem como indicador de qualidade escolar.

[24] Vieira, F. e Oliveira, S. d. M. (2014). Mineração de dados: conceitos e um estudo de caso sobre certificação racial de ovinos. *Embrapa Informática Agropecuária-Capítulo em livro científico (ALICE)*.

[25] Vitelli, R. F., Fritsch, R., e Corsetti, B. (2018). Indicadores educacionais na avaliação da educação básica e possíveis impactos em escolas de ensino médio no município de porto alegre, rio grande do sul. *Revista Brasileira de Educação*, 23.

[26] Witten, I. H., Frank, E., Hall, M. A., Pal, C., e DATA, M. (2005). *Practical machine learning tools and techniques*, volume 2.