UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE EDUCAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS EDUCACIONAIS EM REDE MESTRADO PROFISSIONAL

Alexandre Abreu de Paula

MINERAÇÃO DE DADOS PARA ANÁLISE DE DESEMPENHO DE ALUNOS DO ENSINO FUNDAMENTAL

Santa Maria – RS
2022

Alexandre Abreu de Paula

MINERAÇÃO DE DADOS PARA ANÁLISE DE DESEMPENHO DE ALUNOS DO ENSINO FUNDAMENTAL

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologias Educacionais em Rede, da Universidade Federal de Santa Maria (UFSM, RS) como requisito parcial para a obtenção do título de Mestre em Tecnologias Educacionais em Rede.

Orientadora: Profª. Drª. Solange de Lurdes Pertile

Santa Maria – RS
2022

Paula, Alexandre Abreu de
MINERAÇÃO DE DADOS PARA ANÁLISE DE DESEMPENHO DE ALUNOS DO ENSINO FUNDAMENTAL / Alexandre Abreu de Paula. 2022.
72 p.; 30 cm

Orientadora: Solange de Lurdes Pertile
Dissertação (mestrado) - Universidade Federal de Santa Maria, Centro de Educação, Programa de Pós-Graduação em Tecnologias Educacionais em Rede, RS, 2022

1. Mineração de Dados Educacionais 2. Descoberta de Conhecimento em Bases de Dados 3. Avaliação da Educação Básica no Brasil 4. Ensino Fundamental 5. IDEB I. Pertile, Solange de Lurdes II. Título.

Sistema de geração automática de ficha catalográfica da UFSM. Dados fornecidos pelo autor(a). Sob supervisão da Direção da Divisão de Processos Técnicos da Biblioteca Central. Bibliotecária responsável Paula Schoenfeldt Patta CRB 10/1728.

Declaro, ALEXANDRE ABREU DE PAULA, para os devidos fins e sob as penas da lei, que a pesquisa constante neste trabalho de conclusão de curso (Dissertação) foi por mim elaborada e que as informações necessárias objeto de consulta em literatura e outras fontes estão devidamente referenciadas. Declaro, ainda, que este trabalho ou parte dele não foi apresentado anteriormente para obtenção de qualquer outro grau acadêmico, estando ciente de que a inveracidade da presente declaração poderá resultar na anulação da titulação pela Universidade, entre outras consequências legais.

Alexandre Abreu de Paula

MINERAÇÃO DE DADOS PARA ANÁLISE DE DESEMPENHO DE ALUNOS DO ENSINO FUNDAMENTAL

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologias Educacionais em Rede, da Universidade Federal de Santa Maria (UFSM, RS) como requisito parcial para a obtenção do título de Mestre em Tecnologias Educacionais em Rede.

Aprovada em 31 de março de 2023:

_____

Solange de Lurdes Pertile, Dra. (UFSM)
(Presidente/Orientadora)

_____

Fernando de Jesus Moreira Junior, Dr. (UFSM)
(Coorientador)

_____

Adriana Soares Pereira, Dra. (UFSM)

_____

Edimar Manica, Dr. (IFRS)

Santa Maria – RS
2022

AGRADECIMENTO

Agradeço primeiramente a Deus, por iluminar meu caminho e me dar forças em todos os momentos.

À minha família, pelo amor e apoio dado, assim como aos meus colegas de trabalho do GAM pelo incentivo e aos meus colegas do curso pela aprendizagem e compartilhamento.

Aos meus professores, do PPGTER pelas orientações valiosas e incentivo à pesquisa e quero destacar a Giliane Bernadi coordenadora do curso que sempre esteve auxiliando nas dificuldades.

À minha orientadora, Solange Pertile, por seu constante apoio, dedicação e paciência durante todo o processo de elaboração da minha dissertação.

Agradecimento à banca examinadora, ao Fernando de Jesus Moreira Junior pelo conhecimento passado, assim como Edimar Manica pela dedicação e esforço na avaliação e gostaria de expressar minha profunda gratidão a Adriana Pereira que foi minha professora desde a graduação, transmitindo seu conhecimento. Agradeço a todos pelo tempo e pelas sugestões valiosas fornecidas.

RESUMO

MINERAÇÃO DE DADOS PARA ANÁLISE DE DESEMPENHO DE ALUNOS DO ENSINO FUNDAMENTAL

AUTOR: Alexandre Abreu de Paula
ORIENTADORA: Solange de Lurdes Pertile

Nos dias atuais, o desafio para o campo das políticas sociais é melhorar a qualidade da Educação Básica, e o Índice de Desenvolvimento da Educação Básica (IDEB) é a medida utilizada para mensurar o desempenho do Sistema Educacional Brasileiro. Para enfrentar esse desafio, as Tecnologias de Informação e Comunicação (TIC) têm se mostrado uma plataforma útil de apoio ao processo de aprendizagem, gerando grandes quantidades de dados. Nas escolas estaduais do Rio Grande do Sul, o sistema de Informatização da Secretaria da Educação (ISE) é uma ferramenta que armazena informações dos alunos e gera dados que podem ser usados para a descoberta de novos conhecimentos. O objetivo desta pesquisa é gerar um modelo de predição que permita identificar alunos com potencial de reprovação, de forma que a escola possa traçar ações mais focadas para esses estudantes, possibilitando aos gestores e professores aplicar soluções para melhorar o desempenho dos alunos. O desenvolvimento do trabalho ocorreu em três etapas, com um método de investigação dedutivo, quali-quantitativo, que incluiu uma pesquisa bibliográfica acerca do tema norteador e teve como universo de pesquisa alunos do 6º e 9º ano dos anos de 2016 a 2019. Na primeira etapa, foi realizado um estudo sobre o funcionamento do IDEB, a utilização de técnicas de Mineração de Dados na Educação (MDE) e o processo de descoberta de Conhecimento em Bases de Dados (KDD). Na segunda etapa, foram exploradas etapas do processo de KDD aplicando técnicas e algoritmos de mineração de dados (MD) para a identificação de atributos relacionados ao baixo desempenho dos alunos do 6º e 9º ano. Na terceira etapa, foram avaliados os dados e construído um sistema *web* para a predição de alunos. Os resultados indicaram que, na mineração do experimento dos alunos dos 6º anos, a maior dificuldade encontrada foi na disciplina de matemática, o que está em consonância com os resultados obtidos no IDEB. Já no experimento dos alunos do 9º ano, a disciplina com maior dificuldade foi ciências. Em ambos os experimentos, o padrão encontrado pelo algoritmo como nó raiz foi a nota do 2º trimestre em matemática para o 6º ano e ciências para o 9º, possibilitando aos gestores uma ação necessária para auxiliar alunos que possam ser reprovados.

Palavra-chave: Avaliação da Educação Básica no Brasil. Mineração de Dados Educacionais. Descoberta de Conhecimento em Bases de Dados.

ABSTRACT

DATA MINING FOR PERFORMANCE ANALYSIS OF ELEMENTARY SCHOOL STUDENTS

AUTHOR: Alexandre Abreu de Paula
ADVISOR: Solange Pertile

Nowadays, the challenge for the field of social policies is to improve the quality of basic education, and the Basic Education Development Index (Ideb) is the measure used to measure the performance of the Brazilian educational system. To face this challenge, Information and Communication Technologies (ICT) have proven to be a useful platform to support the learning process, generating large amounts of data. In state schools in Rio Grande do Sul, the Education Department's (ISE) computerization system is a tool that stores student information and generates data that can be used to discover new knowledge. The objective of this research is to generate a prediction model that allows identifying students with the potential to fail, so that the school can outline more focused actions for these students, enabling managers and teachers to apply solutions to improve student performance. The development of the work took place in three stages, with a deductive, qualitative and quantitative research method, which included a bibliographical research on the guiding theme and had as research universe 6th and 9th grade students from 2016 to 2019. At this stage, a study was carried out on the functioning of IDEB, the use of Data Mining techniques in Education (MDE) and the process of discovering knowledge in databases (KDD). In the second stage, stages of the KDD process were explored, applying data mining (DM) techniques and algorithms to identify attributes related to the low performance of 6th and 9th grade students. In the third stage, the data were evaluated and a web system was built to predict students. The results indicated that, in mining the 6th grade students' experiment, the greatest difficulty was found in the mathematics discipline, which is in line with the results obtained in IDEB. In the experiment with 9th grade students, the subject with the greatest difficulty was science. In both experiments, the pattern found by the algorithm as the root node was the 2nd quarter grade in math for the 6th grade and science for the 9th grade, allowing managers to take the necessary action to help students who might fail.

Keyword: Evaluation of Basic Education in Brazil. Educational Data Mining. Discovery of Knowledge in Databases.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ARFF	*Attribute-Relation Files Format*
ASCII	*American Standard Code for Information Interchange*
ASQ	*American Society for Quality*
AVA	Ambientes Virtuais de Aprendizagem
CBIE	Congresso Brasileiro de Informática na Educação
CRISP-DM	*CRoss-Industry Standard Process for Data Mining*
CSV	*Comma-Separated Values*
EDM	*Educational Data Mining*
EEEF	Escola Estadual de Ensino Fundamental
GAM	Escola Estadual de Ensino Fundamental Dr. Gabriel Álvaro de Miranda
GNU	*General Public License*
IDEB	Índice de Desenvolvimento da Educação Básica
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
ISE	Informatização da Secretaria da Educação
KDD	*Knowledge Discovery in Databases*
MD	Mineração de Dados
MDE	Mineração de Dados Educacionais
MOOCS	*Massive Open Online Course*
PDF	*Portable Document Format*
RSL	Revisão Sistemática da Literatura
SAEB	Sistema de Avaliação da Educação Básica
SBIE	Simpósio Brasileiro de Informática na Educação
SIE	Sistemas de Informação Educacionais
STI	Sistemas Tutores Inteligentes
TIC	Tecnologias de Informação e Comunicação
UFRGS	Universidade Federal do Rio Grande do Sul
UFSM	Universidade Federal de Santa Maria

SUMÁRIO

1 INTRODUÇÃO

A Constituição Federal de 1988, em seu Art. 208, inciso I, foi alterada pela Emenda Constitucional 59, de 11 de novembro de 2009, para estabelecer a obrigatoriedade da Educação Básica gratuita dos 4 (quatro) aos 17 (dezessete) anos de idade, com o objetivo de fornecer aos estudantes uma formação escolar como base para o desenvolvimento humano em sua plenitude, em condições de liberdade e dignidade, respeitando e valorizando as diferenças. No entanto, ter acesso à educação não é suficiente, é necessário garantir a permanência com direito à aprendizagem de qualidade (NEUHAUS, 2016).

Em 2022, o Governo Federal estabeleceu metas para melhorar os níveis de desempenho da educação pública brasileira, a fim de atingir o patamar médio dos países mais desenvolvidos. O grande desafio atual no campo das políticas sociais é melhorar a qualidade da Educação Básica. Em 2007, o Índice de Desenvolvimento da Educação Básica (IDEB) foi criado para medir o desempenho do sistema educacional brasileiro, levando em consideração o fluxo escolar e as médias de desempenho nas avaliações, calculado a partir dos dados de aprovação escolar obtidos no Censo Escolar e das médias de desempenho no Sistema de Avaliação da Educação Básica - SAEB (CASTRO, 2018).

Hoje, as Tecnologias de Informação e Comunicação (TIC) têm sido cada vez mais utilizadas no processo educacional como uma plataforma de apoio à aprendizagem. O crescente uso de Sistemas de Informações Educacionais (SIE) e Ambientes Virtuais de Aprendizagem (AVA) tem gerado uma grande quantidade de dados sobre o processo de ensino, permitindo descobrir novos conhecimentos relevantes (RODRIGUES, 2014). Como resultado, uma vasta quantidade de dados é produzida e armazenada pelos mais diversos setores, como Saúde, Educação e Negócios.

Surge então uma nova área de estudo, a Mineração de Dados Educacionais (MDE), a qual é um segmento de pesquisa interdisciplinar que utiliza métodos para estudar dados originados a partir do cenário educacional (ALVES, 2018), combinando as áreas de Computação, Educação e Estatística (RODRIGUES, 2014).

Novas áreas de pesquisa associadas às TIC com a educação têm surgido com o intuito de apresentar novos métodos e técnicas para aperfeiçoar essa relação que ofereça opções para superar os desafios encontrados nas estruturas educacionais. E

uma dessas áreas, está ligada à técnica de mineração de dados que ao ser usada na educação pode auxiliar o gestor, professor ou outros atores do sistema educacional, a desenvolver ações que contribuam no auxílio de alunos com maior dificuldade que possam melhorar seu desempenho.

A MDE tem como objetivo a descoberta de informações que possam contribuir no processo de ensino, no desempenho dos alunos e nos fatores que favoreçam a aprendizagem, representando melhoras na qualidade da educação. Essa qualidade é baseada em valores estatísticos atribuídos por meio de indicadores educacionais, que levam em consideração não apenas o desempenho do aluno, mas também o contexto econômico e social da escola em que estão inseridos (NEVES JUNIOR, 2019).

1.1 PROBLEMA DE PESQUISA

A escola tem como principal missão fornecer uma educação de qualidade, que permita aos alunos ter um bom desempenho ao final do processo educativo. Embora o acesso à Educação Básica obrigatória e gratuita já não seja mais um problema, a taxa de repetência dos estudantes ainda é alta, assim como a baixa proficiência demonstrada nos exames padronizados, como o IDEB. Esse indicador é calculado a partir de dados sobre aprovação escolar obtidos pelo Censo Escolar e a média de desempenho dos alunos nas avaliações da Prova Brasil, aplicadas no 5º e 9º anos do Ensino Fundamental, bem como no 3º ano do Ensino Médio (INEP, 2020).

A MDE está comprometida em explorar tipos de dados exclusivos por meio de métodos que possam fornecer uma compreensão mais profunda dos alunos e dos ambientes de aprendizagem. Para atingir esse objetivo, há disponíveis grandes repositórios de dados que foram criados com informações sobre os alunos. Esses dados podem ser acessados para se obter uma melhor compreensão dos alunos e da sua aprendizagem (ROMERO, 2010).

Esses repositórios de dados contêm informações sobre escolas, diretores, professores e alunos, que são coletados por meio do censo escolar e das avaliações, como a Prova Brasil. Essas informações são armazenadas em uma grande base de dados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). No entanto, muitas vezes, esses dados não são explorados em seu potencial máximo devido ao grande volume de informações envolvidas, o que inviabiliza a interpretação humana (NAMEN, 2013).

Com o avanço tecnológico, surgiram novas ferramentas que impactam significativamente o setor educacional, tais como *softwares* de ensino, o uso da internet como plataforma de aprendizagem e a crescente popularização do *e-learning*. Além disso, a administração digital dos registros dos alunos tem sido, cada vez mais, comum entre os gestores das instituições de ensino. Como resultado, houve um aumento significativo no volume de dados educacionais, tornando essencial o uso de recursos computacionais para análise (SOUZA, 2021).

Nas escolas estaduais do Estado do Rio Grande do Sul (RS), os dados se dão através de uma ferramenta que permite reduzir o tempo de envio de documentos e padronizar os relacionamentos eletrônicos entre escolas que é a Informatização da Secretaria da Educação (ISE), um sistema on-line contemplando a Gestão Escolar, o qual permite a consulta das informações armazenadas que incluem, além dos secretários de escolas, alunos, direções, coordenações pedagógicas e professores (EDUCAÇÃO, 2021).

Dessa forma, é possível aplicar técnicas de MDE para auxiliar tanto as escolas quanto gestores e professores a descobrirem indicadores que possam estar contribuindo para o baixo desempenho dos alunos do Ensino Fundamental no IDEB, utilizando os dados coletados dos alunos do ISE. Essas informações são convertidas de dados brutos em *insights* úteis com o objetivo de analisar e resolver questões relacionadas à pesquisa educacional, possibilitando ações mais direcionadas para melhorar o desempenho desses alunos.

Após analisar o IDEB da Escola Estadual de Ensino Fundamental Dr. Gabriel Álvaro de Miranda (GAM), do 1º ao 5º ano (anos iniciais) e do 6º ao 9º ano (anos finais), verificou-se que nos anos iniciais a escola alcançou as metas estabelecidas pelo governo, conforme demonstrado na Figura 1. Entretanto, nos anos finais, somente em 2009 a escola obteve sucesso nos resultados, como pode ser observado na Figura 2.

Figura 1 - Resultado IDEB da Escola GAM dos anos Iniciais 4ª série/5º ano.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

IDEB
Índice de Desenvolvimento da Educação Básica

IDEB - Resultados e Metas

Parâmetros da Pesquisa

Resultado:	Escola	UF:	RS
Município:	CRUZ ALTA	Nome da Escola:	ESC EST ENS FUN DR GABRIEL ALVARO DE MIRANDA
Rede de ensino:	Estadual	Série / Ano:	4ª série / 5º ano

4ª série / 5º ano

	Ideb Observado								Metas Projetadas							
Escola	2005	2007	2009	2011	2013	2015	2017	2019	2007	2009	2011	2013	2015	2017	2019	2021
ESC EST ENS FUN DR GABRIEL ALVARO DE MIRANDA	4.4	5.1	5.3	5.5	6.8	6.5	6.6	6.9	4.5	4.8	5.2	5.5	5.7	6.0	6.3	6.5

Obs:

* Número de participantes no SAEB insuficiente para que os resultados sejam divulgados.
** Sem média no SAEB: Não participou ou não atendeu os requisitos necessários para ter o desempenho calculado.
*** Solicitação de não divulgação conforme Portaria Inep.
Os resultados marcados em verde referem-se ao Ideb que atingiu a meta.

Pesquisar Novamente

Atualizado em 15/09/2020

Copyright MEC - INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Fonte: http://ideb.inep.gov.br/resultado/

Figura 2 - Resultado IDEB da Escola GAM dos anos Finais 8ª série/9º ano.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

IDEB
Índice de Desenvolvimento da Educação Básica

IDEB - Resultados e Metas

Parâmetros da Pesquisa

Resultado:	Escola	UF:	RS
Município:	CRUZ ALTA	Nome da Escola:	ESC EST ENS FUN DR GABRIEL ALVARO DE MIRANDA
Rede de ensino:	Estadual	Série / Ano:	8ª série / 9º ano

8ª série / 9º ano

	Ideb Observado								Metas Projetadas							
Escola	2005	2007	2009	2011	2013	2015	2017	2019	2007	2009	2011	2013	2015	2017	2019	2021
ESC EST ENS FUN DR GABRIEL ALVARO DE MIRANDA	4.0	3.8	4.4	4.4	4.4	5.0	5.0	5.6	4.0	4.2	4.5	4.9	5.2	5.5	5.7	6.0

Obs:

* Número de participantes no SAEB insuficiente para que os resultados sejam divulgados.
** Sem média no SAEB: Não participou ou não atendeu os requisitos necessários para ter o desempenho calculado.
*** Solicitação de não divulgação conforme Portaria Inep.
Os resultados marcados em verde referem-se ao Ideb que atingiu a meta.

Pesquisar Novamente

Atualizado em 15/09/2020

Copyright MEC - INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

Fonte: http://ideb.inep.gov.br/resultado/

Através da constatação dos resultados e metas apresentados pelo IDEB da escola, a seguinte problemática carece de diretrizes norteadoras: em que medida é possível descobrir novos conhecimentos com base nos dados encontrados no ISE, a fim de gerar uma predição para identificar com antecedência alunos que possam apresentar baixo desempenho?

1.2 OBJETIVOS

O objetivo deste projeto é desenvolver um modelo de previsão de fatores que possam identificar alunos com risco de reprovação nos anos finais do Ensino Fundamental. Ao fornecer ferramentas para ajudar educadores, gestores e professores, podendo implementar soluções eficazes para melhorar o desempenho dos alunos na Escola Estadual de Ensino Fundamental Dr. Gabriel Álvaro de Miranda, localizada em Cruz Alta, RS.

1.2.1 Objetivo Geral

A proposta é utilizar técnicas de mineração de dados para analisar informações do ISE dos alunos entre 2016 e 2019, tendo em vista as mudanças ocorridas no sistema de ensino em função da pandemia em 2020, como a implementação de aulas *online* no RS. O objetivo é criar uma predição que permita antecipar o desempenho insuficiente de alunos, possibilitando o suporte adequado para superar suas dificuldades. Com base nos resultados obtidos, espera-se fornecer subsídios que permitam aos gestores e professores desenvolver soluções para aprimorar o rendimento escolar dos alunos.

1.2.2 Objetivo Específicos

Para que o objetivo geral seja atingido, alguns objetivos específicos devem ser alcançados:

- Realizar buscas de dados através do ISE da escola para a análise;
- Criar uma base de dados dos anos de 2016 a 2019 dos 6º anos e 9º anos;

- Realizar um Pré-processamento dos dados para eliminar ruídos, limpeza, seleção de atributos, integração de dados;
- Transformar os dados brutos em dados compreensíveis a ferramenta *WEKA* para análise;
- Aplicar técnicas de mineração de dados para criar um modelo de predição;
- Usar a ferramenta *WEKA* para obtenção da predição;
- Avaliar qualitativamente os dados disponibilizados pela mineração de dados de forma a avaliar o modelo apresentado;
- Criar um sistema *web* para auxílio da equipe diretiva, para predição de alunos para uma possível revogação da reprovação;

1.3 JUSTIFICATIVA

Com o surgimento do índice para mensurar o desempenho do sistema educacional brasileiro, onde reúne o fluxo escolar obtido no Censo Escolar e as médias de desempenho nas avaliações no SAEB, surgem dados educacionais que podem ser manipulados para entender melhor o desempenho do aluno. Esses dados são armazenados no sistema *online* da escola, ISE, permitindo o acesso às informações e a geração de novos dados que podem ser analisados para identificar relações relevantes.

A Mineração de Dados (MD) “é uma técnica que visa extrair conhecimento valioso através da análise de grandes conjuntos de dados, usando métodos automáticos ou semiautomáticos para detectar padrões” (SILVA, 2016, p. 07). Na área de MD, surgiu uma especialização conhecida como Mineração de Dados Educacionais (MDE), que se concentra na extração de conhecimento útil por meio da análise de dados coletados em ambientes educacionais, tanto presenciais quanto virtuais (BAKER; ISOTANI; CARVALHO, 2011).

Através da utilização de técnicas de MDE e auxílio da ferramenta *WEKA* é possível converter dados brutos obtidos no ISE em uma descoberta de predição, achando um padrão de indicadores que podem prever uma reprovação. Dessa forma, os resultados podem ser divulgados, auxiliando o gestor e professor a tomar uma decisão estratégica, possibilitando com antecedência auxiliar o aluno que tem dificuldades.

2 REFERENCIAL TEÓRICO

Neste Capítulo serão apresentadas as fundamentações teóricas sobre uma breve explanação sobre a Avaliação da Educação Básica no Brasil, Mineração de Dados, Mineração de Dados Educacionais e *WEKA*. Também serão fundamentados teoricamente o processo de descoberta de conhecimento em mineração de dados e suas respectivas etapas.

2.1 AVALIAÇÃO DA EDUCAÇÃO BÁSICA NO BRASIL

Durante os anos 1990, a Educação Básica no Brasil enfrentou desafios significativos, como o acesso limitado ao Ensino Fundamental obrigatório por crianças de famílias pobres, altas taxas de reprovação e abandono escolar. Além disso, os professores que atuavam nesse nível de ensino frequentemente não tinham qualificação adequada. No entanto, a implementação de um sistema de informação e avaliação educacional contribuiu para o desenvolvimento de políticas que buscavam democratizar o acesso à educação e aprofundar o conhecimento sobre os fatores que impactavam a qualidade da educação brasileira (CASTRO, 2018).

A qualidade é um termo em que é utilizada em vários contextos, ar, água, vida e outros. Para afirmar que algo é ou não de qualidade, depende de fatores em inúmeras situações, sendo diferente entre duas pessoas. Um mesmo produto ou serviço pode ter noções de qualidade relacionadas à necessidade e expectativa da pessoa, uns analisam pelo desempenho do produto enquanto outros pelo preço (NEUHAUS, 2016).

Conforme ASQ (*American Society for Quality* – Sociedade Americana para a Qualidade) tem a seguinte definição para o termo qualidade:

> Termo subjetivo para o qual cada pessoa ou setor tem sua própria definição. No uso técnico, qualidade pode ter dois significados: 1) as características de um produto ou serviço que influenciam sua capacidade de satisfazer necessidades declaradas ou implícitas; 2) um produto ou serviço livre de deficiências. De acordo com Joseph Juran, qualidade significa “adequação ao uso”; de acordo com Philip Crosby, significa "conformidade com os requisitos". (ASQ, 2021, s.n., grifos do autor).

Na educação, a definição de qualidade leva em consideração tanto a eficiência quanto a eficácia dos sistemas educacionais. É importante considerar fatores intraescolares e extraescolares, como condições socioeconômicas e culturais, gestão escolar, profissionalização dos professores, acessibilidade e outros aspectos relevantes. Dessa forma, a qualidade da educação pode ser avaliada sob uma perspectiva social e econômica, tendo sucesso quando contribui para a equidade e eficácia dos recursos investidos na educação.

No entanto, um dos maiores desafios enfrentados pelo país no campo das políticas sociais é melhorar a qualidade da Educação Básica. Para isso, é necessário obter um diagnóstico claro das causas dos problemas encontrados e formular estratégias e políticas que possam ajudar a melhorar o quadro atual (CASTRO, 2018).

Somente com uma educação de qualidade para ter um desenvolvimento pessoal e comunitário, fazendo com que busque a realização dos objetivos pessoais e, com isso, alcançar o melhor do seu potencial como seres humanos, onde governos devem garantir gratuitamente uma educação de qualidade para todos os cidadãos (SILVIA, 2016).

Nos dias atuais, a adoção de normas internacionais de certificação de qualidade tem se tornado cada vez mais comum, uma vez que é crucial que a qualidade esteja presente em todos os produtos e serviços oferecidos. No campo da educação, a exigência por qualidade também tem se intensificado, bem como conceitos e padrões de qualidade estão sendo incorporados. É importante considerar a eficiência dos sistemas educacionais e, para isso, existem indicadores disponíveis para análise concreta da qualidade na educação. Com o intuito de coletar dados sobre o desempenho dos alunos, foi criado em 1990 o Sistema Nacional de Avaliação da Educação Básica (SAEB), que realiza avaliações ao final de cada ciclo de estudo (NEUHAUS, 2016).

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) realiza, por meio do SAEB, um diagnóstico da Educação Básica no Brasil, com o objetivo de identificar possíveis fatores que possam afetar o desempenho dos estudantes. A cada dois anos, são aplicados testes e questionários nas escolas públicas e em uma amostra das escolas privadas, que fornecem informações contextuais relevantes com base na avaliação dos estudantes.

O SAEB permite que as escolas avaliem a qualidade da educação que estão oferecendo aos alunos. Os resultados da avaliação são indicadores da qualidade do ensino, fornecendo evidências para monitorar e aprimorar políticas educacionais. O desempenho dos estudantes, avaliado por meio de médias, taxas de aprovação, reprovação e outros indicadores, é usado para compor o Índice de Desenvolvimento da Educação Básica - IDEB (INEP, 2020).

As taxas de aprovação e reprovação são fornecidas pela escola, através do da Informatização da Secretaria da Educação (ISE), um sistema que permite aos secretários, diretores e Coordenadorias de Educação a realizarem consultas em tempo real a todas as informações referentes à escola e aos alunos (EDUCAÇÃO, 2021).

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) utiliza o Sistema de Avaliação da Educação Básica (SAEB) e o indicador de taxa de aprovação do Censo Escolar, que é disponibilizado pelas escolas por meio do sistema ISE, para criar o Índice de Desenvolvimento da Educação Básica (IDEB). Esse índice tem como objetivo mensurar o desempenho do sistema educacional brasileiro (INEP, 2020).

Para promover uma mobilização pela qualidade da educação, é necessário estabelecer metas para o sistema de ensino, de forma que o país possa alcançar níveis comparáveis aos de países desenvolvidos. Nesse sentido, o IDEB tem sido utilizado como um indicador para incentivar ações tanto nas escolas quanto no sistema de ensino como um todo, embora se trate de um indicador de resultados e não de qualidade propriamente dita. Ainda assim, a divulgação dos resultados pode mobilizar ações que visem a melhoria da qualidade da educação (CHIRINEA, 2015).

Um dos recursos disponíveis para aprimorar a qualidade da educação é o uso de técnicas de mineração de dados. Essas técnicas podem explorar volumes de dados provenientes de registros contínuos de informações dos alunos, possibilitando a descoberta de novos conhecimentos e descobertas que poderão contribuir para a melhoria da qualidade do ensino. Com isso, a mineração de dados pode fornecer informações além de relatórios e gerenciamento, ampliando as possibilidades de aprimoramento da educação.

2.2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (KDD)

A aprendizagem de máquina vem sendo amplamente utilizada em diversas áreas devido à grande quantidade de dados gerados e à possibilidade de descobrir conhecimentos por meio da análise desses dados. Anteriormente, a transformação de dados em conhecimento era realizada por meio de análise e interpretação manual, o que era lento, caro e altamente subjetivo, tornando-se inviável em muitos domínios.

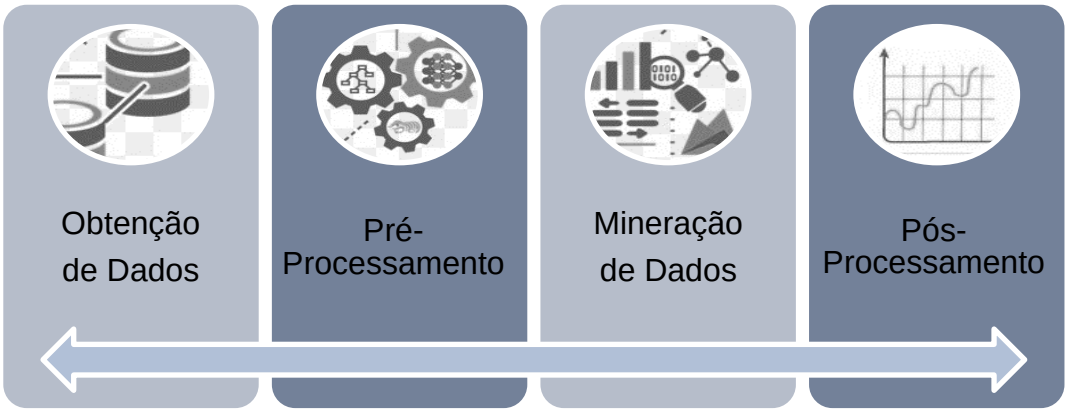
Com a geração de milhões de registros contendo dezenas ou centenas de campos tornou-se impraticável para os seres humanos realizar esse trabalho de análise. Assim, a automação total ou parcial desse processo se concretizou necessária. Esses dados são obtidos para que as empresas possam obter vantagem competitiva, aumentando a eficiência e melhorando a oferta de serviços. Além disso, são utilizados em outras áreas, como ciências, negócios, finanças, saúde, varejo, entre outras, vistas que a era da informação digital nos trouxe uma sobrecarga de dados, e a Descoberta de Conhecimento em Bases de Dados (do inglês *Knowledge Discovery in Databases* -KDD) é uma tentativa de lidar com isso (FAYYAD, 1996).

Em 1995, na conferência internacional sobre KDD, foi decidido que a terminologia *descoberta de conhecimento em bases de dados* seria usada para se referir a todo o processo de extração de conhecimento a partir de dados. Esse processo envolve a descoberta de conhecimentos úteis a partir de dados que, com a evolução da tecnologia, aumentarão consideravelmente, com milhares de novas informações sendo armazenadas em repositórios de dados, muitas vezes, sendo *esquecidas* lá.

O processo de KDD começa com a organização dos dados relevantes para a descoberta do conhecimento, que são submetidos a vários procedimentos de pré-processamento. A Figura 3 apresenta as fases que constituem “o processo de KDD”, conforme mencionado por Silva, Peres e Boscarioli (2016, p. 7). Esse processo envolve a organização das bases de dados em um repositório único, a eliminação de instâncias repetidas ou discrepantes, a seleção dos dados relevantes para a mineração de dados (MD) e a normalização dos dados. Em seguida, a tarefa de MD é iniciada, envolvendo tarefas como predição, agrupamento ou associação. Por fim, os resultados são validados, avaliados e formatados em gráficos, tabelas e relatórios estruturados.

De acordo com Castro (2016, p. 26), o processo KDD é composto por quatro etapas fundamentais. A primeira etapa envolve a criação da base de dados, que consiste em uma coleção de informações referentes a um conjunto de itens. Em seguida, a etapa de pré-processamento dos dados é realizada para preparar as informações para uma análise eficiente e eficaz. Na terceira etapa, a mineração de dados é aplicada por meio de algoritmos capazes de extrair conhecimento relevante após o pré-processamento dos dados. Por fim, a última etapa consiste na avaliação ou validação do conhecimento adquirido, com o objetivo de identificar informações úteis e não triviais (Figura 3).

Figura 3 – Processo de KDD



Fonte: Adaptada de SILVA, PERES e BOSCARIOLI (2016 p. 7).

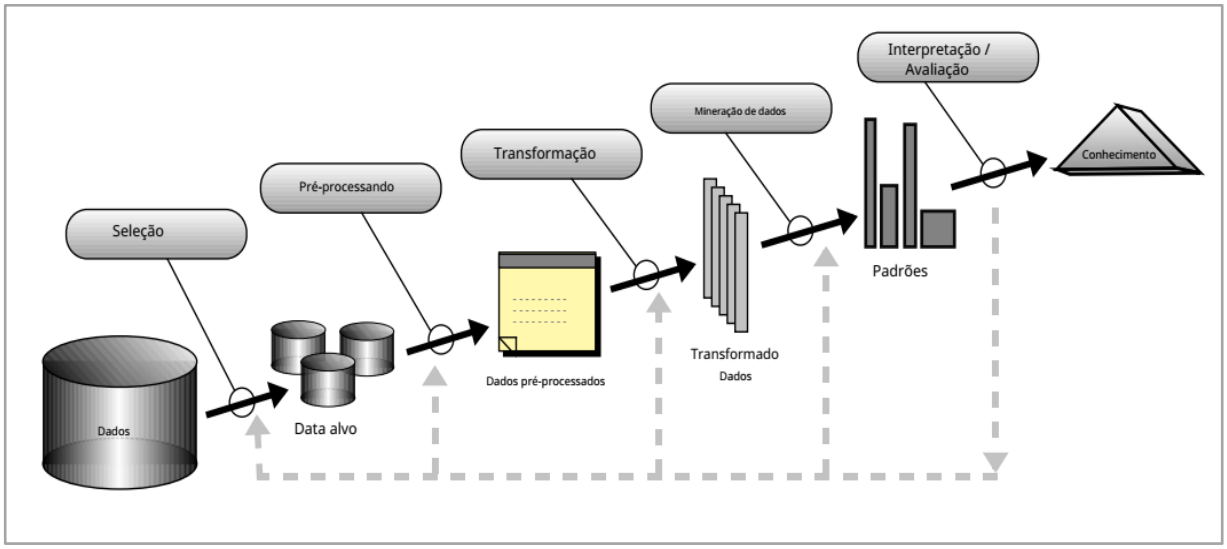
De acordo com Fayyad et al. (1996), o processo de KDD é um processo iterativo e interativo que envolve várias decisões do usuário, dividido em nove etapas:

1. Compreender o domínio da aplicação e a meta do processo do ponto de vista do cliente;
2. Selecionar um conjunto de dados ou focar em um subconjunto de variáveis para o processo de descoberta;
3. Realizar pré-processamento de dados, incluindo limpeza e eliminação de ruído, e decidir qual estratégia usar para lidar com dados ausentes;
4. Encontrar características úteis nos dados por meio de redução ou transformação, reduzindo o número efetivo de variáveis;
5. Combinar os objetivos do processo (etapa 1) com um método de mineração de dados, como classificação, regressão, agrupamento, entre outras;

6. Escolher o algoritmo de mineração de dados e o método a ser usado para buscar padrões nos dados;
7. Realizar a mineração de dados em um conjunto de representações, incluindo regras de classificação, árvores de decisão, regressão e agrupamento, auxiliando o método de mineração de dados a executar corretamente as etapas anteriores;
8. Interpretar os padrões minerados, com a possibilidade de retornar a qualquer etapa anterior para iteração adicional, e visualizar os padrões, modelos extraídos ou dados;
9. Agir sobre o conhecimento descoberto, usando-o em outro sistema para ação ou documentando e gerando relatórios. Também é possível verificar e resolver potenciais conflitos com o conhecimento obtido.

A Figura 4 mostra o processo geral de KDD, que inclui a avaliação e a possível interpretação dos dados minerados para determinar quais podem ser considerados novos conhecimentos. Também inclui todas as etapas adicionais descritas nas seções seguintes.

Figura 4 – Visão geral das etapas que compõem o processo de KDD



Fonte: Fayyad et al. (1996 p. 21).

2.2.1 Seleção de dados

Nesta etapa, realiza-se a seleção dos dados, que podem ser obtidos de diversas estruturas e armazenamentos, tais como bancos de dados, planilhas

eletrônicas, data *warehouses*, arquivos de *log*, *data streams*, dados da *web*, entre outros. Esse processo requer um bom tempo devido à necessidade de compreender não só o tipo de dados armazenados em uma tabela, mas também as relações existentes com outras tabelas, além de exigir conhecimento específico do banco de dados em questão. Um fator crucial para que a modelagem de dados obtenha sucesso é assegurar a qualidade dos dados selecionados.

2.2.2 Pré-processamento

Após concluir a etapa anterior de separação dos dados pertencentes a um determinado domínio, a etapa de pré-processamento é essencial para limpar e tornar os dados confiáveis e consistentes, a fim de melhorar sua qualidade. Durante esta etapa, os dados são analisados e selecionados de acordo com os objetivos do usuário.

O pré-processamento, que também é conhecido como preparação da base de dados, é crucial para limpar e transformar os dados de uma base selecionada, a fim de extrair conhecimento de maneira mais fácil e precisa, o que permite a aplicação das técnicas de mineração (CASTRO, 2016, p. 54).

A qualidade dos dados depende da limpeza realizada, que pode envolver a remoção de atributos desnecessários para o objetivo do usuário. Outra atividade importante é a padronização dos dados faltantes ou inconsistentes, a fim de torná-los viáveis para a mineração de dados. Isso é especialmente importante, pois os dados podem estar em formatos diferentes na base de dados (CRETTON, 2016).

De acordo com Castro (2016), as principais tarefas de pré-processamento são:

> - Limpeza: essa etapa envolve a imputação de valores ausentes, remoção de ruídos e correção de inconsistências para garantir que o processo de KDD não seja comprometido.
> - Integração: em aplicações do mundo real, os dados podem estar distribuídos em estruturas distintas, o que requer a concatenação de todos os dados de múltiplas fontes em um único local, como um armazém de dados (*data warehouse*).
> - Redução: a ampliação do número de objetos e atributos na base de dados pode tornar as medidas matemáticas usadas na análise numericamente instáveis e muitas vezes tornar o processo dos algoritmos de mineração muito complexo. Métodos de redução são necessários para manter a integridade dos dados originais. Agrupar ou eliminar atributos redundantes reduz a dimensão da base de dados ou sumariza para reduzir a quantidade de objetos na base.
> - Discretização: essa tarefa é útil para permitir que métodos que trabalham apenas com atributos nominais sejam aplicados a um conjunto

> maior de problemas. Também reduz a quantidade de valores para um determinado atributo contínuo (CASTRO, 2016, p. 55).

Em resumo, a etapa de pré-processamento é fundamental para garantir a qualidade dos dados e permitir uma análise mais precisa e eficiente por meio das técnicas de mineração.

2.2.3 Transformação dos dados

A etapa de pré-processamento de dados, frequentemente, resulta em complexidade nos dados que requerem transformação para consolidar conjuntos de informações relevantes. As bases de dados brutas e integradas de diferentes fontes podem apresentar problemas, como valores ausentes, ruídos e inconsistências de dados não padronizados. Esses problemas são resolvidos por meio da padronização dos dados, os quais, muitas vezes, requerem conversão de dados e correções de diferenças de unidades e escalas (CASTRO, 2016, p. 70).

Durante a transformação de dados, é possível realizar tratamentos, ajustar conflitos de tipificação e remover ruídos ou conteúdos irrelevantes ao objetivo. Segundo Fayyad et al. (1996), nesta etapa, é possível encontrar características úteis para representar os dados ou representações irrelevantes ao objetivo.

2.2.4 Mineração de Dados

O autor Costa (2012) destaca que a Mineração de Dados (MD) é a fase crucial do processo amplo conhecido como Descoberta de Conhecimento em Bases de Dados (KDD). Durante essa etapa, novos conhecimentos e padrões podem ser extraídos por meio da aplicação de algoritmos específicos que trabalham com os dados.

Para atingir seus objetivos, diversas tarefas, também conhecidas como métodos, podem ser utilizadas. Cada tarefa é única e se aplica a diferentes tipos de padrões, exigindo abordagens e métodos distintos. Conforme Fayyad et al. (1996), as tarefas mais comuns em MD incluem sumarização, classificação, agrupamento, associação e regressão. Na próxima seção, apresentaremos mais informações sobre mineração de dados e essas tarefas em particular.

2.2.5 Interpretação e avaliação

A última etapa do processo KDD tem como objetivo selecionar modelos válidos e úteis para a tomada de decisões. Nessa etapa, os resultados da mineração de dados são interpretados e avaliados para verificar se é possível obter respostas compreensíveis dos algoritmos para a tomada de decisão.

Além disso, essa etapa envolve a apresentação dos resultados da mineração de dados aos usuários por meio de várias estratégias de visualização e GUI (*Graphical User Interface*), como descrito por Fayyad (1996). O objetivo é tornar as informações claras e fáceis de entender para que os usuários possam tomar decisões com base nos *insights* obtidos a partir dos dados minerados.

Em resumo, a última etapa do processo KDD é crucial para garantir que os resultados da mineração de dados sejam apresentados de maneira clara e útil, permitindo que os usuários tomem decisões informadas com base nas informações obtidas.

2.3 MINERAÇÃO DE DADOS

Na década de 1990, a Mineração de Dados emergiu como uma área independente de pesquisa e aplicação, com raízes na matemática, estatística e computação. Sua importância cresceu com o advento do termo *Big Data* e a publicação do relatório *Big Data: The Next Frontier for Innovation, Competition, and Productivity* pelo *McKinsey Global Institute* em 2011 (CASTRO, 2016). A Mineração de Dados é a etapa principal do processo KDD, pois é nessa etapa que novos padrões e conhecimentos são descobertos. Ela envolve a aplicação de algoritmos específicos para a extração desses padrões a partir dos dados.

A MD é necessária devido à dificuldade de interpretar grandes quantidades de dados armazenados, o que leva a excesso de dados desorganizados e sem utilidades. “Esses são conjuntos de dados extensos que não podem ser analisados manualmente devido à presença de muitos registros, atributos, ausência de valores e dados qualitativos e não quantitativos” (CASTRO, 2016, p. 25).

Silva, Peres e Boscarioli (2016, p. 7) define a MD como "encontrar o que era desconhecido, o que estava escondido", o que implica descobrir conhecimento implícito em grandes volumes de dados por meio de uma busca detalhada, um

processo de "mineração" (SILVA; PERES; BOSCARIOLI, 2016, p. 7). A MD é um processo que analisa grandes conjuntos de dados de forma automática ou semiautomática, com o objetivo de descobrir padrões relevantes e importantes para gerar conhecimento.

Na busca por padrões, previsões, erros, associações e outros *insights* em grandes volumes de dados associados à aprendizagem de máquina, a inteligência artificial e algoritmos especializados são capazes de identificar padrões que técnicas triviais de análise ou o "olho nu" não conseguiriam (AMARAL, 2016, p. 2).

O termo MD é uma analogia ao processo de extração de minerais valiosos, onde uma base de dados é explorada usando algoritmos adequados para obter conhecimento valioso. Os dados podem ser apresentados em uma tabela sem significado ou estrutura, e as informações estão contidas nas descrições, dando utilidade aos dados e agregando significado (CASTRO, 2016, p. 25).

De acordo com Castro (2016, p. 26):

> A mineração de dados é parte integrante de um processo mais amplo, conhecido como descoberta de conhecimento em bases de dados (*knowledge discovery in databases*, ou KDD).

O conhecimento é fundamental para a tomada de decisões com o objetivo de agregar valor. A Mineração de Dados (MD) é uma ferramenta que permite extrair informações úteis da base de dados e validar esses conhecimentos sob diferentes perspectivas. Existem tarefas que podem ser classificadas como descritivas, que caracterizam as propriedades gerais dos dados, e preditivas, que fazem deduções a partir dos dados para realizar predições (CASTRO, 2016, p. 27).

Silva, Peres e Boscarioli (2016, p. 9) apresenta “uma taxonomia de dois níveis para as tarefas de MD. No primeiro nível, as tarefas são divididas em descritivas e preditivas”. As tarefas preditivas usam atributos descritivos para predizer valores futuros ou desconhecidos, enquanto as tarefas descritivas têm como objetivo encontrar padrões e descrever informações que possam ser interpretadas pelo ser humano.

No segundo nível, as tarefas de MD são especializadas em descritivas e preditivas. “As tarefas preditivas incluem classificação e regressão, enquanto as tarefas descritivas incluem agrupamento, sumarização, modelagem de dependências e detecção de desvios” (SILVA; PERES; BOSCARIOLI, 2016, p. 9).

Existem diversas opções disponíveis para a extração em mineração de dados, incluindo ferramentas comerciais e de código aberto. “Algumas das ferramentas comerciais mais populares são fornecidas pela Microsoft, SAS, IBM e Oracle. Enquanto isso, algumas das opções de código aberto mais conhecidas incluem o *WEKA* e Orange” (AMARAL, 2016, p. 04).

O *WEKA* (*Waikato Environment for Knowledge Analysis*) é uma ferramenta de código aberto que tem sido desenvolvida desde o início dos anos 90. Com vários algoritmos disponíveis para executar uma ampla variedade de tarefas, “essa ferramenta possui uma interface gráfica fácil de aprender e usar, dispensando a necessidade de digitar códigos” (AMARAL, 2016, p. 05).

As principais tarefas utilizadas para a mineração de dados são citadas a seguir conforme o autor Fayyad (1966):

2.3.1 Classificação

A classificação é uma técnica que utiliza uma base de dados para aprender uma função capaz de mapear e classificar classes. A partir dessa função, é possível construir um modelo que, com base em um conjunto de dados de entrada, é capaz de realizar previsões para novos registros. Essa técnica também conhecida como predição, tem sido amplamente utilizada em diversas áreas para solucionar problemas de classificação e identificação de padrões.

O Romero (2010 p. 7) ressalta que:

> A classificação é um procedimento no qual itens individuais são colocados em grupos com base em informações quantitativas sobre uma ou mais características inerentes aos itens e com base em um conjunto de treinamento de itens previamente rotulados.

A classificação é uma das tarefas mais tradicionais na identificação da classe, a qual um registro pertence. Essa tarefa começa com a indução a partir de um conjunto de registros fornecidos pelo classificador, que contém indicação da classe, a que pertencem. O objetivo é *induzir* como classificar um novo registro (MARTINS, 2017).

Gonçalves (2018) também destaca o objetivo da classificação, que é identificar a classe de novos dados, tentando aprender a generalizar um conceito a partir de uma base de dados com um ou mais atributos preditivos e um atributo classe.

A tarefa de classificação é frequentemente utilizada para classificar perfis de alunos, estilos de aprendizagem e previsão de desempenho. “Os algoritmos mais comuns são: árvores de decisão, máquinas de vetores de suporte, *naive bayes* e redes neurais” (BARROS, 2020, p. 2526).

2.3.2 Regressão

Enquanto na classificação a predição é feita para um atributo classificador que assume valores discretos, nos modelos de regressão a variável alvo “é contínua, ou seja, associa um item de dado a uma ou mais variáveis preditoras que assumem valores reais” (COSTA, 2012, p. 5). Embora semelhante à classificação, a regressão busca prever um valor numérico, e a avaliação do desempenho é baseada na diferença entre o valor previsto e o valor real dos dados históricos (AMARAL, 2016, p. 53).

2.3.3 Agrupamentos

A técnica de MD conhecida como análise de agrupamento tem como objetivo identificar e aproximar registros similares. “Trata-se de uma tarefa de aprendizado de máquina não supervisionada, muito utilizada para classificar as instâncias de acordo com os grupos encontrados” (AMARAL, 2016, p. 101). Os algoritmos mais comuns são o *K-means*, DBSCAN e Hierárquico.

“Essa técnica procura associar um item de dado com um ou vários agrupamentos determinados pelos dados, utilizando principalmente medidas de similaridade” (COSTA, 2012, p. 5). “A análise de agrupamento tem como objetivo dividir o conjunto de dados em grupos homogêneos, maximizando a similaridade dos dados e minimizando aqueles que estão fora do grupo” (SILVA, 2018, p. 26).

2.3.4 Regras de Associação

A tarefa de associação tem como objetivo identificar padrões que relacionam características ou eventos em uma base de dados. Através da análise de subconjuntos de características ou regras de implicação, é possível identificar padrões relevantes que ocorrem com frequência. O desafio dessa tarefa é extrair esses

padrões de forma ágil, dada a enorme quantidade de possibilidades de combinações existentes na base de dados. “Um exemplo prático de aplicação dessa tarefa é o reconhecimento de páginas *web* acessadas simultaneamente” (SILVA, 2018, p. 26). Algoritmos como Apriori, GSP e DHP são exemplos de técnicas utilizadas na tarefa de associação (ZAKI, 2000).

2.4 MINERAÇÃO DE DADOS EDUCACIONAIS (MDE)

A Mineração de Dados Educacionais (*Education Data Mining*) é uma subárea da Mineração de Dados que se concentra em desenvolver métodos para extrair conjuntos de dados coletados em ambientes educacionais. É uma área de pesquisa que combina conhecimentos de Computação, Educação e Estatística, e tem três subáreas principais: *E-learning*, *Data Mining* e *Machine Learning*, além de *Learning Analytics*, que está intimamente relacionada à MDE (ALVES, 2018).

Baker, Carvalho e Isotani (2011, p. 4), define MDE como:

> [...] “Mineração de Dados Educacionais” (do inglês, “*Educational Data Mining*”, ou EDM). A EDM é definida como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais.

A MDE procura não só utilizar métodos já existentes, mas também adaptar algoritmos de mineração para melhor compreender os dados coletados em contextos educacionais. Esses dados são produzidos em ambientes educacionais, principalmente por estudantes e professores que interagem por meio de Ambientes Virtuais de Aprendizagem (AVAs), Sistemas Tutores Inteligentes (STIs) e outros recursos (COSTA, 2012).

Segundo Romero (2010), esses dados coletados dos alunos podem ser utilizados para adquirir novos conhecimentos, ajudar a validar melhorias em aspectos da qualidade da educação e estabelecer as bases para um processo de aprendizagem mais eficaz. Ideias semelhantes já foram aplicadas com sucesso em sistemas de *e-commerce*, os quais utilizam dados para determinar os interesses dos clientes e, assim, aumentar as vendas.

Atualmente, há um aumento no interesse pela aplicação de MD no ambiente educacional. No entanto, essa aplicação envolve algumas questões importantes que a diferenciam de outras áreas, tais como:

- Objetivo: Em cada área de aplicação, o objetivo da MD é diferente. Por exemplo, nos negócios, o principal objetivo é aumentar o lucro, o que pode ser medido em termos tangíveis, como dinheiro e número de clientes. Na área educacional, o objetivo é realizar uma pesquisa aplicada ao processo de aprendizagem, buscando responder questões práticas sobre como melhorar e orientar a aprendizagem dos alunos, bem como objetivos de pesquisa pura, a fim de compreender mais profundamente os fenômenos educacionais. Esses objetivos podem ser difíceis de quantificar e, portanto, requerem técnicas de medição especiais.
- Dados: No ambiente educacional, há vários tipos de dados disponíveis para mineração, que contêm informações essenciais sobre a área educacional, bem como relacionamentos com outros dados e hierarquias significativas em vários níveis. Além disso, é importante levar em consideração os aspectos pedagógicos do aluno e do sistema.
- Técnicas: Os dados e problemas educacionais possuem características especiais que exigem uma abordagem diferente em relação à questão da mineração. Embora a maioria das técnicas tradicionais de MD possam ser aplicadas diretamente, algumas precisam ser adaptadas ao problema educacional específico em questão. Além disso, existem técnicas específicas de mineração de dados que podem ser utilizadas para problemas educacionais específicos.

A Mineração de Dados Educacionais (MDE) busca compreender de forma mais profunda os fenômenos educacionais, incluindo o processo de aprendizagem dos alunos, por meio de pesquisas aplicadas. Essas pesquisas são realizadas em ambientes computacionais de ensino e conteúdos digitais, nos quais as Tecnologias da Informação e Comunicação (TIC) são essenciais para a condução dos processos educacionais.

Com o crescimento dos Ambientes Virtuais de Aprendizagem (AVA) e dos Sistemas de Informações Educacionais (SIE), há um grande volume de dados gerados pelos atores do sistema educacional, como professores, alunos e gestores. Esses dados não se limitam a informações gerenciais ou relatórios de desempenho, mas podem proporcionar novos conhecimentos relevantes por meio da aplicação de técnicas de MDE. Essas técnicas convertem os dados brutos em informações úteis

que podem ser usadas por desenvolvedores de *softwares* educacionais, professores, pesquisadores educacionais e outross profissionais (RODRIGUES, 2014).

O processo de MDE não difere muito de outras áreas de aplicação de mineração de dados, uma vez que se baseia nos mesmos passos do processo de mineração de dados em geral.

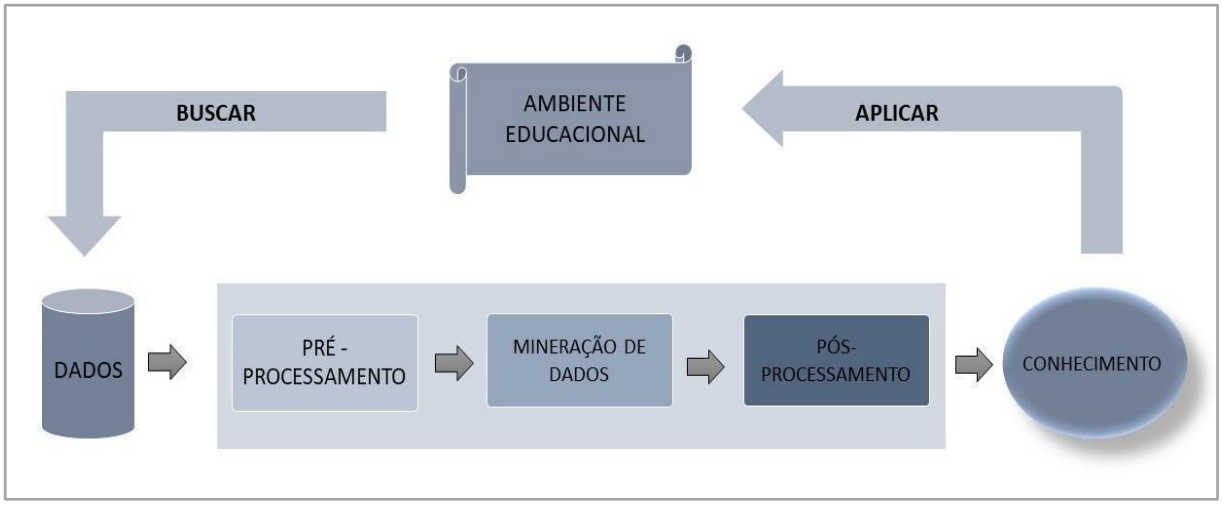
Conforme o autor GARCIA et al. (2011), Figura 5, para realizar a mineração de dados no ambiente educacional, é necessário seguir um processo que envolve três etapas principais: pré-processamento, mineração de dados e pós-processamento.

No pré-processamento, os dados são coletados do ambiente educacional e transformados em um formato adequado para a mineração. Algumas tarefas de pré-processamento incluem limpeza dos dados, seleção e transformação de atributos, integração de dados e outras tarefas necessárias para garantir a qualidade dos dados a serem analisados.

Na etapa de mineração de dados, as técnicas apropriadas são aplicadas para analisar os dados previamente processados. Existem várias técnicas de mineração de dados disponíveis, incluindo visualização, regressão, classificação, agrupamento, mineração de regras de associação, mineração de padrões sequenciais e mineração de texto. A escolha da técnica adequada depende do objetivo específico da análise.

Por fim, na etapa de pós-processamento, os resultados obtidos são interpretados e utilizados para tomar decisões sobre o ambiente educacional. É importante que os resultados sejam apresentados de forma clara e compreensível para facilitar a tomada de decisões informadas.

Figura 5 – Etapas da Mineração de Dados Educacionais.



Fonte: Adaptado GARCIA et al. (2011).

Através deste processo, é possível descobrir novos conhecimentos com base nos dados dos alunos, a fim de melhorar a qualidade da educação, validando e avaliando o sistema educacional. Embora tenha sido aplicado com sucesso em sistemas de *e-commerce*, o progresso na aplicação da MD na educação ainda é limitado, dada as questões específicas que diferenciam sua aplicação nesse contexto (ROMERO, 2010).

O autor Baker (2011) apresenta uma taxonomia de técnicas e métodos aceitos por um grande número de pesquisadores, que podem ser utilizados em ambientes educacionais para descobrir os fatores que influenciam a aprendizagem dos alunos e compreender a forma mais eficaz e adequada de ensino.

Para Barros (2020), as técnicas de MD na educação podem ser divididas em dois tipos: preditivas ou supervisionadas, e análises descritivas ou não supervisionadas. A técnica preditiva é a mais utilizada na prática, e tem como tarefa a classificação e regressão, sendo comumente utilizada para classificar perfis de alunos, estilo de aprendizagem e previsão de desempenho.

A predição é a tentativa de descobrir o que acontecerá em algum momento futuro, usando um modelo construído a partir da base de dados como referência, chamada de base de treinamento, e que contenha um atributo especial com a classe ou valor que se deseja prever (SILVA, 2014).

A meta da técnica preditiva é desenvolver modelos que deduzam aspectos específicos dos dados, através da análise de diversos aspectos presentes nos dados. Para isso, é necessária a codificação manual de uma quantidade de dados, a fim de garantir a correta identificação das variáveis preditoras previamente conhecidas (BAKER, 2011).

> Tarefas Preditivas objetivam predizer o valor de um determinado atributo (variável) baseado nos valores de outros atributos. O atributo a ser predito é comumente conhecido como a variável preditiva, dependente ou alvo, enquanto que os atributos usados para fazer a predição são conhecidos com as variáveis preditoras, independentes ou explicativas (COSTA et al., 2012, p. 3).

2.5 FERRAMENTA WEKA

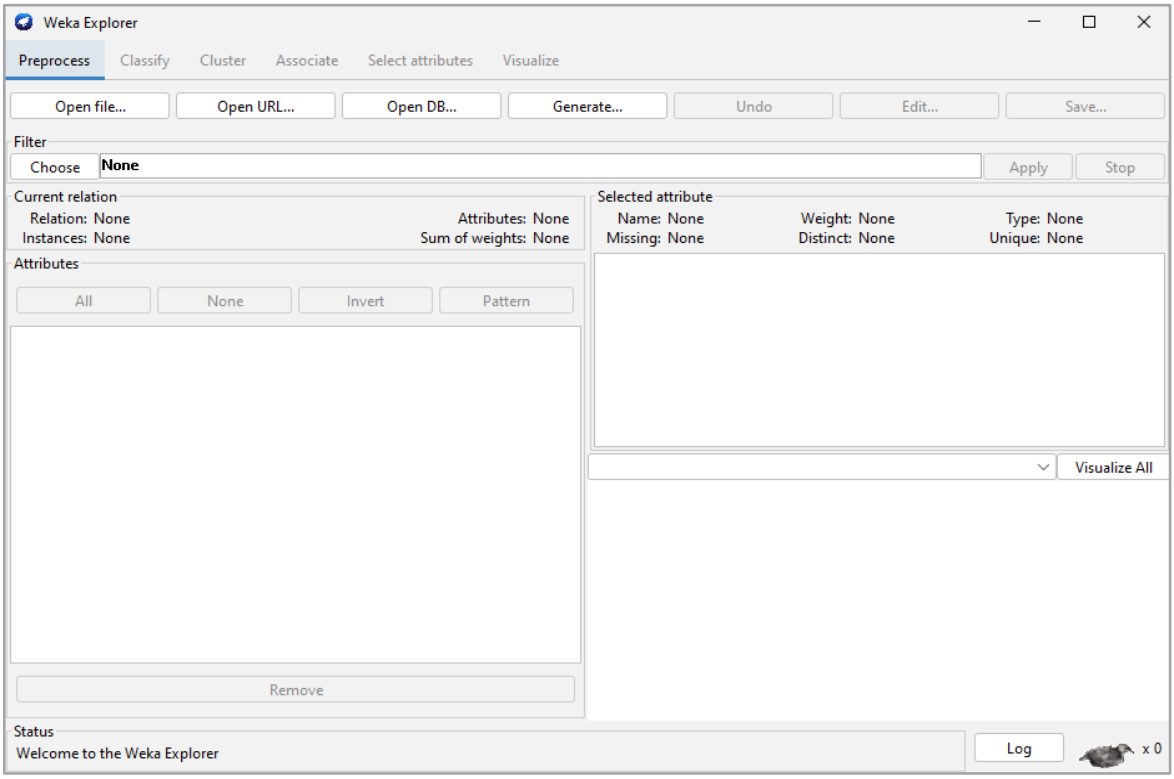
O *WEKA*, acrônimo para *Waikato Environment for Knowledge Analysis*, é uma ferramenta gratuita de código aberto desenvolvida em Java pela Universidade

Waikato, na Nova Zelândia. Com uma interface gráfica intuitiva, o *software* permite aos usuários realizar uma variedade de tarefas, incluindo pré-processamento, classificação, regressão, agrupamento, associação e visualização de dados (CASTRO, 2016, p. 369).

Além de uma ampla variedade de algoritmos de aprendizado de máquina, o *WEKA* também oferece a possibilidade de criar algoritmos personalizados. Na Figura 6, podemos observar os principais algoritmos disponíveis na ferramenta, tais como Classificação (*Classify*), Agrupamento (*Cluster*), Associação (*Associate*) e Seleção de atributos (*Select attributes*).

Para utilizar as técnicas de mineração de dados oferecidas pelo *WEKA*, é necessário que os dados estejam organizados em uma estrutura de dados, planilha ou banco de dados. Com suas diversas técnicas e algoritmos, o *WEKA* é uma excelente opção para análise de dados e desenvolvimento de novos sistemas de aprendizado de máquina.

Figura 6 – Interface gráfica do *WEKA*.



Fonte: *WEKA* (2022).

Para organizar dados no *WEKA*, é possível utilizar o formato de arquivo próprio do software, chamado *Attribute-Relation File Format*, que tem a extensão “.arff”. Esse formato é amplamente utilizado em aprendizado de máquina e consiste em duas seções principais: um cabeçalho com metadados e uma seção com dados separados por vírgula. O arquivo está em formato *ASCII* e pode ser lido em qualquer editor de texto (AMARAL, 2016, p. 14).

No cabeçalho, é necessário definir o nome do conjunto de dados utilizando a marca @RELATION. Os atributos são informados utilizando a marca @ATTRIBUTE e os dados são listados na seção @DATA, em linhas separadas por vírgulas e na mesma ordem definida nos atributos.

“Integrando as bibliotecas do *WEKA* em um ambiente de desenvolvimento Java, como NetBeans ou Eclipse, é possível personalizar ainda mais a mineração de dados” (CASTRO, 2016, p. 369).

3 TRABALHOS RELACIONADOS

Com a definição do tema de pesquisa, realizou-se uma Revisão Sistemática da Literatura para obter informações atualizadas sobre o estado da arte, fornecendo uma base sólida para este trabalho. Neste capítulo, serão apresentados os resultados obtidos a partir da revisão de trabalhos relacionados ao tema.

A pesquisa foi motivada pela análise dos dados do Ideb da escola GAM, que mostrou que, em alguns anos, as séries finais não atingiram as metas estabelecidas pelo governo. Para enfrentar esse desafio, a MDE foi usada para analisar os dados gerados pelos Sistemas Educacionais e encontrar soluções para melhorar o desempenho dos alunos.

Para realizar o estudo, foi realizado um conjunto de pesquisas acadêmicas em livros e no Google Acadêmico, bem como no portal do Simpósio Brasileiro de Informática na Educação (SBIE). Foram estabelecidos critérios para selecionar artigos publicados entre 2015 e 2021 sobre o tema em foco, incluindo *Mineração de Dados, Técnicas de Mineração de Dados, Mineração de Dados Educacionais, Dados Educacionais, Ideb e qualidade na educação*, utilizando diferentes técnicas e ferramentas de mineração de dados.

3.1 DESCRIÇÃO DOS TRABALHOS

No artigo de Barros et al. (2020), os autores investigam a predição do rendimento dos alunos em lógica de programação. Eles propõem o uso de técnicas de MDE para prever o desempenho dos alunos na disciplina de Lógica de Programação, buscando estabelecer uma relação entre o desempenho nas disciplinas dos primeiros períodos dos alunos em 2017 e 2018 e o desempenho subsequente na disciplina em questão. Para isso, foram empregados algoritmos de aprendizado de máquina em *Python*, implementados pela biblioteca *scikit-learn*. Ao analisar os resultados, Barros et al. (2020) indicam que é possível prever o desempenho dos alunos na disciplina de Lógica de Programação com base em suas notas no primeiro período.

Já Souza (2020) realizou um estudo que analisou as contribuições dos métodos de MDE (Mineração de Dados Educacionais) sobre dados gerados no Curso de Esportes e Atividades ao ar Livre, oferecido pela Universidade Federal do Rio Grande

do Sul (UFRGS) na plataforma de cursos *online* MOOCS. O estudo contou com a participação de 702 alunos matriculados, que responderam a um questionário composto por cinco a sete questões de múltipla escolha, baseadas nas videoaulas do curso.

Para realizar a análise dos dados, a autora utilizou a metodologia KDD (*Knowledge Discovery in Databases*), que é amplamente utilizada em pesquisas acadêmicas. Na etapa de mineração de dados, foram aplicados algoritmos de árvore de decisão e agrupamento, que são comuns na área da educação. A aplicação desses algoritmos permitiu a identificação de alguns padrões de comportamento dos alunos, especialmente daqueles com baixo engajamento.

Os resultados da análise revelaram que um grupo de alunos apresentava baixo nível de engajamento no curso. Com base nesses resultados, a autora propôs um modelo de predição baseado em regras, que pode ser utilizado para prever o comportamento de novos alunos. Isso pode ajudar os professores e gestores do curso a identificar e oferecer maior atenção aos alunos que apresentam baixo engajamento.

Em resumo, o estudo de Souza (2020) demonstrou como os métodos de MDE podem ser aplicados para melhorar a qualidade dos cursos *online*, especialmente no que diz respeito ao engajamento dos alunos. A proposta de um modelo de predição baseado em regras pode ser uma ferramenta útil para ajudar os professores e gestores a identificar e oferecer suporte aos alunos com baixo engajamento.

No estudo realizado por Colpo et al. (2020), é apresentada uma Revisão Sistemática da Literatura (RSL) sobre o uso de técnicas de MDE para prever a evasão escolar no contexto de pesquisas brasileiras. A análise foi baseada em trabalhos publicados no Congresso Brasileiro de Informática na Educação (CBIE) e teve como objetivo entender o cenário de pesquisa nacional sobre esse tema.

Os autores examinaram as características contextuais, técnicas e dados que foram abordados nos estudos e identificaram lacunas na exploração dessas informações. Eles mencionaram outros autores que já realizaram um mapeamento sistemático sobre o uso de MDE no Brasil para identificar as causas da evasão escolar. A maioria dos trabalhos analisados utilizou algoritmos de árvore de decisão por serem facilmente interpretáveis. A ferramenta mais utilizada foi o *software WEKA*, seguido pela linguagem R.

Em resumo, o estudo de Colpo et al. (2020) fornece uma visão geral do uso de MDE para prever a evasão escolar em pesquisas brasileiras e destaca a necessidade

de explorar mais profundamente as características contextuais e os dados relevantes para essa área. Os resultados também sugerem a importância de desenvolver algoritmos mais sofisticados para melhorar a precisão das previsões.

No estudo realizado por Noetzold et al. (2021), foram analisados padrões de evasão escolar no Ensino Superior usando dados do Curso de Sistemas de Informação da Universidade Federal de Santa Maria (UFSM). A pesquisa incluiu a utilização do *software WEKA* para minerar dados em cinco etapas: seleção de dados, pré-processamento e limpeza, transformação de dados, aplicação de algoritmos de mineração de dados e interpretação dos resultados. O objetivo foi identificar padrões de evasão entre alunos e investigar aspectos importantes da MD. Os dados foram coletados de cinco planilhas que continham informações de todos os alunos com vínculo entre 2015 e 2019. Os dados considerados incompletos ou irrelevantes foram eliminados durante o pré-processamento e limpeza.

Ramos et al. (2020) desenvolveram um trabalho teórico que propõe uma adaptação do modelo de mineração de dados CRISP-DM (Acrônimo de *CRoss-Industry Standard Process for Data Mining*) para a área da educação, criando assim o modelo CRISP-EDM. Esse modelo foi dividido em seis fases que não são rigorosas, permitindo que os usuários avancem e retrocedam entre elas conforme necessário.

Na primeira fase, denominada *entendimento do negócio*, os problemas educacionais da instituição são levantados. Na segunda fase, *compreensão dos dados*, os dados são coletados e possíveis problemas de qualidade são identificados. Sempre que possível, é recomendado escolher uma teoria educacional já consolidada. Na terceira fase, *preparação dos dados*, os dados brutos são transformados e limpos para construir os dados finais. Essa etapa pode ser executada várias vezes e pode levar entre 50-70% do tempo e esforço do projeto.

Na quarta fase, *modelagem*, o analista de dados ajusta os parâmetros para valores otimizados, podendo retornar à etapa de preparação de dados, se necessário. Na quinta fase, *avaliação*, os resultados obtidos e os conhecimentos descobertos são utilizados para verificar como os modelos desenvolvidos com as técnicas de mineração são executados nos dados reais. Por fim, na sexta fase, *implantação do modelo*, os resultados são apresentados de forma simplificada em relatórios ou *dashboards* para proporcionar melhorias à organização.

Essa adaptação do processo CRISP-DM para o contexto de dados educacionais, denominada CRISP-EDM, possibilita a integração multidisciplinar e completa o processo original, direcionando-o para o ambiente de dados educacionais.

Silva et.al. (2019), realizaram um levantamento bibliográfico sobre MDE através da construção de um Mapeamento Sistemático da Literatura. O objetivo foi classificar e organizar os trabalhos publicados na área. Os resultados identificaram as abordagens mais exploradas, as ferramentas e os algoritmos mais utilizados. O algoritmo *J48* foi o mais empregado nos trabalhos científicos de MDE, evidenciando sua eficiência. A ferramenta mais utilizada para realizar a MDE nos trabalhos pesquisados foi a WEKA. Além disso, foi observado que um número significativo de trabalhos aplicados em ambiente acadêmico, utilizaram a mineração de dados como objetivo principal da MDE.

Para comparar com trabalhos relacionados, foi elaborada uma análise no Quadro 1, que inclui informações sobre a área de aplicação, algoritmos utilizados, ferramentas de KDD, técnicas de MD e atributos empregados. A análise revelou que a mineração de dados é aplicável à área educacional, com foco em dados de estudantes universitários. Na Revisão Sistemática de Literatura realizada por Colpo et al. (2020), observou-se que a maioria dos trabalhos utilizou algoritmos de árvores de decisão e dados acadêmicos.

Com base nessas informações, este trabalho tem como proposta aplicar a mineração de dados em alunos do ensino fundamental de escolas públicas, em particular, os alunos dos 6º e 9º anos. Os dados serão obtidos a partir do ISE da EEEF Dr Gabriel Álvaro de Miranda.

Quadro 1 – Análise dos trabalhos correlatos

Autor(a)	Área	Técnica / Algoritmo utilizado	Ferramenta de KDD	Técnica de MD	Atributos
Barros et al. (2020)	Graduação – Alunos da disciplina de Lógica de Programação	Algoritmos de aprendizado de máquina (MultinominalNB, KNeighborsClassifier, SVM, DecisionTreeClassifier)	Linguagem Python	Classificação	Dados reais de alunos (médias das disciplinas obtidas no primeiro período do curso)
Souza (2020)	Graduação -Cursos MOOCS da UFRGS	Algoritmos de árvore de decisão (*Decision Tree*) e Agrupamento (C*lustering*)	RapidMiner	Classificação e Agrupamento	Questionário com questões de múltipla escolha elaborados com conteúdo presentes nas vídeo aulas.
Colpo et al. (2020)	Revisão Sistemática de Literatura (RSL) - Graduação	Algoritmos de árvores de decisão adotados na maioria dos trabalhos	Ferramenta Weka pela maioria, seguida pela linguagem R	Os trabalhos selecionados nesta RSL consideram a tarefa de Classificação	Dados acadêmicos e sociais são amplamente utilizados, em geral, se baseiam em dados do Censo Escolar
Noetzold et al. (2021)	Graduação	*J48*	Weka	Classificação	Dados dos alunos do curso de Sistemas de Informação da UFSM, campus Frederico Westphalen
Ramos et al. (2020)	Apresentado estudos em Graduação	Não informado	Não Informado	Não Informado	Dados Educacionais
Silva et al. (2019)	Levantamento bibliográfico sobre MDE - Universitário	Mais empregado no trabalho científico de MDE - *J48*	Instrumento mais utilizado - Weka	Não Informado	Dados acadêmicos de Universidade
Paula et al. (2023)	Alunos de Escola Estadual do Ensino Fundamental no RS	*J48*	Weka	Classificação	Dados reais de alunos (notas das disciplinas, medias, bolsa família, sexo)

Fonte: Autor (2022).

4 METODOLOGIA DE PESQUISA

O método utilizado para investigar o projeto consistiu em uma abordagem quali-quantitativa dedutiva, a qual se baseou em pesquisa bibliográfica sobre o tema central. O universo da pesquisa foi a Escola Estadual de Ensino Fundamental Dr. Gabriel Álvaro de Miranda, localizada em Cruz Alta/RS, com foco nos alunos do 6º e 9º anos entre os anos de 2016 a 2019.

Foi constatado que a escola alcançou as metas projetadas pelo governo nos anos iniciais, mas somente em 2009 obteve sucesso nos resultados dos anos finais. Por esse motivo, a pesquisa concentrou-se nos alunos dos anos finais.

A abordagem utilizada é quantitativa, uma vez que os dados dos alunos foram coletados a partir do ISE da escola, e qualitativa, devido à compreensão das descobertas de indicadores que podem prever um possível mau desempenho dos alunos.

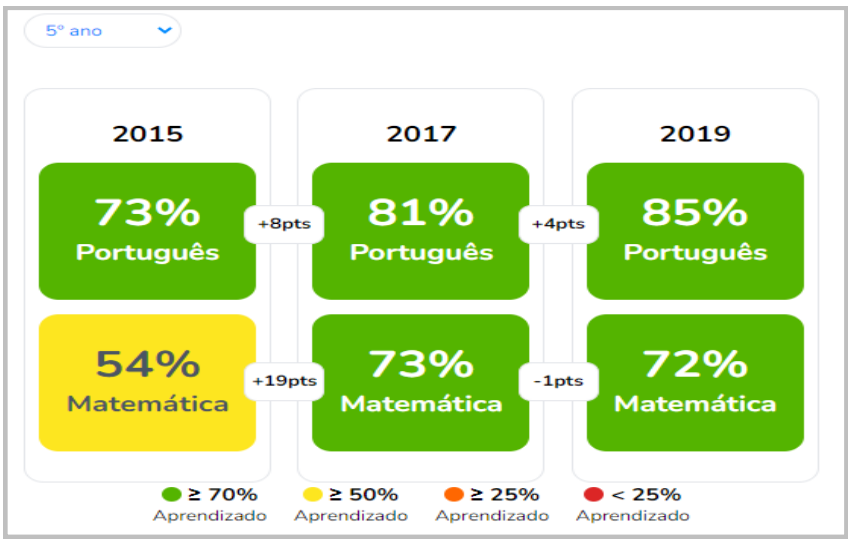
Por meio de pesquisa bibliográfica, buscou-se entender o funcionamento do sistema IDEB, que mede o desempenho do sistema educacional brasileiro. Para o desenvolvimento da pesquisa, foram analisados os índices da Escola nos anos finais do 6º e 9º anos.

De acordo com o portal QEDU, as Figuras 7 e 8 mostram a proporção de alunos com aprendizado adequado em Português e Matemática, avaliados pela Prova Brasil e pelo Sistema Nacional de Avaliação da Educação Básica (SAEB). Essas avaliações têm como objetivo medir a qualidade do ensino oferecido pelo Sistema Educacional Brasileiro.

Conforme ilustrado na Figura 7, os anos iniciais da escola (1º ao 5º ano) têm alcançado objetivos satisfatórios de aprendizagem adequada. No entanto, a Figura 8 indica que, nos anos finais da educação básica (6º ao 9º ano), a aprendizagem está abaixo do esperado.

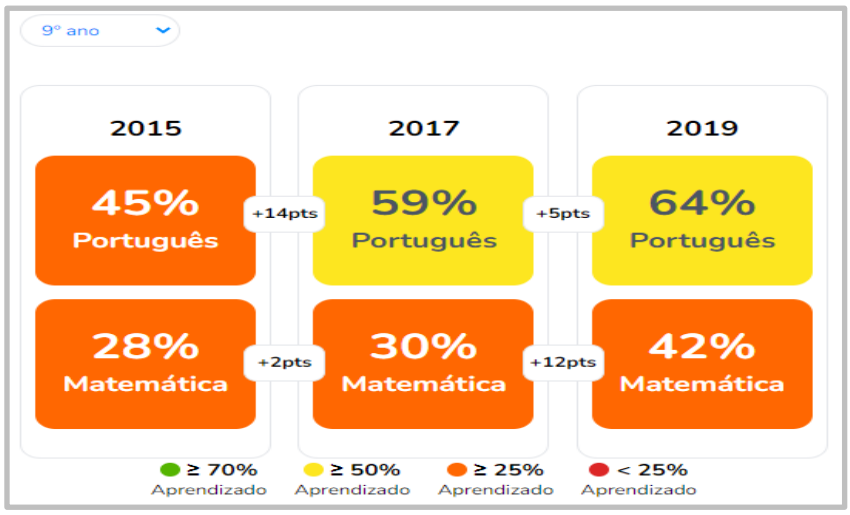
Essa constatação levou à realização de um experimento com alunos dos anos finais, abrangendo o início e o fim dessa fase escolar.

Figura 7 – Proporção de alunos que aprenderam no 5º ano



Fonte: https://qedu.org.br/escola/43050050

Figura 8 – Proporção de alunos que aprenderam no 9º ano



Fonte: https://qedu.org.br/escola/43050050

Através de pesquisa bibliográfica, este trabalho buscou entender como técnicas de Mineração de Dados na Educação (MDE), podem ser utilizadas para predizer alunos com baixo desempenho escolar. O objetivo é que a escola possa tomar decisões com antecedência, melhorando a proporção de alunos com aprendizado adequado para sua etapa escolar.

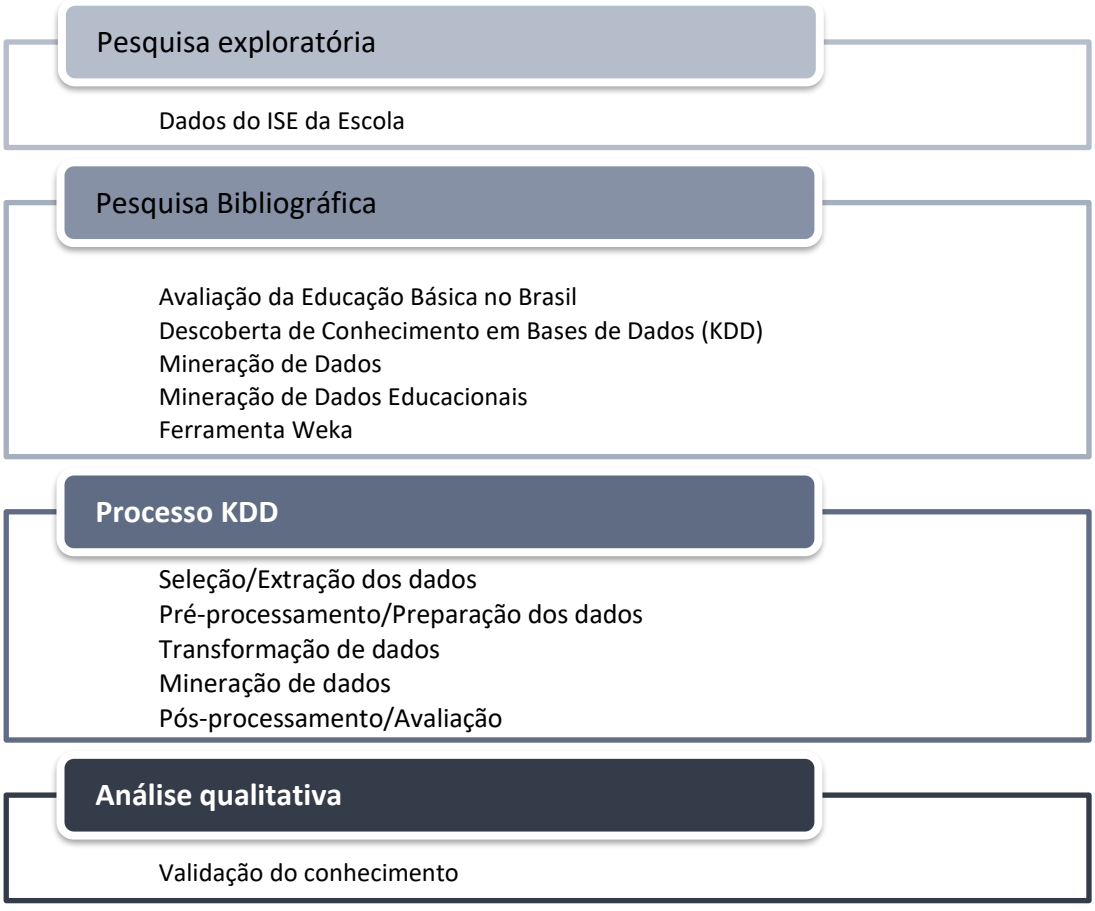
Para alcançar esse objetivo, adotou-se a metodologia do processo de descoberta de conhecimento em bases de dados, KDD, com as seguintes etapas: busca e seleção de dados brutos no Sistema Informatizado da Secretaria da Escola

(ISE); pré-processamento dos dados para uso em mineração de dados, encontrando padrões que levem à avaliação ou validação do conhecimento; e, finalmente, a utilização da ferramenta WEKA, que possui fácil uso e grande variedade de algoritmos disponíveis.

Os dados educacionais foram fornecidos pela secretaria da escola através do ISE e analisados para os anos de 2016 a 2019, pois em 2020 foram realizadas aulas *online* devido à pandemia. Para esta pesquisa, foram utilizadas técnicas de mineração de dados para identificar atributos que possam estar relacionados ao baixo desempenho dos alunos do 6º ao 9º ano da Escola Estadual de Ensino Fundamental Dr. Gabriel Álvaro de Miranda de Cruz Alta - RS.

O estudo foi dividido em dois experimentos: um com todos os alunos do 6º ano e outro com todos os alunos do 9º ano. Na Figura 9, é apresentado graficamente o fluxo das etapas seguidas na pesquisa em cada experimento.

Figura 9 – Etapas de desenvolvimento do projeto



Fonte: Autor (2022).

5 DESENVOLVIMENTO DA PESQUISA

A pesquisa foi conduzida seguindo as etapas do processo de KDD descritas na seção 2.2, as quais serão abordadas a seguir. Inicialmente, os dados brutos dos alunos foram coletados do sistema informatizado utilizado pela escola para armazenar informações, o ISE (Informatização da Secretaria da Educação). Foram utilizados dados dos anos de 2016 a 2019, incluindo alunos dos 6º e 9º anos.

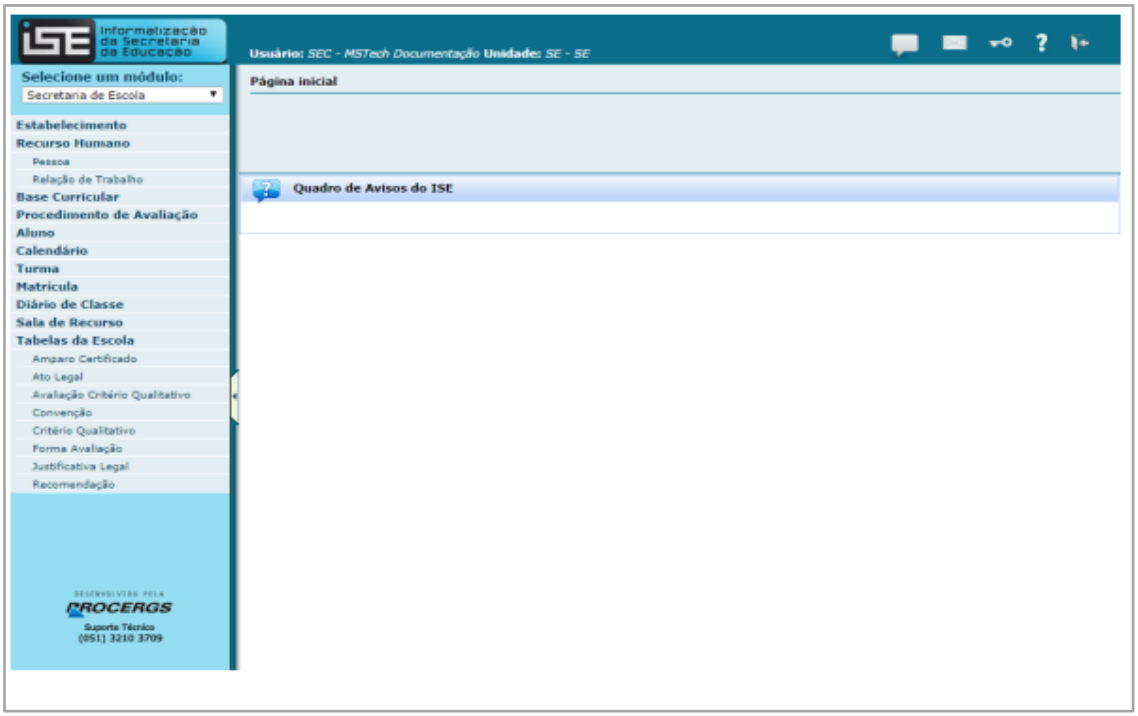
5.1 EXPERIMENTO COM O 6º ANO

Para a realização do experimento, foram utilizados dados de alunos de várias turmas do 6º ano, abrangendo o intervalo mencionado acima. No total, participaram 603 alunos, dos quais 42 foram transferidos ao longo dos anos, resultando em um grupo de 561 alunos para a realização do estudo.

5.1.1 Seleção de Dados

Com o objetivo de prever quais alunos têm maior probabilidade de reprovação, utilizou-se os dados obtidos na pesquisa, que estão armazenados no Sistema de Gestão da Rede Estadual de Ensino do Rio Grande do Sul (ISE). Esse sistema é um suporte tecnológico que auxilia os processos educacionais no Estado do RS e armazena toda a vida escolar de cada aluno. A Figura 10 mostra a interface do ISE, que é um sistema *online* que permite a consulta em tempo real das informações armazenadas. Para este estudo, foram selecionados todos os alunos do 6º ano no período correspondente ao experimento. Os dados utilizados na pesquisa foram disponibilizados pela escola em arquivos no formato pdf (*Portable Document Format*), já que o sistema é desenvolvido pela PROCERGS - Centro de Tecnologia da Informação e Comunicação do Estado do Rio Grande do Sul S. A., um órgão executor da política de informática do Estado do RS que não disponibiliza uma cópia do banco de dados.

Figura 10– Interface do sistema ISE



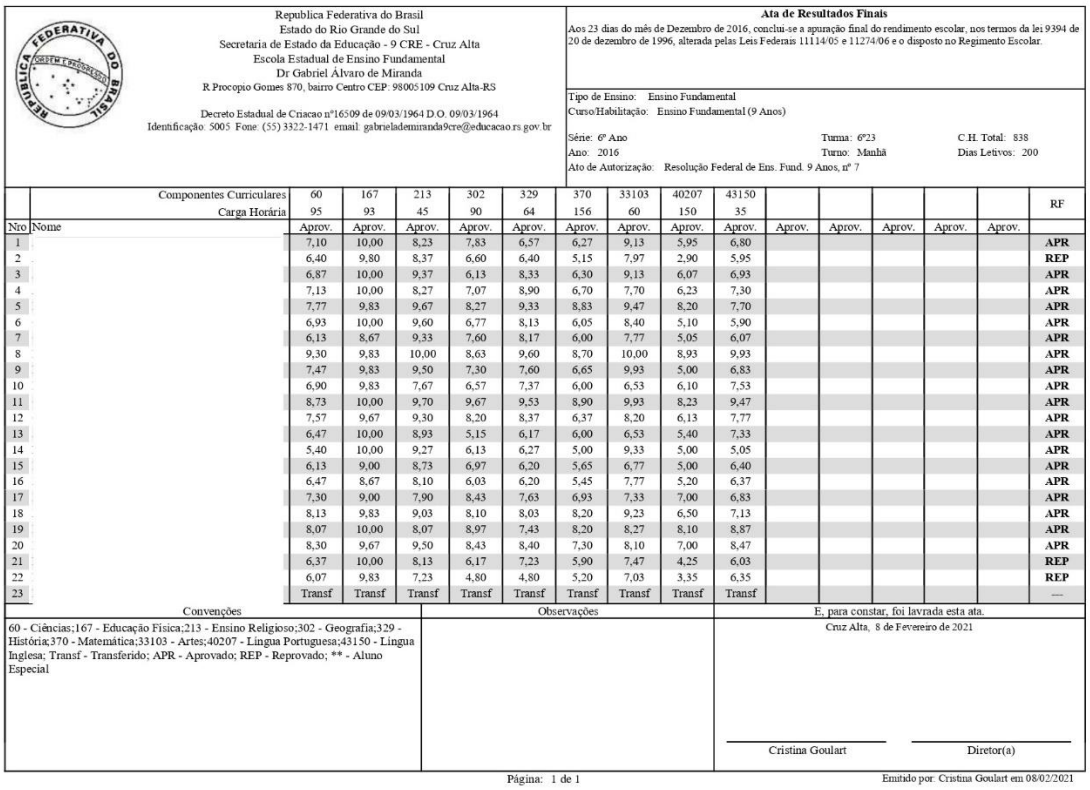
Fonte: https://moodle.educacao.rs.gov.br.

As informações sobre os alunos fornecidas em arquivos em formato PDF são provenientes de diversos documentos, como o boletim de desempenho individual de cada aluno, a lista de famílias beneficiárias do programa Bolsa Família (que é um programa de transferência direta de renda com condicionalidades, destinado a famílias em situação de pobreza e extrema pobreza), a relação dos alunos que utilizam transporte público para chegar à escola e as atas de resultados finais de cada turma do 6º ano.

A Figura 11 apresenta uma ata de resultados finais de uma turma específica de um determinado ano, onde são registrados resultados como aprovação (APR) ou reprovação (REP), média final e transferência de aluno.

Já na Figura 12, temos um boletim de desempenho individual que contém informações detalhadas sobre cada aluno, incluindo notas de cada disciplina por trimestre, média da disciplina, nota de recuperação e resultado final.

Figura 11– Ata de Resultado Final de cada turma em cada ano



Republica Federativa do Brasil
Estado do Rio Grande do Sul
Secretaria de Estado da Educação - 9 CRE - Cruz Alta
Escola Estadual de Ensino Fundamental
Dr Gabriel Álvaro de Miranda
R Procopio Gomes 870, bairro Centro CEP: 98005109 Cruz Alta-RS

Decreto Estadual de Criacao nº16509 de 09/03/1964 D.O. 09/03/1964
Identificação: 5005 Fone: (55) 3322-1471 email: gabrieladelmiranda9cre@educacao.rs.gov.br

Ata de Resultados Finais

Aos 23 dias do mês de Dezembro de 2016, conclui-se a apuração final do rendimento escolar, nos termos da lei 9394 de 20 de dezembro de 1996, alterada pelas Leis Federais 11114/05 e 11274/06 e o disposto no Regimento Escolar.

Tipo de Ensino: Ensino Fundamental
Curso/Habilitação: Ensino Fundamental (9 Anos)

Série: 6º Ano — Turma: 6º23 — C.H. Total: 838
Ano: 2016 — Turno: Manhã — Dias Letivos: 200
Ato de Autorização: Resolução Federal de Ens. Fund. 9 Anos, nº 7

	Componentes Curriculares	60	167	213	302	329	370	33103	40207	43150						RF
	Carga Horária	95	93	45	90	64	156	60	150	35						
Nro	Nome	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	Aprov.	
1		7,10	10,00	8,23	7,83	6,57	6,27	9,13	5,95	6,80						APR
2		6,40	9,80	8,37	6,60	6,40	5,15	7,97	2,90	5,95						REP
3		6,87	10,00	9,37	6,13	8,33	6,30	9,13	6,07	6,93						APR
4		7,13	10,00	8,27	7,07	8,90	6,70	7,70	6,23	7,30						APR
5		7,77	9,83	9,67	8,27	9,33	8,83	9,47	8,20	7,70						APR
6		6,93	10,00	9,60	6,77	8,13	6,05	8,40	5,10	5,90						APR
7		6,13	8,67	9,33	7,60	8,17	6,00	7,77	5,05	6,07						APR
8		9,30	9,83	10,00	8,63	9,60	8,70	10,00	8,93	9,93						APR
9		7,47	9,83	9,50	7,30	7,60	6,65	9,93	5,00	6,83						APR
10		6,90	9,83	7,67	6,57	7,37	6,00	6,53	6,10	7,53						APR
11		8,73	10,00	9,70	9,67	9,53	8,90	9,93	8,23	9,47						APR
12		7,57	9,67	9,30	8,20	8,37	6,37	8,20	6,13	7,77						APR
13		6,47	10,00	8,93	5,15	6,17	6,00	6,53	5,40	7,33						APR
14		5,40	10,00	9,27	6,13	6,27	5,00	9,33	5,00	5,05						APR
15		6,13	9,00	8,73	6,97	6,20	5,65	6,77	5,00	6,40						APR
16		6,47	8,67	8,10	6,03	6,20	5,45	7,77	5,20	6,37						APR
17		7,30	9,00	7,90	8,43	7,63	6,93	7,33	7,00	6,83						APR
18		8,13	9,83	9,03	8,10	8,03	8,20	9,23	6,50	7,13						APR
19		8,07	10,00	8,07	8,97	7,43	8,20	8,27	8,10	8,87						APR
20		8,30	9,67	9,50	8,43	8,40	7,30	8,10	7,00	8,47						APR
21		6,37	10,00	8,13	6,17	7,23	5,90	7,47	4,25	6,03						REP
22		6,07	9,83	7,23	4,80	4,80	5,20	7,03	3,35	6,35						REP
23		Transf	Transf	Transf	Transf	Transf	Transf	Transf	Transf	Transf						---

Convenções

60 - Ciências;167 - Educação Física;213 - Ensino Religioso;302 - Geografia;329 - História;370 - Matemática;33103 - Artes;40207 - Língua Portuguesa;43150 - Língua Inglesa; Transf - Transferido; APR - Aprovado; REP - Reprovado; ** - Aluno Especial

Observações

E, para constar, foi lavrada esta ata.

Cruz Alta, 8 de Fevereiro de 2021

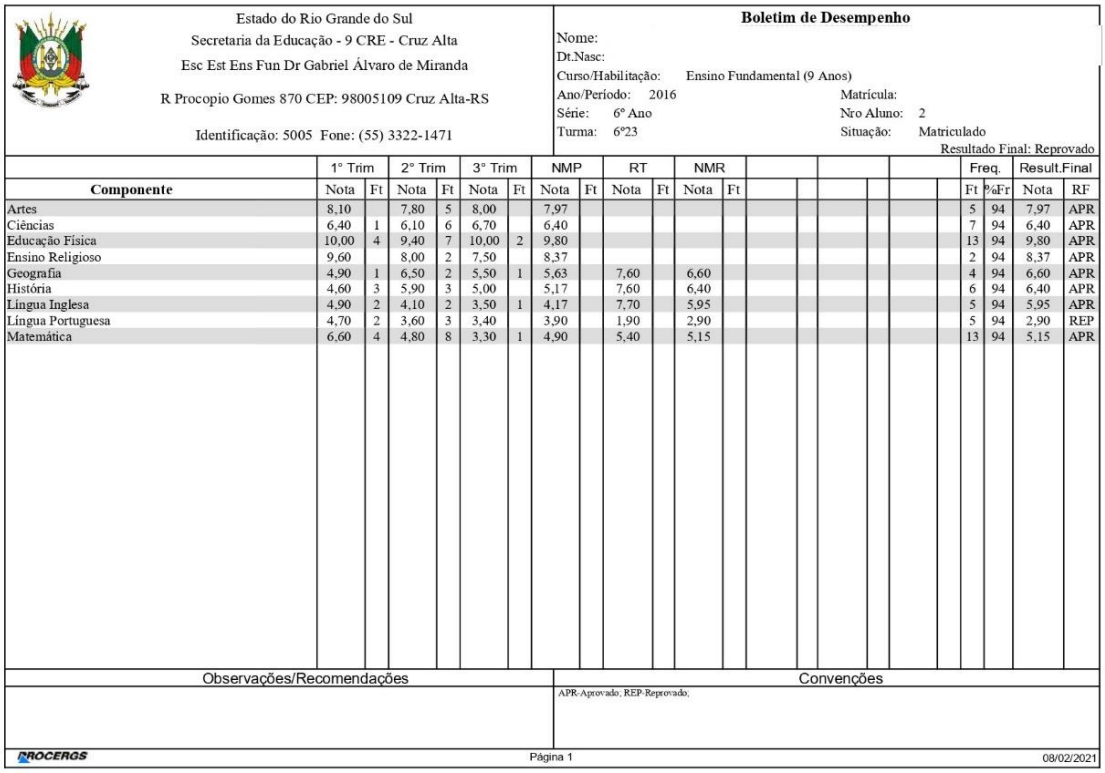
Cristina Goulart — Diretor(a)

Página: 1 de 1

Emitido por: Cristina Goulart em 08/02/2021

Fonte: Autor (ISE da escola)

Figura 12 – Boletim de Desempenho do aluno



Estado do Rio Grande do Sul
Secretaria da Educação - 9 CRE - Cruz Alta
Esc Est Ens Fun Dr Gabriel Álvaro de Miranda

R Procopio Gomes 870 CEP: 98005109 Cruz Alta-RS

Identificação: 5005 Fone: (55) 3322-1471

Boletim de Desempenho

Nome:
Dt.Nasc:
Curso/Habilitação: Ensino Fundamental (9 Anos)
Ano/Período: 2016 — Matrícula:
Série: 6º Ano — Nro Aluno: 2
Turma: 6º23 — Situação: Matriculado

Resultado Final: Reprovado

	1° Trim		2° Trim		3° Trim		NMP		RT		NMR								Freq.		Result.Final	
Componente	Nota	Ft	Nota	Ft	Nota	Ft	Nota	Ft	Nota	Ft	Nota	Ft							Ft	%Fr	Nota	RF
Artes	8,10		7,80	5	8,00		7,97												5	94	7,97	APR
Ciências	6,40	1	6,10	6	6,70		6,40												7	94	6,40	APR
Educação Física	10,00	4	9,40	7	10,00	2	9,80												13	94	9,80	APR
Ensino Religioso	9,60		8,00	2	7,50		8,37												2	94	8,37	APR
Geografia	4,90	1	6,50	2	5,50	1	5,63		7,60		6,60								4	94	6,60	APR
História	4,60	3	5,90	3	5,00		5,17		7,60		6,40								6	94	6,40	APR
Língua Inglesa	4,90	2	4,10	2	3,50	1	4,17		7,70		5,95								5	94	5,95	APR
Língua Portuguesa	4,70	2	3,60	3	3,40		3,90		1,90		2,90								5	94	2,90	REP
Matemática	6,60	4	4,80	8	3,30	1	4,90		5,40		5,15								13	94	5,15	APR

Observações/Recomendações

Convenções

APR-Aprovado; REP-Reprovado;

PROCERGS — Página 1 — 08/02/2021

Fonte: Autor (ISE da escola)

A extração dos dados foi realizada por meio da conversão de arquivos em formato PDF em planilhas de cálculo. Essa conversão permitiu a organização dos dados e a realização de cálculos. Para automatizar esse processo, utilizamos uma ferramenta *online* chamada *www.ilovepdf.com*.

No Quadro 2, apresentamos o boletim de desempenho do aluno convertido em planilha de cálculo. Com a conversão, obteve-se informações como disciplinas, notas, notas de recuperação, frequências e resultado final. Cada *pasta de trabalho* na planilha corresponde a uma turma do 6º ano, contendo planilhas individuais dos alunos identificadas como *Table x*.

Quadro 2 – Dados brutos na planilha de cálculo.

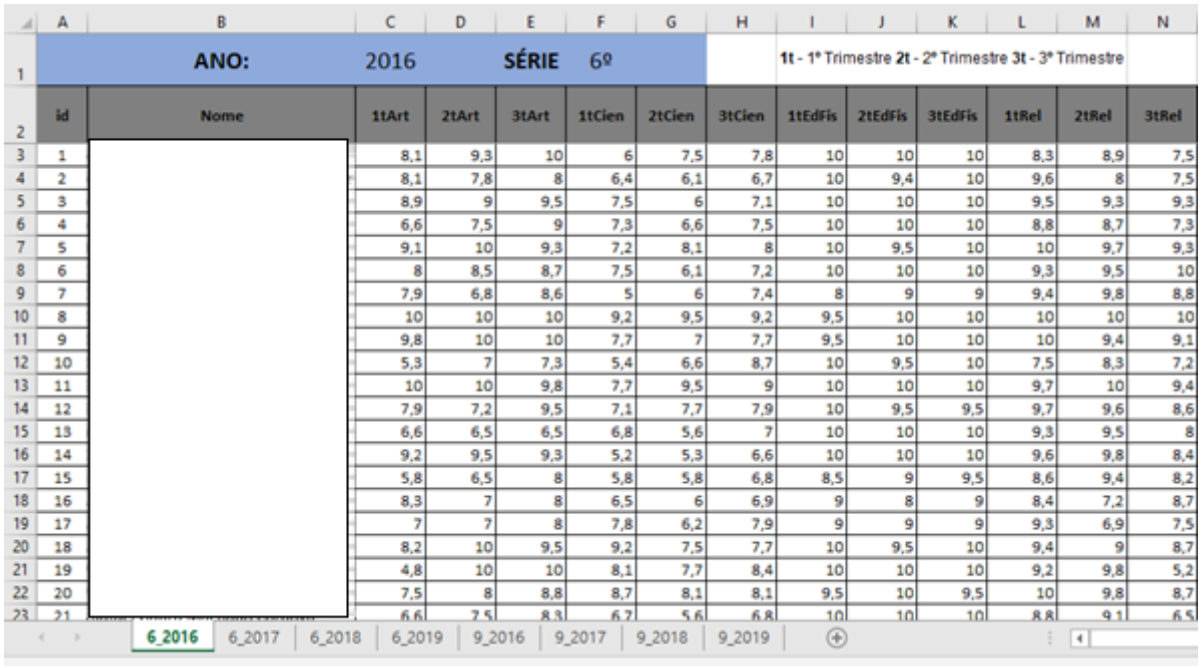
	1°	2°	3°	NMP		RT		NMR								Freq.		Result.Final	
Componente	Nota	Nota	Nota	Nota	Ft	Nota	Ft	Nota	Ft							Ft	%F	Nota	RF
Artes	8,10	9,30	10,00	9,13												5	97	9,13	APR
Ciências	6,00	7,50	7,80	7,10												3	97	7,10	APR
Educação Física	10,00	10,00	10,00	10,00												4	97	10,00	APR
Ensino Religioso	8,30	8,90	7,50	8,23												1	97	8,23	APR
Geografia	6,90	9,20	7,40	7,83												5	97	7,83	APR
História	2,40	8,60	8,70	6,57												3	97	6,57	APR
Língua Inglesa	6,00	7,50	6,90	6,80												2	97	6,80	APR
Língua Portuguesa	5,20	5,50	6,60	5,77		6,10		5,95								6	97	5,95	APR
Matemática	5,10	8,30	5,40	6,27												6	97	6,27	APR

Table 1 | Table 2 | Table 3 | Table 4 | Table 5 | Table 6 | Table 7 | Table 8 | Table 9 | Tal ...

Fonte: Autor

Após coletar as informações de uma turma, uma planilha foi criada para reunir todos os dados dos alunos, como demonstrado no Quadro 3. Cada linha da planilha representa um aluno e exibe suas notas em cada disciplina divididas por trimestre. Essa abordagem permitiu coletar informações detalhadas sobre os alunos e produzir um resultado final, criando assim uma base de dados completa.

Quadro 3 – Dados brutos extraídos por ano.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		ANO:	2016		SÉRIE	6º			1t - 1º Trimestre 2t - 2º Trimestre 3t - 3º Trimestre					
2	id	Nome	1tArt	2tArt	3tArt	1tCien	2tCien	3tCien	1tEdFis	2tEdFis	3tEdFis	1tRel	2tRel	3tRel
3	1		8,1	9,3	10	6	7,5	7,8	10	10	10	8,3	8,9	7,5
4	2		8,1	7,8	8	6,4	6,1	6,7	10	9,4	10	9,6	8	7,5
5	3		8,9	9	9,5	7,5	6	7,1	10	10	10	9,5	9,3	9,3
6	4		6,6	7,5	9	7,3	6,6	7,5	10	10	10	8,8	8,7	7,3
7	5		9,1	10	9,3	7,2	8,1	8	10	9,5	10	10	9,7	9,3
8	6		8	8,5	8,7	7,5	6,1	7,2	10	10	10	9,3	9,5	10
9	7		7,9	6,8	8,6	5	6	7,4	8	9	9	9,4	9,8	8,8
10	8		10	10	10	9,2	9,5	9,2	9,5	10	10	10	10	10
11	9		9,8	10	10	7,7	7	7,7	9,5	10	10	10	9,4	9,1
12	10		5,3	7	7,3	5,4	6,6	8,7	10	9,5	10	7,5	8,3	7,2
13	11		10	10	9,8	7,7	9,5	9	10	10	10	9,7	10	9,4
14	12		7,9	7,2	9,5	7,1	7,7	7,9	10	9,5	9,5	9,7	9,6	8,6
15	13		6,6	6,5	6,5	6,8	5,6	7	10	10	10	9,3	9,5	8
16	14		9,2	9,5	9,3	5,2	5,3	6,6	10	10	10	9,6	9,8	8,4
17	15		5,8	6,5	8	5,8	5,8	6,8	8,5	9	9,5	8,6	9,4	8,2
18	16		8,3	7	8	6,5	6	6,9	9	8	9	8,4	7,2	8,7
19	17		7	7	8	7,8	6,2	7,9	9	9	9	9,3	6,9	7,5
20	18		8,2	10	9,5	9,2	7,5	7,7	10	9,5	10	9,4	9	8,7
21	19		4,8	10	10	8,1	7,7	8,4	10	10	10	9,2	9,8	5,2
22	20		7,5	8	8,8	8,7	8,1	8,1	9,5	10	9,5	10	9,8	8,7
23	21		6,6	7,5	8,3	6,7	5,6	6,8	10	10	10	8,8	9,1	6,5

6_2016 | 6_2017 | 6_2018 | 6_2019 | 9_2016 | 9_2017 | 9_2018 | 9_2019

Fonte: Autor (2022).

Em seguida, todos os alunos do sexto ano dos anos letivos de 2016 a 2019 foram reunidos em uma única planilha contendo todos os dados. Para garantir a privacidade dos alunos, cada um recebeu um ID de referência, no qual seus nomes foram removidos.

Após a seleção da base de dados, os atributos foram filtrados e trabalhados para que o conhecimento pudesse ser extraído da melhor forma possível, conforme descrito no processo de pré-processamento.

5.1.2 Pré-Processamento / Preparação dos dados

De acordo com Castro (2016, p. 55-70), a fase de pré-processamento é crucial para a manipulação e preparação de dados brutos, a fim de extrair conhecimento valioso. Essa etapa deve ser realizada de maneira estruturada e cuidadosa, já que não existem ferramentas automáticas que possam executar essa tarefa. Como resultado, os dados coletados de várias fontes podem conter ruídos, inconsistências e valores ausentes, o que pode ser resolvido por meio da padronização dos dados.

Após a extração dos dados, foi realizada uma limpeza dos mesmos para torná-los confiáveis e consistentes, já que os dados podem estar em diferentes formatos. No total, foram extraídas 603 linhas, cada uma correspondendo a um aluno. Destas, foram descartadas 40 linhas de alunos transferidos e 02 linhas que continham dados fora do padrão (provavelmente erros do sistema), resultando em um total de 561 linhas com 41 atributos.

Quadro 4 – Dados brutos de todos os anos do 6º ano.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	2016		SÉRIE		6º		561 alunos			41 Atributo			23001		dados		488	APR		73	REP
2	1tArte	2tArt	3tArt	MRA	1tCien	2tCien	3tCien	MRC	1tEdFis	2tEdFis	3tEdFis	MRE	1tRel	2tRel	3tRel	MRR	1tGeo	2tGeo	3tGeo	MRG	1tH
3	2,30	5,00	4,40	Sim	3,90	4,70	3,50	Sim	2,30	5,00	4,40	Sim	6,10	9,10	9,40	Nao	2,20	3,00	3,80	Sim	2,2
4	3,00	4,10	2,20	Sim	2,60	1,60	0,70	Sim	3,00	4,10	2,20	Sim	10,00	10,00	10,00	Nao	4,00	3,80	3,10	Sim	4,0
5	3,50	7,30	5,40	Sim	7,00	6,10	6,90	Nao	3,50	7,30	5,40	Sim	9,00	10,00	9,20	Nao	7,30	7,10	4,80	Nao	7,3
6	3,70	4,10	4,20	Sim	5,40	4,20	6,30	Sim	3,70	4,10	4,20	Sim	9,70	8,40	7,00	Nao	5,40	4,60	3,40	Sim	5,4
7	3,70	5,10	3,50	Sim	5,60	4,70	4,80	Sim	3,70	5,10	3,50	Sim	10,00	10,00	10,00	Nao	6,40	5,50	7,30	Nao	6,4
8	4,20	4,80	3,50	Sim	4,40	4,30	6,40	Sim	4,20	4,80	3,50	Sim	8,70	8,90	9,40	Nao	5,00	3,90	4,20	Sim	5,0
9	4,50	7,20	7,60	Nao	4,30	4,10	5,00	Sim	9,00	10,00	10,00	Nao	7,50	8,60	8,20	Nao	2,80	3,00	4,90	Sim	2,0
10	4,50	7,20	7,00	Nao	6,10	6,00	7,30	Nao	9,50	9,50	10,00	Nao	9,40	8,70	6,00	Nao	7,50	8,30	5,90	Nao	7,0
11	4,80	10,00	10,00	Nao	8,10	7,70	8,40	Nao	10,00	10,00	10,00	Nao	9,20	9,80	5,20	Nao	9,70	9,60	7,60	Nao	7,1
12	4,80	8,00	8,30	Nao	6,40	5,10	6,70	Nao	9,50	10,00	10,00	Nao	8,70	6,70	6,30	Nao	4,50	7,50	4,90	Sim	3,7
13	4,80	5,70	3,40	Sim	3,20	5,20	3,70	Sim	4,80	5,70	3,40	Sim	10,00	10,00	10,00	Nao	5,60	4,90	5,00	Sim	5,6
14	4,80	5,80	6,60	Sim	4,80	5,60	7,70	Nao	4,80	5,80	6,60	Sim	5,20	8,80	10,00	Nao	3,80	6,40	6,80	Sim	3,8
15	5,00	6,10	7,00	Nao	5,40	4,30	8,50	Nao	5,00	6,10	7,00	Nao	10,00	10,00	9,70	Nao	6,60	6,60	6,10	Nao	6,6
16	5,20	5,90	3,70	Sim	4,30	4,50	3,90	Sim	5,20	5,90	3,70	Sim	10,00	10,00	10,00	Nao	5,50	6,70	4,10	Sim	5,5
17	5,20	6,20	4,10	Sim	6,20	6,10	3,80	Sim	5,20	6,20	4,10	Sim	9,60	9,80	10,00	Nao	5,50	5,20	5,00	Sim	5,5
18	5,20	4,70	4,90	Sim	5,20	4,00	3,90	Sim	5,20	4,70	4,90	Sim	10,00	10,00	10,00	Nao	7,20	5,20	5,70	Nao	7,2
19	5,30	7,00	7,30	Nao	5,40	6,60	8,70	Nao	10,00	9,50	10,00	Nao	7,50	8,30	7,20	Nao	6,40	6,20	7,10	Nao	7,0
20	5,30	5,80	4,70	Sim	6,10	2,10	7,90	Sim	5,30	5,80	4,70	Sim	10,00	10,00	8,00	Nao	4,80	4,50	7,40	Sim	4,8

Planilha1 Planilha2

Fonte: Autor (2022).

5.1.3 Transformações de Dados

Após o pré-processamento dos dados, a base foi completamente reestruturada com atributos relevantes para o estudo. Para adequar os dados aos algoritmos e ferramentas de mineração utilizadas, “foram transformados em formatos como arquivos CSV (*Comma-Separated Values*), ARFF (*Attribute-Relation File Format*) ou outros formatos especificados” (KAMPFF, 2009, p. 58).

Posteriormente, foram realizados ajustes nas variáveis para melhorar os resultados da mineração de dados. Para as variáveis médias de cada disciplina, os atributos de valor real foram convertidos em atributos nominais (MRA, MRC, MRE, MRR, MRG, MRH, MRI, MRP, MRM) com os valores *sim*, caso o aluno pegou recuperação, e *não*, caso contrário. Conforme apresentado no Quadro 5, a formatação dos atributos ficou da seguinte forma:

Quadro 5 – Relação dos atributos com suas descrições

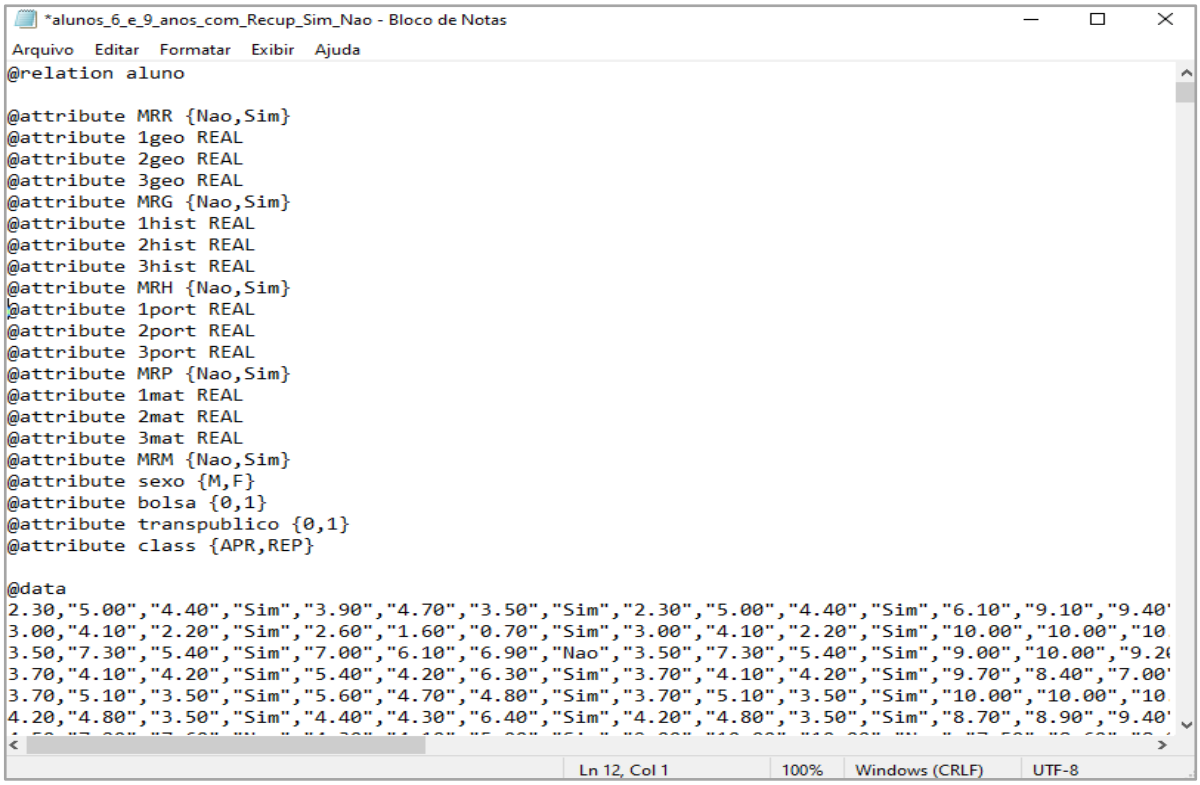
Atributos	Descrição	Valores
1arte	1º Trimestre de Ed. Artística	Real (0,00 – 10,00)
2arte	2º Trimestre de Ed. Artística	Real (0,00 – 10,00)
3arte	3º Trimestre de Ed. Artística	Real (0,00 – 10,00)
MRA	Média de Recuperação em Artística	Nominal (Não, Sim)
1cien	1º Trimestre de Ciências	Real (0,00 – 10,00)
2cien	2º Trimestre de Ciências	Real (0,00 – 10,00)
3cien	3º Trimestre de Ciências	Real (0,00 – 10,00)
MRC	Média de Recuperação em Ciências	Nominal (Não, Sim)
1edfisica	1º Trimestre de Ed. Física	Real (0,00 – 10,00)
2edfisica	2º Trimestre de Ed. Física	Real (0,00 – 10,00)
3edfisica	3º Trimestre de Ed. Física	Real (0,00 – 10,00)
MRE	Média de Recuperação em Física	Nominal (Não, Sim)
1reli	1º Trimestre de Ens. Religioso	Real (0,00 – 10,00)
2reli	2º Trimestre de Ens. Religioso	Real (0,00 – 10,00)
3reli	3º Trimestre de Ens. Religioso	Real (0,00 – 10,00)
MRR	Média de Recuperação em Ens. Religioso	Nominal (Não, Sim)
1geo	1º Trimestre de Geografia	Real (0,00 – 10,00)
2geo	2º Trimestre de Geografia	Real (0,00 – 10,00)
3geo	3º Trimestre de Geografia	Real (0,00 – 10,00)
MRG	Média de Recuperação em Geografia	Nominal (Não, Sim)
1hist	1º Trimestre de História	Real (0,00 – 10,00)
2hist	2º Trimestre de História	Real (0,00 – 10,00)
3hist	3º Trimestre de História	Real (0,00 – 10,00)
MRH	Média de Recuperação em História	Nominal (Não, Sim)
1ingles	1º Trimestre de Inglês	Real (0,00 – 10,00)
2ingles	2º Trimestre de Inglês	Real (0,00 – 10,00)
3ingles	3º Trimestre de Inglês	Real (0,00 – 10,00)
MRI	Média de Recuperação em Inglês	Nominal (Não, Sim)
1port	1º Trimestre de Português	Real (0,00 – 10,00)
2port	2º Trimestre de Português	Real (0,00 – 10,00)
3port	3º Trimestre de Português	Real (0,00 – 10,00)
MRP	Média de Recuperação em Português	Nominal (Não, Sim)
1mat	1º Trimestre de Matemática	Real (0,00 – 10,00)
2mat	2º Trimestre de Matemática	Real (0,00 – 10,00)
3mat	3º Trimestre de Matemática	Real (0,00 – 10,00)
MRM	Média de Recuperação em Matemática	Nominal (Não, Sim)
sexo	Sexo	Nominal (M, F)
bolsa	Bolsa Família	Inteiro (0,1)
transpublico	Transporte Público	Inteiro (0,1)
class	Aprovado e Reprovado	Nominal (APR, REP)

Fonte: Autor (2022).

Com o objetivo de prever antecipadamente a reprovação de um aluno, para que o gestor/professor possa auxiliá-lo antes que a reprovação ocorra, algumas variáveis, como aquelas que são referentes ao 3º trimestre e recuperação, foram consideradas irrelevantes para os experimentos. Dessa forma, os atributos não utilizados incluem: 3arte, MRA, 3cien, MRC, 3edfisica, MRE, 3reli, MRR, 3geo, MRG, 3hist, MRH, 3ingles, MRI, 3port, MRP e 3mat, MRM.

Após a limpeza dos dados, o próximo passo foi transformá-los para o formato *Attribute-Relation File Format* (ARFF), a fim de utilizá-los com a ferramenta WEKA. A Figura 13 apresenta a transformação dos dados para o formato ARFF.

Figura 13 – Arquivo Arff



Fonte: Autor (2022).

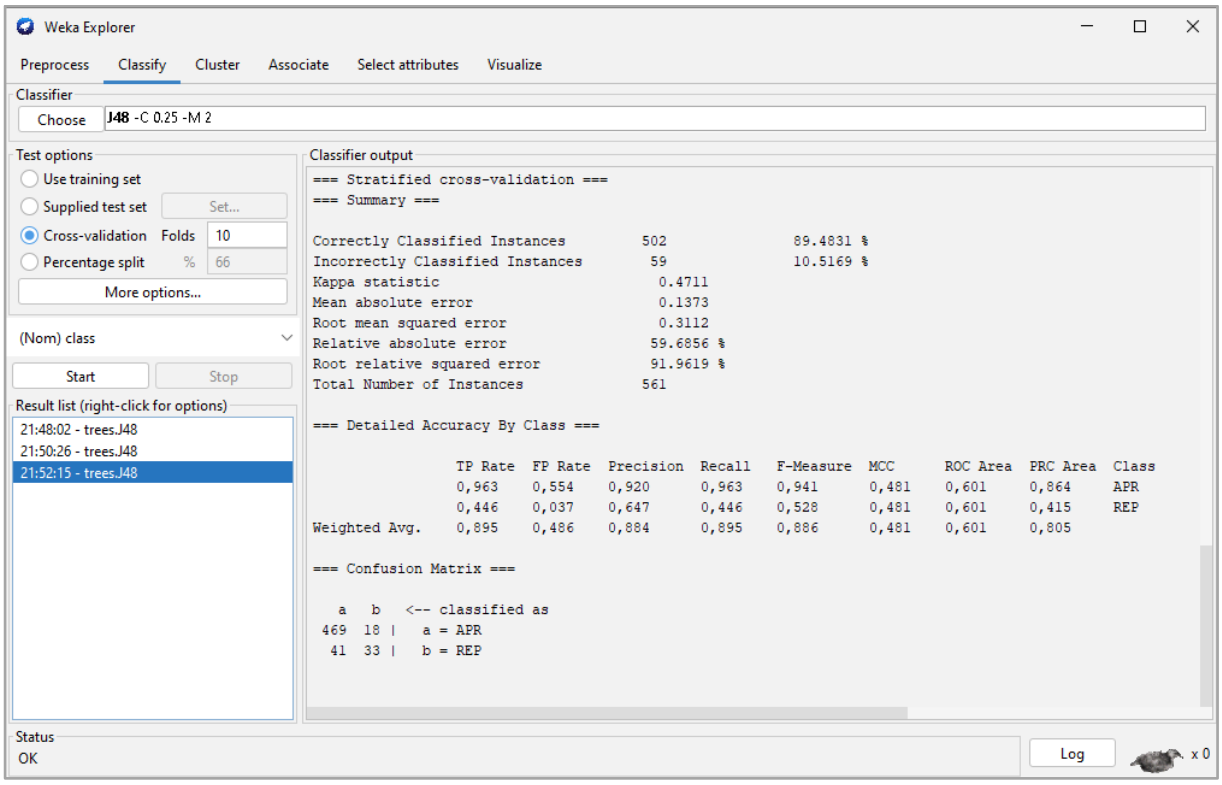
5.1.4 Aplicando Mineração de Dados: experimento 6º ano

De acordo com Kampff (2009, p. 58) “a fase de mineração de dados é essencial para extrair informações implícitas e potencialmente úteis, constituindo a etapa principal do processo de descoberta de conhecimento”. Nessa fase, as técnicas e algoritmos de mineração são aplicados em bancos de dados, analisando e explorando o conjunto de informações em busca de padrões úteis.

Após a seleção dos atributos e a preparação dos dados, a modelagem foi realizada com algoritmos de mineração, sendo escolhida a tarefa de classificação como a mais tradicional e a Árvore de Decisão como a técnica mais intuitiva, devido ao seu formato de árvore, que facilita a compreensão dos padrões encontrados. Para a construção dos modelos preditivos em árvores de decisão, foi utilizada a suíte *WEKA*, que é um *software* livre e contém implementações de algoritmos de diversas técnicas de mineração de dados. O algoritmo classificador escolhido foi o *J48*, e não houve nenhuma alteração nos parâmetros já configurados pelo *WEKA*.

O modo de teste utilizado para a geração dos modelos foi a *cross-validation*, validação cruzada de 10 pastas, onde o conjunto de dados foi dividido em 10 partes iguais, sendo 9 usadas para treinamento e 1 para teste em cada iteração. Na Figura 14, é possível observar a métrica de desempenho do algoritmo *J48* utilizado.

Figura 14 – Métrica de desempenho do 6º ano



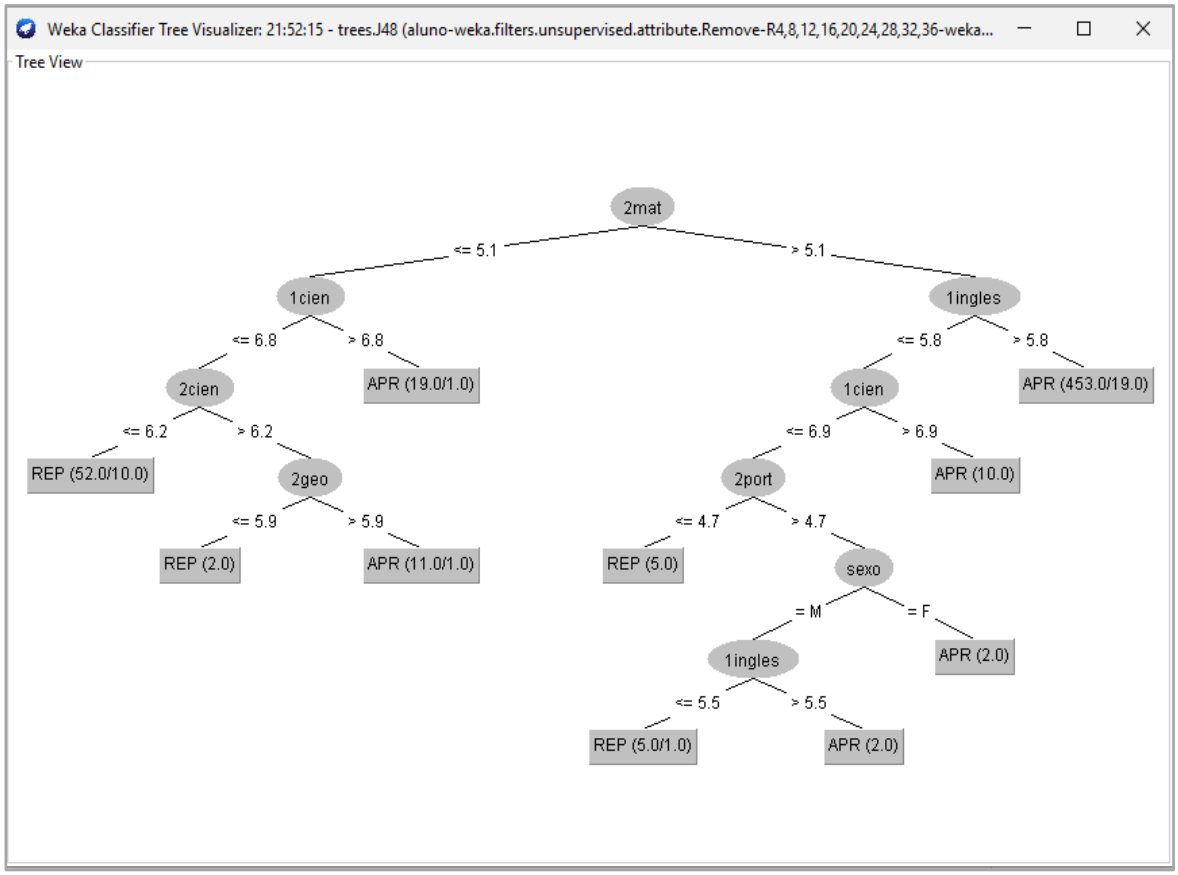
Fonte: Autor (2022).

A partir da análise da Figura 14, constata-se que, das 561 instâncias avaliadas, a acurácia atingiu 89,48%, o que significa que 502 instâncias foram classificadas corretamente e 59 foram classificadas incorretamente (10,52%). Um resultado importante do experimento foi a precisão de 64,7% para o atributo REP. No entanto,

a Matriz de Confusão indica que essa classe foi confundida com APR em 41 casos, o que pode comprometer o objetivo do trabalho, já que é essencial prever corretamente a classe REP.

A Figura 15 mostra a árvore de decisão gerada pelo algoritmo *J48* a partir da mineração de dados de classificação. Essa árvore evidencia os padrões identificados pelo algoritmo e as decisões tomadas com base nos atributos avaliados, incluindo a média da escola, que foi de 6,0.

Figura 15 – Árvore de decisão: todos alunos do 6º ano



Fonte: Autor (2022).

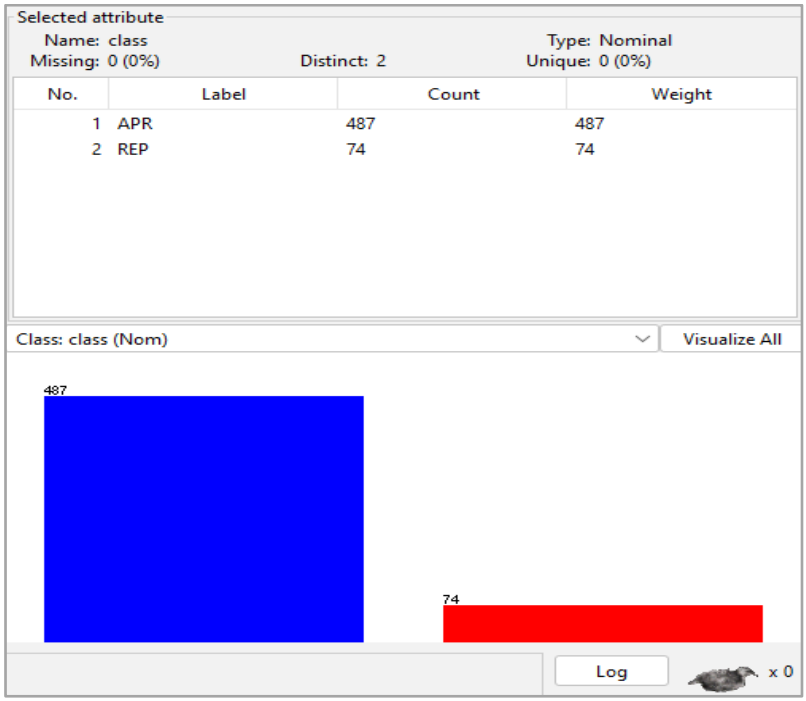
Observando os atributos classificados na árvore de decisão, foi identificado que a nota do 2º trimestre da disciplina de matemática foi escolhida como o nó raiz, dividindo-se em dois ramos - valores menores ou iguais a 5,1 e valores maiores. Nos nós seguintes, novas regras foram criadas, resultando em novos ramos.

Durante o experimento, foi possível notar uma disparidade significativa entre a proporção de instâncias das classes APR e REP, o que é conhecido como o problema

de desbalanceamento de classes. Esse problema limitou e prejudicou a acurácia do algoritmo de classificação. O desbalanceamento de classes ocorre quando há uma clara desproporção entre o número de exemplos de uma ou mais classes em relação às demais classes em um conjunto de dados.

Além disso, constatou-se que o atributo REP apresentou uma precisão baixa, o que pode ter prejudicado a precisão das predições. A Figura 16 ilustra essa disparidade, com a classe *APR* tendo 487 instâncias e a classe *REP* tendo apenas 74 instâncias.

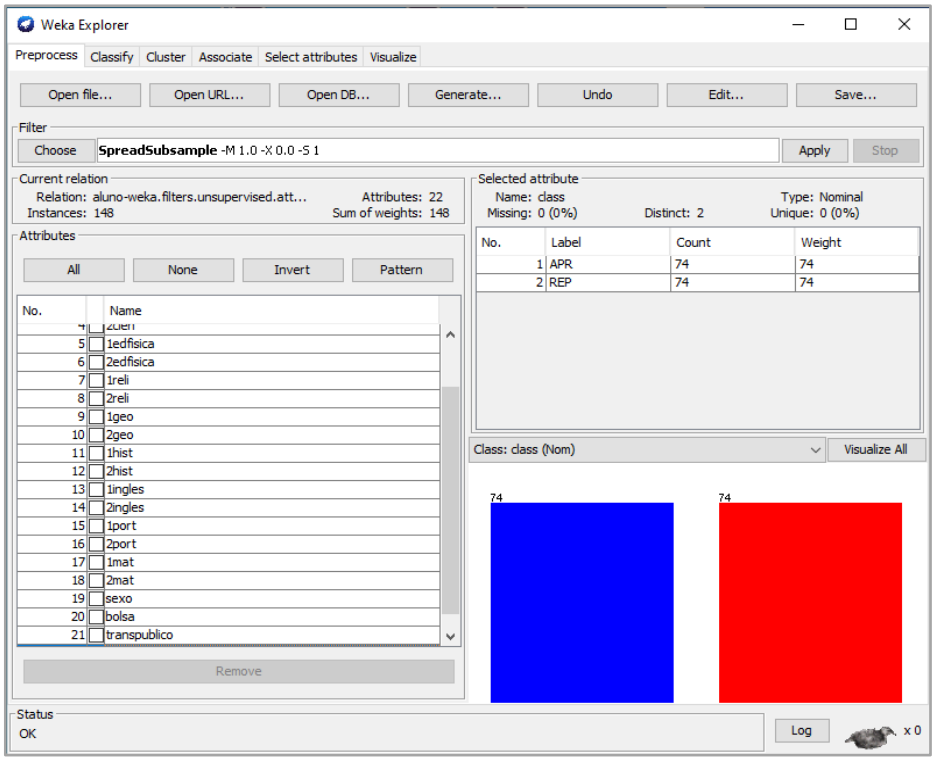
Figura 16 – Dados desbalanceados do 6º ano



Fonte: Autor (2022).

Para resolver esse problema, utilizou-se o filtro *SpreadSubsample* do *WEKA* para balancear os dados. O parâmetro *distributionSpread* foi ajustado para o valor 1, o que produz uma amostra aleatória dos dados e permite definir o máximo *spread* entre as classes, reduzindo as instâncias classificadas como *APR* e equiparando as duas classes. Após o balanceamento das classes, como mostrado na Figura 17, o experimento foi novamente executado, comparando os resultados com a inclusão de todos os atributos usados no experimento anterior.

Figura 17 – Dados balanceados do 6º ano



Fonte: Autor (2022).

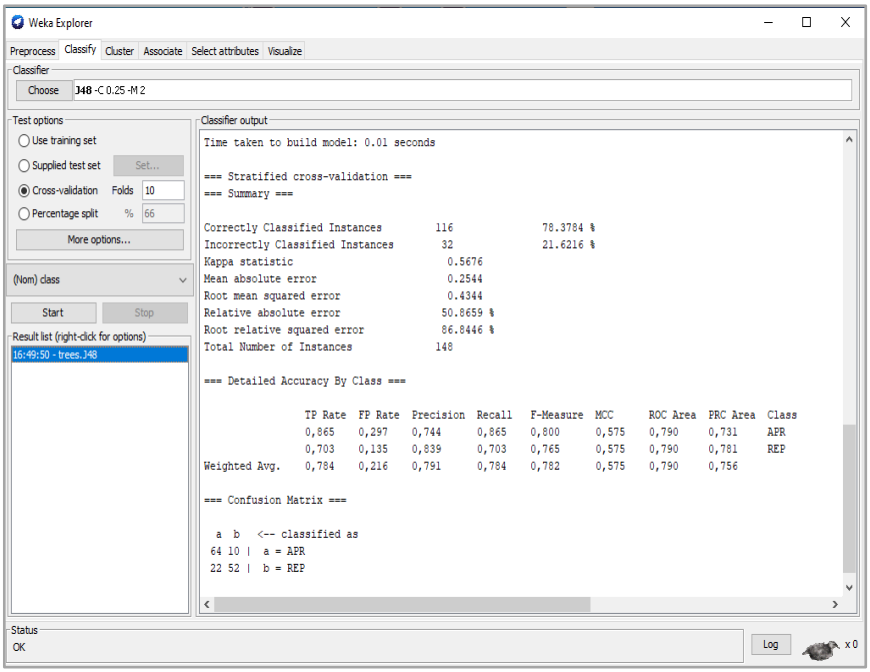
Após o balanceamento das classes, repetiu-se o experimento utilizando o algoritmo *J48* para classificação e obteve-se o resultado apresentado na Figura 18, que indica a métrica alcançada.

Com o balanceamento das classes, foi possível observar que os atributos APR e REP apresentaram a mesma quantidade de instâncias, totalizando 148 amostras onde foi obtida uma acurácia de 78,38%. Dessa forma, 116 instâncias foram classificadas corretamente, enquanto 32 foram classificadas incorretamente, representando 21,62% de erros.

A análise da matriz de confusão indica que, na classe APR, das 74 instâncias, 64 foram corretamente previstas e 10 foram incorretamente classificadas. Já na classe REP, das 74 instâncias, 52 foram previstas corretamente e 22 foram incorretamente classificadas.

É possível perceber que a precisão da classe REP foi melhor em relação ao experimento anterior, atingindo um valor de 83,9%. Além disso, a matriz de confusão mostrou que a classe REP foi menos confundida com a classe APR, com um total de 22 instâncias.

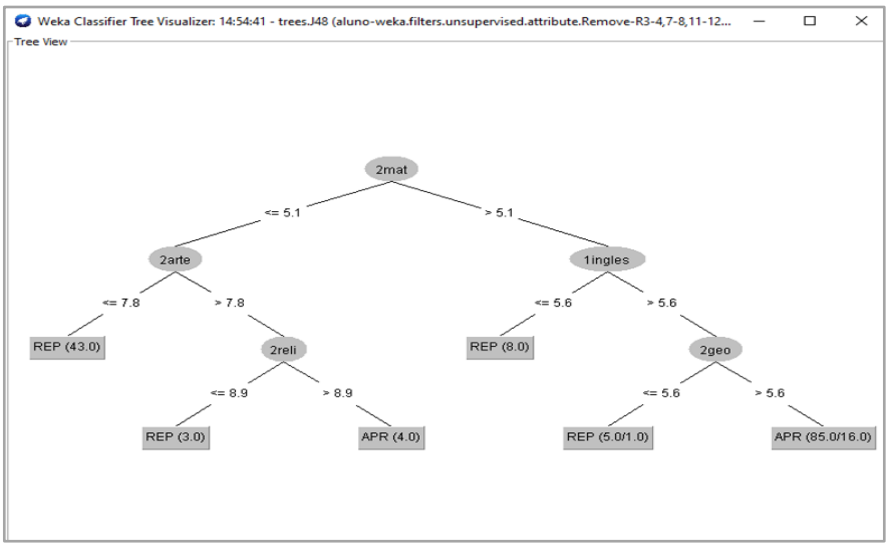
Figura 18 – Métrica de desempenho do algoritmo *J48* após balanceamento do 6º ano



Fonte: Autor (2022).

A Figura 19 mostra a árvore de decisão resultante após a aplicação do balanceamento. Observou-se que, após a realização do processo, alguns atributos não foram considerados na classificação. No entanto, ao criar o padrão utilizando o algoritmo, verificou-se que o atributo que novamente assumiu o papel de nó raiz foi a nota do 2º trimestre da disciplina de matemática (2mat), tornando-se, portanto, o atributo determinante para as decisões subsequentes.

Figura 19 – Árvore de decisão após balanceamento do 6º ano



Fonte: Autor (2022).

Os números entre parênteses abaixo de cada classificação mostram quantos alunos foram corretamente classificados pelo algoritmo como Aprovados (APR) ou Reprovados (REP), e ao lado, quantos foram classificados incorretamente. A árvore gerada mostra alguns padrões interessantes. Dos 148 alunos, 50 tiveram notas iguais ou inferiores a 5,1 na disciplina 2mat, levando à análise da disciplina de 2arte. Nessa disciplina, 43 alunos tiveram notas abaixo ou iguais a 7,8, sendo classificados pelo algoritmo como reprovados, enquanto 3 alunos que obtiveram notas acima de 7,8 em 2arte tiveram a nota de 2reli como determinante, sendo classificados como reprovados se a nota fosse menor ou igual a 8,9, e aprovados se a nota fosse maior que 8,9. Os demais 4 alunos com notas acima de 8,9 em 2reli foram aprovados.

Para os 98 alunos que tiveram notas acima de 5,1 em 2mat, o algoritmo analisou 1ingles, onde 8 alunos que obtiveram notas iguais ou inferiores a 5,6 foram classificados como reprovados. O algoritmo, em seguida, analisou 2geo como determinante para os alunos que obtiveram notas acima de 5,6 em 1ingles. Nesse caso, 5 alunos obtiveram notas iguais ou inferiores a 5,6, sendo classificados como reprovados (um aluno foi classificado incorretamente). Já 85 alunos que obtiveram notas acima de 5,6 em 2geo foram aprovados, embora o algoritmo tenha classificado incorretamente 16 alunos.

5.1.5 Interpretação/Avaliação do Experimento com o 6º Ano

A fase atual envolve a análise e aplicação do conhecimento adquirido para solucionar os problemas que motivaram a mineração. Os conhecimentos gerados na mineração podem ser incorporados por outros sistemas ou disponibilizados para ampliar a base de conhecimento ou auxiliar na tomada de decisão (KAMPFF, 2009, p. 58).

Ao analisar qualitativamente os dados finais, percebeu-se que mesmo com uma acurácia menor na classificação das instâncias, o balanceamento resultou em uma maior precisão na classe REP, que é o atributo preditivo. Esses resultados sugerem que os alunos apresentam maiores dificuldades na disciplina de matemática, o que é consistente com os resultados do IDEB. De fato, a disciplina de matemática, juntamente com outras, é determinante para a aprovação ou reprovação.

Além disso, observou-se que a nota do segundo trimestre em matemática pode ser um indicador para olhar outras disciplinas, tais como linguagem e religião, se a nota for inferior ou igual a 5,1, ou inglês e geografia, se a nota for maior.

O Quadro 6 apresenta a comparação das métricas dos experimentos realizados, que incluem aqueles com e sem balanceamento, bem como a comparação das precisões alcançadas. A precisão se refere ao número de instâncias classificadas corretamente em relação ao total de amostras positivas. Os resultados indicam que a precisão é maior quando é realizado o balanceamento.

Quadro 6 – Métrica do experimento com e sem balanceamento do 6º ano

6º ANO	**Sem balanceamento**		**Com balanceamento**	
		Acurácia		Acurácia
Instâncias classificadas corretamente	502	89,48%	116	78,38%
Instâncias classificadas incorretamente	59	10,52%	32	21,62%
Precisão APR	0,941		0,744	
Precisão REP	0,446		0,839	

Fonte: Autor (2022).

5.2 EXPERIMENTO COM O 9º ANO

Para este experimento, foram utilizados dados de 490 alunos do 9º ano escolar, compreendendo o período de 2016 a 2019. Ao longo desses anos, 43 alunos foram transferidos, resultando em um total de 447 alunos para a realização do experimento.

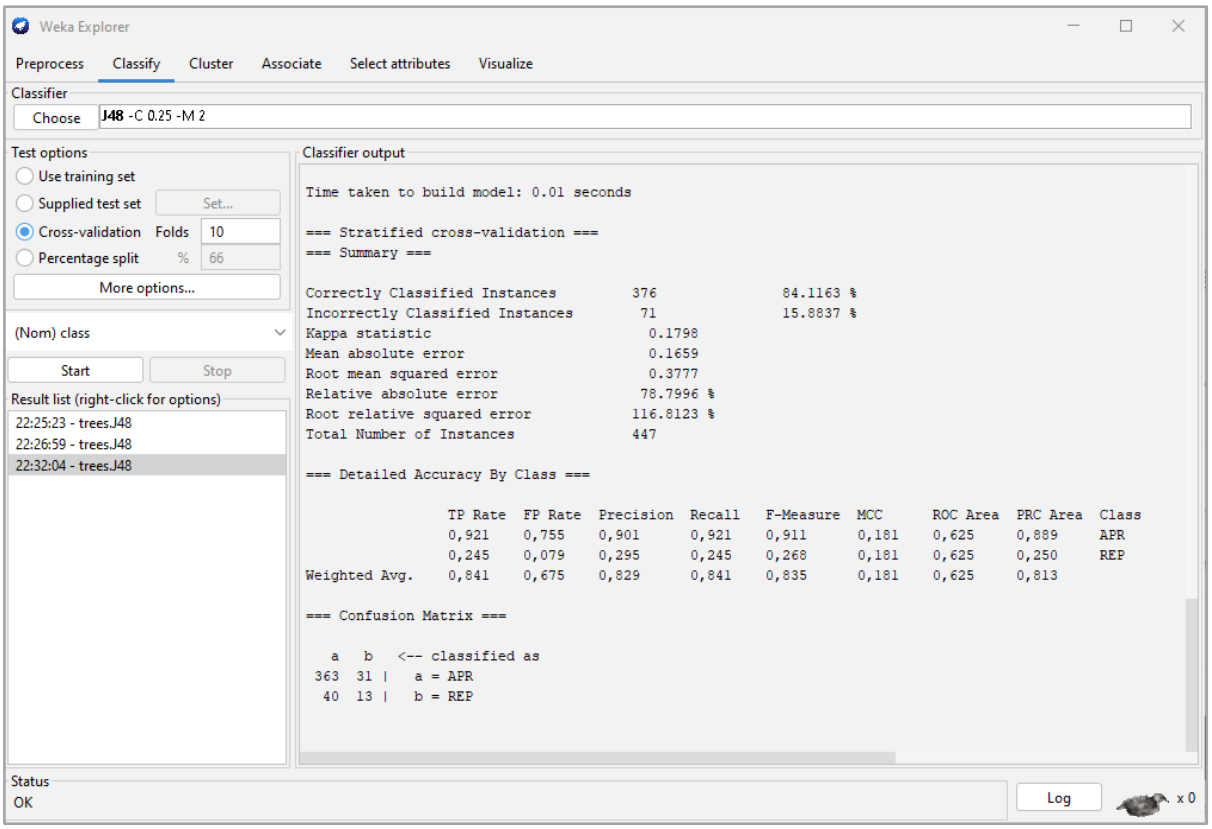
A seleção de dados foi realizada da mesma forma que no experimento do 6º ano, por meio do ISE. O pré-processamento e preparação dos dados foram conduzidos de maneira sistemática, seguindo a mesma metodologia do experimento anterior. A etapa de Transformação de Dados também foi realizada com os mesmos atributos.

5.2.1 Aplicando Mineração de Dados: experimento 9º ano

Neste experimento, utilizou-se o mesmo método empregado anteriormente, o qual consistiu em utilizar a técnica de Árvore de Decisão com o algoritmo classificador *J48*, tendo como tarefa a Classificação em MD. Não foram realizadas alterações nos parâmetros do algoritmo. O modelo foi gerado utilizando-se o método de teste *cross-validation*, com uma validação cruzada de 10 pastas. Nesse método, os dados foram divididos em 9 pastas para treinamento e 1 pasta para teste, sendo que cada iteração utilizou uma pasta diferente para teste.

Ao aplicar esse experimento em todos os alunos do 9º ano, os resultados da mineração de dados foram satisfatórios, obtendo-se uma acurácia de 84,12% com o algoritmo *J48*, o que é evidenciado na Figura 20.

Figura 20 – Métrica de desempenho do 9º ano



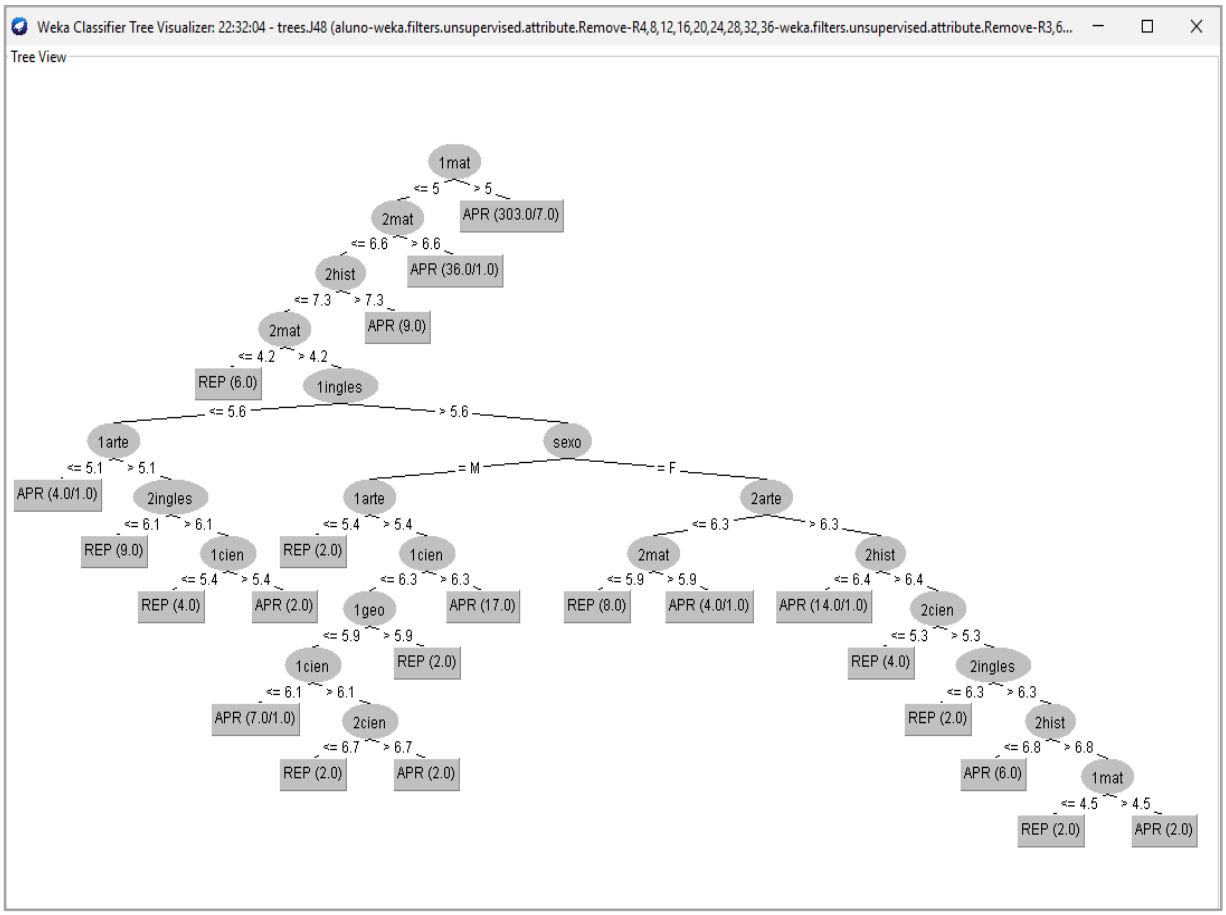
Fonte: Autor (2022).

No experimento realizado com 447 instâncias e 22 atributos, obteve-se uma acurácia de 84,12%, o que significa que 376 instâncias foram classificadas

corretamente e 71 foram classificadas incorretamente, representando 15,88% de erro. É importante destacar que a precisão do atributo REP foi de apenas 29,5% e que a matriz de confusão revelou que, das 53 instâncias, 40 foram previstas de forma equivocada.

A Figura 21 apresenta a árvore de decisão gerada pelo algoritmo *J48*, que permite visualizar os dados obtidos de forma mais clara e intuitiva.

Figura 21 – Árvore de decisão de todos alunos do 9º ano

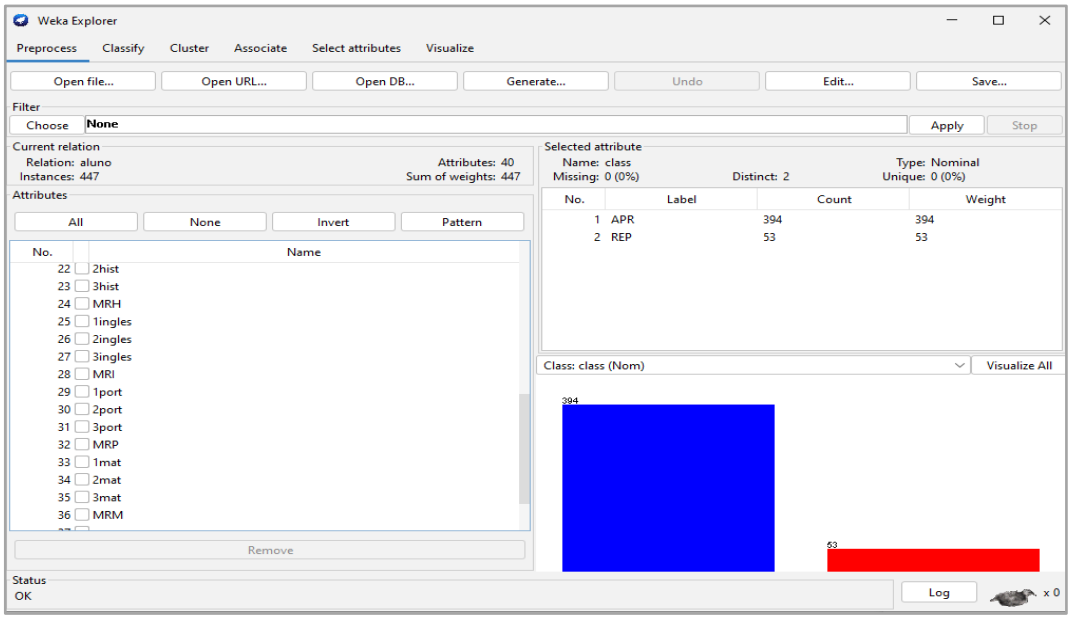


Fonte: Autor (2022).

Neste experimento, utilizou-se a nota do 1º trimestre da disciplina de matemática como nó raiz, sendo esse o atributo que direcionou as demais decisões. Assim como no experimento anterior, houve uma diferença na proporção de instâncias entre as classes APR e REP.

Conforme pode ser observado na Figura 22, o atributo APR apresentou um total de 394 instâncias, enquanto o atributo REP apresentou 53 instâncias.

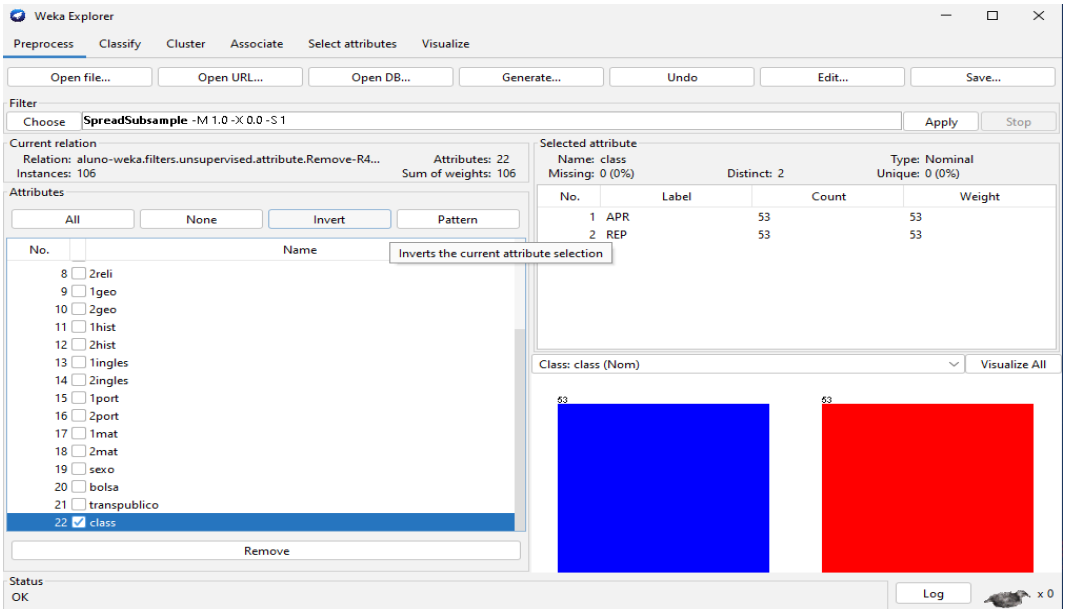
Figura 22 – Dados desbalanceados do 9º ano



Fonte: Autor (2022).

Da mesma forma, foi realizado o balanceamento dos dados utilizando o filtro *SpreadSubsample,* com o campo *distributionSpread* definido como 1, de modo a equalizar as duas classes por meio da redução de instâncias classificadas como APR.

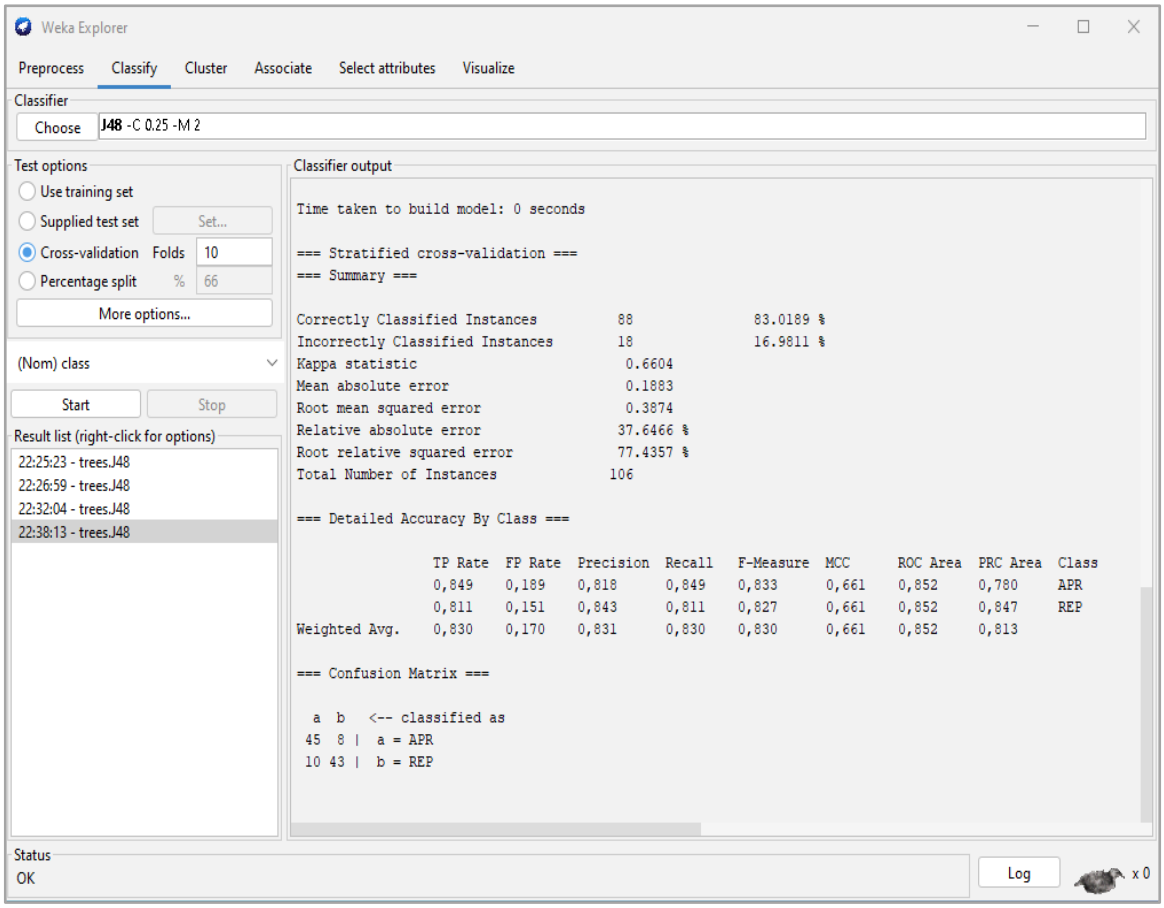
Figura 23 – Dados balanceados do 9º ano



Fonte: Autor (2022).

Após equilibrar as classes, restaram 53 instâncias para repetir o experimento de classificação usando o algoritmo *J48*. Os resultados da métrica e da árvore de decisão podem ser vistos na Figura 24 e Figura 25, respectivamente.

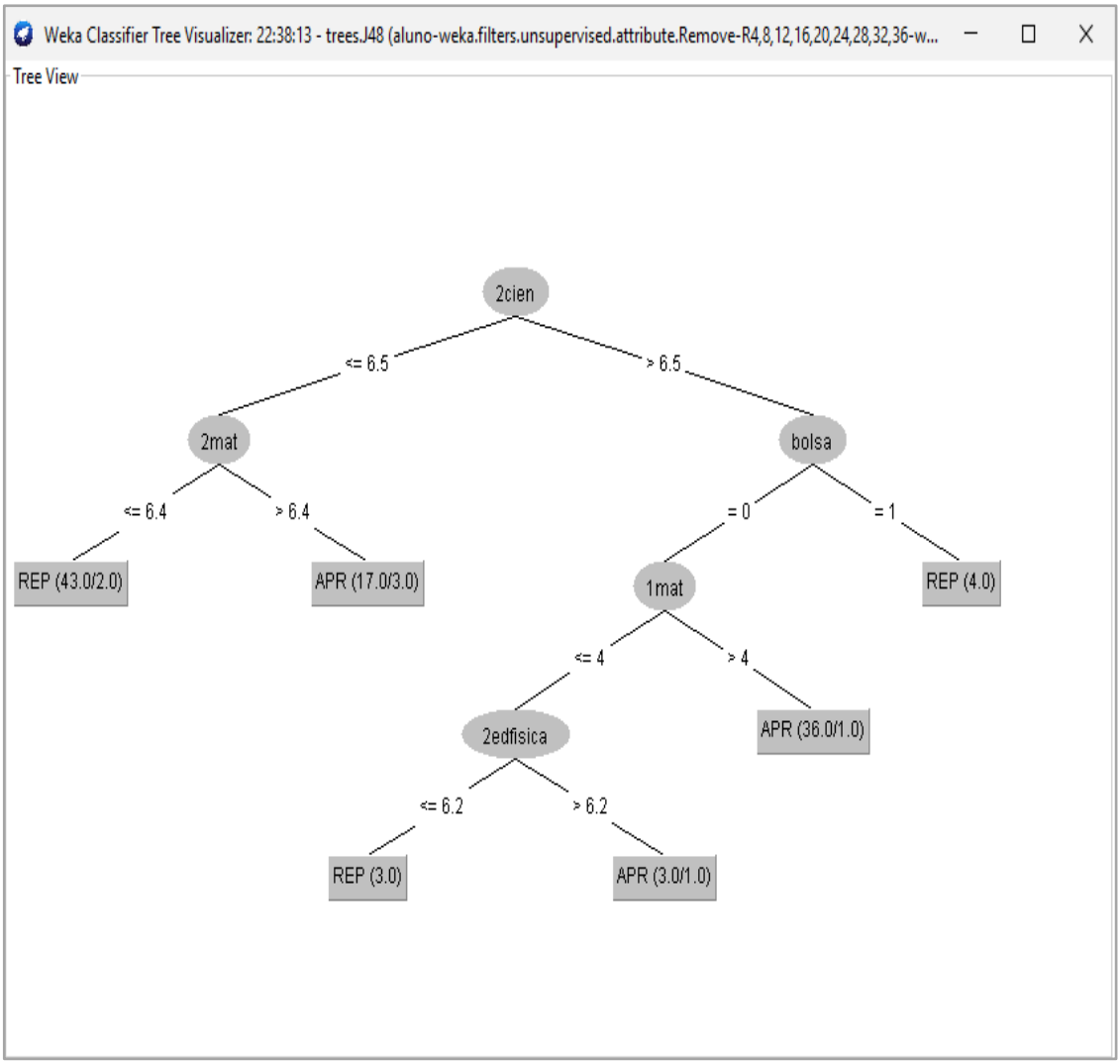
Figura 24 – Métrica do algoritmo *J48* após balanceamento do 9º ano



Fonte: Autor (2022).

Com a aplicação do balanceamento, a acurácia do modelo foi de 83,02%. Das 106 instâncias avaliadas, 88 foram classificadas corretamente e 18 incorretamente, o que representa 16,98% do total. Vale destacar que o atributo REP obteve uma precisão de 84,3%, de acordo com a matriz de confusão. Além disso, é possível observar que essa classe apresentou menos erros, com apenas 10 das 53 instâncias sendo previstas incorretamente.

Figura 25 – Árvore de decisão após balanceamento do 9º ano



Fonte: Autor (2022).

A árvore de decisão revela uma diferença significativa em relação ao 6º ano. O atributo que se tornou o nó raiz foi a nota do 2º trimestre da disciplina de ciências (2cien), que influenciou todas as decisões subsequentes.

De acordo com a árvore de decisão, 60 alunos obtiveram notas iguais ou inferiores a 6,5 na disciplina de 2cien. Para esses alunos, a nota de 2mat foi analisada, e se fosse igual ou inferior a 6,4, eles foram reprovados. Nesse caso, 43 alunos foram reprovados (2 classificados erroneamente), enquanto 17 foram aprovados (3 classificados erroneamente) com notas superiores.

Para os demais alunos que obtiveram notas superiores a 6,5 em 2cien, o algoritmo identificou um padrão: ele verificou se o aluno tinha bolsa família. Se tivesse, o aluno foi reprovado (4 alunos), enquanto se não tivesse, a nota de 1mat foi

analisada, onde 36 alunos (1 classificado erroneamente) com notas superiores a 4,0 foram aprovados, enquanto 6 alunos com notas iguais ou inferiores a 4,0 tiveram suas notas de 2edfisica analisadas. Três desses alunos (1 classificado erroneamente) com notas iguais ou superiores a 6,2 foram aprovados, enquanto os outros três foram reprovados por notas iguais ou inferiores.

5.2.2 Interpretação/Avaliação do Experimento com o 9º Ano

Após o processo de balanceamento, foi observada uma melhoria significativa na precisão da classe REP, conforme demonstrado no Quadro 7. No entanto, em contraste com o desempenho do 6º ano, constatou-se que a disciplina de ciências apresenta maior dificuldade para os alunos. Esta disciplina é determinante para a aprovação ou reprovação dos alunos, dependendo do seu desempenho em outras disciplinas e se possuem benefício do programa Bolsa Família.

Quadro 7 – Métrica do experimento com e sem balanceamento do 9º ano

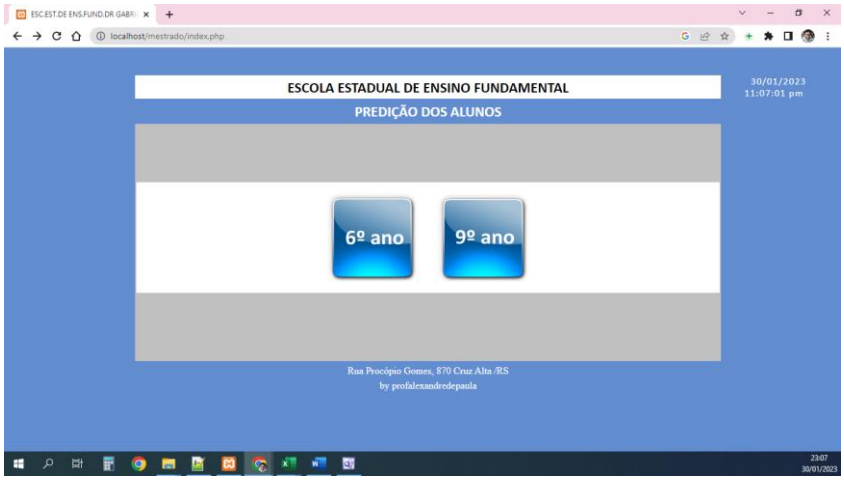
9º ANO	**Sem balanceamento**		**Com balanceamento**	
		Acurácia		Acurácia
Instâncias classificadas corretamente	376	84,12%	88	83,02%
Instâncias classificadas incorretamente	71	15,88%	18	16,98%
Precisão APR	0,901		0,818	
Precisão REP	0,295		0,843	

Fonte: Autor (2022).

6 DESENVOLVIMENTO DO SISTEMA *WEB*

Foram realizados experimentos para desenvolver um sistema *web* que auxilia a equipe diretiva na predição de alunos que possam ter sua reprovação revogada. O sistema foi baseado em padrões identificados nos 6º e 9º anos, como verificado em um estudo de caso realizado na escola. O *layout* do sistema é apresentado na Figura 26, onde é possível selecionar o ano que o aluno está cursando. Após a seleção, são apresentados campos para preenchimento das notas do primeiro e segundo trimestres de algumas disciplinas, conforme ilustrado nas Figuras 27 e 29. Por fim, o sistema fornece uma predição de *Aprovado* ou *Reprovado*, como mostrado nas Figuras 28 e 29.

Figura 26 – *Layout* do sistema de Predição dos alunos.

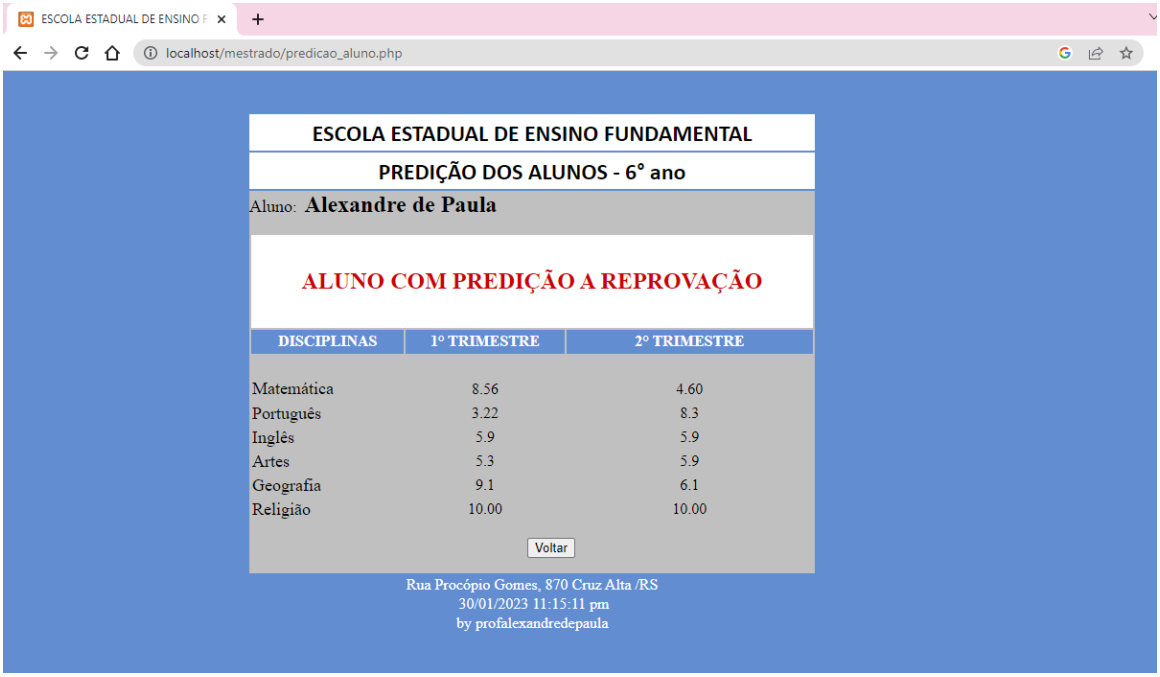


Fonte: Autor (2023).

Figura 27 – *Layout* da Predição dos alunos 6º ano.

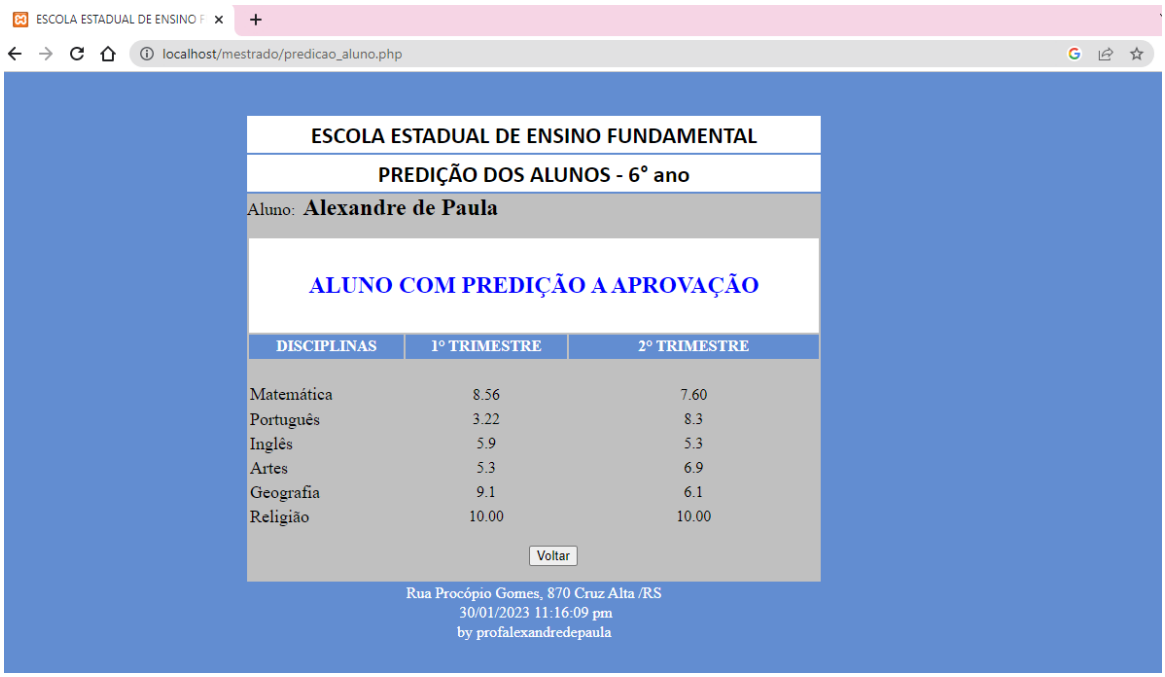
Fonte: Autor (2023).

Figura 28 – *Layout* do resultado da predição - Reprovação



Fonte: Autor (2023).

Figura 29 – *Layout* do resultado da predição - Aprovação.



Fonte: Autor (2023).

A Figura 30 apresenta informações para preenchimento específico do 9º ano, como notas do primeiro e segundo trimestres em algumas disciplinas e a indicação se

o aluno recebe Bolsa Família. Com base nesses dados, o sistema faz uma previsão se o aluno será aprovado ou reprovado.

Figura 30 – *Layout* da Predição dos alunos 9º ano.

Fonte: Autor (2023).

7 CONSIDERAÇÕES FINAIS

A principal tarefa da escola é compartilhar conhecimento e educação, e garantir um ensino de qualidade que resulte em bons desempenhos dos alunos. Para alcançar a eficiência do sistema educacional, é necessário considerar vários fatores, como condição socioeconômica, cultural e acesso, entre outros. A qualidade da educação é medida a partir de uma perspectiva social e econômica e é considerada bem-sucedida quando contribui para a igualdade e a eficácia dos recursos destinados à educação. Indicadores são utilizados para medir a qualidade da educação, sendo um deles o desempenho dos alunos, que é avaliado por meio das taxas de aprovação ou reprovação.

No estado do RS, as escolas estaduais possuem dados dos alunos que podem ser utilizados não só para fins gerenciais, mas também para descobrir informações relevantes utilizando a Mineração de Dados Educacionais. Um estudo foi realizado para gerar um padrão para a predição dos fatores que levam os alunos à reprovação nos anos finais do Ensino Fundamental. Esse estudo buscou auxiliar gestores e professores a aplicar soluções para melhorias no desempenho dos alunos.

Os experimentos foram realizados com alunos do 6º e 9º ano, buscando interações que levassem a um padrão de reprovação. Foram geradas árvores de decisão em ambos os experimentos, que foram balanceados para obter uma melhor acurácia. Houve um aumento na precisão do atributo REP. No experimento do 6º ano, foi constatado que os alunos apresentavam dificuldades na disciplina de matemática, conforme verificado no portal QEDU sobre aprendizado adequado à sua etapa escolar. A árvore de decisão revelou que a nota do 2º trimestre da disciplina de matemática foi o fator determinante para a reprovação, juntamente com outras decisões. Além da matemática, o algoritmo também identificou um padrão nas disciplinas de artes e religião, e em outro ramo, inglês e geografia.

Durante o experimento do 9º ano, o algoritmo identificou a nota do 2º trimestre de ciências como o nó raiz. Caso um aluno apresente uma nota próxima da média (6,0), a disciplina de matemática é considerada. Por outro lado, se um aluno tem uma nota acima da média e é carente (ou seja, tem Bolsa Família), ele é reprovado, enquanto a disciplina de matemática e educação física são consideradas para tomada de decisão.

O experimento do 9º ano revelou a difícil realidade enfrentada pelos alunos carentes e o papel crucial da equipe diretiva em acompanhá-los de perto. Ambos os experimentos realizados demonstraram que o 2º trimestre é um ponto importante, apesar das diferentes disciplinas envolvidas. Isso ressalta a importância de os gestores estarem atentos aos alunos que podem estar em risco de reprovação e oferecer o suporte necessário.

Para trabalhos futuros, recomenda-se a inclusão de outros atributos além das notas das disciplinas, bem como a realização de questionários para coletar informações adicionais e atributos sobre os estudantes. Além disso, sugere-se a realização de experimentos com outros tipos de algoritmos, além das árvores de decisão, e ajustes nos parâmetros do *J48* para comparações mais precisas.

REFERÊNCIAS

AMARAL, F. **Aprenda mineração de dados:** teoria e prática. Rio de Janeiro, Alta Books, 2016. 240 p.

ALVES, R. D. **Predição do desempenho da redação do Enem utilizando técnicas de mineração de dados.** 2018. 67 p. Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Campus Araranguá, Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2018. Disponível em: <https://repositorio.ufsc.br/bitstream/handle/123456789/187760/EnviadoRepositorioTccRafaelDamianiAlves.pdf?sequence=1&isAllowed=y> Acesso em: 22 jun. 2020.

ASQ; American Society for Quality. **O que é qualidade**. 2021 Disponível em: <https://asq.org/quality-resources/quality-glossary/q>. Acesso em: 17 jan. 2021.

BAKER, R. I. S.; CARVALHO, A.; SEIJI ISOTANI. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, v. 19, n. 02, p. 3-13. 2011. Disponível em: <https://www.br-ie.org/pub/index.php/rbie/article/view/1301/0> Acesso em: 07 jan. 2021.

BAKER, R. I. S.; CARVALHO, A. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. Jornada de Atualização em Informática na Educação – **Anais...** JAIE 2012. Disponível em: <https://www.researchgate.net /publication/275344764> Acesso em: 03 fev. 2021.

BARROS, R. P.; JUNIOR, O. V. S.; SILVA, I. R. M.; SANTOS, L. F.; NETO, V.R.C. Predição do rendimento dos alunos em lógica de programação com base no desempenho das disciplinas do primeiro período do curso de ciências e tecnologia utilizando técnicas de mineração de dados. **Brazilian Journal of Development**, Curitiba, v. 6, n. 1, p. 2523-2534 jan. 2020. Disponível em: <https://repositorio.ufrn.br/handle/123456789/30965> Acesso em: 01 fev. 2020.

CASTANHEIRA, L. G. **Aplicação de Técnicas de Mineração de Dados em Problemas de Classificação de Padrões**. Departamento de Engenharia Elétrica – UFMG. Dissertação de Mestrado em Engenharia Elétrica. 2008. Disponível em: <https://www.ppgee.ufmg.br/defesas/349M.PDF> Acesso em: 09 mar. 2021.

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

CASTRO, M. H. G. **O desafio da qualidade.** 2018. Disponível em: <http://www.rededosaber.sp.gov.br/portais/Portals/18/arquivos/DesafioDaQualidade_cr.pdf> Acesso em: 20 jun. 2020.

CHIRINÉA, A. M.; BRANDÃO C. F. O IDEB como política de regulação do Estado e legitimação da qualidade: em busca de significados. **Ensaio: aval. pol. públ. Educ**., Rio de Janeiro, v. 23, n. 87, p. 461-484, abr./jun. 2015 Disponível em: <https://www.scielo.br/pdf/ensaio/v23n87/0104-4036-ensaio-23-87-461.pdf> Acesso em: 18 jul.2020.

COLPO, M. P.; PRIMO T. T.; PERNAS, A. M.; CECHINEL, C. Mineração de Dados Educacionais na Previsão de Evasão: uma RSL sob a Perspectiva do Congresso Brasileiro de Informática na Educação. SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO 2020. **Anais**... 2020. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/12866>. Acesso em: 01 fev. 2021.

COSTA, E.; BAKER, R. S. J.; AMORIM, L.; MAGALHÃES, J. MARINHO, T. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. JORNADA DE ATUALIZAÇÃO EM INFORMÁTICA NA EDUCAÇÃO - JAIE 2012. **Anais...** 2012. Disponível em: <https://www.researchgate.net/publication/275344764> Acesso em: 01 fev. 2021.

CRETTON, N. N.; GOMES, G. R. R. Aplicação de Técnicas de Mineração de Dados na Base de Dados do ENADE com Enfoque nos Cursos de Medicina. **Acta Biomedica Brasiliensia**. V. 7, 2016. Disponível em: <https://www.actabiomedica.com.br/index.php/acta/article/view/130>. Acesso em: 10 Set. 2022.

DECRETO Nº 55.118, DE 16 DE MARÇO DE 2020. **Estabelece medidas complementares de prevenção ao contágio pelo COVID-19 (novo Coronavírus) no âmbito do Estado**. Disponível em: <https://estado.rs.gov.br/upload/arquivos/decreto-55118.pdf> Acesso em: 20 abr. 2022.

EDUCAÇÃO RS; **Manual do ISE** - Informatização da Secretaria da Educação, 2021. Disponível em: <https://moodle.educacao.rs.gov.br/mod/resource/view.php?id=2252> Acesso em: 16 abr. 2022.

FAYYAD, U.; PIATETSKYSHAPIRO, G.; SMYTH, P. From Data Mining to Knowlegde discovery. **American Association for Artificial Intelligence**. v.17, n. 3, p. 39. 1996

GALAFASSI, F. P.; GLUZ, J. C; GALAFASSI, C. Análise Crítica das Pesquisas Recentes sobre as Tecnologias de Objetos de Aprendizagem e Ambientes Virtuais de Aprendizagem**. Revista Brasileira de Informática na Educação**. 2013. Disponível em: <https://www.br-ie.org/pub/index.php/rbie/article/view/2351>. Acesso em: 18 jul. 2020.

GARCÍA, E., ROMERO, C., VENTURA, S., CASTRO. C. A collaborative educational association rule mining tool. **The Internet and Higher Education**, v. 14, n. 2, p. 77-88, 2011. ISSN 1096-7516. Disponível em: <https://www.academia.edu/20870830/A_collaborative_educational_association_rule_mining_tool> > Acesso em: 18 Ago. 2022.

GONÇALVES, T. C.; SILVA, J. C. DA; CORTES, O. A. C. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão. **Revista Brasileira de Computação Aplicada**. V. 10. n. 3, p. 11-20. 2018. Disponível em:

<http://seer.upf.br/index.php/rbca/article/view/8427/114114337>. Acesso em: 09 mar. 2021.

INEP. **Ministério da Educação.** Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. 2020. Disponível em: <http://portal.inep.gov.br/educacao-basica/saeb>. Acesso em: 07 ago. 2020

JUNIOR, R. N.; NASCIMENTO, R. L. S.; FAGUNDES, R. A. A; NETO, P. S. G. Estimação de Índices de Aprovação e Reprovação Escolar do Ensino Médio. VIII CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (CBIE 2019), XXX SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE 2019). **Anais...** 2019. Disponível em: <https://www.br-ie.org/pub/index.php/sbie/article/view/8738> Acesso em: 07 jan. 2020.

KAMPFF, A.J.C. **Mineração de Dados Educacionais para Geração de Alertas em Ambientes Virtuais de Aprendizagem como Apoio à Prática Docente**. PPGIE/UFRGS, Porto Alegre, 2009.

MARTINS, B. C. **Uma discussão sobre diferentes ambientes de software para mineração de dados.** Monografia apresentada ao Curso de Ciência da Computação da Universidade Federal de Maranhão. São Luiz. 2017. Disponível em: <http://monografias.ufma.br/jspui/handle/123456789/3571> Acesso em: 09 mar. 2021.

NAMEN, A. A.; BORGES, S. X. A.; SADALA, M. G. S. Indicadores de qualidade do ensino fundamental: o uso de tecnologias de mineração de dados e de visões multidimensionais para apoio à análise e definição de políticas públicas. **Revista Brasileira de Estudos Pedagógicos**, Brasília, v. 94, n. 238, p. 667-700, set./dez. 2013. Disponível em: <https://www.scielo.br/j/rbeped/a/NccgVmpnQgJwMNL9Kt8jpvD/?format=pdf&lang=pt> Acesso em: 02 abr. 2022.

NEUHAUS, V. H. **Avaliação Externa:** Mecanismo Mobilizador de uma Educação de Qualidade. Monografia, 2016 UFSM Disponível em: <https://portal.ufsm.br/biblioteca/pesquisa/registro.html?idRegistro=442204> Acesso em: 12 mai. 2020.

NOETZOLD, E.; PERTILE, S. Análise e predição de evasão dos alunos do ensino superior da Universidade Federal de Santa Maria Campus Frederico Westphalen por meio da mineração de dados educacionais. ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO (EATI). **Anais...** 2021. Frederico Westphalen – RS, ano 10 n. 1 p. 52-55 jan. 2021. Disponível em: <http://168.228.253.6:8080/index.php/2019/article/view/49/46>. Acesso em: 15 fev. 2021

OLIVEIRA, M. J. S.; CAETANO, G.; DANIEL, E. M. P. Usando a mineração de dados para predição de desempenho de alunos nas disciplinas de português e matemática. **RELVA**, Juara/MT/Brasil, v. 5, n. 2, p. 8-16, jul./dez. 2018. Disponível em: <https://periodicos.unemat.br/index.php/relva/article/viewFile/3403/2720> Acesso em: 06 jan. 2021.

RAMOS, J. L. C.; SILVA, J. C. S.; RODRIGUES R. L.; OLIVEIRA, P. L. S. CRISP-EDM: uma proposta de adaptação do Modelo CRISPDM para mineração de dados educacionais. SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. 2020. **Anais...** 2020 Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/12865>. Acesso em: 15 fev. 2021

RODRIGUES, R. L.; RAMOS, J. L. C.; SILVA, J. C. S; GOMES, A. S. A literatura brasileira sobre mineração de dados educacionais. 3º CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (CBIE 2014) WORKSHOPS (WCBIE 2014). **Anais...** 2014. Disponível em: <https://www.br-ie.org/pub/index.php/wcbie/article/view/3286/2824> Acesso em: 27 mai.2020.

ROMERO, C.; VENTURA, S. **Educational Data Mining:** A Review of the State of the Art. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 40. 601 - 618. 10.1109/TSMCC.2010.2053532. 2010. Disponível em: <https://www.researchgate.net/publication/224160756_Educational_Data_Mining_A_Review_of_the_State_of_the_Art>. Acesso em: 10 jan. 2021.

SEDUC/RS. **Secretaria de Educação do RS**. 2021. Disponível em: <http://moodle.educacao.rs.gov.br/course/view.php?id=78&section=5>. Acesso em: 02 fev. 2021.

SILVA, D. A. **Aplicação de técnicas de pré-processamento e agrupamento na base de dados de benefícios previdenciários do Ministério Público do Trabalho.** Trabalho de conclusão de curso apresentado a Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais. 2018. Disponível em: <https://repositorio.ufu.br/bitstream/123456789/22118/1/AplicacaoTecnicasPreprocessamento.pdf>. Acesso em: 13 out. 2021.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados**: com aplicações em R. 1. ed. – Rio de Janeiro: Elsevier, 2016.

SILVA, L. A.; SILVA, L. Fundamentos de Mineração de Dados Educacionais. 3º CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (CBIE 2014) Workshops (WCBIE 2014). **Anais...** 2014. Disponível em: <https://www.br-ie.org/pub/index.php/wcbie/article/viewFile/3281/2819> Acessado em: 09 mar. 2021

SILVA, L. A; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados:** com aplicações em R. 1 ed. Rio de Janeiro: Elsevier, 2016.

SILVA, J. A. Qualidade na educação**. Revista: EaD & Tecnologias Digitais na Educação**, Universidade Federal da Grande Dourados. Dourados, MS, 2020, v. 8. n. 10. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788522122462/cfi/0!/4/4@0.00:28.4>. Acesso em: 07 ago. 2020

SILVA, M. D.; DIAS, J. L. Levantamento Bibliográfico sobre Mineração de Dados na Educação: Um Mapeamento Sistemático da Literatura. ENCONTRO ANUAL DE TECNOLOGIA DA INFORMAÇÃO (EATI). 2019. **Anais...** 2019. Disponível em:

<http://anais.eati.info:8080/index.php/2019/article/download/45/42>. Acesso em: 12 jan. 2021.

SOUZA, V. F. Mineração de dados educacionais em um mooc brasileiro. **Revista: EaD & Tecnologias Digitais na Educação**, Universidade Federal da Grande Dourados. Dourados, MS, v. 8, n. 10. 2020. Disponível em: <https://ojs.ufgd.edu.br/index.php/ead/article/view/11461>. Acesso em: 01 fev. 2021.

SOUZA, V. F.; SOUZA M.F. Os avanços da mineração de dados educacionais: definições, processo e evolução. **Brazilian Journal of Development**, Curitiba, v.7, n.8, p. 80798-80816 aug. 2021 Disponível em: <https://www.brazilianjournals.com/index.php/BRJD/article/view/34442/pdf> Acesso em: 01 mar. 2021.

PETERMANN, R. J. – **Modelo de Mineração de dados pra classificação de clientes em telecomunicações**. Pontifícia Universidade católica do RS – Faculdade de engenharia – PPGEE – out. 2006.

ZAKI, M. J. **Parallel and Distributed Data Mining**: An Introduction. Large-Scale Parallel Data Mining. Berlin: Springer-Verlag, 2000. Disponível em: <https://doc.lagout.org/Others/Data%20Mining/Large-Scale%20Parallel%20Data%20Mining%20%5BZaki%20%26%20Ho%202000-02-23%5D.pdf>. Acesso em: 01 mar. 2022.