UNIVERSIDADE DO ESTADO DA BAHIA – UNEB
DEPARTAMENTO DE CIÊNCIAS EXATAS E DA TERRA – I
CURSO DE SISTEMAS DE INFORMAÇÃO

JEFERSON SANTOS DE ALMEIDA

Mineração de Dados Educacionais na base de dados SAEB para suporte na tomada de decisão à Educação Básica

Salvador

2023

JEFERSON SANTOS DE ALMEIDA

MINERAÇÃO DE DADOS EDUCACIONAIS NA BASE DE DADOS SAEB PARA SUPORTE NA TOMADA DE DECISÃO À EDUCAÇÃO BÁSICA

Monografia apresentada ao curso de Sistemas de Informação do Departamento de Ciências Exatas e da Terra da Universidade do Estado da Bahia - UNEB, como requisito parcial à obtenção do grau de bacharel em Sistemas de Informação. Área de Concentração: Ciência da Computação

Orientadora: Maria Inês Valderrama Restovic

Coorientadora: Débora Alcina Rego Chaves

Salvador

2023

Termo de Anuência do Orientador

Declaro para os devidos fins que li e revisei este trabalho e atesto sua qualidade como resultado final desta monografia. Confirmo que o referencial teórico apresentado é completo e suficiente para fundamentar os objetivos propostos e que a metodologia científica utilizada e os resultados finais são consistentes e com qualidade suficiente para submissão à banca examinadora final do Trabalho de Conclusão de Curso I do curso de Bacharelado em Sistemas de Informação. Maria Inês Valderrama Restovic

_____

Orientador(a)

AGRADECIMENTOS

Concluir este projeto para mim representa o fim de um ciclo muito importante, trata-se de um simbolismo tal, que por maior que seja o esforço empenhado, não conseguiria expressá-lo em palavras.

Considero acreditar que na vida as coisas dificilmente têm o fim em si mesmas. Portanto, na minha percepção sempre foi muito claro que não se trata “apenas” de uma monografia que representa o fim do ciclo da graduação.

Trata-se na verdade de um conjunto de acontecimentos, decisões, desafios, oportunidades, pessoas que passaram e outras que ficaram, angústias, felicidades, alguns medos e algumas convicções.

Enfim, o ponto é: o produto de todas essas coisas resultou no dia de hoje, construíram o presente e apontam para um possível futuro, onde, vencer essa etapa representa o primeiro passo. Trata-se de uma jornada em que em nenhum momento caminhei sozinho, e parafraseando o célebre filósofo *“A gratidão é a memória do coração”*. E até aqui, são muitas as pessoas às quais serei eternamente grato!

À Deus agradeço pelo privilégio da vida, da saúde e da energia para seguir enfrentando e vencendo os desafios que surgem durante a jornada.

Aos meus Pais, Cátia e Eufrásio, eu sou grato por tudo! Agradeço por todo sacrifício de vida, por conseguirem cuidar da nossa família no meio de tantas dificuldades e adversidades. Sem vocês não existiria nada! Obrigado!

À minha irmã Jéssica, por estar sempre comigo, sendo uma fonte de apoio e motivação incondicional, por segurar a barra em momentos muito difíceis e por me ajudar a permanecer na rota. Obrigado!

À Jucilene pelo o privilégio de compartilhar a vida, pelo apoio e cobranças, por cuidar de nosso pequeno Théo, por toda paciência e compreensão nos momentos em que não pude ser tão presente. Obrigado!

Aos meus amigos Eugênio, Ramon e Ricardo, por estarmos juntos durante toda essa jornada nos dias e nas madrugadas (rs), até aqui enfrentamos e vencemos muitas batalhas. Vocês foram e são uma grande fonte de apoio e inspiração! Obrigado!

As minhas incríveis Orientadoras Professoras, Maria Inês e Débora Chaves, que para além de terem me ensinado muito, sou grato por terem me dado todo o suporte e as condições para que esse projeto fosse concluído, e um agradecimento especial pela paciência e compreensão durante esta etapa final. Obrigado!

Ao professor Antônio Atta, pela orientação e suporte, tendo sido um guia para chegar ao tema desta pesquisa. Obrigado!

Ao professor Alexandre Lenz, que enquanto esteve coordenador, me ajudou a planejar as etapas da conclusão, bem como, criou as condições necessárias para que tudo acontecesse. Obrigado!

A todos os meus professores e professoras de cada ciclo formativo que vivi até aqui. Sou grato a cada um (a) de vocês, por todo conhecimento, aprendizado e ensinamentos, que muitas vezes foram muito além das salas de aulas. Cada um (a), à sua maneira, e no seu momento, teve uma grande parcela de contribuição até aqui. Muito Obrigado!

Por fim, agradeço a mim mesmo por ter persistido, por ter enfrentado, por ter tido coragem e ousadia para recomeçar, por ter encontrado forças para enfrentar as circunstâncias e ter vencido! Obrigado

Nada até aqui ocorreu por acaso, as coisas são como são e acredito que tudo que aconteceu até aqui possui algum tipo de propósito, alguns deles, talvez só o Jeferson do futuro poderá explicar.

Dedico esse trabalho ao meu avô Jair dos Santos (in memoriam).

- Vô, onde quer que esteja, saiba que, seja qual for a circunstância, eu jamais irei desistir!

“O Desafio que você
não enfrenta, vira o seu limite”
Conrado Adolpho

RESUMO

Este trabalho propõe o desenvolvimento de um ambiente de experimentação acessível a educadores e gestores, que permita a extração de informações úteis para aprimorar a educação básica brasileira, com base nos dados do SAEB. A partir da coleta e tratamento dos dados SAEB 2019 e seguinte, foram implementadas técnicas de mineração de dados para extração de conhecimento desta base. A avaliação foi realizada a partir da análise preliminar da ferramenta com a participação de um grupo de educadores e gestores educacionais. O objetivo final foi disponibilizar um ambiente de experimentação com base nos feedbacks recebidos e contribuir para melhorias no sistema educacional brasileiro a partir do uso de tecnologias.

Palavras-chave: Mineração de Dados; Educação Básica; Mineração de Dados Educacionais; SAEB;

ABSTRACT

This research paper presents a proposition for creating an accessible experimentation environment designed for educators and administrators. The primary objective of this environment is to extract valuable information from SAEB data in order to enhance the quality of Brazilian basic education. The proposed methodology involves collecting and processing SAEB 2019 and subsequent data, and subsequently applying data mining techniques to extract knowledge from this dataset. The evaluation of the experimentation environment will be carried out through a preliminary analysis involving a group of educators and educational administrators. The ultimate aim of this study is to refine and finalize the experimentation environment based on the feedback received, ultimately contributing to the improvement of the Brazilian educational system through the effective utilization of technology.

Key-words: Data Mining; Basic education; Educational Data Mining; SAEB;

LISTA DE FIGURAS

Figura 1- Processo KDD 19
Figura 2 - Funcionamento *K-MEANS* 24
Figura 3 - Metodologia DSR 29
Figura 4 - Dados Antes da Normalização 37
Figura 5 - Dados após Normalização 37
Figura 6 - Curva de Elbow (T5_ESCOLAS_BA.csv) 38
Figura 7 - Chamada da Função de Cálculo do Silhouette Score 39
Figura 8 - Valores do Silhoutte Score para cada valor de K 39
Figura 9 - Importação do *K-MEANS* da Biblioteca Scikit Learn 40
Figura 10 - Treinamento *K-MEANS* utilizando 3 clusters 41
Figura 11 - Reversão da escala dos dados 41
Figura 12 - Criação e Exportação de Base de Dados Clusterizada 42
Figura 13 - Curva de Elbow (T5-ESCOLAS-BA-v2.csv) 43
Figura 14 - Cálculo do *Silhoute Score* (T5-ESCOLAS-BA-v2.csv) 43
Figura 15 - Curva de Elbow (T5-ESCOLAS-BA-v3.csv) 44
Figura 16 - Cálculo do Silhoutte Score (T5-ESCOLAS-BA-v3.csv) 44
Figura 17 - Médias LP e MT por Cluster 45
Figura 18 - Visualização com 5 atributos 46
Figura 19 - Resultados clusterização no segundo cenário 49
Figura 20 - Resultados clusterização no terceiro cenário 50
Figura 21 - Arquitetura Plataforma Quadro 52

LISTAS DE QUADROS

Quadro 1 - Técnicas e Tarefas utilizadas na Mineração de Dados 20
Quadro 2 - Principais ferramentas de mineração de dados 26
Quadro 3 - - Campos da tabela afetados na transformação e limpeza dos dados....35
Quadro 4 - Equivalência numérica campo NÍVEL_SOCIO_ECONOMICO 36

LISTA DE GRÁFICOS

Gráfico 1 - Divisão de Escolas por Proficiência em MT.....47
Gráfico 2 - Divisão de Escolas por Proficiência em LP 48
Gráfico 3 - Relevância das Informações da Plataforma 53
Gráfico 4 - Utilidade da Plataforma 54

LISTA DE ABREVIATURAS E SIGLAS

INEP: Instituto Nacional de Estudos Pedagógicos Anísio Teixeira

SAEB: Sistema de Avaliação da Educação Básica

IDEB: Índice de Desenvolvimento da Educação Básica

PNADC: Pesquisa Nacional por Amostra de Domicílios Contínua

DSR: *Design Science Research*

EDM*: Education Data Mining*

LDBEN: Lei das Diretrizes e Bases da Educação Nacional

BNCC : Base Nacional Comum Curricular

GSP: *Generalized Sequential Patterns*

WEKA: *Waikato Environment for Knowledge Analysis*

FGV: Fundação Getúlio Vargas

IDH: Indice de Desenvolvimento Humano

Sumário

1 INTRODUÇÃO 7

1.1 PROBLEMA DE PESQUISA 8

1.2 OBJETIVO GERAL 9

1.3 OBJETIVOS ESPECIFICOS 9

1.4 Brasil, características e reflexos no seu sistema educativo 10

1.5 ESTRUTURA DO TRABALHO 12

2 A Educação Básica, suas Políticas de Avaliação e as potencialidades à extração de conhecimentos diagnósticos 13

2.1 EDUCAÇÃO BÁSICA 13

2.1.1 ENSINO INFANTIL 13

2.1.2 ENSINO FUNDAMENTAL 13

2.1.3 ENSINO MÉDIO 14

2.1.4 EDUCAÇÃO DE JOVENS E ADULTOS 15

2.2 PROVA BRASIL E SAEB 15

2.3 Índice de Desenvolvimento da Educação Básica (IDEB) 17

2.4 MINERAÇÃO DE DADOS 18

2.4.1 TAREFAS, TÉCNICAS E ALGORITMOS PARA MINERAÇÃO DE DADOS 20

2.4.1.1 TAREFAS E ALGORITMOS 20

2.4.1.2 DESCOBERTA DE ASSOCIAÇÃO 21

2.4.1.3 CLASSIFICAÇÃO 21

2.4.1.4 REGRESSÃO 21

2.4.1.5 AGRUPAMENTO (CLUSTERIZAÇÃO) 22

2.4.1.6 O ALGORITMO *K-MEANS* 22

2.4.2 TÉCNICAS DE MINERAÇÃO DE DADOS 24

2.4.2.1 ALGORITMO GENÉTICO 24

2.4.2.2 REDES NEURAIS 25

2.4.2.3 ÁRVORES DE DECISÃO 25

2.4.2.4 REGRAS DE ASSOCIAÇÃO 25

2.4.2.5 ANALÍSE DE VIZINHANÇA 26

2.5 FERRAMENTAS DE MINERAÇÃO DE DADOS 26
2.6 MINERAÇÃO DE DADOS NA EDUCAÇÃO BÁSICA 27
2.8 TRABALHOS CORRELATOS 28
3 METODOLOGIA 29
3.1 DESENVOLVIMENTO 30
3.1.1 Configuração do Ambiente de desenvolvimento 31
3.1.2 Coleta de Dados 32
3.1.3 Seleção dos Dados 33
3.1.4 Pré-processamento e transformação dos dados 33
3.1.5 Mineração de Dados 36
3.1.6 Definição do número ideal de clusters 38
3.1.6.1 O método do Cotovelo 38
3.1.6.2 Silhouette Score 39
3.1.8 Testes Realizados 40
4 Análise de Resultados 45
4.1 Primeiro Cenário – Clusterização considerando todos os atributos da base dados 45
4.2 Segundo Cenário – Clusterização considerando cinco atributos da base dados 49
4.3 Terceiro Cenário – Clusterização considerando três atributos da base dados 50
4.4 Ambiente de Experimentação 51
4.5 Avaliação da Plataforma 52
5 Considerações Finais 55
6 REFERÊNCIAS 56

1 INTRODUÇÃO

O Instituto Nacional de Estudos Pedagógicos Anísio Teixeira (INEP), vinculado ao Ministério da Educação, desempenha um papel fundamental na coleta de dados e na construção de indicadores educacionais no Brasil. Por meio de avaliações e questionários, o INEP busca obter informações relevantes sobre o sistema educacional do país, visando fornecer subsídios para a tomada de decisões e o desenvolvimento de políticas públicas na área da educação. (INEP, 2023)

De forma detalhada, os dados coletados pelo INEP, podem trazer um diagnóstico profundo a respeito do sistema educacional brasileiro a nível da educação básica, sobre os aspectos do desempenho escolar, infraestrutura, formação docente, questões socioeconômicas, entre outros. Como educação básica, compreende-se a educação infantil, ensino fundamental I, II e o ensino médio.

Para coleta dos dados diagnósticos, o INEP realiza ações em colaboração com as secretarias de educação municipais e estaduais, e prevê a participação de todas as escolas públicas e privadas de todas as etapas do ensino básico.

Dentre as diversas iniciativas promovidas pelo INEP, destaca-se a execução do Sistema de Avaliação da Educação Básica (SAEB), pioneira no contexto nacional e desenvolvida com o propósito de investigar minuciosamente o panorama educacional brasileiro.

O SAEB foi desenvolvido no fim dos anos 80 e teve sua primeira aplicação em 1990. Cinco anos depois, o sistema passou por uma reestruturação que permitiu a comparação dos resultados obtidos ao longo dos anos. Desde a sua primeira aplicação, este sistema de avaliação fornece dados sobre a qualidade dos sistemas educacionais do Brasil como um todo, das regiões demográficas e das unidades federadas (Estados e Distrito Federal). (SAEB, 2023).

A partir dos dados gerados por esse sistema, surgiu o Índice de Desenvolvimento da Educação Básica (IDEB). Como um dos mais significativos índices educacionais, o IDEB se destaca como uma ferramenta essencial na avaliação da qualidade da educação básica.

Todos os dados produzidos a partir dessas ações são disponibilizados ao público por meio de uma página de dados abertos, acessível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb/resultados (Acesso em: 30 jun. 2023).

O volume de dados é grande e cresce a cada ano; e a partir da análise desses dados, tem sido possível a descoberta de informação e produção de conhecimento para gestores, de modo a tornar possível a tomada de decisões com base em evidências.

Para subsidiar essas análises, diversas pesquisas têm sido conduzidas no campo da mineração de dados. As técnicas de Mineração de Dados buscam identificar relações entre dados e, a partir desse processo, criar informações que podem produzir novos conhecimentos para o desenvolvimento da ciência, bem como, para a tomada de decisões baseadas em evidências. Dessa forma, subsidiam novas ações e transformam a realidade atual. (BARBOSA, *et al*. 2017).

A complexidade, o tamanho e a vastidão dos dados disponíveis a respeito do sistema educacional brasileiro, deu origem a abordagem de pesquisa especifica: Mineração de Dados Educacionais (MDM). (SOARES, *et al*. 2021).

Ao examinarmos as pesquisas deste campo nos últimos cinco anos (2017 – 2022), constata-se que, apesar da quantidade crescente de trabalhos, é evidente a ausência de padrões específicos para realização dessas análises, bem como, ferramentas acessíveis que possibilitem que Educadores e Gestores consigam tomar decisões a partir de coletas quantitativas de dados sobre o desempenho dos estudantes.

1.1 PROBLEMA DE PESQUISA

Com base no contexto de pesquisa apresentado, este trabalho investigativo é norteado a partir do seguinte questionamento: Como as técnicas de mineração de dados podem ser aplicadas visando o redimensionamento dos dados de resultados da Prova Brasil-SAEB, de forma a suportar a tomada de decisões de gestores e educadores de escolas, a partir das bases de dados históricos públicos de resultados nacionais do exame?

1.2 OBJETIVO GERAL

O problema de pesquisa enunciado a cima, por sua vez, orienta os seguintes objetivos geral e específicos.

Elencar técnicas de mineração de dados aplicáveis aos resultados SAEB a partir de 2019 para subsidiar a tomada de decisão no âmbito da educação básica.

1.3 OBJETIVOS ESPECIFICOS

i. Identificar os principais algoritmos de mineração de dados aplicáveis aos dados SAEB 2019 e seguintes;
ii. Realizar o tratamento e preparação dos dados do SAEB a partir de 2019, garantindo a qualidade e consistência dos mesmos;
iii. Escolher o algoritmo que apresente o melhor resultado para a mineração de dados do SAEB. Com base na análise realizada no objetivo anterior, selecionar o algoritmo mais adequado e eficiente para extração conhecimento a partir dos dados SAEB, visando a tomada de decisão na educação básica;
iv. Construir um ambiente de experimentação acessível a educadores e gestores, que permita a extração de informações úteis para aprimorar a educação básica com base nos dados do SAEB.

1.4 Brasil, características e reflexos no seu sistema educativo

Um país de dimensões continentais, alto e crescente índice populacional, auto suficiente em água e petróleo, uma vasta biodiversidade que carrega consigo riquezas naturais de causar inveja a qualquer país no mundo. Esse, é o Brasil.

Ao descrevermos suas características, imaginamos que estamos nos referindo a um país com todas as condições necessárias para “dar certo” e se destacar no cenário mundial.

Entretanto, quando mergulhamos nos dados do Brasil real, encontramos um país que vive sob a *égide* de ser destaque no mundo por sua desigualdade social. Uma população empobrecida e índices de violência superiores aos de países em guerra.

É sabido que boa parte desses problemas, são reflexos históricos da “Herança Maldita da Escravidão” (ALMEIDA, 2023), e do próprio processo de colonização exploratória no qual esse mesmo Brasil, foi construído.

Ora, é sabido também que um dos remédios para redução das desigualdades é a educação, e por mais repetitivo e por toda obviedade trazida nesta afirmação, um estudo realizado pela Fundação Getúlio Vargas (FGV) em 2019, reforça que a educação foi a principal responsável pela mobilidade social dos últimos 30 anos. (FGV, 2019).

Reconhecer esse pequeno e importante avanço, nos coloca na iminência do debate a respeito dos próximos passos a serem dados no caminho de uma transformação maior.

Nas últimas décadas, a prioridade foi garantir a universalização do ensino básico e o acesso e a permanência escolar, sobretudo, das camadas mais pobres da sociedade. Hoje, para além dessas garantias, faz-se necessária a inclusão do debate a respeito da qualificação dessa educação, para que a mesma corresponda as demandas da era nomeada “Economia do Conhecimento”. (CASTELLS, 1999). Dito isto, um dos caminhos possíveis para alcançarmos esse novo momento é desenvolvermos a compreensão de que existem vários “Brasis”, como defendeu o Antropólogo Darcy Ribeiro. (RIBEIRO, 1995).

Nas continentalidades do nosso Brasil, para além da diversidade étnica, climática e cultural, os níveis de desigualdades podem ser extremamente diferentes entre uma cidade e outra. Segundo a Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC), em nosso País encontramos cidades como São Caetano do Sul em São Paulo, com o Indice de Desenvolvimento Humano (IDH) semelhante ao de Portugal (0,86); e Melgaço no Pará, com IDH (0,42), semelhante ao da Nigéria, um dos países mais pobres do mundo.

Um contraste muito similar é representado adicionalmente, através dos dados da educação da básica (SAEB, 2021), nos quais encontramos cidades com notas no IDEB que podem variar entre 2.8 e 9.1, dentro da máxima de 10 pontos.

Se em um mesmo país é apresentado realidades tão distintas, não é razoável imaginarmos que será possível resolver os desafios colocados, a partir de uma solução geral que funcione em cada um desses “Brasis”.

É necessário compreendermos com profundidade, as problemáticas e complexidades locais e regionais, e com isto, produzirmos soluções que atendam a esta realidade ou seja, tratar diferentes de forma diferente.

Nesse sentido, o poder público e a sociedade precisam estar munidos de tecnologias e ferramentas que possibilitem análises de tal natureza, e a partir das informações e conhecimento produzidos, tornar possível a existência de soluções eficientes baseadas em evidência.

Este não é um trabalho que tem por finalidade pesquisar a área da mineração de dados no contexto educacional brasileiro, mas sim, produzir uma semente que ao florescer, consiga entregar para sociedade um caminho que possibilite iniciarmos uma mudança de pensamento para elaboração de políticas educacionais. E com isto, no meio de tantos desafios conhecidos, enfrentarmos com mais força, energia e dados, o que acredito ser o ponto de partida para outras mudanças.

1.5 ESTRUTURA DO TRABALHO

Esse trabalho está estruturado em 5 capítulos: no Capítulo 1 foi apresentada a introdução, a contextualização do tema, a formulação do problema de pesquisa, os objetivos do trabalho (geral e especifico), a justificativa para a realização da pesquisa e a estrutura da mesma.

O Capítulo 2 apresenta uma revisão de literatura abrangente sobre a educação básica, a Prova Brasil, o SAEB, mineração de dados e a mineração de dados educacionais (EDM), além dos trabalhos relacionados.

O Capítulo 3 descreve em detalhes a metodologia da pesquisa, bem como, as etapas de desenvolvimento do projeto.

O Capítulo 4, descreve a análise dos resultados encontrados, o desenvolvimento; e avaliação do ambiente de experimentação

O Capítulo 5 apresenta as considerações finais; e os resultados alcançados com a conclusão desse estudo.

Por fim, no Capítulo 6, são listadas todas as referências utilizadas para construção desta pesquisa.

2 A Educação Básica, suas Políticas de Avaliação e as potencialidades à extração de conhecimentos diagnósticos

Nesta seção, aprofundaremos a fundamentação teórica que contextualiza a pesquisa e norteia seus encaminhamentos

2.1 EDUCAÇÃO BÁSICA

A educação básica é o nível inicial do sistema educacional brasileiro, compreendendo a educação infantil, o ensino fundamental e o ensino médio. Esta etapa da educação, tem como objetivo assegurar ao indivíduo uma formação comum, indispensável para o exercício da cidadania, além de fornecer meios para progressão no trabalho e em estudos posteriores. (LDB,1996).

A Lei n º 9.394/96 conhecida como Lei de Diretrizes e Bases da Educação Nacional (LDBEN), estabelece de maneira detalhada o modelo educacional adotado no Brasil, atribuindo as competências e responsabilidades da União, Estados e Municípios, além dos objetivos esperados em cada um dos três níveis.

2.1.1 ENSINO INFANTIL

A educação infantil, primeira etapa da educação básica, tem como finalidade o desenvolvimento integral da criança até seis anos de idade, em seus aspectos físico, psicológico, intelectual e social, complementando a ação da família e da comunidade.

De acordo a LBDEN, a educação infantil deverá ser oferecida em: creches, ou entidades equivalentes, para crianças de até três anos de idade e pré-escolas, para as crianças de quatro a seis anos de idade. Nessa etapa do ensino, a avaliação é focada no registro do desenvolvimento da criança, sem o objetivo de promoção, mesmo para o acesso ao ensino fundamental.

2.1.2 ENSINO FUNDAMENTAL

O ensino fundamental, com duração mínima de oito anos, obrigatório e gratuito na escola pública, tem por objetivo:

I. A formação básica do cidadão, mediante o desenvolvimento da capacidade de aprender, tendo como meios básicos o domínio da leitura, da escrita e do cálculo;

II. A compreensão do ambiente natural e social, do sistema político, da tecnologia, das artes e dos valores em que se fundamenta a sociedade;

III. O desenvolvimento da capacidade de aprendizagem, tendo em vista a aquisição de conhecimentos e habilidades e a formação de atitudes e valores, o fortalecimento dos vínculos de família, dos laços de solidariedade humana e de tolerância recíproca em que se assenta a vida social.

Nessa etapa é facultada aos sistemas de ensino, a divisão do ensino fundamental em ciclos, sendo ministrado em língua portuguesa, assegurada às comunidades indígenas a utilização de suas línguas maternas e processos próprios de aprendizagem.

Esta etapa do ensino, deve obrigatoriamente ser presencial, sendo o ensino a distância utilizado como complementação da aprendizagem ou em situações emergenciais. (LDBEN, 1996).

O ensino religioso, é uma das partes integrantes desta etapa da educação. Nesse aspecto, a lei busca tornar possível a diversidade cultural e religiosa do Brasil, sendo vedada quaisquer formas de proselitismo.

Como tentativa de garantir a laicidade constitucional do estado, a LDBEN atribui aos sistemas de ensino a responsabilidade de ouvir as entidades civis, constituída pelas diferentes denominações religiosas para a definição dos conteúdos do ensino religioso.

2.1.3 ENSINO MÉDIO

O ensino médio, etapa final da educação básica, com duração mínima de três anos, tem como finalidades:

I. A consolidação e o aprofundamento dos conhecimentos adquiridos no ensino fundamental, possibilitando o prosseguimento de estudos;
II. A preparação básica para o trabalho e a cidadania do educando, para continuar aprendendo, de modo a ser capaz de se adaptar com flexibilidade a novas condições de ocupação ou aperfeiçoamento posteriores;
III. O aprimoramento do educando como pessoa humana, incluindo a formação ética e o desenvolvimento da autonomia intelectual e do pensamento crítico;
IV. A compreensão dos fundamentos científico-tecnológicos dos processos produtivos, relacionando a teoria com a prática, no ensino de cada disciplina.

Será nessa etapa da educação básica que o educando deverá adquirir o domínio dos princípios científicos e tecnológicos que presidem a produção moderna,

conhecimento das formas contemporâneas de linguagem, o domínio dos conhecimentos de Filosofia e de Sociologia necessários ao exercício da cidadania. (LDBEN, 1996).

2.1.4 EDUCAÇÃO DE JOVENS E ADULTOS

A educação de jovens e adultos é destinada àqueles que não tiveram acesso ou continuidade de estudos no ensino fundamental e médio na idade própria. Os sistemas de ensino deverão assegurar gratuitamente aos jovens e aos adultos, que não puderam efetuar os estudos na idade regular, oportunidades educacionais apropriadas, consideradas as características do aluno, seus interesses, condições de vida e de trabalho. (LDBEN, 1996)

Ao Poder Público caberá o dever viabilizar e estimular o acesso e a permanência do trabalhador na escola, mediante ações integradas e complementares entre si. De modo a garantir o prosseguimento dos estudos em caráter regular, cabe aos sistemas de ensino, a responsabilidade de manter cursos e exames supletivos que compreendam a Base Nacional Comum Curricular. (BNCC). (LDBEN, 1996).

2.2 PROVA BRASIL E SAEB

A Prova Brasil e o Sistema Nacional de Avaliação da Educação Básica (SAEB), são avaliações para diagnóstico em larga escala, desenvolvidas pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep/MEC). Têm o objetivo de avaliar a qualidade do ensino oferecido pelo sistema educacional brasileiro, a partir de testes padronizados e questionários socioeconômicos. (SAEB, 2023).

Para realização da Prova Brasil, são aplicados testes na quarta e oitavas séries (quinto e nono anos) do ensino fundamental. Os estudantes respondem a itens (questões) de língua portuguesa, com foco em leitura, e matemática, com foco na resolução de problemas. No questionário socioeconômico, os estudantes fornecem informações sobre fatores de contexto que podem estar associados ao desempenho.

O SAEB é um conjunto de avaliações externas em larga escala, que permite ao INEP realizar um diagnóstico da educação básica brasileira e de fatores que podem interferir no desempenho do estudante.

Por meio de testes e questionários, aplicados a cada dois anos na rede pública e em uma amostra da rede privada, o SAEB visa mensurar os níveis de aprendizagem demonstrados pelos estudantes avaliados, explicando esses resultados a partir de uma série de informações contextuais.

O SAEB permite que as escolas e as redes municipais e estaduais de ensino, avaliem a qualidade da educação oferecida aos estudantes, e tem como resultado, um indicativo da qualidade do ensino brasileiro fornecendo subsídios para o acompanhamento, desenvolvimento e aprimoramento de políticas educacionais com base em evidências.

As avaliações educacionais de larga escala surgiram na pauta política brasileira no final dos anos 1980, influenciadas pelo processo de redemocratização do país e por tendências internacionais. (COELHO, 2008)**.**

Nesse contexto, o Brasil passou a desenvolver avaliações para mensurar a qualidade da Educação ofertada nas escolas públicas e privadas e, em 1990, implementou o SAEB. Esse sistema foi projetado inicialmente para avaliar três dimensões dos sistemas de ensino por meio de indicadores educacionais, indicadores da escola e indicadores do sistema de gestão educacional. (BASSO,2017).

Desde o seu surgimento, a aplicação do SAEB passou por uma série de aprimoramentos teórico-metodológicos ao longo das edições. A edição de 2019, marca o início de um período de transição entre as matrizes de referência utilizadas desde 2001 e as novas matrizes elaboradas em conformidade com a BNCC.

Nesse novo formato, duas grandes mudanças ocorreram com a aplicação das avaliações. A Educação Infantil passa a ser avaliada também, e os testes foram aplicados aos professores ao invés dos alunos. (GOMES, 2019).

A partir de 2019, os questionários de avaliação foram aplicados na creche, pré-escola, 2º ano, 5º ano, 9º ano do Fundamental e 3º ano do Ensino Médio.

No Fundamental I, a aplicação do SAEB passou a ser feita com os alunos do 2º ano, ao invés do 3º ano, como acontecia, porque a meta para a alfabetização estabelecida pela BNCC, foi antecipada para alunos de sete anos, idade correspondente ao 2º ano.

No Fundamental II, além de Matemática e Língua Portuguesa, os alunos do 9º ano foram avaliados também nas áreas de Ciências na Natureza e Ciências Humanas. Essa inclusão, segundo o INEP, não interferirá na nota do IDEB, principal indicador da educação no país.

Desde a edição de 2017, escolas particulares também podem aderir à avaliação para alunos do Ensino Médio em caráter facultativo. Outra novidade desse novo processo, é o projeto piloto das avaliações em formato digital, além da análise para a elaboração de instrumentos avaliativos de habilidades socioemocionais. (GOMES, 2019).

Apesar do SAEB ter como um dos seus objetivos a utilização dos dados para subsidiar políticas públicas, há um debate importante sobre o uso efetivo dos resultados produzidos na formulação de políticas públicas educacionais. Algo considerado em alguns estudos, como um desafio para Estados e gestores. (KELLAGHAN *et al*., 2011).

ROSISTOLATO *et al*, (2018), corroboram essa análise e apontam que a formação dos gestores para compreensão e uso dos dados é uma das principais dificuldades para a implementação de avaliações de larga escala.

2.3 Índice de Desenvolvimento da Educação Básica (IDEB)

O IDEB foi criado em 2007 e reúne, em um só indicador, os resultados de dois conceitos igualmente importantes para a qualidade da educação: o fluxo escolar e as médias de desempenho nas avaliações. (IDEB, 2023).

O IDEB é calculado a partir dos dados sobre aprovação escolar, obtidos no Censo Escolar, e das médias de desempenho SAEB, sendo, portanto, um dos principais indicadores educacionais. o IDEB tem por objetivo medir a qualidade da educação básica na esfera municipal e estadual.

Esse indicador agrega ao enfoque pedagógico das avaliações em larga escala a possibilidade de resultados sintéticos, facilmente assimiláveis, e que permitem traçar metas de qualidade educacional para os sistemas.

O índice varia de 0 a 10, a combinação entre fluxo e aprendizagem tem o mérito de equilibrar as duas dimensões: se um sistema de ensino retiver seus alunos para obter resultados de melhor qualidade no SAEB, o fator fluxo será alterado, indicando a necessidade de melhoria do sistema. (IDEB, 2023).

A partir das informações do SAEB e da Prova Brasil, o MEC e as secretarias estaduais e municipais de Educação, podem definir ações voltadas ao aprimoramento da qualidade da educação no país e a redução das desigualdades existentes, promovendo, por exemplo, a correção de distorções e debilidades identificadas e

direcionando seus recursos técnicos e financeiros para áreas identificadas como prioritárias.

2.4 MINERAÇÃO DE DADOS

Ao redor do mundo são gerados a cada instante, uma imensa quantidade de conteúdo digital de todos os tipos e formatos, e em variadas fontes disponíveis. Esse contexto faz referência ao termo de *Big Data*, traduzindo no português, Megadados ou Grandes Dados, que consiste no volume, na velocidade de atualização, na variedade de formatos, na veracidade e no valor de toda essa gama de dados que são criados em fotos, vídeos, áudios, textos e etc. (CAVIQUE, 2014).

No tocante ao crescimento cada vez maior de dados, faz-se necessário a aplicação de técnicas e algoritmos que visam investigar esses dados em busca de informações relevantes em repositórios, principalmente no que tange a tomada de decisões, e entre essas técnicas de investigação encontra-se a Mineração de Dados (PATRICIO *et al*. 2018).

A técnica da mineração de dados do Inglês *Data Mining*, é o processo de descoberta de padrões interessantes, inovadores e desconhecidos assim como de modelos descritivos, compreensíveis e preditivos a partir de dados em grande escala. (ZAKY, *et al*. 2014).

De acordo com CÔRTES, *et al*. (2002), a Mineração de dados é uma etapa de um processo maior para descoberta de informações denominado de *Knowledge Discovery in Database* (KDD), ou, Busca de Conhecimento em Banco de Dados.

FAYYAD *et al.* (1996) divide o processo de KDD em seis passos:

I. Preparação dos Dados: consiste em incluir o conhecimento relevante para a aplicação além de definir quais as metas que o processo precisa atingir;
II. Limpeza dos Dados: consiste em retirar os dados que possam distorcer a análise. Assim, utiliza estratégias para remover ruídos, tratar atributos perdidos e até mesmo métodos de transformação para diminuir o número de variáveis envolvidas no processo, visando com isso melhorar o desempenho do algoritmo de análise;
III. Seleção de Dados: consiste em escolher sobre qual conjunto ou subconjunto de dados em que o processo será aplicado;

IV. *Data Mining*: consiste em decidir qual tarefa de *data mining* será aplicada para atingir os objetivos do processo e qual a melhor técnica a ser utilizada (ver seção 3);

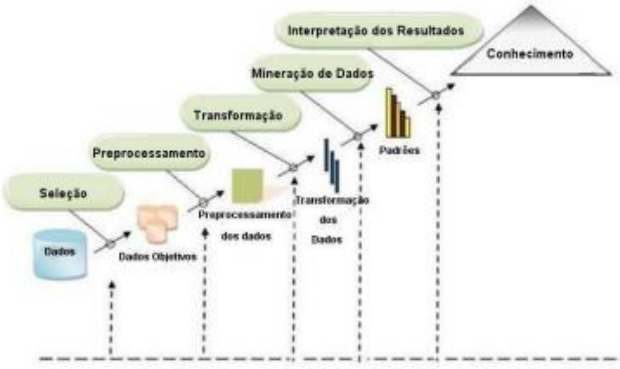
V. Incorporação do conhecimento anterior: consiste em interpretar o modelo descoberto a fim de verificar sua acuracidade em busca de melhorias, possibilitando o retorno para qualquer etapa anterior do processo, retirando padrões redundantes ou irrelevantes;

VI. Interpretação dos resultados: nesse ponto o resultado obtido é incorporado ao sistema, possibilitando a tomada de ações baseadas no conhecimento ou documentando-os e relatando-o às partes interessadas.

As técnicas de Mineração de Dados buscam identificar relações entre dados; e a partir desse processo, criar informações que podem produzir novos conhecimentos para o desenvolvimento da ciência, assim como, também para tomada de decisões baseadas em evidências. E desta forma, subsidiar novas ações e transformar a realidade atual. (BARBOSA *et al*. 2017).

Na figura 1, é ilustrada a mineração de dados como parte de um processo maior de descoberta. (KDD).

Figura 1- Processo KDD



Fonte: FAYYAD *et al* (1996)

2.4.1 TAREFAS, TÉCNICAS E ALGORITMOS PARA MINERAÇÃO DE DADOS

Nesta seção, detalhamos as tarefas, técnicas e algoritmos utilizados para mineração de dados.

2.4.1.1 TAREFAS E ALGORITMOS

A tarefa é um tipo de Mineração de Dados com um propósito particular, da qual existem diversas, ou até dezenas, de implementações distintas por meio de vários algoritmos. Os algoritmos são divididos pelas tarefas levando em consideração o objetivo da implementação, ou seja, os algoritmos de uma mesma tarefa possuem a mesma finalidade. (FURLAN, 2018).

O Quadro 1 exemplifica a relação das tarefas, técnicas e exemplos de algoritmos para mineração de dados.

Quadro 1 - Técnicas e Tarefas utilizadas na Mineração de Dados

Técnica	Descrição	Tarefas	Exemplos
Árvore de Decisão	Baseada em estágios de decisão (nós) e na separação de classes e subconjuntos, organiza os dados de forma hierárquica	-Classificação -Predição	CART, CHAID, C5.0, ID-3.
Redes Neurais	Modelos inspirados na fisiologia do cérebro, nos quais o conhecimento é fruto do mapa de conexões neuronais e dos pesos dessas conexões.	- Classificação - Agrupamento - Predição	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Coumterorioagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognition, Rede BSB.
Raciocínio baseado em casos	Baseado no método do vizinho mais próximo combina e compara atributos para estabelecer hierarquia e semelhança	-Classificação -Agrupamento	BIRCH, CLARANS CLIQUE
Algoritmos Genéticos	Métodos gerais de busca e otimização inspirados na Teoria da Evolução em que a cada nova geração, soluções melhores têm mais chances de ter “descendente”	-Classificação -Agrupamento	Algoritmo Genético Simples,Genitor, G.A-Nuggets, GAPVMINER
Conjuntos Fuzzy	Oferece uma grande vantagem para classificar dados com alto nível de abstração	-Classificação -Agrupamento	*K-MEANS* FCMdd
Regras de Indução	Processo para obter uma hipótese a partir de dados e fatos já existentes	-Classificação -Predição	CART, CHAID
Regras de Associação	Estabelece uma correlação estatística entre atributos de dados e conjuntos de dados	-Associação	Apriori, AprioriTid, Apriori-Hybrid, AIS, SETM

Fonte: (GOLDSCHIMIDT, 2005 – Adaptado pelo Autor)

2.4.1.2 DESCOBERTA DE ASSOCIAÇÃO

Nesta tarefa, os registros do conjunto de dados são chamados de transação, que é formada por um grupo de itens. A tarefa de descoberta de associação é caracterizada por buscar itens que constantemente aparecem juntos em transações do conjunto de dados. Exemplos de algoritmos implementados nessa tarefa são: Apriori*, Generalized Sequential Patterns* (GSP), entre outros. (ZAKI, 2000).

Na vida cotidiana, a descoberta de associações em transações é evidente em situações como compras em supermercados, onde itens são frequentemente adquiridos em conjunto, como pão e leite. Em ambientes online, a análise de padrões de navegação, possibilita sugerir produtos relacionados, demonstrando a aplicabilidade prática dos algoritmos de descoberta de associação.

2.4.1.3 CLASSIFICAÇÃO

Nesta tarefa, existem dois tipos de atributos no conjunto de dados, o primeiro é do tipo atributo previsor e o segundo é chamado de atributo-alvo. Para os valores que diferem do atributo-alvo, existe uma classe que faz referência a um grupo categórico relacionado a um conjunto predefinido.

O objetivo da tarefa de classificação, é descobrir uma função que relacione um conjunto de registros com determinado conjunto de classes. Descoberta essa função, é possível aplicá-la para novos registros. Dessa forma, consegue-se prever a qual classe esse registro pertence. Exemplos de técnicas que podem ser aplicadas na tarefa de classificação são: Rede Neurais, Algoritmos Genéticos e Lógica Nebulosa. (MICHIE *et al*.1994).

2.4.1.4 REGRESSÃO

A tarefa de Regressão é parecida à tarefa de Classificação, onde são buscadas funções que relacionem os registros de uma base de dados em um intervalo de valores reais. A principal diferença, é que o atributo-alvo adota valores numéricos. A tarefa de Regressão utiliza-se da Estatística, Rede Neurais entre outras técnicas que oferecem os recursos para sua implementação. (MICHIE *et al*. 1994).

Na prática cotidiana, a Regressão encontra aplicação em situações como a precificação de imóveis, onde variáveis como localização e tamanho desempenham

papel fundamental na determinação do valor. Da mesma forma, na previsão de vendas, incorporando fatores como publicidade e sazonalidade, a Regressão assume um papel crucial. Esses exemplos evidenciam a importância prática dessa técnica na modelagem e previsão de eventos que envolvem valores numéricos.

2.4.1.5 AGRUPAMENTO (CLUSTERIZAÇÃO)

Compreende um processo de partição dos elementos de um banco de dados em subconjuntos ou *clusters*, de uma maneira que os registros que são semelhantes, fiquem agrupados diferenciando dos registros dos outros subconjuntos. Diferentemente da tarefa de classificação, não existem classes pré-definidas, os elementos são reunidos baseados na similaridade entre eles. (FAYYAD *et al*. 1996).

Os algoritmos *k-Modes, K-MEANS, k-Prototypes* entre outros são usados na implementação dessa tarefa.

2.4.1.6 O ALGORITMO *K-MEANS*

De acordo com a literatura disponível, o algoritmo *K-MEANS,* é considerado ideal para realização de agrupamentos, sendo o mais popular para execução dessa tarefa. Considerando essas particularidades e a disposição dos dados SAEB de maneira numérica e padronizada, o *K-MEANS* foi escolhido como algoritmo utilizado para realização do processo de clusterização nessa base de dados, por esse motivo, essa seção irá detalhar o funcionamento do algoritmo.

O algoritmo *K-MEANS*, também denominada de K-médias, é um popular algoritmo de clusterização utilizado em abordagens de aprendizado de máquina não supervisionado. Segundo JAIN *et. al.* (1999). O algoritmo *K-MEANS* é popular devido a sua facilidade de implementação e sua ordem de complexidade O(n), onde n é o número de padrões.

Segundo PIMENTEL *et. al.* (2003), O *K-MEANS* é uma técnica que usa o algoritmo de agrupamento de dados por K-médias. O objetivo deste algoritmo, é encontrar a melhor divisão de P dados em K grupos Ci, i = 1, ... K, de maneira que a distância total entre os dados de um grupo e o seu respectivo centro, somada por todos os grupos, seja minimizada.

Esse método consiste na utilização de valores iniciais dos primeiros N casos como estimativas temporárias das médias dos k clusters. Os centros iniciais são formados ao redor dos dados mais próximos e, através de um processo contínuo e interativo, encontram-se os centros dos clusters finais. (MONTEIRO *et al*., 2001).

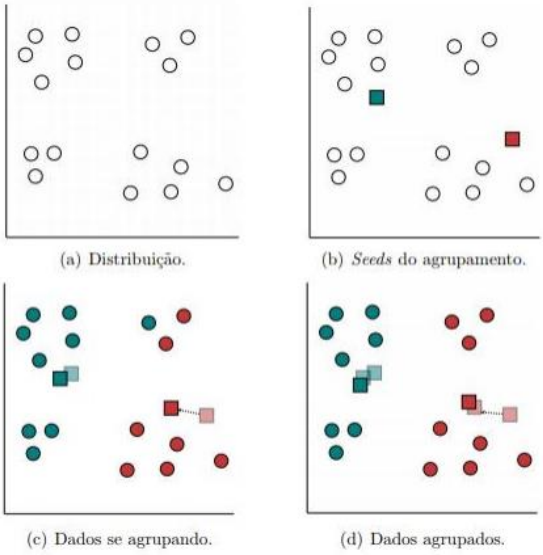
Em outras palavras, o algoritmo atribui aleatoriamente os P pontos a K grupos, e calcula as médias dos vetores de cada grupo. Em seguida, cada ponto é deslocado para o grupo correspondente ao vetor médio do qual ele está mais próximo. Com este novo rearranjo dos pontos em K grupos, novos vetores médios são calculados. O processo de realocação de pontos a novos grupos, cujos vetores médios são os mais próximos deles, continua até que se chegue a uma situação em que todos os pontos já estejam nos grupos dos seus vetores médios mais próximos. (PIMENTEL *et al*, 2003).

O algoritmo *K-MEANS* é aplicado em diversos cenários, onde objetivo é encontrar possíveis padrões em dados não rotulados, como no *marketing* para segmentar clientes e desenvolvimento de campanhas mais assertivas, na medicina para diagnosticar pacientes, em sistemas de recomendação para agrupar usuários com interesses semelhantes, facilitando a recomendação de conteúdo personalizado, tornando-o valioso na análise de dados e na tomada de decisões.

A Figura 2, ilustra o funcionamento do algoritmo *K-MEANS*. A partir da figura é possível observar as etapas do agrupamento, onde inicialmente são atribuídos aleatoriamente centroides para os clusters. A partir disso, calcula-se a distância entre cada ponto e os centroides, associando os pontos ao centroide mais próximo.

Em seguida, recalcula os centroides como a média dos pontos em cada cluster. O processo se repete, reatribuindo pontos aos centroides mais próximos e atualizando os centroides até a convergência do algoritmo.

Figura 2 - Funcionamento *K-MEANS*



Fonte: (PRADO, 2008)

2.4.2 TÉCNICAS DE MINERAÇÃO DE DADOS

As técnicas de mineração de dados são os fundamentos computacionais que possibilitam a construção dos algoritmos que realizam a busca por padrões nos dados. Diversas técnicas podem ser utilizadas para atender a uma tarefa de mineração de dados. Entretanto, cada técnica possui características específicas e é necessário ter o conhecimento do funcionamento e do objetivo das mesmas para interpretar os resultados obtidos. (FRACALANZA, 2009).

A seguir serão apresentadas algumas técnicas e algoritmos para mineração de dados.

2.4.2.1 ALGORITMO GENÉTICO

Simulando o processo natural da evolução, os Algoritmos Genéticos (AGs), têm como objetivo realizar a busca e otimização na descoberta de padrões. Ao contrário dos métodos convencionais com o mesmo propósito, os AGs trabalham simultaneamente em conjuntos de soluções distintas, realizando pesquisas adaptativas nos dados e modelando uma solução específica para um problema em estruturas de dados semelhantes a cromossomos. Operadores são aplicados, recombinando essas estruturas; e gerando novas combinações de regras de associação.

Essa técnica é utilizada na classificação e na segmentação de dados, formulando hipóteses sobre a dependência dos atributos dos dados, com operadores de mutação e cruzamento, desenvolvem várias mutações para a solução do problema. Ao longo do tempo, o algoritmo tende a “aprender” e a se aperfeiçoar, de maneira que somente as soluções com maior poder de acerto na previsão são aceitas. (FRACALANZA, 2009).

2.4.2.2 REDES NEURAIS

As técnicas de Redes Neurais são bastante utilizadas em tarefas de classificação, regressão e segmentação. Os dados são trabalhados com base no funcionamento do cérebro humano, aprendendo a tomar decisões baseadas nas experiências anteriores – nas instâncias anteriores dos dados. Os neurônios do cérebro são representados por nodos que estão conectados em outros nodos por sinapses, formando uma rede de processamento. (FRACALANZA, 2009).

Os valores das entradas são multiplicados nos neurônios pelos pesos de suas sinapses, conforme vão caminhando na rede. Ao final, temos a classificação ou a previsão da entrada.

2.4.2.3 ÁRVORES DE DECISÃO

As árvores de decisão têm como objetivo principal dividir as instâncias em classes. Cada nó da árvore testa o domínio de uma variável da entrada e o redireciona para o nó seguinte. Cada sub-árvore representa o resultado de um teste e a folha é a classificação que aquele registro recebeu. Ao final, cada nó terminal terá os registros da entrada que se adequam as regras regidas por esse nó, representando assim, uma classe. (FRACALANZA, 2009).

2.4.2.4 REGRAS DE ASSOCIAÇÃO

Basicamente, as regras de associação são definidas por uma correlação estatística entre alguns atributos da entrada, com o objetivo de descobrir relações que ocorrem em comum dentro de um conjunto de dados. Cada registro é visto como uma transação e cada variável como um item dessa transação, deixando subentendido que a presença de um item implica necessariamente na presença de outro na mesma transação. (FRACALANZA, 2009).

2.4.2.5 ANALÍSE DE VIZINHANÇA

Através de uma função definida para determinar a “distância“ entre duas instâncias, ou seja, de uma função para identificar um conjunto de registros que estão próximos por determinada característica, essa técnica é empregada na análise de prognósticos e não para descoberta de conhecimento. Não é muito explorada na literatura.

2.5 FERRAMENTAS DE MINERAÇÃO DE DADOS

Dentre as ferramentas disponíveis para Mineração de Dados, uma que mais se destaca é a *Waikato Environment for Knowledge Analysis,* (WEKA), que de acordo com Abernethy (2010), trata-se de um software de código aberto e gratuito que possibilita a transformação de dados em conhecimento útil. (Patrício, *et al*, 2018).

Outra ferramenta voltada para fins acadêmicos, é a RisingMiner, que é uma ferramenta web para mineração de regras de associação baseada no algoritmo *Fuzzy Ontology Generalized Association Rules.* (PATRICIO, *et al*., 2018).

A ferramenta permite ao usuário inserir parâmetros a fim de obter regras customizadas por meio de taxonomias. O Quadro 2 apresenta as seis melhores ferramentas de Mineração de Dados de código aberto.

Quadro 2 - Principais ferramentas de mineração de dados

Ferramenta	**URL**
RAPIDMINER	http://rapidminer.com/products/studio
ORANGE	http://orange.biolab.si/
KNIME	http://www.knine.org/
WEKA	http://www.cs.waikato.ac.nz/ml/weka
KELL	http://www.keel.es/
R	http:///www.r-project.org
SCIKITLEARN	https://scikit-learn.org/

Fonte: Rangra e Bansal, 2014 (Modificado por Patricio, *et, al,Autor*)

2.6 MINERAÇÃO DE DADOS NA EDUCAÇÃO BÁSICA

Apesar das aplicações de Mineração de Dados serem implementadas nos mais diversos setores, como saúde, mercado financeiro, entre outros, o foco deste trabalho é a mineração de dados na educação, especificamente no contexto dos resultados SAEB 2019 e seguintes.

Na literatura, um termo recente tem ganhado força e atenção dos pesquisadores, nomeada de mineração de dados educacionais (EDM). A EDM busca desenvolver ou adaptar métodos e algoritmos de mineração existentes, de tal modo, que se prestem a compreender melhor os dados em contextos educacionais, produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem. (COSTA *et al*., 2012).

Há diversas tarefas envolvidas em EDM, notadamente as que decorrem diretamente da análise de dados gerados nas interações dos estudantes com os ambientes de aprendizagem. (COSTA *et al*, 2012). Segundo o autor, desse campo de análise surgem demandas para responder questões relacionadas a como melhorar a aprendizagem do estudante e como desenvolver ambientes educacionais mais eficazes que contribuam efetivamente para os estudantes aprenderem mais e em menos tempo.

CASTRO *et al*, (2021), identificou uma tendência em grande parte dos artigos que tratam de problemas relacionados ao desempenho escolar de alunos, a partir das bases de dados do INEP. Os artigos analisados pelo autor visam identificar os fatores que influenciam no desempenho, criar modelos de predição, bem como sugerir possíveis soluções que podem ser feitas para melhorar as taxas de desempenho.

Outra tendência identificada, foi a realização de análises exploratórias dos dados sem um método bem definido nos artigos, além da utilização da metodologia CRISP-DM. Quando utilizado aprendizado de máquina nos estudos, os modelos supervisionados de regressão foram os mais utilizados. (Castro, *et al*, 2021).

Do ponto de vista computacional, existem alguns desafios práticos em vários contextos educacionais. Dentre os quais estão relacionados, por exemplo, a falta de padronização dos dados que exigem um grande esforço de pré-processamento. (BAKER, 2011). Além disso, existe a necessidade de adequação dos algoritmos clássicos de mineração de dados para lidar com especificidades inerentes aos dados educacionais, tais como a não independência estatística e a hierarquia dos dados. (BAKER, 2010).

2.8 TRABALHOS CORRELATOS

SOARES, *et al.* (2021), reuniu os principais artigos que tratam sobre mineração de dados utilizando as bases de dados do INEP.

PATRICIO *et al.* (2018), apontou que as análises de *Big Data* para a educação, estão intrinsecamente ligadas à personalização do ensino, em especial no foco da individualidade de cada estudante.

FREITAS, *et al.* (2019), demonstrou a partir do estudo de caso em escolas municipais da cidade de Maceió - AL, como a utilização de técnicas de mineração de dados, pode contribuir com melhorias na gestão escolar.

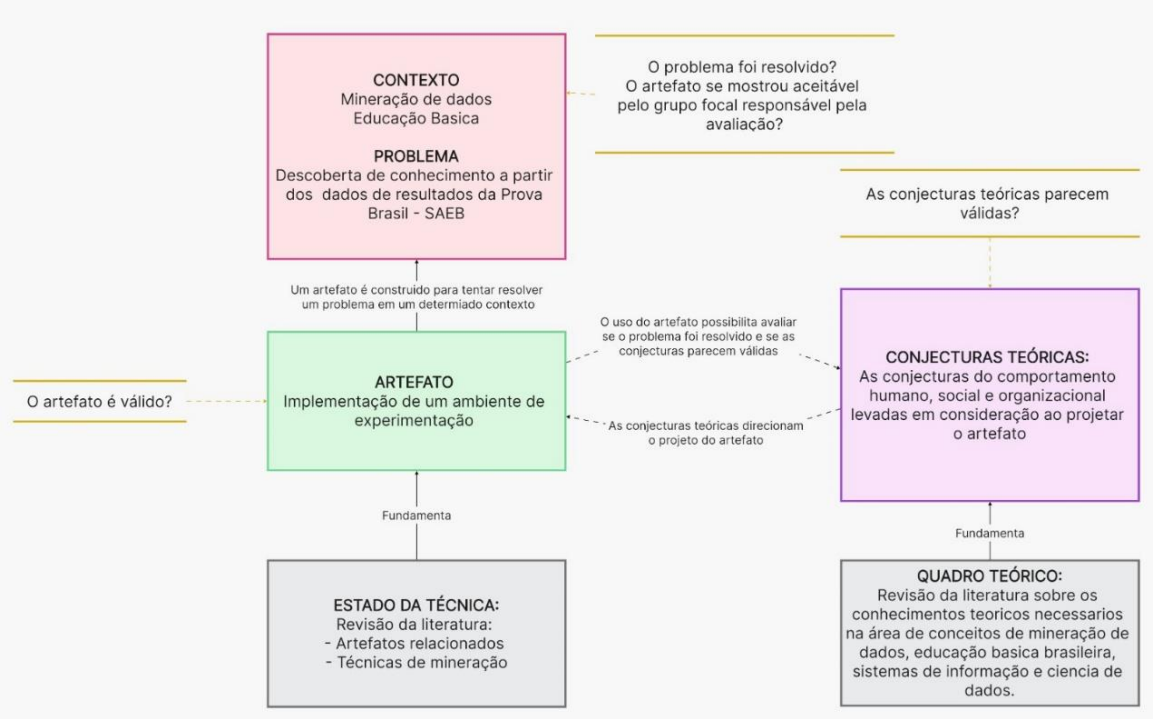
3 METODOLOGIA

A metodologia de pesquisa adotada nesse trabalho, emprega a *Design Science Research* (DSR), uma abordagem que busca criar soluções inovadoras para desafios práticos. Essa metodologia se destaca devido a possibilidade da criação de artefatos com flexibilidade para adaptação, sendo estes artefatos, sistemas, modelos, algoritmos ou métodos projetados para enfrentar desafios do mundo real.

Essa abordagem envolve um ciclo iterativo, em que os pesquisadores identificam um problema, propõem uma solução, constroem o artefato, o avaliam e refinam continuamente para melhorar sua eficácia. (HEVNER *et al*. (2004).

A figura 3 representa graficamente as etapas da metodologia DSR.

Figura 3 - Metodologia DSR



Fonte: O Autor

3.1 DESENVOLVIMENTO

Primeiro, foi realizado uma Revisão Sistemática para levantamento das principais fontes de literatura sobre algoritmos que estão sendo utilizados para mineração de dados nas bases do SAEB.

Com base nessa revisão, foram identificados os principais algoritmos de mineração de dados que foram aplicados a essa base de dados nos últimos anos. Foram analisadas suas características e desempenho, a fim de selecionar o algoritmo mais adequado para aplicação nesses dados.

Considerando a natureza e a organização da base de dados do SAEB e as descobertas experienciadas na revisão de literatura, decidimos pela a utilização de técnicas de agrupamento (clusterização), como estratégia para descoberta de conhecimento nessa base de dados.

Os microdados SAEB não possuem classificações ou rótulos. Desse modo, a partir da clusterização, os dados foram agrupados por similaridade, de maneira que a partir da análise dos clusters foi possível a descoberta de conhecimento no âmbito da educação básica.

Como justificado anteriormente, o *K-MEANS* foi escolhido como algoritmo utilizado para realização do processo de clusterização na base de dados selecionada.

A avaliação de professores e gestores da educação, foi utilizada como métrica para validar os resultados encontrados pelo algoritmo.

Após a definição de métricas, foi realizado a coleta, a limpeza, tratamento e organização dos dados, bem como a resolução de eventuais problemas, como valores ausentes ou inconsistentes, de modo a tornar adequado para ser utilizado no algoritmo selecionado.

Inicialmente foi desenvolvido um ambiente de testes, descrito na seção seguinte. Nesse ambiente foi desenvolvido um programa na linguagem de programação Python, para testar o algoritmo na base de dados tratada.

Para execução do programa desenvolvido, foi definido um cronograma de testes considerando três cenários distintos, onde o pesquisador alterou a quantidade de atributos executados pelo modelo desenvolvido. O detalhamento de cada cenário foi descrito na seção “Testes Realizados”.

Em seguida, foi construído um ambiente de experimentação público com uma interface acessível, que permite aos educadores e gestores a extração de informações úteis a partir dos dados do SAEB, mesmo sem possuírem conhecimentos aprofundados em ferramentas específicas, como o Microsoft Excel ou Weeka.

Nesse ambiente, é possível realizar a visualização e relacionamento de dados a partir do resultado da implementação do algoritmo *K-MEANS*, sendo possível a identificação de possíveis padrões e tendências, que podem ser utilizadas para elaboração de políticas públicas e suporte a decisão.

Os resultados obtidos foram avaliados por gestores e professores da educação básica, analisando-se a eficácia e utilidade das informações extraídas pelo algoritmo *K-MEANS*. Essa avaliação permitiu verificar se os objetivos propostos foram alcançados e se as informações geradas são relevantes para a melhoria da educação básica.

Por fim, espera-se que os resultados e descobertas possam contribuir para a reflexão e discussão sobre a utilização de técnicas de mineração de dados como suporte para a tomada de decisão na área da educação básica.

3.1.1 Configuração do Ambiente de desenvolvimento

O processo de desenvolvimento deste projeto, foi executado no em um ambiente com o sistema operacional Windows em sua versão 10 com a seguinte configuração de hardware :

- Processador Intel Core I5 (4 Núcleos e 4 Threads) com frequência 3.3 Gigahertz (Ghz) de processamento e 6MB de cache;
- Memória de Acesso Randômico (RAM) de 15.9 Gigabytes (GB);
- Memória livre em disco de 318 Gigabytes GB.

Para a implementação do algoritmo *K-MEANS*, foi utilizada a linguagem de programação Python na versão 3.11.4 e bibliotecas específicas para aprendizado de

máquina. Dentre elas, destaca-se a Scikit Learn, por disponibilizar a implementação dos principais algoritmos. O Apêndice B descreve todas as bibliotecas utilizadas nesse projeto, bem como, o objetivo da utilização.

A linguagem de programação Python foi escolhida por sua popularidade, flexibilidade e robustez para o desenvolvimento de soluções voltadas para ciência de dados, dispondo de bibliotecas pertinentes ao contexto do projeto.

O Microsoft Excel, na versão Professional Plus 2016, foi utilizado no processo de limpeza, tratamento, organização e visualização inicial dos resultados dos dados.

Na seção seguinte, são detalhados todos os procedimentos realizados referentes à primeira etapa do processo de KDD, que consiste na seleção inicial dos dados.

3.1.2 Coleta de Dados

Como mencionado anteriormente, o objetivo desse estudo foi analisar as bases de dados do INEP relacionadas à Educação Básica. Esses dados provêm de avaliações em larga escala que fazem parte do Sistema de Avaliação da Educação Básica - SAEB, que tem como finalidade diagnosticar o sistema educacional do Brasil e identificar fatores ligados ao desempenho dos estudantes.

Os resultados desta avaliação estão disponíveis publicamente no site oficial do INEP.

Os arquivos são disponibilizados de duas maneiras: a planilha resultados e os microdados.

A Planilha resultados, é composta por um compilado de todas as informações que integram os resultados das avaliações disponíveis. Nessa planilha, estão inseridos dados que consideram resultados em abrangência nacional, estadual e municipal.

Os microdados, são todas as informações detalhadas relacionados aos resultados das avaliações, separados por arquivos individuais, dedicados a cada ator envolvido na construção dos resultados. são eles: Aluno, Diretor, Escola, Professor e Secretário Municipal. Além dessas informações nesse conjunto de dados, existem os arquivos do dicionário de dados, notas técnicas, matrizes de referência e escalas de proficiência.

Considerando que a finalidade dessa pesquisa é a descoberta de conhecimento, foi realizado o download dos Microdados do SAEB 2019, viabilizando, assim, a análise de conjuntos de dados mais abrangentes e minuciosos.

3.1.3 Seleção dos Dados

Em conformidade com os objetivos desta pesquisa, considerando as diversas bases de dados disponíveis nos arquivos dos Microdados SAEB e considerando a limitação temporal da pesquisa, escolhemos uma das bases disponíveis para realização dos experimentos.

A base selecionada foi o arquivo “TS_ESCOLA.csv”, disponível na pasta DADOS dos Microdados SAEB 2019. A escolha dessa base, se deu por sua relevância no contexto da pesquisa, bem como, o tamanho da base dados, considerando as limitações de hardware do ambiente de desenvolvimento.

O arquivo “TS_ESCOLA.csv”, originalmente possui 70.606 registros, caracterizado por 137 atributos contendo a identificação e as informações de proficiência em português e matemática, em âmbito nacional de todas as escolas avaliadas no SAEB.

Para este estudo, foi selecionado o estado da Bahia como locus para execução da mineração de dados. Nesse sentido, após a consulta ao dicionário de dados presente nos microdados disponibilizados, foram selecionados todos os registros cujo atributo “ID_UF”, tivesse o valor igual a 29. Nesse processo, foram extraídos 6404 registros dando origem a uma nova base de dados, nomeada de “TS_ESCOLA_BA.CSV”.

3.1.4 Pré-processamento e transformação dos dados

Fonseca (2014), destaca que, após a seleção dos dados, é conduzida a etapa de limpeza e pré-processamento. É comum encontrar várias inconsistências nos dados. Nesse sentido, a fase de limpeza de dados desempenha um papel fundamental, uma vez que seu propósito é a eliminação desses problemas, de modo a não afetar os resultados dos algoritmos de mineração que serão aplicados

A etapa de preparação de dados é fundamental no contexto da descoberta de conhecimento. De acordo com Tan, et. al. (2009), a tarefa de preparação de dados é uma das mais dispendiosas e trabalhosas na análise de dados, uma vez que a qualidade dos dados influencia diretamente a qualidade do conhecimento produzido. Portanto, após a conclusão desta fase, é possível avançar no processo de Mineração de Dados no âmbito do KDD.

Nesta etapa de pré processamento, a primeira tarefa executada na base de dados recém criada foi a remoção dos atributos “ID_SAEB”, “ID_REGIÃO” e “ID_UF”, considerados redundantes para realização da análise, no contexto onde a mineração dos dados será realizada apenas com os dados SAEB de 2019 do estado da Bahia.

Por escolha do pesquisador, com objetivo de diminuir a quantidade de dados a serem analisados pelo algoritmo *K-MEANS*, a próxima etapa foi realizada considerando apenas os dados referente ao do 5ª ano. Portanto, os atributos que não correspondem a essa dimensão foram removidos da base de dados.

O Apêndice A, demonstra todos os atributos removidos nesta seleção, bem como, o dicionário referente a cada atributo removido.

Findada a remoção, a base de dados ficou com 33 atributos. Dando continuidade, prosseguimos com o processo de limpeza da base de dados no qual, o foco foi remover os registros que possuíam valores vazios ou nulos em algum dos seus atributos.

O Quadro 3, demonstra os campos da tabela afetados nesta etapa, bem como, a quantidade de registros vazio ou nulos removidas.

Quadro 3 - Campos da tabela afetados na transformação e limpeza dos dados

NOME ATRIBUTO	OPERAÇÃO REALIZADA	EXCLUSÕES
PC_FORMACAO_DOCENTE_INICIAL	Remoção de campos vazios	2213
NIVEL_SOCIO_ECONOMICO	Remoção de campos vazios	160
NU_MATRICULADOS_CENSO_5EF	Remoção de campos vazios	140
NU_PRESENTES_5EF	Remoção de campos valor 0	3
Nivel_0_LP5	Remoção de campos vazios	345
NIVEL_SOCIO_ECONOMICO	Remoção de Outliers	2

Fonte: O Autor

Após a limpeza da base, restaram 3541 registros. A última etapa deste processo consiste na padronização dos dados, com o objetivo da implementação do algoritmo *K-MEANS*.

Considerando que para que este algoritmo funcione, é necessário que a base de dados seja composta com valores numéricos, foi realizado uma alteração nos valores do atributo “NIVEL_SOCIO_ECONOMICO”. Utilizando a função PROC-V do Microsoft Excel, os valores anteriormente em formato texto, foram substituídos por equivalentes numéricos conforme demonstrado no Quadro 4.

Quadro 4 - Equivalência numérica campo NÍVEL_SOCIO_ECONOMICO

VALOR EM TEXTO	EQUIVALENTE NUMÉRICO
Nível I	1
Nível II	2
Nível II	3
Nível IV	4
Nível V	5

Fonte: O Autor

Concluída esta etapa, a base de dados estava pronta para ser utilizada no algoritmo. A próxima etapa descreve o processo de mineração de dados.

3.1.5 Mineração de Dados

Dado a conclusão das etapas anteriores do processo de KDD, iniciou-se a fase da Mineração de Dados. Propriamente, o objetivo da mineração é, a partir da utilização do algoritmo *K-MEANS*, agrupar dados por semelhanças, permitindo a identificação de possíveis padrões que subsidiam decisões para aprimorar a educação.

Para o correto funcionamento do *K-MEANS*, alguns critérios devem ser observados. A.K,*et. al*. (2010), apontam que o desempenho do algoritmo depende da escolha correta do K bem como da normalização dos dados. A normalização dos dados visa garantir que diferentes atributos tenham pesos comparáveis, garantindo assim, análises mais precisas e resultados consistentes. (HAN et. al, 2011).

Considerando esses critérios, utilizamos os recursos das bibliotecas *Python* para implementar métodos que auxiliassem na garantia do cumprimento dos mesmos.

Utilizando os recursos da classe *StandardScale()* , os dados da base de dados “TS_ESCOLAS_BA.csv”, foram padronizados de modo que, todos os registros ficaram na mesma escala. A Figura 4, demonstra a visualização dos dados antes da normalização, a Figura 5, demonstra a visualização dos dados após a normalização.

Figura 4 - Dados Antes da Normalização

	ID_MUNICIPIO	ID_AREA	...	MEDIA_5EF_LP	MEDIA_5EF_MT
0	6312857	2	...	220.63	245.25
1	6312857	2	...	205.08	223.47
2	6312857	2	...	225.83	224.23
3	6312858	2	...	213.72	232.44
4	6312858	2	...	223.54	257.63
...	...	...	...	...	...
3536	6313273	2	...	131.26	161.81
3537	6313273	2	...	160.03	174.56
3538	6313273	2	...	150.89	159.45
3539	6313273	2	...	184.50	188.24
3540	6313273	2	...	155.20	169.74

Fonte: O Autor

Figura 5 - Dados após Normalização

```
Esses sao os dados padronizados
[[-1.80059844  0.28499364 -1.63580484 ... -0.05715594  1.47172691
   2.11002632]
 [-1.80059844  0.28499364 -0.75759499 ... -0.05715594  0.73155166
   1.04664362]
 [-1.80059844  0.28499364  0.88480918 ... -0.05715594  1.71924532
   1.08374972]
 ...
 [ 1.6063212   0.28499364  1.52140358 ... -0.05715594 -1.8478758
  -2.07905705]
 [ 1.6063212   0.28499364  1.58721779 ... -0.05715594 -0.24805006
  -0.6734194 ]
 [ 1.6063212   0.28499364  1.61162292 ... -0.05715594 -1.64272112
```

Fonte: O Autor

Nesta fase inicial da mineração de dados, executamos o *K-MEANS* com um K aleatório apenas para validar o funcionamento do algoritmo na base de dados e verificar possíveis erros oriundos das etapas anteriores.

Durante a execução dos testes não foram encontrados erros de execução, deste modo, avançamos no processo de mineração com a definição do K ideal.

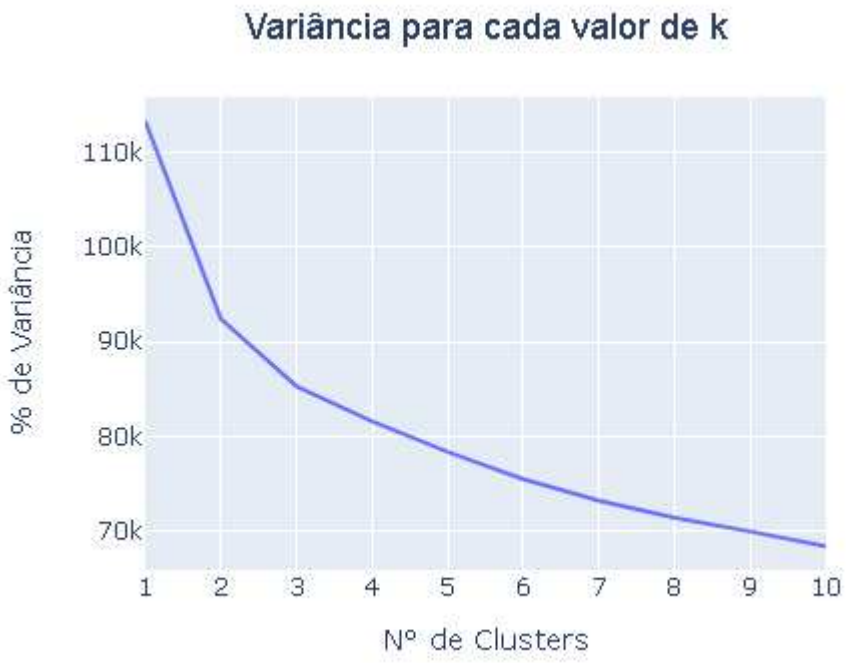
3.1.6 Definição do número ideal de clusters

3.1.6.1 O método do Cotovelo

Para definir o número ideal de *clusters* a ser utilizado para execução do *K-MEANS*, recorremos ao popular “método de cotovelo”. A ideia desse método é identificar o ponto em que a adição de mais *clusters* não melhorou significativamente a qualidade da divisão dos dados em grupos. Isso é feito avaliando a variação intra-cluster em relação ao número de *clusters*. (PANDEY *et al.* 2018).

A Figura 6, demonstra o resultado da aplicação desse método na base de dados “TS_ESCOLA_BA.csv”.

Figura 6 - Curva de Elbow (T5_ESCOLAS_BA.csv)



Fonte: O Autor

Considerando a análise do gráfico com base no método do cotovelo, o número ideal de clusters está entre 3 e 4. Pois, são nesses pontos que a curva do gráfico passa a tomar um formato similar a um “cotovelo”.

A seguir, realizaremos novas validações utilizando o método *“silhouette score”.*

3.1.6.2 Silhouette Score

O *Silhouette Score* é uma métrica de avaliação de qualidade de clusters que mede o quão bem os objetos estão agrupados. Ele varia de -1 a 1, onde valores mais próximos de 1 indicam *clusters* bem definidos, valores próximos de 0 indicam sobreposição entre *clusters* e valores próximos de -1 indicam agrupamentos incorretos.

Ele é calculado usando a distância média entre um objeto e os objetos do mesmo *cluster;* e a distância média entre o objeto e os objetos de outros *clusters* mais próximos.

A Figura 7, demonstra o código fonte da função utilizada para calcular o *silhouette score* dos agrupamentos.

Figura 7 - Chamada da Função de Cálculo do Silhouette Score

```
149  #silhoute score
150  print("esse e o silhuete score: para cluster")
151  labels = modelo_v1.labels_
152  print(silhouette_score(pca, labels, metric='euclidean'))
```

Fonte – O Autor

Para fins de validação, esse método foi executado 4 vezes, no qual o número de *clusters* passado como parâmetro, foi sendo alterado entre 2 e 5. A Figura 8 demonstra os resultados encontrados para cada valor de K.

Figura 8 - Valores do Silhoutte Score para cada valor de K

Fonte: O Autor

Ao examinarmos o gráfico da figura 8, foi possível observar que na medida em que o valor de K aumenta, o valor do score se aproxima do zero. Como dito anteriormente, essa proximidade do zero, indica uma possível similaridade entre os clusters.

3.1.8 Testes Realizados

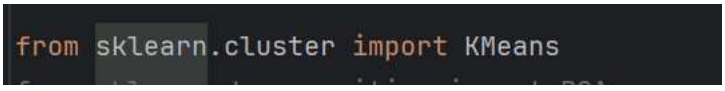
Nesta seção, descrevemos o treinamento realizado para o modelo de clusterização utilizando o algoritmo *K-MEANS*. Em seguida, descrevemos os testes executados que posteriormente foram comparados na fase de análise de resultados.

Para fins de comparação, foram executados testes considerando três cenários distintos. Para cada cenário foi necessário realizarmos modificações na base de dados “T5-ESCOLAS-BA.csv” e nos parâmetros do *K-MEANS,* em função do comportamento do algoritmo na base de dados modificada.

Para execução do treinamento, utilizamos a *Scikit Learn*, uma biblioteca de código aberto utilizada para aprendizado de máquina, implementada para linguagem *Python*. Esta biblioteca disponibiliza a implementação dos principais algoritmos de aprendizado de máquina, simplificando a utilizando destes algoritmos.

A figura 9, demonstra a importação da *Scikit Learn* na linguagem *Python*, considerando a utilização do algoritmo *K-MEANS*.

Figura 9 - Importação do *K-MEANS* da Biblioteca Scikit Learn



Fonte: O Autor.

Para realizarmos o treinamento do primeiro cenário de testes, foram utilizados todos os campos da base de dados “T5-ESCOLAS-BA.csv”. Considerando o resultado apresentado no gráfico de cotovelo. O agrupamento foi realizado utilizando três *clusters*.

A documentação da biblioteca, define o número de *clusters* como único parâmetro obrigatório para utilização do algoritmo *K-MEANS*. Entretanto, considerando o funcionamento do mesmo e de modo a garantir que os resultados dos

treinamentos sejam idênticos a cada execução, adicionamos o parâmetro ***"random_state*" com valor 1**. Segundo a documentação, ao adicionarmos um valor para esse parâmetro, tornamos determinística a geração de números aleatórios para inicialização do centroide.

Definido os parâmetros a serem utilizados, atribuímos o retorno da chamada ao método *K-MEANS* a variável "modelo_v1", em seguida, a partir da chamada do método *fit(),* realizamos o processo de clusterização dos dados do *dataset* armazenados na variável pca.

A Figura 10 demonstra o código fonte da realização desta tarefa.

Figura 10 - Treinamento *K-MEANS* utilizando 3 clusters

```
144  #Criando o modelo com k = 3
145  modelo_v1 = KMeans(n_clusters=3)
146  modelo_v1.fit(pca)
```

Fonte: O Autor

Após o processo de treinamento do modelo clusterizado, revertemos a notação dos dados para facilitar a leitura e análise dos mesmos. Para essa tarefa, utilizamos o método *inverse_transform(),* aplicado aos dados armazenados em pca.

A Figura 11 demonstra a etapa de reversão da escala dos dados.

Figura 11 - Reversão da escala dos dados

```
155  pca = scaler_dataset.inverse_transform(pca)
```

Fonte: O Autor

A etapa final, precedendo a interpretação dos resultados da clusterização, envolveu a criação de uma nova base de dados que possibilitou a atribuição de cada registro ao seu respectivo cluster. Para essa tarefa, primeiro, foi criado o *array "names",* que armazena o nome de todas as colunas da base de dados utilizadas no processo de clusterização, em seguida, utilizando o método *DataFrame()* da biblioteca

Pandas, foi criada a matriz ***cluster_map****,* que armazena os dados de pca e as colunas do *array "names".*

Por fim, o campo *'CLUSTER'* foi adicionado à matriz como campo identificador do cluster a qual cada registro pertence. Esse campo foi preenchido com os valores retornados do método *".**labels_**",* aplicado ao conjunto de dados clusterizados em **modelo_v1**.

Após a identificação, esse novo conjunto de dados foi nomeado como "dump-T5-ESCOLAS-BA", em seguida exportado no formato "csv" a partir da função *to_csv().* A Figura 12 demonstra a etapa de criação e exportação da nova base de dados.

Figura 12 - Criação e Exportação de Base de Dados Clusterizada

```
155  pca = scaler_dataset.inverse_transform(pca)
156  names = ['ID_MUNICIPIO', 'ID_AREA', 'ID_ESCOLA', 'IN_PUBLICA',
157           'ID_LOCALIZACAO', 'PC_FORMACAO_DOCENTE_INICIAL', 'NIVEL_SOCIO_ECONOMICO',
158           'NU_MATRICULADOS_CENSO_5EF', 'NU_PRESENTES_5EF', 'TAXA_PARTICIPACAO_5EF',
159           'Nivel_0_LP5', 'Nivel_1_LP5', 'Nivel_2_LP5', 'Nivel_3_LP5',
160           'Nivel_4_LP5', 'Nivel_5_LP5', 'Nivel_6_LP5', 'Nivel_7_LP5',
161           'Nivel_8_LP5', 'Nivel_9_LP5', 'Nivel_0_MT5', 'Nivel_1_MT5',
162           'Nivel_2_MT5', 'Nivel_3_MT5', 'Nivel_4_MT5', 'Nivel_5_MT5',
163           'Nivel_6_MT5', 'Nivel_7_MT5', 'Nivel_8_MT5', 'Nivel_9_MT5',
164           'Nivel_10_MT5', 'MEDIA_5EF_LP', 'MEDIA_5EF_MT']
165  #Inclui o número de clusters no dataset
166  cluster_map = pd.DataFrame(pca, columns=names)
167  cluster_map['CLUSTER'] = modelo_v1.labels_
168  cluster_map.to_csv('dump-T5-ESCOLA-BA.csv')
```

Fonte: O Autor

O segundo cenário testado, considerou uma alteração no número de campos utilizadas da base dados "T5-ESCOLAS-BA.csv".

Nesse cenário, o agrupamento foi realizado considerando os campos "PC_FORMACAO_DOCENTE_INICIAL","NIVEL_SOCIO_ECONOMICO","TAXA_PARTICIPACAO_5EF", "MEDIA_5EF_LP", "MEDIA_5EF_MT", a partir desta seleção, uma nova base de dados foi criada, nomeada de "T5-ESCOLAS-BA-v2.csv".

O número de *clusters* do mesmo, foi definido a partir da utilização do método de cotovelo e o cálculo do *silhoutte score*. Após a aplicação desses métodos na referida de dados. O agrupamento foi realizado utilizando 3 *clusters.*

As figuras 13 e 14, demonstram o resultado do “método de cotovelo” e do cálculo *silhoutte score* para essa base de dados.

Figura 13 - Curva de Elbow (T5-ESCOLAS-BA-v2.csv)

Fonte: O Autor

Figura 14 - Cálculo do *Silhoute Score* (T5-ESCOLAS-BA-v2.csv)

```
esse e o silhuete score: para 3 cluster
0.19516827913546006
```

Fonte: O autor

Como dito anteriormente; e novamente observado nessa base de dados demonstrado na Figura 14, O resultado do cálculo do *silhoutte score* se aproximou do zero, indicando uma possível similaridade entre os grupos. Os dados deste agrupamento foram exportados para o arquivo “dump2-T5-ESCOLAS-BA.csv”.

O terceiro cenário de teste, considerou apenas os campos “NIVEL_SOCIO_ECONOMICO”, “MEDIA_5EF_LP”, “MEDIA_5EF_MT”, a partir dessa seleção, uma nova base de dados foi criada, nomeada de “T5-ESCOLAS-BA-v3.csv”;

O número de *clusters* desse cenário foi definido a partir da utilização do “método de cotovelo” e o cálculo do *silhoutte score*.

Após a aplicação, o agrupamento foi realizado utilizando 4 *clusters*. As figuras 15 e 16, demonstram o resultado do “método de cotovelo” e do cálculo *silhoutte score* para essa base de dados.

Figura 15 - Curva de Elbow (T5-ESCOLAS-BA-v3.csv)

.Fonte: O Autor

Figura 16 - Cálculo do Silhoutte Score (T5-ESCOLAS-BA-v3.csv)

```
esse e o silhuete score: para 4 cluster
0.40175526938849454
```

Fonte: O Autor

Os dados desse agrupamento foram exportados para o arquivo “dump3-T5-ESCOLAS-BA.csv”.

4 Análise de Resultados

Esta seção, descreve a análise e interpretação dos dados produzidos em cada um dos três cenários descritos na seção anterior. O objetivo desta análise, foi identificar nos agrupamentos, possíveis padrões que pudessem contribuir na fundamentação para formulação de políticas públicas que objetivam melhorais no sistema educacional.

4.1 Primeiro Cenário – Clusterização considerando todos os atributos da base dados

O primeiro cenário testado, adotou a clusterização dos dados em um cenário de 33 atributos, considerando a complexidade da visualização e interpretação deste volume de atributos, o pesquisador adotou uma estratégia de visualização por etapas. Entretanto, é importante destacar, que para o agrupamento o algoritmo considerou todas os atributos, ou seja, o resultado dos agrupamentos não depende do cenário de visualização.

A Figura 17, demonstra a primeira observação, que considerou as médias de português, matemática e o número de escolas de cada *cluster.*

Figura 17 - Médias LP e MT por Cluster

Cluster	MÉDIA MT	MÉDIA LP	ESCOLAS
0	202.56	190.89	1661
1	228.21	216.04	853
2	179.44	165.94	1027
Total Geral	**202.03**	**189.71**	**3541**

Fonte: O Autor

De maneira preliminar, foi observado que o primeiro *cluster* possui médias similares a média global das escolas analisadas, o segundo *cluster* possui média acima da média global e o terceiro possui média abaixo da média global em cada uma das disciplinas.

A Figura 18, demonstra um aprofundamento da visualização desse cenário, no qual foram adicionados os atributos de formação docente e o nível socioeconômico

de cada escola. Onde o 2º representa o menor nível socioeconômico observado e o 5° representa o maior nível socioeconômico observado nas escolas analisadas.

O atributo formação docente, segue a mesma lógica do nível socioeconômico, onde quanto maior o seu valor numérico, maior é o nível de formação dos professores da escola analisada.

Figura 18 - Visualização com 5 atributos

Cluster	MÉDIA MT	MÉDIA LP	ESCOLAS	FORMAÇÃO DOCENTE
0				
2	204.13	191.41	88	56.82
3	202.36	189.84	935	62.73
4	202.61	192.35	629	69.09
5	204.48	192.96	9	61.83
1				
2	229.09	210.41	28	62.69
3	227.21	214.22	293	67.92
4	228.01	216.59	484	73.81
5	235.93	224.89	48	68.35
2				
2	177.22	162.62	277	42.87
3	180.17	167.08	640	52.46
4	180.87	167.84	107	49.86
5	176.74	161.00	3	41.87
Total Geral	**202.03**	**189.71**	**3541**	**61.92**

Fonte: O Autor

Nessa visualização, o primeiro *cluster* é formado majoritariamente por escolas com resultados próximos a média geral em todos os atributos, entretanto destaca-se a presença de *outliers* pertencentes ao nível 5 da classificação socioeconômica. Outro destaque desse *cluster*, encontra-se no grupo socioeconômico de valor 2, que apresenta resultados acima da média geral e o menor resultado do nível de formação docente do *cluster* observado. Ou seja, esse grupo apresenta resultados acima da média, apesar do baixo nível socioeconômico e menor resultado de formação docente, indicando uma possível contradição ao senso comum.

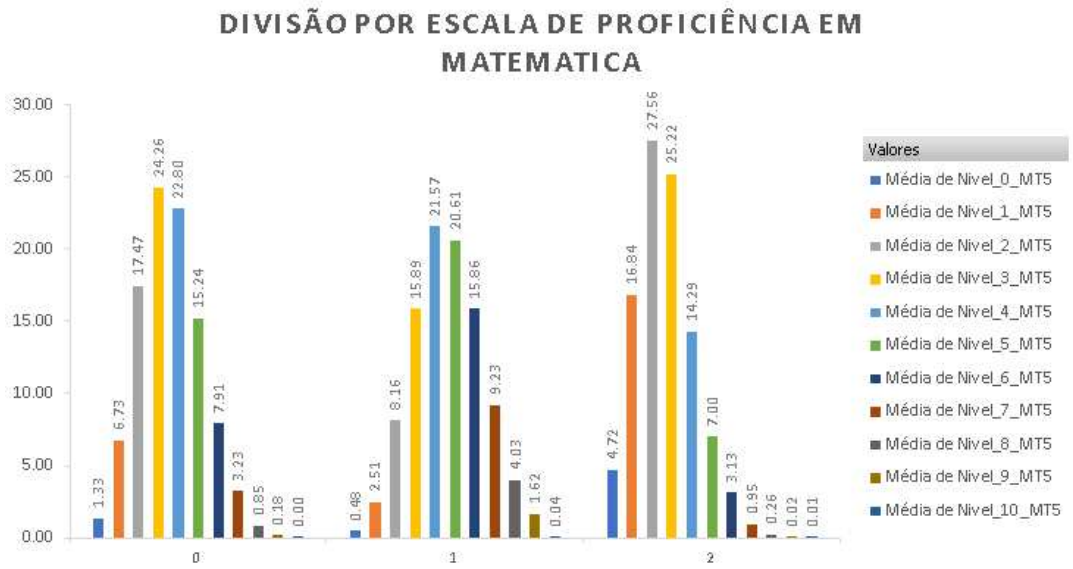
O segundo *cluster*, tem como grupo majoritário, escolas com proficiência acima da média global, maior resultado na formação de professores e nível 4 na classificação socioeconômica. Internamente ocupa a terceira posição nos resultados de matemática e a segunda posição nos resultados de português. Neste *cluster*, o grupo de escolas com as maiores proficiência em ambas as disciplinas pertencem ao grupo socioeconômico 5. Ou seja, apesar desse grupo possui o maior número de escolas

em uma classificação socioeconômica e formação docente alta, a proficiência apresentada nesse grupo não reflete o contexto do mesmo.

O terceiro *cluster,* apresenta como destaque *outliers* que pertencem ao grupo socioeconômico 5, com as menores proficiência nas disciplinas e o menor resultado de formação de professores. No conjunto majoritário, o maior resultado médio de formação de professores do *cluster* e as segundas maiores proficiências intra-cluster. Ou seja, os *outliers* desse grupo, também contrariam o senso comum, pois apresentam um nível socioeconômico alto e resultados que não refletem esse contexto.

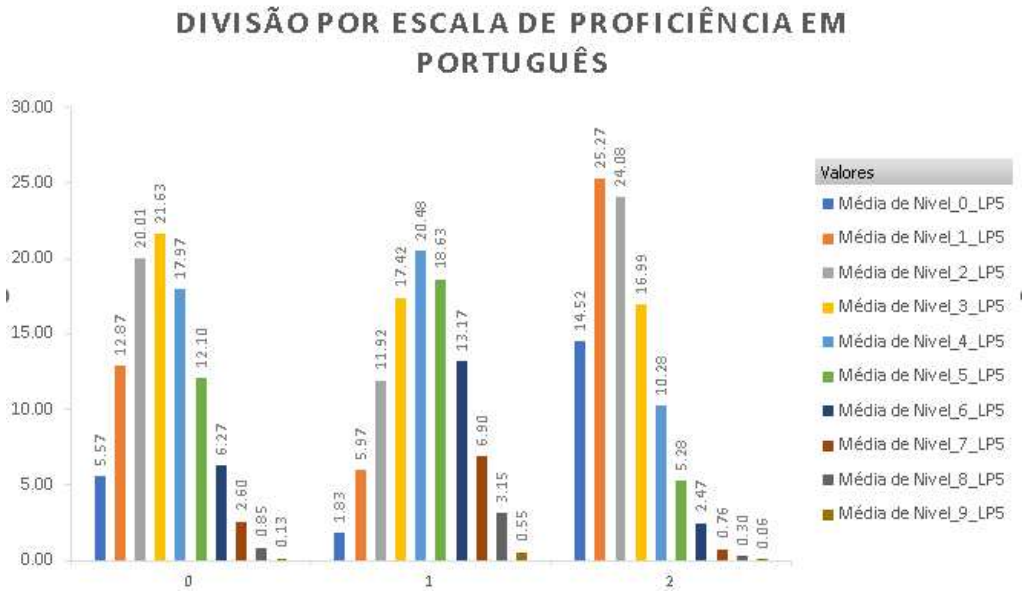
Os Gráficos 1 e 2, apresentam a visualização da divisão de escolas por escala de proficiência em matemática e português presentes em cada *cluster*, nesta visualização, o resultado representa a média da porcentagem de escolas em cada nível.

Gráfico 1 - Divisão de Escolas por Proficiência em MT



Fonte: O Autor

Gráfico 2 - Divisão de Escolas por Proficiência em LP



Fonte: O Autor

No Gráfico 1, o primeiro *cluster* concentra a maior porcentagem de escolas classificadas como nível 4 e 5 na escala de proficiência. O segundo *cluster* está concentrado em escolas majoritariamente dos níveis 5 e 6. O terceiro *cluster,* é formado majoritariamente por escolas de nível 3 e 4, e concentra a maior quantidade de escolas do nível 1 e 2. Ou seja, o primeiro grupo concentra escolas majoritariamente com proficiências próximas a média. O grupo dois, apresenta escolas um pouco acima da média, no terceiro grupo, existe a maior concentração de escolas com proficiência abaixo da média.

No Gráfico 2, o primeiro *cluster* concentra a maior porcentagem de escolas classificadas como nível 3 e 4 na escala de proficiência. O segundo *cluster* está concentrado em escolas majoritariamente dos níveis 5 e 6; e a menor porcentagem de escolas do nível 1. O terceiro *cluster,* é formado majoritariamente por escolas de nível 2 e 3, e concentra a maior quantidade de escolas do nível 1. Ou seja, o primeiro grupo é majoritário em escolas próximas a média. No segundo grupo, a maior parte das escolas estão concentradas em grupos de proficiência próximas ou acima da média. No terceiro grupo, existe a maior concentração de escolas com médias de proficiência muito baixas.

4.2 Segundo Cenário – Clusterização considerando cinco atributos da base dados

O segundo cenário testado, adotou a clusterização dos dados em um cenário de 5 atributos, considerando que complexidade de visualização e interpretação foi reduzida nesse cenário. O pesquisador adotou uma estratégia de visualização em tabela única. A figura 19, demonstra os resultados da tabela com os dados clusterizações neste cenário.

Figura 19 - Resultados clusterização no segundo cenário

Rótulos de Linha	MEDIA MT	MEDIA LP	FORMACAO_DOCENTE	TX_PARTICIPACAO	ESCOLAS
⊟0					
2	230.76	211.50	69.11	96.82%	24
3	222.48	208.77	67.22	96.91%	438
4	219.87	208.94	71.54	95.25%	816
5	231.45	220.55	66.96	96.04%	56
⊟1					
2	193.45	179.64	86.70	90.84%	80
3	194.36	182.48	77.58	91.07%	840
4	193.98	183.28	74.14	90.01%	330
5	183.38	165.91	53.28	89.87%	4
⊟2					
2	181.50	167.24	34.73	95.45%	289
3	187.07	173.68	29.70	95.08%	590
4	185.43	172.96	22.60	95.95%	74
Total Geral	**202.03**	**189.71**	**61.92**	**93.90%**	**3541**

Fonte: O Autor

Nesta visualização, o primeiro *cluster* é composto por escolas acima das médias gerais de proficiência, formação docente e taxa de participação na avaliação superior a 95%. Sendo majoritariamente formado por escolas pertencentes ao nível 4 da classificação socioeconômica, o que indica alunos com maior poder aquisitivo.

O segundo *cluster,* é composto escolas abaixo das médias de proficiências gerais e os maiores resultados de formação docente, com exceção dos *outliers* que apresentam uma média de formação docente abaixo da média geral. A taxa de participação das escolas desse *cluster* foi entre 89.87% e 91.07% e majoritariamente formado por escolas pertencentes ao nível 3 da classificação socioeconômica. Ou seja, nesse grupo, apesar de bons resultados em formação de professores, o grupo concentra escolas com baixo rendimento, contrariando o senso comum.

O terceiro *cluster,* é composto por escolas abaixo das médias gerais de proficiências e os menores resultados de formação docente de todo o agrupamento. A taxa de participação deste *cluster*, foi entre 95.08% e 95.45% e majoritariamente formado por escolas pertencentes ao nível 3 da classificação socioeconômica. Nesse cluster é possível observar que a baixa formação docente, pode refletir no resultado final de proficiência, reafirmando o senso comum.

4.3 Terceiro Cenário – Clusterização considerando três atributos da base dados

O terceiro cenário testado, adotou a clusterização dos dados em um cenário de 3 atributos, considerando a complexidade de visualização e interpretação ter sido reduzida neste cenário, o pesquisador adotou a mesma estratégia de visualização em tabela única utilizada no cenário anterior. A figura 20, demonstra os resultados da tabela com os dados agrupados deste cenário.

Figura 20 - Resultados clusterização no terceiro cenário

Rótulos de Linha	Média MT	Média LP	ESCOLAS
⊟0			
2	213.34	199.16	95
3	208.46	195.61	967
⊟1			
2	178.21	163.78	297
3	182.28	169.41	806
4	166.02	152.98	20
⊟2			
2	272.73	219.77	1
3	237.76	226.33	95
4	228.29	216.84	481
5	237.71	226.67	44
⊟3			
4	200.31	189.73	719
5	202.23	190.08	16
Total Geral	**202.03**	**189.71**	**3541**

Fonte: O Autor

Nesse cenário, os dados foram agrupados em 4 *clusters*. O primeiro *cluster* é composto por escolas dos níveis 2 e 3 da classificação socioeconômica, sendo majoritariamente escolas nível 3.

Os resultados apresentados nesse grupo, ficaram acima das médias gerais de proficiência, mesmo em um cenário de baixo poder aquisitivo.

O segundo *cluster* é composto por escolas abaixo das médias gerais de proficiência, sendo majoritariamente formado por escolas do nível 3 na classificação socioeconômica. Ainda que de maneira minoritária, nesse grupo, existem escolas que pertencem a classificação 4 do nível socioeconômico, ou seja, um maior poder aquisitivo e os menores resultados de proficiência de todo agrupamento.

O terceiro *cluster,* é formado por escolas que possuem os maiores resultados nas médias de proficiência de todo o agrupamento. Nesse grupo, existem escolas de todos os níveis de classificação socioeconômica presentes na base de dados, sendo majoritário no nível 4 de classificação; e com uma única escola assumindo o comportamento de *outlier.* Essa escola apresenta a maior proficiência em matemática de todo o agrupamento e o menor nível na tabela de classificação socioeconômica. Situações dessa natureza, podem indicar potencias campos de estudo e observação, para compreensão e entendimento dos motivadores de bons resultados, apesar do cenário socioeconômico não favorável.

O quarto *cluster,* é composto por escolas um pouco acima da média geral de proficiência em português, e um pouco abaixo da média geral de proficiência em matemática. Essas escolas estão divididas entre os níveis 4 e 5 de classificação socioeconômica, sendo majoritariamente composto por escolas de nível 4. Nesse agrupamento, apesar de um maior poder aquisitivo, as escolas apresentaram resultados próximos a média, o que pode indicar que o fator socioeconômico não é o maior determinante para os resultados acima da média.

4.4 Ambiente de Experimentação

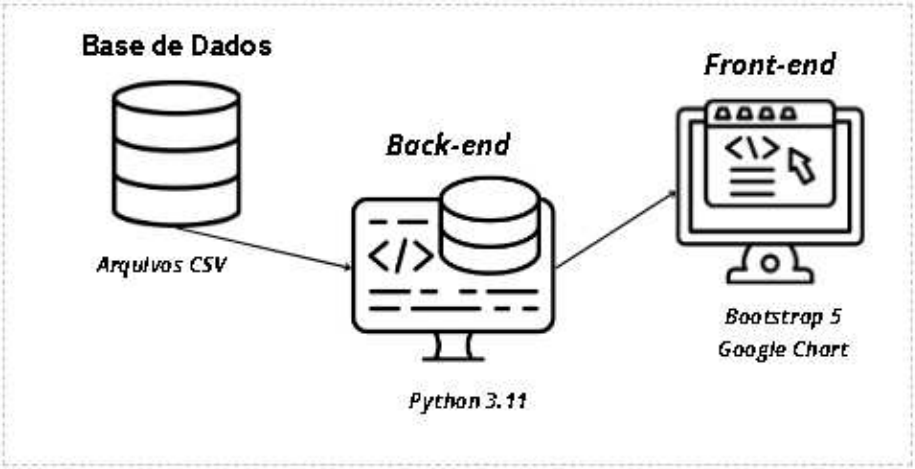
Esta seção, descreve o desenvolvimento de um ambiente de experimentação, acessível a educadores e gestores, que possibilita a extração de informações úteis para aprimorar a educação básica com base nos dados do SAEB.

Dado o contexto, as limitações de tempo e recursos presentes nessa pesquisa, o ambiente de experimentação foi desenvolvido com o objetivo de apresentar a ideia

geral da solução proposta, a análise dos resultados alcançados e as possiblidades mapeadas como trabalhos futuros.

A versão inicial desta plataforma, inicialmente nomeada como “Plataforma Quadro”, foi desenvolvida de acordo a arquitetura apresentada na Figura 21:

Figura 21 - Arquitetura Plataforma Quadro



Fonte: O Autor

Após o seu desenvolvimento, o ambiente de experimentação foi hospedado em um servidor web e disponibilizada através de um domínio especifico.[1]

A versão publicada foi encaminhada para um grupo de professores (as) e gestores (as), que realizaram a avaliação dos resultados encontrados.

4.5 Avaliação da Plataforma

Esta seção, descreve a realização do processo de avaliação do ambiente desenvolvido. A avaliação do artefato produzido, é uma etapa crucial do desenvolvimento considerando a metodologia DSR.

Para executar o processo avaliativo, o pesquisador selecionou um grupo de 12 pessoas, composto por Professores (as), Diretores (as) de Escola e Gestores (as) da área educacional. A avaliação ocorreu a partir do preenchimento de um formulário online disponibilizado pelo pesquisador.

[1] Plataforma Quadro. Disponível em: https://plataformaquadro.tech. Acesso em: 01 Dez, 2023.

O Google Forms foi utilizado como ferramenta para construção e disponibilização do formulário online. Essa escolha seu deu por se tratar de uma plataforma gratuita e que nativamente apresenta os resultados do questionário através de gráficos.

O apêndice C, demonstra todas as perguntas utilizadas no questionário de avaliação. Na primeira coluna é apresentada a pergunta realizada, a segunda coluna representa o tipo de resposta esperada.

O questionário foi divido em perguntas de respostas curtas, caixas de seleção permitindo mais de uma resposta, questões de múltipla escolha e uma questão aberta para críticas, sugestões ou elogios.

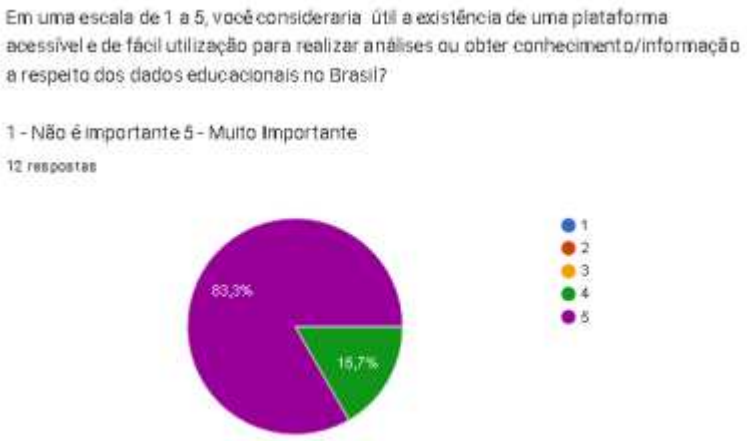
Os Gráficos 3 e 4, demonstram os resultados encontrados a partir do questionário avaliativo. Considerando as respostas observadas, conclui-se que os resultados apresentados no ambiente de experimentação são relevantes e que o desenvolvimento de uma plataforma com esse objetivo será e útil ao contexto de suporte a tomada de decisão para melhorais no sistema educacional.

Os *feedbacks* colhidos a partir do grupo avaliativo, foram analisados para posteriormente serem incorporados como funcionalidades futuras da plataforma.

Gráfico 3 - Relevância das Informações da Plataforma

Fonte: O Autor

Gráfico 4 - Utilidade da Plataforma



Fonte: O Autor

5 Considerações Finais

Ao longo deste projeto, foi possível demonstrar a utilização de estratégias de mineração de dados aplicadas aos dados SAEB. No contexto escolhido pelo pesquisador, foi adotada a estratégia de clusterização utilizando o algoritmo *K-MEANS* no universo de dados da Tabela de Resultados por Escola, considerando as escolas de 5º ano do Estado da Bahia.

Ao examinar os resultados, foi possível identificar alguns padrões conhecidos como a influência de fatores socioeconômicos na proficiência das escolas, bem como, a produção de *insights* não óbvios como a relação não proporcional entre proficiência nas disciplinas avaliadas pelo SAEB, com uma boa qualificação docente encontrada em alguns agrupamentos.

Considerando os resultados observados a partir da avaliação do grupo de professores e gestores, a existência de um ambiente acessível para visualização e análise de dados educacionais, tal qual o ambiente desenvolvido propõe, será útil como suporte a tomada de decisão, visando melhorias contínuas no sistema educacional.

É importante reforçar que o objetivo desta pesquisa, foi a descoberta de conhecimento, a partir da validação da utilização da estratégia de clusterização na base de dados SAEB, como meio de fornecimento de *insights* para fundamentar a tomada de decisão para melhoria do sistema educacional, não sendo objeto desta pesquisa orientar ou emitir conclusões a respeito das ações necessárias para melhorias do sistema educacional.

Para trabalhos futuros, foi identificado a possibilidade de ampliação das bases de dados utilizadas de modo a considerar outros estados, outras turmas e demais tabelas disponíveis nos microdados SAEB. Além disso, considera-se realizar a associação dos resultados da clusterização com outras estratégias no âmbito da mineração de dados, como a classificação, regressão e etc.

Por fim, a possibilidade de realizar a análise dos resultados produzidos de forma automatizada, a partir de integrações com ferramentas de inteligência artificial generativa, como o *Chat GPT,* de modo a fornecer análises, fundamentos e previsões mais robustas para o desenvolvimento e avanço da educação no Brasil.

6 REFERÊNCIAS

FGV, Educação Impulsiona Mobilidade Social. (2019) Disponível em: https://portal.fgv.br/noticias/estudo-revela-educacao-impulsiona-mobilidade-social-brasil. Acesso em: 2 Mai, 2023

BRASIL, Ministro dos Direitos Humanos e Cidadania, (2023), Discurso no segmento de Alto Nível da 52ª Sessão do Conselho de Direitos Humanos da ONU, Genebra, XXVII , fev, 2023. Disponível em: https://www.youtube.com/watch?v=jKQWwqvmpx8 .Acesso em: 24 Abr,2023.

RIBEIRO, Darcy. O povo Brasileiro: A formação e o sentido do Brasil. São Paulo: 1ª Edição, Companhia das Letras,1995.

CASTELLS, Manuel. A sociedade em rede. São Paulo: Paz e Terra, 1999

Silva Babosa, A. A., Andrade, F. S., & de Carvalho, R. N. (2017). MINERAÇÃO DE DADOS EM AMBIENTES VIRTUAIS DE APRENDIZAGEM: APORTES PARA A PESQUISA EM EDUCAÇÃO A DISTÂNCIA. *Interfaces Científicas - Educação*, *6*(1), 125–136

DE CASTRO SOARES, R.; WEBER NETO, N.; REIS COUTINHO, L.; DA SILVA E SILVA, F. J.; VIANA DOS SANTOS, D.; SOARES TELES, A. Mineracão de dados da educação básica brasileira usando as bases do INEP: Uma revisão sistemática da literatura. Revista Novas Tecnologias na Educação, Porto Alegre, v. 19, n. 1, p. 361–370, 2021. DOI: 10.22456/1679-1916.118526. Disponível em: https://seer.ufrgs.br/index.php/renote/article/view/118526. Acesso em: 1 jun. 2023.

HEVNER, A. R., MARCH, S. T., Park, J., & RAM, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75-105.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. AI Magazine, 17(3), 37-54.

Coelho, M. I. de M.. (2008). Vinte anos de avaliação da educação básica no Brasil: aprendizagens e desafios. Ensaio: Avaliação E Políticas Públicas Em Educação, 16(59), 229–258

BRASIL, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).(s.d) Página sobre. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/institucional/sobre. Acesso em: 30 Jun, 2023

BRASIL, Indice de Desenvolvimento da Educação Básica (IDEB).(s.d) Página inicial. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb. Acesso em: 30 Jun, 2023

BRASIL, Sistema de Avaliação da Educação Básica (SAEB). Página inicial. Disponível em:https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb. Acesso em 30 Jun, 2023

BRASIL, Ministério da Educação (MEC). Portal do Mec, artigo Prova. Disponível em: http://portal.mec.gov.br/prova-brasil. Acesso em 30 Jun, 2023

BRASIL. Lei nº 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional (LDBEN). Diário Oficial da União, Brasília, DF, 23 dez. 1996. Disponível em:https://www.planalto.gov.br/ccivil_03/leis/l9394.htm Acesso em 30 Jun, 2023.

BASSO, Flávia Viana. Uso dos resultados do SAEB/Prova Brasil na formulação de políticas educacionais estaduais. 2017. xiii, 138 f., il. Dissertação (Mestrado Profissional em Administração)— Universidade de Brasília, Brasília, 2017.

GOMES, Manoel Messias. Saeb: definição, características e perspectivas. Revista Educação Pública, v. 19, nº 6, 26 de março de 2019. Disponível em: https://educacaopublica.cecierj.edu.br/artigos/19/6/saeb-definicao-caracteristicas-e-perspectivas. Acesso em 30 Jun, 2023.

MICHIE, D.; SLOANE, A.; PEDERSEN, E. D. (1994). Machine learning for user modeling. In: User Modeling: Proceedings of the Ninth International Conference, UM94. John Wiley & Sons, Inc. p. 7-19.

ZAKI, M. J. (2000). Scalable algorithms for association mining. IEEE Transactions on Knowledge and Data Engineering, 12(3), 372-390.

GOLDSCHMIDT, R. Mineração de Dados. 2005. 250 f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo.

Patricio, Thiago Seti, and Maria Da Graça Mello Magnoni. "Mineração De Dados E Big Data Na Educação." Revista GEMInIS 9.1 (2018): 57-75. Web.

KELLAGHAN, T.; GREANEY, V.; MURRAY, T. S. O uso dos resultados da avaliação do desempenho educacional. Brasília, DF: World Bank, 2011. (Pesquisas do banco mundial sobre avaliações de desempenho educacional, v. 5).

ROSISTOLATO, R.; PRADO, A. P.; MARTINS, L. R. A “realidade” de cada escola e a recepção de políticas educacionais. Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 26, n. 98, p. 112-132, jan./mar. 2018. https://doi.org/10.1590/S0104-40362018002601074. Acesso em: 30 Jun, 2023

BASSO, F. V.; FERREIRA, R. R.; OLIVEIRA, A. S. de. Uso das avaliações de larga escala na formulação de políticas públicas educacionais. Ensaio: Avaliação e Políticas Públicas em Educação, v. 30, n. 115, p. 501-519, 2022. Disponível em: https://doi.org/10.1590/S0104-40362021002902436. Acesso em: 30 Jun, 2023

CAVIQUE, L. A. G. Mineração de dados: técnicas, ferramentas e aplicações. São Paulo: Érica, 2014.

FURLAN, F. A. Mineração de dados: conceitos e técnicas. São Paulo: Novatec, 2018.

ZAKY, M. J.; MEIRA JUNIOR, W. (2014). Mineração de dados: conceitos, técnicas, algoritmos, orientações e aplicações. São Paulo: Elsevier.

FRACALANZA, Livia Fonseca. Mineração de dados voltada para recomendação no âmbito de marketing de relacionamento (Mestrado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2009.

COSTA, Evandro et al. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. 2012.

BAKER, R. Data Mining for Education. In: MCGAW, B.; PETERSON, P.; BAKER, E. (Eds.). International Encyclopedia of Education. 3rd ed. Elsevier, Oxford, UK, 2010.

BAKER, R. S. J. d., I. S. d. C. A. Mineração de dados educacionais: Jornada de Atualização em Informática na Educação - JAIE 2012. Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, v. 19, n. 2.

APÊNDICE A - Dicionário de dados com Colunas Removidas

NOME DO CAMPO	DICIONÁRIO
PC_FORMACAO_DOCENTE_FINAL	Indicador de Adequação da Formação Docente (Informação referente ao Grupo 1, para os Anos Finais do EF)
PC_FORMACAO_DOCENTE_MEDIO	Indicador de Adequação da Formação Docente (Informação referente ao Grupo 1, para o Ensino Médio)
NU_MATRICULADOS_CENSO_9EF	Número de alunos matriculados no 9º ano no censo 2019
NU_PRESENTES_9EF	Número de alunos presentes na aplicação
TAXA_PARTICIPACAO_9EF	Razão entre o total de alunos presentes no SAEB (NU_PRESENTES_9EF) e o total de alunos cadastrados no Censo Escolar que são público alvo do SAEB (NU_MATRICULADOS_CENSO_9EF
Nivel_0_LP9 ao Nivel_8_LP9	Proficiência por nível em português do 9º ano
Nivel_0_MT9 ao Nivel_9_MT9	Proficiência por nível em matemática do 9º ano
NU_MATRICULADOS_CENSO_EMT	Número de alunos matriculados na 3ª/4ª série do ensino médio tradicional no censo 2019
NU_PRESENTES_EMT	Número de alunos presentes na aplicação
TAXA_PARTICIPACAO_EMT	Razão entre o total de alunos presentes no SAEB (NU_PRESENTES_EMT) e o total de alunos cadastrados no Censo Escolar

	que são público alvo do SAEB (NU_MATRICULADOS_CENSO_EMT)
Nivel_0_LPEMT ao Nivel_8_LPEMT	Proficiência por nível em português do ensino médio tradicional
Nivel_0_MTEMT ao Nivel_10_MTEMT	Proficiência por nível em matemática do ensino médio tradicional
NU_MATRICULADOS_CENSO_EMI	Número de alunos matriculados na 3ª/4ª série do ensino médio integrado no censo 2019
NU_PRESENTES_EMI	Número de alunos presentes na aplicação do ensino médio integrado
TAXA_PARTICIPACAO_EMI	Razão entre o total de alunos presentes no SAEB (NU_PRESENTES_EMI) e o total de alunos cadastrados no Censo Escolar que são público alvo do SAEB (NU_MATRICULADOS_CENSO_EMI)
Nivel_0_LPEMI ao Nivel_8_LPEMI	Proficiência por nível em português do ensino médio integrado
Nivel_0_MTEMI ao Nivel_10_MTEMI	Proficiência por nível em matemática do ensino médio integrado
NU_MATRICULADOS_CENSO_EM	Número de alunos matriculados na 3ª/4ª série do ensino médio tradicional ou integrado no censo 2019
NU_PRESENTES_EM	Número de alunos presentes na aplicação do ensino médio tradicional ou integrado
TAXA_PARTICIPACAO_EM	Razão entre o total de alunos presentes no SAEB (NU_PRESENTES_EM) e o total de alunos cadastrados no Censo Escolar

	que são público alvo do SAEB (NU_MATRICULADOS_CENSO_EM)
Nivel_0_LPEM ao Nivel_8_LPEM	Proficiência por nível em português do ensino médio tradicional ou integrado
Nivel_0_MTEM ao Nivel_10_MTEM	Proficiência por nível em matemática do ensino médio tradicional ou integrado
MEDIA_9EF_LP	Média em Português 9º ano
MEDIA_9EF_MT	Média em Matemática 9º ano
MEDIA_EMT_LP	Média em Língua Portuguesa 3ª/4ª série do ensino médio tradicional
MEDIA_EMT_MT	Média em Matemática 3ª/4ª série do ensino médio tradicional
MEDIA_EMI_LP	Média em Língua Portuguesa 3ª/4ª série do ensino médio integrado
MEDIA_EMI_MT	Média em Matemática 3ª/4ª série do ensino médio integrado
MEDIA_EM_LP	Média em Língua Portuguesa 3ª/4ª série do ensino médio tradicional ou integrado
MEDIA_EM_MT	Média em Matemática 3ª/4ª série do ensino médio tradicional ou integrado

Apêndice B – Bibliotecas Utilizada para Execução do Projeto

BIBLIOTECA	FUNÇÃO
PANDAS	Fornece uma estrutura robusta para trabalhar com ciência de dados
NUMPY	Operações matemáticas rápidas e manipulações de arrays multidimensionais
MATPLOTLIB.PYPLOT	Interface que permite a criação de gráficos de forma mais conveniente e similar ao estilo MATLAB, o que facilita o processo de plotagem
SKLEARN. PREPROCESSING.STANDARDSCALE	Padroniza os dados, removendo a média e dimensionando para a variação unitária.
SKLEARN. CLUSTER.KMEANS	Disponibiliza a implementação do algoritmo *K-MEANS*
SKLEARN.METRICS. SILHOUETTE_SCORE	Fornece um método para validação e interpretação dos dados de um cluster
SciPy. SPATIAL.DISTANCE	Cálculos de distância espacial
WARNINGS	Controle de Erros

APÊNDICE C – Quadro de Perguntas do Formulário de Avaliação da Plataforma

PERGUNTA REALIZADA	RESPOSTA ESPERADA
1 - Qual seu nome?	Texto Curto
2 - Qual o seu e-mail?	Texto Curto
3 - Você já ocupou ou ocupa algum dos cargos/funções listadas abaixo?	Caixa de Seleção: ☐Professor (a) ☐Coordenador (a) ☐Diretor Escolar (a) ☐Gestor(a) na área da Educação ☐Nenhuma das opções
4 – Você conseguiu acessar a Plataforma Quadro?	Multipla Escola 1 - Sim 2 - Não
5 - Você encontrou alguma dificuldade de acesso a plataforma?	Multipla Escola 1 - Sim 2 - Não
6 - Se sim, o que aconteceu?	Texto Curto:

	Resposta não obrigatória
7 - Por qual dispositivo você acessou a plataforma ?	Caixa de Seleção: ☐Computador / Notebook ☐Celular / Smartphone ☐Tablet ☐Outros
8 - Você possui familiaridade com ferramentas tecnológicas para análise de dados? Por Exemplo: Microsoft Excel	Multipla Escola 1 - Sim 2 - Não
9 - Você conseguiu acessar a visualização e análise de dados disponível na plataforma?	Multipla Escola 1 - Sim 2 - Não
10 - Em uma escala de 1 a 5, quão relevante você considera as informações apresentadas na plataforma? (Sendo um 1 para irrelevante e 5 para muito relevante.)	Multipla Escola 1 – 2 – 3 – 4 – 5
11 - Em uma escala de 1 a 5. Quanto você considera importante a utilização de recursos e ferramentas tecnológicas	Multipla Escola

como suporte para melhoria no sistema educacional? Sendo um 1 para Não é Importante e 5 para Muito Importante	1 – 2 – 3 – 4 – 5
12 - Em uma escala de 1 a 5, você consideraria útil a existência de uma plataforma acessível e de fácil utilização para realizar análises ou obter conhecimento/informação a respeito dos dados educacionais no Brasil? 1 - Não é importante 5 - Muito Importante	Multipla Escola 1 – 2 – 3 – 4 – 5
13 - Espaço Aberto para Criticas / Sugestões / Elogios a respeito da Plataforma Quadro:	Texto curto