UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS RUSSAS

CURSO DE GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MAURÍCIO MATOSO DE PONTES

UM ESTUDO COMPARATIVO DA ANÁLISE DOS ERROS NOS DESCRITORES DAS AVALIAÇÕES EM LARGA ESCALA SPAECE E SAEB COM APLICAÇÃO DA CIÊNCIA DE DADOS

RUSSAS - CEARÁ

2023

MAURÍCIO MATOSO DE PONTES

UM ESTUDO COMPARATIVO DA ANÁLISE DOS ERROS NOS DESCRITORES DAS AVALIAÇÕES EM LARGA ESCALA SPAECE E SAEB COM APLICAÇÃO DA CIÊNCIA DE DADOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Orientador: Prof. Dr. Marcos Vinicius de Andrade Lima.

RUSSAS - CEARÁ

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

P859e Pontes, Maurício Matoso de.
Um estudo comparativo da análise dos erros nos descritores das avaliações em larga escala SPAECE e SAEB com aplicação da ciência de dados / Maurício Matoso de Pontes. – 2023.
47 f.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Ciência da Computação, Russas, 2023.
Orientação: Prof. Dr. Marcos Vinicius de Andrade Lima.

1. Ciência de Dados. 2. Avaliação em Larga Escala. 3. SAEB. 4. SPAECE. 5. FP-Growth. I. Título.

CDD 005

MAURÍCIO MATOSO DE PONTES

UM ESTUDO COMPARATIVO DA ANÁLISE DOS ERROS NOS DESCRITORES DAS AVALIAÇÕES EM LARGA ESCALA SPAECE E SAEB COM APLICAÇÃO DA CIÊNCIA DE DADOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação do Campus Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Ciência da Computação.

Aprovada em: 7 de Dezembro de 2023

BANCA EXAMINADORA

Prof. Dr. Marcos Vinicius de Andrade Lima (Orientador)
Universidade Federal do Ceará (UFC)

Profa. Dra. Anna Beatriz dos Santos Marques
Universidade Federal do Ceará (UFC)

Prof. Ms. Adonias Caetano de Oliveira
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

Ao meu pai, cujo legado vai além do tempo. Sua dedicação à minha educação moldou quem sou hoje. Cada conquista é uma homenagem à sua influência inspiradora.

"Educação não transforma o mundo. Educação muda as pessoas. Pessoas transformam o mundo."

(Paulo Freire)

RESUMO

A geração de dados na área da educação tem crescido significativamente nos últimos anos, impulsionada principalmente pelos dados provenientes de avaliações em larga escala como o Sistema de Avaliação da Educação Básica (SAEB). O SAEB é uma avaliação em larga escala aplicada a alunos de escolas públicas e privadas em todo o Brasil, visando avaliar seu nível de aprendizado. Diante desse contexto, torna-se crucial adotar abordagens que explorem a análise de dados por meio da Ciência de Dados (CD), uma área que nos proporciona diversos recursos para extrair informações cruciais de conjuntos de dados, possibilitando extrair *insights* valiosos para a área da educação. Este trabalho concentra-se na aplicação do algoritmo Frequent Pattern Growth (FP-Growth) para analisar os dados do SAEB. O FP-Growth é amplamente utilizado na área da CD para identificar padrões de associação, aplicando o algoritmo FP-Growth aos dados do SAEB, o objetivo é identificar padrões de associação relacionados aos tópicos em que os alunos enfrentaram dificuldades durante a avaliação. Os *insights* obtidos através dessa abordagem podem fornecer informações valiosas para os educadores e demais responsáveis pela área, permitindo a identificação de possíveis desafios enfrentados pelos alunos que podem ter passado despercebidos anteriormente, ocultos pela massa de dados. Este trabalho visa contribuir para o aprimoramento do processo educacional, permitindo uma intervenção mais direcionada e eficaz em áreas específicas de dificuldade, promovendo, assim, melhorias substanciais no desempenho educacional. Para alcançarmos esse objetivo, primeiramente traçamos todas as características da pesquisa, em seguida, adotamos o processo da CD, composto por seis fases, que se inicia com a definição dos objetivos, passando pela recuperação dos dados, depois a preparação, exploração, modelagem e por fim a apresentação dos resultados. Na última fase do processo da CD, apresentamos os resultados, onde podemos analisar um ranking com as principais dificuldades enfrentadas pelos alunos do Município de Fortaleza, na disciplina de matemática, no SAEB 2019. Além disso, em seguida apresentamos as relações entre essas principais dificuldades dos alunos, demonstrando assuntos que são frequentemente antecedentes e consequentes entre os principais erros apresentados pelos alunos. Após a apresentação dos resultados do SAEB 2019, aplicamos uma análise comparativa com resultados alcançados por outra pesquisa que utilizou os dados do Sistema Permanente de Avaliação da Educação Básica no Ceará (SPAECE) 2019, sendo possível verificar a consistência entre essas duas avaliações. Essa abordagem é fundamental para análise das dificuldades dos alunos, uma vez que ao identificar as relações entre os temas problemáticos, os educadores podem ajustar estratégias de ensino,

desenvolver abordagens mais personalizadas e implementar intervenções pedagógicas eficazes. Dessa forma, a análise das relações entre os assuntos de maior dificuldade não apenas aponta para as deficiências superficiais, mas permite a abordagem do problema na sua origem.

Palavras-chave: avaliação em larga escala; descritores; ciência de dados; FP-Growth; SAEB; SPAECE.

ABSTRACT

Data generation in the field of education has grown significantly in recent years, driven mainly by data from large-scale assessments such as SAEB. SAEB is a large-scale assessment applied to students in public and private schools throughout Brazil, with the aim of evaluating their level of learning. Given this context, it becomes crucial to adopt approaches that explore data analysis through Data Science, a field that provides us with various resources to extract crucial information from data sets, making it possible to extract valuable insights for the field of education. This work focuses on applying the FP-Growth algorithm to analyze SAEB data. FP-Growth is widely used in the field of Data Science (DS) to identify patterns of association. By applying the FP-Growth algorithm to SAEB data, the aim is to identify patterns of association related to the topics in which students faced difficulties during the assessment. The insights gained through this approach can provide valuable information for educators and others responsible for the area, allowing the identification of possible challenges faced by students that may have previously gone unnoticed, hidden by the mass of data. This work aims to contribute to the improvement of the educational process, allowing for more targeted and effective intervention in specific areas of difficulty, thus promoting substantial improvements in educational performance. To achieve this goal, we first outlined all the characteristics of the research, then adopted the DC process, which consists of six phases, starting with defining the objectives, then retrieving the data, then preparing, exploring, modeling and finally presenting the results. In the last phase of the DC process, we present the results, where we can analyze a ranking with the main difficulties faced by students in the Municipality of Fortaleza, in the subject of mathematics, in SAEB 2019. In addition, we then present the relationships between these main difficulties faced by the students, demonstrating issues that are often antecedents and consequents between the main errors presented by the students. After presenting the results of the SAEB 2019, we applied a comparative analysis with the results achieved by another study that used data from the SPAECE 2019, making it possible to verify the consistency between these two assessments. This approach is fundamental for analyzing students' difficulties, since by identifying the relationships between problematic topics, educators can adjust teaching strategies, develop more personalized approaches and implement effective pedagogical interventions. In this way, analyzing the relationships between the most difficult subjects not only points to superficial deficiencies, but allows the problem to be tackled at its source.

Keywords: large-scale assessment; descriptors; data science; FP-Growth; SAEB; SPAECE.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ANEB	Avaliação Nacional da Educação Básica
ANRESC	Avaliação Nacional do Rendimento Escolar
BNCC	Base Nacional Comum Curricular
CD	Ciência de Dados
CSV	*Comma-Separated Values*
Dere	Delegacias Regionais da Educação
EaD	Ensino a Distância
ENEM	Exame Nacional do Ensino Médio
FP-Growth	Frequent Pattern Growth
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LGPD	Lei Geral de Proteção de Dados Pessoais
MEC	Ministério da Educação
SAEB	Sistema de Avaliação da Educação Básica
SEDUC	Secretaria Estadual de Educação do Ceará
SPAECE	Sistema Permanente de Avaliação da Educação Básica no Ceará

SUMÁRIO

1 INTRODUÇÃO

A quantidade de dados gerados nos últimos anos tem crescido exponencialmente em diversas áreas, o que tem impulsionado a necessidade de explorar e extrair valor dessas informações (ALJEHANE, 2020). Com o avanço das tecnologias digitais, como a internet, dispositivos móveis e sensores inteligentes, uma enorme quantidade de dados é gerada a cada instante.

Soterrados sob esses dados estão as respostas para as inúmeras questões que ninguém nunca pensou em perguntar (GRUS, 2016). Essa enorme quantidade de dados, tem transformado a forma como as organizações lidam com informações e tomam decisões. Essa avalanche de dados tem potencial para proporcionar *insights* valiosos e revolucionar práticas e processos em diferentes áreas como as da saúde, economia, educação, *marketing* e governança, que são apenas algumas das muitas áreas que têm se beneficiado dessa enorme quantidade de informações disponíveis (BOYD; CRAWFORD, 2012).

O setor educacional tem visto um aumento significativo na geração de dados, impulsionado pelo uso de sistemas acadêmicos, plataformas de Ensino a Distância (EaD) e avaliações em larga escala. Essa proliferação de dados tem sido explorada em estudos como o de Ferreira *et al.* (2020), em que é abordado o uso do aumento de dados gerados pelas avaliações em larga escala para melhorar a formação dos professores, destacando a importância da análise desses dados como uma ferramenta para auxiliar no planejamento de ações e na melhoria da qualidade da educação.

Nesse contexto, surge a Ciência de Dados (CD), como resultado desse crescimento massivo de dados, tornando-se essencial para lidar com o desafio de extrair conhecimento útil e relevante a partir dos grandes volumes de dados disponíveis (BOYD; CRAWFORD, 2012). Diante desse cenário, pode-se notar a importância de compreender e explorar o crescimento da quantidade de dados nas mais diferentes áreas. Com a compreensão desses dados é possível a aplicação de técnicas e ferramentas avançadas de análise de dados, permitindo a identificação de padrões, correlações e relações que podem ser utilizados para fazer previsões e tomadas de decisão, embasadas em evidências (KHAN; AHMAD, 2016).

Neste trabalho, propomos utilizar os microdados provenientes do Sistema de Avaliação da Educação Básica (SAEB), uma avaliação em larga escala aplicada a cada biênio nas escolas do Brasil.Esses dados podem, por exemplo, ser explorados de maneira estratégica para auxiliar o planejamento e desenvolvimento de ações de formação de professores. Por meio da

análise dos resultados dessas avaliações, é possível identificar padrões, lacunas de conhecimento e áreas de melhoria que podem direcionar programas de capacitação docente. Ao compreender as dificuldades e necessidades dos alunos, evidenciadas pelos dados das avaliações, é possível planejar ações de formação continuada de professores que atendam de forma mais precisa às demandas identificadas. Assim, os dados das avaliações em larga escala podem desempenhar um papel fundamental no processo de formação de professores, direcionando esforços e recursos para as áreas mais necessárias.

No trabalho de Lima (2023), foram abordados os dados da avalialiação em larga escala do SPAECE, trazendo uma nova forma de analisar os resultados dessa avaliação por meio da CD. Os resultados descritos pelo autor poderão ser utilizados para auxiliar o planejamento das ações de formação continuada de professores, pois revelam algumas dificuldades enfrentadas pelos alunos. A partir desse estudo, foi possível levantar a questão de investigação que norteia a presente pesquisa: os dados do SAEB apontam para as mesmas dificuldades verificadas no SPAECE para as disciplinas de Matemática do 9º ano do ensino fundamental para a rede pública do município de Fortaleza para o ano de 2019? Para responder esse questionamento, são apresentados a seguir os objetivos, geral e específicos.

1.1 Objetivo Geral

Comparar os resultados das relações entre descritores verificados na prova em larga escala SPAECE na disciplina de Matemática para o 9º ano do ensino fundamental da rede pública do município de Fortaleza no ano de 2019 com os resultados do SAEB, a fim de avaliar a consistência das avaliações educacionais e fornecer *insights* para possíveis melhorias na qualidade da educação pública.

1.2 Objetivos Específicos

- Extrair a lista de erros e acertos dos estudantes do município de Fortaleza a partir da estrutura dos microdados dos descritores da avaliação em larga escala SAEB referentes à disciplina de Matemática do 9º ano do ensino fundamental no ano de 2019;
- Implementar algoritmo FP-Growth de modo que seja aplicado aos microdados do SAEB 2019;
- Analisar a saída do algoritmo Frequent Pattern Growth (FP-Growth) para a avaliação dos

descritores da disciplina de Matemática do SAEB para o 9º ano do ensino fundamental da rede pública municipal de Fortaleza para o ano de 2019, a fim de identificar os descritores mais frequentes e suas relações.

1.3 Estrutura do Trabalho

O presente trabalho está estruturado em em cinco seções principais. A primeira traz a Introdução, que aborda a contextualização do projeto pesquisa, apresenta a questão de pesquisa e estabelece os objetivos, geral e específicos.

A segunda seção é dedicada à fundamentação teórica, onde são abordados avaliação em larga escala e a Ciência de Dados, explorando especialmente o processo da CD e a técnica de descoberta de padrões por associação.

A terceira seção traz os procedimentos metodológicos, com a definição do paradigma e da abordagem de pesquisa utilizados, perpassando pela fase de coleta de dados, além dos aspectos éticos.

A quarta seção é onde acontece a implementação da metodológia traçada na terceira seção, detalhando os recursos necessários para essa implemetação e os resultados obtidos em cada etapa. Através dessa implementação da metodologia, podemos analisar os resultados dos dados do CD.

A quinta e última seção chegaremos a conclusão dos resultados com a analises dos dados do SAEB, ressaltando a comparação com o trabalho de Lima(2023), além de propor trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção é realizado o embasamento teórico dos pirncipais conceitos necessário para o entendimento deste trabalho. Partindo da Subseção 2.1, onde será explicado o conceito de Avaliação em Larga Escala, origem dos dados manipulados por essa pesquisa, para em seguida, na Subseção 2.2 entendermos conceitos importantes sobre a Ciência de Dados.

2.1 Avaliação em Larga escala

As avaliações em larga escala no Brasil históricamente tiveram íncio na década de 1990, quando aconteceu a primeira iniciativa brasileira, em escala nacional, para se conhecer o sistema educacional brasileiro em profundidade. A primeira edição do Sistema de Avaliação da Educação Básica (SAEB) avaliou os alunos do 5º e do 9º ano do ensino fundamental e do 3º ano do ensino médio (INEP, 2023).

O SAEB foi criado com o objetivo de avaliar o desempenho dos alunos nas disciplinas de Língua Portuguesa e Matemática, além de fornecer informações sobre as condições socioeconômicas e culturais dos alunos e das escolas. Desde então, as avaliações em larga escala se tornaram uma prática comum no Brasil, sendo realizadas regularmente pelo Ministério da Educação (MEC) e pelas secretarias de educação dos estados e municípios. Além do SAEB, destacam-se o Sistema Permanente de Avaliação da Educação Básica do Ceará (SPAECE), que é uma avalição aplicada somente no Estado do Ceará (criada no ano de 1992) e o Exame Nacional do Ensino Médio (ENEM) criado em 1998, aplicado para alunos de todo o Brasil (CASTRO, 2006).

A avaliação em larga escala, uma importante ferramenta de avaliação educacional, busca avaliar o desempenho dos alunos e das escolas em uma determinada região geográfica, geralmente em nível nacional ou estadual. Segundo Hoffmann (2011), a avaliação em larga escala tem como objetivo principal fornecer informações consistentes e precisas para apoiar o planejamento de políticas educacionais e a tomada de decisões. Para Perrenoud e Ramos (1999), a avaliação em larga escala pode ser entendida como um processo sistemático de coleta e análise de informações sobre o desempenho dos alunos em uma determinada área de conhecimento, com o objetivo de avaliar a qualidade da educação.

Encontram-se nesse tipo de avaliação diferentes níveis de decisão, como: formulação de políticas educacionais, orientação do trabalho dos professores, definição de metas e objetivos

para o sistema educacional, entre outros. É uma forma de monitorar a qualidade da educação em uma escala mais ampla e comparar os resultados com outros estados e regiões, permitindo uma análise mais abrangente do desempenho dos alunos e do sistema educacional (CASTRO, 2006).

De acordo com Perrenoud e Ramos (1999), a avaliação em larga escala deve ser capaz de capturar a complexidade do processo educacional, levando em conta não apenas o desempenho dos alunos, mas também as condições em que ocorre o ensino e a aprendizagem, como as condições de formação dos professores, infraestrutura das escolas, recursos didáticos disponíveis, entre outros fatores que influenciam o desempenho dos alunos.

Para que esse tipo de avaliação seja efetiva, é fundamental que ela seja baseada em critérios claros e objetivos, que levem em conta as habilidades e competências que os alunos devem possuir em cada área de conhecimento (CASTRO, 2006). Essas avaliações utilizam como base as Matrizes Curriculates de Referência que são elaboradas de acordo com as Diretrizes Curriculares Nacionais, que por sua vez são orientações gerais para a elaboração dos currículos escolares em todo o país. Essas diretrizes definem as competências e habilidades que os alunos devem adquirir em cada etapa da educação básica, levando em conta as características e necessidades locais. As Matrizes Curriculares de Referência são compostas por um conjunto de descritores, que são indicadores de habilidades e competências que os alunos devem demonstrar em um determinado teste, e que permitem orientar tanto a elaboração do teste quanto o planejamento das atividades de ensino e aprendizagem nas escolas (COTTA, 2014).

Para Locatelli (2002), avaliações em larga escala como o SAEB possuem os seguintes objetivos: monitorar a qualidade, a equidade e efetividade do sistema de educação básica, oferecer às administrações públicas de educação informações que lhes permitam avaliar seus projetos educacionais e formular programas de melhoria da qualidade de ensino e proporcionar aos agentes educacionais e à sociedade informes sobre os resultados dos processos de ensino e dos fatores contextuais a eles associados. Este trabalho tem particular interesse nas avaliações SAEB e SPAECE, detalhadas nas subseções a seguir.

2.1.1 Sistema de Avaliação da Educação Básica

O SAEB é uma avaliação em larga escala nacional organizada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), autarquia vinculada ao Ministério da Educação do Brasil. O SAEB é composto por duas avaliações: a Avaliação Nacional da Educação Básica (ANEB), que avalia o desempenho dos alunos do 5º e 9º anos do ensino

fundamental e 3º ano do ensino médio nas disciplinas de Língua Portuguesa e Matemática como foi descrito no começo desta seção, e a Avaliação Nacional do Rendimento Escolar (ANRESC), que avalia a eficácia das políticas educacionais do país. Por meio de testes e questionários, aplicados a cada dois anos na rede pública e em uma amostra da rede privada, o SAEB reflete os níveis de aprendizagem demonstrados pelos estudantes avaliados, explicando esses resultados a partir de uma série de informações contextuais. Realizado desde 1990, o 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 Base Nacional Comum Curricular (BNCC) (INEP, 2023).

A matriz de referência para a avaliação do 9º ano do ensino fundamental, que orienta a elaboração dos testes do SAEB, é composta por descritores de habilidades e competências que os alunos devem demonstrar em Língua Portuguesa e Matemática. Além disso, também é aplicado um questionário socioeconômico aos alunos, professores e diretores das escolas participantes. Os descritores são organizados em blocos de conteúdos, que abrangem desde conceitos fundamentais até habilidades mais complexas, como a resolução de problemas. A partir dessa matriz, são elaborados os testes do SAEB, que são aplicados a uma amostra representativa de alunos de escolas públicas e privadas de todo o país (COTTA, 2014).

A prova de Matemática do SAEB 2019 aplicada aos alunos do 9º ano é composta por 26 questões, divididas em dois blocos de 13 cada. Os 37 descritores associados a essa prova abrangem diferentes temas, sendo os primeiros 11 relacionados a 'ESPAÇO E FORMA', os seguintes 4 a 'GRANDEZAS E MEDIDAS', e do descritor 16 ao 35 focados em 'NÚMEROS E OPERAÇÕES/ÁLGEBRA E FUNÇÕES'. Os dois últimos descritores abordam o tema 'TRATAMENTO DA INFORMAÇÃO'. Cada descritor especifica competências que os alunos devem demonstrar, proporcionando uma visão abrangente das habilidades avaliadas na prova e orientando a preparação dos estudantes (Ministério da Educação, 2023).

Portanto, o SAEB tem como objetivo principal avaliar a qualidade da educação básica no Brasil, identificar as deficiências e pontos fortes do sistema educacional, e orientar políticas públicas para o setor. Os resultados permitem a comparação do desempenho dos alunos de diferentes estados e regiões do país, bem como a avaliação do progresso da educação ao longo do tempo (COTTA, 2014). O SAEB inspirou a criação de diversos sistemas de estimativas como o SPAECE, detalhado na sequência.

2.1.2 Sistema Permanente de Avaliação da Educação Básica no Ceará

O SPAECE foi criado em 1992 na gestão do governador Ciro Gomes (1991 – 1994) e desde então tem evoluído, ampliando sua abrangência e aperfeiçoando sua estrutura metodológica. A primeira edição contemplou somente a capital Fortaleza, quando foram avaliados 10.590 alunos da 4ª série e 4.010 alunos da 8ª série de 157 escolas estaduais. Os instrumentos continham vinte e cinco questões de múltipla escolha abordando conteúdos de Língua Portuguesa e Matemática. Na segunda edição, em 1993, a amostra contemplou, além de Fortaleza, as sedes das catorze Delegacias Regionais da Educação (Dere), totalizando quinze municípios e um total de 22.886 alunos avaliados. Em 2007, o Spaece passou por uma nova reformulação, voltando a ser realizado anualmente. Entre 2008 e 2019, o sistema tem se fortalecido, passando a integrar o planejamento escolar, planejamento docente e em todas as ações e programas implantados pela Secretaria Estadual de Educação do Ceará (SEDUC), que têm um objetivo comum elevar os indicadores do SPAECE considerando-os como reflexo da melhoria da aprendizagem dos alunos e da qualidade do sistema de ensino (CODED/CED, 2023).

O objetivo do SPAECE é o diagnóstico sobre o estágio das competências e habilidades dos estudantes nas disciplinas de Língua Portuguesa, com foco em leitura, e Matemática dos alunos do ensino fundamental (5º e 9º anos). O SPAECE é composto por duas avaliações: a avaliação externa, realizada por aplicadores externos como o governo ou outras agências de avaliação, e tem como objetivo fornecer dados para políticas públicas de educação e para comparação entre escolas e regiões. E a avaliação interna, aquela realizada pela própria instituição educacional, seja pelos professores, pela coordenação pedagógica ou pela direção, e tem como objetivo acompanhar o desempenho dos alunos e avaliar o processo de ensino-aprendizagem dentro da escola (MAGALHÃES; FARIAS, 2016).

Assim como o SAEB, o SPAECE também possui uma Matriz de Referência, que contém as habilidades e competências que os alunos devem desenvolver nas áreas avaliadas. A Matriz de Referência serve de base para a elaboração das provas. Os resultados do SPAECE são divulgados para as escolas, gestores, professores e sociedade em geral, com o objetivo de subsidiar a tomada de decisões e ações para a melhoria da qualidade da educação no estado do Ceará (MAGALHÃES; FARIAS, 2016).

Além disso, é importante que a avaliação em larga escala seja capaz de fornecer *feedbacks* úteis e significativos para os professores e para os alunos, de forma a orientar a melhoria do desempenho dos alunos e do sistema educacional como um todo. Nesse sentido, é

fundamental que os resultados da avaliação sejam interpretados de forma crítica e reflexiva, e que sejam utilizados para orientar ações pedagógicas mais direcionadas e efetivas.

As avaliações em larga escala geram uma quantidade enorme de informações que precisam ser processadas e analisadas de forma eficiente e eficaz e a CD pode desempenhar um papel fundamental na análise dessas informações. Por meio da CD pode-se utilizar técnicas estatísticas avançadas para analisar grandes conjuntos de dados, como aqueles gerados pelas avaliações em larga escala. Essas técnicas permitem identificar padrões e relações entre as diferentes variáveis avaliadas, e ajudam a identificar fatores que afetam o desempenho dos alunos, permitindo que sejam criadas ações específicas para melhorar o aprendizado e o desempenho escolar. Uma das possibilidades de técnicas de CD é a análise exploratória de dados que consiste em explorar os dados para identificar padrões e relações entre as diferentes variáveis. Através da análise exploratória, é possível obter *insights* iniciais sobre os dados e identificar quais variáveis estão mais relacionadas ao desempenho dos alunos (FERREIRA *et al.*, 2020). A CD será mais detalhada na seção seguinte.

2.2 Ciência de Dados

Para Molina-Markham *et al.* (2019) a Ciência de Dados (CD), em inglês *Data Science*, é uma área interdisciplinar que utiliza métodos estatísticos, computacionais e matemáticos para extrair *insights* e conhecimentos a partir de dados. *Insights* são conclusões ou entendimentos obtidos a partir da análise de dados.

Boyd e Crawford (2012) falam que a CD surgiu como resultado do crescente volume, variedade e velocidade dos dados disponíveis, impulsionado pela evolução da tecnologia e das mídias sociais. Khan e Ahmad (2016) descrevem que a capacidade de coletar, armazenar e processar grandes quantidades de dados tornou possível a análise de informações em escala e detalhe nunca antes alcançados, permitindo a identificação de padrões, correlações e relações que podem ser utilizados para fazer previsões e tomadas de decisões.

Segundo Wu e Zhang (2014), a CD se vale de uma variedade de técnicas, tais como mineração de dados, aprendizado de máquina, análise estatística e visualização de dados, para extrair informações úteis a partir dos dados. Essas técnicas permitem que os cientistas de dados trabalhem com dados estruturados e não estruturados de diversas fontes, incluindo bancos de dados, mídias sociais, dispositivos móveis e sensores (BÜHLMANN; GEER, 2011).

Um dos maiores desafios na Ciência de Dados é lidar com a grande quantidade de

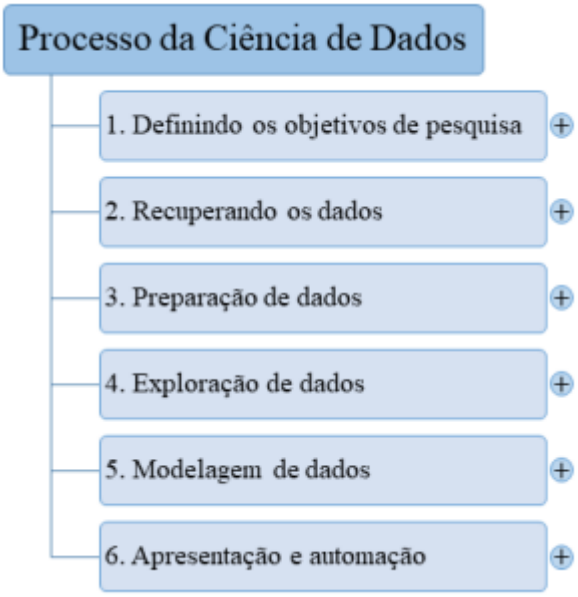
dados não estruturados e não organizados disponíveis. Para enfrentar esse desafio, as técnicas de pré-processamento de dados, como limpeza, transformação e organização, são essenciais para preparar os dados para a análise (BÜHLMANN; GEER, 2011).

A Ciência de Dados pode ser aplicada em diversas áreas, como finanças, saúde, meio ambiente, entre outras, para identificar padrões, tendências e *insights* que possam ser utilizados para tomar decisões estratégicas. Como explica César (2019), a CD envolve diversas etapas, incluindo coleta de dados, limpeza e pré-processamento, análise exploratória de dados, modelagem de dados e validação de modelos. É importante utilizar ferramentas e técnicas avançadas para cada tipo de problema para que seja possível extrair *insights* valiosos. Pode-se entender um pouco mais sobre essas etapas da CD a seguir.

2.2.1 O Processo da CD

O processo da CD que será utilizado neste trabalho, proposto por Cielen *et al.* (2016), é composto por seis fases, que se inicia com a definição dos objetivos da pesquisa e finaliza com a apresentação dos resultados, como podemos ver na Figura 1 e no detalhamento abaixo:

Figura 1 – Fases do processo da Ciência de Dados



Fonte: (CIELEN *et al.*, 2016)

1. **Definição dos objetos de pesquisa:** na fase inicial é criado o termo de abertura para o projeto de CD que é desenvolvido de acordo com as circunstâncias da organização e/ou do departamento. Nesta etapa são definidas algumas informações básicas que incluem

a metodologia da pesquisa, os benefícios esperados, os dados e recursos necessários, o cronograma e o que será entregue.

2. **Recuperação de dados:** no termo de abertura, é informado quais os dados necessários e onde encontrá-los. Nesta etapa é garantida a possibilidade de utilização dos dados no projeto de CD, significa que é verificada a existência dos dados, a qualidade dos dados e o seu acesso.
3. **Preparação de dados:** nesta fase começa o manuseio dos dados, sendo preparados para o uso nas próximas fases. Esta fase está dividida em três subfases: i) limpeza de dados – com a remoção de valores errados e inconsistências; ii) integração de dados – em que dados de várias fontes são combinados; e iii) transformação de dados – momento em que todos os dados são transformados em um formato adequado para utilização pelos modelos.
4. **Exploração de dados:** aqui o foco é no entendimento aprofundado dos dados, que inclui compreensão das relações existentes entre as variáveis, a distribuição dos dados e a presença de valores discrepantes ou pontos fora da curva. Isso geralmente requer o uso de modelagem simples, estatística descritiva e técnicas visuais.
5. **Modelagem de dados:** Os modelos são usados para responder à questão de pesquisa após a compreensão dos dados e do conhecimento do domínio da fase anterior. Nesta fase, os métodos são escolhidos dos domínios da estatística, aprendizado de máquina e pesquisa operacional, entre outros. Como construir um modelo envolve escolher variáveis, executar o modelo e diagnosticar os resultados, essa fase do processo é iterativa.
6. **Apresentação e automação:** Os resultados são apresentados aqui na fase final. Vale ressaltar a importância de automatizar o processo, pois a organização e/ou do departamento poderá utilizar os resultados em outro projeto.

Dentro do processo da CD é necessário escolher uma técnica para ser aplicada na fase de Modelagem dos dados, devendo ser levado em conta uma técnica que possa revelar associações entre itens que não seriam facilmente perceptíveis ou esperadas. Neste trabalho, será aplicada a técnica de Descoberta de Padrões de Associação, descrita a seguir.

2.2.2 Técnica de Descoberta de Padrões de Associação

A técnica de descoberta de padrões de associação é uma das técnicas fundamentais na ciência de dados. Ela é usada para identificar relações e associações entre diferentes itens em grandes conjuntos de dados. O objetivo principal é descobrir regras que indiquem a coocorrência

frequente de itens ou eventos (AMORIM, 2006). Essa técnica é amplamente utilizada em diferentes setores, como varejo, *marketing*, análise de mercado e recomendação de produtos. Ela fornece *insights* valiosos sobre os comportamentos dos consumidores e pode auxiliar na tomada de decisões estratégicas.

A técnica de descoberta de padrões por associação é frequentemente aplicada em conjuntos de transações, onde cada transação consiste em um conjunto de itens. O exemplo clássico é o carrinho de compras de um supermercado, onde cada compra representa uma transação e os itens são os produtos adquiridos. A partir desses dados, a técnica de descoberta de padrões por associação pode identificar quais itens são frequentemente comprados juntos, gerando regras de associação (GOLDSCHMIDT; BEZERRA, 2023).

Um exemplo de regra de associação seria: "Se um cliente compra pão e leite, então é provável que também compre manteiga". Essa regra indica a associação frequente entre esses itens e pode ser útil para ações como a recomendação de produtos relacionados ou o ajuste de estratégias de colocação de produtos nas prateleiras.

A técnica de descoberta de padrões por associação utiliza algoritmos como o Apriori e o FP-Growth para identificar as associações mais significativas nos dados. Esses algoritmos aplicam medidas como suporte, confiança e *lift* para avaliar a importância das regras de associação encontradas. O Apriori é o algoritmo clássico. Ele possui alguns problemas de desempenho, especialmente quando é preciso analisar uma grande quantidade de transações. O FP-Growth é um algoritmo mais recendo, possuindo um desempenho superior ao Apriori porque utiliza uma estrutura de dados em árvore aliada combinando com a técnica de programação dividir-para-conquistar (SILVA *et al.*, 2017).

Pode-se pensar no FP-growth como um detetive de dados especializado em encontrar combinações significativas de itens em grandes conjuntos de informações. As entradas principais do algoritmo FP-Growth são conjuntos de transações. Cada transação representa uma lista de itens que foram adquiridos ou ocorreram juntos em algum contexto. Ele gera conjuntos frequentes, que são grupos de itens que ocorrem juntos com uma frequência acima de um limite definido (suporte mínimo). Esses conjuntos frequentes são as principais descobertas do algoritmo. Para cada conjunto frequente, o algoritmo também fornece a contagem associada, indicando quantas vezes esse conjunto específico aparece nas transações. Essa contagem é útil para avaliar a importância ou popularidade do padrão.

Segundo Mariano (2011), o algoritmo FP-growth é composto por duas fases de

processamento. Na primeira fase é construída é uma representação da base de dados, chamada de FP-Tree, que é feita através da estratégia de busca em profundidade e da contagem de ocorrência de itens. Já na segunda fase o FP-Growth utiliza a FP-Tree para determinar os valores de suporte para todos os *itemsets* frequentes. O processo de contrução da FP-Tree e do FP-Growth são mostrados a seguir:

Figura 2 – Pseudo código da Função FP-Tree

1. **Entrada**: Uma base de dados D e o valor de suporte mínimo *min_sup*.
2. **Saída**: O conjunto L com todos os *itemsets* frequentes.
// Fase 1 - Construção da *FP-Tree*
3. **Função** *FP-Tree*(D, *min_sup*)
4. **percorra** a base de dados D uma vez;
5. **determine** o conjunto de itens frequentes F e seus suportes;
6. **ordene** F em ordem decrescente em função do suporte e chamá-la de L;
7. **crie** a raiz da *FP-Tree* T e coloque como "*null*";
8. **para cada** transação t em D **faça**
9. **selecione e ordene** os itens frequentes em t de acordo com a ordem de L, tornando a lista de itens frequentes em t igual a $[p|P]$, onde p é o primeiro elemento e P é o resto da lista;
10. **execute** *insere_tree*($[p|P]$, T).
11. **se** T tiver um filho N em que $N.nome\_item = p.nome\_item$ **então**
12. **incremente** o contador de N por em 1;
13. **senão**
14. **crie** um novo nó N, e inicie seu contador com 1;
15. **ligue** o seu *parent-link* a T, e seu *node-link* aos nós de mesmo (*nome_item*) através da estrutura dos *node-links*;
16. **fim se**
17. **se** P não for vazio **então**
18. **chame** *insere_tree*(P, N) recursivamente;
19. **fim se**
20. **fim para**

Fonte: (MARIANO, 2011)

Figura 3 – Pseudo código da Função FP-Growth

```
//  Fase 2 - Mineração da FP-Tree
21. Função FP-Growth(Tree, α)
22. se Tree contém apenas um caminho P então
23.       para cada combinação β de nós no caminho P faça
24.             gere o padrão β ∪ α com suporte = min_sup dos nós em β;
25.       fim para
26. senão
27.       para cada a_i na tabela de node-links de Tree faça
28.             gere o padrão β = a_i ∪ α com suporte = a_i.suporte;
29.             construa a base de padrões condicionada de β e crie
                a FP-Tree condicionada de β chamada de Tree_β;
30.             se (Tree_β ≠ ∅) então
31.                   FP-growth(Tree_β, β);
32.             fim se
33.       fim para
34. fim se
```

Fonte: (MARIANO, 2011)

3 PROCEDIMENTOS METODOLÓGICOS

Esta seção descreve a a metodologia a ser aplicada. Portanto, há uma introdução contextualizando o tipo de pesquisa científica aplicada neste trabalho, seguindo para a descrição das etapas executadas para alcançar o objetivo principal deste trabalho.

3.1 Pesquisa científica

A contribuição da pesquisa científica é de extrema importância para o avanço da nossa sociedade. Através dela, adquirimos conhecimento, resolvemos problemas complexos e impulsionamos o progresso em diferentes áreas. A pesquisa científica nos possibilita aprimorar nossa compreensão sobre o mundo em que vivemos, desvendar os fenômenos naturais que nos cercam e desenvolver novas tecnologias para melhorar a saúde e o bem-estar das pessoas. Além disso, ela pode servir como base para a formulação de políticas públicas relevantes (ROSSI-BARBOSA, 2013).

Esta pesquisa está inserida no campo da Computação com aplicação na área da Educação. Especificamente, ela se concentra no auxílio da Ciência de Dados para realizar análises dos resultados de avaliações em larga escala, visando fornecer informações relevantes para a tomada de decisão. Através da aplicação de técnicas e tecnologias da Ciência de Dados, busca-se extrair *insights* valiosos a partir dos dados das avaliações em larga escala, contribuindo para aprimorar as políticas e práticas educacionais.

3.2 Caracterização da pesquisa

Nesta pesquisa, seguimos o paradigma pragmático, no qual defende que a pesquisa deve-se preocupar com as aplicações, o que funciona e as soluções para os problemas, ao invés de se concentrar nos métodos. No pragmatismo busca-se enfatizar o problema, utilizando todas as abordagens disponíveis para o seu entendimento (CRESWELL, 2010).

Em relação a forma da abordagem, trata-se de uma pesquisa quantitativa com aplicação do processo da CD descrito por Cielen *et al.* (2016) que foi detalhado anteriormente na Subseção 2.2.1, com o objetivo de realizar um estudo comparativo entre os resultados da aplicação da análise dos erros dos estudantes na prova do SAEB e do SPAECE no ano de 2019, utilizando uma das técnicas de descoberta de padrões de associação 2.2.2, mais especificamente o algoritmo FP-Growth. Pesquisas quantitativas facilitam a generalização de resultados e

possibilidade de réplicas e comparações entre estudos similares que é exatamente o caso deste trabalho (CRESWELL, 2010).

3.3 Coleta de dados

Em relação aos procedimentos utilizados para coleta de dados, foi feito por meio do acesso aos microdados do SAEB do ano de 2019 fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Esses microdados são informações abrangentes sobre os desempenhos do alunos, as características das escola e outros dados relevantes. O uso desses dados é essencial para aplicação do processo de CD e comparação dos resultados na tese de Lima (2023).

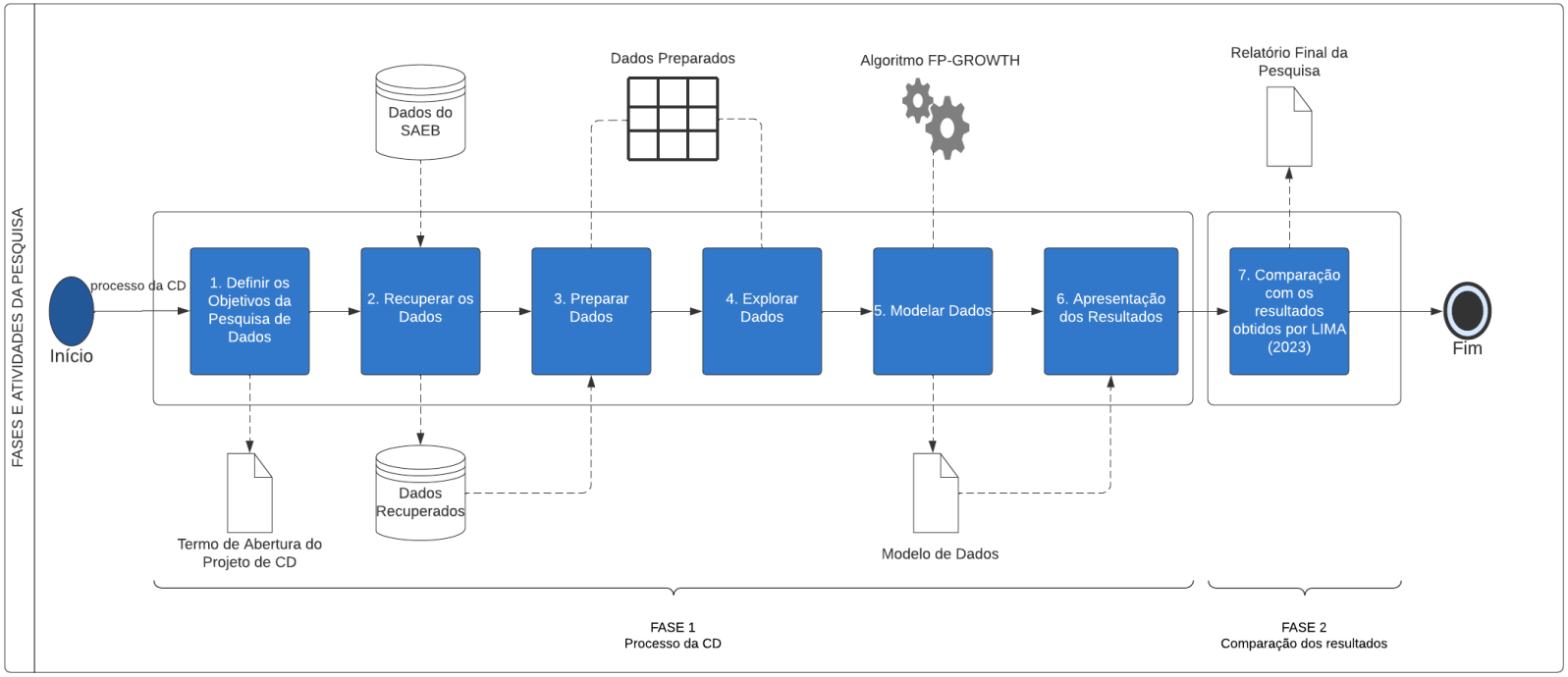
Após a realização da coleta, é importante a análise dos dados, utilizando o trabalho de Lima (2023) como um ponto de referência para essa análise e interpretação dos dados, visto que o trabalho desse autor utilizou dados de outra avaliação em larga escala compatível.

3.4 Fases da pesquisa

O procedimento proposto para realização desta pesquisa científica foi dividido em duas fases principais: o processo da CD e a realização da análise comparativa com os resultados da tese de Lima (2023).

A Figura 4 apresenta a sequência de passos adotados para a realização da pesquisa. Os processos numerados de 1 a 6 fazem parte do processo da CD, descritos na Subseção 2.2.1, enquanto o processo 7 diz respeito a realização da comparação dos resultados desta pesquisa com os valores encontrados por Lima (2023).

Figura 4 – Fases e atividades da pesquisa.



Fonte: Elaborado pelo autor.

A primeira fase consiste na definição dos objetivos da pesquisa, que foram descritos nas Subseções 1.1 e 1.2 respectivamente. A segunda fase é a recuperação dos dados, que também já foi descrita na Subseção 3.3. Dito isto, agora podemos detalhar as próximas fases da pesquisa nas subseções a seguir.

3.4.1 Preparação dos Dados

A preparação dos dados é a terceira fase do processo da CD, que aconteceu da seguinte forma, utilizando uma abordagem prática, implementada no ambiente do Jupyter Notebook, aproveitando a flexibilidade da linguagem de programação Python e a eficácia da biblioteca Pandas[1] para manipulação de dados.

Os microdados do SAEB estavam armazenados em arquivos no formato *Comma-Separated Values* (CSV), o que tornou a tarefa de leitura e manipulação mais acessível. Foi necessário a leitura de dois arquivos, o primeiro contêm as informações dos alunos e suas respostas na avaliação, com tamanho de 866,6 Megabytes e 2.388.931 linhas de dados. O segundo arquivo contém o gabarito de todas as disciplinas, com tamanho bem menor, de apenas 0,0135 Megabytes e 854 linhas.

É importante ressaltar que para a leitura desses arquivos ser possível, foi preciso aplicar uma limpeza de dados, pois durante a manipulação do arquivo CSV, identificou-se um desafio relacionado aos tipos de dados em determinadas colunas, nas quais apresentaram particularidades que demandaram uma abordagem especial. O problema é que elas precisavam ter apenas dados do tipo *string*, porém estavam com dados de outros tipos e isso estava dificultando o processo de leitura do arquivo. Para resolver esse problema, foi implementado um tratamento específico no qual as colunas problemáticas foram configuradas para serem interpretadas como strings. Além disso, foram excluídas as respostas incompletas nos itens da prova de matemática.

3.4.2 Exploração dos Dados

A exploração dos dados é a quarta fase, com a limpeza dos dados e a obtenção de um formato adequado para utilização. Foi possível a aplicação de filtros nesses dados para garantir o foco desta pesquisa. Esses filtros foram projetados para restringir os dados a um conjunto específico de observações que atendesse aos objetivos da pesquisa, no caso para ir de encontro aos dados analisados por Lima (2023) no SPAECE. No *dataset* dos alunos foi aplicado um filtro

[1] Disponível em: <https://pandas.pydata.org/>

para o município de Fortaleza, para a série, o 9º ano do ensino fundamental e para a disciplina de Matemática, isso reduziu o arquivo com os dados dos alunos para 1,1 Megabytes e 21.220 linhas, um arquivo bem menor comparado ao arquivo inicial descrito na etapa anterior e que traz os dados de todos os municípios do Brasil. Já no *dataset* dos gabaritos foi necessário apenas filtrar por série e por município, reduzindo seu tamanho para 0,005 Megabytes e 91 linhas. Esses filtros não apenas refinaram os dados, mas também contribuíram para a precisão e a relevância dos resultados finais, tornando a análise mais específica e interpretável à luz dos objetivos da pesquisa.

Continuando com a exploração dos dados, agora com os dados dos alunos e dos gabaritos refinados, chegamos ao momento de rodar um *script* para que a comparação entre as resposta e o gabarito pudesse ser feita, para que assim seja possível armazenar, em caso de erro, o descritor de cada item errado por cada aluno. O *script* foi desenvolvido para comparar as respostas dos alunos com o gabarito em dois blocos diferentes de questões do SAEB, disponibilizado no APÊNDICE A.

Após a execução do *script* de extração, foi possível obter uma lista com os descritores dos erros dos alunos do 9º ano do ensino fundamental, do município de Fortaleza na avaliação em larga escala SAEB para o ano de 2019. Podemos observar uma pequena parte dessa lista como exemplo na Tabela 1.

Tabela 1 – Exemplo de recorte de dados utilizados como entrada no FP-growth

conj_desc_erro
’D35’, ’D26’, ’D8’, ’D21’, ’D31’
’D23’, ’D10’, ’D8’, ’D21’, ’D12’, ’D3’, ’D29’, ’D6’
’D7’, ’D14’, ’D33’, ’D12’, ’D32’, ’D29’
’D11’, ’D7’, ’D29’, ’D30’, ’D4’, ’D5’, ’D18’, ’D13’

Fonte: Elaborada pelo autor.

3.4.3 Modelagem dos Dados

Com a lista dos descritores obtida, foi possível então chegar a quinta fase do processo da CD que é a modelagem dos dados, onde acontece a aplicação dessa lista de descritores mostrados na Tabela 1 no algoritmo FP-Growth para obter o modelo de dados. O algoritmo FP-Growth é usado para descobrir padrões de associação frequentes em conjuntos de dados transacionais. Abaixo, é descrito o funcionamento desse algoritmo:

1. **Conversão dos Dados em Transações:** inicialmente, os dados de entrada, neste caso, os

descritores de erro obtidos dos alunos, são convertidos em transações. Cada transação é uma lista de itens (neste caso, descritores de erro) associados a um único aluno.

2. **Construção da Árvore FP-tree:** o próximo passo envolve a construção de uma estrutura de dados chamada FP-tree a partir das transações. A árvore FP-tree é usada para representar os padrões de associação encontrados nos dados. A árvore é construída começando com um nó raiz vazio, em seguida, adicionando os itens das transações de acordo com sua frequência.
3. **Ordenação dos Itens:** Os itens nas transações são ordenados em ordem decrescente de frequência. Isso ajuda a otimizar a construção da árvore FP-tree.
4. **Construção da Árvore Condensada (Conditional FP-tree):** Após a construção da FP-tree, é criada uma árvore condensada, conhecida como Conditional FP-tree, para cada item frequente na base de dados. A Conditional FP-tree é criada removendo os itens infrequentes e mantendo apenas os itens frequentes e suas relações.
5. **Extração de Padrões de Associação Frequentes:** O algoritmo realiza uma exploração recursiva na Conditional FP-tree para extrair padrões de associação frequentes. Ele inicia com o item mais infrequente na Conditional FP-tree e constrói padrões frequentes a partir dele, identificando combinações frequentes de itens.
6. **Análise e Filtragem dos Padrões:** Após a extração dos padrões de associação, os resultados são analisados e podem ser filtrados com base em medidas de suporte e confiança para identificar os padrões mais relevantes.

Para a aplicação do algoritmo FP-growth na lista de descritores de erros do SAEB, utilizamos a biblioteca de processamento de dados Apache Spark, aproveitando sua implementação eficiente do algoritmo. A configuração específica empregada para executar o algoritmo pode ser visualizada no Quadro 1:

Quadro 1 – Configuração aplicada no algoritmo FP-Growth.

```
fpGrowth = FPGrowth(itemsCol="descritors", minSupport=0.15, minConfidence=0.7)
```

Fonte: Elaborada pelo autor.

Em que, 'descritors' representa a coluna que contém os descritores das respostas dos alunos, enquanto 'minSupport' e 'minConfidence' foram definidos como 0.15 e 0.7, respectivamente. Isso significa que apenas os padrões que aparecem em pelo menos 15% das transações são considerados frequentes, e as regras de associação só são aceitas se sua confiança for maior ou igual a 70%.

Essa configuração específica nos permitiu identificar os padrões mais relevantes e confiáveis presentes nos descritores de erros do SAEB, o que foi crucial para compreender o desempenho e as tendências dos alunos.

3.5 Aspectos éticos da pesquisa

No contexto dessa pesquisa, é de extrema importância considerar os aspectos éticos relacionados à proteção de dados pessoais, em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD), especificamente a Lei nº 13709/18. Como foi informado na Subseção 3.3, os dados utilizados nesta pesquisa foram fornecidos pelo INEP, e é fundamental respeitar as diretrizes estabelecidas pela legislação para garantir a privacidade e a segurança dos indivíduos envolvidos, no caso os estudantes do 9º ano do ensino fundamental da rede pública do município de Fortaleza.

A LGPD estabelece princípios, direitos e obrigações para o tratamento de dados pessoais, buscando assegurar a privacidade, a transparência e o controle dos indivíduos sobre suas informações pessoais. Como pesquisador, assumo a responsabilidade de garantir a anonimização e a confidencialidade dos dados utilizados, de forma a proteger a identidade dos participantes e preservar a integridade dos dados. Dessa forma, a pesquisa estará em conformidade com a LGPD, assegurando a proteção dos direitos dos indivíduos e contribuindo para o avanço do conhecimento científico de forma ética e responsável.

4 RESULTADOS E DISCUSSÃO

Nesta Seção, apresentamos os resultados dos descritores de erros mais frequentes no SAEB 2019 e das principais relações entre esses descritores. Posteriormente abordaremos a realização da análise comparativa dos dados obtidos nesta pesquisa, que focou no SAEB, com os resultados da tese de Lima (2023), que analisou o SPAECE.

4.1 Apresentação dos Resultados

Após obter o modelo de dados, como foi descrito na Subseção 3.4.3, chegamos à sexta fase do processo da CD proposto por Cielen *et al.* (2016), que é a apresentação dos resultados. Antes de começar a análise dos resultados, o primeiro ponto que vamos apresentar são os descritores com maior frequência na lista dos descritores de erros do SAEB, com isso podemos fazer um levantamento de quais são os descritores que representam os assuntos com maior dificuldade pelos os alunos. Essas informações servirão como base para que posteriomente os resultados trazidos pelo algoritmo FP-growth possam ser analisados com maior embasamento. Outro fator importante, que também destacamos, é que a construção de ranking dos erros de descritores é a técnica padrão utilizada para analisar os resultados das provas em larga escala no Brasil (LIMA, 2023).

Para a análise dos descritores de erro mais frequentes, utilizamos a biblioteca *Collections* em conjunto com a função *Counter*. A biblioteca collections fornece estruturas de dados adicionais e ferramentas para realizar operações úteis em Python. A função *Collections* é uma ferramenta que permite contar a frequência de elementos em uma lista. Aplicando essa abordagem foi possível chegar aos resultados disponíveis na Tabela 2:

Tabela 2 – Frequência dos descritores de erros no SAEB, 2019.

Descritor	Frequência
D20	9012
D10	8487
D29	7979
D28	7449
D13	7304
D7	6963
D4	6706
D24	6615
D11	6300
D3	6128
D32	5755
D14	5517
D21	5317
D23	5052
D6	4665
D12	4621
D1	4194
D26	4017
D18	3945
D8	3896
D22	3870
D36	3768
D16	3683
D37	3456
D34	3431
D2	3074
D31	2955
D9	2903
D33	2855
D35	2446
D30	2336
D25	2195
D5	1833
D19	1479
D15	1398
D27	475

Fonte: Elaborada pelo autor.

Com esses resultados, podemos examinar a tabela de descritores de erros do SAEB, analisando quais são os temas que se destacaram como os assuntos de maiores dificuldades enfrentadas pelos alunos. O Quadro 2 traz a matriz de referência utilizada na prova de matemática para o 9º ano do ensino fundamental, nos quais são apresentados os descritores constituintes.

Quadro 2 – Matriz de referência do SAEB 2019 para Matemática, 9º ano do ensino fundamental

I. ESPAÇO E FORMA	
D1	Identificar a localização/movimentação de objeto em mapas, croquis e outras representações gráficas.
D2	Identificar propriedades comuns e diferenças entre figuras bidimensionais e tridimensionais, relacionando-as com as suas planificações.
D3	Identificar propriedades de triângulos pela comparação de medidas de lados e ângulos.
D4	Identificar relação entre quadriláteros por meio de suas propriedades.
D5	Reconhecer a conservação ou modificação de medidas dos lados, do perímetro, da área em ampliação e/ou redução de figuras poligonais usando malhas quadriculadas.
D6	Reconhecer ângulos como mudança de direção ou giros, identificando ângulos retos e não- retos.
D7	Reconhecer que as imagens de uma figura construída por uma transformação homotética são semelhantes, identificando propriedades e/ou medidas que se modificam ou não se alteram.
D8	Resolver problema utilizando propriedades dos polígonos. (soma de seus ângulos internos, número de diagonais, cálculo da medida de cada ângulo interno nos polígonos regulares).
D9	Interpretar informações apresentadas por meio de coordenadas cartesianas.
D10	Utilizar relações métricas do triângulo retângulo para resolver problemas significativos.
D11	Reconhecer círculo/circunferência, seus elementos e algumas de suas relações.
II. GRANDEZAS E MEDIDAS	
D12	Resolver problema envolvendo o cálculo de perímetro de figuras planas.
D13	Resolver problema envolvendo o cálculo de área de figuras planas.
D14	Resolver problema envolvendo noções de volume.
D15	Resolver problema utilizando relações entre diferentes unidades de medida.
III. NÚMEROS E OPERAÇÕES/ÁLGEBRA E FUNÇÕES	
D16	Identificar a localização de números inteiros na reta numérica.
D17	Identificar a localização de números racionais na reta numérica.
D18	Efetuar cálculos com números inteiros, envolvendo as operações (adição, subtração, multiplicação, divisão, potenciação).
D19	Resolver problema com números naturais, envolvendo diferentes significados das operações (adição, subtração, multiplicação, divisão, potenciação).
D20	Resolver problema com números inteiros envolvendo as operações (adição, subtração, multiplicação, divisão, potenciação).
D21	Reconhecer as diferentes representações de um número racional.
D22	Identificar fração como representação que pode estar associada a diferentes significados.
D23	Identificar frações equivalentes.
D24	Reconhecer as representações decimais dos números racionais como uma extensão do sistema de numeração decimal, identificando a existência de "ordens"como décimos, centésimos e milésimos.
D25	Efetuar cálculos que envolvam operações com números racionais (adição, subtração, multiplicação, divisão, potenciação).
D26	Resolver problema com números racionais envolvendo as operações (adição, subtração, multiplicação, divisão, potenciação).
D27	Efetuar cálculos simples com valores aproximados de radicais.
D28	Resolver problema que envolva porcentagem.
D29	Resolver problema que envolva variação proporcional, direta ou inversa, entre grandezas.
D30	Calcular o valor numérico de uma expressão algébrica.
D31	Resolver problema que envolva equação do 2º grau.
D32	Identificar a expressão algébrica que expressa uma regularidade observada em seqüências de números ou figuras (padrões).
D33	Identificar uma equação ou inequação do 1º grau que expressa um problema.
D34	Identificar um sistema de equações do 1º grau que expressa um problema.
D35	Identificar a relação entre as representações algébrica e geométrica de um sistema de equações do 1º grau.
IV. TRATAMENTO DA INFORMAÇÃO	
D36	Resolver problema envolvendo informações apresentadas em tabelas e/ou gráficos.
D37	Associar informações apresentadas em listas e/ou tabelas simples aos gráficos que as representam.

Fonte: Elaborada pelo autor.

A seguir, destacamos os três descritores de erros mais frequentes, com base nas áreas que demandam atenção prioritária listadas na Tabela 2:

i) D20 - Resolver Problemas com Números Inteiros:

- Um número significativo de alunos enfrentou desafios na resolução de problemas envolvendo números inteiros e suas operações. Adição, subtração, multiplicação, divisão e potenciação foram áreas específicas que exigem reforço. A falta de domínio desses conceitos afeta diretamente o desempenho em diversas disciplinas matemáticas.

ii) D10 - Utilizar Relações Métricas do Triângulo Retângulo:

- Uma parcela considerável dos estudantes demonstrou dificuldades na aplicação das relações métricas do triângulo retângulo para resolver problemas significativos. Essa lacuna de conhecimento impacta diretamente em habilidades essenciais para a compreensão da geometria e suas aplicações práticas.

iii) D29 - Resolver Problema que Envolva Variação Proporcional:

- Outro ponto crítico foi a dificuldade dos alunos em resolver problemas que envolvem variação proporcional, seja direta ou inversa, entre grandezas. Essa competência é crucial para a interpretação de relações proporcionais em diferentes contextos matemáticos e científicos.

A identificação dessas áreas de dificuldade destaca a necessidade de intervenções educativas específicas. Abordar esses pontos críticos é crucial para elevar o desempenho dos alunos, fornecendo suporte adicional, prática direcionada e reforço dos conceitos fundamentais.

Podemos associar cada descritor com a Matriz de referência do SAEB, apresentada no Quadro 2. Com essas informações levantadas, poderemos fazer a comparação com os descritores mais frequentes apontados por Lima (2023) em sua analise dos descritores do SPAECE, essa comparação acontecerá na Seção 4.2.

Visto que os descritores com maior frequência entre os erros dos alunos foram apresentados, agora podemos partir para apresentação dos resultados trazidos pelo algoritmo FP-growth após processar completamente toda a lista de descritores de erros do SAEB. Para que seja possível a compreensão dos dados, apresento a Tabela 3, que traz as principais associações entre os descritores de erros no SAEB, além da relação entre descritor antecedente e consequente, são apresentados a porcentagem de confiança e suporte dessas relações:

Tabela 3 – Associações mais frequentes entre descritores do SAEB para Matemática (erros) no 9º ano do ensino fundamental.

Confiança (%)	Suporte (%)	Antecedente/s	Consequente/s
76,96	24,78	D6	D10
78,78	16,05	D33	D10
75,97	16,04	D14,D20	D10
79,96	15,86	D6,D20	D10
80,55	15,58	D6,D14	D10

Fonte: Elaborada pelo autor.

De acordo com a Tabela 2, os descritores "D20"e "D10"são os que apresentam as maiores frequências de erros cometidos pelos alunos em Matemática. Comparando essa informação com os dados da Tabela 3, podemos perceber que eles aparecem nas cinco associações mais frequentes dos descritores de erros do SAEB.

O decritor "D10"se destaca como consequente nas cinco primeiras associações. Através da análise da Tabela 3 com o Quadro 2, conseguimos levantar algumas associações que merecem destaque, como por exemplo a primeira associação que acontece entre o "D6"e "D10". Pois o "D6"aborda o assunto "reconhecer ângulos como mudança de direção ou giros, identificando ângulos retos e não- retos"e "D10"o assunto "utilizar relações métricas do triângulo retângulo para resolver problemas significativos". Podemos identificar uma certa relação entre esses assuntos, já que a compreensão dos ângulos, especialmente ângulos retos, é fundamental para a aplicação eficaz das relações métricas em triângulos retângulos. No contexto dessas relações, como o Teorema de Pitágoras, a identificação correta dos ângulos retos é crucial para determinar as propriedades métricas dos triângulos. Se um aluno tem dificuldade em reconhecer ângulos e distinguir entre ângulos retos e não retos, isso pode levar a interpretações imprecisas dos problemas que envolvem triângulos retângulos. O uso inadequado de relações métricas devido a uma compreensão limitada dos conceitos de ângulos pode resultar em erros na resolução de problemas.

Na segunda associação, temos o "D33"com o "D10", porém esses dois tópicos não estão diretamente relacionado, uma vez que o "D33"aborda o assunto de "Identificar uma equação ou inequação do 1º grau que expressa um problema". Na terceira associação temos o "D14"juntamente com o "D20"como antecedentes do "D10". O "D14"fala sobre "resolver problema envolvendo noções de volume"que também não está diretamente ligado ao tema do "D10". Já o "D20"merece um destaque, pois aborda o assunto "resolver problema com números inteiros envolvendo as operações (adição, subtração, multiplicação, divisão, potenciação)"que é

um assunto fundamental para que o aluno possa acertas questões dos mais variados assuntos.

Na quarta associação mais frequente, temos o "D6"e "D20"como antecedentes do "D10"e como já foi ressaltado nos paragráfos anteriores, existe uma relação entre esses descritores que merece destaque. Na quinta e última das associações mais frequentes, temos o "D6"juntamente com o "D14"como antecedentes do "D10"que também são descritores que já apresentaram relação com o "D10"na primeira e na terceira associação.

O SPAECE e o SAEB são avaliações educacionais que, embora compartilhem o objetivo de medir o desempenho dos estudantes, apresentam diferenças significativas em suas matrizes de referência e características metodológicas. Portanto, é esperado que os resultados dessas avaliações possam divergir em certa medida, especialmente se houver variações nos conteúdos abordados e nos critérios de avaliação.

Com os resultados da análise dos descritores do SAEB feita, além do levantamento sobre quais os descritores de erro mais frequentes, podemos chegar a última fase da presente pesquisa, que é a comparação desses resultados com os resultados obtidos por Lima (2023) em seu estudo sobre o SPAECE.

4.2 Comparação com os resultados de Lima (2023)

Após o levantamento dos descritores de erros mais frequêntes do SAEB na disciplina de Matemática do 9º ano do ensino fundamental no ano de 2019 e do levantamento das associações mais frequêntes entre esses erros, agora podemos comparar esses resultados com os resultados de Lima (2023), onde foi feita o levantamento dos dados do SPAECE. Para que essa comparação seja possível vamos apresentar primeiro os resultados de Lima (2023):

Tabela 4 – Associações mais frequentes entre descritores do SPAECE para Matemática (erros) no 9º ano do ensino fundamental.

Confiança	Suporte	Antecedente/s	Consequente/s
83,65	48,81	D021	D013
81,99	47,84	D021	D069
89,47	46,19	D015	D067
81,73	42,71	D011	D067
81,66	42,69	D049	D024

Fonte: Lima (2023)

De acordo com Lima (2023), o descritor "D021"é o antecedente de erro mais frequênte. Ele é antecedente do descritor "D013"e também do "D069". O "D021", segundo a

matriz de referência do SPAECE, é sobre "efetuar cálculos com números irracionais, utilizando suas propriedades", porém na matriz de referência do SAEB não aborda assuntos sobre números irracionais. Já o "D013"do SPAECE fala sobre "reconhecer diferentes representações de um mesmo número racional, em situação-problema"e ele equivale ao "D21"do SAEB, no SAEB ele também é um dos erros mais frequêntes entre os alunos (5.317 vezes). O "D069"do SPAECE também possui um descritor equivalente no SAEB que é o "D14", ambos abordam o assunto "resolver problema envolvendo noções de volume", no SAEB ele também é um erro muito frequênte (5.517).

Continuando a análise da Tabela 4, temos o "D015"como antecedente do "D067", o "D015"aborda o assunto "Resolver problema utilizando a adição ou subtração com números racionais representados na forma fracionária (mesmo denominador ou denominadores diferentes) ou na forma decimal". Na matriz de referência do SAEB ele equivale ao "D26"que por sua vez é um descritor que causa muitos erros entre os alunos (4.017). O "D067"do SPAECE equivale ao "D13"do SAEB, eles abordam o tema "resolver problema envolvendo o cálculo de área de figuras planas", no SAEB esse assunto é motivo de muitos erros pelos alunos também (7.304). O que podemos destacar é que esse assunto que fala sobre o cálculo de área de figuras planas é um ponto de dificuldade para os alunos apresentado nos dados de ambas avaliações.

Como assunto que aborda operações com frações foi destaque no trabalho de Lima (2023), sendo ele um dos principais causadores de erros entre os alunos, também podemos destacar algumas associações nos dados do SAEB que envolvem esse assunto como antecedente, destacando os descritores ’D21’, ’D22’, ’D23’, ’D24’ e ’D25’ que estão todos relacionados com esse tema:

Tabela 5 – Associações que envolvem o assunto de frações.

Confiança	Suporte	Antecedente/s	Consequente/s
80,31	13,06	D6,D23	D10
78,00	12,46	D23,D14	D10
76,38	12,08	D23,D3	D20
76,02	11,77	D24,D4,D29	D20
76,57	11,58	D22,D29	D20
75,72	10,67	D22,D3	D20
76,74	10,37	D22,D12	D20
76,41	10,28	D23,D24	D20
78,99	10,06	D37,D24	D20

Fonte: Elaborado pelo autor.

Como foi mostrado na Tabela 5, esse assuntos aparecem acompanhados de outros

descritores como antecedentes. Mas vale ressaltar a recorrencia desses assuntos, visto que foram assuntos com muito destaque também no SPAECE.

5 CONCLUSÕES E TRABALHOS FUTUROS

Neste estudo, propomos a utilização dos microdados do SAEB, uma avaliação em larga escala realizada a cada dois anos nas escolas do Brasil. Esses dados têm o potencial de serem estrategicamente explorados para contribuir no planejamento e desenvolvimento de ações de formação de professores. A análise dos resultados dessas avaliações permite identificar padrões, lacunas de conhecimento e áreas de melhoria que direcionam programas de capacitação docente. O estudo de Lima (2023) sobre os dados da avaliação em larga escala do SPAECE no Ceará inspirou esta pesquisa, apresentando uma nova abordagem analítica dos resultados por meio da Ciência de Dados (CD). Traçamos como foco investigar se os dados do SAEB revelam dificuldades semelhantes às observadas no SPAECE para a disciplina de Matemática no 9º ano do ensino fundamental no município de Fortaleza no ano de 2019.

Durante o desenvolvimento deste estudo, buscamos alcançar os objetivos estabelecidos para a pesquisa. O objetivo geral direcionou nossos esforços para comparar os resultados das relações entre descritores na prova em larga escala SPAECE, especificamente na disciplina de Matemática para o 9º ano do ensino fundamental da rede pública do município de Fortaleza em 2019, com os resultados obtidos pelo SAEB. O propósito era avaliar a consistência das avaliações educacionais, proporcionando insights valiosos para possíveis melhorias na qualidade da educação pública. Esse objetivo foi alcançado, como destacamos no Subseção 4.2, encontramos uma semelhança consideravel entre as duas avaliações, sobre os assuntos relacionados a operações com frações, que demonstrou ser um assunto de grande dificuldade pelos os alunos em ambas avaliações.

Os objetivos específicos direcionaram os passos para alcançar o objetivo geral. Dedicamos esforços para compreender os descritores da avaliação em larga escala SAEB. Em seguida, adquirimos um conhecimento aprofundado sobre as características, entradas, funcionamento e saídas do algoritmo FP-Growth, uma ferramenta essencial para nossa abordagem analítica. Ao aplicar efetivamente esse algoritmo aos microdados do SAEB 2019, conseguimos alcançar o terceiro objetivo específico que era analisar a saída do FP-Growth para os descritores da disciplina de Matemática, identificando as associações mais frequentes entre os descritores de erros dos alunos, todas essas associações foram destacadas na Subseção 3.4.3 que fala sobre os resultados, onde destacamos os principais assuntos que causam erros dos alunos na Tabela 2 de frequencia de erros e também com a Tabela 3 que traz as principais associações entre os descritores.

Ao final deste estudo, podemos afirmar que os objetivos traçados foram cumpridos com sucesso, proporcionando uma análise abrangente e significativa das avaliações em larga escala. Os resultados não apenas contribuem para a consistência das avaliações educacionais, mas também oferecem *insights* valiosos para aprimorar a qualidade da educação pública em Fortaleza. Embora as provas de matemática do SPAECE e SAEB tenham estruturas diferentes, os resultados apresentam convergência na indicação das dificuldades dos alunos do 9º ano do ensino fundamental, principalmente em relação às operações com números racionais.

Sobre trabalhos futuros, considerando que este estudo se concentrou na disciplina de Matemática no 9º ano do ensino fundamental no ano de 2019, podemos sugerir que pesquisas subsequentes estendam essa análise para anos posteriores ao de 2019, permitindo uma compreensão longitudinal das tendências e mudanças ao longo do tempo. Além disso, para uma compreensão mais abrangente do panorama educacional, seria altamente benéfico estender essas análises para outras disciplinas fundamentais, como a de Português. A análise das avaliações em larga escala em disciplinas variadas oferecerá uma visão mais abrangente das lacunas de conhecimento e áreas de melhoria, permitindo um planejamento educacional mais preciso e eficaz.

Essas sugestões para trabalhos futuros não apenas consolidam os resultados deste estudo, mas também destacam a importância contínua de investigações voltadas para o aprimoramento constante do sistema educacional. Ao conduzir análises semelhantes para anos subsequentes e expandir a abordagem para incluir diferentes disciplinas, podemos continuar a construir uma base sólida para orientar estratégias educacionais mais eficazes e abrangentes no futuro.

REFERÊNCIAS

ALJEHANE, N. Big data analytics: challenges and opportunities. In: IEEE. **2020 international conference on computing and information technology (ICCIT-1441)**. [S.l.], 2020. p. 1–4.

AMORIM, T. Conceitos, técnicas, ferramentas e aplicações de mineração de dados para gerar conhecimento a partir de bases de dados. **Universidade Federal de Pernambuco**, 2006.

BOYD, D.; CRAWFORD, K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. **Information, communication & society**, Taylor & Francis, v. 15, n. 5, p. 662–679, 2012.

BÜHLMANN, P.; GEER, S. van de. **Statistics for high-dimensional data: methods, theory and applications**. [S.l.]: Springer Science & Business Media, 2011.

CASTRO, M. H. G. d. **Avaliação educacional: múltiplas abordagens**. São Paulo: Editora Senac São Paulo, 2006.

CIELEN, D.; MEYSMAN, A. D. B.; ALI, M. **Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools**. Shelter Island: Manning, 2016.

CODED/CED, S.-C. **Histórico do SPAECE**. 2023. <https://www.ced.seduc.ce.gov.br/>. Acesso em: 09 maio 2023.

COTTA, T. C. Avaliação educacional e políticas públicas: a experiência do sistema nacional de avaliação da educação básica (saeb). **Revista do Serviço Público**, v. 52, n. 4, p. p. 89–111, fev.
2014. Disponível em: <https://revista.enap.gov.br/index.php/RSP/article/view/316>.

CRESWELL, J. W. **Projeto de pesquisa: métodos qualitativo, quantitativo e misto**. 3. ed.Porto Alegre: Artmed, 2010. 296 p.

CéSAR, P. **Conheça as Etapas do Pré-Processamento de dados**. 2019. Disponível em: https://www.datageeks.com.br/pre-processamento-de-dados/.

FERREIRA, R. V.; OLIVEIRA, V. A. de; ALMEIDA, C. M. Data science aplicada à análise de resultados do enem: estudo de caso em escolas de ensino médio do nordeste brasileiro. **Revista de Informática Teórica e Aplicada**, Sociedade Brasileira de Computação, v. 27, n. 1, p. 119–137, 2020.

GOLDSCHMIDT, R.; BEZERRA, E. **Exemplos de aplicações de Data Mining no mercado brasileiro**. 2023. Disponível em: https://itforum.com.br/noticias/exemplosdeaplicacoes-de-data-mining-no-mercado-brasileiro/. Acessado em 10 de maio de 2023.

GRUS, J. **Data science do zero**. [S.l.: s.n.], 2016.

HOFFMANN, J. M. L. **AVALIAÇÃO MEDIADORA: UMA PRÁTICA EM CONSTRUÇÃO DA PRÉ-ESCOLA À UNIVERSIDADE**. [S.l.]: Mediação, 2011.

INEP, I. N. d. E. e. P. E. A. T. **Sistema de Avaliação da Educação Básica**. 2023. Disponível em: http://portal.inep.gov.br/saeb. Acesso em: 09 maio 2023.

KHAN, N.; AHMAD, S. A systematic review of data science. **Journal of Big Data**, Springer, v. 3, n. 1, p. 1–31, 2016.

LIMA, M. V. d. A. **(Re)estruturação do modelo de planejamento das ações de formação continuada em serviço de professores na rede municipal de Fortaleza: novas possibilidades por meio da Ciência de Dados Educacionais**. 315 p. Tese (Doutorado) — Universidade Estadual do Ceará, Fortaleza, 2023. Tese (Doutorado em Educação) – Programa de Pós-Graduação em Educação.

LOCATELLI, I. Construção de instrumentos para a avaliação de larga escala e indicadores de rendimento: o modelo saeb. **Estudos em Avaliação Educacional**, n. 25, p. 3–21, jun. 2002. Disponível em: <https://publicacoes.fcc.org.br/eae/article/view/2189>.

MAGALHÃES, A. G. J.; FARIAS, M. A. de. Spaece: Uma história em sintonia com avaliação educacional do governo federal. **Revista de Humanidades**, v. 31, n. 2, p. 525–547, 2016.

MARIANO, M. A. Comparação de algoritmos paralelos para a extração de regras de associação no modelo de memória distribuída. 2011.

Ministério da Educação. **Portal do Ministério da Educação - SAEB**. 2023. Acesso em: 25 novembro 2023. Disponível em: <http://portal.mec.gov.br/component/tags/tag/saeb#:~:text=Provas%20%E2%80%94%20Cada%20prova%20do%20Saeb,a%2026%20de%20cada%20disciplina>.

MOLINA-MARKHAM, A.; CHOUDHURY, O.; DIAKOPOULOS, N.; MATTU, S. Privacy and security in data science: A review of methods and practices. **arXiv preprint arXiv:1908.10948**, 2019.

PERRENOUD, P.; RAMOS, P. **Avaliação: da excelência à regulação das aprendizagens : entre duas lógicas**. Artmed, 1999. ISBN 9788573075441. Disponível em: <https: // books.google.com.br/books?id=tRntAAAACAAJ>.

ROSSI-BARBOSA, L. A. R. A importância das pesquisas científicas na graduação. **Renome**, v. 2, n. esp, 2013.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**. [S.l.]: Elsevier Brasil, 2017.

WU, X.; ZHANG, X. **Data mining with big data**. [S.l.]: Springer, 2014.

APÊNDICE A – SCRIPT PARA OBTER LISTA DE DESCRITORES DE ERROS

```
def encontrar_descritores(row, items):
    try:
        desc_erro = []

        if not pd.isnull(row["TX_RESP_BLOCO1_MT"]):
            res_bloco1 = list(row["TX_RESP_BLOCO1_MT"])
            cond_aux1 = (
                (items["BLOCO"] == row["ID_BLOCO_1_MT"])
                & (items["POSICAO"].isin(range(1, len(
                    res_bloco1) + 1)))
            )
            item1 = items[cond_aux1]

            for pos, resp_aluno in enumerate(res_bloco1):
                if resp_aluno not in ['*', '.', ',', ' ']:
                    desc_erro.extend(
                        item1.loc[(item1["GABARITO"] !=
                            resp_aluno) & (item1["POSICAO"]
                            == pos + 1), "
                            NU_DESCRITOR_HABILIDADE"]
                        .tolist()
                    )

        if not pd.isnull(row["TX_RESP_BLOCO2_MT"]):
            res_bloco2 = list(row["TX_RESP_BLOCO2_MT"])
            cond_aux2 = (
                (items["BLOCO"] == row["ID_BLOCO_2_MT"])
                & (items["POSICAO"].isin(range(1, len(
                    res_bloco2) + 1)))
            )
```

```
26              item2 = items[cond_aux2]
27
28              for pos, resp_aluno in enumerate(res_bloco2):
29                  if resp_aluno not in ['*', '.', ',', ' ']:
30                      desc_erro.extend(
31                          item2.loc[(item2["GABARITO"] !=
                                resp_aluno) & (item2["POSICAO"]
                                == pos + 1), "
                                NU_DESCRITOR_HABILIDADE"]
32                          .tolist()
33                      )
34
35          return pd.Series([row["ID_ALUNO"], ','.join(map(str
                , desc_erro))], index=["id_aluno", "
                conj_desc_erro"])
36
37      except Exception as e:
38          print(f"Erro na linha {row.name}: {e}")
39          return pd.Series([row["ID_ALUNO"], ''], index=["
                id_aluno", "conj_desc_erro"])
40
41  desc_erros = alunos.apply(encontrar_descritores, items=
        items, axis=1)
```