UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Análise de dados do ENEM baseada em data warehousing, mineração de dados, estatística inferencial e processamento paralelo e distribuído

Viviana Elizabeth Romero Noguera

Tese de Doutorado do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional (PPG-CCMC)

ICMC USP
SÃO CARLOS

Viviana Elizabeth Romero Noguera

Análise de dados do ENEM baseada em data warehousing, mineração de dados, estatística inferencial e processamento paralelo e distribuído

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Doutora em Ciências – Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientadora: Profa. Dra. Cristina Dutra de Aguiar

USP – São Carlos
Fevereiro de 2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

R763a

Romero Noguera, Viviana Elizabeth
Análise de dados do ENEM baseada em data warehousing, mineração de dados, estatística inferencial e processamento paralelo e distribuído / Viviana Elizabeth Romero Noguera; orientador Cristina Dutra de Aguiar. -- São Carlos, 2023.
155 p.

Tese (Doutorado - Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. ENEM. 2. Análise de dados. 3. Data warehousing. 4. Mineração de dados. 5. Processamento paralelo e distribuído. I. de Aguiar, Cristina Dutra, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Viviana Elizabeth Romero Noguera

ENEM data analysis based on data warehousing, data mining, inferential statistics and parallel and distributed processing

Thesis submitted to the Instituto de Ciências Matemáticas e de Computação – ICMC-USP – in accordance with the requirements of the Computer and Mathematical Sciences Graduate Program, for the degree of Doctor in Science. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Cristina Dutra de Aguiar

USP – São Carlos
February 2023

Dedico este trabalho a minha prima Sandra Noguera (in memorian), que foi como uma irmã para mim.

AGRADECIMENTOS

A Deus por me abençoar com a vida, força, fé e saúde para que eu pudesse concluir mais esta importante etapa da minha vida.

A minha orientadora, a Profa. Dra. Cristina Dutra de Aguiar, pela orientação, pelos conhecimentos compartilhados, por ter compreendido minhas dificuldades e ajudado a realizar este trabalho. A Profa. Dra. Kalinka Regina Lucas Jaquie Castelo Branco, pelo aprendizado, discussões, dicas e suporte durante meu Doutorado.

A meus pais, Eulalia e Ángel pelo apoio incondicional. Também por todo o sacrifício, em todos os aspectos, que fizeram para que eu pudesse alcançar meus objetivos.

A minha madrinha, Graciela, pela força e carinho em todo este tempo.

A meus irmãos, Sofía e Mario, por acreditarem que este grande sonho um dia se realizaria.

Ao Programa Nacional De Becas De Postgrado en el Exterior “Don Carlos Antonio López” (BECAL) e a CAPES, pelo auxílio financeiro que tornou possível o desenvolvimento deste trabalho.

Aos professores do Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional do ICMC, que de alguma forma contribuíram para minha formação.

Aos meus amigos e colegas de doutorado os quais compartilhei alegrias e dificuldades deste percurso.

Enfim, agradeço a todos que contribuíram de forma direta ou indiretamente para a obtenção do título de Doutora em Ciências.

“Conhecimento não é aquilo que você sabe,
mas o que você faz com aquilo que você sabe.”
(Aldous Huxley)

RESUMO

NOGUERA, V. E. R. **Análise de dados do ENEM baseada em data warehousing, mineração de dados, estatística inferencial e processamento paralelo e distribuído**. 2023. 155 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

Em 1998, o Ministério da Educação criou o ENEM, um exame nacional brasileiro padronizado que visa avaliar as competências e habilidades básicas dos alunos. O ENEM é uma avaliação que tem forte influência nas políticas educacionais, nos currículos dos diferentes níveis de ensino e também no futuro dos avaliados. Considerando o escopo da educação, setor fundamental para o crescimento e desenvolvimento de um país, a análise dos dados do ENEM pode revelar informações relevantes para subsidiar a tomada de decisão das instituições de ensino, a definição de investimentos governamentais e planos estratégicos e a formulação de políticas públicas de educação com base no desempenho cognitivo dos alunos. A análise dos dados do ENEM é uma questão desafiadora devido ao grande número de variáveis coletadas e ao grande volume de dados disponibilizados. Essas variáveis referem-se aos dados pessoais e às notas obtidas por cada participante, às respostas das questões de múltipla escolha e às respostas dos questionários. Com o objetivo de auxiliar os gestores educacionais no processo de tomada de decisão educacional, nesta tese é proposta uma arquitetura baseada em *data warehousing, mineração de dados, estatística inferencial e processamento paralelo e distribuído* voltada à análise de dados do ENEM. A arquitetura é composta por cinco camadas: (i) conexão de dados, relacionada com a extração e tratamento dos dados do ENEM; (ii) gerenciamento de dados, voltada ao armazenamento dos dados e metadados relacionados em repositórios especializados, de acordo com as necessidades das análises educacionais; (iii) análise de dados, que tem como objetivo extrair informações úteis e auxiliar na tomada de decisão estratégica; (iv) apresentação de dados, composta por ferramentas de visualização que permitem que cientistas de dados e gestores educacionais visualizem graficamente os resultados de suas análises; e (v) gerenciador de fluxo de trabalho, voltada à automação das tarefas complexas que são executadas na manipulação do grande volume de dados do ENEM. Adicionalmente, são apresentados dois *pipelines*, os quais exemplificam a instanciação da arquitetura proposta com tecnologias e ferramentas de código aberto relacionadas. A arquitetura foi validada por meio de quatro cenários de uso, cada qual com um objetivo de análise específico. Para cada cenário de uso, foi feita uma discussão relacionada aos impactos das análises dentro do contexto educacional. Os resultados demonstraram a aplicabilidade da arquitetura no suporte ao processo da tomada de decisão educacional.

Palavras-chave: ENEM, análise de dados, data warehousing, mineração de dados, processamento paralelo e distribuído.

ABSTRACT

NOGUERA, V. E. R. **ENEM data analysis based on data warehousing, data mining, inferential statistics and parallel and distributed processing**. 2023. 155 p. Tese (Doutorado em Ciências – Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2023.

In 1998, the Ministry of Education created the Brazilian High School National Exam or ENEM, a standardized Brazilian national exam to assess students' essential competencies and skills. The ENEM is an assessment that strongly influences educational policies, the curricula of different levels of education, and the future of the students assessed. Considering the scope of education, an important issue related to the country's growth and development, analyzing the ENEM data can reveal relevant information. For instance, the analyses can support educational decision-making, the definition of government investments and strategic plans, and the formulation of public education policies based on the student's cognitive performance. The analysis of ENEM data is challenging due to the large number of variables collected and the large volume of data available. These variables refer to personal data and the scores obtained by each participant, the answers to the multiple-choice questions, and the answers to the questionnaires. To help educational managers in the educational decision-making process, in this thesis, we propose an architecture based on data warehousing, data mining, inferential statistics, and parallel and distributed processing aimed at analyzing ENEM data. The architecture is composed of five layers: (i) Data connection, related to the extraction and processing of ENEM data; (ii) Data management, aimed at storing data and related metadata in specialized repositories according to the needs of educational analyzes; (iii) Data analysis, which aims to extract useful information and assist in strategic decision-making; (iv) Data presentation, consisting of data visualization tools that allow data scientists and educational managers to graphically visualize the results of their analyses; and (v) Workflow manager, aimed at automating the complex tasks that are performed in the manipulation of the large volume of ENEM data. Additionally, we present two pipelines that exemplify the proposed architecture's instantiation with related open-source technologies and tools. We validated the architecture through four different scenarios, each with a specific analysis objective. For each scenario, we discuss the impact of the analysis on the educational sector. The results demonstrated the architecture's applicability in supporting the educational decision-making process.

Keywords: ENEM, data analysis, data warehousing, data mining, parallel and distributed processing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANN	*Artificial Neural Networks*
AUC	*Area Under Curve*
CRM	Sistemas de gerenciamento de relacionamento com clientes
DAGs	*Directed Acyclic Graphs*
DW	*Data warehouse*
ELT	*Extract, Load and Transform*
ENEM	Exame Nacional de Ensino Médio
ERP	Sistemas integrados de gestão empresarial
ETL	*Extract, Transform and Load*
HDFS	*Hadoop Distributed File System*
HQL	HiveSQL
IBGE	Instituto Brasileiro de Geografia e Estatística
IDEB	Índice de Desenvolvimento da Educação Básica
Inep	Instituto Nacional de Estudos e Pesquisas Educacionais
LGPD	Lei Geral de Proteção de Dados
ML3P	Modelo Logístico de três Parâmetros
MLP	*MultiLayer Perceptron*
NoSQL	*Not Only SQL*
OLAP	*On-Line Analytical Processing*
OLAP	*On-Line Analytical Processing*
ONU	Organização das Nações Unidas
ORC	Optimized Row Columnar
PICO	População, Intervenção, Comparação e Resultados
PISA	Programa Internacional de Avaliação de Estudantes
Prouni	Programa Universidade para Todos
RDD	Dados Distribuídos e Resilientes
REUNI	Programa de Apoio aos Planos de Reestruturação e Expansão das Universidades Federais
ROC	*Receiver Operating Characteristic*
SAEB	Sistema de Avaliação da Educação Básica
Sedap	Serviço de Acesso a Dados Protegidos

SHAP	*SHapley Additive exPlanations*
Sisu	Sistema de Seleção Unificada
SQL	*Structured Query Language*
StArt	State of the Art through Systematic Review
STEM	Ciência, tecnologia, engenharia e matemática
SVM	*Support Vector Machines*
TRI	Teoria de Resposta ao Item
UNESCO	Organização das Nações Unidas para a Educação, a Ciência e a Cultura

SUMÁRIO

CAPÍTULO

1

INTRODUÇÃO

A educação é um setor fundamental para o crescimento e desenvolvimento de um país. A avaliação atual no Brasil indica a necessidade de melhorias neste setor, as quais podem ser alcançadas com o auxílio de Tecnologia da Informação (SILVA; MORINO; SATO, 2015). Um dos principais objetivos da avaliação educacional é garantir a qualidade do ensino. Neste contexto, entende-se avaliação como "um juízo de qualidade sobre dados relevantes, tendo em vista a tomada de decisão" (LUCKESI, 2014).

Em 1998, o Ministério da Educação criou o Exame Nacional do Ensino Médio (ENEM), um exame nacional brasileiro padronizado que visa avaliar competências e habilidades básicas dos alunos. O ENEM é o maior exame do Brasil e o segundo maior vestibular do mundo, apenas atrás do exame chinês Gaokao, o maior vestibular do mundo em número de participantes (SILVEIRA; MAUÁ, 2018). A primeira edição do exame reuniu 115.575 participantes (INEP, 2022a). De 2016 a 2020, o número de participantes do ENEM foi de 8.627.367, 6.731.341, 5.513.747, 5.095.270 e 5.783.109, respectivamente, incluindo treineiros e não treineiros. Um não treineiro é um participante que já concluiu o ensino médio ou que irá completá-lo no mesmo ano da realização do exame. Nestes anos, cerca de 90% dos participantes eram não treineiros.

O número de participantes do ENEM tem crescido muito ao longo do tempo devido às seguintes iniciativas importantes do governo brasileiro: o Programa Universidade para Todos (Prouni) e o Programa de Apoio aos Planos de Reestruturação e Expansão das Universidades Federais (REUNI). O Prouni foi criado em 2004 pelo Ministério da Educação para utilizar a nota do ENEM como base para distribuição de bolsas em universidades privadas (IBGE, 2021). O REUNI foi criado em 2007 para ampliar o acesso e retenção de alunos nas Universidades e Institutos Federais. O REUNI foi responsável por ampliar o número de universidades no Brasil, criando diversos campi pelos estados (PAULA, 2017).

Em muitas instituições de ensino superior públicas e privadas, a pontuação do ENEM é usada como exame de admissão para matrícula nos cursos. Em outras instituições, os resultados

do ENEM são usados para complementar o processo de admissão próprio. Além disso, os programas Sistema de Seleção Unificada (Sisu) e Prouni também utilizam as notas do ENEM como critério de seleção para ingresso nos cursos dentro das vagas disponibilizadas aos participantes do ENEM e oferecidas pelas universidades, assim como para fornecer bolsas de estudo. Atualmente, todas as instituições públicas federais brasileiras preenchem as vagas com base nessas pontuações.

O exame do ENEM contempla 180 questões de múltipla escolha e uma produção de texto (ou seja, *Redação*). As questões são embasadas por uma matriz de referência de eixos cognitivos, que avalia competências nas áreas de conhecimento de: (i) Linguagens, Códigos e suas Tecnologias; (ii) Matemática e suas Tecnologias; (iii) Ciências da Natureza e suas Tecnologias; e (iv) Ciências Humanas e suas Tecnologias. Nesta tese, essas áreas de conhecimento são nomeadas como: (i) *Linguagens e Códigos*; (ii) *Matemática*; (iii) *Ciência da Natureza*; e (iv) *Ciências Humanas*, respectivamente. A Tabela 1 detalha os componentes curriculares de cada área de conhecimento. Em cada área de conhecimento e na *Redação*, a pontuação máxima que pode ser atingida é 1.000 pontos.

Tabela 1 – Descrição das áreas de conhecimento e componentes curriculares do ENEM.

Área do conhecimento	Componentes curriculares
Linguagens, Códigos e suas tecnologias	Língua Portuguesa, Literatura, Língua Estrangeira (Inglês ou Espanhol), Artes, Educação Física e Tecnologias da Informação e Comunicação.
Ciências Humanas e suas tecnologias	História, Geografia, Filosofia e Sociologia.
Ciências da Natureza e suas tecnologias	Química, Física e Biologia.
Matemática e suas tecnologias	Matemática.

Fonte: Inep (2019).

Anualmente, são produzidos quatro tipos diferentes de caderno, um para cada área de conhecimento. Cada tipo de caderno é identificado por cores diferentes, sendo que cada cor contém as mesmas questões, porém organizadas em ordens diferentes. Cada tipo de caderno e cada cor recebe um código de identificação único. Essa diversidade dos tipos de caderno tem como objetivo dificultar fraudes durante a aplicação do exame.

Além das questões objetivas e da produção de texto, os participantes respondem a um questionário que contempla questões relacionadas: (i) aos dados pessoais, como sexo e idade; (ii) ao nível socioeconômico, como acesso à internet, se o participante possui computador e celular e se existe televisor e aparelho de microondas em sua residência; (iii) aos aspectos familiares, como formação e ocupação dos pais, quantidade de pessoas que moram na residência e renda mensal; (iv) aos aspectos educacionais, como dados da escola na qual o participante cursou o ensino médio, intenção de recorrer a auxílios financeiros caso ingresse no ensino superior e motivos para a sua participação no ENEM; e (v) ao trabalho, indicando se o participante exerce atividade remunerada e qual a quantidade de horas trabalhadas. A quantidade de questões presentes no questionário varia a cada ano de realização do exame do ENEM.

Adicionalmente, para cada ano de realização do exame do ENEM, é disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais (Inep) um arquivo no formato .csv, chamado de arquivo de microdados. Cada arquivo contém, para cada participante do ENEM em um determinado ano, a resposta dada pelo participante a cada questão do tipo de caderno, o gabarito de cada questão do tipo de caderno e as notas das questões objetivas e da *Redação* do participante, além das respostas presentes em seu questionário. Os microdados são referentes a até 166 variáveis, dependendo do ano.

As respostas e seus respectivos gabaritos não são armazenados individualmente. Ao contrário, para cada tipo de caderno, existe uma variável do tipo *string* que contém a resposta dada pelo participante a cada questão desse tipo de caderno e uma variável do tipo *string* que contém o gabarito de cada questão. O número total de questões de cada exame do ENEM, 180, é obtido somando-se o número de questões presentes na *string* de resposta de cada tipo de caderno. Os participantes não estão identificados pelos seus nomes. Para cada participante, é atribuído um número de inscrição, garantindo seu anonimato.

1.1 Motivação

Todos os dados coletados pelo ENEM estão disponíveis gratuitamente para análise. Eles são fontes valiosas a serem investigadas. Considerando o escopo da educação, a análise desses dados pode revelar informações relevantes para subsidiar a tomada de decisão das instituições de ensino, a definição de investimentos governamentais e planos estratégicos e a formulação de políticas públicas de educação com base no desempenho cognitivo dos alunos.

Três fatores determinam o desempenho cognitivo dos alunos: fatores associados à estrutura escolar, fatores relacionados à família e fatores que dizem respeito ao próprio aluno. Indicadores educacionais são comumente utilizados para investigar o desempenho cognitivo e, indiretamente, a qualidade da educação escolar. Esses indicadores associam uma nota de desempenho aos alunos considerando os êxitos do processo educacional. Por exemplo, o Índice de Desenvolvimento da Educação Básica (IDEB) avalia semestralmente a educação básica em seus anos iniciais e finais tomando como base as taxas de aprovação dos alunos obtidas no Censo Escolar e as médias em Leitura e Matemática das duas avaliações realizadas pelo Inep e pelo Sistema de Avaliação da Educação Básica (SAEB) (SOARES; ALVES, 2013). Outra prova aplicada a cada três anos, em nível internacional, é o Programa Internacional de Avaliação de Estudantes (PISA) (PISA, 2022), que mede a capacidade dos alunos de usar os seus conhecimentos e habilidades de leitura, matemática e ciências para enfrentar os desafios da vida real.

Analisar os dados do ENEM com o objetivo de investigar o desempenho dos alunos não é uma tarefa trivial de ser realizada, devido à grande quantidade de variáveis consideradas em seus arquivos de microdados e o grande volume de dados disponibilizados. Por um lado, as variáveis podem ser fontes valiosas para se determinar diferentes fatores de análise a serem

investigados dentro do âmbito da educação. Por outro lado, o grande volume de dados não está relacionado apenas à quantidade de participantes inscritos anualmente no exame do ENEM, mas à necessidade de análise das variáveis associadas a esses participantes.

Por exemplo, considere as quatro *strings* de respostas dos participantes. Na Tabela 2 é detalhada a quantidade de respostas a serem investigadas, para os anos de 2016 a 2020. O cálculo da quantidade de respostas foi feito multiplicando-se a quantidade de participantes de um determinado ano pelo número total de questões, que é 180. Os cálculos presentes na Tabela 2 referem-se a apenas 4 das até 166 variáveis disponibilizadas por ano. Portanto, o volume de dados a ser considerado nas análises é ainda muito maior, desde que existe a possibilidade de se investigar o desempenho dos participantes considerando as variáveis individualmente e de forma combinada.

Tabela 2 – Quantidade de respostas por ano de realização do exame do ENEM.

Ano	Quantidade de respostas
2016	1.552.926.060
2017	1.211.641.380
2018	992.474.460
2019	917.148.600
2020	1.040.959.620

Fonte: Elaborada pelo autor.

A análise de dados considerando uma grande quantidade de variáveis e um grande volume de dados pode ser beneficiada pela aplicação de técnicas relacionadas à ciência de dados. A ciência de dados é uma área multidisciplinar voltada a resolver desafios de manipulação e gerenciamento desse gigantescos volumes de dados (*Big data*) (SONG; ZHU, 2016). Ela envolve princípios, processos e técnicas para entender fenômenos por meio da análise de dados (PROVOST; FAWCETT, 2013). Segundo Cielen, Meysman e Ali (2016), os processos da ciência de dados geralmente consistem em seis etapas: (i) estabelecimento de um objetivo de pesquisa; (ii) recuperação dos dados; (iii) preparação dos dados; (iv) exploração dos dados, (v) modelagem dos dados; e (v) apresentação e automação das análises.

Técnicas de *data warehousing*, *mineração de dados*, *estatística inferencial* e *processamento paralelo e distribuído* podem auxiliar no processo de análise de dados. Um ambiente de *data warehousing* provê eficiência e flexibilidade na obtenção de informações relevantes e apropriadas para tomada de decisão (CHAUDHARY; MURALA; SRIVASTAV, 2011). Com relação às técnicas de *mineração de dados*, a classificação permite a identificação e classificação de perfis dos participantes e previsão do desempenho dos participantes de acordo com características personalizadas e pessoais, dentre outros (GOMEDE *et al.*, 2018). Já a *estatística inferencial* pode ser aplicada para investigar o desempenho dos participantes por meio de testes de hipóteses. Por fim, existe uma crescente demanda por ambientes computacionais com alto poder de armazenamento e processamento: os ambientes computacionais paralelos e distribuídos. Esses

ambientes são apropriados para gerenciar grandes volumes de dados, como é o caso do volume de dados disponibilizados nos exames do ENEM. Para abstrair do usuário as complexidades de paralelismo inerentes a esses ambientes, pode-se utilizar *framework* de processamento paralelo e distribuído, como o Apache Spark (ZAHARIA *et al.*, 2010).

O desafio para análise dos dados do ENEM é propor uma arquitetura que engloba técnicas de *data warehousing*, *mineração de dados*, *estatística inferencial* e *processamento paralelo e distribuído* e que ofereça suporte eficiente para auxiliar no processo de tomada de decisão educacional, permitindo aos gestores educacionais propor intervenções baseadas em evidências. Com base na revisão sistemática descrita no Capítulo 3, nenhum estudo existente na literatura considera conjuntamente o uso de todas as técnicas supracitadas para investigar a grande quantidade de variáveis consideradas nos arquivos de microdados do ENEM e o grande volume de dados disponibilizados. Essa lacuna na literatura motiva o desenvolvimento desta tese de doutorado.

1.2 Objetivos

Esta tese tem como objetivo principal propor uma arquitetura baseada em *data warehousing*, *mineração de dados*, *estatística inferencial* e *processamento paralelo e distribuído* e que seja voltada à analise de dados educacionais do ENEM. Usando como base esta arquitetura, alcança-se o objetivo secundário, que é realizar diferentes investigações a respeito do desempenho dos participantes do ENEM.

Com base nesses objetivos, define-se a seguinte hipótese:

É possível auxiliar os gestores educacionais no processo de tomada de decisão educacional utilizando técnicas de data warehousing, mineração de dados, estatística inferencial e processamento paralelo e distribuído no que diz respeito à análise de dados do ENEM.

1.3 Contribuições

Inicialmente é feita a proposta de uma arquitetura para apoiar o processo de tomada de decisão educacional. A arquitetura é composta por cinco camadas: (i) conexão de dados, relacionada com a extração e tratamento dos dados do ENEM; (ii) gerenciamento de dados, voltada ao armazenamento dos dados e metadados relacionados em repositórios especializados, de acordo com as necessidades de análises educacionais a serem realizadas; (iii) análise de dados, a qual tem como objetivo extrair informações úteis e auxiliar na tomada de decisão estratégica; (iv) apresentação de dados, composta por ferramentas de visualização de dados que permitem que os cientistas de dados e os gestores educacionais visualizem graficamente os resultados de suas análises; e (v) gerenciador de fluxo de trabalho, voltada à automação das tarefas complexas que são executadas na manipulação do grande volume de dados do ENEM. Adicionalmente, são

apresentados dois *pipelines*, os quais exemplificam a instanciação da arquitetura proposta com tecnologias e ferramentas de código aberto relacionadas.

Na sequência, a arquitetura é utilizada em diferentes cenários de uso, cada qual com um objetivo de análise específico dentro do contexto da tomada de decisão educacional. Os diferentes períodos de tempo nas análises estão relacionados com a disponibilidade dos dados na época dessas análises. As características de cada cenário de uso são descritas a seguir.

- Primeiro cenário: Identifica os principais indicadores de desempenho dos participantes do ENEM nos anos de 2017 a 2020. A análise desses indicadores é realizada por meio da tarefa de classificação da mineração de dados. Todas as áreas de conhecimento do ENEM são consideradas, a saber: *Matemática*, *Ciência da Natureza*, *Linguagens e Códigos* e *Ciências Humanas*.

- Segundo cenário: Investiga o desempenho dos participantes do ENEM na temática "gêneros[1]", considerando as três áreas de conhecimento relacionadas às ciências exatas, a saber: *Matemática*, *Ciência da Natureza* e *Linguagens e Códigos*. A análise de desempenho considera diferentes perspectivas relativas aos anos de 2013 a 2017, de acordo com as variáveis presentes nos microdados.

- Terceiro cenário: Investiga o desempenho dos participantes do ENEM na temática "gêneros", porém considerando questões classificadas como STEM e não-STEM. Questões STEM se referem àquelas pertencentes às áreas de conhecimento de *Matemática* e *Ciência da Natureza* em adição a algumas questões específicas de *Linguagens e Códigos* (questões identificadas pelos códigos de habilidade 28, 29 e 30). As demais questões são classificadas como não-STEM. A análise de desempenho é feita por meio da estatística inferencial sobre os dados de 2016 a 2020.

- Quarto cenário: Propõe um esquema multidimensional para armazenar os dados do ENEM dos anos de 2016 a 2020 de forma que não somente as análises apresentadas nesta tese podem ser realizadas, mas também diversas outras análises que sejam relevantes às instituições de ensino e aos órgãos governamentais. Adicionalmente, também são ilustradas diferentes consultas analíticas que podem ser executadas contra este esquema.

Os resultados preliminares relacionados às contribuições anteriormente descritas geraram os seguintes artigos científicos:

[1] As definições de sexo e gênero podem ser vistas como: o primeiro se refere às categorias inatas do ponto de vista biológico, enquanto o segundo diz respeito aos papéis sociais relacionados com a mulher e o homem, permitindo uma distinção sociológica (MOORE, 1997). Entretanto, nesta tese de doutorado *sexo* e *gênero* são vistos de forma indiscriminada considerando que o ENEM não oferece essa classificação.

- NOGUERA, V.; BRANCO, K.; CIFERRI, C. Gêneros e suas nuances no ENEM. In: **Anais do XIII Women in Information Technology**. Belém, PA, Brasil: SBC, 2019. p. 41–50. O artigo recebeu o prêmio de melhor artigo do evento XIII Women in Information Technology.
- NOGUERA, V. E. R.; BRANCO, K. R. L. J. C.; CIFERRI, C. D. de A. Análise de desempenho das mulheres no ENEM. **Brazilian Journal of Development**, v. 6, n. 6, p. 35716–35737, 2020.

Adicionalmente, o seguinte artigo científico encontra-se em julgamento:

- NOGUERA, V. E. R.; BRANCO, K. R. L. J. C.; AGUIAR, C. D. Identifying and analyzing key performance indicators from the ENEM data to support educational public policies. **Educational Assessment**.

A arquitetura proposta, embora tenha como objetivo prover suporte para a análise dos dados do ENEM, pode ser aplicada a qualquer contexto no qual dados educacionais sejam coletados.

1.4 Organização da Tese

Além deste capítulo introdutório, esta tese é organizada da seguinte forma.

- No Capítulo 2 são descritos conceitos relacionados à ciência de dados, pré-processamento de dados, ambiente de *data warehousing*, processamento paralelo e distribuído, técnicas de mineração de dados, estatística inferencial e a teoria de resposta ao item.
- No Capítulo 3 é detalhada a revisão sistemática, a qual aborda as fases de planejamento, condução e extração de dados.
- No Capítulo 4 é detalhada a arquitetura proposta em termos de suas diferentes camadas e funcionalidades. Também são descritos dois exemplos de instanciação da arquitetura proposta usando tecnologias e ferramentas de código aberto disponíveis no mercado.
- No Capítulo 5 é detalhado o primeiro cenário de uso da arquitetura. São identificados e analisados os principais indicadores de desempenho para avaliar a eficácia educacional e, deste modo, apoiar políticas públicas educacionais.
- No Capítulo 6 é especificado o segundo cenário de uso da arquitetura. Neste capítulo é analisado o desempenho dos participantes do ENEM na temática dos gêneros e suas nuances na tecnologia da informação. Vários fatores de análise são investigados em termos de participantes do sexo feminino e masculino, considerando as ciências exatas.

- No Capítulo 7 é discutido o terceiro cenário de uso da arquitetura. É investigado o desempenho de participantes com base no gênero e seu impacto em questões classificadas como STEM e não-STEM, de maneira a demostrar a existência de diferença estatística entre as proporções de acertos.
- No Capítulo 8 é detalhado o quarto cenário de uso da arquitetura. É desenvolvido um repositório multidimensional de dados, que oferece diferentes perspectivas de análise e facilita aos gestores educacionais a especificação de consultas analíticas importantes para tomada de decisão educacional.
- No Capítulo 9 são descritas as considerações finais desta tese. São listadas as dificuldades encontradas e trabalhos futuros.

CAPÍTULO

2

FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é feita uma descrição da fundamentação teórica utilizada para o desenvolvimento do trabalho. Na seção 2.1 são sumarizados conceitos relacionados à ciência de dados. Na seção 2.2 são detalhadas técnicas de pré-processamento de dados. Na seção 2.3 são descritos aspectos relacionados aos ambientes de *data warehousing*. Na seção 2.4 são contextualizados os principais conceitos relacionados aos ambientes de computação paralela e distribuída. Na seção 2.5 são descritos os algoritmos de classificação e métricas de avaliação utilizados no processo de mineração de dados. Na seção 2.6 são detalhados conceitos de estatística inferencial. Na seção 2.7 é descrita a Teoria de Resposta ao Item, técnica utilizada para a correção das questões do exame do ENEM. O capítulo é finalizado na seção 2.8 com as considerações finais.

2.1 Ciência de dados

A ciência de dados (*data science*) é uma ciência que oferece metodologias para processar e interpretar grandes volumes de dados coletados por um número crescente de novos dispositivos (SONG; ZHU, 2016), desde que realizar tarefas analíticas usando apenas as metodologias estatísticas estabelecidas há muito tempo não é adequado (VICARIO; COLEMAN, 2019). A ciência de dados é definida como um campo interdisciplinar, que inclui métodos matemáticos, estatísticos, desenvolvimento de algoritmos, análise qualitativa e ciência da computação. Ela é caracterizada por ser uma abordagem prática, que tem como objetivo extrair informações úteis dos dados.

Na literatura, existem várias definições sobre ciência de dados. Por exemplo, segundo Dhar (2013), a ciência de dados é o estudo da extração generalizável do conhecimento a partir dos dados. Saltz e Stanton (2017) definem ciência de dados como uma área de trabalho emergente voltada à coleta, preparação, análise, visualização, gerenciamento e preservação de grandes coleções de informações. Provost e Fawcett (2013) afirmam que a ciência de dados envolve princípios, processos e técnicas para entender fenômenos por meio da análise automatizada de

dados. Nesta tese, considera-se que ciência de dados é uma área interdisciplinar voltada à solução de desafios de manipulação de gigantescos volumes de dados, os quais seguem o conceito de *big data* (SONG; ZHU, 2016).

Duas discussões importantes relacionadas à ciência de dados incluem sua diferença com mineração de dados e a necessidade de uso de linguagens de programação apropriadas. A denominação mineração de dados é frequentemente associada à ciência de dados e, às vezes, até confundida com ela, devido à grande quantidade de métodos matemáticos e de ciência da computação usados em ambos processos. No entanto, são bem diferentes (VICARIO; COLEMAN, 2019). Mineração de dados é um processo para descobrir padrões em grandes conjuntos de dados, extrair informações e transformá-las em uma estrutura compreensível para uso posterior, típica para decisões apropriadas. Por outro lado, a ciência de dados engloba um maior número de processos, incluindo, por exemplo, recuperação, preparação e mineração de dados, além da análise final sobre conjuntos de dados de todos os tipos.

Quanto às linguagens de programação, em teoria, toda linguagem de programação suficientemente poderosa é capaz de expressar qualquer algoritmo a ser executado. Mas, na prática, certas linguagens de programação são muito melhores que outras em tarefas específicas, ou seja, são mais eficientes em termos computacionais de acordo com os requisitos e as características das tarefas. As principais linguagens de programação da ciência de dados são *Python, R, SQL, Matlab, Java, Perl, C/C++ e Mathematica/Wolfram Alpha* (SKIENA, 2017; MUELLER; MASSARON, 2019).

2.2 Pré-processamento de dados

Os dados do mundo real tendem a ser sujos, incompletos e inconsistentes. Neste sentido, existem várias técnicas de pré-processamento de dados utilizadas para minimizar esses problemas (HAN; KAMBER; PEI, 2011). Essas técnicas possibilitam que os dados tornem-se mais adequados para que sejam posteriormente utilizados por algoritmos de mineração de dados.

Dentre as técnicas de pré-processamento dos dados, destacam-se:

- **Limpeza de dados:** pode ser aplicada para preencher valores ausentes, suavizar dados ruidosos, identificar discrepâncias e corrigir inconsistências nos dados.

- **Integração de dados:** integra dados heterogêneos obtidos de diferentes fontes de dados, resolvendo problemas de heterogeneidade semântica e estrutural.

- **Redução de dimensionalidade:** obtém uma representação reduzida de um conjunto de dados, de forma que a integridade dos dados originais seja mantida.

- **Balanceamento dos dados:** balanceia artificialmente um conjunto de dados de forma que não seja utilizada uma classe majoritária quando for feita a aplicação de algoritmos de classificação.

As técnicas de pré-processamento não são mutuamente exclusivas: elas usualmente são aplicadas em conjunto. Além disso, não existe uma ordem fixa para a aplicação das diferentes técnicas.

Um termo genérico bastante utilizado para englobar as técnicas de pré-processamento é o processo de ETL (*Extract, Transform, Load*). Neste processo, primeiramente os dados de interesse são extraídos das fontes de dados. Na sequência, os dados passam por sucessivos processos de transformação, nos quais são aplicadas as técnicas de pré-processamento. Por fim, os dados são armazenados em algum local, como um banco de dados ou um *data warehouse*.

2.3 *Data warehousing*

A grande quantidade de dados produzidos pelas organizações ao longo dos anos motivou o desenvolvimento de ferramentas capazes de extrair informações úteis que podem auxiliar na decisão estratégica das organizações (GOLFARELLI; RIZZI, 2018). Nesse cenário, o ambiente de *data warehousing* surgiu para oferecer suporte para o processo de tomada de decisão. Esse ambiente abrange arquiteturas, algoritmos e ferramentas que possibilitam armazenar, em uma única base de dados, dados de fontes autônomas, heterogêneas e distribuídas (CIFERRI, 2002). Essa base de dados pode então ser usada para a realização de consultas analíticas voltadas à tomada de decisão (GOLFARELLI, 2009; GOLFARELLI; RIZZI, 2018).

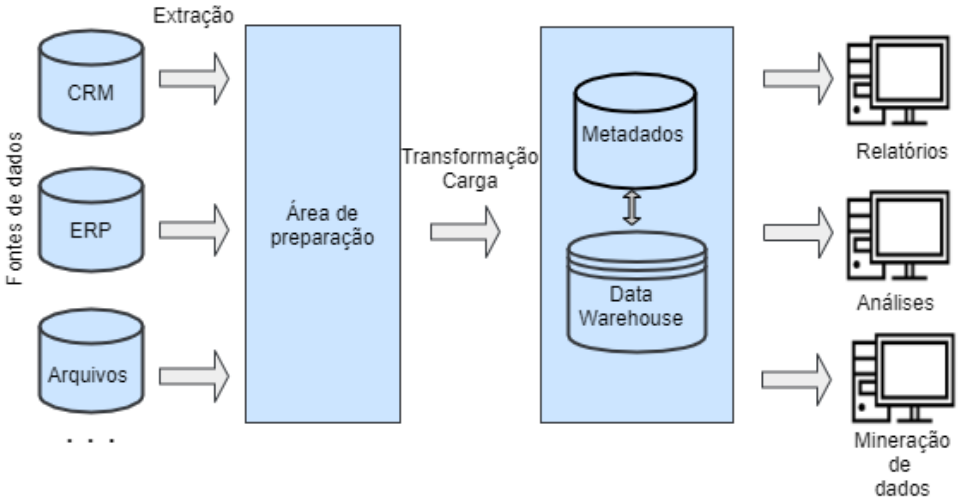
Na seção 2.3.1 é descrita a arquitetura tradicional de um ambiente de *data warehousing*, com destaque para o *data warehouse*. Na seção 2.3.2 é detalhado o *data lake*, que é um componente da arquitetura estendida de *data warehousing*.

2.3.1 *Arquitetura*

Uma arquitetura tradicional de *data warehousing* é mostrada na Figura 1 (BRITO, 2017). Resumidamente, dados de diferentes fontes são extraídos e inseridos em uma área de preparação. Nesta área, os dados são transformados pelo processo de ETL definido de acordo com os interesses da organização (CHANDRA; GUPTA, 2018). Os dados transformados são carregados no *data warehouse* e seus metadados são armazenados no repositório de metadados. Os dados do *data warehouse* são então utilizados para a tomada de decisão estratégica.

A seguir, os componentes da arquitetura ilustrada na Figura 1 são descritos em mais detalhes. As fontes de dados são formadas principalmente por sistemas transacionais, os quais contêm dados operacionais armazenados segundo diferentes modelos e formatos. Por exemplo, pode-se citar sistemas legados e sistemas gerenciadores de banco de dados relacionais. Devido à

Figura 1 – Arquitetura tradicional de um ambiente de *data warehousing*. Os dados são recuperados de fontes heterogêneas. O processo de ETL extrai, transforma e armazena os dados no *data warehouse*, o qual é acessado por ferramentas de geração de relatórios e mineração na tomada de decisão estratégica.



Fonte: Adaptada de Brito (2017).

heterogeneidade dessas inúmeras fontes, o processo de ETL é obrigatório para transformar os dados extraídos em dados confiáveis.

Além de armazenar dados integrados obtidos do processo de ETL, o *data warehouse* também é caracterizado por armazenar dados orientados a assunto, não voláteis e históricos (INMON, 1992). Os dados do *data warehouse* são orientados a assunto, ou seja, relativos aos temas de negócio de maior interesse da corporação. A característica de não-volatilidade está relacionada ao fato de que o conteúdo do *data warehouse* permanece estável por longos períodos de tempo (SILVERS, 2008). Ou seja, praticamente não existe atualização dos dados. Finalmente, os dados do *data warehouse* são históricos, ou seja, eles são referentes a longos períodos de tempo, como dados de 5 a 10 anos. Isso possibilita a realização de análises históricas.

Todas as informações sobre as fontes de dados, quais dados são extraídos de quais fontes, sobre o processo de ETL aplicado, sobre as características do *data warehouse* e das ferramentas são mantidas em um repositório de metadados. O repositório de metadados representa o principal recurso para a administração dos dados no ambiente de *data warehousing*, permitindo automatizar o carregamento e a atualização do *data warehouse*. Isso significa que os metadados são acessados durante o carregamento e atualização e controlam a extração, transformação e carregamento de dados (STAUDT; VADUVA; VETTERLI, 1999).

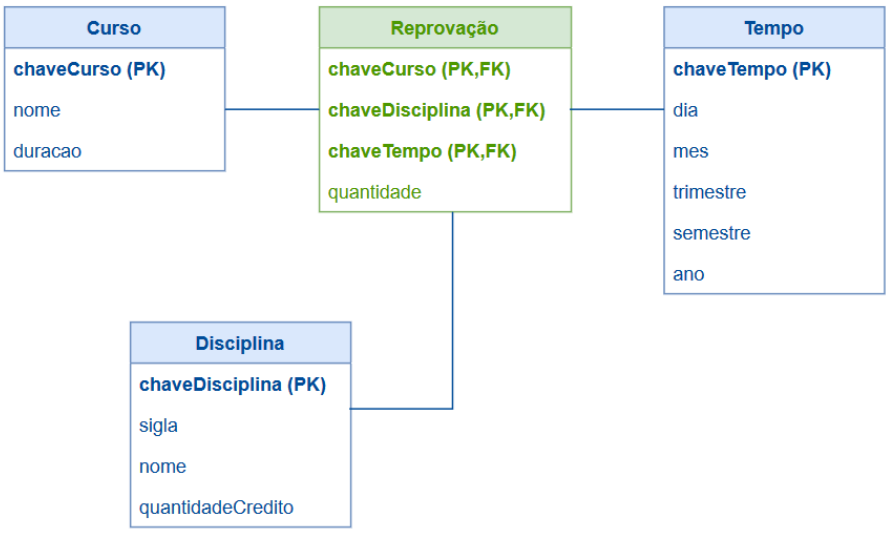
Por fim, a partir dos dados armazenados no *data warehouse*, é possível gerar relatórios estatísticos, aplicar algoritmos de mineração de dados e utilizar ferramentas de *business intelligence* para analisar dados e gerar *insights* sobre o negócio.

Nas implementações relacionais do *data warehouse*, o esquema estrela introduzido em Kimball e Ross (2002) tem sido amplamente adotado para modelagem multidimensional no nível

lógico. O esquema estrela possui dois tipos de tabelas, a tabela de fatos, localizada visualmente no centro, e as tabelas de dimensão, localizadas visualmente nas extremidades. A tabela de fatos contém as medidas numéricas relevantes ao negócio e as referências (chaves estrangeiras) para as tabelas de dimensão. As medidas numéricas refletem os assuntos de interesse importantes para a tomada de decisão estratégica. As tabelas de dimensão contêm uma chave, usada para realizar a junção com a tabela de fatos, assim como os atributos que descrevem aquela dimensão (WEININGER, 2002).

Na Figura 2 é ilustrado um exemplo de esquema estrela. Esse esquema é composto pela tabela de fatos *Reprovação*, cuja medida numérica é quantidade, e pelas tabelas de dimensão *Curso, Disciplina* e *Tempo*. Para fins ilustrativos, foram acrescentados apenas alguns atributos para as dimensões. Esse esquema estrela representa a visão multidimensional *quantidade de reprovados por curso por trimestre por disciplina*, a qual possibilita, por exemplo, que uma universidade analise o índice de reprovação.

Figura 2 – Exemplo de um esquema estrela.



Fonte: Elaborada pelo autor.

2.3.2 Data lake

O *data warehouse* é um componente de destaque da arquitetura tradicional de *data warehousing*. Com a necessidade de manipulação de *big data*, adiciona-se a essa arquitetura o *data lake*. Assim como o *data warehouse*, o *data lake* também armazena dados. Entretanto, conforme definido em Fang (2015), o *data lake* consiste em um enorme repositório de dados baseado em tecnologias de baixo custo que melhora a captura, refinamento, arquivamento e exploração de dados brutos dentro de uma empresa.

Na Figura 3 são detalhados os componentes do *data lake* (RAVAT; ZHAO, 2019). Na área de dados brutos, todos os tipos de dados são inseridos sem processamento e armazenados

em seu formato nativo. Na área de processo, os dados são transformados de acordo com seus requisitos e são armazenados todos os dados intermediários. O processamento de dados inclui o processamento em lote e/ou em tempo real. Esta área permite que o cientista de dados processe dados (seleção, projeção, junção e agregação, dentre outros) para suas análises de dados. Na área de acesso, são armazenados todos os dados disponíveis para análise de dados, além de ser fornecido o acesso aos dados. Esta área permite a realização de diferentes análises, como geração de relatórios, análises estatísticas, análises de inteligência de negócio e aplicação de algoritmos de mineração de dados.

A governança de dados é aplicada em todas as áreas. Ela é responsável por garantir a segurança dos dados, a qualidade dos dados, o ciclo de vida dos dados, o acesso aos dados e o gerenciamento de metadados. Por fim, o repositório de metadados armazena todos os aspectos relacionados aos dados e aos processos aplicados.

Figura 3 – Componentes de um *data lake*.

Fonte: Ravat e Zhao (2019).

Devido às características do *data lake*, surge o processo de ELT (*Extract*, *Load*, *Transform*) (KHINE; WANG, 2018), em contrapartida ao processo tradicional de ETL. O processo de ELT indica que dados gerados são extraídos das fontes, armazenados primeiramente no *data lake* para serem posteriormente transformados sob demanda e carregados no *data warehouse* (LAURENT; LAURENT; MADERA, 2020).

Na Tabela 3 são descritas as diferenças entre *data warehouse* e *data lake*, as quais são discutidas a seguir. O *data warehouse* armazena apenas dados estruturados e consolidados por meio do processo de ETL. No *data lake*, os dados são armazenados de forma não estruturada e desorganizada, o que é mais adequado para *big data*. Adicionalmente, os dados no *data lake* não foram processados e, se algum processamento foi aplicado, ele é muito simples.

Os dados armazenados no *data warehouse* passam pelo processo de ETL, indicando que eles devem ser transformados até que estejam de acordo com o esquema de armazenamento do *data warehouse*. Como exemplo, pode-se citar o esquema estrela que representa o *data warehouse*. Portanto, tem-se uma abordagem *schema-on-write*. As consultas dos usuários devem incidir sobre esse esquema bem definido. Por outro lado, quando os dados encontram-se armazenados no *data lake*, tem-se uma abordagem *schema-on-read*. Isso significa que os dados são transformados

apenas quando consultados, sendo essa transformação feita com base nas características das aplicações dos usuários.

Muitas soluções de *data lake* são implementadas em uma estrutura de código aberto e projetadas para servidores comuns. Portanto, em comparação com as altas taxas de licenciamento do armazenamento dos dados do *data warehouse*, o armazenamento no *data lake* é relativamente mais barato. Com relação à agilidade, embora seja possível alterar o projeto do *data warehouse*, esse processo consome muito tempo e exige um esforço enorme, pois está vinculado a muitos processos de negócios. O *data lake* não possui a estrutura explicitamente definida. Portanto, sua manipulação é mais flexível e ágil.

Adicionalmente, *data warehouses* existem há décadas e possuem aspectos de segurança bem definidos. Como o *data lake* surgiu mais recentemente, aspectos de segurança ainda encontram-se em estágio de amadurecimento. Por fim, até o momento, o uso do *data warehouse* é mais adequado para o processamento analítico e a tomada de decisão estratégica, ou seja, ele é indicado para gerentes e profissionais de negócio que tomam decisões. Já o uso do *data lake* é mais adequado para analistas e cientistas de dados.

Tabela 3 – Comparação entre *data warehouse* e *data lake*.

	Data warehouse	*Data lake*
Dados	Estruturado, dados processados	Estruturado, semi-estruturado, dados não estruturados, dados brutos, dados não processados
Processamento	*Schema-on-write*	*Schema-on-read*
Armazenamento	Caro	Baixo custo
Agilidade	Configuração fixa e menos ágil	Alta agilidade e configuração flexível
Segurança	Bem definida	Em amadurecimento
Usuários	Profissional de negócios	Analistas e cientistas de dados

Fonte: Khine e Wang (2018).

2.4 Processamento paralelo e distribuído

Ambos *data warehouse* e *data lake* podem armazenar volumes de dados dentro do conceito de *big data*. Segundo Laney *et al.* (2001), *big data* é definido em termos de seus atributos de acordo com o modelo de 3Vs: (i) volume, indicando que existe um gigantesco volume de dados; (ii) velocidade, indicando que esse gigantesco volume de dados é gerado muito rapidamente; e (iii) variedade, indicando que o gigantesco volume de dados rapidamente gerado usualmente possui heterogeneidade de formatos e estruturas.

A manipulação de *big data* introduz diversos requisitos, dentre os quais destacam-se a necessidade de se usar sistemas de arquivos distribuídos, como o *Hadoop Distributed File*

System (HDFS), mecanismos de processamento compatíveis com o HDFS, como Apache Hive, e *frameworks* para processamento paralelo e distribuído, como Spark. Conceitos relacionados ao HDFS, Apache Hive e Spark são descritos nas seções 2.4.1, 2.4.2 e 2.4.3, respectivamente.

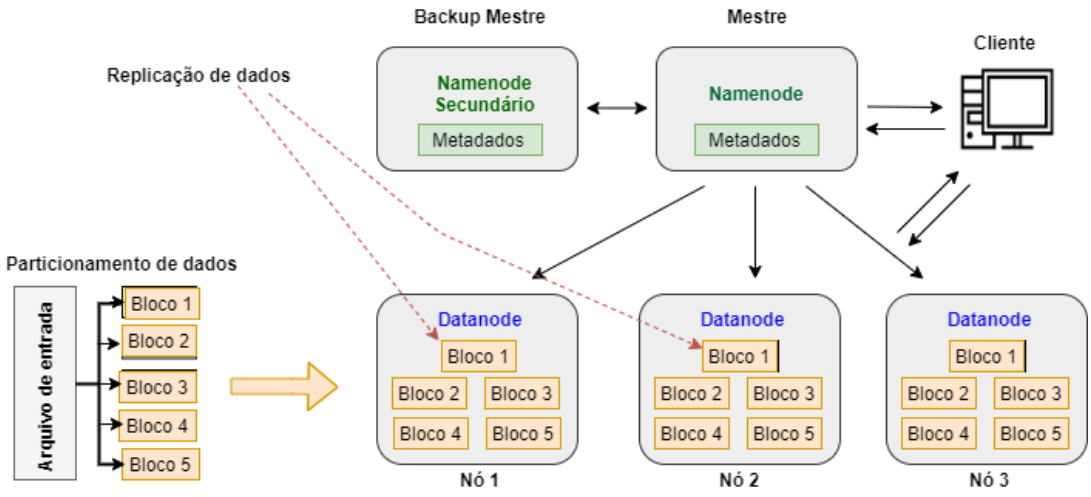
2.4.1 HDFS - Hadoop Distributed File System

O HDFS (SHVACHKO *et al.*, 2010; EDWARDS *et al.*, 2019) é um sistema de arquivos distribuído, altamente escalável, tolerante a falhas e especialmente projetado para armazenar grandes quantidades de dados com o uso de *hardware* de baixo custo (TIWARY; SAHOO; MISRA, 2014). Os arquivos no HDFS são divididos em blocos. O tamanho padrão de cada bloco é 64 MB, sendo que esse tamanho pode ser ajustado de acordo com o ambiente de aplicação. Cada bloco possui três duplicatas em diferentes nós. No caso de alguns blocos danificados, os dados podem ser recuperados da duplicata imediatamente (SHANG; GAN; WU, 2018).

O HDFS é baseado no sistema primário-secundário. Ele é composto por dois tipos de nó: *NameNode* e *DataNode*. O *NameNode*, ou nó primário, possui os metadados sobre a localização dos blocos dos arquivos, enquanto que os *DataNodes*, ou nós secundários, são os nós que armazenam os blocos em si e que são responsáveis pelo processamento de tarefas (SHVACHKO *et al.*, 2010; SHANG; GAN; WU, 2018).

Quando uma aplicação precisa acessar um dado, ela deve primeiro acessar o *NameNode* para saber em qual nó aquele bloco está armazenado e, depois, acessar o *DataNode* correspondente. A arquitetura do HDFS inclui um *NameNode* secundário, utilizado como *backup* do *NameNode* primário. Se o *NameNode* primário falhar, o *NameNode* secundário assume automaticamente a função (SHVACHKO *et al.*, 2010). A Figura 4 ilustra o exemplo de um arquivo de dados dividido em 5 blocos e armazenado de acordo com a arquitetura do HDFS.

Figura 4 – Arquitetura HDFS.



Fonte: Adaptada de Brito (2017).

Por ser um sistema de arquivos, o HDFS pode armazenar arquivos em qualquer formato.

No entanto, para fazer uso adequado da plataforma, o HDFS é utilizado juntamente com mecanismos de processamento que devem ser capazes de manipular os dados de forma eficiente. Assim, o formato de armazenamento define como o mecanismo de processamento interage com o HDFS.

Os mecanismos de processamento compatíveis com HDFS mais conhecidos oferecem suporte para vários formatos de arquivo por meio de APIs apropriadas. Entre os formatos de arquivo, pode-se citar os arquivos binários, de texto, sequenciais, *Optimized Row Columnar* (ORC)[1], Parquet[2] e Avro[3]. Como mecanismos de armazenamento, destacam-se Apache Kudu[4] (formato tabular orientado a colunas), Apache HBase[5] (orientado a colunas) e Apache Hive[6].

2.4.2 Apache Hive

Hive é uma ferramenta de processamento de dados que disponibiliza uma interface que utiliza a linguagem de consulta estruturada, ou do inglês, *Structured Query Language* (SQL), para acessar dados de arquivos armazenados no HDFS (KUMAR; SHINDGIKAR, 2018). A linguagem de consulta do Hive, chamada de HiveSQL (HQL), pode ser executada em diferentes mecanismos, como *Map Reduce*, *Tez* e *Spark*.

A estrutura de metadados do Hive fornece uma estrutura semelhante a uma tabela de alto nível sobre o HDFS. Existem três estruturas de dados principais: tabelas, partições e *buckets*. As tabelas correspondem aos diretórios HDFS e podem ser divididas em partições, nas quais os arquivos de dados podem ser divididos em *buckets* ou blocos.

A estrutura de metadados do Hive é baseada no conceito *schema-on-read*, indicando que não é necessário definir o esquema no Hive antes de se armazenar dados no HDFS (DU, 2015). A aplicação de metadados do Hive após o armazenamento de dados garante flexibilidade e eficiência.

Os tipos de arquivos para os quais Hive oferece suporte são *textfile* (padrão), *Sequencefile*, RCFile, ORC, Parquet, Avro e JSONFile. Já os tipos de compressão de arquivos são *gzip, bzip2, lzo, snappy* e *deflate*.

2.4.3 Spark

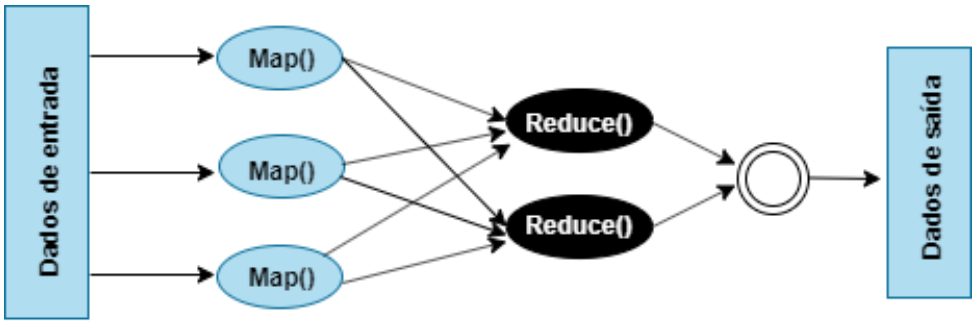
Spark (ZAHARIA *et al.*, 2010) é um *framework* de processamento paralelo e distribuído que visa simplificar a interação da aplicação com as funcionalidades providas pelo HDFS. Além

[1] <https://orc.apache.org/>
[2] <https://parquet.apache.org/>
[3] <https://avro.apache.org/>
[4] <https://kudu.apache.org/>
[5] <https://hbase.apache.org/>
[6] <https://hive.apache.org/>

de utilizar os princípios providos pelo HDFS, Spark também estende o modelo de programação *MapReduce*.

O modelo de programação *MapReduce* é baseado em duas funções: *map* e *reduce* (SOGODEKAR *et al.*, 2016). Na Figura 5 é ilustrado como essas funções trabalham. Dados de entrada são inicialmente processados pela função *map*, a qual tem como objetivo gerar pares no formato chave/valor. A saída produzida pela função *map* é a entrada da função *reduce*, a qual processa os pares no formato chave/valor intermediários para gerar uma saída no formato chave/valor que considera todos os valores para uma mesma chave (SHAIK *et al.*, 2017).

Figura 5 – MapReduce.



Fonte: Adaptada de Shaik *et al.* (2017).

Spark é baseado no uso de RDDs (conjuntos de dados distribuídos e resilientes), os quais consistem em uma coleção de partições de dados distribuídas em vários nós de dados (HUANG *et al.*, 2017; NGUYEN *et al.*, 2017; MOHAMED *et al.*, 2019). A principal característica dos RDDs é que eles são manipulados em memória primária, o que aumenta o desempenho das aplicações.

Adicionalmente, Apache Spark oferece suporte para o agendamento de tarefas na forma de grafos direcionados e acíclicos. Neste sentido, os RDDs podem ser processados por dois tipos de operação (HONG; CHOI; JEONG, 2017). O primeiro tipo é uma transformação que converte um RDD de entrada em um novo RDD. Por exemplo, pode-se citar as funções *map*, *filter*, *reduceByKey* e *join*. O segundo tipo é uma ação, a qual gera os dados de saída a partir de um RDD de entrada. Por exemplo, pode-se citar as funções *reduce, collect, foreach* (GROSSMAN; SARKAR, 2016).

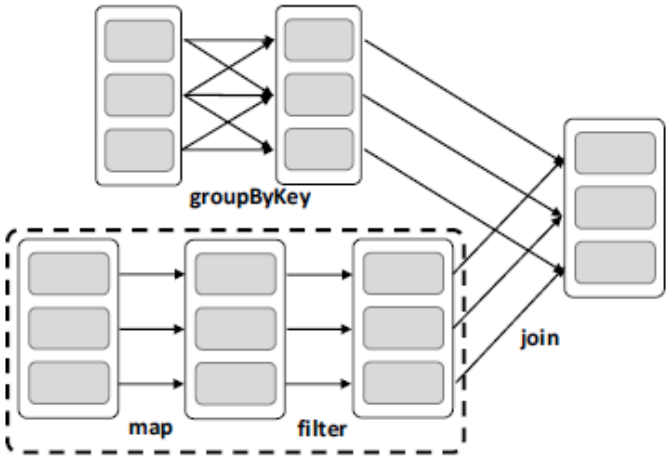
Quando o usuário executa uma ação, o escalonador (*scheduler*) do Spark examina o gráfico de dependência do RDD para a ação, também chamado de *lineage graph*. Em um *lineage graph*, existem dois tipos de dependência entre os RDDs: limitada e ampla (HONG; CHOI; JEONG, 2017). A dependência limitada é quando o RDD filho tem apenas um RDD pai único, enquanto a dependência ampla ocorre quando o RDD filho possui vários RDDs pais.

Na Figura 6 é ilustrado um exemplo de um gráfico de dependência de RDD. Nesta figura, as funções *map* e *filter* são transformações que resultam em uma dependência limitada, enquanto as funções *join* e *groupByKey* resultam em uma dependência ampla. O escalonador do

Spark combina RDDs em uma dependência limitada, chamada *pipeline* de RDD (Figura 6, caixa pontilhada) e os processa como uma única transformação.

O *framework* Spark pode ser utilizado pela implementação de código aberto da *Apache Foundation*: Apache Spark[7].

Figura 6 – Exemplo de gráfico de dependência no processamento de RDDs em Spark.



Fonte: Hong, Choi e Jeong (2017).

2.5 Mineração de dados

Mineração de dados (*data mining*) consiste na aplicação de métodos e técnicas em grandes bases de dados para encontrar tendências ou padrões com o intuito de descobrir conhecimento. Tal conhecimento pode ser usado para ajudar a explicar o comportamento atual ou prever resultados futuros (ROIGER, 2017).

Existem várias técnicas de mineração de dados, dentre as quais se destacam os algoritmos de classificação, descritos na seção 2.5.1. Para analisar os resultados gerados pelas técnicas, devem ser aplicadas métricas de avaliação, as quais são discutidas na seção 2.5.2.

2.5.1 Algoritmos de classificação

A classificação consiste em um processo de aprendizado supervisionado que visa encontrar um modelo (ou função) para descrever e identificar objetos entre classes de dados (HAN; KAMBER; PEI, 2011; SEN; HAJRA; GHOSH, 2020). Este modelo é treinado a partir de um conjunto de objetos rotulados, sendo que existem n rótulos pré-definidos. O modelo construído é usado para prever a classe de novos objetos que possuem rótulos desconhecidos. Esses objetos são então classificados de acordo com um dos n rótulos pré-definidos.

[7] <https://spark.apache.org>

A classificação tem inúmeras aplicações, incluindo detecção de fraude, *marketing*, previsão de desempenho e diagnóstico médico. A seguir são descritos os seguintes classificadores: árvore de decisão, máquinas de vetores de suporte e redes neurais artificiais.

A **árvore de decisão** é um método de classificação eficaz e popular desenvolvido por Leo Breiman em 1984 (BRUCE; BRUCE, 2019). Essas árvores são construídas a partir da raiz por meio de suas folhas. Cada nó da árvore contém um predicado binário, ou seja, uma condição lógica derivada de uma característica específica que pode ser representada como conjuntos de regras do tipo *IF-THEN*, facilitando a leitura e a compreensão humana (QUINLAN, 1986; BRUCE; BRUCE, 2019). Quando uma nova amostra de teste é classificada, cada uma de suas características é analisada, seguindo um caminho existente na árvore até chegar a um nó folha. O rótulo de classe associado a este nó folha é definido como rótulo de classe da nova amostra (NAVADA *et al.*, 2011).

As Máquinas de Vetores de Suporte ou ***Support Vector Machines*** (SVM) são um modelo de classificação popular na literatura de mineração de dados devido à sua eficiência e sucesso. Este modelo representa o limite de decisão usando um subconjunto de amostras de treinamento, conhecido como vetores de suporte. Os conjuntos de dados geralmente possuem mais de um limite de decisão capaz de dividir dados em um espaço n-dimensional. O objetivo do SVM é calcular o hiperplano com a maior margem possível para atuar como limite de decisão (TAN *et al.*, 2009).

Outra abordagem bem conhecida são as Redes Neurais Artificiais ou ***Artificial Neural Networks*** (ANN). O estudo das redes neurais artificiais foi inspirado por tentativas de simular sistemas neurais biológicos. O cérebro humano consiste principalmente de células nervosas chamadas neurônios, conectadas com outras por meio de fios de fibra chamados axônios. O neurônio está conectado aos axônios de outros neurônios por meio de dendritos, que são extensões do corpo celular do neurônio. O ponto de contato entre um dendrito e um axônio é chamado de sinapse. Análoga à estrutura do cérebro humano, uma ANN é composta por um conjunto de nós interligados e conexões direcionadas (TAN *et al.*, 2009).

Um modelo bem conhecido de ANN é chamado de ***MultiLayer Perceptron*** (MLP). Um perceptron é uma estrutura composta por n nós de entrada que recebem os valores das características que descrevem as amostras. Cada nó nesta camada inicial está conectado ao nó de saída por meio de uma aresta com peso w. O desempenho do modelo pode ser melhorado ajustando os valores dos pesos w. O MLP possui camadas ocultas entre as camadas de entrada e saída e trabalha com retro-propagação: os erros de previsão obtidos durante a fase de treinamento são propagados de volta da camada de saída para as camadas anteriores; e esse valor de erro é usado para ajustar os pesos em cada aresta (FACELI *et al.*, 2011; TAN *et al.*, 2009).

2.5.2 Métricas de avaliação

Modelos de classificação devem ser analisados visando identificar seu desempenho, ou seja, a qualidade do resultado obtido. Neste sentido, o método de validação cruzada é amplamente utilizado para avaliar modelos de classificação. O método segmenta os dados em k partições de igual tamanho. Durante cada execução, uma partição é escolhida para teste, enquanto as outras são usadas para treinamento. Como esse procedimento é repetido k vezes, cada partição é usada o mesmo número de vezes para treinamento e apenas uma vez para teste (TAN *et al.*, 2009; GÉRON, 2019).

No centro das métricas de classificação está a matriz de confusão, que é uma tabela que mostra o número de previsões corretas (ou seja, positivas) e incorretas (ou seja, negativas) categorizadas por tipo de resposta (BRUCE; BRUCE, 2019). As células da matriz são preenchidas com quatro valores diferentes. Verdadeiro positivo e verdadeiro negativo indicam amostras positivas e negativas, respectivamente, que foram classificadas corretamente. Falso positivo é o número de amostras negativas classificadas incorretamente como positivas. Finalmente, falso negativo é o número de amostras positivas classificadas incorretamente como negativas (TAN *et al.*, 2009).

Os valores da matriz de confusão são usados para calcular as medidas de desempenho *Accuracy*, *F1-Score*, *Precision* e *Recall*. *Accuracy* é uma medida que geralmente descreve o desempenho do modelo em todas as classes. *F1-Score* tenta encontrar o equilíbrio entre *Precision* e *Recall*. Quanto maior a pontuação *F1-Score*, melhor é o desempenho do modelo. *Precision* indica a fração de instâncias relevantes entre as recuperadas. Por fim, *Recall* é a fração de instâncias relevantes que foram recuperadas.

Outra métrica é a Área Sob Curva (*Area Under Curve - AUC*), que é derivada da curva *Receiver Operating Characteristic* (ROC). A curva ROC é uma abordagem gráfica para exibir o equilíbrio entre a taxa de verdadeiros positivos e a taxa de falsos positivos de um classificador. Em uma curva ROC, a taxa de verdadeiro positivo é mostrada no eixo y e a taxa de falso positivo é mostrada no eixo x. Embora a curva ROC seja uma ferramenta gráfica valiosa, ela não é uma medida única para o desempenho de um classificador. A métrica AUC usa a curva ROC para superar essa limitação. AUC é a área total sob a curva ROC. Quanto maior o valor de AUC, mais eficaz é o classificador. Uma AUC igual a 1 indica um classificador perfeito, ou seja, um classificador que classifica todos os valores 1 corretamente, e não classifica erroneamente nenhum 0 como 1 (BRUCE; BRUCE, 2019; TAN *et al.*, 2009; GÉRON, 2019).

Além de avaliar os modelos de classificação usando métricas, também é importante interpretar as previsões do modelo. O *framework SHapley Additive exPlanations* (SHAP) (LUNDBERG; LEE, 2017) atribui a cada característica um valor de importância para uma predição determinada. Ele calcula a importância dessa característica comparando a previsão do modelo com e sem considerar a característica. Como a ordem em que o modelo investiga as caracte-

rísticas pode afetar as previsões, o SHAP investiga todas as ordens possíveis, garantindo uma comparação justa. Os resultados obtidos pela aplicação do *framework* SHAP são representados graficamente (GARCÍA; AZNARTE, 2020).

2.6 Estatística inferencial

O objetivo da estatística inferencial é produzir afirmações sobre determinada característica da população, a partir de uma amostra dessa população (MORETTIN; BUSSAB, 2017). Essas afirmações devem sempre estar acompanhadas de uma medida de precisão sobre a veracidade. Os testes de hipótese formam parte deste campo da estatística.

O **teste de hipótese** é um procedimento para decisão sobre a veracidade ou falsidade de determinada hipótese (FÁVERO; BELFIORE, 2017). Nos testes de hipóteses, existem duas suposições contraditórias em consideração. O objetivo é decidir, com base nas informações da amostra, qual das duas hipóteses está correta. A hipótese nula (H_0) é a alegação inicial assumida como verdadeira. A hipótese alternativa (H_1) é a afirmação contraditória a H_0.

Os testes de hipóteses se dividem em paramétricos e não paramétricos. Os testes paramétricos são utilizados quando a distribuição do conjunto de dados obedece uma distribuição normal. Neste sentido, podem ser aplicados **testes de normalidade**. Se os resultados sugerirem que nenhuma amostra segue uma distribuição normal considerando uma significância menor do que 0.05, a hipótese de normalidade é rejeitada.

Os testes não paramétricos, por sua vez, são utilizados quando a distribuição da amostra obedece uma distribuição não normal. O teste ***U Mann-Whitney*** é um dos testes não paramétricos mais poderosos. Ele pode ser aplicado às variáveis quantitativas ou qualitativas em escala ordinal, e tem como objetivo verificar se duas amostras não pareadas ou independentes são extraídas da mesma população (FÁVERO; BELFIORE, 2017).

As medidas de **tamanho do efeito** (*effect size*) trazem informações sobre a significância prática dos resultados obtidos em estudos que realizam a comparação de dois grupos. Os valores convencionais de tamanho do efeito, usando como base as regras de Cohen (2013), são 0.01, 0.09, 0.25. Esses valores correspondem ao efeito pequeno, médio e grande, respectivamente.

2.7 Teoria de resposta ao item

O ENEM, alvo de interesse nesta tese, utiliza como base a Teoria de Resposta ao Item (TRI). Essa teoria consiste de um conjunto de modelos matemáticos que representam a chance do participante do ENEM que está respondendo à questão de acertar a resposta desta questão, com base nos parâmetros da questão e de sua habilidade. Ou seja, o modelo é baseado em uma função crescente que indica que, quanto maior o traço latente (ou seja, a habilidade do participante),

maior a chance da questão ser respondida de forma correta (ANDRADE; TAVARES; VALLE, 2000; BAKER; KIM, 2004; KLEIN, 2009).

A TRI utilizada no ENEM inclui o modelo logístico de três parâmetros (ML3P). Este modelo inclui parâmetros de discriminação do item, dificuldade do item e probabilidade de acerto casual, também conhecidos como parâmetros *a, b* e *c*. O parâmetro de discriminação, como o próprio nome indica, é um valor mínimo que garante que respondentes com diferentes proficiências tenham diferentes probabilidades de acertos. Quanto maior a discriminação do item, melhor sua capacidade de diferenciar os participantes que têm dos que não têm a habilidade que se deseja avaliar.

O parâmetro de dificuldade representa a proficiência mínima que um participante deve ter para que sua probabilidade de sucesso seja alta. Assim, esse parâmetro pode ser chamado de "proficiência do item". Por fim, vale destacar que, em avaliações nas quais são possíveis acertos casuais, como o ENEM, o TRI considera não apenas o número de acertos, mas também o padrão de respostas dos alunos. Ou seja, dois alunos com a mesma nota podem receber valores de proficiência diferentes da TRI. O aluno que apresentar as respostas aos itens de forma coerente com o construto que está sendo medido tem maior proficiência.

Sobre os itens, é possível classificá-los em dicotômicos e não dicotômicos. Dicotômicos são itens que podem ter sua resposta classificada como certo (1) ou errado (0), como os itens do ENEM. Esses itens, apesar de serem caracterizados por 5 alternativas, geram apenas uma única alternativa certa como gabarito (1), sendo as demais distrativas (0). Também é importante esclarecer que as respostas aos itens são dicotomizadas como certas ou erradas para a aplicação da TRI no ENEM. Como resultado, não existe diferença com relação a qual alternativa incorreta o participante marcar; o resultado é sempre o mesmo.

As questões que compõem a prova do ENEM são escolhidas por sorteio. Antes de serem aplicadas na prova, as questões são incluídas em provas com alunos do primeiro e segundo anos do ensino médio. Esse processo é confidencial, e os alunos não sabem que estão testando possíveis questões que podem ser incorporadas ao ENEM. Este teste possibilita que o Inep mensure a dificuldade das questões. Questões com alto índice de acertos ou erros são descartadas. Adicionalmente, as questões restantes são classificadas em fácil, médio e difícil. A prova é composta por 25% de questões fáceis, 50% de questões com média dificuldade e 25% de questões difíceis (ENEM, 2013).

2.8 Considerações finais

Neste capítulo foi descrita a fundamentação teórica utilizada como base para o desenvolvimento do trabalho. Foram detalhados os seguintes tópicos: (i) ciência de dados; (ii) técnicas de pré-processamento de dados; (iii) ambientes *data warehousing*, com destaque para o *data warehouse* e o *data lake*; (iv) ambientes computacionais paralelos e distribuídos, englobando

o sistema de arquivos distribuído HDFS, o mecanismo de processamento Apache Hive e o *framework* de processamento paralelo e distribuído Spark; (v) algoritmos de classificação, ou seja, árvore de decisão, máquinas de vetores de suporte e redes neurais artificiais; (vi) métricas de avaliação utilizadas no processo de mineração de dados, ou seja, *Accuracy*, *F1-Score*; *Precision*, *Recall* e *AUC*, além do *framework* SHAP; (vii) estatística inferencial, com destaque para o teste de hipótese; e (viii) teoria de resposta ao item.

No próximo capítulo, Capítulo 3, é detalhado o processo de revisão sistemática que foi realizado no contexto desta tese, o qual engloba conceitos descritos neste capítulo para identificar abordagens na literatura que se relacionam com o objetivo do trabalho.

CAPÍTULO

3

REVISÃO SISTEMÁTICA

Neste capítulo é feita uma descrição da revisão sistemática. Segundo Kitchenham e Charters (2007) e Biolchini *et al.* (2007), o processo de revisão sistemática é dividido em três fases: planejamento, condução e extração de dados. Na seção 3.1 é detalhada a fase de planejamento, a qual define os objetivos, as perguntas de pesquisa, as palavras-chave e as *strings* de pesquisa. Na seção 3.2 é detalhada a fase de condução, na qual as buscas são realizadas e os estudos mais relacionados ao projeto são selecionados com base nas questões de busca e nos critérios de inclusão. Na seção 3.3 é descrita a fase de extração de dados, na qual são sintetizados os estudos selecionados na fase de seleção. O capítulo é finalizado na seção 3.4 com as considerações finais.

3.1 Planejamento

O planejamento é a primeira fase do processo de revisão sistemática e consiste na definição dos seguintes aspectos: objetivos (seção 3.1.1), questões de pesquisa (seção 3.1.2), identificação dos estudos (seção 3.1.3), critérios de seleção (seção 3.1.4) e procedimento de seleção (seção 3.1.5).

Na revisão sistemática descrita nesta tese, utilizou-se a ferramenta computacional denominada StArt (*State of the Art through Systematic Review*) (FABBRI *et al.*, 2012), que tem como objetivo auxiliar o pesquisador nesse processo.

3.1.1 Objetivos

Os objetivos desta revisão sistemática são:

O1: Identificar estudos que realizam análise de dados do ENEM com o objetivo de explorar o desempenho dos participantes.

O2: Identificar estudos que explorem a análise de dados do ENEM com foco no desem-

penho dos participantes e com o auxílio de técnicas da ciência de dados.

3.1.2 Questões de pesquisa

Nesta atividade é feita a definição de questões fortemente relacionadas ao objetivo da revisão. Tais questões justificam a análise dos trabalhos selecionados e, portanto, o foco da pesquisa. As questões de pesquisa definidas para a revisão sistemática são:

Q1: Como se caracterizam os estudos que realizam a análise de dados do ENEM no que se refere ao desempenho dos participantes?

Q2: Quais técnicas da ciência de dados utilizam os estudos que realizam a análise de dados do ENEM, com foco no desempenho dos participantes?

As questões de pesquisa podem ser estruturadas em um conjunto PICO, representando População, Intervenção, Comparação e Resultados (*Outcomes*). A população identifica o grupo que é observado no contexto da revisão sistemática. A intervenção pontua o que deve ser investigado. A comparação define o que deve ser utilizado como base para comparação da investigação. Por fim, os resultados ilustram as conclusões obtidas ao final da revisão sistemática (BIOLCHINI *et al.*, 2007).

Nesta tese, o conjunto PICO é definido da seguinte forma:

- **População:** constituída por participantes do ENEM.
- **Intervenção:** estudos que analisam o desempenho dos participantes do ENEM.
- **Comparação:** estudos não precisam incluir um comparador para inclusão nesta revisão sistemática.
- **Resultados:** características das análises realizadas nos dados correspondentes ao ENEM quanto ao desempenho dos participantes.

3.1.3 Identificação dos estudos

Nesta seção são documentadas as estratégias de seleção para a busca dos estudos relevantes para a revisão, tais como fontes de busca, idioma e período de tempo considerado, palavras-chave e *strings* de busca.

As fontes de busca utilizadas para a revisão sistemática foram definidas com base em um critério de seleção que envolve condições e características que devem ser atendidas. Os critérios observados para a definição das fontes foram: disponibilidade (o texto completo deve estar disponível), abrangência dos estudos (as fontes devem retornar uma quantidade razoável de trabalhos) e atualização (os trabalhos devolvidos devem ser recentes). De acordo com esses

critérios, as seguintes fontes de pesquisa foram definidas: IEEE Xplore[1], Scopus[2], ACM Digital Library[3], Scientific Electronic Library Online - SciELO[4] e Google Scholar[5]. Com relação à fonte DBLP[6], ela não foi considerada porque as publicações indexadas nesta fonte, em geral, já estão indexadas nas fontes escolhidas (BATISTA *et al.*, 2018).

Quanto ao idioma, os estudos primários obtidos devem ser redigidos em português, por ser a língua do país de origem desta tese, e em inglês, por ser a língua internacionalmente aceita para a escrita de trabalhos científicos. Adicionalmente, desde que um dos critérios para seleção de fontes foi a busca por estudos recentes, definiu-se que a busca deve abranger estudos publicados no período dos últimos 5 anos. Desta forma, foram considerados estudos de 2018 até outubro de 2022, que foi a última data de realização da busca.

Foram definidas as seguintes palavras-chave: (i) exame nacional do ensino médio, ENEM, *national high school exam*, *brazilian national high school exam*; (ii) desempenho, *performance*; e (iii) *data science, data warehouse, DW, OLAP, data mining*. A partir dessas palavras-chave, foram definidas as *strings* de busca, as quais foram formadas pela combinação dessas palavras-chave utilizando os operadores `AND` e `OR` e obedecendo às línguas consideradas para a seleção dos trabalhos.

As *strings* de busca em inglês são:

- *((“national high school exam” OR “brazilian national high school exam” OR “ENEM”) AND (“performance”))*
- *(((“national high school exam” OR “brazilian national high school exam” OR “ENEM”) AND (“performance”)) AND (“data science” OR “DW” OR “data warehouse” OR “OLAP” OR “data mining”))*

As *strings* de busca em português são:

- ((“exame nacional do ensino médio” OR “ENEM”) AND (“desempenho”))
- (((“exame nacional do ensino médio” OR “ENEM”) AND (“desempenho”)) AND (“data science” OR “DW” OR “data warehouse” OR “OLAP” OR “data mining”))

Nas *strings* de busca em português, algumas palavras foram mantidas em inglês devido ao fato de que essas palavras são comumente usadas em vez de sua tradução para o português. Por exemplo, pode-se citar *data science* e *data warehouse*. Portanto, foram traduzidos somente

[1] <http://ieeexplore.ieee.org>
[2] <http://scopus.org>
[3] <http://dl.acm.org>
[4] <https://search.scielo.org/>
[5] <https://scholar.google.com/>
[6] <https://dblp.org>

os termos necessários. Outra observação importante é que foi usado `OR` entre *"national high school exam","brazilian national high school exam"* e *"ENEM"* devido ao fato de que as buscas realizadas com `AND` poderiam excluir artigos relevantes.

O período de tempo considerado foi de 2018 até 2022, contudo, esta revisão sistemática foi iniciada no começo do doutorado e atualizada no final de 2022.

3.1.4 Critérios de seleção

Os critérios de seleção de estudos destinam-se a identificar estudos primários que forneçam evidências diretas sobre as questões de pesquisa. Assim, nesta atividade são definidos os critérios de inclusão, os critérios de exclusão e o procedimento de seleção dos estudos.

Foram definidos os seguintes critérios de inclusão:

- O estudo relata as características da análise dos dados do ENEM quanto ao desempenho dos participantes.
- O estudo analisa dados do ENEM com foco no desempenho dos participantes e com suporte de técnicas de ciência de dados.

Com relação aos critérios de exclusão, considerou-se:

- O estudo investiga fatores que estão fora do contexto de análise desta tese, como componentes emocionais e de estresse, comunicação em sala de aula, envolvimento do aluno com a vida universitária e clima escolar.
- O estudo é escrito em uma língua que não seja inglês ou português.
- O estudo não está disponível na íntegra.
- O estudo é de autoria da autora desta tese. Neste caso, o estudo foi excluído porque ele encontra-se descrito neste texto.

3.1.5 Procedimento de seleção

A seleção inicial dos estudos foi realizada da seguinte forma. Com base nas palavras-chave definidas, foram especificadas as *strings* de busca, as quais foram adaptadas de forma apropriada para serem submetidas a cada uma das fontes de pesquisa selecionadas. Os estudos retornados foram avaliados de acordo com a leitura do título e resumo. Se o estudo atendeu a algum critério de inclusão, ele foi selecionado. Em contrapartida, se o estudo atendeu a algum critério de exclusão, ele foi excluído da revisão sistemática.

Os estudos selecionados foram então analisados pela leitura de sua introdução e conclusão, e validados considerando novamente os critérios de inclusão e exclusão. Os estudos que não

foram excluídos pela leitura da introdução e da conclusão foram então lidos na íntegra. Após essa leitura, foram identificados os estudos que estavam de acordo com os critérios de inclusão. Esses estudos foram utilizados para a extração de informações. Para tanto, eles foram sintetizados em termos de suas principais características e seus resultados.

3.2 Condução

Conforme mostrado na Figura 7, foram encontrados 506 estudos. As buscas que utilizaram *strings* em português e que foram submetidas à fonte *Google Scholar* foram as que retornaram mais resultados. A justificativa para este fato é que a maioria dos trabalhos que analisam dados do ENEM são publicados em periódicos, congressos e *workshops* brasileiros. Adicionalmente, muitos desses veículos de comunicação não são indexados por outras fontes de busca.

Dos 506 estudos retornados, 65 foram excluídos por serem duplicados. A avaliação dos títulos e dos resumos resultou na exclusão de mais 355 artigos. Na sequência, a avaliação da introdução e da conclusão resultou na exclusão de 33 artigos. No total, foram lidos e considerados na revisão sistemática 53 estudos completos. A Figura 8 ilustra o número de estudos aceitos por ano na seleção final.

3.3 Extração de dados

Os estudos que foram aceitos na seleção final visam analisar o desempenho dos participantes do ENEM. Eles foram agrupados em estudos que não utilizam ciência de dados na análise (seção 3.3.1) e em estudos que utilizam a ciência de dados (seção 3.3.2).

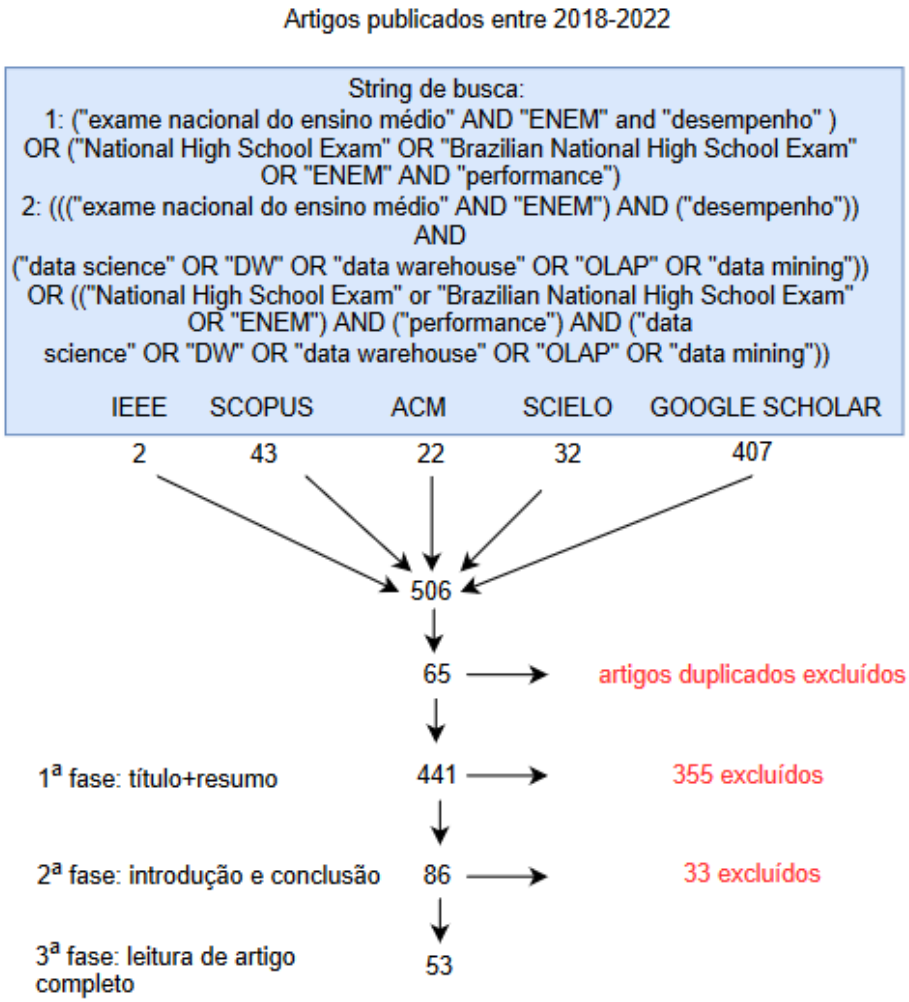
3.3.1 Análise do desempenho dos participantes do ENEM

O objetivo deste grupo é descrever os estudos voltados à analise dos dados do ENEM em relação ao desempenho dos participantes, porém sem empregar técnicas de ciência de dados. Esses estudos são sumarizados de acordo com a área de conhecimento que eles investigam: *Ciência da Natureza* (seção 3.3.1.1), *Matemática* (seção 3.3.1.2) e todas (seção 3.3.1.3). Na seção 3.3.1.4 é feita uma comparação desses estudos na forma de uma tabela. Nesta tabela também são destacados os diferenciais do presente trabalho frente a esses estudos.

3.3.1.1 Considerando Ciência da Natureza

Os trabalhos de Barroso, Rubini e Silva (2018), Nascimento, Cavalcanti e Ostermann (2018), Nascimento (2019), Cestaro, Kleinke e Alle (2020), Gomes, Fernandes *et al.* (2021), Lima e Fraga (2021) e Marques *et al.* (2022) investigam o desempenho dos participantes na área de conhecimento de *Ciência da Natureza*.

Figura 7 – Processos de seleção dos estudos.

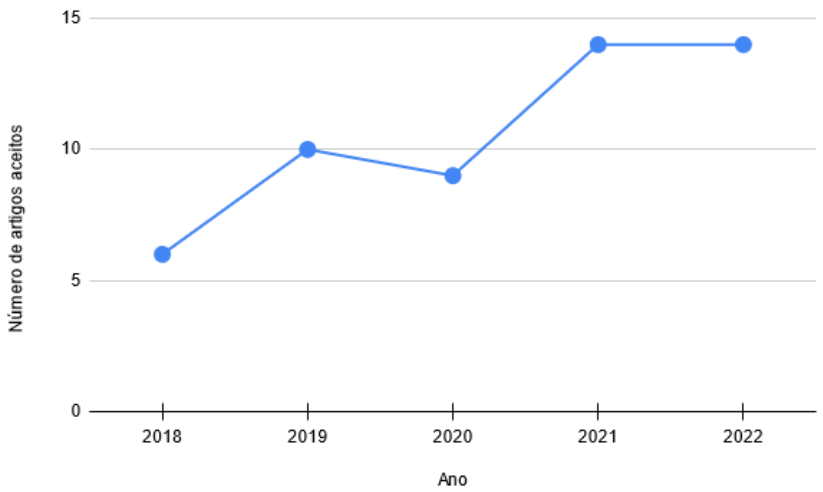


Fonte: Elaborada pelo autor.

No trabalho de Barroso, Rubini e Silva (2018), são descritas 12 questões de física selecionadas dos exames do ENEM no período entre 2009 e 2014, nas quais as respostas alternativas (chamadas de distratores) revelam conceitos não científicos bem definidos na literatura. Nascimento, Cavalcanti e Ostermann (2018) analisam exames dos anos 2009, 2012 e 2015 e avaliam questões da física que são pouco associadas ao nível socioeconômico dos participantes, mas que ainda assim garantem a discriminação por proficiência. Sobre os mesmos dados de física mencionados anteriormente, Nascimento (2019) realiza análises quantitativas sobre o questionário socioeconômico para estudar de que forma o exame do ENEM seleciona os alunos oriundos de contextos sociais favoráveis.

Com o objetivo de compreender e mensurar as dificuldades existentes por parte dos participantes, Cestaro, Kleinke e Alle (2020) descrevem uma pesquisa documental, com abordagens quantitativa e qualitativa, por meio da qual o desempenho dos participantes é analisado considerando 74 questões de biologia referentes aos anos de 2012 a 2016. Gomes, Fernandes *et*

Figura 8 – Quantidade de artigos aceitos por ano.



Fonte: Elaborada pelo autor.

al. (2021) investigam o desempenho dos estudantes de escolas públicas estaduais brasileiras considerando o ENEM de 2017. Os autores comparam o desempenho dos estudantes considerando diferentes tipos de escolas urbanas e rurais, avaliando os fatores que podem causar diferenças nesse desempenho.

Lima e Fraga (2021), inspirados pela sociologia da educação de Pierre Bourdieu chamada de análise de correspondência, investigam o efeito da desigualdade social no desempenho dos estudantes no exame do ENEM de 2012 e 2019. Em Marques *et al.* (2022), os autores analisam 49 questões relacionadas com a educação física nos exames aplicados entre 2009 e 2017, de forma a compreender as apropriações e os usos das relações que se estabelecem entre os saberes, bem como a interdisciplinaridade entre as áreas do conhecimento.

3.3.1.2 Considerando Matemática

Foram identificados sete estudos voltados à análise de desempenho dos candidatos do ENEM considerando a área de conhecimento de *Matemática*: Andrade (2019a), Rocha *et al.* (2022), Holanda *et al.* (2022), Moraes *et al.* (2022a), Moraes *et al.* (2022b), Moraes e Peres (2022) e Feijó e França (2021).

Andrade (2019a) avalia a relação existente entre o desempenho dos estudantes para o estado de Sergipe no ano de 2016 e as condições socioeconômicas dos estudantes e o perfil sociodemográfico do estado. Em Rocha *et al.* (2022), são traçadas reflexões sobre como os fatores sociais influenciam o desempenho dos estudantes na *Matemática*, considerando a região Nordeste e os anos de 2017 e 2019. No trabalho de Holanda *et al.* (2022) é descrita uma análise sobre o desempenho na prova de *Matemática* das alunas nos anos de 2010 a 2019. São consideradas as perspectivas gênero, cor/raça, unidade federativa de residência, deficiência, estado civil e situação de conclusão do ensino médio.

Em Moraes *et al.* (2022a), os autores analisam os dados de 2017 dos municípios com população entre 50 e 500 mil habitantes. São realizadas análises segmentadas por cor/raça, sexo, escolaridade materna e renda familiar para verificar de desigualdade de desempenho nas notas de *Matemática*. Em outro estudo considerando os mesmos dados, Moraes *et al.* (2022b) investigam a relação entre os indicadores educacionais e o desempenho em *Matemática* de alunos de escolas públicas e privadas. Ainda, em Moraes e Peres (2022), os autores propõem um estudo sobre o desempenho em 2017 de alunos de escolas públicas e privadas da região Sudeste. São investigadas variáveis individuais, de infraestrutura escolar e indicadores educacionais para alunos de desempenho alto, médio e baixo de escolas públicas e privadas.

Além de considerar a área de conhecimento de *Matemática*, Feijó e França (2021) também investigam as notas da *Redação* com o propósito de identificar diferenças de desempenho entre alunos das redes públicas e privada ao final do ensino médio. Para atingir esse objetivo, os autores utilizam dados do ano de 2017 e aplicam técnicas de decomposição econométricas.

3.3.1.3 Considerando todas áreas de conhecimento

Os trabalhos descritos nesta seção consideram em suas análises as quatro áreas de conhecimento do ENEM, ou seja, *Matemática*, *Ciência da Natureza*, *Linguagens e Códigos*, e *Ciências Humanas*. Os trabalhos de Carmo, Heckler e Carvalho (2020), Freneda (2020), Gomes e Viana (2022) e Guardieiro, Raimundo e Poco (2022) consideram apenas essas áreas, enquanto os demais trabalhos também incorporam a análise da nota da *Redação* (Santos *et al.* (2019a), Dutra *et al.* (2019), Ferreira *et al.* (2021), Leria *et al.* (2021), Nascimento, Cavalcanti e Ostermann (2020), Santos *et al.* (2019c), Azevedo *et al.* (2018) e Cruz (2022)).

Em Carmo, Heckler e Carvalho (2020), investigam-se resultados do ENEM de 2019 com foco nos dados do estado do Rio Grande do Sul. O estudo analisa e compara o perfil educacional e socioeconômico dos participantes que se enquadram nas 5% melhores e 5% piores médias do estado. Freneda (2020) analisa, por meio de uma abordagem quantitativa-descritiva, o desempenho escolar dos participantes do Distrito Federal nos anos de 2017 e 2018, considerando diferentes variáveis como fator determinante. Já a associação entre a disponibilidade de recursos das tecnologias da informação e comunicação e o desempenho escolar mensurado pela nota do ENEM do ano 2015 é detalhada em Gomes e Viana (2022). No trabalho de Guardieiro, Raimundo e Poco (2022), é investigado se as questões do exame de 2019 demonstram curvas de habilidades semelhantes para subpopulações definidas por sexo, raça e renda, independentemente das habilidades reais do participante.

Com relação aos estudos que também incorporam a análise das notas da *Redação*, em Santos *et al.* (2019a) são investigados os perfis dos alunos do terceiro ano do ensino médio brasileiro da rede escolar pública e privada. O trabalho visa identificar quais fatores extraescolares e interescolares influenciam para que o estudante tenha um desempenho consideravelmente bom no ENEM. Dutra *et al.* (2019) analisam os níveis diferenciados da proficiência obtida pelos

Institutos Federais do Brasil nos exames de 2011 a 2015. A população estudada é composta por todos os Institutos Federais do Brasil que divulgaram dados referentes à proficiência educacional obtida nas médias das áreas do conhecimento do ENEM por escola.

A investigação das notas de *Matemática* no ano de 2019 com as notas das demais áreas de conhecimento e da *Redação* é descrita em Ferreira *et al.* (2021). Já o trabalho de Leria *et al.* (2021) analisa o acesso à educação superior dos candidatos com deficiência visual. São considerados dados das edições de 2017 e 2018. Nascimento, Cavalcanti e Ostermann (2020), por sua vez, examinam quantitativamente o perfil dos candidatos que obtiveram desempenho satisfatório em 2009 apesar da situação econômica e social adversa desses candidatos.

No trabalho de Santos *et al.* (2019c) é feita uma análise do perfil dos candidatos do ENEM de 2016 considerando os dados pessoais desses candidatos e o tipo e a localização da escola, além de fatores socieconômicos. O trabalho emprega Apache Spark para implementar um analisador de *log* distribuído, o qual é usado para melhorar o desempenho na análise dos dados.

Em Azevedo *et al.* (2018), os autores introduzem um *website* que possibilita a análise socieconômica dos candidatos do ENEM do ano 2016. Dentre as variáveis consideradas, pode-se citar renda, tipo de escola, gênero e cor/raça. Avaliar o impacto da pandemia da Covid-19 no desempenho dos participantes do exame de 2020 é o objetivo do estudo de Cruz (2022). Para tanto, os autores analisam os dados das edições de 2019 e 2020 e confirmam a hipótese de um impacto negativo no desempenho médio dos estudantes matriculados no último ano do ensino médio.

3.3.1.4 Comparação entre os estudos

Na Tabela 4 é feita uma comparação entre os estudos voltados à análise de desempenho dos participantes do ENEM sem o emprego de técnicas de ciência de dados. A comparação envolve os anos analisados, a área de conhecimento, o estado, os fatores (ou atributos) considerados e a quantidade de participantes envolvidos na análise. Os estudos são listados na mesma ordem em que eles são descritos no texto.

Tabela 4 – Comparação entre os estudos voltados à análise de desempenho dos participantes do ENEM sem o emprego de técnicas de ciência de dados. *Áreas de conhecimento*: MT: Matemática, CN: Ciências da Natureza, CH: Ciências Humanas, LC: Linguagens e Códigos. *Estados*: RS: Rio grande do Sul, SE: Sergipe, DF: Distrito Federal.

Estudo	Ano analisado	Área de conhecimento	Estado	Atributos	Participantes
Barroso, Rubini e Silva (2018)	2019 a 2014	CN (Física)	Todos	Acertos de questões	860.000 a 1.370.000
Nascimento, Cavalcanti e Ostermann (2018)	2009, 2012 e 2015	CN (Física)	Todos	Notas, Socioeconômico	150.000
Nascimento (2019)	2009	CN (Física)	Todos	Notas Socioeconômico	576.779
Cestaro, Kleinke e Alle (2020)	2012 a 2016	CN (Biología)	Todos	Acertos de questões	6.762.538
Gomes, Fernandes *et al.* (2021)	2017	CN	Todos	Notas, Tipo de escola, Localização	Não especificada
Lima e Fraga (2021)	2012 e 2019	CN	Todos	Socioeconômico, Notas	489.167
Marques *et al.* (2022)	2009 a 2017	CN (Educação Física)	Todos	Notas	Não especificada
Andrade (2019a)	2016	MT	SE	Dados pessoais, Socioeconômico, Notas da matemática	67.821
Rocha *et al.* (2022)	2017 a 2019	MT	Estados da região Nordeste	Dados pessoais, Tipo de escola, Localização, Dados dos pedidos de recursos especializados e específicos	4.135.788
Holanda *et al.* (2022)	2010 a 2019	MT	Todos	Dados pessoais, Socioeconômico	65.407.978

Comparação entre os estudos voltados à análise de desempenho dos participantes do ENEM sem o emprego de técnicas de ciência de dados (continuação). *Áreas de conhecimento*: MT: Matemática, CN: Ciências da Natureza, CH: Ciências Humanas, LC: Linguagens e Códigos. *Estados*: RS: Rio grande do Sul, SE: Sergipe, DF: Distrito Federal.

Estudo	Ano analisado	Área de conhecimento	Estado	Atributos	Participantes
Moraes *et al.* (2022a)	2017	MT	Municípios entre 50 e 500 mil habitantes	Dados pessoais, Socioeconômico, Nota de matemática	375.670
Moraes *et al.* (2022b)	2017	MT	Municípios entre 50 e 500 mil habitantes	Dados pessoais, Socioeconômico, Nota de matemática	218.745
Moraes e Peres (2022)	2017	MT	Estados da região Sudeste	Dados pessoais, Socioeconômico, Notas	218.745
Feijó e França (2021)	2017	MT e Redação	Todos	Dados pessoais, Tipo de escola, Localização, Socioeconômico, Notas	1.786.680
Carmo, Heckler e Carvalho (2020)	2019	MT, CN, CH, LC	RS	Todos	39.055
Freneda (2020)	2017 e 2018	MT, CN, CH, LC	DF	Dados pessoais, Tipo de escola, Localização, Socioeconômico, Notas	50.086
Gomes e Viana (2022)	2015	MT, CN, CH, LC	Todos	Socioeconômico	9228 escolas
Guardieiro, Raimundo e Poco (2022)	2019	MT, CN, CH, LC	Todos	Dados pessoais, Socioeconômico, Notas	1.148.773
Santos *et al.* (2019a)	2016	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Tipo de escola, Localização, Socioeconômico, Notas	1.101.136

Comparação entre os estudos voltados à análise de desempenho dos participantes do ENEM sem o emprego de técnicas de ciência de dados (continuação). *Áreas de conhecimento*: MT: Matemática, CN: Ciências da Natureza, CH: Ciências Humanas, LC: Linguagens e Códigos. *Estados*: RS: Rio grande do Sul, SE: Sergipe, DF: Distrito Federal.

Estudo	Ano analisado	Área de conhecimento	Estado	Atributos	Participantes
Dutra *et al.* (2019)	2011 a 2015	MT, CN, CH, LC e Redação	Todos	Tipo de escola, Notas	Não especificada
Ferreira *et al.* (2021)	2019	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Notas	Não especificada
Leria *et al.* (2021)	2017 e 2018	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Dados dos pedidos de recursos especializados e específicos, Notas	6.952.435
Nascimento, Cavalcanti e Ostermann (2020)	2009	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Socioeconômico	576.779
Santos *et al.* (2019c)	2016	MT, CN, CH, LC e Redação	Todos	Dados pessoais	8.627.265
Azevedo *et al.* (2018)	2016	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Tipo de escola, Socioeconômico	30.000
Cruz (2022)	2019 e 2020	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Tipo de escola, Socioeconômico	2.031.065
Este estudo	2016 a 2020	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Dados da escola, Socioeconômico, Provas, Notas	31.750.834

Fonte: Elaborada pelo autor.

3.3.2 Análise do desempenho dos participantes do ENEM apoiada em técnicas de ciências de dados

O objetivo deste grupo é descrever os estudos voltados à analise dos dados do ENEM em relação ao desempenho dos participantes que empregam técnicas de ciência de dados. Esses estudos são sumarizados da seguinte forma. Na seção 3.3.2.1 são sumarizados estudos que investigam uma área de conhecimento em particular. Na seção 3.3.2.2 são resumidos estudos que investigam todas as áreas de conhecimento, porém sem considerar a nota da prova de *Redação*. Na seção 3.3.2.3 são descritos os estudos que analisam o desempenho em termos de todas as áreas de conhecimento e de *Redação*. Na seção 3.3.2.4 é feita uma comparação desses trabalhos na forma de uma tabela. Nesta tabela também são destacadas os diferenciais do presente trabalho frente a esses estudos.

3.3.2.1 Considerando apenas uma área de conhecimento

Foram identificados cinco estudos que avaliam o desempenho considerando apenas uma área de conhecimento específica: Alves, Cechinel e Queiroga (2018), Gomes *et al.* (2021), Filho, Isotani e Penteado (2021), Alves (2018) e Silva *et al.* (2021).

Em Alves, Cechinel e Queiroga (2018), os autores encontram padrões e geram um modelo preditivo do indicador de desempenho considerando as notas da área de conhecimento de *Matemática* do ano de 2015. Para tanto, os autores empregam algoritmos de classificação, como J48, Naive Bayes e árvore de decisão. Gomes *et al.* (2021) também investigam a área de *Matemática*, porém usam dados de 2011 e empregam a abordagem de regressão em árvore e um modelo com 53 preditores.

O estudo descrito em Filho, Isotani e Penteado (2021) explora a hipótese de predição da nota na área de conhecimento de *Ciências Humanas* com auxílio do algoritmo de regressão linear simples. São analisados dados pedagógicos oriundos de notas de avaliações escolares de estudantes de ensino médio de um colégio particular de São Paulo referentes aos anos de 2017 a 2019.

No trabalho de Alves (2018), são encontrados padrões e é gerado um modelo preditivo do indicador de desempenho das notas da prova de *Redação* referentes aos dados educacionais do ENEM de 2016. São consideradas amostras da cidade de Araranguá, localizada no estado de Santa Catarina. Os modelos foram treinados e testados por meio dos algoritmos de Naive Bayes e J48. Em Silva *et al.* (2021), é introduzida uma abordagem para caracterizar perfis de aprendizagem e estimar notas na avaliação da prova de *Redação*. Para isso, são empregadas a teoria de resposta ao item e técnica de agrupamento K-means, modelo de regressão Bayesiana e linear no conjunto de dados do ano 2019.

3.3.2.2 Considerando todas áreas de conhecimento

Os trabalhos descritos nesta seção consideram em suas análises as quatro áreas de conhecimento do ENEM. O trabalho de Araújo e Silva (2020) utiliza o algoritmo de regras de associação FP-Growth a fim de obter a associação entre os fatores socioeconômicos e as notas dos participantes do estado de Goiás no ENEM dos anos de 2016 a 2018. Partindo de uma amostra de 26.731 participantes de 2012 da cidade de Ribeirão Preto, Novaes (2021) mostra que é possível modelar o desempenho do estudante com ajuda de regressão linear. O estudo também discute o desempenho de estudantes oriundos de escolas públicas e particulares.

Santos *et al.* (2019b) analisam quantitativamente os candidatos do ENEM considerando os anos de 1998 a 2017. Eles propõem um analisador colaborativo e distribuído, com auxílio de Apache Spark, que é capaz de processar quantidades significativas de dados armazenados em arquivos de texto não estruturados. Já investigações relativas à descoberta do perfil dos estudantes encontram-se presentes nos trabalhos de Barcellos *et al.* (2020) e Souza (2021). Em Barcellos *et al.* (2020), usa-se a técnica de agrupamento para verificar se o perfil dos estudantes de 2018 pode ou não impactar positivamente nas suas notas. Em Souza (2021), os perfis dos estudantes de 2019 são analisados para se encontrar *insights* que possam relacionar esses perfis ao desempenho.

Fonseca *et al.* (2022) propõem um modelo multidimensional de DW e apresentam gráficos resultantes de análises solicitadas por coordenadores de escolas brasileiras. Essas análises referem-se aos anos de 2014 e 2015 e incluem dados sobre pessoas com necessidades especiais e sobre questões socioeconômicas. Em Fernandes *et al.* (2022), o conceito de DW é usado para explorar os dados das edições de 2015 a 2020. As análises exibem o desempenho dos alunos por gênero por estado brasileiro por escolaridade e por profissão dos pais.

Em Neto *et al.* (2022), é feita uma análise exploratória para comparar os dados do ENEM de 2019 e 2020 para identificar se a pandemia da Covid-19 impactou no desempenho dos participantes. Em Silva, Brito e Adeodato (2022), é descrito um *framework* para investigar o poder preditivo de características relacionadas aos alunos, professores e escolas, de forma a classificar as escolas no quartil superior e inferior das notas do ENEM. Já Jardim, Delgado e Schneider (2022) introduzem um modelo preditivo orientado por dados que identifica a dificuldade das questões da língua portuguesa usando os dados do ENEM dos anos 2016 e 2017.

3.3.2.3 Considerando todas áreas de conhecimento e redação

Com relação aos estudos que também incorporam a análise das notas da *Redação*, Souza (2019) usa a descoberta de conhecimento para primeiramente classificar as escolas por nota, em seguida selecionar as seis melhores e, por último, traçar um perfil de desempenho dos alunos do ensino médio do Distrito Federal tendo como fonte de pesquisa principal os microdados do ENEM de 2017 associados aos dados socioeconômicos e de infraestrutura escolar. Em Silva *et al.* (2020), usam-se técnicas de mineração de dados com foco na identificação de

desigualdades sociais a partir da análise do desempenho dos estudantes das escolas do estado de Minas Gerais que prestaram o exame em 2019. Com o uso de algoritmos de clusterização e de regras de associação, são mapeadas as variáveis determinantes no desempenho dos estudantes. No trabalho de Motokane (2021) são identificadas variáveis com maior impacto no desempenho dos participantes de 2018 considerando os municípios. São usadas técnicas de estatística espacial para analisar a dinâmica da influência territorial e da perspectiva da sociologia da educação.

Os trabalhos descritos a seguir investigam os perfis dos participantes do ENEM e, a partir desses perfis, identificam fatores que influenciam no desempenho. Santana (2018) usa dados do ano de 2014 e emprega classificação, regras de associação e agrupamento. Os dados do ano de 2014 também são usados em Markoski *et al.* (2019) na aplicação de algoritmos de classificação e na realização de consultas SQL para verificar padrões encontrados. Outros estudo que utiliza classificação é descrito em Rodrigues *et al.* (2019), o qual investiga dados de 2017.

Conte (2019) aplica modelos econométricos com mínimos quadrados ordinários sobre dados do ENEM de 2015 para identificar fatores que influenciam no desempenho de candidatos e verificar se há condições igualitárias entre diferentes classes socioeconômicas e em diferentes dependências administrativas escolares. Já em Lima *et al.* (2020) é utilizado o método de agrupamento K-means para analisar dados entre os anos 2012 e 2017 e investigar as regiões brasileiras, áreas de conhecimento do ENEM, tipo de escola e acessibilidade.

Garcia, Neto e Ribeiro (2021) identificam os fatores escolares que mais influenciam a qualidade do ensino médio no Brasil, considerando dados do Censo Escolar e do ENEM de 2016, 2017 e 2018. Para a seleção dos indicadores, são aplicadas técnicas de mineração de dados, análise de regressão e análise de discriminante. Para a quantificação do impacto dos indicadores, é usado um modelo de regressão logística múltipla. Franco (2021) aplica algoritmos de seleção de atributos e classificadores sobre dados de 1998 a 2019 para elencar os fatores mais importantes para detectar alunos de alto e baixo desempenho.

Em Banni, Oliveira e Bernardini (2021), é realizada uma análise experimental sobre os dados do ENEM de 2018 baseada em mineração de dados educacionais, incluindo o uso de técnicas de visualização de dados univariadas e bivariadas, e construção de modelos preditivos para identificar os atributos mais relacionados ao desempenho dos estudantes. Por fim, em Oliveira, Barwaldt e Lucca (2020), usa-se o algoritmo C4.5 para analisar participantes do ano de 2018 que possuem deficiências físicas e psicológicas.

3.3.2.4 Comparação entre os estudos

Na Tabela 5 é feita uma comparação entre os estudos voltados à análise de desempenho dos participantes do ENEM com o emprego de técnicas de ciência de dados. A comparação envolve os anos analisados, a área de conhecimento, o estado, os fatores (ou atributos) considerados, a quantidade de participantes envolvidos na análise e as técnicas utilizadas. Os estudos são listados na mesma ordem em que eles são descritos no texto.

Tabela 5 – Comparação entre os estudos voltados à análise de desempenho dos participantes do ENEM com o emprego de técnicas de ciência de dados. *Área de conhecimento*: MT: Matemática, CN: Ciências da Natureza, CH: Ciências Humanas, LC: Linguagens e Códigos. *Estado*: SC: Santa Catarina, GO: Goiás. DF: Distrito Federal, MG: Minas Gerais.

Estudo	Ano analisado	Área de conhecimento	Estado	Atributos	Participantes	Técnicas utilizadas
Alves, Cechinel e Queiroga (2018)	2015	MT	Todos	Socioeconômico, Escola, Notas	15.598	J48, Naive Bayes
Gomes *et al.* (2021)	2011	MT	Todos	Dados pessoais, Socioeconômico, Notas	3.670.089	Algoritmo CART
Filho, Isotani e Penteado (2021)	2017 a 2019	CH	Cidade de São Paulo	Socioeconômico, Notas	67	Regressão linear simples
Alves (2018)	2016	Redação	SC	Dados pessoais, Socioeconômico, Notas da Redação	176.891	J48, Naive Bayes
Silva *et al.* (2021)	2019	Redação	Todos	Dados pessoais, Socioeconômico, Notas	3.900.000	K-means, Modelo de regressão Bayesiana e linear
Araújo e Silva (2020)	2016 a 2018	MT, CN, CH, LC	GO	Dados pessoais, Socioeconômico, Notas	468.352	Regras de associação FP-Growth
Novaes (2021)	2012	MT, CN, CH, LC	Cidade de SP: Ribeirão Preto	Dados pessoais, Tipo de escola, Notas	26.731	Regressão linear
Santos *et al.* (2019b)	1998 a 2017	MT, CN, CH, LC	Todos	Dados pessoais, Socioeconômico, Notas	Não especificada	Analisador colaborativo e distribuído, Apache Spark
Barcellos *et al.* (2020)	2018	MT, CN, CH, LC	Todos	Dados pessoais, Socioeconômico, Notas	950.977	SimpleKmeans
Souza (2021)	2019	MT, CN, CH, LC	Todos	Dados pessoais, Tipo de escola, Socioeconômico	3.701.947	Regressão Linear
Fonseca *et al.* (2022)	2014 e 2015	MT, CN, CH, LC	Todos	Dados pessoais, Dados dos pedidos de recursos especializados e específicos	Não especificada	Modelo multidimensional DW
Fernandes *et al.* (2022)	2015 a 2020	MT, CN, CH, LC	Todos	Dados pessoais, Tipo de escola, Socioeconômico, Notas	Não especificada	Modelo multidimensional DW
Neto *et al.* (2022)	2019 e 2020	MT, CN, CH, LC	Todos	Dados pessoais, Tipo de escola, Localização, Notas	6.735.254	Visualização de dados, Matplotlib

Fonte: Elaborada pelo autor.

Comparação entre os estudos voltados à análise de desempenho dos participantes do ENEM com o emprego de técnicas de ciência de dados (continuação). *Área de conhecimento*: MT: Matemática, CN: Ciências da Natureza, CH: Ciências Humanas, LC: Linguagens e Códigos. *Estado*: SC: Santa Catarina, GO: Goiás. DF: Distrito Federal, MG: Minas Gerais.

Estudo	Ano analisado	Área de conhecimento	Estado	Atributos	Participantes	Técnicas utilizadas
Silva, Brito e Adeodato (2022)	2009 a 2019	MT, CN, CH, LC	Todos	Dados pessoais, Socioeconômico, Notas	40.000.000	Regressão logística, Random Forest
Jardim, Delgado e Schneider (2022)	2016 e 2017	MT, CN, CH, LC	Todos	Notas	10.273.178	Convolutional Neural Network
Souza (2019)	2017	MT, CN, CH, LC e Redação	DF	Dados pessoais, Socioeconômico, Notas	32.913	J48, Naive Bayes, Lazy learning
Silva *et al.* (2020)	2019	MT, CN, CH, LC e Redação	MG	Dados pessoais, Tipo de escola, Socioeconômico, Notas	88.659	K-means, Algoritmo apriori
Motokane (2021)	2018	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Socioeconômico, Notas	Não especificada	Regressão linear múltipla, Lisa Maps
Santana (2018)	2014	MT, CN, CH, LC e Redação	RJ	Dados pessoais, Socioeconômico, Dados dos pedidos de recursos especializados, Notas	25.080	JRIP, PART, J48, Random Forest
Markoski *et al.* (2019)	2014	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Socioeconômico, Notas	Não especificada	J48, SQL
Rodrigues *et al.* (2019)	2017	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Socioeconômico, Notas	5.688.295	SVM, Naive Bayes, kNN, Neural Network
Conte (2019)	2015	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Tipo de escola, Socioeconômico, Notas	1.301.766	Modelos econométricos, Mínimos quadrados ordinários
Lima *et al.* (2020)	2012 e 2017	MT, CN, CH, LC e Redação	Todos	Tipo de escola, Dados dos pedidos de recursos especializados e específicos, Notas	8.247.574	K-means
Garcia, Neto e Ribeiro (2021)	2016, 2017 e 2018	MT, CN, CH, LC e Redação	Todos	Tipo de escola, Localização, Notas	5.283.200	Regressão logística multinomial, Análise discriminante

Fonte: Elaborada pelo autor.

Comparação entre os estudos voltados à análise de desempenho dos participantes do ENEM com o emprego de técnicas de ciência de dados (continuação). *Área de conhecimento*: MT: Matemática, CN: Ciências da Natureza, CH: Ciências Humanas, LC: Linguagens e Códigos. *Estado*: SC: Santa Catarina, GO: Goiás. DF: Distrito Federal, MG: Minas Gerais.

Estudo	Ano analisado	Área de conhecimento	Estado	Atributos	Participantes	Técnicas utilizadas
Franco (2021)	1998 a 2019	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Tipo de escola, Localização, Socioeconômico, Notas	Não especificada	XGBoost, LighGBM, Árvore de decisão
Bamni, Oliveira e Bernardini (2021)	2018	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Socioeconômico, Notas	5.513.747	Visualização de dados univariadas e bivariadas, Regressão logística
Oliveira, Barwaldt e Lucca (2020)	2018	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Dados dos pedidos de recursos especializados e específicos, Socioeconômico	23.270	C4.5
Este estudo	2016 a 2020	MT, CN, CH, LC e Redação	Todos	Dados pessoais, Dados da escola, Socioeconômico, Provas, Notas	31.750.834	Árvore de decisão, SVM, ANN, Framework SHAP, Data lake, DW, HDFS, Apache Spark, Apache Hive, R, Python, PySpark Airflow, Metabase

Fonte: Elaborada pelo autor.

3.4 Considerações finais

Neste capítulo foi contextualizada a revisão sistemática realizada nesta tese. Foram descritas as fases de planejamento, condução e extração dos dados. Também foi feito um breve resumo de cada estudo obtido na revisão sistemática que atendeu aos objetivos da mesma e que não foi excluído durante o processo. Os estudos foram analisados por meio de tabelas comparativas.

Na Tabela 6 são listados os estudos obtidos na última etapa da revisão sistemática. Os primeiros 26 trabalhos respondem à questão **Q1**. Comparado ao trabalho desenvolvido nesta tese, esses estudos são limitados. Eles usualmente descrevem poucos fatores de análise considerando anos específicos ou um intervalo pequeno de anos. Muitos desses estudos também consideram áreas de conhecimento específicas e realizam análises de participantes de estados e municípios específicos. Os tipos de análises mais utilizados nesses estudos são: análise descritiva, quantitativa e qualitativa. Em contrapartida, o trabalho desenvolvido nesta tese analisa dados de 31.750.834 participantes entre 2016 e 2020, explora dados pessoais, de escola, socioeconômicos, notas e todas as questões das provas. Adicionalmente, o trabalho desenvolvido é mais genérico e flexível, desde que possibilita o uso de ferramentas de análise estatística, incluindo funções descritivas como mínimo, máximo, média, além da capacidade de definir curva de distribuição, cálculos de percentil, testes de normalidade e testes *U de Mann-Whitney*, dentre outros.

Os 27 estudos restantes respondem à questão **Q2**. Os mesmos possuem escopos limitados quanto comparados ao trabalho realizado, devido ao fato de que nenhum desses estudos consideram conjuntamente o uso de técnicas de *mineração de dados*, *estatística inferencial*, *data warehousing* e *processamento paralelo e distribuído*. O trabalho descrito nesta tese possibilita o uso de ferramentas de análise estatística, mineração de dados e consultas OLAP com suporte de tecnologias de código aberto como Apache Hadoop, Apache Spark, Apache Hive, Airflow e Metabase, com o objetivo de investigar a grande quantidade de variáveis consideradas nos arquivos de microdados do ENEM e o grande volume de dados disponibilizados. As buscas realizadas nesta revisão sistemática foram mais genéricas com o intuito de verificar a existência de uma arquitetura para análise de dados do ENEM. Portanto, como em nenhum estudo foi encontrada a proposta de uma arquitetura, os resultados obtidos nesta tese avançam no estado-da-arte, preenchendo uma lacuna existente na literatura, como mostrado na última linha da Tabela 6.

No próximo capítulo, Capítulo 4, é detalhada a arquitetura proposta, a qual visa apoiar o processo de tomada de decisão educacional para os gestores públicos, assim como auxiliar na elaboração de políticas públicas educacionais, a fim de promover a melhoria da qualidade do ensino público e, consequentemente, do desempenho dos alunos.

Tabela 6 – Comparação entre os estudos selecionados na revisão sistemática.

Estudo	Análise descritiva	Ciência de dados		
		Mineração de dados	DW	Ambiente paralelo e distribuído
Barroso, Rubini e Silva (2018)	✓	✗	✗	✗
Nascimento, Cavalcanti e Ostermann (2018)	✓	✗	✗	✗
Nascimento (2019)	✓	✗	✗	✗
Cestaro, Kleinke e Alle (2020)	✓	✗	✗	✗
Gomes, Fernandes *et al.* (2021)	✓	✗	✗	✗
Lima e Fraga (2021)	✓	✗	✗	✗
Marques *et al.* (2022)	✓	✗	✗	✗
Andrade (2019a)	✓	✗	✗	✗
Rocha *et al.* (2022)	✓	✗	✗	✗
Holanda *et al.* (2022)	✓	✗	✗	✗
Moraes *et al.* (2022a)	✓	✗	✗	✗
Moraes *et al.* (2022b)	✓	✗	✗	✗
Moraes e Peres (2022)	✓	✗	✗	✗
Feijó e França (2021)	✓	✗	✗	✗
Carmo, Heckler e Carvalho (2020)	✓	✗	✗	✗
Freneda (2020)	✓	✗	✗	✗
Gomes e Viana (2022)	✓	✗	✗	✗
Guardieiro, Raimundo e Poco (2022)	✓	✗	✗	✗
Santos *et al.* (2019a)	✓	✗	✗	✗
Dutra *et al.* (2019)	✓	✗	✗	✗
Ferreira *et al.* (2021)	✓	✗	✗	✗
Leria *et al.* (2021)	✓	✗	✗	✗
Nascimento, Cavalcanti e Ostermann (2020)	✓	✗	✗	✗
Santos *et al.* (2019c)	✓	✗	✗	✗
Azevedo *et al.* (2018)	✓	✗	✗	✗
Cruz (2022)	✓	✗	✗	✗
Alves, Cechinel e Queiroga (2018)	✗	✓	✗	✗
Gomes *et al.* (2021)	✗	✓	✗	✗
Filho, Isotani e Penteado (2021)	✗	✓	✗	✗
Alves (2018)	✗	✓	✗	✗
Silva *et al.* (2021)	✗	✓	✗	✗
Araújo e Silva (2020)	✗	✓	✗	✗
Novaes (2021)	✗	✓	✗	✗
Santos *et al.* (2019b)	✗	✗	✗	✓
Barcellos *et al.* (2020)	✗	✓	✗	✗
Souza (2021)	✗	✓	✗	✗
Fonseca *et al.* (2022)	✗	✗	✓	✓
Fernandes *et al.* (2022)	✗	✗	✓	✓
Neto *et al.* (2022)	✗	✓	✗	✗
Silva, Brito e Adeodato (2022)	✗	✓	✗	✗
Jardim, Delgado e Schneider (2022)	✗	✓	✗	✗
Souza (2019)	✗	✓	✗	✗

Fonte: Elaborada pelo autor.

Comparação entre os estudos selecionados na revisão sistemática (continuação).

Estudo	Análise descritiva	Ciência de dados		
		Mineração de dados	DW	Ambiente paralelo e distribuído
Silva *et al.* (2020)	✗	✓	✗	✗
Motokane (2021)	✗	✓	✗	✗
Santana (2018)	✗	✓	✗	✗
Markoski *et al.* (2019)	✗	✓	✗	✗
Rodrigues *et al.* (2019)	✗	✓	✗	✓
Conte (2019)	✗	✓	✗	✗
Lima *et al.* (2020)	✗	✓	✗	✗
Garcia, Neto e Ribeiro (2021)	✗	✓	✗	✗
Franco (2021)	✗	✓	✗	✗
Banni, Oliveira e Bernardini (2021)	✗	✓	✗	✗
Oliveira, Barwaldt e Lucca (2020)	✗	✓	✗	✗
Este estudo	✓	✓	✓	✓

Fonte: Elaborada pelo autor.

CAPÍTULO

4

ARQUITETURA PROPOSTA

Neste capítulo é feita a proposta de uma arquitetura para apoiar o processo de tomada de decisão educacional. A arquitetura relaciona dados obtidos de fontes educacionais com um ambiente de processamento e armazenamento paralelo e distribuído e com o uso de técnicas de ciência de dados. Na seção 4.1 é descrita a arquitetura. Na seção 4.2 são exemplificadas duas instâncias da arquitetura com apoio de tecnologias de código aberto. O capítulo é finalizado na seção 4.3 com as considerações finais.

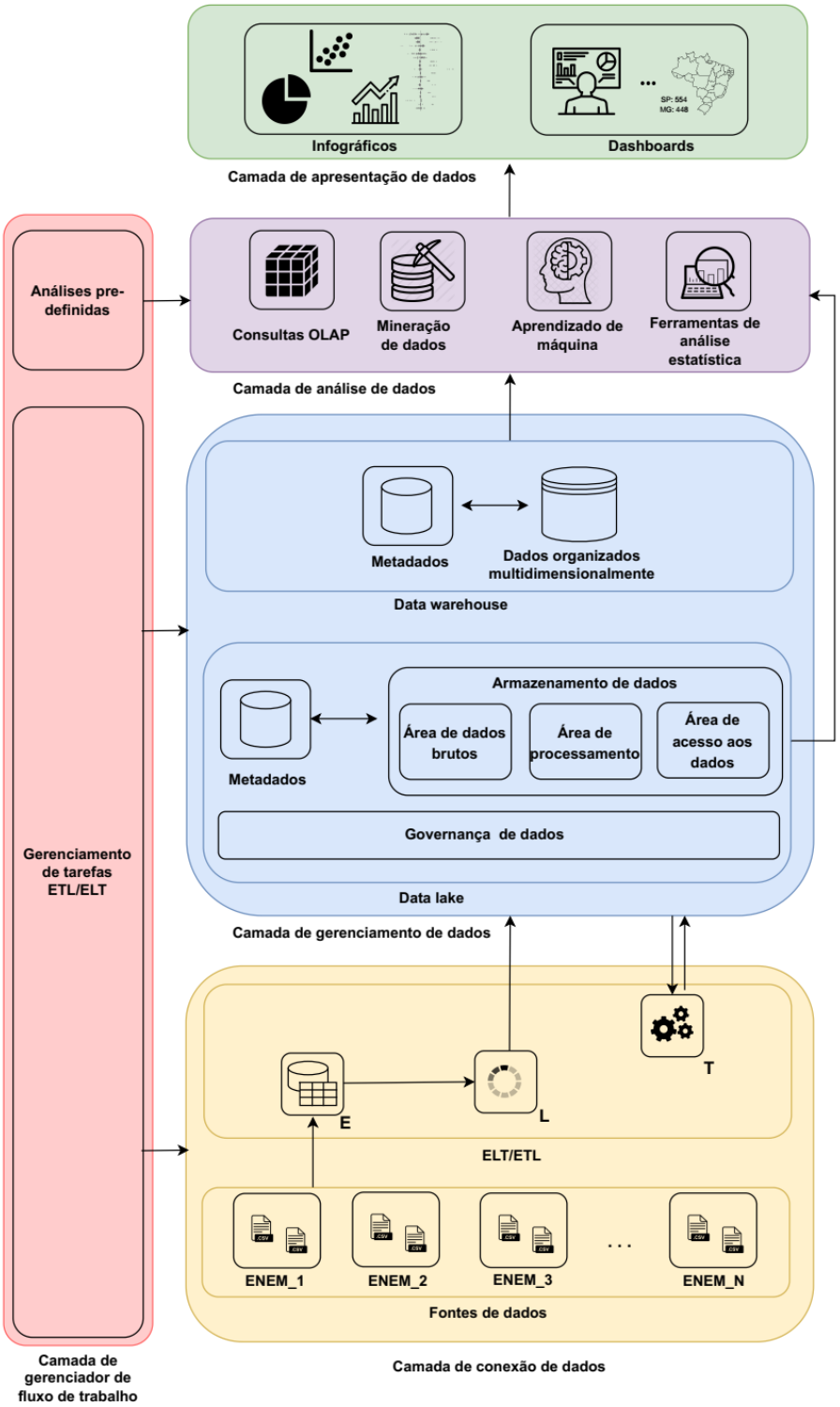
4.1 Arquitetura proposta

A arquitetura proposta é ilustrada na Figura 9 e é composta por cinco camadas: (i) conexão de dados; (ii) gerenciamento de dados; (iii) análise de dados; (iv) apresentação de dados; e (v) gerenciador de fluxo de trabalho. Cada camada é descrita a seguir.

A **camada de conexão de dados** engloba os componentes *fontes de dados* e *processos ELT/ETL*. As *fontes de dados* consistem em dados educacionais. Nesta tese, os dados escolhidos são os dados do **ENEM**, obtidos por meio de arquivos no formato ".csv" e disponibilizados pelo INEP. Para cada ano do exame, são disponibilizados três arquivos: (i) arquivo de microdados, que contém, para cada participante em um determinado ano, a resposta dada pelo participante a cada questão, o gabarito de cada questão, as notas das provas objetivas e da redação, além do questionário respondido pelos participantes; (ii) itens da prova, que contém informações gerais sobre as questões das diferentes provas objetivas; e (iii) dicionário de variáveis, que descreve o significado de cada uma das variáveis contidas nos arquivos de microdados e itens da prova. Apenas os dois primeiros arquivos, arquivo de microdados e itens da prova, são carregados. O dicionário de variáveis é um arquivo complementar, utilizado para obter uma compreensão mais completa dos dados originais.

No componente *processos ELT/ETL*, os dados são preparados para que possam ser

Figura 9 – Visão geral da arquitetura proposta, a qual possui cinco camadas: conexão de dados, gerenciamento de dados, análise de dados, apresentação de dados e gerenciador de fluxo de trabalho.



Fonte: Elaborada pelo autor.

armazenados na camada de gerenciamento de dados. O processo ELT é responsável por extrair (E) os dados das fontes de dados e carregar (L) esses dados no *Data lake*. Neste cenário, as transformações (T) são aplicadas somente quando necessário, ou seja, são aplicadas quando

os dados devem ser armazenados no *Data warehouse* ou devem ser utilizados pela camada de análise de dados. O processo ETL permite que os dados sejam extraídos (E) das fontes de dados ou do *Data lake*, transformados (T) e armazenados (L) no *Data warehouse*.

A **camada de gerenciamento de dados** possui os componentes *Data lake* e *Data warehouse*. O componente *Data lake* engloba os seguintes elementos: *Armazenamento de dados*, *Metadados* e *Governança de dados*, cujas funcionalidades são as mesmas que as especificadas na seção 2.3.2. O componente *Data warehouse* é composto pelo *repositório de dados organizados multidimensionalmente* e pelo *repositório de metadados*. O primeiro repositório armazena dados orientados a assunto, integrados, não voláteis e históricos, além de organizados segundo o modelo multidimensional. O segundo repositório, *metadados*, armazena dados sobre o repositório de dados organizados multidimensionalmente, como sua localização, estrutura e a descrição semântica de seus dados. Tanto no componente *processos ELT/ETL* quanto na **camada de gerenciamento de dados** são utilizados processamento e armazenamento paralelo e distribuído.

A **camada de análise de dados** tem como objetivo principal extrair informações e *insights* úteis e auxiliar na tomada de decisão estratégica. Nesta camada, podem ser utilizadas diferentes técnicas, como: (i) realização de *Consultas OLAP*, as quais permitem manipular os dados de diferentes maneiras com o objetivo de oferecer informações importantes que possam ser usadas para uma possível tomada de decisão; (ii) aplicação de técnicas de *Mineração de dados*, as quais são usadas para auxiliar na extração do conhecimento; (iii) uso de técnicas de *Aprendizado de máquina*, visando automatizar a construção de modelos preditivos e descritivos; e (iv) aplicação de *Ferramentas de análise estatística*, incluindo funções descritivas, como min, max e média, além da capacidade de definir curva de distribuição, cálculos de percentil, testes de normalidade e testes *U de Mann-Whitney*. As análises podem ser realizadas sobre os dados contidos na *área de acesso aos dados* e no repositório de *dados organizados multidimensionalmente*, presentes nos componentes *Data lake* e *Data warehouse*, respectivamente.

A **camada de apresentação de dados** inclui ferramentas de visualização de dados que permitem que os cientistas de dados e os gestores educacionais visualizem graficamente os resultados de suas análises. Nesta camada podem ser utilizadas diferentes ferramentas de visualização, com destaque para: (i) *Infográficos*, adequados para fornecer uma visão geral e fácil de entender sobre um assunto, tais como mapas de calor e gráficos de barras, dentre outros; e (ii) *Dashboards*, também conhecidos como paineis de informações, os quais fornecem uma interface gráfica para que os usuários vejam, de maneira visual, centralizada e dinâmica, as informações mais importantes para auxiliar na tomada de decisão estratégica.

À medida que o volume e a complexidade dos *pipelines* de processamento e análise de dados aumentam, é possível simplificar o processo geral dividindo-o em uma série de tarefas menores e coordenando a execução dessas tarefas como parte de um *workflow*. Para isso, a **camada de gerenciador de fluxo de trabalho** é utilizada. Nesta camada, é possível definir *workflows* que especifiquem a automação de várias etapas, desde as relativas ao processo

ELT/ETL até as presentes na camada de análise.

4.2 *Exemplos de Pipelines*

Um *pipeline* corresponde a uma instanciação de uma arquitetura com tecnologias. O objetivo é mostrar como a arquitetura pode ser utilizada na prática, visto que os componentes da arquitetura a serem usados no desenvolvimento de uma aplicação devem ser escolhidos de acordo com o propósito da aplicação. Adicionalmente, a escolha das tecnologias empregadas depende de vários fatores, como a possibilidade do uso da tecnologia e a experiência do desenvolvedor da aplicação. Nesta seção, são exemplificados dois *pipelines* da arquitetura proposta na seção 4.1 por meio do uso de tecnologias de código aberto.

Na Figura 10 é ilustrada uma instanciação da arquitetura que tem como propósito explorar as funcionalidades do componente *Data lake*. Na **camada de conexão de dados**, a primeira decisão é a seleção dos anos de análise. São escolhidos 5 anos de análise, representados por ENEM_1 a ENEM_5. Para cada ano, são carregados dois arquivos: o arquivo de microdados e o arquivo contendo os itens da prova. O *processo ELT* é realizado usando *Apache Spark*[1], *PySpark* e a linguagem de programação *Python*[2].

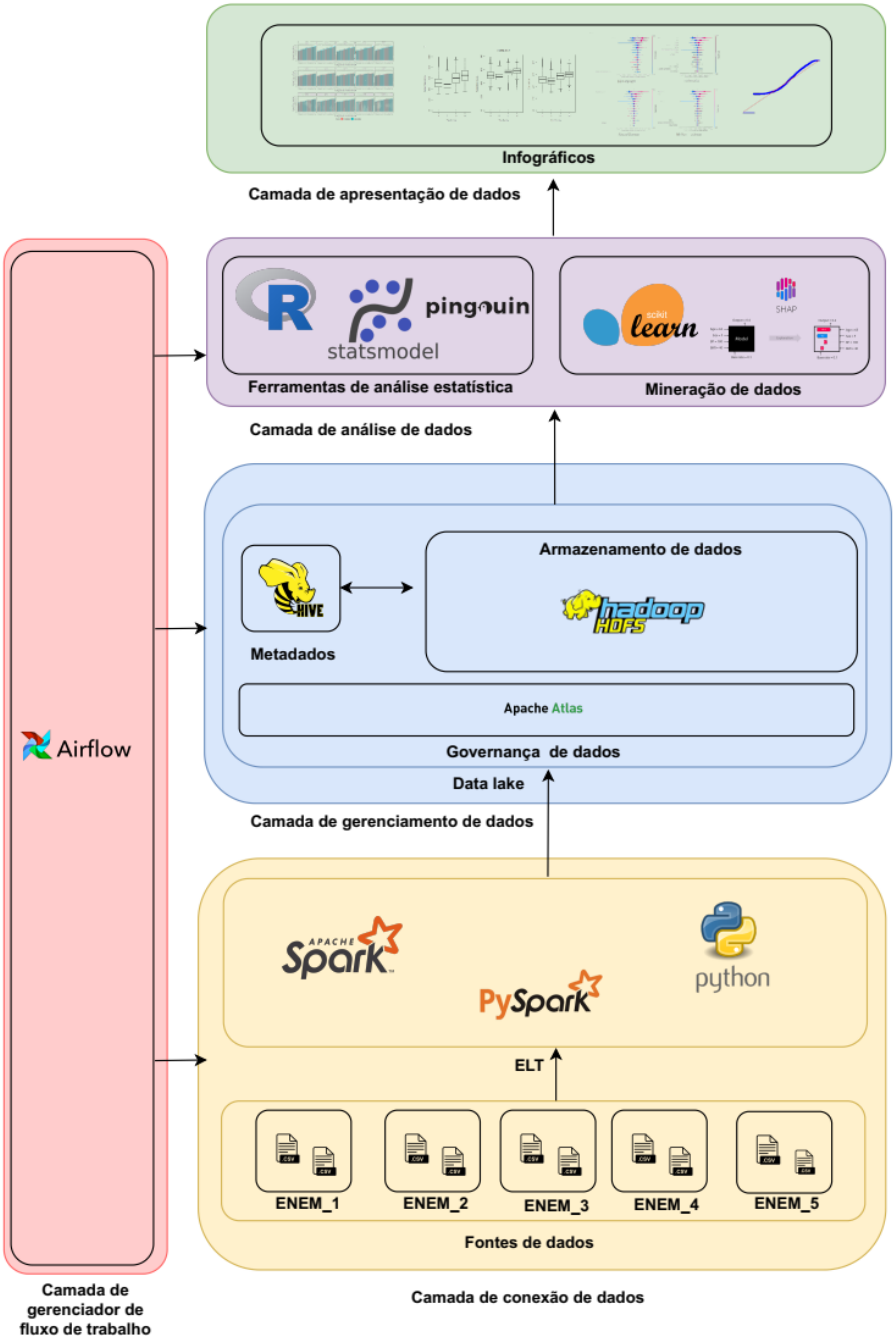
Os dados extraídos são armazenados no HDFS em formato nativo na área de dados brutos da **camada de armazenamento de dados** do *Data lake*. Os metadados são gerenciados em tabelas *Hive* pelo *Hive Metastore* (THUSOO *et al.*, 2009). O *Hive Metastore* é um catálogo do sistema que contém metadados sobre as tabelas armazenadas no *Hive*. Esses metadados são especificados durante a criação das tabelas e reutilizados toda vez que as tabelas são referenciadas. Além disso, emprega-se *Apache Atlas*[3] para a governança dados.

Na **camada de análise de dados** são utilizadas as ferramentas R[4], bibliotecas *statsmodel*[5] e *pingouin*[6] para a análise de informações estatísticas dos dados e provas de testes de hipóteses. Para a mineração dos dados, é empregada a biblioteca *scikit-learn* da linguagem *Python*. Além disso, o *framework* SHAP é usado para analisar as predições dos modelos. Os resultados obtidos são então investigados visualmente na **camada de apresentação de dados**. A visualização é feita por meio da exibição de infográficos e de *summary plots* e *force plots*.

Na **camada de gerenciador de fluxo de trabalho**, a automação das etapas do *processo ELT* até a camada de análise é realizada pelo gerenciador *Apache Airflow*[7]. *Apache Airflow* é uma plataforma *open source* que usa *Directed Acyclic Graphs* (DAGs) para criação, agendamento e

1 <https://spark.apache.org/>
2 <https://www.python.org/>
3 <https://atlas.apache.org/#/>
4 <https://www.r-project.org/>
5 <https://www.statsmodels.org/stable/index.html>
6 <https://pingouin-stats.org/>
7 <http://airflow.apache.org/>

Figura 10 – *Pipeline* para utilização de *Data lake* com ferramentas de análise estatística e mineração de dados com suporte de tecnologias de código aberto.



Fonte: Elaborada pelo autor.

monitoramento de fluxos de trabalho. As tarefas são escritas usando a linguagem de programação *Python*, a qual possui suporte nativo provido pelo *Airflow*.

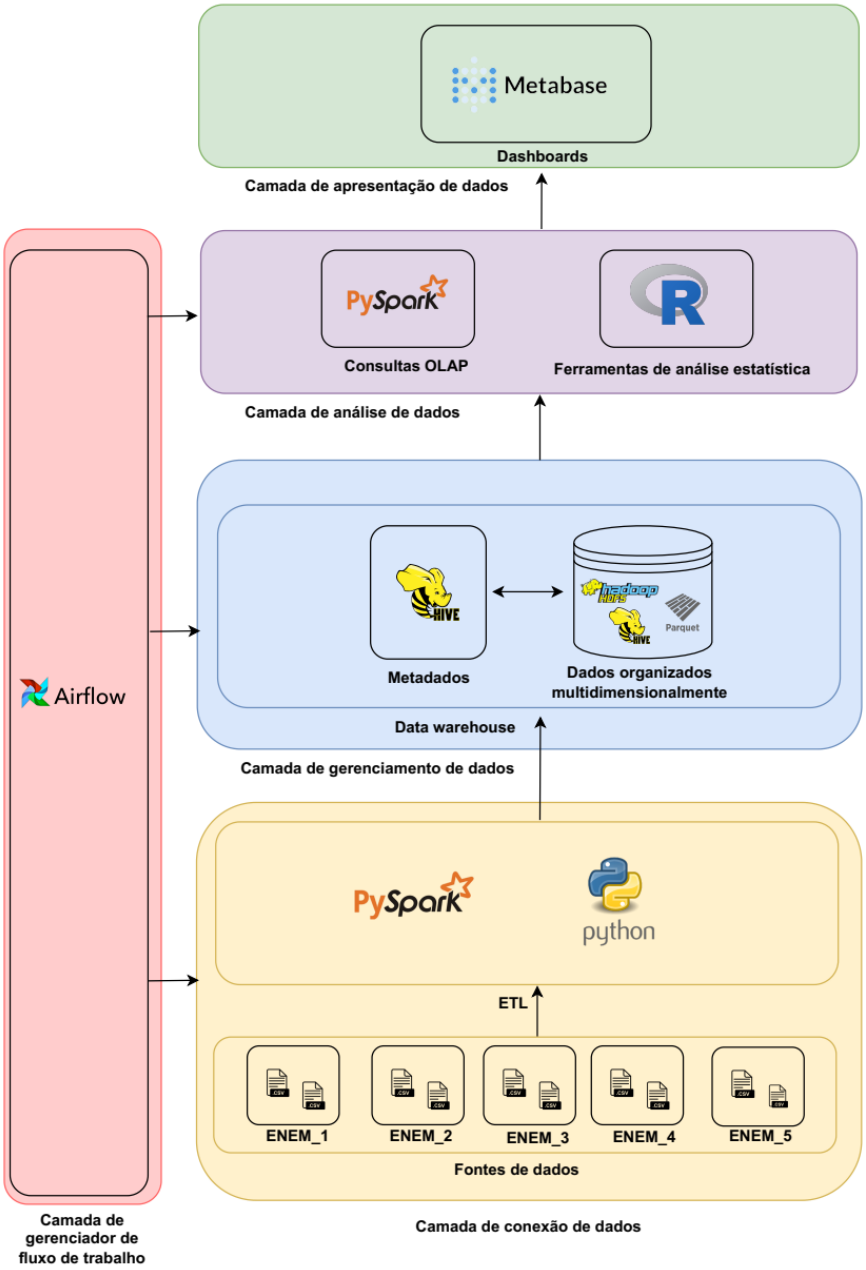
Outro exemplo de *pipeline* é ilustrado na Figura 11. Nesta figura é feita uma instanciação da arquitetura com o propósito explorar as funcionalidades do componente *Data warehouse*. As *fontes de dados* são as mesmas que foram usadas no *pipeline* anterior. O *processo ETL* é

realizado usando *Pyspark* e *Python*. O armazenamento dos dados é efetuado no sistema HDFS, usando o formato *parquet* com compressão *snappy*. *Parquet* é um formato de armazenamento de dados orientado a colunas. As tabelas do repositório de dados estruturados são geradas no *Apache Hive* e os metadados são gerenciados pelo *Hive Metastore*.

Na **camada de análise de dados**, as consultas OLAP são feitas usando *Pyspark* e *SparkSQL* e as análises estatísticas são efetuadas com a ferramenta *R*. Os resultados obtidos são visualizados com a ferramenta Metabase[8], presente na **camada de apresentação de dados**. Novamente, a automação das etapas do *processo ETL* até a camada de análise são realizadas na **camada de gerenciador de fluxo de trabalho** por meio do emprego do *Apache Airflow*.

[8] <https://www.metabase.com/>

Figura 11 – Pipeline para utilização de *Data warehouse* com consultas OLAP e ferramentas de análise estatística com suporte de tecnologias de código aberto.



Fonte: Elaborada pelo autor.

4.3 Considerações finais

Neste capítulo foi proposta uma arquitetura desenvolvida com a finalidade de apoiar o processo de tomada de decisão educacional. A arquitetura relaciona dados obtidos de fontes educacionais, em especial dados do ENEM, com um ambiente de processamento e armazenamento paralelo e distribuído e com o uso de técnicas de ciência de dados. A arquitetura foi descrita em termos de suas cinco camadas, conexão, gerenciamento, análise, apresentação de dados e gerenciador de fluxo de trabalho, e de seus componentes. Finalmente, foram descritos dois exemplos de instância da arquitetura que empregam tecnologias de código aberto.

Nos próximos capítulos, a arquitetura proposta é utilizada para prover suporte para a análise de dados do ENEM. São descritos vários cenários de uso da arquitetura proposta, de acordo com os objetivos específicos da análise. No primeiro cenário, descrito no Capítulo 5 são detalhados todos os processos realizados desde a camada de conexão até a camada de apresentação de dados com o objetivo de investigar e analisar os principais indicadores de desempenho dos participantes do ENEM de 2017 a 2020. São consideradas as quatro áreas de conhecimento do ENEM, ou seja, *Linguagens e Códigos*, *Matemática*, *Ciência da Natureza* e *Ciências Humanas*.

CAPÍTULO

5

PRINCIPAIS INDICADORES DE DESEMPENHO

Neste capítulo é detalhado o primeiro cenário de uso da arquitetura proposta. O objetivo consiste em investigar e analisar os principais indicadores de desempenho dos participantes do ENEM de 2017 a 2020. São consideradas as quatro áreas de conhecimento do ENEM, ou seja, *Linguagens e Códigos*, *Matemática*, *Ciência da Natureza* e *Ciências Humanas*. Cada área é analisada separadamente, desde que possui suas características e também um desempenho médio específico informado pelo Inep. Assim, o desempenho dos participantes de uma determinada área é comparado com o desempenho médio dessa área. Nas análises realizadas, são usados os classificadores árvore de decisão, SVM e MLP. A qualidade dos modelos de classificação é investigada considerando as seguintes métricas: *Accuracy, F1-Score, Precision, Recall* e Área Sob Curva. Adicionalmente, as predições geradas pela árvore de decisão são interpretadas usando o *framework* SHAP. A partir das análises realizadas, é feita uma avaliação da eficácia educacional brasileira, oferecendo suporte para direcionamentos voltados às políticas públicas educacionais.

As contribuições correspondentes aos resultados descritos neste capítulo encontram-se disponíveis no artigo em julgamento Noguera, Branco e Aguiar (2022). Na seção 5.1 é descrito como a arquitetura proposta é instanciada para oferecer suporte para as análises realizadas. Na seção 5.2 são discutidos os resultados obtidos. Na seção 5.3 é apresentada uma discussão sobre os resultados considerando aspectos relacionados aos estudos e políticas educacionais. O capítulo é finalizado na seção 5.4 com as considerações finais.

5.1 Instanciação da arquitetura proposta

A instância da arquitetura proposta empregada para oferecer suporte para as análises descritas neste capítulo faz referência ao *pipeline* para utilização de *Data lake* com ferramentas de mineração de dados ilustrada na Figura 10 (seção 4.2). Na camada de conexão de dados, as

fontes de dados utilizadas correspondem aos dados do ENEM para os anos 2017 até 2020.

Os microdados foram extraídos da página oficial do Inep e armazenados no HDFS em formato nativo na área de dados brutos na camada de armazenamento de dados do *Data lake*. Os conjuntos de dados brutos referentes aos candidatos inscritos nos anos de 2017 a 2020 possuíam 6.731.341, 5.513.747, 5.095.270 e 5.783.109 instâncias, respectivamente. Foram considerados apenas os participantes não treineiros que responderam: (i) às questões de todas as áreas de conhecimento; e (ii) todas as questões presentes no questionário socioeconômico. Assim, os conjuntos de dados gerados na fase de ELT para os anos de 2017 a 2020 foram compostos por 3.990.681, 3.434.771, 3.174.308 e 2.242.169 instâncias, respectivamente.

Na área de processamento na camada de armazenamento de dados do *Data lake* foram realizados diversos processos. O primeiro processo aplicado foi a rotulação das instâncias, o qual ocorreu da seguinte forma. A nota do participante foi rotulada como *desempenho bom* em uma área de conhecimento quando foi igual ou maior que a nota média dessa área informada pelo Inep e o participante não obteve nota zero na *Redação*. Caso contrário, foi rotulada como *desempenho ruim*. Os dados foram rotulados a partir de um *script*, sendo que cada instância foi associada a uma única classe. A nota média informada pelo Inep para cada área de conhecimento em cada ano é descrita na Tabela 7.

Tabela 7 – Média informada pelo Inep para cada área de conhecimento de 2017 a 2020.

Área de conhecimento/Ano	2017	2018	2019	2020
Linguagens, Códigos e suas tecnologias	510,2	526,9	520,9	523,98
Matemática e suas tecnologias	518,5	535,5	523,1	520,73
Ciências da Natureza e suas tecnologias	510,6	493,8	477,8	490,39
Ciências Humanas e suas tecnologias	519,3	569,2	508,0	511,64

Fonte: Elaborada pelo autor.

Após rotular as instâncias, foi aplicado o método embutido de *Random forests* para selecionar as características mais importantes dos microdados. O número de características de 2017 a 2020 foram 136, 137, 137 e 76, respectivamente. A importância das características obtidas segue a seguinte ordem: nota de cada área de conhecimento e *Redação*, idade do participante, renda mensal familiar, possuir carro, ocupação da mãe, possuir computador, dependência administrativa da escola, acesso à internet, ocupação do pai, possuir telefone celular, sexo e estado civil. Todas as atividades relacionadas à extração de dados, rotulagem de dados e seleção de características foram realizadas usando *Apache Spark*.

Na sequência, os tipos de dados alfanuméricos foram transformados em tipos de dados numéricos. Por exemplo, o gênero feminino foi transformado em 0 e o gênero masculino foi transformado em 1. Em seguida, foi atribuída a mediana de valores para os dados incompletos relacionados a uma determinada característica. Em todos os anos, foram encontradas classes desbalanceadas na área de conhecimento de *Matemática*. Adicionalmente, no ano de 2018,

também foram encontradas classes desbalanceadas nas áreas de conhecimento de *Ciência da Natureza* e *Ciências Humanas*. Em todos os casos, a classe maioritária continha cerca de 57% das amostras correspondentes aos participantes com notas abaixo da média. O problema foi solucionado usando o método de subamostragem *Tomek links*. Os processos de transformações e balanceamento de classes foram realizados com ajuda de pacotes da linguagem *Python*. Os resultados finais foram armazenados na área de acesso aos dados do *Data lake*.

Por fim, na camada de análise, com ajuda da biblioteca *scikit-learn* da linguagem *Python*, foram utilizados os modelos de classificação de arvore de decisão, SVM e MLP para fornecer uma ampla investigação e apoiar a predição de *desempenho bom* ou *desempenho ruim* dos participantes. Os hiperparâmetros dos modelos, quando necessários, foram selecionados automaticamente via *Random Search*. Foram verificadas as métricas de *Accuracy, F1-score, Precision, Recall* e *AUC* para avaliar a qualidade dos classificadores. Além disso, o *framework* SHAP foi utilizado para analisar as predições do modelo, que foram investigadas visualmente por meio de *summary plots* e *force plots*.

5.2 Resultados e discussões

Nesta seção são detalhados os resultados dos modelos de classificação para cada área de conhecimento, considerando os anos de 2017 a 2020. Primeiramente, na seção 5.2.1, os dados são analisados e avaliados considerando os modelos de árvore de decisão, SVM e MLP. São aplicadas diversas métricas separadamente, permitindo a investigação de aspectos específicos de desempenho relacionados a cada área. Em seguida, são destacados os resultados da árvore de decisão devido à sua interpretabilidade e ao tempo de execução excessivo dos demais modelos. Para os modelos SVM e MLP, foi executado o interpretador com um método agnóstico de modelo e, mesmo com um subconjunto de amostra, não foi possível interpretar os resultados dos modelos devido ao tempo de execução. A execução foi interrompida após uma semana. Os resultados de interoperabilidade da árvore de decisão são discutidos usando gráficos de *summary plots* (seção 5.2.2) e *force plots* (seção 5.2.3). Como os resultados obtidos considerando os anos de 2017 a 2020 foram muito semelhantes, nesta seção são descritos e discutidos apenas os resultados de desempenho para o ano de 2020.

5.2.1 Modelos preditivos

Os valores para o desempenho dos participantes no ENEM de 2020 para cada área de conhecimento usando os modelos árvore de decisão, SVM e MLP são descritos nas Tabelas 8, 9, 10 e 11. Para este ano, os resultados variaram de 78% a 81%, demonstrando que os modelos aplicados garantem percentual adequado de predições corretas (*Accuracy*), desempenho (*F-score*) e rótulos (*Precision* e *Recall*). Os resultados obtidos para AUC variaram de 0,79 a 0,9, mostrando a eficácia dos classificadores. A inexistência de variação excessiva de valores entre as métricas

permite a análise das diferentes áreas do conhecimento de forma indiscriminada.

Tabela 8 – Métricas de desempenho para *Linguagens e Códigos*.

	Árvore de decisão	SVM	ANN
Accuracy (%)	79	80	81
F1-score (%)	79	80	80
Precision (%)	79	80	80
Recall (%)	79	80	80
AUC	0,87	0,89	0,89

Fonte: Elaborada pelo autor.

Tabela 9 – Métricas de desempenho para *Matemática*.

	Árvore de decisão	SVM	ANN
Accuracy (%)	78	79	79
F1-score (%)	78	79	79
Precision (%)	78	79	79
Recall (%)	78	79	80
AUC	0,86	0,88	0,88

Fonte: Elaborada pelo autor.

Tabela 10 – Métricas de desempenho para *Ciência da Natureza*.

	Árvore de decisão	SVM	ANN
Accuracy (%)	79	80	80
F1-score (%)	78	79	80
Precision (%)	79	80	80
Recall (%)	78	79	79
AUC	0,87	0,88	0,88

Fonte: Elaborada pelo autor.

Tabela 11 – Métricas de desempenho para *Ciências Humanas*.

	Árvore de decisão	SVM	ANN
Accuracy (%)	79	81	81
F1-score (%)	79	81	81
Precision (%)	79	81	81
Recall (%)	79	81	81
AUC	0,79	0,89	0,90

Fonte: Elaborada pelo autor.

5.2.2 Interpretabilidade do modelo de árvore de decisão com summary plot

A distribuição do impacto que cada característica tem na saída do modelo de árvore de decisão usando o gráfico *summary plot* para cada área de conhecimento e considerando o ano de 2020 é ilustrada na Figura 12. Para a produção da figura, foram obtidos os valores de SHAP para as predições negativas e positivas, ou seja, predições relacionadas à classe rotulada como *desempenho ruim* e *desempenho bom*, respectivamente.

O eixo y representa as características ordenadas por importância decrescente, da mais importante na parte superior para a menos importante na parte inferior. O eixo x tem uma escala para valores SHAP com uma linha vertical em zero. O valor do impacto é exibido horizontalmente, variando de -0,6 a 0,6. Um ponto na posição do gráfico representa um valor SHAP para uma característica específica de uma instância. Quanto mais distantes os pontos estiverem à direita e à esquerda, mais forte é a influência da característica para a predição das classes positiva e negativa, respectivamente. Cada ponto no gráfico representa uma amostra e cada cor diferencia o valor da característica relativa entre as amostras. A escala de cores varia do azul ao vermelho, representando os valores altos e baixos, e se acumulam verticalmente para mostrar a densidade (LUNDBERG *et al.*, 2020).

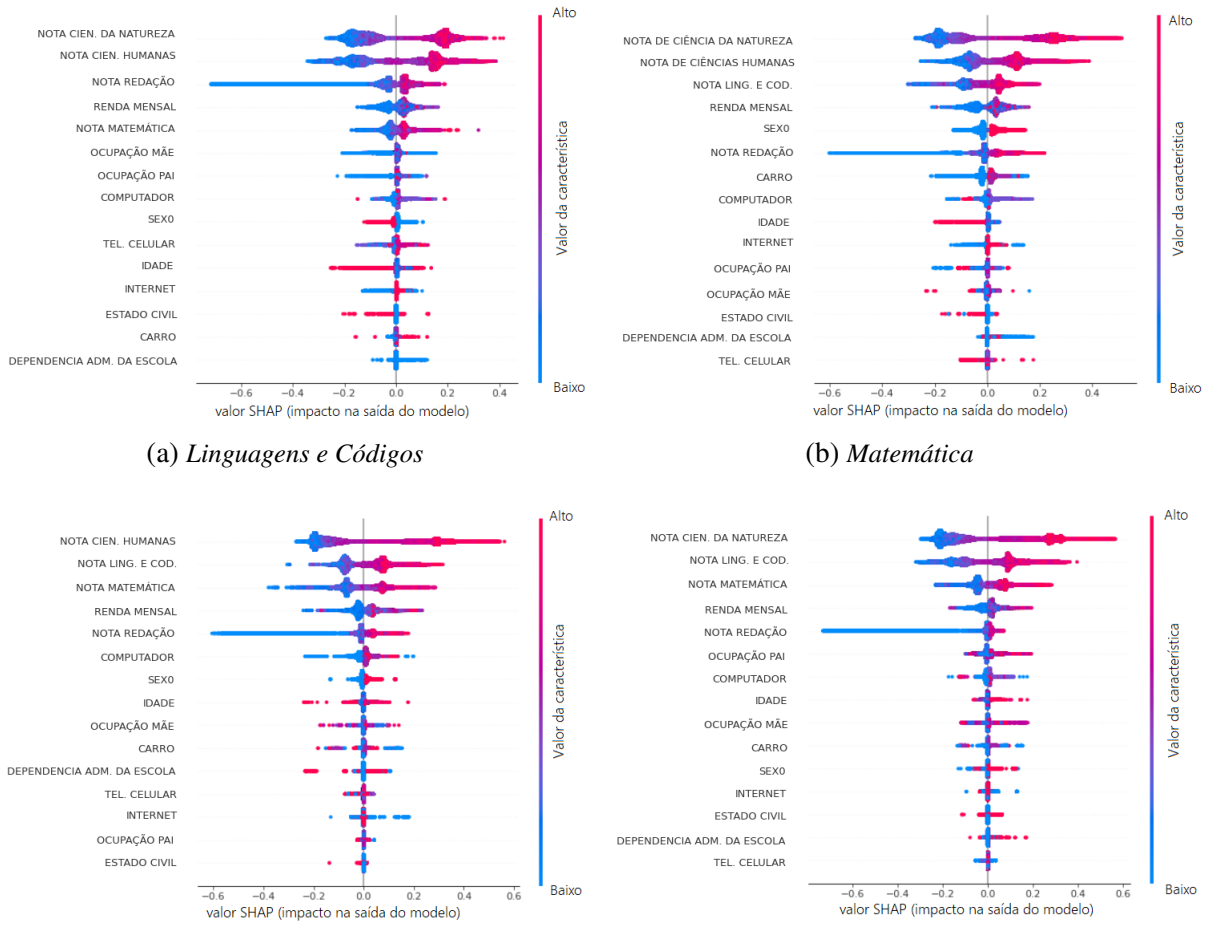
De acordo com os resultados exibidos, é possível especificar os seguintes *insights*:

- Para todas as áreas de conhecimento, a nota da *Ciência da Natureza* é a mais influente na predição, assumindo valores de influência iguais ou superiores a 0,4 para predição positiva. Adicionalmente, os notas relacionadas a *Ciências Humanas*, *Matemática*, *Linguagens e Códigos* e *Redação* também estão entre as de maior influência para a predição. Outras características que também exercem influência na predição do modelo são *renda mensal familiar*, *sexo* e *acesso a um computador*.

- Com relação à nota da *Redação*, pode-se observar que as caudas são mais longas à esquerda, e não à direita. Isso significa que valores baixos para essas notas podem aumentar significativamente a classificação do participante como *desempenho ruim*. Em todos os gráficos da Figura 12, a linha azul desta característica distingue-se das demais por ser a mais próxima do valor de -0,6.

- O fato do participante apresentar maiores notas em *Ciência da Natureza*, *Ciências Humanas*, *Redação*, *Matemática* e *Linguagens e Códigos* e ser do *sexo* masculino (Figura 12b e Figura 12c) aumenta a predição da amostra pertencer à classe positiva. Considerando o atributo *sexo*, as cores vermelha e azul indicam valores masculino e feminino, respectivamente.

Para as características não mencionadas, os valores SHAP são agrupados em torno de



(a) *Linguagens e Códigos* (b) *Matemática*

(c) *Ciência da Natureza* (d) *Ciências Humanas*

Figura 12 – *Summary plot* que resume a predição de desempenho dos participantes (eixo x) para as características mais importantes do ENEM (eixo y), considerando as diferentes áreas de conhecimento:(a) *Linguagens e Códigos*; (b) *Matemática*; (c) *Ciência da Natureza*; e (d) *Ciências Humanas*.

zero. Como resultado, o modelo não consegue distinguir e classificar a amostra com esses valores de características como uma classe positiva ou negativa. Como esses valores não afetam a saída do modelo de árvore de decisão, eles não são tão influentes quanto as características discutidas anteriormente.

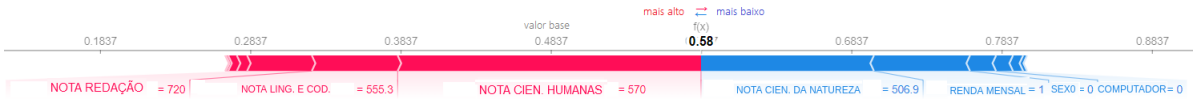
5.2.3 Interpretabilidade do modelo de árvore de decisão com force plot

O gráfico de *force plot* mostra as características do modelo responsáveis por predizer o *desempenho*. Na Figura 13, é ilustrada a predição de uma única amostra aleatória para o modelo de árvore de decisão. O valor base é o valor médio no conjunto de treinamento que seria predito se nenhuma característica da saída do modelo atual fosse conhecida. O gráfico representa em vermelho e azul as características que forçam a predição a ser positiva e negativa, respectivamente. O número em negrito, colocado onde as cores vermelha e azul se cruzam,

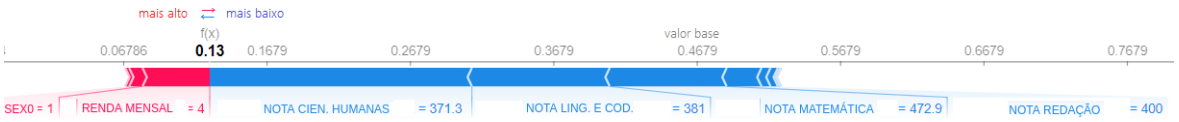
é a *accuracy* da predição local. Notas mais altas e mais baixas levam o modelo a predizer *desempenho bom* e *desempenho ruim*, respectivamente.

De acordo com os resultados exibidos, é possível especificar vários *insights*. Por exemplo, na Figura 13a, a *renda mensal* pertencente à categoria 6 (ou seja, de R$2994,01 a R$3992,00), nota de *Matemática* igual a 601,8, nota de *Ciências Humanas* igual a 610,3 e nota de *Ciência da Natureza* igual a 619,2 contribuem para uma predição positiva para a área de conhecimento de *Linguagens e Códigos*. Por outro lado, na Figura 13d, notas baixas em *Ciência da Natureza*, *Linguagens e Códigos* e *Matemática*, além do participante não ter *carro* e *computador*, contribuem para uma predição negativa para a área de conhecimento de *Ciências Humanas*. Outros *insights* obtidos de forma semelhante aos descritos também podem ser obtidos pela análise dos resultados.

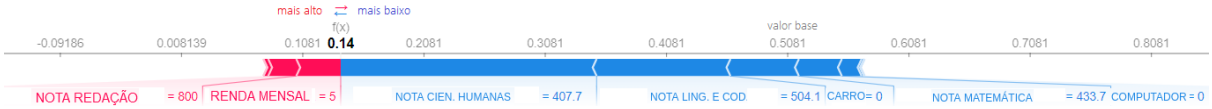
(a) Linguagens e Códigos e suas tecnologias



(b) Matemática e suas tecnologias



(c) Ciências da Natureza e suas tecnologias



(d) Ciências Humanas e suas tecnologias

Figura 13 – *Force plot* que representa a contribuição da característica para uma única predição.As características que estão em vermelho são as que fazem a previsão ter um valor positivo, enquanto as características que estão em azul fazem com que a previsão tenha um valor negativo.

5.3 Análise dos resultados com foco na educação

Os resultados obtidos demonstraram que os principais indicadores de desempenho são *Linguagens e Códigos*, *Matemática*, *Ciências Humanas*, *Ciência da Natureza*, *Redação*, *sexo*, *renda mensal* e *acesso a computador*. Nesta seção são discutidos aspectos educacionais relacionados a esses indicadores. Na seção 5.3.1, os indicadores de desempenho obtidos no presente cenário de uso da arquitetura são avaliados frente a relatórios referentes a estudos educacionais providos pelo IDEB, PISA e a Organização das Nações Unidas para a Educação, a

Ciência e a Cultura (UNESCO). Na seção 5.3.2, os indicadores são utilizados na discussão de políticas educacionais.

5.3.1 Relacionando os resultados obtidos com estudos educacionais

Existem diversos estudos educacionais que indicam problemas relacionados à formação de estudantes. Nesta seção é feita uma relação entre os indicadores de desempenho obtidos e os resultados descritos em relatórios nacionais e internacionais de larga escala providos pelo IDEB, PISA e UNESCO. O objetivo consiste em corroborar a importância dos indicadores identificados. Os últimos resultados providos pelo IDEB, referentes ao ano de 2019, atingiram 4,2 pontos em uma escala de 0 a 10. Esses resultados são muito baixos, dado que o desafio para as escolas era atingir 5,0 pontos (IDEB, 2022). Em relação aos últimos resultados do PISA, coletados em 2018, os alunos brasileiros apresentam baixa proficiência em Leitura, Matemática e Ciências, quando comparados aos alunos de outros 78 países participantes (INEP, 2022b). Os principais indicadores de desempenho obtidos corroboram os resultados dessas duas avaliações em larga escala. A influência positiva e negativa dos indicadores demonstra que o problema não se inicia no ensino médio. Os indicadores são impactados por problemas acumulados em etapas anteriores, como os anos iniciais e finais do ensino fundamental.

Adicionalmente, existe uma correlação entre o desempenho do aluno e o nível socioeconômico (MENDES; KARRUZ, 2015; GIL, 2021; GUSTAFSSON; NILSEN; HANSEN, 2018; THOMSON, 2018), apontada pela *renda mensal* da família. De acordo com Massi (2017) e Koza e Melis (2017), os participantes com pais ricos podem investir mais na educação dos filhos com materiais didáticos, cursos preparatórios, livros e jogos didáticos. Como consequência, essas crianças geralmente obtêm os melhores resultados. Os participantes cujas famílias têm condições financeiras menos favoráveis precisam ajudar os pais nas tarefas domésticas ou ganhar dinheiro. Com isso, a possibilidade desses participantes terem acesso ao computador diminui.

O exame do ENEM também é considerado e divulgado como uma “prova de leitura”. Suas questões visam promover relações interdisciplinares e evitar o conhecimento como mera repetição. O participante que souber ler e interpretar estas questões tende a ter *desempenho bom* (WENCESLAU *et al.*, 2014). Essa relação positiva entre hábitos de leitura e desempenho já foi identificada em outros estudos com delineamento quantitativo (ALVES *et al.*, 2013; GONÇALVES, 2015; INOUK; SUZANNE; JELLE, 2017).

Além disso, a baixa proficiência em Leitura, identificada pelo PISA, pode estar diretamente relacionada ao baixo desempenho na *Redação*, conforme corroborado pelos estudos descritos em Juriati, Ariyanti e Fitriana (2018) e Mailis, Delfi e Erni (2018). De fato, as instituições de ensino têm aumentado o uso da nota da *Redação* como fator principal para o ingresso no ensino superior público. Por exemplo, o participante que obtiver zero na *Redação* não pode participar do Sisu e do Prouni, independentemente de receber boas notas nas áreas de conhecimento (SOUZA, 2020).

Os resultados da investigação realizada também mostraram que os homens superam as mulheres na área de conhecimento de *Matemática*, o que também é vislumbrado nos resultados do PISA. Os padrões de gênero nos interesses de matemática e ciências surgem na primeira infância, se desenvolvem ao longo do tempo e, finalmente, refletem na seleção de curso (STEEGH *et al.*, 2019). Além disso, de acordo com o relatório da UNESCO, as meninas não demonstram o mesmo interesse pelas ciências exatas que os meninos (UNESCO, 2018). As meninas tendem a preferir as *Ciências Humanas* e, com isso, acabam optando por cursos relacionados à educação, saúde e serviços sociais (OECD, 2019). Esses aspectos podem influenciar as escolhas das mulheres para estudar e seguir a carreira em ciências exatas, bem como levar a um desequilíbrio de gênero nos cursos da ciência, tecnologia, engenharia e matemática (STEM).

5.3.2 Relacionando os resultados obtidos com políticas educacionais

Nesta seção é feita uma relação entre os indicadores de desempenho obtidos e a existência de políticas educacionais relacionadas. O objetivo consiste em mostrar se já existem políticas educacionais desenvolvidas e também levantar questionamentos que devem ser considerados quando proposta de novas políticas ou melhoria de políticas já existentes.

Um problema relevante que impacta negativamente no desempenho dos alunos é o apoio financeiro. Famílias de baixa renda que não têm condições de financiamento favoráveis farão menos investimentos no capital humano de seus filhos. Essas crianças terão uma renda futura menor, perpetuando sua condição de pobreza (CHECCHI, 2006; LEE; LEE, 2018). Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), a desigualdade entre ricos e pobres cresce cada vez mais (IBGE, 2022).

Existem várias políticas públicas que visam estabelecer a igualdade de acesso às escolas com educação de qualidade. Entre eles, cita-se o Plano Nacional de Educação (MEC, 2022a), Programa Educação na Prática (MEC, 2022b), Programa Ensino Médio Inovador (MEC, 2022c), cotas sociais e raciais (MEC, 2022d; BERNADO; SILVA, 2019; SIQUEIRA; LARA; LIMA, 2020; COUTINHO; BORGES, 2018). No entanto, os efeitos dessas políticas ainda são insuficientes, refletindo nos resultados de avaliações em larga escala. As escolas devem ter diversos materiais de leitura, e os professores devem planejar atividades regulares de leitura, expondo os alunos aos mais variados gêneros textuais que circulam socialmente (CUNHA; SANTOS, 2019).

O desequilíbrio entre homens e mulheres na *Matemática* impacta no número de mulheres que ingressam e se formam em ciências exatas. Apesar do grande número de cursos e vagas nessa área, esse número está diminuindo (HENN, 2014; GARCÍA; DÍAZ; GARCÍA, 2019). Há também desigualdade de gênero na publicação acadêmica, conforme identificado no exaustivo trabalho de Huang *et al.* (2020), que fornece uma análise bibliométrica das publicações acadêmicas, reconstruindo o histórico completo de publicação de mais de 1,5 milhão de autores identificados por gênero em 83 países.

Vários estudos investigam o desequilíbrio de gênero no STEM em todo o mundo. Esses estudos analisam dados dos Estados Unidos (KUGLER; TINSLEY; UKHANEVA, 2017), Arábia Saudita (PILOTTI, 2021), Reino Unido (PENNINGTON *et al.*, 2021), Hong Kong (TAM; CHAN; LAI, 2020), Chile (VILLASEÑOR *et al.*, 2020), Itália (BERRA; CAVALETTO, 2020), Gana (ANSONG *et al.*, 2020) e Polônia (ZAWISTOWSKA; SADOWSKI, 2019). O Brasil não se comporta de forma diferente, principalmente nas áreas tecnológicas (SANTOS, 2018; ANDRADE, 2019b). Existem diversas iniciativas no Brasil e também no mundo com o objetivo de mudar esse cenário a médio e longo prazo, como Meninas Digitais (MACIEL; BIM; FIGUEIREDO, 2018; GUZMAN *et al.*, 2020), Technovation Girls (TORRES, 2015), She++ (SHE++, 2015), PyLadies (PYLADIES, 2014), Grace Hopper Celebration (TOWNSEND; HARRIGER, 2019), Hour of Code (TOWNSEND; HARRIGER, 2019) e Code Girl (CODEGIRL, 2022).

Estudos focados em habilidades cognitivas confirmam que a compreensão de leitura e o conhecimento de vocabulário estão correlacionados com a *Redação* (JUN; BIN, 2008; ALLEN *et al.*, 2014; BINDER *et al.*, 2017; SOUZA, 2020). Nos resultados obtidos nas análises descritas neste capítulo, a nota da *Redação* é a característica que mais influencia uma amostra a ser classificada como classe negativa. Com base nessas constatações, surgem algumas questões que devem ser investigadas, como: "Os participantes compreendem os temas da *Redação*?"; "Como melhorar as práticas de ensino-aprendizagem na *Redação*?"; "Há ineficiência no ensino das práticas de leitura e escrita na educação básica?". Os alunos devem ser cativados no hábito de leitura e escrita, construindo sua autoeficácia, curiosidade e envolvimento com os textos.

5.4 Considerações finais

Neste capítulo foi descrito o primeiro cenário de uso da arquitetura proposta no Capítulo 4. Foram identificados e analisados os principais indicadores de desempenho dos participantes do ENEM nos anos de 2017 a 2020 com o objetivo de avaliar a eficácia educacional.

Foram utilizados os modelos de classificação de árvore de decisão, SVM e MLP para investigar as notas da *Redação* e das 180 questões de múltipla escolha relacionadas às áreas de conhecimento de: (i) *Linguagens e Códigos*; (ii) *Matemática*; (iii) *Ciência da Natureza*; e (iv) *Ciências Humanas*. Além disso, foi considerado o questionário preenchido pelos participantes detalhando aspectos pessoais, socioeconômicos, familiares, educacionais e de trabalho. Foram verificadas as métricas *Accuracy, F1-score, Precision, Recall* e *Area Under Curve* para avaliar a qualidade dos modelos de classificação. Finalmente, foi utilizado o *framework* SHAP para interpretar as predições fornecidas pelo modelo de árvore de decisão. De acordo com os resultados obtidos, os principais indicadores de desempenho são *Linguagens e Códigos*, *Matemática*, *Ciências Humanas*, *Ciência da Natureza*, *Redação*, *sexo*, *renda mensal* e *acesso a computador*.

Para validar a importância dos indicadores de desempenho identificados, foi feita uma análise desses indicadores frente aos relatórios nacionais e internacionais de larga escala providos

pelo IDEB, PISA e UNESCO. Os indicadores também foram utilizados para a discussão de políticas públicas educacionais existentes. Conclui-se que os indicadores corroboram com problemas identificados pelos relatórios. Adicionalmente, também conclui-se que os indicadores devem ser utilizados como base para a extensão de políticas já existentes e para a criação de novas políticas educacionais.

No próximo capítulo, Capítulo 6, é descrito o segundo cenário de uso da arquitetura proposta, o qual tem como objetivo investigar o desempenho dos participantes do ENEM considerando a temática “gêneros e suas nuances na tecnologia da informação”.

CAPÍTULO

6

GÊNEROS E SUAS NUANCES NO ENEM

Neste capítulo é descrito o segundo cenário de uso da arquitetura proposta. O objetivo consiste em investigar o desempenho dos participantes do ENEM na temática "gêneros e suas nuances na tecnologia da informação". São consideradas as três áreas de conhecimento relacionadas às ciências exatas presentes no ENEM: *Matemática*, *Ciência da Natureza* e *Linguagens e Códigos*. Também são considerados diferentes fatores de análise, de acordo com dois direcionamentos. O primeiro direcionamento refere-se à investigação do desempenho geral dos participantes de sexo feminino e masculino para cada área de conhecimento das ciências exatas no ENEM, considerando o período de 2013 a 2017 e os seguintes fatores de análise: região, cor/raça, tipo de escola de ensino médio e renda salarial mensal. O segundo direcionamento diz respeito à investigação do desempenho dos participantes de sexo feminino e masculino cuja nota média foi igual ou superior à média informada pelo Inep em cada área de conhecimento das ciências exatas no ENEM, considerando o ano de 2017 e, para cada região do Brasil, suas unidades federativas.

As contribuições correspondentes aos resultados descritos neste capítulo encontram-se disponíveis nos artigos publicados Noguera, Branco e Ciferri (2019) e Noguera, Branco e Ciferri (2020). Na seção 6.1 é descrito como a arquitetura proposta é instanciada para oferecer suporte para as análises realizadas. Na seção 6.2 são discutidos os resultados obtidos. O capítulo é finalizado na seção 6.3 com as considerações finais.

6.1 Instanciação da arquitetura proposta

A instância da arquitetura proposta empregada para oferecer suporte para as análises descritas neste capítulo faz referência ao *pipeline* para utilização de *Data lake* com ferramentas de análise estatística ilustrado na Figura 10 (seção 4.2). Na camada de conexão de dados, as fontes de dados utilizadas correspondem aos dados do ENEM para os anos 2013 até 2017. O volume de dados brutos dessas fontes para os anos de 2013 a 2017 é, respectivamente, 5.007.934, 8.724.248, 7.746.427, 8.627.367 e 6.731.341 instâncias correspondentes ao número de inscritos.

Os microdados foram extraídos da página oficial do Inep e armazenados no HDFS em formato nativo na área de dados brutos na camada de armazenamento de dados do *Data lake*. Foram especificadas condições para garantir que as amostras para as análises descritivas fossem constituídas por todos os participantes do ENEM que satisfizesse certos critérios, conforme descrito a seguir:

- A amostra para a análise que considera o desempenho dos participantes detalhada na seção 6.2.1 foi constituída por todos os participantes do ENEM que satisfizeram aos seguintes critérios: (i) ter preenchido o microdado referente ao sexo; (ii) ter informado a região do município de residência; (iii) ter informado a cor/raça; (iv) ter preenchido a categoria de renda mensal de sua família; e (v) ter participado de todas as provas da área de ciências exatas, mesmo que tenha obtido 0 (zero) em uma ou mais dessas provas. Assim, os conjuntos de dados gerados na fase de ELT para os anos de 2013 a 2017 foram compostos por 5.007.934, 5.949.253, 5.604.905, 5.818.446 e 4.426.755 instâncias, respectivamente.

- A amostra para a análise atendendo as médias informadas pelo Inep detalhada na seção 6.2.2 foi constituída por todos os participantes do ENEM no ano 2017 que satisfizeram aos seguintes critérios: (i) ter preenchido o microdado referente ao sexo; (ii) ter informado a região e a unidade federativa do município de residência; e (iii) ter participado e atingido a média igual ou superior à média informada pelo Inep em todas as áreas de conhecimento da ciências exatas. As médias gerais informadas pelo Inep para cada área de conhecimento são: (i) *Matemática*: 518,5; (ii) *Ciência da Natureza*: 510,6; e (iii) *Linguagens e Códigos*: 510,2. Após a aplicação desses critérios, o conjunto de dados foi composto por 1.042.695 instâncias.

As amostras foram armazenadas na área de acesso aos dados na camada de armazenamento de dados do *Data lake*. As mesmas foram constituídas pelas variáveis *co_municipio_residencia, tp_sexo, tp_cor_raca, tp_escola, nu_nota_cn, nu_nota_lc, nu_nota_mt* e *q006*. Os campos *co_municipio_residencia, tp_sexo, tp_cor_raca, tp_escola, nu_nota_cn, nu_nota_lc, nu_nota_mt* correspondem à região de residência do participante, sexo, cor/raça, tipo de ensino médio, notas de *Ciência da Natureza*, *Linguagens e Códigos* e *Matemática*. O campo *q006* corresponde à renda mensal salarial. Todos os campos foram especificados com base no tamanho e tipo de dados informados pelos microdados.

Nas análises foram incluídos todos os participantes, inclusive os treineiros, além de todas as cores de cadernos das provas. O processo ELT foi realizado usando a ferramenta *PySpark* e os metadados do *Data lake* foram gerenciados pelo *Hive Metastore*. Além disso, *Apache Atlas* foi utilizado para governança de dados.

Na camada de análise de dados foi utilizada a ferramenta *R* para a extração das informações estatísticas dos dados e a exibição dos infográficos. Para tanto, foram utilizadas as seguintes

bibliotecas: *dplyr, ggplot2, string e gridExtra*. A automação das etapas do processo ELT até a camada de análise foi realizada com o gerenciador *Apache Airflow*. As tarefas foram escritas usando a linguagem de programação *Python*, pois é suportada nativamente pelo *Apache Airflow*.

6.2 Resultados e discussões

Nesta seção são descritos os resultados obtidos na análise do desempenho dos participantes de sexo feminino e masculino no ENEM, considerando diferentes anos de realização do exame e vários fatores de desempenho, bem como as áreas de conhecimento das ciências exatas. Na seção 6.2.1, discute-se o desempenho geral dos participantes, enquanto que na seção 6.2.2 discute-se o desempenho dos participantes cuja nota média for igual ou superior à média informada pelo Inep em cada área de conhecimento.

6.2.1 Desempenho dos participantes do ENEM

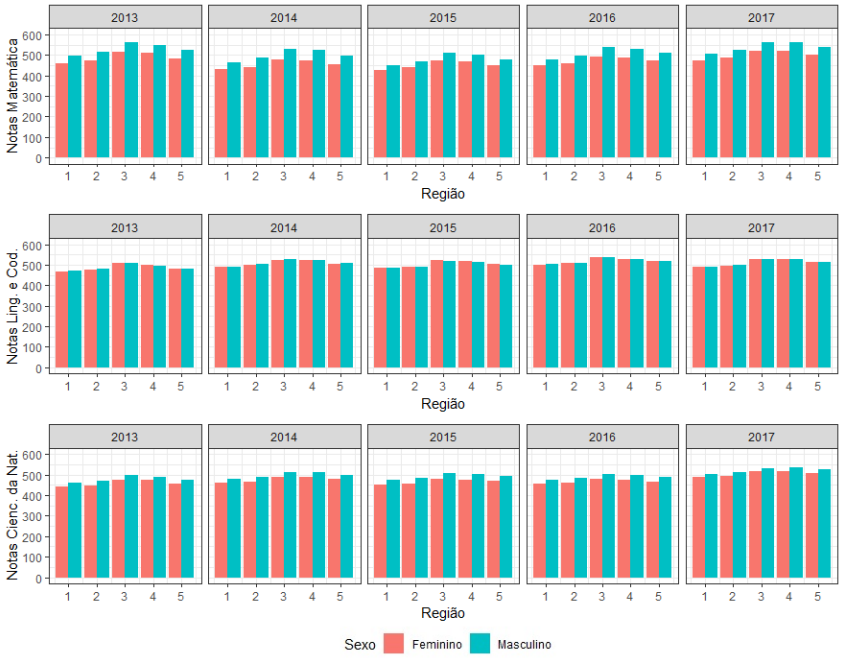
Nos resultados descritos nesta seção, são considerados os seguintes fatores de análise: (i) região; (ii) cor/raça; (iii) tipo de escola de ensino médio; e (iv) renda salarial mensal. Também é considerado um período de 5 anos, abrangendo dados de 2013 a 2017. O desempenho dos participantes foi definido como a média aritmética das notas para cada área de conhecimento das ciências exatas por sexo.

Em geral, os resultados da análise mostraram que o desempenho dos participantes do sexo masculino foi ligeiramente superior ao das participantes do sexo feminino em *Matemática* e *Ciência da Natureza*. Em *Linguagens e Códigos*, o desempenho de ambos os sexos foi quase equivalente. Os resultados detalhados são apresentados e discutidos a seguir.

Na Figura 14 é ilustrado o desempenho por região. Pode-se perceber que participantes das regiões Sudeste, Sul e Centro-Oeste, nessa ordem, obtiveram melhores notas que as regiões Norte e Nordeste, independente do sexo. O desempenho por cor/raça é ilustrado na Figura 15. Nessa figura, foi usada a classificação aplicada no ENEM nos anos de 2013, 2014 e 2017. Nos anos de 2015 e 2016, os microdados também incluíram a categoria (6) não dispõe de informação. No processo de limpeza dos dados, participantes dessa categoria foram agrupados com os participantes da categoria (0) não declarado. Em todos os anos analisados, os participantes de cor/raça branca obtiveram as melhores notas em *Matemática* e *Ciência da Natureza*. Em *Linguagens e Códigos*, independente da cor/raça, as notas foram similares para todos os participantes. Por fim, os participantes indígenas obtiveram as notas mais baixas das ciências exatas.

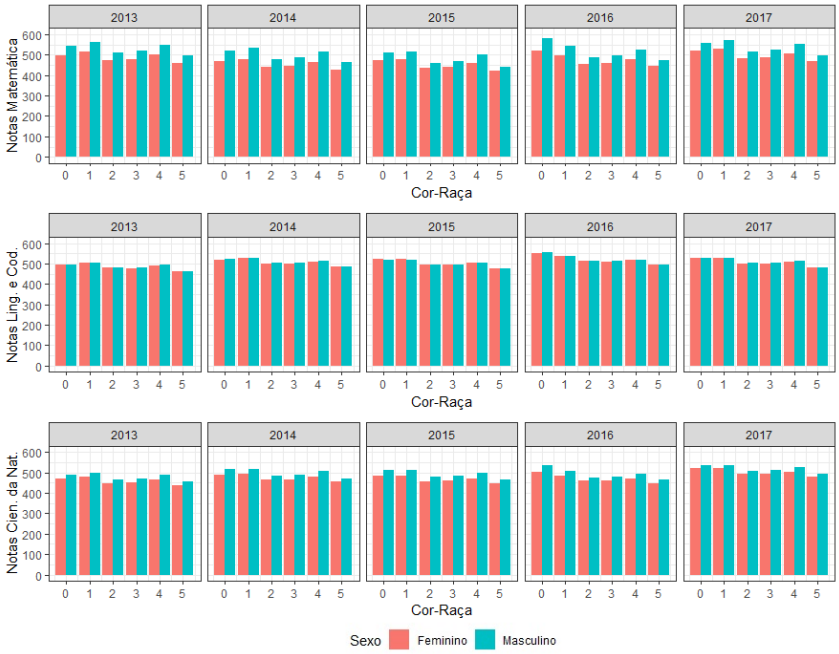
O desempenho por tipo de escola do ensino médio é ilustrado na Figura 16. Em 2013 e 2014, os tipos de escola foram categorizados em (1) pública e (2) privada, enquanto que em 2015, 2016 e 2017 eles foram categorizados como (1) não respondeu, (2) pública, (3) privada, e (4) exterior. No processo de limpeza dos dados, usou-se a categorização mais detalhada visando maior diversidade de análise. Nesse processo, participantes de 2013 e 2014 com valores nulos

Figura 14 – Desempenho por sexo por região nos exames do ENEM nos anos de 2013 a 2017, para cada área de conhecimento nas ciências exatas. Regiões identificadas como: (1) Norte, (2) Nordeste, (3) Sudeste, (4) Sul, (5) Centro-Oeste.



Fonte: Noguera, Branco e Ciferri (2019).

Figura 15 – Desempenho por sexo por cor/raça, nos exames do ENEM nos anos de 2013 a 2017, para cada área de conhecimento nas ciências exatas. Categorias identificadas como: (0) não declarado, (1) branca, (2) preta, (3) parda, (4) amarela, (5) indígena.



Fonte: Noguera, Branco e Ciferri (2019).

em tipo de escola foram reclassificados como da categoria (1) não respondeu. Para esses anos, a categoria (4) exterior não incluiu participantes por não ser possível identificá-los. Nos anos

de 2013 e 2014, os resultados mostraram que os participantes oriundos de escolas privadas obtiveram melhor desempenho do que os de escolas públicas. Para os anos de 2015 a 2017, os resultados foram melhores para participantes de escolas no exterior, depois de escolas privadas e depois de escolas públicas.

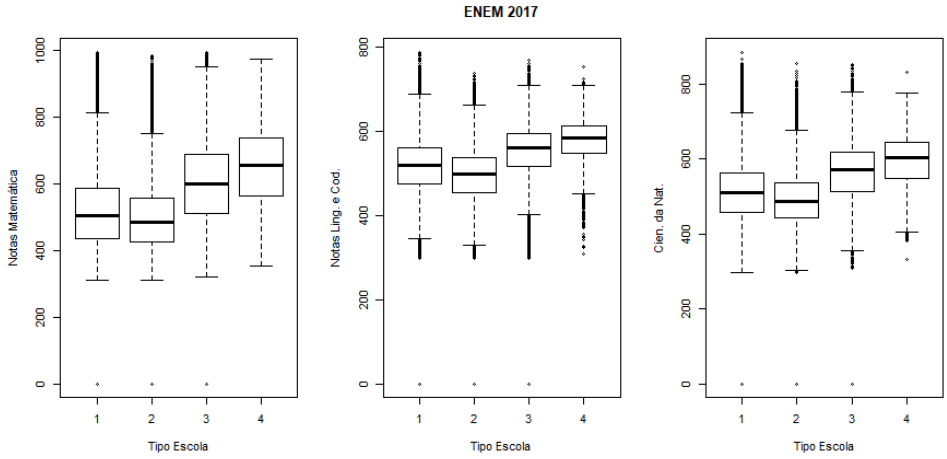
Figura 16 – Desempenho por sexo por tipo de escola do ensino médio nos exames do ENEM nos anos de 2013 a 2017, para cada área de conhecimento nas ciências exatas. Categorias identificadas como: (1) não respondeu, (2) pública, (3) privada, e (4) exterior.

Fonte: Noguera, Branco e Ciferri (2019).

Para investigar a variabilidade dos dados e valores atípicos que podem influenciar a análise dos resultados relativos à categoria (4) exterior, na Figura 17 é ilustrado o desempenho comparativo por tipo de escola do ensino médio no ano de 2017. O retângulo contém 50% da amostra. A linha interna dentro dele indica a mediana, sendo que pelo menos 50% da amostra está acima desse valor e pelo menos 50% está abaixo. Nota-se que: (i) as medianas dos participantes formados no (4) exterior são maiores que as medianas dos participantes formados nos tipos de escola (3) privada e (2) pública; (ii) as notas mínimas dos formados no exterior (traço de linha horizontal mais inferior no gráfico) estão acima das notas mínimas dos formados nas escolas privadas e públicas; (iii) as notas máximas dos formados no exterior (traço de linha horizontal mais superior no gráfico) estão acima das notas máximas dos formados nas escolas privadas e públicas; e (iv) as notas mínimas mostram uma variabilidade menor de valores (pontilhados abaixo do traço horizontal mais inferior), enquanto que as notas máximas mostram uma variabilidade maior de valores (pontilhados acima do traço horizontal mais superior). A justificativa é que, comparada à quantidade de participantes das outras categorias, a categoria (4) possui o menor número de participantes: 625 em 2015, 985 em 2016 e 4.596 em 2017. Assim, à medida que a quantidade de participantes por tipo de escola aumenta, a média das notas diminui.

Na Figura 18 é ilustrado o desempenho por renda salarial mensal, considerando os valores dos salários mensais de 2017. Conforme esperado, os resultados mostraram que as diferenças

Figura 17 – BoxPlot do desempenho comparativo por tipo de escola do ensino médio no exame do ENEM no ano de 2017, para cada área de conhecimento nas ciências exatas. Categorias identificadas como: (1) não respondeu, (2) pública, (3) privada, e (4) exterior.



Fonte: Noguera, Branco e Ciferri (2019).

socioeconômicas influenciam fortemente no desempenho dos participantes, isto é, conforme aumenta a renda mensal, o desempenho dos participantes melhora.

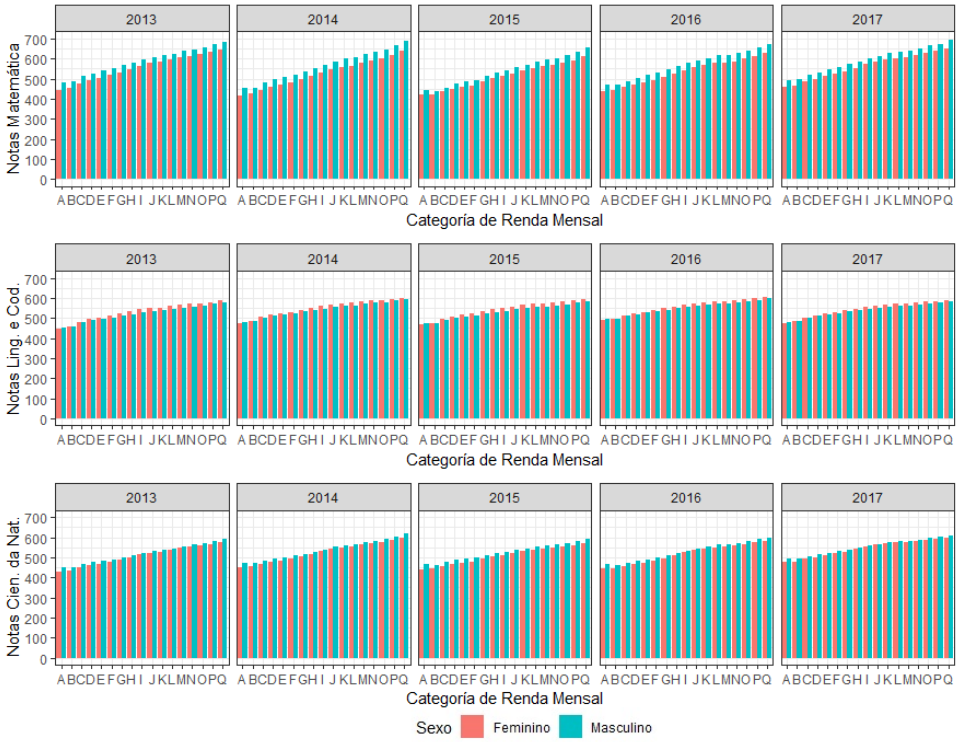
6.2.2 Desempenho dos participantes considerando as médias informadas pelo Inep

Nos resultados descritos nesta seção, é considerado apenas o ano de 2017, o ano mais atual frente ao período que foi analisado no artigo de Noguera, Branco e Ciferri (2020). Também é considerado apenas o fator de análise região, porém são mostrados dados mais detalhados relacionados às unidades federativas dessas regiões. O desempenho dos participantes foi definido como a média aritmética das notas para cada área de conhecimento das ciências exatas por sexo, mas considerando-se apenas as médias iguais ou superiores à média informada pelo Inep em cada área de conhecimento.

Na Figura 19 é ilustrado o desempenho dos participantes cuja nota média foi igual ou superior à média informada pelo Inep em cada área de conhecimento das ciências exatas. Pode-se constatar que em *Matemática* e *Ciência da Natureza* os participantes de sexo masculino tiveram desempenho ligeiramente superior ao das participantes do sexo feminino (de 3,8% e 0,6% respectivamente). Já em *Linguagens e Códigos*, as mulheres proveram desempenho um pouco maior do que os homens (de 0,9%).

Os resultados ilustrados nas Figuras 20 a 24 apresentam os resultados da Figura 19 considerando um nível maior de detalhe, ou seja, mostram o desempenho dos participantes cuja nota média foi igual ou superior à média informada pelo Inep em cada área de conhecimento das ciências exatas, para todas as unidades federativas das regiões Norte (Figura 20), Nordeste (Figura 21), Sudeste (Figura 22), Sul (Figura 23) e Centro-Oeste (Figura 24).

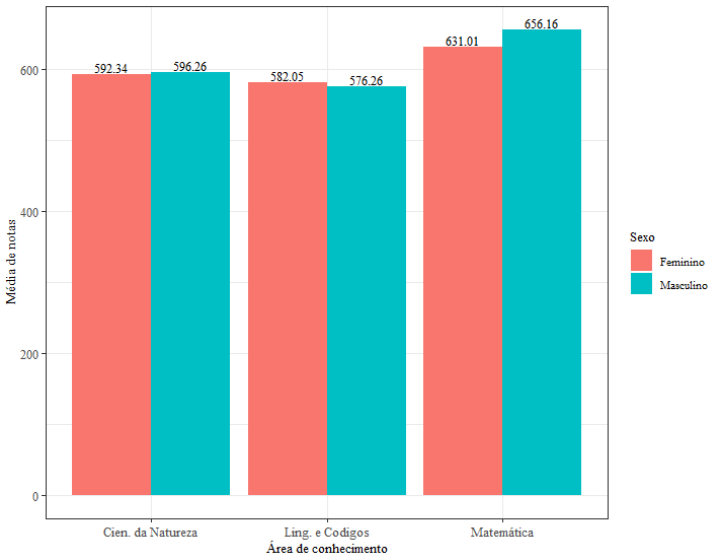
Figura 18 – Desempenho por renda salarial mensal, para as categorias identificadas como: (A) nenhuma renda. (B) até R$ 937,00. (C) de R$ 937,01 até R$ 1.405,50. (D) de R$ 1.405,51 até R$ 1.874,00. (E) de R$ 1.874,01 até R$ 2.342,50. (F) de R$ 2.342,51 até R$ 2.811,00. (G) de R$ 2.811,01 até R$ 3.748,00. (H) de R$ 3.748,01 até R$ 4.685,00. (I) de R$ 4.685,01 até R$ 5.622,00. (J) de R$ 5.622,01 até R$ 6.559,00. (K) de R$ 6.559,01 até R$ 7.496,00. (L) de R$ 7.496,01 até R$ 8.433,00. (M) de R$ 8.433,01 até R$ 9.370,00. (N) de R$ 9.370,01 até R$ 11.244,00. (O) de R$ 11.244,01 até R$ 14.055,00. (P) de R$ 14.055,01 até R$ 18.740,00. (Q) mais de R$ 18.740,00.



Fonte: Noguera, Branco e Ciferri (2019).

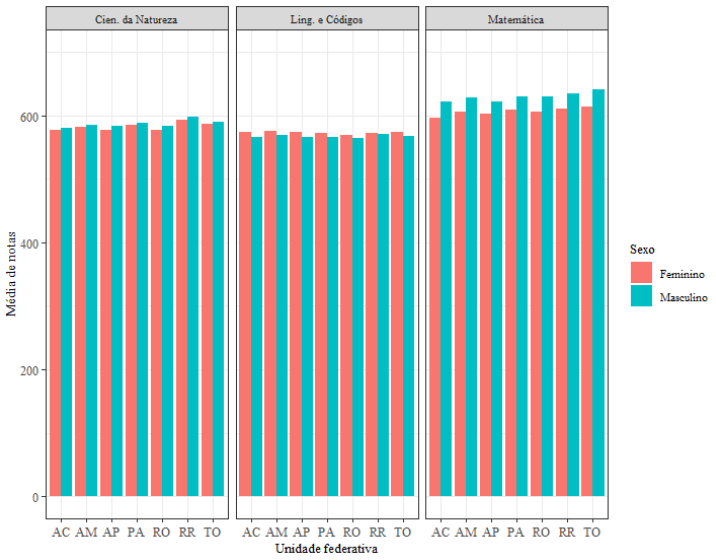
Os resultados mostram, para cada região, a mesma tendência que a observada na Figura 19. Com relação à *Matemática*, o desempenho dos participantes de sexo masculino foi superior ao desempenho dos participantes de sexo feminino, em todas as regiões, em todas as unidades federativas. Considerando este critério, para a região Norte, a diferença favorável mais acentuada e com relevância foi na unidade federativa Tocantins (Figura 20), para a região Nordeste, na unidade federativa Ceará (Figura 21), para a região Sudeste, na unidade federativa Rio de Janeiro (Figura 22), para a região Sul, na unidade federativa Santa Catarina (Figura 23) e para a região Centro-Oeste, na unidade federativa Distrito Federal (Figura 24).

Figura 19 – Desempenho por sexo por área de conhecimento nas ciências exatas no exame do ENEM de 2017, para participantes cuja nota média foi igual ou superior à média informada pelo Inep nas diferentes áreas.



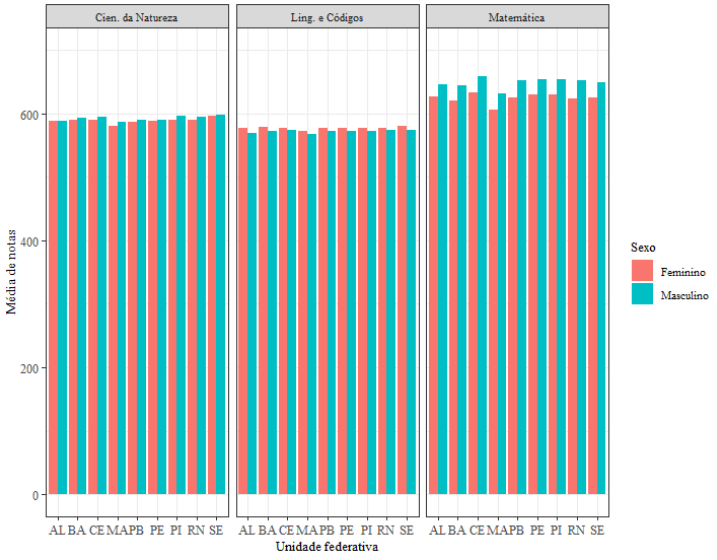
Fonte: Noguera, Branco e Ciferri (2020).

Figura 20 – Desempenho por sexo por unidade federativa para a região Norte no exame do ENEM de 2017, para cada área de conhecimento nas ciências exatas. UF identificadas como: (AC) Acre, (AM) Amazonas, (AP) Amapá, (PA) Pará, (RO) Rondônia, (RR) Roraima e (TO) Tocantins.



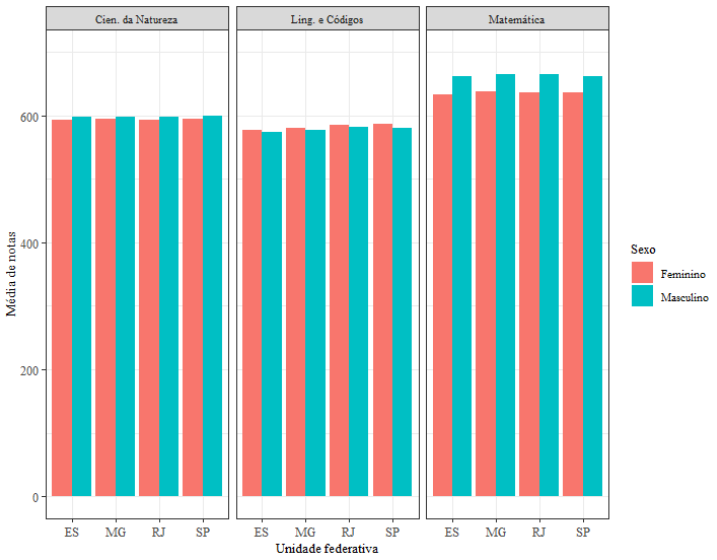
Fonte: Noguera, Branco e Ciferri (2020).

Figura 21 – Desempenho por sexo por unidade federativa para a região Nordeste no exame do ENEM de 2017, para cada área de conhecimento nas ciências exatas. UF identificadas como: (AL) Alagoas, (BA) Bahia, (CE) Ceará, (MA) Maranhão, (PB) Paraíba, (PE) Pernambuco, (PI) Piauí, (RN) Rio Grande do Norte e (SE) Sergipe.



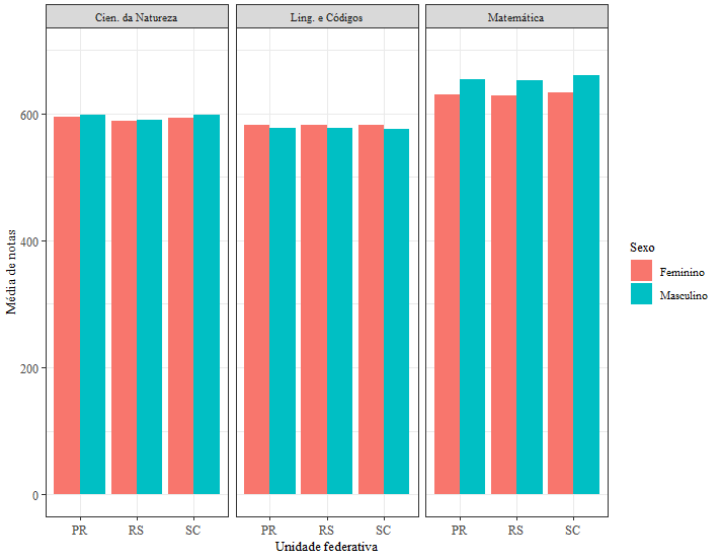
Fonte: Noguera, Branco e Ciferri (2020).

Figura 22 – Desempenho por sexo por unidade federativa para a região Sudeste no exame do ENEM de 2017, para cada área de conhecimento nas ciências exatas. UF identificadas como: (ES) Espírito Santo, (MG) Minas Gerais, (RJ) Rio de Janeiro e (SP) São Paulo.



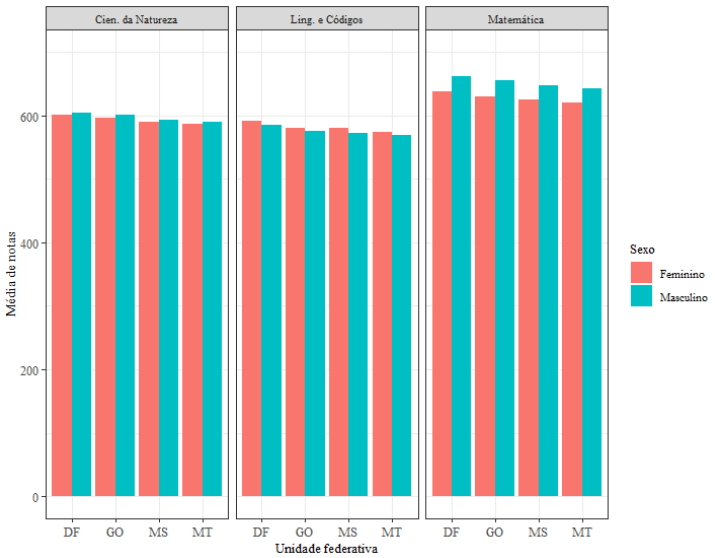
Fonte: Noguera, Branco e Ciferri (2020).

Figura 23 – Desempenho por sexo por unidade federativa para a região Sul no exame do ENEM de 2017, para cada área de conhecimento nas ciências exatas. UF identificadas como: (PR) Paraná, (RS) Rio grande do Sul e (SC) Santa Catarina.



Fonte: Noguera, Branco e Ciferri (2020).

Figura 24 – Desempenho por sexo por unidade federativa para a região Centro-Oeste no exame do ENEM de 2017, para cada área de conhecimento nas ciências exatas. UF identificadas como: (DF) Distrito Federal, (GO) Goiás, (MS) Mato Grosso do Sul e (MT) Mato Grosso.



Fonte: Noguera, Branco e Ciferri (2020).

6.3 Considerações finais

Neste capítulo foi descrito o segundo cenário de uso da arquitetura proposta no Capítulo 4. Foi analisado o desempenho dos participantes do ENEM na temática "gêneros e suas nuances na tecnologia da informação". Foram considerados diferentes fatores de análise e diferentes anos de realização do exame do ENEM, de acordo com as seguintes investigações sobre o desempenho dos participantes de sexo feminino e masculino em cada área de conhecimento das ciências exatas: (i) análise do desempenho de todos os participantes no período de 2013 a 2017 para os fatores de análise região, cor/raça, tipo de escola de ensino médio e renda salarial mensal; e (ii) análise do desempenho dos participantes de sexo feminino e masculino cuja nota média foi igual ou superior à média informada pelo Inep em cada área de conhecimento das ciências exatas, considerando o ano de 2017 e, para cada região do Brasil, suas unidades federativas.

As análises descritas neste capítulo podem ser consideradas como um trabalho inicial voltado à tomada de decisão educacional na proposta de ações específicas no âmbito da educação, com enfoque no equilíbrio entre os gêneros nas ciências exatas. Por meio dessas análises, conclui-se que o desempenho dos participantes de sexo masculino foi discretamente superior ao dos participantes do sexo feminino em *Matemática* e *Ciência da Natureza*, mas para *Linguagens e Códigos* o desempenho de ambos os sexos foi quase equivalente ou as mulheres demonstraram um desempenho apenas um pouco superior ao desempenho dos homens. É possível identificar, portanto, que a diferença de desempenho entre participantes dos sexos femininos e masculinos não é tão acentuada quando comparada com o desequilíbrio desses sexos nos cursos na área de ciências exatas e nos profissionais que atuam nessas áreas. Isso gera o seguinte questionamento que deve ser investigado na tomada de decisão educacional: "Por que as mulheres optam por não seguir carreira nas áreas de ciências exatas, mesmo apresentando desempenho mais ou menos próximo ao desempenho dos homens no exame do ENEM?".

No próximo capítulo, Capítulo 7, é descrito o terceiro cenário de uso da arquitetura proposta, o qual tem como objetivo investigar novamente a temática gênero, porém considerando apenas questões STEM e não-STEM.

CAPÍTULO

7

DESEMPENHO POR GÊNERO NAS QUESTÕES STEM E NÃO-STEM

Neste capítulo é detalhado o terceiro cenário de uso da arquitetura. O objetivo consiste em investigar o desempenho de participantes com base no gênero nas diferentes questões STEM e não-STEM do ENEM dos anos de 2016 a 2020. Questões STEM se referem àquelas pertencentes às áreas de conhecimento de *Matemática* e *Ciência da Natureza* e às questões identificadas pelos códigos de habilidade 28, 29 e 30 assinaladas na matriz de referência de *Linguagens e Códigos* e classificadas como relativas às Tecnologias da Informação e Comunicação (TIC). Em contrapartida, as questões não-STEM correspondem àquelas pertencentes à área de conhecimento de *Ciências Humanas* e às questões da *Linguagens e Códigos* com exceção das classificadas como TIC. Nas análises realizadas, são utilizados testes de hipóteses, tendo sido aplicado o teste *Kolmogorov-Smirnov* para determinar se o conjunto de dados é modelado por uma distribuição normal ou não. Também foi aplicado teste não paramétrico de *U de Mann-Whitney* para demostração rigorosa da existência ou inexistência de diferença estatística entre a proporção de acertos por sexo nas questões STEM e não-STEM. Adicionalmente, é obtida a estimativa de parâmetros dos itens adotados pela metodologia TRI para identificar questões com maiores e menores níveis de dificuldade.

Na seção 7.1 é descrito como a arquitetura proposta é instanciada para oferecer suporte para as análises realizadas. Na seção 7.2 são discutidos os resultados obtidos. Na seção 7.3 é realizada uma análise dos resultados. O capítulo é finalizado na seção 7.4 com as considerações finais.

7.1 Instanciação da arquitetura proposta

A instância da arquitetura proposta empregada para oferecer suporte para as análises descritas neste capítulo faz referência ao *pipeline* para utilização de *Data lake* com ferramentas

de análise estatística ilustrada na Figura 10 (seção 4.2).

Na camada de conexão de dados, as fontes de dados utilizadas correspondem aos dados do ENEM para os anos 2016 até 2020. Os arquivos ITENS_PROVA.csv e MICRODADOS_-ENEM.csv foram armazenados no HDFS em formato nativo na área de dados brutos na camada de armazenamento de dados do *Data lake*. Os conjuntos de dados brutos referentes aos candidatos inscritos nos anos de 2016 a 2020 possuíam 8.627.367, 6.731.341, 5.513.747, 5.095.270 e 5.783.109 instâncias, respectivamente. Foram considerados apenas os participantes não treineiros que participaram: (i) da primeira aplicação da prova; e (ii) de todas as provas das áreas do conhecimento e na *Redação*. A Tabela 12 detalha o número de amostras geradas na fase de ELT para os anos considerados, bem como a porcentagem de participantes do sexo feminino e masculino. A cada ano, a participação das mulheres é maior que a dos homens.

Tabela 12 – Número de participantes analisados a cada ano.

Ano	Quantidade	Feminino	Masculino
2016	4.841.069	57,2%	42,8%
2017	3.986.821	58,1%	41,9%
2018	3.431.317	59,0%	41,0%
2019	3.171.799	59,0%	41,0%
2020	2.157.227	60,0%	40,0%

Na área de processamento na camada de armazenamento de dados do *Data lake* foram realizados diversos processos. Primeiramente, foi obtida a resposta do participante para cada questão para cada área de conhecimento. Também foi obtida a avaliação da resposta, ou seja, foi identificado se o participante acertou, errou, deixou a questão em branco ou fez dupla marcação.

Nos arquivos obtidos do Inep, as respostas e seus respectivos gabaritos não são armazenadas individualmente. Ao contrário, para cada amostra, existe uma variável do tipo *string* que contém a resposta dada pelo participante a cada questão e uma variável do tipo *string* que contém o gabarito de cada questão dentro de sua área de conhecimento. O número total de questões do ENEM, 180, é obtido somando-se o número de questões presentes na *string* de resposta de cada área de conhecimento. Para cada ano da prova e para cada área de conhecimento, foi gerado um novo conjunto de dados, no qual foram armazenadas a resposta do participante para cada questão, assim como a avaliação de sua resposta. A saída dessas transformações foram armazenadas na área de processamento da camada de armazenamento de dados do *Data lake*.

Em outro processo, com base na avaliação do participante para cada questão da prova, foi possível quantificar o total de número de acertos para cada área do conhecimento e para cada questão. Independentemente das cores dos cadernos do ENEM, uma mesma questão em cadernos diferentes, mesmo que identificadas com números diferentes, foi considerada a mesma. Foi calculada a frequência relativa de acertos por sexo, ou seja, o número total de acertos para determinada questão por sexo dividido pelo número total de participantes deste sexo. O resultado da frequência relativa é apresentado na forma de porcentagem.

As questões foram classificadas como STEM e não-STEM, utilizando como base o seguinte critério. Questões STEM se referem às questões: (i) pertencentes às áreas de conhecimento de *Matemática* e *Ciência da Natureza*; e (ii) identificadas pelos códigos de habilidade 28, 29 e 30 assinaladas na matriz de referência de *Linguagens e Códigos*. Questões não-STEM se referem às questões: (i) pertencentes à área de conhecimento de *Ciências Humanas*; e (ii) pertencentes à área de conhecimento de *Linguagens e Códigos*, com exceção daquelas identificadas pelos códigos de habilidade 28, 29 e 30. O conjunto de dados resultante dos últimos processos foi armazenado em um arquivo chamado "analise1" na área de acesso aos dados da camada de armazenamento de dados do *Data lake*.

Para a estimativa dos parâmetros dos itens, ainda na área de processamento da camada de armazenamento de dados do *Data lake*, foram feitos os seguintes processos. Para cada ano de 2016 a 2019, foram utilizadas mais de 3.000 amostras para estimar os parâmetros dos itens da prova do ENEM. De acordo com Santo (2020), uma amostra com 3.000 ou mais participantes apresenta uma convergência de valores dos parâmetros de discriminação e um grau de dificuldade dos itens com uma distância relativa menor do que 10% dos parâmetros estimados para a população.

Cada tipo de caderno é identificado por uma cor diferente, sendo que cada cor contém as mesmas questões, porém dispostas em ordem diferente. Com o objetivo de colocar todas as questões na mesma ordem, o caderno de cor rosa foi adotado como padrão. Colocar todas as questões na mesma ordem é necessário para gerar a matriz de resposta. Essa matriz é posteriormente convertida em uma matriz de resposta binária com base no gabarito, onde 0 significa uma resposta incorreta e 1 uma resposta correta.

Não foram considerados os participantes com nota zero em todas as áreas de conhecimento. Os participantes restantes foram ordenados em ordem decrescente pela soma de acertos de cada um em cada área de conhecimento. Com base nessa ordem, os participantes foram estratificados em 10 grupos, cada um contendo até no máximo 5% da população total. A saída do processo foi armazenada em um arquivo chamado "analise2" na área de acesso aos dados da camada de armazenamento de dados do *Data lake*.

Na camada de análise de dados, para as análises realizadas sobre o arquivo "analise1", foram utilizados testes de hipóteses com o objetivo de demonstrar com rigor a existência ou inexistência de diferenças estatísticas entre proporções de acertos. A decisão sobre usar testes paramétricos ou não paramétricos foi baseada no fato da distribuição das proporções de acertos seguir uma distribuição normal ou não. Para cada ano de análise e para cada grupo de questões STEM e não-STEM, foi utilizado o teste de *Kolmogorov-Smirnov* para testar a hipótese de normalidade. Como as amostras não seguiam a normalidade, foi aplicado o teste não paramétrico de *U de Mann-Whitney*.

Para a estimativa dos parâmetros dos itens, realizadas sobre o arquivo "analise2", foi empregado o pacote *sirt* da linguagem *R* (ROBITZSCH, 2015). O principal critério utilizado

para a escolha do pacote foi conter os métodos de estimativa adotados pelo Inep (*Expected a Posteriori* e Máxima Verossimilhança Marginal).

O processo ELT foi realizado com *Apache Spark* e a governança de dados foi feita por meio do *Apache Atlas*. Para a análise estatística, foram utilizadas a ferramenta *R* e as bibliotecas *statsmodel* e *pingouin*. A automação das etapas do processo ELT até a camada de análise foi realizada com o gerenciador *Apache Airflow*. A linguagem *Python* foi utilizada nas camadas de ELT, análise de dados e gerenciador de fluxo de trabalho.

7.2 Resultados e discussões

Nesta seção são descritos os resultados obtidos nas análises realizadas. Na seção 7.2.1 é resumida a estatística descritiva sobre as proporções de acertos. Na seção 7.2.2 são descritos os resultados obtidos nas aplicações dos testes de hipóteses. Na seção 7.2.3 são mostradas as estimativas dos parâmetros do item.

7.2.1 Estatística descritiva das proporções de acertos

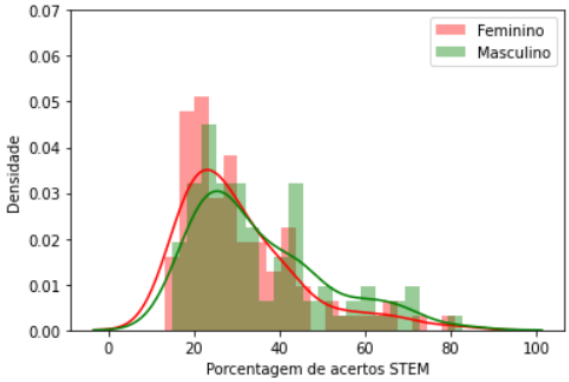
Nas Figuras 25 e 26 são ilustradas as distribuições das proporções de acertos das questões STEM e não-STEM para os participantes de sexo feminino (vermelho) e masculino (verde), para o ENEM de 2020. No eixo x é exibido o valor das proporções de acertos variando de 0 a 100% com intervalos de 20%. No eixo y é exibida a estimativa de densidade do *kernel*, que representa a probabilidade por unidade no eixo x. Por padrão, os diferentes histogramas são sobrepostos uns sobre os outro, a cor marrom indica esta sobreposição. Em ambas figuras, é possível observar que, para os valores de 20% a 40%, as participantes do sexo feminino têm maior estimativa de densidade de probabilidade do que os participantes de sexo masculino, ou seja, existe maior número de observações com essas proporções para as mulheres. Para as proporções acima de 40%, os participantes masculinos têm maior estimativa de densidade de probabilidade.

Na Tabela 13 são descritos diversos valores (média, desvio-padrão, mínimo, mediana e máximo) relativos às proporções de acertos por sexo das questões STEM e não-STEM considerando os anos de análise. As seguintes observações podem ser feitas:

- A diferença na média é de, no máximo, 4%, considerando ambos os sexos e ambas as classificações das questões. Os participantes do sexo masculino são os que têm, em média, maior proporção de acertos.

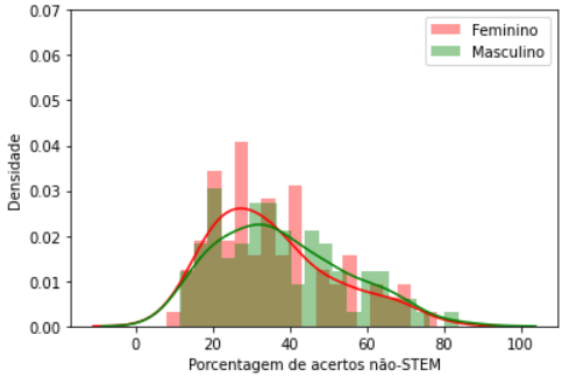
- A diferença no desvio padrão é de, no máximo, 2%, considerando ambos os sexos e ambas as classificações das questões. O desvio padrão pequeno indica que as proporções de acertos em ambos os sexos estão próximos da média, ou seja, têm pouca variabilidade.

Figura 25 – Distribuição das proporções de acertos para questões STEM para os participantes de sexo feminino (vermelho) e masculino (verde).



Fonte: Elaborada pelo autor.

Figura 26 – Distribuição das proporções de acertos para questões não-STEM para os participantes de sexo feminino (vermelho) e masculino (verde).



Fonte: Elaborada pelo autor.

- A proporção mínima de acertos nas questões das áreas STEM é de 6% e 7% para participantes de sexo feminino e masculino, respectivamente. Já a proporção mínima de acertos nas questões das áreas não-STEM é de 8% para ambos os sexos.

- A mediana nas questões das áreas STEM e não-STEM são maiores para os participantes do sexo masculino do que para os participantes do sexo feminino.

- O valor máximo para as proporções de acertos das áreas STEM é de 81% e 83% para participantes do sexo feminino e masculino, respectivamente.

- A proporção máxima de acertos para as questões das áreas não-STEM é de 86% e 85% para participantes do sexo feminino e masculino, respectivamente.

Tabela 13 – Valores médios, desvios-padrão, mínimo, mediana e máximo para os anos de 2016 a 2020 para as proporções de acertos das questões das áreas de STEM e não-STEM por sexo.

Ano	Área	Sexo	Média (%)	Desvio padrão (%)	Mínimo (%)	Mediana (%)	Máximo (%)
2016	STEM	F	25	8	10	29	59
		M	28	10	10	32	70
	Não-STEM	F	33	14	9	39	74
		M	35	14	12	43	78
2017	STEM	F	26	12	6	31	59
		M	28	13	7	34	65
	Não-STEM	F	34	11	11	40	65
		M	36	11	10	42	66
2018	STEM	F	26	10	10	30	69
		M	28	11	12	33	80
	Não-STEM	F	42	18	12	53	86
		M	44	18	12	55	85
2019	STEM	F	28	14	9	33	82
		M	31	15	11	37	81
	Não-STEM	F	35	15	8	47	77
		M	37	15	8	50	66
2020	STEM	F	31	14	13	37	81
		M	35	15	15	44	83
	Não-STEM	F	35	16	8	43	78
		M	38	17	12	49	84

Sexo F: feminino; M: masculino.

Fonte: Elaborada pelo autor.

7.2.2 Testes de hipóteses

Nesta seção são detalhados os resultados obtidos nos testes de hipóteses. Na seção 7.2.2.1 é utilizado o teste de normalidade para determinar se o conjunto de dados analisado é modelado por uma distribuição normal ou não. Na seção 7.2.2.2 é descrito o resultado obtido na aplicação do teste de *U de Mann-Whitney* usado para verificar a igualdade das medianas.

7.2.2.1 Teste de normalidade

Nos testes de hipóteses existem duas suposições contraditórias em consideração. O objetivo é decidir, com base nas informações da amostra, qual das duas hipóteses está correta. A hipótese nula (H_0) é a alegação inicial assumida como verdadeira. A hipótese alternativa (H_1) é a afirmação contraditória à hipótese nula.

O teste de normalidade dos dados foi realizado usando o teste *Kolmogorov-Smirnov*. Foi adotado um nível de significância (α) de 0,05, o qual indica um risco de 5% de se concluir que os dados não seguem a distribuição normal quando eles a seguem. As hipóteses avaliadas para cada ano e para as proporções de acertos das questões classificadas como STEM foram as seguintes:

- H_0: “As amostras correspondentes às proporções de acertos das questões classificadas como STEM seguem a distribuição normal”.
- H_1: “As amostras correspondentes às proporções de acertos das questões classificadas como STEM não seguem a distribuição normal”.

As hipóteses avaliadas para as proporções de acertos das questões classificadas como não-STEM foram as seguintes:

- H_0: “As amostras correspondentes às proporções de acertos das questões classificadas como não-STEM seguem a distribuição normal”.
- H_1: “As amostras correspondentes às proporções de acertos das questões classificadas como não-STEM não seguem a distribuição normal”.

Para todos os anos analisados, independentemente das questões serem classificadas como STEM e não-STEM, os resultados sugerem que nenhuma das amostras segue uma distribuição normal. Ou seja, a significância obtida é menor que 0,05. Como resultado, rejeita-se H_0 e aceita-se H_1.

7.2.2.2 Teste de U de Mann-Whitney

Como nenhuma das amostras apresentou distribuição normal, foi aplicado o teste não paramétrico de *U de Mann-Whitney*. Este teste pode ser considerado a versão não paramétrica do teste *t* para amostras independentes que considera a igualdade das medianas. O objetivo é determinar se há diferença estatística entre as proporções de acertos por sexo por cada grupo de questões. Para cada ano e para o grupo de questões classificadas como STEM, foram avaliadas as seguintes hipóteses:

- H_0: “A proporção de acertos nas questões classificadas como STEM entre os participantes de sexo feminino e masculino não são estatisticamente diferentes”.
- H_1: “A proporção de acertos nas questões classificadas como STEM entre os participantes de sexo feminino e masculino são estatisticamente diferentes”.

Para cada ano e para o grupo de questões classificadas como não-STEM, foram avaliadas as seguintes hipóteses:

- H_0: “A proporção de acertos nas questões classificadas como não-STEM entre os participantes de sexo feminino e masculino não são estatisticamente diferentes”.
- H_1: “A proporção de acertos nas questões classificadas como não-STEM entre os participantes de sexo feminino e masculino são estatisticamente diferentes”.

Na Tabela 14 são detalhados os resultados obtidos pela aplicação do teste de *U de Mann-Whitney*, considerando os anos de 2016 a 2020, a classificação das questões em STEM e não-STEM e o sexo. O valor-p é o nível de significância α mais baixo. As seguintes observações podem ser feitas:

- O valor-p é menor do que 0,05 para os anos de 2016, 2018 e 2020 e as questões classificadas como STEM. Para esses casos, rejeita-se H_0 e aceita-se H_1, ou seja, aceita-se a hipótese de que "A proporção de acertos nas questões classificadas como STEM entre os participantes de sexo feminino e masculino são estatisticamente diferentes".
- Para os demais anos, independentemente da classificação das questões em STEM e não-STEM, o valor-p é maior do que 0,05. Portanto, aceita-se a hipótese H_0, ou seja, aceita-se que "A proporção de acertos entre os participantes de sexo feminino e masculino não são estatisticamente diferentes".

As medidas de tamanho do efeito mostram a significância prática dos resultados obtidos em estudos que realizam a comparação de dois grupos. Foi utilizado o método *eta squared* (η^2) (FRITZ; MORRIS; RICHLER, 2012) para calcular o tamanho do efeito dos testes cujas hipóteses H_1 foram aceitas. Esse método descreve a proporção da variabilidade total nos dados que são contabilizados pelo efeito em consideração. A fórmula de *eta squared* é descrita na Equação 7.1.

$$\eta^2 = \frac{z^2}{N} \tag{7.1}$$

onde z^2 é uma variável padronizada e N é a soma do tamanho dos dois grupos.

Na Tabela 14 são mostrados os valores do tamanho do efeito e a interpretação do mesmo com base nas regras de Cohen (2013). Os valores convencionais de tamanho do efeito 0,01, 0,09, 0,25 correspondem ao efeito pequeno, médio e grande, respectivamente. Para os anos 2016, 2018 e 2020, o tamanho do efeito foi médio.

7.2.3 Estimativa dos parâmetros do item

Desde 2009, o ENEM utiliza o modelo de Teoria de Resposta ao Item (TRI), técnica utilizada para a correção das questões das provas (seção 2.7). Entretanto, os parâmetros do modelo TRI eram acessíveis apenas por meio do Serviço de Acesso a Dados Protegidos (Sedap) do Inep, e somente após cinco anos da aplicação da prova. Na edição de 2020, esses parâmetros foram incluídos pela primeira vez no arquivo ITENS_PROVA_2020.csv. Nesta tese, *item* e *questão* são tratados como sinônimos.

O índice de dificuldade, um dos elementos das estimativas dos parâmetros dos itens disponibilizadas nos microdados do ENEM 2020 é detalhado nas Tabelas 15 e 16. À esquerda,

Tabela 14 – Resultados do teste de *U de Mann-Whitney* para os anos de 2016 a 2020 para as proporções de acertos das questões classificadas como STEM e não-STEM por sexo.

Ano	Grupo	Sexo	Valor-p	Decisão	Valor de tamanho do efeito	Interpretação de tamanho do efeito
2016	STEM	F M	**0,021**	Rejeita-se H_0 e aceita-se H_1.	0,03	Efeito médio
	Não-STEM	F M	0,234	Aceita-se H_0.		
2017	STEM	F M	0,123	Aceita-se H_0.		
	Não-STEM	F M	0,147	Aceita-se H_0.		
2018	STEM	F M	**0,044**	Rejeita-se H_0 e aceita-se H_1.	0,02	Efeito médio
	Não-STEM	F M	0,542	Aceita-se H_0.		
2019	STEM	F M	0,053	Aceita-se H_0.		
	Não-STEM	F M	0,216	Aceita-se H_0.		
2020	STEM	F M	**0,016**	Rejeita-se H_0 e aceita-se H_1.	0,03	Efeito médio
	Não-STEM	F M	0,274	Aceita-se H_0.		

Sexo F: feminino; M: masculino.

Fonte: Elaborada pelo autor.

são mostradas as 10 primeiras questões com maior dificuldade e à direita as 10 primeiras questões com menor dificuldade. Nas tabelas, o campo posição é a localização da questão dentro da prova. A área corresponde às áreas de conhecimento. O código da questão é um código identificador único. O TRI expressa a dificuldade do item em uma escala padrão, a escala de habilidade varia entre -3 e 3 (PASQUALI, 2017). Observa-se que algumas questões com maior dificuldade extrapolam da escala padrão, levando à conclusão de que essas questões são muito difíceis.

Tabela 15 – Estimativa dos parâmetros dos itens para questões classificadas como STEM do ENEM do ano de 2020.

STEM (Dificuldade alta)				STEM (Dificuldade baixa)			
Posição	Área	Cod. Questão	Índice de dificuldade	Posição	Área	Cod. Questão	Índice de dificuldade
95	CN	53.548	3,80025	135	CN	79.124	-0,07547
158	MT	32.975	3,34168	180	MT	82.724	-0,08552
144	MT	48.864	3,29884	31	LC	64.056	-0,16732
167	MT	63.901	3,0197	91	CN	111.501	-0,23526
164	MT	15.248	2,86173	136	MT	111.491	-0,40358
140	MT	83.753	2,82961	6	LC	33.271	-0.5151
176	MT	84.253	2,82281	112	CN	55.948	0,12756
165	MT	86.282	2,77688	114	CN	87.690	0,32436
113	CN	82.243	2,72331	157	MT	88.028	0,44069
161	MT	29.531	2,6803	102	CN	83.584	0,47507

Área MT: *Matemática*, CN: *Ciência da Natureza*, LC: *Linguagens e Códigos*.

Fonte: Elaborada pelo autor.

Tabela 16 – Estimativa dos parâmetros dos itens para as questões classificadas como não-STEM do ENEM do ano de 2020.

Não-STEM (Dificuldade alta)				Não-STEM (Dificuldade baixa)			
Posição	Área	Cod. Questão	Índice de dificuldade	Posição	Área	Cod. Questão	Índice de dificuldade
14	LC	30.842	5,45136	90	CH	88.307	-0,03828
68	CH	111.912	3,51778	25	LC	76.967	-0,11868
58	CH	97.979	3,22867	46	CH	84.333	-0,24666
34	LC	89.327	2,74643	42	LC	89.906	-0,40822
78	CH	85.290	2,28546	15	LC	111.838	-0,44002
23	LC	16.501	2,2551	6	LC	33.271	-0,5151
8	LC	20.652	2,21741	24	LC	44.328	0,02359
49	CH	111.968	2,07959	21	LC	84.295	0,07454
39	LC	112.138	1,91811	67	CH	87.040	0,08927
18	LC	111.989	1,91416	69	CH	112.015	0,10876

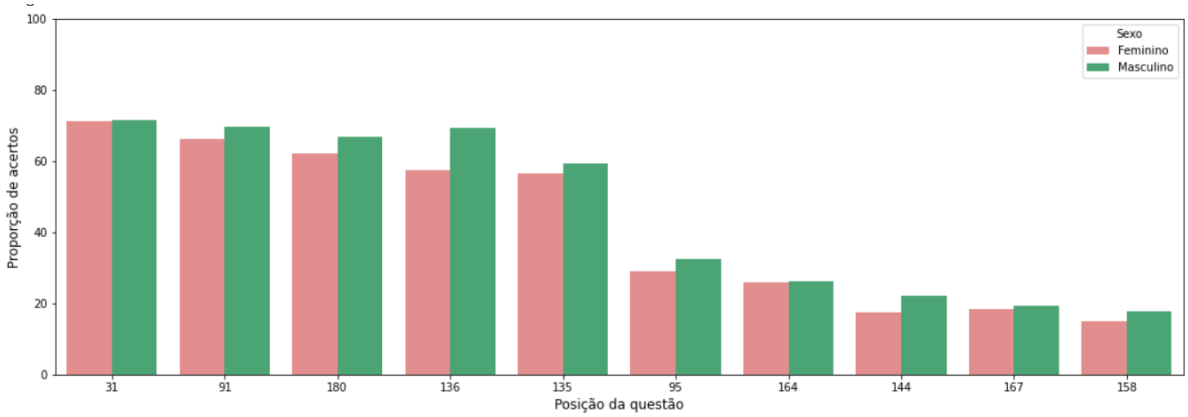
Área LC: *Linguagens e Códigos*; CH: *Ciências Humanas*.

Fonte: Elaborada pelo autor.

Nas Figuras 27 e 28 são exibidas as proporções de acertos das 5 primeiras questões com menor e maior dificuldade presentes nas Tabelas 15 e 16, respectivamente. A proporção refere-se à frequência relativa de acertos por sexo, contabilizando todos os participantes que responderam a questão, independentemente da cor do caderno que responderam. Pode-se observar que a proporção de acertos permanece semelhante nas questões das áreas STEM e não-STEM em ambos os sexos. As questões com maior proporção de acertos variam de 50% a 80%, sendo questões da área não-STEM com maior proporção de acertos. As questões com menor proporção

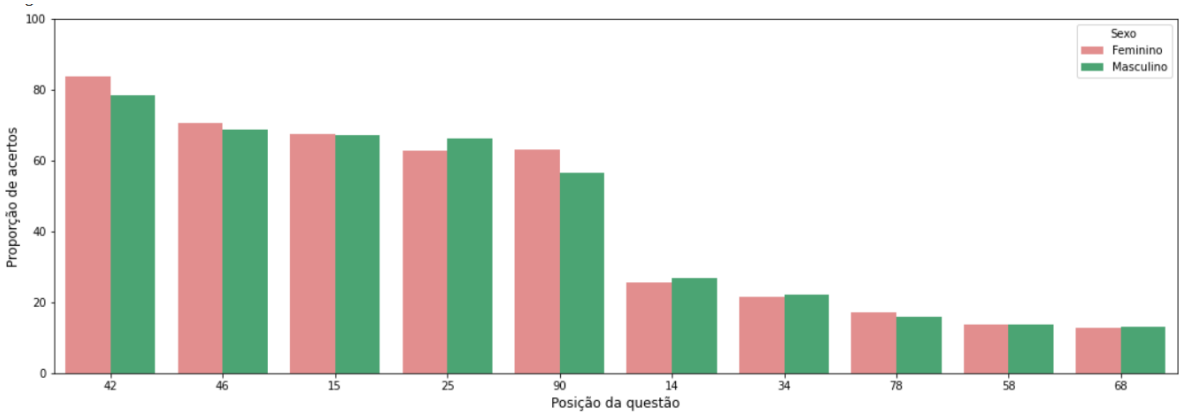
de acertos variam de 10% a 40%. As questões com maior índice de dificuldade apresentam menor proporção de acertos, enquanto as questões com menor índice de dificuldade apresentam maior proporção de acertos.

Figura 27 – Primeiras 5 questões com dificuldade baixa e primeiras 5 questões com dificuldade alta na área STEM.



Fonte: Elaborada pelo autor.

Figura 28 – Primeiras 5 questões com dificuldade baixa e primeiras 5 questões com dificuldade alta na área não-STEM.



Fonte: Elaborada pelo autor.

7.3 Análise dos resultados

Os resultados obtidos demostraram a existência de diferença estatística entre a proporção de acertos nas questões classificadas como STEM entre os participantes de sexo feminino e masculino. Adicionalmente, os resultados mostraram a presença de tamanho do efeito médio, indicando que o tamanho da diferença das medianas entre os grupos analisados é médio. Os resultados corroboram que a diferença de desempenho entre participantes de sexo feminino e masculino não é tão acentuada, assim como obtido nas análises realizadas no Capítulo 6.

O índice de dificuldade do item, um dos elementos da estimativa de parâmetros dos itens adotados pela metodologia TRI, permitiu identificar questões com maiores e menores níveis de dificuldade. Os cálculos envolvendo o índice de dificuldade dos itens possibilitam avaliar pontos críticos de aprendizagem que devem ser considerados no ensino do conteúdo. O alto valor desse índice indica que o assunto a que se refere deve ser abordado com cuidado pelos professores, que devem repensar suas práticas pedagógicas, muitas vezes tradicionais. A análise do conteúdo das questões, por sua vez, pode auxiliar os professores a enxergar as dificuldades pedagógicas dos alunos, apontando possíveis lacunas no aprendizado da disciplina. O índice de dificuldade do item também permite que os professores avaliem se as metodologias de ensino adotadas são realmente eficazes (SOARES; SOARES; SANTOS, 2021). Aos participantes do ENEM, esse índice possibilita a identificação das questões nas quais eles possuem dificuldades, norteando assuntos importantes a serem estudados.

Embora haja diferença estatística no desempenho de participantes do sexo feminino e masculino em diferentes áreas das questões classificadas como STEM, é pertinente refletir se existe um alinhamento entre o conhecimento que vem sendo construído nas escolas e aquele que foi avaliado no ENEM. Além disso, é importante levantar a discussão de que o desempenho do participante não está relacionado apenas ao índice de dificuldade do item, mas que ele pode ter vários fatores adicionais (CESTARO; KLEINKE; ALLE, 2020), como: não dominar a linguagem científica, falta de conhecimento em conteúdo específico, maior familiaridade com as palavras distratoras, uso do bom senso para responder, falta de raciocínio lógico, dificuldade de interpretação e distratores confusos.

7.4 Considerações finais

Neste capítulo foi descrito o terceiro cenário de uso da arquitetura proposta no Capítulo 4. Foi investigado o desempenho de participantes não treineiros com base no gênero nas diferentes questões STEM e não-STEM do ENEM dos anos de 2016 a 2020. Nas análises realizadas, foram utilizados testes de hipóteses, tendo sido usado o teste de normalidade *Kolmogorov-Smirnov* e o teste não paramétrico de *U de Mann-Whitney*. Além disso, foi obtida a estimativa de parâmetros dos itens adotados pela metodologia TRI com ajuda do pacote *sirt* da linguagem *R*.

Os resultados obtidos neste capítulo corroboraram conclusões já identificadas no Capítulo 6, isto é, a diferença de desempenho entre participantes de sexo feminino e masculino não é tão acentuada. Adicionalmente, os resultados obtidos também podem auxiliar na identificação de deficiências dos participantes e, desta maneira, contribuir com os gestores educacionais a traçar planos, metas e projetos que levem a uma melhoria na qualidade do ensino. É importante destacar que recorremos a um profissional estatístico com o problema e a ideia de solução, deste modo, foi tomada a decisão de utilizar as técnicas empregadas. Assim sendo, as técnicas e os resultados descritos neste capítulo foram avaliados por esse profissional, o qual sustenta que os resultados aqui apresentados são pertinentes.

No próximo capítulo, Capítulo 8, é descrito o quarto cenário de uso da arquitetura proposta, o qual tem como objetivo desenvolver um *data warehouse* com os dados do ENEM dos anos de 2016 e 2020. O objetivo é disponibilizar um banco de dados orientado a assunto, histórico, integrado e não-volátil que possa ser utilizado como suporte para prover flexibilidade nas análises realizadas pelos usuários de suporte à decisão educacional.

CAPÍTULO

8

PROPOSTA DE DW PARA TOMADA DE DECISÃO EDUCACIONAL

Neste capítulo é detalhado o quarto cenário de uso da arquitetura. O objetivo consiste em desenvolver um *data warehouse* com os dados do ENEM dos anos de 2016 a 2020. O propósito é oferecer suporte para que esses dados possam ser usados como base para a tomada de decisão educacional. Os dados seguem a modelagem multidimensional. Isso possibilita a realização de análises segundo diferentes perspectivas, as quais são exploradas neste capítulo. Na construção e implementação do *data warehouse* foram utilizadas ferramentas de código aberto e *frameworks* de processamento paralelo e distribuído como Apache Hadoop, Apache Spark, Apache Hive, Apache Airflow e Metabase.

Na seção 8.1 é descrito como a arquitetura proposta é instanciada para oferecer suporte para as análises realizadas. Na seção 8.2 é descrito o esquema lógico construído. Na seção 8.3 são discutidos os resultados obtidos. O capítulo é finalizado na seção 8.4 com as considerações finais.

8.1 Instanciação da arquitetura proposta

A instância da arquitetura proposta empregada para oferecer suporte às análises descritas neste capítulo faz referência ao *pipeline* para utilização de *Data warehouse* com ferramentas de análise estatística e com a especificação de consultas *On-Line Analytical Processing* (OLAP) ilustrada na Figura 11 (seção 4.2). Na camada de conexão de dados, as fontes de dados utilizadas correspondem aos dados do ENEM para os anos 2016 até 2020.

Os microdados do ano 2020 foram extraídos da página oficial do Inep. Para os anos anteriores, os dados brutos do Inep encontravam-se armazenados localmente. Todos os microdados foram posteriormente armazenados no formato *parquet* no HDFS. Os conjuntos de dados brutos referentes aos candidatos inscritos nos anos de 2016 a 2020 possuíam 8.627.367, 6.731.341,

5.513.747, 5.095.270 e 5.783.109 instâncias, respectivamente.

O processo ETL foi realizado com ajuda das linguagens *Python* e *Pyspark*. Devido à Lei Geral de Proteção de Dados (LGPD), o Inep adotou um modelo simplificado de microdados com o objetivo de eliminar da base pública variáveis que facilitem a identificação indevida do participante. Com isso, foram eliminadas variáveis relevantes como município de nascimento, residência do participante e código da escola. Para o ano 2020, o valor da variável município de residência da escola foi utilizado como referência de município, e a escola foi identificada por meio de um código criado com o valor 99.999.

As tabelas do repositório de dados organizados multidimensionalmente foram geradas por meio do uso do *Apache Hive* em conjunto com Spark. *Apache Hive* funciona internamente com o Spark como mecanismo de execução. Quando o Spark é construído para interagir com Hive, ele interage por meio do *Hive Metastore*, o qual é um catálogo do sistema que contém metadados sobre as tabelas armazenadas no Hive. Esses metadados são especificados durante a criação das tabelas e reutilizados toda vez que as tabelas são referenciadas.

Isso significa que o compartilhamento de vários sistemas, como HiveServer2 e Spark ThriftServer, é feito usando o catálogo de metadados. O catálogo de metadados emprega a mesma definição de bancos de dados e tabelas, incluindo a localização dos arquivos de dados (ou seja, os diretórios do HDFS). A configuração do *Hive Metastore* foi feita usando tabelas no banco de dados relacional MySQL com o objetivo de armazenar os metadados como nomes de tabelas, nomes de colunas, tipos de dados e comentários, dentre outros. Com Pyspark, o SparkSQL realiza as consultas OLAP nas tabelas Hive. Os dados foram visualizados usando a ferramenta Metabase e a governança de dados foi feita empregando-se Apache Atlas.

A configuração utilizada na camada de gerenciamento de dados incluiu um *cluster* de 5 computadores comerciais idênticos, sendo 1 mestre e 4 escravos. Todos os computadores executam Hadoop 3.2.2, Apache Spark 3.1.2 e Apache Hive 3.1.2 em uma instalação GNU/Linux (Ubuntu 20.04). Adicionalmente, cada nó possui CPU i5 8400 e 8 GB de memória. YARN foi utilizado como gerenciador do *cluster*.

8.2 Esquema lógico do DW proposto

Nesta seção é descrito o esquema lógico construído. As Figuras 29, 30, 31 e 32 ilustram o esquema lógico do DW proposto. Devido ao tamanho, o esquema foi dividido em diferentes figuras, cada uma referente a uma tabela de fatos distinta. O objetivo é prover suporte para o armazenamento dos dados do ENEM relativos aos anos de 2016 a 2020, sendo que esses dados encontram-se organizados multidimensionalmente.

São propostas quatro tabelas de fatos: (i) Perfil_participante, que armazena a quantidade de participantes com características específicas; (ii) Média, que armazena a média das notas; (iii)

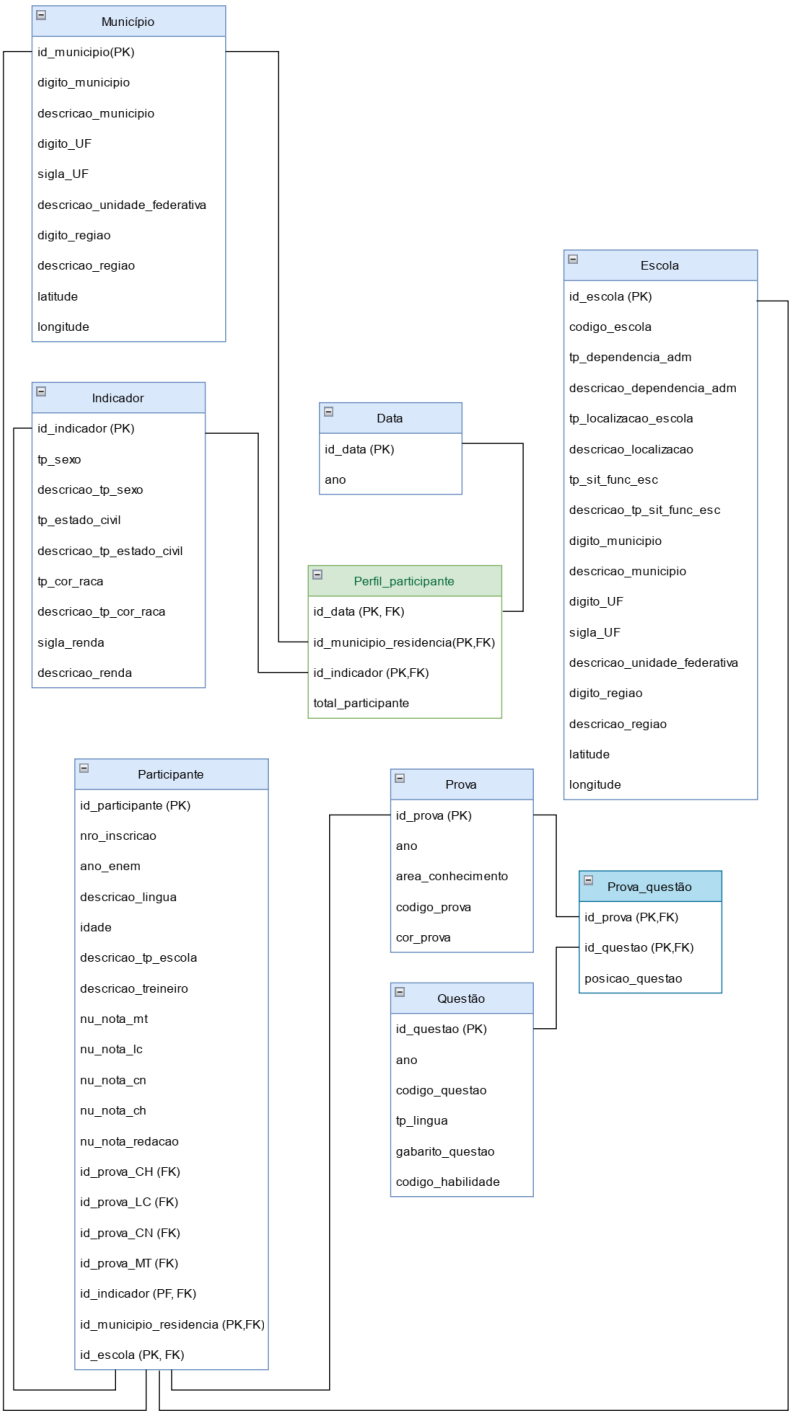
Participante_questões, que armazena para cada participante, o total de questões com acertos, erros, questões deixadas em branco e com dupla marcação; (iv) Questões_acertos, que armazena a porcentagem de acertos por cada questão.

No total, existem 12 tabelas de dimensão no repositório de dados organizados multidimensionalmente:

- Data, que armazena o ano em que ocorreu a prova.
- Participante, que armazena os dados pessoais dos participantes.
- Prova_participante, que armazena os dados particulares das provas realizadas pelos participantes.
- Questionario_socioeconomico_por_ano, que armazena o questionário socioeconômico.
- Resposta_participante_questionario_socioeconomico, que armazena as respostas de cada participante aos distintos itens do questionário socioeconômico.
- Escola, que armazena os dados específicos das diferentes escolas de ensino médio informadas pelo participante.
- Município, que armazena os dados da localização geográfica dos distintos municípios.
- Prova, que armazena os dados das provas.
- Questão, que armazena as diferentes questões de cada prova.
- Prova_questão, que armazena a posição de uma questão em uma prova específica.
- Indicador, que armazena um agrupamento de indicadores de baixa cardinalidade, como sexo, estado civil, cor-raça e renda mensal.
- Indicador_sexo, que armazena o indicador do sexo.

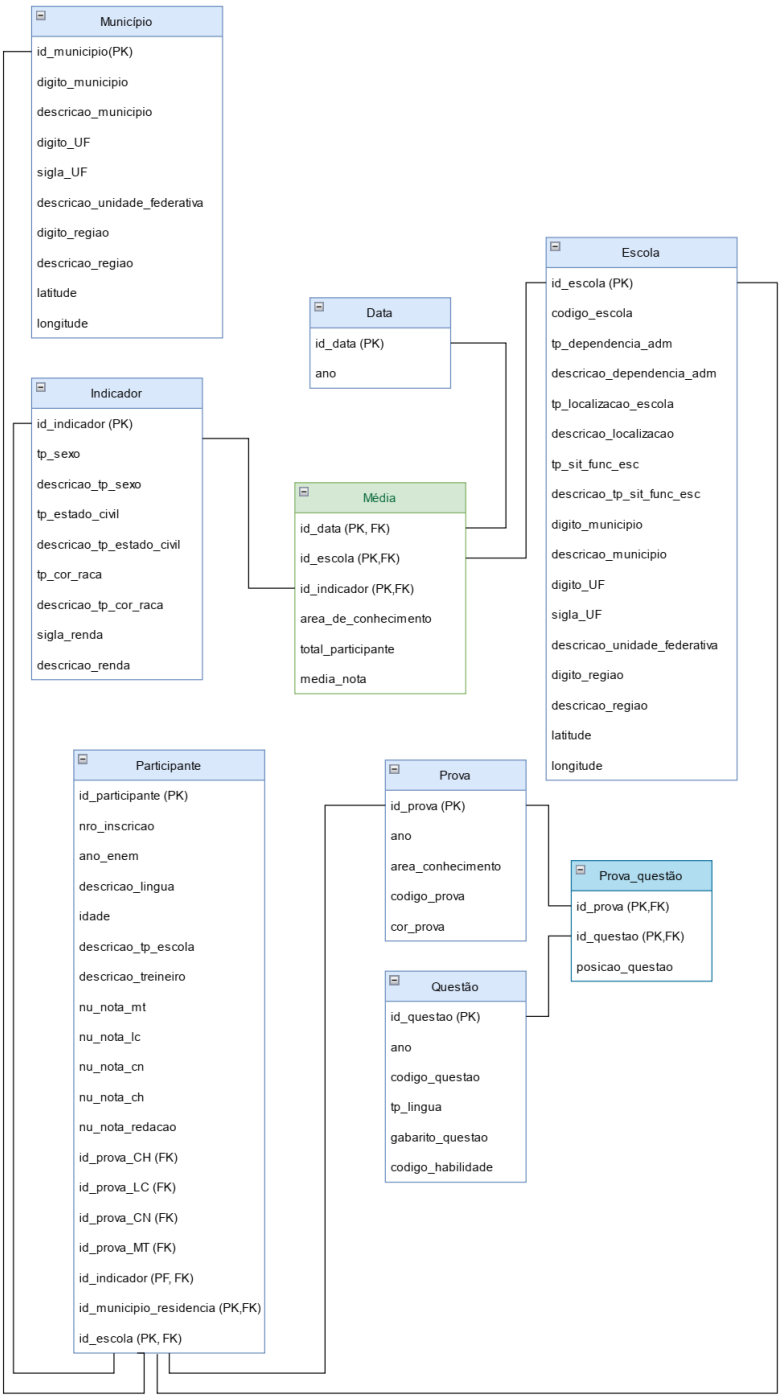
Os esquemas propostos compartilham as dimensões Indicador, Município, Participante, Prova e Questão, formando uma constelação de fatos e possibilitando a realização de consultas analíticas que relacionam medidas numéricas de esquemas diferentes. As características das tabelas do repositório de dados organizados multidimensionalmente são descritas na Tabela 17. São detalhados, para cada tabela gerada no Hive, a quantidade de instâncias da tabela e o tamanho ocupado pela tabela no disco. O volume total do conjunto de dados é de, aproximadamente, 23 GB.

Figura 29 – Esquema lógico do DW.



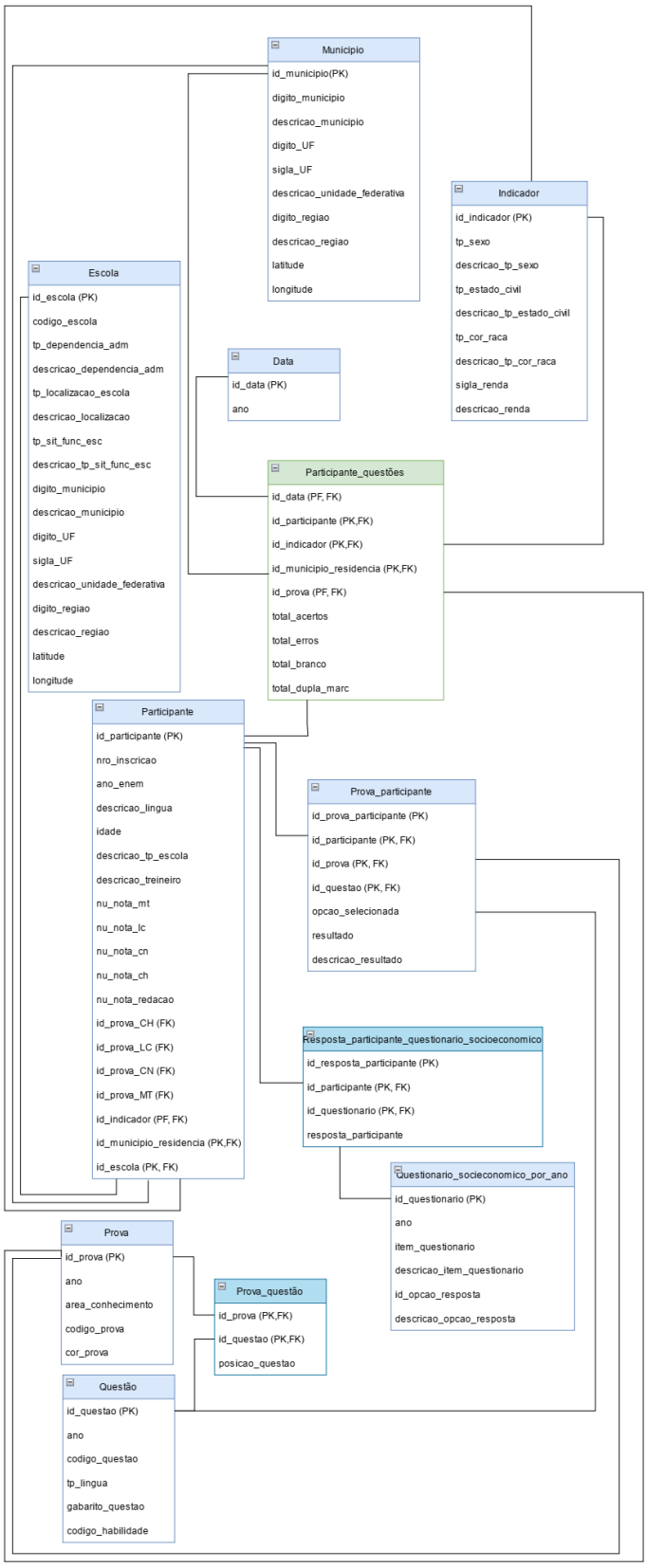
Fonte: Elaborada pelo autor.

Figura 30 – Esquema lógico do DW (Continuação).



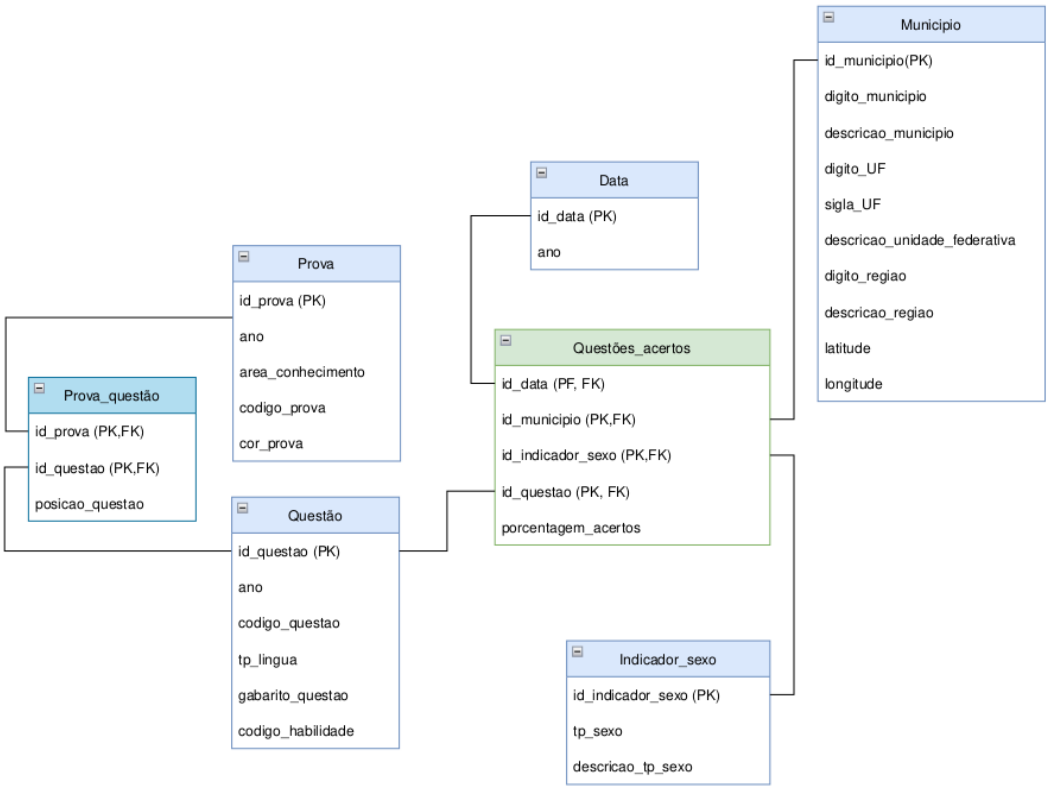
Fonte: Elaborada pelo autor.

Figura 31 – Esquema lógico do DW (Continuação).



Fonte: Elaborada pelo autor.

Figura 32 – Esquema lógico do DW (Continuação).



Fonte: Elaborada pelo autor.

Tabela 17 – Características das tabelas do repositório de dados organizados multidimensionalmente.

Nome da tabela	Quantidade de instâncias	Tamanho
Data	6	210 B
Participante	31.750.825	672.3 MB
Prova_participante	3.212.679.132	16 GB
Questionario_socioeconomico_por_ano	877	38.1 kB
Resposta_participante_questionario _socioeconomico	998.536.005	5.2 GB
Escola	45.939	2.8 MB
Município	5.570	176.4 kB
Prova	290	10.9 kB
Questão	2.284	34.4 kB
Prova_questão	12.650	40.3 kB
Indicador	1.020	7.2 kB
Indicador_sexo	2	40 B
Perfil_participante	2.358.815	8.6 MB
Média	16.052.673	122.8 MB
Participante_questões	60.749.069	581.3 MB
Questões_acertos	9.399.407	35.9 MB

Fonte: Elaborada pelo autor.

8.3 Resultados e discussões

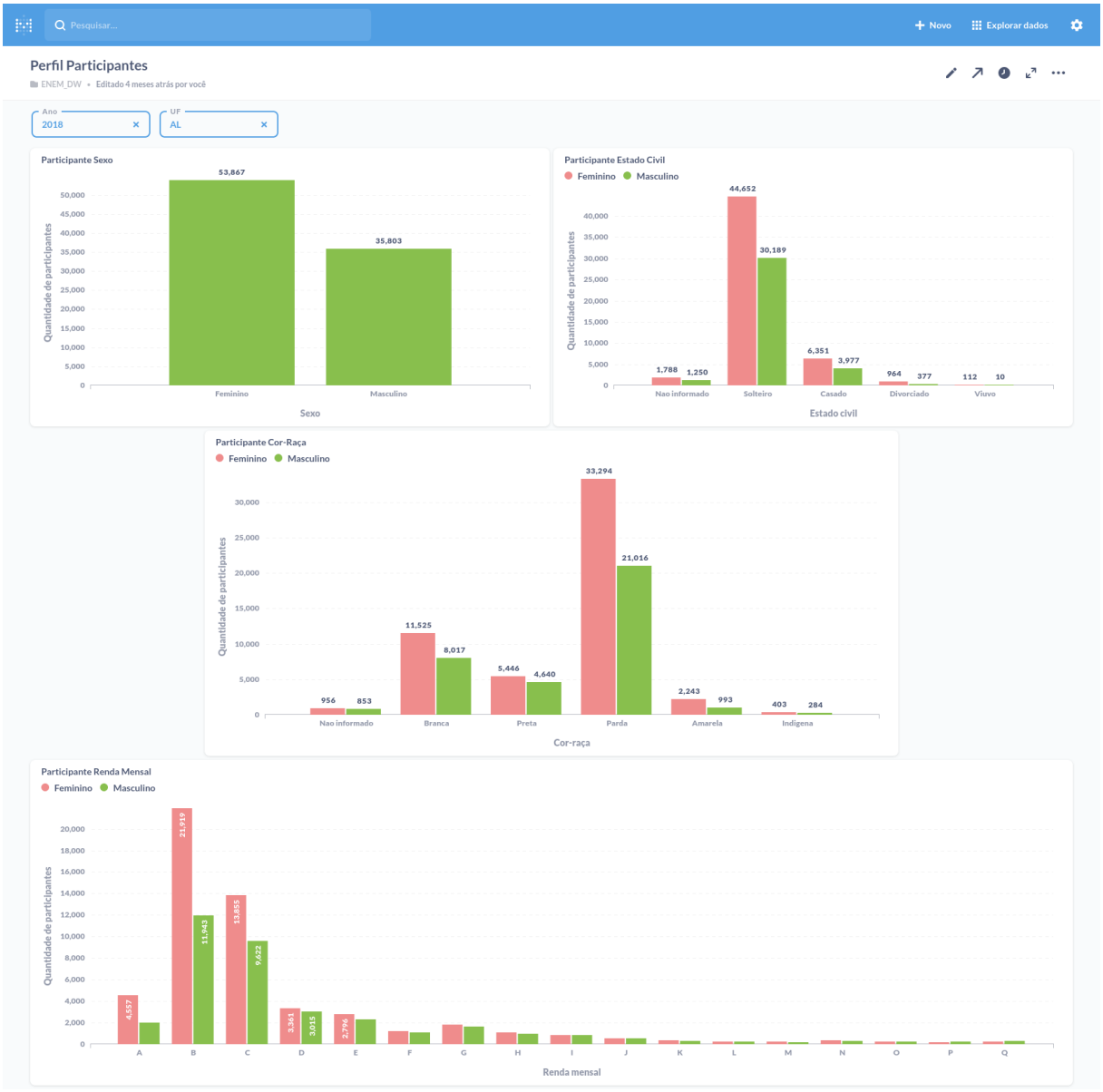
Nesta seção, são descritas diferentes consultas analíticas que podem ser respondidas pelo *Data warehouse*, componente da camada de gerenciamento de dados.

O código mostrado na Consulta 1 retorna a quantidade de participantes do ENEM no ano de 2018 por sexo, considerando a unidade federativa de Alagoas (AL). Para responder à consulta foi utilizado o esquema proposto na Figura 29. Na Figura 33 são mostrados os resultados obtidos, como total de participante por sexo, estado civil, cor e raça e renda mensal. A primeira imagem ilustra a resposta à Consulta 1, sendo 53.867 e 35.803 o total de participantes femininos e masculinos, respectivamente.

```
1: SELECT da.ano, m.sigla_unidade_federativa, d.descricao_tp_sexo,
2: SUM(p.total_participante) quantidade
3: FROM perfil_participante p
4: JOIN indicador d ON p.id_indicador=d.id_indicador
5: JOIN municipio m ON m.id_municipio=p.id_municipio_residencia
6: JOIN data da ON da.id_data=p.id_data
7: WHERE da.ano = 2018 AND m.sigla_unidade_federativa = "AL"
8: GROUP BY da.ano, m.sigla_unidade_federativa,  d.descricao_tp_sexo
9: ORDER BY  d.descricao_tp_sexo ASC
```

Consulta 1 – Qual é o total de participantes do ENEM no ano de 2018 por sexo, considerando a unidade federativa de Alagoas (AL)?

Figura 33 – *Dashboard* que ilustra a resposta à Consulta 1: Qual é o total de participantes do ENEM no ano de 2018 por sexo, considerando a unidade federativa de Alagoas (AL)?



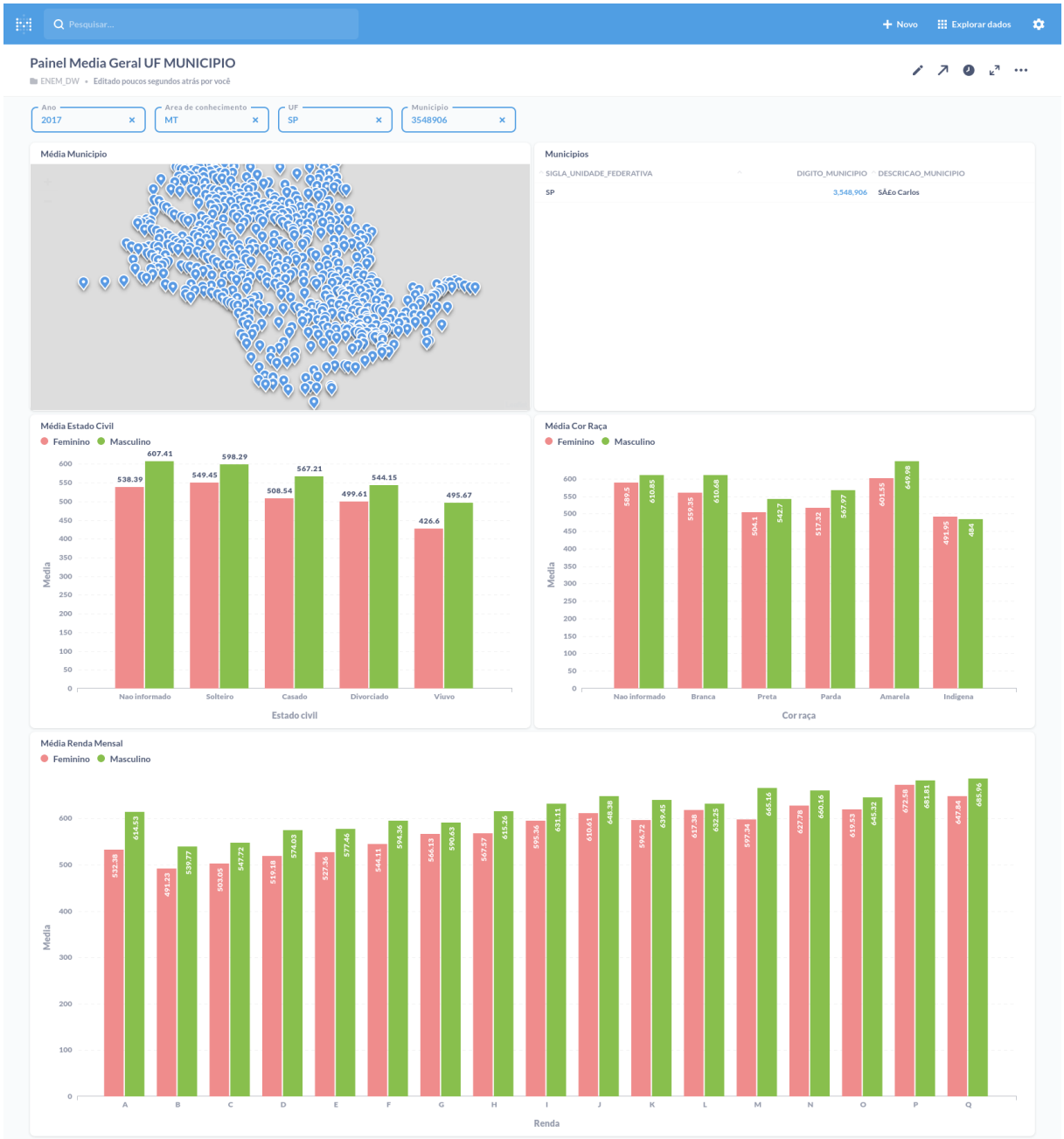
Fonte: Elaborada pelo autor.

O código mostrado na Consulta 2 retorna a média das notas da área de conhecimento de *Matemática* (MT) por renda mensal, considerando as escolas de São Carlos da unidade federativa de São Paulo (digito_municipio = 3548906) no ano 2017. Para responder à consulta foi utilizado o esquema proposto na Figura 30. Na Figura 34 são mostrados os resultados obtidos, como a média dos participantes por sexo, estado civil, cor, raça e renda mensal. A última imagem ilustra a resposta à Consulta 2. A sigla da renda mensal é apresentada de forma crescente, sendo A nenhuma renda e Q a renda mais alta. Pode-se perceber que, à medida que a renda aumenta, o desempenho dos participantes melhora.

```
1: SELECT f.area, e.digito_municipio, e.descricao_municipio, d.tp_sexo,
2: d.descricao_tp_sexo, d.sigla_renda,
3: ROUND(SUM(f.suma_nota)/SUM(f.total_participante),2) media
4: FROM media f
5: JOIN indicador d ON f.id_indicador=d.id_indicador
6: JOIN escola e ON e.id_escola=f.id_escola
7: JOIN data da ON da.id_data=f.id_data
8: WHERE da.ano = 2019 and f.area = "MT" AND e.sigla_UF="SP" AND
9: e.descricao_municipio= "São Carlos"
10: GROUP BY  f.area, e.digito_municipio, e.descricao_municipio,d.tp_sexo,
11: d.descricao_tp_sexo, d.sigla_renda
12: ORDER BY d.descricao_tp_sexo, d.sigla_renda ASC
```

Consulta 2 – Qual é a média das notas da área de conhecimento de *Matemática* (MT) por renda mensal, considerando as escolas de São Carlos da unidade federativa de São Paulo (digito_municipio = 3548906) no ano 2017?

Figura 34 – *Dashboard* que ilustra a resposta à Consulta 2: Qual é a média das notas da área de conhecimento de *Matemática* (MT) por renda mensal, considerando as escolas de São Carlos da unidade federativa de São Paulo (digito_municipio = 3548906) no ano 2017?



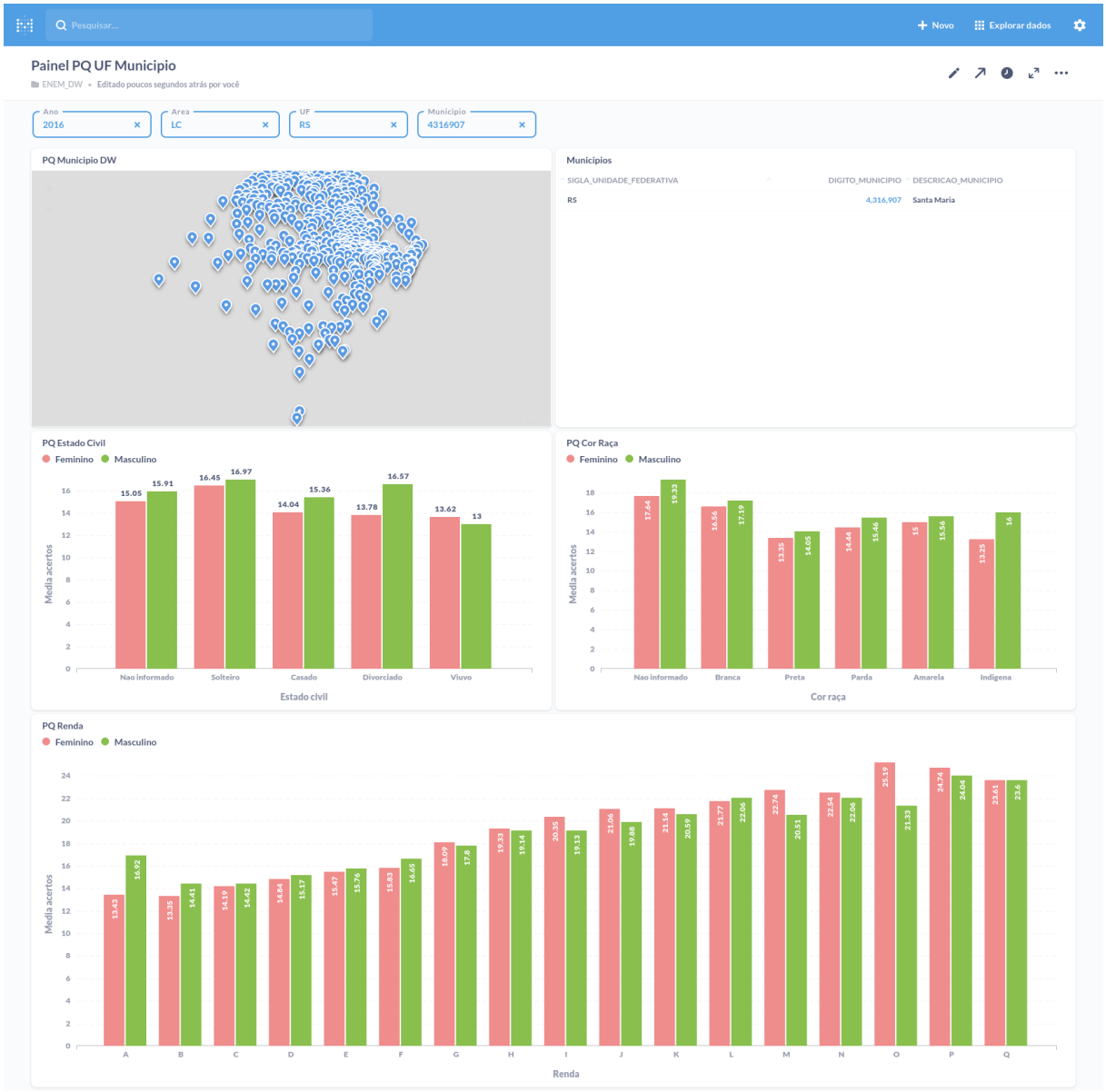
Fonte: Elaborada pelo autor.

O código mostrado na Consulta 3 retorna a média de acertos na prova de *Linguagens e Códigos* por sexo e estado civil, considerando o município de Santa Maria no estado do Rio Grande do Sul (digito_municipio = 4316907) e o ano de 2016. Para responder à consulta foi utilizado o esquema proposto na Figura 31. Na Figura 35 são mostrados os resultados obtidos considerando diferentes perspectivas como sexo, estado civil, cor, raça e renda mensal. A segunda imagem à esquerda ilustra a resposta à Consulta 3. Participantes solteiros femininos e masculinos têm, em média, maior número de acertos do que os participantes dos demais estados civis.

```
1: SELECT  f.ano,p.area, m.digito_municipio, m.descricao_municipio,
2: d.tp_sexo, d.descricao_tp_sexo,
3: d.tp_estado_civil, d.descricao_tp_estado_civil,
4: COUNT(f.id_participante) quantidade_participantes,
5: ROUND(SUM(f.acertos)/COUNT(f.id_participante),2) media_acertos
6: FROM participante_questoes f
7: JOIN indicador d on f.id_indicador=d.id_indicador
8: JOIN municipio m on f.id_municipio_residencia=m.id_municipio
9: JOIN prova p on p.id_prova = f.id_prova AND p.ano=f.ano
10: JOIN data da on da.id_data=f.id_data
11: WHERE da.ano=2016 and p.area="LC" AND m.siglaUF = "RS" AND
12: m.descricao_municpio= "Santa Maria"
13: GROUP BY  f.ano,p.area, m.digito_municipio, m.descricao_municipio,
14: d.tp_sexo, d.descricao_tp_sexo,
15: d.tp_estado_civil,d.descricao_tp_estado_civil
16: ORDER BY d.descricao_tp_sexo, d.tp_estado_civil ASC
```

Consulta 3 – Qual a média de acertos na prova de *Linguagens e Códigos* por sexo e estado civil, considerando o município de Santa Maria no estado do Rio Grande do Sul (digito_municipio = 4316907) e o ano de 2016?

Figura 35 – *Dashboard* que ilustra a resposta à Consulta 3: Qual a média de acertos na prova de *Linguagens e Códigos* por sexo e estado civil, considerando o município de Santa Maria no estado do Rio Grande do Sul e o ano de 2016?



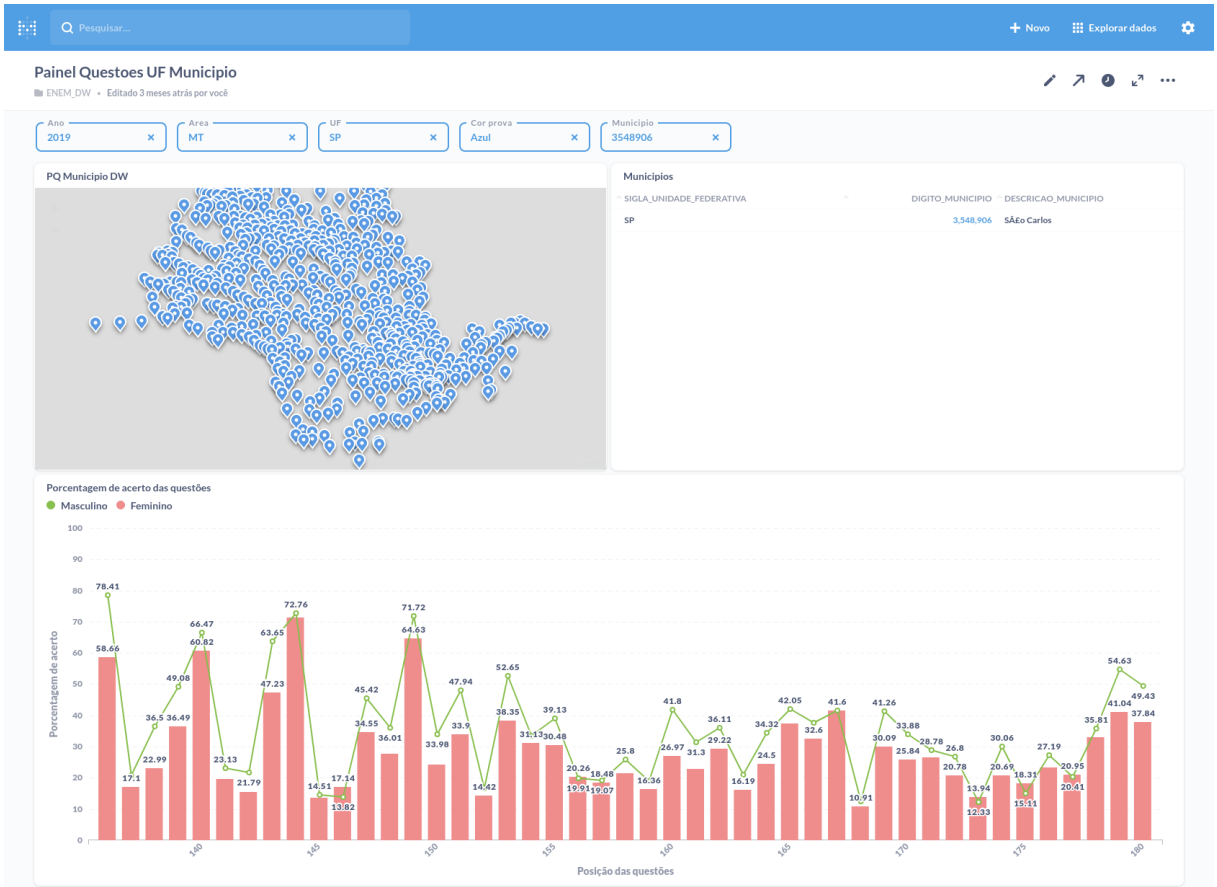
Fonte: Elaborada pelo autor.

O código mostrado na Consulta 4 retorna a porcentagem de acertos nas questões da área de conhecimento de *Matemática*, da prova de cor azul para o município de São Carlos do estado de São Paulo, por sexo, para o ano 2019. Para responder à consulta foi utilizado o esquema proposto na Figura 32. A segunda imagem da Figura 36 responde à pergunta da Consulta 4. A maioria das questões avaliadas em 2019 para *Matemática* revelam que os participantes masculinos acertaram mais questões do que as participantes femininas.

```
1: SELECT d.descricao_tp_sexo, pq.posicao_questao, porc_acertos
2: FROM questoesacertos f
3: JOIN questoes q ON f.id_questao=q.id_questao
4: JOIN prova_questao pq ON pq.id_questao=q.id_questao
5: JOIN prova p ON p.id_prova=pq.id_prova
6: JOIN municipio m ON m.id_municipio=f.id_municipio_residencia
7: JOIN data da ON da.id_data=f.id_data
8: JOIN indicador_sexo d ON f.id_indicador_sexo=d.id_indicador_sexo
9: WHERE da.ano_fqa=q.ano AND da.ano_fqa=2019 AND p.area="MT" AND
10: p.cor_prova="Azul" AND m.sigla_UF="SP" AND
11: m.descricao_municipio= "São Carlos"
12: ORDER BY pq.posicao_questao ASC;
```

Consulta 4 – Qual a porcentagem de acertos nas questões da área de conhecimento de *Matemática*, da prova de cor azul para o município de São Carlos do estado de São Paulo, por sexo, para o ano 2019?

Figura 36 – *Dashboard* que ilustra a resposta à Consulta 4: Qual a porcentagem de acertos nas questões da área de conhecimento de *Matemática*, da prova de cor azul para o município de São Carlos do estado de São Paulo, por sexo, para o ano 2019?



Fonte: Elaborada pelo autor.

8.4 Considerações finais

Neste capítulo foi descrito o quarto cenário de uso da arquitetura proposta no Capítulo 4. Foi proposto o esquema de um *data warehouse* composto por 12 tabelas de dimensões e 4 tabelas de fatos. O esquema foi preenchido com dados dos exames do ENEM realizados nos anos de 2016 a 2020. Também foram especificados vários exemplos de consultas analíticas incidindo sobre esse *data warehouse*, as quais foram visualizadas graficamente por meio de *dashboards* interativos.

O objetivo principal de se implementar e carregar o *data warehouse* com os dados do ENEM é oferecer suporte para que esses dados possam ser usados como base para a tomada de decisão educacional. Os dados seguem a modelagem multidimensional, possibilitando análises segundo diferentes perspectivas. Como resultado, não somente as análises realizadas nesta tese podem ser realizadas, mas também diversas outras análises que sejam relevantes às instituições de ensino. A partir dessas análises, pode-se definir indicadores de qualidade, traçar metas de investimento, realizar planejamentos estratégicos e (re)formular políticas públicas para aprimoramento da qualidade do ensino ofertado.

No próximo capítulo, Capítulo 9, é descrita a conclusão da tese.

CAPÍTULO

9

CONCLUSÕES

Neste capítulo é feita a conclusão da tese de doutorado. Na seção 9.1 é resumido o trabalho desenvolvido e na seção 9.2 são listados os artigos decorrentes do trabalho. Na seção 9.3 são descritas as dificuldades encontradas. Por fim, na seção 9.4 são listados trabalhos futuros.

9.1 Trabalho desenvolvido

Nesta tese foi proposta uma arquitetura baseada em *data warehousing*, *mineração de dados*, *estatística inferencial* e *processamento paralelo e distribuído* voltada à análise de dados educacionais do ENEM. Usando como base esta arquitetura, foi possível realizar diferentes investigações a respeito do desempenho dos participantes do ENEM.

A arquitetura é composta de cinco camadas diferentes. A camada de conexão de dados engloba as fontes de dados e os processos ETL/ELT. A camada de gerenciamento de dados inclui os componentes *Data lake* e *Data warehouse* para armazenamento dos dados, além de funcionalidades relacionadas à Governança de Dados. A camada de análise de dados, por sua vez, possibilita a realização de consultas OLAP, a aplicação de técnicas de mineração de dados e aprendizado de máquina e o uso de ferramentas estatísticas. A camada de apresentação de dados inclui ferramentas de visualização de dados. Por fim, a camada de gerenciador de fluxo de trabalho é utilizada para definir *workflows* que envolvam a automação de diferentes processos realizados na arquitetura.

Com base na arquitetura proposta, foram definidos dois *pipelines*, um no contexto de *Data lake* e outro considerando *Data warehouse*. O objetivo desses pipelines é auxiliar os gestores no processo de implementação da arquitetura de acordo com as necessidades de tomada de decisão educacional. Em especial, esses *pipelines* foram utilizados como base para os cenários de uso desenvolvidos nesta tese, e descritos a seguir.

A arquitetura proposta foi validada por meio de quatro cenários de uso. Os diferentes

períodos de tempo nas análises estão relacionados com a disponibilidade dos dados na época dessas análises. O primeiro deles foi voltado à identificação e análise dos principais indicadores de desempenho dos participantes do ENEM nos anos de 2017 a 2020 com o objetivo de avaliar a eficácia educacional. Foram utilizados os classificadores árvore de decisão, máquinas de vetores de suporte e redes *MultiLayer Perceptron* para investigar dados pessoais, do questionário socioeconômico, das notas da *Redação* e das 180 questões de múltipla escolha relacionadas a todas as áreas de conhecimento. A partir dos resultados obtidos, foi feita uma discussão considerando aspectos relacionados aos estudos e políticas educacionais existentes.

A análise realizada no primeiro cenário de uso não considerou separadamente o desempenho de participantes femininos e masculinos, uma temática atual e de grande interesse. Neste sentido, o segundo cenário de uso investigou o desempenho dos participantes do ENEM na temática "gêneros e suas nuances na tecnologia da informação". O desempenho dos participantes femininos e masculinos para cada área de conhecimento relacionada às ciências exatas foi analisado considerando dois direcionamentos. O primeiro investigou os fatores de análise por região, cor/raça, tipo de escola de ensino médio e renda salarial mensal para os anos de 2013 a 2017. O segundo direcionamento aprofundou a investigação no ano de 2017 e nas regiões do Brasil, deixando em foco as unidades federativas. Os resultados demonstraram que a diferença de desempenho entre participantes dos sexos femininos e masculinos não foi tão acentuada quando comparada com o desequilíbrio desses sexos nos cursos na área de ciências exatas e nos profissionais que atuam nessas áreas.

Visando ratificar os resultados obtidos no segundo cenário de uso, foi desenvolvido um terceiro cenário que também investigou a temática "gêneros". Porém, na análise realizada, foi considerado o desempenho dos participantes em questões STEM e não-STEM do ENEM dos anos de 2016 a 2020. Questões STEM se referem àquelas pertencentes às áreas de conhecimento de *Matemática* e *Ciência da Natureza* em adição a algumas questões específicas de *Linguagens e Códigos* (questões identificadas pelos códigos de habilidade 28, 29 e 30). As demais questões são classificadas como não-STEM. Nas análises realizadas, foram utilizados testes de hipóteses e o índice de dificuldade do item, um dos elementos da estimativa de parâmetros dos itens adotados pela metodologia TRI. Os resultados corroboram que a diferença de desempenho entre participantes de sexo feminino e masculino não foi tão acentuada.

Os dados do ENEM podem ser analisados considerando diferentes perspectivas, de acordo com as necessidades de tomada de decisão estratégica educacional. Desta forma, existe a necessidade de se organizar e disponibilizar esses dados para que eles possam ser amplamente utilizados. Este é o objetivo do quarto cenário de uso da arquitetura, no qual foi projetado um esquema multidimensional para armazenar os dados do ENEM dos anos de 2016 a 2020. O esquema foi desenvolvido de acordo com os seguintes assuntos de interesse: quantidade de participantes com características específicas, média das notas dos participantes, total de acerto por questões de cada participante e porcentagem de acertos de cada questão. Também

foram exemplificadas diferentes consultas analíticas que podem ser executadas sobre os dados armazenados. Como resultado do desenvolvimento do quarto cenário de uso, não somente as análises apresentadas nesta tese podem ser realizadas, mas também diversas outras análises que sejam relevantes às instituições de ensino e aos órgãos governamentais.

9.2 Publicações

Durante o desenvolvimento desta tese, foram publicados os seguintes artigos.

- NOGUERA, V.; BRANCO, K.; CIFERRI, C. Gêneros e suas nuances no ENEM. In: **Anais do XIII Women in Information Technology**. Belém, PA, Brasil: SBC, 2019. p. 41–50. O artigo recebeu o prêmio de melhor artigo do evento XIII Women in Information Technology.
- NOGUERA, V. E. R.; BRANCO, K. R. L. J. C.; CIFERRI, C. D. de A. Análise de desempenho das mulheres no ENEM. **Brazilian Journal of Development**, v. 6, n. 6, p. 35716–35737, 2020.

Adicionalmente, foi submetido o seguinte artigo, o qual se encontra em fase de análise:

- NOGUERA, V. E. R.; BRANCO, K. R. L. J. C.; AGUIAR, C. D. Identifying and analyzing key performance indicators from the ENEM data to support educational public policies. **Educational Assessment**.

9.3 Dificuldades no desenvolvimento do trabalho

Foram encontradas dificuldades relacionadas às mudanças no modelo de microdados do ENEM, ao tempo dispendido para realizar os processos ELT/ELT e ao uso do ecossistema Hadoop. Com relação às mudanças que ocorreram, devido à vigência da LGPD, o Inep adotou um modelo simplificado de microdados com o objetivo de eliminar da base pública variáveis que facilitem a identificação indevida do participante. Com isso, variáveis relevantes para as análises apresentadas nesta tese foram eliminadas. Contudo, esse inconveniente foi contornado com *backups* armazenados localmente no computador da autora da tese e com decisões efetuadas nos processos ETL/ELT, tomando o cuidado de anonimizar esse dados.

Adicionalmente, as variáveis do ENEM mudam ano a ano. Devido a essa diferença na estrutura dos dados, os mesmos passaram por um processo de tratamento completo e meticuloso para realizar a padronização específica ao modelo multidimensional proposto. Com o grande volume de dados considerado nas análises realizadas, o tempo gasto na realização das atividades do processo ETL/ELT foi muito grande.

A manipulação do grande volume de dados do ENEM requereu o uso de ambientes computacionais paralelos e distribuídos. A adaptação das bibliotecas para não criar incompatibilidade entre as diferentes ferramentas utilizadas no ecossistema Hadoop foi trabalhosa e demorada. Além disso, a inicialização do Apache Hive exigiu uma sequência de passos não usuais para conexão com SparkSQL, tais como inicializar o *Metastore*, *Spark ThriftServer* e *HiveServer2*.

9.4 Trabalhos futuros

Existem vários pontos de pesquisa que podem ser explorados em continuação ao trabalho desenvolvido, os quais incluem:

- Avaliar e comparar o desempenho das consultas OLAP, com mecanismos de armazenamento, como Apache Kudu e Apache Kylin[1], e formatos de arquivos *parquet* bem como ORC.
- Expandir o modelo multidimensional do *data warehouse*, considerando diversos indicadores como questões do questionário socioeconômico, para responder um número maior de perguntas.
- Realizar novas análises de dados considerando grupos específicos, como participantes com deficiência, participantes da segunda aplicação da prova e participantes que realizaram a prova digital.
- Explorar as notas da *Redação* considerando as diferentes competências avaliadas, de maneira a entender melhor o baixo desempenho na *Redação*.
- Investigar a aplicação de técnicas de mineração de dados e análises estatísticas não abordadas nesta tese, tais como aplicação de modelo de regressão logística, regressão linear, agrupamentos (FACELI *et al.*, 2011) e técnicas multivariadas exploratórias (FÁVERO; BELFIORE, 2017).
- Estudar o impacto das posições das questões nas diferentes cores de caderno para investigar o efeito fadiga, ou seja, a frequência de respostas corretas deve ser igual, independentemente da posição da questão.
- Integrar outras bases de dados educacionais com os dados do ENEM, tais como os dados do Sisu e ProUni.

[1] <https://kylin.apache.org/>

REFERÊNCIAS

ALLEN, L. K.; SNOW, E. L.; CROSSLEY, S. A.; JACKSON, G. T.; MCNAMARA, D. S. Reading comprehension components and their relation to writing. **LAnnee psychologique**, NecPlus, v. 114, n. 4, p. 663–691, 2014. Disponível em: <https://www.cairn.info/revue-l-annee-psychologique1-2014-4-page-663.htm>. Citado na página 90.

ALVES, M. T. G.; NOGUEIRA, M. A.; NOGUEIRA, C. M. M.; RESENDE, T. d. F. Fatores familiares e desempenho escolar: uma abordagem multidimensional. **Dados**, SciELO Brasil, v. 56, p. 571–603, 2013. Disponível em: <https://www.scielo.br/j/dados/a/5t5Dcx9ZVqTykv6hF8CvHRc/?lang=pt>. Citado na página 88.

ALVES, R. D. **Predição do desempenho da redação do ENEM utilizando técnicas de mineração de dados**. Monografia (Graduação) — Universidade Federal de Santa Catarina, 2018. Disponível em: <https://repositorio.ufc.br/bitstream/riufc/44034/1/2018_eve_rdalves.pdf>. Citado nas páginas 63, 66 e 70.

ALVES, R. D.; CECHINEL, C.; QUEIROGA, E. Predição do desempenho de matemática e suas tecnologias do ENEM utilizando técnicas de mineração de dados. In: CBIE. **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. 2018. v. 7, n. 1, p. 469. Disponível em: <http://ojs.sector3.com.br/index.php/wcbie/article/view/8271>. Citado nas páginas 63, 66 e 70.

ANDRADE, D. F.; TAVARES, H. R.; VALLE, R. da C. Teoria da resposta ao item: conceitos e aplicações. **ABE, Sao Paulo**, 2000. Citado na página 49.

ANDRADE, L. F. **Influência das características socioecônomicas sobre o desempenho no ENEM em Ciências Exatas**. Monografia (Graduação) — Universidade Federal de Sergipe, 2019. Disponível em: <http://ri.ufs.br/jspui/handle/riufs/12212>. Citado nas páginas 57, 60 e 70.

ANDRADE, R. d. O. **A retomada do espaço da mulher na computação.** Revista Pesquisa FAPESP, 2019. Disponível em: <https://revistapesquisa.fapesp.br/a-retomada-do-espaco-da-mulher-na-computacao/>. Citado na página 90.

ANSONG, D.; OKUMU, M.; ALBRITTON, T. J.; BAHNUK, E. P.; SMALL, E. The role of social support and psychological well-being in stem performance trends across gender and locality: Evidence from ghana. **Child Indicators Research**, Springer, v. 13, n. 5, p. 1655–1673, 2020. Disponível em: <https://link.springer.com/article/10.1007/s12187-019-09691-x>. Citado na página 90.

ARAÚJO, E. S. C.; SILVA, H. O. d. M. **Aplicação de mineração de dados na descoberta dos fatores socioeconômicos associados com o desempenho dos participantes do ENEM**. Monografia (Graduação) — Universidade Evangélica de Goiás, 2020. Disponível em: <http://repositorio.aee.edu.br/jspui/handle/aee/17189>. Citado nas páginas 64, 66 e 70.

AZEVEDO, A. R. de; ARA, A.; NOGUTI, M. Y.; BRITO, A. C. de. Aplicação em shiny: Intersecção entre gênero, classe e raça no ENEM de 2016. **Revista do Seminário Internacional de Estatística com R**, v. 3, n. 1, 2018. Disponível em: <https://periodicos.uff.br/anaisdoser/article/view/29220/16951>. Citado nas páginas 58, 59, 62 e 70.

BAKER, F. B.; KIM, S.-H. **Item response theory: Parameter estimation techniques**. CRC Press, 2004. Disponível em: <https://www.taylorfrancis.com/books/mono/10.1201/9781482276725/item-response-theory-frank-baker-seock-ho-kim>. Citado na página 49.

BANNI, M. R.; OLIVEIRA, M. V. d. P.; BERNARDINI, F. C. Uma análise experimental usando mineração de dados educacionais sobre os dados do ENEM para identificação de causas do desempenho dos estudantes. In: SBC. **Anais do II Workshop sobre as Implicações da Computação na Sociedade**. 2021. p. 57–66. Disponível em: <https://sol.sbc.org.br/index.php/wics/article/view/15964/15805>. Citado nas páginas 65, 68 e 71.

BARCELLOS, A. A.; ISOTANI, S.; DIEGO, C.; DAMASCENO, N. Mineração de dados abertos-ENEM 2018. In: **Anais dos Trabalhos de Conclusao de Curso da Pós-Graduaçao em Computaçao Aplicadaa Educaçao**. [s.n.], 2020. Disponível em: <https://especializacao.icmc.usp.br/documentos/tcc/alessandro_barcellos.pdf>. Citado nas páginas 64, 66 e 70.

BARROSO, M. F.; RUBINI, G.; SILVA, T. d. Dificuldades na aprendizagem de física sob a ótica dos resultados do ENEM. **Revista Brasileira de Ensino de Física**, SciELO Brasil, v. 40, 2018. Disponível em: <https://www.scielo.br/j/rbef/a/WgC3RNzBBDTDvdkrfYJfxHQ/?format=pdf&lang=pt>. Citado nas páginas 55, 56, 60 e 70.

BATISTA, N. A.; SOUSA, G. A.; BRANDÃO, M. A.; SILVA, A. P. C. da; MORO, M. M. Tie strength metrics to rank pairs of developers from github. **Journal of Information and Data Management**, v. 9, n. 1, p. 69–83, 6 2018. Disponível em: <https://sol.sbc.org.br/journals/index.php/jidm/article/view/1637>. Citado na página 53.

BERNADO, E. S.; SILVA, F. G. Ensino médio e (m) tempo integral: uma breve análise das políticas públicas para os jovens do século xxi. **REVISTA BRASILEIRA DO ENSINO MÉDIO**, v. 2, p. 100–115, 2019. Disponível em: <https://phprbraem.com.br/ojs/index.php/RBRAEM/article/view/20>. Citado na página 89.

BERRA, M.; CAVALETTO, G. M. Overcoming the STEM gender gap: from school to work. **Italian Journal of Sociology of Education**, v. 12, n. 2, 2020. Disponível em: <http://ijse.padovauniversitypress.it/system/files/papers/IJSE-2020-2-1.pdf>. Citado na página 90.

BINDER, K. S.; COTE, N. G.; LEE, C.; BESSETTE, E.; VU, H. Beyond breadth: The contributions of vocabulary depth to reading comprehension among skilled readers. **Journal of research in reading**, Wiley Online Library, v. 40, n. 3, p. 333–343, 2017. Disponível em: <https://onlinelibrary.wiley.com/doi/epdf/10.1111/1467-9817.12069>. Citado na página 90.

BIOLCHINI, J. C. de A.; MIAN, P. G.; NATALI, A. C. C.; CONTE, T. U.; TRAVASSOS, G. H. Scientific research ontology to support systematic review in software engineering. **Advanced Engineering Informatics**, v. 21, n. 2, p. 133 – 151, 2007. ISSN 1474-0346. Disponível em: <http://www.sciencedirect.com/science/article/pii/S147403460600070X>. Citado nas páginas 51 e 52.

BRITO, J. J. **Data Warehouses na era do Big Data: processamento eficiente de Junções Estrela no Hadoop**. Tese (Doutorado) — Universidade de São Paulo, 2017. Citado nas páginas 37, 38 e 42.

BRUCE, P.; BRUCE, A. **Estatística prática para cientistas de dados - 50 conceitos essenciais**. [S.l.]: Editora Alta Books, 2019. Citado nas páginas 46 e 47.

CARMO, R. V. do; HECKLER, W. F.; CARVALHO, J. V. de. Uma análise do desempenho dos estudantes do rio grande do sul no ENEM 2019. **RENOTE**, v. 18, n. 2, p. 378–387, 2020. Disponível em: <https://www.seer.ufrgs.br/renote/article/viewFile/110257/60030>. Citado nas páginas 58, 61 e 70.

CESTARO, D. C.; KLEINKE, M. U.; ALLE, L. F. Uma análise do desempenho dos participantes e do conteúdo abordado em itens de genética e biologia evolutiva do exame nacional do ensino médio (ENEM): implicações curriculares. **Investigações em Ensino de Ciências**, v. 25, n. 3, 2020. Citado nas páginas 55, 56, 60, 70 e 116.

CHANDRA, P.; GUPTA, M. K. Comprehensive survey on data warehousing research. **International Journal of Information Technology**, Springer, v. 10, n. 2, p. 217–224, 2018. Disponível em: <https://link.springer.com/article/10.1007/s41870-017-0067-y>. Citado na página 37.

CHAUDHARY, S.; MURALA, D. P.; SRIVASTAV, V. A critical review of data warehouse. **Global Journal of Business Management and Information Technology**, v. 1, n. 2, p. 95–103, 2011. Disponível em: <http://www.ripublication.com/gjbmit/gjbmitv1n2_04.pdf>. Citado na página 30.

CHECCHI, D. **The economics of education: Human capital, family background and inequality.** [S.l.]: Cambridge University Press, 2006. Citado na página 89.

CIELEN, D.; MEYSMAN, A.; ALI, M. **Introducing Data Science: Big Data, Machine Learning, and More, Using Python Tools**. 1st. ed. Greenwich, CT, USA: Manning Publications Co., 2016. ISBN 1633430030, 9781633430037. Citado na página 30.

CIFERRI, C. Dutra de A. **Distribuição dos Dados em Ambientes de Data Warehousing: O sistema WebD2W e Algoritmos voltados à fragmentação horizontal dos dados**. Tese (Doutorado), 2002. Disponível em: <https://repositorio.ufpe.br/handle/123456789/1930>. Citado na página 37.

CODEGIRL. **Code girl**. 2022. Accessed: 15.03.2022. Disponível em: <http://www.codegirl.com.br/>. Citado na página 90.

COHEN, J. **Statistical power analysis for the behavioral sciences**. New York, NY, USA: Routledge, 2013. Citado nas páginas 48 e 112.

CONTE, V. d. S. **Mineração de Dados Educacionais para avaliar os fatores que influenciam no desempenho de candidatos do ENEM**. Monografia (Graduação), 2019. Disponível em: <https://app.uff.br/riuff/handle/1/10985>. Citado nas páginas 65, 67 e 71.

COUTINHO, N. C. de A.; BORGES, É. de O. A insuficiência das políticas públicas referentes ao desenvolvimento do ensino público no brasil. **Revista Direitos Sociais e Políticas Públicas (UNIFAFIBE)**, v. 5, n. 2, p. 921–946, 2018. Disponível em: <https://fafibe.br/revista/index.php/direitos-sociais-politicas-pub/article/view/274/pdf>. Citado na página 89.

CRUZ, R. C. **Uma avaliação empírica do Exame Nacional do Ensino Médio–ENEM: impacto da pandemia do Covid-19 no desempenho dos participantes do ENEM 2020**. Dissertação (Mestrado) — Universidade Católica de Brasília, 2022. Citado nas páginas 58, 59, 62 e 70.

CUNHA, N. d. B.; SANTOS, A. Assessment of metatextual awareness and its prediction of reading comprehension. **Psicologia: teoria e prática**, Scielopepsic, v. 21, p. 53 – 68, 04 2019. ISSN 1516-3687. Disponível em: <http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1516-36872019000100003&nrm=iso>. Citado na página 89.

DHAR, V. Data science and prediction. **Commun. ACM**, ACM, New York, NY, USA, v. 56, n. 12, p. 64–73, dez. 2013. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2500499>. Citado na página 35.

DU, D. **Apache Hive Essentials**. [S.l.]: Packt Publishing Ltd, 2015. Citado na página 43.

DUTRA, R. S.; DUTRA, G. B. M.; PARENTE, P. H. N.; PAULO, E. What has changed in the educational performance of federal institutes of brazil? **Ensaio: Avaliação e Políticas Públicas em Educação**, SciELO Brasil, v. 27, p. 631–653, 2019. Disponível em: <https://www.scielo.br/j/ensaio/a/JXwdpkK5Td8xZzcHyxZDHKb/?format=pdf&lang=pt>. Citado nas páginas 58, 62 e 70.

EDWARDS, N. J.; BRAIN, D. T.; JOLY, S. C.; MASUCATO, M. K. Hadoop distributed file system mechanism for processing of large datasets across computers cluster using programming techniques. **International research journal of management, IT and social sciences**, v. 6, n. 6, p. 1–16, 2019. Citado na página 42.

ENEM. **Prova do ENEM – Saiba melhor como é estruturado o exame**. 2013. Disponível em: <https://blogdoenem.com.br/prova-do-enem-exame/>. Citado na página 49.

FABBRI, S.; HERNANDES, E.; THOMMAZO, A. D.; BELGAMO, A.; ZAMBONI, A.; SILVA, C. Managing literature reviews information through visualization. In: **Proceedings of the 14th International Conference on Enterprise Information Systems - Volume 2: ICEIS,**. [S.l.: s.n.], 2012. p. 36–45. ISBN 978-989-8565-11-2. Citado na página 51.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. d. L. F. d. **Inteligência artificial: uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011. Citado nas páginas 46 e 138.

FANG, H. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In: **2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems**. [S.l.: s.n.], 2015. p. 820–824. Citado na página 39.

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®**. [S.l.]: Elsevier Brasil, 2017. Citado nas páginas 48 e 138.

FEIJÓ, J. R.; FRANÇA, J. M. S. d. Diferencial de desempenho entre jovens das escolas públicas e privadas. **Estudos Econômicos (São Paulo)**, SciELO Brasil, v. 51, p. 373–408, 2021. Disponível em: <https://www.scielo.br/j/ee/a/nkypSfcjmwkJj8RbFP9cBkP/?format=pdf&lang=pt>. Citado nas páginas 57, 58, 61 e 70.

FERNANDES, C.; VIEIRA, D.; BARROS, T.; SHAY, A.; FREITAS, N.; VINUTO, T. Eduvizbr: A decision support system for brazilian high school students performance analysis. In: SBC. **Anais do XXXVII Simpósio Brasileiro de Bancos de Dados**. [S.l.], 2022. p. 433–438. Citado nas páginas 64, 66 e 70.

FERREIRA, M. B.; AMORIM, M.; OGASAWARA, E.; BASBASTEFANO, R. A interdisciplinaridade no desempenho da nota de matemática: um olhar para evolução do processo de ensino por meio dos modelos regressivos. In: SBC. **Anais da IV Escola Regional de Informática do Rio de Janeiro**. 2021. p. 41–48. Disponível em: <https://sol.sbc.org.br/index.php/eri-rj/article/view/18773/18603>. Citado nas páginas 58, 59, 62 e 70.

FILHO, J. A. C.; ISOTANI, S.; PENTEADO, B. E. Utilização de notas escolares para predição da nota ENEM em ciências humanas. **Revista Novas Tecnologias na Educação-RENOTE**, 2021. Disponível em: <https://especializacao.icmc.usp.br/documentos/tcc/juvenal_filho.pdf>. Citado nas páginas 63, 66 e 70.

FONSECA, S.; PENEDO, L.; ANTUNES, B.; LIFSCHITZ, S.; CAMPOS, M. L. M.; ALMEIDA, A. C. DW-ENEM: a data warehouse for analytics exploration of national high school exams results. In: SBC. **Anais Estendidos do XXXVII Simpósio Brasileiro de Bancos de Dados**. [S.l.], 2022. p. 34–39. Citado nas páginas 64, 66 e 70.

FRANCO, J. J. **Fatores e evidências sobre o Exame Nacional do Ensino Médio (ENEM): uma abordagem exploratória e experimental com mineração de dados**. Dissertação (Mestrado) — Universidade Federal de Goiás, 2021. Disponível em: <http://repositorio.bc.ufg.br/tede/handle/tede/11248>. Citado nas páginas 65, 68 e 71.

FRENEDA, F. C. B. **Múltiplos fatores do desempenho escolar: uma análise dos microdados do INEP sobre a educação no Distrito Federal**. Dissertação (Mestrado) — Universidade Católica de Brasília, 2020. Disponível em: <https://bdtd.ucb.br:8443/jspui/handle/tede/2763>. Citado nas páginas 58, 61 e 70.

FRITZ, C. O.; MORRIS, P. E.; RICHLER, J. J. Effect size estimates: current use, calculations, and interpretation. **Journal of experimental psychology: General**, American Psychological Association, v. 141, n. 1, p. 2, 2012. Disponível em: <https://psycnet.apa.org/fulltext/2011-16756-001.pdf>. Citado na página 112.

GARCÍA, A.; DÍAZ, A. C.; GARCÍA, F. J. Engaging women into STEM in latin america: W-stem project. In: **Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality**. [s.n.], 2019. p. 232–239. Disponível em: <https://dl.acm.org/doi/pdf/10.1145/3362789.3362902>. Citado na página 89.

GARCÍA, M. V.; AZNARTE, J. L. Shapley additive explanations for no2 forecasting. **Ecological Informatics**, Elsevier, v. 56, p. 101039, 2020. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S1574954119303498>. Citado na página 48.

GARCIA, R. A.; NETO, E. L. G. R.; RIBEIRO, A. d. M. Efeitos rendimento escolar, infraestrutura e prática docente na qualidade do ensino médio no brasil. **Revista Brasileira de Estudos de População**, SciELO Brasil, v. 38, 2021. Disponível em: <https://www.scielo.br/j/rbepop/a/9fjNLP3gPFHzBqpFC75m7Qk/?format=pdf&lang=pt>. Citado nas páginas 65, 67 e 71.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2019. ISBN 9781492032649. Citado na página 47.

GIL, N. d. L. A quantificação da qualidade: algumas considerações sobre os índices de reprovação escolar no brasil. **Sociologias**, SciELO Brasil, v. 23, p. 184–209, 2021. Disponível em: <https://www.scielo.br/j/soc/a/Gs9ZVNbCBj9TczbwmcVpTyB/>. Citado na página 88.

GOLFARELLI, M. Open source bi platforms: A functional and architectural comparison. In: . Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 287–297. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-642-03730-6_23>. Citado na página 37.

GOLFARELLI, M.; RIZZI, S. From star schemas to big data: 20+ years of data warehouse research. **A comprehensive guide through the Italian database research over the last 25 years**, Springer, p. 93–107, 2018. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-61893-7_6>. Citado na página 37.

GOMEDE, E.; GAFFO, F.; BRIGANÓ, G.; BARROS, R. de; MENDES, L. Application of computational intelligence to improve education in smart cities. **Sensors**, Multidisciplinary Digital Publishing Institute, v. 18, n. 1, p. 267, 2018. Disponível em: <https://www.mdpi.com/1424-8220/18/1/267>. Citado na página 30.

GOMES, C.; VIANA, A. B. N. Explorando os efeitos da disponibilidade das tecnologias da informação e comunicação nos resultados do ENEM. **Revista Brasileira de Estudos Pedagógicos**, SciELO Brasil, v. 103, p. 37–60, 2022. Disponível em: <https://www.scielo.br/j/rbeped/a/BHq8pSXH5VXDCxx98Gygy5x/abstract/?lang=pt>. Citado nas páginas 58, 61 e 70.

GOMES, C. M. A.; FLEITH, D. d. S.; MARINHO-ARAUJO, C. M.; RABELO, M. L. Preditores do desempenho em matemática de estudantes do ensino médio. **Psicologia: Teoria e Pesquisa**, SciELO Brasil, v. 36, 2021. Disponível em: <https://www.scielo.br/j/ptp/a/nmFpbxGtkNVM9x96ZSdLLnr/?format=pdf&lang=pt>. Citado nas páginas 63, 66 e 70.

GOMES, G. L. S.; FERNANDES, I. M. *et al.* A influência da infraestrutura escolar e formação docente no desempenho dos estudantes na área de ciências da natureza. SciELO Preprints, 2021. Disponível em: <https://preprints.scielo.org/index.php/scielo/preprint/view/3147/5668>. Citado nas páginas 55, 57, 60 e 70.

GONÇALVES, F. G. d. **Sucesso no campo escolar de estudantes oriundos de classes populares: estrutura e trajetórias.** Dissertação (Mestrado) — Universidade Federal do Rio Grande do Sul, 2015. Citado na página 88.

GROSSMAN, M.; SARKAR, V. Swat: A programmable, in-memory, distributed, high-performance computing platform. In: **Proceedings of the 25th ACM International Symposium on High-Performance Parallel and Distributed Computing**. New York, NY, USA: ACM, 2016. (HPDC '16), p. 81–92. ISBN 978-1-4503-4314-5. Disponível em: <http://doi.acm.org/10.1145/2907294.2907307>. Citado na página 44.

GUARDIEIRO, V.; RAIMUNDO, M. M.; POCO, J. Analyzing the equity of the brazilian national high school exam by validating the item response theory's invariance. In: **15th International Conference on Educational Data Mining**. Durham, United Kingdom: [s.n.], 2022. Disponível em: <https://educationaldatamining.org/edm2022/proceedings/2022.EDM-posters.64/2022.EDM-posters.64.pdf>. Citado nas páginas 58, 61 e 70.

GUSTAFSSON, J.-E.; NILSEN, T.; HANSEN, K. Y. School characteristics moderating the relation between student socio-economic status and mathematics achievement in grade 8. evidence from 50 countries in timss 2011. **Studies in Educational Evaluation**, Elsevier, v. 57, p. 16–30, 2018. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0191491X15300936>. Citado na página 88.

GUZMAN, I.; BERARDI, R.; MACIEL, C.; TAPIA, P. C.; MARIN-RAVENTOS, G.; RODRIGUEZ, N.; RODRIGUEZ, M. Gender gap in it in latin america. 2020. Citado na página 90.

HAN, J.; KAMBER, M.; PEI, J. Data mining: Concepts and techniques third edition [m]. **The Morgan Kaufmann Series in Data Management Systems**, v. 5, n. 4, p. 83–124, 2011. Disponível em: <https://doi.org/10.1016/C2009-0-61819-5>. Citado nas páginas 36 e 45.

HENN, S. When women stopped coding. **NPR Planet Money**, v. 21, 2014. Citado na página 89.

HOLANDA, M.; JÚNIOR, A. L.; SILVA, E. H. M. da; LATERZA, J.; ARAUJO, A.; CASTANHO, C.; KOIKE, C.; OLIVEIRA, R. B. Uma análise comparativa do desempenho em matemática entre gêneros nas provas do ENEM. In: SBC. **Anais do XVI Women in Information Technology**. [S.l.], 2022. p. 145–156. Citado nas páginas 57, 60 e 70.

HONG, S.; CHOI, W.; JEONG, W.-K. Gpu in-memory processing using spark for iterative computation. In: **2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)**. [s.n.], 2017. p. 31–41. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7973686>. Citado nas páginas 44 e 45.

HUANG, C.-S.; TSAI, M.-F.; HUANG, P.-H.; SU, L.-D.; LEE, K.-S. Distributed asteroid discovery system for large astronomical data. **Journal of Network and Computer Applications**, v. 93, p. 27 – 37, 2017. ISSN 1084-8045. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1084804517301157>. Citado na página 44.

HUANG, J.; GATES, A. J.; SINATRA, R.; BARABáSI, A.-L. Historical comparison of gender inequality in scientific careers across countries and disciplines. **Proceedings of the National Academy of Sciences**, v. 117, n. 9, p. 4609–4616, 2020. Disponível em: <https://www.pnas.org/doi/abs/10.1073/pnas.1914221117>. Citado na página 89.

IBGE. **Exame Nacional do Ensino Médio**. 2021. Disponível em: <https://ces.ibge.gov.br/base-de-dados/metadados/inep/exame-nacional-do-ensino-medio-enem.html>. Citado na página 27.

____. **Síntese de Indicadores Sociais**. 2022. Accessed: 15.03.2022. Disponível em: <https://www.ibge.gov.br/estatisticas/multidominio/condicoes-de-vida-desigualdade-e-pobreza/9221-sintese-de-indicadores-sociais.html?=&t=resultados>. Citado na página 89.

IDEB. **Resultados do índice de desenvolvimento da educação básica-2019**. 2022. Accessed: 14.03.2022. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/estatisticas_e_indicadores/resultados_indice_desenvolvimento_educacao_basica_2019_resumo_tecnico.pdf>. Citado na página 88.

INEP. **Microdados ENEM**. 2019. Disponível em: <https://portal.inep.gov.br/web/guest/microdados>. Citado na página 28.

____. **Exame Nacional do Ensino Médio - Histórico**. 2022. Accessed: 12.02.2022. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico>. Citado na página 27.

____. **Relatorio brasil no PISA 2018**. 2022. Accessed: 14.03.2022. Disponível em: <https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/relatorio_brasil_no_pisa_2018.pdf>. Citado na página 88.

INMON, W. H. **Building the Data Warehouse**. New York, NY, USA: John Wiley & Sons, Inc., 1992. ISBN 0471569607. Citado na página 38.

INOUK, B.; SUZANNE, M.; JELLE, J. The role of home literacy environment, mentalizing, expressive verbal ability, and print exposure in third and fourth graders' reading comprehension. **Scientific Studies of Reading**, Routledge, v. 21, n. 3, p. 179–193, 2017. Disponível em: <https://doi.org/10.1080/10888438.2016.1277727>. Citado na página 88.

JARDIM, R.; DELGADO, C.; SCHNEIDER, D. Data science supporting a question classifier model. **Procedia Computer Science**, Elsevier, v. 199, p. 1237–1243, 2022. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050922001582>. Citado nas páginas 64, 67 e 70.

JUN, L.; BIN, S. The role of vocabulary in reading comprehension: The case of secondary school students learning english in singapore. **RELC Journal**, Sage Publications Sage UK: London, England, v. 39, n. 1, p. 51–76, 2008. Disponível em: <https://journals.sagepub.com/doi/pdf/10.1177/0033688208091140>. Citado na página 90.

JURIATI, D. E.; ARIYANTI, A.; FITRIANA, R. The correlation between reading comprehension and writing ability in descriptive text. **Southeast Asian Journal of Islamic Education**, v. 1, n. 1, p. 01–14, 2018. Disponível em: <https://journal.uinsi.ac.id/index.php/SAJIE/article/download/1150/pdf/>. Citado na página 88.

KHINE, P. P.; WANG, Z. S. Data lake: a new ideology in big data era. In: EDP SCIENCES. **ITM web of conferences**. [S.l.], 2018. v. 17, p. 03025. Citado nas páginas 40 e 41.

KIMBALL, R.; ROSS, M. **The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling**. 2nd. ed. New York, NY, USA: John Wiley & Sons, Inc., 2002. ISBN 0471200247, 9780471200246. Citado na página 38.

KITCHENHAM, B.; CHARTERS, S. Guidelines for performing systematic literature reviews in software engineering. Citeseer, 2007. Citado na página 51.

KLEIN, R. Utilização da teoria de resposta ao item no sistema nacional de avaliação da educação básica (SAEB). **Revista Meta: Avaliação**, v. 1, n. 2, p. 125–140, 2009. Disponível em: <https://revistas.cesgranrio.org.br/index.php/metaavaliacao/article/view/38>. Citado na página 49.

KOZA, Ş.; MELIS, F. The effect of socioeconomic status on students' achievement. In: **The factors effecting student achievement**. Springer, 2017. p. 171–181. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-319-56083-0_10>. Citado na página 88.

KUGLER, A. D.; TINSLEY, C. H.; UKHANEVA, O. Choice of majors: Are women really different from men?. National Bureau of Economic Research, 2017. Disponível em: <https://www.nber.org/system/files/working_papers/w23735/w23735.pdf>. Citado na página 90.

KUMAR, V. N.; SHINDGIKAR, P. **Modern Big Data processing with Hadoop: Expert techniques for architecting end-to-end Big Data solutions to get valuable insights**. [S.l.]: Packt Publishing Ltd, 2018. Citado na página 43.

LANEY, D. *et al.* 3d data management: Controlling data volume, velocity and variety. **META group research note**, Stanford, v. 6, n. 70, p. 1, 2001. Citado na página 41.

LAURENT, A.; LAURENT, D.; MADERA, C. Introduction to data lakes: Definitions and discussions. **Data Lakes**, Wiley Online Library, v. 2, p. 1–20, 2020. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119720430.ch1>. Citado na página 40.

LEE, J.-W.; LEE, H. Human capital and income inequality. **Journal of the Asia Pacific Economy**, Routledge, v. 23, n. 4, p. 554–583, 2018. Disponível em: <https://www.econstor.eu/bitstream/10419/190231/1/adbi-wp810.pdf>. Citado na página 89.

LERIA, L. A.; BENITEZ, P.; FERREIRA, L. A.; FRAGA, F. J. *et al.* O acesso do estudante com deficiência visual à educação superior: análise dos microdados do exame nacional do ensino médio (ENEM). SciELO Preprints, 2021. Disponível em: <https://preprints.scielo.org/index.php/scielo/preprint/view/1969/5495>. Citado nas páginas 58, 59, 62 e 70.

LIMA, A.; FLOREZ, A.; LESCANO, A.; NOVAES, J.; MARTINS, N.; JUNIOR, C. T.; SOUSA, E.; JÚNIOR, J. R.; CORDEIRO, R. Analysis of enem's attendants between 2012 and 2017 using a clustering approach. **Journal of Information and Data Management**, v. 11, n. 2, 2020. Disponível em: <https://periodicos.ufmg.br/index.php/jidm/article/view/24835/23240>. Citado nas páginas 65, 67 e 71.

LIMA, P. J.; FRAGA, J. C. J. Qual é o efeito da desigualdade social no desempenho em ciências dos estudantes brasileiros? uma análise do exame nacional do ensino médio (2012-2019). **Investigações em Ensino de Ciência**, Universidade Federal do Rio Grande do Sul, Instituto de Física, v. 26, n. 1, p. 110–126, 2021. Disponível em: <https://www.proquest.com/docview/2524417709?pq-origsite=gscholar&fromopenview=true>. Citado nas páginas 55, 57, 60 e 70.

LUCKESI, C. C. **Avaliação da aprendizagem escolar: estudos e proposições**. [S.l.]: Cortez editora, 2014. Citado na página 27.

LUNDBERG, S. M.; ERION, G.; CHEN, H.; DEGRAVE, A.; PRUTKIN, J. M.; NAIR, B.; KATZ, R.; HIMMELFARB, J.; BANSAL, N.; LEE, S.-I. From local explanations to global understanding with explainable ai for trees. **Nature machine intelligence**, Nature Publishing Group, v. 2, n. 1, p. 56–67, 2020. Disponível em: <https://www.nature.com/articles/s42256-019-0138-9?ref=https://githubhelp.com>. Citado na página 85.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: . Red Hook, NY, USA: Curran Associates Inc., 2017. (NIPS'17), p. 4768–4777. ISBN 9781510860964. Disponível em: <https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf>. Citado na página 47.

MACIEL, C.; BIM, S. A.; FIGUEIREDO, K. da S. Digital girls program: disseminating computer science to girls in brazil. In: **Proceedings of the 1st International Workshop on Gender Equality in Software Engineering**. [S.l.: s.n.], 2018. p. 29–32. Citado na página 90.

MAILIS, M.; DELFI, S.; ERNI, E. **The correlation between reading comprehension and writing ability of the second year students of Sman 1 Muaro Sentajo Teluk Kuantan in recount texts.** Tese (Doutorado) — Riau University, 2018. Disponível em: <https://www.neliti.com/publications/204714/the-correlation-between-reading-comprehension-and-writing-ability-of-the-second>. Citado na página 88.

MARKOSKI, A.; ZANCANARO, L.; GUERRA, P. A. C.; BERTOLINI, C.; SILVEIRA, S. R. Descoberta de indicadores e padrões nos participantes do ENEM. **Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica**, v. 10, n. 1, 2019. Disponível em: <https://periodicos.unifacef.com.br/index.php/resiget/article/view/1668>. Citado nas páginas 65, 67 e 71.

MARQUES, R.; GAMA, J. C. F.; JUNIOR, G. L. O.; NETO, A. F.; SANTOS, W. d. A educação física no ensino médio e os exames estandardizados: uma análise das questões do ENEM. **Movimento**, SciELO Brasil, v. 27, 2022. Disponível em: <https://www.scielo.br/j/mov/a/QmxdKqZypgKXz8hBntFgWPP/abstract/?lang=pt>. Citado nas páginas 55, 57, 60 e 70.

MASSI, F. A matriz de correção da redação do ENEM. **Caminhos em Linguística Aplicada**, v. 16, n. 1, p. 69–89, 2017. Disponível em: <http://periodicos.unitau.br/ojs/index.php/caminhoslinguistica/article/viewFile/2253/1658>. Citado na página 88.

MEC. **Plano Nacional de Educação**. 2022. Accessed: 13.03.2022. Disponível em: <https://pne.mec.gov.br/>. Citado na página 89.

____. **Educação em Prática**. 2022. Accessed: 15.03.2022. Disponível em: <http://portal.mec.gov.br/component/content/article/30000-uncategorised/82221-educacao-em-pratica>. Citado na página 89.

____. **Ensino Médio Inovador**. 2022. Accessed: 15.03.2022. Disponível em: <http://portal.mec.gov.br/component/content/article?id=13439:ensino-medio-inovador>. Citado na página 89.

____. **Educação é o caminho para promover a inclusão e combater a discriminação racial**. 2022. Accessed: 15.03.2022. Disponível em: <http://portal.mec.gov.br/component/tags/tag/cotas-raciais>. Citado na página 89.

MENDES, B. D.; KARRUZ, A. P. Background familiar, desigualdade regional e o desempenho no exame nacional do ensino médio (ENEM). In: **I Encontro Nacional de Ensino e Pesquisa do Campo de Públicas**. [S.l.: s.n.], 2015. Citado na página 88.

MOHAMED, A.; NAJAFABADI, M. K.; WAH, Y. B.; ZAMAN, E. A. K.; MASKAT, R. The state of the art and taxonomy of big data analytics: view from new big data framework. **Artificial Intelligence Review**, Feb 2019. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-019-09685-9>. Citado na página 44.

MOORE, H. Compreendendo sexo e gênero. **Companion Encyclopedia of Anthropology. London: Routledge**, 1997. Citado na página 32.

MORAES, C. P. de; PERES, R. T. Reflexões sobre diferenças de desempenho no ENEM: uma análise socioeconômica e escolar do sudeste do brasil. **Jornal de Políticas Educacionais**, v. 16, n. 1, 2022. Disponível em: <https://revistas.ufpr.br/jpe/article/view/85377>. Citado nas páginas 57, 58, 61 e 70.

MORAES, C. P. de; PERES, R. T.; BARBOSA, M. T. S.; PEDREIRA, C. E. Equidade e desempenho no exame nacional do ensino médio: Um estudo sobre sexo e raça nos municípios brasileiros. **Education Policy Analysis Archives**, v. 30, p. 68–68, 2022. Disponível em: <https://epaa.asu.edu/index.php/epaa/article/view/6971>. Citado nas páginas 57, 58, 61 e 70.

MORAES, C. P. de; PERES, R. T.; BARBOSA, T. S.; PEDREIRA, C. E. Efeito escola a partir de indicadores educacionais: análise entre escolas públicas e privadas no ENEM. **Revista Meta: Avaliação**, v. 14, n. 42, p. 67–93, 2022. Citado nas páginas 57, 58, 61 e 70.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística básica**. [S.l.]: Saraiva Educação SA, 2017. Citado na página 48.

MOTOKANE, M. T. Impacto das variáveis socioeconômicas no desempenho do ENEM: uma análise espacial e sociológica. **Revista de Administração Pública**, v. 55, n. 6, p. 1271–1294, 2021. Disponível em: <https://bibliotecadigital.fgv.br/ojs/index.php/rap/article/download/85021/80363/186338>. Citado nas páginas 65, 67 e 71.

MUELLER, J.; MASSARON, L. **Python for Data Science For Dummies**. Wiley, 2019. ISBN 9781119547624. Disponível em: <https://books.google.com.br/books?id=Uh2EDwAAQBAJ>. Citado na página 36.

NASCIMENTO, M. M. **O acesso ao ensino superior público brasileiro: um estudo quantitativo a partir dos microdados do Exame Nacional do Ensino Médio. 2019. 192 f**. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, 2019. Citado nas páginas 55, 56, 60 e 70.

NASCIMENTO, M. M.; CAVALCANTI, C.; OSTERMANN, F. Uma busca por questões de física do enem potencialmente não reprodutoras das desigualdades socioeconômicas. **Revista Brasileira de Ensino de Física**, SciELO Brasil, v. 40, 2018. Disponível em: <https://www.scielo.br/j/rbef/a/8jPnnXc48zXNmsLHB4JgPWN/?format=pdf&lang=pt>. Citado nas páginas 55, 56, 60 e 70.

____. Sucesso escolar em contextos populares: uma análise a partir do enem. **Estudos em Avaliação Educacional**, v. 31, n. 76, p. 134–163, 2020. Disponível em: <http://publicacoes.fcc.org.br/index.php/eae/article/view/6719/3966>. Citado nas páginas 58, 59, 62 e 70.

NAVADA, A.; ANSARI, A. N.; PATIL, S.; SONKAMBLE, B. A. Overview of use of decision tree algorithms in machine learning. In: IEEE. **2011 IEEE control and system graduate research colloquium**. 2011. p. 37–42. Disponível em: <https://ieeexplore.ieee.org/abstract/document/5991826>. Citado na página 46.

NETO, N. W.; SOARES, R. C.; COUTINHO, L. R.; TELES, A. S. A pandemia da covid-19 impactou o ENEM? uma análise comparativa de dados dos anos de 2019 e 2020. **RENOTE**, v. 20, n. 1, p. 223–232, 2022. Disponível em: <https://www.seer.ufrgs.br/renote/article/view/126655>. Citado nas páginas 64, 66 e 70.

NGUYEN, T.; LARSEN, M. E.; O'DEA, B.; NGUYEN, D. T.; YEARWOOD, J.; PHUNG, D.; VENKATESH, S.; CHRISTENSEN, H. Kernel-based features for predicting population health indices from geocoded social media data. **Decision Support Systems**, v. 102, p. 22 – 31, 2017. ISSN 0167-9236. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167923617301227>. Citado na página 44.

NOGUERA, V.; BRANCO, K.; CIFERRI, C. Gêneros e suas nuances no ENEM. In: **Anais do XIII Women in Information Technology**. Belém, PA, Brasil: SBC, 2019. p. 41–50. Disponível em: <https://sol.sbc.org.br/index.php/wit/article/view/6711>. Citado nas páginas 93, 96, 97, 98 e 99.

NOGUERA, V. E. R.; BRANCO, K. R. L. J. C.; AGUIAR, C. D. d. Identifying and analyzing key performance indicators from the enem data to support educational public policies. ***Artigo em julgamento* no journal Educational Assessment**, 2022. Citado na página 81.

NOGUERA, V. E. R.; BRANCO, K. R. L. J. C.; CIFERRI, C. D. de A. Análise de desempenho das mulheres no ENEM. **Brazilian Journal of Development**, v. 6, n. 6, p. 35716–35737, 2020. Disponível em: <https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/11377>. Citado nas páginas 93, 98, 100, 101 e 102.

NOVAES, A. A. d. **Uma proposta de análise de desempenho de estudantes do ENEM**. Monografia (Graduação) — Universidade Federal de São Carlos, 2021. Disponível em: <https://repositorio.ufscar.br/handle/ufscar/13840>. Citado nas páginas 64, 66 e 70.

OECD. **PISA 2018 Results (Volume II).** [s.n.], 2019. 376 p. Disponível em: <https://www.oecd-ilibrary.org/content/publication/b5fd1b8f-en>. Citado na página 89.

OLIVEIRA, C. G. de; BARWALDT, R.; LUCCA, G. Análise do desempenho de pessoas com deficiência que prestaram o exame nacional do ensino médio-ENEM. **# Tear: Revista de Educação, Ciência e Tecnologia**, v. 9, n. 1, 2020. Disponível em: <https://dev7b.ifrs.edu.br/site_periodicos/periodicos/index.php/tear/article/view/4038>. Citado nas páginas 65, 68 e 71.

PASQUALI, L. **Psicometria: teoria dos testes na psicologia e na educação**. [S.l.]: Editora Vozes Limitada, 2017. Citado na página 113.

PAULA, M. d. F. C. d. Políticas de democratização da educação superior brasileira: limites e desafios para a próxima década. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, SciELO, v. 22, p. 301 – 315, 08 2017. ISSN 1414-4077. Disponível em: <https://www.scielo.br/j/aval/a/KYs6H9L5YpppTCZHpHGd8SK/abstract/?lang=pt>. Citado na página 27.

PENNINGTON, C. R.; KAYE, L. K.; QURESHI, A. W.; HEIM, D. Do gender differences in academic attainment correspond with scholastic attitudes? an exploratory study in a uk secondary school. **Journal of Applied Social Psychology**, Wiley Online Library, v. 51, n. 1, p. 3–16, 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jasp.12711>. Citado na página 90.

PILOTTI, M. A. What lies beneath sustainable education? predicting and tackling gender differences in stem academic success. **Sustainability**, Multidisciplinary Digital Publishing Institute, v. 13, n. 4, p. 1671, 2021. Disponível em: <https://www.mdpi.com/2071-1050/13/4/1671/pdf>. Citado na página 90.

PISA. **Programme for International Student Assessment.** 2022. Accessed: 14.02.2022. Disponível em: <https://www.oecd.org/pisa/>. Citado na página 29.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What You Need to Know About Data Mining and Data-analytic Thinking**. 1st. ed. [S.l.]: O'Reilly Media, Inc., 2013. ISBN 1449361323, 9781449361327. Citado nas páginas 30 e 35.

PYLADIES. **PyLadies Brasil**. 2014. Accessed: 15.03.2022. Disponível em: <https://brasil.pyladies.com/>. Citado na página 90.

QUINLAN, J. R. Induction of decision trees. **Mach. Learn.**, Kluwer Academic Publishers, USA, v. 1, n. 1, p. 81–106, mar. 1986. ISSN 0885-6125. Disponível em: <https://doi.org/10.1023/A:1022643204877>. Citado na página 46.

RAVAT, F.; ZHAO, Y. Data lakes: Trends and perspectives. In: SPRINGER. **International Conference on Database and Expert Systems Applications**. 2019. p. 304–313. Disponível em: <https://hal.archives-ouvertes.fr/hal-02397457/document>. Citado nas páginas 39 e 40.

ROBITZSCH, A. sirt: Supplementary item response theory models. **R package version**, v. 1, n. 0, 2015. Citado na página 107.

ROCHA, F. B. N. da; COSTA, J. E.; MAIA, J. G. R.; FILHO, J. A. de C. Análise dos microdados de matemática do ENEM de 2017–2019 do nordeste. **Research, Society and Development**, v. 11, n. 10, p. e207111032716–e207111032716, 2022. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/32716>. Citado nas páginas 57, 60 e 70.

RODRIGUES, D. de C.; LIMA, M. Dias de; CONCEIÇÃO, M. D. da; SIQUEIRA, V. S. de; BARBOSA, R. M. A data mining approach applied to the high school national examination: Analysis of aspects of candidates to brazilian universities. In: **EPIA Conference on Artificial Intelligence**. Cham: Springer International Publishing, 2019. p. 3–14. ISBN 978-3-030-30241-2. Disponível em: <https://link.springer.com/chapter/10.1007/978-3-030-30241-2_1>. Citado nas páginas 65, 67 e 71.

ROIGER, R. J. **Data mining: a tutorial-based primer**. [S.l.]: Chapman and Hall/CRC, 2017. Citado na página 45.

SALTZ, J. S.; STANTON, J. M. **An Introduction to Data Science**. 1st. ed. Thousand Oaks, CA, USA: Sage Publications, Inc., 2017. ISBN 150637753X, 9781506377537. Citado na página 35.

SANTANA, W. M. d. **Mineração em dados do ENEM para a predição do desempenho acadêmico no âmbito da Rede Federal de Educação Tecnológica**. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2018. Disponível em: <https://repositorio.ufpe.br/handle/123456789/29994>. Citado nas páginas 65, 67 e 71.

SANTO, F. d. E. **Teoria da resposta ao item: influência do tamanho da amostra na estimação dos parâmetros dos itens utilizando os microdados do ENEM**. Dissertação (Mestrado) — Universidade de São Paulo, 2020. Citado na página 107.

SANTOS, A. M. T. B. d. *et al.* **Mineração de dados educacionais: um estudo sobre os dados socioeconômicos na educação na base de dados do INEP**. Dissertação (Mestrado) — Universidade Federal do Pará, 2019. Disponível em: <http://repositorio.ufpa.br/handle/2011/11265>. Citado nas páginas 58, 61 e 70.

SANTOS, B.; OLIVEIRA, C. G.; TOPIN, L. O. H.; MENDIZABAL, O. M.; BARWALDT, R. Analysis of candidates profile for the national entrance exams for admission to brazilian universities. In: IEEE. **2019 IEEE Frontiers in Education Conference (FIE)**. 2019. p. 1–8. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9028381>. Citado nas páginas 64, 66 e 70.

SANTOS, B. S. dos; OLIVEIRA, C. G. de; MENDIZABAL, O. M.; BARWALDT, R. Extração de métricas sobre candidatos do ENEM usando apache spark. In: SBC. **Anais da XIX Escola Regional de Alto Desempenho da Região Sul**. 2019. Disponível em: <https://sol.sbc.org.br/index.php/eradrs/article/view/7041/6930>. Citado nas páginas 58, 59, 62 e 70.

SANTOS, C. M. Por que as mulheres "desapareceram" dos cursos de computação? **Jornal da Universidade de São Paulo**, http://jornal.usp.br/universidade/por-que-as-mulheres-desapareceram-dos-cursos-de-computacao/, 2018. Disponível em: <https://jornal.usp.br/?p=136701>. Citado na página 90.

SEN, P. C.; HAJRA, M.; GHOSH, M. Supervised classification algorithms in machine learning: A survey and review. In: MANDAL, J. K.; BHATTACHARYA, D. (Ed.). **Emerging Technology in Modelling and Graphics**. Singapore: Springer Singapore, 2020. p. 99–111. ISBN 978-981-13-7403-6. Disponível em: <https://link.springer.com/chapter/10.1007/978-981-13-7403-6_11>. Citado na página 45.

SHAIK, J.; JANAHAN, S. K.; ARUN, S.; R, A. Business intelligence and decision support using distinct mapreduce with access patterns (dmrap) in big data analytics. **Journal of Advanced Research in Dynamical and Control Systems**, v. 2017, p. 107–112, 06 2017. Citado na página 44.

SHANG, S.; GAN, Y.; WU, H. An improved distributed file system based on gpu acceleration. In: **2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)**. [S.l.: s.n.], 2018. p. 736–739. Citado na página 42.

SHE++. **She++**. 2015. Accessed: 15.03.2022. Disponível em: <http://www.sheplusplus.org>. Citado na página 90.

SHVACHKO, K.; KUANG, H.; RADIA, S.; CHANSLER, R. The hadoop distributed file system. In: **Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)**. Washington, DC, USA: IEEE Computer Society, 2010. (MSST '10), p. 1–10. ISBN 978-1-4244-7152-2. Disponível em: <http://dx.doi.org/10.1109/MSST.2010.5496972>. Citado na página 42.

SILVA, L.; MORINO, A. H.; SATO, T. M. C. Prática de mineração de dados no exame nacional do ensino médio. **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**, v. 3, n. 1, p. 651, 2015. ISSN 2316-8889. Disponível em: <https://www.br-ie.org/pub/index.php/wcbie/article/view/3289>. Citado na página 27.

SILVA, R. L. C.; BRITO, K. d. S.; ADEODATO, P. J. L. A data mining framework for reporting trends in the predictive contribution of factors related to educational achievement. **Available at SSRN 4215682**, 2022. Citado nas páginas 64, 67 e 70.

SILVA, V. A. A. da; MORENO, L. L. O.; GONÇALVES, L. B.; SOARES, S. S. R. F.; JÚNIOR, R. R. S. Identificação de desigualdades sociais a partir do desempenho dos alunos do ensino médio no ENEM 2019 utilizando mineração de dados. In: SBC. **Anais do XXXI Simpósio Brasileiro de Informática na Educação**. 2020. p. 72–81. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/12763/12617>. Citado nas páginas 64, 67 e 71.

SILVA, W.; OLIVEIRA, E.; CURI, M.; BOURGUET, J. Writing proficiency assessment: Regression analysis of item response theory supported by machine learning techniques. In: IEEE. **2021 XLVII Latin American Computing Conference (CLEI)**. 2021. p. 1–10. Disponível em: <https://ieeexplore.ieee.org/abstract/document/9639903>. Citado nas páginas 63, 66 e 70.

SILVEIRA, I. C.; MAUÁ, D. D. Advances in automatically solving the ENEM. In: **2018 7th Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.: s.n.], 2018. Citado na página 27.

SILVERS, F. **Building and Maintaining a Data Warehouse**. New York, NY, USA: Auerbach Taylor& Francis Group., 2008. ISBN 1420064622. Citado na página 38.

SIQUEIRA, D. P.; LARA, F. C. P.; LIMA, H. F. C. Direitos da personalidade e as políticas públicas de educação: programa educação em prática-a integração entre o ensino fundamental e médio com as universidades. **Revista Húmus**, v. 10, n. 28, 2020. Disponível em: <http://periodicoseletronicos.ufma.br/index.php/revistahumus/article/view/13541/7823>. Citado na página 89.

SKIENA, S. S. **The Data Science Design Manual**. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2017. ISBN 3319554433, 9783319554433. Citado na página 36.

SOARES, J. F.; ALVES, M. T. G. Effects of schools and municipalities in the quality of basic education. **Cadernos de Pesquisa**, SciELO Brasil, v. 43, p. 492–517, 2013. Disponível em: <https://www.scielo.br/j/cp/a/WGhPXprTVJRhtZKc5VVrsdN/?format=pdf&lang=en>. Citado na página 29.

SOARES, T. E. A.; SOARES, D. J. M.; SANTOS, W. D. Medidas de tendência central: Análise da qualidade das questões do ENEM de 2016 a 2018. **Jornal Internacional de Estudos em Educação Matemática**, v. 14, n. 1, p. 119–128, 2021. Disponível em: <http://funes.uniandes.edu.co/30671/>. Citado na página 116.

SOGODEKAR, M.; PANDEY, S.; TUPKARI, I.; MANEKAR, A. Big data analytics: hadoop and tools. In: **2016 IEEE Bombay Section Symposium (IBSS)**. [S.l.: s.n.], 2016. p. 1–6. Citado na página 44.

SONG, I.-Y.; ZHU, Y. Big data and data science: what should we teach? **Expert Systems**, v. 33, n. 4, p. 364–373, 2016. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.12130>. Citado nas páginas 30, 35 e 36.

SOUZA, J. G. d. D. Dificuldades encontradas por educandos para construção da redação no ENEM. **RACE-Revista de Administração do Cesmac**, v. 6, 2020. Disponível em: <https://revistas.cesmac.edu.br/index.php/administracao/article/download/1340/1038>. Citado nas páginas 88 e 90.

SOUZA, K. R. G. **Um estudo sobre o desempenho das escolas públicas do DF sob o ponto de vista do ENEM**. Dissertação (Mestrado) — Universidade Católica de Brasília, 2019. Disponível em: <https://bdtd.ucb.br:8443/jspui/handle/tede/2750>. Citado nas páginas 64, 67 e 70.

SOUZA, T. O. d. **Análise de dados: um estudo do perfil dos participantes do ENEM 2019.** Monografia (Graduação), 2021. Disponível em: <http://repositorio.ufersa.edu.br/handle/prefix/6916>. Citado nas páginas 64, 66 e 70.

STAUDT, M.; VADUVA, A.; VETTERLI, T. **Metadata Management and Data Warehousing**. [S.l.], 1999. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.40.5736&rep=rep1&type=pdf>. Citado na página 38.

STEEGH, A. M.; HÖFFLER, T. N.; KELLER, M. M.; PARCHMANN, I. Gender differences in mathematics and science competitions: A systematic review. **Journal of Research in Science Teaching**, Wiley Online Library, v. 56, n. 10, p. 1431–1460, 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/tea.21580>. Citado na página 89.

TAM, H.-l.; CHAN, A. Y.-f.; LAI, O. L.-h. Gender stereotyping and STEM education: Girls' empowerment through effective ICT training in hong kong. **Children and Youth Services Review**, Elsevier, v. 119, p. 105624, 2020. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0190740920320478>. Citado na página 90.

TAN, P.; STEINBACH, M.; KUMAR, V.; FERNANDES, A. **Introdução ao datamining: mineração de dados**. Ciencia Moderna, 2009. ISBN 9788573937619. Disponível em: <https://books.google.com.br/books?id=69d6PgAACAAJ>. Citado nas páginas 46 e 47.

THOMSON, S. **Achievement at school and socioeconomic background—an educational perspective.** Nature Publishing Group, 2018. 1–2 p. Disponível em: <https://doi.org/10.1038/s41539-018-0022-0>. Citado na página 88.

THUSOO, A.; SARMA, J. S.; JAIN, N.; SHAO, Z.; CHAKKA, P.; ANTHONY, S.; LIU, H.; WYCKOFF, P.; MURTHY, R. Hive: A warehousing solution over a map-reduce framework. VLDB Endowment, v. 2, n. 2, p. 1626–1629, aug 2009. ISSN 2150-8097. Disponível em: <https://doi.org/10.14778/1687553.1687609>. Citado na página 76.

TIWARY, M.; SAHOO, A. K.; MISRA, R. Efficient implementation of apriori algorithm on hdfs using gpu. In: IEEE. **2014 International Conference on High Performance Computing and Applications (ICHPCA)**. [S.l.], 2014. p. 1–7. Citado na página 42.

TORRES, A. Technovation challenge: Introducing innovation and mobile app development to girls around the world. In: **Mobile Media Learning: Innovation and Inspiration**. [s.n.], 2015. p. 171–195. Disponível em: <https://dl.acm.org/doi/abs/10.5555/2811074.2811088>. Citado na página 90.

TOWNSEND, G. C.; HARRIGER, A. 200 vs 20,000: Acm celebrations and grace hopper celebrations of women in computing. In: IEEE. **2019 IEEE Frontiers in Education Conference (FIE)**. 2019. p. 1–8. Disponível em: <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9028415>. Citado na página 90.

UNESCO. **Decifrar o código: educação de meninas e mulheres em ciências, tecnologia, engenharia e matemática (STEM)**. 2018. Citado na página 89.

VICARIO, G.; COLEMAN, S. A review of data science in business and industry and a future view. **Applied Stochastic Models in Business and Industry**, n/a, n. n/a, 2019. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asmb.2488>. Citado nas páginas 35 e 36.

VILLASEÑOR, T.; CELIS, S.; QUEUPIL, J. P.; PINTO, L.; ROJAS, M. The influence of early experiences and university environment for female students choosing geoscience programs: a case study at universidad de chile. **Advances in Geosciences**, Copernicus GmbH, v. 53, p. 227–244, 2020. Disponível em: <https://adgeo.copernicus.org/articles/53/227/2020/>. Citado na página 90.

WEININGER, A. Efficient execution of joins in a star schema. In: **Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data**. New York, NY, USA: ACM, 2002. (SIGMOD '02), p. 542–545. ISBN 1-58113-497-5. Disponível em: <https://doi.org/10.1145/564691.564754>. Citado na página 39.

WENCESLAU, F. d. L. *et al.* **A gramática na leitura pela leitura da gramática: verificação de desempenho em língua portuguesa do ENEM.** Tese (Doutorado), 2014. Citado na página 88.

ZAHARIA, M.; CHOWDHURY, M.; FRANKLIN, M. J.; SHENKER, S.; STOICA, I. Spark: Cluster computing with working sets. In: **Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing**. Berkeley, CA, USA: USENIX Association, 2010. (HotCloud'10),

p. 10–10. Disponível em: <http://dl.acm.org/citation.cfm?id=1863103.1863113>. Citado nas páginas 31 e 43.

ZAWISTOWSKA, A.; SADOWSKI, I. Filtered out, but not by skill: The gender gap in pursing mathematics at a high-stakes exam. **Sex Roles**, Springer, v. 80, n. 11, p. 724–734, 2019. Citado na página 90.


ICMC
USP
SÃO CARLOS