
REVISTA
DataPoint
Volume 01 - 2025
eISSN 3086-433X

EDITORES

José Augusto Theodosio Pazetti (Responsável)
Fatec Rubens Lara

Fábio Pessôa de Sá (Gerente)
Universidade Católica de Santos - UNISANTOS, Fatec Praia Grande

Fernando Ribeiro dos Santos (Textos)
Fatec Rubens Lara, Instituto Federal de São Paulo - IFSP

CONSELHO EDITORIAL

Adélia da Silva Saraiva
Fatec Rubens Lara

Ana Lúcia da Silva Kfouri
Universidade de São Paulo - USP

Denise Durante
Universidade de São Paulo - USP

Fábio Pessôa de Sá
Universidade Católica de Santos UNISANTOS
Fatec Praia Grande

Fernando Ribeiro dos Santos
Fatec Rubens Lara
Instituto Federal de São Paulo - IFSP

Gerson Prando
Universidade Paulista - UNIP

João Paulo Ferreira de Mello
Universidade de São Paulo - USP

José Augusto Theodosio Pazetti
Fatec Rubens Lara

Rafael Fonseca Araujo
Pontifícia Universidade Católica de São Paulo
PUC SP

Rodrigo Luiz Zanethi
Fatec Rubens Lara

ADMINISTRADOR TÉCNICO

Nilton Silva Saraiva • Fatec Rubens Lara

CAPA

Tarcísio Peres • Fatec Carapicuíba

APRESENTAÇÃO

A **Revista Datapoint** (eISSN 3086-433X) é um periódico acadêmico de acesso aberto vinculado ao curso de **Ciência de Dados** da **Fatec Baixada Santista – Rubens Lara**, dedicado à divulgação de produções científicas e tecnológicas que contribuam para o avanço do conhecimento nas áreas de Ciência de Dados, Sistemas Inteligentes e campos interdisciplinares correlatos. Em consonância com o movimento contemporâneo de ciência aberta e com a crescente centralidade dos dados na sociedade digital, a revista busca constituir-se como um espaço de diálogo acadêmico para pesquisadores, profissionais e estudantes interessados em compreender, analisar e aplicar métodos orientados por dados em diferentes contextos.

A **Revista Datapoint** compartilha o compromisso com a disseminação do conhecimento científico, a valorização da interdisciplinaridade e a promoção de investigações que dialoguem com os desafios contemporâneos da sociedade digital. Assim como essas publicações fomentam a circulação de ideias inovadoras e reflexões críticas sobre tecnologia, comunicação e ciência, a **Revista Datapoint** propõe-se a estimular a produção acadêmica que una rigor metodológico, inovação tecnológica e relevância social.

Nesse contexto, a revista acolhe trabalhos de natureza teórica, aplicada e metodológica, incluindo estudos de caso, revisões sistemáticas e relatos técnicos, priorizando pesquisas que apresentem soluções inovadoras baseadas em dados e que ampliem as possibilidades de aplicação da Ciência de Dados em diferentes áreas do conhecimento. Também se valoriza a produção científica de estudantes de graduação e pós-graduação, desde que orientada e desenvolvida dentro dos padrões acadêmicos de qualidade e ética científica.

Ao promover a difusão de pesquisas e experiências que articulam tecnologia, análise de dados e inovação, este periódico busca contribuir para o fortalecimento da cultura científica, incentivando a produção intelectual e a construção de um ambiente acadêmico colaborativo. Dessa forma, pretende consolidar-se como um espaço de reflexão, investigação e compartilhamento de conhecimentos que auxiliem na compreensão dos fenômenos contemporâneos mediados por dados e tecnologias digitais.

José Augusto Theodosio Pazetti
Coordenador do curso de Ciência de Dados

Os impactos da guerra entre Rússia e Ucrânia na economia brasileira: uma análise exploratória das movimentações de importação do porto de Santos

The impacts of the war between Russia and Ukraine on the Brazilian economy: an exploratory analysis of the cargo shipping to Santos port

Guilherme Onorio Pereira da Silva
Fatec Rubens Lara
guilherme.silva594@fatec.sp.gov.br

Igor Silva de Carvalho
Fatec Rubens Lara
igor.carvalho27@fatec.sp.gov.br

Fernando Ribeiro dos Santos
Fatec Rubens Lara
fernando.santos93@cps.sp.gov.br

Revista Datapoint
eISSN 3086-433X
Faculdade de Tecnologia Rubens Lara – FATEC
Ciência de Dados
Peridicidade: Anual
Vol 01, n. 01, 2025
revistadp@fatecrl.edu.br

Recebido: Jun 2025
Aceito: Set 2025
Publicado: Dez 2025

URL: https://www.fatecrl.edu.br/revista/datapoint/index.php/dp/article/view/2
DOI: https://doi.org/10.5281/zenodo.19118353

RESUMO

Esse estudo tem por objetivo verificar a influência da guerra entre Rússia e Ucrânia na importação de cargas pelo Porto de Santos, utilizando dados dos últimos cinco anos (2018-2023) para entender as movimentações durante o conflito iniciado a partir de 2022. O estudo, que é exploratório e de natureza quantitativa, foi feito sobre dados de importação a partir de *datasets* da Agência Nacional de Transportes Aquaviários (ANTAQ). A análise foi sintetizada por meio do processo de integração de dados *Extract, Transform, Load* (ETL – Extrair, Transformar, Carregar) e tabulações de Estatística Descritiva, evidenciando os produtos importados em função do peso de carga bruta. A análise culminou na criação de um *dashboard* interativo para analisar o impacto do conflito nas movimentações portuárias. Foi possível observar as alterações sobre o volume das cargas desembarcadas em território brasileiro, com os dados evidenciando uma aparente substituição da posição ucraniana para certos produtos chave pela Rússia.

PALAVRAS-CHAVE: Ciência de Dados; Porto de Santos; Importação; Mercadoria.

ABSTRACT

This study looks into the influence of the Russian-Ukrainian war on the importation rates at Porto de Santos, using data from the last five years (2018-2023) with the intent to understand the trade before and during the conflict. The study - which is exploratory and of quantitative nature - of the importation data was done with the data from the Agência Nacional de Transportes Aquaviários *(ANTAQ). The analysis was synthesized using the integration process Extract, Transform, Load (ETL) and Descriptive Statistics techniques, highlighting the most imported products in function of raw cargo weight. The analysis ended with the creation of a interactive dashboard, where the impact of the conflict could be more easily explored. It was possible to observe changes in the cargoes volumes that were disembarked at Brazilian territory, with the data suggesting a possible substitution of Ukraine's import position, for certain key-products, by Russia.*

KEY-WORDS: *Data Science; Santos Port; Import; Cargo.*

INTRODUÇÃO

A guerra entre a Rússia e a Ucrânia é resultado de interseções complexas entre fatores históricos, políticos e culturais que se estendem desde o século IX (Marshall, 2018). O conflito tem influenciado nas relações comerciais entre os países dependentes de seus produtos, e o Brasil, sendo um destes países, vem sendo afetado tanto pela instabilidade de preços quanto pela instabilidade geopolítica decorrentes da guerra (Nações Unidas Brasil, 2022).

Segundo informações da Conferência das Nações Unidas de Comércio e Desenvolvimento (UNCTAD) em seu relatório sobre a guerra na Ucrânia, o conflito impactou diretamente nas relações comerciais dos países, criando um aumento geral nos preços de alimentos, combustíveis e fertilizantes. Dessa maneira, o conflito afeta as cadeias de suprimentos, aumenta o custo de produção e gera instabilidade nas economias, principalmente àquelas dos territórios pertencentes à União Europeia (Louise, 2022).

Em 2024, as Nações Unidas projetaram a incerteza no mercado mundial em virtude da tensão geopolítica que se alonga desde 2022, essa incerteza existindo a despeito do avanço nos volumes de comércio global (com um aumento próximo a 7% em importações - Nações Unidas Brasil, 2024). Em contraste, no mesmo período, os números apresentados pela Ministério do Desenvolvimento, Indústria, Comércio e Serviços, revelam uma queda do volume total de importações brasileiras de 11,7% (MDIC, 2024).

A Ucrânia e a Rússia são países estratégicos com sua posição consolidada na comercialização de produtos como combustíveis, adubos potássicos, adubos nitrogenados, grãos e alumínio. Devido a posição dos dois países no comércio global, o conflito impactou diretamente no preço dos produtos, de acordo com o levantamento da Confederação Nacional de Indústria (CNI), aumentando em 51% o preço dos principais produtos. Em 2021, ano anterior a eclosão do conflito, a importação de 21 produtos exclusivos aos dois países correspondeu à 15,2% das importações totais do Brasil, denotando assim uma relação de dependência destes produtos (CNI, 2022).

Os portos são, em geral, o principal canal para o envio e recebimento de mercadorias, com o comércio marítimo permanecendo como um dos principais meios de transporte de produtos, englobando cerca de 80% das movimentações comerciais globais (Corrêia, 2023). No Brasil, o Porto de Santos foi responsável pelo embarque e desembarque de mais de 157 milhões de toneladas apenas em 2023 (APS, 2023), superando em 4,9% a movimentação do ano anterior (150,3 milhões), com algumas destas trocas resultando de acordos com a Rússia e Ucrânia.

Esta pesquisa, a fim de auxiliar na análise dos efeitos do conflito Ucrânia e Rússia no comércio do Porto de Santos, tem como objetivo principal verificar a influência da guerra entre Rússia e Ucrânia na importação de cargas pelo Porto de Santos. Para tal, o estudo visa: (1) analisar os dados de importação do Porto de Santos no período de 2018 a 2023; (2) Identificar as movimentações de importação de produtos da Rússia e Ucrânia para o Porto de Santos.

Foi realizada uma análise exploratória de natureza quantitativa dos dados de importação e exportação do Brasil dentre os anos 2018 e 2023 para contextualizar a influência do conflito nas balanças comerciais do Porto de Santos, a partir de *datasets* da ANTAQ, para os dados de tráfego aquático.

Os dados foram filtrados utilizando-se: o software de planilha, Libre Office Calc, para uma visualização preliminar dos dados brutos; a linguagem de programação Python (com a bibliotecas Pandas por meio do Jupyter Notebook) para o cruzamento de dados entres os *datasets.* Por meio do Power BI os resultados foram elaborados em *dashboards* interativos.

O presente estudo está estruturado nas seguintes seções: esta introdução, o referencial teórico, os procedimentos metodológicos, a análise dos resultados, as considerações finais e as referências.

1. FUNDAMENTAÇÃO TEÓRICA

Apresentam-se a seguir os conceitos necessários para a análise e compreensão dos impactos da guerra entre a Rússia e a Ucrânia tiveram na economia brasileira, assim como observada a partir do Porto de Santos.

1.1 CONTEXTUALIZAÇÃO HISTÓRICA

A guerra entre Rússia e Ucrânia, iniciada em 2014 com a anexação da Crimeia, é um desdobramento de um longo histórico de tensões entre os dois países. O apoio russo a movimentos separatistas no leste da Ucrânia culminou em um conflito armado que, embora tenha começado como uma disputa regional, rapidamente ganhou dimensões geopolíticas (Pedro, 2023).

As relações entre Rússia e Ucrânia são conflituosas desde a formação dos dois estados, passando por confrontos históricos e anexação de território durante o Império Russo (Czarismo) do século XVI, ao estabelecimento e separação da União das Repúblicas Socialistas Soviéticas (URSS) que vigorou entre 1922 e 1991, até chegarmos a 2022, quando um confronto entre as nações interfere diretamente na economia mundial (Pedro, 2023).

A dissolução da União Soviética e a independência ucraniana em 1991 não foram suficientes para evitar a perseguição do país do leste da Europa por parte da Rússia. Por situar-se em uma região estratégica, a Ucrânia manteve-se sob forte vigilância dos russos, ainda que dependesse de insumos e recursos energéticos para a sobrevivência do país, pois qualquer movimento em acordos comerciais ou militares com o ocidente como a Organização do Tratado do Atlântico Norte (OTAN) seria prejudicial (Marshall, 2018).

A busca da Rússia pelo controle da região da Crimeia também é um ponto excepcional na análise da guerra que se desenrola. A utilização das rotas marítimas para realizar a comercialização dos seus produtos são importantes, no entanto, os russos carecem de um porto de águas mornas e longe de regiões congeladas ou sob o domínio de outras nações - como o Japão (Marshall, 2018).

A guerra russo-ucraniana começa em 2014, a partir do momento que a Rússia anexa o território da Crimeia, assim como apoia movimentos de separatistas a favor do país agressor ao leste de seu país, e prolonga o embate ao invadir o território ucraniano em 2022 (CNN Brasil, 2022).

Com o desejo da população de ingressar na União Europeia e evitar acordos que favorecessem uma política pró-Rússia, o acesso ao Mar Negro pela região da Crimeia se tornou o principal objetivo do presidente russo Vladimir Putin. O Porto de Sebastopol é crucial para a frota russa no Mar Negro, oferecendo não apenas uma posição estratégica, mas também facilitando operações militares e comerciais (Marshall, 2018).

1.2 CONTEXTUALIZAÇÃO HISTÓRICA

Seja em blocos econômicos e acordos bilaterais, o Brasil beneficia-se do avanço econômico de países emergentes por compreenderem que há espaço para preencher em soluções para a economia brasileira (GOV, 2024).

Desde o final da URSS, países como Rússia e Ucrânia mantêm o diálogo e relações bilaterais para acordos de nível comercial como para cooperação tecnológica e militar brasileira e dos países parceiros.

As relações russo brasileiras tiveram início no século XIX, após a independência do Brasil, foi estabelecida uma representação diplomática entre os dois países (Rosario, 2000, apud Bacigalup, 2022, p. 59). Passando por um período de guerras e a dissolução da URSS, as relações comerciais entre as duas nações intensificam-se a partir da abertura de mercado (Grieco, 1992).

Assim, Brasil e Rússia encaminharam uma relação amistosa desde 1998, com constante viagens de chefes de Estado e embaixadores entre os países, a criação de acordos econômicos e tecnológicos (Ministério das Relações Exteriores, 2024), e a reunião para o prosseguimento dos acordos de intercâmbio comercial durante o período de conflito (Valor Econômico, 2022).

Por sua vez, a Ucrânia, embora não seja o maior parceiro comercial, tem sua importância no fornecimento de tecnologias e produtos específicos com o Brasil. As relações brasileiras com a Ucrânia iniciaram-se em dezembro de 1991 quando o Brasil reconheceu a independência dela após a dissolução da União Soviética, estabelecendo então relações bilaterais e firmando pactos como Acordo de Cooperação Econômico-Comercial em 1995 (Ministério das Relações Exteriores, 2024).

A Ucrânia mantém uma postura comercial de exportação para com os outros países, repassando ao Brasil commodities, em menor escala, como grãos (Ministério das Relações Exteriores, 2024), produtos farmacêuticos, e maquinários como aquecedores, laminados, rolamentos de esferas, cujo valores de importações não ultrapassaram US$ 105 milhões desde 2021 (ComexStat, 2024). No que diz respeito a importações, a Ucrânia participa do acordo com o recebimento de amendoim, mais de 40%, café, aparelhos mecânicos e matérias brutas de animais (tripas, bexigas e estômago) atingindo mais de 60% da balança comercial do país europeu (ComexStat, 2024).

No entanto, a Rússia acaba sendo o sexto maior destino das exportações brasileiras (Ministério das Relações Exteriores, 2024) e torna-se o principal fornecedor de combustível diesel e fertilizantes nos anos subsequentes à guerra (Valor Econômico, 2024).

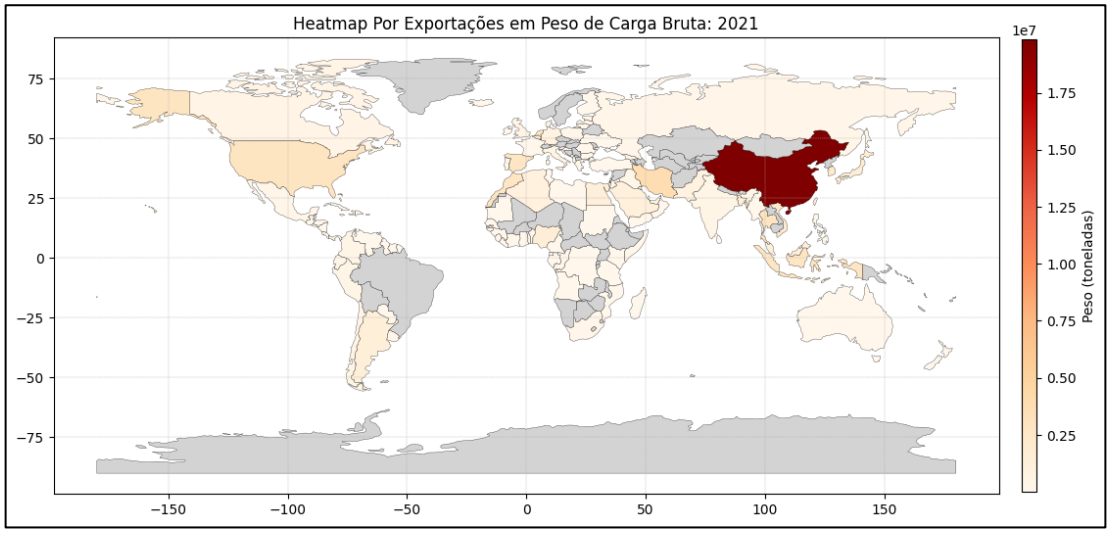
1.3 O PORTO DE SANTOS

O Porto de Santos é considerado o principal ponto de entrada e saída de mercadorias do Brasil, desempenhando um papel crucial na balança comercial do país. Responsável por uma parcela significativa do comércio exterior brasileiro (Figura 1), o porto é vital para a exportação de produtos como soja, café, açúcar e carnes vindos em sua maioria de regiões do Centro-Oeste (Mato Grosso, Mato Grosso do Sul, Goiás) e do Sudeste (São Paulo e Minas Gerais) (APS, 2024).

No entanto, a guerra entre Rússia e Ucrânia, iniciada em 2022, trouxe incertezas para as operações do Porto de Santos. Segundo reportagem publicada no G1 (2022), especialistas apontaram que o funcionamento seria afetado pelo conflito caso existisse a falta de insumos (G1 Santos, 2022).

A partir do porto de Santos, produtos que anteriormente eram exportados com regularidade, como amendoim, carne bovina, milho e café, passaram a enfrentar atrasos constantes, resultando em perdas significativas tanto para os exportadores brasileiros quanto para os importadores estrangeiros (Sant'Ana, 2022). Nesse contexto, o problema alinha-se ao transporte para Rússia e Ucrânia, afetando também a chegada de insumos essenciais, como trigo, milho e fertilizantes, gerando um conflito entre a necessidade de uma logística eficiente e os desafios de garantir o abastecimento para os países (Santimaria, 2022).

Figura 1 - Representação total de carga russa



Fonte: Elaborado a partir dos dados dos estatísticos aquaviários de 2018-2023[1]

[1] Mapa de calor de peso de carga bruta (em toneladas) das cargas saídas a partir do Porto de Santos coloridas de acordo com o destino e o peso exportado a um destino em particular (2021). Quanto mais avermelhado o tom, maior o peso de carga escoada para àquele país.

Apesar desses desafios, o Porto de Santos continua a ser um ativo estratégico para a balança comercial brasileira, com um movimento de carga abrangendo o mundo todo, como pode ser visto pela Figura 1. A movimentação de cargas no porto em 2023 superou a marca histórica do ano anterior, impulsionado pelas mercadorias do agronegócio para exportação e a importação de óleo diesel e gasóleo para importação (APS, 2023).

Até primeiro semestre de 2024, o Porto de Santos teve aumento de 8,42% do transporte em relação ao mesmo período do ano anterior, representando mais de 68 milhões de toneladas - movimentando 10,6% de toda a carga portuária do país em toneladas (G1 ECONOMIA, 2024). Desde 2023, os registros de movimentação do porto superaram a marca de 28% de participação na balança comercial brasileira, sendo no primeiro semestre de 2024 aumento para 28,5% (EBC, 2024).

2. PROCEDIMENTOS METODOLÓGICOS

As pesquisas científicas, segundo Medeiros (2014, p. 33), são assim chamadas "[…] se sua realização for objeto de investigação planejada, desenvolvida e redigida conforme **normas metodológicas** consagradas pela ciência" (grifo nosso). Assim sendo, o presente estudo, caracterizado como pesquisa de objetivo exploratório e de natureza quantitativa compreende, por meio da análise de dados e levantamento de *datasets* relacionados ao comércio exterior por meio do Porto de Santos, no período compreendido entre 2019 e 2023, ou seja, utiliza-se de uma amostra não-probabilística.

2.1 CONJUNTO DE DADOS

A pesquisa em Ciência de Dados requer dados factíveis e de autoridade específica da área (ou do tema, do assunto, do cenário) para sua análise a fim de manter a qualidade esperada (Weitzel, 2000, *apud* Brasileiro, 2022). Os conjuntos de dados escolhidos, de acesso aberto, contêm *datasets* diversos, abrangentes e pesquisados (ou elaborados) por um instituto de pesquisa respeitável, a Agência Nacional de Transportes Aquaviários (ANTAQ) vinculada ao Ministério de Portos e Aeroportos (MPA), e de autoridade na sua respectiva área de estudo.

Seus dados são acessíveis por meio dos seus Estatísticos Aquaviários, publicados periodicamente, nos quais se é possível obter dados e estatísticas diversas sobre navegação marítima.

2.2 TÉCNICAS DE CIÊNCIA DE DADOS

Ao analisar os dados extraídos dos *datasets* da ANTAQ foram utilizados processos e técnicas de Ciência de Dados e de Estatística Descritiva para a organização e interpretação adequada dos dados ali contidos. As técnicas utilizadas foram: *Extract, Transform, Load* (ETL), e tabelas de frequência contínua e medidas de variação.

2.2.1 ETL

Para o processamento geral dos dados foram utilizadas técnicas de ETL, sendo que esta é um processo de: (1) agregação de dados de múltiplas fontes; (2) transformações destes dados com algum intuito específico (como análise, por exemplo); (3) armazenamento em um conjunto de dados significativos (IBM, [2023?]).

Nesta pesquisa, os processos de ETL descritos foram utilizados afim de preparar os dados para sua análise, após as extrações e transformações necessárias, em *dashboards* e gráficos, sendo que o carregamento final dos dados foram feitos para a análise das suas partes relevantes.

2.2.2 Estatística Descritiva

Estatística descritiva é o ramo da Estatística que trabalha com a organização e apresentação dos dados (Akamine; Yamamoto, 2014). Foram utilizadas técnicas de estatística descritiva para a organização e apresentação dos dados.

2.3 FERRAMENTAS DE CIÊNCIA DE DADOS

No que diz respeito a análise dos dados, foram utilizadas, nas etapas de coleta, filtragem e processamento de dados, uma série de ferramentas especializadas, sendo estas:

a) **Python**: uma linguagem de programação interpretativa, interativa, orientada a objetos e funcional. Contém tipos de dados de alto grau de dinamicidade, tendo uma sintaxe clara e poderosa. Além de tudo Python é uma linguagem portável, podendo rodar em vários sistemas operacionais (Python, 2024). A escolha desta ferramenta foi feita levando-se em conta o grande acesso e acessibilidade a ferramentas de análise de dados, além da grande disponibilidade de bibliotecas externas existentes, e facilidade geral de uso.

b) **Pandas**: uma flexível biblioteca código aberto de análise e manipulação de dados feita para a linguagem de programação Python (Pandas, 2024), e estruturada para lidar, principalmente, com dados de *arrays*, de tipo homogêneo (McKinney, 2023, p. 153). Esta pesquisa utiliza a biblioteca Pandas com o objetivo de filtrar e processar os *datasets* coletados. Sendo estes em formato *Comma Separeted Values* (CSV – Valores separados por vírgula), o uso desta biblioteca se demonstra apropriado, uma vez que é possível a importação de dados brutos em CSV e exportação de dados trabalhados em formatos e OpenDocument Spreadsheet (ODS – Planilha DocumentoAberto).

c) **Libre Office Calc**: software de código aberto, parte da suíte do Libre Office, de manipulação e organização de planilhas, e cálculo de dados. Suas funcionalidades permitem a facilidade da extração de informações brutas de banco de dados corporativos, podendo assim tabulá-los e convertê-los em informações significativas. O programa pode lidar tanto com arquivos ODS (Open Document – Documento Aberto) quanto os próprios arquivos do Excel (LibreOffice, [20??]). Nesta pesquisa, o Libre Office Calc é utilizado para uma análise preliminar e inspeção dos dados brutos e disposição e tabulação dos dados trabalhados, com o objetivo de uma análise holística das informações processadas.

d) **Power BI**: *Dashboards*, assim como descritos no contexto do *software* Power BI, são análises dispostas em blocos com visualizações relevantes, contando "histórias", em uma única página, que contenham seus elementos mais importantes (Microsoft, 2023). O Power BI é um uma coleção de softwares de *Business Intelligence* (Inteligência de Negócios), aplicativos e conectores trabalhando em conjunto para transformarem e relacionarem fontes de dados, sejam eles advindos de uma planilha Excel ou de uma data *warehouse* (on-premise – local - ou em nuvem) (Microsoft, 2024). Com os dados carregados e filtrados no Power BI é possível visualização e elaboração de dashboards interativos. Este trabalho utilizou o Power BI para a apresentação dos resultados das análises feitas. Visando a elaboração de um *dashboard* interativo explorando os fluxos dos dados no período da pesquisa.

2.4 EXTRAÇÃO

Os dados dos estatísticos aquaviários da ANTAQ foram extraídos através do Portal de Dados Abertos (PDA) do Gov.br: acessando, no rodapé da página, a seção “conjunto de dados”, e no campo de pesquisa inserindo “EA”, selecionando o resultado “Estatísticos Aquaviários (EA)” da ANTAQ. Logo após “Recursos” e por fim “Acessar o recurso” tanto para o Estatístico quanto para os Metadados – este último para a utilização adequada dos *datasets*[2].

2.5 TRANSFORMAÇÃO

Para a análise dos diversos arquivos foram utilizadas uma gama de procedimentos auxiliados por várias ferramentas de análise de dados, mas, devido ao grande tamanho de muito dos arquivos foram tomadas rotas mais indiretas: houve a criação de arquivos intermediários em alguns dos casos, afim de possibilitar a análise completa dos dados com os recursos computacionais disponíveis.

Os dados retirados dos *datasets* relevantes foram compilados afim de obter o valor de peso de carga bruta para as cargas que foram importados pelo Porto de Santos no período compreendido entre 2018 e 2023. Para todos os *datasets* extraídos da ANTAQ, quando lidos no Pandas, tiveram na função “read_csv” o parâmetro “sep” (separador) com o valor do delimitador “;”[3] e o parâmetro “decimal ”como o valor de “,”.

2.5.1 Transformações Iniciais

A análise do peso de carga bruta foi feita utilizando-se de técnica de tabulação de estatística descritiva. Tanto a Rússia como a Ucrânia tiveram os seus valores de peso de carga bruta isolados para o sentido de importação e tabulados em uma tabela pivô de frequência com as colunas sendo os anos (2018 – 2023) e as linhas correspondendo às mercadorias. A das duas tabelas criadas, uma para cada país, elas mescladas em um único documento de planilha *Open Document Sheets* (ODS) ocupando cada uma um *sheet* (painel) no arquivo.

[2] Os dados podem também serem obtidos diretamente da ANTAQ por meio de https://web3.antaq.gov.br/ea/sense/download.html#pt
[3] Em arquivos CSV, a despeito do nome, pode ocorrer a separação dos valores por delimitadores além da vírgula.

Para a extração e análise dos dados de peso de carga bruta, como foi supracitado, foram utilizados os *datasets* da ANTAQ, retirados do seu estatístico aquaviário sendo, especificamente, utilizados os seguintes *datasets*[4]:

1. Carga – os arquivos de 2018 a 2023;
2. Carga Conteinerizada – os arquivos de 2018 a 2023;
3. E também foram utilizados os arquivos de tabelas auxiliares:
   a) Instalação Origem;
   b) Instalação Destino;
   c) Mercadoria;
   d) Mercadoria Conteinerizada[5]

Todas as mudanças no *Pandas* foram feitas por meio da inserção dos *datasets* em uma estrutura do Pandas, similar a uma tabela, chamada de *DataFrame*.

Para cada ano do período estudado (2018 – 2023), carregou-se através do *Pandas*, em um *DataFrame*, as tabelas de Carga e aplicou-se as seguintes transformações:

I. Manteve-se apenas os registros contidos na coluna "Destino" com o valor "BRSSZ" (sendo esse o código do Porto de Santos).
II. Modificou-se a coluna "Origem" para que ficasse com o valor de "CDBigramaOrigem" (campo com o código bigrama dos países de origem da carga) correspondente - da tabela Instalação Origem – nos casos em que o valor desta última fosse igual a "RU" para Rússia e "UA" para Ucrânia[6], descartando as demais entradas da tabela carga.

Esta junção foi feita por meio dos seguintes processos:

[4] A modelagem de dados pode ser encontrada em: https://web3.antaq.gov.br/ea/sense/download.html#pt em "*Download* modelagem de dados".
[5] Na prática, devido a limitação de recursos computacionais, optou-se por utilizar apenas a tabela de Mercadoria na maioria dos casos, uma vez que os valores relevantes para a análise (códigos de mercadoria numéricos) são idênticos em ambos os datasets.
[6] Código do Porto de Santos e código bigrama dos países conferidos na tabela "Instalação Origem" e confirmados em: https://web3.antaq.gov.br/portalv3/sdpv2servicosonline/ConsultarPorto.aspxCódigo

I. Através da função *merge* do Pandas, juntando a tabela *Carga e Instalação Origem*, com o parâmetro *how* igual a "*left*"[7], e o parâmetro *on* para "Origem" afim de juntar baseado no valor da coluna que ambos têm em comum.
II. Mudou-se o valor da coluna "Origem" da tabela Carga para os correspondentes da coluna "CDBigramaOrigem" recém juntada à tabela.
III. Após a junção aplicou-se, na tabela Carga, a função *drop* para as colunas originadas da tabela Instalação Origem.
IV. Adicionou-se a coluna "Ano" à tabela, contendo o mesmo valor referente ao ano do arquivo, para todos os registros da tabela.
V. Retirou-se os registros que tinha o valor de "ConteinerEstado" como "Vazio".

2.5.2 Importações Segmentadas por Ano e Mercadoria

Da tabela Carga manteve-se apenas os registros contendo na coluna "Sentido" o valor "Desembarque", e "Tipos de Operação da Carga" contendo em sua *string* a palavra "Importação". Converteu-se a coluna "Ano" para o tipo de dados *datetime*.

Juntou-se - com a função "merge" e com seu parâmetro "*how*" para "*left*" - a tabela Carga com as tabelas Carga Conteinerizadas, utilizando a coluna "IDCarga" como referência.

Esta junção foi feita:

a) Filtrando a tabela Carga pela coluna "Ano" para acessar apenas as colunas com o ano correspondente ao da Carga Conteinerizada relevante naquele instante1;
b) Aplicando a função *combine_first* da coluna "VLPesoCargaConteinerizada", recém inserida pela função merge, para a coluna "VLPesoCargaBruta", efetivamente substituindo os valores desta última coluna pelos valores correspondentes – quando não nulos – naquela.
c) Aplicando o mesmo processo acima para as colunas "CDMercadoriaConteinerizada" e "CDMercadoria".
d) Retira-se da tabela Carga as colunas originadas da tabela Carga Conteinerizada.

[7] Equivalente a junção "*Left Outer Join*" em termos de *Structured Query Language* (SQL).

Após a junção adequada dos valores entre os valores de Carga e Carga Conteinerizada, substituiu-se a coluna “CDMercadoria” da tabela Carga pela coluna “Grupo de Mercadoria” da tabela Mercadoria nas entradas correspondentes por meio de um merge pela coluna CDMercadoria.

Há então, para a Rússia e a Ucrânia, a criação de um “*crosstab*” - estrutura do Pandas que modela uma tabela pivô de frequência – da tabela Carga por meio da função homônima, com (1) o índice para da coluna “Grupo de Mercadoria”, (2) colunas para “Anos”, (3) valores de “VLPesoCargaBruta”, e os parâmetros (4) “*margins*” para *True* e (5) “*margins_name*” para “Total”. Por fim, aplica-se um ‘*fillna’* com o valor para 0 (zero). Efetivamente criando uma tabela de frequência acumulada do Peso de Carga Bruta para uma dada mercadoria ano a ano.

Por fim, cria-se um arquivo ODS com as duas tabelas de frequência geradas, cada qual em *sheets* únicas. No Libre Office Calc aplica-se uma formatação visual sobre os valores de peso para exibirem apenas duas casas decimais, e converte as colunas com os anos para o tipo de dados “Data”.

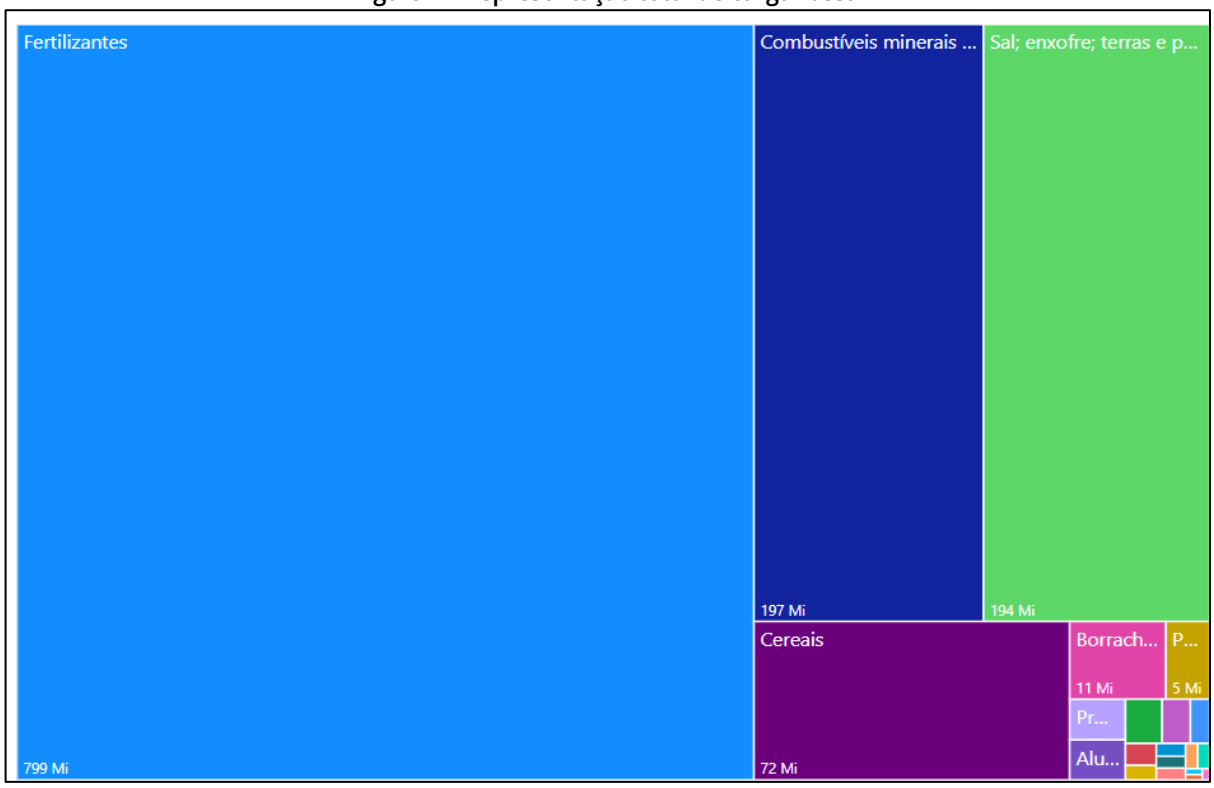
3. RESULTADOS E DISCUSSÃO

A partir do banco de dados relacionadas às importações da Rússia e Ucrânia com o Porto de Santos, entende-se que há uma proporção diferente entre os números apresentados. Isso será discutido em mais detalhes nas próximas subseções.

3.1 RÚSSIA

A análise foi feita em cima de categorias de produtos importados, sendo os números apresentados na figura abaixo em relação ao ponto comum (1 tonelada), declarados por um valor nulo ou que não apresentava valor (simbologia vazia). Dessa forma, foi considerado as cargas que não seriam excluídas da análise propriamente e seriam definidas a partir de um valor comum em zero (0).

Figura 2 - Representação total de carga russa



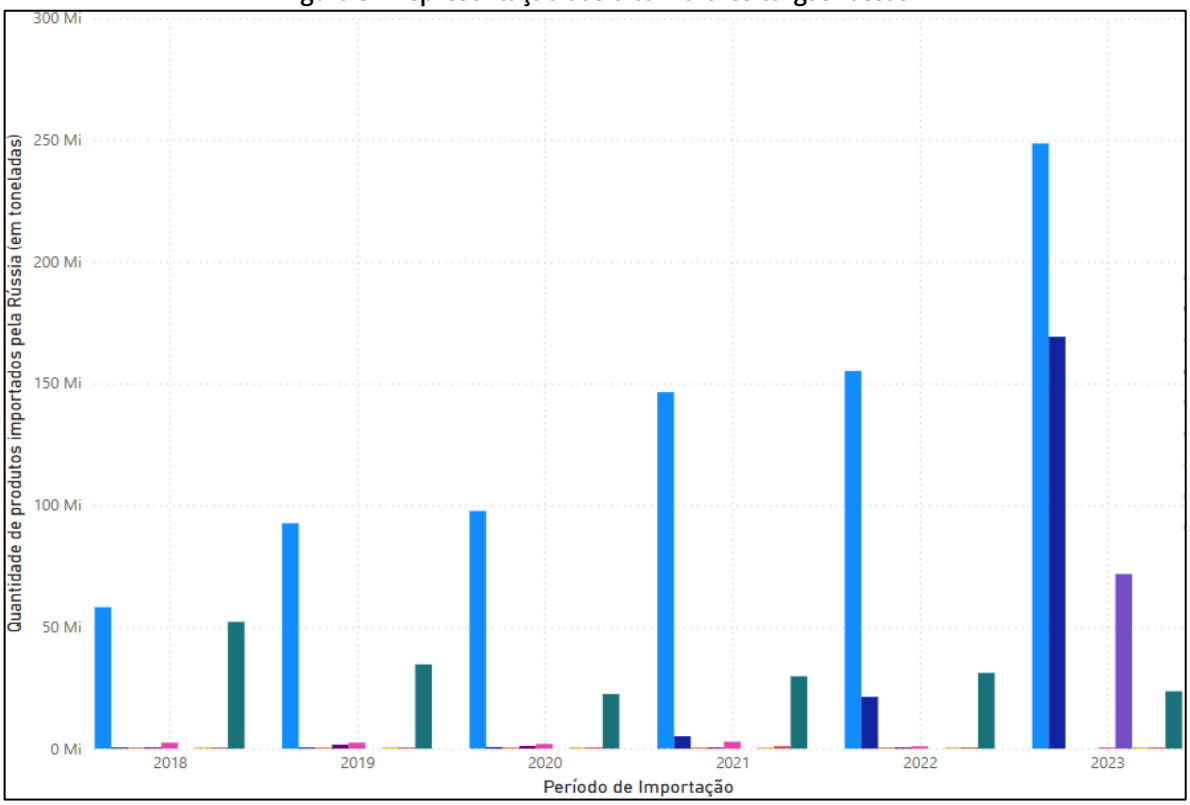
Fonte: Elaborado a partir dos dados dos estatísticos aquaviários de 2018-2023.
Legenda: Azul claro - Fertilizantes; Azul escuro - Combustíveis minerais e derivados; Verde claro - Sal; enxofre; terras e pedras; Roxo escuro - Cereais.

Em relação a Figura 2, é observado que a carga total de importação russa, com sua representação gráfica em comparação aos outros itens, apontando para os principais produtos de uso nacional. Adubos fertilizantes, combustíveis e sedimentos de construção seguem como os principais ativos nessa relação comercial.

Os fertilizantes têm sua importância difundida para o setor de agronegócio brasileiro, e sendo importância nos veículos de comunicação e ações do Governo Federal para que continuassem sendo abastecidos durante o período da pandemia (Valor Econômico, 2022), com crescimento entre 2020 e 2021 em 49,96%, e até mesmo em período de conflito entre Rússia e Ucrânia, com crescimento de carga importada entre 2022 e 2023 em até 60,2%, conforme é visto a visualização na Figura 3. Os combustíveis, que também tiveram um aumento no ano seguinte ao conflito, também é reflexo de influências das relações comerciais da Europa (Valor Econômico, 2024), o que contribuiu com o aumento de mais de 649% em 2023.

Outro ponto importante é sobre o grupo nomeado no gráfico como "cereais", que podem ser entendidos por cereais e outros grãos, sendo a categoria das principais commodities dos países do leste europeu. Ao comparar com os anos anteriores na Figura 3, o Porto de Santos recebeu números que não chegavam a um milhão em carga importada, com índice zero ou quase zerado, sem estabelecer muito impacto na análise, mas alcançou o número superior a 70 mil toneladas até o ano de 2023.

Figura 3 - Representação das oito maiores cargas russas



Fonte: Elaborado a partir dos dados dos estatísticos aquaviários de 2018-2023.
Legenda: Azul claro - Adubos (fertilizantes); Azul escuro - Combustíveis minerais, óleos minerais e produtos da sua destilação; Laranja - Chumbo e suas obras; Vinho - Alumínio e suas obras; Rosa - Borracha e suas obras; Roxo - Cereais; Amarelo - Máquinas, aparelhos e materiais elétricos, e suas partes; aparelhos de gravação ou reprodução de som; Vermelho - Plásticos e suas obras; Verde escuro - Sal; enxofre; terras e pedras; gesso, cal e cimento.

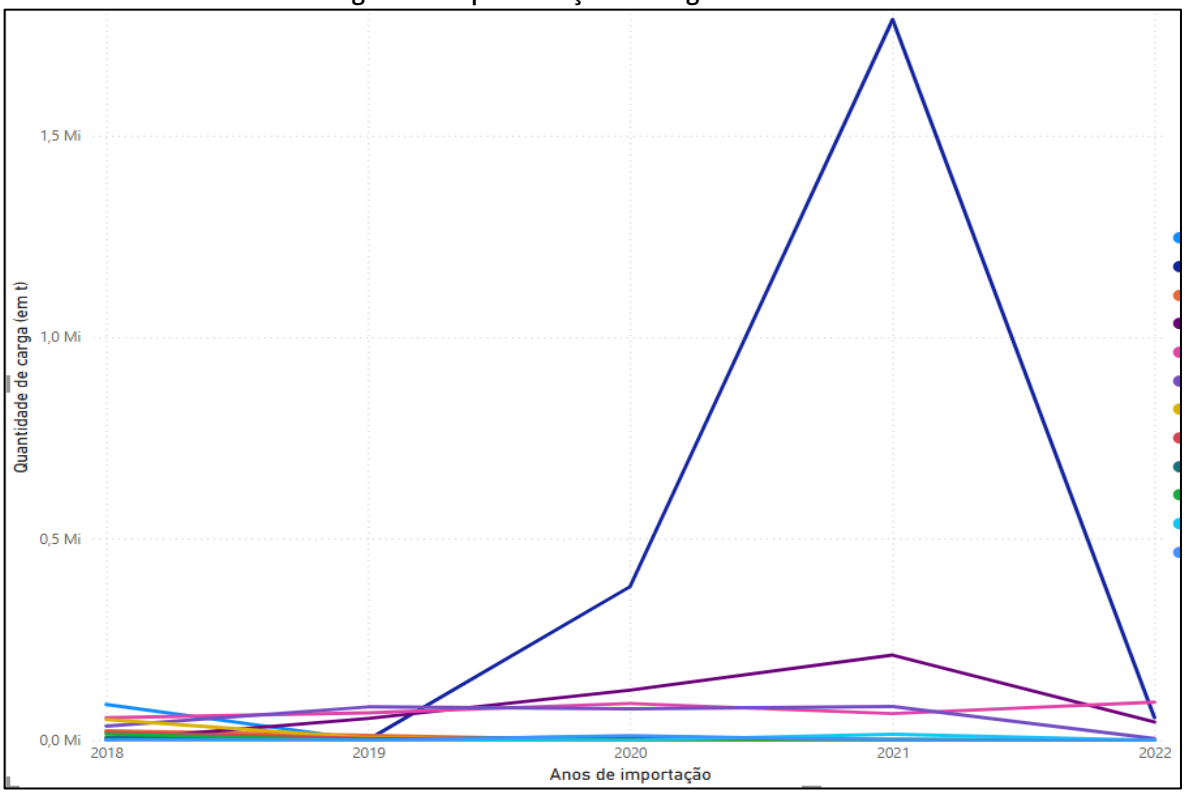
Por outro lado, ao separar os oito principais produtos importados, torna-se evidente (Figura 3) a desproporção do valor de "cereais" em relação aos anos anteriores, que contém valores zerados. Esse fator pode ser explicado pela ausência da Ucrânia no comércio de grãos e a Rússia ocupando o seu espaço de destaque, visto que o Brasil é depende dessa commodity (CNN Brasil, 2023).

Também é observada a tendência de crescimento para os fertilizantes e, principalmente, combustíveis com aumento em cerca de 694% de cargas importadas.

3.2 UCRÂNIA

Salienta-se que, o banco de dados em relação à mercadoria ucraniana além de possuir números de cargas importadas inferiores ao russo, também é visto que não há números disponíveis pelo portal do governo a partir de 2022. Observa-se dentro dos documentos de importação que existem categorias para mineração, maquinário e ferramentas metalúrgicas e matéria-prima em ferro, aço e outros tipos de suprimentos.

Figura 4 - Representação de cargas ucranianas



Fonte: Elaborado a partir dos dados dos estatísticos aquaviários de 2018-2023.
Legenda: Azul claro - Adubos (fertilizantes); Azul escuro - Plásticos e suas obras; Laranja - Cereais; Roxo escuro - Obras de ferro fundido, ferro ou aço; Rosa - Minérios, escórias e cinzas; Vinho - Ferro fundido, ferro e aço; Amarelo - Máquinas, aparelhos e materiais elétricos, e suas partes; Vermelho - Preparações à base de cereais, farinhas, amidos, féculas; Verde escuro - Reatores nucleares, caldeiras, máquinas, aparelhos; Verde claro - Extratos tanantes e tintoriais; taninos e seus derivados; Cinza - Obras de pedra, gesso, cimento, amianto; Azul bebê - Gorduras e óleos animais ou vegetais.

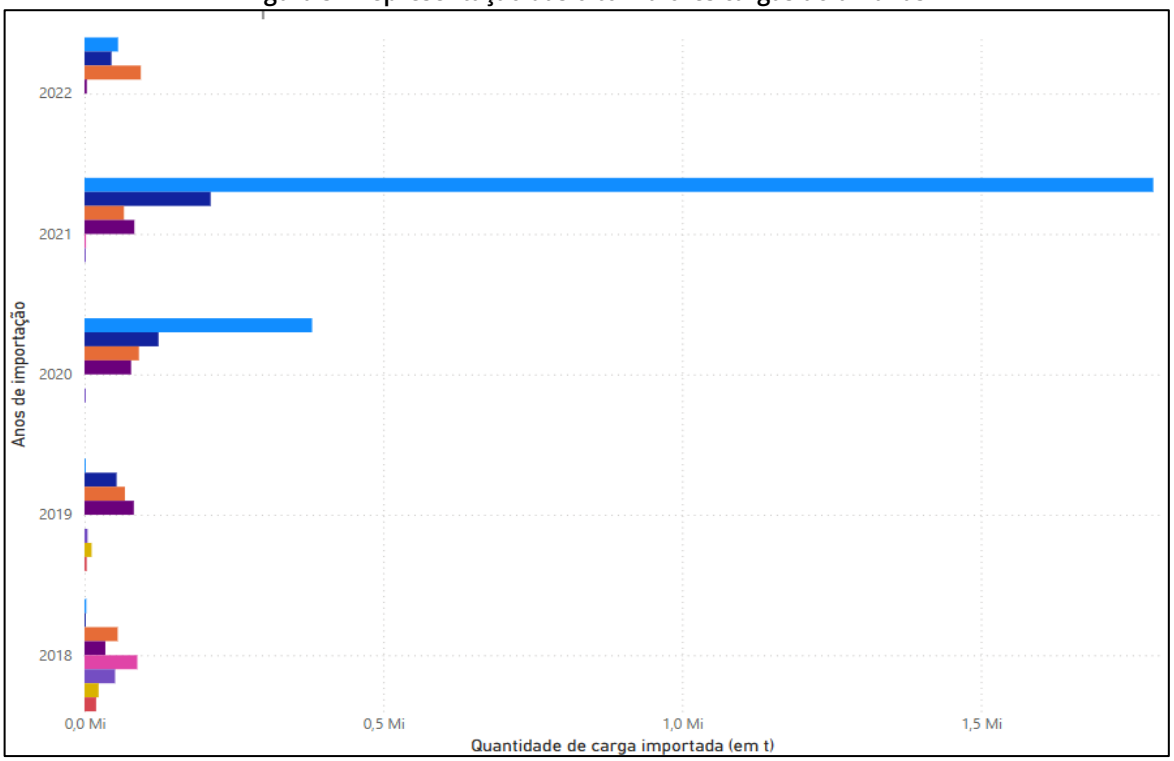
Ao observar a Figura 4 e aplicando os mesmos parâmetros de comparação para os 60 grupos de mercadorias importadas da Ucrânia, houve oito em destaque e, um deles (plástico e suas obras), acaba superando em valores relacionados à sua carga no primeiro ano subsequente ao período de pandemia causado pelo coronavírus (17 mil em 2021) e uma queda de 96,8% (451 toneladas) no ano seguinte relacionando-se diretamente com o período do conflito com a Rússia, e que também que pode ser entendido como uma tendência daquele ano, pois o país passou a importar resinas plásticas (Fontes, 2022) e a produção de plástico nacional não foi possível para suprir as demandas daquele período (Abre, 2022).

Todas outras categorias (Figura 5) seguem a média de carga importada de 120 e 160 mil toneladas de carga bruta, com oscilações pontuais entre 8% e 14% entre os anos de 2018 e 2021 das mercadorias importadas pela Ucrânia, com o maior índice de alteração em nível de queda acontecem em 2022.

O movimento se repete com a categoria “obras de ferro fundido, ferro ou aço” com um aumento de 70,3% (1238 para 2110 toneladas) da importação de 2021 e uma queda de 78,6% (451). O que teve movimento inverso foi “minérios, escórias e cinzas” com seu crescimento de 42,7% em 2022 em comparação ao ano anterior.

Em relação às importações da Ucrânia, existem mercadorias que seriam mais importantes do que suplementares à carga russa. Dos oito destacados por terem maior movimentação, verifica-se que fertilizantes e cereais (milho, cevada, trigo) seguem com quantidades regulares ano a ano e contribuem ativamente ao agronegócio brasileiro, mesmo que, esse espaço tenha sido ocupado pela Rússia com o andamento do conflito.

Figura 5 - Representação das oito maiores cargas ucranianas



Fonte: Elaborado a partir dos dados dos estatísticos aquaviários de 2018-2023.
Legenda: Azul claro - Plásticos e suas obras; Azul escuro - Obras de ferro fundido, ferro ou aço; Laranja - Minérios, escórias e cinzas; Roxo - Ferro fundido, ferro e aço; Rosa - Adubos (fertilizantes); Lilás - Máquinas, aparelhos e materiais elétricos, e suas partes; Amarelo - Cereais; Vermelho escuro - Preparações à base de cereais, farinhas, amidos, féculas.

4. CONSIDERAÇÕES FINAIS

Finaliza-se esta pesquisa sobre as movimentações de importações da Rússia e da Ucrânia por meio do porto de Santos durante o período da guerra na Ucrânia com considerações acerca do impacto da guerra nas relações comerciais brasileiras. A seguinte análise atingiu o objetivo proposto ao, auxiliada por meio dos *dashboards* interativos, verificar uma aparente conexão entre a guerra e o perfil e volume das importações da Ucrânia e principalmente da Rússia.

Ambos os países, nos anos anteriores a guerra, eram um vetor de produtos de matéria-prima, tanto para o setor metalúrgico como para o setor agrário, e esse padrão se manteve mesmo durante a pandemia, mas com a guerra houve algumas mudanças. Denota-se a possível substituição da posição da Ucrânia no que diz respeito a importações de cereais pela Rússia, após o início dos conflitos, elevando os valores anteriormente "zerados" a um dos oito mais importados no Porto de Santos dentre as mercadorias russas.

Já em relação a Ucrânia, o seu perfil de exportação durante a guerra é não caracterizável, mas logo antes o início do conflito, houve um pico em termos de exportação do grupo de mercadoria "plástico e suas obras". Enquanto a categoria em si não constitui uma anormalidade, a quantidade, mesmo levando em o crescimento do ano anterior, constitui uma incógnita para a análise feita neste artigo, exigindo uma pesquisa mais completa afim de corroborar estes resultados.

Destaca-se que este estudo tem suas limitações, uma vez que os dados, retirados dos estatísticos aquaviários da ANTAQ, tendo a variação e consistência destes dados possíveis impactos nos resultados desta pesquisa. Sugere-se pesquisas futuras mais aprofundadas e com análises mais completas, com uma maior gama de variáveis afim de melhor confirmar ou refutar os resultados deste artigo.

Em suma, este artigo contribui para, não somente o entendimento do volume de importações, por peso de carga, do Porto de Santos da Ucrânia e da Rússia com a guerra, como também para as taxas por peso de carga dos dois países antes da guerra também, elucidando as complexas relações de dependência comercial com a Ucrânia e Rússia. O entendimento de como conflitos externos afetam a economia nacional é de grande importância para a melhor qualidade das relações comerciais e dos benefícios por elas geradas.

REFERÊNCIAS

ABRE. **Estudo ABRE macroeconômico da embalagem e cadeia de consumo**. Abre, [mar. 2022]. Disponível em: https://www.abre.org.br/dados-do-setor/2021-2/. Acesso em: 22 nov. 2024.

AKAMINE, C. T.; YAMAMOTO, R. K. **Estatística descritiva**. 3. ed. São Paulo: Érica, 2014.

APS. Autoridade Portuária de Santos. **Movimento acumulado do Porto de Santos cresce 13,5% e atinge 57 milhões até abril**. APS, Santos, 22 maio 2024. Disponível em: https://www.portodesantos.com.br/2024/05/22/movimento-acumulado-do-porto-de-santos-cresce-135-e-atinge-57-milhoes-ate-abril/. Acesso em: 29 set. 2024.

APS. **Movimento acumulado do Porto de Santos cresce 13,5% e atinge 57 milhões até abril**. Porto de Santos: Autoridade Portuária, São Paulo, 22 maio 2024. Disponível em: https://www.portodesantos.com.br/2024/05/22/movimento-acumulado-do-porto-de-santos-cresce-135-e-atinge-57-milhoes-ate-abril/. Acesso em: 03 set. 2024.

APS. **Movimento de cargas no Porto de Santos cresce 29,4% em novembro e aponta para novo recorde de movimentação em 2023**. Porto de Santos: Autoridade Portuária, São Paulo, 26 dez. 2023. Disponível em: https://www.portodesantos.com.br/2023/12/26/movimento-de-cargas-no-porto-de-santos-cresce-294-em-novembro-e-aponta-para-novo-recorde-de-movimentacao-em-2023/. Acesso em: 03 set. 2024.

APS. **Porto de Santos – Área de Influência**. Porto de Santos: Autoridade Portuária, 2024. Disponível em: https://www.portodesantos.com.br/conheca-o-porto/area-de-influencia-2/. Acesso em: 28 set. 2024.

BRASILEIRO, A. M. M. **Como produzir textos acadêmicos e científicos**. 1. ed. São Paulo: Contexto, 2022.

CANAL RURAL. **Guerra da Ucrânia: tensão faz preço do milho disparar**. 2023. Disponível em: https://www.canalrural.com.br/agricultura/guerra-da-ucrania-tensao-faz-preco-do-milho-disparar/. Acesso em: 27 set. 2024.

CHADE, J. **Sob embargo, Rússia quase dobra exportações ao Brasil e negocia diesel**. UOL, [s.l.], 08 ago. 2022. Disponível em: https://noticias.uol.com.br/colunas/jamil-chade/2022/08/08/sob-embargo-russia-quase-dobra-exportacoes-ao-brasil-e-negocia-diesel.htm. Acesso em: 20 set. 2024.

CNI. **Nota econômica 24: Guerra na Ucrânia afeta o preço das importações no Brasil**. Portal da Indústria – Confederação Nacional da Indústria, São Paulo, [nov. 2022]. Disponível em: https://www.portaldaindustria.com.br/publicacoes/2022/11/nota-economica-24-guerra-na-ucrania-afeta-o-preco-das-importacoes-no-brasil/. Acesso em: 03 set. 2024.

CNN BRASIL. **Brasil depende de 5,5 milhões de toneladas de trigo da Ucrânia, diz ministro da Agricultura à CNN**. CNN Brasil, [s.l.], 2023. Disponível em: https://www.cnnbrasil.com.br/economia/macroeconomia/brasil-depende-de-55-milhoes-de-toneladas-de-trigo-da-ucrania-diz-ministro-da-agricultura-a-cnn/. Acesso em: 27 set. 2024.

CNN BRASIL. **Brasil, China e outros 11 países propõem acordo de paz entre Ucrânia e Rússia**. CNN Brasil, [s.l.], 2024. Disponível em: https://www.cnnbrasil.com.br/internacional/brasil-china-e-outros-11-paises-propoem-acordo-de-paz-entre-ucrania-e-russia/. Acesso em: 28 set. 2024.

CNN BRASIL. **Guerra pode "alterar fundamentalmente" ordem econômica e política globais, diz FMI**. CNN Brasil, Londres, 16 mar. 2022. Disponível em: https://www.cnnbrasil.com.br/economia/macroeconomia/guerra-pode-alterar-fundamentalmente-ordem-economica-e-politica-globais-diz-fmi/. Acesso em: 27 set. 2024.

COMEXSTAT. **MDIC – ComexStat**. 2024. Disponível em: https://comexstat.mdic.gov.br/pt/comex-vis/2/831. Acesso em: 27 set. 2024.

CORRÊIA, G. **80% das mercadorias no mundo são transportadas em navios, diz ONU**. Rádio Agência, Maranhão, 03 jan. 2023. Disponível em: https://agenciabrasil.ebc.com.br/radioagencia-nacional/economia/audio/2023-01/80-das-mercadorias-no-mundo-sao-transportadas-em-navios-diz-onu. Acesso em: 03 set. 2024.

EBC. **Setor portuário registra crescimento de 10% no primeiro bimestre**. Agência Gov, [s.l.], 11 abr. 2024. Disponível em: https://agenciagov.ebc.com.br/noticias/202404/setor-portuario-registra-crescimento-de-10-no-primeiro-bimestre. Acesso em: 26 set. 2024.

FONTES, S. **Brasil importa cada vez mais resina plástica**. Valor Econômico, [s.l.], 01 fev. 2022. Disponível em: https://valor.globo.com/empresas/noticia/2022/02/01/brasil-importa-cada-vez-mais-resina-plastica.ghtml. Acesso em: 22 nov. 2024.

G1 ECONOMIA. **Com economia aquecida, transporte de carga pelos portos brasileiros bate recorde no 1º semestre**. 2024. Disponível em: https://g1.globo.com/economia/noticia/2024/08/07/com-economia-aquecida-transporte-de-carga-pelos-portos-brasileiros-bate-recorde-no-1o-semestre.ghtml. Acesso em: 27 set. 2024.

G1 SANTOS. **Baixada Santista é responsável por quase 12% das exportações no estado de São Paulo em 2023, aponta Seade**. G1, Santos, 19 maio 2024. Disponível em: https://g1.globo.com/sp/santos-regiao/noticia/2024/05/19/baixada-santista-e-responsavel-por-quase-12percent-das-exportacoes-no-estado-de-sao-paulo-em-2023-aponta-seade.ghtml. Acesso em: 29 set. 2024.

G1 SANTOS. **Guerra na Ucrânia pode causar impactos no Porto de Santos, diz especialista**. G1, 2022. Disponível em: https://g1.globo.com/sp/santos-regiao/porto-mar/noticia/2022/02/25/ataque-da-russia-a-ucrania-pode-causar-impactos-muito-fortes-para-a-economia-brasileira-diz-especialista.ghtml. Acesso em: 24 set. 2024.

G1. **Em resultado parcial de referendo, Rússia diz que ampla maioria apoia anexação de regiões ucranianas**. G1, [s.l.], 27 set. 2022. Disponível em: https://g1.globo.com/mundo/ucrania-russia/noticia/2022/09/27/esperado-resultado-parcial-dos-referendos-regioes-ucranianas-indicam-anexacao-a-russia.ghtml. Acesso em: 24 set. 2024.

GOVBR. **Comércio exterior brasileiro bate recordes e fecha 2023 com saldo de US$ 98,8 bi**. GOV.BR, [s.l.], 11 jan. 2024. Disponível em: https://www.gov.br/mdic/pt-br/assuntos/noticias/2024/janeiro/comercio-exterior-brasileiro-bate-recordes-e-fecha-2023-com-saldo-de-us-98-8-bi. Acesso em: 03 set. 2024.

GRIECO, F. de A. **O Brasil e a nova geopolítica europeia**. 1. ed. São Paulo: Edições Aduaneiras, 1992.

LIBREOFFICE. **Calc**. LibreOffice, [s.l.], 2024. Disponível em: https://pt-br.libreoffice.org/descubra/calc/. Acesso em: 20 set. 2024.

LOUISE, F. **Conflito entre Rússia e Ucrânia afeta preço das importações no Brasil**. Portal da Indústria, [s.l.], 29 nov. 2022. Disponível em: https://noticias.portaldaindustria.com.br/noticias/internacional/conflito-entre-russia-e-ucrania-afeta-preco-das-importacoes-no-brasil/. Acesso em: 03 set. 2024.

MARSHALL, T. **Prisioneiros da geografia: 10 mapas que explicam tudo o que você precisa saber sobre política global**. 1. ed. Rio de Janeiro: Zahar, 2018.

McKINNEY, W. **Python para análise de dados: tratamento de dados com Pandas, NumPy e Jupyter**. 3. ed. São Paulo: Novatec, 2023.

MDIC. **Comércio exterior brasileiro bate recordes e fecha 2023 com saldo de US$ 98,8 bi**. GOVBR, [s.l.], 05 jan. 2024. Disponível em: https://www.gov.br/mdic/pt-br/assuntos/noticias/2024/janeiro/comercio-exterior-brasileiro-bate-recordes-e-fecha-2023-com-saldo-de-us-98-8-bi. Acesso em: 20 set. 2024.

MEDEIROS, J. B. **Redação científica**. 12. ed. São Paulo: Atlas, 2014. p. 163–167.
MICROSOFT. **Introdução a *dashboards* para designers do Power BI**. Microsoft, [s.l.], 23 nov. 2023. Disponível em: https://learn.microsoft.com/pt-br/power-bi/create-reports/service-dashboards. Acesso em: 20 set. 2024.

MICROSOFT. **O que é Power BI?** Microsoft, [s.l.], 22 mar. 2024. Disponível em: https://learn.microsoft.com/pt-br/power-bi/fundamentals/power-bi-overview. Acesso em: 20 set. 2024.
MINISTÉRIO DAS RELAÇÕES EXTERIORES. **Ucrânia**. GOVBR, [s.l.], 2024. Disponível em: https://www.gov.br/mre/pt-br/assuntos/relacoes-bilaterais/todos-os-paises/ucrania. Acesso em: 27 set. 2024.

NAÇÕES UNIDAS BRASIL. **Comércio global retomou crescimento no primeiro trimestre de 2024**. Nações Unidas Brasil, [s.l.], 2 jul. 2024. Disponível em: https://news.un.org/pt/story/2024/07/1833936. Acesso em: 21 nov. 2024.

NAÇÕES UNIDAS BRASIL. **UNCTAD publica relatório com impactos da guerra da Ucrânia na economia global**. Nações Unidas Brasil, [s.l.], 18 mar. 2022. Disponível em: https://brasil.un.org/pt-br/175249-unctad-publica-relat%C3%B3rio-com-impactos-da-guerra-da-ucr%C3%A2nia-na-economia-global. Acesso em: 03 set. 2024.

PANDAS. ***About Pandas: history of development***. Pandas, [s.l.], 2024. Disponível em: https://pandas.pydata.org/about/. Acesso em: 20 set. 2024.

PEDRO, J. D. L. G. **Perspectivas atuais de guerra Rússia-Ucrânia**: os motivos profundos do conflito. FCHS (DCPC) – Ciência Política e Relações Internacionais, 2023. Disponível em: http://hdl.handle.net/10284/13032. Acesso em: 24 set. 2024.

PYTHON. ***What is Python***. Python.org, [s.l.], 19 set. 2024. Disponível em: https://docs.python.org/3/faq/general.html#what-is-python. Acesso em: 20 set. 2024.

ROSÁRIO, M. E. G. R. do. **Influência da guerra da Ucrânia nas relações comerciais e diplomáticas entre Brasil e Rússia**. Trabalho de Conclusão de Curso (Graduação em Direito) – Universidade Presbiteriana Mackenzie. 2022. Disponível em: https://adelpha-api.mackenzie.br/server/api/core/bitstreams/2114afad-db93-434d-8b68-ad20dd4df8de/content. Acesso em: 20 set. 2024.

SANT'ANA, J. **Amendoim, adubo, máquinas: veja os principais produtos do comércio do Brasil com Rússia e Ucrânia**. G1, Brasília, 01 mar. 2022. Disponível em: https://g1.globo.com/economia/noticia/2022/03/01/amendoim-adubo-maquinas-veja-os-principais-produtos-do-comercio-do-brasil-com-russia-e-ucrania.ghtml. Acesso em: 7 set. 2024.

SANTIMARIA, J. P. M. **Impacto da guerra Rússia/Ucrânia sobre o mercado de fertilizantes brasileiro**. UFSCar, Araras, 2023. Disponível em: https://repositorio.ufscar.br/handle/ufscar/18721. Acesso em: 24 set. 2024.

VALOR ECONÔMICO. **Brasil eleva importação da Rússia com compras de diesel e fertilizantes**. Globo, 2024. Disponível em: https://valor.globo.com/brasil/noticia/2024/07/12/brasil-eleva-importacao-da-russia-com-compras-de-diesel-e-fertilizantes.ghtml. Acesso em: 27 set. 2024.

VALOR ECONÔMICO. **Em conversa com Bolsonaro, Putin promete fornecimento ininterrupto de fertilizantes, diz Kremlin**. Globo, 2022. Disponível em: https://valor.globo.com/politica/noticia/2022/06/27/em-conversa-com-bolsonaro-putin-promete-fornecimento-ininterrupto-de-fertilizantes-diz-kremlin.ghtml. Acesso em: 27 set. 2024.

VALOR ECONÔMICO. **Queda nas importações explica maior parte do avanço do superávit comercial em 2023, diz AEB**. Valor Econômico, [s.l.], 05 jan. 2024. Disponível em: https://valor.globo.com/brasil/noticia/2024/01/05/queda-nas-importacoes-explica-maior-parte-do-avanco-do-superavit-comercial-em-2023-diz-aeb.ghtml. Acesso em: 25 set. 2024.

Um estudo comparativo no crescimento do índice de desenvolvimento humano do município de Santos em relação ao município de Praia Grande

A comparative study on the growth of the human development index of the municipality of Santos in relation to the municipality of Praia Grande

Davi Vitor Silva
Fatec Baixada Santista – Rubens Lara
davi.silva90@fatec.sp.gov.br

Rodrigo Cavalheiro dos Santos
Fatec Baixada Santista – Rubens Lara
rodrigo.santos281@fatec.sp.gov.br

Claudia Maria Sodero Salles
Fatec Baixada Santista – Rubens Lara
claudia.sodero@cps.sp.gov.br

Revista Datapoint
eISSN 3086-433X
Faculdade de Tecnologia Rubens Lara – FATEC
Ciência de Dados
Peridicidade: Anual
Vol 01, n. 01, 2025
revistadp@fatecrl.edu.br

Recebido: Jun 2025
Aceito: Set 2025
Publicado: Dez 2025

URL: https://www.fatecrl.edu.br/revista/datapoint/index.php/dp/article/view/3
DOI: https://doi.org/10.5281/zenodo.19240531

RESUMO

Este estudo comparativo analisa o crescimento do Índice de Desenvolvimento Humano (IDH) de Santos e Praia Grande (PG) em 2000, 2010 e 2022. O objetivo foi verificar se Praia Grande apresenta uma evolução mais rápida, analisando o IDH e seus componentes: saúde, renda e educação. A metodologia utilizou técnicas de ciência de dados (ETL, análise) e modelos de cálculo do PNUD, com dados de fontes oficiais como IBGE e Seade. Devido à ausência de dados consolidados para 2022, foi realizada uma simulação não oficial. Os resultados simulados para 2022 apontam um IDH de 0,896 para Santos (saúde 0,906; educação 0,875; renda 0,907) e um IDH de 0,839 para Praia Grande (saúde 0,905; educação 0,832; renda 0,780). A análise comparativa do crescimento entre 2010 e 2022 mostrou que Praia Grande apresentou um desenvolvimento superior em quase todos os indicadores. O IDH total de PG cresceu 11%, contra 7% de Santos. Na educação, Praia Grande avançou 20% (contra 8% de Santos), e na saúde, cresceu 9% (contra 6% de Santos). O índice de renda foi o único com crescimento equivalente (cerca de 5% em ambos). Os dados indicam que Praia Grande está se desenvolvendo em um ritmo mais acelerado que Santos, especialmente na área da Educação.

PALAVRAS-CHAVE: índice de desenvolvimento humano; expectativa de vida; Baixada Santista; ciência de dados.

ABSTRACT

This comparative study analyzes the growth of the Human Development Index (HDI) of Santos and Praia Grande (PG) in 2000, 2010, and 2022. The objective was to verify if Praia Grande presents a faster evolution, analyzing the HDI and its components: health, income, and education. The methodology used data science techniques (ETL, analysis) and UNDP calculation models, with data from official sources such as IBGE and Seade. Due to the absence of consolidated data for 2022, an unofficial simulation was performed. The simulated results for 2022 indicate an HDI of 0.896 for Santos (health 0.906; education 0.875; income 0.907) and an HDI of 0.839 for Praia Grande (health 0.905; education 0.832; income 0.780). A comparative analysis of growth between 2010 and 2022 showed that Praia Grande exhibited superior development in almost all indicators. Praia Grande's total HDI grew by 11%, compared to 7% for Santos. In education, Praia Grande advanced 20% (compared to 8% for Santos), and in health, it grew by 9% (compared to 6% for Santos). The income index was the only one with equivalent growth (around 5% in both). The data indicate that Praia Grande is developing at a faster pace than Santos, especially in the area of education.

KEY-WORDS: *human development index; life expectancy; Baixada Santista; data science.*

INTRODUÇÃO

O município de Santos, situado no litoral de São Paulo, foi por muitos anos considerada a principal cidade da Região Metropolitana da Baixada Santista, devido ao seu grande papel relevante na economia da região pela atividade portuária no município (MPSP, 2023 p. 4), sendo por muito tempo a cidade com o maior IDH da região. No entanto, o município de Praia Grande, localizado na mesma região, vem apresentando um crescimento constante em seu IDH, tendo apresentado um recente crescimento populacional em cerca de 33% conforme o Ministério Público de São Paulo (MPSP, 2023, p. 4). Considerando essa crescente do indicador do município de Praia Grande, é de grande interesse a comparação do mesmo em relação ao município de Santos, para se observar a velocidade e volume desse crescimento e entender seus principais aspectos.

Assim, esse trabalho investiga e compara o crescimento do IDH tanto de Santos, quanto de Praia Grande no período de 2000, 2010 e 2022, e destaca os componentes do IDH no processo de evolução temporal e aqueles que têm um potencial de crescimento nos respectivos municípios. A pesquisa se concentra no estudo, projeção e análise dos indicadores componentes do IDH, sendo eles: índice de saúde (expectativa de vida), renda e educação, buscando compreender o crescimento do IDH de Santos e Praia Grande, comparando-os.

A análise desses índices foi feita por meio da aplicação de métodos relacionados à ciência de dados, aplicando modelos de cálculos utilizados na formulação do IDH de acordo com o Relatório do Desenvolvimento Humano 2006 (PNUD, 2006, p. 520), com uso das tábuas de vida do Instituto Brasileiro de Geografia e Estatística (IBGE), juntamente com dados oficiais da plataforma Seade, Programa das Nações Unidas (PNUD) e Sistema IBGE de Recuperação Automática (SIDRA).

O objetivo geral da pesquisa consiste na elaboração de análises gráficas a partir dos dados obtidos e trabalhados no escopo da pesquisa. Os dados foram obtidos a partir das plataformas: Sistema Estadual de Análise de Dados (Seade), IBGE e Governo Federal, SIDRA e Seade. Para este estudo, foram selecionados dados dos municípios de Santos e Praia Grande no período de 2000, 2010 e 2022. Para essa finalidade, destacam os seguintes os objetivos específicos: (i) selecionar dados fornecidos pelas plataformas e considerados relevantes para a pesquisa; (ii) realizar o tratamento dos dados obtidos; (iii) aplicar os dados examinados aos cálculos oficiais para obtenção dos índices (saúde, renda e educação) dos municípios de Santos e Praia Grande no ano de 2022, e assim obtendo uma simulação de um IDH recente não oficial;

(iv) elaborar as análises gráficas, examinar e comparar o crescimento dos índices e do IDH entre os municípios de Santos e Praia Grande.

1. FUNDAMENTAÇÃO TEÓRICA

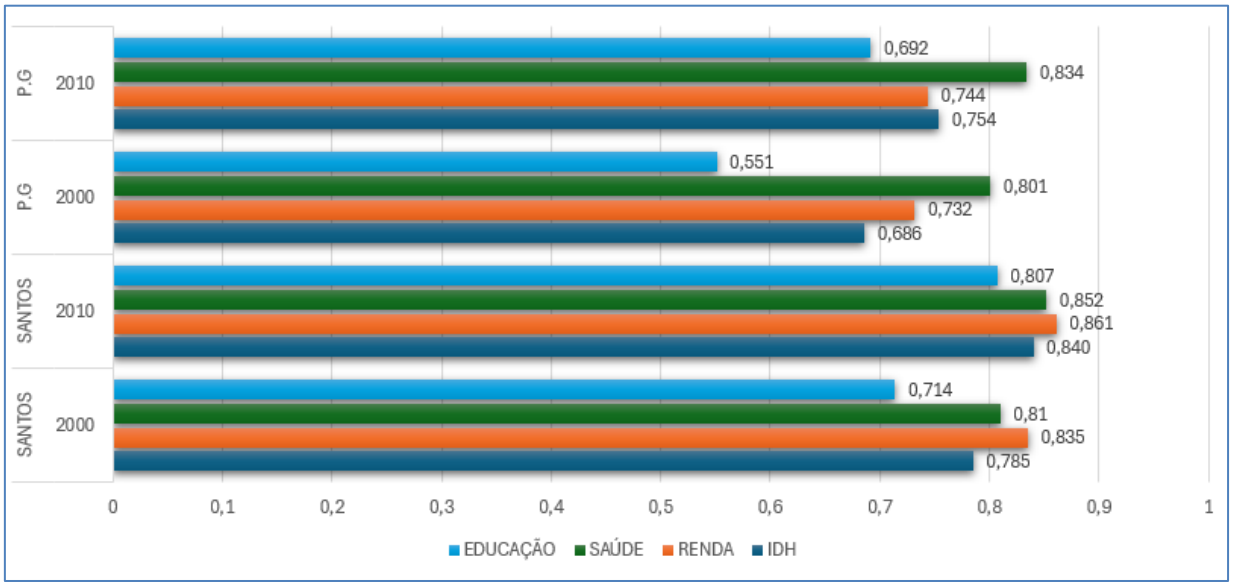
O estudo abordará métodos de cálculo do IDH conforme o Relatório de Desenvolvimento Humano do PNUD, que engloba os índices: saúde (expectativa de vida), educação (média de anos de escolaridade) e renda (PIB per capita em PPC).

1.1 A DINÂMICA DA BAIXADA SANTISTA E A SELEÇÃO DOS MUNICÍPIOS

A Baixada Santista é uma região litorânea do estado de São Paulo, composta por nove municípios interligados por relações econômicas, sociais e urbanas. Dentre eles, destaca-se o município de Santos como centro regional, concentrando infraestrutura em saúde, educação e serviços, além de apresentar o maior IDH da região (MPSP, 2023, p. 4).

O Gráfico 1 ilustra o crescimento de cada indicador que compõe o IDH do município de Santos e de Praia Grande no período de 2000 e 2010. Nele, observa-se que o município de Praia Grande apresentou indicadores inferiores ao município de Santos em relação ao IDH.

Gráfico 1 – Valores dos índices do IDH de Santos e Praia Grande nos anos de 2000 e 2010



Fonte: elaborado pelos autores com dados de Atlas Brasil (2025)

É possível observar também um crescimento significativo no índice de saúde tanto de Santos quanto de Praia Grande, o que já foi observado por Ministério Público de São Paulo (MPSP, 2023, p. 4).

1.2 DADOS SECUNDÁRIOS, ETL E TECNOLOGIAS DE ANÁLISE DE DADOS

Para conduzir a análise proposta nesta pesquisa, foram utilizadas técnicas de ETL, que possibilitam a coleta, limpeza e integração dos dados de diversas fontes. Segundo Kimball e Ross (2013, p. 12), o processo de ETL é fundamental para a preparação de dados em projetos de *Business Intelligence* (BI), pois permite transformar dados brutos em informações úteis para tomada de decisões.

O processo de transformação dos dados foi realizado utilizando-se o Excel e a *Integrated Development Environment* (IDE) Spyder, com o uso da linguagem de programação Python e sua biblioteca Pandas, considerando a necessidade de lidar com um grande volume de dados.

A simulação dos índices e do IDH para o ano de 2022, foi feita utilizando fórmulas adaptadas do PNUD (2006, p. 394), levando em consideração os componentes de saúde, educação e renda. O uso das tábuas de vida do IBGE permitiu calcular a expectativa de vida.

A aquisição dos dados usados tornou possível o cálculo dos índices e a simulação do IDH mais recente dos municípios. Os índices calculados foram: expectativa de vida (saúde), PIB per capita (renda) e grau de instrução (educação). Os dados para calcular o índice de saúde foram obtidos na plataforma Seade, os dados de renda foram obtidos na plataforma do IBGE e os dados relativos à educação foram obtidos nas plataformas SIDRA e Seade, todos para o ano de 2022, com exceção de renda, que os dados mais recentes eram de 2021, mas foram tratados como se fossem de 2022 neste estudo.

1.3 O CÁLCULO DO IDH

Conforme o Relatório do Desenvolvimento Humano de 2006 (PNUD, 2006, p. 394), o Índice de Desenvolvimento Humano (IDH) é calculado como a média aritmética simples de três índices de dimensão:

1) Índice da Saúde (Expectativa de vida)
2) Índice do Grau de Instrução (Educação)
3) Índice do PIB (Renda)

Assim, a fórmula aplicada para o cálculo do IDH foi:

$$\boldsymbol{IDH} = \frac{1}{3} \times (\textit{índice de saúde}) + \frac{1}{3} \times (\textit{índice de educação}) + \frac{1}{3} \times (\textit{índice de renda}) \quad (1)$$

2. PROCEDIMENTOS METODOLÓGICOS

Como já destacado, a seleção dos dados foi realizada a partir das informações disponibilizadas através dos dados abertos em diversas plataformas públicas. Os dados utilizados foram: (i) estimativa da população; (ii) óbitos; (iii) PIB per capita; (iv) fator de conversão em paridade de poder de compra (PPC); (v) matriculados nas modalidades de ensino; (vi) taxa de alfabetização.

2.1 METODOLOGIA DO CÁLCULO DO ÍNDICE DE SAÚDE

Após a obtenção dos dados necessários (população e óbitos) foram aplicados cálculos para a obtenção do valor do índice de saúde de acordo com o livro Tábuas Abreviadas de Mortalidade por Sexo e Idade (IBGE, 2010, p. 15). A expectativa de vida ao nascer (e0) representa o número médio de anos que um indivíduo de determinada população viveria, assumindo que os padrões de mortalidade observados no momento do cálculo permaneçam constantes. Seu cálculo é realizado por meio da tábua de vida, um modelo estatístico-demográfico que sintetiza as taxas de mortalidade por idade.

Para municípios, o método utilizado abreviado do cálculo da expectativa de vida é mais adequado por conta da forma como os dados foram disponibilizados (por faixa etária). O método completo exigiria dados que não estão disponíveis no formato adequado. Portanto, a escolha do método abreviado não só é justificável, como correta dentro da realidade dos dados municipais. A metodologia utilizada para calcular a expectativa de vida de forma abreviada em base nas diretrizes das Tábuas Abreviadas de Mortalidade publicadas pelo IBGE (2013.

O primeiro passo consiste na obtenção de dados populacionais e de mortalidade geralmente fornecidos por: censos demográficos; registros civis (óbitos por idade); estatísticas vitais (nascimentos e mortes).

2.1.1 Construção Da Tábua De Vida

A tábua de vida é composta por colunas que descrevem a mortalidade por faixa etária. Suas principais variáveis estão demonstradas no Quadro 1.

Quadro 1 – Variáveis da Tábua de Vida

Símbolo	Variável	Descrição
q_x	Probabilidade de morte	Chance de um indivíduo de idade x morrer antes de $x+1$
l_x	Sobreviventes	Número de pessoas que atingem a idade x
d_x	Óbitos	Número de mortes entre x e $x+1$
L_x	Anos vividos	Total de anos vividos pela coorte entre x e $x+1$
T_x	Anos totais restantes	Soma dos anos vividos a partir de x até a extinção
e_x	Expectativa de vida	Número médio de anos restantes para idade x

Fonte: Adaptado de IBGE (2023)

Essas variáveis são fundamentais para calcular indicadores demográficos como expectativa de vida e taxas de mortalidade, auxiliando nos estudos populacionais.

2.1.2 Cálculos De Cada Variável Da Tábua De Vida

A primeira variável a ser calculada é a taxa de mortalidade (probabilidade de morte), que mostra a chance de ocorrência de óbitos em cada faixa etária, expressa em percentual por período. Seu cálculo é feito pela relação entre o número de óbitos e a população do mesmo intervalo, conforme a fórmula a seguir:

$$\boldsymbol{Taxa\ de\ Mortalidade} = ({}^{Número\ de\ óbitos}/_{População\ da\ faixa}) \times 100. \quad (2)$$

A taxa de sobrevivência é utilizada para visualizar a probabilidade de sobreviver por sua respectiva faixa etária, representado de forma percentual por cada período. Seu cálculo é realizado a partir da diferença de 100% sobre a taxa de mortalidade apurada por cada intervalo. O cálculo pode ser expresso pela seguinte fórmula matemática:

$$\boldsymbol{Taxa\ de\ Sobrevivência} = 100\% - Taxa\ de\ Mortalidade \quad (3)$$

O cálculo da duração da faixa etária serve para visualizar o total de anos em que se viveu daquele determinado intervalo, considerando que todos fossem viver o tempo do intervalo de anos daquela faixa pela taxa de sobrevivência, obtemos o resultado de anos vividos, o cálculo pode ser expresso pela seguinte fórmula matemática:

$$\boldsymbol{Anos\ Vividos\ na\ Faixa} = Sobrevivência\ (decimal) \times Duração\ da\ Faixa\ (anos) \quad (4)$$

A aplicação do cálculo é repetida para todas as faixas etárias. A conclusão do cálculo se deu para cada um dos intervalos, podendo visualizar o quanto se viveu por cada período em relação a taxa de sobrevivência e prosseguir com o cálculo da expectativa de vida.

Para calcular a soma dos anos vividos, utilizada para visualizar a expectativa de vida em base na relação dos anos em que a população atual viveu, prevendo a sua qualidade de vida respectivamente, foram somados todos os anos vividos. A soma de todos os anos vividos estimados para cada faixa resulta na expectativa de vida ao nascer (e_0), conforme definido na metodologia abreviada (IBGE, 2013, p. 17).

2.1.3 Cálculo Do Índice De Saúde

O índice de saúde é um indicador que buscar medir a qualidade de vida da região, utilizando a expectativa de vida ao nascer como referência. Quanto maior a expectativa de vida, maior se presume o acesso da população a serviços de saúde essenciais, como atendimento médico, vacinas e medicamentos. O cálculo é apurado utilizando a expectativa de vida ao nascer daquela região utilizando a seguinte fórmula matemática:

$$\textbf{Índice de Saúde} = \frac{(\textbf{Expectativa de Vida} - \textbf{Valor Mínimo})}{(\textbf{Valor Máximo} - \textbf{Valor Mínimo})} \quad (5)$$

O índice da saúde pode variar entre 0 e 1, sendo quanto mais próximo de 1 melhor qualidade de vida.

2.2 METODOLOGIA DO ÍNDICE DE EDUCAÇÃO

O Índice de Educação (IE) é um indicador que busca representar o nível educacional de uma população, considerando tanto a taxa de alfabetização quanto a taxa bruta de escolarização em diferentes níveis de ensino. Segundo a metodologia apresentada no Relatório de Desenvolvimento Humano (PNUD, 2006, p. 520), O cálculo do índice usa um método ponderado que dá maior peso à taxa de alfabetização, por sua importância no desenvolvimento educacional, combinando-a à taxa bruta de escolarização para representar o acesso ao ensino. Esse índice é usado para avaliar o desenvolvimento educacional e compor o IDH.

2.2.1 Metodologia do cálculo da Taxa Bruta de Escolarização

A Taxa Bruta de Escolarização (TBE) é calculada como a relação entre o número total de matrículas nos níveis de ensino fundamental (E.F), médio (E.M) e superior (E.S) e a população da faixa etária teórica correspondente. De acordo com a metodologia apresentada pela Relatório do Desenvolvimento Humano (PNUD, 2006, p. 520), a TBE é obtida pela soma das matrículas dividida pela população da faixa etária de 5 a 24 anos, expressa pela fórmula:

$$\textbf{TBE} = \frac{(Matrículas\ no\ E.F + Matrículas\ no\ E.M + Matrículas\ no\ E.S)}{Pupulação\ de\ 05\ a\ 24\ anos} \quad (6)$$

A Taxa Bruta de Escolarização representa a quantidade em percentual dos matriculados em relação com a população do determinado intervalo da faixa etária.

2.2.1.1 Adequação da faixa etária no cálculo da Taxa Bruta de Escolarização

A TBE, quando calculada com base na população de 5 a 24 anos, costuma apresentar valores superiores a 100%. Isso ocorre porque o indicador considera todas as matrículas, incluindo alunos fora da idade adequada para cada nível de ensino, seja por atraso, avanço, ingresso tardio no ensino superior ou por residentes de outras cidades (EAD).

Isso pode ser observado claramente utilizando o caso do município de Santos em 2022 como exemplo. Utilizando os seus seguintes dados de 2022:

- População de 5 a 24 anos: 90.865.
- Matrículas no Ensino Fundamental: 44.435.
- Matrículas no Ensino Médio: 17.087
- Matrículas no Ensino Superior: 40.404.
-

Assim, aplicando a fórmula da TBE (6), obtém-se ≈1,1217:

O cálculo da TBE resultou em aproximadamente 112,17%, um valor que ultrapassa o limite teórico de 100%. Isso evidencia uma clara distorção causada pela faixa etária limitada, especialmente pelo impacto do número de matrículas no ensino superior, que inclui muitos alunos acima dos 24 anos, além de alunos de outros municípios (embora, por falta de dados, estes sejam considerados como um grupo único).

Conforme a metodologia descrita pelo PNUD (2006, p. 520), a TBE original utiliza a faixa de 5 a 24 anos como referência. No entanto, esse recorte etário não é capaz de contemplar precisamente a realidade do acesso ao ensino superior, resultando em distorções. Pesquisa publicada pela Universidade Federal de Uberlândia (Alvarenga, 2023) em base do relatório executivo da V Pesquisa Nacional de Perfil Socioeconômico e Cultural dos (as) Graduandos (as) das IFES (2018) afirmam que 28,9% dos estudantes do Sudeste possuem 25 anos ou mais, evidenciando que uma parcela significativa ingressa ou permanece no ensino superior em idades superiores à faixa tradicional.

Diante dessa diferença, optou-se por ampliar a faixa etária para 5 a 29 anos no cálculo da TBE. Isso permite uma melhor adequação da população de referência, reduzindo distorções e tornando o indicador mais compatível com a realidade educacional contemporânea. Com a nova estimativa de população sendo de 116.314 habitantes, e reutilizando os demais dados de Santos em 2022, obtém-se a partir do novo cálculo, um valor de 0,8763.

Com a ampliação da faixa etária para 5 a 29 anos, a TBE de Santos em 2022 ajusta-se de 112,17% para 87,63%, corrigindo a distorção causada pela presença de alunos fora da faixa de 5 a 24 anos, especialmente no ensino superior. Esse ajuste torna o indicador mais alinhado com a realidade educacional, refletindo de forma mais precisa a proporção de pessoas matriculadas em relação à população potencialmente envolvida no processo educacional.

2.2.1.2 Adequação da Filtragem das Matrículas do Ensino Fundamental

Ao calcular a TBE, é essencial a coerência entre matrículas e população residente. No caso de Santos, verificou-se que o total de matrículas no ensino fundamental (44.435) excede a população estimada de 42.918 pessoas na faixa de 5 a 14 anos, indicando uma distorção. Essa discrepância indica a possível presença de estudantes de outros municípios em escolas de Santos, especialmente nas redes estadual e privada. Como o denominador da TBE usa a população residente (PNUD, 2006), a inclusão de matrículas de não residentes infla a taxa, superestimando o acesso educacional local.

Diante dessa distorção, optou-se por restringir as matrículas do ensino fundamental àquelas da rede municipal, cuja cobertura está mais vinculada à população residente. Essa abordagem visa maior precisão, ajustando o numerador à realidade local e mitigando a distorção do fluxo intermunicipal. Portanto, para coerência metodológica e fidedignidade dos dados, a TBE recalculada utilizará o valor de 19.057 matrículas da rede municipal no ensino fundamental, ao invés do total original de 44.435.

Com a nova filtragem, que usa apenas as 19.057 matrículas da rede municipal, a TBE de Santos em 2022 passa a ser de aproximadamente 65,82%, este resultado é uma estimativa mais conservadora e ajustada à realidade local, pois reduz a distorção causada por alunos de outros municípios.

2.2.2 Cálculo do Índice de Escolaridade

O Índice de Escolaridade (IE) é calculado a partir da combinação ponderada da Taxa de Alfabetização de adultos e da Taxa Bruta de Escolarização, conforme a metodologia do Relatório do Desenvolvimento Humano (PNUD, 2006, p. 520). O cálculo do IE é dado como:

$$\boldsymbol{IE} = \left(\frac{2}{3}\right) \times Taxa\ de\ Alfabetização + \left(\frac{2}{3}\right) \times Taxa\ Bruta\ de\ Escolarização \quad (7)$$

Em que TA é a Taxa de Alfabetização em decimal e TBE é a Taxa Bruta de Escolarização em decimal.

2.3 METODOLOGIA DO ÍNDICE DE RENDA

O Índice de Renda (IR) é o terceiro dos três componentes do IDH, utilizado para mensurar o padrão de vida de uma população com base em sua renda per capita, ajustada pela paridade do poder de compra (PPC). Como será voltada para município, é utilizado apenas o Produto Interno Bruto (PIB) per capita municipal e o fator de conversão do Real para PPC.

A metodologia apresentada segue o modelo de cálculo do IDH adotado pelo PNUD de 2006, que por usar exclusivamente o PIB per capita como descrito no Relatório de Desenvolvimento Humano 2006 (PNUD, 2006, p. 520), o cálculo não é preciso ser ajustado para renda domiciliar per capita (como no modelo de 2010 do PNUD).

2.3.1 Obtenção da Renda Per Capita (RPC) em PPC

A renda per capita utilizada é o PIB per capita municipal em reais, divulgado pelo IBGE. Para possibilitar comparações internacionais, a renda é convertida de reais (R$) para dólares PPC, usando a taxa de conversão da tabela publicada pelo Ministério da Ciência, Tecnologia e Inovação (Brasil, 2023). Em 2022, essa taxa foi de aproximadamente 2,413. A conversão foi feita da seguinte forma:

$$\boldsymbol{RPC\ (USD\ PPC)} = {Renda\ per\ Capita\ (R\$)}\Big/{Taxa\ de\ Conversão\ (PPC)} \quad (8)$$

Assim, a conversão para dólares PPC permite avaliar com maior precisão o padrão de vida da população em relação a outras regiões do mundo.

2.3.2 Cálculo Do Índice De Renda

O Índice de Renda é calculado com base na renda per capita ajustada em PPC seguindo a metodologia do Relatório do Desenvolvimento Humano (PNUD, 2006, p. 520). A fórmula para o cálculo do Índice de Renda é a seguinte:

$$\boldsymbol{IR} = {(\log(RPC) - \log(100))}\Big/{(\log(40.000) - \log(100))} \quad (9)$$

Em que RPC é a renda per capita em dólares PPC. O valor mínimo (100) e o valor máximo (40.000) divulgado no Relatório de Desenvolvimento Humano (PNUD, 2006, p. 520), são os limites utilizados internacionalmente para padronizar a escala do indicador.

2.4 METODOLOGIA DO IDH SIMULADO

A composição do IDH final parte do princípio de que o desenvolvimento de um país não pode ser medido apenas pelo crescimento econômico. Conforme detalhado no Relatório do Desenvolvimento Humano de 2006 do PNUD, o cálculo do IDH é realizado através da média aritmética simples de três índices que representam dimensões fundamentais do desenvolvimento humano. As três dimensões consideradas, cada uma com peso igual na composição do IDH final, são:

1) índice de saúde: Refletida pela expectativa de vida ao nascer
2) índice de educação: Medida pelo grau de instrução da população
3) índice de renda: Avaliada pelo PIB per capita ajustado

Cada uma dessas dimensões é primeiramente convertida em um índice que varia de 0 a 1. O IDH final é, então, o resultado da média desses três índices, conforme expresso na fórmula:

$$\boldsymbol{IDH} = \frac{1}{3} \times (\textit{índice de saúde}) + \frac{1}{3} \times (\textit{índice de educação}) + \frac{1}{3} \times (\textit{índice de renda}) \quad (1)$$

Este método garante que os três índices tenham a mesma importância na avaliação do nível de desenvolvimento de um país ou região.

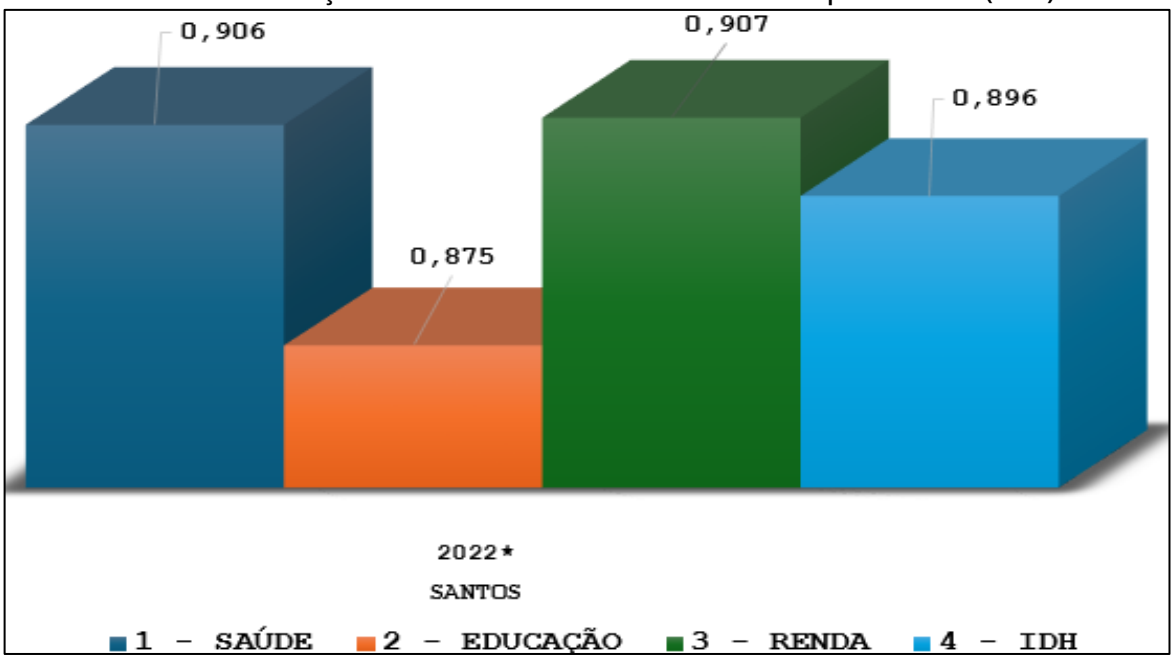
2.4.1 Cálculo final do IDH para Santos (2022)

Para o cálculo do IDH final simulado de Santos em 2022, foram utilizados os resultados obtidos para os índices componentes:

- Índice de Renda: 0,907
- Índice de Saúde: 0,906
- Índice de Educação: 0,875

O Gráfico 2 apresenta o resultado da simulação do IDH para o município de Santos, calculado em 0,896:

Gráfico 2 – Simulação dos valores dos índices do IDH do município de Santos (2022)



Fonte: Elaborado pelos autores (2025)

Um IDH de 0,896 classifica o município de Santos na faixa de "Desenvolvimento Humano Muito Alto" (IDH ≥ 0,800). Este resultado simulado sugere um nível de desenvolvimento humano bastante elevado, refletindo os altos desempenhos nos componentes de longevidade, educação e no índice de renda simulado.

2.4.2 Cálculo Final do IDH para Praia Grande (2022)

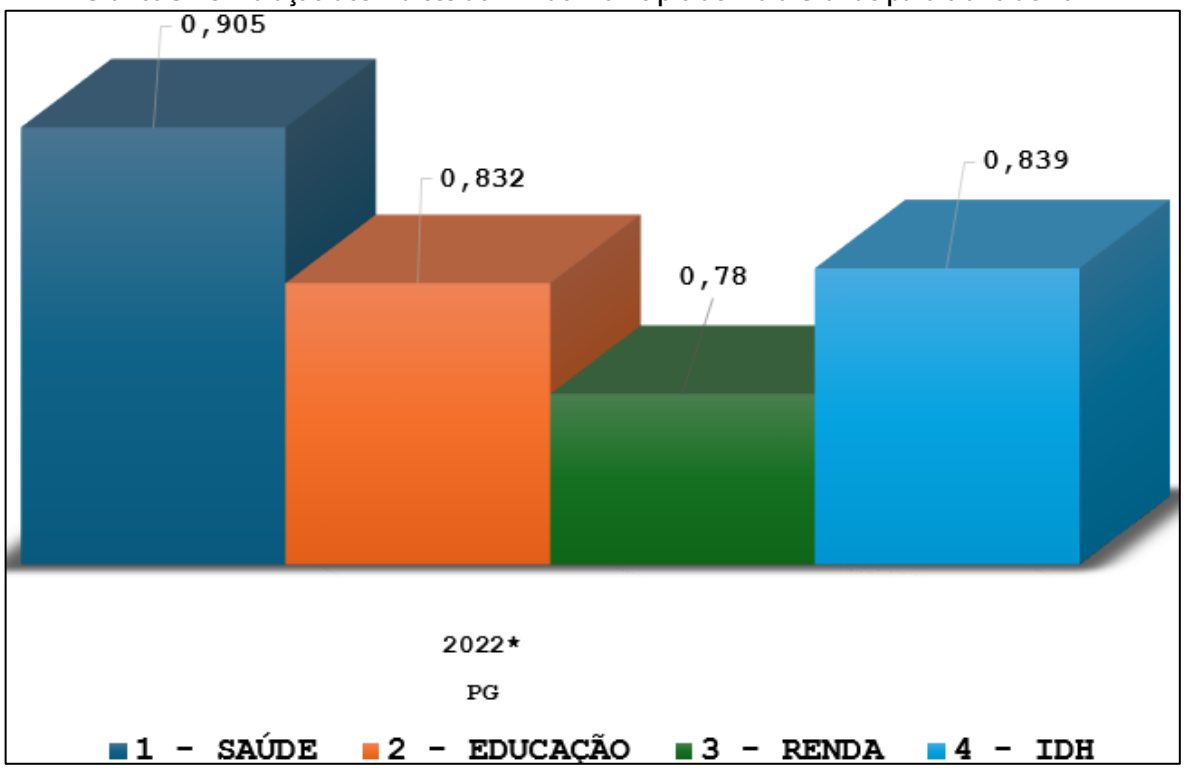
Para o cálculo do IDH final simulado de Praia Grande em 2022, foram utilizados os resultados obtidos para os índices componentes:

- Índice de saúde = 0,905;
- Índice de educação = 0,832;
- Índice de renda = 0,78;

O Gráfico 3 apresenta o resultado da simulação do IDH para o município de Praia Grande, calculado em 0,839:

No gráfico 3, é possível se observar os valores totais da simulação de cada índice que compõe o IDH de um município, nesse caso, o de Praia Grande para o ano de 2022, que apresenta valores considerados "Desenvolvimento Humano Muito Alto" (IDH ≥ 0,800) segundo o Relatório de Desenvolvimento Humano 2006 (PNUD, 2006, p. 520).

Gráfico 3 – Simulação dos índices do IDH do município de Praia Grande para o ano de 2022



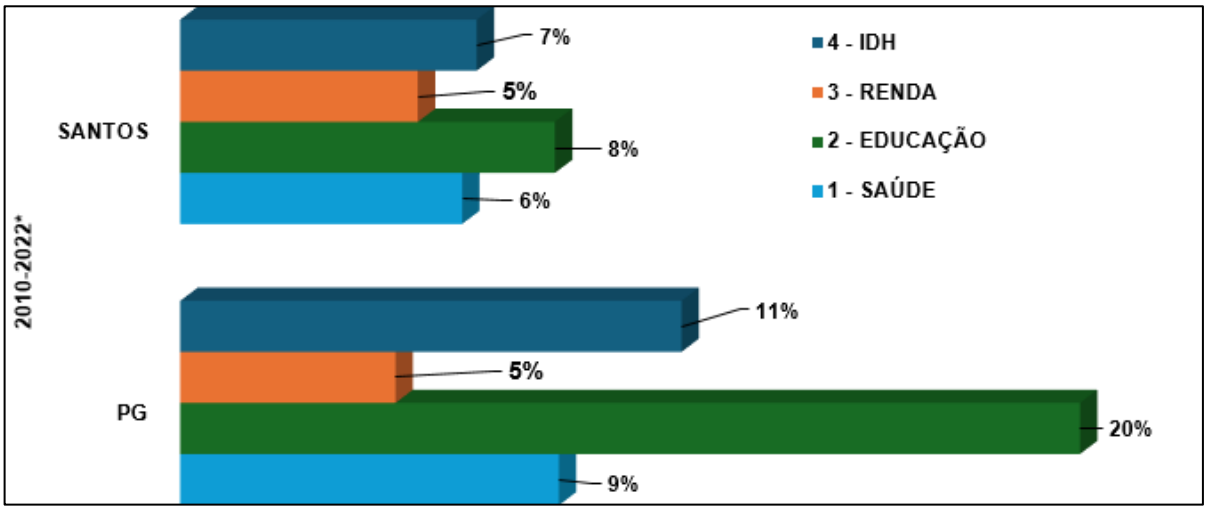
Fonte: Elaborado pelos autores (2025)

3. RESULTADOS E DISCUSSÕES

É importante frisar que este valor do IDH é uma simulação baseada nos índices de renda, saúde e educação que também foram simulados de acordo com os dados fornecidos. O IDHM oficial para os municípios, quando divulgado com base nos dados do Censo 2022 poderá apresentar um valor diferente. A metodologia padrão do IDHM brasileiro, por exemplo, utiliza a média geométrica para agregar os componentes.

Assim, o Gráfico 4 apresenta a evolução do valor de cada índice de cada município analisado nos anos de 2000, 2010 e a simulação de 2022, elaborado para observar e comparar a porcentagem de crescimento de cada índice a cada década, conforme o objetivo da pesquisa. O Gráfico 4 permite comparar o crescimento do IDH de Santos e Praia Grande.

Gráfico 4 – Percentual do aumento dos índices de Santos e Praia Grande (2010 e 2022)



Fonte: Elaborado pelos autores (2025)

Para elaboração de uma análise de um cenário atual, é comparado o crescimento apenas do ano de 2010 a 2022, onde é observado que em todos os índices, o município de Praia Grande apresenta um crescimento maior em comparação ao município de Santos, exceto no índice de renda, que apresenta um crescimento praticamente equivalente a ambos os municípios.

O índice de saúde cresceu cerca de 8,56%, enquanto Santos 6,34%, o de educação 20,23% enquanto Santos 8,43%, o IDH 11,27% e Santos 6,67%. Essa análise responde ao questionamento inicial, se o município de Praia Grande estava realmente crescendo em um ritmo maior ao município de Santos. O município de Praia Grande assume que esse crescimento siga pelas próximas décadas.

Nota-se claramente que o índice ligado à educação, na Praia Grande é a mola propulsora do IDH simulado na cidade. Isso rende um *insight* ótimo para que sejam desenvolvidos novos estudos e pesquisas mais detalhados.

4. CONSIDERAÇÕES FINAIS

Este estudo concluiu que, no período de 2010 a 2022, o município de Praia Grande apresentou um crescimento do IDH significativamente mais acelerado que o de Santos, confirmando a hipótese inicial da pesquisa.

A análise comparativa revelou que Praia Grande superou Santos em quase todos os componentes, com destaque para a Educação, que cresceu 20,23% (contra 8,43% de Santos), e a Saúde (8,56% contra 6,34%). O Índice de Renda foi o único com crescimento equivalente em ambas as cidades. O forte avanço na Educação é apontado como um possível fator para o aumento populacional em Praia Grande, sugerindo uma nova linha de pesquisa sobre a procura por residências no município.

Como principais limitações, é destacado que os dados de 2022 são uma simulação não oficial, criada devido à ausência de dados consolidados do IBGE. Além disso, a metodologia de cálculo utilizada (média aritmética) difere da média geométrica, que é o padrão oficial do IDHM brasileiro.

Para trabalhos futuros, sugere-se a validação destes resultados simulados assim que os dados oficiais de 2022 forem divulgados. Recomenda-se também a investigação de outros fatores socioeconômicos.

REFERÊNCIAS

ALVARENGA, C. **Pesquisa revela perfil do estudante universitário brasileiro**. 2023. Disponível em: https://comunica.ufu.br/noticias/2019/05/pesquisa-revela-perfil-do-estudante-universitario-brasileiro. Acesso em 10 nov. 2025.

BRASIL. MCTI. Ministério da Ciência, Tecnologia e Inovação. **População residente, produto interno bruto (PIB) e fator de conversão para paridade do poder de compra (PPC), 2000-2023**. 2023. Disponível em: https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/indicadores/paginas/dados-socioeconomicos/tabelas/10-1-populacao-residente-populacao-economicamente-ativa-pea-pessoas-ocupadas-produto-interno-bruto-pib-e-fator-de-conversao-para-paridade-do-poder-de-compra-ppc. Acesso em: 4 jun. 2025.

DHNET – DIREITOS HUMANOS NA INTERNET. **Entenda o cálculo do IDH Municipal**: e saiba os indicadores usados. Brasil: DHNet, [S. d]. Disponível em: https://www.dhnet.org.br/direitos/indicadores/idhm/idh_m_entenda_calculo2.pdf. Acesso em: 07 abr. 2025.

FUNDAÇÃO SEADE (São Paulo). **Ensino Superior – Matrículas**. São Paulo: Fundação Sistema Estadual de Análise de Dados, 2022. Disponível em: https://repositorio.seade.gov.br/dataset/educacao-do-ensino-superior/resource/f2a5f819-4d52-4d39-b172-404ab792d351. Acesso em: 14 abr. 2025.

FUNDAÇÃO SEADE (São Paulo). **Óbitos ─ Estado de São Paulo**. São Paulo: Fundação Sistema Estadual de Análise de Dados, 2021. Disponível em: https://repositorio.seade.gov.br/dataset/obitos. Acesso em: 14 abr. 2025.

FUNDAÇÃO SEADE (São Paulo). **População residente - Estado de São Paulo**. São Paulo: Fundação Sistema Estadual de Análise de Dados, 2023. Disponível em: https://repositorio.seade.gov.br/dataset/populacao-residente-estado-de-sao-paulo. Acesso em: 4 abr. 2025.

FUNDAÇÃO SEADE (São Paulo**). Mortalidade no Estado de São Paulo. São Paulo**: Fundação Sistema Estadual de Análise de Dados, 2023. Disponível em: https://repositorio.seade.gov.br/dataset/mortalidade-no-estado-de-sao-paulo. Acesso em: 4 abr. 2025.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Cidades e Estados**: Santos (SP). Brasília: IBGE, 2022. Disponível em: https://cidades.ibge.gov.br/. Acesso em: 10 nov 2025.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Tábuas Abreviadas de Mortalidade por Sexo e Idade: Brasil, Grandes Regiões e Unidades da Federação**. Rio de Janeiro: IBGE, 2013. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv65137.pdf. Acesso em: 13 jun. 2025

IBGE. Instituto Brasileiro de Geografia e Estatística. **Panorama**. [S. l.], 2021. Disponível em: https://cidades.ibge.gov.br/brasil/sp. Acesso em: 3 abr. 2025.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Tábuas completas de mortalidade para o Brasil – 2023**: breve análise da evolução da mortalidade no Brasil. Rio de Janeiro: IBGE, 2024. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/periodicos/3097/tcmb_2023.pdf. Acesso em: 29 mar. 2025.

KIMBALL, R.; ROSS, M. ***The Data Warehouse Toolkit****: the Definitive Guide to Dimensional Modeling*. 3. ed. Indianapolis: John Wiley & Sons, Inc, 2013. Disponível em: https://books.google.com.br/books?hl=pt-BR&lr=&id=4rFXzk8wAB8C&oi=fnd&pg=PR27&dq=kimball+e+ross+2013&ots=3q8NOcYcPM&sig=7CPvh5RGym4aQlth7W2d09lAueU#v=onepage&q&f=false. Acesso em: 7 nov. 2025.

MPSP. Ministério Público do Estado De São Paulo. **Levantamento de Dados**: Área Regional Santos. São Paulo: Ministério Público do Estado de São Paulo, 2023. Disponível em: https://www.mpsp.mp.br/documents/20122/1481760/Levantamento%20dados%20regionas%20NAT%20-%20%20PGA%20Santos%202023%20%281%29.pdf/9cbc485d-3fec-9887-a2d9-8a314296b5bf?t=1701717776959. Acesso em: 31 maio 2025.

PNUD. Programa das Nações Unidas Para o Desenvolvimento. **Atlas do Desenvolvimento Humano no Brasil**. Brasília: PNUD, 2022. Disponível em: http://www.atlasbrasil.org.br/. Acesso em: 23 maio 2025.

PNUD. Programa das Nações Unidas Para o Desenvolvimento. Cálculo dos índices de desenvolvimento humano. **Relatório do Desenvolvimento Humano 20**06: O que está por trás da escassez de água: poder, pobreza e a crise mundial da água. Nova Iorque: PNUD, 2006. Nota Técnica 1, p. 393-524. Disponível em: https://sswm.info/sites/default/files/reference_attachments/PNUD%202006%20Relatorio%20do%20Desenvolvimento%20Humano%202006%20-%20PORTUGUESE.pdf. Acesso em: 29 maio 2025.

SÃO PAULO (Estado). **Plano de Desenvolvimento Urbano Integrado da Região Metropolitana da Baixada Santista** – PDUI-RMBS. São Paulo: Governo do Estado de São Paulo, 2021. Disponível em: https://www.pdui.sp.gov.br/rmbs/. Acesso em: 23 maio 2025.
SIDRA. Sistema IBGE de Recuperação Automática. **Censo Demográfico**: Tabela 9841 - Taxa de alfabetização da população, total e indígena, de 15 anos ou mais de idade, por sexo, idade, quesito de declaração e localização do domicílio. IBGE: SIDRA, 2022. Disponível em: https://sidra.ibge.gov.br/tabela/9841. Acesso em: 05 jun. 2025.

A evolução no perfil dos beneficiários do ProUni: um estudo exploratório dos dados de 2005 a 2020 em Santos – SP

The evolution of the profile of ProUni beneficiaries: an exploratory study of data from 2005 to 2020 in the city of Santos - SP

Daivid Bruno Macedo Silva
Fatec Baixada Santista – Rubens Lara
daivid.silva@fatec.sp.gov.br

Júlia Perez Umbelino
Fatec Baixada Santista – Rubens Lara
julia.umbelino@fatec.sp.gov.br

Kayky Lourenço Martins
Fatec Baixada Santista – Rubens Lara
kayky.martins@fatec.sp.gov.br

Claudia Maria Sodero Salles
Fatec Baixada Santista – Rubens Lara
claudia.sodero@cps.sp.gov.br

Revista Datapoint
eISSN 3086-433X
Faculdade de Tecnologia Rubens Lara – FATEC
Ciência de Dados
Peridicidade: Anual
Vol 01, n. 01, 2025
revistadp@fatecrl.edu.br

Recebido: Jun 2025
Aceito: Set 2025
Publicado: Dez 2025

URL:https://www.fatecrl.edu.br/revista/datapoint/index.php/dp/article/view/4
DOI: https://doi.org/10.5281/zenodo.19240731

RESUMO

Este estudo exploratório analisa a evolução do perfil dos beneficiários do Programa Universidade para Todos (ProUni) na cidade de Santos, estado de São Paulo, entre 2005 e 2020. O ProUni, instituído em 2005, tem como finalidade ampliar o acesso ao ensino superior para estudantes de baixa renda, oferecendo bolsas parciais e integrais em instituições privadas. A pesquisa destaca a importância de compreender o impacto regional do programa, dado que as análises nacionais não fornecem uma visão aprofundada das características locais dos beneficiários. Para tal, foram coletados e tratados dados brutos do Ministério da Educação (MEC), focando em variáveis como sexo, raça/cor, faixa etária e curso escolhido. A metodologia empregou Python (biblioteca Pandas) para tratamento de dados e Power BI para visualização, visando identificar tendências e padrões na transformação do perfil dos bolsistas. O trabalho também contextualiza o município de Santos como um polo universitário em crescimento, impulsionado por investimentos em infraestrutura e instituições de ensino superior. Os resultados obtidos evidenciam a relevância do ProUni na ampliação do acesso ao ensino superior na baixada santista, fortalecendo a formação educacional e profissional da população local. A análise dos dados demonstra avanços significativos na inclusão e na diversidade entre os beneficiários, além de apontar oportunidades de aprimoramento nas políticas públicas voltadas à educação e à equidade social na região.

PALAVRAS-CHAVE: ProUni; Santos; ensino superior; inclusão social; ciência de dados.

ABSTRACT

This exploratory study analyzes the evolution of the profile of beneficiaries of the University for All Program (ProUni) in the city of Santos, state of São Paulo, between 2005 and 2020. Established in 2005, ProUni aims to expand access to higher education for low-income students by offering full and partial scholarships at private institutions. The research highlights the importance of understanding the program's regional impact, as national analyses do not provide an in-depth view of the local characteristics of beneficiaries. To this end, raw data from the Ministry of Education (MEC) were collected and processed, focusing on variables such as gender, race, age group, and chosen course. The methodology employed Python (Pandas library) for data processing and Power BI for visualization, aiming to identify trends and patterns in the transformation of scholarship recipients' profiles. The study also contextualizes the municipality of Santos as a growing university hub, driven by investments in infrastructure and higher education institutions. The results underscore the relevance of ProUni in expanding access to higher education in the Baixada Santista region, strengthening the educational and professional development of the local population. Data analysis reveals significant progress in inclusion and diversity among beneficiaries, as well as opportunities for improving public policies related to education and social equity in the region.

KEY-WORDS: ProUni; Santos; higher education; social inclusion; data Science.

INTRODUÇÃO

As políticas públicas de educação desempenham um papel fundamental na promoção da inclusão e na ampliação do acesso ao ensino superior. O Programa Universidade para Todos (ProUni) foi criado pelo Governo Federal por meio da Lei nº 11.096, de 13 de janeiro de 2005 com o objetivo de ampliar o acesso ao ensino superior no Brasil (BRASIL, 2005). O programa oferece bolsas de estudo parciais (50%) e integrais (100%) em instituições privadas de ensino superior, direcionadas a estudantes de baixa renda que tenham cursado o ensino médio em escolas públicas ou em instituições privadas na condição de bolsistas integrais (BRASIL, 2005).

Os dados disponibilizados sobre o perfil dos bolsistas do ProUni abrangem o programa em nível nacional, com segmentações regionais que permitem comparações detalhadas entre estados e municípios. No entanto, essas análises não são aprofundadas a ponto de fornecer uma visão clara das características dos beneficiários em contextos de locais específicos.

Para que as políticas de inclusão sejam efetivas, é essencial compreender não apenas os resultados imediatos do programa, mas também como eles afetam diferentes grupos da população ao longo do tempo. A realização de estudos detalhados sobre a evolução do perfil dos beneficiários do ProUni na Baixada Santista, por exemplo, pode facilitar a compreensão do impacto regional do programa. Isso otimiza a formulação de estratégias mais direcionadas e eficazes para essa localidade, uma vez que não é possível identificar de forma precisa as transformações no perfil dos estudantes ao longo dos anos.

Esta pesquisa tem como objetivo descrever aspectos da cidade de Santos no contexto educacional, com ênfase no Programa Universidade para Todos (ProUni) Para essa finalidade serão coletados e tratados dados extraídos de *datasets* selecionados, abrangendo o período de 2005 a 2020, garantindo sua consistência e adequação para análise e examinar informações disponibilizadas pelo MEC relacionadas ao ProUni, considerando variáveis como sexo, raça/cor, faixa etária e curso escolhido. A partir dos dados tratados, a pesquisa buscará desenvolver análises gráficas ilustrativas.

A coleta de dados será realizada a partir das informações disponibilizadas nos dados abertos do Ministério da Educação (MEC) sobre o Programa Universidade para Todos (ProUni). Após a obtenção dos dados, será realizado o tratamento utilizando a linguagem Python, por meio da biblioteca Pandas, para limpar, organizar e estruturar as informações

Com os dados tratados, as informações serão importadas para o Power BI, onde serão criados gráficos interativos para análise visual dos resultados.

1. FUNDAMENTAÇÃO TEÓRICA

De acordo com Moino (2023), a cidade de Santos, situada no litoral paulista, tem passado por profundas transformações socioeconômicas nas últimas décadas. Historicamente, a cidade é reconhecida pela importância estratégica de seu porto, que segundo a Comissão Econômica para América Latina e Caribe (CEPAL, 2025) é considerado o maior da América Latina e elemento central no comércio internacional e brasileiro.

Entretanto, a partir dos anos 1980 e 1990, com o declínio de algumas atividades industriais e o fortalecimento da urbanização, Santos iniciou um processo de diversificação de sua economia. Nesse contexto, o setor educacional emergiu como um dos principais vetores de desenvolvimento local. De acordo com Paiva e Righi (2020), a cidade passou por transformações significativas, abandonando gradativamente seu perfil industrial e se adaptando às novas demandas econômicas e sociais, consolidando-se como um eixo estratégico no desenvolvimento regional e nacional. Essa transição permitiu que Santos começasse a construir uma nova identidade, voltada para o ensino superior e a formação de capital humano.

Ainda nesse sentido, a criação do Parque Tecnológico de Santos reforça esse movimento, já que tal iniciativa é capaz de posicionar a cidade como um centro de inovação e pesquisa, ampliando as possibilidades acadêmicas e fomentando a interação entre universidades, setor produtivo e tecnologia (FUNARBE, 2023).

Atualmente, a cidade de Santos, além de preservar sua relevância histórica e econômica como porto estratégico, destaca-se também como referência acadêmica e científica no estado de São Paulo. As políticas de inclusão adotadas pelas universidades locais têm transformado a vida de jovens de comunidades menos favorecidas, possibilitando a quebra do ciclo de desigualdade e contribuindo de forma significativa para o progresso social e econômico da região (UNESCO, 2019).

1.1 DIVERSIDADE E INCLUSÃO NO ENSINO SUPERIOR NA CIDADE DE SANTOS

A diversidade e a inclusão no ensino superior constituem pilares fundamentais para a construção de uma sociedade mais equitativa e a mitigação das desigualdades sociais. De acordo com Freire (1997), o acesso à educação é essencial para que esses grupos possam desenvolver uma consciência crítica e transformar as estruturas sociais que perpetuam a desigualdade. Tais práticas não apenas promovem o acesso à educação para grupos

historicamente marginalizados, mas também fomentam um ambiente de aprendizado enriquecido por uma multiplicidade de perspectivas e experiências.

Ainda segundo o autor, essa perspectiva reforça a importância das universidades como agentes de transformação social, alinhando-se à ideia de que elas desempenham um papel estratégico no desenvolvimento econômico e social das regiões em que estão inseridas. Paiva e Righi (2020), afirmam que as políticas de ações afirmativas, como as cotas raciais e sociais, e programa de bolsas de estudo desempenham um papel essencial na democratização do acesso à educação por indivíduos oriundos de contextos socioeconômicos menos favorecidos no Brasil. Esses instrumentos são indispensáveis para atenuar as barreiras históricas que dificultam o ingresso e a permanência de determinados grupos na esfera acadêmica.

1.2 PROGRAMA UNIVERSIDADE PARA TODOS (PROUNI) E O IMPACTO LOCAL NA CIDADE DE SANTOS

A estrutura do ProUni, de acordo com Portal Único de acesso ao Ensino Superior, baseia-se em um modelo de parceria público-privada, no qual o governo concede isenções fiscais às instituições privadas em contrapartida à oferta de bolsas de estudo.

No período de 2005 a 2020, correspondente ao escopo deste estudo, a implementação do ProUni em Santos gerou transformações significativas no cenário educacional local. Dados fornecidos pelo Ministério da Educação (MEC, 2025) revelam um crescimento expressivo no número de bolsas concedidas, abrangendo beneficiários de diferentes faixas etárias, gêneros e grupos raciais. Essa diversificação dos perfis contribuiu para a formação de capital humano qualificado, atendendo às demandas de setores estratégicos na cidade, como saúde e educação. Ademais, a ampliação do acesso ao ensino superior fortaleceu a coesão social, promovendo inclusão e oportunidades para jovens de baixa renda.

1.3 FERRAMENTAS TECNOLÓGICAS UTILIZADAS NO ESTUDO

O desenvolvimento deste trabalho se deu apoiado na utilização de algumas ferramentas tecnológicas, sendo elas: linguagem Python com a utilização da biblioteca pandas, Excel e Power BI.

Conforme McKinney (2017), Python é uma linguagem de programação amplamente empregada em diversas áreas, sobretudo na ciência de dados e análise computacional. Nesse contexto, a biblioteca Pandas destaca-se como uma ferramenta fundamental para o tratamento de conjuntos de dados. Segundo o autor, suas estruturas principais, como DataFrame e Series, oferecem recursos eficientes para leitura, filtragem e agregação, contribuindo significativamente para a otimização de processos analíticos.

O Microsoft Excel é um dos programas mais utilizados para análise de dados, especialmente no contexto corporativo. Segundo Walkenbach (2013), sua interface baseada em planilhas e facilita a organização e o processamento de grandes conjuntos de dados, além de oferecer recursos avançados, como funções estatísticas e ferramentas de visualização.

O Power BI, desenvolvido pela Microsoft, é uma plataforma de inteligência de negócios que permite a criação de relatórios e dashboards interativos. Conforme enunciado por Davenport (2014), sua capacidade de integrar diferentes fontes de dados e apresentar informações visualmente contribui para tomadas de decisão estratégicas e baseadas em evidências.

2. PROCEDIMENTOS METODOLÓGICOS

O processo de ETL (Extract, Transform, Load) foi essencial para garantir a integridade e padronização dos dados utilizados nesse estudo. Durante a fase de extração, os arquivos obtidos em formato CSV de cada ano de 2005 até 2020 foram coletados do portal de dados abertos do MEC (MEC, 2022), convertidos para XLSX e carregados para a análise.

Os dados referentes a cada ano foram armazenados em um arquivo no formato XLSX. Para viabilizar sua utilização, foram efetuados ajustes nos dados, de modo a adequá-los às necessidades das análises pretendidas. O Quadro 1 lista todas as mudanças realizadas no *dataset*:

Quadro 1 – Listagem das mudanças aplicadas no dataset

O QUE FOI FEITO	COMO FOI FEITO	POR QUE FOI FEITO
Exclusão de linhas	Com base na coluna Municipio_beneficiário_Bolsa, excluídas as linhas cujo conteúdo fosse diferente de SANTOS	Para atendimento ao escopo do estudo
Padronização do conteúdo de coluna	Com base na coluna Sexo_beneficiário_bolsa, foi realizada a substituição dos termos "Masculino" por "M" e "Feminino" por "F"	Em alguns anos, o sexo era registrado como "Masculino/Feminino", enquanto em outros era indicado como "M/F"
	Com base na coluna Beneficiário_deficiente_físico, foi realizada a substituição dos termos "NÃO" por "N" e "SIM" por "S"	Em alguns anos, era registrado como "Não/Sim", enquanto em outros era indicado como "N/S"
	Com base na coluna Tipo_Bolsa, foi realizada a substituição dos termos "BOLSA INTEGRAL" por "INTEGRAL" e "BOLSA PARCIAL" por "PARCIAL"	Em alguns anos, era registrado como "BOLSA INTEGRAL/BOLSA PARCIAL", enquanto em outros era indicado como "INTEGRAL/PARCIAL"
Padronização de *layout* dos *datasets*	Exclusão das colunas “Campus” e “Município” nos arquivos referentes ao ano de 202	Os dados relativos ao ano de 2020 continham duas colunas adicionais. Estas colunas foram excluídas para garantir a conformidade com os demais anos do conjunto de dados
Criação de coluna	foi calculada a idade dos beneficiários utilizando a fórmula DATADIF no Excel e criada coluna Idade_Beneficiario_Bolsa	calcular a idade de acordo com a data de nascimento e o ano da bolsa concedida

Fonte: Elaborado pelos autores (2025)

Após essas modificações, a quantidade de linhas em cada arquivo foi consolidada conforme exposto no Quadro 2.

Quadro 2 – Resultado dos tratamentos aplicados nos *datasets*

Ano-base dos dados	Quantidade original de linhas	Quantidade final de linhas (após aplicação dos tratamentos)
2005	95.629	208
2006	109.025	171
2007	105.574	186
2008	124.621	187
2009	161.369	376
2010	152.733	284
2011	170.766	367
2012	176.764	377
2013	177.326	362
2014	223.598	476
2015	252.650	596
2016	239.262	630
2017	236.636	635
2018	241.032	450
2019	225.555	396
2020	166.830	262

Fonte: elaborado pelos autores, adaptado dos dados abertos do MEC (2025).

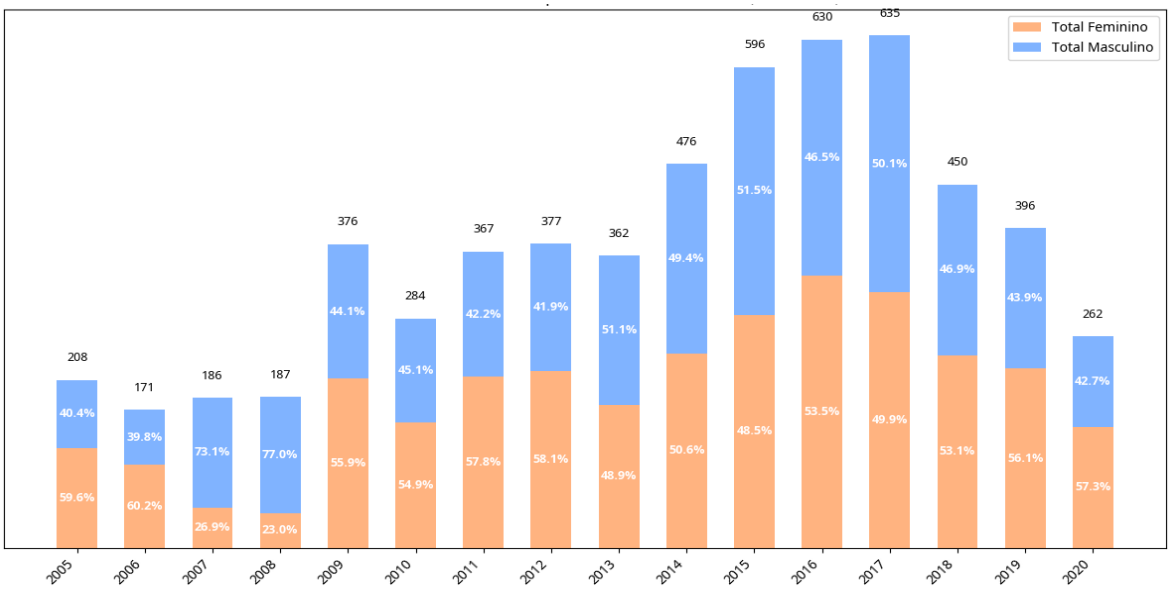
Para facilitar a exploração dos dados e visualização dos resultados, foram desenvolvidas análises gráficas utilizando Excel e Power BI, permitindo uma visualização dinâmica das tendências e distribuições. A análise exploratória revelou diversos padrões e tendências nos dados de bolsas de estudo em Santos entre 2005 e 2020.

3. RESULTADOS E DISCUSSÕES

A análise da série histórica revela uma dinâmica interessante na concessão de bolsas ao longo dos anos cobertos pelos dados, permitindo uma visão mais aprofundada sobre possíveis mudanças e tendências ao longo do tempo.

Observa-se no Gráfico 1 a distribuição das bolsas ao longo do período analisado, divididas por gênero, o que possibilita uma reflexão sobre a equidade de acesso e representatividade entre homens e mulheres.

Gráfico 1 - Total de bolsas concedidas na cidade de Santos-SP



Fonte: elaborado pelos autores, adaptado dos Dados abertos do MEC (2025).

Os primeiros anos (2005-2008) mostram um volume relativamente estável, abaixo de 200 bolsas anuais. Um salto significativo ocorre no ano de 2009, com 376 bolsas, quase o dobro do ano anterior. A partir do período de 2014 a 2017, concentra os maiores volumes anuais, superando 450 bolsas no ano de 2014 e ultrapassando 600 bolsas no ano de 2016 e 2017.

Após o pico no ano de 2017, os anos subsequentes apresentam uma redução recorrente no número de concessões, retornando a patamares vistos entre 2009 e 2013, com 262 bolsas registradas no ano de 2020.

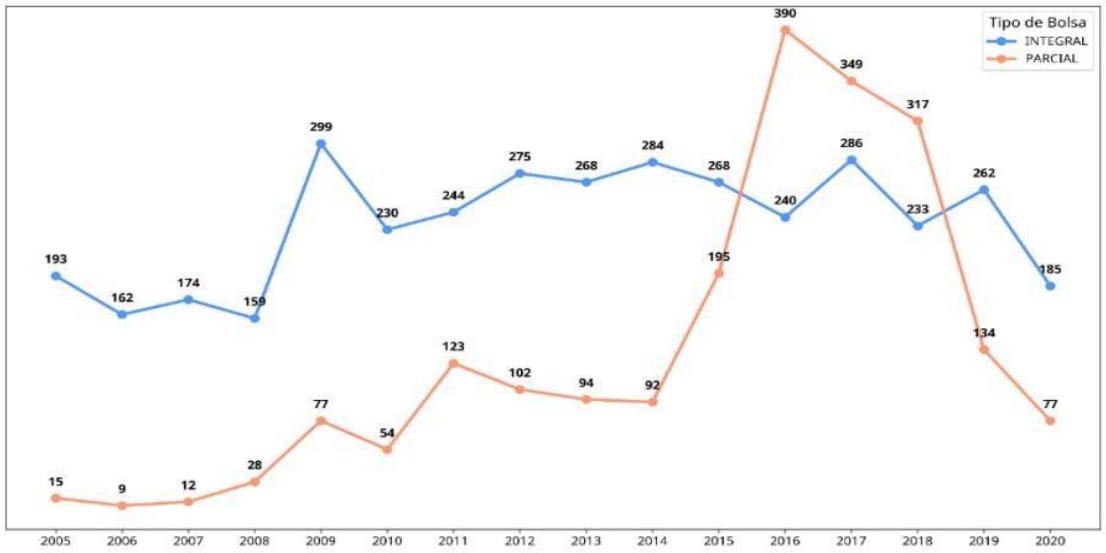
Apesar da oscilação no total de bolsas concedidas, percebe-se uma certa manutenção no padrão de gênero dos beneficiários. Com exceção dos anos de 2007 e 2008 onde houve uma predominância de beneficiários do sexo masculino, os demais anos mostram uma divisão mais igualitária entre gêneros, com leve majoração para o gênero feminino. Há de se considerar que a cidade de Santos tem população de maioria feminina (IBGE, 2022).

3.1 ANÁLISE POR TIPOS DE BOLSA: PARCIAL X INTEGRAL

Os dados classificam as bolsas em Integrais e Parciais. Observa-se no Gráfico 2 uma predominância de bolsas Integrais, totalizando 3746 (aproximadamente 63%) bolsas no período, enquanto as bolsas Parciais somam 2217 (cerca de 37%).

Nos anos de 2015 a 2017, as bolsas Parciais chegaram a superar as Integrais em número absoluto, indicando uma possível mudança na política de concessão ou no perfil da demanda nesse período. Nos anos mais recentes (2018-2020), a proporção voltou a se equilibrar, com uma leve vantagem para as bolsas Integrais em 2019 e 2020.

Gráfico 2 Tipos de Bolsa ao longo dos anos em Santos - SP

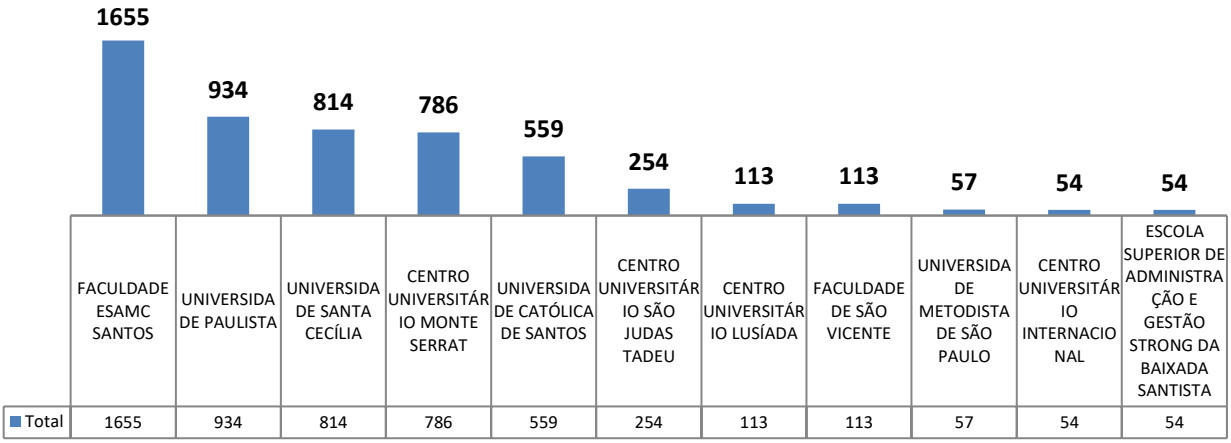


Fonte: dos autores, adaptado do Ministério da Educação (2025)

3.2 ANÁLISE POR INSTITUIÇÃO E CURSOS

A Faculdade ESAMC Santos destacou-se como a instituição com maior número de bolsas concedidas, seguida por Universidade Paulista (UNIP) e Universidade São Judas Tadeu (somadas as bolsas das instituições São Judas Tadeu e Unimonte, uma vez que ambas pertencem ao mesmo grupo educacional), o conforme gráfico 3.

Gráfico 3 - Instituições com mais bolsas



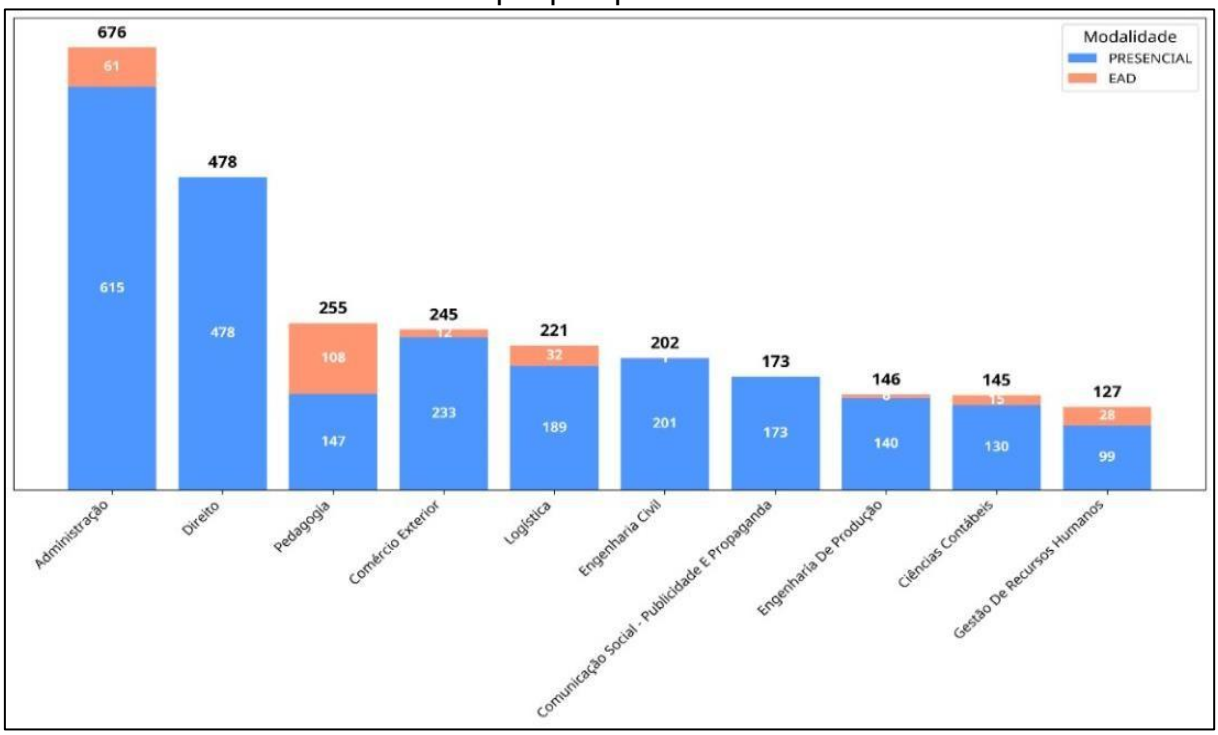
Fonte: dos autores, adaptado do Ministério da Educação (2025)

Os dados analisados, conforme ilustrado no Gráfico 4, revelam que os cursos de Administração, Direito e Pedagogia se destacaram como os mais procurados pelos estudantes ao longo do período. Esta predominância sugere uma demanda consistente por formações tradicionais nessas áreas.

Adicionalmente, observa-se a relevância dos Cursos Superiores de Tecnologia, em particular Comércio Exterior e Logística. A significativa procura por essas modalidades tecnólogas, voltadas para setores operacionais e logísticos, reflete a grande aplicabilidade e a demanda do mercado de trabalho na região.

Outros cursos que figuram entre os dez mais procurados incluem formações como Engenharia Civil, Engenharia de Produção, Ciências Contábeis, Comunicação Social - Publicidade e Propaganda e Gestão de Recursos Humanos. A análise da modalidade de ensino para esses cursos de destaque indica uma clara dominância do formato presencial. Contudo, é notável que cursos como Administração e Pedagogia também apresentaram um número considerável de bolsas na modalidade de Educação a Distância (EAD), apontando para uma diversificação nas preferências de estudo.

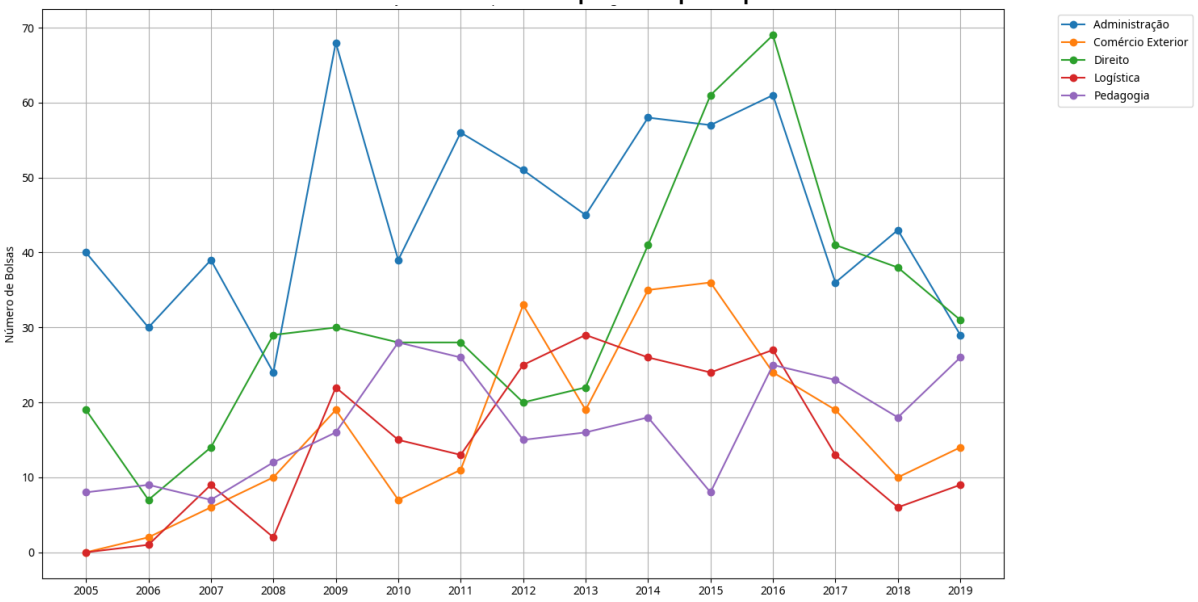
Gráfico 4 - Top 10 principais cursos escolhidos



Fonte: dos autores, adaptado do Ministério da Educação (2025).

Ao examinar a evolução da procura por esses cursos ao longo dos anos, conforme detalhado no Gráfico 5 os cinco principais cursos escolhidos, percebe-se que o curso de Administração se manteve como o mais procurado durante a maior parte do período analisado, embora note-se uma queda acentuada nos últimos anos da análise juntamente com o curso de Direito, enquanto os demais cursos, pode-se verificar uma leva ascensão.

Gráfico 5 – Detalhamento no tempo dos 5 principais cursos escolhidos



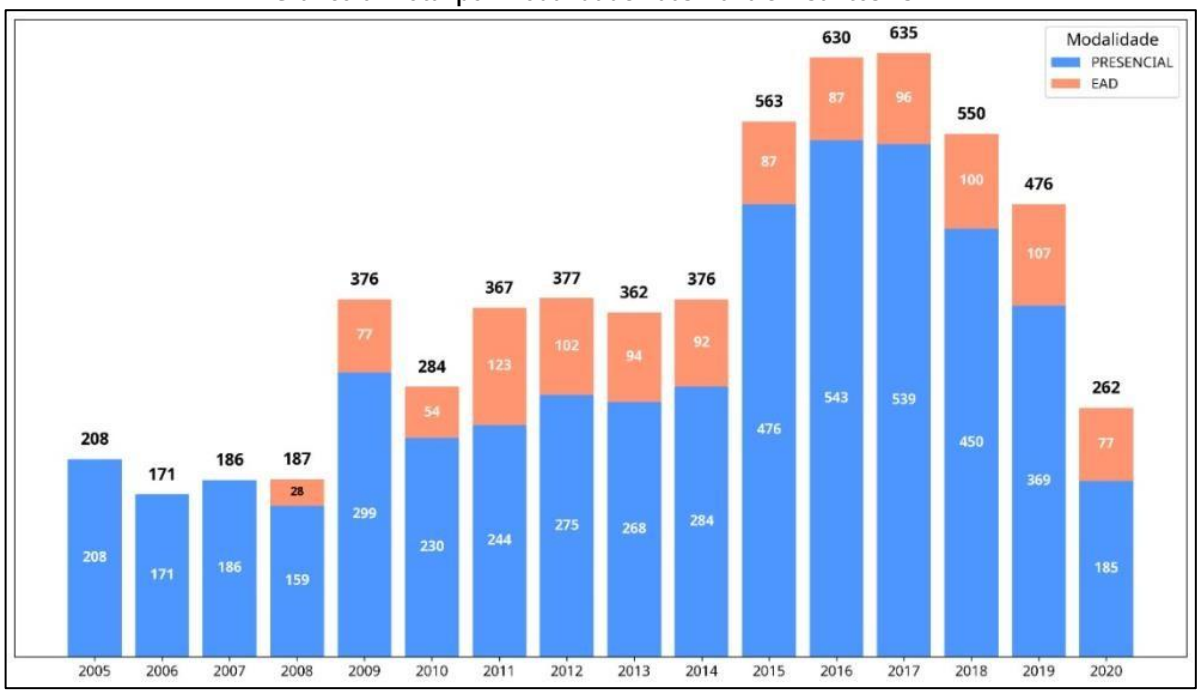
Fonte: dos autores, adaptado do Ministério da Educação (2025)

No entanto, em anos mais recentes, o curso de Direito assumiu um protagonismo crescente, indicando uma mudança nas tendências de escolha.

3.3 ANÁLISE POR MODALIDADE DE ENSINO

A modalidade de ensino é um fator crucial na caracterização das bolsas. Os dados mostram uma predominância expressiva da modalidade Presencial, conforme exposto no Gráfico 6.

Gráfico 6 - Total por modalidade 2005-2020 em Santos - SP



Fonte: dos autores, adaptado dos Dados abertos do MEC

Considerando o total de bolas ofertadas, a modalidade presencial corresponde a aproximadamente 91% do total, já a modalidade de Ensino a Distância (EAD) representa cerca de 9%. A análise da evolução temporal da modalidade indica que a modalidade EAD, embora minoritária, tem ganho alguma representatividade ao longo do tempo.

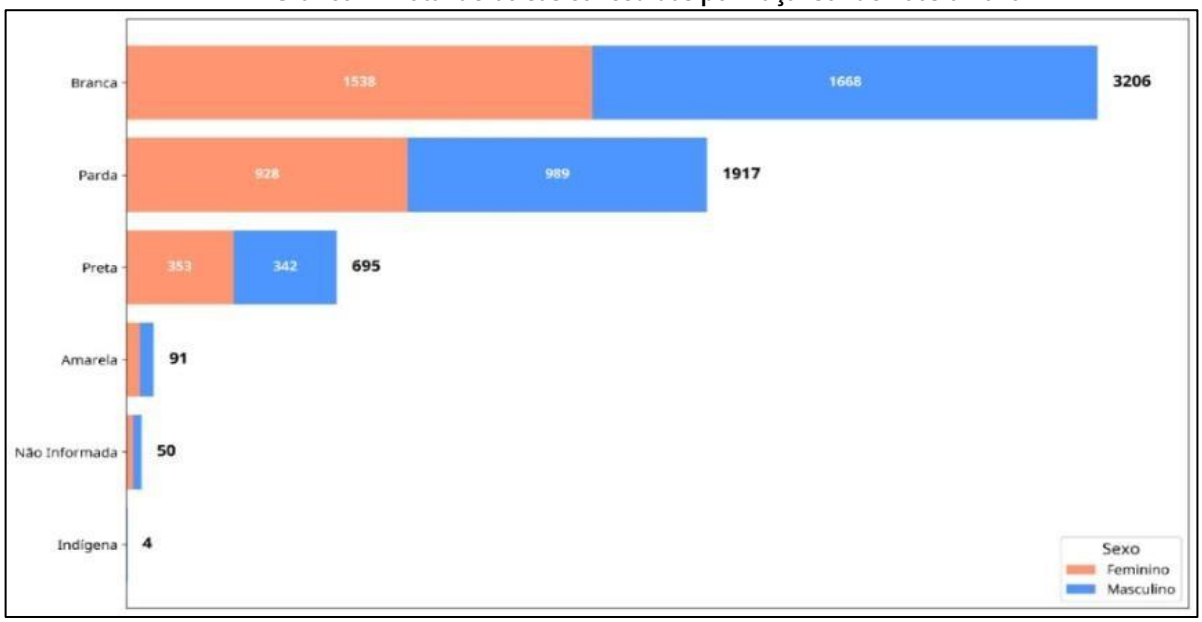
3.4 PERFIL DEMOGRÁFICO DOS BENEFICIÁRIOS

A análise da distribuição de bolsas por raça/cor, considerando as categorias mais frequentes, revela que a maioria dos beneficiários se autodeclarou como Branca (54%). O Gráfico 6 detalha a distribuição total de bolsas concedidas por raça/cor, segmentada por sexo, oferecendo uma compreensão mais detalhada. Observa-se que, para a categoria Branca, há um número ligeiramente maior de bolsas concedidas a homens em comparação com mulheres.

As demais categorias de raça/cor representam uma parcela significativamente menor do total de beneficiários, com a categoria indígena apresentando o menor número de bolsas.

Esta segmentação por sexo e raça/cor é crucial para identificar padrões e potenciais disparidades no acesso às bolsas, sugerindo a necessidade de investigações mais aprofundadas sobre os fatores que influenciam essas distribuições.

Gráfico 7 - Total de bolsas concedidas por Raça/Cor de 2005 a 2020



Fonte: dos autores, adaptado dos Dados abertos do MEC (2025)

A tabela 1 apresenta os dados disponíveis no censo demográfico de Santos 2022:

Tabela 1 – Dados Populacionais de Santos, Censo 2022

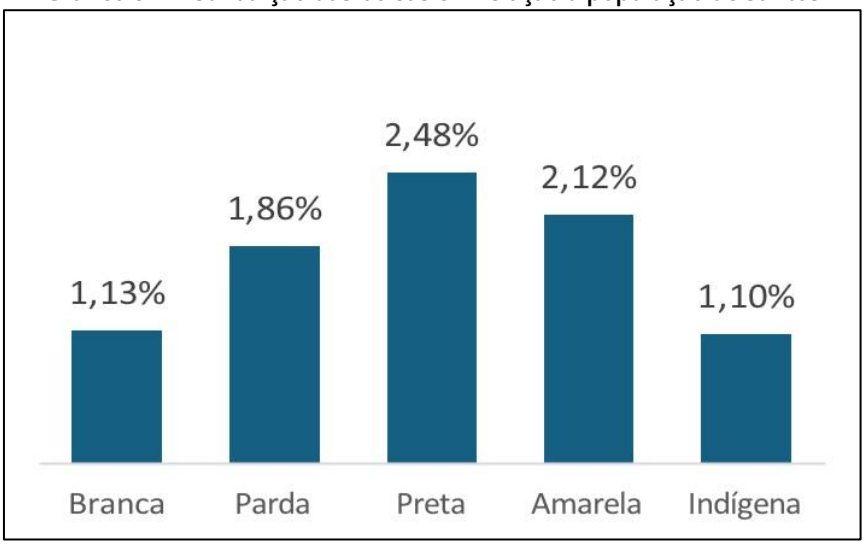
Cor ou raça	População (pessoas)	População %
Branca	282.641	67,52
Parda	103.273	24,67
Preta	28.016	6,69
Amarela	4.290	1,02
Indígena	364	0,09
Total	418.584	100,00

Fonte: dos autores, adaptado do censo do IBGE (2025)

Considerando os dados da população da cidade de Santos, obtidos por meio do censo 2022, a cidade possui 418.608 habitantes, com predominância da população branca que representa aproximadamente 70% do total.

Ao cruzar os dados do censo com os dados do ProUni, foi possível gerar o Gráfico 8 que indica a distribuição percentual das bolsas concedidas em relação a população de Santos.

Gráfico 8 – Distribuição das bolsas em relação a população de Santos



Fonte: dos autores, adaptado do MEC e censo do IBGE (2025)

Embora a maioria das bolsas do Prouni tenha sido concedida a pessoas brancas, com uma diferença significativa em relação ao número de pretos contemplados, a análise proporcional revela uma dinâmica diferente. Quando se considera o percentual de indivíduos de cada grupo racial que obteve acesso às bolsas em relação ao total da sua população, observa-se que a taxa de acesso entre pretos é superior à verificada entre brancos. Isso indica que, proporcionalmente, o Prouni beneficiou mais pretos do que brancos, ainda que o número absoluto de bolsistas brancos permaneça mais elevado.

4. CONSIDERAÇÕES FINAIS

O presente estudo sobre a evolução do perfil dos beneficiários do Programa Universidade para Todos (ProUni) no município de Santos, entre os anos de 2005 e 2020, possibilitou a identificação de transformações significativas no panorama educacional local. Verificou-se um crescimento expressivo na concessão de bolsas ao longo dos anos, impulsionado notadamente pelas bolsas parciais, com destaque para os períodos entre 2014 e 2017.

Contudo, nos dois últimos anos do período analisado, observou-se uma queda relevante em ambos os tipos de bolsa, o que sugere a necessidade de estudos futuros para determinar se tal fenômeno constitui uma tendência ou um evento pontual. A predominância das bolsas integrais foi marcante, sobretudo nos anos iniciais, embora, em determinado intervalo, tenha-

se observado uma inversão pontual com o crescimento relativo das bolsas parciais. A distribuição por modalidade evidenciou a ampla prevalência do ensino presencial, ainda que a modalidade EAD venha conquistando gradualmente maior espaço.

No que corresponde à escolha dos cursos e instituições, destacou-se a preferência acentuada por áreas como Administração e Direito, predominantemente na modalidade presencial, bem como a relevância de cursos tecnólogos voltados à vocação logística da região.

A análise temporal da procura por cursos, evidenciada no Gráfico 5, revela dinâmicas interessantes na preferência dos estudantes. O curso de Administração, embora tenha apresentado flutuações, manteve-se consistentemente entre os mais procurados ao longo de quase todo o período. Notavelmente, o curso de Direito demonstrou uma ascensão significativa, especialmente nos anos mais recentes, chegando a superar Administração em alguns momentos.

Cursos como Comércio Exterior e Logística, apesar de apresentarem picos de interesse em períodos específicos, como o início dos anos 2010 para Comércio Exterior, mostraram uma tendência de estabilização ou menor destaque em comparação com os líderes. A Pedagogia, por sua vez, demonstrou uma procura mais estável, porém em patamares inferiores aos dois primeiros, indicando uma demanda contínua, mas menos volátil

A análise sociodemográfica revelou um retrato diversificado dos beneficiários, com uma distribuição equilibrada entre homens e mulheres. Em números absolutos, constatou-se a predominância da população branca. No entanto, ao analisar os dados relativos à população geral, notou-se uma preponderância das populações preta e parda, o que corrobora o papel do ProUni como uma política de ascensão social por intermédio da educação superior.

REFERÊNCIAS

ATLAS. Atlas do desenvolvimento humano no Brasil. Disponível em: http://www.atlasbrasil.org.br/. Acesso em: 06 abr. 2025.

A TRIBUNA. **Santos é um dos mais importantes polos universitários do estado de São Paulo.** Disponível em: https://www.atribuna.com.br/noticias/educacao/santos-e-um-dos-mais-importantes-polos-universitarios-do-estado-de-s-o-paulo-1.47396. Acesso em: 01 jun. 2025.

BRASIL. Resolução normativa de 13 de janeiro de 2005. **Lei 11.096, Programa Universidade para Todos (ProUni)**. Disponível em: https://www.planalto.gov.br/ccivil_03/_Ato2004-2006/2005/Lei/L11096.htm. Acesso em: 06 abr. 2025.

CEPAL. Comissão econômica para américa latina e caribe. **FAL – boletim 404**, **n.** 5. 2024. Disponível em: https://repositorio.cepal.org/server/api/core/bitstreams/f15aee63-dc86-40b4-922c-51ecc2f3f28d/content. Acesso em: 06 abr. 2025.

COSTA, Danielle Dias da; FERREIRA, Norma-Iracema de Barros. O PROUNI na educação brasileira: Indicadores de acesso e permanência. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 22. 2017. Disponível em: https://www.scielo.br/j/aval/a/q7HhxxZ9MYWXYGS7CyjYPfs/. Acesso em: 06 abr. 2025.

DAVENPORT, T. H. ***Big Data at Work****: Dispelling the Myths, Uncovering the Opportunities*. Harvard *Business Review Press*, 2014.

FREIRE, P. **Pedagogia da esperança:** um reencontro com a pedagogia do oprimido. Rio de Janeiro: Paz e Terra, 1997.

FUNABRE. **Parque Tecnológico: o futuro da inovação**. Disponível em: https://funarbe.org.br/2023/09/01/parques-tecnologicos-o-futuro-da-inovacaohttps://funarbe.org.br/2023/09/01/parques-tecnologicos-o-futuro-da-inovacao-no-brasil/nobrasil/. Acesso em: 06 jun. 2025.

GOMES, Matheus; HIRATA, Guilherme. Determinantes da evasão no ensino superior: Uma abordagem de riscos competitivos. **PPE – Pesquisa e planejamento econômico**, v. 52, n. 3. 2022. Disponível em: https://repositorio.ipea.gov.br/bitstream/11058/13040/1/PPE_v52_n3_artigo1_determinantes_da_evasao.pdf. Acesso em: 06 abr. 2025.

IBGE. Instituto Brasileiro de Geografia e Estatística. **Panorama do Censo 2022.** Disponível em: https://censo2022.ibge.gov.br/panorama/index.html. Acesso em: 05 abr. 2025.

IBGE. Instituto Brasileiro de Geografia e Estatística. Produto Interno Bruto dos Municípios. Disponível em: https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html?=&t=downloads&c=3513504. Acesso em: 05 abr. 2025.

INEP. **Censo da educação superior**. Disponível em: https://www.gov.br/inep/ptbr/areas-de-atuacao/pesquisas-estatisticas-ehttps://upcommons.upc.edu/bitstream/handle/2117/336593/9938-11037-1-SM.pdfindicadores/censo-da-educacaosuperior/resultados. Acesso em: 06 abr. 2025.

MAINARDES, Jeferson. Análise de políticas educacionais: breves considerações teórico-metodológicas. **Contrapontos,** v. 9, n. 1. 2009. Disponível em: https://ri.uepg.br/riuepg/bitstream/handle/123456789/240/ARTIGO_An%C3%A1lisePoliticasEducacionais.pdf?sequence=1. Acesso em: 06 abr. 2025.

McKinney, W. ***Python for Data Analysis****: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly Media, 2017.

MEC. Portal Único de Acesso ao Ensino Superior. **PROUNI**. Disponível em: https://acessounico.mec.gov.br/prouni. Acesso em: 06 abr. 2025.

MICROSOFT CORPORATION. **ETL** (extração, transformação e carregamento). Disponível em: https://learn.microsoft.com/pt-br/azure/architecture/data-guide/relational-data/etl. Acesso em: 1 out. 2025.

MIRANDA, Paula Roberta; AZEVEDO, M. Fies e Prouni na expansão da educação superior brasileira. **Revista Educação & Formação**, v. 5, n 03. 2020. Disponível em: https://revistas.uece.br/index.php/redufor/article/view/1421. Acesso em: 06 abr. 2025.

MOINO, Carolina Ceci. Estudo do crescimento urbano na cidade de Santos e a formação de regiões socioeconômicas e ambientais vulneráveis. **Revista Arte**, v.20, n. 1. 2023. Disponível em: https://revistas.belasartes.br/arte21/article/view/448. Acesso em 06 abr. 2025.

PAIVA FILHO, Hilmar Diniz; RIGHI, Roberto. O desenvolvimento urbano na cidade de Santos, o local, o regional e o nacional e, 400 anos de história. **XII SIIU – Seminário Internacional de Investigação em Urbanismo**, n. 12. 2020. Disponível em: https://upcommons.upc.edu/bitstream/handle /2117/ 336593/ 9938-11037https://datampe.sebrae.com.br/profile/geo/santos?sequence=11SM.pdf?sequence=1. Acesso em: 06 abr. 2025.

SEBRAE. Data MPE Brasil: Dados sobre o município de Santos. Disponível em: https://datampe.sebrae.com.br/profile/geo/santos?selector245id=geo3518701%2Cgeo3541000%2Cgeo3548500%2Cgeo3506359&selector244id=sector1%2C2%2C3%2C4%2C5. Acesso em: 06 abr. 2025.

SILVEIRA, Rogerio Leandro Lima da; DEPONTI, Cidonea Machado. **Desenvolvimento Regional:** Processos, Políticas e transformações Territoriais. São Carlos: Pedro & Jõao Editores, 2020. E-book.

SOUSA, Flavio Eliziario de; FREIESLEBEN, Mariane. A educação como fator de desenvolvimento regional. **Rev. FAE**, Curitiba, v. 21, n. 2, p. 163 – 178. 2018. Disponível em: https://revistafae.fae.emnuvens.com.br/revistafae/article/view/571/483. Acesso em: 06 abr. 2025

UNESCO. Santos UNESCO *creative city-film*. ***Association monitoring report 2015 – 2019.*** Disponível em: https://www.unesco.org/sites/default/files/medias/fichiers/2025/02/Santos_Monitoring_Report_2019.pdf. Acesso em: 06 abr. 2025.

WALKENBACH, J. ***Excel Bible***. New York, NY: Wiley Publishing, 2013.

Modelagem de padrões de sucesso no IMDb (1960-2024) usando aprendizado de máquina e otimização de hiperparâmetros

Modeling success patterns on IMDb (1960-2024) using machine learning and hyperparameter optimization

Robert Richard das Neves Correia dos Santos
Fatec Baixada Santista - Rubens Lara
robert.santos01@fatec.sp.gov.br

Luís Felipe Ruas do Nascimento
Fatec Baixada Santista - Rubens Lara
luis.nascimento20@fatec.sp.gov.br

Victor Barbosa Gonçalves
Fatec Baixada Santista - Rubens Lara
victor.goncalves4@fatec.sp.gov.br

Victor Roma Vianna Ferreira
Fatec Baixada Santista - Rubens Lara
victor.ferreira38@fatec.sp.gov.br

José Augusto Theodósio Pazetti
Fatec Baixada Santista - Rubens Lara
jose.pazetti01@cps.sp.gov.br

Revista Datapoint
eISSN 3086-433X
Faculdade de Tecnologia Rubens Lara – FATEC
Ciência de Dados
Peridicidade: Anual
Vol 01, n. 01, 2025
revistadp@fatecrl.edu.br

Recebido: Jun 2025
Aceito: Set 2025
Publicado: Dez 2025

URL: https://www.fatecrl.edu.br/revista/datapoint/index.php/dp/article/view/5
DOI: https://doi.org/10.5281/zenodo.19240819

RESUMO

Este estudo analisa e modela padrões de sucesso entre os filmes mais bem avaliados no IMDb entre 1960 e 2024, utilizando aprendizado de máquina supervisionado e não supervisionado. O objetivo é compreender os fatores que explicam o desempenho crítico, comparando modelos de regressão e avaliando quanto da variação das notas IMDb pode ser explicada pelos atributos dos filmes. O pipeline foi desenvolvido em Python (Google Colab) e incluiu etapas de pré-processamento com One-Hot Encoding e Standard Scaler (Scikit-learn), redução de dimensionalidade com PCA, e agrupamento com K-Means. Para a modelagem preditiva, aplicaram-se os algoritmos KNN, SVM, Random Forest e XGBoost, com ajuste de hiperparâmetros via Optuna. As visualizações foram geradas com Matplotlib e Seaborn. Os resultados destacam o XGBoost como modelo de melhor desempenho, revelando que indicações a prêmios, duração e número de votos são as variáveis mais associadas a notas elevadas, oferecendo uma visão ampla dos fatores que caracterizam o sucesso cinematográfico ao longo das décadas.

PALAVRAS-CHAVE: Aprendizado de Máquina; IMDb; Otimização de Hiperparâmetros; Sucesso Cinematográfico; Clusterização

ABSTRACT

This study analyzes and models success patterns among the top-rated films on IMDb from 1960 to 2024 using both supervised and unsupervised machine learning approaches. The aim is to understand the factors explaining critical performance by comparing regression models and assessing how much of IMDb rating variability can be explained by film attributes. The pipeline was implemented in Python (Google Colab) and included pre-processing with One-Hot Encoding and Standard Scaler (Scikit-learn), dimensionality reduction via PCA, and clustering with K-Means. For predictive modeling, KNN, SVM, Random Forest, and XGBoost algorithms were applied, with hyperparameter tuning using Optuna. Visualizations were generated through Matplotlib and Seaborn. Results highlight XGBoost as the best-performing model, indicating that award nominations, duration, and number of votes are the strongest predictors of higher ratings, providing comprehensive insights into the factors shaping cinematic success over time.

KEY-WORDS: Machine Learning; IMDb; Hyperparameter Optimization; Cinematic Success; Clustering

INTRODUÇÃO

De acordo com a B_Arco (2023), “o setor audiovisual tem presenciado uma transformação extraordinária ao longo das últimas décadas, impulsionada pelo avanço rápido da tecnologia”.

Nesse cenário, o uso de técnicas de aprendizado de máquina surge como uma ferramenta promissora para compreender os múltiplos fatores que explicam o sucesso cinematográfico.

Este estudo investiga quais características dos filmes, como gênero, país, década, duração e número de votos, influenciam significativamente as notas IMDb e em que medida tais atributos permitem modelar e explicar a variação dessas avaliações. O objetivo central é analisar e comparar o desempenho de diferentes modelos de aprendizado de máquina na identificação e interpretação dos padrões de sucesso cinematográfico, e não apenas prever notas, mas compreender o comportamento das variáveis que mais contribuem para o reconhecimento crítico.

Para alcançar esses objetivos, foi desenvolvido um *pipeline* híbrido de ciência de dados, estruturado em Python (Google Colab) e composto por etapas de pré-processamento (*One-Hot Encoding* e *Standard Scaler*), análise exploratória, redução de dimensionalidade via PCA, agrupamento com *K-Means* e aplicação de modelos supervisionados (KNN, SVM, *Random Forest* e *XGBoost*). O processo incluiu ainda a otimização automática de hiperparâmetros com a biblioteca *Optuna*, garantindo maior precisão e robustez na comparação entre algoritmos.

Com base nesses procedimentos, o estudo busca não apenas mensurar o desempenho dos modelos por meio de métricas como R^2 e RMSE, mas também identificar os fatores mais determinantes do sucesso cinematográfico no IMDb, oferecendo uma perspectiva quantitativa sobre a evolução da recepção crítica e do público ao longo de mais de seis décadas.

1. FUNDAMENTAÇÃO TEÓRICA

A fundamentação teórica deste estudo se divide em duas vertentes principais: a análise de fatores de sucesso cinematográfico e a aplicação de técnicas de aprendizado de máquina no domínio cultural.

1.1 SOBRE O INTERNET MOVIE DATABASE (IMDB)

Segundo Mercado Filho (2022), o *Internet Movie Database* (IMDb) foi criado em 1990 por Col Needham como um banco de dados mantido por fãs e, após migrar para a *web* em 1993, consolidou-se como uma enciclopédia *online* que reúne informações sobre filmes, séries e artistas.

> Além de reunir informações sobre artistas e produções, o site também permite que usuários criem listas e avaliem seus filmes favoritos. Qualquer usuário cadastrado pode dar notas aos títulos. Os votos individuais são agregados e resumidos em uma nota única, exibida com destaque na página principal do título (Melo, 2018 *apud* Coelho; Ferreira; Faustino, 2021, p. 2).

"Com autoridade na indústria de entretenimento, […], a IMDb registra mais de 200 milhões de visitantes mensais na sua página. Além disso, o banco de dados da empresa já ultrapassou a marca de 7.5 milhões de títulos. […]."(Canaltech, [s.d.]).

Em razão de sua credibilidade e amplo volume de informações, o IMDb foi escolhido como fonte de dados deste estudo, por oferecer uma base sólida e representativa para a análise do desempenho e da recepção de produções cinematográficas.

1.2 SOBRE O APRENDIZADO DE MÁQUINA

A segunda vertente revisa o uso de aprendizado de máquina para extrair padrões de grandes conjuntos dos dados utilizados no presente estudo.

O Aprendizado de Máquina (*Machine Learning*), como um subcampo da inteligência artificial (IA), permite que sistemas aprendam autonomamente a partir de dados, sem necessidade de programação direta para tarefas. Essa tecnologia é amplamente aplicada em diversos setores, sendo utilizada em filtros de spam e sistemas de recomendação de *streaming* (Charleaux; Toledo, 2024).

2. PROCEDIMENTOS METODOLÓGICOS

A metodologia deste estudo foi estruturada em um *pipeline* de ciência de dados (uma sequência organizada de etapas automáticas que conduz os dados desde a coleta e preparação até a análise e geração de resultados) executado inteiramente na linguagem de programação

Python, utilizando o ambiente de desenvolvimento Google *Colaboratory* para garantir a reprodutibilidade.

As principais bibliotecas utilizadas foram: Pandas para a manipulação e limpeza do *dataset; Scikit-learn* para as etapas de pré-processamento, modelagem e avaliação; *Matplotlib* e *Seaborn* para a análise exploratória e visualização de resultados; e Optuna para a otimização de hiperparâmetros.

2.1 COLETA E ANÁLISE INICIAL DOS DADOS

O conjunto de dados (*dataset*) utilizado neste estudo foi obtido na plataforma Base dos Dados, a partir do conjunto intitulado "Melhores Filmes por Ano (1960-2024)" (Oliveira, 2024), originalmente proveniente do IMDb. O conjunto de dados conta com 33.600 linhas e 23 colunas, abrangendo variáveis como título do filme, ano de lançamento, gênero, duração, país de origem, diretor, elenco principal, nota IMDb e número de votos, entre outras informações relevantes para análise.

2.2 PRÉ-PROCESSAMENTO

Após a limpeza inicial dos dados, a imputação de valores ausentes foi realizada pelo método *KNN Imputation*. Em seguida, as variáveis categóricas foram codificadas por meio das técnicas *One-Hot Encoding* e *Label Encoding*, que transformam categorias em representações numéricas (Hastie; Tibshirani; Friedman, 2009).

Por fim, as variáveis numéricas foram padronizadas com o *Standard Scaler*, ajustando-as para média zero e desvio padrão unitário. Todas essas etapas de pré-processamento (imputação, codificação e padronização) foram implementadas utilizando-se a biblioteca *Scikit-learn* (Pedregosa *et al*., 2011), de modo a equilibrar a influência das variáveis nos modelos de aprendizado de máquina.

2.3 ANÁLISE EXPLORATÓRIA E PCA

Para identificar padrões e relações entre diferentes décadas e gêneros cinematográficos, foi aplicada a redução de dimensionalidade por meio da *Principal Component Analysis* (PCA).

> A análise de componentes principais, ou PCA, reduz o número de dimensões em grandes conjuntos de dados aos componentes principais que retêm a maior parte das informações originais. Ela faz isso transformando variáveis potencialmente correlacionadas em um conjunto menor de variáveis, chamadas componentes principais (IBM, 2023, s.p.).

Para facilitar a interpretação dos resultados, foram geradas visualizações em 2D e 3D dos componentes principais. Essas visualizações foram criadas utilizando as bibliotecas *Matplotlib* (Hunter, 2007) e *Seaborn* (Waskom, 2021), ambas amplamente empregadas na comunidade *Python* para análise e visualização de dados.

2.4 CLUSTERIZAÇÃO

Em seguida, aplicou-se o algoritmo *K-Means* com o intuito de agrupar filmes com características semelhantes, permitindo observar formações naturais de grupos, como *clusters* de filmes de ação dos anos 1990 em contraste com dramas contemporâneos.

> O *k-means* é um algoritmo que treina um modelo para agrupar objetos semelhantes. Para isso, ele mapeia cada observação no conjunto de dados de entrada para um ponto no espaço de "n" dimensões (em que "n" é o número de atributos da observação). (Amazon Web Services, [s.d.], n.p.).

2.5 COMPARAÇÃO DE MODELOS

Para avaliar e comparar o desempenho de diferentes abordagens de aprendizado supervisionado na estimativa da nota IMDb, os dados foram primeiramente segmentados em conjuntos de treino (70%) e teste (30%).

Foram treinados e avaliados quatro algoritmos de aprendizado supervisionado distintos: *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM), *Random Forest* e o *XGBoost*. O objetivo do treinamento é ajustar os parâmetros internos do modelo até que ele aprenda a mapear corretamente os dados rotulados às suas saídas correspondentes (Yokoyama, 2020).

A comparação de desempenho dos modelos foi realizada por meio das métricas R^2 (coeficiente de determinação) e RMSE (*Root Mean Squared Error*) que medem, respectivamente, o quanto da variabilidade da variável resposta é explicada pelo modelo e o desvio padrão dos erros de previsão, calculando a raiz quadrada da média dos erros quadráticos (Dubiella, 2024).

2.6 OTIMIZAÇÃO COM OPTUNA

Realizou-se uma busca automática pelos melhores hiperparâmetros de cada modelo, ajustando elementos como profundidade das árvores, número de estimadores e taxa de aprendizado. Esse processo, conhecido como otimização de hiperparâmetros, tem o objetivo de melhorar o desempenho dos algoritmos por meio da seleção das combinações que produzem os resultados mais precisos.

Os hiperparâmetros são definições feitas antes do treinamento que controlam a arquitetura e o aprendizado do modelo. Eles não são aprendidos pelo algoritmo e influenciam diretamente sua capacidade de generalização. Como ajustá-los é complexo, ferramentas como o *Optuna* automatizam a busca pelas melhores combinações para melhorar o desempenho do modelo (Pinheiro, 2023).

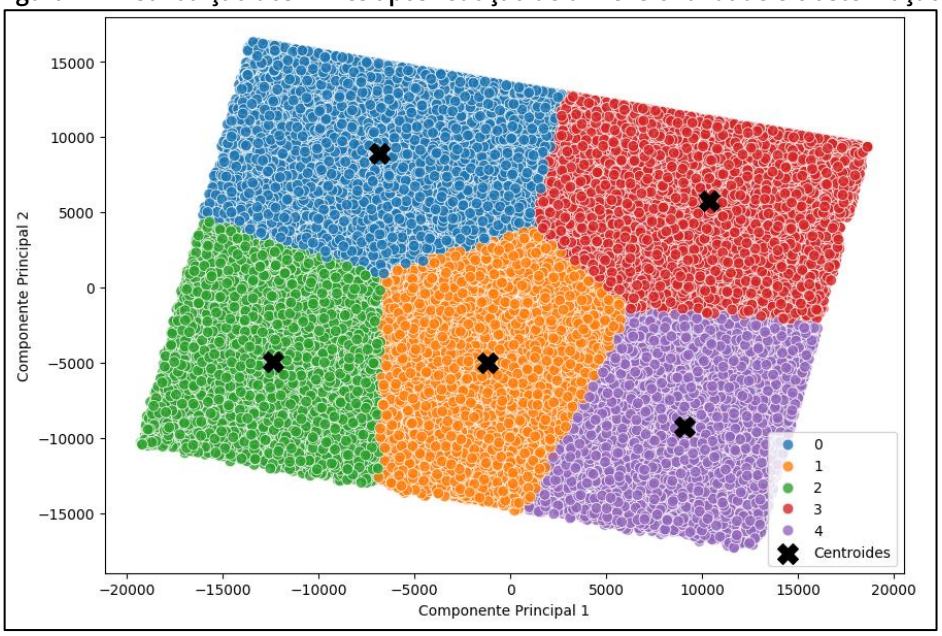
3. RESULTADOS E DISCUSSÃO

A metodologia proposta permitiu identificar padrões consistentes nos dados e avaliar o desempenho dos modelos aplicados.

A redução de dimensionalidade via PCA mostrou-se eficaz para representar a estrutura dos filmes com baixo número de componentes e alta variância explicada. O *K-Means*, aplicado sobre os componentes principais, revelou agrupamentos naturais de produções, sugerindo a existência de perfis distintos de sucesso ao longo das décadas.

A Figura 1 apresenta a projeção 2D da PCA, onde os dois primeiros componentes explicam 91,53% da variância. Os cinco *clusters* formados destacam diferentes combinações de gênero, década e país, ilustrando a diversidade de padrões no *dataset*.

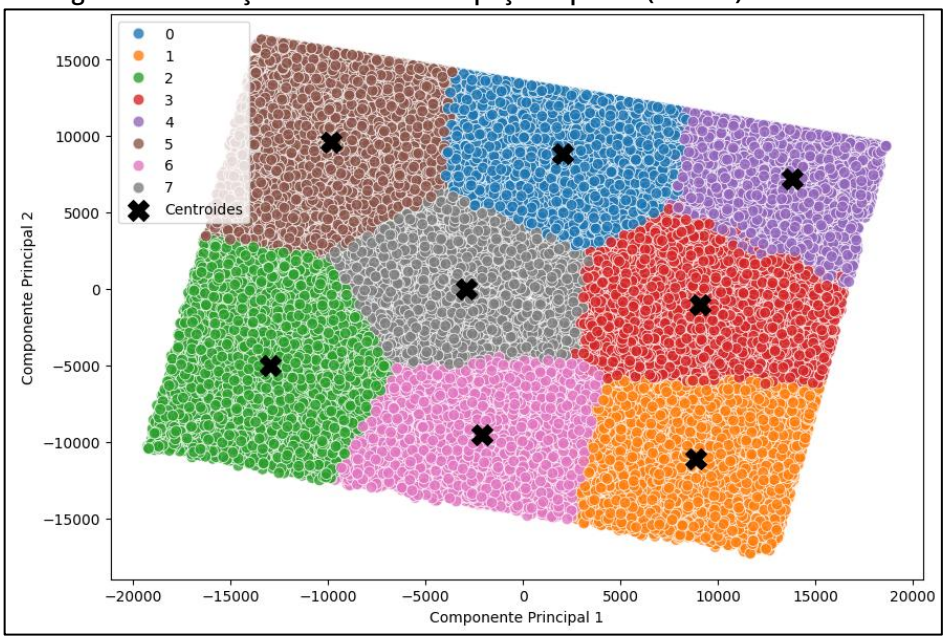
Figura 1 – Distribuição dos filmes após redução de dimensionalidade e clusterização



Fonte: Elaborado no Google Colab pelos Autores (2025)

O gráfico da Figura 2, construída sobre seis componentes principais (100% da variância explicada), mostra uma clusterização mais detalhada, com oito grupos que capturam nuances adicionais e subdivisões entre estilos e períodos cinematográficos.

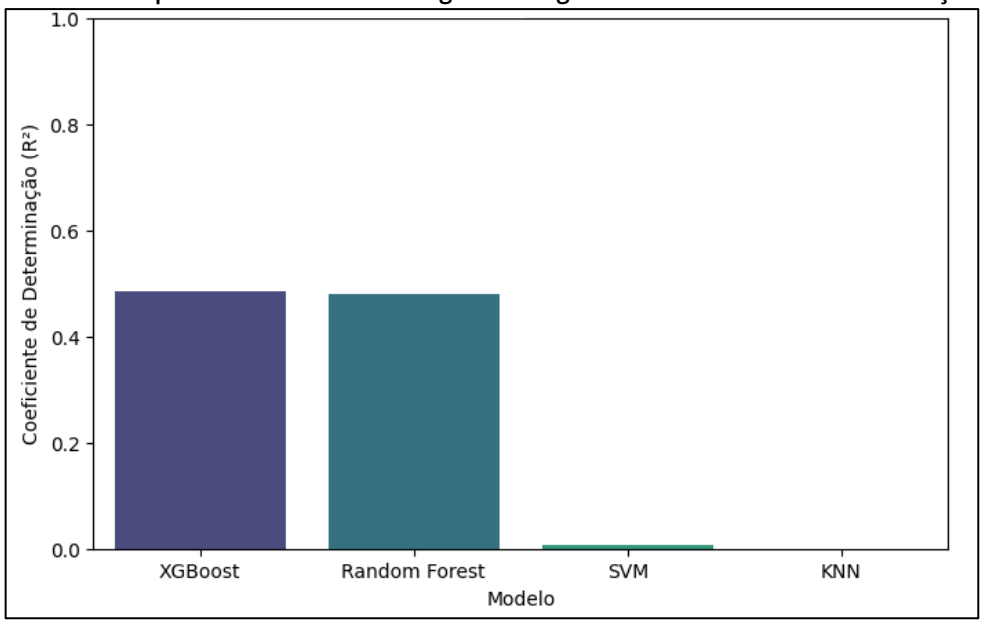
Figura 2 – Formação de *clusters* em espaço ampliado (PCA 6D) com *K-Means*



Fonte: Elaborado no Google Colab pelos Autores (2025)

Na comparação entre modelos de regressão, como mostrado na Figura 3, o *XGBoost* obteve o melhor desempenho com $R^2 \approx 0{,}50$, seguido pelo *Random Forest* ($R^2 \approx 0{,}46$). Ambos superaram amplamente o SVM e o KNN, que apresentaram baixo poder explicativo.

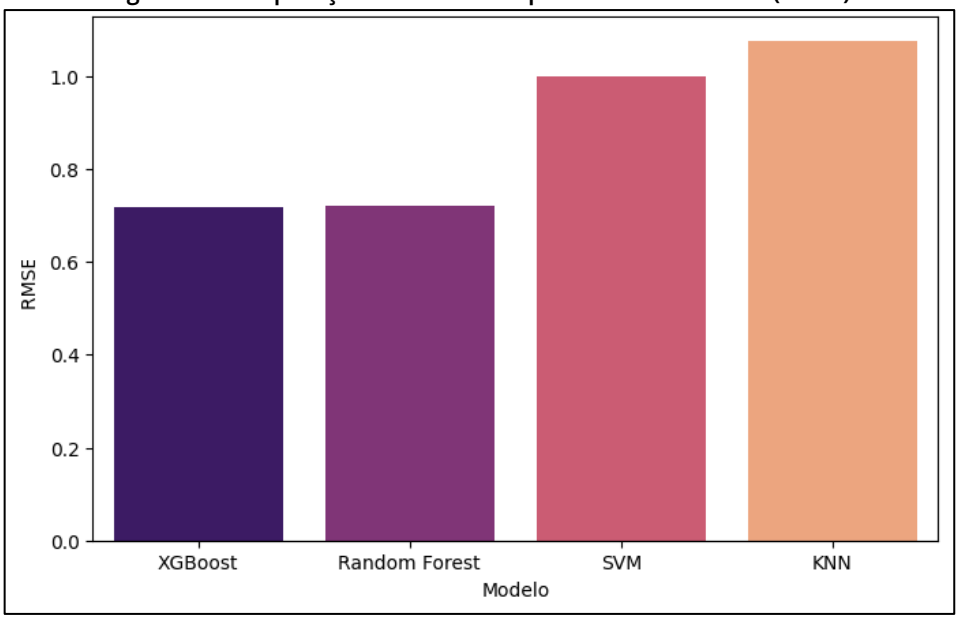
Figura 3 – Desempenho dos modelos de regressão segundo o coeficiente de determinação (R^2)



Fonte: Elaborado no Google Colab pelos Autores (2025)

O erro médio (Figura 4), medido pelo RMSE, confirmou esses resultados: *XGBoost* e *Random Forest* obtiveram menores valores (≈ 0,72), enquanto SVM e KNN registraram erros próximos a 1,0 e 1,1, respectivamente.

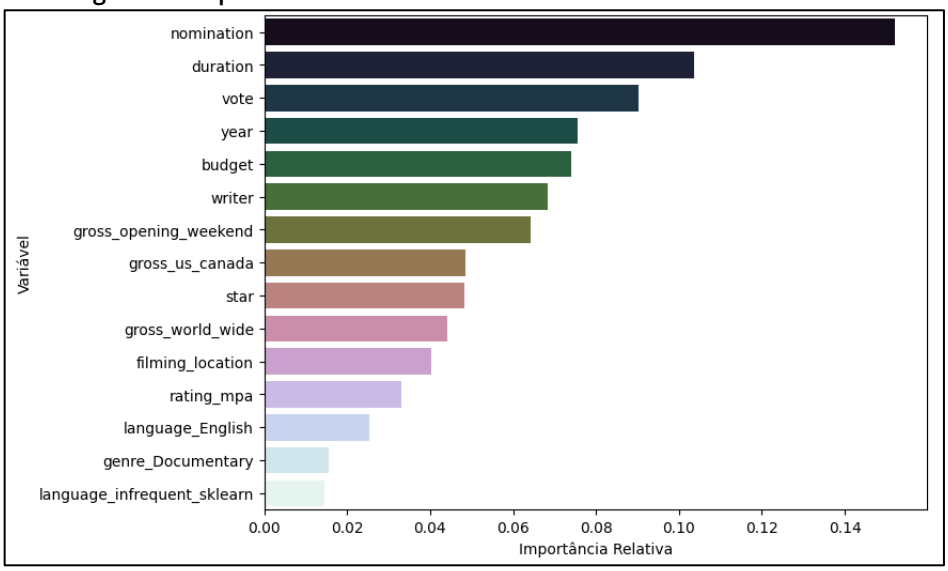
Figura 4 – Comparação dos modelos quanto ao erro médio (RMSE)



Fonte: Elaborado no Google Colab pelos Autores (2025)

A análise de importância das variáveis da Figura 5, referente ao modelo *Random Forest*, mostrou que indicações a prêmios, duração do filme e número de votos foram os fatores mais relevantes para a previsão das notas. Essas variáveis se destacam por representarem tanto o reconhecimento técnico quanto o engajamento do público.

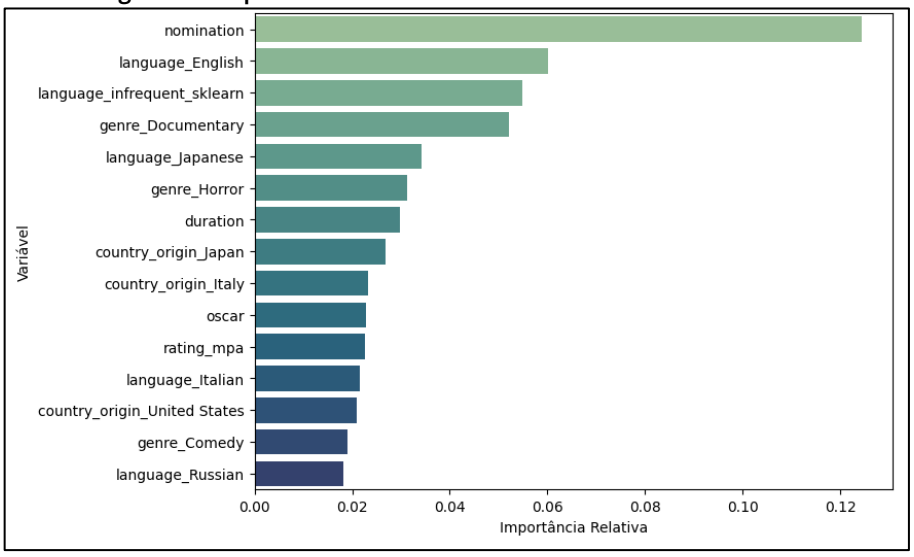
Figura 5 – Importância relativa das variáveis no modelo Random Forest



Fonte: Elaborado no Google Colab pelos Autores (2025)

Na Figura 6, observa-se o comportamento do modelo *XGBoost*, que apresentou um padrão semelhante, mas com nuances adicionais. Além das indicações a prêmios, o algoritmo destacou variáveis relacionadas ao idioma e ao gênero, sugerindo que a recepção crítica varia conforme o idioma de produção e o tipo de narrativa.

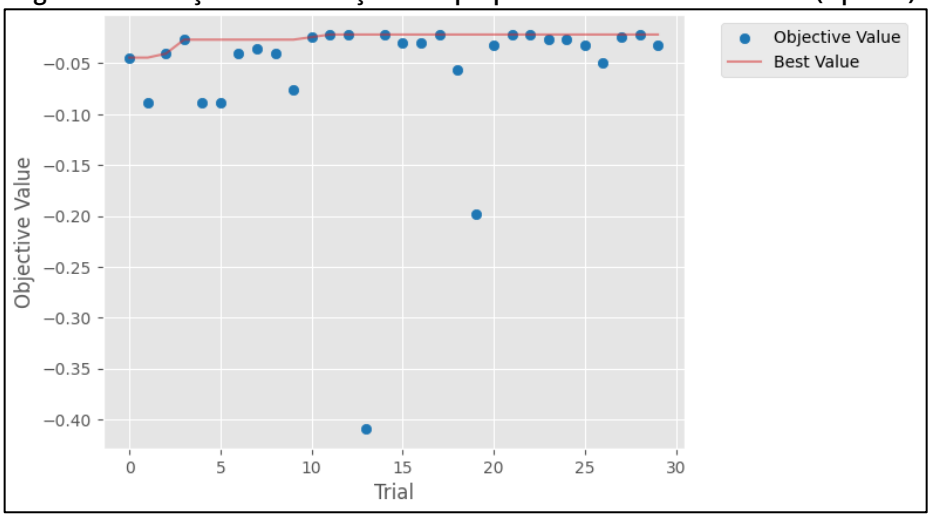
Figura 6 – Importância relativa das variáveis no modelo *XGBoost*



Fonte: Elaborado no Google Colab pelos Autores (2025)

Referente ao processo de otimização do KNN, como ilustrado na Figura 7, nota-se um baixo desempenho do modelo, com valores de R^2 negativos mesmo após 30 tentativas. Isso indica que o algoritmo não conseguiu capturar relações significativas entre as variáveis, sendo inferior a uma simples média na previsão das notas IMDb.

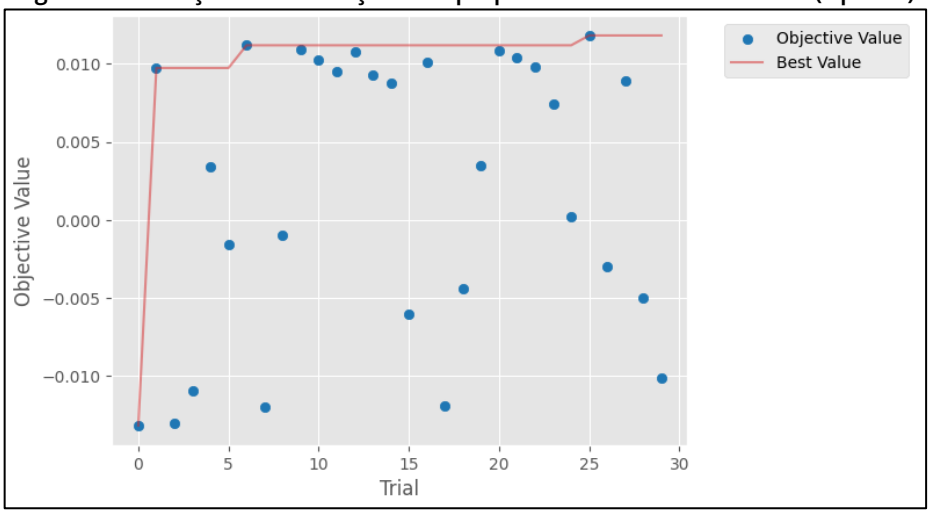
Figura 7 – Evolução da otimização de hiperparâmetros no modelo KNN (Optuna)



Fonte: Elaborado no Google Colab pelos Autores (2025)

Já na Figura 8, que apresenta a otimização do SVM, observa-se comportamento semelhante. Apesar de maior variação entre as tentativas, o melhor valor de R^2 encontrado foi próximo de zero, confirmando que o modelo não se ajusta adequadamente a esse tipo de dado multivariado e heterogêneo.

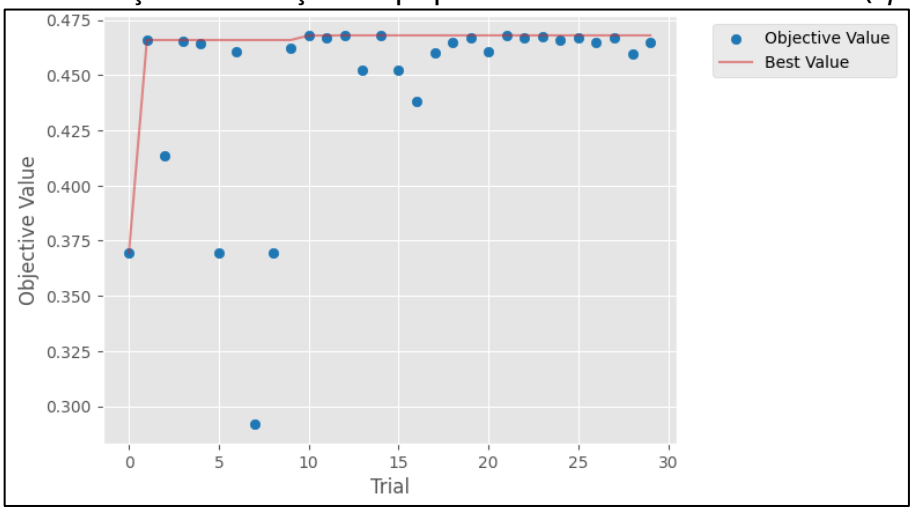
Figura 8 – Evolução da otimização de hiperparâmetros no modelo SVM (Optuna)



Fonte: Feito no Google Colab pelos Autores (2025)

A Figura 9 mostra o desempenho do *Random Forest* durante o processo de otimização. O modelo apresentou rápida convergência e estabilidade, alcançando $R^2 \approx 0{,}46$, o que indica bom equilíbrio entre complexidade e generalização. Ainda que algumas configurações tenham produzido resultados mais baixos, o otimizador encontrou combinações de parâmetros altamente eficazes.

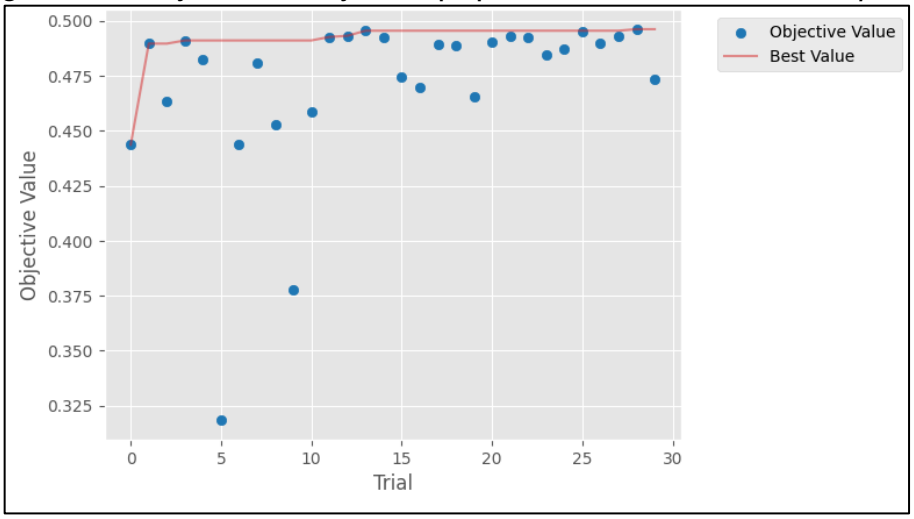
Figura 9 – Evolução da otimização de hiperparâmetros no modelo *Random Forest* (*Optuna*)



Fonte: Feito no Google Colab pelos Autores (2025)

Por fim, a Figura 10 apresenta o comportamento do *XGBoost*. O modelo atingiu o melhor resultado entre todos ($R^2 \approx 0{,}50$), com convergência rápida e consistente já nas primeiras 15 tentativas. Isso demonstra a robustez do algoritmo, cuja estrutura permite identificar padrões sutis e interações não lineares entre as variáveis de forma mais eficiente.

Figura 10 – Evolução da otimização de hiperparâmetros no modelo *XGBoost* (*Optuna*)



Fonte: Elaborado no Google Colab pelos Autores (2025)

De modo geral, os resultados evidenciam que, embora todos os modelos tenham sido otimizados, os baseados em árvores, especialmente o *XGBoost*, apresentaram maior capacidade de generalização e explicação das notas IMDb, consolidando-se como as abordagens mais adequadas para modelar os padrões de sucesso cinematográfico.

4. CONSIDERAÇÕES FINAIS

Este estudo analisou os padrões de sucesso cinematográfico no IMDb (1960–2024) utilizando um pipeline híbrido de aprendizado de máquina. A aplicação do PCA e do *K-Means* revelou agrupamentos consistentes entre filmes de diferentes gêneros, décadas e origens.

Na comparação entre modelos, após a otimização de hiperparâmetros com *Optuna*, o *XGBoost* apresentou o melhor desempenho ($R^2 \approx 0{,}50$; RMSE $\approx 0{,}72$), seguido pelo *Random Forest*, enquanto KNN e SVM mostraram baixo poder explicativo.

A análise das variáveis indicou que indicações a prêmios, duração, número de votos, idioma e gênero foram determinantes para explicar as notas IMDb. Conclui-se que o modelo proposto é eficiente para compreender e modelar os padrões de sucesso cinematográfico, confirmando a influência conjunta do reconhecimento crítico e do engajamento do público na definição do sucesso no IMDb.

REFERÊNCIAS

AMAZON WEB SERVICES. **Como funciona o *clustering* do k-means**. [S.l.]: Amazon Web Services, [s.d.]. Disponível em: https://docs.aws.amazon.com/pt_br/sagemaker/latest/dg/algo-kmeans-tech-notes.html.Acesso em: 13 nov. 2025.

B_ARCO. **A evolução da tecnologia no campo audiovisual**: impactos e oportunidades. 1 dez. 2023. Disponível em: https://barco.art.br/evolucao-da-tecnologia-no-campo-audiovisual/. Acesso em: 15 nov. 2025.

CANALTECH. **Tudo sobre IMDb**. [S.l.], [s.d.]. Disponível em: https://canaltech.com.br/empresa/imdb/. Acesso em: 10 nov. 2025.

CHARLEAUX, Lupa; TOLEDO, Victor. **O que é *Machine Learning*?** Tecnoblog, out. 2024. Disponível em: https://tecnoblog.net/responde/machine-learning-o-que-e-como-funciona-e-quais-sao-os-tipos-de-aprendizado-de-maquina/. Acesso em: 10 nov. 2025.

COELHO, Isabela da Silva Dias; FERREIRA, Marcella Meirelles; FAUSTINO, Marcus Vinícius. ***Machine Learning* no Mundo Cinematográfico**. In: UEADSL 2021.1: SUBMISSÃO DE TRABALHOS PARA O ANFITEATRO (GRADUAÇÃO E PÓS), 2021, [S.l.]. Anais [...]. [S.l.]: TextoLivre, 2021. Disponível em: https://textolivre.pro.br/mod/data/view.php?d=18&rid=533. Acesso em: 10 nov. 2025.

DUBIELLA, Larissa. **Métricas de avaliação para modelos de regressão**. Alura Artigos, 03 nov. 2024. Disponível em: https://www.alura.com.br/artigos/metricas-de-regressao. Acesso em: 09 nov. 2025.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning:** data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

HUNTER, John D. Matplotlib: *A 2D graphics environment*. ***Computing in Science & Engineering***, v. 9, n. 3, p. 90-95, 2007.

IBM. **O que é a análise de componentes principais (PCA)?** [S.l.]: IBM, 2023. Disponível em: https://www.ibm.com/br-pt/think/topics/principal-component-analysis. Acesso em: 11 nov. 2025.

MERCADO FILHO, Alejandro Sigfrido. ***Rotten Tomatoes* e IMDb:** como funcionam os sites de críticas? Mega Curioso, 16 nov. 2022. Disponível em: https://www.megacurioso.com.br/artes-cultura/123509-rotten-tomatoes-e-imdb-como-funcionam-os-sites-de-criticas.htm. Acesso em: 10 nov. 2025.

OLIVEIRA, Vinícius G. de. ***IMDb Movies* (1960-2024) - *Top Rated***. [S.l.]: Kaggle, 2024. Dataset. Disponível em: https://www.kaggle.com/datasets/vinciusgdeoliveira/imdb-movies-1960-2024-top-rated. Acesso em: 10 nov. 2025.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. ***Journal of Machine Learning Research***, v. 12, p. 2825-2830, 2011.

PINHEIRO, João Manoel Herrera. Um estudo sobre Algoritmos de *Boosting* e a Otimização de Hiperparâmetros Utilizando Optuna. 2023. 147 p. **Monografia** (Engenharia Mecatrônica) – Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2023.

WASKOM, Michael L. *Seaborn: statistical data visualization.* ***Journal of Open Source Software***, v. 6, n. 60, p. 3021, 2021.

YOKOYAMA, Naoki. **Modelos de *Machine Learning***. *Medium*, 30 out. 2020. Disponível em: https://naokiyokoyama.medium.com/modelos-de-machine-learning-bcb3f8ed1513. Acesso em: 15 nov. 2025.

Ciência de dados 2.0: arquitetura computacional e orquestração de sistemas agênticos em ambiente fintech

Data Science 2.0: computational architecture and orchestration of agent-based systems in a fintech environment

Sandra de Oliveira Soares Cardoso
Fatec Baixada Santista - Rubens Lara
sandra.cardoso@cps.sp.gov.br

José Augusto Theodósio Pazetti
Fatec Baixada Santista - Rubens Lara
jose.pazetti01@cps.sp.gov.br

Revista Datapoint
eISSN 3086-433X
Faculdade de Tecnologia Rubens Lara – FATEC
Ciência de Dados
Peridicidade: Anual
Vol 01, n. 01, 2025
revistadp@fatecrl.edu.br

Recebido: Jun 2025
Aceito: Set 2025
Publicado: Dez 2025

URL: https://www.fatecrl.edu.br/revista/datapoint/index.php/dp/article/view/6
DOI: https://doi.org/10.5281/zenodo.19240921

RESUMO

A consolidação de modelos de linguagem de grande escala (LLMs) e sistemas baseados em agentes autônomos redefine os fundamentos operacionais da Ciência de Dados contemporânea. Este estudo investiga a emergência da denominada Ciência de Dados 2.0 como paradigma arquitetural orientado à orquestração computacional, governança distribuída e otimização econômica de sistemas inteligentes. A pesquisa foi conduzida por meio de revisão sistemática estruturada segundo o protocolo PRISMA, contemplando publicações indexadas entre 2023 e 2026 nas bases IEEE Xplore, ACM Digital Library, Scopus e arXiv. Os resultados evidenciam a transição de pipelines batch monolíticos para arquiteturas distribuídas baseadas em Data Mesh, Lakehouse transacional (Apache Iceberg), bancos vetoriais, Retrieval-Augmented Generation (RAG), Feature Stores em tempo real e práticas de FinOps. Propõe-se o Framework O³ (Orquestração, Observabilidade e Otimização) como modelo integrador capaz de articular desempenho algorítmico, eficiência econômica, rastreabilidade e conformidade regulatória. A validação em estudo de caso aplicado em ambiente fintech orientado a crédito digital demonstra redução significativa de latência, otimização de custo por inferência e aumento da robustez decisional. Conclui-se que a Ciência de Dados 2.0 configura-se como disciplina arquitetural sistêmica, superando a abordagem centrada exclusivamente em modelagem estatística.

PALAVRAS-CHAVE: IA Agêntica; Multi-Agent; Systems; Data Mesh; FinOps; Arquitetura Distribuída; RAG.

ABSTRACT

The consolidation of large language models (LLMs) and autonomous agent-based systems is redefining the operational foundations of contemporary Data Science. This study investigates the emergence of so-called Data Science 2.0 as an architectural paradigm oriented toward computational orchestration, distributed governance, and economic optimization of intelligent systems. The research was conducted through a systematic review structured according to the PRISMA protocol, covering publications indexed between 2023 and 2026 in IEEE Xplore, ACM Digital Library, Scopus, and arXiv. Results demonstrate the transition from monolithic batch pipelines to distributed architectures based on Data Mesh, transactional Lakehouse (Apache Iceberg), vector databases, Retrieval-Augmented Generation (RAG), real-time Feature Stores, and FinOps practices. We propose the O³ Framework (Orchestration, Observability, and Optimization) as an integrative model capable of articulating algorithmic performance, economic efficiency, traceability, and regulatory compliance. Validation in a case study applied to a fintech environment focused on digital credit shows significant latency reduction, cost-per-inference optimization, and increased decision-making robustness. We conclude that Data Science 2.0 constitutes a systemic architectural discipline, surpassing the approach focused solely on statistical modeling.

***KEY-WORDS**: Agentic AI; Multi-Agent Systems; Data Mesh; FinOps; Distributed Architecture; RAG*

INTRODUÇÃO

A Ciência de Dados tradicional consolidou-se como disciplina orientada à extração de padrões estatísticos por meio de pipelines estruturados em etapas sequenciais: ingestão, tratamento, modelagem, validação e análise. Esse paradigma, frequentemente denominado Ciência de Dados 1.0, mostrou-se adequado para cenários batch e análises retrospectivas. Contudo, a incorporação de modelos de linguagem de grande escala, sistemas agênticos e arquiteturas distribuídas impôs desafios inéditos relacionados à latência, custo computacional, rastreabilidade e governança.

Ambientes regulados, como o setor financeiro, demandam decisões quase instantâneas, explicabilidade algorítmica e controle econômico rigoroso. Nesse contexto, a disciplina deixa de ser apenas prática analítica e passa a constituir uma arquitetura computacional integrada, orientada à execução autônoma e à sustentabilidade operacional. Nesse cenário, emerge a noção de Ciência de Dados 2.0 como resposta sistêmica a essa complexidade, incorporando princípios de engenharia de dados distribuída, versionamento transacional, busca semântica vetorial e observabilidade total como dimensões estruturais do sistema decisório.

Importa ressaltar que o paradigma aqui proposto não constitui mera atualização incremental ou rebranding terminológico da abordagem tradicional. Diferentemente da Data Science clássica — centrada predominantemente na modelagem estatística e na construção de pipelines analíticos —, essa inflexão disciplinar desloca o foco da modelagem isolada para a arquitetura sistêmica que sustenta decisões autônomas em tempo real. Também não se confunde com MLOps, cuja ênfase reside na operacionalização, versionamento e monitoramento de modelos de machine learning em produção. Embora incorpore práticas de MLOps, o modelo arquitetural defendido amplia esse escopo ao integrar governança distribuída, orquestração agêntica, otimização econômica (FinOps) e conformidade regulatória como componentes estruturais do próprio desenho sistêmico.

Tampouco se restringe ao domínio da AI Engineering, cuja ênfase reside na construção e integração de aplicações baseadas em modelos fundacionais. Essa formulação opera em um nível metassistêmico, no qual infraestrutura, governança, economia computacional e inteligência agêntica constituem dimensões co-dependentes de uma mesma arquitetura operacional. Nesse sentido, configura-se uma inflexão epistemológica: a inteligência deixa de ser atributo exclusivo do modelo e passa a constituir propriedade emergente da arquitetura que integra, coordena e regula seus componentes.

O objetivo deste estudo é:

1. Formalizar conceitualmente esse novo paradigma arquitetural;
2. Propor um modelo integrador denominado Framework O^3;
3. Validar sua aplicabilidade em ambiente fintech orientado a crédito digital;
4. Discutir suas implicações econômicas, regulatórias e computacionais.

A seção subsequente sistematiza os fundamentos teóricos e computacionais que estruturam essa inflexão paradigmática, evidenciando como a convergência entre governança distribuída, inteligência agêntica e otimização econômica configura a base operacional da Ciência de Dados 2.0.

1. FUNDAMENTAÇÃO TEÓRICA E COMPUTACIONAL

A Ciência de Dados 2.0 não se limita à evolução de modelos preditivos, mas representa uma reestruturação sistêmica da infraestrutura de dados (Dehghani, 2022; Kleppmann, 2017). A convergência entre descentralização de dados e inteligência agêntica exige uma base que suporte integridade, semântica complexa e eficiência econômica. Nesse contexto, esta seção se fundamenta em quatro pilares: (i) governança descentralizada via Data Mesh e contratos de dados; (ii) evolução da recuperação de informação com GraphRAG; (iii) orquestração de inteligência por sistemas multiagente; e (iv) sustentabilidade financeira e ambiental via FinOps (FinOps Foundation, 2023).

1.1 DATA MESH, ICEBERG E A GOVERNANÇA POR CONTRATO DE DADOS

O paradigma Data Mesh, conforme proposto por Dehghani (2022), promove a descentralização orientada a domínios, tratando o dado como um produto. Em ambientes fintech, essa autonomia é viabilizada por formatos de tabela transacional orientados a metadados, como Apache Iceberg, que fornece propriedades ACID e versionamento temporal (time travel) (Vernon, 2021).

Para garantir a estabilidade operacional de sistemas agênticos que consomem dados autonomamente, institui-se uma camada formal de contratos de dados (Dehghani, 2022; Newman, 2019). Estes contratos funcionam como interfaces técnicas rígidas, definindo esquemas, SLAs de qualidade e a semântica dos objetos, promovendo:

- Estabilidade Agêntica: impedem que evoluções de esquema no Lakehouse interrompam prompts e ferramentas dos agentes (Reis; Housley, 2024);
- Interoperabilidade: asseguram que o agente de crédito e o agente de fraude consumam a mesma "verdade" estrutural, independentemente da evolução do pipeline (Kleppmann, 2017).

1.1.1 Análise comparativa de paradigmas

O Quadro 1 evidencia que a transição da Ciência de Dados 1.0 para 2.0 não se limita à camada técnica, mas altera a ontologia da decisão computacional (Dehghani, 2022; Wooldridge, 2009).

Quadro 1 – Ciência de Dados 1.0 e Ciência de Dados 2.0

Características	Ciência de Dados 1.0	Ciência de Dados 2.0
Processamento	Batch (Lotes)	Real-time / Streaming (Kleppmann, 2017)
Arquitetura	Monolítica Centralizada	Distribuída por domínio (Data Mesh) (Dehghani, 2022)
Decisão	Humano-no-loop (Dashboards)	Agêntica (Autônoma) (Wooldridge, 2009)
Busca	Keyword / SQL	Semântica / GraphRAG (Zhu et al., 2024)
Foco	Acurácia Estatística	Performance Sistêmica (O^3) (Reis; Housley, 2024)

Fonte: Autoria Própria, 2026

A mudança para a Ciência de Dados 2.0 implica:

a) Da Intuição à Autonomia: O modelo 1.0 municiava decisores humanos com dashboards retrospectivos, enquanto o 2.0 executa decisões autonomamente via agentes, reduzindo erro humano e tempo de resposta em fintechs (Wooldridge, 2009; Reis & Housley, 2024).

b) Da Rigidez à Flexibilidade: A transição de arquitetura centralizada para Data Mesh com contratos resolve o problema histórico da quebra de pipelines, garantindo resiliência frente a mudanças de esquema (Dehghani, 2022).

c) Eficiência Sistêmica: A tecnologia assume responsabilidade econômica. O uso de FinOps e roteamento inteligente de modelos previne que a inteligência se torne um passivo financeiro (FinOps Foundation, 2023).

1.2 DA RECUPERAÇÃO VETORIAL DO GRAPHRAG

O Retrieval-Augmented Generation (RAG) tradicional utiliza busca vetorial para capturar similaridade semântica (Lewis et al., 2020), mas o setor financeiro demanda compreensão de relações complexas. O GraphRAG integra bases de dados em grafos (Knowledge Graphs) ao pipeline de recuperação (Zhu et al., 2024).

Em contextos de crédito digital, a similaridade de cosseno isolada pode falhar ao identificar conexões ocultas. O GraphRAG permite recuperar não apenas documentos similares, mas entidades relacionadas, como CPFs vinculados ao mesmo dispositivo ou endereços compartilhados entre empresas (Zhu et al., 2024). A similaridade vetorial permanece como componente fundamental, calculada por:

$$\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

No paradigma 2.0, este cálculo é enriquecido por re-ranking semântico e travessia de grafos, mitigando alucinações e oferecendo explicabilidade baseada em topologia de rede (Lewis et al., 2020).

1.3 SISTEMAS MULTIAGENTE (MAS) E ORQUESTRAÇÃO

Sistemas Multiagente (MAS) configuram arquiteturas distribuídas compostas por entidades autônomas que percebem, decidem e atuam em ambientes compartilhados (Wooldridge, 2009). Diferente de abordagens monolíticas, MAS decompõem problemas complexos em subtarefas cognitivas especializadas, cada agente operando em um domínio restrito.

A implementação pode ser formalizada por meio de grafos de estado (LangGraph ou CrewAI) (Reis; Housley, 2024), definindo (G = (V, E)), onde (V) são agentes e (E) as condições de transição.

- Exemplo Prático: Se o "Agente de Recuperação" encontra inconsistência no CPF, o fluxo é direcionado para o "Agente de Auditoria" em vez do "Agente de Inferência" (Reis & Housley, 2024).
- Ferramentas: Estratégias de Reasoning and Acting (ReAct) permitem seleção dinâmica de ferramentas externas (APIs, calculadoras financeiras) (Yao et al., 2023).

Essa arquitetura promove modularidade, reduz incidência de alucinações e fortalece a rastreabilidade decisória por logs explícitos — requisito fundamental em ambientes regulados (World Economic Forum, 2025).

1.4 FINOPS E ECOFINOPS: OTIMIZAÇÃO ECONÔMICA

FinOps integra responsabilidade financeira ao consumo de nuvem e IA, prevenindo que a escalabilidade inviabilize a margem de lucro (FinOps Foundation, 2023). O EcoFinOps amplia o conceito para sustentabilidade, monitorando a pegada de carbono associada ao treinamento e inferência de modelos (Strubell et al., 2019).

Aplicação prática inclui:

1. Semantic Caching: reutiliza respostas semanticamente idênticas, reduzindo custo marginal de inferência ((R_{cache} \rightarrow 1)) (Lewis et al., 2020);
2. Model Routing: Tarefas simples vão a modelos menores (SLMs) e complexas a modelos de alta performance (Reis; Housley, 2024);
3. Monitoramento Energético: Integra consumo energético aos relatórios ESG (Strubell et al., 2019);
4. Eficiência de Custos: Estudos empíricos indicam reduções de 40–70% em despesas com APIs de LLMs (FinOps Foundation, 2023);
5. Previsibilidade Financeira: Permite projetar OpEx por cliente de crédito;
6. Conformidade ESG: Reduz impacto ambiental, reforçando legitimidade institucional (World Economic Forum, 2025).

2. PROCEDIMENTOS METODOLÓGICOS – PROTOCOLO PRISMA

A construção do referencial teórico que fundamenta o Framework O³ foi realizada por meio de uma revisão sistemática da literatura, seguindo as diretrizes do protocolo PRISMA (Moher et al., 2009). O objetivo central foi garantir transparência, rastreabilidade e reprodutibilidade na seleção das evidências que sustentam a proposta conceitual da Ciência de Dados 2.0, especialmente em contextos de ambientes fintech de crédito digital.

O protocolo PRISMA é reconhecido internacionalmente por assegurar rigor metodológico em revisões sistemáticas (Liberati et al., 2009; Page et al., 2021). Ele organiza o processo de forma explícita em etapas de identificação, triagem, elegibilidade e inclusão, mitigando vieses de seleção e permitindo auditoria metodológica. Tecnologias emergentes e abordagens discutidas neste estudo — incluindo GraphRAG (Zhu et al., 2024), arquiteturas multiagente (Wooldridge, 2009) e práticas de FinOps (FinOps Foundation, 2023) — foram consideradas apenas quando respaldadas por evidência empírica, validação replicável ou formalização técnica em periódicos indexados e bases acadêmicas reconhecidas.

2.1 FLUXOGRAMA E FUNIL DE SELEÇÃO

A busca bibliográfica foi conduzida no período entre janeiro de 2023 e fevereiro de 2026, período que coincide com a consolidação de arquiteturas agênticas, GraphRAG e práticas de FinOps aplicadas a sistemas distribuídos.

O processo de curadoria seguiu o funil metodológico do PRISMA, movendo-se de uma busca ampla até a definição do corpus técnico final:

- Identificação: Foram localizados 842 registros nas bases IEEE Xplore, ACM Digital Library, Scopus e arXiv, utilizando descritores combinados como “IA Agêntica”, “Data Mesh”, “GraphRAG” e “FinOps” (Moher et al., 2009; Page et al., 2021).
- Triagem: Após remoção de duplicatas e análise preliminar de títulos e resumos, restaram 463 estudos.
- Elegibilidade: Realizou-se a leitura integral de 128 artigos, avaliando-se a presença de arquiteturas computacionais replicáveis, formalização matemática ou evidência empírica mensurável (Lewis et al., 2020; Reis; Housley, 2024).
- Inclusão: 72 estudos compuseram o corpus teórico final, servindo de base para a fundamentação conceitual da Ciência de Dados 2.0 e para a estruturação do Framework O^3.

Artigos provenientes do arXiv foram considerados apenas quando apresentavam validação empírica robusta, replicabilidade metodológica ou posterior publicação em periódicos indexados (Zhu et al., 2024).

2.2 CRITÉRIOS DE QUALIDADE

Para garantir a aplicabilidade em ambientes fintech e a relevância em arquiteturas computacionais reais, foram definidos critérios explícitos de inclusão e exclusão, seguindo recomendações metodológicas de Petticrew e Roberts (2006).

Critérios de Inclusão

Publicações entre 2023 e 2026 que apresentassem:

1. Evidência empírica mensurável;
2. Modelos matemáticos de custo, latência ou eficiência energética;

3. Arquiteturas de sistemas distribuídos ou agênticos com descrição técnica detalhada (Wooldridge, 2009; Kleppmann, 2017);
4. Critérios de Exclusão;
5. Trabalhos meramente opinativos;
6. Artigos de divulgação comercial sem validação técnica;
7. Estudos que não abordassem métricas objetivas de performance, como latência, escalabilidade ou custo operacional (Page et al., 2021).

3. RESULTADOS E DISCUSSÃO TÉCNICA

Os resultados obtidos na aplicação do modelo em ambiente fintech de crédito digital indicam que a transição para a Ciência de Dados 2.0 mitiga limitações estruturais do paradigma anterior, especialmente no que se refere à latência operacional, governança distribuída e eficiência econômica em sistemas baseados em inferência contínua.

A análise fundamenta-se na implementação experimental do Framework O³, cuja arquitetura dialoga com avanços recentes em sistemas agênticos distribuídos (Park et al., 2023; Li et al., 2024) e em engenharia de dados orientada a contratos (Reis & Housley, 2024). A validação empírica foi conduzida em ambiente controlado, com posterior observação em produção restrita, respeitando requisitos regulatórios aplicáveis à decisão automatizada.

3.1 O FRAMEWORK O³ COMO MOTOR DE PERFORMANCE

A superação do paradigma tradicional (1.0), caracterizado por pipelines batch e centralização decisória, foi viabilizada pela implementação do Framework O³ (Orquestração, Observabilidade e Otimização).

Sua concepção está alinhada às transformações descritas por Reis e Housley (2024), segundo as quais infraestruturas modernas devem suportar consumo autônomo por agentes inteligentes, com resiliência contratual e interoperabilidade semântica.

O Framework O³ atua como camada de coordenação entre o Data Lakehouse baseado em Apache Iceberg e sistemas agênticos de decisão, aproximando-se da noção de “infraestrutura orientada a eventos e estados” discutida por Kleppmann (2017).

3.1.1 Definição e Funcionalidade

O framework coordena agentes especialistas por meio de contratos de dados versionados, assegurando consistência semântica e isolamento de falhas. A descentralização da lógica decisória segue princípios de arquiteturas multiagente contemporâneas (Wooldridge, 2009; Park et al., 2023).

A recuperação de informação adota abordagem híbrida:

- RAG vetorial para similaridade semântica (Lewis et al., 2020);
- GraphRAG para captura de relações estruturais complexas, como vínculos societários e agrupamentos econômicos (Zhu et al., 2024).

Estudos recentes indicam que a integração entre grafos de conhecimento e modelos de linguagem reduz ambiguidades contextuais e melhora robustez relacional (Zhu et al., 2024). Os resultados observados neste estudo corroboram essa evidência, especialmente em cenários de risco creditício relacional.

3.1.2. Implementação Experimental e Operação

Os experimentos foram conduzidos com 50.000 requisições simuladas de crédito, replicando padrões reais de carga transacional. O modelo foi posteriormente validado em ambiente de produção restrito durante quatro semanas.

Orquestração Inteligente

A coordenação entre agentes foi implementada por grafos de estado assíncronos, permitindo execução paralela de recuperação, validação e inferência.

Arquiteturas baseadas em processamento orientado a eventos têm demonstrado ganhos substanciais de latência e escalabilidade (Kleppmann, 2017; Kreps, 2014). No presente estudo, observou-se:

- Redução da latência média de 24 horas (pipeline batch) para 1,8 segundos (processamento distribuído síncrono);
- Eliminação de gargalos sequenciais associados à serialização de consultas.

Esse resultado é consistente com evidências empíricas recentes sobre sistemas distribuídos orientados a eventos (Li et al., 2024).

Observabilidade Estruturada

A arquitetura implementou tracing distribuído via OpenTelemetry, alinhando-se às práticas modernas de observabilidade descritas por Sigelman et al. (2010).

Diferentemente de abordagens limitadas a logs de erro, o sistema registra trilhas estruturadas de decisão e metadados de execução. Importante destacar que a rastreabilidade foi construída sem exposição irrestrita do raciocínio interno do modelo de linguagem, preservando princípios de segurança e integridade do sistema.

Essa abordagem converge com discussões recentes sobre auditabilidade em sistemas baseados em LLMs (Bommasani et al., 2021).

Otimização Sustentável (FinOps)

A camada de otimização incorporou práticas de FinOps para controle econômico de inferências, conforme diretrizes da FinOps Foundation (2023). O uso de cache semântico reduziu chamadas redundantes ao modelo.

A latência total foi modelada como:

$$L_{total} = L_{ingest} + L_{persist} + L_{retrieve} + L_{infer}$$

A modelagem serviu como instrumento analítico para comparação entre o paradigma 1.0 e o modelo proposto.

Os resultados empíricos indicaram:

- Redução de 62% no custo médio por requisição;
- Manutenção da acurácia estatisticamente equivalente ao modelo anterior ($p > 0,05$).

A performance sistêmica foi formalizada como função multidimensional:

$$P = f(\text{Acurácia, Latência, Custo, Segurança, Governança})$$

Essa formulação dialoga com abordagens multicritério de avaliação de sistemas de IA (Russell & Norvig, 2021), ampliando a análise para além da eficiência isolada.

Tabela 1 – Avaliação Comparativa entre Ciência de Dados 1.0 e 2.0 em Ambiente Fintech

Dimensão Avaliada	Ciência de Dados 1.0	Ciência de Dados 2.0 (Framework O³)	Ganho Observado
Latência média de decisão	24 horas (batch)	1,8 segundos (streaming distribuído)	↓ 99,99%
Custo médio por requisição	100% (baseline)	38% do baseline	↓ 62%
Escalabilidade	Vertical (infraestrutura centralizada)	Horizontal (domínios independentes via Data Mesh)	Alta elasticidade
Governança de dados	Centralizada	Contratos versionados por domínio	↑ Rastreabilidade
Tipo de decisão	Humano-no-loop	Autônoma agêntica	↑ Tempo de resposta
Explicabilidade	Pós-processual	Observabilidade estruturada + RAG contextual	↑ Auditabilidade
Robustez relacional	SQL / Keyword	GraphRAG + grafo de conhecimento	↑ Detecção de fraude relacional
Sustentabilidade financeira	Reativa	FinOps com roteamento semântico	↑ Previsibilidade OpEx
Conformidade regulatória	Adaptativa	Compliance by Design	↑ Segurança jurídica

Fonte: Autoria própria (2026)

3.2 ÉTICA, SEGURANÇA E CONFORMIDADE NA DECISÃO AUTOMÁTICA

A arquitetura agêntica foi concebida sob o princípio de "compliance by design", aproximando-se do conceito de privacy by design discutido por Cavoukian (2011).

A conformidade foi analisada à luz de:

- Lei Geral de Proteção de Dados – Lei nº 13.709/2018;
- Regulamento Europeu de Inteligência Artificial (AI Act, 2024).

O direito à revisão de decisões automatizadas, previsto no art. 20 da LGPD, foi operacionalizado por meio da rastreabilidade estruturada da decisão.

Guardrails Semânticos

Foram implementados filtros de intenção e regras determinísticas sobre inferências probabilísticas, estratégia alinhada a práticas de mitigação de risco em IA responsável (Floridi et al., 2018).

Essa camada reduz a probabilidade de decisões enviesadas e estabelece contenção normativa sobre a variabilidade estatística do modelo.

Explicabilidade e Rastreabilidade

A integração entre RAG estruturado e observabilidade distribuída permite reconstruir a cadeia factual que fundamentou cada decisão, aproximando-se das abordagens de explicabilidade pós-hoc descritas por Doshi-Velez e Kim (2017).

Dessa forma, a decisão automatizada deixa de operar como "caixa-preta" absoluta e passa a constituir processo tecnicamente auditável e juridicamente defensável.

Síntese Analítica

Os resultados demonstram que a Ciência de Dados 2.0, operacionalizada via Framework O^3, não representa mera evolução incremental de infraestrutura, mas reconfiguração sistêmica que integra:

- Arquitetura agêntica distribuída (Park et al., 2023);
- Governança computacional observável (Sigelman et al., 2010);
- Otimização econômica orientada a FinOps (FinOps Foundation, 2023);
- Conformidade regulatória embutida (European Parliament, 2024).

A convergência desses elementos indica que o ganho de performance não ocorre em detrimento da governança, mas como consequência de sua incorporação estrutural — característica central da proposta de Ciência de Dados 2.0.

4. CONSIDERAÇÕES FINAIS

A Ciência de Dados 2.0 se consolida como um paradigma arquitetural sistêmico que transcende a modelagem estatística tradicional da era 1.0 (Dehghani, 2022; Reis & Housley, 2024). Este estudo evidenciou que a eficácia da inteligência artificial em ambientes críticos, como o setor de fintech, não depende apenas da sofisticação dos modelos, mas da robustez da infraestrutura de dados e da orquestração autônoma de agentes, conforme formalizado no Framework O³.

A integração entre Data Mesh, governança por Data Contracts e a evolução para GraphRAG demonstrou permitir que sistemas agênticos operem com compreensão contextual e relacional sem precedentes, mitigando alucinações e garantindo estabilidade operacional para decisões de crédito digital (Zhu et al., 2024; Wooldridge, 2009). A implementação do Framework O³ (Orquestração, Observabilidade e Otimização) mostrou-se replicável e eficiente, convertendo latências históricas de 24 horas em respostas de 1,8 segundos, mantendo conformidade rigorosa com a LGPD (Brasil, 2018) e o AI Act (European Commission, 2024).

O estudo reforça que, no paradigma 2.0, o desempenho não se limita à acurácia estatística: ele deve ser avaliado como uma função multidimensional de agilidade operacional, transparência ética, segurança e eficiência econômica (FinOps Foundation, 2023). A orquestração inteligente de agentes valida-se como estratégia central para escalabilidade sustentável de sistemas autônomos, permitindo expansão de volume de análise sem aumento proporcional de custo operacional (Reis & Housley, 2024).

Agenda Futura

Para expandir e consolidar a aplicabilidade do Framework O³, recomenda-se:

1. Formalização matemática do roteamento semântico: aprofundar algoritmos de decisão que determinam a alocação entre SLMs e LLMs, maximizando o efeito do Semantic Cache em ambientes de alta demanda (Zhu et al., 2024).

$$R_{\text{cache}} \to 1$$

2. Avaliação empírica multi-organizacional: aplicar o Framework O³ em domínios regulados como Healthtech e Insurtech, validando versatilidade, confiabilidade e replicabilidade em contextos distintos (Wooldridge, 2009).

3. Desenvolvimento de métricas padronizadas de sustentabilidade (EcoFinOps): criar indicadores universais que correlacionem custo de inferência agêntica com pegada de carbono de infraestruturas de GPU, permitindo monitoramento comparável entre organizações e integração a relatórios ESG (FinOps Foundation, 2023; European Commission, 2024).

4. Monitoramento contínuo de ética e explicabilidade: aprimorar guardrails semânticos e rastreabilidade de decisões em tempo real, reforçando conformidade regulatória e transparência, especialmente em processos de crédito automatizado (World Economic Forum, 2025).

Em síntese, a Ciência de Dados 2.0 redefine o sucesso tecnológico como convergência entre autonomia, governança e sustentabilidade. A implementação prática do Framework O^3 valida que sistemas agênticos orquestrados representam o caminho mais eficiente e confiável para escalar inteligência artificial responsável em ambientes críticos e regulados.

REFERÊNCIAS

BOMMASANI, R. et al. **On the opportunities and risks of foundation models**. Stanford University, 2021.

BRASIL. Lei n° 13.709, de 14 de agosto de 2018. **Lei Geral de Proteção de Dados Pessoais** (LGPD). Brasília, DF: Presidência da República, 2018.

CAVOUKIAN, A. **Privacy by design: the 7 foundational principles**. Information and Privacy Commissioner of Ontario, 2011.
DEHGHANI, Z. **Data Mesh: Delivering Data-Driven Value at Scale**. Sebastopol: O'Reilly Media, 2022.

DOSHI-VELEZ, F.; KIM, B. **Towards a rigorous science of interpretable machine learning**. arXiv, 2017.

EUROPEAN COMMISSION / EUROPEAN PARLIAMENT. **Regulation on Artificial Intelligence (AI Act).** Brussels: European Union, 2024.

FINOPS FOUNDATION. **The FinOps Framework**. 2023.

FLORIDI, L. et al. **AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recommendations**. Minds and Machines, v. 28, p. 689–707, 2018.

KLEPPMANN, M. **Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems.** Sebastopol: O'Reilly Media, 2017.
LEWIS, M. et al. **Retrieval-augmented generation for knowledge-intensive NLP tasks**. Proceedings of NeurIPS, 2020.

LIBERATI, A. et al. **The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions**. BMJ, v. 339, b2700, 2009.

MOHER, D. et al. **Preferred Reporting Items for Systematic Reviews and Meta-Analyses**: The PRISMA Statement. PLoS Medicine, v. 6, n. 7, e1000097, 2009.

PAGE, M.J. et al. **PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews**. BMJ, v. 372, n. 71, 2021.

PETTICREW, M.; ROBERTS, H. **Systematic Reviews in the Social Sciences: A Practical Guide**. Malden: Wiley-Blackwell, 2006.

REIS, J.; HOUSLEY, M. **Fundamentals of Data Engineering: Plan and Build Robust Data Systems**. Sebastopol: O'Reilly Media, 2024.

SIGELMAN, B.; BARROS, A. et al. **Dapper, a large-scale distributed systems tracing infrastructure**. Google Research, 2010.

STRUBELL, E.; GANESH, A.; MCCALLUM, A. **Energy and policy considerations for deep learning in NLP**. ACL, 2019.

WOOLDRIDGE, M. **An Introduction to MultiAgent Systems**. Wiley, 2009.

YAO, S.; BUTTERWICK, J.; et al. **ReAct: Synergizing reasoning and acting in language models.** ACM Transactions on Intelligent Systems, 2023.

ZHU, X. et al. **GraphRAG: Integrating knowledge graphs with retrieval-augmented generation.** Journal of Artificial Intelligence Research, 2024.

WORLD ECONOMIC FORUM. **Future of Jobs Report 2025**. Geneva: WEF, 2025.