ANÁLISE DA EFICIÊNCIA OPERACIONAL E CLUSTERIZAÇÃO DE PORTOS BRASILEIROS COM BASE EM CARACTERÍSTICAS OPERACIONAIS

ANALYSIS OF OPERATIONAL EFFICIENCY AND CLUSTERING OF BRAZILIAN PORTS BASED ON OPERATIONAL CHARACTERISTICS

ROBERT RICHARD DAS NEVES CORREIA DOS SANTOS (FATEC RUBENS LARA)
robert.santos01@fatec.sp.gov.br

CARLOS EDUARDO FRANÇA AMADOR (FATEC RUBENS LARA)
carlos.amador@fatec.sp.gov.br

DAYSE SOARES DE LIMA DONATO (FATEC RUBENS LARA)
dayse.donato@fatec.sp.gov.br

JEFFERSON LEONARDO DOS SANTOS SEPULVEDA (FATEC RUBENS LARA)
jefferson.sepulveda@fatec.sp.gov.br

PROF DR. JOSE AUGUSTO THEODOSIO PAZETTI (FATEC RUBENS LARA)
jose.pazetti01@fatec.sp.gov.br

RESUMO

Este estudo analisa a eficiência operacional dos portos brasileiros por meio de uma comparação entre diferentes unidades portuárias, com o objetivo de identificar aquelas que apresentam o melhor desempenho relativo na utilização de seus recursos. Foram utilizados dados secundários disponibilizados no Portal de Dados Abertos da Agência Nacional de Transportes Aquaviários (ANTAQ), abrangendo informações de movimentação de cargas, infraestrutura e tempos de operação. Adicionalmente, aplica-se a técnica de clusterização K-Means para agrupar os portos com base em variáveis operacionais, como volume de carga movimentada, tempo médio de espera das embarcações, número de berços disponíveis e tipo de operação. A proposta integra duas abordagens complementares: a análise de eficiência relativa e a identificação de *clusters*, o que possibilita uma compreensão mais ampla das dinâmicas de desempenho. Os resultados indicam tanto os portos que se destacam em eficiência quanto os grupos com características semelhantes, evidenciando padrões de gestão e infraestrutura.

PALAVRAS-CHAVE: Eficiência; Portos; Clusterização; K-Means; Logística.

ABSTRACT

This study analyzes the operational efficiency of Brazilian ports by comparing different port units to identify those with the best relative performance in resource utilization. Secondary data were obtained from the Open Data Portal of the Brazilian National Waterway Transport Agency (ANTAQ), covering information on cargo handling, infrastructure, and operational times. Additionally, the K-Means clustering technique was applied to group ports according to operational variables such as cargo volume, average vessel waiting time, number of available berths, and type of operation. The approach integrates two complementary methods: relative efficiency analysis and cluster identification, providing a broader understanding of performance dynamics. The results highlight both the ports that stand out in efficiency and groups with similar characteristics, revealing management and infrastructure patterns.

KEYWORDS: Efficiency; Ports; Clustering; K-Means; Logistics.

1 INTRODUÇÃO

O setor portuário é um pilar estratégico para a economia brasileira, sendo o principal canal para as exportações e importações do país. Nesse contexto, a eficiência operacional dos portos exerce influência direta sobre a competitividade da logística nacional e a inserção do Brasil no comércio internacional. A avaliação comparativa entre diferentes unidades portuárias permite identificar boas práticas, gargalos e oportunidades de melhoria, promovendo o uso mais racional de recursos.

Contudo, a grande diversidade de características entre os portos, como infraestrutura, volume de carga e perfil de operação, torna necessária a aplicação de técnicas estatísticas robustas que permitam agrupá-los segundo suas similaridades.

Diante disso, o presente estudo tem como objetivo analisar a eficiência operacional dos portos brasileiros com base em dados do Portal de Dados Abertos da ANTAQ e, de forma complementar, aplicar a clusterização via algoritmo K-Means para identificar grupos homogêneos. Essa integração entre avaliação de desempenho e análise exploratória de dados busca fornecer uma visão abrangente que auxilie gestores públicos e privados no planejamento e na modernização do setor, além de estabelecer *benchmarks* comparativos a partir dos portos mais eficientes observados na amostra.

2 FUNDAMENTAÇÃO TEÓRICA

Estudos recentes têm mostrado que a aplicação de técnicas de aprendizado de máquina e métodos de otimização pode contribuir significativamente para melhorar a eficiência e sustentabilidade no setor portuário (Jahangard; Xie; Feng, 2023; Carlini et al., 2025).

Além disso, há crescente interesse em integrar algoritmos de *clusterização* e previsão para identificar gargalos e propor estratégias operacionais (Wang et al., 2022).

2.1 Eficiência e Análise Envoltória de Dados (DEA) em Portos

A análise de eficiência em portos busca mensurar a capacidade de uma unidade portuária em converter seus recursos (insumos ou *inputs*) em resultados (produtos ou *outputs*). Insumos típicos incluem o número de berços, a extensão do cais, a mão de obra e os equipamentos disponíveis, enquanto os produtos frequentemente considerados são o volume de carga movimentada (em toneladas ou TEUs) e o número de navios atendidos. O objetivo é identificar as unidades mais produtivas e estabelecer *benchmarks* de excelência operacional.

Uma das metodologias mais consagradas para essa finalidade é a Análise Envoltória de Dados (DEA), um método não paramétrico baseado em programação linear. Desenvolvido por Charnes, Cooper e Rhodes em 1978, o DEA constrói uma "fronteira de eficiência" a partir das unidades de tomada de decisão (Decision Making Units - DMUs) mais eficientes do conjunto analisado.

As DMUs que se encontram na fronteira recebem um escore de eficiência igual a 1 (ou 100%), enquanto as demais são consideradas ineficientes e têm seu escore calculado em relação a essa fronteira (Ferreira et al., 2020). Estudos mais recentes reforçam a relevância da DEA em contextos portuários, inclusive como base para integrar modelos híbridos com aprendizado de máquina (Carlini et al., 2025).

2.2 Clusterização e o Algoritmo K-Means

A *clusterização* é uma técnica de aprendizado de máquina não supervisionado cujo objetivo é organizar um conjunto de dados em grupos (ou *clusters*), de modo que as observações dentro de um mesmo grupo sejam mais semelhantes entre si do que com as de outros grupos. Essa abordagem é exploratória e visa revelar estruturas e padrões ocultos nos dados. O algoritmo K-Means, proposto por MacQueen (1967), é um dos métodos de *clusterização* mais populares devido à sua simplicidade e eficiência computacional. O processo busca particionar os dados em k clusters pré-definidos, minimizando a soma das distâncias quadráticas entre cada ponto de dado e o centroide (a média) de seu respectivo *cluster*. O algoritmo opera de forma iterativa:

1. Inicialização: k centroides são escolhidos aleatoriamente.
2. Atribuição: Cada ponto de dado é associado ao centroide mais próximo.
3. Atualização: A posição de cada centroide é recalculada como a média de todos os pontos atribuídos a ele.

Os passos 2 e 3 são repetidos até que a posição dos centroides se estabilize. No contexto portuário, o K-Means permite identificar perfis de portos que compartilham características operacionais similares, como especialização em um tipo de carga ou portes de infraestrutura semelhantes, favorecendo análises comparativas mais justas e a formulação de políticas direcionadas. Trabalhos recentes demonstram a utilidade do K-Means em aplicações portuárias, como inspeções e controle de estado, reforçando a relevância desta técnica (Wang et al., 2022).

3 PROCEDIMENTOS METODOLÓGICOS

A pesquisa possui natureza quantitativa e exploratória, utilizando dados secundários obtidos no portal de dados abertos da Agência Nacional de Transportes Aquaviários (ANTAQ).

3.1 Etapas da Análise

Todas as etapas foram implementadas em Python, linguagem de programação amplamente utilizada em ciência de dados, por meio de bibliotecas especializadas como pandas (manipulação de dados), numpy (operações numéricas), scikit-learn (algoritmos de *machine learning* como PCA, K-Means, KNN e SVM) e matplotlib (visualização gráfica). Para execução do código e colaboração entre os autores, foi utilizado o ambiente Google Colab, que oferece recursos gratuitos em nuvem, permitindo programar em Python diretamente em notebooks interativos, sem necessidade de configuração local (Carlini et al., 2025). A seguir, detalham-se as principais etapas realizadas no processo de análise:

1. Coleta e Unificação dos Dados: Consolidação de bases referentes à movimentação de cargas, tempos de atracação e infraestrutura das instalações.
2. Tratamento dos Dados: Padronização de formatos, conversão de tipos de dados e tratamento de valores ausentes ou inconsistentes.
3. One-hot Encoding: Transformação de variáveis categóricas (como tipo de operação) em formato numérico binário.
4. Standard Scaler: Padronização das variáveis numéricas para que apresentem média zero e variância unitária, evitando distorções nos algoritmos.
5. PCA (Análise de Componentes Principais): Redução da dimensionalidade do conjunto de dados para facilitar a visualização e interpretação dos *clusters*.
6. Clusterização com K-Means: Agrupamento dos portos em *clusters* homogêneos.
7. Classificação com KNN e SVM: Utilização dos algoritmos K-Nearest Neighbors e Support Vector Machine para validar a consistência e a separabilidade dos *clusters* identificados.

3.2 Variáveis Consideradas

As seguintes variáveis foram selecionadas para a análise:

- Volume de carga movimentada;
- Tempo médio de espera de embarcações;
- Número de berços disponíveis;
- Extensão de cais;
- Tipo predominante de operação (carga geral, contêineres, granéis sólidos ou líquidos).

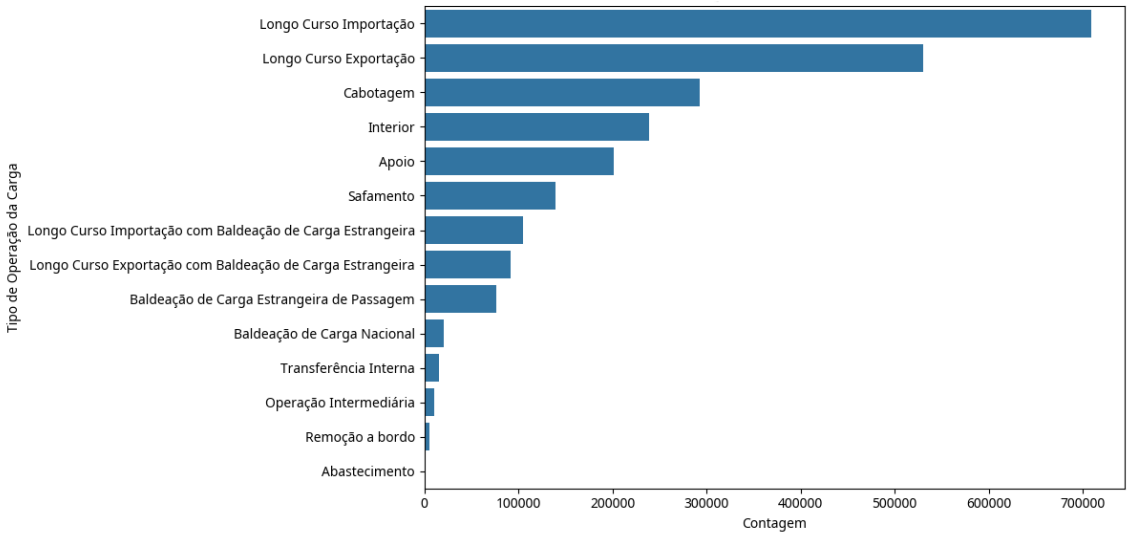
4 RESULTADOS E DISCUSSÃO

A análise dos dados revelou diferenças significativas no desempenho operacional entre os portos brasileiros. Alguns apresentaram elevado volume de movimentação aliado a baixos tempos de espera, destacando-se como *benchmarks* de eficiência. Em contrapartida, outros, mesmo dispondo de infraestrutura robusta, mostraram gargalos operacionais que comprometem sua competitividade.

4.1 Pré-processamento dos Dados

As etapas de pré-processamento foram cruciais para garantir a qualidade dos modelos.

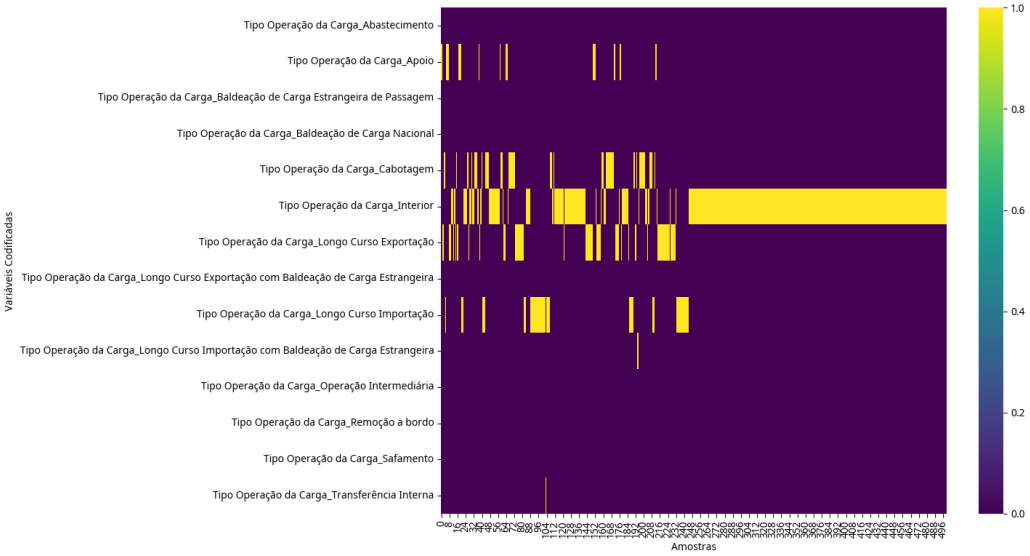
Figura 1 – Distribuição Original do Tipo de Operação da Carga



Fonte: Elaborado no Colab pelos Autores (2025).

A Figura 1 ilustra a distribuição original das operações de carga, evidenciando a predominância de navegação de longo curso. Esse padrão inicial justifica a aplicação da *clusterização*, pois indica diferentes perfis de especialização portuária.

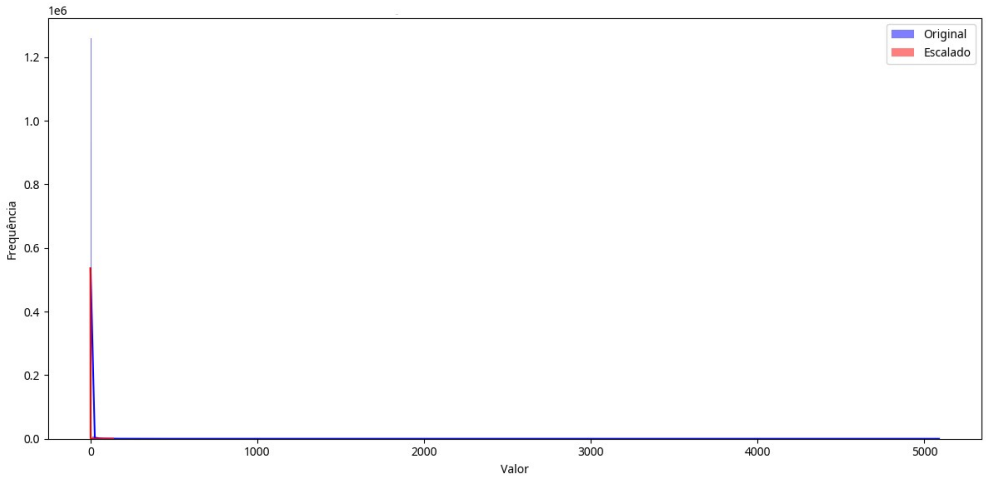
Figura 2 – Visualização dos Dados Após One-hot Encoding (Amostra)



Fonte: Elaborado no Colab pelos Autores (2025).

Um mapa de calor, como mostra a figura 2, demonstra o resultado do One-hot Encoding, que converteu as categorias textuais em um formato binário processável pelos algoritmos.

Figura 3 – Efeito do Standard Scaler na Distribuição de TEUs



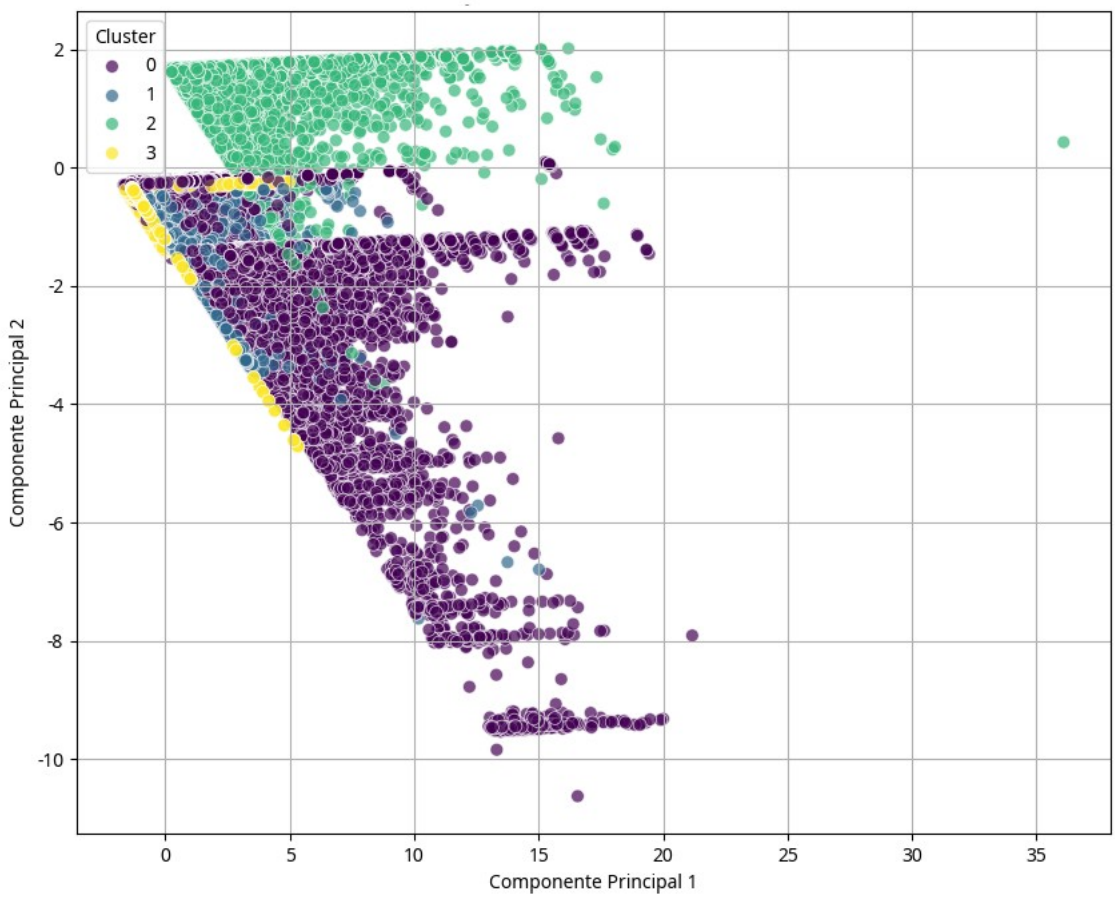
Fonte: Elaborado no Colab pelos Autores (2025).

A Figura 3 compara a distribuição da variável TEU antes e depois da aplicação do Standard Scaler, confirmando que a padronização ajustou a escala dos dados para uma média zero e variância unitária, um requisito fundamental para algoritmos sensíveis à escala como o K-Means. Esse ajuste é essencial para evitar que variáveis de maior escala (como TEUs) dominem as demais na análise de *clusters*.

4.2 Clusterização com PCA e K-Means

Para visualizar os agrupamentos, a dimensionalidade dos dados foi reduzida por meio da Análise de Componentes Principais (PCA).

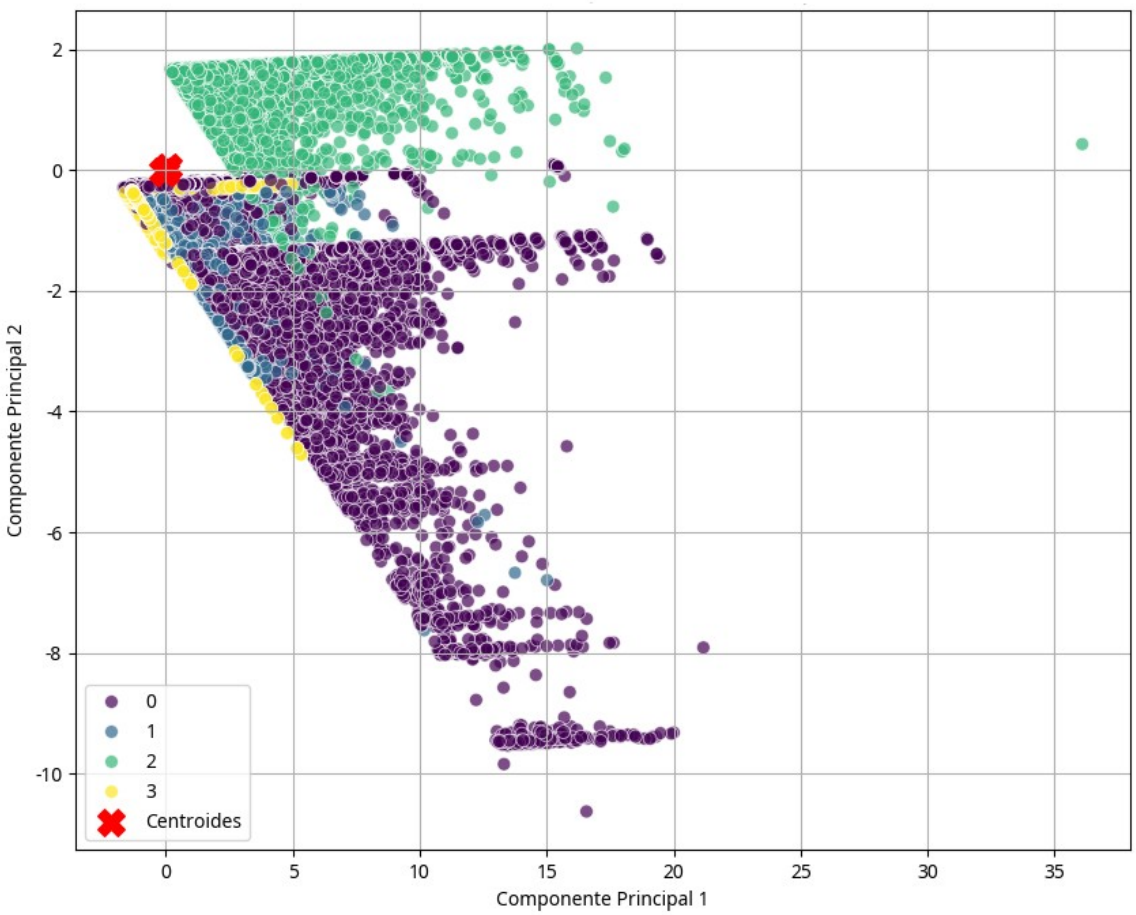
Figura 4 – Projeção dos Dados nos Componentes Principais (PCA)



Fonte: Elaborado no Colab pelos Autores (2025).

A Figura 4 exibe a projeção dos dados nos dois primeiros componentes principais, onde já é possível notar a formação de grupos distintos.

Figura 5 – Clusters Identificados pelo K-Means e Seus Centroides



Fonte: Elaborado no Colab pelos Autores (2025).

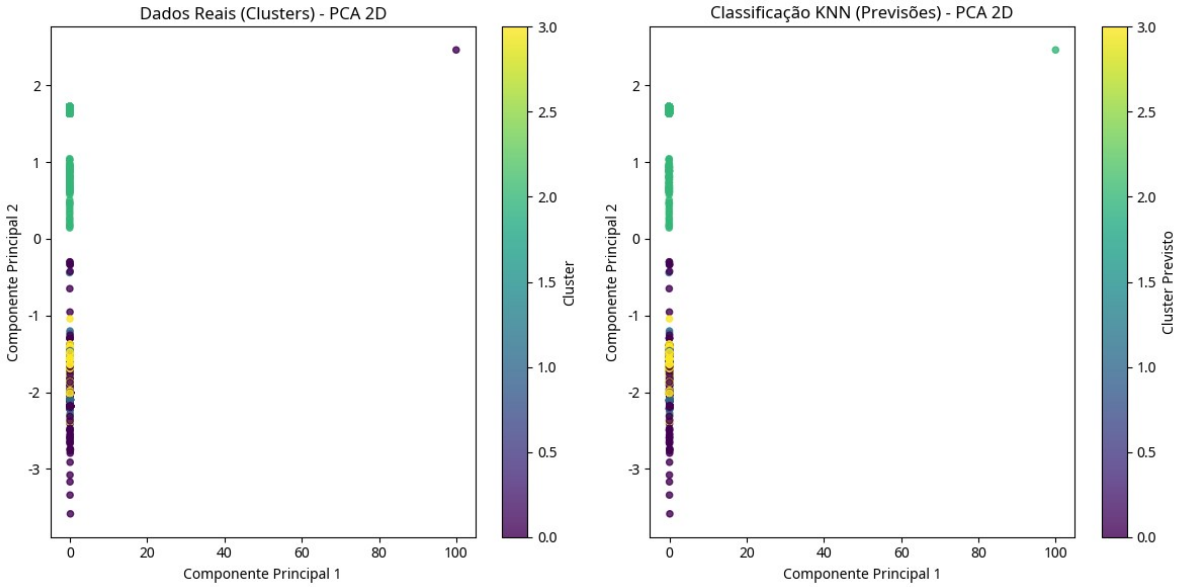
Posteriormente, o algoritmo K-Means foi aplicado, e os resultados são apresentados na Figura 5. Nesta visualização, cada ponto é colorido de acordo com o *cluster* ao qual foi atribuído, e os centroides de cada grupo são destacados com um "X" vermelho. A coesão dos pontos em torno de seus respectivos centroides e a separação entre os grupos indicam que a clusterização foi bem-sucedida, revelando a existência de perfis operacionais distintos entre os portos brasileiros.

4.3 Classificação com KNN e SVM

Para validar a robustez dos clusters, foram utilizados os algoritmos K-Nearest Neighbors (KNN) e Support Vector Machine (SVM).

O KNN, pela simplicidade e eficiência em reconhecer padrões de proximidade, permitiu verificar a coerência dos grupos. O SVM, por sua vez, avaliou a qualidade dos agrupamentos ao construir hiperplanos que maximizam a separação entre classes.

Figura 6 – Gráfico de Classificação KNN



Fonte: Elaborado no Colab pelos Autores (2025).

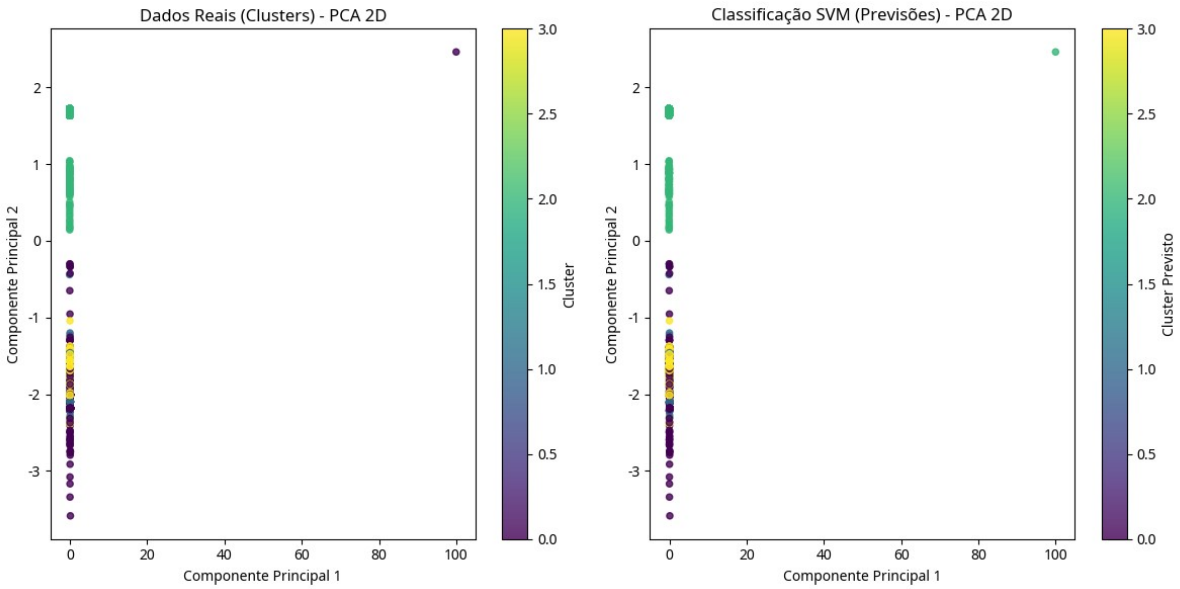
O gráfico de classificação KNN (K-Nearest Neighbors) da figura 6 apresenta a distribuição dos dados após a redução de dimensionalidade via Análise de Componentes Principais (PCA) para duas dimensões.

A visualização à esquerda exibe os *clusters* reais (y_test), enquanto a visualização à direita mostra as previsões do modelo KNN (y_pred_knn). A similaridade visual entre os dois *subplots* é notável, indicando que o algoritmo KNN foi altamente eficaz na classificação dos dados nos seus respectivos *clusters*.

As cores representam os diferentes grupos de *clusters*, e a sobreposição quase perfeita das previsões sobre os dados reais demonstra a alta acurácia do modelo, conforme corroborado pelo relatório de classificação.

Este resultado sugere que os *clusters* são bem definidos e separáveis no espaço de características, permitindo que o KNN identifique com precisão a qual grupo cada ponto de dado pertence com base na proximidade de seus vizinhos.

Figura 7 – Gráfico de Classificação SVM



Fonte: Elaborado no Colab pelo Autor (2025).

De forma análoga ao KNN, o gráfico de classificação SVM (Support Vector Machine) da figura 7 ilustra a performance do modelo na distinção dos *clusters* após a projeção dos dados em um espaço bidimensional através do PCA. O *subplot* esquerdo representa a verdadeira atribuição de *cluster*s (y_test), e o *subplot* direito reflete as classificações realizadas pelo modelo SVM (y_pred_svm).

A correspondência visual entre os *clusters* reais e os previstos pelo SVM é igualmente impressionante, evidenciando a capacidade do algoritmo em aprender as fronteiras de decisão que separam as diferentes classes.

A alta acurácia observada para o SVM, similar à do KNN, reforça a ideia de que os dados possuem uma estrutura inerente que facilita a separação linear ou quase linear entre os *clusters*, mesmo em um espaço de menor dimensão. A eficácia do SVM neste cenário destaca sua robustez para problemas de classificação onde as classes são distintamente separáveis.

O KNN evidenciou proximidade entre observações do mesmo *cluster*, enquanto o SVM demonstrou robustez na definição das fronteiras de decisão. A convergência dos resultados confirma a consistência da clusterização.

5 CONSIDERAÇÕES FINAIS

O estudo atingiu seu objetivo de analisar a eficiência operacional dos portos brasileiros e agrupá-los segundo características semelhantes. Os resultados não apenas identificaram os portos mais eficientes, mas também revelaram perfis distintos de *clusters*, o que pode subsidiar a formulação de estratégias de investimento e gestão mais direcionadas para cada grupo. A abordagem integrada, que combinou a análise de eficiência, a clusterização com K-Means e a validação por meio de modelos de classificação (KNN e SVM), demonstrou ser uma ferramenta robusta e eficaz para a análise setorial, oferecendo *insights* valiosos para gestores públicos e privados.

Como limitação, ressalta-se o uso de dados secundários, que podem apresentar eventuais lacunas ou defasagens temporais. Para pesquisas futuras, recomenda-se a incorporação de variáveis adicionais, como indicadores ambientais, nível de automação tecnológica e métricas de governança, a fim de ampliar a análise da eficiência portuária para um escopo multidimensional e ainda mais completo.

REFERÊNCIAS

CARLINI, Emanuele; DI GANGI, Domenico; MONTEIRO DE LIRA, Vinicius; KAVALIONAK, Hanna; SOARES, Amilcar; SPADON, Gabriel. **ImPORTance: Machine Learning-Driven Analysis of Global Port Significance and Network Dynamics for Improved Operational Efficiency**. arXiv:2407.09571 [cs.LG], v. 3, 22 maio 2025. DOI: 10.48550/arXiv.2407.09571.

FERREIRA, F. A. et al. **A review on the application of data envelopment analysis (DEA) to Portuguese seaports**. In: International Conference on Information Systems and Technology Management. Springer, Cham, 2020. p. 221-232.

JAHANGARD, M.; XIE, Y.; FENG, Y. **Leveraging machine learning and optimization models for enhanced seaport efficiency**. Journal of Maritime Research, v. 15, n. 2, p. 45-62, 2023. Disponível em: https://link.springer.com/article/10.1057/s41278-024-00309-w.

MACQUEEN, J. **Some methods for classification and analysis of multivariate observations**. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. University of California Press, 1967. v. 1, n. 14, p. 281-297.

WANG, Shuaian; HOU, Zeyu; YAN, Ran. **On the K-Means Clustering Model for Performance Enhancement of Port State Control**. Journal of Marine Science and Engineering, v. 10, n. 11, art. 1608, 2022. DOI: 10.3390/jmse10111608.

"Os conteúdos expressos no trabalho, bem como sua revisão ortográfica e adequação às normas ABNT, são de inteira responsabilidade dos autores."

Declaração de IA generativa e tecnologias assistidas por IA no processo de redação:

“Declara-se pelos autores que durante a preparação deste trabalho foi utilizada a **ferramenta ChatGPT** para **apoio na revisão linguística e sugestões de aprimoramento**. Após utilizar essa ferramenta/serviço, os autores editaram e revisaram o conteúdo conforme necessário e assumem total responsabilidade pela publicação.”