Agrupamento de Padrões de Vento

Antonio Paulo A. de Barros e Silva[1], Clara Machado de Araújo[1], Davi Gomes F. R. de Almeida[1], Heloísa Tanaka Fernandes[1], João Pedro Fontes Ferreira[1], Larissa Sobrinho Santos[1], Leonardo C. de Carvalho Guedes[1]

[1]C.E.S.A.R School – Centro de Estudos e Sistemas Avançados do Recife

{apabs,cma3,dgfra,htf,jpff2,lss2,lccg}@cesar.school

Abstract. *The analysis of massive meteorological data demands scalable infrastructures. This paper proposes a containerized architecture to process data from INMET, applied to wind monitoring in Caruaru-PE. The solution orchestrates a continuous pipeline integrating ingestion via FastAPI, Data Lake (MinIO), and structuring in PostgreSQL. Using the K-Means algorithm, behavioral wind clusters (direction, speed, and gust) were identified, with experiments managed via MLFlow. The results validate the integration between Data Engineering and Machine Learning, providing dashboards in ThingsBoard that facilitate the interpretation of atmospheric phenomena.*

Resumo. *A análise de dados meteorológicos massivos demanda infraestruturas escaláveis. Este artigo propõe uma arquitetura conteinerizada para processar dados do INMET, aplicada ao monitoramento de ventos em Caruaru-PE. A solução orquestra um pipeline contínuo integrando ingestão via FastAPI, Data Lake (MinIO) e estruturação em PostgreSQL. Utilizando o algoritmo K-Means, identificaram-se clusters comportamentais de vento (direção, velocidade e rajada), com experimentos geridos via MLFlow. Os resultados validam a integração entre Engenharia de Dados e Machine Learning, fornecendo dashboards no ThingsBoard que facilitam a interpretação de fenômenos atmosféricos.*

1. Introdução

Este projeto desenvolve um pipeline completo de Business Intelligence (BI) para dados meteorológicos do Instituto Nacional de Meteorologia (INMET), com foco no estado de Pernambuco. O objetivo central é transformar dados brutos, coletados por meio de arquivos CSV, em informações estruturadas, tratadas e visualmente interpretáveis, utilizando uma arquitetura moderna baseada em contêineres. O pipeline integra serviços como **FastAPI** para ingestão, **MinIO** para armazenamento, **Postgres** local para banco de dados dos outros serviços, **Jupyter Notebook** para análise e modelagem, **MLFlow** para versionamento e **ThingsBoard** para visualização interativa.

Além da construção da infraestrutura, o projeto aplica técnicas de análise exploratória e aprendizado de máquina para identificar padrões meteorológicos relevantes. No escopo deste trabalho, a ênfase está no agrupamento de padrões de vento, combinando direção, velocidade e rajada para revelar comportamentos atmosféricos semelhantes ao longo do tempo. O resultado final demonstra como a integração entre

engenharia de dados, modelagem analítica e dashboards interativos pode gerar inteligência aplicável e apoiar a interpretação de fenômenos climáticos regionais.

2. Objetivos do Trabalho

O presente trabalho tem como objetivo desenvolver e avaliar um pipeline completo de processamento, análise e modelagem preditiva aplicado a dados meteorológicos provenientes de arquivos CSV do INMET. Busca-se estruturar, limpar e padronizar os dados brutos, integrar técnicas de aprendizado de máquina para predição de comportamentos atmosféricos e disponibilizar os resultados por meio de visualizações analíticas capazes de apoiar a interpretação dos fenômenos climáticos regionais. Para isso, propõe-se uma arquitetura baseada em serviços conteinerizados, contemplando ingestão, armazenamento, tratamento, modelagem e exposição dos resultados em dashboards, de modo a demonstrar a aplicabilidade prática de soluções de Inteligência Artificial no contexto de monitoramento meteorológico.

2.1. Objetivos Específicos

- Coletar dados horários de direção, velocidade e rajada do vento a partir de arquivos CSV obtidos do INMET.
- Tratar inconsistências, remover valores ausentes e padronizar as variáveis de vento.
- Criar métricas e representações adequadas para análise, incluindo a conversão da direção do vento em componentes vetoriais (u/v).
- Aplicar o algoritmo de clusterização K-Means para identificar padrões de comportamento do vento.
- Avaliar a qualidade dos agrupamentos utilizando métricas como o índice de silhouette.
- Registrar experimentos e versões de modelos no MLFlow.
- Exibir os resultados do agrupamento em dashboards interativos no ThingsBoard.
- Documentar e validar o pipeline completo utilizando Docker Compose, garantindo reprodutibilidade e organização da solução.

3. Arquitetura do Pipeline

Figure 1. Flow Diagram

3.1. Visão Geral da Arquitetura

O Pipeline inicia em **FastAPI,** pegando as informações diretamente do csv do INMET, extraindo apenas as colunas necessárias para o presente trabalho, sendo elas: “VENTO,

DIREÇÃO HORÁRIA (gr)", "VENTO, RAJADA MÁXIMA (m/s)", "VENTO, VELOCIDADE HORÁRIA (m/s)". Após a extração dos dados nas colunas necessárias.

O próximo passo é, com os dados extraídos, a geração do arquivo **.parquet** para armazenar no serviço **MinIO**. A partir do arquivo salvo no **MinIO**, é feito três processos, sendo eles:

1. Processo de clusterização dos dados armazenados no arquivo **.parquet** para armazenar no serviço de banco de dados **Postgres** na tabela **vento_clusters_diarios**.

2. Exploração inicial dos dados, realizando os primeiros tratamentos (conversão de tipos, verificação de nulos), e o agrupamento dos dados horários por dia, testando diferentes valores de k no algoritmo K-Means que serão usados para o versionamento no MLFlow.

3. Leitura da tabela **vento_clusters_diarios**, realizando uma análise dos clusters e gerando gráficos que servirão de base para o dashboard, como por exemplo a rosa dos ventos.

Após de realizar todos esses processos de tratamentos de dados, os dados estão prontos para serem visualizados no ThingsBoard, sendo exibido em 6 gráficos diferentes, sendo eles 3 histogramas e 3 line charts. Além disso, as visualizações dos dados tratados estão sendo mostrados através de 3 histogramas e 3 line charts.

3.2. Descrição dos Componentes

A arquitetura do projeto foi composta por um conjunto integrado de ferramentas que sustentam todo o fluxo de ingestão, armazenamento, processamento e visualização dos dados. A ingestão inicial foi realizada por meio do **FastAPI**, que recebeu e processou automaticamente os dados meteorológicos disponibilizados pelo INMET. Esses dados, em sua forma bruta, juntamente com os modelos gerados ao longo do projeto, foram armazenados no **MinIO**, que atuou como repositório central de objetos.

Para a camada estruturada, utilizou-se um banco **Postgres**, responsável por organizar e persistir as tabelas tratadas. As estampas de tratamento, exploração e modelagem foram conduzidas no ambiente do **Jupyter Notebook**, onde foram desenvolvidos os experimentos e análises iniciais. O versionamento desses experimentos ficaram sob responsabilidade do **MLflow**, garantindo rastreabilidade e reprodutibilidade dos modelos.

A etapa de visualização foi implementada por meio de um dashboard construído em **ThingsBoard**, que permitiu apresentar os clusters e demais resultados de forma clara e interativa. Por fim, todos os serviços foram executados de maneira padronizada através do **Docker Compose**, que orquestrou os containers necessários ao funcionamento da solução.

3.3. Fluxo de Dados

- Entrada → Ingestão → Storage → ETL → Modelagem → Versionamento → Dashboard.

4. Coleta de Dados

A coleta de dados foi realizada a partir de um **arquivo CSV** previamente baixado, contendo registros meteorológicos consolidados. O **período trabalhado** corresponde ao intervalo definido no escopo do projeto, abrangendo todas as medições disponíveis nesse intervalo.

Para este estudo, foram selecionadas exclusivamente as **estações meteorológicas localizadas no estado de Pernambuco**, assegurando uniformidade regional nas observações. As variáveis consideradas nesta etapa foram: **data**, **hora**, **velocidade do vento**, **velocidade máxima de rajada**, **direção do vento por hora** e **velocidade horária**.

A ingestão dos dados foi automatizada por meio de rotinas desenvolvidas em **FastAPI**, responsáveis por ler o arquivo CSV, validar os campos e disponibilizar os registros para as etapas subsequentes. Esse procedimento padronizou o fluxo de entrada, reduziu a necessidade de intervenção manual e garantiu consistência no processamento das informações.

5. Análise Exploratória

A análise exploratória de dados (AED) foi uma etapa crucial para entender a estrutura dos dados meteorológicos e identificar padrões significativos antes da aplicação do clustering. Durante essa fase, foram realizadas diversas visualizações e estatísticas descritivas para examinar as distribuições das variáveis, como direção, velocidade e rajada de vento. Gráficos de dispersão, histogramas e lines charts foram utilizados para observar a variação dessas variáveis ao longo do tempo, detectar outliers e verificar a presença de correlações entre elas. Além disso, foi feita uma análise de valores ausentes e inconsistentes, que foram tratados adequadamente para garantir a qualidade dos dados. Através dessa abordagem, foi possível obter uma compreensão mais profunda dos dados, o que facilitou a escolha do algoritmo de clustering mais adequado e o pré-processamento necessário para a modelagem dos dados. A análise exploratória de dados também forneceu insights importantes sobre a sazonalidade e tendências meteorológicas, essenciais para o sucesso da análise de agrupamento.

6. Metodologia de Clustering

A metodologia de clustering aplicada neste projeto foi focada no uso do algoritmo **K-means**, com o objetivo de identificar padrões semelhantes no comportamento do vento em Pernambuco, com base nas variáveis de **direção do vento**, **velocidade do vento** e **rajada de vento**. O K-means é um algoritmo de aprendizado não supervisionado amplamente utilizado para agrupar dados com base em características

similares. Sua aplicação foi fundamental para identificar padrões e entender as dinâmicas atmosféricas da região de Pernambuco. Além disso, utilizamos o **Índice de Silhueta** para avaliar a qualidade dos clusters formados, garantindo que o agrupamento fosse significativo e bem definido.

7. Resultados

A aplicação da metodologia proposta permitiu identificar padrões consistentes no comportamento do vento na região de Caruaru-PE ao longo do período analisado. Após o pré-processamento dos dados e a conversão das direções de vento em componentes vetoriais, diferentes valores de k foram testados no algoritmo K-Means, com avaliação por meio do Índice de Silhouette. O melhor desempenho foi obtido com k = 4, representando o ponto de equilíbrio entre separabilidade dos grupos e coesão interna dos padrões formados.

Os quatro clusters resultantes apresentaram assinaturas meteorológicas distintas, refletindo tanto regimes típicos de circulação de ventos quanto episódios mais intensos observados ao longo do ano. A seguir, descrevem-se os comportamentos predominantes identificados:

• Cluster 0 — Ventos Fracos e Direção Predominante Leste-Sudeste

Esse cluster concentrou registros de menor intensidade, com velocidades geralmente inferiores a 2 m/s. A direção se manteve majoritariamente entre 90° e 140°, associada ao regime de brisas e condições atmosféricas estáveis. Os gráficos de série temporal e histogramas mostram baixa variabilidade e comportamento relativamente previsível ao longo dos meses.

• Cluster 1 — Ventos Moderados e Estáveis

Caracterizado por velocidades entre 2 e 4 m/s, esse grupo apresentou fluxo mais consistente ao longo do ano, com dispersão direcional moderada. Os dados sugerem um regime intermediário, associado a padrões de circulação regionais que se mantêm mesmo com mudanças pontuais de temperatura ou pressão.

• Cluster 2 — Ventos Fortes com Baixa Variabilidade Direcional

Registrou velocidades mais elevadas (4 a 6 m/s) e rajadas frequentes, porém com direção relativamente concentrada — conforme evidenciado na rosa dos ventos. Esse comportamento indica a atuação de ventos mais persistentes, provavelmente associados a episódios atmosféricos de maior estabilidade dinâmica, como corredores de vento mais definidos.

• Cluster 3 — Ventos de Alta Variabilidade Direcional e Rajadas Intensas

Esse cluster agrupou eventos de vento com maior irregularidade, tanto na intensidade quanto na direção. Observam-se pontos dispersos em diferentes quadrantes na rosa dos

ventos, indicando oscilações mais abruptas e ocorrência de rajadas intensas, muitas vezes superiores a 8–10 m/s. Trata-se do padrão mais turbulento entre os quatro grupos, associado a instabilidades atmosféricas e possíveis eventos sinóticos.

Figure 2. Gráficos de dados brutos

Figure 3. Gráficos de dados tratados

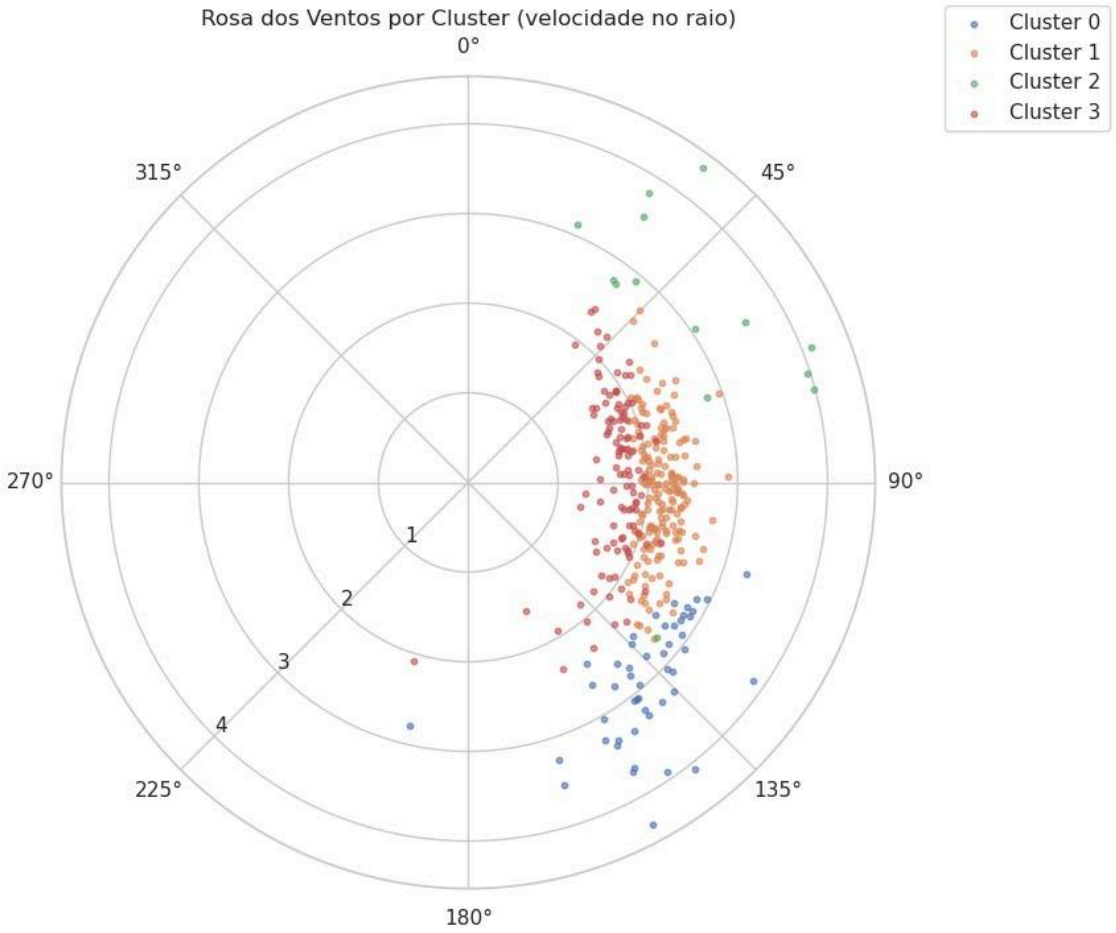

Figure 4. Gráfico Rosa dos Ventos

Referências

INSTITUTO NACIONAL DE METEOROLOGIA. Dados históricos do INMET. 2024. Disponível em: https://portal.inmet.gov.br/dadoshistoricos

MINIO. MinIO Documentation. Disponível em: https://docs.min.io/enterprise/aistor-object-store/

POSTGRESQL. PostgreSQL Documentation. Disponível em:

JUPYTER. Project Jupyter Documentation. Disponível em: https://docs.jupyter.org/pt-br/latest/

MLFLOW. MLFlow Documentation. Disponível em: https://mlflow.org/docs/3.3.1/

THINGS BOARD. Documentação da Plataforma ThingsBoard. Disponível em: https://thingsboard.io/docs/

DOCKER. Docker Documentation. Disponível em: https://docs.docker.com/