Inteligência Artificial e Computacional

Global Solution – 1º Semestre

Daniel Okudaira Carapeto	93180
Gabriel Gonçalves	93069
Rodrigo Lima de Carvalho	96247

Identificação de Padrões de Poluição Plástica por Região

Introdução

A poluição plástica é uma das principais preocupações ambientais da atualidade. Com grandes quantidades de resíduos plásticos contaminando nossos oceanos, rios e terras, é essencial entender os padrões de poluição para desenvolver estratégias eficazes de mitigação. Este relatório apresenta uma abordagem detalhada para identificar padrões de poluição plástica em diferentes regiões geográficas utilizando técnicas de clustering.

“Clustering, ou agrupamento, consiste na implementação de técnicas computacionais para separar um conjunto de dados em diferentes grupos com base em suas semelhanças.” (“Conceitos básicos, principais algoritmos e aplicações - Medium”) Diferentemente de algoritmos de classificação e regressão, o Agrupamento faz parte do universo da Aprendizagem Não Supervisionada, na qual os algoritmos devem entender as relações entre dados sem estarem rotulados a nenhuma categoria prévia.

Metodologia

1. Análise Exploratória de Dados (EDA)
2. Pré-processamento dos Dados
3. Aplicação de Técnicas de Clustering
4. Análise e Visualização dos Clusters
5. Interpretação dos Resultados

1. Análise Exploratória de Dados (EDA)

A Análise Exploratória de Dados (EDA) é um passo crucial para entender a estrutura e as características dos dados. Aqui, exploramos o conjunto de dados para obter insights iniciais sobre a distribuição e variabilidade da poluição plástica.

Dados Utilizados

Os dados utilizados neste estudo podem incluir:

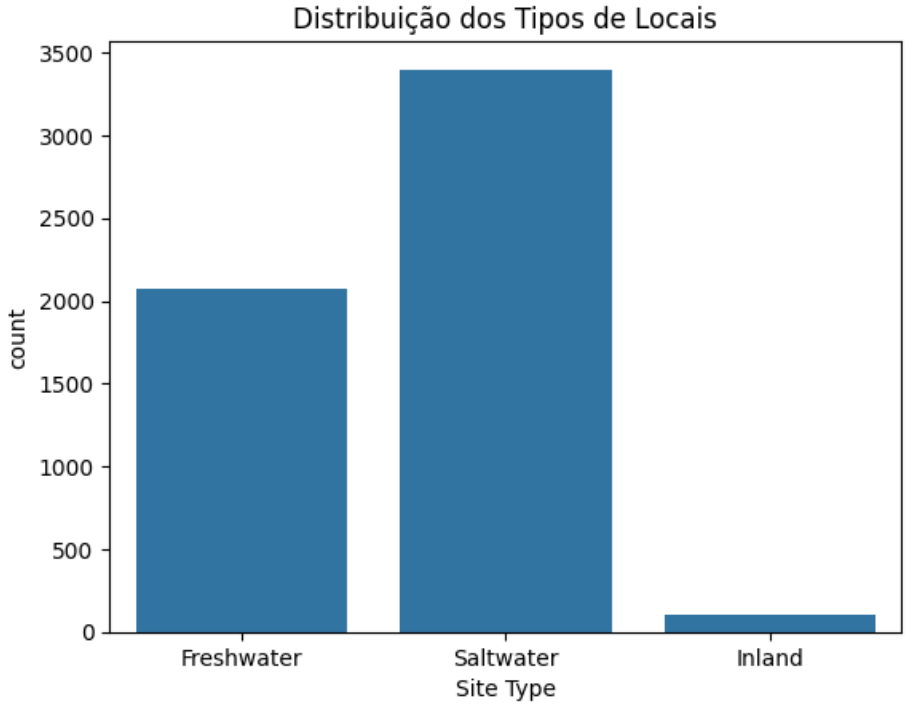
- Localização geográfica (latitude e longitude)
- Quantidade de resíduos plásticos
- Tipos de plásticos (micro plásticos, macro plásticos etc.)
- Fontes de poluição (doméstica, industrial, marítima etc.)
- Dados temporais (datas de coleta)

Principais Técnicas de EDA

- Estatísticas descritivas
- Visualizações gráficas (histogramas, boxplots, mapas de calor)
- Análise de correlação entre variáveis

Exemplos de Visualizações

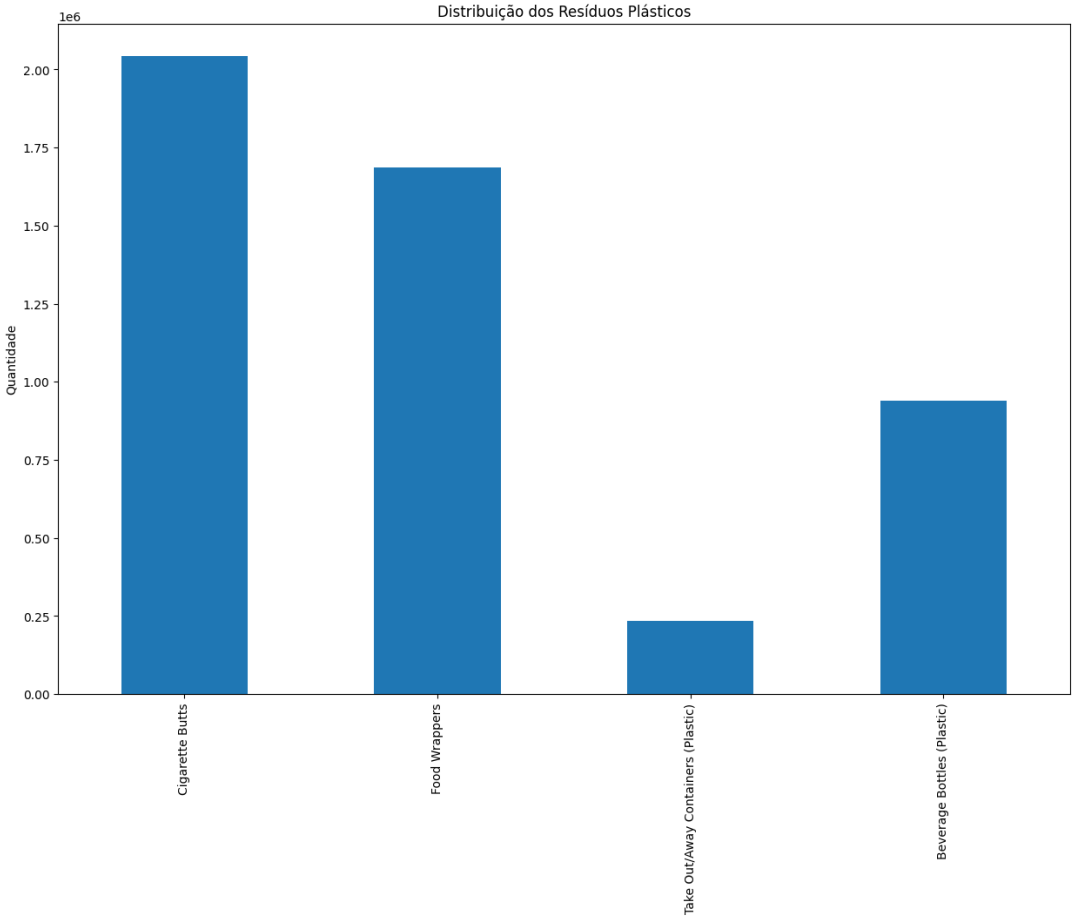
- Mapa de calor mostrando a densidade de poluição plástica em diferentes regiões
- Histogramas da quantidade de resíduos plásticos por tipo e fonte
- Gráficos de dispersão geográfica



2. Pré-processamento dos Dados

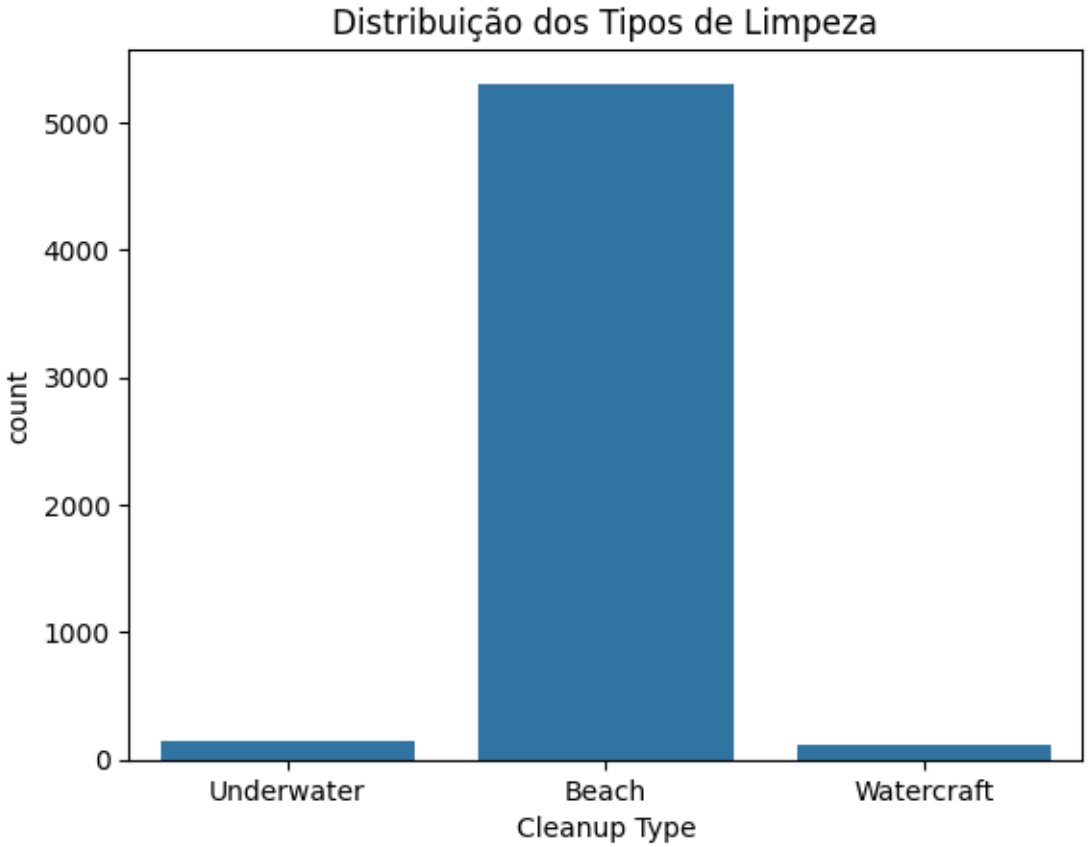
Antes de aplicar técnicas de clustering, é necessário preparar os dados. Isso envolve:

- Tratamento de valores ausentes
- Normalização ou padronização das variáveis
- Conversão de variáveis categóricas em numéricas
- Remoção de outliers, se necessário



Passos Detalhados

- Tratamento de Valores Ausentes: Imputação de valores ausentes utilizando médias, medianas ou técnicas mais avançadas como KNN-imputation.
- Normalização/Padronização: Ajuste das escalas das variáveis para garantir que nenhuma variável domine o clustering.
- Codificação de Variáveis Categóricas: Utilização de técnicas como One-Hot Encoding para converter categorias em valores numéricos.



3. Aplicação de Técnicas de Clustering

Clustering é uma técnica de aprendizado não supervisionado que agrupa dados em clusters baseados em similaridades. As técnicas comuns incluem:

- K-means
- DBSCAN
- Hierarchical Clustering
- Escolha do Método de Clustering

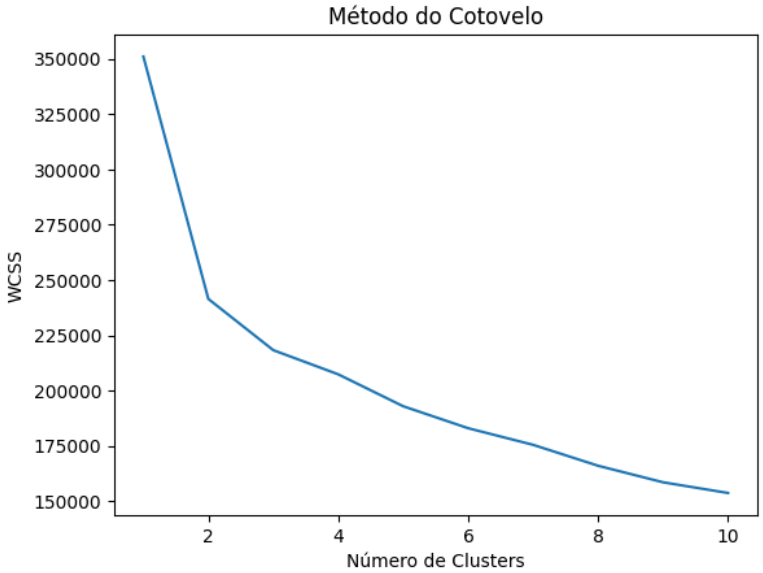
Para este estudo, utilizaremos o método K-means devido à sua simplicidade e eficácia em muitos cenários. No entanto, técnicas como DBSCAN podem ser exploradas para detectar clusters de forma não esférica.

K-Means

“Um dos algoritmos mais conhecido assim como bem antigo (proposto em 1967), esse algoritmo baseado em centroides, tem como objetivo, encontrar os agrupamentos nos dados nos quais a variância dentro de cada cluster seja mínima.” (“Conceitos básicos, principais algoritmos e aplicações - Medium”)

Determinação do Número de Clusters

Para utilizarmos o K-Means precisamos determinar o número de clusters. A determinação do número adequado de clusters (k) pode ser feita utilizando o método do cotovelo (Elbow Method) ou o método da silhueta (Silhouette Method).

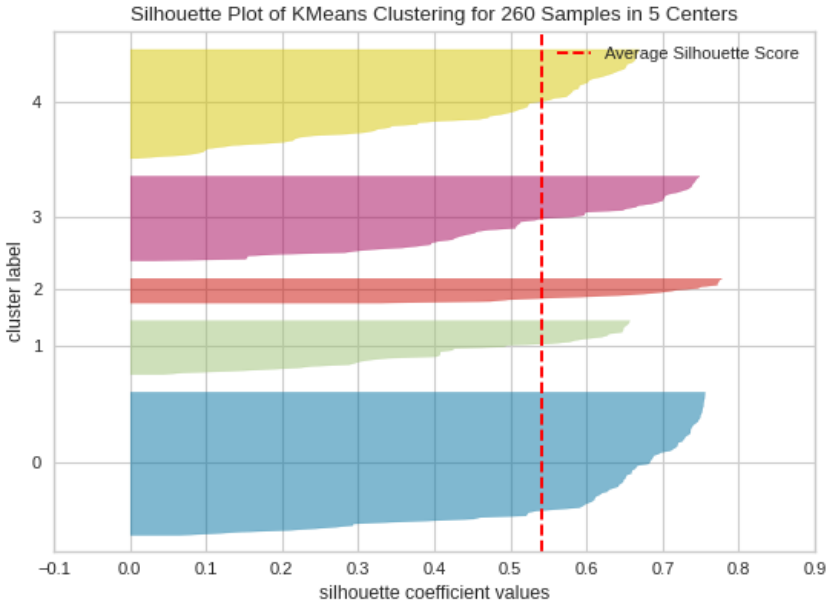


“O Método do Cotovelo é uma técnica utilizada em análise de dados e aprendizado de máquina para determinar o número ótimo de clusters em um conjunto de dados. Envolve a plotagem da variância explicada por diferentes números de clusters e a identificação do ponto onde a taxa de redução da variância se estabiliza, sugerindo o número apropriado de clusters para análise ou treinamento do modelo.” (Elbow Method for Finding the Optimal Number of Clusters in K-Means)

Este método é uma técnica visual usada para determinar o melhor valor de K para um algoritmo de clusterização K-Means. No gráfico do cotovelo, são plotados os valores da soma dos quadrados intra-cluster (WCSS) contra vários valores de K. O valor ótimo de K é identificado no ponto onde o gráfico forma um cotovelo.

O Método da Silhueta oferece uma melhor alternativa quando o método do cotovelo não mostra um ponto claro. A pontuação da silhueta varia de -1 a 1:

- 1: Pontos perfeitamente atribuídos e clusters distinguíveis.
- 0: Clusters sobrepostos.
- -1: Pontos atribuídos erroneamente.



A pontuação da silhueta é calculada usando a fórmula (b-a)/max(a,b), onde 'a' é a distância média intra-cluster e 'b' é a distância média inter-cluster. A implementação prática é demonstrada com exemplos de código Python, utilizando o conjunto de dados Iris. A visualização da silhueta ajuda a determinar o número de clusters mais adequado, mostrando que o valor ideal de K para este conjunto de dados é 3.

“Os gráficos de curva de cotovelo e silhueta são técnicas muito úteis para encontrar o K ideal para agrupamento de k-médias. Em conjuntos de dados do mundo real, você encontrará muitos casos em que a curva do cotovelo não é suficiente para encontrar o ‘K’ correto. Nesses casos, você deve usar o gráfico de silhueta para descobrir o número ideal de clusters para o seu conjunto de dados.” (Stop Using Elbow Method in K-Means Clustering)

Execução do Clustering

- Inicialização: Definir o número de clusters (k).
- Iteração: Atribuir pontos aos clusters mais próximos e recalcular os centroides.
- Convergência: Repetir o processo até que os centroides se estabilizem.

4. Análise e Visualização dos Clusters

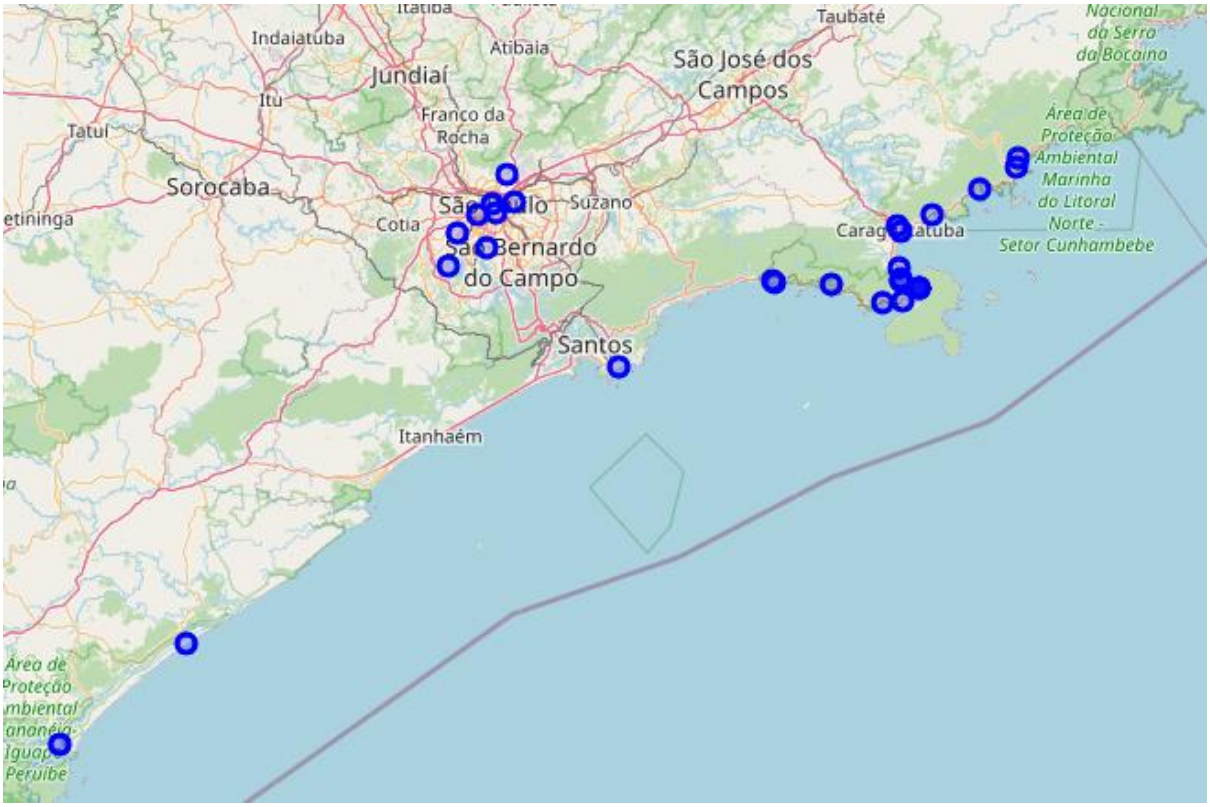
Após a aplicação do clustering, a análise dos clusters resultantes é essencial para entender os padrões de poluição.

Visualizações

- Mapas geográficos mostrando a distribuição dos clusters
- Gráficos de barras comparando características dos clusters
- Análise de centroides para identificar padrões centrais

Ferramentas

- Bibliotecas de visualização como Matplotlib, Seaborn e Plotly
- Ferramentas GIS (Geographic Information System) para mapas detalhados



5. Interpretação dos Resultados

A interpretação dos resultados envolve a análise das características de cada cluster para entender os padrões de poluição plástica em diferentes regiões.

Perguntas Chave

- Quais regiões apresentam os maiores níveis de poluição plástica?
- Existem padrões específicos de fontes de poluição em diferentes clusters?
- Como as características geográficas influenciam a distribuição da poluição plástica?

Uso Prático

Os insights obtidos podem ser utilizados para:

- Direcionar esforços de limpeza para áreas mais críticas
- Desenvolver políticas ambientais focadas nas regiões mais afetadas
- Promover campanhas de conscientização específicas para as características de cada região

Conclusão

Este relatório apresenta uma abordagem detalhada para identificar padrões de poluição plástica utilizando técnicas de clustering. Através da análise exploratória de dados, pré-processamento, aplicação de clustering e interpretação dos resultados, é possível obter uma compreensão aprofundada da distribuição da poluição plástica, permitindo ações mais eficazes na mitigação desse problema ambiental crítico.

Bibliografia

- Clustering — Conceitos básicos, principais algoritmos e aplicações: https://medium.com/turing-talks/clustering-conceitos-b%C3%A1sicos-principais-algoritmos-e-aplica%C3%A7%C3%A3o-ace572a062a9
- O que é clustering?: https://developers.google.com/machine-learning/clustering/overview?hl=pt-br
- Silhouette Visualizer: https://www.scikit-yb.org/en/latest/api/cluster/silhouette.html
- Stop Using Elbow Method in K-Means Clustering: https://builtin.com/data-science/elbow-method#:~:text=The%20elbow%20method%20is%20a%20graphical%20method%20for%20finding%20the,the%20graph%20forms%20an%20elbow.
- Elbow Method for Finding the Optimal Number of Clusters in K-Means: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/
- Vídeo do Projeto: https://youtu.be/DoU4viLpWFE