

CENTRO DE PESQUISA E DESENVOLVIMENTO TECNOLÓGICO EM
INFORMÁTICA E ELETROELETRÔNICA DE ILHÉUS (CEPEDI)
TRILHA DE CIÊNCIA DE DADOS

CAIO CORDEIRO MATOS
FERNANDO NARDES FERREIRA NETO

Human Activity Recognition Using Smartphones

Vitória da Conquista – BA
03 de dezembro de 2024

RESUMO

Este projeto tem como objetivo explorar e agrupar atividades humanas com base nos dados do dataset *Human Activity Recognition Using Smartphones*. Para isso, utilizou-se o algoritmo de K-means, com técnicas de normalização, análise exploratória e redução de dimensionalidade via PCA. O número ideal de clusters foi determinado usando métodos como o cotovelo (*elbow method*) e o *silhouette score*. Os resultados mostraram padrões claros que correspondem a diferentes tipos de atividades registradas, oferecendo insights significativos sobre os agrupamentos.

1 INTRODUÇÃO

O reconhecimento de atividades humanas é um campo em crescente desenvolvimento, com aplicações em saúde, esportes e interação homem-máquina. Este projeto utiliza o dataset *Human Activity Recognition Using Smartphones*, que contém dados coletados por sensores de acelerômetro e giroscópio durante atividades como caminhar, subir escadas e ficar em pé. O objetivo é identificar padrões subjacentes nessas atividades usando o algoritmo de K-means, uma técnica amplamente utilizada para agrupamento não supervisionado.

2 METODOLOGIA

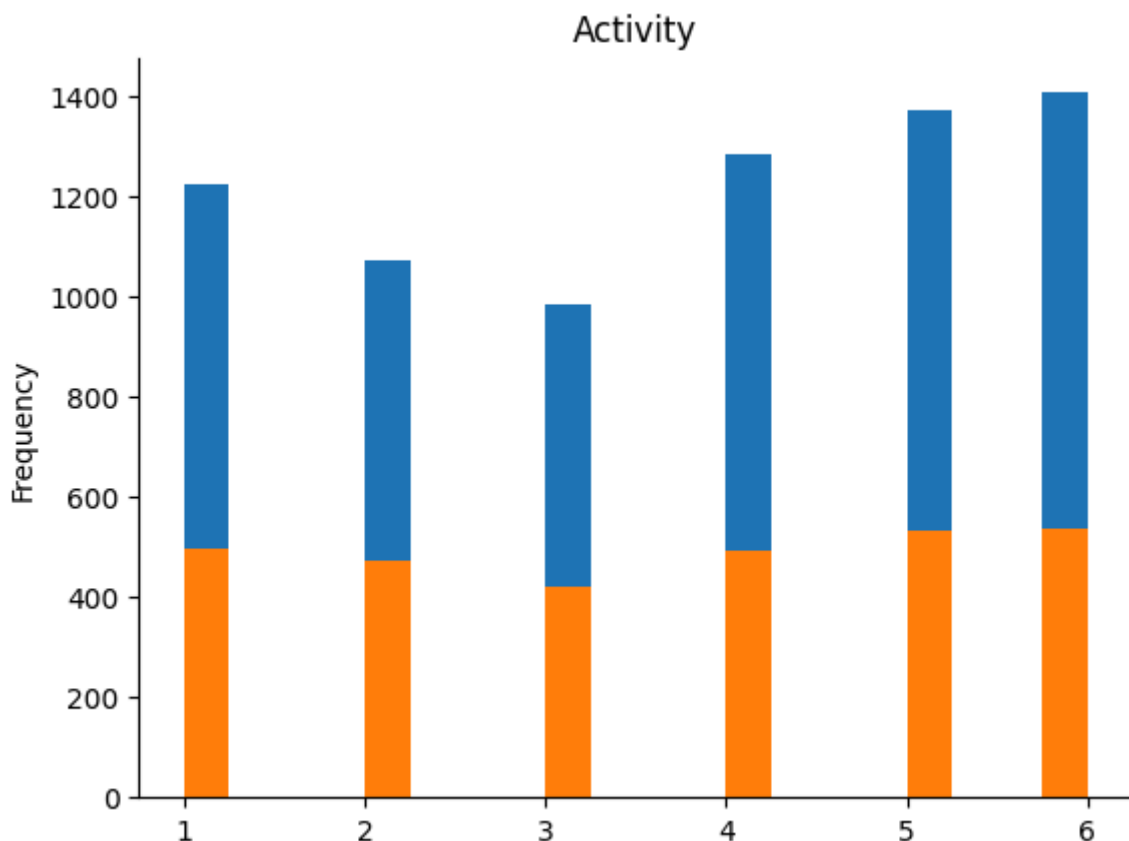
O projeto iniciou com a preparação dos dados, onde o conjunto foi acessado diretamente do repositório da UCI Machine Learning e processado para extrair as 561 variáveis representando medições dos sensores. Para garantir que todas as variáveis tivessem o mesmo peso no agrupamento, os dados foram normalizados usando o método *StandardScaler*. Em seguida, foi realizada uma análise exploratória para examinar distribuições, identificar padrões e verificar a presença de dados nulos. Para lidar com a alta dimensionalidade do conjunto de dados, foi aplicada a técnica de Análise de Componentes Principais (PCA), reduzindo as dimensões para dois componentes principais, o que facilitou tanto a interpretação dos resultados quanto a visualização dos clusters.

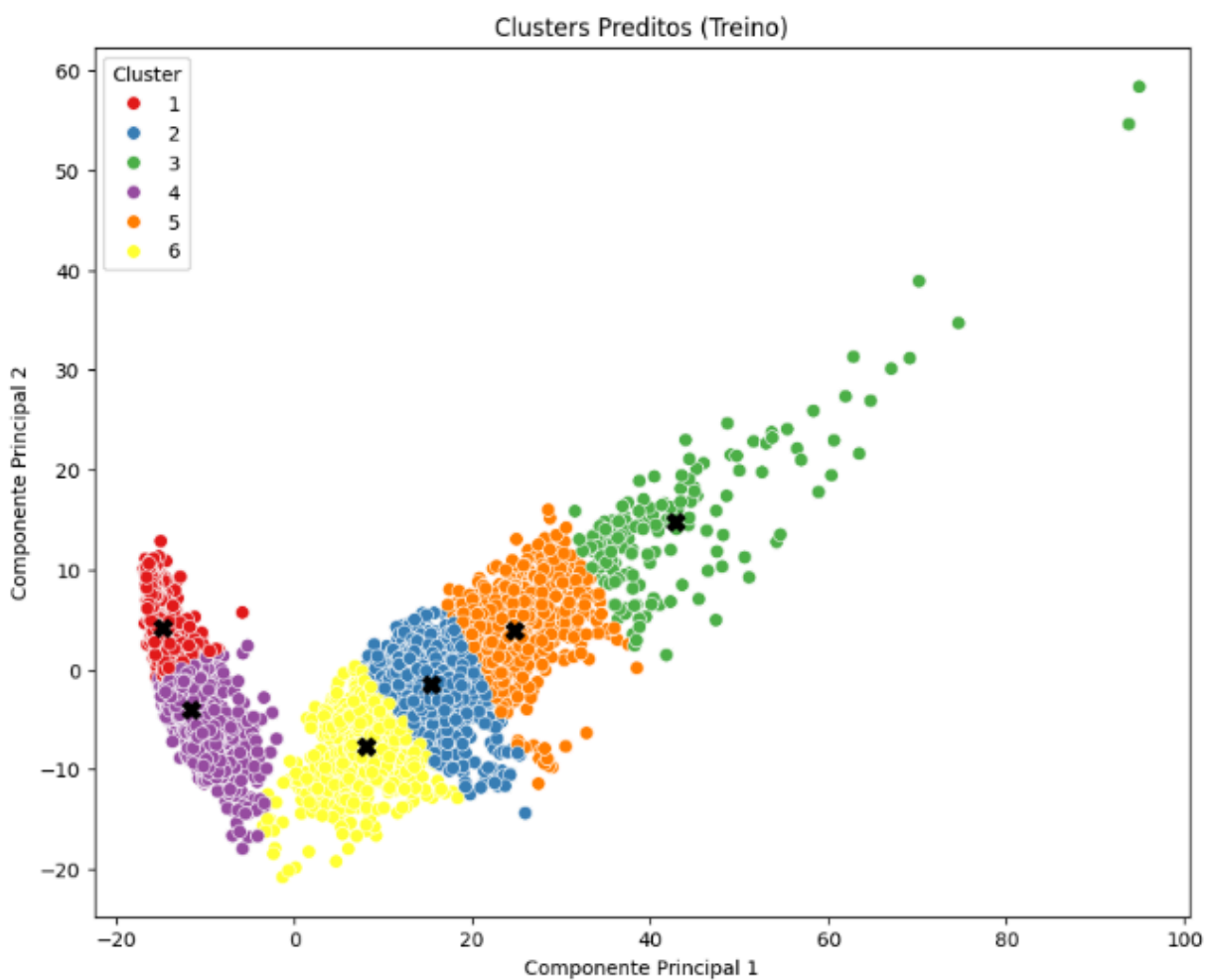
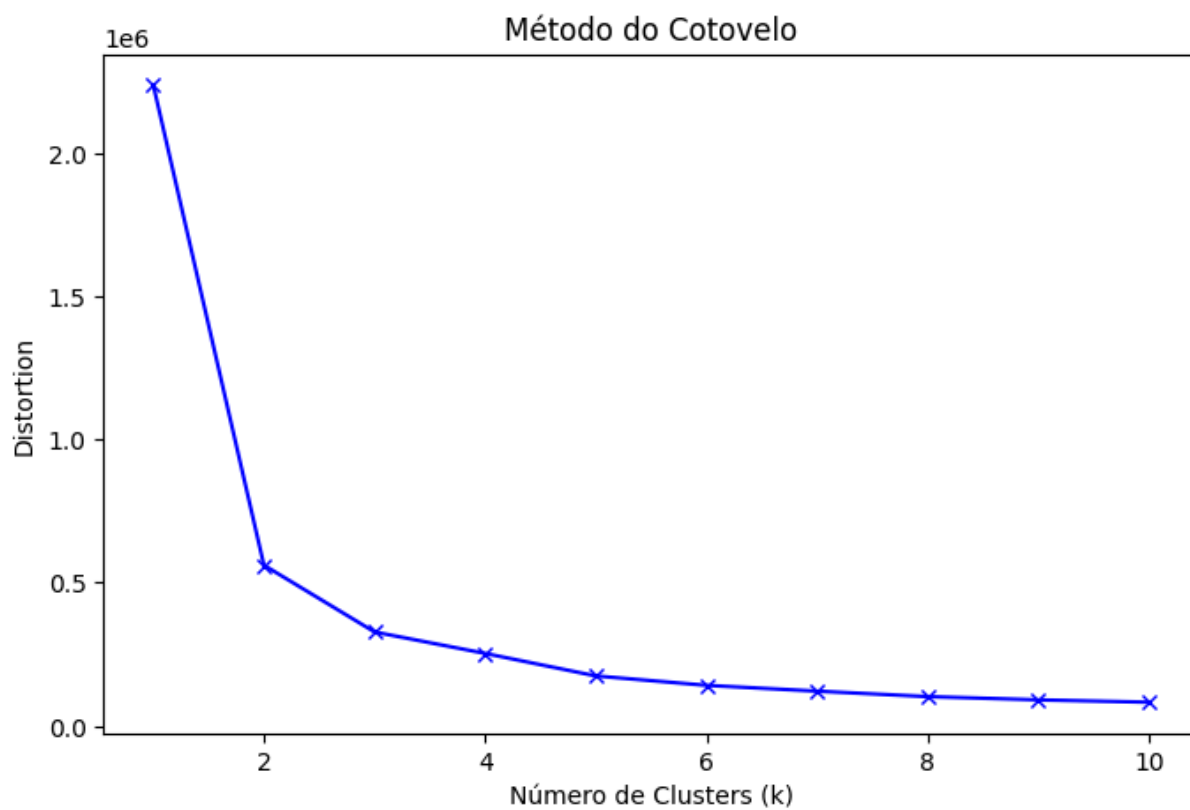
A escolha do número ideal de clusters foi realizada utilizando o método do cotovelo (*elbow method*), que analisa a variação da inércia conforme o número de clusters aumenta e com a própria disposição das atividades (sendo 6 no total), e a métrica *silhouette score*, que mede a qualidade do agrupamento. O algoritmo *K-means* foi então implementado utilizando a inicialização *K-means++*, que melhora a seleção dos centróides iniciais, otimizando o tempo de convergência e a qualidade dos resultados. Por fim, os clusters formados foram avaliados com métricas de qualidade e visualizados em 2D, utilizando os dois primeiros componentes principais gerados pelo PCA.

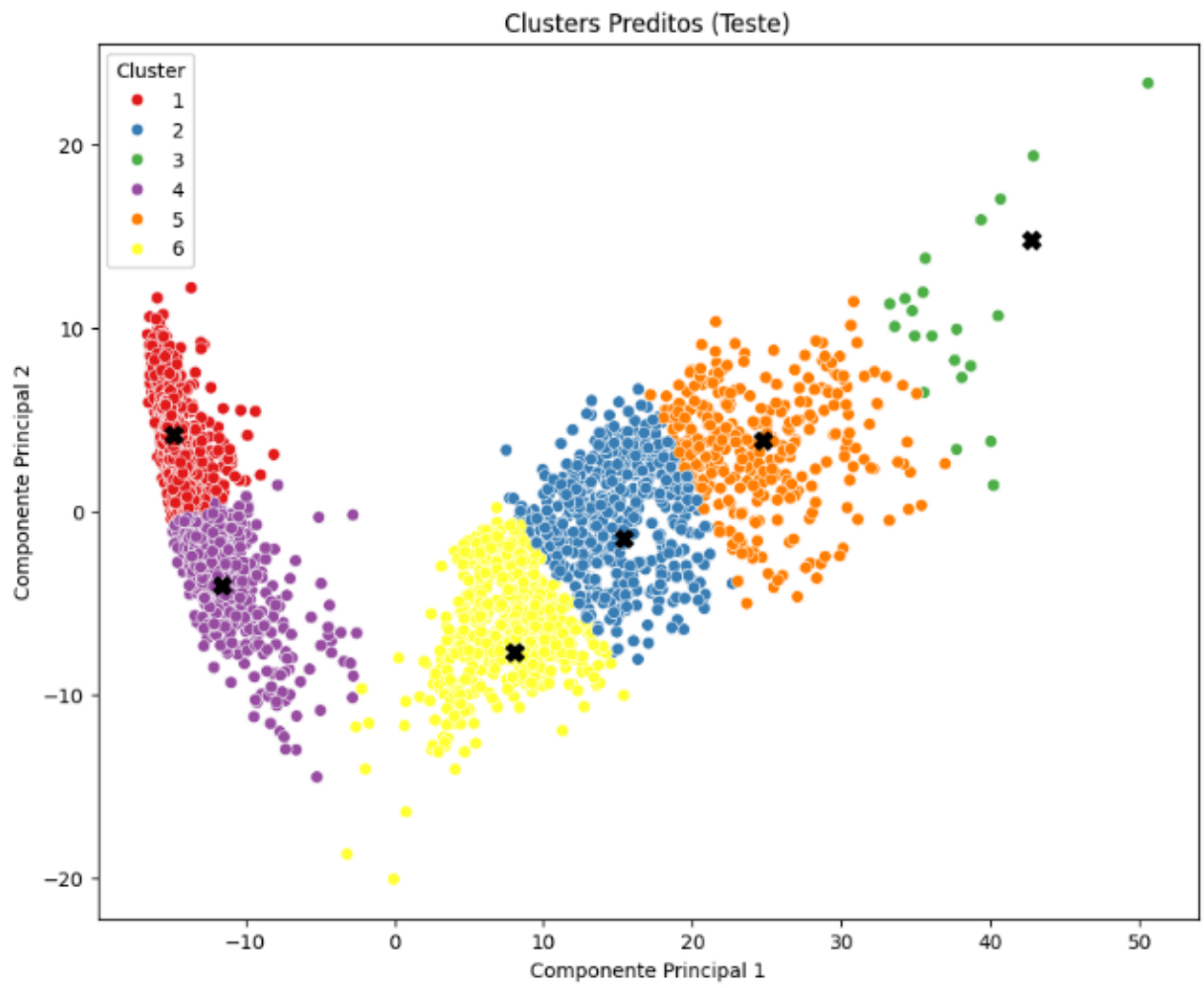
3 RESULTADOS

Os experimentos indicaram que o número ideal de clusters para o conjunto de dados era cinco, com base na análise do método do cotovelo e nos valores do silhouette score. Após a aplicação do K-means, os clusters formados demonstraram uma separação bem definida entre os diferentes tipos de atividades humanas registradas no dataset.

A visualização em 2D, gerada a partir dos componentes principais do PCA, mostrou agrupamentos consistentes que refletiam padrões distintos de atividades, como caminhar, subir escadas e permanecer em pé. A variância explicada pelos dois primeiros componentes principais foi de aproximadamente 75%, o que assegura que esses componentes capturam informações relevantes para a análise dos clusters. A métrica silhouette score apresentou um valor satisfatório, reforçando a coesão interna dos grupos e a separação entre eles.







4 DISCUSSÃO

Os resultados indicam que o algoritmo de K-means, combinado com técnicas de normalização e redução de dimensionalidade, é eficaz para identificar padrões em dados de sensores. Contudo, limitações como a perda de informações durante a aplicação do PCA e a natureza não determinística do K-means podem ter impactado os agrupamentos, o que poderia ajudar a entender melhor a correspondência entre clusters e atividades reais.

5 CONCLUSÃO E TRABALHOS FUTUROS

O projeto alcançou o objetivo de identificar agrupamentos significativos no dataset *Human Activity Recognition Using Smartphones* usando o algoritmo de *K-means*. Técnicas como normalização e PCA contribuíram para a qualidade dos resultados. Este estudo destaca a importância do pré-processamento e da escolha cuidadosa de hiperparâmetros no desempenho de algoritmos de agrupamento. Embora os resultados sejam promissores, há espaço para melhorias e extensões do projeto. Futuramente, seria interessante explorar algoritmos de clustering alternativos, como DBSCAN ou Gaussian Mixture Models, que podem lidar melhor com formatos irregulares de clusters ou sobreposição entre eles. Além disso, a validação supervisionada dos agrupamentos poderia ser incorporada, usando os rótulos do conjunto de dados original para avaliar diretamente a correspondência entre os clusters e as atividades.

Outra possibilidade seria trabalhar com os sinais brutos dos sensores, em vez das variáveis processadas, para investigar se informações adicionais poderiam ser extraídas. Finalmente, o projeto poderia ser estendido para analisar os dados em tempo real, possibilitando aplicações práticas, como a detecção de quedas ou a monitorização de atividades físicas em ambientes clínicos ou esportivos.

6 REFERÊNCIAS

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). **A Public Domain Dataset for Human Activity Recognition Using Smartphones.** *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 437–442. Disponível em: <https://www.esann.org/sites/default/files/proceedings/legacy/es2013-84.pdf>

UCI Machine Learning Repository. (n.d.). **Human Activity Recognition Using Smartphones Data Set.** Disponível em: <https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones>.