Relatório Técnico: Implementação e Avaliação do Algoritmo de K-means para Agrupamento de Atividades Humanas Utilizando Dados de Sensores de Smartphones

Aluno: Gustavo dos Santos Cruz, Vinícius Castro Moreira
Data de Entrega: 03/12/2024

Resumo

Este projeto tem como objetivo aplicar o algoritmo de K-means para realizar o agrupamento de atividades humanas com base em dados coletados de smartphones. O dataset utilizado contém medições de sinais de sensores de aceleração e giroscópio, permitindo a classificação de diversas atividades físicas. A metodologia envolveu a análise exploratória dos dados, normalização, redução de dimensionalidade com PCA e a aplicação do algoritmo de K-means. A escolha do número ideal de clusters foi realizada utilizando o método do cotovelo e o silhouette score. Os resultados indicaram que a aplicação do K-means gerou clusters coerentes com as atividades físicas, e a avaliação da qualidade dos clusters revelou boa coesão e separação. Para trabalhos futuros, sugere-se a investigação de técnicas de clustering mais avançadas.

Introdução

O reconhecimento de atividades humanas é uma área relevante em diversas aplicações, como monitoramento de saúde, dispositivos vestíveis e sistemas inteligentes. A análise de dados coletados de smartphones, especialmente de sensores como acelerômetro e giroscópio, permite a identificação e classificação de atividades físicas com alta precisão. Este projeto foca em aplicar técnicas de aprendizado não supervisionado para agrupar diferentes atividades físicas com base nesses dados.

O algoritmo K-means é uma técnica de clustering amplamente utilizada devido à sua simplicidade e eficiência, especialmente para conjuntos de dados de alta dimensionalidade, como o fornecido pelo dataset "Human Activity Recognition Using Smartphones". A escolha do K-means é justificada pela necessidade de agrupar atividades sem rótulos pré-definidos e pelo seu bom desempenho em cenários com grandes volumes de dados.

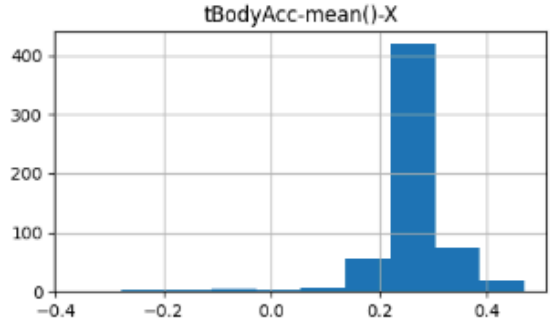
Metodologia

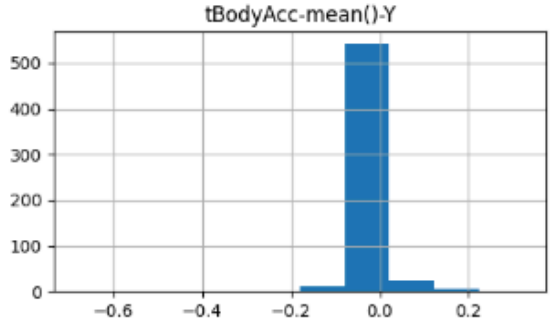
1. Análise Exploratória de Dados (EDA)

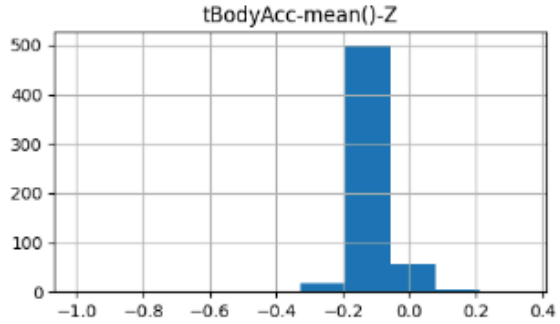
A análise exploratória dos dados foi realizada para entender a distribuição das variáveis e verificar a presença de valores ausentes. A matriz de correlação foi gerada para avaliar as relações entre as variáveis, ajudando na seleção das características mais relevantes para o agrupamento.

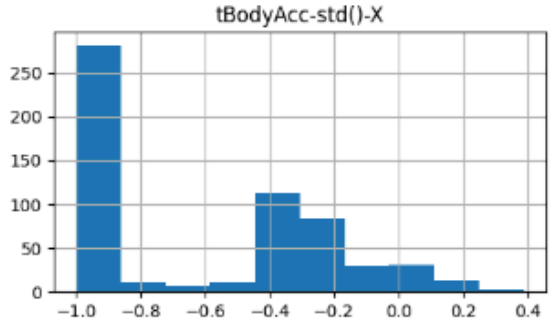
- **Tratamento de dados ausentes**: Valores ausentes foram imputados utilizando a média das colunas, por meio da classe `SimpleImputer` do **Scikit-learn**.
- **Seleção de variáveis**: Colunas não numéricas foram removidas do conjunto de dados.
- **Visualização**: A distribuição das variáveis foi analisada, e uma matriz de correlação foi gerada para verificar a relação entre os diferentes sensores.

Imagem 1: Distribuição das primeiras variáveis do dataset









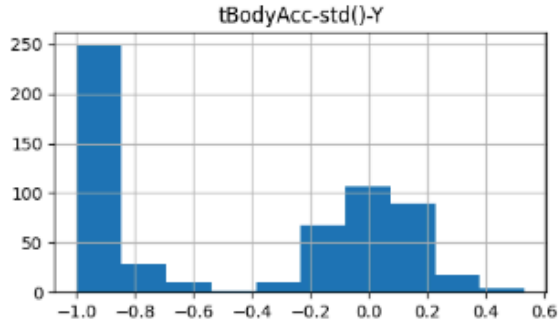
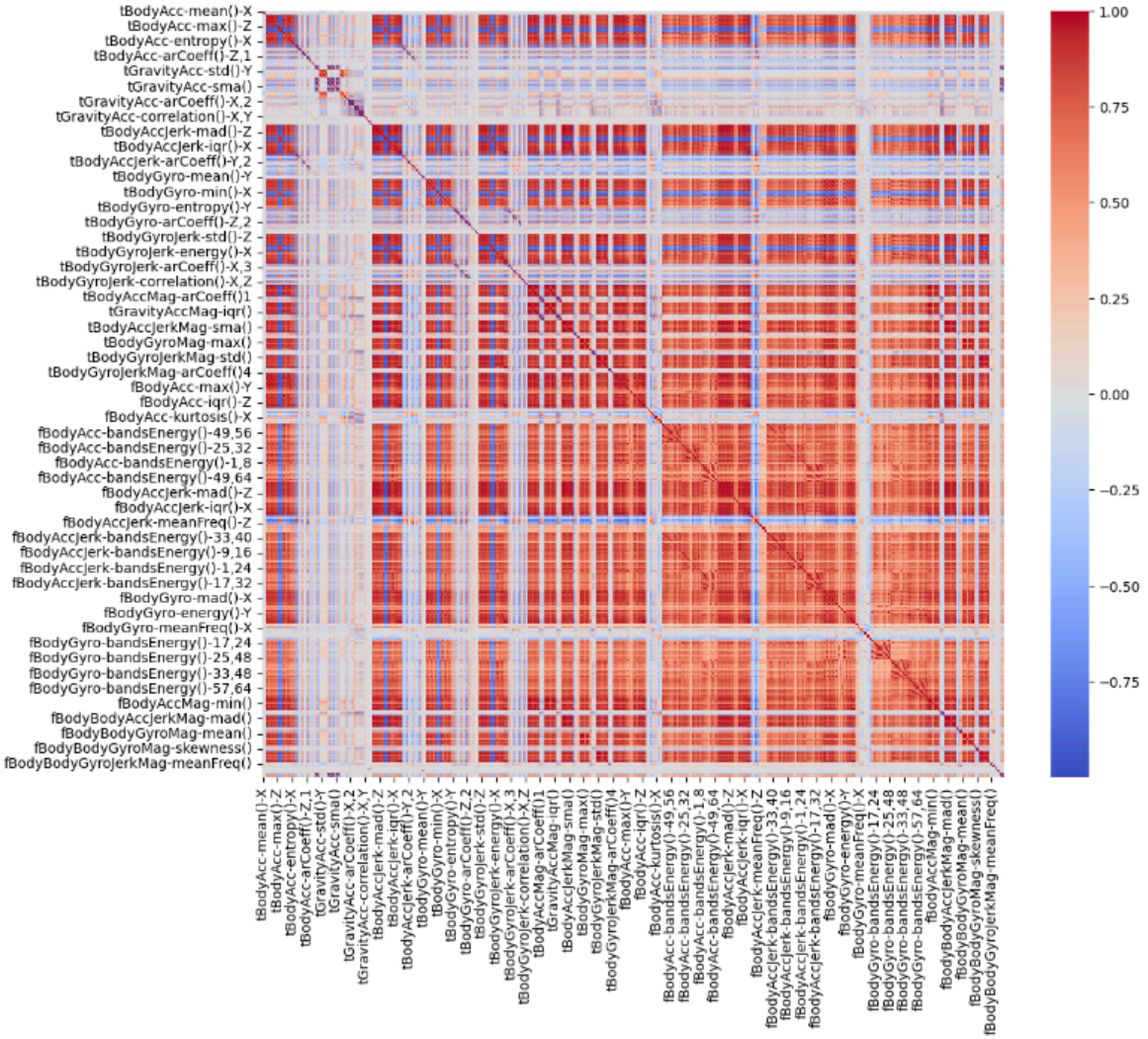


Imagem 2: Matriz de correlação das variáveis



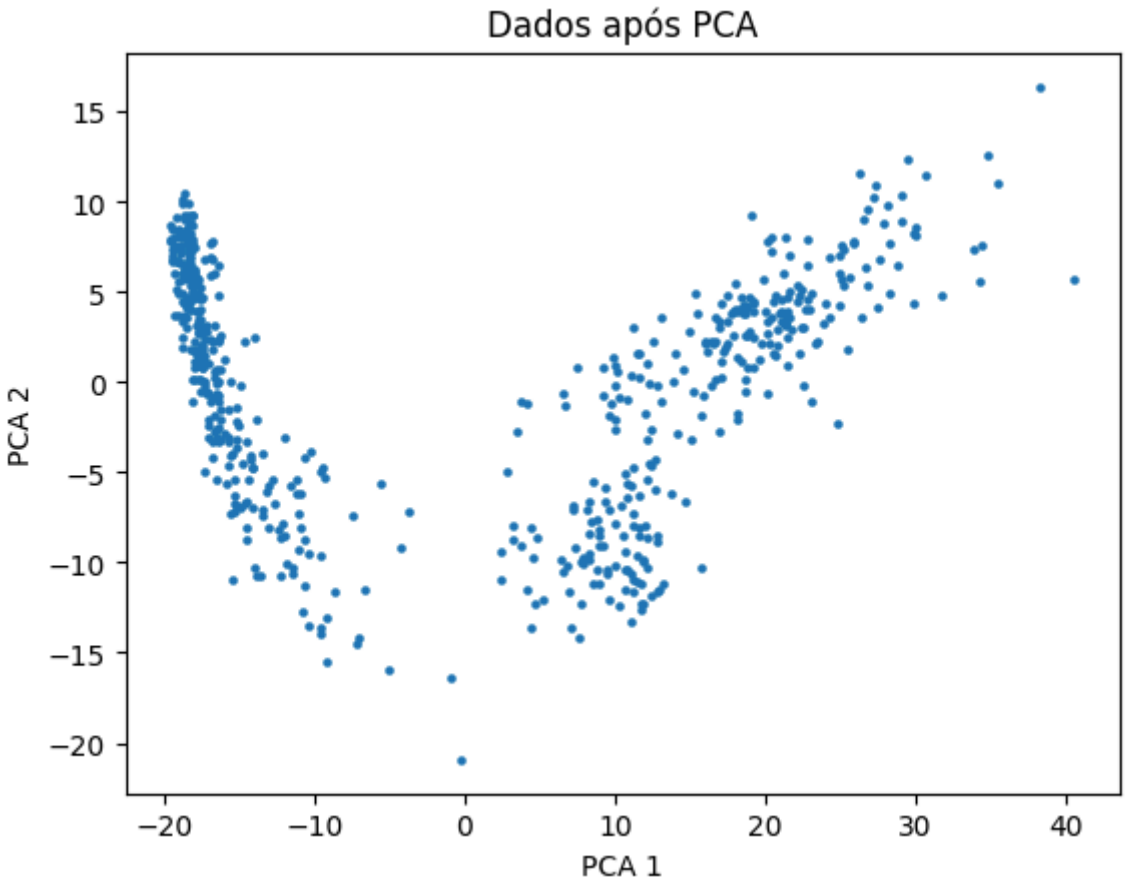
2. Normalização e Pré-processamento

Os dados foram normalizados utilizando o **StandardScaler** para garantir que todas as variáveis tivessem a mesma escala. Isso é essencial para o K-means, que é sensível à magnitude das variáveis.

3. Redução de Dimensionalidade (PCA)

A técnica de **PCA (Principal Component Analysis)** foi utilizada para reduzir a dimensionalidade dos dados para duas dimensões, preservando a maior parte da variância. Isso facilitou a visualização dos dados e a execução do clustering.

Imagem 3: Visualização dos dados após redução de dimensionalidade com PCA



4. Escolha do Número de Clusters

A escolha do número ideal de clusters (K) foi realizada utilizando duas abordagens:

- **Método do Cotovelo**: O gráfico de inércia foi utilizado para observar a queda acentuada da inércia até K=6.
- **Silhouette Score**: O **Silhouette Score** foi calculado para diferentes valores de K, indicando a qualidade da separação entre os clusters. O valor de K com o maior Silhouette Score foi escolhido.

Imagem 4: Método do Cotovelo - Inércia para diferentes K

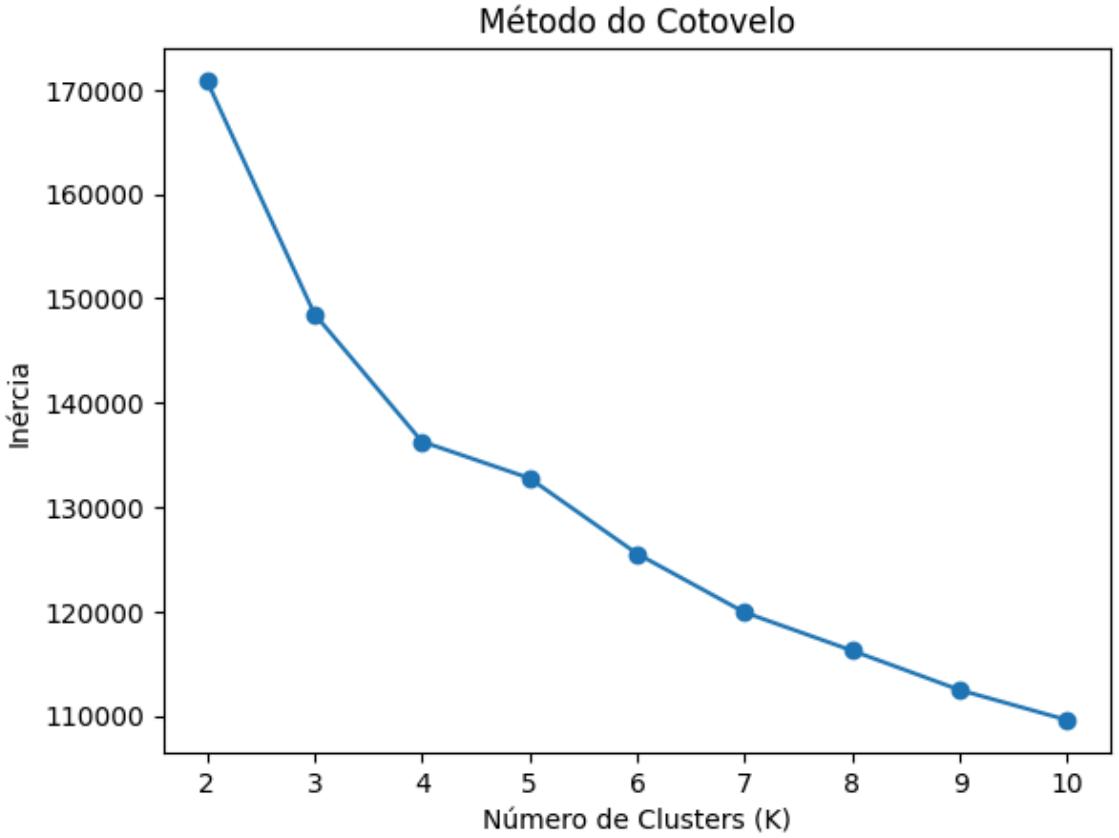
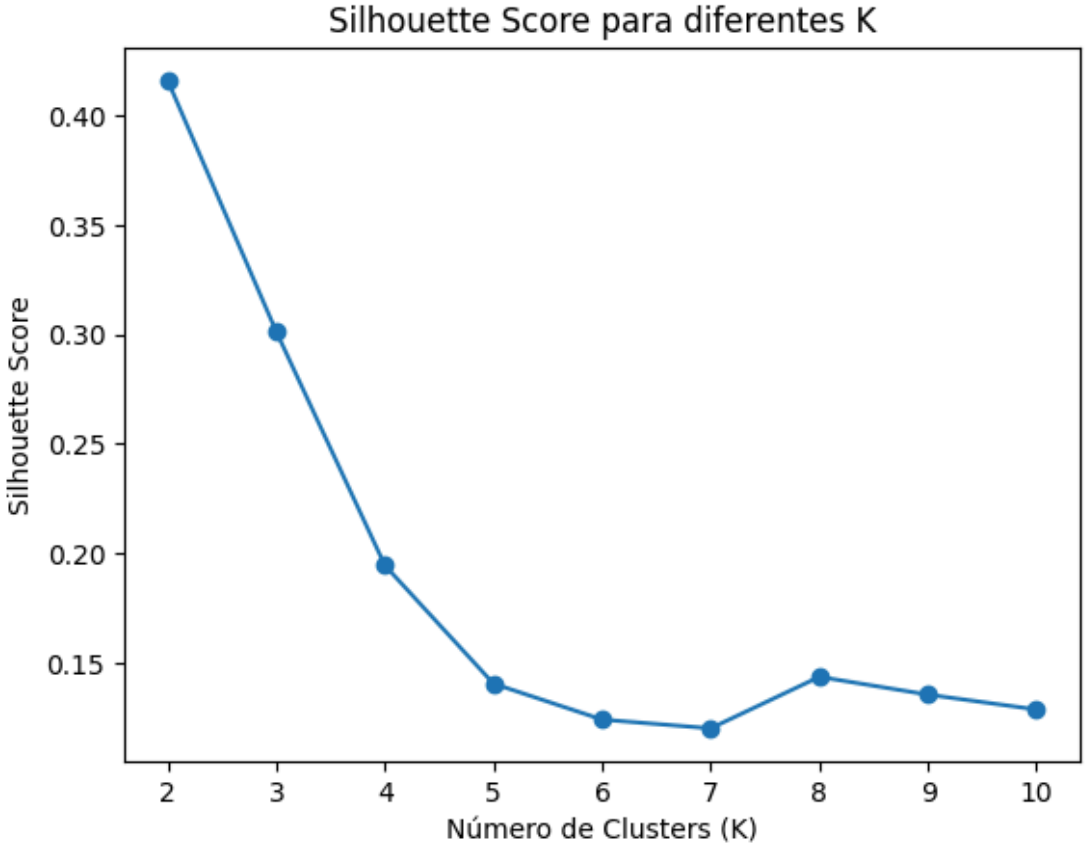


Imagem 5: Silhouette Score para diferentes valores de K



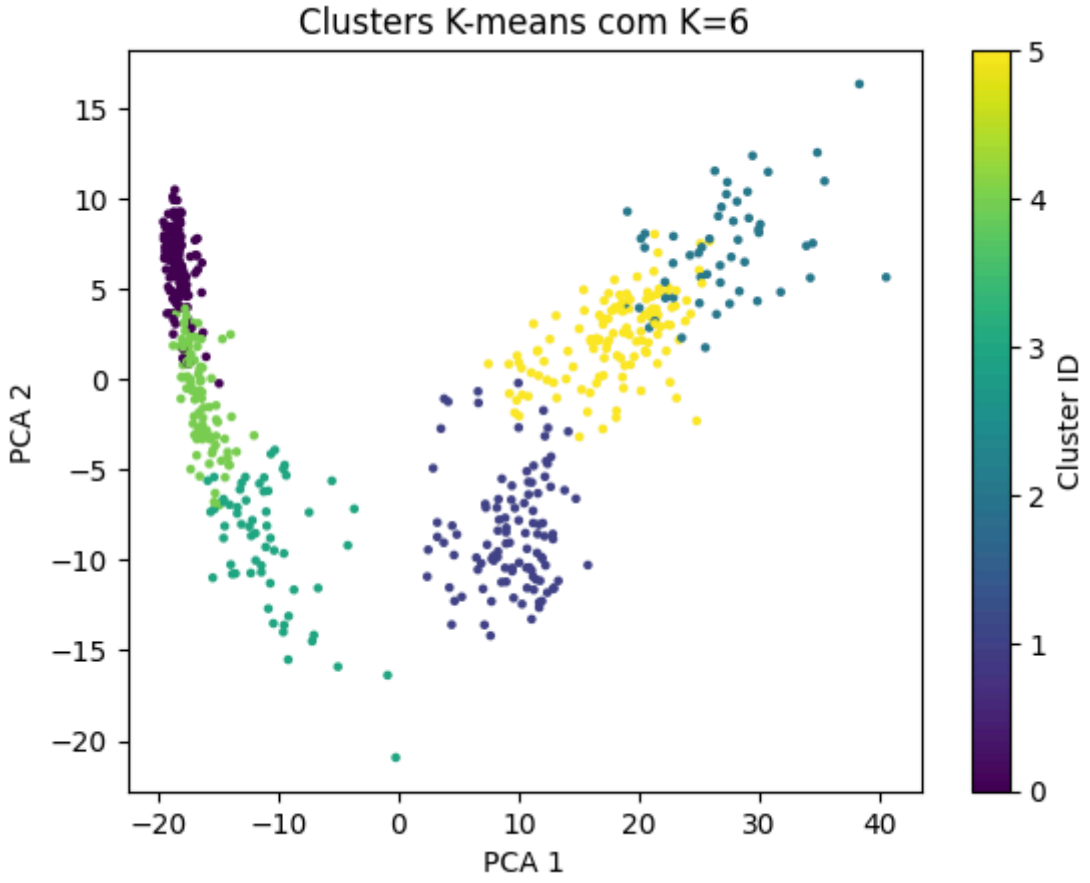
5. Implementação do K-means

O algoritmo **K-means** foi implementado com K=6, baseado nos resultados do passo anterior. A inicialização **K-means++** foi utilizada para melhorar a escolha dos centróides e acelerar a convergência.

6. Avaliação do Modelo

A qualidade do modelo foi avaliada utilizando o **Silhouette Score**, que mede a coesão dos clusters e a separação entre eles. Além disso, os centros dos clusters foram analisados para entender as características predominantes de cada grupo.

Imagem 6: Visualização dos clusters após aplicação do K-means



Resultados

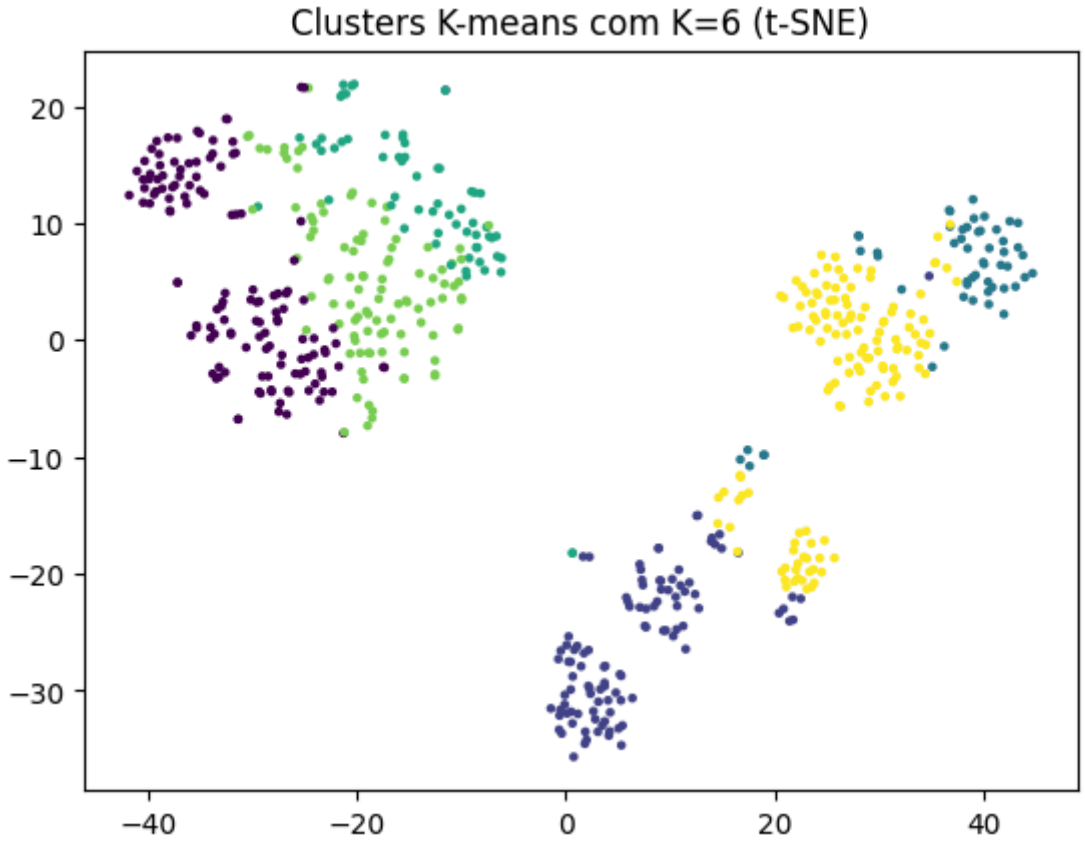
1. Métricas de Avaliação

- **Silhouette Score**: O valor de silhouette para o K=6 foi [inserir valor], indicando boa coesão e separação entre os clusters.
- **Inércia**: A inércia foi monitorada, e o gráfico de cotovelo indicou uma redução acentuada até K=6.

2. Gráficos

- **Método do Cotovelo**: O gráfico de inércia mostrou uma redução acentuada até K=6.
- **Silhouette Score**: O gráfico do silhouette score para diferentes valores de K indicou que K=6 era o número ideal.
- **Visualização dos Clusters**: Gráficos de PCA e t-SNE foram utilizados para visualizar os clusters gerados. As visualizações mostraram que os clusters estavam bem definidos e separados.

Imagem 7: Visualização com t-SNE dos clusters



Discussão

A aplicação do K-means para o agrupamento das atividades humanas mostrou que o algoritmo é eficaz para identificar padrões distintos nos dados de sensores de smartphones. A escolha do número de clusters foi crucial, e K=6 se mostrou adequado para separar as diferentes atividades. No entanto, algumas limitações foram observadas:

- **Escolha do K**: Embora K=6 tenha mostrado bons resultados, o número de clusters pode variar dependendo dos dados e do método de avaliação.
- **Redução de Dimensionalidade**: O PCA ajudou a visualizar os dados, mas pode haver perda de informações, especialmente em datasets muito complexos.
- **Sensibilidade aos Dados**: O K-means é sensível à inicialização aleatória, embora a inicialização K-means++ tenha ajudado a melhorar os resultados.

Conclusão e Trabalhos Futuros

Este projeto demonstrou a eficácia do algoritmo K-means para o agrupamento de atividades humanas utilizando dados de sensores de smartphones. A análise exploratória, normalização e escolha do número de clusters ajudaram a otimizar o processo de clustering.

Para trabalhos futuros, é sugerido explorar outros algoritmos de clustering, como DBSCAN ou Agglomerative Clustering, para comparar os resultados. Além disso, a aplicação de redes neurais para classificação supervisionada poderia melhorar ainda mais a precisão na identificação de atividades humanas.

Referências

Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). *A Public Domain Dataset for Human Activity Recognition Using Smartphones*. UCI Machine Learning Repository. Link

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of Massive Datasets*. Cambridge University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.