Relatório Técnico: Análise de Clustering com K-means

Grupo 37

Ana Fernanda Marinho Azevedo Alves

Reynaldo de Oliveira Santos

Data de entrega: 03 de dezembro de 2024

Resumo

Este relatório documenta o projeto de aplicação do algoritmo K-means no dataset “Human Activity Recognition Using Smartphones”. Este dataset é um Banco de dados de reconhecimento de atividade humana criado a partir de gravações de 30 indivíduos realizando atividades da vida diária (AVD) enquanto carregavam um smartphone preso à cintura com sensores inerciais incorporados. O objetivo foi explorar e agrupar atividades humanas baseadas em dados de sensores. O projeto envolveu análise exploratória, normalização dos dados, escolha do número ideal de clusters e avaliação dos resultados. Os principais resultados incluem uma análise de coesão dos clusters e insights sobre a separação das atividades humanas no espaço vetorial.

Introdução

O reconhecimento de atividades humanas a partir de dados de sensores tem aplicações relevantes em áreas como saúde, fitness e monitoramento remoto. Os experimentos foram realizados com um grupo de 30 voluntários na faixa etária de 19 a 48 anos. Cada pessoa realizou seis atividades (Andar, Andar_Para_Cima, Andar_Para_Baixo, Sentar, Ficar_De_Pé, Deitar) usando um smartphone (Samsung Galaxy S II) na cintura. O dataset utilizado contém medições de sensores acelerômetros e giroscópios coletados de smartphones.

Este projeto utiliza o algoritmo de clustering K-means para agrupar os dados e explorar a separação entre diferentes atividades. A escolha do K-means é justificada pela sua simplicidade e eficiência em dados de alta dimensionalidade. Esse algoritmo divide os dados em um número pré-definido de grupos (clusters), tentando garantir que os pontos dentro de cada grupo sejam os mais semelhantes possível e os grupos sejam bem distintos entre si. A aplicação de técnicas como

o método do cotovelo e o coeficiente de silhueta permite determinar o número ideal de clusters.

Metodologia

O desenvolvimento do projeto seguiu as etapas descritas abaixo:

1. **Análise Exploratória dos Dados:**

   - Verificação da estrutura do dataset e identificação de features duplicadas.
   - Avaliação da distribuição das atividades humanas registradas.

2. **Pré-processamento:**

- Normalização dos dados com `StandardScaler` para garantir igualdade de peso entre as variáveis, dada a sensibilidade do K-means à escala.

3. **Escolha do Número de Clusters:**

- Utilização do método do cotovelo e do coeficiente de silhueta para determinar o número ideal de clusters.

4. **Implementação do K-means:**

- Aplicação do algoritmo K-means ao conjunto de treinamento normalizado com diferentes valores de `k` (número de clusters).

5. **Avaliação e Visualização:**

- Avaliação dos clusters usando coeficiente de silhueta.
- Visualização das distribuições com PCA e mapeamento entre clusters e atividades reais.

Resultados

O conjunto de dados usado (Reconhecimento de Atividade Humana) tem 6 atividades, o que sugere que idealmente o número de clusters deve estar em torno de 6.

Se o método do conjunto sugere menos clusters, isso pode indicar que algumas atividades têm características semelhantes, sendo agrupadas juntas.

O método do cotovelo (Elbow Method) é uma técnica amplamente utilizada para determinar o número ideal de clusters para o algoritmo K-means. Ele avalia a soma das distâncias quadradas dentro do cluster (Within-Cluster Sum of Squares - WCSS) para diferentes números de clusters e identifica o ponto em que o WCSS diminui mais lentamente (o "cotovelo").

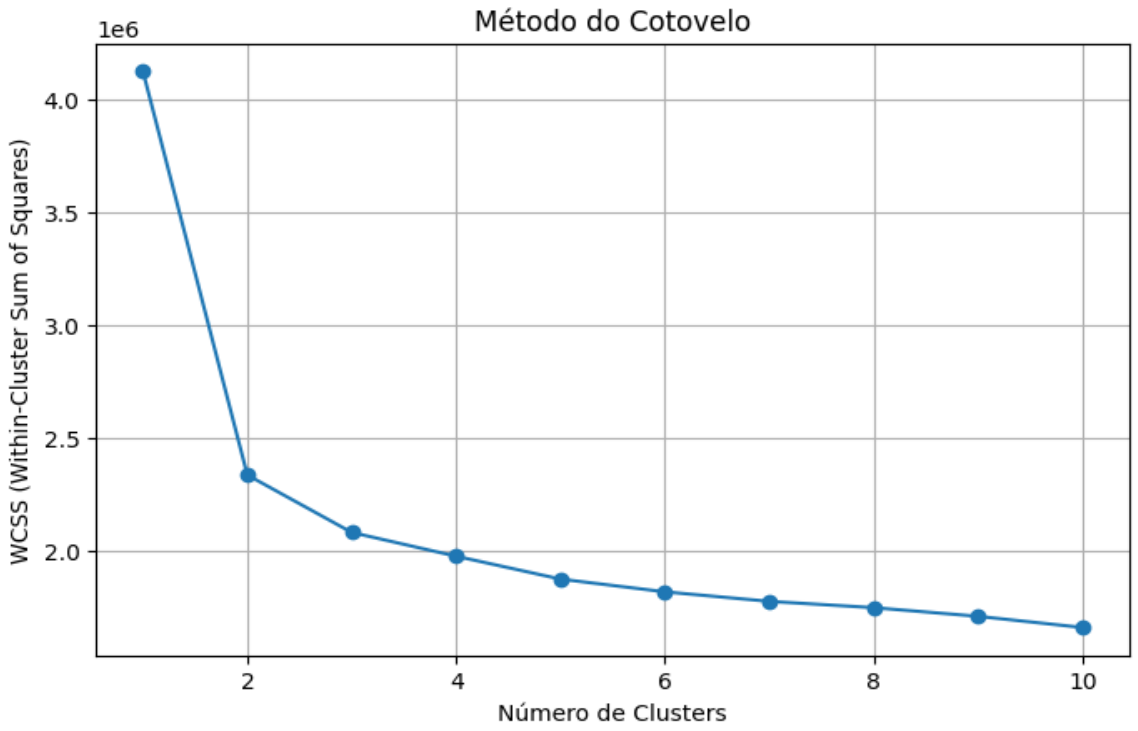

Figura 1 - Método do Cotovelo (Fonte: COLAB)

Os resultados do K-means mostraram uma separação moderada entre as atividades humanas. Os resultados do K-means mostraram:

Coeficiente de Silhueta:

- Para 2 clusters: 0.329 (melhor separação entre clusters).
- Para 6 clusters: 0.109 (moderada correspondência com as 6 atividades reais).

Os clusters foram avaliados usando o coeficiente de silhueta que indicou uma qualidade de clustering aceitável (0,109). A imagem de silhueta (Figura 2), é uma representação visual que avalia a qualidade de clusters gerados pelo K-means. Cada ponto no gráfico representa uma amostra, com seu valor inferior e o quão bem ela se ajusta ao seu cluster em comparação com clusters vizinhos.

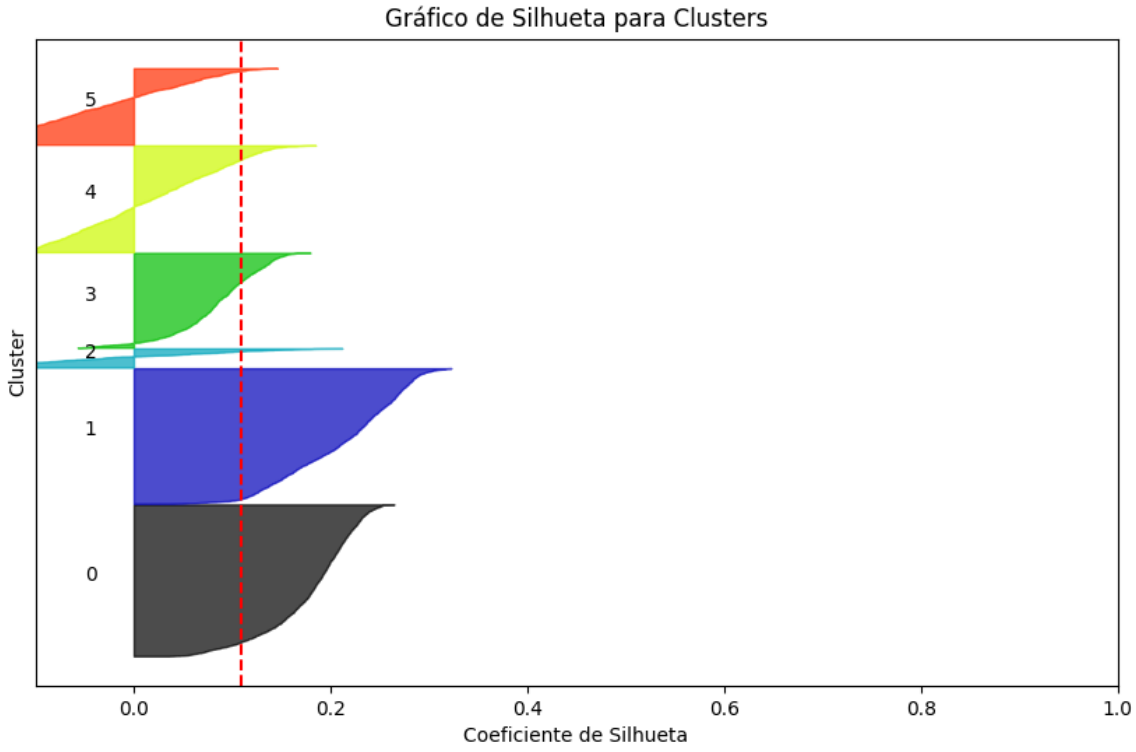

Figura 2 - Gráfico de silueta para clusters (Fonte: COLAB)

A visualização com PCA (Figura 3) revelou uma boa separação para atividades específicas como “Laying” e “Walking”, enquanto outras atividades apresentaram sobreposição significativa.

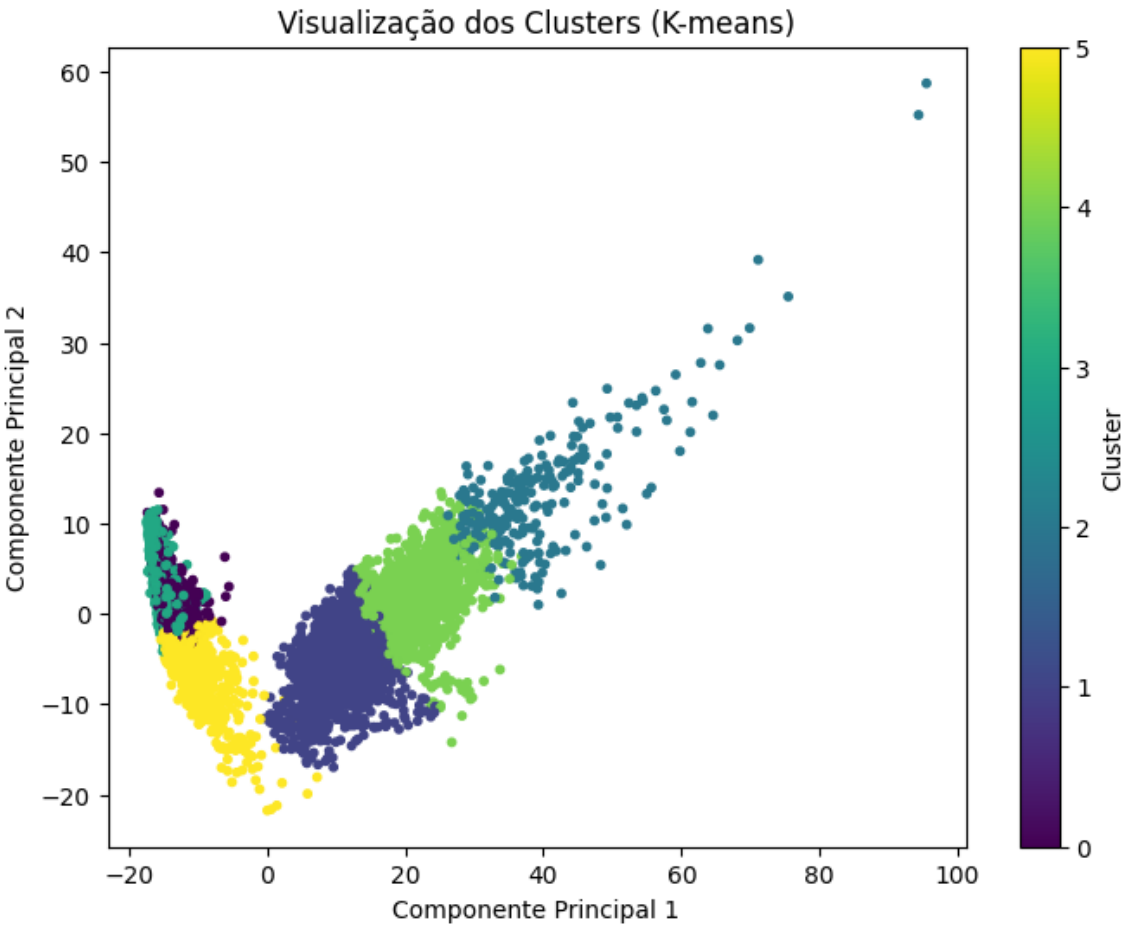

Figura 1 - Visualização dos Clusters (Fonte: COLAB)

Discussão

Os resultados (Figura 4) indicam que o K-means é eficaz para agrupar atividades humanas bem definidas, como “Laying” e “Walking”, mas apresenta limitações na separação de atividades semelhantes. O pré-processamento foi essencial para garantir a eficiência do modelo, mas mesmo com a normalização, as atividades

que envolvem posturas estáticas (ex.: “Sitting” e “Standing”) mostraram-se difíceis de distinguir.

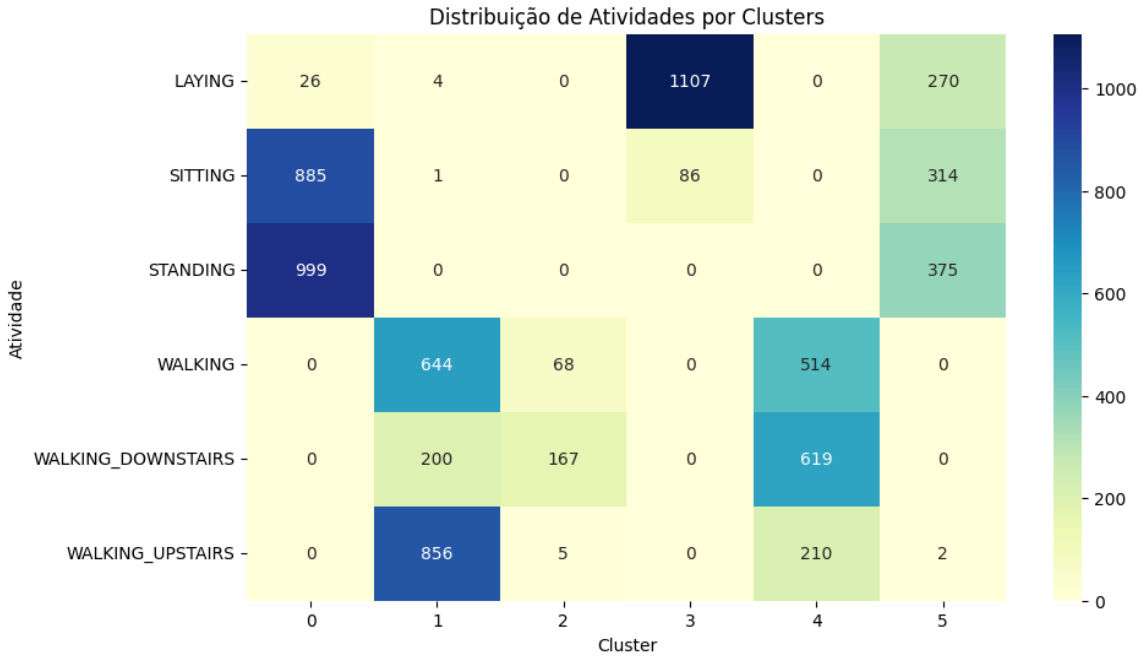

Figura 2 - Correspondência entre atividades reais e clusters (Fonte: COLAB)

Limitações:

Limitações do projeto incluem a sensibilidade do K-means à inicialização dos clusters e a natureza não supervisionada do algoritmo, que dificulta a correspondência direta com as atividades reais.

- O modelo assume relações lineares, o que pode não capturar padrões mais complexos.
- A normalização melhorou a convergência, mas não alterou significativamente as previsões.

Conclusão e Trabalhos Futuros

O projeto demonstrou o uso do K-means para explorar a separação de atividades humanas em dados de sensores. Os principais aprendizados incluem:

- A importância do pré-processamento e normalização dos dados.

- O impacto de técnicas de avaliação, como coeficiente de silhueta e método do cotovelo, na escolha do número de clusters.

Trabalhos futuros podem incluir a aplicação de algoritmos mais sofisticados, como modelos de mistura gaussiana, e a incorporação de métodos de redução de dimensionalidade, como t-SNE, para melhorar a visualização e separação dos clusters.

Referências

Bishop, C. M., **Pattern Recognition and Machine Learning**, Springer, 2006.

Grus, Joel. **Data Science do Zero – Primeiras Regras com o Python.** Traduzido por Wellington Nascimento. – Rio de Janeiro: Alta Books, 2016.