Relatório Técnico

Implementação e Análise do Algoritmo de K-means referente às atividades Conjunto de dados Reconhecimento de atividade humana

Integrantes

Jeferson Braga Luz

Rafael Rodrigues Souza

Vitória da Conquista - BA

03 de dezembro de 2024

Sumário

Resumo

Neste trabalho, foi aplicado o algoritmo K-means para agrupar dados de um banco de dados de reconhecimento de atividades humanas. Após a análise exploratória e a redução de dimensionalidade com PCA, o algoritmo foi aplicado e o número ideal de clusters foi determinado utilizando o método do cotovelo e o silhouette score. Os resultados mostraram que o K-means foi capaz de identificar grupos distintos de atividades, com base nas características dos sensores.

Introdução

O reconhecimento de atividades humanas envolve a coleta e análise de dados de sensores para identificar padrões de movimento. Uma etapa crucial nesse processo é a agrupamento de dados, que permite organizar as informações em grupos coerentes. O algoritmo K-means, por sua simplicidade e eficiência, é amplamente utilizado para essa tarefa. Ao dividir os dados em clusters, o K-means revela insights valiosos sobre as diferentes atividades realizadas, facilitando a criação de modelos mais precisos e robustos para o reconhecimento de atividades.

Metodologia

Banco de dados:

Banco de dados extraído da página oficial informada e exportada para o drive o qual foi compartilhado em link público. Diante disso utilizamos a biblioteca “gdown” para fazer o download no próprio repositório e fazer a extração.

Análise Exploratória:

Evidenciado no retorno do código a dimensão do banco de dados com número de colunas e linhas, também foi mostrado as primeiras informações do banco (constando apenas 5 linhas) e por fim mostrado a distribuição de classes e suas respectivas contagens.

Normalização dos Dados:

Utiliza o escalonador “StandardScaler” para normalizar os dados de treino (X_train). A normalização transforma os dados em uma escala padrão, o que pode ser importante para o algoritmo K-means funcionar melhor.

Redução de Dimensionalidade com PCA:

Aplica a técnica PCA (Análise de Componentes Principais) para reduzir a dimensionalidade dos dados para 2 componentes principais. O PCA ajuda a identificar as variáveis que mais contribuem para a variância do dataset, permitindo a visualização em um espaço bidimensional. O qual os pontos foram coloridos de acordo com a classe.

Implementação do K-means:

Após analisar os histogramas dos principais valores e a matriz de correlação entre variáveis. Realizamos descrição detalhada do algoritmo, incluindo a inicialização dos centroides e o processo iterativo. Aplicamos os cálculos de inércia e silhouette score.

Escolha do número de clusters:

Utiliza o método do cotovelo para identificar o número ideal de clusters.

- O algoritmo K-means é executado para diferentes valores de K (de 2 a 10).
- A inércia (soma quadrada das distâncias entre os pontos e seus centroides) é calculada para cada valor de K.
- Um gráfico é plotado mostrando a inércia em função do número de clusters.
- O valor de K onde a curva da inércia começa a se estabilizar (cotovelo) é considerado o ideal.

Também calcula o Silhouette Score para cada valor de K. O Silhouette Score é uma medida de quão bem os pontos estão agrupados em seus respectivos clusters.

O código seleciona o valor de K com o melhor Silhouette Score como o número ideal de clusters (K=6 neste caso).

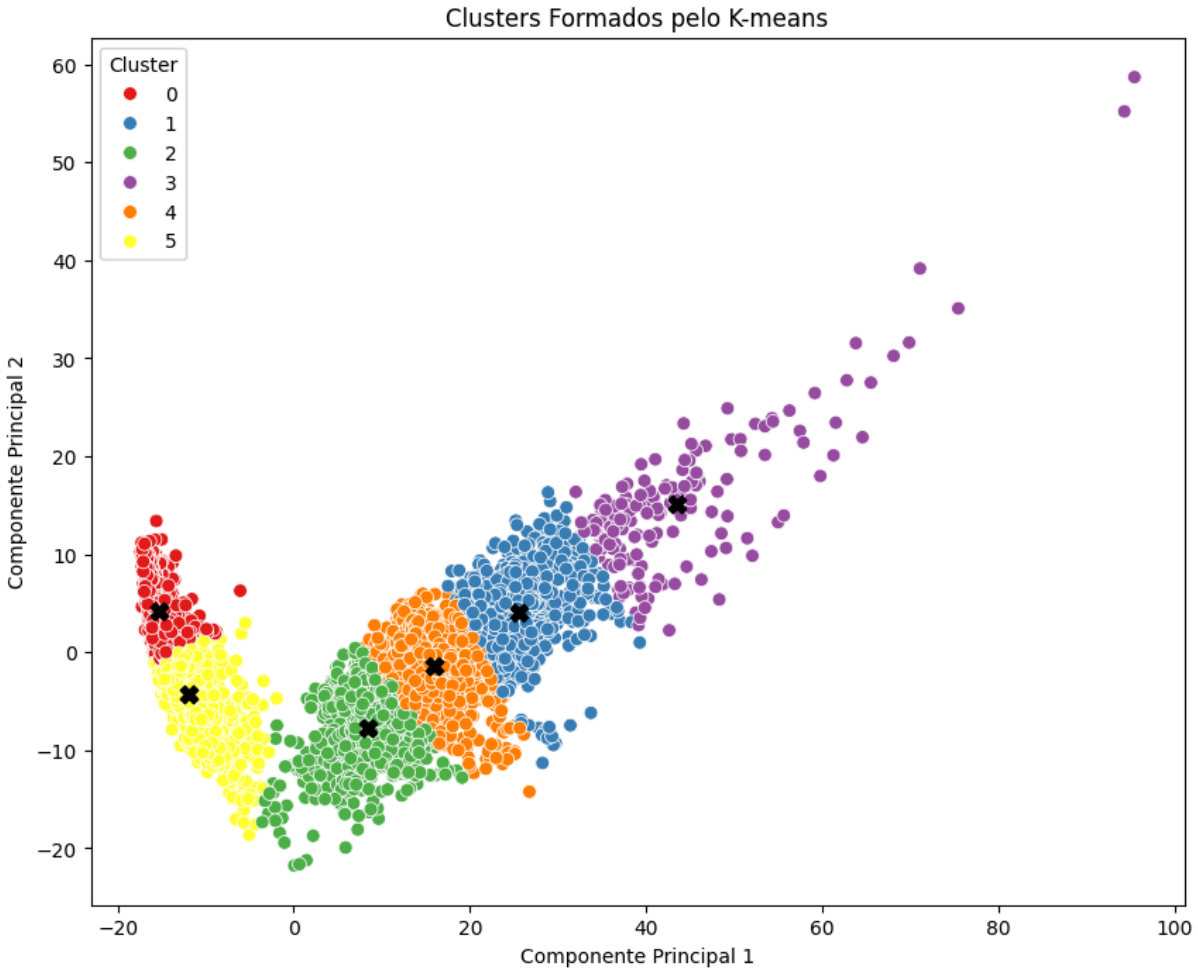
K-means e Visualização:

Diante a escolha do número de cluster executa-se o K-means com o número de clusters selecionado (K=6). Repetindo 10 vezes para validar a estabilidade do

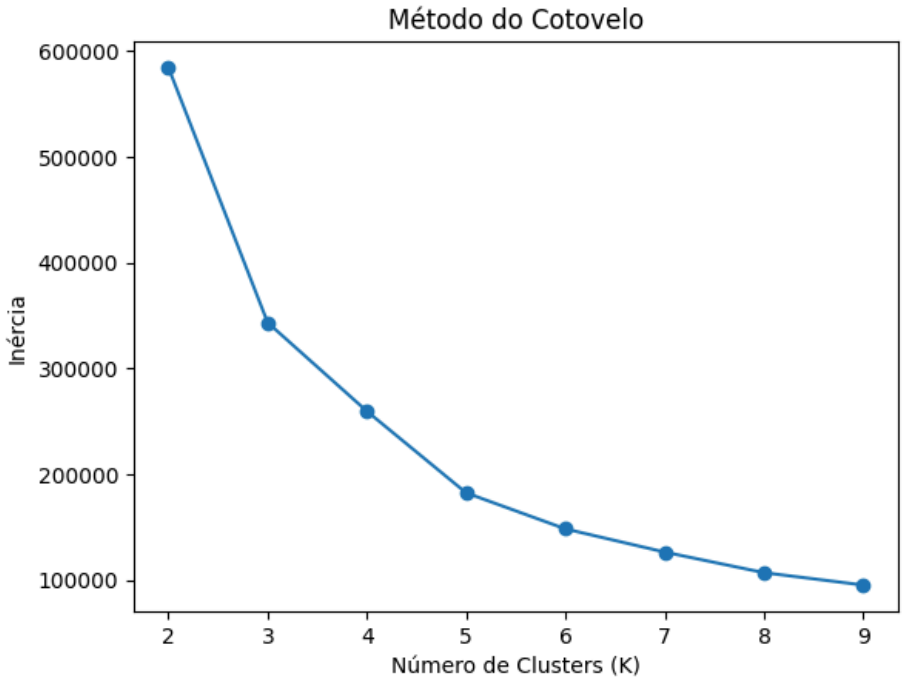
resultado. Visualiza os cluster formados pelo K-means no espaço bidimensional e calculamos o Silhouette Score final.

Resultados

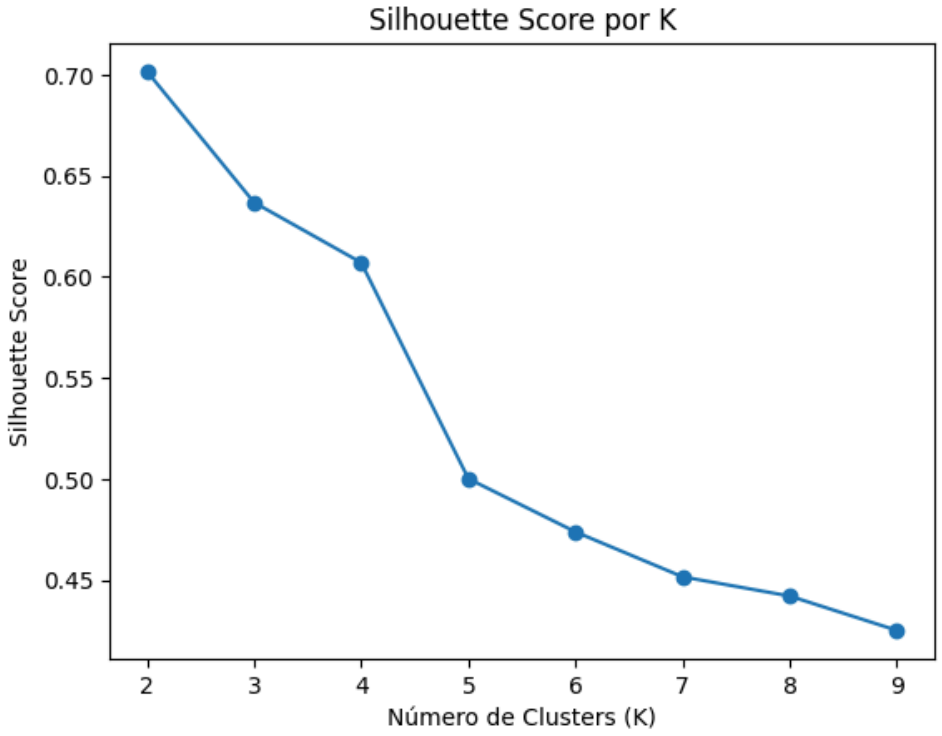
Métricas de Avaliação utilizada foi Silhouette Score. No qual o código calcula o Silhouette Score para diferentes valores de K (número de clusters) e para o valor final selecionado (K=6), varia entre -1 e 1, sendo 1 o melhor valor. Valores próximos a 1 indicam que os pontos estão bem agrupados dentro de seu próprio cluster e distantes de outros clusters.



Gráficos de Método do Cotovelo mostra a relação entre o número de clusters (K) e a inércia (soma das distâncias quadradas entre os pontos e seus respectivos centroides).



E o gráfico Silhouette Score por K mostra a relação entre o número de clusters (K) e o Silhouette Score. Valores mais altos de Silhouette Score indicam melhor separação entre os clusters.



Discussão

A escolha do número de clusters (K) é um aspecto crucial na aplicação do algoritmo K-means. Neste trabalho, o método do cotovelo e o Silhouette Score foram utilizados para determinar o valor ideal de K. No entanto, a interpretação do método do cotovelo pode ser subjetiva, e o Silhouette Score pode não ser a métrica mais adequada para todos os tipos de dados. Futuras pesquisas poderiam explorar outras métricas, como o índice de Davies-Bouldin, para avaliar a qualidade dos clusters. Além disso, a utilização de técnicas de validação cruzada poderia ajudar a estimular a generalização do modelo.

Conclusão

O presente estudo teve como objetivo aplicar o algoritmo K-means para agrupar dados de um conjunto de dados de reconhecimento de atividades humanas. A análise exploratória permitiu compreender as características dos dados, e a redução de dimensionalidade por meio do PCA facilitou a visualização e a interpretação dos resultados. O algoritmo K-means, por sua vez, demonstrou ser eficaz em identificar grupos distintos de atividades, baseando-se nas características dos sensores.

Os resultados obtidos evidenciam a importância da escolha adequada do número de clusters, que foi realizada com base em métricas como o método do cotovelo e o Silhouette Score. A análise dos clusters resultantes permitiu identificar padrões interessantes nos dados e obter insights valiosos sobre as diferentes atividades realizadas pelos indivíduos.

Este trabalho apresenta uma contribuição significativa para o campo do reconhecimento de atividades humanas ao demonstrar a eficácia do algoritmo K-means na identificação de padrões em dados de sensores. Além disso, a pesquisa propõe uma metodologia robusta para a seleção do número ideal de clusters e oferece uma análise detalhada dos resultados, contribuindo para a interpretabilidade dos modelos e a compreensão dos comportamentos humanos.

Com base nos resultados obtidos, diversas direções promissoras podem ser exploradas em futuras pesquisas. A incorporação de informações temporais e espaciais aos dados pode aprimorar a precisão dos modelos. A exploração de algoritmos de aprendizado profundo, como redes neurais recorrentes e convolucionais, pode permitir a extração de características mais complexas dos dados. Além disso, a aplicação dos modelos desenvolvidos em cenários reais, como monitoramento de pacientes ou análise de comportamento em ambientes de trabalho, pode gerar impactos significativos em diversas áreas. A comunidade científica pode se beneficiar da exploração de novos dados, da comparação com outros algoritmos de clustering e da desenvolvimento de interfaces mais intuitivas para a visualização e interpretação dos resultados.

Referências

1. Repositório GitHub:

LUZ, J. B. Projeto K-means. Disponível em: https://github.com/JefersonBLuz/ProjetoK-means. Acesso em: dia mês ano.

2. Conjunto de Dados UCI:

HARMON, D.; MONTERO, J.; REAL, J.; ÁLVAREZ, J.; TAPAS, M. Human activity recognition using smartphones. **UCI Machine Learning Repository** [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. 2013.

3. Artigo de Conferência:

PLATT, J. C.; COLTESCU, V. Large-margin DAGs for multiclass classification. In: **EUROPEAN SYMPOSIUM ON ARTIFICIAL NEURAL NETWORKS**, 13., 2005, Bruges. Proceedings... Bruges: d-side publications, 2005. p. 511-516.

4. Documentação Online:

SCIkit-learn. **Scikit-learn: Machine Learning in Python** [http://scikit-learn.org/stable/user_guide.html]. Acesso em: dia mês ano.