Relatório do Projeto de Análise de Atividades Humanas

Introdução

O projeto visa desenvolver um modelo de clustering para a análise de um conjunto de dados de atividades humanas coletados com sensores de smartphones. O dataset "Human Activity Recognition Using Smartphones", disponível no repositório da UCI Machine Learning, contém medições de 561 variáveis extraídas de acelerômetros e giroscópios de 30 voluntários em atividades cotidianas como caminhar, subir escadas e ficar em pé. Este estudo busca identificar padrões de comportamento e agrupar as atividades usando o algoritmo de K-means, destacando a importância da escolha do número de clusters, a normalização dos dados e a análise dos resultados. O uso do K-means é justificado por sua capacidade de segmentar grandes volumes de dados de forma eficiente e permitir a exploração visual e quantitativa dos clusters formados. A aplicação de técnicas como PCA para redução de dimensionalidade facilita a visualização e a interpretação dos grupos, enquanto a avaliação de métricas como o silhouette score e a inércia proporciona uma análise crítica da qualidade dos agrupamentos.

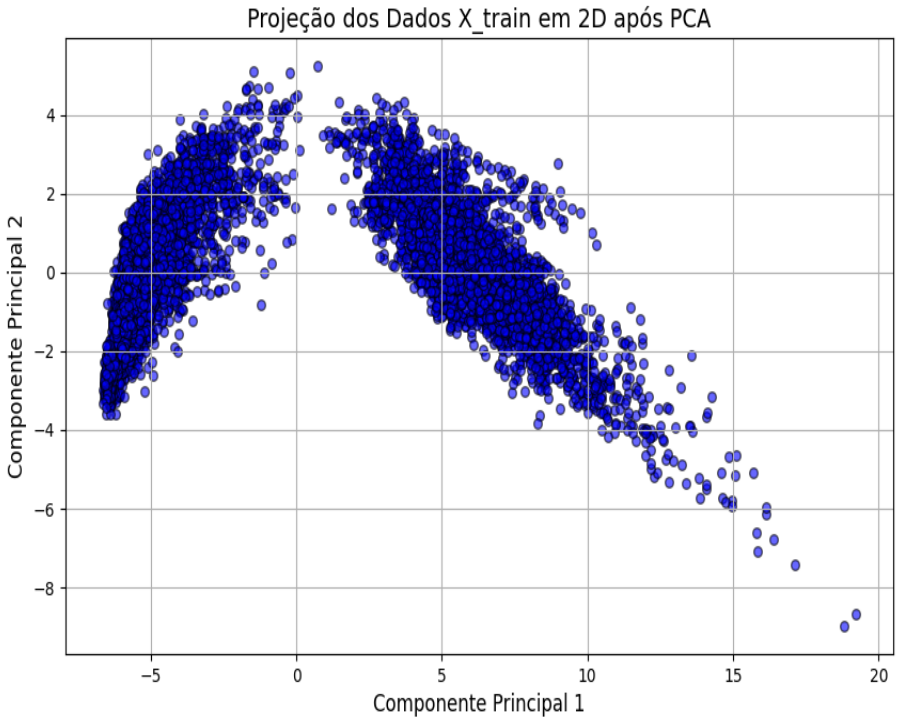
Metodologia

1. Análise Inicial do Dataset

A análise inicial do dataset revelou que os dados possuem o tipo float64, o que é comum para medições de sensores de alta precisão. O dataset contém um total de 7352 linhas e 561 colunas, representando um grande número de observações e variáveis. Importante destacar que, conforme descrito no README do dataset, os dados já estavam normalizados desde o início, o que assegura que todas as variáveis contribuam de forma equilibrada para as análises subsequentes.

2. Visualização 2D dos Dados

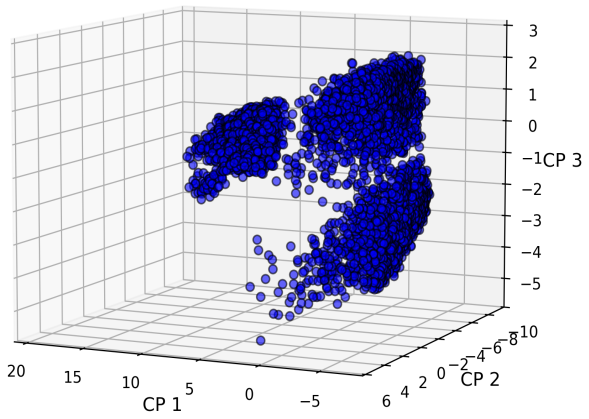
A visualização em 2D foi realizada para examinar a distribuição dos dados em um espaço bidimensional. A projeção dos dados após a aplicação da técnica de PCA revelou dois conjuntos distintos, evidenciando a separação de grupos dentro do dataset. Essa visualização fornece uma visão clara da estrutura dos dados, permitindo a identificação de possíveis agrupamentos e padrões que poderão ser explorados pelo algoritmo K-means.



3. Visualização 3D dos Dados

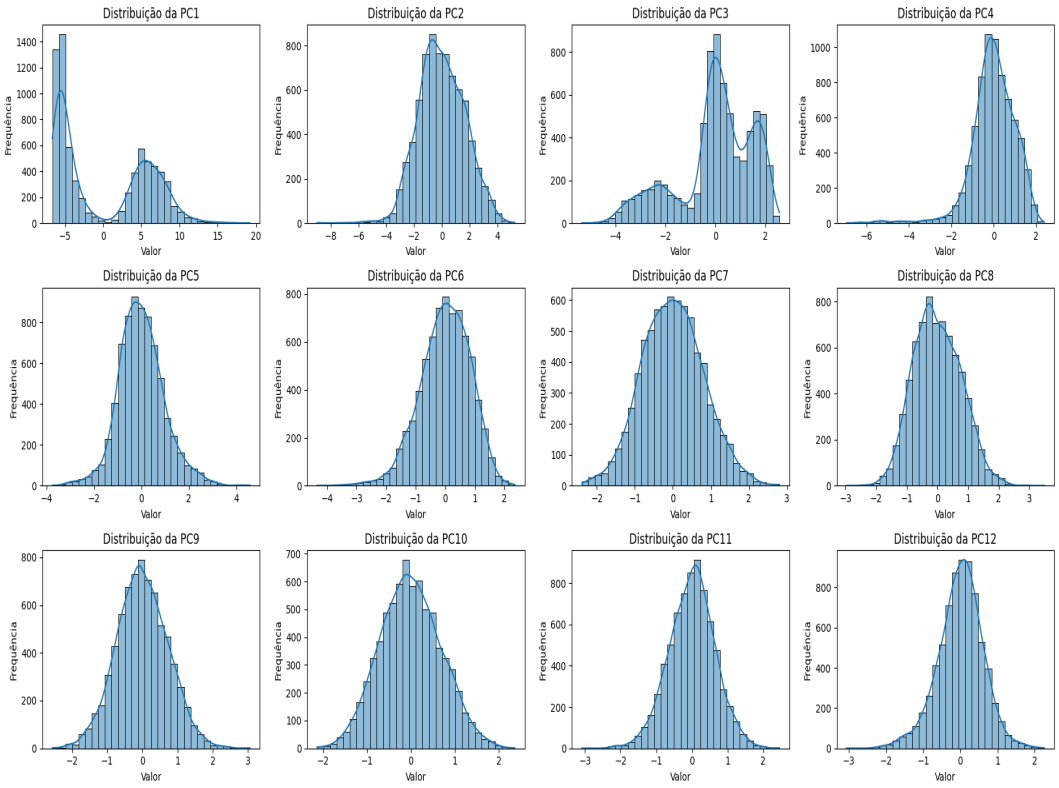
A análise em 3D, também baseada na técnica de PCA, foi usada para examinar os dados em três dimensões, revelando três agrupamentos claros. Essa projeção tridimensional proporciona uma melhor compreensão da separação dos grupos e da distribuição dos dados, mostrando que há uma estrutura que justifica a aplicação do algoritmo K-means para identificação de clusters. A visualização em 3D ajuda a perceber a complexidade dos dados e a interpretar as relações entre as variáveis de forma mais detalhada.

Projeção 3D dos Dados X_train após PCA



4. Redução de Dimensionalidade com PCA e Distribuição dos Componentes

A técnica de PCA foi aplicada para reduzir a dimensionalidade dos dados para 12 componentes principais. Essa redução tornou a visualização e a interpretação dos dados mais gerenciáveis e revelou insights interessantes. Para cada uma das 12 componentes principais, foi gerado um histograma para examinar a distribuição dos dados.



Os histogramas mostraram que, entre os 12 componentes, pelo menos 9 seguem uma distribuição normal, indicando uma simetria nos dados que pode facilitar a modelagem. No entanto, 3 componentes apresentaram distribuições mais irregulares, sugerindo que esses componentes podem ter características diferentes ou serem mais influenciados por ruído nos dados. Essas observações são cruciais para a análise de clusters e podem impactar a interpretação e a eficácia do algoritmo K-means na identificação de agrupamentos.

5. Implementação do Modelo

O algoritmo de K-means foi implementado com diferentes valores de K (número de clusters): K=2, K=3, K=4. Para cada configuração, foram gerados gráficos ilustrando os agrupamentos e calculados os Silhouette Scores para avaliar a qualidade do agrupamento.

6. Escolha do Número de Clusters

Foi aplicado o seguinte método para escolher K:

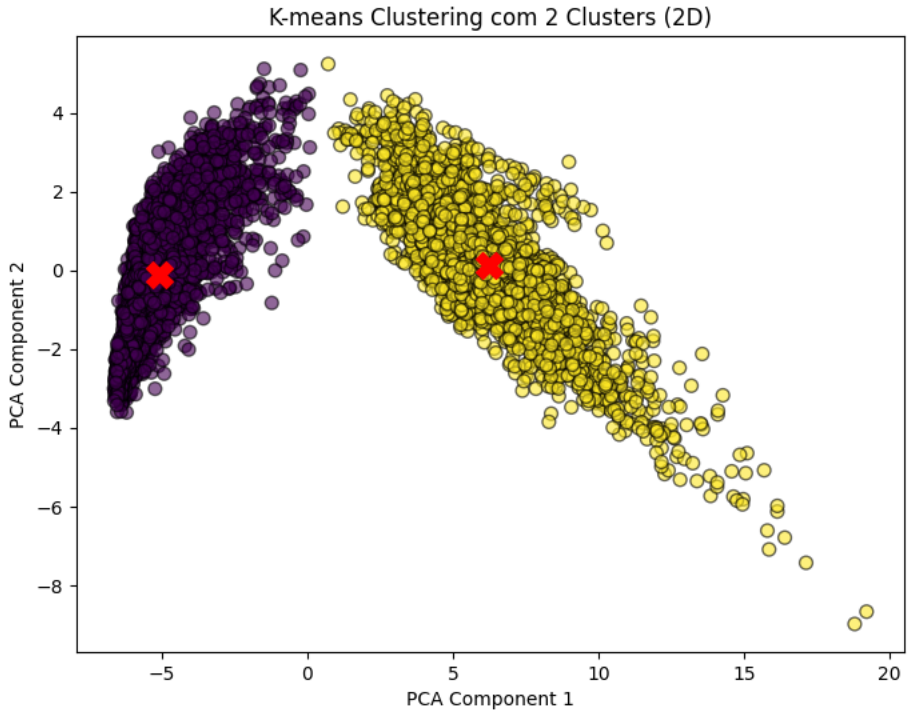
Silhouette Score:

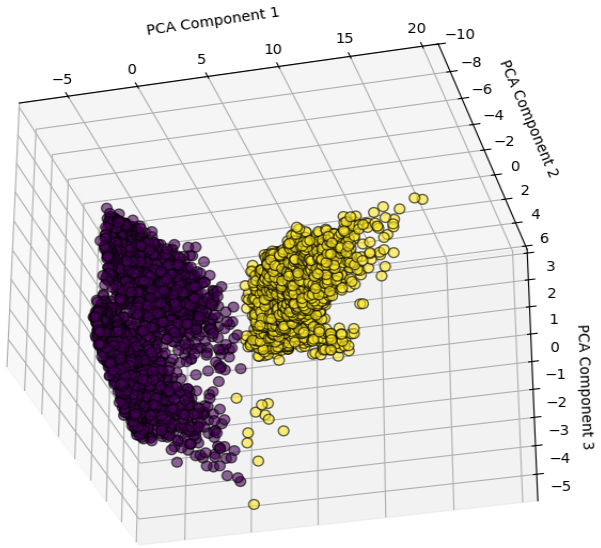
Valores de Silhouette Score foram avaliados para cada K:
- K=2: 0.757 (2D), 0.701 (3D)
- K=3: 0.638 (2D), 0.570 (3D)
- K=4: 0.506 (2D), 0.480 (3D)

Resultados e Discussão

Resultados para K=2

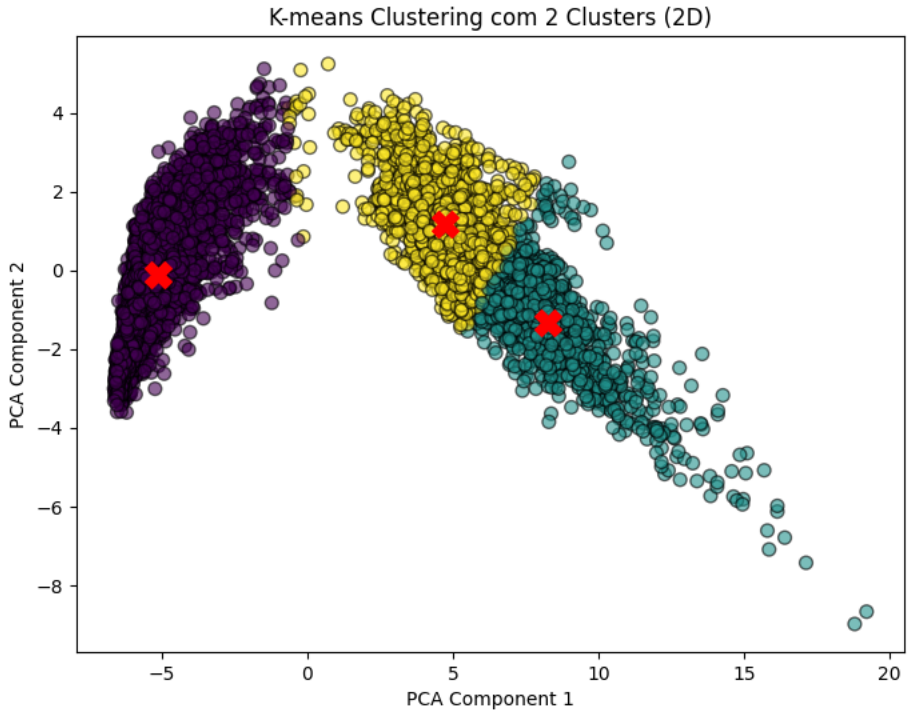
Silhouette Score: 0.757 (2D) e 0.701 (3D) Observação: O agrupamento separa os dados em dois conjuntos principais, correspondendo a duas grandes divisões naturais nos dados. Este modelo reflete bem os padrões gerais, com alta coesão intracluster e boa separação entre clusters. Indicado para quando queremos uma visão simplificada e geral dos dados.

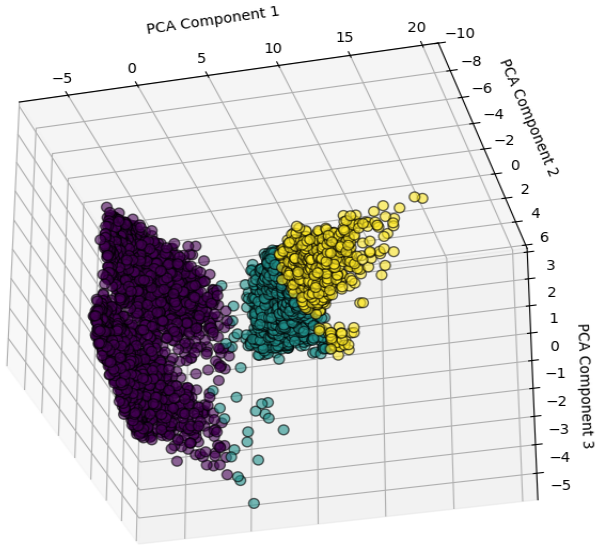




Resultados para K=3

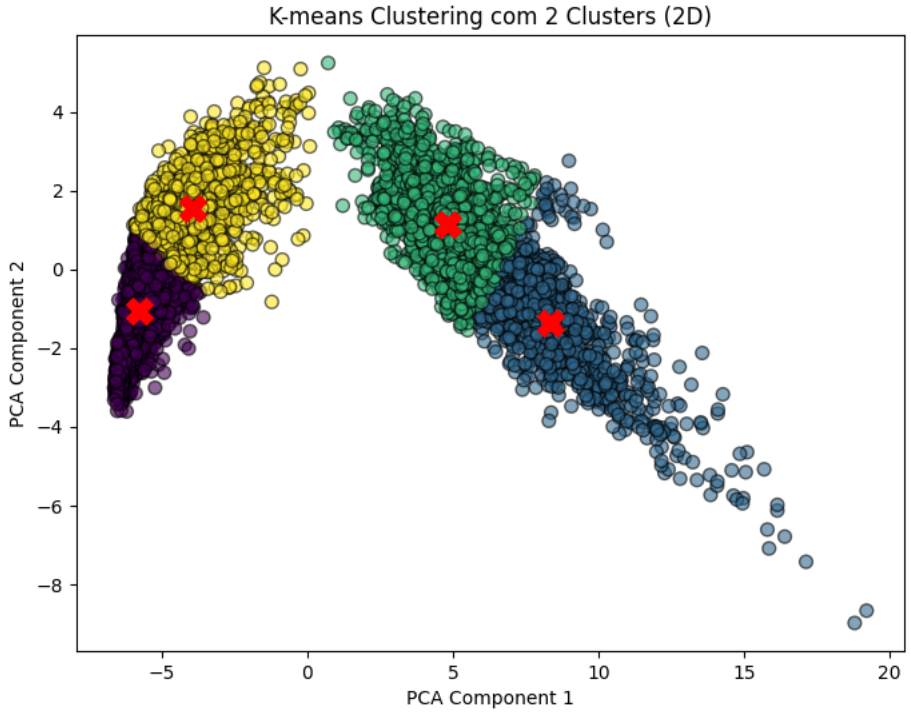
Silhouette Score: 0.638 (2D) e 0.570 (3D) Observação: O agrupamento identifica um cluster muito grande e dois menores. Embora tenha reduzido a qualidade do agrupamento (medida pelo Silhouette Score), essa configuração pode ser útil em situações onde há necessidade de um nível intermediário de granularidade.

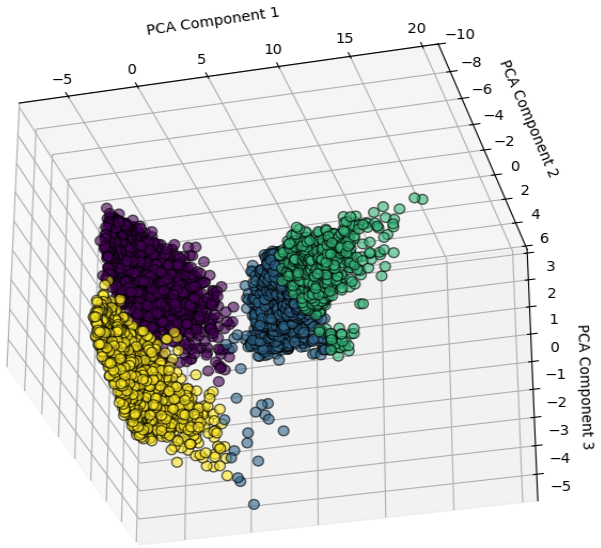




Resultados para K=4

Silhouette Score: 0.506 (2D) e 0.480 (3D) Observação: O modelo divide os dados em quatro clusters de tamanhos semelhantes. Apesar de fornecer maior granularidade, o baixo Silhouette Score indica que os clusters podem estar menos bem definidos, com maior sobreposição entre eles.





Conclusão e Trabalhos Futuros

Este projeto demonstrou a eficácia da aplicação do algoritmo K-means para identificar padrões em dados de atividades humanas coletados com sensores de smartphones. A análise revelou que K=2 é a configuração mais adequada para este conjunto de dados, garantindo uma boa separação intracluster e maior simplicidade interpretativa. As técnicas de redução de dimensionalidade como PCA foram cruciais para a visualização e análise dos dados de alta dimensionalidade. Trabalhos futuros podem incluir a exploração de algoritmos alternativos de clustering, como DBSCAN ou GMM, para lidar melhor com distribuições irregulares e sobreposição de clusters. Além disso, o uso de métodos de validação cruzada pode ajudar a garantir a robustez dos resultados. Uma análise mais detalhada das variáveis com distribuições irregulares também pode fornecer insights adicionais sobre os dados.

Referências

1. Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. International Workshop of Ambient Assisted Living (IWAAL 2012). Springer, 2013.

2. Repositório UCI Machine Learning: Human Activity Recognition Using Smartphones Dataset. Disponível em: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

3. Sklearn Documentation: Clustering Metrics. Disponível em: https://scikit-learn.org/stable/modules/clustering.html