CEPEDI

CIÊNCIA DE DADOS

ALEX REGINALDO ABREU CAMPOS DOS SANTOS

SARA SACRAMENTO DE MELLO

RELATÓRIO TÉCNICO

Implementação e Análise do Algoritmo de K-means com o Dataset Human Activity Recognition

Data de Entrega: 03/12/2024

RESUMO

O objetivo deste projeto é implementar e avaliar o algoritmo de K-means utilizando o dataset "Human Activity Recognition Using Smartphones". Este relatório documenta todas as etapas do projeto, incluindo a análise exploratória dos dados, a implementação do algoritmo de K-means, a escolha do número ideal de clusters e a visualização dos resultados. Os principais resultados mostram a eficácia do K-means na identificação de padrões de atividades humanas a partir dos dados de sensores.

INTRODUÇÃO

O reconhecimento de atividades humanas é uma área de pesquisa importante, com aplicações em saúde, fitness e interfaces inteligentes. Este projeto utiliza dados de sensores de smartphones para agrupar atividades humanas usando o algoritmo de K-means. A escolha do K-means se deve à sua simplicidade e eficácia na formação de clusters em dados de alta dimensionalidade.

METODOLOGIA E RESULTADO

1. Definição e Preparação do Problema

- Acesso ao Conjunto de Dados: O dataset foi acessado através do repositório UCI Machine Learning. Ele contém medições de 561 variáveis calculadas a partir dos sinais brutos dos sensores.
- Análise Exploratória: Foram examinadas as distribuições das variáveis e avaliadas possíveis correlações entre elas. Em seguida, foi realizada a normalização dos dados para garantir que todas as variáveis contribuam de forma equilibrada para o agrupamento.

2. Análise Exploratória dos Dados

A matriz de correlação foi visualizada para identificar relações entre as variáveis:

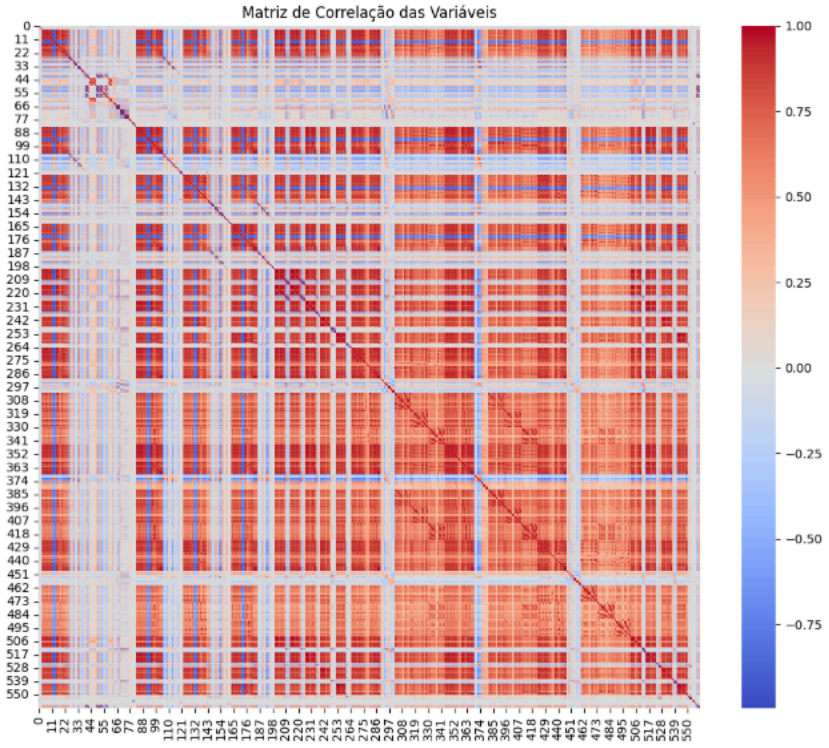

Figura 1 - Matriz de Correlação das Variáveis

Esta matriz de calor (heatmap) ilustra a correlação entre as variáveis do dataset original. O gradiente de cores varia de azul (correlação negativa) a vermelho (correlação positiva).

A matriz de correlação destaca padrões de relação entre as 561 variáveis dos sensores. Variáveis altamente correlacionadas (em tons vermelhos) sugerem

redundância nos dados, enquanto áreas com baixa correlação (tons azulados ou neutros) indicam maior independência. A presença de blocos distintos na matriz indica que grupos de variáveis estão mais fortemente relacionadas, o que pode ser reflexo das características específicas capturadas pelos sensores em diferentes condições de atividade.

3. Redução de Dimensionalidade com PCA

Para facilitar a visualização e interpretação dos clusters, foi utilizada a Análise de Componentes Principais (PCA). O PCA é uma técnica estatística que transforma os dados originais em um novo conjunto de variáveis, as componentes principais, que são combinações lineares das variáveis originais. Essas componentes principais são ordenadas de forma que a primeira componente retém a maior parte da variância total presente nos dados, seguida pela segunda, e assim por diante.

O gráfico abaixo mostra a variância explicada acumulada por cada componente principal:

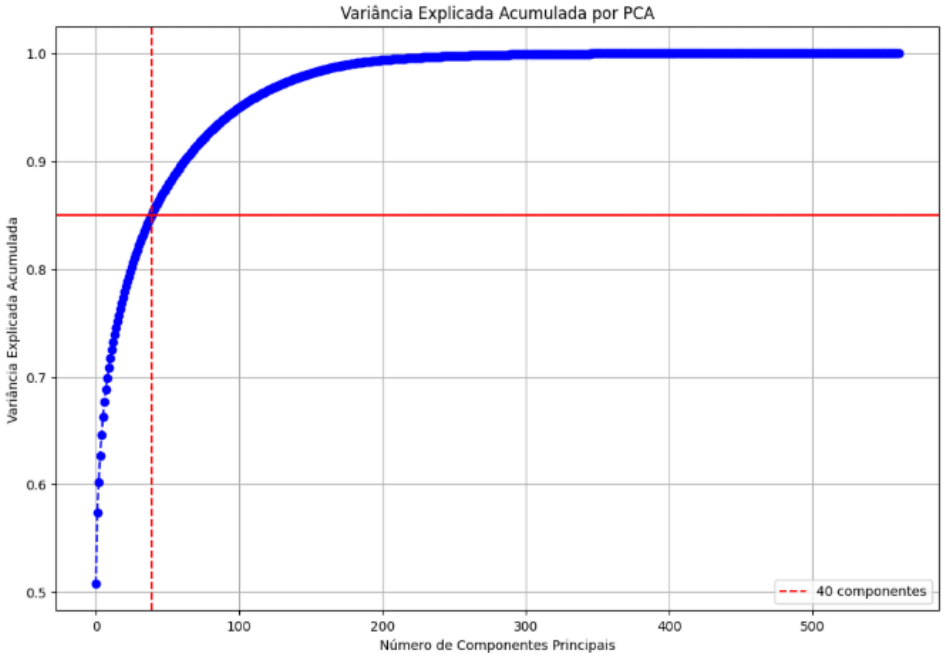

Figura 2 - Variância Explicada Acumulada por PCA

No gráfico, a linha vermelha horizontal indica o limiar de 85% da variância explicada, que é frequentemente utilizado como critério para selecionar o número de componentes principais. A linha vermelha vertical marca o ponto em que a variância explicada acumulada atinge este limiar.

Com base no gráfico, podemos observar que:

- 40 componentes principais são suficientes para explicar aproximadamente 85% da variância total nos dados.
- Este resultado sugere que, ao reduzir a dimensionalidade dos dados para 40 componentes, ainda conseguimos preservar a maior parte da informação original presente no conjunto de dados.

A escolha de 40 componentes principais foi então utilizada para transformar os dados originais, reduzindo a complexidade e facilitando o processo de clustering. A aplicação do PCA não apenas simplificou os dados, mas também ajudou a mitigar problemas relacionados à multicolinearidade, melhorando a performance do algoritmo de K-means.

4. Escolha do Número de Clusters

A escolha do número de clusters é uma etapa crucial na aplicação do K-means. Para determinar o valor mais adequado de K, utilizamos dois métodos principais: o método do cotovelo e o silhouette score.

- Método do Cotovelo:

O método do cotovelo consiste em executar o algoritmo de K-means para diferentes valores de K e plotar a soma das distâncias quadradas dentro dos clusters (inércia) para cada valor de K. O ponto em que a taxa de diminuição da inércia começa a se estabilizar indica o número ideal de clusters. O gráfico abaixo mostra a inércia para valores de K variando de 2 a 14:

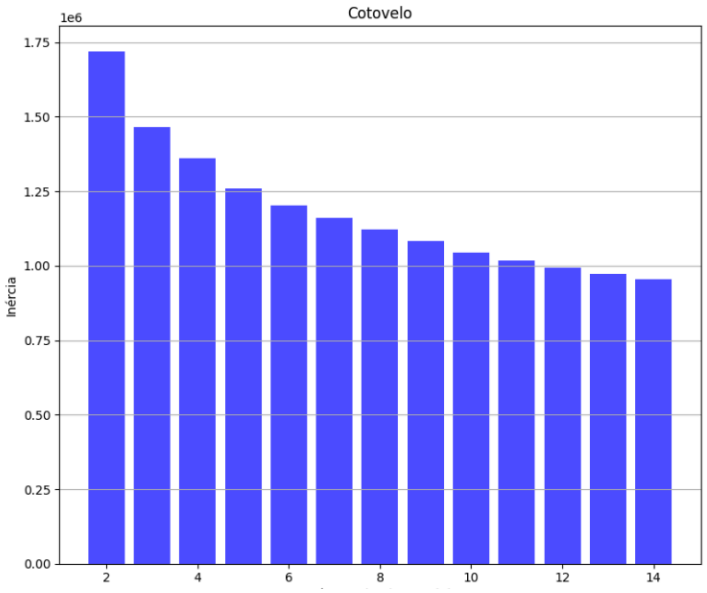

Figura 2 - Cotovelo

No gráfico do método do cotovelo, podemos observar que a inércia diminui rapidamente à medida que K aumenta de 2 para 4, mas a taxa de diminuição se estabiliza para valores maiores de K. Este comportamento sugere que a escolha de K = 2 ou K = 3 poderia ser apropriada, pois após esses pontos, os ganhos adicionais na redução da inércia são pequenos.

- Silhouette Score

O silhouette score mede a qualidade do agrupamento, avaliando o quão similar um ponto é ao seu próprio cluster comparado ao ponto mais próximo de um cluster diferente. O score varia de -1 a 1, onde valores mais altos indicam clusters mais bem definidos. O gráfico abaixo mostra o silhouette score para valores de K variando de 2 a 14:

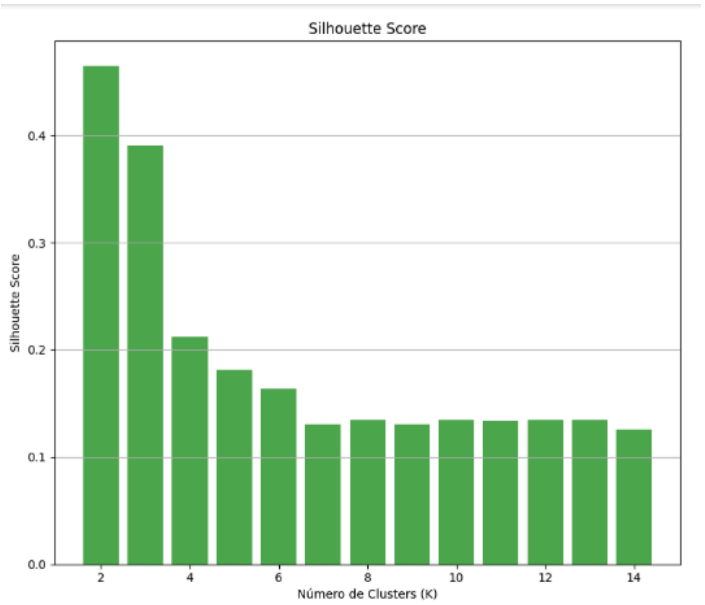

Figura 3 - Silhouette Score

No gráfico de silhouette score, podemos observar que o silhouette score é mais alto para K = 2, indicando que este valor de K proporciona a melhor coesão e separação dos clusters. O score diminui à medida que K aumenta, sugerindo que clusters adicionais não estão bem definidos.

- Escolha Final de K

Com base nos resultados do método do cotovelo e do silhouette score, o método do cotovelo sugere que K = 2 ou K = 3 pode ser apropriado. O silhouette score indica que K = 2 é o valor ideal.

Portanto, foi decidido utilizar K = 2 como o número de clusters para o modelo final, considerando que oferece a melhor definição de clusters segundo o silhouette score.

5. Visualização dos Clusters em 3D usando PCA

Após determinar o número ideal de clusters e aplicar o algoritmo de K-means, utilizamos novamente a Análise de Componentes Principais (PCA) para reduzir os dados a três componentes principais, permitindo a visualização dos clusters em 3D. O gráfico abaixo mostra a distribuição dos dados transformados em três componentes principais, com os pontos coloridos de acordo com o cluster ao qual pertencem:

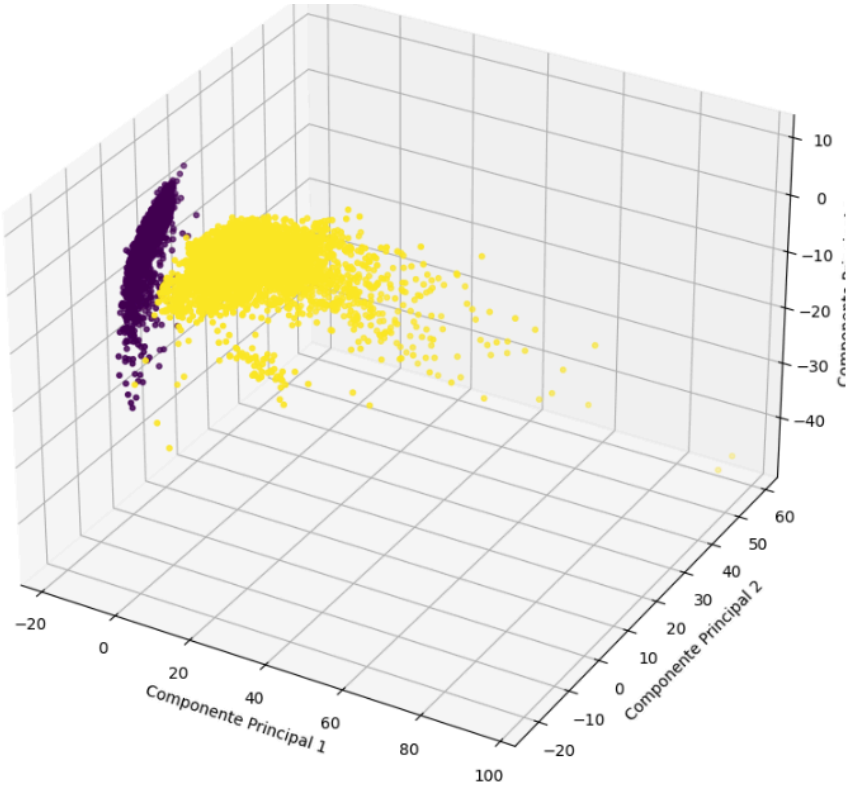

Figura 4 - Visualização dos Clusters em 3D usando PCA

Podemos observar que os pontos amarelos e roxos representam os dois clusters identificados pelo K-means. A maioria dos dados parece estar bem

separada entre os dois clusters, o que sugere que o K-means foi capaz de identificar agrupamentos significativos nas atividades humanas. A separação entre os clusters não é perfeita, o que é esperado devido à complexidade e à sobreposição inerente nos dados de atividades humanas. No entanto, os clusters são razoavelmente distintos, indicando que o modelo conseguiu capturar diferenças importantes nas características das atividades.

6. Análise Final dos Clusters

Para entender melhor as características dos clusters formados, realizamos uma análise adicional verificando a distribuição das atividades dentro de cada cluster. O gráfico abaixo mostra a contagem de cada atividade em cada cluster:

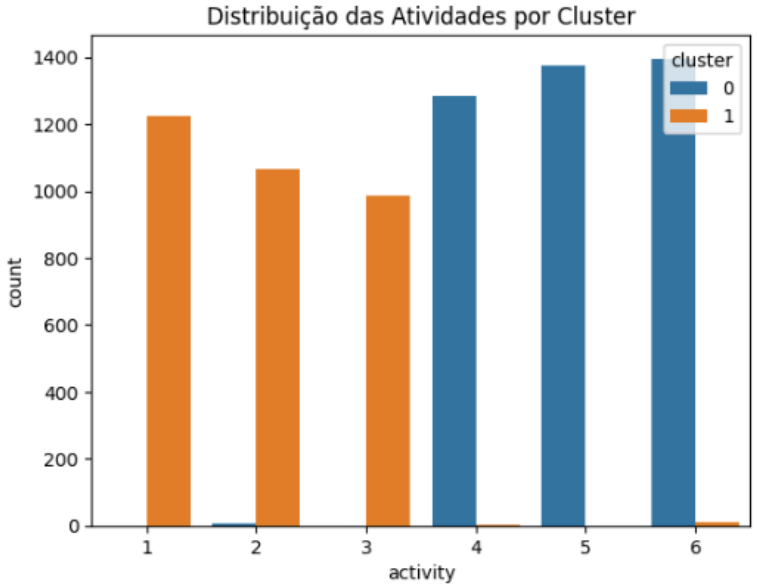

Figura 5 - Distribuição das Atividades por Cluster

O gráfico revela como o K-means agrupou as atividades, indicando o número de observações atribuídas a cada cluster para cada atividade. Observa-se que algumas atividades possuem uma distribuição mais equilibrada entre os clusters, enquanto outras são dominadas por um único cluster.

DISCUSSÃO

Os resultados do projeto demonstram que o K-means, aliado ao PCA, foi uma abordagem eficiente para reduzir a dimensionalidade e identificar padrões relevantes nos dados. Os clusters formados refletem grupos distintos, embora haja alguma sobreposição devido à complexidade do conjunto de dados e à semelhança entre certas atividades. O uso do PCA ajudou a otimizar o desempenho computacional e a facilitar a visualização, mas pode ter resultado em perda de detalhes importantes para a separação entre atividades. Além disso, a escolha de K com base no silhouette score se mostrou sólida, mas as interpretações práticas ainda são limitadas pela falta de uma correspondência direta entre clusters e atividades específicas.

CONCLUSÃO E TRABALHOS FUTUROS

O projeto destacou o potencial do K-means para agrupamento de dados complexos, demonstrando a eficácia de técnicas como normalização e PCA para lidar com alta dimensionalidade. A combinação dessas técnicas permitiu identificar padrões relevantes, embora o modelo não tenha capturado perfeitamente todas as nuances das atividades humanas.

Trabalhos futuros incluem avaliar algoritmos alternativos que podem lidar melhor com sobreposições de dados; Explorar abordagens supervisionadas para validar e refinar os clusters; Analisar dados temporais para capturar melhor a transição entre atividades.

REFERÊNCIAS

ANGUITA, D., GHIO, A., ONETO, L., PARRA, X., & ORTIZ, J. L. (2013). A Public Domain Dataset for Human Activity Recognition Using Smartphones. Bruges (Belgium), 24-26 April 2013. Disponível em: https://www.esann.org/sites/default/files/proceedings/legacy/es2013-84.pdf. Acesso em: 28 nov. 2024.

ANÁLISE DE COMPONENTES PRINCIPAIS In: WIKIPEDIA, a enciclopédia livre. Disponível em: https://pt.wikipedia.org/wiki/An%C3%A1lise_de_componentes_principais. Acesso em 28 nov. 2024.

K-MEANS In: WIKIPEDIA, a enciclopédia livre. Disponível em: https://pt.wikipedia.org/wiki/K-means. Acesso em 28 nov. 2024.