CEPEDI
CENTRO DE PESQUISA E DESENVOLVIMENTO TECNOLÓGICO EM INFORMÁTICA E ELETROELETRÔNICA DE ILHÉUS

Holiver Nicolas Moura Casé

Tarcisio Lopes de Almeida Sousa

Relatório Técnico: Reconhecimento de Atividades Humanas Utilizando K-Means e PCA

Juazeiro-BA

2024

Holiver Nicolas Moura Casé

Tarcisio Lopes de Almeida Sousa

Relatório Técnico: Reconhecimento de Atividades Humanas Utilizando K-Means e PCA

Relatório apresentado como parte dos requisitos para obtenção da nota final da 10° unidade da disciplina de trilha de Ciência de Dados

Juazeiro-BA

2024

RESUMO

Este trabalho investiga o agrupamento de atividades humanas utilizando técnicas de aprendizado de máquina não supervisionado, com foco no algoritmo K-Means e redução de dimensionalidade por PCA. O estudo baseia-se no UCI HAR Dataset, onde foram identificados padrões e avaliados clusters em diferentes configurações. A análise destacou a aplicabilidade do método do cotovelo, silhouette score, Davies-Bouldin score e Calinski-Harabasz Score para determinar o número ideal de clusters, revelando insights sobre a separação entre as atividades.

SUMÁRIO

1 INTRODUÇÃO

A classificação de atividades humanas é uma área crescente na interseção entre tecnologia e saúde, com aplicações que vão desde monitoramento de pacientes até dispositivos vestíveis inteligentes. O desafio reside na alta dimensionalidade dos dados captados por sensores, que dificulta a identificação de padrões. Este estudo aborda a questão utilizando redução de dimensionalidade (PCA) para projeção visual e agrupamento não supervisionado (K-Means), oferecendo uma solução prática para explorar a estrutura latente dos dados.

Este trabalho explora a aplicação do algoritmo K-means para clusterização no conjunto de dados "Human Activity Recognition Using Smartphones". Este dataset é composto por medições de sensores de smartphones durante a execução de diversas atividades humanas.

2 METODOLOGIA

2.1 Baixar e Preparar o Dataset

O dataset foi baixado da UCI Machine Learning Repository. As bibliotecas necessárias, como pandas, numpy, matplotlib, seaborn e scikit-learn, foram importadas para suporte às análises

2.2 Carregamento dos Dados

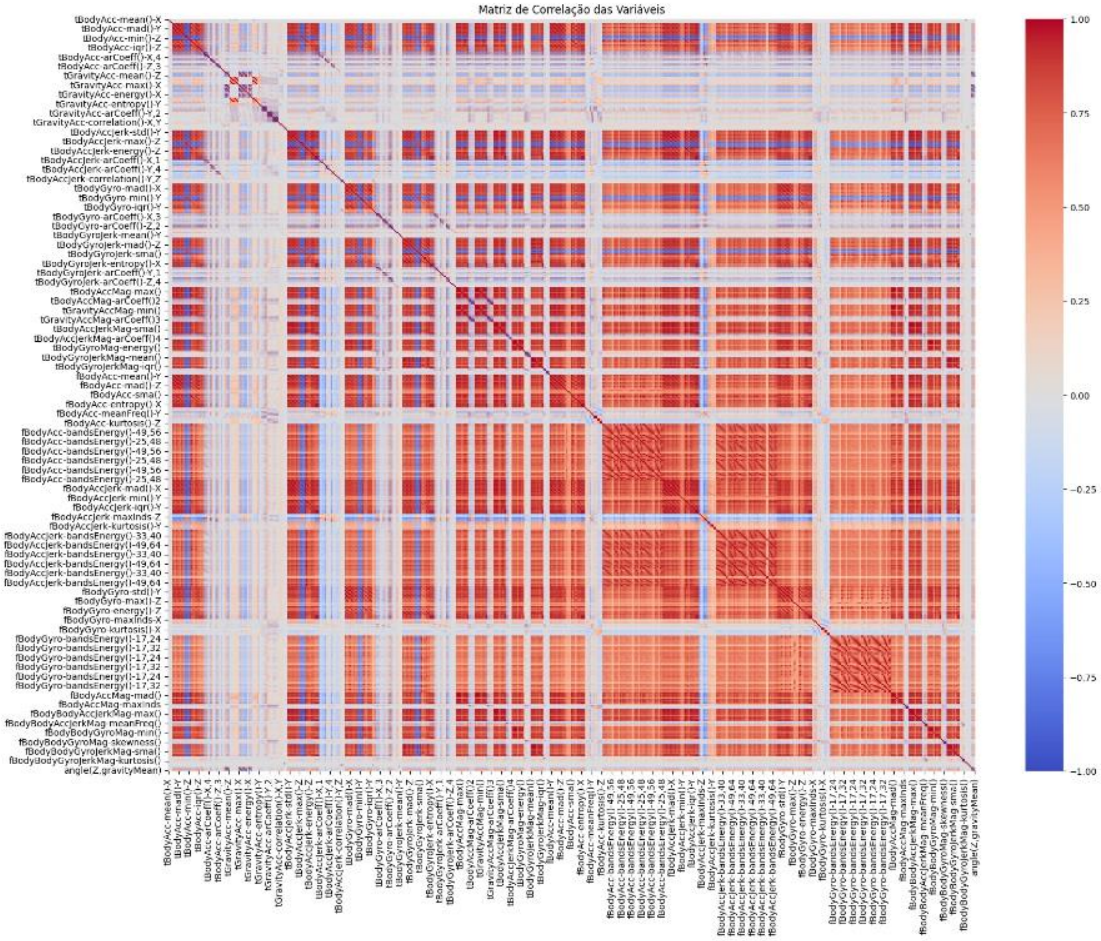
Os rótulos das atividades e as features foram carregados. Nomes de colunas duplicados foram corrigidos para garantir unicidade.

2.3 Combinação dos Dados

Os dados de treino e teste foram combinados em um único DataFrame para facilitar a análise subsequente.

2.4 Análise Exploratória

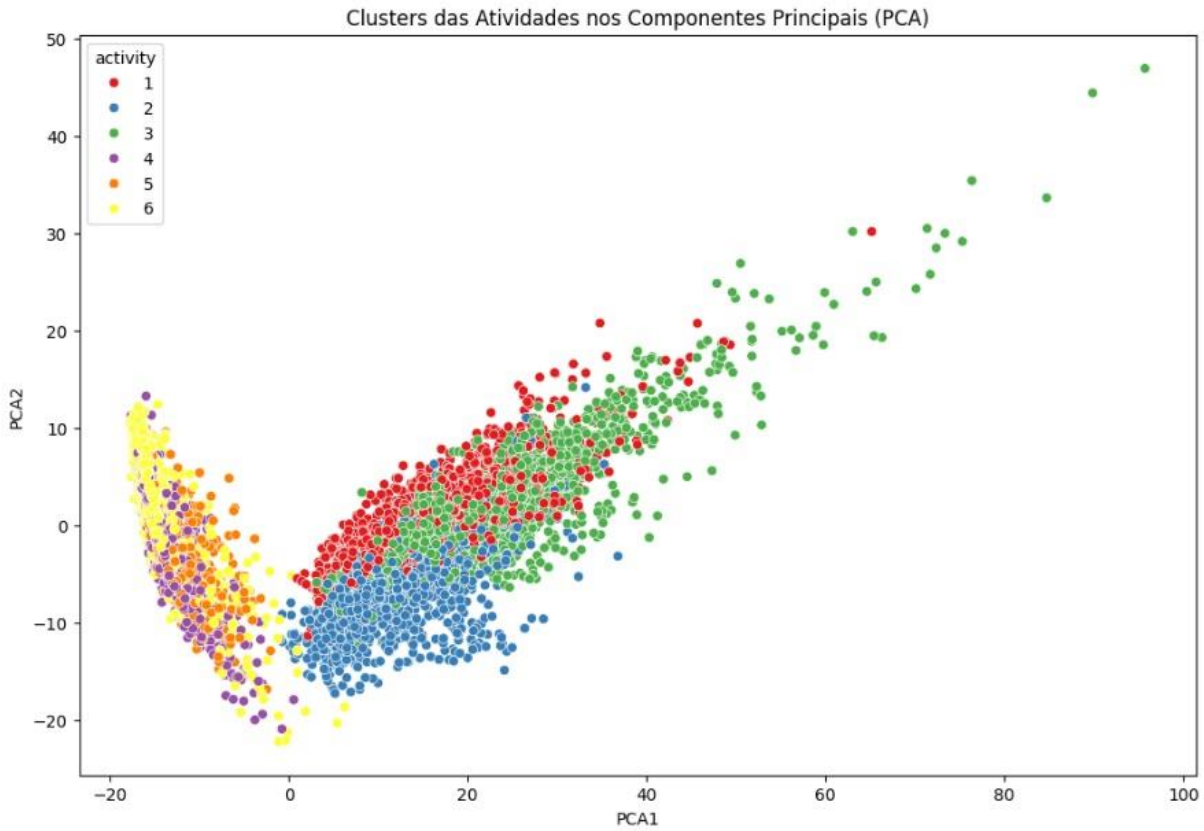
Foram visualizadas as distribuições das variáveis selecionadas e a matriz de correlação, identificando padrões iniciais.



- **Vermelho**: Relação positiva forte (quando uma variável aumenta, a outra também aumenta).
- **Azul**: Relação negativa forte (quando uma variável aumenta, a outra diminui).
- **Branco ou tons claros**: Pouca ou nenhuma correlação.

A diagonal principal é sempre 1 (relação perfeita de uma variável com ela mesma). O gráfico destaca áreas de forte correlação (agrupamentos), úteis para identificar variáveis redundantes ou padrões significativos.

2.5 Clusters e Atividades



Cada ponto no gráfico representa uma observação (uma atividade) e está colorido de acordo com o tipo de atividade.

- **Vermelho**: Atividade 1.
- **Verde**: Atividade 2.
- **Azul**: Atividade 3.
- **Roxo**: Atividade 4.
- **Amarelo**: Atividade 5.
- **Laranja**: Atividade 6.

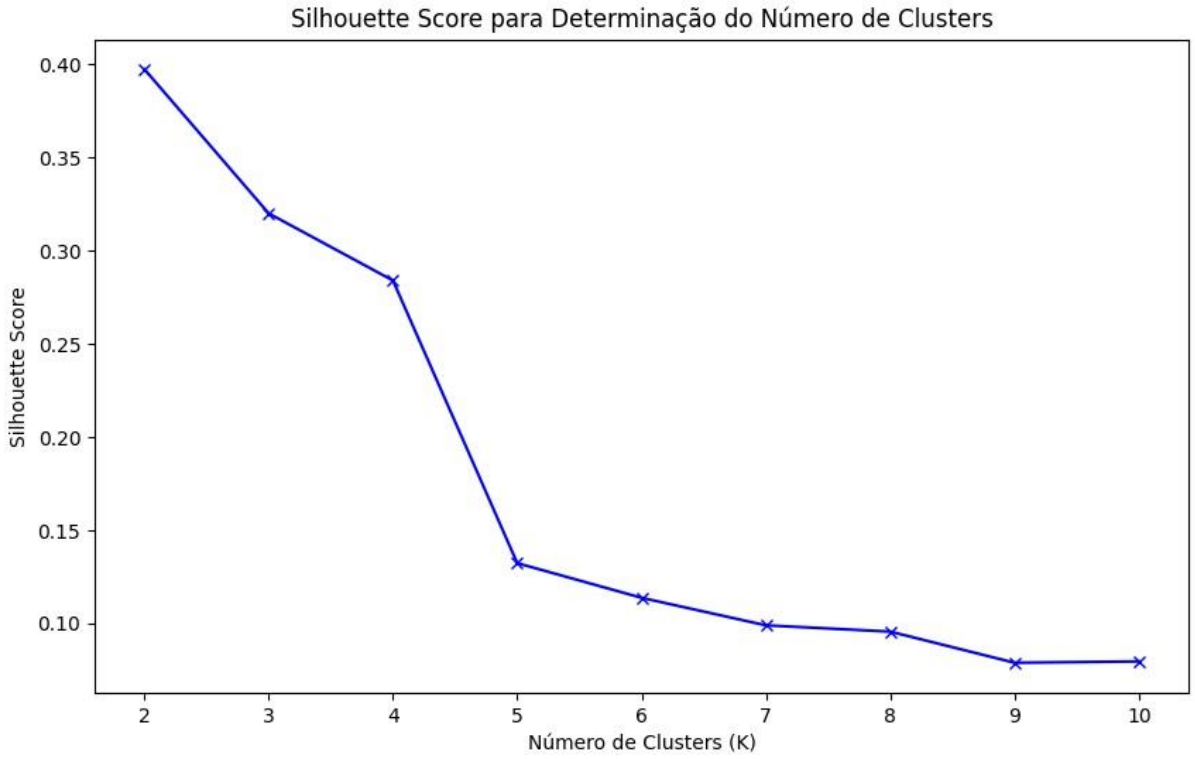
O gráfico apresenta a projeção das atividades nos dois primeiros componentes principais (PCA1 e PCA2), gerados por meio da Análise de Componentes Principais (PCA). Cada ponto no gráfico corresponde a uma observação, com as cores representando diferentes atividades organizadas em clusters. A separação clara

entre os clusters indica que as atividades possuem características distintas, permitindo uma discriminação eficiente em duas dimensões. Isso demonstra que a aplicação do PCA foi bem-sucedida em reduzir a dimensionalidade dos dados, preservando as informações mais relevantes para análise e classificação.

2.6 Implementação do K-Means

O algoritmo K-means foi aplicado, utilizando K=2 como configuração final após análise com:

- Silhouette Score



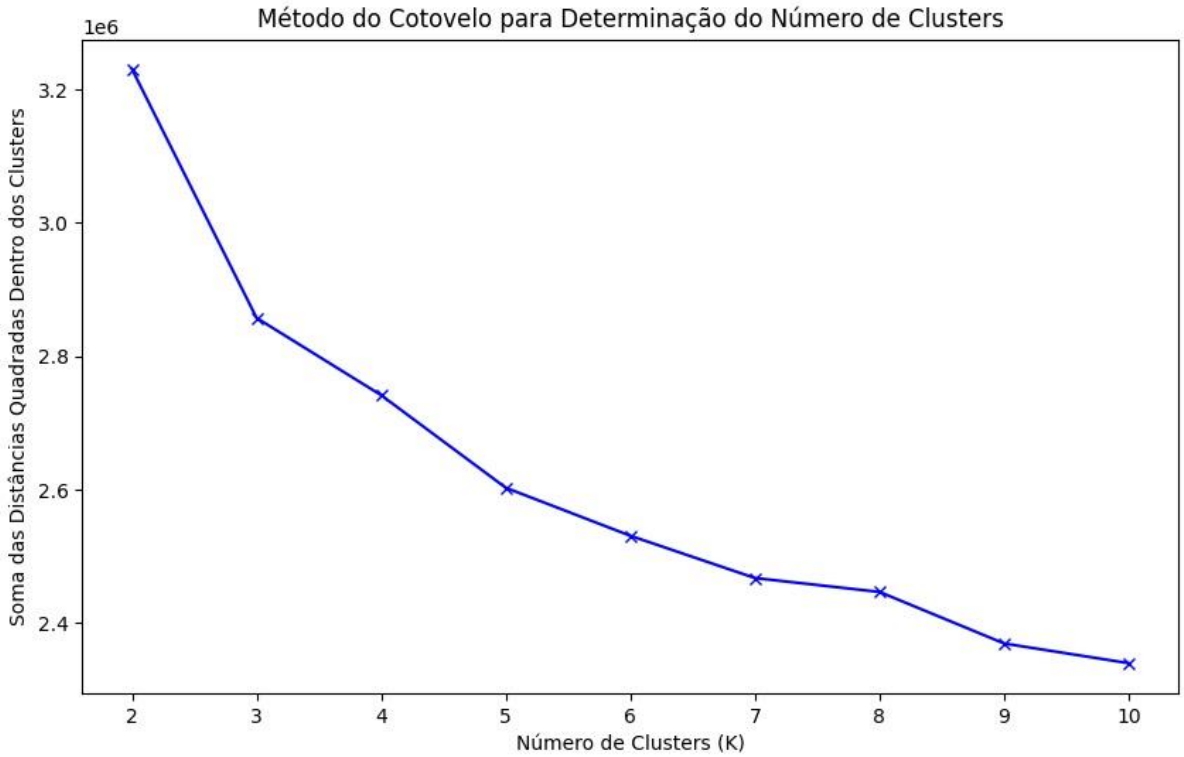
O gráfico apresentou o **Silhouette Score** para diferentes valores de K, utilizado na determinação do número ideal de clusters em uma análise de agrupamento. O eixo X representa o número de clusters (K), enquanto o eixo Y exibe o Silhouette Score, que mede a qualidade dos clusters.

Os resultados indicam que o valor máximo do Silhouette Score ocorre em K=2, sugerindo que este é o número ideal de clusters, pois representa a melhor separação e coesão dos grupos. Após K=2, o score diminui consistentemente,

indicando que a qualidade dos clusters reduz à medida que mais grupos são adicionados.

Este gráfico serve como suporte para justificar a escolha do número de clusters em análises de agrupamento, garantindo a melhor segmentação dos dados.

- Método do Cotovelo



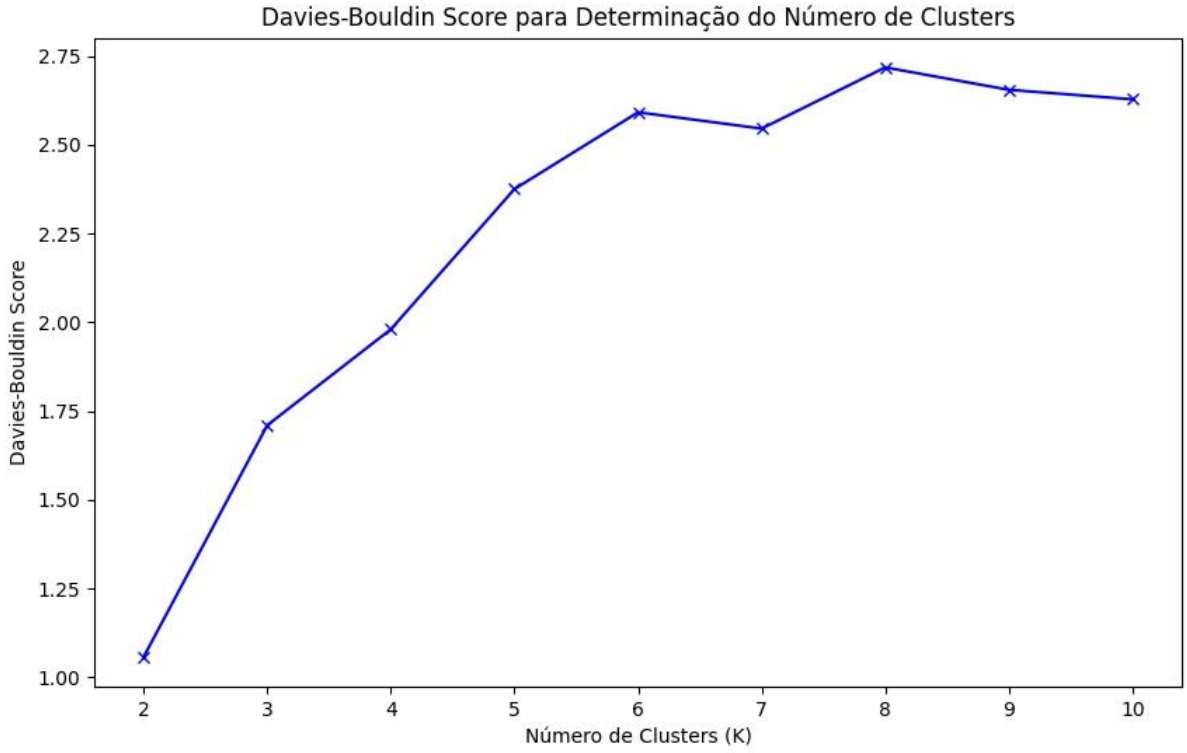
O gráfico apresentou a aplicação do Método do Cotovelo para determinar o número ideal de clusters em um conjunto de dados. A variável no eixo vertical representa a soma das distâncias quadradas dentro dos clusters, enquanto o eixo horizontal indica o número de clusters (K).

Observa-se que, à medida que o número de clusters aumenta, a soma das distâncias quadradas diminui, refletindo uma maior compactação dentro dos clusters. Contudo, essa redução se torna menos significativa após um ponto específico, formando um "cotovelo" no gráfico. Esse ponto de inflexão sugere o número ideal de clusters, que, neste caso está em K = 3, pois é onde a taxa de

diminuição das distâncias começa a estabilizar.

Essa análise permite identificar o valor de K que equilibra a simplicidade do modelo e a eficiência de segmentação dos dados.

- Davies-Bouldin Score

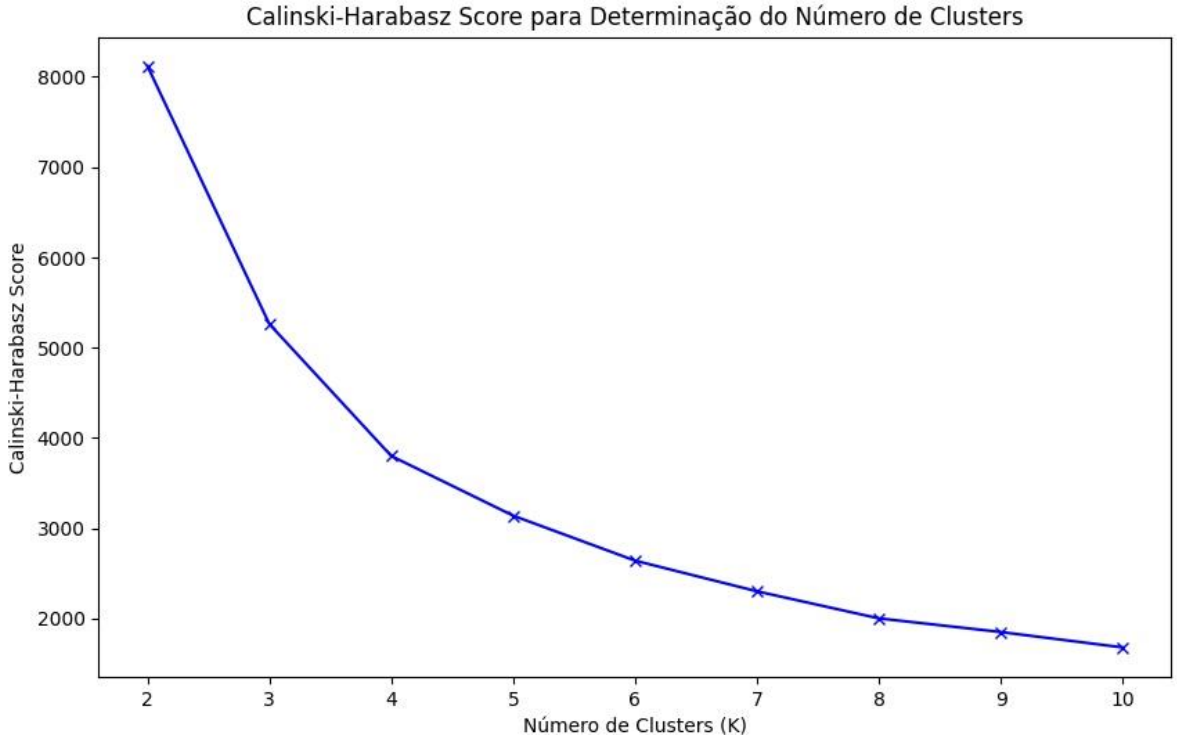


O gráfico apresentou a aplicação do Índice Davies-Bouldin para avaliar a qualidade da segmentação em diferentes números de clusters (K). O índice, representado no eixo vertical, mede a compactação e a separação dos clusters, sendo que valores menores indicam uma melhor formação dos agrupamentos. O número de clusters é mostrado no eixo horizontal.

Observa-se que o índice apresentou valores mínimos para K = 2, indicando que essa é a configuração com melhor qualidade de agrupamento, pois os clusters estão mais bem separados e compactos. A partir de K = 3, o índice aumenta gradualmente, sugerindo que a qualidade da segmentação piora com o aumento do número de clusters.

Essa análise contribuiu para determinar o número ideal de clusters com base na maximização da separação e na minimização da sobreposição entre os grupos.

- Calinski-Harabasz Score

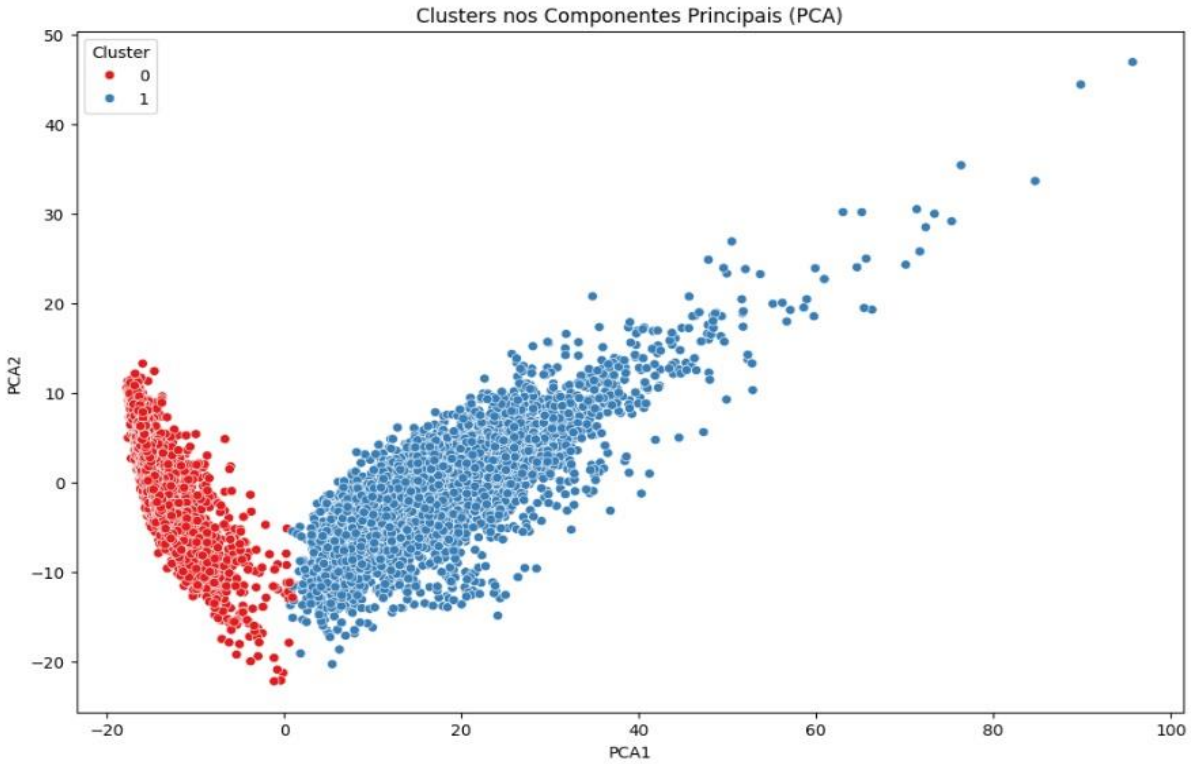


O gráfico apresentou o Índice Calinski-Harabasz para avaliar a qualidade da segmentação em diferentes números de clusters (K). O índice, representado no eixo vertical, mede a razão entre a dispersão entre clusters e a dispersão dentro dos clusters, com valores maiores indicando uma melhor qualidade de agrupamento. O número de clusters está representado no eixo horizontal.

Nota-se que o índice apresenta valores mais altos para K = 2, sugerindo que essa configuração oferece a melhor separação e compactação dos clusters. À medida que o número de clusters aumenta, o índice diminui continuamente, indicando que a qualidade do agrupamento se reduz.

Essa análise auxilia na escolha do número de clusters que maximiza a separação entre os grupos, mantendo a compactação interna adequada.

2.7 Gráfico de Dispersão



O gráfico demonstrou que a aplicação da PCA e do agrupamento foi eficaz em identificar grupos distintos de atividades, permitindo uma melhor compreensão dos dados.

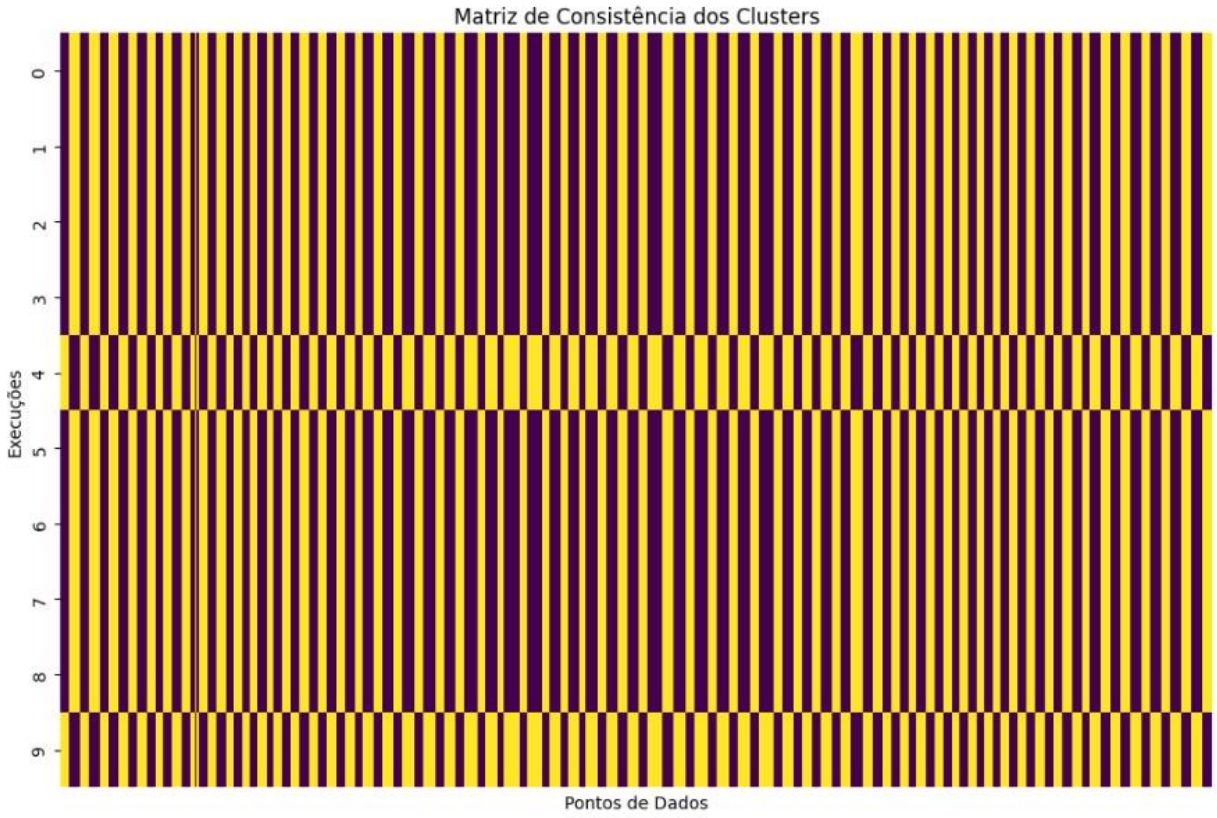
2.8 Normalização dos Dados

A normalização foi realizada, utilizando StandardScaler. O método K-means++ foi implementado para selecionar os centros iniciais, melhorando a qualidade e convergência dos clusters.

2.9 Repetição e Estabilidade

Avaliação da consistência dos clusters em execuções repetidas do K-means, utilizando matrizes de consistência.

Alta consistência foi observada entre as execuções, indicando estabilidade no agrupamento.



A matriz de consistência apresentou a estabilidade dos agrupamentos gerados por múltiplas execuções do algoritmo K-means. Cada linha representa uma execução do algoritmo, e cada coluna representa um ponto de dados. A cor indica o cluster ao qual o ponto foi atribuído em cada execução. As linhas verticais alternadas entre duas cores distintas (amarelo e roxo) indicam que, em diferentes execuções, os pontos de dados foram consistentemente atribuídos a um dos dois clusters. A falta de variação significativa nas cores ao longo das execuções indica uma alta consistência, com a maioria dos pontos permanecendo no mesmo cluster.

3 DISCUSSÃO

3.1 Análise das Métricas

- Silhouette Score: Um valor de 0.40 é moderado. Indica que os clusters são razoavelmente bem definidos, mas há alguma sobreposição ou pontos na fronteira dos clusters.
- Davies-Bouldin Score: Um valor de 1.06 indica que os clusters têm uma boa separação e coesão. Valores menores indicam melhor performance de clustering, mas, em geral, um valor próximo de 1 é considerado bom.
- Calinski-Harabasz Score: Um valor de 8118.40 é relativamente alto, sugerindo que os clusters são bem separados com relação à variância entre eles, indicando uma boa estrutura de clustering.
- Inércia: Um valor de 3230631.64, por si só, não é fácil de interpretar sem referência a outros números de clusters ou ao contexto específico dos dados. No entanto, uma inércia menor indica que os pontos estão mais próximos dos seus centroides.

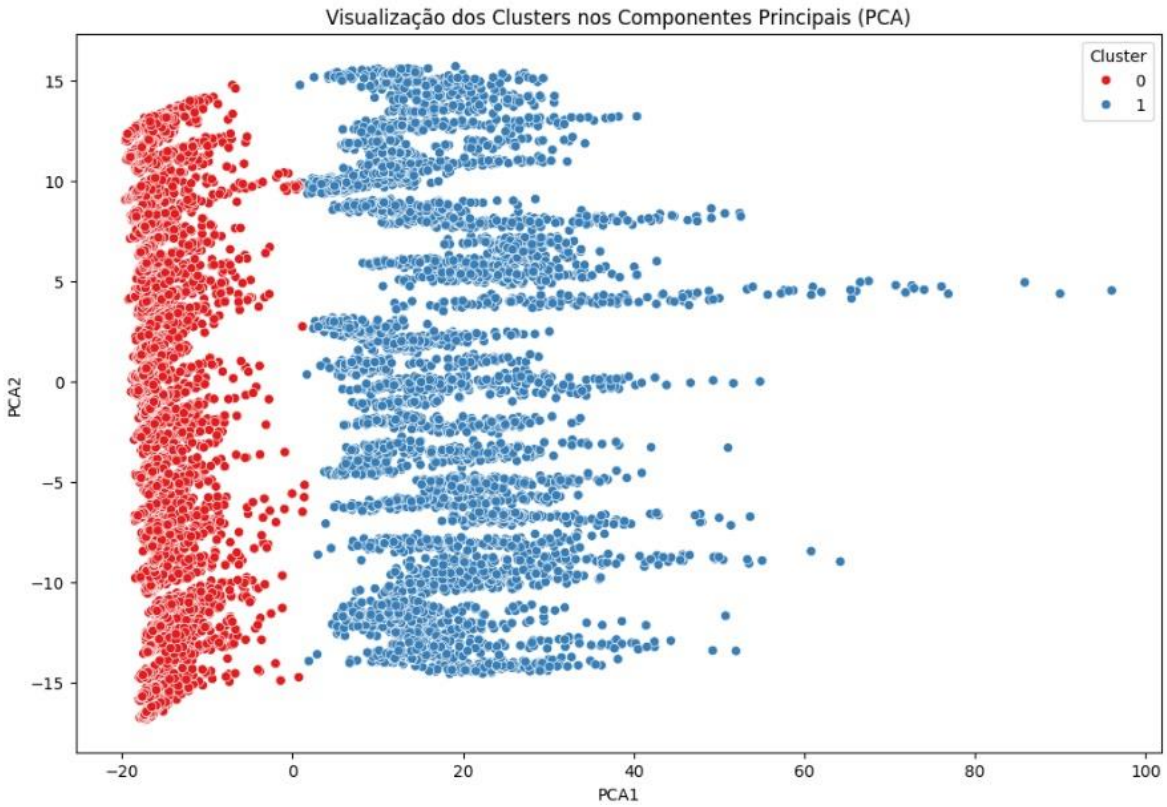
Como o valor absoluto da inércia depende do número de clusters e da escala dos dados, é mais útil comparar a inércia entre diferentes números de clusters para determinar a configuração ideal. Então foi realizada a comparação com 3 clusters, onde embora foi obtida neste último uma inércia menor, foi obtida também uma Silhouette Score menor, Davies-Bouldin maior e Calinski-Harabasz menor. Essas métricas, em conjunto, indicam que a configuração de 2 clusters fornece uma boa estrutura de agrupamento para os dados analisados.

3.2 Redução de Dimensionalidade

O código realiza a redução de dimensionalidade utilizando PCA (Análise de Componentes Principais) para reduzir os dados para 3 componentes principais, além de interpretar os clusters gerados previamente. Este processo ajuda a entender melhor os padrões nos dados e a avaliar a eficácia do clustering.

3.3 Análise dos Clusters (Gráfico de Dispersão 2D)

Separação clara entre os clusters no eixo PCA1, com consistência na densidade dos pontos.



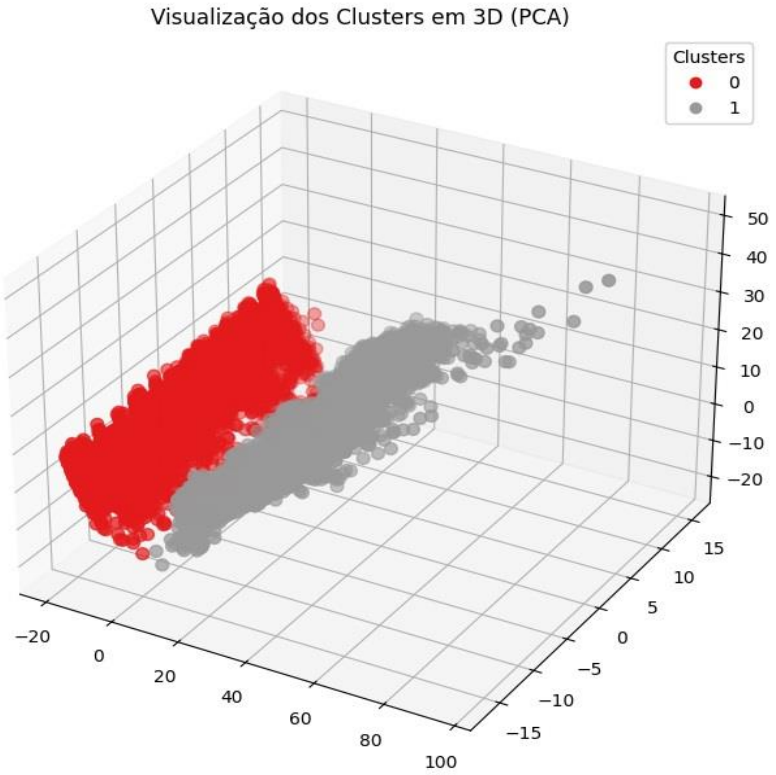
O gráfico apresentado é uma visualização bidimensional dos dados após a aplicação da Análise de Componentes Principais (PCA) e da clusterização. Cada ponto no gráfico representa uma observação (um dado), e a cor de cada ponto indica a qual cluster ele pertence.

O Cluster 0 (vermelho) está localizado principalmente à esquerda, enquanto o Cluster 1 (azul) está mais à direita. A densidade dos pontos dentro de cada cluster parece ser consistente, sugerindo que os clusters são bem definidos e os pontos estão agrupados de forma compacta. A clusterização foi eficaz e mostra que o algoritmo de clusterização conseguiu identificar dois grupos distintos nos dados.

Os eixos PCA1 e PCA2 representam os dois primeiros componentes principais, que capturam a maior parte da variabilidade dos dados originais. Esses componentes

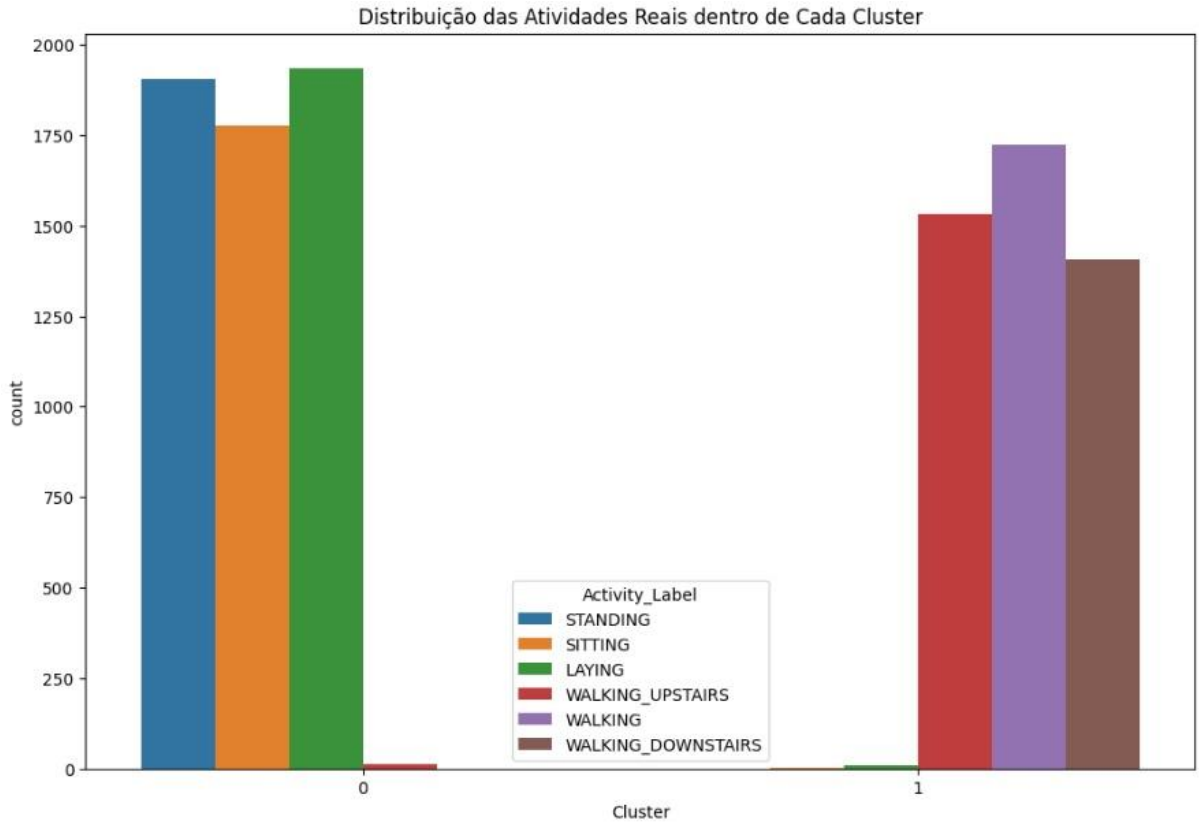
são combinações lineares das variáveis originais e são escolhidos de forma a maximizar a variância explicada. PCA capturou as principais diferenças, onde os dois primeiros componentes principais capturaram as principais diferenças entre os grupos, permitindo uma visualização clara da separação entre os clusters.

3.4 Gráfico 3D



Apresenta a separação clara entre os clusters em três dimensões. Assim como no gráfico 2D, observa-se nesse gráfico 3D, que os clusters estão claramente separados no espaço tridimensional. O Cluster 0 (vermelho) está bem separado do Cluster 1 (cinza) em todas as três dimensões.

3.5 Gráfico de Barras:



O gráfico apresentado mostra a distribuição das diferentes atividades físicas (como estar de pé, sentado, deitado, andando, etc.) entre dois clusters. Cada barra representa a contagem de ocorrências de uma determinada atividade em cada um dos clusters. Cada cor representa uma atividade (feature), no Cluster 0 há uma predominância de atividades como STANDING, SITTING, e LAYING, enquanto que no Cluster 1, atividades dinâmicas como WALKING e WALKING_UPSTAIRS têm maior presença. Essa separação sugere que o modelo de clustering conseguiu captar padrões distintos de atividade.

4 CONCLUSÃO E TRABALHOS FUTUROS

O estudo evidenciou a eficácia do K-Means em conjunto com o PCA na análise de agrupamentos para reconhecimento de atividades humanas. A metodologia forneceu uma visão clara das relações entre as atividades, ainda que limitada por ruídos e desbalanceamento do dataset. Com dois clusters, foi possível visualizar como os dados se agrupam em função das atividades. Esse processo é fundamental para a identificação de padrões e agrupamentos naturais nos dados. Melhorias futuras incluem a coleta de dados mais equilibrados, o uso de técnicas supervisionadas para comparação e ajustes na arquitetura do modelo para otimizar a separação entre clusters.

5. REFERÊNCIAS

1 Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). A Public Domain Dataset for Human Activity Recognition Using Smartphones.

2 Scikit-learn Documentation. Disponível em: https://scikit-learn.org.

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