Centro de Pesquisa e Desenvolvimento Tecnológico em Informática e Eletrônica - CEPEDI

Mateus Lisboa

Solana Bonfim Lemos

Relatório Técnico: Implementação e Análise do Algoritmo K-Means

Ilhéus - BA

2024

Resumo

O projeto de reconhecimento de atividades humanas com K-means busca identificar padrões em dados de acelerômetro e giroscópio coletados por smartphones. O processo começa com a limpeza e normalização dos dados, seguida pela redução de dimensionalidade usando métodos como PCA, análise de correlação e Random Forest, reduzindo as 561 variáveis do dataset para 50.

A escolha do número de clusters foi feita com base em uma análise do método do cotovelo e do silhouette score, concluindo que 3 clusters são os mais apropriados. A análise qualitativa das atividades confirmaram que esses 3 clusters correspondem a atividades distintas: em movimento, parado em pé e parado deitado.

O algoritmo K-means foi executado 5 vezes, apresentando resultados consistentes, e as visualizações gráficas dos clusters mostraram que o modelo conseguiu agrupar as atividades de forma eficaz. O código pode ser executado no Google Colab, e os resultados fornecem uma base sólida para o entendimento de como aplicar técnicas de clusterização em dados de atividades humanas.

Sumário

1. Introdução

O reconhecimento de atividades humanas por meio de sensores está se tornando uma ferramenta essencial em diversas áreas, como saúde, segurança e monitoramento de desempenho físico. Com o advento de tecnologias como smartphones e dispositivos vestíveis, é possível coletar grandes volumes de dados de atividades humanas, como movimento, postura e interação com o ambiente. O objetivo deste projeto é aplicar técnicas de aprendizado de máquina para realizar a classificação dessas atividades, utilizando o algoritmo K-means de agrupamento.

O uso do K-means foi escolhido para este estudo devido à sua simplicidade, eficiência e aplicabilidade em cenários de dados sem rótulos (não supervisionados), como os dados de sensores de atividades. O K-means é capaz de agrupar dados de forma que objetos semelhantes sejam colocados no mesmo grupo, o que se alinha com o objetivo de classificar atividades humanas de forma automática.

2. Metodologia

2.1 Análise Exploratória

O primeiro passo para a análise foi explorar o dataset de atividades humanas, que contém medidas provenientes de sensores. Para garantir que as atividades estivessem balanceadas, foi gerado um histograma do número de atividades, o que revelou que a distribuição estava bem equilibrada entre as classes, facilitando a análise e a modelagem.

Devido à alta dimensionalidade do dataset, com 561 variáveis, foi necessário aplicar técnicas de redução de dimensionalidade que não envolvessem análise gráfica, por exemplo, uma matriz de correlação teria 561 x 561 para ser plotada, gerando uma imagem grande e confusa, um biplot de PCA teria 561 vetores sobrepostos, o que também não ajudaria na análise. Então decidimos usar tresholds para filtrar automaticamente as variáveis correlacionadas, e analisar os componentes de maior importância do PCA pelo peso dos loadings, sem necessitar de plotar os gráficos de ambos . As seguintes técnicas foram utilizadas:

- Análise de colunas duplicadas: haviam algumas colunas duplicadas no dataset, coseguimos remover 84 colunas analisando essa redundância.

- Correlação: Foi calculada uma matriz de correlação para identificar features altamente correlacionadas, as quais poderiam ser removidas para reduzir a redundância dos dados. Definimos o limite para correlação de 0.95 para excluir as variáveis muito correlacionadas, passamos de 561 para 197 colunas após esse processo.

- PCA (Análise de Componentes Principais): Utilizamos o PCA para reduzir as dimensões do dataset, extraindo os componentes principais que explicam a maior parte da variabilidade dos dados. No entanto, o PCA não nos foi útil pois os loadings das features eram muito próximos.

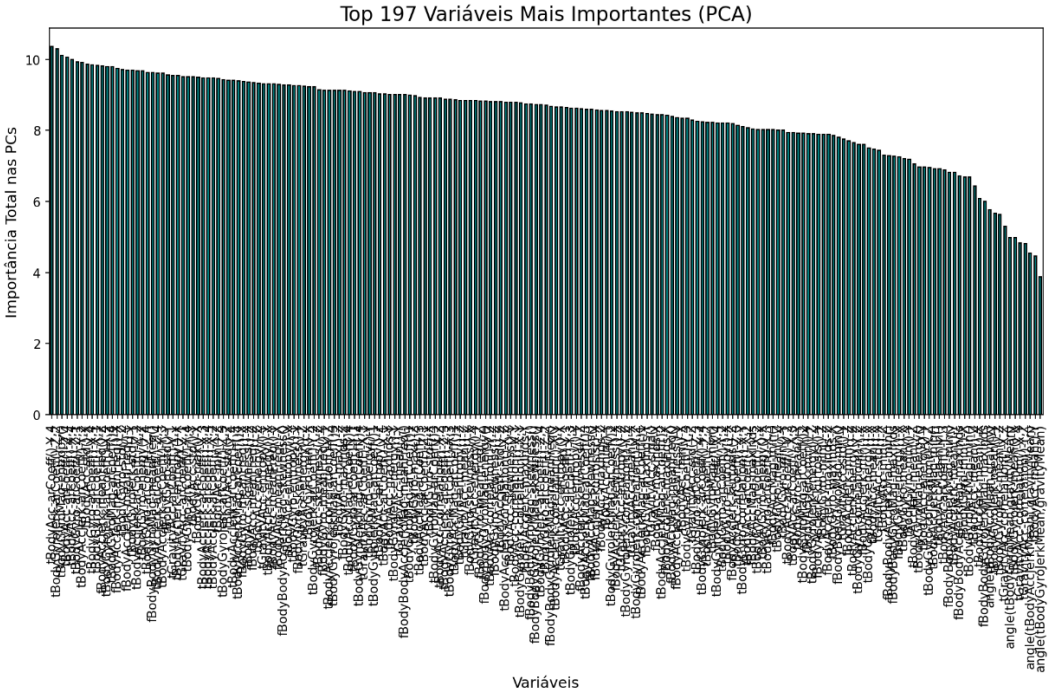

Figura 1:Top 197 variáveis mais importantes do PCA. Fonte: autores.

- Random Forest: O Random Forest foi empregado para calcular a importância das variáveis, permitindo a eliminação das menos relevantes. Essa técnica também foi aplicada para reduzir o número de features do dataset, resultando em uma redução de 197 para 50 variáveis mais significativas.

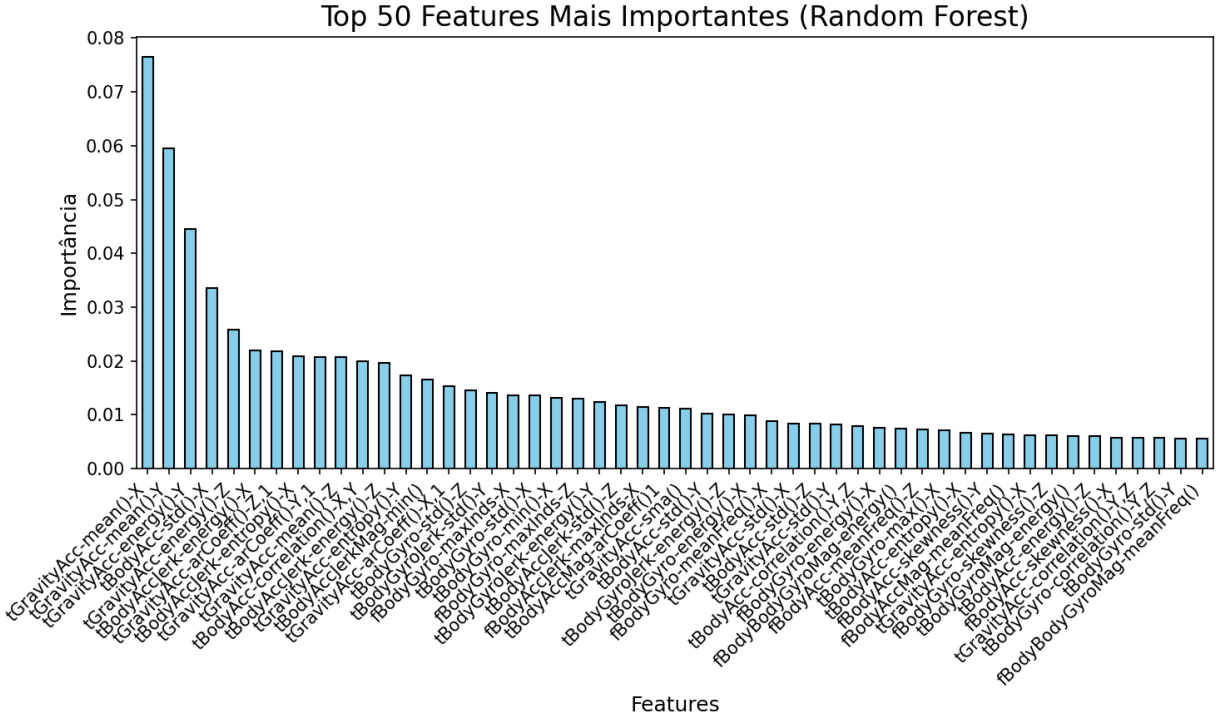

Figura 2: Top 50 features mais importnates da Random Forest. Fonte: autores.

A combinação dessas abordagens garantiu que as variáveis mais relevantes fossem preservadas, enquanto as menos importantes foram descartadas, resultando em um conjunto de dados reduzido a 50 variáveis.

2.2 Implementação do Algoritmo

Após a redução da dimensionalidade, foi aplicado o algoritmo K-means para realizar a

clusterização dos dados. A escolha do número de clusters (K) é uma etapa crucial no processo de modelagem, e foi feita utilizando dois métodos principais:

- Método do Cotovelo (Elbow Method): Embora o método do cotovelo tenha mostrado uma diminuição na inércia (distância intra-cluster) à medida que K aumentava, a escolha do número ideal de clusters foi difícil devido à alta dimensionalidade (50 dimensões). A "maldição da dimensionalidade" tornou a análise das distâncias mais desafiadora, pois em dimensões altas as distâncias entre os pontos de dados tendem a se homogeneizar.

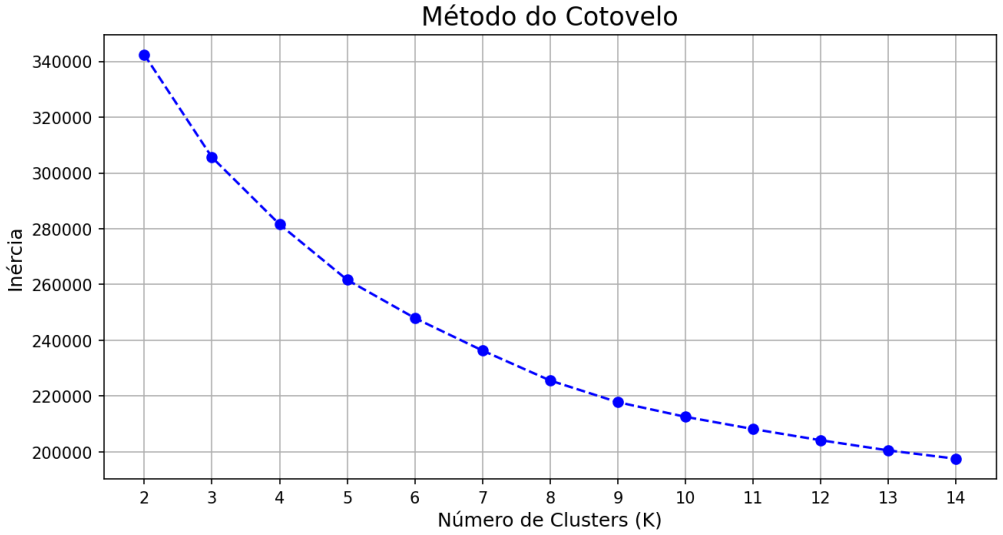

Figura 3: gráfico do Método do Cotovelo. Fonte: autores.

- Silhouette Score: Para melhorar a escolha do número de clusters, utilizamos o silhouette score, que mede a coesão e a separação entre os clusters. Os resultados mostraram que K = 3 foi o valor ideal, pois valores mais baixos (como K = 2) simplificariam demais a classificação, enquanto K = 3 ofereceu uma segmentação mais clara das atividades. As atividades puderam ser divididas em três grupos: atividades em movimento, atividades parado em pé e atividades parado deitado.

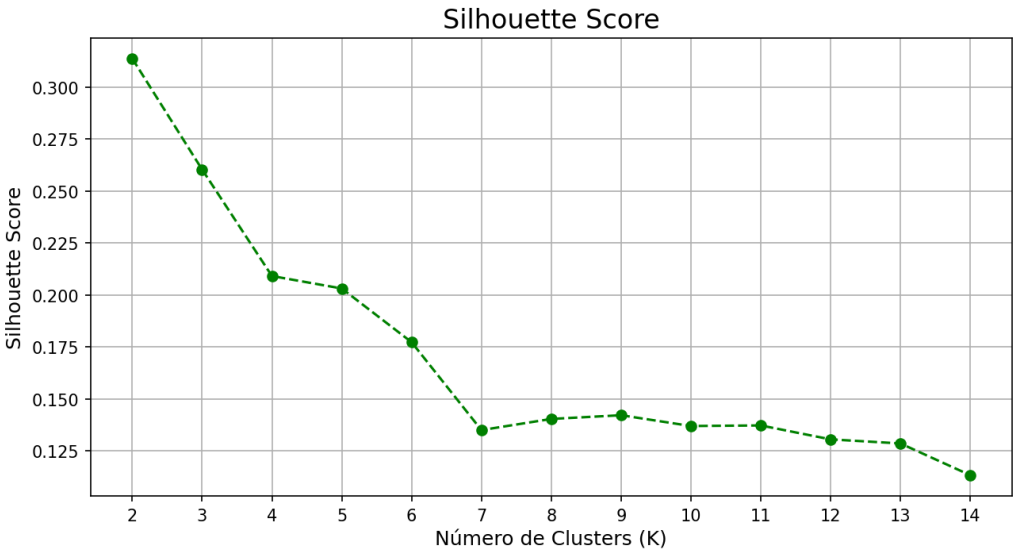

Figura 4: Gráfico do Silouette Score. Fonte: autores.

Além disso, o método de inicialização "K-means++" foi utilizado para garantir que os centros iniciais dos clusters fossem selecionados de forma eficiente, o que ajudou a otimizar a

convergência do algoritmo.

3. Resultados

Após treinar o modelo, realizamos 5 rodadas do K-means para verificar a consistência dos clusters. Os resultados indicaram que a clusterização foi bastante consistente entre as execuções, com os clusters sendo formados de maneira similar em todas as rodadas.

A normalização dos dados foi uma etapa importante, pois as variáveis possuíam escalas muito distintas. Utilizamos o método de normalização MinMax para garantir que todas as variáveis tivessem a mesma escala, contribuindo de forma equilibrada para o processo de agrupamento.

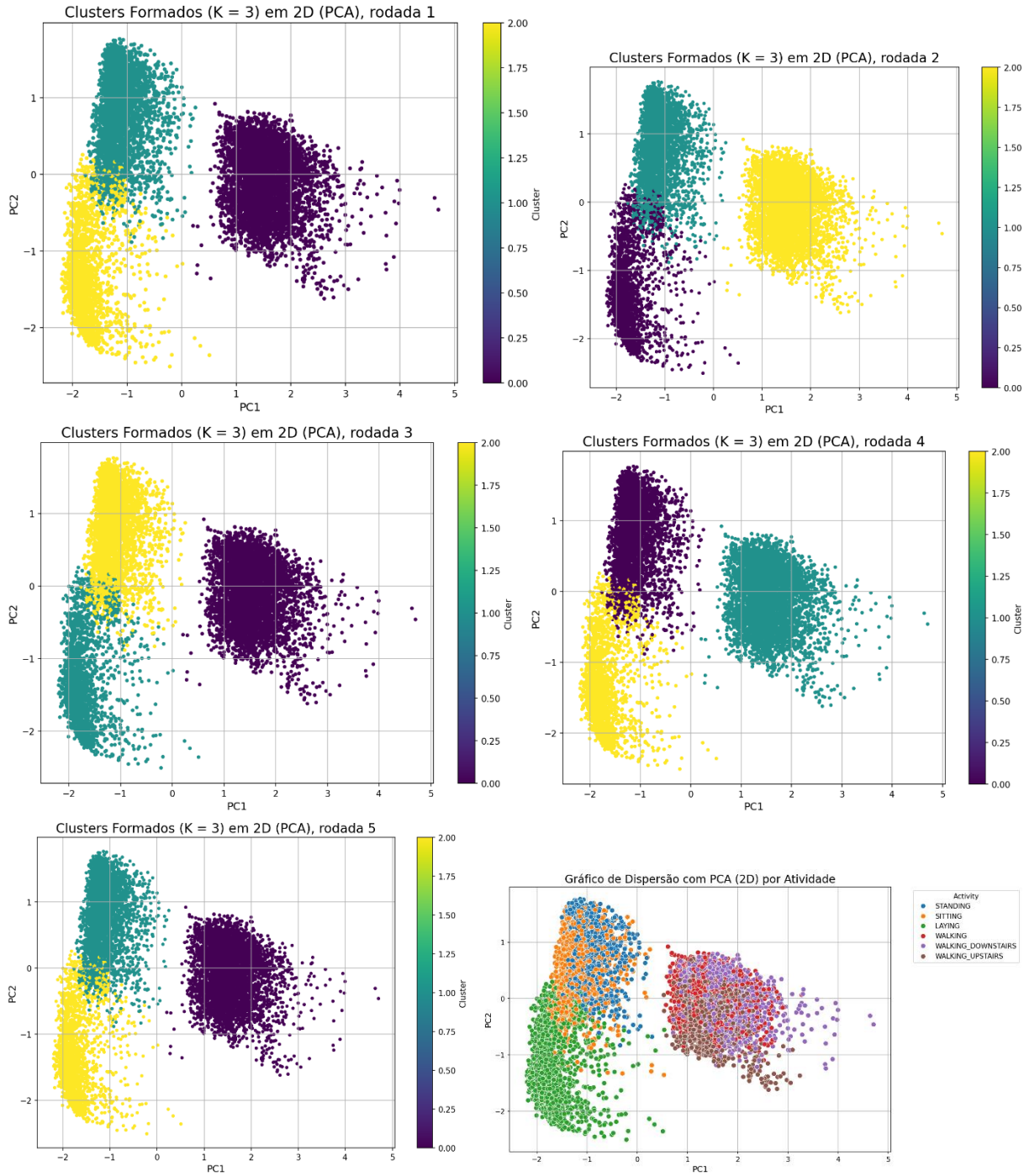

Figura 5: Gráficos de disperção das 5 rodadas do K-means e por fim o gráfico de dispersão com os rótulos originais dos dados. Fonte: autores.

Esses resultados indicam que o modelo conseguiu formar clusters relativamente coesos e bem separados.

A visualização dos clusters em 2D, após a redução de dimensionalidade usando PCA, foi feita com sucesso. O gráfico de dispersão confirmou que o modelo foi eficaz na segmentação das atividades:

Atividades em movimento: Representadas por pontos vermelhos, roxos e marrons.
Atividades parado em pé: Representadas por pontos azuis e amarelos.
Atividades parado deitado: Representadas por pontos verdes.

4. Conclusão e Trabalhos Futuros

Embora o método do cotovelo tenha apresentado dificuldades devido à alta dimensionalidade, o uso do silhouette score foi crucial para identificar o número de clusters ideal. A decisão de utilizar 3 clusters foi bem justificada, pois permitiu uma segmentação das atividades de forma mais granular e interpretável. Além disso, foi realizada uma análise qualitativa dos dados, que ajudou significativamente na escolha do número de clusters. Observamos que as 6 atividades presentes no dataset poderiam ser agrupadas em três grandes categorias:

- Atividades em movimento: Essas atividades envolvem mudanças rápidas de posição ou movimentos contínuos, como caminhar ou correr.
- Atividades parado em pé: Atividades onde o indivíduo permanece em uma posição estática, mas em pé, como ficar em pé sem se mover.
- Atividades parado deitado: Atividades em que o indivíduo está imóvel e deitado, como descansar ou dormir.

Essa análise qualitativa das atividades corroborou a escolha de 3 clusters, uma vez que as atividades naturais podem ser facilmente agrupadas nessas três categorias amplas. Essa segmentação foi vista como intuitiva, pois se alinha com a compreensão geral dos comportamentos humanos em termos de movimento, postura e repouso.

5. Referências

ANGUITA, D. et al. A public domain dataset for human activity recognition using smartphones. 2013. Disponível em: https://www.semanticscholar.org/paper/A-Public-Domain-Dataset-for-Human-Activity-Using-Anguita-Ghio/9de254d9b174c92f8249e8cd2e2c694eced515a3. Acesso em: 1 dez. 2024.

SONAWANE, N. et al. Activity recognition using smartphones: A review of challenges and methods. arXiv preprint, 2018. Disponível em: https://arxiv.org/abs/2404.02869. Acesso em: 1 dez. 2024.

UCI Machine Learning Repository. Human Activity Recognition Using Smartphones Dataset. Disponível em: https://archive.ics.uci.edu/ml/datasets/Smartphone-Based+Recognition+of+Human+Activities+and+Postural+Transitions. Acesso em: 1 dez. 2024.

YOUTUBE. Project 18: Human Activity Recognition with Smartphones | End to End Machine Learning Projects. Disponível em: https://www.youtube.com/watch?v=0AxDJPp7ssE. Acesso em: 1 dez. 2024.