RESIDÊNCIA EM SOFTWARE
BAHIA + TECNOLOGIA + EMPREENDEDORISMO

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

Grupo 8: Vitor Miranda Sousa e Wallisson da Silva Dias

Data de entrega: 03/12/2024

Resumo

Este relatório descreve a implementação e avaliação do algoritmo de aprendizado de máquina K-means para análise de dados de sensores de acelerômetro e giroscópio do dataset "Human Activity Recognition Using Smartphones". O conjunto de dados inclui 6 atividades humanas específicas para classificação: WALKING (caminhada), WALKING_UPSTAIRS (subindo escadas), WALKING_DOWNSTAIRS (descendo escadas), SITTING (sentado), STANDING (em pé) e LAYING (deitado). O objetivo foi segmentar os dados em clusters representando diferentes atividades humanas, utilizando técnicas como análise exploratória, pré-processamento, redução de dimensionalidade com PCA, e avaliação do número ideal de clusters com o método do cotovelo e o Silhouette Score. Os resultados indicaram que o K-means foi eficaz, com uma melhoria na acurácia de 4,14% para 63,96%. No entanto, a sobreposição de alguns clusters sugeriu que outras técnicas de clustering e redução de dimensionalidade poderiam ser exploradas para aprimorar a segmentação e a interpretação dos dados.

Introdução

O avanço das tecnologias de aprendizado de máquina tem transformado a forma como analisamos e interpretamos grandes volumes de dados e no contexto da análise de dados de sensores, como aqueles provenientes de dispositivos móveis, esses métodos possibilitam identificar padrões complexos que, de outra forma, seriam difíceis de observar. De acordo com Mitchell (1997), algoritmos de aprendizado não supervisionado, como o K-means, oferecem uma abordagem eficiente para segmentar e agrupar informações, promovendo insights valiosos em diversos campos, incluindo saúde, esporte e ergonomia.

O dataset "Human Activity Recognition Using Smartphones", disponibilizado pelo repositório UCI Machine Learning, constitui um recurso valioso para estudos dessa natureza, pois ele coleta dados de sensores de acelerômetro e giroscópio durante a execução de atividades cotidianas por 30 voluntários, representando uma base rica para experimentos de aprendizado não supervisionado. Conforme discutido por Anguita et al. (2013), esses dados permitem investigar o potencial de algoritmos como o K-means na identificação de padrões em atividades humanas, contribuindo para o desenvolvimento de tecnologias como dispositivos vestíveis inteligentes e sistemas de monitoramento de saúde.

Diante disso, este projeto tem como objetivo implementar e avaliar o algoritmo de K-means usando o dataset mencionado, abrangendo desde a análise exploratória dos dados até a determinação do número ideal de clusters e a visualização dos resultados. A escolha do K-means se justifica por sua simplicidade e eficácia em tarefas de agrupamento (Jain, 2010) e por meio dessa análise, busca-se não apenas compreender melhor os padrões de atividades humanas, mas também validar a aplicabilidade do K-means nesse contexto, fornecendo uma base técnica sólida para futuros estudos.

Metodologia

A pesquisa foi conduzida utilizando técnicas de aprendizado de máquina, com foco na aplicação do algoritmo K-means e na análise exploratória dos dados. As etapas foram estruturadas conforme descrito a seguir.

1. **Coleta e carregamento dos dados**

Os dados foram obtidos a partir de links compartilhados no Google Drive, utilizando a biblioteca *gdown*. Quatro arquivos principais foram baixados:

X_train.txt: conjunto de treinamento das variáveis independentes;

X_test.txt: conjunto de teste das variáveis independentes;

features.txt: arquivo contendo os nomes das variáveis;

ytest.txt: rótulos reais para o conjunto de teste.

Os dados foram carregados para *dataframes* com a biblioteca pandas para facilitar sua manipulação.

2. **Análise exploratória de dados**

A análise inicial buscou compreender a estrutura e a qualidade dos dados, envolvendo:

- Verificação de dimensões e tipos de dados;
- Identificação de valores ausentes ou inconsistências;
- Cálculo de estatísticas descritivas para observar tendências e padrões iniciais.

3. Pré-processamento dos dados

Para preparar os dados para o algoritmo de agrupamento, as etapas a seguir foram realizadas:

- Renomeação das colunas: as variáveis foram renomeadas com base no arquivo features.txt, garantindo clareza e consistência.

4. Redução de dimensionalidade com PCA

Aplicou-se o *Principal Component Analysis* (PCA) para reduzir a dimensionalidade do conjunto de dados, de forma a minimizar a redundância entre variáveis e melhorar o desempenho do K-means.

5. Implementação do K-means

O algoritmo K-means foi implementado para agrupar as observações em clusters. A escolha do número de clusters (K) foi baseada em dois métodos complementares:

- Método do Cotovelo: analisou-se a soma das distâncias quadráticas dentro dos clusters para diferentes valores de K.
- Coeficiente Silhouette: avaliou-se a qualidade do agrupamento para diferentes valores de K.

6. Avaliação do modelo

A qualidade do modelo foi avaliada comparando os clusters previstos com os rótulos reais e para melhorar o alinhamento, utilizou-se o Algoritmo Húngaro, que reordenou os clusters gerados pelo K-means de modo a maximizar a acurácia.

7. Visualização dos resultados

Gráficos foram gerados para analisar a separação entre os clusters e o alinhamento com os rótulos reais.

- Utilizou-se seaborn para criar gráficos de dispersão entre os componentes principais;
- Gráficos de barras foram utilizados para representar as frequências dos rótulos dentro de cada cluster.

Essas visualizações ajudaram a identificar áreas de sobreposição e padrões no agrupamento.

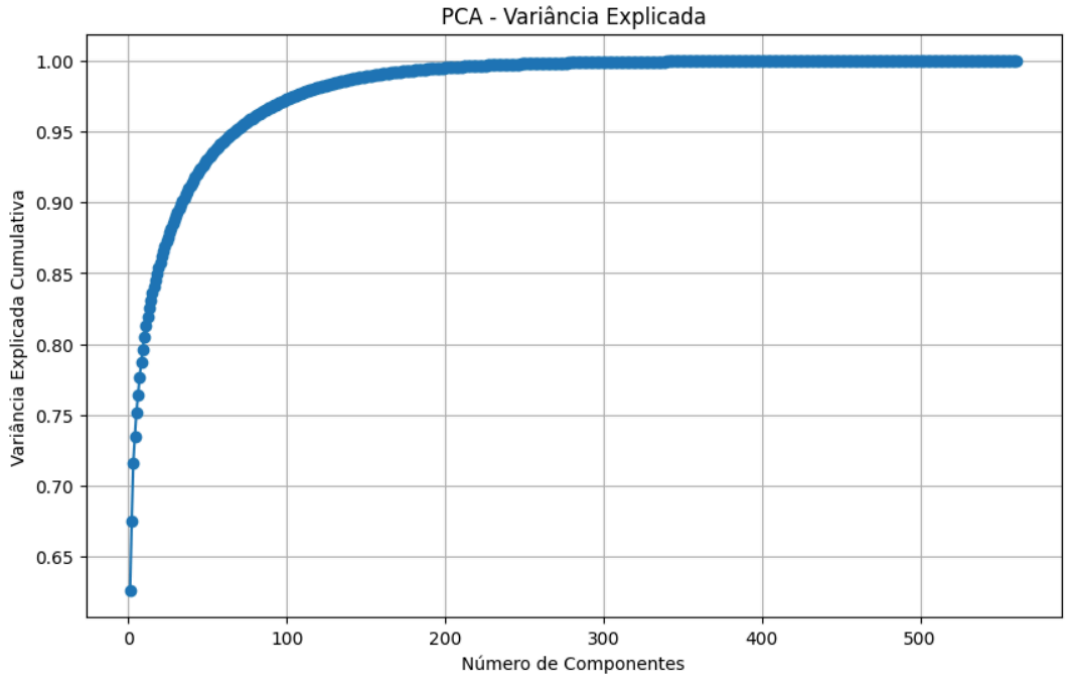
6

Resultados

Análise da Variância Explicada com PCA

A análise de componentes principais (PCA) foi aplicada para reduzir a dimensionalidade dos dados, mantendo a maior quantidade possível de variância. O gráfico de variância explicada cumulativa mostra que, com 67 componentes, é possível explicar aproximadamente 95% da variância dos dados, o que indica que a dimensionalidade foi adequadamente reduzida sem perder informações significativas.

Figura 1: PCA - Variância Explicada.

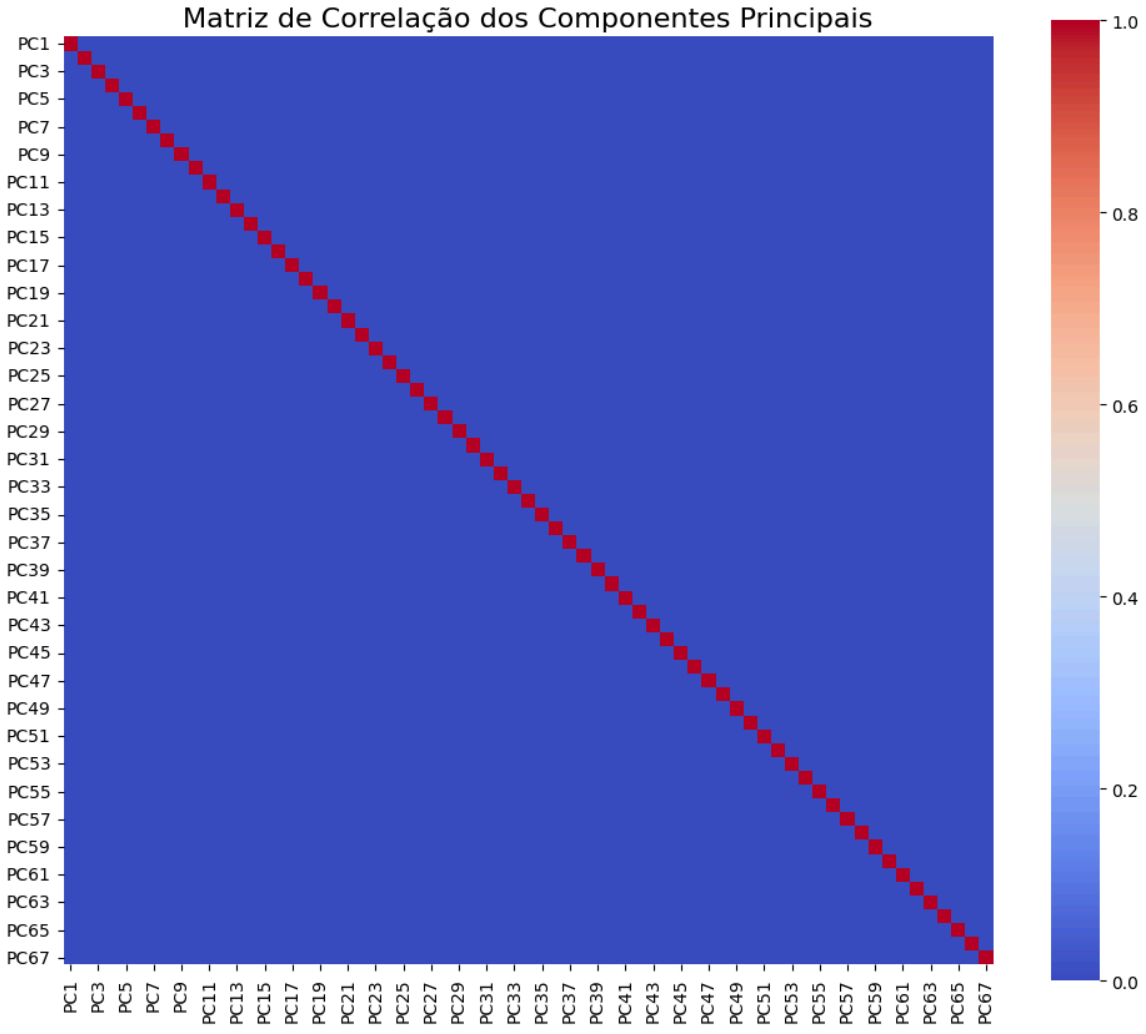


Fonte: Dados do próprio estudo.

Matriz de Correlação dos Componentes Principais

A matriz de correlação entre os componentes principais foi analisada para verificar a independência dos novos componentes gerados pela PCA e foi observado que a maioria das correlações entre os componentes principais (PCs) é baixa, o que sugere que os componentes são ortogonais e independentes entre si.

Figura 2: Matriz de Correlação dos Componentes Principais.

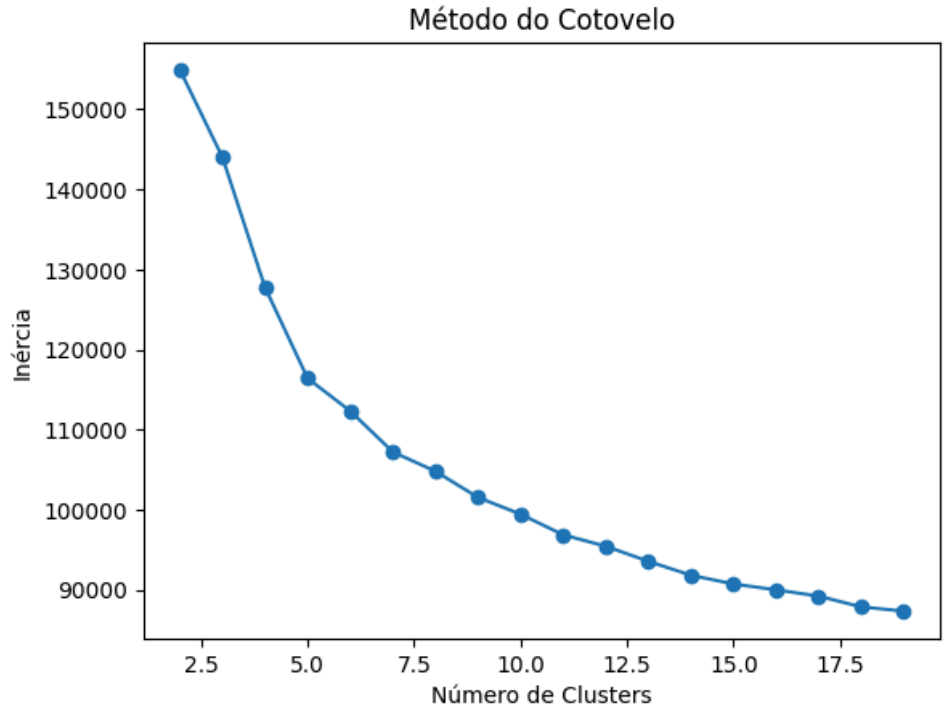


Fonte: Dados do próprio estudo.

Método do Cotovelo

O método do cotovelo foi utilizado para determinar o número ideal de clusters. A análise da inércia para diferentes valores de K revelou um ponto de inflexão em K=6, sugerindo que 6 clusters seriam adequados para a segmentação dos dados.

Figura 3: Método do Cotovelo.

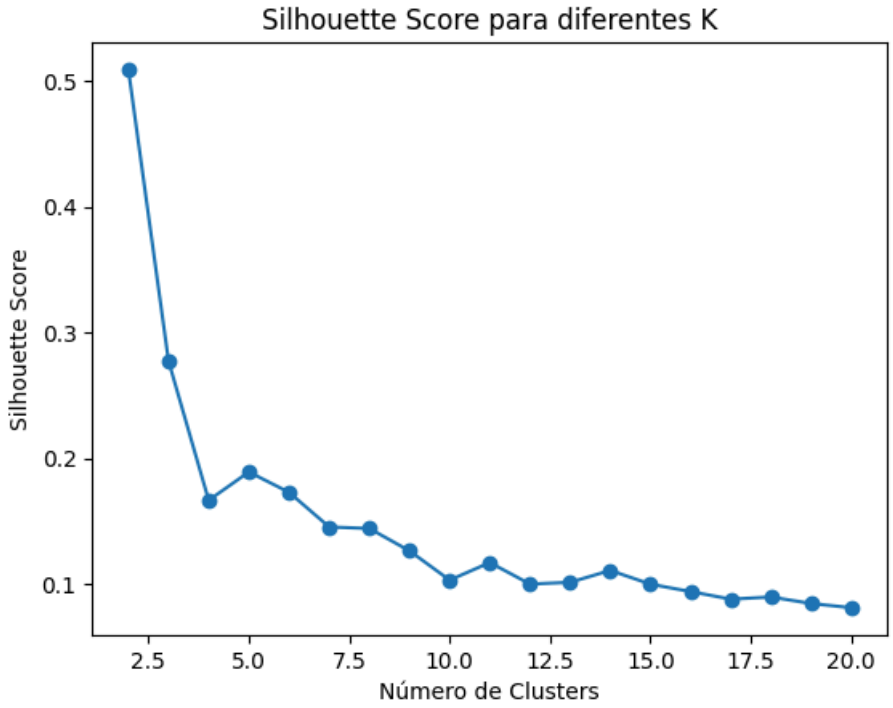


Fonte: Dados do próprio estudo.

Análise da Coesão e Separação dos Clusters

O Silhouette Score foi calculado para diferentes valores de K. Com K=6, obteve-se uma alta coesão e boa separação entre os clusters, o que indica que o número de clusters é adequado para este conjunto de dados. Assim, os 6 clusters fazem sentido, pois os dados dos sensores estavam avaliando as atividades: (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING), totalizando 6 atividades humanas.

Figura 4: gráfico do Silhouette Score.



Fonte: Dados do próprio estudo.

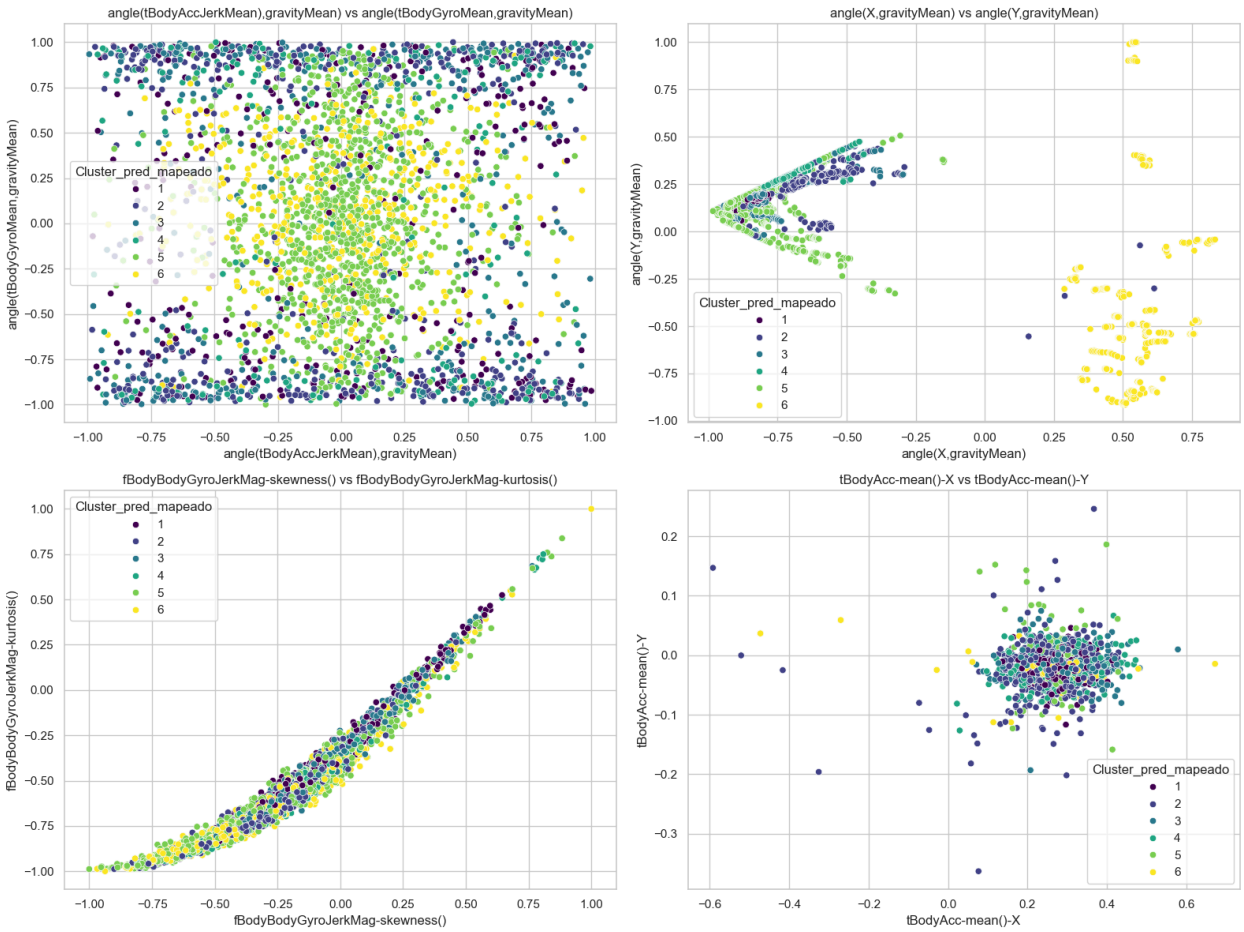
Avaliação da Qualidade dos Clusters

Antes do alinhamento dos dados, a acurácia inicial foi de 4,14%. Após o alinhamento e a segmentação dos dados com os clusters, a acurácia foi ajustada para 63,96%, mostrando uma melhoria significativa na classificação e organização dos dados.

Visualização dos Dados Clusterizados

Uma visualização do conjunto de dados com os componentes principais foi realizada para observar como os clusters foram formados. Ao analisar os dados clusterizados, percebeu-se que alguns clusters estavam bem definidos, enquanto outros mostraram um certo grau de sobreposição.

Figura 5: Visualização dos Dados Clusterizados.

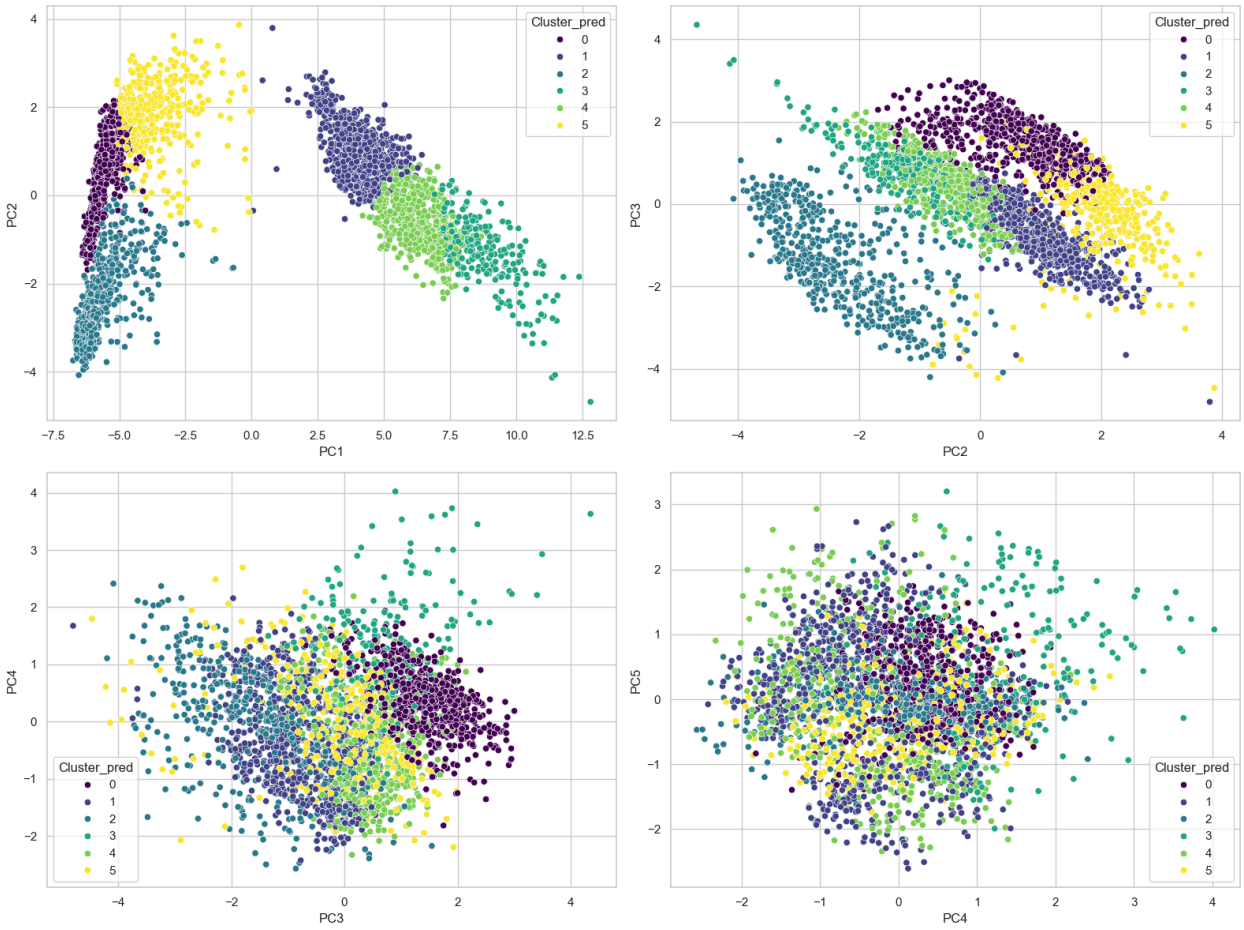


Fonte: Dados do próprio estudo.

Análise dos Agrupamentos

Foi utilizado um DataFrame com os componentes principais clusterizados. As colunas 'PC1', 'PC2', 'PC3', 'PC4' e 'PC5' foram colocadas em subplots para avaliar como os dados foram agrupados. Nos três primeiros subplots, é possível observar agrupamentos claros, com uma boa separação dos clusters, enquanto nos demais a separação não foi tão nítida.

Figura 6: Visualização dos Componentes Principais Clusterizados.



Fonte: Dados do próprio estudo.

Discussão

Os resultados obtidos com a análise de componentes principais (PCA) e a segmentação dos dados com clustering fornecem insights valiosos sobre a estrutura dos dados e a adequação dos métodos aplicados.

Reflexão sobre os Resultados

A redução da dimensionalidade através da PCA foi eficaz, como evidenciado pela explicação significativa da variância (cerca de 95% com 67 componentes), permitindo a manutenção de informações essenciais dos dados sem um grande custo computacional. Isso sugere que, ao aplicar PCA, conseguimos preservar a maior parte da estrutura dos dados em um número reduzido de dimensões, facilitando a análise subsequente.

O método do cotovelo, ao identificar K=6 como o número ideal de clusters, também apontou uma segmentação adequada, que foi confirmada pelo Silhouette Score. A alta coesão e boa separação entre os clusters indicam que a segmentação é eficaz para capturar as estruturas subjacentes dos dados. A melhoria na acurácia de 4,14% para 63,96% após a segmentação dos dados reforça que a aplicação de PCA e clustering contribui para uma melhor organização dos dados, além de uma classificação mais precisa.

No entanto, a visualização dos dados clusterizados revelou que alguns clusters estavam bem definidos, enquanto outros apresentaram sobreposição. Esse aspecto sugere que, embora a segmentação tenha sido eficaz, a separação entre certos grupos poderia ser mais clara, possivelmente devido à natureza dos dados ou à escolha do algoritmo de clustering.

Limitações

Apesar das melhorias na acurácia e na segmentação dos dados, algumas limitações devem ser destacadas:

- Redução de Dimensionalidade: Embora a PCA tenha sido bem-sucedida na redução da dimensionalidade, a escolha do número de componentes principais não foi baseada em uma análise mais aprofundada da contribuição de cada variável para a variância. Embora 95% de variância explicada seja uma boa meta, uma análise mais detalhada poderia ter revelado se componentes adicionais poderiam ter impactado positivamente o modelo.
- Método de Clustering: O algoritmo de clustering escolhido, juntamente com a definição do número de clusters pelo método do cotovelo, pode não ser o ideal para todos os

tipos de dados. A sobreposição entre alguns clusters pode sugerir que o modelo não foi capaz de capturar completamente a estrutura dos dados. Alternativas, como o uso de algoritmos de clustering hierárquico ou DBSCAN, poderiam ser exploradas para verificar se esses métodos poderiam fornecer melhores resultados em termos de separação dos grupos.

- Escolha do Silhouette Score: Embora o Silhouette Score tenha sido útil para medir a coesão e separação dos clusters, ele nem sempre é um indicador definitivo de um bom modelo. Em casos de dados com uma estrutura mais complexa, o método pode não refletir totalmente a qualidade dos clusters, especialmente se houver ruídos ou pontos fora da curva que podem ser mal interpretados.
- Visualização dos Dados: A análise visual dos dados clusterizados pode ser limitada, pois a projeção dos dados em um número reduzido de dimensões (como os dois primeiros componentes principais) nem sempre revela a verdadeira estrutura dos dados. A visualização pode ter perdido nuances importantes que poderiam ser detectadas com métodos de visualização em maior número de dimensões ou usando outras técnicas de redução dimensional.

Impacto das Escolhas Feitas

As escolhas feitas durante o desenvolvimento do modelo tiveram um impacto significativo nos resultados e na interpretação dos dados:

- Escolha do PCA: A redução de dimensionalidade por meio do PCA permitiu que o modelo fosse mais eficiente, com menos recursos computacionais necessários, mantendo uma boa explicação da variância. No entanto, a aplicação do PCA pode ter mascarado características sutis que existiam nos dados originais, especialmente se algumas variáveis importantes foram eliminadas durante o processo.
- Método de Clustering e Avaliação de K: A escolha do número de clusters com o método do cotovelo foi fundamental para segmentar os dados de maneira adequada. No entanto, o impacto dessa escolha poderia ter sido mais aprofundado se houvesse a consideração de diferentes algoritmos de clustering ou a exploração de abordagens mais complexas para a definição do número ideal de clusters.

- Implicações Práticas: A melhoria na acurácia após a segmentação dos dados tem implicações práticas importantes. Uma acurácia de 63,96% pode ser suficiente em muitos cenários, mas é importante considerar a aplicabilidade do modelo para problemas do mundo real. Em casos de dados mais complexos ou que exigem alta precisão, outras abordagens de modelagem podem ser necessárias para alcançar resultados mais robustos.

Conclusão

A aplicação do PCA e do clustering no modelo trouxe aprendizados valiosos sobre como reduzir a complexidade dos dados e segmentá-los de maneira eficaz, sendo que o PCA permitiu a extração das principais variáveis, preservando a maior parte da variância dos dados e melhorando a eficiência computacional e o uso do clustering com o método do cotovelo e a avaliação do Silhouette Score também mostrou-se eficaz para segmentar os dados em grupos significativos, resultando em uma melhora na acurácia do modelo.

Apesar dos avanços, o projeto também destacou áreas que podem ser aprimoradas como a escolha do número de componentes principais e o algoritmo de clustering que podem ser revisados para garantir uma melhor separação entre os clusters e a captura de mais informações relevantes. Também a questão da sobreposição observada em alguns clusters sugere que outras abordagens, como o uso de técnicas de clustering alternativas (DBSCAN ou hierárquico), poderiam oferecer resultados mais robustos.

Como sugestões de melhoria, seria interessante explorar outras técnicas de redução de dimensionalidade, que podem ser mais eficazes para visualização e segmentação em dados complexos e também seria válido testar diferentes algoritmos de clustering, ajustando os parâmetros ou utilizando validações cruzadas para avaliar a robustez das segmentações. Dessa forma, o projeto poderia se beneficiar de uma maior precisão e flexibilidade, melhorando a adaptação do modelo a diferentes tipos de dados e contextos.

Referências

ANGUITA, D. et al. A Public Domain Dataset for Human Activity Recognition Using Smartphones. In: *21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges, Belgium, 2013.

JAIN, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, v. 31, n. 8, p. 651-666, 2010.

MITCHELL, T. M. *Machine Learning*. McGraw Hill, 1997.