RELATÓRIO TÉCNICO

Implementação e Análise do Algoritmo de K-Means

Marya de Souza Fernandes Matos
Daiane Gomes Meira
03/dez/2024

Marya de Souza Fernandes Matos
Daiane Gomes Meira

Resumo

Este relatório busca contribuir com uma abordagem sistemática para análise exploratória, tratamento de dados e clusterização, para possíveis aplicações em reconhecimento de padrões, análise de comportamento e personalização de serviços. O objetivo principal desse projeto foi explorar e analisar o conjunto de dados “Human Activity Recognition Using Smartphones”, para identificar padrões relevantes e realizar uma segmentação eficiente utilizando o algoritmo K-Means para, de forma não supervisionada, entender as relações entre essas atividades humanas.

A metodologia envolveu a aplicação de padronização das variáveis buscando mantê-las na mesma faixa de valor; técnicas de tratamento de outliers; redução de dimensionalidade para facilitar a visualização dos dados; análise da variância e correlação entre as variáveis e identificação e remoção de multicolinearidade por meio do cálculo do Variance Inflation Factor (VIF). Além disso, foram utilizados métodos de teste e visualização para encontrar o valor ideal de K.

Os resultados obtidos mostram que as técnicas de redução de dimensionalidade foram capazes de explicar quase 54% da estrutura dos dados com apenas dois componentes principais. Esse valor pode ser considerado limitado, mas foi suficiente para visualização e clustering. O K-Means conseguiu captar melhor as informações de uma parte das classes do que da outra, mas ainda assim, obteve uma média de 72% de definição adequada dos clusters.

Introdução

O reconhecimento de atividades humanas é um campo crescente no desenvolvimento de sistemas inteligentes, com aplicações em áreas como saúde, esportes e segurança, por exemplo. A capacidade de identificar padrões em dados provenientes de sensores possibilita monitorar atividades diárias, detectar

comportamentos anormais e oferecer serviços personalizados. Um bom exemplo de aplicação é o uso de smartphones/smartwatches que podem capturar dados de movimentos e utilizá-los no rastreamento de atividades físicas.

O conjunto de dados “Human Activity Recognition Using Smartphones” representa um cenário típico desse problema, contendo medições de acelerômetros e giroscópios obtidos através de dispositivos móveis. Esses dados oferecem uma oportunidade de explorar e identificar padrões associados a diferentes atividades humanas, como caminhar, subir escadas ou ficar em pé, utilizando técnicas de análise de dados e machine learning.

A eficiência e simplicidade na segmentação de grandes volumes de dados contínuos, como é o caso do problema desse projeto, fez com que o algoritmo K-Means fosse escolhido para esse estudo. O K-Means é útil para agrupar dados de alta dimensionalidade em clusters que compartilhem características semelhantes, de forma não supervisionada.

O conjunto de dados utilizado neste estudo foi obtido a partir do repositório UCI Machine Learning, e contém medições de 30 indivíduos que realizaram atividades da vida diária (andar, subir escadas, descer escadas, sentar, ficar em pé, ficar e deitar) enquanto carregavam um smartphone preso à cintura com sensores inerciais incorporados. O dataset apresenta 561 variáveis calculadas a partir dos sinais brutos desses sensores, sendo ideal para tarefas de agrupamento e análise de atividades. Cada uma dessas variáveis foi pré-processada e transformada conforme necessário, buscando facilitar o entendimento e o manuseio desses dados, garantindo a qualidade deles e a robustez das previsões.

O principal desafio na aplicação do K-Means no contexto desse problema é garantir que as etapas de pré-processamento e redução de dimensionalidade preservem as informações críticas para a segmentação. Este trabalho aborda esses desafios e explora como o K-Means pode ser utilizado de forma eficiente para identificar padrões em atividades humanas, oferecendo insights valiosos sobre os dados analisados.

Metodologia

A etapa inicial do projeto constituiu na pré-análise dos dados, com o objetivo de carregar os dados brutos de forma correta e eficiente através da URL do repositório. Depois de carregados, deu-se início à análise exploratória dos dados, com o objetivo de entender a estrutura e as características do conjunto de dados e das variáveis disponíveis e identificar padrões relevantes. Foram gerados boxplots para verificar a presença de outliers de algumas variáveis, que foram tratados por Windsorização, limitando valores extremos detectados a um intervalo aceitável. Como o K-Means é um algoritmo bem sensível à escala, o *RobustScaler* foi utilizado para normalizar os dados, padronizando as variáveis em torno da mediana e escalando-as com base no intervalo interquartil.

Buscando pelo melhor desempenho, alguns modelos foram trabalhados. No primeiro modelo a ser testado, a variância das variáveis foi verificada, buscando por aquelas que pudessem ser insignificantes para o modelo. Como nenhuma variável com baixa variância foi identificada, todas elas seguiram para o próximo passo, que foi uma análise de correlação. Como as altas correlações não são, necessariamente, um problema nesse caso em específico, já que a redução da dimensionalidade reduz redundâncias automaticamente, todas as variáveis continuaram sendo mantidas. O *variance_inflation_factor* do *statsmodels* não é otimizado para datasets grandes, buscando analisar as variáveis com alta multicolinearidade, então o VIF pôde ser calculado através da matriz de correlação e sua inversa. Depois do filtro de multicolinearidade, foi utilizada a técnica de Análise de Componentes Principais (PCA) para lidar com a alta dimensionalidade dos dados, reduzindo-os para 2 dimensões. Nessas circunstâncias obteve-se uma variância explicada pelos dois componentes de menos de 33%, podendo indicar que a maioria das informações nos dados originais pode ter sido descartada. Além disso, nesse caso, também foi necessário um tratamento extra de outliers nesses 2 componentes.

Em um segundo teste, ao invés de fazer uma seleção e aplicar filtros nas variáveis, a técnica do PCA foi logo aplicada no conjunto total de variáveis. Nesse contexto, a variância explicada pelos 2 componentes foi de quase 54%, mostrando que menos informações foram perdidas na redução da dimensionalidade. O

tratamento dos outliers não precisou ser repetido e as distribuições dos componentes também pareceu ser mais normalizada do que antes.

Em um terceiro modelo, a técnica utilizada para reduzir a dimensionalidade foi o Mapeamento Uniforme de Aproximação e Projeção (UMAP), que pode capturar melhor as relações não lineares nos dados.

O K-Means foi escolhido como o algoritmo a ser trabalhado por sua eficiência computacional e capacidade de segmentar os dados em clusters baseados em similaridades. A definição do número ideal de clusters foi realizada com algumas métricas de avaliação. A primeira delas foi o método do cotovelo (*Elbow Method*), que avalia a soma das distâncias quadradas dentro dos clusters em função do número de clusters. Matematicamente, o ponto que indica o k ideal é o ponto da curva mais distante de uma reta traçada entre o primeiro e o último ponto. Uma segunda métrica de avaliação é o índice de silhueta (*Silhouette Score),* que mede a coesão interna e a separação externa dos clusters para identificar o número de k que maximizasse a qualidade da segmentação. Em posse dessas informações de k ideal, o modelo de K-Means foi implementado usando a biblioteca *scikit-learn*, que permite uma integração eficiente com diversas outras ferramentas de análise. O modelo foi inicializado com o método K-Means++, que tende a melhorar a convergência e, geralmente, reduz a variabilidade entre as execuções, resultando em um melhor desempenho do algoritmo e clusters mais estáveis. Além disso, o parâmetro *n_init* controla o número de vezes que o algoritmo será executado com diferentes inicializações de centroides. Nesse projeto, o algoritmo foi executado 20 vezes e depois o modelo foi ajustado e retornou os rótulos classificados.

Para avaliar os resultados, inicialmente a separação e a coesão dos grupos formados foram verificadas através da visualização dos clusters em duas dimensões (obtidas pelo PCA e pelo UMAP); e foi feita uma avaliação qualitativa da consistência dos clusters formados com métricas como o Coeficiente de Silhouette, o índice de Calinski-Harabasz, a homogeneidade e a completude. Para complementar a avaliação dos resultados, foi utilizada uma matriz de contingência, que é essencial para comparar os clusters gerados pelo K-Means com as classes reais do conjunto de dados.

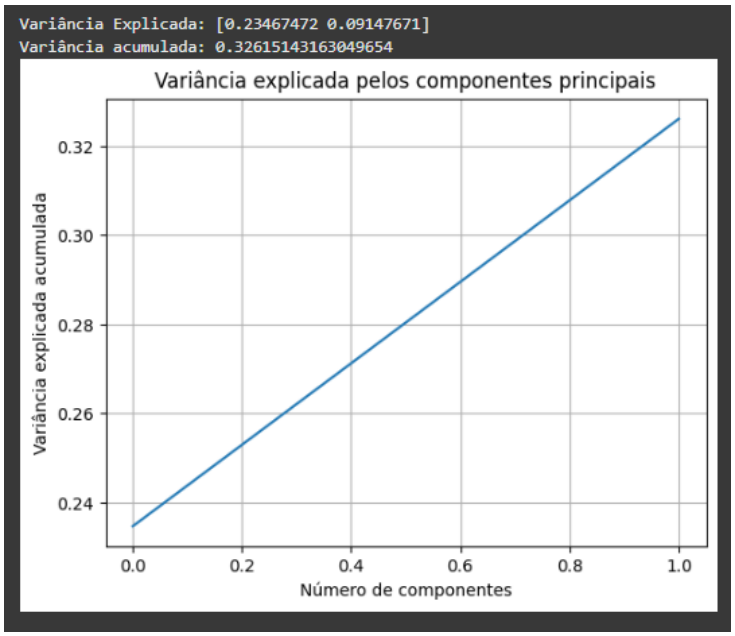
A escolha do melhor modelo foi determinada com base nas melhores métricas e no melhor conjunto de características observadas.

Resultados

Durante o processo do projeto, alguns modelos foram testados para identificar a solução que apresentava o melhor equilíbrio entre desempenho e simplicidade. De uma forma geral, de acordo com as técnicas escolhidas, eles iam performando um pouco melhor a cada análise.

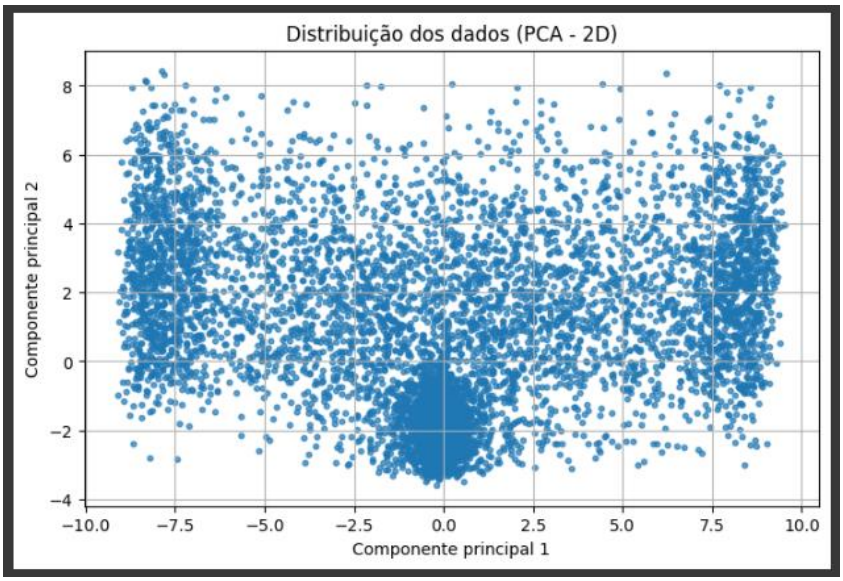
O primeiro modelo testado apresentou uma variância explicada pelos componentes principais abaixo do que poderia ser considerado bom, com uma taxa de menos de 34% e uma distribuição de dados bem dispersa depois da aplicação do PCA, como mostram as Figuras 1 e 2.

Figura 1 – Variância explicada pelos 2 componentes do modelo 1 testado.



Fonte: Próprio Autor (2024).

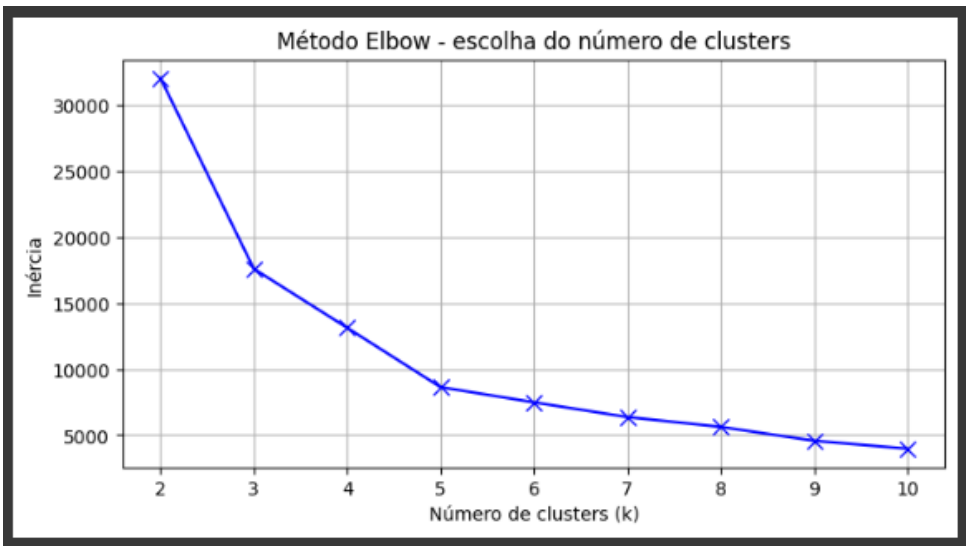
Figura 2 – Dispersão dos dados do modelo 1 testado.



Fonte: Próprio Autor (2024).

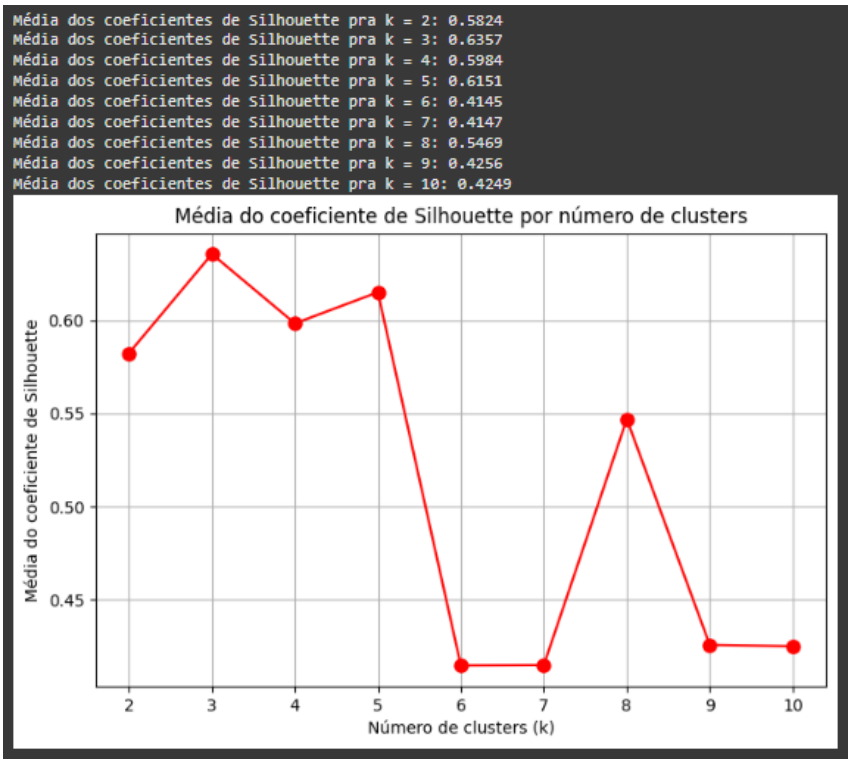
Os componentes passaram por um novo tratamento de outliers com a mesma técnica de windsorização utilizada no pré-processamento e depois disso, deu-se início à definição do número ideal de k para o modelo ser treinado. Inicialmente buscando o “cotovelo” pelo método de Elbow (Figura 3) e, depois, pelas médias dos coeficientes de Silhouette (Figura 4).

Figura 3 – Dispersão dos dados do modelo 1 testado.



Fonte: Próprio Autor (2024).

Figura 4 – Dispersão dos dados do modelo 1 testado.

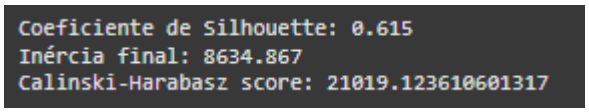


Fonte: Próprio Autor (2024).

O método de Elbow identificou o k=5 como o ponto de equilíbrio entre homogeneidade dentro dos clusters e a separação entre os clusters. Já as médias de Silhouette foram maiores pra um k=3, mas apresenta um valor que não é muito menor para um k=5, o que significa que os clusters podem continuar bem definidos. Dessa forma, foi conveniente seguir com o k=5 pro treinamento do modelo de K-Means.

O coeficiente de Silhouette varia entre -1 e 1, sendo que 1 significa que os clusters são bem definidos, com os pontos de cada cluster bem próximo do seu centroide e bem afastado de outros clusters. Depois de treinado, o modelo forneceu um coeficiente de Silhouette igual a 0.615, que é um resultado moderado. Não existe um valor absoluto ideal para o Índice de Calinski-Harabasz, mas pode-se dizer que valores maiores indicam clusters bem separados entre si e internamente coesos. Nesse modelo, o valor desse índice foi de 21019.124. A ideia é guardar esse valor para servir de comparação com outros modelos testados. As métricas alcançadas com esse modelo estão na Figura 5.

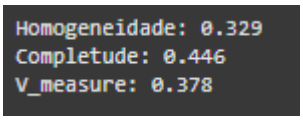
Figura 5 – Métricas de avaliação do modelo 1 testado.



Fonte: Próprio Autor (2024).

Outras métricas de avaliação foram os índices de homogeneidade, que mede o grau em que cada cluster contém apenas membros de uma única classe verdadeira, e de completude, que mede o grau em que todos os membros de uma classe verdadeira estão alocados no mesmo cluster. Esses valores variam entre 0 e 1, e quanto mais perto de 1, indicam uma melhor classificação e agrupamento dos clusters. O modelo obteve um índice de homogeneidade de 0.329 e de completude de 0.446. O balanceamento entre essas duas métricas é chamado de v-measure, e o valor obtido nesse modelo foi de 0.378, o que significa que o modelo ainda pode melhorar bastante. Essas métricas podem ser observadas na Figura 6.

Figura 6 – Métricas de avaliação do modelo 1 testado.

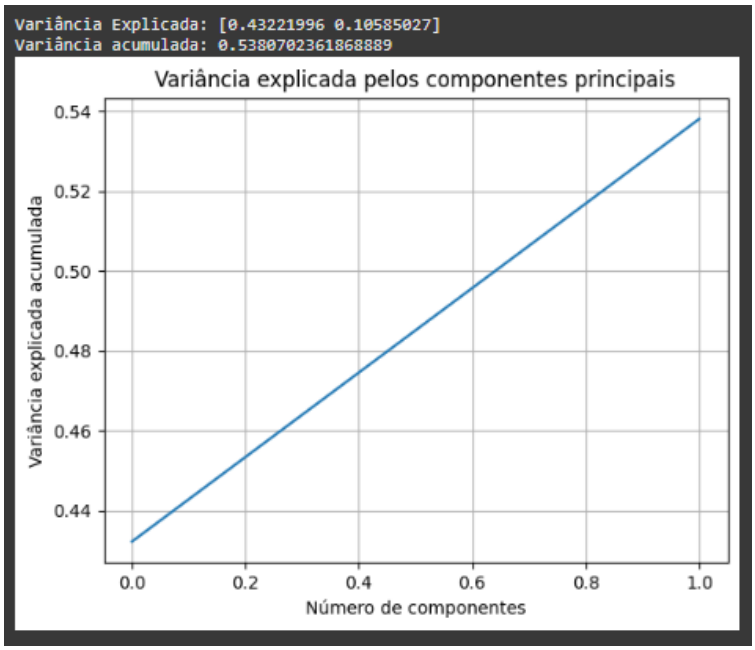


Fonte: Próprio Autor (2024).

Na matriz de contingência dá pra perceber que o modelo consegue captar relativamente bem as informações das atividades 4, 5 e 6, mas confunde bastante as informações das atividades 1, 2 e 3. Os clusters também parecem bem heterogêneos, tendo dados de quase todas as classes em todos eles.

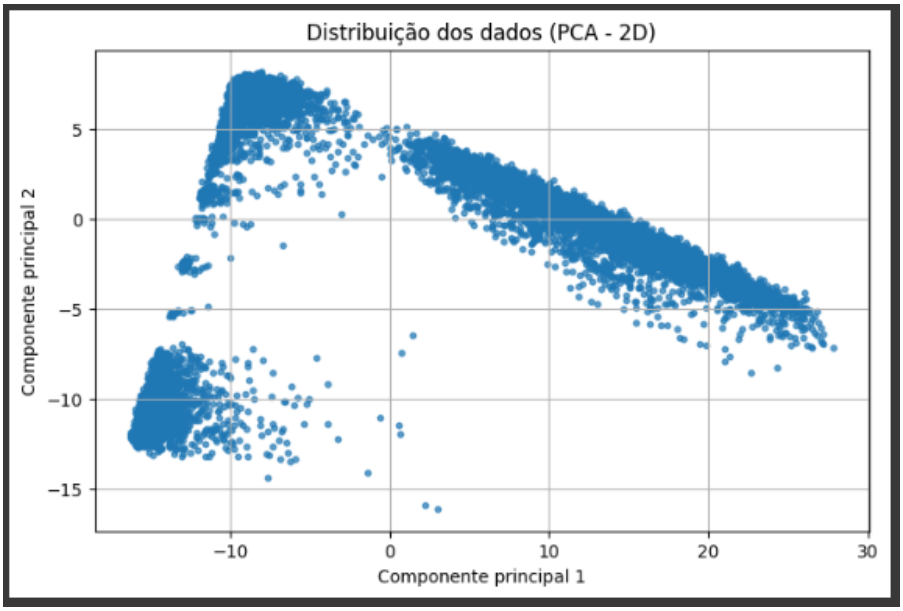
No segundo modelo testado, a variância explicada pelos componentes depois do PCA já aumentou para quase 54%, o que mostra que menos informações foram perdidas na redução de dimensionalidade. A variância pode ser vista na Figura 7. A distribuição dos dados já parece menos dispersa e pode ser vista na Figura 8.

Figura 7 – Variância explicada pelos 2 componentes do modelo 2 testado.



Fonte: Próprio Autor (2024).

Figura 8 – Dispersão dos dados do modelo 2 testado.



Fonte: Próprio Autor (2024).

Em relação à definição do valor ideal de k, o comportamento foi bem semelhante ao modelo 1. O método de Elbow identificou como ideal um k=4, enquanto que as médias dos coeficientes de Silhouette foram melhores pra um k=3, mas ficaram

bem semelhantes para um k=4. Nesse modelo, as métricas de avaliação melhoraram significativamente. O coeficiente de Silhouette subiu de 0.615 para 0.722, o índice de Calinski-Harabasz passou de 21019.124 para 78099.305 e o v-measure passou de 0.378 para 0.672, mostrando que os clusters já estão melhores divididos.

Figura 9 – Métricas de avaliação do modelo 2 testado.

```
Coeficiente de Silhouette: 0.722
Inércia final: 87436.317
Calinski-Harabasz score: 78099.30445305367
```

Fonte: Próprio Autor (2024).

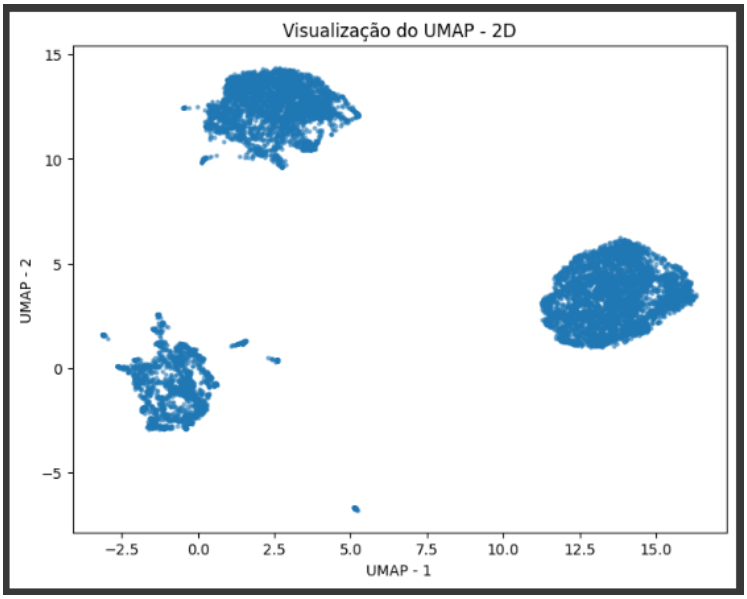
Figura 10 – Métricas de avaliação do modelo 2 testado.

```
Homogeneidade: 0.592
Completude: 0.778
V_measure: 0.672
```

Fonte: Próprio Autor (2024).

Finamente, o modelo 3 já mostrou uma distribuição de dados mais uniforme, como mostra a Figura 11. E no caso desse modelo, os métodos de Elbow e dos coeficientes de Silhouette suportaram a mesma decisão de um k ideal sendo igual a 3.

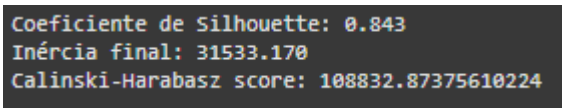
Figura 11 – Métricas de avaliação do modelo 3 testado.



Fonte: Próprio Autor (2024).

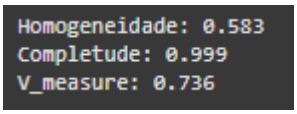
As métricas de avaliação obtidas com esse modelo foram significativamente superiores, com um coeficiente de Silhouette igual a 0.843, o que indica uma boa separação entre os clusters e uma boa coesão dentro de cada cluster. Além disso, apresenta um v-measure igual a 0.736.

Figura 12 – Métricas de avaliação do modelo 3 testado.



Fonte: Próprio Autor (2024).

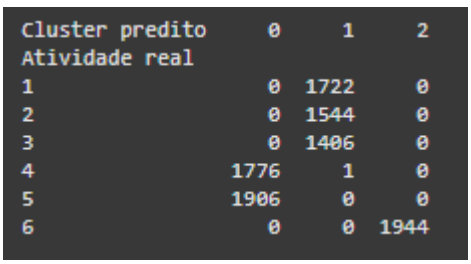
Figura 13 – Métricas de avaliação do modelo 3 testado.



Fonte: Próprio Autor (2024).

A matriz de contingência sugere que o modelo conseguiu agrupar as atividades reais em 3 grupos coerentes, como mostra a Figura 14. Entretanto, algumas atividades diferentes foram colocadas no mesmo cluster, como é o caso das atividades 2, 4 e 6 no cluster 0. Provavelmente, as atividades reais que compartilham clusters possuem padrões similares nas variáveis analisadas. Há pouca confusão evidente entre clusters, mas a união de múltiplas atividades em um único cluster pode impactar a utilidade prática do modelo, apesar das boas métricas.

Figura 14 – Métricas de avaliação do modelo 2 testado.



Cluster predito	0	1	2
Atividade real			
1	0	1722	0
2	0	1544	0
3	0	1406	0
4	1776	1	0
5	1906	0	0
6	0	0	1944

Fonte: Próprio Autor (2024).

Discussão

Durante o desenvolvimento deste projeto, três abordagens distintas foram implementadas para avaliar o impacto de diferentes estratégias de pré-processamento e redução de dimensionalidade no desempenho do modelo K-Means. O primeiro modelo incorporou um pipeline mais detalhado de pré-processamento com exclusão de algumas variáveis. Embora robusto, o modelo apresentou desempenho inferior em termos de métricas calculadas, possivelmente devido à eliminação excessiva de variáveis no processo de filtragem, o que pode ter resultado na perda de informações relevantes para a formação dos clusters.

O segundo modelo simplificou o pipeline ao omitir etapas intermediárias, como a análise de correlações e de multiclinearidade, focando diretamente na aplicação do PCA após o pré-processamento inicial. Apesar de apresentar um desempenho melhor ao ser comparado ao primeiro modelo, o desempenho também não foi suficientemente satisfatório, indicando que, mesmo com menos etapas, a redução linear via PCA pode não ter conseguido capturar de forma eficiente as relações dos dados para separar as atividades humanas.

O terceiro modelo utilizou o UMAP para a redução da dimensionalidade, o que permitiu capturar relações não lineares nos dados de forma mais eficiente. Este modelo apresentou os melhores resultados com métricas significativas. Um Coeficiente de Silhouette igual a 0.843, indicando uma boa separação entre os clusters formados; índices de homogeneidade de 0.583 e de completude de 0.999, sugerindo que, embora os clusters não sejam perfeitamente homogêneos, os dados reais podem estar bem representados pelos clusters formados.

A escolha do UMAP no terceiro modelo foi determinante para o desempenho superior, destacando a importância de métodos de redução de dimensionalidade que preservam relações não lineares. Comparado ao PCA, que é limitado à variância linear, o UMAP conseguiu representar melhor os padrões complexos dos dados.

Por outro lado, as etapas de filtragem do primeiro modelo, apesar de relevantes em termos de análise exploratória, podem ter causado perda de informações importantes. Essa perda pode ter impactado negativamente na formação dos clusters, mostrando ser crucial o balanceamento entre simplificação e retenção de informações.

Apesar dos bons resultados apresentados pelo terceiro modelo, vale considerar algumas limitações, como por exemplo: a capacidade do algoritmo K-Means de assumir clusters esféricos, o que pode não corresponder à natureza dos dados; e a influência do parâmetro de *n_neighbors* do UMAP, que pode influenciar nos resultados.

Os resultados obtidos demonstram que a escolha de técnicas de pré-processamento e redução de dimensionalidade tem impacto significativo no desempenho de modelos de clustering. Os resultados obtidos com o terceiro modelo destacaram a importância de métodos que preservem relações complexas em dados de alta dimensionalidade.

Conclusão e Trabalhos Futuros

Esse projeto demonstrou a viabilidade e os desafios do uso do algoritmo K-Means para o reconhecimento de atividades humanas com base em dados de sensores. Os resultados obtidos mostraram que, com o pré-processamento e com a escolha de técnicas apropriadas, é possível identificar padrões relevantes e formar clusters representativos das atividades.

A identificação e o tratamento dos outliers, além da padronização dos dados, foram passos cruciais para garantir a robustez do modelo. Além disso, a comparação entre PCA e UMAP destacou a superioridade de abordagens não lineares para dados complexos e de alta dimensionalidade. E, por último, uma avaliação completa de métricas como homogeneidade, completude e coeficiente de Silhouette, combinadas com a matriz de contingência, fornecem insights valiosos sobre a qualidade dos clusters formados.

Todas as decisões tomadas durante o projeto impactam diretamente no desempenho do modelo. Isso reforça a necessidade de fundamentar e basear todas as escolhas em evidências e não fazer escolhas arbitrárias.

O modelo desenvolvido apresentou bons resultados em termos de desempenho geral, mas também relevou algumas limitações do K-Means, como a

suposição de clusters esféricos. Ele oferece uma ferramenta valiosa para prever a taxa de engajamento no Instagram.

Para melhorar ainda mais os resultados e superar as limitações identificadas, algumas sugestões poderiam ser exploradas em trabalhos futuros, como a exploração de novos algoritmos de clustering, como *DBSCAN*, que podem lidar melhor com clusters de formas arbitrárias e oferecer mais flexibilidade; o ajuste fino de hiperparâmetros como *o n_neighbors*, que podem influenciar diretamente os resultados; a incorporação de informações temporais, que poderiam melhorar a identificação de padrões e transições entre as atividades; ou ainda a integração entre técnicas, adotando uma abordagem semisupervisionada que poderia testar a força do K-Means com classificadores supervisionados.