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

Residentes: Yuri Oliveira dos Santos e Rafael Santos Souza

Data de entrega: 03/12/2024

1.Resumo

Machine learning não supervisionada é uma maneira computacional de identificar padrões e agrupar dados sem a necessidade de variáveis dependentes ou rótulos pré-definidos, permitindo descobrir estruturas ocultas ou relações intrínsecas nos dados.

Neste projeto desenvolvemos um modelo preditivo não supervisionado utilizando o algoritmo K-means para resolver um problema de inferência na classificação de diferentes ADLs (Activities of Daily Living) que foram registrados em um dataset de domínio público.

Durante o processo desse trabalho foi utilizado técnicas de normalização, padronização e engenharia de características para melhorar a performance do modelo, no decorrer do trabalho é abordado a análise das métricas do modelo como silhouet score, e de hiperparâmetros do modelo.

2.Introdução

O Reconhecimento de Atividade Humana (HAR) tem como objetivo identificar as ações realizadas por uma pessoa com base em um conjunto de observações dela mesma e do ambiente ao seu redor. O reconhecimento pode ser realizado explorando as informações obtidas de várias fontes, como sensores ambientais ou sensores vestíveis no corpo. Este dataset compõe os dados tratados dos sensores de um Samsung galaxy S II colocados na cintura das pessoas que realizaram o experimento. As seis ADLs selecionadas foram: “ficar em pé, sentar, deitar, andar, descer escadas e subir escadas” (ESANN, 2013).

Os smartphones estão abrindo novas oportunidades de pesquisa para aplicações centradas no ser humano, onde o usuário se torna uma fonte rica de informações contextuais, e o dispositivo atua como uma ferramenta de sensoriamento direto. Os modelos mais recentes vêm equipados com sensores embutidos, como microfones, câmeras duplas, acelerômetros, giroscópios, entre outros. O uso de smartphones surge como uma solução alternativa para o Reconhecimento de Atividade Humana (HAR).

A coleta de dados para o Reconhecimento de Atividade Humana (HAR) por meio de smartphones enfrenta desafios significativos, como a variabilidade das condições ambientais, diferenças entre dispositivos e a necessidade de rotulagem manual, que pode ser demorada e suscetível a erros. Nesse contexto, o uso de *machine learning* não supervisionada se mostra ideal, pois permite a identificação de padrões e a criação de agrupamentos de atividades sem a necessidade de rótulos pré-definidos. Isso facilita a análise de grandes volumes de dados brutos, reduzindo a dependência de intervenção humana e tornando o processo mais eficiente e escalável.

As leituras dos sinais de sensores foram pré-processadas por filtro de ruídos, criando amostras em janelas de 2.56 segundos e sobrepostas em 50% (128 leituras por janela).
O dataset trás 17 leituras de sensores, sendo 8 deles triaxiais (contendo uma leitura por eixo), sendo no total 33 variáveis a serem consideradas por 17 métricas, conforme a tabela 1, diferentes gerando as 561 variáveis aleatórias encontradas no dataset.

Tabela 1: Funções e descrições

Function	Description
mean	Mean value
std	Standard deviation
mad	Median absolute value
max	Largest values in array
min	Smallest value in array
sma	Signal magnitude area
energy	Average sum of the squares
iqr	Interquartile range
entropy	Signal Entropy
arCoeff	Autorregresion coefficients
correlation	Correlation coefficient
maxFreqInd	Largest frequency component
meanFreq	Frequency signal weighted average
skewness	Frequency signal Skewness
kurtosis	Frequency signal Kurtosis
energyBand	Energy of a frequency interval
angle	Angle between two vectors

Fonte: (ESANN, 2013).

O uso do algoritmo K-means para tarefas de Reconhecimento de Atividade Humana (HAR) é vantajoso devido à sua simplicidade e eficiência na identificação de padrões em grandes volumes de dados. O K-means agrupa os dados com base em suas características, permitindo diferenciar atividades como caminhar, correr ou sentar, mesmo sem rótulos pré-definidos. Além disso, sua capacidade de lidar com dados de sensores contínuos, como acelerômetros e giroscópios, facilita a segmentação das atividades em clusters distintos. Essa

abordagem reduz a necessidade de rotulagem manual e melhora a generalização em cenários onde as atividades humanas apresentam variações naturais, como em diferentes ambientes ou entre usuários.

Ao implementar um algoritmo não supervisionado a partir de um dataset de domínio público, estimula se a análise crítica de dados e a prática das técnicas de tratamentos estudadas no curso de residência de software RESTIC 36. Essa abordagem de projeto favorece o exercício de habilidades desejáveis pro curso como inovação e criatividade, não mensuráveis em avaliações teóricas fechadas.

> "Um projeto gera situações problemáticas, ao mesmo tempo, reais e diversificadas. Possibilita, assim, que os educandos, ao decidirem, opinarem, debaterem, construam sua autonomia e seu compromisso com o social". (Lúcia, 1996)

3.Metodologia

A análise exploratória iniciou com uma visualização geral da estatística descritiva do dataset, com o objetivo de avaliar a escala dos valores e identificar contextos superficiais.
o método .describe() do pandas permitiu uma confirmação geral do estado dos dados no dataset, como a normalização feita nos dados de leitura no intervalo de -1 à 1 apresentadas como mínimo e máximo (ESANN, 2013). Foi procurado por valores nulos no dataset.
Com a impressão do gráfico boxplot foi possível verificar a distribuição dos outliers de cada coluna onde foi impresso no gráfico o percentual de dados outliers em comparação ao total.
Essa etapa investigativa apesar do dataset vir com arquivos descritivos sobre a coleta e pré processamento dos dados se mostrou importante para entendimento e consolidações de inferências sobre o contexto dos valores encontrados no dataset. Segundo Aggarwal (2016), "a análise exploratória de dados é uma etapa crítica para entender as características subjacentes do conjunto de dados, especialmente em aplicações complexas como reconhecimento de atividades humanas" (p. 123).

Os percentuais de outliers em algumas colunas se mostraram significativos o que levantou a hipótese de que não seriam outliers mas valores válidos e representativos da leitura dos sensores, em reunião com a tutora Luciene Torquato entendemos que poderíamos usar algumas técnicas em que poderia ser feito o tratamento desses outliers sem "agredir" o comportamento dados por estes dados. Assim sendo utilizamos a Winsorização para compactar os dados que estão fora dos 5% dos limiares de quantis de cada coluna.

Outra hipótese levantada foi a de melhoria dos resultados juntando os dados dos dois arquivos de treino e teste mas não se comprovou com a aplicação do PCA, seguindo orientações optamos pelo uso apenas do dataset de treino ignorando os 30% dos resultados dos experimentos que foi particionado no teste.

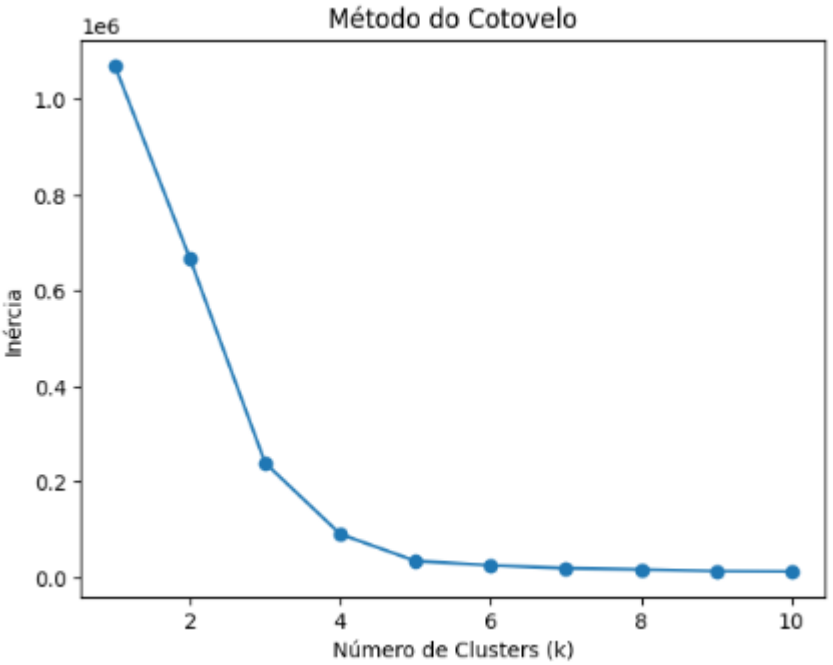
O k-means usa distâncias euclidianas, então é importante que todas as variáveis tenham a mesma escala. Para isso, usamos o método RobustScaler() da biblioteca scikit-learn em Python para padronizar, ele remove a mediana e escala os dados de acordo com o intervalo interquartil (IQR), ou seja, a diferença entre o primeiro quartil (Q1, 25%) e o terceiro quartil (Q3, 75%).

Para lidar com valores extremos no conjunto de dados, foi utilizada a técnica de winsorização, que consiste em limitar os valores outliers substituindo-os pelos valores dos percentis definidos. Neste trabalho, foram ajustados os valores inferiores ao 5º percentil e superiores ao 95º percentil, garantindo que as análises fossem menos suscetíveis a distorções causadas por outliers. Essa abordagem preserva a estrutura geral dos dados, reduzindo o impacto de valores atípicos sem eliminar informações relevantes.

Neste trabalho, a análise de componentes principais (PCA) foi empregada para reduzir a dimensionalidade do conjunto de dados, composto por 561 variáveis. Dada a alta dimensionalidade, havia o risco de multicolinearidade e redundância entre as variáveis, o que poderia dificultar a interpretação dos modelos e aumentar o custo computacional. A PCA foi utilizada para projetar as variáveis originais em um novo espaço de menor dimensão, preservando a maior parte da variabilidade dos dados. Optou-se por reter componentes que explicassem pelo menos 95% da variância total, garantindo que as informações mais relevantes fossem mantidas. A implementação foi realizada utilizando a biblioteca scikit-learn em Python, permitindo uma integração eficiente com as etapas subsequentes de modelagem e análise.

Para avaliar a qualidade e a eficiência dos agrupamentos gerados, foram aplicadas diversas métricas e técnicas de validação. O método do cotovelo foi utilizado para identificar o número ideal de clusters, analisando a relação entre o número de clusters e a inércia, que representa a soma das distâncias quadradas entre os pontos e seus respectivos centróides, como mostrado na figura 1. O ponto em que a redução da inércia começa a ser marginal (convergir) indicou o número adequado de clusters, garantindo um equilíbrio entre simplicidade e precisão.

Figura 1: Método do Cotovelo



Fonte: Autor

Adicionalmente, o Silhouette Score foi aplicado para medir a coesão e a separação dos clusters, avaliando a proximidade dos pontos em relação ao cluster ao qual pertencem e aos clusters vizinhos. Com valores variando de -1 a 1, um Silhouette Score próximo a 1 indica uma boa formação dos clusters.

O dataset já trás um arquivo com as etiquetas geradas manualmente (ESANN, 2013) mas a quantidade de amostras (linhas) entre a tabela com etiquetas e o arquivo com os vetores de características não tem o mesmo valor, por isso optamos por gerar as etiquetas a partir do k-means e utilizar a classificação gerada como etiqueta real.

Para complementar a análise, o Rand Index foi utilizado para comparar os agrupamentos obtidos com uma classificação de referência, mensurando a proporção de acertos na atribuição dos pontos aos clusters corretos.

Uma outra métrica utilizada para comparação do resultado obtido a partir de etiquetas foi Índice Calinski-Harabasz. Segundo Caliński e Harabasz (1974), "valores elevados do índice indicam uma boa separação entre os clusters e uma baixa dispersão dentro dos mesmos, sendo ideal para determinar a qualidade dos agrupamentos em algoritmos de clustering".

O Índice avalia a qualidade dos agrupamentos ao medir a dispersão entre clusters em comparação à dispersão dentro dos clusters. O valor obtido foi de 67.493,07, considerado na

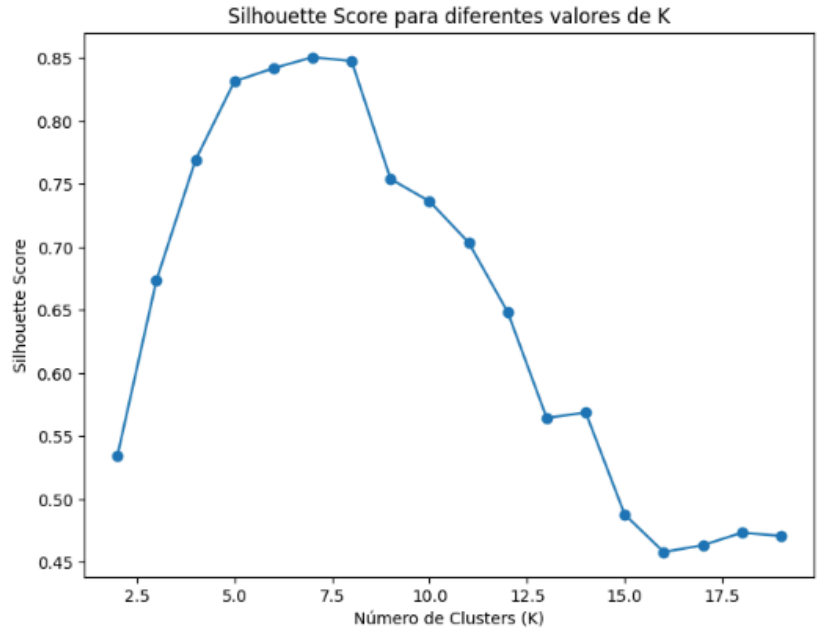
literatura, um valor elevado indica uma separação clara entre os clusters (alta dispersão intercluster) e baixa variabilidade dentro de cada cluster (baixa dispersão intracluster), o que demonstra a homogeneidade dos dados em cada grupo. Esse resultado reforça a escolha de K=7, evidenciando que o modelo é eficiente em capturar a estrutura subjacente dos dados.

4.Resultados

Os resultados obtidos para o modelo K-Means indicam uma configuração de alto desempenho no agrupamento dos dados, considerando as análises avaliadas. A seguir, detalhamos cada uma dessas considerações e seus significados relevantes:

O número K=7 foi definido como ideal com base no Silhouette Score conforme na figura 2, refletindo a estrutura natural dos dados. Essa escolha demonstra que a divisão em sete agrupamentos fornece o melhor equilíbrio entre a coesão interna e a separação entre os clusters. Na figura 3 mostra que essa configuração é indicativa de uma estrutura subjacente clara nos dados, o que é fundamental para análises interpretativas e decisões estabelecidas nos agrupamentos.

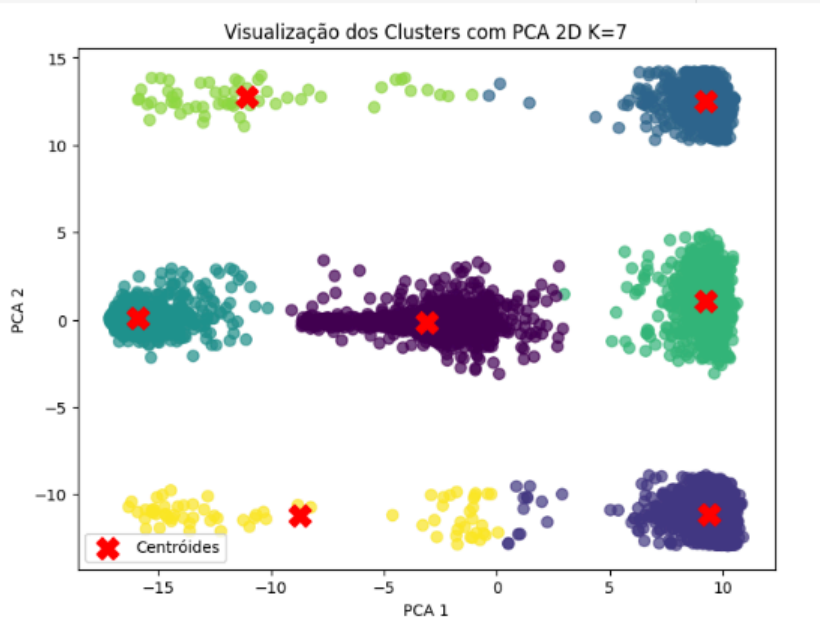
Figura 2 : Silhouette Score para diferentes valores de K




```
Número ideal de clusters com base no Silhouette Score: K=7
```

Fonte: Autor

Figura 3:Visualização dos Clusters com PCA 2D K=7



Fonte: Autor

O Silhouette Score é uma métrica de avaliação que varia de -1 a 1 e mede a qualidade do agrupamento considerando:

- Coesão interna: A proximidade dos pontos dentro de cada cluster é um aspecto fundamental na avaliação de agrupamentos. Clusters altamente coesos indicam que os pontos estão bem agrupados em torno de seus centróides, apresentando baixa dispersão interna. Isso significa que os elementos dentro do cluster são semelhantes entre eles, o que resulta num agrupamento bem definido.
- Separação intercluster: A distância entre clusters diferentes. Com um valor de 0,8506, o modelo apresentado.

Possui boa separação : Existe uma clara distinção entre os clusters, diminuindo que há pouca ou nenhuma sobreposição entre eles. Este resultado evidencia que os agrupamentos gerados possuem uma qualidade excepcional, reforçando a adequação do modelo para segmentação de dados.

A inércia foi definida como a soma das distâncias quadradas entre os pontos e os centroides de seus respectivos clusters, é um indicativo de compacidade.

Um valor de 19081.31, apesar de ser relativo ao volume de dados, mostra que os clusters formados são compactos. A redução da inércia com o aumento de K até 7 demonstra

que a escolha deste valor como número ideal de clusters otimiza a relação entre simplicidade do modelo e precisão dos agrupamentos.

O Índice Calinski-Harabasz , ou índice de razão de variância, mede a dispersão entre clusters em relação à dispersão interna. Um valor elevado como 67493,07 indica:

- Dispersão intercluster significa que os clusters estão bem separados.
- Baixa variabilidade intracluster : Os dados dentro de cada cluster são homogêneos. Esse índice reforça a escolha de K = 7, evidenciando que o modelo é capaz de capturar as estruturas dos dados.

5.Discussão

Algumas das escolhas de análise e tratamento do dataset usado assim como metodologias aplicadas foram resumidas ao escopo de tempo e objetivo deste trabalho, sendo assim alguns desses caminhos não tomados poderiam produzir diferentes resultados, abaixo citamos alguns desses.

O dataset de domínio público usado neste trabalho contém um arquivo com cada variável calculada para vetor de características, esta tabela não foi explorada em mais detalhes neste trabalho mas tem o potencial de trazer maior contexto e melhor modelagem ao dataset de treino usado.

Outro arquivo não explorado são os sinais RAW (crus) dos sensores inerciais também disponíveis no dataset.

Entendemos que as características das variáveis disponíveis podem afetar diretamente no resultado do PCA que consecutivamente interfere no resultado do modelo do K-means, logo é de extrema relevância o tratamento dado as features e características como outliers, neste trabalho optamos por usar o método de Winsorização, mas existem outras técnicas que poderiam ser aplicadas trazendo diferentes resultados, o próprio fato de usar anterior ao PCA ou não é uma opção não explorada neste trabalho.

Utilizamos uma técnica para padronizar os dados do dataset antes de aplicar o K-means, mas deve se ressaltar que o dataset utilizado já tem os dados escalados no intervalo de -1 à 1, a diferença entre padronizar novamente este intervalo não é explorado neste trabalho.

O dataset já trás um arquivo com as etiquetas geradas manualmente (ESANN, 2013) mas a quantidade de amostras (linhas) entre a tabela com etiquetas e o arquivo com os vetores

de características não tem o mesmo valor, por isso optamos por gerar as etiquetas a partir do k-means e utilizar a classificação gerada como etiqueta real. A compreensão do contexto desta tabela de etiquetas e o tratamento necessário para utilizar estes dados em métricas como rand index e Índice Calinski-Harabasz não é explorado neste trabalho.

6.Conclusão

Os resultados obtidos confirmam qualidade satisfatória do modelo K-Means com K=7, evidenciada pelas notas de avaliação. O Silhouette Score , com valor de 0,8506, demonstra que os agrupamentos são altamente coesos e bem separados, estabelecendo uma clara distinção entre os clusters. A inércia , com valor de 19081,31, reflete a compactação dos dados dentro de cada agrupamento, reforçando a homogeneidade interna dos clusters. Por sua vez, o Índice Calinski-Harabasz , com valor de 67493,07, destaca a nitidez das fronteiras entre os clusters, demonstrando a eficácia do modelo na separação dos grupos. Esses resultados reforçam a robustez do modelo de segmentação de dados e sua adequação para aplicações que exigem uma análise satisfatória de padrões e comportamentos. No entretanto, o modelo não atingiu o resultado ótimo obtido no paper de referência adotando Support Vector Machine (SVM) (ESANN, 2013).

Essa configuração é ideal para aplicações em segmentações, como análises de dados para o Reconhecimento de Atividade Humana (HAR) por meio de smartphones, que enfrentam desafios significativos. Os resultados obtidos fornecem uma base sólida para gerar insights estratégicos e orientar ações fundamentadas nos agrupamentos identificados.

Esta prática fornece um roteiro de trabalho enriquecedor na área de machine learning não supervisionada, pois exemplifica técnicas e insights para reprodutibilidade e falseabilidade. Além disso Fornece ao leitor valores referência para métricas e hiperparâmetros em trabalhos dentro do mesmo contexto.

7.Referências

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013,

i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*, *3*(1), 1-27. https://doi.org/10.1080/03610927408827101

LEITE, Lúcia Helena Alvarez, Pedagogia de Projetos: intervenção no presente. Presença Pedagógica, Belo Horizonte: Dimensão, 1996. pp. 24-33.

Aggarwal, C. C. (2016). *Data mining: The textbook*. Springer.