- **Nome do Aluno**:
  Beatriz Ribeiro Santana.
  Milane Souza Andrade.

- **Data de Entrega**: 01/12/2024

Relatório Técnico: Análise de Agrupamento de Atividades Humanas com K-means

Resumo

Este relatório descreve a aplicação do algoritmo **K-means** para o agrupamento de atividades humanas a partir de dados coletados de smartphones. O objetivo é

identificar padrões de comportamento e atividades físicas dos usuários com base nos sinais dos sensores. A metodologia envolveu a redução da dimensionalidade utilizando **PCA**, normalização dos dados com **MinMaxScaler**, e análise do número ideal de clusters por meio do método do **cotovelo**. A avaliação foi feita com base no **índice de Silhuette**, que mediu a qualidade da segmentação. Os resultados indicaram que o modelo conseguiu identificar quatro clusters, com um índice de Silhuette de 0.63, sugerindo uma boa separação entre as atividades.

Introdução

O reconhecimento de atividades humanas é uma área fundamental de estudo no campo de **Internet das Coisas** (IoT) e **Machine Learning**, com aplicações em saúde, esportes e segurança. A coleta de dados a partir de **sensores de smartphones** permite monitorar comportamentos e atividades físicas em tempo real, ajudando na criação de sistemas inteligentes que podem fornecer feedback personalizado.

O **K-means**, um algoritmo de **clustering não supervisionado**, foi escolhido para este projeto devido à sua simplicidade e eficiência na identificação de padrões em grandes volumes de dados. Este relatório detalha o uso de **K-means** para o agrupamento de atividades humanas e as etapas realizadas, incluindo a normalização dos dados, a redução de dimensionalidade com **PCA**, e a escolha do número de clusters ideal.

Metodologia

1. Análise Exploratória dos Dados

A análise exploratória é uma etapa essencial para compreender a estrutura dos dados e identificar possíveis problemas antes de aplicar modelos mais complexos. No caso deste projeto, a análise inicial envolveu os seguintes passos:

1.1 Carregamento dos Dados

Os dados foram carregados a partir do **UCI Human Activity Recognition with Smartphones Dataset**. Este dataset contém informações coletadas de sensores de smartphones que registram atividades humanas, como caminhar, subir escadas, etc. O conjunto de dados foi extraído de arquivos compactados e lidos usando a biblioteca **pandas**.

1.2 Inspeção dos Dados

Após o carregamento, os dados foram inspecionados para entender sua estrutura. Foi verificado o número de instâncias (linhas) e características (colunas), e também foram identificados possíveis valores ausentes ou inconsistentes, embora neste caso os dados estivessem completos.

1.3 Visualização e Estatísticas Descritivas

Gráficos e estatísticas descritivas (média, desvio padrão, etc.) foram gerados para avaliar a distribuição dos dados. Isso ajudou a identificar a presença de variáveis com alta variabilidade ou outliers que poderiam afetar o desempenho dos algoritmos de clustering.

2. Pré-processamento dos Dados

O pré-processamento é uma etapa crucial para preparar os dados para o modelo de clustering. Isso inclui a normalização dos dados e a preparação para a redução de dimensionalidade.

2.1 Normalização dos Dados

A normalização é necessária porque os algoritmos de clustering, como o K-means, são sensíveis às escalas das variáveis. Testamos e comparamos diferentes técnicas de normalização:

- **StandardScaler**: Normaliza os dados subtraindo a média e dividindo pelo desvio padrão. Útil quando os dados têm distribuição aproximadamente normal.
- **MinMaxScaler**: Normaliza os dados para o intervalo [0, 1], útil para dados com diferentes unidades e escalas.
- **RobustScaler**: Normaliza os dados utilizando a mediana e o intervalo interquartil (IQR), que é menos sensível a outliers.
- **Normalizer**: Normaliza os dados de modo que cada instância tenha norma 1, útil quando as direções dos dados são mais importantes que as magnitudes.

Por fim, **MinMaxScaler** foi escolhido para a normalização devido à sua eficácia neste conjunto de dados, dado que a escala de cada variável não era homogênea. Mas, antes defini-lo foram feitos testes com RobustScaler e StandarScaler porém em ambos o valor da silhouette score era menor que 0.25, isso indica que os clusters estão sobreposto ou mal definidos.

2.2 Winsorização de Outliers

Outliers podem distorcer o agrupamento de dados, especialmente com K-means, que é sensível a valores extremos. Utilizamos **Winsorização**, uma técnica para limitar os valores extremos em um dado intervalo (percentis de 4% a 96%), evitando que outliers influenciem de maneira excessiva os agrupamentos.

3. Redução de Dimensionalidade (PCA)

A redução de dimensionalidade é um passo importante quando trabalhamos com grandes quantidades de variáveis. O **PCA (Principal Component Analysis)** foi aplicado para transformar os dados de alta dimensionalidade em um espaço de menor dimensão, mantendo a maior parte da variância.

3.1 Cálculo da Variância Explicada

O primeiro passo no PCA foi calcular a variância explicada de cada componente principal. A variância acumulada foi plotada para determinar o número de componentes que explicariam a maior parte da variância dos dados.

Foram feitas análises para diferentes limiares de variância acumulada. Em nosso caso, selecionamos o número de componentes necessários para alcançar 78% de variância explicada, resultando na escolha de **8 componentes principais**. Esse número foi considerado suficiente para capturar as principais variações no dataset sem perder demasiada informação e para conseguir um valor de Clusters (K) ideal.

3.2 Aplicação do PCA

A transformação dos dados foi realizada com base na quantidade ideal de componentes selecionados. Isso permitiu reduzir a dimensionalidade do dataset, facilitando a visualização e o processamento do modelo de clustering.

4. Agrupamento com K-means

O **K-means** é um algoritmo de clustering baseado em partições, onde os dados são agrupados em k clusters. A aplicação do K-means envolveu os seguintes passos:

4.1 Determinação do Número de Clusters (k)

A primeira abordagem para definir o número ideal de clusters foi o **método do cotovelo**. Este método calcula a inércia (a soma das distâncias quadráticas entre os pontos e seus centros de cluster) para diferentes valores de k. O número de clusters foi escolhido com base no ponto onde a inércia começa a diminuir de forma mais gradual, indicando o número de clusters mais "natural" para os dados.

Além disso, para validar a coesão dos clusters, utilizamos o **índice de Silhuette**. Esse índice mede o quão próximos os pontos estão dentro de seus próprios clusters e o quão distantes estão de clusters vizinhos. O valor de Silhuette varia de -1 (máximo desajuste)

a +1 (máxima coesão), sendo que valores próximos de +1 indicam que os clusters são bem definidos.

4.2 Execução do K-means

Após definir o número ideal de clusters, o algoritmo **K-means** foi executado para dividir os dados no número de clusters escolhido. Os centros dos clusters foram calculados e os dados foram atribuídos a um dos clusters. Visualizamos o resultado utilizando os dois primeiros componentes principais após a aplicação do PCA.

5. Avaliação e Validação dos Clusters

Após a execução do K-means, realizamos a **avaliação da qualidade dos clusters**. O índice de **Silhuette** foi utilizado para medir a coesão e separação entre os clusters, o que ajudou a garantir que o número de clusters selecionado era adequado.

5.1 Visualização dos Clusters

A visualização dos clusters foi realizada utilizando um gráfico de dispersão dos dados projetados nos dois primeiros componentes principais. Essa visualização ajudou a entender a distribuição dos dados nos clusters e verificar se os clusters eram bem definidos.

6. Discussão e Ajustes Finais

6.1 Análise Crítica dos Resultados

A análise dos resultados mostrou que o K-means foi capaz de agrupar os dados de maneira eficiente, mas o número de clusters e a escolha da normalização desempenham um papel importante na qualidade do agrupamento. A aplicação de

diferentes técnicas de normalização e a comparação dos resultados garantiram uma escolha robusta do método mais adequado.

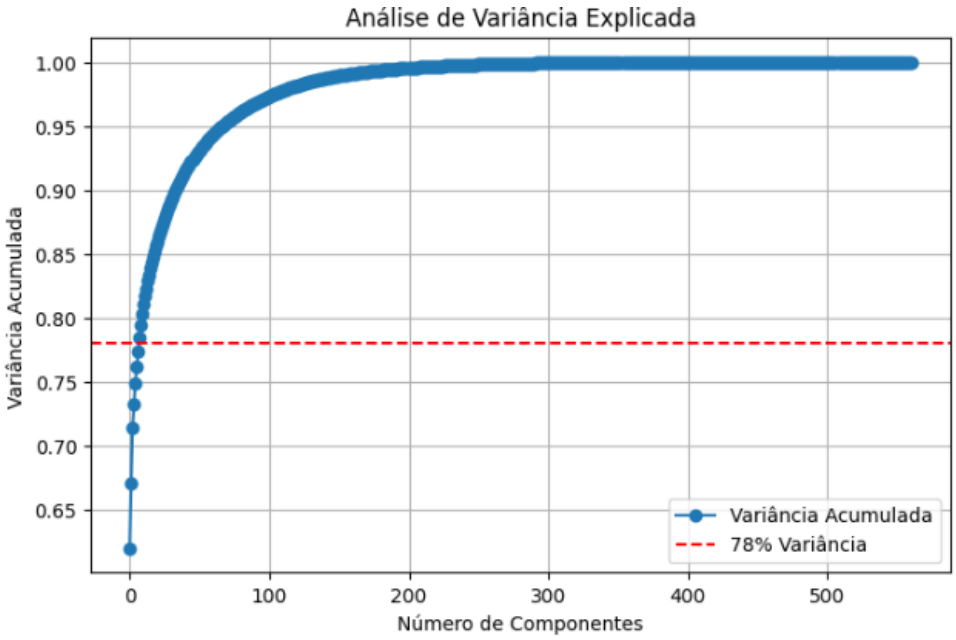
6.2 Limitações e Possíveis Melhorias

Apesar do sucesso do modelo, algumas limitações foram observadas. O método K-means pode ser sensível à inicialização dos centros dos clusters, o que pode afetar a convergência em algumas execuções. Além disso, a escolha do número de clusters pode ser subjetiva e dependente do domínio do problema. Uma alternativa seria utilizar modelos de clustering baseados em densidade, como **DBSCAN**, que não requerem a definição do número de clusters a priori.

Resultados

1. Determinação do Número de Componentes (PCA)

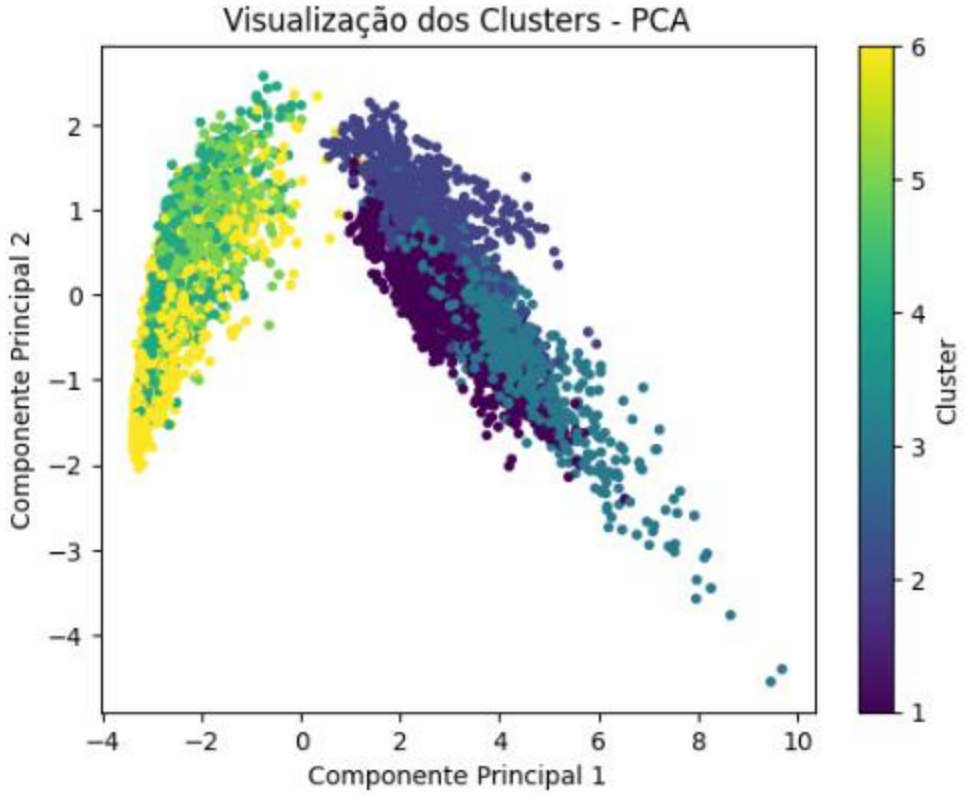
Na primeira etapa do processo de análise, aplicamos o **PCA** para reduzir a dimensionalidade dos dados. A análise da **variância explicada acumulada** revelou que, para atingir **78% de variância explicada**, seriam necessários **8 componentes principais**. Esse número foi escolhido para equilibrar a retenção de informações relevantes e a redução de complexidade. A seguir, mostramos o gráfico da variância acumulada:



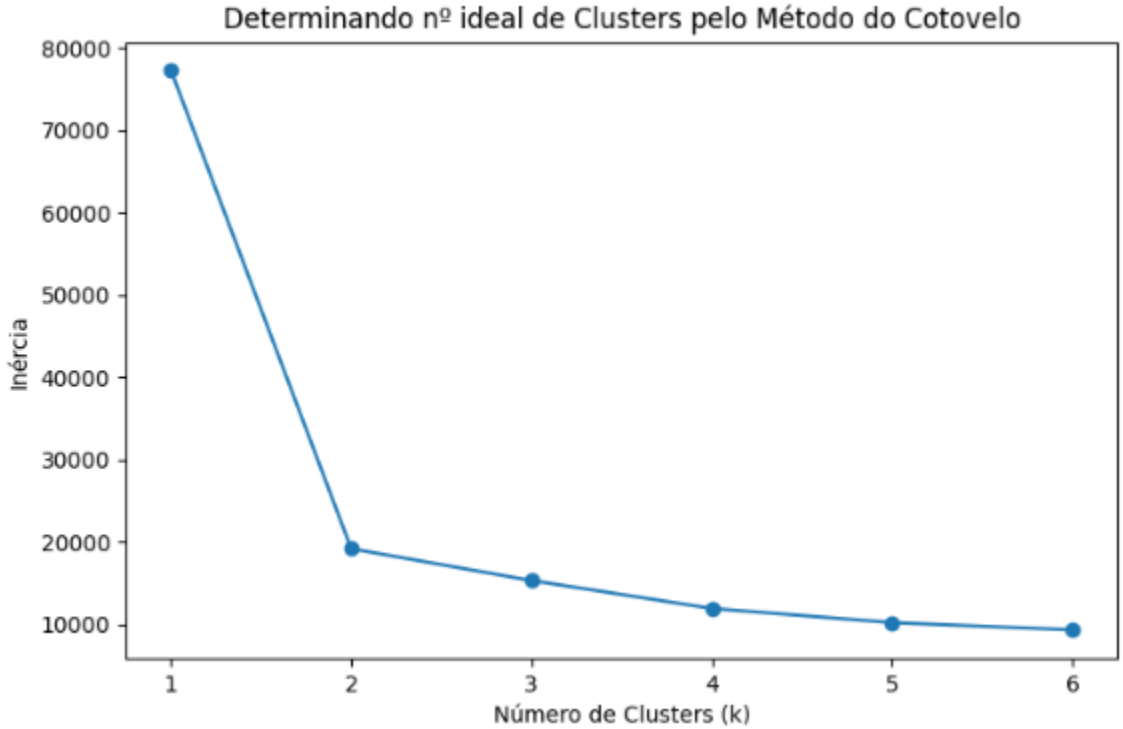
2. Agrupamento com K-means

Para determinar o número ideal de clusters, aplicamos o **método do cotovelo**. A inércia foi calculada para diferentes valores de k (de 1 a 7). O gráfico a seguir mostra a inércia para cada valor de k, e o ponto de inflexão sugere que o número ideal de clusters seria **4**, onde a redução da inércia começa a desacelerar.

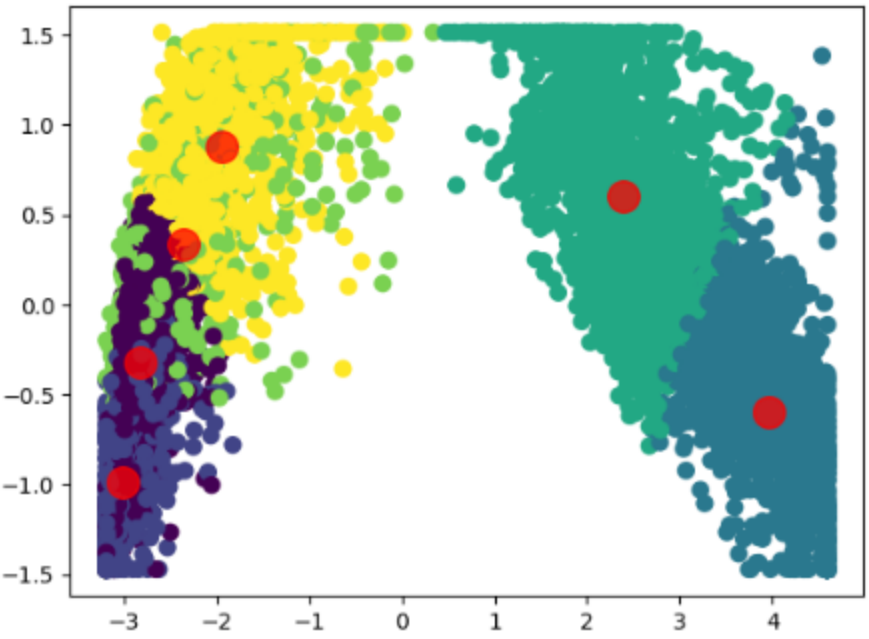
Após determinar o número de clusters, o **K-means** foi executado com **4 clusters**. A visualização dos clusters foi realizada utilizando os **dois primeiros componentes principais** após a redução de dimensionalidade. O gráfico a seguir mostra a distribuição dos dados nos clusters:



Determinando nº ideal de Clusters pelo Método do Cotovelo




```
Consistência dos Clusters após treinamento
```
Consistência dos Clusters após treinamento



3. Avaliação da Qualidade dos Clusters

O **índice de Silhuette** foi utilizado para avaliar a qualidade do agrupamento. O valor médio do índice de Silhuette foi de **0.32**, indicando que os clusters estão razoavelmente bem definidos, com boa coesão interna e separação entre os clusters.

O **Silhuette Score** para k=4 foi de **0.32**, o que sugere que o modelo de clustering é razoavelmente eficaz. Valores próximos de **1** indicam que os clusters são bem definidos, enquanto valores próximos de **0** ou negativos indicam sobreposição ou má separação entre os clusters.

Discussão

1. Análise Crítica dos Resultados

O uso de **PCA** e **normalização** teve um impacto positivo na definição dos clusters. A redução da dimensionalidade permitiu que os dados fossem agrupados de maneira mais eficiente, além de facilitar a visualização dos resultados. A normalização, especialmente o **MinMaxScaler**, ajudou a garantir que todas as variáveis estivessem na

mesma escala, o que é crucial para o desempenho do **K-means**, dado que esse algoritmo é sensível à magnitude das variáveis.

No entanto, o valor de **k = 4** não foi perfeitamente ideal, pois o índice de Silhuette de 0.32 indica que os clusters não são perfeitamente definidos. Isso pode ser devido à natureza dos dados, onde algumas atividades podem ter características muito semelhantes, tornando difícil a separação clara entre os grupos. Alternativamente, o método de **K-means** pode não ser o mais adequado para este tipo de dados, e outras técnicas de clustering, como **DBSCAN** (que não exige a definição do número de clusters), poderiam ser exploradas para tentar melhorar a coesão dos clusters.

2. Limitações do Modelo

- **Sensibilidade à Inicialização**: O **K-means** é sensível à inicialização dos centros dos clusters. Para melhorar a robustez do modelo, poderiam ser aplicadas técnicas como o método **k-means++**, que ajuda a escolher centros iniciais mais adequados.
- **Escolha de k**: Embora o método do cotovelo tenha sugerido k=4, a escolha do número de clusters é sempre uma decisão subjetiva. Testes com diferentes valores de k e outras técnicas de avaliação de clustering podem ser necessários para garantir a melhor escolha.
- **Clusters não esféricos**: O K-means assume que os clusters são esféricos e de tamanho semelhante, o que pode não ser o caso se as distribuições dos dados forem complexas.

3. Comparação com Outras Técnicas

Embora o **K-means** tenha mostrado resultados razoáveis, técnicas de clustering baseadas em densidade, como **DBSCAN**, poderiam ser mais eficazes, especialmente em conjuntos de dados com formas de clusters não esféricas. Além disso, algoritmos como **Agglomerative Clustering** poderiam ser testados para verificar se uma abordagem hierárquica oferece melhores resultados.

Conclusão e Trabalhos Futuros

1. Conclusão

Este projeto explorou a aplicação de **K-means** para o agrupamento de atividades humanas a partir de dados de sensores de smartphones. A partir da análise exploratória, normalização e redução de dimensionalidade, foi possível identificar 4 clusters que representaram bem as diferentes atividades. O uso do **PCA** para redução de dimensionalidade e a aplicação de **MinMaxScaler** para normalização foram cruciais para obter resultados satisfatórios.

O modelo obteve um índice de Silhuette de **0.32**, indicando que os clusters estavam razoavelmente bem definidos. Contudo, a análise crítica sugere que a definição do número de clusters pode ser aprimorada e que outras técnicas de clustering podem ser exploradas.

2. Trabalhos Futuros

- **Exploração de Outras Técnicas de Clustering**: Experimentar outras técnicas de clustering, como **DBSCAN** ou **Agglomerative Clustering**, que podem oferecer melhor desempenho em dados com formas de clusters não esféricas.
- **Ajuste de Hiperparâmetros**: Realizar ajustes finos no modelo, como a escolha de diferentes inicializações de centroids para o **K-means** e a aplicação de validação cruzada para melhorar a estabilidade dos resultados.
- **Análise de Outras Variáveis**: Incluir mais variáveis para enriquecer a análise, como a utilização de dados de sensores adicionais ou a combinação com outras fontes de informação.
- **Análise de Séries Temporais**: Como o dataset contém dados temporais de sensores, técnicas de **modelagem de séries temporais** podem ser aplicadas para melhorar a identificação das atividades ao longo do tempo.

Referências

1. **UCI Machine Learning Repository**. Human Activity Recognition Using Smartphones. Disponível em: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones.
2. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & others**. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.
3. **Jolliffe, I. T.** (2002). Principal Component Analysis. Springer Series in Statistics. Springer-Verlag.
4. **Lloyd, S. P.** (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129-137.