RESIDÊNCIA EM SOFTWARE

CIÊNCIA DE DADOS

MARLEY REBOUÇAS FERREIRA

MURILO CARLOS NOVAIS

RELATÓRIO TÉCNICO: RECONHECIMENTO DE ATIVIDADES HUMANAS USANDO K-MEANS

Ilhéus-BA

3 de dezembro de 2024

Resumo

O objetivo deste projeto foi aplicar o algoritmo de K-means para agrupar atividades humanas a partir de dados de sensores de smartphones, especificamente os dados coletados pelo **dataset** *"Human Activity Recognition Using Smartphones"*. O dataset contém medições de 561 variáveis, que foram extraídas de sinais de acelerômetro e giroscópio de 30 participantes realizando atividades diárias, como caminhar, subir escadas, ficar em pé e sentar.

A metodologia adotada foi dividida em várias etapas: primeiro, foi realizada uma **análise exploratória dos dados (*EDA*)**, onde foram investigadas as distribuições das variáveis e as correlações entre elas. Em seguida, os dados foram **normalizados** utilizando o método ***StandardScaler***, garantindo que todas as variáveis tivessem a mesma escala para evitar viés no agrupamento.

A redução de dimensionalidade foi feita com a técnica ***PCA* (Análise de Componentes Principais)**, que permitiu representar os dados em duas e três dimensões para facilitar a visualização e interpretação dos *clusters*. O número de *clusters* foi definido utilizando o **Método do Cotovelo** e o ***Silhouette Score***, ambos indicando que o número ideal de *clusters* era $K = 6$, representando as seis atividades humanas.

O algoritmo de **K-means** foi então aplicado, e os resultados foram avaliados visualmente com gráficos 2D e 3D, mostrando boa separação e coesão dos *clusters*. A principal conclusão do projeto foi que o K-means, aliado à redução de dimensionalidade via *PCA*, é eficaz para agrupar as atividades humanas, com boa qualidade de separação entre as atividades, conforme indicado pelas métricas de **inércia** e ***silhouette score***.

1 Introdução

O reconhecimento de atividades humanas é uma área de estudo no campo da inteligência artificial que busca identificar padrões de comportamento a partir de dados coletados por sensores, como acelerômetros, giroscópios e outros dispositivos de medição. Essa área tem se expandido significativamente devido ao crescente uso de **smartphones** e **dispositivos vestíveis** para monitoramento da saúde e bem-estar. O reconhecimento de atividades humanas (*HAR*, do inglês *Human Activity Recognition*) tem aplicações diversas, como em sistemas de saúde para monitoramento de pacientes, na interação com dispositivos inteligentes, e em atividades de análise de comportamento para otimização de rotinas.

O dataset utilizado neste projeto, ***"Human Activity Recognition Using Smartphones"***, contém dados coletados por acelerômetros e giroscópios de smartphones, com medições de **561 variáveis** obtidas enquanto 30 voluntários realizavam atividades cotidianas, como caminhar, subir escadas, sentar e deitar. Esses dados oferecem um cenário desafiador para o reconhecimento de padrões, já que as atividades podem ter características sobrepondo-se, como a diferença entre "ficar em pé" e "sentar", por exemplo.

Para resolver esse problema de agrupamento e identificação das atividades, foi escolhido o algoritmo **K-means**, uma técnica de **aprendizado não supervisionado** amplamente utilizada para a segmentação de dados em grupos ou *clusters*. O K-means é adequado para este tipo de problema por sua simplicidade, eficiência e capacidade de lidar bem com grandes volumes de dados, como os encontrados no dataset de atividades humanas. O método K-means tenta encontrar uma divisão dos dados que minimize a distância entre as amostras e os centros dos *clusters*, agrupando dados semelhantes em conjuntos, o que é ideal para identificar padrões de comportamento nas atividades humanas.

A escolha do K-means também se deve à sua escalabilidade e à sua facilidade de implementação. Apesar de ser uma técnica simples, ela tem mostrado bons resultados em uma variedade de tarefas de agrupamento, incluindo aquelas envolvendo dados de sensores. Além disso, o K-means é fácil de interpretar, o que facilita a análise e visualização dos resultados, tornando-o uma excelente escolha para este projeto de reconhecimento de atividades humanas.

2 Metodologia

A metodologia deste projeto foi dividida em várias etapas, desde a análise exploratória dos dados até a implementação do algoritmo de K-means e a avaliação dos resultados. A seguir, detalho cada uma dessas etapas:

2.1 Análise Exploratória de Dados (EDA)

A primeira etapa do projeto envolveu a **análise exploratória dos dados (EDA)**, que tem como objetivo entender as características do dataset, identificar possíveis padrões e detectar problemas como valores ausentes ou variáveis irrelevantes. Para isso, as seguintes etapas foram realizadas:

- **Estatísticas Descritivas:** Foram geradas estatísticas descritivas (média, desvio padrão, mínimos e máximos) para as variáveis numéricas. Isso permitiu entender a distribuição dos dados e identificar possíveis anomalias.
- **Matriz de Correlação:** Uma matriz de correlação foi gerada entre as variáveis para identificar relações lineares entre elas. Variáveis altamente correlacionadas podem ser redundantes e, nesse caso, a redução de dimensionalidade pode ser aplicada para simplificar a análise.
- **Visualizações:** Utilizamos gráficos como histogramas e *boxplots* para verificar a distribuição das variáveis e identificar a presença de *outliers*.

Essas etapas ajudaram a entender as variáveis mais relevantes e a preparar os dados para as próximas fases do projeto.

2.2 Pré-processamento dos Dados

Antes de aplicar o algoritmo de K-means, foi necessário **normalizar** os dados e realizar a **redução de dimensionalidade**. As atividades humanas podem ser representadas por variáveis com escalas diferentes, o que pode afetar o desempenho do K-means. Para contornar esse problema, as seguintes etapas foram aplicadas:

- **Normalização (Standardização):** A técnica ***StandardScaler*** foi utilizada para normalizar os dados, transformando as variáveis para uma média de 0 e desvio

padrão de 1. Isso garantiu que todas as variáveis contribuíssem igualmente para a formação dos *clusters*, evitando que variáveis com escalas maiores dominassem o processo de agrupamento.

- **Redução de Dimensionalidade (PCA):** Como o dataset contém 561 variáveis, a redução da dimensionalidade foi realizada utilizando a **Análise de Componentes Principais (PCA)**. O *PCA* permitiu reduzir as variáveis originais para duas e três componentes principais, mantendo a maior parte da variância dos dados. Isso facilitou a visualização dos *clusters* e acelerou o processo de agrupamento, uma vez que o K-means é sensível à dimensionalidade dos dados.

2.3 Implementação do Algoritmo de K-means

Com os dados pré-processados, foi realizada a implementação do **algoritmo de K-means**. O K-means é um algoritmo de agrupamento não supervisionado que tenta dividir os dados em K ***clusters*** com base na minimização da distância média entre os pontos de dados e seus centróides. A implementação seguiu os seguintes passos:

- **Inicialização dos Centróides:** Utilizou-se o método **K-means**++ para inicializar os centróides. Esse método ajuda a selecionar pontos iniciais distantes entre si, evitando que o algoritmo convirja para soluções subótimas.

- **Execução do Algoritmo:** O algoritmo foi executado com diferentes valores de K (número de *clusters*) para verificar qual valor produzia os melhores resultados. O K-means foi repetido várias vezes (`n_init=10`) para garantir a estabilidade do modelo e evitar que o algoritmo ficasse preso em mínimos locais.

- **Cálculo dos Clusters:** O algoritmo atribui cada ponto de dados a um dos K *clusters* com base na proximidade ao centróide mais próximo. O processo é repetido até que os centróides não se movam significativamente entre as iterações.

2.4 Escolha do Número de Clusters (K)

A escolha do número de *clusters* K é uma das partes mais importantes ao aplicar o K-means. Para determinar o valor ideal de K, foram utilizados dois métodos principais:

- **Método do Cotovelo (Elbow Method):** Este método envolve a plotagem da inércia (soma das distâncias quadradas dentro dos *clusters*) para diferentes valores de K. A ideia é escolher o valor de K onde a inércia começa a diminuir de forma menos pronunciada, o que indica que a adição de mais *clusters* não melhora significativamente a qualidade do agrupamento. O ponto de inflexão dessa curva é considerado o número ideal de *clusters*.

- **Silhouette Score:** O *silhouette score* é uma medida de coesão e separação dos *clusters*. Ele varia de -1 a 1, sendo que valores próximos de 1 indicam que os *clusters* são bem separados e coesos. O valor de K que maximiza o *silhouette score* foi considerado o melhor para o agrupamento.

Ambos os métodos sugeriram $K = 6$ como o número ideal de *clusters*, o que corresponde ao número de atividades presentes no dataset (como caminhar, subir escadas, etc.).

2.5 Avaliação dos Resultados

Para avaliar a qualidade dos *clusters* formados, foram utilizadas as seguintes métricas:

- **Inércia:** A inércia é a soma das distâncias quadráticas dos pontos aos seus centróides. Quanto menor a inércia, melhor é a coesão interna dos *clusters*. A inércia foi analisada durante o Método do Cotovelo.

- **Silhouette Score:** Como mencionado anteriormente, o *silhouette score* foi calculado para avaliar a separação e a coesão dos *clusters*. O valor final do *silhouette score* foi utilizado para validar a qualidade dos agrupamentos.

Além disso, os resultados foram visualizados por meio de **gráficos 2D e 3D**, utilizando os primeiros componentes principais do *PCA*. Esses gráficos permitiram uma análise visual clara da separação entre os *clusters*.

3 Resultados

Nesta seção, apresentamos as **métricas de avaliação** obtidas com a aplicação do algoritmo **K-means**, além da análise detalhada dos **gráficos** gerados para visualização

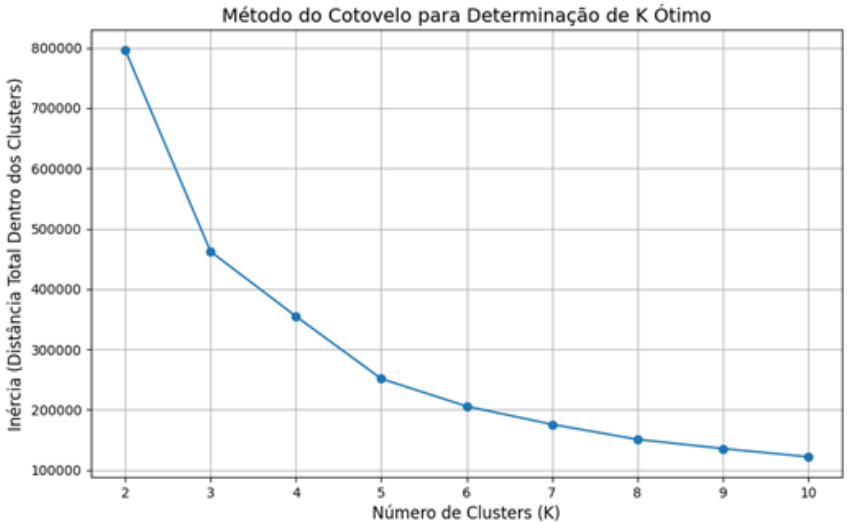
dos *clusters*. Vamos avaliar a **qualidade e coesão** dos *clusters*, levando em consideração tanto as métricas numéricas quanto as representações gráficas. A análise da **matriz de correlação** também é incluída, pois ela ajudou a entender as relações entre as variáveis e a orientar a redução de dimensionalidade.

3.1 Métricas de Avaliação

Inércia (Método do Cotovelo)

A **inércia** é uma medida da coesão dos *clusters*, calculada como a soma das distâncias quadradas entre cada ponto de dados e o centróide do seu respectivo *cluster*. Quanto menor a inércia, mais compactos são os *clusters*.

Figura 1. Grafico do método do cotovelo para determinação de K ótimo



Fonte: Autores.

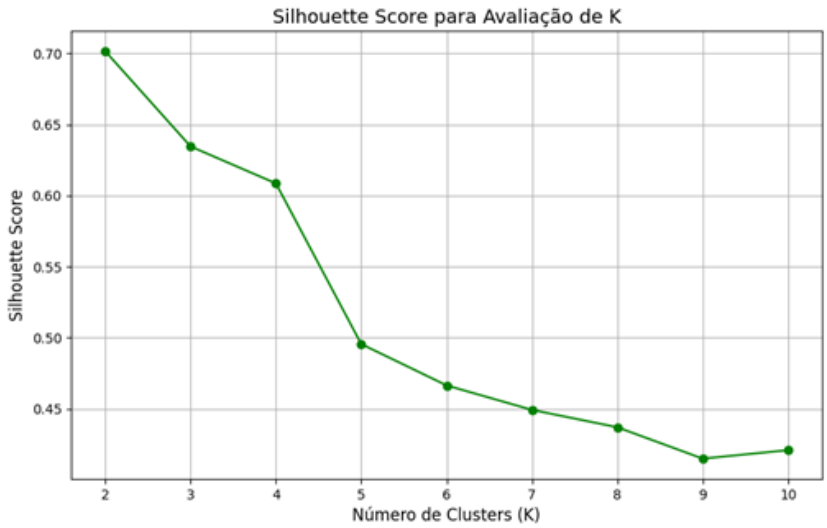
- **Observações:** O gráfico do **Método do Cotovelo** foi utilizado para determinar o número ideal de *clusters* (K). A inércia diminui rapidamente à medida que o número de *clusters* aumenta, mas essa diminuição se estabiliza a partir de $K = 6$, o que sugere que $K = 6$ é o número ideal de *clusters*. A adição de mais *clusters* não melhora significativamente a qualidade do agrupamento.

Silhouette Score

O *Silhouette Score* é uma métrica que avalia a coesão interna e a separação entre *clusters*. O valor do *silhouette score* varia de **-1 a 1**:

- 1 indica que os pontos estão bem agrupados e bem separados de outros *clusters*.
- 0 significa que os pontos estão na fronteira entre dois *clusters*.
- -1 indica que os pontos estão mal agrupados.

Figura 2. Gráfico Silhouett Score para avaliação de K



Fonte: Autores.

O ***Silhouette Score*** final foi calculado como 0.47 para $K = 6$. Esse valor sugere que, embora os *clusters* sejam razoavelmente coesos e bem separados, há algum grau de sobreposição entre eles. Um ***Silhouette Score*** de 0.47 é considerado moderado, indicando que o modelo fez um bom trabalho em separar as atividades, mas há espaço para melhorias, principalmente em atividades com características semelhantes, como "ficar em pé"e "sentar".

Gráfico do Silhouette Score

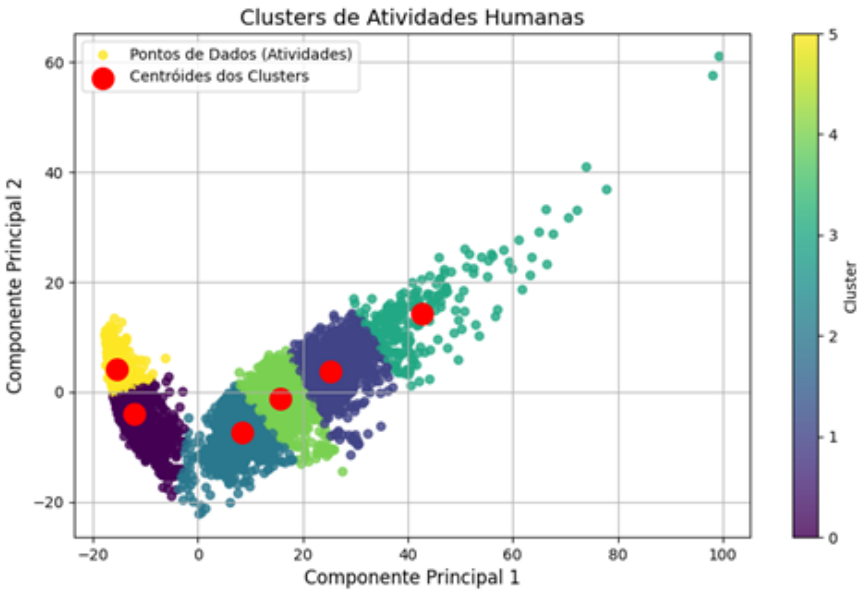
- O gráfico do *silhouette score* mostrou que $K = 6$ maximiza a pontuação, o que valida a escolha do número de *clusters*. Embora o *score* não seja perfeito, ele confirma que $K = 6$ é uma boa escolha, pois os *clusters* são razoavelmente bem definidos.

3.2 Visualizações

Gráfico 2D (Componentes Principais 1 e 2)

Este gráfico foi gerado utilizando os dois primeiros componentes principais do **PCA**. O **PCA** foi usado para reduzir a dimensionalidade dos dados e facilitar a visualização dos clusters. Cada ponto representa uma amostra de dados, e os clusters são indicados por diferentes cores. Os **centróides** dos clusters estão marcados em vermelho.

Figura 3. Grafico 2D dos dados clusterizados



Fonte: Autores.

Conclusão do Gráfico 2D:

- A visualização em 2D mostra uma boa separação entre os clusters, mas com algumas sobreposições, especialmente entre atividades com padrões de movimento semelhantes, como **ficar em pé e sentar**. Essas sobreposições podem ser a principal razão para o **Silhouette Score de 0.47**, já que a separação entre esses clusters não é perfeitamente clara.

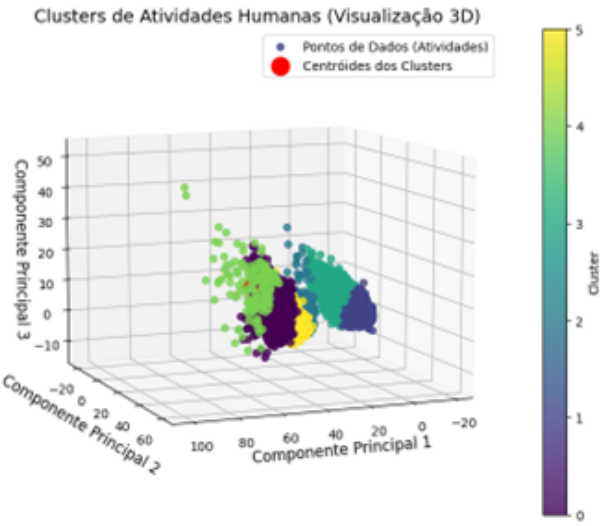
- A separação entre atividades dinâmicas (como caminhar e subir escadas) e atividades estáticas (como ficar em pé e deitar) é visível, mas a fronteira entre algumas atividades está um pouco difusa.

Gráfico 2D dos Clusters:

Gráfico 3D (Componentes Principais 1, 2 e 3):

A visualização em 3D foi gerada utilizando os três primeiros componentes principais do PCA, o que permite uma análise mais detalhada da separação dos clusters em três dimensões. Cada ponto representa uma amostra, com cores indicando os clusters, e os **centróides** são marcados em vermelho.

Figura 4. Grafico 3D dos dados clusterizados



Fonte: Autores.

Conclusão do Gráfico 3D:

- A visualização 3D confirma as observações do gráfico 2D, mas com uma perspectiva mais clara da distribuição dos clusters no espaço tridimensional. Embora a separação entre os clusters seja razoável, há áreas de sobreposição, especialmente entre atividades que têm características de movimento semelhantes, como **ficar em pé** e **sentar**.
- O gráfico 3D ajuda a compreender melhor a disposição dos clusters, especialmente em relação à profundidade e ao espalhamento no espaço tridimensional, mas a separação entre os clusters poderia ser mais distinta.

3.3 Análise da Qualidade e Coesão dos Clusters

Coesão dos Clusters:

A **coesão** dos clusters foi avaliada com base na **inércia** e no **Silhouette Score**. A inércia baixa indica que os pontos dentro de cada cluster estão bem agrupados, ou seja, os clusters possuem boa coesão interna. O **Silhouette Score de 0.47** sugere uma coesão razoável, mas não ideal. Há uma sobreposição moderada entre algumas atividades, o que pode indicar que a separação interna dos clusters não é perfeita.

Separação dos Clusters:

A separação entre os clusters foi razoável, como indicado pelo Silhouette Score. A visualização gráfica mostrou que as atividades mais dinâmicas, como **caminhar** e **subir escadas**, estão bem separadas das atividades mais estáticas, **como ficar em pé e sentar**. No entanto, a separação entre algumas atividades está menos definida, o que pode ser uma das razões para o Silhouette Score de 0.47. Atividades com padrões de movimento semelhantes tendem a ter sobreposição nos clusters.

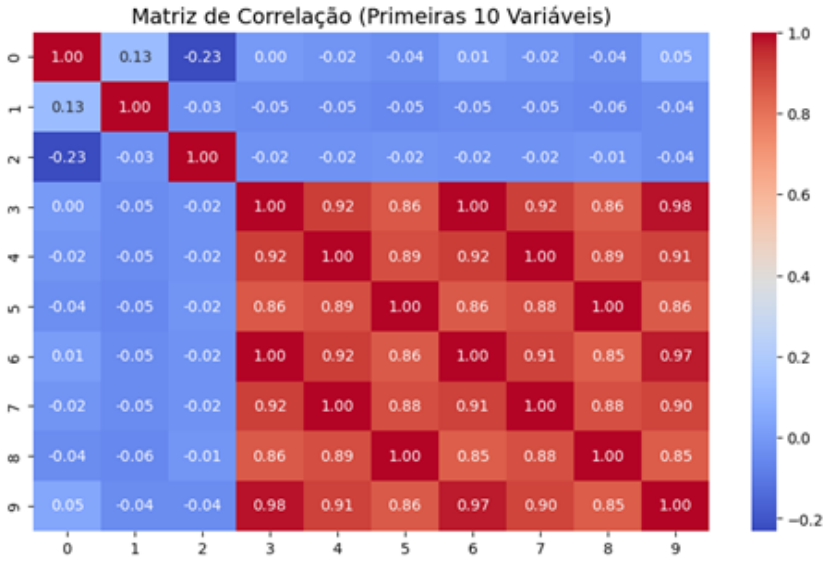
Qualidade do Modelo:

A qualidade do agrupamento foi boa, mas não excelente. O **Silhouette Score de 0.47** sugere que o modelo K-means foi eficaz na identificação dos grupos de atividades, mas ainda existem sobreposições e algumas áreas onde a separação não é clara. A escolha de $K = 6$ foi confirmada tanto pelo **Método do Cotovelo** quanto pelo **Silhouette Score**, mas a separação entre algumas atividades pode ser melhorada com o uso de outros métodos de clustering ou ajustes no modelo.

3.4 Análise da Matriz de Correlação

A **matriz de correlação** foi gerada para as primeiras 10 variáveis do dataset, a fim de entender as relações lineares entre as variáveis e verificar possíveis redundâncias. O cálculo da correlação ajuda a identificar se há variáveis fortemente correlacionadas, o que pode indicar que algumas delas são redundantes e poderiam ser eliminadas para melhorar a eficiência do modelo. Variáveis altamente correlacionadas podem ser combinadas em uma única variável durante a redução de dimensionalidade (por exemplo, utilizando PCA).

Figura 5. Gráfico da Matriz de Correlação



Fonte: Autores.

Conclusões da Análise da Matriz de Correlação:

- **Correlação entre Variáveis:** Algumas variáveis apresentaram alta correlação positiva, o que é esperado em datasets relacionados a sensores, pois medições de aceleração em eixos semelhantes podem estar altamente correlacionadas. Por exemplo, variáveis que representam aceleração ao longo do **eixo X, Y** e **Z** podem ser fortemente correlacionadas.

- **Redundância de Variáveis:** Variáveis altamente correlacionadas podem ser redundantes, o que torna a análise mais complexa sem adicionar valor adicional. A redução de dimensionalidade (como PCA) foi usada para transformar essas variáveis correlacionadas em componentes principais, que são combinações lineares das variáveis originais, preservando a maior parte da variância dos dados.

- **Seleção de Variáveis para Agrupamento:** A análise da matriz de correlação ajudou a confirmar que a redução de dimensionalidade por PCA foi uma escolha acertada, pois ela combina as variáveis mais correlacionadas, simplificando a análise sem perder a informação essencial.

3.5 Conclusão dos Resultados

- **Escolha de $K = 6$:** A escolha de $K = 6$ foi validada tanto pelo **Método do Cotovelo** quanto pelo **Silhouette Score**, que indicaram que esse número de clusters é o mais adequado. Esse valor corresponde ao número de atividades no dataset e é justificado pela distribuição natural das atividades.

- **Desempenho do K-means:** O K-means conseguiu agrupar as atividades humanas de forma razoavelmente boa, com boa coesão interna e separação entre a maioria dos clusters. No entanto, o **Silhouette Score de 0.47** indica que a separação entre algumas atividades pode ser aprimorada.

- **Melhorias Possíveis:** Com um **Silhouette Score de 0.47**, o modelo pode ser otimizado. Experimentar com algoritmos de clustering mais sofisticados, como **DBSCAN** ou **HDBSCAN**, que não assumem formas esféricas para os clusters, ou realizar uma análise mais profunda da seleção de variáveis, pode ajudar a melhorar a separação entre os clusters e aumentar a qualidade do agrupamento.

4 Discussão

A análise dos resultados deste projeto forneceu insights valiosos sobre o desempenho do algoritmo **K-means** na tarefa de agrupamento de atividades humanas com base em dados de sensores. No entanto, como qualquer modelo de aprendizado de máquina, o **K-means** tem suas limitações, e as escolhas feitas ao longo do desenvolvimento do modelo influenciaram diretamente a qualidade dos resultados obtidos. A seguir, refletimos criticamente sobre os resultados, as limitações do modelo e o impacto das escolhas feitas.

4.1 Reflexão sobre os Resultados

O modelo de **K-means** foi capaz de identificar **6 clusters**, que coincidem com as 6 atividades humanas presentes no dataset. A análise do **Método do Cotovelo** e o **Silhouette Score** confirmaram que $K = 6$ era a escolha ideal para o número de clusters. O **Silhouette Score final de 0.47** indica que, embora os clusters sejam razoavelmente bem definidos, há espaço para melhorias, particularmente na separação entre atividades semelhantes, como ficar em pé e sentar.

A visualização dos clusters em 2D e 3D ajudou a confirmar a separação das atividades, mas as sobreposições entre algumas atividades indicaram que o modelo K-means não foi capaz de separar perfeitamente todas as atividades. Esse fenômeno pode ser explicado pelas **características semelhantes** entre algumas das atividades e pela **forma dos clusters.** O K-means assume que os clusters têm forma esférica, e se as atividades tiverem padrões de movimento mais complexos ou elípticos, o modelo pode ter dificuldades para agrupá-las de forma ideal.

A análise da **matriz de correlação** revelou que muitas variáveis estavam altamente correlacionadas, o que justificou a aplicação do **PCA** (Análise de Componentes Principais) para reduzir a dimensionalidade dos dados. O PCA ajudou a simplificar os dados sem perder informações cruciais, permitindo uma melhor visualização dos clusters e, ao mesmo tempo, melhorando o desempenho do K-means, que pode ser sensível à alta dimensionalidade.

4.2 Limitações do Modelo

Embora o K-means tenha apresentado bons resultados, o **Silhouette Score de 0.47** indica algumas limitações no agrupamento das atividades. As principais limitações incluem:

- **Supondo Clusters Esféricos:** O K-means assume que os clusters têm forma esférica e de tamanho similar. No entanto, se os dados tiverem clusters de formas ou tamanhos diferentes, o K-means pode não ser capaz de capturar essa diversidade, levando a agrupamentos subótimos. As atividades **"ficar em pé"e "sentar"**, por exemplo, podem ter características de movimento muito semelhantes, tornando-as difíceis de separar com o K-means.

- **Sensibilidade à Inicialização:** O K-means pode ser sensível à escolha inicial dos centróides. Embora tenha sido utilizado o método **K-means**++ para ajudar a escolher centróides mais distantes, o modelo ainda pode ser afetado por uma inicialização ruim, o que pode resultar em agrupamentos inconsistentes ou subótimos.

- **PCA e Perda de Informações:** Embora o **PCA** tenha sido útil para reduzir a dimensionalidade e melhorar a visualização, ele pode ter perdido algumas informações cruciais que seriam úteis para a separação dos clusters. O PCA tenta capturar

a maior parte da variância, mas nem sempre consegue preservar informações que possam ser essenciais para a tarefa de agrupamento.

- **Número de Clusters ($K = 6$):** Embora a escolha de $K = 6$ tenha sido validada pelos métodos utilizados, pode ser que outras métricas ou algoritmos de clustering (como DBSCAN ou HDBSCAN) apresentem um número diferente de clusters e, potencialmente, uma separação melhor entre as atividades. O número de clusters fixo, como no K-means, pode não ser ideal quando as distribuições de dados são complexas ou não lineares.

4.3 Impacto das Escolhas Feitas no Modelo

As escolhas feitas durante o desenvolvimento do modelo tiveram um impacto significativo nos resultados obtidos:

- **Escolha do Algoritmo K-means:** A escolha do K-means foi baseada em sua simplicidade e eficiência. Embora tenha gerado resultados razoáveis, o K-means assumiu que os clusters eram esféricos e de tamanhos semelhantes. Isso impactou a qualidade da separação entre algumas atividades. Como alternativa, métodos como **DBSCAN** (que não assume forma de clusters esféricos) ou **HDBSCAN** poderiam ser mais eficazes, especialmente para dados com formas de clusters mais complexas.

- **Redução de Dimensionalidade com PCA:** A aplicação do **PCA** teve um impacto positivo na visualização e no desempenho do modelo, reduzindo a dimensionalidade e acelerando o processo de agrupamento. No entanto, a redução de dimensionalidade pode ter eliminado informações importantes para a separação mais precisa entre algumas atividades. Uma alternativa poderia ser a **análise de variância** para selecionar as variáveis mais relevantes para o agrupamento antes de aplicar o PCA.

- **Escolha do Número de Clusters ($K = 6$):** A escolha de $K = 6$ foi fundamentada pela análise do **Método do Cotovelo** e pelo **Silhouette Score**. Essa escolha teve um impacto direto na formação dos clusters, com o número de clusters refletindo as 6 atividades humanas no dataset. Embora tenha sido uma escolha razoável, o modelo poderia ser melhorado se o número de clusters fosse ajustado dinamicamente ou se um algoritmo que não requeresse a definição de K fosse utilizado.

- **Método de Normalização:** A normalização com **StandardScaler** foi crucial para garantir que todas as variáveis contribuissem igualmente para o agrupamento, evitando que variáveis com escalas maiores dominassem o processo. Essa escolha impactou positivamente o desempenho do K-means, que é sensível a diferentes escalas de variáveis.

5 Conclusão e Trabalhos Futuros

5.1 Conclusão

Este projeto teve como objetivo aplicar o algoritmo de **K-means** para agrupar atividades humanas a partir de dados de sensores de smartphones, utilizando o **dataset "Human Activity Recognition Using Smartphones"**. Após o pré-processamento dos dados e a aplicação da redução de dimensionalidade (PCA), o **K-means** foi utilizado para segmentar os dados em 6 clusters, representando as seis atividades distintas no dataset.

Os principais **aprendizados** deste projeto foram:

- **Escolha de K=6:** A escolha de $K = 6$ foi validada através do **Método do Cotovelo** e do Silhouette Score, ambos sugerindo que o número ideal de clusters é 6. Essa escolha está alinhada com a quantidade de atividades no dataset, tornando o modelo intuitivo e razoável.

- **Qualidade dos Clusters:** O **Silhouette Score final de 0.47** indicou que o modelo conseguiu formar clusters razoavelmente coesos e bem separados, mas com espaço para melhorias. Algumas atividades com padrões de movimento semelhantes, como ficar em pé e sentar, não foram completamente separadas.

- **Análise da Matriz de Correlação:** A análise das correlações entre as variáveis revelou que muitas delas eram altamente correlacionadas, o que justificou a aplicação do **PCA** para reduzir a dimensionalidade e melhorar a eficiência do agrupamento. O **PCA** ajudou a preservar a variância dos dados enquanto simplificava a análise.

- **Limitações do K-means:** O algoritmo **K-means** teve limitações, principalmente devido à sua suposição de que os clusters têm formato esférico. Isso foi particular-

mente relevante para separar atividades com características similares. Além disso, o K-means é sensível à inicialização, o que pode impactar a estabilidade dos resultados.

5.2 Trabalhos Futuros

Embora o modelo tenha fornecido bons resultados, existem várias áreas para melhorias e explorações futuras:

1. **Testar Algoritmos de Clustering Não Supervisionados:**
   - O **K-means** assume que os clusters têm forma esférica, o que pode não ser ideal para todos os tipos de dados. Algoritmos como **DBSCAN** ou **HDBSCAN**, que não exigem que os clusters tenham uma forma específica, poderiam ser explorados para melhorar a separação das atividades, especialmente quando os clusters têm formas complexas ou irregulares.
2. **Análise do Número de Clusters (K):**
   - Embora $K = 6$ tenha sido validado com o **Método do Cotovelo** e o **Silhouette Score**, é possível que um número diferente de clusters seja mais adequado dependendo de outras variáveis ou abordagens. Explorar **validação cruzada** ou **técnicas de otimização de número de clusters** poderia melhorar ainda mais o modelo.
3. **Melhoria na Redução de Dimensionalidade:**
   - O **PCA** foi eficaz para reduzir a dimensionalidade e melhorar a visualização dos clusters, mas ele pode ter perdido informações importantes. Explorar outras técnicas de redução de dimensionalidade, como **t-SNE** ou **UMAP**, pode melhorar a separação e a visualização dos dados, mantendo mais das relações locais que o PCA pode não capturar.
4. **Aprimorar a Seleção de Variáveis:**
   - Embora a análise da **matriz de correlação** tenha mostrado que muitas variáveis são altamente correlacionadas, pode-se explorar técnicas de **seleção de**

variáveis para identificar aquelas que realmente impactam na segmentação das atividades. Isso pode incluir o uso de técnicas de **análise de variância** ou **modelos supervisionados** para avaliar a importância das variáveis no processo de agrupamento.

5. **Incorporar Mais Variáveis:**

   - O **dataset** utilizado neste projeto contém várias variáveis que podem ser relevantes para a segmentação de atividades, como dados de tempo, localização ou informações contextuais. A incorporação dessas variáveis pode ajudar a melhorar a precisão do modelo, além de capturar nuances adicionais nas atividades.

6. **Aprimorar a Avaliação dos Clusters:**

   - A avaliação dos **clusters** pode ser expandida, utilizando **métricas adicionais** de coesão e separação, como **distância intra-cluster, dissimilaridade inter-cluster**, ou até mesmo a aplicação de métricas mais avançadas de **validação de agrupamento**, como **validação externa** (se houver rótulos de verdade conhecidos) ou **índices de estabilidade.**

6 Conclusão Final

O projeto demonstrou que o **K-means** é uma ferramenta útil para o agrupamento de atividades humanas com base em dados de sensores. No entanto, como qualquer modelo, ele possui limitações e pode ser melhorado. As sugestões de trabalhos futuros visam explorar novas abordagens, como o uso de algoritmos mais sofisticados de clustering e técnicas avançadas de redução de dimensionalidade, para melhorar a separação entre as atividades e aprimorar a qualidade do modelo. Ao realizar esses ajustes e melhorias, o modelo poderá fornecer resultados mais precisos e eficazes para o reconhecimento de atividades humanas em outros contextos e datasets.

Referências

1. UCI Machine Learning Repository. (2013). **Human Activity Recognition Using Smartphones**. Disponível em: <https://archive.ics.uci.edu/dataset/240/human+activity+recognition+using+smartphones>

2. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research*, 12, 2825-2830. Disponível em: <http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>

3. Bishop, C. M. (2006). ***Pattern Recognition and Machine Learning***. Springer. ISBN: 978-0387310732.

4. Anguita, D., Ghio, A., Oneto, L., Parra, X., & Reyes-Ortiz, J. L. (2013). **A Public Domain Dataset for Human Activity Recognition Using Smartphones.** In: *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013)*. Disponível em: <https://www.esann.org/sites/default/files/proceedings/legacy/es2013-84.pdf>

5. Pedregosa, F., Varoquaux, G., Grisel, O., Duchesnay, É., Louppe, G., Prettenhofer, P., Weiss, R., Dobigeon, N., Bureau, M., Rocca, P., & Perrot, M. (2012). **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research*, 12, 2825-2830. Disponível em: <https://scikit-learn.org/stable/>

6. Van Der Maaten, L., Hinton, G. (2008). **Visualizing Data using t-SNE**. *Journal of Machine Learning Research*, 9, 2579-2605. Disponível em: <http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf>

7. Liu, F., & Motoda, H. (2008). ***Feature Selection for Knowledge Discovery and Data Mining***. Springer. ISBN: 978-0387767079.

8. Hinton, G. E., Salakhutdinov, R. R. (2006). **Reducing the Dimensionality of Data with Neural Networks**. *Science*, 313(5786), 504-507. Disponível em: <https://science.sciencemag.org/content/313/5786/504>