CENTRO DE PESQUISA E DESENVOLVIMENTO TECNOLÓGICO EM INFORMÁTICA E ELETROELETRÔNICA DE ILHÉUS - CEPEDI

Relatório Técnico: Implementação e Análise do Algoritmo k-Means Aplicado ao dataset *"Human Activity Recognition Using Smartphones"*

Weslei Ferreira
Daniel Morais Pereira

01 de dezembro de 2024

Resumo

Este relatório consiste em descrever o passo a passo do desenvolvimento de projeto que analisa as atividades humanas utilizando o algoritmo de clusterização K-means. Foram realizadas etapas de análise exploratória pré-processamento e validação do modelo. A metodologia aplicada garantiu uma identificação coerente dos padrões presentes nos dados. Os resultados destacam as devidas formação de partições com base nas atividades analisadas.

Introdução

A crescente utilização de dispositivos móveis e sensores em dispositivos vestíveis tem permitido a coleta em massa de dados para o reconhecimento de atividade AR, Human Activity Recognition). Essa tarefa é essencial em áreas como saúde, esportes e interação humano-computador.

Este trabalho aplica o algoritmo K-means para segmentar os dados em partições que representam diferentes atividades humanas. A escolha do K-means foi baseada em sua simplicidade e eficiência para cenários não supervisionados, além de sua ótima eficiência para identificar padrões inerentes em grandes conjuntos de dados.

Metodologia

1. Análise Exploratória de Dados (EDA)

Descrição dos Dados

O conjunto de dados utilizado para o projeto foi o *Human Activity Recognition Using Smartphones* (HAR), que contém informações coletadas de sensores de acelerômetro e giroscópio em dispositivos móveis. As principais características do dataset são:

- **Número de amostras**: 10.299 registros, representando medições de 30 voluntários durante a realização de 6 diferentes atividades.
- **Número de features**: 561 variáveis contínuas, correspondendo a estatísticas extraídas de sinais brutos coletados pelos sensores.
- **Classes (Atividades)**: Caminhar, Subir escadas, Descer escadas, Sentado, Em pé e Deitado.

Visualizações

Foram realizadas visualizações para entender melhor a estrutura dos dados e identificar padrões iniciais:

1. **Distribuição das variáveis**: Histogramas foram gerados para analisar a dispersão dos valores. A maioria das variáveis apresentou uma distribuição aproximadamente normal, facilitando a aplicação de técnicas de normalização.
2. **Correlação entre features**: Um heatmap de correlação revelou a presença de algumas variáveis altamente correlacionadas, indicando a possibilidade de redundância de informações em determinados casos.
3. **Gráficos de dispersão**: Utilizando t-SNE e PCA, visualizamos os dados em 2 dimensões, observando indícios de separação entre classes.

2. Implementação do K-means

Pré-processamento

Para garantir que o algoritmo K-means funcionasse de forma eficiente, foram realizadas as seguintes etapas de pré-processamento:

- **Normalização dos dados**: As variáveis foram escaladas utilizando o método ***StandardScaler***, que centraliza os valores em torno da média (0) e normaliza o desvio padrão para 1.
- **Seleção de Features**: Para reduzir a dimensionalidade e eliminar possíveis redundâncias, foi utilizado o método ***SelectKBest*** com o critério de informação mútua, selecionando as 50 variáveis mais relevantes para o agrupamento.

Escolha do Número de Clusters

Foram testados diferentes valores de K (número de clusters), utilizando os seguintes métodos para definir o valor ideal:

1. **Método do Cotovelo (Elbow Method)**: O gráfico da soma das distâncias quadráticas internas aos clusters (inertia) indicou que o valor k=6k = 6k=6 seria adequado, refletindo as 6 atividades no dataset.
2. **Silhouette Score**: Para validar a qualidade dos clusters, foi calculado o Silhouette Score para cada K. O valor mais alto foi obtido com k=6k = 6k=6, confirmando a escolha do Método do Cotovelo.

Configuração do Algoritmo

O K-means foi configurado com os seguintes parâmetros:

- **Número de clusters**: k=6.
- **Inicialização dos centroides**: Utilizamos o método **k-means + +** para melhorar a eficiência e evitar inicializações ruins.
- **Métrica de distância**: Distância Euclidiana, apropriada para dados normalizados.
- **Número máximo de iterações**: O valor padrão de 300 foi suficiente para garantir a convergência.

Resultados

Nesta seção, apresentamos os principais resultados obtidos durante a aplicação do algoritmo K-means e sua análise comparativa com métodos supervisionados.

1. Métricas de Avaliação do K-means

Para avaliar a qualidade dos clusters formados pelo K-means, utilizamos as seguintes métricas:

- **Silhouette Score**: O valor obtido foi **0.6086**. Esse resultado indica que há uma separação razoável entre os clusters e uma boa coesão interna.Além disso, com este resultado, é sugerido que as atividades humanas possuem padrões distintos, permitindo que o algoritmo identifique agrupamentos relevantes.
- **Adjusted Rand Index (ARI)**: O valor apresentado foi **0.5403**. Com esse resultado podemos refletir que há uma correspondência moderada entre os clusters formados pelo K-means e as classes reais.
- **Normalized Mutual Information (NMI)**: A métrica NMI resultou em **0.6611**. Tal resultado evidencia uma boa correspondência entre os clusters e as categorias reais do dataset.

2. Visualizações

- **Matriz de Confusão Normalizada**: Foi gerada uma matriz de confusão normalizada para avaliar a associação entre as atividades reais e os clusters formados pelo K-means. A matriz revelou que as atividades "Deitado" e "Ficar em Pé" foram agrupadas corretamente na maior parte dos casos, enquanto as atividades dinâmicas, como "Caminhar", apresentaram maior sobreposição entre clusters.
- **Representação Visual com t-SNE**: Uma visualização dos dados em 2D foi realizada utilizando t-SNE para reduzir a dimensionalidade. Os clusters formados pelo K-means foram destacados em diferentes cores. É possível evidenciar, visualizando o gráfico que as atividades "Deitado" e "Sentado" formaram grupos bem separados, enquanto "Subir Escadas" e "Descer Escadas" apresentaram maior proximidade.
- **Distribuição de Atividades por Cluster**: Para analisar com houve a distribuição das atividades reais entre os clusters, foi gerada uma tabela, localizada abaixo. Com uma rápida análise é possível concluir que clusters associados a atividades estáticas ("Deitado" e "Sentado") apresentaram alta coesão, enquanto atividades dinâmicas exibiram maior dispersão entre os clusters.

  Abaixo, segue os gráfico das representações citadas acima

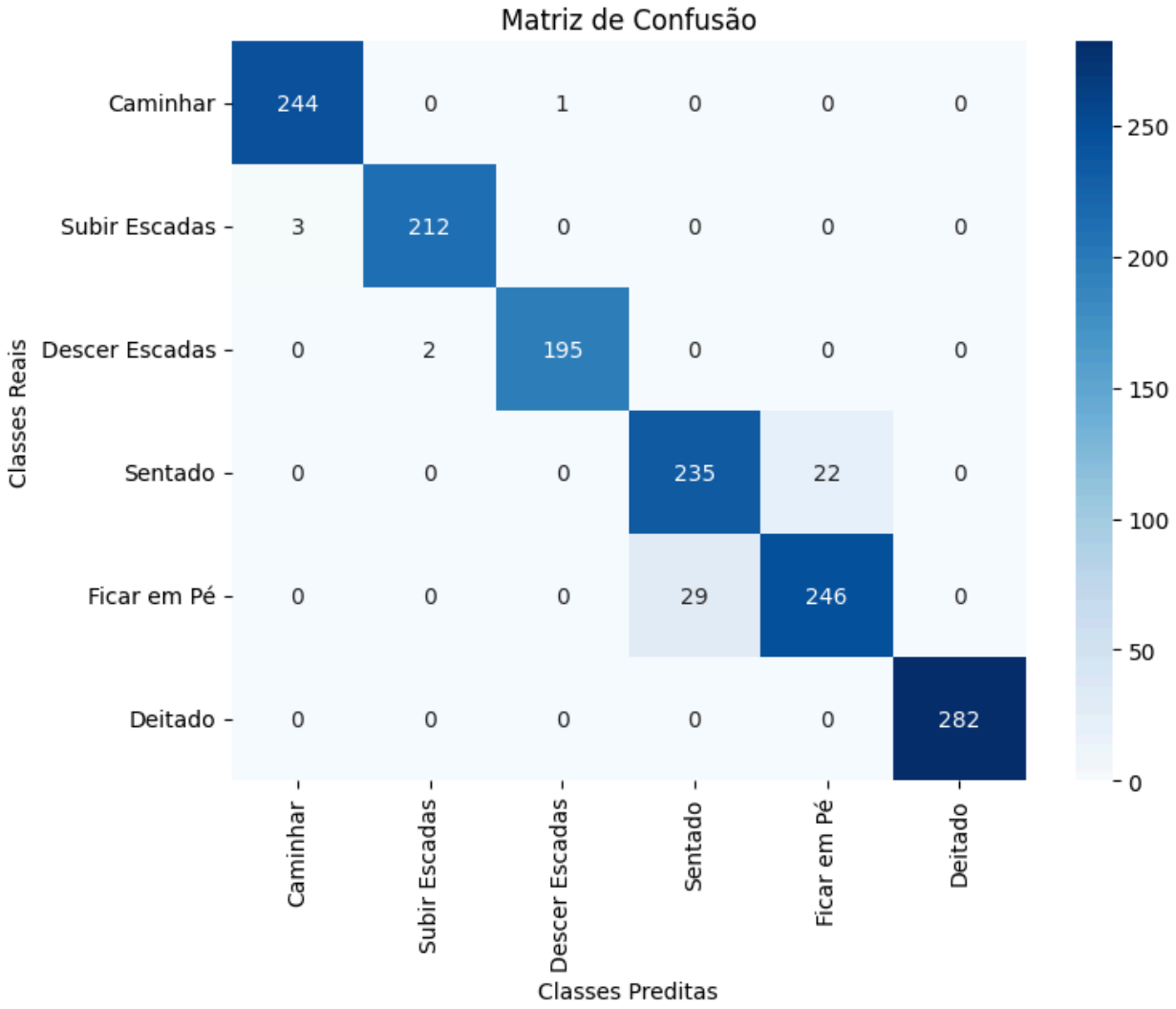


Figura 1: Matriz de confusão feita a partir do modelo DNN. Fonte: dos autores

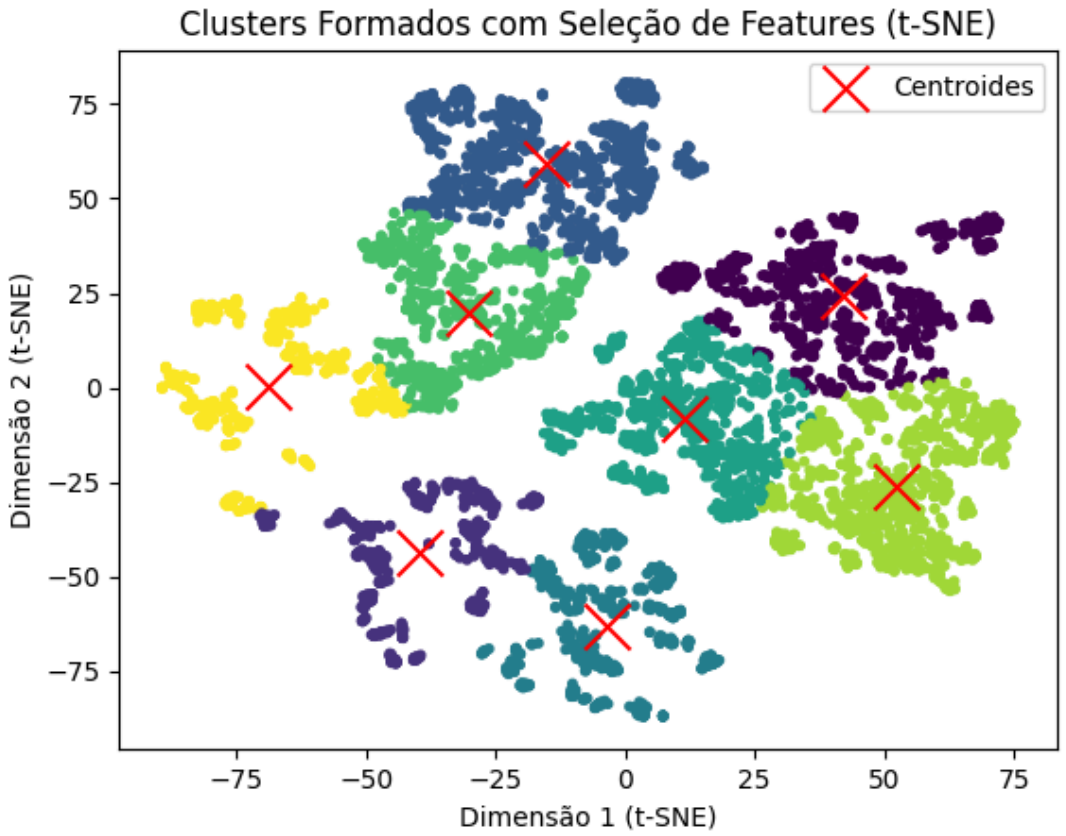
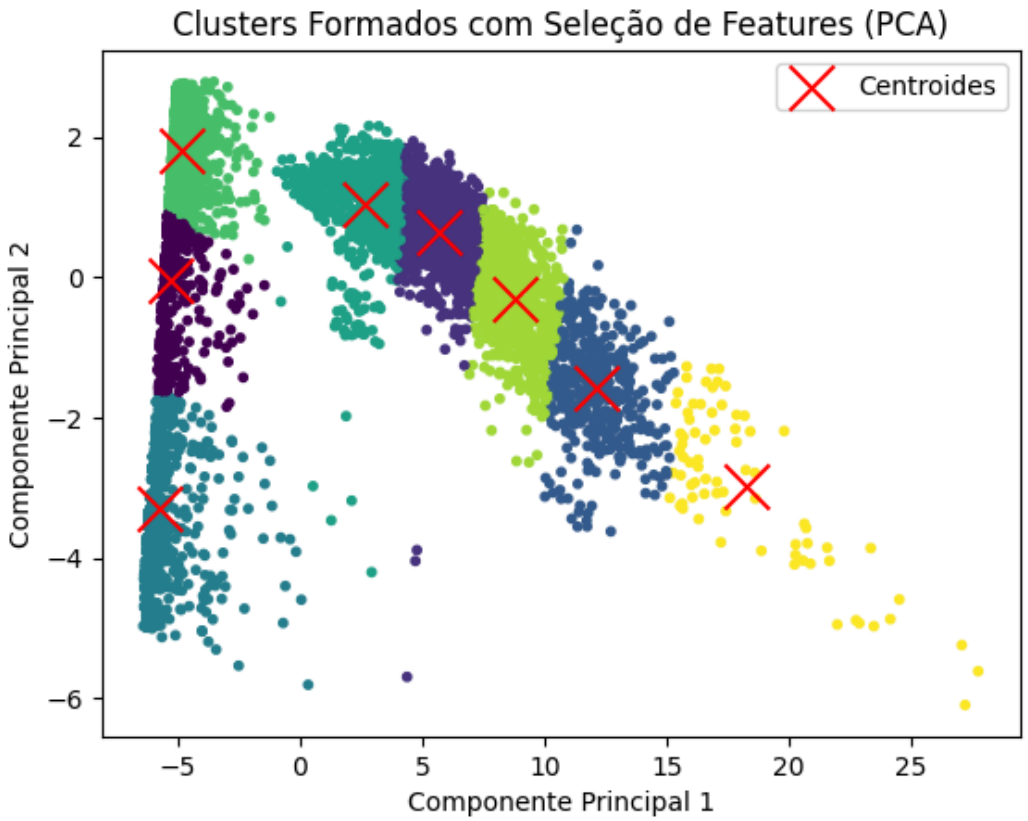
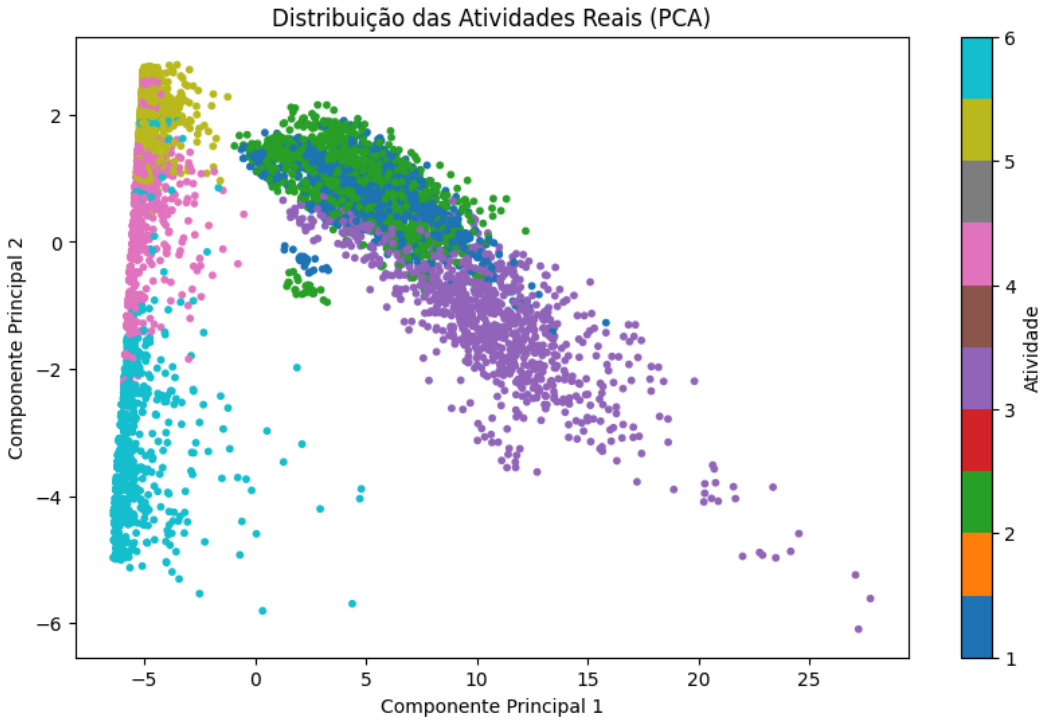


Figura 2: Formação dos Clusters com t-SNE. Fonte: dos autores





Figuras 3 e 4: Formação das atividades com PCA Fonte: Dos Autores

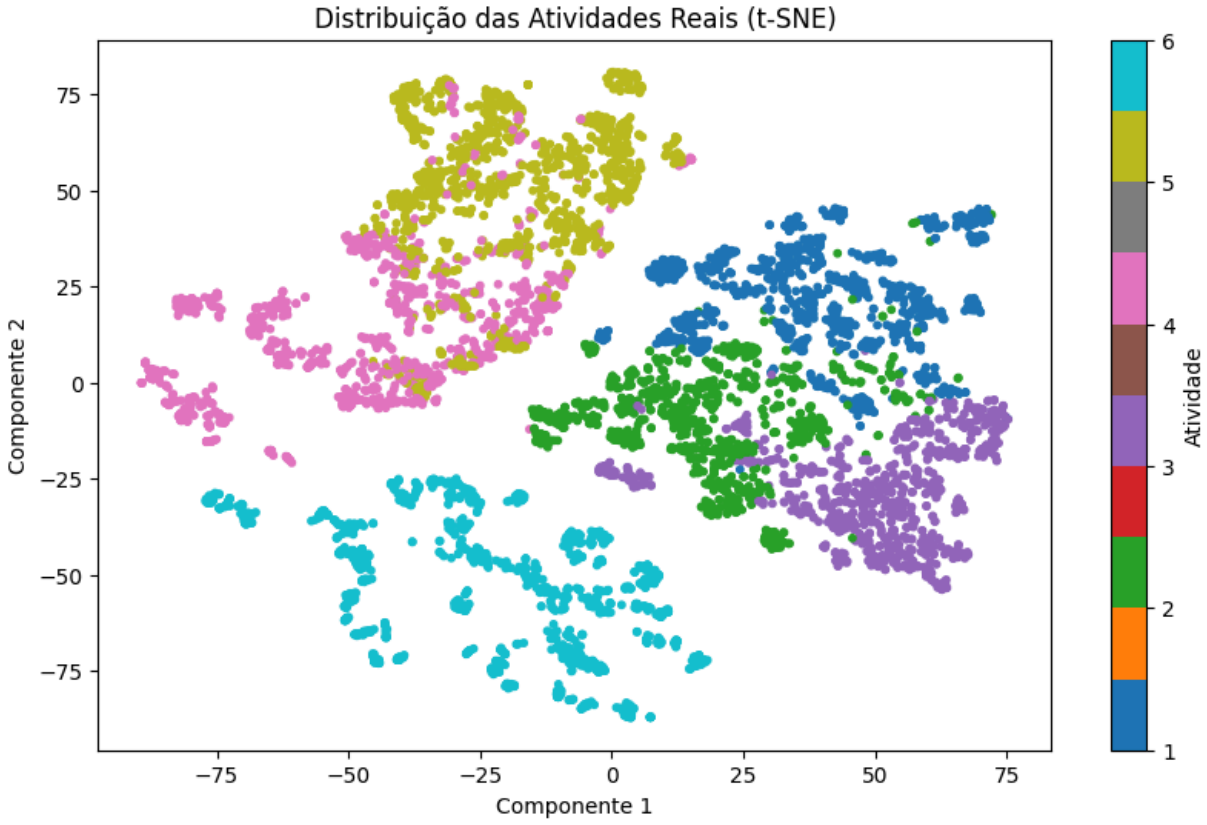


Figura 5 : Formação da distribuição das atividades com t-SNE. Fonte: dos autores

Tabela sobre a distribuição de Atividades por Cluster

Atividade	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Caminhar	241	0	4	0	0	0
Subir escadas	2	204	9	0	0	0
Descer escadas	0	0	197	0	0	0
Sentado	0	0	0	223	34	0
Ficar em pé	0	0	0	29	246	0
Deitado	0	0	0	0	0	282

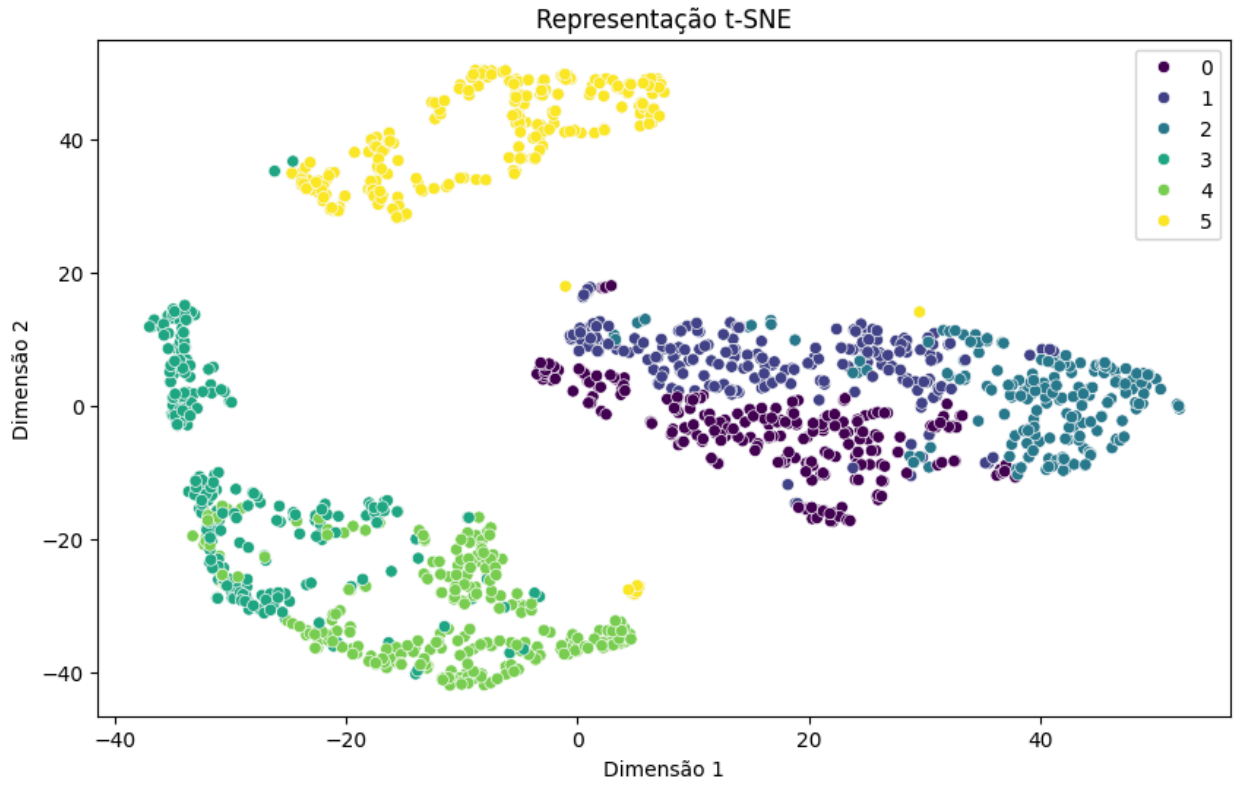


Figura 6: Representação com t-SNE obtida através da classificação do DNN

Discussão

Os resultados obtidos demonstraram que o algoritmo K-means é uma ferramenta útil para identificação inicial de padrões em dados de alta dimensionalidade, como no caso do reconhecimento de atividades humanas. No entanto, algumas limitações naturais do método foram observadas como separação de atividades estáticas como "Deitado" e "Sentado", foram identificadas de forma clara pelo K-means, com clusters bem definidos. Por outro lado, atividades dinâmicas, como "Caminhar", "Subir Escadas" e "Descer Escadas", apresentaram maior sobreposição, refletindo a proximidade entre os padrões gerados pelos sensores nessas classes.

A normalização dos dados foi essencial para evitar que variáveis com escalas maiores dominassem o cálculo das distâncias no K-means. Como se tratava de uma base de dados usada em modelos supervisionados, nós achamos que seria interessante incluir um modelo de rede neural supervisionado como DNN para ter uma base de comparação e como foi exposto acima o DNN teve um ótimo desempenho.

Conclusão

Este projeto explorou o algoritmo K-means e sua aplicação para o reconhecimento de atividades humanas usando dados de sensores. As principais etapas para a execução foram a análise exploratória, pré-processamento e a escolha do número de clusters juntamente com a avaliação dos resultados, bem como a comparação com a rede neural supervisionada, DNN. Mesmo as limitações do K-means, como sensibilidade ao número de clusters e incapacidade de lidar com clusters que não são esféricos, o trabalho mostrou o potencial como primeira abordagem para a exploração de dados e análise não supervisionada.

Referências

SKLEARN. User Guide: Clustering. Disponível em: https://scikit-learn.org/stable/modules/clustering.html. Acesso em: 25 nov. 2024.

SKLEARN. User Guide: Metrics and scoring: assessing performance. Disponível em: https://scikit-learn.org/stable/modules/model_evaluation.html. Acesso em: 25 nov. 2024.

KERAS. **Getting Started: Introduction to Keras for Researchers**. Disponível em: https://keras.io/getting_started/. Acesso em: 25 nov. 2024.

UC IRVINE. **Human Activity Recognition Using Smartphones Data Set**. Disponível em: https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones. Acesso em: 25 nov. 2024.