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

Daniel de Souza Pereira
Victor Cesar do Rosario Calheiros

01 de dezembro de 2024

Resumo:

Esse relatório tem como objetivo mostrar todas as etapas do projeto em que desta vez, focamos em desenvolver uma implementação e análise do algoritmo de K-means com o Dataset Human Activity Recognition, um experimento com 30 voluntários que tiveram que executar seis atividades físicas usando o smartphone e tem um banco de dados com as informações de cada atividade, como aceleração tri-axial do acelerômetro e aceleração corporal estimada, velocidade angular tri-axial do giroscópio e um vetor de 561 características com variáveis de domínios do tempo e da frequência de cada participante, para o código conseguir passar informações de Inércia média e Pontuação de Silhueta, visando achar o melhor número de clusters na análise de silhueta, trazendo resultados dos Clusters em PCA tanto em gráficos 2D quando 3D.

Introdução:

O reconhecimento de atividades humanas ou Human Activity Recognition, é uma área de estudo focada em identificar e classificar atividades realizadas por indivíduos, com base em dados coletados por sensores. Sensores de aceleração e giroscópios, geralmente embutidos em smartphones ou dispositivos vestíveis, geram dados que podem ser usados para distinguir atividades como caminhar**,** correr**,** sentar-se ou ficar de pé**.** Nesse contexto, foi feito um experimento com 30 pessoas em executar 6 atividades físicas para adquirir dados dos 3 eixos (X,Y,Z) dessas atividade. Para conseguir agrupar esses dados de forma ordenada foi utilizada o K-means, uma técnica de aprendizado que agrupa os dados em K-Clusters, com base na similaridade entre os pontos dos dados, sendo adequada para este projeto por ajudar na interpretação e análise desses dados de alta dimensionalidade e muitos padrões complexos.

Metodologia:

Basicamente, faremos uma explicação sobre o passo a passo que o nosso código faz:

Ele inicia importando todas as bibliotecas que serão necessárias para o funcionamento do código e a exibição dos resultados graficamente como desejamos, fazendo o carregamento dos dados de X_train, Y_train, X_test e Y_test diretamente dos links raw do github, também fazendo o carregamento do arquivo activity_labels.txt, para mapear as atividades de acordo com os seus rótulos. Depois é feito a normalização dos dados através do StandardScaler, garantindo que todas as variáveis contribuam para o algoritmo k-means, tendo em vista que os dados dos sensores podem ter escalas muito diferentes. Em seguida, o algoritmo K-means++ é executado 10 vezes para inicializar os centroides de forma eficiente, ajudando a otimizar o tempo de convergência e fazendo a verificação de estabilidade. Após isso, é usado o método do cotovelo para encontrar o número ideal de clusters, o qual nos atentamos em demonstrar graficamente também. Então, é feito a análise de silhueta para escolher o melhor número de clusters e se realiza o treinamento do K-means final com a melhor configuração para obter os rótulos dos clusters e centroides. Por fim, é mostrada as métricas de avaliação no conjunto de treino e de teste, fazendo a comparação dos clusters com y_test e terminando com a visualização em PCA, tanto em 2D quanto em 3D.

Resultados:

Dando um resumo geral, como resultados desse projeto, conseguimos adquirir informações da média de Inércia e Pontuação de Silhueta, também fizemos a análise de silhueta para achar o maior valor de pontuação de silhueta relacionado ao número de clusters.

Dimensões:

Dimensões de X_train: (7352, 561)

Dimensões de X_test: (2947, 561)

Dimensões de Y_train: (7352,)

Dimensões de Y_test: (2947,)

Resultados das Execuções do K-means:

Execução 1: Inércia = 207069.80, Silhouette Score = 0.70

Execução 2: Inércia = 207069.80, Silhouette Score = 0.70

Execução 3: Inércia = 207069.80, Silhouette Score = 0.70

Execução 4: Inércia = 207069.80, Silhouette Score = 0.70

Execução 5: Inércia = 207069.80, Silhouette Score = 0.70

Execução 6: Inércia = 207069.80, Silhouette Score = 0.70

Execução 7: Inércia = 207069.80, Silhouette Score = 0.70

Execução 8: Inércia = 207069.80, Silhouette Score = 0.70

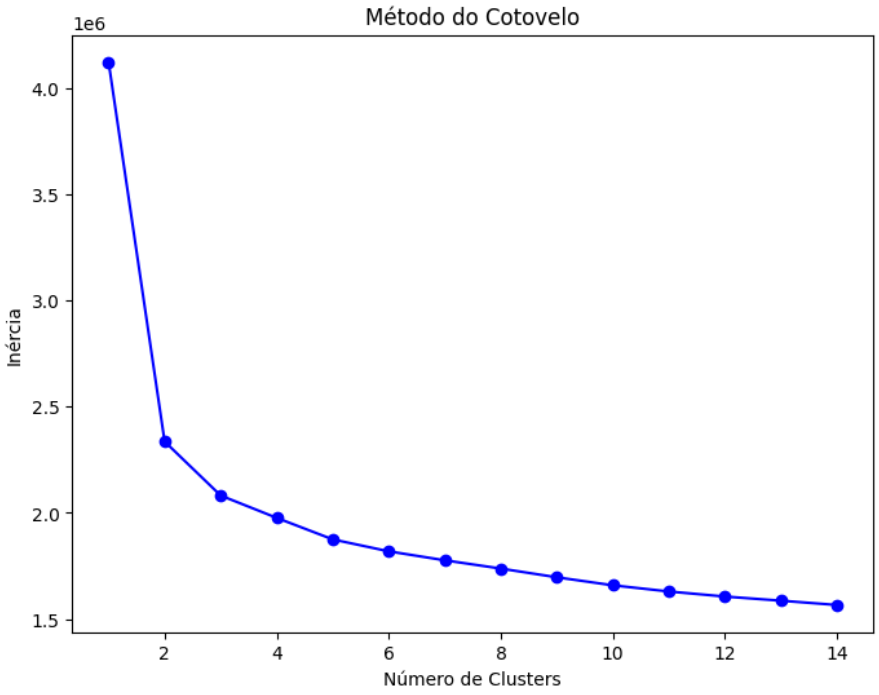
Execução 9: Inércia = 207069.80, Silhouette Score = 0.70

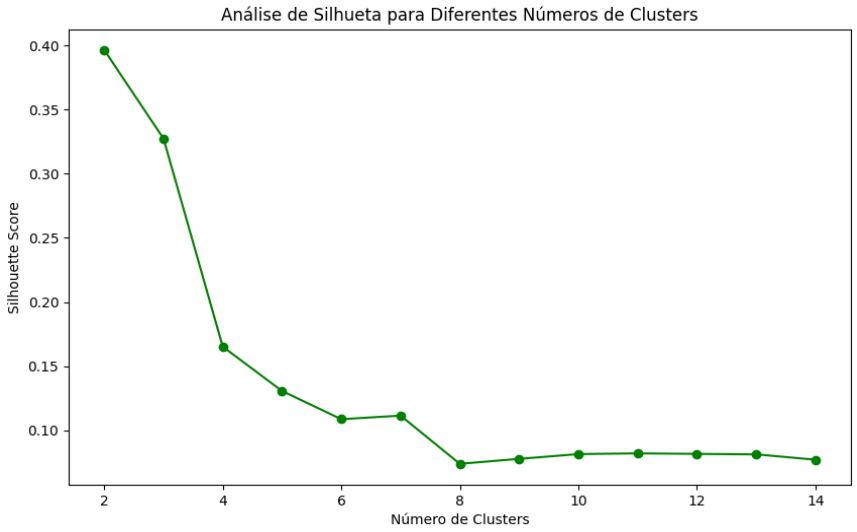
Execução 10: Inércia = 207069.80, Silhouette Score = 0.70

Média e Desvio Padrão das Metricas:

Inércia média: 207069.80, Desvio padrão 0.00

Silhouette Score médio: 0.70, Desvio padrão: 0.00





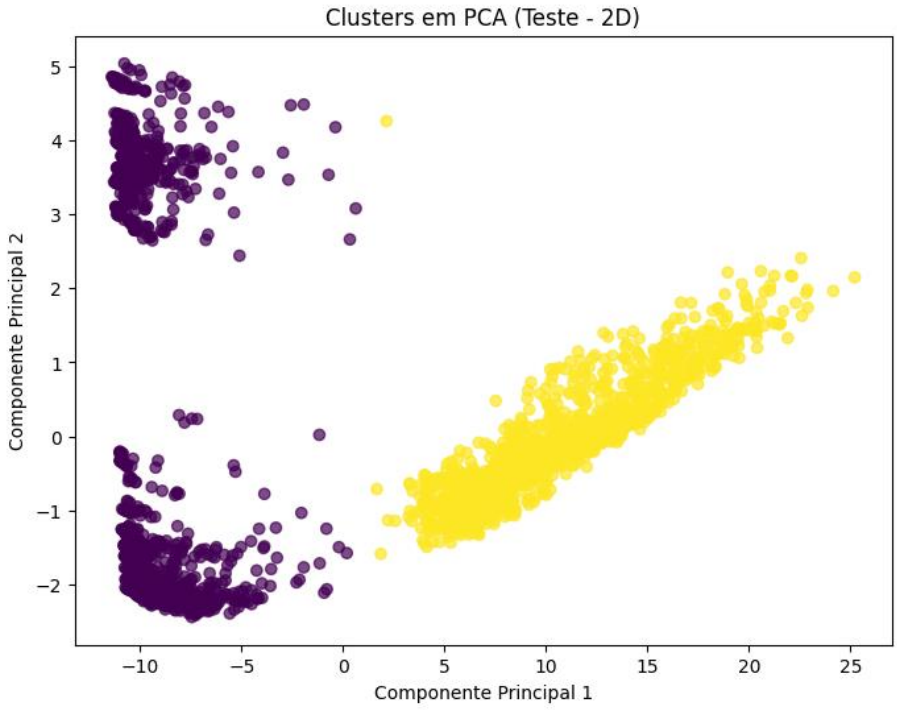
Observando esse gráfico de análise de silhueta para diferentes números de Clusters, percebemos que o número de clusters com a melhor pontuação de silhueta é 2 clusters, tendo aproximadamente 0.40 de Silhouette Score.

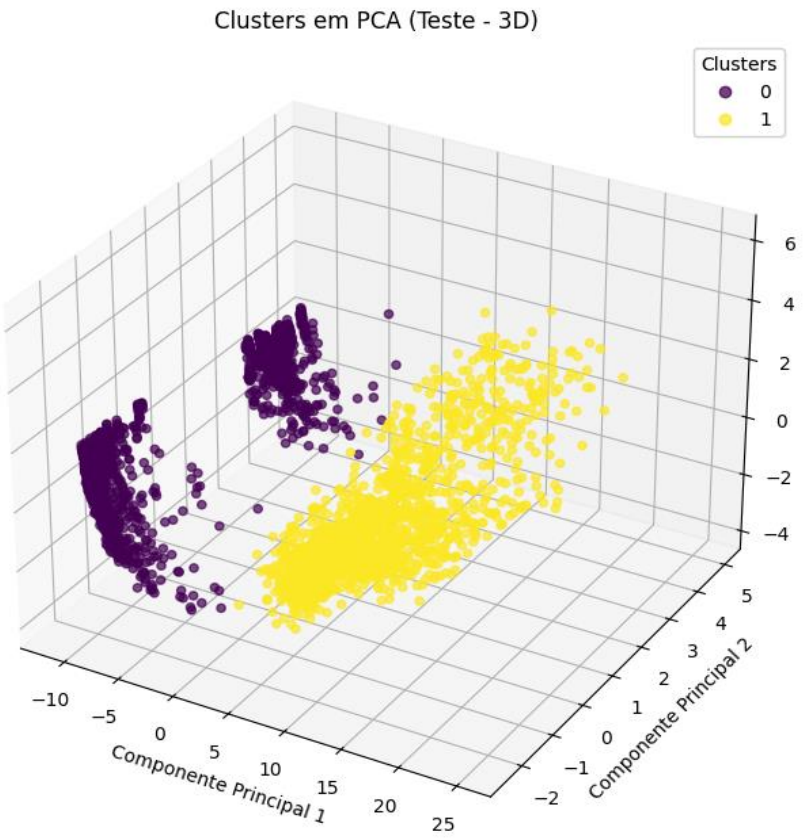
Então como valores finais de Inercia e Silhouette Score, temos:

Inércia (Treino Final): 207069.80464675982

Silhouette Score (Treino Final): 0.7029330769783646

Resultados de Clusters em PCA em 2D e 3D:





Discussão:

Tendo em vista todo o processo da metodologia do modelo, imaginamos que o projeto tenha apresentado resultados interessantes, talvez faltando alguns dados mais detalhados e mais desenvolvidos. Isso se deu por conta que optamos por apresentar resultados mais diretos e concretos ligados à técnica de aprendizado e ao banco de dados no qual foi escolhido para ser aprendido, além disso nos gráficos 2D e 3D dos clusters, houve alguns que acabaram se sobrepondo uns dos outros, mas isso se deu por conta do Silhouette Score baixo que foi obtido.

Conclusão e Trabalhos Futuros:

Esse projeto teve bastante importância para o aprendizado, por nos colocar para trabalhar com um conjunto de dados bem mais extenso e complexo comparado ao projeto anterior, além de nos fazer trabalhar com uma técnica de aprendizado diferente dessa vez, com o K-means, nos dando uma noção de como o campo da

ciência de dados é extremamente versátil quando se trata de formas para ler dados e informações. Como possíveis melhorias para o projeto, diria sobre que poderia haver uma exatidão na passagem dos dados do K-Clusters, tentando deixar um pouco mais claro para quem fosse utilizar dessas informações. Mas de um grande panorama geral, ainda imagino que o projeto conseguiu cumprir como o prometido de passar as informações dos dados da forma mais precisa possível utilizando do K-means.

Referências:

FCS Fonseca. WAR Beltrame. UFES Vitoria. **Aplicações Práticas dos Algoritmos de Clusterização Kmeans e Bisecting K-means.** 2010. Disponível em: https://www.researchgate.net/profile/Walber-Beltrame/publication/327121358_Aplicacoes_Praticas_dos_Algoritmos_de_Clusterizacao_K-means_e_Bisecting_K-means/links/5b7b53a6299bf1d5a718d785/Aplicacoes-Praticas-dos-Algoritmos-de-Clusterizacao-K-means-e-Bisecting-K-means.pdf. Acessado em: 30 nov. 2024.

SOUSA, Maria. **Uma análise do algoritmo K-means como introdução ao Aprendizado de Máquinas.** 2023. Disponível em: https://repositorio.uft.edu.br/handle/11612/4511. Acessado em: 30 nov. 2024.