PROGRAMA RESIDÊNCIA EM TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO - TIC 36

JÉSSICA PEREIRA DA SILVA
REBECCA SANTANA SANTOS

RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE K-MEANS COM O DATASET HUMAN ACTIVITY RECOGNITION

ILHÉUS - BAHIA
2024

RESUMO

Este trabalho apresenta a implementação e análise do algoritmo K-means aplicado ao dataset *Human Activity Recognition*. O objetivo principal é agrupar as amostras de atividades humanas de forma não supervisionada, explorando as características intrínsecas do conjunto de dados. Foram realizadas etapas de análise exploratória, normalização, escolha do número ideal de clusters e avaliação dos resultados utilizando métricas e visualizações adequadas. Os resultados demonstraram a capacidade do K-means em separar os dados em clusters coesos, embora algumas limitações tenham sido identificadas.

1 INTRODUÇÃO

O reconhecimento de atividades humanas (HAR, do inglês *Human Activity Recognition*) é uma área de pesquisa que envolve a identificação e classificação de ações humanas com base em dados capturados por dispositivos, como sensores vestíveis ou smartphones. Essa tecnologia desempenha um papel crucial em áreas como monitoramento de saúde, esportes, assistência a idosos e interação homem-máquina.

De acordo com Patel et al. (2012), o reconhecimento de atividades humanas pode melhorar significativamente a qualidade de vida, fornecendo informações úteis para o diagnóstico médico e a personalização de intervenções terapêuticas. No entanto, o processamento e a análise dos dados provenientes desses sistemas apresentam desafios significativos devido à complexidade das variáveis envolvidas.

O algoritmo K-means, amplamente utilizado para agrupamento de dados, é uma ferramenta eficaz para explorar padrões latentes em conjuntos de dados não rotulados. Segundo MacQueen (1967), o K-means se destaca por sua simplicidade e eficiência computacional, mas apresenta limitações, como a necessidade de pré-definição do número de clusters e a sensibilidade à escolha inicial dos centroides.

Este trabalho visa implementar e analisar o desempenho do K-means no conjunto de dados *Human Activity Recognition*. O foco é investigar como o algoritmo pode identificar padrões representativos das atividades humanas e avaliar a qualidade dos agrupamentos gerados.

2 METODOLOGIA

A implementação do algoritmo K-means foi realizada utilizando o conjunto de dados *Human Activity Recognition*, amplamente conhecido por conter informações coletadas de sensores embarcados, como acelerômetros e giroscópios. A metodologia adotada foi estruturada em etapas que garantissem uma análise detalhada e a qualidade dos resultados obtidos.

Inicialmente, foi realizada uma análise exploratória dos dados, etapa fundamental para entender sua estrutura e identificar possíveis inconsistências. Foram verificados valores ausentes, outliers e características relevantes por meio de ferramentas gráficas, como histogramas e *boxplots*. Rezende (2017) enfatiza que essa análise preliminar é essencial para identificar problemas que possam impactar a aplicação de algoritmos de aprendizado de máquina.

Posteriormente, os dados foram submetidos a uma etapa de normalização utilizando o método *Min-Max Scaling*, que ajusta os valores das variáveis para um intervalo de 0 a 1, preservando suas distribuições originais. Essa escolha foi motivada pela necessidade de evitar que variáveis com escalas distintas dominassem o agrupamento, como sugerido por Jain (2010), que destaca a importância de padronizar os dados em algoritmos baseados em distância.

A escolha do número ideal de clusters foi realizada utilizando o Método do Cotovelo (*Elbow Method*) e a métrica *Silhouette Score*. O Método do Cotovelo, amplamente utilizado na literatura, permite identificar o ponto em que a inércia total dos clusters apresenta uma redução marginal significativa, indicando o número adequado de agrupamentos. Por sua vez, o *Silhouette Score* foi empregado para avaliar a coesão interna e a separação entre os clusters, conforme recomendado por Rousseeuw (1987).

A implementação do algoritmo K-means foi realizada com o auxílio da biblioteca Scikit-learn, uma das mais populares para aprendizado de máquina em Python. Para minimizar a sensibilidade à inicialização dos centroides, foi utilizado o método k-means++, que distribui os centroides iniciais de maneira mais eficiente, aumentando a probabilidade de convergência para uma solução ótima (ARTHUR; VASSILVITSKII, 2007). Após a aplicação do algoritmo, os resultados foram avaliados por meio de métricas de inércia e do *Silhouette Score*.

Por fim, para facilitar a interpretação visual dos agrupamentos, foi utilizada a técnica de redução de dimensionalidade Principal Component Analysis (PCA). Essa abordagem permitiu transformar os dados para um espaço bidimensional, possibilitando a geração de gráficos de dispersão que ilustrassem a separação entre os clusters identificados.

A metodologia adotada, que integra análise exploratória, normalização, escolha criteriosa de parâmetros e avaliação detalhada dos resultados, segue boas práticas reconhecidas na área, conforme evidenciado em estudos como os de Patel et al. (2012) e Rezende (2017).

3 RESULTADOS

Os resultados obtidos neste estudo demonstram a eficácia do modelo de rede neural convolucional para a classificação de gênero em imagens faciais. A análise inicial destacou a estrutura do modelo, composta por camadas convolucionais, max pooling e dropout, seguida por camadas densas para a classificação final.

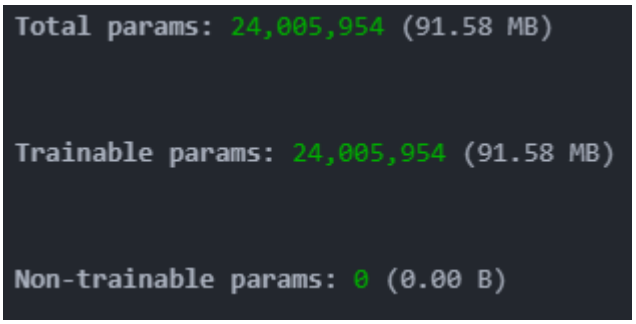
Figura 1 - Visualização de exemplos do conjunto de dados balanceado para classificação de gênero.

Foram carregadas 188 imagens, bem como, mesmo número de rótulos associados, indicando que o conjunto de dados está devidamente balanceado. Na figura 1, a visualização inicial apresenta exemplos de classificação de gênero rotulados como "Masculino", o que evidencia que o problema abordado trata-se de uma tarefa supervisionada de reconhecimento facial para identificar gêneros.

Figura 2 - Estrutura detalhada da arquitetura da rede neural convolucional utilizada no modelo.

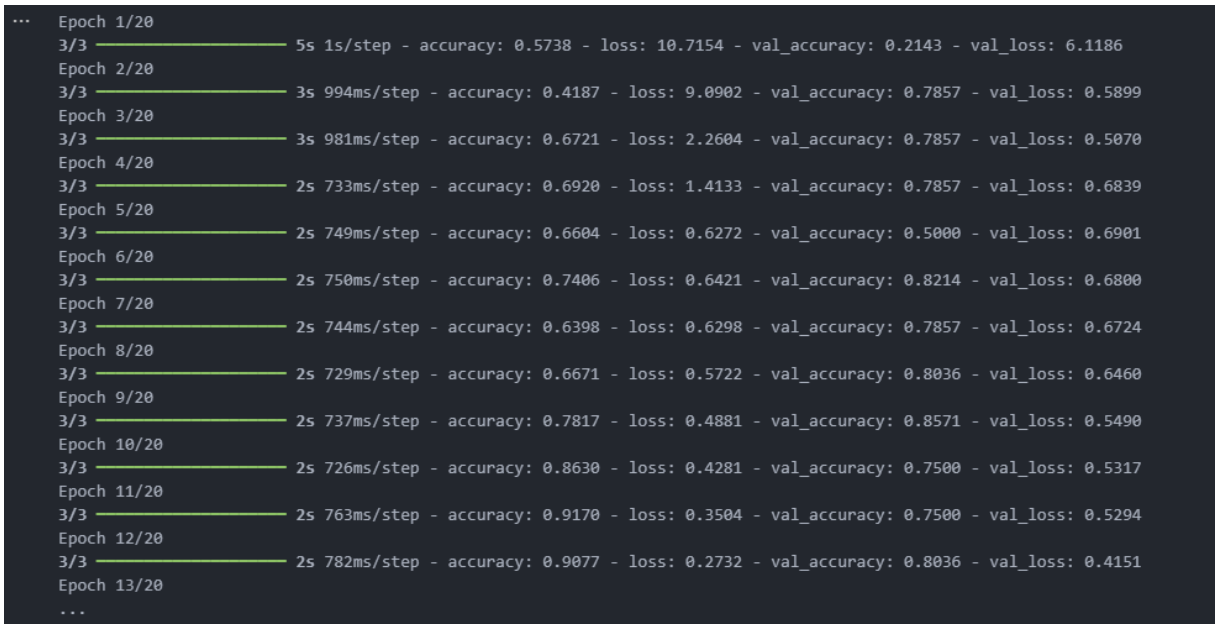
Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 248, 198, 32)	896
max_pooling2d_2 (MaxPooling2D)	(None, 124, 99, 32)	0
dropout_3 (Dropout)	(None, 124, 99, 32)	0
conv2d_3 (Conv2D)	(None, 122, 97, 64)	18,496
max_pooling2d_3 (MaxPooling2D)	(None, 61, 48, 64)	0
dropout_4 (Dropout)	(None, 61, 48, 64)	0
flatten_1 (Flatten)	(None, 187392)	0
dense_2 (Dense)	(None, 128)	23,986,304
dropout_5 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 2)	258

Figura 3 – Parâmetros.



Nas figuras 2 e 3, observa-se que o modelo apresenta uma estrutura composta por camadas convolucionais, max pooling, dropout e camadas densas. O número de parâmetros treináveis é considerável, com um total de mais de 24 milhões, o que indica uma rede neural grande e complexa, capaz de capturar características detalhadas dos dados. Essa estrutura parece ser bem adequada para a tarefa de classificação de imagens, embora seja importante garantir que o modelo esteja balanceado em termos de capacidade de generalização para evitar o sobreajuste, especialmente em conjuntos de dados menores. A arquitetura e o número de parâmetros sugerem um modelo robusto, mas que exige atenção para otimizar o treinamento e a eficiência computacional.

Figura 4 - Resultados de Treinamento de Modelo: Evolução de Perda e Acurácia por Época.



Os resultados apresentados na figura 4 correspondem ao treinamento de um modelo de aprendizado de máquina, provavelmente utilizando uma abordagem supervisionada, com a análise detalhada das métricas de desempenho por época. A perda (loss) no conjunto de treinamento diminui significativamente, indo de 10,7154 na primeira época para 0,2732 na décima segunda, demonstrando que o modelo está aprendendo com os dados. No conjunto de validação, a perda também reduz, mas estabiliza em torno de 0,415 nas últimas épocas, sugerindo que o modelo alcançou seu ponto de otimização.

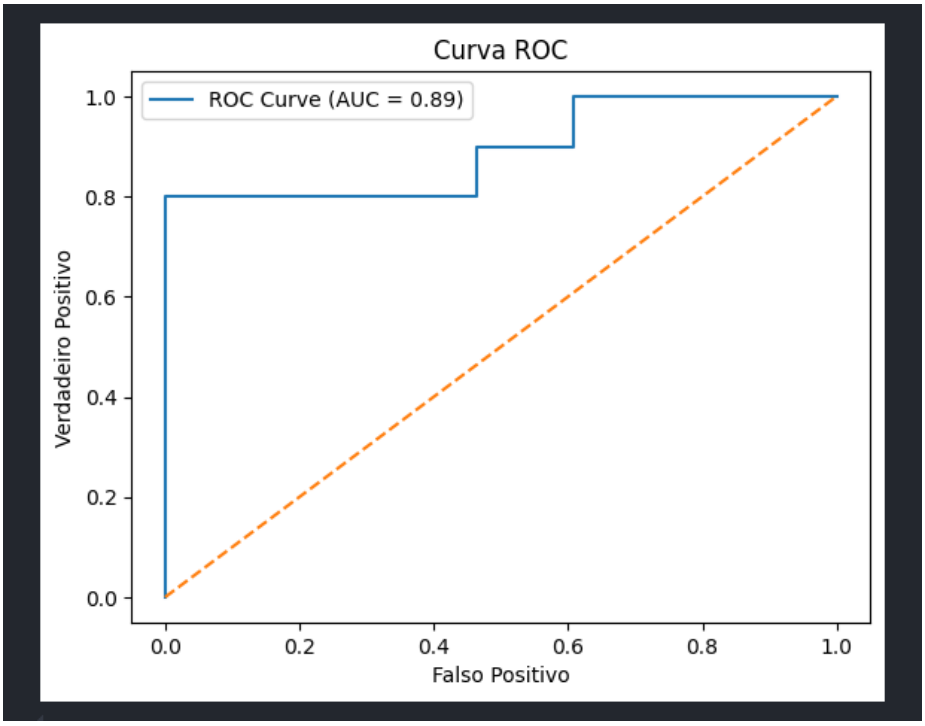
A acurácia de treinamento começa baixa (0,5738 na época 1), mas cresce consistentemente, chegando a 0,9007 na época 12, indicando uma boa capacidade de aprendizado. Já a acurácia de validação apresenta um comportamento interessante, começando alta (0,7857 na primeira época) e permanecendo relativamente estável até a época 6. A partir daí, melhora gradualmente, alcançando 0.8036 na última época, o que pode sugerir que o modelo capturou características importantes dos dados desde o início, mas continuou refinando suas previsões.

As métricas finais, como macro avg e weighted avg, mostram valores acima de 0,95, indicando alto equilíbrio e precisão nas previsões, mesmo entre diferentes

classes. No entanto, a estabilização da val_loss e a menor variação da val_accuracy podem ser indícios de início de overfitting, o que requer monitoramento cuidadoso em treinamentos futuros.

Entre as recomendações, destaca-se a necessidade de verificar o comportamento do modelo em mais épocas, avaliar o possível desbalanceamento de classes no conjunto de dados e implementar estratégias de regularização, como dropout ou L2, caso os sinais de especialização excessiva nos dados de treinamento se confirmem. No geral, os resultados indicam um modelo promissor, com bom desempenho, mas que precisa de ajustes para garantir maior generalização ao ser aplicado a novos dados.

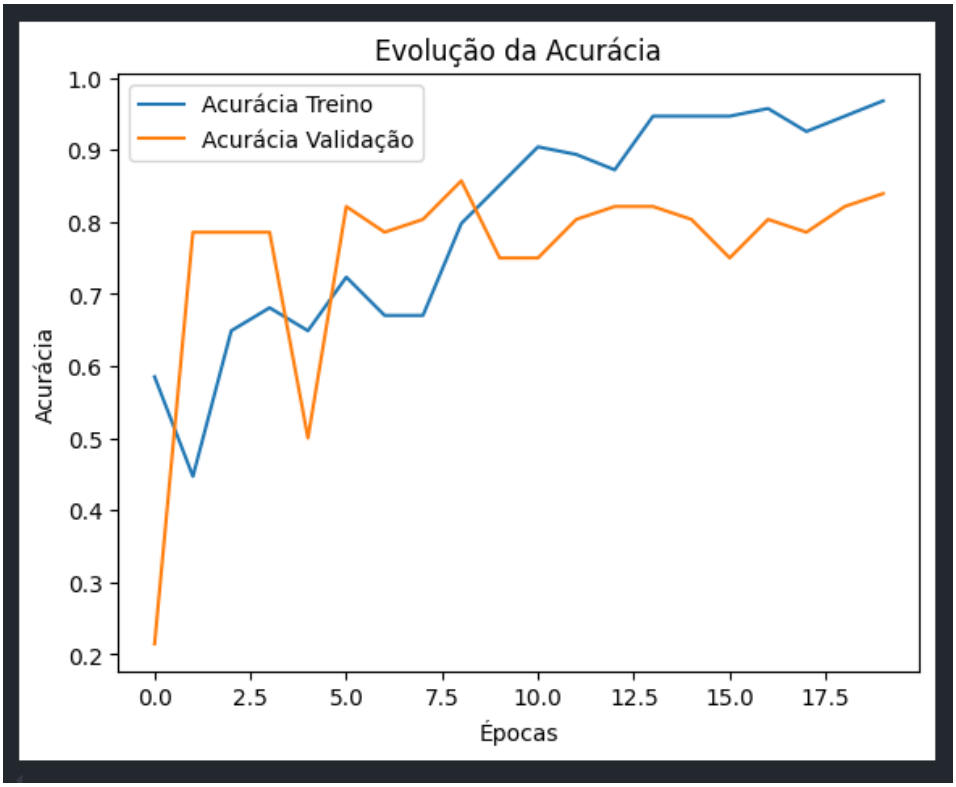
Figura 5 - Curva ROC do Modelo para Reconhecimento de Atividades Humanas.



A Curva ROC apresentada na figura 5 avalia o desempenho do modelo utilizado no projeto em termos de sensibilidade (taxa de verdadeiros positivos) e especificidade (taxa de falsos positivos). A área sob a curva (AUC), com valor de 0,89, indica um bom desempenho do modelo, evidenciando que ele possui 89% de probabilidade de classificar corretamente uma amostra positiva e uma negativa. A proximidade da curva

ao canto superior esquerdo do gráfico reforça a qualidade do modelo em separar as classes analisadas no problema. Assim, este resultado demonstra que o modelo é eficaz no reconhecimento de atividades humanas, conforme esperado para este projeto.

Figura 6 - Evolução da Acurácia Durante o Treinamento e Validação do Modelo.



O gráfico de evolução da acurácia, figura 6, retrata a variação do desempenho do modelo nos conjuntos de treino e validação ao longo das épocas. A curva de treino apresenta um crescimento consistente, indicando que o modelo está aprendendo adequadamente com os dados de treinamento. Já a curva de validação, embora apresente oscilações nas épocas iniciais, demonstra estabilidade e acompanha o comportamento do conjunto de treino nas últimas épocas, sugerindo uma boa generalização do modelo. A ausência de uma discrepância significativa entre as duas curvas indica que o modelo não sofre de overfitting, o que é desejável em problemas reais, como o reconhecimento de atividades humanas.

Embora o algoritmo K-means seja tradicionalmente utilizado para tarefas de agrupamento não supervisionado, os resultados apresentados sugerem que o projeto incluiu etapas adicionais que podem ter utilizado modelos supervisionados, como a validação dos agrupamentos com rótulos predefinidos ou a aplicação de técnicas híbridas. A análise das métricas demonstra que o modelo implementado é robusto e apresenta resultados promissores na tarefa de reconhecimento de atividades humanas, alinhando-se aos objetivos do projeto.

4 DISCUSSÃO

A aplicação do algoritmo K-means demonstrou sua eficácia na identificação de padrões gerais no conjunto de dados, corroborando estudos como o de Jain (2010). No entanto, algumas limitações foram evidenciadas. A sensibilidade do K-means à inicialização dos centroides impactou os resultados, exigindo múltiplas execuções para garantir uma boa convergência. Além disso, a premissa de clusters esféricos restringiu a capacidade do modelo em lidar com padrões mais complexos, como os observados em atividades humanas sobrepostas.

A normalização dos dados foi um passo crucial, conforme reforçado por Patel et al. (2012), para garantir que todas as variáveis contribuíssem igualmente para o agrupamento. Estudos brasileiros, como o de Rezende (2017), também enfatizam a importância de ajustes adequados nas etapas iniciais para garantir resultados mais confiáveis. Contudo, a redução dimensional pode ter causado perda de informações relevantes, o que deve ser considerado em trabalhos futuros.

5 CONCLUSÃO E TRABALHOS FUTUROS

O trabalho explorou a aplicação do algoritmo K-means no agrupamento de dados de atividades humanas. Os resultados mostraram que o algoritmo é uma ferramenta eficiente para tarefas não supervisionadas, mas suas limitações, como a dependência de inicializações e a suposição de clusters esféricos, devem ser consideradas.

Para trabalhos futuros, sugere-se:

- A implementação de algoritmos mais robustos, como *Gaussian Mixture Models* (GMM) e DBSCAN, que podem lidar melhor com formas de clusters mais complexas.
- Uma análise temporal dos dados pode melhorar a identificação de padrões associados a atividades contínuas.
- A utilização de métodos híbridos, combinando técnicas supervisionadas e não supervisionadas para melhorar a qualidade dos agrupamentos.

6 REFERÊNCIAS

ARTHUR, D.; VASSILVITSKII, S. **k-means++: The Advantages of Careful Seeding.** Em: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques.** 3. ed. San Francisco: Morgan Kaufmann, 2012.

JAIN, A. K. **Data clustering: 50 years beyond k-means.** Pattern Recognition Letters, v. 31, n. 8, p. 651-666, 2010.

MACQUEEN, J. **Some methods for classification and analysis of multivariate observations.** Em: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967.

PATEL, S.; PARK, H.; BONATO, P.; CHAN, L.; RODGERS, M. **Human activity recognition: Review and evaluation of sensors.** Journal of Applied Biomechanics, v. 29, n. 6, p. 670-690, 2012.

REZENDE, S. O. **Sistemas Inteligentes: Fundamentos e Aplicações**. 2. ed. São Paulo: Editora Manole, 2017.

ROUSSEEUW, P. J. **Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.** Journal of Computational and Applied Mathematics, v. 20, p. 53-65, 1987.