Instituto Federal Do Mato Grosso(IFMT)-Campus Cuiabá - Cel. Octayde Jorge da Silva

RELATÓRIO TÉCNICO: ANÁLISE COMPARATIVA DE ALGORITMOS DE CLUSTERIZAÇÃO

Beatriz Aparecida Dutra Da Silva
Namem Rachid Jaudy Neto
Prof. Me Matheus Candido

Cuiabá- 10 Novembro 2025

RESUMO

Este relatório detalha os procedimentos e resultados de duas atividades de mineração de dados focadas na clusterização. O objetivo principal foi comparar o desempenho dos algoritmos KMeans e Clusterização Hierárquica Aglomerativa em dois cenários distintos. A Atividade 1 utilizou um conjunto de dados não rotulado, aplicando normalização e métricas de validação interna (Coeficiente de Silhouette, Índice de Davies-Bouldin) para determinar o número ótimo de clusters, identificado como k=4. A Atividade 2 utilizou um conjunto de dados binário com rótulos verdadeiros, permitindo uma validação externa (Adjusted Rand Score, Jaccard, Pureza). Nesta segunda atividade, o número ideal de clusters foi k=3, e os perfis de grupo resultantes foram analisados e interpretados. Ambos os algoritmos foram avaliados quanto à sua capacidade de formar grupos coesos e distintos em cada cenário.

Palavras-chave: Clusterização. KMeans. Clusterização Aglomerativa. Validação Interna. Validação Externa. Mineração de Dados.

SUMÁRIO

1 INTRODUÇÃO

Este documento apresenta um relatório técnico sobre duas atividades práticas de clusterização de dados. O objetivo central foi conduzir uma análise comparativa entre dois dos algoritmos de agrupamento mais comuns: o **KMeans**, um método particional, e a **Clusterização Hierárquica Aglomerativa**, um método hierárquico.

As atividades foram estruturadas para abordar dois cenários fundamentais na análise de dados:

1. **Validação Interna:** Análise de um conjunto de dados sem rótulos pré-definidos (data_1.csv), onde a qualidade dos clusters é avaliada por métricas intrínsecas de coesão e separação.
2. **Validação Externa:** Análise de um conjunto de dados com rótulos verdadeiros conhecidos (data_2.csv), permitindo uma avaliação supervisionada da precisão dos algoritmos em replicar as classes existentes.

Este relatório detalha as metodologias aplicadas em cada atividade, os resultados obtidos e as conclusões sobre o desempenho comparativo dos algoritmos.

2 ATIVIDADE 1: CLUSTERIZAÇÃO COM VALIDAÇÃO INTERNA

A primeira atividade focou na aplicação de técnicas de clusterização em um cenário não supervisionado.

2.1 Metodologia

A metodologia desta atividade seguiu quatro etapas principais:

1. **Carregamento e Pré-processamento:** Os dados foram carregados do arquivo data_1.csv. Foi aplicado um pré-processamento de normalização usando StandardScaler do Scikit-learn, garantindo que todas as variáveis tivessem média zero e variância unitária.
2. **Modelagem:** Os algoritmos KMeans e AgglomerativeClustering foram aplicados aos dados normalizados.
3. **Avaliação de 'k':** Ambos os modelos foram testados iterativamente para um número de clusters (k) variando de 2 a 10.
4. **Métricas de Validação:** A qualidade dos clusters foi avaliada usando métricas de validação interna:
    - **SSE (Inércia):** Utilizado para o "Método do Cotovelo" no KMeans.
    - **Coeficiente de Silhouette:** Mede a separação e coesão dos clusters (valores mais próximos de 1 são melhores).
    - **Índice de Davies-Bouldin:** Mede a similaridade entre clusters (valores mais próximos de 0 são melhores).
5. **Visualização:** Os resultados dos clusters foram visualizados graficamente usando a técnica de redução de dimensionalidade PCA (Análise de Componentes Principais).

2.2 Resultados e Discussão

A análise das métricas nos gráficos (SSE, Silhouette e Davies-Bouldin) levou à seleção de k=4 como o número ótimo de clusters para a análise detalhada.

- **Quantidade de Clusters:** O número de clusters definido para a análise final foi 4.
- **Distribuição de Pontos:** Foi analisada a contagem de pontos em cada um dos 4 clusters para ambos os métodos.
- **Métricas para k=4:** Os scores de Silhouette e Davies-Bouldin foram registrados.
- **Comparação de Performance:** Os scripts realizaram uma comparação direta entre os valores das métricas do KMeans e do AgglomerativeClustering, concluindo se houve ou não diferença significativa de performance entre os algoritmos para este conjunto de dados.

3 ATIVIDADE 2: CLUSTERIZAÇÃO COM VALIDAÇÃO EXTERNA

A segunda atividade explorou um cenário onde os rótulos verdadeiros dos dados eram conhecidos, permitindo uma avaliação de desempenho supervisionada (validação externa).

3.1 Metodologia

1. **Carregamento de Dados:** Os dados foram lidos do arquivo data_2.csv. As colunas de *features* foram separadas da coluna label (rótulos verdadeiros). Nenhum pré-processamento de normalização foi aplicado, dado o-caráter binário dos dados.
2. **Modelagem:** Os algoritmos KMeans e AgglomerativeClustering foram aplicados.
3. **Avaliação de 'k':** Os modelos foram testados para k variando de 2 a 10.
4. **Métricas de Validação:** Foram utilizadas métricas de validação externa para comparar os clusters gerados (cluster_labels) com os rótulos verdadeiros (true_labels):
    - **Adjusted Rand Score (Rand Ajustado)**
    - **Jaccard Score**
    - **Fowlkes-Mallows Score**
    - **Pureza (Purity):** Uma função customizada foi implementada para calcular a pureza média dos clusters.
5. **Seleção do Melhor 'k':** O "melhor k" foi determinado identificando o valor de k que maximizou o Adjusted Rand Score para cada algoritmo.

3.2 Resultados e Discussão

A análise dos gráficos de métricas e a seleção do melhor 'k' indicaram que k=3 era o número ideal de clusters, correspondendo aos rótulos reais.

- **Quantidade de Clusters:** O número de clusters usado para a análise final foi 3.
- **Métricas para k=3:** Os valores para Rand Ajustado, Jaccard, Fowlkes-Mallows e Pureza foram calculados e comparados.
- **Comparação de Performance:** Os resultados dos dois algoritmos para k=3 foram comparados, e o script avaliou se existiam diferenças perceptíveis em seus desempenhos.
- **Análise de Perfis (k=3):** A etapa mais relevante desta atividade foi a análise das características de cada grupo. Ao calcular a média das *features* para cada cluster, foi possível interpretar e definir perfis claros para os três grupos identificados:
    - **Perfil 1 (Agglo 1 / KMeans 0):** Homens Jovens (19-29), Solteiros, Com Filhos. Renda Média-Alta (4-8k). Local: Perto.
    - **Perfil 2 (Agglo 0 / KMeans 1):** Mulheres Maduras (30+), Casadas, Sem Filhos. Renda Média-Alta (4-8k). Local: Perto.
    - **Perfil 3 (Agglo 2 / KMeans 2):** Mulheres Jovens (19-29), Solteiras, Com Filhos. Renda Alta (8k+). Local: Longe.

4 CONCLUSÃO

Este relatório documentou a execução de duas atividades de clusterização, comparando os algoritmos KMeans e Hierárquico Aglomerativo em cenários de validação interna e externa.

Na Atividade 1, foi demonstrada a importância do pré-processamento (normalização) e o uso de métricas internas (Silhouette, Davies-Bouldin) para identificar uma estrutura de agrupamento ótima (k=4) em dados não rotulados.

Na Atividade 2, a disponibilidade de rótulos verdadeiros permitiu uma avaliação mais direta da eficácia dos algoritmos. Ambos os métodos foram capazes de identificar corretamente os três perfis de clientes subjacentes no conjunto de dados, conforme validado pelas métricas de Rand, Jaccard e Pureza. A análise final dos perfis em k=3 demonstrou a capacidade dos algoritmos não apenas de agrupar, mas de fornecer *insights* de negócio interpretáveis.

Conclui-se que ambos os algoritmos são ferramentas eficazes, e a escolha entre eles, bem como o método de validação, depende fundamentalmente da natureza dos dados e da disponibilidade de rótulos verdadeiros.