Trabalho prático 1 - Mineração de Dados

Arthur Antunes

UFSJ
Universidade Federal de São João del-Rei

Introdução

- Objetivo do trabalho: Explorar um conjunto de dados de músicas do Spotify (músicas populares de 2023) para extrair características e padrões.
- Ferramentas utilizadas: Pandas, Matplotlib, Seaborn e Scikit-Learn (Python).
- Conjunto de dados: Inclui métricas de dançabilidade, energia, popularidade e outras características musicais.

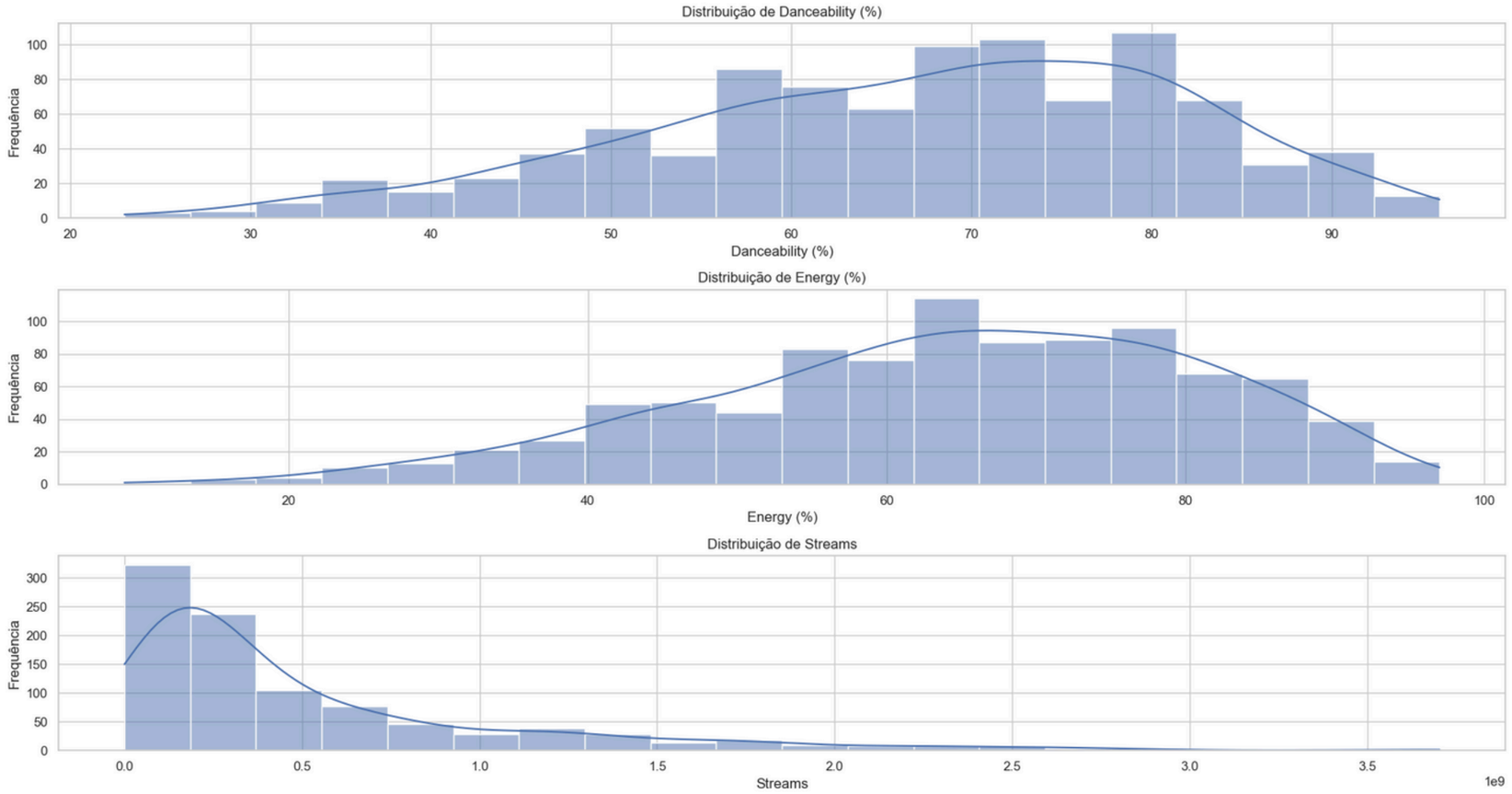
UFSJ
Universidade Federal de São João del-Rei

Exploração Inicial e Pré-processamento

- Análise Exploratória de Dados (EDA): Gráficos para distribuição das variáveis (danceability, energy, streams, etc.).
- Principais observações: A maioria das músicas são dançantes e enérgicas; distribuição de streams é assimétrica, com algumas músicas dominando.
- Tratamento de valores ausentes: Remoção e substituição de dados nulos em variáveis como streams.
- Conversão de tipos: Conversão de valores numéricos em formato texto.
- Normalização: Ajuste das variáveis numéricas para uma escala comum, necessário para alguns algoritmos de mineração.

UFSJ
Universidade Federal de São João del-Rei

Exploração Inicial e Pré processamento

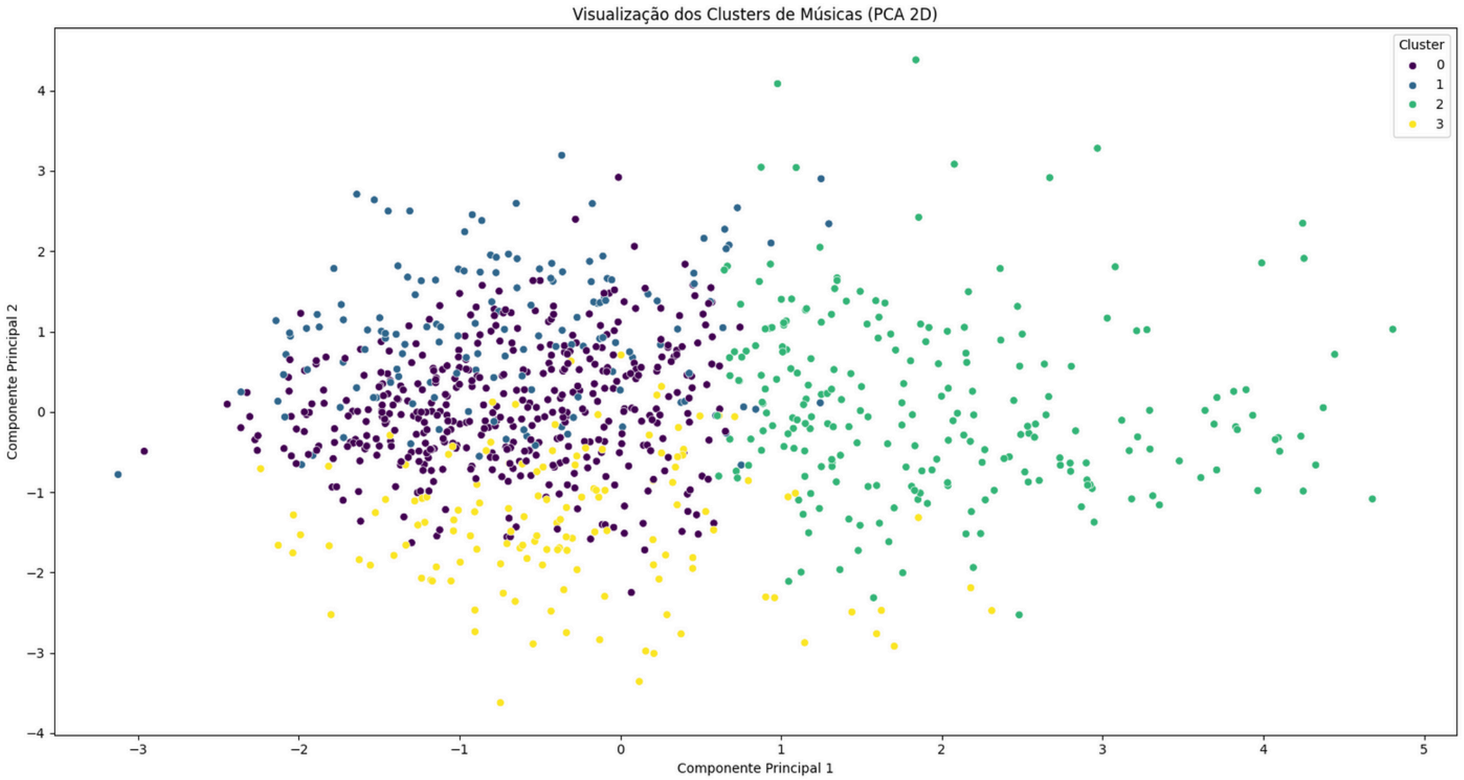


Plot de Distribuição de Danceability, Energy e Streams

Universidade Federal de São João del-Rei

Algoritmo de Agrupamento (K-means Clustering)

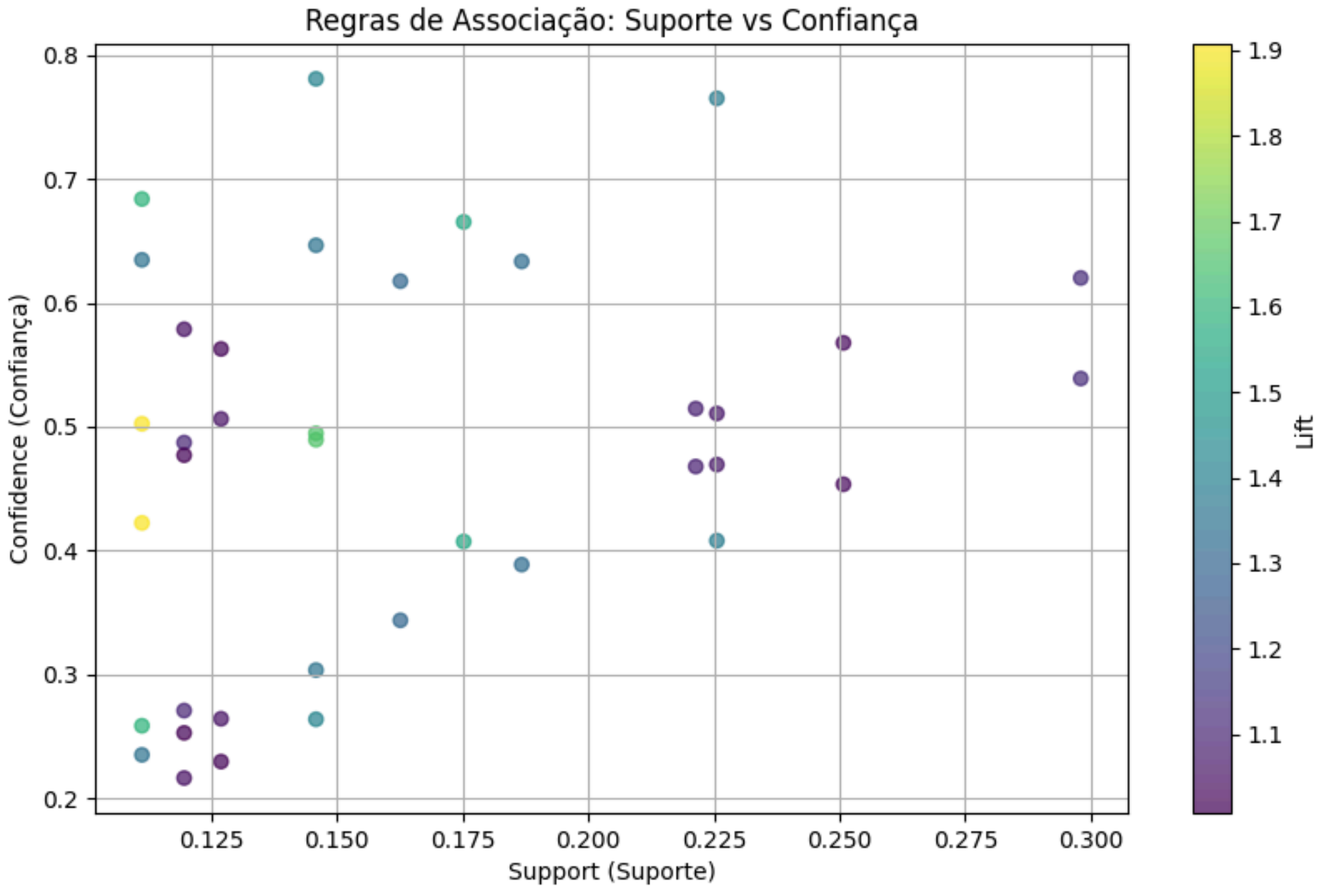
- **Objetivo**: Agrupar músicas com características semelhantes.
- **Configuração do K-means**: 4 clusters definidos, baseados em variáveis como danceability, energy e valence.
- **Interpretação dos Clusters:** Identificação de grupos distintos, por exemplo, músicas dançantes e energéticas em um cluster e músicas acústicas em outro.
- **Gráfico de dispersão:** Representação dos clusters com diferentes cores.
- **Análise**: Clusters mostram músicas com características similares, útil para recomendações de playlists.



Universidade Federal de São João del-Rei

Algoritmo de Associação (Apriori)

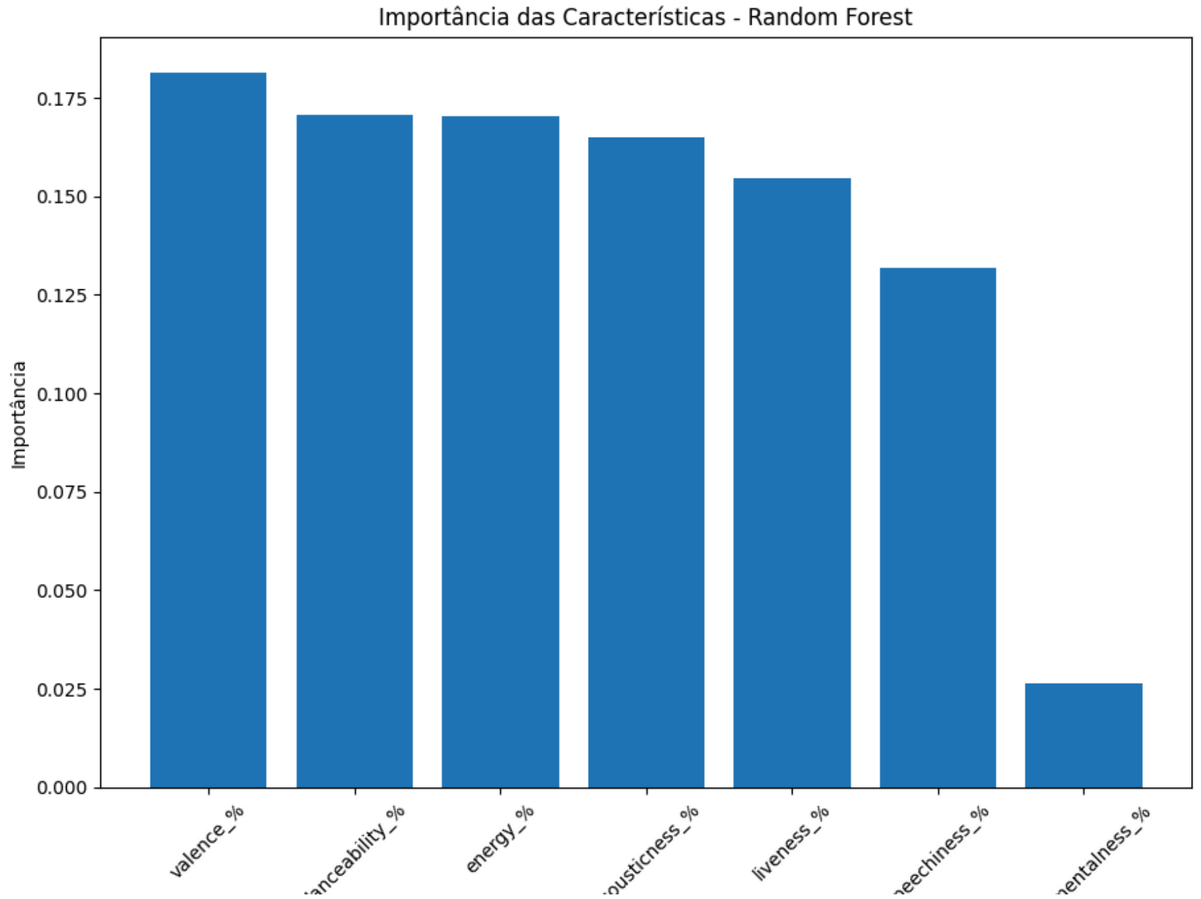
- Objetivo: Identificar padrões e correlações entre características musicais.
- Funcionamento: Algoritmo Apriori identifica conjuntos frequentes e regras de associação, como "alta dançabilidade" associada a "alta energia".
- Regras significativas: Exemplo - músicas com alta danceability têm alta energy em 80% dos casos.
- Gráfico de suporte, confiança e lift: Mostra a qualidade e a relevância das regras encontradas.
- Análise: Regras revelam correlações importantes para personalizar recomendações musicais.



Universidade Federal de São João del-Rei

Algoritmo de Classificação (Support Vector Machine - SVM)

- **Objetivo:** Classificar músicas em categorias de energia (alta ou baixa).
- **Processo:** Treinamento com variáveis como danceability e valence; separação com hiperplano e uso de kernel RBF para separação não-linear.
- **Resultados:** Boa acurácia para categorizar músicas em níveis de energia.
- **Métricas:** Acurácia, Precisão, Recall e F1-Score.
- **Análise:** O modelo SVM mostrou-se eficaz para prever se uma música pertence a uma categoria de alta ou baixa energia.



Universidade Federal de São João del-Rei

Conclusão

- Resumo: Técnicas de mineração de dados ajudaram a identificar padrões e características em músicas populares.
- Aplicabilidade: Insights úteis para sistemas de recomendação, como o do Spotify.
- Relevância: Mineração de dados aprimora a experiência de usuários em plataformas de streaming, permitindo recomendações mais precisas.

UFSJ
Universidade Federal de São João del-Rei

Obrigado!

- [1]Karydis, Ioannis, et al. "Musical track popularity mining dataset: Extension & experimentation."Neurocomputing 280 (2018): 76-85.
- [2]Silva, Mariana O., Laıs M. Rocha, and Mirella M. Moro. "MusicOSet: An enhanced open dataset for music data mining."XXXII Simpósio Brasileiro de Banco de Dados: Dataset Showcase Workshop, SBBD. 2019.
- [3]DA, P.; MOREIRA, C. UNIVERSIDADE FEDERAL DO PARANÁ MINERAÇÃO DE DADOS APLICADA À CLASSIFICAÇÃO AUTOMÁTICA DE GÊNEROS MUSICAIS CURITIBA 2017.
- [4]MacQueen, J. B. (1967). "Some Methods for Classification and Analysis of Multivariate Observations". Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.
- [5]Goethals, B., & Zaki, M. J. (2003). "Advances in frequent itemset mining implementations: report on FIMI'03 workshop". ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 109–117.
- [6]Savasere, A., Omiecinski, E., & Navathe, S. (1995). "An efficient algorithm for mining association rules in large databases". Proceedings of the 21st International Conference on Very Large Data Bases (VLDB), pp. 432–444.
[7]Cortes, C., & Vapnik, V. (1995). "Support-vector networks". Machine Learning, vol. 20, no. 3, pp. 273–297.

UFSJ

Universidade Federal de São João del-Rei