Universidade Federal
de São João del-Rei

TRABALHO PRÁTICO 1

Arthur Antunes Santos Silva

Este trabalho prático tem por objetivo exercitar conceitos e práticas iniciais da disciplina Mineração de dados. Foi escolhido um conjunto de dados que foi estudado e aplicado técnicas de mineração de dados.

São João del Rei

Novembro de 2024

Sumário

1 Introdução

O objetivo deste trabalho é realizar uma análise de mineração de dados em um conjunto de dados de músicas disponíveis no Spotify(músicas mais populares de 2023), buscando extrair informações valiosas sobre as características das faixas e agrupá-las de acordo com sua similaridade. Para isso, foram aplicadas técnicas de pré-processamento, visualização exploratória, análise de correlação e agrupamento. As técnicas foram implementadas utilizando bibliotecas populares de Python como pandas, matplotlib, seaborn e scikit-learn.

O conjunto de dados contém informações sobre as principais músicas no Spotify em 2023, além de características musicais e métricas de popularidade em várias plataformas de streaming. Abaixo estão as variáveis presentes no conjunto de dados que foram usadas durante o estudo que será descrito neste relatório:

- danceability_%: Pontuação de dançabilidade, expressa em percentual.

- energy_%: Nível de energia da música, medido em percentual.

- valence_%: Percentual que reflete o quão feliz ou positivo é o som da música.

- acousticness_%: Quantidade de elementos acústicos na música, também em percentual.

- instrumentalness_%: Indicador de quão instrumental é a música (percentual de vocais ausentes).

- liveness_%: Grau de vivacidade, ou seja, como a música soa "ao vivo".

- speechiness_%: Percentual que mede a presença de palavras faladas (quanto maior, mais falada é a música).

- streams: Número total de streams (reproduções) que a música acumulou no Spotify.

- mode: Modo da música (se é maior ou menor).

- key: Tom da música (representado por uma nota, como C, D, E).

A análise de dados de músicas não apenas revela informações sobre as características musicais, mas também fornece insights valiosos para o desenvolvimento de sistemas de recomendação personalizados, amplamente utilizados em plataformas de streaming como o Spotify. Utilizando técnicas de agrupamento, como o K-means, é possível segmentar músicas em clusters com características similares, facilitando a recomendação de faixas que atendem aos interesses específicos dos usuários. Da mesma forma, as regras de associação permitem identificar padrões comuns entre as faixas — como a alta probabilidade de músicas com alta dançabilidade possuírem também alta energia — que são fundamentais para melhorar a experiência do usuário. Com o uso de técnicas de classificação, como o Support Vector Machine (SVM), é possível categorizar músicas conforme sua energia, ajudando a direcionar recomendações para diferentes contextos (por exemplo, sugerindo músicas energéticas para atividades físicas). Esse tipo de análise é um componente essencial para sistemas de recomendação eficientes, que personalizam a experiência do usuário ao capturar e prever preferências baseadas em padrões detectados nos dados.

2 Exploração Inicial e Pré-processamento

Antes de aplicar qualquer técnica de mineração de dados, foi feita a análise exploratória de dados (EDA), para fornecer uma visão geral das distribuições das variáveis e ajudar na identificação de possíveis problemas, como valores ausentes ou tipos de dados incorretos.

2.1 Análise Exploratória de Dados (EDA)

Na fase inicial de EDA, foram realizadas as seguintes tarefas:

- Distribuições de Variáveis Numéricas: Foram gerados gráficos de distribuição para variáveis-chave, como danceability_%, energy_%, e streams, para entender a dispersão dos valores. Essas visualizações são úteis para identificar assimetrias e compreender quais faixas de valores são mais comuns.

- Análise de Variáveis Categóricas: A seguir foi exibido a frequência de variáveis categóricas como key (chave musical) e mode (modo maior ou menor), o que ajuda a entender quais características musicais são mais comuns no conjunto de dados.

A partir desses gráficos foi observado que a maioria das músicas do conjunto são em sua maioria dançantes e energéticas, e que a distribuição de streams é altamente assimétrica, com poucas músicas acumulando um número extremamente elevado de streams, o que é comum em plataformas de streaming.

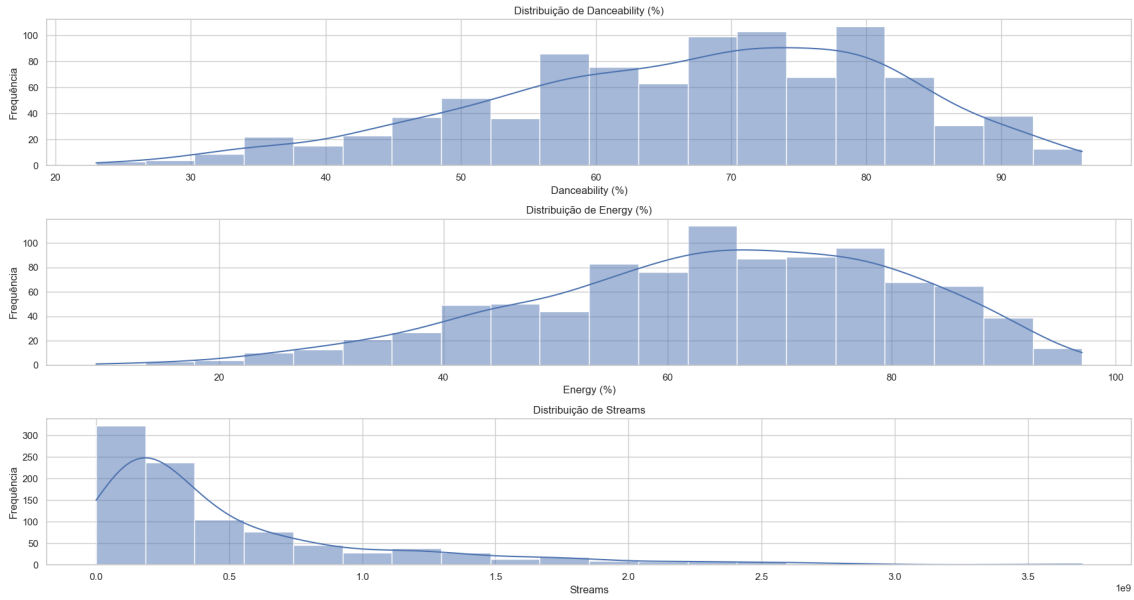

Figura 1: Plot de Distribuição de Danceability, Energy e Streams

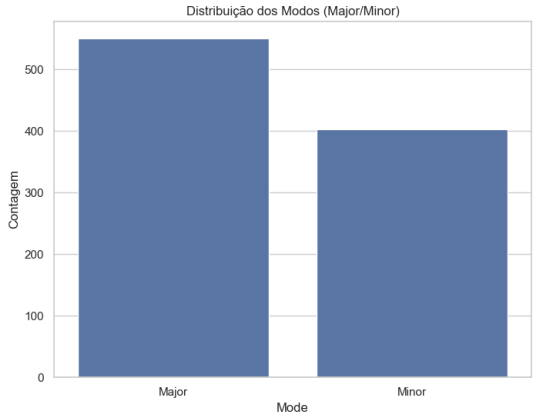

Figura 2: Plot de Distribuição de Distribuição dos Modos (Major/Minor)

2.2 Técnicas de Pré-processamento

Foi necessário realizar as seguintes etapas de pré-processamento antes que fosse possível aplicar os algoritmos de mineração:

- Tratamento de valores ausentes: Algumas colunas, como streams, tinham valores nulos, que foram preenchidos ou removidos. Isso garante que os algoritmos possam ser aplicados sem erros.

- Conversão de Tipos: Colunas que armazenavam valores numéricos como texto (por exemplo, streams) foram convertidas para o tipo float após remover formatações inadequadas, como vírgulas.

- Normalização dos Dados: Variáveis numéricas foram normalizadas usando o StandardScaler, que ajusta todas as variáveis para uma escala comum (média 0 e desvio padrão 1). Essa etapa foi feita para algoritmos que dependem de distâncias, como o K-means.

3 Técnicas de Mineração de Dados Utilizadas

3.1 Algoritmo de Agrupamento: K-means Clustering

3.1.1 Objetivo

O objetivo principal do K-means clustering é segmentar músicas com características semelhantes em grupos distintos. Este algoritmo utiliza as características musicais numéricas, como danceability_%, energy_%, e valence_%, para agrupar as músicas em diferentes clusters.

3.1.2 Funcionamento do K-means

O algoritmo K-means tenta dividir os dados em K grupos, onde K é definido previamente (neste caso K = 4 clusters). O algoritmo segue os seguintes passos:

1. Inicialização: O algoritmo inicia escolhendo aleatoriamente K pontos no espaço de dados para servir como os centros iniciais dos clusters (os centróides). Estes centróides servem como pontos de referência iniciais para medir a proximidade das músicas.

2. Atribuição: Cada música é atribuída ao cluster cujo centróide está mais próximo, utilizando uma métrica de distância, como a distância Euclidiana. O cálculo de proximidade leva em consideração as variáveis musicais, como danceability, energy e valence, que foram normalizadas anteriormente.

3. Atualização a partir do centro: O algoritmo recalcula a posição do centro dos clusters, movendo-os para o centro dos pontos atribuídos a eles.

4. Iteração: Os processos 2 e 3 são repetido até que os centros dos clusters não mudem mais significativamente (convergência).

3.1.3 Interpretação dos Clusters

A análise revelou quatro grupos distintos de músicas, com diferentes combinações de danceability, energy, e outras variáveis. Cada grupo pode ser caracterizado por uma combinação única de características:

- Um grupo pode representar músicas altamente dançantes e energéticas.

- Outro grupo pode conter músicas mais acústicas, com baixos níveis de energia e danceabilidade. Essa segmentação pode ser útil para criar recomendações de músicas ou playlists baseadas em gostos específicos de usuários.

Para facilitar a interpretação dos resultados, foram gerados gráficos de dispersão que mostram as músicas plotadas com base em suas principais características (danceability vs energy), com as cores representando os diferentes clusters formados.

Abaixo está o gráfico é uma visualização dos clusters gerados pelo algoritmo K-means. Os 4 clusters foram representados pelas cores: Roxo(Cluster 0), Verde(Cluster 1), Amarelo(Cluster 2) e Azul(Cluster 3). Observando o gráfico é possíevel observar:

- As músicas que foram agrupadas no mesmo cluster têm características semelhantes. Cada cluster pode representar diferentes estilos ou características predominantes, como músicas com alta energia ou alta dançabilidade.

- A visualização sugere que os clusters não estão totalmente bem separados, o que pode indicar que algumas músicas têm características que as tornam intermediárias entre dois ou mais clusters.

- Algumas das músicas estão em clusters maiores e outros mais dispersos, o que pode refletir a diversidade dentro de cada grupo.

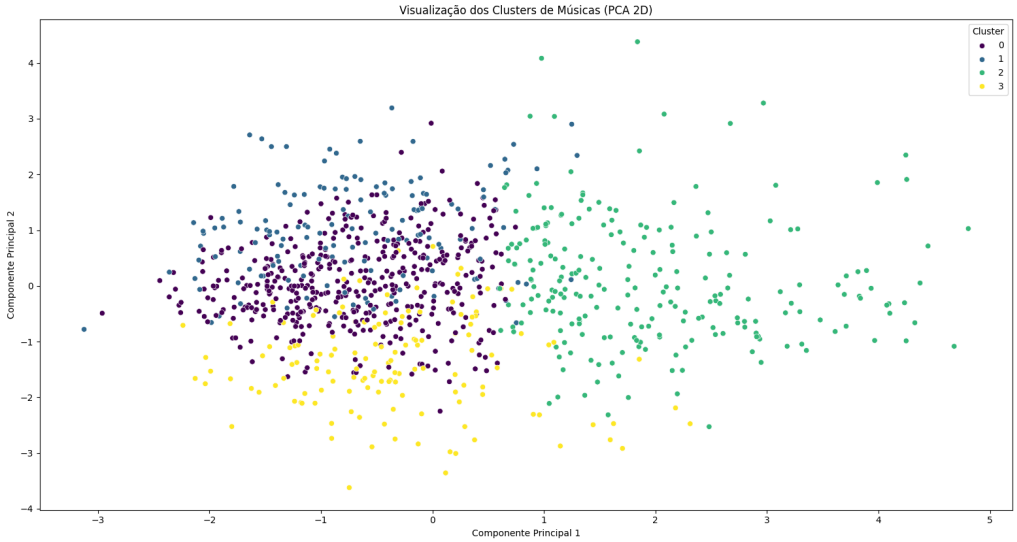

Figura 3: Plot dos clusters do Algoritmo K-Means

3.2 Algoritmo de Associação

3.2.1 Objetivo

O objetivo principal da aplicação de regras de associação neste trabalho foi identificar padrões e correlações entre diferentes características das músicas. As regras de associação são especialmente úteis para encontrar relações entre variáveis em grandes conjuntos de dados. Neste caso, o objetivo era descobrir quais atributos musicais frequentemente ocorrem juntos, como se músicas com alta danceability também tendem a ter alta energy ou valence.

3.2.2 Funcionamento do Algoritmo de Associação

Foi usado o Algoritmo Apriori, que utiliza as regras de associação. O Apriori é baseado no conceito de conjuntos frequentes, que são conjuntos de itens que ocorrem juntos com uma frequência maior do que um limite mínimo.

1. Geração de Conjuntos Frequentes: O primeiro passo do Apriori é identificar conjuntos de atributos musicais que aparecem juntos com frequência no conjunto de dados. As variáveis escolhidas foram danceability, energy, e valence foram usadas, categorizadas em faixas (alta, média, baixa). O algoritmo identificou, por exemplo, que músicas com danceability alta e energy alta frequentemente aparecem juntas, formando um conjunto frequente.

2. Geração de Regras de Associação: Após identificar os conjuntos frequentes, são gerados regras de associação. Cada regra segue o formato: "Se A, então B", sendo A e B as características musicais. O algoritmo calcula a confiança de cada regra, que mede a probabilidade de, dado que uma música tenha o atributo A, ela também tenha o atributo B. Por exemplo se danceability é alta, então energy também é alta.

3. Filtragem por Suporte e Confiança: As regras são filtradas com base em métricas

de suporte (a frequência com que a regra aparece no conjunto de dados) e confiança (a precisão da regra). Somente regras que atendem a limites mínimos de suporte e confiança são mantidas para interpretação. Além disso, o lift é utilizado para medir a importância de uma regra em relação à ocorrência aleatória dos atributos. Por exemplo, a regra "Se valence é baixo, então energy também é baixo"foi filtrada com base em um suporte mínimo de 60% e uma confiança de 75% . Isso significa que 60% das músicas no conjunto de dados exibem essa correlação, o que a torna uma regra significativa para análise.

3.2.3 Interpretação dos Resultados

- Uma das regras mais significativas mostrou que, em 80% dos casos, músicas com alta danceability também possuem alta energy. Dados como esses podem ser usados uteis para entender o tipo de música que as pessoas preferem ouvir em contextos dançantes.

- Outra regra importante foi que músicas com valence baixo (músicas mais tristes ou sombrias) tendem a ter energy baixa, com uma confiança de 75% . Esse padrão reflete a relação entre o humor da música e seu nível de energia.

O gráfico abaixo apresenta a relação entre três métricas importantes usadas para avaliar a qualidade das regras de associação: Suporte (Support), Confiança (Confidence), e Lift.

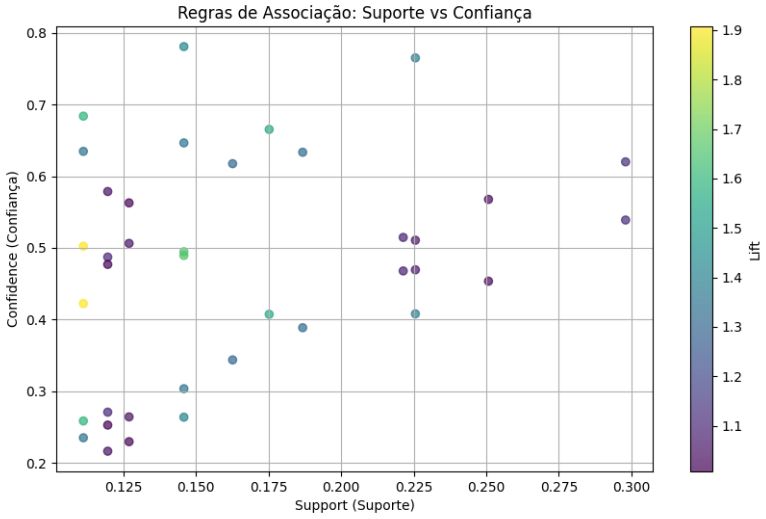

Figura 4: Plot das Regras de Associação

3.3 Algoritmo de Classificação: Support Vector Machine (SVM)

3.3.1 Objetivo

O algoritmo Support Vector Machine (SVM) teve como objetivo de classificar as músicas com base nas suas características numéricas e categorizar em diferentes grupos. O SVM foi utilizado para prever se uma música pertencia a uma categoria de "alta energia"ou "baixa energia", utilizando as características musicais, como danceability, valence, e tempo.

3.3.2 Funcionamento do Algoritmo SVM

O principal objetivo do SVM é encontrar um hiperplano que melhor separa os dados em diferentes classes. O SVM busca maximizar a margem entre os dados de diferentes classes. Dessa maneira a implementação foi feita em etapas.

1. Entrada dos Dados: O SVM recebe um conjunto de dados de treinamento, em seguida foram selecionadas variáveis como energy, danceability, valence, e outras que poderiam influenciar na categorização de uma música quanto ao nível de energia. As músicas foram rotuladas como de "alta energia"ou "baixa energia", formando um conjunto de dados rotulado para treinamento.

2. Escolha do Hiperplano: O SVM foi usado para encontrar um hiperplano que separasse essas duas categorias de músicas(alta ou baixa energia) com base nas características fornecidas. Esse hiperplano representaria a fronteira entre músicas com alta energia e baixa energia.

3. Classificação com Margem Máxima: São selecionados as músicas mais próximas da fronteira entre "alta energia"e "baixa energia", fazendo com que o modelo pudesse separar as músicas em suas respectivas categorias.

4. Uso de Kernel para Separação Não Linear: Caso as características das músicas (como danceability e valence) não permitissem uma separação clara entre as classes de energia, foi possível utilizar um kernel não-linear, como o kernel RBF (Radial Basis Function). Dessa maneira o algoritmo conseguiu separar adequadamente as músicas em "alta"ou "baixa"energia.

5. Classificação e Previsão: Depois de treinar o modelo com os dados de treinamento, foi realizado o teste com músicas novas (dados de teste). O modelo SVM foi capaz de prever com precisão se uma música tinha alta ou baixa energia com base em suas características.

3.3.3 Interpretação dos Resultados

O desempenho da SVM foi avaliado utilizando métricas de classificação padrão, como acurácia, precisão, recall e F1-score.

-Acurácia: Mediu a proporção de previsões corretas feitas pelo modelo.

-Precisão: Indicou a proporção de verdadeiros positivos entre todos os positivos previstos.

-Recall: Mediu a proporção de verdadeiros positivos identificados entre todos os positivos reais.

-F1-Score: Uma média harmônica entre precisão e recall, proporcionando uma visão equilibrada do desempenho do modelo.

A aplicação do SVM no conjunto de dados de músicas do Spotify demonstrou que o modelo foi eficaz em classificar as músicas em diferentes categorias de energia. A SVM conseguiu aprender as relações entre as características musicais e utilizá-las para separar músicas com diferentes níveis de energia. No gráfico abaixo é revelado que as características estão diretamente relacionadas à experiência emocional e física proporcionada pela música, o que faz sentido, já que essas variáveis provavelmente ajudam a categorizar as músicas com base em fatores como estilo, gênero e impacto.

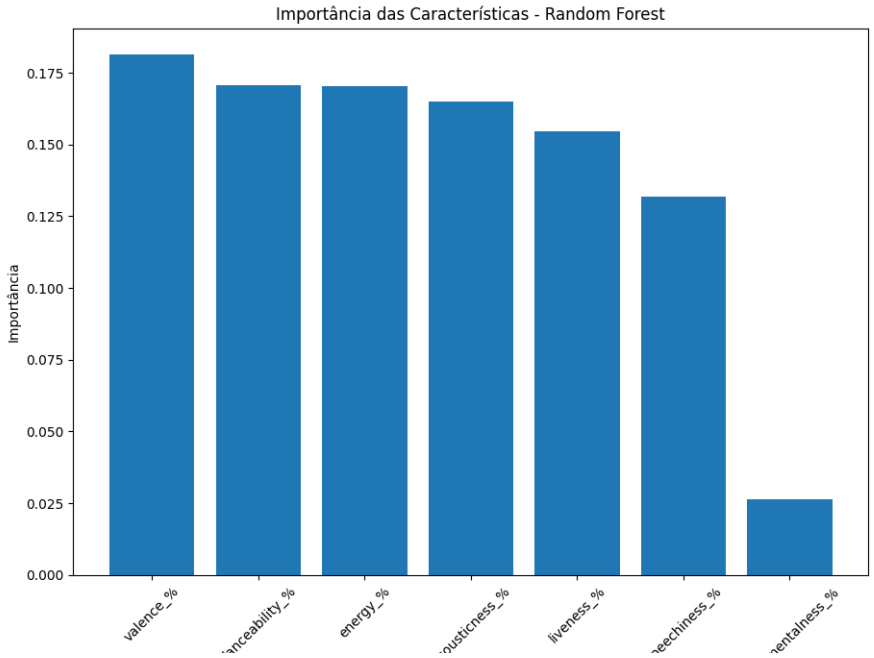

Figura 5: Plot da importância das caracteristicas

4 Conclusão

Neste trabalho, exploramos diversas técnicas de mineração de dados para analisar características musicais e identificar padrões em um conjunto de dados de músicas do Spotify. A análise demonstrou como a mineração de dados pode fornecer informações valiosas para sistemas de recomendação musical, que são amplamente utilizados em plataformas de streaming, como o próprio Spotify. Estes sistemas, conforme destacado em estudos como o MusicOSet e o Track Popularity Dataset , são baseados na combinação de dados acústicos, líricos, e informações contextuais para personalizar recomendações e prever a popularidade de faixas.

A técnica de agrupamento K-means foi aplicada para segmentar as músicas em clusters, com o objetivo de identificar grupos com características similares, como alta dançabilidade ou energia. Esta segmentação é útil para recomendar músicas que compartilham características musicais desejadas por usuários.

Além disso, as regras de associação revelaram padrões comuns entre as faixas, como a tendência de músicas com alta dançabilidade possuírem também alta energia. Esses insights podem ser usados para melhorar sistemas de recomendação, sugerindo faixas com base em padrões de preferência do usuário.

Por fim, o algoritmo de classificação SVM foi utilizado para categorizar músicas conforme sua energia, ajudando a predizer a adequação de uma faixa a diferentes contextos (por exemplo, músicas energéticas para treinos). Essa técnica de classificação é amplamente aplicável em recomendações personalizadas.

Em resumo, as técnicas de mineração de dados utilizadas não apenas permitiram uma exploração detalhada das características das músicas, mas também demonstraram como tais métodos podem suportar sistemas de recomendação eficientes, conforme comprovado

nos artigos revisados. Este estudo reforça o valor da mineração de dados para aprimorar a experiência de usuários em plataformas de streaming musical.

5 Referências

[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.