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 trabalho foi dividido em 4 etapas que serão mostradas a seguir.

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.

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: Escolhe aleatoriamente K pontos que representam o centro de cada cluster.

2. Atribuição: Cada ponto (música) é atribuído ao cluster cujo centro está mais próximo.

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 Aplicação no Conjunto de Dados do Spotify

4 Conclusão

5 Referências

`https://www.geeksforgeeks.org/longest-prefix-also-suffix/`

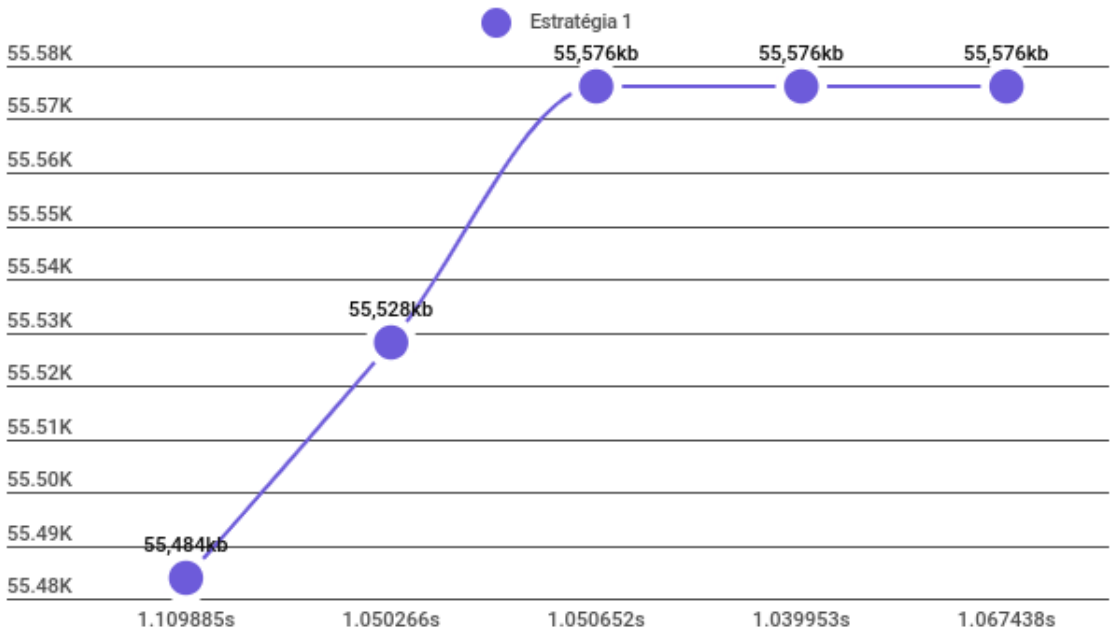

Figura 1: Estrategia 1