Descoberta de Conhecimento em Bases de Dados de Redes Sociais

Mineração e Análise de Redes Sociais

Henrique Azevedo Andrade Silva

Base de dados

A base de dados foi estruturada a partir de dados extraídos da API do Spotify. Ela foi organizada em tabelas relacionais para representar diferentes entidades, suas propriedades e relações.

É estruturada em quatro tabelas principais: Genres, Artists, Albums e Tracks, representando os gêneros musicais, artistas, álbuns e músicas, respectivamente. A tabela Genres armazena os gêneros musicais associados às músicas, com cada gênero identificado por um genre_id. A tabela Artists contém informações sobre os artistas, incluindo artist_id, artist_name e o link para seu perfil no Spotify (artist_uri). A tabela Albums relaciona álbuns aos artistas, armazenando o album_id, album_name, album_release_date, album_type, o número total de faixas (total_tracks) e a chave estrangeira artist_id para conectar ao artista correspondente. A tabela Tracks representa as músicas, com atributos como track_id, track_name, track_duration_ms (duração em milissegundos), track_popularity (popularidade em uma escala de 0 a 100), is_explicit (indicador de conteúdo explícito), além de chaves estrangeiras album_id e genre_id para conectar cada música a um álbum e a um gênero.

Descrição Semântica dos Atributos

Os atributos investigados na base de dados representam informações essenciais sobre gêneros, artistas, álbuns e músicas do Spotify. O genre_id e o genre_name identificam e nomeiam os estilos musicais, como Rock ou Pop, enquanto os artistas são descritos pelos atributos artist_id, que é o identificador único fornecido pelo Spotify, artist_name, o nome do artista, e artist_uri, que é um link para seu perfil na plataforma. Os álbuns, vinculados aos artistas por meio do artist_id, são caracterizados pelo album_id (identificador único), album_name (nome do álbum), album_release_date (data de lançamento), album_type (tipo do álbum, como "single" ou "album") e total_tracks (quantidade de faixas no álbum).

Já as músicas, associadas aos álbuns e gêneros por meio dos atributos album_id e genre_id, são descritas por track_id (identificador único), track_name (nome da faixa), track_duration_ms (duração em milissegundos), track_popularity (popularidade em uma escala de 0 a 100), e is_explicit (indicador de conteúdo explícito). Esses atributos capturam tanto informações objetivas, como duração e popularidade, quanto relações entre as entidades, permitindo análises detalhadas e enriquecedoras.

Estatísticas dos Atributos

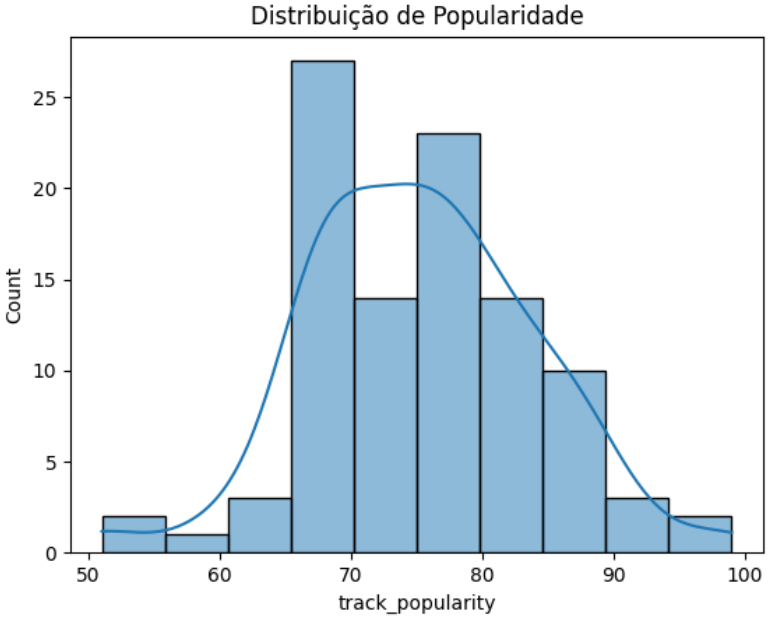
A análise estatística dos atributos track_duration_ms e track_popularity revelou que o conjunto de dados contém 99 músicas. O atributo track_duration_ms, que representa a duração das músicas em milissegundos, tem uma média de 218.011 ms e um desvio padrão de 59.279 ms, indicando uma variação moderada na duração das faixas. O valor mínimo é 119.880 ms, enquanto o máximo atinge 425.800 ms, com a mediana em 208.106 ms, o que sugere que a maior parte das músicas tem duração próxima à média. Já o atributo track_popularity, que mede a popularidade em uma escala de 0 a 100, apresenta uma média de 75,27 e um desvio padrão de 8,63, indicando que a maioria das músicas tem popularidade elevada. O valor mínimo é 51, e o máximo é 99, com a mediana em 75, o que confirma uma distribuição concentrada em faixas populares, como evidenciado pelos valores do primeiro quartil (68,5) e terceiro quartil (80). Esses resultados refletem uma coleção de músicas majoritariamente populares e com durações moderadamente homogêneas.

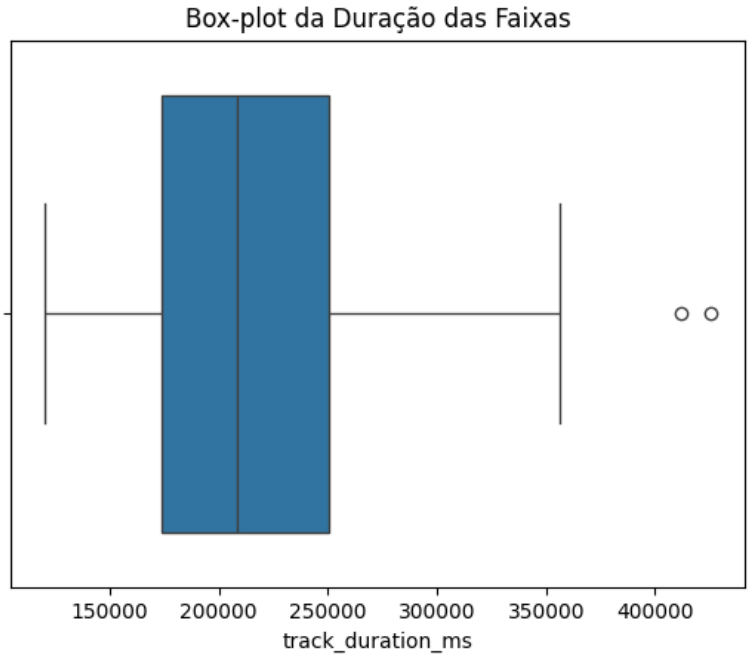
	track_duration_ms	track_popularity
count	99.000000	99.000000
mean	218011.222222	75.272727
std	59279.314588	8.628167
min	119880.000000	51.000000
25%	173785.500000	68.500000
50%	208106.000000	75.000000
75%	250600.000000	80.000000
max	425800.000000	99.000000

Distribuição dos valores dos atributos

A análise da distribuição dos atributos track_popularity e track_duration_ms revela insights importantes sobre as características das músicas no dataset. O histograma da popularidade das músicas mostra uma distribuição concentrada entre os valores 60 e 90, com uma clara tendência centralizada em torno da mediana de 75, corroborada pelo pico na faixa de 70 a 80. Essa distribuição sugere que a maioria das músicas no conjunto de dados é bastante popular.

Já o box-plot da duração das faixas indica que a maioria das músicas tem durações entre aproximadamente 150.000 ms e 300.000 ms, com a mediana próxima de 208.000 ms. Apesar de algumas observações se destacarem como outliers, alcançando até 425.000 ms, a distribuição da duração é relativamente homogênea, com poucas faixas ultrapassando o intervalo interquartil. Essas visualizações reforçam a percepção de que o dataset é composto por músicas populares e de duração moderada, com variações controladas.

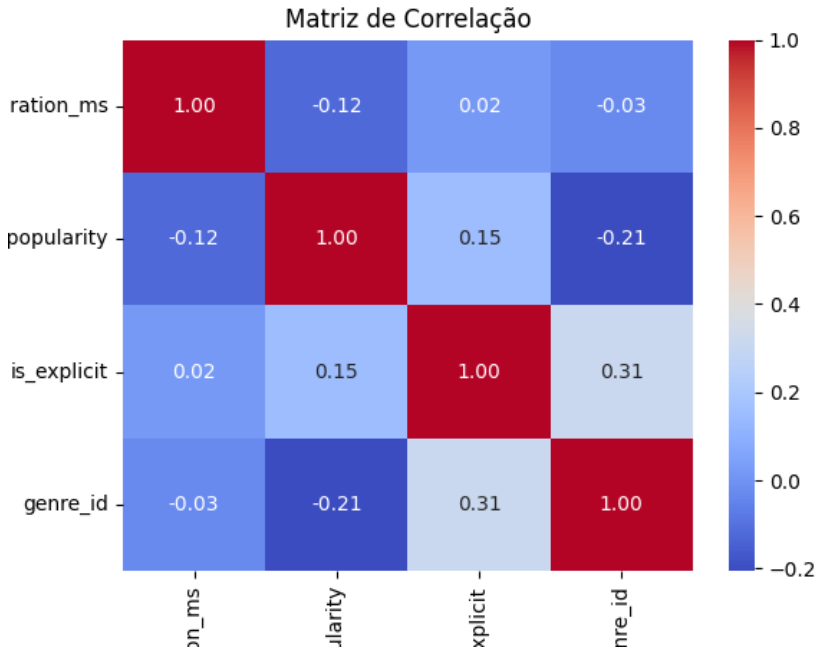




Compreensão da distribuição projetada dos atributos

A matriz de correlação apresentada fornece insights sobre as relações entre os atributos numéricos e categóricos do dataset, como track_duration_ms, track_popularity, is_explicit, e genre_id. Observa-se que a duração das faixas (track_duration_ms) tem correlações muito fracas com os demais atributos, como popularidade (track_popularity), com um coeficiente de -0,12, indicando que a duração não exerce uma influência significativa na popularidade das músicas. Por outro lado, track_popularity apresenta uma correlação ligeiramente positiva com o atributo is_explicit (0,15), sugerindo que músicas

com conteúdo explícito tendem a ter uma popularidade um pouco maior. O atributo genre_id mostra uma correlação moderada com is_explicit (0,31), indicando que certos gêneros musicais podem ter maior propensão a incluir conteúdo explícito. No entanto, sua correlação com track_popularity é levemente negativa (-0,21), sugerindo que o gênero pode influenciar de forma marginal a popularidade. Essas correlações, ainda que não muito altas, oferecem direções úteis para investigações futuras, como análise da popularidade com base no conteúdo explícito ou influência de gêneros específicos, enquanto confirmam que a duração das músicas não é um fator decisivo na popularidade ou classificação por gênero.

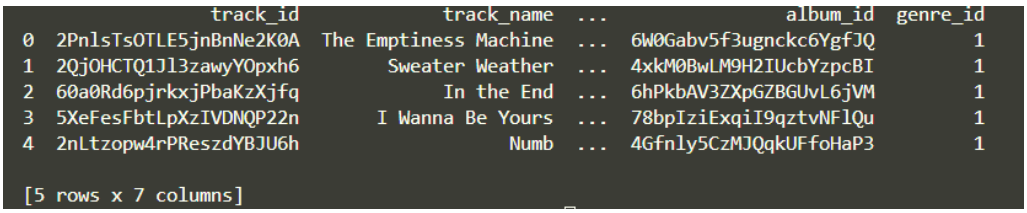


Normalização dos dados

A normalização dos dados é uma etapa essencial no pré-processamento de datasets, especialmente quando os atributos têm escalas diferentes e serão usados em análises baseadas em distância, como clustering ou aprendizado de máquina. No dataset fornecido, que inclui atributos como track_duration_ms e track_popularity, a normalização foi aplicada utilizando a técnica **Min-Max Scaling**. Essa técnica transforma os valores dos atributos para o intervalo [0, 1], garantindo que o menor valor de cada atributo seja mapeado para 0 e o maior para 1, enquanto os demais valores são escalonados proporcionalmente. Por exemplo, os valores originais de track_duration_ms, que variam de 119.880 a 425.800 ms, e track_popularity, que varia de 51 a 99, foram convertidos para uma escala comum, mantendo suas relações relativas.

Essa transformação é particularmente importante para evitar que atributos com escalas maiores dominem as análises, como a popularidade sendo mascarada pela alta variação na duração das músicas. A normalização também facilita a comparação e visualização de atributos diferentes, além de melhorar a convergência de algoritmos de aprendizado de máquina. Com base na tabela fornecida, os atributos categóricos, como

genre_id e identificadores (track_id, album_id), permanecem inalterados, pois não são escaláveis numericamente e representam apenas chaves para relacionamentos ou classificações. Essa abordagem permite que os atributos relevantes sejam processados de forma consistente, preservando as propriedades qualitativas do dataset.



	track_id	track_name	...	album_id	genre_id
0	2PnlsTsOTLE5jnBnNe2K0A	The Emptiness Machine	...	6W0Gabv5f3ugnckc6YgfJQ	1
1	2QjOHCTQ1Jl3zawyYOpxh6	Sweater Weather	...	4xkM0BwLM9H2IUcbYzpcBI	1
2	60a0Rd6pjrkxjPbaKzXjfq	In the End	...	6hPkbAV3ZXpGZBGUvL6jVM	1
3	5XeFesFbtLpXzIVDNQP22n	I Wanna Be Yours	...	78bpIziExqiI9qztvNFlQu	1
4	2nLtzopw4rPReszdYBJU6h	Numb	...	4Gfnly5CzMJQqkUFfoHaP3	1

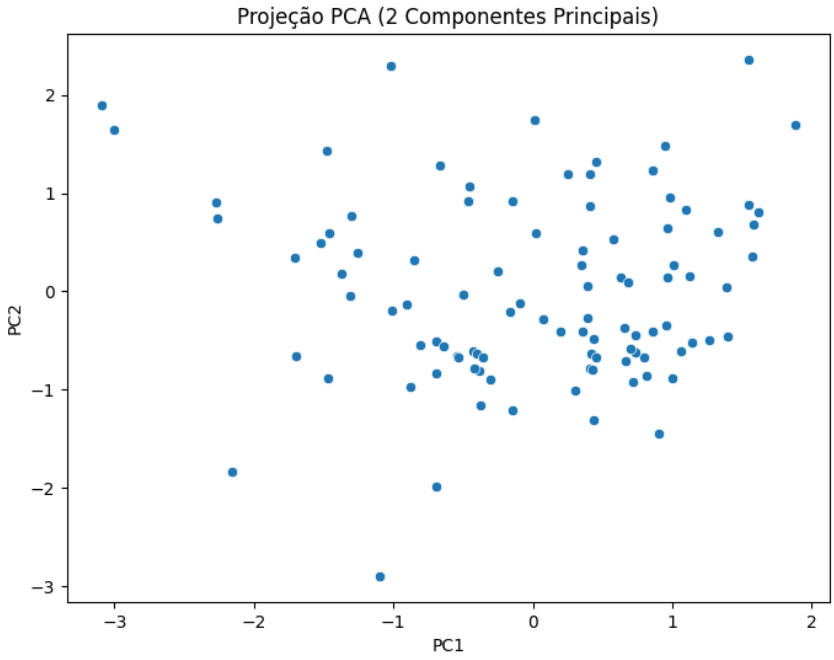
[5 rows x 7 columns]

Redução da Dimensionalidade

A redução da dimensionalidade do conjunto de dados foi realizada utilizando a técnica de Análise de Componentes Principais (PCA), que projeta os dados originais em um espaço de menor dimensão enquanto retém a maior parte da variância explicada pelos atributos. No caso apresentado, os dois primeiros componentes principais foram escolhidos para a projeção, que juntos explicam aproximadamente 100% da variância total do conjunto de dados, sendo o primeiro responsável por 56,1% e o segundo por 43,9%.

Essa redução permite representar as informações mais relevantes em apenas duas dimensões, simplificando a análise visual e mitigando problemas de alta dimensionalidade, como redundância e ruído. O gráfico de dispersão gerado a partir da projeção dos dados nos dois componentes principais (PC1 e PC2) revela uma distribuição dispersa, indicando que os dados têm variabilidade significativa e que não há uma concentração clara em um único padrão ou cluster específico. Essa abordagem é especialmente útil em cenários como este, onde os dados originais incluem múltiplos atributos numéricos com possível redundância.

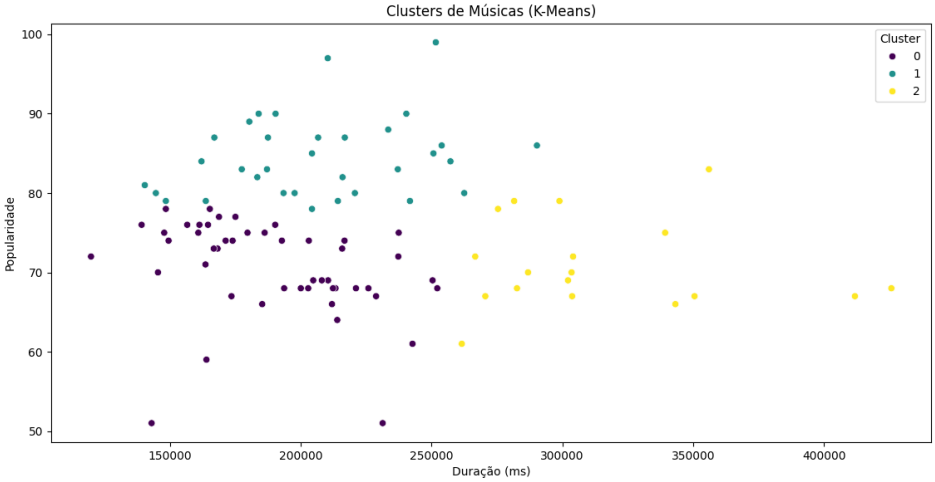
Ao reduzir para dois componentes principais, mantemos a maior parte das informações necessárias para análises futuras, enquanto facilitamos a interpretação e visualização dos dados. Além disso, a projeção pode ser utilizada como entrada para algoritmos de clustering ou classificação, garantindo que os resultados sejam baseados em informações otimizadas e livres de redundâncias.




```
Contribuição de variância explicada por componente:
[0.5609663 0.4390337]
```

Explicação do Kmeans Clustering

A técnica de mineração de dados adotada foi o K-Means Clustering, que é amplamente utilizada para agrupar dados em clusters com base em suas características. O algoritmo funciona inicializando um número predefinido de centróides (neste caso, 3 clusters) e ajustando iterativamente suas posições para minimizar a soma das distâncias quadradas entre os pontos de dados e seus centróides correspondentes. No gráfico apresentado, os dados foram agrupados com base na duração das músicas (em milissegundos) e na popularidade (em uma escala de 0 a 100), resultando em três clusters distintos. O cluster roxo (Cluster 0) representa músicas de menor popularidade e duração mais curta, o cluster azul claro (Cluster 1) contém músicas de popularidade intermediária e duração moderada, enquanto o cluster amarelo (Cluster 2) agrupa músicas de maior duração e popularidade variável. Esses agrupamentos permitem identificar padrões importantes, como a relação entre a duração e a popularidade, além de fornecer insights que podem ser úteis para categorização de músicas, criação de playlists ou personalização de recomendações para usuários. A visualização clara dos clusters evidencia a separação dos dados com base nas características selecionadas, validando a eficácia da técnica utilizada.



Estratégias para minimizar o overfitting

Embora o K-Means Clustering seja menos suscetível ao overfitting do que modelos supervisionados, ele ainda pode ser influenciado por ruídos nos dados, escolha inadequada de parâmetros, ou um número excessivo de clusters.

Escolha apropriada do número de clusters, remoção de outliers, aumento no número de inicializações, e redução de dimensionalidade são algumas das estratégias para minimizar o overfitting.

Avaliação do resultado da classificação

A classificação realizada por meio do algoritmo K-Means Clustering apresentou resultados promissores ao agrupar as músicas em três clusters distintos com base em sua duração e popularidade. A separação visual no gráfico indica que os clusters estão relativamente bem definidos, com grupos identificáveis e poucos pontos de sobreposição. Essa estrutura reflete a eficácia do K-Means em identificar padrões dentro dos dados, mesmo em um espaço de duas dimensões.