Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Luis Henrique da Rocha Troscianczuk

Classificação de gênero musical utilizando características de arquivos de áudio

Curitiba
2024

Luis Henrique da Rocha Troscianczuk

Classificação de gênero musical utilizando características de arquivos de áudio

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Marco Antonio Zanata

Curitiba
2024

Classificação de gênero musical utilizando características de arquivos de áudio

Music genre classification using audio file features

Luis Henrique da Rocha Troscianczuk[1], Prof. Marco Antonio Zanata[2]

[1]Aluno do programa de Especialização em Data Science & Big Data, luis@trosciancz.uk
[2]Professor do Departamento de Informática - DInf/UFPR, mazalves@inf.ufpr.br

A automatização do processo de classificação de gênero musical apresenta diversos desafios, desde a captura de características de sinais digitais, até a definição de critérios objetivos que diferenciem gêneros. Este trabalho visa analisar tais desafios e avaliar a viabilidade de tal processo. Para isso, foram treinados modelos de aprendizado de máquina com um *dataset* extraído com técnicas de *webscraping* e transformações dos sinais obtidos para obtenção dos MFCCs. Os resultados foram moderados e atingiram até 75% de acurácia avaliando a moda dos resultados de trechos de um mesmo álbum. O código desenvolvido está disponível em: `https://github.com/luis951/sound-similarity/`
Palavras-chave: classificação de gênero musical, processamento de sinais digitais, MFCCs, aprendizado de máquina

The automation of the musical genre classification process presents several challenges, from capturing digital signal characteristics to defining objective criteria that differentiate genres. This work aims to analyze such challenges and evaluate the viability of such a process. For that, machine learning models were trained with a *dataset* extracted using *webscraping* techniques and transformations of the obtained signals to obtain the MFCCs. The results were moderate and reached up to an accuracy of 75% when evaluating the mode of results from samples from the same album. The developed code is available in: `https://github.com/luis951/sound-similarity/`
Keywords: musical genre classification, digital signal processing, MFCCs, machine learning

1. Introdução

O gênero musical é uma caraterística importante para um usuário que deseja encontrar uma música em um grande acervo ou mesmo como um critério para um algoritmo de recomendação. Em 2022, executivos dos principais serviços de streaming digital estimaram que cerca de 100 mil músicas eram adicionadas aos catálogos todos os dias, que, com uma duração média de 3 minutos, resultariam em 5000 horas de música, o que exigiria um enorme esforço para ser catalogado. [1]

Humanos utilizam critérios subjetivos para classificar gêneros musicais, no entanto, existem características sonoras que são usadas para descrever músicas que podem refletir esses critérios, como: ritmo, dinâmica, melodia, harmonia, timbre e textura. [2] [3]

- Ritmo é um elemento relativo ao tempo na música, como a velocidade das batidas, seus padrões e sua uniformidade;
- Dinâmica se refere ao volume e intensidade da música;
- Melodia é um componente da frequência que refere à sua variação ao longo do tempo;
- Harmonia se refere a frequências que acontecem simultâneamente à melodia, podendo ser consoantes ou dissonantes em relação a ela;
- Timbre é a qualidade referente do instrumento (ou voz) que origina a frequência;
- Textura é referente a quantidade de diferentes melodias que soam simultaneamente em uma música.

Abaixo está uma transcrição para piano de um solo de uma apresentação de John Coltrane, famoso saxofonista de *Jazz*. A partitura é uma linguagem que consegue codificar todos as características acima e é possí-

vel que uma pessoa com treinamento musical consiga identificar o genêro baseando-se nesses dados.

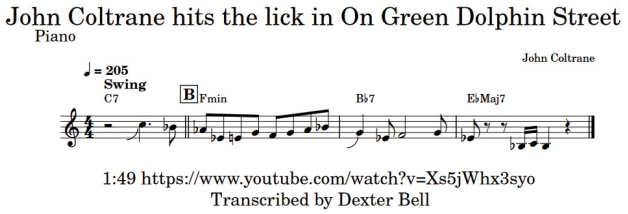


Figura 1: Transcrição de um solo de saxofone.

No entanto, um arquivo de áudio, carrega em si uma digitalização de um sinal de áudio. Sinal digital é a descrição da variação de um parâmetro em relação a outro, para sinais de áudio isso se reflete na intensidade e frequência de ondas sonoras a serem reproduzidas por um dispositivo.

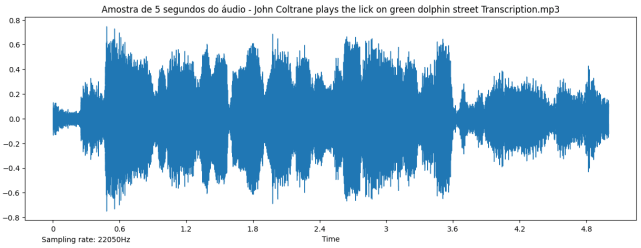


Figura 2: Forma de onda da amostra

Abaixo uma visualização dos primeiros 10 milissegundos do sinal acima onde é possível observar os incrementos quantizados de tempo e amplitude da onda, relacionados à taxa de amostragem e profundidade de bits do sinal, respectivamente:

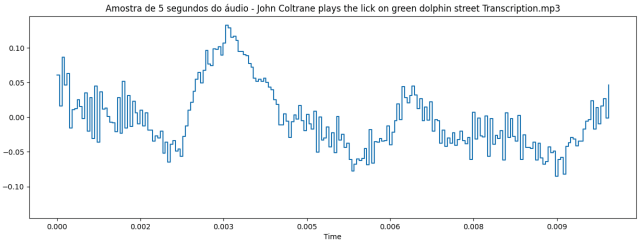


Figura 3: Forma de onda da amostra ampliada

No entanto, apenas com a forma de onda do sinal não é possível estimar nenhuma informação sobre frequência, para isso é utilizada a transformada de Fourier, que converte a informação do domínio do tempo para o domínio da frequência.

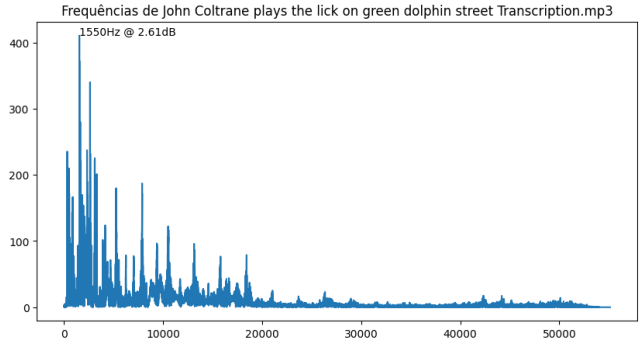


Figura 4: Frequências da onda da amostra

A informação acima nos ajuda a estimar as frequências prevalentes globalmente no sinal analisado, sem levar em consideração o tempo que essas frequências ocorrem. Para isso, é utilizada a Transformada de Fourier de tempo curto, obtendo as frequências em janelas que dividem todo o sinal analisado, coletando assim os dados de frequência e magnitude ao longo do eixo do tempo. A escala de frequência original obtida é linear e apresenta uma distância muito grande entre tons, a escala em Mels, por ser logarítmica corrige essa distorção e corresponde melhor à percepção humana da distância entre os tons.

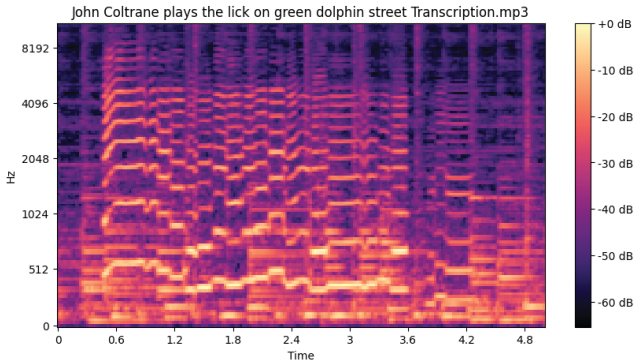


Figura 5: Espectrograma de Mel da amostra

Por fim, os MFCCs (Coeficientes Cepstrais de frequências Mel) são um conjunto de atributos extraídas a partir da Transformada Discreta de Coseno. O método foi proposto inicialmente em um artigo em 1980 para reconhecimento de palavras monossilábicas e continua altamente relevante para pesquisas relacionadas a classificação de sons.

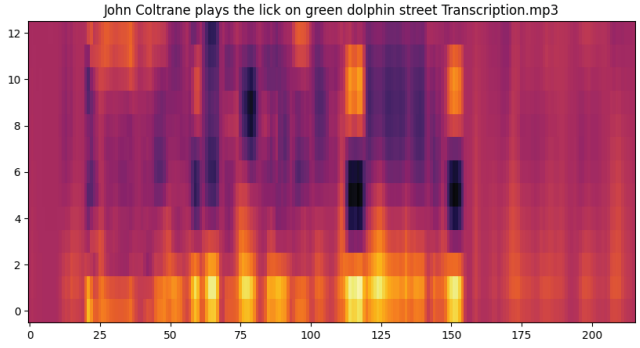


Figura 6: MFCCs da amostra

2. Discussão

Os MFCCs são característica do som utilizada em diversos projetos de classificação de dados sonoros, com muito sucesso em alguns casos. Existem artigos que utilizam outras características do som, como o espectrograma em si [6] ou padrões binários locais [5], uma característica textural que pode ser extraído do espectrograma em escala linear que também atingiram resultados muito bons. Nesse artigo foi optado pela utilização do MFCC pelo motivo da redução de dimensionalidade, de uma matriz de 1025x216 para uma de 13x10 para um trecho de 5 segundos, mais adequado para o processamento em um computador pessoal e também pelo fato dessa *feature* codificar duas características importantes para a descrição musical: timbre, como descrito em e também em um experimento anterior feito com sons de piano gerados automaticamente que resultou no seguinte gráfico, feito com redução de dimensionalidade TSNE aplicada nos MFCCs extraídos das amostras:

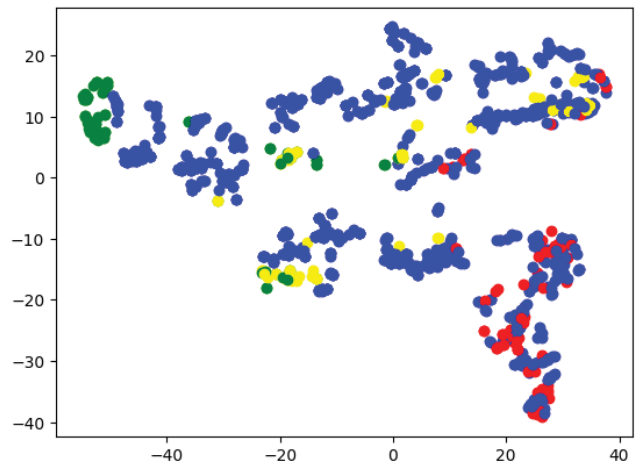


Figura 7: Análise de tonalidades de amostras com MFCC e TSNE

É possível observar que há um distanciamento das amostras conforme seu tom, com as amostras destacadas em vermelho sendo em C/dó, as em amarelo em F#/fá sustenido (6 semitons acima) e as em verde em B/si. (11 semitons acima de C/dó)

3. Materiais e métodos

A construção da base de dados utilizada na construção desse trabalho foi obtida em duas etapas: a primeira delas, a coleta dos dados de classificação, foi através do *webscraping* do site *Rate Your Music*,[7] uma espécie de catálogo musical público onde usuários da comunidade podem categorizar, descrever e avaliar álbums e canções. Foram coletadas informações dos 80 álbums mais bem classificados de 5 gêneros diferentes: *Folk, Hip Hop, R&B, Jazz e Rock*, totalizando 400. É importante resaltar que foram utilizados filtro especiais de busca para evitar intersecção de álbums que pudessem ser classificados em diferentes gêneros. A segunda etapa foi a coleta dos arquivos de som dos álbums, feita com a biblioteca *Savify* [8] do *Python*, onde foi possível obter apenas 384 dos álbums selecionados na etapa anterior, devido a problemas de disponibilidade.

Os álbums foram divididos nas categorias de treino, teste e validação na proporção de 50%, 25% e 25%, resultando na seguinte distribuição considerando a segregação das classes:

X	Folk	Hip Hop	Jazz	R&B	Rock	Total
Trn	31	37	39	39	39	185
Tst	15	18	19	18	19	89
Vld	15	18	19	18	19	89

O processamento dos dados para uniformização do tamanho das amostras para ingestão nos modelos de machine learning se deu primeiramente na concatenação de todos os arquivos de cada álbum e na separação em janelas de 5 segundos. Para cada janela de 5 segundos foi extraído o espectrograma de Mel com as janelas e intervalos de metade ao *sampling rate*, totalizando 10 janelas no período completo. Desses espectogramas foram extraídos os MFCCs, que foram agrupados em arquivos respectivos a seus álbums.

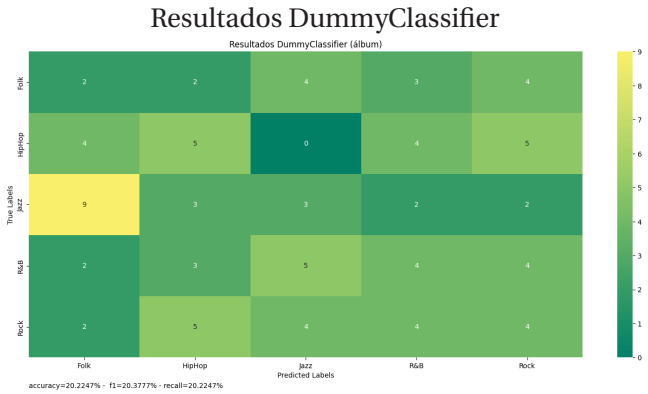
Os métodos de *machine learning* de classificação escolhidos para os testes foram: *Nearest Neighbors, Decision Tree, Random Forest, XGBoost*, disponibilizados pela biblioteca *Scikit Learn* [9] e Redes Neurais Convolucionais, através da biblioteca *Keras* [10].

Devido à restrição de formato bidimensional para as entradas nos modelos de *machine learning*, clássicos, as matrizes de *input* foram redimensionadas com

uma função de *flatten*, alterando-se suas dimensões de 10x39 para 1x390.

4. Resultados

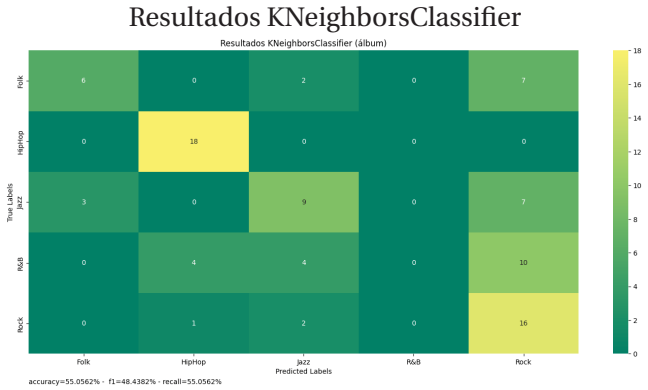
Como base de comparação foi utilizado o resultado de um modelo classificador *Dummy*, que classifica as amostras de maneira aleatória e uniforme, obtendo um resultado de 20.22% de acurácia:



accuracy=20.22% - f1=20.37% - recall=20.22%

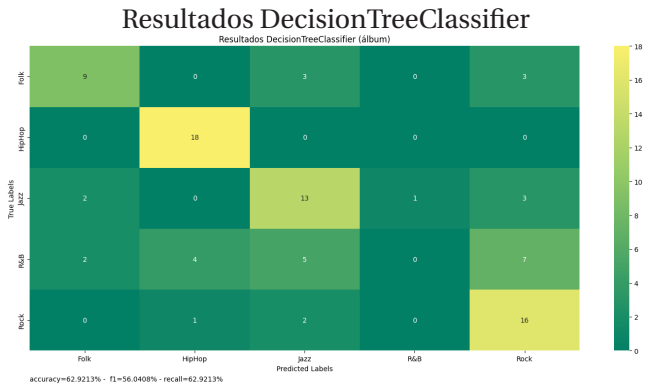
Figura 8: Matriz de confusão do modelo Dummy Classifier

Em seguida os modelos testados foram *Nearest Neighbors, Decision Tree, Random Forest, XGBoost*:



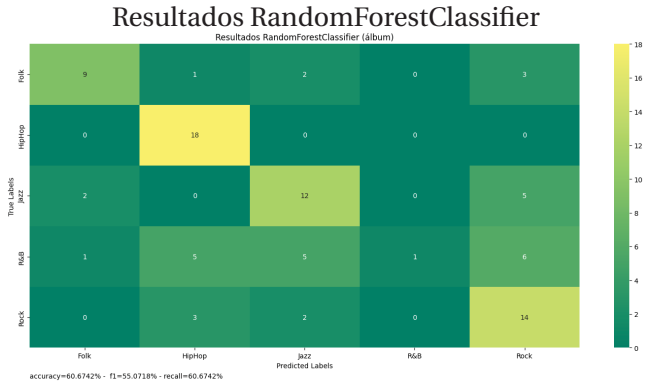
accuracy=55.06% - f1=48.43% - recall=55.06%

Figura 9: Matriz de confusão do modelo KNeighborsClassifier



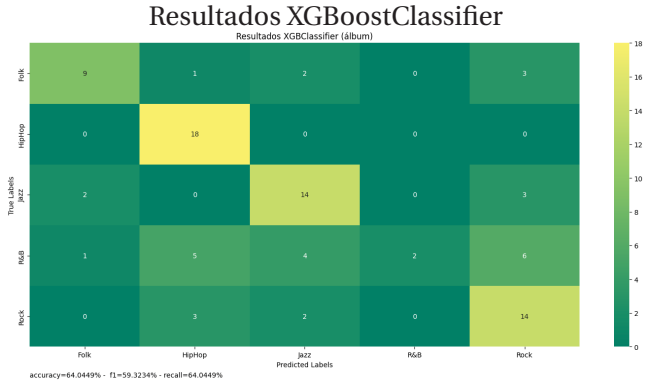
accuracy=62.92% - f1=56.05% - recall=62.92%

Figura 10: Matriz de confusão do modelo DecisionTreeClassifier



accuracy=60.67% - f1=55.07% - recall=60.67%

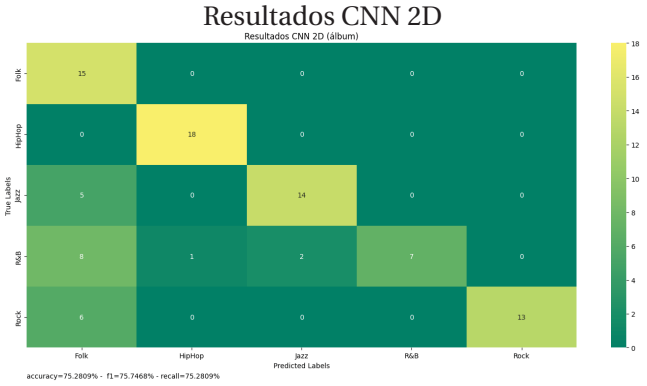
Figura 11: Matriz de confusão do modelo RandomForestClassifier



accuracy=64.04% - f1=59.32% - recall=64.04%

Figura 12: Matriz de confusão do modelo XGBoostClassifier

Os resultados foram muito próximos, com um leve viés para a classe *Rock*, afetando a medida de F1 devido a quantidade de falsos positivos dessa categoria. Outro ponto a se observar é o de que a classe *Hip Hop* teve uma pontuação perfeita em todos os modelos.



accuracy=75.28% - f1=75.74% - recall=75.28%

Figura 13: Matriz de confusão do modelo de classificação com redes neurais convolucionais

O modelo de Redes Neurais Convolucionais teve um resultado levemente superior aos demais modelos, dessa vez com o viés tendendo levemente ao *Folk*, com diversos falsos positivos.

5. Conclusão

Os modelos de *machine learning* tiveram um resultado muito superior ao modelo base, porém suas classificações ainda não tiveram um resultado satisfatório o suficiente para o uso em um ambiente real. Além disso, as condições de uso dos modelos treinados são bem limitadas considerando que é feita apenas a classificação de álbum completos, sem levar em conta músicas lançadas individualmente. Outro fator a se ressaltar é o de que músicas, em geral, têm seções mais e menos relevantes a seu gênero, com algumas incluindo momentos de silêncio, trechos de conversa e outros sons não-musicais que podem representar ruído para o modelo. Muitas outras *features* poderiam ter sido extraídas dos arquivos para complementar o modelo de dados e trazer um resultado melhor. Todos esses são fatores a se considerar em um trabalhos futuros sobre o tema.

Referências

[1] Disponível em: *https://variety.com/2022/music/news/new-songs-100000-being-released-every-day-dsps-1235395788/*.

[2] AHRENDT, P. Music Genre Classification Systems -A Computational Approach. [s.l: s.n.]. Disponível em: *https://core.ac.uk/download/pdf/13734609.pdf*.

[3] SCHMIDT-JONES, C. The Basic Elements of Music. [s.l.] Orange Groove Books, 2009.

[4] TERASAWA, H.; SLANEY, M.; BERGER, J. Perceptual Distancia in Timbre Space. 1 jul. 2005.

[5] COSTA, Y. M. G. et al. Music genre classification using LBP textural features. Signal Processing, v. 92, n. 11, p. 2723–2737, nov. 2012.

[6] COSTA, Y. M. G.; OLIVEIRA, L. S.; SILLA, C. N. An evaluation of Convolutional Neural Networks for music classification using spectrograms. Applied Soft Computing, v. 52, p. 28–38, mar. 2017.

[7] Rate Your Music. Disponível em: <https://rateyourmusic.com>.

[8] RAWLINGS, L. LaurenceRawlings/savify. Disponível em: <https://github.com/LaurenceRawlings/savify>. Acesso em: 16 jul. 2024.

[9] SCIKIT-LEARN. scikit-learn: machine learning in Python. Disponível em: <https://scikit-learn.org/stable/>.

[10] KERAS. Home - Keras Documentation. Disponível em: <https://keras.io/>.