Reconhecimento de Gênero Musical

Geovana Teixeira Camargo
Instituto de Informática (INF)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
geovanateixeira@discente.ufg.br

Jair Martins Santiago Neto
Instituto de Informática (INF)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
jairneto@discente.ufg.br

Lidia Rakel Alcantara do Vale
Instituto de Informática (INF)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
lidia.vale@discente.ufg.br

Mirmila Sócrates do Nascimento
Instituto de Informática (INF)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
mirmilasocrates@discente.ufg.br

Natã Milhomem Lemos
Instituto de Informática (INF)
Universidade Federal de Goiás (UFG)
Goiânia, Brasil
natamilhomem@discente.ufg.br

***Resumo* — Este trabalho consistiu na construção de um classificador de gêneros musicais utilizando aprendizado de máquina. Após análise exploratória e pré-processamento dos dados acústicos, foram treinados modelos de Random Forest, CatBoost, Gradient Boosting e uma rede neural, esta última implementada com TensorFlow. A avaliação incluiu o cálculo da acurácia, do F1-Score e da precisão top-3, além da visualização da matriz de confusão. O modelo que teve melhor desempenho foi o CatBoost, com 83% de acurácia, 83% de F1-Score e 96% de precisão top-3.**

Palavras-chave* — *classificação de gêneros musicais, aprendizado de máquina, redes neurais

I. Introdução

Este trabalho aborda a classificação automática de músicas em gêneros como jazz, rock e pop, utilizando técnicas de processamento digital de sinais e aprendizado de máquina. Com o aumento da quantidade de músicas disponíveis em plataformas de streaming, a categorização automática em gêneros facilita a organização de bibliotecas, a personalização de recomendações e a busca em arquivos de áudio [1]. O objetivo do trabalho é identificar padrões dentre as características extraídas dos sinais do áudio, como intensidade de energia e espectro de frequência, de forma que possa ser construído um modelo de reconhecimento musical.

Para isso, foi utilizado como base de dados o GTZAN Music Genre Dataset [2], amplamente reconhecido em pesquisas de classificação de gêneros musicais, e foi realizada uma revisão da literatura sobre extração de características e algoritmos de classificação com aprendizado de máquina. A metodologia do trabalho inclui a análise de características do dataset, pré-processamento dos dados e treinamento de modelos de aprendizado. O desempenho dos modelos foi avaliado com as métricas como acurácia, matriz de confusão e F-score, além de benchmarks da literatura, buscando identificar as abordagens mais eficazes para a classificação de gêneros musicais.

II. Fundamentos teóricos

O reconhecimento de gênero musical é uma tarefa complexa que combina processamento avançado de sinais de áudio e aprendizado de máquina, sendo amplamente aplicado em sistemas de recomendação e organização de bibliotecas musicais digitais. São abordados conceitos fundamentais que sustentam as técnicas empregadas, detalhando mecanismos, algoritmos e estratégias de avaliação.

A biblioteca Librosa desempenha um papel central na análise e visualização dos sinais de áudio, permitindo a extração de características relevantes, como coeficientes Mel-Frequency Cepstral Coefficients (MFCCs), energia, ritmo e espectrogramas, que servem como base para a classificação de gêneros musicais [4].

Os MFCCs representam a envoltória espectral de um sinal, capturando o timbre e a tonalidade. Essa técnica é amplamente reconhecida por seu alinhamento com a percepção humana de sons [3]. A análise demonstra que a Transformada Discreta do Cosseno é adequada para correlacionar espectros tanto de fala quanto de música. No processo de extração das características MFCC o sinal de áudio é segmentado em quadros, aplicando-se uma função de janela em intervalos regulares. Esse procedimento visa modelar pequenas partes do sinal (geralmente com duração de 20 ms), que são consideradas estatisticamente estacionárias. Como resultado, é gerado um vetor de características cepstrais (estruturas periódicas em espectros de frequência) para cada quadro processado.

Os espectrogramas são representações visuais da energia do áudio em diferentes frequências ao longo do tempo. As Redes Neurais Convolucionais (CNNs) são eficazes na tarefa de classificação de áudios por espectrograma porque podem capturar padrões espaciais e temporais, associando características específicas a gêneros musicais. O seu uso permite que o modelo aproveite informações detalhadas sobre o timbre, ritmo e frequência, fundamentais para diferenciar gêneros musicais [5]. Essa abordagem melhora a precisão na classificação quando comparada a métodos baseados apenas em características numéricas.

Quanto aos métodos de aprendizado de máquina, o CatBoost é um algoritmo que tem ganhado popularidade nos últimos anos, especialmente na área de classificação de gêneros musicais. O nome "CatBoost" é uma abreviação de "Category Boosting", o que destaca sua capacidade de lidar eficientemente com variáveis categóricas, um tipo de dado comum nesse tipo de classificação [8]. Pode ser usado para construir modelos que classificam músicas em diferentes gêneros com base em suas letras. As letras são processadas e transformadas em características numéricas que servem de entrada para o algoritmo. Após o treinamento, o modelo pode prever o gênero de novas músicas com base em suas letras.

Outro algoritmo é o Random Forest, que é baseado em um conjunto de árvores de decisão. Ele funciona gerando múltiplas árvores de decisão independentes a partir de diferentes amostras aleatórias do conjunto de dados e, em seguida, combina os resultados das árvores individuais (por meio de votação, no caso de classificação, ou média, no caso de regressão) para produzir uma previsão final. Este método reduz o risco de overfitting, aproveitando a diversidade entre as árvores, e melhora a generalização do modelo. A inclusão de variáveis aleatórias na divisão de nós das árvores também aumenta a robustez contra ruído nos dados.

Ainda, O XGBoost (Extreme Gradient Boosting) é uma técnica de aprendizado baseada no método de boosting, onde múltiplas árvores de decisão são construídas sequencialmente, cada uma tentando corrigir os erros das anteriores. Diferentemente de outros algoritmos de boosting, o XGBoost utiliza otimizações avançadas, como paralelismo e regularização L1/L2, para melhorar a eficiência computacional e reduzir o risco de overfitting. Ele é amplamente utilizado em problemas de classificação e regressão devido à sua capacidade de lidar com dados desbalanceados, alta dimensionalidade e cenários complexos.

A avaliação de modelos em tarefas de classificação, como a de gêneros musicais, é uma etapa essencial para medir sua eficácia e identificar áreas de melhoria. Em problemas de aprendizado de máquina, a avaliação permite verificar se o modelo treinado é capaz de generalizar adequadamente para dados desconhecidos, que representam casos reais.

Dentre as métricas comumente utilizadas, a acurácia é uma das mais populares e mede a proporção de previsões corretas em relação ao total de amostras testadas. Formalmente, é expressa como a razão entre o número de acertos e o número total de testes. Embora simples e intuitiva, a acurácia pode ser enganosa em conjuntos de dados desbalanceados. Por exemplo, se um gênero musical domina o conjunto de dados, o modelo pode obter uma alta acurácia apenas por prever consistentemente a classe majoritária, mesmo sem aprender padrões significativos das outras classes [6].

Por isso, métricas adicionais são frequentemente utilizadas para complementar a análise, como visualizações de desempenho, que podem ser essenciais. Gráficos que mostram a evolução da acurácia e do erro para os conjuntos de treinamento e validação são amplamente utilizados. Esses gráficos ajudam a diagnosticar problemas como overfitting (quando o modelo se ajusta excessivamente aos dados de treinamento e perde capacidade de generalização) e underfitting (quando o modelo é incapaz de capturar padrões significativos nos dados). Por exemplo, uma grande diferença entre a acurácia no treinamento e na validação pode indicar overfitting, enquanto baixos valores de acurácia em ambos os conjuntos sugerem underfitting. Esses insights podem orientar ajustes nos hiperparâmetros, como o número de árvores em algoritmos de boosting, a taxa de aprendizado ou mesmo os métodos de pré-processamento e seleção de características.

III. Metodologia

Primeiramente, a base de dados foi composta por trechos de áudio de 30 segundos, categorizados em 10 diferentes gêneros musicais. Cada amostra possuía características extraídas previamente, como coeficientes mel-freq (MFCCs), energia, espectro de frequência e tempo entre batidas, que foram fundamentais para representar as propriedades acústicas dos gêneros. A manipulação dos dados e a realização das análises exploratórias foram possibilitadas pelas bibliotecas pandas e numpy para organização e cálculos, além de matplotlib e seaborn para visualizações detalhadas.

Na etapa de análise exploratória, foi verificada a distribuição das amostras para cada gênero musical utilizando funções como value_counts() do pandas. Foi confirmado o balanceamento da base, ou seja, cada gênero possuía a mesma quantidade de amostras (100 amostras por gênero). Em seguida, para compreender melhor as características dos áudios, formas de onda e espectrogramas foram gerados com a biblioteca librosa. As formas de onda destacaram a amplitude das amostras ao longo do tempo, enquanto os espectrogramas forneceram uma representação visual da energia em diferentes frequências ao longo do tempo, permitindo identificar padrões distintos entre os gêneros. As Figuras 1 e 2 apresentam o gráfico de onda e o espectrograma, respectivamente do gênero Pop a título de exemplo.

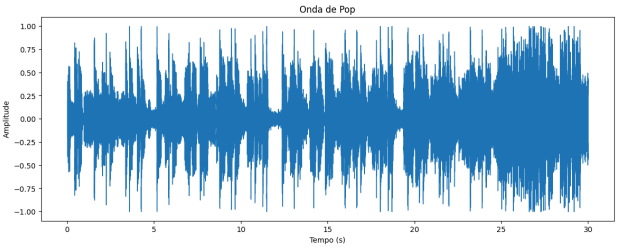

Fig. 1. Forma de onda do gênero Pop

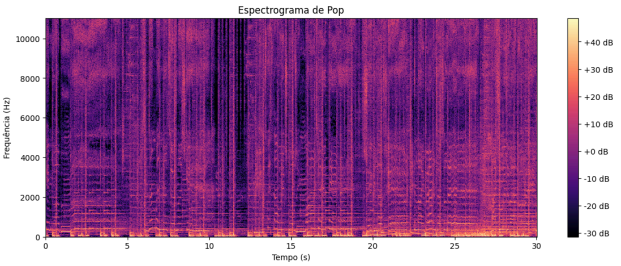

Fig. 2. Espectrograma do gênero Pop

Ainda na exploração dos dados, foi analisada a relação entre as colunas do dataset através de em heatmap, apresentado na Figura 3, que representa a correlação entre as variáveis que capturam características médias do áudio. As cores mais escuras indicam correlações fortes (próximas de 1 ou -1), enquanto as cores claras representam correlações fracas ou inexistentes. Observa-se que os coeficientes MFCC, especialmente os consecutivos, possuem alta correlação entre si, o que é esperado, pois eles descrevem padrões relacionados no domínio da frequência. Já variáveis como chroma_stft_mean e zero_crossing_rate_mean mostram menor correlação com os MFCCs, sugerindo que capturam aspectos distintos do áudio, como tonalidade e transições do sinal. Isso pode indicar redundância em algumas variáveis e complementaridade em outras, o que é relevante para seleção ou redução de features.

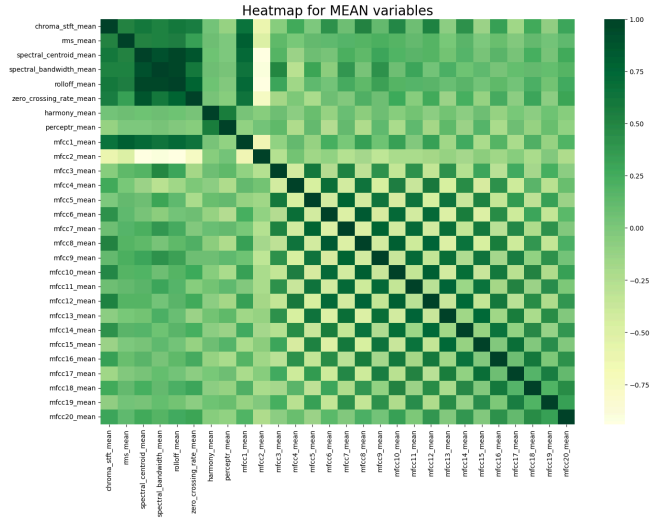

Fig. 3. Heatmap de correlação das características acústicas

Em seguida foi realizado o pré-processamento dos dados. Inicialmente, as etiquetas de gênero foram convertidas em valores numéricos utilizando o LabelEncoder da biblioteca scikit-learn, possibilitando o uso das classes como saída nos algoritmos. Além disso, as características numéricas foram escaladas com o MinMaxScaler para normalizá-las em um intervalo comum, evitando que variáveis com amplitudes maiores dominassem o aprendizado. Colunas irrelevantes, como o nome dos arquivos, foram removidas para eliminar dados redundantes e desnecessários. Os dados foram então divididos em conjuntos de treino (70%) e teste (30%) usando a função train_test_split, garantindo a validação durante a etapa de treinamento.

Para o treinamento dos modelos, foram escolhidos três algoritmos com diferentes abordagens de aprendizado supervisionado para classificação: Random Forest, CatBoost, e XGBoost. O Random Forest foi configurado com 1000 árvores, profundidade máxima de 10, e uma semente aleatória para reprodutibilidade, sendo eficiente em capturar relações não lineares entre variáveis. O CatBoost, otimizado para lidar com variáveis categóricas, foi configurado com uma métrica de avaliação de acurácia e uma função de perda para classificação multiclasse. Já o XGBoost, que utiliza boosting para combinar árvores de decisão, foi configurado com 1000 estimadores e uma taxa de aprendizado de 0,05, ideal para evitar overfitting enquanto constrói árvores sequencialmente.

A metodologia também incluiu a implementação de uma rede neural utilizando a biblioteca TensorFlow, com uma arquitetura simples, porém eficiente. A rede foi definida através do modelo sequencial, contendo as seguintes camadas: uma camada de entrada Flatten, que transforma os dados de entrada em uma única dimensão, seguida por uma camada densa (Dense) com 256 neurônios e ativação ReLU. Para melhorar o desempenho e a estabilidade do treinamento, foi adicionada uma camada de normalização em lote (BatchNormalization). Em seguida, outra camada densa com 128 neurônios e ativação ReLU foi incorporada, acompanhada por uma camada de dropout (Dropout) com taxa de 0,3 para evitar o overfitting. Por fim, a camada de saída, com 10 neurônios e ativação softmax, foi configurada para realizar a classificação dos gêneros musicais em categorias exclusivas.

Para cada modelo, foi feita a avaliação do seu desempenho com métricas calculadas, que incluem a Acurácia, que mede a proporção de previsões corretas; o F1-Score (Macro), que avalia o equilíbrio entre precisão e revocação para todas as classes de maneira uniforme; e a Precisão Top-3, que verifica se a classe correta está entre as três principais previsões do modelo. Além disso, foi gerada uma matriz de confusão para visualizar como o modelo classificou corretamente ou confundiu as diferentes classes.

IV. Resultados e Conclusões

Os resultados obtidos no trabalho demonstraram o desempenho dos diferentes algoritmos de aprendizado de máquina na classificação de gêneros musicais. Durante os testes, foi possível observar que os modelos apresentaram resultados variados, refletindo suas características específicas e adaptações aos dados fornecidos. A Tabela 1 apresenta os resultados obtidos.

TABELA I. Resultado das métricas de avaliação dos modelos

Modelo	Acurácia	F1-Score (Macro)	Precisão Top-3
Random Forest	0,78	0,78	0,96
CatBoost	0,83	0,83	0,96
Gradient Boosting	0,78	0,78	0,96
Rede Neural	0,77	0,76	0,95

a.

O algoritmo Random Forest mostrou um bom desempenho geral, alcançando uma acurácia de aproximadamente 78%. Isso é esperado devido à sua abordagem de combinar múltiplas Árvores de Decisão, reduzindo o risco de overfitting e aumentando a robustez do modelo. A Figura 4 apresenta a matriz de confusão do modelo com Random Forest.

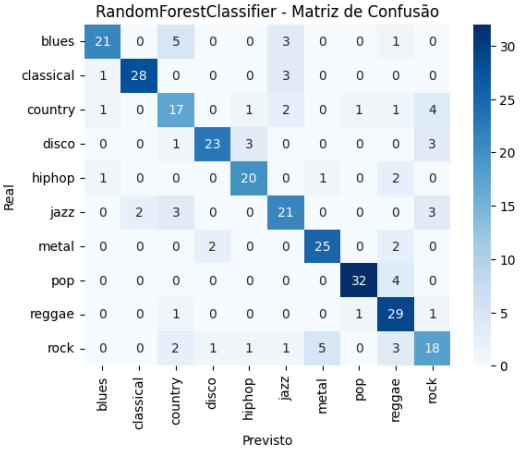

Fig. 4. Matriz de confusão do Random Forest

O CatBoost, por sua vez, destacou-se como o mais preciso, com uma acurácia em torno de 83%, o que evidencia sua eficiência em lidar com dados categóricos e complexidades inerentes ao conjunto de características acústicas. A Figura 5 apresenta a matriz de confusão do modelo com CatBoost.

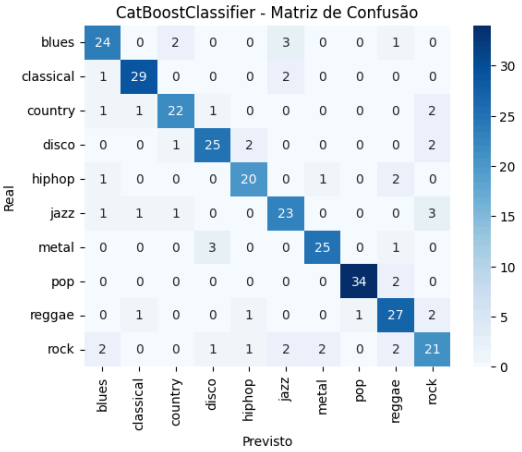

Fig. 5. Matriz de confusão do CatBoost

O Gradient Boosting também obteve um desempenho sólido, atingindo cerca de 79%, beneficiando-se de sua abordagem sequencial de otimização. A Figura 6 apresenta a matriz de confusão do modelo com Gradient Boost.

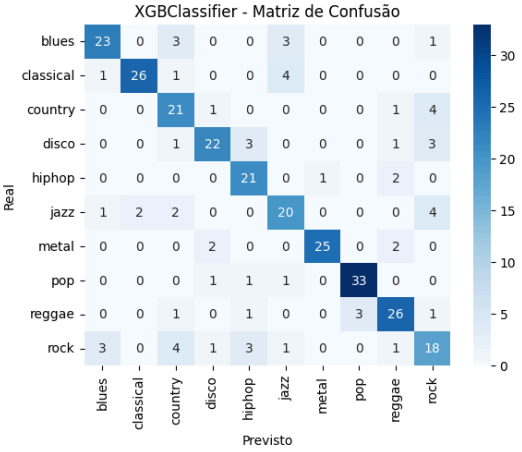

Fig. 6. Matriz de confusão do Gradient Boost

A rede neural alcançou uma acurácia de teste de aproximadamente 77%. Embora inferior ao CatBoost, a rede demonstrou ser promissora para tarefas futuras, considerando sua capacidade de se adaptar a dados maiores ou mais complexos. A Figura 7 apresenta a matriz de confusão da rede neural e a Figura 8 apresenta a evolução da acurácia e do erro durante o treinamento e o teste do modelo com a rede neural.

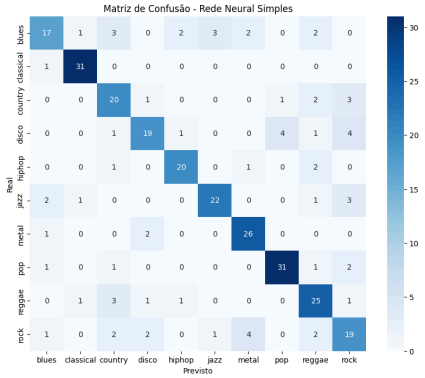

Fig. 7. Matriz de confusão da Rede Neural

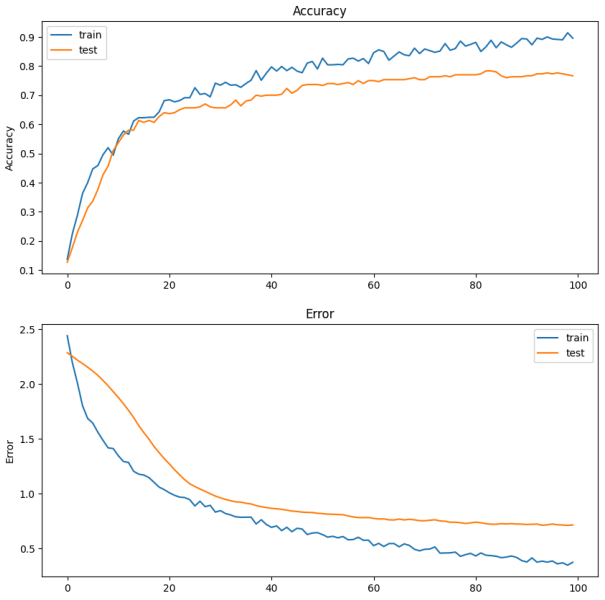

Fig. 8. Evolução da acurácia e do erro a Rede Neural

Os resultados deste trabalho mostram a eficácia de técnicas de processamento de sinais e aprendizado de máquina na classificação de gêneros musicais, destacando o CatBoost como o algoritmo mais eficiente, seguido por Random Forest, Gradient Boosting e Rede Neural. A análise de características presentes no dataset revelou padrões úteis para o treinamento de modelos, enquanto a rede neural desenvolvida mostrou potencial para cenários mais complexos. O estudo reforça a importância da escolha de características e algoritmos, mas aponta limitações na base GTZAN devido a inconsistências entre gêneros. Futuras pesquisas devem explorar bases mais amplas e arquiteturas avançadas, como CNNs, para melhorias significativas e aplicações práticas na indústria musical.

Referências

[1] X. Cai e H. Zhang, “Music genre classification based on auditory image, spectral and acoustic features”. Multimedia Syst., janeiro de 2022. Acesso em 05/11/2024. Disponível: https://doi.org/10.1007/s00530-021-00886-3

[2] “GTZAN Dataset - Music Genre Classification”. Kaggle: Your Machine Learning and Data Science Community. Acesso em 05/11/2024. Disponível: https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification/data

[3] Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. International Symposium on Music Information Retrieval, DOI:https://www.researchgate.net/publication/2552483_Mel_Frequency_Cepstral_Coefficients_for_Music_Modeling

[4] Athulya K., M. & Sindhu, S. "Deep Learning Based Music Genre Classification Using Spectrogram" in Proceedings of the 2nd International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS 2021), Kerala, India, 2021. DOI: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3883911

[5] G. Tzanetakis, A. Phan, and P. Cook, "A machine learning approach to music genre classification," Electronics Letters, vol. 56, no. 10, pp. 1234-1236, 2019. DOI: 10.1049/el.2019.4202.

[6] N. Ndou, R. Ajoodha, e A. Jadhav, "Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches," em 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), Toronto, Canadá, 21-24 de abril de 2021. DOI: https://doi.org/10.1109/IEMTRONICS52119.2021.9422487

[7] SINGH, Y.; BISWAS, A. Robustness of musical features on deep learning models for music genre classification. Expert Systems With Applications, v. 199, p. 116879, 2022. DOI: https://doi.org/10.1016/j.eswa.2022.116879

[8] Salihu, S. A., Lawal, I. O., Abikoye, O. C., Balogun, A. O., Mojeed, H. A., Usman-Hamza, F. E., & Akintola, A. G. (2023). Classification of Music Genres Using Catboost Algorithm. Sule Lamido University Journal of Science and Technology (SLUJST), 7(2), 17–26. https://doi.org/10.56471/slujst.v7i.472.