Análise comparativa de classificadores supervisionados clássicos para classificação automática de gêneros musicais

Lucas Henrique T. Ribeiro[1], **Marcelo R. Trierveiler**[1], **Ayslan T. Possebom**[1]

[1]Campus Paranavaí – Instituto Federal do Paraná (IFPR)
CEP 87703-536 – Paranavaí – PR – Brasil

lucashtribeiro2@gmail.com, marcelo.trierveiler@ifpr.edu.br

ayslan.possebom@ifpr.edu.br

Abstract. *Automatic music genre classification supports the organization of audio collections. This study compared five classical supervised classifiers: Perceptron, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Gaussian Naive Bayes. The experiments used 7,984 nonredundant tracks from FMA Small, covering eight genres and 518 previously extracted acoustic features. Performance was assessed through stratified ten-fold cross-validation, with macro F1-score as the primary metric. Differences were examined using the Friedman test and pairwise Wilcoxon tests with Holm correction. SVM achieved the highest mean macro F1-score (0.6129) and ranked first in all folds. It outperformed the other evaluated configurations, while no statistically significant difference was found between Perceptron and KNN.*

Resumo. *A classificação automática de gêneros musicais auxilia na organização de coleções de áudio. Este estudo comparou cinco classificadores supervisionados clássicos: Perceptron, máquina de vetores de suporte (SVM), método dos k-vizinhos mais próximos (KNN), árvore de decisão e Naive Bayes gaussiano. Os experimentos utilizaram 7.984 faixas não redundantes da FMA Small, distribuídas em oito gêneros e representadas por 518 características acústicas previamente extraídas. O desempenho foi avaliado por validação cruzada estratificada com dez partições, tendo a medida F1 macro como métrica principal. As diferenças foram examinadas pelos testes de Friedman e Wilcoxon, com correção de Holm. O SVM obteve a maior F1 macro média (0,6129) e ficou em primeiro lugar em todas as partições. O modelo superou as demais configurações avaliadas, enquanto Perceptron e KNN não apresentaram diferença estatisticamente significativa entre si.*

1. Introdução

O crescimento dos acervos musicais digitais ampliou a necessidade de métodos capazes de organizar e recuperar faixas em grande escala. A Recuperação de Informação Musical (do inglês *Music Information Retrieval* - MIR) estuda formas de representar e analisar esse conteúdo por meios computacionais [Downie 2003]. Entre suas aplicações está a classificação automática de gêneros, que pode auxiliar na indexação de coleções musicais e reduzir a dependência da rotulação manual [Tzanetakis e Cook 2002].

Essa classificação apresenta dificuldades próprias. Os limites entre gêneros não são rígidos, e uma mesma faixa pode reunir características associadas a categorias diferentes [Tzanetakis e Cook 2002]. Além disso, os algoritmos supervisionados

aprendem a partir dos atributos fornecidos e podem responder de maneiras distintas à mesma representação do áudio [Silla et al. 2008]. Por isso, comparar seus resultados sob condições experimentais comuns ajuda a avaliar a influência da escolha do classificador.

Este trabalho tem como objetivo comparar o desempenho de cinco classificadores supervisionados clássicos na classificação de gêneros musicais a partir de características acústicas previamente extraídas. Busca-se verificar se as diferenças de desempenho observadas são estatisticamente significativas. A comparação oferece uma referência para a escolha de modelos nesse conjunto de dados e nas condições adotadas, sem pressupor que os resultados se estendam a outras bases musicais.

O artigo está organizado da seguinte forma: a Seção 2 apresenta o referencial teórico; a Seção 3 apresenta os trabalhos correlatos; a Seção 4 descreve a base de dados e o procedimento experimental; a Seção 5 apresenta e discute os resultados; e a Seção 6 reúne as conclusões.

2. Referencial Teórico

A Recuperação de Informação Musical (do inglês *Music Information Retrieval* - MIR) estuda métodos computacionais para representar, organizar, buscar e analisar informações relacionadas à música [Downie 2003]. Essas informações podem provir de metadados, como título e artista, ou do próprio conteúdo sonoro. A análise do áudio permite descrever aspectos da música mesmo quando os dados textuais são insuficientes para uma tarefa de busca ou organização.

Uma das aplicações da MIR é a classificação automática de gêneros musicais. Nessa tarefa, um sistema recebe informações sobre uma faixa e lhe atribui um rótulo de gênero. Os gêneros auxiliam na organização de coleções, mas não possuem limites estritamente definidos: músicas de categorias diferentes podem compartilhar instrumentação, padrões rítmicos e elementos harmônicos [Tzanetakis e Cook 2002]. Por isso, as propriedades do áudio oferecem indícios para a classificação, sem determinar de forma inequívoca o gênero de uma faixa.

Para que o conteúdo sonoro possa ser processado por classificadores, ele é representado por atributos numéricos extraídos do sinal. Coeficientes cepstrais na escala Mel (MFCC), por exemplo, descrevem aspectos do espectro associados ao timbre; características cromáticas representam a distribuição da energia entre classes de notas; e medidas como energia quadrática média e centroide espectral descrevem, respectivamente, a intensidade do sinal e a distribuição de sua energia nas frequências. Esses descritores podem ser resumidos por estatísticas calculadas ao longo da faixa, formando um vetor de atributos para cada música [Tzanetakis e Cook 2002, McFee et al. 2015].

Os classificadores supervisionados utilizam exemplos com gêneros conhecidos para relacionar os atributos das músicas aos rótulos que devem prever. O Perceptron combina os atributos por meio de pesos, ajustados durante o treinamento quando ocorrem erros de classificação. Essa combinação produz fronteiras de decisão lineares, que delimitam as regiões atribuídas às classes [Rosenblatt 1958]. A máquina de vetores de suporte (SVM) busca uma separação com margem ampla entre as classes, orientada pelas amostras mais próximas da fronteira. Funções de núcleo permitem representar separações não lineares; o núcleo de função de base radial (RBF), por exemplo, calcula similaridades a partir da distância entre amostras [Cortes e Vapnik 1995].

O método dos k-vizinhos mais próximos (KNN) compara uma nova faixa com os exemplos de treinamento e atribui a classe predominante entre seus k vizinhos mais próximos no espaço de atributos [Cover e Hart 1967]. A árvore de decisão utiliza uma sequência de perguntas sobre os valores dos atributos, conduzindo cada amostra até uma classificação. Árvores muito complexas podem se ajustar excessivamente às particularidades do treinamento, fenômeno denominado sobreajuste, prejudicando a classificação de novos exemplos [Breiman et al. 1984].

O Naive Bayes gaussiano estima a probabilidade de cada gênero a partir dos atributos observados e seleciona a classe mais provável. Para simplificar esse cálculo, considera os atributos condicionalmente independentes dado o gênero e modela os valores de cada atributo por uma distribuição gaussiana em cada classe [John e Langley 1995]. Como os cinco modelos utilizam estratégias distintas, podem apresentar desempenhos diferentes diante da mesma representação acústica.

A avaliação verifica como os classificadores se comportam em faixas que não participaram de seu treinamento. Na validação cruzada, os dados são divididos sucessivamente em conjuntos de treino e teste; a estratificação preserva aproximadamente a proporção das classes em cada divisão [Kohavi 1995]. A acurácia mede a proporção total de classificações corretas, enquanto precisão e revocação examinam diferentes aspectos dos acertos em cada classe. A medida F1 combina precisão e revocação, e sua média macro atribui o mesmo peso a todos os gêneros [Sokolova e Lapalme 2009].

Além das métricas, testes estatísticos permitem examinar as diferenças de desempenho observadas entre os modelos. O teste de Friedman compara suas posições relativas nas partições avaliadas. Caso indique uma diferença global, comparações pareadas podem identificar entre quais modelos ela ocorre; a correção de Holm ajusta os valores de p para considerar as múltiplas comparações [Demšar 2006, Garcia e Herrera 2008].

3. Trabalhos correlatos

Um sistema baseado em características de timbre e espectrotemporais combinou seleção de atributos por um ranqueador SVM e classificação por SVM com núcleo RBF [Lim et al. 2012]. Os experimentos utilizaram as bases GTZAN e ISMIR2004, esta última associada à Sociedade Internacional para Recuperação de Informação Musical (do inglês *International Society for Music Information Retrieval* - ISMIR), nas quais o método alcançou acurácias de 87,4% e 89,9%, respectivamente.

A combinação de características acústicas com descritores extraídos de representações visuais do áudio, como espectrogramas, gammatonegramas e imagens rítmicas, foi investigada em tarefas de classificação de áudio [Nanni et al. 2017]. Diferentes classificadores SVM foram combinados em um conjunto, avaliado na Base de Dados de Música Latina (do inglês *Latin Music Database* - LMD), ISMIR2004 e GTZAN. Os resultados indicaram que a fusão de representações complementares pode superar o uso isolado de características acústicas ou visuais.

Outra abordagem combinou características acústicas, espectrais e imagens inspiradas no sistema auditivo humano [Cai e Zhang 2022]. O estudo comparou inicialmente SVM, KNN e classificação por representação esparsa, identificando melhor desempenho do SVM. Com a fusão das características, foram alcançadas acurácias de 91,8% na GTZAN, 91,2% na ISMIR2004 e 65,2% na Homburg.

Em comum, esses trabalhos demonstram que o desempenho depende tanto da representação do áudio quanto do classificador utilizado, com recorrente destaque para o SVM. Diferentemente das abordagens voltadas à criação ou fusão de características, o presente estudo concentra-se na comparação controlada de cinco classificadores supervisionados clássicos sobre os mesmos 518 atributos da *FMA Small*. A utilização das mesmas partições, de métricas macro e de testes estatísticos busca identificar não apenas qual modelo apresenta a maior média, mas se as diferenças observadas são significativas.

4. Metodologia e Configuração Experimental

4.1. Caracterização da pesquisa

O estudo consiste em uma pesquisa aplicada, de abordagem quantitativa e caráter experimental e comparativo. Foram utilizados dados secundários da FMA Small para avaliar cinco classificadores supervisionados por meio de métricas de classificação e testes estatísticos.

4.2. Base de dados e características acústicas

O experimento utilizou o subconjunto Small do Arquivo de Música Livre (do inglês *Free Music Archive* - FMA), denominado FMA *Small* [Defferrard et al. 2017, Defferrard et al. 2018]. Em sua configuração original, esse subconjunto contém 8.000 faixas musicais distribuídas igualmente entre oito gêneros: *Electronic*, *Experimental*, *Folk*, *Hip-Hop*, *Instrumental*, *International*, *Pop* e *Rock*. Cada classe possui originalmente 1.000 faixas.

Para cada faixa, foram utilizados o identificador da música, o rótulo de gênero e 518 atributos acústicos previamente extraídos pelos autores da base com o auxílio da biblioteca Python *Librosa*. Dessa forma, não foi necessário processar diretamente os arquivos de áudio nem realizar uma nova extração de características. Todas as características acústicas são numéricas e constituem os atributos de entrada dos classificadores. O identificador da faixa não foi utilizado como atributo, e o gênero corresponde ao rótulo categórico a ser previsto.

Os atributos estão organizados nos grupos *chroma_cens*, *chroma_cqt*, *chroma_stft*, *mfcc*, *rmse*, *spectral_bandwidth*, *spectral_centroid*, *spectral_contrast*, *spectral_rolloff*, *tonnetz* e *zcr*. Esses atributos incluem estatísticas agregadas, como média, desvio padrão, assimetria, curtose, mediana, mínimo e máximo.

Optou-se pela manutenção das 518 características disponíveis, sem a realização de uma seleção manual de atributos, pois esse procedimento poderia introduzir critérios arbitrários na comparação.

4.3. Pré-processamento dos dados

Inicialmente, foi realizada uma inspeção da estrutura do conjunto de dados. Foram verificados a quantidade de registros e atributos, os tipos das variáveis, a distribuição das classes, a presença de valores ausentes ou infinitos, a existência de identificadores repetidos e a ocorrência de atributos constantes. Não foram encontrados valores ausentes, valores infinitos, identificadores de faixas duplicados ou características constantes.

Em cada grupo de registros com valores idênticos nas 518 características acústicas e no rótulo de gênero, foi mantida a primeira ocorrência na ordem do arquivo de entrada,

desconsiderando o identificador da faixa na comparação. Foram removidos 16 registros excedentes, distribuídos em 12 grupos, antes da criação das partições de validação. O conjunto final passou a conter 7.984 instâncias, com distribuição aproximadamente balanceada, variando entre 993 e 1.000 registros por gênero.

Foram mantidos registros com atributos idênticos e rótulos diferentes, pois não atendiam ao critério de redundância adotado. Foi identificado um par nessas condições, cujas duas faixas foram alocadas à mesma partição de teste.

Os atributos foram padronizados com *StandardScaler* nos modelos Perceptron, SVM, KNN e Naive Bayes gaussiano. Em cada partição, a média e o desvio padrão foram calculados exclusivamente no treinamento, dentro de um *pipeline*, e utilizados para transformar os conjuntos de treinamento e teste. A árvore de decisão dispensou essa transformação por utilizar regras de particionamento que não dependem diretamente da escala dos atributos.

4.4. Classificadores e parametrização

Foram avaliados cinco classificadores supervisionados clássicos: Perceptron, SVM, KNN, árvore de decisão e Naive Bayes gaussiano.

Foram comparadas configurações de referência com hiperparâmetros fixos, sem otimização, apresentadas na Tabela 1. Essa escolha pode favorecer determinados modelos, pois eles respondem de maneiras distintas à parametrização. As conclusões restringem-se, portanto, às configurações avaliadas, sem estabelecer o desempenho máximo de cada algoritmo.

Tabela 1. Configuração dos classificadores utilizados no experimento.

Classificador	Configuração principal
Perceptron	Máximo de 1.000 iterações, tolerância de 0,001, embaralhamento dos registros e semente aleatória igual a 42.
SVM	Núcleo RBF, parâmetro de regularização $C = 1,0$, parâmetro γ definido como *scale* e função de decisão apresentada no formato *one-versus-rest*.
KNN	Cinco vizinhos, pesos uniformes e distância de Minkowski com $p = 2$, correspondente à distância euclidiana.
Árvore de decisão	Critério de impureza de Gini, estratégia de divisão *best*, sem limitação prévia de profundidade e semente aleatória igual a 42.
Naive Bayes gaussiano	Parâmetro de suavização de variância igual a 10^{-9}.

No SVC, a classificação multiclasse utiliza internamente comparações entre pares de classes (*one-versus-one*); o parâmetro `decision_function_shape='ovr'` define o formato de apresentação dos escores de decisão.

4.5. Protocolo de validação

O desempenho dos modelos foi avaliado por meio de validação cruzada estratificada com dez partições. Foi utilizado o procedimento *StratifiedKFold*, com embaralhamento dos registros e semente aleatória igual a 42.

A divisão oficial de treinamento, validação e teste da FMA não foi utilizada. As partições foram construídas por estratificação dos gêneros, sem agrupamento por artista. Assim, faixas de um mesmo artista podem estar presentes simultaneamente nos conjuntos de treinamento e teste. O protocolo não assegura, portanto, a avaliação da generalização para artistas não observados durante o treinamento.

A estratificação preservou aproximadamente a proporção dos oito gêneros musicais em cada partição. Ao longo das dez iterações, cada registro foi utilizado uma vez no conjunto de teste e nove vezes nos conjuntos de treinamento.

As partições foram geradas uma única vez e reutilizadas por todos os classificadores. Dessa forma, cada modelo foi treinado e avaliado exatamente sobre os mesmos conjuntos, permitindo uma comparação pareada e reduzindo a influência de diferentes divisões dos dados.

Como foram utilizados cinco classificadores e dez partições, o protocolo resultou em 50 execuções de treinamento e teste. Em cada execução, uma nova instância do *pipeline* foi criada, evitando o compartilhamento de parâmetros aprendidos entre modelos ou partições.

As predições realizadas sobre os conjuntos de teste foram armazenadas como predições *out-of-fold*. Assim, para cada classificador, foi obtida uma predição de teste para cada uma das 7.984 faixas utilizadas no experimento.

4.6. Métricas de avaliação

A avaliação foi realizada por meio das métricas de acurácia, precisão macro, revocação macro e medida F1 macro. A acurácia representa a proporção total de previsões corretas. A precisão indica a proporção de classificações corretas entre os registros atribuídos a uma classe, enquanto a revocação representa a proporção de registros corretamente identificados entre aqueles que pertencem efetivamente à classe.

Nas versões macro, as métricas foram calculadas individualmente para cada gênero e posteriormente agregadas por média aritmética, atribuindo o mesmo peso a todas as classes. A medida F1 macro foi definida previamente como a métrica principal do experimento, por combinar precisão e revocação e permitir a avaliação equilibrada das oito classes.

Os resultados obtidos nas dez partições foram consolidados por meio de média, desvio padrão, mediana, valor mínimo e valor máximo. Também foram calculadas métricas individuais por gênero.

As predições *out-of-fold* foram utilizadas para a geração das matrizes de confusão de cada classificador. Foram produzidas matrizes absolutas e matrizes normalizadas pela classe verdadeira, permitindo analisar tanto o número de classificações quanto a proporção de acertos e erros para cada gênero musical.

4.7. Análise estatística

O teste de Friedman foi aplicado aos valores de medida F1 macro obtidos pelos cinco classificadores nas dez partições. Nesse procedimento, os classificadores foram ordenados dentro de cada partição, atribuindo-se o rank 1 ao melhor resultado. Em casos de empate, foi utilizado o rank médio das posições correspondentes.

A hipótese nula considerada no teste de Friedman estabeleceu que os classificadores apresentavam desempenho equivalente. A hipótese alternativa estabeleceu que pelo menos um dos modelos apresentava desempenho diferente dos demais.

Foi adotado nível de significância de $\alpha = 0,05$. Dessa forma, valores de p inferiores a 0,05 foram utilizados como critério para rejeição da hipótese nula.

O protocolo estatístico estabeleceu que, caso o teste de Friedman indicasse diferença estatisticamente significativa, seriam realizadas comparações pareadas por meio do teste de postos sinalizados de Wilcoxon. Considerando os cinco classificadores, esse procedimento corresponde a dez comparações entre pares.

O teste de Wilcoxon foi aplicado na modalidade bilateral, desconsiderando os pares com diferença nula. O método de cálculo do valor de p foi selecionado automaticamente pela biblioteca SciPy.

Para controlar o aumento da probabilidade de erro do tipo I decorrente das múltiplas comparações, os valores de p obtidos pelo teste de Wilcoxon foram ajustados pelo procedimento de Holm. A significância das diferenças entre pares foi determinada a partir dos valores ajustados.

Os resultados estatísticos foram interpretados especificamente no contexto da base FMA Small e do protocolo experimental adotado. Considerou-se, ainda, que as partições de uma validação cruzada compartilham parcialmente registros de treinamento e, portanto, não constituem conjuntos experimentais completamente independentes.

4.8. Ambiente de implementação e reprodutibilidade

O experimento foi implementado em linguagem Python 3.11. Foram utilizadas as bibliotecas *NumPy* e *pandas* para manipulação dos dados, *scikit-learn* para implementação dos classificadores, dos *pipelines*, da validação cruzada e das métricas, *SciPy* para os testes estatísticos, *statsmodels* para a correção de Holm e *Matplotlib* para a geração das visualizações.

O código foi organizado em módulos para carregamento e validação dos dados, definição dos modelos, validação cruzada, análise estatística e visualização. Os resultados foram armazenados em arquivos CSV, e as figuras foram exportadas em PNG.

Foram utilizados valores fixos de semente aleatória nos procedimentos que possuíam componentes pseudoaleatórios. Além disso, as versões das dependências utilizadas foram registradas em um arquivo específico, permitindo a reprodução do ambiente experimental.

O código-fonte, o script de preparação dos dados e os resultados do experimento estão disponíveis em repositório público[1], acompanhado de instruções para instalação das dependências e execução das etapas.

5. Resultados e Discussão

5.1. Desempenho geral dos classificadores

A Tabela 2 apresenta o desempenho médio dos classificadores nas dez partições da validação cruzada. Os resultados são expressos como média ± desvio padrão. O rank

[1] https://github.com/Luqui05/classificacao-generos-musicais-fma

médio representa a posição ocupada por cada modelo nas partições, de modo que valores menores indicam melhor desempenho.

Tabela 2. Desempenho médio dos classificadores nas dez partições.

Classificador	Acurácia	Precisão macro	Revocação macro	F1 macro	Rank médio
SVM	$\mathbf{0{,}6137 \pm 0{,}0055}$	$\mathbf{0{,}6147 \pm 0{,}0057}$	$\mathbf{0{,}6137 \pm 0{,}0054}$	$\mathbf{0{,}6129 \pm 0{,}0050}$	**1,0**
Perceptron	$0{,}4792 \pm 0{,}0135$	$0{,}4830 \pm 0{,}0154$	$0{,}4792 \pm 0{,}0135$	$0{,}4766 \pm 0{,}0139$	2,4
KNN	$0{,}4820 \pm 0{,}0144$	$0{,}4920 \pm 0{,}0171$	$0{,}4818 \pm 0{,}0144$	$0{,}4710 \pm 0{,}0160$	2,6
Árvore de decisão	$0{,}3659 \pm 0{,}0275$	$0{,}3677 \pm 0{,}0279$	$0{,}3658 \pm 0{,}0275$	$0{,}3660 \pm 0{,}0278$	4,0
Naive Bayes gaussiano	$0{,}2989 \pm 0{,}0143$	$0{,}4052 \pm 0{,}0186$	$0{,}2985 \pm 0{,}0143$	$0{,}2509 \pm 0{,}0096$	5,0

O SVM apresentou o melhor desempenho em todas as métricas, com medida F1 macro média de 0,6129 e rank médio igual a 1,0, ocupando a primeira posição em todas as partições. O baixo desvio padrão de 0,0050 indica baixa variação da medida F1 macro nas diferentes partições.

Perceptron e KNN apresentaram resultados intermediários e próximos. O KNN alcançou acurácia, precisão e revocação ligeiramente superiores, enquanto o Perceptron obteve maior medida F1 macro. A diferença de aproximadamente 0,0057 entre os dois modelos é pequena em relação à variação observada entre as partições.

A árvore de decisão apresentou desempenho inferior e a maior dispersão entre os modelos. O Naive Bayes gaussiano obteve a menor F1 macro. Sua precisão macro (0,4052) foi superior à própria revocação macro (0,2985), mas inferior à precisão do SVM, do Perceptron e do KNN.

5.2. Desempenho por gênero e padrões de confusão

A Figura 1 apresenta a medida F1 de cada classificador nos oito gêneros musicais, calculada a partir das predições *out-of-fold*.

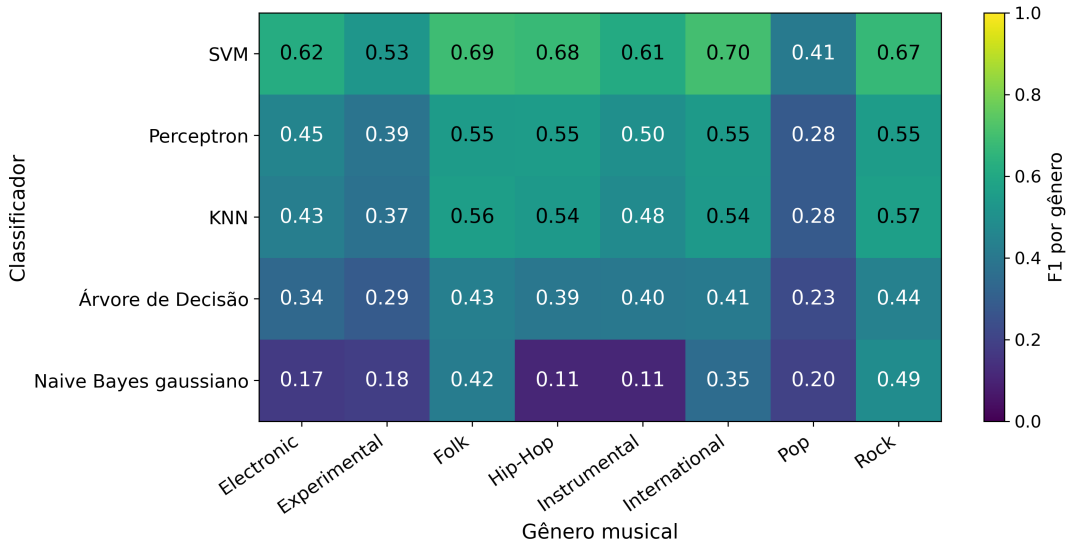


Figura 1. Medida F1 por gênero musical e classificador.

O SVM obteve o melhor resultado em todas as classes. Seus maiores valores foram observados em *International*, *Folk*, *Hip-Hop* e *Rock*, com resultados entre 0,67 e

0,70. O menor desempenho ocorreu em *Pop*, com aproximadamente 0,41, seguido por *Experimental*, com 0,53.

O gênero *Pop* apresentou resultados reduzidos em todos os modelos. KNN e Perceptron tiveram vantagens em gêneros distintos: o KNN obteve maior F1 em *Folk* e *Rock*, enquanto o Perceptron se destacou em *Electronic*, *Experimental*, *Hip-Hop*, *Instrumental* e *International*. Em *Pop*, ambos apresentaram F1 de aproximadamente 0,28.

A árvore de decisão permaneceu abaixo dos três primeiros modelos em todos os gêneros. O Naive Bayes gaussiano apresentou o comportamento mais desigual, com resultados relativamente elevados para *Rock* e *Folk*, mas valores reduzidos em classes como *Hip-Hop*, *Instrumental* e *Electronic*.

A Figura 2 apresenta a matriz de confusão normalizada do SVM, selecionado por ter obtido o melhor desempenho geral. Os valores da diagonal representam a proporção de classificações corretas em cada gênero.

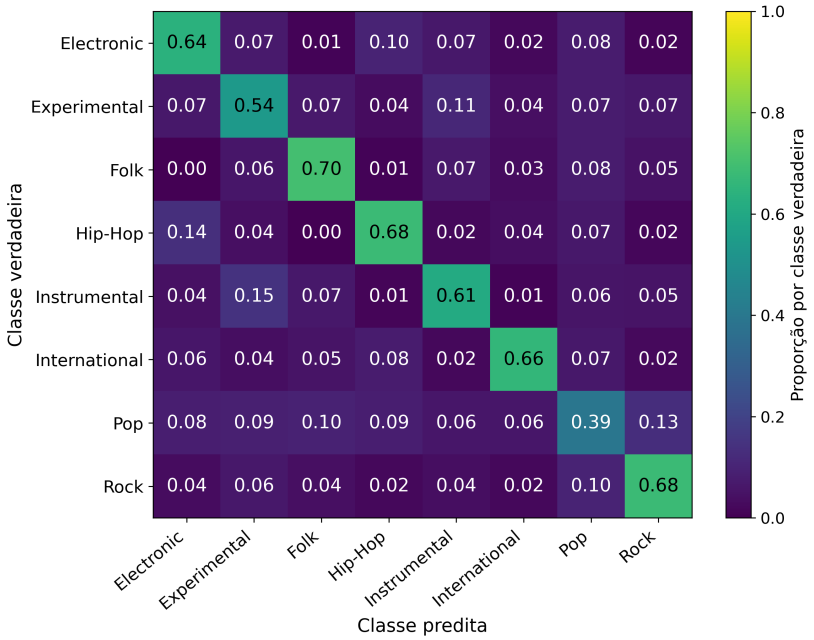


Figura 2. Matriz de confusão normalizada do classificador SVM.

O SVM classificou corretamente 70% das faixas de *Folk*, 68% de *Hip-Hop* e *Rock*, 66% de *International*, 64% de *Electronic* e 61% de *Instrumental*. Os menores valores ocorreram para *Experimental*, com 54%, e *Pop*, com 39%.

Entre as confusões mais frequentes, 15% das faixas de *Instrumental* foram classificadas como *Experimental*, enquanto 14% das faixas de *Hip-Hop* foram atribuídas a *Electronic*. A classe *Pop* apresentou erros distribuídos entre diferentes gêneros, principalmente *Rock*, *Folk*, *Experimental* e *Hip-Hop*. Esse comportamento indica que os erros não ocorreram de forma uniforme e que determinadas classes apresentaram maior sobreposição dentro da representação acústica utilizada.

5.3. Análise estatística

O teste de Friedman foi aplicado aos valores de medida F1 macro obtidos nas dez partições. O resultado apresentou estatística igual a 38,08 e valor de p aproximadamente igual a $1{,}08 \times 10^{-7}$. Como $p < 0{,}05$, a hipótese nula de equivalência entre os classificadores foi rejeitada.

As comparações pareadas de Wilcoxon, com correção de Holm, identificaram diferenças significativas em todos os pares, exceto entre Perceptron e KNN, cujo valor de p ajustado foi 0,4316. O SVM apresentou desempenho superior aos demais modelos nas dez partições.

Nas nove comparações significativas, as diferenças mantiveram o mesmo sentido nas dez partições, sem diferenças nulas ou empates nos valores absolutos. Nessas condições, o teste bilateral exato produz o valor mínimo de p igual a 2/1024, aproximadamente 0,00195. A correção de Holm para as dez comparações resultou em valores ajustados iguais a aproximadamente 0,0195, explicando sua repetição.

Os resultados formaram três níveis principais de desempenho: o SVM na primeira posição; Perceptron e KNN em um grupo intermediário sem diferença estatisticamente significativa; e a árvore de decisão e o Naive Bayes gaussiano nas posições inferiores.

5.4. Síntese e limitações

A superioridade do SVM é compatível com a capacidade do núcleo RBF de representar fronteiras de decisão não lineares em um espaço de alta dimensionalidade. Além de alcançar o melhor desempenho, o modelo apresentou a menor variação entre as partições. Entretanto, sua medida F1 macro próxima de 0,61 indica que a classificação dos oito gêneros permaneceu uma tarefa complexa.

O desempenho semelhante entre Perceptron e KNN foi alcançado por estratégias distintas. O primeiro utiliza uma separação linear, enquanto o segundo depende da proximidade entre as instâncias. A alta dimensionalidade pode ter dificultado o uso de distâncias pelo KNN. A árvore de decisão, utilizada sem restrição de profundidade ou poda, apresentou maior variação. Já o Naive Bayes gaussiano pode ter sido prejudicado pela presença de relações e correlações entre as características acústicas, contrariando parcialmente sua hipótese de independência.

Os resultados também evidenciaram que a dificuldade depende do gênero analisado. *Pop* foi a classe mais problemática, enquanto *Folk*, *International* e *Rock* apresentaram melhor separação em diferentes modelos. Essas diferenças podem decorrer da sobreposição acústica e conceitual entre gêneros, embora o experimento não permita determinar isoladamente suas causas.

O destaque do SVM converge qualitativamente com resultados anteriores [Lim et al. 2012, Cai e Zhang 2022]. Entretanto, os valores de desempenho não são diretamente comparáveis, pois as bases, as características e os protocolos diferem. Esses estudos exploraram seleção ou combinação de atributos, enquanto outro trabalho investigou a fusão de características acústicas e visuais [Nanni et al. 2017]. Essas abordagens contextualizam a limitação do presente experimento, que manteve as 518 características da FMA Small sem seleção ou redução.

Entre as limitações do estudo estão o uso de apenas uma base de dados, a utilização de características previamente extraídas, a ausência de seleção ou redução de atributos e a adoção de hiperparâmetros fixos. Também foi utilizada uma única configuração de embaralhamento dos dados. Além disso, os folds compartilham registros de treinamento e não representam experimentos completamente independentes.

A ausência de separação por artista também pode favorecer os resultados, pois faixas de um mesmo artista podem compartilhar características de produção e estilo. Uma avaliação com agrupamento por artista permitiria investigar o desempenho em músicas de artistas não observados no treinamento.

Por fim, os gêneros musicais possuem limites parcialmente subjetivos, e uma mesma faixa pode apresentar características associadas a diferentes categorias. Assim, os resultados indicam que o SVM foi o classificador mais adequado entre as configurações avaliadas, mas essa conclusão permanece restrita à base *FMA Small*, às 518 características utilizadas e ao protocolo experimental adotado.

6. Conclusão

Este trabalho comparou cinco classificadores supervisionados na classificação de gêneros musicais, usando 518 características da base FMA Small e validação cruzada estratificada em dez partições. O SVM obteve o melhor resultado, com F1 macro médio de 0,6129 e primeira colocação em todas as partições. As análises estatísticas apontaram diferenças significativas entre o SVM e os demais modelos, enquanto Perceptron e KNN não diferiram significativamente. Pop foi a classe mais difícil para a maioria dos classificadores.

Entre as configurações avaliadas, o SVM apresentou o desempenho mais adequado. Essa conclusão se limita à base, às características e aos hiperparâmetros utilizados. Trabalhos futuros podem avaliar a divisão dos dados por artista, testar outras bases musicais e otimizar os hiperparâmetros com validação cruzada aninhada.

Referências

Breiman, L., Friedman, J., Olshen, R. A., e Stone, C. J. (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, 1st edition.

Cai, X. e Zhang, H. (2022). Music genre classification based on auditory image, spectral and acoustic features. *Multimedia Systems*, 28(3):779–791.

Cortes, C. e Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cover, T. e Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Defferrard, M., Benzi, K., Vandergheynst, P., e Bresson, X. (2017). FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 316–323.

Defferrard, M., Mohanty, S. P., Carroll, S. F., e Salathé, M. (2018). Learning to recognize musical genre from audio: Challenge overview. In *Companion proceedings of the web conference 2018*, pages 1921–1922.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30.

Downie, J. S. (2003). Music information retrieval. *Annual review of information science and technology*, 37(1):295–340.

Garcia, S. e Herrera, F. (2008). An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of machine learning research*, 9(89):2677–2694.

John, G. H. e Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1145, Montreal, Canada. Morgan Kaufmann.

Lim, S.-C., Lee, J.-S., Jang, S.-J., Lee, S.-P., e Kim, M. Y. (2012). Music-genre classification system based on spectro-temporal features and feature selection. *IEEE Transactions on Consumer Electronics*, 58(4):1262–1268.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., e Nieto, O. (2015). librosa: Audio and music signal analysis in Python. In *Proceedings of the 14th Python in Science Conference*, pages 18–25.

Nanni, L., Costa, Y. M. G., Lucio, D. R., Silla, C. N., e Brahnam, S. (2017). Combining visual and acoustic features for audio classification tasks. *Pattern Recognition Letters*, 88:49–56.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408.

Silla, C. N., Koerich, A. L., e Kaestner, C. A. A. (2008). A Machine Learning Approach to Automatic Music Genre Classification. *Journal of the Brazilian Computer Society*, 14(3):7–18.

Sokolova, M. e Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

Tzanetakis, G. e Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302.