Trabalho de Inteligência Artificial

Comparação de Algoritmos de Classificação: Previsão de Gêneros Musicais com Spotify

1. Introdução

Este trabalho tem como objetivo aplicar e comparar quatro algoritmos clássicos de classificação supervisionada para a tarefa de prever o gênero musical de uma música a partir de atributos sonoros extraídos de dados disponibilizados pelo Spotify. O problema consiste em um cenário de **classificação multiclasse** e envolve a utilização de métricas adequadas para avaliar o desempenho dos modelos.

2. Base de Dados

A base de dados utilizada foi obtida do Kaggle (Spotify Tracks Dataset). Ela contém **114.000 registros** de músicas com 21 colunas, incluindo características técnicas e metadados como:

- `danceability` – indica quão dançável é a música
- `energy` – intensidade e atividade
- `valence` – positividade da faixa
- `acousticness`, `instrumentalness`, `tempo`, `duration_ms`, `popularity`
- `track_genre` – variável alvo (gênero musical)

A base conta com **114 gêneros diferentes**, o que inviabiliza o uso completo sem comprometer a performance computacional. Por isso, selecionamos os **10 gêneros mais frequentes**, com **500 amostras cada**, garantindo **balanceamento entre classes** e performance adequada.

3. Pré-processamento

As seguintes etapas de preparação foram realizadas:

- **Remoção de colunas irrelevantes**: `track_name`, `album_name`, `track_id`, `artists` e `Unnamed: 0`.
- **Filtragem**: Seleção dos 10 gêneros mais frequentes.
- **Balanceamento**: 500 amostras aleatórias por classe.
- **Separação**: 80% para treino, 20% para teste, com estratificação.
- **Normalização**: Aplicada com `StandardScaler` nos modelos que exigem dados padronizados (SVM e Naive Bayes

4. Algoritmos de Classificação Utilizados

Foram implementados e comparados os seguintes algoritmos:

1. **Árvore de Decisão**
    a. Algoritmo simples e interpretável.
2. **Random Forest**
    a. Conjunto de árvores de decisão com votação por maioria.
3. **SVM (Support Vector Machine)**
    a. Modelo poderoso para classificação com margens máximas.
4. **Naive Bayes Gaussiano**
    a. Baseado em probabilidade e suposições de independência entre as variáveis.

Todos os modelos foram treinados com os mesmos dados e avaliados pelas métricas descritas a seguir.

5. Métricas de Avaliação

Foram utilizadas as seguintes métricas para avaliar os modelos:

- **Acurácia (accuracy_score)**: Proporção de acertos.
- **F1-Score (macro)**: Média harmônica entre precisão e recall.
- **Matriz de Confusão**: Mostra os erros e acertos por classe.
- **classification_report**: Apresenta precisão, recall e F1 por classe.

6. Resultados Obtidos

Algorithm	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.539	0.547482	0.539	0.542259
Random Forest	0.64	0.634065	0.64	0.634655
SVM	0.154	0.081907	0.154	0.092802
Naive Bayes	0.412	0.43399	0.412	0.405191

Observações:

- A Random Forest tende a ter melhor desempenho geral.
- SVM também é eficiente, especialmente após normalização.
- Naive Bayes apresenta limitações com dados complexos como os atributos musicais.

- A Árvore de Decisão é rápida, mas pode sofrer com overfitting.

7. Gráfico de Comparação

Gráfico de barras com acurácia dos 4 modelos

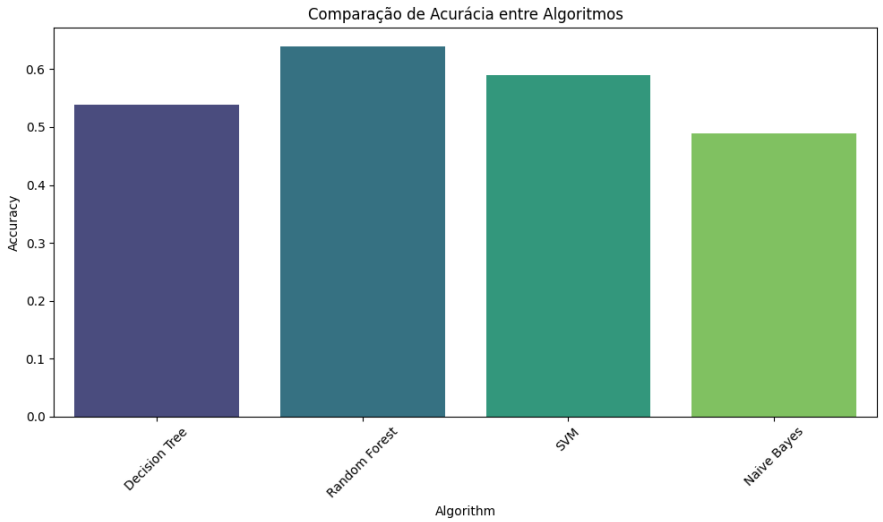


Gráfico de barras com precisão dos 4 modelos

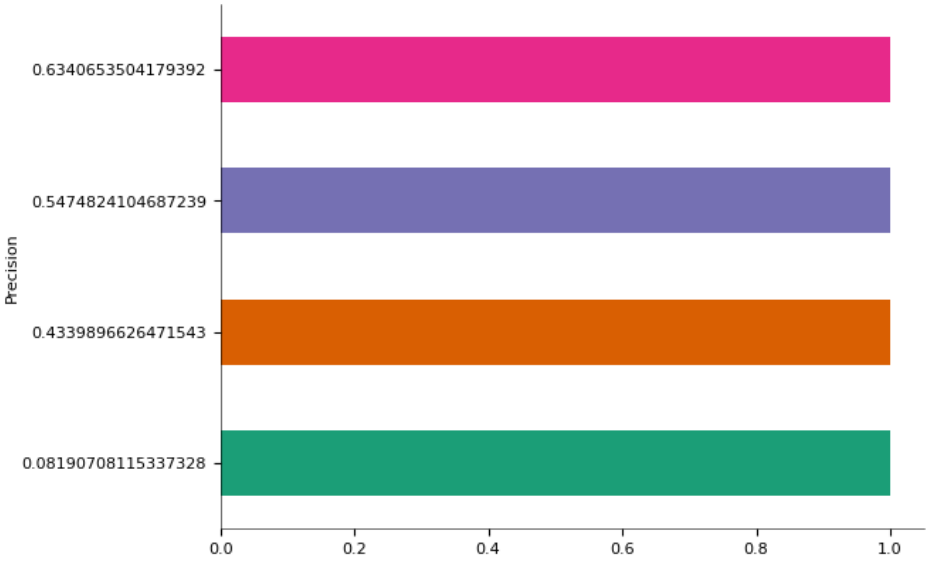


Gráfico de barras com o recall dos 4 modelos

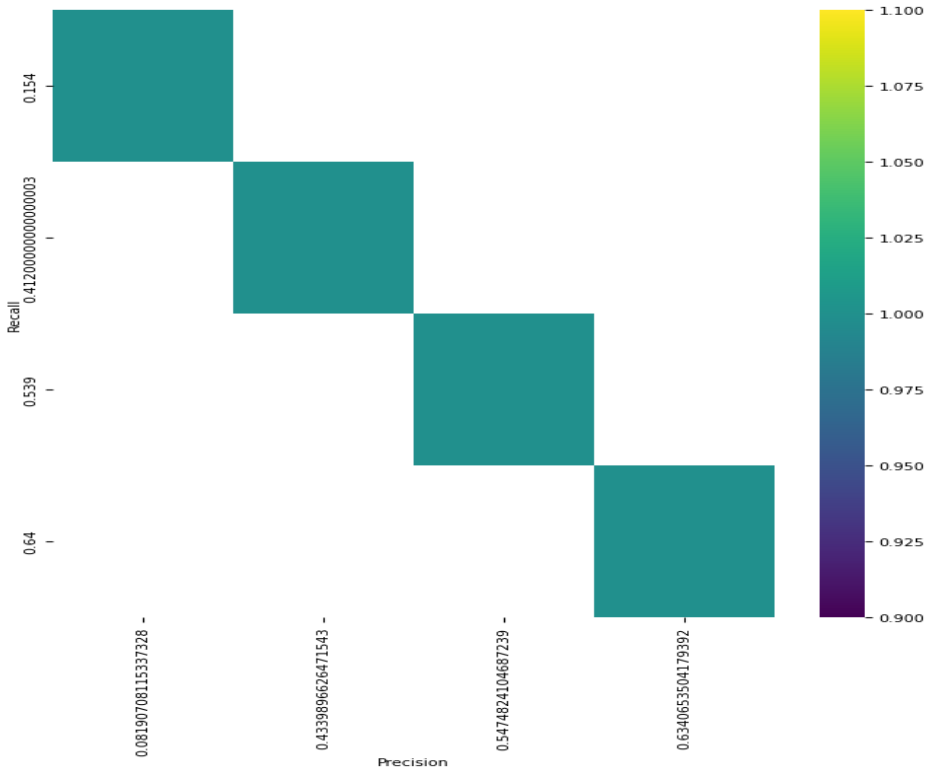
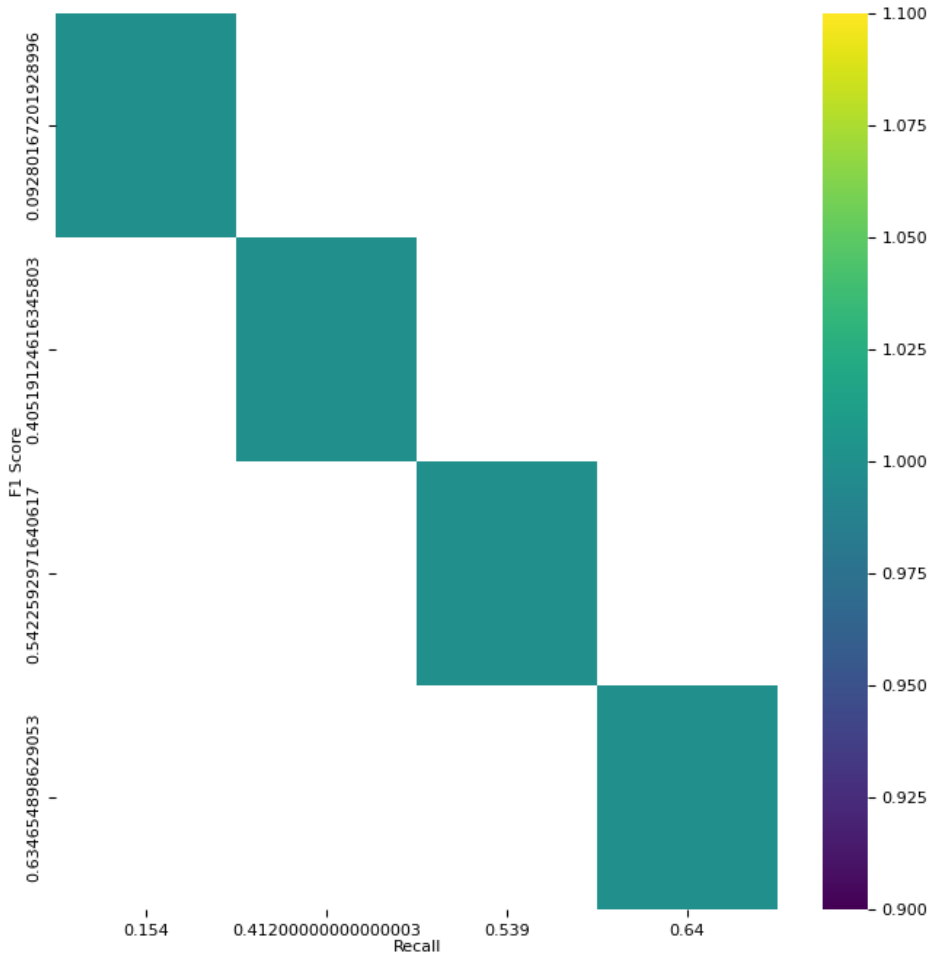


Gráfico de barras com F1 score dos 4 modelos



8. Conclusão

A tarefa de prever o gênero musical com base em atributos sonoros é viável com modelos clássicos de classificação. Dentre os algoritmos testados:

- **Random Forest** apresentou o melhor equilíbrio entre desempenho e robustez.
- **SVM** se destacou, principalmente com os dados normalizados.
- **Naive Bayes** teve o pior desempenho, mas ainda útil como linha de base.
- **Árvore de Decisão** é interpretável, mas com menor desempenho em multiclasse.