Relatório Técnico - Trabalho em Grupo

Amanda Barbosa Camilotti e Letícia Garrido Lorrenz

Junho 2025

Sumário

1 Introdução

No dinâmico mercado da música digital, compreender os fatores que levam uma canção ao sucesso é um desafio central para artistas, gravadoras e plataformas de streaming. Este trabalho propõe uma análise comparativa aprofundada entre diferentes modelos de regressão, combinando abordagens da estatística clássica e do aprendizado de máquina, para prever o nível de popularidade de músicas na plataforma Spotify.

Utilizando um vasto conjunto de dados, o objetivo é desenvolver e avaliar modelos capazes de prever a popularidade de uma faixa — uma variável contínua — com base em suas características técnicas e contextuais. Ao fazer isso, busca-se responder à seguinte questão: Qual abordagem preditiva oferece o melhor balanço entre precisão, interpretabilidade e eficiência computacional para este problema?

2 Descrição do Dataset

O estudo foi conduzido utilizando um dataset público disponível na plataforma Kaggle, intitulado "Top Spotify Songs". A base de dados é composta por aproximadamente 1.7 milhão de amostras e 25 features, o que atende plenamente aos critérios de volume e complexidade exigidos para esta análise. Para garantir a execução em tempo hábil, este estudo utilizou uma amostra de 100.000 registros.

A escolha deste dataset justifica-se por sua relevância cultural, grande dimensionalidade e pela rica combinação de variáveis preditoras, que incluem:

- **Características Sonoras (Áudio Features):** Atributos quantitativos extraídos diretamente do áudio das faixas, como *danceability*, *energy*, *acousticness*, *instrumentalness*, *valence* e *tempo*.
- **Variável-Alvo (Target):** *popularity*, uma métrica de 0 a 100 atribuída pelo próprio Spotify, que reflete o quão "quente"uma música está no ecossistema da plataforma. Esta será nossa variável dependente contínua.
- **Metadados Contextuais:** Informações como *country* (país do ranking), *album_release_date* (data de lançamento) e *is_explicit* (se a faixa contém conteúdo explícito).

3 Métricas de Avaliação

Para garantir uma comparação justa e multifacetada, a performance de cada modelo foi avaliada sob três perspectivas distintas, representadas pelas seguintes métricas:

- **RMSE (Root Mean Squared Error):** A raiz do erro quadrático médio foi escolhida como nossa métrica de erro principal. Ela mede a magnitude média dos erros de previsão na mesma unidade da variável-alvo (pontos de popularidade). O RMSE penaliza erros maiores de forma quadrática, o que o torna sensível a previsões muito distantes do valor real. Quanto menor o RMSE, mais preciso é o modelo.
- **R^2 (Coeficiente de Determinação):** Esta métrica (R^2) indica a proporção da variância da popularidade que é explicada pelo modelo. Um R^2 de 0.85, por exemplo, significa que 85% da variação na popularidade das músicas pode ser atribuída

às features utilizadas. Valores mais próximos de 1 indicam um modelo com maior poder explicativo.

- **MPIW (Mean Prediction Interval Width):** A largura média do intervalo de predição (com 95% de confiança) avalia a certeza do modelo. Um intervalo estreito (MPIW baixo) sugere que o modelo não só acerta a previsão, mas o faz com alta confiança. Um intervalo largo, por outro lado, indica grande incerteza. Quanto menor o MPIW, mais confiáveis e precisas são as estimativas.

4 Modelos Implementados

A seguir, está uma descrição do funcionamento, vantagens e desvantagens de cada modelo utilizado neste estudo.

4.1 Regressão Linear

A Regressão Linear busca encontrar a "melhor linha reta"que descreve a relação entre as features e a popularidade. O modelo calcula uma equação matemática onde os coeficientes representam o peso de cada feature, minimizando o erro entre as predições e os valores reais.

- **Vantagens:** Máxima interpretabilidade dos coeficientes. Por exemplo, um coeficiente positivo para *danceability* indica que, mantendo as outras variáveis constantes, um aumento na "dançabilidade"leva a um aumento na popularidade.
- **Desvantagens:** Assume que a relação entre as variáveis é linear, possui baixa performance em problemas complexos e é sensível a outliers.

4.2 Random Forest Regressor

Este é um modelo de *ensemble* que constrói uma "floresta"com centenas de árvores de decisão. A predição final é a média das predições de todas as árvores, o que torna o modelo mais robusto e preciso.

- **Vantagens:** Alta precisão, robustez contra overfitting e capacidade de ranquear a importância das features.
- **Desvantagens:** Menor interpretabilidade e maior custo computacional para treinar.

4.3 XGBoost Regressor

Baseado na técnica de *Gradient Boosting*, o XGBoost constrói modelos de forma sequencial, onde cada novo modelo é treinado para corrigir os erros do modelo anterior.

- **Vantagens:** Desempenho de ponta, alta velocidade e robustez contra overfitting através de regularização nativa.
- **Desvantagens:** Muitos hiperparâmetros para ajustar e é conceitualmente mais complexo.

4.4 SVR (Support Vector Regression)

O SVR busca ajustar uma linha que mantenha o máximo de pontos dentro de uma margem de erro definida ("tubo"), usando "kernels"para encontrar relações não-lineares.

- **Vantagens:** Eficaz com muitas features e capaz de capturar padrões não-lineares.
- **Desvantagens:** Extremamente lento em grandes datasets e sensível à escala das features.

4.5 Árvore de Decisão (Decision Tree Regressor)

Funciona como um fluxograma, fazendo perguntas sobre as features. A predição é a média da popularidade das músicas que terminam em um mesmo "nó folha".

- **Vantagens:** Altamente interpretável, rápido e não exige normalização de dados.
- **Desvantagens:** Propenso a overfitting e instável.

4.6 KNN (K-Nearest Neighbors Regressor)

Prevê a popularidade de uma música com base na média da popularidade das suas 'K' vizinhas mais próximas.

- **Vantagens:** Simples de entender e não possui um "treino"demorado.
- **Desvantagens:** Lento na predição, performance degrada com muitas features e é sensível à escala das variáveis.

4.7 MLP (Multi-layer Perceptron Regressor)

É uma Rede Neural Artificial que aprende padrões complexos através de camadas de "neurônios"interconectados, ajustando "pesos"para minimizar o erro.

- **Vantagens:** Altíssimo poder preditivo e capacidade de modelar relações complexas.
- **Desvantagens:** É um modelo "caixa-preta", computacionalmente caro e exige muitos dados e ajustes.

5 Análise Comparativa dos Resultados

Os modelos foram treinados com uma amostra de 100.000 registros para prever a popularidade das músicas. Os resultados de performance e custo computacional foram consolidados na Tabela 1.

Tabela 1: Resultados Comparativos dos Modelos de Regressão

Modelo	RMSE	R^2	MPIW (95%)	Tempo (s)
Regressão Linear	13.23	0.3989	51.46	0.5
Random Forest	11.31	0.5614	43.58	4.0
XGBoost	9.60	0.6836	36.36	0.3
SVR	9.64	0.6810	36.49	467.6
Árvore de Decisão	14.05	0.3230	54.64	0.5
KNN Regressor	5.74	0.8869	17.79	0.0
MLP Regressor	6.49	0.8554	22.02	50.6

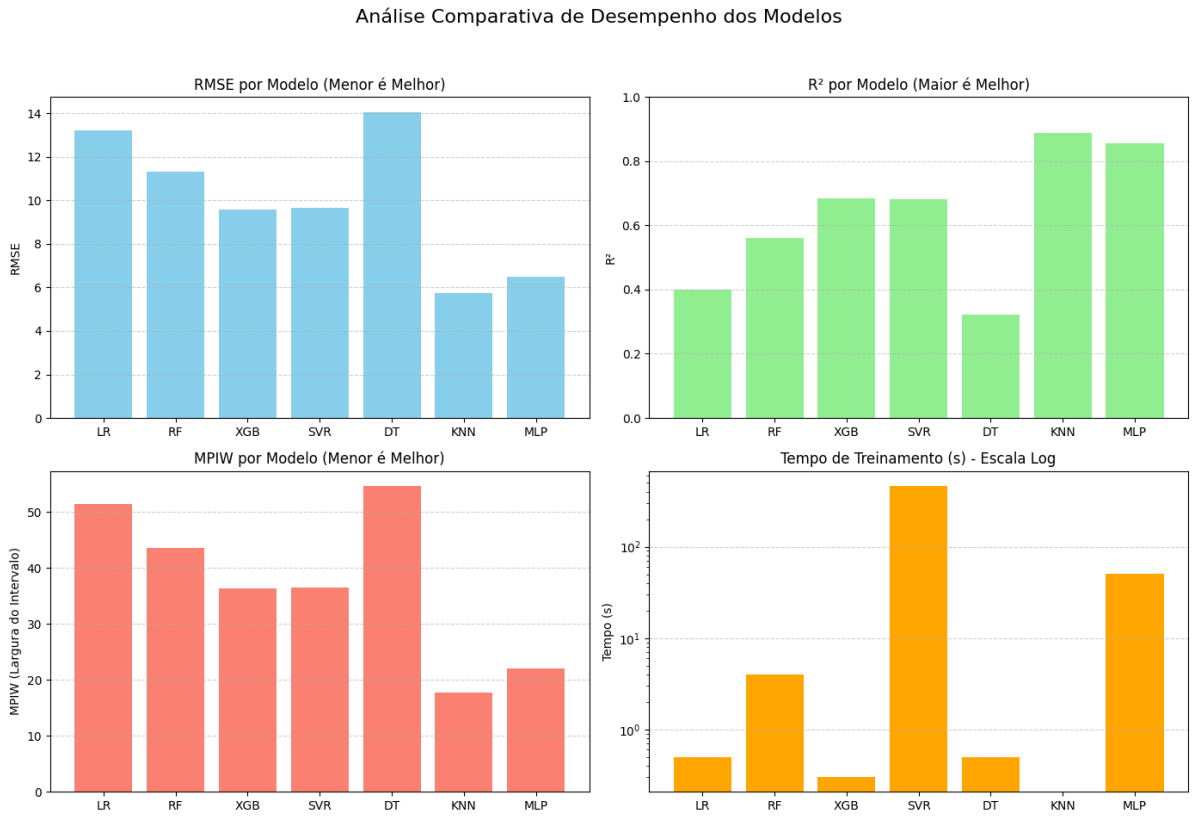

Figura 1: Gráficos Comparativos de Desempenho dos Modelos

Discussão dos Resultados

A análise dos resultados revela um vencedor surpreendente e claros trade-offs entre os modelos.

O Campeão Inesperado (KNN): Contrariando as expectativas, o KNN Regressor foi o modelo com o desempenho superior em todas as métricas de precisão: apresentou o menor RMSE (5.74), o maior R^2 (0.8869), indicando que explica quase 89% da variabilidade da popularidade, e o intervalo de confiança mais estreito (MPIW de 17.79). Seu tempo de "treino"foi nulo, pois o modelo apenas armazena os dados, embora o custo computacional real ocorra na predição.

Alto Desempenho (MLP): A Rede Neural (MLP Regressor) foi a segunda mais precisa, com um R^2 de 0.8554, mas com um tempo de treino considerável (50.6s). É im-

portante notar que o modelo emitiu um *ConvergenceWarning*, indicando que ele poderia ter alcançado uma performance ainda melhor se tivesse mais tempo para treinar.

O Mais Eficiente (XGBoost): O XGBoost Regressor se consagrou como o modelo mais equilibrado. Ele entregou uma performance muito boa (R^2 de 0.6836), quase idêntica à do SVR, mas em um tempo de treinamento de apenas 0.3 segundos, o que é mais de 1500 vezes mais rápido que o SVR.

Os Modelos de Árvore: A Árvore de Decisão única teve o pior desempenho geral, o que era esperado devido à sua propensão a overfitting. O Random Forest, seu análogo de ensemble, teve uma performance significativamente melhor, ilustrando o poder do método.

As Decepções: A Regressão Linear confirmou sua limitação em problemas não-lineares. O SVR, apesar de sua performance decente, teve um custo computacional muito alto, tornando-o a escolha menos prática de todas.

6 Conclusão

Com base nos resultados apresentados, o modelo que demonstrou o melhor desempenho preditivo geral foi o KNN Regressor, superando todos os outros em acurácia e confiança. No entanto, considerando um cenário prático onde tanto a velocidade de treinamento quanto a performance são cruciais, o XGBoost Regressor se destaca como a escolha mais pragmática e eficiente. A análise confirma que, para o problema de predição de popularidade musical, abordagens não-lineares são fundamentais, com os modelos baseados em instância (KNN) e em boosting (XGBoost) apresentando os resultados mais promissores. Uma limitação do estudo é a ausência de um ajuste fino de hiperparâmetros, que poderia, por exemplo, melhorar ainda mais a performance do MLP, como sugerido pelo seu alerta de convergência.