
Universidade Federal do Rio de Janeiro

Escola Politécnica

MBA em Big Data, Business Intelligence e Business Analytics (MB3B)

# O PRÓXIMO *HIT*:

## Análise e predição de sucessos do áudio *streaming* no Brasil

Autor:

Júlia Tavares Canedo Machado

Orientador:

Claudio Miceli de Farias D. Sc.

Examinador:

Cláudio Luiz Latta de Souza, M. Sc.

Examinador:

Nilton José Rizzo, D. Sc.

Examinador:

Norberto Ribeiro Bellas, M. Sc.

**Rio de Janeiro**
**Dezembro/2021**

# Declaração de Autoria e de Direitos

Eu, **Júlia Tavares Canedo Machado** CPF 141.420.927-40, autor da monografia ***O PRÓXIMO HIT*: Análise e predição de sucessos do áudio streaming no Brasil**, subscrevo para os devidos fins, as seguintes informações:

1. O autor declara que o trabalho apresentado na defesa da monografia do curso de Pós-Graduação, Especialização MBA em Big Data, Business Intelligence e Business Analytics da Escola Politécnica da UFRJ é de sua autoria, sendo original em forma e conteúdo.

1. Excetuam-se do item 1 eventuais transcrições de texto, figuras, tabelas, conceitos e ideias, que identifiquem claramente a fonte original, explicitando as autorizações obtidas dos respectivos proprietários, quando necessárias.

1. O autor permite que a UFRJ, por um prazo indeterminado, efetue em qualquer mídia de divulgação, a publicação do trabalho acadêmico em sua totalidade, ou em parte. Essa autorização não envolve ônus de qualquer natureza à UFRJ, ou aos seus representantes.

2. O autor declara, ainda, ter a capacidade jurídica para a prática do presente ato, assim como ter conhecimento do teor da presente Declaração, estando ciente das sanções e punições legais, no que tange a cópia parcial, ou total, de obra intelectual, o que se configura como violação do direito autoral previsto no Código Penal Brasileiro no art.184 e art.299, bem como na Lei 9.610.

2. O autor é o único responsável pelo conteúdo apresentado nos trabalhos acadêmicos publicados, não cabendo à UFRJ, aos seus representantes, ou ao(s) orientador(es), qualquer responsabilização/ indenização nesse sentido.

1. Por ser verdade, firmo a presente declaração.

Rio de Janeiro, 04 de dezembro de 2021.

_____

Júlia Tavares Canedo Machado

**UNIVERSIDADE FEDERAL DO RIO DE JANEIRO**

Av. Athos da Silveira, 149 - Centro de Tecnologia, Bloco H, sala - 212, Cidade Universitária Rio de Janeiro – RJ - CEP 21949-900.

## DEDICATÓRIA

Dedico esta pesquisa a todos que dedicam sua vida à arte e à cultura, sem as quais não sentimos nem experienciamos a nossa existência neste mundo.

*"A ciência descreve as coisas como são; a arte, como são sentidas, como se sente que são".*
(Fernando Pessoa)

## AGRADECIMENTO

Agradeço primeiramente aos meus pais, que são a base de tudo que eu sou e tudo que eu tenho; à minha vó Penha, que correu para que hoje eu possa caminhar; à minha irmã Laura, que compartilha comigo diariamente sua força, inspiração e apoio e, por fim, ao meu marido Lucas, que caminha ao meu lado ao longo de todos os desafios, conquistas e aventuras desta vida.

Agradeço aos professores e profissionais da educação de todo o Brasil que, apesar do negacionismo e todos os boicotes à ciência e à cultura, permanecem firmes na luta pela emancipação social promovida pela educação. Em especial aos professores da Universidade Federal do Rio de Janeiro, que me forma hoje pela terceira vez: a primeira como ser humano, a segunda como jornalista, e hoje como especialista.

# RESUMO

O objetivo deste trabalho é prever o potencial de sucesso de uma música baseada em dados sobre sua performance inicial e aspectos técnicos. É utilizado como base para esse estudo o *chart* Top 200 Brasil do *Spotify* de todos os dias do ano de 2020. Dessa forma, a análise pretende determinar as variáveis que mais contribuem para o sucesso de uma música que está entre as mais ouvidas do país e compreender o potencial de sucesso de um produto musical. O *chart* do *Spotify* é hoje uma das principais metas comerciais de artistas e suas gravadoras e distribuidoras, além de uma importante referência sobre o consumo de música no país e também um parâmetro para observação do surgimento de novos talentos. Nesse contexto, são utilizadas as variáveis disponibilizadas pela própria plataforma no sentido de estabelecer padrões de consumo, relações entre os aspectos das faixas e uma análise geral sobre a natureza do *chart* de 2020, e em seguida as faixas são categorizadas em grupos por critério de desempenho final. Por fim, são implementados os modelos preditivos de classificação escolhidos e testadas as suas acurácias a fim de atuarem como uma aplicação primária do uso de aprendizado de máquina no mercado fonográfico para a predição de resultados, que é o objetivo central desta pesquisa.

Palavras-Chave: Aprendizado de Máquina. *Charts*. Classificação. *Streaming*.

# ABSTRACT

This work intends to forecast the success potential of a song using data about its initial performance and technical aspects. The Top 200 Brazil chart from Spotify of 2020's is subject to studies. Thus, the analysis intends to determine the variables that most contribute to the success of a most-streamed in the country and comprehend the potential for success of a music product. The Spotify chart is currently one of the main commercial goals of artists and their record labels and distributors, in addition to being an important reference on the consumption of music in Brazil and a parameter for observing the emergence of new talents. In this context, the features provided by the platform are used in order to establish consumption patterns, relationships between aspects of the tracks and a general analysis of the nature of the 2020's chart, and then the bands are categorized into groups by final performance criterion. Finally, the chosen predictive classification models are implemented and their accuracy tested in order to act as a primary application of the use of machine learning in the phonographic market for the prediction of results, which is the main objective of this research.

Keywords: Machine Learning. Charts. Classification. Streaming.

## SIGLAS

| | |
|---|---|
| IFPI | International Federation of the Phonographic Industry |
| SVC | Support Vector Classifier |
| SVM | Support Vector Machines |
| MLP | Multilayer Perceptron |
| KNN | k-Nearest-Neighbors |
| TVP | Taxa de Verdadeiros Positivos |
| TFP | Taxa de Falsos Positivos |
| ROC | Receiver Operating Characteristic Curve |
| ROC AUC | Area Under Receiver Operating Characteristic Curve |
| LVQ | Learning Vector Quantization |
| RNA | Redes Neurais Artificiais |
| PMC | Percpetron Multicamadas |

## LISTA DE FIGURAS

Figura 2.1 Diagrama de Venn demonstrando as etapas de análise exploratória 31
Figura 2.2 Espaço ROC com três classificadores 36
Figura 2.3 Espaço de Máquina de Vetores de Suporte: hiperplano para classificação 41
Figura 2.4 Representação de uma Máquina de Vetor de Suporte Classificadora utilizando um kernel não linear 41
Figura 3.1 Etapas do projeto 46
Figura 4.1 Relação de resultados utilizando MLP Classifier 72

## LISTA DE TABELAS


| | | |
|---|---|---|
| Tabela 4.1 | Frequência de gênero | 60 |
| Tabela 4.2 | Relatório de classificação do modelo KNN | 67 |

## LISTA DE GRÁFICOS


| | | |
|---|---|---|
| Gráfico 3.1 | Quantidade de dados por coluna | 52 |
| Gráfico 4.1 | Histograma de dispersão da variável Total_Streams | 54 |
| Gráfico 4.2 | Dispersão da variável Maturação | 55 |
| Gráfico 4.3 | Contagem de valores positivos e negativos da variável Maturação | 55 |
| Gráfico 4.4 | Histograma da variável de frequência de chart | 56 |
| Gráfico 4.5 | Diagrama de dispersão | 57 |
| Gráfico 4.6 | Matriz de Correlação | 58 |
| Gráfico 4.7 | Histograma das variáveis de fan metrics | 59 |
| Gráfico 4.8 | Boxplots das variáveis consideradas na função de classificação | 64 |
| Gráfico 4.9 | Acurácia do modelo KNN | 66 |
| Gráfico 4.10 | Taxa de erro por número de vizinhos (KNN) | 67 |
| Gráfico 4.11 | Acurácia do modelo de Árvore de Decisão | 68 |
| Gráfico 4.12 | Matriz de Confusão de predição da Floresta Aleatória | 69 |
| Gráfico 4.13 | ROC Curve do modelo de Regressão Logística Binária | 73 |
| Gráfico 5.1 | Grau de importância das variáveis nos modelos preditivos | 77 |
| Gráfico 5.2 | Performance dos modelos preditivos | 77 |

## LISTA DE QUADROS


| | | |
|---|---|---|
| Quadro 2.1 | Resumo dos modelos de aprendizado de máquina | 43 |
| Quadro 3.1 | Descrição dos atributos da base de dados inicial | 47 |
| Quadro 3.2 | Relação inicial de variáveis | 48 |
| Quadro 3.2 | Relação de variáveis utilizada nos modelos | 62 |

# Sumário

**Capítulo 1: Introdução ..... 14**

1.1 – Objetivos Gerais e Específicos ..... 17
1.2 - Descrição ..... 17

**Capítulo 2: Referencial Teórico ..... 19**

2.1 – A Industrialização do Entretenimento ..... 19
2.2 – O Mercado Fonográfico e os desafios na era digital ..... 21
2.3 – Inteligência Competitiva e Big Data ..... 25
2.4 – Análises de Mercado para o Processo Decisório ..... 28
2.5 – Análise e Modelagem de Dados ..... 30
2.6 – Aprendizado de Máquina. ..... 32
2.6.1 – Avaliação dos modelos: overfitting e validação cruzada ..... 34
2.6.2 – Modelos de aprendizado de máquina ..... 37

**Capítulo 3: Metodologia ..... 45**

3.1 – Coleta de Dados. ..... 46
3.2 – Tratamento de Dados. ..... 48

**Capítulo 4: Análises e Resultados ..... 54**

4.1 – Análise Exploratória. ..... 54
4.2 – Correlação. ..... 58
4.3 – Análise por gênero. ..... 60
4.4 – Delimitação de dataset. ..... 62
4.5 – Aplicação de modelos preditivos. ..... 64
4.5.1 – KNN. ..... 66
4.5.2 – Árvore de Decisão. ..... 68
4.5.3 – Floresta Aleatória. ..... 69
4.5.4 – Naive Bayes. ..... 70
4.5.5 – Máquina de Vetores de Suporte - Classificador. ..... 71
4.5.6 – Perceptron Multicamadas ..... 71
4.5.7 – Classificador por regressão logística binária. ..... 73

**Capítulo 5: Conclusão e Trabalhos Futuros ..... 76**

**Referências Bibliográficas ..... 81**

# Capítulo 1

# Introdução

Ao longo das duas últimas duas décadas, é indiscutível que a tecnologia vem moderando de forma cada vez mais dinâmica a produção e o consumo. No mercado da música, essa inovação tecnológica junto às transformações sociais fundamentaram o surgimento de plataformas digitais de *streaming*. Essa forma de consumo oferece acesso sob demanda, com opção ilimitada via assinatura, permitindo que os usuários consumam um extenso catálogo com uma enorme variedade de músicas (Wikström, 2009, p. 104).

O *streaming* se tornou a melhor alternativa de geração de lucro para o mercado fonográfico nos últimos 10 anos, após significativa queda de receita com a democratização do acesso à internet – que ocasionou na distribuição ilegal de conteúdo. Atualmente, segundo o relatório do IFPI[1] de 2021 (com dados de 2020), os serviços de *streaming* geraram 13,4 bilhões de dólares, 62,1% da receita global do mercado fonográfico, e vêm conquistando a cada ano uma maior fatia dessa receita.

Entre as empresas que oferecem serviço de *streaming* de música, o *Spotify* é o que possui maior destaque atualmente. Fundada em 2006, a empresa detém hoje a maior fatia de público do mercado, tendo cerca de 381 milhões de usuários mensalmente ativos e 172 milhões de usuários *premium* (*Spotify Technology S.A. Announces Financial Results for Third Quarter 2021*)[2]. O lançamento do *Spotify* no Brasil aconteceu em 2014 (Moschetta e Vieira, 2018), e a plataforma divulga diariamente e semanalmente, de forma pública e gratuita, as faixas mais ouvidas na plataforma a nível Global e local (por país).

O chamado Top 200 de faixas mais ouvidas no *Spotify* se tornou um importante medidor de sucesso para gravadoras, distribuidoras e artistas, substituindo os rankings de

[1] Federação Internacional da Indústria Fonográfica (*International Federation of the Phonographic Industry*, IFPI). Formada no ano de 1933, representa hoje mais de 1450 empresas discográficas, em 75 países diferentes.

[2] Disponível em https://investors.*Spotify*.com/financials/press-release-details/2021/*Spotify*-Technology-S.A.-Announces- Financial-Results-for-Third-Quarter-2021/default.aspx Acesso em: 18/11/2020.

mais vendidos na era dos discos. “Rankear” nesse *chart* se tornou uma meta a ser atingida pelos artistas e, mais do que isso, uma escala de potencial de sucesso a ser observada por gestores de gravadoras para a contratação e desenvolvimento de novos talentos.

No contexto do mercado fonográfico, tem-se, portanto, a necessidade constante de produção de *hits* e desenvolvimento de artistas que figurem nos *charts* com certa frequência. Isso desencadeia no cenário de produção artística uma aceleração das mudanças dos estilos e tendências, além de uma imprevisibilidade de performance e distribuição desigual do sucesso. Nessa “economia dos extremos", a indústria musical acompanha esse processo e busca implementar a lógica de produção em massa para oferecer uma enorme variedade de produções a seus ouvintes e assim obter mais chance de êxito com seus produtos (Lipovetsky e Serroy, 2015, p. 101).

O estudo sobre a dinâmica de sucesso nos *chart*s do Reino Unido, realizado por Strobl e Tucker, é usado como base para este trabalho e comprova que a desigualdade de sucesso no mercado musical é antecedente à era digital de consumo. Com base em dados do *chart* do *New Musical Express* (NME)[3] de 1980 a 1993, o trabalho analisa e distribuição e incidência de sucesso entre os artistas e álbuns e conclui que:

> Nossa análise dos *charts* de álbuns do Reino Unido usando a listagem da NME no período de 1980-1993 fornece evidências de que a extensão do sucesso entre artistas de música popular não é distribuída igualmente. Especificamente, descobrimos que o grau de listagem do gráfico incidência, seja medida em termos de número de álbuns, média de semanas por álbum, ou semanas passadas no *chart*, é altamente enviesado, um fato que não é surpreendente em um mercado onde algumas “superestrelas” dominam e onde as gravadoras tendem a se concentrar em artistas já consagrados (STROBL; TUCKER, 2000, p.130)

A performance de produtos da indústria fonográfica e a dinâmica desse mercado criativo é uma discussão que vem ganhando impacto à medida que o meio digital lidera cada vez mais o consumo de música. Na perspectiva teórica, o livro Hitmakers - Como Nascem as Tendências, de Derek Thompson, trazem uma discussão com exemplos atuais sobre a imprevisibilidade de sucesso dos produtos de entretenimento. Hennig-Thurau e Houston em sua obra sobre *Entertainment Science* também abordam a discussão da aplicabilidade de algoritmos e técnicas relacionadas a *big data* aos produtos da indústria criativa, e citam que, por enquanto, a capacidade de efetivamente desenvolver produções

[3] Revista britânica de música publicada semanalmente desde 1952

artísticas ainda é realidade somente em histórias *cyberpunk,* que descrevem um cenário futurístico em que robôs são praticamente humanos.

De fato, no mercado criativo existe uma limitação maior quanto ao uso de métodos matemáticos e estatísticos relacionados à previsão e criação de novos produtos, isso porque os avanços da tecnologia e da inteligência artificial ainda não alcançaram a capacidade criativa, imagética e sensível inerente ao ser humano. A realidade subjetiva da percepção de obras relacionadas à arte faz do mercado criativo uma conjuntura de infinitas possibilidades de inovação que está em frequente atualização.

Ainda assim, tanto as ferramentas analíticas de *business intelligence* quanto o uso de linguagens de programação para finalidades estatísticas vêm se expandido no mercado, e a discussão sobre a aplicabilidade de aprendizado de máquina com fins preditivos começa a ser difundida na indústria criativa.

> A previsão de *hits* é útil para músicos, gravadoras e distribuidoras de música porque canções populares geram receitas maiores e permitem que os artistas compartilhem sua mensagem com um público amplo. Por exemplo, se uma gravadora deseja aumentar os lucros, ela pode optar por investir seus recursos limitados (campanhas publicitárias, equipamento de estúdio etc.) em faixas que provavelmente se tornarão populares. Por outro lado, se um artista deseja incorporar uma estética desprovida de características musicais convencionais, eles podem escolher lançar faixas que provavelmente não se tornarão populares (MIDDLEBROOK; SHEIK, 2019, p.1)

Em 2019, Middlebrook e Sheik utilizaram computadores de alta performance e um *dataset* de 1,8 bilhão de faixas, sendo 12 mil faixas que estiveram no *Billboard Hot 100*[4] *hits* para elaborar um modelo preditivo que alcançou 87% de acurácia. Seguindo uma semelhante linha de pesquisa, Pastore, Teixeira e Rezende, em 2020, elaboraram um trabalho sobre predição de volume de *streams* utilizando dados sobre músicas que estiveram no Top 200 do *Spotify* entre 2017 e 2020. A estimação foi feita por meio de regressão linear múltipla, mas obteve um baixo coeficiente de determinação.

Os dois trabalhos trazem importantes informações teóricas e práticas sobre como o cenário musical e o mercado criativo podem ser impulsionados com o avanço da utilização de inteligência de dados. Esta pesquisa busca, portanto, ampliar esse debate ao acrescentar novos estudos sobre a indústria fonográfica utilizando análises e modelos preditivos

[4] Ranking semanal publicado pela revista americana *Billboard* que classifica as melhores 100 faixas baseado no volume de vendas, execuções em rádio e volume de *streams*

capazes de evidenciar as tendências de mercado no Brasil, além de contribuir ativamente com a pesquisa acadêmica sobre inteligência artificial, servindo de base para futuros posteriores sobre a emergente utilização dessas ferramentas no setor cultural.

Neste trabalho serão utilizados os *charts* das 200 faixas mais executadas por dia no ano 2020 – publicamente divulgados pelo *Spotify* – para a análise das métricas da própria plataforma e a geração de novas variáveis que possibilitem o estudo do comportamento do *chart* ao longo do ano, da natureza das músicas que compuseram este ranking e busca de padrões em tendências. Em seguida, há a divisão das faixas em grupo por desempenho ao final e o teste de modelos de aprendizado de máquina para a previsão a fim de responder: em qual grupo essas músicas estarão ao final do ano considerando suas características e dados iniciais de performance?

## 1.1 – Objetivos Gerais e Específicos

O objetivo geral deste trabalho foca em obter a previsão sobre o potencial de sucesso de uma faixa com base nas características (variáveis) técnicas e de performance inicial da música. O universo de dados utilizado é o Top 200 diário do *Spotify* no Brasil.

Para esta finalidade, os objetivos específicos que tangem a abordagem central desta pesquisa são:

1. Buscar conclusões sobre o comportamento das músicas que estiveram no *chart* para encontrar padrões e tendências;
2. Encontrar correlações entre variáveis técnicas de produção e os resultados de performance alcançados pelas músicas;
3. Aplicar modelos preditivos utilizando aprendizado de máquina e testar e comparar suas acurácias;

## 1.2 – Descrição

Os conceitos abordados neste trabalho e seus estudos estão estruturados da seguinte forma: o capítulo 2 aborda o referencial teórico de fundamentação para a elaboração deste trabalho, apresentando o desenvolvimento social e tecnológico que

possibilitou a construção do cenário atual em que se fazem necessárias as aplicações finais do trabalho.

O capítulo 3 se reserva a explicar a metodologia desenvolvida, os passos e procedimentos que compõem a estruturação prática de aplicação do aprendizado de máquina para o problema principal relacionado – predizer se uma faixa tem potencial de sucesso ou não baseada em sua performance inicial e seus aspectos técnicos.

No capítulo 4 estão registradas as análises e resultados, tanto da análise exploratória – que permite uma visão detalhada sobre o comportamento das músicas e a natureza do *chart* – quanto dos resultados de previsão obtidos pelos modelos preditivos selecionados.

Por fim, o capítulo 5 consolida as conclusões obtidas e disserta sobre os resultados, sugerindo mudanças e novas aplicações para os trabalhos futuros.

# Capítulo 2

# Referencial Teórico

Para entender melhor o contexto em que se fazem necessárias as aplicações discutidas neste trabalho, é necessária uma abordagem teórica capaz de explicar os processos sociais e tecnológicos que fundamentaram o estudo e análises elaboradas. Neste capítulo busco explicar como se deu o processo de transformação do consumo e produção de música e quais são os principais pontos de desenvolvimento do mercado fonográfico no Brasil e no mundo.

Além disso, são abordadas também ao longo deste capítulo as perspectivas teóricas dos procedimentos de tratamento, análise e preparação dos dados para aplicações de modelos de *machine learning* que são objeto de pesquisa neste trabalho, buscando acrescentar na pesquisa sobre o uso de tecnologias de *big data* no mercado criativo.

## 2.1 – A Industrialização do Entretenimento

A industrialização é um processo econômico da produção em massa de bens de consumo. No entanto, entre o início da Revolução Industrial no fim do século XVII e os dias atuais, o termo "indústria" passou a ser utilizado para se referenciar a processos não somente de produção, mas de estetização, propaganda e distribuição de bens e serviços, incluindo os bens não-materiais, como a arte e a cultura (Wikström, 2009, p. 46).

Theodor W. Adorno e Max Horkheimer em 1947 cunharam a expressão "Indústria Cultural" para diferenciar a arte que surgia de forma espontânea no meio popular ("cultura das massas") daquela que surgia sistematicamente através da dinâmica capitalista de produção para reforçar as conjunturas sociais vigentes. A indústria cultural é, segundo Costa (2013, p. 136)," é fruto da oportunidade de expansão da lógica do capitalismo sobre a cultura".

A indústria do entretenimento é um fenômeno da indústria cultural, ao passo em que busca o lucro através da produção massiva de bens culturais e artísticos. Hennig-

Thurau e Houston (2019, p. 41) definem o entretenimento como qualquer produto oferecido ao consumidor cujo principal objetivo é proporcionar prazer e divertimento, à frente de oferecer alguma funcionalidade prática útil; e pode ser oferecido na forma de diversos conteúdos - filmados, escritos, gravados, programados etc.

Enquanto manifestação do capitalismo, a indústria cultural se adapta às rápidas mudanças culturais, políticas e tecnológicas do sistema ao mesmo tempo em que consegue ditar novas tendências. No entanto, diferentemente dos bens de consumo que visam a conveniência de utilização e o aumento da produtividade, nos produtos do entretenimento prevalecem as dimensões artísticas, criativas e imaginárias que têm como principal objetivo atingir o consumidor por meio dos sentidos e sentimentos. Assim, a indústria do entretenimento participa do desenvolvimento do que Lipovetsky e Serroy chamam de capitalismo artista, o conceito que define a lógica mercantilista da arte, da estética, das imagens e do sensível.

> Como triunfo do regime artista ou criativo, o capitalismo não se torna "menos" capitalista: muito pelo contrário, ele o é cada vez mais e numa escala vastíssima, como atestam a magnitude crescente dos investimentos financeiros, a mundialização dos mercados do consumo, da moda e do luxo, o desenvolvimento das multinacionais da cultura, a predominância do marketing e da comunicação, os lucros consideráveis que são gerados (LIPOVETSKY e SERROY, 2015, p.20).

Entre os pilares da indústria do entretenimento estão os livros, o cinema e a música. A música e o cinema são as artes que mais foram impactadas pelo desenvolvimento de novas tecnologias, reestruturações financeiras e configurações de distribuição. Também as que melhor triunfaram com a reinvenção nas formas de atingir o público, uma vez que são produtos mais acessíveis às massas - em contrapartida à pintura, da escultura e das artes plásticas.

O cinema e a música, por mais que produzidos industrialmente, detém a singularidade presente em cada roteiro, som e imagem. Além disso, cada obra do cinema e da música combina o padrão e a originalidade na construção de um produto de entretenimento único. Tanto o cinema quanto a música triunfaram na era do capitalismo artista não apesar, mas em razão do avanço tecnológico que possibilitou: a produção em nível industrial e ainda assim singular; e um consumo coletivo, mas ao mesmo tempo a nível experiencial de cada consumidor.

> Muito comparável com o desenvolvimento do cinema e aliás ligada a ele por laços ao mesmo tempo industriais e artísticos, a música gravada, outra forma de arte industrial, muda radicalmente a situação do mundo musical. Até então limitada ao instante da sua interpretação, a obra, graças à gravação, se vê subitamente fixada num suporte que possibilita sua escuta contínua, repetitiva, praticamente sem fim. Ela se abre, com isso, a um público imensamente mais vasto do que as pessoas presentes ao concerto, privado ou público, que eram seus únicos ouvintes. (LIPOVETSKY e SERROY, 2015, p.116)

## 2.2 – O Mercado Fonográfico e os desafios na era digital

No século XVIII, a indústria do entretenimento voltada à música se resumia a organizar, imprimir e vender partituras de música em catálogos e livros. A única forma de ouvir música fora dos concertos era tocando você mesmo a música em um instrumento utilizando esses livros de partitura. A indústria da música efetivamente nasceu a partir da evolução do "*phonograph*" (fonograma) de Thomas Edison - que foi a primeira forma de se gravar música - até a difusão do gramofone, inventado por Emile Berliner, que trouxe a difusão dos discos de vinil, comercializados até hoje. Fáceis de serem produzidos e armazenados, a comercialização dos vinis deu início ao mercado fonográfico como conhecemos hoje (Smith e Telang, 2016, p. 18).

No final dos anos 1990 os CDs tomaram o espaço dos vinis e se tornaram um formato de mídia musical extremamente lucrativa:

> No final de 1995, a I*nternational Federation of the Phonographic Industry* relatou que "as vendas anuais de música pré-gravada atingiram um recorde histórico, com vendas de cerca de 3,8 bilhões de unidades, avaliadas em quase US $ 40 bilhões". "As vendas de unidades estão atualmente 80% mais altas do que há uma década", continuou o relatório, "e o valor real do mercado mundial de música mais que dobrou no mesmo período (SMITH; TELANG, 2016, p. 24).

Ao longo das décadas, gravadoras e distribuidoras de música surgiram e foram englobadas por empresas maiores. Se adaptaram a diversos formatos de armazenar e distribuir música, de desenvolver e divulgar artistas e suas obras. Atualmente, a maior fatia do mercado musical se concentra em três *major labels*: *Universal Music Group*, *Warner Music Group* e *Sony Music Entertainment*.

O fenômeno da digitalização da cultura que teve início no século XX ocasionou uma série de rápidas mudanças no mercado musical. O declínio dos CDs e o avanço da

arquitetura de redes *peer-to-peer* entre o fim do século XX e o início do século XXI mudaram completamente a dinâmica do mercado fonográfico. A possibilidade de se distribuir arquivos de música digitais online e ignorar os direitos de produção e distribuição ameaçou o controle das gravadoras e seus artistas.

A pirataria abalou a geração de receita musical. Segundo Wikström (2009, p. 64), esse foi um dos principais motivos para o início da queda na arrecadação do mercado fonográfico no final dos anos 1990. Por outro lado, a internet promoveu certa democratização do acesso aos meios de produção e distribuição de música, favorecendo os chamados artistas independentes.

Com a massiva oferta de músicas em inúmeros sites na internet, a pirataria enfraqueceu a indústria fonográfica e descentralizou o poder da produção musical. No entanto, no início dos anos 2010, tanto a evolução das tecnologias quanto uma mudança expressiva no comportamento do consumidor favoreceram a última grande revolução na forma de consumo de entretenimento: as plataformas de *streaming*.

A expansão da banda larga e o acesso à internet de qualidade e velocidade cada vez mais apropriadas, combinadas à oferta de conteúdo certa forma desorganizada abastecida pela pirataria e produções independentes, causaram aos consumidores "a demanda e a conveniência de acessar diferentes produtos em um só lugar" (SMITH; TELANG, 2016, p. 76). Assim, a indústria fonográfica, antes surpreendida pela descentralização da produção musical, conseguiu com as plataformas de *streaming* desenvolver estratégias para se beneficiar da distribuição na era digital.

> Em apenas dez anos, a participação da receita global gerada por produtos musicais imateriais cresceu de 11% para 59% em 2016, com uma tendência de seguir crescendo ainda mais (HENNIG-THURAU; HOUSTON, 2019, p. 172).

Segundo o *Global Music Report* do IFPI de 2020, a receita anual produzida pela indústria fonográfica chegou a 21.6 bilhões de dólares. Além de produzir valor econômico substancial, a indústria do entretenimento também é a pioneira em estratégias que hoje são utilizadas em negócios de vários ramos, como criatividade, inovação, *storytelling*, o trabalho realizado no pré-lançamento de produtos e, acima de tudo, a adaptação às disrupturas da era digital.

> Como o conteúdo dos produtos de entretenimento consiste essencialmente em informação, a indústria do entretenimento, embora certamente nem sempre intencionalmente, se adaptou a um papel pioneiro ao lidar com os desafios da digitalização. [...] A digitalização agora afeta todas as indústrias de uma forma fundamental, desafios semelhantes existem, ou estão prestes a chegar, para gerentes em outras áreas além do entretenimento. (HENNIG-THURAU; HOUSTON, 2019, p. 48)

O desenvolvimento tecnológico de *streaming* possibilitou a expansão da oferta de um acervo de músicas na casa das dezenas de milhões. O mesmo relatório aponta que o *streaming* já representa 62,1% da receita do mercado global de música. Embora a internet tenha multiplicado de forma abrupta a quantidade de produtos musicais oferecidos, a atenção do público e a concentração do lucro permanece em um pequeno número de referências do mercado fonográfico.

> O capitalismo artista é, portanto, o sistema no qual se observa uma distribuição extremamente desigual do sucesso, uma espiral dos desempenhos extremos. (...) Nesse sistema, um produto que tem êxito absorve as perdas da maioria: é uma lógica de cassino que estrutura a economia das indústrias culturais. (LIPOVETSKY & SERROY, 2015, p. 102)

Dada a imprevisibilidade do mercado criativo devido à linha tênue entre sucesso e fracasso, é preciso se submeter à conjuntura comercial em que para grandes sucessos existe um cemitério de ideias que fracassaram. Projetos irrelevantes, frustrados e até os com grande potencial, mas que se desenvolveram em uma época desfavorável são situações comuns nas indústrias fundamentadas na criação e na criatividade humana.

Segundo Thompson (2018), as histórias de fracasso na indústria criativa e do entretenimento se repetem em diversos negócios devido à inconstância do ramo, e o êxito de uma pequena parte dessas ideias

> Acima de tudo, requer um modelo de negócios que suporte a inevitabilidade de que a maior parte das coisas fracassam; as ideias mais promissoras com frequência atraem um coro de céticos; e um grande *hit* pode pagar por mil fracassos (THOMPSON, 2018, p. 222)

Portanto, fica evidente que a oferta massiva de música em diversas plataformas digitais ao mesmo tempo alavancou a competitividade do mercado de entretenimento não somente a nível de inovação criativa, mas também tecnológica.

Além da revolução na forma de ser levado a consumir música, o usuário também é diretamente influenciado no processo de descoberta de novas músicas. Antes da internet, todo o processo de conhecer novos artistas, produções e acompanhar os novos lançamentos e tendências demandava a dedicação e o custo de frequentar lojas de discos e apresentações ao vivo. Atualmente as plataformas de *streaming* atendem a essa demanda ao proporcionar o consumo ilimitado de música por um custo mensal razoável e oferecendo curadorias instrumentalizadas por especialistas e inteligência artificial.

> O *Spotify* utiliza uma combinação de curadoria humana e algorítmica para sugerir e apresentar músicas compatíveis com os gostos e preferências musicais passadas dos utilizadores, funcionando como um fio condutor da experiência de consumo no presente (MOSCHETTA; VIEIRA, 2018, p.265)

O aperfeiçoamento dos algoritmos de recomendação e a possibilidade coletar dados que podem ser traduzidos em informações valiosas de mercado modificaram completamente a dinâmica da distribuição musical no meio digital. Nesse contexto, o uso dessas técnicas de coleta, tratamento e interpretação de *big data* e sofisticados sistemas de recomendação por parte das grandes instituições que oferecem o serviço de *streaming* de música (*Spotify, Deezer, Apple Music* e os mais recentes *Amazon Music e Youtube Music*) passaram a impactar ainda mais na oferta de música aos usuários, sendo, portanto "cruciais para as formas como a música a distribuição desenvolveu-se ao longo do tempo na cultura de *streaming*" (MAASØ; HAGEN, 2020, p. 19).

Atualmente as plataformas digitais de música conseguem mapear o comportamento dos usuários de forma detalhada. Cada *stream* (vez que a música é tocada por mais de 30 segundos) é computado, sendo possível identificar quando e em que dispositivo a faixa foi ouvida, em que lugar do mundo, o número de vezes em que a faixa foi pulada, repetida, além do número de *playlists* em que a faixa está, e o mais importante: entender como o usuário chegou àquela música - busca na plataforma, direcionada por posts em redes sociais, divulgação online em blogs e sites, *playlists* editoriais ou de outros usuários, etc.

Anteriormente à era digital, os mediadores culturais eram majoritariamente perceptíveis, como em forma de propaganda, recomendações de especialistas, disposição das mídias em lojas físicas e tempo de execução na TV e no rádio. Os avanços tecnológicos possibilitaram fazer dos dados e das ciências matemáticas e estatísticas ferramentas "invisíveis" que constroem os guias do usuário à música, e o sistemas de

recomendação "se tornaram os novos intermediários culturais, absorvidos pelos maiores grupos de multimídia e gigantes da tecnologia" (SANTINI; SALLES, 2020, p. 85).

Os modernos processos de análise de dados de consumo transformam a experiência dos usuários, e consequentemente afetam a dinâmica de produção e distribuição de música. O custo reduzido e a facilidade da disponibilização online permitiram que as gravadoras e distribuidoras, com base nos dados e nas interpretações destes, estudassem e testassem novas estratégias de lançamento para cada um dos seus projetos e artistas.

Visto que o *Spotify* é a companhia com maior *market share* (32%) entre as plataformas de *streaming* segundo relatório do MIDiA Research de 2021[5], uma das principais metas na indústria musical é ter o artista e suas produções como um dos mais consumidos pelo público da plataforma. Os *charts* do *Spotify* inauguraram a era digital do que, no século XX, foram os rankings de discos mais vendidos da *Billboard*: o indicador de sucesso de uma música/artista e a síntese das principais tendências de consumo a nível global e local. Além disso, a participação no *chart* também funciona como uma ferramenta de divulgação que tem potencial para alavancar a performance não apenas do projeto que se destacou, mas também dos trabalhos futuros do artista (Strobl e Tucker, 2000, p.113).

O estudo de dados de consumo digital é, portanto, uma estratégia de negócio para se aproximar cada vez mais do entendimento pleno do comportamento da audiência. Atualmente, a nova economia da música é impactada pelo grande número de informações disponíveis e uma maior conexão dos consumidores entre eles, o que pautou a transição da relação pré-digital – de controle da informação por parte das grandes empresas – para a relação pós-digital, em que os ouvintes participam ativamente do mercado de música, principalmente por meio de compartilhamentos nas redes sociais, trocando opiniões, críticas e influência.

## 2.3 – Inteligência Competitiva e Big Data

Identificar os mecanismos que guiam os ouvintes até a música e entender as influências de cada um deles na performance de música no meio digital é hoje o maior desafio das gravadoras e distribuidoras. Para traçar esse caminho, a inteligência

[5] Disponível em https://www.midiaresearch.com/blog/global-music-subscriber-market-shares-q1-2021. Acesso em 15/11/2021.

competitiva é um importante diferencial que permite o estudo do posicionamento da empresa no mercado, a identificação de oportunidades de negócio, o monitoramento de ações e resultados da concorrência e a estruturação de estratégias a curto, médio e longo prazo.

> A inteligência competitiva trata de um processo alternativo, com metodologia específica, para o desenvolvimento de práticas que tem como base a informação que circula no ambiente de negócios (GOMES; BRAGA, 2017, p. 24)

Na constituição da inteligência competitiva estão os entendimentos dos conceitos de dados, informação e conhecimento. Davenport e Prusak (2003, p. 4) definem que dados são "um conjunto de dados distintos e objetivos", ou seja, são registros do estado ou ação de algo ou alguém, que podem ser obtidos, estruturados e transferidos automaticamente, sem interferência humana, e são os insumos para a estruturação da informação.

A informação é o produto final da significação dos dados, ou seja, é constituída a partir da modelagem dos dados para uma determinada finalidade. Entre os processos de construção da informação estão a definição do objetivo ao qual a informação pode servir, cálculos e análises matemáticas e estatísticas dos dados, a eliminação de erros e a condensação dos dados para uma forma mais concisa (Davenport e Prusak, 2003, p. 5).

A concepção das percepções, ideias e conclusões a partir das informações compõem o que chamamos de conhecimento. O conhecimento é um recurso essencialmente humano, edificado a partir de vivências, experiências, observações e métodos de compreensão de informações.

> Uma das razões pelas quais achamos o conhecimento valioso é que ele está próximo - mais do que os dados e as informações - da ação. O conhecimento pode e deve ser avaliado pelas decisões ou tomadas de ação às quais ele leva. Um conhecimento melhor pode levar, por exemplo, a eficiência mensurável em desenvolvimento de produtos e na sua produção. Podemos usá-lo para tomar decisões mais acertadas com relação a estratégia, concorrentes, clientes, canais de distribuição e ciclos de vida de produto e serviço (DAVENPORT; PRUSAK 2003, p. 6)

A inteligência competitiva é fundamentada na gestão da informação e do conhecimento. Enquanto a gestão da informação se refere à administração de fluxos formais de informação - referentes às informações que circulam entre as áreas da própria empresa -, a gestão do conhecimento atua no sentido de gerenciar a ciência produzida a

partir de ideias, informações e conhecimento gerado internamente pelos recursos humanos nas organizações.

Em um contexto em que é possível as organizações terem o mesmo nível de acesso à informação sobre o mercado que a concorrência, a inteligência competitiva adquire cada vez mais relevância ao passo em que idealiza novas forma de gerenciar e viabilizar essas

informações para a obtenção de insights e o apoio à análise estratégica e ao processo decisório. Nesse sentido, o papel da inteligência competitiva é modelar um cenário de informações direcionadas ao negócio e munir as equipes de insumos suficientes para análises que sejam conclusivas e respondam assertivamente às questões de negócios.

A era do *big data* transforma ativamente a inteligência competitiva nas organizações devido ao grande volume, velocidade e variedade de dados gerados a partir do desenvolvimento tecnológico e amplo acesso à internet. Nessa perspectiva, segundo Braga e Gomes (2017, p. 66), é necessário o aprofundamento dos processos de inteligência competitiva para manipular grandes conjuntos de dados e gerar análises mais complexas.

Como visto anteriormente, para a indústria da música os dados mais relevantes atualmente são os gerados pelas plataformas de *streaming*, que concentram a maior parte da receita do mercado. De acordo com Smith e Telang (2016, p. 3), para os negócios criativos estes "são os melhores e piores tempos" porque, ao mesmo tempo em que o desenvolvimento tecnológico viabilizou poderosas formas de produzir novos trabalhos e compreender a audiência, o novo cenário competitivo "forçou os líderes a fazerem difíceis compensações entre os velhos modelos e as novas oportunidades de negócio".

Portanto, a competitividade alavancada pela tecnologia vem tornando essencial o investimento em diferenciais de inteligência de mercado no mercado da música. Entre as principais possibilidades das análises de inteligência competitiva na indústria musical e criativa estão: elencar as prioridades que diminuam o risco histórico de fracasso da maior parte dos produtos musicais, posicionar os próximos artistas e lançamentos no mercado musical, prever sucessos e antecipar investimentos.

Segundo Braga e Gomes (2017, p. 66), as opções analíticas aplicáveis em processo de inteligência competitiva são divididas em três tipos:

- Análise descritiva: agregação e mineração de dados para o entendimento de uma

visão do passado, permitindo o estudo do que aconteceu anteriormente;

- Análise preditiva: que utiliza métodos de previsão e análises estatísticas para antever possíveis cenários e questões de negócio;

- Análise prescritiva: que utiliza algoritmos e aprendizado de máquina para prever resultados sobre possíveis decisões a serem tomadas.

Essas análises apoiam o processo ao identificar tendências e gerar insights sobre o mercado e a concorrência, fornecendo "uma indicação direta ou indireta de suas intenções, motivos ou metas" e tornando-se "um instrumento estratégico de apoio à gestão que alterará a maneira como as empresas lidam com seu mundo (BRAGA e GOMES, 2017, p. 51).

## 2.4 – Análises de mercado para o processo decisório

Segundo Hennig-Houston e Thurau (2019, p. 2), a indústria do entretenimento há décadas adota a abordagem definida por William Goldman como "*Nobody Knows Anything*" (Ninguém Sabe de Nada, tradução livre). Como tratado no início do capítulo, de fato a imprevisibilidade de sucesso é um desafio enfrentado pela indústria fonográfica devido ao caráter sensível humano dos produtos culturais. No entanto,

> em um competitivo cenário digital, com uma grande quantidade de informação disponível, os líderes não podem mais justificar importantes tomadas de decisão baseadas somente em seus instintos pessoais - fazer isso seria restritivo e os e os tornaria vítimas da armadilha "Ninguém Sabe de Nada" (HENNIG-HOUSTON; THUREAU, 2019, p. 2).

Apesar disso, também é possível que a modelagem de dados seja feita até que os líderes consigam observar fortes relações entre os dados e o produto estudado, tornando a análise "na melhor das hipóteses, impressionante, mas idiossincrática, de valor para curto prazo e, no pior dos casos, enganosas e até mesmo contraproducentes" (HENNIG-HOUSTON; THUREAU, 2019, p. 2). Estabelecer uma relação entre o conhecimento empírico, experiencial e teórico com o uso de análises de dados é o que os autores definem como *Entertainment Science.*

Algumas das possíveis análises de mercado que podem ser conduzidas nas grandes

companhias da indústria fonográfica utilizando fundamentos de *Entertainment Science* são no sentido de:

- Acompanhar os lançamentos e estratégias de empresas concorrentes;
- Monitorar a performance dos produtos de seus artistas para antecipar as estratégias dos próximos lançamentos;
- Avaliar o potencial de desenvolvimento de novos artistas ao longo do tempo;
- Desenvolver novos produtos com base no valor gerado historicamente à produtos semelhantes;
- Aperfeiçoar técnicas criativas e operacionais no desenvolvimento de produtos e artistas;
- Identificar potenciais oportunidades de campanhas, parcerias e destaques nas plataformas de *streaming*.

Para cada uma das possíveis análises é necessária uma metodologia que aborde os dados necessários e se adeque ao modelo (descritivo, prescritivo ou preditivo) de geração de resultados. Como tratado durante o capítulo, apesar do valor substancial dos dados e informações, estes isoladamente não agregam potencial a decisões. "A tomada de decisão com base em informação proporciona maior grau de gestão de risco, mas ainda não é tão segura quanto aquelas decisões com base em conhecimentos acionáveis" (GOMES; BRAGA, 2017p. 47).

Essa perspectiva considera que, apesar de a tecnologia da informação ser cada vez mais eficiente em recolher, tratar e consolidar elementos capazes de constituir uma ideia, os processos de construção e uso do conhecimento raramente são promovidos pela tecnologia (DAVENPORT; PRUSAK 2003, p. 68). Isso porque se atentar a um grande volume de informações e fontes sem potencial de oferecer significado real à análise de negócio não necessariamente agrega valor ao processo de tomada de decisão, podendo acarretar um desperdício de recursos.

> A complexidade do cenário de negócios atual, combinada com expectativas crescentes de desempenho e a velocidade com a qual tomadas de decisões devem ser feitas, são uma receita potencial para o desastre do tomador de decisão nos dias atuais, a menos que uma

> metodologia definida para a tomada de decisão seja estabelecida (GOMES; BRAGA, 2017, p. 49).

Sendo assim, é possível afirmar que a construção de boas análises de mercado deve se sustentar em estratégias bem definidas, utilizando bases de dados confiáveis, buscando objetivos específicos, e estabelecendo uma gestão de conhecimento na equipe de inteligência que possibilite a determinação de metodologias capazes de responder às questões de negócio. Dessa forma, os tomadores de decisão estarão mais preparados ao usufruir de um processo estruturado, baseado em fatos e insights que definem o mercado com mais precisão e fundamentam decisões mais assertivas.

## 2.5 – Análise e modelagem de dados

Entre definir as questões de negócio a serem resolvidas e obter conhecimento útil capaz de responder a essas questões com base em dados estão os processos de preparação, análise e modelagem, que de fato transformam os dados em importantes ativos estratégicos.

A preparação dos dados é geralmente a etapa mais demorada, já que requer a coleta de dados de diversas fontes e conformação de formatos e estruturas para que a implementação das análises seja feita de forma adequada. Myatt (2007, p. 2) destaca que qualquer tarefa em um projeto de *data science* direcionado à tomada de decisão está em uma das três categorias:

- Sintetizar os dados: quando os dados são resumidos a uma determinada interpretação sem comprometer informações importantes, ou seja, extrair de uma grande quantidade de dados as principais informações com o objetivo de alcançar uma visão ampla do que aqueles dados podem informar;

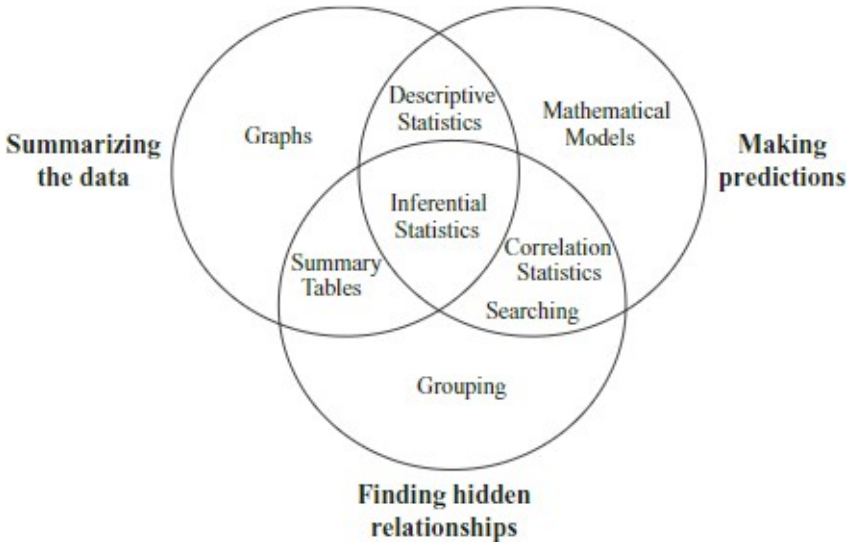
- Encontrar relações: quando são identificadas importantes associações, anomalias, correspondências e tendências ao cruzar os dados entre eles, adquirindo informações que não são tangíveis no processo de sintetização;

- Fazer predições: o processo de antever resultados ou situações com base em eventos já ocorridos, estimando cenários futuros;

A análise exploratória e a mineração de dados são os conjuntos de técnicas estatísticas e matemáticas aplicáveis a dados para realizar essas tarefas que visam sintetizar e relacionar dados e fazer previsões. Esses processos estão internamente ligados, uma vez que uma análise de dados muitas vezes pode combinar os métodos utilizados nessas três grandes categorias de

tarefas para oferecer múltiplas visões do cenário que os dados podem refletir. Esses métodos devem ser definidos considerando o problema de negócio a ser resolvido.

**Figura 2.1 -** Diagrama de Venn demonstrando as etapas de análise exploratória




Fonte: Myatt (2007, p. 3)

A estatística descritiva é a área da estatística responsável por aplicar técnicas que explicam as variáveis de diversas formas, permitindo observar e descrever quantitativamente e com precisão os valores que compõem os atributos envolvidos em um problema de negócios. Esse conjunto de técnicas é fundamental a qualquer trabalho de *data science*, uma vez que possibilita a primeira visão da distribuição e comportamento dos dados.

A partir da análise descritiva é que se torna também mais amplo e preciso o conhecimento sobre a correlação entre as variáveis. Entre os principais resultados da

análise exploratória e descritiva estão: a identificação de anomalias (como, por exemplo, outliers), a possível ocorrência de multicolinearidade entre os dados e a detecção de relações que, à primeira vista, não parecem óbvias.

As plataformas de *streaming*, enquanto principal forma de consumo de música atual, detém um controle do fluxo de informações entre o público e os artistas que reflete a realidade do mercado fonográfico de forma cada vez mais precisa e detalhada, fazendo dessa dinâmica

um elemento "crucial para as maneiras pelas quais a distribuição de música se desenvolveu ao longo do tempo na cultura de *streaming*". Além disso, a manipulação e análise adequada dessas informações pode gerar métricas que têm um papel central na elaboração de planejamento estratégico em atividades dentro e fora dos serviços de *streaming* (Maasø e Hagen, 2020, p. 19).

Um dos mais modernos métodos de gerenciamento de informações e projeção de resultados é a predição. A análise preditiva é a aplicação de técnicas para a elaboração de modelos de predição, usados em diversos problemas de negócio para prever eventos futuros. Os modelos preditivos englobam a análise exploratória, a mineração de dados e o uso de algoritmos de aprendizado de máquina, "podendo assumir várias formas e tamanhos, a depender da complexidade da aplicação para o qual eles são elaborados" (NYCE, 2007, p.9).

Nesse contexto, a ascensão do *big data* permite que as companhias tenham acesso a maiores e diversas amostragens, com dados mais detalhados e que possibilitam análises complexas e efetivas para os negócios. Com o *big data* aliado à expansão tecnológica e à disseminação de estudos sobre *machine learning*, o mercado do entretenimento – que ainda suporta grande parte das decisões de negócio forma subjetiva – abre as portas para a análise preditiva e, dessa forma, aumenta as oportunidades de desenvolvimento de conteúdo, estratégias de divulgação, descoberta e desenvolvimento de artistas.

## 2.6 – Aprendizado de Máquina

A expansão do *big data* e o crescimento das possibilidades de obtenção de dados

em grande velocidade, variedade, valor e volume ocasionou na necessidade do desenvolvimento de soluções que pudessem resolver questões complexas. Essas ferramentas deveriam ser capazes "de criar por si próprias, a partir da experiência passada, uma hipótese, ou função, capaz de resolver o problema que deseja se tratar" (FACELI et. al, 2011, p. 2). O conceito de *machine learning* (aprendizado de máquina) é, portanto, esse conjunto de processos computacionais sofisticados que aprendem com situações previamente passadas por meio de algoritmos.

Atualmente as aplicações de *machine learning* têm tido rápido e constante desenvolvimento, ao passo em os algoritmos são otimizados, modificados e continuamente propostos a diversas utilidades nos mais variados problemas de negócio.

> Um requisito importante para algoritmos de AM é que eles sejam capazes de lidar com dados imperfeitos. Muitos conjuntos de dados apresentam algum tipo de problema, como presença de ruídos, dados inconsistentes, dados ausentes e dados redundantes. Al-goritmos de AM devem, idealmente, ser robustos aos problemas presentes nos dados, minimizando sua influência no processo de indução de hipóteses. (FACELI et al 2011, p.4)

Cada algoritmo possui os vieses de representação e busca. O viés de representação é a maneira através da qual o algoritmo representa a hipótese que induz uma relação entre os dados apresentados; enquanto o viés de busca é o método utilizado para a procura da hipótese que melhor descreve a forma como os dados de treinamento estão relacionados (FACELI et al, 2011, p. 5).

As tarefas de aprendizado mais comuns são referentes a predição e descrição dos conjuntos de dados. As tarefas preditivas utilizam algoritmos capazes de prever um valor futuro baseado nas variáveis apresentadas anteriormente, exigindo atributos de entrada e saída. Já as tarefas descritivas têm o objetivo de analisar e caracterizar o conjunto de dados, encontrando associações entre os dados.

> As tarefas supervisionadas se distinguem pelo tipo dos rótulos dos dados: discreto, no caso de classificação; e contínuo, no caso de regressão. As tarefas descritivas são genericamente divididas em: agrupamento, em que os dados são agrupados de acordo com sua similaridade; sumarização, cujo objetivo é encontrar uma descrição simples e compacta para um conjunto de dados; e associação, que consiste em encontrar padrões frequentes de associações entre os atributos de um conjunto de dados (FACELI et. al, 2011, p. 6)

Os dados são submetidos a modelos de aprendizagem após as etapas de processamento e análise exploratória – que inclui a integração, limpeza, escalonamento, transformação, tratamento de dados ausentes e *outliers* etc. A preparação dos dados tem como principal objetivo reduzir os problemas e otimizar o desempenho dos algoritmos, e nesta etapa é importante considerar que tipo de estimador será utilizado, a depender do problema de negócios. Se o resultado a ser previsto pertence a um conjunto de valores nominais, é necessário

gerar um estimador de classificação; caso o valor alvo seja um conjunto infinito e ordenado de valores, o estimador gerado deve ser um regressor.

### 2.6.1 – Avaliação dos modelos: overfitting e validação cruzada

O objetivo dos modelos de *machine learning* é encontrar as funções que melhor ajustam os objetos de entrada dos dados de treino para prever com eficácia geral os valores de saída em um conjunto de teste. Modelos de aprendizado de máquina supervisionado estão suscetíveis a performar de maneira indesejada caso haja problemas com o conjunto de dados usado na etapa de treinamento. Um dos problemas mais recorrentes nesse sentido é o *overfitting* (ou, em português, super ajustamento), que acontece quando o modelo acaba ter uma performance de predição praticamente impecável e sem erros, se comparada a acurácia nos *dataset*s de treino e teste. De acordo com Ying (2019, p. 2), podemos estabelecer categorias para as causas do *overfitting*, das quais duas principais, que serão abordadas a seguir:

- Ruídos no conjunto de dados de treino: quando o *dataset* usado para treinar os modelos são pequenos, contém dados pouco representativos ou ruidosos. Nesse caso, há chances de os ruídos dos dados serem aprendidos pelos modelos e reproduzidos como base de predição;
- Complexidade de hipótese do modelo: viés e variância são resultados mais importantes a serem observados em um modelo de aprendizado supervisionado, e podem auxiliar no ajuste e na avaliação de desempenho. De forma resumida, o viés é a diferença entre o resultado esperado e as previsões do modelo, enquanto a variância é a medida do quanto as previsões variam de uma realização de previsão

para a outra. É preciso ter um equilíbrio suficiente entre o viés e a variância, e o processo de ajuste a fim de minimizar esses erros é chamado de *trade-off bias-variance*, mais essencial à medida que o modelo se torna complexo (quando mais parâmetros são adicionados).

A fim de evitar o *overfitting* podem ser implementadas algumas técnicas para ajuste dos conjuntos de dados de treino e avaliação do modelo. Para os casos em que o *dataset* se apresenta em um tamanho reduzido, a expansão do conjunto de dados pode ser possível por meio, por exemplo, da aquisição de mais dados ou geração de novos dados baseada na distribuição dos dados já existentes.

> Um conjunto de dados expandido pode melhorar a precisão de previsão em grande medida, especialmente em modelos complicados. É por isso que o aumento de dados é amplamente utilizado e provou ser eficaz como uma estratégia geral para melhorar o desempenho da generalização dos modelos em muitas áreas de aplicação, como reconhecimento de padrões e processamento de imagem (YING, 2018, p. 3)

Para os casos em que o conjunto de dados apresenta muitas variáveis, aumentando a complexidade do modelo, pode ser feita a regularização. Ainda segundo Ying (2018, p. 5), um modelo que desempenha uma previsão com *overfitting* tende a levar todas as variáveis em consideração, ainda que algumas tenham pouca ou nenhuma influência para o resultado final. Essa situação pode ser ajustada ao selecionar somente as variáveis que são úteis ao modelo, ou associar pesos às variáveis, atribuindo pesos menores às menos vantajosas.

Outra forma de reduzir o *overfitting* é acionar parâmetros que configuram o modelo a interromper o aprendizado ao atingir um determinado ponto onde o seu desempenho não é aumentado – ou até mesmo pode ser prejudicado. Essa técnica é usada principalmente em modelos de redes neurais artificiais.

Além do *trade-off* de viés e variância, podem ser usados outros métodos de medição de eficácia dos modelos. Neste trabalho, é utilizado o *score* (cálculo de porcentagem de acertos e erros para os dados de treino e teste), e também são implementadas a validação cruzada, a análise de curva ROC e a matriz de confusão.

A validação cruzada é um método simples e de fácil compreensão, que estima o desempenho do modelo de forma menos tendenciosa que o score de acurácia de treino e

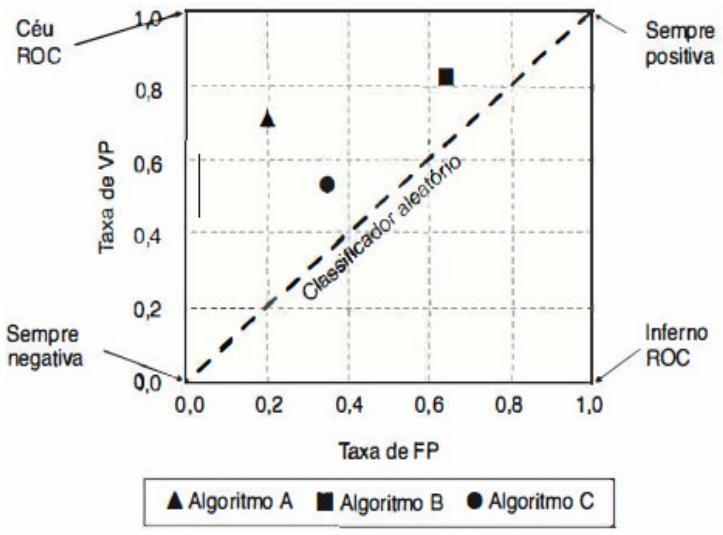
teste. Na validação cruzada (*k-fold cross validation*), o conjunto de dados de treino é dividido em k subconjuntos aproximadamente do mesmo tamanho. O processo é repetido k vezes e, em cada ciclo, os objetos de cada subconjunto são utilizados para treinar o modelo preditor que, em seguida, é testado na repartição restante. O desempenho final do preditor é então definido pela média dos desempenhos de cada um dos subconjuntos de teste.

Para problemas de classificação, o *k-fold cross validation* é estratificado, ou seja,

> mantém em cada partição a proporção de exemplos de cada classe semelhante à proporção contida no conjunto de dados total. Se, por exemplo, o conjunto de dados original tem 20% dos objetos na classe c1 e 80% na classe c2, cada partição também procura manter essa proporção, apresentando 20% de seus exemplos na classe c1 e 80% na classe c2 (FACELI et al., 2011, p. 162).

Para a regressão logística binária, um dos métodos para se avaliar o desempenho do classificador é o das curvas ROC (*Receiving Operating Characteristics*). O gráfico da ROC é bidimensional, e relaciona os parâmetros de TFP (taxa de falsos positivos) e TVP (taxa de verdadeiros positivos) variando os pontos de corte na probabilidade estimada (*threshold*). A figura abaixo ilustra os principais aspectos de uma análise de curva ROC.

**Figura 2.2 –** Espaço ROC com três classificadores




Fonte: Retirado de FACELI et al. (2011, p. 166)

No ponto (0,0), as classificações são sempre negativas, enquanto no ponto (1,1) representa as classificações sempre positivas. O ponto (0,1) representa as classificações perfeitas, onde todos os exemplos são classificados perfeitamente e não há falsos positivos (céu ROC); já no ponto (1,0) é onde todas as classificações são erradas (inferno ROC).

Um classificador é considerado melhor que um outro se seu ponto no espaço ROC encontra-se acima e à esquerda do ponto correspondente ao segundo classificador (Prati, 2006 apud Faceli et al., 2011, p. 166). Um bom modelo, portanto, é o que classifica corretamente os verdadeiros positivos à medida que o *threshold* é diminuído e antes que cresça muito o número de falsos positivos.

A matriz de confusão também é outro método simples e de fácil interpretação utilizado na avaliação para modelos de classificação, que mostra a frequência com que as classes são preditas corretamente. Dessa forma, a matriz de confusão auxilia na visualização do desempenho de classificação, mostrando a proporção de erro para cada classe testada e o quanto o modelo acertou entre as previsões possíveis.

### 2.6.2 – Modelos de aprendizado de máquina

Entre os principais modelos de aprendizado, destaco os que foram testados neste trabalho e serão abordados posteriormente nos próximos capítulos: Regressão Logística (GONZALEZ, 2018), Máquina de Vetores de Suporte (GUNN, 2018), Floresta Aleatória (PONTE; CAMINHA; FURTADO, 2020), *Perceptron* Multicamadas, *Naive Bayes*, Árvore de Decisão e KNN (*k-Nearest Neighbors*) (FACELI et al., 2011).

O KNN é um algoritmo baseado em distâncias, que classifica um novo objeto com base nas amostras do conjunto apresentado no treinamento que são próximos a ele e pode ser utilizado para problemas de classificação e regressão. É um algoritmo simples e baseado em memória, que pode ser incrementado com novos conjuntos de treinamento. Apesar disso, como o KNN calcula a distância entre o novo objeto e todos os outros objetos de treinamento, o processo pode ser demorado em um conjunto muito grande de dados.

Uma grande quantidade de atributos também pode prejudicar o desempenho do KNN, uma vez que são os atributos que definem a dimensão ocupada pelos objetos no espaço em que são calculadas as distâncias entre eles. Dessa forma, com o aumento da dimensão, a distância do objeto ao vizinho mais próximo acaba por se aproximar da distância ao vizinho mais distante (Beyer et al. 1999, apud FACELI et al. 2011, p. 64).

A árvore de decisão é outro modelo usado para problemas de classificação e regressão que visualmente se assemelha a um fluxograma, técnica tradicional de tomada de decisão a partir da divisão de problemas em problemas menores. É um modelo recursivo baseado em procura, e utilizam os chamados nós (que podem ser raiz, de divisão ou folha) que se relacionam entre si. O nó raiz é o ponto de partida, sendo o atributo que melhor divide o conjunto de dados. A partir dele são estabelecidos os nós de divisão, que contém testes condicionais que são

baseados nos valores das variáveis do conjunto de dados. Os nós folha são os nós terminais das árvores, e representam a decisão a ser tomada.

Em uma árvore de decisão, o critério a ser estabelecido para a subdivisão deve ser o quão bem um atributo consegue discriminar as classes. Portanto, o atributo que tem o maior ganho de informação é selecionado como a base para a partição da árvore. Para o cálculo desse ganho de informação podem ser usadas a Entropia, que mede a aleatoriedade (ou falta de homogeneidade dos dados) e o índice Gini, que mede a heterogeneidade dos dados. As duas medidas calculam o grau de impureza dos atributos que serão selecionados para a divisão. “A construção da árvore de decisão é guiada pelo objetivo de reduzir a entropia, isto é, a aleatoriedade (dificuldade para predizer) da variável alvo” (FACELI et al. 2011, p.87).

Outra forma de reduzir o erro das árvores de decisão é a poda, que resolve os problemas de árvores muito grandes (de difícil compreensão) que podem ter em seus nós mais profundos casos *overfitting* e decisões de baixo nível de importância devido à pouca quantidade de dados que chega a esses nós folha. Os métodos podem ser divididos entre os que finalizam a árvore quando alguma condição é atingida (pré-poda), e os fazem a poda após a construção da árvore (pós-poda). Na pós-poda, os erros do nó de decisão e dos nós seguintes são calculados, e caso o erro do nó seja menor ou igual à soma dos erros dos próximos nós, o nó de decisão é substituído por um nó folha.

Uma outra forma de reduzir o risco de *overfitting* é desenvolver mais de uma árvore de decisão, construindo as chamadas florestas aleatórias. A ideia principal é que a multiplicação de árvores de decisão aumente a aleatoriedade de seleção dos atributos e assim evite o sobreajuste dos resultados em relação aos dados de treinamento.

> Tipicamente, cada árvore que forma uma floresta tem igual colaboração na predição final. Quando se trata de classificação, a predição final é decidida por voto majoritário. Para regressão, a decisão final é uma média entre as decisões individuais (PONTE; CAMINHA; FURTADO, 2020, p. 1).

Diferentemente dos modelos abordados até então, a Regressão Logística é uma técnica de estimativa de probabilidade de ocorrência, sendo uma alternativa à Regressão Linear quando o problema envolve uma variável alvo que se apresenta de forma qualitativa. Nesse caso, a variável categórica é apresentada como 0 ou 1 e representa a ocorrência ou não de um fenômeno. Nos casos em que há somente um evento a ser previsto em questão, a regressão

logística é binária; já nos casos em que se pretende estimar a probabilidade de ocorrência de dois ou mais eventos, a regressão logística é multinomial.

> Ao se ter uma variável qualitativa como fenômeno a ser estudado, fica inviável a estimação do modelo por meio do método de mínimos quadrados ordinários estudados [...] uma vez que esta variável dependente não apresenta média e variância e, portanto, não há como minimizar a somatória dos termos de erro ao quadrado sem que seja feita uma incoerente ponderação arbitrária (FÁVERO; BELFIORE, 2017, p. 612).

Na regressão logística binária, a variável dependente segue a distribuição de Bernoulli, onde apenas dois resultados são possíveis: sucesso ou fracasso – em que a probabilidade de ocorrência do evento é p. A regressão logística estima este valor para a combinação linear de variáveis independentes utilizando a função logit, que mapeia esta combinação que poderia retornar qualquer valor em uma distribuição de probabilidades de Bernoulli, com valores de 0 e 1 (GONZALEZ, 2018, p. 16).

Para ajustar o modelo de regressão logístico é necessário estimar os parâmetros da equação de regressão por meio do método de estimação da máxima verossimilhança, ou seja, encontrar os valores desses parâmetros que maximizam a probabilidade para a ocorrência ou não de um evento. Em resumo, o modelo probabilístico de predição da

regressão logística é um método de aprendizado supervisionado baseado na observação de todos os dados do conjunto de treinamento que calcula os melhores valores de parâmetros dos atributos desses dados para prever as probabilidades de cada classe para dados futuros.

O classificador Naive Bayes também é uma alternativa de modelo preditivo probabilístico, em que os valores dos atributos de um objeto são independentes entre si. Nesse sentido, o Naive Bayes não considera as relações entre esses atributos, tratando apenas sobre a probabilidade condicional, ou seja, a probabilidade de um evento A ocorrer, dado a ocorrência ou não do evento B. Por esse motivo, a funcionalidade do Naive Bayes tem melhor desempenho nos casos em que a independência entre os atributos é assumida. Além disso, o algoritmo tem melhor performance ao ser treinado com um grande conjunto de dados.

No caso de variáveis categóricas, o algoritmo mantém um contador para cada valor de variável por classe, já que o conjunto de possíveis valores é enumerável. Já no caso de variáveis numéricas, o Naive Bayes assume uma distribuição para os valores da variável, que geralmente

é a distribuição normal. Outra solução é tornar discretas as variáveis numéricas na fase de preparação dos dados.

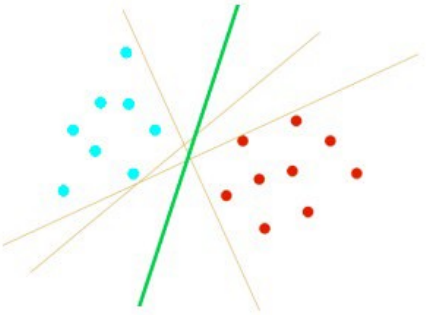
Outro algoritmo bastante funcional para a separação dos dados em classes que vem sendo cada vez mais utilizado são os SVM - *Support Vector Machines*, ou Máquinas de Vetores de Suporte. As SVM são embasadas pela teoria de aprendizado estatístico (Faceli et al., 2011,
p. 122) e podem ser utilizados para problemas de regressão, mas sua usabilidade é tradicionalmente voltada para a classificação - os SVC *(Support Vector Classifier*) e é esse o modelo utilizado neste trabalho.

Em síntese, o algoritmo do SVC linear separa os objetos em classes e busca a melhor reta de separação entre eles, analisando o objeto de uma classe que se encontra mais próximo de um objeto de outra classe. Dessa forma, o algoritmo encontra a linha – chamada de hiperplano – que maximiza a distância entre os grupos, ou seja, a que se distancia mais de cada um. A partir do exemplo de Gunn (2018):

> Existem muitos classificadores lineares possíveis que podem separar os dados, mas há apenas um que maximiza a margem (maximiza a

> distância entre ele e o ponto de dados mais próximo de cada classe). Este classificador linear é denominado como *optimal separate hyperplane*. Intuitivamente, esperaríamos que esse limite fizesse uma boa generalização quando oposto às outras fronteiras possíveis (GUNN, 2018, p.17)

**Figura 2.3 –** Espaço de Máquina de Vetores de Suporte: hiperplano para classificação

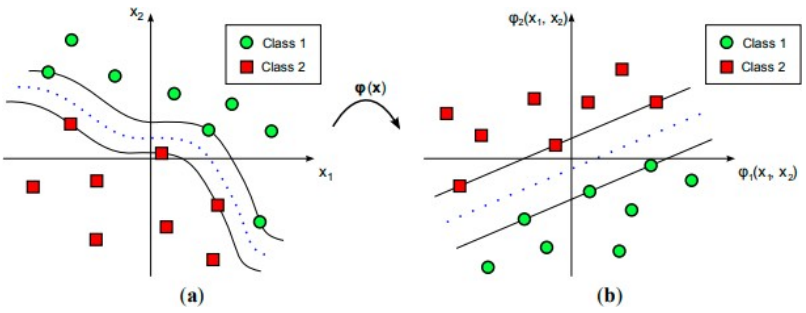


Fonte: Retirado de GUNN (1998, p.17)

Em problemas em que os objetos não são linearmente separáveis, são utilizadas as SVM não lineares, que traçam retas ou curvas que melhor separam e as classes e fazem uma transformação não linear do espaço antes de efetivamente separar esses grupos de forma linear.

> As SVMs lidam com problemas não lineares mapeando o conjunto de treinamento de seu espaço original, referenciado como de entradas, para um novo espaço de maior dimensão, denominado espaço de características (*feature space*) (HEARST et al., 1998 apud FACELI et al., 2011, p. 130)

**Figura 2.4 –** Representação de uma Máquina de Vetor de Suporte Classificadora utilizando um kernel não linear




Fonte: Retirado de RUIZ-GONZALEZ et al. (2014, p. 20720)

As Redes Neurais Artificiais (RNA) são modelos de aprendizagem baseados em otimização e inspiradas pelo funcionamento do sistema nervoso humano, com o objetivo de desenvolver sistemas de aprendizado razoavelmente semelhantes ao biológico. São formadas por unidades de processamento simples - os chamados neurônios - que computam funções matemáticas e estão interligados por conexões, sendo são dispostos em uma ou mais camadas.

> Na maioria das arquiteturas, essas conexões, que simulam as sinapses biológicas, possuem pesos associados, que ponderam a entrada recebida por cada neurônio da rede. Os pesos podem assumir valores positivos ou negativos, dependendo de o comportamento da cone-xão ser excitatório ou inibitório, respectivamente. Os pesos têm seus valores ajustados em um processo de aprendizado e codificam o conhecimento adquirido pela rede (BRAGA et al., 2007 apud FACELI et al., 2011, p. 110).

As RNA são representadas pelas definições de arquitetura e aprendizado. O primeiro termo se refere aos tipos, à quantidade e à forma de conexão entre os neurônios; enquanto o

segundo especifica as regras sobre para o ajuste dos pesos e quais as informações serão utilizadas por essas regras.

A primeira rede neural a ser implementada foi a rede *perceptron*, desenvolvida por Rosenblatt (1958). Ela possui apenas uma camada de neurônios e é treinada por um algoritmo supervisionado de correção de erros. No entanto, a existência de apenas uma camada nessas redes limita a classificação a somente objetos linearmente separáveis. Para problemas não lineares, são utilizadas mais camadas intermediárias, que definem as *MultiLayer Perceptron* (MLP) ou *Perceptron* Multicamadas (PMC).

> Em uma MLP, cada neurônio realiza uma função específica. A função implementada por um neurônio de uma dada camada é uma combinação das funções realizadas pelos neurônios da camada anterior que estão conectados a ele. À medida que o processamento avança de uma camada intermediária para a camada seguinte, o processamento realizado (e a função correspondente) se torna mais complexo (FACELI et al., 2011, p. 116).

No processo de treinamento das MLP é utilizado o algoritmo *backpropagation*, que possui as fases *forward* e *back*. Na fase *forward*, os objetos são apresentados à rede, ponderado pelos pesos das conexões de entrada e, a partir desse processamento, cada

neurônio da camada aplica a função de ativação e produz um valor de saída. Os valores de saída são utilizados como valor de entrada para os neurônios da próxima camada sucessivamente, até a camada de saída, onde o valor de saída de cada neurônio é comparado ao valor de saída desejado.

A partir da diferença entre o valor de saída do neurônio e o valor desejado é que é calculado o erro cometido, e o valor de erro é então utilizado na fase *backward*, onde os pesos de entrada são ajustados desde a camada de saída até a primeira camada. Como os valores desses erros são determinados somente para os neurônios localizados na última camada (camada de saída), o *backpropagation* utiliza os erros da camada posterior para estimar os valores dos erros das camadas intermediárias. A estimação se dá pela "soma dos erros dos neurônios da camada seguinte, cujos terminais de entrada estão conectados a ele, ponderados pelo valor do peso associado a essas conexões" (FACELI et al., 2011, p.117).

O resumo sobre a abordagem de cada modelo de aprendizado de máquina utilizado neste trabalho é descrito no quadro a seguir.

**Quadro 2.1** - Resumo dos modelos de aprendizado de máquina

| Modelo | Tipo | Descrição |
|---|---|---|
| Regressão Logística | Classificação | Modelo probabilístico de predição para variáveis categóricas, que prevê a classificação ao estimar os valores dos parâmetros que mais influenciam na ocorrência ou não de um evento. |
| Árvore de Decisão | Classificação ou Regressão | Modelo recursivo baseado em procura, que estabelece divisões para o problema principal escolhendo o atributo que faz a melhor divisão da árvore. |
| KNN | Classificação ou Regressão | Aprendizado indutivo baseado em distâncias, onde considera que objetos com características semelhantes pertencem ao mesmo grupo. |
| Floresta Aleatória | Classificação ou Regressão | Modelo que elabora diversas árvores de decisão, evitando o *overfitting* devido ao aumento da aleatoriedade de seleção dos atributos e à construção de árvores menores que as elaboradas em um modelo de árvore única. |
| Máquinas de Vetores de Suporte | Classificação ou Regressão | Algoritmo baseado na teoria de aprendizado estatístico, que traça uma reta de separação entre as classes de objeto buscando o hiperplano que define a maior distância entre os objetos de cada classe. Podem ser usados em problemas com classes separadas linearmente ou não. |

| | | |
|---|---|---|
| Perceptron Multicamadas | Classificação ou Regressão | Rede Neural que utiliza duas ou mais camadas de neurônio com pesos definidos para os objetos de entrada e utiliza o algoritmo de *backpropagation* para calcular o erro das camadas de saída e ajustar os pesos dos neurônios da rede. |
| Naive Bayes | Classificação | Modelo probabilístico de predição baseado no Teorema de Bayes, que assume a independência das variáveis do conjunto de dados e considera a probabilidade condicional para realizar predições. |

Fonte: Elaboração da autora

Os métodos, técnicas e ferramentas abordados neste capítulo são a base de construção da análise e dos modelos testados para a proposta deste trabalho, cujo objetivo, metodologia e resultados serão apresentados ao longo dos próximos capítulos.

# Capítulo 3

# Metodologia

Este trabalho tem como principal objetivo a previsão de resultado de uma nova música a partir da aplicação de modelos de *machine learning*. Como foi visto em trabalhos anteriores, devido à natureza não linear das relações entre o total de *streams* – que é o indicador de sucesso de uma faixa no meio digital – e as outras variáveis, o uso de regressão linear para a predição de sucesso baseado somente no número total de *streams* de uma faixa se mostrou pouco efetivo. Dessa forma, esse trabalho pretende ser uma iniciativa de utilização dos modelos preditivos de classificação para realizar a previsão de sucesso de uma faixa.

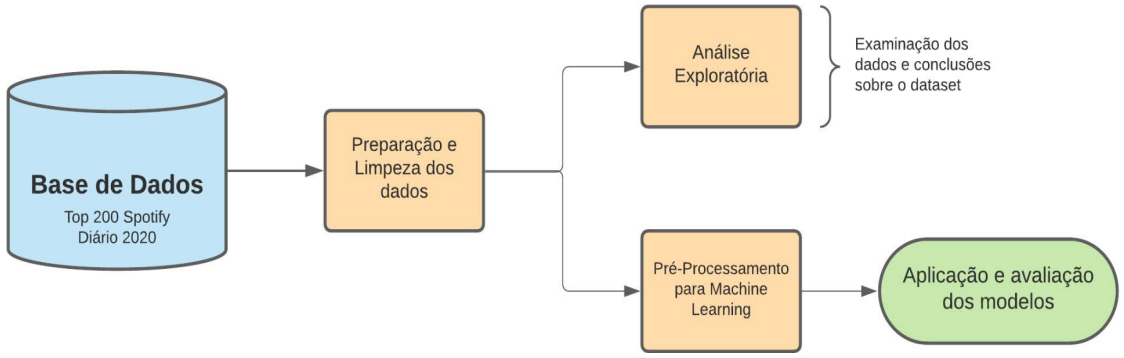
Devido à condição confidencial dos dados relacionados ao desempenho das faixas por parte das gravadoras e distribuidoras, a pesquisa utilizou dados do Top 200 diário do *Spotify* no Brasil, que disponibiliza diariamente as músicas mais ouvidas e o número de *streams* de cada faixa que foram acumulados por dia. Os dados utilizados são referentes às faixas que estiveram no Top 200 do *Spotify* ao longo de todo o ano de 2020.

O projeto é construído tendo como base atributos técnicos e artísticos próprios do meio musical, levando em consideração o formato e o conteúdo das músicas, além de variáveis geradas a partir dos dados disponíveis de forma pública. A limitação do trabalho está, portanto, relacionada à confidencialidade dos dados e à insuficiência de recursos para extração, tratamento e processamento de dados em grande volume. No entanto, em uma realidade de grandes companhias do mercado fonográfico, o potencial dessa pesquisa pode ser alavancado com o investimento em equipe, acesso a dados comerciais e tecnologia para análises e predições mais complexas e assertivas.

Esta pesquisa tem um caráter experimental, prático e aplicável. Os dados extraídos têm bastante representatividade de mercado, uma vez que o *Spotify* possui o maior *market share* entre as plataformas de *streaming*s de música. O foco principal deste trabalho é, portanto, uma continuação na busca de possíveis soluções para problemas da realidade do mercado fonográfico.

A estrutura desta proposta está dividida da forma clássica de um projeto de análise de dados, o que pode ser observado na figura 5.

**Figura 3.1 –** Etapas do projeto




Fonte: Elaboração da autora

## 3.1 – Coleta de dados

Em um primeiro momento, a ideia desse projeto se concentrava em criar um modelo que auxiliasse estratégias de *business intelligence* em uma gravadora ou distribuidora para a descoberta de novos talentos e predição de resultados a partir de dados históricos. Semanalmente a associação de produtores musicais (Pró Música) envia às instituições do mercado fonográfico um relatório comercial em que consta as 5.000 músicas mais executadas nas principais plataformas de *streaming*. No entanto, sendo um relatório restrito às empresas e artistas, a utilização desse relatório foi vetada para o uso nesta pesquisa, ainda que fossem datados de anos anteriores.

Foi feita, portanto, a tentativa de coleta de dados da página do *Spotify* onde ficam os *chart*s da plataforma por meio de um *script* de *web scrapping* em R. O *script* funcionou para os dados principais (posição no *chart*, artista, nome da faixa, data do *chart* e número de *streaming*s), mas falhou ao acessar a *tag* de identificador da faixa, sendo um empecilho para a coleta posterior de outros dados relacionados às faixas que teriam esse identificador como chave de busca.

A equipe de *business intelligence* da Sony Music Brasil gentilmente cedeu uma

planilha no formato de Excel (xlsx) com 73000 linhas contendo as faixas do Top 200 de todo o ano de 2020, entendendo que não haveria impedimentos jurídicos por se tratar de dados públicos disponíveis na página do *Spotify*[6]. Esta base inicial é composta por 6 colunas, que são descritas no quadro 3.1.

**Quadro 3.1 -** Descrição dos atributos da base de dados inicial

| Coluna | Descrição | Tipo |
|---|---|---|
| POSITION | Posição da faixa no *chart* | int |
| ARTIST | Nome do artista da faixa | str |
| TRACK | Título da faixa | str |
| *STREAMS* | Contagem de *streams* acumulados nas últimas 24h | int |
| DATE | Data do *chart* | str |
| ID | Identificador da faixa na plataforma | str |

Fonte: Elaboração da autora

Foi então desenvolvido um código em *Python* para buscar informações das faixas e os códigos identificadores dos artistas na API do *Spotify*, utilizando os métodos descritos na documentação do *Spotify for Developers*[7] (2020). Para complementar os dados, foi utilizado um outro código, também em Python, para acessar a API do *Chartmetric*, plataforma de dados musicais que reúne informações sobre faixas, artistas e playlists de várias plataformas de *streaming*. O acesso à API tem um custo mensal de US$80 (oitenta dólares americanos), mas foi gentilmente cedido por aproximadamente 3 meses e de forma gratuita pela equipe do *Chartmetric* para a elaboração desta pesquisa.

## 3.2 – Tratamento de dados

O id de cada faixa no *Spotify* é gerado de acordo com o produto em que a faixa se

[6] Os dados estão disponíveis no link: https://*Spotifycharts*.com/regional/br/daily. Acesso em 29/08/2020.

[7] A documentação do Spotify for Developers está disponível em https://developer.spotify.com/documentation/web-api/reference/#/.Acesso em 03/09/2020.

apresenta na plataforma. Por esse motivo, a mesma faixa pode se apresentar com vários ids, já que é possível (e bastante comum), por exemplo, que a faixa seja lançada como single e também dentro de um álbum. Por esse motivo, foi extraída por meio da API a informação de ISRC (código único internacional de identificação de fonogramas), para evitar a consideração de uma mesma faixa duas ou mais vezes no processo de transformação em um *dataset* de faixas únicas.

Durante a coleta, a API do *Spotify* não retornou resultados para 3 ids, resultando na obtenção de 41 linhas com valores nulos. Outra questão observada foi que algumas faixas indicavam o código de artista para o perfil *Various Artists*, página de artista *Spotify* destinada a lançamentos que contém mais de 3 artistas principais. Nesse caso, foram excluídas mais 1625 linhas referentes a 42 faixas. Além disso, a mesma situação de lançamento de uma faixa em dois ou mais produtos também poderia ter resultado em datas de lançamento diferentes. Por esse motivo, foi considerada apenas a primeira data de lançamento para cada faixa.

Por conta da ausência do acesso a um banco de dados unificado e mais bem estruturado para a extração de dados, toda a consolidação da base e geração de novas métricas da base foi feita manualmente em *Python.* A partir dos dados da base inicial foram geradas novas métricas e novas informações sobre as faixas e artistas foram coletadas via API do *Spotify* e *Chartmetric*, entre elas as *audio features*, variáveis de áudio atribuídas pelo *Spotify* para avaliação de som das músicas disponibilizadas na plataforma[8]. A estrutura da base final está descrita no quadro 3.2, e considera somente as datas de 2020.

**Quadro 3.2 –** Relação inicial de variáveis

| **Variável** | **Descrição** |
|---|---|
| TRACK | Título da faixa |
| ARTIST | Artista da faixa |
| ISRC | Código internacional de identificação da faixa |

[8] Informações sobre as *audio features* estão na documentação da API, disponível em: https://developer.*Spotify*.com/documentation/web-api/reference/#/operations/get-several-audio-features. Acesso em 03/09/2021.

| | |
|---|---|
| LABEL | Distribuidora da faixa na plataforma |
| GENRE | Gênero da faixa atribuído pela plataforma |
| RELEASE_DATE | Data de lançamento da faixa |
| POPULARITY | Popularidade da faixa (de 0 a 100) |
| ACOUSTICNESS | Medida de confiança para classificar se a faixa é acústica ou não. Valor entre 0 e 1, onde 1 representa alta confiança de que a faixa é acústica |
| DANCEABILITY | Descreve o quão adequada uma faixa é para dançar com base em uma combinação de elementos musicais, incluindo tempo, estabilidade do ritmo, força da batida e regularidade geral. Um valor de 0 é menos dançante e 1 é mais dançante |
| ENERGY | Energia é uma medida de 0 a 1 e representa uma medida perceptual de intensidade e atividade. Normalmente, as faixas energéticas parecem rápidas, altas e barulhentas |
| INSTRUMENTALNESS | Prevê se uma faixa não contém vocais. Quanto mais próximo o valor da instrumentalidade estiver de 1, maior será a probabilidade de a faixa não conter conteúdo vocal Valores acima de 0,5 representam faixas instrumentais, mas a confiança é maior à medida que o valor se aproxima de 1 |
| LOUDNESS | O volume geral de uma faixa em decibéis (dB) |
| SPEECHINESS | Detecta a presença de palavras faladas em uma faixa, mais próximo de 1 será o valor do atributo |
| TEMPO | O tempo estimado geral de uma faixa em batidas por minuto (BPM). Na terminologia musical, o tempo é a velocidade ou ritmo de uma determinada peça e deriva diretamente da duração média do tempo |
| TIME_SIGNATURE | Uma estimativa de fórmula de compasso geral de uma faixa. *Time signature* é uma convenção notacional para especificar quantas batidas existem em por compasso |

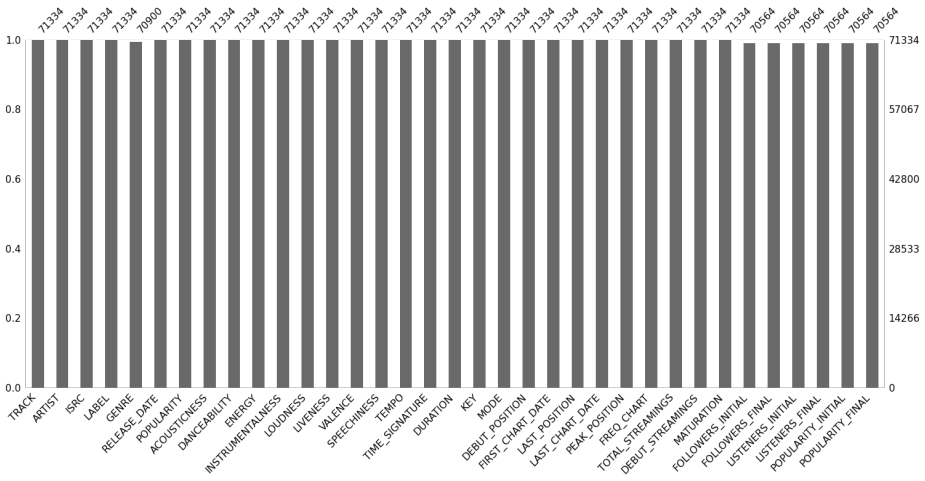
| | |
|---|---|
| DURATION | Duração da faixa em milisegundos |
| KEY | O tom musical da faixa. Aqui são mapeados usando a notação padrão de classe de tom. Por exemplo. 0 = C, 1 = C♯ / D ♭, 2 = D e assim por diante |
| MODE | Indica o tipo de escala da qual seu conteúdo melódico é derivado (maior ou menor). O maior é representado por 1 e o menor é 0 |
| LIVENESS | Detecta a presença de um público na gravação. Valores de mais altos representam um aumento na probabilidade de a trilha ter sido executada ao vivo |
| VALENCE | Uma medida de 0 a 1 que descreve a positividade musical transmitida por uma faixa. Faixas com alta valência soam mais positivas (por exemplo, feliz, alegre, eufórico), enquanto faixas com baixa valência soam mais negativas (por exemplo, triste, deprimido, com raiva). |
| DEBUT_POSITION | Posição de estreia da faixa no *chart* |
| FIRST_CHART_DATE | Primeira data em que a faixa esteve no *chart* |
| LAST_POSITION | Última posição da faixa no *chart* |
| LAST_CHART_DATE | Última data em que a faixa apareceu no *chart* |
| PEAK_POSITION | Posição de pico da faixa |
| FREQ_CHART | Quantidade de dias em que a faixa esteve no *chart* |
| TOTAL_*STREAMS* | Total de *streaming*s acumulados pela faixa em 2020, considerando a soma dos *streaming*s que dias em que a faixa esteve no *chart* |
| DEBUT_*STREAMS* | Quantidade de *streaming*s no primeiro dia em que faixa apareceu no *chart* |
| MATURATION | Diferença entre [FIRST_CHART_DATE] e [RELEASE_DATE] |

| | |
|---|---|
| FOLLOWERS_INITIAL | Número de seguidores no *Spotify* na primeira data em que o artista esteve no *chart* |
| FOLLOWERS_FINAL | Número de seguidores no *Spotify* na última data em que o artista esteve no *chart* |
| LISTENERS_INITIAL | Número de ouvintes no *Spotify* na primeira data em que o artista esteve no *chart* |
| LISTENERS_FINAL | Número de ouvintes no *Spotify* na última data em que o artista esteve no *chart* |
| POPULARITY_INITIAL | Popularidade do artista no *Spotify* na primeira data em que o artista esteve no *chart* (de 0 a 100) |
| POPULARITY_FINAL | Popularidade do artista no *Spotify* na última data em que o artista esteve no *chart* (de 0 a 100) |

Fonte: Elaboração da autora

Como o modelo de previsão deste trabalho leva em consideração a medida de sucesso e metas atingidas por faixa, na etapa de tratamento de dados foi realizada a limpeza e considerados apenas os registros únicos. A relação da quantidade de valores de cada atributo do *dataset* mostrou que a coluna com o gênero continha 434 linhas com valores nulos, enquanto nas colunas de seguidores, ouvintes e popularidade dos artistas, tanto dos números iniciais quanto finais, a situação acontecia em 770 linhas. Isso significa que as APIs não retornaram essas informações, sendo a do *Spotify* no caso de gênero e o *Chart*metric no caso das métricas de popularidade do artista (*fan metrics*).

**Gráfico 3.1 -** Quantidade de dados por coluna




Fonte: Elaboração da autora

No caso do campo de gênero, por se tratar de uma variável categórica, não foi possível fazer uma estimação para os valores nulos. Na avaliação global de valores nulos, foi constatado que esses valores nulos afetavam o total de 75 faixas únicas, correspondendo a aproximadamente 7% do *dataset*. Portanto, optou-se pela exclusão desses casos.

O atributo gênero, inicialmente, seguia um padrão pouco útil para a análise, contendo uma grande variedade de classificações que não seguiam um formato muito restrito. Por exemplo, as faixas do gênero sertanejo poderiam aparecer como sertanejo universitário, as de funk também apareciam como funk carioca etc. Para reduzir a especificidade de gênero e aprimorar a análise, foi aplicado um tratamento na coluna para a transformação e redução de 82 para 21 gêneros.

Para seguir com as análises de faixas únicas, foi feita a exclusão por duplicidade com base no ISRC, resultando em um *dataset* completo com **1.012 faixas únicas** e um total de 35 variáveis.

# Capítulo 4

# Análises e Resultados

Neste capítulo serão demonstradas as análises que consideraram o aspecto geral do Top 200 no *Spotify* em 2020. Em primeiro lugar foi feita uma análise exploratória da base de dados para o estudo da distribuição das variáveis e das possíveis correlações entre elas, com o intuito de gerar conclusões sobre a natureza e o comportamento dessas faixas na plataforma.

Após a análise exploratória, os dados foram submetidos a um pré-processamento para a aplicação dos modelos de *machine learning* explicados de forma teórica anteriormente neste trabalho.

## 4.1 – Análise Exploratória

A análise exploratória é uma etapa crucial para a análise de dados, uma vez que possibilita a investigação e o entendimento do conjunto de dados. John W. Tukey (1978) considera que a análise exploratória de dados nunca poderá ser um estudo completo, mas nada além dessa etapa pode servir como o primeiro passo para um trabalho *de data science*.

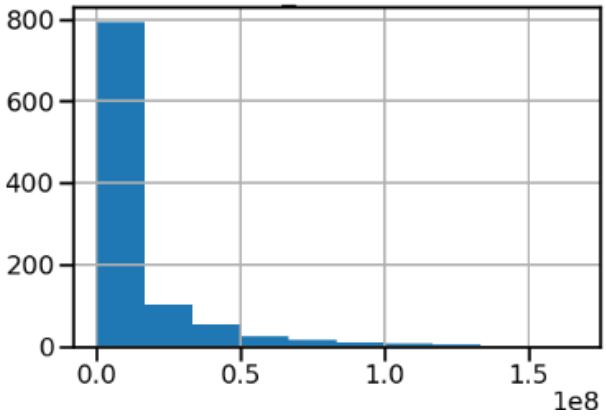
> O exame inicial dos dados, antes da aplicação da técnica de análise multivariada em si, trata-se de uma parte essencial do processo, pois permite o desenvolvimento de uma visão crítica das características dos dados a partir de uma compreensão primária da base e das relações entre as variáveis (HAIR et al., 2005, apud PASTORE; REZENDE; TEIXEIRA, 2020, p. 42).

O objetivo desta etapa foi gerar conclusões, estudar a distribuição dos dados, encontrar relações entre as variáveis, possíveis outliers e valores que poderiam não fazer sentido no contexto da base de dados. Uma análise simples e efetiva comprovou alguns pontos sobre a realidade do mercado fonográfico que foram discutidos ao longo deste trabalho e serão abordados a seguir.

Ainda que se tratando de um recorte restrito de faixas que estiveram entre as 200 mais ouvidas no *Spotify*, 71% dessas faixas atingiram no máximo 10 milhões de *streams* acumulados no ano de 2020, e só 6% teve mais de 50 milhões de *streams*. A variável número de *streams* acumulados em 2020 teve um desvio padrão de 21,68 milhões, indicando uma alta dispersão desses dados. Além disso, apenas 8% das faixas atingiram as 10 primeiras posições, e só 10% levaram 15 dias ou menos entre a data de lançamento e a data de estreia no *chart.*

Pode-se inferir, portanto, que existem casos de eventos extremos, como é visualmente notório no histograma da variável. Como pode ser observado, a imensa maioria dos valores se concentra até 50 milhões de *streams*, e existem faixas que atingem mais de 100 milhões, o que comprova que muito poucas músicas triunfam de forma excepcional.

**Gráfico 4.1** - Histograma de dispersão da variável Total_Streams




Fonte: Elaboração da autora

Durante a análise de dispersão da variável maturação, foram identificados alguns outliers. Valores muito altos de maturação podem indicar faixas antigas que entraram em voga por algum motivo bastante tempo após o lançamento, geralmente porque foram usadas em produtos audiovisuais (filmes, séries, jogos, propagandas etc.), em caso de reuniões de bandas ou, em muitos casos, porque viralizaram em alguma rede social.

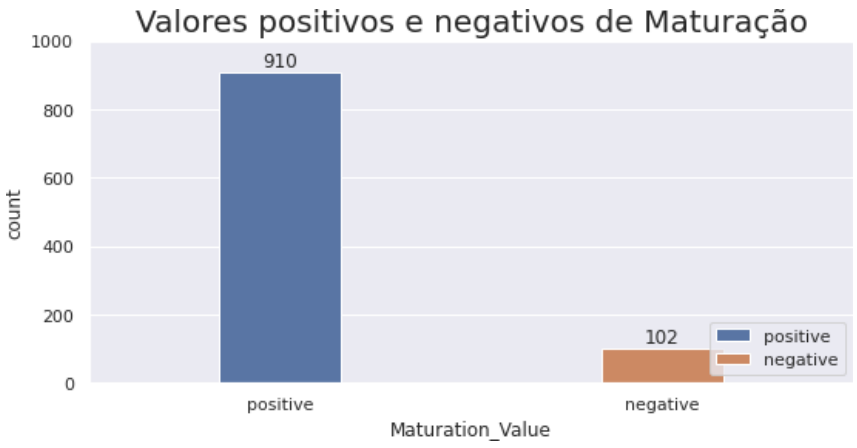
Alguns valores negativos também foram localizados, indicando que a faixa teria sido lançada após a sua estreia no *chart*. Esses valores identificam um erro ou estratégia de registro da release date por parte da distribuidora que entrega a faixa às plataformas. Alguns produtos podem ser lançados com o que se considera “data de estreia retroativa”. Nesses casos, os produtos assumem uma data de lançamento anterior à data que de fato

estrearam na plataforma, e esse procedimento é feito quando se deseja que o produto não receba o destaque como último lançamento (latest release) na página do artista, geralmente devido à escolha comercial de priorizar outros lançamentos.

**Gráfico 4.2** - Gráfico de dispersão da variável Maturação


Fonte: Elaboração da autora

**Gráfico 4.3** - Gráfico de contagem de valores positivos e negativos da variável Maturação

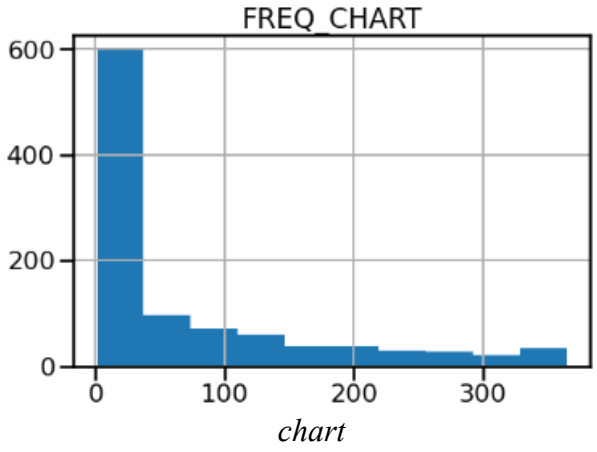



Fonte: Elaboração da autora

A frequência de *chart* é outra métrica considerada neste trabalho para o sucesso de uma faixa. Assim como a quantidade de *streams*, a quantidade de dias em que a faixa aparece no Top 200 durante o ano também é um indicador sobre a qualidade do ciclo de vida da faixa ao apontar o tempo em que ela esteve entre as mais ouvidas do país. No histograma da variável, a conclusão prática acompanha a teoria sobre poucos produtos fazerem muito sucesso, mostrando que a maior parte das faixas tem uma frequência baixa de *chart*. Esse comportamento comprova

a situação de músicas de sucesso que tiveram uma ótima performance inicial – isso geralmente acontece na estreia, se a música conta com um bom plano de marketing para divulgação – mas ficaram poucos dias em evidência.

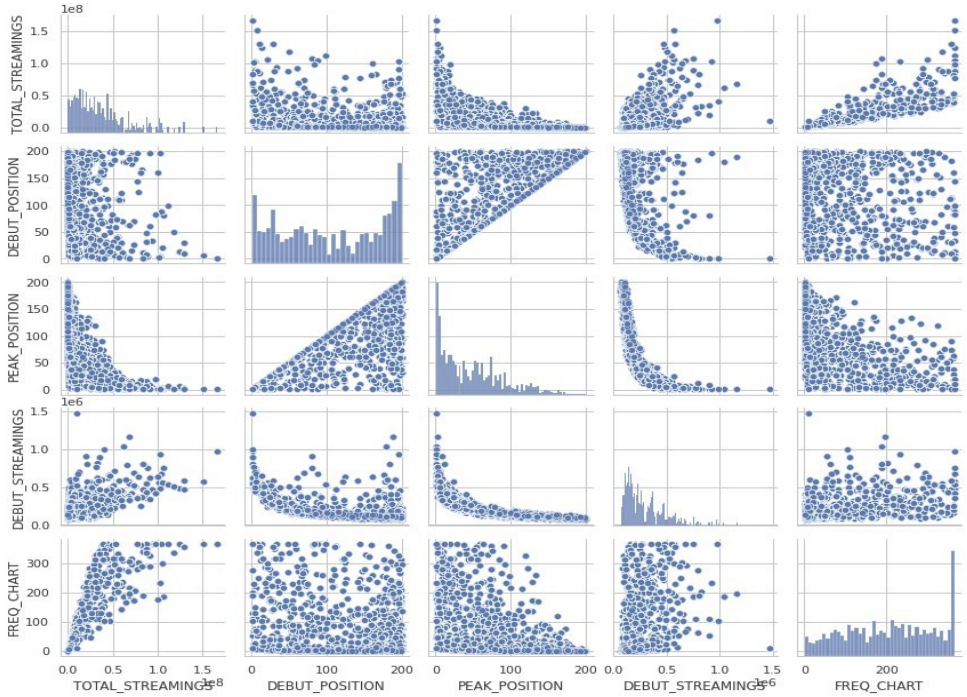
**Gráfico 4.4** - Histograma da variável de frequência de




Fonte: Elaboração da autora

Para avaliação mais apurada sobre a influência da performance inicial das faixas no desempenho total (ao final do ano), foi desenvolvido um diagrama de dispersão que combina as variáveis de posição e *streams* de estreia com o total de *streams*, posição de pico e frequência de *chart*.

**Gráfico 4.5** - Diagrama de dispersão




Fonte: Elaboração da autora

Por apresentar uma distribuição assimétrica, os dados apresentaram uma dispersão desigual das variáveis. De acordo com o diagrama, é possível perceber que a variável de *streams* acumulados não tem relação específica com a posição de estreia da faixa, enquanto apresenta uma relação próxima do linear com as variáveis de quantidade de *streams* na estreia e a frequência de *chart*.

No entanto, algumas observações também podem comprovar o caráter imprevisível do mercado fonográfico. A frequência de *chart* apresenta uma distribuição de situações bem diversas quando relacionada às posições de pico e estreia, mostrando que existem: 1) faixas que alcançaram as primeiras posições, mas permaneceram no *chart* por pouco tempo; 2) faixas que estrearam e alcançaram posições mais baixas, mas permaneceram bastante tempo em evidência.

## 4.2 – Correlação

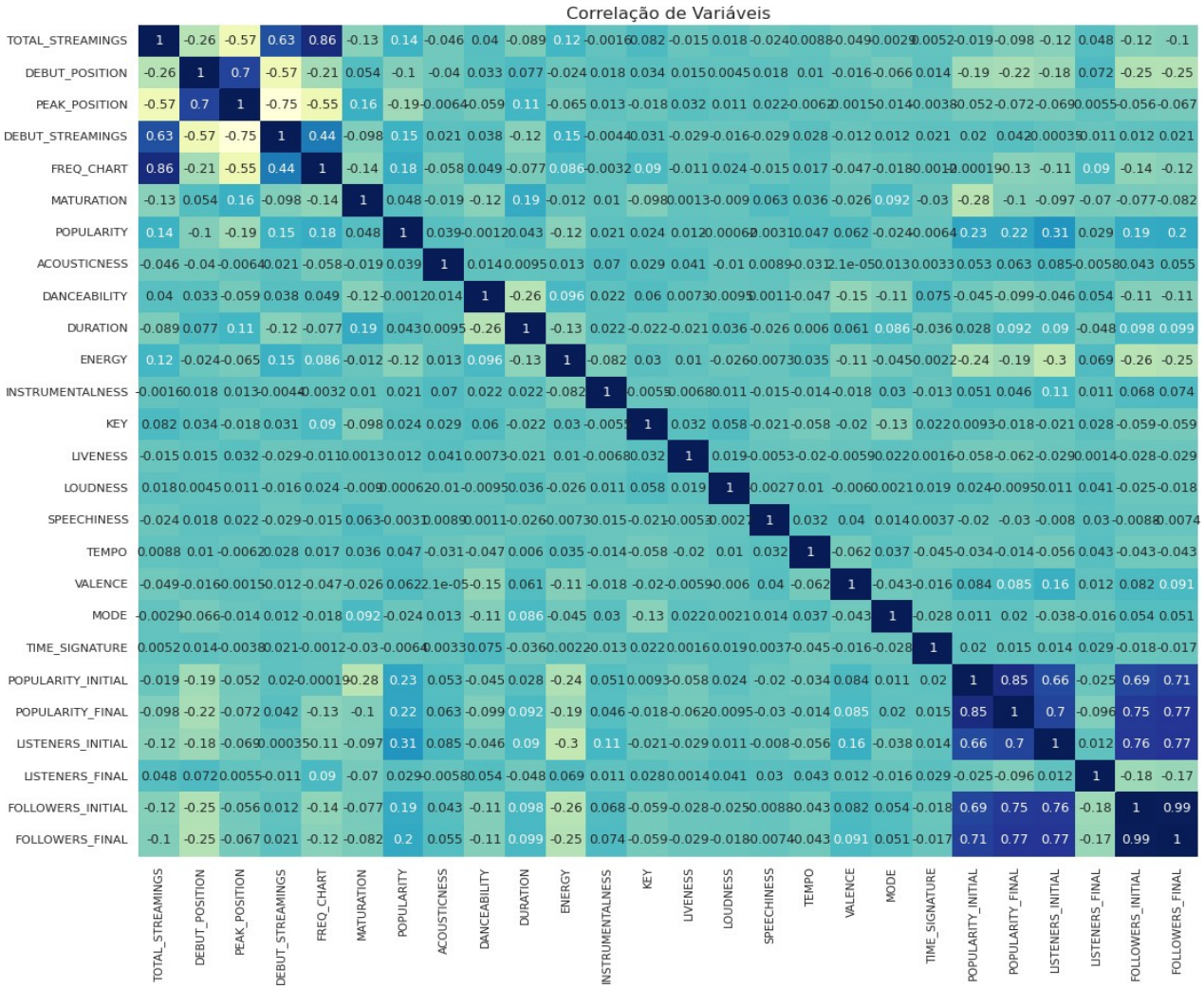
Em seguida, iniciou-se os estudos sobre as *audio features*, com o intuito de estudar as relações entre os aspectos técnicos das músicas e as métricas de performance das faixas.

Para facilitar a visualização e identificar as relações, foi aplicado ao *dataset* o teste de correlação para testar a relação entre as variáveis e, assim, concluir sobre a influência linear de cada uma das *features* no resultado sobre o desempenho das músicas. O coeficiente de correlação utilizado é o de Pearson, calculado segundo a fórmula:

$$p = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{r}(y_i-\bar{y})^2}} = \frac{cov(X,Y)}{\sqrt{var(X)\cdot var(Y)}}$$

A partir do cálculo das correlações foi gerada a matriz de correlação para todas as variáveis.

**Gráfico 4.6** - Matriz de Correlação

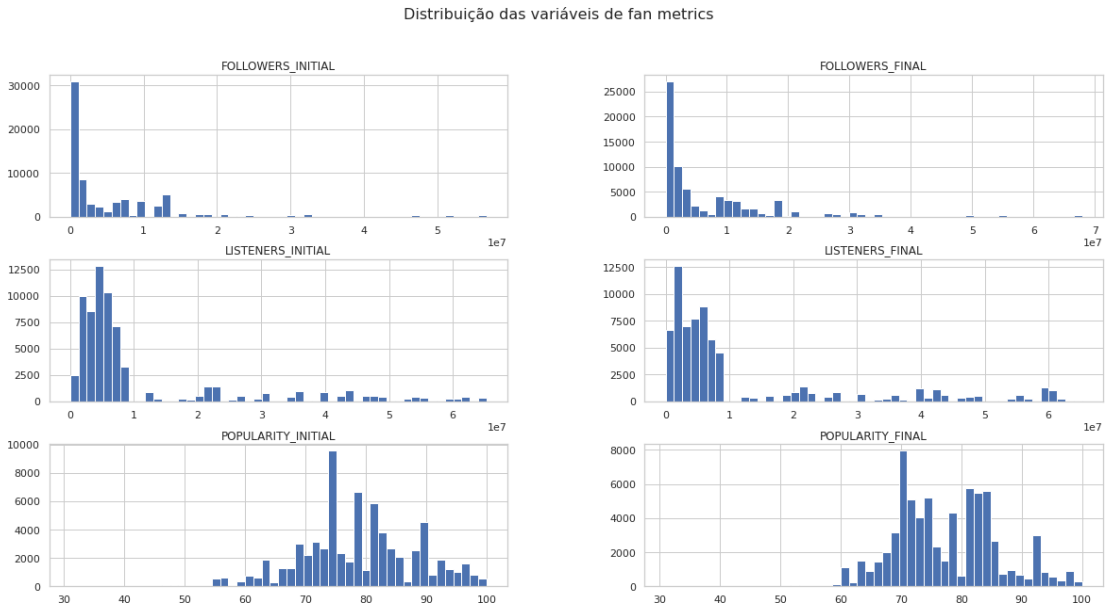



Fonte: Elaboração da autora

Entre as possíveis conclusões, o principal a ser destacado é que as audio features possuem relações lineares pouco significativas com a performance das faixas, indicando que esses dados disponibilizados pelo *Spotify* pouco auxiliam nos insights de performance em relação à produção técnica da música em si. A correlação mais forte entre as *audio features* e as métricas de performance foi observada para a variável *energy,* que indica a quão rápida e dinâmica é uma música. Isso aponta uma possibilidade de que faixas mais enérgicas tenham melhor desempenho.

Comprovou-se, portanto, que as relações mais fortes (acima de 0,5) estão nas relações apresentadas anteriormente no diagrama de dispersão. Além disso, é importante destacar uma alta colinearidade entre as *fan metrics* (variáveis de desempenho do artista). Ao estudar melhor a relação, concluiu-se que a alta colinearidade de métricas iniciais e finais é devido à pouca diferenciação quanto à distribuição.

**Gráfico 4.7 -** Histograma das variáveis de fan metrics




Fonte: Elaboração da autora

## 4.3 – Análise por gênero

Visto que o Top 200 do *Spotify* é um *ranking* que tem como único critério o volume de *streams* diário, era esperado que o conjunto apresentasse uma distribuição variada de características, o que também foi observado no trabalho de Pastore, Teixeira

e Rezende.

> Isso porque essas faixas estão relacionadas tanto a diferentes gêneros musicais, quanto à artistas de diferentes tamanhos, além de representarem tanto conteúdo nacional como internacional, por exemplo. Tais distinções, bem como a natureza da distribuição do consumo em um *chart* musical são, assim, responsáveis por gerar uma grande heterocedasticidade dos dados (PASTORE; TEIXEIRA; REZENDE, 2020, p. 53).

Seguindo essa observação, foi realizada uma análise de correlação por gênero na tentativa de encontrar possíveis relações entre métricas de *audio features* e variáveis de desempenho das faixas que, em teoria, compartilham semelhanças entre si. Para isso, foram considerados os gêneros mais frequentes no *dataset*: sertanejo, pop, funk e forró, que representam quase 70% do conjunto de dados.

**Tabela 4.1 -** Frequência de gênero

| | Gênero | Contagem | Frequência (%) |
|---|---|---|---|
| 0 | pop | 273 | 26.976 |
| 1 | sertanejo | 174 | 17.194 |
| 2 | funk | 165 | 16.304 |
| 3 | forró | 85 | 8.399 |
| 4 | hip-hop | 56 | 5.534 |
| 5 | pagode | 42 | 4.150 |
| 6 | latin | 37 | 3.656 |
| 7 | electronic music | 36 | 3.557 |
| 8 | rap | 29 | 2.866 |
| 9 | outros | 25 | 2.470 |
| 10 | r&b | 23 | 2.273 |
| 11 | brasileira | 22 | 2.174 |
| 12 | alternative | 11 | 1.087 |
| 13 | axe | 10 | 0.988 |
| 14 | rock | 10 | 0.988 |
| 15 | gospel | 8 | 0.791 |
| 16 | mpb | 5 | 0.494 |
| 17 | bossa_nova | 1 | 0.099 |

Fonte: Elaboração da autora

No entanto, a correlação entre *audio features* e variáveis de desempenho não aumentou significativamente após a clusterização por gênero. Apesar disso, algumas observações interessantes valem ser destacadas:

- No gênero sertanejo, a maturação apresentou uma correlação de 0,43 em relação à posição de pico, indicando que faixas desse gênero alcançam melhores posições em um menor intervalo de tempo entre a data de lançamento e a data de estreia no *chart*. Essa característica comprova o caráter viral das faixas do sertanejo e o potencial desse gênero no Brasil, o que já é de comum conhecimento no mercado fonográfico brasileiro;
- Em relação ao pop, *energy* apresentou uma correlação de 0,24 com os *streams* de estreia no *chart*. Apesar de, nesse *dataset*, o gênero pop ser bastante homogêneo (por considerar qualquer vertente da música pop), isso indica uma tendência de que faixas pop mais enérgicas têm um desempenho mais significativo no *chart*;
- No forró, a *key* apresentou uma correlação de 0,22 com a soma de *streams* acumulados. Mesmo sendo relativamente fraca, essa associação demonstra a tendência de que as faixas de sucesso do forró apresentam notas mais agudas.

## 4.4 – Delimitação de dataset

Após as análises feitas previamente neste capítulo e a fim de aplicar os modelos de *machine learning*, foram retiradas do conjunto as métricas de *fan metrics*, considerando que a análise descritiva destas variáveis apresentou dados muito semelhantes e a correlação entre elas apresentou índices muito altos. A colinearidade entre as variáveis dependentes reduz o poder preditivo dos modelos e, por este motivo, optou-se pela exclusão destas.

> Quando a colinearidade aumenta, a variância única explicada por cada variável independente diminui e o percentual da previsão compartilhada aumenta. [...] Para maximizar a previsão a partir de um dado número de variáveis independentes, o pesquisador deve procurar variáveis independentes que tenham baixa multicolinearidade com as outras variáveis independentes, mas também apresentam correlações elevadas com a variável dependente. (HAIR et al, 2005 apud PASTORE; TEIXEIRA; REZENDE, 2020, p. 55)

Além disso, para o início da aplicação de modelo preditivos, foram consideradas como variáveis preditoras as que os possíveis gestores de uma gravadora/distribuidora teriam acesso no momento de estreia da faixa no *chart* ou em relatórios comerciais. Essas variáveis seriam as que ajudariam a prever o potencial comercial de uma música, auxiliando na tomada de decisão sobre a contratação de novos artistas e investimento de

divulgação, estratégias de marketing e busca de parcerias.

Por esse motivo, a variável popularity foi também descartada do *dataset*, já que se refere à popularidade da faixa somente no momento da coleta dos dados na API do *Spotify*. Não foi encontrada uma maneira de recolher essa informação de forma retroativa para avaliar o desempenho histórico e evolução de popularidade da faixa.

Portanto, as variáveis utilizadas a partir deste momento para a aplicação dos modelos preditivos - que serão abordados no próximo subcapítulo - estão relacionadas a seguir:

**Quadro 4.1** - Relação de variáveis utilizada nos modelos

| **Variáveis Preditoras (independentes)** | |
|---|---|
| **Variável** | **Tipo** |
| DEBUT_POSITION | numérica |
| TOTAL_*STREAMING*S | numérica |
| GENRE | categórica |
| AUDIO FEATURES | numérica |
| MATURATION | numérica |
| **Variáveis de Resposta (dependentes)** | |
| TOTAL_*STREAMING*S | numérica |
| PEAK_POSITION | numérica |
| FREQ_CHART | numérica |

Fonte: Elaboração da autora

## 4.5 – Aplicação de modelos preditivos

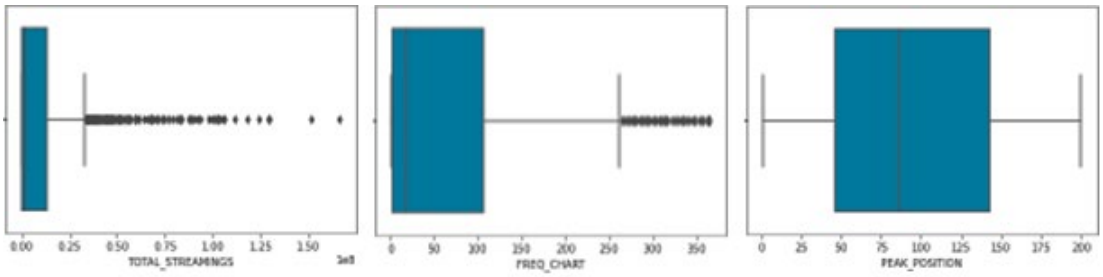
A próxima etapa do projeto consiste em desenvolver os modelos preditivos e avaliar o desempenho a fim de encontrar a melhor aplicação na predição de sucesso de uma faixa, levando em conta as variáveis dependentes e independentes que foram relacionadas no capítulo anterior.

Com a conclusão sobre o conjunto de dados ser assimétrico, os modelos apresentam maior dificuldade de predição de valores numéricos. Este trabalho, portanto, foca em elaborar um modelo predição que estima o potencial de sucesso, e para isso foi gerada uma classificação que dividiu o conjunto de dados em 3 grupos: os das faixas que tiveram um excelente desempenho, os das faixas que tiveram menos êxito mas que foram consideradas destaques, e as faixas que obtiveram um desempenho mais tímido.

Apesar de entendido que o universo de faixas é bastante reduzido, o trabalho busca estudar a possibilidade de oferecer aos gestores e equipes do mercado fonográfico uma facilidade em obter uma visão futura capaz de qualificar o potencial de desempenho de seus produtos e artistas.

Para a divisão dos dados nessas categorias, foi aplicada uma função que considerou as variáveis de resposta – total de *streams* acumulado, frequência de *chart* e posição de pico. Isso porque ficou entendido que essas são as principais características de desempenho de uma música que diferenciam os verdadeiros *hits* das faixas que, apesar de conseguirem entrar no Top 200, apresentaram uma performance mais tímida e com menor destaque. Essa classificação foi aplicada a partir da observação do comportamento dessas variáveis por meio dos *boxplots* gerados.

**Gráfico 4.8 -** Boxplots das variáveis consideradas na função de classificação




Fonte: Elaboração da autora

A função aplicou uma classificação de *tier* (em grupos) que considerou os seguintes critérios:

- Receberam a classificação de *tier* 1 às faixas que tiveram pelo menos 500 milhões de *streams*, ou estiveram por 60 dias ou mais no *chart*, ou alcançaram as 50 primeiras posições;
- Receberam a classificação de *tier* 2 as faixas que tiveram 100 milhões de *streams*, ou estiveram por 30 dias ou mais no *chart*, ou alcançaram as 125 primeiras posições;
- Receberam a classificação de *tier* 3 as demais faixas que não alcançaram os critérios de *tier* 1 e 2.

Ao final, as faixas assumiram essas três categorias, sendo 444 assumindo o *tier* 1, 287 classificadas como *tier* 2 e 281 classificadas como *tier* 3. Esse método de discretização dos valores foi, portanto, supervisionado e não paramétrico, uma vez que o intervalo de dados foi definido utilizando somente as informações presentes nos valores dessas variáveis. “As técnicas supervisionadas geralmente levam a melhores resultados, uma vez que a definição dos intervalos sem conhecimento das classes pode levar à mistura de classes” (FACELI et al., 2011, p. 44).

Após a classificação dos dados em *tier*, iniciou-se a técnica de preparação dos dados para a aplicação dos modelos. Essa etapa incluiu os seguintes passos:

1. Remoção das variáveis alvo utilizadas na função de classificação de *tier* (total de

*streams* acumulados, frequência de *chart* e posição de pico);

2. Aplicação do *one-hot encoding* para a transformação da variável preditora gênero em *dummy*, "que assumem valores iguais a 0 ou 1, de forma a estratificar a amostra da maneira que for definido determinado critério, evento ou atributo, para, aí assim, serem incluídas no modelo em análise" (FÁVERO; BELFIORE, 2017, p. 541).

3. Aplicação do *standard scaler* para padronizar os dados e atribuí-los média 0 e desvio padrão 1, tornando-os mais "digeríveis" aos algoritmos. Segundo a documentação desse método[9], essa padronização é um requisito para aplicação de modelos de aprendizado de máquina, uma vez que eles podem ter o comportamento afetado caso as variáveis não assumirem mais ou menos a distribuição padrão normal.

4. Transformação da variável categórica *tier* em valores entre 0 e n-1, sendo n o número de classes, tornando mais legível a classificação categórica aos modelos de aprendizado.

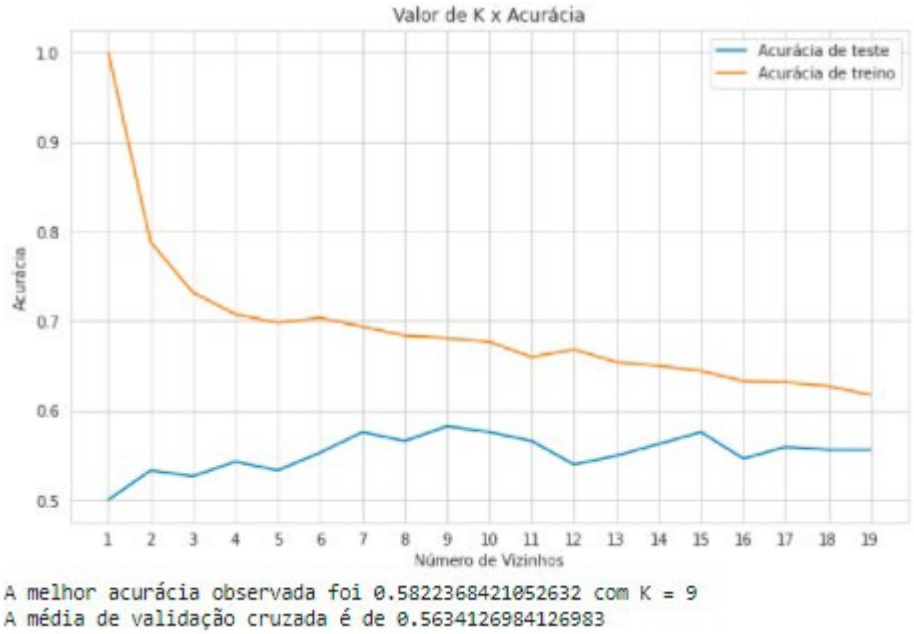
5. Finalizando a etapa de preparação dos dados para o modelo, foi feita a separação dos dados em variáveis preditoras (X) e variável alvo (Y) e a divisão em *dataset*s de treino (70%) e teste (30%).

[9] Disponível em https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html. Acesso em 15/11/2021.

### 4.5.1 – KNN

O primeiro modelo aplicado foi o KNN (*k-nearest neighbor*), sendo utilizado um *range* de 1 a 20 vizinhos cuja acurácia foi testada para cada um e seu desempenho pode ser observado no gráfico 1:

**Gráfico 4.9** - Acurácia do modelo KNN




Fonte: Elaboração da autora

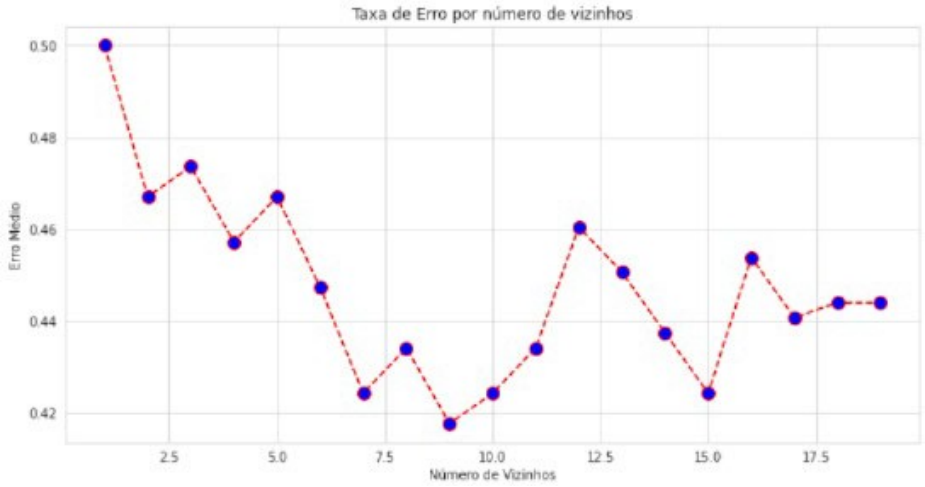
A melhor acurácia observada foi de **58% para 9 vizinhos mais próximos**, um resultado bem próximo da média de validação cruzada para 10 novas subdivisões treino/teste (56%). O KNN aponta a melhor acurácia considerando não somente a porcentagem de acertos do modelo como também o menor número de estimadores (vizinhos) utilizados. O relatório de classificação mostra uma boa precisão do modelo para estimar a classificação de *tier* 1 e um erro médio que decai até a estimação com 9 vizinhos, voltando a crescer após esse número.

**Tabela 4.2 -** Relatório de classificação do modelo KNN

| **Relatório de Classificação KNN** | | | |
|---|---|---|---|
| *tier* | Precision | Recall | F1-Score |
| 1 | 0,69 | 0,61 | 0,64 |
| 2 | 0,41 | 0,48 | 0,44 |
| 3 | 0,51 | 0,55 | 0,53 |

Fonte: Elaboração da autora

**Gráfico 4.10 -** Taxa de erro por número de vizinhos (KNN)




Fonte: Elaboração da autora

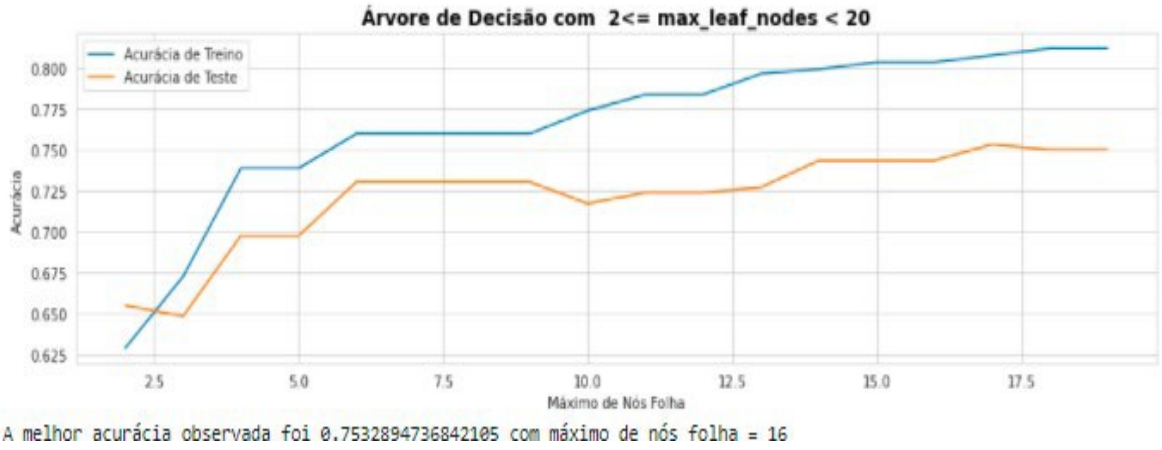
## 4.5.2 – Árvore de Decisão

O segundo modelo de previsão aplicado foi a árvore decisão, testado com entre 2 e 20 nós folhas e utilizando a entropia como critério de divisão dos dados a cada nó. Quanto aos parâmetros, a árvore foi testada utilizando a máxima profundidade para os valores 100, 20 e nenhum, e para os critérios entropia e *gini*, não havendo significativas mudanças de performance na comparação dos resultados. A melhor acurácia observada foi de **72% com 16 nós folhas**, e a validação cruzada para 10 novas divisões apresentou valores semelhantes de acurácia com média de 71%. Por fim, a taxa de erro para os valores de nós

folhas testados não variou muito, sendo observados valores entre 34% e 27%.

Uma modificação significativa no modelo foi em relação ao parâmetro *splitter* - que é a estratégia do algoritmo de selecionar a divisão em cada um dos nós[10]. Após a mudança de *random* (estratégia de melhor divisão aleatória) para *best* (melhor divisão), o modelo apontou a melhor performance também para 16 nós, mas com uma acurácia de 75%, valores também semelhantes entre si para a validação cruzada e erro médio de 25%.

**Gráfico 4.11 -** Acurácia do modelo de Árvore de Decisão




Fonte: Elaboração da autora

## 4.5.3 – Floresta Aleatória

Como abordado anteriormente, a floresta aleatória é um conjunto de árvores de decisão que utiliza a média de acurácia de cada árvore para melhorar o modelo de previsão e evitar o *overfitting*. A floresta aleatória desenvolvida para este trabalho utilizou o número padrão de estimadores (árvores) do algoritmo[11] (100), a entropia como critério de divisão e nenhum valor para o parâmetro de máxima profundidade da árvore.

As mudanças testadas nos valores dos parâmetros (50 árvores, índice gini como

[10] A documentação da Árvore de Decisão para classificação está disponível em https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. Acesso em 15/11/2021

[11] A documentação da Floresta Aleatória para classificação está disponível em https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. Acesso em 15/11/2020.

critério de divisão e 50 no valor de máxima profundidade) não alterou a performance do modelo de

forma significativa, o que também aconteceu na aplicação da árvore de decisão. O modelo obteve uma acurácia de teste de aproximadamente 74%, bem próximo do valor de acurácia da árvore de decisão.

**Gráfico 4.12 -** Matriz de Confusão de predição da Floresta Aleatória


Fonte: Elaboração da autora

A matriz de confusão demonstra que para as faixas *tier* 2 a floresta aleatória teve mais dificuldade de fazer a predição, o que pode indicar que seja necessária uma melhor modelagem dos dados a fim de aperfeiçoar a precisão do modelo no sentido de fazer uma melhor diferenciação entre as especificações que discriminam uma música de um *tier* para outro.

## 4.5.4 – Naive Bayes

Para o teste do algoritmo Naive Bayes já era esperada uma performance ruim devido à independência das variáveis não ser assumida e, principalmente, em razão da baixa quantidade e distribuição assimétrica dos dados. Ainda assim, o modelo foi testado como uma opção probabilística de predição, mas alcançou somente 37,5% de chance de acertar as predições quando utilizado o modelo Gaussiano (de distribuição normal).

A performance do Naive Bayes indica que as relações de probabilidade entre as ocorrências de uma ou outra variável não auxilia de forma útil a predição de uma nova música, dada as devidas características, ter um desempenho melhor ou pior no *chart*. Isso pode ser explicado pelas características de *audio features* terem apresentado pouca relação com as variáveis resposta, mas uma dependência considerável entre elas.

### 4.5.5 – Máquina de Vetor de Suporte - Classificador

A máquina de vetor de suporte (*support vector machine* - SVM) para classificação (*support vector classifier* - SVC), como abordado anteriormente, é um modelo baseado na teoria de aprendizado estatístico que traça um hiperplano – a melhor reta possível – a para separar os objetos das classes (FACELI et al., 2011, p. 126).

No modelo desenvolvido, o SVC com kernel linear teve uma a melhor acurácia de 67% e média de 73% para validação cruzada (10 novas divisões). Uma análise preliminar indica que esse bom resultado pode ser devido à correlação próxima do linear entre as variáveis alvo e as variáveis que mais influenciam na decisão do algoritmo de predição da classe – posição de estreia e *streams* de estreia. Uma análise futura mais aprofundada deste modelo pode investigar as causas dessa performance e verificar se essa diferença de 5% entre a validação cruzada e a acurácia de teste indicam que o modelo é tendencioso.

Apesar disso, no teste do SVC não-linear utilizando o kernel RBF (*Radial Basis Function*), usado para problemas não linearmente separáveis, o resultado também foi satisfatório (62% com média de validação cruzada de também 62%) dada a grande quantidade, a complexidade de variáveis do *dataset* e a pouca correlação linear entre quase todas as variáveis. Por conta da nenhuma diferença entre a acurácia de teste e a validação cruzada, esse modelo de SVM foi considerado o melhor a ser utilizado neste caso.

### 4.5.6 – Percpetron Multicamadas (Multilayer Perceptron – MLP)

Para a construção da rede neural *perceptron* multicamadas foram utilizadas duas

ferramentas disponíveis na biblioteca *scikit-learn*: o MLP Classifier[12], para a construção da rede neural; e uma importante ferramenta de avaliação de parâmetros, o Grid Search CV[13].

O Grid Search CV implementa a aplicação e a avaliação do modelo por meio de testes usando parâmetros previamente estabelecidos. Antes de aplicar o modelo é criado um dicionário de parâmetros que recebem vários valores, e em seguida o Grid Search CV testa a aplicação do modelo utilizando todos os valores do dicionário e instancia o modelo com a combinação de valores para os parâmetros que trazem a melhor performance de teste com a menor perda.

O modelo foi testado utilizando o seguinte dicionário de parâmetros:

```
params =  {
   'hidden_layer_sizes': [3,
   2, 1],
   'activation': ['tanh', 'relu',
   'logistic'], 'solver': ['lbfgs'],
}
```

O Grid Search CV fez então a procura pela melhor combinação que trazia o melhor resultado ao modelo. Aos parâmetros que não tiveram valores especificados no dicionário, o algoritmo utilizou as configurações padrão. O MLP Classifier utiliza a função Softmax como função de saída, o que permite o suporte à classificação multiclasse - nesse caso, a previsão entre 3 classes de *tier*.

O melhor resultado observado foi uma média de validação cruzada (para 10 subdivisões de treino e teste) de 73% de acurácia. Os melhores parâmetros escolhidos foram: dois neurônios, ativação logística - que usa a função sigmoide para transformar o sinal de entrada da rede neural - e o otimizador de pesos lbfgs, indicado pela biblioteca do *scikit-learn* como o melhor a ser utilizado em *dataset*s menores. Quando testada a predição do modelo utilizando o dataset de teste, a acurácia foi de 70%, indicando um bom desempenho da rede neural. As outras combinações retornaram acurácias médias parecidas.

[12] A documentação do MLPClassifier disponível em https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html. Acesso em 29/10/2021

[13] A documentação do Grid Search CV está disponível em https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. Acesso em 29/10/2021.

**Figura 4.1 -** Relação de resultados utilizando MLP Classifier

```
Melhores parâmetros encontrados: {'activation': 'logistic', 'hidden_layer_sizes': 2, 'solver': 'lbfgs'}, melhor score = 0.734466800804829

0.703 (+/-0.099) for {'activation': 'tanh', 'hidden_layer_sizes': 3, 'solver': 'lbfgs'}
0.730 (+/-0.076) for {'activation': 'tanh', 'hidden_layer_sizes': 2, 'solver': 'lbfgs'}
0.701 (+/-0.115) for {'activation': 'tanh', 'hidden_layer_sizes': 1, 'solver': 'lbfgs'}
0.716 (+/-0.079) for {'activation': 'relu', 'hidden_layer_sizes': 3, 'solver': 'lbfgs'}
0.722 (+/-0.082) for {'activation': 'relu', 'hidden_layer_sizes': 2, 'solver': 'lbfgs'}
0.675 (+/-0.106) for {'activation': 'relu', 'hidden_layer_sizes': 1, 'solver': 'lbfgs'}
0.729 (+/-0.112) for {'activation': 'logistic', 'hidden_layer_sizes': 3, 'solver': 'lbfgs'}
0.734 (+/-0.093) for {'activation': 'logistic', 'hidden_layer_sizes': 2, 'solver': 'lbfgs'}
0.698 (+/-0.110) for {'activation': 'logistic', 'hidden_layer_sizes': 1, 'solver': 'lbfgs'}

Resultados na predição de teste:
              precision    recall  f1-score   support

           0       0.73      0.79      0.76       126
           1       0.67      0.52      0.58       100
           2       0.69      0.78      0.73        78

    accuracy                           0.70       304
   macro avg       0.69      0.70      0.69       304
weighted avg       0.70      0.70      0.69       304
```

Fonte: Elaboração da Autora

### 4.5.7 Classificador por Regressão Logística Binária

O último modelo testado foi o classificador probabilístico de regressão logística binária. Para a aplicação desse modelo foi necessária uma nova classificação de *tier* para as faixas do *dataset*, onde foram consideradas como *tier* 1 as faixas que alcançaram a marca de pelo menos 50 milhões de *streams* acumulados, ou estiveram no *chart* por pelo menos 30 dias ou alcançaram as 100 primeiras posições. As demais receberam 0 para o valor de *tier*, criando uma classificação binária que, simplificando, significa se tiveram ou não sucesso. As demais etapas de preparação para a aplicação do modelo seguiram os mesmos passos abordados no capítulo anterior.

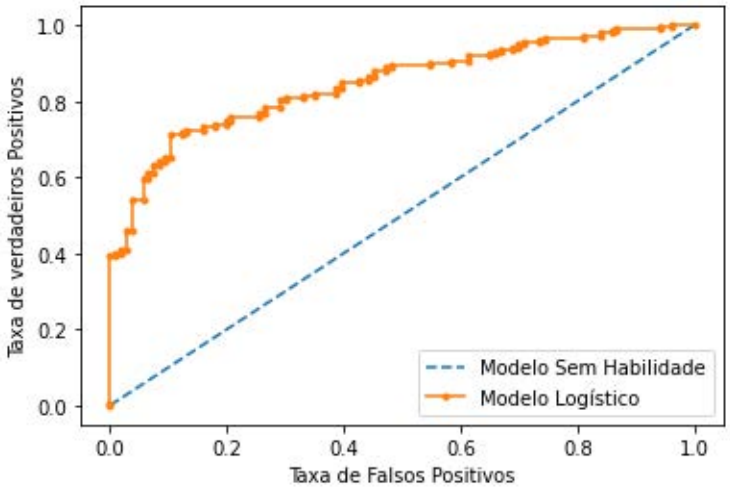
O algoritmo de regressão logística utilizou valores padrão para todos os parâmetros indicados na documentação do modelo[14], com exceção do otimizador, que foi configurado como linear - apontado pelo *scikit-learn* como o adequado para *dataset*s menores.

O resultado foi uma acurácia de 76% no teste de classificação. Também foi medida a acurácia nula – cálculo de acurácia de um modelo "burro", que prevê sempre 0 – , e o resultado foi de 65% de acerto, o que era esperado devido ao tamanho reduzido do *dataset* e todas as limitações envolvidas neste trabalho.

[14] A documentação do scikit-learn está disponível em https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. Acesso em 29/10/2021.

A figura mostra a *ROC Curve* do modelo, que compara as predições de probabilidade de falsos eventos positivos e verdadeiros eventos positivos. A linha diagonal indica uma classificação aleatória, e a curva na área superior esquerda demonstra que o modelo consegue, a cada ponto de corte, aumentar a probabilidade de acertar as faixas com *tier* 1 enquanto controla a taxa de falsos positivos (quando uma faixa *tier* 0 é classificada como 1).

**Gráfico 4.13** – ROC Curve do modelo de Regressão Logística Binária




Fonte: Elaboração da autora

Outros métodos de acurácia envolvendo a ROC Curve foram utilizados, como a ROC AUC (*Area Under ROC Curve*), que indicou que esse modelo probabilístico alcançou 84% de chance de prever corretamente as classes. Esse valor tende ao *overfitting* devido também ao tamanho reduzido do *dataset* e isso foi observado na validação cruzada de 10 subdivisões de treino e teste, que apresentou valores com até aproximadamente 20% de diferença de acurácia.

Uma análise prévia mostra que os resultados obtidos pelos modelos preditivos de classificação abordados funcionam relativamente, tanto os baseados em decisão (árvore de decisão e floresta aleatória), quanto o SVC e a rede neural MLP.

A regressão logística, apesar de aparentar uma melhor performance, tende ao *overfitting* na medida em que apresentou grande variação de resultados (20%) quando comparadas as acurácias de validação cruzada. O desempenho fraco do Naive Bayes também

sugere que as classificações probabilísticas talvez não sejam as mais aplicáveis neste caso de classificação, dada a distribuição assimétrica dos dados e uma dependência das variáveis.

O KNN, baseado em procura, também apresentou um desempenho abaixo do ideal, dado que o algoritmo assume que elementos similares assumem uma proximidade entre eles. No caso do conjunto de dados utilizado, uma análise preliminar sugere que as *audio features* podem ser responsáveis por esse baixo desempenho devido a músicas com essas características técnicas similares não terem o mesmo desempenho, como foi visto no gráfico de correlação.

Em todos os casos, os modelos podem apresentar tendências devido ao baixo volume de dados treinados. Portanto, são necessários ajustes de parâmetros nos modelos, novas preparações e análises mais aprofundadas para a melhoria do estudo e a previsão mais assertiva dos resultados, o que será abordado no próximo capítulo.

# Capítulo 5

# Conclusão e Trabalhos Futuros

Por meio da análise do Top 200 do *Spotify* foi possível extrair informações sobre as músicas mais bem sucedidas na plataforma em 2020 e gerar outras variáveis capazes de medir o desempenho e elaborar modelos de previsão de sucesso com base em 1012 faixas de 284 artistas diferentes, que acumularam o total de 12 bilhões *streams*.

Os principais obstáculos identificados foram a indisponibilidade de recursos tecnológicos, a restrição de algumas informações às gravadoras e distribuidoras sobre os desempenhos de seus artistas e produtos, e tempo hábil para um melhor detalhamento e tratamento do *dataset*. No entanto, apesar das adversidades desafiadoras, o trabalho cumpre o propósito de ser uma sugestão prévia de utilização de ferramentas de inteligência de mercado, *big data* e *machine learning* para negócios tão imprevisíveis como os do entretenimento.

Em um primeiro momento, a suposição era de que as *audio features* apresentassem alguma boa correlação com as variáveis que apontam o sucesso da faixa. Essas mesmas *features* foram utilizadas no artigo de Middlebrook e Sheik (2019) para prever *hits* da *Billboard* e obtiveram um bom resultado de predição com a floresta aleatória. No entanto, neste trabalho, para essas variáveis não foram observadas correlações lineares e importância significativas para comprovar essa suposição, provavelmente devido ao baixo volume de dados. Como também visto na introdução deste trabalho, o trabalho anterior de Pastore, Teixeira e Rezende (2020) não obteve grande sucesso em prever a quantidade de *streams* acumulados por meio de regressão linear múltipla.

Nesse sentido, a alternativa utilizada para prever o potencial de sucesso das faixas foi fazer uma classificação supervisionada para dividir essas músicas em grupos (aqui chamados de *tier*s) conforme o seu desempenho guiado pela distribuição das variáveis resposta e do conhecimento empírico do mercado. Essa classificação prévia tem um caráter arbitrário e não seria possível em um real cenário em que não se teria conhecimento sobre os sobre a distribuição futura (ao final do ano, no caso deste trabalho)

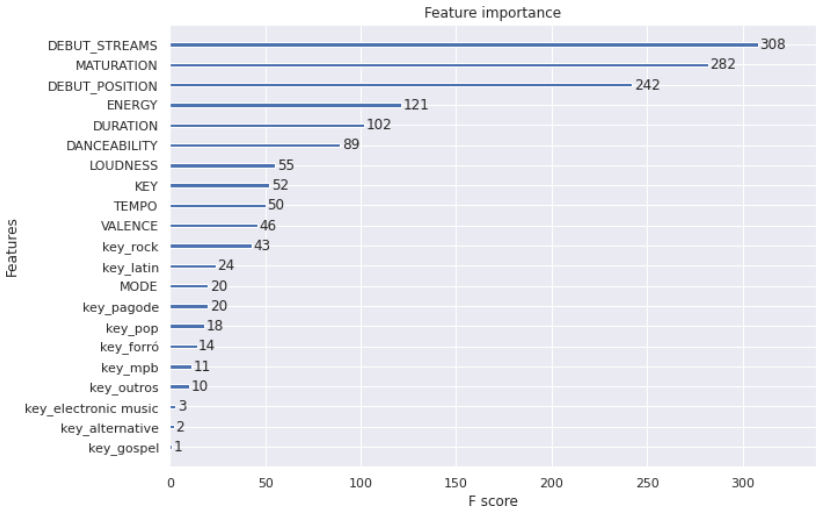
das métricas de resposta para gerar essa categorização por *tier*.

Também há a possibilidade de utilizar um volume maior de dados históricos e aplicar pesos para a classificação de *tier* conforme a ponderação para as métricas de sucesso de um ano e de outro. Por exemplo, um grande sucesso de 2015 provavelmente teve um volume de *streams* menor do que um de 2020, o que não significa necessariamente que o primeiro foi menos bem sucedido do que o segundo, e sim que a diferença de *streams* tem influência do avanço da tecnologia e a crescente disseminação dos serviços de *streaming* dos últimos anos.

Essas aplicações também podem ser utilizadas para analisar dados de relatórios comerciais, como os do Pró Música, que informam semanalmente às gravadoras as 5 mil faixas mais ouvidas no país nas plataformas de *streaming*. Empresas com acesso a bancos de dados sobre música – que geralmente são pagos para acesso ilimitado, como é o caso do *Chartmetric* –, com equipes treinadas e tecnologia mais avançada podem enriquecer o *dataset* adicionando mais músicas e variáveis, inserindo nos modelos de *machine learning* dados de treino mais precisos que melhorem o aprendizado e, por consequência, a predição. O tratamento de gêneros musicais também é uma questão a ser melhorada, já que por questões de tempo reduzido não foi possível elaborar um sistema que agrupasse melhor as faixas.

Ainda com as limitações, neste trabalho foi observada a grande influência das variáveis de *streams* de estreia, posição de estreia e maturação para a predição de classificação de *tier*. A importância dessas variáveis mostra que a performance inicial da faixa interfere positivamente no seu sucesso final e é mais considerada pelos modelos para predizer o sucesso das músicas no *chart*. As *audio features* apresentaram um grau de importância bastante reduzido, e as classificações de gênero não apresentaram importância significativa.

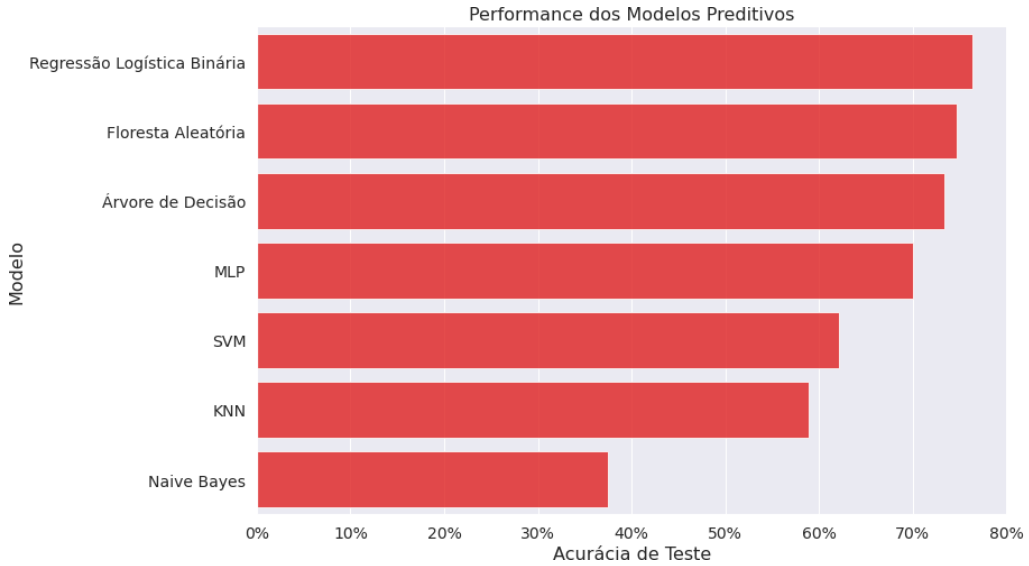
**Gráfico 5.1 -** Grau de importância das variáveis nos modelos preditivos




Fonte: Elaborado pela Autora

Os modelos de aprendizado de máquina para classificação por *tier* apresentaram boa performance, apesar de serem utilizados como testes iniciais de predição e uma estudo geral sobre como cada modelo funciona para prever sucessos com base nos dados e variáveis disponíveis de forma pública e acessível, tendo potencial de ser otimizado com mais variáveis, maior volume e melhor tratamento de dados e maior capacidade de processamento e análise.

**Gráfico 5.2** - Performance dos modelos preditivos




Fonte: Elaboração da Autora

Como principais objetivos para trabalhos futuros ficam então indicados:

1) Utilização dos modelos para um volume maior de dados: com acesso a um volume maior de dados de performance – que inclua faixas que não estiveram no *chart*, por exemplo –, é possível elaborar um modelo com desempenho mais assertivo quanto à classificação de *tier* para novas faixas. Para isso é necessário: o estudo do uso de uma pré-classificação com fronteiras de classes mais bem definidas para os *tiers* e a geração de novas classes intermediárias, o que em tese melhoraria o sistema de previsão; e a coleta de novas variáveis utilizando outros bancos de dados sobre as músicas também pode ajudar a especificar e definir melhor os *clusters* de faixas, também ajudando na eficácia da diferenciação entre os *tiers*.
2) Pesquisas mais aprofundadas sobre os modelos de *machine learning* de classificação, que incluam a realização de ajustes com diferentes valores para os parâmetros dos modelos preditivos utilizados e testes com outros que não foram abordados neste trabalho, como por exemplo o LVQ (*Learning Vector Quantization*) e o uso do *Bagging e Boosting* e para a melhoria de aprendizagem.

O LVQ é basicamente um grupo de algoritmos que funcionam de forma semelhante ao KNN (reconhecimento de padrões), mas pode ser usado neste trabalho para um conjunto maior de dados com mais agilidade. Isso porque no sistema LVQ as classes são descritas por um pequeno número de protótipos que são colocados em zonas e fazem a aproximação de fronteiras do espaço de variáveis com base na regra do vizinho mais próximo. (KOHONEN et al., 1996, p. 1). Por não precisar considerar todo o conjunto de dados, a necessidade de recursos computacionais é reduzida.

O *Bagging*, por sua vez, é "um método para gerar diversas versões de modelos preditores a fim de obter um preditor agregado" (BREIMAN, 1996, p. 1). Nele, são usados n conjuntos de treinamento iguais e replicar esses conjuntos aleatoriamente, cosntruindo redes independentes utilizando re-amostragem com reposição (*bootstrap*), e em seguida essas redes são agregadas de forma a construir um melhor classificador pela maioria de votos. Já o *Boosting* funciona de forma semelhantes, mas é nele a importância do voto de cada classificador é definida a depender do desempenho de cada modelo, ou seja, não são atribuídos o mesmo peso para todos os votos dos classificadores (LANTZ, 2013, p. 343).

A aplicação dos conceitos aqui desenvolvidos pode funcionar para predição de

performances para produtos de outros segmentos (filmes, séries etc.) e fora do universo do *streaming*: em redes sociais e projeção de receita, por exemplo. Os métodos abordados e os trabalhos futuros a serem desenvolvidos se propõem a gerar informações precisas para a observação e geração de *insights* sobre o potencial de talentos, colocando negócios e artistas um passo à frente na estratégia de desenvolver e projetar o sucesso de seus produtos. E mais do que isso: existem infinitas possibilidades para desenvolver as técnicas apresentadas neste trabalho que pretende ser, para o mercado criativo, uma demonstração da promissora enotável relação entre a cultura e a inteligência artificial.

# REFERÊNCIAS


BRAGA, F.; GOMES, E. **Inteligência competitiva em tempos de *Big data*:** Analisando informações e identificando tendências em tempo real. Rio de Janeiro: Alta Books Editora, 2001.

BREIMAN, Leo. Bagging predictors. **Machine learning**, v. 24, n. 2, p. 123-140, 1996.

COSTA, J. A atualidade da discussão sobre a indústria cultural em Theodor W. Adorno.**Trans/Form/Ação,** v. 36, n. 2, p. 135-154, mai/ago 2013.

DAVENPORT, Thomas H.; PRUSAK, Laurence. **Conhecimento empresarial: como as organizações gerenciam o seu capital intelectua**l. Rio de Janeiro: Elsevier, 2003.

FACELI, et al. **Inteligência Artificial–uma abordagem de aprendizado de máquina**. Rio de Janeiro: LTC, 2011.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de Análise de Dados**: Estatística e Modelagem Multivariada com Excel®, SPSS® e Stata®. Elsevier Brasil, 2017.

GONZALEZ, Leandro. **Regressão logística e suas aplicações**. Monografia (Ciência da Computação) - Universidade Federal do Maranhão, São Luís, 2018.

GUNN, Steve R. et al. Support vector machines for classification and regression. **ISIS technical report**, v. 14, n. 1, p. 5-16, 1998.

HENNING-THURAU, T. e HOUSTON, M. **Entertainment Science:** Data Analytics and Practical Theory for Movies, Games, Books, and Music. Editora Springer, 2018.

IFPI, **Reports and Resources.** IFPI, 2021. Disponível em: <https://www.ifpi.org/resources/>. Acesso em: 25/08/2021.

KOHONEN, Teuvo et al. LVQ PAK: **The learning vector quantization program package**. Technical report, 1996.

LANTZ, Brett**. Machine learning with R: expert techniques for predictive modeling**. Packt publishing ltd, 2019.

LIPOVETSKY, G.; SERROY, J. A expansão econômica dos mundos transestéticos. *In:* **A estetização do mundo: viver na era do capitalismo artista.** São Paulo: Companhia das Letras, 2015.

MAASØ, Arnt; HAGEN, Anja Nylund. Metrics and decision-making in music streaming. **Popular Communication**, v. 18, n. 1, p. 18-31, 2020.

MIDDLEBROOK, K.; SHEIK, K. **Song Hit Prediction: Predicting Billboard Hits Using *Spotify* Data.** Department of Math & Statistics University, San Francisco, 2019.

MIDiA Research. **Global music subscriber market shares Q1 2021.** MIDiA Research, 2021. Disponível em: <https://www.midiaresearch.com/blog/global-music-subscriber-market- shares-q1-2021>. Acesso em: 25/08/2020.

MOSCHETTA, P.; VIEIRA, J. Música na era do *streaming*: curadoria e descoberta musical no *Spotify*. **Sociologias**, Porto Alegre, ano 20, n. 49, p. 258-292, set/dez 2018.

MYATT, Glenn J. **Making sense of data: a practical guide to exploratory data analysis and data mining**. John Wiley & Sons, 2007.

NYCE, C. **Predictive Analytics White Paper**. American Institute for CPCU. Malvern, PA, 2007. Disponível em:<https://www.the-digital-insurer.com/wp-content/uploads/2013/12/78- Predictive-Modeling-White-Paper.pdf/>.

PASTORE, Júlia; TEIXEIRA, Júlia; REZENDE, Mariana. **Prevendo o Consumo de Áudio Streaming no Brasil**: Uma análise sobre o sucesso comercial de novos lançamentos nos charts do *Spotify*. Monografia (Administração) - Escola Superior De Propaganda e Marketing, Rio de Janeiro, 2020.

PONTE, Caio; CAMINHA, Carlos; FURTADO, Vasco. Otimização de Florestas Aleatórias através de ponderação de folhas em árvore de regressão. In: **Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional**. SBC, 2020. p. 698-708.

RUIZ-GONZALEZ, Ruben et al. An SVM-based classifier for estimating the state of various rotating components in agro-industrial machinery with a vibration signal acquired from a single point on the machine chassis. **Sensors**, v. 14, n. 11, p. 20713-20735, 2014.

SANTINI. R. M.; SALLES, D. O impacto dos algoritmos no consumo de música: uma revisão sistemática de literatura. **Signos do Consum**o, v. 12, n. 1, p. 83-93, jan/jun, 2020.

SMITH, Michael D.; TELANG, Rahul. **Streaming, sharing, stealing: Big data and the future of entertainment**. Mit Press, 2016.

STROBL, Eric A.; TUCKER, Clive. The dynamics of chart success in the UK pre-recorded popular music industry. **Journal of Cultural Economics**, v. 24, n. 2, p. 113-134, 2000.

THOMPSON, Derek. **Hit Makers: Como Nascem as Tendências**. Rio de Janeiro: Harper Colins, 2018.

WIKSTRÖM, Patrik. **The music industry: music in the cloud**. John Wiley & Sons, 2020.

YING, Xue. An overview of overfitting and its solutions. In: **Journal of Physics: Conference Series**. IOP Publishing, 2019. p. 0220