

Universidade Federal do Paraná

Bruna Davies Wundervald

Rogério de Jesus Hultmann Filho

**Engenharia de Características Baseadas em
Cifras para a Classificação de Gêneros na
Música Brasileira**

Curitiba

2018

Bruna Davies Wundervald
Rogério de Jesus Hultmann Filho

Engenharia de Características Baseadas em Cifras para a Classificação de Gêneros na Música Brasileira

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Graduação em Estatística da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Walmes M. Zeviani

Curitiba
2018

Agradecimentos

À Deus.

Aos familiares, amigos e companheiros pelo total apoio e compreensão nos momentos de ausência.

Ao nosso orientador Professor Doutor Walmes Marques Zeviani por todos os conhecimentos a nós repassados e pela amizade.

"Se fracassar, ao menos que fracasse ousando grandes feitos, de modo que a sua postura não seja nunca a dessas almas frias e tímidas que não conhecem nem a vitória nem a derrota."

(Theodore Roosevelt)

Resumo

Música e Estatística sempre foram duas áreas do conhecimento muito distintas. Entretanto, é possível uni-las, gerando a área de *Music Information Retrieval*, que desenvolve e aplica ferramentas computacionais combinadas com teoria e conceitos musicais no intuito de ampliar a compreensão e utilidade dos dados de música. A abordagem apresentada tem como principais objetivos reunir um grande conjunto de cifras de músicas brasileiras, analisar a estrutura harmônica delas, e elencar conjuntos de características baseadas nestas cifras, com as quais seja possível realizar a classificação das músicas em gêneros musicais e obter seus respectivos graus de importância nestas previsões. A dificuldade observada em se conseguir esse tipo de dados motivou o desenvolvimento de um pacote para o software estatístico R, chamado de (*chordrds*), que torna a extração automatizada. Com os dados em mãos, as características são extraídas através de, principalmente, expressões regulares, transformando as *strings* de cada acorde em informações sobre as músicas avaliadas. Algumas características são obtidas com auxílio da API do software de *streaming* de música Spotify. Estas informações servem como recurso para a classificação das músicas em seus respectivos gêneros musicais. A classificação, por sua vez, é feita por meio de modelos de aprendizado de máquina não paramétricos, conhecidos como florestas aleatórias. O método consiste em uma combinação de diversas árvores de classificação para a previsão de classes, e seu uso é motivado pela natureza do problema, que envolve especialmente a obtenção de medidas interpretáveis de importância para as variáveis preditoras. Os principais resultados consistem na avaliação de quanto cada característica é relevante para a classificação dos gêneros estudados. As três características mais significativas para a classificação foram: a porcentagem de acordes com a sétima nota em cada música, a porcentagem de acordes menores e com sétima nota, e porcentagem de acordes menores. Desta forma, observa-se que elementos da composição harmônica são suficientes para prever os gêneros das músicas presentes na amostra, desde que suas características sejam devidamente extraídas.

Palavras-chave: Engenharia de Características, Florestas Aleatórias, Web Scraping, Música.

Sumário

1	INTRODUÇÃO	6
1.1	Música e Harmonia	6
1.2	<i>Music Information Retrieval</i>	7
1.3	Formatos de Dados Musicais	7
1.4	Gêneros Musicais	8
1.5	Objetivos	9
2	REVISÃO DE LITERATURA	10
2.1	<i>Music Information Retrieval</i>	10
2.2	Engenharia de Características	11
2.3	Classificação de gêneros musicais	12
3	MATERIAL E MÉTODOS	14
3.1	Web Scraping	14
3.1.1	Artistas	15
3.1.2	Dados	15
3.1.3	O círculo das quintas	16
3.2	Engenharia de Características	17
3.3	Análise Exploratória	22
3.4	Árvores de Classificação	22
3.5	Florestas Aleatórias	25
3.5.1	Importância de Variáveis	26
4	RESULTADOS E DISCUSSÃO	27
4.1	O pacote <i>chordrds</i>	27
4.2	Análise Exploratória	27
4.2.1	Acordes	27
4.2.2	Variáveis Extraídas	34
4.3	Modelagem	38
4.3.1	Conjuntos de Variáveis	38
4.3.2	Florestas Aleatórias	39
5	CONSIDERAÇÕES FINAIS	47
	REFERÊNCIAS	48

1 Introdução

1.1 Música e Harmonia

A música é um elemento cultural intrínseco em nossa sociedade, sendo difícil encontrar quem não goste de escutá-la e ou mesmo criá-la. De acordo com SEEGER 2008, uma definição geral de música deve incluir tanto sons quanto seres humanos. Música é um sistema de comunicação que envolve sons estruturados produzidos por membros de uma comunidade que se comunicam com outros membros. Em sua definição mais simples, é uma combinação de harmonias, melodias e ritmos.

Para entendermos o que é harmonia, é preciso primeiro definir alguns elementos que a compõem, como notas, acordes e melodia. Notas musicais são sinais gráficos e sonoros criados para representar as variações da altura do som musical. A partir delas são criados os acordes, uma combinação de três ou mais notas que soam simultaneamente. Tais combinações ou acordes são criadas sempre segundo alguma regra. Por exemplo, uma tríade maior, ou seja, um acorde de três notas, é composta por uma nota **raíz**, sua **terça** (quatro semitons mais alta que a raíz) e sua **quinta** (três semitons mais alta que a terça), sendo que um semitom é a menor diferença entre dois sons na escala melódica ocidental, que é a mais utilizada.

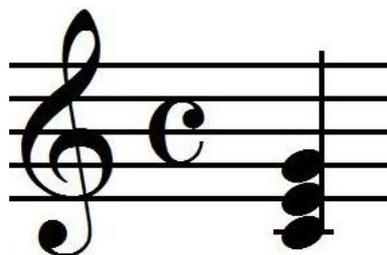


Figura 1 – Representação do acorde de Dó maior em uma partitura ou pentagrama. O símbolo a esquerda é uma clave de Sol. O sinal semelhante à letra C representa tempo de andamento do compasso. Por fim, o acorde de Dó maior é representado pelos três pontos fechados que são suas 3 notas constituintes: dó (raíz, a primeira de baixo para cima), mi (terça) e sol (quinta).

A melodia é composta pela sucessão de notas, que podem diferir em duração, altura e intensidade. Mas as notas não acontecem de forma meramente casual: elas estão vinculadas ao acorde que acontece em cada momento, que por sua vez está conectado à estrutura harmônica em progresso. Além disso, tem-se subsídios para afirmar que músicas são caracterizadas principalmente por sua sequência melódica. A melodia é a parte da música que conseguimos 'cantar' ou reproduzir com assovios.

A harmonia é a ciência dos sons que são expressos simultânea e sequencialmente. Não se deve confundir harmonia com polifonia, que representa a sobreposição de uma ou mais melodias. Ela é o que justifica a sequência de acordes nas músicas, que acontecem, geralmente, seguindo alguma estrutura harmônica. Isso porque cada acorde pode ter uma função diferente na música, e a harmonia trata de organizá-los de alguma forma, produzindo no ouvinte determinadas sensações, por exemplo, algumas progressões podem remeter à suspense, outras à triunfo. Essas sensações são bastante exploradas em trilhas sonoras de filmes, novelas, espetáculos em geral, onde as melodias são usadas para intensificar os sentimentos do espectador.

1.2 *Music Information Retrieval*

Uma das áreas que se ocupa da música é a de *Music Information Retrieval*, que desenvolve e aplica ferramentas computacionais combinadas com teoria e conceitos musicais, no intuito de ampliar a compreensão e utilidade dos dados de música. Esta área é responsável pela criação de diferentes mecanismos de busca e recomendação para música, composições automáticas geradas por aprendizado de máquina, descrição de características musicais utilizando sinais de áudio ou representação simbólica de sons, e assim por diante.

Transcrições automáticas de áudios de músicas para partituras ou arquivos MIDI também são objeto de estudos das técnicas de *MIR*, através de detecção de ritmo, instrumentos, notas, etc. Resultados como estes são de grande valia para desde estudantes e entusiastas de música, até para quem depende dela profissionalmente, como compositores, produtores e músicos.

1.3 Formatos de Dados Musicais

Sons em geral podem ser representados de diversas formas, mas serão trabalhadas apenas as formas que concernem a música em si. Entre os formatos nos quais uma música pode ser representada, os mais comuns são: MIDI, cifras, partituras, e, no caso de músicas letradas, as próprias letras.

O MIDI (*Musical Instrument Digital Interface*) descreve protocolos de comunicação, interface digital e conectores elétricos, permitindo que uma variedade de instrumentos musicais eletrônicos, computadores e *softwares* de música se conectem. Dados em MIDI são extremamente ricos, porque contém informações sobre a notação, notas e suas alturas, velocidade, volume, e outras especificações técnicas. Porém, mesmo sabendo quais notas são tocadas em cada tempo a partir de um arquivo em MIDI, normalmente não se conhece o acorde em questão. As informações sobre os acordes de uma música podem ser trazidos pelas cifras, que basicamente são representação, em formato de texto, da sequência de acordes que estruturam cada canção. As cifras geralmente acompanham a letra das canções

indicando em qual parte da letra é feita a mudança de acorde. Uma limitação desta representação é que não se sabe ao certo em qual tempo exato os acordes estão ocorrendo, o que exige do instrumentista conhecimento da canção.

Partitura, por sua vez, é a representação padronizada mundial de música. Ela possui seus próprios símbolos, sendo que cada um representa uma nota, altura, tempo, pausa, entre outros, assim como todo sistema de escrita. É como se fosse uma linguagem específica para música, compreensível por qualquer pessoa que tenha conhecimento sobre seus significados. Em um ambiente digital, utilizando tecnologias como o MIDI, é possível traduzir partituras para formatos legíveis pelo computador e instrumentos eletrônicos.

Figura 2 – Trecho de uma partitura do Hino Nacional Brasileiro com as seguintes representações utilizadas: claves (de sol e de fá), pausas, tremolos, ligaduras e sinais de intensidade sonora. Fonte: (SILVA; ESTRADA, 2004)

Em um senso mais amplo, a música possui diversas formas de representação. Cada forma tem suas especificidades, contendo diferentes níveis de informação sobre a mesma música. Assim, cada formato desses exige um certo tratamento e possibilita (ou não) a aplicação de diferentes técnicas visando a análise dos dados.

1.4 Gêneros Musicais

Gêneros musicais são as categorias que contêm peças musicais que compartilham elementos em comum. Estes elementos podem ser: função, estrutura, contextualização, instrumentação, mídia utilizada e outros. Pode-se dizer que cada estilo musical tem uma assinatura harmônica típica que o representa, mas que também existem variações e afastamentos dela. Todavia, é muito difícil definir o que exatamente qualifica um gênero. Por esse motivo, dependendo do critério, algumas músicas podem ser encaixar em vários gêneros, até mesmo por terem referências em vários estilos musicais. Alguns gêneros são até considerados vagos, como o 'post-rock', que foi criado pelo crítico musical REYNOLDS.

Neste trabalho, foram considerados os gêneros musicais mais comuns na música brasileira. Por eles serem de difícil definição exata, como dito anteriormente, esse tópico não será aprofundado. Os gêneros considerados são: Reggae, Pop, Forró, Bossa Nova, Sertanejo, MPB, Rock e Samba.

1.5 Objetivos

Objetivos Específicos

Os objetivos principais deste trabalho são definidos em seis partes:

- Propor um método eficaz de extração de cifras musicais da *web* e covariáveis relacionadas a elas;
- Disponibilizar um pacote em R que implemente o método proposto;
- Utilizar o método para extração de dados;
- Criar um acervo de canções cifradas dos estilos musicais pretendidos;
- Realizar uma breve análise da estrutura harmônica de composições brasileiras do acervo;
- Avaliar, através de modelagem, quais são as características mais relevantes para a classificação das músicas extraídas em seus respectivos gêneros musicais.

Para o primeiro ponto ser cumprido, utiliza-se técnicas de *webscraping* e engenharia de covariáveis (*feature engineering*). Em seguida, a forma como isso foi realizado é implementada em um pacote para o R e disponibilizada à comunidade, de forma que as análises aqui apresentadas sejam integralmente reproduzíveis. No terceiro ponto, realiza-se a extração de dados, com auxílio do pacote proposto. Enfim, avalia-se quais covariáveis são necessárias para descrever de forma adequada a classificação de gênero musical que cada música recebe. Vale lembrar que esta classificação, de forma geral, pode levar em consideração não apenas a estrutura harmônica das músicas, mas diversas outras particularidades delas como número de acordes distintos, proporção de acordes maiores, menores, origem, data de lançamento das músicas, entre outros aspectos.

2 Revisão de Literatura

2.1 *Music Information Retrieval*

Quando fala-se sobre acordes, deve-se considerar especialmente as progressões deles. Fazer inferência sobre progressões de acordes é essencial para modelar relações de dependência entre acordes.

A conexão entre música e estatística não é trivial. Este é um campo de estudo mais comum nas áreas de ciência da computação, mas muitos estudos fazem uso de ferramentas e modelos estatísticos para a análise de dados musicas. Pode-se encontrar um exemplo em (RAPHAEL; STODDARD, 2003), que utilizaram modelos ocultos de Markov para análises harmônicas utilizando notas e ritmo como base. A propriedade Markoviana, nesse caso, é apenas uma suposição dos autores dado a sua simplicidade. Com isso, eles fazem a suposição de que a tônica do acorde em um período de tempo tem uma grande probabilidade de ser a mesma que era em um período anterior. O modelo oculto de Markov para esse caso assume que toda vez que um estado é visitado (rótulo harmônico), os novos dados são obtidos retirando um conjunto de notas de uma distribuição característica desse rótulo harmônico.

Motivados em analisar relações musicais, Paiement, Eck e Bengio (2005) capturam estruturas harmônicas usando modelos grafos e dados MIDI, defendendo a vantagem do método em ser flexível, podendo ser usado para análise ou uma ferramenta de geração de progressões harmônicas. A ideia principal é converter distâncias Euclidianas entre as representações de acordes em probabilidades de troca de acordes, onde esses podem ser representados como eventos discretos individuais. Modelos grafos, de acordo com Lauritzen (1996), são úteis para descrever distribuições de probabilidades, onde os grafos são usados como uma fatoração das probabilidades conjuntas. Nessa representação, as variáveis aleatórias estão associadas com os vértices. Por exemplo: se dois vértices não estão ligados por uma aresta, são consideradas variáveis aleatórias independentes. Uma característica do modelo grafo é que as variáveis podem ter sido observadas na base de dados ou estarem ocultas, e para solucionar isso utiliza-se o algoritmo EM (*Expectation-Maximization*). O primeiro passo do algoritmo consiste em calcular o valor esperado da variável não observada dados os parâmetros do modelo e uma observação aleatória da base de dados. Assim, valores dos parâmetros são atualizados para maximizar a verossimilhança. Depois, Paiement, Bengio e Eck (2009) utilizam derivações dos modelos ocultos de Markov de entrada e saída (ou IOHMMs - *Input/Output Hidden Markov Models*) para realizar predições melódicas. Quatro diferentes representações abstratas de acordes foram consideradas como técnicas de suavização para uma melhor generalização: considerando todos os possíveis

acordes; considerando apenas a tônica e a raiz (se é maior ou menor); olhando apenas se o acorde é maior, menor ou com sétima, independente de sua tônica e a quarta opção é considerar as tônicas e as três modulações anteriores (maior, menor e sétima). Os IOHMMs, segundo Chiappa e Bengio (2003), são uma extensão dos modelos ocultos de Markov para o caso no qual a distribuição das variáveis de saída e dos estados estão associadas a um conjunto de variáveis de entrada. Para fins de classificação, as variáveis de entrada estão associadas à sequências observadas e as variáveis de saída às classes.

Nos artigos citados, um número pequeno de observações e um gênero específico foram selecionados. Neste estudo, não é preciso lidar com tais limitações, uma vez que metodologia aplicada possibilita o uso de quantas observações e gêneros se queira. Além disso, quanto mais observações disponíveis, maior é o número de técnicas aplicáveis.

2.2 Engenharia de Características

O processo de extração das características mais relevantes e criação de novas variáveis de um banco de dados inicial chama-se *feature extraction*, ou extração de covariáveis. Essa técnica é usada para levantar informações relevantes, reduzir a dimensão dos dados, aumentar a acurácia da inferência ou mesmo para o entendimento do comportamento dos dados. Segundo Nargesian et al. (2017), a engenharia de covariáveis é a principal tarefa na preparação de dados para aprendizado de máquina. É a prática de construção de recursos adequados a partir de determinados recursos que levam a um melhor desempenho preditivo. Geralmente é conduzida por um cientista de dados, e depende de seu conhecimento, tentativa e erro e avaliação do modelo. Vários métodos de aprendizado de máquina executam a extração de recursos ou o aprendizado indireto. Embora eles não trabalhem explicitamente com transformações de dados de entrada, podem gerar novos recursos como meio de resolver outro problema. As técnicas de extração de características reduzem o comprimento do vetor de características por transformação ou projeção de todos os recursos do vetor de características de menor dimensão. Isto é, mapeia-se os dados de alta dimensão no espaço de baixa dimensão.

Guyon e Elisseeff (2006) concluem que a expertise humana, muitas vezes necessária para converter o dado bruto em um conjunto de recursos úteis, pode ser complementada por um grupo de métodos automáticos de construção de características. Em algumas abordagens, a construção destas características está integrada ao processo de modelagem. Logo, muitas são as aplicações da engenharia de características, estando dentre elas o processamento de imagem, processamento de texto, reconhecimento de voz, diagnóstico de falhas, entre outros.

Na música também pode-se ter vários tipos de informações, por exemplo: quantidade de acordes diferentes, quantidade de acordes maiores, menores, com sétima, com baixo em outra nota que não a tônica da tríade, números de transições do primeiro para o

quinto grau, etc. Na busca de descobrir padrões em sequências melódicas, Rolland (1999) apresenta um algoritmo de extração de padrões chamado FLEXPAT (“*FLExible Extraction of Patterns*”), em português, “Extração Flexível de Padrões”, utilizando principalmente dados MIDI e representações de partituras. O algoritmo funciona em duas partes: comparação de pares de passagens de melodia e depois a categorização. Após listar todas as características possíveis, dentre elas tempo, tônica, momentos de tensão, crescimento ou decrescimento da intensidade do timbre, notas ligadas ou espaçadas, o algoritmo é colocado em prática. Na primeira fase são identificados todos os pares de passagem de maneira computacionalmente otimizada e gera um gráfico de similaridade. Na segunda etapa, extrai os padrões atuais do gráfico de similaridade.

Para Scaringella, Zoia e Mlynek (2006), os gêneros musicais são os principais fatores usados por comerciantes de música e bibliotecários ao organizar suas coleções de músicas. Isso evidencia o interesse em agrupar músicas por características semelhantes. A definição de gênero musical ainda continua algo muito pobre, sendo um desafio para a engenharia de características. Existem diversas conjecturas diferentes sobre os gêneros da música brasileira e seus padrões. Por exemplo, sertanejo usa sempre a mesma sequência de variações da tônica? O número de acordes das músicas foram diminuindo ao longo do tempo? A MPB é o gênero que mais possui acordes diferentes? Essas e outras perguntas podem ter respostas bem embasadas utilizando as variáveis corretas.

2.3 Classificação de gêneros musicais

Em um trabalho anterior, Correa, Costa e Saito (2010) utilizaram 70 músicas em MIDI de 4 diferentes gêneros musicais: blues, bossa nova, reggae e rock para propor um método de classificação automática de gênero. Os autores acreditam que esse processo baseia-se principalmente em três pilares: uma boa representação da música, a extração de características e por fim a classificação. A característica de interesse foi o comportamento rítmico e, usando uma abordagem Markoviana, matrizes de probabilidades de transição são calculadas a partir dos ritmos de cada música. A estimação do gênero leva em consideração as medições obtidas dos dígrafos construídos a partir das matrizes de transição. Dígrafos, nesse caso, são gráficos com as bordas orientadas para representar os ritmos musicais, algumas das medidas de interesse foram extraídas desses gráficos como: o grau total do vértice, o coeficiente de robustez e o de classificação. Para uma melhor extração de características e eliminação de redundâncias foi utilizada a análise de componentes principais, essa abordagem utiliza autovalores e autovetores da matriz de covariância para aplicar transformações geométricas ao espaço original, criando novas características ortogonais não correlacionadas. A classificação supervisionada é realizada por um classificador Bayesiano, que recebe como priori as propriedades das matrizes de transição, bem como as medidas obtidas diretamente pelos dígrafos, produzindo como

posteriori o gênero com maior probabilidade.

Ainda sobre métodos de classificação em gêneros musicais, Basili, Serafini e Stellato (2004) utilizaram vários algoritmos de *machine learning* com o objetivo de encontrar o melhor classificador:

- Naive Bayes: permite a realização de análises estatísticas a partir do conjunto de dados de treinamento. Produz estimadores de máxima verossimilhança e maximiza as probabilidades condicionais nos valores observados das características como um critério de decisão.
- VFI: (*Voting Feature Intervals*), esse algoritmo trabalha com discretização de atributos, construindo primeiramente intervalos para cada um deles e em seguida utiliza uma técnica de votação para avaliar o aprendizado do modelo.
- J48: constrói uma árvore de decisão, onde os nós representam regras de discriminação entre as características selecionadas.
- PART: semelhante ao J48, onde o algoritmo seleciona o melhor ramo em cada iteração, construindo assim uma árvore de decisão parcialmente otimizada.
- NNge: um sistema baseado em regras que constrói uma espécie de modelo hipergeométrico com regras do tipo “se” e “então”.
- JRip: um classificador baseado em regras que implementa um aprendizado de regras. O modelo de aprendizado é desenvolvido por iteração em um subconjunto de treinamento e pela otimização da estrutura para minimizar a taxa de erros.

Após a avaliação de todos os modelos citados, a conclusão é de que o mais promissor é o *Naive Bayes*, o que pode ser explicado pela heterogeneidade das características observadas no conjunto de músicas, pois cada música tem suas próprias características. Modelos baseados em regras ou em árvores tendem a agrupar características realmente diferentes para produzir os classificadores, acabando por impor uma generalização das covariáveis. Os autores utilizaram 300 músicas em MIDI, porém não especificaram as variáveis utilizadas.

3 Material e Métodos

3.1 Web Scraping

Os dados utilizados neste trabalho foram obtidos através da técnica conhecida como *web scraping*. O método, vindo da engenharia de software, baseia-se em recolher informações de *websites* de maneira automática. Isso pode ser feito de alguns modos diferentes, dependendo do tipo de dado coletado, que comumente estão em arquivos *XML*, *HTML* e *JSON*.

Entre os dois primeiros não existe muita diferença: *XML* é uma linguagem de marcação parecida com o *HTML*, mas o primeiro tem o objetivo de armazenar dados, enquanto o segundo é usado na construção de páginas *web*. Já o *JSON* é uma maneira de representar objetos complexos, não sendo tão flexível quanto as duas anteriores.

Os dados coletados são provenientes do *website* Cifraclub (<https://www.cifraclub.com.br/>), que contém um acervo de canções cifradas para violão, teclado, cavaquinho e guitarra. O acesso ao site é livre ao público geral, sendo que grande maioria das canções cifradas nele contidas são contribuições dos próprios usuários. Em outras palavras: os acordes e letras do site, em sua maioria, são fornecidos pelas pessoas que o utilizam. Isso confere ao dados disponíveis uma grande variabilidade, tanto em termos de notação quanto em qualidade, já que o mesmo acorde pode ser definido com diversas simbologias em músicas diferentes, uma vez que eles não foram inseridos pela mesma pessoa.

O método em si consiste em acessar o código fonte da página de cada canção e dele extrair os dados. Cada dado é encontrado de acordo com o elemento *HTML* no qual ele está contido. Neste trabalho, foi feita a extração dos acordes, da escala e dos nomes, de cada música e do artista.

```

      Gm7  A7  Dm7
Eu amei,   amei demais
           Bb7M           C7           A7
O que eu sofri por causa de amor ninguém sofreu
      Gm7  A7  Dm7
Eu chorei,   perdi a paz
           F           Am7
Mas o que eu sei é que ninguém nunca teve mais
           Dm7  G/D
Mais do que eu

      Gm7  C7           F7+  F7           Bb7
Capoeira me mandou dizer que já chegou

```

Figura 3 – Trecho da cifra da música Berimbau de Vinicius de Moraes. Fonte: CifraClub

3.1.1 Artistas

A seleção das fontes de dados foi feita de forma que representasse os artistas mais populares do site CifraClub. Justamente por isso, as bandas e cantores selecionados são aqueles presentes na seção de 'Mais Acessados' dos gêneros Rock, Sertanejo, MPB e Samba, dos quais foram mantidos apenas os artistas brasileiros.

3.1.2 Dados

Ao total, foi obtida através de *web scraping* uma base de dados com 483.991 linhas e 6 colunas. As 2 colunas adicionais são referentes à data de lançamento da música e sua popularidade. Estas duas informações extras são obtidas através de consultas à *API* do software de streaming de música Spotify (VER APENDICE).

Estes dados estão no formato longo. Nas linhas da coluna de acordes estão os acordes de cada música em sequência, sendo que em uma linha está um acorde. Dessa forma, uma canção ocupa um número variável de linhas, tantas quantos forem os acordes que aparecem na canção. Para exemplificação, a tabela a seguir mostra uma parte da tabela dos dados ainda não tratados.

Tabela 1 – 10 linhas da tabela de dados não tratados.

ano	artista	música	tonalidade	acorde
1966	chico Buarque	chico buarque a banda	D	D6/9
1966	chico Buarque	chico buarque a banda	D	A7
1966	chico Buarque	chico buarque a banda	D	F#m7
1966	chico Buarque	chico buarque a banda	D	B7
1966	chico Buarque	chico buarque a banda	D	E7(9)
1966	chico Buarque	chico buarque a banda	D	A7
1966	chico Buarque	chico buarque a banda	D	D6/9
1966	chico Buarque	chico buarque a banda	D	A7
1966	chico Buarque	chico buarque a banda	D	F#m7
1966	chico Buarque	chico buarque a banda	D	B7

A Tabela 1 mostra 10 linhas referentes à música "A Banda", do cantor Chico Buarque. Os acordes estão na sequência original de aparecimento na música, o que de certa forma captura a estrutura temporal de acontecimento deles. A próxima tabela nos traz um resumo geral dos dados.

Com este resumo, não agrupado por músicas, verifica-se a existência de uma maior frequência dos anos de 2015 e 2016. Além disso, nota-se quais são as músicas com mais acordes, quais são as características da distribuição da popularidade, quantos acordes diferentes existem, e as escalas e artistas mais comuns.

Ao fim do processo, disponibilizamos um pacote para o R, que faz tanto a extração de dados quanto das covariáveis.

Tabela 2 – Resumo da base original.

data		musica	
2015: 31490		legião urbana faroeste caboclo: 371	
2016: 29852		chico buarque construcao deus lhe pague: 368	
2014: 26588		jair rodrigues disparada: 344	
2017: 25276		chico buarque morena de angola: 331	
2012: 24491		jorge ben jor pais tropical spyro gyro : 331	
2013: 23416		charlie brown jr vem ser minha: 300	
(Outros): 322878		(Outros): 481946	
acorde	escala	artista	
Quantidade: 3175	C :132696	chico buarque : 26424	
Classe: caractere	G: 116118	caetano veloso : 20212	
	D : 89524	zeze di camargo e luciano: 18730	
	A : 83041	bruno e marrone : 16946	
	E : 61781	engenheiros do hawaii : 15082	
	F : 52999	fundo de quintal : 13044	
	(Outros): 81396	(Outros) : 507117	

3.1.3 O círculo das quintas

O círculo das quintas é um instrumento fundamental para o estudo de harmonia musical. Em (ALMADA, 2012), a definição proposta é

O círculo das quintas [...] é, na realidade, uma maneira de apresentar as doze notas do total cromático distanciadas entre si por intervalos de 4J (quarta justa) ou 5J (quinta justa), dependendo do sentido observado.

Aonde as quartas justas são notas que estão a 5 semitons da nota de partida, e as quintas justas estão a 7 semitons. A Figura 4 mostra a organização dos acordes no círculo dos quintas.

Em teoria musical, o círculo das quintas é útil para indicar quais são os acordes mais prováveis nas músicas. Por exemplo, se uma música está na escala de Dó (ou C), os acordes que ocorrerão com maior frequência são o Sol (G) e Fá (F). Eles são os acordes que estão, respectivamente, à direita e à esquerda do Dó no círculo das quintas, e são a 5J e 4J dele (como indicado na definição). Faz sentido esperar que os acordes de maior ocorrência, em cada música, sejam os vizinhos do acorde raiz, de acordo com o círculo. Estes vizinhos compõem o “campo harmônico” das canções. De forma complementar, quanto mais distantes os acordes estão de sua raiz, mais a canção é destoante, ou incomum, em questões de complexidade harmônica.

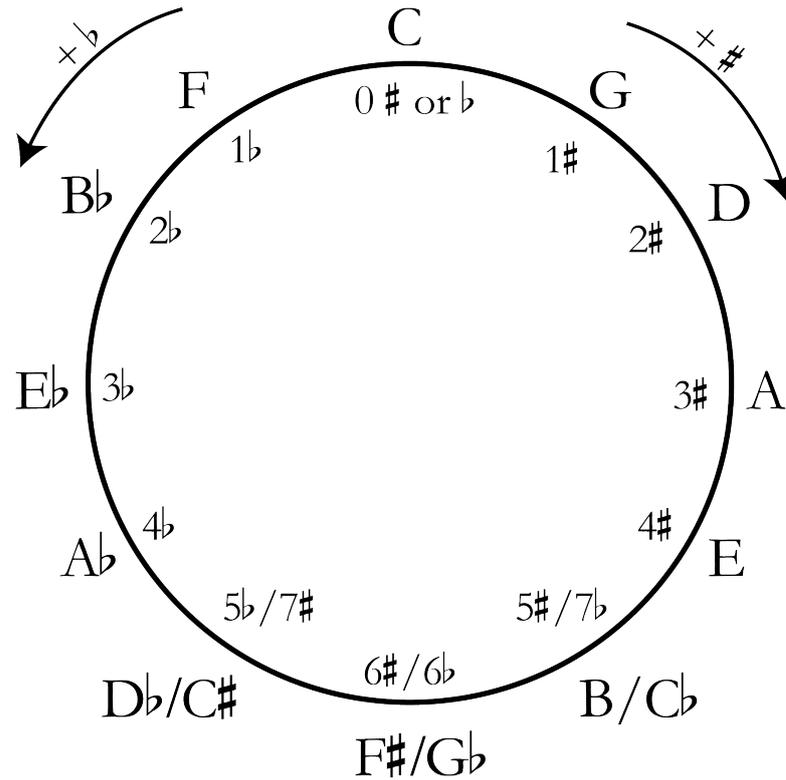


Figura 4 – Representação gráfica do círculo das quintas. Fonte: *Open Music Theory*

3.2 Engenharia de Características

Em tradução livre de um trecho do livro *Feature Engineering Made Easy* (OZDEMIR; SUSARLA, 2018)

“Engenharia de características é o processo de transformação dos dados em características que melhor representam o problema adjacente, resultando em um aperfeiçoamento do processo de aprendizagem.”

A engenharia de características representa uma parte significativa da ciência de dados. Em geral, o objetivo deste processo é obter características que sejam mais úteis para os algoritmos, de forma que eles consigam extrair padrões e obter resultados de forma mais eficiente. Com “característica”, o autor quer dizer um atributo dos dados que é relevante para o processo de aprendizado de máquina. Esta será a terminologia utilizada daqui para frente, sendo que, neste contexto, as “características” trabalhadas são as covariáveis criadas ou extraídas a partir das já existentes.

A motivação para realizar engenharia de características é avaliar qual é o conjunto de características mais importante para a classificação de gêneros musicais. Com isso, algumas questões podem ser levantadas, como: o quanto saber que um acorde possui uma nota extra é interessante, como um acorde com sétimo ou segundo grau é informativo sobre a música da qual ele faz parte, quanto saber que existe um acorde “acidente” (ou

seja, que não está presente na escala da música) na música pode dizer sobre sua estrutura harmônica, e assim por diante.

O banco de dados original não possui muitas variáveis explicativas. Mas pela natureza do problema, sabe-se que é possível extrair novas características das já existentes que sejam de grande utilidade. Das diversas possibilidades, pode-se trabalhar com as seguintes medidas calculadas por música:

1. Porcentagem de acordes suspensos (e.g. Gsus).
2. Porcentagem de acordes com a sétima nota menor (e.g. C7, Em7).
3. Porcentagem de acordes com a sétima nota e menor (interação entre ter sétima nota e ter a terça menor, e.g. Em7, C#m7).
4. Porcentagem de acordes com a sétima maior (e.g. F7+, Am7+).
5. Porcentagem de acordes menores (Em, C#m).
6. Porcentagem de acordes diminutos (e.g. B°).
7. Porcentagem de acordes aumentados (e.g. Baug).
8. Porcentagem de acordes com a quarta nota (e.g. D4).
9. Porcentagem de acordes com a sexta nota (e.g. E6).
10. Porcentagem de acordes com a nona nota (e.g. G9).
11. Porcentagem de acordes com a quinta nota diminuta (e.g. C5- ou C5b).
12. Porcentagem de acordes com a quinta nota aumentada (e.g. C5+ ou C5#).
13. Porcentagem da transição entre acordes mais comum na música.
14. Porcentagem da segunda transição entre acordes mais comum na música.
15. Porcentagem da terceira transição entre acordes mais comum na música.
16. Porcentagem de acordes com baixo variante (e.g. C/E, C/G, C/Bb).
17. Quantidade total de acordes não distintos.
18. Quantidade absoluta do acorde mais comum da música.
19. Popularidade da canção, proveniente da API do Spotify.
20. Média de distâncias dos acordes ao dó, pelo círculo das quintas.
21. Média de distâncias dos acordes ao dó em semitons.
22. Ano de lançamento do álbum que contém a música.
23. Indicador de diferença se a escala e o acorde mais comum da música são os mesmos.

A obtenção das variáveis de porcentagem (1 a 16) é feita de forma simples. Os passos são: 1. identificar quais acordes possuem os atributos, formando uma variável indicadora para cada um deles; 2. encontrar a soma, por música, destas variáveis; 3. dividir essa soma pela total de acordes de cada música.

A identificação destes atributos, por sua vez, é feita através do texto de cada acorde. Por exemplo, o acorde “Cm7” é um acorde de Dó menor com sétima menor. O acorde “F5+” é um acorde de Fá maior com quinta aumentada. Encontrar a presença ou não das características é feita através de um conjunto de expressões regulares que as representem. Todavia, isso não é algo trivial, já que os mesmos acordes ou atributos algumas vezes são escritos de diferentes formas já que as cifras são contribuições do público e existe variação de adoção de nomenclatura, mas significam o mesmo. A Tabela 3 mostra quais são as expressões regulares para cada caso.

Tabela 3 – Expressões regulares para a extração de características.

Exp. Regular	Acordes	Exemplo
m	Menores	Cm
(dim °)	Diminutos	Edim, E°
aug	Aumentados	Eaug
sus	Suspensos	Fsus
7	Sétima	A7
7(M \\+)	Sétima maior	F7M, C7+
(6 13)	Sexta	D6,13
(4 11)	Quartas	B4, B11
5(# \\+)	Quinta aumentada	A5+, A5#
5(b -)	Quinta diminuta	B5b, B5-
(9 2)	Nona	E9, E2
(?<=/.)*	Acordes com baixo variante	C/G

De outro ponto de vista, ressalta-se a relação dos cálculos de variáveis extraídas e estatísticas suficientes. Pode-se considerar que, para cada música, as observações da variável indicadora de presença de atributos são independentes e seguem uma distribuição de probabilidades $Bernoulli(\theta_j)$. Logo, o somatório destas observações, $\sum_{i=1}^{n_j} x_i$, para cada música, segue uma distribuição $Binomial(n_j, \theta_j)$. Este somatório, $T = \sum_{i=1}^{n_j} x_i$ é uma estatística suficiente para a variável indicadora, sendo que isto quer dizer que essa estatística contém toda a informação que a amostra possui sobre θ_j , que é probabilidade de cada acorde possuir os atributos, na j -ésima música. Assim, é razoável resumir a variável à esta medida. Apesar deste exemplo ser natural, estatísticas suficientes são tópicos comumente discutidos quando se fala sobre redução de dimensionalidade, justamente pela propriedade de condensarem as informações relevantes da amostra (REAL, 1996a). Vínculos parecidos podem ser estabelecidos também entre estatísticas suficientes e programação paralela (ZHANG; HSU; FORMAN, 2000).

A variável sobre a quantidade total de acordes não distintos (17) apenas conta quantos eles são em cada música. Esta variável traz uma noção da complexidade da música, uma vez que essa informação não está disponível na base. A característica (18) indica quantas vezes o acorde mais comum da música aconteceu. Já a popularidade (19) é uma medida extraída com auxílio da API do Spotify (SPOTIFY...), podendo variar de 0 a 100, sendo que o 100 indica que a música é muito popular.

A média das distâncias ao Dó (20), segundo o círculo das quintas, é calculada usando algumas noções de harmonia. O círculo das quintas indica o campo harmônico mais provável para as músicas, ou seja, os acordes mais próximos neste círculo acontecem com mais frequência. Por exemplo, se temos uma música na escala de Dó (C), o acorde dominante será o próprio Dó, e acordes muito comuns serão o Sol (G), vizinho direito do Dó no círculo, e o Fá (F), o vizinho esquerdo. Estes acordes terão, segundo a medida proposta, a distância 1 do Dó, considerando o círculo das quintas, já que eles são vizinhos de primeira ordem. Como o Dó, na verdade, é o acorde mais comum em música, ele é o ponto 0, ou seja, a referência para o cálculo das distâncias. Essa medida tem a utilidade de dar um indicativo do quão atípicos os acordes em uma canção são, em relação ao que é habitual em harmonia.

Outra forma de saber a distância ao Dó, agora à nota Dó e não ao acorde, é contar quantos semitons existem dele até a nota raiz do acorde (21). Por exemplo, para o acorde Ré (D), cuja primeira nota é o Ré, esta distância é de dois semitons. Para o acorde de Fá por outro lado, ela é 7 e não 8, como poderia ser esperado, por ser uma nota 4 posições à frente do dó. Ambas as medidas de distância estão na Tabela 4, com os valores atribuídos a cada acorde.

Tabela 4 – Distâncias à nota Dó, pelo círculo das quintas e por semitons.

Acorde	Círculo	Semitons
C	0	0
G	1	7
D	2	2
A	3	9
E	4	4
B	5	11
F#	6	6
C#	5	1
Ab	4	8
Eb	3	3
Bb	2	10
F	1	5

A característica calculada é definida como a média das distâncias dos acordes ao Dó, pelo círculo das quintas (20) ou por semitons (21), para cada música, ou seja

$$Dist = \frac{\sum_{i=1}^{n_j} d_i}{n_j} \quad (3.1)$$

aonde os d_i são as distâncias para os i -ésimos acordes, $i = 1, \dots, n_j$, da j -ésima música, que tem tamanho n_j .

O ano de lançamento do álbum (22) que contém a música também é obtido através da API do Spotify. A técnica é simples: basta extrair, para cada artista, os anos de seus álbuns e verificar em qual álbum cada música foi lançada. Acredita-se que essa variável discrimina bem entre os gêneros, dado que alguns deles foram mais populares em épocas passadas do que nas recentes. A última característica (23) sugerida trata da indicadora sobre se o acorde mais comum na música é aquele que define a escala. Em música, faz sentido pensar que, se uma canção está na escala de Dó, um acorde de grande ocorrência será o próprio Dó. Mas isso nem sempre é verdade, podendo ser outro indicativo de que a música não segue completamente as regras empíricas de harmonia.

3.3 Análise Exploratória

Antes de avançar para métodos de exploração dos dados mais sofisticados, inicia-se com a análise exploratória gráfica. Os métodos utilizados são: gráficos de pontos, gráficos de segmentos e diagramas de cordas (CHIASSON; GREGORY, 2014).

Os gráficos de pontos são utilizados quando é necessário avaliar a relação entre duas variáveis, e aparecem comumente em combinação com alguma forma de descrever a relação entre as variáveis, como uma reta de regressão linear simples ou de um suavizador não paramétrico. Os gráficos de segmentos, por sua vez, trazem medidas resumo, como o máximo, mínimo e mediana, sobre uma variável de interesse. E os diagramas de cordas trazem uma forma interessante de visualizar interações entre indivíduos, ou grupos, de alguma população ou amostra. Os arcos do diagrama representam a força destas relações.

3.4 Árvores de Classificação

Um dos objetivos deste trabalho é avaliar quais características são relevantes na descrição de gêneros musicais, ou seja, em sua classificação. Para isso, vamos partir do algoritmo de árvores de classificação (Hastie, Trevor, Tibshirani, Robert, Friedman, 2009).

Árvores de regressão são similares à modelos aditivos no que elas representam o compromisso entre um modelo linear e uma abordagem completamente não paramétrica. Na literatura, a metodologia de árvores é aplicada em problemas de decisão aonde não há uma estrutura estocástica, mas existe interesse em construir regras para fazer as decisões corretas (FARAWAY, 2005).

Árvores de regressão são aquelas nas quais há interesse em fazer previsões de uma variável contínua. O método consiste na partição do espaço das covariáveis em regiões retangulares e no ajuste de um modelo simples em cada uma delas, sem fazer suposições de distribuição sobre a variável resposta. Este ajuste pode ser simplesmente uma constante, como a média da variável a ser prevista em cada região. O algoritmo de construção da árvore é:

1. Em uma variável preditora é feita uma partição, escolhendo-se um ponto desta variável que seja a quebra.
2. Para cada partição, encontra-se a média da variável resposta naquela região. A partir dessa média, encontra-se

$$SQR_{partição} = SQR_{part1} + SQR_{part2} \quad (3.2)$$

que é o critério a ser minimizado pela árvore. Assim, a partição final da variável preditora é aquela que produz a menor soma de quadrados de resíduos.

3. As partições são sub-partidas, de forma recursiva. São feitas novas partições a partir das anteriores, como ramos de uma árvores.

Durante o processo de seleção das partições, muitos pontos de quebra são avaliados. Apesar disso, os cálculos em cada partição são simples, o que permite que o modelo seja ajustado sem maiores esforços computacionais. Em caso de preditoras categóricas, as partições são feitas nos níveis da variável. O método é também apropriado para encontrar interações entre variáveis. Quando uma variável é segmentada a partir de outra, isso configura uma interação entre elas. Com o crescimento da árvore, o número de interações também aumenta.

Uma árvore de classificação é similar a uma de regressão, com a diferença de que o interesse agora é prever classes. O algoritmo é muito parecido com o de árvores de regressão, exceto pelo critério a ser minimizado, que pode ser a taxa de classificações incorretas. Em cada partição, a predição é feita com base na classe mais comum para as duas regiões produzidas. Ou seja, o objetivo é encontrar partições nas quais uma classe da variável resposta seja prevalente, tornando a previsão mais assertiva. A representação gráfica de um exemplo simples de árvore de classificação está na Figura 4.

As divisões binárias nos ramos das árvores são feitas considerando as opções que minimizam algum critério de heterogeneidade, como o critério de Gini. Essa medida é dada por

$$Gini = 1 - \sum_{i=1}^r p_i^2 \quad (3.3)$$

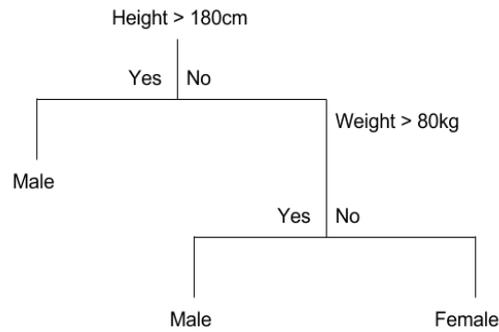


Figura 5 – Representação gráfica de uma árvore de classificação com resposta binária.

aonde cada p_i é a proporção com que a categoria i é observada na base. Essa medida é minimizada quando os indivíduos pertencem à uma mesma categoria (existe uma classe predominante), e, de forma contrária, ela é a maior possível quando as proporções das classes da variável resposta são iguais na região. O critério de Gini pode ser chamado também de impureza. As divisões das variáveis predictoras são feitas de forma a minimizar a diferença entre a impureza da divisão original e a seguinte, ou seja, minimizar

$$\Delta Imp = Imp_O - \left(\frac{n_1}{n_O} Imp_1 + \frac{n_2}{n_O} Imp_2 \right) \quad (3.4)$$

aonde 1 e 2 são as duas novas divisões candidatas, que partem da região O .

Existem algumas vantagens em usar o método de árvores. Regiões de decisão não lineares entre variáveis explicativas são melhor capturadas por ele do que por modelos lineares, por exemplo. A Figura 6 mostra como isso é contornado pelas árvores, para duas classes. Na ilustração, temos um comportamento que não seria tão bem capturado por um modelo linear, mas a árvore consegue lidar com isso, por meio de suas divisões por regiões.

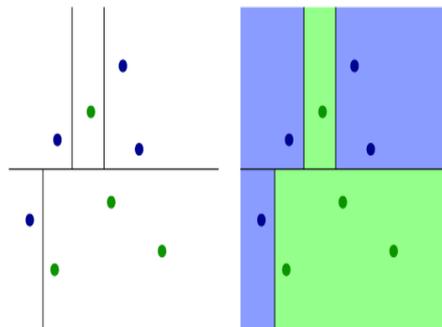


Figura 6 – Divisão de regiões por uma árvore, com limiares não-lineares para as classes.

Adicionando à lista anterior, os resultados das árvores, que formam um conjunto de regras de decisão, são próximos do que o julgamento humano faria. Isso torna-os facilmente interpretáveis mesmo para quem não é familiarizado com métodos estatísticos.

A avaliação da qualidade de predição de cada árvore é feita através da comparação entre o que foi observado na base de teste e o que o modelo prediz para ela. Assim, quanto maior for a proximidade entre o vetor de valores preditos e o vetor de valores observados em dados não vistos pelo modelo, melhor o algoritmo está performando. Esta medida é conhecida como “acurácia”, e é popular em problemas que envolvem classificação. A medida de acurácia pode ser escrita como:

$$\text{Acurácia} = \frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i) \quad (3.5)$$

aonde I indica se a predição feita pelo modelo, \hat{y}_i , é compatível com o que foi observado, y_i , e n é o tamanho amostral. A acurácia é algo que está no intervalo $[0, 1]$, sendo que, quanto mais perto de 1, melhor o modelo está performando, e vice-versa. Mesmo em situações nas quais o foco não seja encontrar um algoritmo que prediz o melhor possível, ter uma árvore com boa capacidade preditiva é uma qualidade sempre desejável.

3.5 Florestas Aleatórias

Uma vez que conhecida a definição de árvores de classificação, pode-se falar sobre as florestas aleatórias (Hastie, Trevor, Tibshirani, Robert, Friedman, 2009). Neste caso, diversas árvores são construídas, formando uma coleção que é combinada para a produção dos resultados.

Árvores de decisão podem ter alta variância. Isso quer dizer que, se a base original for separada em duas e em cada uma delas ajustar-se um modelo, os resultados serão discordantes. Florestas aleatórias utilizam “*bagging*” para a redução da variância. Este procedimento baseia-se na propriedade de que, se existe um conjunto de observações independentes Z_1, \dots, Z_n , com variância σ^2 , a média delas, \bar{Z} , tem variância σ^2/n . Uma forma natural de reduzir a variância é selecionar diversos conjuntos de treinamento, provenientes da população original, construir um modelo para cada um deles, e computar a média das predições destes modelos. O objetivo final é obter um modelo de baixa variância para predição, dado pela média dos outros modelos, ou seja

$$\hat{f}_{\text{média}}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(x) \quad (3.6)$$

aonde os f^b representam os modelos construídos com cada conjunto b , gerados a partir da população original. O problema é que, em geral, estes B conjuntos não estão disponíveis. Para contornar esta dificuldade, os conjuntos são formados com amostras *bootstrap*, ou seja, amostras de mesmo tamanho que são retiradas com reposição. Os algoritmos são treinados em cada amostra *bootstrap*, gerando o que é conhecido como modelo de *bagging*:

$$\hat{f}_{\text{bagging}}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x) \quad (3.7)$$

Em florestas aleatórias, são construídas B árvores, uma para cada amostra *bootstrap*, e os B resultados são combinados através da média das predições. No caso de classificação, esta média é substituída pela classe mais escolhida entre as B árvores. Ou seja: a predição final será aquela que for mais comum entre as B predições, das B árvores de classificação geradas. Assim, o grupo de árvores com alta variância acaba formando uma floresta, que tem baixa variância e é mais acurada do que as árvores sozinhas.

Os modelos de que compõem as florestas aleatórias têm algumas peculiaridades. A cada momento em que uma divisão é considerada, m preditoras, dentre as p totais ($m < p$), são escolhidas de forma aleatória e sem reposição como candidatas à divisão. Logo, só são avaliadas m preditoras em cada passo do algoritmo. Tipicamente, m é escolhido como aproximadamente \sqrt{p} .

Esse passo é usado para descorrelacionar as árvores que serão combinadas. Suponha que existe uma preditora muito forte, enquanto as outras são moderadas. Sem a seleção das m variáveis a cada partição, a maioria das árvores vai usar esta variável forte na primeira divisão, tornando as predições altamente correlacionadas. Em particular, isso significa que a redução na variância não será tão eficaz. Logo, ter um método que descorrelaciona as árvores torna os resultados mais confiáveis.

3.5.1 Importância de Variáveis

A combinação de várias árvores, a floresta, é de mais difícil interpretação do que apenas uma árvore. Por outro lado, é possível obter uma medida de importância para as variáveis incluídas no modelo. A diminuição na soma dos quadrados dos resíduos resultante de cada nova divisão é contabilizada para as árvore da floresta. Ao fim do algoritmo, a média desta diminuição, para cada variável, é acessada, tornando-se a medida que informa o quanto cada variável é importante para o modelo. Em os problemas que envolvem classificação, o método é parecido, mas o critério considerado é de Gini, mencionado anteriormente. Ou seja, o foco é na redução do critério de Gini por cada quebra.

4 Resultados e Discussão

Neste capítulo, são apresentados os resultados e discussões. As primeiras seções são dedicadas à extração e análise exploratória dos dados, enquanto as seções finais discutem os resultados da modelagem.

4.1 O pacote *chorrds*

A extração dos dados foi, de forma resumida, feita através de técnicas de *web scraping*. O intuito é pesquisar, no *html* das páginas do site CifraClub, aonde se encontram os acordes das músicas e capturá-los. Após isso, esses dados são armazenados e retornados para que possam ser utilizados. Um total de 8261 músicas foram capturadas, de uma base total com 483.991 linhas.

Para a extração dos dados, utilizou-se o pacote *chorrds*, construído especificamente para isso. Ele está disponível para download e uso a partir do CRAN (*The Comprehensive R Archive Network*), que é o repositório oficial de pacotes para o software R, desde Abril de 2018. Atualmente, está na versão 0.1.1, e vai continuar a ser aprimorado.

Entende-se que a reprodutibilidade das análises realizadas, e a produção de ferramentas úteis provenientes delas, como o pacote para o R, é de grande relevância. Além disso, o código fonte é disponibilizado no endereço <<https://github.com/brunaw>> (Wundervald (2018)) Além disso, os códigos geradores das análises estão disponíveis em <<https://github.com/r-music>> (Wundervald e Trecenti (2018)).

4.2 Análise Exploratória

4.2.1 Acordes

De forma geral, muito se fala sobre o quanto as músicas no Brasil vêm se tornando mais simples, ou que alguns gêneros musicais são mais ricos que outros. Com os dados disponíveis é possível verificar, por exemplo, a quantidade média de acordes distintos, por música, ao longo dos anos.

Na Figura 7, os gêneros musicais mais tradicionalmente brasileiros, como o samba, MPB e bossa nova, têm uma quantidade média de acordes distintos por música maior do que os demais. Eles têm uma maior variação ao longo dos anos também. O conhecimento popular que diz que esses gêneros são mais ricos em diversidade de acordes é sustentado por este resultado. Além disso, de forma complementar, os outros gêneros são mais uniformes harmonicamente. Estes, por sua vez, são conhecidamente mais simples e apenas apareceram no Brasil em períodos posteriores aos outros gêneros (CALDAS, 2010).

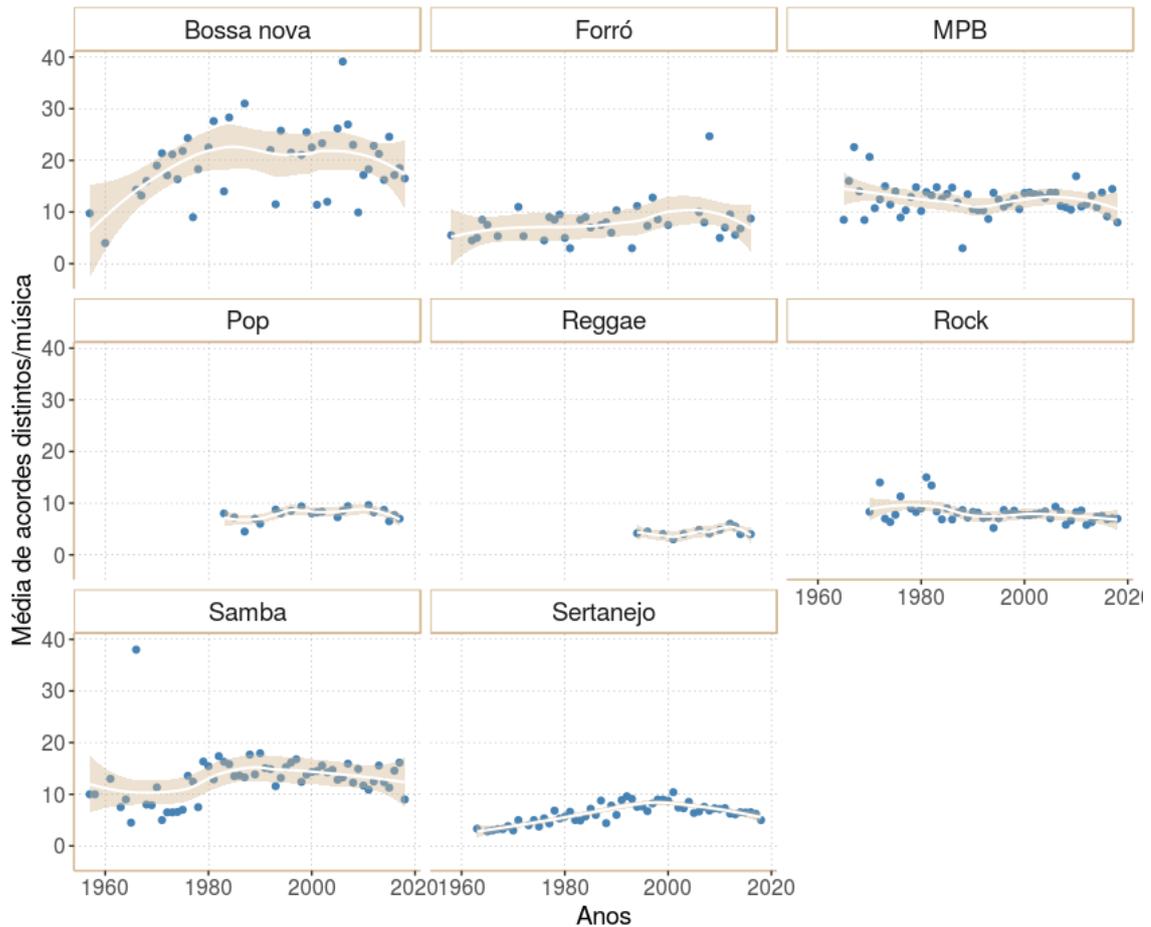


Figura 7 – Quantidade média de acordes distintos nas músicas em cada ano, com separação por gêneros musicais. A linha acompanhada do envelope representa o ajuste de um modelo de regressão local de suavização não paramétrica considerado para destacar a tendência sobre a média da distribuição.

Os valores apresentados no gráfico anterior não revelam a variação entre artistas. Um mesmo artista pode apresentar considerável variação com relação o número de acordes por canção. Ainda sob a ótica de acordes, observa-se como os artistas se comportam em relação à quantidade de acordes distintos em suas obras.

Na Figura 8, a primeira posição é ocupada por um dos principais musicistas brasileiros, que é também poeta e referência internacional em questões de refinamento musical: Chico Buarque. Logo em seguida, temos o Reinaldo, um dos maiores sambistas que o Brasil já conheceu, e o Pixinguinha, conhecido como gênio do choro. A lista segue com artistas principalmente da bossa nova, samba e MPB. Artistas de outros gêneros como o rock não estão no topo da lista, e a primeira representante deste gênero é a Rita Lee. Dos representantes do gênero sertanejo, nenhum aparece no gráfico, mostrando que a variedade harmônica desse gênero musical é baixa.

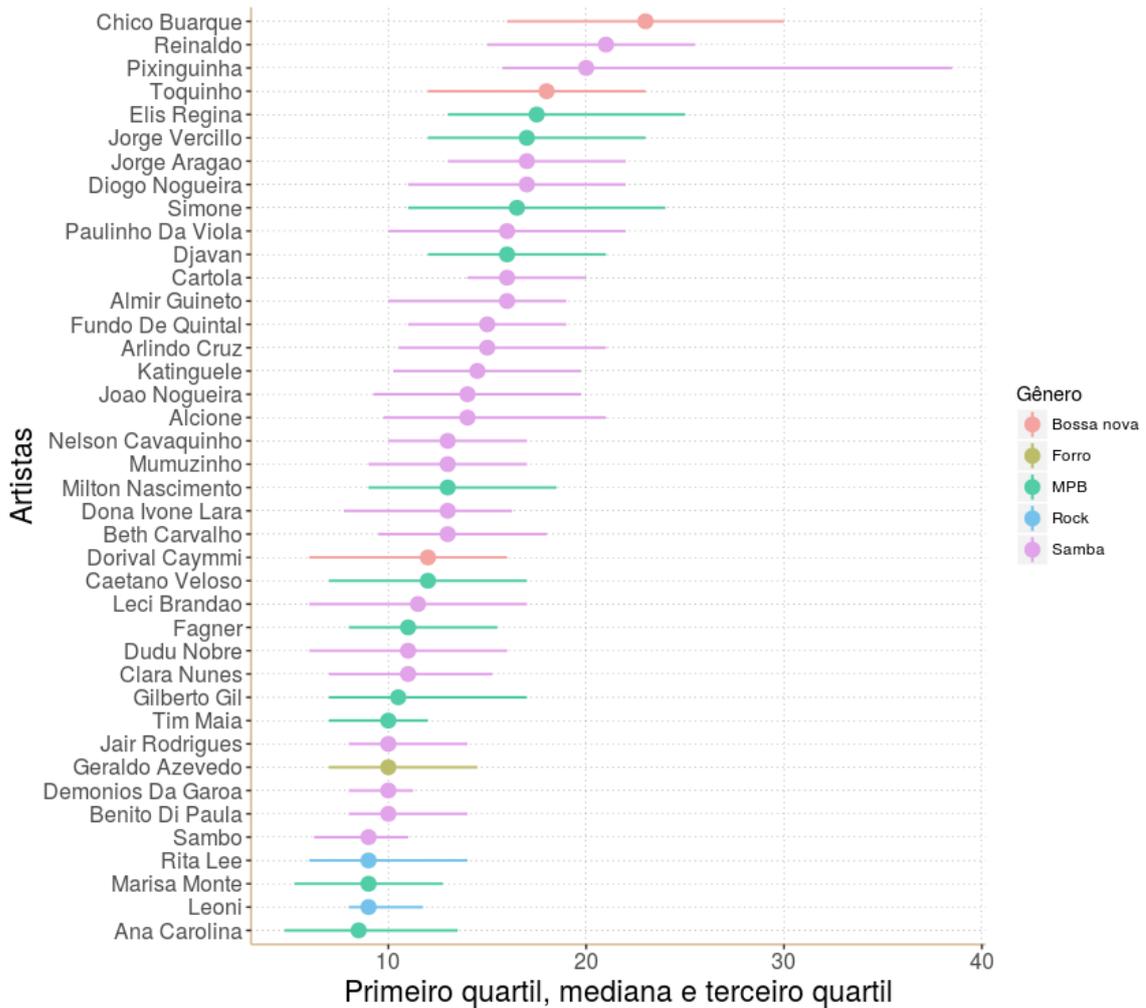


Figura 8 – Mediana de acordes distintos por artista, com intervalo quantílico. São apresentados os artistas com mediana superior a 8 acordes distintos por música em ordem decrescente.

É possível explorar também a estrutura harmônica dos gêneros de forma mais geral, sem considerar fatores temporais ou artistas específicos. Uma forma de fazer isso é encadear as transições entre os acordes em um diagrama de cordas, explicitando as relações entre elas. Neste caso, a transição significa um acorde ser seguido do outro em algum momento da música. Os diagramas de cordas mostram as porcentagens de vezes, através de arcos, que um acorde foi seguido do outro. Para a construção dos gráficos, utiliza-se apenas a forma fundamental do acorde, sem informações de adição de notas. Esta aproximação é feita para que os acordes possam ser dispostos de acordo com o círculo das quintas. As próximas figuras mostram estes diagramas para cada gênero.

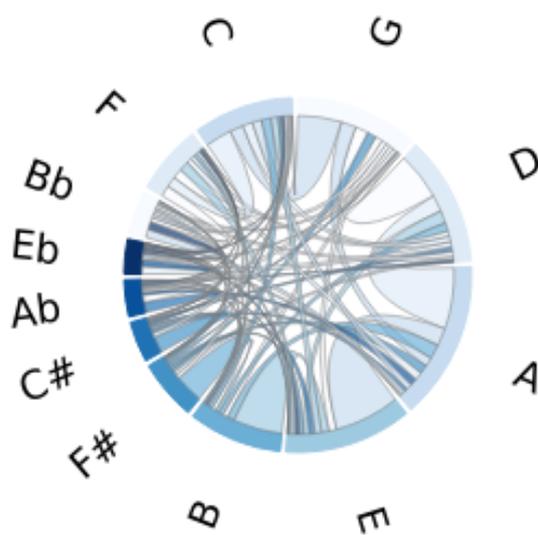


Figura 9 – Diagrama de Cordas - Bossa Nova

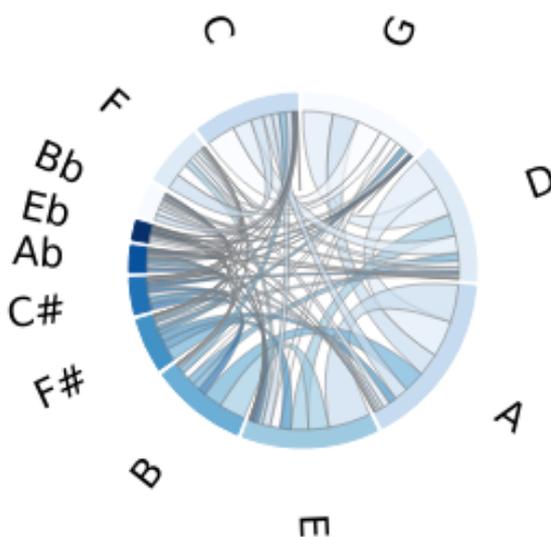


Figura 10 – Diagrama de Cordas - Forró

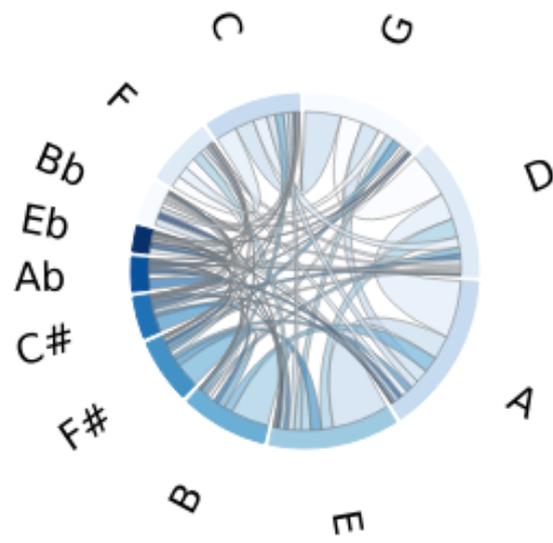


Figura 11 – Diagrama de Cordas - MPB

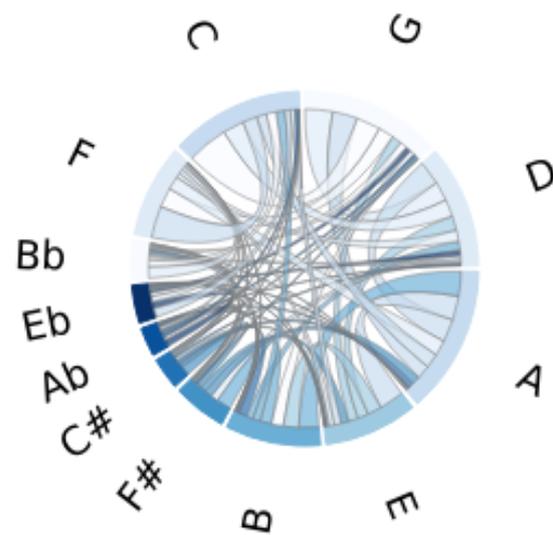


Figura 12 – Diagrama de Cordas - Pop

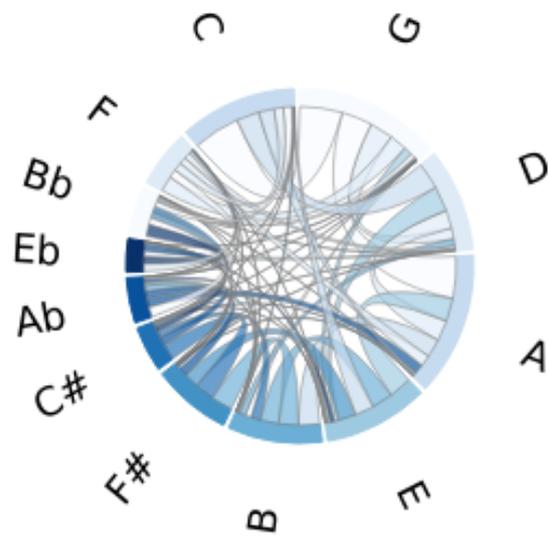


Figura 13 – Diagrama de Cordas - Reggae

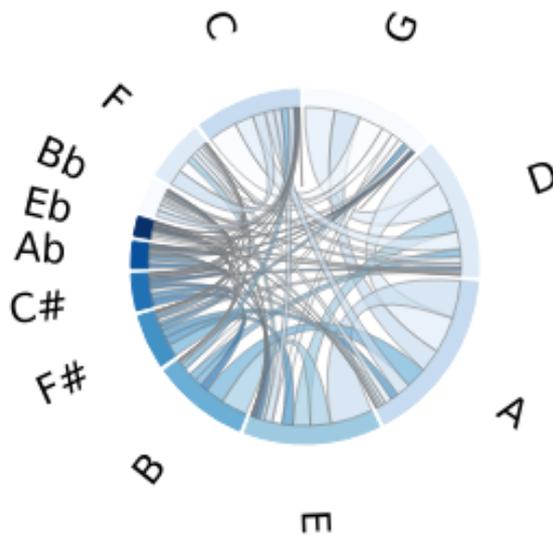


Figura 14 – Diagrama de Cordas - Rock

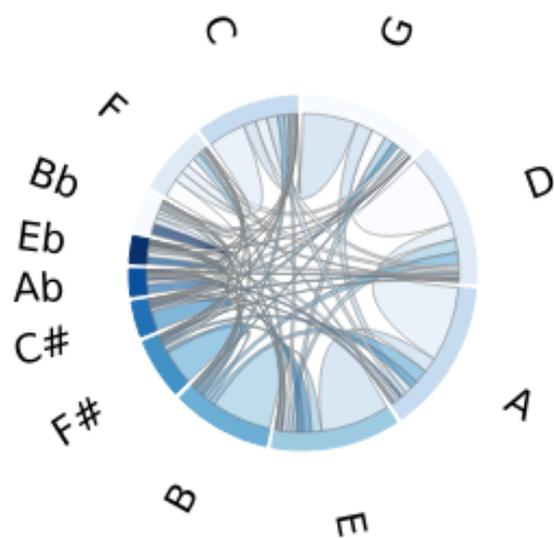


Figura 15 – Diagrama de Cordas - Samba

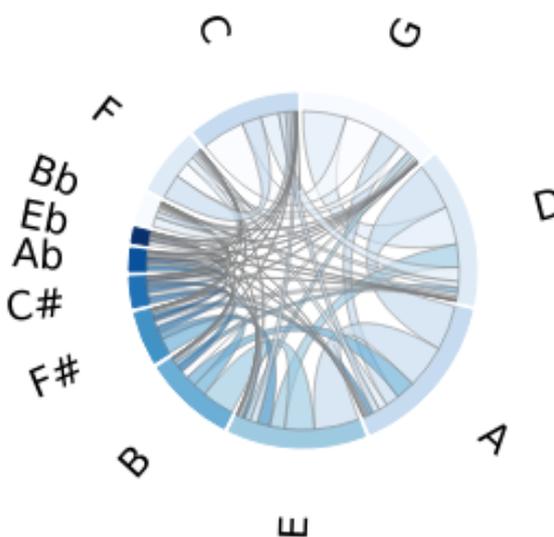


Figura 16 – Diagrama de Cordas - Sertanejo

As Figuras 9, 10, 11, 12, 13, 14, 15 e 16 representam as frequências nas quais as transições dos acordes aconteceram por gênero, ordenadas pelo círculo das quintas, que é um instrumento fundamental para o entendimento de sequências harmônicas. Esta configuração é usada uma vez que as transições mais frequentes são as das tônicas para suas vizinhas.

Para a Bossa Nova, há uma maior riqueza de transições, visto que as mais incomuns ocorrem com bastante frequência nesse gênero, ao compararmos com os outros. Este cenário é parecido com o Rock e Forró. Para o Samba, por outro lado, as ocorrências mais frequentes são de acordes próximos no círculo das quintas, mostrando o quanto este gênero segue as regras de teoria musical. Diferenças assim ajudam a evidenciar se a classificação por gêneros dadas às músicas brasileiras pode ser caracterizada através de estruturas harmônicas.

4.2.2 Variáveis Extraídas

As variáveis extraídas por meio de engenharia de características foram divididas em grupos temáticos. Estas divisões são:

- Conjunto 1 - Tríades e tétrades simples
 1. Porcentagem de acordes suspensos (e.g. Gsus).
 2. Porcentagem de acordes com a sétima nota menor (e.g. C7, Em7).
 3. Porcentagem de acordes com a sétima nota e menor (interação entre ter sétima nota e ter a terça menor (e.g. Em7, C#m7).
 4. Porcentagem de acordes com a sétima maior (e.g. F7+, Am7+).
 5. Porcentagem de acordes menores (e.g. Em, C#m).
 6. Porcentagem de acordes diminutos (e.g. B°).
 7. Porcentagem de acordes aumentados (e.g. Baug).
- Conjunto 2 - Tétrades dissonantes
 1. Porcentagem de acordes com a quarta nota (e.g. D4).
 2. Porcentagem de acordes com a sexta nota (e.g. E6).
 3. Porcentagem de acordes com a nona nota (e.g. G9).
 4. Porcentagem de acordes com a quinta nota diminuta (e.g. C5- ou C5b).
 5. Porcentagem de acordes com a quinta nota aumentada (e.g. C5+ ou C5#).
- Conjunto 3 - Transições e frequências absolutas
 1. Porcentagem da transição entre acordes mais comum na música.

2. Porcentagem da segunda transição entre acordes mais comum na música.
 3. Porcentagem da terceira transição entre acordes mais comum na música.
- Conjunto 4 - Informações extra
 1. Porcentagem de acordes com baixo variante (e.g. C/E, C/G, C/Bb).
 2. Quantidade total de acordes não distintos.
 3. Quantidade absoluta do acorde mais comum da música.
 4. Popularidade da canção, proveniente da API do Spotify.
 5. Média de distâncias dos acordes ao dó, pelo círculo das quintas.
 6. Média de distâncias dos acordes ao dó em semitons.
 7. Ano de lançamento do álbum que contém a música.
 8. Indicador de diferença se a escala e o acorde mais comum da música são os mesmos.

A Tabela 5 apresenta um pequeno resumo das variáveis construídas que são contínuas.

Tabela 5 – Resumo das variáveis extraídas contínuas.

Variável	Média	DP	Min	Max
Sus	0.00097	0.013	0	0.46
Sétima	0.3	0.31	0	1
Sétima e Menor	0.0028	0.022	0	1
Sétima Maior	0.038	0.088	0	1
Menores	0.27	0.2	0	1
Diminuto	0.01	0.033	0	0.58
Aumentado	0.022	0.068	0	1
Quarta	0.016	0.016	0	0.68
Sexta	0.025	0.072	0	0.85
Nona	0.059	0.13	0	1
Quinta Dim.	0.0061	0.024	0	0.27
Quinta Aum.	0.0033	0.02	0	0.39
Transição (1 ^a)	0.19	0.081	0.0078	0.8
Transição (2 ^a)	0.17	0.075	0.016	0.9
Transição (3 ^a)	0.15	0.073	0.01	0.88
Baixo	0.081	0.15	0	1
Total de Acordes	57.12	31.97	1	370
Qtde. acorde mais comum	18.72	11.54	1	202
Popularidade	28.54	14.6	1	83
Dist. ao Dó	2.67	1.18	0	6
Dist. em Semitons	5.48	1.27	0	1

Sobre as variáveis discretas, a Tabela 6 traz mais informações.

Tabela 6 – Resumo das variáveis extraídas discretas.

Variável	Classes	Mais comum
Data	1957 a 2017	2015 (7%)
Diferença entre escala e acorde mais comum	Sim ou Não	Não (67%)

A próxima figura apresenta os *boxplots* de linhas das variáveis extraídas. Com ela, percebe-se os contrastes entre os comportamentos de cada variável entre os gêneros. Estes são possíveis evidências de que as variáveis extraídas podem ser capazes de permitir a distinção entre as classificações por gêneros.

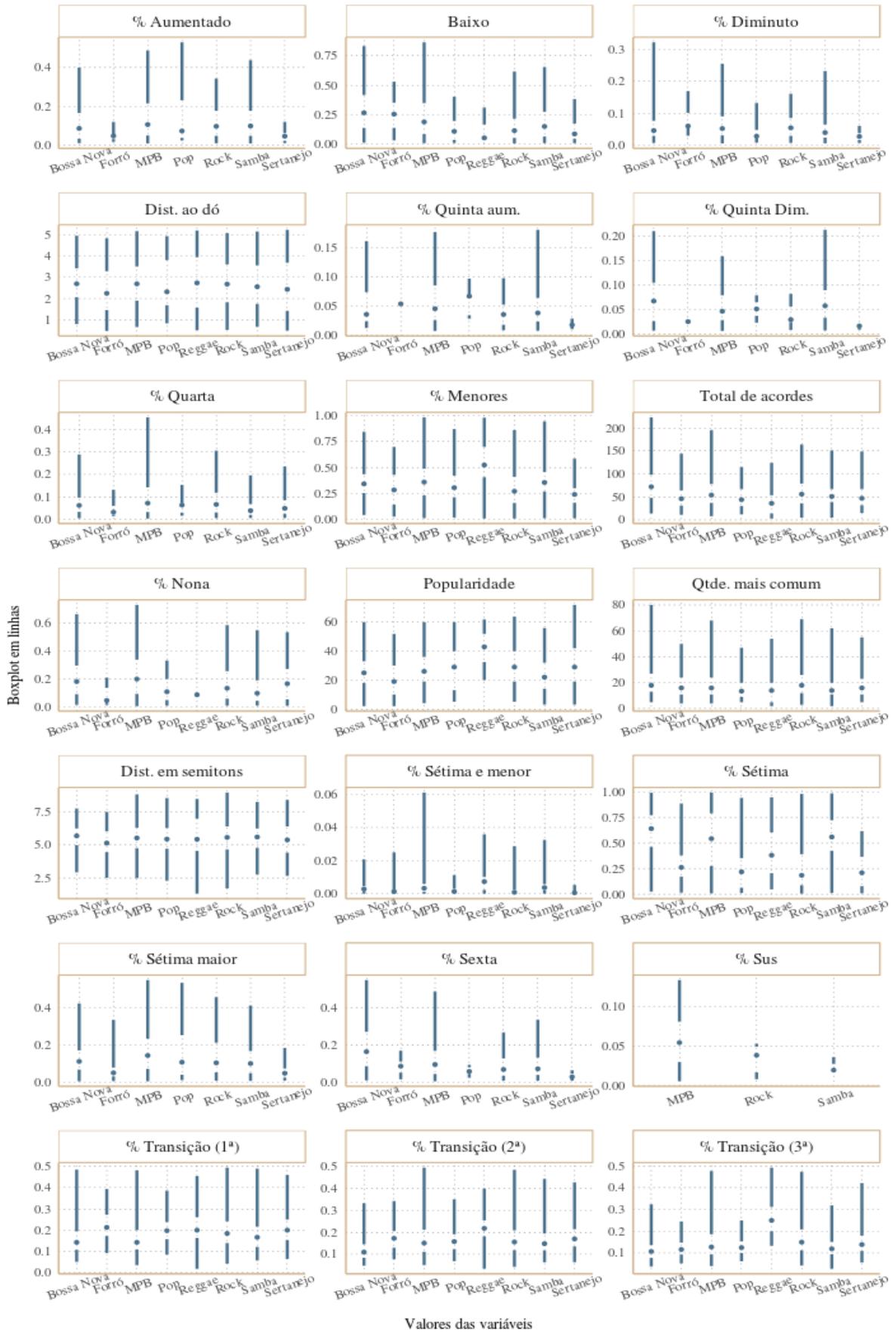


Figura 17 – Boxplots das variáveis extraídas.

Os gráficos mostram como algumas variáveis são discrepantes entre os gêneros. Por exemplo, para a variável de baixo, porcentagem de quarta nota, acordes diminutos e aumentados, os comportamentos variam entre os gêneros. Isso é ainda mais forte em variáveis como a quantidade de sétimas em acordes menores, a a porcentagem de acordes sus e com sexta nota, e assim por diante.

4.3 Modelagem

4.3.1 Conjuntos de Variáveis

O foco principal do trabalho é a extração e construção de características para a criação de um classificador de gêneros musicais baseado em cifras de músicas. Considerando isso, diante da quantidade de variáveis extraídas, a avaliação da qualidade do classificador como função das mesmas, estabelece-se através da separação das variáveis em conjuntos. Desta forma, os modelos utilizados **vão crescendo em quantidade de variáveis**, seguindo uma estrutura lógica.

1. Conjunto 1 (tríades): apenas as variáveis relacionadas a acordes menores, com sétima, aumentados ou diminuto (ou seja, as tríades e acorde com sétima).
2. Conjunto 2 (tétrades): o conjunto acima acrescido de variáveis sobre as quartas, sextas, sétimas maiores e nonas notas.
3. Conjunto 3 (transições): os anteriores com a inclusão das frações entre as transições entre acordes.
4. Conjunto 4 (informações extras): inclusão do ano e popularidade da canção, que são informações além do que é pode ser tirado da cifra, quantidade de acordes e médias de distâncias.

As variáveis de cada conjunto são mostradas na Tabela 7.

Tabela 7 – Variáveis em cada conjunto

Conjunto 1	Conjunto 2	Conjunto 3	Conjunto 4
% Menores	% Quartas		Ano
% Sétima	% Sextas	% 1 ^a Trans. comum	Popularidade
% Sétima e menor	% Nona	% 2 ^a Trans. comum	% Baixo
% Aumentados	% Sétima maior	% 3 ^a Trans. comum	Dist. média ao dó em semitons
% Sus	% Quinta aum.		Dist. média ao dó pelo círculo
% Diminutos	% Quinta dim.		I(Escala = Acorde mais comum)
			Qtde. acorde mais comum)
% Sétima maior			Qtde. acordes distintos

Os conjuntos de variáveis são adicionados *sequencialmente*, produzindo ao fim quatro modelos distintos. Uma analogia boa é a de “modelos aninhados”, ou seja: a cada modelo, novas variáveis são incluídas e as comparações são realizadas. Mas as florestas aleatórias não podem ser estritamente chamadas de “aninhadas”, já que são baseadas em um procedimento não paramétrico. Nos resultados, analisa-se as performances de cada classificador.

4.3.2 Florestas Aleatórias

O modelo de florestas aleatórias foi escolhido para ser o algoritmo de classificação dos gêneros musicais. Todas as variáveis extraídas participam do algoritmo. A escolha deste métodos justifica-se em três pontos principais:

- Obtenção de medidas simples de importância das variáveis preditoras;
- Não é preciso realizar transformações/normalização das variáveis, uma vez que as árvores são invariante em relação à escala dos preditores.
- Acomodação de relações não-lineares entre as variáveis explicativas e a resposta, que são contornadas pelo modelo.

A base foi particionada aleatoriamente em dados de treino (70%) e de teste (30%), com balanceamento por gênero. A primeira partição é utilizada no treinamento do modelo, enquanto a segunda serve para a a verificação do desempenho na classificação de dados não utilizados no treinamento. O descritivo dos resultados das partições entre os gêneros está na Tabela 8.

Tabela 8 – Quantidade de músicas para cada gênero entre as partições.

Gênero	Treino	Teste	Representatividade do gênero
Bossa Nova	305 (68%)	133 (32%)	438 (5.3%)
Forró	115 (73%)	48 (27%)	163 (2%)
MPB	1196 (67.8%)	476 (32.2%)	1679 (20.3%)
Pop	104 (66.4%)	39 (33.6%)	143 (1.7%)
Reggae	46 (68.1%)	24 (31.9%)	70 (0.8%)
Rock	1127 (69.8%)	552 (30.2%)	1679 (20.4%)
Samba	877 (70.8%)	378 (29.2%)	1255 (15.1%)
Sertanejo	1992 (68.2%)	849 (31.8%)	2841 (34.4%)

Um extrato da base analítica final é exposto na Tabela 9.

Tabela 9 – Extrato da base final de modelagem

Música	Qtde	% m	% Dimi.	% Sus	% Aum.	% 7	%7 e menor
adoniran barbosa aguenta mao joao	19	0.11	0.00	0.00	0.00	0.00	0.00
adoniran barbosa apaga fogo mane	31	0.48	0.00	0.00	0.00	0.45	0.01
adoniran barbosa as mariposa	26	0.27	0.04	0.00	0.00	0.46	0.00
adoniran barbosa bom dia tristeza	27	0.33	0.00	0.00	0.07	0.74	0.01
adoniran barbosa conselho de mulher	33	0.24	0.00	0.00	0.00	0.39	0.00
adoniran barbosa despejo na favela	62	0.48	0.00	0.00	0.00	0.48	0.00

% 7 ^a	% 6 ^a	% 5 ^a Aum.	% 5 ^a Dim.	% 4 ^a	% 9 ^a	Baixo	Gênero	Ano	Popul.	Dist. semitons
0.00	0.00	0.00	0.00	0.00	0.00	0.00	Samba	1965	32	4.84
0.00	0.00	0.00	0.00	0.00	0.00	0.00	Samba	2010	25	4.84
0.00	0.00	0.00	0.00	0.00	0.00	0.00	Samba	2013	26	4.35
0.07	0.00	0.00	0.00	0.00	0.00	0.07	Samba	2013	23	6.07
0.00	0.00	0.00	0.00	0.00	0.00	0.00	Samba	1975	16	3.58
0.00	0.00	0.00	0.00	0.00	0.00	0.00	Samba	2005	30	7.55

Escala	Dist. no círculo	Qtde. acorde mais comum	+Comum = Escala	Partição
D	1.89	5	1	treino
F	1.87	13	0	teste
C	1.58	7	0	treino
F	2.15	9	0	teste
C	1.33	11	1	teste
G	4.84	26	0	treino

Em ambas as quatro florestas aleatórias, todos os gêneros estiveram presentes. Estes gêneros são: Bossa Nova, Forró, Sertanejo, MPB, Pop, Reggae, Rock e Samba. As medidas de qualidade dos ajustes estão na Tabela 10.

A estatística *Kappa* é uma métrica que compara a acurácia observada com a acurácia esperada. Esta acurácia esperada, também chamada de *Non Information Rate*, ou *N.I.R.*, é a proporção do gênero mais comum na base, que neste caso é 34%. Ou seja,

Tabela 10 – Comparativo de medidas de qualidade dos quatro modelos.

Modelo	Acurácia Geral	L.I. Acurácia	L.S. Acurácia	Kappa	Valor P (Acur. < <i>N.I.R</i>)
Modelo 1	0.53	0.51	0.55	0.37	< 0.000
Modelo 2	0.57	0.54	0.59	0.42	< 0.000
Modelo 3	0.59	0.56	0.60	0.44	< 0.000
Modelo 4	0.62	0.60	0.64	0.49	< 0.000

esta estatística serve para decidir se a classificação do modelo construído acerta mais do que dizer que todas as observações pertencem ao gênero mais comum, o que acarretaria em um acerto de aproximadamente 34%. Essa estatística é calculada como

$$kappa = \frac{p_0 - p_e}{1 - p_e} \quad (4.1)$$

onde p_0 é a acurácia (ou aceitação relativa) p_e é a taxa de aceitação esperada (ou *Non Information Rate*). Esta estatística segue, assintoticamente, uma distribuição aproximadamente Normal. O valor p , exposto na última coluna, indica se a acurácia dos modelos é mais alta do que a *N.I.R.* Segundo estes resultados, existem evidências de que, para os quatro modelos, a acurácia é significativa.

Observa-se que a inclusão dos novos conjuntos de variáveis vai, progressivamente, aumentando a acurácia dos modelos. Vale ressaltar que a medida de acurácia não se comporta como o R^2 , utilizado na avaliação do ajuste de modelos lineares em geral. O R^2 cresce a medida que novas variáveis são incluídas, invariavelmente, uma vez que ele não é penalizado pela quantidade de parâmetros. A acurácia, por outro lado, pode ser afetada negativamente com a inserção de características não informativas, que introduzem ruído aos dados, tornando mais difícil a distinção entre as classes. Em virtude desta propriedade, há mais segurança em afirmar que as variáveis incorporadas a cada novo modelo estão mesmo aumentando a capacidade de predição.

Este aumento na acurácia se dá uniformemente entre os modelos. A cada novo conjunto de variáveis, a acurácia melhora em aproximadamente 3%. Ao término dos ajustes, o modelo 4, que contém todas as características, possui uma capacidade de predição de aproximadamente 62%, sendo quase o dobro da *N.I.R.*. Por consequência, pode-se afirmar que tanto a estrutura harmônica das músicas, como propriedades extra, são capazes de caracterizar as classificações em gêneros musicais atribuídas às músicas presentes em nossa amostra. As matrizes de confusão, em porcentagens, estão a seguir.

Tabela 11 – Matriz de confusão para o modelo com o primeiro conjunto de variáveis.

	Bossa Nova	Forró	MPB	Pop	Reggae	Rock	Samba	Sertanejo
Bossa Nova	0.14	0.00	0.33	0.00	0.00	0.05	0.33	0.15
Forró	0.00	0.00	0.10	0.00	0.00	0.15	0.12	0.62
MPB	0.03	0.00	0.41	0.00	0.00	0.14	0.23	0.20
Pop	0.00	0.00	0.15	0.00	0.00	0.26	0.23	0.36
Reggae	0.00	0.00	0.25	0.00	0.00	0.50	0.04	0.21
Rock	0.01	0.00	0.11	0.00	0.00	0.34	0.07	0.47
Samba	0.02	0.00	0.26	0.00	0.00	0.05	0.57	0.11
Sertanejo	0.00	0.00	0.02	0.00	0.00	0.12	0.02	0.84

Tabela 12 – Matriz de confusão para o modelo com o primeiro e segundo conjuntos de variáveis.

	Bossa Nova	Forró	MPB	Pop	Reggae	Rock	Samba	Sertanejo
Bossa Nova	0.29	0.00	0.35	0.00	0.00	0.05	0.19	0.14
Forró	0.00	0.00	0.10	0.00	0.00	0.15	0.12	0.62
MPB	0.03	0.00	0.49	0.00	0.00	0.13	0.17	0.18
Pop	0.00	0.00	0.15	0.00	0.00	0.31	0.18	0.36
Reggae	0.00	0.00	0.17	0.00	0.00	0.50	0.12	0.21
Rock	0.00	0.00	0.13	0.00	0.00	0.36	0.06	0.44
Samba	0.02	0.00	0.20	0.00	0.00	0.04	0.63	0.10
Sertanejo	0.00	0.00	0.02	0.00	0.00	0.12	0.02	0.84

Tabela 13 – Matriz de confusão para o modelo com o primeiro, segundo e terceiro conjuntos de variáveis.

	Bossa Nova	Forró	MPB	Pop	Reggae	Rock	Samba	Sertanejo
Bossa Nova	0.29	0.00	0.35	0.00	0.00	0.05	0.17	0.13
Forró	0.00	0.00	0.06	0.00	0.00	0.21	0.08	0.65
MPB	0.03	0.00	0.55	0.00	0.00	0.12	0.15	0.15
Pop	0.00	0.00	0.23	0.00	0.00	0.13	0.21	0.44
Reggae	0.00	0.00	0.38	0.00	0.04	0.46	0.04	0.08
Rock	0.00	0.00	0.14	0.00	0.00	0.35	0.06	0.45
Samba	0.02	0.00	0.21	0.00	0.00	0.03	0.66	0.08
Sertanejo	0.00	0.00	0.02	0.00	0.00	0.09	0.02	0.86

Tabela 14 – Matriz de confusão para o modelo todas as variáveis.

	Bossa Nova	Forró	MPB	Pop	Reggae	Rock	Samba	Sertanejo
Bossa Nova	0.28	0.00	0.40	0.00	0.00	0.05	0.16	0.12
Forró	0.00	0.00	0.12	0.00	0.00	0.12	0.10	0.65
MPB	0.01	0.00	0.59	0.00	0.00	0.11	0.13	0.15
Pop	0.00	0.00	0.13	0.00	0.00	0.28	0.15	0.44
Reggae	0.00	0.00	0.25	0.00	0.08	0.46	0.08	0.12
Rock	0.00	0.00	0.16	0.00	0.00	0.43	0.05	0.35
Samba	0.01	0.00	0.20	0.00	0.00	0.03	0.66	0.10
Sertanejo	0.00	0.00	0.02	0.00	0.00	0.07	0.02	0.89

Da Tabela 11 para a 12, houve um acréscimo considerável na taxa de classificação correta especialmente para a Bossa Nova, seguida da MPB e Samba. Isso significa que com a inserção de variáveis sobre as tétrades, os gêneros clássicos da música popular brasileira são mais facilmente identificados pelo modelo. Na Tabela 13, o aumento também ocorre para a MPB e Samba, mas é mais chamativo para o Reggae. Antes, esta classe era completamente mal classificada. Agora, com as informações sobre transições mais comuns, pelo menos uma parte dela é distinguida pelo modelo, o que já é um ganho significativo. Finalmente, na Tabela 14, o acréscimo da acurácia continua sendo para a MPB e Reggae,

mas também acontece para o Rock com uma taxa alta, e para o Sertanejo. Anteriormente, os acréscimos do Rock e Sertanejo haviam sido pequenos. Com isso, verifica-se que as últimas variáveis, as extras, sobre popularidade, ano da música e distâncias ao dó, por exemplo, são notavelmente relevantes na classificação destes gêneros em especial.

Após estas conclusões, avalia-se especificamente os níveis de importância para as variáveis em cada modelo. As Figuras 18 a 21 ilustram estas medidas.

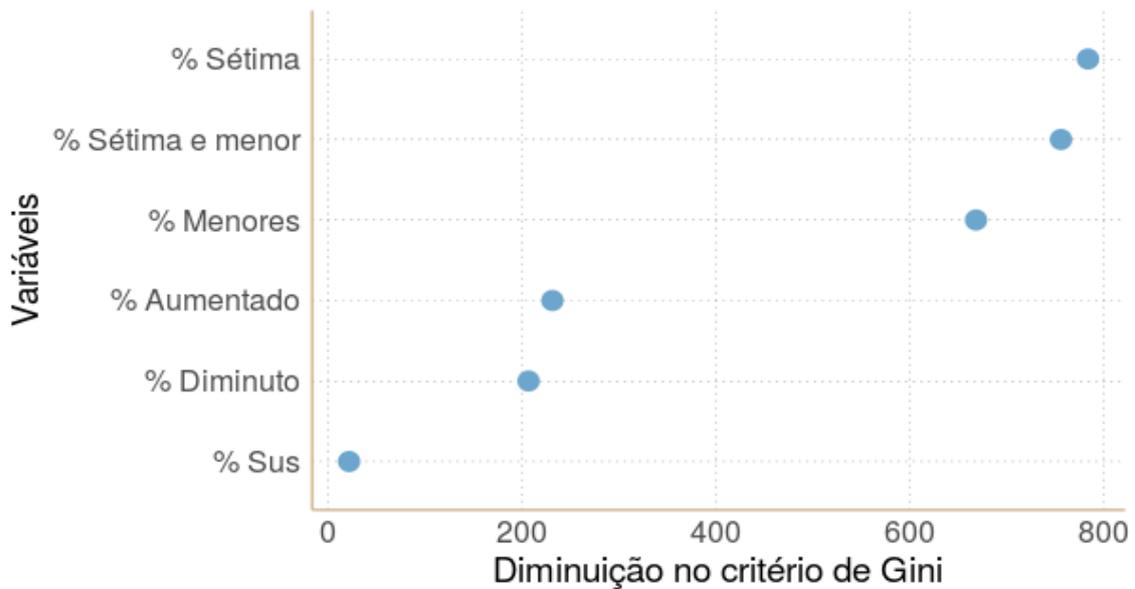


Figura 18 – Importância das variáveis na classificação dos gêneros, para o primeiro modelo.

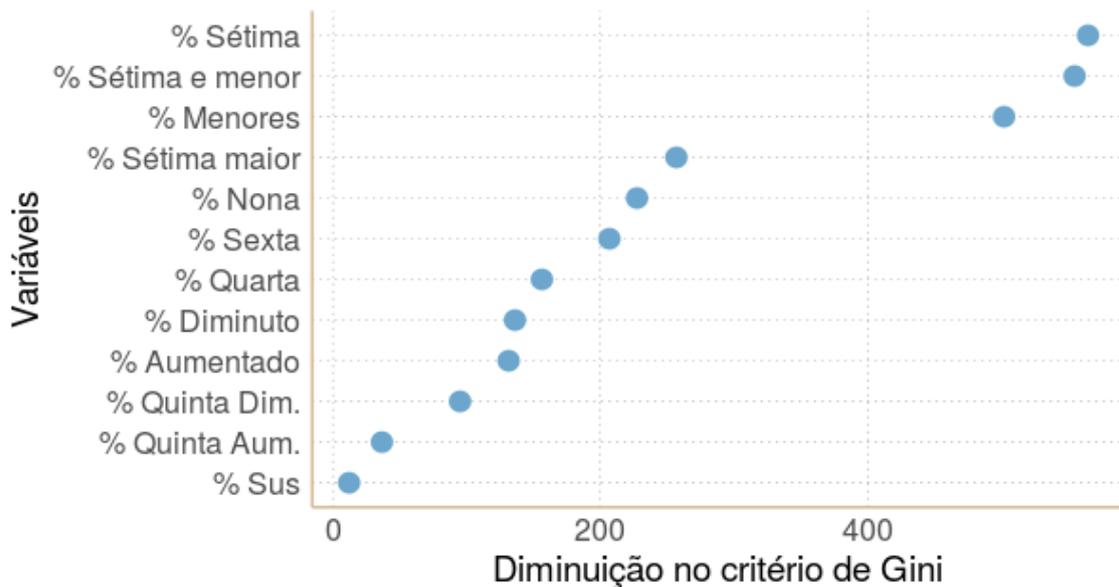


Figura 19 – Importância das variáveis na classificação dos gêneros, para o segundo modelo.

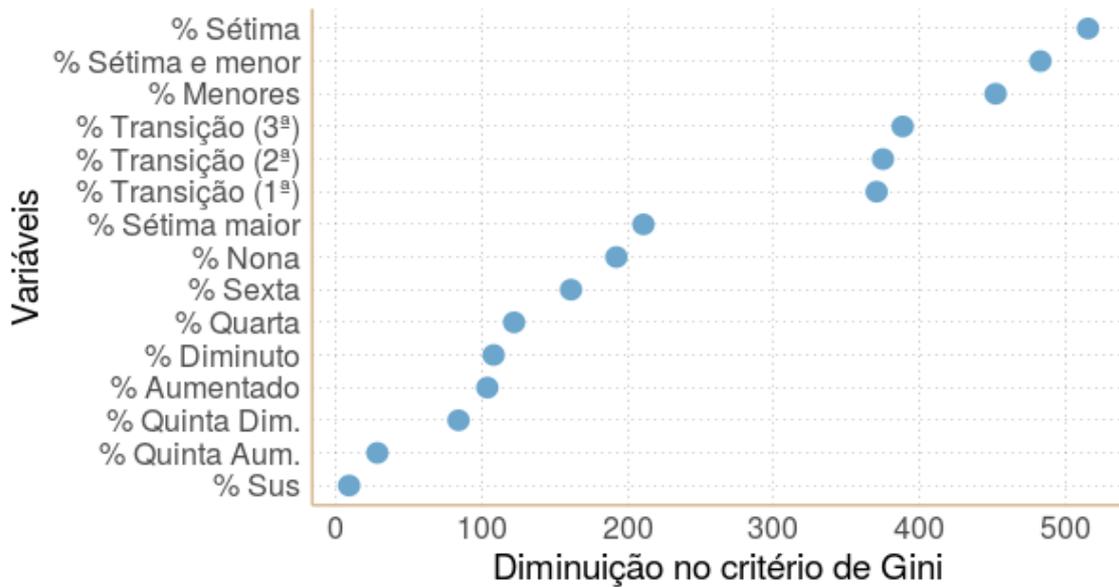


Figura 20 – Importância das variáveis na classificação dos gêneros, para o terceiro modelo.

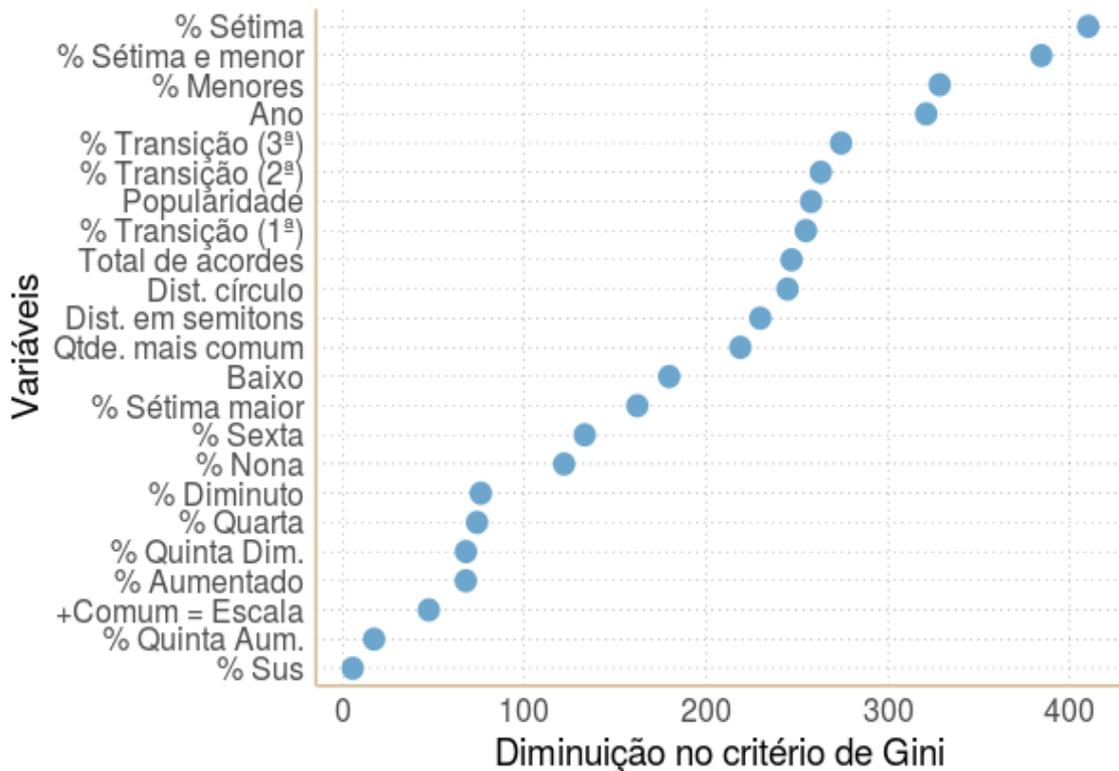


Figura 21 – Importância das variáveis na classificação dos gêneros, para o quarto modelo.

O primeiro modelo tem três variáveis bem significativas: a porcentagem de acordes com sétima, com sétima e menor e de acordes menores. Na adição do primeiro conjunto de variáveis ao segundo, esta situação se manteve. Logo, pode-se entender que o segundo conjunto de características não é tão discriminante quanto o primeiro, para este conjunto de dados. Do segundo ajuste para o terceiro, as mesmas variáveis continuaram nas três

primeiras posições, mas as novas variáveis, do terceiro conjunto, tomaram o lugar das que haviam entrado no segundo passo. Este aspecto mostra que as variáveis do terceiro conjunto trazem mais informação para o modelo do que as do segundo. Por fim, com a inclusão do último conjunto de variáveis, a mais relevante entre estas é o ano de lançamento do álbum das músicas, tomando o quarto lugar no gráfico de importância. Algumas outras variáveis do quarto conjunto ocuparam lugares de relevância, antes ocupados por variáveis de outros conjuntos, como a sétima (Popularidade) e nona (Total de acordes) posição.

Conclui-se, então, que é possível caracterizar os gêneros da música brasileira pela estrutura harmônica de suas músicas. As variáveis mais importantes para isso são as porcentagens de acordes com sétima, acordes com sétima e menores, acordes menores, o ano das músicas, a popularidade, a distribuição das transições mais comuns das canções e a quantidade absoluta de acordes. Neste grupo, prevalecem as características que podem ser extraídas puramente com as informações sobre quais acordes estão presentes em cada música. As variáveis restantes são aquelas obtidas por meio de consultas à API do Spotify, o que é bem fundamentado, uma vez que este é o principal software de *streaming* musical da atualidade e suas informações são confiáveis.

5 Considerações Finais

Os objetivos principais foram obter dados de forma automatizada, construir um pacote que implemente a extração, e realizar a análise exploratória e modelagem deles. Estes dados são as cifras de uma amostra de músicas brasileiras, complementados com algumas variáveis da API do Spotify. Com isso, o foco era mostrar que gêneros musicais brasileiros podem ser caracterizados por sua estrutura harmônica. Estes objetivos foram alcançados de forma satisfatória.

O pacote *chordrds* se manterá em constante desenvolvimento. Sua disponibilização à comunidade, feita através do CRAN, é de fundamental importância tanto para o próprio pacote, como para o crescimento das aplicações de análise musical com o R de forma geral.

O planejamento sobre trabalhos futuros envolve melhorias no pacote e, especialmente, estudos complementares sobre outras possibilidades de criação e interpretação de variáveis.

Vale destacar que a construção do pacote e uso das técnicas de análise exploratória, extração de características e modelagem são importantes não só para a obtenção das conclusões, mas porque motivaram um profundo aprendizado sobre os métodos. Apesar das aplicações práticas do estudo ainda serem limitadas, pretende-se, com ele, ampliar a percepção sobre sua utilidade.

Referências

- AGARWAL, B.; MITTAL, N. *Prominent Feature Extraction for Sentiment Analysis*. [S.l.]: Springer International Publishing, 2015. (Socio-Affective Computing). ISBN 9783319253435.
- ALMADA, C. *Harmonia Funcional*. [S.l.: s.n.], 2012. v. 1.
- ASHARI, A.; PARYUDI, I.; TJOA, A. M. Performance comparison between naïve bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Citeseer, v. 4, n. 11, 2013.
- BASILI, R.; SERAFINI, A.; STELLATO, A. Classification of musical genre: a machine learning approach. In: *ISMIR*. [S.l.: s.n.], 2004.
- BELLOT, D. *Learning Probabilistic Graphical Models in R*. [S.l.]: Packt Publishing Ltd, 2016.
- BEN-GAL, I. et al. *Bayesian networks, encyclopedia of statistics in quality and reliability*. [S.l.]: Wiley & Sons, 2007.
- BINDER, J. et al. Adaptive probabilistic networks with hidden variables. *Machine Learning*, Springer, v. 29, n. 2-3, p. 213–244, 1997.
- CALDAS, W. *Iniciação à Música Popular Brasileira*. [S.l.: s.n.], 2010. v. 1.
- CHERLA, S. et al. Learning distributed representations for multiple-viewpoint melodic prediction. 2013.
- CHIAPPA, S.; BENGIO, S. *HMM and IOHMM modeling of EEG rhythms for asynchronous BCI systems*. [S.l.], 2003.
- CHIASSON, T.; GREGORY, D. *Data + Design: A Simple Introduction to Preparing and Visualizing Information*. [S.l.]: Infoactive, 2014. ISBN 9781787286474.
- CIFRA Club - seu site de cifras e tablaturas. <<https://www.cifraclub.com.br/>>.
- CORREA, D. C.; COSTA, L. d. F.; SAITO, J. H. *Tracking the Beat: Classification of Music Genres and Synthesis of Rhythms*. [S.l.]: IWSSIP, 2010.
- CRUZ, F. W. et al. A brazilian popular music digital library oriented to musical harmony e-learning. 2004.
- Ese Xxxx, T. N. et al. Probabilistic Models for Music JEAN-FRANÇOIS PAIEMENT. 2008.
- FARAWAY, J. J. *Extending the Linear Model with R (Texts in Statistical Science)*. [S.l.]: Chapman & Hall/CRC, 2005. ISBN 158488424X.
- FARID, D. M.; HARBI, N.; RAHMAN, M. Z. Combining naive bayes and decision tree for adaptive intrusion detection. *arXiv preprint arXiv:1005.4496*, 2010.

- FOGEL, A. R. et al. Studying musical and linguistic prediction in comparable ways: The melodic cloze probability method. *Frontiers in psychology*, Frontiers, v. 6, p. 1718, 2015.
- FRIEDMAN, N. et al. Learning probabilistic relational models. In: *IJCAI*. [S.l.: s.n.], 1999. v. 99, p. 1300–1309.
- GUYON, I.; ELISSEEFF, A. An introduction to feature extraction. In: *Feature extraction*. [S.l.]: Springer, 2006. p. 1–25.
- Hastie, Trevor, Tibshirani, Robert, Friedman, J. *The Elements of Statistical Learning The Elements of Statistical Learning Data Mining, Inference, and Prediction, Second Edition*. [S.l.: s.n.], 2009. 282 p. ISSN 00111287. ISBN 978-0-387-84858-7.
- KOLLER, D.; FRIEDMAN, N. *Probabilistic Graphical Models: Principles and Techniques*. [S.l.: s.n.], 2009. v. 2009. 1231 p. ISSN 15582264. ISBN 0262013193.
- KUHN, M.; JOHNSON, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. [S.l.: s.n.], 2018.
- LAURITZEN, S. *Graphical Models*. [S.l.]: Clarendon Press, 1996. (Oxford Statistical Science Series). ISBN 9780191591228.
- LEÓN, P. J. P. D.; IÑESTA, J. M.; RIZO, D. Mining digital music score collections: melody extraction and genre recognition. In: *Pattern Recognition Techniques, Technology and Applications*. [S.l.]: Intech, 2008.
- MASON, D. J. The beatles genome project: Cluster analysis and visualization of popular music.
- MERKELBACH, K. Feature extraction for musical genre classification mus-15. Citeseer, 2015.
- MIERSWA, I.; MORIK, K. Automatic feature extraction for classifying audio data. *Machine learning*, Springer, v. 58, n. 2-3, p. 127–149, 2005.
- NARGESIAN, F. et al. Learning feature engineering for classification. In: AAAI PRESS. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. [S.l.], 2017. p. 2529–2535.
- NOLL, T.; MONTIEL, M. Mathematics and computation in music. *Glarean's Dodecachordon Revisited. Lecture Notes in Computer Science*, Springer, v. 7937.
- OZDEMIR, S.; SUSARLA, D. *Feature Engineering Made Easy: Identify unique features from your dataset in order to build powerful machine learning systems*. [S.l.]: Packt Publishing, 2018. ISBN 9781787286474.
- PAIEMENT, J.-F. Probabilistic models for music. EPFL, 2008.
- PAIEMENT, J.-F.; BENGIO, S.; ECK, D. Probabilistic models for melodic prediction. *Artificial Intelligence*, Elsevier, v. 173, n. 14, p. 1266–1274, 2009.
- PAIEMENT, J.-F.; ECK, D.; BENGIO, S. A probabilistic model for chord progressions. In: *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR)*. [S.l.: s.n.], 2005.

- RAPHAEL, C.; STODDARD, J. Harmonic analysis with probabilistic graphical models. Johns Hopkins University, 2003.
- REAL, E. Feature extraction and sufficient statistics in detection and classification. v. 6, p. 3049 – 3052 vol. 6, 06 1996.
- REAL, E. C. Feature extraction and sufficient statistics in detection and classification. In: IEEE. *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on.* [S.l.], 1996. v. 6, p. 3049–3052.
- REYNOLDS, S. Post-rock. *Audio culture: Readings in modern music.* New York: Continuum, 1995.
- RIZO, D. et al. A pattern recognition approach for melody track selection in midi files. 2006.
- ROLLAND, P.-Y. Discovering patterns in musical sequences. *Journal of New Music Research*, Taylor & Francis, v. 28, n. 4, p. 334–350, 1999.
- ROSCA, M. et al. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.
- SCARINGELLA, N.; ZOIA, G.; MLYNEK, D. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, IEEE, v. 23, n. 2, p. 133–141, 2006.
- SEEGER, A. Etnografia da música. *Cadernos de Campo (São Paulo, 1991)*, v. 17, n. 17, p. 237–260, 2008.
- SILVA, F.; ESTRADA, J. *Hino Nacional do Brasil.* [S.l.]: Fundação Nacional de Artes - Funarte, 2004.
- SPILIOPOULOU, A.; STORKEY, A. Comparing probabilistic models for melodic sequences. In: SPRINGER. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases.* [S.l.], 2011. p. 289–304.
- SPOTIFY AB, Regeringsgatan 19, SE-111 53 Stockholm, Sweden. [S.l.: s.n.].
- TASKAR, B.; ABBEEL, P.; KOLLER, D. Discriminative probabilistic models for relational data. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence.* [S.l.], 2002. p. 485–492.
- TEAM, R. R Development Core Team. *R: A Language and Environment for Statistical Computing*, v. 55, p. 275–286, 2013. ISSN 3-900051-07-0.
- VIRTANEN, T. Unsupervised learning methods for source separation in monaural music signals. In: *Signal Processing Methods for Music Transcription.* [S.l.]: Springer, 2006. p. 267–296.
- WILKINSON, L. The grammar of graphics. In: *Handbook of Computational Statistics.* [S.l.]: Springer, 2012. p. 375–414.
- WUNDERVALD, B. *GitHub - Bruna Wundervald.* 2018. Disponível em: <<https://github.com/brunaw/chorrrds>>.

WUNDERVALD, B.; TRECENTI, J. *R-Music*. 2018. Disponível em: <<https://github.com/r-music>>.

ZHANG, B.; HSU, M.; FORMAN, G. Accurate recasting of parameter estimation algorithms using sufficient statistics for efficient parallel speed-up. In: SPRINGER. *European Conference on Principles of Data Mining and Knowledge Discovery*. [S.l.], 2000. p. 243–254.