UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM

INSTITUTO DE COMPUTAÇÃO - ICOMP

PROGRAMA PÓS-GRADUAÇÃO EM INFORMÁTICA - PPGI

Um Modelo para Previsão do Sucesso no Mercado Musical

Carlos Vicente Soares Araujo

Manaus - AM

Dezembro de 2019

Carlos Vicente Soares Araujo

Um Modelo para Previsão do Sucesso no Mercado Musical

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Amazonas, como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

Orientador

Prof. Dr. Rafael Giusti

Universidade Federal do Amazonas - UFAM

Instituto de Computação - IComp

Manaus - AM

Dezembro de 2019

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

A663u

Araújo, Carlos Vicente Soares
Um Modelo para Previsão do Sucesso no Mercado Musical / Carlos Vicente Soares Araújo. 2019
93 f.: il. color; 31 cm.

Orientador: Rafael Giusti
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Sucesso Musical. 2. Aprendizagem de Máquina. 3. Música. 4. Ciência dos Dados. I. Giusti, Rafael II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Um Modelo para Previsão do Sucesso no Mercado Musical"

CARLOS VICENTE SOARES ARAUJO

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Rafael Giusti - PRESIDENTE

Prof.ª Eulanda Miranda dos Santos - MEMBRO INTERNO

Prof. Diego Furtado Silva - MEMBRO EXTERNO

Manaus, 13 de Dezembro de 2019

Para meus pais.

Agradecimentos

Estes agradecimentos não poderiam começar com outras pessoas que não fossem meus pais. Se esse texto tornou-se realidade muito devo aos dois que fizeram tudo ao alcance para me proporcionarem a melhor educação dentro do possível! Todo meu esforço é para recompensar tudo o que fizeram e fazem por mim! Não à toa esse texto está dedicado a eles.

Outra pessoa que merece todos os agradecimentos possíveis é o meu orientador Prof. Dr. Rafael Giusti. Não nego que quando começamos a trabalhar juntos eu tinha minhas reticências por não o conhecer, entretanto estas preocupações se mostraram infundadas. Construímos juntos uma incrível parceria, que rendeu frutos a partir de nosso árduo trabalho. Se essa pesquisa hoje é realidade muito devo à nossa colaboração!

Também agradeço a todo o corpo docente do Instituto de Computação da universidade, mesmo aqueles com quem nunca tive aula ou qualquer interação, pois sei que a colaboração entre eles que traz excelência a esse instituto apesar de suas limitações. Destaco principalmente os professores Eduardo Nakamura, que me orientou na graduação, e Marco Cristo, que foi meu professor de Aprendizagem de Máquina e que também ofereceu suporte em alguns aspectos desta pesquisa.

Na reta final de minha caminhada sofri um grave acidente enquanto dirigia, só cheguei ao fim dessa trajetória graças ao inventor do cinto de segurança e todos aqueles que ajudaram a aprimorar esse item. Para todos eles ficam também os meus agradecimentos.

Fico feliz que você tá substituindo esse trauma por outro.

Cerginho da Pereira Nunes

Um Modelo para Previsão do Sucesso no Mercado Musical

Autor: Carlos Vicente Soares Araujo

Orientador: Prof. Dr. Rafael Giusti

Resumo

O mercado musical é extremamente competitivo e movimenta bilhões de dólares todos os anos. Só nos Estados Unidos, existem mais de 1400 selos musicais atualmente registrados. Destacar-se nesse cenário é uma árdua missão. Neste trabalho, apresentamos um modelo de previsão do sucesso no mercado musical que pode ser usado por artistas e gravadoras para focar seus esforços em músicas com maior tendência a obter retorno comercial. O modelo proposto utiliza informações sobre as músicas para prever, antes mesmo de seus lançamentos, se irão ou não aparecer no *ranking* Top 50 Global da plataforma de *streaming* Spotify. Para validação do modelo, nós adotamos como *baseline* o trabalho mais semelhante ao nosso já estabelecido na literatura científica. Esse *baseline* utiliza o mesmo tipo de informação que utilizamos, mas com uma abordagem distinta em relação à preparação da base. Nossos resultados chegaram a ser 920% superiores aos obtidos pelo *baseline*.

Palavras-chave: Sucesso Musical, *Hit Song Science*, Aprendizagem de Máquina.

Um Modelo para Previsão do Sucesso no Mercado Musical

Autor: Carlos Vicente Soares Araujo

Orientador: Prof. Dr. Rafael Giusti

Abstract

The music market is extremely competitive and moves billions of dollars every year. In the United States alone, there are over 1400 music labels currently registered. Standing out in this scenario is an arduous mission. In this dissertation, a model for success prediction in the music market is presented, which artists and record labels may use to direct their efforts into songs with higher potential to return profit. The proposed model makes avail of information about the songs to predict, even before their release, whether they will appear or not in the Top 50 Global ranking from the streaming platform Spotify. To validate this model, we chose as a baseline the most similar model already consolidated in the scientific literature. The baseline employs the kind of information we have employed in our model, but with a distinct approach with respect to the data preparation. Our results are 920% better than those achieved by the baseline.

Keywords: Music Success, Hit Song Science, Machine Learning.

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AUC	Área Abaixo da Curva de Característica de Operação do Receptor
BNB	Naive Bayes Bernoulli
DT	Árvore de Decisão
EQM	Erro Quadrático Médio
ENIAC	Encontro Nacional de Inteligência Artificial e Computacional
FDA	Função de Distribuição Acumulada
GNB	Naive Bayes Gaussiano
HSS	*Hit Song Science*
ICMLA	IEEE International Conference on Machine Learning and Applications
IFPI	International Federation of the Phonographic Industry
LR	Regressão Logística
MCC	Coeficiente de Correlação de Matthews
MFCC	*Mel-Frequency Cepstral Coefficients*
MIR	*Music Information Retrieval*
MLP	Perceptron Multicamadas
NAR	Redes Neurais Autorregressivas Não-Lineares
NARX	Redes Neurais Autorregressivas Não-Lineares com Entrada Externa
NB	Naive Bayes

OCC	Official Charts Company
P-P	Probabilidade-Probabilidade
RF	Floresta Aleatória
ROC	Característica de Operação do Receptor
SBCM	Simpósio Brasileiro de Computação Musical
SVM	Máquina de Vetores de Suporte
VPN	Valor Preditivo Negativo

Sumário

1 Introdução

A forma como as pessoas vêm consumindo música está mudando. Em 2018, pela primeira vez, o *streaming* passou a ser a principal forma de consumo de música, representando 47% do mercado musical, segundo o relatório anual da International Federation of the Phonographic Industry (IFPI)[1]. Logo, o *streaming* passou a ser fundamental para que artistas e gravadoras obtenham bons resultados comerciais.

Uma possível forma para auxiliar esses artistas e gravadoras a maximizar o retorno é utilizar um modelo que consiga prever se essas músicas obterão destaque nas plataformas de *streaming*. Para se ter uma noção do impacto comercial que este modelo pode apresentar, basta-se dizer que o mercado musical, considerando somente o arrecadado com consumo e licenciamento de músicas, movimentou US$ 19,1 bilhões ao redor do globo em 2018, segundo a IFPI. Desenvolvemos um modelo para esse fim e o seu processo de criação e validação será apresentado neste trabalho.

O restante deste capítulo está organizado da seguinte forma: na Seção 1.1 é apresentado o contexto em que esta pesquisa está inserida; a motivação para tal estudo está na Seção 1.2. A descrição do problema é dada na Seção 1.3 e a hipótese na qual este estudo foi baseado está na Seção 1.4. Na Seção 1.5, os objetivos são apresentados, e as contribuições obtidas são vistas na Seção 1.6.

1.1 Contexto

Esta pesquisa está inserida na área de *Hit Song Science* (HSS), termo ainda sem tradução para o português[2]. A HSS estuda formas de prever o sucesso de músicas antes mesmo delas serem disponibilizadas no mercado. Logo, é uma

[1] <www.ifpi.org/downloads/GMR2019.pdf>

[2] Sem tradução oficial. Uma possível tradução livre seria “ciência de músicas de sucesso”.

importante área para artistas, selos e gravadoras musicais para que possam traçar ações que obtenham maior retorno financeiro (PACHET, 2011).

A HSS teve uma história inicial conturbada. O termo foi introduzido em 2003 pela empresa "Polyphonic HMI", que afirmava conseguir prever o sucesso de um álbum com meses de antecedência em relação ao seu lançamento. Atualmente, é difícil encontrar informações sobre a empresa, o que pode indicar que não conseguiram obter resultados satisfatórios. Além disso, Pachet e Roy (2008) afirmaram ser impossível realizar previsões de popularidade de músicas, concluindo que a HSS não poderia ser considerada ciência até aquele momento. Esse trabalho será melhor explanado no Capítulo 3.

Entretanto, tal entendimento já não é mais válido na atual década devido a mudanças na forma do consumo de música e avanços da área de aprendizagem de máquina, tornando viável a previsão de sucesso no mercado musical. Essa conclusão foi dada por um dos autores do trabalho apresentado anteriormente (PACHET; ROY, 2008) no capítulo específico sobre HSS (PACHET, 2011) no livro "Music Data Mining" (LI; OGIHARA; TZANETAKIS, 2011), um dos principais da área de *Music Information Retrieval* (MIR), que significa recuperação de informações musicais em português e onde HSS está inserida. Além disso, Ni et al. (2011) defendem que HSS é ciência ao apresentar um modelo que prevê se uma música ficará entre as cinco primeiras posições do *ranking* da Official Charts Company (OCC)[3]. Esse modelo foi treinado com características acústicas das músicas não utilizadas no trabalho de Pachet e Roy (2008) e obteve acurácia sempre superior ao de uma decisão aleatória. No Capítulo 3 também abordaremos esse trabalho.

1.2 Motivação

A motivação para a realização deste trabalho foi dada levando em consideração dois diferentes fatores: o comercial e o científico.

[3] <https://www.officialcharts.com/>

O fator comercial é decorrente do fato do mercado musical ter movimentado US$ 19,1 bilhões de dólares em 2018 e ser extremamente competitivo. Para se ter uma noção, a Wikipedia cataloga 1411 selos musicais somente nos Estados Unidos[4]. Além disso, o site Data Usa afirma que há 138000 artistas em terras americanas[5], isso considerando somente aqueles registrados em algum órgão competente. Ao utilizar um modelo de previsão, eles podem dar foco maior para as músicas que apresentam maior tendência a obterem bons rendimentos.

Tratando-se agora do aspecto científico, HSS é uma área de pesquisa recente, onde diferentes abordagens vêm sendo utilizadas buscando gerar modelos que obtenham previsões cada vez melhores. Em HSS é comum que trabalhos sejam publicados com o intuito de indicar metodologias que necessitam de maior aprofundamento ou que não sejam recomendadas (REIMAN; ÖRNELL, 2018; ARAKELYAN et al., 2018). Neste trabalho, tomamos como base a metodologia de um desses trabalhos e a aprimoramos de forma a conseguir realizar melhores previsões.

1.3 Descrição do Problema

O escopo de HSS é bem específico. Um trabalho só pode ser considerado da área se apresenta um modelo que realiza previsões de popularidade antes mesmo do lançamento da música ou álbum. Por exemplo, uma pesquisa que busca identificar se um álbum já lançado irá ou não cair em um *ranking* (ABEL et al., 2010) não pode ser considerada HSS.

Modelos de HSS são em geral de aprendizagem de máquina, logo são necessários dados para treinar esses modelos. Em nossos estudos, identificamos três tipos de fontes de dados como as mais comuns, sendo elas: características acústicas das músicas (LEE; LEE, 2018; KARYDIS et al., 2018), informações

[4] <https://en.wikipedia.org/wiki/Category:American_record_labels>. Acesso às 18:46 do dia 05/11/2019.

[5] <https://datausa.io/profile/soc/272040/>. Acesso às 18:47 do dia 05/11/2019.

de redes sociais (KIM; SUH; LEE, 2014; DHAR; CHANG, 2009) e dados sobre shows e festivais (ARAKELYAN et al., 2018; STEININGER; GATZEMEIER, 2013).

Entretanto, nenhum desses despontam como a principal fonte de dados para pesquisas na área. Logo, decidir que tipo de fonte utilizar é um dos principais desafios em HSS. Em nosso modelo, decidimos utilizar informações que poderiam ser obtidas diretamente de uma plataforma de *streaming*, as quais não são nenhuma das três fontes anteriormente mencionadas. Esses dados representam atributos de alto nível das músicas que indicam, por exemplo, se elas são dançantes, alegres, instrumentais, dentre outras possibilidades. Embora esses atributos possam ser extraídos com base em características acústicas, eles também podem ser determinados por um especialista ou pelo próprio artista.

1.4 Hipótese

A seguinte hipótese foi utilizada como ponto de partida desta pesquisa: **É possível prever se uma música irá ou não aparecer no *ranking* Top 50 Global do Spotify antes mesmo de ser lançada.** O Spotify foi escolhido como nosso estudo de caso por ser o serviço global de *streaming* de músicas com maior número de usuários, atrás somente do Soudcloud, que não conta com músicas de artistas e gravadoras renomadas[6,7]. O Top 50 Global é um *ranking* diário que apresenta as 50 músicas que tiveram o maior número de reproduções no dia anterior na plataforma.

Uma etapa importante de trabalhos em HSS é estabelecer um critério para determinar o que é sucesso no mercado musical. Neste trabalho, determinamos que uma música alcança sucesso ao aparecer em um *ranking* qualificado, pois é uma abordagem já consolidada em outros trabalhos da área (HERREMANS; MARTENS; SöRENSEN, 2014; REIMAN; ÖRNELL, 2018). Especificamente, consideramos que uma música obtém sucesso ao aparecer

[6] <http://bit.ly/2KwJmGu>
[7] <http://bit.ly/2QrRGLs>

no *ranking* Top 50 da plataforma Spotify.

1.5 Objetivos

O objetivo geral desta pesquisa é **realizar previsões de sucesso de músicas em plataformas de *streaming* antes mesmo de serem lançadas**.

Para atingir tal objetivo, os seguintes objetivos específicos são necessários:

- Desenvolver e validar um modelo de previsão que utiliza informações sobre características das músicas;
- Desenvolver um modelo baseado em metodologia de trabalho semelhante e consolidado da área a fim de comparar os resultados obtidos. Isso é necessário devido à impossibilidade de aplicarmos dados coletados em modelos já desenvolvidos e publicados.

1.6 Contribuições

Do desenvolvimento do modelo proposto, as seguintes contribuições foram obtidas na realização deste trabalho:

- Criamos um modelo que consegue prever se uma música viral irá se tornar popular no Spotify e vice-versa;
- Criamos um modelo que prevê se uma música já popular continuará a ser popular;
- Também desenvolvemos um modelo que consegue prever quais músicas dentro de um conjunto se tornarão populares;
- Mostramos que é possível realizar previsões de sucesso de músicas em plataformas de *streaming* ao desenvolver e validar um modelo que obteve resultados superiores ao alcançado em outro trabalho da literatura.

Além dessas contribuições, como parte dos estudos desenvolvidos ao longo desta pesquisa, demonstramos que mensagens com polaridade positiva no Twitter têm correlação com a popularidade de um álbum no Spotify. Também identificamos que o surgimento de um ato musical ou um movimento de destaque é o principal fator de influência na popularidade de um gênero musical e observamos que um artista ou banda podem alavancar sozinhos a popularidade de um gênero por um período de até cinco anos. Além disso, constatamos que o Pop foi o gênero que apresentou maior tendência a ter a música mais popular de um ano durante as décadas de 1960 a 2010. Por fim, verificamos que, no Rap, novos artistas não apresentaram tendência a colaborarem com artistas consagrados. Discutimos esses estudos no Apêndice A.

1.7 Organização do Documento

O restante deste documento está organizado da seguinte forma:

- No Capítulo 2 estão fundamentos teóricos que possibilitam a melhor compreensão do trabalho aqui descrito;
- No Capítulo 3 abordamos trabalhos relacionados ao realizado nesta pesquisa;
- No Capítulo 4 expomos modelos de previsão preliminares que resultaram em artigos publicados;
- No Capítulo 5 apresentamos o modelo proposto, com a metodologia utilizada para seu desenvolvimento e os resultados que obtivemos;
- Por fim, no Capítulo 6 realizamos considerações finais e apontamos possíveis trabalhos futuros.

2 Fundamentação Teórica

Neste capítulo, conceitos e definições necessários para melhor entendimento do trabalho são expostos, sendo estes divididos em duas seções. Na Seção 2.1 estão os fundamentos musicais, enquanto os fundamentos matemáticos e computacionais estão na Seção 2.2.

2.1 Fundamentos Musicais

HSS é uma área de estudo da computação, afinal os modelos de previsão são modelos computacionais. Mas, para a geração destes, conceitos da área musical devem ser compreendidos:

- ***Single*:** Música lançada separadamente em que se espera bom resultado comercial (ATWOOD, 2017).
- **Selo Musical:** Marca ou empresa responsável pelo lançamento e divulgação de músicas de artistas associados a ela. Pode também ser responsável por gerenciar carreiras de seus artistas, assim como também pode ser filiada a gravadoras maiores (KLEIN, 2003).
- **Gênero Musical:** Categoria que identifica músicas como pertencentes a uma tradição compartilhada ou um conjunto de convenções (SAMSON, 2001).
- **Metadados Musicais:** Informações que descrevem gravações de músicas específicas. De forma geral, a maioria dos arquivos de áudio suporta uma estrutura conhecida como ID3, que é projetada para armazenar informações como o nome do artista, nome da faixa, descrição da música e título do álbum (LI; LI, 2011).
- **Características Acústicas:** Quaisquer propriedades acústicas de um áudio que podem ser gravadas e analisadas. Por exemplo, quando uma

orquestra sinfônica está tocando a 9ª Sinfonia de Beethoven, cada instrumento musical, com exceção de algumas percussões, produz diferentes vibrações periódicas. Em outras palavras, os sons produzidos pelos instrumentos musicais são o resultado da combinação de diferentes frequências (ORIO, 2006). As características acústicas utilizadas neste trabalho são listadas na Tabela 1. Elas foram selecionadas por serem características já consolidadas em outros trabalhos da literatura (KARYDIS et al., 2018; HERREMANS; MARTENS; SöRENSEN, 2014).

Tabela 1 – Algumas características acústicas.

Característica Acústica	**Descrição**
Mel-Frequency Cepstral Coefficients (MFCC)	Obtido a partir do espectro da representação mel comprimida do sinal. O MFCC é provavelmente um dos recursos mais usados no processamento de fala e é uma representação expressiva de baixa dimensão de um sinal (MCFEE et al., 2015).
Centroide Espectral	O centroide de cada quadro de um espectrograma de magnitude que foi normalizado e tratado como uma distribuição sobre intervalos de frequência (MCFEE et al., 2015).
Coeficiente de Tonalidade	Uma medida de quanto um som se parece com um tom, em vez de ruído (DUBNOV, 2004).
Passagem por Zero	O número de vezes que uma forma de onda muda de sinal (MCFEE et al., 2015).
Tempo	Número de batidas por minuto (MCFEE et al., 2015).

- ***Tags* Sociais:** Uma coleção de informações textuais que anotam diferentes itens musicais, como álbuns, músicas e artistas. *Tags* sociais são criadas por fãs, contendo uma grande quantidade de informações, incluindo qualidade, emoção ou uma descrição simples. Essas *tags* são geralmente utilizadas para facilitar a busca, encontrar músicas semelhantes e ouvintes com mesmos interesses. O *website* Last.fm[1] tem um dos sistemas de marcação social mais conhecidos (LAMERE, 2008).

Além dos *rankings* do Spotify, mencionamos aqui também os da revista Billboard. Os *rankings* da Billboard são publicados desde 1940 e são um dos

[1] <http://last.fm/>

mais relevantes na indústria. Muitos trabalhos relacionados a serem apresentados utilizam-se desses rankings como parâmetros de sucesso. Destacam-se o *Hot 100 Songs*, lançado semanalmente desde 1955, contendo as 100 músicas mais populares da semana, e o *Billboard 200*, que conta com os 200 álbuns mais populares da semana, sendo lançado desde 1967. Há ainda versões de fim-de-ano delas, além de dezenas de outros *rankings* voltados a gêneros, países e formas de consumo específicas[2] (MOLANPHY, 2013).

O ponto negativo dos *rankings* da Billboard é que eles são feitos considerando apenas o mercado estado-unidense. Em contraste, os *rankings* do Spotify utilizam dados providos por seus clientes em todos os países onde operam. Tal fator foi decisivo na escolha dos *rankings* do Spotify como parâmetros de sucesso neste trabalho.

2.2 Fundamentos Matemáticos e Computacionais

Nesta seção, fundamentos matemáticos e computacionais utilizados nesta pesquisa são apresentados, estando ela dividida em três subseções. A Subseção 2.2.1 é dedicada a apresentar coeficientes de correlação que foram utilizados em trabalhos relacionados e prévios ao desenvolvimento do modelo. Na Subseção 2.2.2 discutimos os algoritmos de aprendizagem de máquina utilizados nesta pesquisa. Por fim, na Subseção 2.2.3 estão as métricas utilizadas para avaliação dos modelos.

2.2.1 Coeficientes de Correlação

Coeficientes de correlação são empregados como ferramentas para identificar associações entre duas variáveis aleatórias na pesquisa experimental (BHATTACHARJEE, 2014). Embora o modelo final não faça uso desses coeficientes, entender como se dão as associações entre dados de redes sociais e do mercado

[2] <https://www.billboard.com/charts>

musical foi necessário para pesquisas anteriores ao desenvolvimento do modelo proposto em si, estas são apresentadas no Apêndice A. Além disso, trabalhos relacionados que serão discutidos no Capítulo 3 também fazem uso desses coeficientes.

Nós empregamos três coeficientes de correlação, que podem ser diferenciados por parametrização e tipo de correlação capturada.

A correlação de Pearson (1895) quantifica a correlação linear entre duas variáveis, variando entre -1 (correlação linear total e negativa) e $+1$ (correlação linear total e positiva). O valor 0 indica que não há correlação linear, mas não garante a independência entre as variáveis. Ela pode ser calculada como

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{2.1}$$

onde $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_n\}$, cov(X,Y) é a covariância entre X e Y, σ_X é o desvio padrão de X e σ_Y o desvio padrão de Y.

O coeficiente de correlação de postos de Spearman (1904) é uma estatística de classificação não paramétrica, proposta como medida da força da associação entre duas variáveis. É uma medida de associação monótona, que avalia o quão bem uma função monotônica arbitrária pode descrever a relação entre duas variáveis, sem fazer hipóteses sobre a distribuição de frequência das variáveis, e é definida por

$$\rho = \frac{6 \sum d_i^2}{n^3 - n}, \tag{2.2}$$

onde d_i é a diferença entre cada posto (posição) de valor correspondentes de X e Y, e n é o número dos pares dos valores. Seu valor varia de -1 (função monótona decrescente perfeita) a $+1$ (função monótona crescente perfeita). O valor 0 indica que não há uma função monótona que relacione as duas variáveis, mas também não garante a independência das variáveis.

A correlação de distância (SZÉKELY et al., 2007) é uma medida de dependência entre vetores aleatórios e serve para verificar se há alguma relação (não necessariamente linear) entre duas variáveis (X e Y). Além disso, X e Y podem ser vetores de diferentes tamanhos. A correlação de distância é dada

por

$$\mathrm{dCor}(X,\,Y) = \frac{\mathrm{dCov}(X,\,Y)}{\sqrt{\mathrm{dVar}(X)\,\mathrm{dVar}(Y)}}, \tag{2.3}$$

onde dCov(X,Y) é a distância de covariância entre (X,Y) e é definida pela raiz quadrada de

$$\mathrm{dCov}^2 = \frac{1}{n^2}\sum_{k=1}^{n}\sum_{l=1}^{n} A_{kl}B_{kl}, \tag{2.4}$$

onde $A_{kl} = \|X_k - X_l\|$, $B_{kl} = \|Y_k - Y_l\|$, que são funções lineares simples da distância entre os pares dos elementos da base de dados (SZÉKELY; RIZZO et al., 2009), e n é o número de elementos dos vetores. A distância de variância $\mathrm{dVar}(X)$ e $\mathrm{dVar}(Y)$ também são definidas pela equação 2.4, porém utilizando o mesmo vetor de dados para A e B. A Correlação de Distância é mais robusta que as anteriores no sentido de que não assume linearidade, nem monotonicidade, e não depende do tipo de distribuição. O valor da Correlação de Distância varia de 0 a 1, sendo igual a 0 se, e somente se, as variáveis forem independentes.

2.2.2 Algoritmos de Aprendizagem de Máquina

Diferentes algoritmos de aprendizagem de máquina foram utilizados para geração do modelo proposto e também de modelos prévios. Estes serão apresentados nesta seção.

O método de Máquinas de Vetores de Suporte (SVM) foi utilizado em sua versão binária. O SVM tenta encontrar o hiperplano que melhor separa as classes, dividindo o espaço de decisão em dois subespaços. Ao classificar uma nova amostra, o SVM a projeta nesse mesmo espaço de decisão e verifica em qual subespaço a instância projetada "cai" e a atribui à classe associada a esse subespaço. Dada uma coleção de N valores, o SVM resolve o seguinte problema quadrático:

$$\min_{\omega,\varepsilon_i,\rho}\frac{1}{2}\|\omega\|^2 + \frac{1}{\nu N}\sum_i \varepsilon_i - \rho, \tag{2.5}$$

sujeito a

$$(\omega\Phi(x_i)) \geq \rho - \varepsilon_i, i = 1..N \; e \; \varepsilon_i \geq 0, i = 1..N, \tag{2.6}$$

onde ω, ε e ρ são os parâmetros para construir os hiperplanos de separação, ν serve como limite superior dos itens que não fazem parte do conjunto e o limite inferior das amostras usadas como vetores de suporte, e Φ é a função que mapeia os dados em um espaço de produto interno, de forma que os dados projetados possam ser modelados por algum *kernel* (GIUSTI; SILVA; BATISTA, 2016; SCHöLKOPF et al., 2001). As funções *kernel* utilizadas foram: RBF, polinomial, linear e sigmoide.

Um exemplo de separação entre classes utilizando SVM, em que um hiperplano de separação ótimo foi determinado ao maximizar a margem M, está representado na Figura 1.

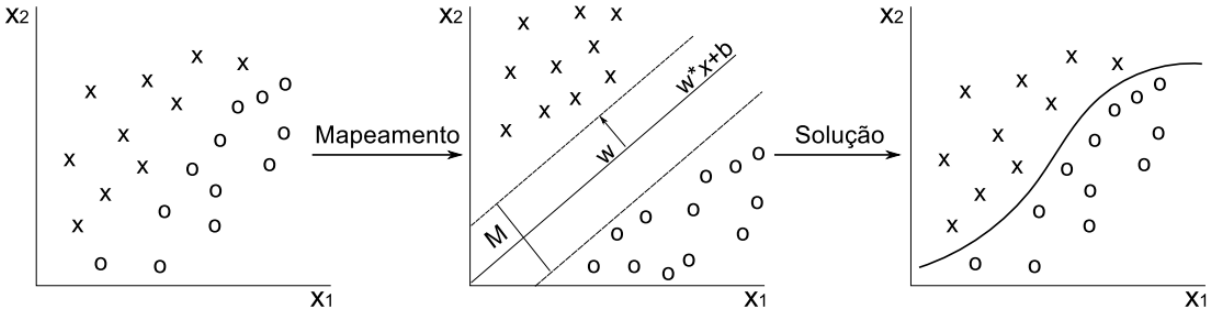


Figura 1 – Exemplo de hiperplano de separação ótima, mapeamento e solução com SVM. Retirado de (COLONNA et al., 2012).

Os métodos Naive Bayes (NB) são um conjunto de algoritmos de aprendizado supervisionado baseados na aplicação do teorema de Bayes com a suposição "ingênua" de independência entre os atributos. Os diferentes classificadores Naive Bayes diferem principalmente pelas premissas que fazem sobre a distribuição dos atributos. No algoritmo Naive Bayes Gaussiano (GNB), a distribuição dos recursos é assumida como gaussiana (ZHANG, 2004), enquanto o Naive Bayes Bernoulli (BNB) assume que os dados seguem a distribuição multivariada de Bernoulli (MANNING; RAGHAVAN; SCHÜTZE, 2010).

O classificador k-vizinhos mais próximos (KNN, do inglês k-*nearest neighbors*) é um modelo baseado em instâncias e não generalizante. Em vez de

aprender um modelo que generaliza os dados, ele memoriza todas as instâncias do conjunto de treinamento (ou um subconjunto delas). A classificação de uma nova instância em geral se dá por voto majoritário simples dos vizinhos mais próximos de cada ponto: um ponto de consulta recebe a classe que tem mais representantes nos vizinhos mais próximos ao ponto. O número de vizinhos consultados é um hiperparâmetro do modelo[3].

A Regressão Logística (LR), apesar de seu nome, é um modelo linear para classificação, em vez de regressão. Nesse modelo, as probabilidades que descrevem os possíveis resultados de um único estudo são modeladas usando uma função logística[4].

AdaBoost é um meta-classificador para *boosting* e produz modelos que são comitês de classificadores. Ele induz uma série de classificadores em uma versão ponderada da base de treinamento. A cada iteração, as instâncias incorretamente classificadas por um modelo têm seu peso aumentado, a fim de que o modelo induzido na iteração seguinte concentre-se nos exemplos mais difíceis (FREUND; SCHAPIRE, 1997). O poder preditivo do conjunto pode ser substancialmente maior que o de um classificador individual, contanto que o classificador base seja um aprendiz fraco—isto é, que seu desempenho seja pelo menos marginalmente melhor que a escolha aleatória. Em nossos experimentos, usamos a implementação AdaBoost-SAMME (HASTIE et al., 2009), que emprega pequenas Árvores de Decisão como classificador base.

As Árvores de Decisão (DT) são modelos de aprendizado supervisionado não paramétrico. Uma DT é uma estrutura de árvore cujos nós são relações sobre os atributos e o caminho da raiz até um nó folha equivale a uma regra de decisão. Árvores de Decisão são tipicamente induzidas por particionamento do conjunto de treinamento visando maximizar uma medida de pureza[5].

Uma Floresta Aleatória (RF) é um meta-estimador que produz um conjunto de DT's utilizando várias sub-amostras do conjunto de treinamento e usa a média para melhorar a precisão e controlar o ajuste excessivo (BREIMAN,

[3] <https://scikit-learn.org/stable/modules/neighbors.html>
[4] <https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression>
[5] <https://scikit-learn.org/stable/modules/tree.html>

2001).

2.2.3 Métricas de Avaliação

Aqui são apresentadas as métricas utilizadas para avaliação dos modelos.

Utilizando o trabalho de Olson e Delen (2008) como base, as equações e definições de precisão, acurácia, revocação, especificidade, valor-f1 e Valor Preditivo Negativo serão apresentadas.

Considerando tp como o número de positivos verdadeiros, tn o número de negativos verdadeiros, fp o de falsos positivos e fn o de falsos negativos retornados pelos modelos, temos:

- **Precisão =** $\frac{tp}{tp+fp}$, a porcentagem de valores positivos corretamente previstos.
- **Acurácia =** $\frac{tp+tn}{tp+tn+fp+fn}$, a porcentagem de instâncias corretamente previstas.
- **Revocação =** $\frac{tp}{tp+fn}$, a porcentagem de positivos reais.
- **Especificidade =** $\frac{tn}{tn+fp}$, a porcentagem de negativos reais.
- **Valor-f1 =** $2\frac{\text{Precisão}*\text{Revocação}}{\text{Precisão}+\text{Revocação}}$, uma métrica que combina os valores de precisão e revocação.
- **Valor Preditivo Negativo (VPN) =** $\frac{tn}{fn+tn}$, a porcentagem de valores negativos corretamente previstos.
- ***Fall-out* =** $\frac{fp}{fp+tn}$, a porcentagem de falsos positivos (STOREY, 2003).
- **Taxa de Perda =** $\frac{fn}{tp+fn}$, a porcentagem de falsos negativos, quanto menor melhor é o resultado (STOREY, 2003).
- **Área abaixo da curva de Característica de Operação do Receptor (AUC) =** $\int_{\infty}^{-\infty} \text{Revocação(T)Especificidade'(T)}\ d\text{T}$, probabilidade do classificador classificar uma instância positiva escolhida aleatoriamente mais

alta do que uma negativa aleatoriamente escolhida, quando utiliza-se de unidades normalizadas (FAWCETT, 2006). Uma curva de Característica de Operação do Receptor (ROC) é criada ao plotar a revocação contra o *fall-out* em diferentes configurações de limite (HANLEY; MCNEIL, 1982).

- **Coeficiente de Correlação de Matthews (MCC) =** $\frac{tp*tn-fp*fn}{\sqrt{(tp+fp)*(tp+fn)*(tn+fp)*(tn+fn)}}$, mede a qualidade de uma classificação binária (MATTHEWS, 1975).

- **Erro quadrático médio (EQM) =** $\frac{1}{n}\sum_{i-1}^{n}(\mathrm{Y}_i - \mathrm{Y'}_i)^2$, onde $\mathrm{Y'}$ é um vetor de n previsões geradas a partir de uma amostra de n pontos de dados em todas as variáveis e Y é o vetor dos valores observados da variável sendo prevista (LEHMANN; CASELLA, 2006).

3 Trabalhos Relacionados

Neste capítulo, trabalhos correlatos ao realizado nesta pesquisa são apresentados e discutidos, o que ocorre nas seções 3.1 e 3.2, respectivamente.

3.1 Estado da Arte

Pachet e Roy (2008) afirmaram que HSS ainda não poderia ser considerada uma ciência, mas deixam claro que mesmo não tendo conseguido realizar previsões de sucesso, não seria válido concluir que esse tipo de previsão é impossível. Os autores utilizaram um conjunto de dados de 32000 músicas provenientes da base de dados da empresa HiFind[1]. Cada música da base possuía um indicador de popularidade que poderia assumir três valores: *low* (baixa), *medium* (média) e *high* (alta). Os autores buscavam prever esse campo. Para realizar essa previsão utilizaram um classificador SVM com *kernel* RBF treinado com três diferentes conjuntos de características sobre essas músicas. O primeiro continha 49 características acústicas das músicas extraídas do próprio arquivo delas, enquanto o segundo também contava com características acústicas, mas extraídas usando um *software* proprietário da Sony. No terceiro estavam 632 *tags* sociais extraídas da plataforma Pandora[2]. Os autores avaliaram seus resultados somente utilizando o valor-F, que, no melhor caso, atingiu 41% ao prever a popularidade *low* utilizando as *tags* sociais.

Em contraponto a esse resultado, Ni et al. (2011) buscam provar que HSS pode ser considerada ciência. Para isso, coletaram o pico alcançado por 5947 músicas no *ranking* da OCC entre as décadas de 1960 a 2010. Utilizando a API do The Echo Nest[3], coletaram seis diferentes características acústicas das

[1] Não conseguimos encontrar maiores informações sobre essa companhia, o site da empresa presente no artigo está fora do ar. A tentativa de acesso ocorreu às 19:19 do dia 13/11/2019. <http://www.hifind.com/>

[2] <https://www.pandora.com/>

[3] O The Echo Nest é a principal empresa de inteligência musical, oferecendo a desenvolvedores profundo conhecimento sobre músicas e seus fãs. <http://the.echonest.com/>

faixas. As músicas coletadas tiveram seu pico máximo ou no Top 5 do *ranking* ou entre as posições 30 a 40. O modelo criado buscava prever em quais desses dois intervalos uma música estaria. O algoritmo escolhido para realizar os experimentos foi o *Shifting Perceptron*. Os autores obtiveram acurácia mínima de 54% e máxima de 57% na tarefa de prever as músicas populares das décadas de 1960 e 1990, respectivamente. Como a quantidade de instâncias em cada uma das classes estava balanceada, os resultados obtidos foram superiores ao de uma decisão aleatória (50%), ainda que apenas marginalmente. Todavia, esse trabalho sugeriu que HSS poderia ser considerada uma ciência.

Em relação ao uso de informações sobre shows para realizar previsões, Arakelyan et al. (2018) coletaram dados a partir do site SongKick[4]. Tais dados continham a localização, lista de artistas participantes, nome do evento e um valor em popularidade do evento dado pela plataforma. Para os autores, um artista pode ser considerado popular se possui contrato com uma das seguintes gravadoras: Sony BMG, Universal Music Group ou Warner. Os selos afiliados a essas companhias também eram considerados para a obtenção de sucesso. Os autores aplicaram o método de regressão logística nos dados buscando prever se um artista faria ou não sucesso. A precisão máxima obtida foi de 39%.

Outro trabalho que também utilizou dados sobre shows, festivais e afins foi o de (STEININGER; GATZEMEIER, 2013). Para cada evento, os autores obtiveram cerca de 20 parâmetros identificados por colaboradores do Amazon Mechanical Turk (um serviço oferecido pela Amazon para contratação de colaboradores humanos para realizar tarefas de forma virtual[5]). A partir desses dados, buscavam prever se as músicas dos artistas que participaram desses eventos iriam ou não aparecer em uma lista das 500 músicas mais populares da Alemanha em 2011, ano em que os experimentos foram executados. Não há a informação de onde tal lista é publicada. Os autores conseguiram mostrar que havia correlação entre os dados com 95% de certeza estatística. Entretanto, a precisão máxima obtida foi de 43,5% utilizando a abordagem PLS-SEM.

[4] <https://www.songkick.com/>
[5] <https://www.mturk.com/>

Quanto ao uso de dados de redes sociais, Kim, Suh e Lee (2014) coletaram mensagens na rede social Twitter com as *tags*: #nowplaying, que significa "tocando agora", sua versão abreviada (#np) e #itunes, plataforma digital de venda de músicas. A partir desses dados, buscavam prever se uma música iria fazer sucesso. Para os autores, o sucesso é obtido quando a canção aparece até determinada posição no *Hot 100* da Billboard (essa posição foi variada nos experimentos). Os autores calcularam diferentes coeficientes de correlação entre o número de mensagens coletadas e o sucesso de cada canção. O valor máximo foi de 0,41, o que pode indicar que não há correlação entre estes. Mesmo com tal entrave, os autores aplicaram o classificador RF, obtendo acurácia de 90% no modelo em que uma música só é considerada como de sucesso se estiver entre as dez primeiras posições.

Por fim, em relação ao uso de características acústicas das músicas, Herremans, Martens e Sörensen (2014) criaram um modelo de previsão de popularidade de músicas do gênero Dance. Para uma música ser considerada popular nessa pesquisa ela deveria estar até certa posição do Top 40 Dance Music da OCC, assim como no trabalho anterior, essa posição também foi sendo variada nos experimentos. Os autores coletaram meta-dados e informações das características acústicas das faixas que apareceram nesse *ranking* entre 2009 e 2013 utilizando o The Echo Nest. Três experimentos distintos foram realizados, onde diferentes parâmetros para uma música ser considerada popular foram testados. Os melhores resultados foram obtidos ao utilizar o classificador Naive Bayes (não foi dito qual premissa sobre a distribuição foi tomada) no experimento onde a música deveria estar entre as 10 primeiras posições do *ranking* para ser considerada popular e entre as posições 31 a 40 para ser considerada impopular, as músicas nas posições 11 a 30 foram descartadas desse experimento. Nesse experimento os autores obtiveram acurácia e AUC de 65%.

Além desse trabalho, Karydis et al. (2018) coletaram informações de 9193 músicas que apareceram em ao menos um *ranking* da Billboard, Last.fm

ou Spotify. Os dados tratam do período entre 28 de abril de 2013 e 28 de dezembro de 2014. Os autores também coletaram informações de outras 14192 canções contidas nos álbuns das músicas coletadas na etapa anterior. Desse total de 23385 músicas, os autores obtiveram informações das canções a partir de três diferentes fontes, sendo elas: iTunes[6], Spotify e 7digital[7]. Além disso, usando quatro diferentes ferramentas extraíram características acústicas dessas canções a partir de amostras de 30 segundos das mesmas. O objetivo da pesquisa era prever qual música seria a mais popular de um álbum. Utilizando Redes Neurais Autorregressivas Não-Lineares (NAR) e sua variação com entrada externa (NARX), os autores relataram acurácia de 52,26% e precisão de 45,92%.

A pesquisa que mais se assemelha ao aqui realizado é a de (REIMAN; ÖRNELL, 2018). Nesse trabalho, Reiman e Örnell (2018) coletaram dados de 287 músicas que apareceram no *ranking* Billboard Hot 100 entre 2016 e 2018. Elas também coletaram dados de outras 322 músicas que nunca apareceram nesse *ranking*, escolhidas aleatoriamente a partir de 13 gêneros musicais diferentes. Essas informações foram coletadas usando a API do Spotify e tratam de particularidades dessas músicas, isto é, se são alegres, dançantes, instrumentais, dentre outras. Para uma música ser considerada popular nessa pesquisa, ela deveria estar presente no Hot 100.

Reiman e Örnell (2018) utilizaram quatro diferentes algoritmos para realizar suas previsões, sendo eles: KNN, SVM, NBG e LR. A avaliação experimental foi feita com base em validação por *hold-out* (80% treino e 20% para teste), com acurácia máxima 60,17%, obtida pelo Naive Bayes Gaussiano. A conclusão apresentada em (REIMAN; ÖRNELL, 2018) é que os experimentos não demonstraram ser possível prever se uma música será ou não um sucesso.

[6] <https://affiliate.itunes.apple.com/resources/documentation/itunes-store-web-service-search-api/>

[7] <http://docs.7digital.com/>

3.2 Discussão sobre os Trabalhos Relacionados

No último parágrafo de seu trabalho, Pachet e Roy (2008) também afirmam que o principal ponto a se trabalhar para ser possível realizar previsões no mercado musical é o de determinar que tipo de fonte de dados utilizar em pesquisas na área. Como visto anteriormente, diferentes tipos de fontes já foram utilizados, mas nenhum desponta como principal.

Das fontes apresentadas, a utilização de informações sobre shows e festivais (ARAKELYAN et al., 2018; STEININGER; GATZEMEIER, 2013) mostrou não ser a melhor escolha. A precisão obtida pelos modelos que usaram tais dados foi inferior a 50%.

Sobre dados de redes sociais, um dos trabalhos apresentados (KIM; SUH; LEE, 2014) alcançou acurácia superior a 90% em sua previsão quando utilizou dados dessas redes. Enquanto, uma outra pesquisa (DHAR; CHANG, 2009) afirma que a quantidade de postagens sobre um álbum em redes sociais e *blogs* é o principal fator de influência na popularidade desse disco. Portanto, o uso de informações de redes sociais para realizar previsões no mercado musical pode ser um bom caminho a ser seguido em futuras pesquisas em HSS.

As características acústicas das músicas também despontam como um bom caminho a ser seguido. Os trabalhos apresentados (NI et al., 2011; HERREMANS; MARTENS; SöRENSEN, 2014; KARYDIS et al., 2018) obtiveram todos acurácia superior a 50%. Dois dos trabalhos apresentados utilizam informações coletadas do The Echo Nest, mas lamentavelmente não é mais possível utilizar a plataforma para coletar tais dados. A empresa foi comprada pelo Spotify em 2014 e hoje provê suas análises apenas para companhias que são seus clientes[8]. Importante salientar que algumas das informações anteriormente providas pelo The Echo Nest agora podem ser acessadas pela API do Spotify.

Nossa abordagem se distingue das apresentadas anteriormente. Utilizamos informações sobre características das músicas que também foram usados

[8] <http://bit.ly/32RtRPz>

por Reiman e Örnell (2018). Os dados de redes sociais não foram utilizados, pois estes não apresentam correlação com *rankings* de músicas populares, conforme trabalho prévio desenvolvido e que está apresentado no Apêndice A. Destacamos que, apesar de tal indicação, pretendemos futuramente analisar se a incorporação de tais dados ao modelo desenvolvido fará com que consigamos melhores resultados, pois esse trabalho foi feito com base em um *ranking* da Billboard, que é específico do mercado americano. Enquanto as características acústicas foram empregadas em modelos preliminares (apresentados no Capítulo 4), mas sua utilização não trouxe aumento de desempenho significativo, logo elas não estão presentes em nosso modelo final. Entretanto, para o cálculo automatizado das informações por nós utilizadas são necessários os valores de algumas dessas características acústicas.

O trabalho de Reiman e Örnell (2018) utiliza os valores exatos das características. Em nossa pesquisa nosso foco está em dar um novo significado e representação aos dados. Essa nova representação possibilita que artistas e gravadoras consigam realizar previsões para suas novas músicas antes de serem lançadas sem a necessidade de realizar cálculos ou de conhecimento técnico específico.

4 Modelos Preliminares

Até chegarmos ao modelo que estamos propondo nesta dissertação realizamos o desenvolvimento de modelos preliminares. Esses modelos não podem ser considerados como de HSS por realizarem previsões para músicas já lançadas. Mostraremos aqui três modelos prévios desenvolvidos que resultaram em artigos publicados e serão explanados em ordem cronológica. Assim, pode-se analisar como conhecimentos aprendidos em um trabalho foram utilizados nos modelos seguintes. Esse aprendizado foi fundamental para desenvolvermos o modelo proposto, pois nos auxiliou a encontrar a melhor forma de se realizar previsões no mercado musical e também indicou caminhos a serem evitados. Na Seção 4.1 discutiremos o modelo apresentado no Simpósio Brasileiro de Computação Musical (SBCM) (ARAUJO; CRISTO; GIUSTI, 2019a), enquanto na Seção 4.2 mostraremos o defendido no Encontro Nacional de Inteligência Artificial e Computacional (ENIAC) (ARAUJO; CRISTO; GIUSTI, 2019c). Por fim, na Seção 4.3 abordaremos o modelo aceito no IEEE International Conference on Machine Learning and Applications (ICMLA) (ARAUJO; CRISTO; GIUSTI, 2019b).

4.1 *Predicting Music Popularity on Streaming Platforms*

Assim como nos demais modelos, os dados utilizados nessa pesquisa foram coletados a partir da API do Spotify[1]. Esses dados eram referentes aos *rankings* diários Top 50 Global e Viral 50 Global e foram coletados entre novembro de 2018 e janeiro de 2019. O Top 50 conta com as 50 músicas mais ouvidas na plataforma no dia anterior ao lançamento da lista. Enquanto o Viral 50 contém as 50 músicas que tiveram maior acréscimo no número de ouvintes em relação

[1] <https://developer.spotify.com/documentation/web-api/>

ao dia anterior[2]. Assim, para um artista popular é mais fácil chegar ao Top 50 do que no Viral 50, enquanto para um artista menos popular a dificuldade é contrária. Nosso objetivo era desenvolver um modelo que previsse se uma música que está no Top 50 iria aparecer no Viral 50 e vice-versa.

Para cada entrada em uma edição do *ranking*, coletamos informações de nove campos diretamente da API. São eles: a posição da música no *ranking*, a data do *ranking*, os nomes dos artistas e da faixa, a data de lançamento da canção e sua duração em milissegundos. Há também um indicador de popularidade, definido pela plataforma como um valor inteiro no intervalo $[0, 100]$, e um campo binário que indica se a música contém palavras de baixo calão.

Além desses atributos, cada entrada possui uma URL para uma amostra de 30 segundos da música. Com base nessas amostras, extraímos cinco características acústicas, sendo elas: MFCC (em 13 coeficientes), centroide espectral, coeficiente de tonalidade, passagens por zero e tempo (apresentados na Tabela 1, pág. 23). Esse processo foi feito utilizando o pacote para Python LibROSA (MCFEE et al., 2015).

Visto que o objetivo era prever se uma música presente em um *ranking* iria aparecer no outro, então era necessário que as entradas de um *ranking* apresentassem informações sobre essa mesma música no outro, logo adicionamos novos campos. Esses campos foram: a posição da música na outra lista, o número de dias consecutivos que a música permaneceu na lista, a data da primeira aparição. Também incorporamos um campo binário que indicava se a música estava presente no outro *ranking* no dia seguinte, sendo esse campo a classe de nosso problema.

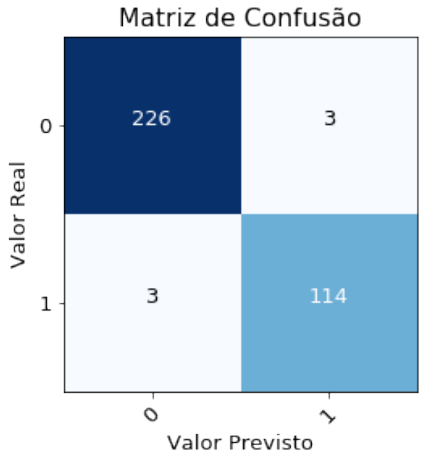
Induzimos quatro modelos. Três variam com respeito aos atributos: o modelo de "informações prévias" foi induzido apenas com dados obtidos diretamente a partir do *ranking* (posição na lista, datas, popularidade etc.). O modelo de "características acústicas" foi induzido apenas com os dados extraídos a

[2] Conforme explicado por Kevin Goldsmith, ex-vice-presidente de engenharia do Spotify, em informação na plataforma Quora. Resposta disponível em <http://bit.ly/33fXg67>, é preciso estar logado para acessá-la. Consulta feita às 11:32 do dia 08/11/2019.

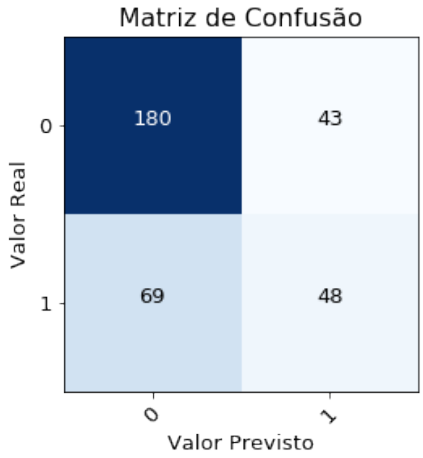
partir das amostras de 30 segundos. O modelo "global" foi induzido com os dados que estavam disponíveis para os dois modelos anteriores. Finalmente, o modelo "*baseline*" foi induzido com base na mediana da quantidade de vezes que uma música aparecia no outro *ranking*. Se essa música aparecesse com frequência igual ou superior à mediana, esse modelo dizia que essa música estaria presente em todos os dias do teste.

Com exceção do *baseline*, os modelos foram treinados com o classificador SVM utilizando *kernel* RBF. Os dados de treinamento foram extraídos dos *rankings* coletados em novembro e dezembro de 2018. Os testes foram realizados com dados coletados em cada semana de janeiro de 2019, de modo que cada modelo foi avaliado em quatro conjuntos de teste.

O melhor resultado, em termos de AUC, foi obtido pelo modelo de "informações prévias" ao prever o Top 50 da quarta semana. A matriz de confusão desse experimento é mostrada na Figura 2a, na qual pode-se observar que apenas seis instâncias foram classificadas incorretamente. Em contrapartida, o pior resultado foi obtido pelo modelo de "características acústicas" ao prever o Viral 50 da quarta semana. A matriz de confusão desse caso é apresentada na Figura 2b.



(a) Matriz de confusão do modelo de "informações prévias" ao prever o Top 50 da quarta semana de janeiro.



(b) Matriz de confusão do modelo de "características acústicas" ao prever o Viral 50 da quarta semana de janeiro.

Figura 2 – Matrizes de confusão (a) do melhor e (b) pior modelos segundo os valores de AUC. Extraído de (ARAUJO; CRISTO; GIUSTI, 2019a).

Os resultados obtidos pelo *baseline* estão na Tabela 2, enquanto os do modelo de "informações prévias" estão na Tabela 3. Na Tabela 4 estão os valores atingidos pelo modelo de "características acústicas". Por fim, aqueles referentes ao modelo "global" estão na Tabela 5.

Tabela 2 – Desempenho do *baseline*. Extraído de (ARAUJO; CRISTO; GIUSTI, 2019a).

	Viral				**Top**			
	1ª Semana	**2ª Semana**	**3ª Semana**	**4ª Semana**	**1ª Semana**	**2ª Semana**	**3ª Semana**	**4ª Semana**
Acurácia	0,7062	0,7560	0,6932	0,6416	0,9475	0,9296	0,8675	0,8526
AUC	0,6970	0,7465	0,6691	0,6217	0,9182	0,8856	0,8070	0,7862
MCC	0,3715	0,4662	0,3293	0,2196	0,8746	0,8247	0,7148	0,6741
Revocação	0,6731	0,7228	0,5982	0,5214	0,8447	0,7835	0,6140	0,5812
Valor-F1	0,5858	0,6404	0,5630	0,4959	0,9063	0,8636	0,7609	0,7273
Precisão	0,5185	0,5748	0,5317	0,4729	0,9775	0,9620	1,0000	0,9714
Especificidade	0,7210	0,7702	0,7401	0,7031	0,9917	0,9877	1,0000	0,9913

Tabela 3 – Desempenho do modelo de "informações prévias". Extraído de (ARAUJO; CRISTO; GIUSTI, 2019a).

	Viral				**Top**			
	1ª Semana	**2ª Semana**	**3ª Semana**	**4ª Semana**	**1ª Semana**	**2ª Semana**	**3ª Semana**	**4ª Semana**
Acurácia	0,9436	0,9375	0,9410	0,9353	0,9650	0,9824	0,9849	0,9827
AUC	0,9113	0,9073	0,9107	0,9161	0,9473	0,9784	0,9781	0,9806
MCC	0,8683	0,8496	0,8689	0,8558	0,9163	0,9568	0,9668	0,9613
Revocação	0,8269	0,8317	0,8214	0,8547	0,9029	0,9691	0,9561	0,9744
Valor-F1	0,9005	0,8889	0,9020	0,9009	0,9394	0,9691	0,9776	0,9744
Precisão	0,9885	0,9545	1,0000	0,9524	0,9789	0,9691	1,0000	0,9744
Especificidade	0,9957	0,9830	1,0000	0,9776	0,9917	0,9877	1,0000	0,9869

Tabela 4 – Desempenho do modelo de "características acústicas". Extraído de (ARAUJO; CRISTO; GIUSTI, 2019a).

	Viral				**Top**			
	1ª Semana	**2ª Semana**	**3ª Semana**	**4ª Semana**	**1ª Semana**	**2ª Semana**	**3ª Semana**	**4ª Semana**
Acurácia	0,8101	0,8185	0,7552	0,6706	0,8630	0,8798	0,8253	0,8121
AUC	0,7695	0,7714	0,6996	0,6087	0,7718	0,7886	0,7456	0,7264
MCC	0,5482	0,5582	0,4237	0,2333	0,6743	0,7030	0,6299	0,5816
Revocação	0,6635	0,6535	0,5357	0,4103	0,5437	0,5773	0,4912	0,4615
Valor-F1	0,6832	0,6839	0,5911	0,4615	0,7044	0,7320	0,6588	0,6243
Precisão	0,7041	0,7174	0,6593	0,5275	1,0000	1,0000	1,0000	0,9643
Especificidade	0,8755	0,8894	0,8634	0,8072	1,0000	1,0000	1,0000	0,9913

Tabela 5 – Desempenho do modelo "global". Extraído de (ARAUJO; CRISTO; GIUSTI, 2019a).

	Viral				**Top**			
	1ª Semana	**2ª Semana**	**3ª Semana**	**4ª Semana**	**1ª Semana**	**2ª Semana**	**3ª Semana**	**4ª Semana**
Acurácia	0,9080	0,8690	0,8584	0,7971	0,9417	0,9179	0,8554	0,8295
AUC	0,8510	0,7822	0,7857	0,7132	0,9057	0,8619	0,7895	0,7541
MCC	0,7871	0,6895	0,6868	0,5463	0,8614	0,7961	0,6888	0,6193
Revocação	0,7019	0,5644	0,5714	0,4444	0,8155	0,7320	0,5789	0,5214
Valor-F1	0,8249	0,7215	0,7273	0,6012	0,8936	0,8353	0,7333	0,6740
Precisão	1,0000	1,0000	1,0000	0,9286	0,9882	0,9726	1,0000	0,9531
Especificidade	1,0000	1,0000	1,0000	0,9821	0,9958	0,9918	1,0000	0,9869

Ao analisar os resultados, verificamos que o modelo de "informações prévias" obteve os melhores resultados de forma geral. A exceção ocorre nos valores de precisão e especificidade, pois os outros modelos tendem a prever mais instâncias como negativas, diminuindo a quantidade de falsos positivos (ou até mesmo a zerando), que é um fator utilizado na fórmula da precisão e especificidade. Ao analisarmos as outras métricas, observamos que os resultados obtidos nesses experimentos foram piores, demonstrando a importância de se ter um conjunto de métricas para avaliação dos modelos.

Outro importante fator que pode ser observado no modelo de "informações prévias" é o fato dos seus resultados não se degradarem conforme os dados de testes ficaram mais distantes aos de treino. Curiosamente, o maior valor em AUC obtido nos casos das duas listas foi obtido na previsão da quarta semana de janeiro.

Cenário distinto ocorre nos outros dois modelos. O modelo de "características acústicas" chega a declinar 42,56% em MCC quando comparados às previsões da primeira e quarta semana do Viral 50. Esse modelo também é o que obteve os piores resultados em nossos experimentos, que chegam a ser inferiores aos obtidos pelo *baseline* no caso da previsão do Top 50. O modelo "global" também apresentou declínio nos seus resultados. Nesse caso, a queda atingiu 63,93% em revocação quando comparado os resultados das previsões da primeira e quarta semana do Top 50.

Entretanto, o modelo de "características acústicas" e o "global" obtiveram precisão e especificidade igual a 1 em alguns dos experimentos realizados. Isso

implica que quando esses modelos preveem uma instância como positiva ela garantidamente é uma previsão correta. Esse cenário não ocorre ao prever uma instância como negativa. Porém, um modelo que acerta 100% de suas previsões de uma certa classe pode ser útil quando aplicado em casos reais, mesmo que de forma geral seus resultados sejam inferiores.

Uma possível explicação para esse cenário é que ao utilizar as características acústicas, o modelo decorou quais músicas são positivas no treino e as prevê como positivas quando estão no teste. Conforme a janela de teste fica mais distante a de treino, essas músicas tendem a não mais aparecer e, portanto, o modelo acerta menos instâncias de músicas positivas. Por outro lado, o modelo de "informações prévias" deve ter aprendido aspectos que fazem uma música ser considerada positiva em nosso experimento, assim consegue fazer previsões corretas mesmo para músicas que não estiveram no treino.

4.2 *Will I Remain Popular? A Study Case on Spotify*

Nesse segundo trabalho o objetivo era identificar se uma música já presente no Top 50 iria nela permanecer após um determinado período de tempo. Portanto, não utilizamos informações do Viral 50. Os dados foram coletados entre novembro de 2018 a abril de 2019.

No trabalho anterior, as previsões foram feitas para o dia seguinte ao da entrada. Nesse segundo trabalho buscamos uma forma de aumentar essa janela de previsão, pois é mais interessante prever se uma música obterá sucesso em longo alcance em relação à previsão de sucesso imediato.

Para aumentar essa janela fizemos três rodadas de previsão, em que os resultados obtidos nessas rodadas eram inseridos como se fossem dados reais nas rodadas seguintes, permitindo previsões em janelas maiores. Esses dados eram inseridos em campos *lagged*, que contavam com informações sobre a presença dessas músicas em dias posteriores do *ranking*. Cada entrada tinha quatro campos *lagged* que representavam a presença da música nos *rankings*

de 20, 40, 60 e 80 dias após a data do dado de entrada. Uma representação dessas rodadas pode ser vista na Figura 3.

No trabalho anterior utilizamos somente um algoritmo de aprendizagem de máquina, que poderia não ser o mais indicado. Portanto, nesse trabalho decidimos utilizar uma gama de algoritmos para identificar aquele que melhor se adéque ao problema, são eles: Ada Boost, Naive Bayes Bernoulli, Naive Bayes Gaussiano, Floresta Aleatória e SVM com *kernels* linear, polinomial, RBF e sigmoid. Todos foram explanados na Subseção 2.2.2. Os testes eram feitos com entradas referentes aos dias 20 de janeiro a 08 de fevereiro para previsão dos dias 10 a 29 de abril.

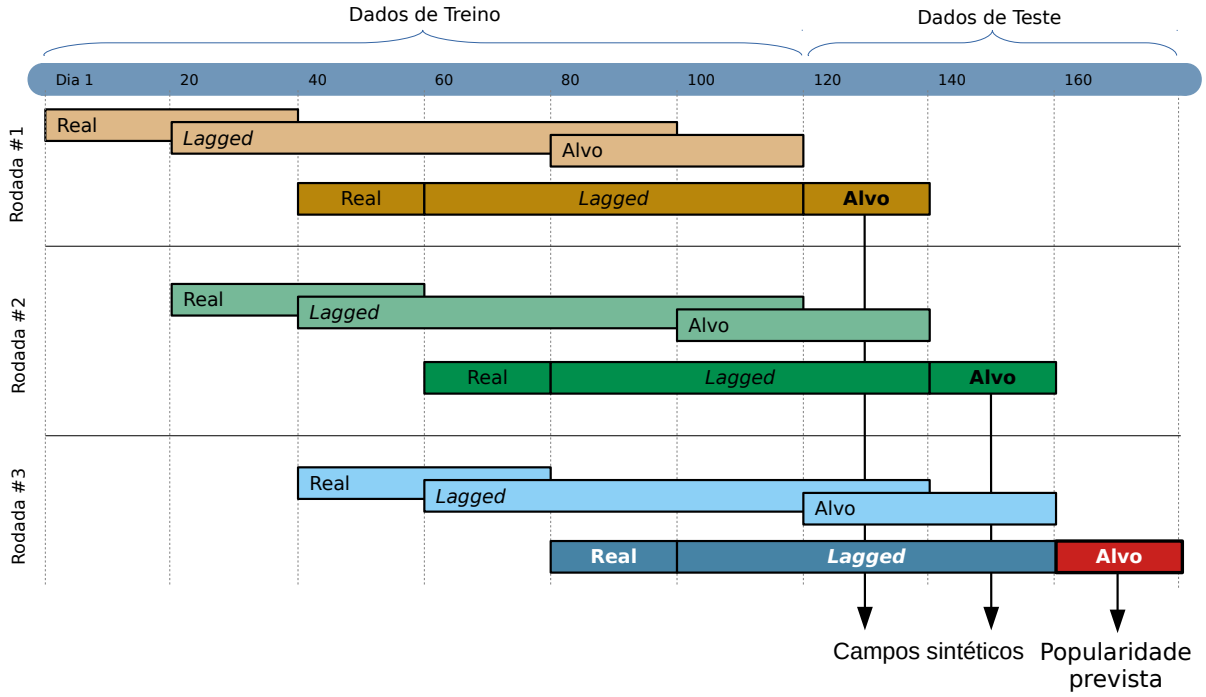


Figura 3 – Descrição das rodadas de classificação. Em cada rodada, treinamos um classificador e prevemos informações de popularidade em mais 20 dias no futuro. As caixas mais claras de cada rodada representam os dados usados para treinar o classificador, e as caixas mais escuras representam os dados usados para prever as novas informações de popularidade, que são usadas para treinar os classificadores subsequentes. Isso nos permite melhorar o nosso horizonte de previsão três vezes. Extraído de (ARAUJO; CRISTO; GIUSTI, 2019b).

Como o trabalho anterior mostrou que o uso de características acústicas de forma isolada não trazia bons resultados, então aqui foram desenvolvidos somente dois modelos, sendo que os dois utilizam informações prévias sobre

a presença das músicas no *ranking*, um deles também contendo informações sobre as características acústicas e o outro não. Os resultados obtidos pelo que não utiliza essas informações está na Tabela 6, enquanto na Tabela 7 estão os valores alcançados pelo outro modelo. Os melhores resultados atingidos estão evidenciados em vermelho.

Tabela 6 – Desempenho dos modelos que utilizam somente dos dados prévios do Top 50. Extraído de (ARAUJO; CRISTO; GIUSTI, 2019c).

Dados Prévios	**Ada Boost**	**NB Bernoulli**	**NB Gaussiano**	**Floresta Aleatória**	**SVM Linear**	**SVM Poli**	**SVM RBF**	**SVM Sigmoid**
Acurácia	0,6924	0,7278	0,5172	0,6903	0,7372	0,6455	0,7143	0,7049
Precisão	0,5641	0,5945	0,3975	0,5978	0,5984	0,5146	0,5757	0,5708
VPN	0,7962	0,8566	0,6789	0,7267	0,8872	0,7180	0,8665	0,8354
Revocação	0,6914	0,8000	0,6257	0,4629	0,8514	0,5029	0,8257	0,7714
Especificidade	0,6929	0,6864	0,4548	**0,8210**	0,6716	0,7274	0,6502	0,6667
Valor-F1	0,6213	0,6821	0,4861	0,5217	0,7028	0,5087	0,6784	0,6561
AUC	0,6922	0,7432	0,5403	0,6419	0,7615	0,6151	0,7380	0,7190
Fall-out	0,3071	0,3136	0,5452	**0,1790**	0,3284	0,2726	0,3498	0,3333
Taxa de Perda	0,3086	0,2000	0,3743	0,5371	0,1486	0,4971	0,1743	0,2286

Tabela 7 – Desempenho dos modelos que utilizam todos os dados disponíveis. Extraído de (ARAUJO; CRISTO; GIUSTI, 2019c).

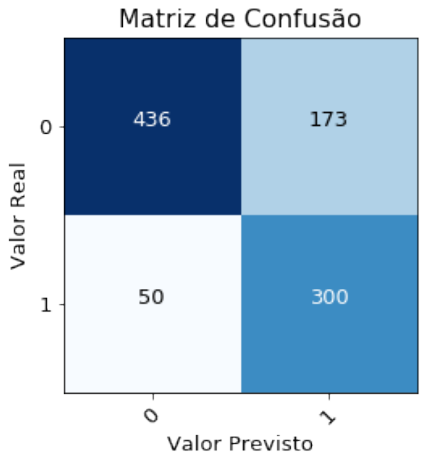
Todos os Dados	**Ada Boost**	**NB Bernoulli**	**NB Gaussiano**	**Floresta Aleatória**	**SVM Linear**	**SVM Poli**	**SVM RBF**	**SVM Sigmoid**
Acurácia	0,7372	0,7247	0,3889	0,6538	**0,7675**	0,6945	0,7143	0,6840
Precisão	0,6099	0,5915	0,3405	0,5313	**0,6342**	0,5899	0,5805	0,5512
VPN	0,8480	0,8528	0,5525	0,7064	**0,8971**	0,7461	0,8439	0,8060
Revocação	0,7771	0,7943	0,7200	0,4371	**0,8571**	0,5343	0,7829	0,7229
Especificidade	0,7143	0,6847	0,1987	0,7783	0,7159	0,7865	0,6749	0,6617
Valor-F1	0,6834	0,6780	0,4624	0,4796	**0,7290**	0,5607	0,6667	0,6255
AUC	0,7457	0,7395	0,4593	0,6077	**0,7865**	0,6604	0,7289	0,6923
Fall-out	0,2857	0,3153	0,8013	0,2217	0,2841	0,2135	0,3251	0,3383
Taxa de Perda	0,2229	0,2057	0,2800	0,5629	**0,1429**	0,4657	0,2171	0,2771

O classificador NBG foi o pior em nosso experimento. Mas, essa situação era esperada, visto que o foco dos algoritmos Naive Bayes está na eficiência, isto é, na rápida execução de sua classificação (ZHANG, 2004). Entretanto, os resultados obtidos por esses classificadores são importantes, pois podem servir como *baseline*.

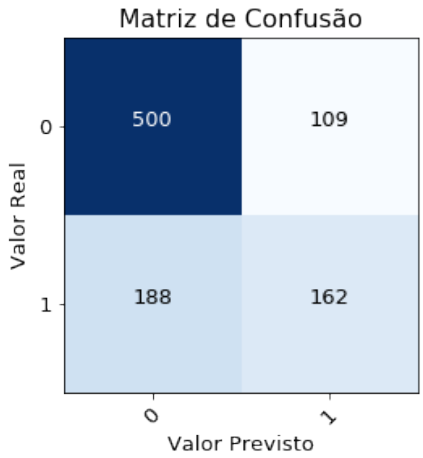
De forma geral, o melhor modelo em nossa pesquisa foi treinado utilizando todos os dados em um classificador SVM com *kernel* linear. Seus resultados foram 97,35% superiores em acurácia e 260,29% em especificidade

em relação aos obtidos pelo NBG. A superioridade mínima encontrada entre esses dois classificadores foi de 19,04% em revocação. A matriz de confusão obtida pelo melhor modelo está representada na Figura 4a.

Observa-se por essa matriz que o maior ponto a se desenvolver nesse modelo é o dos falsos positivos, isto é, instâncias que o modelo prevê que estariam presentes na lista, mas que na realidade não estão. Tal situação explica o fato desse modelo não apresentar o melhor valor em especificidade em nosso experimento. Ele foi superado pelos classificadores RF e SVM com *kernel* polinomial treinados tanto com todos os dados, quanto somente com os dados prévios. O melhor resultado obtido segundo essa métrica foi utilizando somente os dados prévios treinados no classificador RF. A matriz de confusão obtida nesse caso está na Figura 4b. Observamos por essa matriz que esse classificador previu 71,74% das instâncias como negativas, o que diminuiu o número de falsos positivos, aumentando o valor em especificidade.



(a) Matriz de confusão para o modelo treinado com todos os dados utilizando classificador SVM com *kernel* linear.



(b) Matriz de confusão para o modelo treinado somente com dados prévios utilizando o classificador Floresta Aleatória.

Figura 4 – Matrizes de confusão (a) do melhor modelo segundo sete das métricas utilizadas e (b) do melhor modelo segundo os valores de especificidade e *fall-out*. Extraído de (ARAUJO; CRISTO; GIUSTI, 2019c).

O alto valor em especificidade obtido pelo RF utilizando somente os dados prévios poderia ser um indicativo para sua escolha como melhor modelo.

Entretanto, devido às particularidades de nosso caso de teste onde 63,5% das instâncias são negativas, então as métricas revocação e AUC são as mais relevantes. Comparativamente, o modelo SVM com *kernel* linear treinado com todos os dados obteve revocação e AUC 85,16% e 22,53% superiores em relação ao RF, respectivamente.

Diferentemente do trabalho apresentado na seção 4.1, ao utilizar as características acústicas das faixas o modelo SVM com *kernel* linear apresentou uma pequena melhoria de 5,98% e 6,60% em precisão e especificidade, respectivamente. Entretanto, consideramos que esse reduzido aumento de eficiência não vale o custo computacional de se extrair tais informações.

4.3 *Predicting Music Popularity Using Music Charts*

Esse trabalho estende o trabalho anterior, realizando previsões tanto para músicas populares quanto para não populares, enquanto no anterior as previsões foram feitas somente para músicas já populares. Os dados coletados foram os mesmos e o processo de rodadas também foi realizado a fim de estender o horizonte preditivo. Também foram utilizados os mesmos algoritmos de classificação e os testes foram feitos na mesma janela de tempo. A diferença é que aqui levantamos todo o conjunto de músicas distintas que apareceram no período de coleta e buscávamos prever quais delas iriam aparecer em cada um dos dias dos *rankings* do teste. Assim, nossa base contava com 274 músicas distintas.

Ao aprender a melhor forma de se realizar esse tipo de previsão, então o próximo passo seria o de realizar para músicas ainda não lançadas, como fazemos no modelo final proposto.

Aqui também foram desenvolvidos dois modelos, um com e um sem os dados referentes às características acústicas das faixas, além das informações sobre a presença prévia dessas músicas no *ranking*. O resultado obtido nessa terceira pesquisa foi essencial para definição sobre o uso das características

acústicas em nosso modelo final proposto. Pois, com seu uso foram alcançados resultados piores no primeiro trabalho, enquanto no segundo houve aumento de desempenho, mesmo que reduzido.

Na Tabela 8 estão os valores calculados das métricas de avaliação para o modelo que utiliza somente os dados prévios das listas. Enquanto para aquele que usa também as informações de características acústicas das faixas é possível ver seu desempenho na Tabela 9. Os melhores resultados atingidos estão evidenciados em vermelho.

Tabela 8 – Desempenho do modelo que não utiliza as características acústicas. Extraído de (ARAUJO; CRISTO; GIUSTI, 2019b).

Dados Prévios	**Ada Boost**	**NB Bernoulli**	**NB Gaussiano**	**Floresta Aleatória**	**SVM Linear**	**SVM Poli**	**SVM RBF**	**SVM Sigmoid**
Acurácia	0,8828	0,8849	0,8359	0,8730	0,8849	0,8849	0,8849	0,7856
Precisão	0,6762	0,6697	0,5255	0,6599	0,6697	0,6697	0,6697	0,3750
VPN	0,9241	0,9316	0,9281	0,9113	0,9316	0,9316	0,9316	0,8608
Revocação	0,6397	0,6802	0,6843	0,5722	0,6802	0,6802	0,6802	0,3302
Especificidade	0,9347	0,9285	0,8683	0,9371	0,9285	0,9285	0,9285	0,8827
Valor-F1	0,6574	0,6749	0,5945	0,6129	0,6749	0,6749	0,6749	0,3512
AUC	0,7872	0,8043	0,7763	0,7546	0,8043	0,8043	0,8043	0,6064
Fall-out	0,0653	0,0715	0,1317	0,0629	0,0715	0,0715	0,0715	0,1173
Taxa de Perda	0,3603	0,3198	0,3157	0,3401	0,3198	0,3198	0,3198	0,6698

Tabela 9 – Desempenho do modelo que utiliza todos os dados disponíveis. Extraído de (ARAUJO; CRISTO; GIUSTI, 2019b).

Todos os Dados	**Ada Boost**	**NB Bernoulli**	**NB Gaussiano**	**Floresta Aleatória**	**SVM Linear**	**SVM Poli**	**SVM RBF**	**SVM Sigmoid**
Acurácia	0,8867	0,8869	0,8286	0,8847	0,8849	0,8901	**0,8909**	0,7887
Precisão	0,6804	0,6774	0,5092	0,6918	0,6697	**0,6935**	0,6915	0,3818
VPN	0,9298	0,9318	0,9284	0,9207	0,9316	0,9305	**0,9328**	0,8608
Revocação	0,6698	0,6802	0,6895	0,6199	0,6802	0,6719	**0,6843**	0,3271
Especificidade	0,9329	0,9309	0,8583	0,9411	0,9285	**0,9367**	0,9349	0,8871
Valor-F1	0,6750	0,6788	0,5858	0,6539	0,6749	0,6825	**0,6879**	0,3523
AUC	0,8013	0,8055	0,7739	0,7805	0,8043	0,8043	**0,8096**	0,6071
Fall-out	0,0671	0,0691	0,1417	0,0589	0,0715	**0,0633**	0,0651	0,1129
Taxa de Perda	0,3302	0,3198	0,3105	0,3801	0,3198	0,3281	**0,3157**	0,6729

Como nossa base de dados é bastante desbalanceada, com cerca de 4,6 vezes mais instâncias negativas que positivas, métricas mais robustas a tal situação são mais apropriadas para avaliação. Assim, focamos mais em revocação, taxa de perda, valor-F1 e AUC do que em precisão e acurácia. Considerando esses fatores, o melhor modelo para o nosso problema foi o que

utilizou todos os dados e foi treinado com o classificador SVM com *kernel* RBF. Conforme destacado em vermelho, o modelo treinado com *kernel* polinomial atingiu melhores resultados em precisão, especificidade e *fall-out*, pois previu corretamente mais instâncias negativas que o *kernel* RBF, entretanto obteve piores resultados ao prever instâncias positivas.

Ao utilizar as características acústicas, o modelo que utiliza o classificador SVM com *kernel* RBF previu corretamente 29 instâncias negativas e 4 instâncias positivas a mais em relação a quando esses dados não foram utilizados. Entretanto, essas 33 instâncias representam um acréscimo de desempenho de apenas 5,23%. Logo, assim como ocorreu no trabalho apresentado na seção 4.2, as características acústicas não trouxeram aumento significativo de desempenho nos modelos. Portanto decidimos não as utilizar em nosso modelo proposto.

O principal ponto a se desenvolver nesse modelo é achar uma forma de diminuir a quantidade de falsos positivos, isto é, instâncias que o modelo prevê que estariam presentes na lista, mas que na realidade não estão. O modelo que obteve os melhores resultados previu incorretamente 31,57% das instâncias positivas.

5 O Modelo

Com o aprendizado adquirido no desenvolvimento dos modelos preliminares, estabelecemos uma metodologia que eventualmente nos permitiu alcançar o modelo proposto. O modelo final promove dois avanços substanciais com relação aos anteriores. Primeiramente, esse modelo pode ser caracterizado como HSS, visto que as previsões podem ser realmente feitas para músicas antes mesmo de serem lançadas. Além disso, nosso modelo necessita de apenas um treinamento para realizar previsões para músicas em datas distantes das de treino, não exigindo diversas rodadas de previsão para estender o horizonte.

Como será explanado nas seções a seguir, o modelo final não utiliza diretamente as características acústicas. Pois, concluímos que essas características não contribuíram substancialmente nos modelos anteriores. Portanto, empregamos atributos de mais alto nível que podem ser definidos sem necessidade de analisar acusticamente a música. Nosso modelo utiliza esses atributos para prever se uma música irá ou não aparecer no Top 50.

Este capítulo está dividido em três seções, sendo a Seção 5.1 dedicada à apresentação da metodologia utilizada. Na Seção 5.2 estão os resultados alcançados, que são discutidos na Seção 5.3.

5.1 Metodologia

Nossa metodologia inicia com a coleta de dados a partir da API do Spotify. Os dados coletados geram nossa base de dados em que utilizamos diferentes algoritmos de aprendizagem de máquina para identificar aquele que alcança os melhores resultados. A metodologia será aprofundada nas próximas subseções e uma representação gráfica dela está na Figura 5.

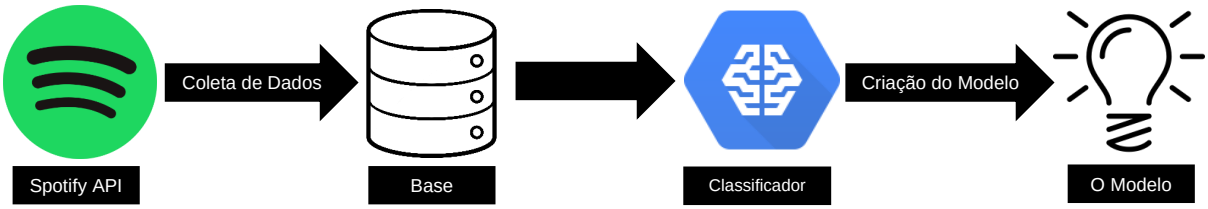


Figura 5 – Metodologia utilizada.

5.1.1 Coleta de Dados e Preparação da Base

Para criação e teste do nosso modelo, coletamos diariamente dados dos *rankings* Top 50 e Viral 50 utilizando a Web API do Spotify durante os meses de novembro de 2018 a julho de 2019.

Neste trabalho, consideramos as músicas presentes no Top 50 como populares, em uma abordagem já utilizada em outras pesquisas de HSS (HERREMANS; MARTENS; SöRENSEN, 2014; REIMAN; ÖRNELL, 2018). As músicas não populares da base são aquelas presentes no Viral 50, excluindo-se as músicas também presentes no Top 50, assim evitando entradas duplicadas com classes distintas. Optamos por construir nosso conjunto de músicas não populares a partir do *ranking* Viral 50, pois a plataforma só fornece informações de popularidade para músicas que aparecem em *rankings*.

As informações desses *rankings* foram coletadas usando a função "Get a Playlist's Tracks" da API. Assim, pegamos os nomes dos artistas e das faixas que as compõem, os ID's dessas músicas dentro da plataforma, além de um campo que traduzimos como 'Explicitude', que indica se a música contém palavras de baixo calão.

Como substitutos das características acústicas utilizadas nos trabalhos apresentados na Seção 4 e que trouxeram resultados pouco animadores, aqui utilizamos *audio features* disponibilizadas pelo Spotify. As *audio features* que utilizamos em nosso experimento estão descritas a seguir, com explanação baseada na documentação da API:

- ***Danceability*:** descreve o quanto uma faixa é adequada para dançar com base em uma combinação de elementos musicais, incluindo tempo,

estabilidade do ritmo, força da batida e regularidade geral. Um valor de 0,0 indica que a música é menos dançável e 1,0 indica uma mais dançável;

- ***Energy*:** uma medida de 0,0 a 1,0 que representa uma medida perceptiva de intensidade e atividade. Normalmente, as faixas enérgicas parecem rápidas, altas e barulhentas. Por exemplo, *death metal* tem alta energia, enquanto um prelúdio de Bach é baixo na escala. Os recursos que contribuem para esse atributo incluem faixa dinâmica, volume percebido, timbre, taxa de início e entropia geral;

- ***Speechiness*:** detecta a presença de palavras faladas em uma faixa. Quanto mais a gravação parecer exclusivamente com um discurso (por exemplo, poesia, audiolivros e *talk-shows*), mais próximo a 1,0 é o valor do atributo. Segundo a documentação da API, valores acima de 0,66 descrevem faixas que provavelmente são compostas inteiramente por palavras faladas, enquanto valores entre 0,33 e 0,66 descrevem faixas que podem conter música e fala em seções, como no Rap. Por fim, valores abaixo de 0,33 provavelmente representam música e outras faixas que não são unicamente de fala;

- ***Acousticness*:** uma medida de 0,0 a 1,0 que indica se a faixa é acústica. 1,0 representa alta confiança de que a faixa é acústica;

- ***Instrumentalness*:** indica se a música não contém vocais. Os sons “Ooh” e “aah” são tratados como instrumentais nesse contexto. Faixas de rap, por exemplo, são claramente “vocais”. Quanto mais próximo o valor for de 1,0, maior a probabilidade da faixa não conter conteúdo vocal;

- ***Liveness*:** detecta a presença de uma audiência na gravação. Um valor acima de 0,8 oferece uma forte probabilidade da faixa ter sido gravada ao vivo;

- ***Valence*:** uma medida de 0,0 a 1,0 que descreve a positividade musical transmitida por uma faixa. Faixas com valores mais altos soam mais

positivas (por exemplo, felizes, eufóricas), enquanto àquelas com valores baixos soam mais negativas (por exemplo, tristes, deprimidas e zangadas).

Todas essas *audio features* são campos *float* e a documentação não informa como são calculados. Logo, não temos como computar esses valores para músicas que não estão na plataforma, o que impossibilitaria realizar previsões para músicas ainda não lançadas. Para tornar viável essas previsões, decidimos binarizar esses campos. Na binarização dos dados coletados, o campo era considerado positivo se seu valor fosse superior a 0,5. As exceções foram em *speechiness* e *liveness*, em que usamos os valores 0,33 e 0,8 como base, respectivamente, devido a própria descrição desses campos na documentação.

Para músicas que não foram lançadas, por mais que não saibamos o valor exato atingido por elas nas *audio features*, o próprio artista pode indicar se ela é alegre, ao vivo, dançante, dentre outras características que esses campos representam. Dessa forma é possível representar músicas ainda não lançadas como instâncias da nossa base, possibilitando prever o sucesso delas.

Além disso, mesmo que soubéssemos como essas *audio features* são calculadas o processo de binarização ainda assim apresentaria serventia. Em um cenário onde esse processo não fosse realizado, as gravadoras e artistas necessitariam de apoio especializado para a realização dos cálculos dos valores a serem utilizados como entrada no modelo, impossibilitando o utilizar diretamente.

Para nossos experimentos, montamos duas bases. Na primeira, cada entrada representava uma música em um determinado dia, podendo haver várias entradas para uma mesma música se ela figurar em mais de um *ranking*. Na segunda, as entradas que apresentavam mesmo nome de música e artista eram englobadas em uma só. Nesse caso uma música só era considerada popular se aparecesse mais que uma certa quantia de vezes no Top 50 durante o tempo de coleta. As entradas de janeiro de 2019 foram descartadas da primeira base para aumentar ainda mais a distância em tempo dos dados de treino e teste. Após esse processo, descartamos os campos de nome das faixas

e dos artistas das duas, além dos ID's.

Durante o período natalino é comum que músicas temáticas apareçam no Top 50 entre os dias 23 a 26 de dezembro. De forma a evitar que essas músicas fossem consideradas populares no segundo experimento, estabelecemos que para uma canção ser considerada popular ela deveria ter aparecido mais que quatro vezes no Top 50.

Para fins de comparação, montamos um modelo baseado na metodologia utilizada por Reiman e Örnell (2018). Usaremos a sigla ROM (Reiman e Örnell *Model*) quando tratarmos desse modelo daqui em diante. Nesse trabalho não foi utilizado o campo "Explicitude", mas todos as *audio features* disponíveis na API foram. Logo, além dos previamente apresentados, temos ainda:

- **Duration_ms:** a duração da faixa em milissegundos;
- ***Key*:** a nota musical em que a faixa está. Os números inteiros são mapeados usando a notação padrão da classe de afinação. Por exemplo, 0 = C, 1 = C♯/D♭, 2 = D e assim sucessivamente;
- ***Mode*:** indica a modalidade (*major* ou *minor*) da faixa. *Major* é representado por 1 e *minor* por 0;
- ***Tempo*:** o tempo geral estimado de uma faixa em batidas por minuto (BPM). Na terminologia musical, tempo é a velocidade ou ritmo de uma determinada peça e deriva diretamente da duração média da batida;
- **Time_signature:** uma assinatura de tempo global estimada de uma faixa. A assinatura de tempo é uma convenção notacional para especificar quantas batidas existem em cada barra;
- ***Loudness*:** o volume geral de uma faixa em decibéis.

Daqui em diante empregaremos a sigla MP quando tratarmos de nosso modelo proposto. No MP não utilizamos todas as *audio features* disponíveis na API, pois só usamos aquelas em que era possível fazer o processo de binarização. No caso específico do campo "mode" que é binário, não o utilizamos pois o que

ele representa está diretamente associado a nota musical, que é representada no campo "key" e que não foi usada por não poder ser binarizada.

O ROM não é uma replicação da metodologia utilizada por Reiman e Örnell (2018), mas um modelo criado com base nesse texto, assim modificações foram feitas para que ele se adequasse aos nossos experimentos. A primeira diferença está na fonte dos dados populares e não populares, neste trabalho utilizamos o Top 50 e Viral 50 como fontes das músicas populares e não populares, respectivamente. Enquanto Reiman e Örnell (2018) utilizam o Hot 100 da Billbaord como sua fonte de obras populares e músicas coletadas aleatoriamente como não populares. Além disso, nesse trabalho as *audio features* não foram extraídas diretamente da API do Spotify, como nós realizamos, mas sim utilizando a biblioteca para Python Spotipy[1]. Logo, pode haver diferenças na forma como as *audio features* são calculadas nesses dois casos.

No ROM, assim como feito por Reiman e Örnell (2018), não realizamos o processo de binarização dos campos e também não normalizamos os dados. Nesse trabalho é afirmado que só foram utilizadas as instâncias em que as *audio features* estavam no mesmo intervalo. Entretanto, não há a informação de qual intervalo foi utilizado, logo em nossos experimentos todo o conjunto de dados foi empregado. Nós também montamos duas bases para o ROM com a finalidade de realizar a comparação dos resultados obtidos com nossa metodologia. As instâncias dessas duas bases representam as mesmas entradas dos bancos do MP. Por fim, nessa pesquisa só são apresentados os resultados segundo as métricas acurácia, precisão e revocação, logo só temos como comparar os resultados aqui obtidos e os obtidos no texto em relação a essas três métricas.

[1] <https://spotipy.readthedocs.io/en/latest/>

5.1.2 Experimentação

Para realização de nossos experimentos utilizamos diferentes algoritmos de aprendizagem de máquina. Logo, foi necessário dividir nossas bases em grupos de treino e teste. Para os experimentos com repetições de músicas utilizamos os dados de novembro e dezembro de 2018 no treino. Para os que englobam as músicas repetidas os dados de janeiro de 2019 também foram utilizados. Os testes sempre foram realizados nos dados de junho e julho de 2019. Logo, há uma diferença mínima de ao menos cinco meses entre as datas dos dados de treino e teste.

Ainda visando tornar os resultados mais comparáveis, restringimos o número de algoritmos utilizado em nossos experimentos àqueles que também foram utilizados em (REIMAN; ÖRNELL, 2018). Isto é, utilizamos os algoritmos Naive Bayes Gaussiano, KNN, Regressão Logística e Máquina de Vetores de Suporte com *kernel* RBF, todos eles já apresentados na Subseção 2.2.2. A forma como realizamos a preparação dos dados fez com que alcançássemos resultados superiores em relação ao ROM, conforme será demonstrado nas próximas seções deste capítulo.

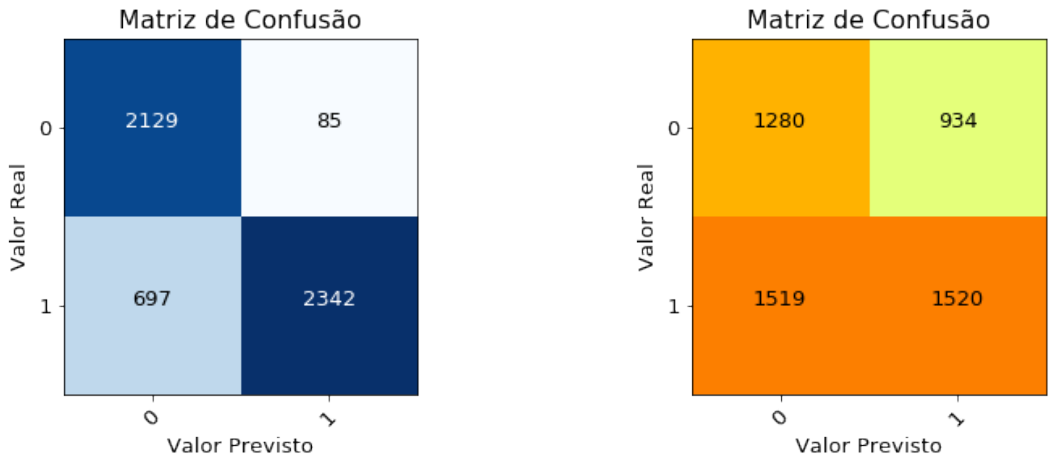
Nós utilizamos a biblioteca “scikit-learn” (PEDREGOSA et al., 2011) para a execução de nossos experimentos. Essa biblioteca contém implementações de todos os algoritmos utilizados, além de ser uma das mais utilizadas tanto na academia quanto no mercado. Utilizamos os valores padrão em todos os parâmetros.

Para avaliação dos modelos foram utilizadas as seguintes métricas: acurácia, precisão, Valor Preditivo Negativo, revocação, especificidade, valor-F1, Área Abaixo da Curva ROC e Coeficiente de Correlação de Matthews. Todas foram apresentadas na Subseção 2.2.3.

5.2 Resultados

As matrizes de confusão obtidas no experimento onde há entradas que representam as mesmas músicas estão nas Figuras 6, 7, 8 e 9. Na Tabela 10 estão os valores alcançados nas métricas de avaliação nesse experimento.

Em relação ao MP, nesse experimento o melhor resultado foi obtido, em termos de acurácia, ao utilizar o classificador SVM. Esse caso apresenta a menor quantia de falsos positivos do experimento com um valor 2,44 vezes menor em relação ao caso que utiliza de KNN, que tem a segunda menor quantia destas instâncias incorretamente previstas. Entretanto, ao utilizar SVM também foi obtida o maior número de falsos negativos, com uma quantia 0,69 vezes superior ao caso do uso de GNB, que apresentou a menor quantidade destas instâncias.



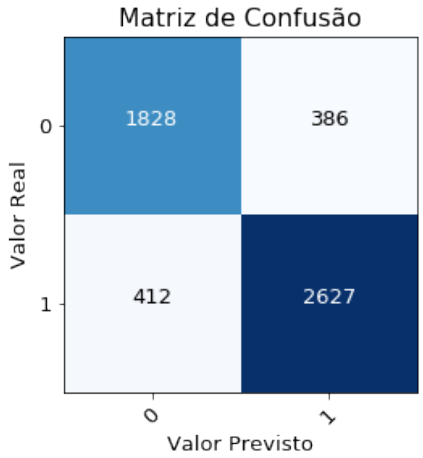
(a) Matriz de confusão obtida pelo MP. (b) Matriz de confusão obtida pelo ROM.

Figura 6 – Matrizes de confusão obtidas no experimento com repetição de músicas utilizando o classificador SVM.

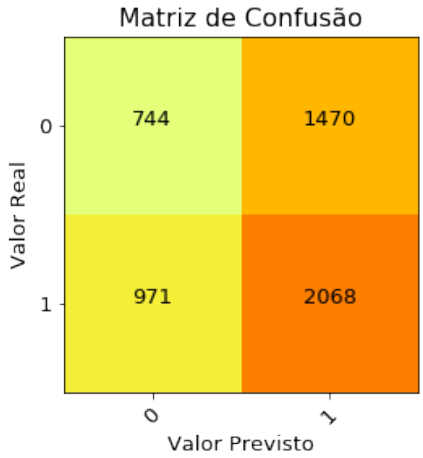
Devido a esses fatores, o caso com classificador SVM não obteve os melhores resultados em todas as métricas utilizadas para avaliação dos modelos. Entretanto, ele obteve o maior valor em MCC, que avalia a qualidade de uma classificação binária, o que indica que o resultado obtido por esse classificador foi o melhor de forma geral.

Tabela 10 – Desempenho dos modelos para o experimento onde há repetições de músicas na entrada.

	SVM		**GNB**		**LR**		**KNN**	
	MP	**ROM**	**MP**	**ROM**	**MP**	**ROM**	**MP**	**ROM**
Acurácia	**0,8511**	0,5330	0,8481	0,5353	0,8403	0,5336	0,8395	0,5433
Precisão	**0,9650**	0,6194	0,8719	0,5845	0,8734	0,5843	0,8947	0,6293
VPN	0,7534	0,4573	0,8161	0,4338	**0,7980**	0,4329	0,7774	0,4667
Revocação	0,7706	0,5002	0,8644	0,6805	**0,8467**	0,6719	0,8190	0,5123
Especificidade	**0,9616**	0,5781	0,8257	0,3360	0,8315	0,3437	0,8677	0,5858
Valor-F1	0,8569	0,5534	0,8681	0,6289	**0,8598**	0,6250	0,8552	0,5648
AUC	**0,8661**	0,5391	0,8450	0,5083	0,8391	0,5078	0,8433	0,5491
MCC	**0,7253**	0,0775	0,6890	0,0174	0,6748	0,0164	0,6793	0,0971

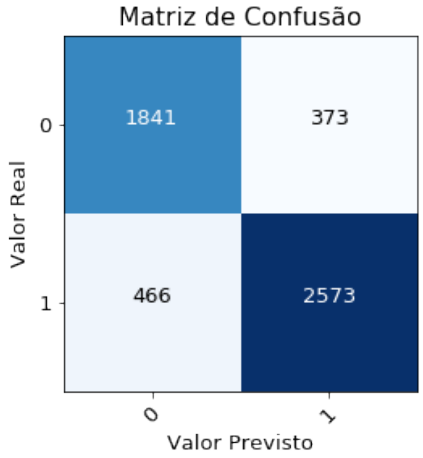


(a) Matriz de confusão obtida pelo MP.

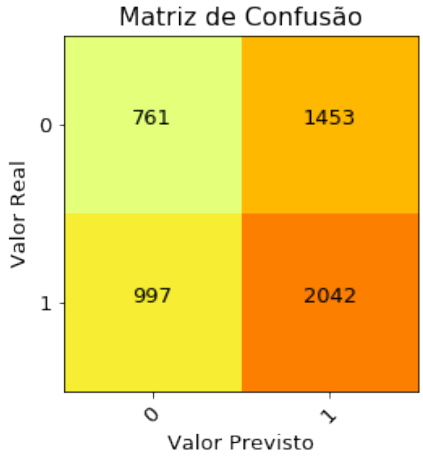


(b) Matriz de confusão obtida pelo ROM.

Figura 7 – Matrizes de confusão obtidas no experimento com repetição de músicas utilizando o classificador Gaussian Naive Bayes.

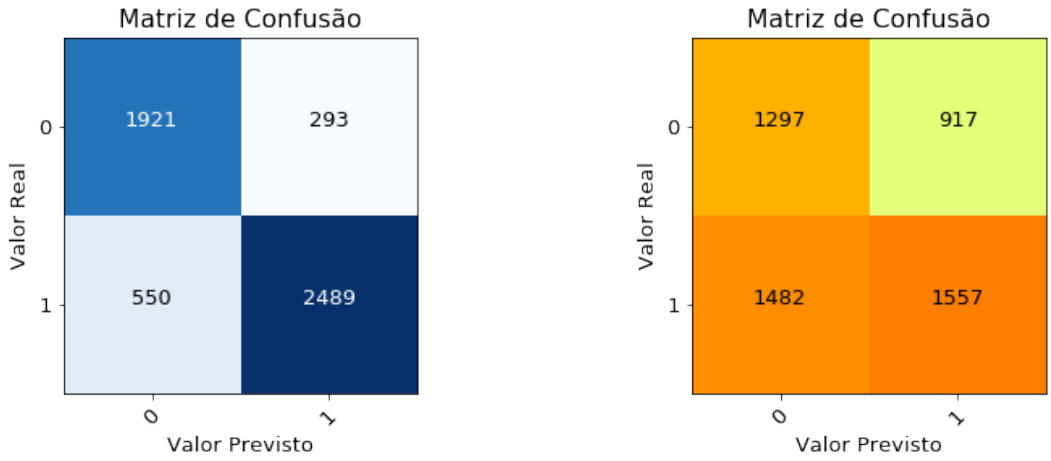


(a) Matriz de confusão obtida pelo MP.



(b) Matriz de confusão obtida pelo ROM.

Figura 8 – Matrizes de confusão obtidas no experimento com repetição de músicas utilizando o Regressão Logística.



(a) Matriz de confusão obtida pelo MP. (b) Matriz de confusão obtida pelo ROM.

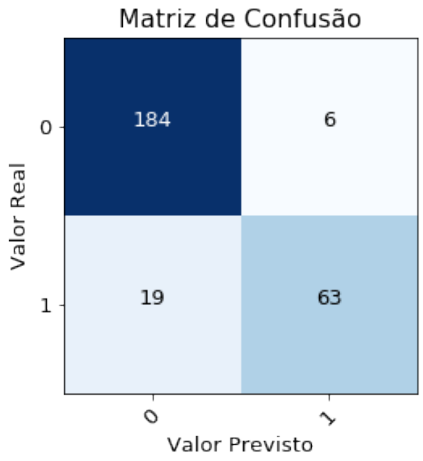
Figura 9 – Matrizes de confusão obtidas no experimento com repetição de músicas utilizando KNN.

As matrizes de confusão obtidas no experimento onde cada entrada representa uma música distinta estão nas Figuras 10, 11, 12 e 13. Na Tabela 11 estão os valores alcançados nas métricas de avaliação nesse experimento, os melhores resultados obtidos em cada uma das métricas estão evidenciados em vermelho.

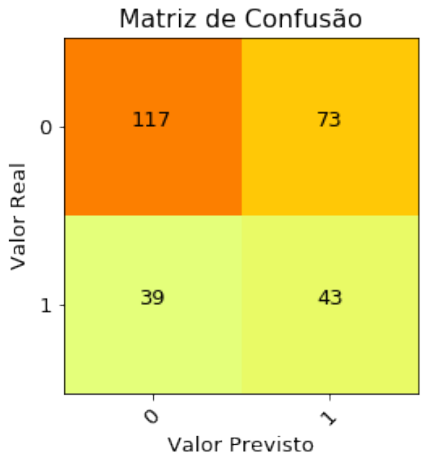
Nesse experimento, em relação ao MP, novamente o SVM obteve o maior valor em MCC e acurácia, o que indica que foi o que obteve o melhor resultado de forma geral. Nesse caso, a quantidade de falsos positivos e também a de falsos negativos foi a segunda menor em relação aos outros modelos. Assim, diferentemente do primeiro experimento, o maior valor-F1 também foi obtido pelo SVM.

Tabela 11 – Performance dos modelos para o experimento onde cada entrada representa uma música distinta.

	SVM		**GNB**		**LR**		**KNN**	
	MP	**ROM**	**MP**	**ROM**	**MP**	**ROM**	**MP**	**ROM**
Acurácia	**0,9081**	0,5882	0,8456	0,6324	0,8235	0,6838	0,8713	0,6360
Precisão	0,9130	0,3707	0,7381	0,3200	0,6667	0,3333	**0,9273**	0,3651
VPN	**0,9064**	0,7500	0,8936	0,7027	0,9176	0,7000	0,8571	0,7177
Revocação	0,7683	0,5244	0,7561	0,1951	**0,8293**	0,0488	0,6220	0,2805
Especificidade	0,9684	0,6158	0,8842	0,8211	0,8211	0,9579	**0,9789**	0,7895
Valor-F1	**0,8344**	0,4343	0,7470	0,2424	0,7391	0,0851	0,7445	0,3172
AUC	**0,8684**	0,5701	0,8603	0,5081	0,8560	0,5033	0,8004	0,5350
MCC	**0,7770**	0,1301	0,6360	0,0192	0,6164	0,0149	0,6866	0,0761

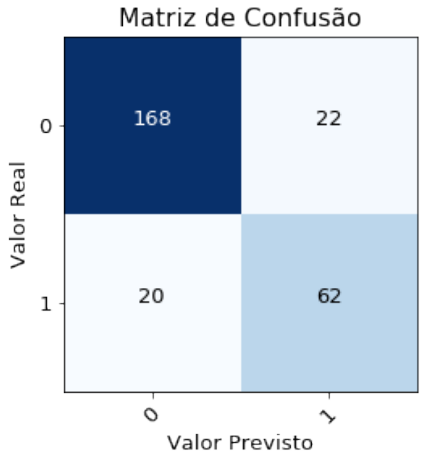


(a) Matriz de confusão obtida pelo MP.

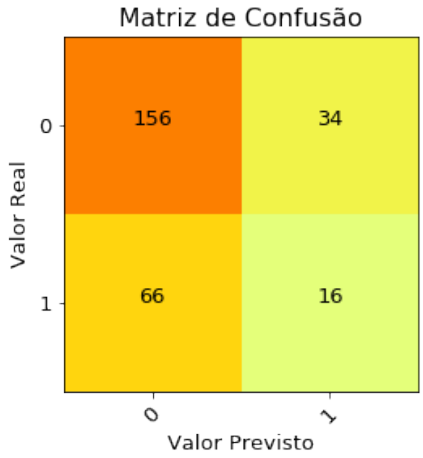


(b) Matriz de confusão obtida pelo ROM.

Figura 10 – Matrizes de confusão obtidas no experimento sem repetição de músicas utilizando o classificador SVM.



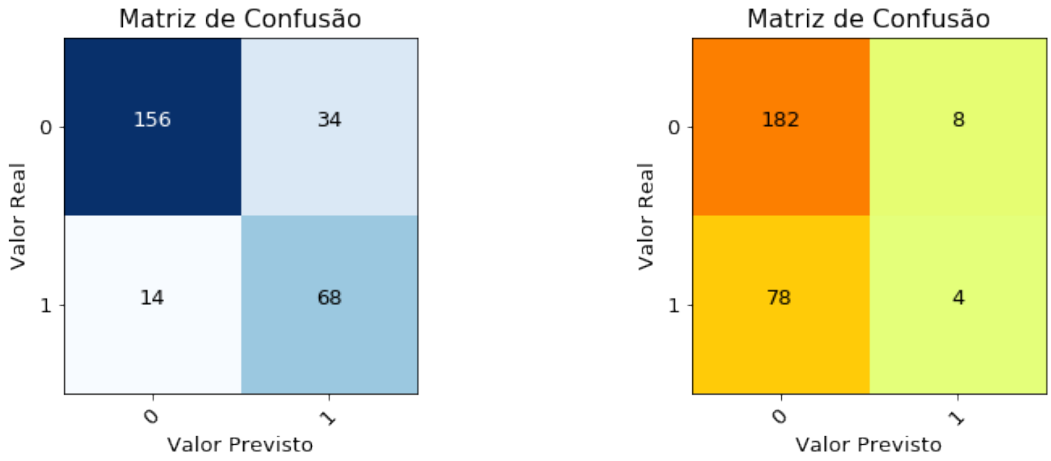
(a) Matriz de confusão obtida pelo MP.



(b) Matriz de confusão obtida pelo ROM.

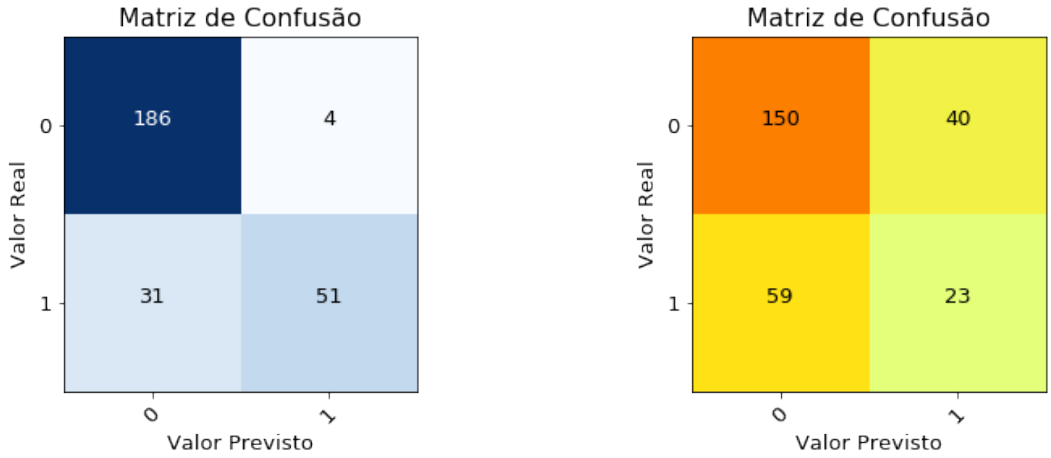
Figura 11 – Matrizes de confusão obtidas no experimento sem repetição de músicas utilizando o classificador Gaussian Naive Bayes.

Como já abordado anteriormente, o ROM não é uma replicação do modelo apresentado por Reiman e Örnell (2018), mas um modelo criado com base nesse texto em que modificações foram necessárias para que se adequasse aos experimentos realizados. Uma preocupação era a de que os resultados obtidos pelo ROM não necessariamente representariam os resultados que o modelo de Reiman e Örnell (2018) pudesse a vir obter. Entretanto, a diferença máxima obtida em pontos percentuais entre os resultados obtidos pelo ROM em nosso primeiro experimento e os resultados apresentados no texto base foi de apenas 6,64 pontos percentuais em acurácia quando utilizado o classificador GNB.



(a) Matriz de confusão obtida pelo MP. (b) Matriz de confusão obtida pelo ROM.

Figura 12 – Matrizes de confusão obtidas no experimento sem repetição de músicas utilizando o Regressão Logística.



(a) Matriz de confusão obtida pelo MP. (b) Matriz de confusão obtida pelo ROM.

Figura 13 – Matrizes de confusão obtidas no experimento sem repetição de músicas utilizando KNN.

Reiman e Örnell (2018) afirmaram não ser possível realizar previsões no mercado musical utilizando as informações levantadas. Ao criar um modelo baseado na metodologia proposta nesse trabalho, também não conseguimos realizar boas previsões, o MCC nos experimentos não ultrapassou 0,14 em nenhum dos casos, indicando uma classificação binária insatisfatória. Entretanto, nossa metodologia permitiu previsões com MCC superior a 0,7. Esse resultado indica que as *audio features* permitem, de fato, realizar previsões no mercado musical antes mesmo do lançamento das músicas.

5.3 Discussão

Nesta pesquisa desenvolvemos um modelo baseado em aprendizagem de máquina para prever se uma música ainda não lançada irá ou não se tornar popular. Especificamente, nós prevemos se uma música irá ou não aparecer *ranking* Top 50 do Spotify, porém nossa metodologia poderia ser reproduzida em qualquer plataforma de *streaming*. Realizamos dois experimentos, sendo que no primeiro cada instância representava uma música em um dia específico dos *rankings* coletados (Top 50 e Viral 50 do Spotify), de modo que havia várias instâncias idênticas que representavam as mesmas músicas. No segundo, as instâncias que representavam uma mesma música foram englobadas em uma só entrada. Assim, no primeiro experimento, os algoritmos foram treinados com 5389 instâncias e no segundo com 405.

No primeiro experimento, as previsões eram feitas para edições individuais do *ranking*. Isto é, uma música era considerada popular em um dia específico se aparecesse no Top 50 daquele dia. Em contraste, no segundo experimento as previsões eram feitas para um conjunto de edições do *ranking*. Nesse caso, para uma música ser considerada popular, ela deveria aparecer um certo número de vezes no Top 50. Decidimos estabelecer esse valor em quatro aparições, pois esse valor impede que as músicas que se destacaram tão somente entre os dias 23 a 26 de dezembro fossem consideradas populares.

Apesar dessa discrepância na quantidade de instâncias para treino nos experimentos, o MP obteve resultados semelhantes nos dois experimentos, mostrando que consegue realizar bons aprendizados mesmo com uma pequena quantidade de dados. O classificador SVM com *kernel* RBF obteve os maiores valores em MCC, AUC e acurácia nesses experimentos. Comparando os resultados dos experimentos nesse caso, a diferença em acurácia foi de 5,7 pontos percentuais, enquanto ficou em 0,23 em AUC e 5,17 em MCC, onde o segundo experimento obteve os maiores valores nestas métricas.

Os resultados obtidos pelo MP diferem-se daqueles obtidos pelo ROM. O MP apresentou, no melhor caso dos dois modelos, acurácia 56,65% superior

no primeiro experimento e MCC 921,02% superior no segundo. Uma possível explicação para o resultado ruim obtido pelo ROM é o fato de não haver qualquer tipo de preparação dos dados nessa metodologia. Reiman e Örnell (2018) não realizaram normalização dos atributos utilizados em sua pesquisa, assim dificultando o aprendizado de seu modelo por estarem expostos a valores atípicos e com grande variação. Por outro lado, em nosso modelo, além de não utilizarmos todo o conjunto de informações disponível pela API do Spotify, nós transformamos os atributos em campos binários. Esse processo retira a necessidade de normalização, facilita o aprendizado e ainda nos permitiu realizar previsões para músicas não lançadas.

Um estudo (PERCINO; KLIMEK; THURNER, 2015) já demonstrou que músicas populares tendem a soar parecidas. Esse estudo analisou 500 mil álbuns de 15 gêneros diferentes. Os autores avaliaram a complexidade de cada uma das músicas, calculada a partir das características acústicas das faixas, e compararam esses valores com a quantidade de vendas desses álbuns. Ao aplicar o coeficiente de correlação de Pearson nos dados, os autores obtiveram o valor de $-0,69$ com valor-p igual a $0,001$, o que demonstra a significância estatística desse resultado. Assim, demonstraram que há uma correlação linear negativa entre os dados. Tal resultado indica que quanto mais complexa uma música é, menos ela tende a obter maiores vendas. Essa situação pode explicar como nosso modelo proposto obteve resultados tão animadores, visto que como as músicas populares apresentam características semelhantes e nosso modelo aprende algumas dessas, logo consegue realizar boas previsões.

6 Considerações Finais e Trabalhos Futuros

Este capítulo está dividido em duas seções, sendo a Seção 6.1 destinada às considerações finais e a Seção 6.2 dedicada a apresentar limitações de nossa pesquisa e possíveis trabalhos futuros.

6.1 Considerações Finais

Neste trabalho foi apresentado um modelo de previsão para o sucesso no mercado musical. Tal modelo tem importante valor comercial, pois pode ser utilizado por artistas e gravadoras para decidir em quais músicas investir maior esforço para que se obtenha melhor retorno comercial.

Esta pesquisa está inserida na área de *Hit Song Science*, que tem como objetivo estudar formas de se fazer previsões de popularidade de músicas antes mesmo de serem lançadas. De forma geral, trabalhos dessa área utilizam dados sobre as características acústicas das músicas, de redes sociais e/ou de shows e festivais. Entretanto, nosso modelo não utiliza nenhuma destas com o intuito de analisar os resultados obtidos com um novo tipo de fonte.

Atualmente, o *streaming* é a principal forma de consumo de música. Portanto, decidimos desenvolver nosso modelo para realizar previsões nesse âmbito. Especificamente, nosso modelo prevê se uma música irá ou não aparecer no Top 50 Global do Spotify. O Spotify é um dos maiores serviços de *streaming* de música nos dias atuais.

Para criação de nosso modelo, montamos uma base contendo músicas que já haviam aparecido no Top 50 e outras que nunca estiveram lá. Usando a própria API da plataforma extraímos informações sobre as músicas da base. Essas informações tratam de características das músicas, que indicam se elas são dançantes, acústicas, instrumentais, dentre outras. Esses dados são

coletados em números reais a partir da própria plataforma. Para permitir a inclusão de músicas ainda não lançadas, decidimos binarizar esses atributos. Dessa forma, o artista ou a gravadora podem determinar se suas músicas apresentam ou não essas características sem necessitar fazer uso da API.

Além de nosso modelo proposto, também desenvolvemos outro baseado na metodologia utilizada por Reiman e Örnell (2018) de forma a comparar os resultados obtidos.

Nós realizamos dois experimentos. No primeiro as previsões foram feitas por dia, isto é, buscávamos prever quais músicas seriam populares em um dia específico. Assim, uma música que só aparecesse no Top 50 uma única vez era considerada popular naquele dia específico. Portanto, esse experimento contava com instâncias repetidas com possíveis classes distintas. Por outro lado, no segundo as previsões foram feitas por música, assim cada instância representava uma canção distinta e ela só era considerada popular se tivesse aparecido ao menos quatro vezes no período de teste. Apesar das diferenças, o modelo obteve resultados parecidos nos dois experimentos com diferença máxima de 5,7 pontos percentuais em acurácia.

O modelo proposto obteve acurácia, precisão e AUC acima de 80% em todos os casos. No melhor caso, ao utilizar o classificador SVM com *kernel* RBF, o resultado chegou a ser mais de 920% superior, segundo o Coeficiente de Correlação de Matthews, em relação ao modelo baseado em Reiman e Örnell (2018).

6.2 Limitações e Trabalhos Futuros

O objetivo esperado neste trabalho de **realizar previsões de sucesso de músicas em plataformas de *streaming* antes mesmo de serem lançadas** foi alcançado. Para isso desenvolvemos um modelo baseado em informações sobre as músicas que obteve resultados superiores comparado a outro da literatura.

Entretanto, nosso estudo apresenta uma limitação. Ela é decorrente do

fato de utilizarmos músicas provenientes do *ranking* de virais do Spotify como músicas não populares. A presença nesse *ranking* indica que essas músicas atingiram uma certa popularidade, mesmo que mais baixa que as presentes no Top 50. Idealmente, o conjunto de músicas não populares deveria ser composto por músicas que nunca alcançaram o Viral 50. Todavia, não foi possível gerar tal conjunto devido a limitações da API do Spotify que não permite a coleta das informações utilizadas de músicas não presentes em *rankings*.

Como trabalhos futuros, delineamos alguns pontos a melhorar no modelo proposto. O principal ponto a se desenvolver é o de encontrar uma forma de diminuir a quantidade de falsos negativos. Ao utilizar o algoritmo SVM foram obtidos os maiores valores em MCC, mas não em revocação, em comparação aos outros algoritmos. No pior caso, a quantia de falsos negativos foi 0,69 vezes superior.

No Apêndice A apresentamos trabalhos realizados no âmbito de identificação de fatores de influência no mercado musical. Em um desses trabalhos mostramos que mensagens no Twitter apresentam correlação com a popularidade de um álbum no Spotify (ARAUJO et al., 2017a). Logo, pretendemos avaliar se a utilização de informações provenientes de redes sociais podem auxiliar o modelo proposto a realizar melhores previsões.

Por fim, um desejo futuro é acertar parcerias com gravadoras e artistas para aplicação do modelo proposto em músicas antes de serem lançadas. Em nosso experimento, por não termos acesso as essas músicas, os testes foram feitos considerando músicas já lançadas como se fossem canções inéditas.

Referências

ABEL, F.; DIAZ-AVILES, E.; HENZE, N.; KRAUSE, D.; SIEHNDEL, P. Analyzing the blogosphere for predicting the success of music and movie products. In: *2010 International Conference on Advances in Social Networks Analysis and Mining*. [S.l.: s.n.], 2010. p. 276–280. 18

ARAKELYAN, S.; MORSTATTER, F.; MARTIN, M.; FERRARA, E.; GALSTYAN, A. Mining and forecasting career trajectories of music artists. *CoRR*, abs/1805.03324, 2018. Disponível em: <http://arxiv.org/abs/1805.03324>. 18, 19, 32, 35

ARAUJO, C. V.; CRISTO, M. A. P.; GIUSTI, R. Predicting music popularity on streaming platforms. In: *Proceedings of the 17th Brazilian Symposium on Computer Music*. [S.l.]: Sociedade Brasileira de Computação, 2019. 37, 39, 40, 41

ARAUJO, C. V.; CRISTO, M. A. P.; GIUSTI, R. Predicting music popularity using music charts. In: *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.: s.n.], 2019. 37, 43, 47

ARAUJO, C. V.; CRISTO, M. A. P.; GIUSTI, R. Will i remain popular? a study case on spotify. In: *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*. [S.l.]: SBC, 2019. 37, 44, 45

ARAUJO, C. V.; NETO, R. M.; NAKAMURA, F. G.; NAKAMURA, E. F. Predicting music success based on users' comments on online social networks. In: *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*. ACM, 2017. (WebMedia '17), p. 149–156. ISBN 978-1-4503-5096-9. Disponível em: <http://doi.acm.org/10.1145/3126858.3126885>. 9, 65, 71, 72, 74, 82

ARAUJO, C. V.; NETO, R. M.; NAKAMURA, F. G.; NAKAMURA, E. F. Using complex networks to assess collaboration in rap music: A study case of dj khaled. In: *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*. ACM, 2017. (WebMedia '17), p. 425–428. ISBN 978-1-4503-5096-9. Disponível em: <http://doi.acm.org/10.1145/3126858.3131605>. 9, 75, 77, 78

ARAUJO, C. V. S.; NAKAMURA, E. F. Identification of most popular musical genres and their influence factors. In: *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*. ACM, 2018. (WebMedia '18), p. 233–236. ISBN 978-1-4503-5867-5. Disponível em: <http://doi.acm.org/10.1145/3243082.3264665>. 79, 80, 91, 92

ATWOOD, B. *The History of the Music Industry's First-Ever Digital Single For Sale, 20 Years After Its Release*. 2017. Disponível em: <https://www.billboard.com/articles/business/7964771/history-music-industry-first-ever-digital-single-20-years-later>. 22

BHATTACHARJEE, A. Distance correlation coefficient: An application with bayesian approach in clinical data analysis. *Journal of Modern Applied Statistical Methods*, v. 13, n. 1, p. 23, 2014. 24

BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, IOP Publishing, v. 2008, n. 10, p. P10008, 2008. 76

BRANDES, U. A faster algorithm for betweenness centrality. *Journal of mathematical sociology*, Taylor & Francis, v. 25, n. 2, p. 163–177, 2001. 76

BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, Oct 2001. ISSN 1573-0565. Disponível em: <https://doi.org/10.1023/A:1010933404324>. 29

COHEN, J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd. ed. [S.l.]: Routledge, 1988. 400 p. 73

COLONNA, J. G. et al. *Uma abordagem para classificação de anuros baseada em vocalizações*. Dissertação (Mestrado), 2012. Instituto de Computação. Disponível em: <http://tede.ufam.edu.br/handle/tede/2964>. 27

DEISENROTH, M. P.; FAISAL, A. A.; ONG, C. S. *Mathematics For Machine Learning*. [S.l.]: Cambridge University Press, 2020. 73

DHAR, V.; CHANG, E. A. Does chatter matter? the impact of user-generated content on music sales. *Journal of Interactive Marketing*, v. 23, n. 4, p. 300 – 307, 2009. ISSN 1094-9968. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1094996809000723>. 19, 35

DUBNOV, S. Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters*, v. 11, n. 8, p. 698–701, Aug 2004. ISSN 1070-9908. 23

FAWCETT, T. An introduction to roc analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861 – 874, 2006. ISSN 0167-8655. ROC Analysis in Pattern Recognition. Disponível em: <http://www.sciencedirect.com/science/article/pii/S016786550500303X>. 30

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, v. 55, n. 1, p. 119 – 139, 1997. ISSN 0022-0000. Disponível em: <http://www.sciencedirect.com/science/article/pii/S002200009791504X>. 28

GIUSTI, R.; SILVA, D. F.; BATISTA, G. E. A. P. A. Improved time series classification with representation diversity and svm. In: *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. [S.l.: s.n.], 2016. p. 1–6. 27

HANLEY, J. A.; MCNEIL, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, v. 143, n. 1, p. 29–36, 1982. PMID: 7063747. Disponível em: <https://doi.org/10.1148/radiology.143.1.7063747>. 30

HASTIE, T.; ROSSET, S.; ZHU, J.; ZOU, H. Multi-class adaboost. *Statistics and its Interface*, International Press of Boston, v. 2, n. 3, p. 349–360, 2009. 28

HERREMANS, D.; MARTENS, D.; SöRENSEN, K. Dance hit song prediction. *Journal of New Music Research*, Routledge, v. 43, n. 3, p. 291–302, 2014. Disponível em: <https://doi.org/10.1080/09298215.2014.881888>. 19, 23, 33, 35, 50

KARYDIS, I.; GKIOKAS, A.; KATSOUROS, V.; ILIADIS, L. Musical track popularity mining dataset: Extension & experimentation. *Neurocomputing*, v. 280, p. 76 – 85, 2018. ISSN 0925-2312. Applications of Neural Modeling in the new era for data and IT. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0925231217317666>. 18, 23, 33, 35

KIM, Y.; SUH, B.; LEE, K. #nowplaying the future billboard: Mining music listening behaviors of twitter users for hit song prediction. In: *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*. New York, NY, USA: ACM, 2014. (SoMeRA '14), p. 51–56. ISBN 978-1-4503-3022-0. Disponível em: <http://doi.acm.org/10.1145/2632188.2632206>. 19, 33, 35

KLEIN, A. *How Record Labels Work*. HowStuffWorks, 2003. Disponível em: <https://entertainment.howstuffworks.com/record-label.htm>. 22

LAMERE, P. Social tagging and music information retrieval. *Journal of New Music Research*, Routledge, v. 37, n. 2, p. 101–114, 2008. Disponível em: <https://doi.org/10.1080/09298210802479284>. 23

LEE, J.; LEE, J. Music popularity: Metrics, characteristics, and audio-based prediction. *IEEE Transactions on Multimedia*, v. 20, n. 11, p. 3173–3182, Nov 2018. ISSN 1520-9210. 18

LEHMANN, E. L.; CASELLA, G. *Theory of point estimation*. [S.l.]: Springer Science & Business Media, 2006. 30

LI, T.; LI, L. Music data mining: An introduction. *Music data mining*, p. 1, 2011. 22

LI, T.; OGIHARA, M.; TZANETAKIS, G. *Music data mining*. [S.l.]: CRC Press, 2011. 17

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. *Natural Language Engineering*, Cambridge university press, v. 16, n. 1, p. 100–103, 2010. 27

MATTHEWS, B. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, v. 405, n. 2, p. 442 – 451, 1975. ISSN 0005-2795. Disponível em: <http://www.sciencedirect.com/science/article/pii/0005279575901099>. 30

MCFEE, B.; RAFFEL, C.; LIANG, D.; ELLIS, D. P.; MCVICAR, M.; BATTENBERG, E.; NIETO, O. librosa: Audio and music signal analysis in python. In: *Proceedings of the 14th python in science conference*. [S.l.: s.n.], 2015. p. 18–25. 23, 38

MOLANPHY, C. *How The Hot 100 Became America's Hit Barometer*. 2013. Disponível em: <https://www.npr.org/sections/therecord/2013/08/16/207879695/how-the-hot-100-became-americas-hit-barometer>. 24

NETO, R. M.; SOUZA, B. A.; ALMEIRA, T. G.; NAKAMURA, F. G.; NAKAMURA, E. F. Uma abordagem para identificação de entidades influentes em eventos comentados nas redes sociais online. *Simpósio Brasileiro de Sistemas Colaborativos (SBSC)*, 2017. 76

NEWMAN, M. E. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, National Acad Sciences, v. 103, n. 23, p. 8577–8582, 2006. 76

NI, Y.; SANTOS-RODRíGUEZ, R.; MCVICAR, M.; BIE, T. D. Hit song science once again a science? In: *4th International Workshop on Machine Learning and Music*. [S.l.: s.n.], 2011. 17, 31, 35

OLSON, D. L.; DELEN, D. *Advanced data mining techniques*. [S.l.]: Springer Science & Business Media, 2008. 29

ORIO, N. Music retrieval: A tutorial and review. *Foundations and Trends® in Information Retrieval*, v. 1, n. 1, p. 1–90, 2006. ISSN 1554-0669. Disponível em: <http://dx.doi.org/10.1561/1500000002>. 23

PACHET, F. Hit song science. *Music data mining*, Chapman & Hall/CRC Press Boca Raton, FL, p. 305–326, 2011. 17

PACHET, F.; ROY, P. Hit song science is not yet a science. In: *ISMIR*. [S.l.: s.n.], 2008. p. 355–360. 17, 31, 35

PEARSON, K. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, JSTOR, v. 58, p. 240–242, 1895. 25

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. 55

PERCINO, G.; KLIMEK, P.; THURNER, S. Instrumentational complexity of music genres and why simplicity sells. *PLOS ONE*, Public Library of Science, v. 9, n. 12, p. 1–16, 12 2015. Disponível em: <https://doi.org/10.1371/journal.pone.0115255>. 62

REIMAN, M.; ÖRNELL, P. Predicting hit songs with machine learning. In: . [S.l.: s.n.], 2018. (TRITA-EECS-EX, 2018:202). 18, 19, 34, 36, 50, 53, 54, 55, 59, 60, 62, 64

SAMSON, J. *Genre*. Oxford University Press, 2001. Disponível em: <http:////www.oxfordmusiconline.com/grovemusic/view/10.1093/gmo/9781561592630.001.0001/omo-9781561592630-e-0000040599>. 22

SCHöLKOPF, B.; PLATT, J. C.; SHAWE-TAYLOR, J.; SMOLA, A. J.; WILLIAMSON, R. C. Estimating the support of a high-dimensional distribution. *Neural Computation*, v. 13, n. 7, p. 1443–1471, 2001. Disponível em: <https://doi.org/10.1162/089976601750264965>. 27

SPEARMAN, C. The proof and measurement of association between two things. *The American journal of psychology*, JSTOR, v. 15, n. 1, p. 72–101, 1904. 25

STEININGER, D. M.; GATZEMEIER, S. Using the wisdom of the crowd to predict popular music chart success. In: *ECIS*. [S.l.: s.n.], 2013. p. 215. 19, 32, 35

STOREY, J. D. The positive false discovery rate: a bayesian interpretation and the q -value. *Ann. Statist.*, The Institute of Mathematical Statistics, v. 31, n. 6, p. 2013–2035, 12 2003. Disponível em: <https://doi.org/10.1214/aos/1074290335>. 29

SZÉKELY, G. J.; RIZZO, M. L.; BAKIROV, N. K. et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 35, n. 6, p. 2769–2794, 2007. 25

SZÉKELY, G. J.; RIZZO, M. L. et al. Brownian distance covariance. *The annals of applied statistics*, Institute of Mathematical Statistics, v. 3, n. 4, p. 1236–1265, 2009. 26

THELWALL, M.; BUCKLEY, K.; PALTOGLOU, G.; CAI, D.; KAPPAS, A. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, v. 61, n. 12, p. 2544–2558, 2010. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21416>. 72

ZHANG, H. The optimality of naive bayes. *AA*, v. 1, n. 2, p. 3, 2004. 27, 44

APÊNDICE A – Trabalhos Prévios

Anterior ao desenvolvimento do modelo de previsão, realizamos pesquisas com o intuito de identificar fatores de influência no mercado musical. Considerávamos que o entendimento de como se dava a influência desses fatores seria necessário em nosso modelo. Entretanto, nossa pesquisa seguiu por um caminho distinto, logo esses fatores acabaram por não serem utilizados. Mas, devido à relevância científica dos resultados obtidos nessa etapa, decidimos dedicar este apêndice à apresentação dessas pesquisas. Cada uma das seções neste capítulo abordam um diferente trabalho publicado. Na Seção A.1 tratamos do impacto de mensagens no Twitter na popularidade de álbuns, enquanto na Seção A.2 estudamos como se dá a colaboração entre artistas no Rap, por fim na Seção A.3 analisamos os fatores de influência na popularidade de gêneros musicais.

A.1 *Predicting Music Success Based on Users' Comments on Online Social Networks*

Nessa pesquisa (ARAUJO et al., 2017a) apresentamos o impacto de mensagens na rede social Twitter (tweets) na popularidade de álbuns no Spotify e na revista americana Billboard. Em nossa análise, estudamos se mensagens com polaridade positiva ou negativa, isto é, mensagens que falam bem ou mal de um álbum, apresentam maior impacto ou se é tão somente o volume de mensagens que importa.

A primeira etapa da metodologia deste trabalho consistiu em coletar o número de unidades atingidas no Billboard Top 200[1] e o valor em popularidade no Spotify uma semana após o lançamento de álbuns de destaque de 2016 e 2017. Para cada um desses álbuns foram apanhados tweets relacionados a eles

[1] A quantia em unidades atingida por um álbum é um valor que leva em conta o número de vendas físicas e digitais, além da quantidade de *streamings*, conforme apresentado em <http://www.billboard.com/articles/columns/chart-beat/6320099/billboard-200-makeover-streams-digital-tracks>.

e seus artistas publicados nos 30 dias que antecederam os seus lançamentos. Dessas mensagens, somente aquelas em inglês foram selecionadas, possibilitando a detecção de polaridade que foi realizada utilizando o Sentistrength (THELWALL et al., 2010).

Por fim, foi feita a análise estatística dos dados, onde foram aplicados os coeficientes de correlação apresentados na Subseção 2.2.1 e também a análise Probabilidade-Probabilidade (P-P). Esta verifica a similaridade entre as funções de probabilidades acumuladas das duas variáveis contrapostas. Assim, foi possível identificar se havia relação entre as mensagens e a popularidade desses álbuns. Uma representação gráfica dessa metodologia pode ser vista na Figura 14.

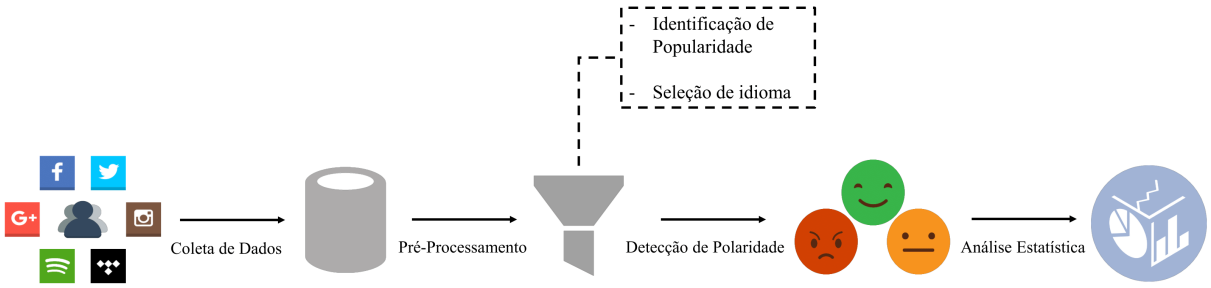


Figura 14 – Metodologia utilizada em (ARAUJO et al., 2017a).

Um descritivo dos dados utilizados nessa pesquisa está no Apêndice B. A Figura 15a apresenta o gráfico de dispersão da Popularidade Spotify vs. Quantidade de Tweets. O gráfico inclui a curva de regressão linear dos pontos (em vermelho) com o respectivo intervalo de confiança (em cinza). Nesse caso, ao serem considerados apenas Tweets Positivos, os pontos se aproximam mais da curva de regressão.

Os testes estatísticos consideram a hipótese nula de que há correlação entre as variáveis. Os resultados mostram que, considerando Tweets Positivos, o valor-p de todos os testes estão abaixo de $\alpha = 0{,}05$. Portanto, os três coeficientes de correlação calculados são estatisticamente significativos. Consequentemente, há uma correlação positiva entre a quantidade de Tweets Positivos, publicados nos 30 dias que antecedem o lançamento do álbum, e sua Popularidade Spotify, após o lançamento.

O coeficiente de Pearson de $0{,}665 > 0{,}5$ indica uma forte correlação

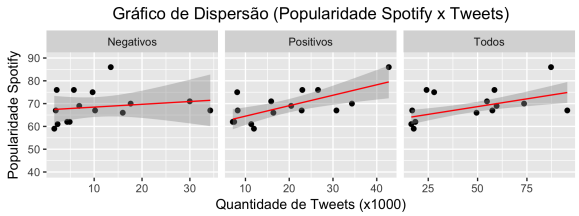
linear positiva, conforme explica Cohen (1988). Além disso, o teste-t com valor-p igual a 0,007 $< \alpha =$ 0,05, mostra que esse resultado possui significância estatística.

Ao confrontar as Funções de Distribuição Acumulada (FDAs), que indicam a probabilidade de uma variável aleatória ser menor ou igual a um valor real x (DEISENROTH; FAISAL; ONG, 2020) (Figura 15b), observa-se que a distribuição dos Tweets Positivos é a que mais se assemelha à distribuição da Popularidade Spotify. Nesse caso, o erro quadrático médio é 0,019 com variância residual 0,004 (Tabela 12a). Esses resultados são estatisticamente suportados pelos valores-p superiores a $\alpha =$ 0,05, aceitando a hipótese do teste de Wilcoxon de que as distribuições são equivalentes.

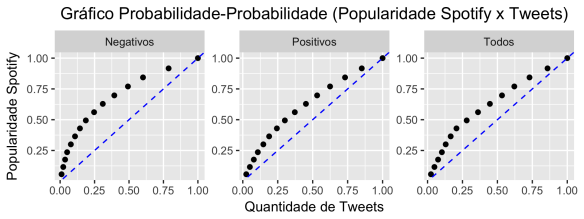
As mesmas observações são válidas considerando Todos os Tweets. Contudo, as estatísticas para Todos os Tweets são um pouco menos significativas do que as para Tweets Positivos. Considerando os Tweets Negativos, a hipótese de que há correlação é rejeitada para todos os testes feitos, pois o valor-p deles é superior a $\alpha =$ 0,05 (Tabela 12a).

Ao analisar os gráficos P-P da Figura 15d, observa-se que as FDAs das Unidades Billboard e das Quantidades de Tweets aproximam-se do comportamento teórico. Essa análise visual é confirmada pelos valores de erro quadrático médio e variância residual, todos menores ou iguais a 0,003 e 0,001, respectivamente (Tabela 12b).

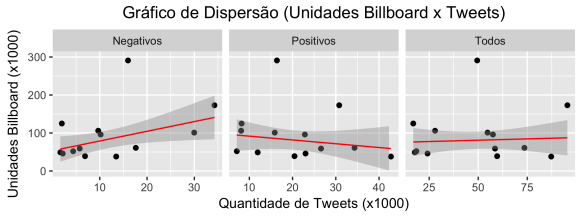
Por outro lado, a correlação entre as variáveis "Unidades Billboard vs. Tweets" é menos acentuada, principalmente para Tweets Positivos e Todos os Tweets, quando comparada com a análise da Popularidade Spotify. Esse fato é ratificado pelos testes de Wilcoxon, cujos valores-p são todos superiores a $\alpha =$ 0,05, aceitando a hipótese de que o comportamento das distribuições é semelhante (em termos de FDA). Nesse caso, não há evidência estatística suficiente para suportar a afirmação de que há correlação entre a quantidade de tweets publicados 30 dias antes do lançamento de um álbum e seu sucesso no *ranking* da Billboard.



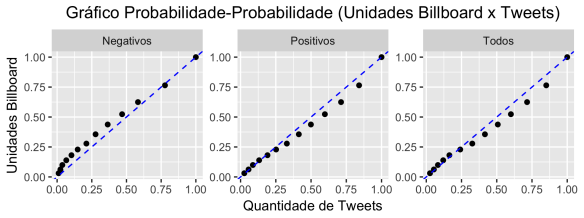
(a) Análise de Dispersão da *Popularidade Spotify vs. Quantidade de Tweets* (negativos, positivos e total) com a reta de regressão em vermelho e intervalo de confiança ($\alpha = 0{,}05$) em cinza escuro.



(b) Análise Probabilidade-Probabilidade da *Popularidade Spotify vs. Quantidade de Tweets* (negativos, positivos e total). A reta tracejada representa o comportamento no caso de equivalência.



(c) Análise de Dispersão da *Unidades Billboard vs. Quantidade de Tweets* (negativos, positivos e total) com a reta de regressão em vermelho e intervalo de confiança ($\alpha = 0{,}05$) em cinza escuro.



(d) Análise Probabilidade-Probabilidade da *Unidades Billboard vs. Quantidade de Tweets* (negativos, positivos e total). A reta tracejada representa o comportamento no caso de equivalência.

Figura 15 – Análises gráficas: (a) Dispersão para Popularidade Spotify vs. Tweets, (b) Probabilidade-Probabilidade para Popularidade Spotify vs. Tweets, (c) Dispersão para Unidades Billboard vs. Tweets, (d) Probabilidade-Probabilidade para Unidades Billboard vs. Tweets. Extraído de (ARAUJO et al., 2017a).

Tabela 12 – Avaliação de correlação estatística entre: (a) Popularidade Spotify vs. Tweets e (b) Unidades Billboard vs. Tweets. Análise P-P verifica a similaridade entre as funções de probabilidades acumuladas das duas variáveis contrapostas, sendo EQM o erro quadrático médio e σ_r^2 a variância residual. Para Pearson, Spearman e Distância, a hipótese nula é H_0: as variáveis não estão correlacionadas. Para Wilcoxon, a hipótese nula é H_0: as distribuições são equivalentes. Extraído de (ARAUJO et al., 2017a).

(a) Popularidade Spotify vs. Tweets

	Pearson		Spearman		Distância		Análise P-P		Wilcoxon	
	Estatística	valor-p	Estatística	valor-p	Estatística	valor-p	EQM	σ_r^2	Estatística	valor-p
Tweets Negativos	0,501	0,189	0,186	0,361	0,433	0,227	0,053	0,011	160,00	0,050
Tweets Positivos ⋆	0,665	**0,007**	0,592	**0,020**	0,636	**0,021**	**0,019**	**0,004**	**140,000**	**0,267**
Todos os Tweets	0,519	0,048	0,596	0,019	0,580	0,032	0,022	0,004	144,000	0,202

(b) Unidades Billboard vs. Tweets

	Pearson		Spearman		Distância		Análise P-P		Wilcoxon	
	Estatística	valor-p	Estatística	valor-p	Estatística	valor-p	EQM	σ_r^2	Estatística	valor-p
Tweets Negativos	0,429	0,144	0,407	0,170	0,545	0,168	0,003	0,001	96,500	0,555
Tweets Positivos	−0,176	0,566	−0,253	0,404	0,365	0,727	0,002	0,001	82,500	0,939
Todos os Tweets	0,113	0,712	−0,093	0,765	0,341	0,822	0,003	0,001	84,000	1,000

Um possível motivo para existir correlação com a popularidade do Spotify, mas não com as unidades da Billboard, é que o peso da quantidade de *streamings* no valor em unidades é muito menor ao de uma venda de um álbum. Durante a coleta dos dados, observou-se que o número de vendas dos discos continua sendo o que mais influencia no valor em unidades. Álbuns lançados somente em serviços digitais nunca haviam conseguido alcançar a primeira posição do *ranking*, até então. Por exemplo, o álbum *Coloring Book*, de Chance the Rapper, foi lançado somente em serviços de *streaming*, obtendo 38.000 unidades da Billboard em sua primeira semana, equivalentes a 57,3 milhões de *streamings*[2]. Esse é o álbum com a maior Popularidade Spotify (no estudo realizado) e tem o maior valor na escala positiva de sentimentos. Até mesmo artistas que priorizaram o lançamento por *streaming* só conseguiram atingir o topo ao somar as vendas. Esse cenário ocorreu com o rapper Kanye West, cujo disco foi o primeiro a atingir o topo do *ranking* da Billboard com prioritariamente o número de *streams*[3].

A.2 *Using Complex Networks to Assess Collaboration in Rap Music: A Study Case of DJ Khaled*

Nesse segundo trabalho (ARAUJO et al., 2017b), o objetivo era estudar o funcionamento da colaboração entre artistas no Rap. Para isso, realizamos um estudo de caso das colaborações que ocorrem nas músicas do DJ Khaled, músico popularmente conhecido pela grande quantidade de artistas participando em suas faixas. Por exemplo, a música "Welcome to My Hood (Remix)" do seu álbum "We the Best Forever" conta com a participação de 12 cantores distintos.

Utilizando a API do serviço de *streaming* Napster[4], coletamos os dados dos artistas que realizam colaborações entre si nas músicas do DJ Khaled. Utilizando-se dessas informações montamos grafos onde cada artista era um nó e havia uma aresta entre eles caso já houvessem colaborado em uma mesma

[2] <http://www.billboard.com/articles/columns/chart-beat/7378360/drakes-views-no-1-billboard-200-album-chart-meghan-trainor-thank-you>

[3] <http://www.billboard.com/articles/columns/chart-beat/7326493/kanye-wests-the-life-of-pablo-debuts-at-no-1-on-billboard-200>

[4] <https://developer.napster.com/api/v2.1>

música. O peso das arestas foi dado pela quantidade de colaborações entre os dois artistas por elas conectados.

Para estudarmos o impacto da influência de determinados músicos ao longo do tempo aplicamos à rede a medida de centralidade *betweenness* (BRANDES, 2001). Utilizamos essa medida, pois apresenta a frequência que um vértice i aparece nos caminhos mínimos entre todos os pares de vértices da rede (NETO et al., 2017). No contexto estudado pudemos verificar qual colaborador possuía uma ligação mais central em relação aos outros. Logo, esse artista apresenta maior destaque na rede, por interligar o maior número de músicos.

Para detectar comunidades na rede e estudar a interação entre elas utilizamos a modularidade, uma função que mede a qualidade da divisão de uma rede em grupos ou comunidades (NEWMAN, 2006). Nesse trabalho aplicamos o algoritmo de *Louvain*, um método heurístico que se baseia na maximização da modularidade para guiar o processo de construção das comunidades. Matematicamente, o ganho da função de modularidade é definido por:

$$\Delta Q = \frac{\sum_{in} + k_{i,in}}{2m} - \left(\frac{\sum_t + k_i}{2m}\right)^2 - \frac{\sum_{in}}{2m} + \left(\frac{\sum_t}{2m}\right)^2 + \left(\frac{k_i}{2m}\right)^2, \tag{A.1}$$

onde $\sum_{in}$ é a soma dos pesos das arestas na comunidade C_x, $\sum_t$ é a soma dos pesos das arestas incidentes nos nós em C_x, k_i é a soma dos pesos das arestas incidentes no nó i, $k_{i,in}$ é a soma dos pesos das arestas partindo do nó i para todos os nós em C_x e m é a soma dos pesos de todas as arestas da rede. Todo esse processo é aplicado repetidas vezes, até que a modularidade não possa mais ser aprimorada (BLONDEL et al., 2008).

Montamos um total de nove redes distintas (uma para cada álbum lançado por Khaled até o ano de 2016). Cada uma dessas redes incorpora a rede anterior, acrescentando as novas colaborações de cada álbum. Por exemplo, a rede do segundo álbum conta com as colaborações do primeiro e segundo disco e assim sucessivamente. Dessa forma, pôde-se analisar a evolução da rede e perceber a diferença de influência das entidades. Uma representação gráfica dessa metodologia pode ser vista na Figura 16.

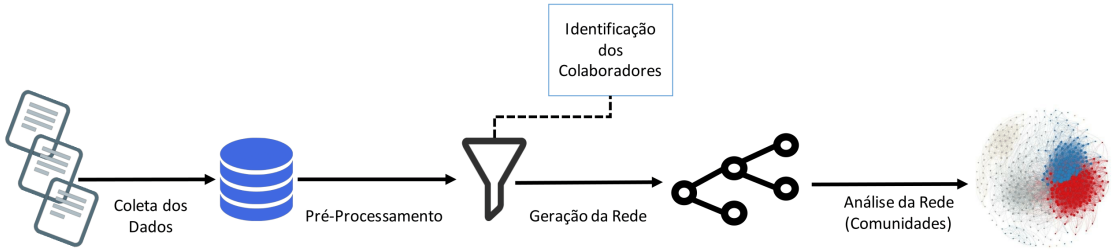


Figura 16 – Metodologia utilizada em (ARAUJO et al., 2017b).

Somente a partir da rede contando com as colaborações dos seis primeiros álbuns foi possível identificar características bem-definidas nas comunidades evidenciadas. Os artistas que colaboraram nos três álbuns seguintes e que não haviam aparecido nos discos anteriores geralmente passavam a fazer parte de uma das comunidades que já existiam. A nona rede, que engloba os dados de todos os discos até então lançados, pode ser visualizada na Figura 17. As outras redes geradas estão no Apêndice C.

Na rede final foram identificadas seis comunidades representativas, sendo elas: artistas famosos mais frequentes, artistas famosos menos frequentes, artistas pouco conhecidos, artistas do sub-gênero Gangsta Rap, artistas de outros gêneros e artistas do selo musical Young Money Cash Money Billionares.

O coeficiente de modularidade da rede foi de 0,502, o que indica que as conexões entre nós de uma mesma comunidade é mais densa, enquanto é mais esparsa em relação a nós de diferentes grupos. Logo, pôde-se dizer que, no Rap, os artistas pouco conhecidos não tendem a colaborar com artistas famosos.

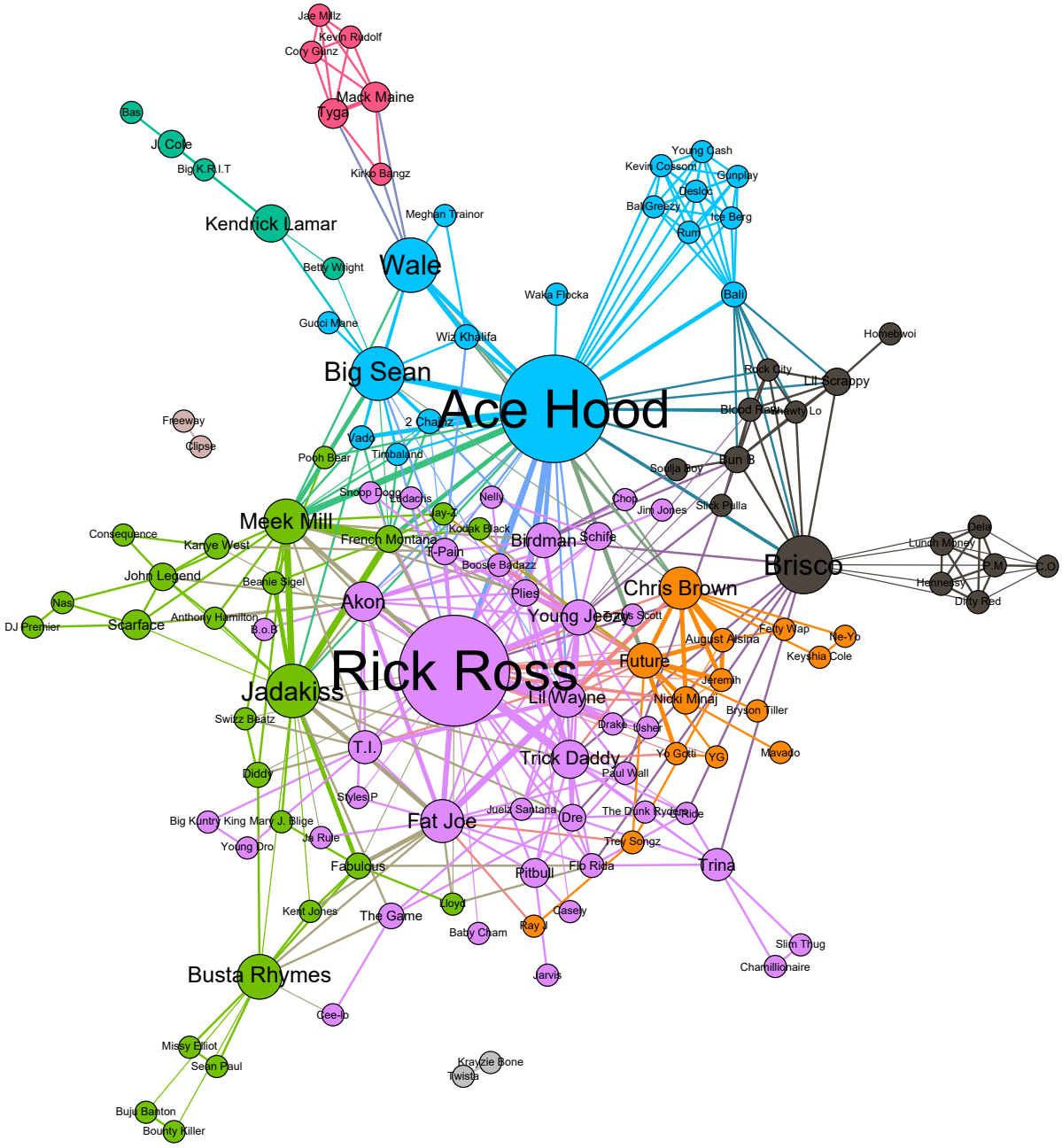


Figura 17 – Rede englobando todos os artistas colaboradores dos nove primeiros discos do DJ Khaled. Extraído de (ARAUJO et al., 2017b).

A.3 *Identification of Most Popular Musical Genres and their Influence Factors*

No terceiro artigo da parte de identificação de fatores de influência (ARAUJO; NAKAMURA, 2018) o foco foi em estudar o que impacta na popularidade de gêneros musicais.

Nesse trabalho analisamos os gêneros musicais das 100 músicas mais populares dos anos de 1959 a 2017 segundo ranqueamentos da revista americana Billboard[5] para identificar os fatores de influência. Como no site da revista não há a informação dos *rankings* de todos esses anos, recorremos a um *crawler* de informações da Wikipedia[6] para nossa coleta de dados. Essas informações foram validadas em dois sites que também mantém tais registros, são eles: o "Billboard Top 100 of"[7] e o site do Bob Borst[8].

Para cada década (1959 ficou junto a década de 60) montamos representações gráficas contendo a quantidade de músicas de cada gênero musical por ano. Quando observamos mudanças abruptas nessa quantidade, então buscávamos informações sobre marcos históricos no mundo da música que ocorreram naquele ano. Dessa forma, obtivemos os fatores de influência na popularidade de gêneros musicais.

A Figura 18 apresenta a quantidade de músicas de cada gênero musical nos anos 70. É interessante observar que mesmo contendo a maior quantidade de músicas em toda a década o Rock só conta com duas músicas na primeira posição, conforme apresentado na Tabela 13.

Mesmo obtendo destaque na quantidade total de músicas somente nas décadas de 1960 e 2010, o Pop apresentou a maior quantia de músicas na primeira posição em todas as décadas. A década de 1980 é a única exceção. Nesse período o Rock chega a ter mais da metade das músicas nas listas, conforme a Figura 19. Nessa década, seis músicas de Rock foram a mais popular do ano, enquanto o Pop teve três músicas. Os gráficos das décadas de 1960, 1990, 2000 e 2010 estão no Apêndice D.

[5] <https://www.billboard.com/charts/year-end/2017/hot-100-songs>
[6] <https://en.wikipedia.org/wiki/Billboard_Year-End>
[7] <http://billboardtop100of.com/>
[8] <http://www.bobborst.com/>

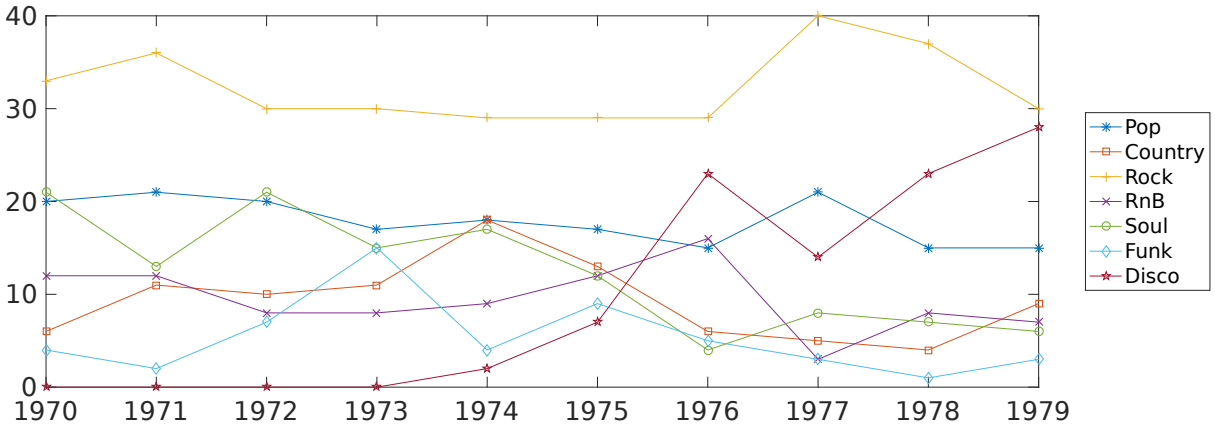


Figura 18 – Número de músicas de cada gênero por ano na década de 70. Extraído de (ARAUJO; NAKAMURA, 2018).

Tabela 13 – Dados referentes as músicas mais populares da década de 1970.

Anos 70	Soma Total	Qnt. Nº 1
Pop	179	5
Rock	323	2
Soul	124	1
Disco	97	2

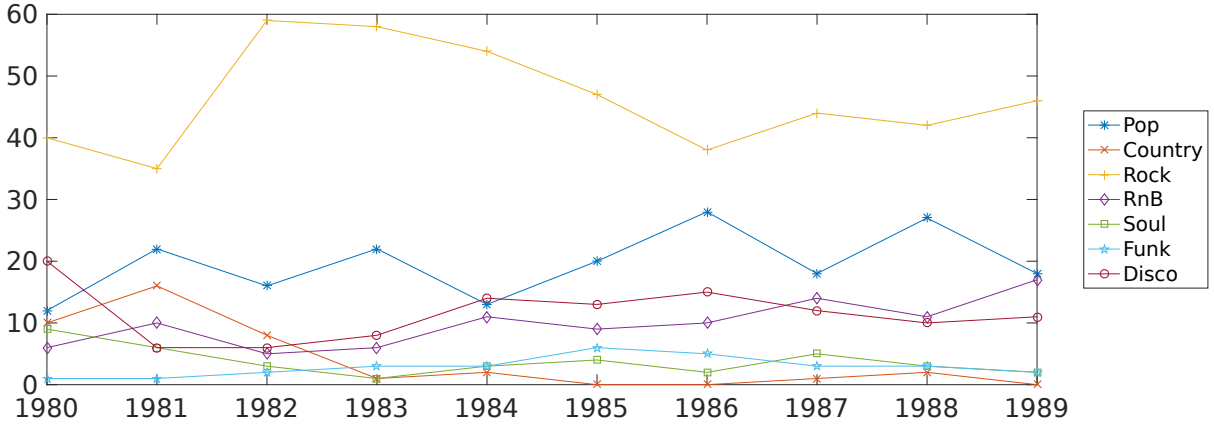


Figura 19 – Número de músicas de cada gênero por ano na década de 80. Extraído de (ARAUJO; NAKAMURA, 2018).

De modo geral, o principal fator de influência na popularidade de um gênero musical é o surgimento de um novo artista ou de um movimento de destaque nesse gênero. Como ocorreu no Rock nas décadas de 60 e 80, com a invasão britânica capitaneada pelos Beatles e do movimento New Wave, respectivamente. E, também, no RnB com Mariah Carey nos anos 90 e no Rap com o Trap, a partir de 2015.

Entretanto, não basta apenas um artista para mudar o panorama da música por um longo período. Por exemplo, o Country obteve bons números

após as primeiras músicas de sucesso das cantoras Faith Hill e Taylor Swift, mas não conseguiu manter-se em destaque pela falta de outros grandes nomes a surgir. Assim como com a banda Bee Gees que iniciou uma popularização da música Disco no final dos anos 70, mas que não conseguiu se manter pelos mesmos motivos. Enquanto nos anos 90, após a primeira aparição de Mariah Carey, bandas de destaque no RnB como Boyz II Men e TLC também surgiram. Observou-se que o período máximo que um ato musical consegue alavancar sozinho um gênero é de cinco anos.

Outros fatores identificados, mas com menor influência são: a migração de artistas já conhecidos para outro gênero, como ocorreu com o Pop nos anos 2010, fazendo aumentar a sua popularidade, e situações trágicas, como ocorrido em 1996 e 1997, com os assassinatos de 2Pac e Notorius B.I.G., afetando a ascensão do Rap.

APÊNDICE B – Dados utilizados em Araujo et al. (2017a)

Tabela 14 – Dados obtidos após coleta e análise de sentimento (ordem decrescente de Popularidade no Spotify). Extraído de (ARAUJO et al., 2017a).

Artista	Disco	# Tweets Negativos	# Tweets Positivos	# Total de Tweets	Popularidade Spotify	Unidades Billboard
Chance The Rapper	Coloring Book	13.469	42.749	87.214	86	38.000
Lukas Graham	Lukas Graham	5.758	26.580	58.547	76	59.000
The XX	I See You	2.175	22.995	24.143	76	46.000
Rick Ross	Rather You Than Me	9.672	8.109	27.986	75	106.000
Childish Gambino	Awaken, My Love!	29.970	15.878	54.759	71	101.000
Macklemore	This Unruly Mess I've Made	17.621	34.331	73.564	70	61.000
The Chainsmokers	Collage	6.883	20.447	59.629	69	39.000
Radiohead	A Moon Shaped Pool	34.262	30.785	95.458	67	173.000
Dj Khaled	Major Key	10.188	22.844	57.500	67	96.000
The Lumineers	Cleopatra	1.974	8.217	16.865	67	125.000
Metallica	Hardwired... to Self-Destruct	15.971	16.381	49.469	66	291.000
Deadmau5	W:/2016ALBUM/	4.882	7.431	18.446	62	-
Incubus	8	4.366	7.106	18.453	62	52.000
Tom Odell	Wrong Crowd	2.354	11.321	16.408	61	-
Lindsey Stirling	Brave Enough	1.665	11.937	17.644	59	49.000

APÊNDICE C – Redes criadas em Araujo et al. (2017b)

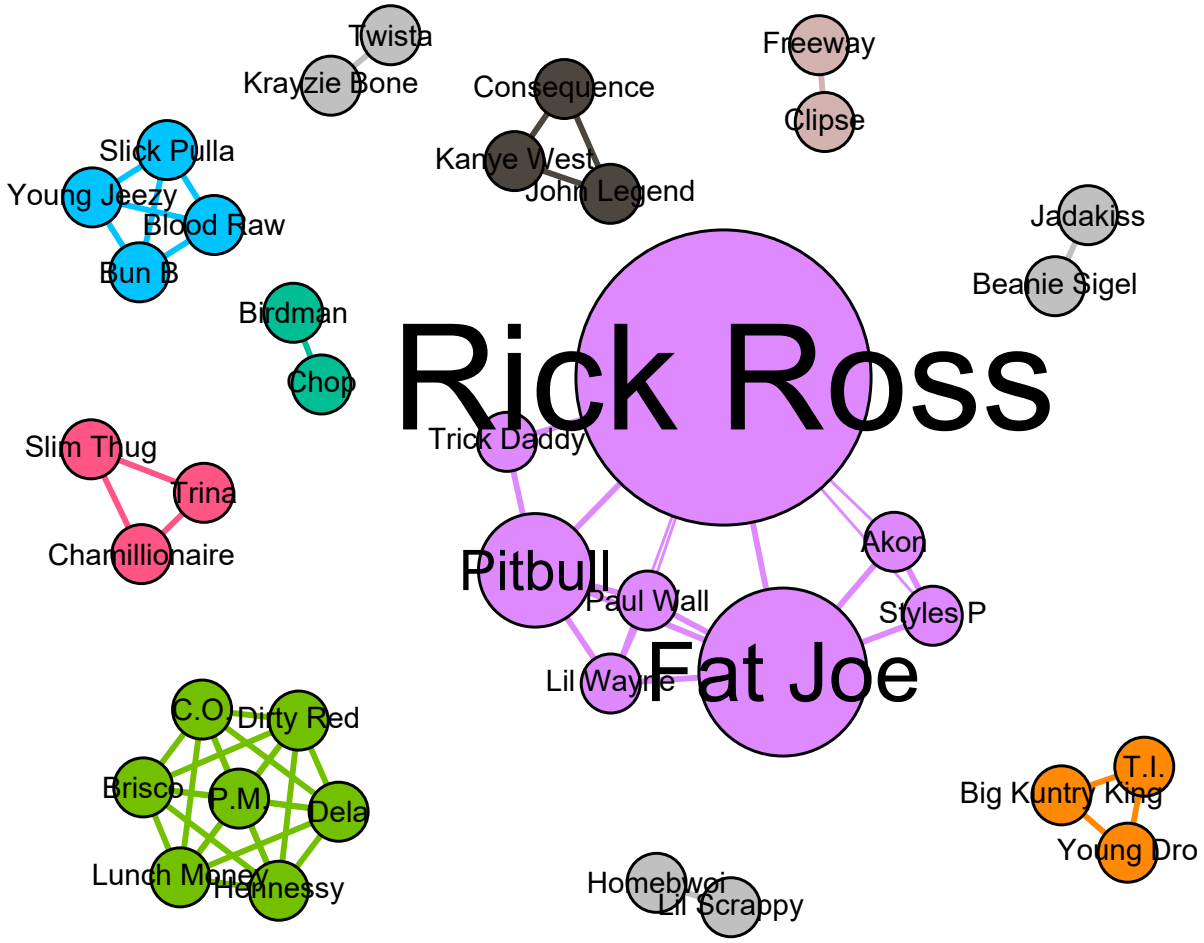


Figura 20 – Rede englobando todos os artistas colaboradores primeiro disco do DJ Khaled.

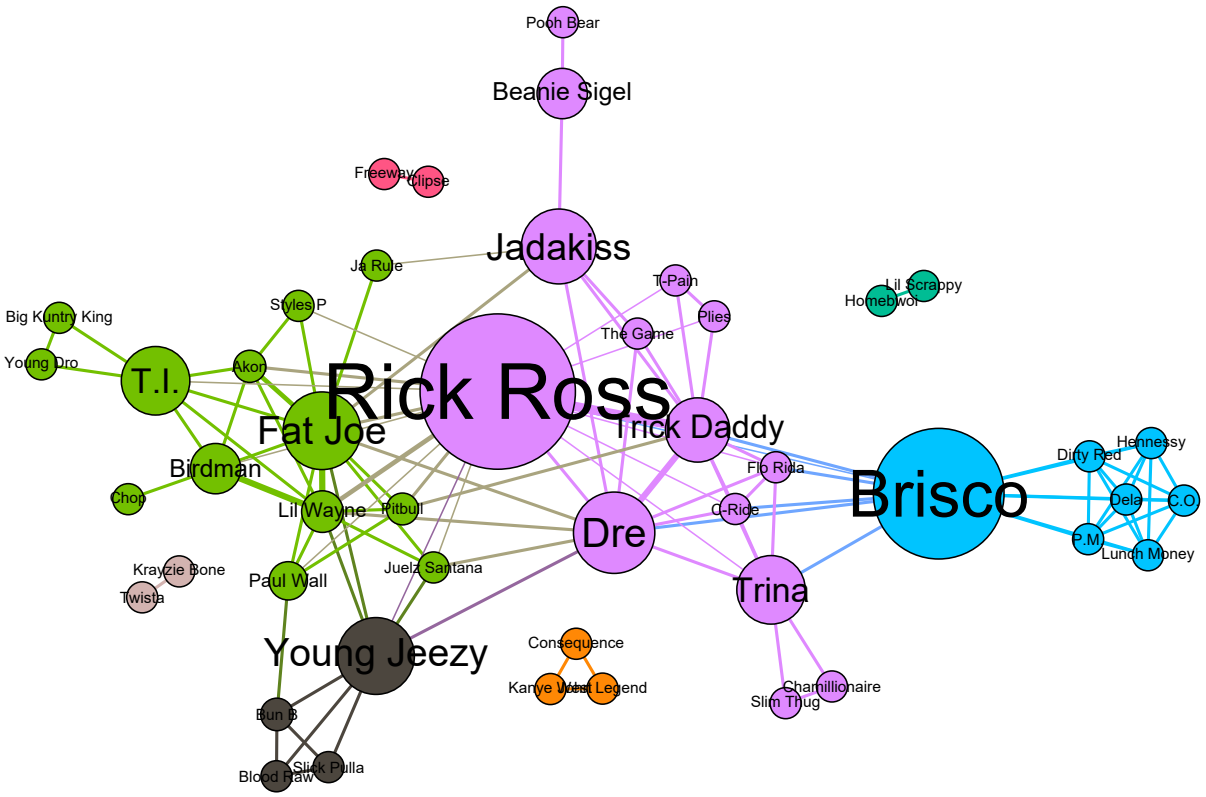


Figura 21 – Rede englobando todos os artistas colaboradores dos dois primeiros discos do DJ Khaled.

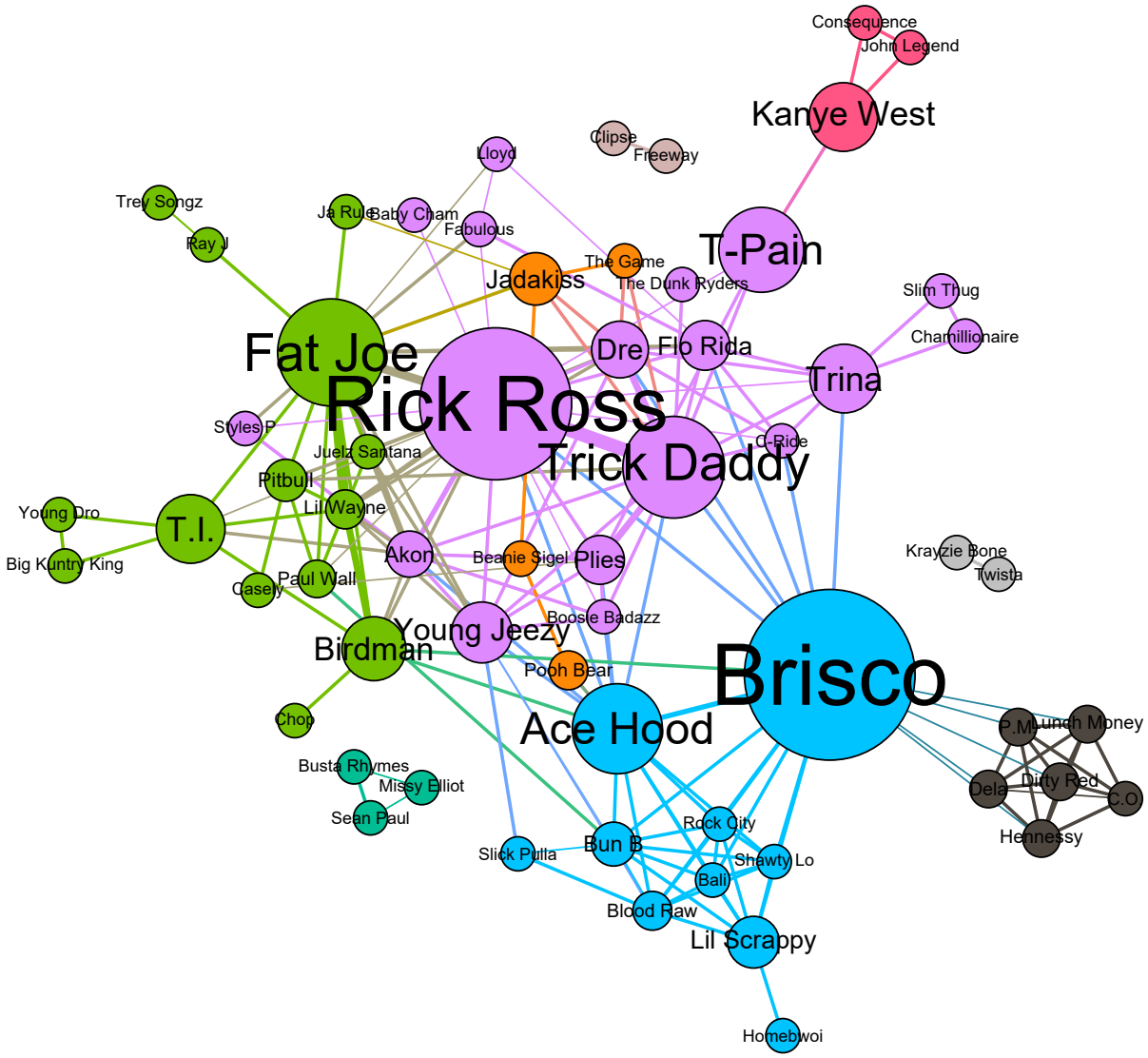


Figura 22 – Rede englobando todos os artistas colaboradores dos três primeiros discos do DJ Khaled.

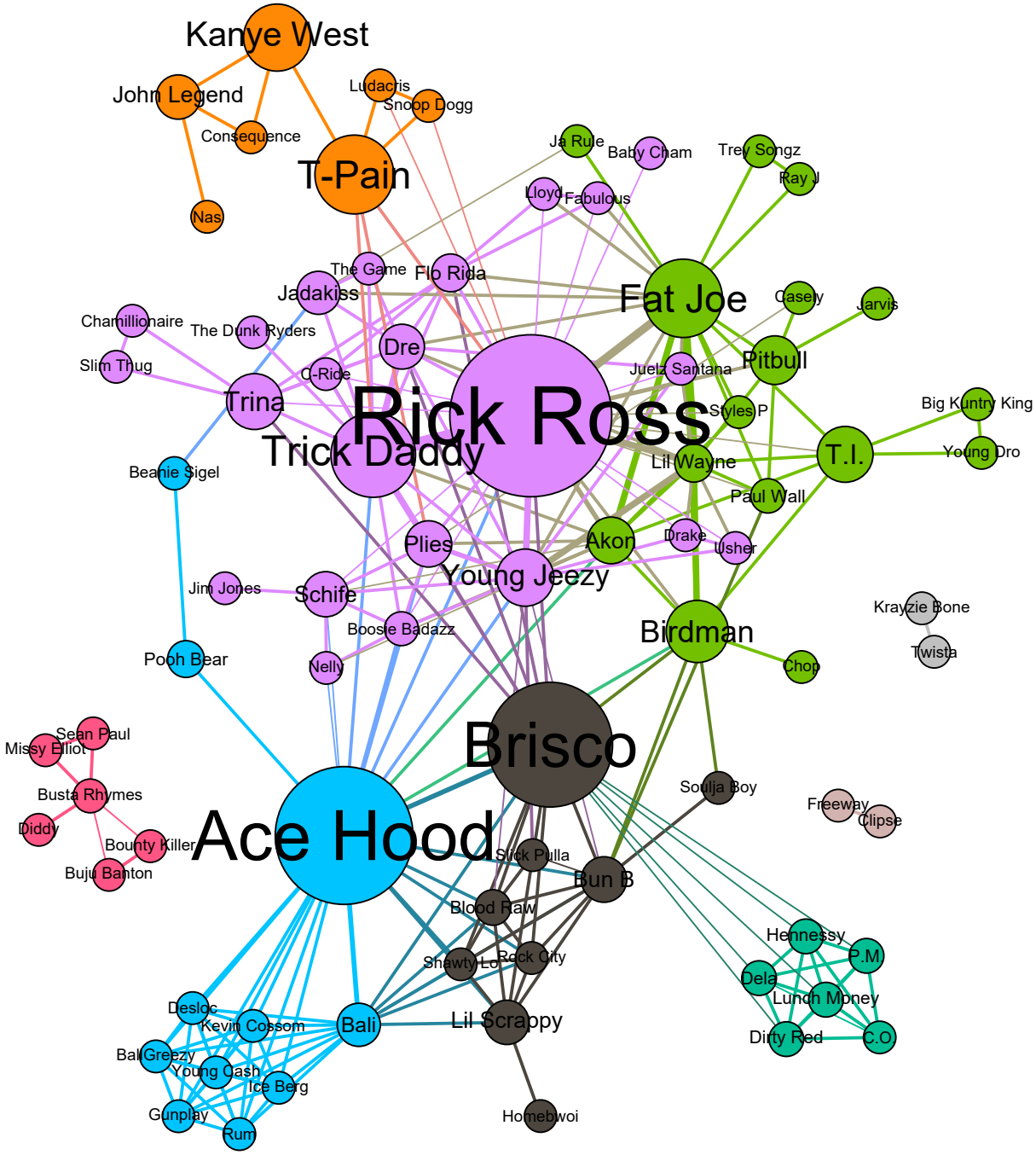


Figura 23 – Rede englobando todos os artistas colaboradores dos quatro primeiros discos do DJ Khaled.

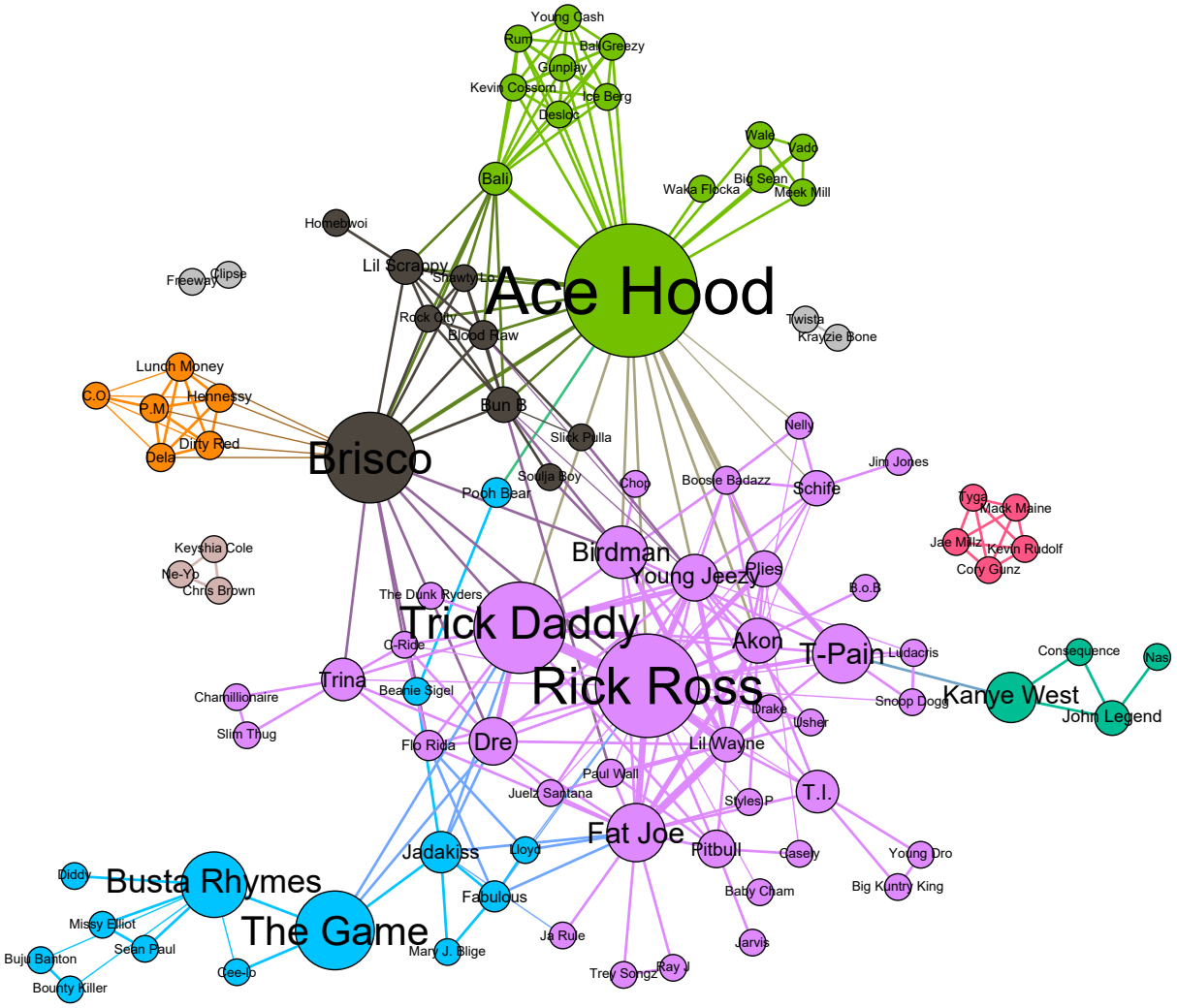


Figura 24 – Rede englobando todos os artistas colaboradores dos cinco primeiros discos do DJ Khaled.

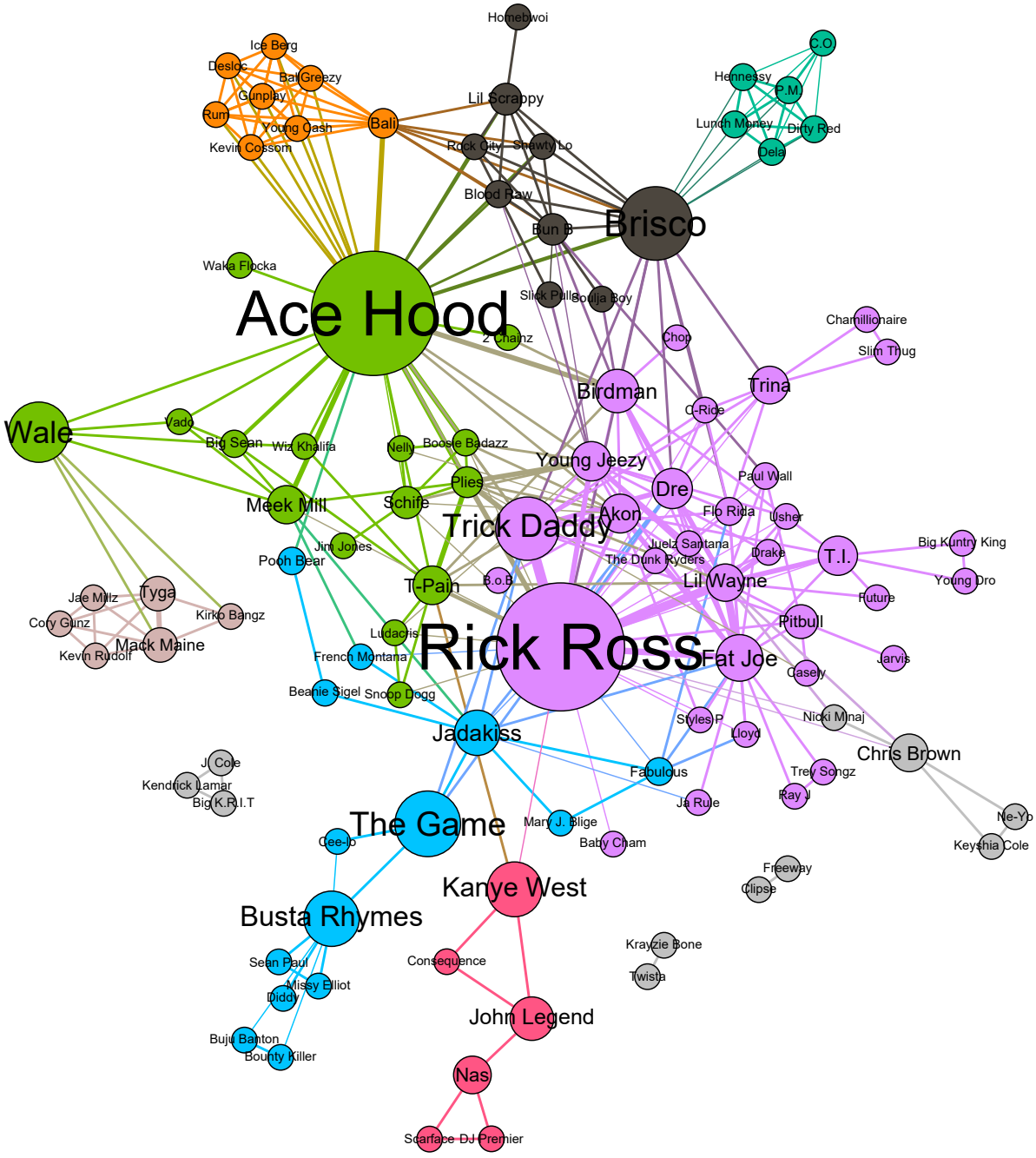


Figura 25 – Rede englobando todos os artistas colaboradores dos seis primeiros discos do DJ Khaled.

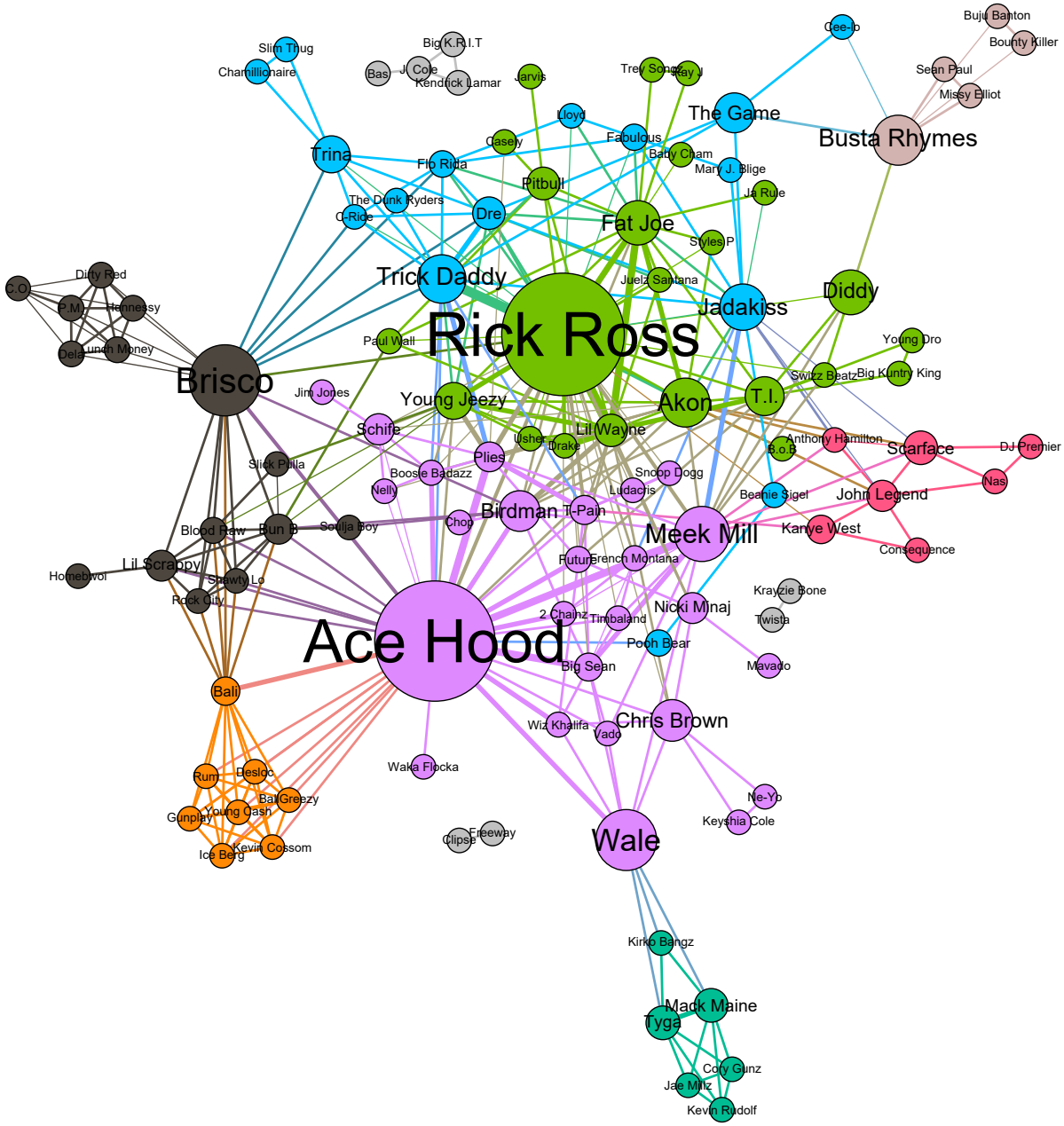


Figura 26 – Rede englobando todos os artistas colaboradores dos sete primeiros discos do DJ Khaled.

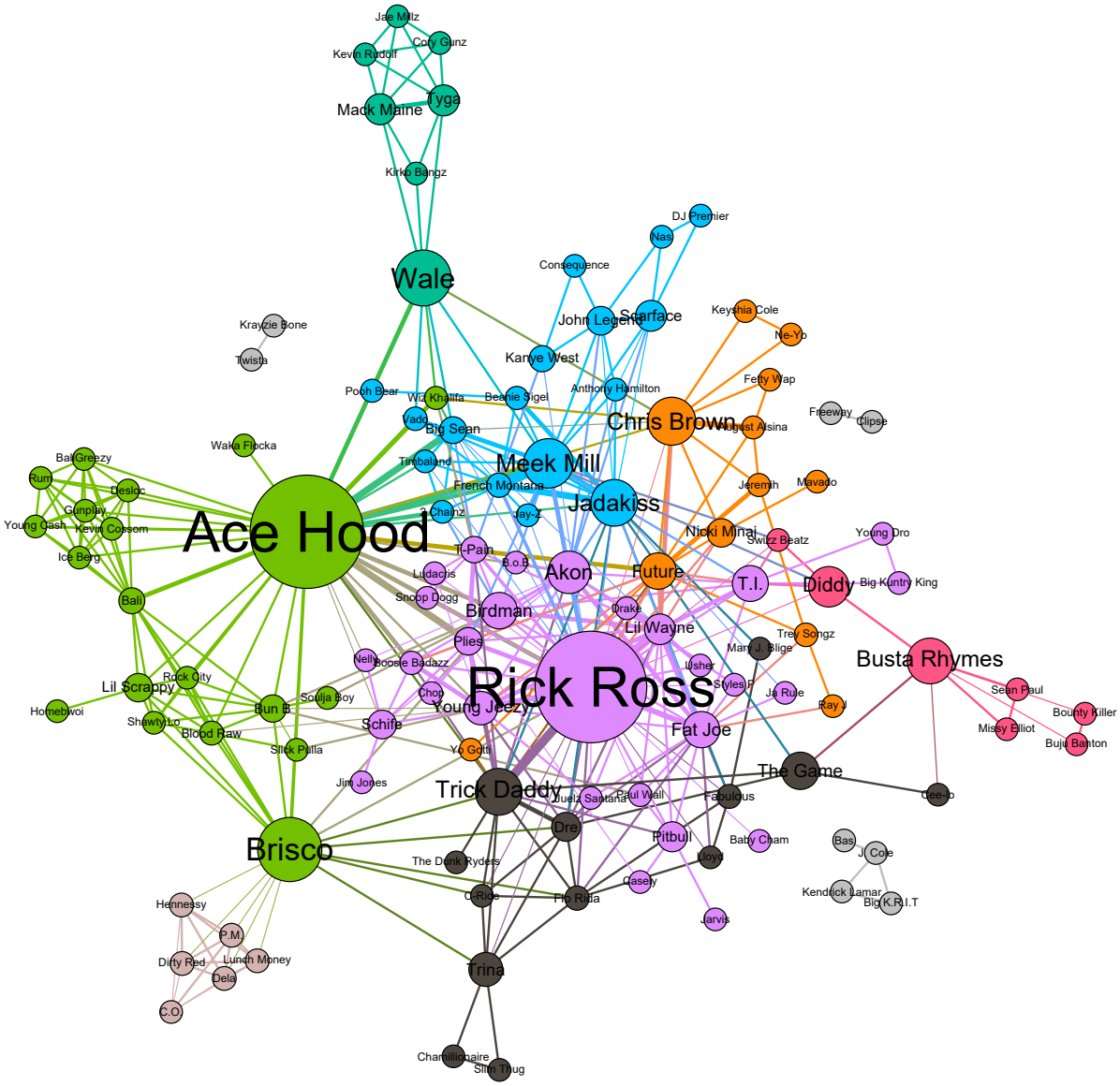


Figura 27 – Rede englobando todos os artistas colaboradores dos oito primeiros discos do DJ Khaled.

APÊNDICE D – Gráficos obtidos em Araujo e Nakamura (2018)

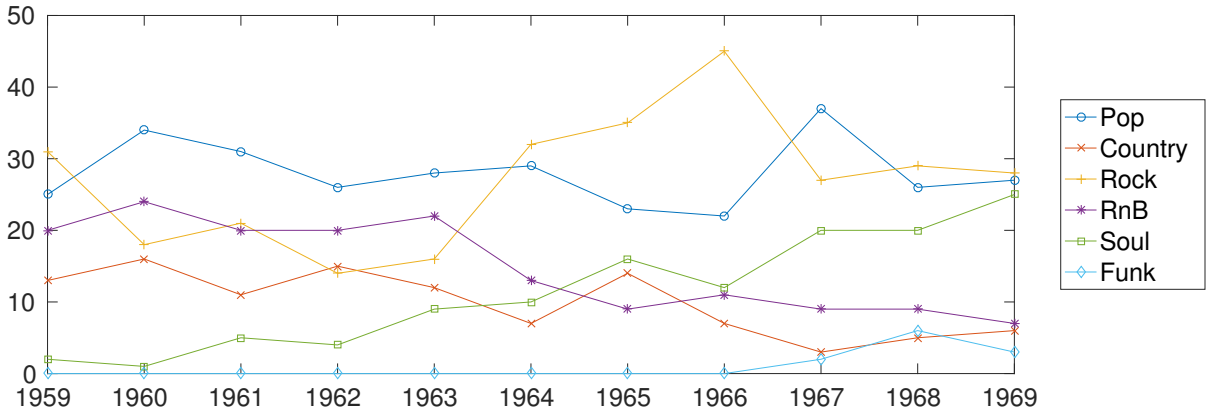


Figura 28 – Número de músicas de cada gênero por ano na década de 60. Extraído de (ARAUJO; NAKAMURA, 2018).

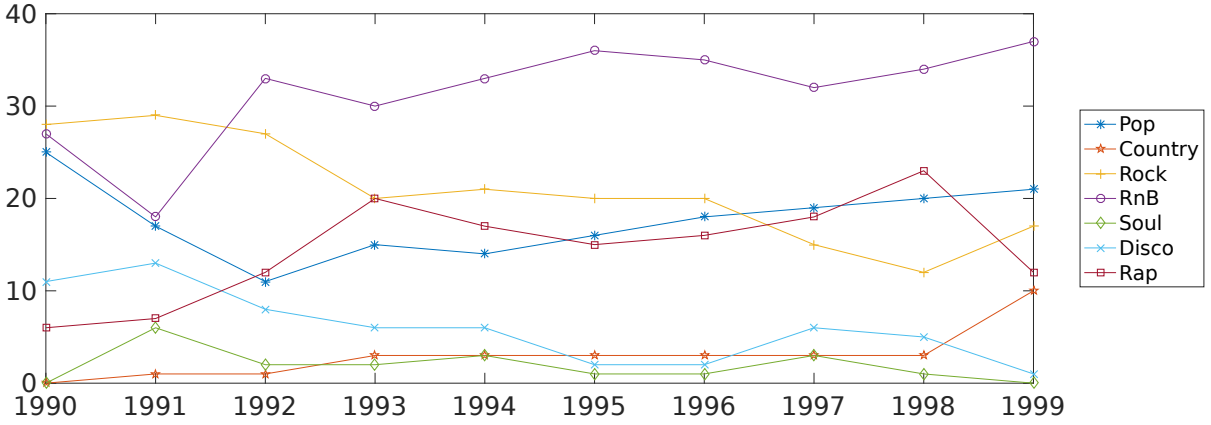


Figura 29 – Número de músicas de cada gênero por ano na década de 90. Extraído de (ARAUJO; NAKAMURA, 2018).

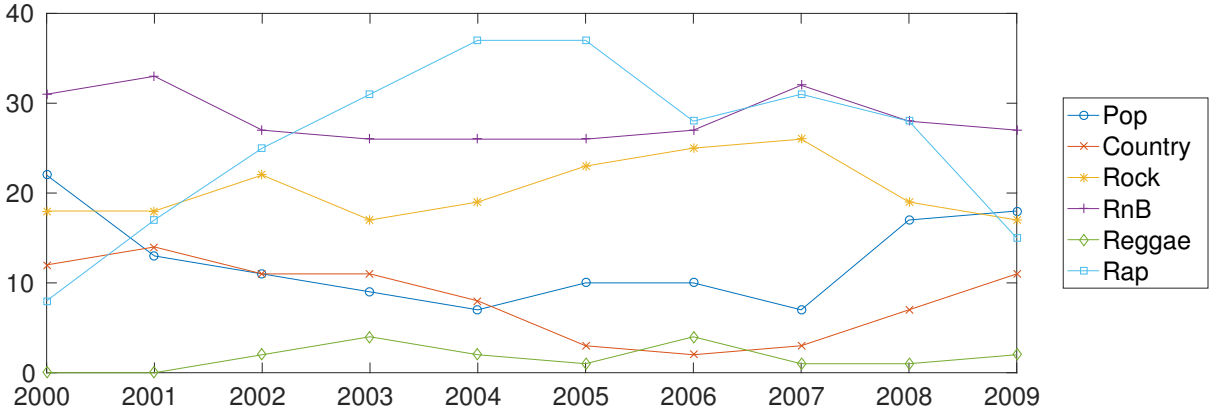


Figura 30 – Número de músicas de cada gênero por ano na década de 2000. Extraído de (ARAUJO; NAKAMURA, 2018).

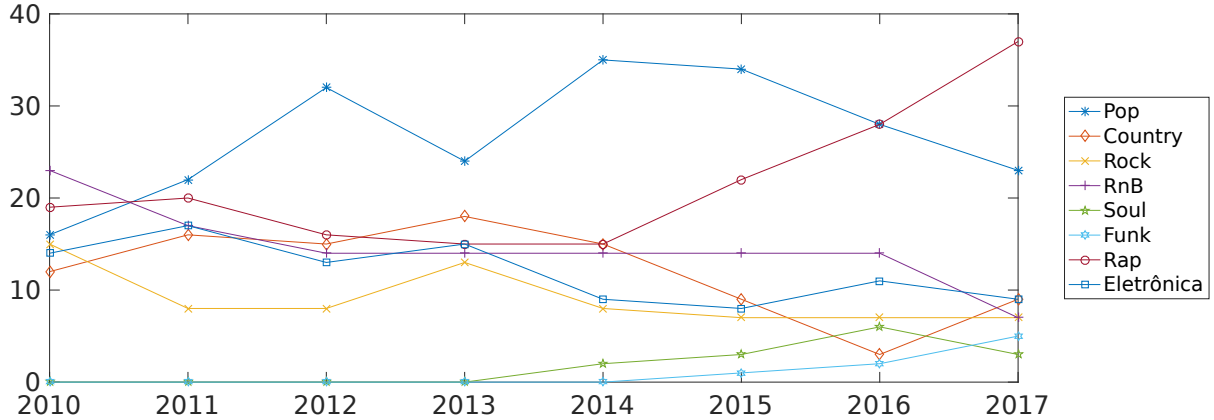


Figura 31 – Número de músicas de cada gênero por ano na década de 2010. Extraído de (ARAUJO; NAKAMURA, 2018).