



Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Faculdade de Engenharia

Gabriela Siqueira Eduardo

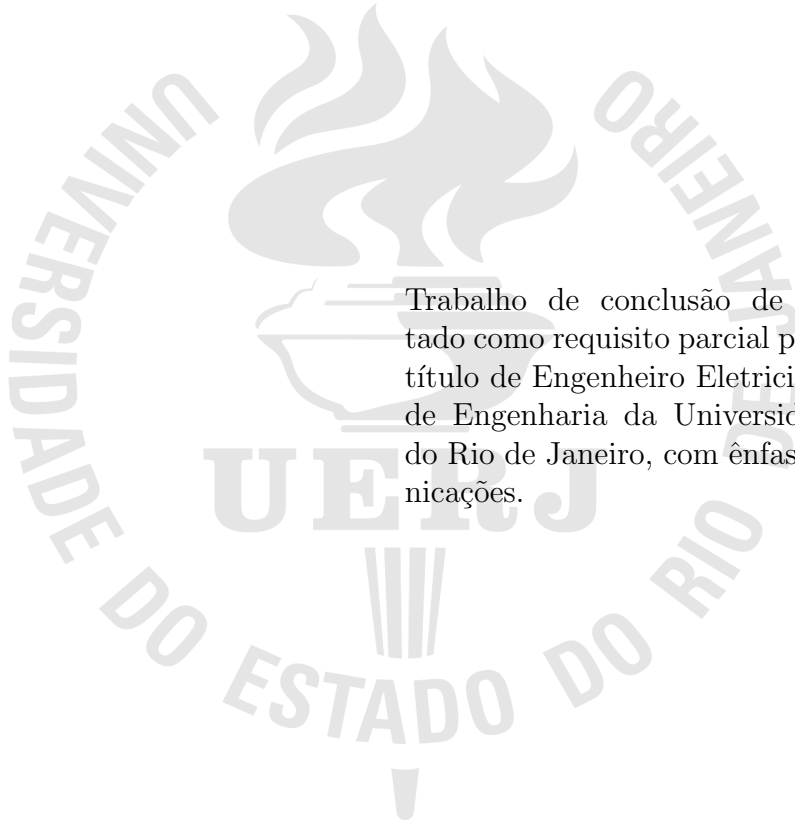
**Classificação de instrumentos musicais baseada em aprendizado
de máquina**

Rio de Janeiro

2022

Gabriela Siqueira Eduardo

Classificação de instrumentos musicais baseada em aprendizado de máquina



Trabalho de conclusão de curso apresentado como requisito parcial para obtenção do título de Engenheiro Eletricista à Faculdade de Engenharia da Universidade do Estado do Rio de Janeiro, com ênfase em Telecomunicações.

Orientador: Prof. Dr. Michel Pompeu Tcheou

Rio de Janeiro

2022

CATALOGAÇÃO NA FONTE

S237

UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

Sobrenome, Nome do Autor

Título / Nome completo do autor. – 2012.

105 f.

Orientadores: Nome completo do orientador1;

Nome completo do orientador2

Dissertação(Mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia.

Texto a ser informado pela biblioteca.

CDU 621:528.8

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta dissertação, desde que citada a fonte.

Assinatura

Data

Gabriela Siqueira Eduardo

Classificação de instrumentos musicais baseada em aprendizado de máquina

Trabalho de conclusão de curso apresentado como requisito parcial para obtenção do título de Engenheiro Eletricista à Faculdade de Engenharia da Universidade do Estado do Rio de Janeiro, com ênfase em Telecomunicações.

Aprovado em: x de x de 2022

Banca Examinadora:

Prof. Dr. Nome do Professor 1 (Orientador)

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Prof. Dr. Nome do Professor 2

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Prof. Dr. Nome do Professor 3

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Prof. Dr. Nome do Professor 4

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Prof. Dr. Nome do Professor 5

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Rio de Janeiro

2022

AGRADECIMENTO

Aqui entra seu agradecimento.

RESUMO

EDUARDO, Gabriela Siqueira. *Classificação de instrumentos musicais baseada em aprendizado de máquina*. 2022. 49 f. Trabalho de Conclusão de Curso (Graduação em Engenharia Elétrica) - Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, 2022.

Classificação de instrumentos musicais em gravações polifônicas, utilizando algoritmos de aprendizado de máquina supervisionado. Extração de informações temporais e espectrais de sinais de áudio disponibilizados pela base de dados do IRMAS. Apresentação de três métodos de classificação de múltiplas classes, sendo eles: Máquinas de Vetores de Suporte, Floresta Aleatória e Redes Neurais Artificiais. Sugestões de propostas para aprimoramento posterior de classificadores.

Palavras-chave: Reconhecimento de instrumentos musicais, Aprendizado de máquina, Multiclasse, RNA, SVM, FA, Espectrograma Mel, Extração de preditores.

ABSTRACT

EDUARDO, Gabriela Siqueira. *Musical instruments classification based on machine learning*. 2022. 49 f. Trabalho de Conclusão de Curso (Graduação em Engenharia Elétrica) - Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, 2022.

Musical instruments classification in polyphonic recordings, using supervised machine learning algorithms. Temporal and spectral feature extraction from audio signals provided by the IRMAS dataset. Presentation of three multiclass classification methods, them being: Support Vector Machine, Random Forest and Artificial Neural Networks. Proposal suggestions for classifiers' further improvements.

Keywords: Musical Information Recognition, Machine Learning, ANN, SVC, RF, Mel Spectrogram, Feature Extraction.

LISTA DE FIGURAS

Figura 1 - Representação no domínio do tempo.	15
Figura 2 - Representação no domínio do tempo e da frequência.	16
Figura 3 - Guitarra [1].	17
Figura 4 - Violino.	18
Figura 5 - Piano.	19
Figura 6 - Flauta transversal [2].	20
Figura 7 - Trompete [2].	20
Figura 8 - Saxofone [2].	21
Figura 9 - Esquematização do SVM [3]	24
Figura 10- Esquematização da floresta aleatória.	25
Figura 11- Funções de ativação [4]	26
Figura 12- Esquematização da rede neural	27
Figura 13- Proposta de projeto	28
Figura 14- Exemplo de nome de arquivo da base de dados	29
Figura 15- Distribuição do RMS	30
Figura 16- Distribuição do SC	31
Figura 17- Distribuição da SB	32
Figura 18- Distribuição da frequência de <i>rolloff</i>	32
Figura 19- Distribuição do ZCR	33
Figura 20- Coeficientes Cepstrais de Mel	34
Figura 21 - ANN projetada.	37
Figura 22- Classificações SVC.	38
Figura 23- Influência dos <i>top</i> 10 preditores do SVC.	39
Figura 24- Classificações RF.	40
Figura 25- Influência dos <i>top</i> 10 preditores da RF.	41
Figura 26- Classificações ANN.	42
Figura 27- Acurácia da ANN x épocas.	43
Figura 28- Influência dos <i>top</i> 10 preditores da ANN.	44

LISTA DE TABELAS

Tabela 1 - Matriz de confusão.....	23
Tabela 2 - Quantidade de amostras para cada instrumento	29
Tabela 3 - Quantidade de amostras para cada instrumento para base de treino e de teste.	34
Tabela 4 - Representação numérica e vetorial das classes.	35
Tabela 5 - Métricas resultantes do SVC.	38
Tabela 6 - Métricas resultantes da RF.....	40
Tabela 7 - Métricas resultantes da ANN.	42
Tabela 8 - Acurácia de cada modelo projetado.	45
Tabela 9 - Resumo das métricas de cada instrumento para cada classificador.....	45

LISTA DE SIGLAS

AD	Árvore de decisão
ADSR	Attack Decay Sustain Release
ANN	Artificial Neural Network
FFT	Fast Fourier Transform
flu	Flauta
FN	Falso Negativo
FP	Falso Positivo
gel	Guitarra
IRMAS	Instrument Recognition in Musical Audio Signals
ISMIR	International Society for Music Information Retrieval
MFCC	Mel Frequency Cepstral Coefficients
ML	Machine Learning
pia	Piano
rbf	Radial Basis Function
ReLU	Rectified Linear Unit
RF	Random Forest
RMS	Root Mean Square
sax	Saxofone
SB	Spectral Bandwidth
SC	Spectral Centroid
SHAP	SHapley Additive exPlanations
SVC	Support Vector Classifier
SVM	Support Vector Machine
tru	Trompete
vio	Violino
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
wav	Waveform Audio File Format
ZCR	Zero Crossing Rate

LISTA DE SÍMBOLOS

f	Frequência
N	Quantidade de classes
Φ	Função de Ativação
W_n	Peso
x_n	Entrada
b_n	Viés
y_n	Saída
N	Quantidade de amostras em um áudio

SUMÁRIO

	INTRODUÇÃO	12
1	A FÍSICA DA MÚSICA	14
1.1	Onda sonora	14
1.2	Instrumentos	16
1.2.1	<u>Guitarra elétrica</u>	17
1.2.2	<u>Violino</u>	17
1.2.3	<u>Piano</u>	18
1.2.4	<u>Flauta</u>	19
1.2.5	<u>Trompete</u>	20
1.2.6	<u>Saxofone</u>	20
2	APRENDIZADO DE MÁQUINA	22
2.1	Fundamentação teórica do aprendizado de máquina	22
2.2	Aprendizado supervisionado	23
2.2.1	<u>Máquinas de vetores de suporte</u>	24
2.2.2	<u>Floresta Aleatória</u>	25
2.2.3	<u>Redes Neurais Artificiais</u>	26
3	PROJETO	28
3.1	Base de dados	29
3.2	Extração de informações	30
3.2.1	<u>Valor quadrático médio</u>	30
3.2.2	<u>Centróide espectral</u>	31
3.2.3	<u>Largura de banda espectral</u>	31
3.2.4	<u>Frequência de Rolloff</u>	32
3.2.5	<u>Zero Crossing Rate</u>	33
3.2.6	<u>Coefficientes Cepstrais de Frequência Mel</u>	33
3.3	Preparação da base	34
3.4	Classificadores	35
3.4.1	<u>Projeto do SVM</u>	36

3.4.2	<u>Projeto da RF</u>	36
3.4.3	<u>Projeto da ANN</u>	36
4	RESULTADOS	38
4.1	Resultado do SVC projetado	38
4.2	Resultado da RF projetada	40
4.3	Resultado da ANN projetada	42
	CONCLUSÃO	45
	REFERÊNCIAS	47

INTRODUÇÃO

Já na Pré-História, a música era um elemento fundamental da cultura humana. Desde os primórdios, os homens produziam diversas formas de sonoridade com variados objetivos, entre os quais para celebrar a caça, para realizar rituais de agradecimento, para aplacar a fúria ou para fazer pedidos aos deuses.

A música é um tipo de arte que trabalha com a harmonia entre os sons, o ritmo, a melodia, a voz. Todos esses elementos são importantes e podem transportar as pessoas para outro tempo e espaço, resgatar memórias e reacender emoções.

Com o decorrer dos anos, houve um grande aumento na disponibilidade de músicas, sobretudo com o advento dos canais digitais, tornando bem mais fácil o acesso a elas, levando ao consequente crescimento do número de ouvintes.

Os grandes responsáveis pela facilidade de distribuição e de consumo de músicas da atualidade são as plataformas de *streaming*, como, por exemplo, *Spotify*, *Apple Music*, *Deezer*, entre muitas outras opções.

A compreensão do timbre de instrumentos musicais é uma questão importante para a transcrição automática de música e recuperação de informações musicais. Essas plataformas podem utilizá-las em sistemas de classificação para a categorização do catálogo de músicas, bem como em sistemas de recomendação, a fim de aprimorar a experiência dos usuários, sugerindo estilos semelhantes ao ouvinte, a depender do seu gosto pessoal.

O presente trabalho tem como o objetivo principal classificar instrumentos musicais presentes em composições de estilos variados, utilizando-se de algoritmos computacionais de aprendizado de máquina supervisionado.

Para tal, serão estudadas algumas das características espectrais de um sinal de áudio - como largura de banda de frequência, centróide espectral, coeficientes cepstrais de frequência-Mel, entre outras, extraídas através de algoritmos próprios -, para posterior aplicação em modelos.

Além disso, buscar-se-á explorar o universo dos algoritmos de aprendizado de máquina escolhidos - *Support Vector Machine*, *Random Forest* e Redes Neurais -, abordando o seu funcionamento e os seus parâmetros.

O que motivou à realização deste estudo foi o fato de existir uma grande dificuldade no reconhecimento de cada um dos múltiplos instrumentos que compõem uma canção,

devido à sobreposição de tempo e de frequência, à variação de timbres e à falta de dados classificados. A isso, soma-se o fato de que, na realidade, as componentes espectrais de um mesmo instrumento não se mantêm constantes, mesmo que se esteja estudando uma mesma nota - o que eleva o grau de dificuldade no seu reconhecimento.

Ainda, a classificação de áudio de instrumentos, de gêneros, de notas, entre outros, faz-se interessante na automatização de consultas de peças musicais, de criação de catálogo, de transcrição de músicas, bem como na criação de sistemas de recomendação.

Após este capítulo introdutório, reservou-se o seguinte para a fundamentação teórica dos sinais de áudio utilizados.

Já o Capítulo 2 apresenta os pressupostos teóricos que norteiam o aprendizado de máquina, com enfoque nos algoritmos utilizados neste projeto.

Em seguida, no Capítulo 3, foi apresentada a metodologia utilizada, abrangendo desde a obtenção dos dados até a aplicação dos modelos de aprendizado de máquina utilizados.

O Capítulo 4, por sua vez, dedica-se à exposição dos resultados obtidos.

Em seguida, apresenta-se a conclusão geral do trabalho, além do fornecimento de propostas para um posterior aprimoramento do classificador projetado.

Por fim, seguem-se as referências bibliográficas sobre o assunto.

1 A FÍSICA DA MÚSICA

No presente capítulo, serão apresentados alguns fundamentos teóricos dos sinais de áudio e as características dos instrumentos escolhidos para estudo.

1.1 Onda sonora

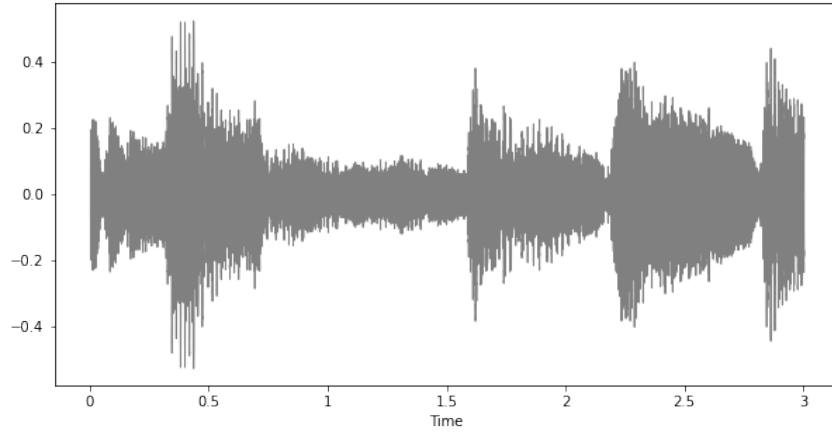
Áudio é um sinal correspondente aos sons, e música é a arte de combinar os sons.

Chamamos de sinal de áudio uma série de compressões e rarefações alternadas do ar, causadas pelas vibrações das moléculas do meio, que se propagam em ondas sonoras, cujo efeito mecânico é captado pelo tímpano [5] [6].

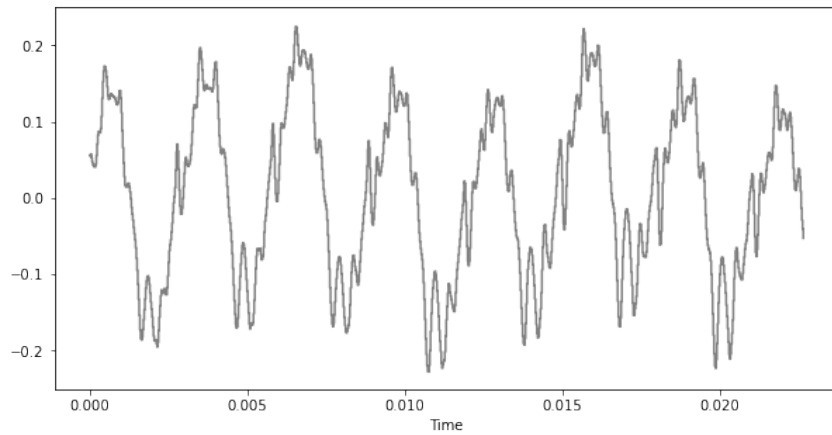
O contínuo aumento e diminuição dessa pressão formam uma onda com forma senoidal, as chamadas ondas sonoras. A proporção da mudança de pressão do ar indica a amplitude, que nada mais é do que a intensidade sonora - quantidade de energia emitida por uma fonte. Já a velocidade em que o sinal se repete - ciclo vibratório completo - indica a frequência da onda [7]. A onda sonora também dispõe de uma propriedade chamada timbre, que diferencia sons diferentes possuidores de uma mesma frequência e amplitude.

É interessante observar que os limiares de frequência da audibilidade humana são de 20Hz até aproximadamente 20kHz. Os sinais que possuem frequências fora dessa faixa, chamados de infrassom e ultrassom, respectivamente, não são possíveis de ser ouvidos. Outra observação a ser feita é que, como o som é uma forma de energia, ele não pode simplesmente deixar de existir; portanto a explicação do decaimento dos sons se dá pela absorção deles pelas superfícies dos objetos no espaço - que podem incluir móveis, pessoas e ar, transformando a energia em calor [8].

A Figura 1a apresenta graficamente a amplitude em relação ao tempo de uma amostra de sinal de áudio de 3 segundos de duração, e a Figura 1b exibe a mesma amostra, porém com uma duração de aproximadamente 0.025 segundos para a exemplificação da característica senoidal.



(a) Amostra de 3 segundos.



(b) Amostra de 0.025 segundos.

Figura 1: - Representação no domínio do tempo.

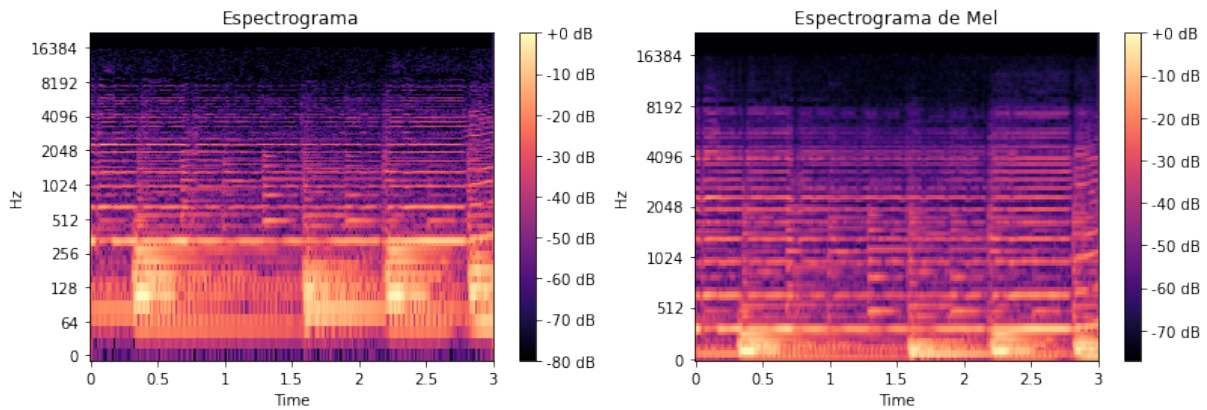
O espectrograma é uma representação visual do som, tanto no domínio da frequência, como no do tempo. Para criá-lo, é necessário converter as amostras no domínio do tempo individuais para o domínio da frequência, utilizando a Transformada Rápida de Fourier (FFT), configurada pela equação 1.1.

$$S_n = \sum_{k=0}^{N-1} s_k e^{-j \frac{2kn\pi}{N}}, n = 1, 2, \dots, N-1 \quad (1.1)$$

No lugar da FFT, também pode ser aplicado o escalonamento de frequência Mel, que é uma aproximação da percepção humana de sons. Ela apresenta uma melhor resolução em baixas frequências e uma pior em altas. A obtenção da representação da escala Mel a partir da frequência em Hertz pode ser dada pela equação 1.2 [9].

$$Mel(f) = \frac{1000}{\log 2} \log \left(1 + \frac{f}{700} \right) \quad (1.2)$$

Ambos os tipos de espectrogramas trazem a distribuição da frequência no tempo e, em uma terceira dimensão, indica a amplitude de determinada frequência em um certo instante de tempo. A Figura 2, mostra esses dois tipos de representação a partir da mesma amostra de áudio mostrada na Figura 1.



(a) Espectrograma.

(b) Espectrograma de Mel.

Figura 2: - Representação no domínio do tempo e da frequência.

1.2 Instrumentos

Como já apresentado na seção 1.1, uma forma de diferenciar um sinal com mesma amplitude e frequência é através do timbre.

Uma nota musical define-se apenas pela sua frequência fundamental. Quando uma nota é tocada em um instrumento real, uma série de frequências harmônicas também soam. A amplitude dos harmônicos determinam a qualidade do tom produzido, já que elas se diferenciam entre instrumentos distintos - o que representa a característica espectral do timbre [5] [10]. Alguns dos fatores responsáveis pela diferenciação do timbre em um mesmo instrumento são: material de construção (tipo de madeira, metal), material das cordas, espessura delas, entre outros.

Em termos temporais, é considerada a envoltória sonora, composta pelo Ataque, Decaimento, Sustentação e Relaxamento (ADSR) para diferenciação do timbre. Essas características representam a forma em que um som evolui no tempo [10]. Sendo:

- **Ataque:** como um som se inicia, tempo entre silêncio e intensidade total do mesmo.
- **Decaimento:** como um som se estabiliza, tempo até que a intensidade chegue ao valor desejado.

- **Sustentação:** duração do som, tempo em que a intensidade desejada se mantém.
- **Relaxamento:** como um som termina, tempo em que a intensidade diminui até desaparecer.

Nas subseções a seguir, será brevemente apresentado o que se espera ouvir dos instrumentos estudados neste projeto.

1.2.1 Guitarra elétrica

Apesar de muito parecida com o violão, a guitarra elétrica é um instrumento completamente diferente, principalmente pela sua forma de captação de som, que, em vez de ser através de uma caixa acústica, é feita por captadores eletromagnéticos.

As guitarras compõem-se principalmente por um corpo sólido de madeira, um braço -também de madeira- e por cordas. O som é captado através das vibrações das cordas quando tocadas, que causam uma mudança no fluxo magnético através da bobina do ímã permanente do captador, induzindo um sinal elétrico nela [6].

A frequência do sinal define-se pelo tamanho das cordas ao pressioná-las e pela tensão em cada corda. Como é um instrumento que depende de equipamentos eletrônicos, a intensidade do som e o timbre não são de total dependência da guitarra.



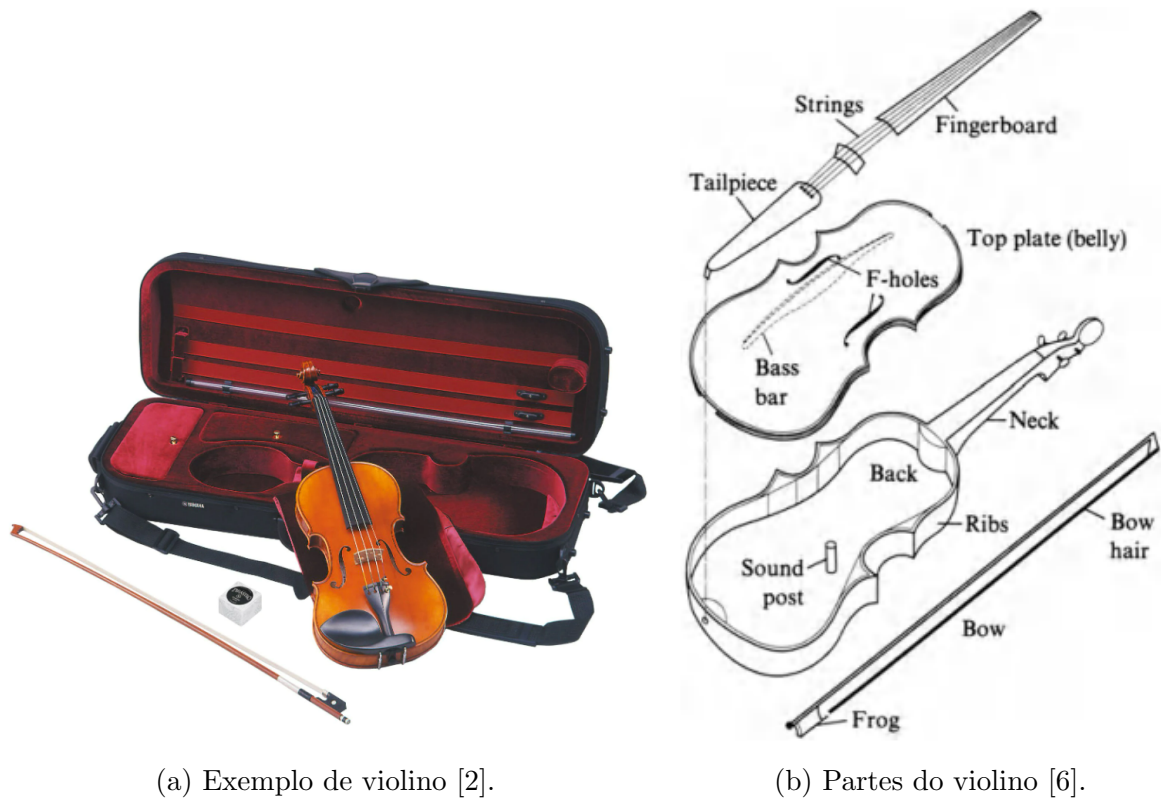
Figura 3: - Guitarra [1].

1.2.2 Violino

O violino acústico compõe-se especialmente por um corpo oco de madeira com aberturas, cordas e um arco de madeira e crina. O som é produzido através do atrito

(possibilitado pelo breu passado no arco) entre a crina do arco e as cordas, que resulta em uma vibração das cordas e amplificado pelo corpo.

Determina-se a intensidade do som principalmente pela velocidade e pela pressão do arco sobre as cordas. A frequência depende da tensão das cordas e do tamanho delas, que é alterado ao pressioná-las. O aspecto temporal também depende da forma como se manuseia o instrumento, como, por exemplo, a pressão aplicada no arco, a velocidade deste e a posição em que ele é mantido [6].



(a) Exemplo de violino [2].

(b) Partes do violino [6].

Figura 4: - Violino.

1.2.3 Piano

O piano compõe-se principalmente por teclas, martelos, cordas, pedais e uma caixa de ressonância. Nele, o som é produzido da seguinte forma: quando uma tecla é pressionada, ela ativa o martelo que toca nas cordas referentes à tecla pressionada e, então, o som é amplificado pela sua caixa de ressonância. Se o pedal de sustentação não estiver pressionado, haverá um amortecedor que impedirá a corda de vibrar quando o martelo não estiver sobre ele; se ele estiver pressionado, o som fluirá até a corda parar de vibrar naturalmente.

A característica temporal do sinal também depende do suprimento de ar, como o ataque, que depende da pressão do sopro, podendo ser abrupto, gradual e até plosivo [6].



Figura 6: - Flauta transversal [2].

1.2.5 Trompete

O trompete constitui-se principalmente pelo seu corpo metálico recurvado sobre si mesmo, um bocal e pistões. O som é produzido através da vibração labial junto com o sopro em seu bocal, que deve ter uma frequência próxima da nota desejada.

Assim como na flauta, a intensidade do som é estabelecida pela pressão do sopro. Já a frequência do som determina-se pelo tamanho do percurso - designado pela posição dos pistões - bem como pela frequência de vibração labial do trompetista [6].



Figura 7: - Trompete [2].

1.2.6 Saxofone

O saxofone constitui-se por um tubo metálico curvado, com buracos para os dedos - mecânica semelhante à da flauta - e por uma palheta de madeira na boquilha. O seu som é produzido a partir da vibração da palheta, resultante da coluna de ar gerada pelo sopro. Curiosamente, isso determina que o saxofone seja classificado como um instrumento da família das madeiras.

A intensidade do sinal é estabelecida pela força do sopro, e a frequência é determinada pela frequência de vibração da palheta e pelo tamanho do corpo - assim como na flauta. Por consequência de sua composição e formato, o saxofone produz um som com todos os harmônicos presentes [6].



Figura 8: - Saxofone [2].

2 APRENDIZADO DE MÁQUINA

Neste capítulo, será apresentada uma breve fundamentação teórica dos conceitos básicos de aprendizado de máquina e dos métodos utilizados no presente projeto.

2.1 Fundamentação teórica do aprendizado de máquina

Estudiosos da Universidade de Berkeley, nos Estados Unidos, definem que um algoritmo de aprendizado de máquina consiste, basicamente, de três partes principais [11]:

- **Processo de decisão:** passos que um algoritmo toma para realizar uma generalização dos dados de entrada, o que possibilita encontrar padrões para realizar previsões.
- **Função erro:** cálculos que retornam a avaliação da previsão, como, por exemplo, taxa de erro ou variação.
- **Processo de otimização do modelo:** métodos que levam em consideração a minimização do erro durante o processo de aprendizado do modelo.

O aprendizado de máquina, geralmente, pode ser classificado de quatro formas diferentes, de acordo com a sua forma de aprendizado. São elas [12]:

- **Aprendizado supervisionado:** o processo de aprendizado da máquina se dá pelas entradas e saídas de dados pareadas, o que chamamos de dados rotulados. A máquina identifica padrões e aprende através das suas observações, podendo então realizar previsões para futuras entradas de dados.
- **Aprendizado semi-supervisionado:** são utilizados ambos os dados - rotulados e não rotulados -, sem uma saída conhecida. Acredita-se que esse método generaliza melhor novos dados, já que pode modificar ou repriorizar as hipóteses criadas apenas com os dados rotulados [13].
- **Aprendizado não supervisionado:** empregam-se apenas dados não pareados. A máquina utiliza dados de entrada para tentar interpretar e encontrar padrões intrínsecos neles.

- **Aprendizado por reforço:** os dados de entrada não são pareados com os de saída; porém contam com um sinal de recompensa, que deve ser maximizado com o tempo.

Neste projeto, será utilizado o aprendizado supervisionado.

2.2 Aprendizado supervisionado

O aprendizado supervisionado pode ser classificado em dois tipos de problemas: o de classificação (a saída desejada é um dado categórico) e o de regressão (quando a saída desejada é um dado contínuo) [14].

Para a realização do modelo de classificação proposto neste projeto, foram escolhidos três algoritmos de classificação com aprendizado supervisionado: máquina de vetores de suporte, floresta aleatória e redes neurais artificiais.

Um classificador generaliza as informações de entrada e atribui uma probabilidade para cada saída, as chamadas classes. Esses modelos podem ser binários (quando só apresentam duas classes) ou multiclasse (quando apresentam mais de duas classes). A classificação final é escolhida a partir da classe à qual foi atribuída uma maior probabilidade de pertencimento.

A avaliação do modelo é realizada através de métricas que comparam os valores reais com os preditos. A Tabela 1 mostra como os dados são classificados em relação às suas respectivas predições [15].

Tabela 1: - Matriz de confusão.

	Classe real positiva	Classe real negativa
Classe predita positiva	Verdadeiro positivo (VP)	Falso negativo (FN)
Classe predita negativa	Falso positivo (FP)	Verdadeiro negativo (VN)

A acurácia, representada pela equação 2.1, mede a razão entre as predições corretas e o total de observações.

$$Acurácia = \frac{VP + VN}{VP + FP + VN + FN} \quad (2.1)$$

A sensibilidade, como mostra a equação 2.2, representa a fração de valores da classe positiva que foram corretamente classificados.

$$Sensibilidade = \frac{VP}{VP + FN} \quad (2.2)$$

Por fim, a precisão - equação 2.3 - mostra a relação entre os valores da classe positiva que foram corretamente classificados e a quantidade total que foi predita na classe positiva, tanto corretamente, como incorretamente.

$$Precisão = \frac{VP}{VP + FP} \quad (2.3)$$

2.2.1 Máquinas de vetores de suporte

O algoritmo de aprendizado supervisionado de *support vector machine* (SVM) tem como objetivo, no caso da classificação, a diferenciação de pontos em um hiperplano em um espaço n-dimensional, sendo n o número de preditores. Essa diferenciação é realizada através da obtenção do hiperplano que apresenta uma distância maior entre as margens dos vetores de cada classe dos dados de entrada [3].

A Figura 9 esquematiza como se faz o processo de decisão do SVM em um hiperplano com apenas dois preditores.

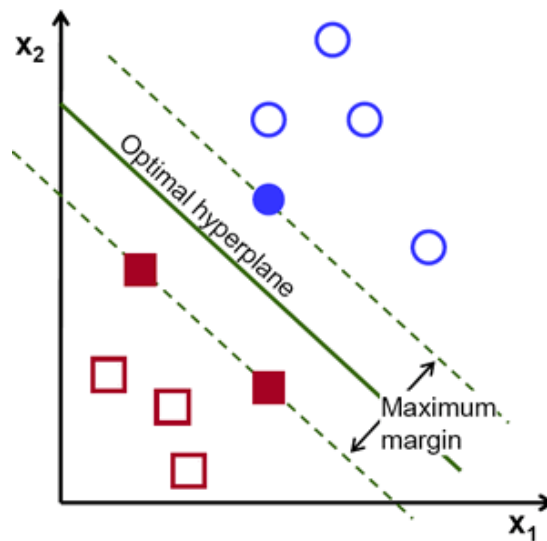


Figura 9: - Esquematização do SVM [3]

Tal método possui um bom desempenho quando existe pouca quantidade de amostras para cada classe, porém performa mal quando há muitos dados - já que exige muita capacidade computacional - e quando existem muitos *outliers* [16].

Um único classificador SVM não consegue realizar uma classificação de múltiplas classes, mas, sim, "quebrar" os dados em apenas duas classes. Para realizar múltiplas classificações binárias, utiliza-se mais de um modelo de classificação, o que é chamado de

one vs one [17].

A quantidade de vetores de suporte necessários é determinada a partir da quantidade de rótulos, como mostra a equação 2.4.

$$SVM = \frac{n(n-1)}{2} \quad (2.4)$$

2.2.2 Floresta Aleatória

O algoritmo de aprendizado de máquina *random forest* (RF) é constituído por um conjunto classificadores de árvores de decisão, que recebem como entrada vetores aleatórios independentes e identicamente distribuídos [18].

As árvores de decisão (AD) são compostas por nós, ramos e folhas, que representam o percurso que os dados de entrada percorrem para que seja realizada a predição. O nó realiza um teste em cada atributo dos dados, o ramo corresponde ao valor do atributo testado pelo nó e a folha representa a classificação [19]. Como saída da AD, há uma probabilidade de cada dado específico pertencer a uma classe.

O RF trata cada AD de forma independente, atribuindo uma amostra dos dados de entrada para cada uma e escolhendo a moda do resultado delas como classe final, de forma a aprimorar a sua acurácia e impedir o *overfitting*.

A Figura 10 mostra como esse processo é realizado.

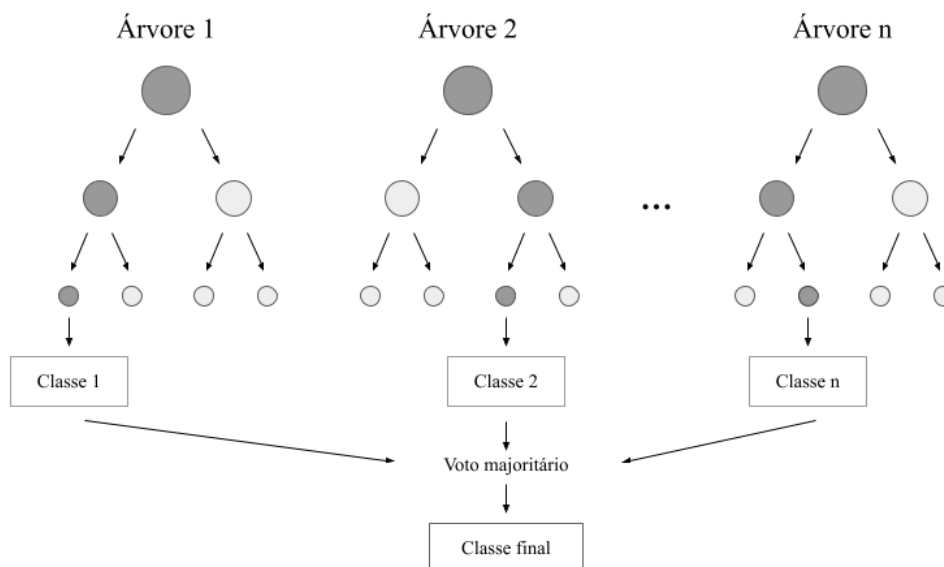


Figura 10: - Esquemática da floresta aleatória.

2.2.3 Redes Neurais Artificiais

O objetivo das redes neurais artificiais (ANN) é simular o funcionamento do sistema nervoso biológico computacionalmente, de forma que uma máquina replique - dentro de suas capacidades - o processo de aprendizagem de um cérebro [4].

Uma rede neural consiste de uma camada de entrada, de uma ou mais camadas ocultas e de uma camada de saída. No caso de mais de uma camada oculta, a rede passa a se denominar de rede neural profunda, e as implementações desse tipo de rede se classificam como aprendizado profundo [4]. A quantidade de neurônios na camada de entrada corresponde à quantidade de preditores que a base de dados possui; já na saída, esse número representa a quantidade de classes em que se deseja realizar a classificação.

Como entrada de cada neurônio da camada oculta ou de saída, deve-se considerar a soma da multiplicação de um peso pelo valor dessa entrada, que representa a saída de cada neurônio da camada anterior, mais um valor constante final, chamado de viés.

A função de ativação entra para transformar o sinal de entrada, de forma a gerar uma saída para a próxima camada de neurônios. Sem essa etapa, a relação entre uma camada e outra seria linear, o que não é ideal em situações reais. A utilização da função de ativação garante uma não linearidade entre as relações, possibilitando a execução de tarefas mais complexas.

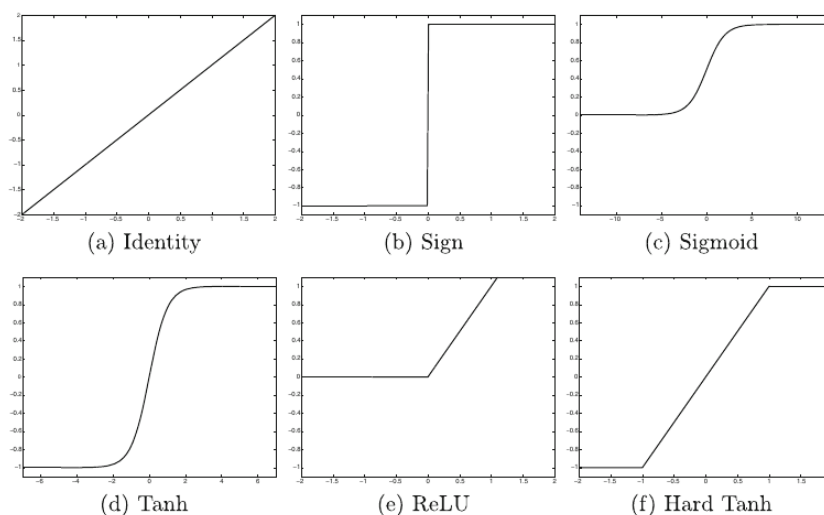


Figura 11: - Funções de ativação [4]

Além das funções de ativação representadas na Figura 11, também existe a *Soft-Max*, que é utilizada em problemas multiclasse. Nela, a saída da camada da rede são

valores de probabilidades para cada classe, que, quando somados, resultam em 1. A classificação final é atribuída ao rótulo que obteve o maior valor de probabilidade no momento da predição.

A saída de um neurônio é dada pela equação 2.5.

$$y_n = \Phi(\Sigma(W_n x_n) + b_n) \quad (2.5)$$

Onde n representa o neurônio, Φ a função de ativação, W_n o peso, x_n a entrada e b_n o viés. A Figura 12 mostra a esquematização de uma rede neural.

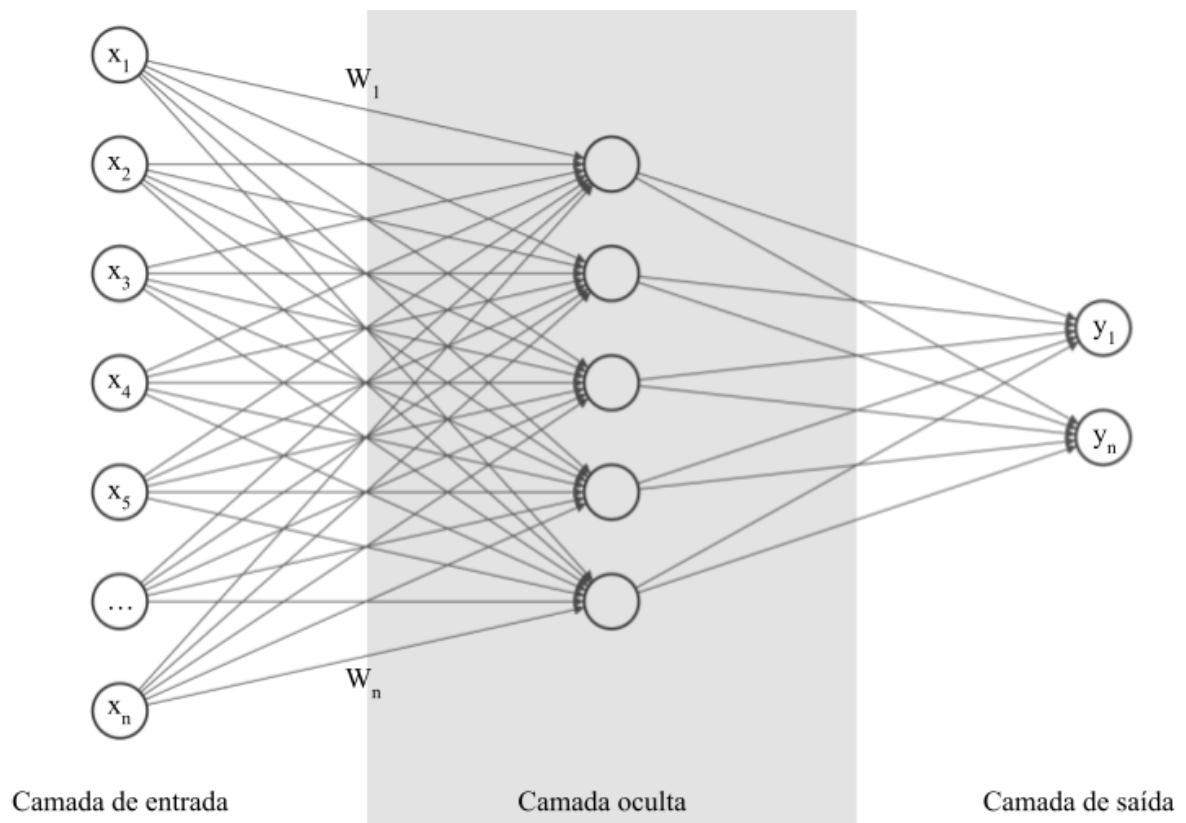


Figura 12: - Esquematização da rede neural

3 PROJETO

No presente capítulo, será apresentada a proposta do classificador.

O objetivo deste projeto é criar um modelo de aprendizado supervisionado de classificação que tenha capacidade de distinguir o instrumento principal em uma amostra de áudio polifônica.

A proposta desse projeto se dá a partir de uma base de dados de amostras reais de músicas, da qual serão extraídas informações temporais e espectrais dos sinais, cujos dados serão tratados e analisados. Para efeito de comparação, o classificador desenvolver-se-á em três modelos de aprendizado de máquina diferentes, e os resultados deles serão analisados para a escolha do algoritmo com melhor desempenho geral.

A Figura 13 esquematiza essa proposta.

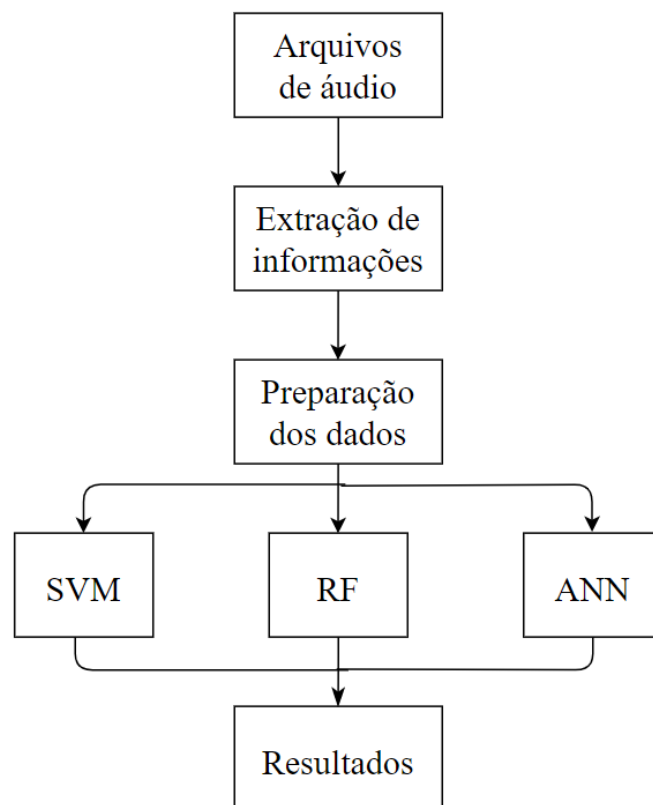


Figura 13: - Proposta de projeto

3.1 Base de dados

Para este projeto, escolheu-se a base de dados IRMAS [20], frequentemente utilizada em estudos de reconhecimento de instrumentos musicais [21].

O conjunto de dados compõe-se de 3.716 arquivos de áudio em formato wav estéreo de 16 bits, amostrados em 44,1kHz. Eles apresentam trechos de 3 segundos de gravações polifônicas distintas, incluindo músicas reais gravadas tanto no período atual como em diversas décadas do passado, além de diferentes qualidades de áudio, estilos, artistas e tipos de instrumentos.

O título de cada arquivo de áudio traz diversas informações, tais como: a identificação do instrumento predominante, o estilo da música e o código de identificação desta. Sendo assim, uma determinada gravação será identificada através de um código, este seguido de um subcódigo referente a cada uma das amostras que compõem a música em questão. É importante salientar que cada gravação conterà até três amostras diferentes, cada qual com a duração de três segundos.

Como exemplificação, a Figura 14 mostra um exemplo esquematizado:

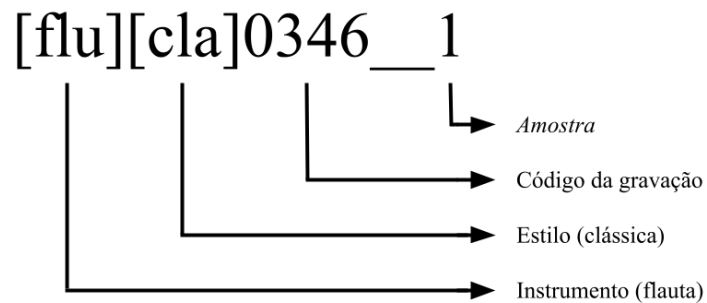


Figura 14: - Exemplo de nome de arquivo da base de dados

A seguir, a Tabela 2 mostra a distribuição de amostras de áudios para cada instrumento, junto com as siglas que os representam.

Tabela 2: - Quantidade de amostras para cada instrumento

Sigla	Instrumento	Quantidade
flu	Flauta	451
gel	Guitarra	760
pia	Piano	721
sax	Saxofone	626
tru	Trompete	577
vio	Violino	580

3.2 Extração de informações

Os áudios da base de dados foram processados no *python* [22] através da biblioteca *librosa* [23] de forma que o sinal de áudio com dois canais de reprodução (estéreo) foi reduzido para apenas um (monofônico).

Como realizado por Racharla, K. et al., as informações do sinal escolhidas para serem extraídas foram: valor quadrático médio (RMS), centróide espectral (SC), largura de banda espectral (SB), frequência de *roll-off*, taxa de cruzamento do zero (ZCR) e 20 coeficientes cepstrais de frequência-Mel (MFCCs). Esses dados foram dispostos em formato tabular, sendo utilizada a média desses valores para cada áudio, já que o retorno dessas funções será o valor de cada quadro analisado [24].

3.2.1 Valor quadrático médio

O valor de RMS é utilizado para representar a energia do sinal, carregando o conceito de altura no sinal de áudio. [25].

A Figura 15 mostra a distribuição da média do valor de RMS da base de dados de forma segmentada por cada instrumento. Nela, é possível observar uma diferença de comportamento principalmente nas ocorrências de guitarra, que apresenta valores geralmente mais altos em relação aos outros instrumentos. Também percebe-se que o RMS do piano, saxofone, trompete e violino é concentrado em valores mais baixos.

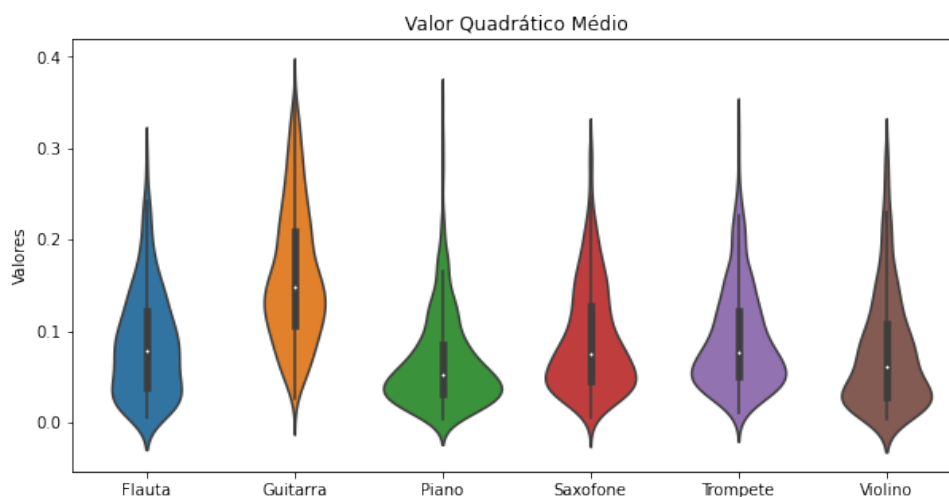


Figura 15: - Distribuição do RMS

3.2.2 Centróide espectral

O centróide espectral é a frequência média do centro de gravidade do espectrograma. Esse valor é uma boa representação do timbre do instrumento, já que é um bom indicador do "brilho" do som [24].

Na distribuição segmentada da Figura 16, observa-se que o piano possui frequências mais baixas. Já a guitarra se diferencia dos outros instrumentos de forma que seus valores estão concentrados acima da média deles.

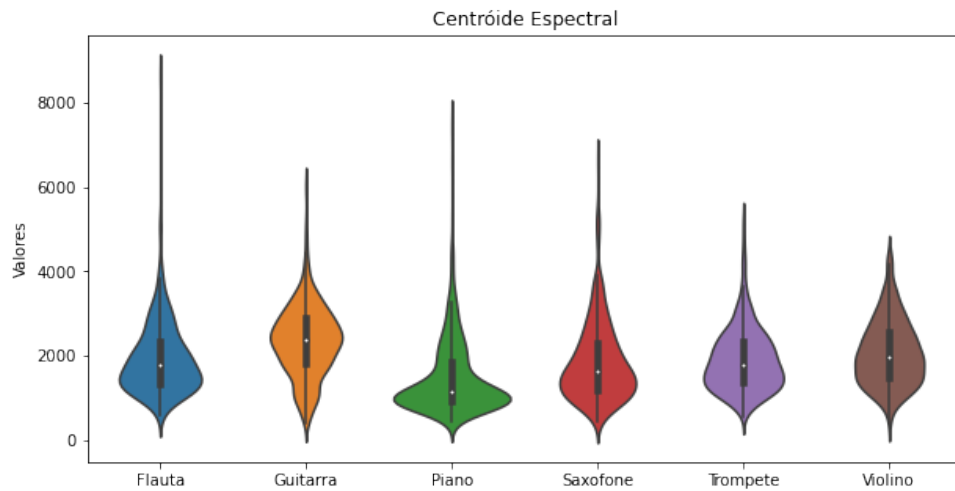


Figura 16: - Distribuição do SC

3.2.3 Largura de banda espectral

A largura de banda espectral é descrita pelo espalhamento espectral, é a média ponderada das frequências em torno do centróide espectral do seu quadro. Esse valor constitui um bom indicador de timbre [25] [24].

Na Figura 17, percebe-se uma forte semelhança entre o piano, o saxofone e o trompete, visto que seus valores são majoritariamente mais baixos. A flauta e o violino possuem uma distribuição mais bem distribuída e, por fim, a guitarra tem valores de frequências um pouco mais altos.

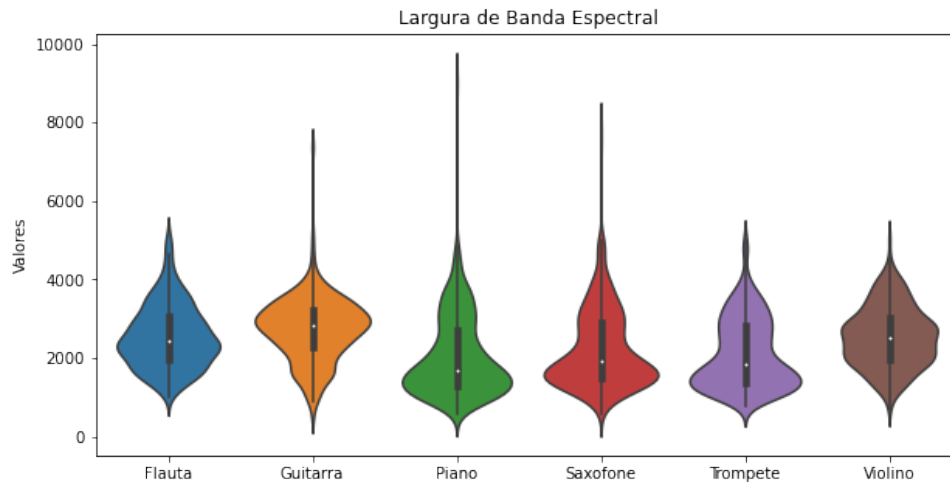


Figura 17: - Distribuição da SB

3.2.4 Frequência de Rolloff

A frequência de *rolloff* é o valor em que a energia do sinal chega a 85% (valor predefinido pela librosa) do seu valor total. Essa fração de energia é considerada a mais substancial, enquanto as frequências que representam os 15% restantes são interpretadas como interferências ou ruídos [23].

A distribuição da Figura 18 mostra que a energia do piano se concentra majoritariamente em valores mais baixos; no caso da guitarra, os valores de frequência são mais altos. Os demais instrumentos possuem comportamentos mais parecidos entre si.

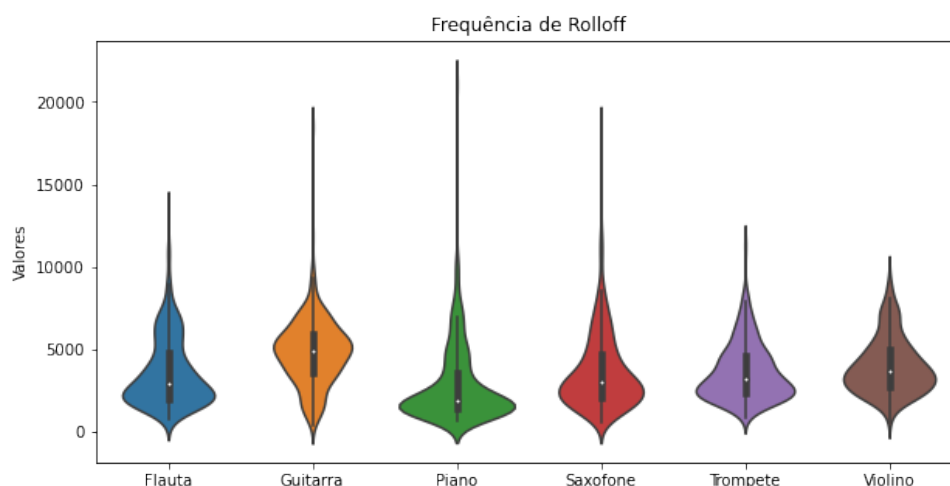


Figura 18: - Distribuição da frequência de *rolloff*

3.2.5 Zero Crossing Rate

O ZCR representa a taxa da quantidade de vezes que o sinal passa pelo zero, ou seja, que ele muda sua direção (positivo para negativo e vice-versa). Esse dado também é uma forma de representação do "brilho" do som, já que taxas maiores indicam uma frequência mais alta. [25].

A distribuição segmentada por instrumento da Figura 19 mostra comportamentos muito parecidos entre a flauta e o piano, que possuem taxas bem distribuídas em torno da sua média. O trompete também apresenta a maioria dos seus valores na média, porém essa média é um pouco mais alta. A guitarra e o violino apresentam taxas mais distribuídas em torno dos seus limites.

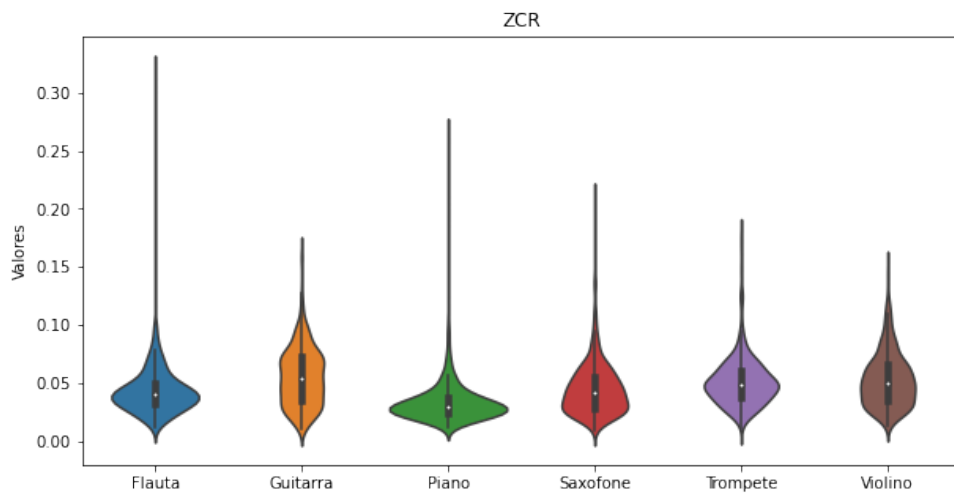


Figura 19: - Distribuição do ZCR

3.2.6 Coefficientes Cepstrais de Frequência Mel

Os MFCCs são os coeficientes da escala Mel, que é uma melhor forma de representação da audição humana, como mostrado na seção 1.1 deste trabalho.

Eles são obtidos através da realização da transformada discreta do cosseno do espectro logarítmico do sinal na escala Mel. Essa informação representa o timbre do som e a qualidade dele, sendo um bom indicador da forma como o som foi gerado [25] [26].

O padrão das distribuições dos coeficientes variaram bastante entre os instrumentos, porém, a partir do décimo e terceiro MFCC, as distribuições se assemelharam muito.

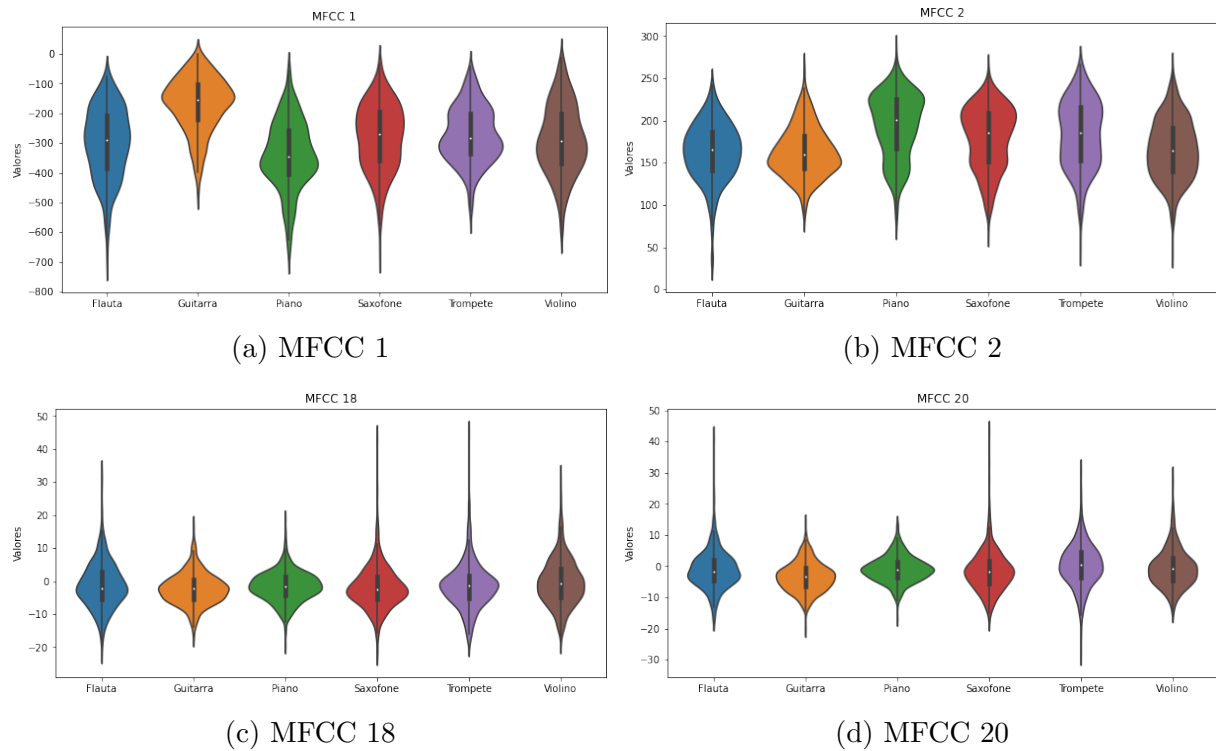


Figura 20: - Coeficientes Cepstrais de Mel

3.3 Preparação da base

Os dados foram preparados para treino utilizando os pacotes do *python*: *scikit-learn* [27] e *numpy* [28].

Após a obtenção dos dados em forma tabular, esses dados foram embaralhados aleatoriamente, para que cada classe não estivesse muito próxima uma da outra. Em seguida, os dados de treino e teste foram separados, também de forma aleatória, para o processo de aprendizado e avaliação dos modelos. Os dados de treino representam 70% da base, enquanto os de teste representam os outros 30%.

A Tabela 3 mostra a quantidade de dados para cada instrumento na base de treino e de teste.

Tabela 3: - Quantidade de amostras para cada instrumento para base de treino e de teste.

Instrumento	Treino	Teste
Flauta	301	150
Guitarra	519	241
Piano	527	194
Saxofone	445	181
Trompete	388	189
Violino	420	160

Os dados de treino foram normalizados utilizando o *Standard Scaler* [27], que

padroniza os preditores individualmente, transformando a média em 0 e escalonando a variância deles a uma unidade. Após a aplicação nos dados de treino, o mesmo modelo de padronização foi aplicado nos dados de teste, a fim de evitar um vazamento de dados (*data leakage*).

Certos modelos de aprendizado de máquina precisam que os dados categóricos estejam vetorizados, para que o treino possa ser realizado, o que foi o caso da ANN. Então utilizou-se o *LabelEncoder* [27] e o *to_categorical* [28] para realizar essa separação. Dessa forma, as classes foram representadas como mostra a Tabela 4.

Tabela 4: - Representação numérica e vetorial das classes.

Instrumento	Numérico	Vetor
Flauta	0	100000
Guitarra	1	010000
Piano	2	001000
Saxofone	3	000100
Trompete	4	000010
Violino	5	000001

3.4 Classificadores

Como já citado anteriormente, foram escolhidos 3 modelos de classificação de aprendizado de máquina supervisionado, são eles: SVM, RF e ANN.

No caso de SVM e RF, foram realizadas buscas de melhores hiperparâmetros maximizando a métrica de acurácia geral do modelo, utilizando o pacote *GridSearchCV* [27].

O *grid search* recebe um modelo, uma lista de valores para seus diversos hiperparâmetros e a métrica que se deseja maximizar para serem testados. Em seguida, é realizado um treino com validação cruzada para cada combinação possível desses hiperparâmetros, retornando, então, os valores deles (hiperparâmetros) que resultaram em uma melhor avaliação do modelo.

A validação cruzada é um procedimento que divide a base de dados de treino em subconjuntos (chamados de *folds*) de treino e validação e retorna as métricas para cada um deles. Essa técnica é utilizada a fim de evitar um *overfitting*, quando o modelo não consegue encontrar um padrão em seus dados de entrada.

Nos modelos desse projeto (SVM e RF), utilizaram-se 5 *folds* para a validação cruzada e a acurácia como métrica a ser maximizada.

3.4.1 Projeto do SVM

Para o desenvolvimento de um modelo de classificação SVM, aplicou-se o *Support Vector Classifier* (SVC), com os seguintes hiperparâmetros obtidos pelo *grid search*, além dos valores predeterminados pela classe do modelo:

- **C**: 10, parâmetro de regularização;
- **kernel**: *radial basis function* (rbf), função utilizada para diminuir a complexidade do cálculo do hiperplano;
- **gamma**: 0,1, coeficiente do *kernel*.

3.4.2 Projeto da RF

Para a criação da RF, além dos valores padrão, foram utilizados os hiperparâmetros testados resultantes do *grid search*:

- **n_estimators**: 100, quantidade de árvores;
- **max_depth**: 10, profundidade da árvore;
- **max_features**: 100%, porcentagem de *features* consideradas em cada *split*;
- **min_samples_leaf**: 2, quantidade mínima de amostras em cada folha da árvore;
- **min_samples_split**: 8, quantidade mínima de amostras para realizar o *split* de um nó interno.

3.4.3 Projeto da ANN

Para projetar a ANN, empregou-se a biblioteca *Keras* [29] do *python*.

A rede neural foi construída com apenas uma camada oculta, a fim de não ser utilizado o aprendizado profundo. Sendo assim, a arquitetura da ANN é composta de:

- Uma **camada de entrada** com 25 neurônios, valor correspondente à quantidade de preditores;
- uma **camada oculta** densa com 128 neurônios, com função de ativação *ReLU*;

- uma **camada de saída** com 6 neurônios, que correspondem a cada classificação de instrumento, utilizando a função de ativação *SoftMax*.

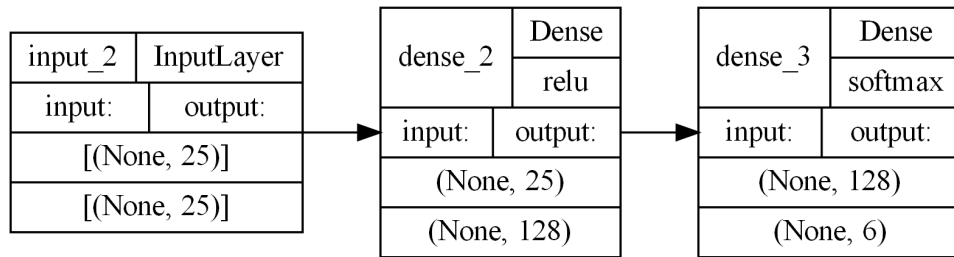


Figura 21: - ANN projetada.

4 RESULTADOS

Neste capítulo, serão apresentadas as métricas de teste (precisão e sensibilidade) para cada instrumento, a acurácia de teste geral, bem como os 10 preditores mais influentes de cada modelo, através do pacote SHAP [30] do *python*.

4.1 Resultado do SVC projetado

O modelo SVC apresentou uma acurácia geral de 72%, além das precisões e sensibilidades mostradas na Tabela 5. Observa-se, em geral, resultados muito bons para todos os instrumentos, principalmente para a guitarra, o trompete e o piano.

Tabela 5: - Métricas resultantes do SVC.

Instrumento	Precisão	Sensibilidade
Flauta	62%	65%
Guitarra	78%	84%
Piano	74%	73%
Saxofone	67%	65%
Trompete	78%	69%
Violino	69%	70%

A Figura 22 mostra um mapa de calor da matriz de confusão, que relaciona a classe real e a predita. As cores mais escuras - apenas na diagonal principal - indicam um alto índice de classificação realizada corretamente.

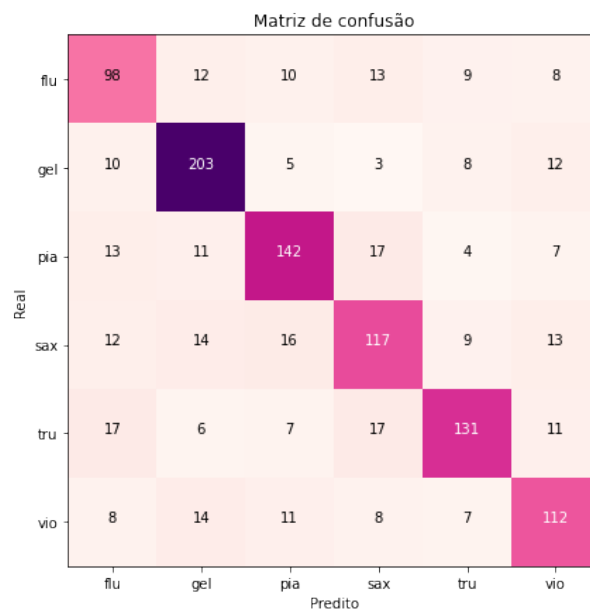
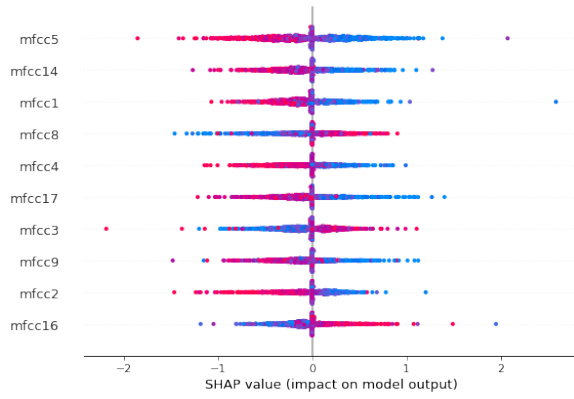
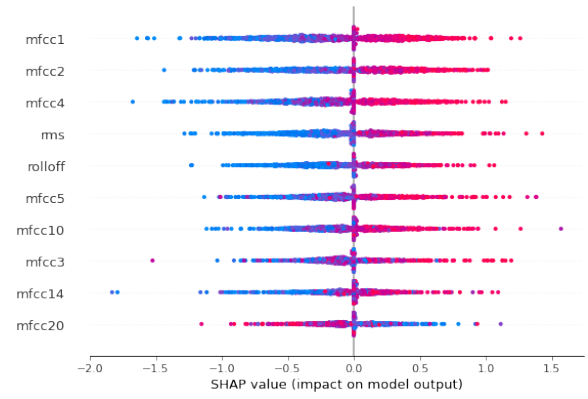


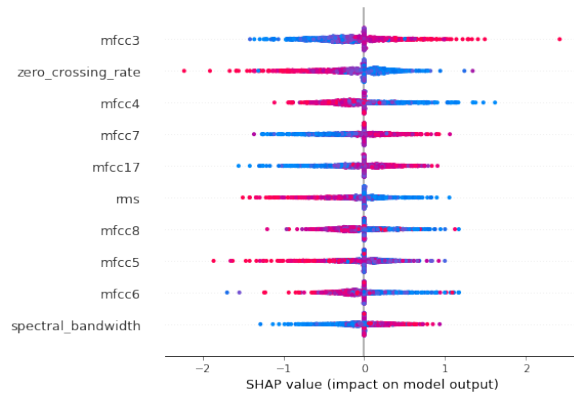
Figura 22: - Classificações SVC.



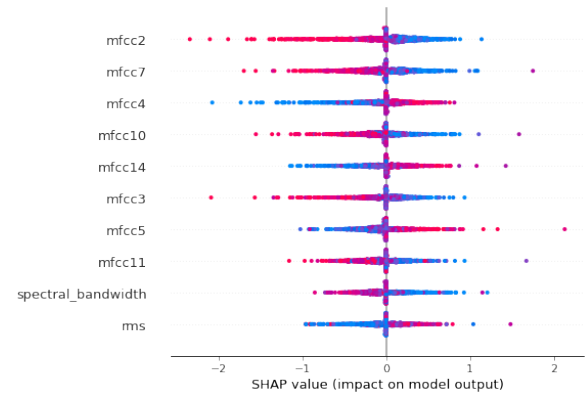
(a) Flauta



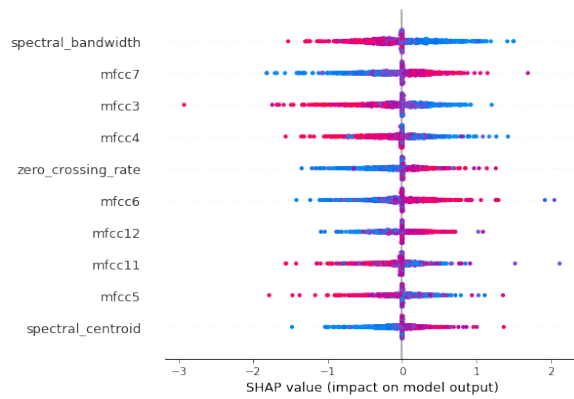
(b) Guitarra



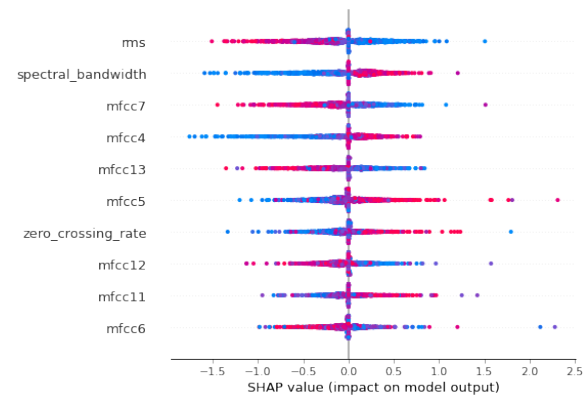
(c) Piano



(d) Saxofone



(e) Trompete



(f) Violino

Figura 23: - Influência dos *top* 10 preditores do SVC.

Os gráficos do SHAP, na Figura 23, mostram uma influência positiva e negativa bem delimitada para valores altos (vermelho) e baixos (azul) de cada preditor. Também percebe-se que, para todos os instrumentos, os MFCCs tiveram bastante importância na predição.

4.2 Resultado da RF projetada

A RF teve uma acurácia geral razoável de 57%. As precisões e sensibilidades expostas na Tabela 6 apresentam resultados medianos para todos os instrumentos, exceto para o caso da guitarra, que obteve métricas bem altas.

Tabela 6: - Métricas resultantes da RF.

Instrumento	Precisão	Sensibilidade
Flauta	60%	35%
Guitarra	88%	76%
Piano	52%	75%
Saxofone	52%	43%
Trompete	66%	53%
Violino	62%	51%

Importante observar que a flauta possui uma sensibilidade bem baixa para uma precisão boa, ou seja, existe um alto índice de acerto para a predição da classe flauta; no entanto há uma alta incidência de erro quando se trata de identificar todas as amostras cujo instrumento predominante é a flauta, como mostra a Figura 24.

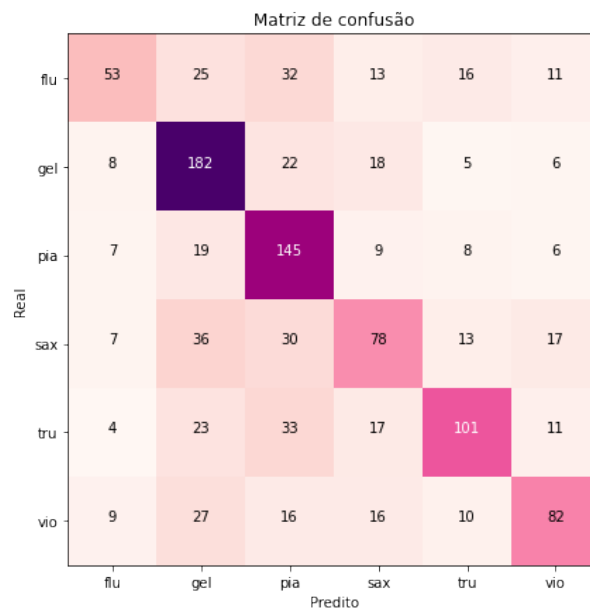
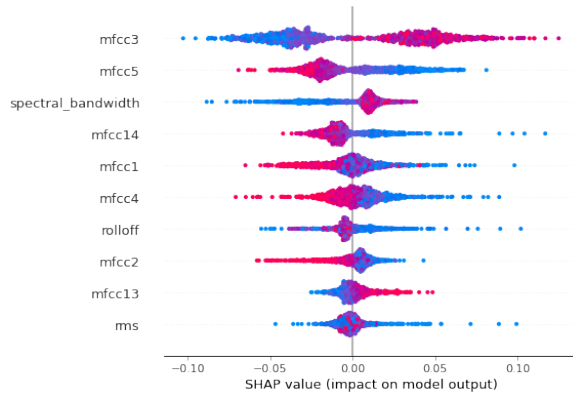
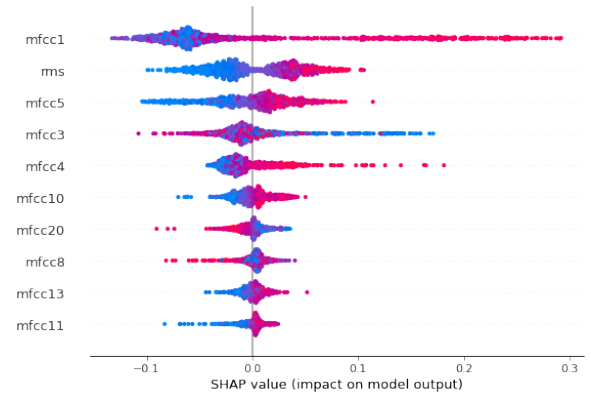


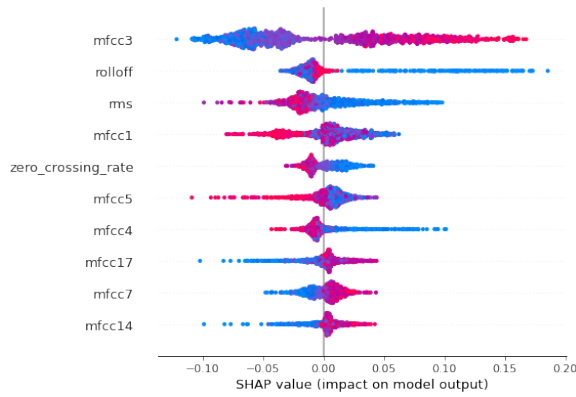
Figura 24: - Classificações RF.



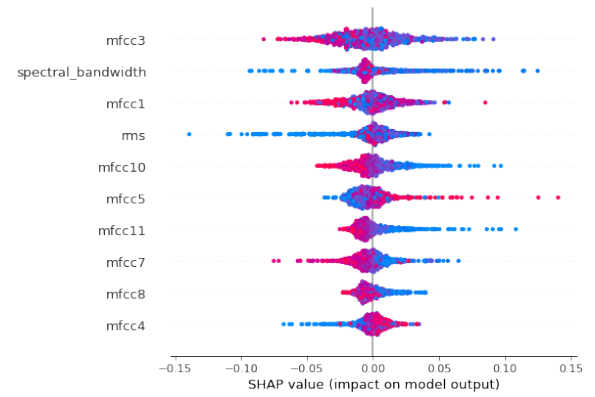
(a) Flauta



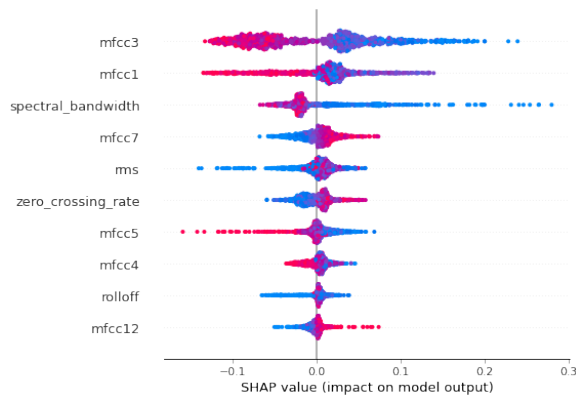
(b) Guitarra



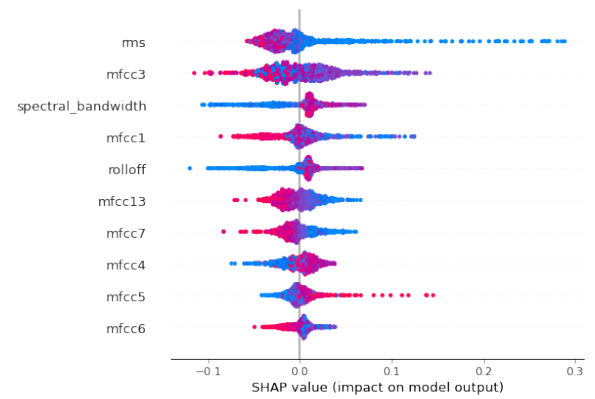
(c) Piano



(d) Saxofone



(e) Trompete



(f) Violino

Figura 25: - Influência dos *top* 10 preditores da RF.

Já o SHAP da Figura 25 mostra tanto influências distintas para valores altos e baixos - as cores azul e vermelho se misturam pouco -, quanto a não distinção desses valores. Os MFCCs continuam sendo, predominantemente, os preditores mais influentes.

4.3 Resultado da ANN projetada

A rede neural projetada apresentou uma acurácia bem baixa, de 48%, o que era esperado devido a sua simplicidade - apenas uma camada oculta - e à pequena quantidade de amostras para cada instrumento contida na base de dados.

A Tabela 7 mostra que, assim como a acurácia, as métricas de sensibilidade e precisão também foram bem baixas, exceto para a guitarra e para o piano, que foram razoáveis na precisão e altas para a sensibilidade.

Tabela 7: - Métricas resultantes da ANN.

Instrumento	Precisão	Sensibilidade
Flauta	46%	25%
Guitarra	48%	81%
Piano	51%	72%
Saxofone	40%	24%
Trompete	49%	29%
Violino	51%	41%

A matriz de confusão da Figura 26 traduz as métricas apresentadas, mostrando um alto índice de classificações corretas para a guitarra e o piano e um baixo para os outros instrumentos.

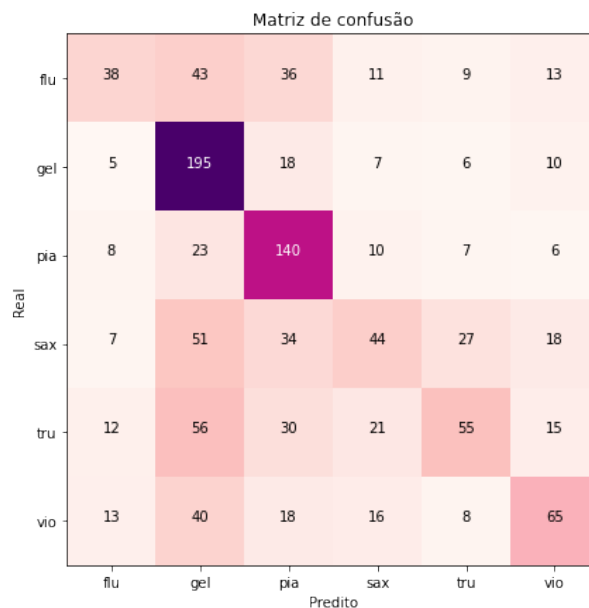


Figura 26: - Classificações ANN.

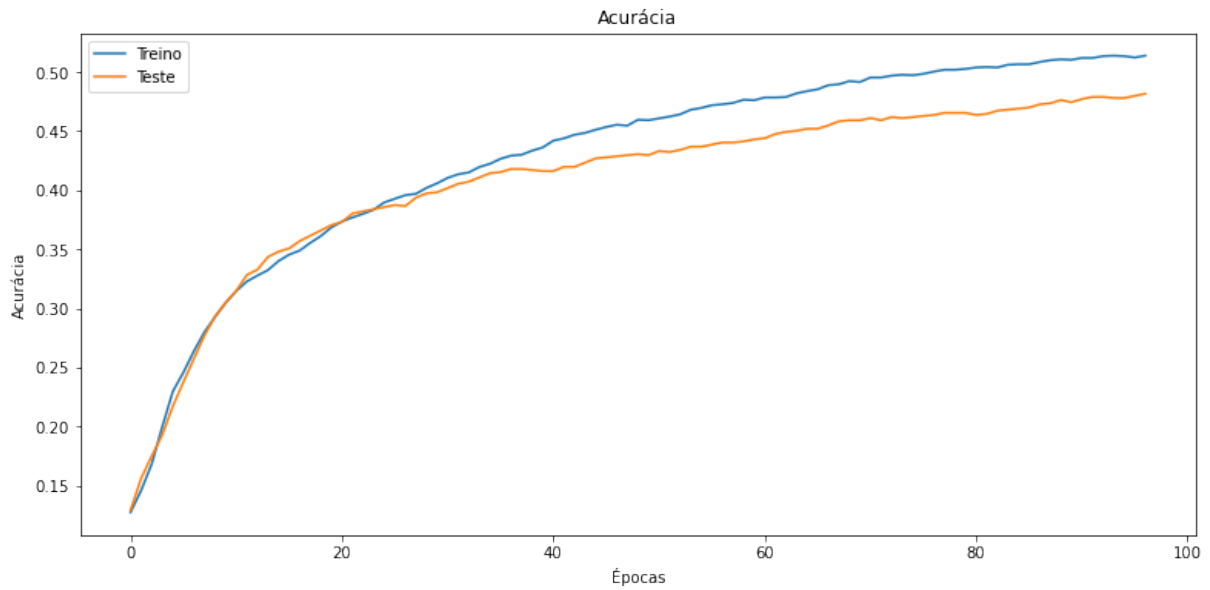
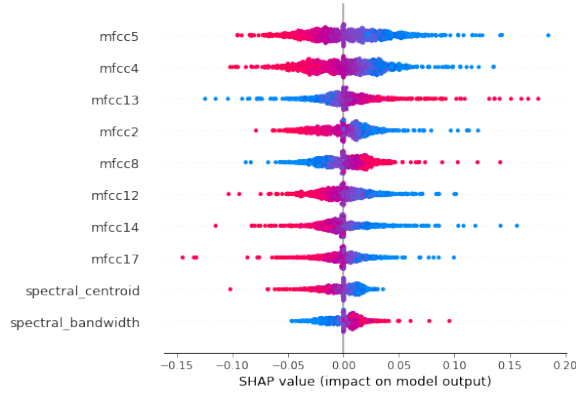


Figura 27: - Acurácia da ANN x épocas.

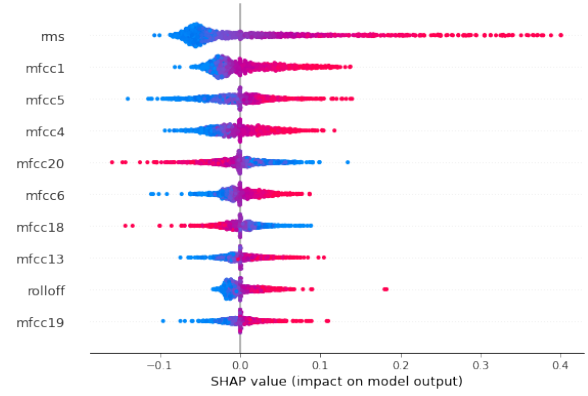
No modelo de rede neural, é possível avaliar como ocorreu o processo de aprendizado no decorrer das épocas, que indicam quantas vezes o algoritmo utilizou a base de dados inteira em seu treinamento.

Na rede projetada, como indica a Figura 27, foram necessárias 97 épocas para se chegar ao valor ótimo entre a acurácia de treino e de teste - 51% e 48%, respectivamente.

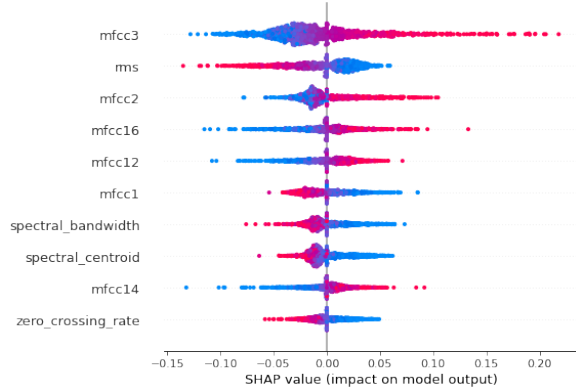
Assim como nos outros modelos, o SHAP da rede neural considerou uma influência bem alta para os preditores de MFCC, como mostra a Figura 28. Apesar de ainda haver uma diferenciação entre as cores dos valores altos e baixos dos preditores (vermelho e azul), nesse caso, elas são consideravelmente mais misturadas do que as apresentadas na Figura 23 e na Figura 25, o que indica uma dificuldade na classificação dos instrumentos, comprovada pelas métricas baixas.



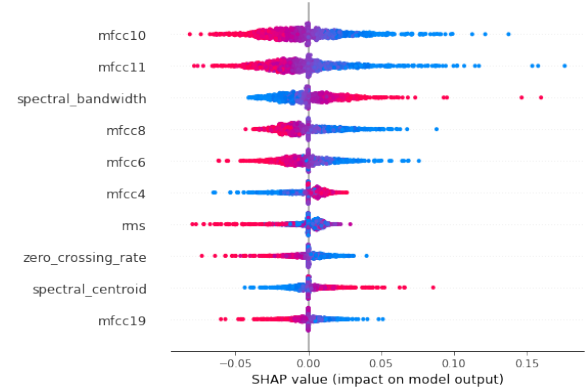
(a) Flauta



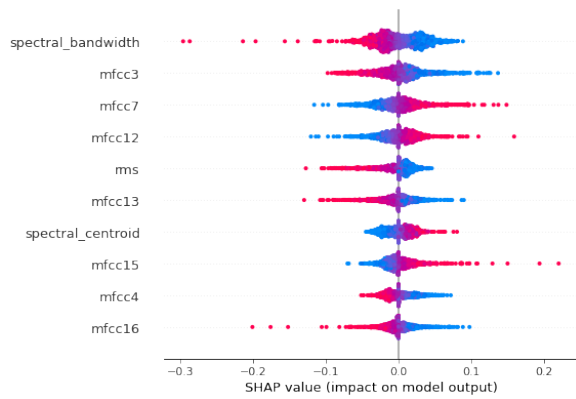
(b) Guitarra



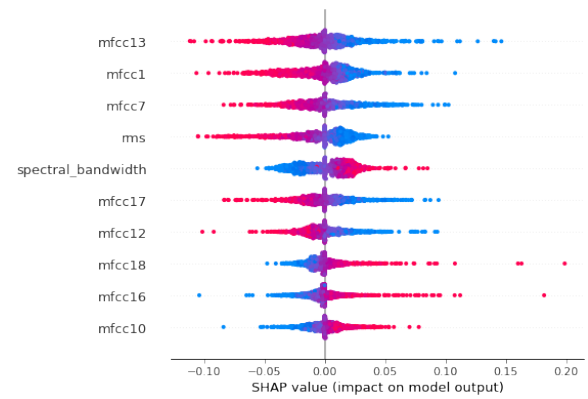
(c) Piano



(d) Saxofone



(e) Trompete



(f) Violino

Figura 28: - Influência dos *top* 10 preditores da ANN.

CONCLUSÃO

Após a obtenção dos resultados de métricas, de acertos totais e de influenciadores, expostos no capítulo 4, procede-se a uma comparação entre eles.

Tabela 8: - Acurácia de cada modelo projetado.

Modelo	Acurácia
SVC	72%
RF	57%
ANN	48%

Na Tabela 8, que resume a acurácia geral de todos os modelos testados, observa-se que o classificador baseado em SVC obteve um resultado muito melhor em comparação aos outros, com uma diferença de 15% para RF e 24% para ANN. Esse resultado já era esperado, pois, como citado na subseção 2.2.1, o SVC trabalha melhor com poucas amostras, o que é a realidade deste projeto.

Tabela 9: - Resumo das métricas de cada instrumento para cada classificador.

Instrumento	SVC		RF		ANN	
	Precisão	Sensibilidade	Precisão	Sensibilidade	Precisão	Sensibilidade
Flauta	62%	65%	60%	35%	46%	25%
Guitarra	78%	84%	88%	76%	48%	81%
Piano	74%	73%	52%	75%	51%	72%
Saxofone	67%	65%	52%	43%	40%	24%
Trompete	78%	69%	66%	53%	49%	29%
Violino	69%	70%	62%	51%	51%	41%

Em geral, os instrumentos mais facilmente identificados em uma música foram a guitarra e o piano, como mostram as métricas da Tabela 9. Tal fato pode ser comprovado através das análises dos dados, realizadas na seção 3.1, que demonstraram que a distribuição deles difere bastante entre si e entre os demais instrumentos.

Ainda, ao comparar os mapas de calor da Figura 22, da Figura 24 e da Figura 26, observa-se que os modelos encontraram uma dificuldade maior em identificar as amostras de flauta, saxofone, trompete e violino, confundindo-as, principalmente, com as classes de guitarra e de piano. Esse fato pode ser explicado pela semelhança entre a distribuição dos dados, dificultando a diferenciação no momento do treino dos modelos. Uma outra explicação para isso pode ser o fato de a base de dados não ser balanceada, tendo uma quantidade significativamente maior de amostras para guitarra e piano, como demonstra a Tabela 2.

Diante do exposto, conclui-se que o objetivo deste estudo de projetar um classificador de instrumentos musicais foi atingido, ao utilizar o modelo de aprendizado super-

visionado de máquina baseado em máquinas de vetores de suporte, o SVC. Apesar desse tipo de algoritmo exigir uma maior capacidade computacional, ele obteve resultados consideravelmente superiores quando comparados aos outros baseados em florestas aleatórias (RF) e em redes neurais simples (ANN).

Por fim, são sugeridos como próximos passos para aprimoramento do classificador:

- a obtenção de um conjunto mais amplo de amostras musicais, com o objetivo de apresentar casos mais diferenciados para o modelo no momento da aprendizagem;
- o balanceamento da base de dados das amostras, com o propósito de evitar o enviesamento de classes no momento do treino;
- a extração de mais informações dos áudios, para tentar minimizar o monopólio dos MFCCs como principais influenciadores;
- o estudo da aplicação de redes neurais com mais camadas, *deep learning*, com o intuito de se realizar um melhor aprendizado;
- a análise do uso de imagens dos espectrogramas do sinal como preditores para a distinção entre os instrumentos.

REFERÊNCIAS

- [1] Gibson. *Jerry Cantrell "Wino" Les Paul Custom (Aged Signed)*. Disponível em: <https://www.gibson.com/en-US/Electric-Guitar/CUSKBN534/Wine-Red>. Acesso em: 14 de agosto de 2022.
- [2] Yamaha USA. *Yamaha: Make Waves*. Disponível em: <https://usa.yamaha.com/>. Acesso em: 14 de agosto de 2022.
- [3] Rohith Gandhi. *Support Vector Machine — Introduction to Machine Learning Algorithms*. Disponível em: <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>. Acesso em: 23 de julho de 2022.
- [4] AGGARWAL, C. C. *Neural Networks and Deep Learning*. 1. ed. [S.l.]: Springer Cham, 2018.
- [5] FENG, J. Q. Music in terms of science. *ArXiv*, abs/1209.3767, 2012.
- [6] FLETCHER, N. H.; ROSSING, T. D. *The Physics of Musical Instruments*. 1. ed. [S.l.]: Springer New York, NY, 1991.
- [7] DOBRIAN, C. Msp: The documentation. *Cycling '74 and IRCAM*, Dezembro 1997.
- [8] LLOYD, L. S. *Music and Sound*. [S.l.]: Ayer Publishing.
- [9] VIRTANEN, T.; PLUMBLEY, M. D.; ELLIS, D. *Computational Analysis of Sound Scenes and Events*. [S.l.]: Springer.
- [10] Multiple Contributors. *Audio Representation*. Disponível em: https://musicinformationretrieval.com/audio_representation.html. Acesso em: 13 de agosto de 2022.
- [11] datascience@berkeley. *What Is Machine Learning (ML)?* Disponível em: <https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/>. Acesso em: 12 de julho de 2022.

- [12] Katrina Wakefield. *A guide to the types of machine learning algorithms and their applications*. Disponível em: <https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html>. Acesso em: 12 de julho de 2022.
- [13] ZHU, X. *Semi-Supervised Learning Literature Survey*. [S.l.], 2005.
- [14] IBM Cloud Education. *What is Supervised Learning?* Disponível em: <<https://www.ibm.com/cloud/learn/supervised-learning>>. Acesso em: 13 de julho de 2022.
- [15] HOSSIN, M.; M.N, S. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining Knowledge Management Process*, v. 5, p. 01–11, Março 2015.
- [16] MINGHUI, M.; CHUANFENG, Z. Application of support vector machines to a small-sample prediction. *Advances in Petroleum Exploration and Development*, Canadian Research Development Center of Sciences and Cultures, v. 10, n. 2, p. 72–75, Dezembro 2015.
- [17] Hucker Marius. *Multiclass Classification with Support Vector Machines (SVM), Dual Problem and Kernel Functions*. Disponível em: <<https://towardsdatascience.com/multiclass-classification-with-support-vector-machines-svm-kernel-trick-kernel-f>>. Acesso em: 23 de julho de 2022.
- [18] BREIMAN, L. Random forests. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, Janeiro 2001.
- [19] BREIMAN, L. et al. *Classification And Regression Trees*. 1. ed. [S.l.]: Routledge, 1894. 246–280 p.
- [20] BOSCH, J. J. et al. A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals. *Proc. ISMIR*.
- [21] GURURANI, S.; SHARMA, M.; LERCH, A. An attention mechanism for musical instrument recognition. *ArXiv*, abs/1907.04294, 2019.
- [22] PYTHON. Disponível em: <<https://docs.python.org/3/>>. Acesso em: 23 de agosto de 2022.

- [23] LIBROSA: Audio and music signal analysis in python. Disponível em: <<https://librosa.org/doc/latest/index.html>>. Acesso em: 23 de agosto de 2022.
- [24] RACHARLA, K. et al. Predominant musical instrument classification based on spectral features. In: *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*. [S.l.: s.n.], 2020. p. 617–622.
- [25] KLAPURI, A.; DAVY, M. *Signal Processing Methods for Music Transcription*. 1. ed. [S.l.]: Springer, 2006.
- [26] SELL, G.; MYSORE, G. J.; CHON, S. H. *Musical Instrument Detection Detecting instrumentation in polyphonic musical signals on a frame-by-frame basis*. 2006.
- [27] SCIKIT-LEARN: Machine Learning in Python. Disponível em: <<https://scikit-learn.org/stable/>>. Acesso em: 23 de agosto de 2022.
- [28] NUMPY: The fundamental package for scientific computing with Python. Disponível em: <<https://numpy.org/>>. Acesso em: 23 de agosto de 2022.
- [29] KERAS: a deep learning API written in Python. Disponível em: <<https://keras.io/>>. Acesso em: 30 de setembro de 2022.
- [30] SHAP: (SHapley Additive exPlanations). Disponível em: <<https://shap.readthedocs.io/en/latest/index.html>>. Acesso em: 3 de outubro de 2022.