



Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Faculdade de Engenharia

Gabriela Siqueira Eduardo

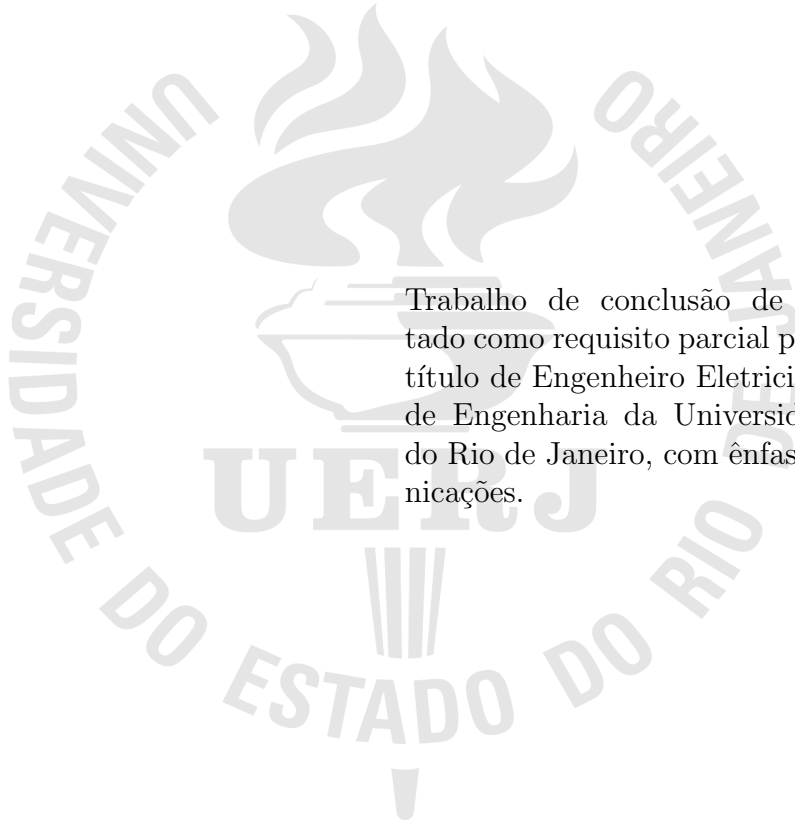
**Classificação de instrumentos musicais baseada em aprendizado
de máquina**

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. 000 f. Trabalho de Conclusão de Curso (Graduação em Engenharia Elétrica com Ênfase em Telecomunicações) - Departamento de Engenharia Eletrônica e de Telecomunicações, Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro, 2022.

Devido ao aumento da disponibilidade e da distribuição de músicas através das plataformas de *streaming*, torna-se relevante o uso de ferramentas de aprimoramento da experiência do usuário, como a classificação de áudios para criação, categorização e recomendação de catálogos. Neste trabalho, serão apresentadas informações temporais e espectrais de um sinal de áudio, uma introdução a modelos de aprendizado de máquina bem como a análise dos dados extraídos da base de composições polifônicas do IRMAS para cada instrumento musical selecionado. Além disso, serão projetados três métodos de classificação de múltiplas classes de instrumentos, sendo eles: Máquinas de Vetores de Suporte, Florestas Aleatórias e Redes Neurais Artificiais. Por fim, avaliar-se-á cada modelo através de suas métricas e principais preditores.

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 - Formas de onda de diferentes instrumentos para uma mesma nota musical.	15
Figura 2 - Faixas de frequência.	17
Figura 3 - Representação no domínio do tempo.	18
Figura 4 - Representação no domínio do tempo e da frequência.	19
Figura 5 - Efeito dos harmônicos em um sinal senoidal.	20
Figura 6 - Evolução do sinal ao longo do tempo.....	21
Figura 7 - Partes da guitarra - adaptação de <i>Gibson</i> [1].	22
Figura 8 - Violino.	23
Figura 9 - Piano.....	24
Figura 10- Flauta transversal [2].	25
Figura 11- Trompete, adaptado de <i>Yamaha</i> [2].	25
Figura 12- Saxofone, adaptado de <i>Yamaha</i> [2].	26
Figura 13- Esquematização do SVM [3]	29
Figura 14- Esquematização da floresta aleatória.	31
Figura 15- Funções de ativação [4] (Referenciar e exeplicar figura no texto.).....	32
Figura 16- Esquematização da rede neural	33
Figura 17- Proposta de projeto	34
Figura 18- Exemplo de nome de arquivo da base de dados	35
Figura 19- Distribuição do RMS	36
Figura 20- Distribuição do SC	37
Figura 21- Distribuição da SB	38
Figura 22- Distribuição da frequência de <i>rolloff</i>	38
Figura 23- Distribuição do ZCR	39
Figura 24- Coeficientes Cepstrais de Mel	40
Figura 25- ANN projetada.....	43
Figura 26- Classificações SVC.	44
Figura 27- Influência dos <i>top</i> 10 preditores do SVC.	45
Figura 28- Classificações RF.....	46
Figura 29- Influência dos <i>top</i> 10 preditores da RF.....	47

Figura 30- Classificações ANN.	48
Figura 31 - Acurácia da ANN x épocas.	49
Figura 32- Influência dos <i>top</i> 10 preditores da ANN.....	50

LISTA DE TABELAS

Tabela 1 - Faixa de frequências dos instrumentos.	26
Tabela 2 - Matriz de confusão.....	28
Tabela 3 - Quantidade de amostras para cada instrumento	35
Tabela 4 - Quantidade de amostras para cada instrumento para base de treino e de teste.	40
Tabela 5 - Representação numérica e vetorial das classes.	41
Tabela 6 - Métricas resultantes do SVC.....	44
Tabela 7 - Métricas resultantes da RF.....	46
Tabela 8 - Métricas resultantes da ANN.	48
Tabela 9 - Acurácia de cada modelo projetado.	51
Tabela 10- Resumo das métricas de cada instrumento para cada classificador.....	51

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
STFT	Short Term Fourier Transform
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

N	Quantidade de amostras em um áudio
$w[k]$	Função de janelamento
f	Frequência
n	Quantidade de classes
Φ	Função de Ativação
W_n	Peso
x_n	Entrada
b_n	Viés
y_n	Saída

SUMÁRIO

	INTRODUÇÃO	14
1	A FÍSICA DA MÚSICA	17
1.1	Onda sonora	17
1.2	Instrumentos	19
1.2.1	<u>Guitarra elétrica</u>	21
1.2.2	<u>Violino</u>	22
1.2.3	<u>Piano</u>	23
1.2.4	<u>Flauta</u>	24
1.2.5	<u>Trompete</u>	25
1.2.6	<u>Saxofone</u>	25
1.2.7	<u>Considerações</u>	26
2	APRENDIZADO DE MÁQUINA	27
2.1	Fundamentação teórica do aprendizado de máquina	27
2.2	Aprendizado supervisionado	28
2.2.1	<u>Máquinas de vetores de suporte</u>	29
2.2.2	<u>Floresta Aleatória</u>	30
2.2.3	<u>Redes Neurais Artificiais</u>	31
3	PROJETO	34
3.1	Base de dados	35
3.2	Extração de informações	36
3.2.1	<u>Valor quadrático médio</u>	36
3.2.2	<u>Centróide espectral</u>	37
3.2.3	<u>Largura de banda espectral</u>	37
3.2.4	<u>Frequência de Rolloff</u>	38
3.2.5	<u>Zero Crossing Rate</u>	39
3.2.6	<u>Coefficientes Cepstrais de Frequência Mel</u>	39
3.3	Preparação da base	40
3.4	Classificadores	41

3.4.1	<u>Projeto do SVM</u>	42
3.4.2	<u>Projeto da RF</u>	42
3.4.3	<u>Projeto da ANN</u>	42
4	RESULTADOS	44
4.1	Resultado do SVC projetado	44
4.2	Resultado da RF projetada	46
4.3	Resultado da ANN projetada	48
	CONCLUSÃO	51
	REFERÊNCIAS	53

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, dentre os quais celebrar a caça, realizar rituais de agradecimento, aplacar a fúria ou fazer pedidos aos deuses [5].

A música é um tipo de arte que trabalha com a harmonia entre os sons, com o ritmo, com a melodia, com 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 [6]. 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 características específicas de sinais de áudio, como a largura de banda de frequência, o centróide espectral, os coeficientes cepstrais de frequência-Mel, entre outras, extraídas através de algoritmos próprios, para posterior aplicação em modelos [7].

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 a 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

A seguir, a Figura 1 apresenta as formas de ondas resultantes de quatro instrumentos diferentes – flauta, saxofone, trompete e violino – tocando uma única nota em uma mesma frequência fundamental, bem como o resultante da soma desses sinais. Nela, é possível observar como as sobreposições de frequências, relacionadas ao timbre, afetam e diferenciam os sinais de cada instrumento.

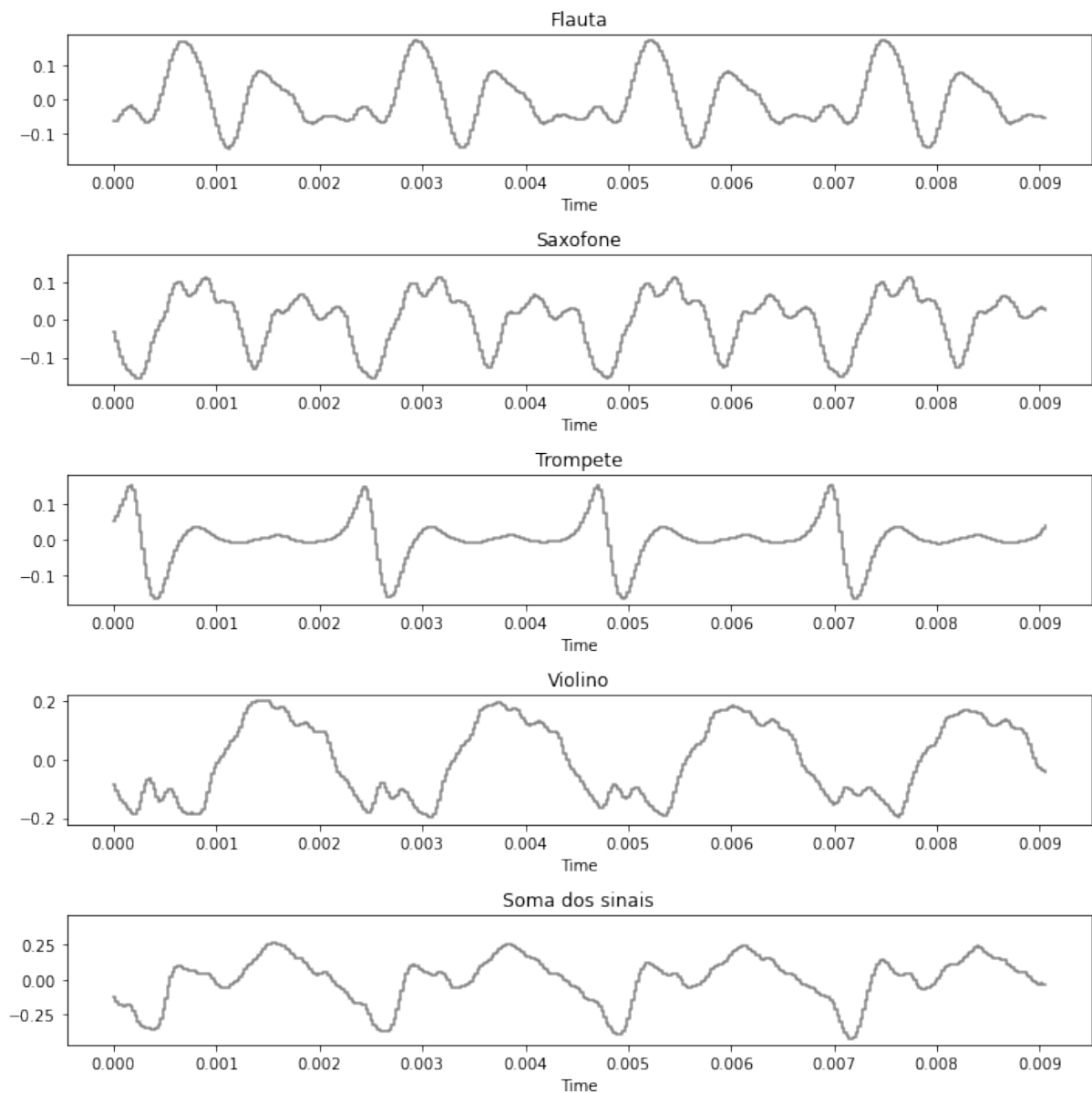


Figura 1: - Formas de onda de diferentes instrumentos para uma mesma nota musical.

Ainda, uma outra motivação para que este estudo fosse realizado se refere ao fato de que 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 [7].

Organização do Texto

Após este capítulo introdutório, reserva-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, é apresentada a metodologia utilizada, abrangendo desde a obtenção dos dados até a aplicação dos modelos de aprendizado de máquina selecionados.

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

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

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

O sinal de áudio é um sinal correspondente aos sons, e a música é a arte de combinar os sons. O sinal de áudio resulta de 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 [8] [9].

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 [10]. A onda sonora também dispõe de uma propriedade chamada timbre, que diferencia sons distintos possuidores de uma mesma frequência e amplitude.

É interessante observar que os limiares de frequência da audibilidade humana são de 20 Hz até aproximadamente 20 kHz. Os sinais que possuem frequências fora dessa faixa, chamados de infrassom e ultrassom, respectivamente, não são possíveis de ser ouvidos. A Figura 2 exemplifica essas faixas de frequência.

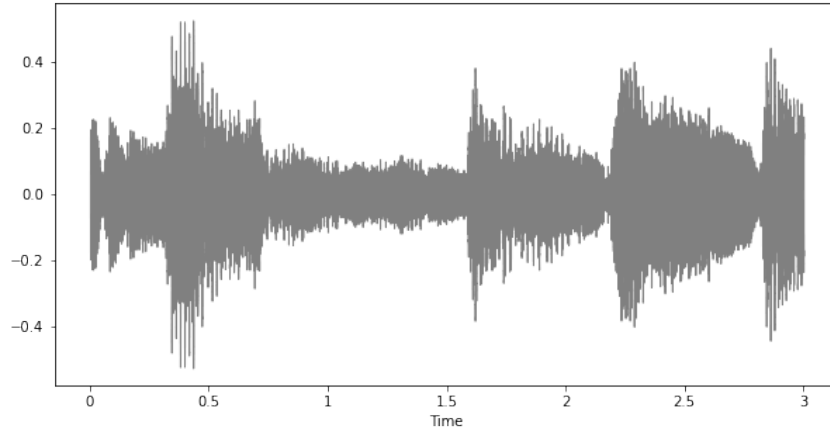
Faixa de frequência auditiva humana		
Infrassom	Som audível	Ultrassom
	20 Hz	20 kHz

Figura 2: - Faixas de frequência.

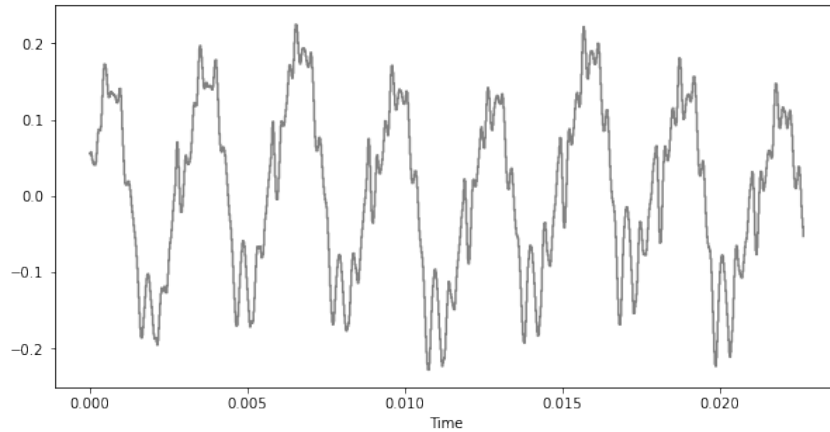
Observa-se ainda que, como o som é uma forma de energia, ele não pode simplesmente deixar de existir; portanto a explicação do decaimento dos sons se dá pela absorção deles pelas superfícies dos objetos no espaço - que podem incluir móveis, pessoas e ar, transformando a energia em calor [11].

A Figura 3a apresenta graficamente a amplitude em relação ao tempo de uma amostra de sinal de áudio de 3 segundos de duração, e a Figura 3b exibe a mesma amos-

tra, porém com uma duração de aproximadamente 0,025 segundos para a exemplificação da característica senoidal e **estacionária** - esse gráfico é de um sinal de música, não é estacionária.



(a) Amostra de 3 segundos.



(b) Amostra de 0,025 segundos.

Figura 3: - Representação no domínio do tempo.

O espectrograma é uma representação visual do sinal sonoro, tanto no domínio da frequência, como no do tempo. Para criá-lo, é necessário converter as amostras em janelas no domínio do tempo individuais para o domínio da frequência, utilizando a Transformada Rápida de Fourier (FFT), definida como

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

No caso da onda não estacionária, deve-se aplicar a FFT em pequenas janelas de tempo - curtas o suficiente para que não haja grandes variações estatísticas do sinal-

utilizando a Transformada de Fourier de curto termo (STFT), definida por

$$X[t, n] = \sum_{k=0}^{N-1} w[k]x[tN + k]e^{-j\frac{2kn\pi}{N}} \quad (1.2)$$

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 por [12]

$$Mel(f) = \frac{1000}{\log 2} \log \left(1 + \frac{f}{700} \right) \quad (1.3)$$

Ambos os tipos de espectrogramas indicam o comportamento espectral ao longo do tempo. A Figura 4 mostra esses dois tipos de representação a partir da mesma amostra de áudio mostrada na Figura 3.

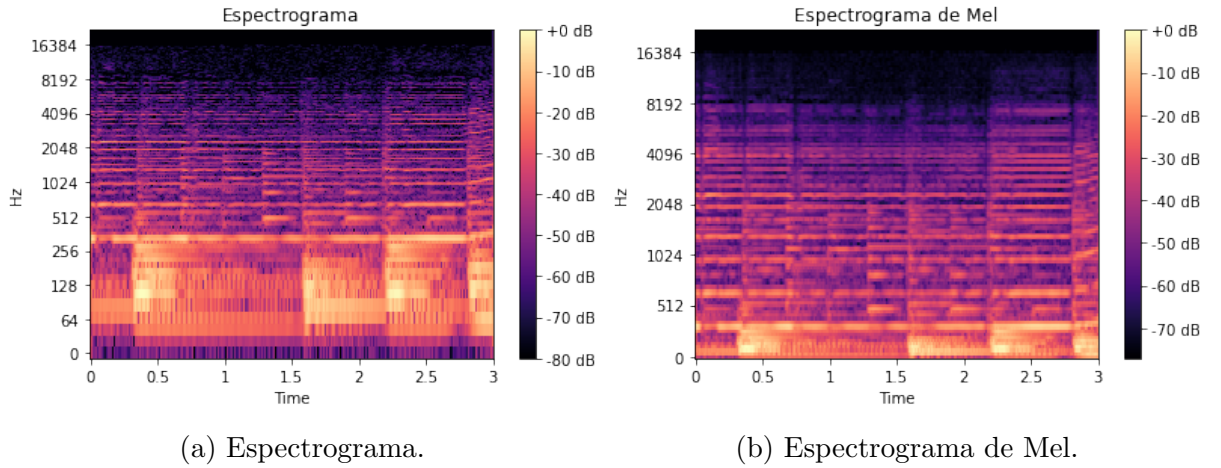


Figura 4: - Representação no domínio do tempo e da frequência.

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 [8] [13]. Alguns dos fatores responsáveis pela diferenciação do timbre em um mesmo instrumento são: material de construção (tipo de madeira ou metal), material das cordas, espessura delas, entre outros.

A Figura 5 apresenta uma simulação realizada através do *PhET: Interactive Simulations* [14]. Nela, são somadas as ondas da frequência fundamental e os próximos três harmônicos de mesma amplitude. Ou seja, toca-se uma nota Lá (A) em sua frequência fundamental de aproximadamente 438 Hz e, em seguida, adicionam-se os seus harmônicos, que representam multiplicações desse valor, sendo eles 876 Hz, 1752 Hz e 3504 Hz, respectivamente. É importante observar que todas essas diferentes frequências ainda representam a nota Lá. Dessa forma, é possível observar como os harmônicos deformam o sinal.

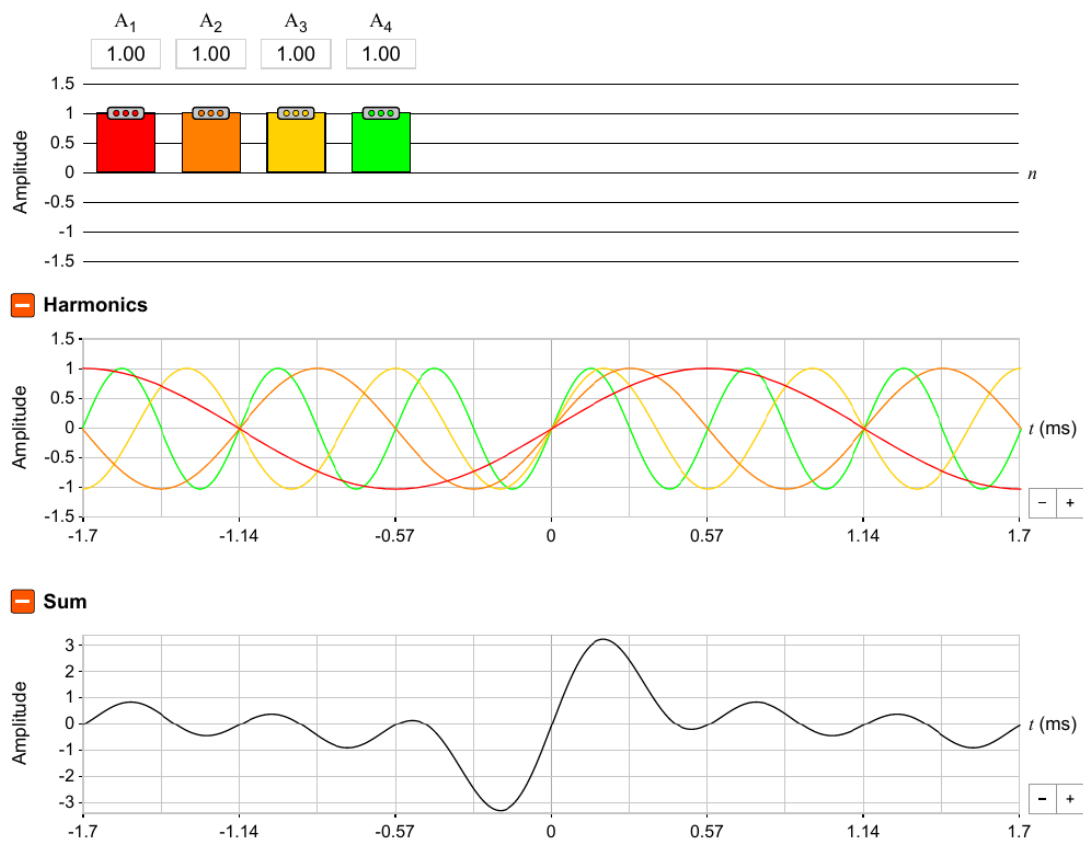


Figura 5: - Efeito dos harmônicos em um sinal senoidal.

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 [13]. 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.

A Figura 6 exemplifica, graficamente, como a amplitude varia em cada etapa do ADSR ao longo do tempo.

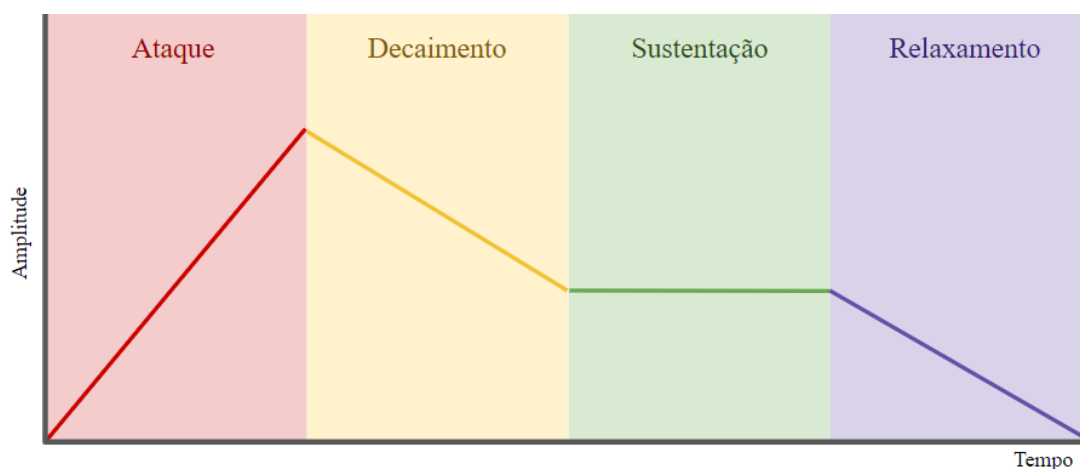


Figura 6: - Evolução do sinal ao longo do tempo.

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 são compostas 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 provocam uma mudança no fluxo magnético através da bobina do ímã permanente do captador, induzindo um sinal elétrico. Para o captador do braço, a amplitude da frequência fundamental é mais favorecida, enquanto para o captador da ponte (mais distante do braço), os harmônicos são mais presentes [9].

Define-se a frequência do sinal pelo tamanho e pela tensão das cordas pressionadas. Por ser um instrumento que depende de equipamentos eletrônicos, a intensidade do som e o timbre não dependem totalmente da guitarra.

Além disso, as guitarras elétricas são suscetíveis à captação da frequência de 60 Hz da linha de energia de corrente alternada, que se torna um ruído para esse instrumento [9].

A Figura 7 mostra as partes de uma guitarra, detalhando aquelas citadas anteriormente.

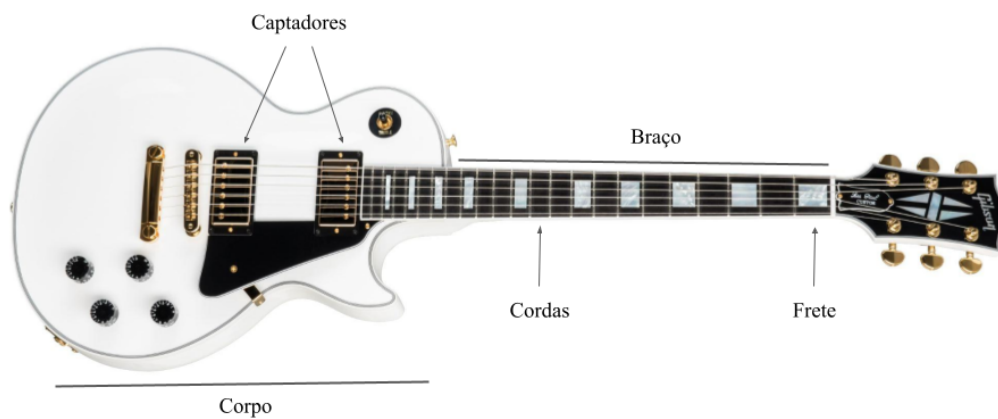


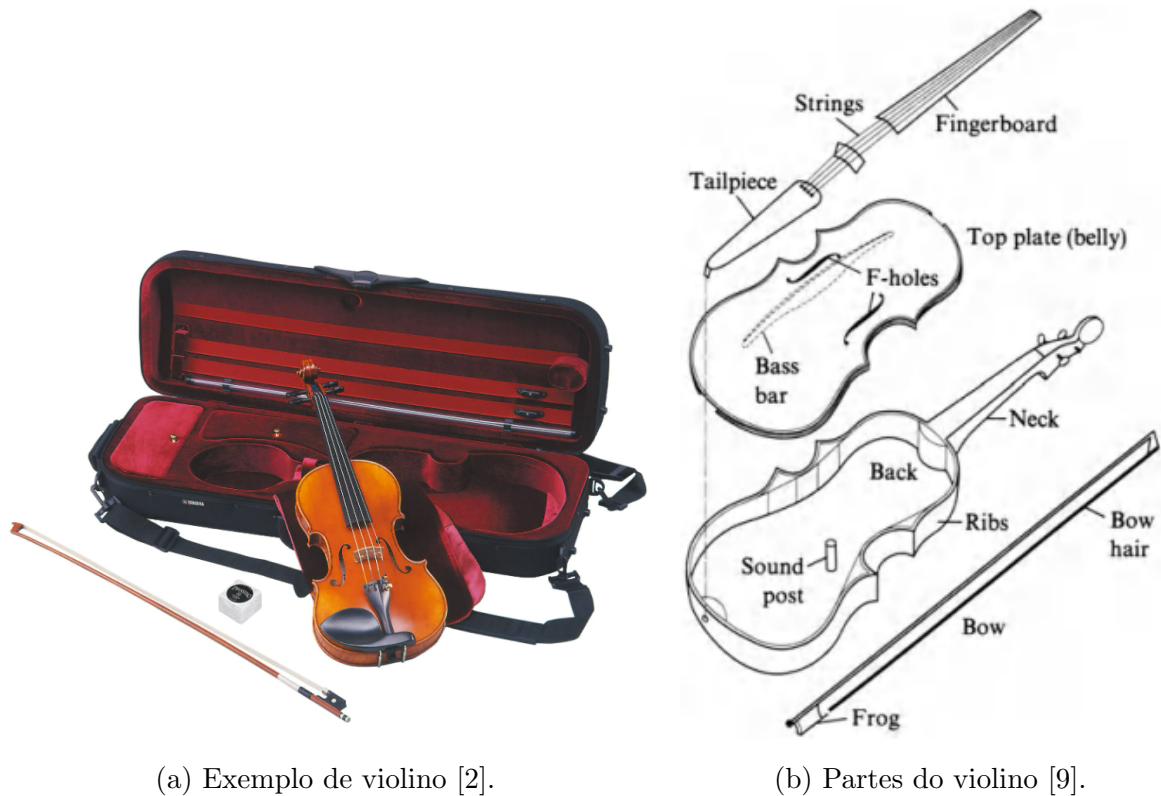
Figura 7: - Partes da guitarra - adaptação de *Gibson* [1].

1.2.2 Violino

O violino acústico é composto, especialmente, por um corpo oco de madeira com aberturas, cordas e um arco de madeira, além de 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, que é amplificada 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 [9].

A Figura 8 mostra um violino bem como as suas partes de forma detalhada.



(a) Exemplo de violino [2].

(b) Partes do violino [9].

Figura 8: - Violino.

1.2.3 Piano

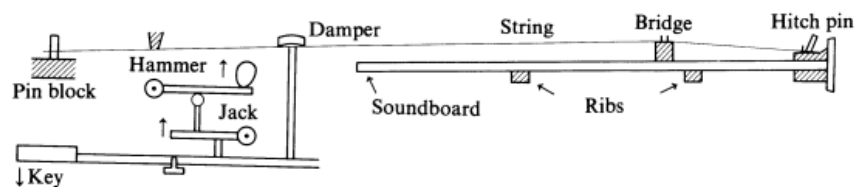
O piano é composto principalmente por teclas, martelos, cordas, pedais e uma caixa de ressonância. Nele, o som é produzido da seguinte forma: quando uma tecla é pressionada, ela ativa o martelo, que toca nas cordas referentes à tecla pressionada e, então, o som é amplificado pela sua caixa de ressonância. 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 Figura 9b esquematiza as partes do piano que fazem parte desse processo.

Determina-se a intensidade do som pela força aplicada ao pressionar as teclas, enquanto a frequência é definida a partir da afinação das cordas - levando em conta a tensão e o tamanho delas.

Já o timbre é dominado por sons transientes, que caracterizam o ataque, que, no caso do piano, inclui sons mecânicos - o martelo batendo nas cordas [9].



(a) Exemplo piano [2].



(b) Esquematização do piano [9].

Figura 9: - Piano.

1.2.4 Flauta

A flauta transversal, assim como a representada pela Figura 10, é constituída de um corpo oco, com buracos para os dedos ao longo do seu comprimento, além de uma abertura para entrada do sopro e de uma outra no ponto mais distante do sopro. Cada combinação de buracos fechados representa uma nota.

Na flauta, o som é produzido através de um jato de ar dentro do seu corpo oco. A velocidade (pressão) do jato define a intensidade do áudio, e a direção desse jato dentro do instrumento, junto com o tamanho do seu corpo - definido pelo fechamento dos seus buracos - determina a frequência da nota produzida.

A característica temporal do sinal também depende do suprimento de ar, como o ataque, que se sujeita à pressão do sopro, podendo ser abrupto, gradual e até plosivo [9].



Figura 10: - Flauta transversal [2].

1.2.5 Trompete

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

A Figura 11 mostra um trompete e as suas partes citadas, responsáveis pela produção do som.

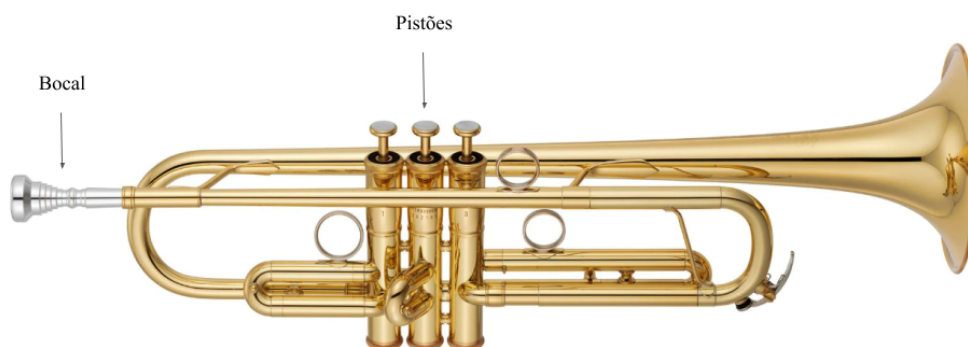


Figura 11: - Trompete, adaptado de *Yamaha* [2].

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 da vibração labial do trompetista [9].

1.2.6 Saxofone

O saxofone é constituído por um tubo metálico curvado, com buracos para os dedos - mecânica semelhante à da flauta – e por uma palheta de madeira na boquilha, como mostra a Figura 12. O seu som é produzido a partir da vibração dessa palheta, resultante da coluna de ar gerada pelo sopro. Curiosamente, essa forma de produção de som, dependente de um elemento composto de madeira (a palheta), determina que o saxofone seja classificado como um instrumento da família das madeiras.



Figura 12: - Saxofone, adaptado de *Yamaha* [2].

A intensidade do sinal estabelece-se pela força do sopro, e a frequência define-se pela frequência da 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, ou seja, todos os múltiplos inteiros da frequência fundamental afetam a onda sonora, até que, no decorrer do tempo, suas amplitudes decaiam e, por fim, desapareçam [9].

1.2.7 Considerações

A Tabela 1 resume a faixa de frequência que cada instrumento citado abrange, incluindo as frequências dos seus maiores harmônicos [15].

Tabela 1: - Faixa de frequências dos instrumentos.

Instrumento	Frequência mínima (Hz)	Frequência máxima (Hz)	Frequência do maior harmônico (Hz)
Guitarra	82,4	1318,5	5274,04
Violino	196	3136	15804,26
Piano	27,5	4186	10549,08
Flauta	261,63	2349,3	1175,3
Trompete	164,81	1046,5	9397,27
Saxofone	110	880	8372,02

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 [16]:

- **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. (Está confuso. A função erro não tem a ver com a função objetivo do processo de otimização? Variação de que?)
- **Processo de otimização do modelo:** métodos que levam em consideração a minimização do erro durante o processo de aprendizado do modelo. (Que tipos de algoritmos de otimização são empregados?)

O aprendizado de máquina, geralmente, pode ser classificado de quatro formas diferentes, de acordo com a sua forma de aprendizado. São elas [17]:

- **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 (?? tente exemplificar...) as hipóteses criadas apenas com os dados rotulados [18].
- **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 (?? tente exemplificar...), 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) [19].

Para a realização do modelo de classificação proposto neste projeto, foram escolhidos três algoritmos de classificação com aprendizado supervisionado: **Máquina de Vetores de Suporte (SVM – Support Vector Machine)**, **Floresta Aleatória (RF – Random Forest)** e **Redes Neurais Artificiais (ANN – Artificial Neural Networks)**.

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 2 mostra como os dados são classificados em relação às suas respectivas predições [20].

Tabela 2: - 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.

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

A sensibilidade, como mostra a equação 2.2, representa a fração de valores da classe

positiva que foram corretamente classificados.

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

Por fim, a precisão, [como mostra a 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.

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

2.2.1 Máquinas de vetores de suporte

O algoritmo de aprendizado supervisionado 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 13 [ilustra](#) o processo de decisão do SVM em um hiperplano com apenas dois preditores. [\(Incluir mais detalhes do passo-a-passo do SVM por meio ou de diagrama de blocos, ou enumeração dos passos, ou psuedo-código.\)](#)

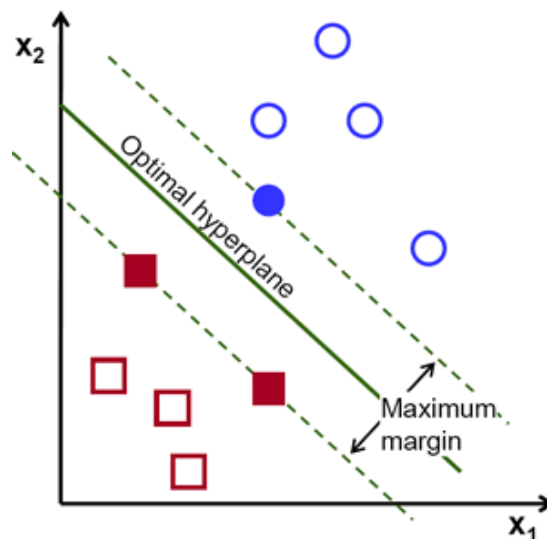


Figura 13: - Esquematização do SVM [3]

Tal método possui um bom desempenho quando existe pouca quantidade de amos-

tras para cada classe, porém **desempenha** mal quando há muitos dados - já que exige muita capacidade computacional - e quando existem muitos *outliers* [21].

Um único classificador SVM não consegue realizar uma classificação de múltiplas classes, mas, sim, “**quebrar**” os dados em apenas duas classes. (Não entendi, reescrever.) Para realizar múltiplas classificações binárias, utiliza-se mais de um modelo de classificação, o que é chamado de *one vs one* [22]. (Não entendi, reescrever.)

A quantidade de vetores de suporte necessários é determinada a partir da quantidade de rótulos, da seguinte forma

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

onde n é (completar...)

2.2.2 Floresta Aleatória

O algoritmo de aprendizado de máquina RF é constituído por um conjunto classificadores de árvores de decisão, que recebem como entrada vetores aleatórios independentes e identicamente distribuídos [23].

As árvores de decisão (AD) são compostas por nós, ramos e folhas, que representam o percurso que os dados de entrada **realizam** para **fins de** 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 [24]. 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 14 mostra como esse processo é realizado. (Explicar com mais detalhes a figura para melhor entendimento do RF, por favor.)

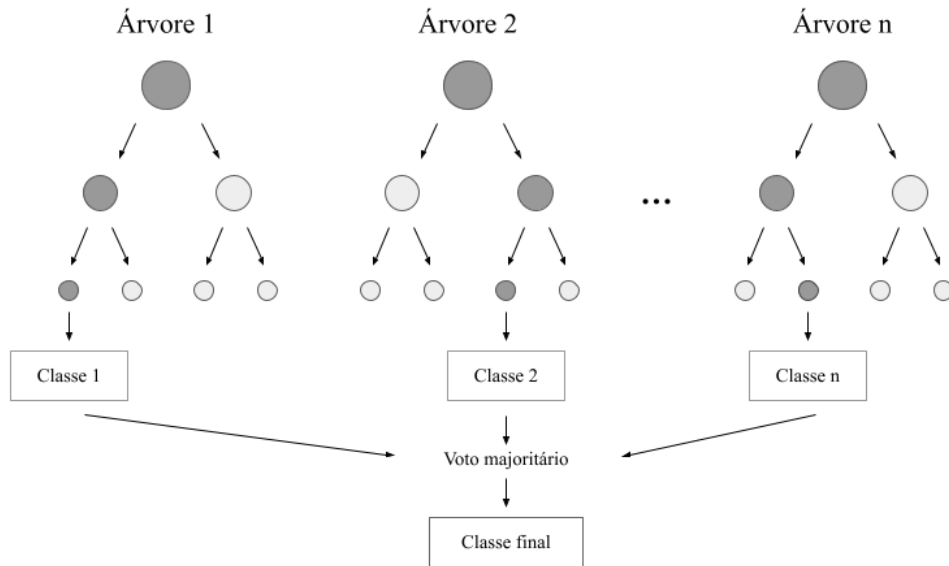


Figura 14: - Esquemática da floresta aleatória.

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 saída de um neurônio é dada pela seguinte equação

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

onde n representa o neurônio, $\Phi(\cdot)$ a função de ativação, W_n o peso, x_n a entrada e b_n o viés.

A função de ativação transforma 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. (Que tipo de tarefas complexas?)

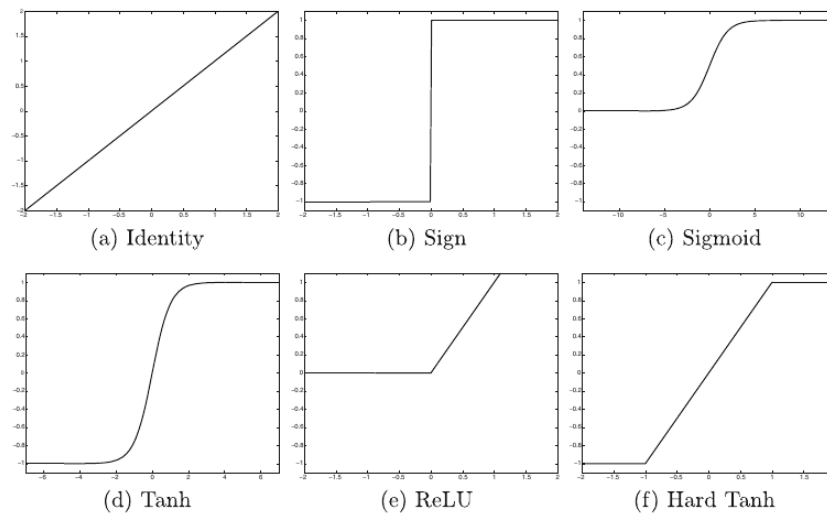


Figura 15: - Funções de ativação [4] (Referenciar e exeplicar figura no texto.)

Além das funções de ativação representadas na Figura 15, 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 Figura 16 mostra a esquematização de uma rede neural. (Apresentar mais detalhes/informações a respeito da figura.)

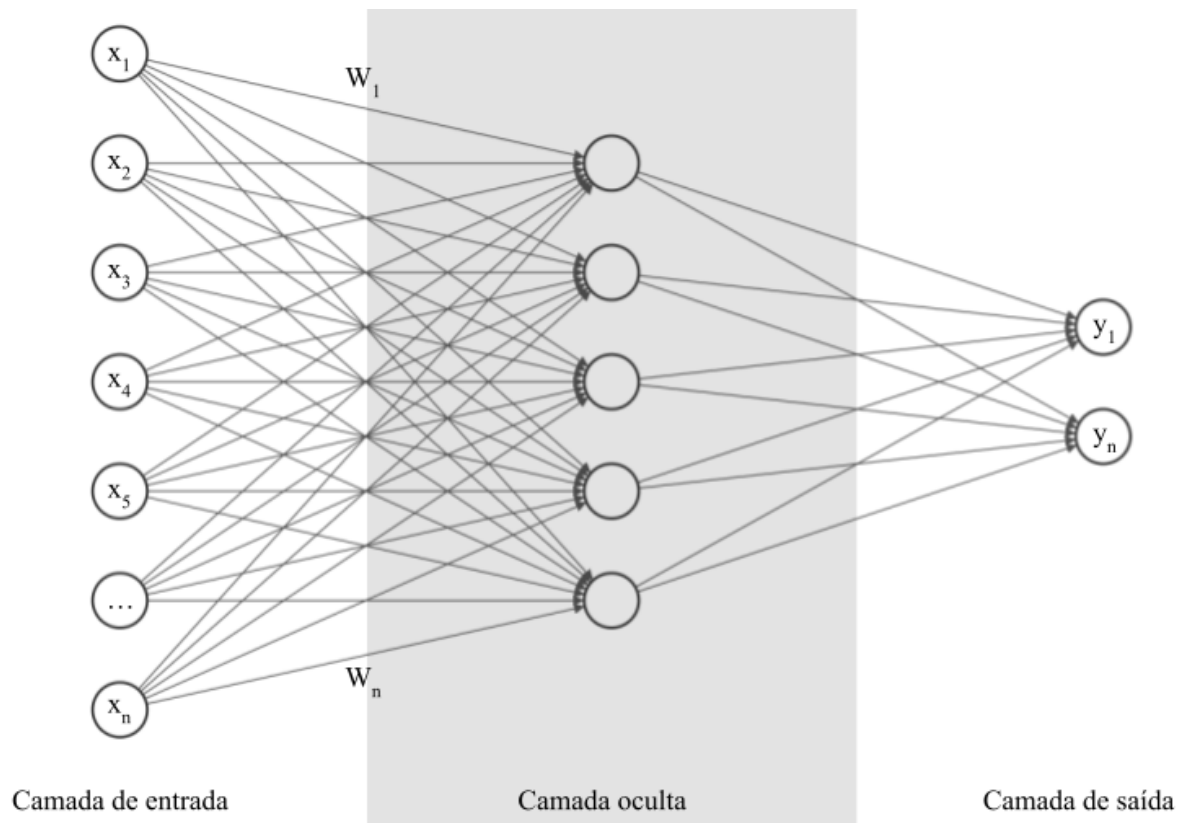


Figura 16: - Esquematização da rede neural

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 17 esquematiza essa proposta.

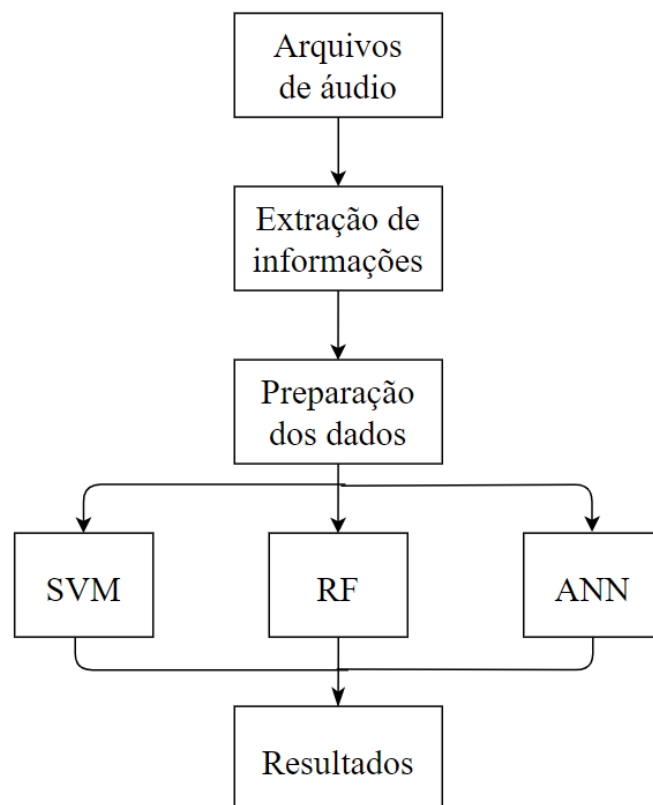


Figura 17: - Proposta de projeto

3.1 Base de dados

Para este projeto, escolheu-se a base de dados IRMAS [25], frequentemente utilizada em estudos de reconhecimento de instrumentos musicais [6].

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 18 mostra um exemplo esquematizado:

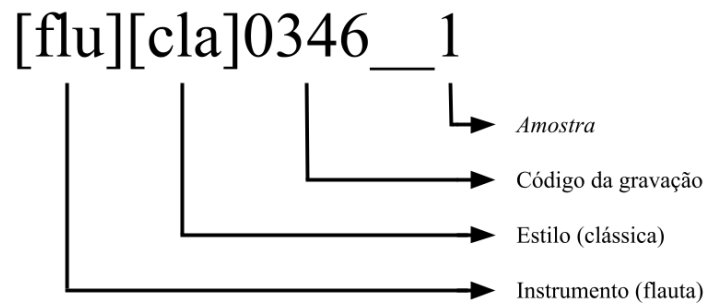


Figura 18: - Exemplo de nome de arquivo da base de dados

A seguir, a Tabela 3 mostra a distribuição de amostras de áudios para cada instrumento, junto com as siglas que os representam.

Tabela 3: - 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* [26] através da biblioteca *librosa* [27] 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 [7].

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. [28].

A Figura 19 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.

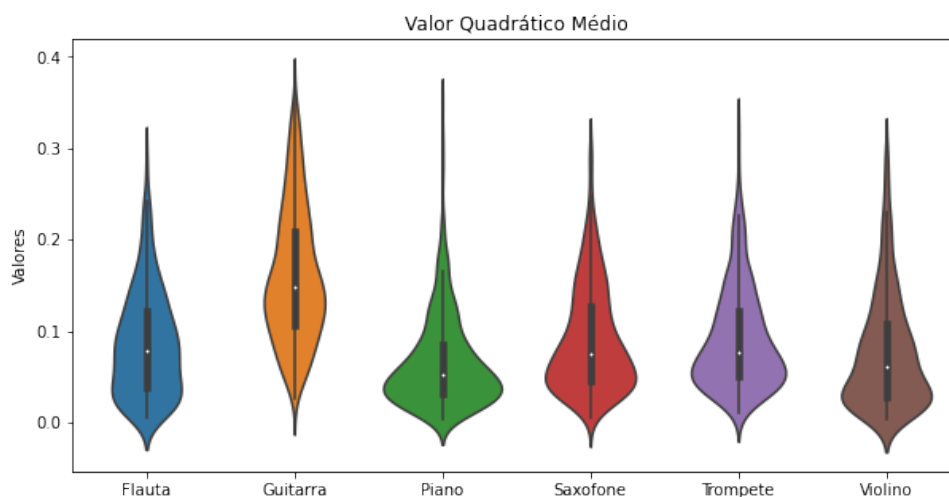


Figura 19: - Distribuição do RMS

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 [7].

Na distribuição segmentada da Figura 20, 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.

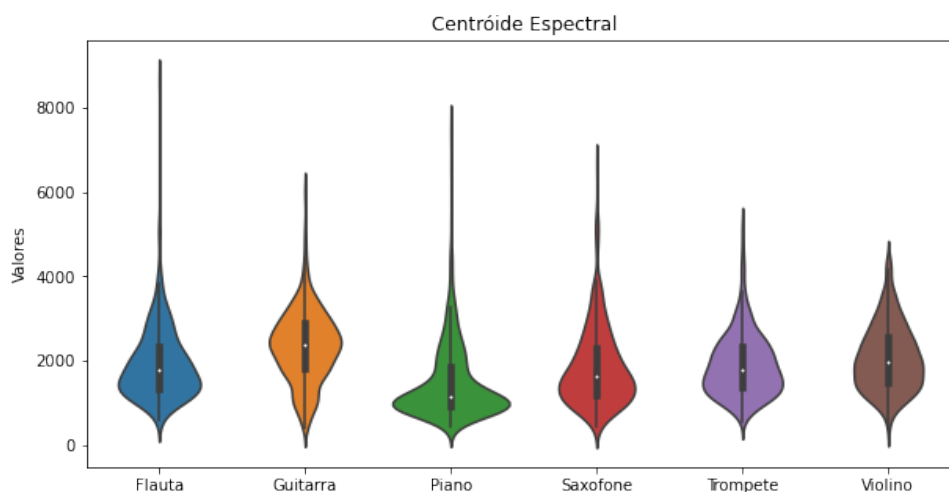


Figura 20: - Distribuição do SC

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 [28] [7].

Na Figura 21, 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.

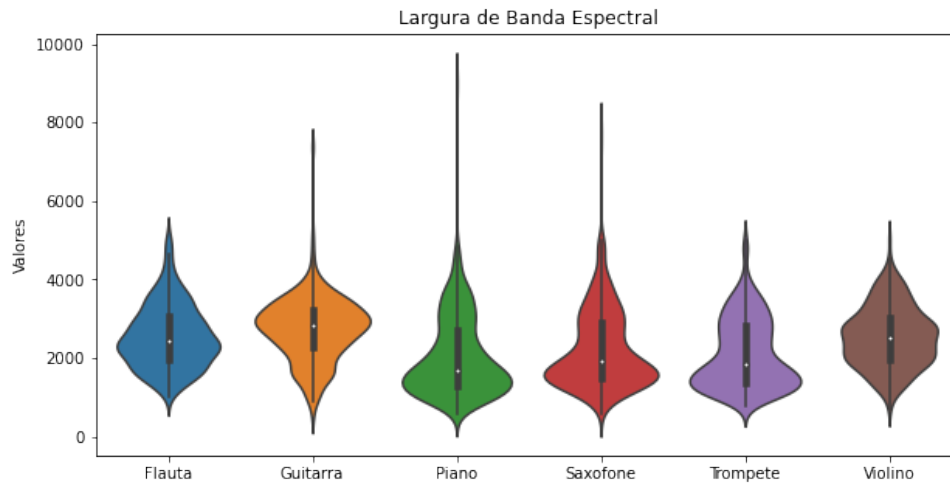


Figura 21: - Distribuição da SB

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 [27].

A distribuição da Figura 22 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.

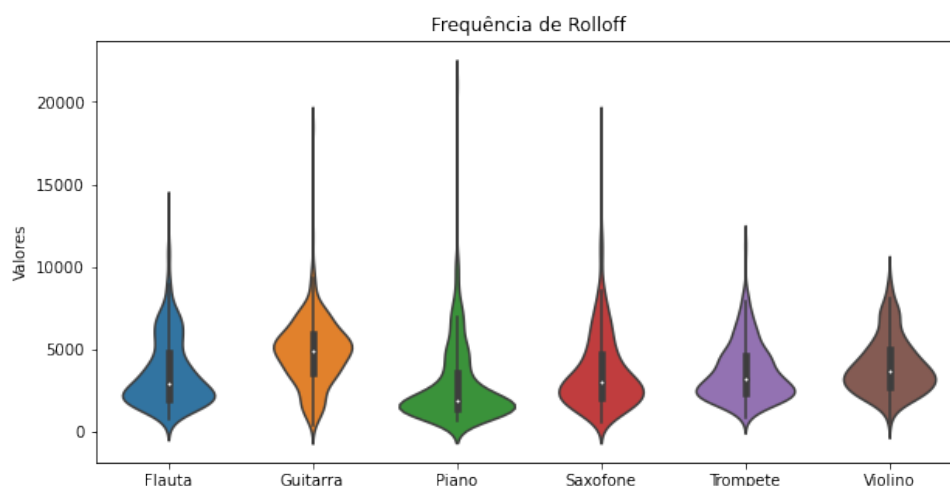


Figura 22: - Distribuição da frequência de *rolloff*

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. [28].

A distribuição segmentada por instrumento da Figura 23 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.

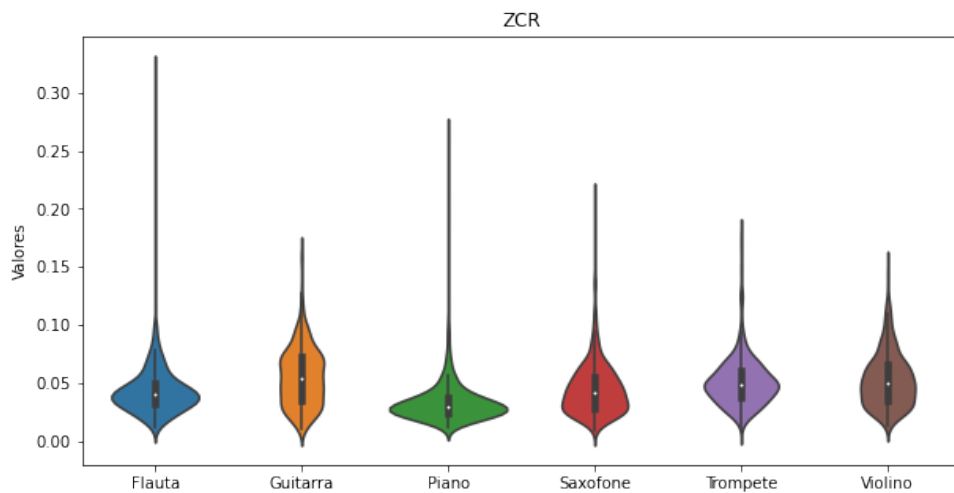


Figura 23: - Distribuição do ZCR

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 [28] [29].

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

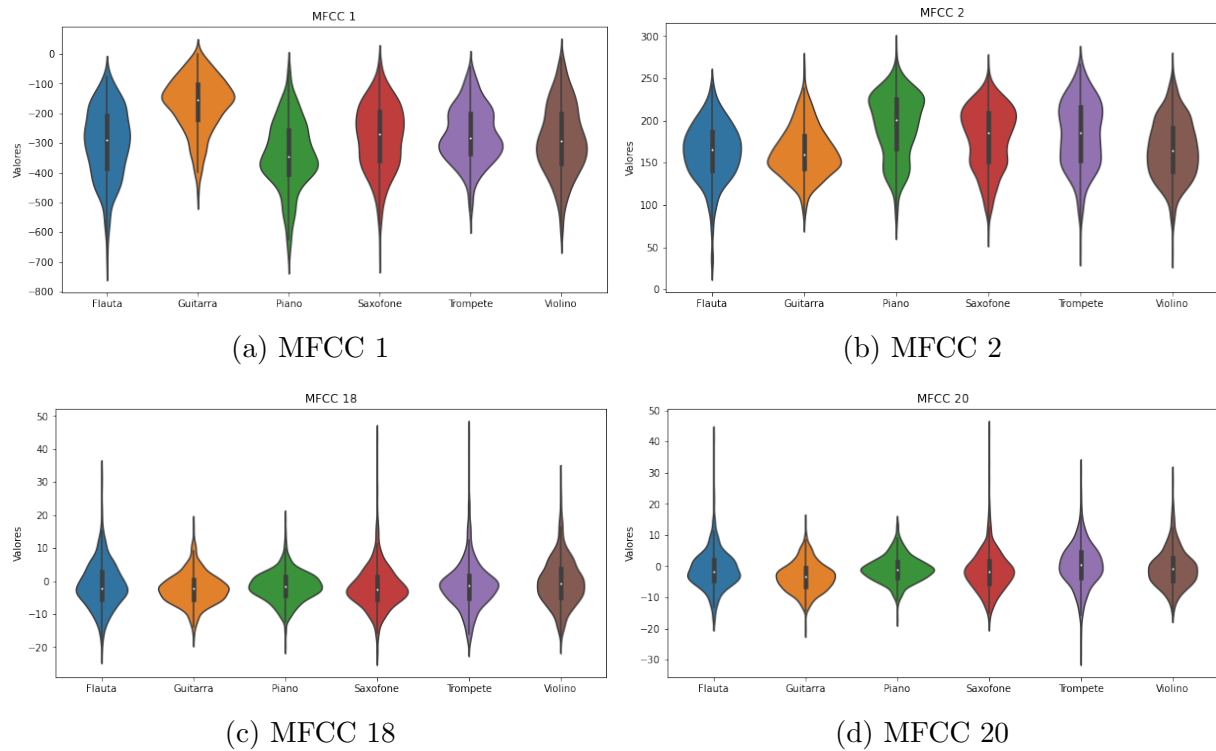


Figura 24: - Coeficientes Cepstrais de Mel

3.3 Preparação da base

Os dados foram preparados para treino utilizando os pacotes do *python*: *scikit-learn* [30] e *numpy* [31].

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 4 mostra a quantidade de dados para cada instrumento na base de treino e de teste.

Tabela 4: - 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* [30], 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* [30] e o *to_categorical* [31] para realizar essa separação. Dessa forma, as classes foram representadas como mostra a Tabela 5.

Tabela 5: - 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* [30].

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* [32] 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*.

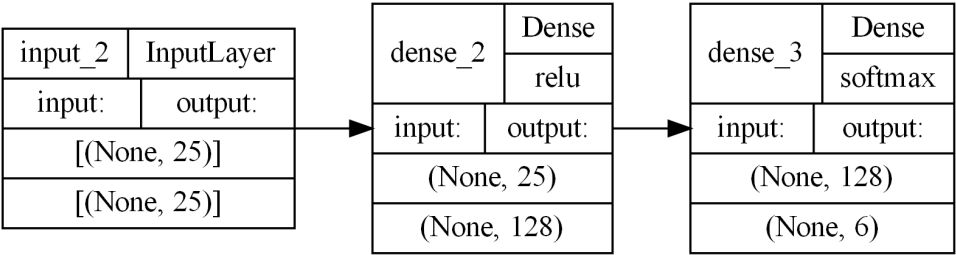


Figura 25: - ANN projetada.

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 [33] 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 6. Observa-se, em geral, resultados muito bons para todos os instrumentos, principalmente para a guitarra, o trompete e o piano.

Tabela 6: - 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 26 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.

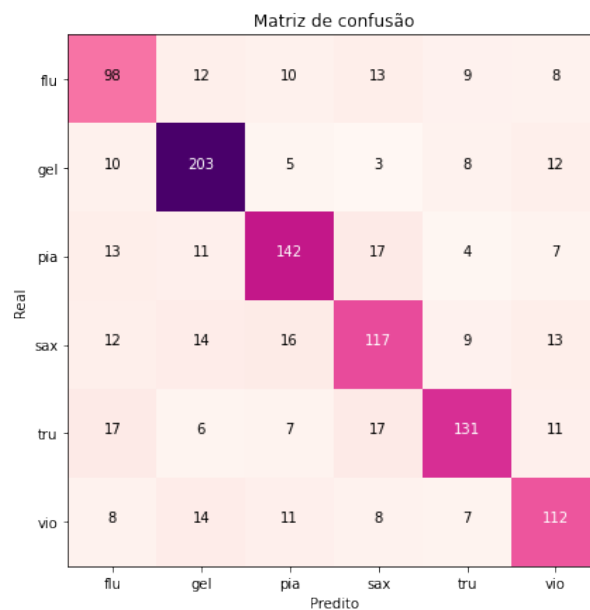
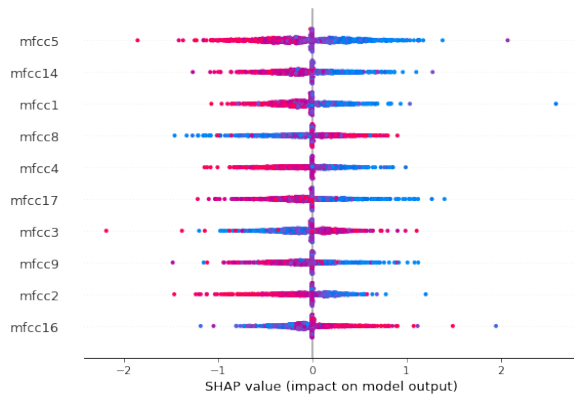
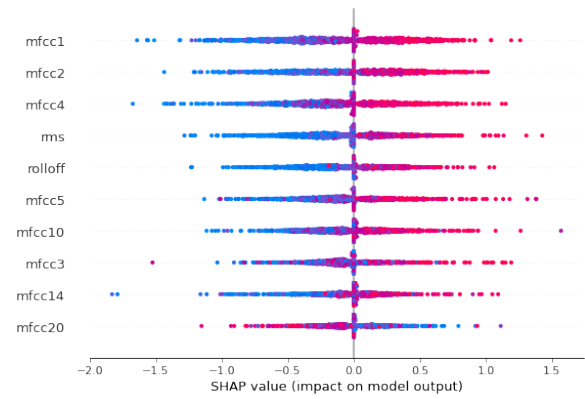


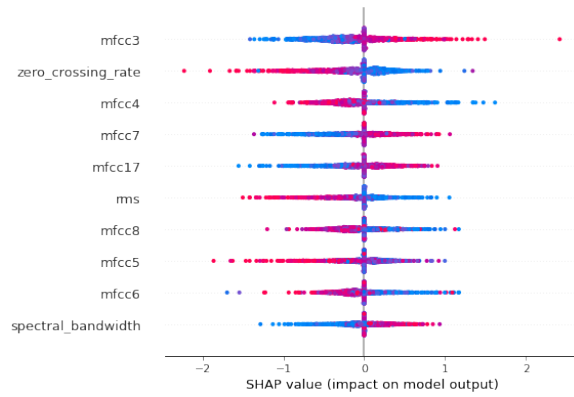
Figura 26: - Classificações SVC.



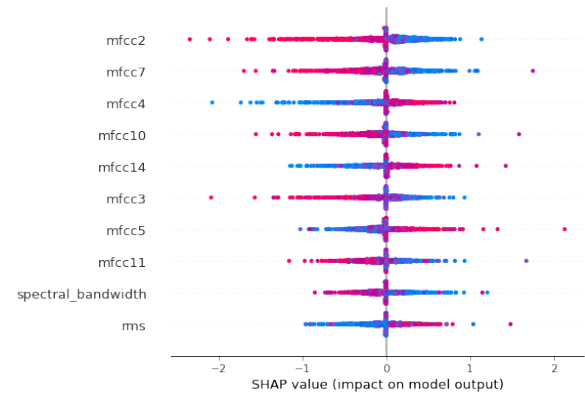
(a) Flauta



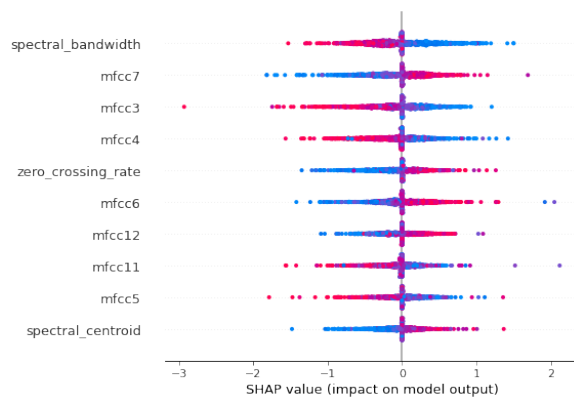
(b) Guitarra



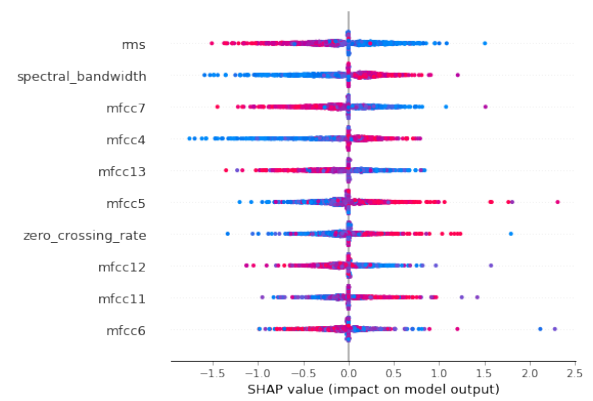
(c) Piano



(d) Saxofone



(e) Trompete



(f) Violino

Figura 27: - Influência dos *top* 10 preditores do SVC.

Os gráficos do SHAP, na Figura 27, 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 7 apresentam resultados medianos para todos os instrumentos, exceto para o caso da guitarra, que obteve métricas bem altas.

Tabela 7: - 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 28.

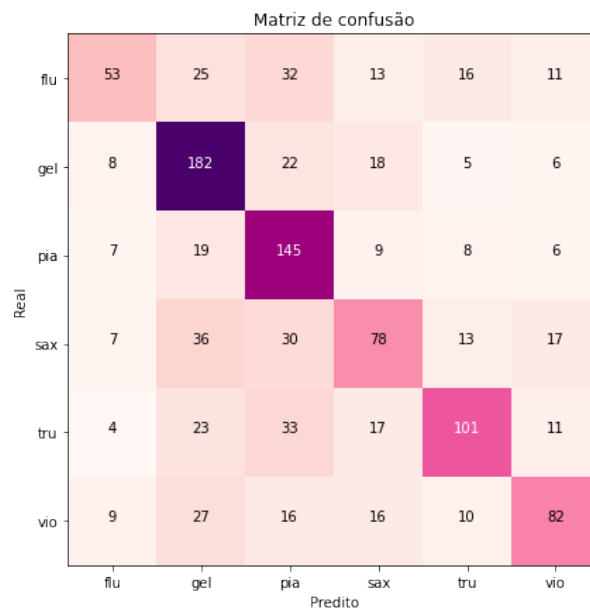
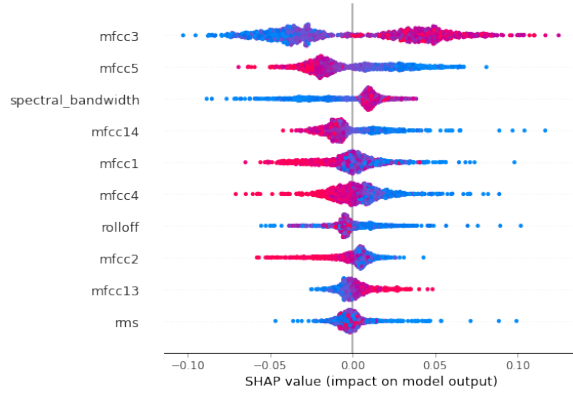
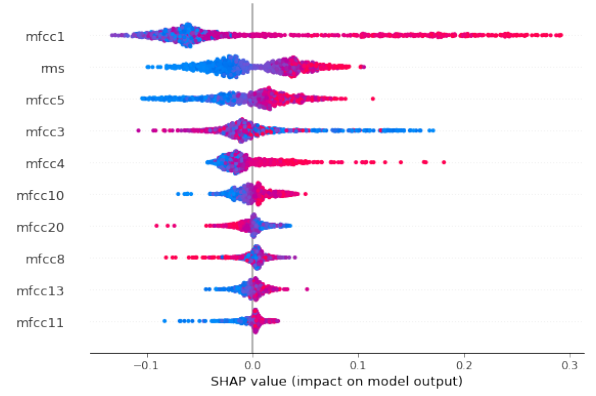


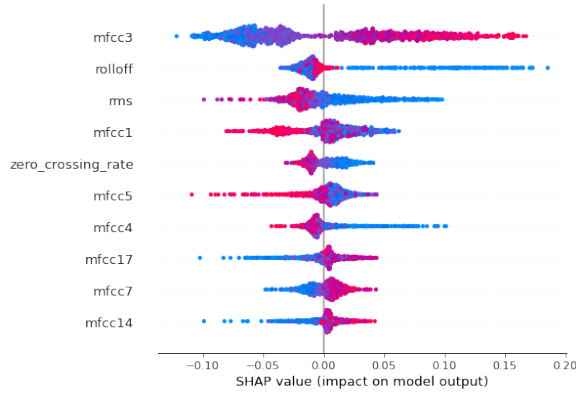
Figura 28: - Classificações RF.



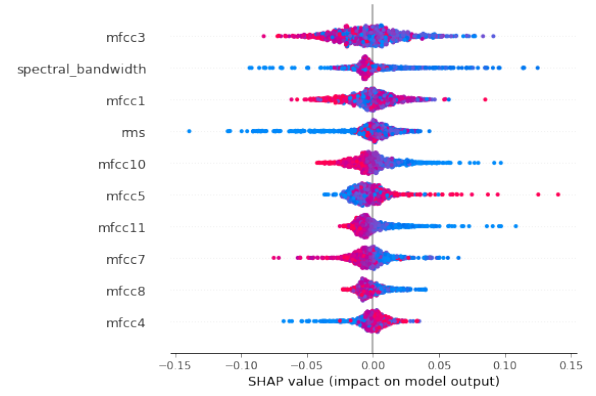
(a) Flauta



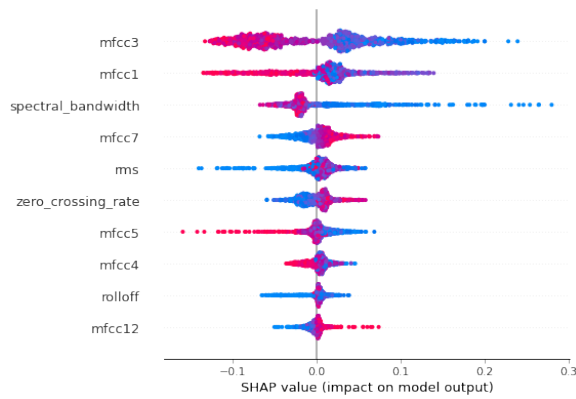
(b) Guitarra



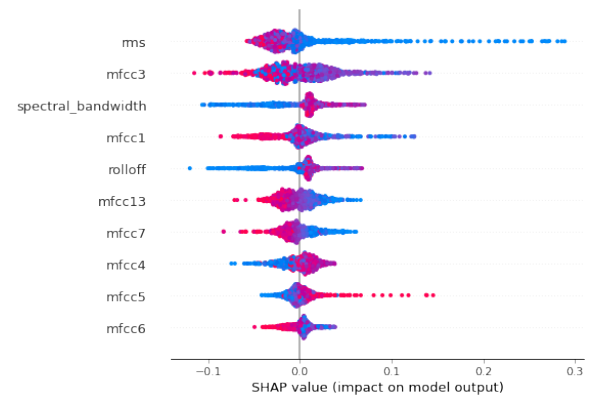
(c) Piano



(d) Saxofone



(e) Trompete



(f) Violino

Figura 29: - Influência dos *top* 10 preditores da RF.

Já o SHAP da Figura 29 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 8 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 8: - 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 30 traduz as métricas apresentadas, mostrando um alto índice de classificações corretas para a guitarra e o piano e um baixo para os outros instrumentos.

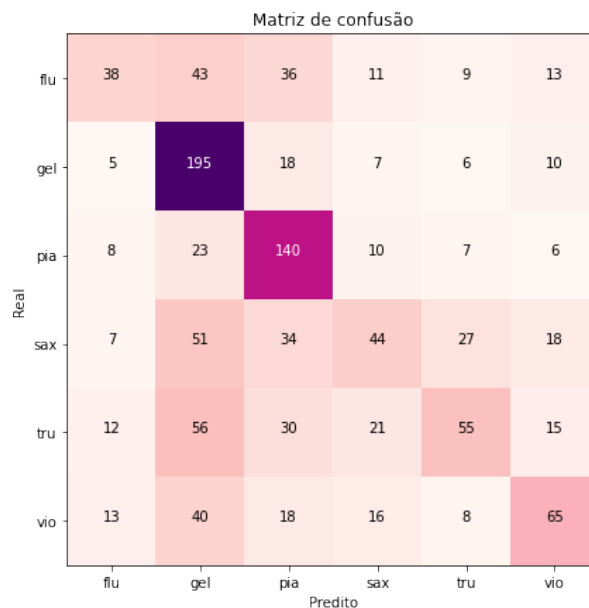


Figura 30: - Classificações ANN.

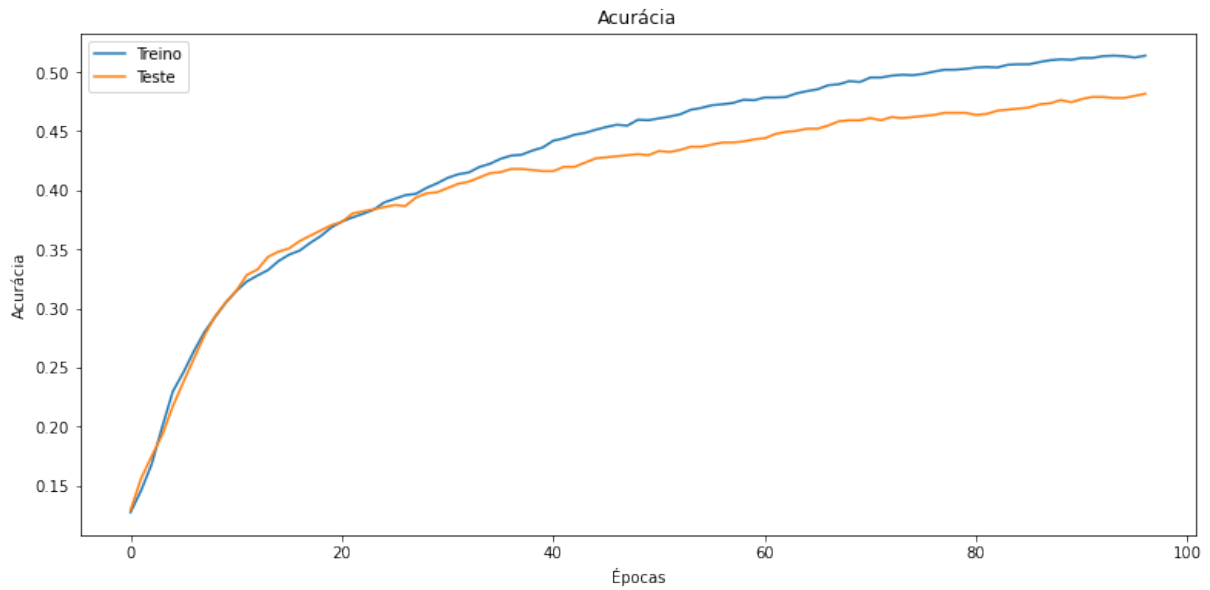
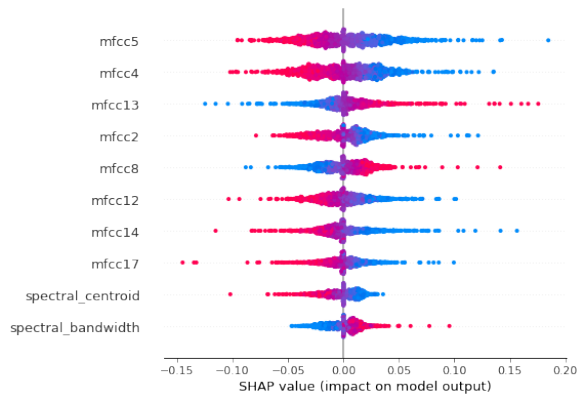


Figura 31: - Acurácia da ANN x épocas.

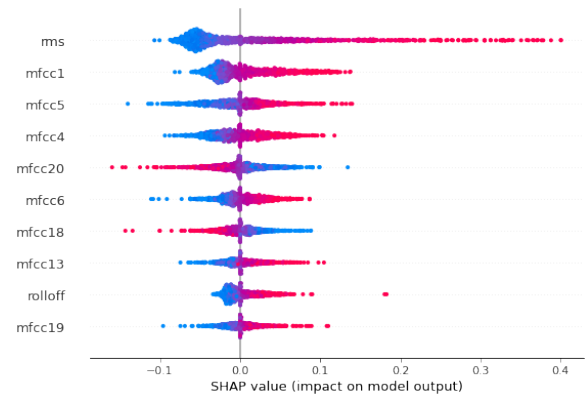
No modelo de rede neural, é possível avaliar como ocorreu o processo de aprendizado no decorrer das épocas, que indicam quantas vezes o algoritmo utilizou a base de dados inteira em seu treinamento.

Na rede projetada, como indica a Figura 31, foram necessárias 97 épocas para se chegar ao valor ótimo entre a acurácia de treino e de teste - 51% e 48%, respectivamente.

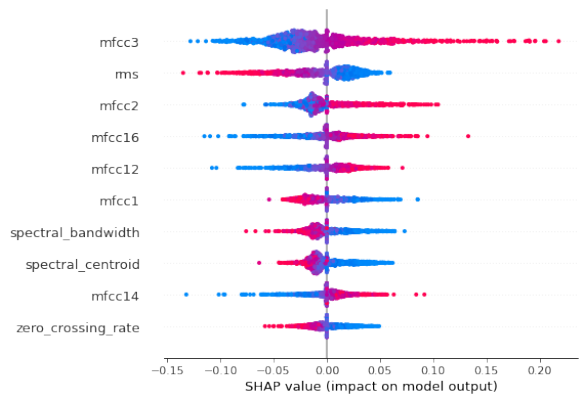
Assim como nos outros modelos, o SHAP da rede neural considerou uma influência bem alta para os preditores de MFCC, como mostra a Figura 32. Apesar de ainda haver uma diferenciação entre as cores dos valores altos e baixos dos preditores (vermelho e azul), nesse caso, elas são consideravelmente mais misturadas do que as apresentadas na Figura 27 e na Figura 29, o que indica uma dificuldade na classificação dos instrumentos, comprovada pelas métricas baixas.



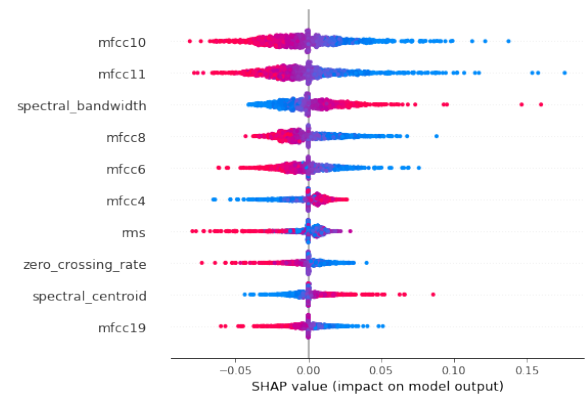
(a) Flauta



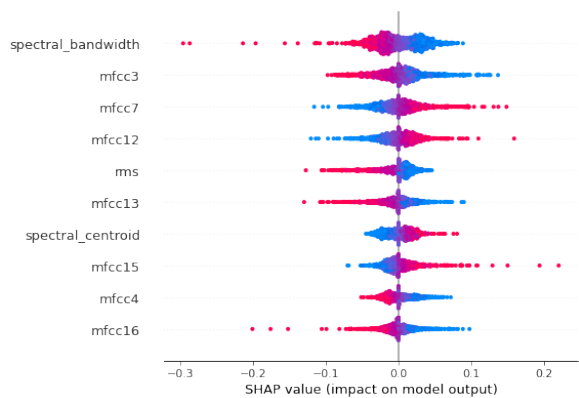
(b) Guitarra



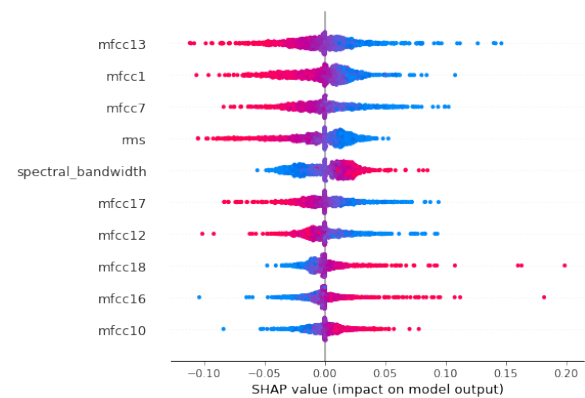
(c) Piano



(d) Saxofone



(e) Trompete



(f) Violino

Figura 32: - 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 9: - Acurácia de cada modelo projetado.

Modelo	Acurácia
SVC	72%
RF	57%
ANN	48%

Na Tabela 9, 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 10: - 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 10. 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 26, da Figura 28 e da Figura 30, 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 3.

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. *Les Paul Custom w/ Ebony Fingerboard Gloss*. Disponível em: <<https://www.gibson.com/en-US/Electric-Guitar/CUSZJG839/Alpine-White>>. 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] Laura Aydar. *História da Música*. Disponível em: <<https://www.todamateria.com.br/historia-da-musica/>>. Acesso em: 07 de maio de 2022.
- [6] GURURANI, S.; SHARMA, M.; LERCH, A. An attention mechanism for musical instrument recognition. *ArXiv*, abs/1907.04294, 2019.
- [7] 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.
- [8] FENG, J. Q. Music in terms of science. *ArXiv*, abs/1209.3767, 2012.
- [9] FLETCHER, N. H.; ROSSING, T. D. *The Physics of Musical Instruments*. 1. ed. [S.l.]: Springer New York, NY, 1991.
- [10] DOBRIAN, C. Msp: The documentation. *Cycling '74 and IRCAM*, Dezembro 1997.
- [11] LLOYD, L. S. *Music and Sound*. [S.l.]: Ayer Publishing, 1970. 169 p.

- [12] VIRTANEN, T.; PLUMBLEY, M. D.; ELLIS, D. *Computational Analysis of Sound Scenes and Events*. [S.l.]: Springer, 2018.
- [13] Multiple Contributors. *Audio Representation*. Disponível em: <https://musicinformationretrieval.com/audio_representation.html>. Acesso em: 13 de agosto de 2022.
- [14] University of Colorado Boulder. *PhET Interactive Simulations: Fourier - Making Waves*). Disponível em: <https://phet.colorado.edu/sims/html/fourier-making-waves/latest/fourier-making-waves_en.html>. Acesso em: 19 de outubro de 2022.
- [15] Rory Seydel. *EQ Cheat Sheet: How to Use An Instrument Frequency Chart*. Disponível em: <<https://blog.landr.com/eq-cheat-sheet/>>. Acesso em: 19 de outubro de 2022.
- [16] 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.
- [17] 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.
- [18] ZHU, X. *Semi-Supervised Learning Literature Survey*. [S.l.], 2005.
- [19] IBM Cloud Education. *What is Supervised Learning?* Disponível em: <<https://www.ibm.com/cloud/learn/supervised-learning>>. Acesso em: 13 de julho de 2022.
- [20] 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.
- [21] 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.

- [22] 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-functions-f9d5377d6f02>>. Acesso em: 23 de julho de 2022.
- [23] BREIMAN, L. Random forests. *Machine Learning*, Springer Science and Business Media LLC, v. 45, n. 1, p. 5–32, Janeiro 2001.
- [24] BREIMAN, L. et al. *Classification And Regression Trees*. 1. ed. [S.l.]: Routledge, 1894. 246–280 p.
- [25] BOSCH, J. J. et al. A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals. *Proc. ISMIR*, p. 559–564, 2012.
- [26] PYTHON. Disponível em: <<https://docs.python.org/3/>>. Acesso em: 23 de agosto de 2022.
- [27] LIBROSA: Audio and music signal analysis in python. Disponível em: <<https://librosa.org/doc/latest/index.html>>. Acesso em: 23 de agosto de 2022.
- [28] KLAPURI, A.; DAVY, M. *Signal Processing Methods for Music Transcription*. 1. ed. [S.l.]: Springer, 2006.
- [29] SELL, G.; MYSORE, G. J.; CHON, S. H. *Musical Instrument Detection Detecting instrumentation in polyphonic musical signals on a frame-by-frame basis*. 2006.
- [30] SCIKIT-LEARN: Machine Learning in Python. Disponível em: <<https://scikit-learn.org/stable/>>. Acesso em: 23 de agosto de 2022.
- [31] NUMPY: The fundamental package for scientific computing with Python. Disponível em: <<https://numpy.org/>>. Acesso em: 23 de agosto de 2022.
- [32] KERAS: a deep learning API written in Python. Disponível em: <<https://keras.io/>>. Acesso em: 30 de setembro de 2022.
- [33] SHAP: (SHapley Additive exPlanations). Disponível em: <<https://shap.readthedocs.io/en/latest/index.html>>. Acesso em: 3 de outubro de 2022.