UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ENGENHARIA DE SÃO CARLOS

Bruno Gianesi

Classificação de Gênero via Análise de Áudio Utilizando Métodos de Aprendizado de Máquina Tradicionais

São Carlos

2021

Bruno Gianesi

Classificação de Gênero via Análise de Áudio Utilizando Métodos de Aprendizado de Máquina Tradicionais

Monografia apresentada ao Curso de Engenharia Mecatrônica, da Escola de Engenharia de São Carlos da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Engenheiro Mecatrônico.

Orientadora: Profa. Sandra Aluísio

São Carlos

2021

Ficha catalográfica elaborada pela Biblioteca Prof. Dr. Sérgio Rodrigues Fontes da EESC/USP com os dados inseridos pelo(a) autor(a).

Gianesi, Bruno Honorio do Carmo
G433c Classificação de Gênero via Análise de Áudio Utilizando Métodos de Aprendizado de Máquina Tradicionais / Bruno Honorio do Carmo Gianesi; orientadora Sandra Maria Aluísio. São Carlos, 2021.

Monografia (Graduação em Engenharia Mecatrônica) -- Escola de Engenharia de São Carlos da Universidade de São Paulo, 2021.

1. reconhecimento automático de gênero. 2. classificação de gênero. 3. métodos de aprendizado de máquina tradicionais. I. Título.

Eduardo Graziosi Silva - CRB - 8/8907

Folha de Aprovação

Candidato: Bruno Honorio do Carmo Gianesi

Título do TCC: Classificação de Gênero via Análise de Áudio Utilizando Métodos de Aprendizado de Máquina Tradicionais

Data da Defesa: 21/12/2021

Comissão Julgadora	Resultado
Professora Dra. Sandra Maria Aluísio (orientador)	Aprovado 
Instituição: ICMC/USP	
Professor Dr. Arnaldo Cândido Junior	Aprovado 
Instituição: UTFPR	
Dr. Sidney Evaldo Leal	Aprovado 
Instituição: ICMC/USP	

Presidente da Banca: **Professora Dra. Sandra Maria Aluísio**

Este trabalho é dedicado à minha família e amigos.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a minha família por todo o incentivo dado a minha formação.

Também gostaria de agradecer à República Poltergeist por ser a minha segunda casa durante toda a minha graduação.

Por fim, agradeço à professora Sandra Maria Aluísio pelo tempo dedicado a minha orientação no desenvolvimento deste trabalho.

RESUMO

GIANESI, B. **Classificação de Gênero via Análise de Áudio Utilizando Métodos de Aprendizado de Máquina Tradicionais** . 2021. 60p. Monografia (Trabalho de Conclusão de Curso) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2021.

A tarefa de classificação de gênero baseada na voz identifica automaticamente uma voz como masculina ou feminina, a partir de um arquivo de áudio. Ela tem sido amplamente utilizada e abordada por vários métodos computacionais diferentes, tanto os de aprendizado de máquina tradicionais como os de redes neurais profundas (*Deep Learning*). Neste trabalho, avaliamos métodos de aprendizado de máquina tradicionais e comparamos o uso de diferentes *features* e modelos em *datasets* distintos a fim de determinar a combinação que tem o melhor desempenho na tarefa de classificação de gênero baseada na voz. Além disso, também avaliamos o quanto os modelos treinados em um *dataset* de áudios em português do Brasil gravados em um ambiente controlado generalizam para outros contextos, como outros idiomas (escolhemos o inglês) e ambientes com ruído. Ainda, buscamos também entender se a duração do áudio utilizado no treinamento influencia na precisão dos modelos. Como resultado, constatamos que a combinação do modelo treinado com o método *gradient boosting*, utilizando a agregação das *features* de frequência fundamental, *Mel-frequency cepstral coefficients* (MFCCs) e estatísticas de frequência foi a que teve melhor acurácia no geral (94,1% no CETUC, 91,5% no MLS, 75,3% no Common Voice em português, 90,8% no MLS com ruído e 82,4% no Common Voice em inglês). Também inferimos que os modelos conseguiram generalizar para áudios em língua inglesa e com ruído, com bom desempenho. Por fim, em relação à duração do áudio, entendeu-se que, dependendo da *feature* utilizada, a precisão do modelo pode ser afetada negativamente com a redução da duração do dado.

Palavras-chave: reconhecimento automático de gênero. classificação de gênero. métodos de aprendizado de máquina tradicionais. *Português do Brasil*.

ABSTRACT

GIANESI, B. **Gender Classification via Audio Analysis Using Traditional Machine Learning Algorithms**. 2021. 60p. Monografia (Trabalho de Conclusão de Curso) - Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2021.

The task of classifying gender based on voice identifies automatically a voice as masculine or feminine, based on an audio file. It has been widely addressed by different computing methods, either by traditional machine learning algorithms or by deep learning. In this paper, we evaluate traditional machine learning methods and compare the use of distinct features and models applied on different datasets in order to determine the combination which best performs on the task of gender classification based on voice. In addition, we also assessed whether the models generalize to other contexts, such as other languages (English) or noisy environments, when trained on a Brazilian Portuguese dataset that was recorded in a controlled environment. Also, we try to understand if the duration of the audio file used in the training has an influence on the accuracy obtained by the models. As a result, we found that the combination of the trained model with the gradient boosting algorithm, using the aggregation of the fundamental frequency, Mel-frequency cepstral coefficients (MFCCs) and the frequency statistics was the combination with better accuracy overall (94,1% on CETUC, 91,5% on MLS, 75,3% on Common Voice in Portuguese, 90,8% on MLS with noise and 82,4% on Common Voice in English). We also inferred that the models were able to generalize to audios in English and with noise, with good performance. Lastly, regarding the duration of the audio, it was understood that, depending on the feature used, the accuracy of the model can be negatively affected by reducing the duration of the data.

Keywords: automatic gender recognition. gender classification. traditional machine learning algorithms.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A identificação automática de gênero baseada na voz de um falante é uma tarefa que consiste na classificação binária de um arquivo de áudio em falante do sexo masculino ou feminino. Esta tarefa possui várias aplicações como, por exemplo, facilitar a criação de métodos de reconhecimento de locutores, personalizar interações humano-computador, reconhecer emoções e análisar a voz para investigação de crimes (NAIR; VIJAYAN, 2019; ALKHAWALDEH, 2019; NAIR; SAVITHRI, 2021).

De acordo com Harb e Chen (2003) e também com Kabil, Muckenhirn e Magimai-Doss (2018), duas classes gerais de *features* podem ser usadas para a identificação automática de gênero que estão relacionadas com dois aspectos fisiológicos do sistema de produção de fala de homens e mulheres: a diferença na constituição das pregas vocais e o tamanho do trato vocal. Enquanto que no homen as pregas vocais são mais grossas e mais elásticas e vibram em torno de 125 vezes por segundo (125Hz), na mulher, as pregas são mais finas e tensas e, assim, vibram com maior frequência (250Hz). Os homens geralmente têm um trato vocal mais longo do que as mulheres e, como consequência, as localizações da frequência dos formantes[1] mudam.

A primeira classe usa a extração da altura (*pitch*) de uma onda sonora, sendo que a altura está relacionada com a frequência. De acordo com Cristófaro-Silva (2021), o *pitch* é uma propriedade do som que permite sua classificação como grave ou agudo: quanto maior a frequência, mais agudo é o som, e quanto menor a frequência, mais grave é o som. Além do pitch, nesta classe também é utilizada a feature frequência fundamental (F0), que é medida em Hz (Hertz). Os limites inferior e superior de percepção de ondas sonoras por seres humanos são, respectivamente, 20Hz e 20.000Hz (ou 20kHz). É importante observar que, embora o pitch esteja correlacionado com a frequência fundamental, as escalas do pitch e da frequência fundamental são diferentes.

A outra classe usa medições espectrais de tempo curto que são normalmente realizadas ao longo de janelas de 20 ms, avançando a cada 10 ms como nos *Mel-Frequency Cepstral Coefficients* (MFCC). Um MFCC é primeiramente um coeficiente extraído de um espectrograma representado na Escala de Mel. Isso é feito para que o espectro seja mais representativo quando se considera a maneira em que o sistema auditivo humano compreende as frequências.

Há a possibilidade de usar uma combinação das duas abordagens acima como, por

[1] Segundo Gusmão, Campos e Maia (2010), o formante é representado pelas frequências naturais de ressonância do trato vocal. Geralmente, eles são expressos através de seu valor médio em Hertz (Hz), ou ciclos por segundo, e designados por F1, F2, F3... Fn, de modo progressivo. Cada formante ocorre em uma posição do trato vocal.

exemplo, em Slomka e Sridharan (1997). Além disso, alguns trabalhos utilizam estatísticas (o mínimo, o máximo, a mediana, a média e o desvio padrão, dentre outras) retiradas das frequências de um áudio processado via Transformada de Fourier de tempo discreto como em Nair e Vijayan (2019) ou estatísticas da frequência fundamental (F0) junto com *features* da MFCC como em Levitan, Mishra e Bangalore (2016).

Além de explorar métodos tradicionais de aprendizado de máquina, abordagens mais atuais para a identificação automática de gênero representam sinais de áudio por meio de imagens de espectrogramas para serem analisadas por redes neurais profundas como em Alkhawaldeh (2019), Ferreira et al. (2021). Um espectrograma consiste na representação visual do áudio para indicar a densidade espectral de energia, obtida a partir de transformada de Fourier sobre o sinal do áudio. Os valores são apresentados considerando a relação de tempo e frequência, na qual diferentes cores indicam intensidade da densidade espectral de energia.

É importante notar também, que, em sua maioria, os trabalhos que abordaram essa tarefa não o fizeram utilizando dados em português, sendo esse idioma ainda pouco explorado na área, se comparado a outros idiomas como o inglês.

Neste trabalho em específico, buscou-se entender quais as combinações de *features* e algoritmos/métodos de aprendizado de máquina tradicionais que trazem o melhor desempenho para a tarefa. Também avaliou-se o quanto os modelos treinados em um *dataset* em português e em um ambiente controlado generalizam para outros contextos, como outros idiomas e ambientes com ruído. Por fim, tentou-se entender qual a influência da duração do áudio na precisão dos modelos.

Para a avaliação deste trabalho, foram utilizados quatro *datasets*: CETUC (ALENCAR; ALCAIM, 2008), MLS (PRATAP et al., 2020) e Common Voice (ARDILA et al., 2020) em suas versões em inglês e em português. Sendo o *corpus* CETUC utilizado para treinamento e os demais apenas para teste. E foram utilizados seis algoritmos de aprendizado de máquina tradicionais treinados a partir das *features*: frequência fundamental, *Mel-Frequency Cepstral Coefficients* (MFCCs) e estatísticas retiradas das frequências extraídas do áudio. As métricas utilizadas para a avaliação dos modelos foram a acurácia, precisão, *recall* e *f1-score*.

O repositório oficial do projeto no qual realizou-se todo o processo de extração de *features* e de teste dos modelos se encontra no Github[2].

Este trabalho está organizado da seguinte forma: o Capítulo 2 traz a fundamentação teórica a respeito das *features*, algoritmos e métricas utilizados no trabalho, bem como algumas bibliotecas e *datasets* citados em outros capítulos. O Capítulo 3 traz uma breve revisão da literatura com trabalhos que abordaram a tarefa de reconhecimento de gênero

[2] <https://github.com/BrunoGianesi/Speaker-Gender-Recognition>

pela voz de maneira semelhante, os quais foram utilizados como base para este. O Capítulo 4 detalha a metodologia utilizada, trazendo detalhes dos experimentos realizados (datasets usados, *features* avaliadas, métodos de aprendizado de máquina, métricas de avaliação), os resultados e uma breve discussão acerca destes. O Capítulo 5, por fim, conclui o trabalho, resumindo os resultados e apontando os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Pré-processamento do sinal de áudio

O processamento do sinal de áudio é um processo muito importante para a tarefa de reconhecimento automático de gênero pela voz. Existem diversas maneiras de processar o áudio a fim de extrair as informações desejadas e utilizáveis no processo de aprendizado de máquina. Um método comum de processamento de áudio pode ser descrito em algumas etapas: primeiro, o sinal, que será exibido graficamente em função da amplitude pelo tempo (Figura 1), deverá ser dividido em partes; depois, em cada parcela do sinal, será aplicada a transformada de Fourier de curto termo (ou transformada de Fourier de tempo curto), transformando cada janela em um conjunto de valores representados pela magnitude *vs* a frequência (Figura 2); por fim, converte-se a magnitude para decibels e, agregando todas a janelas, tem-se uma ultima visualização em que o eixo horizontal é o tempo, o vertical a frequência e a intensidade da cor da imagem é a magnitude (Figura 3).

A partir dos espectrogramas é possível não só retirar uma série de informações que podem ser utilizadas em modelos de aprendizado de máquina como também utilizar o próprio espectrograma, de forma a transformar a tarefa em um problema de reconhecimento de imagem.

Desta maneira, um espectograma pode ser entendido como um gŕafico, sendo o tempo no eixo horizontal, as faixas de frequência no vertical e a coloração do ponto como intensidade (VALENTIM; CÔRTES; GAMA, 2010) e a transformada de Fourier como uma operação que transforma um sinal no domínio do tempo para o domínio da frequência.

A Equação 2.1 descreve a função matemática usada pela transformada de Fourier de tempo discreto.

Além disso, muitas vezes é utilizado o espectograma Mel, que é, basicamente, um espectograma transformado para a escala Mel. Esta é, basicamente, uma transformação logarítmica da frequência do sinal do som. Ela é usada para que sons de diferentes distâncias na escala Mel, sejam percebidos de maneira igual por humanos. Por exemplo, os humanos têm problemas para diferenciar frequências mais altas; assim, com a escala Mel, as frequências representam melhor a percepção que temos sobre os sons.

A Equação 2.2 descreve a função matemática usada pela escala Mel.

$$X(\omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-i\omega n} \tag{2.1}$$

$$m = 2595 * log_{10}(1 + \frac{f}{700}) \tag{2.2}$$

Figura 1 – Exemplo de sinal de áudio

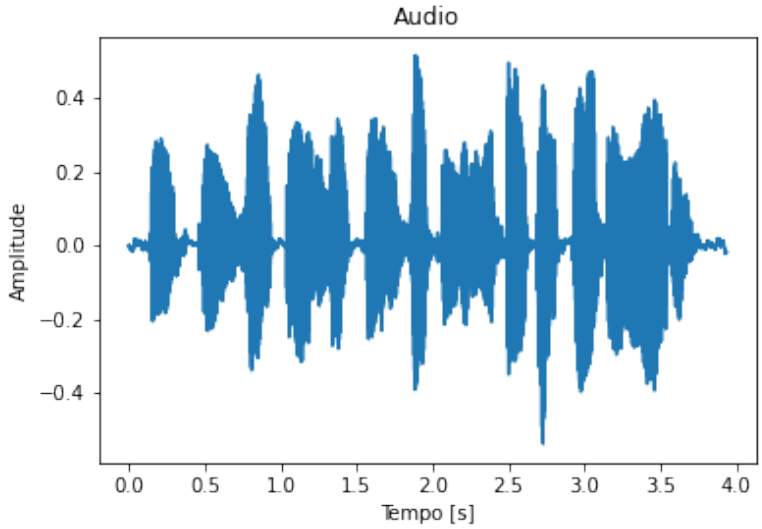
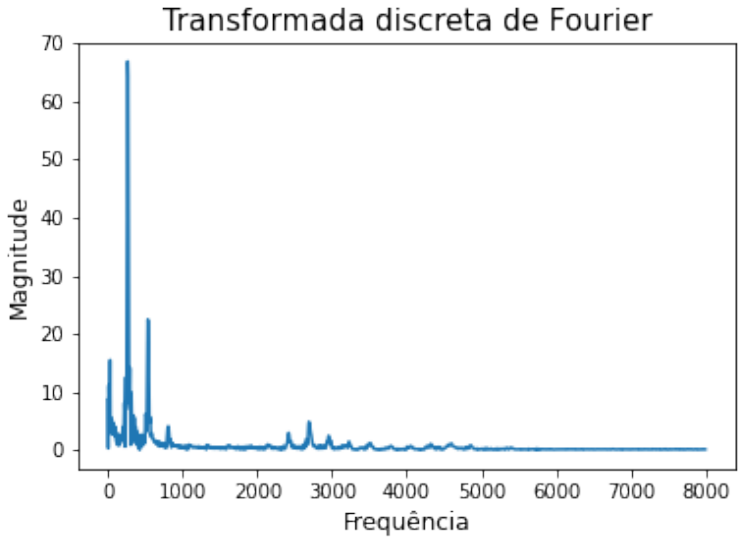


Figura 2 – Exemplo de janela de sinal de áudio após a aplicação da transformada de Fourier de curto termo



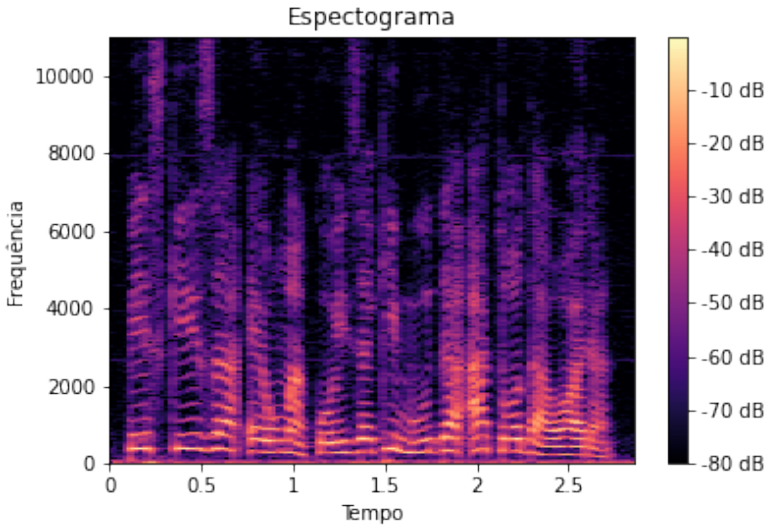
2.2 *Features* utilizadas para a tarefa

Nesta seção, trazemos as *features* mais recorrentes encontradas na literatura para abordar a tarefa de classificação de gênero baseada em áudio. Uma breve descrição destas segue abaixo.

2.2.1 *Mel-Frequency Cepstral Coefficients*

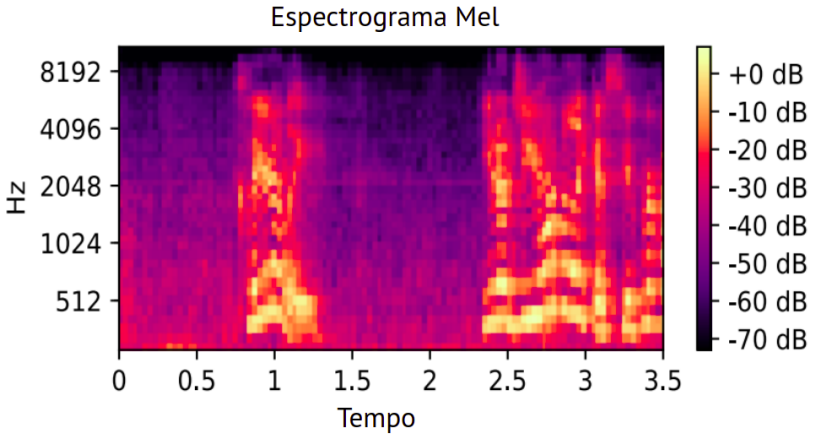
Os MFCCs (*Mel-Frequency Cepstral Coefficients*) são obtidos quando o Espectrograma Mel (Figura 4) (que é um espectrograma construído na escala Mel) sofre uma transformação discreta de cosseno. Esses valores são amplamente utilizados em diversas tarefas de reconhecimento de fala, como identificação do falante/locutor (HASAN et al.,

Figura 3 – Exemplo de espectrograma



2004) ou intentificação do idioma (KOOLAGUDI; RASTOGI; RAO, 2012).

Figura 4 – Exemplo de espectrograma Mel



Traduzido de: (MENG et al., 2019)

2.2.2 Frequência Fundamental

A frequência fundamental (F0) é a primeira frequência produzida pela glote, através dela é possível entender variações de entonação da voz, tal qual sua intensidade (KREMER; GOMES, 2014).

A frequência fundamental é medida em Hz (Hertz), e é definida pelo número de vezes por segundo em que as pregas vocais completam um ciclo de vibração (LUCENTE, 2017). Ainda de acordo com Lucente (2017), uma voz feminina com F0 média de 400Hz indica que a passagem do ar faz as pregas vocais vibrarem 400 vezes em um segundo; e se

a mesma voz passa a vibrar numa frequência de 450Hz indica que a vibração das pregas aumentou.

Alguns autores (SPAZZAPAN et al., 2018; SPAZZAPAN et al., 2020) sinalizam que esta *feature* é uma das mais distintivas na tarefa de reconhecimento de gênero por voz.

2.2.3 Estatísticas de Frequência

As estatísticas de frequência são os valores estatísticos capturados a partir das frequências dominantes[1] de um arquivo de áudio.

Essa extração ocorre a partir da aplicação da transformada de Fourier de tempo discreto e da divisão do áudio em janelas, obtendo o conjunto das maiores frequências. Neste conjunto, então, são realizadas operações para se extrair alguns valores estatísticos, como: média, mediana, moda, desvio padrão, assimetria, curtose, máximo, mínimo, Q25, Q75, e variação interquartil (IQR).

2.3 Métodos de aprendizado de máquina tradicionais abordados na tarefa

As próximas seções trazem um resumo das características de vários métodos de aprendizado de máquina utilizados em trabalhos de identificação automática de gênero da literatura e que foram utilizados neste trabalho de pesquisa:

1. Árvores de Decisão
2. *Random Forest*
3. Support Vector Machines (SVM)
4. Gradient Boosting
5. Multi-layer Perceptron (MLP)
6. Regressão Logística

2.3.1 Árvore de Decisão

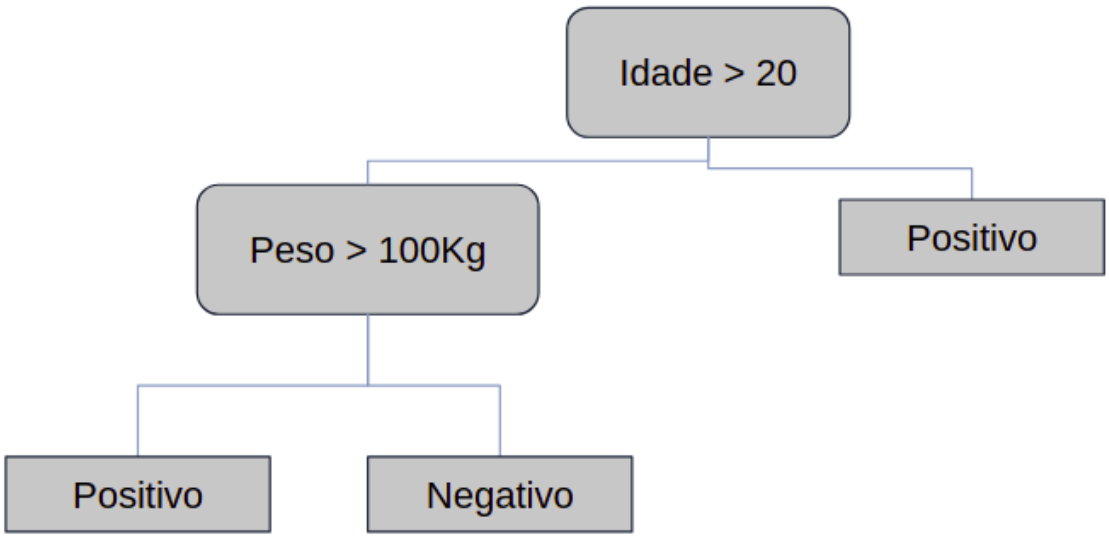
Classificadores baseados em árvores de decisão são modelos de aprendizado não parametrizados que se utilizam de regras simples para separar as *features* e, a partir disso, fazer a predição das classes. Esse tipo de classificador é amplamente utilizado principalmente por ser um modelo de fácil interpretação e visualização.

[1] A frequência dominante é aquela que carrega a energia máxima dentre todas as frequências encontradas no espectro. As frequências podem ser ordenadas pelas suas energias, sendo chamadas de segunda dominante, terceira dominante, etc. (TELGARSKY, 2013).

Um ponto de atenção é que este modelo, dependendo da tarefa, pode acabar gerando árvores muito complexas, que não generalizam os dados do *dataset*, causando *overfitting*.

Um exemplo de árvore de decisão bem simples que classifica a possbilidade de um doença considerando a idade e o peso de uma pessoa é ilustrado na Figura 5.

Figura 5 – Exemplo de fluxograma de árvore de decisão.



2.3.2 *Random Forest*

O *Random Forest* é um algoritmo de *machine learning* supervisionado utilizado para classificação ou regressão.

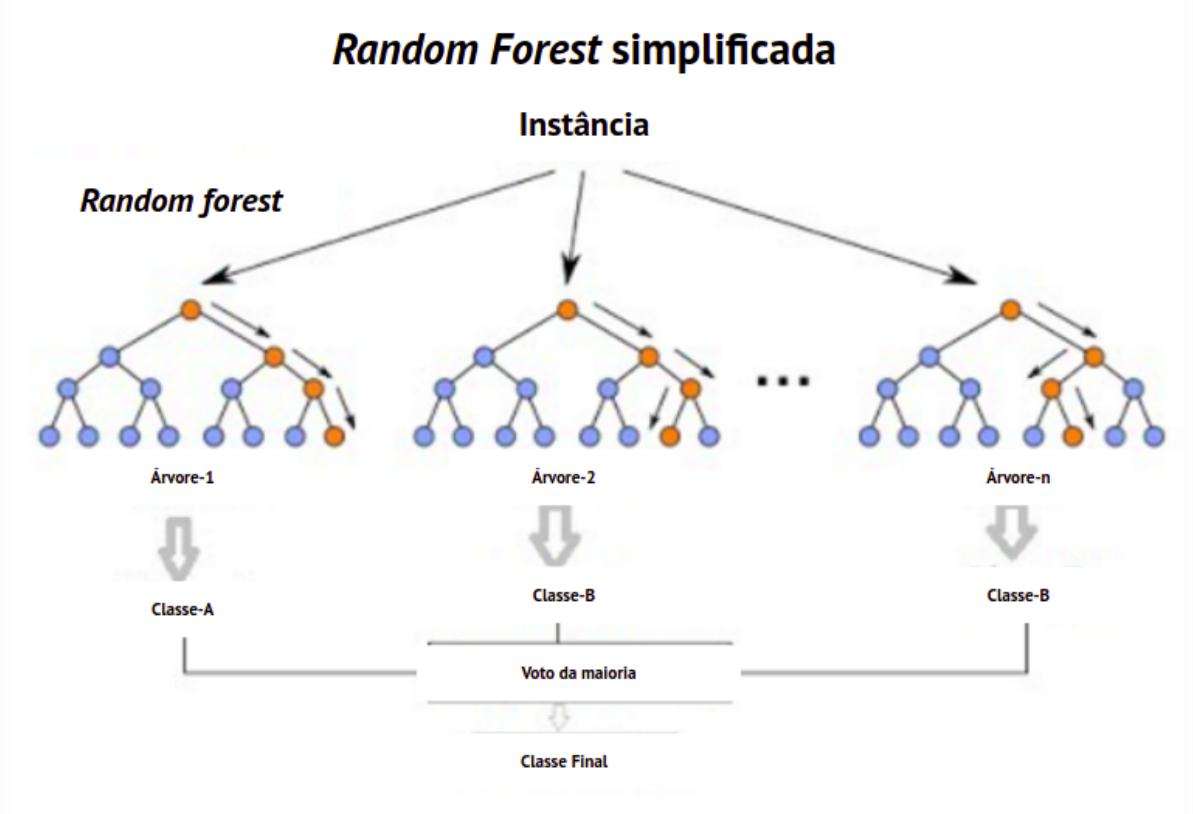
Este modelo se utiliza de subamostras retiradas com teor aleatório do dataset original para gerar diversas árvores de decisão. Estas árvores, então, retornam as previsões das classes e a classe mais prevista é o retorno do modelo. Desta maneira é possivel evitar, em certo grau, o *overfitting* que pode ocorrer nos modelos de árvore de decisão.

A Figura 6 ilustra o modelo de maneira simplificada.

2.3.3 Gradient Boosting

O algoritmo Gradient Boosting é um método de aprendizado de máquina utilizado em problemas de regressão e classificação. Este algoritmo se baseia na produção de um modelo que realiza sua predição se utilizando de uma série de modelos “fracos”, geralmente árvores de decisão. O modelo visa reduzir o valor do erro entre a previsão do modelo anterior e o valor real, para um número predefinido de iterações.

Figura 6 – Ilustração simplificada do modelo *Random Forest*



Traduzido de: (Venkata Jagannath, 2020)

2.3.4 *Support Vector Machine* (SVM)

Máquinas de vetores de suporte ou *SVM*, como são comumente chamadas, são discriminantes lineares as quais buscam encontrar uma linha ou plano de separação entre classes (chamado de hiperplano). Essa linha, por sua vez, busca maximizar a distância dos pontos de dados mais próximos entre duas classes diferentes através de uma margem.

Um exemplo de um caso com classes genéricas, mas com as *features* idade e saldo é ilustrado na Figura 7.

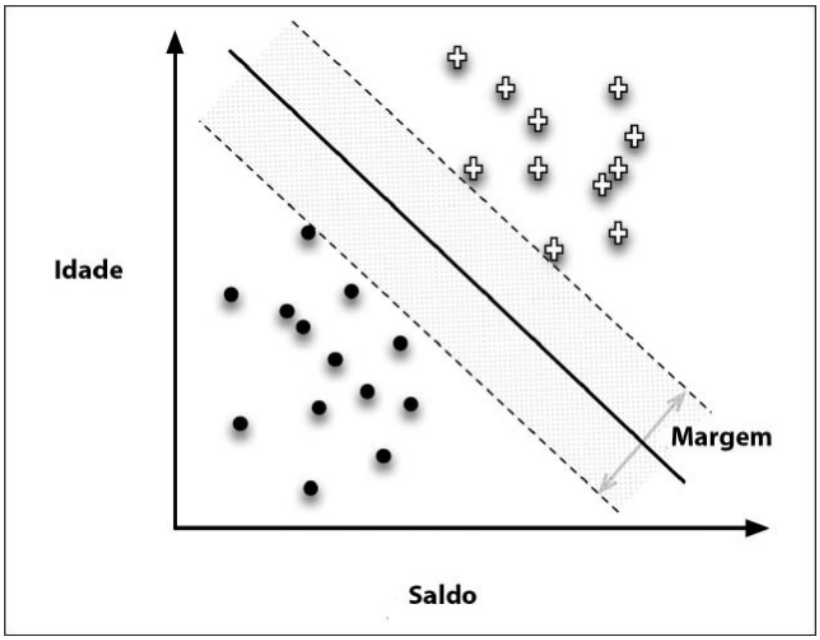
2.3.5 *Multi-layer Perceptron* (MLP)

O método MLP é um algoritmo de aprendizado supervisionado. Ele é uma rede neural que conta com uma ou mais camadas ocultas.

O modelo é treinado a partir de um método chamado retropropagação. Primeiramente, são atribuídos valores aleatórios aos neurônios. Depois, são calculados os valores de saída a partir de uma operação entre os valores de entrada com os valores atribuídos aos neurônios. Por fim, o resultado final é comparado com o esperado e o erro é usado para atualizar os valores dos neurônios, e assim iterativamente.

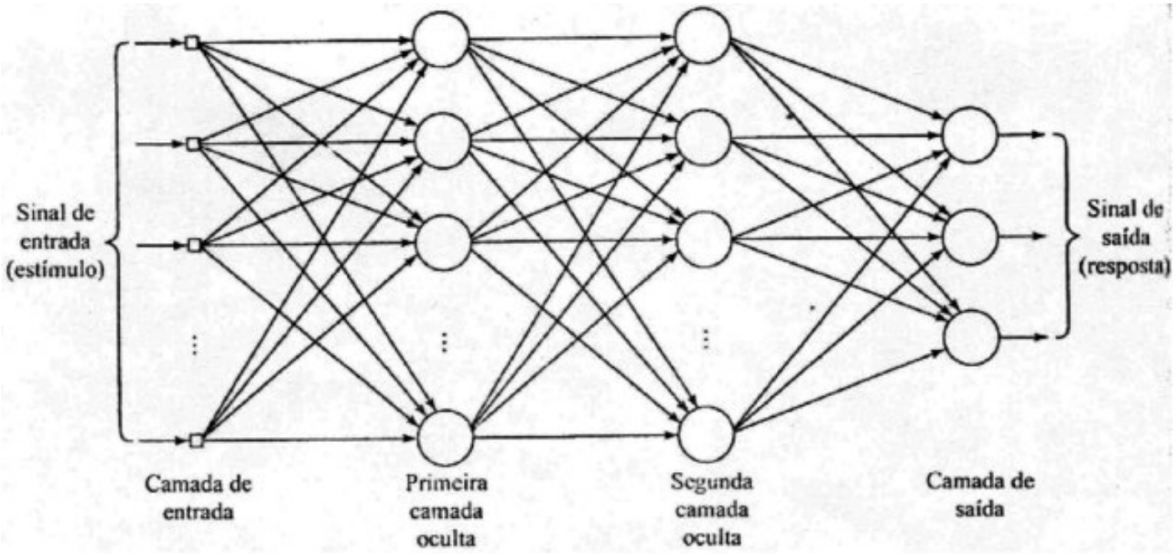
A Figura 10 é um exemplo de um modelo de *multi-layer perceptron*.

Figura 7 – Ilustração do separador linear e sua margem do modelo *SVM*



Fonte: (PROVOST; FAWCETT, 2013)

Figura 8 – Grafo arquitetural de um perceptron de múltiplas camadas com duas camadas ocultas



Fonte: (HAYKIN, 2007)

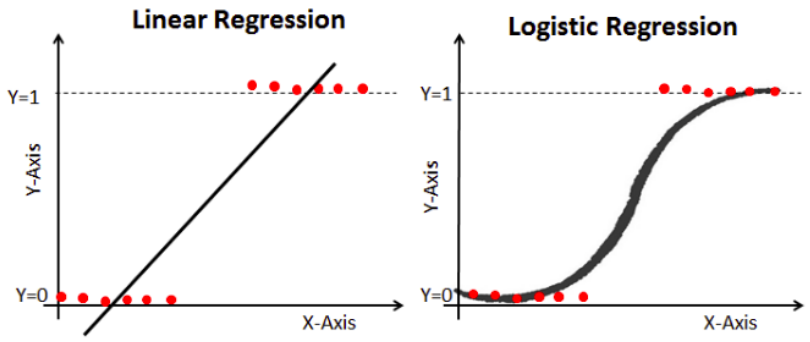
2.3.6 Regressão Logística

O método de regressão logística se utiliza de uma função logística para determinar a probabilidade de um de uma classe ser de um determinado valor a partir de um conjunto de *features*, sendo estas binárias ou não. Um exemplo de função logística é ilustrada na

Equação 2.3

A Figura 9 mostra como se comporta a curva da regressão logistica quando comparada à regressão linear.

Figura 9 – Ilustração do gráfico de regressão linear e regressão logística



Fonte: (Divyansh Chaudhary, 2018)

$$f(x) = \frac{1}{1 + e^{-(x)}} \tag{2.3}$$

2.4 Métricas de Avaliação

Para a classificação, a acurácia e o erro são as medidas básicas de desempenho. Quanto maior a acurácia do modelo, maiores são os acertos e menores são os erros cometidos.

2.4.1 Acurácia

A acurácia pode ser calculada como o número de classificações corretas do modelo dividido pelo número pelo número de classificações totais.

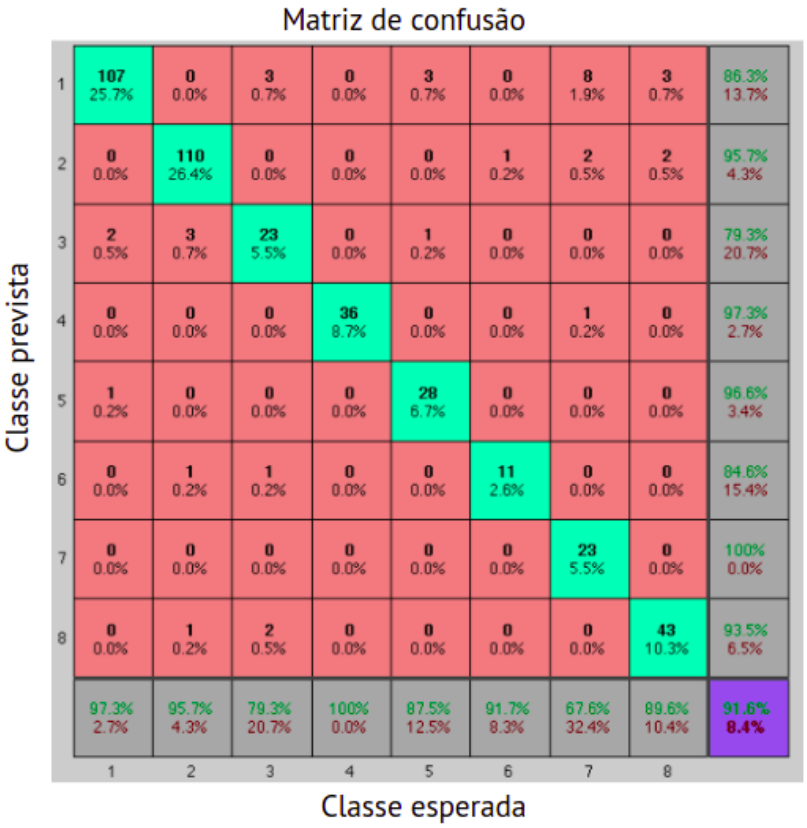
$$Acurácia = \frac{Número\ de\ predições\ corretas}{Número\ total\ de\ predições} \tag{2.4}$$

$$Erro = 1 - Acurácia \tag{2.5}$$

2.4.2 Matriz de confusão

A matriz de confusão é uma tabela que indica os falsos positivos, falsos negativos, verdadeiros positivos e verdadeiros negativos de cada classe. Ela ajuda a ter uma melhor noção da qualidade do classificador, principalmente em datasets com classes desbalanceadas. É a partir dela que são extraídas as métricas de precisão, *recall* e *f1-score*.

Figura 10 – Exemplo de matriz de confusão



Traduzido de: (VISA et al., 2011)

2.5 Bibliotecas utilizadas

2.5.1 Pandas

A biblioteca pandas[2] (MCKINNEY; TEAM, 2015), criada em 2008, é um pacote *open source* construído para a linguagem Python para facilitar a manipulação de dados.

2.5.2 Librosa

Librosa[3] (MCFEE et al., 2015) é uma biblioteca amplamente utilizada na linguagem Python que facilita a análise e manipulação de arquivos de áudio.

2.5.3 scikitlearn

Scikit-learn[4] (PEDREGOSA et al., 2011) é uma biblioteca em Python que disponibiliza um grande conjunto de algoritmos de *machine learning* e outras funções estatísticas que visam facilitar o trabalho na área de aprendizado de máquina.

[2] <https://pandas.pydata.org/>
[3] <https://librosa.org/>
[4] <https://scikit-learn.org>

2.5.4 Parselmouth

Parselmouth[5] (JADOUL; THOMPSON; BOER, 2018) é uma biblioteca que permite o uso, em python, das funcionalidades do *software* Praat[6] (BOERSMA; WEENINK, 2021), uma ferramenta de análise de voz, para pesquisas fonéticas e fonológicas.

2.5.5 Scipy

SciPy[7] (VIRTANEN et al., 2020) é uma coleção de algoritmos matemáticos e funções de conveniência baseadas na extensão numpy do Python. Ele fornece ao usuário comandos e classes de alto nível para manipulação e visualização de dados.

2.6 Datasets

2.6.1 VoxForge

VoxForge[8] é um *corpus* de áudios de fala baseado no conceito de *open source*, ou seja, depende de colaboradores para a obtenção de dados. Os colaboradores podem contribuir gravando sua própria voz lendo algum texto ou gravando um capítulo de um *audiobook*. Pode se dizer também que por causa da flexibilidade das gravações (diferentes *setups* e ambientes), não se pode garantir a qualidade destas.

O *dataset* conta com dados de 17 idiomas incluindo inglês, mandarim e português.

2.6.2 HMIHY ("How May I Help You")

O *corpus* HMIHY ("How May I Help You") (GORIN; RICCARDI; WRIGHT, 1997) conta com mais de dez mil diálogos entre clientes e agentes humanos e foi elaborado com o intuito de automatizar a tarefa de roteamento de chamadas.

Um exemplo de interação presente neste *corpus* é mostrado na Figura 11.

2.6.3 aGender

O corpus aGender[9] contém áudios em alemão de sentenças predefinidas e de texto livre falados por 945 falantes alemães de diferentes idades e gêneros. Cada sentença falada é rotulada com um dos quatro grupos etários: criança, jovem, adulto, idoso e com três grupos de gênero: homem, mulher e criança. O *dataset* conta com quarenta e sete horas de diálogos divididos em áudios de dois a seis segundos cada.

---

5 <https://parselmouth.readthedocs.io/en/stable/>
6 <https://www.fon.hum.uva.nl/praat/>
7 <https://scipy.org/>
8 <http://www.voxforge.org/>
9 <https://paperswithcode.com/dataset/agender>

M: How may I help you?
U: *I tried calling my Aunt and got a wrong number.*
M: You need a billing credit.
U: *(silence)*
M: Please speak the number that you dialed.
U: *908 582 2778*
M: Was the call billed to the phone that you're calling from?
U: *Yes it was.*
M: You will be given credit, thank you for calling.

Figura 11 – Exemplo de diálogo presente no *dataset* HMIHY

Fonte: (GORIN; RICCARDI; WRIGHT, 1997)

3 TRABALHOS RELACIONADOS

Para a realização desse trabalho, foi feita uma revisão de alguns artigos que embasaram as decisões tomadas no decorrer da pesquisa. Abaixo são detalhados três artigos atuais da literatura cujos objetivos eram o reconhecimento do gênero do falante através de métodos de *machine learning* tradicionais.

3.1 Voice based Gender Recognition

O artigo de Nair e Vijayan (2019) tem como objetivo comparar diversos modelos de *machine learning* na tarefa de classificação de gênero através da fala.

Para isso, se utilizaram do *dataset* VoxForge (2021), em inglês e processaram os dados de maneira a remover ruídos e extrair as frequências dominantes em janelas de 200 ms a partir do uso da Transformada de Fourier de tempo discreto. A partir desses valores, são obtidas as *features* estatísticas (média, mediana, moda, desvio padrão, assimetria, curtose, máximo, mínimo, Q25, Q75, IQR) que são utilizadas como entrada nos modelos de Support Vector Machine, Decision Trees, Gradient Tree Boosting e Random Forest.

O modelo com melhor desempenho no trabalho foi o Random Forest, que obteve uma acurácia de 98,7% nos dados de treino. As *features* mais descriminativas da tarefa foram o desvio padrão, a curtose e a assimetria.

3.2 Automatic Identification of Gender from Speech

Em Levitan, Mishra e Bangalore (2016), são utilizados cinco modelos de *machine learning* para abordar a tarefa de classificação automática de gênero baseada em voz. Para isso, foram avaliadas duas *features*, a frequência fundamental (F0) e os MFCCs (*Mel-frequency cepstral coefficients*). No caso da frequência fundamental foram extraídos valores estatísticos relacionados com esta *feature*, como o mínimo, o máximo, a mediana, a média e o desvio padrão, e estes valores serviram como entrada para os modelos de Regressão Linear, Regressão Logística, AdaBoost, Random Forest e Máxima Entropia (MaxEnt) LLAMA (HAFFNER, 2006).

Em relação aos *datasets* utilizados, foram usados o HMIHY ("How May I Help You") corpus (GORIN; RICCARDI; WRIGHT, 1997) em inglês e o aGender corpus[1] (BURKHARDT et al., 2010) em alemão, para explorar a robustez dos classificadores treinados com o dataset em inglês em dados do alemão.

[1] <https://paperswithcode.com/dataset/agender>

Além disso, foi avaliado o impacto em se utilizar um *dataset* monolíngue ou multilíngue e se a duração dos áudios poderia influenciar na acurácia dos modelos. Também foram realizados testes avaliando não só as classes masculino e feminino, mas também uma terceira classe de crianças.

Por fim, concluíu-se que a combinação das *features* F0 e MFCCs, extraídas dos áudios completos e utilizadas no modelo de Regressão Logística, foi a que retornou os resultados mais satisfatórios no teste com as classes masculino *vs* feminino, chegando a um valor de acurácia de 95,2%.

Já nos testes com o classificador ternário (masculino *vs* feminino *vs* criança) o melhor resultado foi obtido a partir da combinação dos MFCCs com a F0 combinada com as estatísticas mínimo, máximo, mediana e desvio padrão da energia, avaliada pelo modelo de *Random Forest*, com um resultado de 85% de acurácia.

3.3 Voice based gender classification using machine learning

Neste artigo, Raahul et al. (2017) avaliaram cinco algoritmos de *machine learning* no contexto de reconhecimento de gênero por fala através de oito diferentes visualizações da acurácia dos modelos.

Os autores se utilizaram de um *dataset* com 1584 áudios de vozes masculinas e 1985 áudios de vozes femininas com diferentes medidas acústicas (média, mediana e desvio padrão da F0, por exemplo), para avaliar os modelos: Análise Discriminante Linear (LDA), *K-Nearest Neighbour* (KNN), Árvore de Classificação e Regressão (CART), *Random Forest* (RF) e *Support Vector Machine* (SVM).

A partir das visualizações *Scatter plot*, *R-plot*, *Parallel plot*, *Pairwise plot*, *Dot plot*, *Density plot*, foi concluído que o *Support Vector Machine* foi o modelo melhor avaliado para a tarefa, mas foi apontado que esse resultado pode variar dependendo do *dataset*.

4 DESENVOLVIMENTO

Este trabalho teve como objetivo avaliar métodos de aprendizado de máquina tradicional e vários conjuntos de *features* (*features* individuais e combinadas) a fim de determinar a melhor combinação de método e *features* que resulte no melhor desempenho da tarefa de classificação de gênero **(Questão de Pesquisa 1)**.

O foco deste trabalho é no Português do Brasil, assim, escolhemos os datasets CETUC (ALENCAR; ALCAIM, 2008), MLS (PRATAP et al., 2020) e *Common Voice* (ARDILA et al., 2020), que são exemplos de discurso lido, sendo que o CETUC e MLS contém áudios limpos, sem barulho e no Common Voice o nível de ruído e a qualidade da gravação são heretogêneos.

Este trabalho também visou avaliar qual o efeito do ruído e da variação na duração dos áudios na classificação do gênero **(Questões de Pesquisa 2 e 3)** e, por fim, entender como é o desempenho dessas combinações de métodos e *features* nesses casos.

Embora o foco do trabalho seja no Português do Brasil, também demos os primeiros passos na avaliação de quais features utilizadas neste trabalho são independentes de língua **(Questão de Pesquisa 4)** e assim podem ser utilizadas em outros trabalhos que se propõem avaliar a tarefa. A língua escolhida para esta avaliação foi o inglês.

Este capítulo está organizado em três seções: (i) que detalham os experimentos que respondem cada uma das questões de pesquisa elencadas acima, trazendo os datasets escolhidos, as features utilizadas, os métodos de aprendizado de máquina selecionados e os métodos de avaliação utilizados (Seção 4.1), (ii) que trazem os resultados em si (Seção 4.2), e (iii) que trazem as discussões sobre os resultados (Seção 4.3).

4.1 Experimentos

4.1.1 Datasets

Para a tarefa, foram utilizados quatro *datasets*, sendo um deles utilizado para treino (CETUC), e os restantes para teste (MLS, Common Voice em sua versão em português e Common Voice em sua versão em inglês). Os datasets são descritos nas próximas seções.

4.1.1.1 CETUC

O dataset CETUC (ALENCAR; ALCAIM, 2008) conta com 145 horas de áudio de 100 falantes divididos igualmente entre os sexos masculino e feminino. Cada falante conta com 1000 sentenças lidas em um ambiente controlado, sem ruído; as sentenças foram retiradas do CETENFolha (Corpus de Extractos de Textos Eletrônicos NILC/Folha de S. Paulo (CENTENFOLHA, 2014)). Este dataset tem uma *sampling rate* de 16kHz.

Para a tarefa, foram utilizados 80% dos falantes para treino e 20% para teste. Os resultados de acurácia, precisão, *recall* e *f1-score* foram obtidos após uma média dos resultados da validação cruzada *5-fold*.

4.1.1.2 MLS (Multilingual LibriSpeech)

O MLS (PRATAP et al., 2020) é um *corpus* que conta com 8 línguas diferentes, entre elas: inglês, alemão, holandês, francês, espanhol, italiano, português e polonês. Sobre os dados em português, que estão relacionados ao contexto deste trabalho, estes estão divididos entre 36 falantes masculinos e 26 falantes femininos, totalizando 62 falantes, que leram 1.261 sentenças em ambiente controlado e sem ruído, retiradas dos audiobooks da LibriVox (LIBRIVOX, 2021).

Para este trabalho, a fim de igualar o número de sentenças lidas por falantes do sexo feminino e masculino, foram utilizadas 871 sentenças, sendo elas 424 de falantes femininos e 447 de falantes masculinos. Além disso, também utilizou-se deste *corpus* para avaliar se existe diferença no desempenho dos modelos quanto à duração dos áudios: foram usados os dados originais vs. três segundos cada áudio. Essa divisão na duração dos áudios foi feita pela biblioteca *pydub* (ROBERT; WEBBIE et al., 2018).

4.1.1.3 Common Voice

Common Voice (ARDILA et al., 2020) é um *corpus* de áudios coletados pela Mozilla Common Voice a partir do website do *Common Voice*[1], disponível publicamente e de forma gratuita, com o objetivo de criar uma amostra que represente a diversidade de pessoas reais.

É um *corpus* de fala lida, sendo as sentenças obtidas de textos de blogs, livros antigos, filmes, e outras fontes de texto públicas. Seu principal propósito é permitir o treinamento e teste de modelos para a tarefa de reconhecimento automático de voz. Existem várias versões do *dataset* em português, sendo que a mais atual é o Common Voice Corpus 7.0, de 21/07/2021, composto de 112 horas das quais 84 foram validadas. A versão usada neste projeto é a Common Voice Corpus 6.1, com 63 horas de áudio, sendo 50 destas horas consideradas validadas. Na versão 6.1, o conjunto de dados compreende 1.120 falantes, 81% dos quais são homens e 3% mulheres[2]. Os áudios foram coletados usando o site Common Voice ou usando um APP para Iphone. Os falantes lêem em voz alta as frases apresentadas na tela. Cada par áudio/transcrição é então analisado por um máximo de 3 colaboradores e é aplicada a votação simples: dois votos para aceitação validam o áudio; dois votos para rejeição invalidam o áudio. Como a coleta desse *dataset* pode ser gravada em qualquer ambiente que o usuário deseje, não há garantias de que os áudios não tenham ruído.

[1] <http://voice.mozilla.org/>

[2] Alguns áudios não apresentam rótulos de gênero

Para este trabalho, o fato do dataset Common Voice 6.1 apresentar uma discrepância no número de dados rotulados entre os sexos masculino e feminino, foi preciso reduzir o número de sentenças para 526, sendo 263 para cada sexo. Também utilizou-se a versão em inglês do dataset (Common Voice Corpus 6.1), com 799 passagens de 408 falantes femininos e 391 falantes masculinos.

4.1.1.4 MLS (Multilingual LibriSpeech) com ruído

A fim de avaliar como os modelos treinados no CETUC desempenhariam em um *dataset* ruidoso, foram mesclados, usando a biblioteca *audiomentations*[3], 842 ruídos diferentes ao MLS (PRATAP et al., 2020). Estes ruídos foram obtidos a partir do *corpus* MUSAN (SNYDER; CHEN; POVEY, 2015).

4.1.2 *Features*

Para a realização da pesquisa, foram escolhidos três conjuntos de *features* que foram avaliados isoladamente e combinados: (i) as estatísticas extraídas da frequência, (ii) a frequência fundamental (F0) e (iii) os MFCCs (*Mel-Frequency Cepstral Coefficients*).

4.1.2.1 Estatísticas de frequência

Para a extração deste conjunto de *features*, foi aplicada a transformada de Fourier nos arquivos de áudio em janelas de tempo de 0.2 segundos e retirada a frequência de maior valor de cada janela. Com esses valores, utilizando-se a biblioteca scipy, foram calculadas diversas medidas estatísticas representativas, como: número de observações, média, assimetria[4], curtose[5], mediana, moda, mínimo, máximo, desvio padrão, Q25, Q75, IQR.

Estes valores estatísticos foram utilizados como entrada para o treinamento dos modelos.

4.1.2.2 *Mel Frequency Cepstral Coefficients* (MFCCs)

Para recuperar os MFCCs, foi utilizada uma função pertencente à biblioteca librosa (MCFEE et al., 2015) com todos os atributos não obrigatórios padrões. A partir dela, foram recuperados 20 coeficientes os quais foram utilizados como entrada para o treinamento dos modelos.

---

[3] <https://github.com/iver56/audiomentations>

[4] A assimetria (em Inglês "skewness") é uma medida da falta de simetria de uma determinada distribuição de frequência (<https://pt.wikipedia.org/wiki/Assimetria_(estatística)>

[5] A curtose é uma medida de forma que caracteriza o achatamento da curva da função de distribuição de probabilidade (<https://pt.wikipedia.org/wiki/Curtose)>

4.1.2.3 Frequência Fundamental (F0)

A frequência fundamental foi extraída dos arquivos de áudio utilizando-se uma função da biblioteca *parselmouth* (JADOUL; THOMPSON; BOER, 2018). A partir do resultado desta função, foram calculadas, utilizando-se a biblioteca scipy, as mesmas medidas estatísticas calculadas na seçao anterior, para serem utilizadas como entrada pelos modelos: número de observações, média, assimetria, curtose, mediana, moda, mínimo, máximo, desvio padrão, Q25, Q75, IQR.

4.1.3 Métodos de aprendizado de máquina

Neste trabalho, foram avaliados seis métodos distintos de aprendizado de máquina: árvore de decisão, *random forest*, *gradient boosting*, regressão logística, *support vector machines* e *multi-layer perceptron*. Todos estes métodos foram usados da biblioteca scikit-learn (PEDREGOSA et al., 2011) sem alterações nos hiperparâmetros.

4.1.4 Métodos de avaliação

Para a avaliação dos modelos, decidiu-se utilizar os métodos de acurácia, precisão, *recall* e *f1-score*, extraídos da matriz de confusão. Todas essas métricas foram recuperadas a partir de funções da biblioteca scikit-learn (PEDREGOSA et al., 2011).

4.2 Resultados

4.2.1 Avaliação das combinações de métodos de aprendizado de máquina com diferentes *features* **(Questão de Pesquisa 1 - features individuais)**

Os resultados obtidos a partir da avaliação dos modelos citados na Seção 4.1.3 pelas *features* descritas na Seção 4.1.2 estão organizados nas Tabelas 1, 2 e 3.

4.2.2 Avaliação das combinações de métodos de aprendizado de máquina com a junção de diferentes *features* **(Questão de Pesquisa 1 - features combinadas)**

Os resultados obtidos a partir da avaliação dos modelos citados na Seção 4.1.3 pela combinação das *features* descritas na Seção 4.1.2 estão organizados nas Tabelas 4, 5, 6 e 7.

4.2.3 Avaliação dos modelos obtidos em um dataset com ruído **(Questão de Pesquisa 2)**

Os resultados obtidos a partir da avaliação dos modelos citados na Seção 4.1.3 no dataset *MLS* (PRATAP et al., 2020) com adição de ruído estão organizados nas Tabelas 8 e 9.

Tabela 1 – Resultados de teste e treino para o dataset CETUC

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência	Decision Tree	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.750	Teste: 0.766	Teste: 0.742	Teste: 0.753
	Random Forest	Treino: 0.986	Treino: 0.986	Treino: 0.986	Treino: 0.986
		Teste: 0.782	Teste: 0.795	Teste: 0.776	Teste: 0.784
	Gradient Boosting	Treino: 0.868	Treino: 0.858	Treino: 0.873	Treino: 0.866
		Teste: 0.814	Teste: 0.831	Teste: 0.805	Teste: 0.816
	Logistic Regression	Treino: 0.699	Treino: 0.735	Treino: 0.681	Treino: 0.707
		Teste: 0.732	Teste: 0.763	Teste: 0.720	Teste: 0.740
	SVM	Treino: 0.851	Treino: 0.860	Treino: 0.842	Treino: 0.851
		Teste: 0.800	Teste: 0.822	Teste: 0.789	Teste: 0.804
	MLP	Treino: 0.886	Treino: 0.892	Treino: 0.880	Treino: 0.886
		Teste: 0.800	Teste: 0.820	Teste: 0.789	Teste: 0.803
F0	Decision Tree	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.884	Teste: 0.851	Teste: 0.910	Teste: 0.879
	Random Forest	Treino: 0.993	Treino: 0.993	Treino: 0.994	Treino: 0.993
		Teste: 0.898	Teste: 0.863	Teste: 0.927	Teste: 0.893
	Gradient Boosting	Treino: 0.953	Treino: 0.952	Treino: 0.953	Treino: 0.952
		Teste: 0.907	Teste: 0.873	Teste: 0.937	Teste: 0.902
	Logistic Regression	Treino: 0.934	Treino: 0.931	Treino: 0.935	Treino: 0.933
		Teste: 0.896	Teste: 0.857	Teste: 0.928	Teste: 0.889
	SVM	Treino:0.950	Treino: 0.948	Treino: 0.950	Treino: 0.949
		Teste: 0.900	Teste: 0.861	Teste: 0.931	Teste: 0.894
	MLP	Treino: 0.957	Treino: 0.960	Treino: 0.953	Treino: 0.956
		Teste: 0.903	Teste: 0.875	Teste: 0.926	Teste: 0.899
MFCCs	Decision Tree	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.796	Teste: 0.802	Teste: 0.792	Teste: 0.796
	Random Forest	Treino: 0.998	Treino: 0.998	Treino: 0.998	Treino: 0.998
		Teste: 0.855	Teste: 0.862	Teste: 0.849	Teste: 0.855
	Gradient Boosting	Treino: 0.968	Treino: 0.965	Treino: 0.969	Treino: 0.967
		Teste: 0.886	Teste: 0.887	Teste: 0.884	Teste: 0.885
	Logistic Regression	Treino: 0.936	Treino: 0.931	Treino: 0.938	Treino: 0.934
		Teste: 0.907	Teste: 0.907	Teste: 0.906	Teste: 0.906
	SVM	Treino: 0.995	Treino: 0.995	Treino: 0.996	Treino: 0.995
		Teste: 0.900	Teste: 0.900	Teste: 0.899	Teste: 0.898
	MLP	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.885	Teste: 0.898	Teste: 0.875	Teste: 0.885

4.2.4 Avaliação dos modelos com a duração dos áudios reduzida **(Questão de Pesquisa 3)**

Os resultados obtidos a partir da avaliação dos modelos citados na Seção 4.1.3 no dataset *MLS* (PRATAP et al., 2020) com os dados originais divididos em áudios de três

Tabela 2 – Resultados de teste com diferentes *features* para o dataset MLS

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência	Decision Tree	0.608	0.682	0.445	0.539
	Random Forest	0.700	0.769	0.595	0.671
	Gradient Boosting	0.753	0.830	0.653	0.731
	Logistic Regression	0.545	0.623	0.289	0.394
	SVM	0.536	0.586	0.327	0.420
	MLP	0.485	0.499	0.855	0.630
F0	Decision Tree	0.804	0.795	0.832	0.813
	Random Forest	0.879	0.826	0.969	0.892
	Gradient Boosting	0.804	0.807	0.812	0.809
	Logistic Regression	0.876	0.814	0.982	0.890
	SVM	0.868	0.895	0.841	0.867
	MLP	0.816	0.918	0.705	0.797
MFCCs	Decision Tree	0.713	0.794	0.595	0.680
	Random Forest	0.783	0.964	0.600	0.739
	Gradient Boosting	0.817	1.000	0.644	0.784
	Logistic Regression	0.798	0.996	0.609	0.756
	SVM	0.781	0.989	0.579	0.731
	MLP	0.765	0.996	0.544	0.703

Tabela 3 – Resultados de teste com diferentes *features* para o dataset Common Voice

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência	Decision Tree	0.500	0.500	0.452	0.475
	Random Forest	0.475	0.474	0.456	0.465
	Gradient Boosting	0.511	0.512	0.475	0.493
	Logistic Regression	0.540	0.553	0.418	0.476
	Support Vector Machine	0.456	0.455	0.437	0.446
	Multilayer perceptron	0.454	0.458	0.498	0.477
F0	Decision Tree	0.738	0.659	0.985	0.790
	Random Forest	0.751	0.671	0.985	0.798
	Gradient Boosting	0.759	0.677	0.989	0.804
	Logistic Regression	0.700	0.626	0.992	0.768
	Support Vector Machine	0.766	0.684	0.989	0.809
	Multilayer perceptron	0.764	0.683	0.985	0.807
MFCCs	Decision Tree	0.690	0.719	0.624	0.668
	Random Forest	0.732	0.818	0.597	0.690
	Gradient Boosting	0.740	0.926	0.521	0.667
	Logistic Regression	0.635	0.918	0.297	0.448
	Support Vector Machine	0.679	0.862	0.426	0.570
	Multilayer perceptron	0.671	0.959	0.357	0.521

Tabela 4 – Resultados de teste e treino para o dataset CETUC com features combinadas - Parte 1

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência + MFCCs	Decision Tree	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.821	Teste: 0.801	Teste: 0.835	Teste: 0.817
	Random Forest	Treino: 0.998	Treino: 0.998	Treino: 0.998	Treino: 0.998
		Teste: 0.873	Teste: 0.852	Teste: 0.891	Teste: 0.870
	Gradient Boosting	Treino: 0.974	Treino: 0.971	Treino: 0.975	Treino: 0.973
		Teste: 0.923	Teste: 0.912	Teste: 0.934	Teste: 0.923
	Logistic Regression	Treino: 0.934	Treino: 0.929	Treino: 0.936	Treino: 0.933
		Teste: 0.924	Teste: 0.928	Teste: 0.922	Teste: 0.924
	Support Vector Machine	Treino: 0.991	Treino: 0.991	Treino: 0.992	Treino: 0.991
		Teste: 0.904	Teste: 0.894	Teste: 0.912	Teste: 0.903
	Multilayer Perceptron	Treino: 0.999	Treino: 0.999	Treino: 0.999	Treino: 0.999
		Teste: 0.858	Teste: 0.842	Teste: 0.870	Teste: 0.855
Estatísticas da frequência + F0	Decision Tree	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.906	Teste: 0.905	Teste: 0.908	Teste: 0.906
	Random Forest	Treino: 0.996	Treino: 0.996	Treino: 0.996	Treino: 0.996
		Teste: 0.922	Teste: 0.912	Teste: 0.932	Teste: 0.922
	Gradient Boosting	Treino: 0.963	Treino: 0.962	Treino: 0.964	Treino: 0.963
		Teste: 0.931	Teste: 0.918	Teste: 0.944	Teste: 0.930
	Logistic Regression	Treino: 0.934	Treino: 0.930	Treino: 0.936	Treino: 0.933
		Teste: 0.915	Teste: 0.914	Teste: 0.917	Teste: 0.915
	Support Vector Machine	Treino: 0.965	Treino: 0.965	Treino: 0.964	Treino: 0.964
		Teste: 0.923	Teste: 0.915	Teste: 0.931	Teste: 0.923
	Multilayer Perceptron	Treino: 0.978	Treino: 0.976	Treino: 0.980	Treino: 0.978
		Teste: 0.919	Teste: 0.901	Teste: 0.938	Teste: 0.918
MFCCs + F0	Decision Tree	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.924	Teste: 0.924	Teste: 0.924	Teste: 0.924
	Random Forest	Treino: 0.998	Treino: 0.998	Treino: 0.998	Treino: 0.998
		Teste: 0.946	Teste: 0.943	Teste: 0.950	Teste: 0.946
	Gradient Boosting	Treino: 0.987	Treino: 0.985	Treino: 0.988	Treino: 0.986
		Teste: 0.951	Teste: 0.945	Teste: 0.958	Teste: 0.951
	Logistic Regression	Treino: 0.978	Treino: 0.976	Treino: 0.979	Treino: 0.977
		Teste: 0.955	Teste: 0.950	Teste: 0.960	Teste: 0.954
	Support Vector Machine	Treino: 0.995	Treino: 0.994	Treino: 0.996	Treino: 0.995
		Teste: 0.953	Teste: 0.950	Teste: 0.958	Teste: 0.953
	Multilayer Perceptron	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.930	Teste: 0.933	Teste: 0.931	Teste: 0.931

segundos cada, estão organizados nas Tabelas 10 e 11.

Tabela 5 – Resultados de teste e treino para o dataset CETUC com features combinadas - Parte 2

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência + MFCCs + F0	Decision Tree	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.894	Teste: 0.853	Teste: 0.930	Teste: 0.889
	Random Forest	Treino: 0.998	Treino: 0.998	Treino: 0.998	Treino: 0.998
		Teste: 0.926	Teste: 0.901	Teste: 0.950	Teste: 0.924
	Gradient Boosting	Treino: 0.984	Treino: 0.983	Treino: 0.983	Treino: 0.983
		Teste: 0.941	Teste: 0.916	Teste: 0.964	Teste: 0.939
	Logistic Regression	Treino: 0.971	Treino: 0.968	Treino: 0.973	Treino: 0.970
		Teste: 0.954	Teste: 0.934	Teste: 0.973	Teste: 0.952
	Support Vector Machine	Treino: 0.994	Treino: 0.994	Treino: 0.994	Treino: 0.994
		Teste: 0.941	Teste: 0.915	Teste: 0.967	Teste: 0.939
	Multilayer Perceptron	Treino: 1.000	Treino: 1.000	Treino: 1.000	Treino: 1.000
		Teste: 0.919	Teste: 0.898	Teste: 0.939	Teste: 0.918

Tabela 6 – Resultados de teste com diferentes combinações de *features* para o dataset MLS

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência + MFCCs	Decision Tree	0.743	0.856	0.600	0.705
	Random Forest	0.739	0.816	0.635	0.714
	Gradient Boosting	0.883	0.981	0.787	0.873
	Logistic Regression	0.804	0.756	0.913	0.827
	SVM	0.569	0.575	0.617	0.595
	MLP	0.642	0.612	0.823	0.702
Estatísticas da frequência + F0	Decision Tree	0.797	0.807	0.794	0.800
	Random Forest	0.882	0.836	0.957	0.893
	Gradient Boosting	0.887	0.832	0.978	0.899
	Logistic Regression	0.859	0.800	0.966	0.875
	SVM	0.677	0.867	0.438	0.582
	MLP	0.634	0.635	0.673	0.654
MFCCs + F0	Decision Tree	0.881	0.858	0.919	0.888
	Random Forest	0.890	0.855	0.946	0.898
	Gradient Boosting	0.886	0.836	0.969	0.897
	Logistic Regression	0.894	0.862	0.946	0.902
	SVM	0.710	0.644	0.969	0.774
	MLP	0.854	0.820	0.917	0.866
Estatísticas da frequência + MFCCs + F0	Decision Tree	0.817	0.780	0.897	0.835
	Random Forest	0.856	0.876	0.839	0.857
	Gradient Boosting	0.915	0.889	0.953	0.920
	Logistic Regression	0.908	0.896	0.928	0.912
	SVM	0.571	0.546	0.966	0.698
	MLP	0.784	0.968	0.600	0.740

Tabela 7 – Resultados de teste com diferentes combinações de *features* para o dataset Common Voice

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência + MFCCs	Decision Tree	0.624	0.643	0.555	0.596
	Random Forest	0.662	0.692	0.582	0.632
	Gradient Boosting	0.717	0.803	0.574	0.670
	Logistic Regression	0.641	0.940	0.300	0.455
	SVM	0.662	0.870	0.380	0.529
	MLP	0.654	0.748	0.464	0.573
Estatísticas da frequência + F0	Decision Tree	0.732	0.652	0.992	0.787
	Random Forest	0.734	0.655	0.989	0.788
	Gradient Boosting	0.722	0.644	0.996	0.782
	Logistic Regression	0.681	0.610	1.000	0.758
	SVM	0.719	0.656	0.920	0.766
	MLP	0.753	0.672	0.989	0.800
MFCCs + F0	Decision Tree	0.690	0.651	0.821	0.726
	Random Forest	0.743	0.663	0.989	0.794
	Gradient Boosting	0.749	0.669	0.985	0.797
	Logistic Regression	0.673	0.613	0.939	0.742
	SVM	0.740	0.660	0.989	0.791
	MLP	0.768	0.692	0.966	0.806
Estatísticas da frequência + MFCCs + F0	Decision Tree	0.677	0.672	0.692	0.682
	Random Forest	0.732	0.670	0.913	0.773
	Gradient Boosting	0.753	0.677	0.966	0.796
	Logistic Regression	0.804	0.784	0.840	0.811
	SVM	0.791	0.752	0.867	0.806
	MLP	0.728	0.729	0.726	0.728

4.2.5 Avaliação dos modelos obtidos em um dataset em inglês **(Questão de Pesquisa 4)**

Os resultados obtidos a partir da avaliação dos modelos citados na Seção 4.1.3 no dataset *Common Voice* (ARDILA et al., 2020) em sua versão em inglês, estão organizados nas Tabelas 12 e 13.

4.2.6 Comparação das acurácias obtidas pelo modelo *gradient boosting*, utilizando a combinação das três *features*

Os resultados obtidos a partir da avaliação do modelo de *gradient boosting* utilizando-se da frequência fundamental, MFCCs e estatísticas da frequência como *features* em todos os *datasets*, estão organizados na Tabela 14.

Tabela 8 – Resultados de teste com diferentes *features* para o dataset MLS com a adição de ruído

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência	Decision Tree	0.630	0.704	0.483	0.573
	Random Forest	0.699	0.776	0.582	0.665
	Gradient Boosting	0.757	0.871	0.617	0.723
	Logistic Regression	0.560	0.663	0.291	0.404
	SVM	0.498	0.511	0.528	0.519
	MLP	0.533	0.538	0.640	0.584
F0	Decision Tree	0.811	0.804	0.834	0.819
	Random Forest	0.879	0.834	0.955	0.891
	Gradient Boosting	0.799	0.805	0.803	0.804
	Logistic Regression	0.870	0.819	0.960	0.884
	SVM	0.858	0.895	0.819	0.855
	MLP	0.809	0.924	0.685	0.787
MFCCs	Decision Tree	0.674	0.730	0.579	0.646
	Random Forest	0.742	0.899	0.559	0.690
	Gradient Boosting	0.768	0.966	0.568	0.715
	Logistic Regression	0.730	0.986	0.481	0.647
	SVM	0.727	0.934	0.503	0.654
	MLP	0.710	0.975	0.445	0.611

4.2.7 Médias das acurácias obtidas pelos algoritmos de machine learning

As médias dos resultados de acurácias obtidas pelos algoritmos de machine learning para cada *dataset* estão organizadas nas Tabelas 18, 19, 20, 22, 24 e 26.

4.2.8 Médias das acurácias obtidas pelas *features*

As médias dos resultados de acurácias obtidas pelas *features* para cada *dataset* estão organizadas nas Tabelas 15, 16, 17, 21, 23, 25.

4.3 Discussão

4.3.1 Melhor combinação de método e *feature*

4.3.1.1 Melhor *feature* isolada

Depois de analisar os resultados de todos os testes realizados, ficou claro que a frequência fundamental (F0) é a *feature* mais descritiva entre as avaliadas. Em segundo lugar ficaram os MFCCs e, por último, as estatísticas retiradas da frequência. Essa ordem se manteve em todos os *datasets* avaliados e suas variações.

Tabela 9 – Resultados de teste com diferentes combinações de *features* para o dataset MLS com a adição de ruído

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência + MFCCs	Decision Tree	0.736	0.834	0.606	0.702
	Random Forest	0.720	0.812	0.591	0.684
	Gradient Boosting	0.851	0.965	0.736	0.835
	Logistic Regression	0.816	0.877	0.747	0.807
	SVM	0.599	0.598	0.667	0.631
	MLP	0.682	0.692	0.685	0.688
Estatísticas da frequência + F0	Decision Tree	0.809	0.822	0.803	0.812
	Random Forest	0.870	0.833	0.935	0.881
	Gradient Boosting	0.881	0.833	0.960	0.892
	Logistic Regression	0.873	0.821	0.962	0.886
	SVM	0.741	0.817	0.638	0.716
	MLP	0.720	0.712	0.763	0.737
MFCCs + F0	Decision Tree	0.835	0.832	0.850	0.841
	Random Forest	0.889	0.857	0.940	0.896
	Gradient Boosting	0.885	0.844	0.953	0.895
	Logistic Regression	0.884	0.884	0.890	0.887
	SVM	0.695	0.632	0.969	0.765
	MLP	0.832	0.799	0.899	0.846
Estatísticas da frequência + MFCCs + F0	Decision Tree	0.804	0.790	0.841	0.815
	Random Forest	0.830	0.845	0.819	0.832
	Gradient Boosting	0.908	0.891	0.935	0.913
	Logistic Regression	0.874	0.940	0.805	0.867
	SVM	0.649	0.603	0.922	0.729
	MLP	0.749	0.967	0.528	0.683

4.3.1.2 Melhores *features* combinadas

Em relação à melhor combinação entre as três *features*, atestou-se que o uso da frequência fundamental com os MFCCs desempenhou, em média, acima de todas as outras combinações, e, inclusive, acima de todas as *features* isoladas. Em segundo lugar ficou o uso da frequência fundamental, dos MFCCs e das estatísticas da frequência combinadas, em terceiro as estatísticas da frequência com a frequência fundamental e, por último, o uso das estatísticas da frequência com os MFCCs.

Nota-se que esses resultados refletem, de certa forma, os resultados discutidos anteriormente, na Seção 4.3.1.1.

Estas conclusões podem ser ilustradas pelas Tabelas 15, 16, 17, 21, 23, 25.

É importante notar, também, que embora essa ordem tenha sido obtida a partir da média das acurácias de todos os modelos experimentados, existem alguns casos em

Tabela 10 – Resultados de teste com diferentes *features* para o dataset MLS com seus dados divididos em áudios de três segundos cada

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência	Decision Tree	0.507	0.531	0.281	0.367
	Random Forest	0.542	0.631	0.243	0.351
	Gradient Boosting	0.542	0.688	0.186	0.293
	Logistic Regression	0.585	0.766	0.268	0.397
	SVM	0.618	0.791	0.339	0.475
	MLP	0.598	0.675	0.407	0.508
F0	Decision Tree	0.853	0.819	0.913	0.863
	Random Forest	0.851	0.812	0.920	0.863
	Gradient Boosting	0.855	0.812	0.932	0.868
	Logistic Regression	0.849	0.797	0.946	0.865
	SVM	0.858	0.824	0.917	0.868
	MLP	0.861	0.815	0.940	0.873
MFCCs	Decision Tree	0.710	0.765	0.621	0.685
	Random Forest	0.735	0.859	0.573	0.688
	Gradient Boosting	0.802	0.961	0.637	0.766
	Logistic Regression	0.784	0.967	0.596	0.737
	SVM	0.776	0.955	0.588	0.728
	MLP	0.756	0.984	0.529	0.688

que uma desempenha melhor do que a outra, por exemplo, na Tabela 7, a combinação das três *features* desempenhou pior do que a combinação da frequência fundamental com os MFCCs nos modelos *decision tree*, *random forest* e *multilayer perceptron*. Entretanto, nos modelos *gradient boosting*, *logistic regression* e *support vector machines*, ela foi a com melhor desempenho. Esse resultado também ocorreu em outros *datasets*.

Em contraponto à ordem apresentada, percebeu-se que a junção das três *features* foi a que trouxe o melhor resultado em todos os *datasets* que foram usados para teste, com exceção do *Common Voice* em sua versão em inglês, sendo que, os modelos que chegaram a esses melhores resultados, são distintos.

4.3.1.3 Análise de acurácia dos métodos de *machine learning*

Sobre os métodos de *machine learning*, o que apresentou a maior média de acurácia entre todas as *features* e suas combinações foi o *gradient boosting*, em segundo lugar ficou o modelo de regressão logística. Logo após, os modelos *random forest* e *support vector machines* apresentaram resultados médios bem parecidos. E, por fim, o único modelo que não teve um desempenho melhor nenhuma vez foi o de árvore de decisão.

Em termos de melhor desempenho, a combinação do modelo *gradient boosting*, treinado e testado a partir da junção das três *features* avaliadas, foi o que teve o melhor

Tabela 11 – Resultados de teste com diferentes combinações de *features* para o dataset MLS com seus dados divididos em áudios de três segundos cada

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência + MFCCs	Decision Tree	0.637	0.746	0.437	0.551
	Random Forest	0.636	0.769	0.409	0.534
	Gradient Boosting	0.719	0.976	0.459	0.624
	Logistic Regression	0.748	0.957	0.529	0.682
	SVM	0.729	0.971	0.483	0.645
	MLP	0.711	0.881	0.501	0.639
Estatísticas da frequência + F0	Decision Tree	0.817	0.778	0.896	0.833
	Random Forest	0.852	0.801	0.944	0.867
	Gradient Boosting	0.857	0.805	0.948	0.871
	Logistic Regression	0.833	0.764	0.973	0.856
	SVM	0.776	0.820	0.718	0.765
	MLP	0.823	0.791	0.887	0.836
MFCCs + F0	Decision Tree	0.813	0.772	0.898	0.830
	Random Forest	0.844	0.800	0.923	0.857
	Gradient Boosting	0.868	0.835	0.924	0.877
	Logistic Regression	0.869	0.845	0.910	0.877
	SVM	0.825	0.766	0.946	0.846
	MLP	0.789	0.746	0.888	0.811
Estatísticas da frequência + MFCCs + F0	Decision Tree	0.718	0.732	0.705	0.718
	Random Forest	0.785	0.839	0.715	0.772
	Gradient Boosting	0.902	0.942	0.861	0.900
	Logistic Regression	0.869	0.917	0.817	0.864
	SVM	0.844	0.865	0.821	0.843
	MLP	0.771	0.905	0.615	0.732

desempenho, resultando nos maiores valores de acurácia nos *datasets* MLS (com ruído, sem ruído e com seus dados divididos em áudios de três segundos cada). Já no CETUC, o melhor desempenho veio da combinação do modelo *logistic regression* com a junção da frequência fundamental com os MFCCs. No *Common Voice* em inglês, o melhor desempenho veio da utilização das estatísticas da frequência combinadas com a frequência fundamental também no modelo *gradient boosting*. Finalmente, no *Common Voice* em português, o modelo com melhor desempenho foi o *logistic regression*, treinado a partir da combinação das três *features*.

Estas conclusões podem ser ilustradas pelas Tabelas 18, 19, 20, 22, 24 e 26.

4.3.2 Influência da língua na identificação de gênero pela voz

Com relação à influência da língua, houve algumas divergências em relação aos resultados obtidos. A acurácia média dos modelos que utilizaram as estatísticas da frequência

Tabela 12 – Resultados de teste com diferentes *features* para o dataset Common Voice na sua versão em inglês

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência	Decision Tree	0.502	0.495	0.849	0.625
	Random Forest	0.476	0.479	0.806	0.601
	Gradient Boosting	0.514	0.502	0.974	0.663
	Logistic Regression	0.552	0.530	0.749	0.621
	SVM	0.506	0.497	0.760	0.601
	MLP	0.462	0.472	0.854	0.608
F0	Decision Tree	0.870	0.825	0.931	0.875
	Random Forest	0.882	0.831	0.954	0.888
	Gradient Boosting	0.886	0.836	0.954	0.891
	Logistic Regression	0.856	0.786	0.969	0.868
	SVM	0.895	0.850	0.954	0.899
	MLP	0.899	0.852	0.959	0.903
MFCCs	Decision Tree	0.512	0.501	0.575	0.536
	Random Forest	0.542	0.526	0.657	0.584
	Gradient Boosting	0.579	0.555	0.711	0.623
	Logistic Regression	0.513	0.503	0.437	0.468
	SVM	0.481	0.474	0.550	0.509
	MLP	0.531	0.521	0.512	0.516

e da frequência fundamental como *features* separadas, foi mais alta na versão do Common Voice (ARDILA et al., 2020) em inglês, quando comparada à acurácia obtida a partir do *dataset* em português. Já os MFCCs tiveram o melhor desempenho na sua versão em português (Tabelas 17 e 25).

Os resultados das combinações das *features* se manteve fiel aos resultados descritos no parágrago anterior, sendo que a combinação das estatísticas da frequência com os MFCCs tiveram o pior desempenho no *dataset* em inglês, assim como a combinação das três *features*.

Contudo, em vista dos altos valores de acurácia obtidos ao se testar os modelos treinados em um conjunto de dados em português, no *dataset* em inglês, é possível concluir que estes foram capazes de generalizar os dados para a tarefa, principalmente se utilizada a frequência fundamental como entrada para os modelos de treino e teste.

4.3.3 Influência da duração do áudio na identificação do gênero pela voz

A partir dos modelos utilizados no *dataset* MLS (PRATAP et al., 2020) com seus dados divididos em áudios de 3 segundos cada, percebeu-se uma diminuição nas médias de acurácia obtidas pelos modelos ao se utilizar as estatísticas da frequência e dos MFCCs, separadamente, se comparados ao conjunto de dados original. Em contraponto, observou-se

Tabela 13 – Resultados de teste com diferentes combinações de *features* para o dataset Common Voice na sua versão em inglês

		Accuracy	Precision	Recall	F1-Score
Estatísticas da frequência + MFCCs	Decision Tree	0.503	0.493	0.581	0.533
	Random Forest	0.472	0.475	0.744	0.580
	Gradient Boosting	0.532	0.512	0.923	0.659
	Logistic Regression	0.544	0.526	0.691	0.597
	SVM	0.516	0.504	0.673	0.576
	MLP	0.467	0.472	0.744	0.577
Estatísticas da frequência + F0	Decision Tree	0.835	0.806	0.872	0.838
	Random Forest	0.881	0.841	0.934	0.885
	Gradient Boosting	0.890	0.839	0.959	0.895
	Logistic Regression	0.852	0.787	0.957	0.864
	SVM	0.768	0.871	0.619	0.723
	MLP	0.875	0.876	0.867	0.871
MFCCs + F0	Decision Tree	0.671	0.632	0.785	0.700
	Random Forest	0.857	0.797	0.951	0.867
	Gradient Boosting	0.867	0.816	0.941	0.874
	Logistic Regression	0.850	0.822	0.885	0.852
	SVM	0.804	0.759	0.877	0.814
	MLP	0.780	0.721	0.898	0.800
Estatísticas da frequência + MFCCs + F0	Decision Tree	0.626	0.583	0.824	0.683
	Random Forest	0.738	0.674	0.900	0.771
	Gradient Boosting	0.824	0.751	0.957	0.841
	Logistic Regression	0.766	0.690	0.946	0.798
	SVM	0.656	0.593	0.944	0.729
	MLP	0.665	0.624	0.790	0.698

Tabela 14 – Comparação das acurácias entre os gêneros para o modelo de *gradient boosting* utilizando os três conjuntos de *features*

	Masculino	Feminino
CETUC (treino)	0.985	0.986
CETUC (teste)	0.875	0.965
MLS	0.953	0.875
MLS (com ruído)	0.935	0.879
MLS (segmentado)	0.860	0.945
Common Voice (Português)	0.965	0.540
Common Voice (Inglês)	0.956	0.696
Média	0.932	0.841

um aumento nas médias dos resultados dos modelos treinados e testados pela *feature* frequência fundamental (Tabelas 16 e 23).

Tabela 15 – Média das acurácias entre as *features* no CETUC

	Média
Estatísticas da frequência	0.791
F0	0.899
MFCCs	0.886
Estatísticas da frequência + MFCCs	0.889
Estatísticas da frequência + F0	0.921
MFCCs + F0	0.949
Estatísticas da frequência + MFCCs + F0	0.934

Tabela 16 – Média das acurácias entre as *features* no MLS

	Média
Estatísticas da frequência	0.577
F0	0.842
MFCCs	0.782
Estatísticas da frequência + MFCCs	0.741
Estatísticas da frequência + F0	0.828
MFCCs + F0	0.884
Estatísticas da frequência + MFCCs + F0	0.837

Tabela 17 – Média das acurácias entre as *features* no Common Voice

	Média
Estatísticas da frequência	0.488
F0	0.755
MFCCs	0.685
Estatísticas da frequência + MFCCs	0.658
Estatísticas da frequência + F0	0.727
MFCCs + F0	0.742
Estatísticas da frequência + MFCCs + F0	0.743

Tabela 18 – Média das acurácias entre os algoritmos no CETUC

	Média
Decision Tree	0.853
Random Forest	0.886
Gradient Boosting	0.907
Logistic Regression	0.897
SVM	0.902
MLP	0.887

Tabela 19 – Média das acurácias entre os algoritmos no MLS

	Média
Decision Tree	0.766
Random Forest	0.818
Gradient Boosting	0.849
Logistic Regression	0.812
SVM	0.673
MLP	0.711

Tabela 20 – Média das acurácias entre os algoritmos no Common Voice

	Média
Decision Tree	0.664
Random Forest	0.690
Gradient Boosting	0.707
Logistic Regression	0.668
SVM	0.688
MLP	0.685

Tabela 21 – Média das acurácias entre as *features* no MLS com ruído

	Média
Estatísticas da frequência	0.595
F0	0.835
MFCCs	0.729
Estatísticas da frequência + MFCCs	0.728
Estatísticas da frequência + F0	0.840
MFCCs + F0	0.860
Estatísticas da frequência + MFCCs + F0	0.817

Tabela 22 – Média das acurácias entre os algoritmos no MLS com ruído

	Média
Decision Tree	0.757
Random Forest	0.804
Gradient Boosting	0.836
Logistic Regression	0.801
SVM	0.681
MLP	0.719

Tabela 23 – Média das acurácias entre as *features* no MLS com seus áudios divididos em segmentos de 3 segundos cada

	Média
Estatísticas da frequência	0.564
F0	0.854
MFCCs	0.766
Estatísticas da frequência + MFCCs	0.715
Estatísticas da frequência + F0	0.828
MFCCs + F0	0.835
Estatísticas da frequência + MFCCs + F0	0.815

Tabela 24 – Média das acurácias entre os algoritmos no MLS com seus áudios divididos em segmentos de 3 segundos cada

	Média
Decision Tree	0.722
Random Forest	0.749
Gradient Boosting	0.792
Logistic Regression	0.791
SVM	0.775
MLP	0.758

Tabela 25 – Média das acurácias entre as *features* no Common Voice em inglês

	Média
Estatísticas da frequência	0.504
F0	0.884
MFCCs	0.522
Estatísticas da frequência + MFCCs	0.510
Estatísticas da frequência + F0	0.864
MFCCs + F0	0.827
Estatísticas da frequência + MFCCs + F0	0.702

Tabela 26 – Média das acurácias entre os algoritmos no Common Voice em inglês

	Média
Decision Tree	0.646
Random Forest	0.693
Gradient Boosting	0.727
Logistic Regression	0.705
SVM	0.661
MLP	0.668

Em relação à combinação das *features*, todas tiveram um desempenho igual ou melhor no conjunto de dados original.

Assim, foi possível observar que, no contexto da tarefa abordada, as *features* das estatísticas da frequência e os MFCCs tiveram pior desempenho em áudios mais curtos.

4.3.4 Influência do ruído na identificação do gênero pela voz

Com relação à questão da influência do ruído na tarefa proposta, verificou-se pouca mudança nos resultados se comparados aos valores obtidos no *dataset* original do MLS (PRATAP et al., 2020). Notou-se, porém, uma ligeira diminuição na média das acurácias obtidas entre os modelos (Tabelas 16 e 21).

Assim, pode-se afirmar que esses métodos podem ser usados em um *dataset* com ruído, embora isto cause uma perda de acurácia.

4.3.5 Diferenças na acurácia entre as classes

Por fim, ao se comparar os resultados da combinação modelo-*features* que teve melhor desempenho na maioria dos *datasets*, ou seja, o modelo *gradient boosting* com as *features* de frequência fundamental, MFCCs e estatísticas da frequência, tem-se melhores resultados na previsão da classe masculina em relação à feminina.

5 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, foram comparados diversos modelos de aprendizado de máquina tradicionais em quatro conjuntos de dados — CETUC, MLS, Common Voice em português e Common Voice em inglês — sendo o primeiro *dataset* utilizado para treinamento. Foram utilizadas três conjuntos de *features* (estatísticas da frequência, frequência fundamental e os MFCCs) para descobrir os melhores modelos e *features* (ou a combinação delas) para a tarefa de identificação automática de gênero a partir de áudios de voz. Todo o desenvolvimento feito para a realização deste projeto pode ser visto no repositório oficial do projeto no Github[1].

Concluiu-se que a frequência fundamental foi a mais discriminativa dentre as três, sendo a mais recomendada para a tarefa em português.

Em relação à combinação das *features*, foi concluído que a combinação do modelo treinado com o método gradient boosting, utilizando a agregação das features de frequência fundamental, Mel-frequency cepstral coefficients (MFCCs) e estatísticas da frequência foi a que teve melhor desempenho no geral.

Entendeu-se também que, embora os melhores resultados gerais tenham sido, em sua maioria, gerados pela união de todas as *features*, é importante ressaltar que, em média, a agregação da frequência fundamental com os MFCCs foi a que apresentou maior acurácia e, tendo em vista que tempo de processamento e extração de *features* pode ser um fator relevante, talvez o ideal seja a utilização de uma quantidade menor de *features*, dependendo da aplicação e da quantidade de dados.

Também avaliou-se o desempenho destes modelos em dados de testes em um idioma diferente dos dados de treino. A comparação foi feita utilizando duas versões do *dataset* Common Voice, a primeira em português e a segunda em inglês. Como conclusão, entendeu-se que, no contexto abordado por este trabalho, foi possível generalizar os dados de treino e teste, mesmo estes sendo de idiomas distintos.

Além disso, foi avaliada a influência do tamanho dos áudios na acurácia dos modelos. Para isso foi gerada uma nova versão do *dataset* do MLS em que seus dados foram divididos em segmentos de três segundos cada. Os resultados mostraram uma perda de desempenho no conjunto de dados segmentados ao se utilizar as *features* MFCCs e as estatísticas da frequência, mas não ao se utilizar a frequência fundamental.

Por fim, analisou-se se a adição de ruído em um conjunto de dados de treino traria uma diminuição significativa de desempenho. Para tal, foi adicionado ruído externo no *dataset* MLS a fim de realizar uma comparação direta da influência do ruído. Como

[1] <https://github.com/BrunoGianesi/Speaker-Gender-Recognition>

resultado, foi observado que houve apenas uma ligeira diminuição nas acurácias obtidas pelo conjunto de dados original. Dessa forma, concluí-se que os modelos treinados no CETUC, também generalizam para dados de teste mais ruidosos nesta tarefa.

Este trabalho abordou um escopo específico no tema de aprendizado de máquina aplicado ao reconhecimento de gênero a partir de áudios. Em trabalhos futuros, seria pertinente um aprofundamento nas avaliações a partir do ajuste de hiperparâmetros dos modelos. Também seria interessante avaliar a influência da idade nos resultados obtidos neste trabalho (usando uma terceira classe com áudios de crianças como no trabalho de Levitan, Mishra e Bangalore (2016) ou um dataset com áudios de idosos), assim como a influência de sotaques das diversas regiões do país.

REFERÊNCIAS

ALENCAR, V. F. S.; ALCAIM, A. LSF and LPC - Derived Features for Large Vocabulary Distributed Continuous Speech Recognition in Brazilian Portuguese. In: **2008 42nd Asilomar Conference on Signals, Systems and Computers**. [S.l.: s.n.], 2008. p. 1237–1241.

ALKHAWALDEH, R. S. Dgr: Gender recognition of human speech using one-dimensional conventional neural network. **Sci. Program.**, v. 2019, p. 7213717:1–7213717:12, 2019.

ARDILA, R. et al. Common voice: A massively-multilingual speech corpus. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 4218–4222. ISBN 979-10-95546-34-4. Disponível em: <https://www.aclweb.org/anthology/2020.lrec-1.520>.

BOERSMA, P.; WEENINK, D. **Praat: doing phonetics by computer [Computer program]**. 2021. Version 6.1.38, retrieved 2 January 2021 <http://www.praat.org/>.

BURKHARDT, F. et al. A database of age and gender annotated telephone speech. In: **Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)**. Valletta, Malta: European Language Resources Association (ELRA), 2010. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2010/pdf/262_Paper.pdf>.

CENTENFOLHA. **CETENFolha**. 2014. Disponível em: <https://www.linguateca.pt/cetenfolha/index_info.html>.

CRISTóFARO-SILVA, T. **Fonética e Fonologia**. 2021. <https://fonologia.org/>. Accessed: 2021-11-16.

Divyansh Chaudhary. **Logistic Regression: Machine Learning in Python**. 2018. <https://medium.com/swlh/logistic-regression-machine-learning-in-python-a2c95c980ca2>, Last accessed on 2021-12-19.

FERREIRA, A. et al. A comparison of deep learning architectures for automatic gender recognition from audio signals. In: **Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional**. Porto Alegre, RS, Brasil: SBC, 2021. p. 715–726. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/eniac/article/view/18297>.

GORIN, A.; RICCARDI, G.; WRIGHT, J. How may I help you? **Speech Communication**, v. 23, n. 1, p. 113–127, 1997. ISSN 0167-6393. Disponível em: <https://www.sciencedirect.com/science/article/pii/S016763939700040X>.

GUSMãO, C. d. S.; CAMPOS, P. H.; MAIA, M. E. O. O formante do cantor e os ajustes laríngeos utilizados para realizá-lo: uma revisão descritiva. **Per Musi**, SciELO Brasil, n. 21, p. 43–50, 2010.

HAFFNER, P. Scaling large margin classifiers for spoken language understanding. **Speech Commun.**, v. 48, p. 239–261, 2006.

HARB, H.; CHEN, L. Gender identification using a general audio classifier. In: **2003 IEEE International Conference on Multimedia and Expo. ICME03. Proceedings.** [S.l.: s.n.], 2003. v. 2, p. 733–736.

HASAN, M. R. et al. Speaker identification using mel frequency cepstral coefficients. **variations**, v. 1, n. 4, p. 565–568, 2004.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007.

JADOUL, Y.; THOMPSON, B.; BOER, B. de. Introducing Parselmouth: A Python interface to Praat. **Journal of Phonetics**, v. 71, p. 1–15, 2018.

KABIL, S. H.; MUCKENHIRN, H.; MAGIMAI-DOSS, M. On learning to identify genders from raw speech signal using cnns. In: **Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018**. [s.n.], 2018. p. 287–291. Disponível em: <https://doi.org/10.21437/Interspeech.2018-1240>.

KOOLAGUDI, S. G.; RASTOGI, D.; RAO, K. S. Identification of language using mel-frequency cepstral coefficients (mfcc). **Procedia Engineering**, Elsevier, v. 38, p. 3391–3398, 2012.

KREMER, R. L.; GOMES, M. L. d. C. A eficiência do disfarce em vozes femininas: uma análise da frequência fundamental. **ReVEL, 12 (23)**, p. 28–34, 2014.

LEVITAN, S. I.; MISHRA, T.; BANGALORE, S. Automatic identification of gender from speech. In: **Proceeding of speech prosody**. [S.l.: s.n.], 2016. p. 84–88.

LIBRIVOX. **LibriVox**. 2021. Disponível em: <https://librivox.org/>.

LUCENTE, L. Introdução à análise entoacional. In: FREITAG, R. M. K.; LUCENTE, L. (Ed.). **Prosódia da fala: pesquisa e ensino**. São Paulo, SP, Brasil: Editora Blucher, 2017. cap. 1, p. 7—26.

MCFEE, B. et al. librosa: Audio and music signal analysis in python. In: CITESEER. **Proceedings of the 14th python in science conference**. [S.l.], 2015. v. 8, p. 18–25.

MCKINNEY, W.; TEAM, P. Pandas-powerful python data analysis toolkit. **Pandas—Powerful Python Data Anal Toolkit**, v. 1625, 2015.

MENG, H. et al. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. **IEEE access**, IEEE, v. 7, p. 125868–125881, 2019.

NAIR, A.; SAVITHRI, S. Classification of pitch and gender of speakers for forensic speaker recognition from disguised voices using novel features learned by deep convolutional neural networks. **Traitement du Signal**, Scopus, v. 38, n. 1, p. 221–230, 2021.

NAIR, R. R.; VIJAYAN, B. Voice based gender recognition. **International Research Journal of Engineering and Technology (IRJET)**, v. 06, n. 05, p. 2109 – 2112, 2019.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PRATAP, V. et al. MLS: A large-scale multilingual dataset for speech research. **arXiv preprint arXiv:2012.03411**, 2020.

PROVOST, F.; FAWCETT, T. **Data Science for Business: What you need to know about data mining and data-analytic thinking**. [S.l.]: O'Reilly Media, Inc., 2013.

RAAHUL, A. et al. Voice based gender classification using machine learning. In: IOP PUBLISHING. **IOP Conference Series: Materials Science and Engineering**. [S.l.], 2017. v. 263, n. 4, p. 042083.

ROBERT, J.; WEBBIE, M. et al. **Pydub**. GitHub, 2018. Disponível em: <http://pydub.com/>.

SLOMKA, S.; SRIDHARAN, S. Automatic gender identification optimised for language independence. In: **IEEE TENCON '97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications**. [S.l.: s.n.], 1997. v. 1, p. 145–148.

SNYDER, D.; CHEN, G.; POVEY, D. **MUSAN: A Music, Speech, and Noise Corpus**. 2015. ArXiv:1510.08484v1.

SPAZZAPAN, E. A. et al. Acoustic characteristics of healthy voices of adults: from young to middle age. In: SCIELO BRASIL. **CoDAS**. [S.l.], 2018. v. 30.

_____. Acoustic characteristics of the voice for brazilian portuguese speakers across the life span. **Journal of Voice**, Elsevier, 2020.

TELGARSKY, R. **Dominant Frequency Extraction**. 2013.

VALENTIM, A. F.; CÔRTES, M. G.; GAMA, A. C. C. Análise espectrográfica da voz: efeito do treinamento visual na confiabilidade da avaliação. **Revista da Sociedade Brasileira de Fonoaudiologia**, SciELO Brasil, v. 15, p. 335–342, 2010.

Venkata Jagannath. **Random Forest Template for TIBCO Spotfire®**. 2020. <https://community.tibco.com/wiki/random-forest-template-tibco-spotfire>, Last accessed on 2021-08-09.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. **Nature Methods**, v. 17, p. 261–272, 2020.

VISA, S. et al. Confusion matrix-based feature selection. **MAICS**, v. 710, p. 120–127, 2011.

VOXFORGE. **VoxForge**. 2021. Disponível em: <http://www.voxforge.org/>.