INSPER

Ciência dos Dados

Projeto 3

Classificação de gênero baseado em propriedades acústicas.

Luiza Silveira, Jose Antônio Bechara.

São Paulo

2018

Introdução:

Objetivo

No projeto de ciências dos dados, foi proposto conduzir uma análise de dados com grau elevado de autonomia e liberdade de escolha de tema e de técnica. Neste trabalho foi avaliado o algoritmo de classificação aplicado a classificação de gênero de acordo com as propriedades acústicas da voz.

Organização do documento :

1 - Introdução : objetivos do trabalho

2 – Metodologia e materiais : Técnicas e recursos usados para a classificação

3 – Resultados : Resultados apresentados pela analise exploratória dos dados e pelo modelo

4 – Conclusão : Conclusões extraídas do classificador e das analises do modelo.

5 – Referencia: Códigos e referencias adquiridos da internet

Metodologia e materiais:

Dataset

O dataset escolhido para a análise continha 3168 amostras de vozes gravadas, entre falantes masculinos e femininos, dessas amostras foram extraídas 21 propriedades acústicas e a classificação da voz como female (mulher) e male (homem), porém no projeto foram usados apenas 6 dessas propriedades, sendo elas:

Meanfreq: Frequência média (kHz)
sd: Desvio padrão da frequência
Median: frequência mediana (kHz)
Q25: Primeiro quartil (kHz)
Q75: Terceiro quartil (kHz)
IQR: Intervalo interquartílico (kHz)

Técnica

O algoritmo utilizado para realizar a classificação foi a regressão logística, essa técnica é recomendada para situações em que a variável dependente é de natureza binaria, ou seja, possui apenas duas categorias. A regressão logística busca estimar a probabilidade dessa variável assumir um determinado valor em função de outras variáveis independentes.

Desenvolvimento

Primeiramente, pegamos o dataset e transformamos em um *Dataframe* usando a biblioteca **Pandas** do python e retiramos as *features* extras que não usaríamos no modelo. Para verificar qual delas melhor diferencia a voz feminina da masculina, fizemos uma análise exploratória dos dados obtidos.

Para a análise, dividimos o *Dataframe* em outros 2, um com os dados só das vozes femininas e outro com os dados das vozes masculinas, na análise foi usado *boxplots* e histogramas.

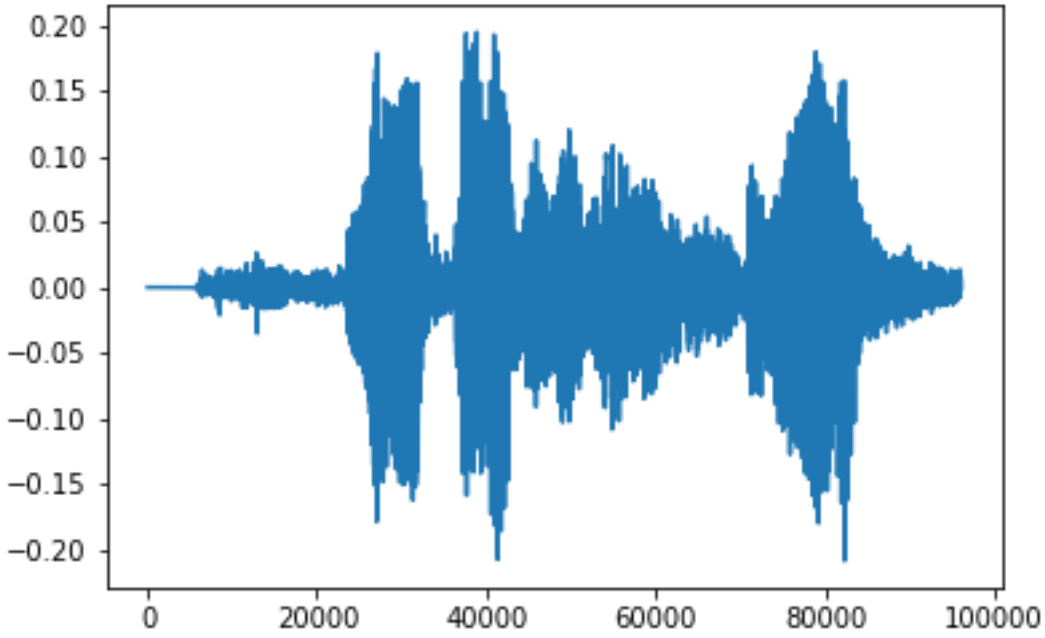
Tendo isso feito, dividimos o Dataframe em 2, um com os valores das features usadas e o outro com a classificação da voz entre homem e mulher, e atribuímos o valor True para mulher e False para homem, após isso dividimos aleatoriamente os 2 *Dataframes* em 60 % treinamento e 40 % teste.

Usando a função ***LogisticRegression*** da biblioteca ***sklearn.linear_model*** treinamos o modelo com os dois Dataframes de treinamento (***model = LogisticRegression()*** **e** ***model.fit(X_train, y_train)***) e após o treinamento, aplicamos o modelo com os Dataframes de teste (***y_pred=*** ***model.predict(X_test)***) e com isso foi possível calcular a acurácia do modelo (***accuracy_score(y_test, y_pred***)**.**

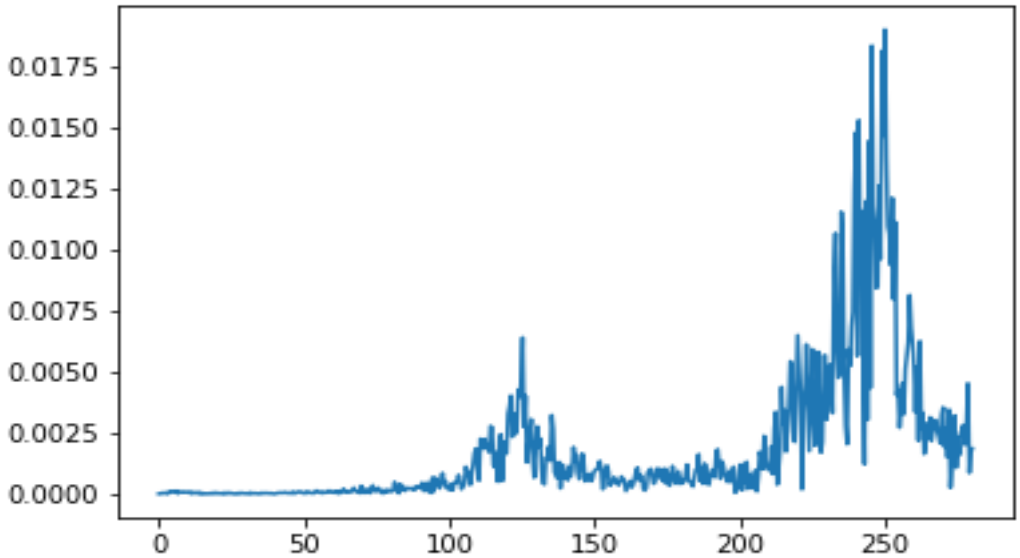
Após testar o modelo, com a biblioteca ***PyAudio*** usamos o microfone do computador para gravar uma amostra da voz da pessoa a ser avaliada.

O programa lê a energia gerada no microfone, e grava esses valores em uma determinada frequência (no caso 48000 vezes por segundo), e assim gera um gráfico da energia por frequência instantânea.

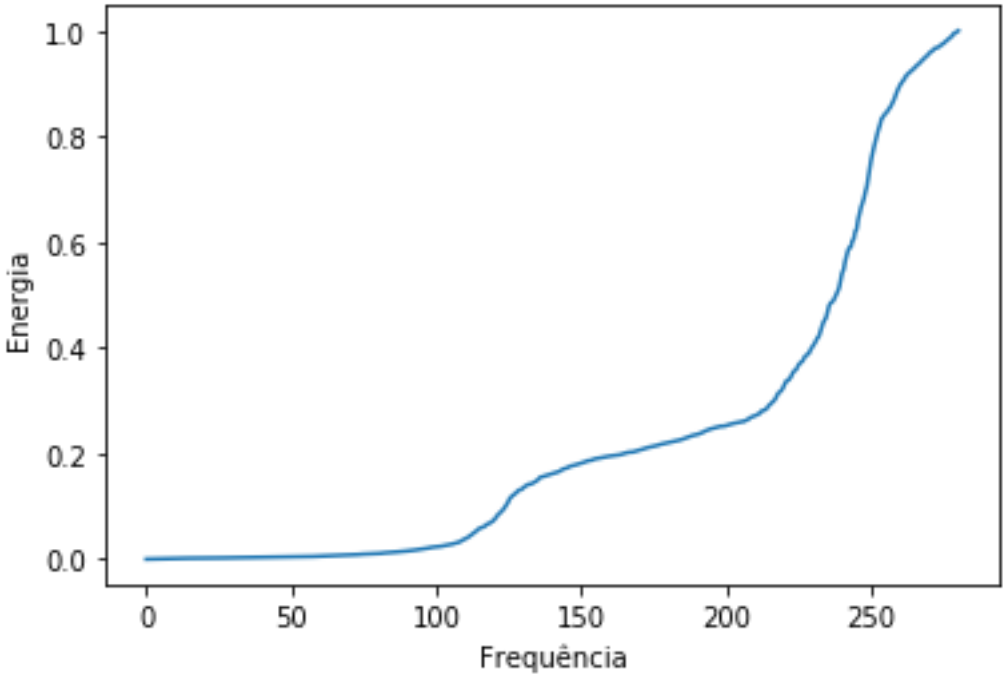
Segue um exemplo desse gráfico :



Em seguida, foi feita a transformada de Fourier que permite analisar de forma mais adequada o gráfico de onda pois o transforma em suas componentes *sinusoidais*, assim obtém-se o gráfico abaixo :



Para simplificar ainda mais a análise da onda, foi realizado a soma acumulativa dos dados obtidos, dessa maneira, é exibido a soma total dos dados conforme crescem com o tempo.

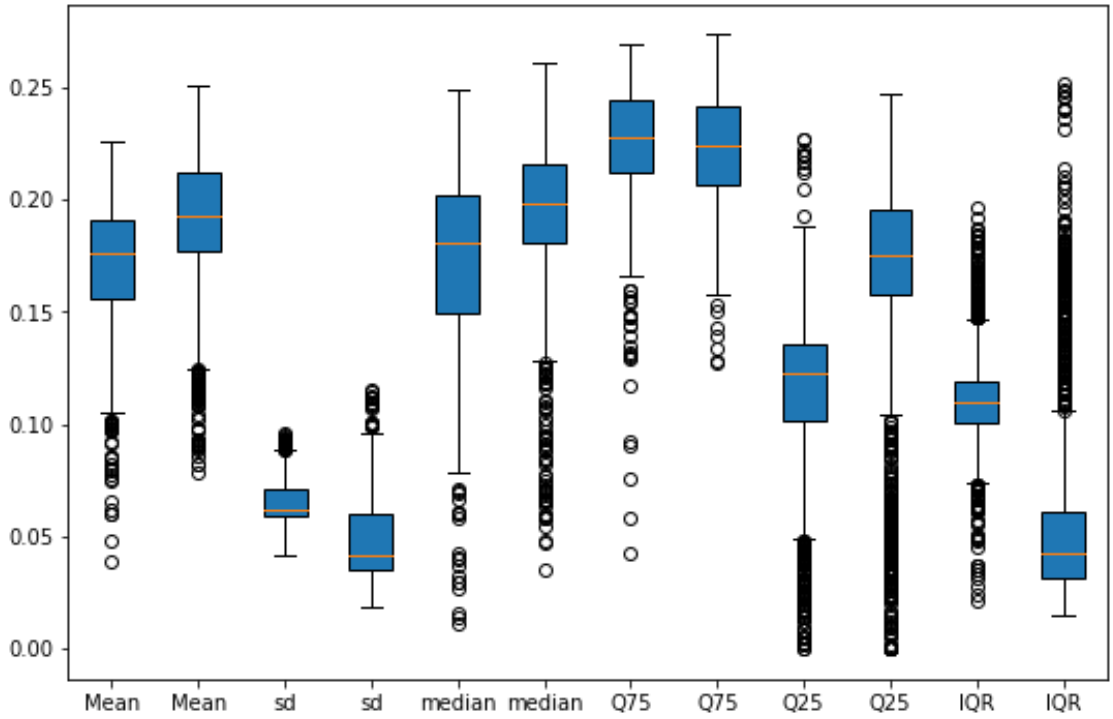


A partir desse gráfico, foi possível extrair as *features* usadas no modelo (média, desvio padrão, mediana, Q25, Q75, IQR) e dessa maneira prever em tempo real se a voz falada no microfone do computador é feminina ou masculina usando o **(*model.predict ([mean, std , median , Q25, Q75 , IQ*]).** O modelo retorna True se a voz for classificada como feminina e False se for masculina.

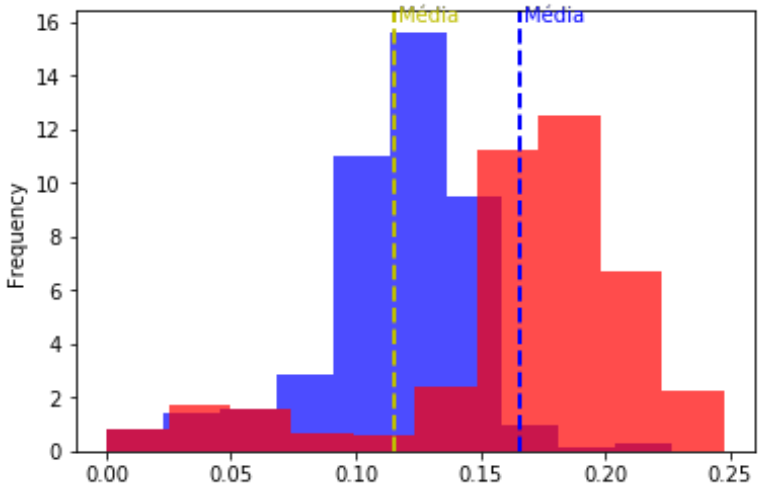
Resultados:

O classificador de regressão logística obteve uma acurácia de 87%

Foi feita uma análise exploratória acerca das 6 propriedades acústicas, o gráfico abaixo possui pares de boxplots, cada um para homens e mulheres respectivamente, pelos *boxplots* é possível ver a distribuição e valores discrepantes (*outliers*) dos dados e compará-los.

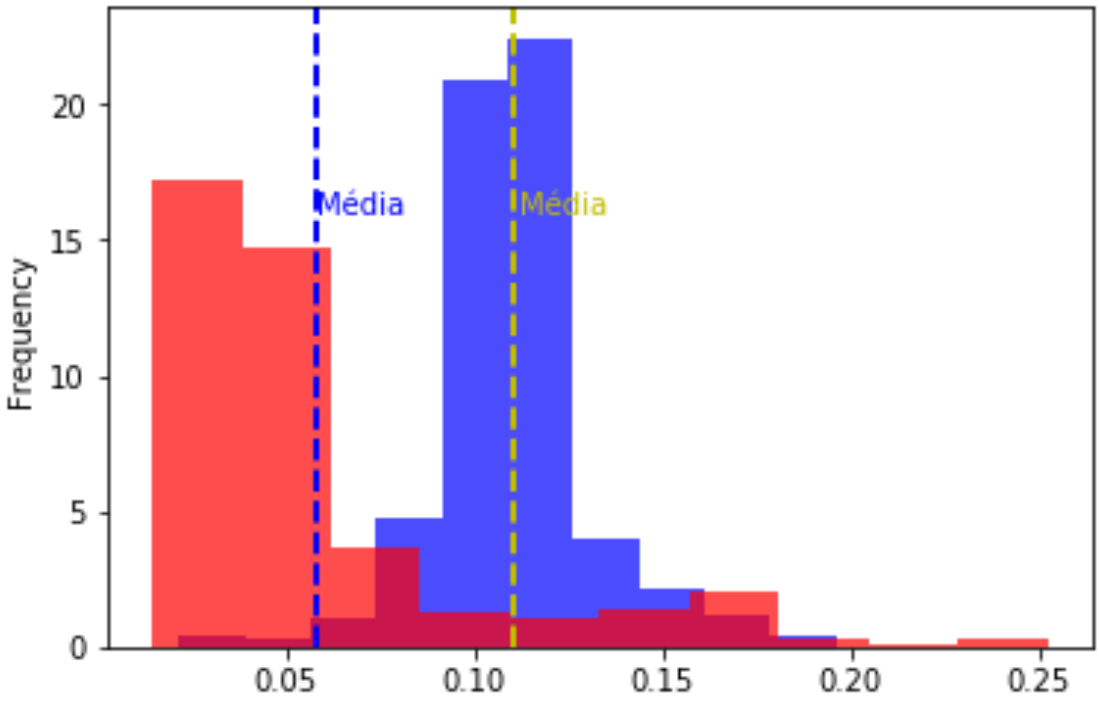


Analisando os dados acima, percebe-se que há uma leve diferença entre as médias, a mediana, o desvio padrão e o terceiro quartil das frequências das vozes masculinas e femininas, porém essas diferenças não são tão expressivas quanto os dados do primeiro quartil e do intervalo interquartil. Então, para analisar mais precisamente essas duas variáveis, segue abaixo um histograma delas:



Em vermelho está os dados do Q25 das mulheres e em azul está os dados do Q25 dos homens, percebe-se que os valores para os homens são menores que os das mulheres e que em média esse

valor é de 0,11 khz, enquanto os das vozes femininas são de 0,16 kHz.



Esse histograma mostra os dados do intervalo interquartil (Q75 - Q25), onde há a maior discrepância de valores. A media desses valores para as vozes femininas são de 0,05 kHz enquanto a masculina é de 0,11 kHz.

Conclusão:

O classificador de regressão logística possuiu uma alta acurácia, fazendo dele um bom classificador para esse modelo.

É possível concluir também que para ter um acerto maior na hora de classificar o sexo de uma pessoa em detrimento da sua voz, a melhor opção é calcular o primeiro quartil (*Q25*) e o Intervalo Interquartil (*IQR*) da voz.

Referencias :

Audio recording:

https://python-sounddevice.readthedocs.io/en/0.3.12/usage.html#recording

Transformada de *fourier*:

https://docs.scipy.org/doc/numpy-1.15.0/reference/routines.fft.html

Regressão logistica:

https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8