***Voice-based gender identification***

1. **Abstract:**

Gender recognition by voice is one of the attractive and interesting technologies when we can catch terrorists hiding without knowing their identity, voice recognition is a recognition system that uses Use frequencies and Acoustic Properties Measured to determine whether the person is male or female. In this project we used ML algorithms including Logistic Regression, GaussianNB, Decision Tree, Random Forest, and achieved 98.78% accuracy on the dataset.

1. **Introduction**:

Determining a person’s gender as male or female, based upon a sample of their voice seems to initially be an easy task. Often, the human ear can easily detect the difference between a male or female voice within the first few spoken words. However, designing a computer program to do this turns out to be a bit trickier

The model is constructed using 3,168 recorded samples of male and female voices, speech, and utterances. The samples are processed using acoustic analysis

One of the reasons is that it can improve human-machine interaction. For example, ads can be specialized based on the age and gender of the person on the phone. It can also help identify suspects in criminal cases or at least it can minimize the number of suspects. Some of the other uses of this system could be to tailor queue music, where a different type of music can be played according to the person's age and gender.

[Speaker Gender Recognition Based on Deep Neural Networks and ResNet50 (hindawi.com)](https://www.hindawi.com/journals/wcmc/2022/4444388/)

[DGR: Gender Recognition of Human Speech Using One-Dimensional Conventional Neural Network (hindawi.com)](https://www.hindawi.com/journals/sp/2019/7213717/)

[GitHub - PrathamSolanki/gender-recognition-by-voice: Identify a voice as male or female.](https://github.com/PrathamSolanki/gender-recognition-by-voice)

<https://github.com/primaryobjects/voice-gender>

1. **Project Goal:**

The research goal was to design a system that could accurately identify a person’s gender using their voice. A set of experiments were conducted in order to determine the most appropriate classifier for gender classification. Choosing a classifier for the gender detection problem in multimedia applications is based on identifying genders in data.

1. **Data**

We won't be using raw audio data since audio samples can be of any length and can be problematic in terms of noise. As a result, we need to perform some feature extraction before feeding it into the neural network.

We'll be using ***Mozilla's Common Voice Dataset***, a corpus of speech data read by users on the ***Common Voice website.*** Its purpose is to enable the training and testing of automatic speech recognition. However, after I took a look at the dataset, many of the samples were labeled in the genre column. Therefore, we can extract these labeled samples and perform gender recognition.

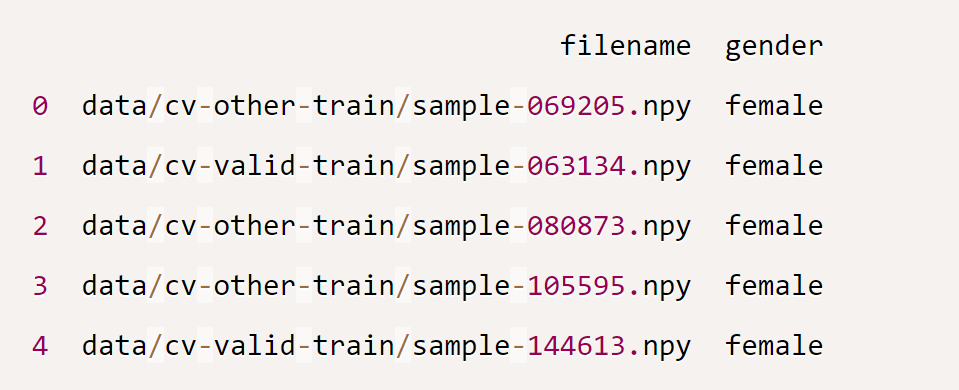
Here is what I did to prepare the dataset for gender recognition:

* First, I only filtered the labeled samples in the genre field.
* After that, I balanced the dataset so that the number of female samples is equal to male samples; this will help the neural network not overfit on a particular gender.
* Finally, I've used the [Mel Spectrogram](https://translate.google.com/website?sl=en&tl=vi&hl=vi&anno=2&client=webapp&u=https://en.wikipedia.org/wiki/Mel_scale) extraction technique to get a vector of the length 128 from each voice sample.

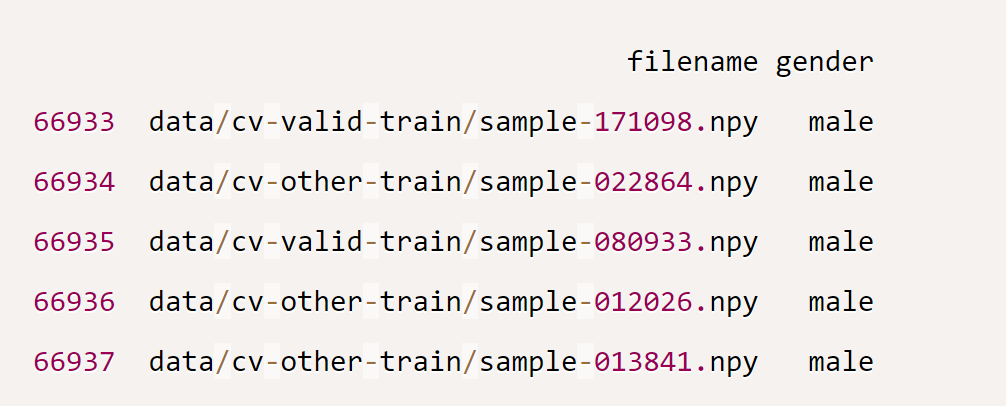
Now to get the gender of each sample, there is a CSV metadata file:



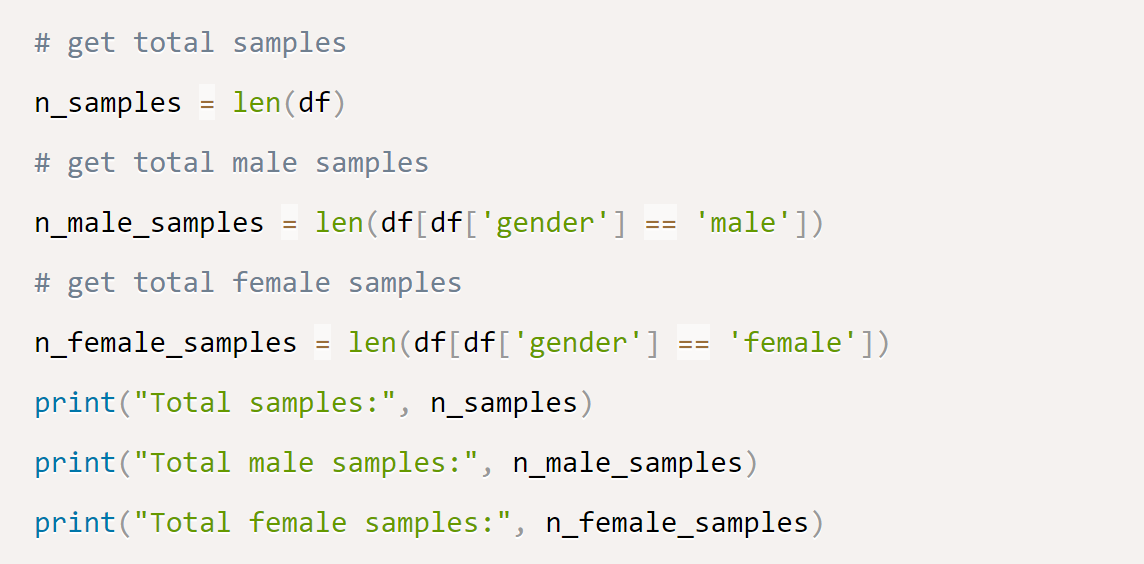
Output:



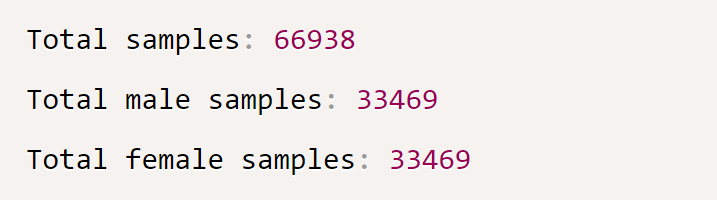
Dataframe ends:



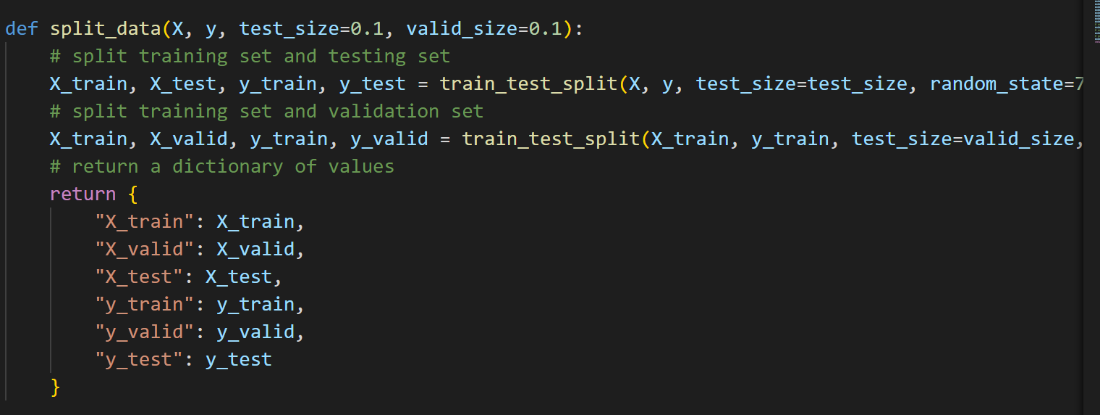
The number of samples of each gender:



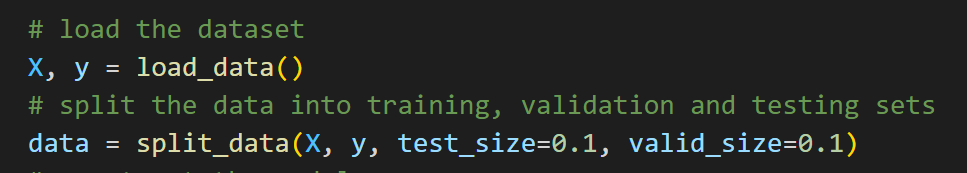
Output:



Next, Split our dataset into training, testing, and validation sets:

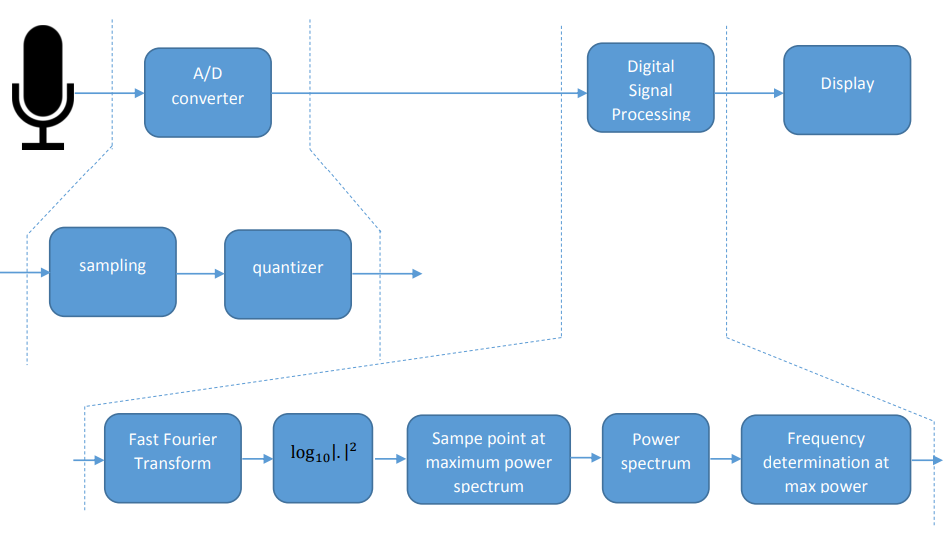


Using ***sklearn's train\_test\_split()*** convenient function, which will shuffle our dataset and split it into training and testing sets. Then rerun it on the training set to get the validation set.



This data dictionary contains everything we need to fit model.

1. **System Design & Implementation**



1. Analog to digital converter (A/D converter)

An analog to digital converter consists of two main blocks:

* + - Sampling block: This block samples the analog signal at the input. The output outputs a discrete signal.
    - Quantization block: This block takes the input signal as a discrete signal at the output of the sampler. The signal at the output of the quantizer is a discrete signal with amplitudes that have been reduced to definite levels.

1. Digital Signal Processing

The voice gender recognition program's digital signal processor consists of five blocks main:

* + - FFT transform block: This block takes an input signal that is a discrete signal in the time domain. The output of the FFT block is the frequency domain response of the input signal.
    - Log10|.|2 block: calculate the spectral power density (in dB) of the voice signal from the frequency domain response
    - Block [sample point at maximum power spectrum] + [power spectrum] + [frequency determination at max power]: determines the frequency at which the spectral power density is highest. From there determine whether this is a male or female voice.

1. Display block:

Displays whether the speaker is male or female.

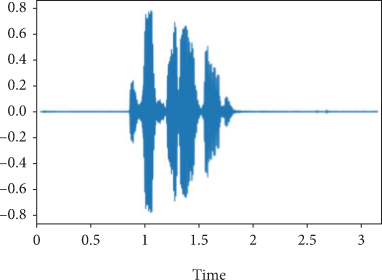


Figure: Original time domain signal

1. **Method**
2. **Fourier Transforms**

A Fourier transform (FT) is a [mathematical](https://en.wikipedia.org/wiki/Mathematics) [transform](https://en.wikipedia.org/wiki/Integral_transform) that decomposes [functions](https://en.wikipedia.org/wiki/Function_(mathematics)) depending on [space](https://en.wikipedia.org/wiki/Space) or [time](https://en.wikipedia.org/wiki/Time) into functions depending on [spatial frequency](https://en.wikipedia.org/wiki/Spatial_frequency) or [temporal frequency](https://en.wikipedia.org/wiki/Frequency). That process is also called [*analysis*](https://en.wikipedia.org/wiki/Analysis). An example application would be decomposing the [waveform](https://en.wikipedia.org/wiki/Waveform) of a musical [chord](https://en.wikipedia.org/wiki/Chord_(music)) into terms of the [intensity](https://en.wikipedia.org/wiki/Sound_intensity) of its constituent [pitches](https://en.wikipedia.org/wiki/Pitch_(music)). The term *Fourier transform* refers to both the [frequency domain](https://en.wikipedia.org/wiki/Frequency_domain) representation and the [mathematical operation](https://en.wikipedia.org/wiki/Operation_(mathematics)) that associates the frequency domain representation to a function of space or time.



Audio signals can be complicated combinations of several sound components. First, it is decomposed into building components that can be processed more efficiently to understand a signal better. When these building blocks are exponential functions, the procedure is known as Fourier analysis. The Fourier transform [1] converts a time-dependent signal to a frequency-dependent role, revealing the original signal’s frequency spectrum. The Fourier transform gives the signal’s frequencies as well as their magnitudes. The inverse Fourier transform converts the frequency-domain representation of a given signal into the original signal

1. **Discrete Fourier Transform (DFT)**

Discrete Fourier Transform called the finite Fourier transform, is a transform in Fourier analysis for discrete time signals. The input to this transform is a finite series of real or complex numbers. This transform can be computed quickly by the fast Fourier transform algorithm

*DFT's formula*:

k = 0,…,N-1

*IDFT’s formula:*

k = 0,…,N-1

1. **Fast Fourier Transform (FFT)**

The fast Fourier transform (FFT) is a mathematical procedure frequently used to estimate the discrete Fourier transform of any sequence (DFT). Using the FFT technique, each frame of those samples is converted from a time-domain discrete signal to a frequency-domain signal. The FFT is considered an effective computing implementation of the DFT method, which is specified on a set of samples as follows:

1. **Chroma-STFT (Short-Time Fourier Transform)**

Used short-term Fourier transformation to compute Chroma features. STFT represents information about the classification of pitch and signal structure. It depicts the spike with high values. An octave is defined as the distance of 12 pitches in our scale. Tone height and Chroma are two components of a pitch. Chroma STFT, as illustrated in Figure 1(a), generates a chromagram from a waveform or power spectrogram.

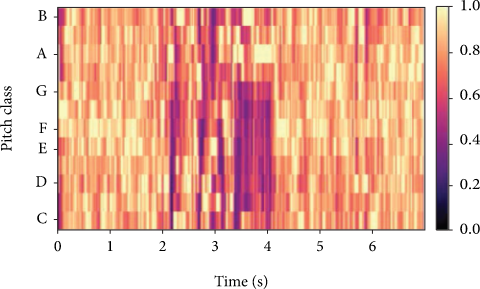


Figure 1(a): Chromogram

1. **Mel Spectrogram**

A spectrogram is a visual representation of a signal’s frequency spectrum. The Mel scale [3] is a mathematical representation of how the human ear works, demonstrating that people do not perceive frequencies on a linear scale. Humans are more sensitive to differences at lower frequencies than at higher frequencies. In mathematical terms, the Mel scale is the outcome of a nonlinear transformation of the frequency scale [4]. The term “Mel-frequency scale” refers to a scale that is defined as:

The term “Mel spectrogram” refers to a spectrogram that has been converted to the Mel scale. For example, Mel spectrogram returns a power spectrogram coefficient that has been Mel scaled. Mel spectrogram object is shown in Figure 1(b)

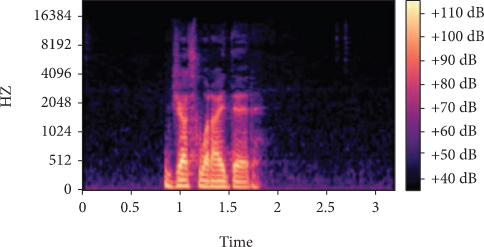


Figure 1(b): Mel spectrogram

1. **MFCC**

MFCC accurately portrays the vocal tract, a filtered shape of a human voice, and a short-time power spectrum envelope. MFCC is nothing but the coefficients that make up the Mel-frequency cepstrum as shown in Figure 1(c)

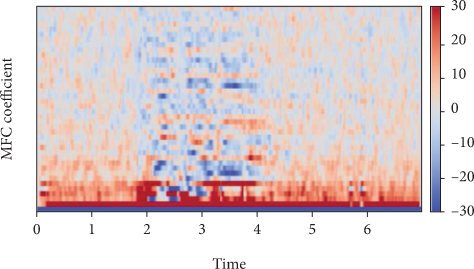


Figure 1(c): MFC coefficients

The first 13 coefficients of MFCC are considered features because they reflect the spectral envelope. And the higher dimensions that have been deleted express the spectral subtleties

1. **Progress**

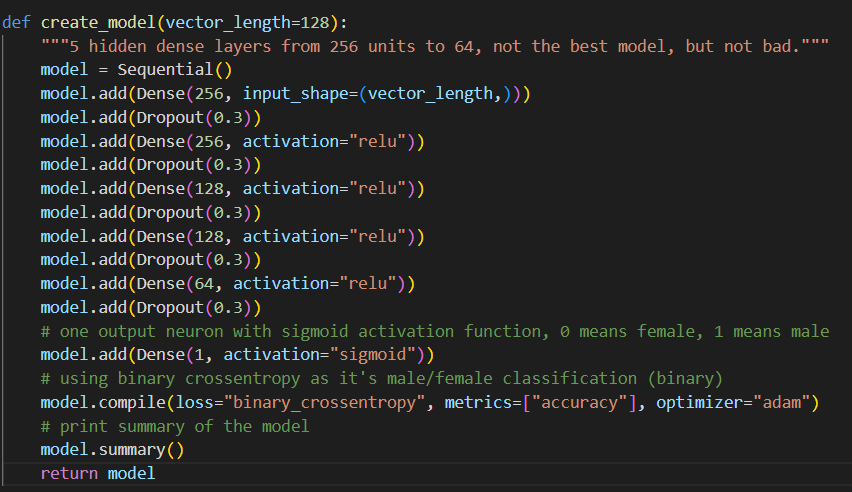
## *Training the Model*:

We defined two callbacks that will get executed after the end of each epoch:

The first is the [tensorboard](https://www.tensorflow.org/tensorboard); we will use it to see how the model goes during the training in terms of loss and accuracy.

The second callback is [early stopping](https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping); this will stop the training when the model stops improving. I've specified a patience of 5, which means it will stop training after 5 epochs of not improving, setting restore\_best\_weights to True will restore the optimal weights recorded during the training and assign them to the model weights.

*Building the Model (Utils):*



We Use a deep feed-forward neural network with 5 hidden layers. We use a 30% [dropout](https://www.thepythoncode.com/article/dropout-regularization-in-pytorch) rate after each fully connected layer; this type of regularization will hopefully prevent overfitting on the training dataset.

We use a single output unit (neuron) with a [sigmoid activation function](https://en.wikipedia.org/wiki/Sigmoid_function) in the output layer; the model will output the scalar 1 (or close to it) when the audio's speaker is a male, and female when it's closer to 0

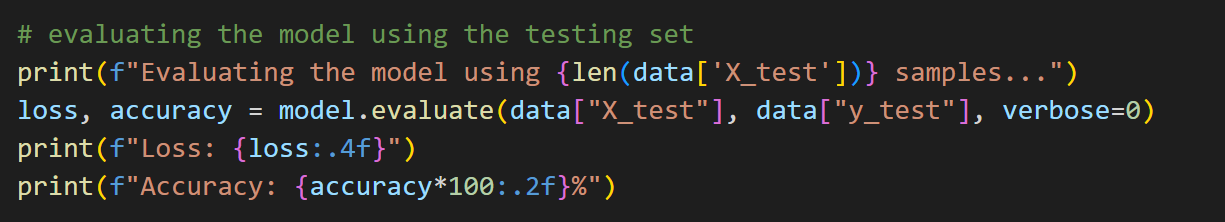
*Test the model with Voice:*

for the Test Model class, first thing we will use PyAudio to use for our voice input. Second, we will record the input voice into a file using the record\_to\_file command. Next we will have a function that will be responsible for loading the audio file and extracting features from it. Finally, we will predict

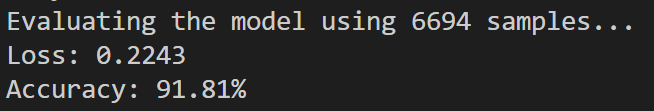
1. **Results and Discussion**

**Testing the Model.**

Since the model now is trained and the weights are optimal, let's test it using the testing set we created earlier:



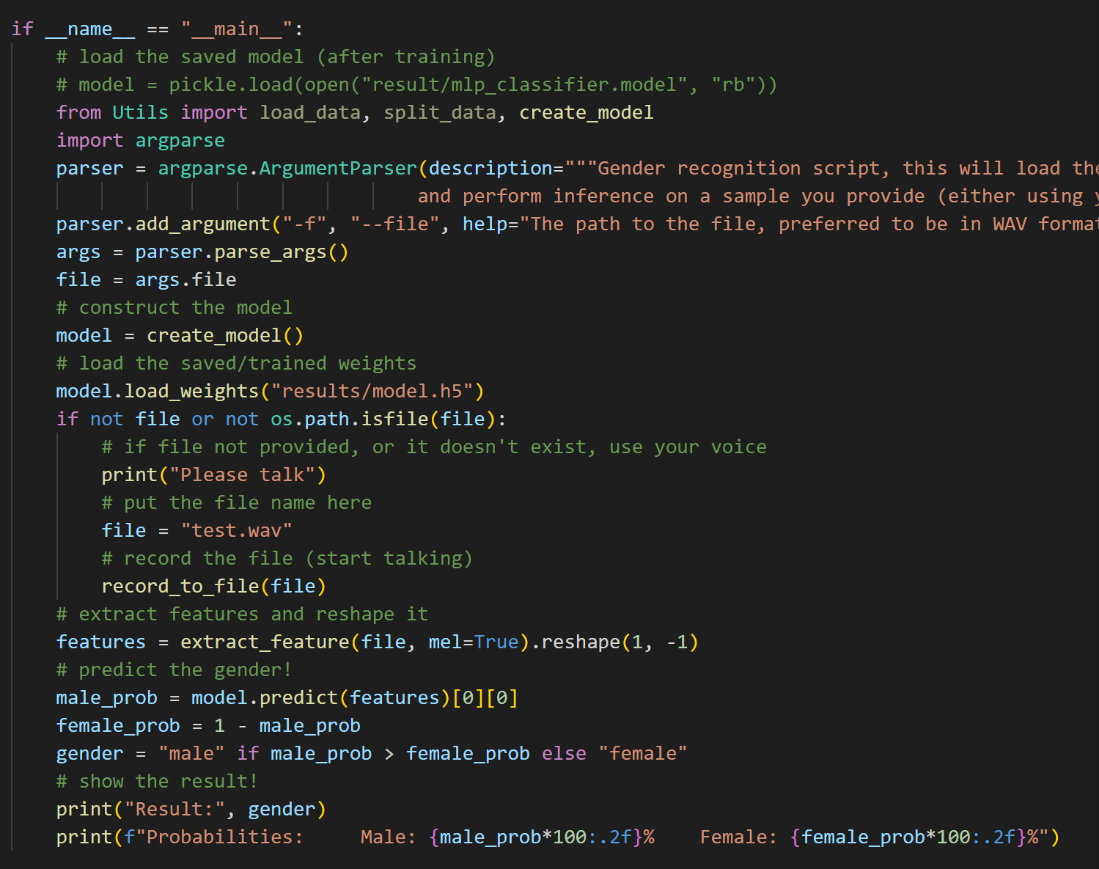
Output:

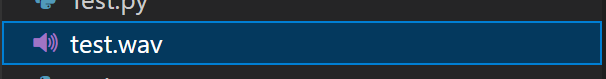


We've reached nearly 92% accuracy on samples at this time.

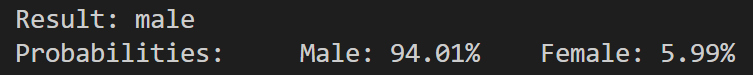
### **Testing the Model with my Voice.**

The below lines of code will use the ***argparse*** module to parse an audio file path passed from the command line and make inference on it:





Output:



And the result is correct. This is the male voice in my group.

1. **Conclusion and Perspectives**

In short, this project brings many benefits to today's society and with RandomForest and DecisionTree method will bring higher performance in Machine Learning Algorithm when training data with performance up to 0.99. Some algorithms can perform better than others for example with Decision Tree it has the strength that Input data can be missing data, no need to normalize or create dummy variables but can do work with both numerical and categorical data.

Regarding our project, there are still some limitations that for people who are livid, it is difficult for the machine to distinguish whether the voice signal is male or female, because we predict based on % and frequency if AI predicts that frequency is more similar to female frequency, then AI will decide it is female and vice versa. If we have time, we suggest exploring the opposite aspects of easy-to-communicate people. For example, the dumb people, the stammers, to be able to exploit the maximum efficiency of this project

**References**

1. D. Gabor, “Theory of communication. Part 1: the analysis of information,” *The journal of the Institution of Electrical Engineers. Radio and communication engineering*, vol. 93, no. 26, pp. 429–441, 1946.View at: [Publisher Site](https://doi.org/10.1049/ji-3-2.1946.0074) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Theory%20of%20communication.%20Part%201:%20the%20analysis%20of%20information&author=D.%20Gabor&publication_year=1946)
2. K.-I. Kanatani, *Group-Theoretical Methods in Image Understanding*, Springer Science & Business Media, vol. 20, 2012.
3. S. S. Stevens and J. Volkmann, “The relation of pitch to frequency: a revised scale,” *The American Journal of Psychology*, vol. 53, no. 3, pp. 329–353, 1940.View at: [Publisher Site](https://doi.org/10.2307/1417526) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=The%20relation%20of%20pitch%20to%20frequency:%20a%20revised%20scale&author=S.%20S.%20Stevens%20&author=J.%20Volkmann&publication_year=1940)
4. T. Qiao, S. Zhang, Z. Zhang, S. Cao, and S. Xu, “Sub-spectrogram segmentation for environmental sound classification via convolutional recurrent neural network and score level fusion,” in *2019 IEEE International Workshop on Signal Processing Systems (SiPS)*, pp. 318–323, Nanjing, China, 2019.View at: [Publisher Site](https://doi.org/10.1109/SiPS47522.2019.9020418) | [Google Scholar](https://scholar.google.com/scholar_lookup?title=Sub-spectrogram%20segmentation%20for%20environmental%20sound%20classification%20via%20convolutional%20recurrent%20neural%20network%20and%20score%20level%20fusion&author=T.%20Qiao&author=S.%20Zhang&author=Z.%20Zhang&author=S.%20Cao&author=&author=S.%20Xu)

Chart, box and whisker chart

Description automatically generated

Figure:Waveform for the Female voice

Chart, bar chart

Description automatically generated

Figure: Chroma for the Female Voice

Chart, box and whisker chart

Description automatically generated

Figure: Waveform for the Male voice

Chart

Description automatically generated

Figure: Chroma for the Male voice