Predicting Gender Based on Voice

By Group 18: Kartik, Advait, Nikhil, Harshvardhan

Can we predict the gender of a person based on their voice?

Dataset

2 kinds of data

For 2 kinds of analysis

- VoiceGender Dataset

  For training and testing models for classification on the basis of extracted features.

  Csv file consisting of features and labels

- VoxCeleb Dataset

  To implement our own feature engineering.

  .wav files for each gender.

Experiments

VoiceGender

1. The input vectors were 20 dimensional consisting of features like mean frequency, kurtosis etc. Roughly 3000 samples in all. Around 1600, or 53% of which were male samples

2. There was a unique label, either ‘male’ or ‘female’

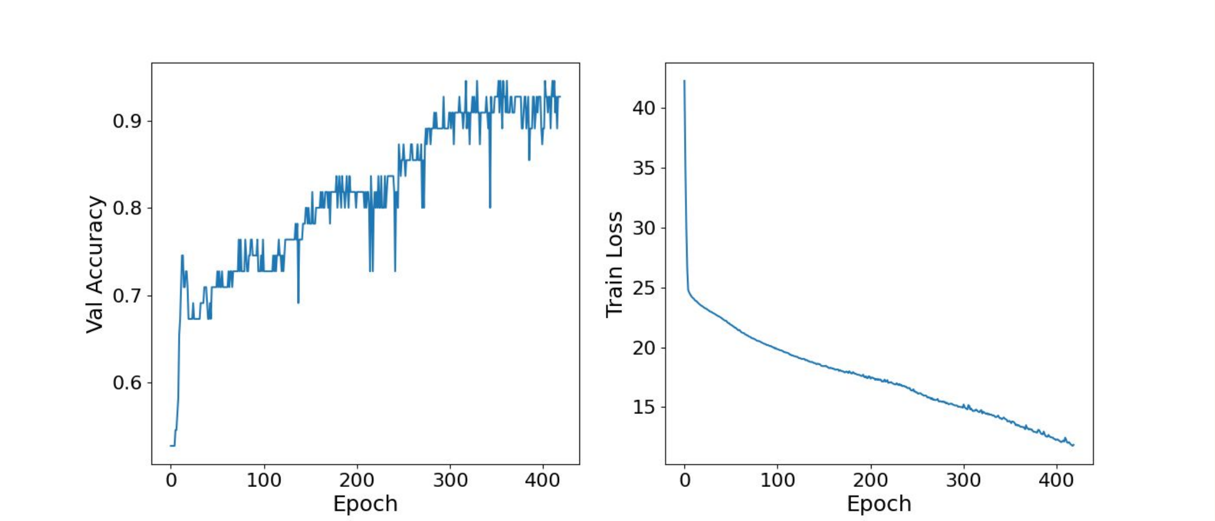
3. Thus, our problem was finding a hypothesis H, such that H(X) = Y for X as vector in the feature space and Y being the label

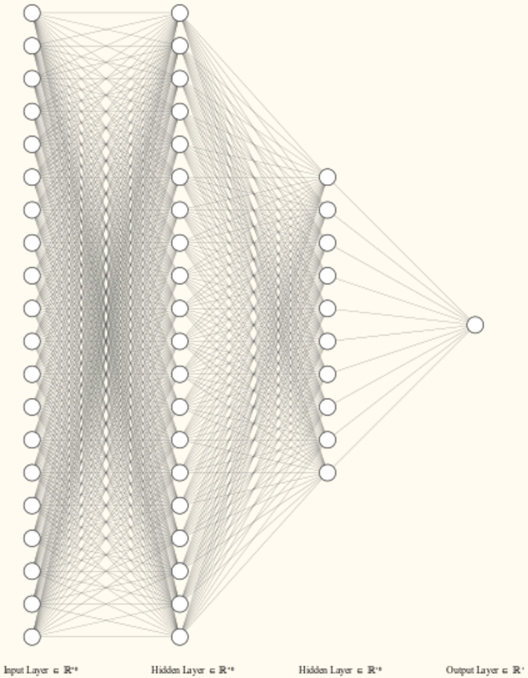
4. We used a multilayer perceptron to model the hypothesis function H.

5. We also used Support Vector Machines using Kernels to transform and linearly separate the samples of each label in the feature space.

VoiceGender - MultiLayer Perceptron

We used a multilayer (2 hidden layers) perceptron to model the hypothesis function H. Using the binary cross entropy loss function and the Adam optimiser, we attained an accuracy of 93% on the validation dataset.



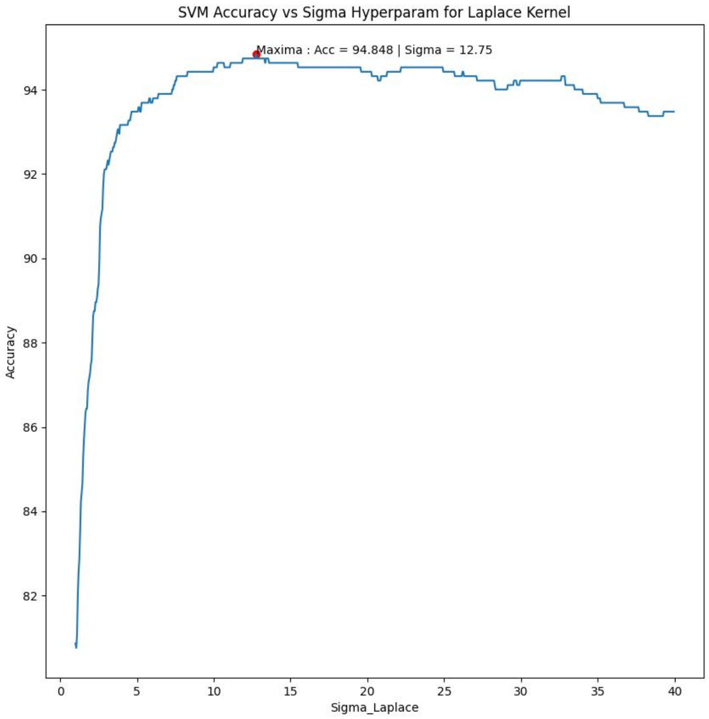

Input Layer ∈ $\mathbb{R}^{20}$
Hidden Layer ∈ $\mathbb{R}^{20}$
Hidden Layer ∈ $\mathbb{R}^{10}$
Output Layer ∈ $\mathbb{R}^{1}$

VoiceGender - SVM with kernels

We used 4 different kernels to transform and separate the data.

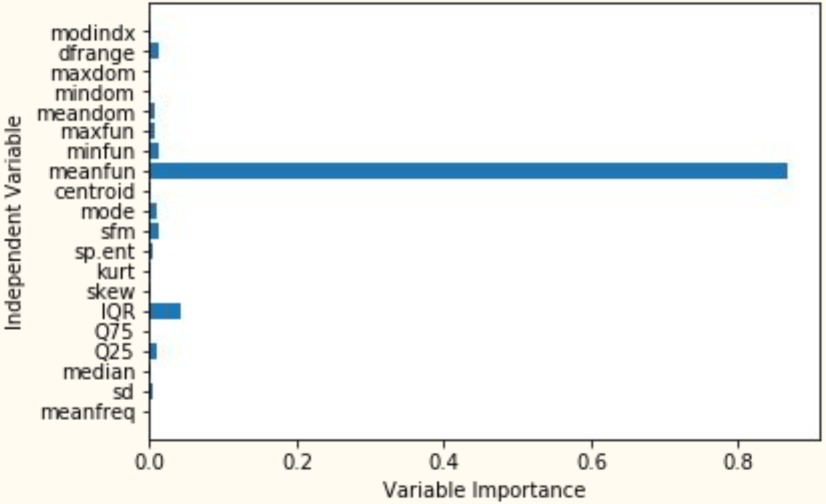
- RBF
- Laplace
- Gaussian
- Polynomial

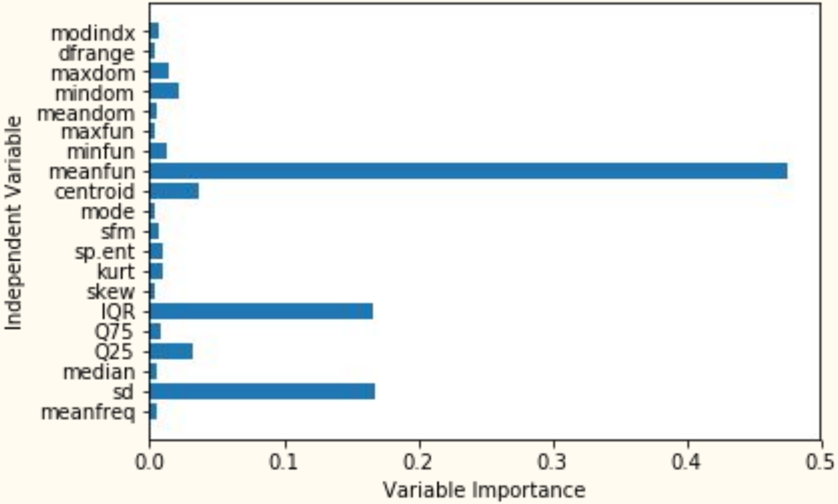
Accuracy obtained: 94% validation accuracy



VoiceGender - Decisions Trees & Random Forests

We receive roughly 96% and 97% validation accuracy on the dataset using decision trees and random forests respectively. The results on the left are by decision trees and on the right by random forests.





VoxCeleb - Feature Engineering

- The MFCC are state-of-the-art features for analysing audio sample data
- Start by taking a short window frame (20 to 40 ms) in which we can assume that the audio signal is stationary. We then select a frame step of around 10 ms.
- We then compute the power spectrum of each frame through a periodogram. To do so, start by taking the Discrete Fourier Transform of the frame.
- We then take the logarithm of the all those 26 series of energy of those filter-banks since we do not perceive loudness linearly, but close to logarithmically.
- We finally apply a Discrete Cosine Transform to the 26 log filterbank energies in order to decorrelate the overlapping filterbanks energies. This gives us 26 coefficients, called the MFCC.
- We extract the mean, variance, quartiles, median etc as a descriptive statistic at the end of an audio sample, and compare several audio samples on this basis.

VoxCeleb - Model

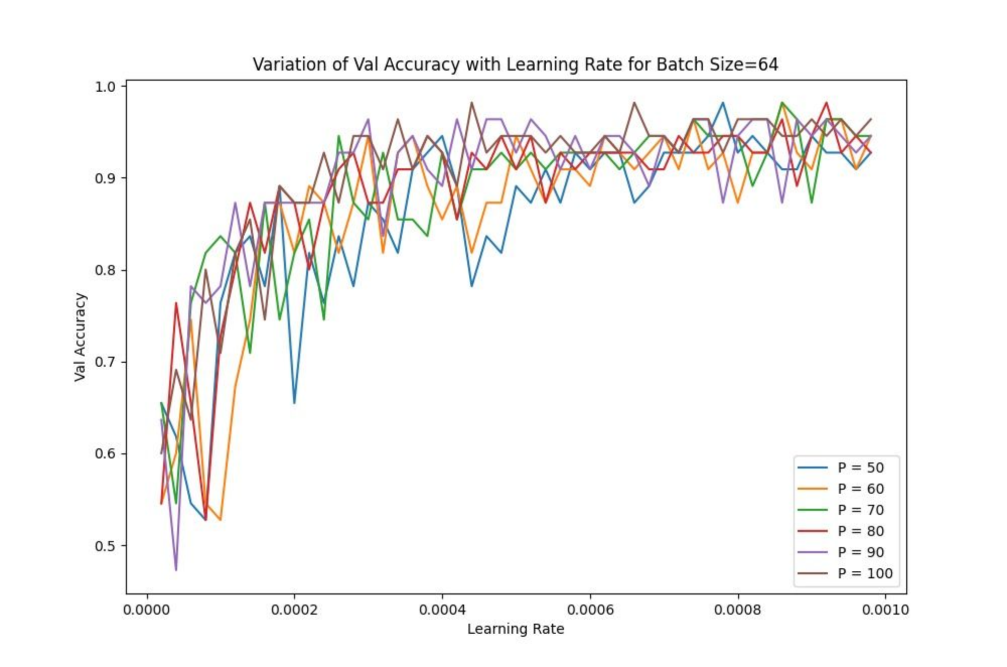
Once the feature engineering was completed, we had vectors in the feature space. We simply trained a classification model on these feature vectors.

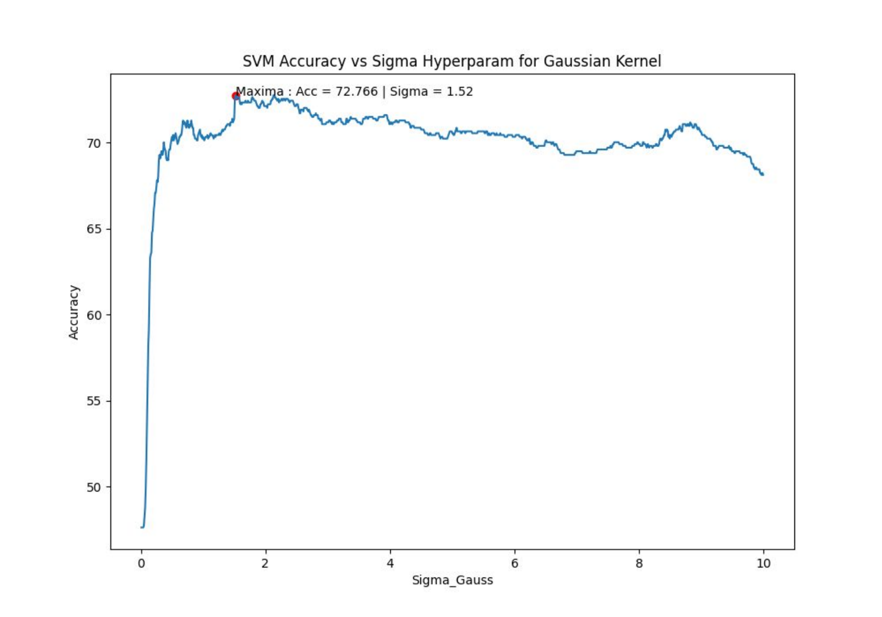
Models tried: -

- Decision tree
- K Nearest Neighbors
- Random forest
- Support Vector Machine
- Multilayer Perceptron

Hyperparameter Tuning

Using methods like grid search, we tuned the models across various parameters





Results

Summarising all Results

VoiceGender

- Multilayer Perceptron: 0.93
- SVM: 0.94
- Decision Trees: 0.96
- Random Forests: 0.97

VoxCeleb

- Decision tree accuracy: 0.76
- Logistic regression accuracy: 0.89
- Hard voting accuracy: 0.89
- K Nearest Neighbors accuracy: 0.73
- Random forest accuracy: 0.88
- svm accuracy: 0.79

Note: All values are measured on the validation set

Conclusion

The voice of a person is a complicated data sample, which is affected by many factors. However, there is a distinct correlation between gender and certain extractable features of voice which makes it possible for us, with reasonable accuracy, to predict gender on the basis of voice.

This ability to predict gender becomes increasingly accurate with more number of features and the model's ability to learn and generalise the provided data, as seen in both datasets.

Thank You!