REPORT: Identifying the Gender of a Voice using Unsupervised learning approach

By Emuejevoke Eshemitan

Objectives

The objective of this project is to use several unsupervised machine learning algorithm to group voices. We would be focused on clustering the dataset into male or female.

About Dataset

This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech. The dataset consists of 3,168 recorded voice samples, collected from male and female speakers.

Data Source: https://www.kaggle.com/datasets/primaryobjects/voicegender

Attribute Information: The following acoustic properties of each voice are measured:

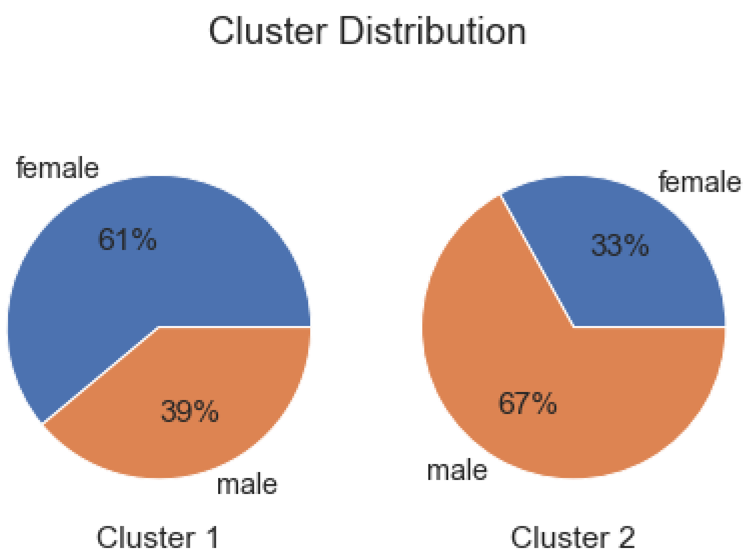
- duration: length of signal
- meanfreq: mean frequency (in kHz)
- sd: standard deviation of frequency
- median: median frequency (in kHz)
- Q25: first quantile (in kHz)
- Q75: third quantile (in kHz)
- IQR: interquantile range (in kHz)
- skew: skewness (see note in specprop description)
- kurt: kurtosis (see note in specprop description)
- sp.ent: spectral entropy
- sfm: spectral flatness
- mode: mode frequency
- centroid: frequency centroid (see specprop)
- peakf: peak frequency (frequency with highest energy)
- meanfun: average of fundamental frequency measured across acoustic signal
- minfun: minimum fundamental frequency measured across acoustic signal
- maxfun: maximum fundamental frequency measured across acoustic signal
- meandom: average of dominant frequency measured across acoustic signal
- mindom: minimum of dominant frequency measured across acoustic signal
- maxdom: maximum of dominant frequency measured across acoustic signal
- dfrange: range of dominant frequency measured across acoustic signal
- modindx: modulation index. Calculated as the accumulated absolute difference between
- abel: male or female

Summary of Data Exploration and actions taken for data cleaning

- Some features in the data was skewed so I had to perform logrithmic transformation on highly skewed data.
- Because we are using unsupervised approach, I also had to drop the target column before clustering.
- Other than the above actions taken, the dataset used for this project didn't require further cleaning.

Summary of training at least three variations of the unsupervised model

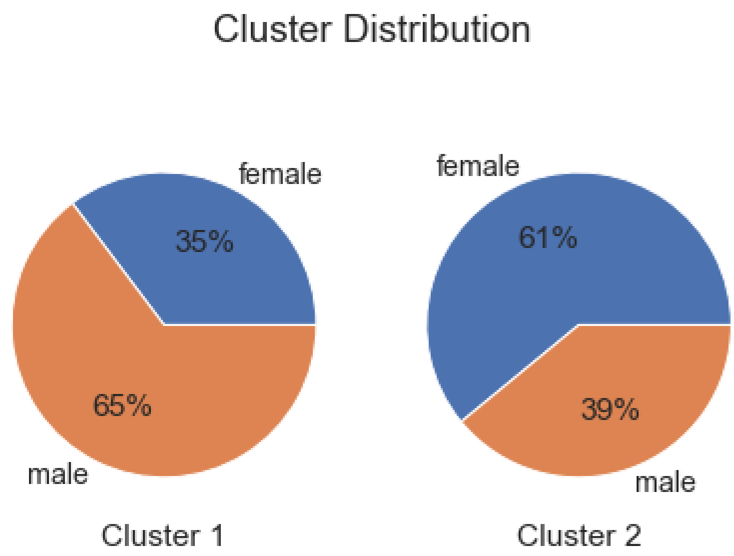
> In this project, we applied three(3) unsupervised machine learning algorithm
>
> - **KMeans Algorithm:** Here we selected the number of cluster to be 2 with random state of 42. The result for clustering the data is visualized below:



3]:

kmeans	label	number
0	female	1170
	male	746
1	female	414
	male	838

- **Agglomerative Clustering:** We can deduce from the above information that although the Agglomerative algorithm has done a good job in clustering this data, but the Kmeans algorithm has done a better job. The result for clustering the data is visualized below:



[19]:

agglom	label	number
0	female	477
	male	875
1	female	1107
	male	709

- **DBSCAN:** After several hyperparemeter changing, I discovered that we cannot use dbscan for this problem because we cannot explicitly specify the number of cluster we want

Recommended Unsupervised Learning model best fits the need

- We can visualize from our analysis that the KMeans algorithm did better at clustering the data with
  - Cluster 1 = 61% female and 39% male and
  - Cluster 2 = 33% female and 67% male
- KMEANS algorithm is considered the best for this project.

Summary Key Findings and Insights

- Most people have maximum of dominant frequency less than 10
- Some of the features are correlated with each other
- KMEANS algorithm is considered the best for this project

Suggestions for next steps in analyzing this data

- Maybe Clustering the data into more than 2 clusters might be better
- Using PCA might make the model perform better.