STAT 451 Project Proposal - Henry N, Gavin J, Kaitlyn S, Kyle Z, Laura K

Proposal/Background:

Our group will be utilizing data collected from a historically important acoustics study from Peterson and Barney (1952) that measured several different vowel sounds from 72 different speakers. 1520 sounds were recorded in total. Formant measurements were taken from each sound. These measurements are local maximums in the spectrum of speech being produced. They can also be thought of as overtones above the fundamental frequency or pitch being produced by a human's voice. Our goal is to use these measurements to predict which vowel sound was produced by those frequencies.

Data Source:

(R Package): https://rdrr.io/cran/phonTools/man/pb52.html

R code for installing package and saving data as a .csv file:

```
install.packages("phonTools")
library(phonTools)
data(pb52)
write.csv(pb52, "/home/henrynomeland/pb52.csv", row.names=F) #any file location
```

Python code for loading in the .csv file as a pandas dataframe:

```
import pandas as pd
df = pd.read_csv("/home/henrynomeland/pb52.csv") #previous file location
```

Questions:

Can the frequency measurements from recordings of vowel sounds accurately determine which vowel sounds are being spoken?
Which models/hyperparameters best categorize sounds into vowel labels?

Variables:

Vowel - target/class/dependent variable - sounds represented by the X-SAMPA phonetic alphabet
Sex - m (male) / f (female)
Speaker - integer representing the id of the particular subject being recorded
Repetition - within each speaker, the numbered time that they said a particular sound
F0 - the fundamental frequency - measured in Hz
F1/F2/F3 - features - the three lowest formant measurements for each vowel sound - measured in Hz

Methods:

We will begin by evaluating a subset of the vowel sounds (perhaps a, i, and u), which will then be split up into training, validation, and testing subsets.
We will train several models (perhaps SVM, KNN, Individual Decision Tree, Random Forests), tuning hyperparameters for optimal validation accuracy. We will then compare model results using test data.