Report

1. Introduction

The chosen dataset - Gender Recognition by Voice
The link to Kaggle dataset is:
https://www.kaggle.com/primaryobjects/voicegender

We plan to use the following csv file: **voice.csv** file for our project.

2. Problem Definition and Algorithm

2.1 Task Definition

This database was created to identify a voice as male or female, based upon acoustic properties of the voice and speech.

2.2 Dataset Description

Number of the features: 20

Number of instances: 3,168 recorded voice samples, collected from male and female speakers.

Data distribution:

The given dataset contains the following attributes:

- **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 adjacent measurements of fundamental frequencies divided by the frequency range
- **label**: male or female

Pre-Processing: The dataset has no null valued attributes, null values are not handled moreover we also checked for near zero variance and found none.
The label is non numeric thus we factor the label column. The results of the above are placed in preprocess.txt file

3. Experimental Evaluation

3.1 Methodology

We have trained and compared the results of 5 calssifiers namely :- Svm, RandomForest, Knn, Ann and Boosting. On the basis of this analysis, the one with the best results is considered to be the most appropriate for our dataset. The analysis is in the result below.

3.2 Results

CLASSIFIER	Parameter1	Parameter2	Best Accuracy
SVM	cost=1	gamma=0.5	97.47
Random Forest	ntree=500	mtry=4	97.79
k-NN	k=7	NA	97.47
ANN	size=7	decay=0.001	99.05
Boosting	mfinal=50	maxdepth=10	98.39

3.3 Discussion

As Ann gives the best results on the above analysis, we thus conclude that for gender recognition on voice input Artificial Neural Networks work the best.

4. Related Work

Most of the related work is done python and is mainly on using set of kernels for data visualization.

5. Conclusion

We as a team achieved the goal of implementing the above mentioned classifiers on our dataset and recorded the one that works the best.

6. Reference

We used the help feature in R studio to overcome our hurdles.