Status Report:

Data Mining on Voice Disorder

JIARUI LI • 04.10.2021

Overview

Data Processing

- The demographic sheet
- Voice Conditions Measures
- Data Visualization

Simple Model Fitting

- Support Vector Machine
- Random Forest
- LogisticRegression
- K-Fold Cross Validation

Next Step

- . . .

Progress - Data Processing

Part-One [Information Understanding]

- ***The demographic sheet*** (Age, Gender, Education Level, Smoking Experience, Medical history, Choir Experience, etc.)

- ***Voice Condition measures*** (Picture Description, Grandfather Reading Passage, Conversation

Progress - Data Processing

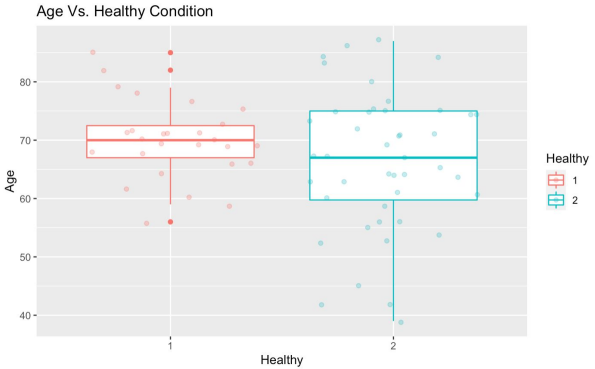
The demographic Sheet

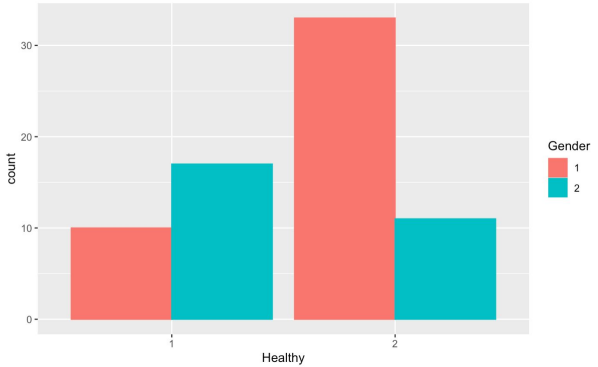
- Drop uncorrelated attributes (Date, Education-Level, Medication status, etc.). Next, data errors are cleaned

- To better find the relationship between Ages and Healthy conditions, various ages are divided into different age groups.

- Since work types may trigger voice-related problems, different categories are created.

Data Visualization





Discover-One

1. In this dataset, people in the Healthy group are greater than in the disease groups.
2. Mainly people who have the disease are in their old age.

Discover-Two

1. Male who has been infected the disease are outnumbered than women who have infected the disease from this dataset.

Measurements on Voice

Scenarios - 2

Picture Description

- Min/Max Energy (dB)
- Mean Frequency (HZ)
- etc.

Scenarios - 4

Conversation

- F0-Conversion
- Range
- etc.

Scenarios - 1

Pitch prolonged /ah

- Min/Max Pitch (HZ)
- Mean Frequency (HZ)
- etc.

Scenarios - 3

Reading Passage

- Min/Max Energy (dB)
- Min/Max Frequency (HZ)
- Periodicity, etc.

Dataset

ID	Healthy	Choir	Age	age_group	Gender	Height	Ethnicity	Work	Smoke	singing activity	Exercise	UPDRS_Motor	UPDRS_Non_m
1	2	1.50	80	3	2	178	1	1	2	2	1	0	
2	2	0.00	86	3	2	166	2	1	2	2	1	0	
3	2	0.00	87	3	1	150	2	2	2	2	1	0	
4	1	1.00	71	2	1	160	1	2	2	2	0	7	
5	1	1.00	70	2	1	157	2	2	2	2	0	12	
6	2	0.00	64	2	1	166	1	2	2	2	0	0	
7	1	1.00	68	2	2	180	1	1	1	2	0	17	
8	2	0.00	71	2	1	161	1	1	2	2	0	0	
9	2	0.75	67	2	1	158	1	2	2	2	0	0	
10	1	0.00	60	2	2	173	2	1	2	2	0	5	

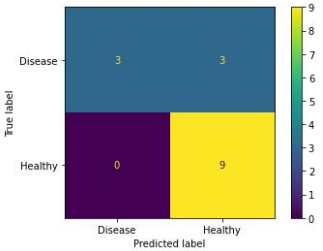
...

	Periodicity	Semitone.Range	Std..Deviation..Semitones.	Maximum.Pitch..Hz..1	Minimum.Energy..dB..1	Maximum.Energy..dB..1	Mean.Frequency..Hz..1
23	4.19	18.0	4.02	366.00	25.10	79.18	137.47
30	1.9	21.1	5.10	349.10	31.29	87.27	159.47
25	0.72	17.0	4.52	371.20	34.00	78.30	171.00
42	2.6	26.0	6.96	346.40	32.60	76.80	156.80
30	2.6	21.0	6.55	367.50	25.80	80.98	180.44
11	10.29	21.0	2.29	384.74	28.77	78.73	189.73
23	1.96	29.0	4.16	295.67	32.83	83.07	126.52
30	0.31	25.0	5.77	395.62	28.61	81.51	186.93
39	0.4	26.0	7.85	392.48	29.03	78.09	180.39
39	1.4	29.0	7.99	356.13	31.29	78.13	115.00

Machine Learning Techniques

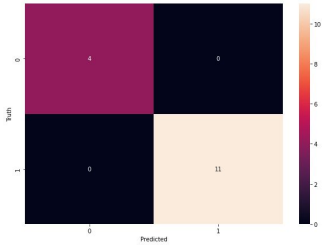
Support Vector Machine (SVM)

- “C” = 0.5
- “Kernel” = “Linear”



Random Forest Algorithm (RF)

- “N_estimator” = 10



LogisticRegression

- C = 1

K-Fold Cross Validation

- CV = 3...

Next steps

- Come up with a suitable model to determine which features are significant for explaining the results (R).

- ...