Cough Data Classification for COVID

Link to the paper: [**https://www.nature.com/articles/s41597-021-00937-4#Sec7**](https://www.nature.com/articles/s41597-021-00937-4#Sec7)

**Problem Identification / Introduction:**

The COUGHVID dataset provides over 25,000 crowdsourced cough recordings representing a wide range of participant ages, genders, geographic locations, and COVID-19 statuses. Over the course of this pandemic many people have contracted COVID-19 and developed a cough so machine learning can help screen individuals in order to relieve some of the strain on the medical field. The goal is to build a cough detection model with the recorded sound datasets that characterizes whether a person has a respiratory illness or not based on provided personal information. The benefit of using this dataset is that there are 2,800 recordings labeled by four experienced physicians who diagnosed medical abnormalities present in the cough datasets.

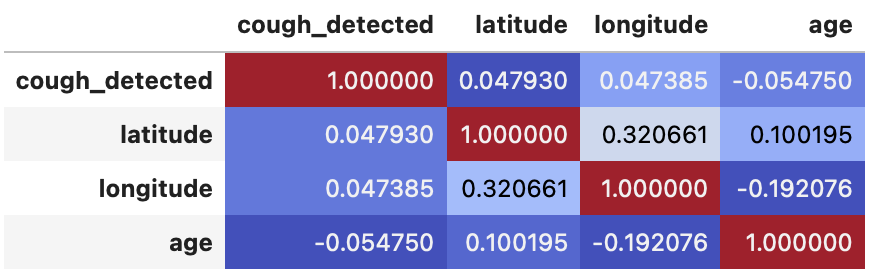
**Data Wrangling:**

The data came in a folder structure with a json file for the patients information adjacent to a ‘.ogg’ or ‘.webm’ file for the sound data. The json data was first imported into a dataframe and the sound files loaded into a list and then converted into .wav format. Cleaning the raw data the expert labeled columns had variables in a dictionary format that I expanded into multiple columns in order to reduce dimensionality.

**Exploratory Data Analysis:**

Sound data is a mixture of different wave frequencies at different intensities so in order to understand it I plotted a single .wav file to explore with spectra analysis in order to perform feature extraction. The spectra centroid found the center mass of the audio and the spectral bandwidth is the range of interest around the centroid. The mel-frequency cepstral coefficients model the characteristics of the human voice. The spectral rolloff analyzes the waveform suddenly dropping from high to 0, it helps to distinguish between harmonic and noisy sounds. The root mean square is used to understand the pressure of the wave and relates to the energy carried by the sound wave. Once this data for these sound features was collected we took a random sample of the audio data and found the averages for each and the maximum value.

The longitude and latitude of the participants' personal information were extracted to plot a correlation matrix looking at age, latitude, longitude, and cough\_detected. The cough\_detected column was provided by the researchers who cleaned up the data to provide a column to distinguish if it was cough sound or not. Overall what we saw was no correlation between any of these features.



**Preprocessing:**

The training data was:

* The target variable (y) was the COVID diagnosis
* The feature variables (X) was all the information minus COVID diagnosis (the four expert labeling, sound features, and personal data about the patients)

**Modeling Procedures:**

Experiment one:

* Ran the training data through a few model types: Random Forest Classifier, Ridge Classifier, KNN, Decision Tree, Naive Bayes
* The AUC scores where used as the primary evaluation matrix with random forest classifier and Naive Bayes having the highest scores
* Ridge Regression and Decision Trees had 0.99> AUC score meaning that these models were able to perfectly classify observations into classes, showing signs of feature leakage

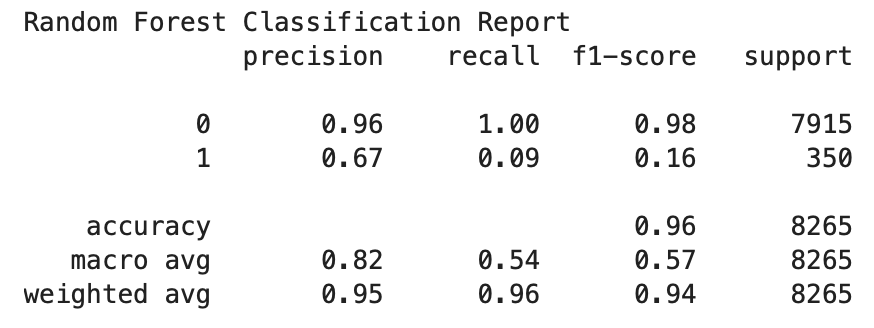
Experiment two:

Fixing the data that is leaking through by removing the status\_healthy column and moving to better performance metrics

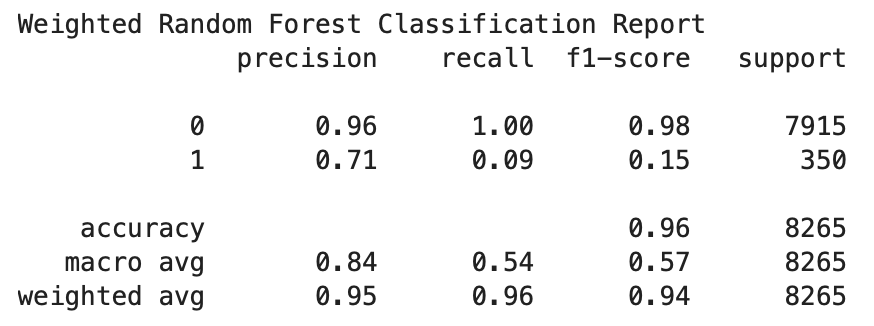
* Previously we used accuracy to determine model effectiveness which would not work since the data is unbalanced
* Instead this time we used a classification report looking at the f1-score as the harmonic mean of precision and recall giving us better intuition of prediction results
* Based on this score the ideal models for this data would be random forest classifier and decision tree

Confusion Matrix & Classification Reports:

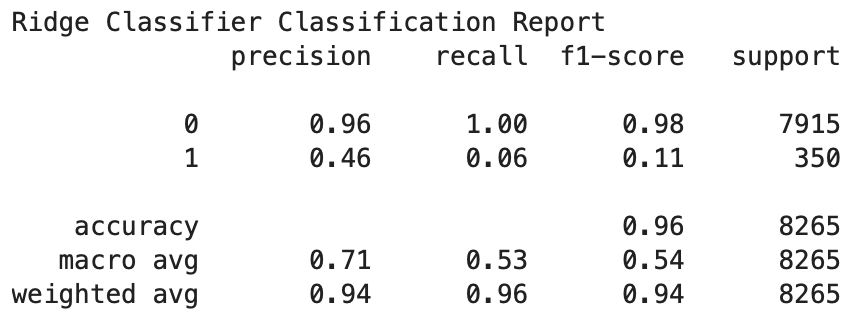
| **Random Forest Classifier** |  | **Actual** | |
| --- | --- | --- | --- |
|  |  | Not COVID (0) | COVID (1) |
| **Predicted** | Not COVID (0) | 7900 | 320 |
| COVID (1) | 15 | 31 |



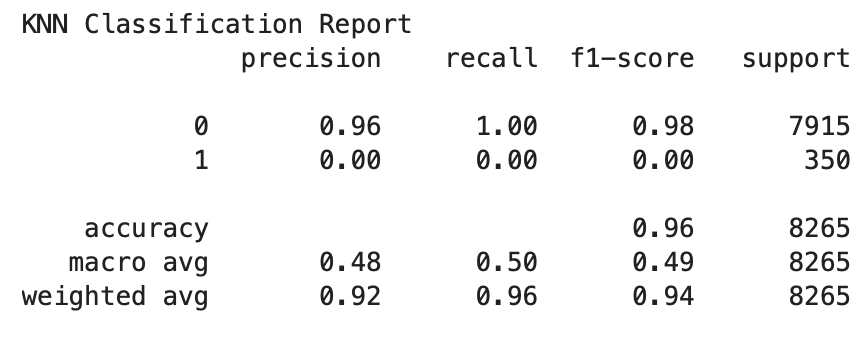
| **Weighted Random Forest Classifier** |  | **Actual** | |
| --- | --- | --- | --- |
|  |  | Not COVID (0) | COVID (1) |
| **Predicted** | Not COVID (0) | 7900 | 330 |
| COVID (1) | 12 | 30 |



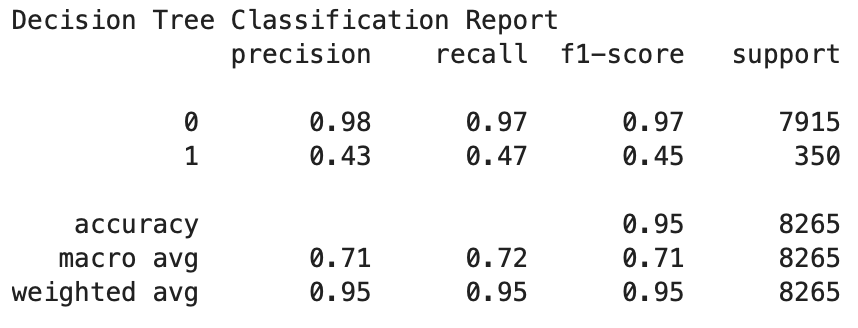
| **Ridge Classifier** |  | **Actual** | |
| --- | --- | --- | --- |
|  |  | Not COVID (0) | COVID (1) |
| **Predicted** | Not COVID (0) | 7900 | 330 |
| COVID (1) | 26 | 22 |



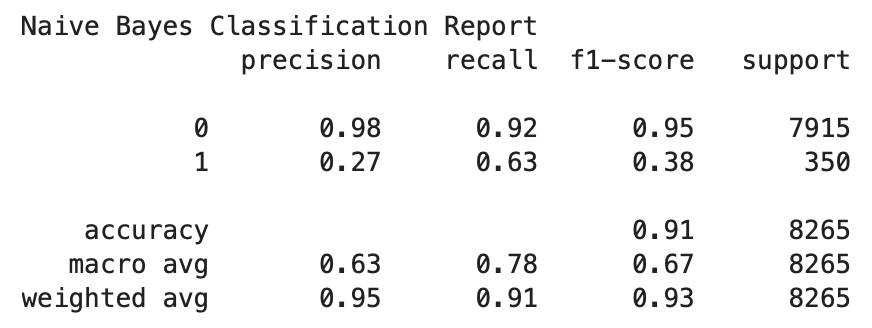
| **KNN** |  | **Actual** | |
| --- | --- | --- | --- |
|  |  | Not COVID (0) | COVID (1) |
| **Predicted** | Not COVID (0) | 7900 | 350 |
| COVID (1) | 2 | 0 |



| **Decision Tree** |  | **Actual** | |
| --- | --- | --- | --- |
|  |  | Not COVID (0) | COVID (1) |
| **Predicted** | Not COVID (0) | 7700 | 190 |
| COVID (1) | 210 | 160 |



| **Naive Bayes** |  | **Actual** | |
| --- | --- | --- | --- |
|  |  | Not COVID (0) | COVID (1) |
| **Predicted** | Not COVID (0) | 7300 | 130 |
| COVID (1) | 600 | 220 |



**Findings and Recommendations:**

Initially testing the dataset against the models gave us some overfitting due to data leakage from the one-hot-encoded status column. To overcome this, some of the data had to be excluded from the training set. While using accuracy score as the first metric of choice seemed reliable, after further inspection of the feature importance it was revealed that we aren't dealing with a balanced dataset so it would be better to use a classification report. In the classification report, the precision is the ability of the classifier not to label something as positive when it is negative, and the recall is the ability of the classifier to find all the positive samples.

Ranking the models based on the f-score, Decision Tree, Naive Bayes, and Random Forest are the most accurate for this dataset. Comparing the unweighted Random Forest classifier to the weighted, the only difference is the precision score, with 67% for unweighted vs 71% for weighted in predicting that a person has COVID. But overall the f-score is one percent higher for the unweighted, compared to the weighted. By looking at the cost-sensitive learning instead of trying to optimize the accuracy, the problem now is to minimize the total misclassification cost for the imbalanced dataset. Other metrics that will benefit to understand the current results from the models are sensitivity and specificity.

For future directions I would recommend testing other combinational variables, where we work on hyperparameter optimization with features from the sound data. To process the spectrogram images of the sound data, we can build a CNN classification and generate a feature map to then input into a linear classifier layer. During the training we may vary our learning rate with functions that process a batch per iteration. The goal is to be able to classify sound as well as expert diagnostics.