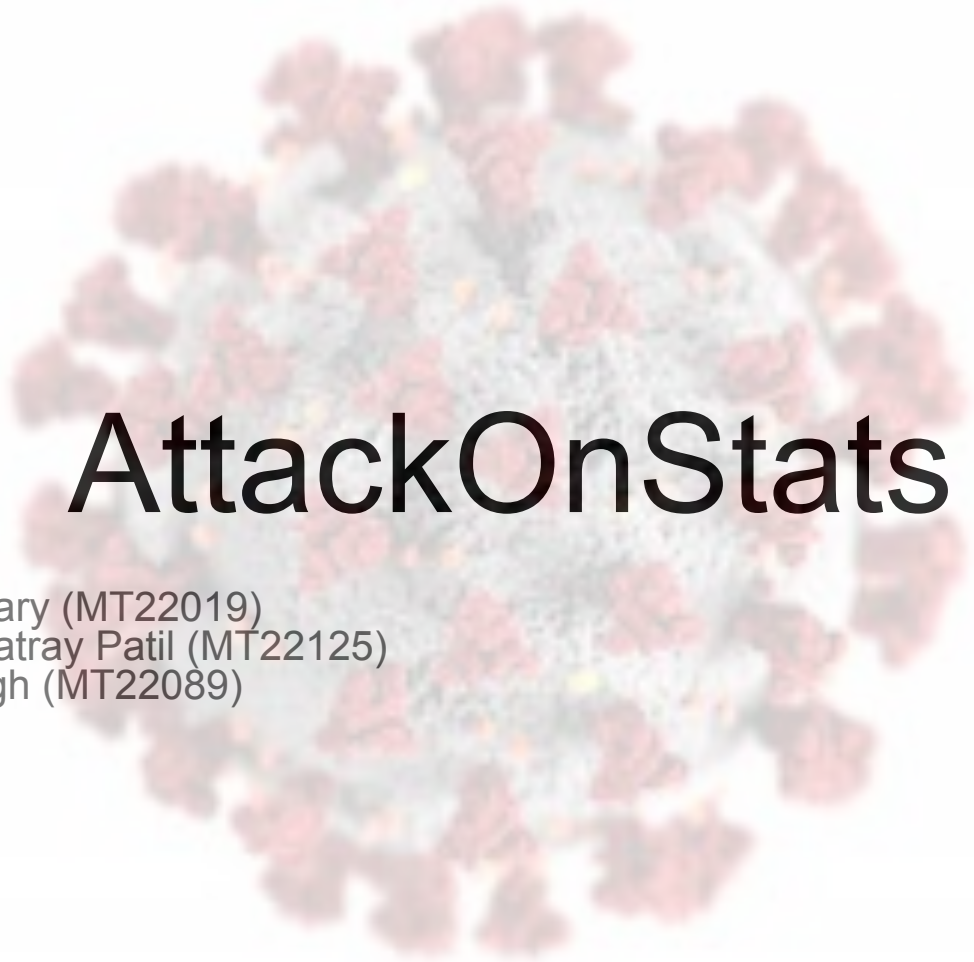


AttackOnStats

- Aryan Chaudhary (MT22019)
- Shubham Dattatray Patil (MT22125)
- Vimal Kirti Singh (MT22089)

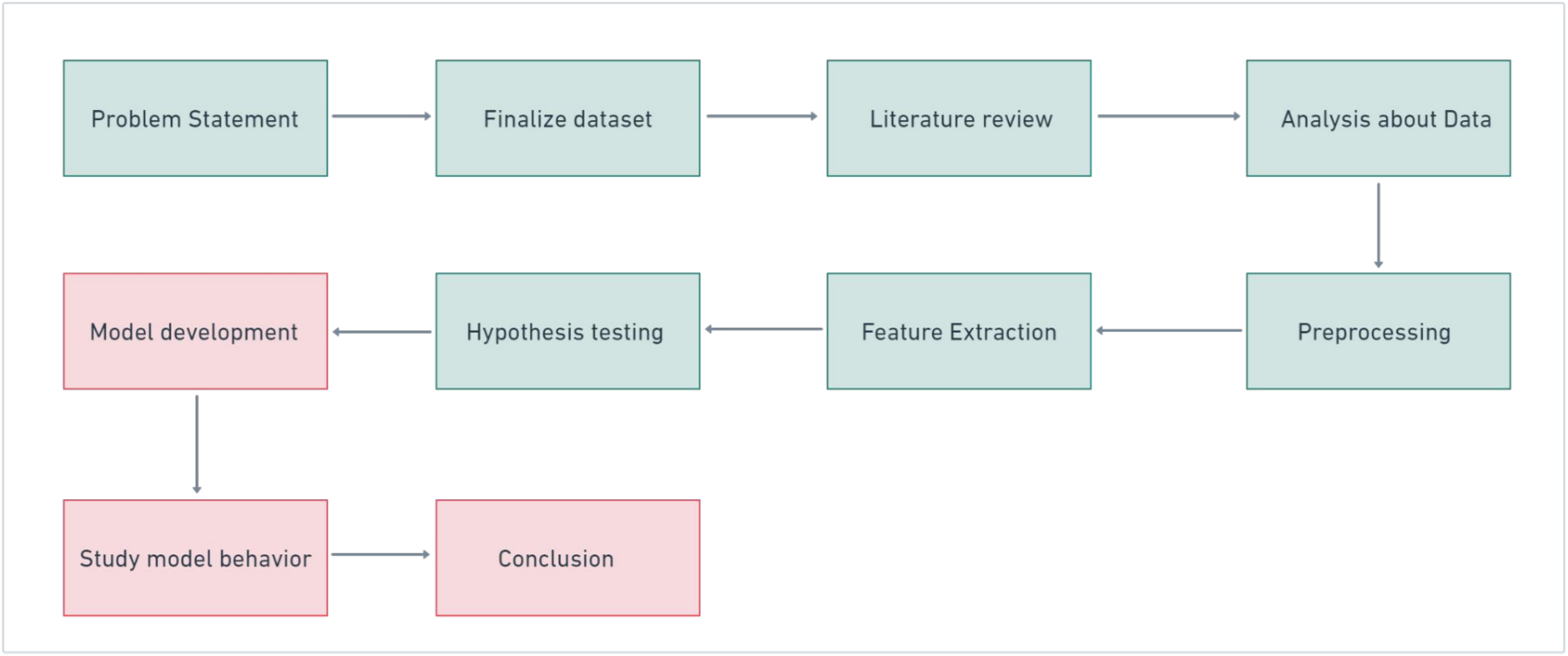
Problem Statement

Predicting whether a person has COVID-19 based on cough audio and diagnostic data.

Importance

1. Widespread testing became a significant bottleneck.
2. Swab tests are invasive, expensive, and time consuming
3. The time required to receive test results is significant
4. Contamination risk is high when individuals travel to testing sites to obtain their tests
5. Tests need to be administered by trained clinicians, severely limiting throughput

Plan of action



Hypothesis 11

Hypothesis: Cough Detection and COVID-19 Status

- Null Hypothesis (H0): There is no statistically significant difference in the "Cough_detected" between individuals with COVID-19 and those without COVID

- Alternative Hypothesis (H1): There is a statistically significant difference in the "Cough_detected" between individuals with COVID-19 and those without COVID

ANOVA results:

	df	sum_sq	mean_sq	F	PR(>F)
status	2.0	10.207364	5.103682	47.282369	3.555422e-21
Residual	11311.0	1220.914816	0.107940	NaN	NaN

Tukey's HSD results:

Multiple Comparison of Means - Tukey HSD

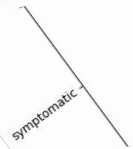
group1	group2				
COVID-19	hea				
COVID-19 symptom				54	True
healthy symptoma				0.023	0.0175 False

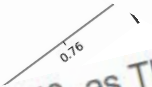
Hypothesis: Cough Type and COVID-19 Status

- Null Hypothesis (H0): Cough Type is not associated with COVID-19 status.

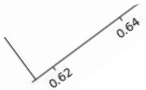
- Alternative Hypothesis (H1): There is a statistically significant association between Cough Type and COVID-19 status

Hypothesis 6





We applied ANOVA test and we cannot reject H0, as There is **no** statistical difference between mean of the groups of Zero Crossing Rate and COVID-19 status

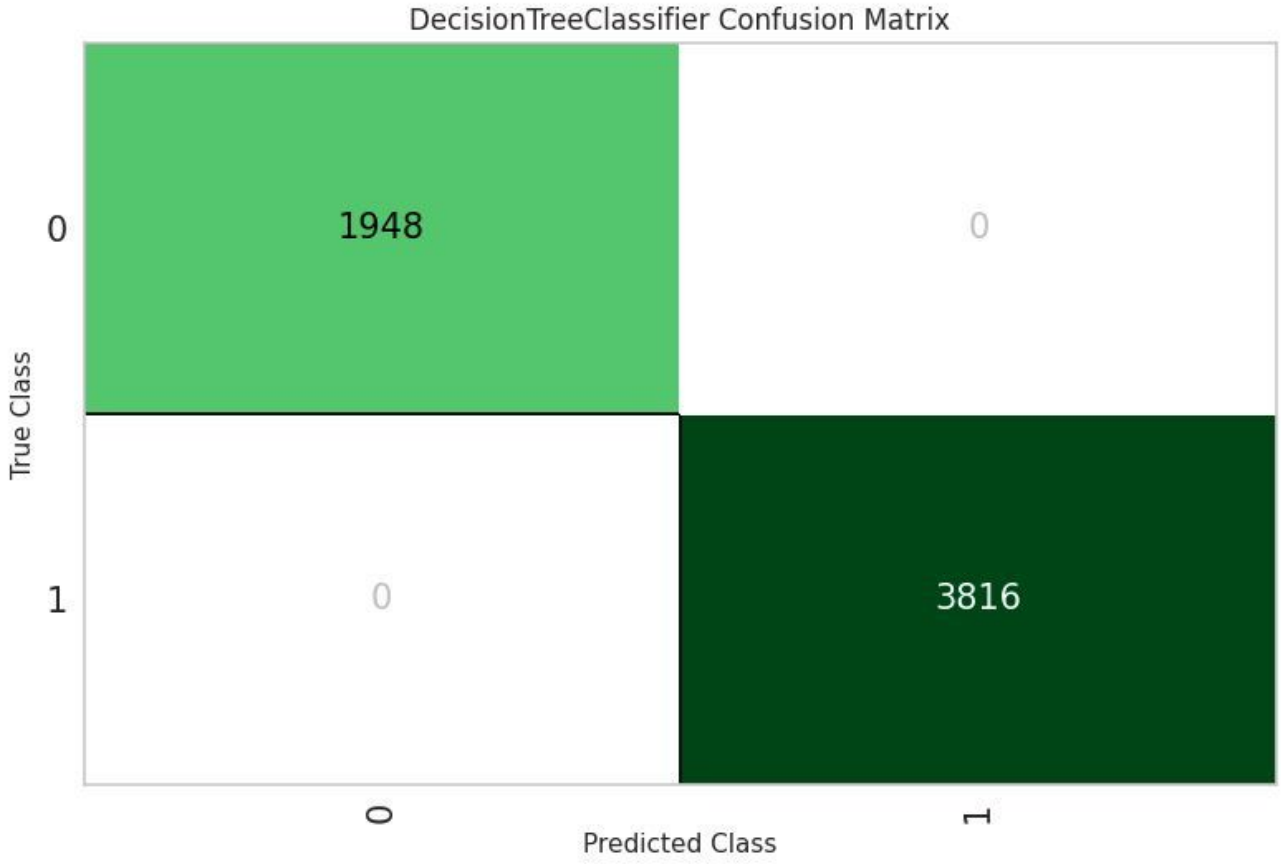


Important Features - Post Hypothesis Testing

1. Season
2. Age Group
3. Geographical Clusters
4. Weather Data
5. Severity of Symptoms
6. Spectrogram
7. MFCC
8. Spectral Bandwidth
9. Chroma Feature
10. Spectral Features ……. And 58 other features

Model Development: Cough Detection

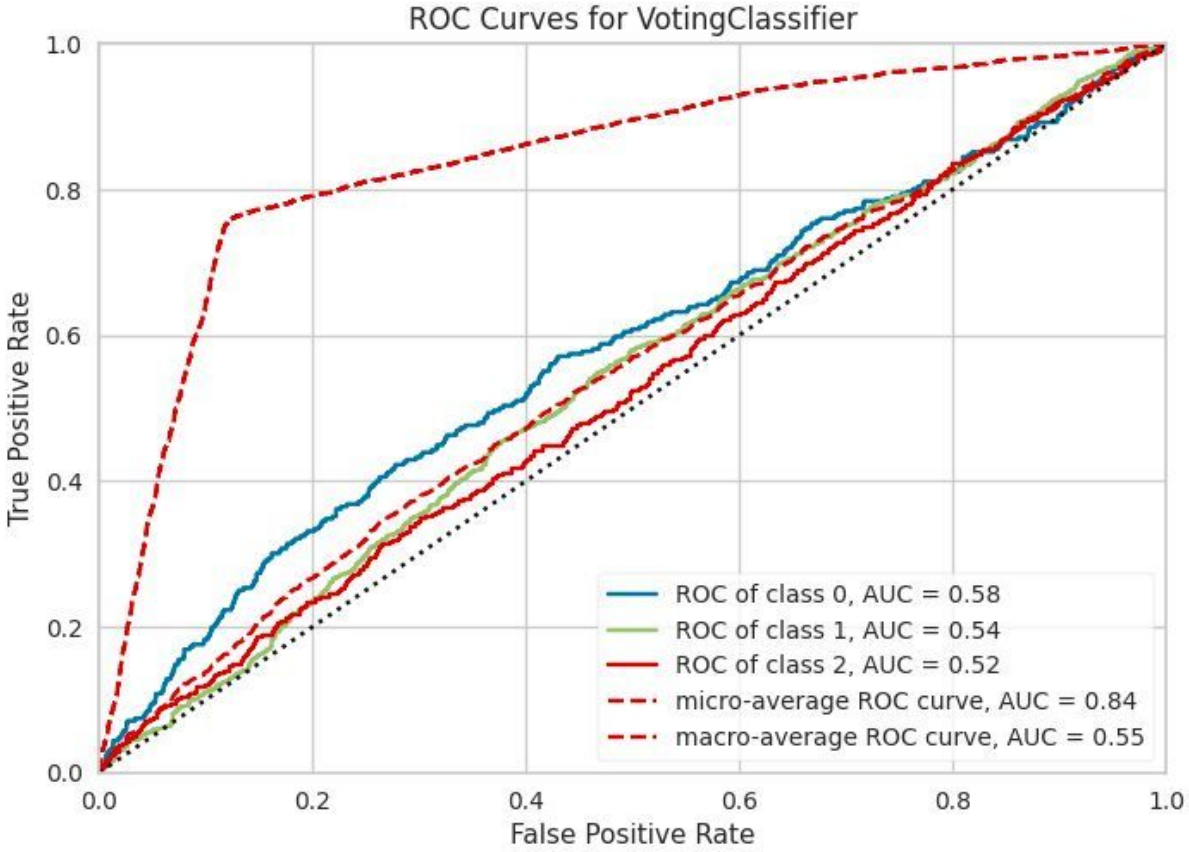
Model	Accuracy	AUC	Recall	F1	Kappa
Decision Trees	0.9999	0.9999	0.9999	0.9999	0.9998
Random Forest	0.9999	0.9999	0.9999	0.9999	0.9997
AdaBoost	0.9999	0.9999	0.9999	0.9999	0.9998
Gradient Boosting	0.9999	0.9999	0.9999	0.9999	0.9998
Logistic Regression	0.8816	0.9431	0.8816	0.8797	0.7277
Naive Bayes	0.7296	0.8065	0.7296	0.7134	0.3428
KNN	0.7162	0.6842	0.7162	0.7024	0.3186


DecisionTreeClassifier Confusion Matrix
True Class
0
1
1948
0
0
3816
0
1
Predicted Class

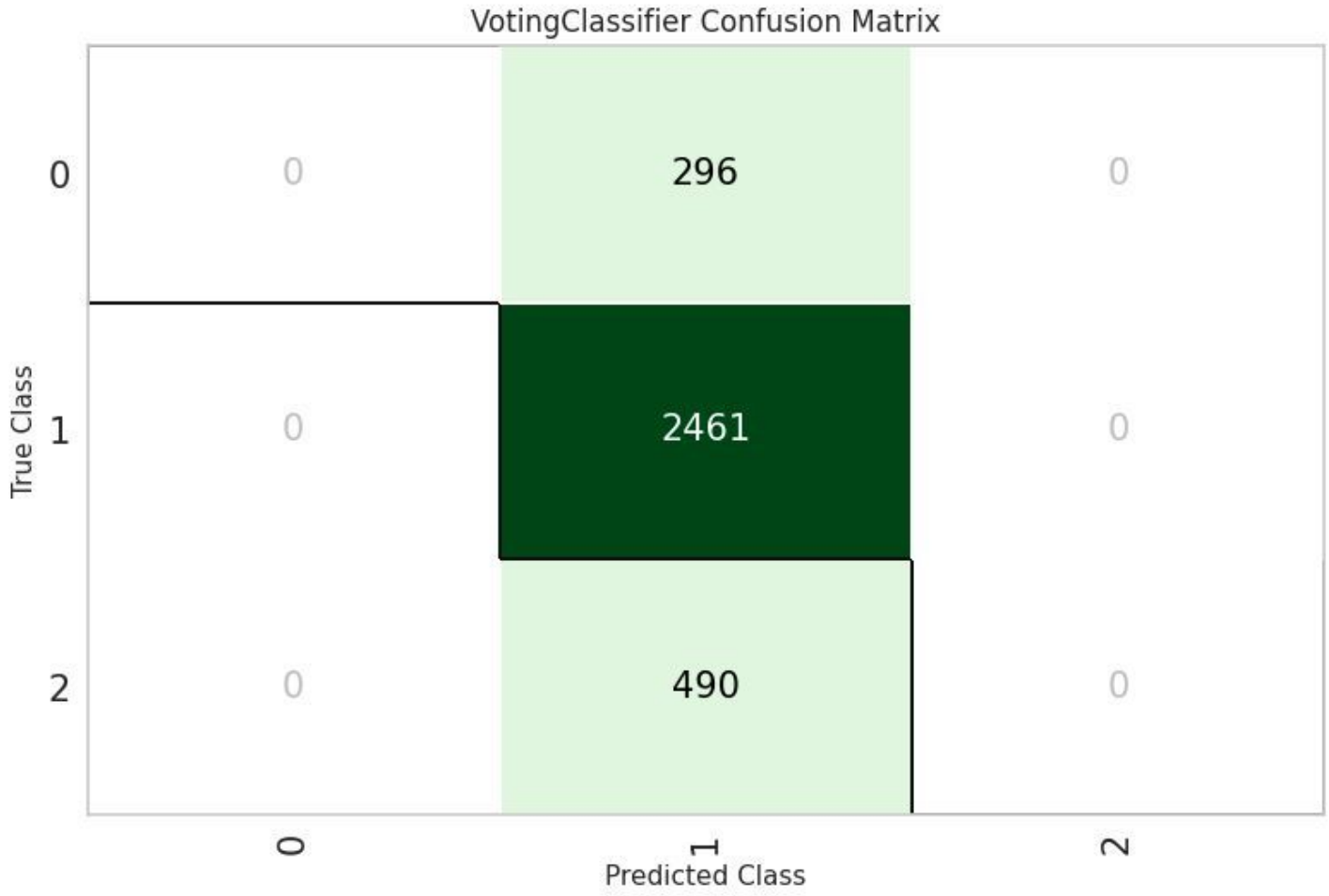
Model Development: Covid19 Classification

Model	Accuracy	AUC	Recall	Precision	F1
Blended Model (LR, RF, Ada)	0.7581	0.5416	0.7581	0.5746	0.6537
Random Forest	0.7567	0.5447	0.7567	0.6162	0.6544
AdaBoost	0.7548	0.5412	0.7548	0.6034	0.6546
Gradient Boosting	0.7536	0.5563	0.7536	0.5972	0.6541
Logistic Regression	0.7578	0.5426	0.7578	0.5747	0.6536
Stacked Model (Meta=LR, RF, Ada)	0.7578	0.5425	0.7578	0.5747	0.6536
KNN	0.7196	0.4962	0.7196	0.6027	0.6459

Model Behavior




VotingClassifier Classification Report
2
1
0
0.000
0.000
0.000
490
0.758
1.000
0.862
2461
0.000
0.000
0.000
296
precision
recall
f1
support
1.0
0.8
0.6
0.4
0.2
0.0


VotingClassifier Confusion Matrix
0
296
0
0
2461
0
0
490
0
True Class
0
1
2
0
1
2
Predicted Class

Conclusion

After considering features from Audio(cough recordings) and metadata(JSON files), We can conclude that, out of all the Models shown before **Blended Model (LR, RF, Ada)** performs best for us with an overall accuracy of **0.7581**

Source Code

Future Scope

We can use this methodology in order to predict and classify any chronic disease like:

1. Asthma
2. Cancer
3. Tuberculosis
4. Pneumonia

Thank You