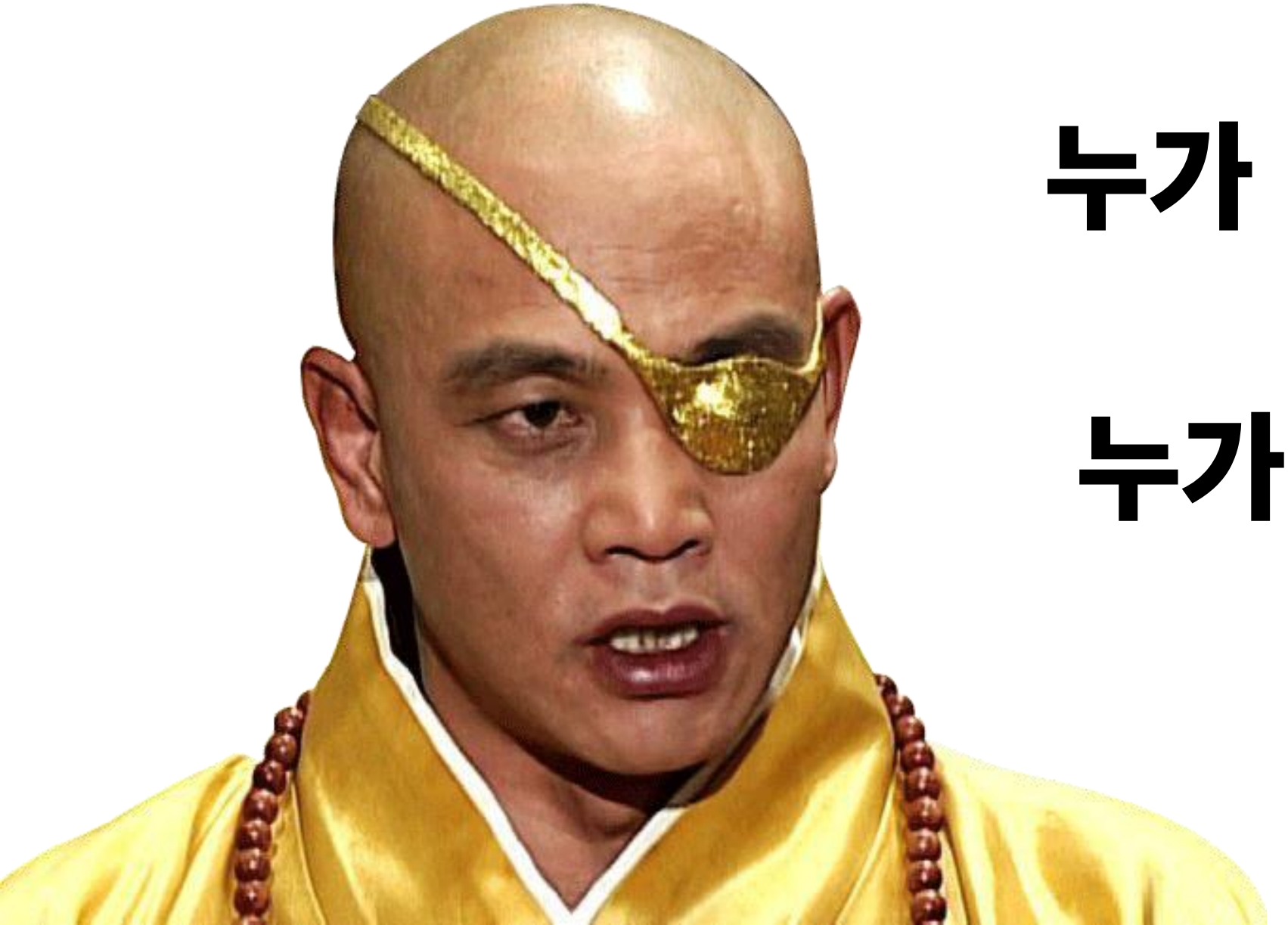


누가 기침소리를 내었는가-? 누가 기침소리를 내었어-?

박규호 박준혁 서효정 선은지 장해식

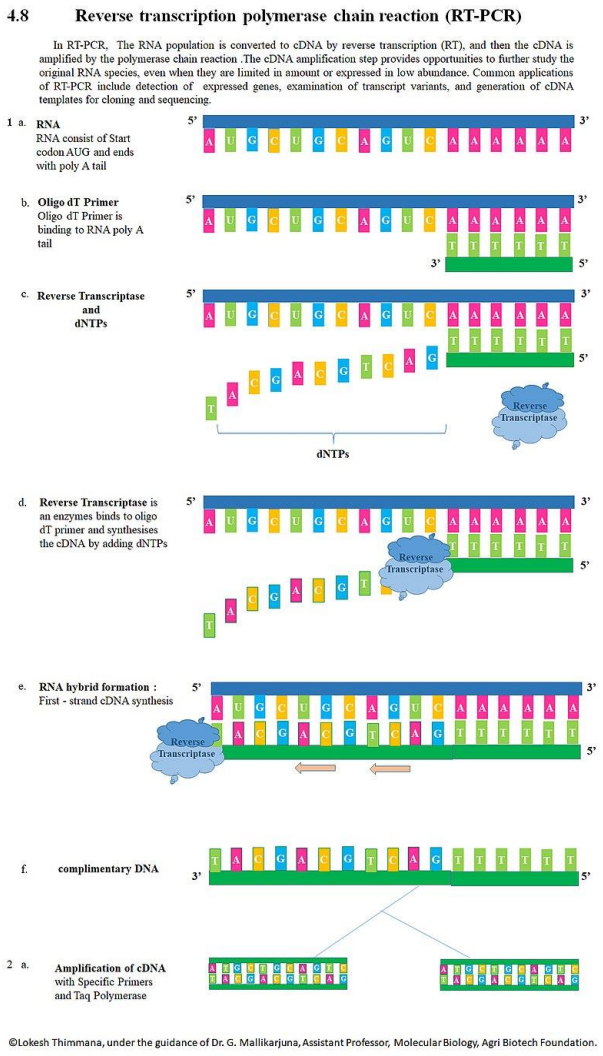
‘양성’입니다

COVID-19 COUGH DETECTOR

코로나19 기침소리 자가진단

‘음성’입니다

COVID-19 검출 표준 방법



역전사 중합효소 연쇄 반응(RT-PCR) 검사

- 증폭 과정을 통해 많은 수의 DNA 서열을 만들기 위해 분자생물학에서 일반적으로 사용하는 실험기법

- RNA가 먼저 역전사 효소에 의해 역전사되어 cDNA를 만들고, 만들어진 cDNA가 기존의 중합효소연쇄반응이나 실시간 중합효소연쇄반응을 통해 증폭

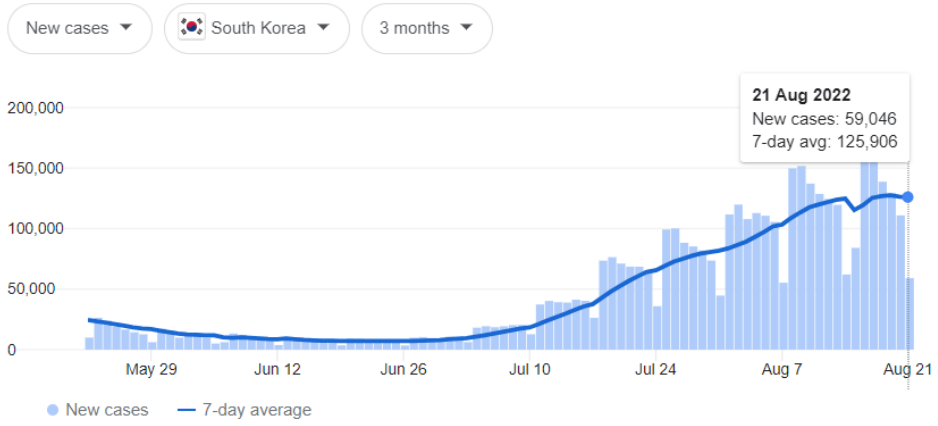
- **민감도와 특이도가 가장 높아** 전 세계적으로 코로나19 감염의 표준 검사법으로 사용되고 있다

COVID-19 검출 표준 방법

그러나,

전용 장비와 시약, 숙련된 전문인력이 필요하며
비용이 많이 들고 시간이 많이 걸린다

기존 검사 방식 개선의 필요성



COVID-19 높은 재유행 가능성

전문가들의 의견 및 코로나 양상 증감 그래프에 따르면 9-12월 사이 COVID-19의 재유행 우려

재유행으로 인해 감염자가 다시 급증하여
검사량이 몰리게 된다면
자칫 **의료체계가 무너질 가능성**도 있다

" 대규모로 배포할 수 있으며,
기존의 한계점을 해결할 수 있는
대체 진단 도구가 필요 "

프로젝트의 방향과 목적

COVID-19의 두드러진 증상

기침

호흡 곤란

+

AI 기술의 접목

기침 소리로부터
COVID-19에 대한
유용한 통찰력

=

프로젝트의 목적

새로운 진단 도구의 설계 가능

기침 소리로부터 COVID-19를
검출하는 AI 모델

데이터 출처 및 구성

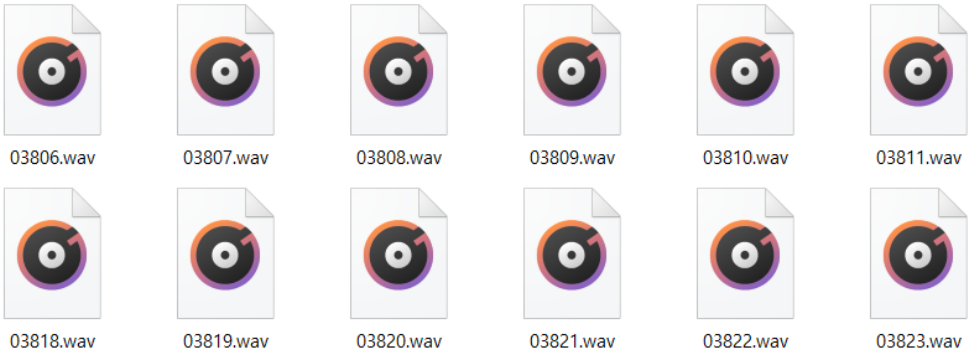
DACON

'월간 데이콘 음향 데이터 COVID-19 검출 AI 경진대회'

- train
  - train_data.csv
  - wav (음성파일)
- test
  - test_data.csv
  - wav (음성파일)
- Unlabeled (* 미검사자)

https://dacon.io/competitions/official/235910/overview/description

wav 파일



csv 파일

train (3805, 38), test (5732, 37), unlabeled (1867, 5)

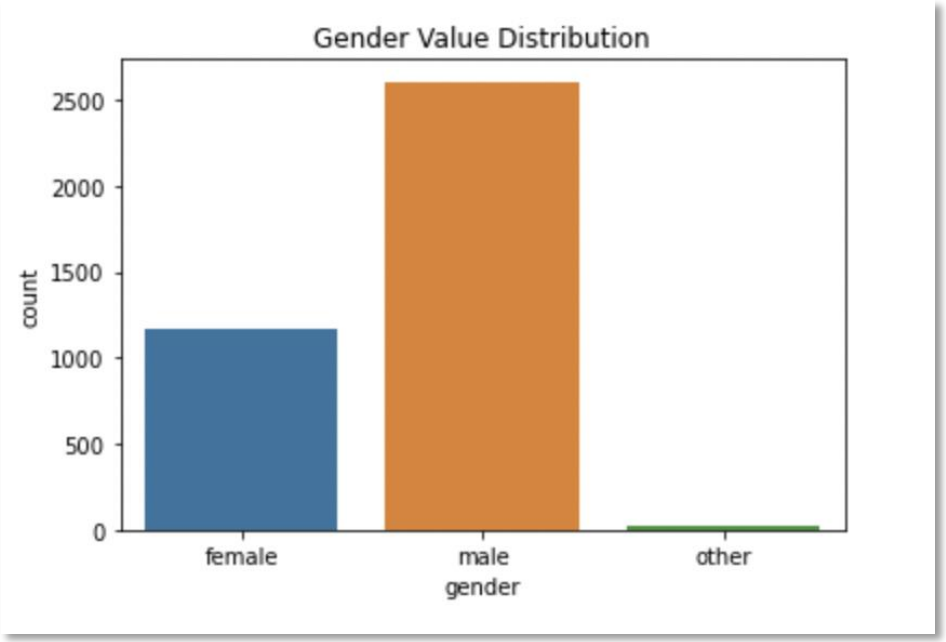
	id	age	gender	respiratory_condition	fever_or_muscle_pain	covid19
0	1	24	female	0	1	0
1	2	51	male	0	0	0
2	3	22	male	0	0	0
3	4	29	female	1	0	0
4	5	23	male	0	0	0

	id	age	gender	respiratory_condition	fever_or_muscle_pain
0	3806	48	female	1	0
1	3807	24	female	0	0
2	3808	29	male	0	0
3	3809	39	female	0	0
4	3810	34	male	0	0

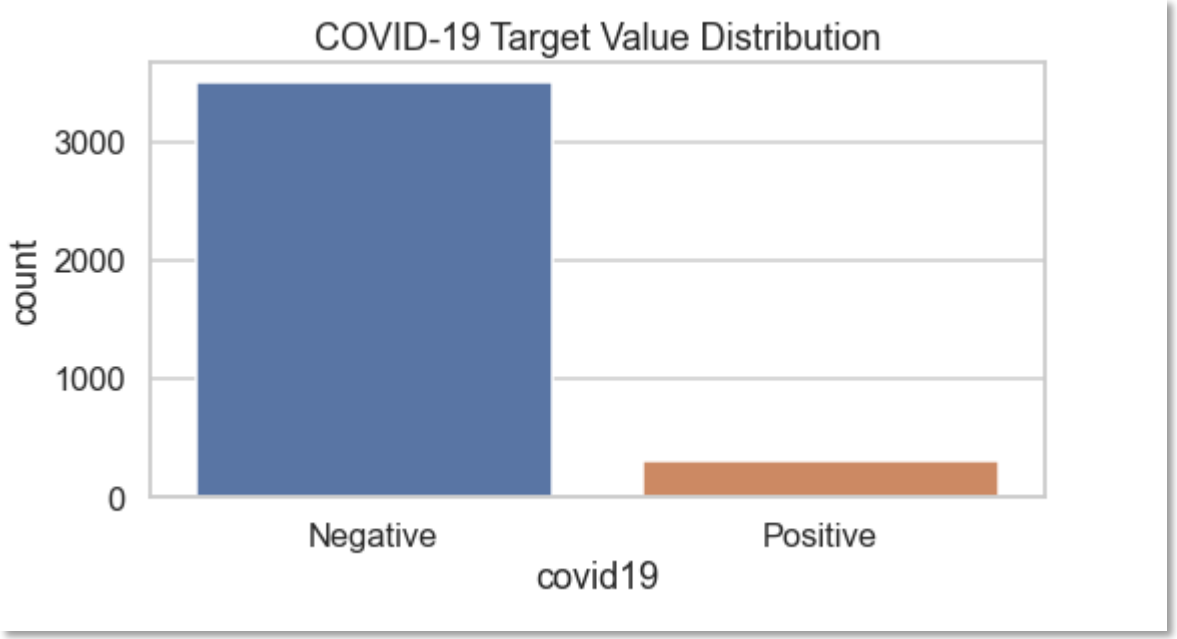
	id	age	gender	respiratory_condition	fever_or_muscle_pain
0	9538	35	male	1	0
1	9539	40	female	0	1
2	9540	33	male	0	0
3	9541	35	male	0	0
4	9542	54	female	0	0

EDA - Train Data Set

Gender Value Distribution



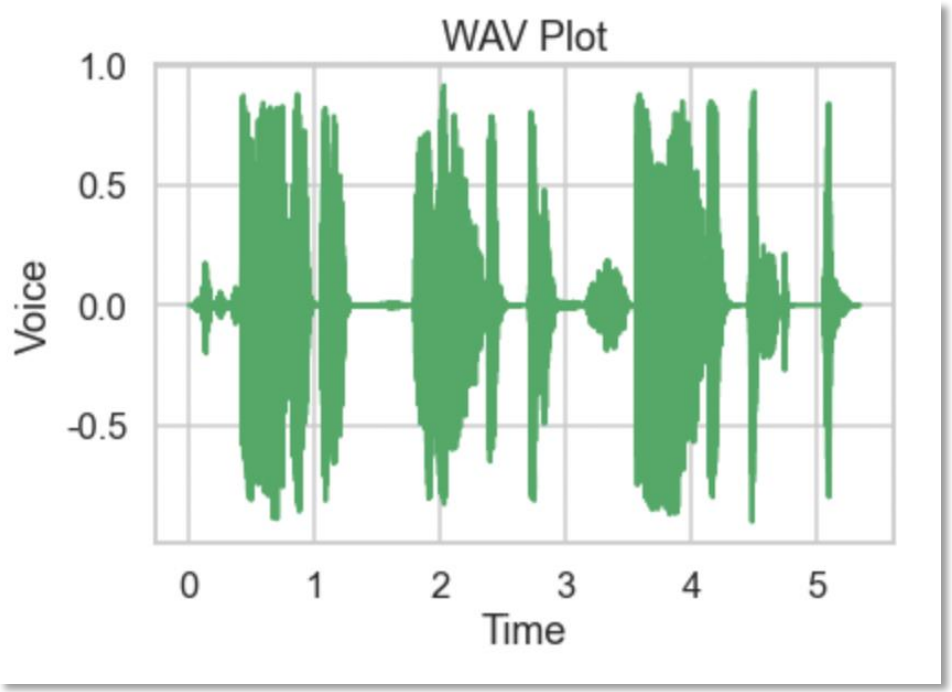
Covid-19 Value Distribution



* 양성 데이터의 비율이 매우 적음 (3499, 306)

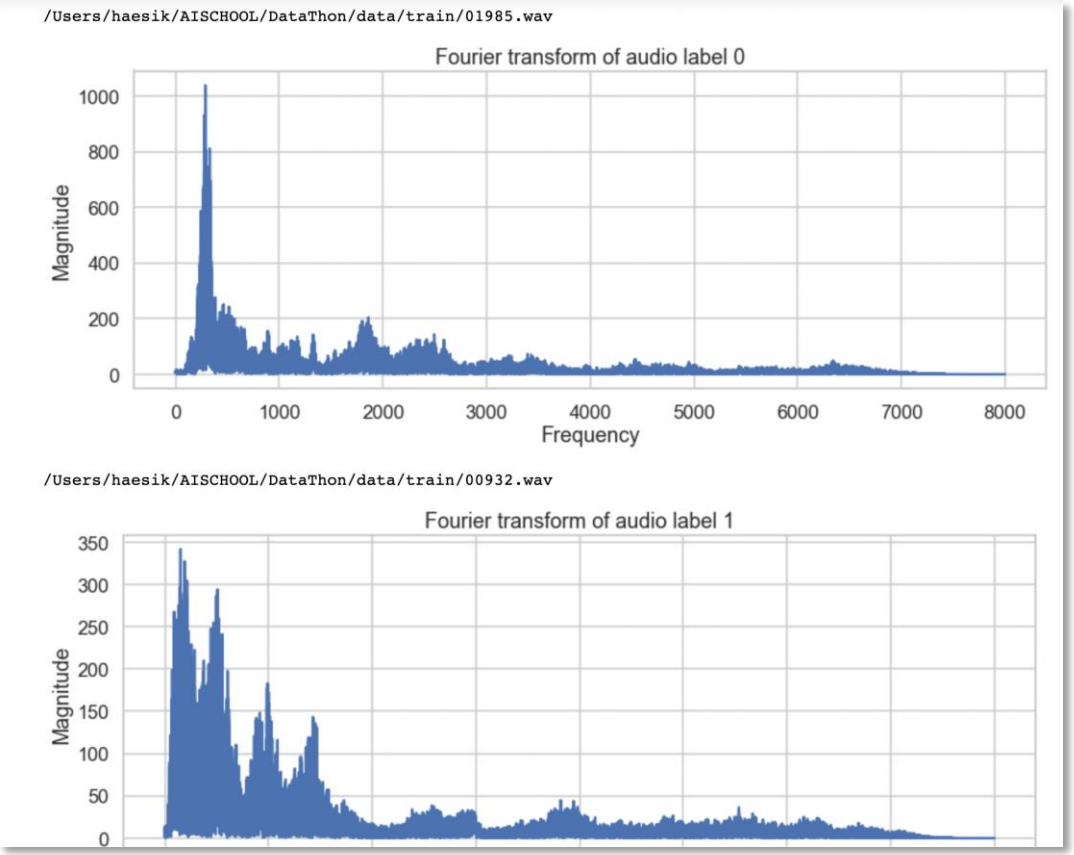
EDA - Sound Wave

Single Sample WAV Visualization



오디오는 **시간(Time)**에 따른 **음압(Voice)**의 표현
= **시간영역(Time Domain)**의 표현

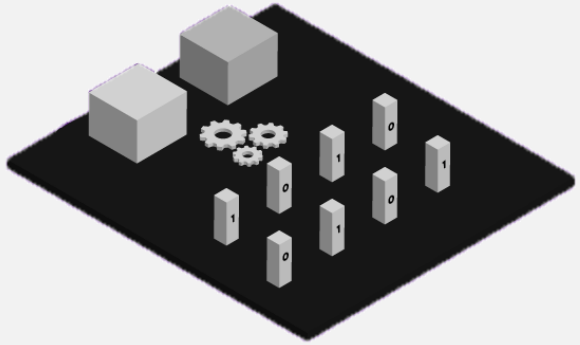
Single Sample WAV Visualization



Data Preprocessing

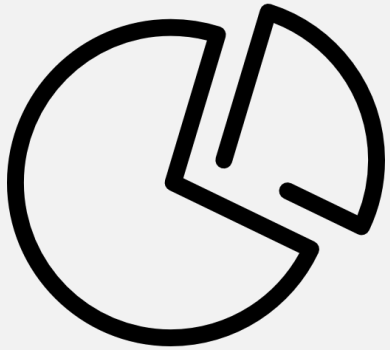
1

One-Hot-Encoding



2

Data split



3

MFCC
Feature Extraction

Data Preprocessing

1

2

ata split

MFCC란?

오디오 신호에서 추출할 수 있는 feature

소리의 고유한 특징을 나타내는 수치

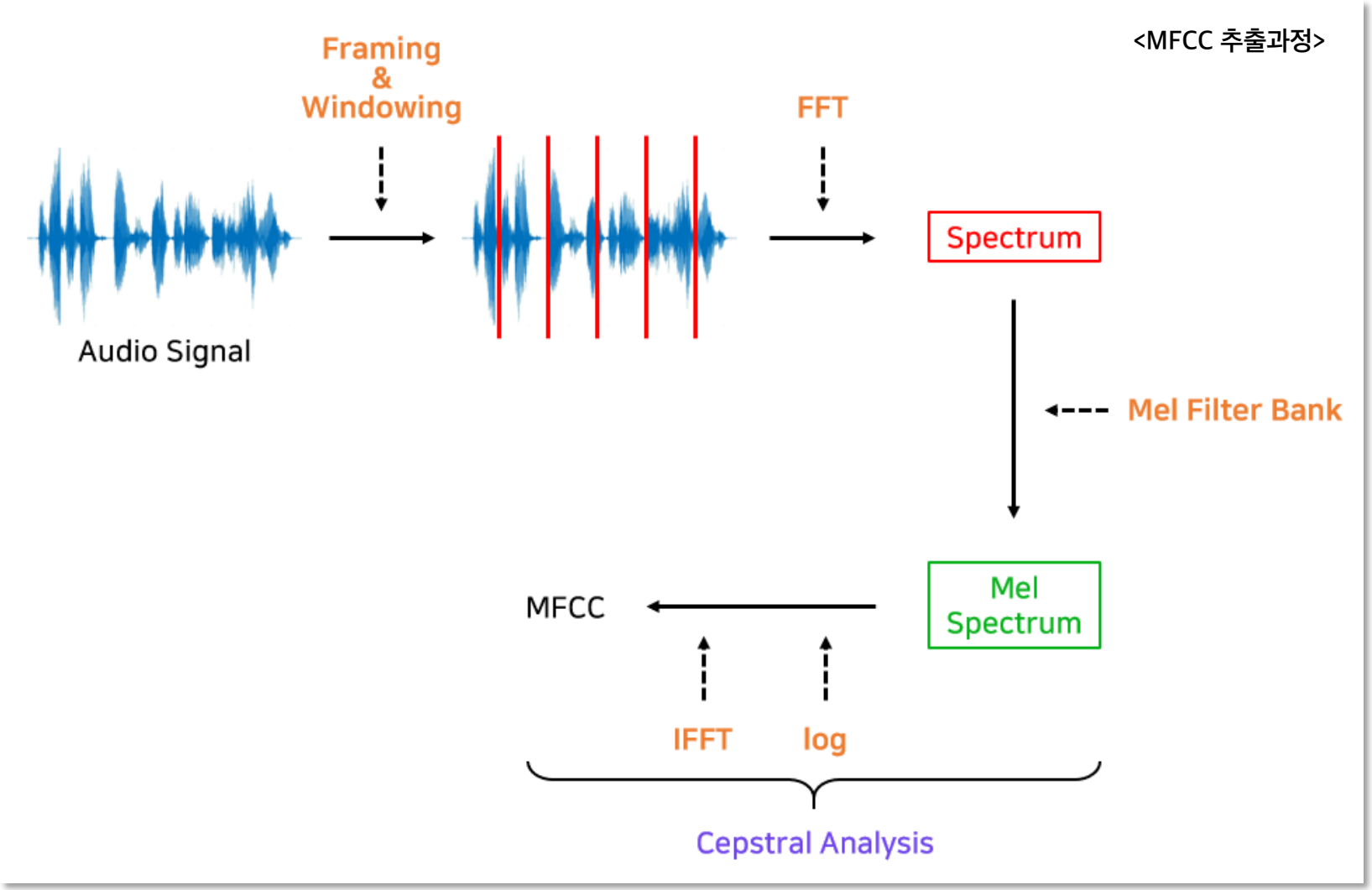
MFCC 추출 과정

1. 오디오 신호를 프레임별로 나누어 FFT를 적용
   > Spectrum을 구한다
2. Spectrum에 Mel Filter Bank를 적용
   > Mel Spectrum을 구한다
3. Mel Spectrum에 Cepstral 분석을 적용
   > MFCC를 구한다

MFCC
Feature Extraction

?

Data Preprocessing



ML Modeling Framework

Preprocessed Data

ML Model Search

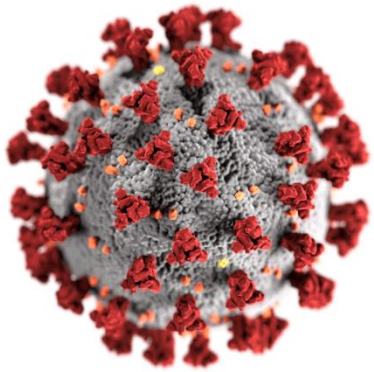
Semi Supervised Learning

MLPClassifier Modeling

Performance Evaluation

Post Tuning

Covid-19
Cough Detector



모델링 - ML Model Search

① ML Search

총 11종의 Machine Learning Model Search

	name	acc	f1	precision	recall
0	GaussianNB()	0.885677	0.314961	0.303030	0.327869
1	(DecisionTreeClassifier(max_depth=1, random_st...	0.918528	0.162162	0.461538	0.098361
2	DecisionTreeClassifier(random_state=42)	0.859396	0.144000	0.140625	0.147541
3	([DecisionTreeRegressor(criterion='friedman_ms...	0.918528	0.114286	0.444444	0.065574
4	MLPClassifier(random_state=42)	0.905388	0.100000	0.210526	0.065574
5	LGBMClassifier(random_state=42)	0.919842	0.061538	0.500000	0.032787
6	LogisticRegression(random_state=42)	0.919842	0.031746	0.500000	0.016393
7	(DecisionTreeClassifier(max_features='auto', r...	0.919842	0.031746	0.500000	0.016393
8	SVC(random_state=42)	0.919842	0.000000	0.000000	0.000000
9	KNeighborsClassifier()	0.918528	0.000000	0.000000	0.000000
10	LinearSVC(random_state=42)	0.919842	0.000000	0.000000	0.000000

② HyperParameter Tuning

GaussianNB 모델에 HyperParameter Tuning 수행

	params	mean_test_score	rank_test_score
5	{'var_smoothing': 1e-07}	0.270246	1
7	{'var_smoothing': 1e-09}	0.270051	2
8	{'var_smoothing': 1e-10}	0.270051	2

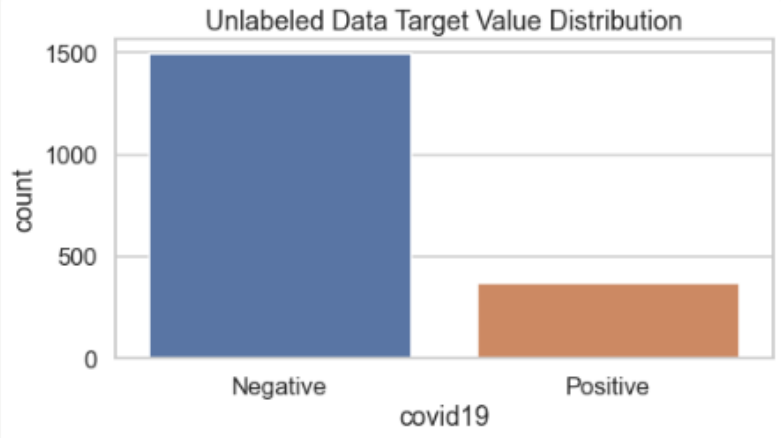
```
# Best HyperParameter & Score
best HyperParameter : {'var_smoothing': 1e-07}
best Grid_CV Score : 0.2702

# GaussianNB Model Attributes
best_est.sigma_.shape : (2, 38)
best_est.var_.shape : (2, 38)
best_est.theta_.shape : (2, 38)
```

모델링 - Semi Supervised Learning

③ Unlabeled Data Labeling

사전 모델을 통해 Unlabeled Data에 Label 값을 부여

Unlabeled Data Labeling

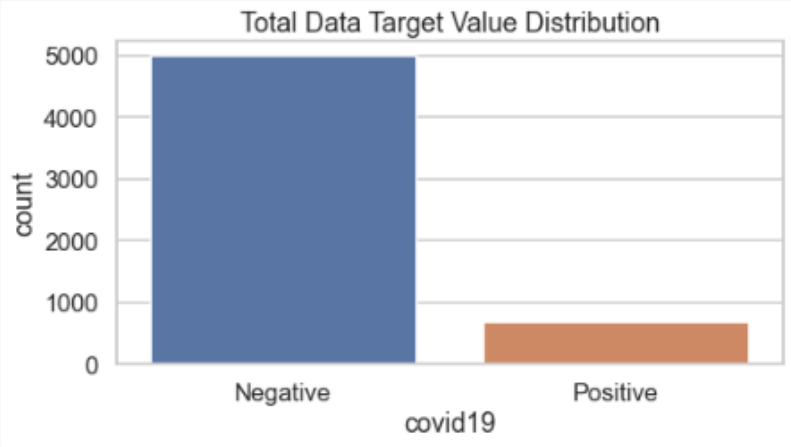



```
# Unlabeled Data Target Value Distribution
0    1494
1     373
Name: covid19, dtype: int64
```

④ Augmented Train Data

기존 학습 데이터셋과 병합 > 최종 데이터셋 구축

Total Data Labeling Distribution




```
# Total Data Target Value Distribution
0    4993
1     679
Name: covid19, dtype: int64
```

모델링 - MLPClassifier Modeling

❺ Multi-layer Perception Classifier Modeling

sklearn에서 제공하는 MLPClassifier 클래스를 사용한 Voice Detecting Classifier Modeling

Multi-layer Perceptron Classifier Modeling

[Model Structure]

```
# MLPClassifier
model = MLPClassifier(activation = 'relu', solver = 'adam',
                      random_state=42)
```

```
# Get MLPClassifier Format
n_features : 38
n_iter : 156
n_layers : 3
n_outputs : 1
out_activation : logistic
```

```
Neural Network Parameter's format (38, 100) (100, 1)
```

[Learning Curve]

[Confusion Matrix]

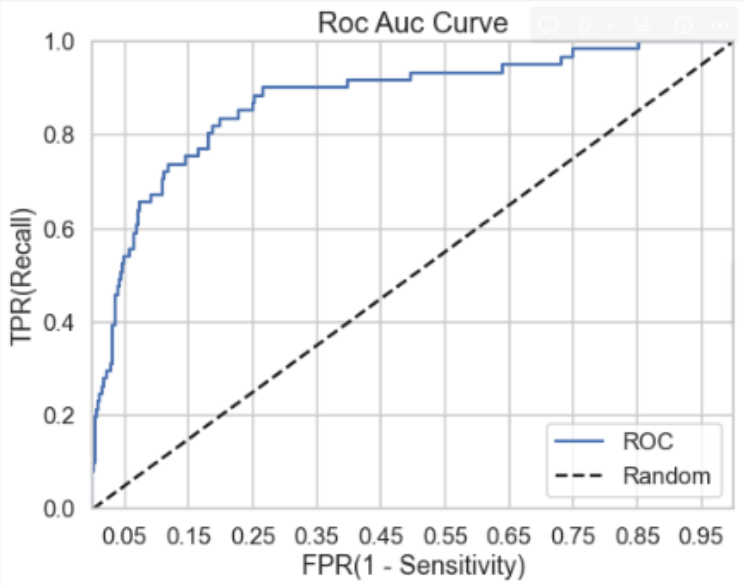

```
# Get Classification Metrics
[[683  17]
 [ 43  18]]
정확도 :0.9212, 정밀도 :0.5143, 재현율 :0.2951, F1 :0.3750
```

모델링 - Performance Evaluation

6 ROC Auc Curve

MLPClasssifer 학습 결과,
0.8761으로 준수한 수치의 ROC AUC SCORE이 나타남

[ROC Curve]

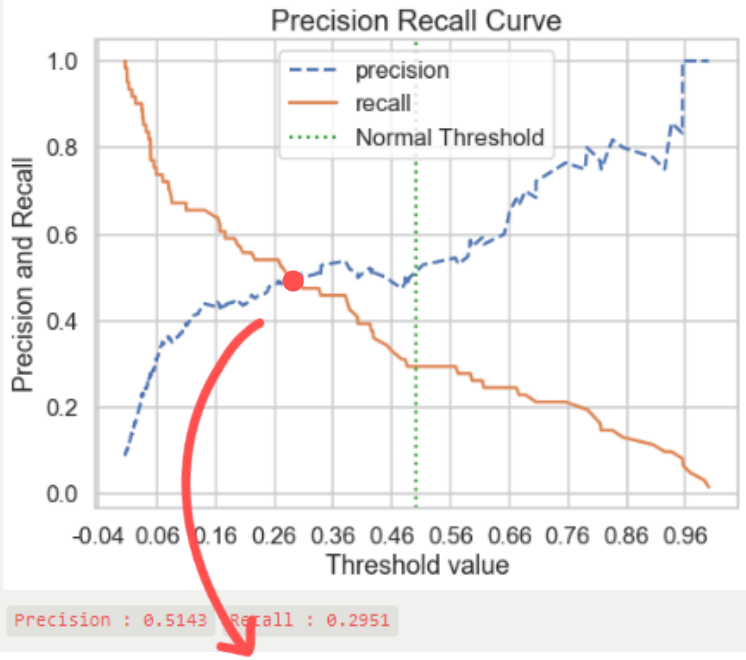


Roc Auc Score : 0.8761

7 Precision Recall Curve

MLPClassifier에 사후 튜닝을 수행하기 전,
Precision Recall Curve를 통해 대략적인 튜닝 방향을 확인 가능

[Precision Recall Curve]



Precision : 0.5143 Recall : 0.2951

정확도가 재현율이 만나는 접점이 prob 0.5 좌측에 위치
사후튜닝 과정에서 prob이 낮아질 것으로 예측

모델링 - Post Tuning

8 Post Tuning Result

[Post Tuning Result]

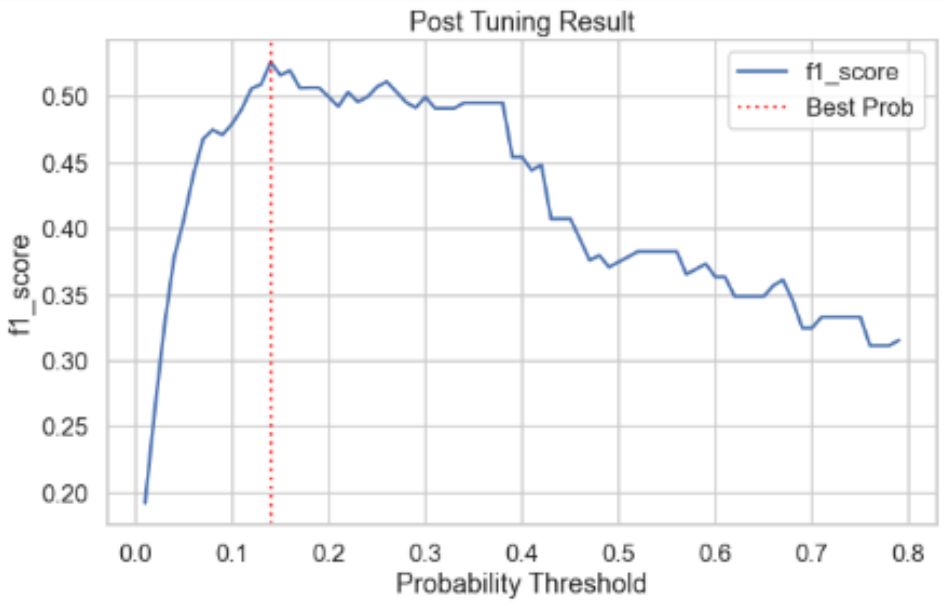
```
# Post Tuning
post_result = mlp_post_tuning(y_pred_proba_1,0.01,0.8)
```

```
[[216 484]
 [  3  58]]
정확도 :0.3601, 정밀도 :0.1070, 재현율 :0.9508, F1 :0.1924
[[390 310]
 [  5  56]]
정확도 :0.5861, 정밀도 :0.1530, 재현율 :0.9180, F1 :0.2623
[[479 221]
 [  6  55]]
정확도 :0.7017, 정밀도 :0.1993, 재현율 :0.9016, F1 :0.3264
[[538 162]
 [  9  52]]
정확도 :0.7753, 정밀도 :0.2430, 재현율 :0.8525, F1 :0.3782
[[577 123]
 [ 14  47]]
정확도 :0.8200, 정밀도 :0.2765, 재현율 :0.7705, F1 :0.4069
[[602  98]
 [ 16  45]]
정확도 :0.8502, 정밀도 :0.3147, 재현율 :0.7377, F1 :0.4412
[[617  83]
 [ 17  44]]
정확도 :0.8686, 정밀도 :0.3465, 재현율 :0.7213, F1 :0.4681
```

	proba	acc	pre	rec	f1
13	0.14	0.905388	0.43956	0.655738	0.526316
15	0.16	0.905388	0.438202	0.639344	0.52
14	0.15	0.904074	0.433333	0.639344	0.516556
25	0.26	0.917214	0.485294	0.540984	0.511628
12	0.13	0.898817	0.416667	0.655738	0.509554

9 Post Tuning Result Visualization

[Post Tuning Result Visualization]



모델링 - Final Result

⑩ Final Result & Sample Simulation

Public Sample Submission

[Dacon Public Score]

WINNER 1% 4% 10%

#	팀	팀 멤버	점수	제출수
56	하이든	하이	0.58917	2
1	dsjoh		0.63021	22
2	inha_lx		0.63015	9
3	alstjrdltzz	al ri	0.62921	79
4	뮤쉬잇		0.62891	8

TOP SCORE: 0.63021

SCORE: 0.58917

Single Sample Simulation

[Load the Trained Model]

```
# MLP Model Load
model = pickle.load(open(filename, 'rb'))
```

[Single Image Sample : Positive]

0:00 / 0:00

	age	respiratory_condition	fever_or_muscle_pain	female	male	other
184	31.0	1.0	1.0	0.0	1.0	0.0

data/train/00185.wav

sample_return : Positive

[Single Image Sample : Negative]

0:00 / 0:00

	age	respiratory_condition	fever_or_muscle_pain	female	male	other
106	31.0	1.0	0.0	1.0	0.0	0.0

data/train/000107.wav

sample_return : Negative

COVID-19 COUGH DETECTOR

직접 사용해보러 갑시다 !