Multimodal Data Analysis of Covid-19 Patients

By

Jishnu Biswas

Introduction:

COVID-19, a communicable respiratory infection caused by severe acute respiratory syndrome coronavirus 2 (SARS-COV2) virus which originated from Wuhan, China, has a cumulative incidence of over 80 million cases worldwide causing over 1.70 million deaths as of late December 2020. In March 2020, the World Health Organization (WHO) declared COVID19 as a pandemic. The high incidence of cases can be attributed to the high infectivity of the virus, low mortality among the hosts and the existence of asymptomatic carriers causes the continuous and rapid spread of the virus. Several countries were compelled to go for lockdown to restrict the rapid spread of the virus which leads to suspension of services across industries and economic slowdown.

This unprecedented situation propels the scientific community to probe the covid-19 patients' entire journey in their illness, to be precise, from being tested positive to final fate through having been hospitalized or not. This work will come as an aid to the health care sectors to take quick decisions on a patient and decide accordingly since the excessive pressure on the hospitals needs to be efficiently controlled.

In this project, the three major Objective of the Project:

- Prediction of final outcome of the patient(Death/Discharged).
- Duration of Hospitalization.
- ICU Requirement.

Preliminary Analysis for Understanding Data

About Provided Datasets

- **Patient Dataset** - 7063 samples
- **About Features**

The Original Dataset contains lots of Raw data as it was directly collected at the time of admission/Covid-test of the patient. Columns(like Patient personal Information(mobile no, country, address,etc) and some other details taken at airport) were straightway

eliminated. The attributes that has been considered for further analysis are listed below.

Columns:

1. Age- Integer type variable, determining the Patient's age.
2. Gender - Categorical Variable (Male/Female)
3. Condition_patient - Categorical Variable (Symptomatic/Asymptomatic)
4. Comorbidities Factor-
    a. COPD(Chronic Obstructive Pulmonary Disease): Categorical (Abs/Pre)
    b. Hypertension: Categorical (Abs/Pre)
    c. Asthma: Categorical (Abs/Pre)
    d. Heart Disease: Categorical (Abs/Pre)
    e. Diabetes:Categorical (Abs/Pre)
5. Symptoms at the time of Admission:
    a. Fever
    b. Sore Throat
    c. General Weakness
    d. Breathlessness
    e. Headache
    f. Cough

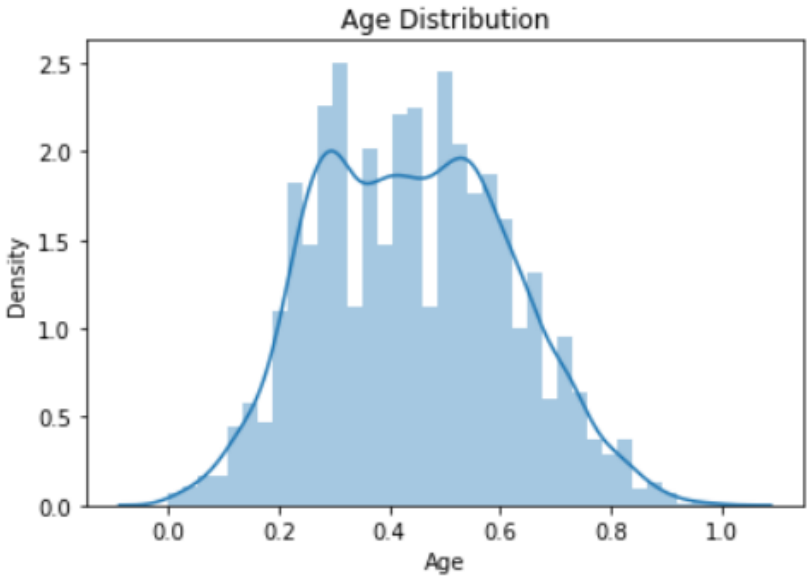
Finally these 14 features were retained out of which all are categorical ones except age.

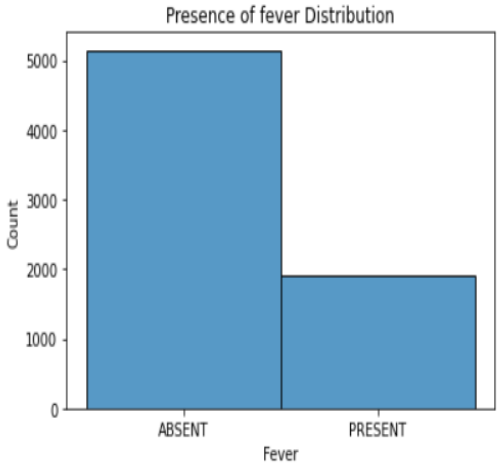
Target Classes:

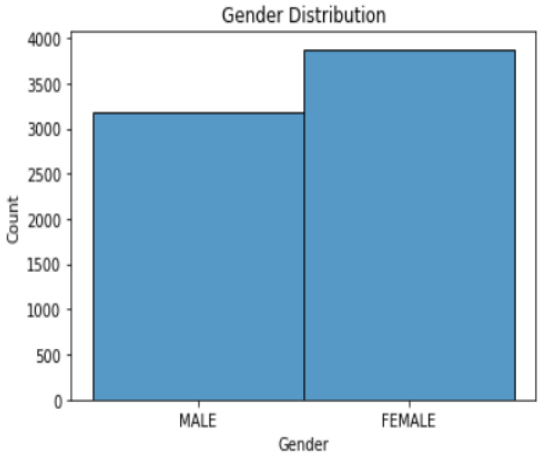
Prediction of Final Outcome of Patient(Death/Discharged)

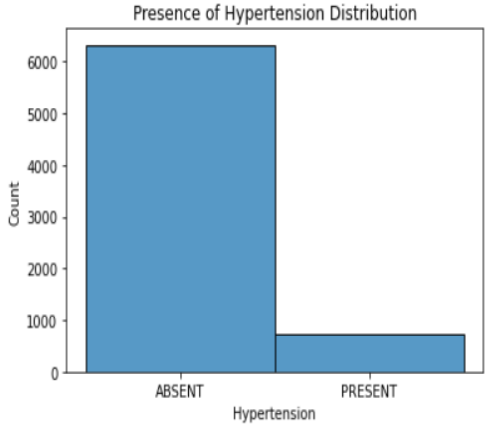
Exploratory Data Analysis:

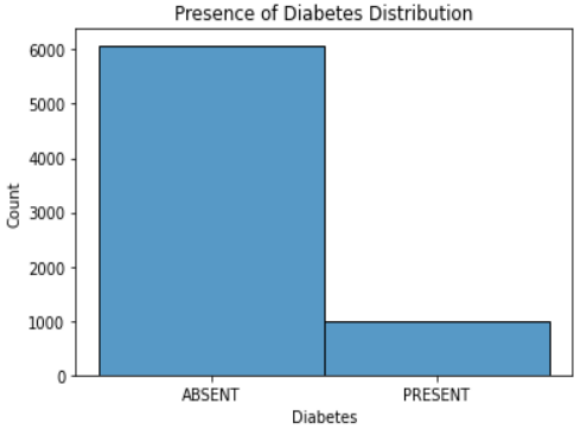
Distributions of Age and some features are given below:

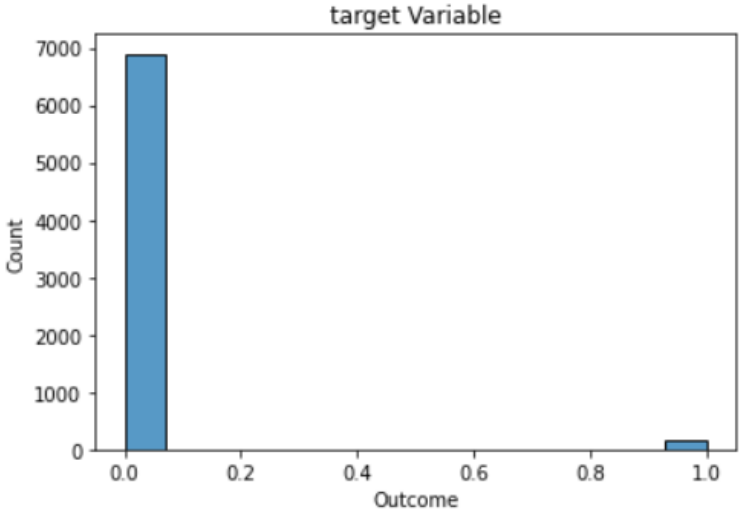




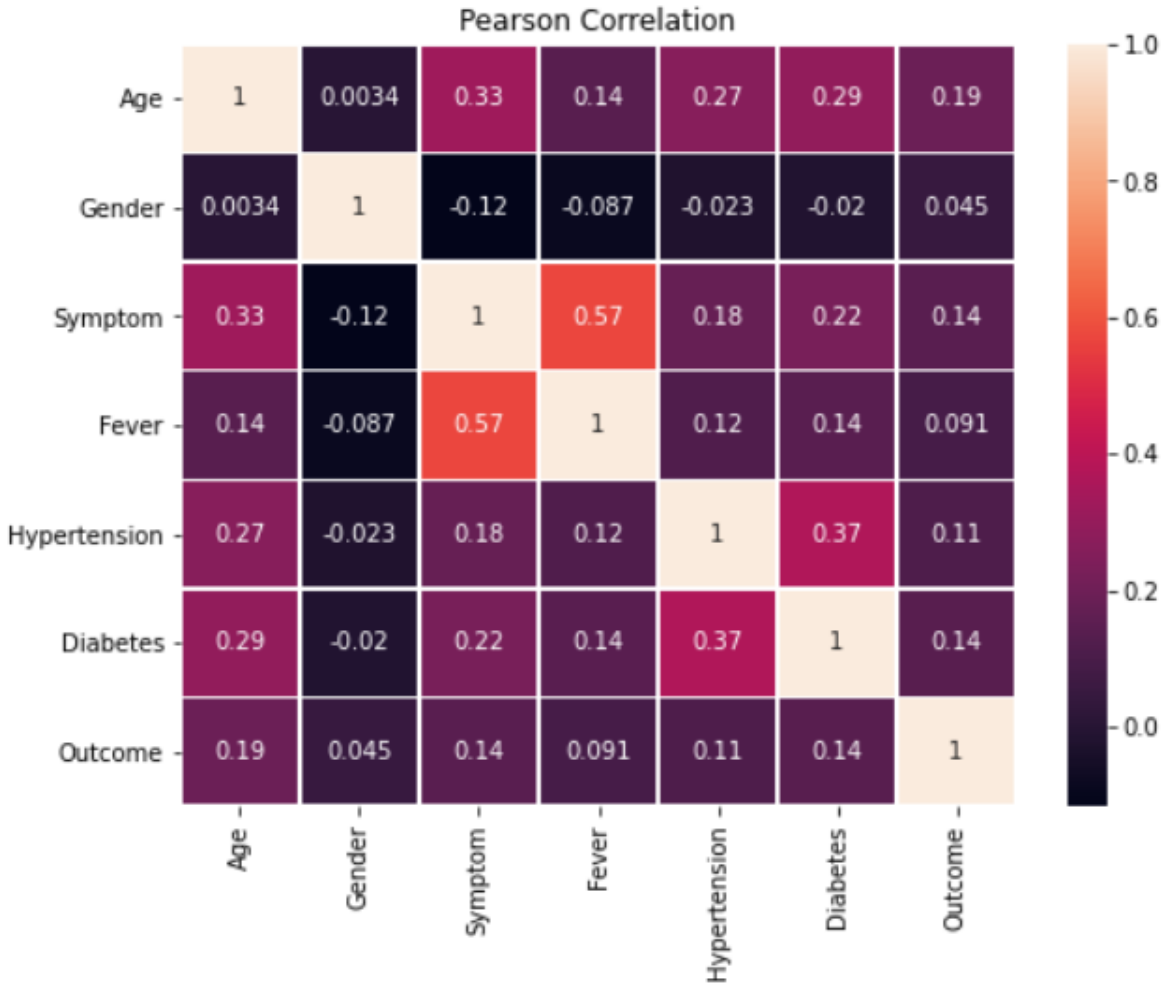








From the Target Variable(Outcome) Hist Plot, the imbalancing of the dataset is evident. Pearson correlation matrix for some comorbidities and symptoms with outcome is shown below.



Model Development:

As a naive approach, standard Machine Learning based Classification Models have been tried(SVMs, Random Forest, Bayesian classifier and multi layer Perceptron). Finally Extreme Gradient Boosting(XgBoost) algorithm has been adopted and its detailed hyperparameters and results and importance score has been reported below.

Model Name: XgBoost (85% training data, 15% test data)
Hyparameters have been finalized through Randomised Search with cross validation.

[Scale_pos-weight=40, n_estimators=100, reg_lambda=1.0,learning_rate=0.1,
max-depth=9, gamma=92, min_child_weight=1]

Visualization of Model Performance:

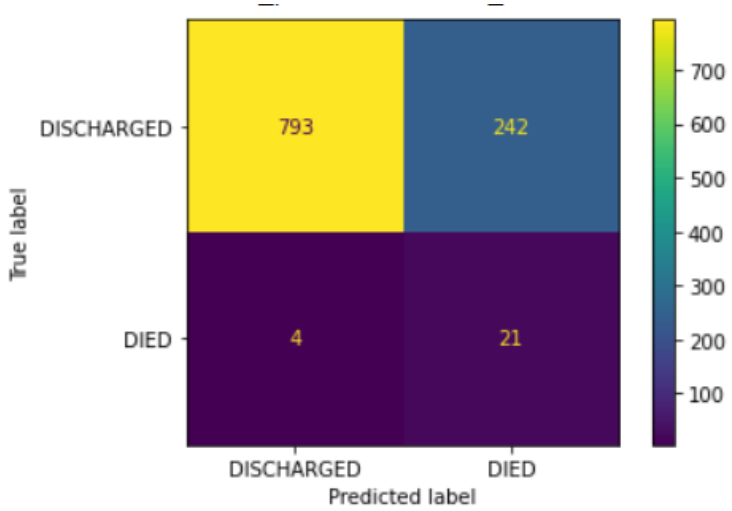
Classification Report on Training Set:

	precision	recall	f1-score	support
0	1.00	0.81	0.89	5860
1	0.10	0.92	0.19	143
accuracy			0.81	6003
macro avg	0.55	0.86	0.54	6003
weighted avg	0.98	0.81	0.88	6003

Classification Report on Test Set:

	precision	recall	f1-score	support
0	0.99	0.77	0.87	1035
1	0.08	0.84	0.15	25
accuracy			0.77	1060
macro avg	0.54	0.80	0.51	1060
weighted avg	0.97	0.77	0.85	1060

Confusion Matrix on Test Data:



Conclusion:

Since almost all features for our Covid-19 outcome prediction task were of the categorical type(except age), the analysis has been planned accordingly. We used the insights gained from visualizations and statistical analysis for primary understanding on the dataset. Modelling is initiated with some naive algorithms and finally adopted Xgboost Model, it successfully satisfied our objective of prediction of correct outcome of the patients along with a uniformly distributed recall, in a robust manner. The Model, though satisfying our objective efficiently, is unable to cross a certain ~0.85 mark despite exhaustive hyperparameter tuning, which is an indicator that the given features of the dataset are not capable enough to dictate outcome alone and demand more impactful attributes, for instance, incorporation of Laboratory test parameters(e.g, CRP, Ferritin, D-Dimer etc) would improve the result and along with this, keeping in mind, the imbalancing dataset wrt death cases(which it will be always), more training data will also help to apply more robust algorithm though Deep NeuralNet architectures.