COVID 19: Patient Pre-condition Analysis

By Abhijit Gokhale and Shubham Sharma

Introduction

A health crisis of massive proportions such as the current COVID-19 pandemic provides us with an opportunity to ponder and reflect over what we can do better in the way we deal with healthcare to make us humans be more prepared and enabled to combat such an event in the future.

During the entire course of the pandemic, one of the main problems that healthcare providers have faced is the shortage of medical resources and a proper plan to efficiently distribute them.

They have been in the dark failing to understand how much resource they could even in the very next week as the COVID-19 curve has swayed very unpredictably. In these tough times, being able to predict what kind of resource an individual might require at the time of being tested positive or even before that will be of great help to the authorities as they would be able to procure and arrange for the resources necessary to save the life of that patient.

Objective

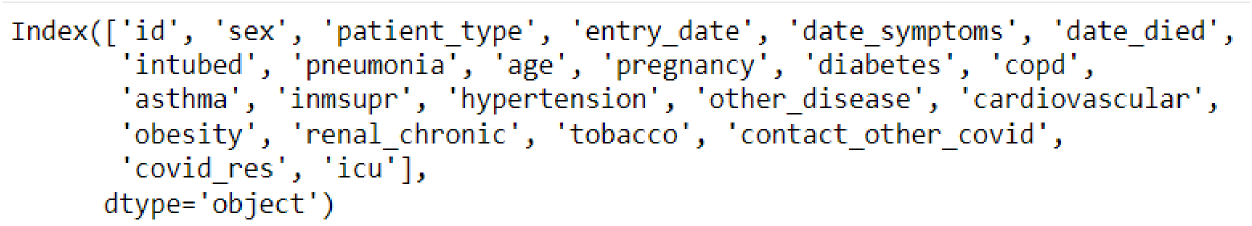
Our main objective was to analyze the factors that influence mortality in patients who do not need an intensive care unit and those who do. What are the underlying comorbidities that patients are likely to test positive for COVID? Which factors influence the need for an intensive care unit? This can help with patient triage, optimal distribution of vaccinations (if needed) in countries with limited resources, or prevention in countries susceptible to the virus.

Data Description

The data obtained is from the Mexican government and hence, the analysis is valid for Mexico or maybe North America. The pandemic stats and behaviours are extremely different for Asian countries when compared to North American or European countries owing to far lower case fatality rate for Asia.

We have data of 566k patients related to their health conditions such as whether they have pneumonia, cardiovascular problems, obesity, copd, asthma etc.

These are our columns -

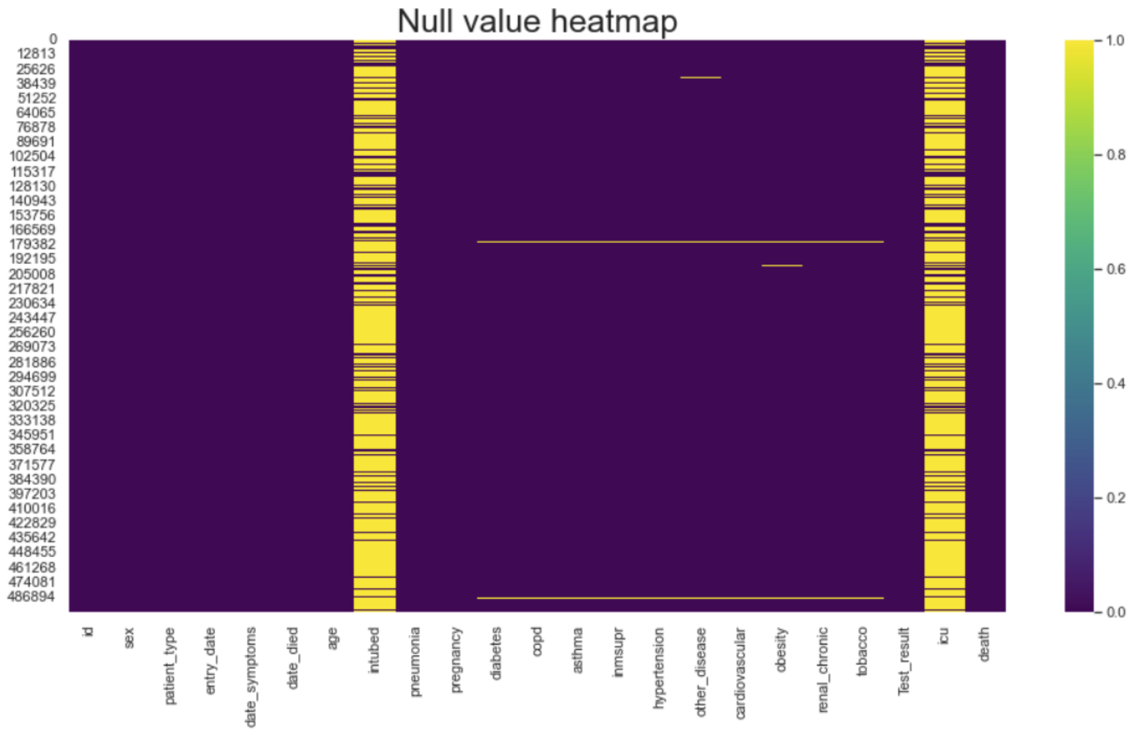


```
Index(['id', 'sex', 'patient_type', 'entry_date', 'date_symptoms', 'date_died',
       'intubed', 'pneumonia', 'age', 'pregnancy', 'diabetes', 'copd',
       'asthma', 'inmsupr', 'hypertension', 'other_disease', 'cardiovascular',
       'obesity', 'renal_chronic', 'tobacco', 'contact_other_covid',
       'covid_res', 'icu'],
      dtype='object')
```

Most of them contain values no or yes to indicate whether the patients are suffering from that particular comorbidity.

There are a lot of target variables that we can study based on the segmentation of the data but first let's check out the correlation heatmap and get rid of missing values. In healthcare data analysis, it's critical to get rid of missing data rather than impute them because of the cruciality of the analysis. We can't generate synthetic healthcare data. Hence, **we decided to get rid of all rows having missing data**.

From the Null values heatmap we can see, there is a lot of missing data on the requirement of ICU by a patient. Hence, we will analyze only those patients for whom we know whether they required ICU or not.



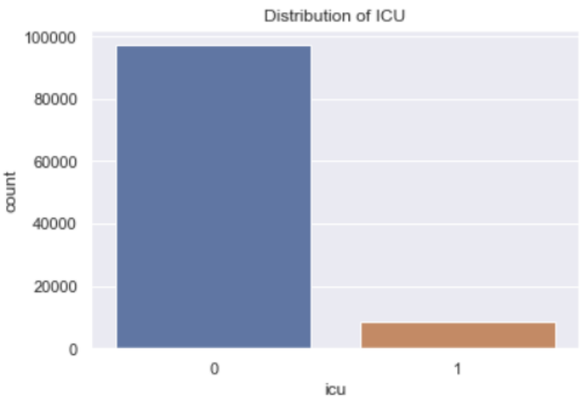
Based on our existing columns, we feature engineered some new columns that were useful in time series analysis and building models.

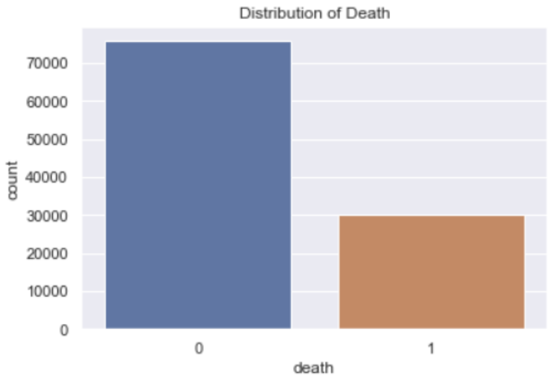
Feature Engineering -

"Death" indicates whether a patient has died or not. We discretized age to better study infants, children, adults, senior citizens and really old people. Using datetime columns such as date_symptoms, date_entry and date_died, we created columns like the difference between date of hospitalization and first onset of symptoms etc. to understand some important trends within the data.

Exploratory Data Analysis

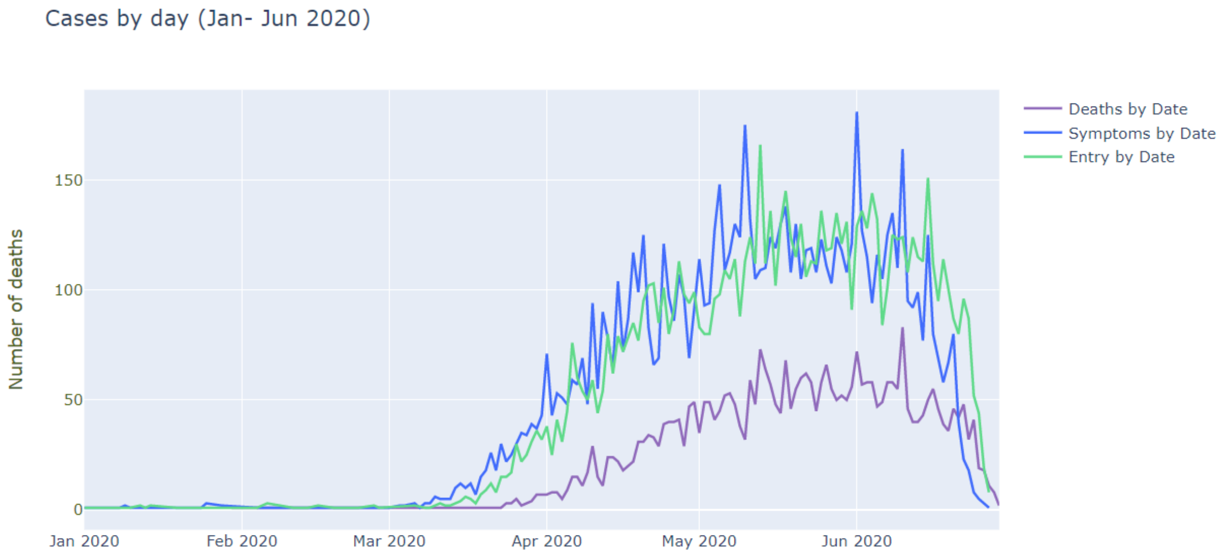
After data cleaning and feature engineering, here is what the distribution of ICU and death variables look like -



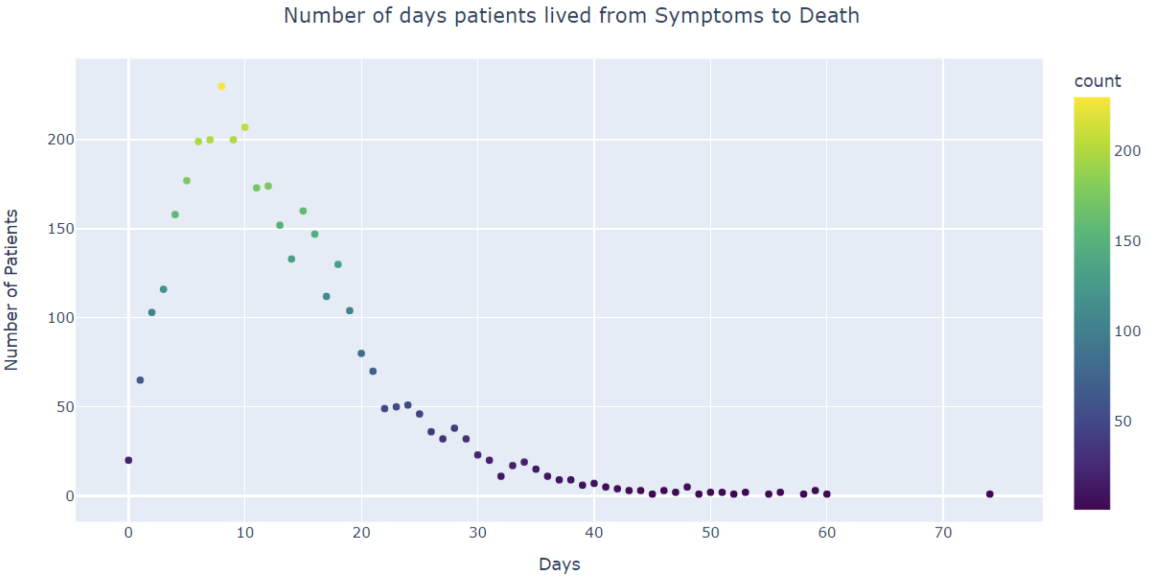


So, not a lot of patients require ICU but the percentage of patients who have died is more. It might be possible that the severity of COVID for these patients were very high and they died as soon as they were admitted before the doctors could figure out the requirement for ICU.

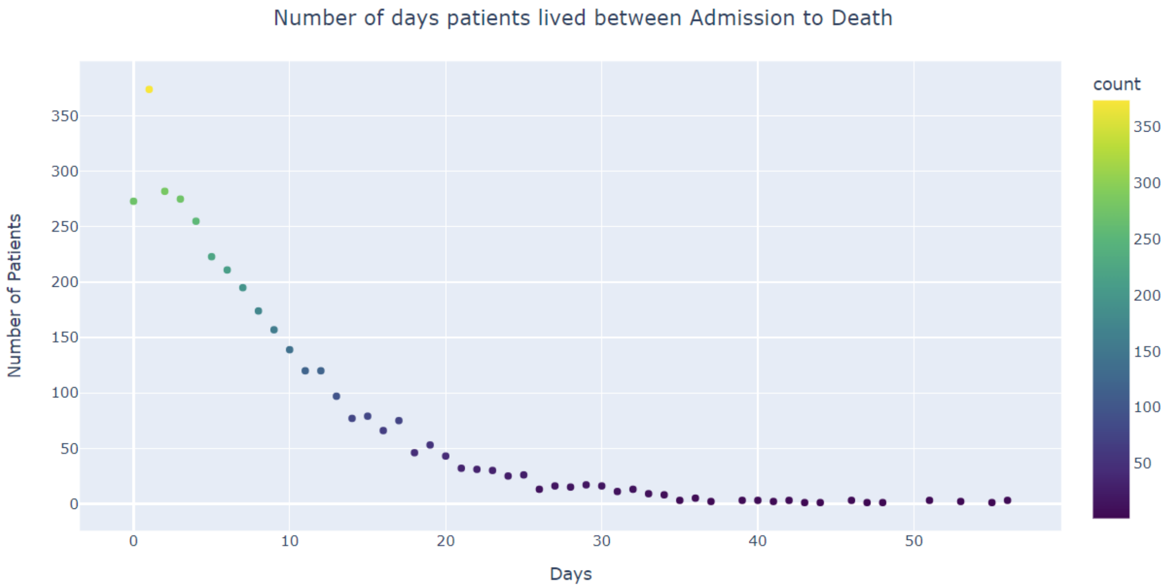
Time Series Analysis -



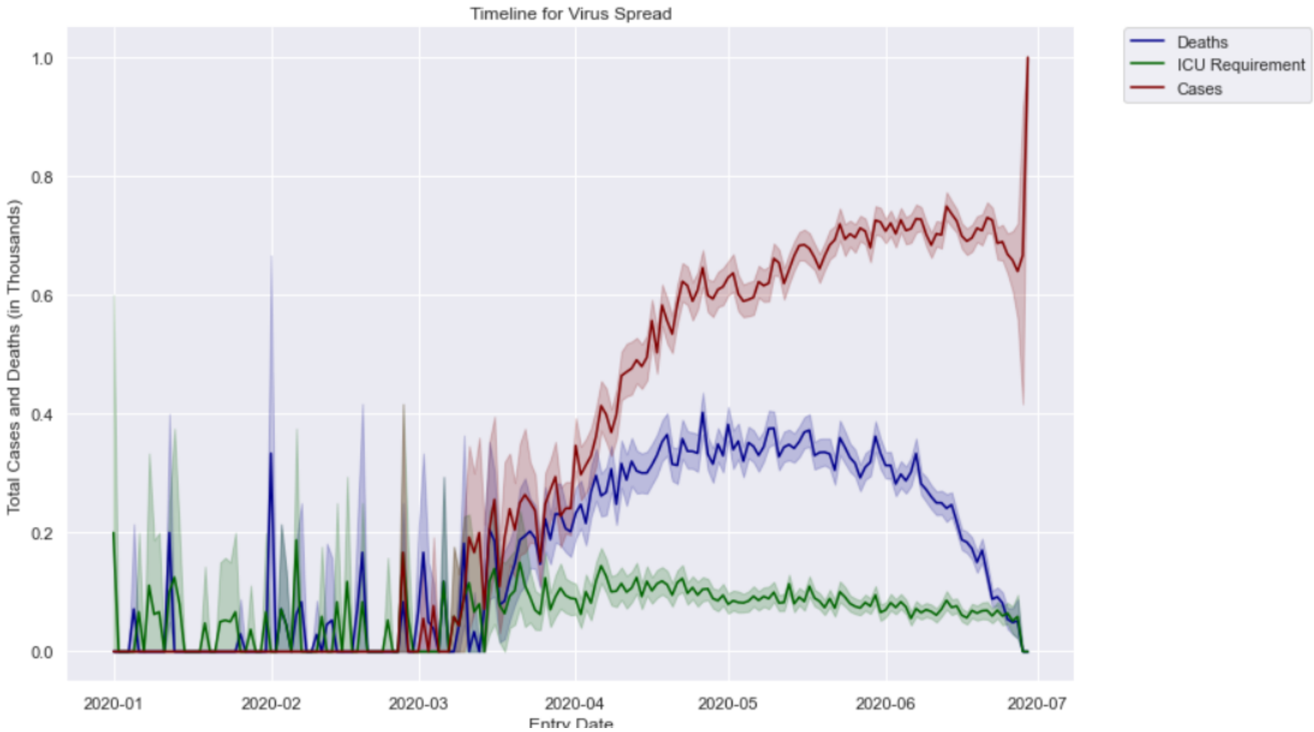
We can see the cases started rising around March and peaked around May and then gradually started to come down when restrictions were imposed by the Mexican government.

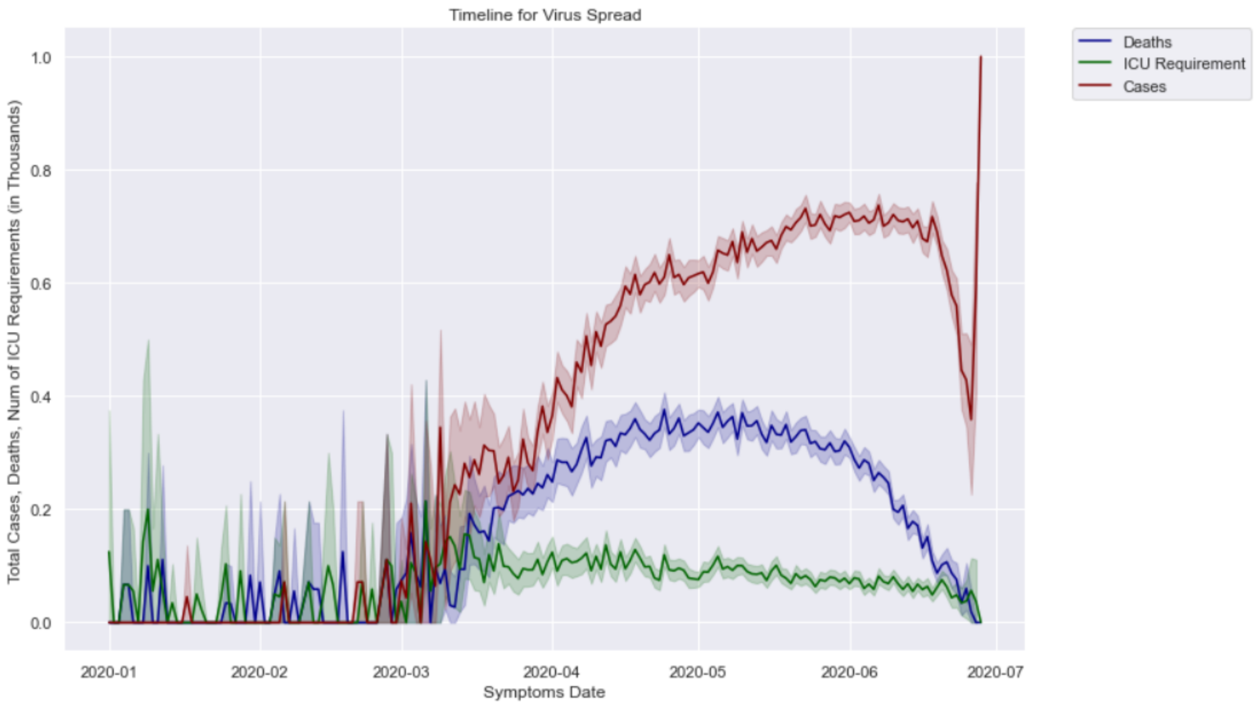


Patients who died after suffering a month are below 30. This could be because they had milder virus and the medical attention over weeks was working well to bring down deaths.



We can see the graph going down meaning the longer they are in the hospital the number of patients who died is reducing. It is likely that the medical care given was working and the medical team was in control of the situation.

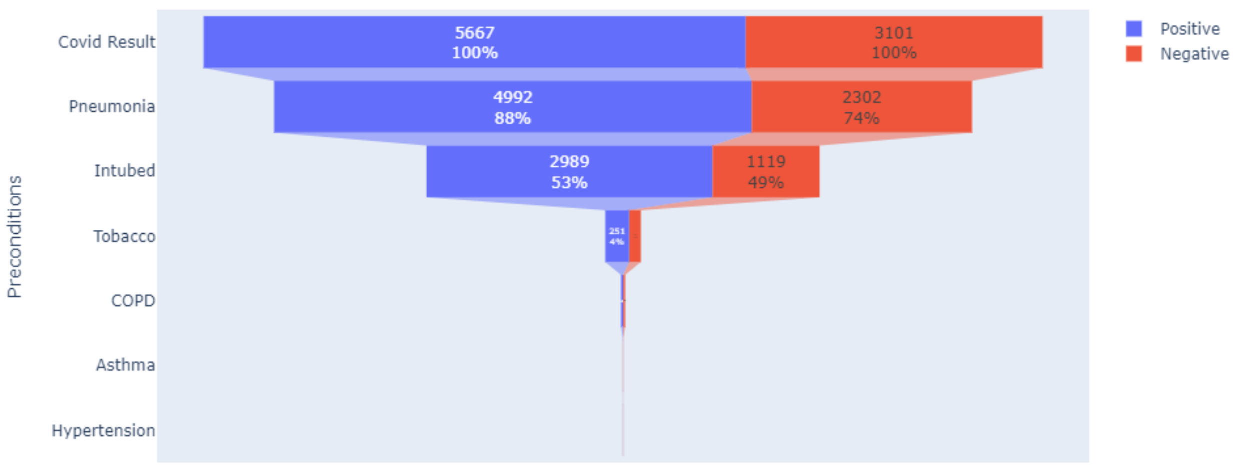




In the above two visualizations, we can observe that the peaks and downs in the ICU requirement, death cases, and constant for Covid cases suggests that they are due to other comorbidities the patients must be having. In addition to this, ICU requirement increases with Covid Cases and Deaths. The sudden rise in the Covid cases somewhere around the start of July could be due to the 2nd wave that had appeared. In Spite of the sudden rise in the Covid cases, we see a consistent decrease in deaths. Also, ICU requirement has a constant flow and seems to be decreasing slowly.

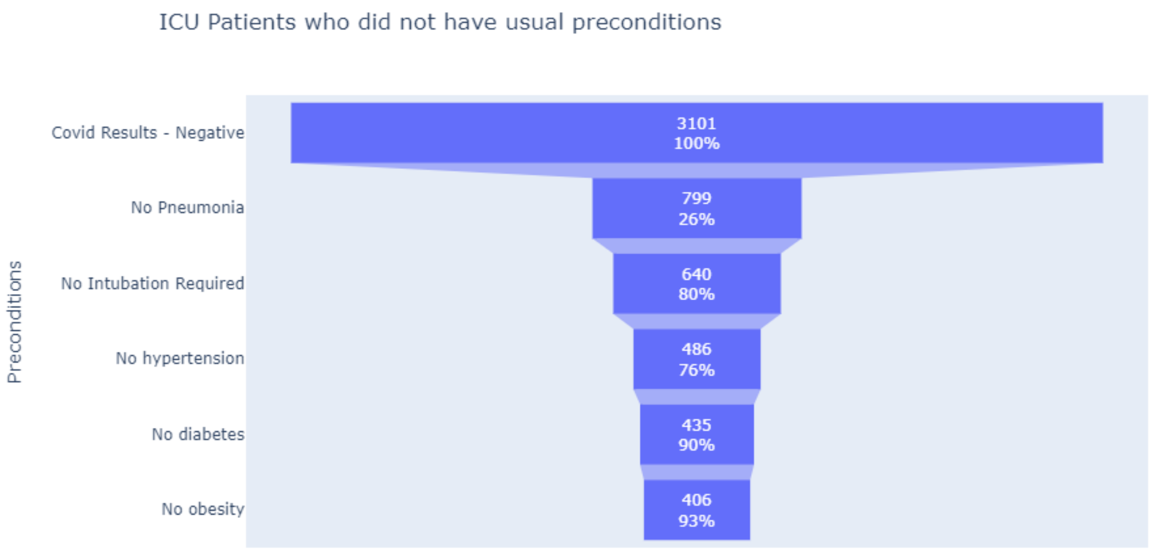
Funnel Plots -

ICU Patients who have most common preconditions



As we can see in the above visualization the distribution of ICU patients who have usual conditions is as follows:-

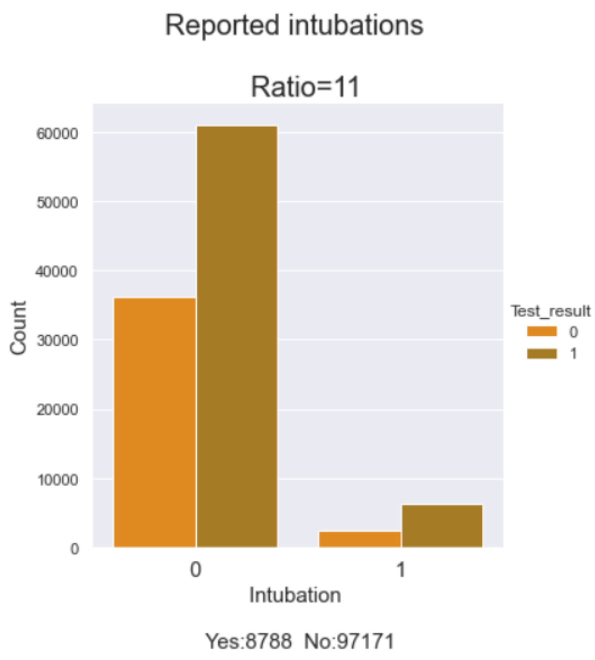
- Let us look at Lung related diseases or infections because corona virus is known to decapitate the lungs.
- There were 8768 ICU patients. 5667 patients are COVID +ve and 3101 patients are COVID-ve.
- Around 88% of COVID positive have pneumonia, similarly 74% of COVID negative have pneumonia.
- Around 53% of above pneumonia positive people needed intubation.
- Around 4% of the above people who needed intubation, use tobacco.
- So, Pneumonia appears like a strong reason to need ICU whether the COVID result is positive or negative

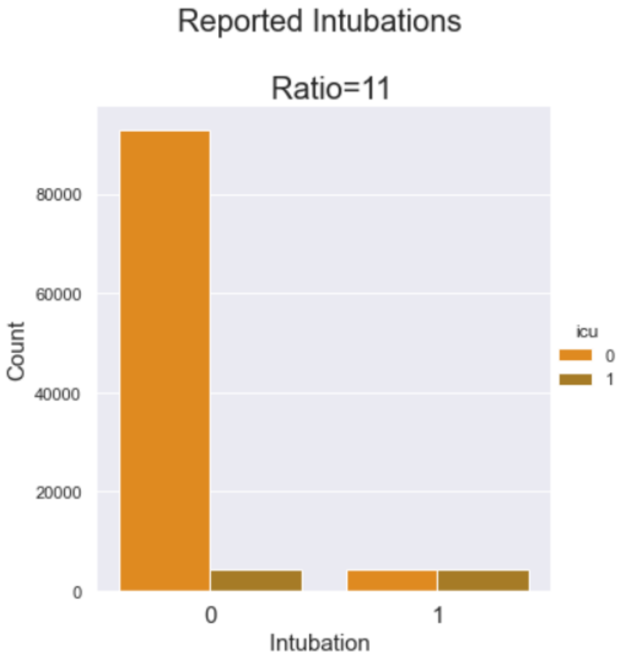


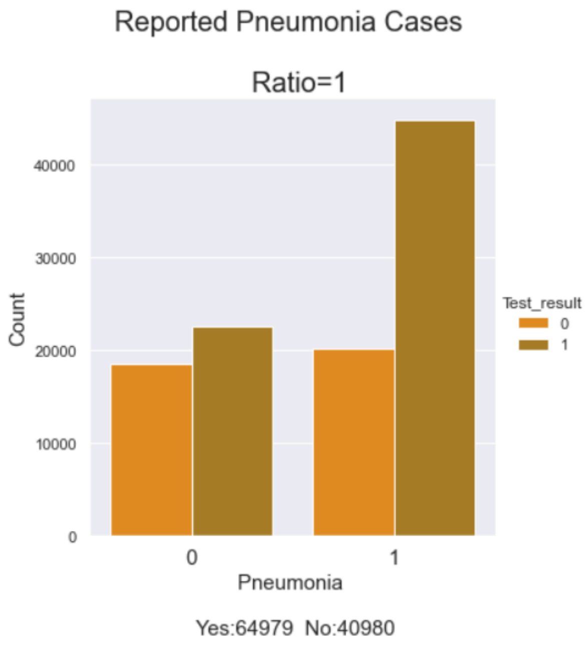
As we can see in the above visualization the distribution of ICU patients who did not have usual conditions is as follows:-

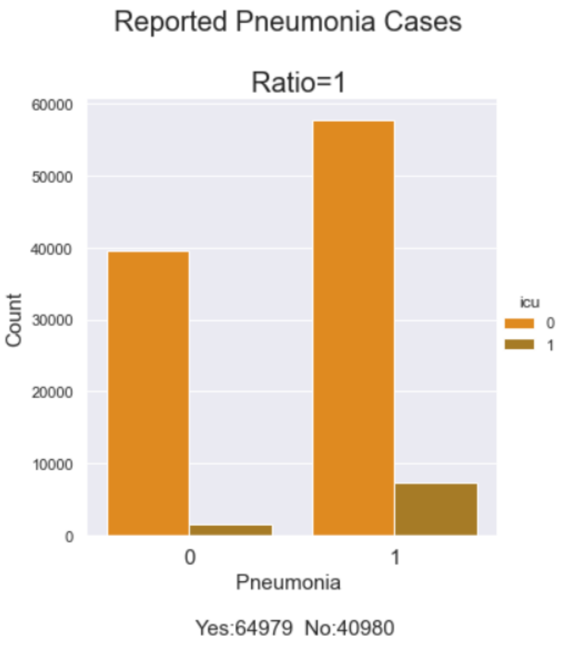
- We have 3101 patients that had a requirement for ICU but weren't tested positive for COVID.
- Around 26% of COVID negative have no pneumonia (primary reason to have an ICU admission).
- Around 80% of above pneumonia negative people needed NO intubation.
- As seen in the bottom bar, 406 patients who tested covid negative - are free of usual suspects (preconditions /co-morbidities)

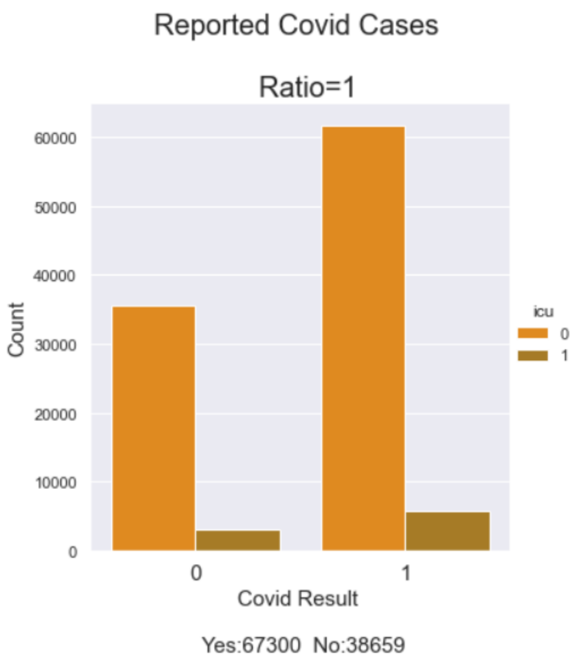
Below we have visualized **count plots** of some of the important features we found in our model -

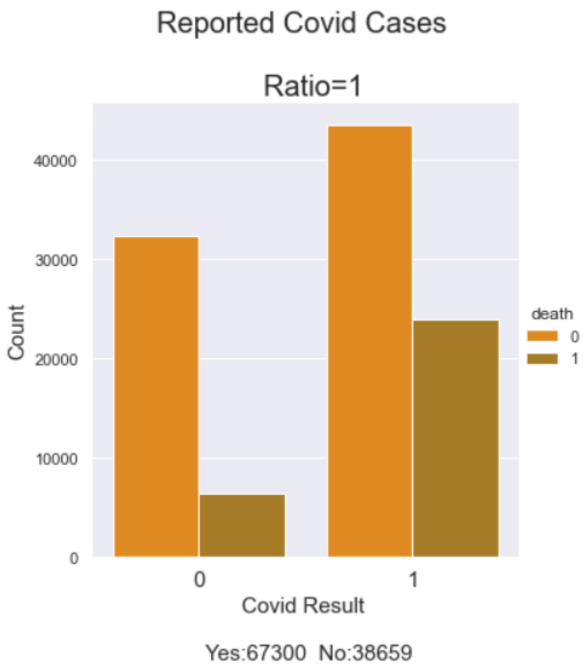






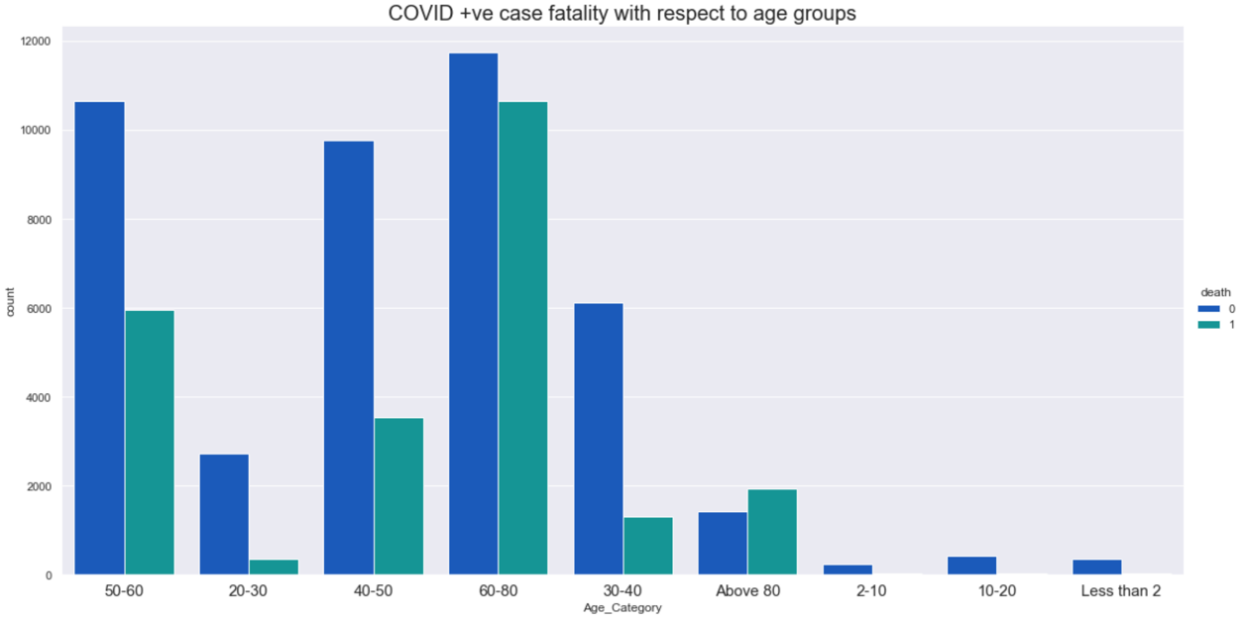






These count plots give us better understanding over patients requiring ICU. We can see and thus speculate that Patients with pre-conditions such as Pneumonia, Intubations because of surgery, or Covid have a high chance of ICU requirement.

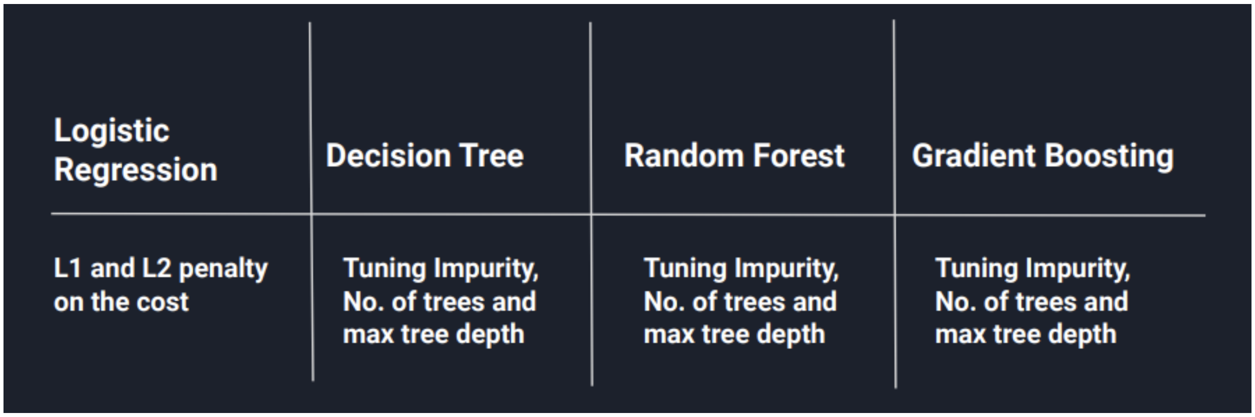
Let's understand the fatality rate distribution with respect to different Age groups.



As we can see from the above distribution of Age Category, Elderly people and senior citizens in the age range of 50 to 80 have the highest fatality rate because of Covid.

Machine Learning Models

Below are the 4 supervised machine learning algorithms that we have used to classify ICU requirements, likelihood of testing positive and factors that affect fatality rate.



Logistic Regression	Decision Tree	Random Forest	Gradient Boosting
L1 and L2 penalty on the cost	Tuning Impurity, No. of trees and max tree depth	Tuning Impurity, No. of trees and max tree depth	Tuning Impurity, No. of trees and max tree depth

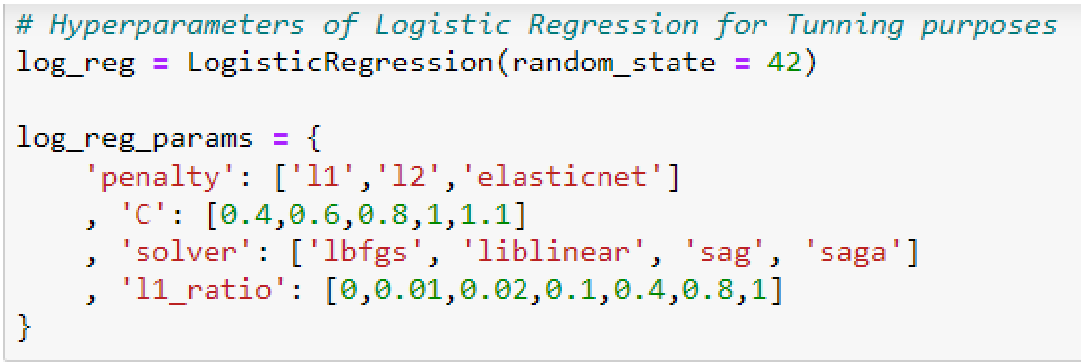
Hyperparameter tuning was performed on each model using grid search and 3-fold cross validation. Since the classes were imbalanced, we tried to focus more on precision and recall. We have used sklearn for building models. Below, we present the best performing model and important features in each case as well as consolidated results at the end.

We decided to build above models using below case statements in order to better understand the ICU requirements based on patients who have died and are still hospitalized, utilizing various pre-conditions as a basis: -

- Case1: Predicting death for patients requiring ICU
- Case2: Predicting death for patients not requiring ICU
- Case3: Predicting ICU for died patients
- Case4: Predicting ICU for hospitalized patients
- Case5: Predicting Covid

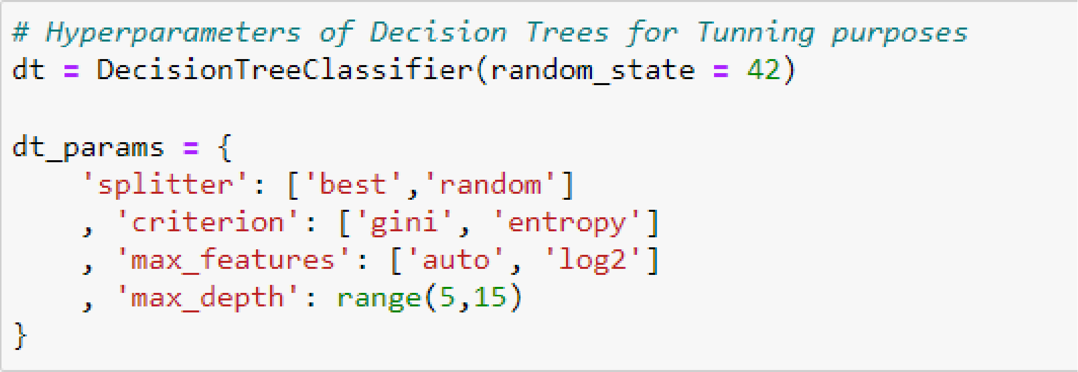
Using the results obtained from above cases and exploratory analysis will help us to give suggestions to the hospital on how to handle the patient's ICU requirement of ICU based on certain pre-conditions.

Below are the hyperparameters considered for each model -



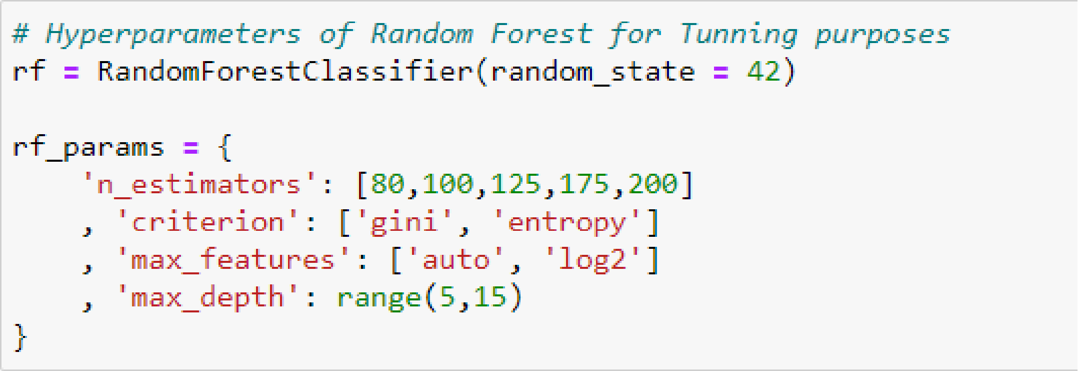
```
# Hyperparameters of Logistic Regression for Tunning purposes
log_reg = LogisticRegression(random_state = 42)

log_reg_params = {
    'penalty': ['l1','l2','elasticnet']
    , 'C': [0.4,0.6,0.8,1,1.1]
    , 'solver': ['lbfgs', 'liblinear', 'sag', 'saga']
    , 'l1_ratio': [0,0.01,0.02,0.1,0.4,0.8,1]
}
```



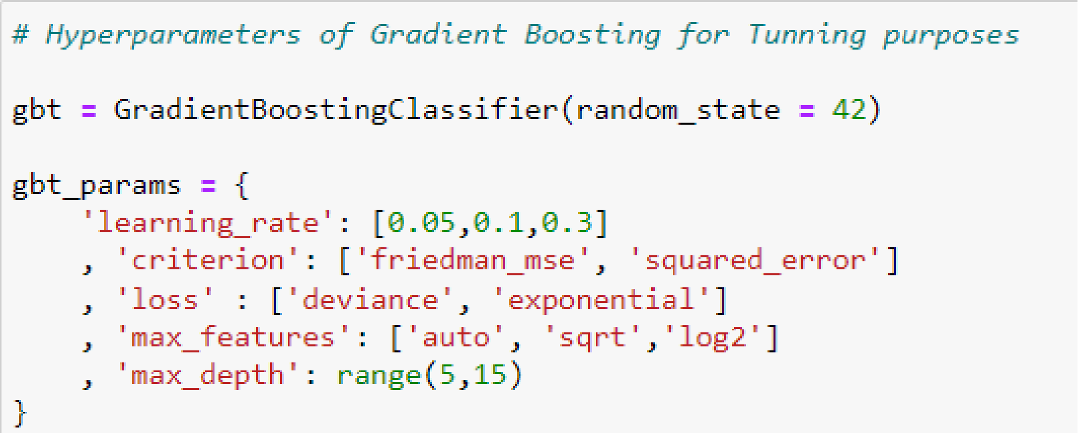
```
# Hyperparameters of Decision Trees for Tunning purposes
dt = DecisionTreeClassifier(random_state = 42)

dt_params = {
    'splitter': ['best','random']
    , 'criterion': ['gini', 'entropy']
    , 'max_features': ['auto', 'log2']
    , 'max_depth': range(5,15)
}
```



```
# Hyperparameters of Random Forest for Tunning purposes
rf = RandomForestClassifier(random_state = 42)

rf_params = {
    'n_estimators': [80,100,125,175,200]
    , 'criterion': ['gini', 'entropy']
    , 'max_features': ['auto', 'log2']
    , 'max_depth': range(5,15)
}
```



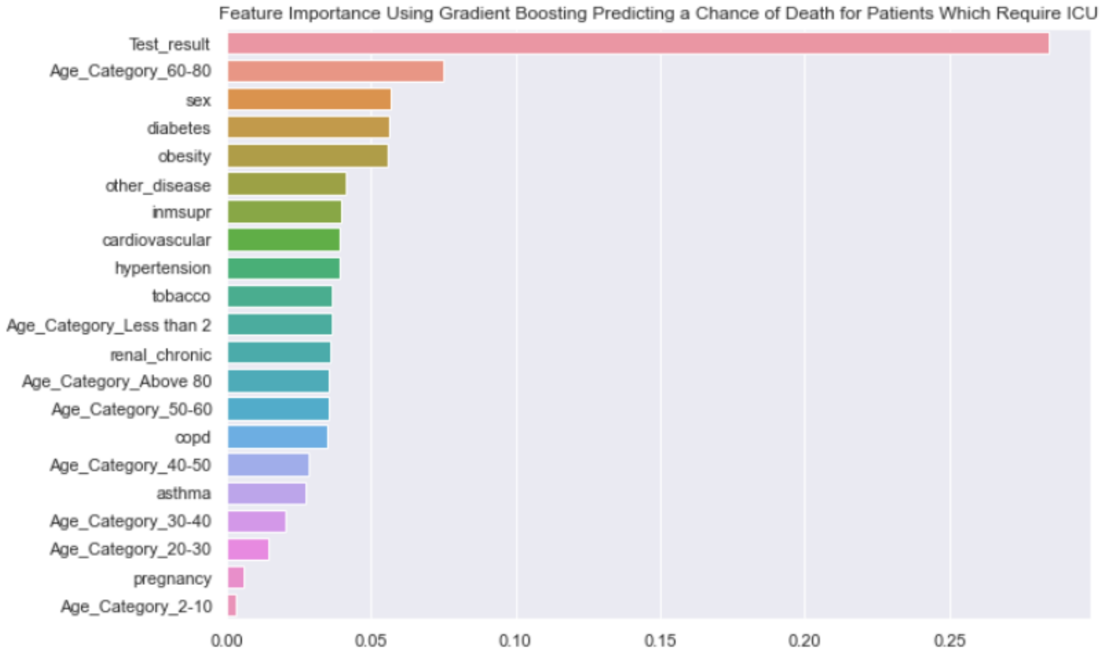
```
# Hyperparameters of Gradient Boosting for Tunning purposes

gbt = GradientBoostingClassifier(random_state = 42)

gbt_params = {
    'learning_rate': [0.05,0.1,0.3]
    , 'criterion': ['friedman_mse', 'squared_error']
    , 'loss' : ['deviance', 'exponential']
    , 'max_features': ['auto', 'sqrt','log2']
    , 'max_depth': range(5,15)
}
```

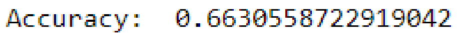
❖ Case 1: Death as Target variable for patients who required ICU

Below are the important features obtained by applying Gradient Boosting.

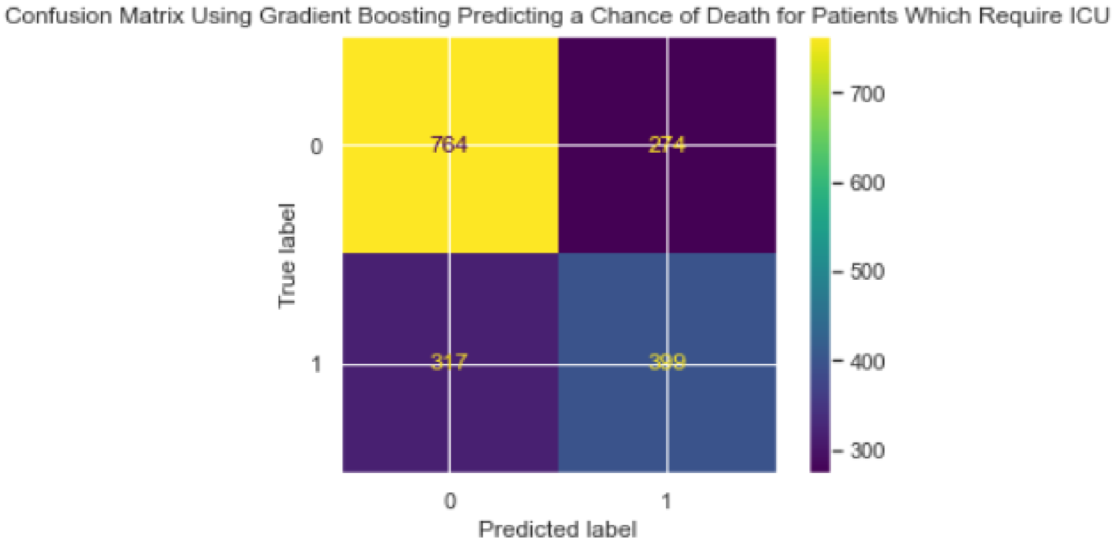


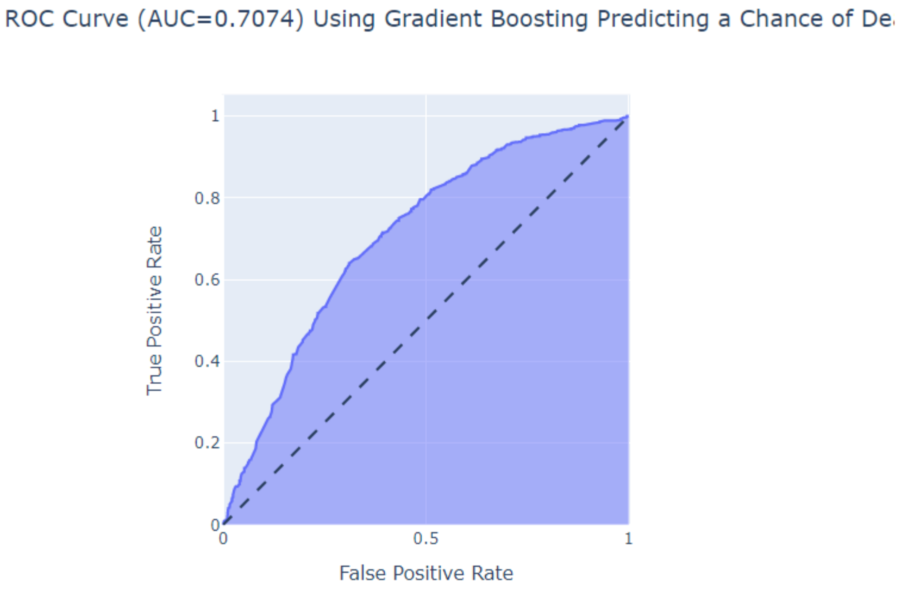
We can see from the top 2 important features, that for patients who required ICU, having COVID and if the patient's age is between 60 and 80 could be critical cases leading to a patient's death.

Accuracy of our model is 66.30% and area under ROC Curve is 0.7074



```
Accuracy:  0.6630558722919042
```

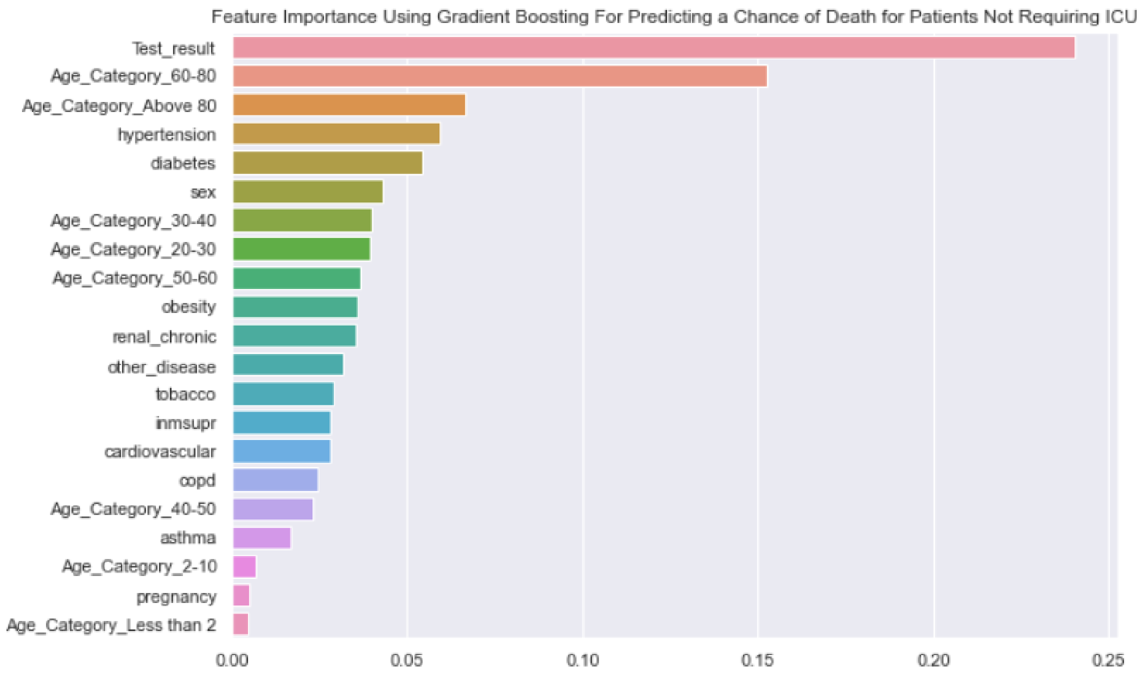




Accuracy and Area under ROC curve tells us that if a patient requires ICU then there is a roughly 66% chance of predicting whether the medical pre-conditions could lead that patient to death.

- **Case 2: Death as Target variable for patients who did not require ICU (trying to study what factors caused death even though patients did not require ICU).**

Below are the important features obtained by applying Gradient Boosting.

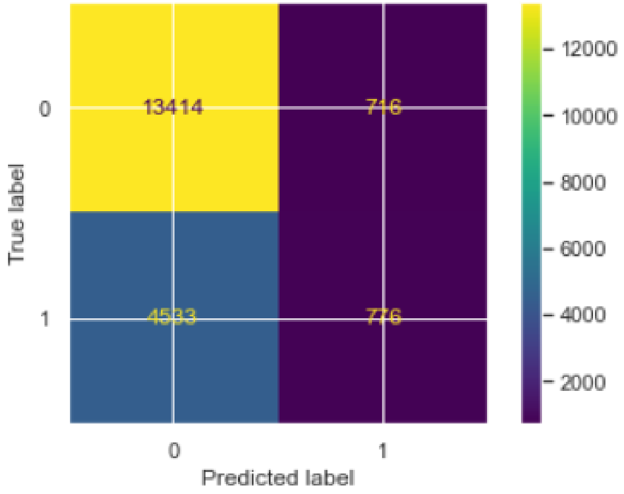


We can see from the top 5 important features, that for patients who don't require ICU, having COVID and if the patient's age is between 60 and 80 or above 80 could be critical cases leading to a patient's death. On top of this, for patients having hypertension and diabetes as medical pre-conditions could lead to death as well.

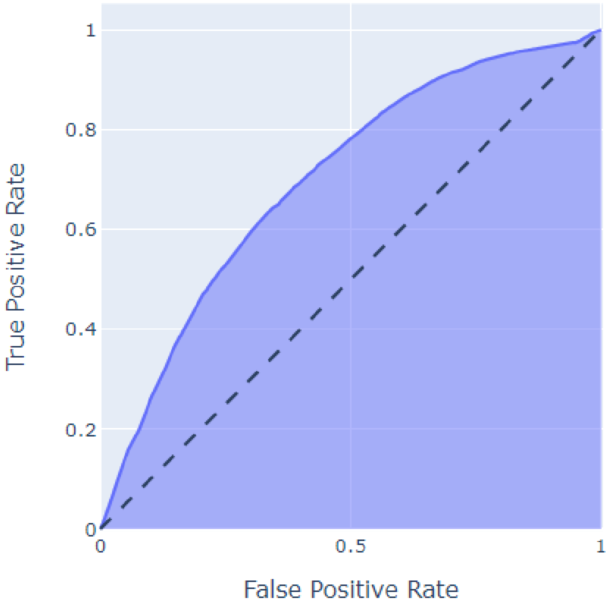
Accuracy of our model in this case is 73% and area under ROC curve is 0.6994.

```
Accuracy:  0.729975821801533
```

Confusion Matrix Using Gradient Boosting For Predicting a Chance of Death for Patients Not Requiring ICU



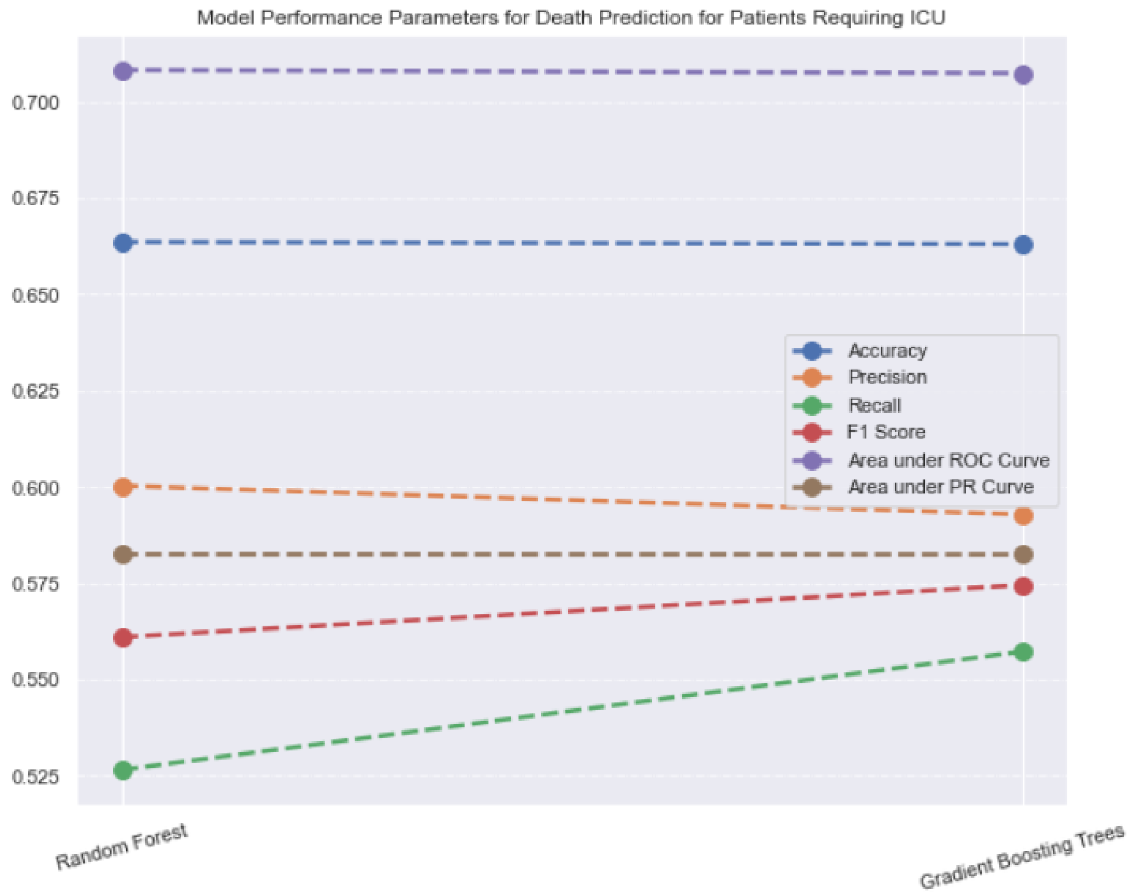
ROC Curve (AUC=0.6994) Using Gradient Boosting For



Accuracy and Area under ROC curve tells us that if a patient doesn't require ICU then there is a roughly 73% chance of predicting whether the medical pre-conditions, especially for elderly patients could lead to patient's death.

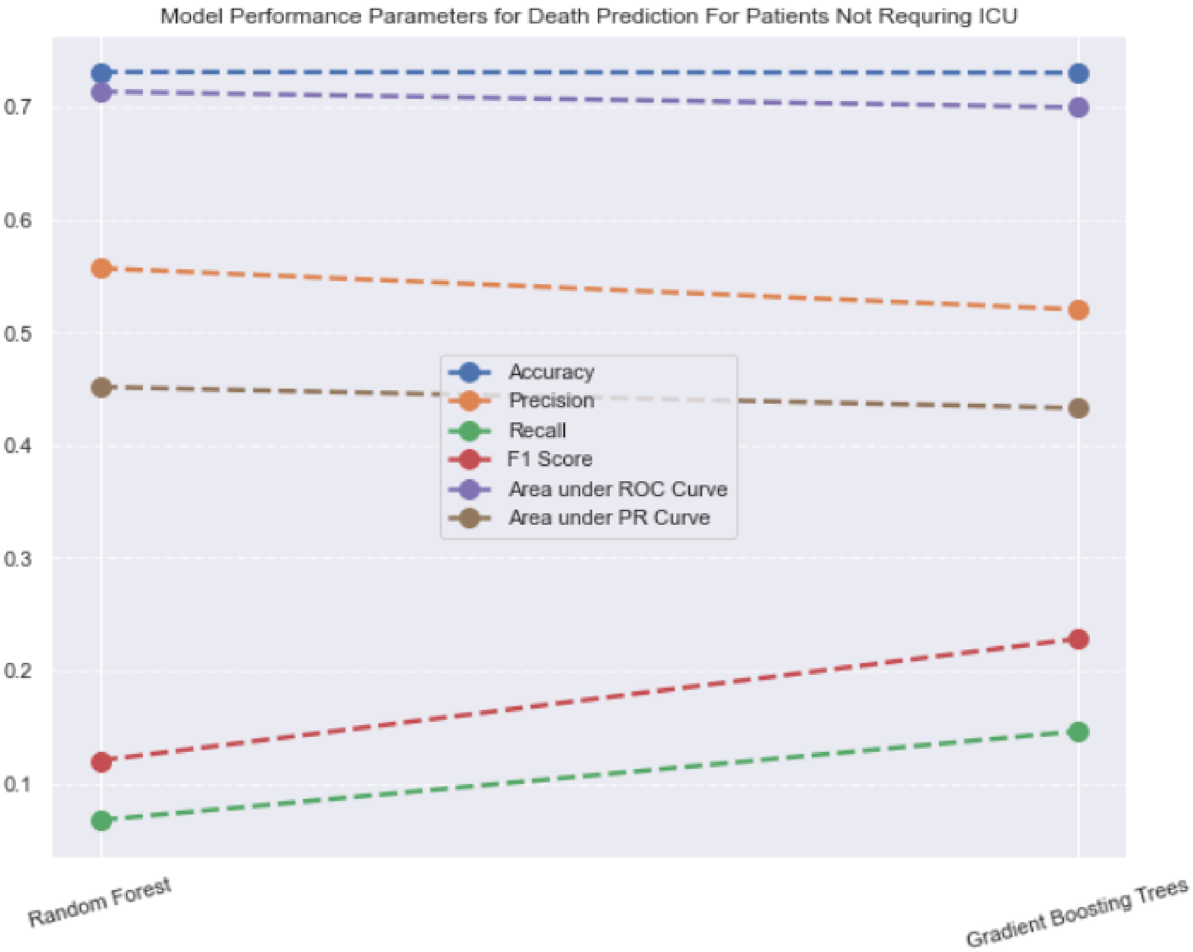
❖ Model Evaluations for Case 1 and Case 2:

Case 1 Model: -



The higher recall given by the Gradient Boosting model confirms that we don't want to classify that the patient which is dead, is still alive given that the patient requires ICU. This made us consider it as the best model.

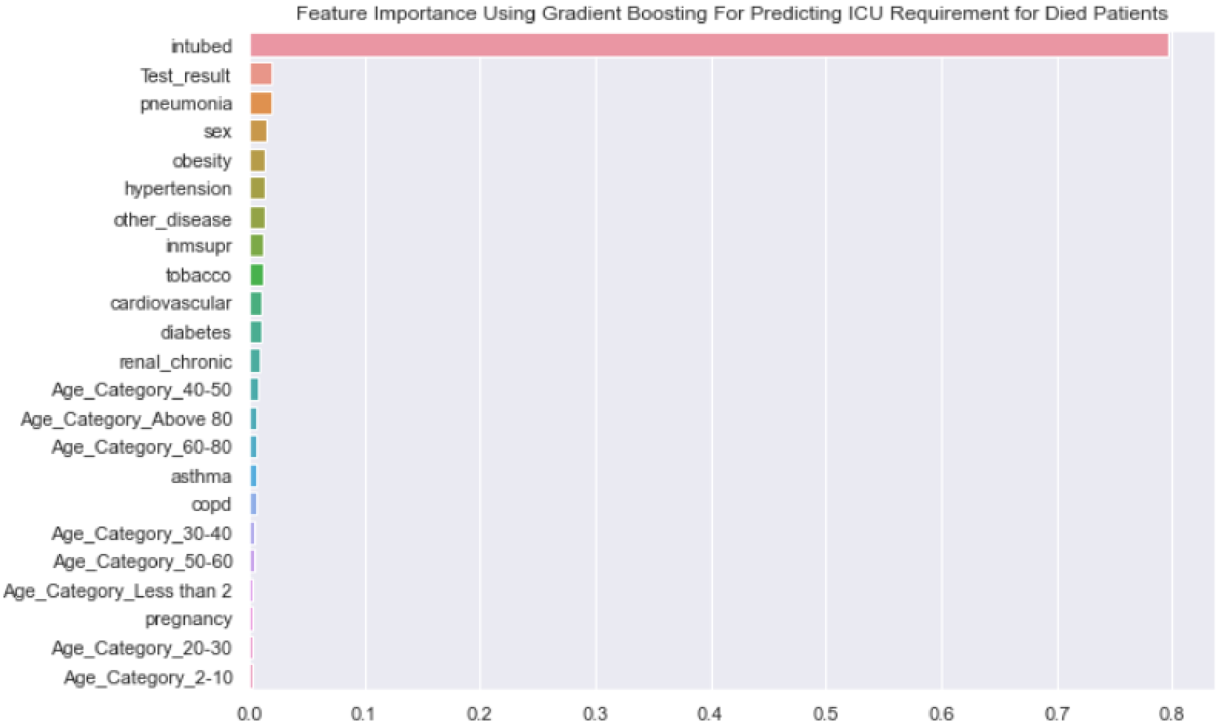
Case 2 Model: -



The higher recall given by the Gradient Boosting model confirms that we don't want to classify that the patient which is dead, is still alive given that the patient does not require ICU. This made us consider it as the best model.

❖ Case 3: Analyzing whether patients who died could have been saved if given ICU treatment

Below are the important features obtained by applying Gradient Boosting.

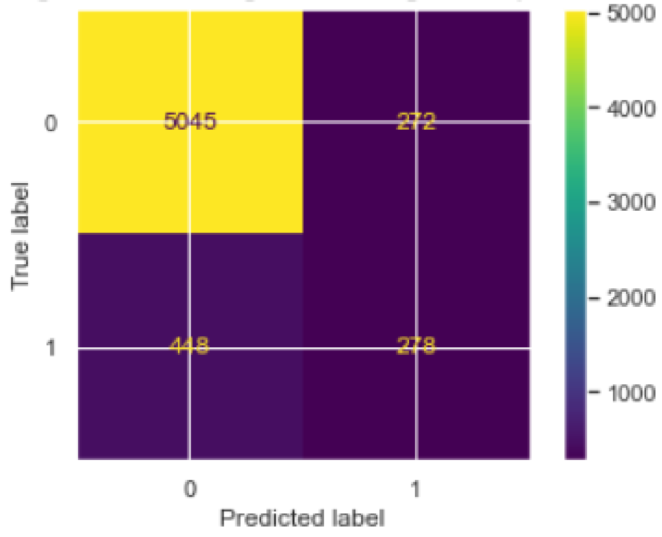


We can see from the top features, that for patients who died having previous intubation and is a Covid positive patient and on top of that suffered from pneumonia could have been saved if given ICU treatment.

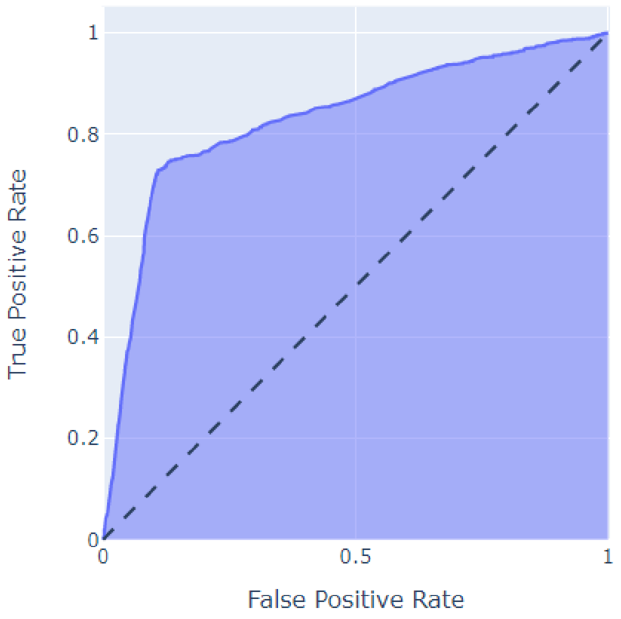
Accuracy of our model in this case is 88% and area under ROC curve is 0.8304.

```
Accuracy:  0.8808538805229191
```

Confusion Matrix Using Gradient Boosting For Predicting ICU Requirement for Died Patients



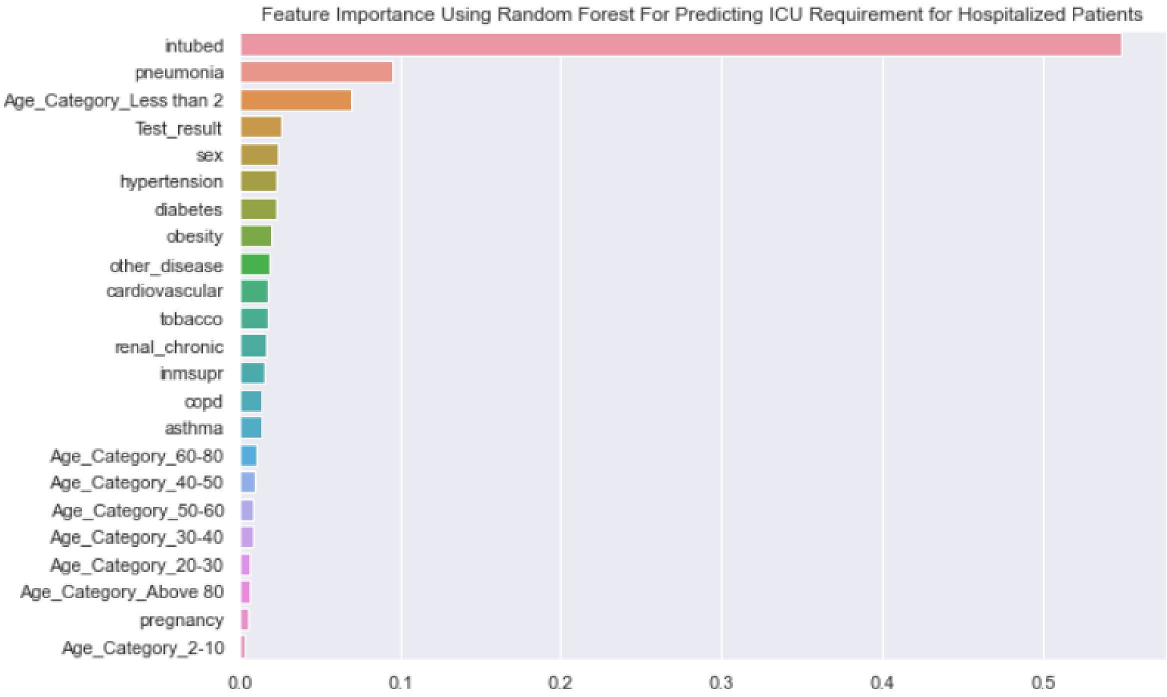
ROC Curve (AUC=0.8307) Using Gradient Boosting For Predicting ICU



Accuracy and Area under ROC curve tells us that based on died patient data, current hospitalized patients having previous intubations, Covid or pneumonia could be saved if given ICU treatment.

❖ Case 4: Predicting ICU for hospitalized patients

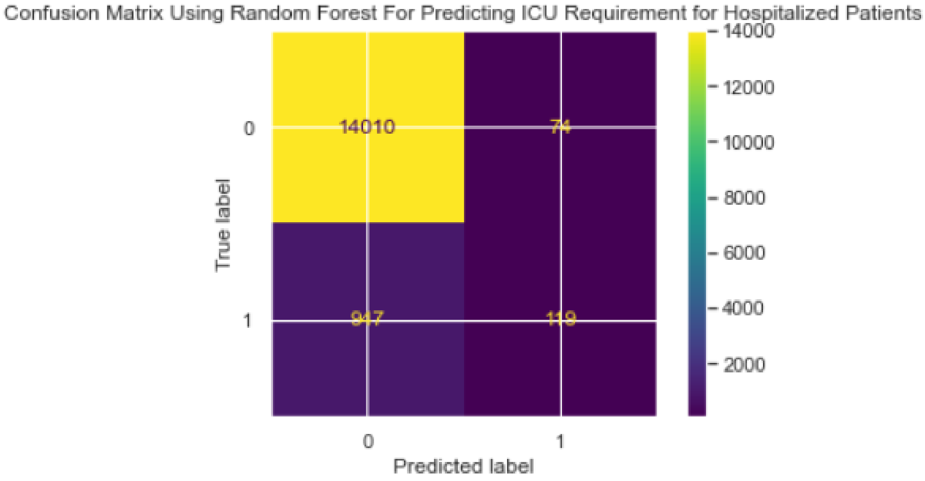
Below are the important features obtained by applying Random Forest..

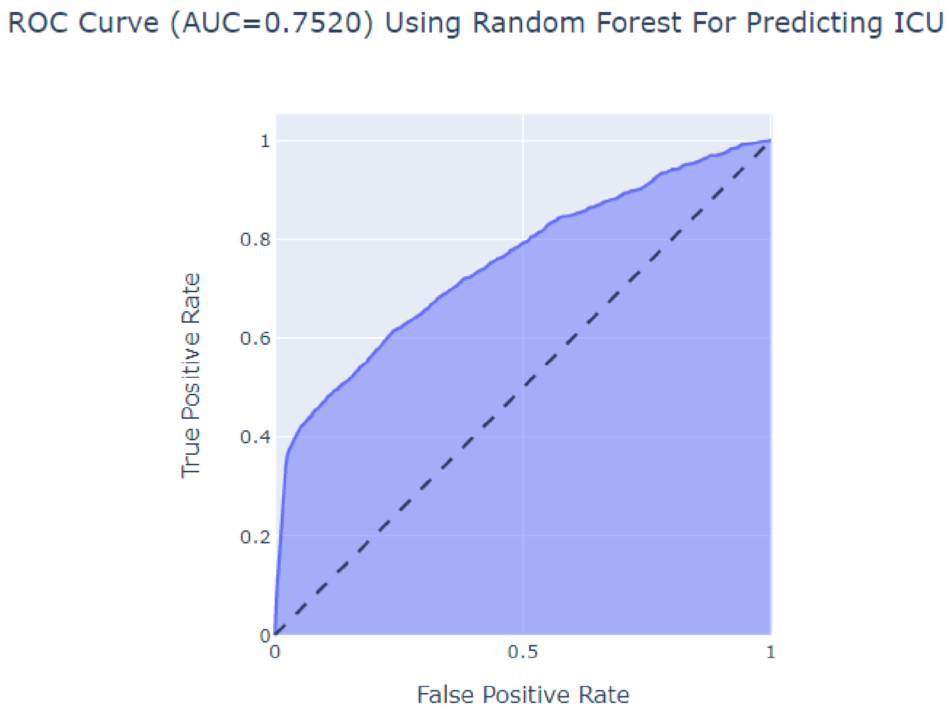


We can see from the top features that for patients who are hospitalized having previous intubation and suffering from Pneumonia or had pneumonia need ICU care. In addition to this, if a patient lies in the age category less than 2 or if a patient is Covid positive should be given ICU treatment.

Accuracy of our model in this case is 93% and the area under the ROC curve is 0.7520.

```
Accuracy:  0.9326072607260726
```

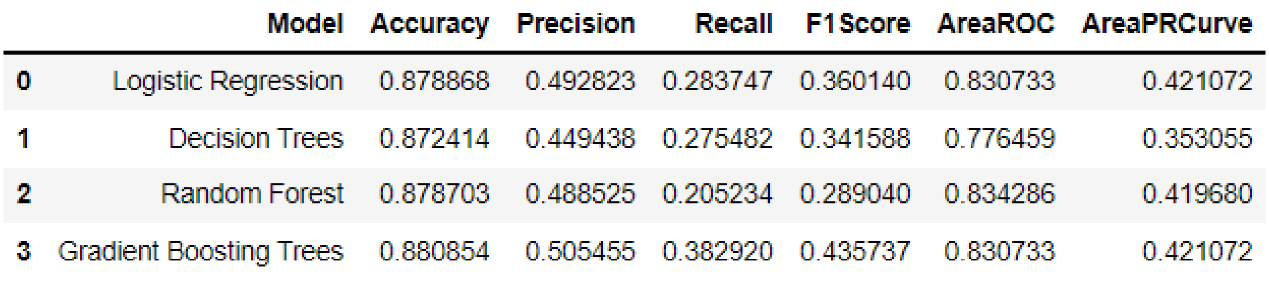




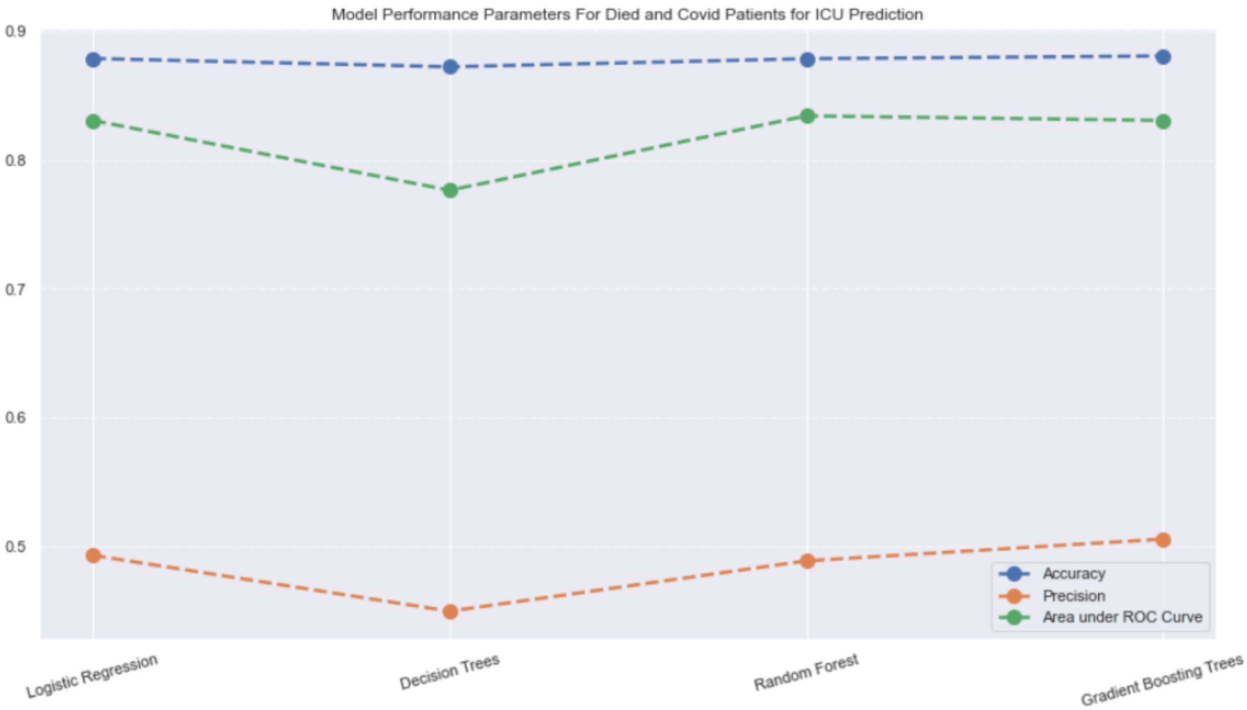
Accuracy and Area under ROC curve tells us that based on hospitalized patient data, patients having previous intubations, pneumonia, infants or age less than 2 years and Covid patients should be given ICU treatment.

❖ Model Evaluations For Case 3 and Case 4:

<u>Case 3 Model: -</u>

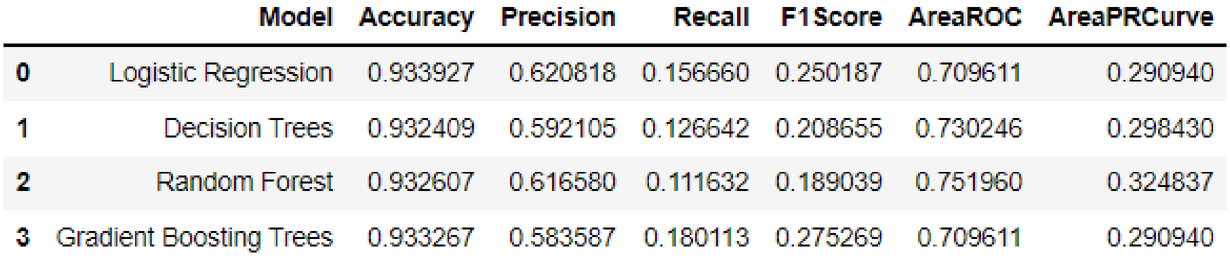


	Model	Accuracy	Precision	Recall	F1Score	AreaROC	AreaPRCurve
0	Logistic Regression	0.878868	0.492823	0.283747	0.360140	0.830733	0.421072
1	Decision Trees	0.872414	0.449438	0.275482	0.341588	0.776459	0.353055
2	Random Forest	0.878703	0.488525	0.205234	0.289040	0.834286	0.419680
3	Gradient Boosting Trees	0.880854	0.505455	0.382920	0.435737	0.830733	0.421072

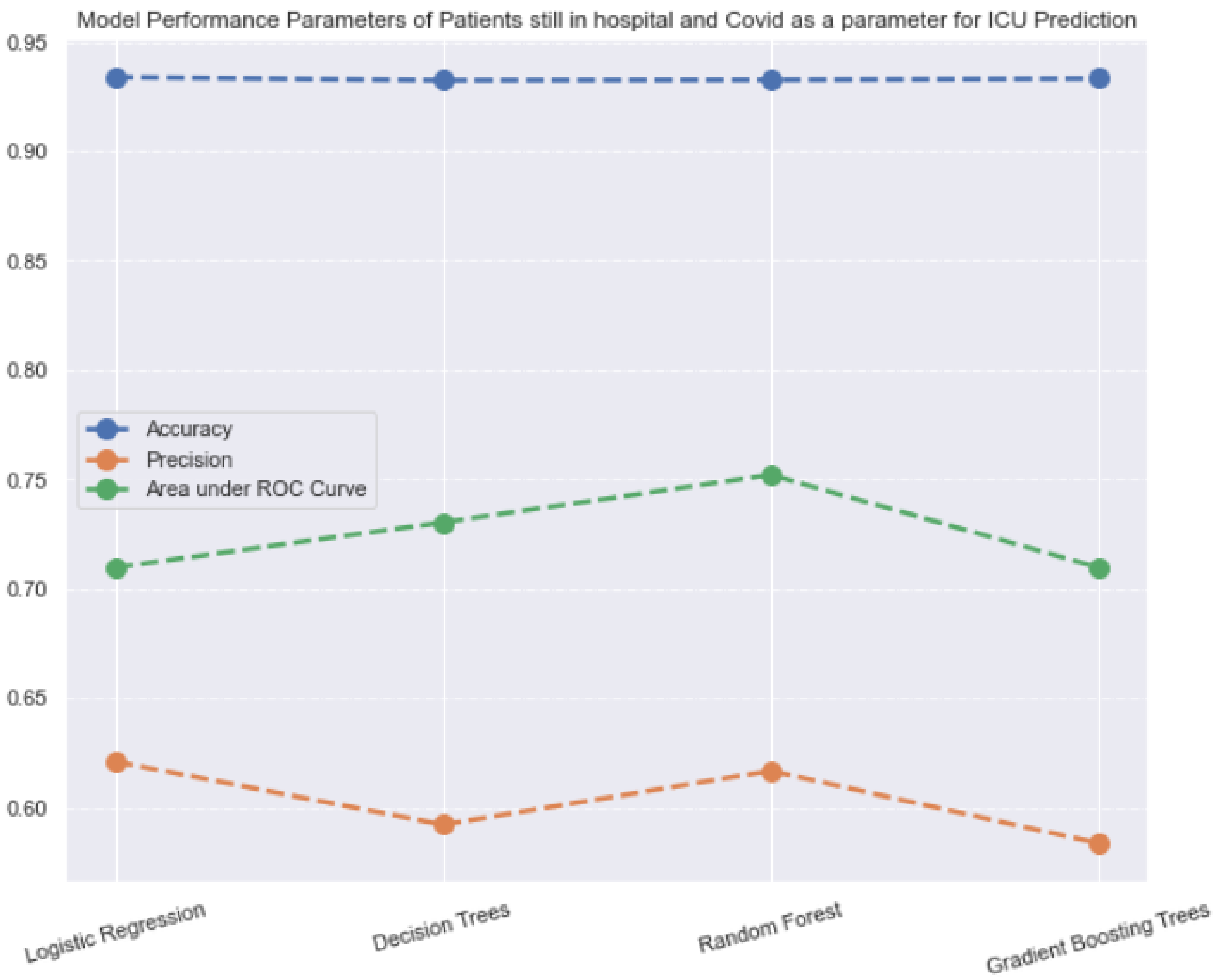


High precision given by the Gradient Boosting model shows us that on the basis of the feature importance we are able to predict the requirement of ICU to a very substantial point which could have saved patient’s lives.

Case 4 Model: -



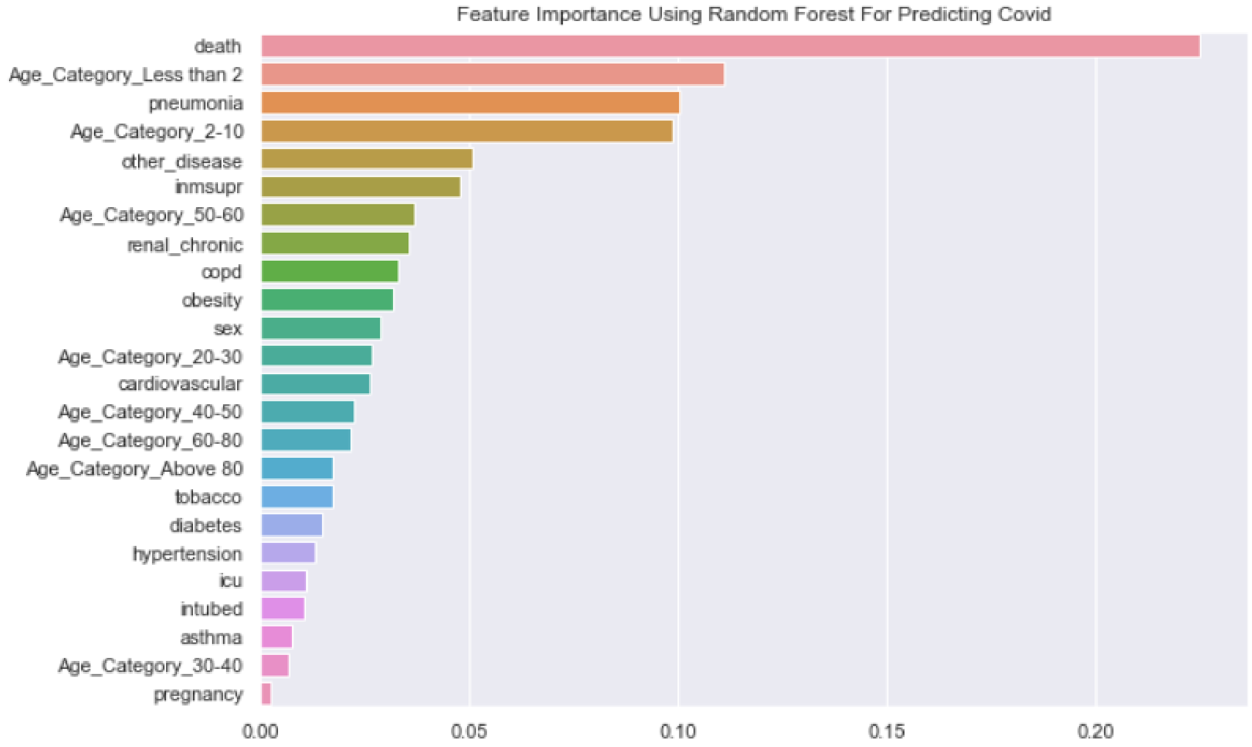
	Model	Accuracy	Precision	Recall	F1Score	AreaROC	AreaPRCurve
0	Logistic Regression	0.933927	0.620818	0.156660	0.250187	0.709611	0.290940
1	Decision Trees	0.932409	0.592105	0.126642	0.208655	0.730246	0.298430
2	Random Forest	0.932607	0.616580	0.111632	0.189039	0.751960	0.324837
3	Gradient Boosting Trees	0.933267	0.583587	0.180113	0.275269	0.709611	0.290940



High precision given by the Random Forest model shows us that on the basis of the feature importance we are able to predict the requirement of ICU to a very substantial point for hospitalized patients.

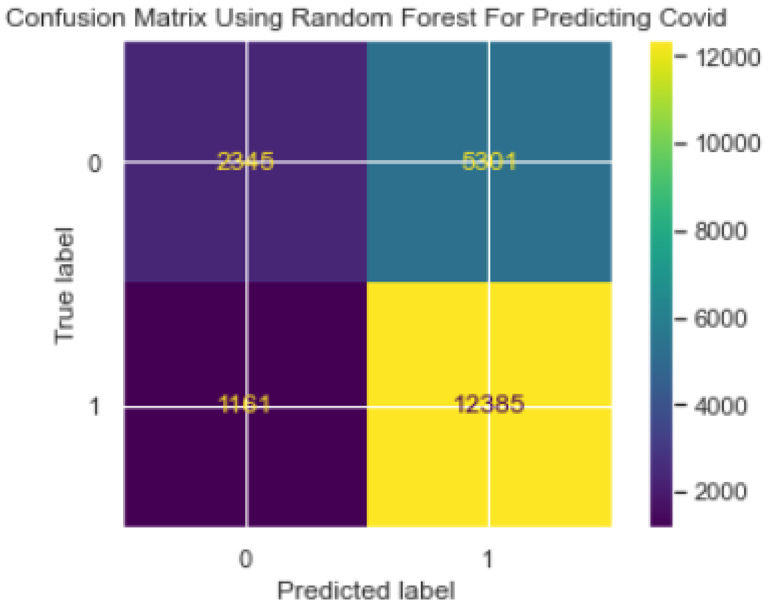
❖ Case 5: Predicting whether a patient is likely to test positive for COVID

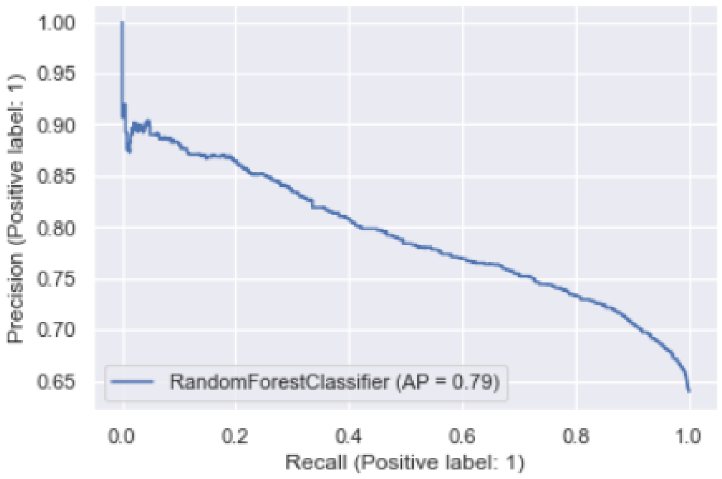
Below are the important features obtained by applying Random Forest..



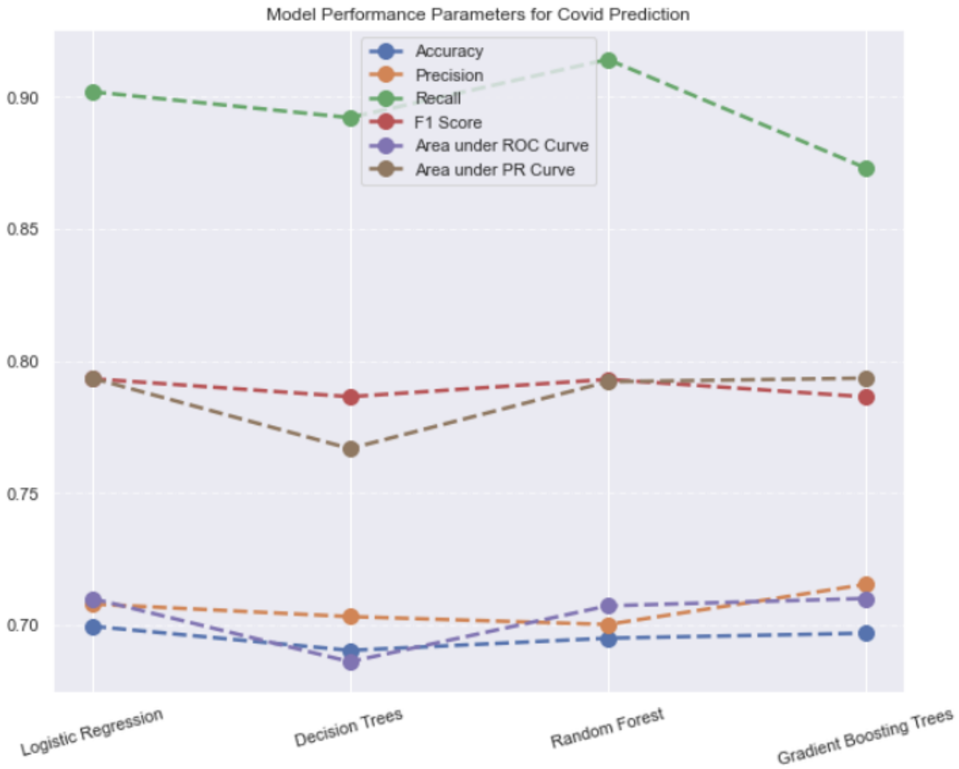
Accuracy of our model is 69% and the area under Precision-Recall Curve is 0.79.

```
Accuracy:  0.6950736126840317
```





Below plot represents consolidated results for every algorithm used -



We can see that our recall is high which was our main target since we do not want to classify any patients having covid as not having covid.

Accuracy and Area under precision recall curve tells us that based on the patient data, the death rate is higher because of Covid patients. Also if a patient lies in the age of 0 to 10 or has pneumonia, other diseases or immunosuppression has a high chance of having Covid result as positive.

Conclusion and Suggestions:

The exploratory analysis and the model evaluation results led us to actionable insights in order to decide whether a patient needs to be given ICU or not considering the reasons that had caused fatality and the importance of Covid and other pre-conditions to get to that result.

- If any patient is hospitalized with a positive test result, he should be given ICU urgently if they have pneumonia or they are infants.

- Large number of deaths are caused by COVID instead of other comorbidities.

- Patients in the age range of 0 to 10 and in the age range of 50 to 60 are more prone to Covid if they are suffering from medical preconditions such as pneumonia, immunosuppression and other diseases.

- Mortality rate among patients who were hospitalized the day they started showing symptoms is higher, maybe due to higher severity of Covid than other pre-conditions based on the feature Importance.

- The longer patients are in the hospital the mortality rate is reducing. It is likely that the medical care given was working.

Future Scope:

- Data acquisition for Criticality for all the pre-conditions including Covid.

- Discharge data (DateTime) from the hospital would really help us to understand the reduction in the Mortality rate and Covid positive rate.

- Applying these models over different country's patients data to gather understanding over effects of patients pre-conditions on factors such as death rate, ICU requirement and criticality of Covid.

- Data generated in the year 2021 to understand the developments happening in the Covid-19 Variants such as delta and omicron.

References

- https://www.gob.mx/salud/documentos/datos-abiertos-152127
- https://www.hopkinsmedicine.org/coronavirus/articles/icu-recovery.html
- https://medicalxpress.com/news/2021-12-severe-covid-patients-mortality-icu.html
- https://scikit-learn.org/stable/index.html