**Features preparation**

* Patient file : include gender and age features
* Observation file- Filter for covid patients and include top 20 occuring observation in covid positive patients which showed positive feature importance while feeding them into decision tree.
* Condition file - Include conditions such as 'Acute deep venous thrombosis (disorder)', 'Acute pulmonary embolism (disorder)', 'Acute respiratory distress syndrome (disorder)', 'Acute respiratory failure (disorder)', 'Hypoxemia (disorder)'and 'Sepsis caused by virus (disorder) since these features helped in decreasing MAE(Mean absolute error) while feeding them into decision tree. Consider number of days a patient suffered from a particular condition, if end date is not mentioned then consider maximum date value from given train data.

**Model**

Split data into train:test=70:30 ratio

Decision tree using Grid\_Search\_CV for hyperparameter tuning.

Model metric:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Baseline(Compare with median value from train data) | Train data | Test data |
| RMSE(Root mean square error) | 5.01 | 3.16 | 3.23 |
| MAE(Mean absolute error) | 4.10 | 2.32 | 2.35 |
|  |  |  |  |

Feature importance and SHAP summary plot:

|  |  |
| --- | --- |
|  |  |

**Risk Factors**

* Some features identified as important from train data might not be present which might lead to have zeros included.