Indian institute of information technology Allahabad

Course :- Introduction to machine learning

project :- Covid-19 severity prediction based on symptoms

Guide : - Dr.K.P.singh

Student Name	*Student ID*
Spandan Roy	MIT2021021
Akshay Jain	MIT2021038
Shivarajkumar G	MIT2021042

Contents

1 Abstract

Accurately assessing the severity of COVID-19 patients at an early stage is an effective way to increase patient survival. Identifying and triaging the people at highest risk of complications that can result in patient mortality is indeed a difficult problem, particularly in developing countries around the world. This problem is exacerbated by a scarcity of specialists. Using machine learning (ML) techniques to predict the severity of people with COVID-19 in the initial screening process can be an effective method for sorting and treating patients, allowing them to receive appropriate clinical management while making the best use of medical facilities.

In this project, we will use symptoms to predict the severity of the covid-19 disease. We are using different techniques of machine learning to predict the Severity of Covid-19 disease, and analyzing their performances for the predcition. As it is a classification task at the end we also tried implementing ROC graph and analysed using AUC which really is one of the best way to analyse any classification task.

2 introduction

The SARS-CoV-2-caused novel coronavirus disease 2019 (COVID-19) pandemic remains a critical and urgent threat to global health. The outbreak began in early December 2019 in the People's Republic of China's Hubei province and has since spread throughout the world.

As of October 2020, the total number of confirmed patients with the disease had surpassed 39,500,000 in over 180 countries, though the actual number of people infected is likely much higher. COVID-19 has killed over 1,000,000 people.

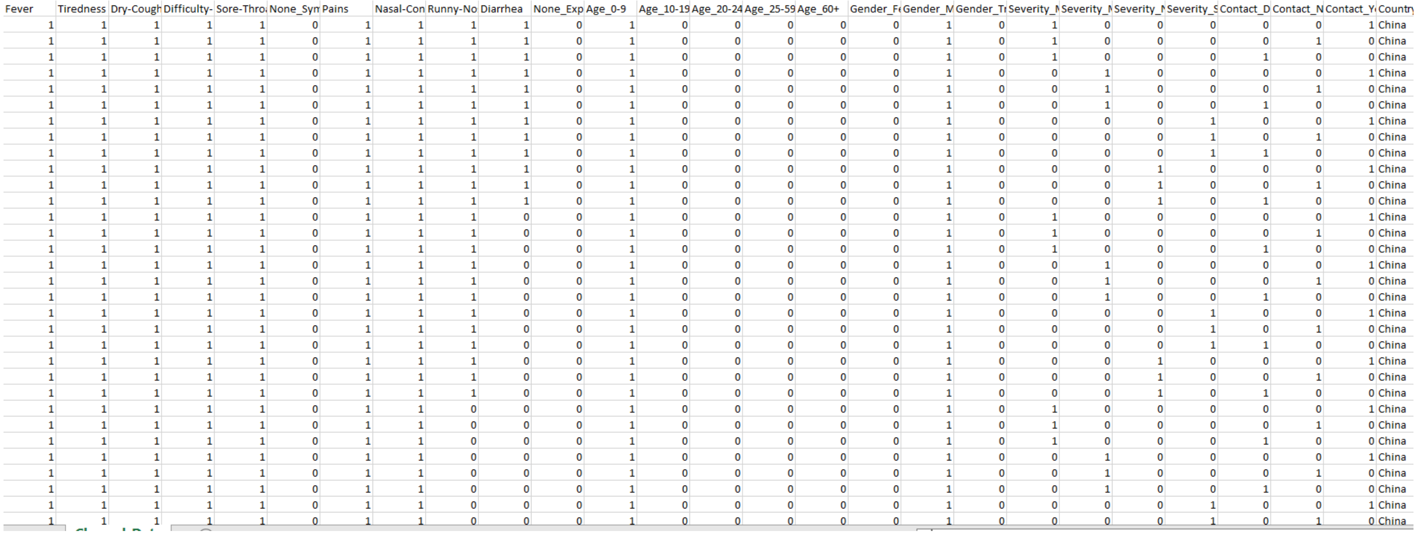
This global pandemic continues to strain medical systems around the world in a variety of ways, including increased demand for hospital beds and critical shortages of medical equipment, while many healthcare workers have become infected. As a result, the ability to make immediate clinical decisions and use medical resources effectively is critical. The most validated COVID-19 diagnosis test, reverse transcriptase polymerase chain reaction (RT-PCR), has long been in short supply in developing countries. This contributes to higher infection rates and causes critical preventive measures to be delayed.

Effective screening allows for the rapid and accurate diagnosis of COVID-19, reducing the burden on healthcare systems. In the hope of assisting medical staff worldwide in triaging patients, prediction models that combine several features to estimate the risk of infection have been developed. Computer tomography (CT) scans, clinical symptoms, laboratory tests, and an integration of these features are used in these models. However, because most previous models relied on data from hospitalised patients, they are ineffective for screening for SARS-CoV-2 in the general population.

In this project, we will use symptoms to predict the severity of the covid-19 disease. We have a data set of various country records from which our model will predict how drastic the illness will be, which ailments contribute to greater severity, and which are the most common.

3 Data sets

1. Cleaned.csv : - it's from kaggle it contained the covid symptoms with marking 1 means present and 0 means absent of various countries
2. data set URL : - https://www.kaggle.com/iamhungundji/covid19-symptoms-checker
3. Number of features: 27 4.used features: source=17 ,target=1 we eliminated some of features which are not required. used features are mentioned below.
source=$[Fever, Tiredness, Dry-Cough, Difficulty-in-Breathing, Sore-Throat, Pains, Nasal-Congestion, Runny-Nose, Diarrhea, Age_0-9, Age_10-19, Age_20-24, Age_25-59, Age_60+, Gender_Male, Gender_Female, Gender_Transgender]$
$target = severity - none$



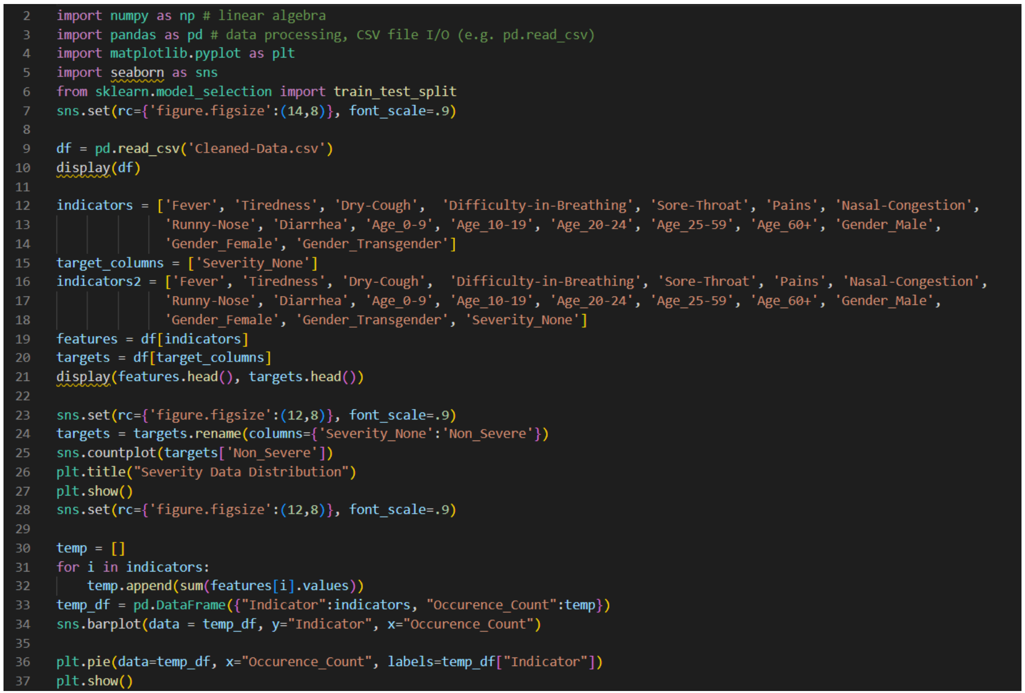
Fever	Tiredness	Dry-Cough	Difficulty-	Sore-Thro	None_Sym	Pains	Nasal-Con	Runny-No	Diarrhea	None_Exp	Age_0-9	Age_10-19	Age_20-24	Age_25-59	Age_60+	Gender_F	Gender_M	Gender_Tr	Severity_M	Severity_M	Severity_N	Severity_S	Contact_D	Contact_N	Contact_Y	Country
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	1	0	0	0	1	0	0	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	China
1	1	1	1	1	0	1	1	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	1	0	0	0	1	0	0	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	1	0	China
1	1	1	1	1	0	1	1	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	1	0	0	China
1	1	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	China
1	1	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	China
1	1	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	1	0	0	China
1	1	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	China
1	1	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	1	0	China
1	1	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	China
1	1	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	China
1	1	1	1	1	0	1	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	China

fig:-data set

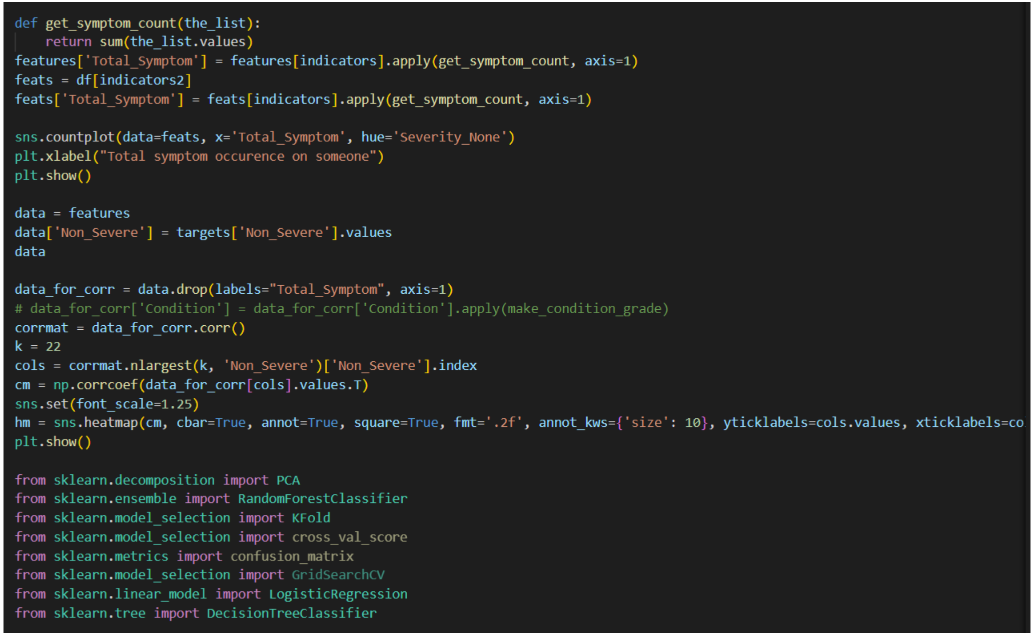
4 machine learning models used

1.Decision tree classifier
2.Random forest classifier
3.Logistic Regression
4. principal component Analysis

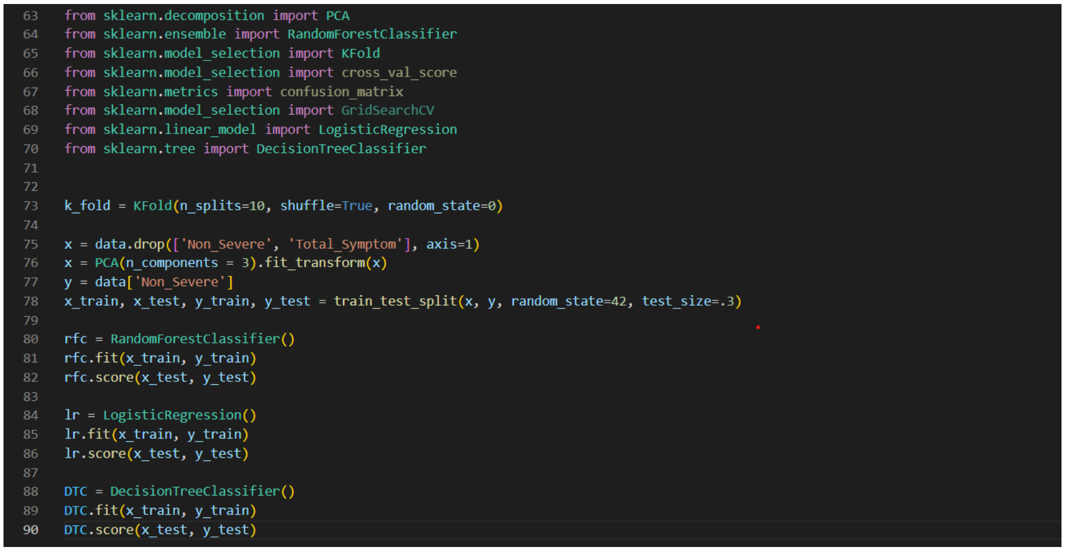
5 code



```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
sns.set(rc={'figure.figsize':(14,8)}, font_scale=.9)

df = pd.read_csv('Cleaned-Data.csv')
display(df)

indicators = ['Fever', 'Tiredness', 'Dry-Cough',  'Difficulty-in-Breathing', 'Sore-Throat', 'Pains', 'Nasal-Congestion',
              'Runny-Nose', 'Diarrhea', 'Age_0-9', 'Age_10-19', 'Age_20-24', 'Age_25-59', 'Age_60+', 'Gender_Male',
              'Gender_Female', 'Gender_Transgender']
target_columns = ['Severity_None']
indicators2 = ['Fever', 'Tiredness', 'Dry-Cough',  'Difficulty-in-Breathing', 'Sore-Throat', 'Pains', 'Nasal-Congestion',
              'Runny-Nose', 'Diarrhea', 'Age_0-9', 'Age_10-19', 'Age_20-24', 'Age_25-59', 'Age_60+', 'Gender_Male',
              'Gender_Female', 'Gender_Transgender', 'Severity_None']
features = df[indicators]
targets = df[target_columns]
display(features.head(), targets.head())

sns.set(rc={'figure.figsize':(12,8)}, font_scale=.9)
targets = targets.rename(columns={'Severity_None':'Non_Severe'})
sns.countplot(targets['Non_Severe'])
plt.title("Severity Data Distribution")
plt.show()
sns.set(rc={'figure.figsize':(12,8)}, font_scale=.9)

temp = []
for i in indicators:
    temp.append(sum(features[i].values))
temp_df = pd.DataFrame({"Indicator":indicators, "Occurence_Count":temp})
sns.barplot(data = temp_df, y="Indicator", x="Occurence_Count")

plt.pie(data=temp_df, x="Occurence_Count", labels=temp_df["Indicator"])
plt.show()
```

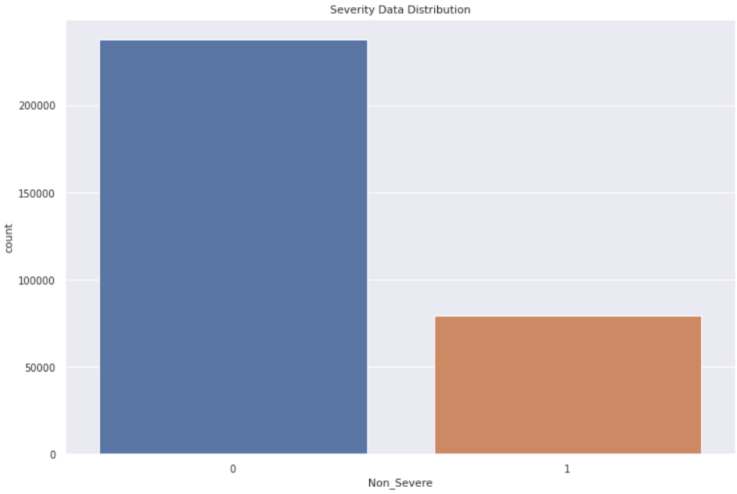


```
def get_symptom_count(the_list):
    return sum(the_list.values)
features['Total_Symptom'] = features[indicators].apply(get_symptom_count, axis=1)
feats = df[indicators2]
feats['Total_Symptom'] = feats[indicators].apply(get_symptom_count, axis=1)

sns.countplot(data=feats, x='Total_Symptom', hue='Severity_None')
plt.xlabel("Total symptom occurence on someone")
plt.show()

data = features
data['Non_Severe'] = targets['Non_Severe'].values
data

data_for_corr = data.drop(labels="Total_Symptom", axis=1)
# data_for_corr['Condition'] = data_for_corr['Condition'].apply(make_condition_grade)
corrmat = data_for_corr.corr()
k = 22
cols = corrmat.nlargest(k, 'Non_Severe')['Non_Severe'].index
cm = np.corrcoef(data_for_corr[cols].values.T)
sns.set(font_scale=1.25)
hm = sns.heatmap(cm, cbar=True, annot=True, square=True, fmt='.2f', annot_kws={'size': 10}, yticklabels=cols.values, xticklabels=co
plt.show()

from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
```

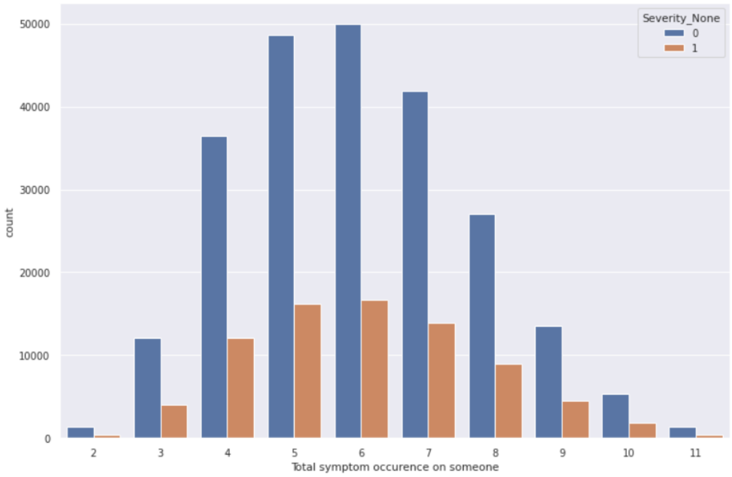


```
from sklearn.decomposition import PCA
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.metrics import confusion_matrix
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier


k_fold = KFold(n_splits=10, shuffle=True, random_state=0)

x = data.drop(['Non_Severe', 'Total_Symptom'], axis=1)
x = PCA(n_components = 3).fit_transform(x)
y = data['Non_Severe']
x_train, x_test, y_train, y_test = train_test_split(x, y, random_state=42, test_size=.3)

rfc = RandomForestClassifier()
rfc.fit(x_train, y_train)
rfc.score(x_test, y_test)

lr = LogisticRegression()
lr.fit(x_train, y_train)
lr.score(x_test, y_test)

DTC = DecisionTreeClassifier()
DTC.fit(x_train, y_train)
DTC.score(x_test, y_test)
```

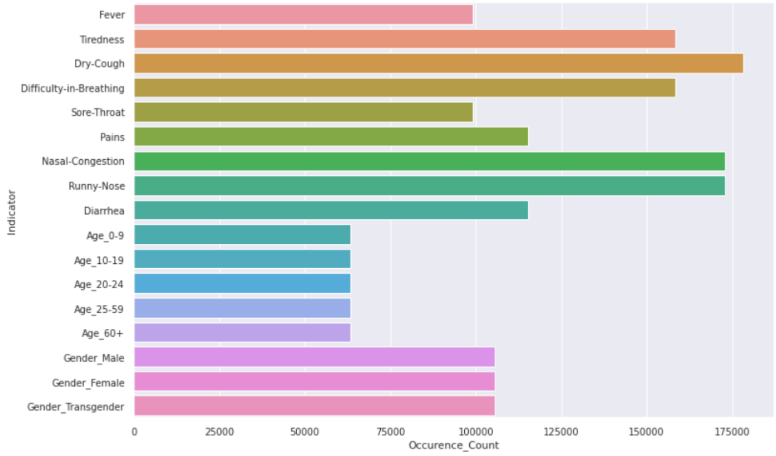
6 output description



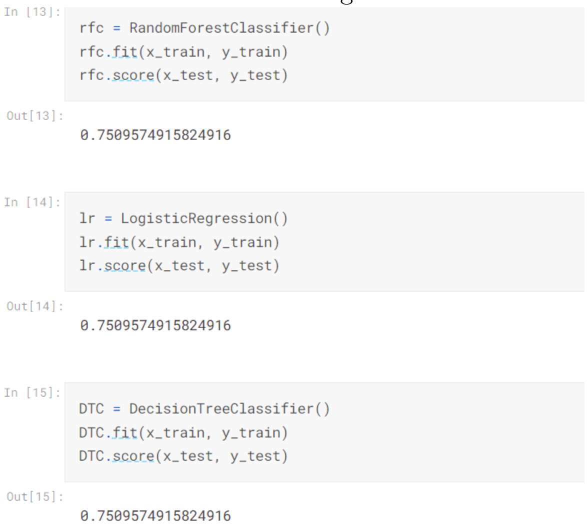
In above chart we are showing the count of non-severe patients with denim blue color and count of severe patients with brown color.



In above chart we are showing based on number of symptoms the count distribution of severe and non-severe patients.



In above chart we are showing the contribution of each symptom to cause disease and lead's to severity.



```
In [13]:
rfc = RandomForestClassifier()
rfc.fit(x_train, y_train)
rfc.score(x_test, y_test)

Out[13]:
0.7509574915824916
```

```
In [14]:
lr = LogisticRegression()
lr.fit(x_train, y_train)
lr.score(x_test, y_test)

Out[14]:
0.7509574915824916
```

```
In [15]:
DTC = DecisionTreeClassifier()
DTC.fit(x_train, y_train)
DTC.score(x_test, y_test)

Out[15]:
0.7509574915824916
```

In above figure we are showing various machine learning models used and their accuracy score.

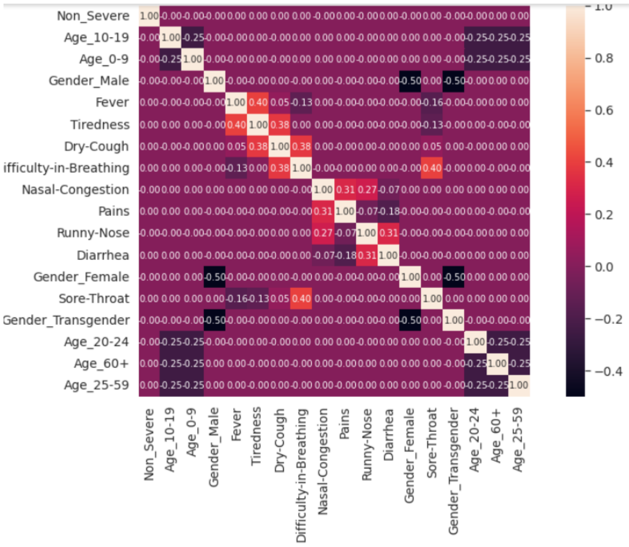
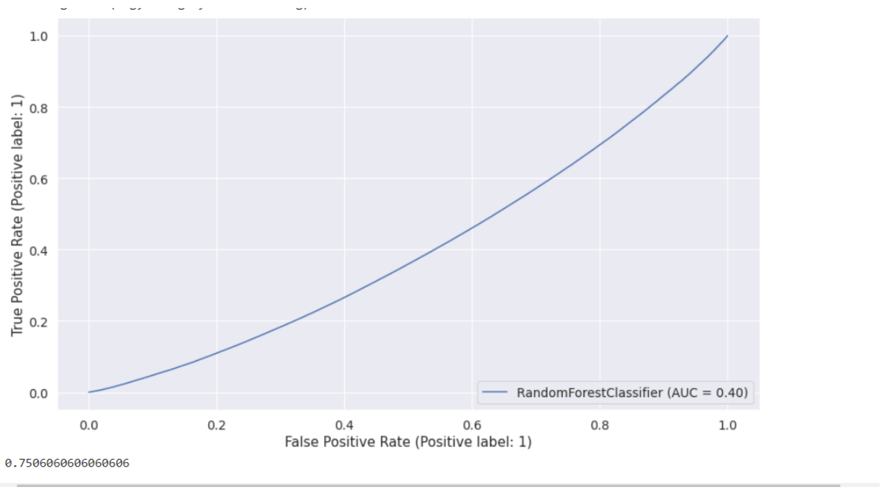
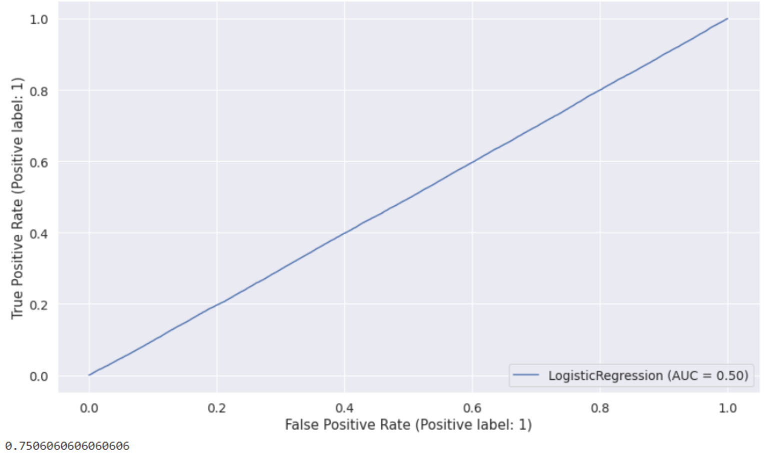


fig:- confusion matrix



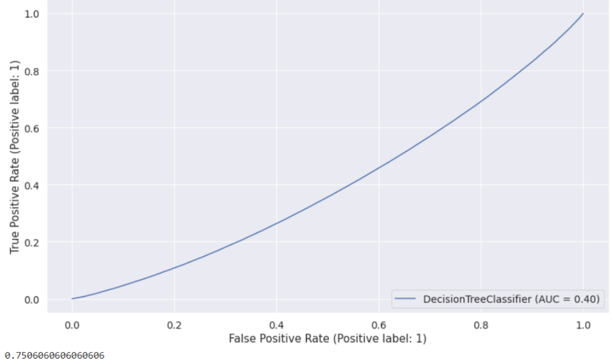
```
0.7506060606060606
```

fig:- ROC Graph of Random Forest



```
0.7506060606060606
```

fig:- ROC Graph of Logistic Regression



```
0.7506060606060606
```

fig:- ROC Graph of Decision Classifier

7 conclusion

we used various machine learning models like linear regression, logistic regression, decision tree, PCA(principal component Analysis) etc. and we represented different accuracy levels and various charts which are related to severity of covid-19. so it conclude various aspects related severity measure.