Galena Wagdy Zareef 20399124

CSAI 801 PROJECT: COVID-19 OUTCOME PREDICTION

The data is available from 22 Jan, 2020. Data is in “data.csv”. The dataset contains 14 major variables that will be having an impact on whether someone has recovered or not, the description of each variable are as follows,

1. Country: where the person resides
2. Location: which part in the Country
3. Age: Classification of the age group for each person, based on WHO Age Group Standard
4. Gender: Male or Female
5. Visited_Wuhan: whether the person has visited Wuhan, China or not
6. From_Wuhan: whether the person is from Wuhan, China or not
7. Symptoms: there are six families of symptoms that are coded in six fields.
8. Time_before_symptoms_appear:
9. Result: death (1) or recovered (0)

	location	country	gender	age	vis_wuhan	from_wuhan	symptom1	symptom2	symptom3	symptom4	symptom5	symptom6	diff_sym_hos	result
0	104	8	1	66.0	1	0	14	31	19	12	3	1	8	1
1	101	8	0	56.0	0	1	14	31	19	12	3	1	0	0
2	137	8	1	46.0	0	1	14	31	19	12	3	1	13	0
3	116	8	0	60.0	1	0	14	31	19	12	3	1	0	0
4	116	8	1	58.0	0	0	14	31	19	12	3	1	0	0

Describing the data:

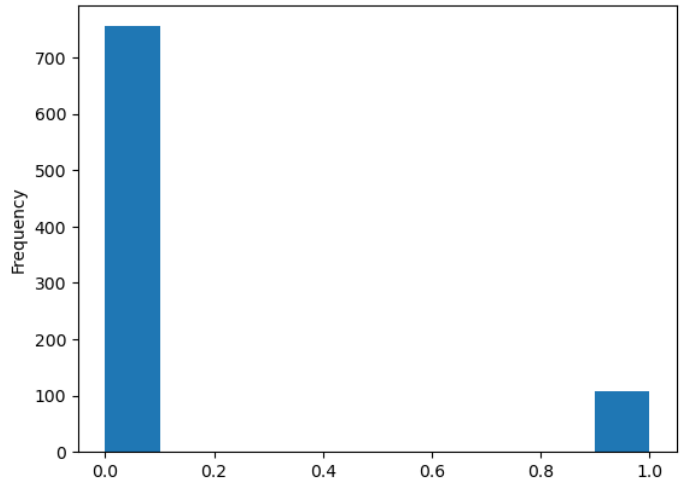
	location	country	gender	age	vis_wuhan	from_wuhan	symptom1	symptom2	symptom3	symptom4	symptom5	symptom6	diff_sym_hos	result
count	863.000000	863.000000	863.000000	863.000000	863.000000	863.000000	863.00000	863.000000	863.000000	863.000000	863.000000	863.000000	863.000000	863.000000
mean	76.645423	16.995365	0.849363	49.400000	0.181924	0.107764	12.13905	28.002317	18.298957	11.840093	2.993048	0.998841	0.995365	0.125145
std	39.200264	7.809951	0.726062	15.079203	0.386005	0.310261	3.99787	7.473231	2.864064	1.183771	0.127251	0.034040	2.358767	0.331075
min	0.000000	0.000000	0.000000	2.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	-5.000000	0.000000
25%	45.000000	11.000000	0.000000	40.000000	0.000000	0.000000	14.00000	31.000000	19.000000	12.000000	3.000000	1.000000	0.000000	0.000000
50%	87.000000	18.000000	1.000000	49.400000	0.000000	0.000000	14.00000	31.000000	19.000000	12.000000	3.000000	1.000000	0.000000	0.000000
75%	110.000000	24.000000	1.000000	57.000000	0.000000	0.000000	14.00000	31.000000	19.000000	12.000000	3.000000	1.000000	1.000000	0.000000
max	138.000000	33.000000	2.000000	96.000000	1.000000	1.000000	24.00000	31.000000	19.000000	12.000000	3.000000	1.000000	15.000000	1.000000

Information of the data:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 863 entries, 0 to 862
Data columns (total 14 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   location      863 non-null    int64
 1   country       863 non-null    int64
 2   gender        863 non-null    int64
 3   age           863 non-null    float64
 4   vis_wuhan     863 non-null    int64
 5   from_wuhan    863 non-null    int64
 6   symptom1      863 non-null    int64
 7   symptom2      863 non-null    int64
 8   symptom3      863 non-null    int64
 9   symptom4      863 non-null    int64
 10  symptom5      863 non-null    int64
 11  symptom6      863 non-null    int64
 12  diff_sym_hos  863 non-null    int64
 13  result        863 non-null    int64
dtypes: float64(1), int64(13)
memory usage: 101.1 KB
```

Shape of the features: (863, 12)
Shape of the target column: (863,)

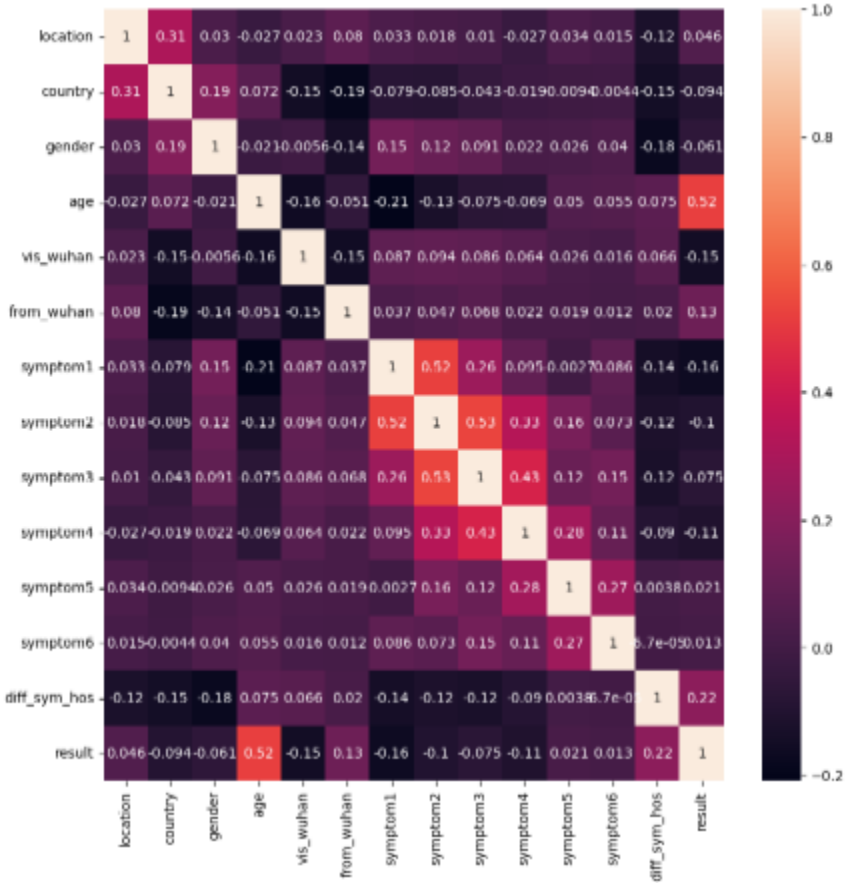
Count 0 755
Count 1 108



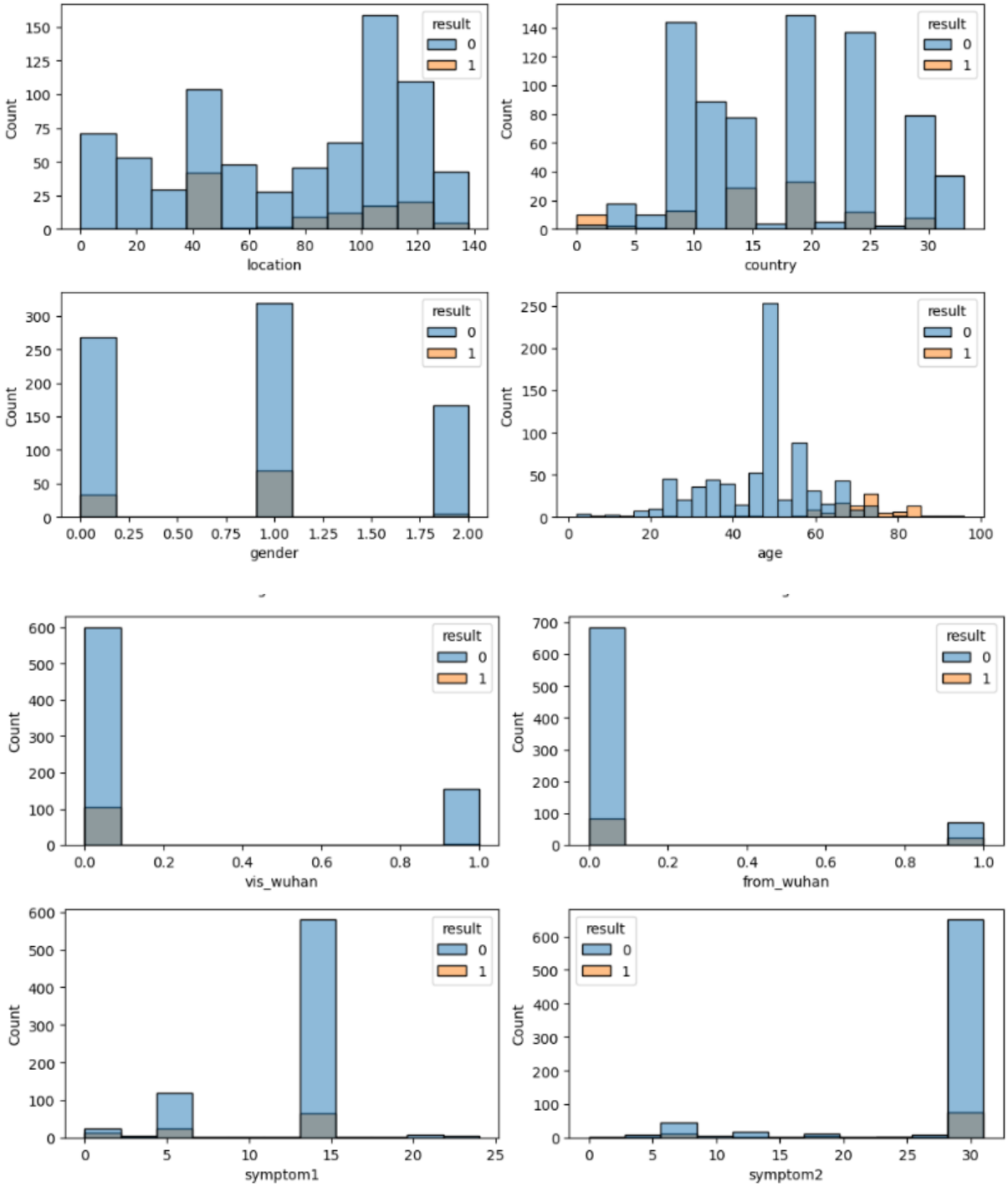
Correlation with target column:

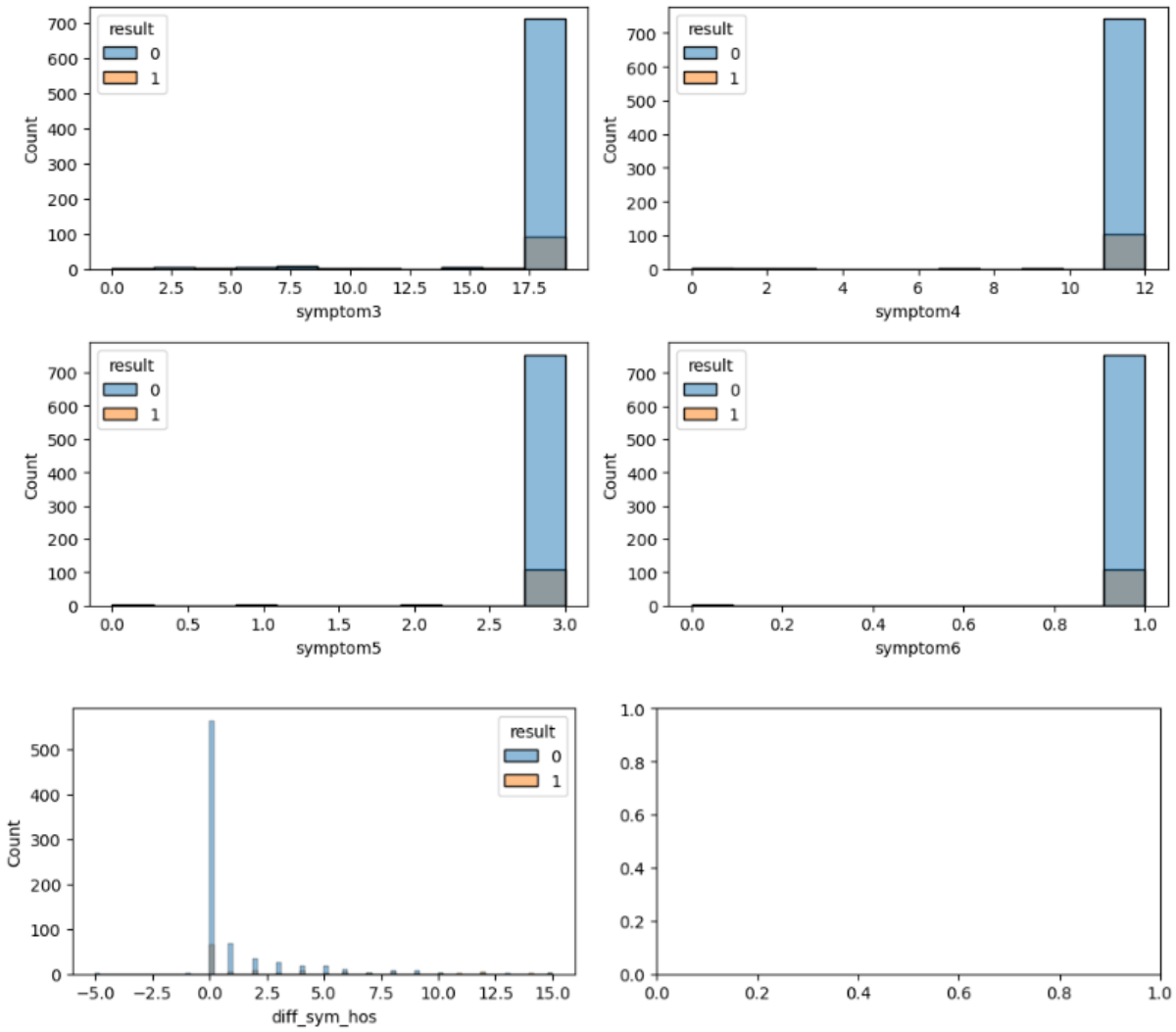
```
location         0.046240
country         -0.094443
gender          -0.061441
age              0.515127
vis_wuhan       -0.151122
from_wuhan       0.128314
symptom1        -0.163039
symptom2        -0.103270
symptom3        -0.074982
symptom4        -0.108723
symptom5         0.020676
symptom6         0.012882
diff_sym_hos     0.219116
result           1.000000
Name: result, dtype: float64
```

Correlation with each column:



Visualizing the data with respect of the target column:





Machine learning algorithms:

1) KNN:

a) Classification report and roc for the training:

	precision	recall	f1-score	support
0	0.96	0.96	0.96	24
1	0.50	0.50	0.50	2
accuracy			0.92	26
macro avg	0.73	0.73	0.73	26
weighted avg	0.92	0.92	0.92	26

0.7291666666666666

b) Classification report and roc for the validation:

```
Fitting 10 folds for each of 30 candidates, totalling 300 fits
              precision    recall  f1-score   support

           0       0.98      0.96      0.97       204
           1       0.75      0.83      0.79        29

    accuracy                           0.94       233
   macro avg       0.86      0.89      0.88       233
weighted avg       0.95      0.94      0.95       233

0.894185260311021
```

c) Classification report and roc for the testing:

```
              precision    recall  f1-score   support

           0       0.96      0.96      0.96        24
           1       0.50      0.50      0.50         2

    accuracy                           0.92        26
   macro avg       0.73      0.73      0.73        26
weighted avg       0.92      0.92      0.92        26

0.7291666666666666
```

2) Logistic Regression

a) Classification report and roc for the training:

```
              precision    recall  f1-score   support

           0       0.96      0.96      0.96        24
           1       0.50      0.50      0.50         2

    accuracy                           0.92        26
   macro avg       0.73      0.73      0.73        26
weighted avg       0.92      0.92      0.92        26

0.7291666666666666
```

b) Classification report and roc for the validation:

	precision	recall	f1-score	support
0	0.96	0.97	0.96	204
1	0.74	0.69	0.71	29
accuracy			0.93	233
macro avg	0.85	0.83	0.84	233
weighted avg	0.93	0.93	0.93	233

0.8276707234617986

c) Classification report and roc for the testing:

	precision	recall	f1-score	support
0	0.96	0.96	0.96	24
1	0.50	0.50	0.50	2
accuracy			0.92	26
macro avg	0.73	0.73	0.73	26
weighted avg	0.92	0.92	0.92	26

0.7291666666666666

3) Naive bayes:

a) Classification report and roc for the training:

	precision	recall	f1-score	support
0	1.00	0.29	0.45	24
1	0.11	1.00	0.19	2
accuracy			0.35	26
macro avg	0.55	0.65	0.32	26
weighted avg	0.93	0.35	0.43	26

0.6458333333333333

b) Classification report and roc for the validation:

	precision	recall	f1-score	support
0	0.96	0.97	0.96	204
1	0.74	0.69	0.71	29
accuracy			0.93	233
macro avg	0.85	0.83	0.84	233
weighted avg	0.93	0.93	0.93	233

0.8276707234617986

c) Classification report and roc for the testing:

	precision	recall	f1-score	support
0	0.96	0.96	0.96	24
1	0.50	0.50	0.50	2
accuracy			0.92	26
macro avg	0.73	0.73	0.73	26
weighted avg	0.92	0.92	0.92	26

0.7291666666666666

4) Decision tree:

a) Classification report and roc for the training:

	precision	recall	f1-score	support
0	0.96	0.96	0.96	24
1	0.50	0.50	0.50	2
accuracy			0.92	26
macro avg	0.73	0.73	0.73	26
weighted avg	0.92	0.92	0.92	26

0.7291666666666666

b) Classification report and roc for the validation:

```
Fitting 4 folds for each of 50 candidates, totalling 200 fits
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       204
           1       0.78      0.72      0.75        29

    accuracy                           0.94       233
   macro avg       0.87      0.85      0.86       233
weighted avg       0.94      0.94      0.94       233

0.8473630831643002
```

c) Classification report and roc for the testing:

```
              precision    recall  f1-score   support

           0       0.96      0.96      0.96        24
           1       0.50      0.50      0.50         2

    accuracy                           0.92        26
   macro avg       0.73      0.73      0.73        26
weighted avg       0.92      0.92      0.92        26

0.7291666666666666
```

5) Support vector machine:

a) Classification report and roc for the training:

```
              precision    recall  f1-score   support

           0       0.96      1.00      0.98        24
           1       1.00      0.50      0.67         2

    accuracy                           0.96        26
   macro avg       0.98      0.75      0.82        26
weighted avg       0.96      0.96      0.96        26

0.75
```

b) Classification report and roc for the validation:

```
Fitting 4 folds for each of 50 candidates, totalling 200 fits
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       204
           1       0.78      0.72      0.75        29

    accuracy                           0.94       233
   macro avg       0.87      0.85      0.86       233
weighted avg       0.94      0.94      0.94       233

0.8473630831643002
```

c) Classification report and roc for the testing:

```
              precision    recall  f1-score   support

           0       0.96      0.96      0.96        24
           1       0.50      0.50      0.50         2

    accuracy                           0.92        26
   macro avg       0.73      0.73      0.73        26
weighted avg       0.92      0.92      0.92        26

0.7291666666666666
```