Supervised Machine learning to predict COVID-19 Lab result

DATA602 – Introduction to Machine learning

Presented by:
Sharath Srinivas

Why Covid-19 is important?

- With the worst coronavirus (COVID-19) outbreak still ongoing, the necessity for action to prevent ourselves from contracting and transmitting the virus has become an essential aspect of our daily life.
- However, some segments of the population are still not taking the threat as seriously as they should. Whether it's due to denial, living in an echo chamber of misinformation, or simply being unaware of the hazards, significant swaths of the population aren't taking efforts to prevent the spread.

Why testing is important

- Effective testing is critical in slowing the spread of the virus by identifying persons infected and allowing treatment or isolation.
- Testing is also necessary to learn more about how the virus spreads and how widespread it is in a particular community.
- But, for the same reasons, health officials must ensure that their tests are as efficient as possible.
- In other words, with continued shortages, the imperative of ensuring that those most in need are tested means that not everyone should be checked.

Introduction

- In this project supervised machine learning techniques are used to develop predictive models for the COVID-19 infection, using an epidemiology labeled dataset for positive and negative COVID-19 cases in Mexico, with supervised learning algorithms such as decision tree, logistic regression, Random Forest and naive Bayes, support vector machine, and k-nearest neighbors

Problem Statement

- The global pandemic 2019-nCoV or COVID-19 is caused by the latest pandemic outbreak of the novel Severe Acute Respiratory Syndrome-Coronavirus two (SARS-CoV-2). The virus was first discovered in bats in late December 2019 in the Chinese province of Wuhan, with evidence suggesting it spread to people via intermediary hosts such as raccoons, dogs etc.
- Significant COVID-19 symptoms include fever (98%) cough (76%) and diarrhea (3%), which are generally more severe in older persons with chronic conditions , and many patients have reported shortness of breath, which in many cases appears to be a sign of the flu. , 2019 Since its discovery in late 2019, Covid has spread rapidly throughout the world.

Objective

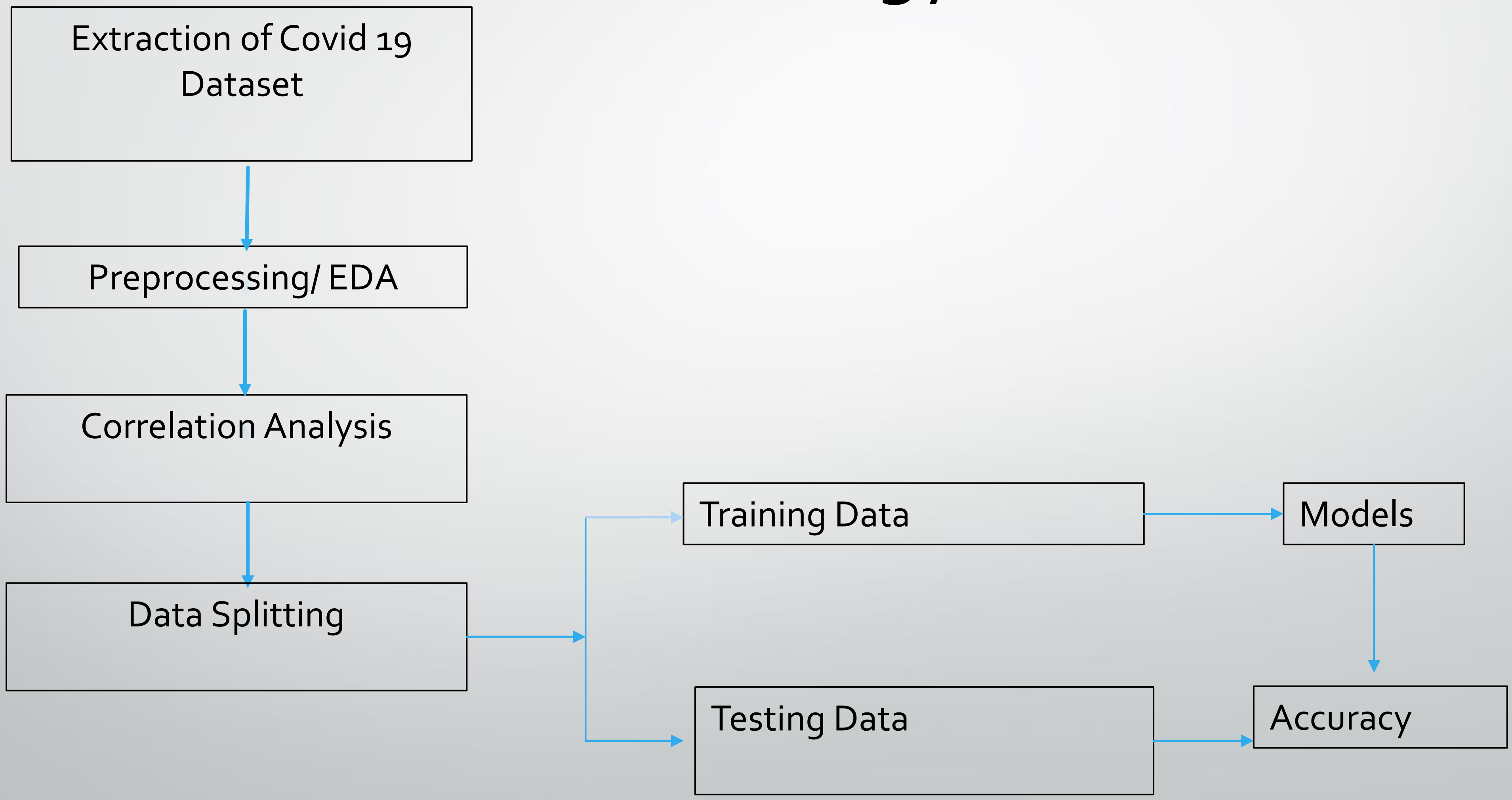
- To predict covid 19 results based on the symptoms of the patient.
- In this project we will be using the Lab result of covid antigen test as our target variables using symptoms features.

Data Description

The data set used was provided by Mexico's Epidemiological Surveillance System for Viral Respiratory Diseases and includes information on all COVID19-related cases, including positive and negative tested individuals. This data set is available for downloading.

Sl no	Columns	Description
1	UPDATE DATE	The database is fed daily, this variable allows identifying the date of the last update.
2	REGISTRATION ID	Case identifier number
3	ORIGIN	Sentinel surveillance is carried out through the system of health units monitoring respiratory diseases (USMER). The USMER include medical units of the first, second or third level of care, and third level units also participate as USMERs, which due to their characteristics contribute to broadening the epidemiological information panorama, including those with a specialty in pulmonology, infectology or pediatrics. .
4	SECTOR	Identifies the type of institution of the National Health System that provided the care
5	ENTITY UM	Identifies the entity where the medical unit that provided the care is located.
6	SEX	SEX Identifies the sex of the patient.
7	NAC ENTITY	Identifies the patient's birth entity.
8	RES ENTITIY	Identifies the patient's residence entity.
9	MUNICIPALITY RES	Identifies the municipality of residence of the patient.
10	PATIENT TYPE	Identifies the type of care the patient received in the unit. It is called an outpatient if you returned home or it is called an inpatient if you were admitted to the hospital.
11	ENTRY DATE	Identifies the date of admission of the patient to the care unit.
12	SYMPTOMS DATE	Identifies the date on which the patient's symptoms began.
13	DATE DEF	Identifies the date the patient died.
14	INTUBATED	Identifies if the patient required intubation.
15	PNEUMONIA	Identifies if the patient has been diagnosed with pneumonia.
16	AGE	Identifies the age of the patient.
17	NATIONALITY	Identifies if the patient is Mexican or foreign.
18	PREGNANCY	Identifies if the patient is pregnant
19	INDIGENOUS LANGUAGE SPEAKS	Identifies if the patient speaks an indigenous language
20	DIABETES	Identifies if the patient has a diagnosis of diabetes.
21	COPD	Identifies if the patient has a diagnosis of COPD.
22	ASTHMA	Identifies if the patient has a diagnosis of asthma.
23	INMUSUPR	Identifies if the patient is immunosuppressed.
24	HYPERTENSION	Identifies if the patient has a diagnosis of hypertension.
25	OTHER COM	Identifies if the patient has a diagnosis of other diseases.
26	CARDIOVASCULAR	Identifies if the patient has a diagnosis of cardiovascular disease.
27	OBESITY	Identifies if the patient has a diagnosis of obesity.
28	CHRONIC KIDNEY	Identifies if the patient has a diagnosis of chronic renal failure.
29	SMOKING	Identifies if the patient has a smoking habit.
30	OTHER CASE	Identifies if the patient had contact with any other case diagnosed with SARS CoV-2
31	RESULT	Identifies if the patient is a case of COVID-19
32	MIGRANT	Identifies if the patient is a migrant person
33	COUNTRY_NATIONALITY	Identifies the nationality of the patient.
34	COUNTRY_ORIGEN	Identifies the country from which the patient left for Mexico.
35	ICU	Identifies if the patient required admission to an Intensive Care Unit.

Methodology



Extraction of Covid 19 Dataset

Preprocessing/ EDA

Correlation Analysis

Data Splitting

Training Data

Testing Data

Models

Accuracy

After importing the data frames dropped the all DATE columns along with nationality and information related to province/state because it will not help for our analysis

The columns "ENTITY RES" and " ENTITY UM" refer to the state of residence and treatment place respectively. These are coded by numbers so that they can be replaced by their respective names.

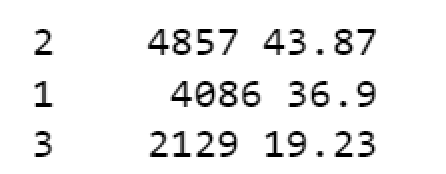
Symptoms features selected for analysis:

- PNEUMONIA
- INDIGENOUS_LANG
- DIABETES
- COPD
- ASTHMA
- INMUSUPR
- HYPERTENSION
- OTHERCOM
- CARDIOVASCULAR
- OBESITY
- CHRONIC_KIDNEY
- SMOKING
- OTHERCASE

Figure 1

EDA Observation

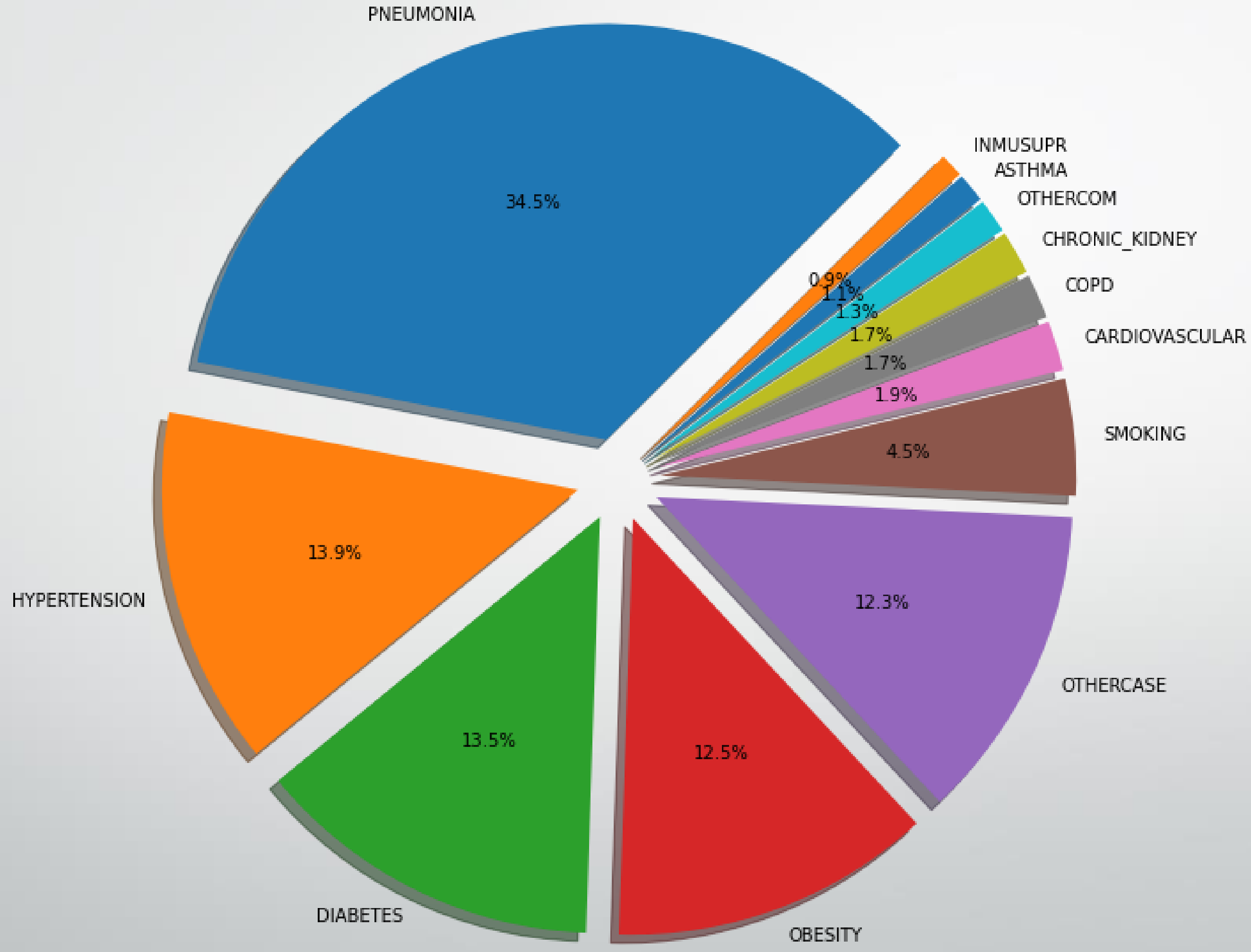
1. Only Pregnancy, Other case and ICU column has significant amount of Not applicable and Not specified values. In our dataset for symptoms columns, 2 = No, and 1 = Yes. For our analysis we removed all not applicable and not specified and ignored values
2. Result column had three values 2 is negative and 1 is positive and 3 is result pending. We will be dropping the those values which are pending results and convert 2 to 0

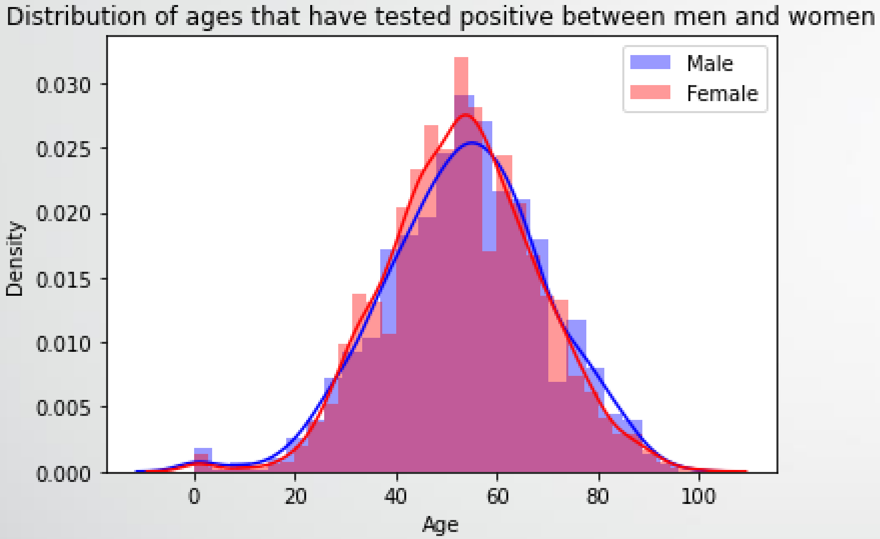


```
2    4857 43.87
1     4086 36.9
3    2129 19.23
```

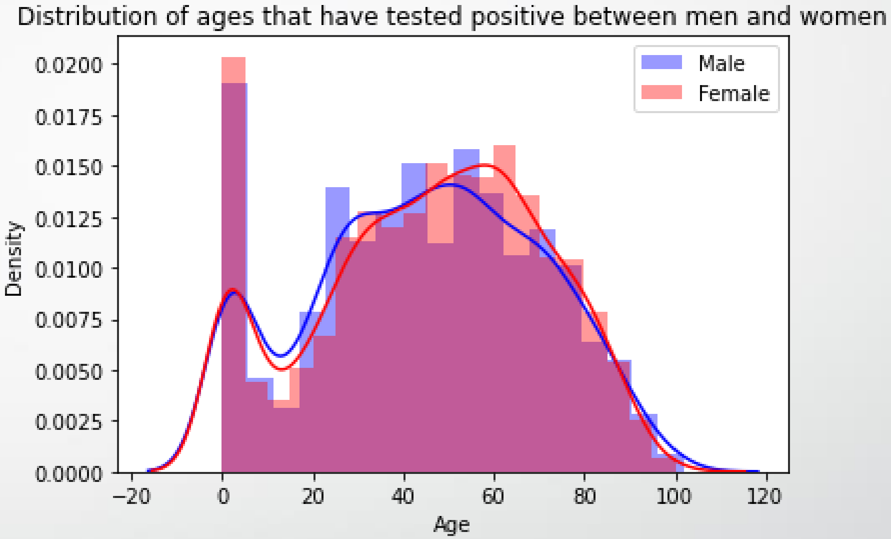
3. Cudad de Mexico state has highest number of positive and negative cases
4. Observed that patient male and female between age 40 to 60 are more likely to get Covid-19
5. top 4 symptoms are Pneumonia, Hypertension, Obesity and diabetes this indicates patients who have history of these decease must be carefully.

Percentage of Illness for positive cases



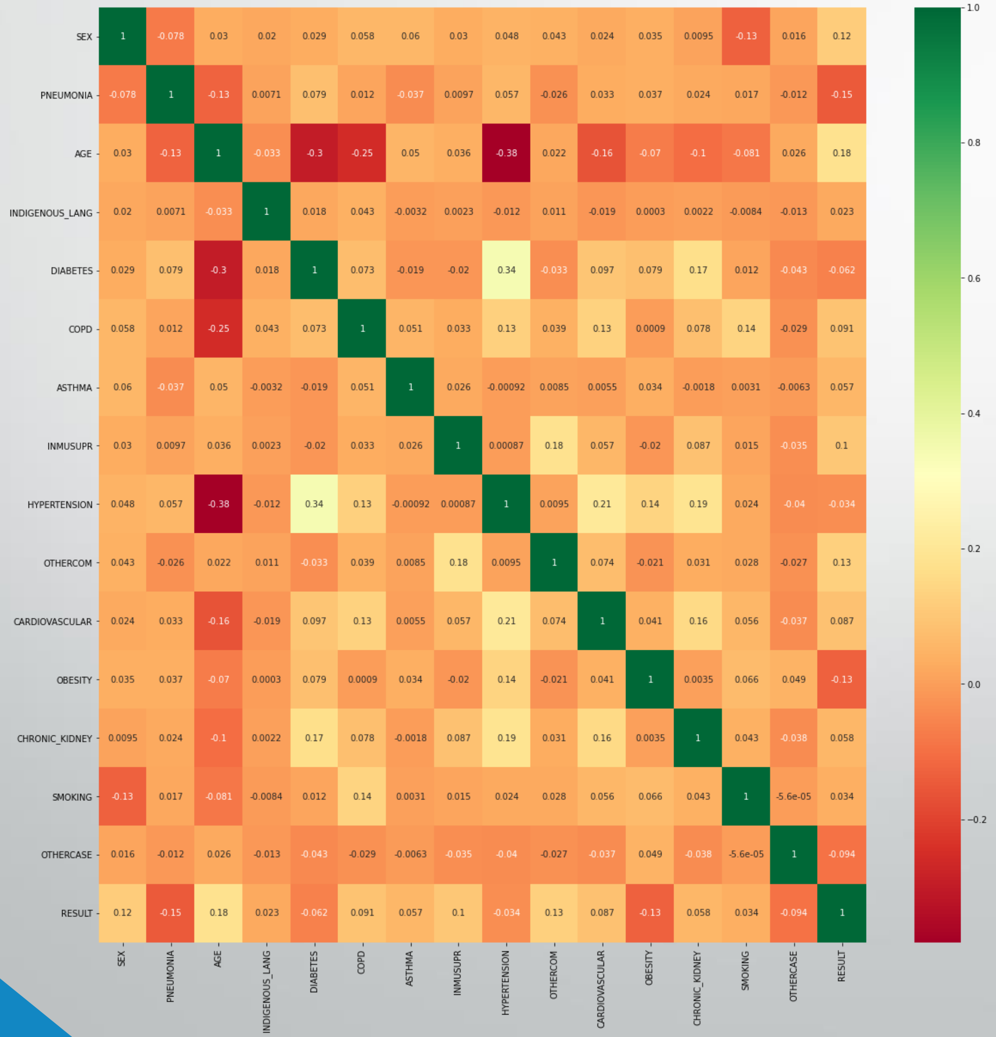


Distribution of ages that have tested positive between men and women



Distribution of ages that have tested positive between men and women

Correlation Matrix



	SEX	PNEUMONIA	AGE	INDIGENOUS_LANG	DIABETES	COPD	ASTHMA	INMUSUPR	HYPERTENSION	OTHERCOM	CARDIOVASCULAR	OBESITY	CHRONIC_KIDNEY	SMOKING	OTHERCASE	RESULT
SEX	1	-0.078	0.03	0.02	0.029	0.058	0.06	0.03	0.048	0.043	0.024	0.035	0.0095	-0.13	0.016	0.12
PNEUMONIA	-0.078	1	-0.13	0.0071	0.079	0.012	-0.037	0.0097	0.057	-0.026	0.033	0.037	0.024	0.017	-0.012	-0.15
AGE	0.03	-0.13	1	-0.033	-0.3	-0.25	0.05	0.036	-0.38	0.022	-0.16	-0.07	-0.1	-0.081	0.026	0.18
INDIGENOUS_LANG	0.02	0.0071	-0.033	1	0.018	0.043	-0.0032	0.0023	-0.012	0.011	-0.019	0.0003	0.0022	-0.0084	-0.013	0.023
DIABETES	0.029	0.079	-0.3	0.018	1	0.073	-0.019	-0.02	0.34	-0.033	0.097	0.079	0.17	0.012	-0.043	-0.062
COPD	0.058	0.012	-0.25	0.043	0.073	1	0.051	0.033	0.13	0.039	0.13	0.0009	0.078	0.14	-0.029	0.091
ASTHMA	0.06	-0.037	0.05	-0.0032	-0.019	0.051	1	0.026	-0.00092	0.0085	0.0055	0.034	-0.0018	0.0031	-0.0063	0.057
INMUSUPR	0.03	0.0097	0.036	0.0023	-0.02	0.033	0.026	1	0.00087	0.18	0.057	-0.02	0.087	0.015	-0.035	0.1
HYPERTENSION	0.048	0.057	-0.38	-0.012	0.34	0.13	-0.00092	0.00087	1	0.0095	0.21	0.14	0.19	0.024	-0.04	-0.034
OTHERCOM	0.043	-0.026	0.022	0.011	-0.033	0.039	0.0085	0.18	0.0095	1	0.074	-0.021	0.031	0.028	-0.027	0.13
CARDIOVASCULAR	0.024	0.033	-0.16	-0.019	0.097	0.13	0.0055	0.057	0.21	0.074	1	0.041	0.16	0.056	-0.037	0.087
OBESITY	0.035	0.037	-0.07	0.0003	0.079	0.0009	0.034	-0.02	0.14	-0.021	0.041	1	0.0035	0.066	0.049	-0.13
CHRONIC_KIDNEY	0.0095	0.024	-0.1	0.0022	0.17	0.078	-0.0018	0.087	0.19	0.031	0.16	0.0035	1	0.043	-0.038	0.058
SMOKING	-0.13	0.017	-0.081	-0.0084	0.012	0.14	0.0031	0.015	0.024	0.028	0.056	0.066	0.043	1	-5.6e-05	0.034
OTHERCASE	0.016	-0.012	0.026	-0.013	-0.043	-0.029	-0.0063	-0.035	-0.04	-0.027	-0.037	0.049	-0.038	-5.6e-05	1	-0.094
RESULT	0.12	-0.15	0.18	0.023	-0.062	0.091	0.057	0.1	-0.034	0.13	0.087	-0.13	0.058	0.034	-0.094	1

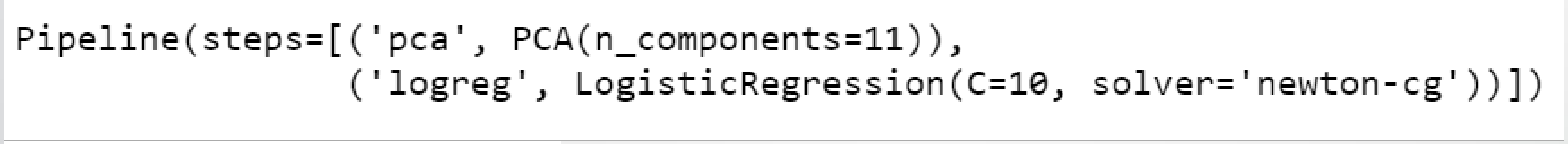
Figure 5

There are no strong positive correlation between the features but there are weak positive correlation between symptoms and result features.

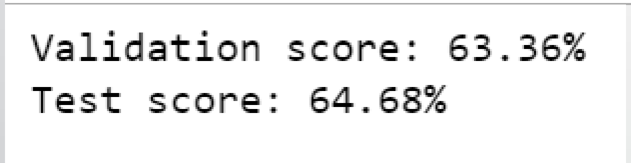
Model

Data split was done for 60% of train data and 40% of test data

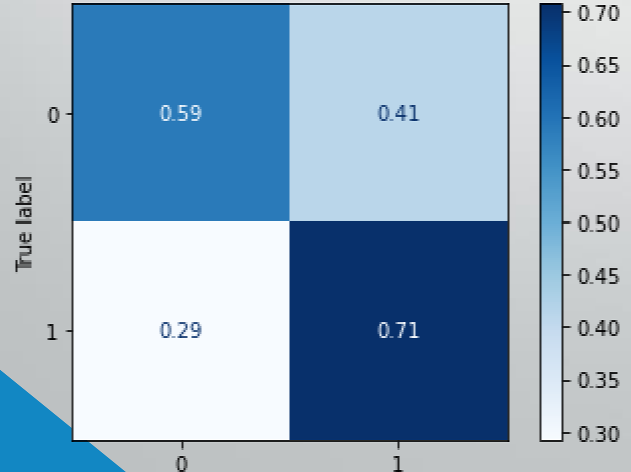
Logistic Regression : For Logistic Regression we used C and solver has a hyperparameter with PCA



```
Pipeline(steps=[('pca', PCA(n_components=11)),
                ('logreg', LogisticRegression(C=10, solver='newton-cg'))])
```



```
Validation score: 63.36%
Test score: 64.68%
```



Logistic Regression Confussion matrix on how the model can predict covid positive

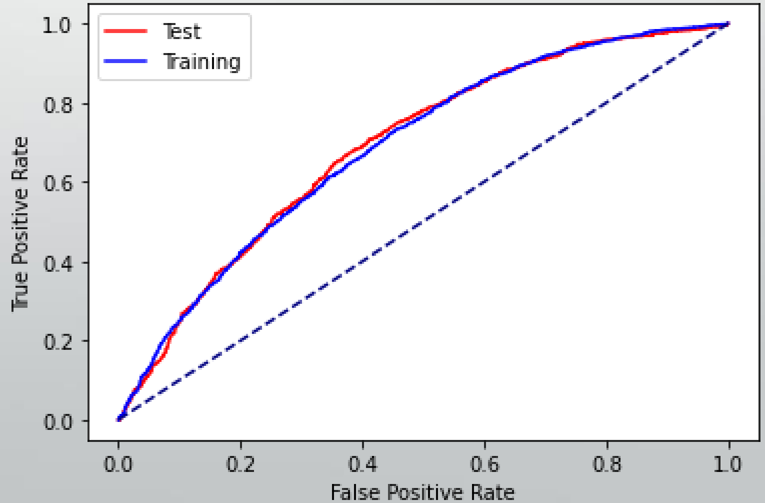
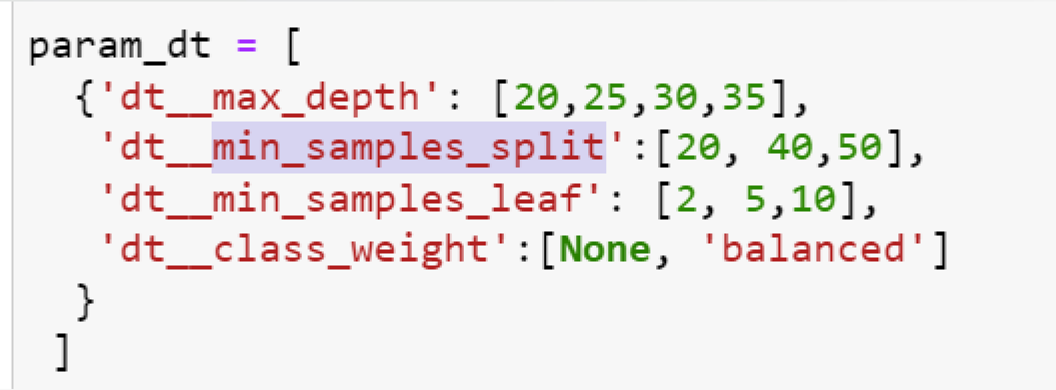
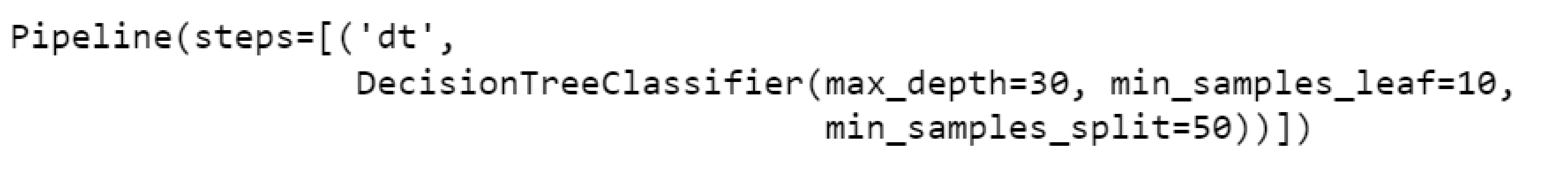


Figure 7

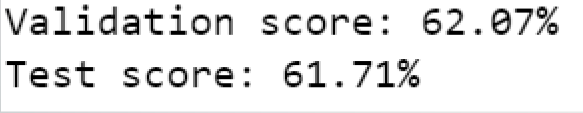
Decision Tree



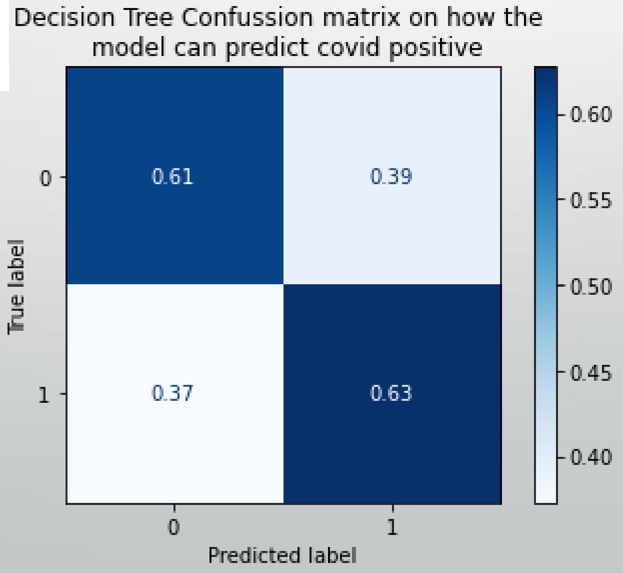
```
param_dt = [
  {'dt__max_depth': [20,25,30,35],
   'dt__min_samples_split':[20, 40,50],
   'dt__min_samples_leaf': [2, 5,10],
   'dt__class_weight':[None, 'balanced']
  }
 ]
```

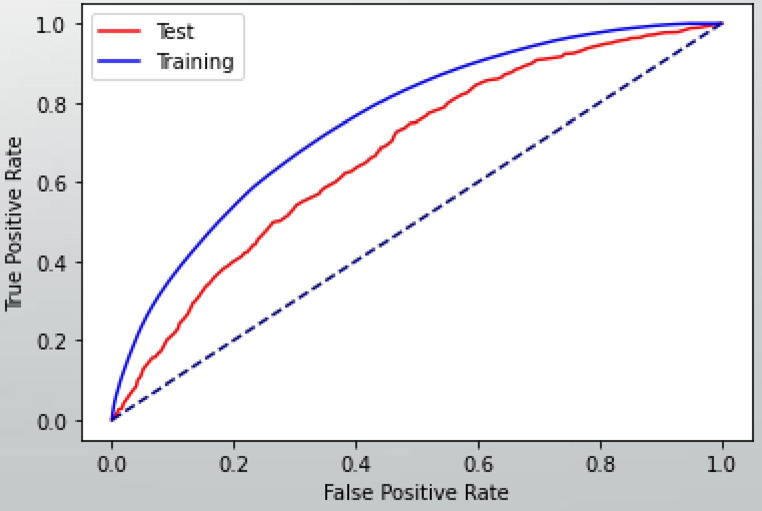


```
Pipeline(steps=[('dt',
                 DecisionTreeClassifier(max_depth=30, min_samples_leaf=10,
                                        min_samples_split=50))])
```

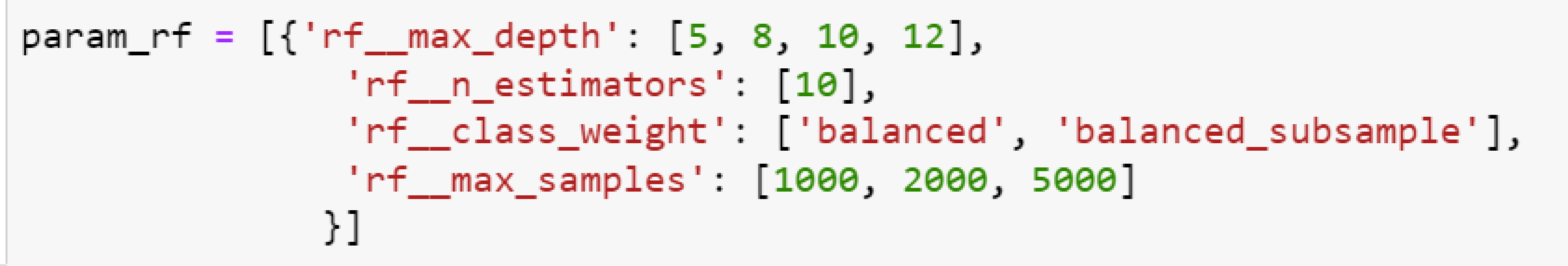


```
Validation score: 62.07%
Test score: 61.71%
```

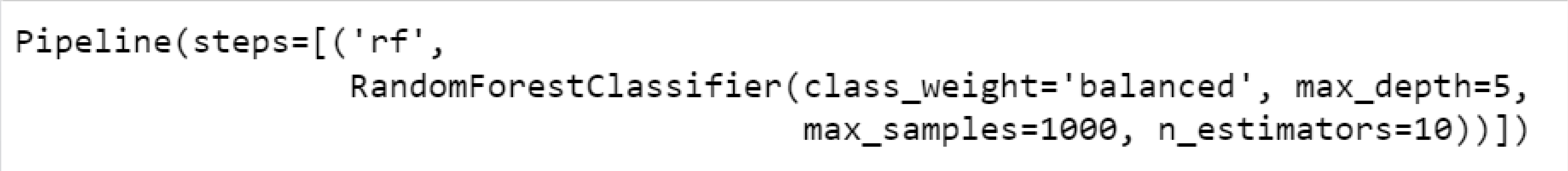




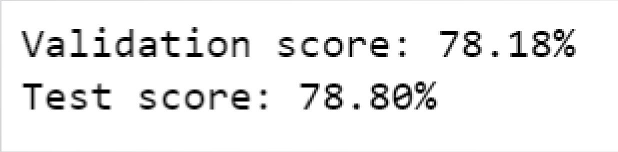
Random forest



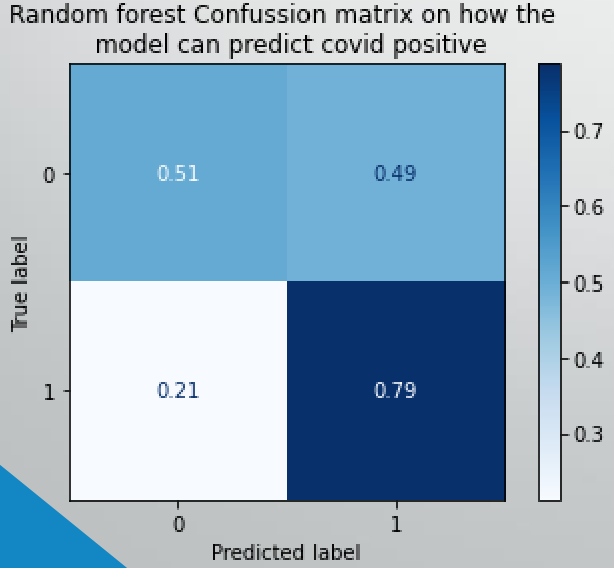
```
param_rf = [{'rf__max_depth': [5, 8, 10, 12],
               'rf__n_estimators': [10],
               'rf__class_weight': ['balanced', 'balanced_subsample'],
               'rf__max_samples': [1000, 2000, 5000]
              }]
```



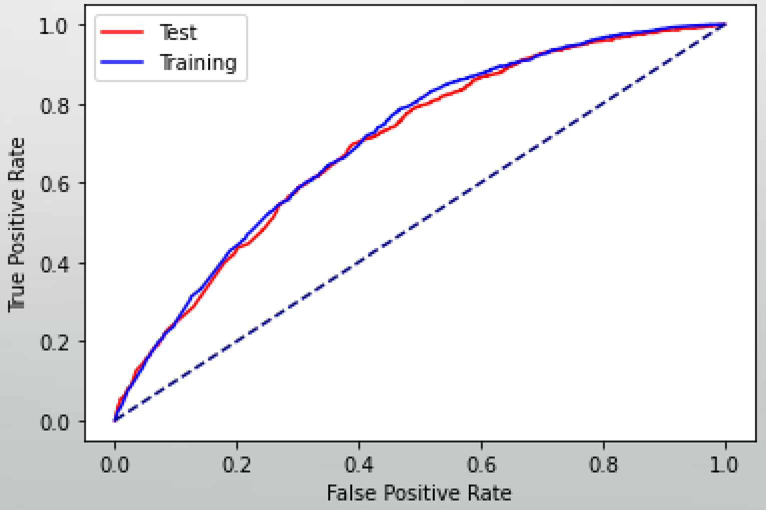
```
Pipeline(steps=[('rf',
                 RandomForestClassifier(class_weight='balanced', max_depth=5,
                                        max_samples=1000, n_estimators=10))])
```



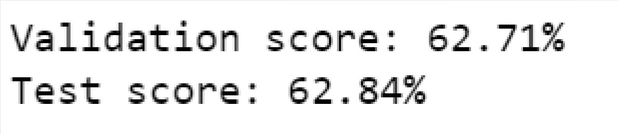
```
Validation score: 78.18%
Test score: 78.80%
```



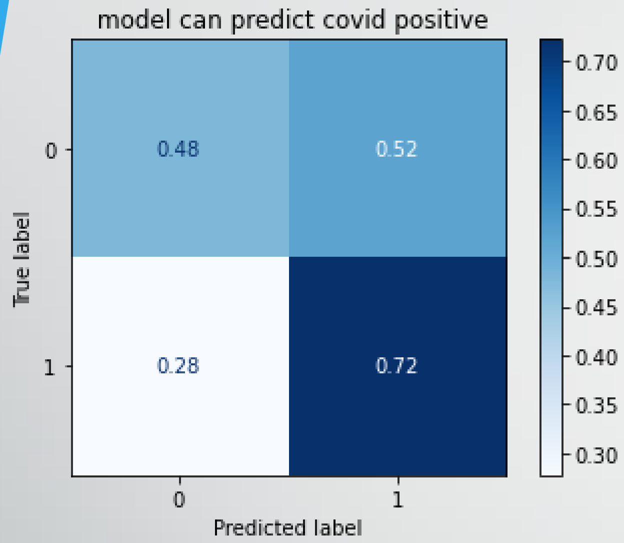
Random forest Confussion matrix on how the model can predict covid positive



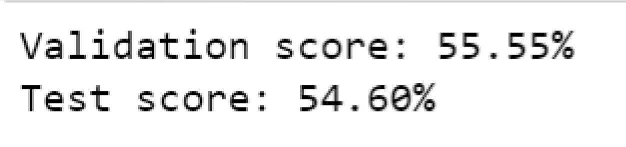
SVM



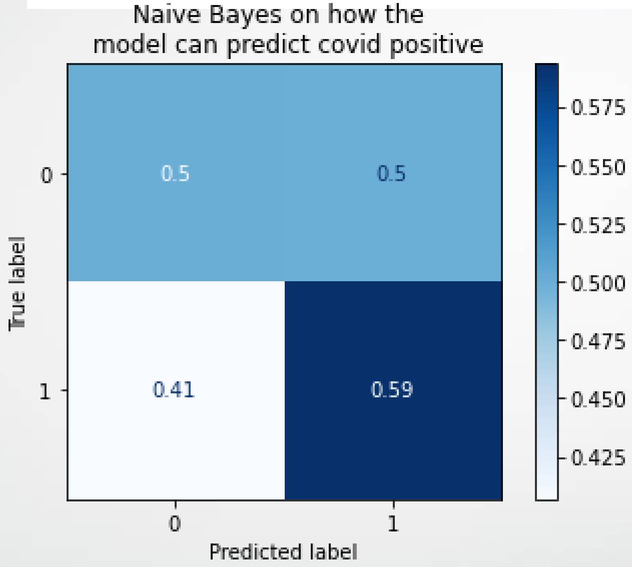
```
Validation score: 62.71%
Test score: 62.84%
```



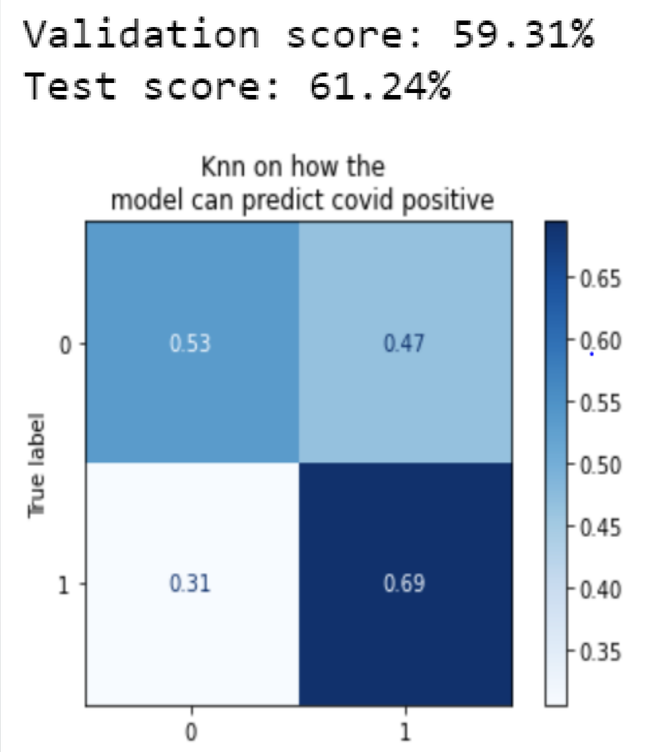
Naive Bayes



```
Validation score: 55.55%
Test score: 54.60%
```



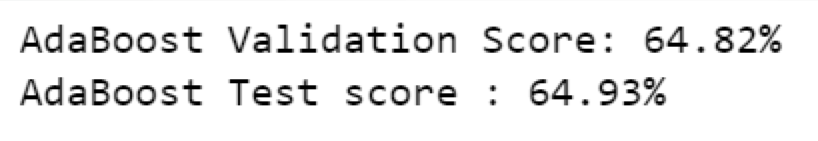
KNN



```
Validation score: 59.31%
Test score: 61.24%
```

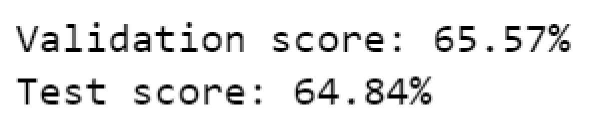
Knn on how the
model can predict covid positive

ADA Boost



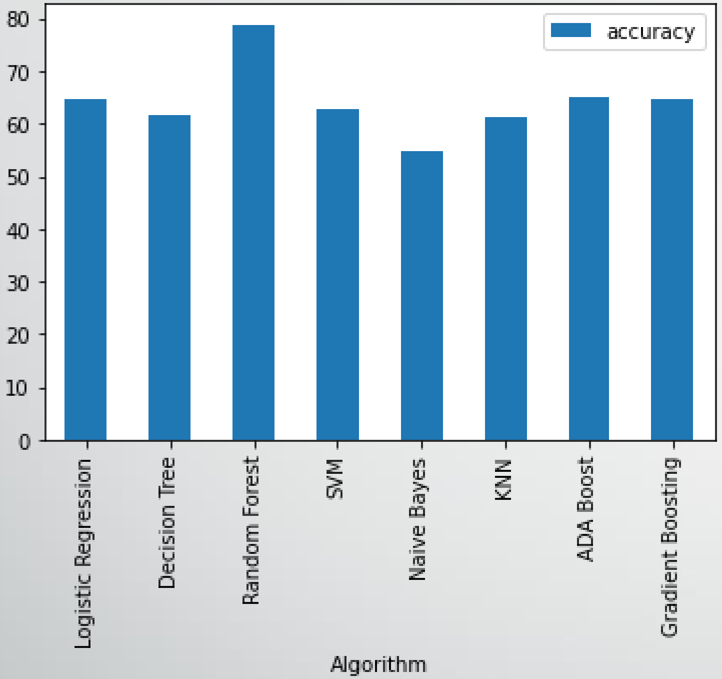
```
AdaBoost Validation Score: 64.82%
AdaBoost Test score : 64.93%
```

Gradient Boosting



```
Validation score: 65.57%
Test score: 64.84%
```

Conclusion



Early COVID-19 prediction can aid in minimizing the undue burden on healthcare systems by assisting in the diagnosis of COVID-19 patients. In this project prior to creating the models, the correlation coefficient analysis between various dependent and independent features was performed to establish the strength of the association between each dependent and independent feature of the dataset. 60% of the training dataset was utilized to train the models, while the remaining 40% was used to test the models. The Random forest model has the highest accuracy of 78.80% percent followed by logistic regression, ADA boost and Gradient Boosting which lies around 65%.The performance of all models was evaluated based on accuracy parameters

Thank you !!