# PREDICT COVID-19 WITH BEST MODEL BASED ON SYMPTOM

Md Fahim Faez Abir[1]
2018-1-60-032

Nazmus Sakib[2]
2018-1-60-104

Md. Tanbin Hossain Himel[3]
2018-1-60-151

Kamruzzaman Leeon[4]
2018-1-60-252

MAY 26, 2021

## 1 Introduction

Coronavirus disease (COVID-19) is a newly found coronavirus that causes an infectious illness. The majority of patients infected with the COVID-19 virus will have mild to moderate respiratory symptoms and will recover without needing any specific therapy. People over the age of 65 and infants, as well as those with underlying medical conditions such as cardiovascular disease, diabetes, chronic respiratory disease, and cancer, are at a higher risk of developing serious illness [1]. It was first detected in humans by scientists in 1965. Later in 2019, COVID was found in Wuhan city, china and renamed as COVID-19. And now it has become a global pandemic all over the world [2].

Therefore, our problem is to find the best estimator model and find the dependency of the best estimator model so that, we can distinguish COVID-19 patient. We intended to do the project because at present COVID-19 test is time-consuming and costly. Therefore, it is convenient for a person if he can get a pre-result before final test.

We have found some existing work on this problem. The motivation as well as the objective of these works are as same as ours but most of them have some lacking i.e. lack of real data-set, lack of authentic data set, complexity of the algorithms. We also found some image processing work on COVID-19 but seems less convenient at the present scenario in terms of cost [7].

The necessity of the project on this topic is very important at this moment, because currently we are currently undergoing a pandemic situation. The only prevention of COVID-19 is available at this moment is to take care of the patient after identifying symptoms. Therefore, it is convenient for us if we can distinguish COVID patient as early as possible and take initiative for taking care of the patient.

## 2 Methodology

Our first task was to collect real data of patients. After collecting several data-set we have to find relevant data-set, we preprocessed the data. After that, we analyze the importance of the features. When we finalize the features, we divide the data-set into 3 types and create five model i.e. Random Forest, Logistic Regression, Neural Network, Decision Tree  Naïve Bayes. Finally, we train our models with the data-sets and conclude our result.

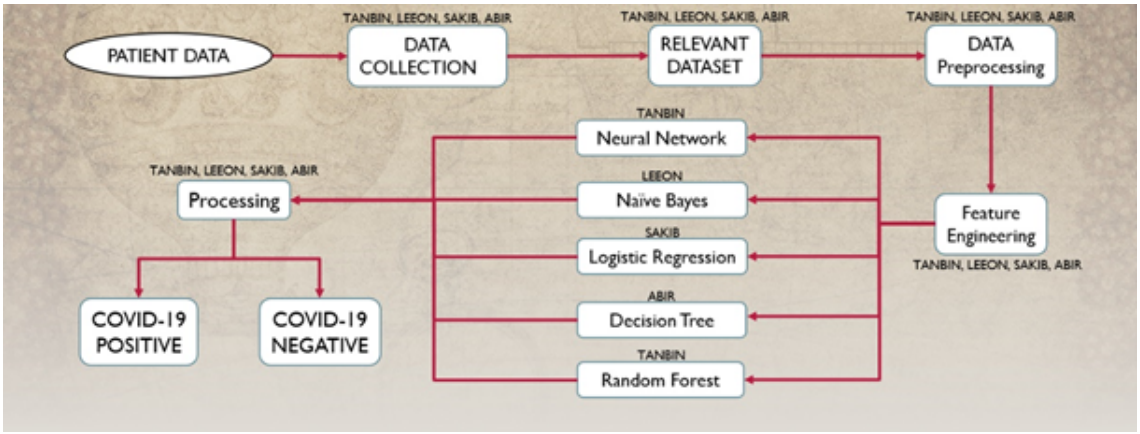



Figure 1: our methodology

## 3 Implement

### 3.1 Data collection

We have collected our data from Kaggle. We have found several data-sets on COVID-19. But we select a data-set titled "COVID-19 symptom checker" [8]. Because the data provider assures us that it was originally collected from WHO without any manipulation. Therefore it can be assumed that, it is an authentic data-set with real values.
Our dataset has 22 attributes i.e. 'Fever','Tiredness', 'Dry-Cough', 'Difficulty-in-Breathing', 'Sore-Throat', 'None_Sympton', 'Pains', 'Nasal-Congestion', 'Runny-Nose', 'Diarrhea', 'None_Experiencing', 'Age_0-9', 'Age_10-19', 'Age_20-24', 'Age_25-59', 'Age_60+', 'Gender_Female', 'Gender_Male', 'Gender_Transgender', 'Contact_Dont-Know', 'Contact_No', 'Contact_Yes'. Almost 316800 peoples data have been collected here. From the dataset we can observe that, gender is divided into 3

sub-columns, age is divided into 5 sub-columns. Also, we found that all the columns are related to COVID-19. We also notice that, the deciding factor of the dataset is divided into four sub-columns i.e Severity_Mild, Severity_Moderate, Severity_None, Severity_Severe.

### 3.2 Data Processing

Our data was not fully organized before. Here we observe that our data-set has a country column which has 10 different values. Therefore, we convert it into 10 sub-columns i.e. 'Country_China', 'Country_France', 'Country_Germany','Country_Iran','Country_Italy', 'Country_Other', 'Country_ Other-EUR', 'Country_Republic of Korean', 'Country_Spain' 'Country_UAE'. Since our deciding factor is divided into 4 sub-columns, we decided to make three types of data-sets here based on the deciding factor.

Table 1: Every table needs a caption.

| Value | Type-01 | Type-02 | Type-03 |
|---|---|---|---|
| None | 0 | 0 | 0 |
| mild | 1 | 0 | 1 |
| moderate | 1 | 1 | 2 |
| severe | 1 | 1 | 3 |

- **Type 1:** Here, we decided to convert the deciding factor into binary values 0 1 where 1 means 'yes' and 0 means 'no'. If we ignore the level of severity, 'Mild', 'moderate' 'severe' means that the person actually has COVID-19. Therefore, we convert these three values into 'yes'. On the other hand, None means the person does not have COVID-19. Therefore, we convert it into 'no'.
- **Type 2:** Here also, we decided to convert the deciding factor into binary values 0 1 where 1 means 'yes' and 0 means 'no'. 'None' 'Mild', means that the person does not have COVID-19 or has a few symptoms of COVID-19. Therefore, we assume it as negative and convert these two values into 'no'. On the other hand, 'moderate' 'severe' means the person has symptoms of COVID-19 but the severity of the symptoms is higher than 'None' and 'Mild'. Therefore, we assume it as positive and convert these two values into 'yes'.
- **Type 3:** Here we divided the deciding factor according to its level of severity. The more the values the higher the level of severity. Therefore, 'none' is converted into 0, 'mild' is 1, 'moderate' is 2 and severe is 3.

With these three types of data, we decided to train and test our model separately and compare them.

## 4 Model Development

We have worked with five models i.e. Random forest, Logistic Regression, Decision Tree, Gaussian Naive Bayes Neural networks so that we can compare the accuracy between the models and find

the best estimator.

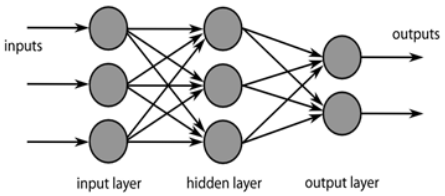
- **Random forest:**A random forest is a meta estimator that uses averaging to improve projected accuracy and control over-fitting by fitting several decision tree classifiers to different sub-samples of the dataset. The sub-sample size is controlled by the max samples parameter if bootstrap=True (default); otherwise, the whole dataset is used to generate each tree [3].
- **Logistic Regression:**Logistic regression is a classification model, not a regression model, despite its name. In the literature, logistic regression is also known as logit regression, maximum-entropy classification (MaxEnt), or the log-linear classifier. In this approach, a logistic function is employed to characterize the likelihood of the likely outcomes of a single experiment [4].
- **Decision Tree:** The decision Tree is a supervised learning method for classification and regression that is non-parametric. The goal is to learn basic decision rules from data attributes and build a model that predicts the value of a target variable. A piecewise constant approximation is a tree. The decision criteria become more complex as the tree goes deeper and the model becomes more accurate [5]. The necessary equation to calculate decision tree:

$$Info(D) = -\sum_{i=1}^{m} p_i \lg(p_i)$$
$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} Info(D_j)$$
$$Gain(A) = Info(D) - Info_A(D)$$

- **Gaussian Naive Bayes:**Naive Bayes approaches are supervised learning methods based on Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the class variable value. The Gaussian Naive Bayes approach is used in GaussianNB, a classification tool[6]. The features' probability is considered to be Gaussian:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

- **Neural Network:** A neural network is a reliable method for forecasting real, discrete, or vector data. It's based on the biological learning system and can tackle biological as well as non-biological challenges.




Nodes, edges, edge widgets, and bias are the representative elements.

## 5 Result

The results of the models is given below:

| Value | Type-01 | | Type-02 | | Type-03 | |
|---|---|---|---|---|---|---|
| Name | accuracy | F1 accuracy | accuracy | F1 accuracy | accuracy | F1 accuracy |
| Random Forest | 68.96% | 40.81% | 16.35% | 16.35% | 0.21% | 0.21% |
| logistic Regression | 74.91% | 53.33% | 49.45% | 49.45% | 24.05% | 23.87% |
| Decision Tree | 53.33% | 34.83% | 16.35% | 14.95% | 0.21% | 0.15% |
| Gaussian Naive Bayes | 74.91% | 42.83% | 49.58% | 49.33% | 24.29% | 24.06% |
| Neural Network | 74.91% | 42.83% | 42.83% | 43.41% | 23.53% | 15.22% |

Here we observe that the accuracy level depends on the classification of the deciding factor. The accuracy of the Type-03 dataset is ranging from 0 to 25%, the Type-02 dataset is ranging from 16-50%, and the Type-01 is ranging from 53-74%. After observing the ROC curve we can see that, false positive rate is much more higher than true positive rate. As a reason we can observe that, the value of the dataset is not sufficient since COVID-19 is a new case for the world. As well as there is a lot of anomaly in the dataset. For that reason, the model is confused to truly identify positive cases and ROC curve value is negative. We can also conclude that the accuracy is dependent on the model. Here we can see that Random Forest  Decision Tree has less accuracy in every type of dataset. On the other hand, Logistic Regression, Naive Bayes, and Neural Network have the same type of higher accuracy. In term of f1 score, we can see that f1 scores value is higher in type-02 dataset compared to type-01 dataset. From the above discussion we can conclude that, Logistic Regression is work well among all types of dataset since its accuracy and f1 accuracy are higher than other models. Therefore our best estimator model is Logistic Regression for distinguish COVID-19 patient.

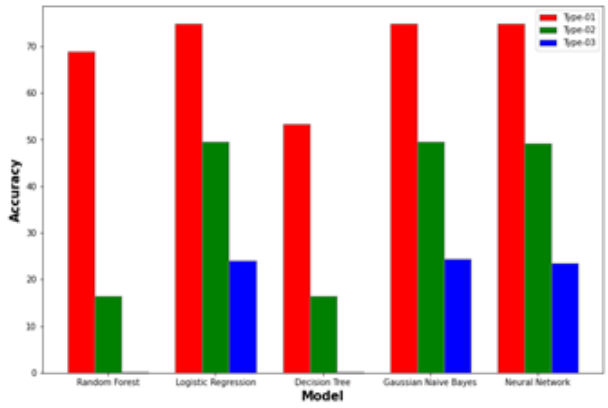



Figure 2: Accuracy

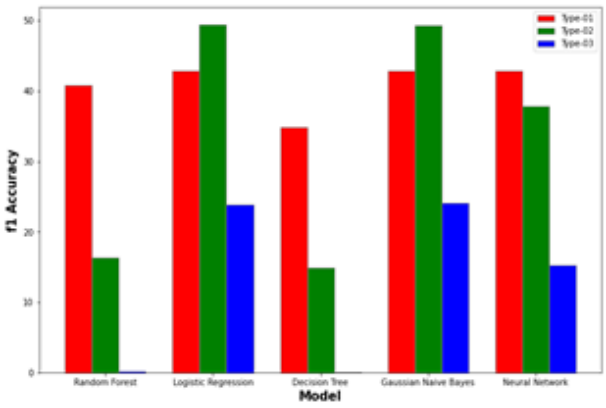



Figure 3: F1

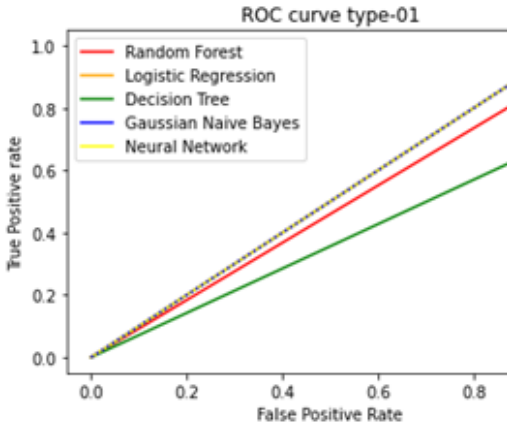



Figure 4: ROC curve type-01

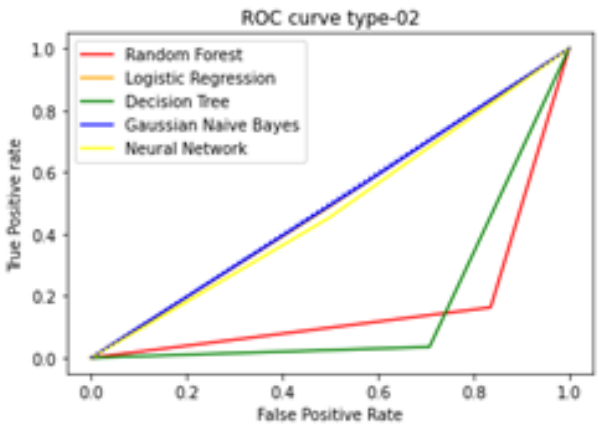



Figure 5: ROC curve type-02

# 6 Conclusions

## 6.1 challenges

The first challenges we have faced to collect the authentic dataset with real values. There are several datasets available but either they are found not authentic or the dataset is manipulated. Since we are trying to estimate the best model with real values, we must avoid these datasets. Another challenge was to find the higher accuracy of the best estimators. Since COVID-19 is a new disease we do not have enough data to take the accuracy at its peak.

## 6.2 Limitations

In the Type-3 dataset, we divided the deciding factor according to its level of severity. As a result, It has 4 values ranging from 0 to 3. As a result, its accuracy is too less than our expectation and we are not able to make any perfect model using it.

## 6.3 Future direction

We have some learning from our project. Our first learning is to work with the project with sufficient amount of real values. Otherwise it is not possible for the model to predict the COVID-19 in real life scenario. Since, We found that Type-01 Type-02 dataset has higher accuracy than Type-03 dataset, Therefore our recommendation is to classify the deciding factor into binary values. Also, we recommend to use dataset with having less anomaly so that the model can distinguish COVID-19 patient.

## References


[1] World Health Organization: Coronavirus,
https://www.who.int/health-topics/coronavirus#tab=tab_1

[2] WebMD: Coronavirus History,
https://www.webmd.com/lung/coronavirus-history

[3] scikit-learn: Random Forest Classifier,
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.h

[4] scikit-learn: Logistic Regression,
https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

[5] scikit-learn: Decision Trees,
https://scikit-learn.org/stable/modules/tree.html

[6] scikit-learn: Gaussian Naive Bayes,
https://scikit-learn.org/stable/modules/naive$_b$*ayes.html*/*gaussian* − *naive* − *bayes*

[7] Kaggle: COVID-19 X-ray image classification,
https://www.kaggle.com/c/stat946winter2021

[8] Kaggle: COVID-19 Symptoms Checker,
https://www.kaggle.com/iamhungundji/covid19-symptoms-checker