Neha Kale 2018130018

Shivangi Kochrekar 2018130020

Rishita Mote 2018130029

**Machine Learning Lab Report**

Machine Learning(ML) can be explained as automating and improving the learning process of computers based on their experiences without being actually programmed i.e. without any human assistance. The process starts with feeding good quality data and then training our machines(computers) by building machine learning models using the data and different algorithms. The choice of algorithms depends on what type of data we have and what kind of task we are trying to automate.

**Supervised Machine Learning**

Data scientists use many different kinds of machine learning algorithms to discover patterns in big data that lead to actionable insights. At a high level, these different algorithms can be classified into two groups based on the way they “learn” about data to make predictions: supervised and unsupervised learning.

**Supervised Machine Learning:** The majority of practical machine learning uses supervised learning. Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output Y = f(X) . The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data. Supervised learning requires that the data used to train the algorithm is already labeled with correct answers.

### **Linear Regression**

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

***Aim:***

To understand and implement a linear regression model.

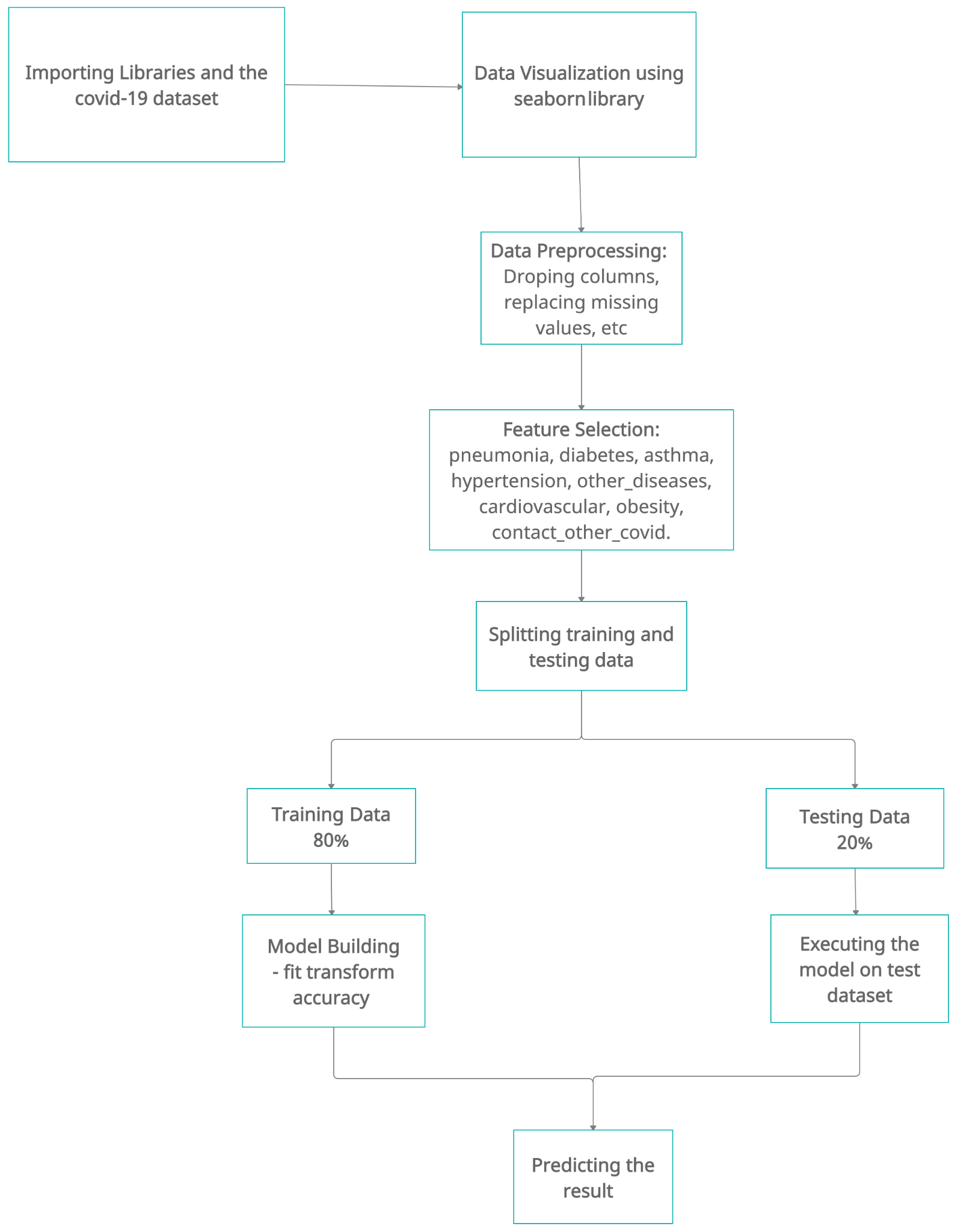
***Problem Statement:***

To predict whether a covid positive person needs to be admitted to an ICU or not based on prevailing health and other factors.

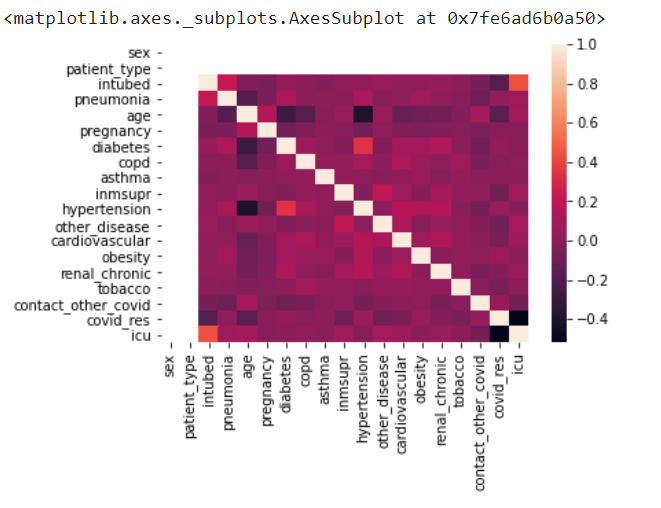
***Objective:***

1. To understand sklearn library and its functions
2. To understand data preprocessing techniques
3. To find out the relevant features for the model
4. To understand linear regression

#### ***Methodology:***



***Correlation Matrix:***

******

From the correlation matrix, we selected important features for the model. The features selected are: ‘sex', 'patient\_type', 'intubed', 'pneumonia', 'pregnancy',

'diabetes', 'copd', 'asthma', 'inmsupr', 'hypertension',

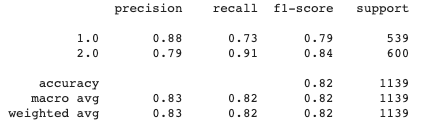
'other\_disease', 'cardiovascular', 'obesity', 'renal\_chronic',

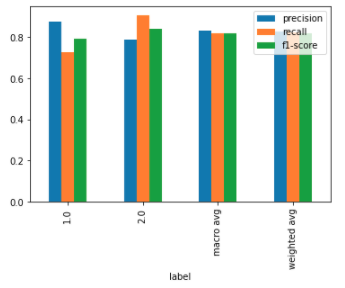
'tobacco', 'contact\_other\_covid', 'covid\_res'.

Using these features we trained our regression model & the result is as follows.

***Results:***

We got an accuracy of **82.1%** and successfully found out the important features.

******

******

#### ***Conclusion:***

1. We learned about various models provided by ScikitLearn
2. Understood the Data processing techniques
3. We got an accuracy of **82.1%** and successfully found out the important features.
4. We found out the relevant features for the model which were pneumonia, diabetes, asthma, hypertension, other\_diseases, cardiovascular, obesity, contact\_other\_covid.

## **Logistic regression**

Logistic regression is a statistical model that uses Logistic function to model the conditional probability. This is read as the conditional probability of Y=1, given X or conditional probability of Y=0, given X. An example of logistic regression can be to find if a person will default their credit card payment or not.

***Aim:***

To understand and implement a linear regression model.

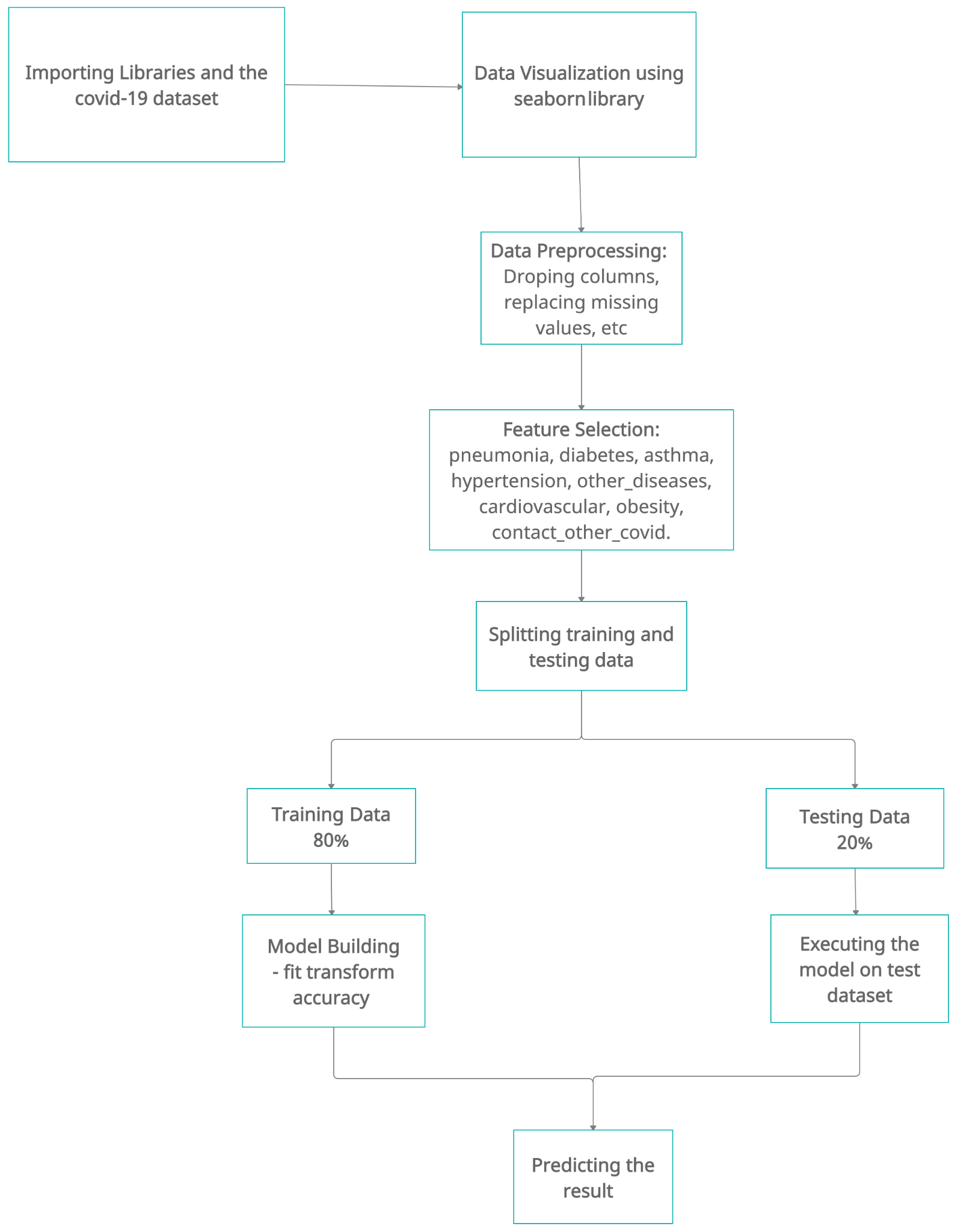
***Problem Statement:***

To predict whether a covid positive person needs to be admitted to an ICU or not based on prevailing health and other factors.

***Objective:***

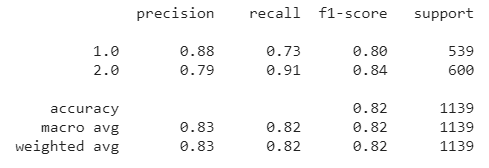
1. To understand sklearn library and its functions
2. To understand data preprocessing techniques
3. To find out the relevant features for the model
4. To understand logistic regression

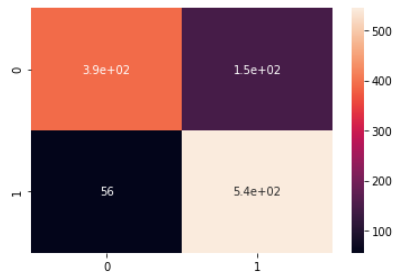
#### ***Methodology:***



***Results:***

We got an accuracy of 83.3% and successfully found out the important features.

******

******

#### ***Conclusion:***

1. We got a clear understanding of logistic regression and implemented the same using sklearn
2. The algorithm gave 83.3% accuracy
3. Enhanced accuracy using data preprocessing
4. We found out the relevant features for the model which were pneumonia, diabetes, asthma, hypertension, other\_diseases, cardiovascular, obesity, contact\_other\_covid.

## **Classification**

A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. A classification model attempts to draw some conclusions from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.

For example, when filtering emails “spam” or “not spam”, when looking at transaction data, “fraudulent”, or “authorized”. In short Classification either predicts categorical class labels or classifies data (construct a model) based on the training set and the values (class labels) in classifying attributes and uses it in classifying new data. There are a number of classification models. Classification models include decision tree, random forest, gradient-boosted tree, multilayer perceptron, one-vs-rest, and Naive Bayes.

***Aim:***

To understand and implement a classification model.

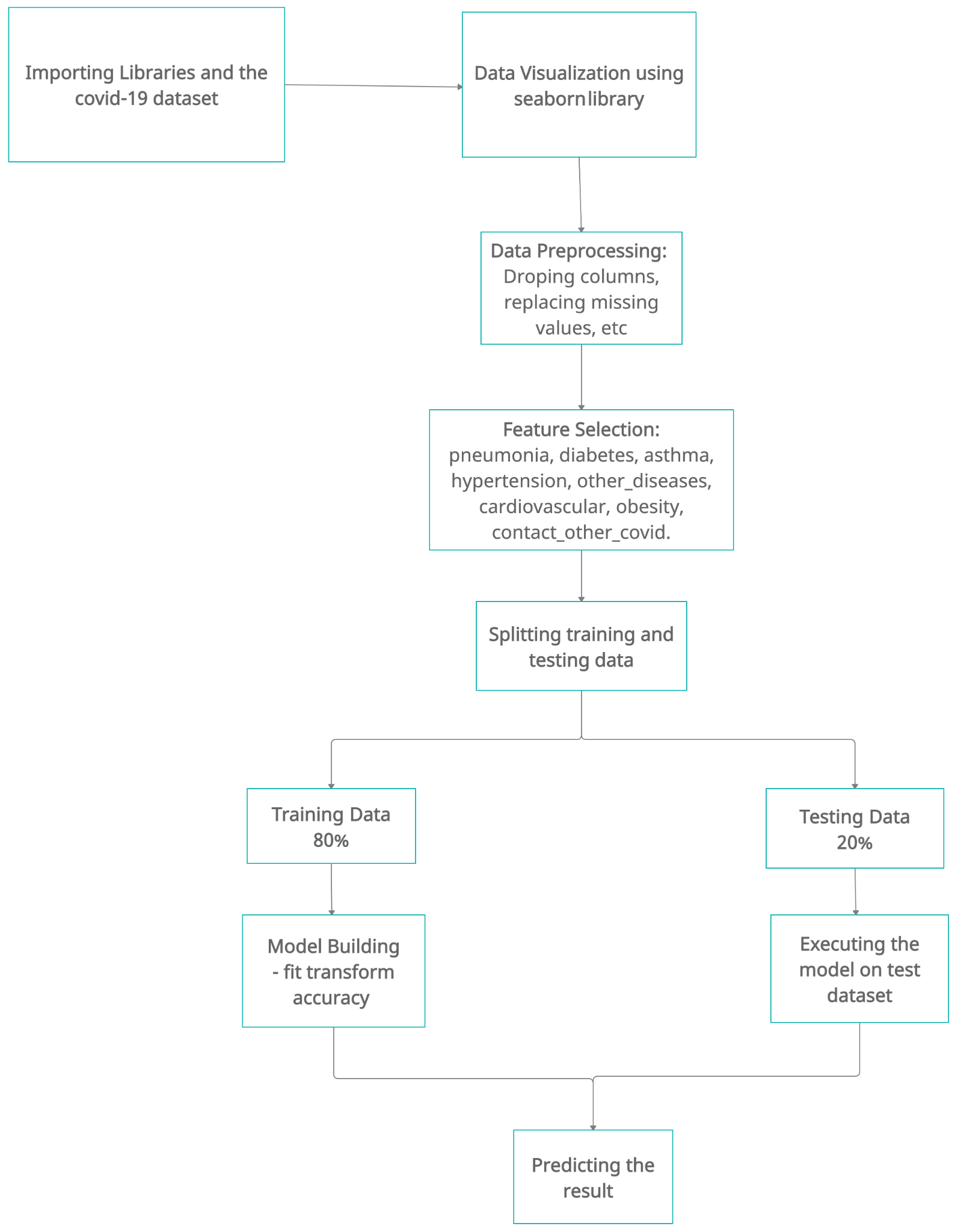
***Problem Statement:***

To predict whether a covid positive person needs to be admitted to an ICU or not based on prevailing health and other factors and testing of covid reports

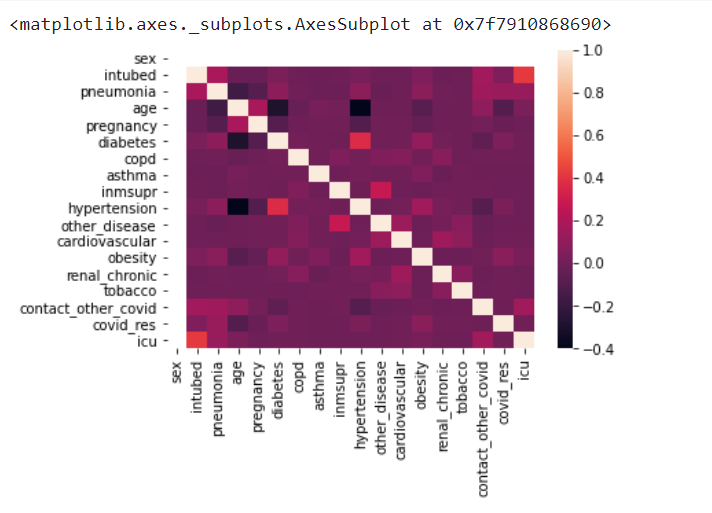
***Objective:***

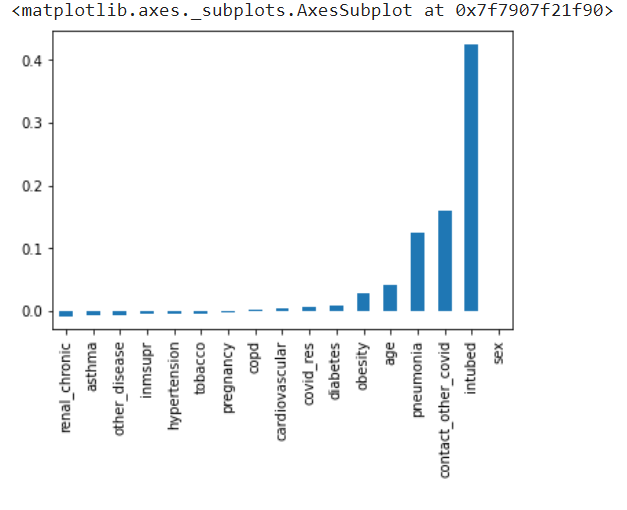
1. To understand sklearn library and its functions
2. To understand classification models
3. To understand data preprocessing techniques

#### ***Methodology:***



Correlation Matrix to select features for our model.



******

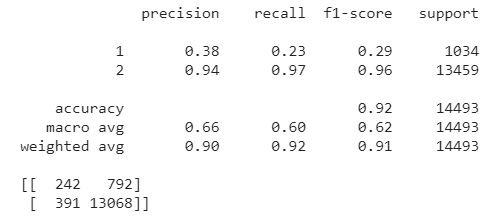
***Results:***

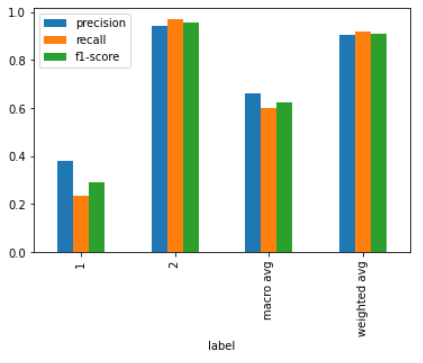
Random Forest: 92%

Decision Tree: 90%

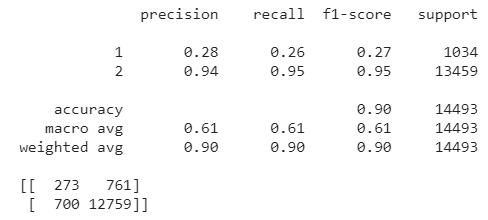
Gradient Boosting: 93%

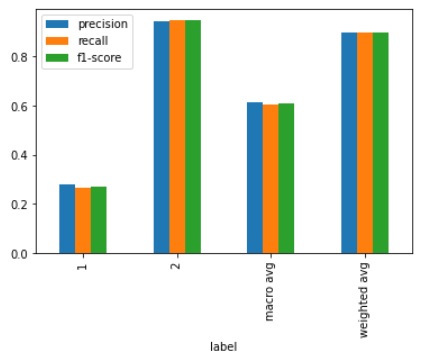
**Random Forest Classifier:**

******

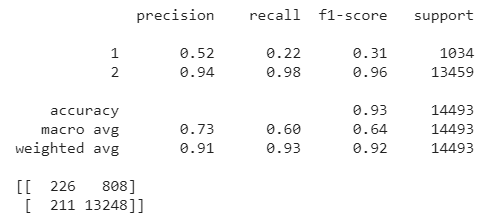
******

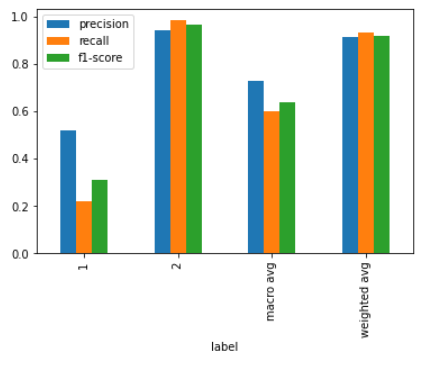
**Decision Tree Classifier:**





**Gradient Boosting Classifier:**





#### 

#### 

#### 

#### 

#### 

#### 

#### 

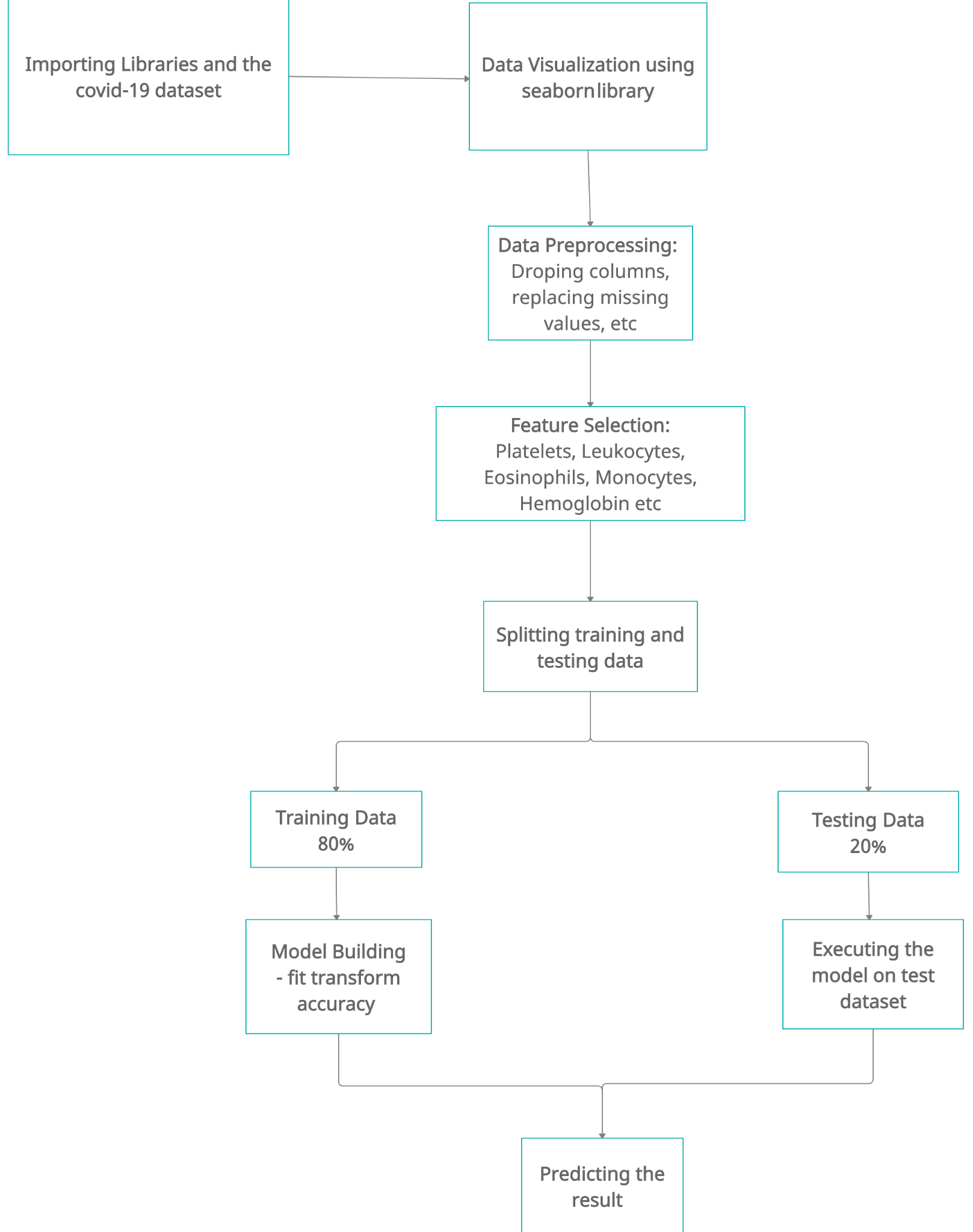
#### 

#### 

#### 

#### 

#### ***Methodology for KNN:***



#### 

#### 

#### 

#### 

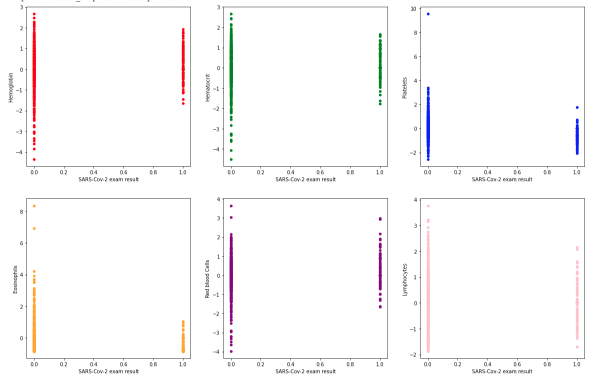
#### 

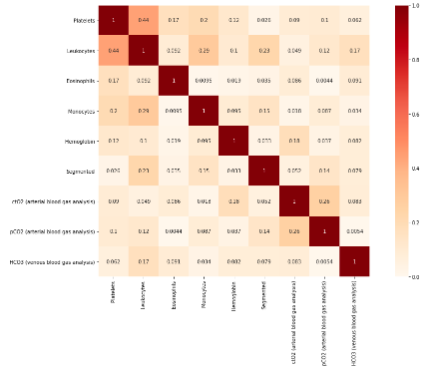
#### 

#### 

#### 

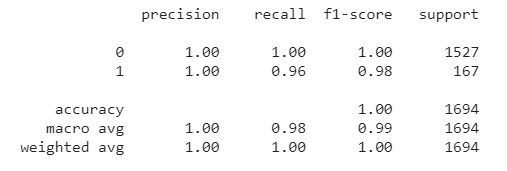
#### ***Scatter Plots and Correlation figure for the features:***

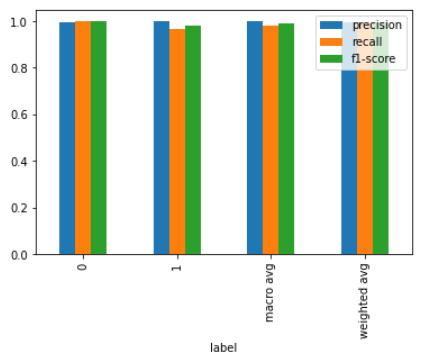


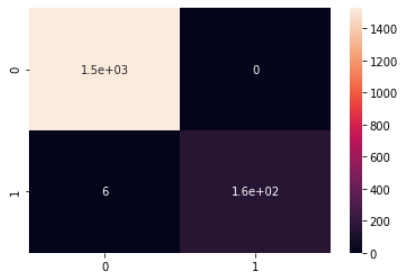


***Results:***

We got an accuracy of 99.6% and successfully found out the important features.

******

******

******

#### ***Conclusion:***

1. We understood classification models and implemented Random Forest Classifier, Decision Tree Classifier and Gradient Boosting Classifier.
2. The accuracies were found out to be as follows:

Random Forest: 92%

Decision Tree: 90%

Gradient Boosting: 93%

1. Gradient Boosting proved to be the most effective algorithm.
2. We also implemented a KNN classifier.
3. We got an accuracy of 99.6% and successfully found out the important features.
4. The relevant features were Platelets, Leukocytes, Eosinophils, Monocytes, Hemoglobin, Segmented, ctO2 (arterial blood gas analysis), pCO2 (arterial blood gas analysis), HCO3 (venous blood gas analysis).

**Unsupervised Machine Learning**

Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.

## **Clustering**

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

***Aim:***

To understand and implement clustering model.

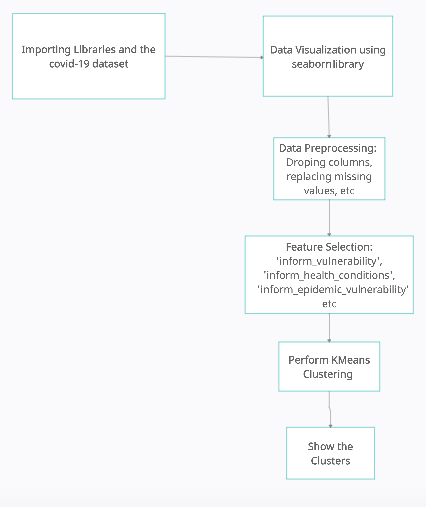
***Problem Statement:***

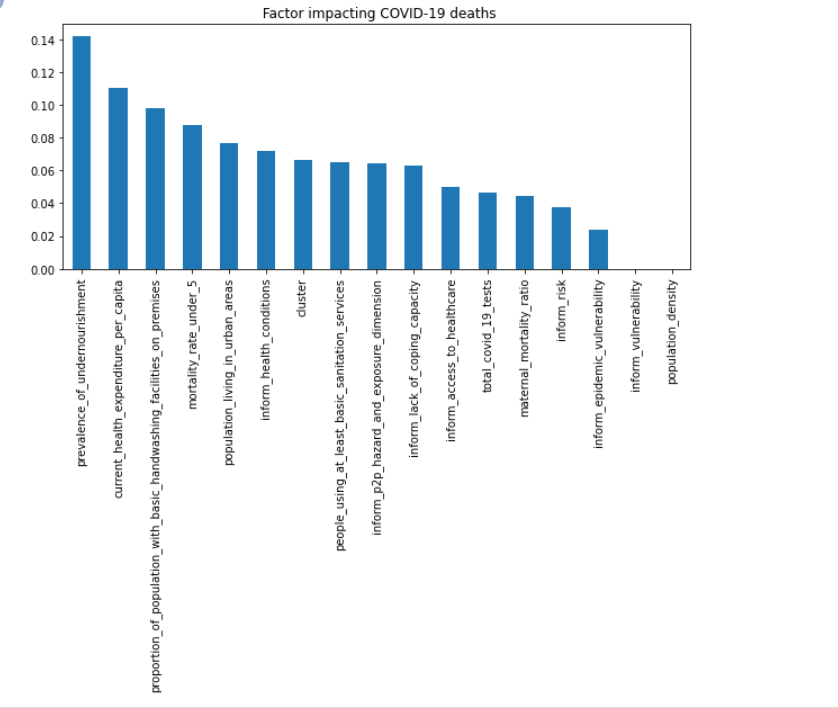
To make clusters based on different health care quality, number of COVID-19 tests performed and numbers of COVID-19 cases confirmed in various countries.

***Objective:***

1. To understand KMeans Clustering
2. To understand data preprocessing techniques
3. To understand clustering model

#### ***Methodology:***





#### 

#### 

#### 

#### 

#### 

#### 

#### 

#### 

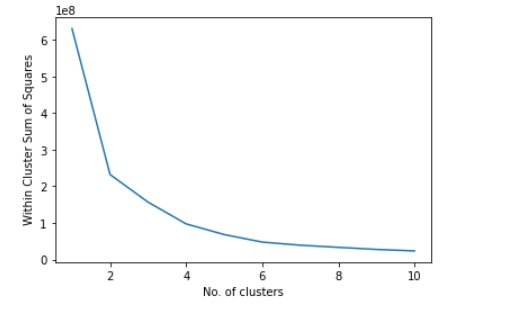
#### 

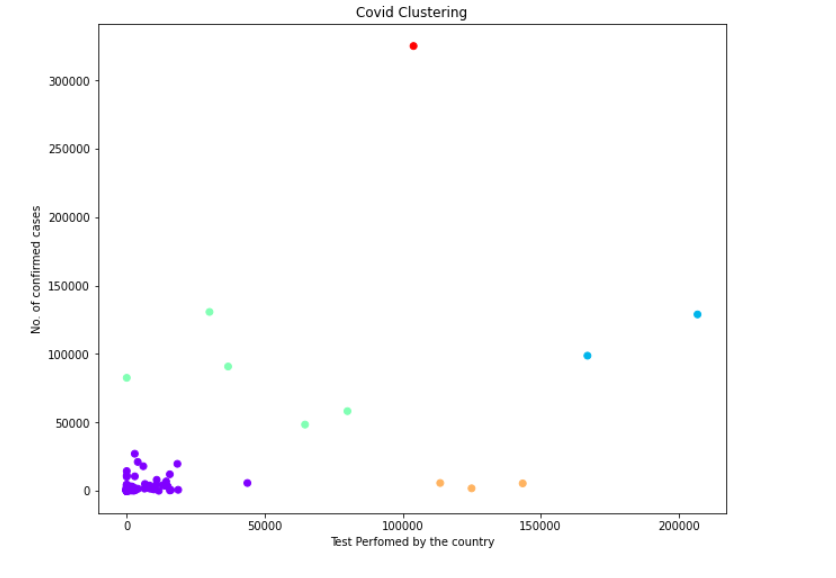
#### 

#### 

#### ***Results:***

We formed 5 clusters. The elbow graph and the clusters observed are shown below.





#### 

#### ***Conclusion:***

1. We understood clustering models and implemented KMeans Clustering.
2. We found out the factors impacting covid 19 deaths and then made 5 clusters on the basis of risks faced by the countries.

#### 

#### ***Future Scope:***

Hence, we can use the prediction of whether the patient needs to go to ICU or not to further detect whether the case is mild or severe. We can use the results of covid testing model to produce faster and more accurate results of covid testing and help the health sector be more efficient. The testing and severity models can be combined to make a combined prediction. The clustering can help us find out which countries need more vaccine dosage and healthcare equipment.

#### ***Conclusion:***

1. We found out that Gradient boosting gives the best accuracy of 93% for the prediction of ICU for similar features than the regression and other classification models.
2. The KNN model gives an accuracy of 99% for the prediction of covid testing which is because of all the important features selected from all the given parameters.
3. The clusters were made according to the risks faced by each country.
4. Hence, the datasets were worked on thoroughly and good resulting models were created.

#### ***References:***

1. <https://link.springer.com/article/10.1007/s41870-020-00552-3>
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7128942/>
3. <https://www.kaggle.com/tanmoyx/covid19-patient-precondition-dataset>
4. <https://www.kaggle.com/einsteindata4u/covid19>
5. <https://www.kaggle.com/roche-data-science-coalition/uncover>