****

**Master of Data Science (Semester 1 – 2020/2021)**

**Faculty of Computer Science & Information Technology**

**WQD7006 MACHINE LEARNING FOR DATA SCIENCE**

**Assignment Title:**

**Coronaviruses (COVID-19) Prediction based on Symptoms**

**Student Details:**

**Mun Siu Hou**

**17218313**

**17218313@siswa.um.edu.my**

## Introduction

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness. The best way to prevent and slow down transmission is be well informed about the COVID-19 virus, the disease it causes and how it spreads. Protect yourself and others from infection by washing your hands or using an alcohol-based rub frequently and not touching your face. The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes, so it is important that you also practice respiratory etiquette (for example, by coughing into a flexed elbow). At this time, there are no specific vaccines or treatments for COVID-19. Therefore, it is a need to provide a symptoms self-checker system for us to determine whether we are infected with this disease or not. This self-checker system not only allow the user to do a pre-testing but also helping the health care workers to reduce their workload.

## Objective

1. To provide help for the user to check whether is infecting with the coronavirus disease or not.
2. To help to reduce the workload of the heath care workers where those medical supplies are limited.

## Methodology

There will be several stages for this assignment such as data preparation, data cleaning, data visualization and etc. Each stage will be discussed in each section. Figure 1 showed the general flow of this project.

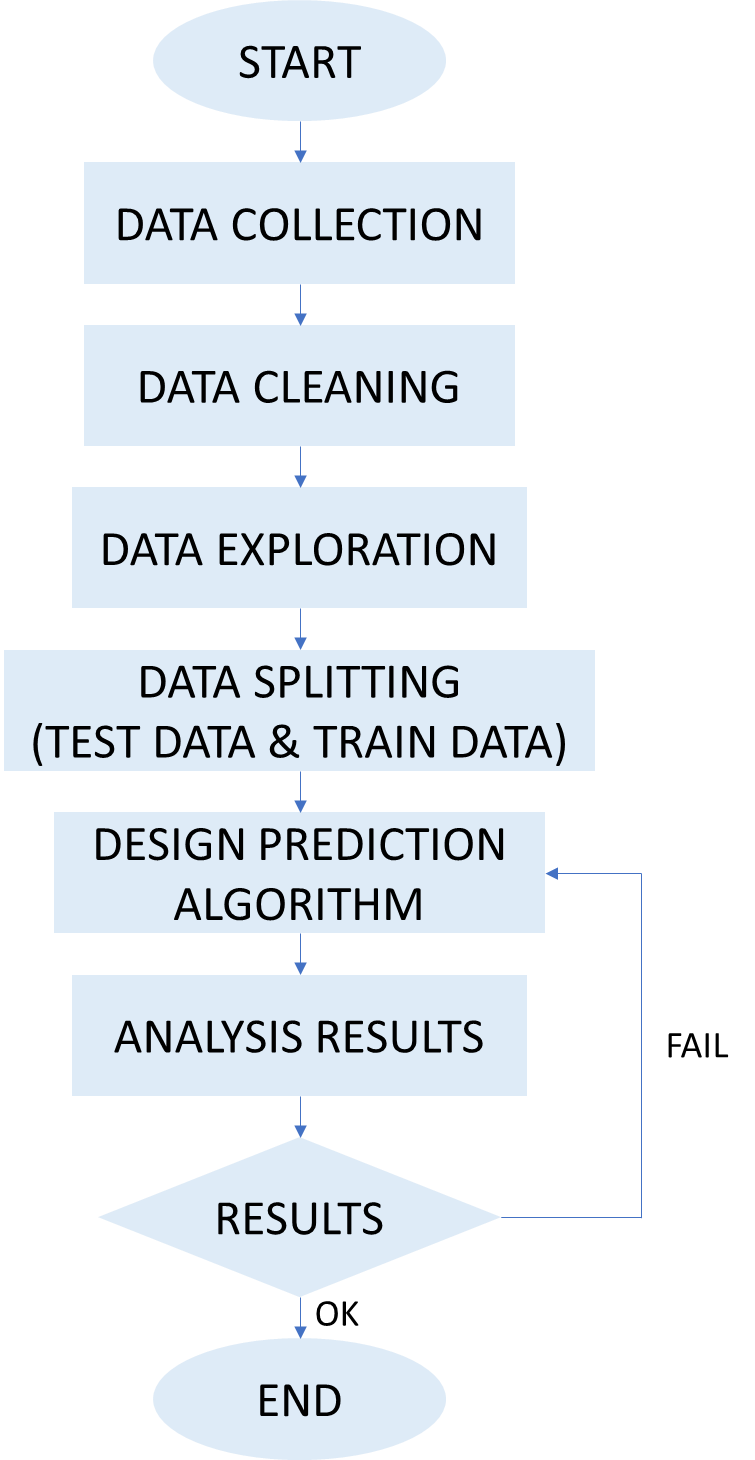


Figure 1: General flow of the Project

First, this assignment starts with data collection. The data that this project looking for is patient data with symptoms and detecting results. Next, this assignment will continue with data cleaning, data exploration and data splitting. After that, will research about suitable algorithms to do prediction. Next, will go through training and testing process and then only go for analysis on results. If the results are not as expecting, will redesign a better algorithm.

## Data Collection Stage

First, this assignment will be using the data set from the [website](https://data.gov.il/dataset/covid-19/resource/d337959a-020a-4ed3-84f7-fca182292308?inner_span=True). This website is from open data from government Israel. Therefore, it required to translate the language before we access the website. Because of the privacy issue, a lot of hospital will not be sharing their patient information and symptoms details.

## Data Cleaning Stage

After downloading the dataset from the website, this project will need to do the data cleaning process to make sure the data are ready to be used. Before the data cleaning process, this project also has to make sure that what type of data will be using later. Therefore, at first, this project will check the data shape, structure, and their features. Next, this project will go through each column to check for the missing data. If the missing data is only a very small portion of the whole dataset, this project will directly remove it. Besides for the symptoms of the patients, this dataset also included the activities details for the patients whereby it also recorded the patient’s records on 14 days traveling history and get contact history.

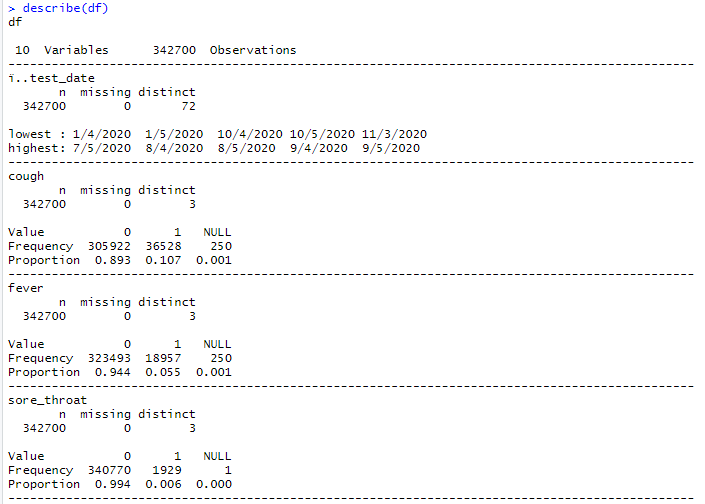


Figure 2: Result for data description

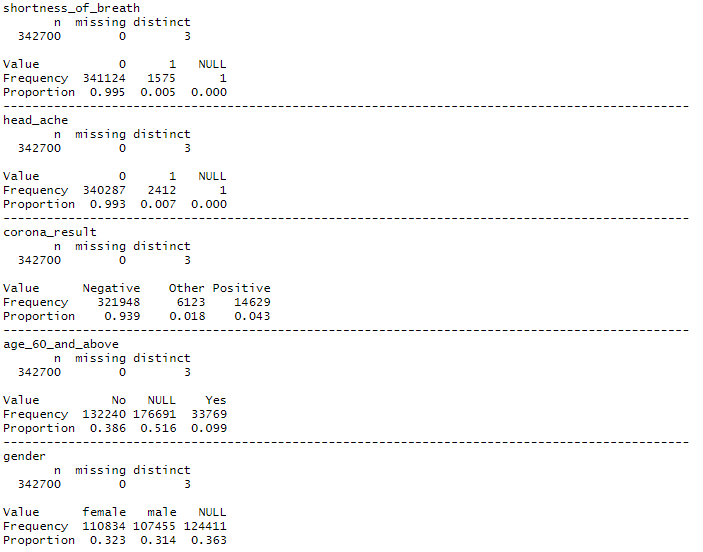


Figure 3: Result for data description

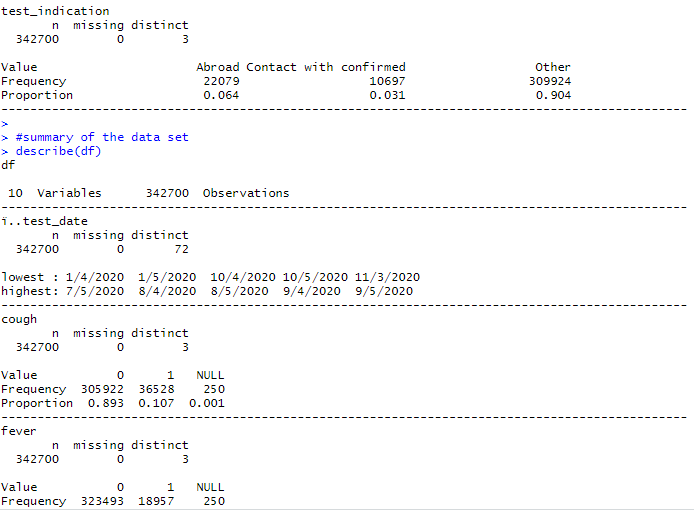


Figure 4: Result for data description

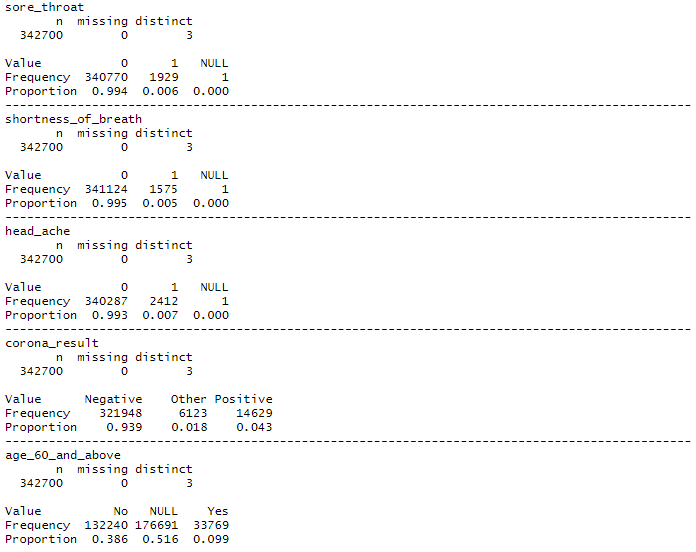


Figure 5: Result for data description

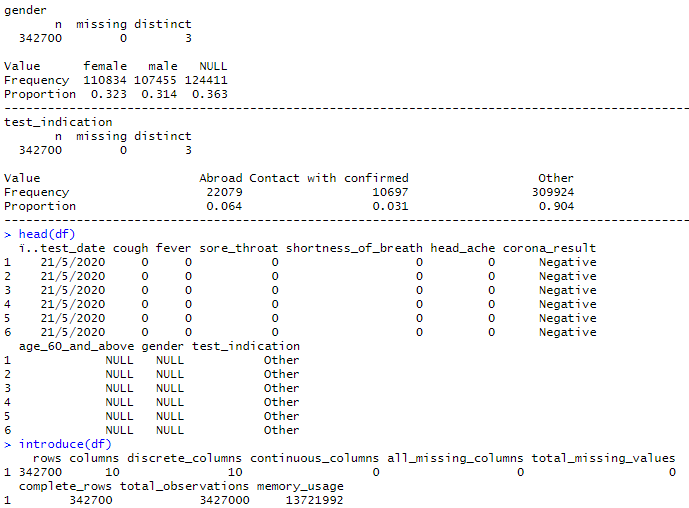


Figure 6: Result for data description

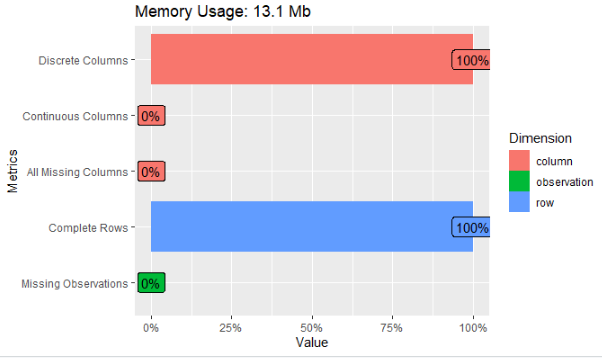


Figure 7: Plot of the dataset

This dataset does not have any missing column because all the missing columns are considered as NULL value.

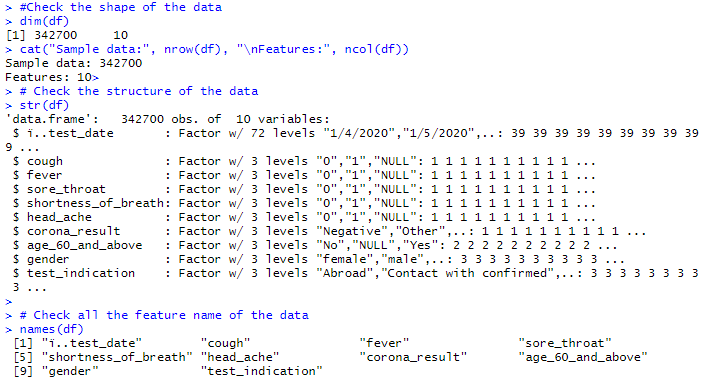


Figure 8: Check the structure of the dataset

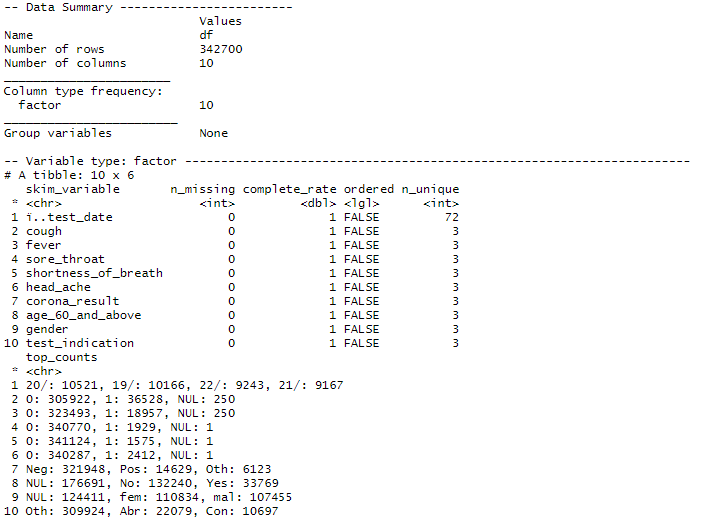


Figure 9: Data Summary

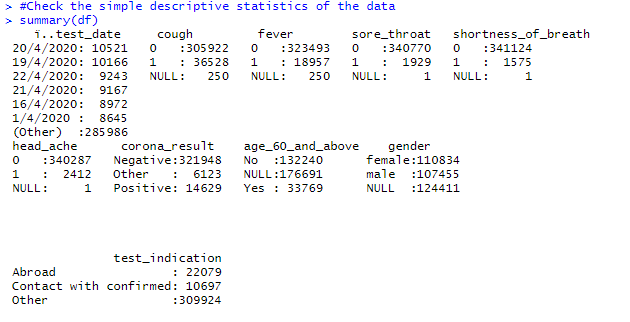


Figure 10: Simplify version of descriptive statistics for the dataset

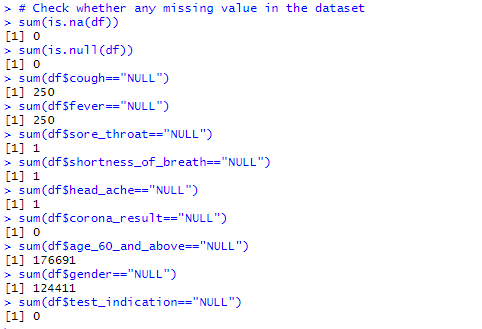


Figure 11: Check Missing value

After proceed with all the data cleaning process, the clean dataset will be looking like figure 12.

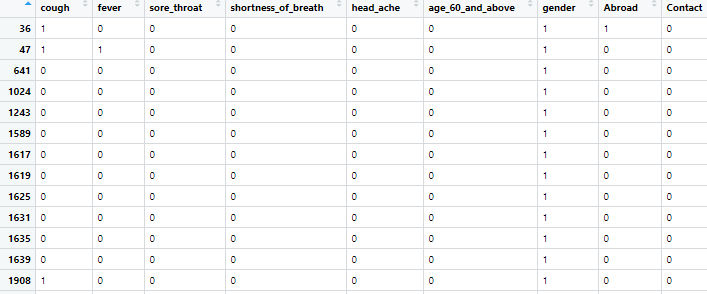


Figure 12: Clean data

## Data Exploration Stage

In this data exploration stage, this project will do some analysis on the data which include some comparison between the symptoms and the results.

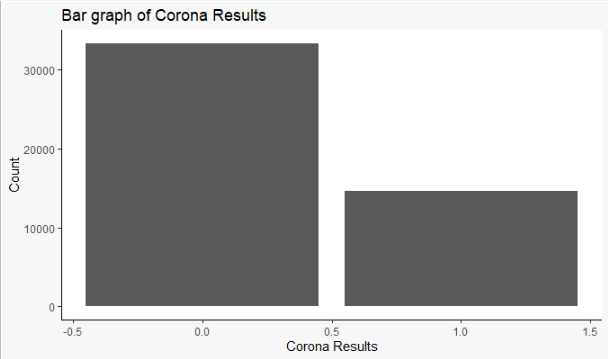


Figure 13: Overall corona result of this dataset

From figure 13, the dataset showed the patient that does not infect with the virus is lesser than who infected.

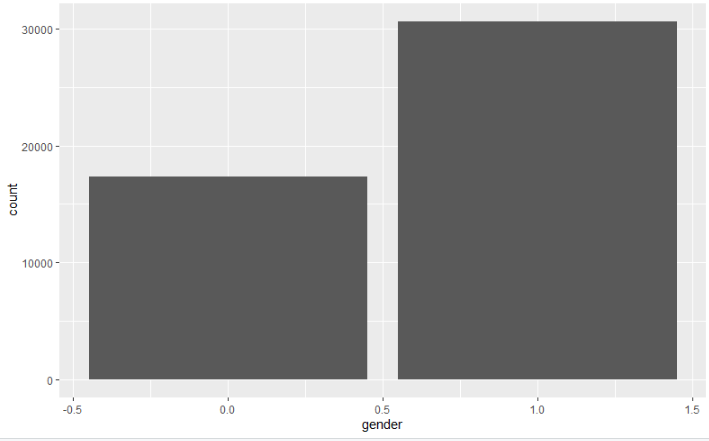


Figure 14: Comparison between gender and corona result

From figure 14, this dataset has more male patient who infected this corona virus than female patient.

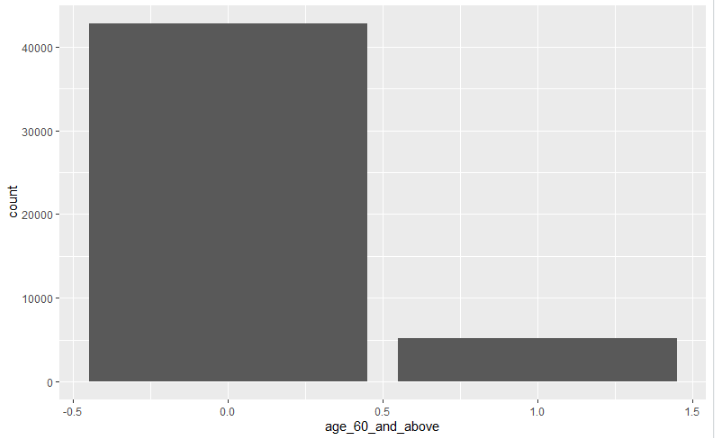


Figure 15: Comparison between age and corona result

Figure 15 showed that generally there are more young patient who has the corona virus than the older one.

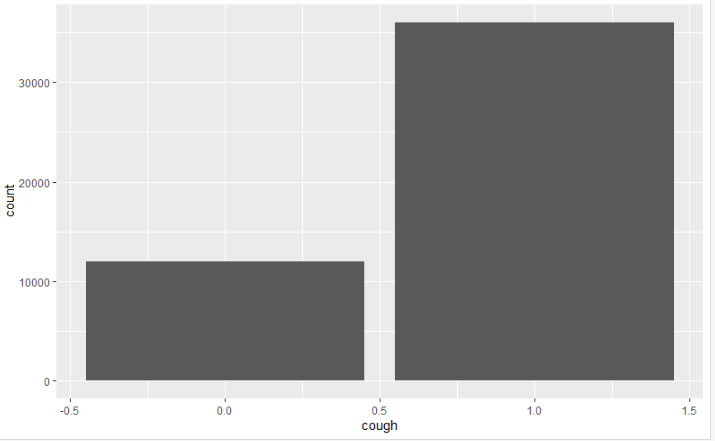


Figure 16: Comparison between cough and corona result

Figure 16 showed the patient who having symptom like cough have higher chances to get infected with this corona virus.

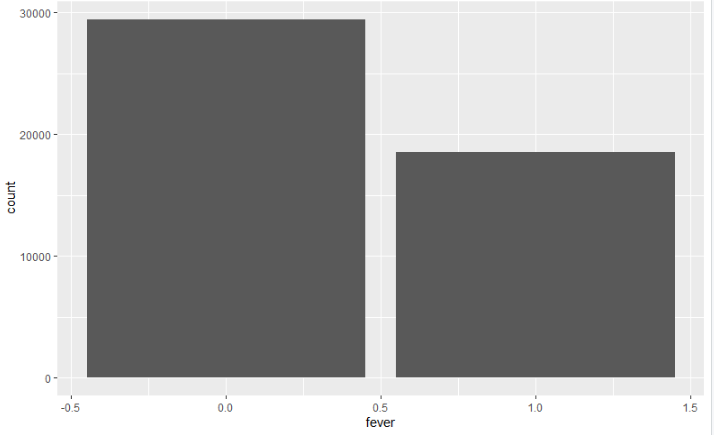


Figure 17: Comparison between fever and corona result

Figure 17 showed the patient who having symptom like fever are not necessary to get infected with this corona virus.

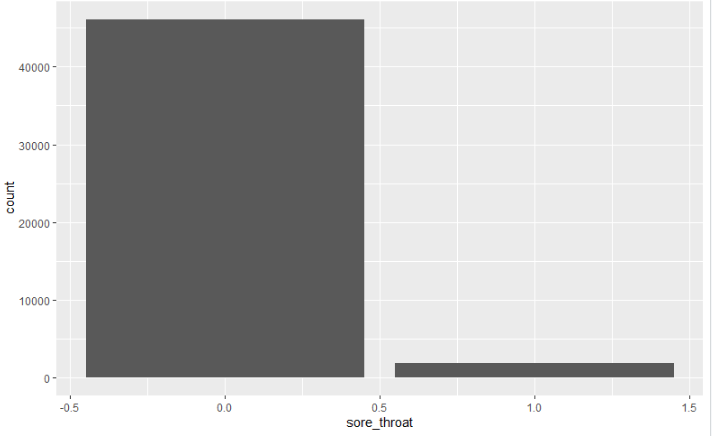


Figure 18: Comparison between sore throat and corona result

Figure 18 showed the patient who having symptom like sore throat are not necessary to get infected with this corona virus.

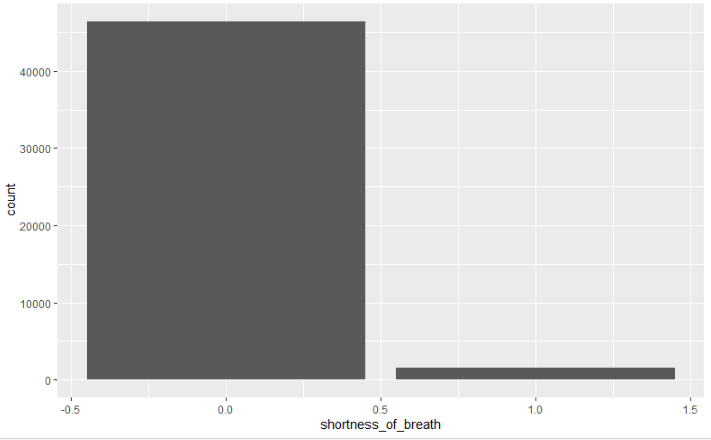


Figure 19: Comparison between shortness of breath and corona result

Figure 19 showed the patient who having symptom like shortness of breath are not necessary to get infected with this corona virus.

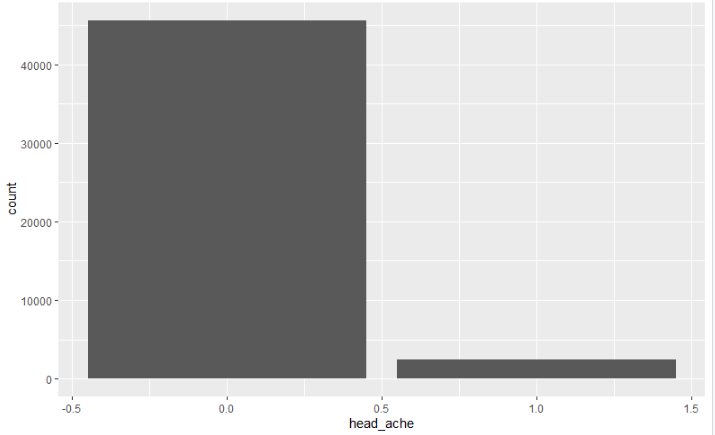


Figure 20: Comparison between headache and corona result

Figure 20 showed the patient who having symptom like headache are not necessary to get infected with this corona virus.

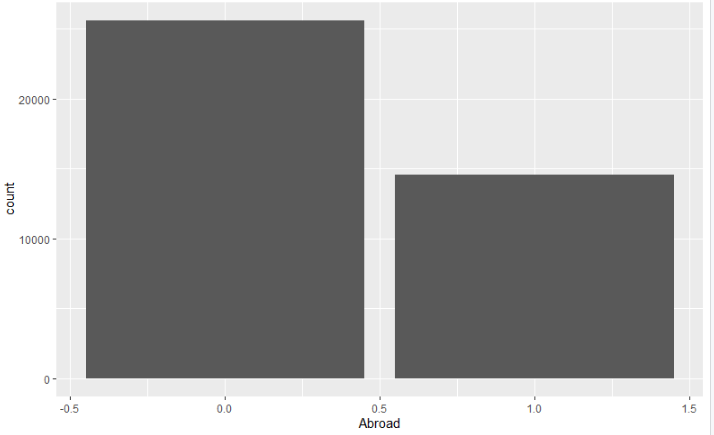


Figure 21: Comparison between abroad and corona result

Figure 21 showed the patient who having recent activity like abroad are not necessary to get infected with this corona virus.

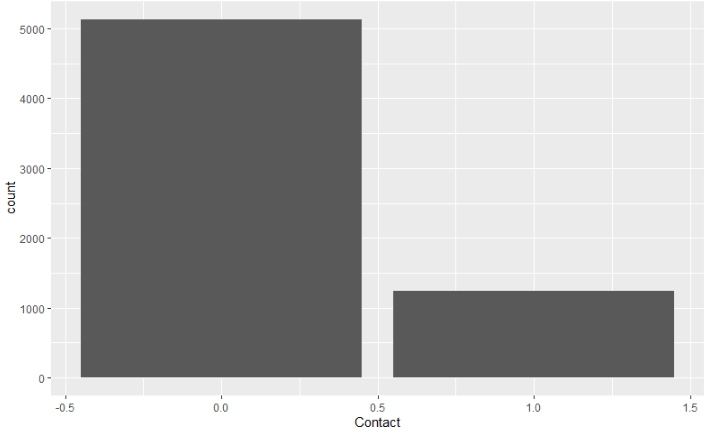


Figure 22: Comparison between contact and corona result

Figure 22 showed the patient who having recent activity like close contact with patient are not necessary to get infected with this corona virus.

## Data Modelling stage

For this data modelling stage, this project will be using two different algorithms as the prediction algorithm which are Naive Bayes Classification and Decision Tree Algorithm. After that, this project will compare the results for both algorithms and select the best algorithm.

### Naïve Bayes Classification

For this naïve bayes classification, there are two different methods to apply in R language. The first one is without any splitting data where just use the naïve bayes function from the library and train the data. And the second one required to split the data into two which representing train data and test data. After the model training the train data will use the test data to do prediction accuracy testing.

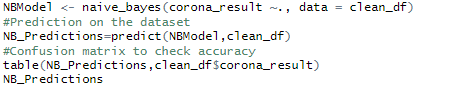


Figure 23: Naïve Bayes Function (Without splitting data)

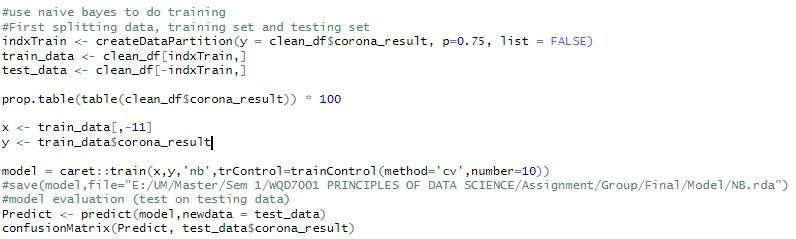


Figure 24: Naïve Bayes Function (With splitting data)

### Decision Tree Algorithm

For decision tree algorithm, the data will also split into two part which are train data and test data. After that, the train data will be used to training with this decision tree algorithm. After the training process is done, this project will be testing the prediction accuracy using the test data.

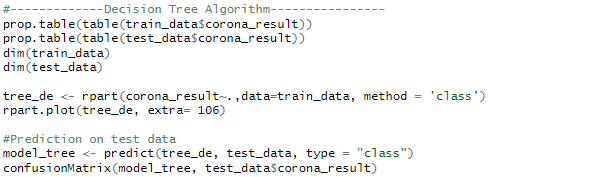


Figure 25: Decision tree algorithm (With splitting data)

## Analysis Results Stage

After all the algorithms being trained and tested, this project comparing those three algorithms methods’ results which are naïve bayes function without splitting data, naïve bayes function with splitting data and decision tree algorithm with splitting data. This project will use the confusion matrix to determine the accuracy for the algorithms.



Figure 26: Confusion Matrix for naïve bayes function without splitting data

Based on figure 26, we can calculate the accuracy of this naïve bayes function without splitting data are 0.8475.

For the naïve bayes function with splitting data and decision tree algorithm with splitting data, we put 0.75 as the splitting ratio.

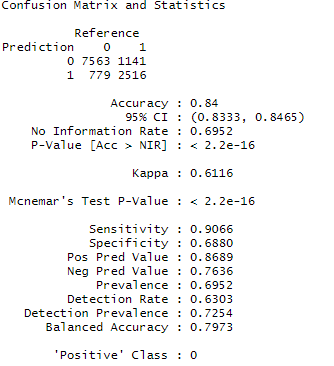


Figure 27: Results of naïve bayes function with splitting data

As from figure 27, we can clearly see that the accuracy of naïve bayes function with splitting data is also around 0.84.

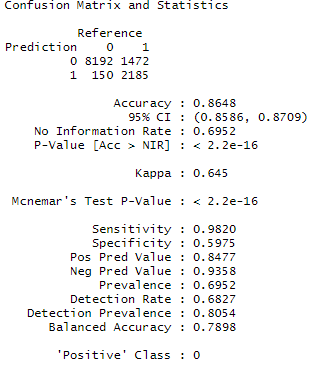


Figure 28: Results of decision tree algorithm with splitting data

As from figure 27, the prediction accuracy result of decision tree algorithm with splitting data is also around 0.8648 which are much higher than the naïve bayes algorithms method.

## Conclusion and limitation

After we compared all the results for each algorithm, we conclude that decision tree algorithm with splitting data has higher prediction accuracy compare to naïve bayes algorithm. However, there are some limitation that need to be mentioned. Because of the limitation of data sources, this dataset cannot represent for all the possibility of infecting this corona virus. Besides, based on figure 29, we can clearly see that this dataset has a lot of unbiased variable which might cause the results to become unbiased. Therefore, in future, we hope we can get more different dataset to enlarge the sample size.

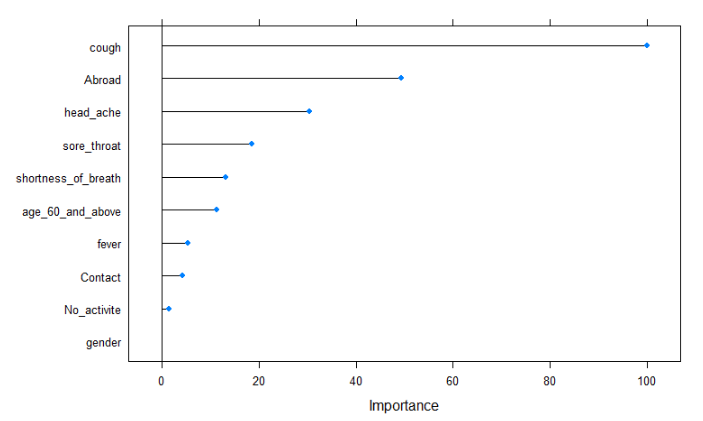


Figure 29: Variable performance for each category