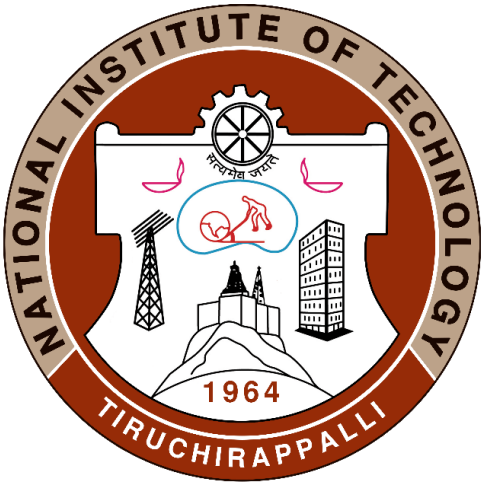
NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI



DEPARTMENT OF COMPUTER APPLICATIONS

COVID-19 DATA ANALYSIS AND VISUALIZATION

PROJECT WORK

submitted by

AMIT GUPTA
ROLL NO:- 205119010

Under the guidance of

Dr. CHITRA BASKAR

Submitted in fulfillment of the project in DATA MINING.

NATIONAL INSTITUTE OF TECHNOLOGY TIRUCHIRAPPALLI

CERTIFICATE

This is certify that *AMIT GUPTA* roll no:- *205119010*, Student of THIRD semester MCA (batch 2019-2022) of NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI has successfully completed the project *COVID-19 ANALYSIS & VISUALIZATION* under The guidance of *Dr. Chitra Baskar.*

Signature

(Dr. Chitra Baskar)

DATA MINING PROJECT

Covid-19 Data Analysis & Visualization

Submitted By

AMIT GUPTA
ROLL NO:-205119010

CONTACT & MAIL:-

guptaamit8602@gmail.com
205119010@nitt.edu

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION TO COVID-19

On 31st December 2019, in the city of Wuhan (CHINA), a cluster of cases of pneumonia of unknown cause was reported to World Health organisation. In January 2020, a previously unknown new virus was identified, subsequently named 2019 novel corona virus. WHO has declared the COVID-19 as a pandemic. A pandemic is defined as disease spread over a wide range of geographical area and that has affected high proportion of population.

CHAPTER NO. 2

PROJECT DESCRIPTION & ANALYSIS

The pandemic has already taken grip over people's life. Since the start of the pandemic, some countries are facing problem of ever-increasing cases. Through the data analysis of cases one can analyse how countries all over the world are doing in the term of controlling the pandemic. Analysing data leads to adapt the prevention model of the countries that are doing great in terms of lowering the graph. Predictions are made with the dataset available to the individual/country/organisations, thus helping them to decide how far they are able to control the pandemic or up to how much extent they should guide preventive measures. Through this project, a step towards helping people to understand the spread and predict the cases in their country is done. This project also gives an insight of how a country is doing in terms of limiting the speed.

The aim of the project is to provide data analysis and visualisation of covid-19 (a pandemic started in December -19). Through plotting of data, various cases have been studied like most affected countries due to this pandemic. Study of data from various countries is combined to show the growth of cases and recovery graph. In this project, the predictions on various cases has been done and finally, the different algorithms are performed using weka for decision making purposes. Comparison graphs has also been plotted to analyse how much person is getting affected/recovered in certain time intervals.

CHAPTER NO. 3

TECHNOLOGY AND CONCEPTS

3.1 For Data Visualisation :-

1. **Basic Python:-**

   Requires basic knowledge about Python DS.

2. **Basic Python and Machine learning Modules:-**

   a). **Pandas** :- pandas is a software library written for data manipulation and analysis. It offers data structures and operations for manipulating numeric tables and time series.

   b). **Plotly** :- Plotly is a python graphics library which makes interactive publication –quality graphs.

   c). **Matplotlib** :- Matplotlib is a comprehension library for creating static, animated and interactive visualizations in python.

3.2 For Data Analysis :-

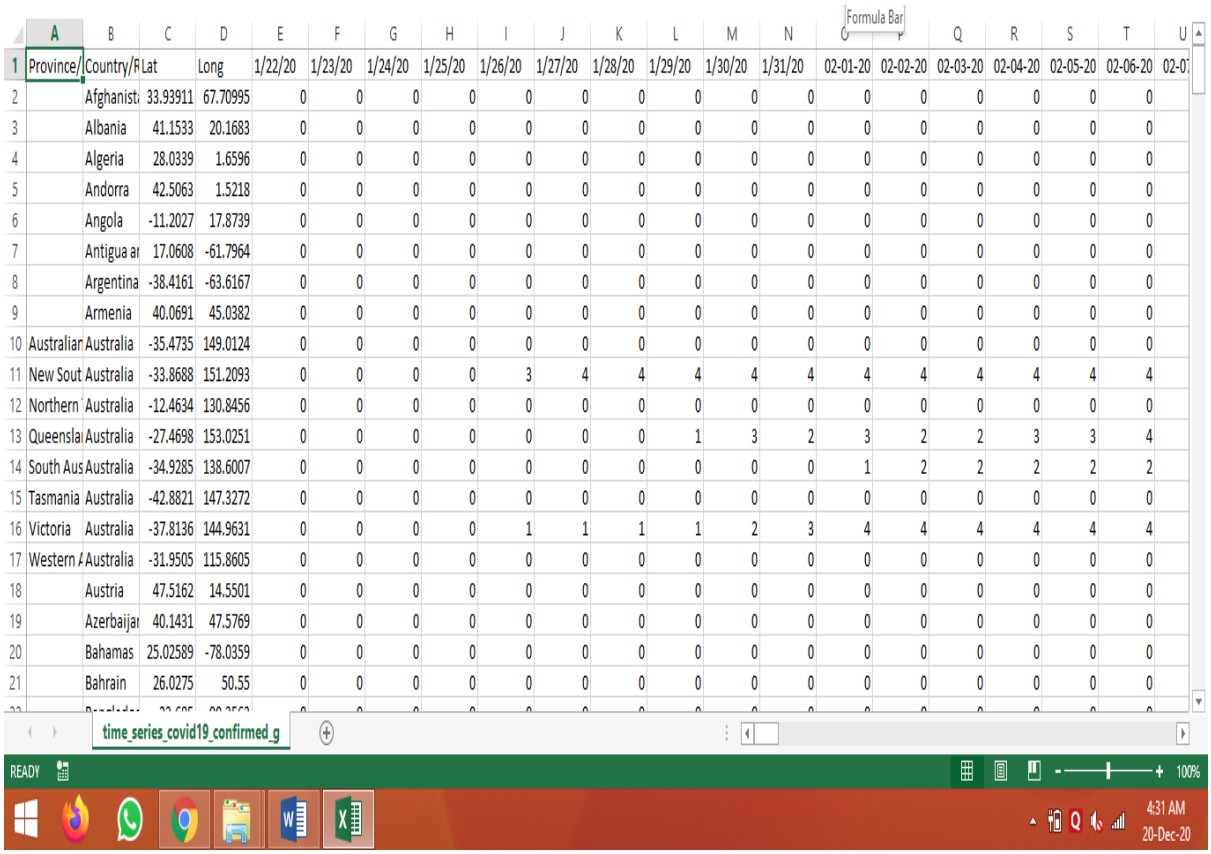
WEKA :-

Weka is a collection of machine learning algorithms for Data mining tasks. The algorithm can directly we applied to a dataset or called from your own java code. Weka contains tool for data preprocessing, classification, regression, clustering, association rules and visualization.

CHAPTER NO. 4

DATASET USED

In this project I am using the dataset "Covid-19_ Cases.csv". the format of our dataset is .csv.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	Q	R	S	T	U
1	Province/	Country/R	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	02-01-20	02-02-20	02-03-20	02-04-20	02-05-20	02-06-20	02-07
2		Afghanist	33.93911	67.70995	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
3		Albania	41.1533	20.1683	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
4		Algeria	28.0339	1.6596	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
5		Andorra	42.5063	1.5218	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
6		Angola	-11.2027	17.8739	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
7		Antigua ar	17.0608	-61.7964	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
8		Argentina	-38.4161	-63.6167	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
9		Armenia	40.0691	45.0382	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
10	Australian	Australia	-35.4735	149.0124	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
11	New Sout	Australia	-33.8688	151.2093	0	0	0	0	3	4	4	4	4	4	4	4	4	4	4	4	
12	Northern	Australia	-12.4634	130.8456	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
13	Queensla	Australia	-27.4698	153.0251	0	0	0	0	0	0	0	1	3	2	3	2	2	3	3	4	
14	South Aus	Australia	-34.9285	138.6007	0	0	0	0	0	0	0	0	0	0	1	2	2	2	2	2	
15	Tasmania	Australia	-42.8821	147.3272	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	Victoria	Australia	-37.8136	144.9631	0	0	0	0	1	1	1	1	2	3	4	4	4	4	4	4	
17	Western A	Australia	-31.9505	115.8605	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18		Austria	47.5162	14.5501	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
19		Azerbaijar	40.1431	47.5769	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20		Bahamas	25.02589	-78.0359	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
21		Bahrain	26.0275	50.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

This dataset basically includes the province, Contry, Latitude, Longitude columns with all the confirmed cases from the date 1/22/20 to 12/16/20 with all over the countries across the globe.

This dataset basically includes the 1000 rows and 300 columns. It is a huge dataset in which we are working with huge dataset processing.

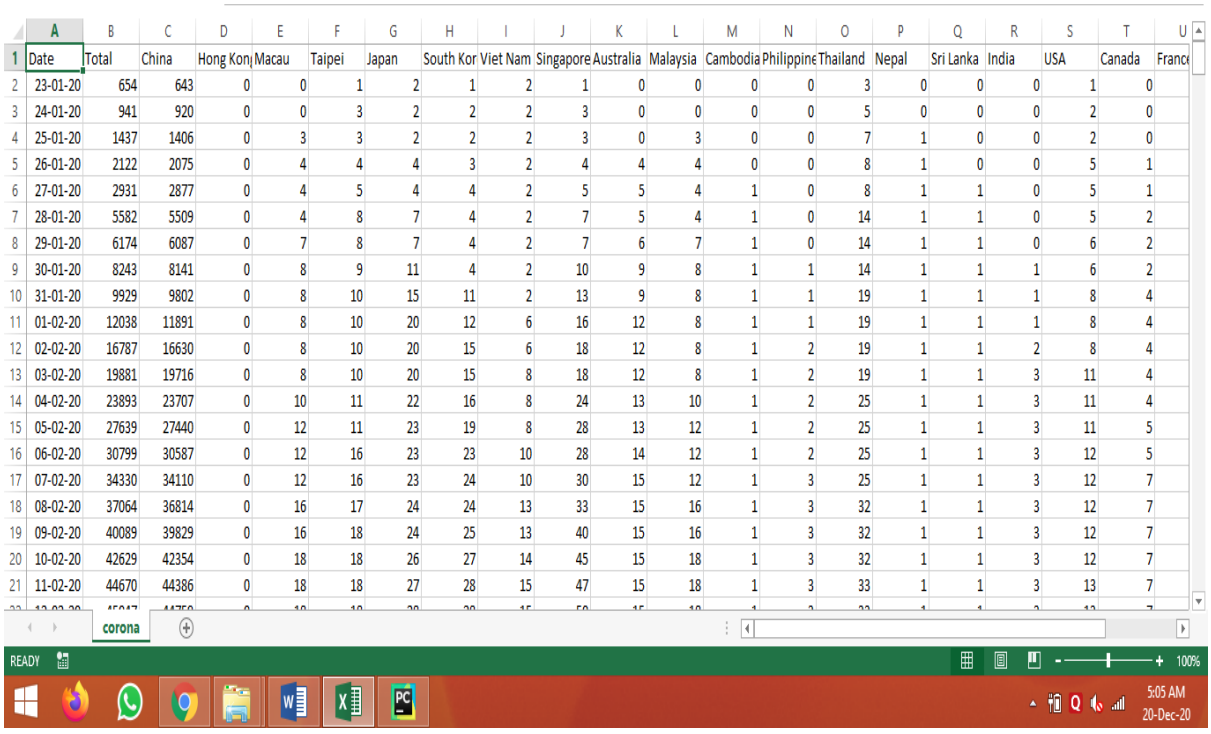
CHAPTER NO. 5

DATASET PRE-PROCESSING

There are a lot of datasets are present on the internet regarding the Covid-19. It is hard to find the dataset which fulfills our need sometime it may happen that the dataset contains unwanted rows and cols which are not nessecary for our analysis. In the dataset preprocessing we deals with the proper formatting of datasets according to our needs.

```
import pandas as pd
data=pd.read_csv('/content/time_series_covid19_confir
med_global.csv')
data=data.groupby('Country/Region').sum()
data=data.drop(columns=['Lat','Long'])
data_transposed=data.T
```

After preprocessing Our dataset will look like:-



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Date	Total	China	Hong Kon	Macau	Taipei	Japan	South Kor	Viet Nam	Singapore	Australia	Malaysia	Cambodia	Philippine	Thailand	Nepal	Sri Lanka	India	USA	Canada	France
2	23-01-20	654	643	0	0	1	2	1	2	1	0	0	0	0	3	0	0	0	1	0	
3	24-01-20	941	920	0	0	3	2	2	2	3	0	0	0	0	5	0	0	0	2	0	
4	25-01-20	1437	1406	0	3	3	2	2	2	3	0	3	0	0	7	1	0	0	2	0	
5	26-01-20	2122	2075	0	4	4	4	3	2	4	4	4	0	0	8	1	0	0	5	1	
6	27-01-20	2931	2877	0	4	5	4	4	2	5	5	4	1	0	8	1	1	0	5	1	
7	28-01-20	5582	5509	0	4	8	7	4	2	7	5	4	1	0	14	1	1	0	5	2	
8	29-01-20	6174	6087	0	7	8	7	4	2	7	6	7	1	0	14	1	1	0	6	2	
9	30-01-20	8243	8141	0	8	9	11	4	2	10	9	8	1	1	14	1	1	1	6	2	
10	31-01-20	9929	9802	0	8	10	15	11	2	13	9	8	1	1	19	1	1	1	8	4	
11	01-02-20	12038	11891	0	8	10	20	12	6	16	12	8	1	1	19	1	1	1	8	4	
12	02-02-20	16787	16630	0	8	10	20	15	6	18	12	8	1	2	19	1	1	2	8	4	
13	03-02-20	19881	19716	0	8	10	20	15	8	18	12	8	1	2	19	1	1	3	11	4	
14	04-02-20	23893	23707	0	10	11	22	16	8	24	13	10	1	2	25	1	1	3	11	4	
15	05-02-20	27639	27440	0	12	11	23	19	8	28	13	12	1	2	25	1	1	3	11	5	
16	06-02-20	30799	30587	0	12	16	23	23	10	28	14	12	1	2	25	1	1	3	12	5	
17	07-02-20	34330	34110	0	12	16	23	24	10	30	15	12	1	3	25	1	1	3	12	7	
18	08-02-20	37064	36814	0	16	17	24	24	13	33	15	16	1	3	32	1	1	3	12	7	
19	09-02-20	40089	39829	0	16	18	24	25	13	40	15	16	1	3	32	1	1	3	12	7	
20	10-02-20	42629	42354	0	18	18	26	27	14	45	15	18	1	3	32	1	1	3	12	7	
21	11-02-20	44670	44386	0	18	18	27	28	15	47	15	18	1	3	33	1	1	3	13	7	

CHAPTER NO. 6

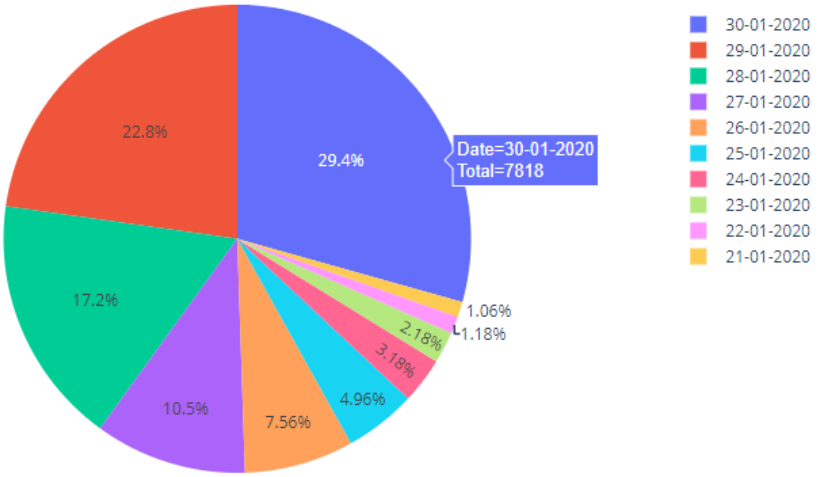
COVID-19 DATA VISUALISATIONS

6.1 PIE CHART REPRESENTATION:-

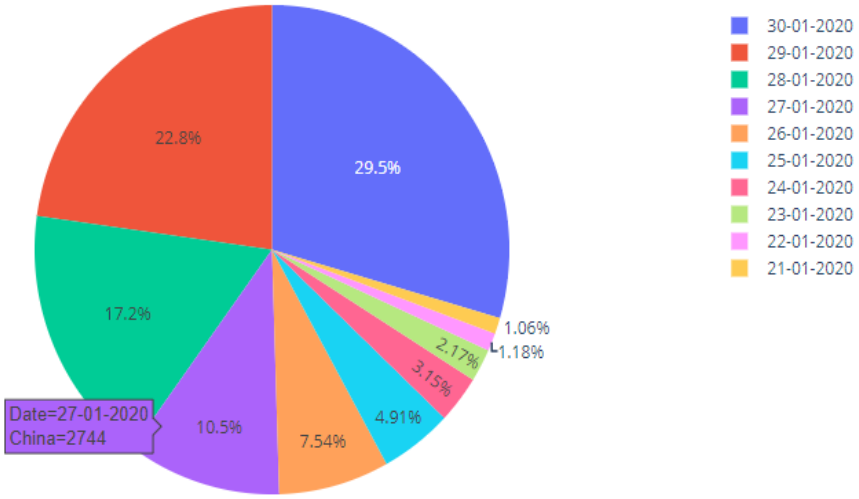
Basically representing the Infected count of all the countries and total in the form of PIE chart of a range of date.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import plotly.express as px
import os
for dirname,_,filenames in os.walk('/content'):
    for filename in filenames:
        print(os.path.join(dirname,filename))

df=pd.read_csv('/content/corona_virus.csv')

countries=df.drop('Date',axis=1)
def plot_growth(dates,country):
    plt.rcParams['figure.figsize']=(20,10)
    plt.plot_date(dates,country)
    plt.xlabel("Date",fontsize=18)
    plt.ylabel("Infected people",fontsize=16)

    for country in countries.columns:
        plot_growth(df['Date'],df[country])

for country in countries.columns:

fig=px.pie(df,values=country,names='Date',title='Coun
t Infected of '+country)
    fig.show()
```

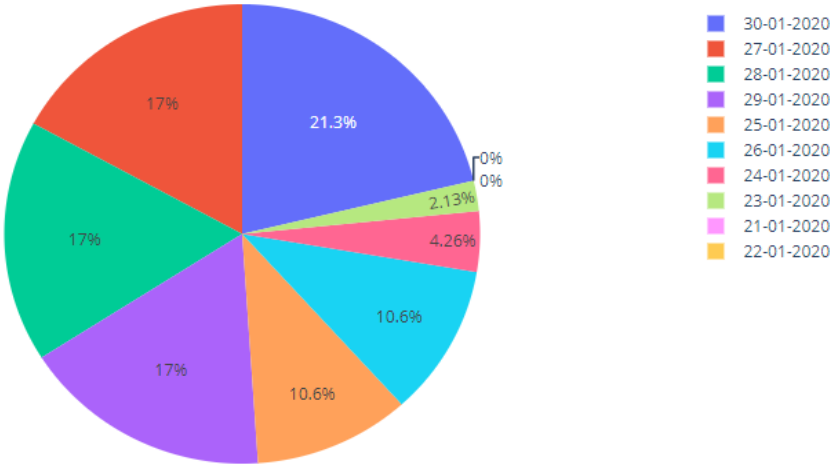
Count Infected of Total



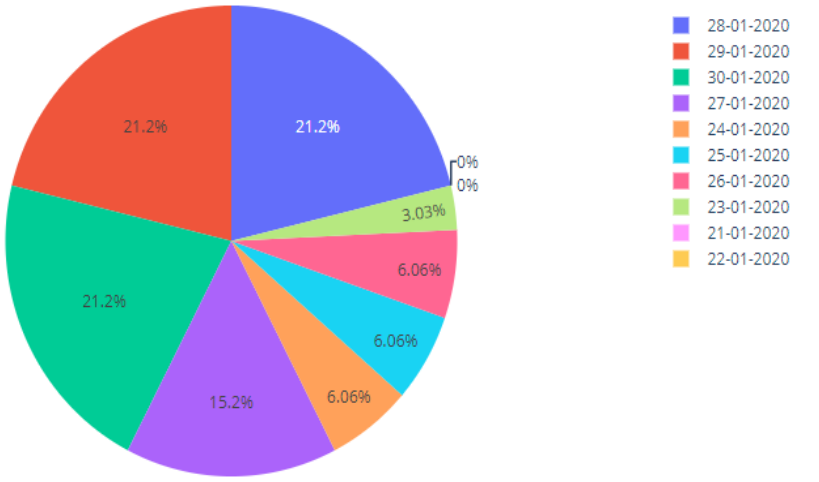
Count Infected of China



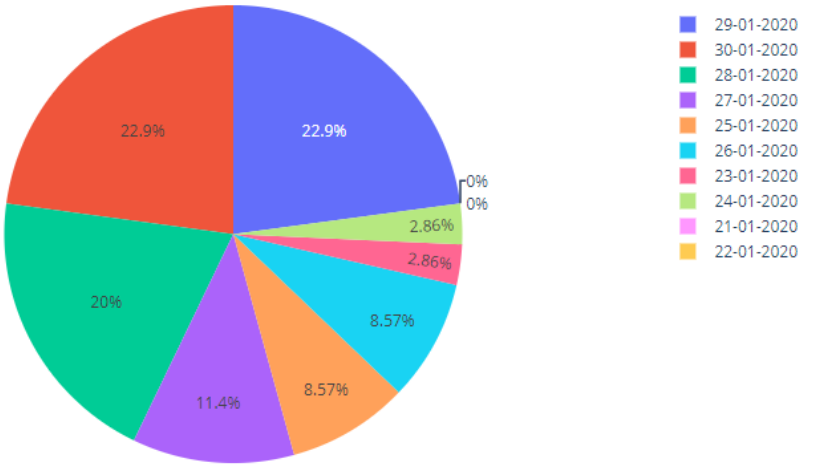
Count Infected of Hong Kong



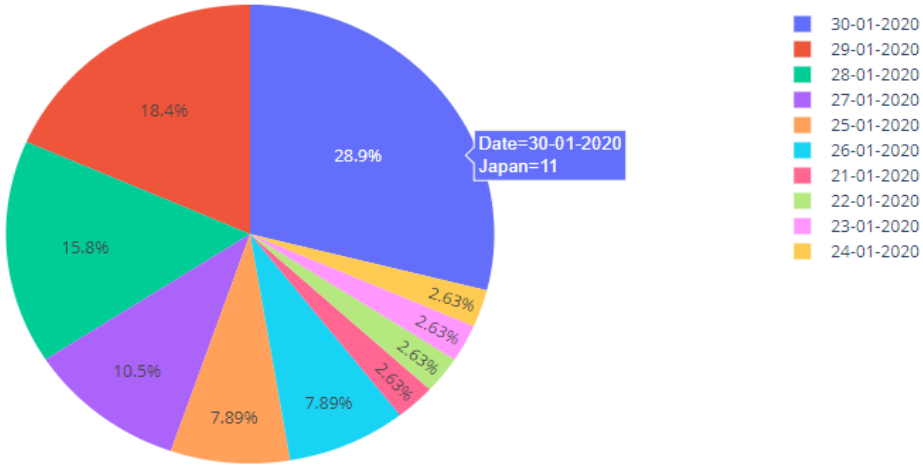
Count Infected of Macau



Count Infected of Taipei



Count Infected of Japan



6.2 RUNNING BAR GRAPH :-

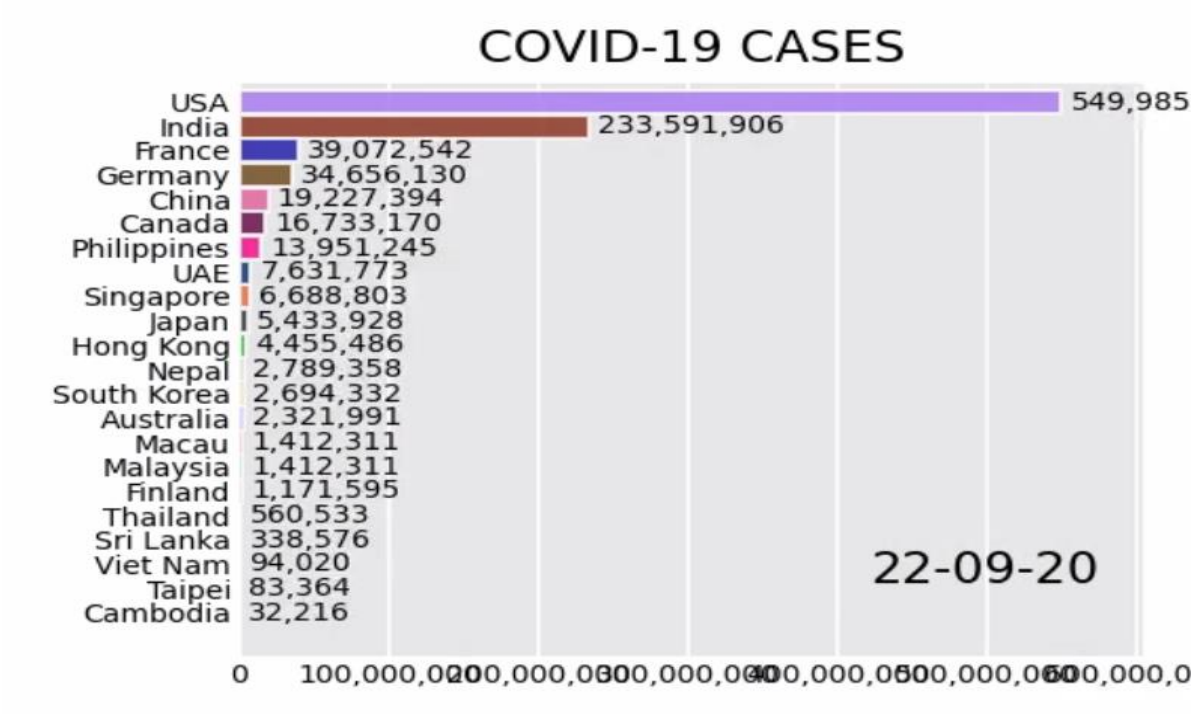
Basically representing the Infected count of all the countries and total count in the form of Running bar graph chart of a January month.

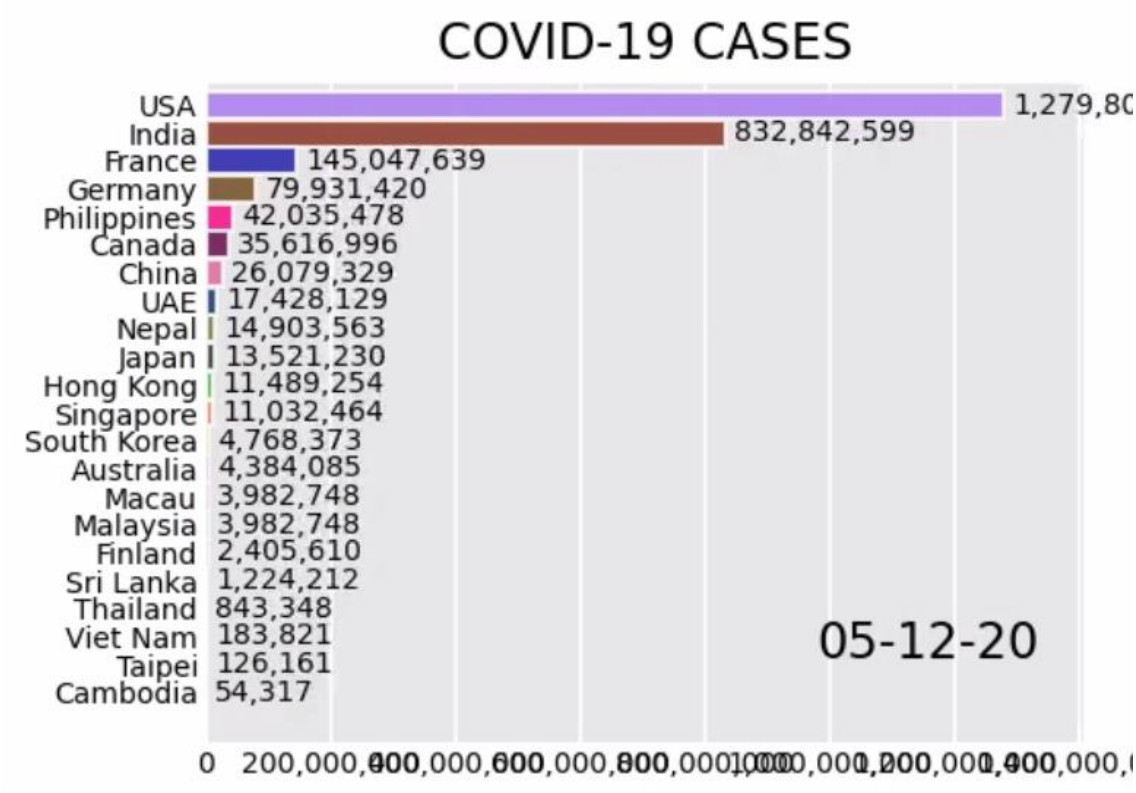
```
import pandas as pd
import os
os.chdir("/content")
data_corona=pd.read_csv("/content/corona_virus.csv")
data_corona.head()
cols=['Date','Total','China','Hong
Kong','Macau','Taipei','Japan','South Korea','Viet
Nam','Singapore','Australia','Malaysia','Cambodia','P
hilippines','Thailand','Nepal','Sri
Lanka','India','USA','Canada','France','Finland','Ger
many','UAE']
subsetdf=data_corona[cols]
subsetdf.set_index("Date",inplace=True)
cum_sum_df=subsetdf.cumsum(axis=0)
cum_sum_df.tail(10)
import bar_chart_race as bcr
bcr.bar_chart_race(df=cum_sum_df,filename=None,figsiz
e=(3.5,3),title='COVID-19 CASES')
```

RUNNING BAR GRAPH VIDEO:-

If the above video does not get played, then go to my drive link and you can watch it.

https://drive.google.com/file/d/1_Vi1UzARvrdm3VQcy6u58C2yf0drk4Q6/view?usp=sharing


COVID-19 CASES
USA 549,985
India 233,591,906
France 39,072,542
Germany 34,656,130
China 19,227,394
Canada 16,733,170
Philippines 13,951,245
UAE 7,631,773
Singapore 6,688,803
Japan 5,433,928
Hong Kong 4,455,486
Nepal 2,789,358
South Korea 2,694,332
Australia 2,321,991
Macau 1,412,311
Malaysia 1,412,311
Finland 1,171,595
Thailand 560,533
Sri Lanka 338,576
Viet Nam 94,020
Taipei 83,364
Cambodia 32,216
22-09-20
0
100,000,0
200,000,0
300,000,0
400,000,0
500,000,0
600,000,0


COVID-19 CASES
USA 1,279,8
India 832,842,599
France 145,047,639
Germany 79,931,420
Philippines 42,035,478
Canada 35,616,996
China 26,079,329
UAE 17,428,129
Nepal 14,903,563
Japan 13,521,230
Hong Kong 11,489,254
Singapore 11,032,464
South Korea 4,768,373
Australia 4,384,085
Macau 3,982,748
Malaysia 3,982,748
Finland 2,405,610
Sri Lanka 1,224,212
Thailand 843,348
Viet Nam 183,821
Taipei 126,161
Cambodia 54,317
05-12-20
0
200,000,
400,000,
600,000,
800,000,
1,000,000,
1,200,000,
1,400,000,

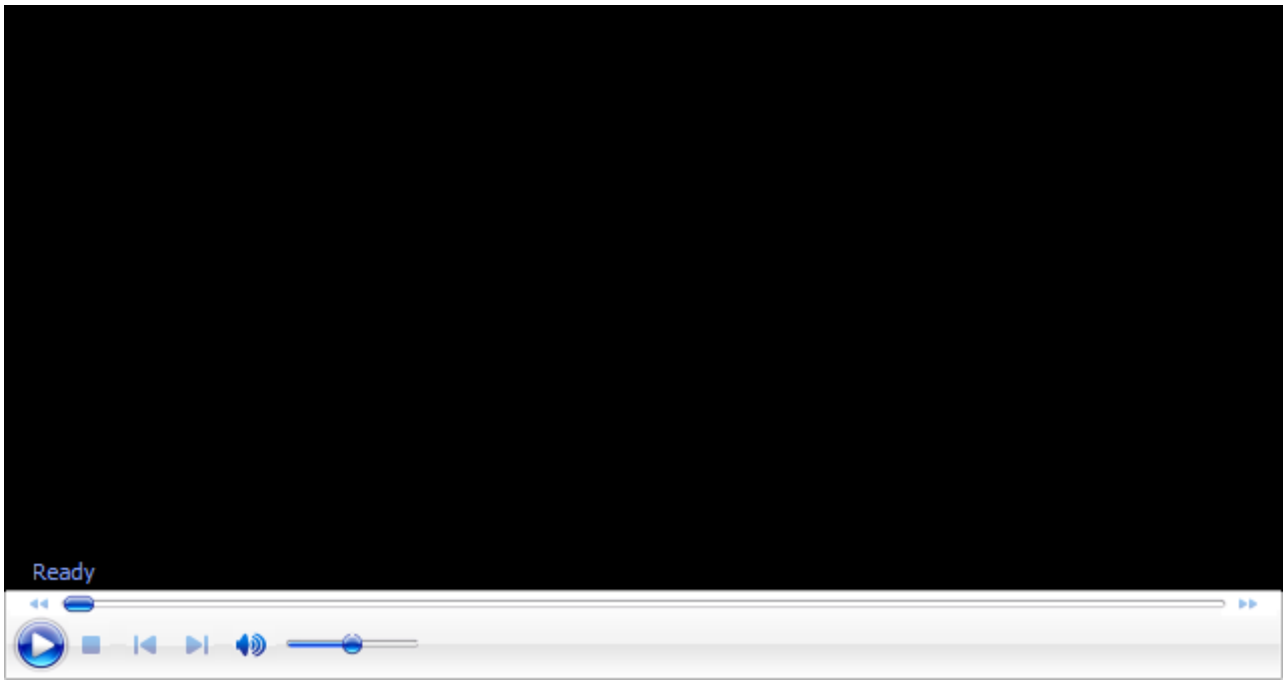
6.3 DYNAMIC MAPPING OF COVID 19 :-

Basically representing the Infected count of all the countries in the form of dynamic map.

```
import pandas as pd
import geopandas as gpd
import PIL
import io
data=pd.read_csv('/content/time_series_covid19_confir
med_global.csv')
data=data.groupby('Country/Region').sum()
data=data.drop(columns=['Lat','Long'])
data_transposed=data.T
#data_transposed.head(10)
#data_transposed.plot(y=['Australia','China','US','It
aly'],use_index=True,figsize=(8,8),marker='*')
world=gpd.read_file('/content/World_Map.shp')
world.replace('Viet Nam','Vietnam',inplace=True)
world.replace('Brunei
Darussalam','Brunei',inplace=True)
world.replace('Cape Verde','Cabo Verde',inplace=True)
world.replace('Democratic Republic of the
Congo','Congo (Kinshasa)',inplace=True)
world.replace('Congo','Congo
(Brazzaville)',inplace=True)
world.replace('Czech
Republic','Czechia',inplace=True)
world.replace('Iran (Islamic Republic
of)','Iran',inplace=True)
world.replace('Korea, Republic of','Korea,
South',inplace=True)
world.replace("Lao People's Democratic
Republic",'Laos',inplace=True)
world.replace('Libyan Arab
Jamahiriya','Libya',inplace=True)
world.replace('Republic of
Moldova','Moldova',inplace=True)
world.replace('The former Yugoslav Republic of
Macedonia','North Macedonia',inplace=True)
world.replace('Syrian Arab
```

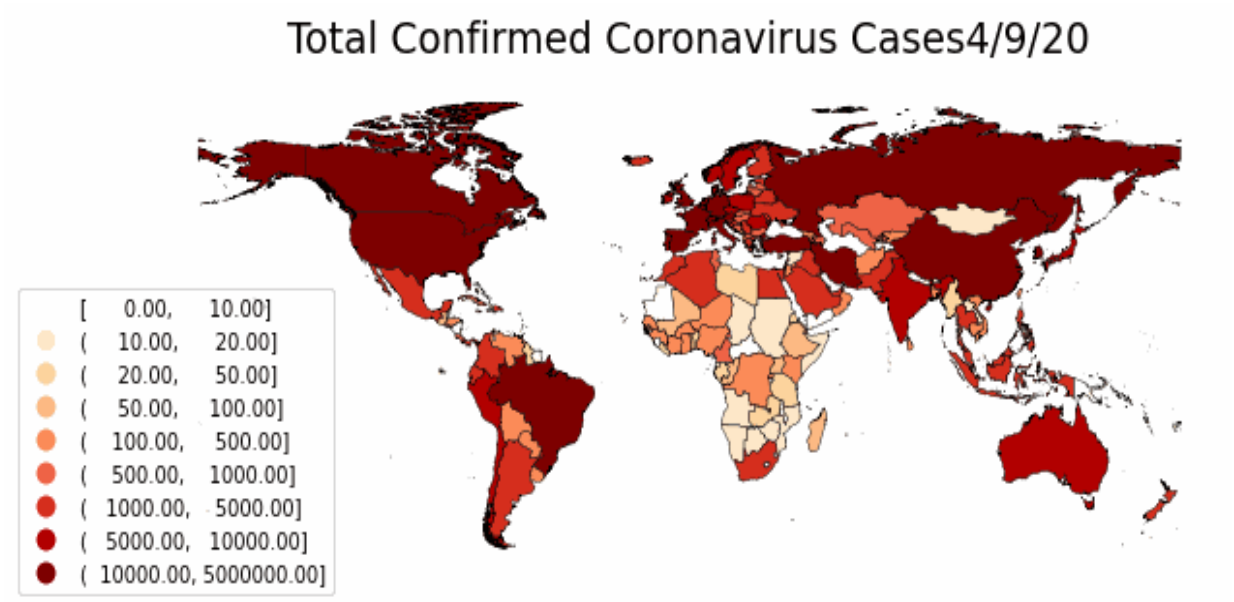
```
Republic','Syria',inplace=True)

world.replace('Taiwan','Taiwan*',inplace=True)
world.replace('United Republic of
Tanzania','Tanzania',inplace=True)
world.replace('United States','US',inplace=True)
world.replace('Palestine','West Bank and
Gaza',inplace=True)
merge=world.join(data,on='NAME',how='right')
image_frames=[]
for dates in merge.columns.to_list()[2:87]:

ax=merge.plot(column=dates,cmap='OrRd',figsize=(14,14
),legend=True,scheme='user_defined',

classification_kwds={'bins':[10,20,50,100,500,1000,50
00,10000,5000000]},
                  edgecolor='black',linewidth=0.4,)
    ax.set_title('Total Confirmed Coronavirus Cases'
+ dates,fontdict={'fontsize':20},pad=12.5)

    ax.set_axis_off()

    ax.get_legend().set_bbox_to_anchor((0.18,0.6))
    img=ax.get_figure()
    f=io.BytesIO()
    img.savefig(f,format='png',bbox_inches='tight')
    f.seek(0)
    image_frames.append(PIL.Image.open(f))
image_frames[0].save('/content/Dynamic Covid 19
map.gif',format='GIF',
                     append_images=image_frames[1:],
                     save_all=True,duration=300,
                     loop=3)

f.close()
```

VIDEO OF DYNAMIC COVID MAPPING :-



If the above video does not get played, then go to my drive link and you can watch it

https://drive.google.com/file/d/1C2K2LN8rQc9uJzQ_TzxVQoRg_ouLk1WK/view?usp=sharing

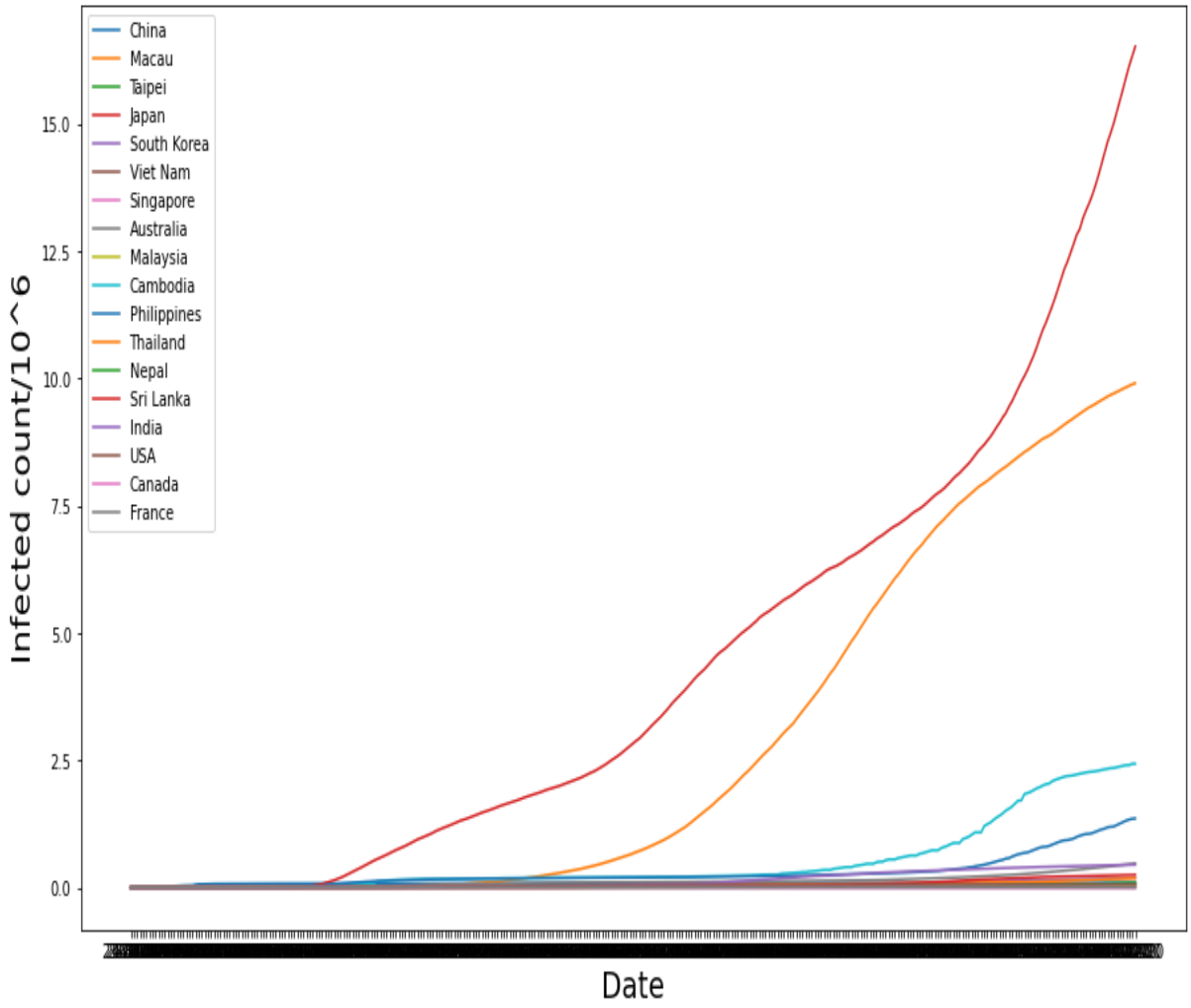


6.4 POINT GRAPH OF COVID-19 :-

Basically representing the Infected count of all the countries in the form of points on the XY plane where X axis denotes the dates and Y axis denotes the infected count.

```
import pandas as pd
from matplotlib import pyplot as plt
sample__data=pd.read_csv('/content/corona.csv')
plt.figure(figsize=(15,9))
plt.plot(sample__data.Date,sample__data.China/10**6)
plt.plot(sample__data.Date,sample__data.India/10**6)
#plt.plot(sample__data.Date,sample__data.Hong Kong)
plt.plot(sample__data.Date,sample__data.Macau/10**6)
plt.plot(sample__data.Date,sample__data.USA/10**6)
plt.plot(sample__data.Date,sample__data.UAE/10**6)
plt.plot(sample__data.Date,sample__data.Australia/10**6)
plt.plot(sample__data.Date,sample__data.Cambodia/10**6)
plt.plot(sample__data.Date,sample__data.Canada/10**6)
plt.plot(sample__data.Date,sample__data.Finland/10**6)
plt.plot(sample__data.Date,sample__data.France/10**6)
plt.plot(sample__data.Date,sample__data.Germany/10**6)
plt.plot(sample__data.Date,sample__data.Japan/10**6)
plt.plot(sample__data.Date,sample__data.Malaysia/10**6)
plt.plot(sample__data.Date,sample__data.Nepal/10**6)
plt.plot(sample__data.Date,sample__data.Philippines/10**6)
plt.plot(sample__data.Date,sample__data.Singapore/10**6)
plt.plot(sample__data.Date,sample__data.Taipei/10**6)
plt.plot(sample__data.Date,sample__data.Thailand/10**6)
plt.legend(['China','Macau','Taipei','Japan','South
Korea','Viet
Nam','Singapore','Australia','Malaysia','Cambodia','Philippine
s','Thailand','Nepal','Sri
Lanka','India','USA','Canada','France','Finland','Germany','UA
E'])
plt.xlabel('Date',size=20)
plt.ylabel('Infected count/10^6',size=20)
plt.show()
```

DIAGRAM OF POINT GRAPH:-



CHAPTER NO. 7

COVID-19 DATA ANALYSIS

7.1 SUPERVISED LEARNING ALGORITHMS:-

Supervised learning uses labeled training data to learn the mapping function that turns input variables (X) into the output variable (Y). In other words, it solves for *f* in the following equation:

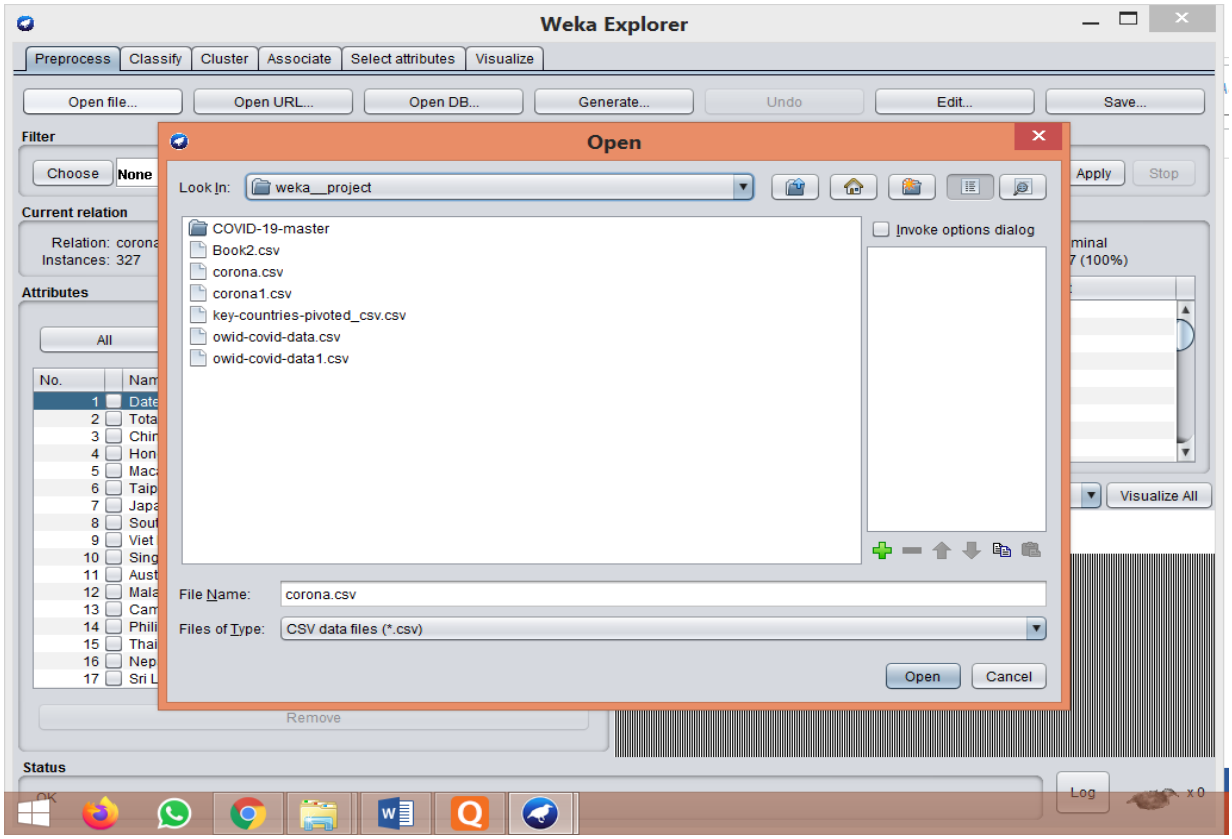
$$Y = f(X)$$

This allows us to accurately generate outputs when given new inputs.

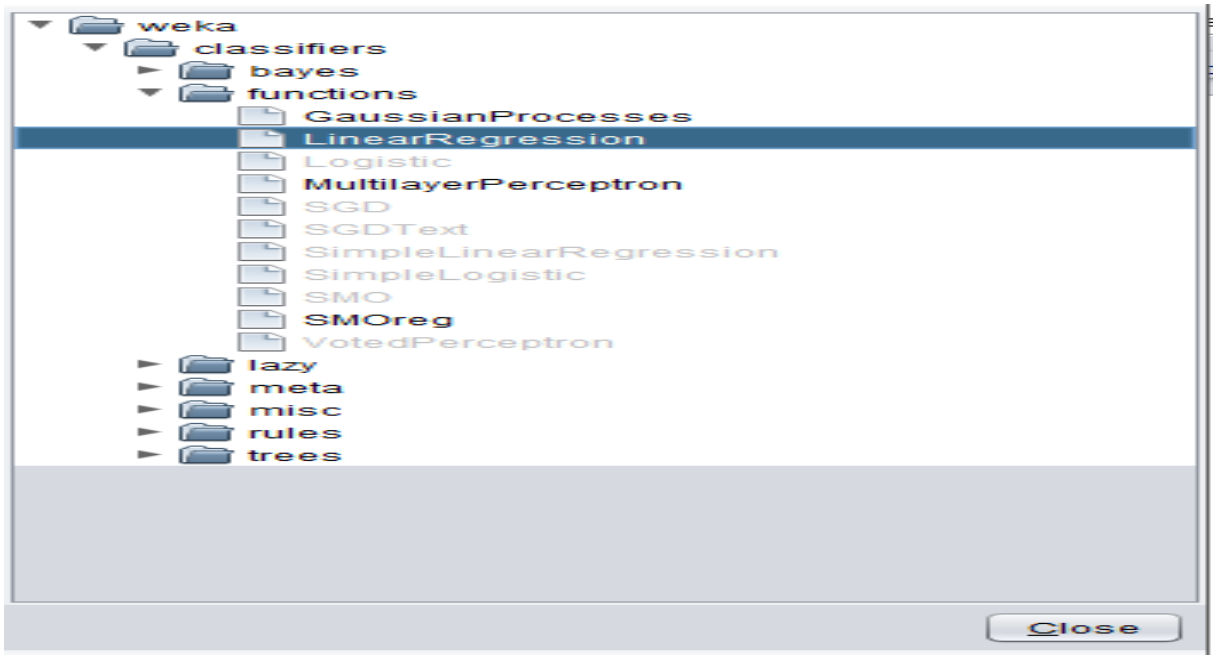
APPLIED ALGORITHMS:-

A) . LINEAR REGRESSION:-

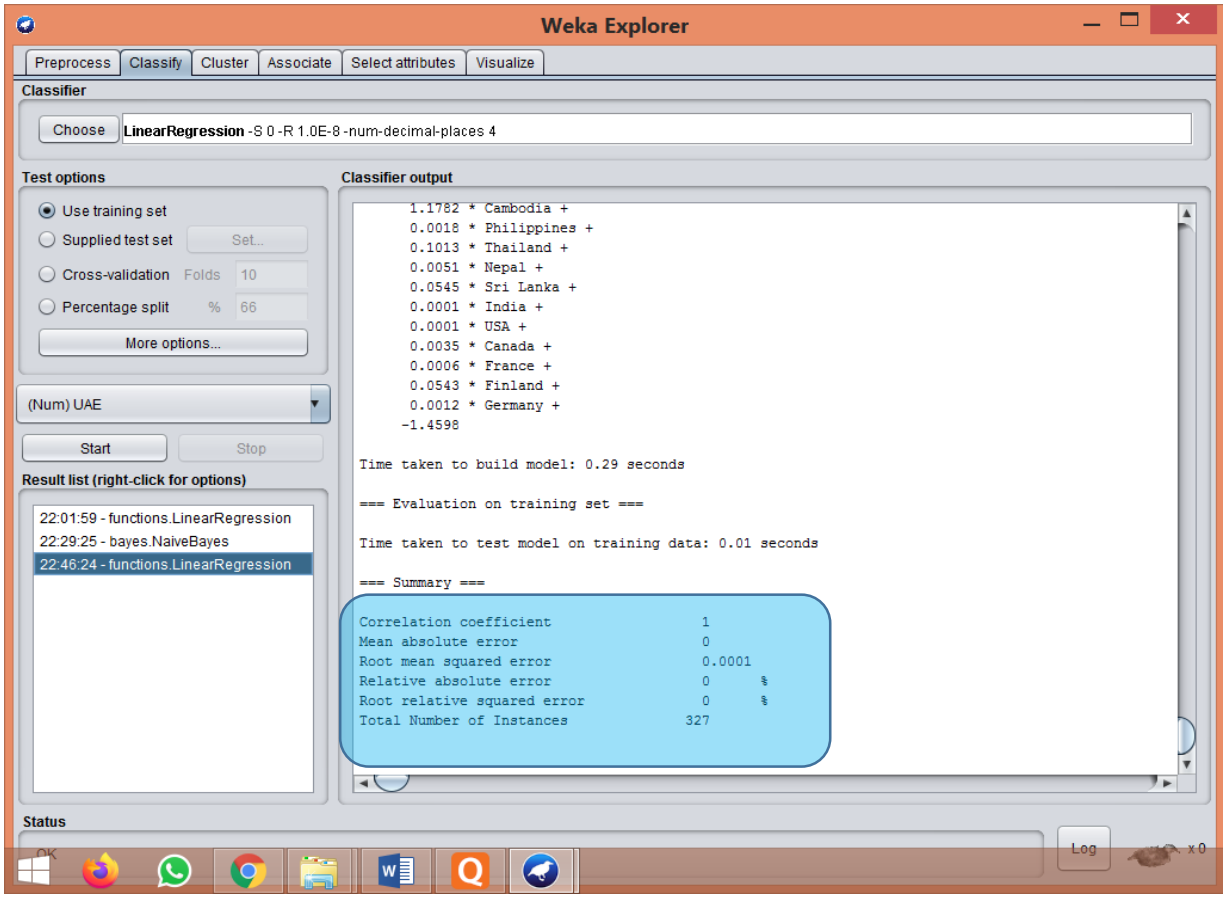
step 1:- Open Weka, Load dataset by selecting the **Open File** option from Preprocess Tab.



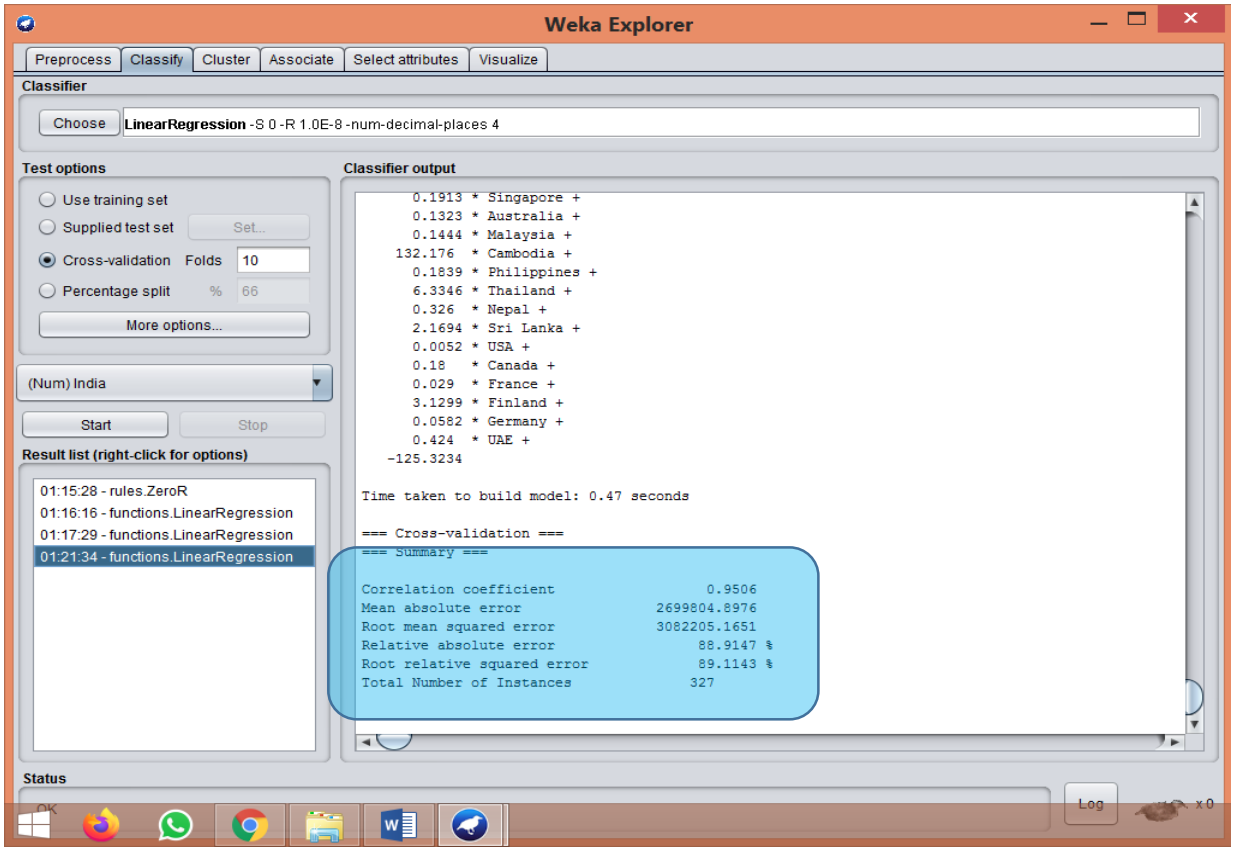
Step 2:- Click on the **Choose** option from **Classify** Tab. A new window will open select **Linear Regression** from **Function**.



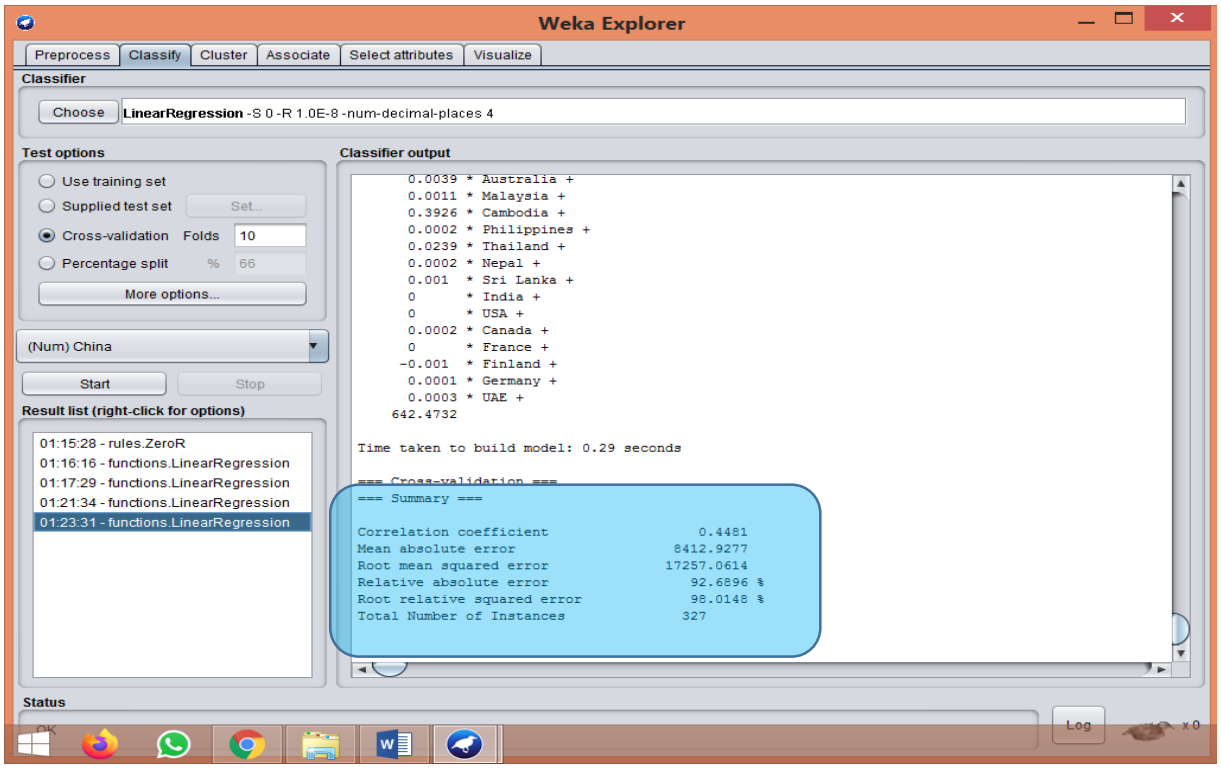
Step 3 :- Click on **Start**. (For UAE)



Step 4 :- Click on **Start**. (For INDIA)

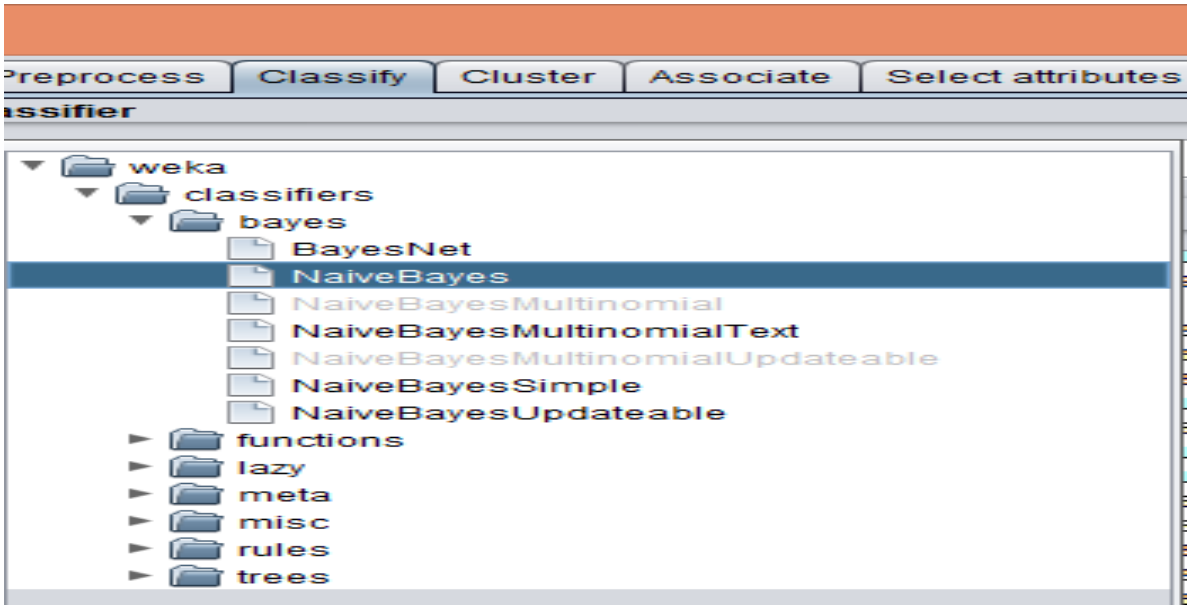


Step 4 :- Click on **Start**. (For CHINA)

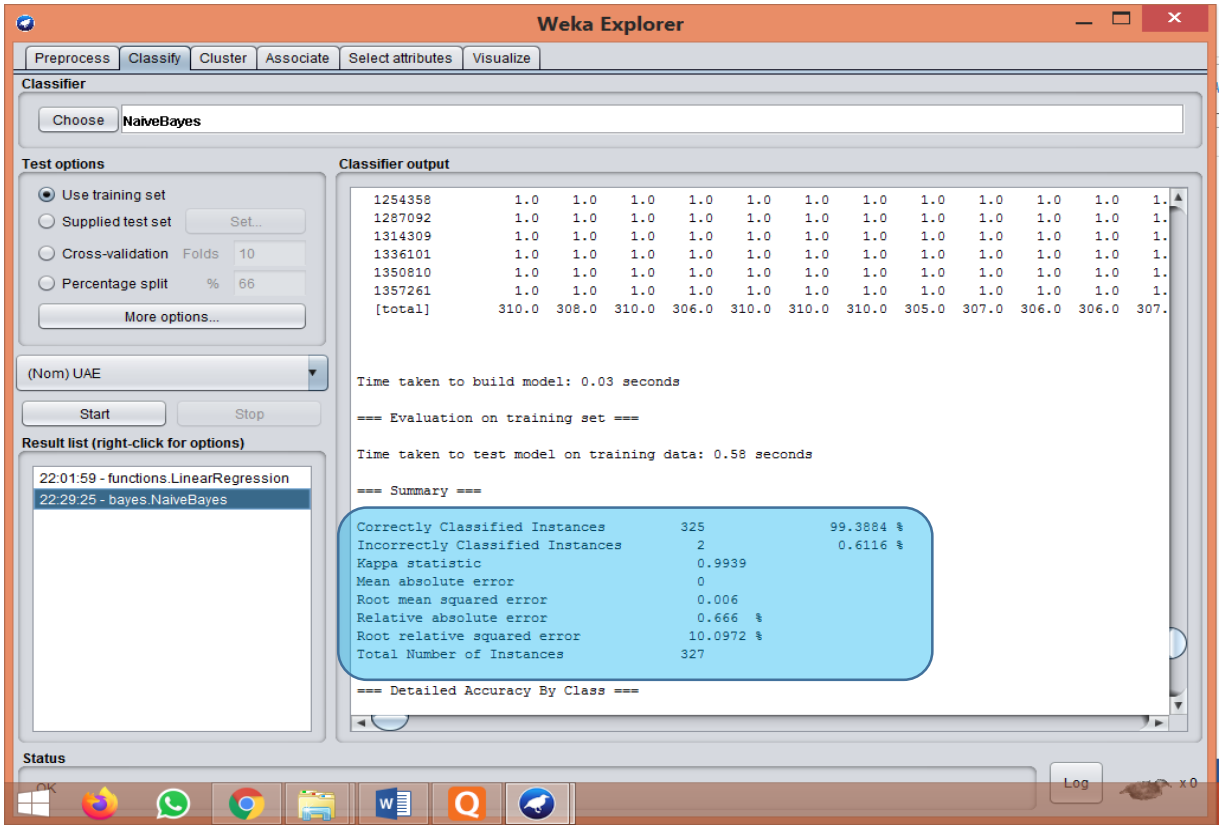


B). NAÏVE BAYES :-

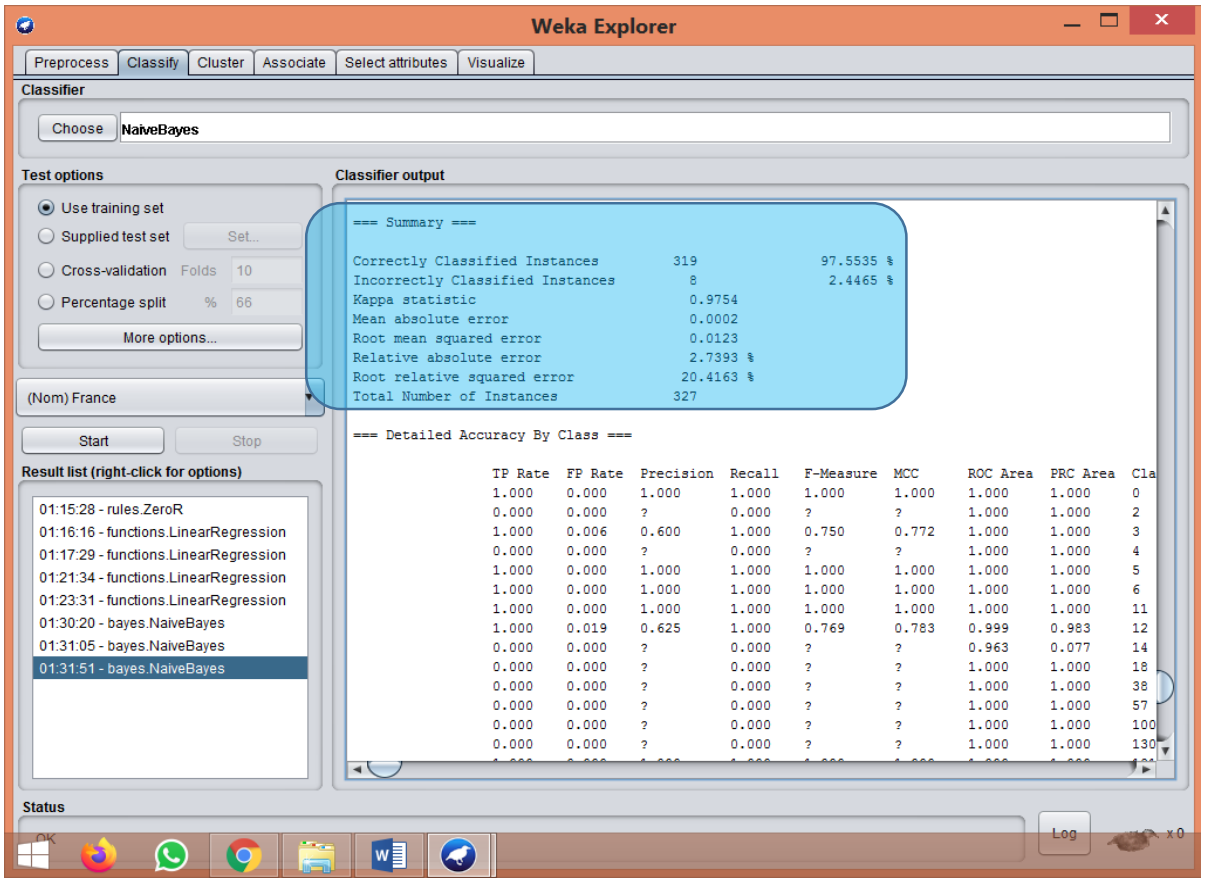
Step 1:- Click on the **Choose** option from **Classify** Tab. A new window will open select **NaveBayes** from **Bayes**.



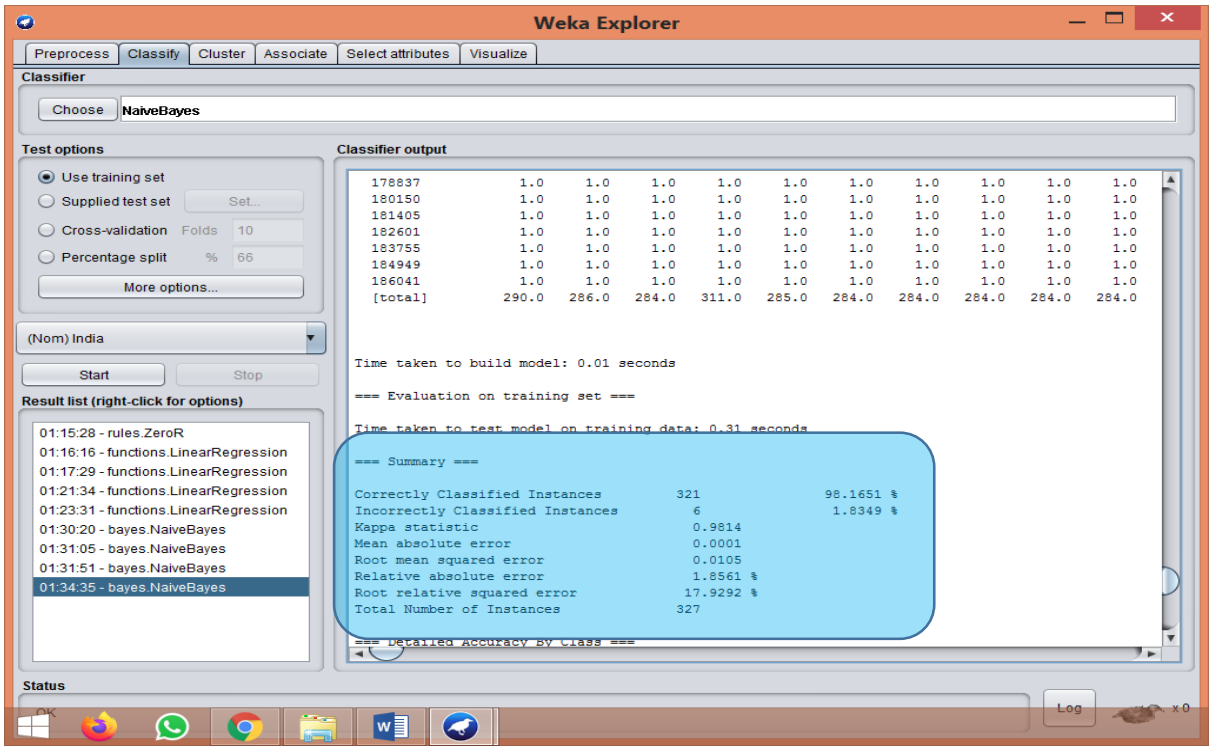
Step 2 :- Click on **Start**. (For UAE)



Step 3 :- Click on **Start**. (For FRANCE)

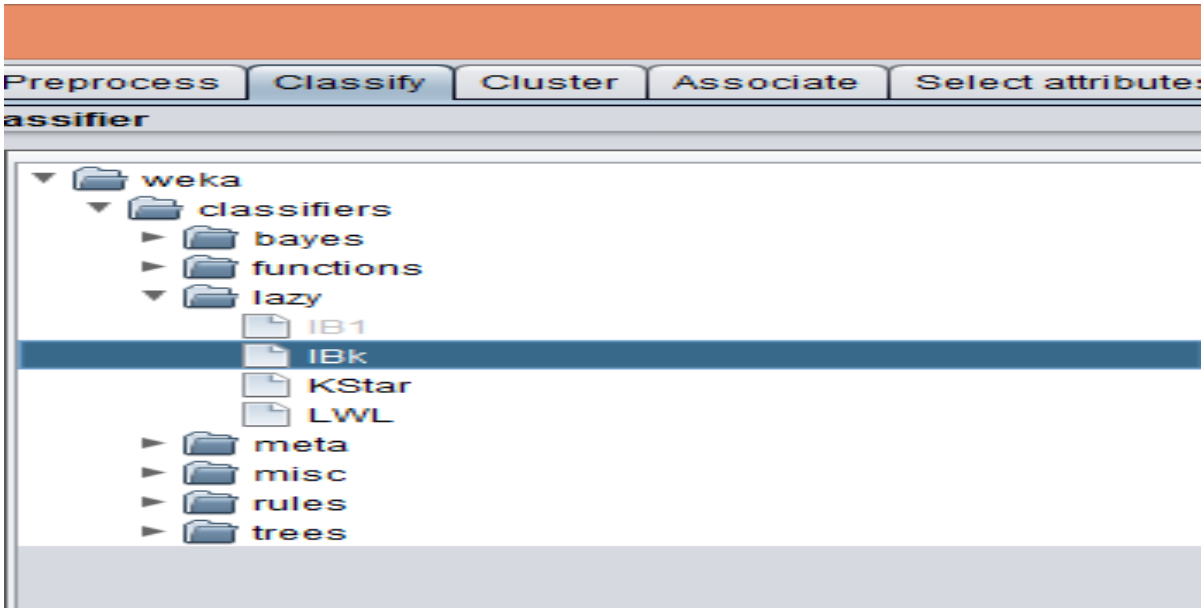


Step 4 :- Click on **Start**. (For INDIA)

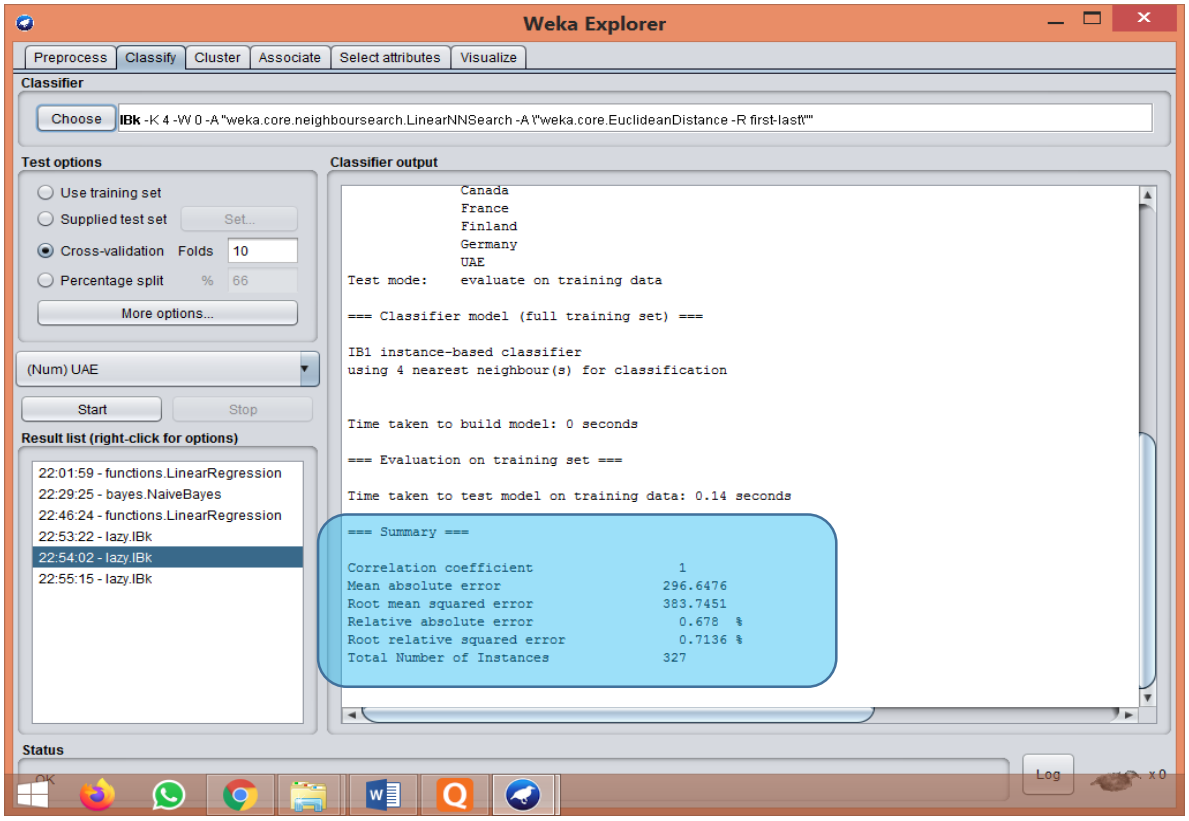


C). KNN (K NEAREST NEIGHBOURS) ALGORITHM :-

Step 1:- Click on the **Choose** option from **Classify** Tab. A new window will open select **IBK** from **Lazy**.



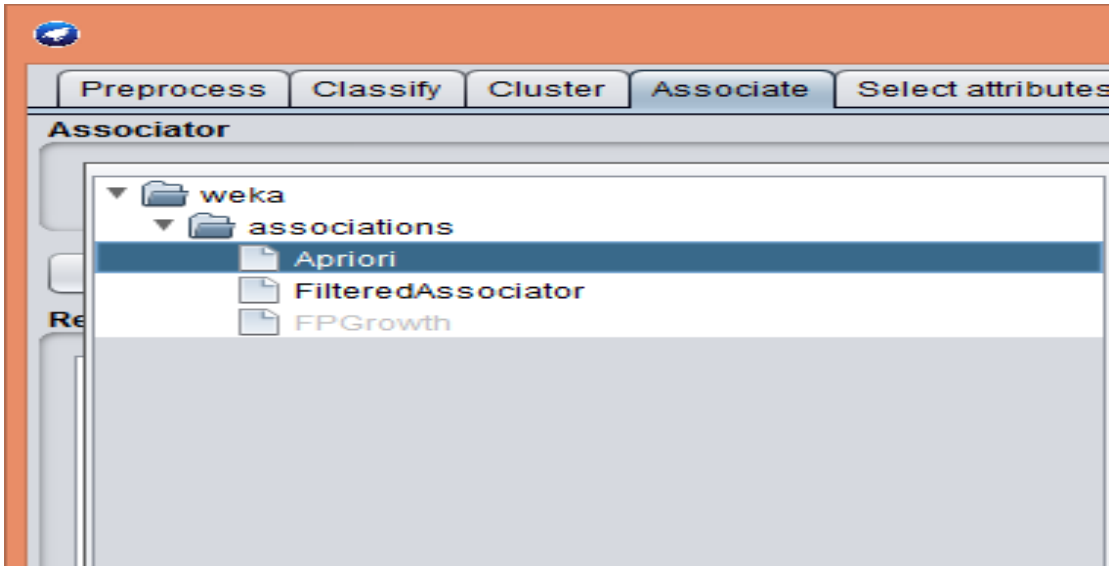
Step 2 :- Click on **Start**. (For UAE)



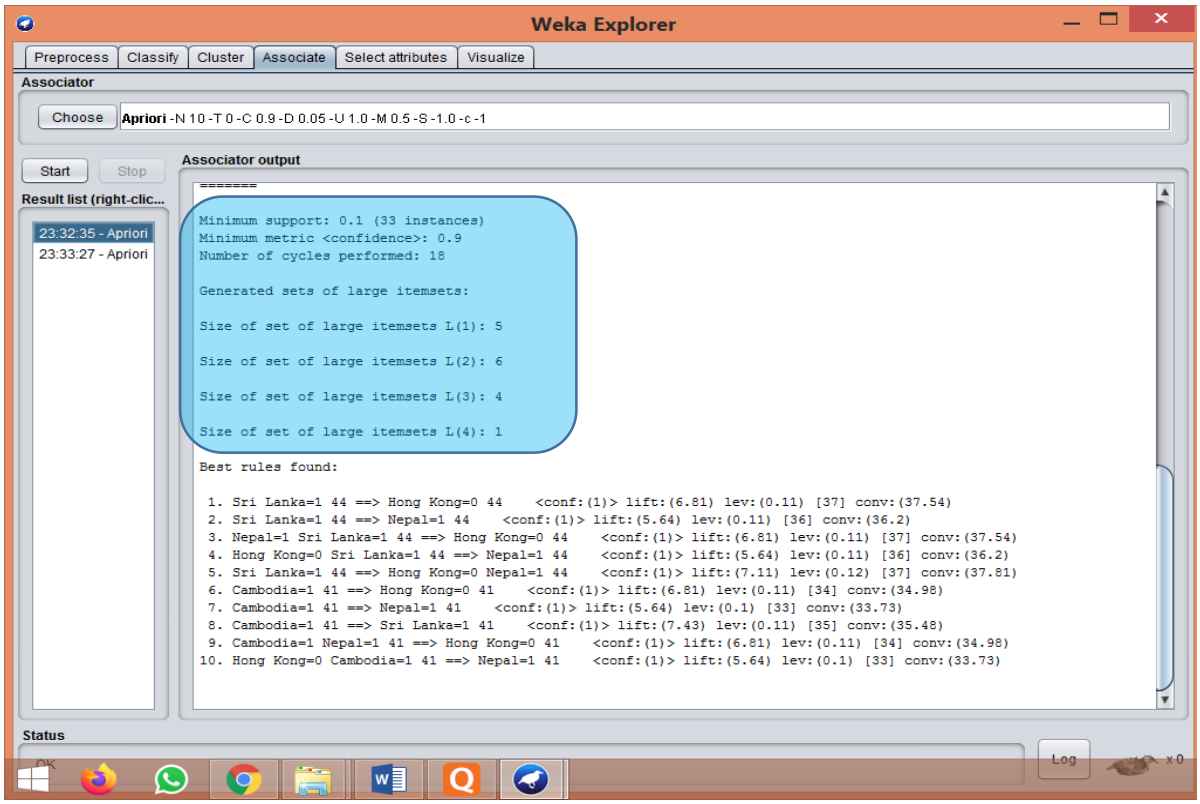
7.2 UNSUPERVISED LEARNING ALGORITHMS:-

A). APRIORI ALGORITHM:-

Step 1:- Click on the **Choose** option from **Associate** Tab. A new window will open select **Apriory** from **Associations**.

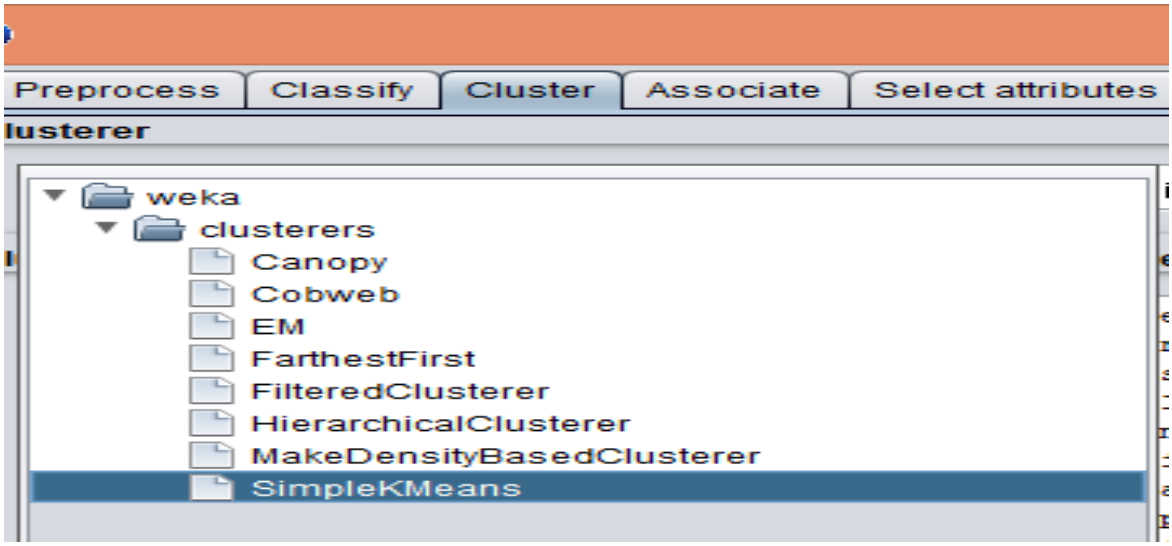


Step 2 :- Click on **Start**. (For UAE)

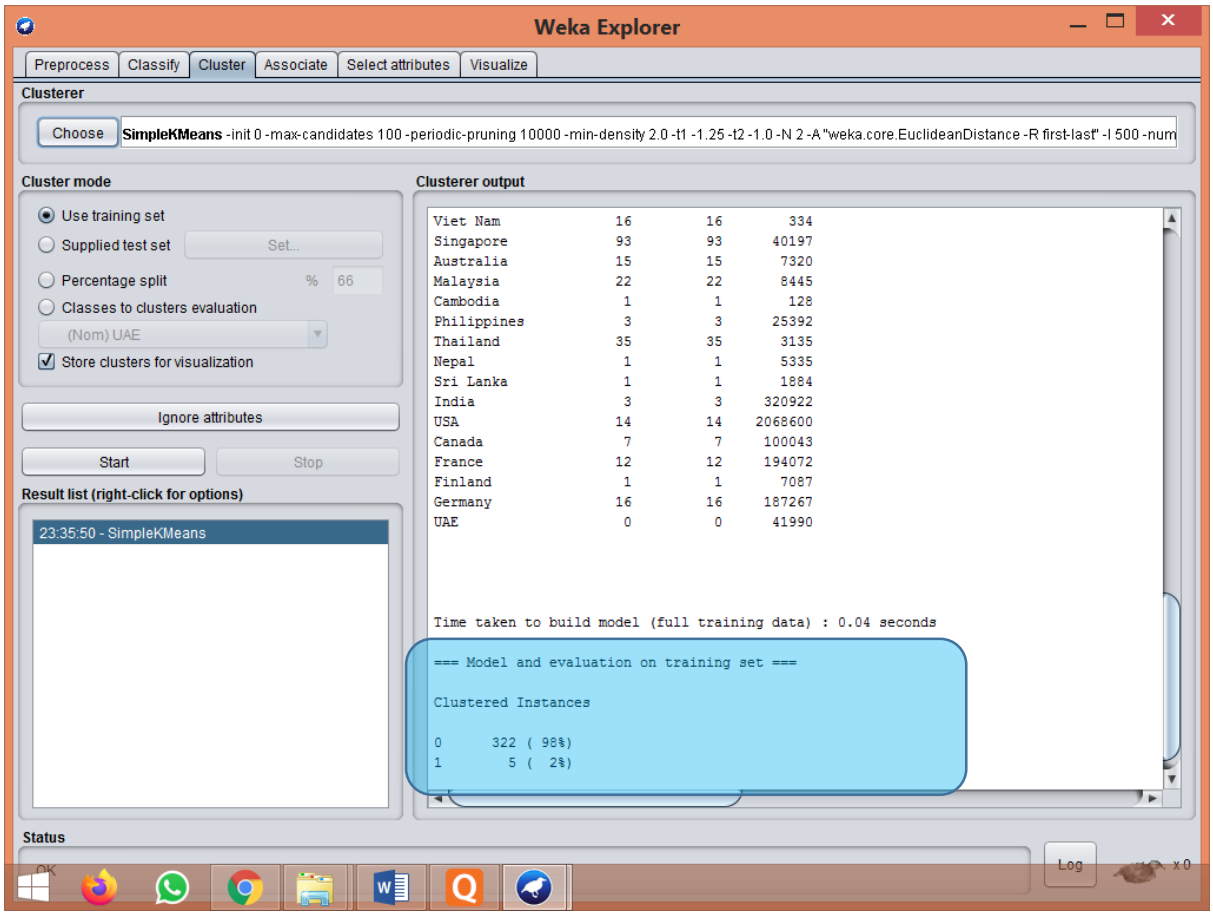


B). SIMPLE K-MEANS ALGORITHM:-

Step 1:- Click on the **Choose** option from **Cluster** Tab. A new window will open select **K-means** from **Clusterers**.



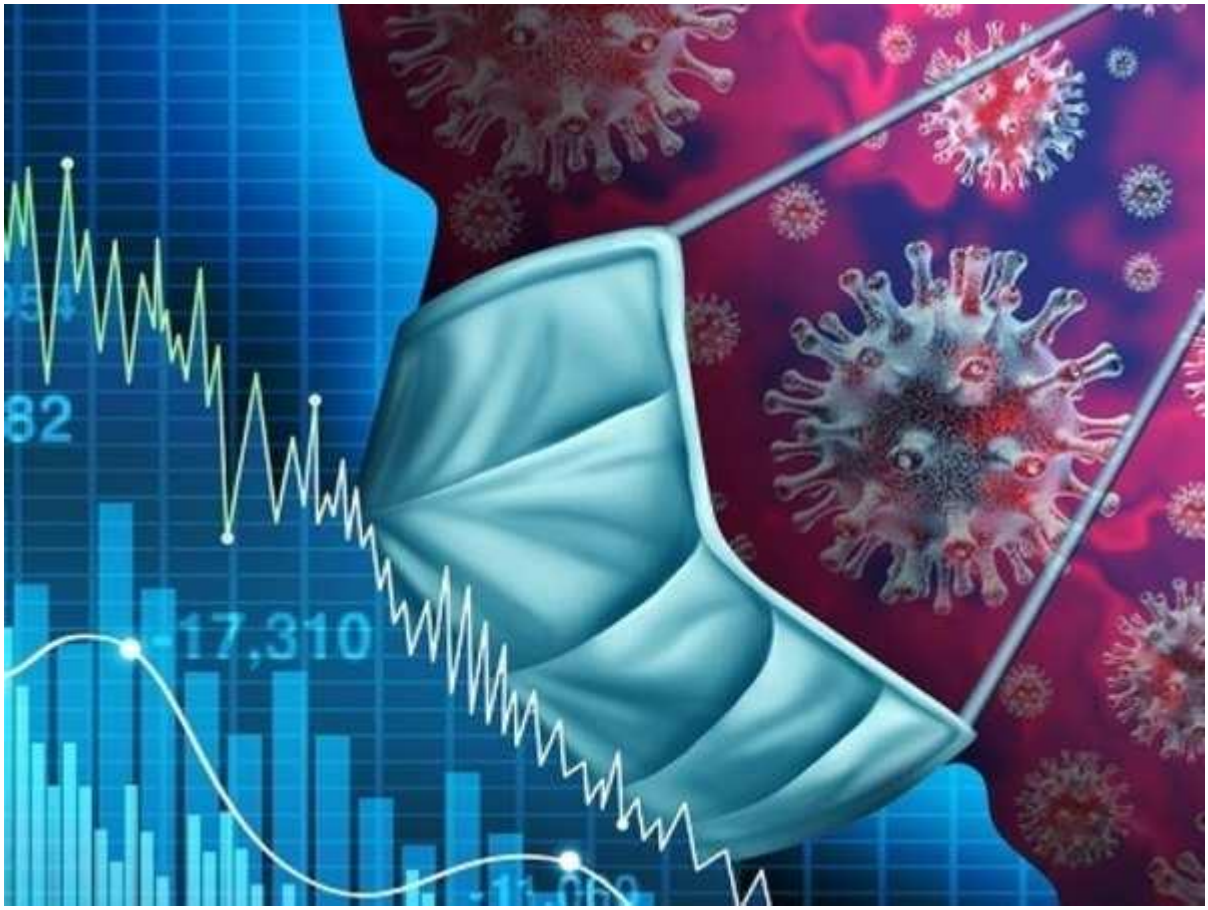
Step 2 :- Click on **Start**. (For UAE)



CHAPTER NO. 8

CONCLUSION

Through this project, the analysis on COVID-19 data has been performed successfully. The analysis on this pandemic spread has been done and compared between different countries. The analysis of confirmed cases, active cases are done to give a clear look on how the virus is spreading, which countries are getting affected mostly and how different countries are recovering. A separate analysis on cases of INDIA has been done and predictions of different cases both around the world and INDIA has been done. At last, the accuracy check using different Algorithms is performed over all the analysis done in this project.



Chapter 9

REFERENCES AND BIBLIOGRAPHY

1. www.kaggle.com.

2. www.stackoverflow.com

3. Data Mining: Concepts and Techniques by Jiawei Han

4. WEKA software

THANKS STAY SAFE STAY HOME