EAST WEST UNIVERSITY

COVID-19 IMPACT ANALYSIS AND DEATH PREDICTION USING MACHINE LEARNING

Prepared by:

SALMAN AHAD

ID: 2017-2-50-011

This Thesis paper submitted in Partial Fulfillment of the requirements for the degree of Bachelor of Science in Information and Communication Engineering

Supervised by:

Dr. Mohammad Arifuzzaman

Associate Professor and Chairperson

Department of Electronics & Communications Engineering

APPROVAL

The thesis paper titled "COVID-19 IMPACT ANALYSIS and DEATH PREDICTION USING MACHINE LEARNING" submitted by SALMAN AHAD (ID:2017-2-50-011) to the Department of Electronics and Communications Engineering, East West University, Dhaka, Bangladesh has been accepted as satisfactory for the partial fulfillment of the requirement for the degree of Bachelor of Science in Information and Communications Engineering and approved as to its style and contents.

Approved by

____

(Supervisor)

Dr. Mohammad Arifuzzaman

Assistant Professor and Chairperson

ECE Department

East West University

Dhaka, Bangladesh

DECLARATION

We declare that our work has not been previously submitted and approved for the award of a degree by this or any other University. As per of my knowledge and belief, this paper contains no material previously published or written by another person except where due reference is made in the paper itself. We hereby, declare that the work presented in this thesis paper is the outcome of the investigation performed by us under the supervision of Dr. Mohammad Arifuzzaman, Assistant Professor, Department of Electronics & Communications Engineering, East West University, Dhaka, Bangladesh.

Countersigned

(Supervisor)

Dr. Mohammad Arifuzzaman

Signature

Salman Ahad

ID: 2017-2-50-011

ACKNOWLEDGEMENT

I would like to express our sincere gratitude to our supervisor, Dr. Mohammad Arifuzzaman for his valuable guidance and advice throughout the investigation of the topic and our experiment on the detection techniques. His inclination to offer his effort and time is greatly appreciated. We would also like to thank him for his friendship, empathy, and great sense of humor. Through our research, we learned from him many valuable lessons and concepts about machine learning techniques. His relentless encouragement gave us the confidence to do our work.

Lastly, I thank the Almighty God, my parents, and friends for the much encouragement and support I have received during the period of this research. Their unconditional support made it possible for us to finish this thesis.

ABSTRACT

The scale of the COVID-19 epidemic, as well as the global lockdown consequences, are still unknown. However, as events unfold, there has been a rapid fall in social connections, a looming global economic downturn, deaths, and a growing fear of the "unknown," all of which have resulted in a shift in the status quo. Furthermore, the COVID-19 pandemic has had far-reaching consequences around the world, including a significant strain on various countries' healthcare systems, deaths, and other diseases/health difficulties. In this thesis paper I used various Covid19 data to analyze the present situation of the world and also the death of covid19. Then I used various machine learning model to predict the death of covid19 patients. This data is collected from Kaggle. I do analysis with almost 97k data. I used decision tree algorithm, logistic regression, KNN, and random forest classifier model. But I got the highest accuracy of 91.09% from Logistic Regression. I also did survey on the awareness of Covid19 among the students.

Table of Content

CHAPTER ONE - INTRODUCTION

Coronaviruses are encased, positive single-stranded big RNA viruses that can infect humans as well as a variety of other animals. Tyrell and Bynoe, who developed the viruses from patients with common colds, initially characterized coronaviruses in 1966 2. They have named coronaviruses (Latin: corona = crown) because of their shape as spherical versions with a core-shell and surface projections resembling a solar corona. Coronaviruses are divided into four subfamilies: alpha, beta, gamma, and delta. While alpha and beta coronaviruses are thought to have originated in mammals, particularly bats, gamma and delta viruses are thought to have originated in pigs and birds. The genome size ranges from 26 to 32 kb. Beta-coronaviruses, one of seven subtypes of coronaviruses that can infect humans, can cause severe sickness and death, whereas alpha-coronaviruses cause asymptomatic or mild disease.

On the Huanan seafood market in Wuhan, China, SARSCoV2 appears to have made the transfer from animals to humans. Attempts to identify probable intermediate hosts, on the other hand, appear to have been neglected in Wuhan, and the actual route of transmission must be identified as soon as possible.

Pneumonia was the first clinical symptom of the SARS-CoV-2 interconnected disease COVID19 that enabled case discovery. More recent findings, particularly among young children, indicate gastrointestinal symptoms and silent infections. So far, observations indicate a median incubation length of three days (range: 0–24 days) and a mean incubation period of five days. The percentage of people infected with SARS-CoV-2 who are asymptomatic throughout the duration of infection has yet to be determined. The clinical features of the disease, which include fever, cough, nasal congestion, lethargy, and other indicators of upper respiratory tract infections, usually begin after less than a week in symptomatic patients. In around 75% of patients, the infection can proceed to severe illness, with dyspnoea and severe chest symptoms resembling pneumonia, as revealed on computed tomography on admission. Pneumonia usually develops in the second or third week after the onset of symptoms. Reduced oxygen saturation, blood gas deviations, changes visible on chest X-rays and other imaging techniques, including ground glass abnormalities, patchy consolidation, alveolar exudates, and interlobular involvement, all of which eventually indicate deterioration, are all common signs of viral pneumonia. Inflammatory markers (C-reactive protein and proinflammatory cytokines) are increased, and lymphopenia appears to be prevalent.

Dense communities are particularly vulnerable, and Africa is undoubtedly the most vulnerable region because to the high volume of transportation between China and Africa. Few African countries have enough and appropriate diagnostic capabilities, and dealing with outbreaks poses clear obstacles. In fact, the virus may soon spread to Africa. The World Health Organization has identified 13 high-priority nations (Algeria, Angola, Cote d'Ivoire, DRC, Ethiopia, Ghana, Kenya, Mauritius, Nigeria, South Africa, Tanzania, Uganda, and Zambia) that have direct linkages to China or a significant amount of travel to China.

New and reemerging disease outbreaks, such as the present COVID19 epidemic, can possibly paralyze health systems in most Sub-Saharan African nations, at the expense of primary healthcare requirements. In the countries hit by the Ebola virus, the economic and healthcare effects are still being felt five years later. Effective epidemic responses and preparedness during large-scale events are difficult to come by in Africa and other low- and middle-income nations. Only by strengthening current regional and sub-Saharan African health structures can such circumstances be partially alleviated.

CHAPTER-TWO- LITERATURE REVIEW

“COVID-19” this pandemic has made us seen the unseen era in this world since 2020. A tiny virus has competed with the entire creation of this universe and affected every inch of our regular life. The researchers worldwide took it as a challenge and worked relentlessly to find the insights out of the deadly virus. Many researchers have find out the pattern of covid patients, prediction of covid exposure, number of affected patients and deaths etc.

This was really a new scenario for all the people around the world even for the researchers who may come up with some analysis. It took quite a time to analyze this Covid-19 data and reach a conclusion. Yet, many researchers have worked on this and have published many useful analytical journals and papers that is worth of mentioning. In this chapter, the relevant research works on covid-19.

M. Zivkovic et. al. [1] basically describes the improved algorithms like the current time-series prediction algorithms based on the hybrids merging the machine learning and nature-inspired algorism to analysis the data of the COVID-19. The main focus was to predict the number of the newer infected cases of the patients so that predicted cases may prove accurate and so adequate measure can be taken. Here, the proposed prediction model was based on a hybrid model involving the ML model and neuro-fuzzy inference system and also improved beetle antennae search swarm intelligence metaheuristics. Initially, an improved beetle antennae search was applied and been observed that it overcomes the drawbacks of its main or original version. Later on, the enhanced algorithm has been implement for the COVID-19 data and gave a really significant output. The proposed model CESBAS-ANFIS attained a R^2 score of 0.9763, which is comparatively higher when compared to the R^2 value of 0.9645, that was attained by FPASSA-ANFIS. The proposed hybrid method outperformed other sophisticated approaches tested on the same datasets, proving to be a useful tool for time-series prediction, according to simulation results and comparative analysis.

F. A. B. Hamzah et. al. [2] in their research paper, performed a real time data analysis for the COVID-19 data. Firstly, they visualized website data and then prepared a queried data that was used for analyzing with the Susceptible-Exposed-Infectious-Recovered (SEIR) predictive modelling model. That model was used to analyze the outbreak of the infected cased within and outside of China. During the time of this research, The number of confirmed cases is expected to exceed 76000, with the outbreak peaking before February 20, 2020. On January 20, 2020, the average Infected-Suspected ratio was found to be 2.399, which they used to obtain the number of Exposed people as a product of the number of Infected people. This outbreak is expected to peak in late May 2020 and then begin to subside in early July 2020.

R. Vaishya et. al. in their paper [3], they emphasize on using the artificial intelligence (AI), Internet of Things (IoT), Big Data and Machine Learning models and techniques to deal and analyze the news health related issues. Hence, the COVID-19 data and also in other health crisis models related to these algorithms can be proven beneficial as per the authors' analysis. In this paper, initially, a quick review on the literature is done on the data of PubMed, Scopus and Google Scholar using the keyword of COVID-19 or Coronavirus and Artificial Intelligence or AI. By this, they collected the latest information. By collecting and analyzing all previous data, this technology played an important role in detecting a cluster of cases and predicting where this virus will affect in the future.

S. Lalmuanawma et. al. [4] implemented machine learning method and artificial intelligence technology on Covid-19. This research is basically on rapid and critical analysis of the cases. This paper looks at recent studies that use machine learning and artificial intelligence to help researchers in a variety of ways. It also addresses a few common pitfalls and difficulties encountered when applying such algorithms to real-world problems. The paper also discusses model designers', medical experts', and policymakers' recommendations for dealing with the Covid-19 pandemic now and in the future. To avoid the human intervention in clinical trials of covid 19 prediction, treatment, vaccination, vaccine development etc., AI and ML has shown proven efficiency.

S. Kushwaha et. at. [5] mainly tries to emphasize on the significance of machine learning models in solving the covid-19 pandemic. Here the main was to discuss the importance and how the machine learning models as an evolving sector of artificial intelligence can be implemented to the data of Covid-19 and come with an analytical report. All type of ML models like supervised models, unsupervised models and semi supervised learning and also self-learning models have been implemented and came up with a significant efficient result on the data analysis of Covid-19.

In short, they have focused a merged report on all the ML models implementation report on the covid-19 data.

K. B. Prakash et. al. [6] in their paper, they performed an analytical prediction and evaluation of the COVID-19 using the ML models. Different machine learning models have been implemented to the targeted dataset such as Random Forest Regressor and Random Forest Classifier outperformed the other machine learning models like SVM, KNN+NCA, Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, Multilinear Regression, Logistic Regression and XGBoost Classifier. SVM, KNN+NCA, DT Classifier, GNB Classifier, Multilinear Regression, Logistic Regression, Random Forest Classifier, Random Forest Regressor, XGB Classifier models have given accuracy of 0.96667, 0.93333, 0.96667, 0.83333, 0.93333, 0.93333, 0.96667, N/A, 0.93333 respectively in predicting the number of cases.

H. Yun et. al. [7] performed an analytical research on the highly infectious Covid-19. They practically collected hematology and nucleic acid data from 2510 patients for the infectious retrospective analysis of the Covid-19. COVID-19 and influenza A and B infection rates were 1.3%, 3% and 3% respectively in their obtained result. Among 32 patients who were COVID-19 positive, 47% and 50% patients have been shown decreased lymphocyte count and lymphocyte ratio respectively, 66% and 75% showed decreased eosinophil count and eosinophil ratio, and 56% patients signs an increased level of C-reactive Protein. The positive rate of influenza A and B infection was much higher than that of COVID-19. Final conclusion can be added from their research is that, COVID-19 causes some hematological indicator changes in human body.

E. Ong et. al. [8] developed a machine learning-based Vaxign-ML reverse vaccinology tools to predict 22 COVID-19 vaccine candidates. The model perfectly associates with the clinical trials of the COVID-19 vaccine. This research states that coronaviruses are 64 positively stranded RNA viruses with its genome packed inside the nucleocapsid (N) protein and 65 enveloped by the membrane (M) protein, envelope (E) protein, and the spike (S) protein (Li, 66 2016). There had been many vaccine researches for coronaviruses but all were stopped soon after the outbreak of SARS and MERS. ML was used to analysis the existing trials and came up with a more effective vaccine development.

Y. Zoabi et. al. [9] proposed a machine learning model that was trained on a 51831 tested individuals in Israel. The model used eight binary features and acquired high accuracy. Predictions from this research were generated using a gradient-boosting machine learning model built with decision-tree base-learners. The predictor made with gradient-boosting model was trained with LightGBM Python package The model was scored on the test set using the auROC. auROC is a performance measure tool that is used for validation set in this paper. This paper shows a result of 95% accuracy in prediction of covid infected people.

CHAPTER-THREE-COVID-19 EFFECTS

IMPACT OF THE GLOBAL LOCKDOWN ON FOOD SECURITY

People's access to food has been hampered by the global lockdown, which has also hampered the manufacturing of certain foods. Overall, the global agricultural market is stable, however output of major essentials such as rice, wheat, and maize has increased. 18 However, the main sources of food insecurity during this time include limits on people and vehicular movement, interruption of the food supply chain, and export restrictions. As a result of these factors, the price of some cash crops has fallen due to lower global demand, while the price of other foods has risen due to rising demand within the locations. This is especially evident in Sub-Saharan Africa, where border closures have resulted in reduced labor and interrupted the flow of food and other items, affecting the cost of living, health, and nutrition. Another aspect of food insecurity as a result of the lockdown, particularly in Africa, is that many people rely on enormous open markets, which are generally packed, for their food and essentials, as well as much-needed funds for subsistence farmers. In several African countries, these markets have been largely shut down. According to the United Nations World Food Program, over 265 million people could face extreme food insecurity by the end of 2020 as a result of income and remittance losses. Food insecurity or inaccessibility, on the other hand, is a growing problem in the developed world.

IMPACT OF THE GLOBAL LOCKDOWN ON THE GLOBAL ECONOMY

It should come as no surprise that the shutdown is having an impact on the global economy also. Restrictive policies in many countries have caused swings and volatility in international trade, finance, and investments. Following months of lockdown and restriction, countries in the global South and North saw swings in trade of goods and services. For example, in the United Kingdom, the lockdown harmed trade, resulting in a drop in both imports and exports in the second quarter (April to June) of 2020, followed by a surge in imports and exports of trade in products in the third quarter (July to September) once the restrictions were eased. For emerging economies like Kenya, the worldwide trade embargo resulted in a 12 percent increase in exports and a 28 percent decline in imports. The sudden shutdown in economic activity impacted commodities related to agriculture and pharmaceuticals across several countries' economy. Government restrictions and stockpiles became a source of concern for food security and commodity prices. Similarly, the oil sector was severely impacted, owing to the Organization of Petroleum Exporting Countries (OPEC) and its allies' inability to agree on production cuts necessary to stabilize oil prices, with Russia refusing to reduce oil production, causing Saudi Arabia to flood the market with excess products at a reduced price, resulting in the steepest drop in oil prices since 1991. These two causes combined to cause a sharp decline in oil prices, with forecasts for a gradual recovery below $40 per barrel by 2022. Nigeria, for example, which relied on oil sales on the worldwide market, was severely impacted when the crude oil price fell from $60 per barrel to $30 per barrel in March 2020. This result was expected, given the sharp drop in demand for aviation and automobile gasoline. As a result, the country's budget was severely impacted, with no revenue to serve it.

The impact of the lockdown was also visible in the global market, with stock indices around the world falling in March 2020, just as the lockdown began. Several of the world's largest firms saw their stock prices plummet at this time. The value of the dollar has risen against the majority of international currencies, impacting the ability of many African countries to conduct commerce. For example, on March 23rd, the value of the global equities market was wiped out by $26 trillion, resulting in massive losses for those who hold shares as well as pension and insurance funds. The capacity to access credit, loans, and mortgages were also hampered as a result of the abrupt decline in economic activity, which disrupted incoming cash flows.

Impact of the global lockdown on education

As part of the global lockdown effort to stop the spread of the virus, most governments have closed schools at all levels. This has had an impact on formal education, with 143 countries imposing nationwide school closures. This has impacted 1,184,126,508 (67.6%) of enrolled learners in pre-primary, primary, lower and upper secondary, and university education levels around the world. The closure of educational facilities, particularly for children, is appropriate since they have lower immunity levels and a higher tendency to transmit symptomatic infectious disease, as evidenced in the transmission of influenza among children versus adults. The effectiveness of school closures in combating the COVID-19 outbreak, however, has been questioned.

Many universities throughout the world are quickly adopting online education to conduct lectures and other academic activities. Some traditional universities, such as Cambridge University in the United Kingdom, have advocated that all classes be moved online until the next academic session.

Furthermore, as a result of the disturbance in their education, school closure has many and different effects on young people. The lockdown has caused worry among some young people in some parts of Africa, as there are unanswered questions about how education would continue after the lockdown, due to lack of family income, repetition of the school year, or even failure in national tests. Working on the farm or completing home chores can also create a loss of enthusiasm and concentration when it comes to studying.

IMPACT OF THE GLOBAL LOCKDOWN ON TOURISM, HOSPITALITY, SPORTS AND LEISURE

The global embargo has had a substantial impact on tourism, with international tourism down 22% in Q1 2020 and a forecast fall of 60%–80% by the end of 2020. Overall, there were 67 million fewer international tourists in the first quarter of 2020, from January to March, resulting in a $80 billion loss in exports. This decline in tourism is expected to threaten around 100–120 million employment directly tied to tourism, with a loss of $910 billion to $1.2 trillion in export income. Notably, countries believed to be hotspots for the COVID-19 epidemic have lost millions of dollars in revenue due to the lockdown, as people who would normally visit vacation destinations have canceled their plans. The severity of this loss might be attributed to the fact that the lockdown's tactics, such as movement restrictions and social separation, have a direct impact on tourism.

The COVID-19 pandemic has forced the aviation industry to find safety measures that may be taken to restore normalcy and prevent the virus from spreading, particularly at airports.

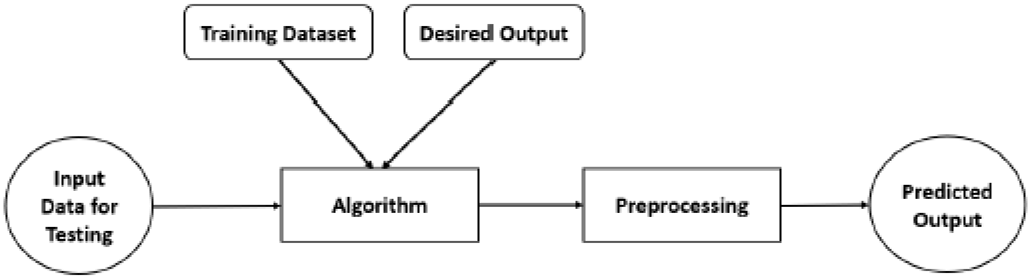
With event cancellations, hotel and accommodation closures, and the shutdown of leisure parks, restaurants, and services, the hospitality value chain was also impacted. Many parks, gyms, and pools for leisure activities were closed, as was tourism. Sporting events are being postponed or cancelled all across the world. The Olympic Games in Tokyo, which will take place in 2020, are a significant example. The global event has been postponed until 2021, with all indications that, given the recent surge and growth in the number of infected cases, it may be postponed even further. This is due to the large crowds and high crowd density associated with such events, as well as the increased risk of viral transmission.

CHAPTER FOUR -MACHINE LEARNING MODEL

MACHINE LEARNING TECHNIQUES FOR DETECTION

We see in our previous chapter that COVID-19 is rapidly increasing day by day, and also the death from this virus is still going. So, it is necessary to detect and prevent the spread of this virus. We can use machine learning algorithms to detect death prediction. In our paper, we use a machine learning algorithm for the detection of COVID-19.

We used supervised learning for our detection techniques. In supervised learning there is input variables denoted by (x) and output variable denoted by (y) and an algorithm is used to learn the mapping function from the input to the output. In supervised learning, the training set we feed to the algorithm contains the wanted solutions, called labels.



Supervised Learning can be divided by two ways of problem solving:

1. Classification

2. Regression

Regression analysis mainly understand the relationship between dependent and independent variable and in Classification technique the algorithm learns from the input given to it and then uses this learning to classify new observation. We will use these algorithms:

K- Nearest Neighbor (KNN)

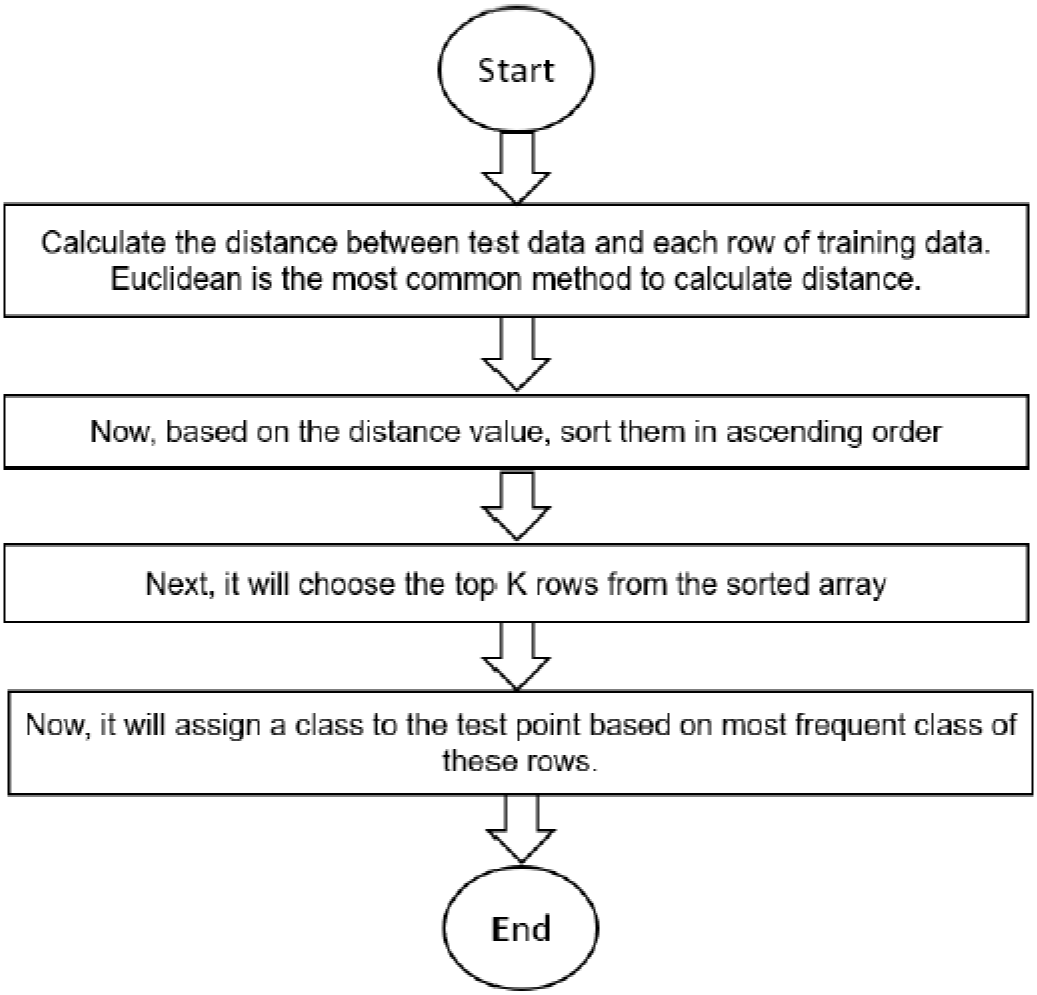
Decision Tree Algorithm

Random Forest Classifier

Logistic Regression

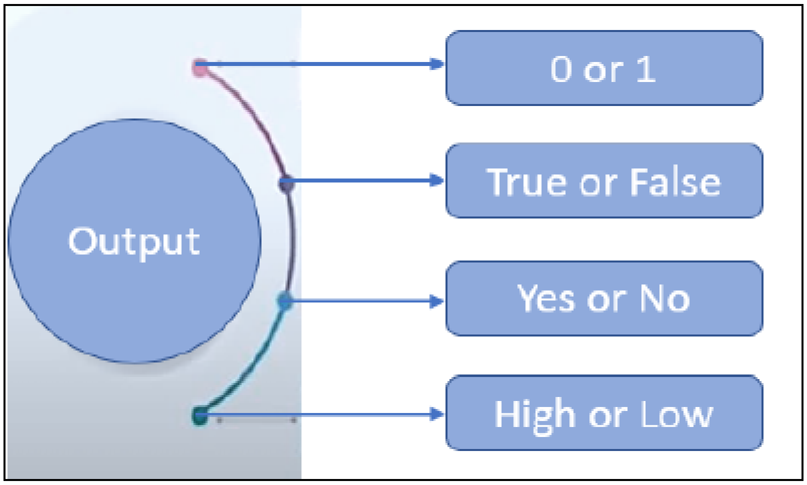
KNN

K Nearest Neighbor algorithm classifies the new data based on a similarity measure and stores all the available possible cases. Here the number of nearest neighbors is represented by k. KNN algorithm uses ‘feature similarity’ to predict the values of output. Working of KNN model:


Start
Calculate the distance between test data and each row of training data. Euclidean is the most common method to calculate distance.
Now, based on the distance value, sort them in ascending order
Next, it will choose the top K rows from the sorted array
Now, it will assign a class to the test point based on most frequent class of these rows.
End

LOGISTIC REGRESSION

Logistic Regression produces output in a binary format which is used to predict the outcome of a categorical dependent variable. Such as:

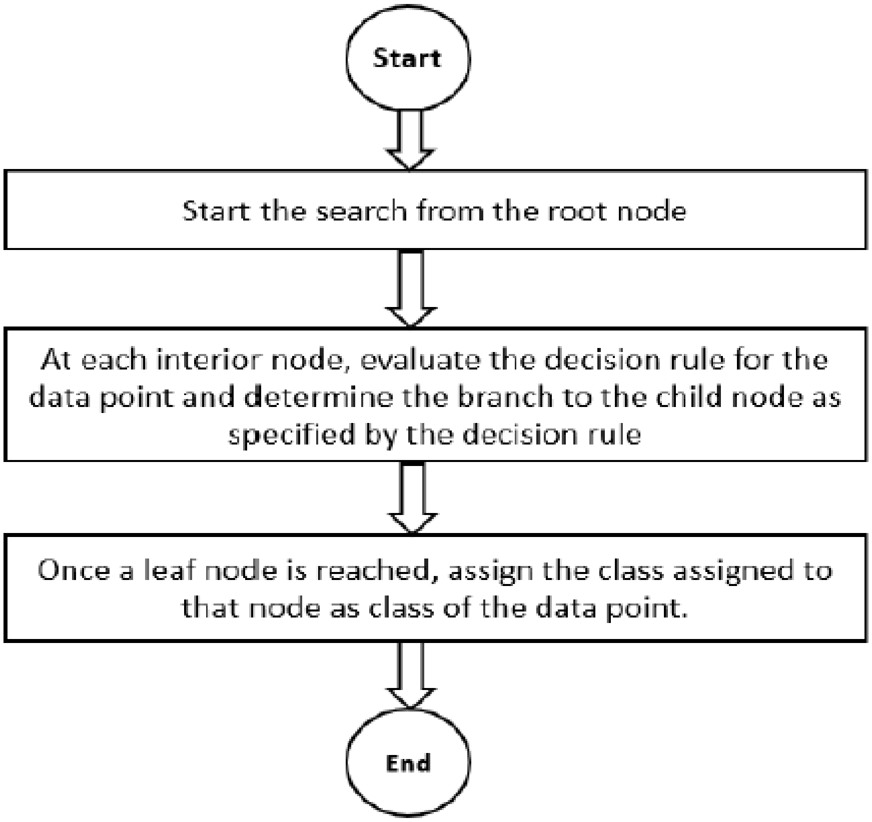


The dependent variable is also called target variable. Logistic regression predicts the outcome probability by using log function. Normally sigmoid function is used to predict the output value. Sigmoid function: ***p*** **= 1 / 1 + e-y.**

The threshold value of sigmoid function decides the outcome

DECISION TREE ALGORITHM

For different conditions Decision tree can spilt dataset in different ways. To formulate a set of decision rules decision tree algorithms, use training data. The classes of the test data are estimated based on the set of decision rules. This is represented as a tree structure with each non-leaf node turns as a decision maker and each leaf node signifies a class.



Start
Start the search from the root node
At each interior node, evaluate the decision rule for the data point and determine the branch to the child node as specified by the decision rule
Once a leaf node is reached, assign the class assigned to that node as class of the data point.
End

The stopping criteria of this algorithm:

- All the leaf nodes are labeled.
- Maximal node depth is attained.
- There is no information gain on splitting of any node

ALGORITHM

Input: Dataset file for each feature

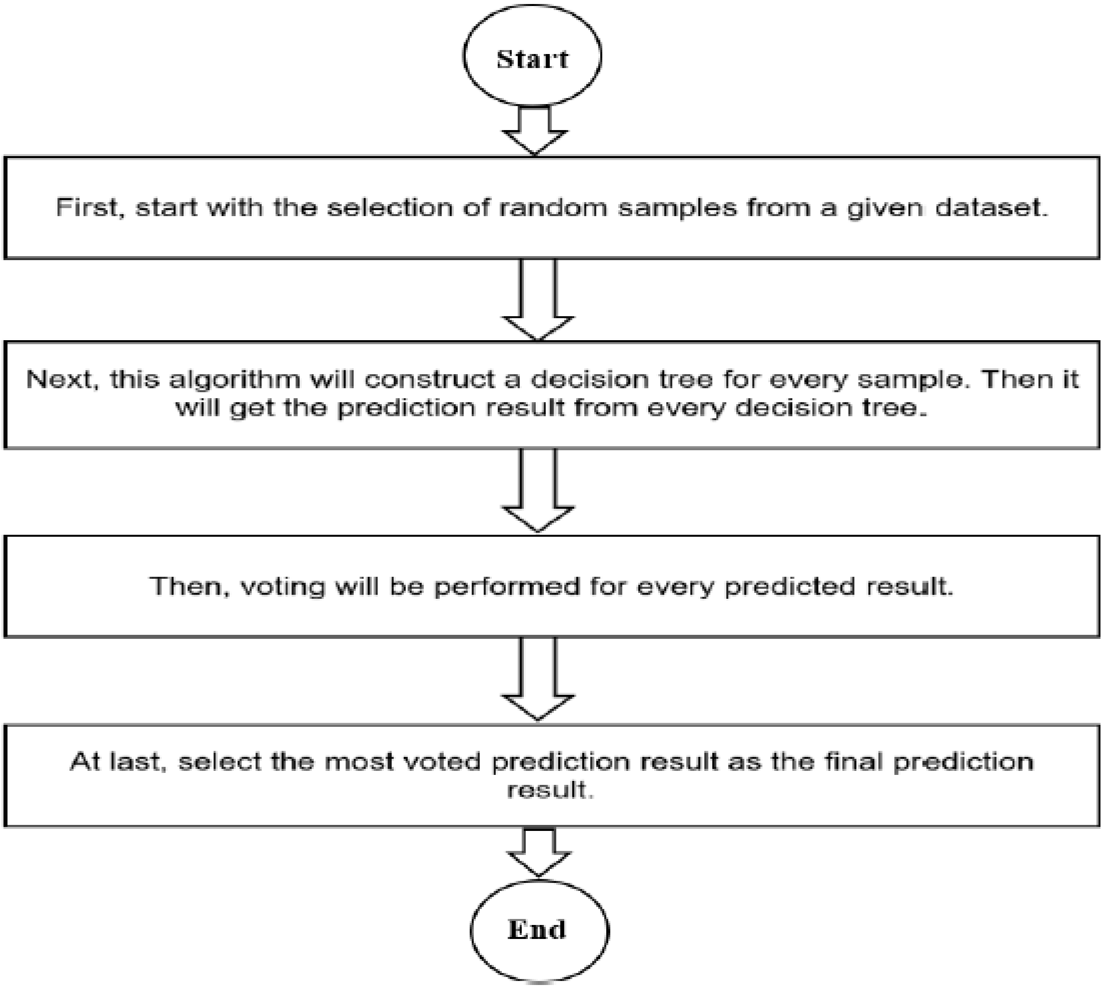
Output: Classified samples with labels allocated

1. if leaf nodes do not satisfy the early stopping criteria then
2. select attribute that gives the “best” split; assign it as the root node
3. Repeat
4. Begin from the root node as the parent node
5. Split parent node at some feature xi to maximize the information gain

6. Assign training samples to new child nodes
7. Until for each new child node
8. end if

Random Forest Classifier

Random Forest Classifier used for both classification and regression techniques. It decreases the overfitting in the region of the result. Working model of Random Forest Classifier:


Start
First, start with the selection of random samples from a given dataset.
Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
Then, voting will be performed for every predicted result.
At last, select the most voted prediction result as the final prediction result.
End

SOME IMPORTANT FEATURES OF MACHINE LEARNING

True Positive (TP): It characterizes the value of accurate estimations of positives out of actual positive cases.

False Positive (FP): It characterizes the value of wrong positive estimations that means the number of negatives value which gets falsely estimated as positive.

True Negative (TN): True negative characterizes the value of accurate estimations of negatives out of actual negative cases.

False Negative (FN): False negative characterizes the value of wrong negative estimations that means the number of positives value which gets falsely estimated as negative.

Confusion Matrix: The confusion matrix is used to determine the performance of the classification models for a given set of test data. It signifies a tabular representation of Actual vs Estimated values.

Precision: Precision score signifies the model's capability to correctly expect the positives out of all the positive prediction it made. It signifies the ratio of true positive to the sum of true positive and false positive.

Precision Score = True Positive/True Positive + False Positive

Recall: Recall score signifies the model's capability to correctly expect the positives out of actual positives. It signifies the ratio of true positive to the sum of true positive and false negative.

Recall Score = True Positive/True Positive + False Negative

Accuracy score: It signifies the model's ability to accurately expect both the positives and negatives out of all the predictions and it signifies the ratio of sum of true positive and true negatives out of all the predictions.

Accuracy Score = (True Positive + True Negative)/(True Positive + False Negative + True Negative +False Positive)

F1 Score: F1 score signifies the model score as a function of precision and recall score.

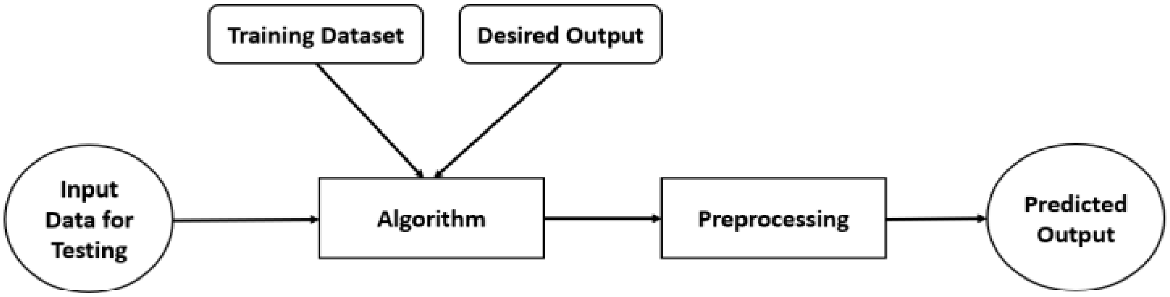
F1 Score = 2∗ Precision Score ∗ Recall Score/Precision Score + Recall Score

In our detection technique we will not detect anything manually. In python there are built in machine learning model in SKLearn’s library and we will use it. Without it we will import some necessary libraries for our detection experiment like pandas, NumPy, matplotlib, seaborn.

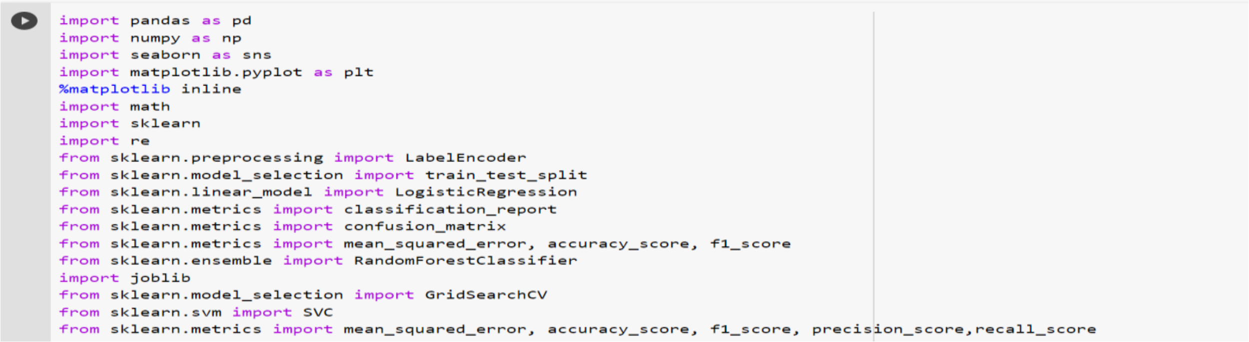
CHAPTER FIVE - MACHINE LEARNING ANALYSIS

In many cases, arriving patients to intensive care units (ICUs) do not have validated medical histories. A distressed patient, especially one who has been brought in disoriented or unresponsive, may be unable to offer information on chronic diseases including heart disease, traumas, or diabetes. Transferring medical records might take days, especially if the patient comes from another medical provider or system. Knowledge of chronic illnesses can help clinicians make better judgments regarding patient treatment and, as a result, increase patient survival rates.

We used machine learning model to predict the death and also do some visualization. We used supervised learning for our detection techniques. In supervised learning there is input variables denoted by (x) and output variable denoted by (y) and an algorithm is used to learn the mapping function from the input to the output. In supervised learning, the training set we feed to the algorithm contains the wanted solutions, called labels.

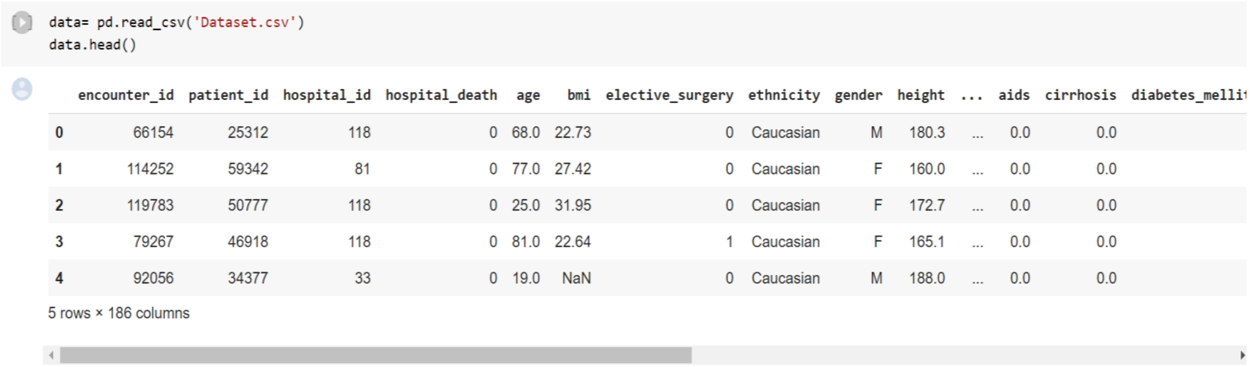


Data Collection: This data is collected from Kaggle. We do analysis with almost 97k data. At first we import all the libraries that we need for our analysis.



```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import math
import sklearn
import re
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import mean_squared_error, accuracy_score, f1_score
from sklearn.ensemble import RandomForestClassifier
import joblib
from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVC
from sklearn.metrics import mean_squared_error, accuracy_score, f1_score, precision_score,recall_score
```

Using Pandas, we read our dataset.



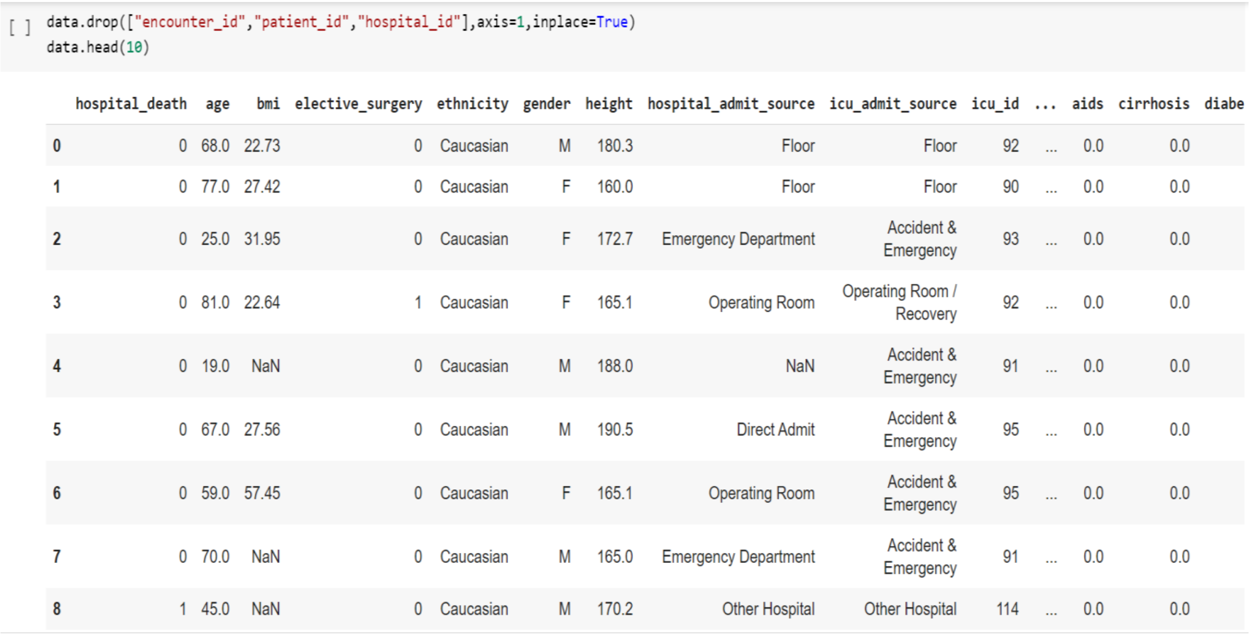
```python
data= pd.read_csv('Dataset.csv')
data.head()
```

	encounter_id	patient_id	hospital_id	hospital_death	age	bmi	elective_surgery	ethnicity	gender	height	...	aids	cirrhosis	diabetes_mellit
0	66154	25312	118	0	68.0	22.73	0	Caucasian	M	180.3	...	0.0	0.0	
1	114252	59342	81	0	77.0	27.42	0	Caucasian	F	160.0	...	0.0	0.0	
2	119783	50777	118	0	25.0	31.95	0	Caucasian	F	172.7	...	0.0	0.0	
3	79267	46918	118	0	81.0	22.64	1	Caucasian	F	165.1	...	0.0	0.0	
4	92056	34377	33	0	19.0	NaN	0	Caucasian	M	188.0	...	0.0	0.0	

5 rows × 186 columns

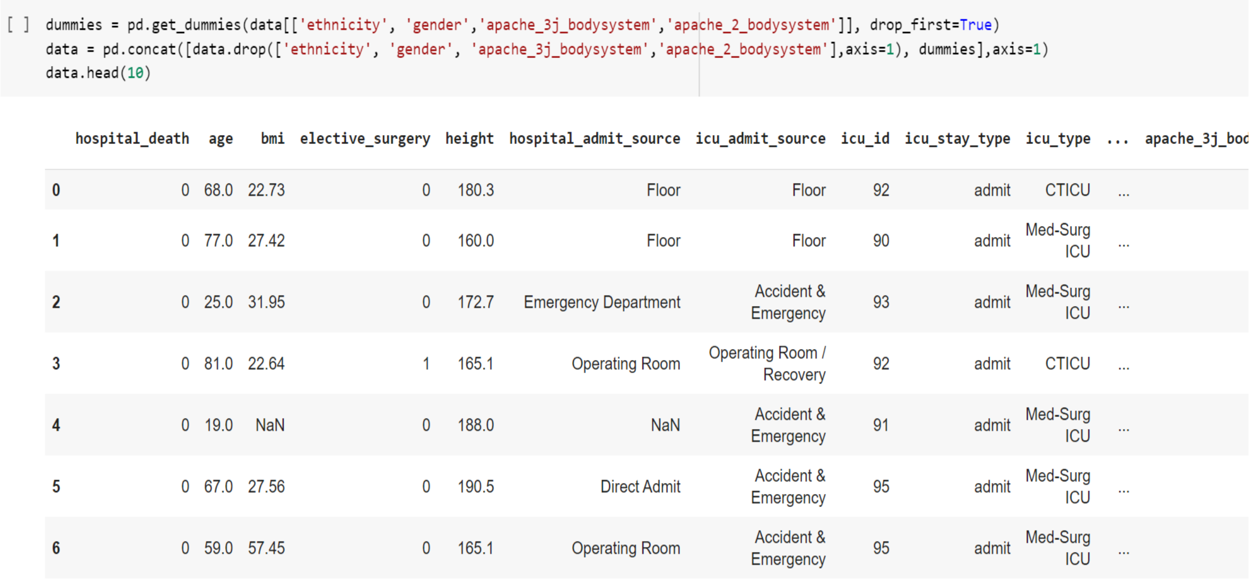
Then we preprocessed our data by dropping the columns that we don’t need and dummies the data because we can’t do analysis with string data. So, we do dummies on

'ethnicity','gender', 'apache_3j_bodysystem','apache_2_bodysystem' - column



```python
data.drop(["encounter_id","patient_id","hospital_id"],axis=1,inplace=True)
data.head(10)
```

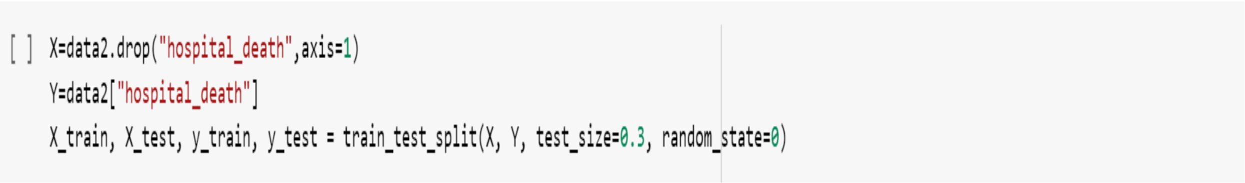
	hospital_death	age	bmi	elective_surgery	ethnicity	gender	height	hospital_admit_source	icu_admit_source	icu_id	...	aids	cirrhosis	diabe
0	0	68.0	22.73	0	Caucasian	M	180.3	Floor	Floor	92	...	0.0	0.0	
1	0	77.0	27.42	0	Caucasian	F	160.0	Floor	Floor	90	...	0.0	0.0	
2	0	25.0	31.95	0	Caucasian	F	172.7	Emergency Department	Accident & Emergency	93	...	0.0	0.0	
3	0	81.0	22.64	1	Caucasian	F	165.1	Operating Room	Operating Room / Recovery	92	...	0.0	0.0	
4	0	19.0	NaN	0	Caucasian	M	188.0	NaN	Accident & Emergency	91	...	0.0	0.0	
5	0	67.0	27.56	0	Caucasian	M	190.5	Direct Admit	Accident & Emergency	95	...	0.0	0.0	
6	0	59.0	57.45	0	Caucasian	F	165.1	Operating Room	Accident & Emergency	95	...	0.0	0.0	
7	0	70.0	NaN	0	Caucasian	M	165.0	Emergency Department	Accident & Emergency	91	...	0.0	0.0	
8	1	45.0	NaN	0	Caucasian	M	170.2	Other Hospital	Other Hospital	114	...	0.0	0.0	



```
[ ] dummies = pd.get_dummies(data[['ethnicity', 'gender','apache_3j_bodysystem','apache_2_bodysystem']], drop_first=True)
    data = pd.concat([data.drop(['ethnicity', 'gender', 'apache_3j_bodysystem','apache_2_bodysystem'],axis=1), dummies],axis=1)
    data.head(10)
```

	hospital_death	age	bmi	elective_surgery	height	hospital_admit_source	icu_admit_source	icu_id	icu_stay_type	icu_type	...	apache_3j_boc
0	0	68.0	22.73	0	180.3	Floor	Floor	92	admit	CTICU	...	
1	0	77.0	27.42	0	160.0	Floor	Floor	90	admit	Med-Surg ICU	...	
2	0	25.0	31.95	0	172.7	Emergency Department	Accident & Emergency	93	admit	Med-Surg ICU	...	
3	0	81.0	22.64	1	165.1	Operating Room	Operating Room / Recovery	92	admit	CTICU	...	
4	0	19.0	NaN	0	188.0	NaN	Accident & Emergency	91	admit	Med-Surg ICU	...	
5	0	67.0	27.56	0	190.5	Direct Admit	Accident & Emergency	95	admit	Med-Surg ICU	...	
6	0	59.0	57.45	0	165.1	Operating Room	Accident & Emergency	95	admit	Med-Surg ICU	...	

Then we train our dataset and test. Our train test ratio is 70:30.



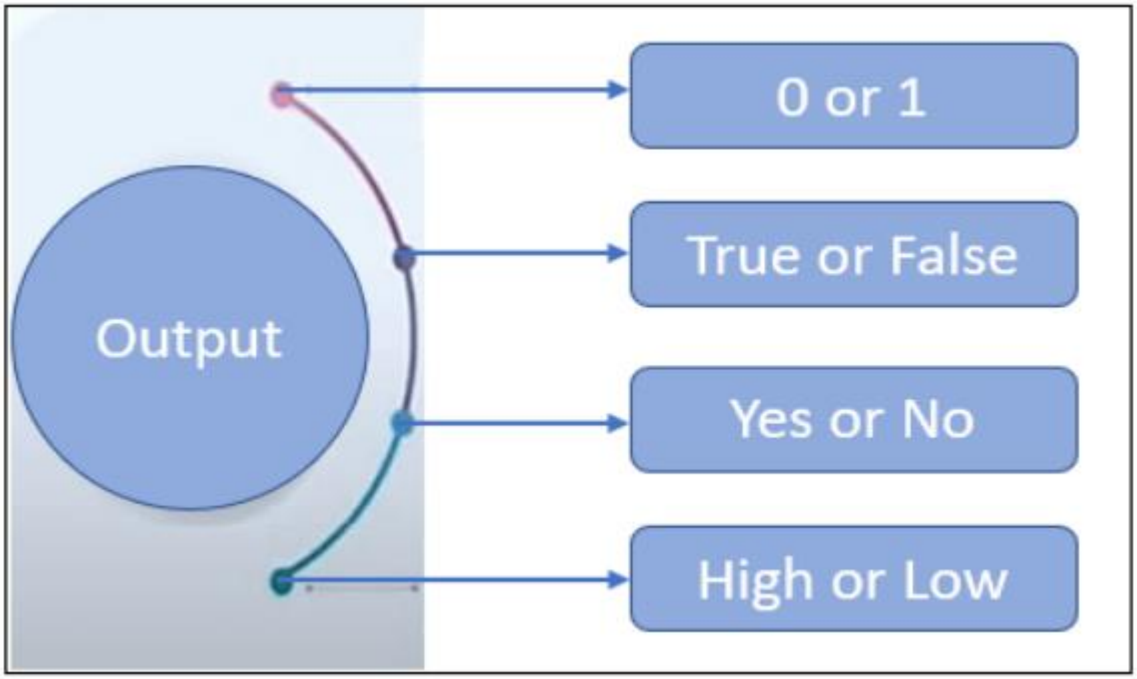
```
[ ] X=data2.drop("hospital_death",axis=1)
    Y=data2["hospital_death"]
    X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=0)
```

Result:

We used four model and got the highest accuracy of 91.09% from Logistic Regression.

Model Name	**Accuracy**
Random Forest	87.58%
Logistic Regression	91.09%
KNN	90.54%
Decision Tree Algorithm	86.21%

Logistic Regression produces output in a binary format which is used to predict the outcome of a categorical dependent variable. Such as:



The dependent variable is also called target variable. Logistic regression predicts the outcome probability by using log function. Normally sigmoid function is used to predict the output value. Sigmoid function: p = 1 / 1 + e-y.

PLOTS:

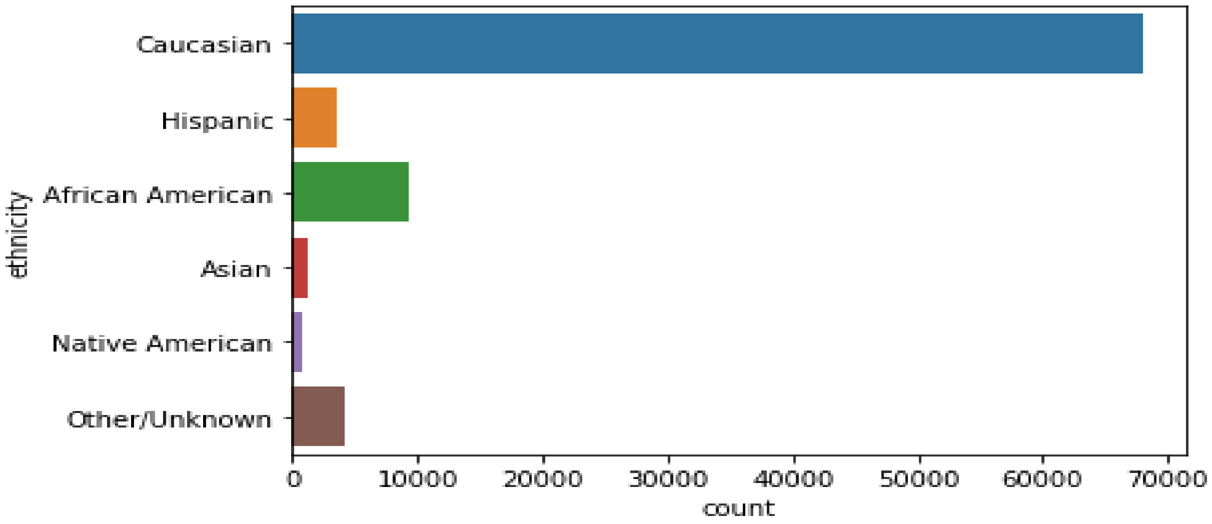


Figure: Ethnicity by numbers

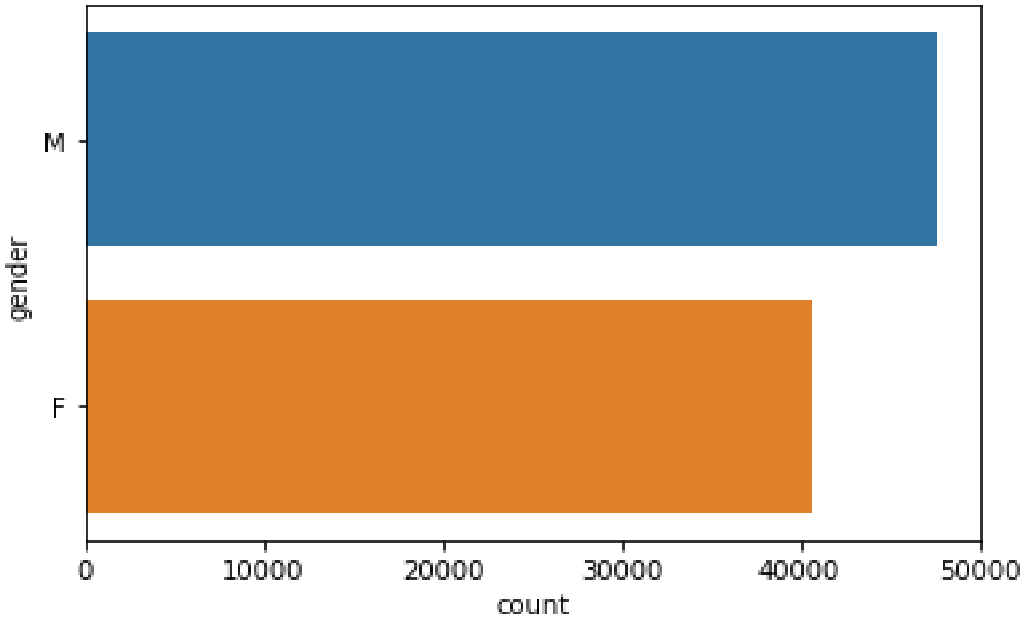


Figure: Total Death by male and female

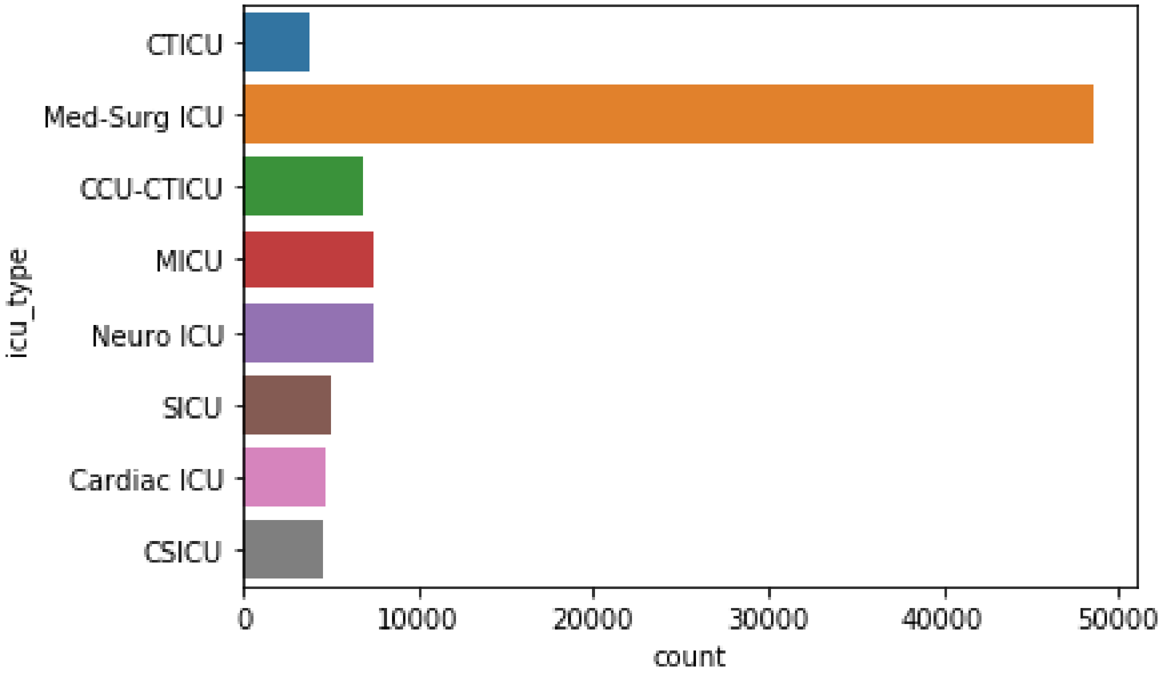


Figure: Total number of ICU Type where patient admitted

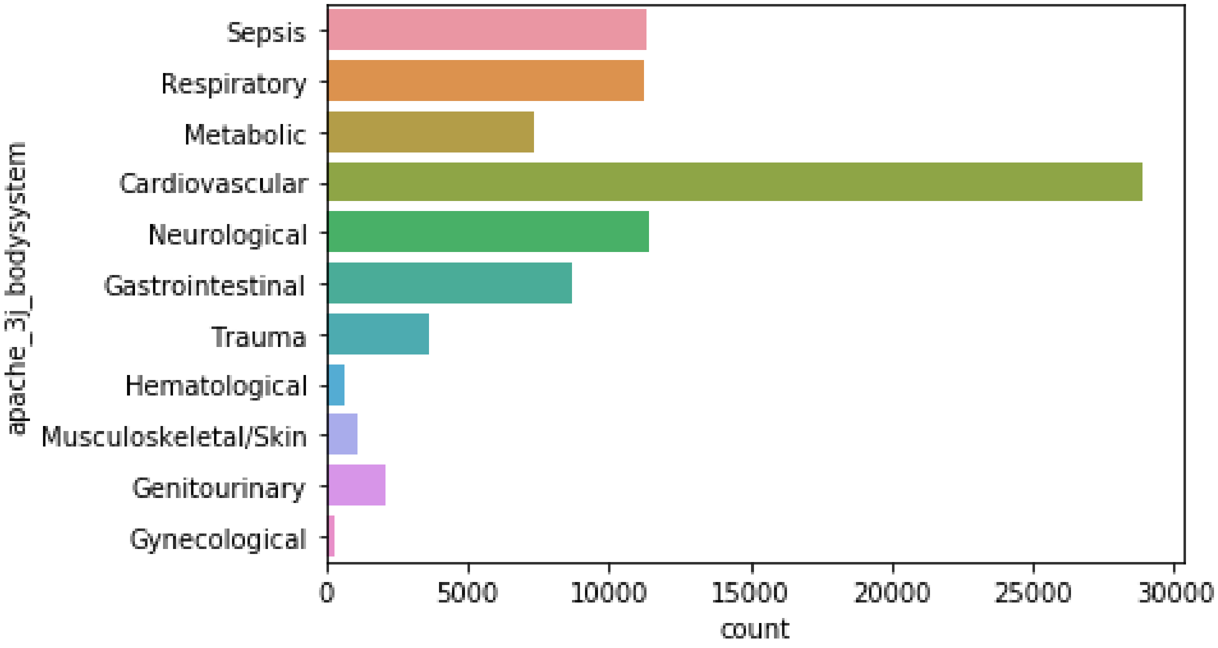


Figure: Apache_3j_bodysystem

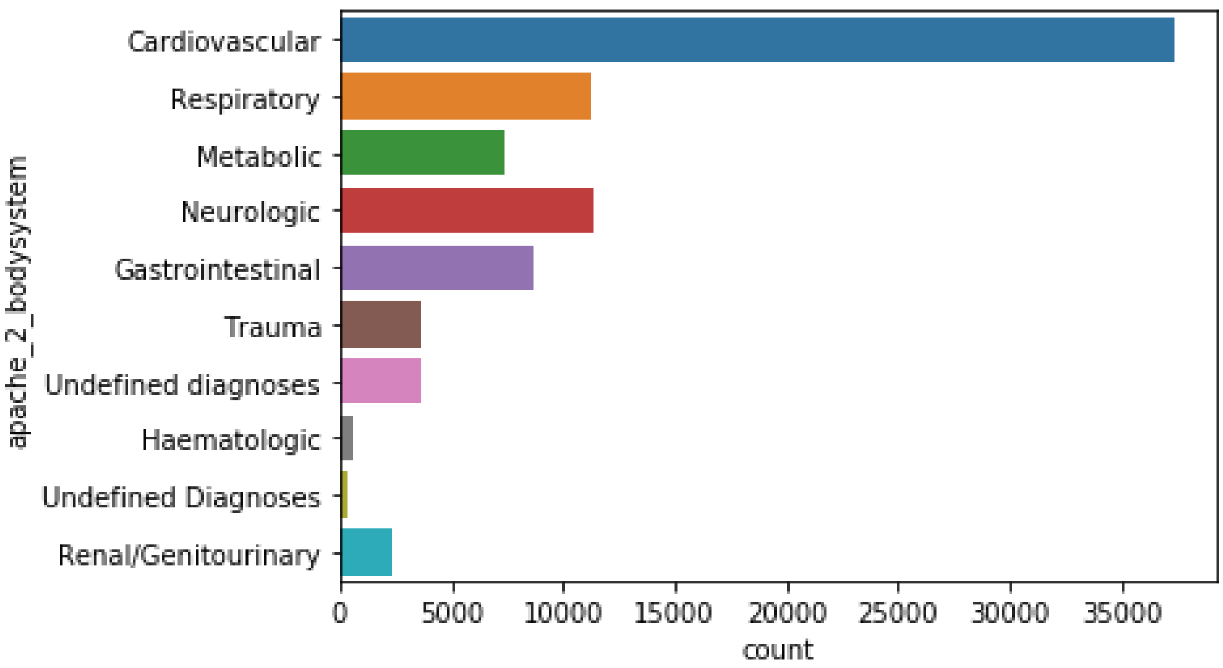
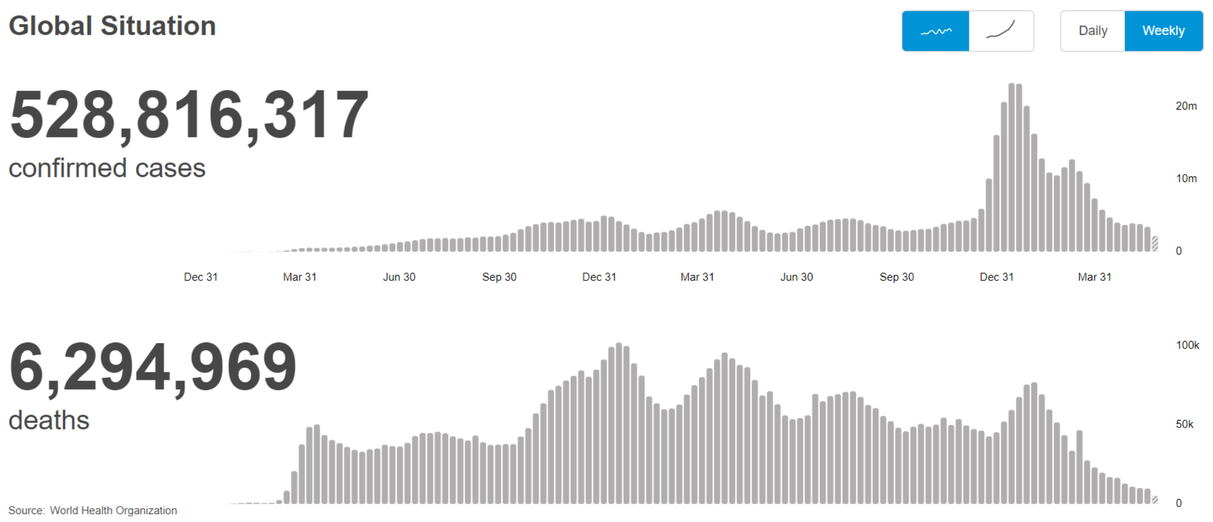


Figure: Apache_3j_bodysystem

CHAPTER-SIX - COVID-19 DATA ANALYSIS AND SURVEY REPORT

As per global data from WHO there are 528,816,317 covid cases are confirmed and 6,294,969 are deaths. But the actual number will be more forsure. Because everyone was not tested and not go to the hospital for covid reason. Or may be some deaths are with covid suspicious and not tested positive. Here is the data chart from WHO-



Global Situation

528,816,317

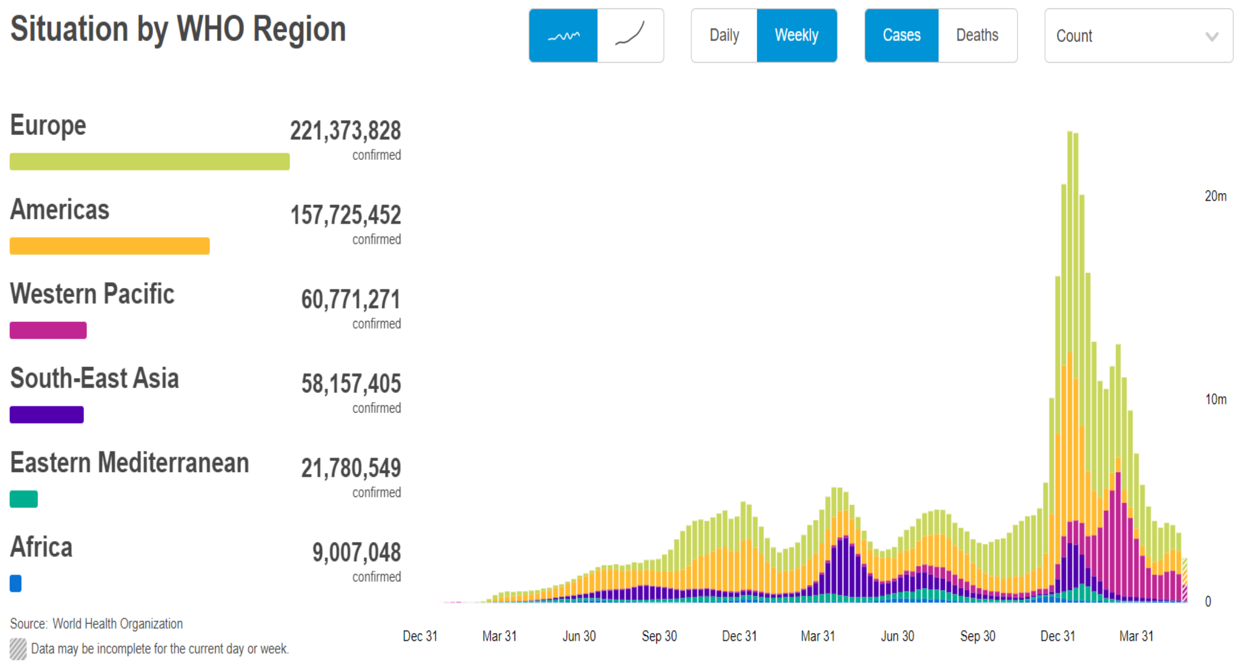
confirmed cases

6,294,969

deaths

Source: World Health Organization

The most number of Covid19 cases are in Europe 221,373,828. Then the second position is for America 157,725,452.



Situation by WHO Region

Europe 221,373,828 confirmed

Americas 157,725,452 confirmed

Western Pacific 60,771,271 confirmed

South-East Asia 58,157,405 confirmed

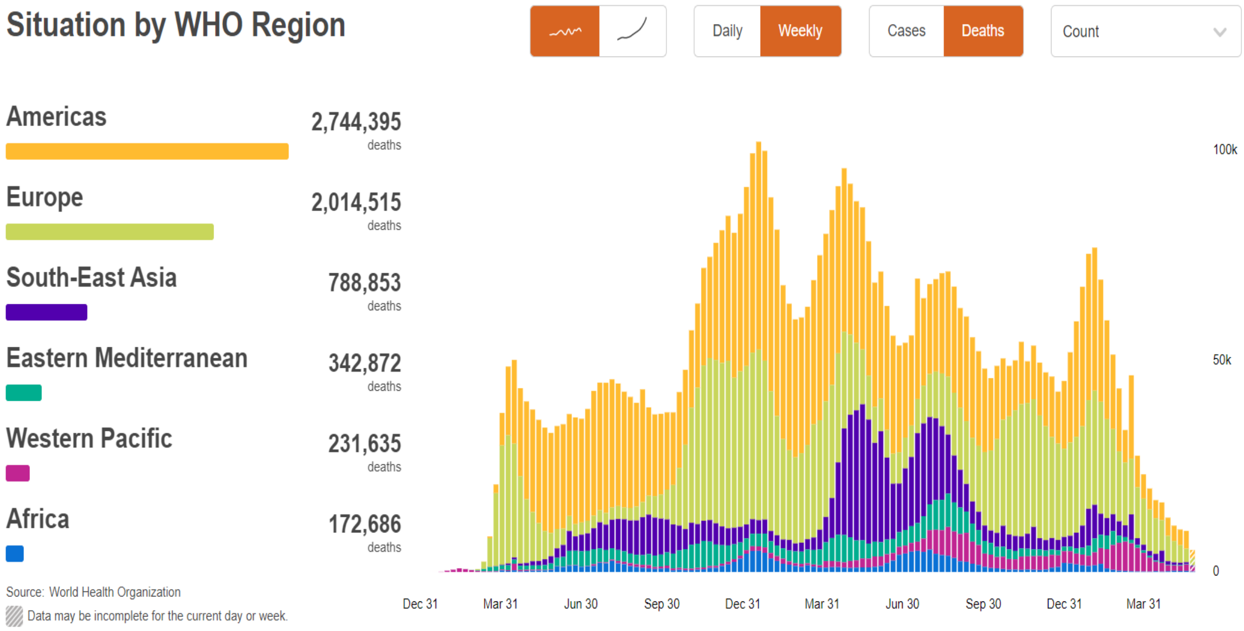
Eastern Mediterranean 21,780,549 confirmed

Africa 9,007,048 confirmed

Source: World Health Organization

▨ Data may be incomplete for the current day or week.

But the number of death in America is the highest 2,744,495.



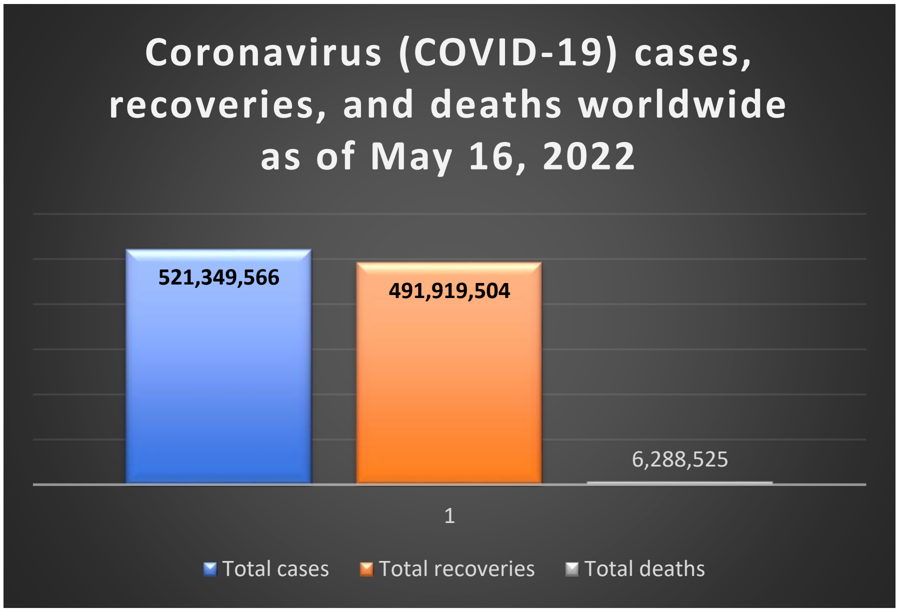
Situation by WHO Region

Americas	2,744,395 deaths
Europe	2,014,515 deaths
South-East Asia	788,853 deaths
Eastern Mediterranean	342,872 deaths
Western Pacific	231,635 deaths
Africa	172,686 deaths

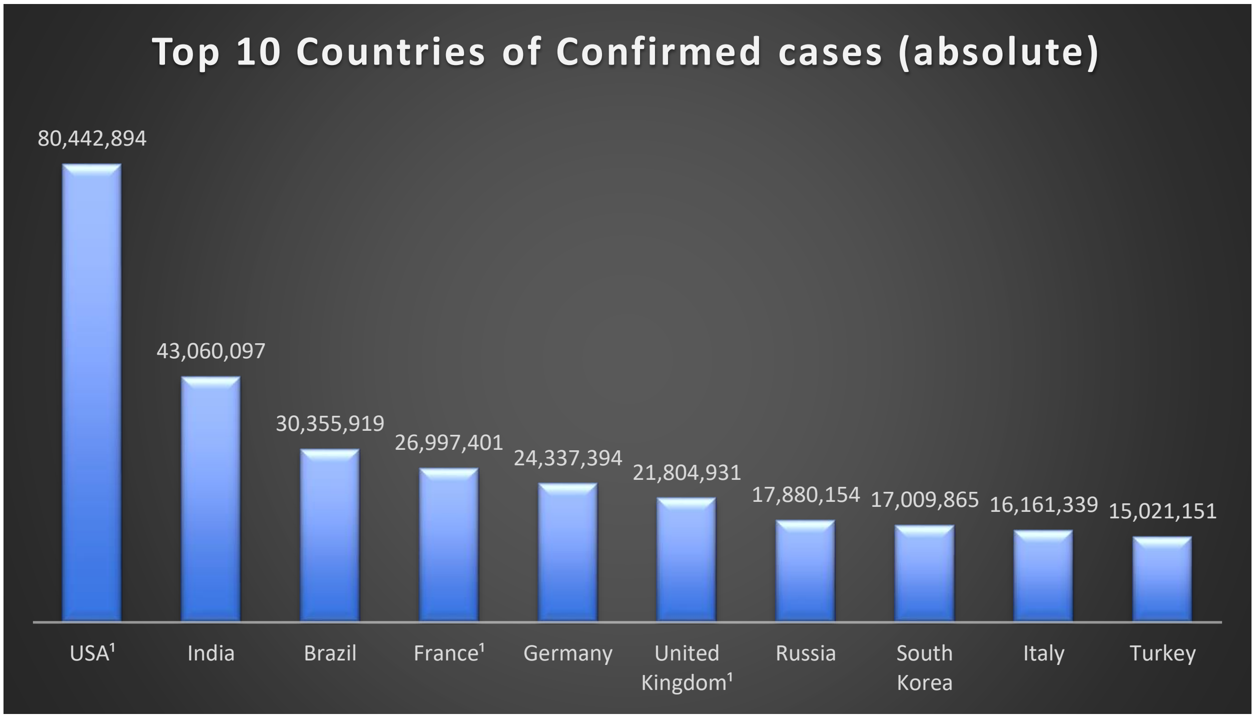
Source: World Health Organization

▨ Data may be incomplete for the current day or week.

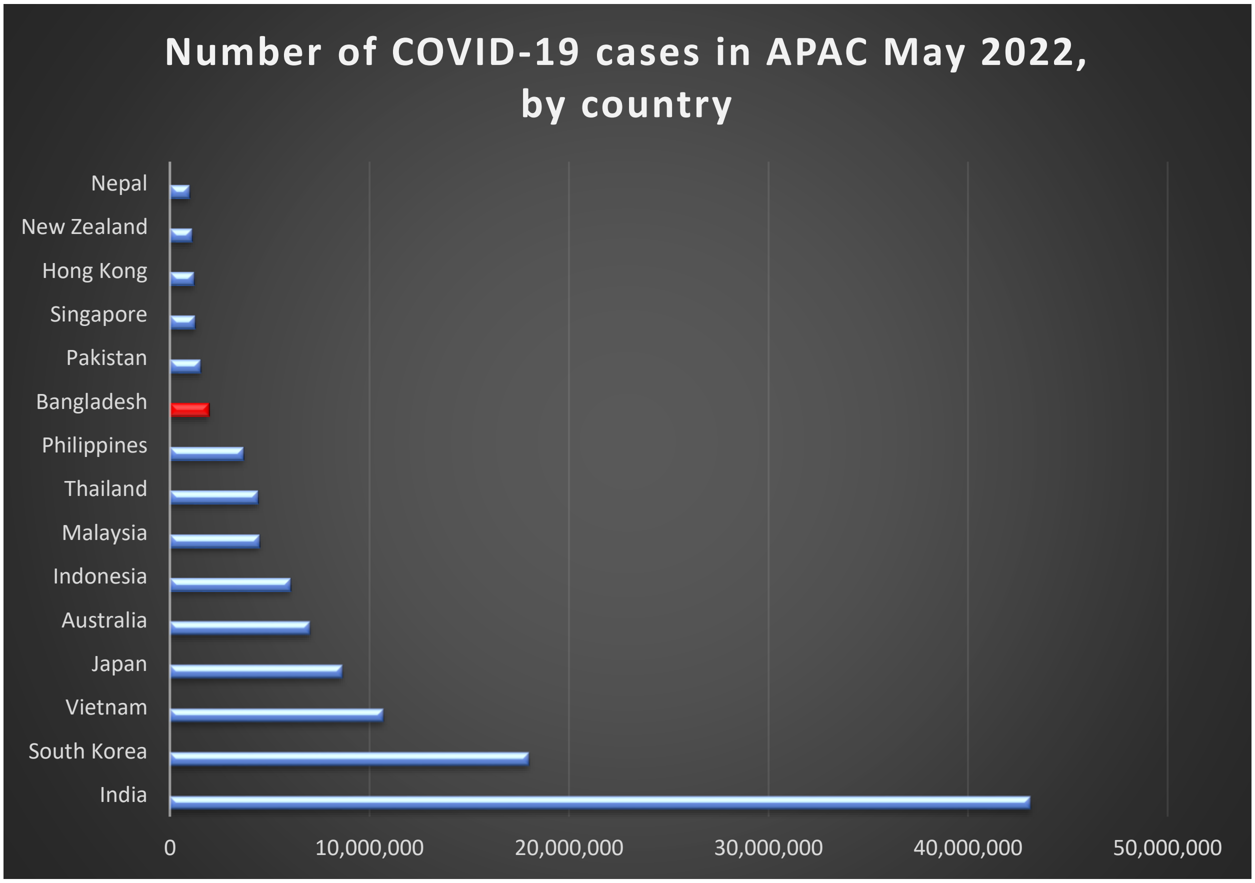
Also from Johns Hopkins University (CSSE); World Bank and Worldometer we get-



Coronavirus (COVID-19) cases, recoveries, and deaths worldwide as of May 16, 2022

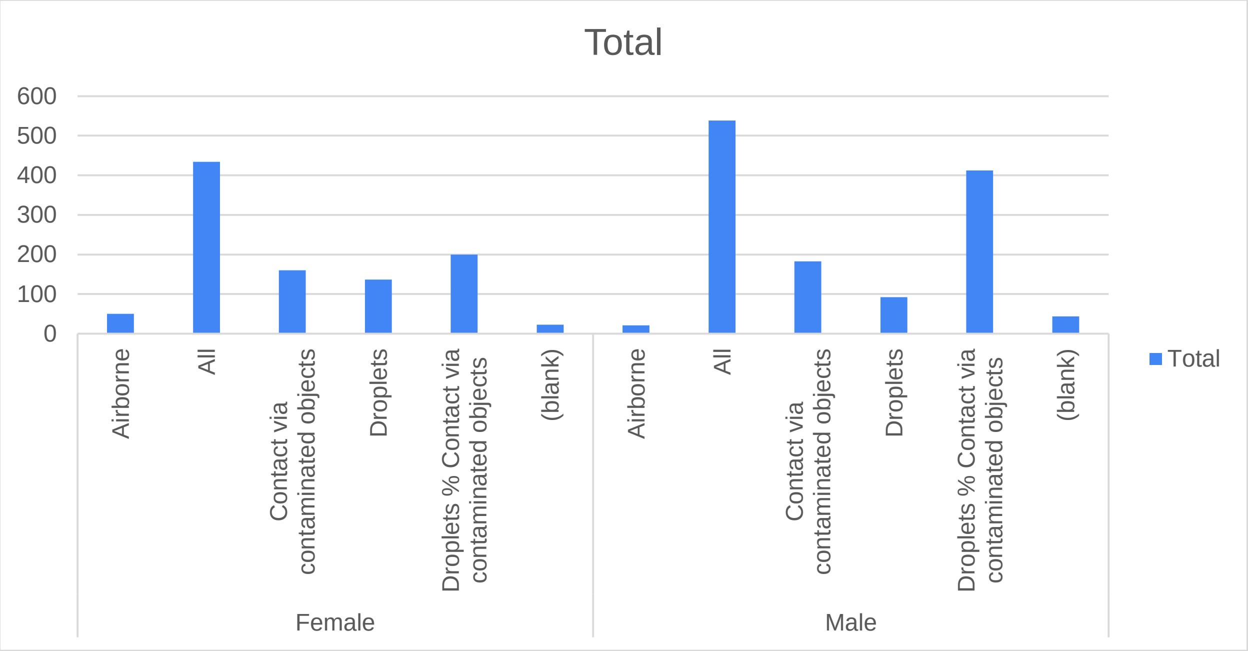


But the number of covid-19 cases in APAC May 2022-



SURVEY REPORT

Here is the percentage of what people think about how does the COVID-19 transmit through



Here is the percentage of few other survey questions:

Do you think a "mask" can reduce the risk of getting infected with coronavirus?

Yes – 86%

No – 14%

The percentage of carrying sanitizer as a precautionary step for the COVID-19 is

Yes - 73%

No – 27%

The percentage of infected family members with the virus from the respondent is

Getting infected – 39%

Not infected - 61%

Percentage of tested positive for COVID-19

Yes – 23%

No – 77%

CHAPTER SEVEN - CONCLUSION

LIMITATIONS OF RESEARCH

This topic shows the data analysis and prediction of infected cases of COVID-19. There were many challenges of this research. Firstly, it is a topic where newer and newer data is being generated each day around the world. Yet, a dataset of almost one lakh entries has been chosen to do the analytical research on the data. It was a challenge to predict the cases as newer variant of the virus was being mutated almost every day. That’s why, predicting cases was a bit challenging because all on a sudden, we have seen the peak is changing abruptly in real life. So, it was a bit difficult. Rather from this, there were only prediction of the infected cases. It would have been better if there are more features. As more features will tend us to do more accurate and real time analysis of the data. There can be added another limitation that is only prediction on newer cases can be achieved, but apart from this the level of the infected or more features like clustering on the data depending on the infected cases, or death rate vs the infected rate etc. could have been done. Time is the main limitation because if more time has been provided then, the research could have been conducted keeping more aspects on mind like then the above-mentioned limitations could have been overcome.

FUTURE SCOPE

There can be many future scope of this research topic. Firstly, on the targeted dataset, a few Machine learning models have been applied to record the desired accuracy. Many more efficient ML models can also be applied to have a comparative analysis on the results. Apart from ML models, deep learning algorithm and models can also be applied to get a sight of the accuracy. Thus, a perfect comparative analysis can be achieved. Also with this analysis, any similar variant disease can also be predicted. In the website, or using many countries data, it can be observed that there are image data and cluster data. Image processing and unsupervised data can also be taken into account to have a more accurate analysis. Apart from these few features, many more features can be collected and added to the data to make the dataset more informative. This research can be extended on adding deep learning analysis and this type of models give much more precise accurate result so in future prediction on similar diseases can be done if a best accuracy can be derived. Apart from the built in ML models and deep learning models, a customized model or hybrid model can also be developed depending on the data set to do a liberal analysis on the data. This type of future scope can be achieved on this research topic.

CONCLUSION

The scale of the COVID-19 epidemic, as well as the global lockdown consequences, are still unknown. However, as events unfold, there has been a rapid fall in social connections, a looming global economic downturn, deaths, and a growing fear of the "unknown," all of which have resulted in a shift in the status quo. Furthermore, the COVID-19 pandemic has had far-reaching consequences around the world, including a significant strain on various countries' healthcare systems, deaths, and other diseases/health difficulties. Apart from the COVID-19's health implications, there are other consequences of the subsequent lockdown that have impacted the world, which this review has described across numerous life stages.

REFERENCE

1. M. Zivkovic , N. Bacanin , K. Venkatachalam , A. Nayyar ,A. Djordjevic , I. Strumberger, F. A. Turjman, “COVID-19 cases prediction by using hybrid machine learning and beetle antennae search approach” , 2021.
2. F. A. B. Hamzaha, C. H. Laub, H. Nazric, D. V. Ligotd, G. Leee, C. L. Tanf, M. K. B. M. Shaibg, U. H. B. Zaidonh, A. B. Abdullahi, M. H. Chungj, C. H. Ongk, P. Y. Chewl and R. E. Salungam, “Corona Tracker: World-wide COVID-19 Outbreak Data Analysis and Prediction”, 19 March 2020.
3. R. Vaishya , M. Javaid , I. H. Khan , A. Haleem, “Artificial Intelligence (AI) applications for COVID-19 pandemic” , 2020.
4. S. Lalmuanawma, J. Hussain, L. Chhakchhuak, “Applications of Machine Learning and Artificial Intelligence for Covid-19 (SARS-CoV-2) pandemic: A review”, 23 June 2020
5. S. Kushwaha, S. Bahl, A. K. Bagha, K. S. Parmar, M. Javaid, A. Haleem, R. P. Singh, “Significant Applications of Machine Learning for COVID-19 Pandemic”, 28 October 2020
6. K. B. Prakash, S. S. Imambi, M. Ismail, T. P. Kumar, Y. N. Pawan, “Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms”, 5 May 2020
7. H. Yun, Z. Sun, J. Wu, A. Tang, M. Hu, Z. Xiang, “Laboratory data analysis of novel coronavirus (COVID-19) screening in 2510 patients” , 16 April 2020
8. E Ong, MU Wong, A Huffman, Y.He, COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. bioRxiv. 2020.doi:10.1101/2020.03.20.000141.
9. Y. Zoabi, S. Deri-Rozov, N. Shomron, Machine learning-based prediction of COVID-19 diagnosis based on symptoms, 2021, npj Digital Medicine. 4: 1-5
10. Ripa, S. P., Islam, F., & Arifuzzaman, M. (2021, July). The Emergence Threat of Phishing Attack and The Detection Techniques Using Machine Learning Models. In *2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI)* (pp. 1-6). IEEE.