Anna University Regional Campus Coimbatore

Anna University: Chennai-600 025

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING



IBM Naan Mudhalvan Phase 4 Submission

**Development Part-I**

**Title**: Covid 19 Vaccine Analysis

Name : THIRUKUMARAN G

Register Number : 710021106038

Department : ECE

Sem/year : 5,3year

Covid 19 Vaccine Analysis

**Objective:**

COVID 19 pandemic caused due to the Corona virus devastated the world by causing several fatalities around the world. This virus originated in Wuhan, China in 2019 and was later spread throughout the world due to human contact in one way or the other. The disease showed symptoms as basic as mild fever and cold but also caused life threatening symptoms like breathing problems caused by damage to the lungs. As this virus was new to the world and there was no vaccine or cure to it at the initial period there were several deaths around the world. The countries around the world were forced to shut themselves to others in order to avoid the further spread of the virus and people were stuck inside their houses and faced many issues with their finances, mental health etc., and felt like animals in a cage. An effort was made to find a cure or vaccine by several health organizations to bring a stop to this pandemic.

**Description of dataset:**

The link for the chosen dataset is mentioned below,

<https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress>

The above dataset contains Covid 19 Vaccine Analysis.

The Google Health COVID-19 Open Data Repository is one of the most comprehensive collections of up-to-date COVID-19-related information. Comprising data from more than 20,000 locations worldwide, it contains a rich variety of data types to help public health professionals, researchers, policymakers and others in understanding and managing the virus.

**1.Data Importing**

In power BI desktop with the help of the get data option import the CSV data which is named as country\_vaccinations and clicked load option.

**2.Data Cleaning**

After loading the data and after analyzing the data | understood that there are 86512 rows and 15 columns. And in that some of the columns contained null values I have replaced the null values by 0 with the use of replace functions and started working on the data.

**3.Visualizations**

In visualization part with the help of power BI desktop software I have used different kinds of charts, graphs, cards and table to display the data in the format which will be easy to understand.

**4.Insights**

Here we analyzed the top 10 fully vaccinated countries in which India tops the list which indicates that people in the country where showing lots of interests to get vaccinated.

* And also analyzed top 5 vaccinated countries here also India tops the list.
* And then analyzed top 5 daily vaccinating countries and here China tops the list.
* And also we analyse the sum of daily vaccinating details, fully vaccinating and vaccinating people details.
* And our year wise analyse shows that 2021 was the peak year for every vaccination details.

**5.DNA vaccines**

DNA vaccines can enter cells like viral infections and use the host protein translation system to generate target antigens. As an endogenous immunogen, it can induce humoral and cellular immune responses at the same time. Given the advantages of nucleic acid vaccines, DNA vaccines do not require live viruses, so safety is improved. DNA vaccines insert genes encoding foreign antigens into plasmids containing eukaryotic expression elements and then directly introduce the plasmids into humans or animals, allowing them to express antigen proteins in host cells and induce immune responses to prevent diseases.

**6.Inactivated vaccines**

Inactivated vaccines are the most classic form of vaccines. They are easy to prepare and can efficiently cause humoral immune responses. They are often the first choice for new infectious diseases. Inactivated vaccines are mainly obtained through three inactivation methods, such as formaldehyde, β-propiolactone, and ultraviolet. SARS and MERS inactivated vaccines can cause mice, hamsters, ferrets, and monkeys to produce high-titer neutralizing antibodies. The SARS-inactivated vaccine has completed phase I clinical trials, proving that it is safe in humans and can induce neutralizing antibodies’ production.

**7.Subunit vaccines**

Subunit vaccines are composed of purified recombinant proteins and are considered to be the safest vaccines. There are currently several subunit vaccines on the market, including hepatitis B, hepatitis E, and human papillomavirus vaccines. SARS and MERS subunit vaccines can produce high-titer neutralizing antibodies in mice, and nasal or oral vaccination can also induce a mucosal immune response, thereby more effectively blocking the virus transmission through the respiratory tract. The data also prove the protective efficacy of mucosal vaccination better than intramuscular inoculation.

**8.Travel immunization**

If the epidemic situation is well controlled, and the future epidemic situation is mainly imported, entry and exit personnel should be the target of implementing the immunization strategy, and close contacts of entry personnel should be used as vaccinations.

**9.Pre-exposure immunization**

For subjects who may be exposed to COVID-19 patients or high-risk infections, such as medical staff in fever clinics, COVID-19 pathogen testing personnel, contact persons from COVID-19 endemic countries, etc., should take exposure pre-immune prevention strategies.

**Model Training:**

* Choose a machine learning algorithm
* There are number of different machine learning algorithms that can be used for air quality analysis such as linear regression, KNN, Lasso Regression and Random Forests.

**Model Evaluation:**

* Model Evaluation is the process of assessing the performance of a machine learning model on unseen data. This is important to ensure that the model will generalize well to the new data.
* There are number of different metrics that can be used to evaluate the performance of air quality analysis and prediction model
* Some of the most common metrics include:
* **Mean Squared Error (MSE):** This metric measures the average squared difference between the different areas/cities in the World.
* **Root Mean Squared Error (RMSE):** This metric is the square root of the MSE
* **Mean Absolute Error (MAE):** This metric measures the average absolute difference between the different areas/cities in the World.
* **R-Squared:** This metric measures how well the model explains the variation in the different areas.

## **What is linear regression?**

Linear regression is a data analysis technique that predicts the value of unknown data by using another related and known data value. It mathematically models the unknown or dependent variable and the known or independent variable as a linear equation. For instance, suppose that you have data about your expenses and income for last year. Linear regression techniques analyze this data and determine that your expenses are half your income. They then calculate an unknown future expense by halving a future known income.

## **Why is linear regression important?**

Linear regression models are relatively simple and provide an easy-to-interpret mathematical formula to generate predictions. Linear regression is an established statistical technique and applies easily to software and computing. Businesses use it to reliably and predictably convert raw data into business intelligence and actionable insights. Scientists in many fields, including biology and the behavioral, environmental, and social sciences, use linear regression to conduct preliminary data analysis and predict future trends. Many data science methods, such as machine learning and artificial intelligence, use linear regression to solve complex problems.

## **How does linear regression work?**

At its core, a simple linear regression technique attempts to plot a line graph between two data variables, x and y. As the independent variable, x is plotted along the horizontal axis. Independent variables are also called explanatory variables or predictor variables. The dependent variable, y, is plotted on the vertical axis. You can also refer to y values as response variables or predicted variables.

### ****Steps in linear regression****

For this overview, consider the simplest form of the line graph equation between y and x; y=c\*x+m, where c and m are constant for all possible values of x and y. So, for example, suppose that the input dataset for (x,y) was (1,5), (2,8), and (3,11). To identify the linear regression method, you would take the following steps:

1. Plot a straight line, and measure the correlation between 1 and 5.
2. Keep changing the direction of the straight line for new values (2,8) and (3,11) until all values fit.
3. Identify the linear regression equation as y=3\*x+2.
4. Extrapolate or predict that y is 14 when x is

## **What is linear regression in machine learning?**

In machine learning, computer programs called algorithms analyze large datasets and work backward from that data to calculate the linear regression equation. Data scientists first train the algorithm on known or labeled datasets and then use the algorithm to predict unknown values. Real-life data is more complicated than the previous example. That is why linear regression analysis must mathematically modify or transform the data values to meet the following four assumptions.

### ****Linear relationship****

A linear relationship must exist between the independent and dependent variables. To determine this relationship, data scientists create a scatter plot—a random collection of x and y values—to see whether they fall along a straight line. If not, you can apply nonlinear functions such as square root or log to mathematically create the linear relationship between the two variables.

### ****Residual independence****

Data scientists use residuals to measure prediction accuracy. A residual is the difference between the observed data and the predicted value. Residuals must not have an identifiable pattern between them. For example, you don't want the residuals to grow larger with time. You can use different mathematical tests, like the Durbin-Watson test, to determine residual independence. You can use dummy data to replace any data variation, such as seasonal data.

### ****Normality****

Graphing techniques like Q-Q plots determine whether the residuals are normally distributed. The residuals should fall along a diagonal line in the center of the graph. If the residuals are not normalized, you can test the data for random outliers or values that are not typical. Removing the outliers or performing nonlinear transformations can fix the issue.

### ****Homoscedasticity****

Homoscedasticity assumes that residuals have a constant variance or standard deviation from the mean for every value of x. If not, the results of the analysis might not be accurate. If this assumption is not met, you might have to change the dependent variable. Because variance occurs naturally in large datasets, it makes sense to change the scale of the dependent variable. For example, instead of using the population size to predict the number of fire stations in a city, might use population size to predict the number of fire stations per person.

## **What are the types of linear regression?**

Some types of regression analysis are more suited to handle complex datasets than others. The following are some examples.

### ****Simple linear regression****

Simple linear regression is defined by the linear function:

Y= β0\*X + β1 + ε

β0 and β1 are two unknown constants representing the regression slope, whereas ε (epsilon) is the error term.

You can use simple linear regression to model the relationship between two variables, such as these:

* Rainfall and crop yield
* Age and height in children
* Temperature and expansion of the metal mercury in a thermometer

### ****Multiple linear regression****

In multiple linear regression analysis, the dataset contains one dependent variable and multiple independent variables. The linear regression line function changes to include more factors as follows:

Y= β0\*X0 + β1X1 + β2X2+…… βnXn+ ε

As the number of predictor variables increases, the β constants also increase correspondingly.

 Multiple linear regression models multiple variables and their impact on an outcome:

* Rainfall, temperature, and fertilizer use on crop yield
* Diet and exercise on heart disease
* Wage growth and inflation on home loan rates

### ****Logistic regression****

Data scientists use logistic regression to measure the probability of an event occurring. The prediction is a value between 0 and 1, where 0 indicates an event that is unlikely to happen, and 1 indicates a maximum likelihood that it will happen. Logistic equations use logarithmic functions to compute the regression line.

These are some examples:

* The probability of a win or loss in a sporting match
* The probability of passing or failing a test
* The probability of an image being a fruit or an animal

**Program:**

#importing files

from sklearn.model\_selection import train\_test\_split

import pandas as pd

from sklearn.impute import SimpleImputer

from sklearn.preprocessing import StandardScaler

import matplotlib.pyplot as plt

import numpy as np

import seaborn as sns

from sklearn import datasets

from sklearn.preprocessing import LabelBinarizer

from sklearn.linear\_model import LinearRegression

#readfile

df=pd.read\_csv("/content/country\_vaccinations.csv")

df.head()

x=df[['total\_vaccinations','people\_vaccinated']]

x

y=df[['daily\_vaccinations', 'total\_vaccinations\_per\_hundred']]

y

#simpleimputer

imputer=SimpleImputer(missing\_values=np.nan,strategy="mean")

imputer=imputer.fit(x)

x=imputer.transform(x)

x

imputer=SimpleImputer(missing\_values=np.nan,strategy="mean")

imputer=imputer.fit(y)

y=imputer.transform(y)

y

#one\_hot\_encoding

one\_hot=LabelBinarizer()

one\_hot.fit\_transform(x)

one\_hot.fit\_transform(y)

xtrain,xtest,ytrain,ytest=train\_test\_split(x,y)

plt.scatter(x,y)

plt.plot(xtest,ypred)

lr=LinearRegression()

model=lr.fit(xtrain,ytrain)

ypred=model.predict(xtest)

from sklearn.metrics import mean\_absolute\_error,mean\_squared\_error,r2\_score

mean\_absolute\_error(ytest,ypred)

#root mean squared error

from math import sqrt as sqrt

sqrt(mean\_squared\_error(ytest,ypred))

r2\_score(ytest,ypred)

## **CONCLUSION**

In short, the successful development of the COVID-19 vaccine concerns almost all countries and people in the world. We must do an excellent job of researching the immunogenicity and immune reactivity of the vaccines.