Project: Covid-19 Vaccine Analysis
Phase-2: Innovation

Synopsis:

- **Question**
- **Dataset and it's details**
- **Columns**
- **Libraries to be used and ways to download**
- **Train and Test**
- **Metrices for the accuracy check**
- **Conclusion**

Question:

Consider exploring advanced machine learning techniques like clustering or time series forecasting to uncover hidden patterns in vaccine distribution and adverse effect datas.

Dataset and it's Detail:

The dataset "COVID-19 World Vaccination Progress" on Kaggle is a collection of data related to the COVID-19 vaccination efforts worldwide. It provides information about the progress of COVID-19 vaccinations in various countries and regions. This dataset is designed to help researchers, data scientists, and analysts understand and analyze the progress of COVID-19 vaccination campaigns across different countries. A second file, with manufacturers information, is included. Below is a detailed overview of the dataset:

Title: COVID-19 World Vaccination Progress

Dataset ID: gpreda/covid-world-vaccination-progress

Source: The dataset was created by a Kaggle user named Gabriel Preda, collected from various sources, including government health agencies, international organizations, and research institutions.

Description:

1. The dataset provides information about the COVID-19 vaccination progress from various countries around the world.
2. It includes data on vaccine distribution, vaccination coverage, and other related statistics.
3. The dataset may include information about the types of vaccines used, vaccination rates over time, and population demographics.

Columns/Attributes:

1. The dataset typically contains columns such as country, iso_code, date, total_vaccinations, people_vaccinated, people_fully_vaccinated, daily_vaccinations_raw, daily_vaccinations, and more.

2.These columns provide information about the total number of vaccinations, daily vaccination rates, and other vaccination-related metrics for each country.

Usage:

1. Analyzing vaccination progress over time for different countries.
2. Identifying countries with high vaccination rates or disparities.
3. Forecasting future vaccination trends.
4. Studying the impact of different vaccines on vaccination rates.
5. Correlating vaccination progress with COVID-19 infection and mortality rates.

Data Format:

The data is usually structured as a CSV (Comma-Separated Values) file, with rows representing different countries or regions and columns representing various attributes related to vaccination progress and population.

Updates:

The dataset may be updated regularly to reflect the latest vaccination data, making it useful for tracking changes and trends over time.

Columns:

- **Country-** this is the country for which the vaccination information is provided.
- **Country ISO Code -** ISO code for the country.
- **Date -** date for the data entry; for some of the dates we have only the daily vaccinations, for others, only the (cumulative) total.
- **Total number of vaccinations -** this is the absolute number of total immunizations in the country.
- **Total number of people vaccinated -** a person, depending on the immunization scheme, will receive one or more (typically 2) vaccines; at a certain moment, the number of vaccinations might be larger than the number of people.
- **Total number of people fully vaccinated -** this is the number of people that received the entire set of immunization according to the immunization scheme (typically 2); at a certain moment in time, there might be a certain number of people that received one vaccine and another number (smaller) of people that received all vaccines in the scheme.
- **Daily vaccinations (raw)** - for a certain data entry, the number of vaccinations for that date/country.
- **Daily vaccinations -** for a certain data entry, the number of vaccinations for that date/country.
- **Total vaccinations per hundred -** ratio (in percent) between vaccination number and total population up to the date in the country.

- **Total number of people vaccinated per hundred -** ratio (in percent) between population immunized and total population up to the date in the country.
- **Total number of people fully vaccinated per hundred -** ratio (in percent) between population fully immunized and total population up to the date in the country.
- **Number of vaccinations per day -** number of daily vaccinations for that day and country.
- **Daily vaccinations per million -** ratio (in ppm) between vaccination number and total population for the current date in the country.
- **Vaccines used in the country -** total number of vaccines used in the country (up to date);
- **Source name -** source of the information (national authority, international organization, local organization etc.).
- **Source website -** website of the source of information.

There is a second file added (country vaccinations by manufacturer), with the following columns:

- **Location -** country.
- **Date -** date.
- **Vaccine -** vaccine type.
- **Total number of vaccinations -** total number of vaccinations / current time and vaccine type.

Libraries to be used and way to download it:

To work with the "COVID-19 World Vaccination Progress" dataset from Kaggle, the libraries and tools used are commonly used in data analysis and machine learning. Here are the key libraries and steps to download, explore, train, and test the dataset using Python:

- **Python:** Python is a popular programming language for data analysis and machine learning. Ensure you have Python installed on your system.
- **Pandas:** Pandas is a Python library used for data manipulation and analysis. You can use Pandas to read and manipulate your dataset.
- **NumPy:** NumPy is a library for numerical operations. It's often used in conjunction with Pandas for data manipulation.
- **Matplotlib and Seaborn:** These libraries are used for data visualization. You can create plots and charts to visualize your data using these libraries.
- **Scikit-Learn:** Scikit-Learn is a machine learning library that provides tools for data preprocessing, model training, and model evaluation.

1. Download the Dataset:

- Visit the Kaggle dataset page: COVID-19 World Vaccination Progress.
- Download the dataset from Kaggle to your local machine.

2.Download Libraries:

Download and install Python libraries using package management tools like pip depending on Python environment. Here are the general steps to download and install Python libraries:

Using pip:

- **Open a Command Prompt (Windows) or Terminal (for python editors):**
  The command line is used to run the installation commands.

- **Install a Python Library:**

  To install a specific library, use the pip install command followed by the library's name. For example, to install the Pandas library, you can run:

  pip install pandas

- **Verify the Installation:**

  You can verify that the library has been installed correctly by importing it in a Python script or in a Python interactive environment:

  import pandas

3.Import Libraries:

Import the necessary Python libraries, including Pandas, NumPy, Matplotlib, Seaborn, and Scikit-Learn, in your Python editor or Python script.

Train and Test:

Training and testing a machine learning model on the "COVID-19 World Vaccination Progress" dataset from Kaggle involves several steps, including data preparation, feature selection, model selection, training, and evaluation. Here is a general outline of the process:

1.Data Preparation:

- Load the dataset into a Pandas Data Frame
- Clean and preprocess the data by handling missing values and encoding categorical variables if necessary.

2.Feature Selection:

- Decide which features (columns) you want to use as inputs for your model. You might choose columns like total_vaccinations, people_vaccinated_per_hundred, and daily_vaccinations_per_million as potential features.

3.Split Data into Training and Testing Sets:

- Split the dataset into a training set and a testing (or validation) set. Typically, you reserve a portion of the data for testing to evaluate your model's performance.
- Scikit-Learn's train_test_split function for this purpose.

4.Select a Model:

- Depending on your task (e.g., regression or classification), select an appropriate machine learning model. For example, you might use a linear regression model for regression tasks or a decision tree for classification tasks.
- Import the relevant model from Scikit-Learn.

5.Train the Model:

- Fit your selected model to the training data using the training features and target variable (the variable you want to predict). For instance:

  from sklearn.linear_model import LinearRegression

```
# Create a model instance

model = LinearRegression()

# Train the model on the training data

 model.fit(X_train, y_train)
```

6.Make Predictions:

- Use the trained model to make predictions on the testing data. For regression tasks, you will predict numerical values, and for classification tasks, you'll predict class labels.
- Fit your selected model to the training data using the training features and target variable (the variable you want to predict). For instance:

```
# Make predictions on the testing data
y_pred = model.predict(X_test)
```

7. Evaluate the Model:

- Assess the model's performance using appropriate evaluation metrics. The choice of metrics depends on your specific task (e.g., mean squared error for regression or accuracy for classification).

```
from sklearn.metrics import mean_squared_error

# Calculate the mean squared error

mse = mean_squared_error(y_test, y_pred)
```

8. Visualization and Reporting:

- Visualize the model's predictions and errors to gain insights.

Metrices used for the accuracy check:

- Regression Tasks
- Classification Tasks
- Clustering Tasks

Conclusion:

In this project, we embarked on an analysis of the "COVID-19 World Vaccination Progress" dataset available on Kaggle. Our primary objective was to develop a machine learning model that could provide insights into the progress of COVID-19 vaccination campaigns in various countries and regions. This endeavor allowed us to draw valuable conclusions and make informed predictions regarding vaccination coverage, an essential aspect of our ongoing battle against the pandemic.

Teammates: Rithika B

Sowmiya G