Machine Learning Techniques in COVID-19 Data Analysis

# 1. ARIMA for Time Series Forecasting

**Purpose:** To predict future values of total COVID-19 cases.  
**Methodology:**   
**- Data Preparation:** The dataset is grouped by date, summing the total cases.  
**- Model:** Used the **'auto\_arima'** function to automatically select the best ARIMA model.  
**- Forecasting:** The model predicts the next 30 days of total cases.  
**Reason for Choice:** ARIMA is effective for time series forecasting, especially when data shows a clear trend or seasonality, which is common in pandemic data.

# 2. PCA (Principal Component Analysis)

**Purpose:** To reduce the dimensionality of the dataset while retaining most of the variance. It simplifies the complexity in high-dimensional data while retaining trends and patterns.  
**Methodology:**   
**- Data Selection:** Chose 'total\_cases\_per\_million' and 'people\_vaccinated\_per\_hundred' for PCA.  
**- Standardization:** Used **'StandardScaler'** to standardize the data.  
**- PCA Application:** Applied PCA to reduce dimensions to 2 principal components.  
**- Explained Variance**: The proportion of dataset's variance that lies along each principal component.  
**Reason for Choice:** PCA helps in visualizing high-dimensional data and can uncover hidden patterns.

# 3. K-Means Clustering

**Purpose:** To segment the data into clusters for further analysis.  
Methodology:   
**- Optimal Clusters:** Used the **Elbow method** to find the optimal number of clusters.  
**- Clustering:** Applied K-Means clustering on the PCA results.  
**Reason for Choice:** K-Means is a widely-used clustering technique that is efficient and effective for identifying distinct groups in the data.

# Comparison of Machine Learning Techniques

**1. ARIMA for Time Series Forecasting**  
**- Effectiveness:** Highly effective for forecasting time-series data.  
**- Accuracy:** Good for short-term forecasting, dependent on data patterns.  
**- Output:** Predicted total COVID-19 cases for the next 30 days.  
  
**2. PCA (Principal Component Analysis)**  
**- Effectiveness:** Excellent for reducing dimensionality and uncovering patterns.  
**- Accuracy**: High if the explained variance ratio is high for the principal components used.  
**- Output:** Transformed dataset with reduced dimensions.  
  
**3. K-Means Clustering**  
**- Effectiveness:** Great for segmenting data into distinct groups.  
**- Accuracy:** Depends on initial centroids and data distribution.  
**- Output:** Segmentation of the data into optimal clusters.  
  
**Best Technique for the Project**  
**- For Forecasting:** ARIMA.  
**- For Data Reduction and Pattern Discovery:** PCA.  
**- For Data Segmentation and Group Analysis:** K-Means Clustering.