**COVID Vaccines Analysis**

**Phase- 2 (Innovation)**

**Title:**

Exploring advanced machine learning techniques like clustering or time series forecasting to uncover hidden patterns in vaccine distribution and adverse effects data.

**Dataset link:**

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

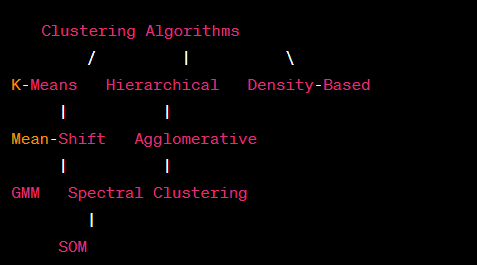
**Abstract:**

The rapid distribution and administration of vaccines are of paramount importance in the global effort to combat infectious diseases, and the comprehensive analysis of vaccine distribution data can provide valuable insights for optimizing public health strategies. With the ongoing challenges posed by the COVID-19 pandemic and the necessity to manage vaccine supply chains effectively, the utilization of advanced machine learning techniques has emerged as a key innovation.

This document explores the application of cutting-edge machine learning methodologies, particularly clustering and time series forecasting, to analyze vaccine distribution and adverse effects data. We aim to uncover hidden patterns, gain a deeper understanding of distribution dynamics, and enhance decision-making processes.

**Key Objectives:**

1. **Clustering Analysis:** This document investigates how clustering techniques can help identify unique groups within the vaccine distribution dataset. By segmenting the data into meaningful clusters, we aim to improve the allocation of resources and optimize vaccine distribution strategies.

****

1. **Time Series Forecasting:** Time series forecasting enables us to predict future vaccine distribution trends, helping healthcare systems plan for contingencies and allocate resources more efficiently. By analyzing historical data, we can develop predictive models that offer valuable insights into the vaccine distribution process.



**Benefits:**

**.** **Improved resource allocation:** By understanding distribution patterns, authorities can allocate vaccines where they are needed most.

**.** **Enhanced planning:** Time series forecasting can help healthcare systems prepare for future demand and optimize distribution logistics.

**.** **Informed decision-making:** Data-driven insights enable public health officials to make well-informed decisions regarding vaccine distribution and management.

**Source Code:**

**1) Clustering:**

import pandas as pd

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

import warnings

import matplotlib.pyplot as plt

import seaborn as sns

# Load the vaccine distribution data

data = pd.read\_csv('country\_vaccinations.csv')

data\_copy = data.copy()

# Select relevant features for clustering

features = data[['daily\_vaccinations', 'daily\_vaccinations\_per\_million']]

# Data Preprocessing

# 1. Handling missing values (if any)

features.fillna(0, inplace=True)  # Replace missing values with zeros

# 2. Standardization (optional but recommended)

scaler = StandardScaler()

scaled\_features = scaler.fit\_transform(features)

# 3. Dimensionality Reduction (optional but recommended for high-dimensional data)

# Using Principal Component Analysis (PCA) to reduce dimensionality

pca = PCA(n\_components=2)  # Adjust the number of components as needed

reduced\_features = pca.fit\_transform(scaled\_features)

# At this point, 'reduced\_features' contains the preprocessed data suitable for clustering

from sklearn.cluster import KMeans

kmeans = KMeans(n\_clusters=3)

clusters = kmeans.fit\_predict(reduced\_features)

# Add cluster labels to the original DataFrame

data['Cluster'] = clusters

selected\_attributes = data[['daily\_vaccinations', 'daily\_vaccinations\_per\_million', 'Cluster']]

selected\_attributes.to\_csv('vaccine\_distribution\_clusters.csv', index=False)

# Print the cluster assignments

print(data[['country', 'daily\_vaccinations', 'daily\_vaccinations\_per\_million', 'Cluster']])

# Visualize the clusters

sns.scatterplot(x='daily\_vaccinations', y='daily\_vaccinations\_per\_million', hue='Cluster', data=data)

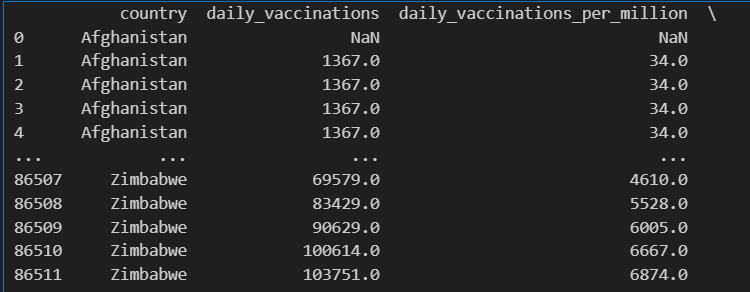
plt.xlabel('Daily Vaccinations')

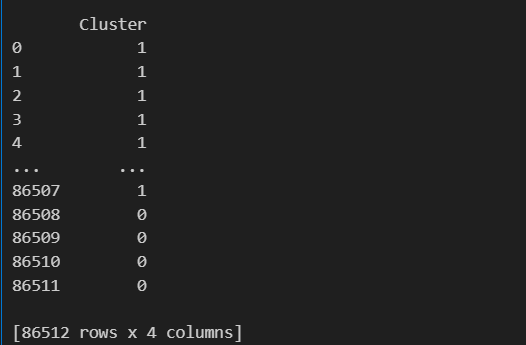
plt.ylabel('Daily Vaccinations per Million')

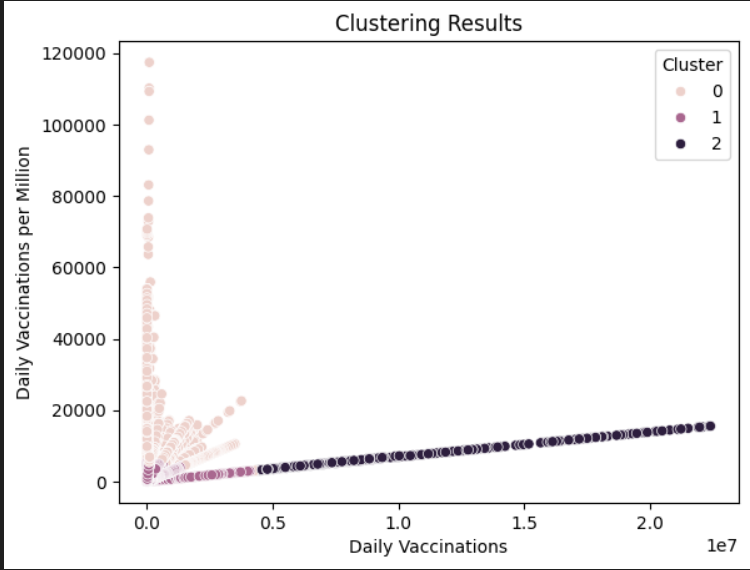
plt.title('Clustering Results')

plt.show()

**Output:**

****

****

****

**2) Time series forecasting:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

from statsmodels.tsa.arima.model import ARIMA

from statsmodels.graphics.tsaplots import plot\_acf, plot\_pacf

from statsmodels.tsa.stattools import adfuller

import warnings

warnings.filterwarnings('ignore')

# Load your time series data

data = pd.read\_csv('country\_vaccinations\_by\_manufacturer.csv')

data['date'] = pd.to\_datetime(data['date'])

data.set\_index('date', inplace=True)

# Check for stationarity

result = adfuller(data['total\_vaccinations'])

print(f'ADF Statistic: {result[0]}')

print(f'p-value: {result[1]}')

# If the data is non-stationary, difference it to make it stationary

if result[1] > 0.05:

    data['total\_vaccinations'] = data['total\_vaccinations'].diff().dropna()

# Plot the ACF and PACF to determine model orders (p and q)

plot\_acf(data['total\_vaccinations'])

plot\_pacf(data['total\_vaccinations'])

plt.show()

# Fit the ARIMA model

model = ARIMA(data['total\_vaccinations'], order=(1, 1, 1))  # Adjust p, d, and q as needed

model\_fit = model.fit()

# Forecast future values

forecast\_steps = 10  # Number of steps to forecast

forecast = model\_fit.forecast(steps=forecast\_steps)

# Create a date range for the forecasted values

forecast\_index = pd.date\_range(start=data.index[-1], periods=forecast\_steps + 1)

# Plot the original data and forecasted values

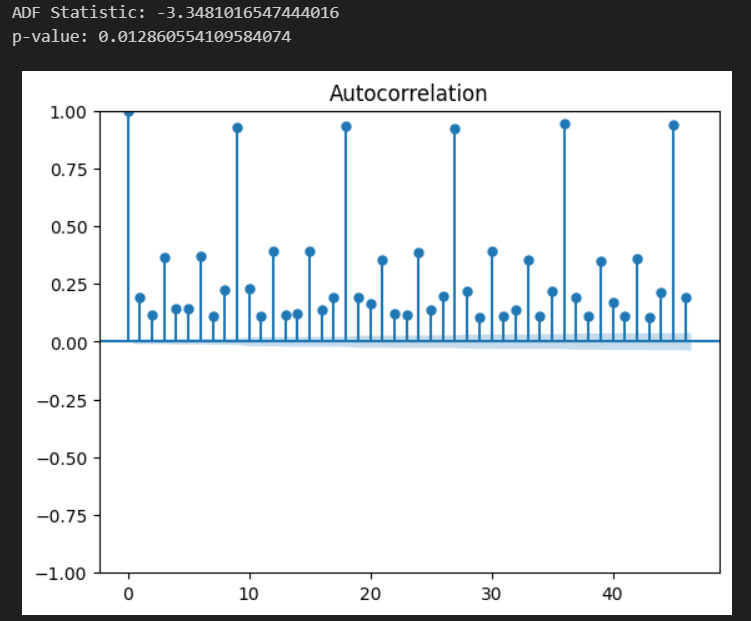
plt.plot(data['total\_vaccinations'], label='Original Data')

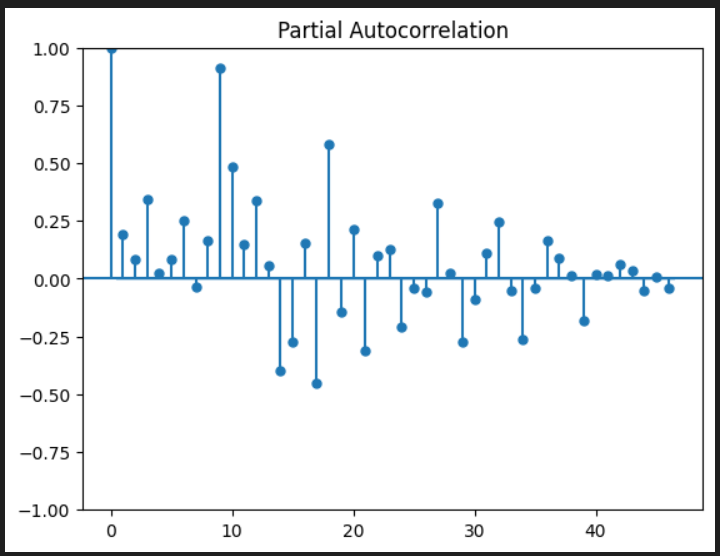
plt.plot(forecast\_index[1:], forecast, label='Forecast', color='red')

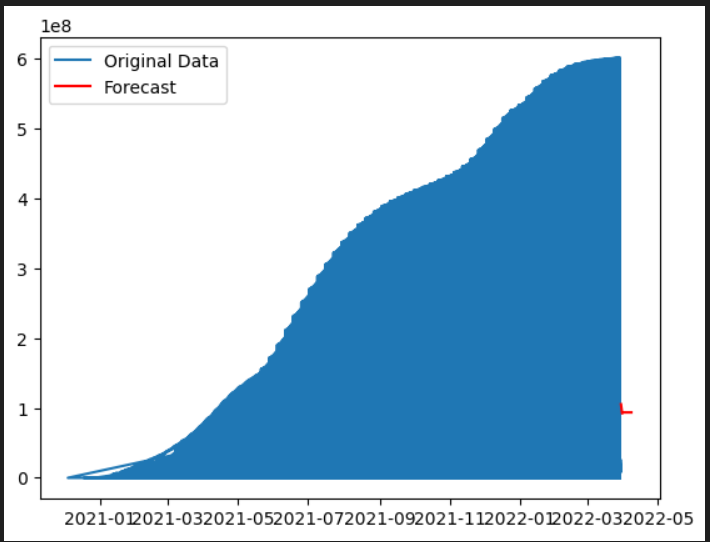
plt.legend()

plt.show()

**Output:**

****

****

****

**Conclusion:**

In the face of global health crises like the COVID-19 pandemic, advanced machine learning techniques offer innovative solutions to the challenges of vaccine distribution. This document explores the application of clustering and time series forecasting to uncover hidden patterns in vaccine distribution and adverse effects data. By doing so, we aim to contribute to more effective vaccine allocation, better planning, and informed decision-making in the realm of public health.