**COVID-19 Vaccine Data Analysis**

## **Introduction:**

The COVID-19 pandemic has emphasized the need for data-driven decision-making in vaccine deployment strategies. This project focuses on analyzing COVID-19 vaccine data to provide actionable insights for optimizing vaccine deployment. The analysis covers data preprocessing, feature selection, encoding, and splitting for further modeling.

**Step 1: Import Necessary Libraries**

**Python:**

**import** **pandas** **as** **pd**

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

**from** **sklearn.preprocessing** **import** StandardScaler

**from** **sklearn.preprocessing** **import** OneHotEncoder

- We begin by importing the required Python libraries. These libraries will be used for data manipulation, data splitting, and data transformation.

**Step 2: Load the Dataset**

**Python:**

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

- In this step, we load the dataset from a CSV file. We set the Dataset as `'country\_vaccinations.csv'` with the path to your dataset file.

**Step 3: Data Preprocessing**

**Python:**

**#Remove duplicates**

data = data.drop\_duplicates()

**# Handle missing values (example: fill missing numerical values with the mean)**

data.fillna(data.mean(), inplace=True)

- Data preprocessing starts with the removal of duplicate rows in the dataset.

- Missing values are addressed by filling missing numerical values with their respective column's mean.

**Step 4: Feature Selection**

**Python:**

selected\_features = ['feature1', 'feature2', 'categorical\_feature', 'target\_column']

**if** all(feature **in** data.columns **for** feature **in** selected\_features):

data = data[selected\_features]

**else**:

**print**("Selected columns do not exist in the dataset. Please verify column names.")

- Feature selection involves choosing specific columns or features that are relevant for the analysis.

- Ensure that the selected columns exist in the dataset to avoid errors.

**Step 5: Encoding Categorical Variables (One-Hot Encoding)**

**Python:**

encoder = OneHotEncoder(sparse=False, drop='first')

encoded\_categorical\_features = encoder.fit\_transform(data[['categorical\_feature']])

encoded\_categorical\_feature\_names = encoder.get\_feature\_names(['categorical\_feature'])

data\_encoded = pd.concat([data, pd.DataFrame(encoded\_categorical\_features, columns=encoded\_categorical\_feature\_names)], axis=**1**)

data\_encoded.drop(['categorical\_feature'], axis=**1**, inplace=True)

- Categorical variables are transformed into a numerical format using one-hot encoding.

- The one-hot encoded features are concatenated with the original dataset, and the original categorical feature is dropped.

**Step 6: Split Data into Training, Validation, and Test Sets**

**Python:**

train\_data, test\_data = train\_test\_split(data\_encoded, test\_size=0.2, random\_state=42)

validation\_data, test\_data = train\_test\_split(test\_data, test\_size=0.5, random\_state=42)

- The dataset is split into training, validation, and test sets.

- In this example, 80% of the data is used for training, 10% for validation, and 10% for testing.

**Step 7: Scaling Numerical Features (StandardScaler)**

**Python:**

scaler = StandardScaler()

train\_data[['feature1', 'feature2']] = scaler.fit\_transform(train\_data[['feature1', 'feature2']])

validation\_data[['feature1', 'feature2']] = scaler.transform(validation\_data[['feature1', 'feature2']])

test\_data[['feature1', 'feature2']] = scaler.transform(test\_data[['feature1', 'feature2']])

- Numerical features, 'feature1' and 'feature2,' are scaled using the StandardScaler. This step ensures that numerical features have a mean of 0 and a standard deviation of 1.

**Conclusion:**

The preprocessing steps outlined in this document ensure that the COVID-19 vaccine data is clean, relevant, and prepared for further analysis. These steps are vital for optimizing vaccine deployment strategies. The preprocessed data can now be used for machine learning model development and in-depth analysis to provide insights and recommendations for policymakers and health organizations.