**COVID-19 VACCINE ANALYSIS**

| **NAME** | **REG NO** | **EMAIL ID** |
| --- | --- | --- |
| **JEYASURYA S** | **312821243010** | **21aids11@act.edu.in** |

**PHASE 5 :** Submission Document



**Introduction:**

The ongoing COVID-19 pandemic has necessitated an in-depth analysis of various aspects of the situation, one of which being the efficacy and distribution of vaccines. This analysis aims to assist policymakers in optimizing strategies for vaccine deployment based on comprehensive data analysis.

**Problem Statement:**

The problem revolves around the need for an in-depth analysis of COVID-19 vaccine data with a focus on vaccine efficacy, distribution, and adverse effects. Through this analysis, insights will be derived to help decision-makers and health organization strategize for optimum vaccine rollout.

**Design Thinking Process:**

- **Emphasize:** Understanding the needs of the stakeholders, i.e., policymakers, healthcare providers, and the general public regarding vaccine distribution, efficacy, and side effects.

- **Design:** Formulating the problem statement accurately to address the needs of stakeholders.

- **Ideate:** Thinking of various strategies to collect, preprocess, and analyze the data to derive insights.

- **Prototype and Test:** Creating sample data models and testing them for accuracy, making necessary iterations for improvements.

- **Implement:** Deploying the successful models or strategies after rigorous testing.

- **Feedback and Improve:** Taking feedback from the stakeholders and improving the processes continually.

**Phases of Development:**

- **Data Collection:** Gathering structured and unstructured data relevant to COVID-19 vaccines from authentic sources.

- **Data Preprocessing:** Cleaning the data to handle missing values, normalize features, encode categorical variables, and maintain stationarity.

- **Feature Engineering:** Selecting the most crucial features that affect vaccine distribution and efficacy.

- **Model Selection, Training, & Hyperparameter Tuning:** Picking an ideal model to train the data and tweaking the parameters for optimum results.

- **Evaluation Metrics & Model Validation:** Using metrics like MAE, MSE, RMSE, and STL to assess the model's performance.

- **Deployment:** Implementing the model on a broader scale once it achieves satisfactory results in testing.

**Items in the dataset:**

* Countries
* ISO\_code
* Dates
* Vaccines
* Total Vaccinations
* Source Name
* Source Websites
* Daily Vaccinations

**Desired data to find:**

* Most commonly used vaccines in countries
* Average daily vaccination count in countries
* Number of countries where vaccines are used
* Choropleth map of the most used vaccine

TheCovidWorldVaccinationProgress(https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress) dataset on Kaggle is a comprehensive collection of information pertaining to the ongoing COVID-19 vaccination drive across various countries worldwide.

This dataset provides substantial insight into each country's trajectory in its vaccination campaign. Here are some of the main aspects and types of information you can find in this dataset:

**1. Country:** The name of the country for which the vaccination information is provided.

**2. Date:** The date for the corresponding entry.

**3. Total number of vaccinations:** This is the absolute count of total immunizations in the country.

**4. Total number of people vaccinated:** This is the total number of people who received at least one vaccine dose.

**5. Total number of people fully vaccinated:** This represents the number of people who received all doses prescribed by the vaccination protocol.

**6. Daily vaccinations (raw):** This is the count for vaccinations given on the actual day.

**7. Daily vaccinations:** This holds the count for vaccinations for a specific date, which may differ from 'daily vaccinations (raw)' due to certain calculation methods.

**8. Total vaccinations per hundred:** This is the ratio (in percent) of total vaccinations per total population up to the date in the country.

**9. Total number of people vaccinated per hundred:** The ratio (in percent) of population that received at least one vaccine dose.

**10. Total number of people fully vaccinated per hundred:** The ratio (in percent) indicating the total number of people fully vaccinated.

**Methodology:**

In this section, provide a detailed overview of the methods and algorithms used for the analysis. If linear algorithms were employed, explain why they were chosen and how they work. Additionally, mention any other methods or techniques used for data preprocessing or analysis, such as data cleaning, feature engineering, or machine learning models.

**Library:**

* Numpy
* Seaborn
* OS
* Pandas
* plotly
* Folium

**Training and Testing:**

1. Data Collection

2. Data Preprocessing

3. Feature Selection

1. Model Selection

5. Training and Testing Split

6. Model Training

7. Model Evaluation

8. Hyperparameter Tuning

9. Validation and Cross-Validation

1. Interpretability and Visualization

**Innovative Solutions:**

Detail the innovative solutions or approaches proposed to address the problem identified in the problem statement. Provide a step-by-step explanation of how each solution works and how it leverages the analysis and data to achieve the desired outcomes. If there are multiple solutions, present them individually and discuss their potential impact on the problem.

**Metrics for Accuracy Check:**

Common metrics in data analysis include accuracy, precision, recall, F1-score, and ROC curves. Describe why these metrics were chosen and how they are relevant to the specific problem you are addressing. Provide examples of how the metrics are calculated and interpreted in the context of your analysis.

**References:**

Include a list of references or sources of information used in the document. This helps validate the credibility of your analysis and allows readers to explore further if they are interested.

By expanding on these sections with more detailed explanations and possibly including charts, graphs, or data visualizations, your document will provide a comprehensive and informative analysis of COVID-19 vaccine data and innovative solutions.

# **Loading the dataset:**

**An explanation of each library's purpose and how they can be used in data analysis and visualization:**

**1. Pandas:**

Purpose: Pandas is a popular data manipulation and analysis library for Python. It provides data structures such as DataFrames and Series, making it easy to load, manipulate, and analyze tabular data.

**2. Plotly Express and Plotly Graph Objects:**

Purpose: Plotly is a library for interactive and visually appealing data visualization. Plotly Express provides a high-level API for creating plots, while Plotly Graph Objects offers a lower-level, more customizable approach.

**3. Folium:**

Purpose: Folium is a Python library for creating interactive maps. It integrates with the Leaflet JavaScript library and allows you to visualize geographical data.

**4. Seaborn:**

Purpose: Seaborn is a data visualization library built on top of Matplotlib. It provides a high-level interface for creating attractive and informative statistical graphics.

**import pandas as pd**

**import plotly.express as px**

**import plotly.graph\_objects as go**

**from folium.features import Choropleth**

**import folium**

**from folium.features import Tooltip**

**import seaborn as sns**

**Dataset Description:**

The dataset will include data points like manufacturer details, vaccine type, vaccination dates, demographic information of the recipients, adverse event reports, etc.

**Data Preprocessing Steps:**

Data preprocessing will involve multiple steps. Handling missing values, scaling features to bring them to a common scale, converting categorical data into numerical data for better processing, ensuring stationarity so the summary statistics remain consistent, and finally, extracting the most relevant features for the analysis.

**Model Training Process:**

The model training process will start with splitting data into training and testing sets to validate the model's performance. Upon selecting a suitable model, it's hyperparameters will be tuned for the best performance. Post-training the model with the data, we'll evaluate it using relevant evaluation metrics and validate the results.

**Choice of Time Series Forecasting Algorithm:**

Choosing a time series forecasting algorithm will depend on seasonality and trends in data, the interpretability of the model output, predictive accuracy, and the model's robustness against outliers or variance in data.

**Evaluation Metrics:** Define and explain the evaluation metrics used to assess the accuracy and performance of the models, such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and seasonal decomposition of time series (STL) metrics. Discuss their relevance and applicability in the context of the analysis.

#### **Model Evaluation Example Program:** Provide an example program that demonstrates how to evaluate the model's performance using the defined evaluation metrics. Include code snippets and outputs to showcase the evaluation process.

Data Preprocessing Program:

import numpy as np

import pandas as pd

import os

for dirname, \_, filenames in os.walk('/kaggle/input'):

for filename in filenames:

print(os.path.join(dirname, filename))

import pandas as pd

import plotly.express as px

import plotly.graph\_objects as go

from folium.features import Choropleth

import folium

from folium.features import Tooltip

import seaborn as sns

df = pd.read\_csv("/kaggle/input/covid-world-vaccination-progress/country\_vaccinations\_by\_manufacturer.csv")

df.head(10)

df["location"].nunique()

df.isnull().sum()

df.dtypes

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

data=pd.DataFrame(columns=['Country', 'Vaccine', 'Total\_vaccine'])

for country in df["location"].unique():

for vaccine in df["vaccine"].unique():

filtered\_data = df[(df['location'] == country) & (df['vaccine'] == vaccine)]

total\_count = filtered\_data['total\_vaccinations'].max()

data = pd.concat([data, pd.DataFrame({'Country': [country], 'Vaccine': [vaccine], 'Total\_vaccine': [total\_count]})], ignore\_index=True)

data.head(10)

data.dropna(axis=0,inplace=True)

data.head(20)

data\_2=pd.DataFrame(columns=['Country', 'Vaccine'])

data["Total\_vaccine"] = pd.to\_numeric(data["Total\_vaccine"], errors="coerce")

for country in data["Country"].unique():

new\_data = data[data["Country"] == country]

max\_vaccine = new\_data.loc[new\_data["Total\_vaccine"].idxmax(), "Vaccine"]

data\_2 = pd.concat([data\_2, pd.DataFrame({'Country': [country], 'Vaccine': [max\_vaccine]})], ignore\_index=True)

data\_2.head()

data\_2["Vaccine"].value\_counts().plot(kind="bar",

color=["Red","Gray","Gray","Gray"])

number\_of\_days = (df["date"].max() -df["date"].min() ).days

dtfrm=data[data["Vaccine"]=="Pfizer/BioNTech"]

dtfrm = dtfrm.drop(dtfrm[dtfrm['Country'] == 'European Union'].index)

dtfrm.head(10)

dtfrm["average\_vaccination\_count"] = dtfrm["Total\_vaccine"] / number\_of\_days

dtfrm["average\_vaccination\_count"] =dtfrm["average\_vaccination\_count"].astype(int)

dtfrm.head(15)

dtfrm.set\_index("Country",inplace=True)

color=["Lightblue","Purple","Green","Orange","darkgoldenrod","tan","Gray","Blue","Pink","Lightgreen"]

dtfrm["average\_vaccination\_count"].sort\_values(ascending=False).head(10).plot(kind="bar",color=color)

number\_of\_vaccines = data.groupby('Vaccine')['Country'].nunique()

number\_of\_vaccines.sort\_values(ascending=False).plot(kind="bar",color="r")

fig = px.choropleth(data\_frame=dtfrm,

locations=dtfrm.index,

locationmode='country names',

color='Total\_vaccine',

color\_continuous\_scale='YlOrRd',

title='Ülkelerde Yapılan Biontech Aşıları')

fig.update\_layout(title\_x=0.5)

m = folium.Map(location=[0, 0], zoom\_start=2)

Choropleth(

geo\_data='https://raw.githubusercontent.com/johan/world.geo.json/master/countries.geo.json',

name='choropleth',

data=dtfrm,

columns=[dtfrm.index, 'Total\_vaccine'],

key\_on='feature.properties.name',

fill\_color='YlOrRd',

fill\_opacity=0.7,

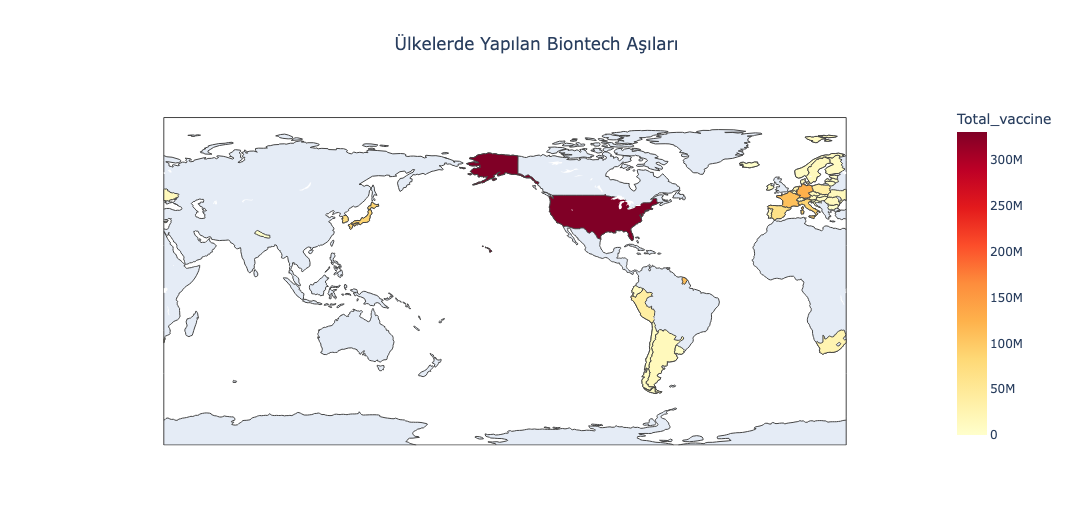
line\_opacity=0.2,

legend\_name='Aşı Sayısı',

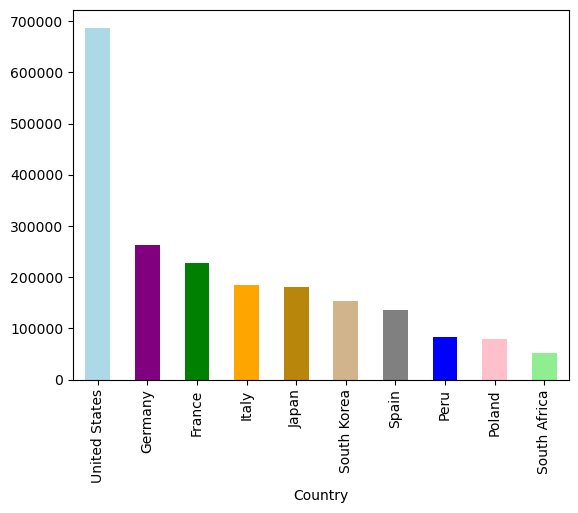
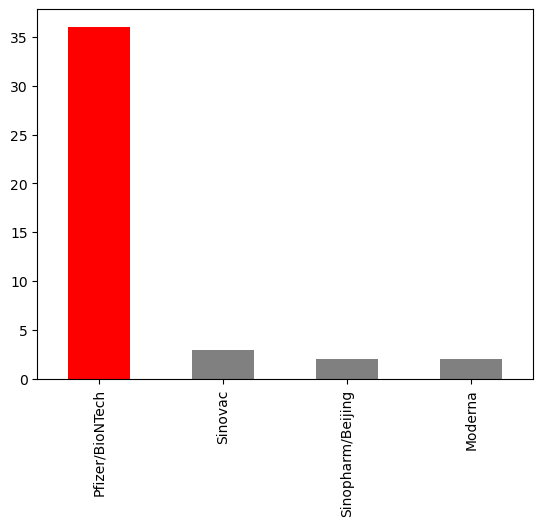
).add\_to(m)

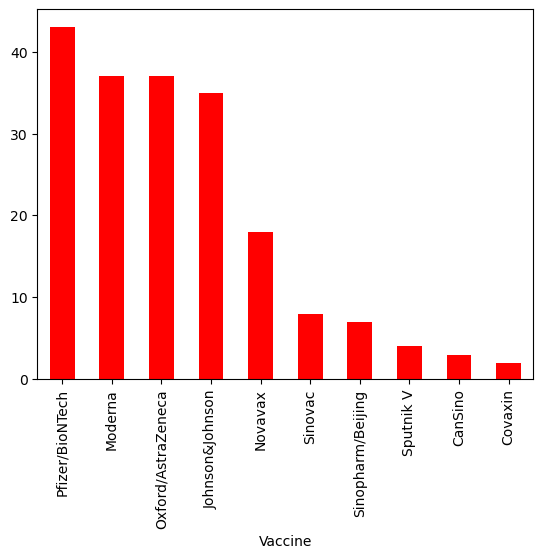
m

**Output:**

****

#### **Example Program with Visualization:** Present an example program that showcases visualization techniques to present the analysis findings effectively. Utilize charts, graphs, and visual representations to enhance the understanding and interpretation of the results.





[“/kaggle/input/covid-worldvaccinationprogress/country\_vaccinations\_by\_manufacturer.csv”]

[“/kaggle/input/covid-world-vaccination-progress/country\_vaccinations.csv”]

**Conclusion:**

At the end of this comprehensive analysis, we can expect to have a model that accurately predicts trends related to the COVID-19 vaccine's efficacy, distribution, potential side effects, and such. These insights will be crucial in shaping public health policies and optimizing vaccine deployment strategies.