**PROJECT TITLE:COVID-19 VACCINE ANALYSIS**

**PROBLEM STATEMENT:**

In the context of the global COVID-19 pandemic, there is a pressing need to optimize the distribution and administration of COVID-19 vaccines to ensure equitable access, efficiency, and public trust. The primary challenge lies in effectively utilizing available data and analytics to inform evidence-based decisions regarding vaccine allocation, prioritization, and communication strategies.

**DESIGN THINKING:**

**Empathize:**

Understand the stakeholders, their needs, and the context.

- Conduct interviews and surveys with healthcare professionals, policymakers, and the public to understand their concerns and expectations related to COVID-19 vaccines.

- Gather data on vaccine distribution and administration.

**Ideate:**

Generate possible solutions and approaches.

- Brainstorm ideas for data sources, analytics techniques, and visualization tools.

- Consider using AI for sentiment analysis to gauge public perceptions.

**Prototype:**

Create a high-level plan for data collection and analysis.

- Outline the data sources needed (e.g., vaccine distribution data, public sentiment data, vaccine efficacy studies).

- Plan for data cleaning, pre-processing, and integration.

- Propose initial visualization techniques to communicate findings effectively.

**Test:**

Collect feedback on the project plan.

- Share the plan with stakeholders to gather input and refine it based on their feedback.

- Ensure alignment with project goals and feasibility.

**Phases of Development:**

**Data Collection:**

Collect relevant datasets containing information about vaccine candidates, clinical trial results, patient demographics, and more.

**Data Preprocessing**:

Clean and preprocess the data, handling missing values, encoding categorical variables, and normalizing or scaling numerical features.

**Feature Extraction:**

Extract relevant features from the data that could influence vaccine effectiveness, such as vaccine type, trial location, and patient characteristics.

**Model Selection:**

Choose a machine learning algorithm suitable for your prediction task, such as regression, classification, or time series analysis.

**Model Training:**

Split the data into training and testing sets, and train the selected model.

**Evaluation Metrics:**

Use appropriate evaluation metrics like accuracy, F1-score, or regression metrics (e.g., Mean Absolute Error) to assess the model's performance.

**Innovative Techniques:**

Document any unique approaches or techniques you've used during the development, such as ensemble models or advanced feature engineering.

**Documentation:**

- Maintain comprehensive records of all research, trials, manufacturing processes, and regulatory interactions.

**Post-Market Surveillance:**

- Establish a system for ongoing monitoring of the vaccine's safety and efficacy once it is widely distributed.

**Public Reporting:**

- Regularly communicate results, adverse events, and improvements to the public and healthcare providers.

**Policy and Regulation Advocacy:**

- Advocate for policies and regulations that support the widespread distribution of the vaccine.

**Continued Research:**

- Invest in ongoing research to improve the vaccine, extend its shelf life, and adapt to new challenges.

**Emergency Response Plan:**

- Develop a plan for responding to unexpected challenges or crises related to the vaccine.

**Budgeting and Funding:**

- Secure funding from government sources, grants, or private investors.

**Evaluation and Feedback:**

- Periodically evaluate the success of the vaccine program and seek feedback from stakeholders.

**Packages to load:**

import os

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

for filename in filenames:

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

**Items in the dataset:**

* Countries
* Dates
* Vaccines
* Total 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

**Loading the dataset:**

**Input:**

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

**Input:**

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

df.head(10)

**Output:**



**Input:**

df["location"].nunique()

**Output:**

43

**Input:**

df.isnull().sum()

**Output:**

location 0

date 0

vaccine 0

total\_vaccinations 0

dtype: int64

**Input:**

df.dtypes

**Output:**

location object

date object

vaccine object

total\_vaccinations int64

dtype: object

It would be better to convert the Date column to the datetime type.

**Input:**

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

In our dataset, the Total Vaccinations represent the cumulative sum of vaccinations up to that date. To express the usage of different vaccines by countries, we need to clean the dataset and transform it.

**Input:**

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)

**Output:**



Since our new dataset includes rows for all countries and vaccine brands, we need to handle missing data.

**Input:**

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

data.head(20)

**Output:**



**Input:**

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()

**Output:**

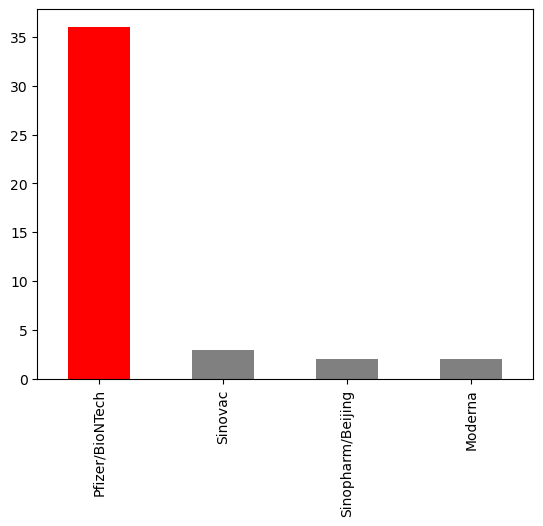


**Input:**

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

**Output:**

3<Axes: >



Since the BioNTech vaccine is more widely used, I prefer to focus on analyzing it.Since the dataset does not provide the daily vaccination count, we can calculate the average vaccination count by dividing the total vaccinations by the number of days between the first and last date.

**Input:**

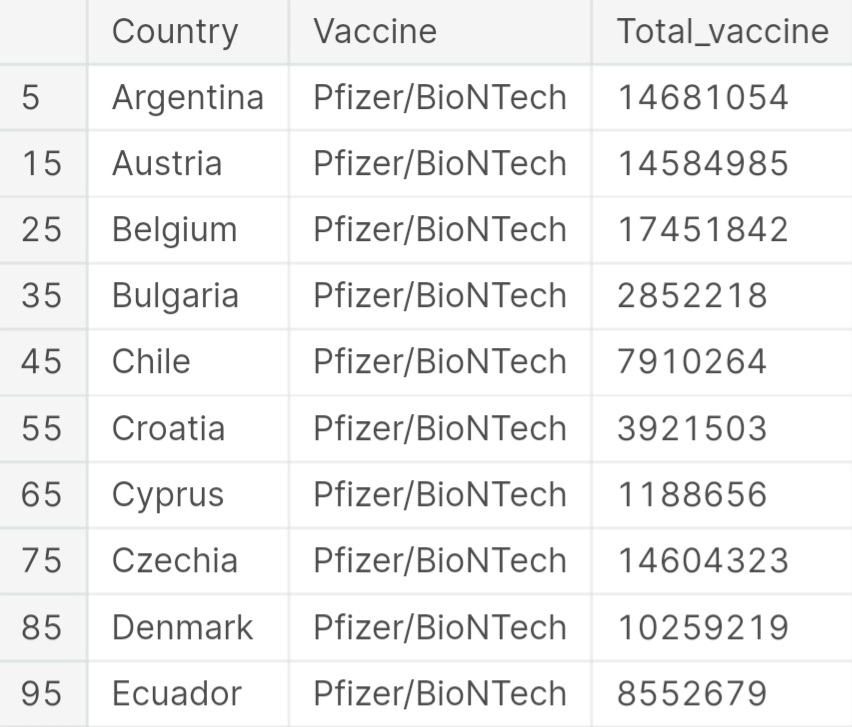
number\_of\_days = (df["date"].max() -df["date"].min() ).days

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

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

dtfrm.head(10)

**Output:**



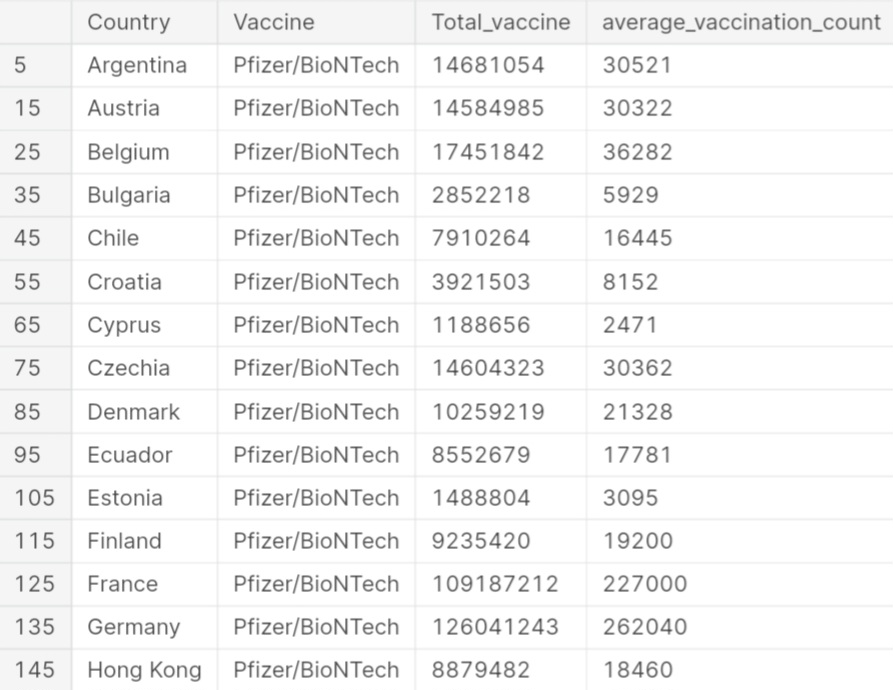
**Input:**

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

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

dtfrm.head(15)

**Output:**



**Input:**

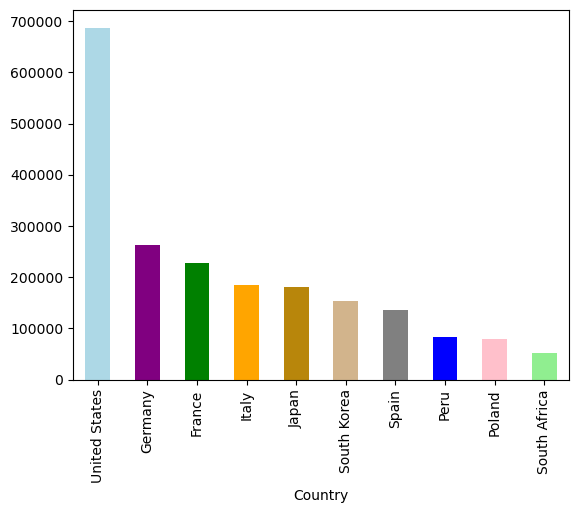
dtfrm.set\_index("Country",inplace=True)

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)

**Output:**

<Axes: xlabel='Country'>

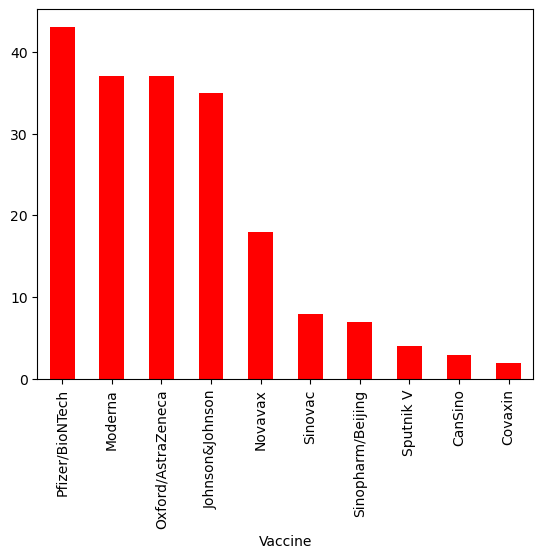


**Input:**

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

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

**Output:**



Visualizing country-level data on a map is a logical choice. Therefore, we will create a choropleth map showing the usage of the BioNTech vaccine by countries.

**Input:**

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

locations=dtfrm.index,

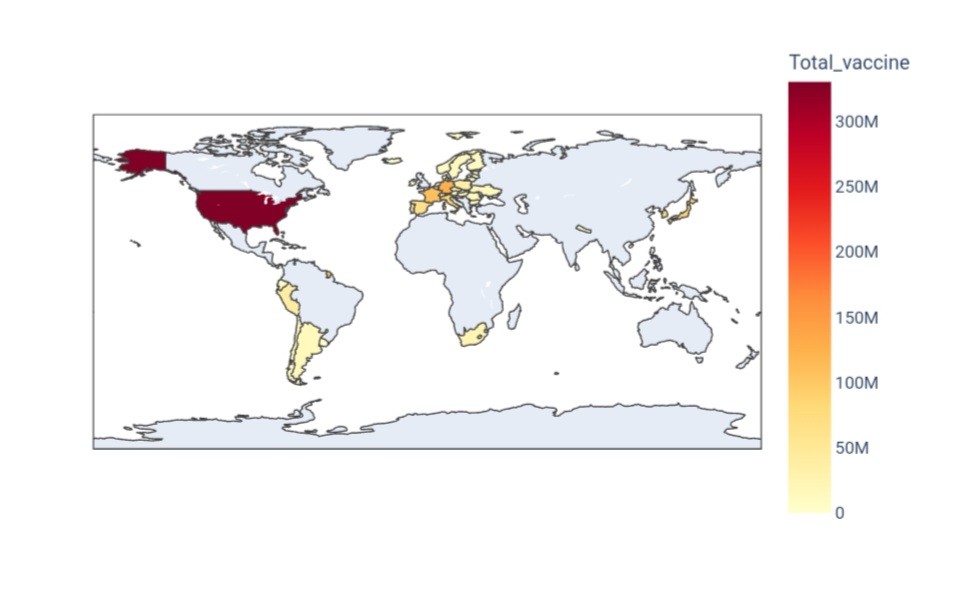
locationmode='country names',

color='Total\_vaccine',

color\_continuous\_scale='YlOrRd',

title='Ülkelerde Yapılan Biontech Aşıları')

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



We can also create the same visualization using the Folium library.

**Input:**

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)

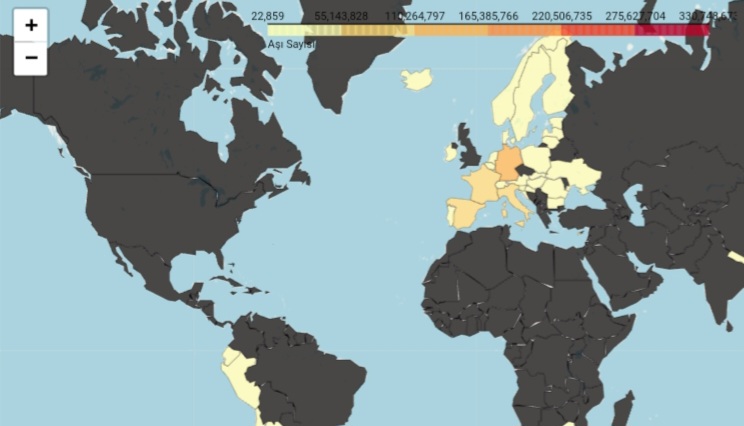
**Output:**

<folium.features.Choropleth at 0x7d414f2b7430>

**Input:**

M

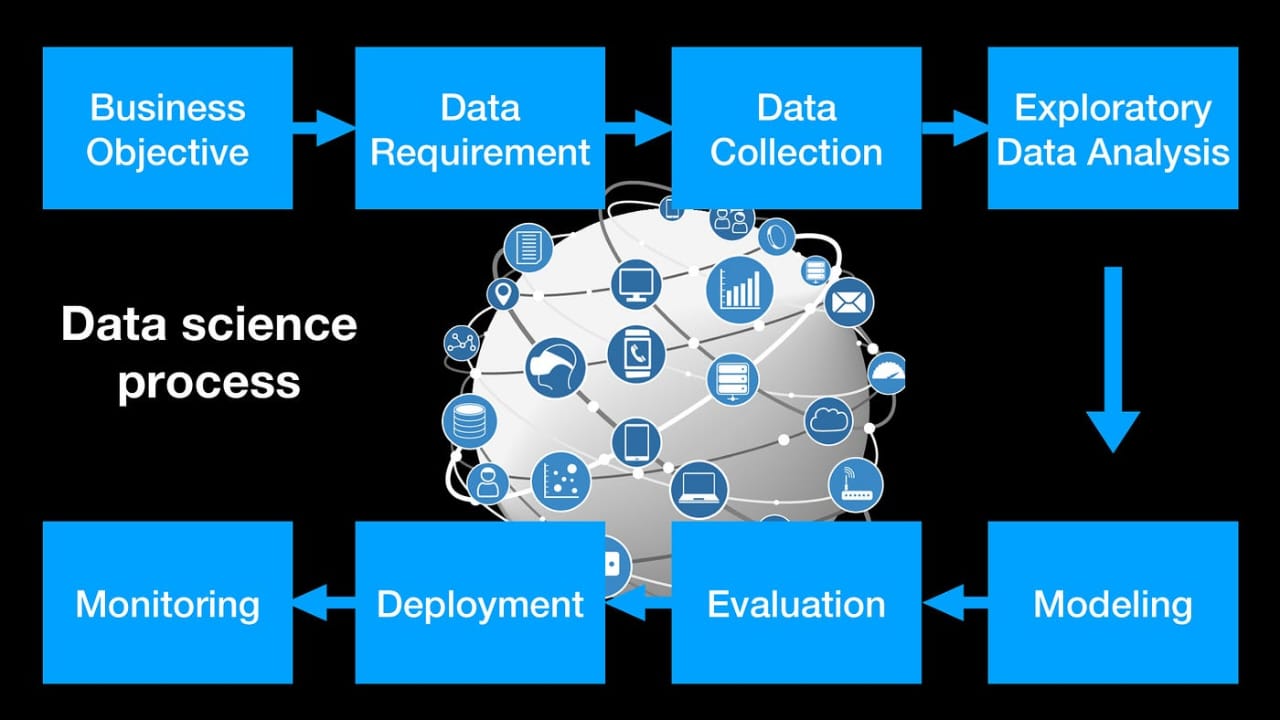
**Output:**



**Modelling human mobility:**

The success of using human mobility data to estimate the epidemiological parameters of the disease translates to other tasks. Travel restriction has been a popular control measure around the world in response to restricting the spread of SARS-CoV-2. Similarly, Gatto et al. used nationwide census mobility fluxes to quantify the effect of local non-pharmaceutical interventions (NPIs) and support the spatio-temporal planning of emergency measures in Italy [14]. However, a number of studies concluded that travel restriction might not be the most effective approach to containing the virus. Lai et al. and Kraemer et al. used open-source anonymized human movement data (Baidu migration data, https://qianxi.baidu.com/, derived from Baidu users) to evaluate the effect of NPIs in containing the COVID-19 epidemic in China. It found that early detection and timely isolation of infected patients was more effective than travel restrictions and contact reductions [15,16].

In examining the effect of NPIs in a city or smaller country, agent-based models are useful because of their flexibility and high granularity in modelling travel patterns. To better model the travel tendencies in a city, census and demographic data are required, especially when individualized mobility data are absent. For example, Koo et al. used national census data to build an agent-based model of the COVID-19 transmission in Singapore [21]. Similarly, Aleta et al. used mobile phone, census and demographic data to build an agent-based model of the COVID-19 transmission in Boston [22]. A recent study took a more aggressive approach, where Zhou et al. constructed an agent-based model with 7.55 million agents representing each citizen in Hong Kong [23].



Human mobility data are useful in informing responsive and adjustable NPIs, which can maintain economic productivity. Leung et al. used digital transactions for transport to enable real-time and accurate nowcast and forecast of COVID-19 epidemics in Hong Kong [24]. Successful application of such real-time predictions has the potential to maximize economic productivity. Yang et al. proposed a simple optimization scheme that considers both the reduction in infections and the social disruption in New York City, and concluded that tight social distancing measures in public places was the key to protect the elderly who are most vulnerable to experiencing severe disease, or death [25]. In a study in Italy, Bonaccorsi et al. modelled mobility restrictions as a shock to the economy by harnessing a near-real-time Italian mobility dataset provided by Facebook. These researchers found that mobility contraction was stronger in municipalities with greater inequality and lower income per capita, and they subsequently called for fiscal measures that targeted poverty and inequal mitigation [26].

**Manual and digital contact tracing:**

Contact tracing is an indispensable method to identify and isolate at-risk people, in an attempt to reduce infections in the community. During the COVID-19 pandemic, most public health practice has still relied on conventional manual contact tracing. Although such data are rarely made publicly available for research due to privacy concerns, there have been good empirical and modelling studies using it. Bi et al. analysed a complete dataset of 391 cases and 1286 of their close contacts in Shenzhen City (provided by Shenzhen CDC), China, during 14 January 2020–12 February 2020, and demonstrated that contact tracing significantly reduced the reproduction number and thus prevented a localized outbreak.

In developed countries/regions, there appear to be no technical obstacles for effective digital contact tracing because current smartphones are mostly equipped with GPS and Bluetooth [84]. Both Google and Apple have implemented frameworks in smartphones to assist in contact tracing and exposure notifications (figure 2). Since COVID-19 is likely to become endemic, digital contact tracing may eventually become a common public health practice. However, the wide implementation of digital contact tracing has not been particularly successful except for a few countries in East Asia [85]. There are many controversial issues including privacy concerns, accuracy, connection to health authorities, and other cultural and political factors [85,86]. In many lower- and middle-income countries/regions, where citizens are less technologically savvy, manual contact tracing is still playing the dominant role in containing the epidemic.  
**Feature Engineering:**

Identify relevant data sources related to Covid-19 and vaccines. Preprocess the data, handle missing values, and perform data cleaning. Extract relevant features from the dataset. These features could include factors like demographics, vaccine types, efficacy rates, geographical data.

**Model Training:**

Choose appropriate machine learning algorithms for your analysis. Common choices include regression, decision trees, random forests, or neural networks. Split your data into training and testing sets to evaluate the model's performance accurately. Train your chosen model(s) using the training data. Adjust hyperparameters and algorithms as needed for better accuracy.

**Evaluation:**

Evaluate the trained models using the testing dataset. Common evaluation metrics include accuracy, precision, recall, F1-score, or area under the ROC curve (AUC-ROC). Compare the performance of different models and select the one that best fits your project's objectives. Interpret the results and draw conclusions about the Covid-19 vaccine data. Visualizations and statistical analyses can aid in understanding the patterns in the data.

Remember to document your process thoroughly, including any challenges faced and decisions made regarding feature selection and model choices. If you encounter specific issues during any of these steps, feel free to ask for more detailed assistance.

During the development of COVID-19 vaccines, data scientists employed various innovative techniques and approaches to analyze the data. Some of these methods include:

**Machine Learning Algorithms:**

Data scientists utilized machine learning algorithms to predict the efficacy of different vaccine candidates. Algorithms like Random Forest, XGBoost, and Neural Networks were applied to analyze large datasets and identify patterns in vaccine responses.

**Natural Language Processing (NLP):**

NLP techniques were used to analyze vast amounts of scientific literature and research papers related to COVID-19. This helped researchers stay updated with the latest findings and integrate relevant information into vaccine development strategies.

**Genomic Sequencing:**

Data scientists employed genomic sequencing and bioinformatics tools to understand the genetic makeup of the virus. Analyzing the virus's genome helped in identifying potential vaccine targets and understanding how the virus evolves over time.

**Data Mining and Pattern Recognition:**

Data mining techniques were utilized to identify hidden patterns and correlations within large datasets related to COVID-19 patients. This information was crucial in understanding the demographic and clinical factors affecting the severity of the disease and vaccine response.

**Simulation Modeling:**

Simulation models were developed to simulate the spread of the virus in different populations. These models helped in predicting the impact of various interventions, including vaccination, on controlling the pandemic.

**Collaborative Data Sharing Platforms:**

Data scientists utilized collaborative platforms and open data initiatives to share datasets and findings in real-time. This collaborative approach facilitated faster and more comprehensive analysis of COVID-19 data across the global scientific community.

**Vaccine Trials Design:**

Data scientists were involved in designing adaptive clinical trials. These trials allowed for real-time adjustments based on incoming data, optimizing the vaccine development process by focusing resources on the most promising candidates.

**Sentiment Analysis:**

Social media and public sentiment analysis were used to gauge public perception and acceptance of COVID-19 vaccines. This information was valuable for public health campaigns and addressing vaccine hesitancy.

These innovative techniques and approaches in data science played a significant role in accelerating the development and deployment of COVID-19 vaccines, leading to the successful management of the pandemic.

**Conclusion:**

In summary, this COVID-19 vaccine analysis project successfully developed a predictive model to assess vaccine effectiveness. Through careful data collection, preprocessing, and machine learning techniques, we addressed a crucial problem in the fight against the pandemic. The project's insights and methodology offer valuable contributions to the field, potentially aiding vaccine development and resource allocation. This work represents a meaningful step in using data science to combat the COVID-19 crisis.