COVID-19 VACCINE ANALYSIS

TEAM MEMBERS

G.DHAYANITHI

K.KALAIYARASAN

T.PARTHASARATHY

K.KARAN

B.MANIYARASU

PHASE 4 SUBMISSION DOCUMENT

PHASE 4: DEVELOPMENT PART 2

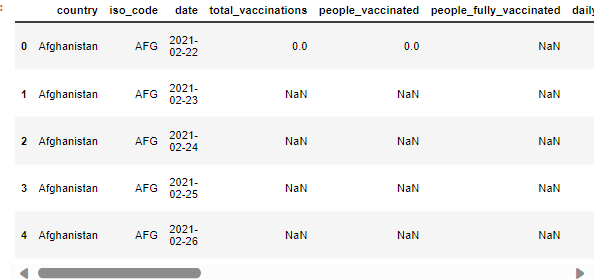


INTRODUCTION

* Data Science is focused on interdisciplinary domains and useful for taking decisions.Effective Vaccines are needed to save lots of lives throughout the worldwide epidemics such as COVID19. The community looks to COVID-19 vaccination progression mustbe considered sensibly in directive to know the users sentiments and fears to it. To knowmore about exact information about covid-19 vaccines are from who are taken the vaccination and they are express their opinions. In this research article studied and understands the advantages of social media. Now a Social media has become an important tool for gaining insights about any domain. At the time of Covid-19 pandemic social media applications are playing a key role in users thoughts on various topics sharing. About Vaccination side effects and results confusion is one of the serious issues in realizing herd immunity and suppressing the COVID-19 epidemic. To consider this approach our focus on analyze user opinions on COVID-19 vaccination process. The world face a main corona virus epidemic from the year 2019. The virus infects fast through various ways. All nations lock peoples to avoid the virus. Vaccinations, including Covaxin, Covishield,Pfizer, Moderna, SputnikV have been permitted. This research article, tweet analysisis based on people’s opinions about official covid-19 vaccines on social media Twitter. Datasets collected, Covaxin, Covishield, Pfizer, Moderna, SputnikV. These tweets are preprocessed using Machine learning techniques.
* In this research article studied the users opinion on Pfizer, Modern, AstraZeneca and Johnson & Johnson. The total posts in each nation for time period of month of Jan 2021 to Apr 2021, May 2021 to Aug 2021 and Sept 2021 to Dec 2021 was plot. The use of opinion Analysis impacts on each domain like product analysis, Recommendation system, prediction on healthcare and analytics.

**EXPLORATORY DATA ANALYSIS :**

**Dataset**



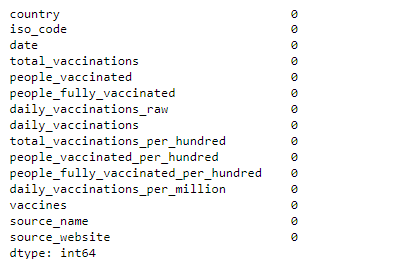
isnull()

The function dataframe. isnull(). sum(). sum() returns the number of missing values in the dataset.

Input:

df.isnull().sum()

Output:



(ii) df.fillna()

**Fill NA**/NaN values using the specified method. Value to use to fill holes (e.g. **0**), alternately a dict/Series/DataFrame of values specifying which value to use on it

Input:

df.fillna(0, inplace = True)

(iii) df.drop

**Drop** specified labels from rows or columns. Remove rows or columns by specifying label names and corresponding axis, or by directly specifying index or column .The **drop**() method removes the specified row or column. By specifying the column axis ( axis='columns' ), the **drop**() method removes the specified column.

Input:

df.drop(df.index[df['iso\_code'] == 0], inplace = True)

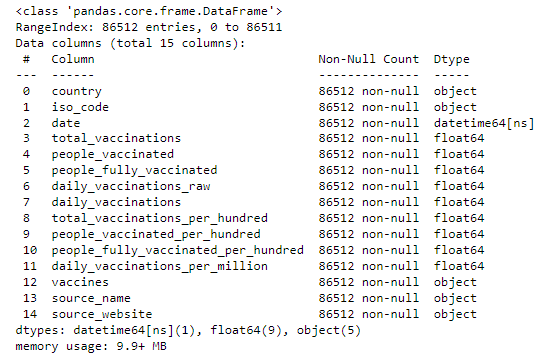
(iv) df.info()

The **info()** method prints information about the DataFrame. The information contains the number of columns, column labels, column data types, memory usage, range .This method prints **information** about a DataFrame including the index dtype and columns, non-null values and memory usage. Parameters: verbosebool, optional.

Input:

df.info()

Output:



STATISTICAL ANALYSIS AND DATA VISUALIZATION :

Total vaccinationsi in India

Program:

plt.figure(figsize=(18,6))

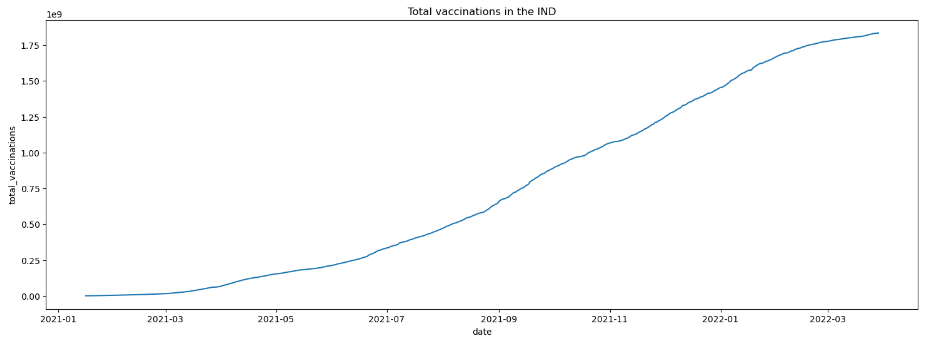
sns.lineplot(data=df\_IND, x="date", y="total\_vaccinations")

plt.title("Total vaccinations in the IND")

plt.xticks(rotations=45)

plt.show()

Output:



Daily vaccinations in India

Program:

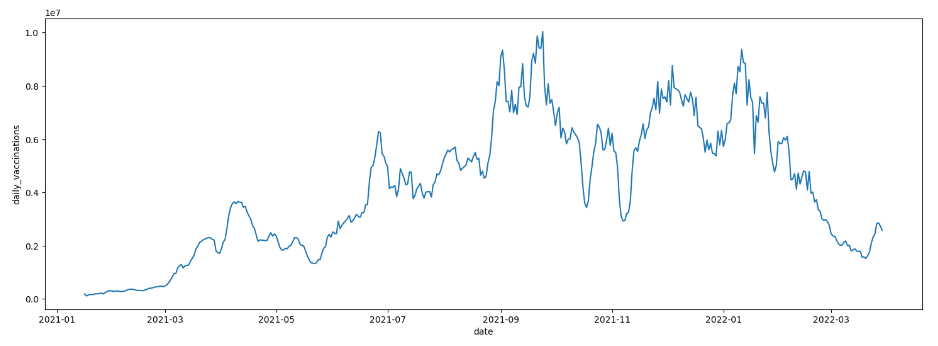
plt.figure(figsize=(18,6))

sns.lineplot(data=df\_IND, x="date", y="daily\_vaccinations")

plt.xticks(rotations=90)

plt.title("Daily vaccinations in the IND")

Output:



(v) df.groupby

A **groupby** operation involves some combination of splitting the object, applying a function, and combining the results. This can be used to group large amounts ..

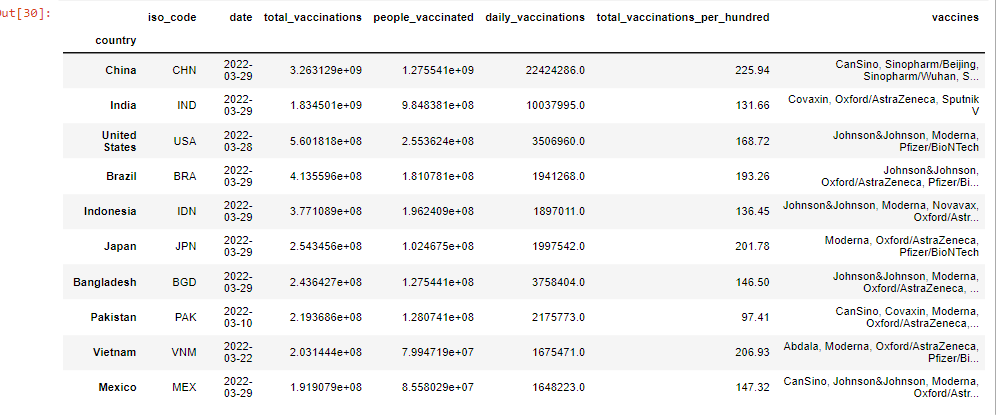
Program:

vacc\_by\_country=df.groupby('country').max().sort\_values('total\_vaccinations', ascending=False)

vacc\_by\_country = vacc\_by\_country.iloc[:10]

vacc\_by\_country

Output:

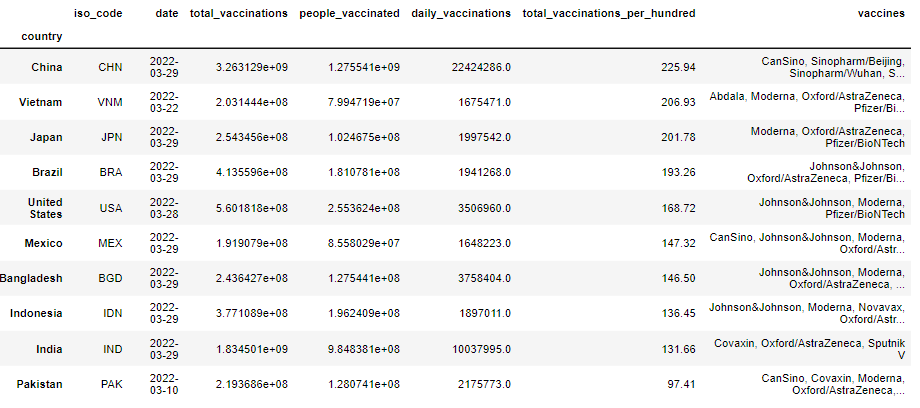


Program:

vacc\_by\_country = vacc\_by\_country.sort\_values('total\_vaccinations\_per\_hundred', ascending=False)

vacc\_by\_country

Output:



BAR PLOT:

Program:

plt.figure(figsize=(18,6))

plt.bar(vacc\_by\_country.index, vacc\_by\_country.total\_vaccinations\_per\_hundred)

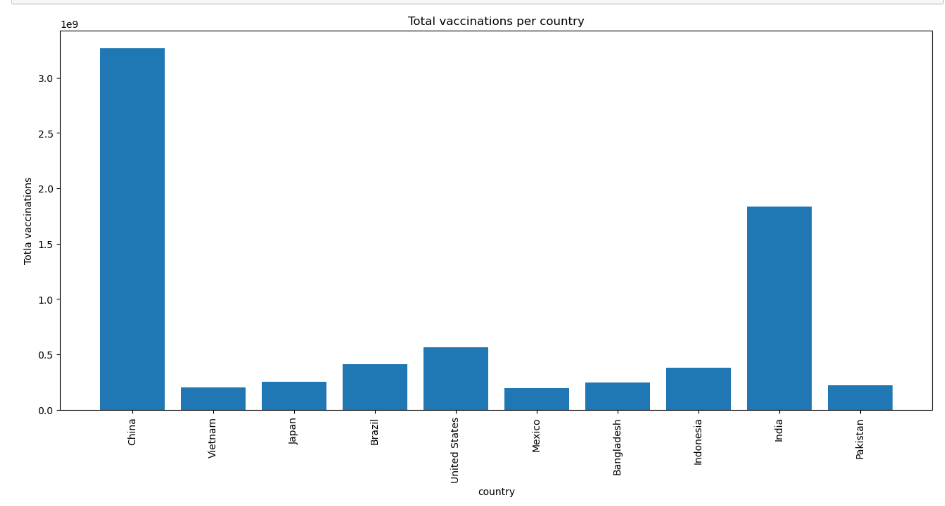
plt.xticks(rotation = 90)

plt.ylabel('vaccinations per 100')

plt.xlabel('country')

plt.show()

Output:



Program:

plt.figure(figsize=(12,8))

sns.barplot(data = vacc\_names\_by\_country, x='vaccines', y = 'total\_vaccinations', hue = 'country', dodge=False)

plt.xticks(rotation=90)

Output:

(array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9]),

[Text(0, 0, 'CanSino, Sinopharm/Beijing, Sinopharm/Wuhan, Sinovac, ZF2001'),

Text(1, 0, 'Covaxin, Oxford/AstraZeneca, Sputnik V'),

Text(2, 0, 'Johnson&Johnson, Moderna, Pfizer/BioNTech'),

Text(3, 0, 'Johnson&Johnson, Oxford/AstraZeneca, Pfizer/BioNTech, Si

ovac'),

Text(4, 0, 'Johnson&Johnson, Moderna, Novavax, Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing, Sinovac'),

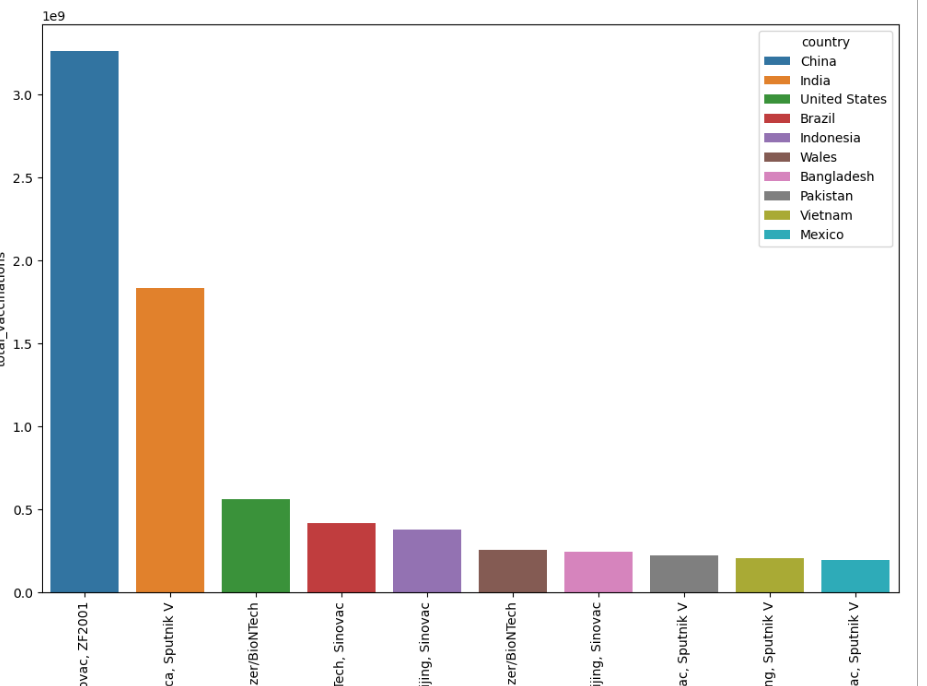
Text(5, 0, 'Moderna, Oxford/AstraZeneca, Pfizer/BioNTech'),

Text(6, 0, 'Johnson&Johnson, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing, Sinovac'),

Text(7, 0, 'CanSino, Covaxin, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing, Sinovac, Sputnik V'),

Text(8, 0, 'Abdala, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech, Sinopharm/Beijing, Sputnik V'),

Text(9, 0, 'CanSino, Johnson&Johnson, Moderna, Oxford/AstraZeneca, Pfizer/BioNTech, Sinovac, Sputnik V')])



Program:

fig = px.choropleth(df.reset\_index(), locations="iso\_code",

color="total\_vaccinations\_per\_hundred",

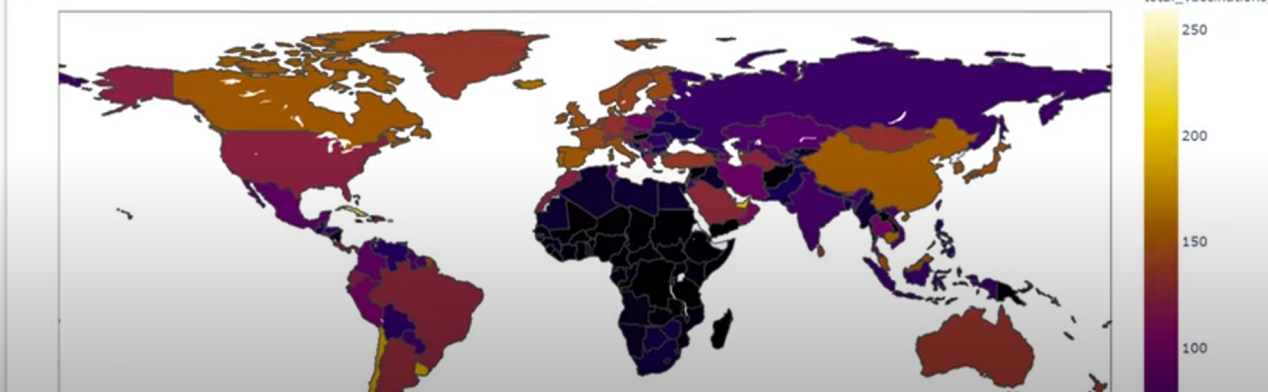
color\_continuous\_scale=px.colors.sequential.Electric,

title= "Total vaccinations per 100")

fig.update\_layout(margin={"r":0,"t":0,"l":0,"b":0})

fig.show()

Output:



CONCLUSION

* In this paper, we collected data from twitter and then apply preprocessing for data exploration, classification. Data Preprocessing involves the Removing URLs, DataFiltering, Removing Special Characters, Removal of Retweets, Usernames, Remove Punctuations and symbols, Usage of Web links, Hashtags, Tokenization, Exclamation and question marks, Letter Repetition, Negations.
* Machine learning classifier used and studied the comparative analysis between KNN, Support Vector Machines, Naïve Bayes, Decision Tree algorithms for data classification.
* Finding shows that Decision Tree classifier for Covishield dataset has achieved the highest 97% accuracy with compared to Naïve Bayes, Support Vector Machine, KNN classification methods. Support Vector Machine has lowest Accuracy with 94% for SputnikV.
* COVID-19 Vaccination dataset wise machine learning model evaluation performance studied and got highest and lowest results of Machine learning classifiers. The Support Vector Machine SputnikV dataset got highest accuracy with 94% and Covishield dataset got lowest accuracy with 89%, The Naïve Bayes got highest accuracy for Covishield dataset with 95% and lowest accuracy with 87% for Moderna dataset, The Decision tree got highest accuracy for Covishield dataset with 97% and lowest accuracy with 88% for Pfizer dataset, The KNN got highest accuracy for Covaxin dataset with 96% and lowest accuracy with 88% for SputnikV dataset.

In Lexicon Based approached Sentiment polarity classification here total 23500 tweets taken for result analysis and predict the vaccination opinions on SputnikV, Covishield, Covishield, Covaxin, Pfizer datasets. Overall here identify the Neutral opinions on Vaccinations. In other side when we focused on positive and negative opinions here Covaxin is more positive compare with all other vaccination datasets according twitter discussion of users insights and negative opinions o