Using Machine Learning Techniques for Predicting and Visualizing Covid-19 Vaccinations Process

University of Missouri – Kansas City
College of Computing & Engineering

Python and Deep Learning

Final Report
Project

Ahmed Alanazi
Rasheed Alhazmi

Introduction:

COVID-19 has become a world pandemic in the past two years. COVID-19 is a virus that affect a respiratory system in human's body. COVID-19 stands for Coronavirus Disease 2019. The issue of this virus, it transmits from one to another very fast with a short distance. COVID-19 has several common symptoms, but its symptoms are similar to another disease which make it more difficult to detected [1]. In December 2019, the first case of Coronavirus was reported in China. It is started to spread to all over the world which causing a global pandemic. Almost, the total cases have reached over 125 million cases globally. Vaccine is the only to stop this pandemic. In the beginning of Dec 2020, number is vaccines were funded and approved by the FDA [2]. In Machine Learning, analyzing and predicting can be reached with using available techniques. Therefore, this project is going to take a chance in processing a dataset from a source over the online websites. The Python/Deep Learning course is the opportunity to apply gained knowledge for predicting possible affected people through symptoms by using unsupervised machine learning techniques.
Video Link: https://youtu.be/L70qNQhXx6E

Objectives:

To achieve the goal of predicting vaccines process, finding a dataset contain all of reported given vaccines so far for the COVID-19, is the key to that goal. The dataset in this project was found on the internet form a website (Kaggle) that matched the requirement of the project. After reading the dataset on to the program, some command had run on the dataset which is Info(), print sample(), and describe(). Attached below is the output for those commands.

Our objectives are to use different supervised and unsupervised machine learning techniques to analyst the information in a dataset that was chosen. By using number different algorithms such as the K-means, for example, K-means will allow us to run it many times on the data, so we can get an accurate result of predicting the Corona Virus given vaccines from the output. We will use what we have learned in the class form python and machine learning.

Procedure

There are different steps needed to follow in order to complete this project. As a first step began with reading and learning about Machine/Deep Learning techniques with the help of CS5590 course. The CS5590 has given us a very deep knowledge to take a role of completing a Machine/Deep Learning project using the Python programming language, also, what algorithms and techniques can rely on.

A dataset with .CSV file was found on Kaggle to be used in this project. The dataset holds useful information can be processed or trained for visualizing and predicting the case of vaccinations around the world. Preparing a dataset is challenging because it must be reviewed and analyzed before applying into an algorithm such KNN. In this project, the dataset that is used needed many work and review to use it. The dataset cleaning process took several days due to leaning new techniques and programming for getting the dataset ready for evaluations.

An Executive Summary of the Dataset

The previous dataset wasn't understandable enough to train it or use it which resulted in not clear evaluations and predictions, therefore, the dataset was changes it this project. This dataset holds number of countries vaccinations statistics. The main attributes in the datasets are Total Vaccinations, Daily Vaccinations, People Vaccinated and People Fully Vaccinated. There are more attributes but our focus on the mention attributes. The mentioned attributes will visualize perfect illustrations of the case of vaccinations process. Each country has reported updates on vaccinations through their official websites. For accuracy, the brand of the vaccine was reported as well.

Proposed Approach

The approach focuses on testing different algorithms and pick out the best one for visualizing and evaluations. Several algorithms were evaluated at once. The help of cleaning and training the data was perfect technique to fit the data into different algorithms at once.

Available Used Tools

Python programing language is the main tool in this project which contains available libraries to complete tasks, trains data, visualize result etc.

```
# Importing Required Libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline

# Read dataset file .CSV through Pandas
dataset = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/country_vaccinations.csv')

# Rertrive first 5 rows of dataset
dataset.head(5)
```

	country	iso_code	date	total_vaccinations	people_vaccinated	people_fully_vaccinated	daily_vacci
0	Afghanistan	AFG	22/02/2021	0.0	0.0	NaN	
1	Afghanistan	AFG	23/02/2021	NaN	NaN	NaN	
2	Afghanistan	AFG	24/02/2021	NaN	NaN	NaN	
3	Afghanistan	AFG	25/02/2021	NaN	NaN	NaN	
4	Afghanistan	AFG	26/02/2021	NaN	NaN	NaN	

In this screenshot, it shows the imported libraries as a start of processing the dataset. Pandas is applied to read a data from a CSV file and with the use of head() which retrieve small part of the

dataset to ensure programming codes are working as needed. Oher libraries will be explained as going through this report.

Approach and Evaluations

Programming is the main aspect of this project, however, in this section the focus will be on reviewing and evaluating the dataset, also, what algorithms are used. There is different type of data such as integer, string, float etc. Defining the type of data is important before applying into an algorithm.

```
country                                0
iso_code                               0
date                                   0
total_vaccinations                    14
people_vaccinated                     15
people_fully_vaccinated               29
daily_vaccinations_raw                24
daily_vaccinations                     1
total_vaccinations_per_hundred        14
people_vaccinated_per_hundred         15
people_fully_vaccinated_per_hundred   29
daily_vaccinations_per_million         1
vaccines                               0
source_name                            0
source_website                         0
dtype: int64
```

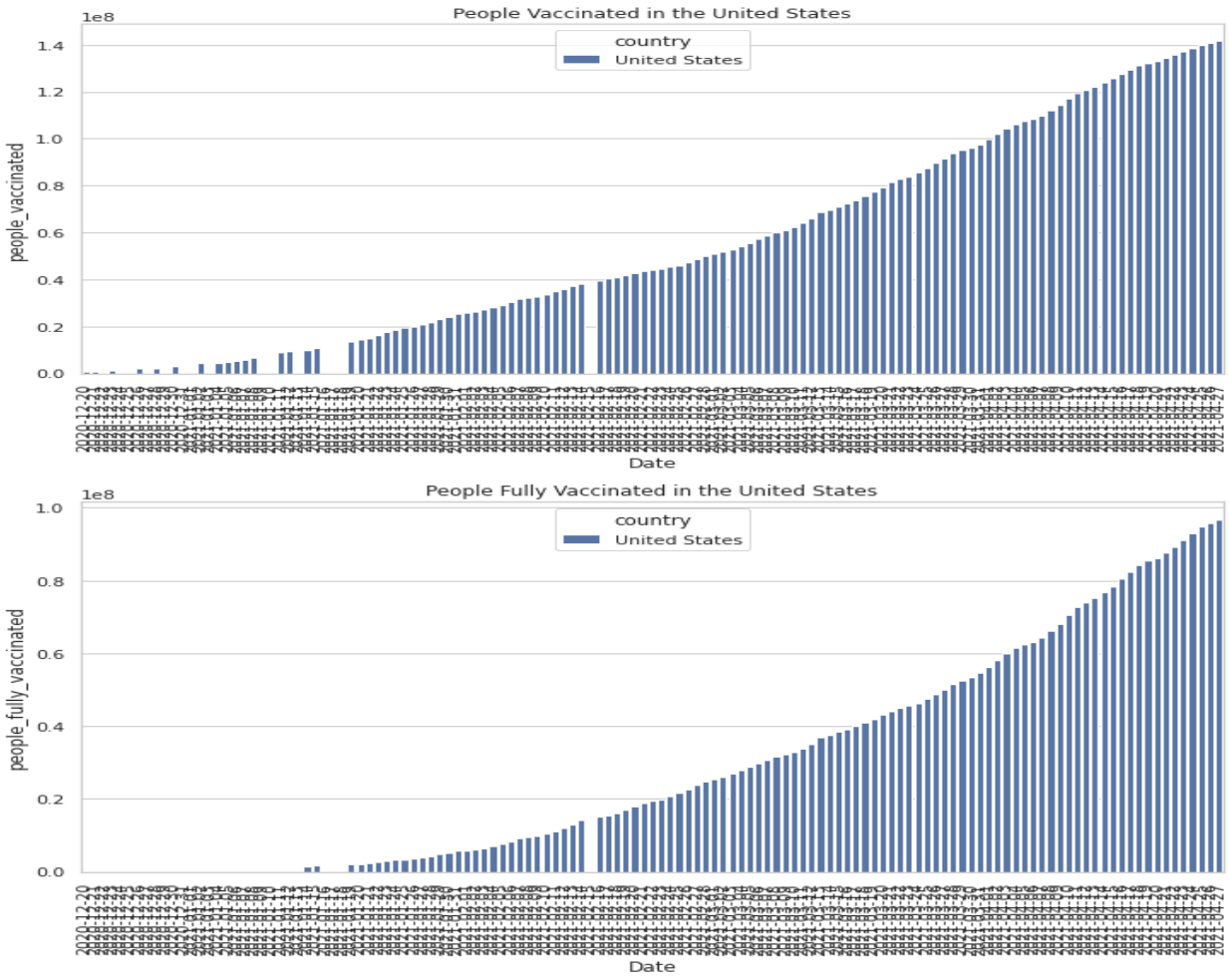
In this screenshot, it lists the attributes and if there are any null values. Algorithms cannot deal with null values which result in not accurate evaluation or not running the process in order to avoid that it is useful to fill null values with mean.

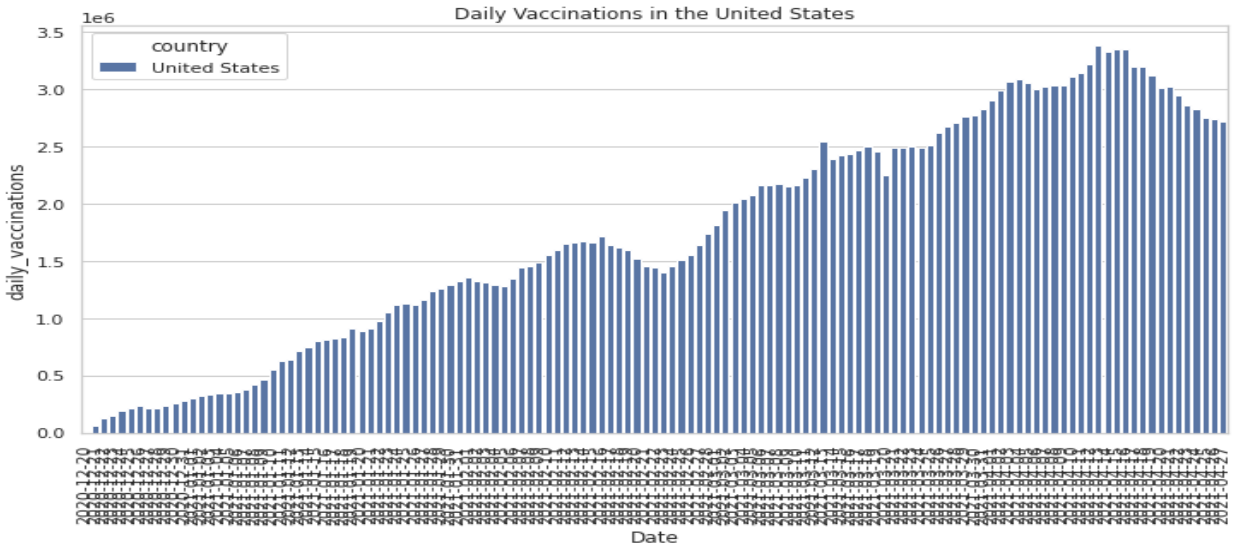
The dataset has a large data regarding vaccinations so picking matched data to be trained is a perfect use for accurate result. In this project, United States was chosen to be evaluated as shown below.

```
# Group data by country specifically United States
UnitedStates = dataset.loc[dataset.country == 'United States']
print(UnitedStates.head())
print(UnitedStates.isna().sum())
```

This line of code, hold only data related to the US every time is runs. This data is temporary saved into UnitedStates variable. The sum() function helps counting how many raw UnitedStates variable holds. As a result it holds 129 raw which mean the data will be processed faster.

As the vaccinations runs daily, the US government reports it. It reported Daily Vaccinations, People Vaccinated, Total Vaccinations, People Fully Vaccinated etc. The goal is to vaccinate as much as they can in a short time. Predicting and evaluating will help to see if the process of vaccinations is proceeding as planned or it has a delay, also, when it can be done.

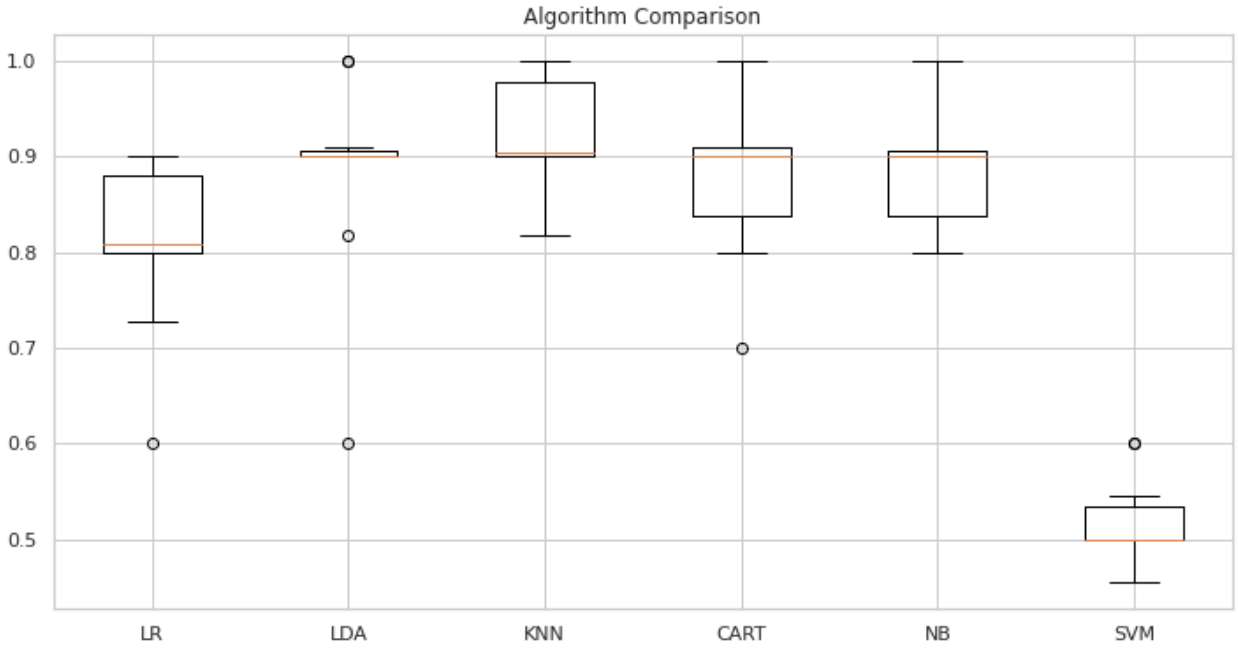




The above illustrations show how the daily vaccinations, people vaccinated and people fully vaccinated are growing daily. However, this data still contains some errors which can result in inaccuracy of the evaluation and predictions of the final results. In this case, different algorithms will be run and used to pick the highest accuracy. Also, the dataset will be trained to fit models such K-means.

KNeighbors Classifier, Logistic Regression, Decision Tree Classifier, Linear Discriminant Analysis, GaussianNB and SVC are used to pick the best one which has a higher accuracy. After running the above algorithms, the best one was KNN as shown below.

```
Algorithm Comparison: _____________________
LR: 0.806364 (0.086917)
LDA: 0.882727 (0.106674)
KNN: 0.923636 (0.055922)
CART: 0.883636 (0.086129)
NB: 0.894545 (0.065707)
SVM: 0.515455 (0.048795)
```



The next step was to train the data and run it into the chosen algorithm which KNN because it had the best result. With training the data, some null values prevented the algorithm from running, therefore, some hard coding was applied to fill the null values with means as shown below.

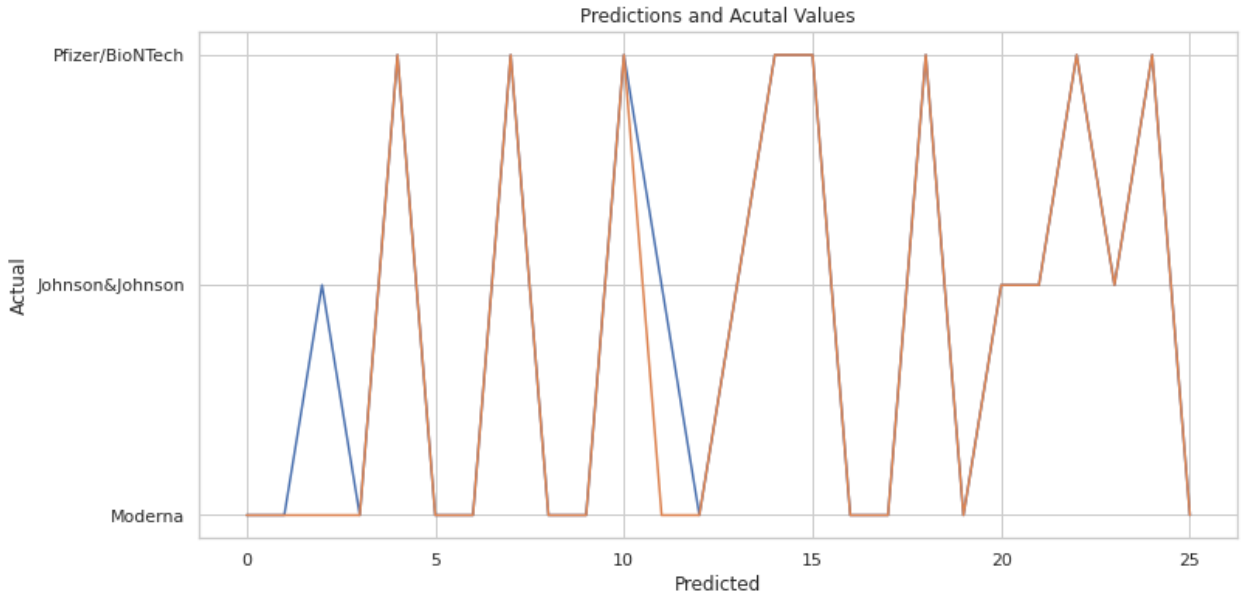
```
vals = pd.to_numeric(data2['daily_vaccinations_raw'], errors='coerce')
data2["daily_vaccinations_raw"] = vals.fillna(vals.mean())

vals = pd.to_numeric(data2['daily_vaccinations'], errors='coerce')
data2["daily_vaccinations"] = vals.fillna(vals.mean())

vals = pd.to_numeric(data2['total_vaccinations_per_hundred'], errors='coerce')
data2["total_vaccinations_per_hundred"] = vals.fillna(vals.mean())
```

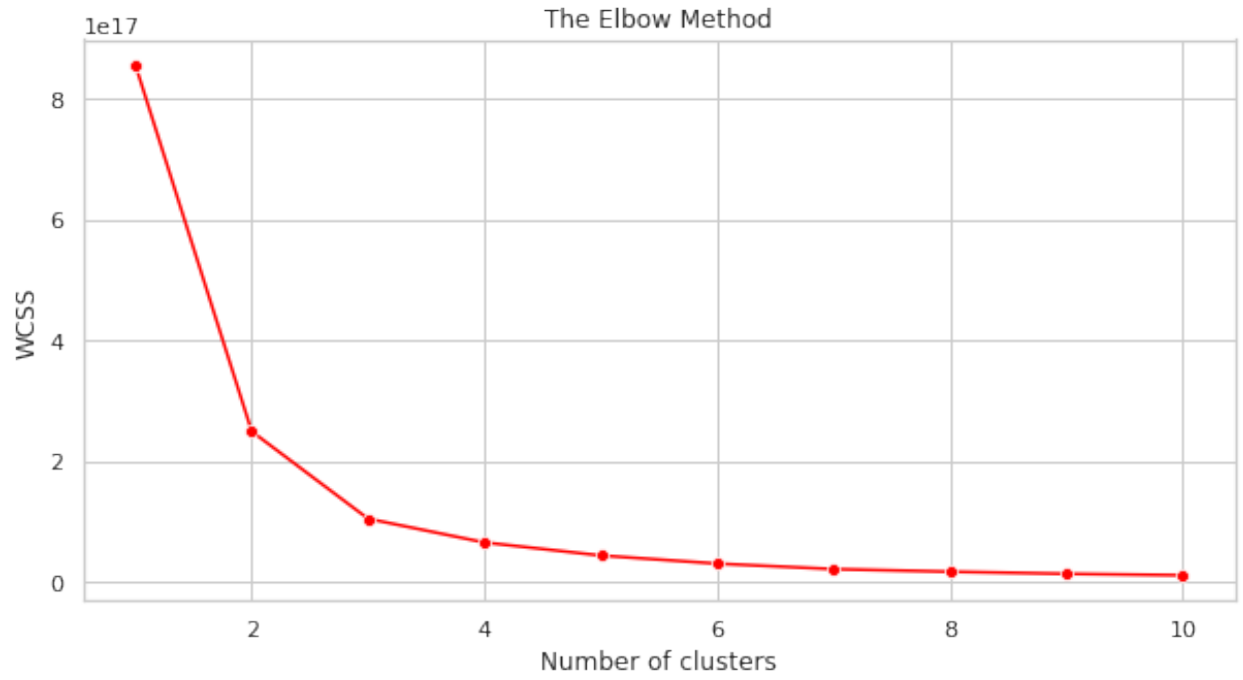
Fitting the data into the KNN then predict it. The KNN takes the data as trained and tested through Split Test after that predict the result which also can be compared with the actual values as shown below.

```
model = KNeighborsClassifier(n_neighbors=5)
model.fit(X_train, Y_train)
predictions = model.predict(X_validation)
```

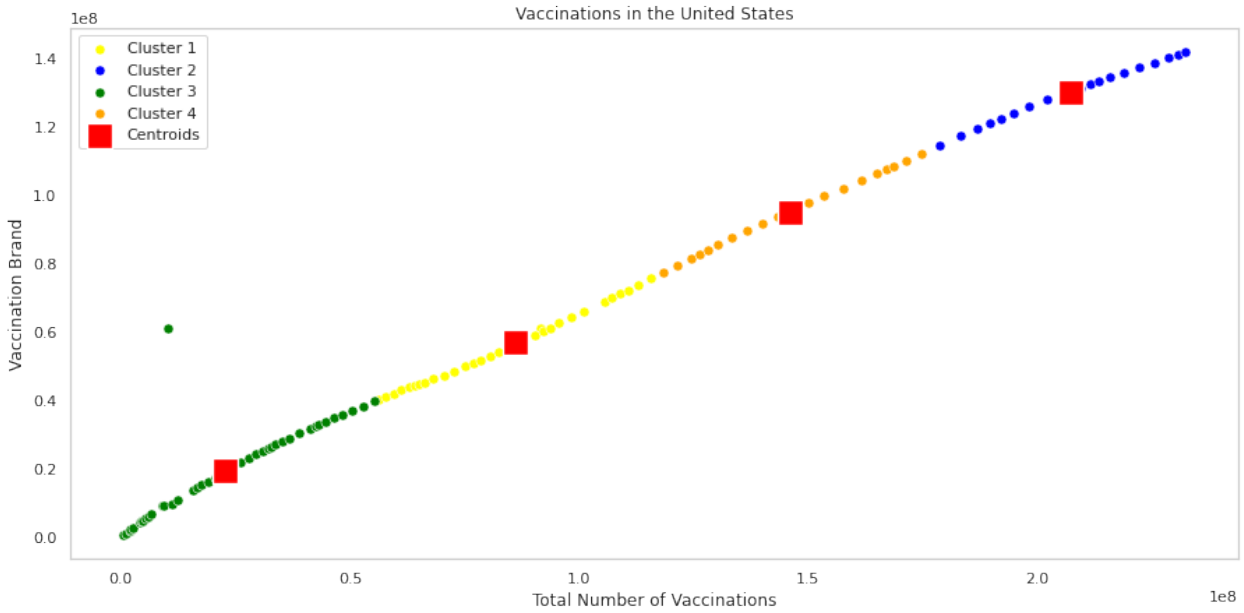


In this graph, it shows the actual and predicted values are close. The values of Daily Vaccinations, Total Vaccinations, People Vaccinated and People Fully Vaccinated with brand. The brand played a role in getting people to vaccinated which shows people preferred moderna and Pfizer vaccines.

On other hand, the evaluation and prediction does not seem accurate even though it has a 92 accuracy. In this case, an unsupervised learning was applied to take a better understanding of the result. K Means Clustering was applied to this dataset also the elbow technique was used to pick the right n_clustering as shown below.



The n_clustering was 4 as this plot shows after applying it the result four groups of data was divided as shown.



This graph shows there are four groups taking vaccines brands in terms of daily and total vaccinations.

Conclusion:

As a conclusion, this project is completed, however, it shows the accuracy not fully 100 percent due some issues such as the dataset, training the dataset, or choosing an algorithm. It was very challenging to find a clean or perfect data that way we ended up having hard time spent in understanding and training the dataset. Overall, this project gave us a big opportunity to learn a new programming language and machine/deep learning.

References:

[1] What is Coronavirus. (n.d.). Retrieved May 05, 2021, from https://www.who.int/health-topics/coronavirus#tab=tab_1

[2] Pfizer–BioNTech COVID-19 vaccine. (2021, May 02). Retrieved May 05, 2021, from https://en.wikipedia.org/wiki/Pfizer%E2%80%93BioNTech_COVID-19_vaccine

[3] Brownlee, J. (2020, August 19). Your first machine learning project in python Step-By-Step. Retrieved May 06, 2021, from https://machinelearningmastery.com/machine-learning-in-python-step-by-step/

[4] Your machine learning and data science community. (n.d.). Retrieved May 06, 2021, from https://www.kaggle.com/

[5] Amolbhivarkar. (2018, March 26). KNN for classification using scikit-learn. Retrieved May 06, 2021, from https://www.kaggle.com/amolbhivarkar/knn-for-classification-using-scikit-learn

[6] Sklearn.cluster.kmeans¶. (n.d.). Retrieved May 06, 2021, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html