COVID-19 Prediction model using Machine Learning

Amr Jadi

a.jadi@uoh.edu.sa
Department of Computer Science and Information
College of Computer Science and Engineering,
University of Ha'il, Ha'il, Saudi Arabia

Summary

The outbreak of the deadly virus COVID-19 is said to infect 17.3Cr people around the globe since 2019. This outbreak is continuously affecting a lot of new people till this day and, most of it is said to under control. However, vaccines introduced around the world can help mitigate the risk of the virus. Apart from medical professionals, prediction models are also said to combinedly help predict the risk of infection based on given datasets. This paper is based on publication of a machine learning approach using regression models to predict the output based on dataset which have indictors grouped based on active, tested, recovered and critical cases along with regions and cities covering most of it from Dubai. Hence, the active cases are tested based on the other indicators and other attributes. The coefficient of the determination (r^2) is 0.96, which is considered promising. This model can be used as an frame work, among others, to predict the resources related to the dangerous outbreak.

Keywords— *COVID-19, Linear Regression, Machine Learning, Saudi Arabia.*

I. INTRODUCTION

All over the world, every country, state, district, city and town are suffering from this dangerous Corona Virus for the past two years. World Health Organization (WHO) identified this virus on 31st December 2019. The start of Corona virus is from Wuhan city located in China. The actual meaning of Covid-19 is "CO" is Corona, "VI" is Virus and "D" is Disease. Among all virus and infections, Covid-19 is a deadly virus [6]. It starts from sickness, cold, cough and it leads to several other dangerous diseases. A fifty years old man is the first person in China who got contrasted with Covid-19 and it was approximately on November 17, 2019. After one-month doctors noticed the increase in cases. Wuhan city in China is more popular for various kinds of seafood. They first have a suspicion that something has been sold in the market due to which the people are getting infected [5].

But some people affected by the disease have no connection with the market. China has declared that Covid-19 is spreading because of bats and it contains the virus which leads to corona. All these assumptions of China went wrong and WHO confirmed that the virus is spreading among people, not with any animal. From Wuhan, it starts spreading faster to other states and cities in China. The deadly virus also attacked the countries which are near to China they are North and South Korea and Japan. Likewise, it started spreading all over the world and reached around 200 countries. Saudi Arabia got the first case of corona virus in the year 2020 on 2nd March. Around 2.1 million people got infected in two months during the peak spread of Covid-19. By April 5 2020, this disease has infected over 2,400 individual's people in Saudi Arabia [1]. It took exactly 32 days to spread from the person who got infected on March 2nd 2020. Every country decided to have a lockdown as this is only the possible way to stop the spread of the corona virus. So Saudi Arabia has announced a lockdown on March 25, 2020. From the first day of lockdown, the cases in Saudi Arabia are around 400,000. As the population of humans has double over the past five decades is the main reason for the rapid spread of the disease and it takes few seconds to spread from one person to another person [4]. There is no proper infrastructure and even doctors don't have an idea how to stop this virus. More than three-quarters of the world populations are living in urban areas.

It will be very difficult for such people and even more complex to survive in lockdown situations. The population of Saudi Arabia is around 38.8 million and stands at 25th position among all countries in regards to income rate [2]. This country is not yet developed well and it is still under the category of a developing country. It is very difficult for them to determine the exact number of cases and spread of the disease at different geographical locations. In April 2020 the number of cases remains increasing and all the hospitals of Saudi Arabia came under enormous pressure for per peak cases. By having a proper lockdown, the spread of disease is lowered down to less than 0.4 million in Saudi [3]. To have proper vaccination and cure, cities are making more efforts to stop this pandemic situation.

A. Aim and objectives

AIM: to predict the COVID-19 cases using machine learning
OBJECTIVES:

- To find an appropriate dataset.
- To apply preferable cleansing methods on the Raw dataset.
- To apply a suitable machine learning algorithm for continuous type if values in the dataset.
- To examine the predicted output, accuracy and other parameters.

II. METHODOLOGY

The paper is related to the prediction of the deadly disease COVID-19 using machine learning algorithm. The flow of the design can be seen in the Fig.1. the following steps

Manuscript received August 5, 2021
Manuscript revised August 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.8.33

involved in achieving the output is discussed here in this section as shown in Fig.1. Steps like data cleansing, understanding the attributes in the dataset and carrying the models are also discussed.

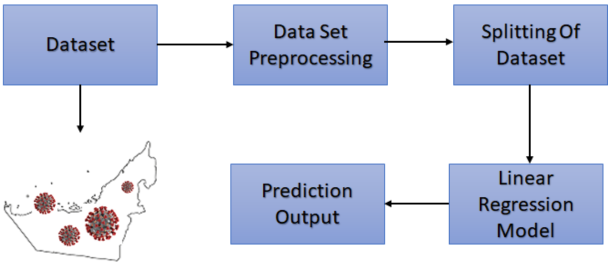


Figure 1 steps to carry out the model.

B. Data pre-processing

This stage indicates all the activities associated with the cleansing of the raw dataset. This processed dataset is necessary to perform efficiently. This process is done as the first step. This includes modifications on the raw data like removal of null values and removing the unwanted attributes which can affect the accuracy of the model. This is important as some datasets are corrupted with unwanted datatypes and corrupted values. This is somewhere related to the data debugging as said by Krishnan *et al* [7]. The methods to cleanse the data merely depends on the type of the dataset and the expected outputs. However, many other automated methods for data cleansing are discussed by [8], [9].

C. Data splitting

The data is split into training and testing sets. Usually, the sets are splitted into 80% and 20%. Where one set is used to test and the other to train the prediction model. In this way, one specifies enough data to train and test the models.

D. Application of the prediction models

After the dataset is cleansed and splitted a suitable model is imported to predict the output based on the input. In this case, linear regression model.

E. Predicted output

The last step after fitting the applicable model to the splitted sets is to predict the outputs for the given input and this step determines how accurate the model predicts the output.

F. Linear Regression

The model which is used in the prediction of analysis can be referred to as the linear regression model. It will predict the value of one variable depending on the value of another variable [10]. After the process of correlation, linear regression is considered to be the next step. Variables that are predicted by this model will be considered as dependent variables and outcome variables. The idea behind the linear regression model is to observe the set of predictor variables [16]. In such a way that variables are performing their assigned task in predicting the outcome of the variable. It also observes which variable will be considered as the most important predictor of the outcome variable. Linear regression helps in presenting the outcome variables by the magnitude and the sign of beta. It also checks the impact of magnitude and sign of beta on outcome variables. The relationship between two variables can be built by the linear equation of the linear regression model and the linear equation is

$$Y = a + bX. \qquad (1)$$

In this equation (1), ‘X’ is the explanatory variable and ‘Y’ is the dependent variable. In equation ‘b’ will be the slope of the line and ‘a’ will be considered as intercept [11]. Outcome variables and criterion variable are called regressions dependent variables. Exogenous variables and predictor variables are called independent variables. To determine the strength of the predictors the linear regression model is used. For forecasting an effect and trending the forecast this model is used. Linear regression is classified into two types that are simple linear regression model and multiple linear regressions. The process of finding the relationship between the independent variable and a corresponding dependent variable is called simple linear regression [13]. The equation of simple linear regression model is

$$Yi = \beta0 + \beta1\, Xi + €i \qquad (2)$$

In the above equation (2), YI is the dependent variable, β0 is the population of the Y-intercept and β1 is the population of the slope coefficient. XI is the independent variable and €i is the random error term. This simple linear regression model has a wide range of applications that are depending on the amount of rainfall predicting the crop yield. In this application, the yield is the dependent variable and the rainfall is the independent variable. Depending on the professor or lecture teaching the marks scored by the student. In this professor, teaching is the independent variable and the student marks are dependent variables. Multiple linear regressions are one of the types of a linear regression model. The process of finding the relationship between the two or more independent variables and the corresponding dependent variables is called multiple linear regressions [12]. In this, the independent variables will be in continuous form. The equation of multiple linear regressions is

$$\mathbf{Yi = \beta0 + \beta1\, Xi1 + \beta2\, Xi2 + \cdots \ldots + \beta p - 1\, Xip - 1 + €i} \qquad (3)$$

In the above equation (3), Yi is the response of the dependent variable. The variables β0+β1 Xi1+ β2 Xi2+……+ βp-1 Xip-1+€i can be referred to as linear predictors, β is the coefficient and X is the predictor the independent variable. The application of multiple linear regression models is it helps in predicting the trends for the shopkeepers and shareholders, etc. Future values also predicted by the multiple linear regression models.

Impact of any changes can be done with this model. It also estimates the change that takes place in the dependent variable when there is a change in the independent variable. The advantage of the linear regression model is work progress will be excellent irrespective of the dataset size [14]. Information can be gathered about the significance of the features. There will be an investigation process between the relationship of Y and numerous independent variables (X). Any standard software can use this model. The measurement for the input variables can be at any level and the interval variable will be considered as a target. The disadvantages of the linear regression model are there will be more assumptions in predicting the output variables.

When the difficult levels of the models increase at that time disturbance will be raised in the programming process of the model [15].

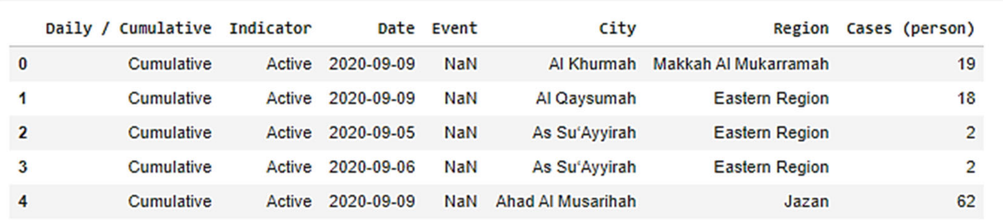
To overcome the disadvantages of the linear regression model lasso regression and ridge regression were introduced. Lasso (Least absolute shrinkage and selection operator) regression uses shrinkage and the word shrinkage means the whole process of the model. It is the place where the data values will be shrunk towards the central point [18]. The process of approximating the coefficients of the multiple-regression model is called ridge regression.

This approximation is possible only when the correlation of the independent variables takes place in a large amount [17].

III. RESULT

A. Dataset Description

Dataset provides information about Covid-19 cases in Saudi Arabia. In the year 2020 coronavirus has reached all over the world and count of people got affected by this disease is in millions and more. This dataset provides the Covid-19 cases in Saudi Arabia with some particular columns and rows as shown in Fig. 2. Those are date, city, region, cases (person) and cumulative cases. The date column is very important to know how many people got affected by this disease on that particular date. In this dataset, the information is from 4 June 2020 and it was spitted into two sheets. In the second sheet, the date is from March 02, 2020 and it has the same columns as in sheet-1. In sheet-1 from 4 June 2020, the person infected rate and cumulative cases rate is also the same that is 11. The city and region of that particular date are Al Wajh and Tabuk. On 12 June 2020 in Al Madinah Al Munawwarah the person infected rate is 11189 and the cumulative cases are 11900. From that day there is a rise and fall in persons infected rate but the cumulative cases have increased. In sheet-2 only one person in Al Qatif in the Eastern Region of Saudi Arabia got infected with the virus and the cumulative case is also one. From March 8th 2020 for the same region and city, the infected person rate has increase to 11 and cumulative cases to 60.



	Daily / Cumulative	Indicator	Date	Event	City	Region	Cases (person)
0	Cumulative	Active	2020-09-09	NaN	Al Khumah	Makkah Al Mukarramah	19
1	Cumulative	Active	2020-09-09	NaN	Al Qaysumah	Eastern Region	18
2	Cumulative	Active	2020-09-05	NaN	As Su‘Ayyirah	Eastern Region	2
3	Cumulative	Active	2020-09-06	NaN	As Su‘Ayyirah	Eastern Region	2
4	Cumulative	Active	2020-09-09	NaN	Ahad Al Musarihah	Jazan	62

Figure 2 Head of the Dataset

B. Implementation of the model

This part shows the step-by-step implantation of the dataset towards the prediction model.

- To predict the Covid-19 cases of Saudi Arabia the linear regression model is used.
- To implement that model Google Collaboratory is used as a notebook and drive to mount and access the dataset.
- Next step is to import some libraries that are required to continue the process. Those are numpy as np, pandas as Pd, Matplotlib. pyplot as plt, import plotly. graph_objects as go, plotly.express as px, seaborn as sns, and finally, datetime.
- All these libraries are imported and after this os is also imported to locate the path of the dataset using os.chir (“path location).
- To read that file data=pd.read_csv (“name of the file”).
- To visible the dataset data.head () is used.
- After this EDA (exploratory data analysis is carried out to explore the attributes and instances of the data.
- Finally, the data-processing helps to get the best quality out of the raw data to be used by the model.
- The last step involves the model fitting and calculating the r^2 score of the model to justify the performance.

C. EDA (Exploratory data analysis)

Type of dataset, number of columns, rows and null values are represented and the investigation process is done by the means of EDA. It summarizes the main characters of the dataset and some data visualization methods are also offered. EDA is applied to the dataset.

Firstly, the names of the column are group along with their percentages representing the null values. The column "Event" is filled with 98% of the null values so these are replaced with the word "UNKNOWN".

Updating of dates column is done and then the index column will be updated. Date, cases (persons) and index columns are arranged in ascending order. The evolution of cases in Saudi Arabia is large. Day-to-day the graph of covid-19 patients is increasing in a wide range as seen in Fig. 3. From day 0 to 50 there are 51, 361.47 cases and after 100 days it was 100, 1.583194 million people got infected. By 200 days it was 22, 8.702904 million cases.

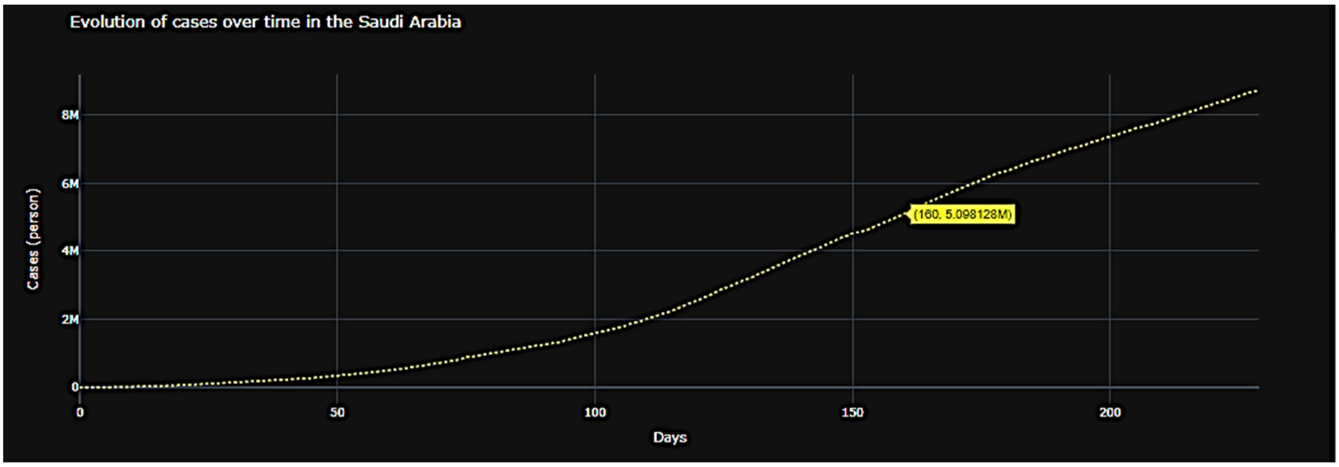


Figure 3 Evolution of cases over the time.

From April 2020 to October 2020 there is a wide increase in the cases in each day. In April it is just less than one million and in July it was 4.558867 million. On 1st October 2020, the cases were 7.9832242 million and by 17 October it reached 8.702904 million as seen in Fig.4.

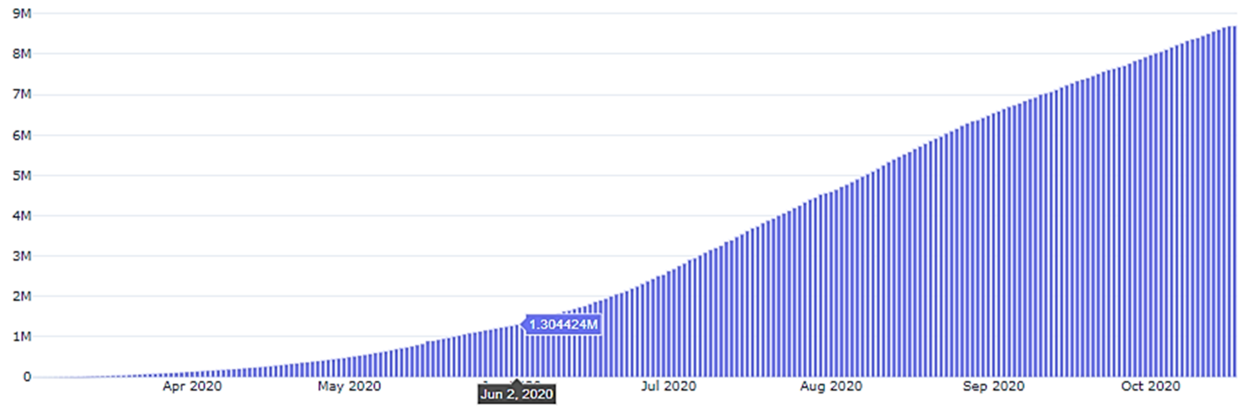


Figure 4 Cases per month

Cases in cities of Saudi Arabia, in a Namas there are 1,723 cases and 131 cases in Uqlat as Suqur. 10,941 cases in Al Jubayl and Ar Riyad region 114,061 people got infected with this disease. There are the highest cases in this city and the least are in Uqlat as Suqur. In the Al Jawf region of Saudi Arabia, they are 246 cases. 6059 cases in Al Bahah and 51,005 cases in Al Madinah Al Munawwaarah. Makkah Al Mukarramah is the region that got infected with more Covid-19 cases and the number is 159,837. Al Jawf is the region with the least number of cases and the percentages are shown in Fig. 5,6.

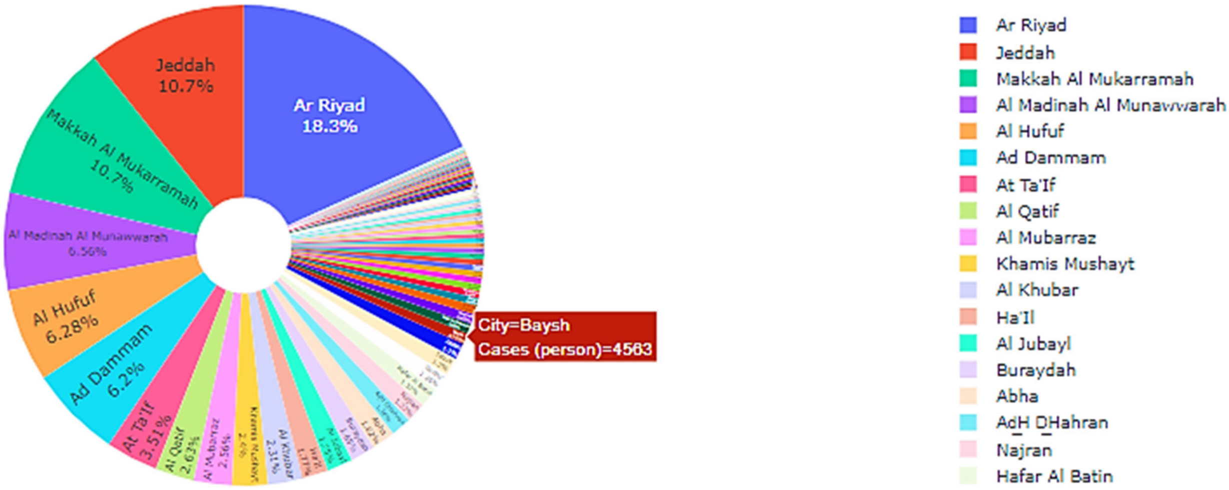


Figure 5 Donut chart representing the cases per city.

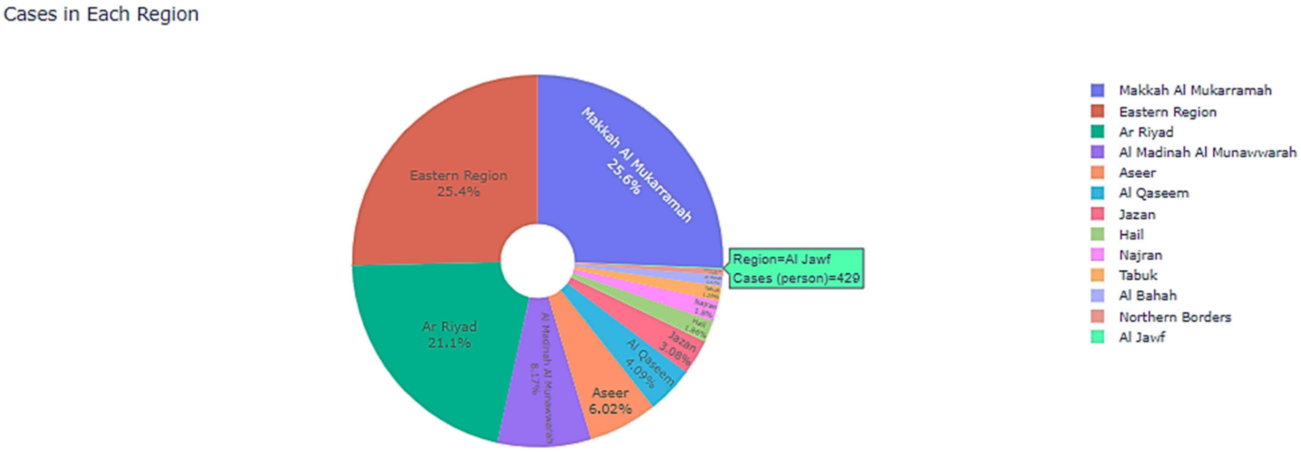
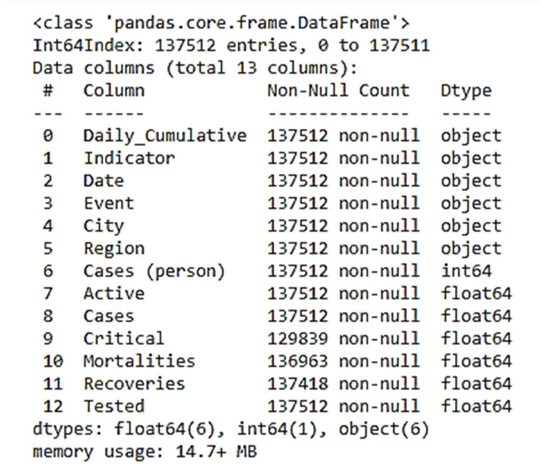


Figure 6 Donut chart representing the cases per region.

D. Data Pre-processing

In data pre-processing pandas is imported as pd and the dataset is extracted with columns and rows. They are”Daily_Cumulative”, ‘Indicator’, "Date", "Event", "City", "Region", "Cases (person)”. The length of dataset is 230. Interaction process and the value count of event column are done. Curfew lifted (all regions), this is assigned to all rows of event using mode function. Data is copied and extracted. Indicator and date column is merged the data inside the indicator column is created with new column name as Active. Another df2 is merged with df and df1. The shape of the df2 is 137512, 13 and the information of df2 is seen in Fig. 7.

Manuscript received August 5, 2021
Manuscript revised August 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.8.33

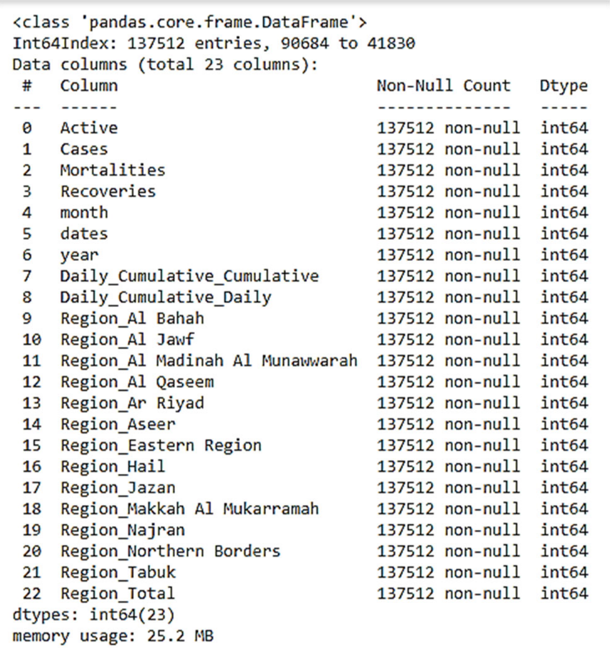


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 137512 entries, 0 to 137511
Data columns (total 13 columns):
 #   Column           Non-Null Count   Dtype
---  ------           --------------   -----
 0   Daily_Cumulative 137512 non-null  object
 1   Indicator        137512 non-null  object
 2   Date             137512 non-null  object
 3   Event            137512 non-null  object
 4   City             137512 non-null  object
 5   Region           137512 non-null  object
 6   Cases (person)   137512 non-null  int64
 7   Active           137512 non-null  float64
 8   Cases            137512 non-null  float64
 9   Critical         129839 non-null  float64
 10  Mortalities      136963 non-null  float64
 11  Recoveries       137418 non-null  float64
 12  Tested           137512 non-null  float64
dtypes: float64(6), int64(1), object(6)
memory usage: 14.7+ MB
```

Figure 7 information about the Df2 data frame.

However, in df2 the date argument is converted to date time, month, dates, and year. The memory usage is 17.8+ MB. Finally, when drop of event and indicator takes place and the memory usage will be reduced to 15.7+MB. Again, in df2 the drop of cases (person) and city takes and the memory will be 13.6+ MB. Sorting of values by date is done in df2.

The df2 I copied to df3 and the nulls are found in critical, mortalities and recoveries. A mortalities null value is replaced with median and it was 73.0. Count is 136963.000000 and mean is 66.003256. In the same way recoveries and critical null values columns are replaced with median. Median value of recoveries is 285.0. In df2, drop of tested cases and date is done and all null values are removed. Location of active in df3 is assigned to target and the new data is created with some dummy's columns daily Cumulative and Region. New data with integer types is assigned to new data as seen in Fig. 8.



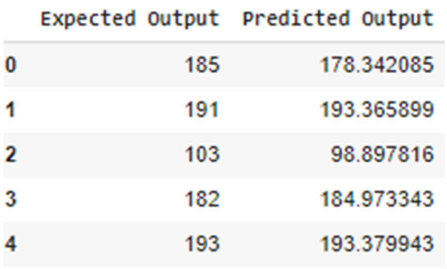
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 137512 entries, 90684 to 41830
Data columns (total 23 columns):
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   Active                            137512 non-null  int64
 1   Cases                             137512 non-null  int64
 2   Mortalities                       137512 non-null  int64
 3   Recoveries                        137512 non-null  int64
 4   month                             137512 non-null  int64
 5   dates                             137512 non-null  int64
 6   year                              137512 non-null  int64
 7   Daily_Cumulative_Cumulative       137512 non-null  int64
 8   Daily_Cumulative_Daily            137512 non-null  int64
 9   Region_Al Bahah                   137512 non-null  int64
 10  Region_Al Jawf                    137512 non-null  int64
 11  Region_Al Madinah Al Munawwarah   137512 non-null  int64
 12  Region_Al Qaseem                  137512 non-null  int64
 13  Region_Ar Riyad                   137512 non-null  int64
 14  Region_Aseer                      137512 non-null  int64
 15  Region_Eastern Region             137512 non-null  int64
 16  Region_Hail                       137512 non-null  int64
 17  Region_Jazan                      137512 non-null  int64
 18  Region_Makkah Al Mukarramah       137512 non-null  int64
 19  Region_Najran                     137512 non-null  int64
 20  Region_Northern Borders           137512 non-null  int64
 21  Region_Tabuk                      137512 non-null  int64
 22  Region_Total                      137512 non-null  int64
dtypes: int64(23)
memory usage: 25.2 MB
```

Figure 8 new data dataframe.

Finally, after this the linear regressor model is imported from sklearn and the training and testing sets which are 80% and 20% respectively are fitted to predict the output. However, the r^2 score is 0.96 and the coefficients are calculated.

IV. Conclusions and Discussions

However, linear regression model from machine learning was selected to predict the output based on the dataset. New data from 1st column is assigned to X and from 0 column is assigned to Y variable. It is seen that a satisfactory "good" r^2 score is produced and the predicted outputs are shown in the Fig. 9.



	Expected Output	Predicted Output
0	185	178.342085
1	191	193.365899
2	103	98.897816
3	182	184.973343
4	193	193.379943

Figure 9 Prediction output verses Expected output

To conclude, excepted output is 185 but with 0.96 of r2 score the model is predicting it as 178.34, which is considered good and satisfactory. However, to have a close look over all the values a graph can be crafted as shown in Fig. 10. The blue lines indicate the predicted outputs and the black line indicates the expected ones. The lines in the graph shows a lot of correspondences with each other and can be concluded to work better.

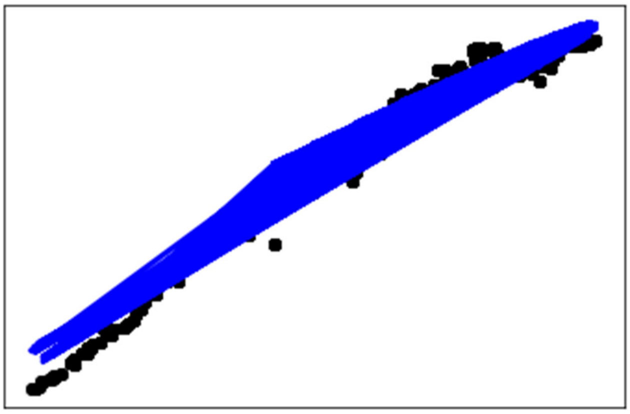


Figure 10 Graph to evaluate the expected and predicted outputs by the model

REFERENCES

[1] D. Rafiq, A. Batool, and M.A. Bazaz. "Three months of COVID-19: A systematic review and meta-analysis." *Reviews in Medical Virology* 30, no. 4 (2020): e2113.

[2] G. Barsoum. "Arab youth: the challenges of education, employment and civic paricipation." *OIDA International Journal of Sustainable Development* 5, no. 10 (2012): 39-54.

[3] H. Alahdal, F. Basingab, and R. Alotaibi. "An analytical study on the awareness, attitude and practice during the COVID-19 pandemic in Riyadh, Saudi Arabia." *Journal of infection and public health* 13, no. 10 (2020): 1446-1452.

[4] J.W. Lai, and K.H. Cheong. "Superposition of COVID-19 waves, anticipating a sustained wave, and lessons for the future." *BioEssays* 42, no. 12 (2020): 2000178.

[5] M. Sironi, S.E. Hasnain, T. Phan, F. Luciani, M.A. Shaw, M.A. Sallum, M.E. Mirhashemi, S. Morand, and F. González-Candelas. "SARS-CoV-2 and COVID-19: A genetic, epidemiological, and evolutionary perspective." *Infection, Genetics and Evolution* (2020): 104384.

[6] T. Acter, N. Uddin, J. Das, A. Akhter, T.R. Choudhury, and S. Kim. "Evolution of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) as coronavirus disease 2019 (COVID-19) pandemic: A global health emergency." *Science of the Total Environment* (2020): 138996.

[7] S. Krishnan, M. J. Franklin, K. Goldberg, J. Wang, and E. Wu. "Activeclean: An interactive data cleaning framework for modern machine learning." In Proceedings of the 2016 International Conference on Management of Data, pp. 2117-2120. 2016.

[8] M. Dallachiesa, A. Ebaid, A. Eldawy, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, and N. Tang. "NADEEF: a commodity data cleaning system."

[9] In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pp. 541-552. 2013.

[10] J. Van den Broeck, S. A. Cunningham, R. Eeckels, and K. Herbst. "Data cleaning: detecting, diagnosing, and editing data abnormalities." *PLoS Med* 2, no. 10 (2005): e267.

[11] D.C. Montgomery, E.A. Peck, and G.G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[12] R. Aggarwal, and P. Ranganathan. "Common pitfalls in statistical analysis: Linear regression analysis." *Perspectives in clinical research* 8, no. 2 (2017): 100.

[13] M. Tranmer, and M. Elliot. "Multiple linear regression." *The Cathie Marsh Centre for Census and Survey Research (CCSR)* 5, no. 5 (2008): 1-5.

[14] N. Altman, and M. Krzywinski. "Simple linear regression." (2015): 999-1000.

[15] M. Li. "Moving beyond the linear regression model: Advantages of the quantile regression model." *Journal of Management* 41, no. 1 (2015): 71-98.

[16] T. Fang, and R. Lahdelma. "Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system." *Applied energy* 179 (2016): 544-552.

[17] B. Jann. "The Blinder–Oaxaca decomposition for linear regression models." *The Stata Journal* 8, no. 4 (2008): 453-479.

[18] G.C. McDonald. "Ridge regression." *Wiley Interdisciplinary Reviews: Computational Statistics* 1, no. 1 (2009): 93-100.

[19] R. Tibshirani. "Regression shrinkage and selection via the lasso: a retrospective." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73, no. 3 (2011): 273-282.

Amr Jadi is an Associate Professor of Software Engineering at Collage of Computer Science and Engineering, University of Ha'il. Dr.Jadi received PhD degree from De Montfort University and Master Degree from Bradford University, UK. The author is specialized in with an area interest in Software Engineering, Artifical Intellegence, Early warning systems, Risk management and Critical Systems. Presently the author is also involved in various development activities within the University of Hail and abroad as a consultant.