Predicting COVID-19 Cases

Subtitle: Using Linear Regression Process to Predict COVID 19 Cases.

Tanveer Mahmood Hasan, Tasnim Rafia, Billah Syed Mashkur, K.M Zubair

Department of Computer Science, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Kuala Lampur, Malaysia.

Email: tanveer.iium@gmail.com, tasnim.rafia777@gmail.com, contactKMZubair@gmail.com, mashkurbillah@outlook.com

Abstract:

COVID-19 has already had an immense effect globally, and more than 75 million people are contaminated in more than 218 nations. A number of countries have implemented protection steps to curb its spread. However, it is not clear when the outbreak will stop in different countries or internationally. Predicting the COVID-19 pattern is a big challenge. It is one of the world's greatest safety problems ever. For potential hospital centres, public health policymakers need to accurately forecast reported events. Machine learning algorithms learn from past data to forecast incidents and can be used to estimate the amount of verified Covid-19 events. This is why doctors, physicians and health professionals around the world want to strive for innovative technologies to better combat the Covid-19 pandemic using Machine learning algorithms because its application on the past epidemic allows researchers to use a different method to tackle the new outbreak of coronavirus. A predictive disease model can assist the distribution of patient services and more accurately assess social distancing interventions. In this paper, we have used real datasets and Linear Regression Machine learning algorithms which works based on the relationship between two features by fitting a linear equation for observing the data.

Keywords: Covid-19, Linear Regression, Correlation, Error rate, Sampling.

I. INTRODUCTION:

Covid-19 was first discovered in the human body on 31 December 2019 in Wuhan city, China. But this is the first time Covid-19 was detected in the human body. In the past, it was discovered in animals' bodies for instance, bats, cats, chicken etc. It is not the same disease as pneumonia, flu, or any other cold fever. Its reproduction is very strong and during the reproduction process in the human body it can change the RNA. So, it is difficult to invent a new vaccine which can destroy the chain of reproduction. But our scientist is working hard to invent the vaccine. For the daily cases and updates visit: Covid-19.

II. BACKGROUND:

Covid-19, the cause of death of more than one million people, so far, is the most alarmed fact these days. It has become the origin of psychological disruption that is extremely terrifying. Getting rid of the pandemic has become the most widely known desire. To predict the period of discarding these dreadful days, the prediction of the number of patients in next days is genuinely a primary mission. The projection of this project is to forecast the number of a discrete number of days that will enable to anticipate the pandemic session.

III. PROBLEM OVERVIEW:

Coronavirus (COVID-19) is a disease triggered by a virus that can transmit from individual to individual. As indicated by the Wuhan Municipal Health Commission, tests from China's Wuhan's Huanan Seafood Wholesale Market were positive for covid-19. Cases showed side effects, including flu, dry hack, dyspnoea; reciprocal lung infiltrations were indicated by radiological discovery. The China CDC announced on 9 January 2020 that for 15 of the 59 cases of pneumonia, a novel coronavirus (later renamed SARS-CoV-2, the infection that triggers COVID-19) was

distinguished as the causative specialist. The main novel structure of the coronavirus genome has been made publicly available on 10 January 2020. By 20 January 2020, records of affirmed cases had been issued from three countries outside China: Thailand, Japan, and South Korea. These cases were all dispatched from China.

WHO released revised guidelines on the use of masks in households on 5 June 2020, in the field of home care and in the field of health care in areas where COVID-19 cases have been identified. This advice is consistent with the opinion of the ECDC, which was released on 8 April 2020. On 13 June 2020, in Beijing, the People's Republic of China, officials from the National Health Commission and the Beijing Health Commission announced a new cluster of COVID-19 incidents. The European Commission adopted a European approach on 17 June 2020 to speed up the turn of events, to assemble and to organise antibodies against COVID-19. ECDC circulated a rapid risk evaluation on the reappearance in European/EAE, UK and EU competitors and potential applicants of coronavirus disease 2020 instances (COVID-19) on 2 July 2020, indicating that while the pattern of disease occurrence is largely decreasing or stable in Europe, the network remains in depth.

So, we decided to utilize the COVID-19 infection as a model to our project. We accept our displaying beginning from around March until July 2020. We have taken four nations in record to do some investigation and computations which are the United States of America. Our information regularly incorporates dates of the first case, at that point how it builds, number of all out dynamic cases, number of new cases, number of recovered and also some data of dead ones among these countries.

IV. THE RESEARCH QUESTION OR HYPOTHESIS:

The number of Covid-19 patients is evidently detectable to be climbing up most recently. If it is assumed based on the number of cases of the current year, it will be found that "Next year is going to be bringing about more cases".

V. THE RESEARCH OBJECTIVES:

- Using Data Pre-processing in order to reform the covid-19 dataset to ease the research in the next steps.
- Understanding the data on covid-19 cases for data visualization.
- Extracting out the features useful in order to predict covid-19 cases in the upcoming days.
- Building up a model to point out the risk factors in order to predict covid-19 cases on the upcoming days.
- Finally, testing the model for any errors involved for better prediction and checking out its performance.

VI. THE RESEARCH SIGNIFICANCES:

- Predicting the possibility of new covid-19 cases in the upcoming days would allow us to be prepared in order to take possible safety needed.
- Pointing out the risk factors would highly assist to predict covid-19 cases on the next day that could allow the governments of the countries to take necessary steps.
- Predicting new covid-19 cases could also help doctors at the hospitals to assume approximately how many patients could be admitted and take preparations accordingly.
- Advance alert of ventilation requirements may be used to enhance patient outcomes of COVID-19.
- Substantially reducing the transmission of Covid-19.
- Prediction of covid-19 cases would also assist to point out the safe zones or locations to travel to.

VII. THE RELEVANT WORKS/LITERATURE REVIEW:

For this section, we are presenting relevant recent papers discussing the implementation of the process and methodologies to solve a pandemic problem. Through scanning the online directories, we retrieved, screened, and interpreted the material of certain journal papers by an iterative approach to theory. The results of the survey are stated by detailing the contents of the papers about key machine learning methods, source data and predictive output consistency.

In data science, we use a hypothetical function, f′, $\hat{y} = f'(x)$ is used to formulate various techniques which are eventually modified to make the difference between y, the actual expected output and $\hat{y}$, the output of the function, as little as possible. The f′ function is generally referred to as an ML model. Most of the techniques create and refine a model f′ by providing a function constructing several of the pairs ⟨x, y⟩, which is a collection of input (x) data that is already

named and associate them to their right y (being y either yes or no, 1 or 0). If the weight (y) is predicted by height (x). Our variant of linear regression for this problem will be:

$$y = B0 + B1 * x1$$

Gitanjali and Asmita along with others [1] did a research in 2020 on forecasting models for Covid 19 to control the global outreach of pandemic and found out that The Mathematical models inferred 44% of the transmission happened without showing any prior symptoms. The Data Science models forecasted suspected members of Covid-19, death counts, Impact on Public Health, Identify individuals who are at paramount risk and predicted the infection rate in China. Population Statistics helped to make reliable predictions having less computational overhead. They have used the best Machine Learning models for prediction and used natural progression of disease for better accuracy. Big Data Analysis and for the study of large data, Deep Learning is used from WHO or other national resources. Used Population Statistics to view data. These models and predictions will help to prepare against probable tragedies and consequences in other similar pandemics in the future.

Additionally, Akib Mohi, Syed Tanzeel and his group [2] implemented the Machine learning-based methods to utilizing clinical text knowledge to identify COVID-19. They used machine learning algorithms such as Naïve multinomial (MNB), decision tree, vector machine help (SVM), Logistic regression, Bagging, Classification of Random Tree, Adaboost and Stochastic gradients to complete this analysis. The outcome was interesting. Their logistic regression and multi-installed Naive Bayes have 94% accuracy, 96% recording, 95% f1 score and 96,2% test accuracy and algorithms such as random forest and gradient improves 94,3% accuracy. In future, recurrent neural networks may be used for greater accuracy and the authors believe that the more data they get the more accurate their models will become. They also tend to use a Deep learning approach in the future.

Moreover, Dasari Naga Vinod and Prabaharan [3] worked on using Data Science to achieve fast diagnosis of Covid-19. They used AI for the X-Ray chest and the CT scan to identify Covid-19. Using tree scores for faster prediction. Usage of X-ray photos and CT images via deep learning techniques. In CT scanning photos, the suggested method achieved 93 percent precision and x-ray of chest pictures produced 88 percent accuracy rate and also offers quicker outcomes during preparation and inference. This research can be used to detect other diseases or infections that are related to lungs.

Besides, Peipei Wang, Xinqi Zheng and others [4] published a research on prediction with the logistic method and machine pattern learning methods for COVID-19. The most revised epidemiological information from COVID-19 was incorporated into the Logistic Model before 16 June 2020 to meet the epidemic pattern and then the cap value was added to the FbProphet Model, a machine-based time series prediction model that outlines the disease curve and forecasts an event. These models could help to forecast potential cases that are verified if the distribution of the virus does not shift beyond expectation. It tends to carry out tests with the global pandemic and also in some countries in order to estimate the disease high, the fast-rising point and the recovery point.

Last but not least, Aboul Ella Hassanien, Guesh Dagnew and their team [5] researched on using weather data with Machine learning to predict mortality rate. The authors used the machine training model to consider the impact of temperature and moisture on the distribution of Covid-19, Regression Analysis (LR), Regression Models, Elastic Net, Lowest Angle Reversion, Lowest Angular Regression, Sequential Pattern Pursuit, Probabilistic Ridge, Automatic Significance Estimation and Random Sample Consensus. Experimentally, it has been shown that the atmosphere, the variables of humidity and temperature have an inverse relation to the number of incidents confirmed. The ML model predicts the relationship on the basis of the number of cases reported as of 17 May 2020 in Italy. For each unit rise in individual variable temperature, the number of COVID-19 cases rises by around 143 percent. The authors are planning to consider additional features like wind speed and rainfall to improve their model's accuracy. Essential measures can be taken by predicting mortality rate on similar pandemics in the future.

The aforementioned research papers used different techniques that can be used for predicting Covid-19 outbreaks. The most commonly used technique is machine learning and the model is regression. As linear regression has a higher accuracy, we're planning to use these algorithms for our project.

VIII. METHODOLOGY:

In this project we are trying to do our analysis by using real data sets. So, we collect our data from different online dataset sources. Moreover, this data set can give more or less accuracy to analyse the real outcomes. To avoid less accuracy, we endeavour to collect more data.

A. Data Collection and Splits:

Dataset is collected from Kaggle. Data is uploaded by srk from covid-19 Tracking project and New York Times.

we split data according to State. There are 49 States. And we will work on three states only. procedures are quite straightforward.

B. Tools:

R, RStudio and Python.

C. Machine Learning Algorithm:

We are going to use Linear Regression. Linear Regression works based on the relationship between two features by fitting a linear equation for observing the data. One feature is considering the independent variable and other is considering the defendant variable.

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{xip} + \varepsilon_i = x_i^T \beta + \varepsilon_i$$
$$i = 1, \ldots, n$$

Formally, the terminologies of Linear Regression are as follows:

- T represents the transpose.
- $x_i^T \beta$ represent the inner value between vector x_i and β

it can also write as $y = X\beta + \varepsilon$ *where,*

- y is a dependent variable.
- x is an independent or Random variable.

D. Training and testing Datasets:

80% of the U.S.A (State_CA) Dataset will be used for training and 20% for validation. Which consider as the standard of splitting the data.

E. Feature Selection:

In our dataset we can see there are 15 features. So, we have selected three columns only to make our prediction. All the features are interrelated. To ensure our model with the best accuracy, we have selected 'Date', 'New case' and 'Comulative_Case'. From the dataset we split our data according to the state. We consider 'Date' and 'New Case' as the independent variable and 'Cumulative _case' as the dependent variable.

F. Sampling:

Since we have collected our data recently, our data is updated. In the dataset the record of COVID-19 cases has collected from 22nd January 2020 to 11th December 2020. This is specially a feature where new cases are more likely to be recent cases. We conduct our research and analysis based on the state.

IX. MODELING:

A. Exploratory Data Analysis (EDA):

EDA is made for understanding data better. It helps to understand data with statistical measures. From our dataset we select only three states which are California, New York, and Washington. EDA will be implemented to visualize the circumstance better.

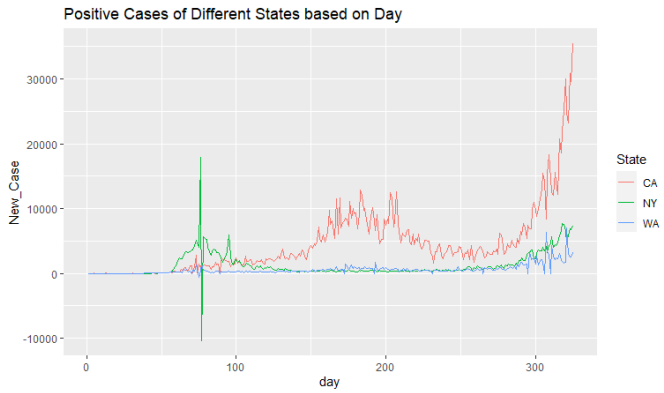


Figure: Number of cases based on days

The figure above is showing the overall situation of different states regarding Covid-19 positive cases since the very beginning of 2020. Where red line indicates the positive cases of covid-19 in California, Green line indicates the positive cases of covid-19 in New York and blue line indicates the positive cases of covid-19 in Washington over the month. Furthermore, California started to get more and more affected cases at day 80 approximately. New York was in pretty big range at the same time but that reduced all of a sudden. Whereas

Washington was in a reasonable range during the whole period of 2020.

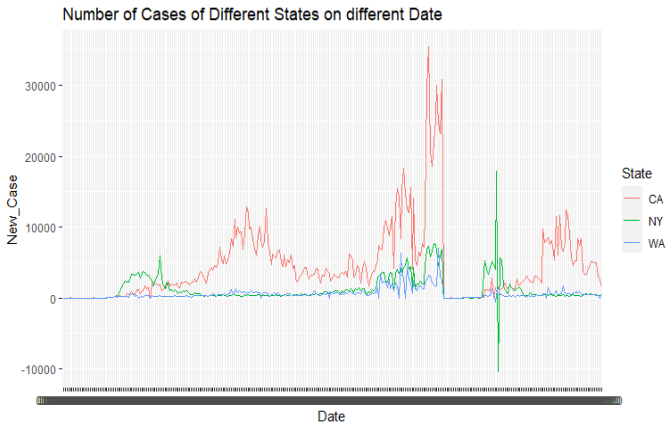


Figure: Number of cases in different date

The line graph above is explaining the cases in 3 different states in each date in brief.

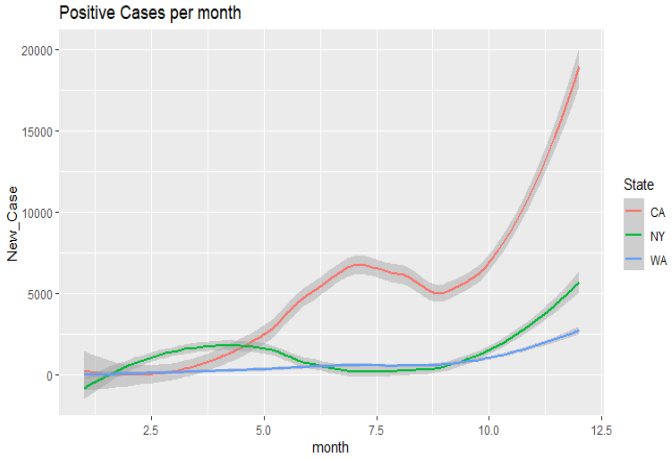


Figure: New Positive Cases Per Month

The third graph over here shows the new cases according to the month and state. From the graphs it is extremely notified that the increment of new cases is higher in California rather than New York and Washington. The transmission of COVID-19 in California is overall greater than other states.

The purpose of the project is to predict the number of positive Covid-19 cases in these following states. To determine the number of future cases, it is absolutely important to observe the past and present cases in these places. In order to observe the past, exploratory data analysis was occupied in addition with some graphs from which the condition is detectable for further prediction.

B. Building The model:

```
> summary(California)

Call:
lm(formula = New_Case ~ day, data = state_CA)

Residuals:
    Min      1Q  Median      3Q     Max
-6533.2 -1692.6  -293.6  1038.4 24756.1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1606.974    444.036  -3.619 0.000343 ***
day            37.904      2.361  16.054  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3993 on 323 degrees of freedom
  (30 observations deleted due to missingness)
Multiple R-squared:  0.4438,    Adjusted R-squared:  0.4421
F-statistic: 257.7 on 1 and 323 DF,  p-value: < 2.2e-16
```

We applied a Linear Regression method. From Linear Regression we got this equation.

Number of cases = intercept + (day * Nth days)

```
> Prediction_nDays
[1] 9764.226
```

here we predict the new cases of day 300th. So now let's check the actual new cases of day 300th.

```
> state_CA$New_Case[300]
[1] 9890
>
```

This number is pretty close. Although the model is quite simple, it has so much flexibility. All we need to do is to run a loop to find out the number of cases after some given days.

```
> print(pred)
 [1] 10711.83 10749.73 10787.63 10825.54 10863.44 10901.35 10939.25
 [8] 10977.15 11015.06 11052.96 11090.87 11128.77 11166.67 11204.58
[15] 11242.48 11280.39 11318.29 11356.19 11394.10 11432.00 11469.91
[22] 11507.81 11545.71 11583.62 11621.52 11659.43 11697.33 11735.23
[29] 11773.14 11811.04
>
```

After checking with the actual numbers, we found that they are close to each other.

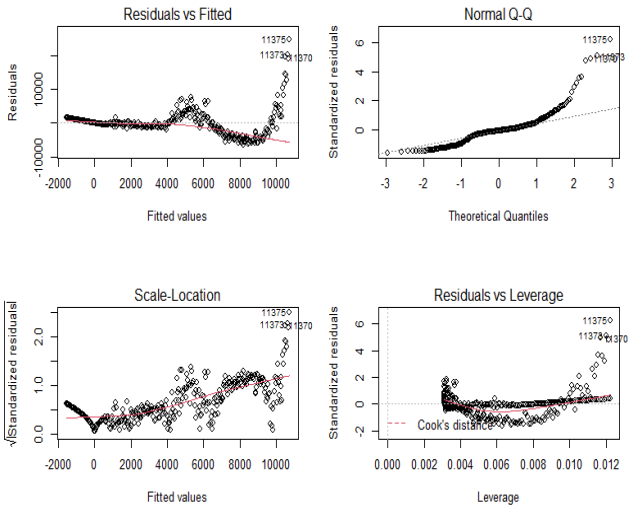
Going back to our model we can notice the following:

- RSE is 3993, that means we can make a prediction mistake about 3993 up or down on average.
- R-squared is 0.4438, that means our model can explain almost 45% of the variance in the response, which is not quite a good percentage.

- Both the intercept and the Day variable have a significant relationship with the response because they both have p-value smaller than 0.05.

- The p-value for F-statistic is quite small, indicating that the relationship is pretty strong.

Now, let us figure out more about our model, by visualizing the results:



We can see that the model has an obvious non-linearity issue, and we want to come over this problem, so maybe we can try transformation of the predictor.

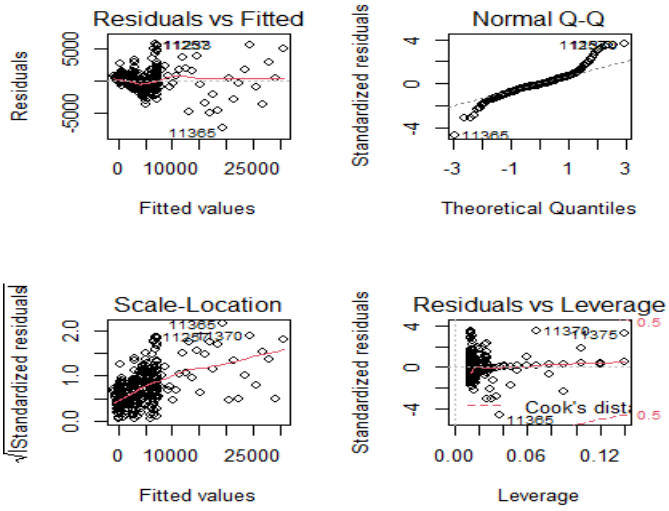
```
Call:
lm(formula = New_Case ~ day + I(day^2) + I(day^3) + I(day^4) +
    I(day^5) + I(day^6), data = state_CA)

Residuals:
    Min      1Q  Median      3Q     Max
-7263.4  -702.9  -163.9   722.8  5593.8

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.762e+02  6.470e+02  -1.354   0.1766
day          1.085e+02  5.457e+01   1.989   0.0475 *
I(day^2)    -3.022e+00  1.450e+00  -2.084   0.0380 *
I(day^3)     3.025e-02  1.666e-02   1.816   0.0703 .
I(day^4)    -7.996e-05  9.283e-05  -0.861   0.3897
I(day^5)    -1.296e-07  2.471e-07  -0.525   0.6003
I(day^6)     5.422e-10  2.518e-10   2.153   0.0321 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1605 on 318 degrees of freedom
  (30 observations deleted due to missingness)
Multiple R-squared:  0.9115,    Adjusted R-squared:  0.9099
F-statistic:   546 on 6 and 318 DF,  p-value: < 2.2e-16
```

We have tuned our model for acquiring the best accuracy. Now we can see the differences. We got Residual Standard Error (RSE) is 1605 and the multiple of R-Squared is 0.9115. That means the prediction is better, and the percentage of prediction is pretty good. So, now we can visualize our latest model.



We can see now that we came over the problem of non-linearity. To sum up, our inferences show that, a number of new cases can be represented by the number of days passed from a certain date.

X. RESULTS:

A. Correlation testing:

This dataset included many variables. But we can only see the correlation with our targeted variable and day. So, we check whether there is any correlation between day and New case. It has been observed that most variables are categorical which also include the targeted variable.

```
> cor(State_CA$day,State_CA$'New_Case')
[1] 0.6661933
>
```

Above figure shows the correlation between Days and New cases. Where we can see that the correlation score is near 70% which is pretty good. It is near to the 1. Therefore, the linear model is statistically significant and

B. P value testing and Hypothesis testing:

From our data set we got the P value 2.2 e-16 which is less than 0.05. So, it is clear that our prediction model is linearly significant.

```
Adjusted R-squared:  0.9099
DF,  p-value: < 2.2e-16
```

It can also be said that we easily reject the null hypothesis (H_o).

C. Accuracy:

As we mentioned before that the model was fitted with the train data of 80% and test data of 20%. which is standard. The data point selected randomly.

```
> correlation_accuracy
             actuals predicteds
actuals    1.0000000  0.9203675
predicteds 0.9203675  1.0000000
```

So, after doing the correlation accuracy testing between the actual and predicted values, we can get an accuracy of 92% which is acceptable. Moreover, the actual and predicted values have similar directional movement. This implies that it is a linear relation, and the linear model is intact to be used.

D. Error Rate:

We used Mean Absolute Percentage Error (MAPE) in order to calculate the error rate from our prediction model.

```
> MAPE(actuals_preds$actuals, actuals_preds$predicteds)
[1] 0.7130458
>
```

As we can see from the above figure that our Mean Absolute Percentage Error (MAPE) rate is approximately 70%. Which is a moderate approach to predict the results. However, as our rest of the results were good such as accuracy, prediction and so on. Therefore, we can state that the final model is not ignorable.

On the other hand, we evaluated the prediction model with the application of R-squared/R2 in order to estimate how well the model is fitted on the linear regression line.

```
> R2(actuals_preds$predicteds, actuals_preds$actuals, form = "traditional")
[1] 0.8395898
> summary(California2)$r.squared
[1] 0.911523
>
```

From the above figure the result for both R2 and R-Squared or Coefficient of Determination are 0.83 and 0.91 which is around 90%. This proves that the data which have been taken for creating the model are highly fitted to the regression line and so the linear regression model fits our data well concluding it to be strong.

XI. DISCUSSION:

The purpose of this analysis is to estimate the number of new Covid-19 cases in future on the upcoming days. If we see the statistics, we can see the increment of COVID-19 is significantly rising. By using the dataset of United States of America covid-19 cases, selecting a specific state (California) for evaluation/testing, and applying the linear regression model for prediction, we came up with some acceptable results which are interesting as well. In order to evaluate the performance, we used various processes such as Mean Absolute Percentage Error (MAPE), Coefficient of Determination (R-Squared) and also the correlation accuracy. On the other hand, correlation function is used wisely and properly to verify the linear relation between the variables that are taken to build the LR model. In order to build the model, we took two variables as our independent ('days' variable) and non-independent ('New_Cases' variable) values as we are going to predict the number of new cases in California based on the upcoming days. However, we have applied the model for the possibility of the new case at certain days and the model gave us the amazing output which is pretty close to the actuals and the accuracy of our model is above 90 percent which cannot ignorable.

XII. FUTURE WORKS:

In the probable future, we are intrigued to increase the accuracy of our Linear Regression classifiers by adjusting the tuneable parameters, applying some classifier combination techniques like bagging, boosting or ensembling. Also, we need to brush-up our basic parsing, refine the selected features from data and obviously add more data for more desirable results. Apart from that, we will use Fisher method to optimize our Linear Regression classifier since it can measure the probability of a class with each aspect of a report and estimates the combined probabilities with the probability of a random collection of features.

Although this is not the 100% accuracy of our prediction because there is a shortage of resource and data. Moreover, we cannot make it 100%. However, for our analysis we just give a concept using some sort of prediction and calculation, which could help people in order to maintain a safe and healthy lifestyle in the future.

XIII. REFERENCES:

[1] R., G., B., A., N, P., & Dey, N. (2020). Forecasting Models for Coronavirus Disease. Retrieved November 09, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7289234/

[2] F. Wu, S., N. Chen, M., A. Kumar, V., P. Verma, A., S. Chakraborti, A., A. Sarwar, M., Friedman, J. (1970, January 01). Machine learning based approaches for detecting COVID-19 using clinical text data. Retrieved November 09, 2020, from https://link.springer.com/article/10.1007/s41870-020-00495-9

[3] Vinod, D., & Prabaharan, S. (2020, July 30). Data science and the role of Artificial Intelligence in achieving the fast diagnosis of Covid-19. Retrieved November 09, 2020, from https://www.sciencedirect.com/science/article/pii/S0960077920305786

[4] Wang, P., Zheng, X., Li, J., & Zhu, B. (2020). Prediction of epidemic trends in COVID-19 with logistic models and machine learning techniques. Chaos, Solitons & Fractals, 139, 110058. https://doi.org/10.1016/j.chaos.2020.110058

[5] Malki, Z., Atlam, E., Hassanien, A., Dagnew, G., Elhosseini, M., & Gad, I. (2020, July 17). Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. Retrieved November 09, 2020, from https://www.sciencedirect.com/science/article/pii/S0960077920305336

[6] Dataset: https://www.kaggle.com/sudalairajkumar/covid19-in-usa?select=us_covid19_daily.csv