IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

Covid-19 Analysis and Future Forecasting

Aditi Vadhavkar[1], Pratiksha Thombare[2], Utkarsha Auti[3], Priyanka Bhalerao[4], Prof. Sagar Dhanake[5]
Students, Computer Engineering[1, 2, 3,4]
Assistant Professor, Computer Engineering[5]
D Y Patil Institute of Engineering and Technology, Ambi, Pune, Maharashtra, India

Abstract: *Forecasting Mechanisms such as Machine Learning (ML) models have been demonstrating their value in predicting perioperative outcomes in the domain of future course of action decision making. Many application domains have seen the usage of machine learning models for identification and prediction. A threat's unfavourable factors are prioritised COVID-19 has proven to be a serious hazard to humans. A worldwide pandemic has been declared by mankind. Many assets around the world have had problems. This sickness has a high infectivity and contagiousness. Take a look at the diagram of undermining elements. We used four Machine Learning Models in COVID-19: Linear Regression (LR), Support vector machine (SVM) and Bayesian Ridge. The results show that of the four models used in this work, the LR performs the best, followed by the Bayesian and SVM, which both perform well in forecasting newly confirmed cases, death rates, and recovery rates.*

Keywords: COVID-19, Supervised Machine Learning, Forecasting, Machine Learning

I. INTRODUCTION

Machine Learning has consistently demonstrated its importance as a subject of research over the last decade by tackling a wide range of complex and sophisticated real-world challenges. Healthcare, autonomous vehicles (AV), business applications, natural language processing (NLP), intelligent robotics, gaming, climate, modelling, voice, and image processing are some of the real-world sectors where machine learning is used. Unlike traditional algorithms, which follow programming instructions, machine learning algorithms function on a trial-and-error basis. Forecasting is one of the most important fields of machine learning, and many standard algorithms have been used to direct future actions in areas such as weather forecasting, disease prognosis, and other forecasting domains. This research focuses on the realtime prediction of COVID-19 confirmed cases and outbreaks, as well as early reaction. These forecasts can aid decision-making in dealing with current conditions and advise early efforts to effectively limit the pandemic's spread.

II. DESCRIPTION OF THE PROBLEM

Problem Definition

To investigate future COVID-19 epidemic forecasts, with a focus on the number of newly infected cases, deaths, and recoveries.

Motivation

Millions of individuals throughout the world have been affected by the increasing spread of new corona virus, popularly known as SARS-CoV-2 and formally termed COVID-19 by the World Health Organization. It has influenced people's lifestyles, work cultures, education, and so on. Studying the outbreak and early response based on the number of new positive cases, deaths, and recoveries can aid us in developing better methods to combat the global pandemic.

II. SYSTEM DESIGN AND FLOW

2.1 Dataset

The number of positive cases, the number of deaths, and the number of COVID-19 recoveries are the subject of this study. The dataset used in this study is available on covidtracking.com, which is the main COVID-19 statistics website in the United States.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

2.2 Models of Supervised Machine Learning

Supervised Machine Learning is a kind of machine learning that uses labelled datasets to train algorithms that reliably classify data or predict outcomes. The models used are:

- Linear Regression (LR)
- Support Vector Machine (SVM)
- Bayesian Ridge

A. Linear Regression

Linear Regression is a machine learning approach based on supervised learning that conducts a regression job using independent variables supporting the target prediction value. It's most likely utilised for determining the relationship between variables and forecasting.

$$y = \beta 0 + \beta 1x + \varepsilon$$

$$E(y) = \beta 0 + \beta 1x$$

Where, ε is the linear regression error term, and (x, y) represents the input training dataset and sophistication labels within the input dataset, respectively.

B. Support Vector Machine (SVM)

SVM is a basic machine learning method that can be applied to classification and regression problems in the domain. The most function maps to the pliability to tell apart between two classes. It's a model that can generalise between two separate classes if the set of labelled data is provided within the training set to the algorithm. It is the ability to solve linear and nonlinear issues and perform well in a variety of situations.

C. Bayesian Ridge

By modelling linear regression using probability distributors rather than point estimates, Bayesian regression permits a natural mechanism to survive limited or poorly distributed data. Instead of being estimated as a single value, the output or response 'y' is supposed to be chosen from a probability distribution.

$$p(y| X,w,\alpha)=N(y| Xw,\alpha)$$

2.3 Evaluation Parameters

We'll use assessment measures like R squared score, Adjusted R-square, MAE, MSE, and root mean square error to assess each learning model's performance. These measurements show us how accurate our forecasts are and how much they differ from the actual numbers in our data.

A. R-Squared Score

In simple terms, the R-squared score corresponds to the percentage of variance in the dependent variables that can be explained by the freelance variables. As a result, if it's 100 percent, the two variables are completely connected with no variance.

B. Adjusted R-square

This has been adjusted for the number of predictors in the model because it is a modified form of R-square.It increases (rises) when a predictor improves the model by less than predicted and drops (decreases) when the new term witnesses the model more than would be expected by chance.

C. Mean Absolute Error

In statistics, mean absolute error can be a development and can be used to put up a model analysis metric for regression models. With regard to a look at set, the mean of all the values of the individual prediction errors on a look at set is relevant. It calculates the average magnitude of errors among a set of projections without taking into account their direction.

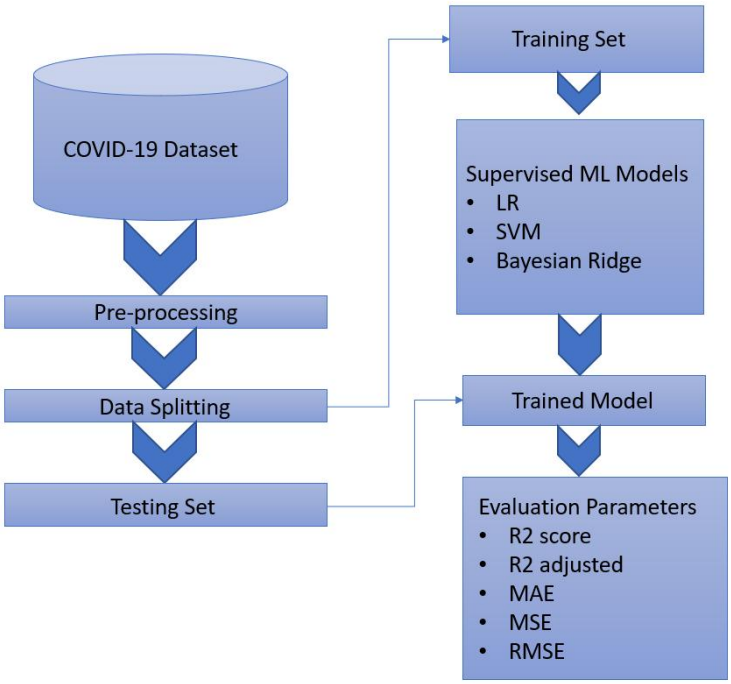
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

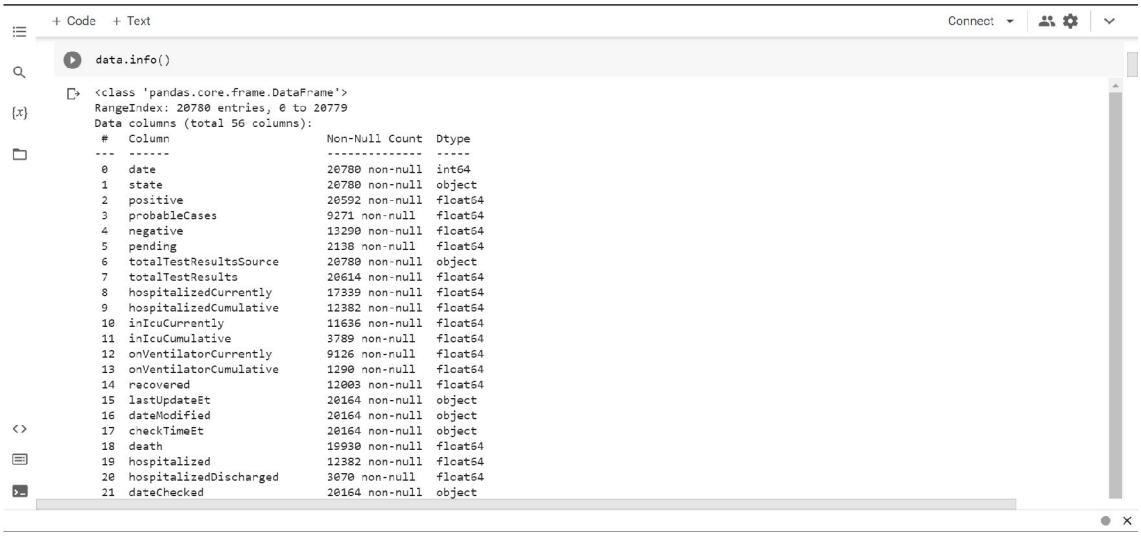
D. Root Mean Square Error

The base of the mean of the square of all errors gives a good live of accuracy, but it's only used to look at prediction errors of different models. In general, a lower RMSD number is preferred over the preceding one.

III. PROJECT IMPLEMENTATION

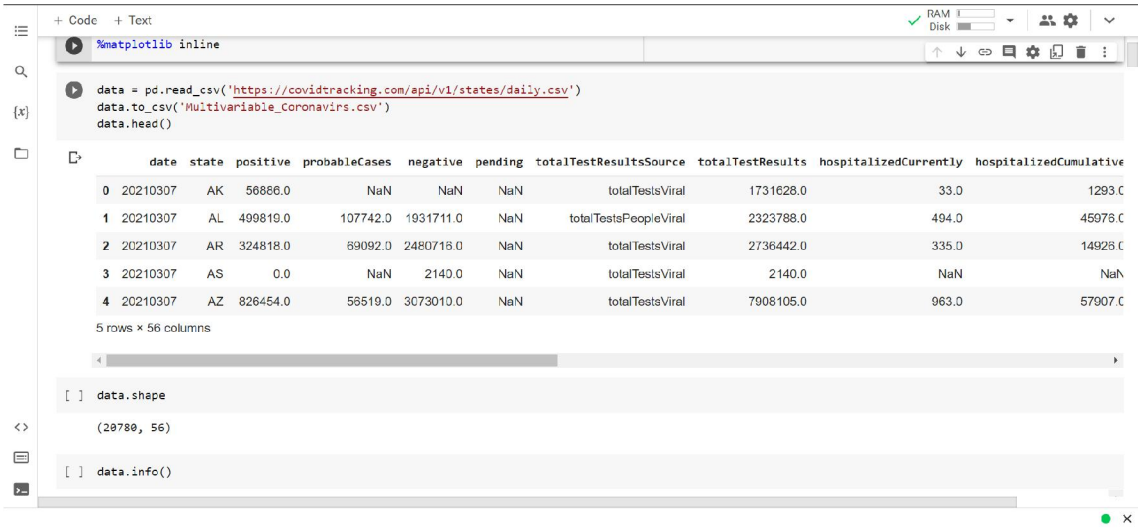
COVID-19 predictions, also known as new corona virus, are the goal of this investigation. With tens of thousands of deaths per day and death rates rising dramatically over the world, it has shown to be a potential threat to humanity. This study tries to forecast the amount of newly infected patients, death rates, daily number of cases, and recovery each day in order to help with the pandemic.





```
+ Code  + Text                                                    Connect

data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20780 entries, 0 to 20779
Data columns (total 56 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   date                        20780 non-null  int64
 1   state                       20780 non-null  object
 2   positive                    20592 non-null  float64
 3   probableCases               9271 non-null   float64
 4   negative                    13290 non-null  float64
 5   pending                     2138 non-null   float64
 6   totalTestResultsSource      20780 non-null  object
 7   totalTestResults            20614 non-null  float64
 8   hospitalizedCurrently       17339 non-null  float64
 9   hospitalizedCumulative      12382 non-null  float64
 10  inIcuCurrently              11636 non-null  float64
 11  inIcuCumulative             3789 non-null   float64
 12  onVentilatorCurrently       9126 non-null   float64
 13  onVentilatorCumulative      1290 non-null   float64
 14  recovered                   12003 non-null  float64
 15  lastUpdateEt                20164 non-null  object
 16  dateModified                20164 non-null  object
 17  checkTimeEt                 20164 non-null  object
 18  death                       19930 non-null  float64
 19  hospitalized                12382 non-null  float64
 20  hospitalizedDischarged      3670 non-null   float64
 21  dateChecked                 20164 non-null  object
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252



```
%matplotlib inline

data = pd.read_csv('https://covidtracking.com/api/v1/states/daily.csv')
data.to_csv('Multivariable_Coronavirs.csv')
data.head()
```

	date	state	positive	probableCases	negative	pending	totalTestResultsSource	totalTestResults	hospitalizedCurrently	hospitalizedCumulative
0	20210307	AK	56866.0	NaN	NaN	NaN	totalTestsViral	1731628.0	33.0	1293.0
1	20210307	AL	499819.0	107742.0	1931711.0	NaN	totalTestsPeopleViral	2323788.0	494.0	45976.0
2	20210307	AR	324818.0	69092.0	2480716.0	NaN	totalTestsViral	2736442.0	335.0	14928.0
3	20210307	AS	0.0	NaN	2140.0	NaN	totalTestsViral	2140.0	NaN	NaN
4	20210307	AZ	826454.0	56519.0	3073010.0	NaN	totalTestsViral	7908105.0	963.0	57907.0

5 rows × 56 columns

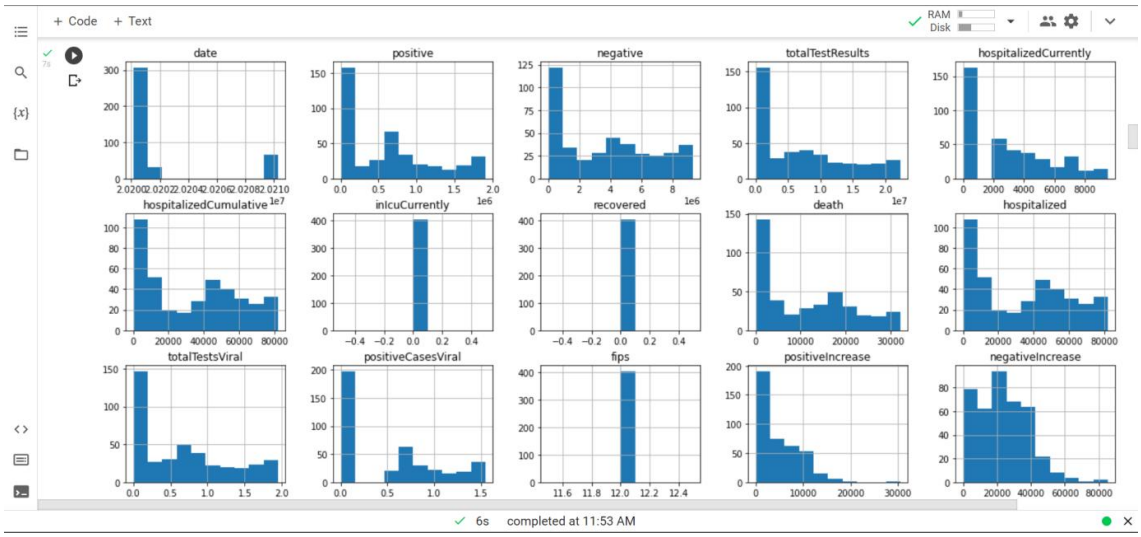
```
data.shape
```

```
(20780, 56)
```

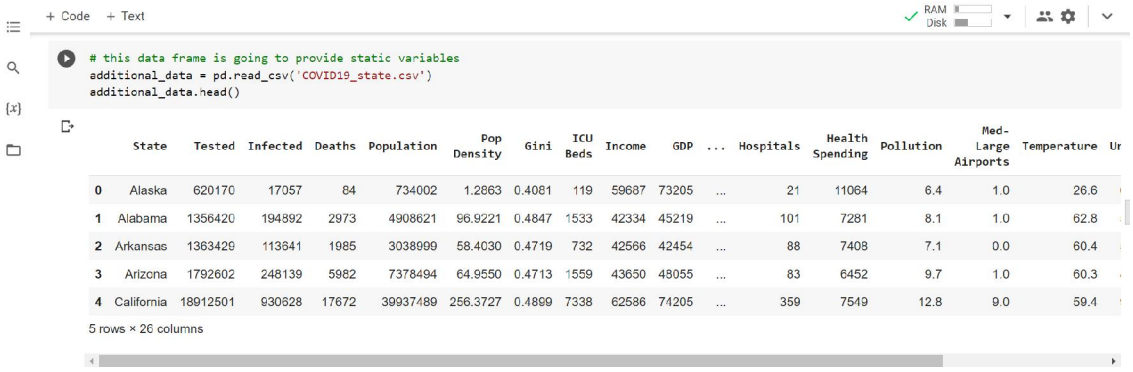
```
data.info()
```

These above images are of the dataset that we took from covidtracking.com only for the United States. We handle all the null values from the dataset to get a more accurate model. All the data is arranged chronologically. For more precision we drop out the columns who has more than 50% null values.

Below bar chart is of the state Florida which depicts the cases with all different parameters like positive cases, negative cases, total test results, total number of patients hospitalised currently, number of recovered people, total number of deaths, total number of hospitalised cases, etc. This gives us an idea of how the data is contained within the dataset. But there are some features which have few or no variations which we'll drop.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252



```
# this data frame is going to provide static variables
additional_data = pd.read_csv('COVID19_state.csv')
additional_data.head()
```

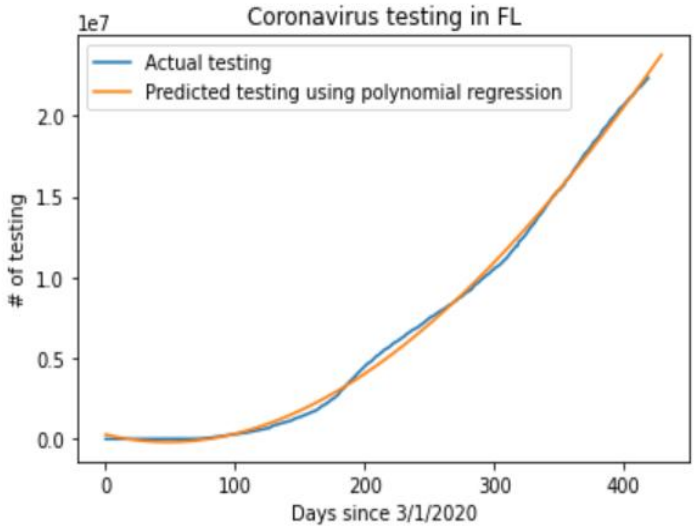
	State	Tested	Infected	Deaths	Population	Pop Density	Gini	ICU Beds	Income	GDP	...	Hospitals	Health Spending	Pollution	Med-Large Airports	Temperature	Ur
0	Alaska	620170	17057	84	734002	1.2863	0.4081	119	59687	73205	...	21	11064	6.4	1.0	26.6	
1	Alabama	1356420	194892	2973	4908621	96.9221	0.4847	1533	42334	45219	...	101	7281	8.1	1.0	62.8	
2	Arkansas	1363429	113641	1985	3038999	58.4030	0.4719	732	42566	42454	...	88	7408	7.1	0.0	60.4	
3	Arizona	1792602	248139	5982	7378494	64.9550	0.4713	1559	43650	48055	...	83	6452	9.7	1.0	60.3	
4	California	18912501	930828	17672	39937489	256.3727	0.4899	7338	62586	74205	...	359	7549	12.8	9.0	59.4	

5 rows × 26 columns

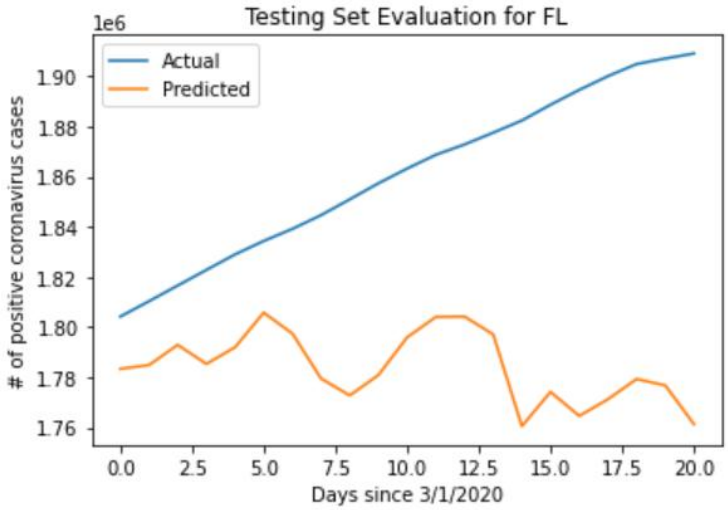
We use one more dataset for the population density of the United States.

```
[ ]  #Let's take a look at California's population density

[ ]  additional_data[additional_data.State=='CA']['Pop Density']

4    256.3727
Name: Pop Density, dtype: float64
```

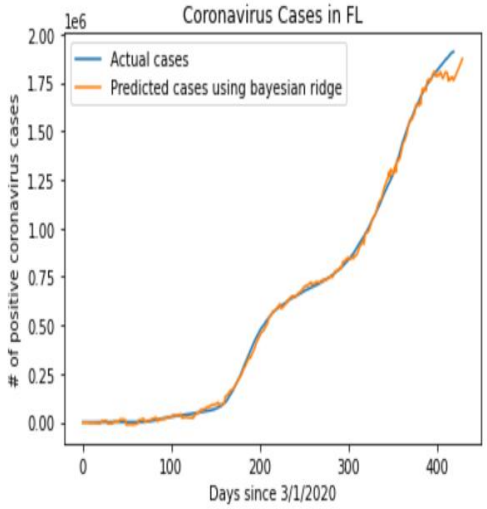
For example, we check the population density of California. It is always better to normalise the data between 0 and 1 for better processing. The mobility dataset is used for having the current update to our dataset in terms of everyday changing values. After cleaning, handling the nulls and arranging them in order according to the date, we then proceed to the future forecasting part of it. In this study, we are predicting the next 10 days of COVID-19 variation in the States. The data is divided into 80%-20% as training and testing set. The models are trained by using supervised Machine Learning Algorithms Viz. Linear Regression, Bayesian Ridge and Support Vector Machine. We use the evaluation parameters as Mean Absolute error and Mean Squared Error. We plot the Actual Testing and Predicted Testing using Bayesian and Linear Regression.

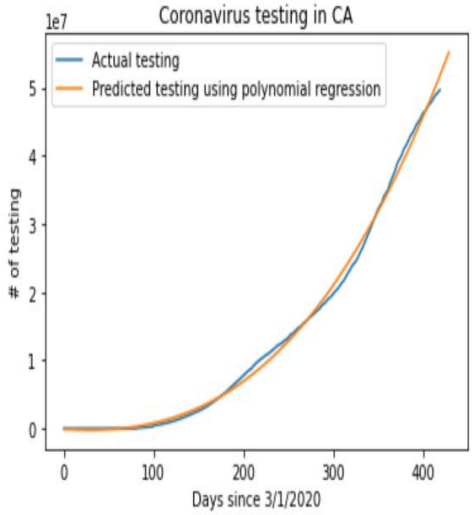


IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252




```
MAE: 76940.04920524596
MSE: 7589196634.891494
Weight: [ 8.67012183e+03 -3.44805958e+03  3.00058501e+02  5.52175530e+03
 -4.46858198e+03 -3.52666537e+03  3.70691292e+01 -1.20636572e+02
  2.11307652e+04  3.85452699e+03  1.22326102e+03  1.19762842e+01
  1.91572581e+02 -1.54843686e+02 -1.21900142e+02 -1.88627268e+03
  2.20950846e+03  5.44666573e+03 -1.11958422e+03 -2.93029253e+03
 -1.24957412e+04  1.59662497e-01  1.23656539e+00 -2.16932947e+02
 -3.61506363e+01 -6.05640277e+01 -3.92759167e+00  7.33845387e+02
  1.34157999e+02  4.27647818e+01  3.61167928e+04  2.02488449e+04
  1.15772920e+04  1.25501514e+03 -3.43400411e+03  5.08533709e+03
  3.83128156e-01  6.64195859e+00 -5.37520890e+00 -4.24426870e+00
 -6.54718243e+01  7.67161031e+01  1.89069219e+02 -3.88430935e+01
 -1.01721927e+02 -4.33760341e+02  1.01511677e+03 -8.14448533e+02
 -1.88853413e+03  4.93280110e+02  2.11417099e+03  5.36217491e+03
  6.06946094e+03  1.61040769e+03 -5.33434932e+03 -1.33488532e+04]
```





This shows the Actual vs the Predicted testing for Florida with its Mean Absolute Error, Mean Squared Error and Weights

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 6, May 2022
Impact Factor: 6.252

VII. CONCLUSION& FUTURE WORK

The COVID-19 pandemic's randomness has the potential to produce a catastrophic worldwide calamity. Some studies throughout the world have discovered that a pandemic might harm a huge section of humanity. In this paper, a prediction system based on machine learning techniques is provided for historical and prospective COVID-19 pandemic predictions.Using machine learning algorithms, the system analyses a dataset containing historical data and predicts future events. The results show that LR outperforms Bayesian Ridge and SVM in prediction in the current forecasting domain. Overall, we believe that the model prediction is accurate based on the dataset values, which may serve as a guiding light for better decision making throughout.

ACKNOWLEDGEMENTS

The completion of our project brings with it a sense of satisfaction, but it is never complete without those people who made it possible and whose constant support has crowned our efforts with success. One cannot even imagine our completion of the project without guidance and neither can we succeed without acknowledging it. It is a great pleasure that we acknowledge the enormous assistance and excellent co-operation to us by the respected personalities.

REFERENCES

[1]. G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," in Proc. Eur. Bus. Intell. Summer School. Berlin, Germany: Springer, 2012, pp. 62–77.

[2]. F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, "Regression models for prognostic prediction: Advantages, problems, and suggested solutions," Cancer Treat. Rep., vol. 69, no. 10, pp. 1071–1077, 1985

[3]. P. Lapuerta, S. P. Azen, and L. Labree, "Use of neural networks in predicting the risk of coronary artery disease," Comput. Biomed. Res., vol. 28, no. 1, pp. 38–52, Feb. 1995.

[4]. K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "cardiovascular disease risk profiles," Amer. heart J., vol. 121, no. 1, pp. 293–298, 1991.

[5]. F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID 19," PLoS ONE, vol. 15, no. 3, Mar. 2020, Art. no. e0231236.

[6]. G. Grasselli, A. Pesenti, and M. Cecconi, "Critical care utilization for the COVID-19 outbreak in lombardy, italy: Early experience and forecast during an emergency response," JAMA, vol. 323, no. 16, p. 1545, Apr. 2020.

[7]. WHO. Naming the Corona virus Disease (Covid-19) and the Virus That Causes it. Accessed: Apr. 1, 2020. [Online]. Available: https:// www.who.int/emergencies/diseases/novel coronavirus-2019/technic% alguidance/ naming-the-coronavirus-disease-(covid-2019)- and-the-virusthat-cau%ses-it

[8]. C. P. E. R. E. Novel, "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (Covid-19) in China," Zhonghua Liu Xing Bing XueZaZhi= Zhong hua Liuxingb in gxueZazhi, vol. 41, no. 2, p. 145, 2020.

[9]. M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, "Analytics defined," in Information Security Analytics, M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, Eds. Boston, MA, USA: Syngress, 2015, pp. 1–12. [Online]. Available: http://www. sciencedirect.com/science/article/pii/ B9780128002070000010.