MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
V.N. Karazin Kharkiv National University
School of Mathematics and Computer Science
Department of Theoretical and Applied Informatics

Master's Thesis

Prediction of the dynamics COVID19 epidemic process of using the Ridge regression model

Author:
Final year Master's Program student,
group 63
specialty - Computer Sciences and
Information Technologies,
educational program: "Informatics"
Zhang Dongxing
Supervisor: Ievgen Meniailov
Reviewer: Kseniia Bazilevych
Adviser: Stas Kachanov

Kharkiv, 2024

1. INTRODUCTION

Challenges:

The COVID-19 pandemic has posed significant challenges to public health systems, resource allocation, and policymaking worldwide. The unpredictability of the virus, combined with its rapid spread, has made it extremely difficult for governments and healthcare organizations to effectively plan and allocate resources. One of the most complex aspects of managing the pandemic has been the accurate prediction of daily case numbers. Several factors contribute to the difficulty in making these predictions. Data variability is a major issue, as the quality and consistency of reported case numbers can fluctuate across regions and time periods. Additionally, the limited availability of real-time data, due to delays in reporting or inconsistencies in data collection practices, further complicates predictions. Moreover, the rapidly evolving nature of the pandemic itself, with new variants of the virus emerging and changes in human behavior or public health interventions (such as lockdowns and vaccination campaigns), leads to an environment of constant change. These challenges have made it critical to develop advanced forecasting methods that can adapt to the shifting dynamics of the COVID-19 crisis and help inform timely, evidence-based decisions.

Motivation:

Given the challenges in predicting the trajectory of the COVID-19 pandemic, there is an urgent need for accurate and reliable predictive models that can provide actionable insights. By leveraging machine learning (ML) techniques, governments and healthcare organizations can better anticipate future trends, optimize resource distribution, and plan for potential surges in cases. Data-driven predictive models, particularly those that use historical data to forecast future cases, can provide valuable foresight for decision-making. This study is motivated by the desire to contribute to this effort by developing such predictive models for COVID-19. Specifically, it focuses on the application of regression models to forecast short-term trends in the pandemic. The ability to predict daily case numbers with a reasonable degree of accuracy is vital for enhancing the preparedness of healthcare systems, minimizing healthcare burden, and guiding effective policymaking. Furthermore, this study highlights the practical applications of machine learning in public health, demonstrating how advanced computational techniques can assist in managing a global health crisis.

Goals:

The primary objective of this study is to develop and implement a predictive model for COVID-19 case numbers in a specific country. To achieve this, the study will begin by gathering COVID-19 case data for the selected country, ensuring that the dataset is comprehensive, accurate, and up-to-date. The historical data will be used to train a regression model that can learn the patterns and trends in the spread of the virus. The model will then be used to predict daily case numbers for a three-day period: from September 28 to September 30, 2024. By focusing on this short-term forecast, the study aims to evaluate the model's ability to capture the near-term dynamics of the pandemic, which can be critical for immediate public health response.

Additionally, the study will assess the model's accuracy by calculating both absolute and relative errors for the predicted values on September 30, 2024. These error metrics will help evaluate the precision of the model's predictions and provide insight into its overall effectiveness. The comparison of predicted and actual case numbers will also offer guidance for improving the model's performance in future predictions. Ultimately, the goal is to develop a reliable, data-driven tool that can be used to inform public health decision-making and support more effective management of the ongoing pandemic. This study contributes to the broader effort to use machine learning for public health forecasting, with the aim of enhancing pandemic preparedness and response strategies.

---

2. MAIN CONCEPTS

Workplan:
The project was divided into the following phases:

1. **Data Collection**: Gathering COVID-19 case data from the selected country using sources such as the Johns Hopkins ArcGIS Dashboard and WHO Public Emergency platform[1][2].

This study aims to forecast the future spread of COVID-19, focusing on predicting the number of new positive cases, deaths, and recoveries. The dataset utilized in this research was sourced from the GitHub repository maintained by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University [12]. This repository was initially created to support the university's 2019 Novel Coronavirus visual dashboard, with additional backing from the ESRI Living Atlas Team.

The dataset files are organized within a folder named csse_covid_19_time_series on the repository. This folder contains daily time series summary tables detailing confirmed cases, deaths, and recoveries. These data entries are compiled from daily case reports and updated with a frequency of once per day. Samples from the dataset files are presented in Tables 1, 2, and 3, respectively.

TABLE 1. COVID-19 patient death cases time-series worldwide.

Province/S	Country/R	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20
	Afghanista	33.93911	67.70995	0	0	0	0
	Albania	41.1533	20.1683	0	0	0	0
	Algeria	28.0339	1.6596	0	0	0	0
	Andorra	42.5063	1.5218	0	0	0	0
	Angola	-11.2027	17.8739	0	0	0	0
	Antarctica	-71.9499	23.347	0	0	0	0
	Antigua ar	17.0608	-61.7964	0	0	0	0
	Argentina	-38.4161	-63.6167	0	0	0	0
	Armenia	40.0691	45.0382	0	0	0	0
Australian	Australia	-35.4735	149.0124	0	0	0	0
New Soutl	Australia	-33.8688	151.2093	0	0	0	0
Northern T	Australia	-12.4634	130.8456	0	0	0	0

TABLE 2. COVID-19 new confirmed cases time-series worldwide.

Yunnan	China	24.974	101.487	1	2	5
Zhejiang	China	29.1832	120.0934	10	27	43
	Colombia	4.5709	-74.2973	0	0	0
	Comoros	-11.6455	43.3333	0	0	0
	Congo (Bra	-0.228	15.8277	0	0	0
	Congo (Kir	-4.0383	21.7587	0	0	0
	Costa Rica	9.7489	-83.7534	0	0	0
	Cote d'Ivoi	7.54	-5.5471	0	0	0
	Croatia	45.1	15.2	0	0	0
	Cuba	21.52176	-77.7812	0	0	0
	Cyprus	35.1264	33.4299	0	0	0
	Czechia	49.8175	15.473	0	0	0
Faroe Islan	Denmark	61.8926	-6.9118	0	0	0
Greenland	Denmark	71.7069	-42.6043	0	0	0

TABLE 3. COVID-19 recovery cases time-series worldwide.

Anhui	China	31.8257	117.2264	0	0	0
Beijing	China	40.1824	116.4142	0	0	1
Chongqing	China	30.0572	107.874	0	0	0
Fujian	China	26.0789	117.9874	0	0	0
Gansu	China	35.7518	104.2861	0	0	0
Guangdon	China	23.3417	113.4244	0	2	2
Guangxi	China	23.8298	108.7881	0	0	0
Guizhou	China	26.8154	106.8748	0	0	0
Hainan	China	19.1959	109.7453	0	0	0
Hebei	China	37.8957	114.9042	0	0	0
Heilongjiar	China	47.862	127.7615	0	0	0
Henan	China	33.882	113.614	0	0	0
Hong Kong	China	22.3	114.2	0	0	0
Hubei	China	30.9756	112.2707	28	28	31

2. **Model Development**:

A supervised learning model is built to make a prediction when it is provided with an unknown input instance. Thus in this learning technique, the learning algorithm takes a dataset with input instances along with their corresponding regressor to train the regression model. The trained model then generates a prediction for the given unforeseen input data or test dataset. This learning method may use regression techniques and classication algorithms for predictive models' development.Four regression models have been used in this study of COVID-19 future forecasting: Linear Regression,LASSO Regression and Exponential Smoothing.

Linear Regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables[3].The main goal of linear regression is to fit a line (or hyperplane in the case of multiple variables) that best predicts the dependent variable based on the independent variables.

This is the most basic form of linear regression, where the relationship between the dependent variable YY and the independent variable XX is modeled as a straight line.The equation for simple linear regression is:

$$Y = \beta_0 + \beta_1 X + \varepsilon Y$$

Where: Y is the dependent variable (target),X is the independent variable (predictor),β0 is the intercept,β1 is the slope of the line (the change in Y for a one-unit change in X), ε is the error term (accounting for the difference between the predicted and actual values).

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be applied to both classification and regression tasks[4]. In the case of SVM regression, it operates as a non-parametric method that relies on specific mathematical functions. These functions, known as kernels, transform the input data into a required format. SVM regression addresses regression problems by employing a linear function. However, for non-linear regression scenarios, it first maps the input vector (x) to a higher-dimensional space called the feature space (z) using non-linear mapping techniques. Once this mapping is complete, linear regression is performed in the transformed feature space.

In the context of machine learning and a multivariate training dataset (xn) containing N observations with yn representing the corresponding set of observed outputs, the linear function used for regression can be expressed as:

$$f(x) = x'\beta + b$$

The objective is to make it as flat as possible thus to find the value of $f(x)$ as minimal norm values. So the problem fits in minimization function as:

$$J(\beta) = \frac{1}{2}\beta'\beta$$

with a special condition of the values of all residuals not more than ε, as in the following equation:

$$\forall n : | y_n - (x_n'\beta + b) | \leq \varepsilon$$

LASSO is a linear regression technique that applies shrinkage[5], which involves reducing the influence of extreme data values towards the central values, making the model more stable and less prone to error. It is particularly effective in scenarios with multicollinearity, as it uses L1 regularization, penalizing the magnitude of the regression coefficients. This penalty encourages the model to use fewer features, automatically eliminating those that do not significantly contribute to the regression result by setting their coefficients to zero. Unlike ordinary regression, which includes all features and assigns each a coefficient, LASSO adds features one by one and only keeps those that improve the model enough to outweigh the penalty. This results in a sparse model with fewer coefficients, making feature selection an integral part of the regularization process. The objective of LASSO is to minimize the following:

$$\sum_{i=1}^{n}(y_j - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} | \beta_j |$$

In the series of exponential smoothing (ES) methods, forecasting is performed using data from previous time periods. The influence of past observations diminishes exponentially as they become older, meaning the weights assigned to earlier values decrease geometrically.

ES is a simple yet powerful forecasting technique, particularly effective for univariate time series data [6].The forecast for the current period (Ft) in exponential smoothing can be expressed as:

$$F_t = \alpha A_{t-1} + (1-\alpha) F_{t-1}$$

Here, α represents the smoothing parameter, where 0≤α≤10 is the actual value from the previous time period, and Ft−1 is the forecasted value from the prior period.

3. Evaluating Regression Models

To assess how well the linear regression model fits the data, several metrics are used:

R-squared : This is a measure of how well the independent variables explain the variability in the dependent variable[7]. It ranges from 0 to 1, where 1 indicates perfect prediction.

$$R^2 = 1 - \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\sum_i (Y_i - \bar{Y})^2}$$

Where Yi is the actual value, Yi^ is the predicted value, and Y¯is the mean of the actual values.

Mean Squared Error (MSE): This measures the average of the squared differences between the actual and predicted values[8]. A lower MSE indicates better model fit.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Root Mean Squared Error (RMSE): The square root of MSE, which is useful because it has the same units as the dependent variable, making it more interpretable[9].

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}$$

4. Methodology

This study focuses on predictions related to the novel coronavirus, also known as COVID-19. COVID-19 has emerged as a significant threat to human life, resulting in tens of thousands of deaths worldwide, with the mortality rate continuing to rise daily. To aid in managing this global pandemic, the study aims to forecast key metrics, including the death rate, the daily number of confirmed cases, and recoveries, over future time frames of 3, 7, 10, 14, 21, and 30 days.

The forecasting is conducted using four machine learning approaches specifically suited to this context. The dataset utilized for the analysis comprises daily time series summary tables that record the number of confirmed cases, deaths, and recoveries since the start of the pandemic. As a preparatory step, the dataset underwent preprocessing to calculate global statistics for daily deaths, confirmed cases, and recoveries, ensuring the data is ready for forecasting.

Following the initial data preprocessing, the dataset was split into two subsets: a training set for model training and a testing set for evaluation. The study employed learning models, including Linear Regression (LR), Support Vector Machine (SVM), and LASSO. These models were trained using the patterns of daily confirmed cases, recoveries, and deaths.The performance of the trained models was assessed using key evaluation metrics such as the R^2 score and Mean Absolute Error (MAE), with the results documented and analyzed in the study.

5. Prediction and Evaluation:

The dataset utilized includes daily records of newly infected cases, recoveries, and deaths worldwide in this study. With the confirmed cases and death rates rising daily, this poses a significant global concern. The exact number of individuals who may be affected by the pandemic in various countries remains uncertain.

Four machine learning models—Linear Regression (LR), LASSO, Support Vector Machine (SVM), and Exponential Smoothing (ES), were applied to make these predictions.

A. Forecasting Death Rates

The study conducted forecasts on the death rate, and the results indicated that the ES model outperformed the others. LR and LASSO performed comparably, achieving nearly identical R^2 scores, while SVM exhibited the weakest performance in this scenario. These findings are summarized in Table 4.

Figures 1, 2, 3 and 4 illustrate the performance of the LR, LASSO, SVM, and ES models, respectively, through graphs. The graphs across these figures consistently predict a rise in the death rate over the coming days, highlighting an alarming trend.

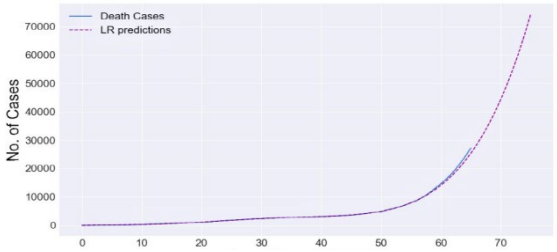

FIGURE 1. Death prediction by LR for the upcoming 10 days.

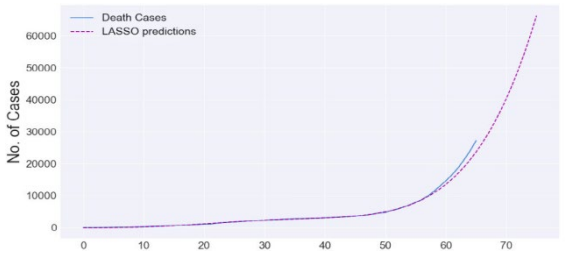

FIGURE 2. Death prediction by LASSO for the upcoming 10 days.

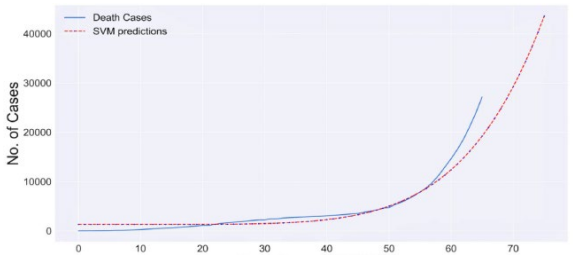

FIGURE 3. Death prediction by SVM for the upcoming 10 days.

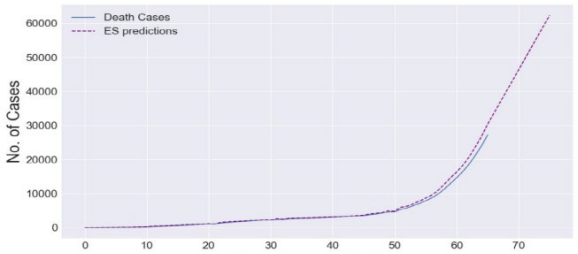

FIGURE 4. Death prediction by ES for the upcoming 10 days.

TABLE 4. Models performance on future forecasting for death rate.

Model	R2 Score	MSE	MAE	RMSE
LR	0.97	841250.29	724.58	920.54
LASSO	0.82	3245885.52	1460.14	1795.53
SVM	0.57	16047050.41	3527.58	3984.76
ES	0.95	654025.74	416.57	820.74

B. Forecasting New Confirmed Cases

The number of newly confirmed COVID-19 cases continues to rise daily. Table 5 presents the forecasting results of the models applied in this study. Among them, ES and LASSO demonstrate the best performance, while LR also performs well. However, SVM shows significantly poor results across all evaluation metrics. Figures 5, 6, 7, and 8 illustrate the predictions made by these learning models.

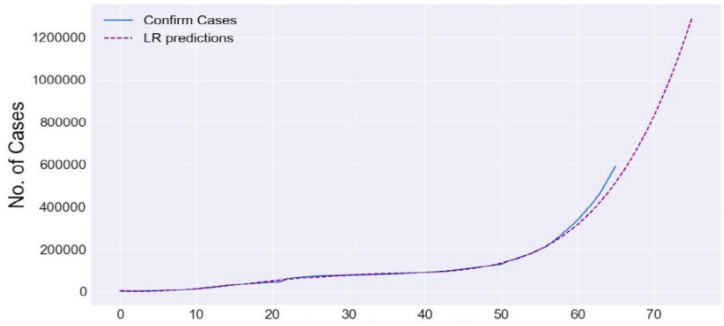

FIGURE 5. New infected confirm cases prediction by LR for the upcoming 10 days.

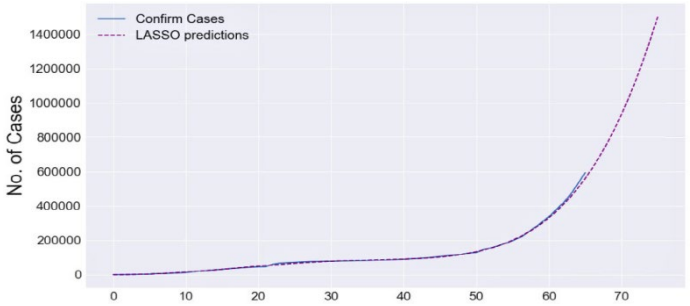

FIGURE 6. New infected confirm cases prediction by LASSO for the upcoming 10 days.

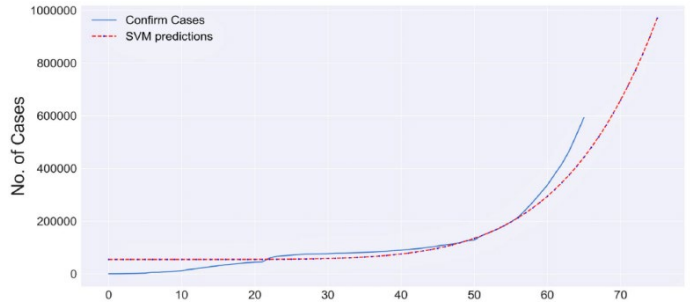

FIGURE 7. New infected confirm cases prediction by SVM for the upcoming 10 days.

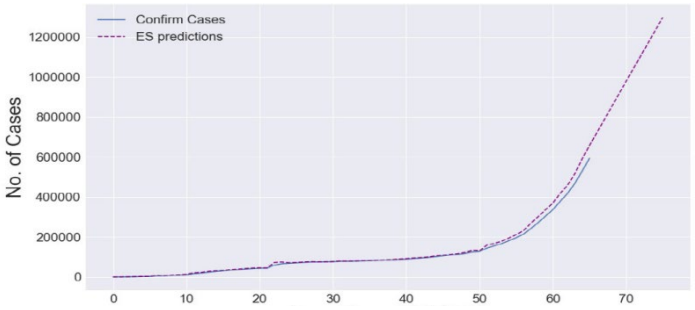

FIGURE 8. New infected confirm cases prediction by ES for the upcoming 10 days.

TABLE 5. Models performance on future forecasting for new infected confirm cases.

Model	R2 Score	MSE	MAE	RMSE
LR	0.83	1473142988.47	30311.51	38404.55
LASSO	0.95	234965710.64	11701.84	15715.63
SVM	0.62	5780064738.51	60543.34	78434.96
ES	0.94	283445604.2	8769.43	17868.57

C. Forecasting Recovery Rates

For forecasting the recovery rate, the ES model once again outperforms all other models. The remaining models show comparatively poor performance, with their ranking from best to worst being ES, followed by LR, LASSO, and SVM. This ranking reflects how effectively each model handles the available time-series data. The predicted trends for the upcoming days are depicted in Figures 9, 10, 11, and 12. The performance results of the learning models are summarized in Table 6 below.

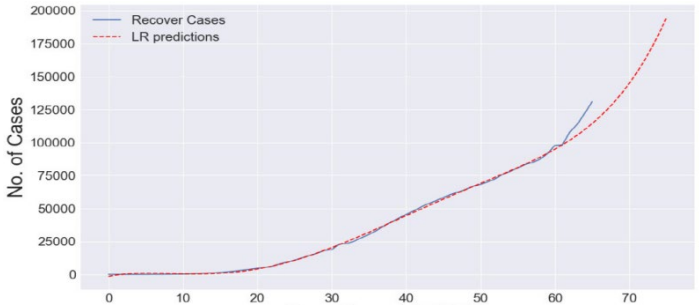

FIGURE 9. Recovery rate prediction by LR for the upcoming 10 days.

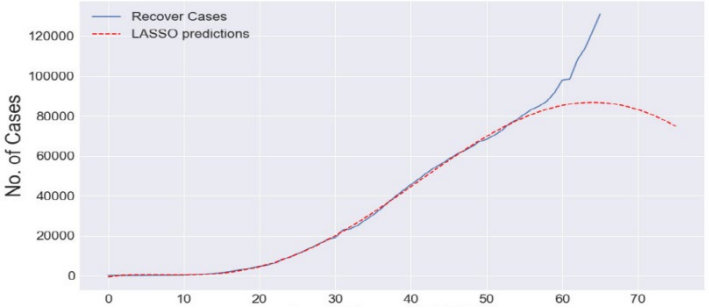

FIGURE 10. Recovery rate prediction by LASSO for the upcoming 10 days.

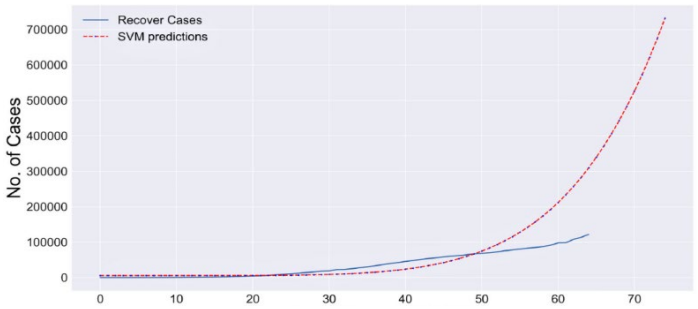

FIGURE 11. Recovery rate prediction by SVM for the upcoming 10 days.

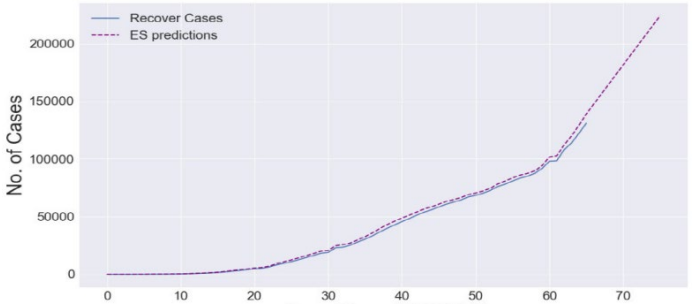

FIGURE 12. Recovery rate prediction by ES for the upcoming 10 days.

TABLE 6. Models performance on future forecasting for recovery rate.

Model	R2 Score	MSE	MAE	RMSE
LR	0.36	484126209.59	17264.38	26894.51
LASSO	0.27	1432498785.12	30460.91	36852.64
SVM	0.25	14105194087.41	106529.68	135781.67
ES	0.96	580625.64	1819.73	2547.97

3. CONCLUSIONS

The uncertainty surrounding the COVID-19 pandemic has the potential to trigger a severe global crisis. Researchers and government agencies worldwide have expressed concerns that the pandemic could impact a significant portion of the global population. This study introduces a machine-learning-based prediction system designed to forecast the risk of COVID-19 outbreaks on a global scale. The system utilizes a dataset containing day-wise historical data and employs machine learning algorithms to predict future trends.

The study's findings demonstrate that the Exponential Smoothing (ES) model outperforms others in this forecasting context, given the dataset's nature and size. While Linear Regression (LR) and LASSO also show satisfactory performance in predicting death rates and confirmed

cases, their results indicate an increase in death rates and a slowdown in recovery rates in the coming days. On the other hand, the Support Vector Machine (SVM) model performs poorly across all scenarios, likely due to fluctuations in the dataset values, which make it challenging to establish an accurate hyperplane for predictions.
Overall, the results confirm that the model predictions align with the current situation, offering valuable insights into potential future developments. These forecasts could aid authorities in implementing timely measures and making informed decisions to mitigate the COVID-19 crisis. The study will continue to evolve, with future efforts focusing on refining the prediction methodology using updated datasets and employing more accurate and suitable machine learning techniques. Real-time, live forecasting will be a key area of emphasis in future work.

---

4. REFERENCES

[1]ArcGIS COVID-19 Dashboard: https://gisanddata.maps.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6

[2]WHO Public Emergency Database: https://extranet.who.int/publicemergency/

[3] DeVore C, Kaukis N. Linear regression analysis[J]. STAT, 2003, 3013(11/16): 15.

[4] Suthaharan S, Suthaharan S. Support vector machine[J]. Machine learning models and algorithms for big data classification: thinking with examples for effective learning, 2016: 207-235.

[5]Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society Series B: Statistical Methodology, 1996, 58(1): 267-288.

[6] Gardner Jr E S. Exponential smoothing: The state of the art[J]. Journal of forecasting, 1985, 4(1): 1-28.

[7] Miles J. R-squared, adjusted R-squared[J]. Encyclopedia of statistics in behavioral science, 2005.

[8] Error M S. Mean squared error[J]. MA: Springer US, 2010: 653-653.

[9]Willmott C J, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance[J]. Climate research, 2005, 30(1): 79-82.
