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

Qualification work

Master

on the topic:

«Prediction of the dynamics COVID19 epidemic process of using the Stochastic Gradient Descent model »

Done by: 2-th year student, group MCS-64
specialty: Computer science and Information Technologies
education program: «Computer Sciences»
Liu Xianglei

Supervisor: Victoriya Kuznietcova
Reviewer: Kseniia Bazilevych
Adviser: Ruslan Borodai

Kharkiv, 2024

Table of Contents

1 INTRODUCTION

1.1 Challenges

Despite the remarkable advancements in human technology over the past few decades, the shadow of epidemics has continued to loom over human civilization and society. Especially with the acceleration of globalization and the increase in population mobility, the speed and scope of infectious disease transmission have expanded significantly, posing unprecedented challenges to public health systems.

A coronavirus is a type of virus that can cause a range of illnesses. Over time, new variants of the coronavirus have emerged, including SARS-CoV-2, which causes COVID-19. Because there was no pre-existing immunity to SARS-CoV-2 in humans, it had the potential for widespread transmission. Symptoms of COVID-19 typically appear within 1 to 14 days after exposure and are often similar to those of the flu. Most of the sufferers will experience a fever above 37.8 ℃ and a dry cough, but it is usually accompanied by dizziness, diarrhea, and so on. About 20% of sufferers, on the seventh day, will experience shortness of breath, pneumonia, and lung inflammation [1]. The COVID-19 outbreak that broke out in December 2019 quickly spread globally, according to data provided by the World Health Organization, as of November 24, 2024, there were a total of 776841264 confirmed cases, resulting in 7075468 deaths [2].

WHO has classified the level of global risk from low risk to high risk [3]. COVID-19 has had an unprecedented impact on communities around the world [4]. So that the COVID-19 outbreak is a concern for all stakeholders around the world to control and anticipate its spread [5].

1.2 Motivation

Accurate prediction of epidemic dynamics is vital for effective prevention, control strategies, and optimal resource allocation. With the advancement of machine learning technology, forecasting epidemic trends and progression has become increasingly feasible and reliable.

The COVID-19 pandemic has brought unprecedented challenges to global public health systems and socio-economic stability. This pandemic not only resulted in hundreds of millions of infections and millions of deaths but also severely impacted the stability and development of the global economy. In order to effectively respond to this public health crisis, early warning of the pandemic's development trends has become particularly important. Accurate epidemic predictions can help governments and health departments take timely and effective measures, reduce the spread of the virus, alleviate the burden on healthcare systems, and maximize the protection of public health.

Although existing epidemic prediction models have achieved some success, there is still significant room for improvement in their accuracy and reliability. Many models face challenges such as low computational efficiency and slow convergence when handling large-scale datasets, which limits their effectiveness in real-world applications. When selecting a prediction model, we considered various factors, including computational efficiency, convergence speed, and the ability to handle large-scale datasets. Stochastic Gradient Descent (SGD) stands out as an efficient optimization algorithm, demonstrating significant advantages in processing large-scale datasets. By updating model parameters using only one sample or a small batch of samples per iteration, SGD significantly reduces the required computational resources and memory usage, making it particularly suitable for real-time applications and online learning scenarios. Additionally, SGD can quickly find local optima on large datasets and achieve faster

convergence by appropriately adjusting the learning rate. Its simple linear solution is not only easy to understand and implement but also enhances the model's interpretability, facilitating the analysis of feature importance. These characteristics make SGD an ideal choice for improving prediction accuracy while maintaining high computational efficiency. Therefore, this study aims to enhance the precision and efficiency of epidemic prediction by introducing a Stochastic Gradient Descent (SGD) regression model. As an efficient optimization algorithm, SGD can converge quickly on large datasets and provide more accurate prediction results.

Through this study, we hope to provide scientific evidence to assist governments and health departments in formulating more precise prevention and control measures. Specifically, we will use the SGD regression model to model the epidemic data from Bangladesh and predict the daily cumulative death toll. We will also ensure the model's generalization ability and predictive performance by constructing lag features and dividing the dataset into training and testing sets.

In conclusion, this study not only aims to enhance the accuracy and reliability of epidemic predictions but also seeks to provide strong support for future global pandemic prevention efforts using scientific methods and advanced technologies.

1.3 Goals

Considering the efficiency of model development and the potential for future implementation and application, this study has chosen Python as a comprehensive solution for implementing the Stochastic Gradient Descent (SGD) model and analyzing prediction accuracy. The entire mathematical modeling process is realized in a simple and easy-to-learn programming language

environment, making it accessible to a wide range of users, including public health researchers and practitioners.

The main objective of this study is to train the SGD model with different datasets to predict the development of an epidemic over various periods. By comparing the relative and absolute errors under different conditions, we aim to analyze the model's prediction accuracy for epidemic trends. This approach provides public health researchers with a straightforward and accurate modeling solution, facilitating quick mastery of mathematical modeling techniques. It also reduces the barriers to entry for mathematical modeling, enabling researchers to study the impact of epidemic-related parameters and control measures on epidemic trends. The insights gained from this study can have significant public health implications, offering valuable information for epidemic prevention and control.

To achieve the research goals, the following tasks have been outlined:

1. Data collection and data cleaning: Gather relevant data on the epidemic, including daily new confirmed cases, cumulative confirmed cases, and other related parameters. Clean the data to remove any inconsistencies, missing values, or outliers that could affect the model's performance.

2. Data review and key feature identification: Conduct a thorough review of the collected data to understand its distribution and characteristics. Identify key features that are most relevant to the prediction of epidemic trends, such as the number of new cases, cumulative cases, and other demographic or environmental factors.

3. Use the data from January 4, 2020, to September 30, 2024, as the first dataset, providing a long-term perspective on the epidemic. Use the data from January 1, 2024, to September 30, 2024, as the second dataset, focusing on more recent trends and patterns.

4. Determine the prediction period: Define the time periods for which predictions will be made. This could include short-term, medium-term, and long-term predictions to evaluate the model's performance across different horizons.

5. Create the Stochastic Gradient Descent (SGD) regression model: Implement the SGD regression model using Python. This includes defining the model architecture, setting hyperparameters, and training the model on the prepared datasets. Use the cumulative number of confirmed cases from the previous three days as input features to capture the latest trends and improve prediction accuracy.

6. Analyze the prediction results: Evaluate the prediction results based on different datasets and prediction periods. Compare the relative and absolute errors of the predictions to assess the model's accuracy and reliability.

This study makes two significant contributions:

First, it investigates how the choice of dataset (e.g., long-term vs. short-term data) affects the performance of the SGD regression prediction model, providing insights into the optimal dataset selection for different types of epidemic trend predictions.

Second, it evaluates the prediction performance of the SGD regression model over various prediction periods (short-term, medium-term, and long-term), offering recommendations on the most suitable prediction horizon for achieving accurate and reliable forecasts.

By addressing these tasks and contributions, this study aims to provide a robust and practical solution for predicting epidemic trends, ultimately supporting public health efforts in preventing and controlling epidemics.

2 MAIN CONCEPTS

2.1 Current Research Analysis

Kermack and McKendrick's research laid the foundation for the modeling method used to determine the incidence rate of infectious diseases [6]. Building on Ronald Ross's infection rate modeling method [7], they proposed the classical modeling method for the incidence rate of infectious diseases.

This approach employs a system of differential equations to model the transmission dynamics of infectious diseases within a specific population. While it is widely utilized, it has certain limitations, including high computational demands and limited scalability.

Machine learning algorithms have demonstrated exceptional effectiveness in predicting COVID-19 transmission. Their strength lies in integrating diverse approaches, optimizing model parameters, and adapting flexibly to different scenarios and data formats. These capabilities have positioned them as a key tool in epidemiological modeling.

Machine learning algorithms have been effectively utilized for both time-series data (e.g., LSTM) and regression tasks (e.g., multilayer perceptron, MLP). Research indicates that LSTM-based time-series analyses are particularly popular for forecasting COVID-19 transmission trends, followed by MLP and other regression techniques. Studies on regression and classification using ML algorithms have demonstrated that many models achieve highly accurate predictions. For example, MLP-based methods have shown excellent accuracy and regression performance in regions such as the United States and India.

To further enhance predictive performance, numerous studies have combined different ML algorithms. For instance, hybrid approaches integrating deep neural networks (DNN), LSTM, and CNN have proven effective in improving accuracy by merging predictive outputs. Additionally, evolutionary

algorithms have been employed to optimize the hyperparameters of ML models, significantly boosting their performance.

In the study [8], support vector regression (SVR) and the Bayesian regression model were employed to analyze COVID-19 spread rates, recovery rates, and mortality. Using data from March 6 to March 26-27, 2020, the results revealed that the Bayesian Ridge method produced simulation outcomes that aligned more closely with actual data, outperforming the support vector machine (SVM) approach.

In reference [9], multi-view machine learning techniques achieved an accuracy of 95.5%, while deep convolutional and recurrent neural networks reported an accuracy of 86.7% [10]. Reference [11] demonstrated a 99.7% accuracy using an ensemble of deep convolutional neural networks, and polynomial regression model achieved about 93.0% accuracy in reference [12]. Additionally, logistic regression analysis yielded an AUC of 96.2% in reference [13]. In reference [14], eight different machine learning classifiers were employed. These findings highlight that machine learning-based regression approaches are highly effective for predicting the dynamic progression of epidemics.

However, machine learning (ML) algorithms have certain limitations. They rely on large datasets for effective training, but epidemiological data for COVID-19 may be limited due to underreporting or inconsistent data quality across different regions. Additionally, some ML methods, particularly deep learning models like LSTM, involve intricate parameter tuning processes. This complexity can result in overfitting or biases when training data is insufficient [15].

To overcome these challenges, this study introduces the following predictive methods:

1. Utilize publicly available data provided by the WHO, focusing on Bangladesh as the study subject to analyze the daily cumulative case numbers of COVID-19.

2. Apply the Stochastic Gradient Descent (SGD) optimization algorithm to optimize the parameters of the linear regression prediction model.

3. Perform a comparative analysis of prediction accuracy for different forecasting time intervals within the same dataset using relative error and absolute error metrics.

4. Conduct a comparative analysis of prediction accuracy for the same forecasting time interval across different datasets using relative error and absolute error metrics.

2.2 Data Analysis and Method

This section primarily introduces the selection of data sets, the choice of models, and the underlying mathematical principles.

During data collection and storage, various sources of uncertainty can lead to errors, such as measurement inaccuracies, recording mistakes, and data loss. As a result, selecting a suitable dataset is essential to ensure accurate predictions.

Regression analysis remains a widely used method for forecasting time series across different domains and is easily implementable with modern computational tools. While non-adaptive models are capable of predicting incidence dynamics over any time period, they often fail to account for local fluctuations in epidemic indicators, limiting their effectiveness for short-term forecasts. In contrast, adaptive models are specifically designed to provide predictions for several weeks ahead, making them more suitable for capturing long-term trends.

2.3 Data Description

The data used in this study is sourced from official WHO records. The dataset is updated daily by country, providing information on daily confirmed cases, cumulative confirmed cases, daily deaths, and cumulative deaths. However, since August 2023, the WHO has ceased requiring daily reporting of epidemic data, and most countries now update their data on a weekly basis. For better research outcomes, this study focuses on infection data from Bangladesh, as it remains one of the few countries that continues to update relevant data on a daily basis. From January 4, 2020, to September 30, 2024, Bangladesh has reported 2,051,430 confirmed cases of COVID-19 and 29,499 deaths .

2.3.1 Stochastic Gradient Descent (SGD) Regression Model

Stochastic Gradient Descent (SGD) is an optimization algorithm that is particularly well-suited for large-scale and online learning scenarios. Unlike batch gradient descent, which uses the entire dataset to compute the gradient at each step, SGD updates the model parameters using a single data point at a time. This makes the training process faster and more computationally efficient, especially when dealing with large datasets. The key advantage of using the SGD model is its ability to handle large datasets efficiently and its suitability for online learning, where the model can be updated incrementally as new data becomes available. This makes it particularly useful for real-time prediction and monitoring of the epidemic situation.

The SGD method aims to minimize the loss function by iteratively adjusting the model parameters in the direction of the negative gradient of the loss function. The update rule for the parameters can be expressed as:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t, x_i, y_i) \quad (1)$$

Where

θ_t:the model parameters at iteration t

η: the learning rate, controlling the step size for updates.

$\nabla_\theta L(\theta_t, x_i, y_i)$: the gradient of the loss function L with respect to the parameters θ_t calculated using a single data point (x_i, y_i) or a small batch.

In the context of this study, we applied the SGD model to predict the incidence dynamics of COVID-19.

The model was configured with the following settings:

Loss Function: Mean Squared Error (MSE) was chosen as the loss function to measure the difference between the predicted and actual values.

Regularization: L2 regularization was applied to prevent overfitting.

Learning Rate: Set to adaptive, to ensure effective convergence of the model during training.

Number of Iterations: The maximum number of iterations was set to 10000 to ensure the model is sufficiently trained.

In this study, we use the SGD regression model, which at its core employs the Stochastic Gradient Descent (SGD) algorithm to minimize the mean squared error (MSE) loss function of a linear regression. By continuously updating the model's parameters, the model is able to gradually improve its prediction accuracy.

Specifically, the working principle of the SGD regression is as follows:

The linear regression model assumes a linear relationship between the target value and the input features, which can be expressed as:

$$y_i = \theta^T x_i + \varepsilon_i \tag{2}$$

θ: a parameter of the model (which needs to be learned through training).

ε_i: an error term.

loss function can be expressed as:

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m} (y_i - \theta^T x_i)^2 \tag{3}$$

m: number of training samples.

y_i: actual values.

$\theta^T x_i$:The predicted value of the model.

The partial derivative of the gradient of the loss function can be expressed as:

$$\nabla_\theta L(\theta_t, x_i, y_i) = -2(y_i - \theta^T x_i)x_i \tag{4}$$

Therefore, the update rule becomes:

$$\theta_{t+1} = \theta_t + 2\eta(y_i - \theta^T x_i)x_i \tag{5}$$

Prediction Accuracy Evaluation

In this study, the prediction accuracy of the SGD regression model will be evaluated using absolute error and relative error, with the formulas as follows:

Absolute Error (AE):

$$\mathrm{AE} = \left|\mathrm{y}_{\mathrm{true}} - \mathrm{y}_{\mathrm{pred}}\right| \tag{6}$$

Relative Error (RE):

$$\mathrm{RE} = \left|\frac{\mathrm{y}_{\mathrm{true}} - \mathrm{y}_{\mathrm{pred}}}{\mathrm{y}_{\mathrm{true}}}\right| \times 100\% \tag{7}$$

Absolute error directly calculates the difference between the predicted and actual values, making it suitable for small ranges or high-precision scenarios, as it provides an intuitive reflection of the actual deviation. Relative error, on the other hand, normalizes the absolute error by the true value, making it suitable for comparing errors across different scales or magnitudes, as it emphasizes the relative size of the error and is useful for assessing the relative accuracy of predictions.

2.4 Description of the software solution

This chapter primarily focuses on illustrating the program structure and presenting the model's running results, along with a comprehensive evaluation of those results. To achieve this, we utilize Unified Modeling Language (UML) diagrams, which are a standard tool for specifying, visualizing, constructing, and documenting the various components and artifacts of software systems.

Unified Modeling Language (UML) is a widely recognized and standardized language that provides a visual way to represent the development activities of software. UML diagrams serve as a powerful communication tool among developers, stakeholders, and other team members, enabling them to understand the system's architecture, design, and behavior more clearly. These diagrams can be used throughout the software development lifecycle, from requirements gathering and analysis to design, implementation, and maintenance.

By using these UML diagrams, we aim to provide a clear and structured view of the program, making it easier to understand the system’s architecture, design, and behavior. Additionally, we will present the running results of the model, including any relevant metrics and performance indicators. We will also conduct a thorough evaluation of these results, discussing their implications and providing insights into the effectiveness and efficiency of the implemented solution.

This approach not only enhances the clarity and comprehensibility of the project but also ensures that all stakeholders have a shared understanding of the system, facilitating better collaboration and decision-making.

2.4.1 UML diagrams

The software activity diagram for this study is shown in Figure 1. The diagram provides a detailed illustration of each step in the process, including data acquisition, data cleaning, feature selection, lag feature extraction, dataset partitioning, SGD model construction, model training, and accuracy analysis. Subsequently, in Section 3.2, we will provide a detailed description of each of these steps.

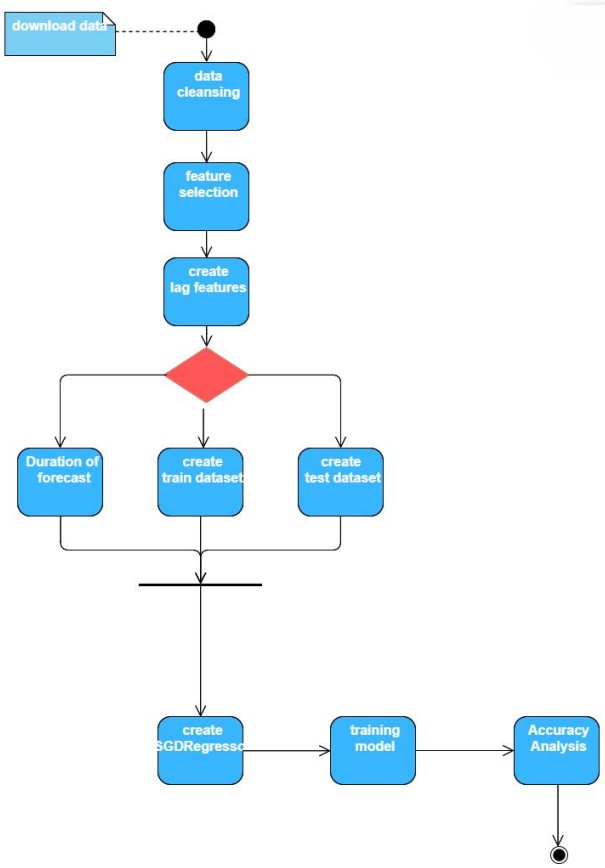

Figure 1. UML activity diagram

2.4.2 Experimental Results

To establish an effective predictive model, we first conducted a thorough observation and analysis of the data. Figure 2 shows the distribution of daily new confirmed cases over time, while Figure 3 illustrates the distribution of

cumulative confirmed cases over time. After analysis, it was found that both the daily new confirmed cases and the cumulative confirmed cases exhibit a nonlinear relationship with time. Therefore, based on the distribution characteristics of the data, we decided to use a specific method to predict the daily cumulative number of deaths, which involves using the cumulative number of confirmed cases from the previous three days as input features. The reasons for choosing this method are as follows: by utilizing the data from the most recent few days, we can capture the latest trends of the epidemic, thereby enhancing the accuracy of the predictions.

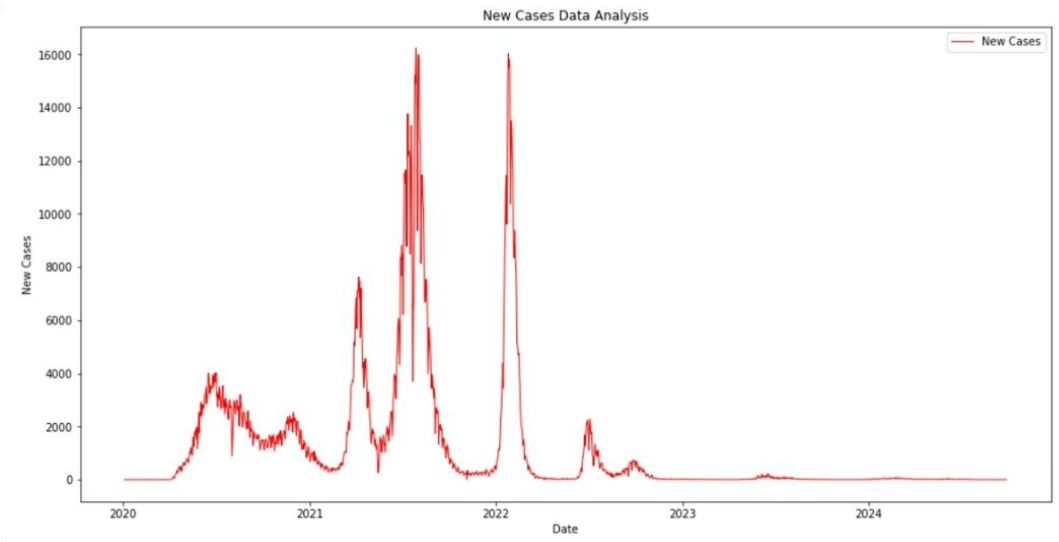

Figure 2. time distribution of new cases

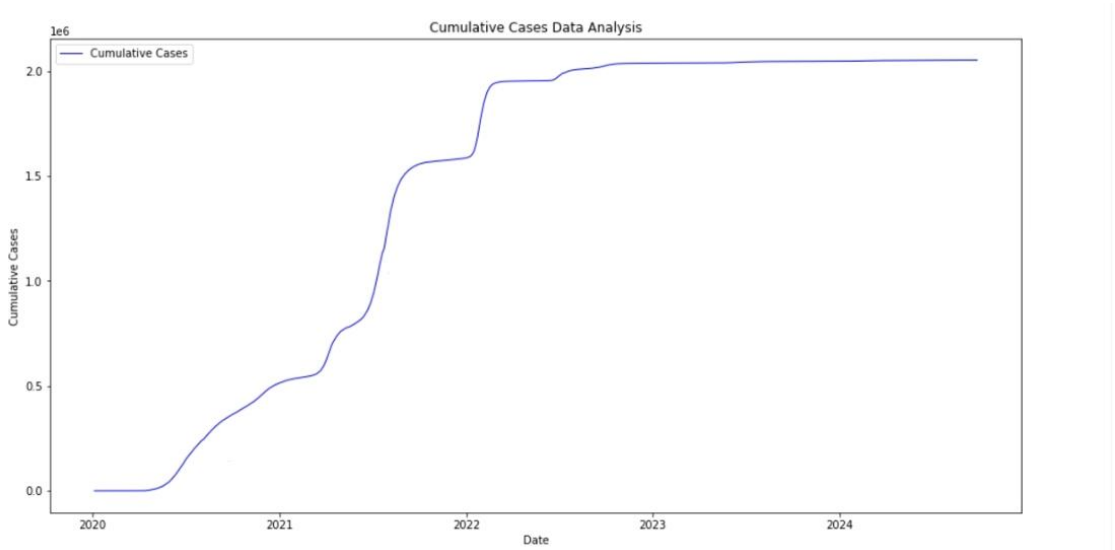

Figure 3. time distribution of cumulative cases

After determining the above characteristics, our next work is divided into several steps.

After determining the features, our subsequent work is divided into several steps. First, we clean the data by filling in missing values with the most recent valid data. Then, we split the dataset into two subsets: the first dataset includes epidemic data from Bangladesh from January 4, 2020, to September 30, 2024, and the second dataset includes epidemic data from Bangladesh from January 1, 2024, to September 30, 2024.

Next, we construct lag features by shifting the cumulative number of confirmed cases in the dataset back by three days to form new lag features. This helps capture temporal dependencies in the data.

Following this, we divide the dataset into training and testing sets based on the prediction horizons (3 days, 5 days, 7 days, 10 days, 14 days, 21 days, and 30 days). This ensures that the model can be effectively validated on unseen data.

Subsequently, we build the SGD regression prediction model. Finally, we evaluate the prediction results using the absolute error and relative error metrics mentioned earlier. Figure 4 shows the time distribution of new confirmed cases in Dataset 2, and Figure 5 illustrates the time distribution of cumulative

confirmed cases in Dataset 2. The time distribution curves of new confirmed cases and cumulative confirmed cases for Dataset 1 have already been presented in Figures 2 and 3.

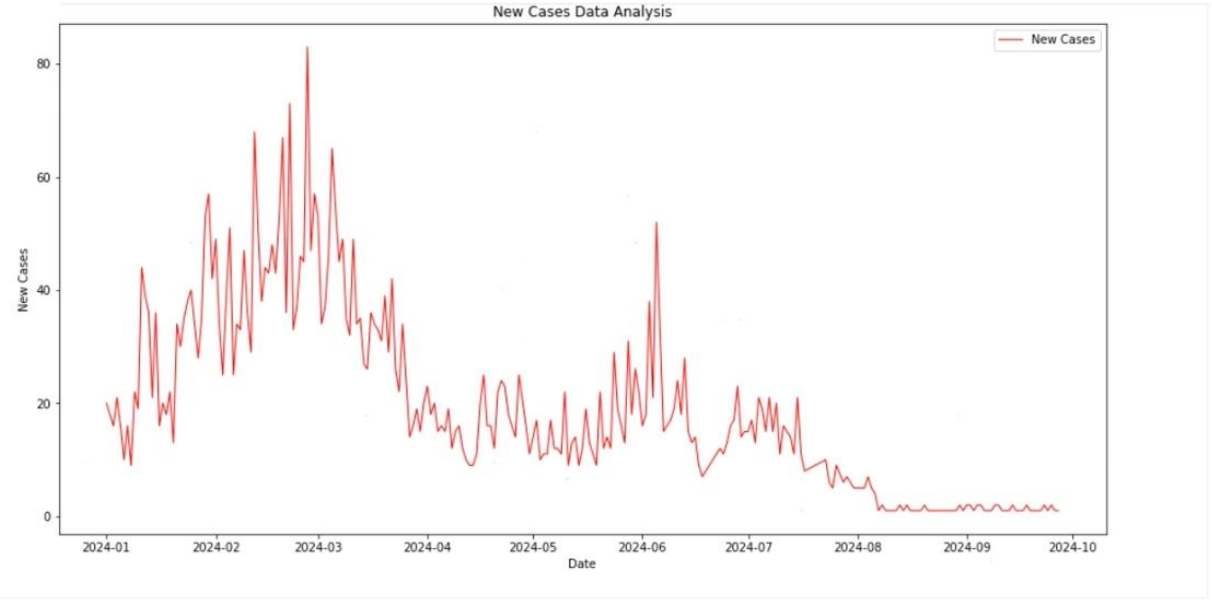

Figure 4. time distribution of new cases

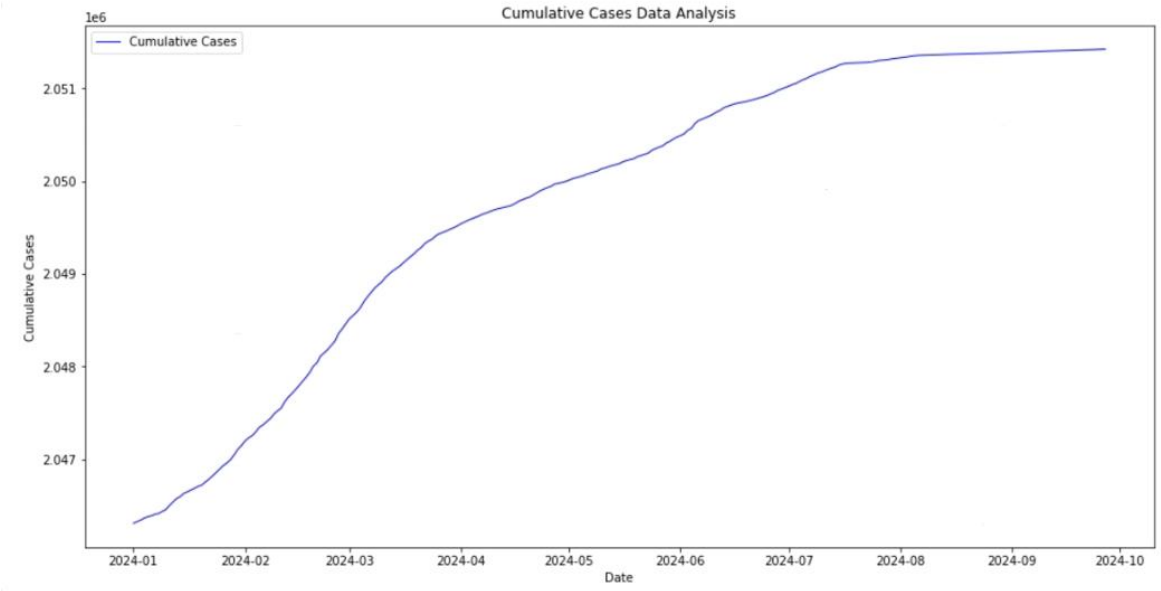

Figure 5. time distribution of cumulative cases

The model prediction results are shown in the figures below.

Figures 6 through 12 illustrate the predicted cumulative number of confirmed cases using the SGD regression model with the first dataset, for

prediction horizons of 3 days, 5 days, 7 days, 10 days, 14 days, 21 days, and 30 days, respectively.

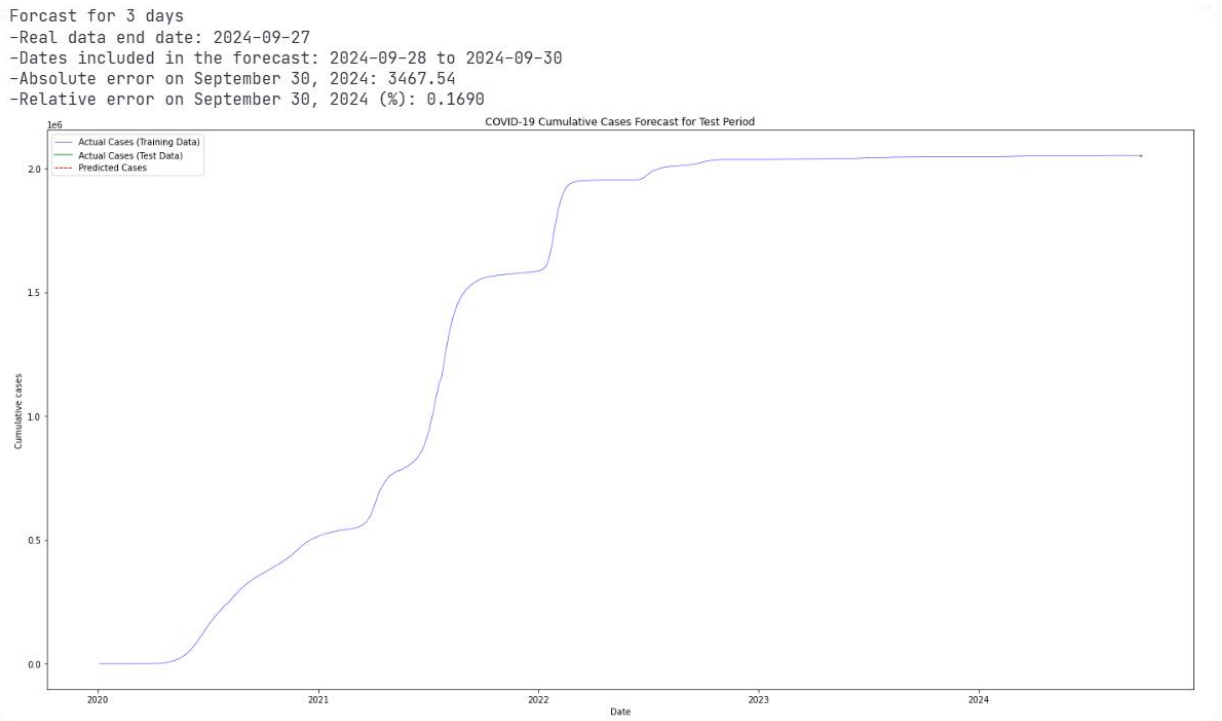

Figure 6. Forecast 3 days Cumulative Cases by dataset 1.

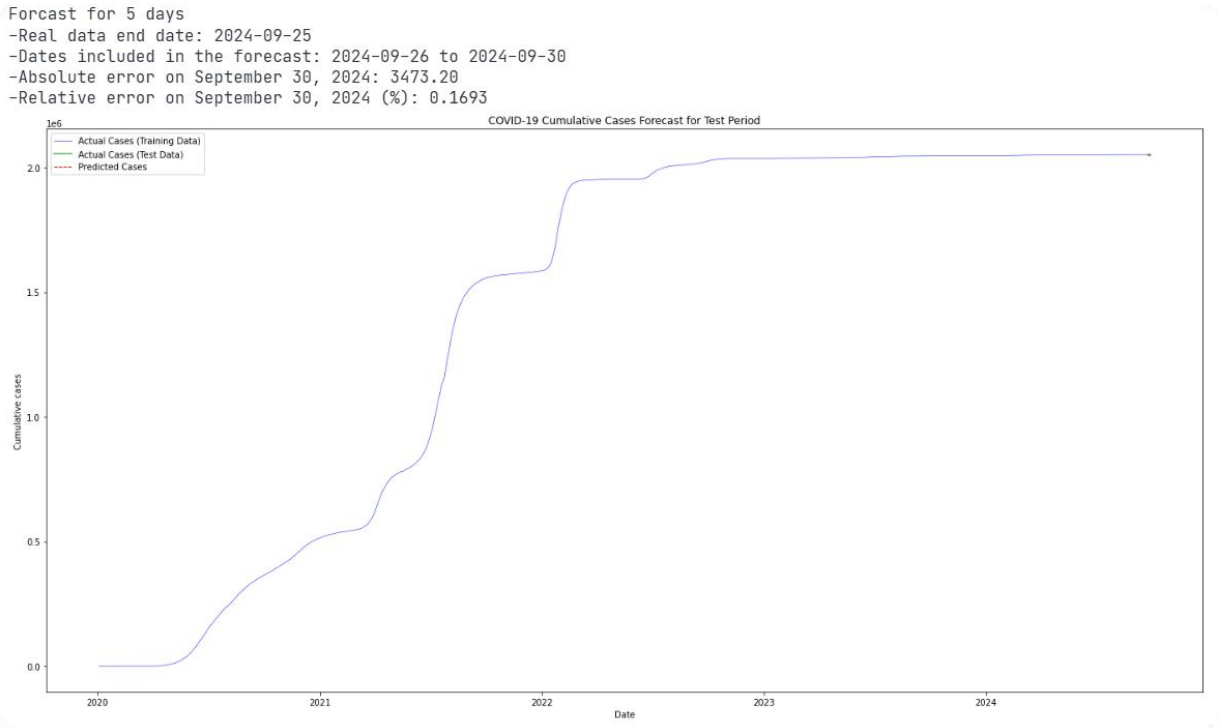

Figure 7. Forecast 5 days Cumulative Cases by dataset 1.

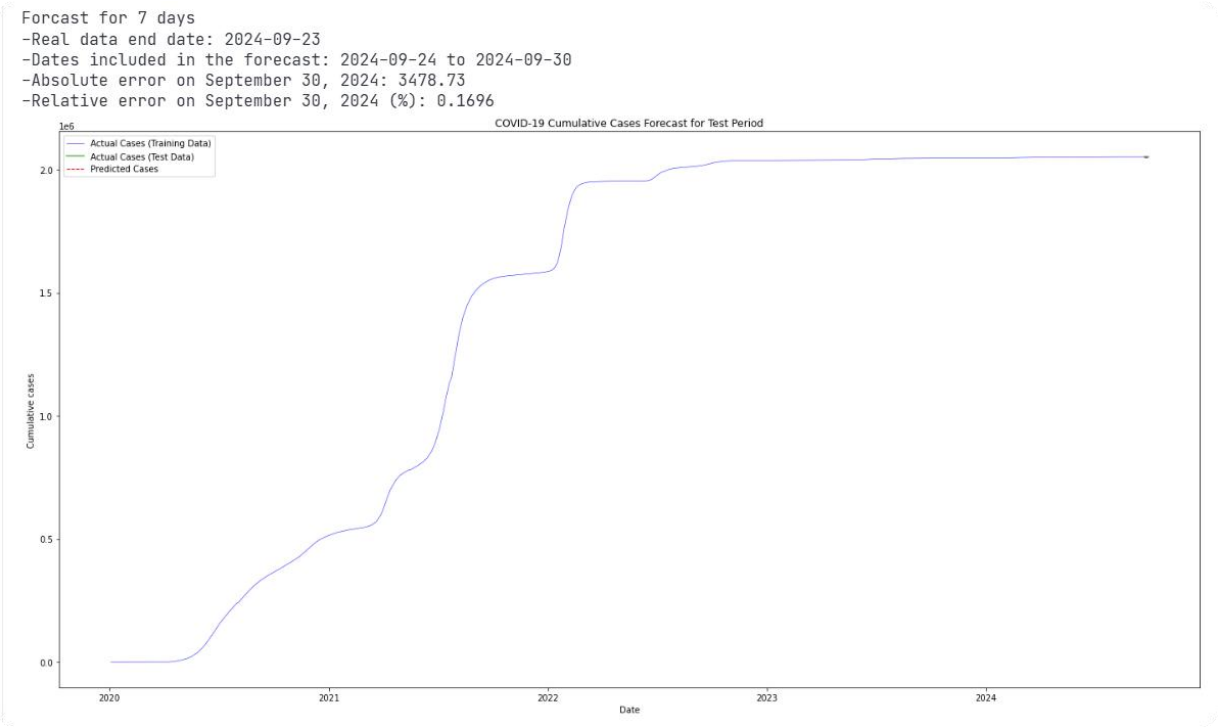

Figure 8. Forecast 7 days Cumulative Cases by dataset 1.

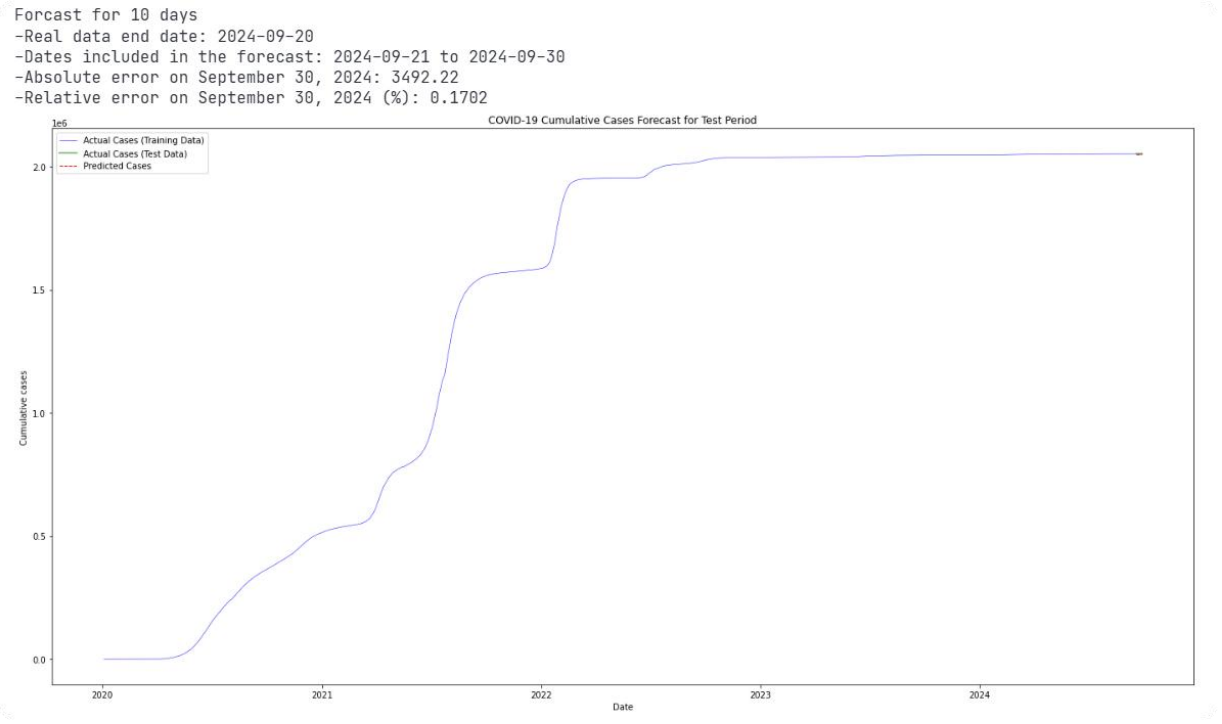

Figure 9. Forecast 10 days Cumulative Cases by dataset 1.

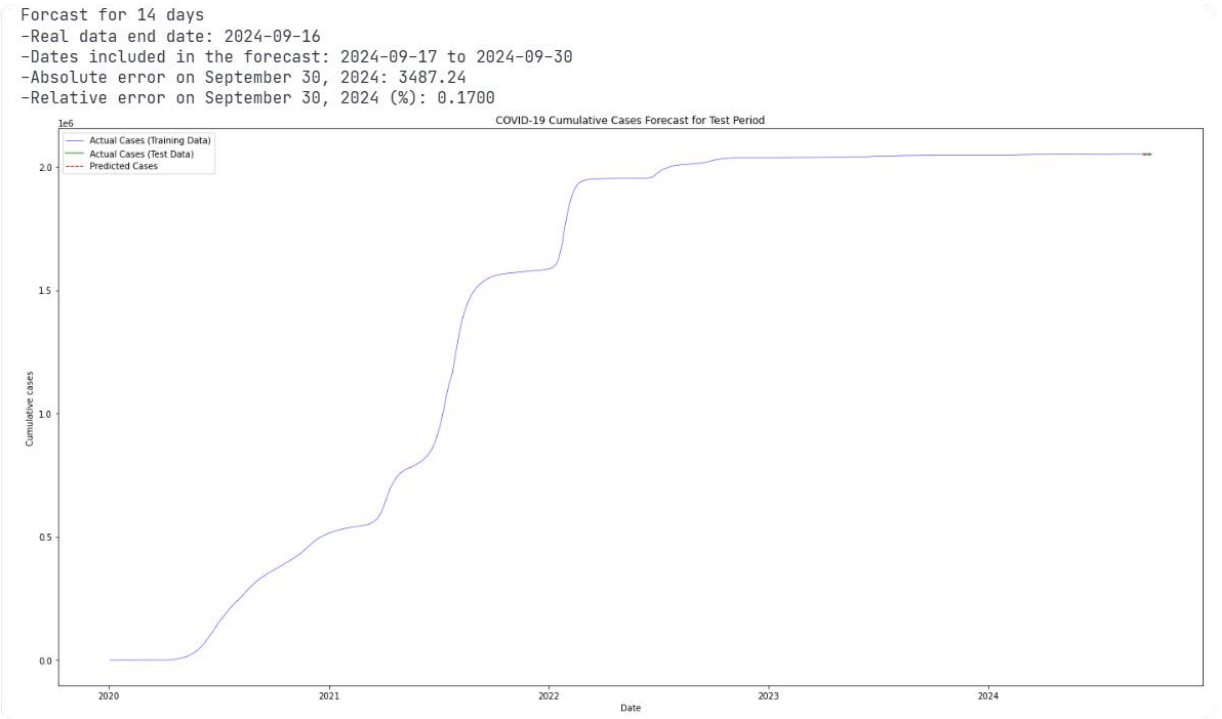

Figure 10. Forecast 14 days Cumulative Cases by dataset 1.

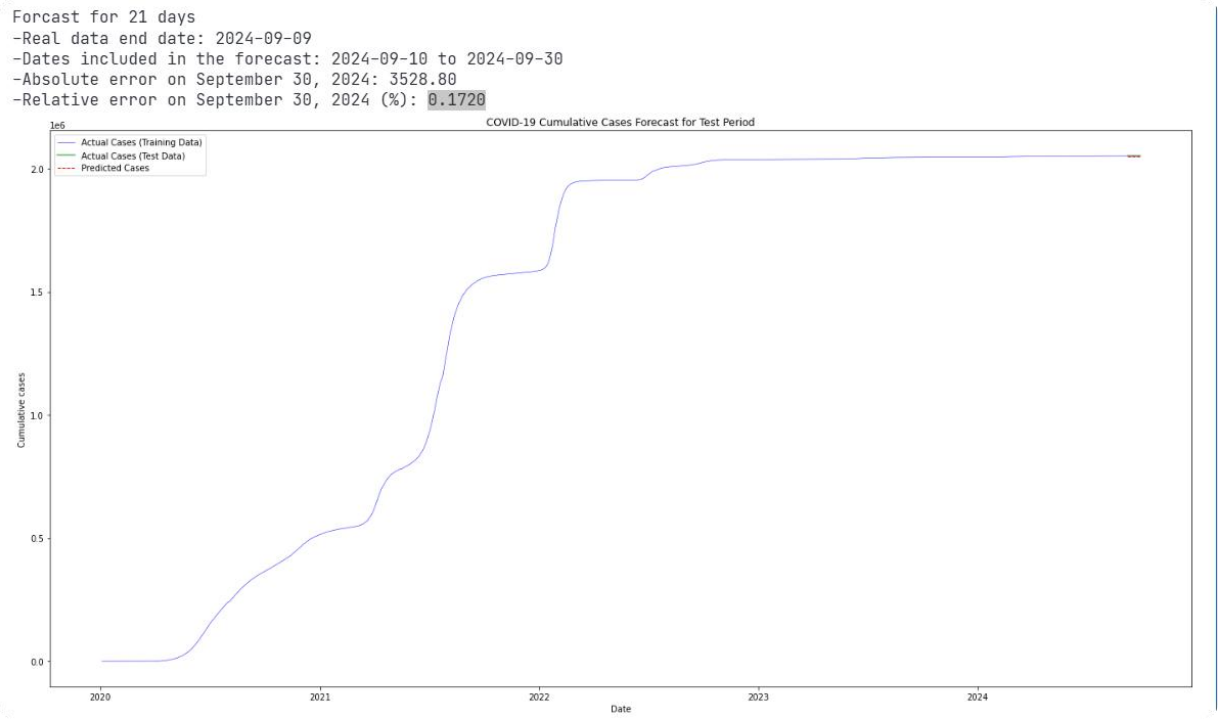

Figure 11. Forecast 21 days Cumulative Cases by dataset 1.

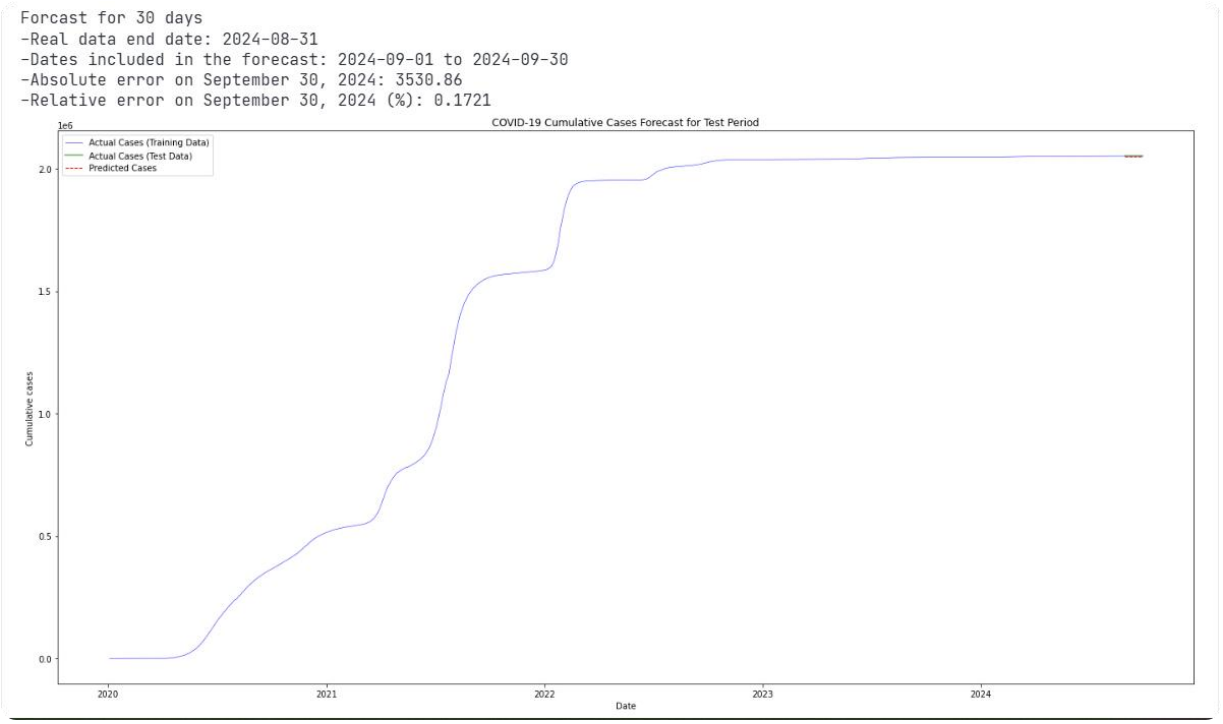

Figure 12. Forecast 30 days Cumulative Cases by dataset 1.

Figures 13 through 18 show the predicted cumulative number of confirmed cases using the SGD regression model with the second dataset, for the same prediction horizons (3 days, 5 days, 7 days, 10 days, 14 days, 21 days, and 30 days).

```
Forcast for 3 days
-Real data end date: 2024-09-27
-Dates included in the forecast: 2024-09-28 to 2024-09-30
-Absolute error on September 30, 2024: 17.90
-Relative error on September 30, 2024 (%): 0.0009
```

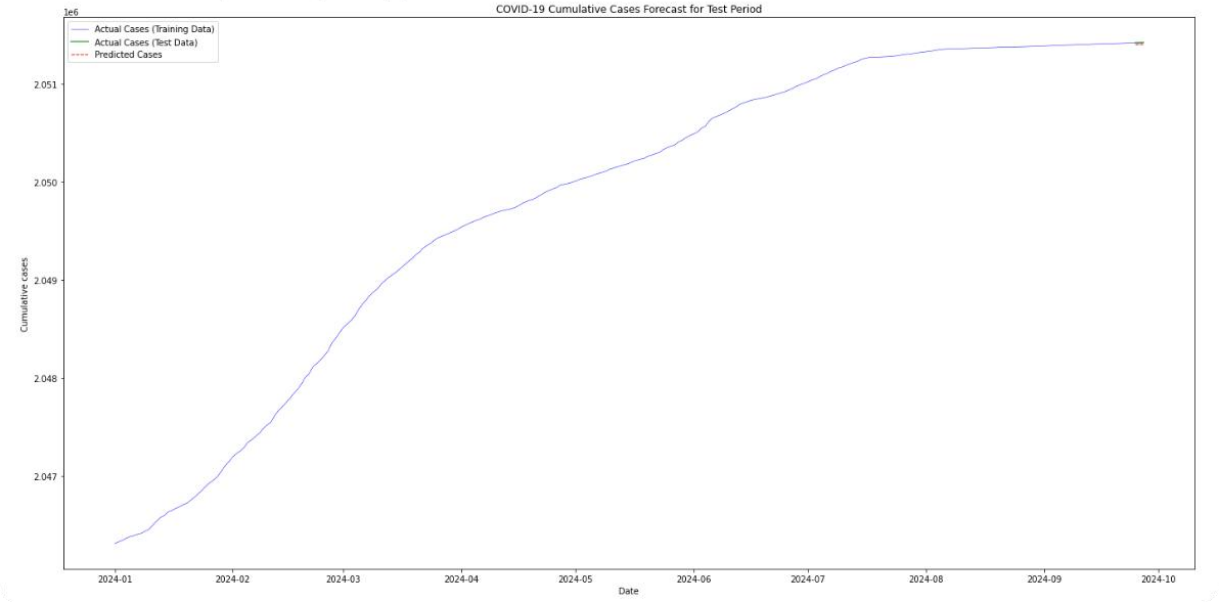

Figure 13. Forecast 3 days Cumulative Cases by dataset 2.

```
Forcast for 5 days
-Real data end date: 2024-09-25
-Dates included in the forecast: 2024-09-26 to 2024-09-30
-Absolute error on September 30, 2024: 14.75
-Relative error on September 30, 2024 (%): 0.0007
```

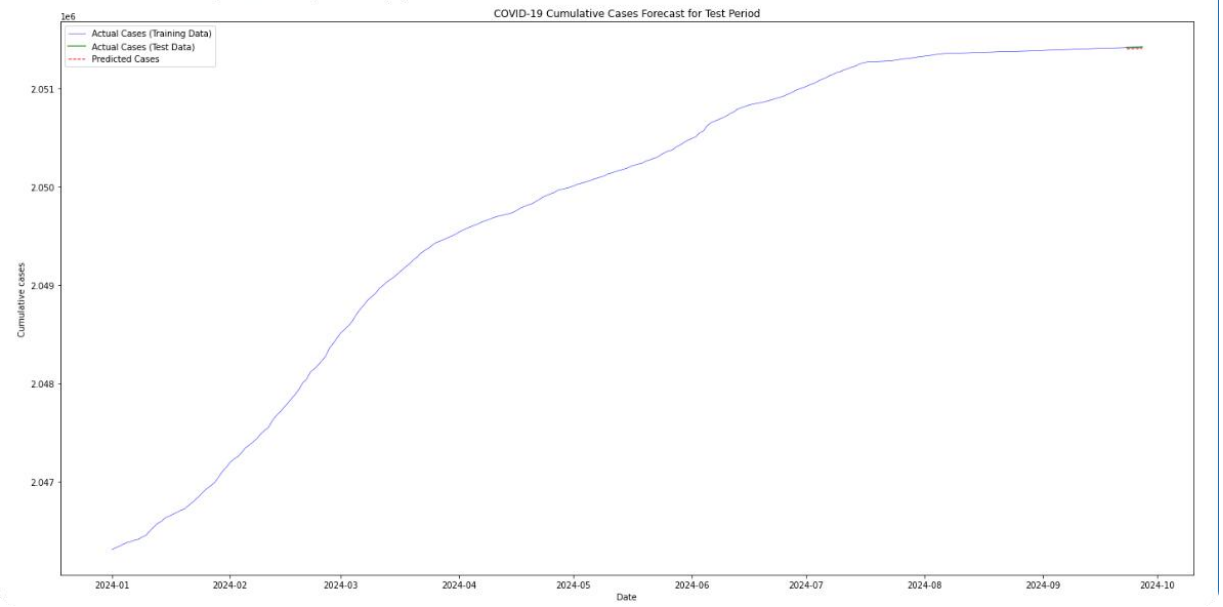

Figure 14. Forecast 5 days Cumulative Cases by dataset 2.

```
Forcast for 7 days
-Real data end date: 2024-09-23
-Dates included in the forecast: 2024-09-24 to 2024-09-30
-Absolute error on September 30, 2024: 15.40
-Relative error on September 30, 2024 (%): 0.0008
```

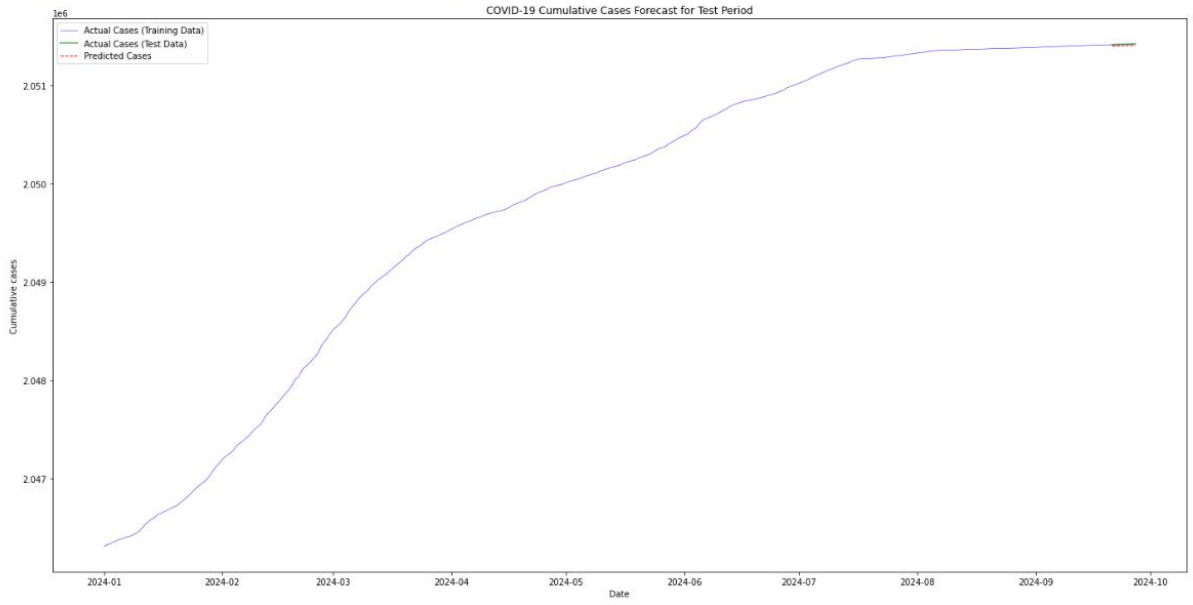

Figure 15. Forecast 7 days Cumulative Cases by dataset 2.

```
Forcast for 10 days
-Real data end date: 2024-09-20
-Dates included in the forecast: 2024-09-21 to 2024-09-30
-Absolute error on September 30, 2024: 18.42
-Relative error on September 30, 2024 (%): 0.0009
```

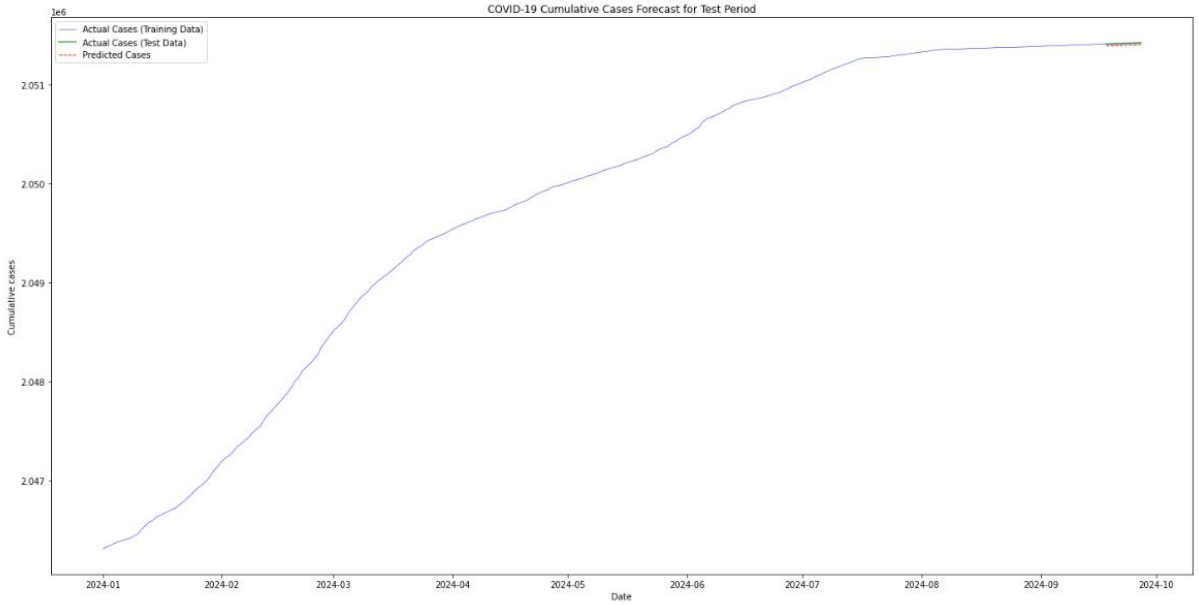

Figure 16. Forecast 10 days Cumulative Cases by dataset 2.

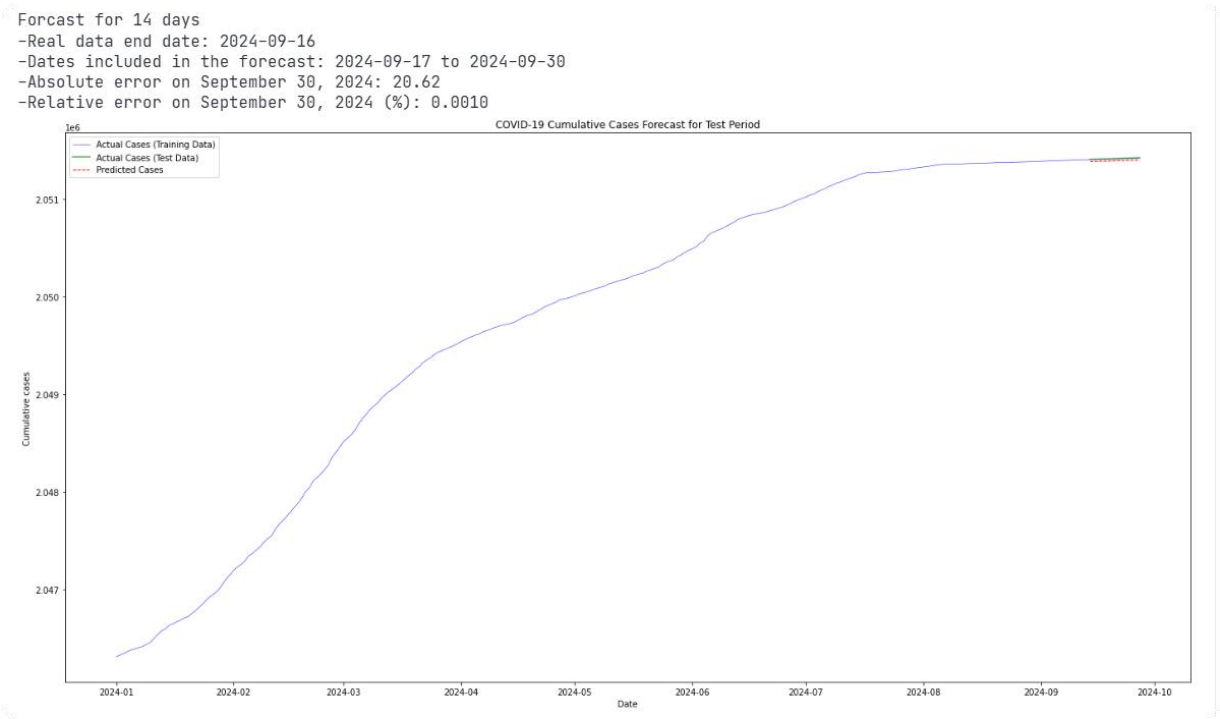

Figure 17. Forecast 14 days Cumulative Cases by dataset 2.

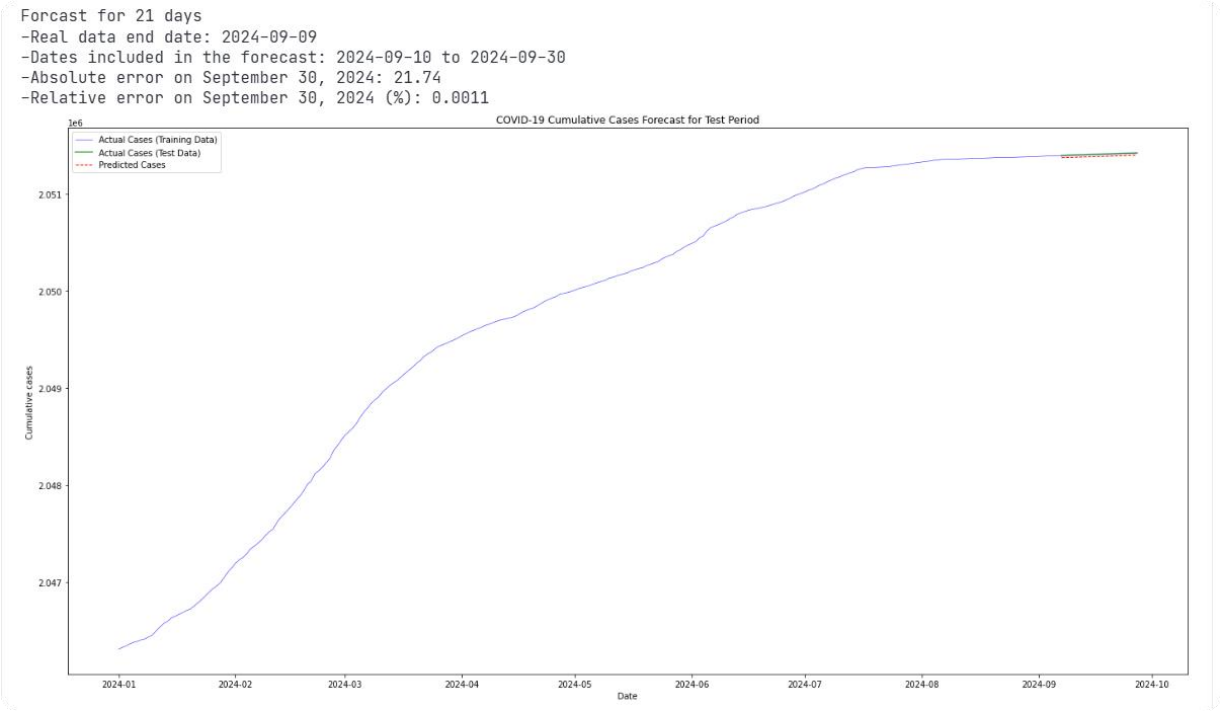

Figure 18. Forecast 21 days Cumulative Cases by dataset 2.

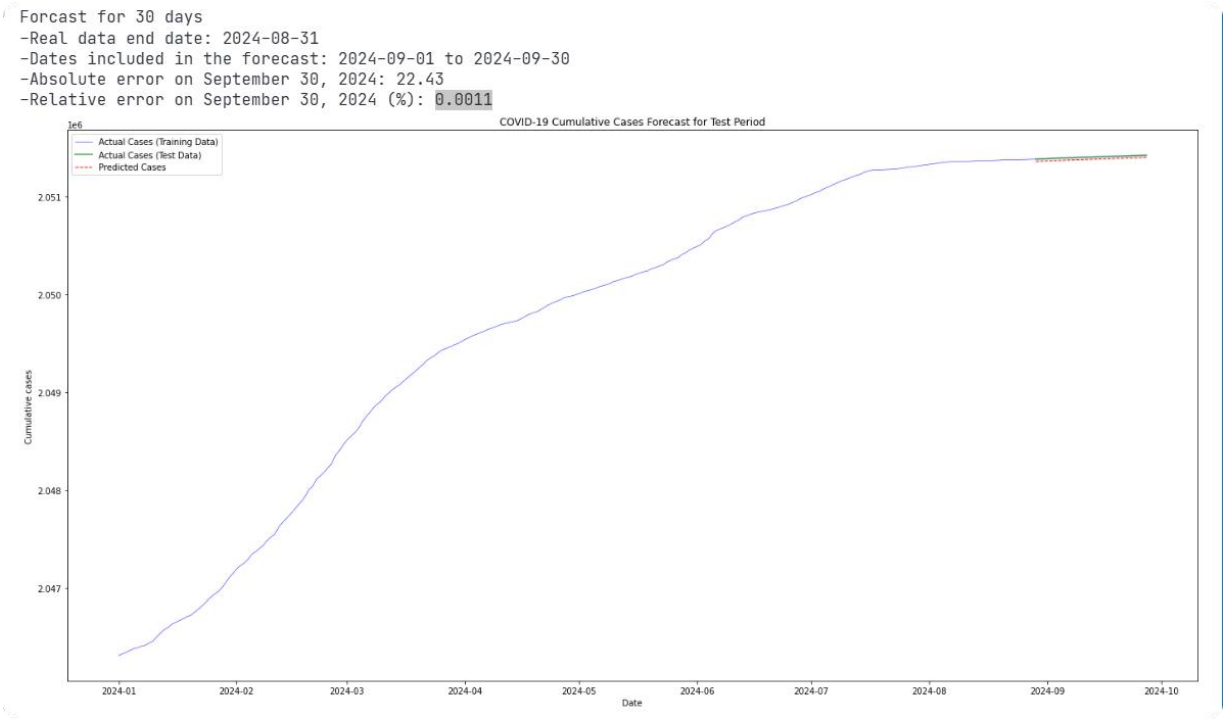

Figure 19. Forecast 30 days Cumulative Cases by dataset 2.

To evaluate the prediction accuracy of the SGD regression model, we calculated the absolute and relative errors for different prediction horizons (3 days, 5 days, 7 days, 10 days, 14 days, 21 days, and 30 days). These error metrics help us understand the model's performance across various time ranges.

Table 1 shows the absolute errors of the predictions made by the SGD regression model using the two datasets. Through these data, we can visually see the prediction accuracy of the model for different prediction period and with different datasets.

Table 1. Absolute Errors for Different Prediction Duration

Duration of forecast	Dataset1	Dataset2
3	3467.54	17.90
5	3473.20	14.75
7	3478.73	15.40
10	3492.22	18.42
14	3487.24	20.62
21	3528.80	21.74
30	3530.86	22.43

Table 2 shows the relative errors of the predictions made by the SGD regression model using the two datasets. Relative error normalizes the absolute error into a percentage, providing a better reflection of the relative difference between the predicted and actual values. The data in Table 2 help us understand the relative prediction accuracy of the model for different prediction period and with different datasets.

Table 2. Relative Errors for Different Prediction Duration

Duration of forecast	Dataset1	Dataset2
3	0.1690	0.0009
5	0.1693	0.0007
7	0.1696	0.0008
10	0.1702	0.0009
14	0.1700	0.0010
21	0.1720	0.0011
30	0.1721	0.0011

2.4.3 Accuracy Analysis

This study evaluated the prediction accuracy of the Stochastic Gradient Descent (SGD) regression model across different datasets and prediction horizons. Tables 1 and 2 show the absolute and relative errors of the SGD prediction model using two datasets for different prediction horizons. Through these data, we can comprehensively assess the performance of the SGD regression model across different datasets and prediction horizons, and determine its reliability and accuracy in practical applications.

The experimental results indicate that the characteristics of the dataset and the length of the prediction horizon significantly influence the model's accuracy. When trained on Dataset 1, the model's absolute prediction errors ranged from 3467.54 to 3530.86, with relative errors between 0.1690% and 0.1721%. In contrast, the model trained on Dataset 2 showed significantly lower absolute

errors, ranging from 14.75 to 22.43, and relative errors reduced to between 0.0007% and 0.0011%.

A deeper analysis reveals that Dataset 1 (January 2020 to September 2024) covers the entire period from the early stages of the pandemic to 2024, including multiple outbreaks and significant fluctuations. The complexity and instability of this dataset increase the noise in the data, making it difficult for the model to capture stable trends. Consequently, the model's performance on Dataset 1 is poorer, with higher absolute and relative errors. Dataset 2 (January 2024 to September 2024) includes only the data from 2024, which is a post-pandemic period characterized by relatively stable and consistent trends. This stability reduces the noise in the data, allowing the model to more accurately identify underlying patterns, resulting in higher prediction accuracy on Dataset 2.

From the above analysis, we find that the characteristics of the dataset, such as data stability and consistency, are key factors influencing model performance. Using a stable and consistent dataset (such as Dataset 2) can significantly enhance the model's prediction accuracy. Additionally, the prediction horizon also affects model performance, with short-term predictions generally being more accurate than medium- to long-term predictions.

In conclusion, the predictive performance of the SGD model is influenced not only by the prediction horizon but, more importantly, by the quality and stability of the training dataset. The experimental results highlight the critical importance of dataset selection and analysis in building high-precision predictive models.

3 CONCLUSIONS

This study assessed the efficacy of the Stochastic Gradient Descent (SGD) regression model in forecasting cumulative confirmed COVID-19 cases across various prediction horizons, ranging from 3 to 30 days. The findings revealed that the model's prediction accuracy is significantly influenced by the length of the prediction horizon. Shorter prediction horizons generally resulted in higher accuracy, likely due to the reduced complexity and uncertainty of near-term forecasts. However, the variations in accuracy across different horizons were relatively minor, suggesting that the model retains a strong degree of reliability even for longer forecasting periods. This robustness makes it a viable option for both short- and medium-term epidemic prediction tasks.

A key determinant of the model's performance was the choice of the training dataset. The analysis highlighted that training the model using the second dataset, covering the period from January 1, 2024, to September 30, 2024, led to improved prediction accuracy and reduced errors compared to the first dataset. This improvement is likely attributable to the relative stabilization of the COVID-19 pandemic during this period, which resulted in more consistent and predictable data patterns. The stabilization phase allowed the model to better capture the underlying dynamics of case growth, enhancing its ability to generalize and produce reliable forecasts. In contrast, the first dataset may have included periods of heightened volatility and uncertainty, which can pose challenges for model training and prediction accuracy.

Moreover, the study underscores the critical importance of data stability and relevance in model development. The second dataset not only provided a more stable training environment but also reflected the most recent trends in the pandemic, ensuring that the model remained responsive to current dynamics. These findings suggest that the careful selection of datasets, particularly those

representing stabilized periods of an epidemic, can significantly enhance the performance and reliability of predictive models.

In summary, the SGD regression model demonstrated strong potential as a robust tool for forecasting cumulative COVID-19 cases. Its consistent performance across varying prediction horizons, coupled with its improved accuracy when trained on stable, recent data, highlights its adaptability and reliability for epidemic forecasting. These results emphasize the necessity of leveraging high-quality, stable datasets and accounting for data characteristics such as volatility and trend stability when developing predictive models. By doing so, researchers and policymakers can improve the accuracy and utility of epidemic forecasting tools, ultimately supporting more informed public health decision-making.

Future research can further advance the performance and practicality of epidemic prediction models by integrating cutting-edge machine learning techniques, such as deep learning and ensemble methods, to enhance prediction accuracy. Expanding the range of input features to include additional relevant factors—such as demographic data, mobility patterns, and environmental variables—and leveraging multi-modal data fusion can offer a more holistic understanding of epidemic dynamics. Furthermore, developing systems capable of integrating real-time data and improving the model's online learning capabilities to dynamically adapt to evolving trends is essential. Comparing the proposed model with state-of-the-art approaches will help evaluate its performance in specific contexts, while establishing a systematic model selection framework can guide the choice of the most suitable model for different scenarios. Extending the prediction horizon to 60 or 90 days and incorporating uncertainty analysis will also provide more reliable risk assessments for decision-makers. Lastly, applying the model to diverse countries and regions to test its generalizability and adaptability across varying geographical and

epidemiological conditions, as well as validating it through multi-center studies, will ensure its robustness and consistency. These research directions aim to improve the accuracy and utility of epidemic prediction models, offering stronger and more dependable support for global public health decision-making.

4 REFERENCES

1. Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., & Lessler, J. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine*.

2. World Health Organization. (2024). COVID-19. Retrieved November 24, 2024, from https://covid19.who.int/

3. World Health Organization. (2020). Coronavirus disease 2019 (COVID-19): Situation report-94.

4. Nechyporenko, A., Alekseeva, V., Nazaryan, R., & Gargin, V. (2021). Biometric recognition of personality based on spiral computed tomography data. In *2021 IEEE 16th International Conference on the Experience of Designing and Application of CAD Systems (CADSM)* (pp. 11–15). IEEE.

5. Mattiuzzi, C., & Lippi, G. (2020). Which lessons shall we learn from the 2019 novel coronavirus outbreak? *Annals of Translational Medicine, 8*(3), 48.

6. Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 115*(772), 700–721.

7. Bacaër, N. (2011). Ross and malaria (1911). In *A short history of mathematical population dynamics* (pp. 65–69). Springer.

8. Parhusip, H. A. (2020). Menelusuri Covid-19 di dunia dan di Indonesia dengan model regresi SVM, Bayesian dan Gaussian. *Jurnal Ilmiah Sains, 20*(2), 49–57.

9. Kang, H., et al. (2020). Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning. *IEEE Transactions on Medical Imaging, 39*(8), 2606–2614.

10. Dastider, A. G., Sadik, F., & Fattah, S. A. (2021). An integrated autoencoder-based hybrid CNN-LSTM model for COVID-19 severity prediction from lung ultrasound. *Computers in Biology and Medicine, 132*, 104296.

11. Bhosale, Y. H., & Patnaik, K. S. (2023). PulDi-COVID: Chronic obstructive pulmonary (lung) diseases with COVID-19 classification using ensemble deep convolutional neural network from chest X-ray images to minimize severity and mortality rates. *Biomedical Signal Processing and Control, 81*, 104445.

12. Gambhir, E., Jain, R., Gupta, A., & Tomer, U. (2020). Regression analysis of COVID-19 using machine learning algorithms. In *2020 International Conference on Smart Electronics and Communication* (pp. 65–71). IEEE.

13. Bao, S., et al. (2022). A diagnostic model for serious COVID-19 infection among older adults in Shanghai during the Omicron wave. *Frontiers in Medicine (Lausanne), 9*, 1018516.

14. Ustebay, S., Sarmis, A., Kaya, G. K., & Sujan, M. (2023). A comparison of machine learning algorithms in predicting COVID-19 prognostics. *Internal and Emergency Medicine, 18*(1), 229–239.

15. Zhao, Z., Nehil-Puleo, K., & Zhao, Y. (2020). How well can we forecast the COVID-19 pandemic with curve fitting and recurrent neural networks? *medRxiv*, 5, 1–5.

5 APPENDIX

The following code implements the SGD regression model used for predicting the cumulative number of COVID-19 cases.

Code for SGD Model Implementation:

```
import pandas as pd
import numpy as np
from sklearn.linear_model import SGDRegressor
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.pipeline import make_pipeline
from datetime import datetime, timedelta
#Divide the training and test sets based on the prediction horizon.
def split_train_test(X, y,total_samples_, test_size_):
    x_train = X.iloc[:total_samples_ - test_size_]
    y_train_ = y.iloc[:total_samples_ - test_size_]
    x_test = X.iloc[total_samples_ - test_size_:]
    y_test_ = y.iloc[total_samples_ - test_size_:]
    return x_train, y_train_, x_test, y_test_
#load data source
data = pd.read_csv('/data/notebook_files/.private/cases_per_year.csv')
data['Date_reported'] = pd.to_datetime(data['Date_reported'], format='%Y/%m/%d')
#Construct lag features.
cases = data['Cumulative_cases']
data['lag_1'] = cases.shift(1)
data['lag_2'] = cases.shift(2)
data['lag_3'] = cases.shift(3)
data = data.dropna()
data.sort_values(by='Date_reported', inplace=True)

x_data = data[['lag_1', 'lag_2', 'lag_3']]
y_data = data[['Cumulative_cases']]
test_size = 30
total_samples = len(x_data)
dates = data['Date_reported'].values
dates_test = dates[total_samples - test_size:]

X_train, y_train, X_test, y_test = split_train_test(x_data, y_data,total_samples, test_size)
X_train = X_train.values
y_train = y_train.values.ravel()
X_test = X_test.values
#Build the SGDRegressor model.
model = make_pipeline(
    StandardScaler(),
    SGDRegressor(max_iter=10000,tol=1e-3,learning_rate='adaptive',eta0=0.01,alpha=0.01,random_state=42)
)
```

```
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
#Calculate the absolute error and relative error.
absolute_error = np.abs(y_test.values[-1] - y_pred[-1])[0]
relative_error = absolute_error / y_test.values[-1][0] * 100
base_date = datetime(2024, 9, 30)

enddate = base_date - timedelta(days=test_size)
startdate = enddate + timedelta(days=1)

print(f"Forcast for {test_size} days")
print(f"-Real data end date: {enddate.strftime('%Y-%m-%d')}")
print(f"-Dates included in the forecast: {startdate.strftime('%Y-%m-%d')} to 2024-09-30")
print(f"-Absolute error on September 30, 2024: {absolute_error:.2f}")
print(f"-Relative error on September 30, 2024 (%): {relative_error:.4f}")
plt.figure(figsize=(24, 12))
plt.plot(data['Date_reported'], data['Cumulative_cases'], label='Actual Cases (Training Data)',
color='blue',alpha=1,linewidth=0.5)
plt.plot(dates_test, y_test, label='Actual Cases (Test Data)', color='green',linestyle='-', alpha=0.5,linewidth=2)
plt.plot(dates_test, y_pred, label='Predicted Cases', color='red', alpha=1,linestyle='--',linewidth=1)
plt.title('COVID-19 Cumulative Cases Forecast for Test Period')
plt.xlabel('Date')
plt.ylabel('Cumulative cases')
plt.legend()
plt.show()
```