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 Least Angle Regression model

Author:
Final year Master's Program student,
group 63
specialty - Computer Sciences and
Information Technologies,
educational program: "Informatics"
Gao Ya
Supervisor: Victoriya Kuznietcova
Reviewer: Kseniia Bazilevych
Adviser: Stas Kachanov

Kharkiv, 2024

Prediction of the dynamics COVID-19 epidemic process of using the Bayesian Regression model

Abstract

To predict the COVID-19 situation, this paper employs Bayesian regression methods to construct a model that effectively forecasts the pandemic. The model is trained using data from the first day of the pandemic in the selected country and data from the first day of 2024 in the same country. Subsequently, the absolute and relative errors of predictions made by these two models for 3-day, 5-day, 7-day, and 10-day forecasts are summarized, and an evaluation of both models is provided

Keywords COVID-19, Prediction, Polynomial ,Bayesian Ridge regression

1. Introduction

Since December 31, 2019, the COVID-19 pandemic has frequently occurred, significantly impacting global health and economic development. Although vaccination efforts have led to a general control of the pandemic, a small number of cases continue to emerge. Therefore, building prediction models for COVID-19 cases is crucial for analyzing the trends in case data. Such models provide a solid foundation for countries to implement appropriate public health strategies.

Due to the nonlinear changes in COVID-19 cases numbers, it is essential to choose an appropriate method for model construction that accurately predicts future cases and quantifies the uncertainty of those predictions. In the classical statistical framework, regression analysis involves mathematically processing a large amount of statistical data to determine the relationship between the input independent variable X and the output dependent variable y, establishing a regression equation (functional expression) with a good correlation. Based on this, the method is used for extrapolation to predict future changes in the dependent variable y. Among various regression analysis methods, linear regression

is a significant approach due to its simplicity and effectiveness, consistently playing an extremely important role in regression analysis. However, despite its simplicity and effectiveness in practical use, linear regression has substantial limitations, which can sometimes lead to many inaccurate or even contradictory conclusions. Generally, linear models can capture linear relationships between variables; however, if there exists a certain nonlinear functional relationship between the variables, and the expected function is continuously differentiable, then according to Taylor's theorem, this function can be approximated by a polynomial function. Specifically, as long as the changes in the independent variable are small enough, we can use a linear model for approximation.

Through data analysis, we found that the COVID-19 case data we collected typically exhibits significant patterns of increase, decrease, and relative stability over time. The number of cases may be influenced by multiple factors, resulting in nonlinear growth or fluctuations in the data. For instance, the case numbers started from just 1 and surged to millions, indicating a vast range and high volatility. Given these data characteristics, a direct linear model cannot capture the complex patterns. Therefore, I chose to construct a nonlinear prediction model using polynomial expansion of time features and Bayesian Ridge regression, along with Bayesian formulas for parameter updating and uncertainty quantification.The constructed Bayesian nonlinear regression model demonstrates good predictive accuracy on the COVID-19 datasets and provides reasonable confidence intervals.

2. Theoretical Background

The statistical framework used in this study is the Bayesian Ridge Regression model combined with polynomial feature expansion. This framework leverages the prior and posterior inference capabilities of Bayesian methods, along with the advantages of polynomial features to capture nonlinear relationships, to construct a nonlinear prediction model through several key steps: data loading, feature engineering, data segmentation, model training, and prediction.First, the study loads COVID-19 case data and converts the date column to a datetime format for time series analysis. During the feature engineering phase, several important features are generated, including days, months, years, and case increments (cases_diff). These features form the model's inputs, helping to capture the temporal evolution of the pandemic.In the data segmentation process, key moments of pandemic fluctuations are identified by detecting changes in case numbers. Although these change points are not directly used to adjust the model's priors, they provide important references for defining the scope of data analysis and

modeling. Subsequently, the data is divided into training and prediction sets, and the training data is standardized to ensure that different features have the same scale, thereby avoiding bias in the model due to differences in the numerical ranges of features.

During the model training phase, this study fits the polynomial-expanded training data using Bayesian Ridge Regression. Bayesian Ridge Regression introduces prior assumptions to constrain the distribution of model parameters, thus alleviating overfitting issues when data is scarce or when the feature dimensions are high. The update of the posterior distribution is accomplished through Bayes' theorem, allowing the model to optimize parameter estimates based on existing data.Finally, by predicting future dates, this study generates estimates of COVID-19 case numbers and quantifies the uncertainty of predictions using Bayesian methods, providing a more reliable basis for decision-making in epidemic prevention and control. This series of steps ensures that the model effectively utilizes data features, combining the Bayesian framework to deliver accurate predictions with explanations of uncertainty.

2.1 Standardizing Time Features

Standardizing time features is a common technique in data preprocessing. It involves transforming features to have a mean of 0 and a standard deviation of 1. This process ensures that the contribution of each feature to the model's performance is equal.

Purpose of Standardization:

- Equal Contribution: Different features may have different scales (e.g., time measured in seconds and distance in kilometers). Standardization brings all features onto a common scale, ensuring that no single feature dominates the model.
- Improved Convergence: For optimization algorithms, standardizing data can accelerate the convergence speed during training, as gradients become more uniform.
- Enhanced Interpretability: Standardized features make the model coefficients easier to interpret, as they reflect the relative importance of each feature on the same scale.

Standardization transforms features using the following formula:

(1) Calculate Mean and Standard Deviation:

For the features in the training set X_{train}, compute the mean μ_X, and standard deviation σ_X.

Mean formulas can be mathematically expressed as:

$$\mu_X = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Where:
μ_X is the mean of the feature X.
n is the number of samples.
$X_i$is the feature value for each sample.

The purpose of this formula is to compute the average value of the feature, representing the central position of that feature. The mean is an important statistic that helps us understand the overall trend of the data.

Standard Deviation formulas can be mathematically expressed as:

$$\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)^2}$$

Where:
σ_X is the standard deviation of feature X.

This formula calculates the degree of variation of feature values from the mean. A larger standard deviation indicates that the data is more dispersed, while a smaller standard deviation indicates that the data is more concentrated.

(2) Standardize Features
Use the mean and standard deviation to standardize the features, resulting in the standardized feature $\tilde{X}$,formulas can be mathematically expressed as:

$$\tilde{X} = \frac{X - \mu_X}{\sigma_X}$$

Where:
$\tilde{X}$is the standardized feature value.
$X$is the original feature value.

Using this formula, we subtract the mean from each feature value and then divide by the standard deviation. The result is the standardized data will have a mean of 0. the standardized data will have a standard deviation of 1.

2.2 Polynomial Features

Polynomial features are a machine learning and statistical modeling technique that, after standardizing features, generates higher-order features by introducing polynomial expansions of the input features, effectively transforming the original feature space into a higher-dimensional space. This transformation allows models such as linear regression to effectively capture nonlinear relationships in the data. In our study of the actual data analysis of COVID-19 cases in South Africa, the relationship between the input features and the output target is nonlinear. Traditional linear models assume a linear relationship between features and the target variable, which may not adequately reflect the complex

patterns in the data. Therefore, by constructing nonlinear features and inputting them into the model, we can better fit the data and improve model performance.In summary, polynomial expansion is a method for generating nonlinear features by computing higher powers of the original features and the interaction terms between features. These new features provide the model with a greater capacity to represent the original data in a higher-dimensional space.

Let the original input feature vector be $\tilde{X}$ (the standardized features), with a dimension of n, i.e., $\tilde{X} = [x_1, x_2, \ldots, x_n]^\top$. We can generate the polynomial feature formula $\Phi(\tilde{X})$:

Where: $$\Phi(\tilde{X}) = [1, \tilde{X}, \tilde{X}^2, \ldots, \tilde{X}^d]$$

- $\Phi(\tilde{X})$ represents the polynomial feature vector.
- $\tilde{X}$ is the standardized feature.
- 1 represents the constant term (i.e., the intercept term).
- $\tilde{X}^2, \ldots, \tilde{X}^d$ are the higher-order terms, up to the highest degree d.
- d is the highest degree of the polynomial, referred to as the "degree" of the polynomial.

Example：

Suppose $\tilde{X} = [x_1, x_2]$ and d = 2. The generated polynomial features are：

For d = 3, the features include:

$$\Phi(\tilde{X}) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]$$
$$\Phi(\tilde{X}) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2, x_1^3, x_2^3, x_1^2 x_2, x_1 x_2^2]$$

Application Scenarios

(1) Nonlinear Regression

In nonlinear regression tasks, polynomial features enable simple linear regression models to fit complex nonlinear relationships. For example, given data (x, y) with the true relationship:

$$y = a + bx + cx^2$$

We can construct features $[1, x, x^2]$ and use a linear model:

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2$$

(2) Classification Tasks

In classification problems, polynomial features can improve a model's ability to distinguish complex data distributions. For instance, logistic regression with polynomial features can effectively fit datasets with nonlinear decision boundaries.

(3) Higher-Order Interaction Features

For problems involving multiple variables, polynomial expansion captures higher-order interactions between features. For example, $[x_1 x_2, x_1^2 x_2^3]$ describes complex interactions between x_1 and x_2.

Considerations in Polynomial Expansion

While polynomial feature expansion is very powerful, there are several considerations to keep in mind:

(1) Degree Selection: Choosing the highest degree ddd is crucial. A lower degree may fail to capture the complexity of the data, while a higher degree may lead to overfitting. Techniques such as cross-validation can be used to determine the optimal degree.

(2) Feature Scaling: Since the magnitude of polynomial features can grow rapidly, it is important to ensure that features are appropriately scaled (e.g., standardized) during model training to maintain numerical stability.

(3) Interpretability: The interpretability of high-degree polynomial models may decrease. As the number of terms increases, it becomes more challenging to understand the impact of each feature on the predictions.

After standardizing the data and performing polynomial expansion, we can effectively prepare the data for the Bayesian ridge regression model. Based on the standardized and expanded features, Bayesian ridge regression can be used for parameter estimation and prediction. Next, we will elaborate on Bayesian ridge regression, Bayes' theorem, and ridge regression in detail.

2.3 Bayesian Ridge Regression

Bayesian ridge regression is a linear regression model that combines Bayesian statistics and ridge regression. By introducing a prior distribution for the model parameters, it allows for the simultaneous consideration of data and prior information during parameter estimation. The process of assuming the prior distribution of parameters to be Gaussian and performing Bayesian inference to obtain the posterior distribution of the parameters naturally introduces the regularization effect of ridge regression. This approach effectively reduces model complexity and helps avoid overfitting, particularly after features have been standardized and polynomially expanded, enabling a better capture of complex relationships in the data. The following is a detailed description of Bayesian ridge regression. Assuming the model is:

$$y = \Phi(\tilde{X})^\top \theta + \epsilon$$

where:

- y is Target variable.
- $\Phi(\tilde{X})$ is feature matrix after polynomial expansion.
- θ is regression coefficients.

- $\epsilon \sim N(0, \sigma^2)$ independent and identically distributed Gaussian noise.

Bayesian ridge regression applies a Gaussian prior distribution to the parameter θ:

$$p(\theta \mid \lambda) = N(0, \lambda^{-1} I)$$

Where λ is a hyperparameter that represents the strength of regularization

2.3.1 Prior Distribution

Bayesian ridge regression is a statistical method that incorporates prior knowledge into regression analysis. Unlike traditional least squares methods, Bayesian approaches model the regression coefficients by introducing a prior distribution, which allows for better handling of multicollinearity and overfitting issues. In Bayesian ridge regression, we assume that the regression coefficients θ follow a normal distribution, typically set as:

$$\theta \sim N(0, \lambda^{-1} I)$$

Where:

- $N(0, \lambda^{-1} I)$: This represents the prior distribution of the regression coefficients which follow a Gaussian distribution. Specifically, the regression coefficients follow a multivariate normal distribution with a mean of 0 and a covariance matrix of $\lambda^{-1} I$.
- Mean of 0: This setting reflects our initial belief about the regression coefficients, indicating that before observing the data, we assume the coefficients are close to zero. This reflects a bias towards simpler models, helping to avoid overestimating the impact of features on the target variable.This assumption is reasonable because, in many practical applications, we expect the influence of most features on the target variable to be weak.
- Variance of λ^{-1} : Introducing variance allows us to model the uncertainty of the regression coefficients. A larger λ means a smaller variance, indicating a stronger belief in the coefficients, while a smaller λ implies a larger variance, allowing for greater fluctuations.
- I is the identity matrix: This part indicates that we assume the different regression coefficients are independent and have the same variance. This assumption simplifies calculations, making the model easier to handle.
- $\lambda^{-1} I$ is covariance matrix: This part indicates that independence assumption,the regression coefficients of different features are independent of each other. Homoscedasticity assumption, all regression coefficients have the same prior variance, which is λ^{-1}.

The hyperparameter λ controls the magnitude of the variance, when λ is large, the variance is small, leading to stronger constraints on the coefficients, which results in a preference for simpler models. When λ is small, the variance is large, resulting in weaker constraints on the coefficients, allowing for more flexible fitting.

The introduction of the prior distribution has several key impacts on Bayesian Ridge Regression:

(1) Regularization Effect, By incorporating a prior distribution, Bayesian Ridge Regression can effectively regularize the regression coefficients, suppressing overfitting. This is because the prior distribution imposes a certain penalty on the regression coefficients, resulting in a smoother model.

(2) Handling Multicollinearity, In the presence of multicollinearity, traditional least squares methods may lead to unstable estimates. Bayesian Ridge Regression can alleviate this issue to some extent by introducing a prior distribution, providing more robust parameter estimates.

(3) Incorporating Prior Information, In some cases, researchers may have prior knowledge regarding certain regression coefficients. By adjusting the parameters of the prior distribution, this prior information can be integrated into the model, thereby enhancing its predictive capability.

2.3.2 Likelihood Function

Assuming the observed data y follows a Gaussian distribution given the feature matrix $\Phi(\tilde{X})$ and parameters θ:

$$p(y \mid \theta, \sigma^2) = N(y \mid \Phi(\tilde{X})^T \theta, \sigma^2 I)$$

Where:

σ^2 is variance of the noise.

Next, we will briefly explain the mathematical derivation of the probability density function.

Assuming each observation y_i in y(i.e., components of the target variable) is independent given $\Phi(\tilde{X})$ and θ, and follows the same normal

distribution $N(\Phi(\tilde{X})^T\theta, \sigma^2 I)$. For n samples, the overall likelihood function is:

$$p(y \mid \theta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \Phi(\tilde{X})_i^T\theta)^2\right)$$

Expressing y in vector form, the probability density function of the likelihood function can be written as:

$$p(y \mid \theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y - \Phi(\tilde{X})^T\theta\|^2\right)$$

where:

$$\|y - \Phi(\tilde{X})^T\theta\|^2 = (y - \Phi(\tilde{X})^T\theta)^T(y - \Phi(\tilde{X})^T\theta)$$

represents the squared error.

This form indicates that $y$follows a Gaussian distribution：

$$y \sim N(\Phi(\tilde{X})^T\theta, \sigma^2 I)$$

Where:

- Mean $\Phi(\tilde{X})^T\theta$ represents the model's predicted values.
- Covariance Matrix $\sigma^2 I$ indicates the independence of observations (no correlation among the noise).

From the least squares perspective of maximum likelihood estimation, optimizing $p(y \mid \theta, \sigma^2)$is equivalent to minimizing the negative log-likelihood:

$$-\log p(y \mid \theta, \sigma^2) \propto \|y - \Phi(\tilde{X})^T\theta\|^2$$

This is consistent with the objective function of ordinary least squares, fitting parameters by minimizing the sum of squared residuals.

Thus, the role of the likelihood function in Bayesian Ridge Regression is as follows:

(1) Data Constraint that describes how the target variable y is generated from the feature matrix $\Phi(\tilde{X})$ and parameters θ.

(2) Connection Between Prior and Posterior that $p(\theta \mid y) \propto p(y \mid \theta)p(\theta)$ the likelihood function is central to constructing the posterior distribution.

(3) Model Uncertainty that The Gaussian noise assumption provides a quantification of uncertainty in the model, where the noise variance σ^2affects the prediction range in the predictive distribution.

2.3.3 Bayesian Theory

Bayesian Theory was first proposed by the British theologian Bayes to describe the relationship between two conditional probabilities. It is a

statistical method that uses probability to express uncertainty. Based on the basic formula of Bayes' theorem:

$$P(\theta|Y) = \frac{P(Y|\theta) \cdot P(\theta)}{P(Y)} \tag{1}$$

where:

- $P(\theta|Y)$ is the posterior distribution, representing the probability of the parameter θ given the data Y;
- $P(Y|\theta)$ is the likelihood function, indicating the probability of the data under parameter θ;
- $P(\theta)$ is the prior distribution, reflecting prior beliefs about θ;
- $P(Y)$ is marginal likelihood, acting as a normalizing constant;

This formula provides the foundation for parameter estimation in Bayesian regression, where prior knowledge is combined with observed data to generate better models to predict data.

Since our features x have been processed through polynomial transformation, the updated Bayes' theorem is as follows:

$$p(\theta \mid y, \Phi(\tilde{X})) = \frac{p(y \mid \theta, \Phi(\tilde{X}))p(\theta)}{p(y \mid \Phi(\tilde{X}))}$$

Where:

- $p(\theta \mid y, \Phi(\tilde{X}))$ is Posterior distribution, representing the probability distribution of the parameter θ given the data y.
- $p(y \mid \theta, \Phi(\tilde{X}))$ is Likelihood function, representing the probability of observing the data y given the parameter θ.
- $p(\theta)$ is Prior distribution, representing the belief about the parameter θ before observing the data.
- $p(y \mid \Phi(\tilde{X}))$ is Model evidence, also known as marginal likelihood, representing the overall probability of observing the data y given the model, used as a normalization constant.

The posterior distribution $p(\theta \mid y, \Phi(\tilde{X}))$ represents the combined information from the data y and the prior $p(\theta)$. It is our final inference about the parameter θ.

The likelihood function $p(y \mid \theta, \Phi(\tilde{X}))$ reflects the probability of generating the data y given the parameter θ and the model structure $\Phi(\tilde{X})$. It measures the degree of fit between the parameter θ and the observed data.

The prior distribution $p(\theta)$ captures our beliefs or assumptions about θ before observing the data. The prior can be based on empirical subjective assumptions or can be non-informative.

The model evidence $p(y \mid \Phi(\tilde{X}))$ is the normalization term for the posterior distribution, obtained by integrating over all possible parameter values θ, $p(y \mid \Phi(\tilde{X})) = \int p(y \mid \theta, \Phi(\tilde{X}))p(\theta)d\theta$ Although it is often ignored as a constant in parameter estimation, it plays a crucial role in model comparison (e.g., Bayesian model selection).

2.3.4 Posterior Distribution

The posterior distribution in Bayesian Ridge Regression is a core component that combines the observed information from the data (likelihood function) and prior information. It derives the distribution of the regression coefficients θ using Bayes' theorem. According to Bayes' theorem, a posterior distribution can be expressed as:

$$p(\theta \mid y, \Phi(\tilde{X})) \propto p(y \mid \theta, \Phi(\tilde{X})) \cdot p(\theta)$$

This indicates that for role of the Likelihood Function which adjusts our belief about the parameter θ based on the observed data y. For role of the Prior Distribution that provides the initial assumption about the parameter θ. Through likelihood function and prior distribution, we get the result Posterior Distribution. The posterior distribution integrates data and prior information, serving as the final basis for parameter inference.

From a mathematical standpoint, Assuming a parameter θ describes the probability distribution of a phenomenon, we initially believe it follows $p(\theta)$. Once we observe the data y, we update our belief about θ using the likelihood function $p(y \mid \theta)$ The updated distribution $p(\theta \mid y)$ is the posterior distribution.

Bayes' theorem is recursive. When new data y' is available, the current posterior distribution $p(\theta \mid y, \Phi(\tilde{X}))$ can serve as the new prior distribution, combined with the new likelihood function $p(y' \mid \theta, \Phi(\tilde{X}))$ to compute the updated posterior distribution.

2.4 Summary of Bayesian Ridge Regression

Bayesian Ridge Regression is a regression method that combines the ideas of Bayesian inference and ridge regression. It imposes constraints on model parameters by introducing prior distributions while retaining the theoretical advantages of Bayesian methods.

The core concepts are the Bayesian Method and the Ridge Regression Idea. The Bayesian Method is based on observed data and prior knowledge; the Bayesian formula combines the likelihood function and prior distribution to obtain the posterior distribution of the parameters. The Ridge Regression Idea introduces L_2 regularization (corresponding

to a Gaussian prior) on the model parameters, which reduces excessive fluctuations in the parameters and enhances the robustness of the model.

Mathematical Modeling, Assume the model is given by：

$$y = \Phi(\tilde{X})^\top \theta + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

(1) Likelihood Function

Given the model parameters θ and the design matrix $\Phi(\tilde{X})$, the target variable y follows a Gaussian distribution:

$$p(y \mid \theta, \Phi(\tilde{X}), \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\|y - \Phi(\tilde{X})^\top \theta\|^2\right)$$

(2) Prior Distribution

Assume the prior distribution of the parameter θ is a zero-mean Gaussian distribution:

$$p(\theta \mid \lambda) = \frac{1}{(2\pi)^{d/2}|\lambda^{-1} I|^{1/2}} \exp\left(-\frac{\lambda}{2}\|\theta\|^2\right)$$

(3) Posterior Distribution

By applying Bayes' theorem to combine the prior distribution and the likelihood function:

$$p(\theta \mid y, \Phi(\tilde{X}), \sigma^2, \lambda) \propto p(y \mid \theta, \Phi(\tilde{X}), \sigma^2) \cdot p(\theta \mid \lambda)$$

The posterior distribution takes the form of a Gaussian distribution:

$$p(\theta \mid y, \Phi(\tilde{X}), \sigma^2, \lambda) = N(\mu_\theta, \Sigma_\theta)$$

The parameters of the posterior distribution are posterior mean and posterior covariance.

Posterior Mean：

$$\mu_\theta = \left(\lambda I + \sigma^{-2}\Phi(\tilde{X})\Phi(\tilde{X})^\top\right)^{-1} \left(\sigma^{-2}\Phi(\tilde{X})y\right)$$

Posterior Covariance：

$$\Sigma_\theta = \left(\lambda I + \sigma^{-2}\Phi(\tilde{X})\Phi(\tilde{X})^\top\right)^{-1}$$

Properties of Bayesian Ridge Regression, for regularization, the Gaussian prior naturally introduces regularization, similar to the L_2 penalty in ridge regression, which controls model complexity. For Uncertainty Quantification, The posterior distribution not only provides point estimates of the parameters (posterior mean μ_θ) but also quantifies the uncertainty of the parameters (posterior covariance Σ_θ). For Robustness,By incorporating the prior distribution, Bayesian Ridge Regression can maintain stable performance even when the dataset is small or when features are highly correlated.

Ridge Regression explicitly incorporates regularization through a penalty term $\lambda\|\theta\|_2$, while Bayesian Ridge Regression naturally introduces regularization via a Gaussian prior with the same penalty term. In terms of results, Ridge Regression provides a single point estimate, whereas Bayesian Ridge Regression yields a posterior distribution that includes both the mean and the quantification of uncertainty. The theoretical foundation of Ridge Regression is based on optimization methods, whereas Bayesian Ridge Regression relies on Bayesian inference. Additionally, model uncertainty quantification is absent in Ridge Regression but is present in Bayesian Ridge Regression.

Bayesian Ridge Regression offers a range of significant advantages that make it a powerful tool in statistical modeling. One of its primary strengths is the theoretical rigor that stems from Bayesian principles, providing a unified framework for analysis that integrates prior knowledge with observed data. This approach not only enhances the robustness of the model but also allows for a more comprehensive understanding of the underlying relationships in the data. Furthermore, Bayesian Ridge Regression improves model interpretability by delivering valuable uncertainty information about the parameters through the posterior distribution. This feature helps practitioners gauge the confidence levels of the estimates, facilitating informed decision-making based on the model's outputs.The method's flexibility is another notable benefit, as it allows for the adjustment of regularization strength based on the chosen prior distribution. This adaptability makes it suitable for a wide variety of problem domains, accommodating different modeling needs and complexities. Additionally, Bayesian Ridge Regression is particularly effective in addressing issues related to multicollinearity and small sample sizes. By maintaining stable performance even under these challenging conditions, it ensures that the model remains reliable and valid, making it an essential tool for data analysts and researchers alike.

Bayesian Ridge Regression is particularly suited for several scenarios. First, it effectively addresses feature multicollinearity, where high correlations among features can distort model estimates. The regularization inherent in Bayesian Ridge Regression helps mitigate the effects of multicollinearity, leading to more stable and reliable parameter estimates.Second, it is advantageous in situations involving small sample sizes. In cases where data is limited, the use of prior distributions can impose necessary constraints on the model, enhancing its robustness despite the scarcity of information. Lastly, Bayesian Ridge Regression is ideal for contexts where quantifying uncertainty is crucial. The posterior distribution’s covariance provides valuable insights into prediction intervals and associated risks, allowing practitioners to make more

informed decisions based on the model's outputs. This combination of features makes Bayesian Ridge Regression a versatile choice for various analytical challenges, particularly in fields where data quality and quantity may be compromised.

So,Bayesian Ridge Regression integrates the principles of Bayesian statistics with those of ridge regression, effectively combining regularization with a comprehensive characterization of parameter uncertainty. This method offers a mathematically rigorous framework and a solid theoretical foundation, making it a robust choice for regression analysis. One of the key strengths of Bayesian Ridge Regression is its ability to provide not only point estimates of parameters but also to quantify the uncertainty associated with these estimates through the posterior distribution. This dual capability enhances model interpretability, allowing practitioners to understand the confidence levels of their predictions and make informed decisions based on the results. Moreover, Bayesian Ridge Regression excels in addressing common challenges encountered in data analysis, such as noise, multicollinearity among features, and limited sample sizes. The regularization effect inherent in the model helps mitigate the adverse impacts of multicollinearity, leading to more stable and reliable estimates. In situations where data is scarce, the incorporation of prior distributions allows the model to impose necessary constraints, enhancing its robustness. Overall, Bayesian Ridge Regression stands out as a practical and theoretically sound regression method, offering significant advantages in various analytical contexts. Its ability to handle complex data scenarios while providing a clear understanding of uncertainty makes it a valuable tool for researchers and practitioners in fields that require rigorous statistical modeling.

3. Model Implementation

3.1 Data Preprocessing

Data preprocessing is a critical step in constructing machine learning models, directly influencing their predictive performance and generalization capability. For predicting COVID-19 cases, raw data typically includes dates and cumulative case counts. Preprocessing ensures that date values are converted into numerical features, and input and target variables are cleaned, standardized, and engineered.

For Feature Engineering for Time Variables, time is a key variable for predicting the growth of COVID-19 cases. It must be transformed into a numerical format that models can interpret. In this study, time is

expressed as the number of days elapsed since the earliest date in the dataset. This feature represents the overall trend of case growth and serves as the primary input for the model.
The implementation is as follows:

```
# 1. Data preprocessing
df = pd.read_csv(data_path)
df['date'] = pd.to_datetime(df['date'])

# 2. Feature engineering
df['days'] = (df['date'] - df['date'].min()).dt.days
df['month'] = df['date'].dt.month
df['year'] = df['date'].dt.year
df['cases_diff'] = df['cases'].diff().fillna(0)
```

For Target Variable Cleaning, the target variable y, which represents the case count, may contain missing or anomalous values. To ensure model stability, the data must be cleaned and filled. Common outlier detection methods include statistical approaches such as the three-sigma rule, the IQR method, or manual detection based on historical trends.

3.2 Data Standardization

For Standardization, the input features might have different scales and distributions. Feeding such unprocessed features directly into the model could result in unequal contributions to the regression weights. To address this, standardization is applied to transform features into distributions with a mean of 0 and a standard deviation of 1:

$$x' = \frac{x - \mu}{\sigma}$$

Where μ is the mean and σ is the standard deviation.The implementation is as follows:

The standardized time feature $X$is then prepared for subsequent The standardized time feature X is then prepared for subsequent polynomial

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
```

ability of linear models to capture nonlinear relationships. By generating higher-order combinations of features, the model can better fit complex growth trends. For example, given a time feature X, its polynomial expansion includes terms like x^2, x^3, and so on.
Assume the input feature matrix is X, and the polynomial expansion generates the feature matrix:

$$\Phi(X) = [X, X^2, X^3, \ldots, X^d]$$

where d is the highest degree of the polynomial. The expanded feature matrix can be expressed as:

$$\Phi(X) = \begin{bmatrix} x_1 x_2 \cdots x_n \\ x_1^2 x_2^2 \cdots x_n^2 \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ x_1^d x_2^d \cdots x_n^d \end{bmatrix}$$

For Balancing Higher-Order Features and Regularization, the polynomial degree d is a critical hyperparameter. A degree that is too low may result in underfitting, while an excessively high degree may incorporate noise, increase model complexity, and lead to overfitting. In this study, $d = 3$ is chosen as a balance between fitting capability and model complexity. Additionally, regularization methods (such as the L2 regularization in Bayesian Ridge Regression) are particularly important for controlling parameter fluctuations in high-order feature scenarios.
The implementation is as follows:

```
from sklearn.preprocessing import PolynomialFeatures

poly = PolynomialFeatures(degree=degree, include_bias=False)
X_poly_train = poly.fit_transform(X_train_scaled)
```

The expanded matrix X_{poly} includes high-order features of the time variable, providing rich information for the model to capture the nonlinear growth trends of COVID-19 cases.

3.4 Model Training With Bayesian Ridge Regression

Bayesian Ridge Regression incorporates Bayesian inference and ridge regression regularization to estimate parameters while accounting for both prior information and the uncertainty in observed data. The model computes regression coefficients by maximizing the posterior distribution. For Mathematical Foundations, assume that the observed data follows a normal distribution and the parameters θ follow a Gaussian prior distribution with zero mean. The objective function can be expressed as:

$$L(\theta) = \frac{1}{2\sigma^2}\|y - \Phi(\tilde{X})\theta\|^2 + \lambda\|\theta\|^2$$

The first term represents the fitting error, measuring how well the model explains the observed data. The second term is the L2 regularization term, penalizing large parameter fluctuations.Using Bayesian inference, the posterior distribution of the parameters can be expressed as:

$$p(\theta \mid y, \Phi(\tilde{X}), \sigma^2, \lambda) \propto p(y \mid \theta, \Phi(\tilde{X}), \sigma^2) \cdot p(\theta \mid \lambda)$$

The goal of training the model is to maximize the posterior probability $p(\theta \mid y, \Phi(\tilde{X}), \sigma^2, \lambda)$, yielding the optimal estimates of the regression coefficients. The Bayesian Ridge Regression model is implemented as follows:

```
from sklearn.linear_model import BayesianRidge

model = BayesianRidge()
model.fit(X_poly_train, y_train)
```

After training, the regression coefficients not only include point estimates but also their distributional characteristics, enabling the quantification of parameter uncertainty. This capability is crucial for scenarios like pandemic predictions, where uncertainty must be explicitly accounted for.

3.5 Forecasting Future Data

Once the model is trained, it is used to predict future case counts for specified dates. The forecasting process involves the following steps, First, for Time Data Transformation and Expansion, future date inputs are first converted into standardized day features and then expanded into polynomial feature matrices:

```
future_days_scaled = scaler.transform(future_days)
X_poly_future = poly.transform(future_days_scaled)
```

Second, using the trained Bayesian Ridge Regression model, predictions for future dates are calculated. Since the model leverages posterior distributions, the predictions include not only point estimates but also confidence intervals:

```
y_future_pred = model.predict(X_poly_future)
```

The model output also provides information about the uncertainty in parameters and predictions (e.g., confidence intervals or variances), which enhances robustness and interpretability in practical applications.

3.6 Conclusion

The outlined methodology provides a robust framework for training a COVID-19 case prediction model, demonstrating a systematic integration of data preprocessing, polynomial feature expansion, and Bayesian Ridge Regression. The preprocessing stage ensures the input data is properly transformed and standardized, reducing scale inconsistencies and improving model stability. By incorporating polynomial feature expansion, the model effectively captures nonlinear growth patterns in case trends, enhancing its ability to reflect real-world complexities. Bayesian Ridge Regression further strengthens the model by leveraging regularization and probabilistic inference, which not only prevents overfitting but also provides posterior distributions that quantify prediction uncertainty. This uncertainty estimation is especially valuable for public health applications, where confidence intervals and risk assessment play a crucial role in planning and response. Finally, the model's prediction phase demonstrates its practical utility, offering actionable insights into future case trajectories. By combining accuracy with interpretability, the model serves as a reliable tool for informed

decision-making in the management and mitigation of pandemic scenarios.

4. Results

4.1 Data Description

The model training process for analyzing COVID-19 in South Africa is built on two essential datasets, each serving a unique purpose in understanding the pandemic's dynamics. The first dataset, South_Africa_total_amount.csv, is a comprehensive record that captures all confirmed COVID-19 case numbers from the date of the first reported case in South Africa until September 30, 2024. This dataset encompasses a total of 1,670 data samples, meticulously curated to provide a detailed overview of the situation over time. For dataset structure,the dataset is structured with five distinct features that are crucial for analysis. Of these, four are numerical attributes: days, month, year, and cases_diff. The days attribute counts the number of days since the first confirmed case, which allows for a straightforward chronological analysis of the pandemic's progression. This feature is particularly valuable when assessing how the number of cases changes over time, as it provides a clear timeline against which to measure the impact of interventions, public health measures, and societal behavior.The month and year attributes serve to categorize the data temporally, facilitating seasonal analysis and understanding how the pandemic's impact may vary with time. For instance, months with higher case counts may indicate seasonal trends or spikes in transmission rates, which can inform future public health strategies. The cases_diff attribute is particularly significant as it indicates the daily change in the number of confirmed cases compared to the previous day. This feature is essential for tracking the outbreak's dynamics, allowing us to observe patterns such as increases or decreases in new cases, which can signal the effectiveness of mitigation efforts.Accompanying these numerical features is a date feature labeled date, which specifies the exact date corresponding to each data entry. This is crucial for establishing a timeline of the pandemic's progression. Lastly, the dataset includes one label, cases, which quantifies the number of confirmed cases reported on each date. This label is the target variable for our predictive modeling efforts, as we seek to understand and forecast the number of cases based on the passage of days since the first reported case.The second dataset, cases_per_year.csv, complements the first by focusing specifically on COVID-19 case numbers in South Africa from January 1, 2024, to September 30, 2024. This dataset contains 274 data samples and shares the same structural features as the South_Africa_total_amount.csv file. By using two datasets with overlapping time frames, we can create a more robust model that captures a wide range of trends and patterns, enhancing our predictive capabilities.Both datasets were downloaded from the World Health

Organization (WHO) website, ensuring that we are working with reliable and up-to-date information. The WHO provides comprehensive data on pandemic-related statistics, making it a trusted source for researchers and public health officials alike.

For the purpose of predictive modeling, we utilized Bayesian Ridge Regression, a powerful method that extends linear regression by incorporating Bayesian principles. This approach allows us to perform linear regression while also accounting for uncertainty in the estimates, making it particularly suitable for datasets with inherent variability, such as those related to infectious diseases.In this modeling approach, we selected days as the feature and cases as the target label for training and prediction. By focusing on the number of days since the first confirmed case, we gain insights into how time influences the number of reported cases. This is a fundamental aspect of epidemiological modeling, as it enables us to observe how the disease spreads over time and how various factors, such as government interventions or public compliance with health guidelines, impact the progression of the outbreak. Before we could proceed with training the model, it was essential to prepare the data appropriately. This preparation involved several steps, including standardization and polynomial expansion of the feature values. Standardization is a critical preprocessing step that transforms the feature values to have a mean of zero and a standard deviation of one. This normalization helps improve the performance of many machine learning algorithms by ensuring that all features contribute equally to the model's learning process. Polynomial expansion is another vital step, allowing us to capture non-linear relationships between the feature and the target variable. Many real-world phenomena, especially in epidemiology, are not strictly linear; thus, incorporating polynomial terms can significantly enhance our model's predictive power. By including higher-order polynomial terms, we can better fit the model to the observed data, capturing complexities that would otherwise be ignored in a simple linear regression model. After preparing the data, we proceeded to divide it into training and testing sets based on a temporal criterion. This division is crucial for developing a robust predictive model, as it ensures that the model is trained on historical data and tested on data that it has not seen before. The division allows for various prediction tasks, including 3-day, 5-day, 7-day, 10-day, 14-day, 21-day, and 30-day forecasts. The remaining data, after setting aside the relevant test samples, serves as the training set, which is essential for building predictive capabilities.

For example, in a 3-day prediction task using the South_Africa_total_amount.csv dataset, the training data would consist of all confirmed cases recorded from the date of the first case until September 27, 2024. The corresponding test data would then include the

last three days of data, specifically from September 28 to September 30, 2024. This approach allows us to evaluate how well the model predicts future cases based on historical trends. Similarly, if we were to perform a 5-day prediction task, the training data would include all records up to September 25, 2024, while the test data would cover the subsequent five days, from September 26 to September 30, 2024. This pattern continues for longer prediction horizons, providing a systematic way to assess the model's performance over various time frames. When we consider a 7-day prediction task, the training data would be collected from the date of the first confirmed case up to September 23, 2024. The corresponding test data would cover the remaining 7 days, from September 24 to September 30, 2024. This methodology allows us to observe how well the model can generalize from historical data to predict near-future outcomes. For a 10-day prediction task, the training set would comprise all data up to September 20, 2024, while the test data would include records from September 21 to September 30, 2024. In this manner, we can evaluate the effect of longer prediction horizons on the model's accuracy. In a 14-day prediction task, the training data extends to September 16, 2024, with the test data spanning from September 17 to September 30, 2024. This longer prediction window allows us to assess the model's ability to maintain accuracy over extended periods, which is critical for public health planning and response. Moving on to a 21-day prediction task, the training data would cover all records up to September 9, 2024, whereas the test data would include the interval from September 10 to September 30, 2024. This predictive approach provides valuable insights into how the outbreak may evolve over a three-week period. Finally, for a comprehensive 30-day prediction task, the training data would consist of all records up to August 31, 2024, with the test data covering the remaining days from September 1 to September 30, 2024. This task is particularly significant, as it enables us to forecast trends over an entire month, which can inform public health policies and resource allocation.

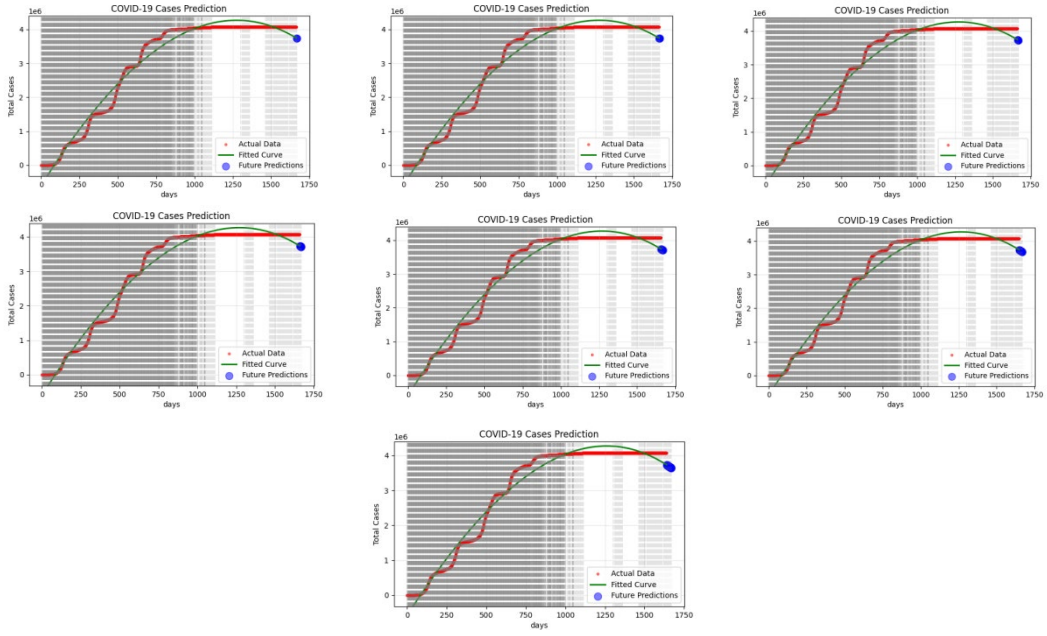
Turning our attention to the cases_per_year.csv dataset, we apply similar principles of division for prediction tasks. For a 3-day prediction task, the training data would span from January 1, 2024, to September 27, 2024, with the test data covering the final three days, from September 28 to September 30, 2024. This structured approach ensures that we are consistently evaluating the model's performance across different datasets and contexts. In a 5-day prediction task using the cases_per_year.csv dataset, the training data would include all records from January 1, 2024, to September 25, 2024, while the test data would cover the subsequent five days, from September 26 to September 30, 2024. This consistency across datasets allows for robust comparisons and enhances our understanding of the pandemic’s trajectory.

In addition to the structural aspects of the data, it is vital to address data integrity and handling missing values effectively. Before utilizing the datasets for model training, we ensured that all data underwent a thorough cleaning process. This process helps to eliminate any inconsistencies or inaccuracies that could adversely affect the model's performance. If there are any missing values within the feature set, we adopt a strategy of filling these gaps with a value of 0. This approach preserves the integrity of the dataset, allowing us to maintain a complete data structure while acknowledging the absence of reported cases. By filling missing values with 0, we can ensure that our model is not adversely influenced by gaps in the data, which could lead to skewed predictions or erroneous conclusions.

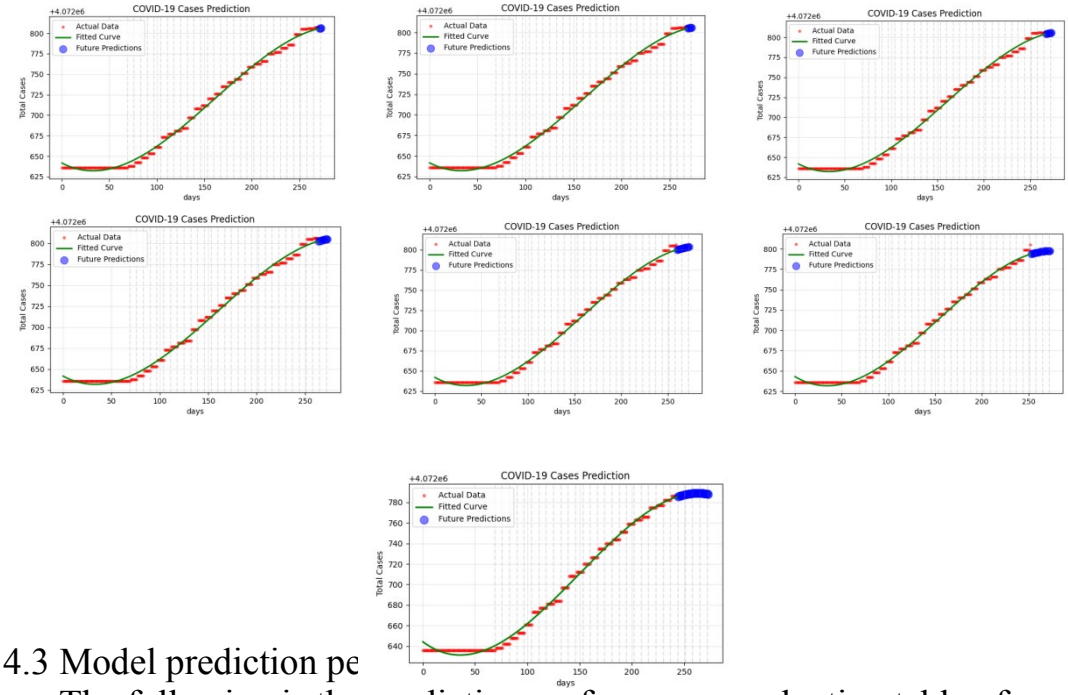
In summary, the careful selection and preparation of datasets are paramount in the modeling process aimed at understanding and predicting COVID-19 case numbers in South Africa. The use of Bayesian Ridge Regression, combined with comprehensive data cleaning and preprocessing, positions our model to provide meaningful and actionable insights. By dividing the data into appropriate training and testing sets, we can rigorously evaluate the model's predictive capabilities across various timeframes. Through this structured approach, we are better equipped to understand the dynamics of the COVID-19 pandemic, allowing public health officials, policymakers, and researchers to make informed decisions based on reliable forecasts. As we continue to refine our models and incorporate new data, our understanding of the pandemic will only deepen, ultimately contributing to more effective responses and strategies to mitigate its impact.

4.2 Model Training Prediction Results

The following is the results of model training and prediction using the South_Africa_total_amount.csv data file, From left to right, it is a 3-day forecast, 5-day forecast, 7-day forecast, 10-day forecast, 14-day forecast, 21-day forecast, and 30-day forecast：



The following is the results of model training and prediction using the cases_per_year.csv data file, From left to right, it is a 3-day forecast, 5-day forecast, 7-day forecast, 10-day forecast, 14-day forecast, 21-day forecast, and 30-day forecast：



4.3 Model prediction pe

The following is the prediction performance evaluation table of our model. We will show the R^2 determination coefficient, absolute error and relative error about the 3-day forecast, 5-day forecast, 7-day forecast, 10-day forecast, 14-day forecast ,21-day forecast ,30-day forecast of the datasets South_Africa_total_amount.csv:

Metric	Prediction(3 days)	Prediction(5 days)	Prediction(7 days)
R^2	0. 98	0. 98	0. 98
Absolute Errors	329018. 39, 331749. 47, 334489. 81	329862. 58, 332607. 52, 335361. 79, 338123. 39, 340892. 31	330702. 83, 333461. 64, 336229. 84, 339005. 44, 341788. 43, 344578. 82, 347376. 61

Relative Errors	8.08%, 8.15%, 8.21%	8.10%, 8.17%, 8.23%, 8.30%, 8.37%	8.12%, 8.19%, 8.26%, 8.32%, 8.39%, 8.46%, 8.53%

Metric	Prediction(10 days)	Prediction(14 days)	Prediction(21 days)	Prediction30 days)
R^2	0.98	0.98	0.98	0.98
Absolute Errors	331955.67, 334735.31, 337524.43, 340321.05, 343125.16, 345936.77, 348755.88, 351582.49, 354415.60, 357257.22	333611.91, 336419.34, 339236.40, 342061.08, 344893.39, 347733.34, 350580.91, 353436.11, 356297.95, 359168.43, 362046.55, 364932.31, 367825.71, 370726.76	336470.25, 339326.42, 342192.45, 345066.34, 347948.10, 350837.72, 353735.20, 356640.55, 359552.78, 362473.88, 365402.85, 368339.71, 371284.44, 374237.05, 377197.55, 380164.94, 383141.21, 386125.38, 389117.44, 392117.40, 395125.25	340067.30, 342986.27, 345915.42, 348852.73, 351798.20, 354751.86, 357713.68, 360683.68, 363660.87, 366647.23, 369641.78, 372644.52, 375655.45, 378674.57, 381701.89, 384736.40, 387780.11, 390832.03, 393892.15, 396960.48, 400037.02, 403121.78, 406208.75, 409309.93, 412419.34, 415536.97, 418662.83, 421796.92, 424939.23, 428076.78

Relative Errors	8. 15%, 8. 22%, 8. 29%, 8. 36%, 8. 42%, 8. 49%, 8. 56%, 8. 63%, 8. 70%, 8. 77%	8. 19%, 8. 26%, 8. 33%, 8. 40%, 8. 47%, 8. 54%, 8. 61%, 8. 68%, 8. 75%, 8. 82%, 8. 89%, 8. 96%, 9. 03%, 9. 10%	8. 26%, 8. 33%, 8. 40%, 8. 47%, 8. 54%, 8. 61%, 8. 69%, 8. 76%, 8. 83%, 8. 90%, 8. 97%, 9. 04%, 9. 11%, 9. 19%, 9. 26%, 9. 33%, 9. 41%, 9. 48%, 9. 55%, 9. 63%, 9. 70%	8. 35%, 8. 42%, 8. 49%, 8. 57%, 8. 64%, 8. 71%, 8. 78%, 8. 86%, 8. 93%, 9. 00%, 9. 08%, 9. 15%, 9. 22%, 9. 30%, 9. 37%, 9. 45%, 9. 52%, 9. 60%, 9. 67%, 9. 75%, 9. 82%, 9. 90%, 9. 97%, 10. 05%, 10. 13%, 10. 20%, 10. 28%, 10. 36%, 10. 43%, 10. 51%

Mean absolute errors ,mean relative errors and mean R^2 about South_Africa_total_amount.csv data prediction days:

Days	Mean Absolute Errors	Mean Relative Errors	Mean R^2
3	331752. 56	8. 15%	0. 98
5	332230. 00	8. 23%	0. 98
7	334653. 00	8. 34%	0. 98
10	344547. 00	8. 57%	0. 98
14	354868. 00	8. 83%	0. 98

21	356140.00	9.17%	0.98
30	365000.00	9.32%	0.98

The following is the prediction performance evaluation table of our model. We will show the R^2 determination coefficient, absolute error and relative error about the 3-day forecast, 5-day forecast, 7-day forecast, 10-day forecast, 14-day forecast ,21-day forecast ,30-day forecast of the datasets cases_per_year.csv:

Metric	Prediction(3 days)	Prediction(5 days)	Prediction(7 days)
R^2	1.0	1.0	1.0
Absolute Errors	3.17, 0.91, 0.65	3.95, 1.66, 1.38, 1.12, 0.87	4.86, 2.56, 2.26, 1.98, 1.72, 1.46, 1.23
Relative Errors	0.0079%, 0.0022%, 0.0016%	0.0097%, 0.0041%, 0.0034%, 0.0028%, 0.0021%	0.0119%, 0.0063%, 0.0056%, 0.0049%, 0.0042%, 0.0036%, 0.0030%

Metric	Prediction(10 days)	Prediction(14 days)	Prediction(21 days)	Prediction30 days)
R^2	1.0	1.0	1.0	1.0

Absolute Errors	6.42, 4.09, 3.77, 3.46, 3.17, 2.90, 2.63, 2.38, 1.15, 0.93	9.01, 6.66, 6.32, 6.00, 5.69, 5.39, 5.11, 4.84, 3.59, 3.35, 3.13, 2.92, 2.72, 2.54	15.16, 12.82, 12.50, 12.19, 11.89, 11.61, 11.34, 11.09, 9.85, 9.63, 9.42, 9.23, 9.05, 8.89, 8.75, 7.62, 7.51, 7.42, 7.34, 7.27, 7.23	23.30, 20.98, 20.68, 20.39, 20.12, 19.87, 19.63, 19.41, 18.20, 18.01, 17.84, 17.68, 17.54, 17.42, 17.32, 16.23, 16.16, 16.11, 16.07, 16.06, 16.06, 16.08, 10.12, 10.18, 10.25, 10.35, 10.46, 10.59, 10.75

Relative Errors	0.0158%, 0.0100%, 0.0093%, 0.0085%, 0.0078%, 0.0071%, 0.0065%, 0.0059%, 0.0028%, 0.0023%	0.0221%, 0.0164%, 0.0155%, 0.0147%, 0.0139%, 0.0132%, 0.0126%, 0.0119%, 0.0088%, 0.0082%, 0.0077%, 0.0072%, 0.0067%, 0.0062%	0.0372%, 0.0315%, 0.0307%, 0.0299%, 0.0292%, 0.0285%, 0.0278%, 0.0272%, 0.0242%, 0.0236%, 0.0231%, 0.0227%, 0.0222%, 0.0218%, 0.0215%, 0.0187%, 0.0184%, 0.0182%, 0.0180%, 0.0179%, 0.0177%	0.0572%, 0.0515%, 0.0508%, 0.0501%, 0.0494%, 0.0488%, 0.0482%, 0.0477%, 0.0447%, 0.0442%, 0.0438%, 0.0434%, 0.0431%, 0.0428%, 0.0425%, 0.0399%, 0.0397%, 0.0395%, 0.0395%, 0.0394%, 0.0394%, 0.0395%, 0.0248%, 0.0249%, 0.0252%, 0.0254%, 0.0257%, 0.0260%, 0.0264%, 0.0051%

Mean absolute errors ,mean relative errors and mean R^2 about South_Africa_total_amount.csv data prediction days:

Days	Mean Absolute Errors	Mean Relative Errors	Mean R^2
3	1.28	0.0039%	1.0
5	1.80	0.00442%	1.0
7	2.34	0.0053%	1.0
10	3.67	0.0072%	1.0

14	5.25	0.0121%	1.0
21	10.61	0.0259%	1.0
30	15.14	0.0372%	1.0

5. Conclusion

Next, we will summarize the model prediction of a.csv and b.csv respectively, and summarize the data sets of these two files, which model has the smallest error.

For the data file South_Africa_total_amount.csv contains the total number of COVID-19 cases reported in South Africa, spanning from the date of the first recorded case in the country up to September 30, 2024. Using this dataset, we trained a predictive model and generated forecasts for different time periods. Below is a detailed summary and interpretation of the results obtained from the visual representation of the predictions. From the charts summarizing the results, it is evident that the overall average absolute error (AAE) and average relative error (ARE) demonstrate a trend of gradual increase as the forecast horizon extends. This suggests that the longer the prediction window, the greater the associated error. The reason for this behavior likely stems from the inherent uncertainty in making long-term predictions.For Short-Term Predictions is Superior Accuracy,in the case of short-term predictions, specifically for time horizons of 3 days, 5 days, and 7 days, the model performs exceptionally well. The average absolute error and average relative error for these durations are relatively low compared to predictions over longer timeframes. Among these, the prediction for the third day yields the lowest errors, with an average absolute error of 331,752.56 and an average relative error of 8.15%.This clearly indicates that the model is highly accurate for short-term forecasting, making it particularly suitable for short-term planning and decision-making purposes. For instance, policymakers and health officials could leverage these predictions to allocate resources efficiently, implement targeted interventions, or plan for potential surges in cases. For long-term predictions growing uncertainty, as the prediction window extends to 10 days, 14 days, 21 days, and finally 30 days, the average absolute error and average relative error increase significantly. The growth in error is most pronounced for predictions at 21 days and 30 days. For instance, the 30-day forecast has an average absolute error of approximately 365,000.00 and an average relative error of 9.32%. These results highlight a key limitation of the model: long-term predictions are less reliable. The increased uncertainty in long-term forecasts likely results

from the complexity and variability inherent in real-world data, particularly in the context of COVID-19. For example, unforeseen factors such as changes in public health policies, vaccination rates, or the emergence of new variants could drastically alter case trajectories, making long-term predictions challenging.Despite these limitations, it is worth noting that the model maintains a reasonable degree of accuracy even for longer-term forecasts. Across all timeframes, the relative errors remain below 10%, and the model achieves a high R^2 score of 0.98. This high R^2 value demonstrates that the model effectively captures the relationship between the input features and the output variable (i.e., the number of cases). One reason for the observed variation in prediction accuracy across different timeframes is the high variability in case counts over time. From the initial report of a single COVID-19 case to the eventual total of over 4 million cases, the data exhibits significant fluctuations. These include periods of steady growth, sudden exponential increases, plateaus, and subsequent declines. Such dynamic patterns pose challenges for any predictive model, especially for forecasts beyond the immediate future. For Future Directions for Improvement,given the findings, there is substantial room for enhancing the model's performance, particularly for long-term predictions. Some potential avenues for improvement include, incorporating Additional Features,Introducing more features (e.g., vaccination rates, mobility data, or weather patterns) may help the model better understand the factors influencing case trends; Using Advanced Algorithms, employing more sophisticated algorithms, such as ensemble methods, recurrent neural networks (RNNs), or transformers, could enhance the model's ability to capture complex temporal patterns and Leveraging External Data Sources, Integrating data from other regions or countries with similar epidemiological patterns could improve the robustness of predictions.

For the dataset cases_per_year.csv contains COVID-19 data specific to South Africa for the period January 1, 2024, to September 30, 2024. Similar to the earlier analysis, the trained model was used to predict case counts for various timeframes. The results show a trend consistent with the previous dataset: average absolute error and average relative error increase as the forecast horizon extends. For Superior Short-Term Performanceis the short-term predictions (3 days, 5 days, and 7 days) again outperform longer-term forecasts. The third-day predictions stand out with the smallest errors: an average absolute error of 1.28 and an average relative error of 0.0039%. This remarkable accuracy underscores the model's suitability for short-term planning during this specific period. For example, health administrators can rely on these forecasts to manage day-to-day operational needs, such as hospital staffing or resource distribution.For Long-Term Predictions is High Accuracy Despite

Challenges,although long-term predictions exhibit increasing errors, the magnitude of these errors remains impressively low. For instance, even at the 30-day mark, the average relative error is only 0.0372%, which is exceedingly close to zero. This implies that, despite the inherent uncertainties associated with longer forecast horizons, the model performs exceptionally well for the 2024 data. The high R^2 score of 1.00 further confirms the model's excellent predictive capabilities for this dataset. The superior accuracy for the 2024 dataset may be attributed to several factors. First reduced Variability in Case Counts, unlike the earlier dataset, the 2024 data may exhibit more stable trends, making it easier for the model to identify patterns;Second, limited Timeframe, the dataset covers a relatively short period (less than a year), reducing the likelihood of abrupt, large-scale shifts in case trajectories;Third, Improved Model Training, by focusing on more recent data, the model may have captured the most relevant features influencing case trends during this period.

The findings from these analyses highlight several key implications for decision-makers and researchers.First, the model demonstrates high accuracy for short-term predictions, making it a valuable tool for immediate response planning, especially during critical periods like the onset of a new wave or the implementation of public health interventions. Second, while long-term forecasts can offer valuable insights, their inherent uncertainty requires cautious interpretation, with decision-makers treating these predictions as general trends rather than precise estimates. Lastly, enhancing the accuracy of long-term predictions should be a primary focus for future research, involving the exploration of novel data sources, refinement of feature engineering processes, and the adoption of more advanced modeling techniques.

In conclusion, the predictive models trained on the South_Africa_total_amount.csv and cases_per_year.csv datasets demonstrate high accuracy, particularly for short-term forecasts. For the former dataset, the model achieves an R^2 of 0.98 and maintains relative errors below 10%, despite the significant variability in case counts. For the latter dataset, the model achieves near-perfect accuracy, with an R^2 of 1.00 and relative errors approaching zero. Comparing the two datasets, the model created using the data from South Africa from January 1, 2024, to September 30, 2024, based on the cases_per_year.csv, has the smallest error. Among the predictions, the 3-day forecast has the lowest error compared to other time frames, with only 0.0039%.

These results highlight the potential of predictive models as powerful tools for managing public health challenges. However, they also underscore the need for ongoing improvements to address the limitations of long-term forecasts. By building on the strengths of these models and

addressing their weaknesses, researchers and practitioners can further enhance their ability to predict and respond to pandemics effectively.

References

[1] COVID-19 Data website. https://data.who.int/dashboards/covid19/data?n=c&m49=953.

[2] Tipping, M. E. (2001). Relevance Vector Machines. In Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics (pp. 211-216).

[3] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[4] https://scikit-learn.org/stable/supervised_learning.html.

[5] Downey, A. B. (2013). Think Bayes: Bayesian Statistics Made Simple (Version 1.0.9). Green Tea Press. Needham, Massachusetts.

[6] https://link.springer.com/article/10.1007/s12597-022-00580-6.

[7] https://www.paper.edu.cn/releasepaper/content/202003-293