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 ElasticNet model

Done by: 2-th year student, group MCS-641
specialty - Computer Sciences and Information Technologies,
educational program: "Informatics"
Wang Xinping
Supervisor: Ievgen Meniailov
Reviewer: Kseniia Bazilevych
Adviser: Stas Kachanov

Kharkiv, 2024

Abstract

This study predicts the dynamics of the COVID-19 epidemic based on the ElasticNet regression model, exploring the relationship between epidemic transmission and control strategies. COVID-19, as a global infectious disease, has had a profound impact on public health, society, the economy, and individual lives since its outbreak at the end of 2019. Through data-driven modeling methods, this research aims to address the limitations of traditional epidemiological models in handling high-dimensional data and dynamic complexity. ElasticNet, combining the strengths of Lasso and Ridge regressions, performs feature selection and prevents overfitting, making it suitable for predicting the complex dynamics of epidemic development.

The main tasks of this study include data collection and preprocessing, training and validation of the ElasticNet model, and forecasting future epidemic developments over multiple time horizons based on historical data. Data from China, including daily new cases, cumulative cases, deaths, and other indicators, were used to build a regression model centered around ElasticNet. Experimental results show that the model performs well across various prediction scenarios, particularly in periods with significant changes in epidemic transmission, where it adapts to provide reliable forecasts.

The main findings of the study are as follows: The ElasticNet model excels at handling high-dimensional data with strong correlations among variables and can automatically select the most influential factors on epidemic transmission. Forecasts for different time horizons (e.g., 3 days, 7 days, and 30 days) indicate that the model's error is lower for short-term predictions, but longer-term forecasts are still crucial for resource allocation and policy decision-making. Additionally, the study emphasizes

the importance of dynamic data updates and model optimization to keep pace with the rapid changes in the epidemic.

However, some limitations of this study remain, such as the variability in data quality and the impact of regional differences on the model's prediction performance. Furthermore, the accuracy of long-term forecasts could be improved. Future research could incorporate multimodal data (such as social behavior, climate factors, etc.) and more efficient machine learning algorithms to enhance prediction accuracy. This study not only provides data support for COVID-19 control measures but also offers theoretical foundations and practical experience for predicting other similar infectious diseases.

Keywords: ElasticNet regression, COVID-19 epidemic prediction, high-dimensional data modeling, machine learning, public health decision-making.

Table of Contents

Abstract ..... **I**

Table of Contents ..... **III**

Table of Figures ..... **V**

1. INTRODUCTION **1**

1.1 Background and significance of the study 1

1.2 The research status 5

2. MAIN CONCEPTS **9**

2.1 Workplan 10

2.2 Tasks 10

2.2.1 Data Collection 10

2.2.2 Model Preparation 11

2.2.3 Model Training and Forecasting 11

2.2.4 Error Analysis 11

2.2.5 Visualization 12

2.2.6 Secondary Study 12

2.2.7 Paper Writing 12

2.3 Milestones 12

2.4 Notations 13

2.4.1 Linear Regression Model 13

2.4.2 Regularization 15

2.4.3 ElasticNet Regression 16

2.4.4 Summary 17

3. EXPERIMENTS **18**

3.1 Dataset 18

3.2 Prediction result 20

3.2.1 Experiment design 20
3.2.2 Code implementation 22
3.2.3 Real data start date: January 4, 2020 24
3.2.4 Real data start date: January 1, 2024 26
4. CONCLUSIONS 29
4.1 Concluding remarks 29
4.2 Recommendations for future work 31
5 REFERENCES 34
6 APPENDIX 37
7. ACKNOWLEDGEMENTS 45

Table of Figures

Figure 1 The selected country is China, code to filter the data 23

Figure 2 Build the ElasticNet model 24

Figure 3 output the prediction results.....24

Figure 4 The experimental results with Real data start date: January 4, 2020 25

Figure 5 The Cumulative cases and predictions over time (the entire period) with Real data start date: January 4, 2020.....26

Figure 6 The Cumulative cases and predictions over time (the last month) with Real data start date: January 4, 2020.....26

Figure 7 The experimental results with Real data start date: January 1, 2024 27

Figure 8 The Cumulative cases and predictions over time (the entire period) with Real data start date: January 1, 2024.....27

Figure 9 The Cumulative cases and predictions over time (the last month) with Real data start date: January 1, 2024.....28

1. INTRODUCTION

1.1 Background and significance of the study

COVID-19 (Coronavirus Disease 2019) is a global infectious disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Since its discovery in late 2019, COVID-19 rapidly developed into a global pandemic, posing a severe challenge to public health systems worldwide and having a profound impact on society, the economy, and individual health. SARS-CoV-2 is a coronavirus belonging to the β-genus of the Coronaviridae family, closely related to SARS-CoV (the severe acute respiratory syndrome virus that emerged in 2003) and MERS-CoV (the Middle East respiratory syndrome virus that emerged in 2012). Its genome is composed of single-stranded positive-sense RNA, which has a high mutation rate and strong transmissibility [1]. SARS-CoV-2 primarily spreads through respiratory droplets, but can also be transmitted via contact with contaminated surfaces or aerosols in the air [2].

COVID-19 exhibits exponential growth in its spread, with the reproduction number (R0) estimated to be between 2 and 4 [3]. This rapid transmission pattern places enormous demands on healthcare resources and public health systems. For example, during peak periods of the pandemic, many countries faced shortages of medical equipment and healthcare personnel. The pandemic also caused far-reaching economic impacts, including supply chain disruptions, rising unemployment, and slower economic growth. The International Monetary Fund (IMF) predicted that COVID-19 would cause a 3% global economic contraction in 2020 [4]. The virus has a relatively long incubation period, typically ranging from 2 to 14 days, which increases the difficulty of controlling its spread. After infection, symptoms can range

from mild to severe, with common symptoms including fever, dry cough, fatigue, and, in severe cases, acute respiratory distress syndrome, organ failure, and other complications [5].

Governments worldwide have adopted various non-pharmaceutical interventions (NPIs) to control the spread of the virus, including: social distancing measures such as lockdowns and public gathering restrictions; travel restrictions such as border closures and reduction in international travel; and personal protective measures such as promoting the wearing of masks, maintaining social distancing, and hand hygiene [6], [7], [8]. While these measures have partially curtailed the spread of the virus, they have also led to negative impacts such as social isolation and mental health issues.

Since the outbreak of the pandemic, the global scientific community has accelerated vaccine development. By 2021, multiple COVID-19 vaccines had been approved and vaccination campaigns were underway worldwide. Vaccines have significantly reduced severe illness and mortality, becoming a key tool in controlling the pandemic. In terms of treatment, although there is no specific antiviral drug, several treatment options, including antiviral drugs, antibody therapies, and corticosteroids, have shown effectiveness in alleviating symptoms. For example, dexamethasone has been shown to be highly effective in severe cases [9].

The motivation for COVID-19 epidemic prediction is to control the spread of the virus, optimize healthcare resource allocation, mitigate economic and social burdens, and respond to viral mutations and vaccine distribution. As of now, the pandemic has caused millions of deaths and infections, while severely impacting the global economy. In this context, epidemic prediction has become a critical task in public health. Accurate epidemic prediction can help identify high-risk regions and

populations in the future, supporting governments and public health agencies in taking effective prevention and control measures. For example, by predicting daily new cases, governments can decide whether to implement lockdowns, social distancing, and other interventions [10]. The outbreak of the epidemic has put tremendous pressure on healthcare systems, including shortages of hospital beds, ventilators, and medical personnel. Predictions can help estimate future resource demands, optimize distribution plans, and prevent resource waste or shortages. Epidemic control measures (such as lockdowns) have significantly impacted economic and social activities. By predicting the development of the epidemic, policies can be formulated to balance public health and economic development, reducing unnecessary economic losses. The continuous mutation of SARS-CoV-2 has raised new demands for public health policies. Prediction models can analyze the transmission patterns of different variants and guide vaccine distribution strategies to maximize public health benefits from vaccination [11].

The goal of COVID-19 epidemic prediction is to use data and modeling techniques to simulate and predict the dynamic process of the epidemic, providing scientific evidence for public health decision-making. Prediction of epidemic trends includes forecasting the number of new confirmed cases, deaths, and recoveries over a future period. These predictions can provide estimates for the peak time, duration, and end of the epidemic [12]; Evaluation of intervention measures simulates the impact of different control measures (such as lockdowns, travel restrictions, and vaccination) on epidemic development, helping decision-makers select the optimal intervention plans. Regional epidemic risk assessment, through analysis of epidemic data from different regions, identifies high-risk areas, supporting resource allocation and policy implementation [8]; Supporting vaccination strategies, based on epidemic prediction results, optimizes the order and distribution of vaccine allocation to ensure

priority coverage in high-risk populations [13]; Long-term epidemic management planning forecasts the long-term trend of the epidemic, providing support for long-term planning of public health systems, including the construction of medical infrastructure and adjustments to public health policies [14].

COVID-19 epidemic prediction faces multiple challenges: inconsistent data quality, where epidemic data may have errors, omissions, and delays. Different countries and regions use different standards for counting confirmed cases and deaths, making it difficult to compare and analyze data [15]; insufficient data dimensions, where many key data (such as personal behavior data, socio-economic indicators) may not have been fully collected, limiting the accuracy of prediction models [16]; dynamic characteristics of epidemic transmission, where transmission patterns are influenced by various factors (such as viral mutations, policy interventions, and public behavior). These dynamic factors increase the complexity of modeling [17]; complex effects of policy interventions, where the impact of different policies may vary by time and region, making it difficult for models to comprehensively quantify their effects [18]; regional differences, where significant variations in healthcare resources, population density, and cultural background make models difficult to generalize and require adjustments for specific cases; the impact of viral mutations, where the transmissibility and pathogenicity of new variants may significantly affect the accuracy of predictions made by existing models; heterogeneous data integration, where epidemic predictions require integrating various types of data (such as epidemiological, environmental, and social media data), but these data come from diverse sources with differing formats and qualities, complicating data processing; real-time prediction demands, where epidemic predictions require rapid responses, but the integration and processing of multimodal data may cause delays in timeliness [19]; public behavior and social psychological impacts, as epidemic transmission is

closely related to public behavior, such as wearing masks and maintaining social distance. However, public behavior is influenced by complex factors such as policies, culture, and psychology, making it difficult to model accurately; ethical and privacy issues, as epidemic prediction requires extensive use of personal data (such as location information and health records), making it a challenge to protect privacy while enabling data sharing and efficient prediction.

Based on ElasticNet, the COVID-19 epidemic dynamic prediction method provides a new approach to epidemic forecasting, combining the strengths of statistics, machine learning, and epidemiology. It demonstrates strong adaptability and predictive capability in handling complex and variable epidemic data. This method is not only suitable for COVID-19 but can also be extended to other similar infectious disease predictions, providing data support and theoretical basis for global public health response strategies.

1.2 The research status

The ElasticNet model is a linear regression model designed to address the variable selection problem in high-dimensional data. It was introduced by Zou and Hastie in 2005, combining the advantages of Lasso regression (L1 regularization) and Ridge regression (L2 regularization) to overcome the limitations of each. Lasso regression performs variable selection, while Ridge regression handles multicollinearity issues. In practice, ElasticNet balances the strengths and weaknesses of both regression methods by adjusting two regularization parameters (L1 and L2), enabling it to select strongly correlated features in high-dimensional data while avoiding overfitting.

The primary motivation for the development of ElasticNet was to address the "curse of dimensionality" often encountered in high-dimensional statistical modeling,

where there are many features but only a few truly influential variables. ElasticNet uses regularization techniques (particularly the combination of L1 and L2) to ensure model simplicity while enhancing predictive accuracy. With the advent of the big data era, ElasticNet has been widely applied across various fields, including gene selection, financial forecasting, image processing, and epidemiology.

Since the COVID-19 pandemic first emerged in late 2019, it has rapidly spread worldwide, presenting unprecedented challenges to public health systems, economies, and societies. The rapid transmission, complex dynamic characteristics, and uncertain future trends of the pandemic have made traditional epidemiological models, such as the SEIR (Susceptible-Exposed-Infected-Recovered) and SIR (Susceptible-Infected-Recovered) models, inadequate for accurately describing and predicting the course of the disease. As a result, increasing numbers of researchers have turned to data-driven methods, particularly machine learning algorithms, to improve the accuracy of pandemic predictions.

The introduction of the ElasticNet regression model allows for handling multi-dimensional pandemic-related data and selecting the most relevant features through feature selection, thereby improving the predictive performance of the model. Unlike traditional models based on fixed parameters or infection rates, the ElasticNet model does not rely on assumed values but instead learns patterns from historical pandemic data. Specifically, ElasticNet can consider a range of factors, including government policies, social behavior, medical resources, and population density, to fully account for the dynamic changes and transmission patterns of the pandemic. Therefore, ElasticNet has shown significant advantages in predicting the COVID-19 pandemic, especially when handling large, complex datasets, as it can automatically select the most important variables, improving prediction accuracy.

Predicting the dynamic process of the COVID-19 pandemic faces multiple challenges, one of which is how to handle complex high-dimensional data. The transmission of the virus is influenced by various factors, including population movement, government interventions, social behavior changes, and medical resource availability, which have complex non-linear relationships. Traditional epidemiological models often assume linear relationships among these factors, making it difficult to adapt to the evolving pandemic data. In contrast, the ElasticNet model uses regularization to automatically select relevant variables in high-dimensional data and effectively address multicollinearity, making it a powerful tool for pandemic prediction.

An important advantage of ElasticNet is its regularization mechanism, which reduces the risk of overfitting. In the context of COVID-19 data, many factors may interact in complex ways and be highly correlated, causing traditional linear regression models to overfit. However, through the combination of L1 and L2 regularization, ElasticNet can effectively reduce the risk of overfitting during feature selection, improving the model's generalization ability. [20]; Additionally, ElasticNet can adaptively adjust model parameters, allowing it to update predictions when new pandemic data arrives. This self-adjusting ability is particularly important for dynamic processes like the COVID-19 pandemic, where new variables and information continually alter the transmission patterns. ElasticNet's adaptability enables it to track real-time changes in the pandemic and adjust predictions accordingly as the model is updated.

Ribeiro et al. (2020) used the ElasticNet regression model to make short-term predictions for the COVID-19 pandemic in Brazil. By incorporating government policy interventions, population movement, medical resources, and social behavior into the model and training it using ElasticNet regression, they found that ElasticNet

could effectively capture key factors influencing pandemic transmission, providing accurate predictions [21]; Yang et al. (2020) applied the ElasticNet model to dynamic predictions of the COVID-19 pandemic in China. By combining ElasticNet with an improved SEIR model, they proposed a new multi-factor modeling approach that considered the impacts of government policies, changes in social behavior, and medical interventions, achieving good prediction results [22]; Global epidemic trend analysis by Flaxman et al. (2020) used the ElasticNet model to model the COVID-19 pandemic across multiple European countries. They found that the ElasticNet model performed well across different countries, automatically adjusting predictions based on the data of each country, providing scientific support for government policymaking [3].

2. MAIN CONCEPTS

With the accumulation of epidemiological data, prediction models based on statistics and machine learning have been widely applied to forecast the spread of COVID-19, case numbers, and healthcare burdens. These methods can capture nonlinear relationships between complex data, improving prediction accuracy and reliability [17], [23]. In epidemic forecasting, traditional linear regression methods are widely used due to their ease of understanding and implementation, but they perform poorly when faced with high-dimensional data and multicollinearity issues. To address these challenges, researchers have adopted improved regression methods, such as Ridge Regression, Lasso Regression, Bayesian Regression, and Least Angle Regression (LAR), each with its unique characteristics and application scenarios.

The outbreak of COVID-19 has triggered a high demand for data-driven research, due to the complexity of virus transmission and the sharp increase in healthcare resource requirements. Data ranging from confirmed cases, death rates, vaccine coverage, to the implementation of non-pharmaceutical interventions provide the foundation for modeling and prediction [14]; By analyzing historical data, researchers can identify key driving factors and provide scientific evidence for policymakers. COVID-19 epidemic forecasting is a crucial part of public health decision-making. Epidemic prediction models aim to simulate the dynamic process of virus spread, forecast future case numbers, healthcare resource needs, and the end time of the epidemic. Accurate epidemic forecasts provide key support for resource allocation and policy adjustments. For example, early prediction models, such as the SIR and SEIR models, used mathematical modeling techniques to predict the trajectory of the epidemic and its peak period, offering valuable references for decision-makers [10].

This paper predicts the dynamic process of the COVID-19 pandemic based on the ElasticNet model. The rapid spread and high volatility of COVID-19 present numerous challenges for traditional epidemiological models in terms of prediction accuracy and practical application. To overcome these limitations, the ElasticNet regression model is employed. This model combines the advantages of Ridge Regression and Lasso Regression, making it effective in maintaining good prediction performance when faced with multicollinearity and high-dimensional data.

2.1 Workplan

The primary objective of this workplan is to develop a regression model based on the ElasticNet algorithm to predict the dynamics of COVID-19 cases. Using data from a chosen country, this project involves data collection, model building, statistical forecasting, and error analysis. The work will be implemented using Python, with a focus on statistical accuracy and model performance evaluation.

2.2 Tasks

2.2.1 Data Collection

1) Choose a target country for analysis.
2) Collect COVID-19 statistics (e.g., total cases) from the pandemic's start to September 30, 2024.
3) Save this data in a file named NameCountry-total_amount.csv.
4) Collect data specifically for the period January 1, 2024, to September 30, 2024, in cases_per_year.csv.

2.2.2 Model Preparation

1) Familiarize yourself with the ElasticNet regression model using resources from scikit-learn.
2) Prepare datasets for training and testing the model.
3) Implement data preprocessing steps, including handling missing data and feature scaling.

2.2.3 Model Training and Forecasting

1) Train the ElasticNet regression model using historical data.
2) Generate forecasts for various scenarios:
   - 3-day forecast: Use data up to September 27, 2024.
   - 5-day forecast: Use data up to September 25, 2024.
   - 7-day forecast: Use data up to September 23, 2024.
   - 10-day forecast: Use data up to September 20, 2024.
   - 14-day forecast: Use data up to September 16, 2024.
   - 21-day forecast: Use data up to September 9, 2024.
   - 30-day forecast: Use data up to August 31, 2024.

2.2.4 Error Analysis

1) Calculate absolute and relative errors for each forecast.
2) Analyze which forecast duration provides the lowest error.

2.2.5 Visualization

1) Create charts for each forecasting scenario using the Matplotlib library.

2) Compare predictions with real data and visualize error trends.

2.2.6 Secondary Study

1) Conduct additional forecasts using data from January 1, 2024.

2) Repeat the forecasting and error analysis process for comparison.

2.2.7 Paper Writing

1) Summarize findings and conclusions regarding the model's accuracy and its dependence on dataset selection.

2) Discuss the scenarios with the lowest error and highlight key insights.

2.3 Milestones

1) Week 1-2: Data Collection. Complete data extraction and preprocessing for the selected country.

2) Week 3-4: Model Preparation. Familiarize yourself with ElasticNet and implement the preprocessing pipeline.

3) Week 5-6: Model Training and Forecasting. Train the ElasticNet regression model and generate forecasts for various durations.

4) Week 7: Error Analysis. Calculate and analyze errors for all forecast scenarios.

5) Week 8: Visualization. Create comprehensive visualizations for all scenarios.

6) Week 9: Secondary Study. Repeat the forecasting process using the alternative dataset (from January 1, 2024).

7) Week 10: Paper Compilation. Write the final report, including methodology, results, and conclusions.

2.4 Notations

2.4.1 Linear Regression Model

We begin with the most basic linear regression model. Suppose we have n samples, and each sample has p features. The goal of linear regression is to predict the target variable y using a linear equation.

For the i-th sample, the linear regression model is:

$$y_i = x_i^T \beta + \text{ò}_i \tag{1}$$

where:

$y_i \in \mathbf{R}$ is the target value of the i-th sample.

$x_i \in \mathbf{R}^p$ is the feature vector of the i-th sample (with p features).

$\beta \in \mathbf{R}^p$ is the regression coefficient vector (the value we want to estimate).

ò_i is the error term (the deviation or noise).

In matrix form, the equation for all samples is:

$$y = X\beta + \grave{o} \tag{2}$$

where:

$y \in \mathbf{R}^n$ is the vector of target variables.

$X \in \mathbf{R}^{n \times p}$ is the feature matrix, where each row represents a sample's features.

$\beta \in \mathbf{R}^p$ is the regression coefficient vector.

Objective Function: Ordinary Least Squares. OLS seeks to find the regression coefficients β by minimizing the error between the predicted values and the true values. The objective function of OLS is the sum of squared residuals (the squared difference between the predicted values and the actual values):

$$J_{OLS}(\beta) = \frac{1}{2n} \left\| y - X\beta \right\|_2^2 \tag{3}$$

where:

y is the true target values.

$X\beta$ is the predicted values from the model.

$\left\| y - X\beta \right\|_2^2$ is the squared Euclidean distance between the predicted and true values.

The goal is to minimize $J_{OLS}(\beta)$ to find the optimal β.

By differentiating the objective function and setting the gradient equal to zero, we obtain the closed-form solution for OLS:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y \tag{4}$$

This formula gives the optimal regression coefficients β under OLS.

2.4.2 Regularization

Linear regression is simple, but it can be prone to overfitting, especially when there is noise in the data or multicollinearity among features. To prevent overfitting, we introduce regularization, which adds a penalty term to the objective function to constrain the complexity of the model.

Regularization involves adding an additional term to the objective function to penalize large coefficients. The most common regularization methods are Lasso Regression and Ridge Regression.

Lasso regression adds an L1 regularization term, which encourages sparsity in the regression coefficients, forcing some of them to be zero and thus performing feature selection. The objective function for Lasso regression is:

$$J_{Lasso}(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{5}$$

where:

$\|\beta\|_1 = \sum_{i=1}^{p}|\beta_i|$ is the L1 norm of the regression coefficients, i.e., the sum of the absolute values of all the coefficients.

λ is the regularization parameter that controls the strength of the L1 regularization term.

Lasso regression encourages some coefficients to become exactly zero, leading to feature selection.

Ridge regression adds an L2 regularization term to the OLS objective function, constraining the size of the regression coefficients. The objective function for Ridge regression is:

$$J_{Ridge}(\beta) = \frac{1}{2n} \| y - X\beta \|_2^2 + \lambda \|\beta\|_2^2 \tag{6}$$

where:

$\|\beta\|_2^2 = \sum_{i=1}^{p} \beta_i^2$ is the L2 norm of the regression coefficients, i.e., the sum of the squares of all the coefficients.

λ is the regularization parameter that controls the strength of the L2 regularization term.

2.4.3 ElasticNet Regression

ElasticNet regression combines both L1 and L2 regularization, allowing the model to both perform feature selection and prevent overfitting by constraining the size of the coefficients.

The objective function for ElasticNet combines both L1 and L2 regularization terms:

$$J_{ElasticNet}(\beta) = \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \frac{\lambda_2}{2} \|\beta\|_2^2 \tag{7}$$

where:

$\frac{1}{2n} \|y - X\beta\|_2^2$ is the least squares error term, which measures the goodness of fit of the model.

$\lambda_1 \|\beta\|_1$ is the L1 regularization term, controlling the sparsity of the coefficients (feature selection).

$\frac{\lambda_2}{2}\|\beta\|_2^2$ is the L2 regularization term, controlling the size of the coefficients (preventing overfitting).

ElasticNet has two regularization parameters, λ_1 and λ_2, which control the strength of the L1 and L2 regularization terms, respectively. In practice, we typically use cross-validation to select the optimal values of these parameters, ensuring that the model performs well on the validation set. The optimization problem for ElasticNet is to minimize the objective function, Numerical optimization methods such as gradient descent or coordinate descent are typically used to solve for the optimal β.

To sum up, ElasticNet regression is a regularization technique that combines the strengths of **Lasso regression** (L1 regularization) and **Ridge regression** (L2 regularization). Its main goal is to prevent overfitting by regularizing the model, especially when dealing with multicollinearity or high-dimensional datasets, while also performing feature selection for irrelevant variables. ElasticNet's regularization strategy makes it an effective choice in many practical problems, particularly when there is strong correlation between features. ElasticNet allows for the best of both Lasso and Ridge, adjusting the parameters λ_1 and λ_2 to achieve an optimal balance between feature selection and overfitting prevention. The optimization goal of the ElasticNet model is to minimize the objective function, which consists of the **loss function** and the **regularization term**.

2.4.4 Summary

Linear Regression (OLS): The simplest regression method, minimizing the sum of squared residuals to estimate the coefficients, yielding the closed-form solution $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$.

Lasso Regression: Adds an L1 regularization term to OLS, encouraging sparsity in the coefficients and enabling feature selection.

Ridge Regression: Adds an L2 regularization term to OLS, preventing the coefficients from becoming too large, yielding the closed-form solution $\widehat{\beta}_{Ridge} = (X^T X + n\lambda I)^{-1} X^T y$

ElasticNet Regression: Combines L1 and L2 regularization, balancing feature selection and coefficient shrinkage, and is particularly useful for high-dimensional data and multicollinearity.

3. EXPERIMENTS

3.1 Dataset

The selected country is China. The data resource used is from the "Prediction of the dynamics of COVID-19 task tutorial" Link2, and the data download address is:

https://srhdpeuwpubsa.blob.core.windows.net/whdh/COVID/WHO-COVID-19-global-daily-data.csv

Daily frequency reporting of new COVID-19 cases and deaths by date reported to WHO. WHO stopped requiring daily basis data reporting since August 2023 and the data presented in this dashboard are weekly frequency. However, a few number of countries continues to report at daily frequency which will be updated in this release and Users can find the data at daily granularity mainly for early years of the pandemic in this statistical release.

The dataset consists of 8 features, listed from left to right as follows: Date_reported, Country_code, Country, WHO_region, New_cases, Cumulative_cases, New_deaths, and Cumulative_deaths.

1) Date_reported: Represents the date. The dates start from January 4, 2020, and are updated daily.

2) Country_code: Represents the country code. Since the selected country is China, the country code is CN.

3) Country: Represents the country. This feature is consistently China.

4) WHO_region: Represents one of the six regions defined by the World Health Organization (WHO), specifically the Western Pacific Region (WPR). This feature is consistently WPR.

5) New_cases: Represents the number of new infections reported daily.

6) Cumulative_cases: Represents the cumulative number of infections up to that date.

7) New_deaths: Represents the number of new deaths reported daily.

8) Cumulative_deaths: Represents the cumulative number of deaths up to that date.

In a separate file cases_per_year.csv, collect statistics of COVID-19 cases from January 1, 2024, to September 30, 2024. Use the same resources listed above.

3.2 Prediction result

3.2.1 Experiment design

Build the model specified in the topic using the resources at scikit-learn. The programming language is Python. For each forecast scenario, create charts using the matplotlib library. Make conclusions about where the model's error is the lowest and on which of the two datasets.

Forecast for 3 days:

1) Real data end date: September 27, 2024.
2) Dates included in the forecast: September 28-30, 2024.
3) Absolute error: The absolute difference between the real data and the model's predicted data for September 30.
4) Relative error: The percentage ratio of the absolute error to the real data for September 30. Relative Error=｜Real Data on September 30−Predicted Data on September 30｜/Real Data on September 30×100%.

Forecast for 5 days:

1) Real data end date: September 25, 2024.
2) Dates included in the forecast: September 26-30, 2024.
3) Absolute error: The absolute difference between the real data and the model's predicted data for September 30.
4) Relative error: The percentage ratio of the absolute error to the real data for September 30.

Forecast for 7 days:

1) Real data end date: September 23, 2024.
2) Dates included in the forecast: September 24-30, 2024.
3) Absolute error: The absolute difference between the real data and the model's predicted data for September 30.
4) Relative error: The percentage ratio of the absolute error to the real data for September 30.

Forecast for 10 days:

1) Real data end date: September 20, 2024.
2) Dates included in the forecast: September 21-30, 2024.
3) Absolute error: The absolute difference between the real data and the model's predicted data for September 30.
4) Relative error: The percentage ratio of the absolute error to the real data for September 30.

Forecast for 14 days:

1) Real data end date: September 16, 2024.
2) Dates included in the forecast: September 17-30, 2024.
3) Absolute error: The absolute difference between the real data and the model's predicted data for September 30.
4) Relative error: The percentage ratio of the absolute error to the real data for September 30.

Forecast for 21 days:

1) Real data end date: September 9, 2024.

2) Dates included in the forecast: September 10-30, 2024.

3) Absolute error: The absolute difference between the real data and the model's predicted data for September 30.

4) Relative error: The percentage ratio of the absolute error to the real data for September 30.

Forecast for 30 days:

1) Real data end date: August 31, 2024.

2) Dates included in the forecast: September 1-30, 2024.

3) Absolute error: The absolute difference between the real data and the model's predicted data for September 30.

4) Relative error: The percentage ratio of the absolute error to the real data for September 30.

3.2.2 Code implementation

The selected country is China, and the code to filter the data is shown in Figure 1. The code to build the ElasticNet model is shown in Figure 2, and the code to output the prediction results is shown in Figure 3.

main.py

```python
#WangXinping
import pandas as pd

# Read data
data = pd.read_csv("WHO-COVID-19-global-daily-data.csv")

# Filter data for required countries (China)
country_name = "China"
country_data = data[data['Country'] == country_name]

# Filter date range: from the first case to September 30, 2024
country_data['Date_reported'] = pd.to_datetime(country_data['Date_reported'])
filtered_data = country_data[country_data['Date_reported'] <= "2024-09-30"]

# Save
filtered_data[['Date_reported', 'New_cases', 'Cumulative_cases']].to_csv(f"{country_name}-total_amount.csv", index=False)
print(f"{country_name}-total_amount.csv  *SAVE!")

```

```python
import pandas as pd
import numpy as np

# Read data
#china_total = pd.read_csv('China-total_amount.csv')
cases_2024 = pd.read_csv('cases_per_year.csv')

# Data preprocessing
#china_total['Date_reported'] = pd.to_datetime(china_total['Date_reported'])
cases_2024['Date_reported'] = pd.to_datetime(cases_2024['Date_reported'])

#china_total = china_total.sort_values(by='Date_reported')
cases_2024 = cases_2024.sort_values(by='Date_reported')
```

```python
from sklearn.linear_model import ElasticNet


# Prepare data features and targets
1 usage
def prepare_data(data, train_end_date, predict_days):
    # Filter training data
    train_data = data[data['Date_reported'] <= train_end_date]
    X = (train_data['Date_reported'] - train_data['Date_reported'].min()).dt.days.values.reshape(-1, 1)
    y = train_data['Cumulative_cases'].values

    # Build prediction time
    last_day = train_data['Date_reported'].max()
    predict_dates = [last_day + pd.Timedelta(days=i) for i in range(1, predict_days + 1)]
    X_pred = (pd.Series(predict_dates) - train_data['Date_reported'].min()).dt.days.values.reshape(-1, 1)

    return X, y, X_pred, predict_dates
```

Figure 1 The selected country is China, code to filter the data

```python
# Establish an ElasticNet Model and make predictions
1 usage
def train_and_predict(data, train_end_date, predict_days):
    X, y, X_pred, predict_dates = prepare_data(data, train_end_date, predict_days)
    model = ElasticNet(alpha=0.1, l1_ratio=0.5)
    model.fit(X, y)
    predictions = model.predict(X_pred)
    predictions = [round(pred) for pred in predictions]  # 四舍五入为整数
    return predict_dates, predictions

#
1 usage
def calculate_daily_errors(data, predict_dates, predictions):
    # Filter between true and predicted values
    true_values = data[data['Date_reported'].isin(predict_dates)]['Cumulative_cases'].values
    daily_absolute_errors = abs(true_values - predictions)
    daily_relative_errors = (daily_absolute_errors / true_values) * 100  # 百分比计算

    # Calculate the daily average error
    daily_relative_errors = np.round(daily_relative_errors, decimals: 5)

    # Calculate the average error
    avg_absolute_error = np.round(daily_absolute_errors.mean(), decimals: 5)
    avg_relative_error = np.round(daily_relative_errors.mean(), decimals: 5)

    return daily_absolute_errors, daily_relative_errors, avg_absolute_error, avg_relative_error
```

Figure 2 Build the ElasticNet model

```python
# Result output
results = []
for days, end_date in zip([3, 5, 7, 10, 14, 21, 30],
                          ['2024-09-27', '2024-09-25', '2024-09-23',
                           '2024-09-20', '2024-09-16', '2024-09-09', '2024-08-31']):
    predict_dates, predictions = train_and_predict(cases_2024, end_date, days)
    daily_absolute_errors, daily_relative_errors, avg_absolute_error, avg_relative_error = calculate_daily_errors(
        cases_2024, predict_dates, predictions
    )
    results.append({
        'Predict the number of days': days,
        'Daily absolute error': daily_absolute_errors,
        'Daily relative error': daily_relative_errors,
        'Average absolute error': avg_absolute_error,
        'Average relative error': avg_relative_error
    })
```

Figure 3 output the prediction results

3.2.3 Real data start date: January 4, 2020

Real data start date: January 4, 2020, The daily absolute error, Relative error, and the average absolute error, average Relative error are shown in Figure 4. The

Cumulative cases and predictions over time (the entire period) are shown in Figure 5. The Cumulative cases and predictions over time (the last month) are shown in Figure 6.

	(2020)PredictionRange													
Date	3days		5days		7days		10days		14days		21days		30days	
	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors
2024/9/30	8575206	8.63%	8613802	8.67%	8651935	8.71%	8708254	8.76%	8781670	8.84%	8905409	8.96%	9055285	9.11%
2024/9/29	8494351	8.55%	8532914	8.59%	8571015	8.62%	8627285	8.68%	8700637	8.75%	8824270	8.88%	8974017	9.03%
2024/9/28	8413497	8.47%	8452026	8.50%	8490094	8.54%	8546315	8.60%	8619604	8.67%	8743130	8.80%	8892748	8.95%
2024/9/27			8371138	8.42%	8409173	8.46%	8465346	8.52%	8538571	8.59%	8661991	8.72%	8811480	8.87%
2024/9/26			8290386	8.34%	8328387	8.38%	8384511	8.44%	8457674	8.51%	8580986	8.63%	8730346	8.78%
2024/9/25					8247611	8.30%	8303686	8.36%	8376785	8.43%	8499991	8.55%	8649222	8.70%
2024/9/24					8166690	8.22%	8222716	8.27%	8295752	8.35%	8418851	8.47%	8567954	8.62%
2024/9/23							8141747	8.19%	8214719	8.27%	8337712	8.39%	8486685	8.54%
2024/9/22							8060777	8.11%	8133686	8.18%	8256572	8.31%	8405417	8.46%
2024/9/21							7979808	8.03%	8052653	8.10%	8175433	8.23%	8324148	8.38%
2024/9/20									7971621	8.02%	8094293	8.14%	8242880	8.29%
2024/9/19									7890588	7.94%	8013154	8.06%	8161611	8.21%
2024/9/18									7809724	7.86%	7932183	7.98%	8080512	8.13%
2024/9/17									7728691	7.78%	7851044	7.90%	7999244	8.05%
2024/9/16											7769904	7.82%	7917975	7.97%
2024/9/15											7688765	7.74%	7836707	7.89%
2024/9/14											7607625	7.66%	7755438	7.80%
2024/9/13											7526486	7.57%	7674170	7.72%
2024/9/12											7445604	7.49%	7593160	7.64%
2024/9/11											7364695	7.41%	7512121	7.56%
2024/9/10											7283555	7.33%	7430853	7.48%
2024/9/9													7349584	7.40%
2024/9/8													7268316	7.31%
2024/9/7													7187052	7.23%
2024/9/6													7106145	7.15%
2024/9/5													7024877	7.07%
2024/9/4													6943909	6.99%
2024/9/3													6862641	6.91%
2024/9/2													6781373	6.82%
2024/9/1													6700104	6.74%
Average	8494351	8.55%	8452053	8.50%	8409272	8.46%	8344045	8.40%	8255170	8.31%	8094364	8.14%	7877532	7.93%

Figure 4 The experimental results with Real data start date: January 4, 2020

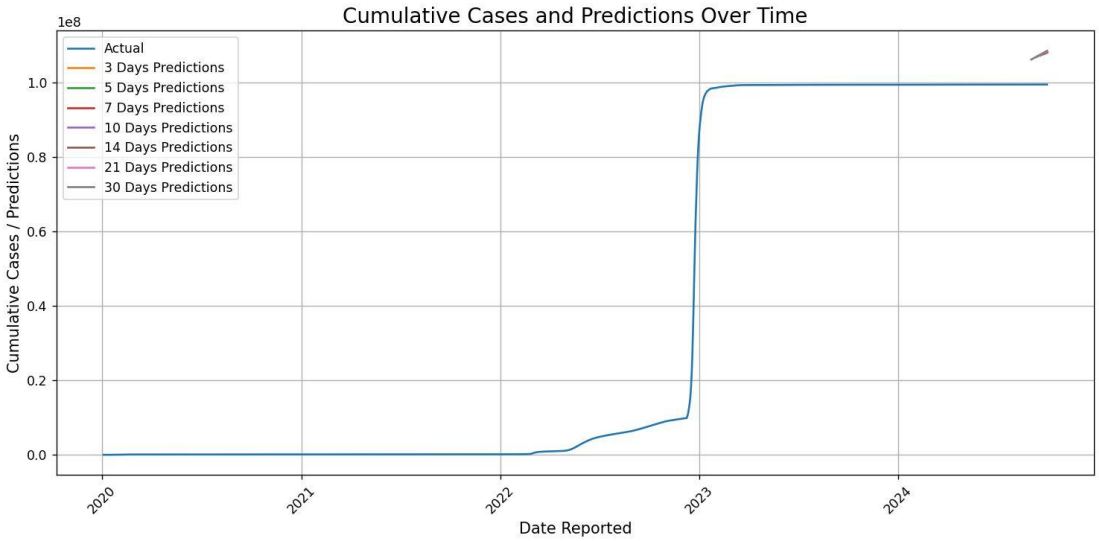

Figure 5 The Cumulative cases and predictions over time (the entire period) with Real data start date: January 4, 2020

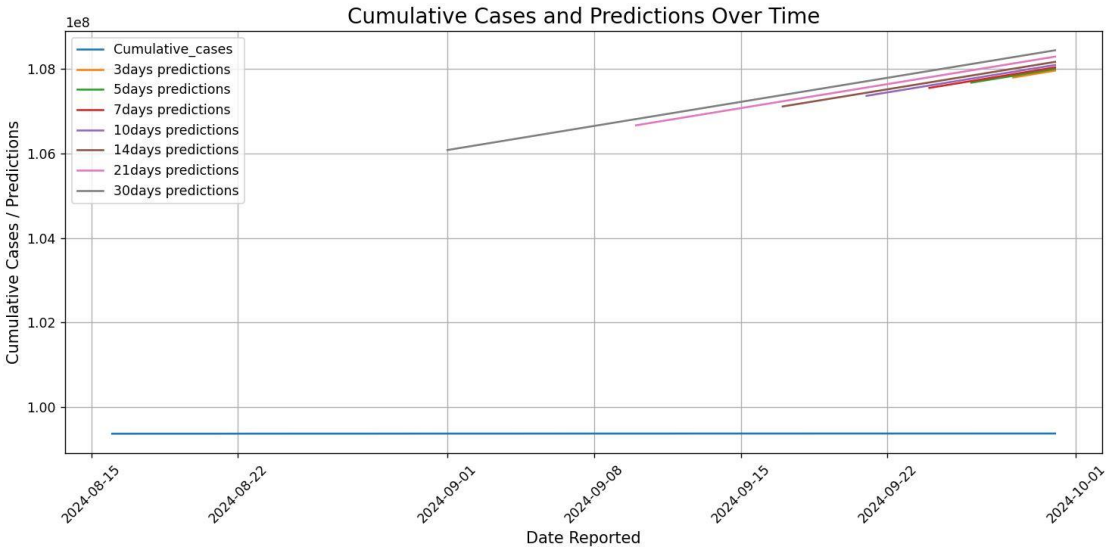

Figure 6 The Cumulative cases and predictions over time (the last month) with Real data start date: January 4, 2020

3.2.4 Real data start date: January 1, 2024

Real data start date: January 1, 2024, The daily absolute error, Relative error, and the average absolute error, average Relative error are shown in Figure 7. The Cumulative cases and predictions over time (the entire period) are shown in Figure 8. The Cumulative cases and predictions over time (the last month) are shown in Figure 9.

In conclusion, the ElasticNet model has the lowest average prediction error on the dataset from the Real data start date: January 1, 2024, to the Real data end date: August 31, 2024, with an average absolute error of 3,731 people and an average relative error of 0.00375%.

	(2024)PredictionRange													
Date	3days		5days		7days		10days		14days		21days		30days	
	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors	Absolute_errors	Relative_errors
2024/9/30	5160	0.00519%	5296	0.00533%	5431	0.00546%	5619	0.00565%	5844	0.00588%	6169	0.00621%	6474	0.00651%
2024/9/29	4934	0.00496%	5069	0.00510%	5203	0.00524%	5390	0.00542%	5614	0.00565%	5938	0.00598%	6240	0.00628%
2024/9/28	4708	0.00474%	4842	0.00487%	4976	0.00501%	5162	0.00519%	5384	0.00542%	5706	0.00574%	6007	0.00604%
2024/9/27			4615	0.00464%	4748	0.00478%	4933	0.00496%	5154	0.00519%	5474	0.00551%	5774	0.00581%
2024/9/26			4523	0.00455%	4655	0.00468%	4839	0.00487%	5060	0.00509%	5378	0.00541%	5676	0.00571%
2024/9/25					4572	0.00460%	4755	0.00478%	4974	0.00501%	5290	0.00532%	5587	0.00562%
2024/9/24					4344	0.00437%	4526	0.00455%	4744	0.00477%	5058	0.00509%	5353	0.00539%
2024/9/23							4297	0.00432%	4514	0.00454%	4827	0.00486%	5120	0.00515%
2024/9/22							4068	0.00409%	4284	0.00431%	4595	0.00462%	4887	0.00492%
2024/9/21							3840	0.00386%	4054	0.00408%	4363	0.00439%	4654	0.00468%
2024/9/20									3824	0.00385%	4132	0.00416%	4420	0.00445%
2024/9/19									3594	0.00362%	3900	0.00392%	4187	0.00421%
2024/9/18									3533	0.00356%	3838	0.00386%	4123	0.00415%
2024/9/17									3303	0.00332%	3606	0.00363%	3890	0.00391%
2024/9/16											3374	0.00340%	3656	0.00368%
2024/9/15											3143	0.00316%	3423	0.00344%
2024/9/14											2911	0.00293%	3190	0.00321%
2024/9/13											2706	0.00272%	2957	0.00298%
2024/9/12											2704	0.00272%	2981	0.00300%
2024/9/11											2679	0.00270%	2978	0.00300%
2024/9/10											2472	0.00249%	2745	0.00276%
2024/9/9													2512	0.00253%
2024/9/8													2278	0.00229%
2024/9/7													2049	0.00206%
2024/9/6													2178	0.00219%
2024/9/5													1945	0.00196%
2024/9/4													2013	0.00203%
2024/9/3													1779	0.00179%
2024/9/2													1546	0.00156%
2024/9/1													1313	0.00132%
Average	4934	0.00496%	4869	0.00490%	4847	0.00488%	4743	0.00477%	4563	0.00459%	4203	0.00423%	3731	0.00375%

Figure 7 The experimental results with Real data start date: January 1, 2024

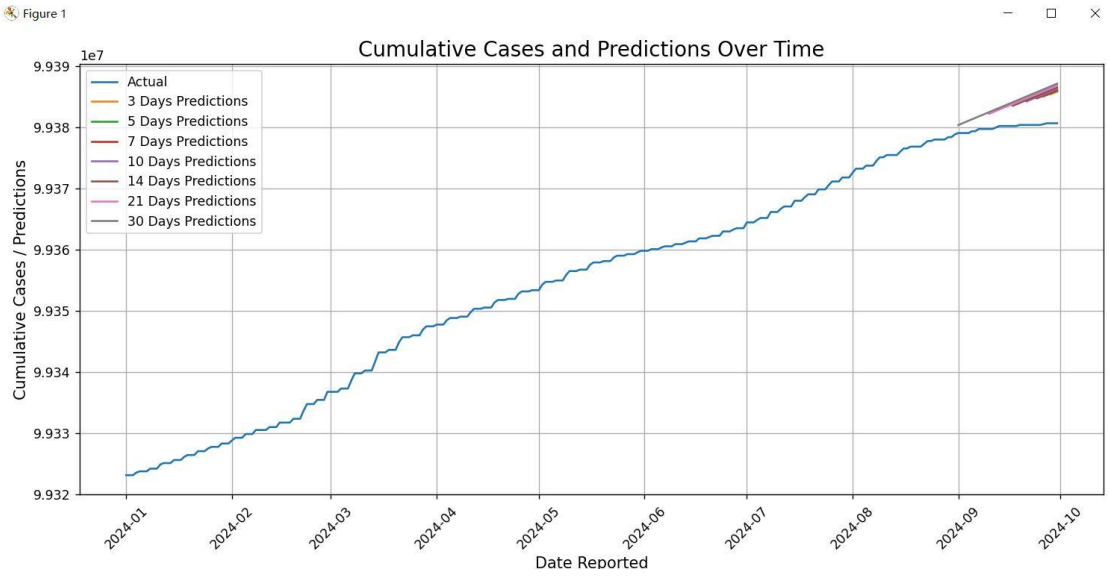

Figure 8 The Cumulative cases and predictions over time (the entire period) with Real data start date: January 1, 2024

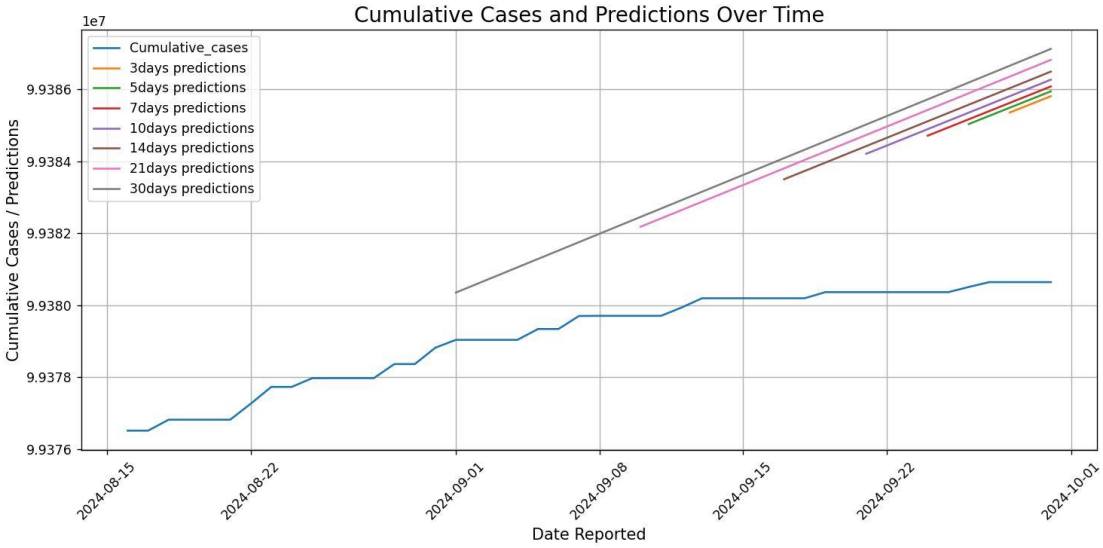

Figure 9 The Cumulative cases and predictions over time (the last month) with Real data start date: January 1, 2024

4. CONCLUSIONS

4.1 Concluding remarks

This study explores the dynamic prediction of the COVID-19 pandemic using the ElasticNet regression model. With the global spread of COVID-19, traditional epidemiological models face significant challenges, mainly due to their difficulty in handling high-dimensional data, complex transmission characteristics, and the impact of policy interventions. Data-driven approaches, particularly machine learning-based modeling techniques, offer a new perspective to address these challenges. The main contribution of this study lies in efficiently modeling and accurately predicting the spread of COVID-19 using the ElasticNet regression model.

The ElasticNet regression model combines the advantages of Lasso and Ridge regression by integrating L1 and L2 regularization. This allows it to effectively handle multicollinearity in high-dimensional data while performing feature selection and avoiding model overfitting. The model not only extracts key variables from historical epidemic data but also enables dynamic prediction. By modeling the COVID-19 data from China, this study conducted predictions over various time horizons (e.g., 3 days, 5 days, 7 days, 10 days, 14 days, 21 days, and 30 days) and evaluated prediction accuracy and error rates.

The experimental results indicate that the ElasticNet regression model demonstrates strong predictive capabilities when handling COVID-19 data. The model accurately reflects the trends of the epidemic in both short-term and long-term forecasts. Short-term predictions provide data support for timely adjustments of epidemic control measures (such as lockdowns and social distancing), while long-term predictions assist governments in making strategic decisions regarding

healthcare resource allocation and vaccine distribution, balancing public health and economic development. This study also found that prediction errors are closely related to data selection and processing methods, and reasonable data preprocessing and feature selection are key factors in improving prediction accuracy. Furthermore, the regularization mechanism of ElasticNet regression effectively prevents overfitting and enhances the model's generalization ability, while the model can adapt to dynamic data changes, continually updating as new data arrives, providing real-time predictions.

This study employed the ElasticNet regression model for dynamic COVID-19 predictions, validating the potential of machine learning methods in epidemic forecasting and providing scientific evidence for public health decision-making. Although the model still has some limitations, future research could further improve prediction accuracy by combining multiple data sources, optimizing the model structure, and enhancing adaptability. With advancements in data science, artificial intelligence, and public health technologies, epidemic forecasting will play an important role in addressing future pandemic challenges. Through global collaboration and technological innovation, we are likely to respond more effectively to future public health crises.

However, there are several limitations in this study. First, the quality and integrity of data are critical factors influencing the model's accuracy. Second, COVID-19 itself is highly uncertain, with its transmission dynamics influenced by various factors such as policy interventions, social behavior, and viral mutations, which are difficult to fully capture using static models. Finally, this study focused on basic epidemic data (such as new and cumulative cases) and did not consider other important factors that may influence epidemic transmission, such as social behavior

and climate changes. Therefore, future research should optimize the model further by incorporating more dimensions of data to improve prediction accuracy.

4.2 Recommendations for future work

Although this study successfully predicted the dynamics of the COVID-19 pandemic using the ElasticNet regression model, addressing the complex challenges of epidemic forecasting requires further research and model optimization. Future research can be improved and expanded in the following areas:

1) Multimodal Data Fusion and Model Optimization

This study primarily relied on basic epidemic statistics (such as new cases and deaths) for prediction. However, in reality, the spread of the epidemic is influenced by a variety of factors, including social behavior, population movement, policy interventions, and weather changes. Future research could integrate multiple types of data, such as social media data, mobility data, and meteorological data, for multimodal data fusion, which would improve prediction accuracy and reliability. By considering different data sources, we can better capture the dynamic characteristics of epidemic transmission.

In terms of model optimization, with the development of machine learning and deep learning technologies, more complex models, such as deep neural networks (DNN)[24], [25] and long short-term memory networks (LSTM)[26], [27], [28], have been proposed. While these methods typically require large datasets and significant computational resources, they have significant advantages in handling nonlinear relationships and time-series data. Future research could explore combining ElasticNet with these advanced machine learning methods to build more complex hybrid models, further improving prediction performance.

2) Enhancing Real-time Data Updates and Model Adaptability

One notable characteristic of the COVID-19 pandemic is its high uncertainty and dynamic nature. The spread of the virus is significantly influenced by factors such as mutations, policy changes, and shifts in public behavior. Therefore, the real-time updating capability of prediction models is crucial. Future research should focus on improving the adaptability of models so they can quickly update and adjust predictions as new data arrives. To achieve this, real-time data processing and online learning technologies will be key areas of focus. Online learning methods based on stream data can effectively handle the challenges of large and rapidly changing datasets, enabling the model to respond quickly to the evolving epidemic.

3) Incorporating the Impact of Policy Interventions and External Factors

This study did not fully account for the timing of policy interventions and regional policy differences in relation to the spread of the epidemic. However, government policies, such as lockdowns, social distancing, and travel restrictions, play a critical role in controlling the epidemic. Therefore, future research should focus more on the quantification of policy factors within the model. For example, a causal model incorporating policy interventions could be constructed to assess the impact of various measures on epidemic progression. Additionally, external factors such as social behavior (e.g., mobility and gathering patterns) and climate change could also influence the spread of the epidemic. These factors should be included in the model to improve prediction accuracy.

4) Cross-Regional and Cross-National Data Sharing and Collaboration

COVID-19 is a global pandemic, and its transmission characteristics vary across countries and regions. To better predict the pandemic's progression worldwide, cross-regional and cross-national data sharing and collaboration are crucial. By

integrating and comparing data from multiple countries, researchers can better understand the commonalities and differences in epidemic transmission, thus providing scientific evidence for more precise prevention strategies in each country. Ensuring data privacy and security during data sharing remains a critical issue that needs to be addressed. Therefore, establishing a global data-sharing platform with appropriate privacy protection policies will be an important area of focus in future epidemic prediction research.

5) Long-term Public Health Management and Resource Planning

The long-term impacts of COVID-19 are not only reflected in current epidemic control but also in the profound effects on global public health systems, economies, and societies. Therefore, future research should focus on integrating epidemic forecasting with long-term public health management. By forecasting the long-term trends of future epidemics, governments can better plan for medical resource reserves, vaccine distribution strategies, and be prepared for potential epidemic fluctuations. Moreover, epidemic forecasting models can also be used to support post-epidemic recovery phases, such as assessing the speed of socioeconomic recovery and rebuilding healthcare systems.

5 REFERENCES

[1] N. Zhu *et al.*, “A novel coronavirus from patients with pneumonia in China, 2019,” *New England Journal of Medicine*, vol. 382, no. 8, pp. 727–733, 2020.

[2] C. Huang *et al.*, “Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China,” *The Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[3] Y. Liu, A. A. Gayle, A. Wilder-Smith, and J. Rocklöv, “The reproductive number of COVID-19 is higher compared to SARS coronavirus,” *Journal of Travel Medicine*, vol. 27, no. 2, p. taaa021, 2020.

[4] IMF, *World Economic Outlook, April 2020: The Great Lockdown*. International Monetary Fund, 2020.

[5] W. J. Guan *et al.*, “Clinical characteristics of coronavirus disease 2019 in China,” *New England Journal of Medicine*, vol. 382, no. 18, pp. 1708–1720, 2020.

[6] N. M. Ferguson *et al.*, “Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand,” 2020, [Online]. Available:
https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/

[7] A. W. H. Chin, J. T. S. Chu, M. R. A. Perera, K. P. Y. Hui, and others., “Stability of SARS-CoV-2 in different environmental conditions,” *The Lancet Microbe*, vol. 1, no. 1, pp. e10–e13, 2020, doi: 10.1016/S2666-5247(20)30003-3.

[8] S. K. Brooks *et al.*, “The psychological impact of quarantine and how to reduce it: rapid review of the evidence,” *The Lancet*, vol. 395, no. 10227, pp. 912–920, 2020.

[9] RECOVERY Collaborative Group, “Dexamethasone in hospitalized patients with Covid-19,” *New England Journal of Medicine*, vol. 384, no. 8, pp. 693–704, 2021.

[10] W. O. Kermack and A. G. McKendrick, “A contribution to the mathematical theory of epidemics,” *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 772, pp. 700–721, 1927.

[11] S. Pei, S. Kandula, and J. Shaman, “Differential effects of intervention timing on COVID-19 spread in the United States,” *Science Advances*, vol. 6, no. 49, p. eabd6370, 2020.

[12] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, and L. dos Santos Coelho, “Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil,” *Chaos, Solitons & Fractals*, vol. 135, p. 109853, 2020.

[13] F. P. Polack *et al.*, “Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine,” *New England Journal of Medicine*, vol. 383, no. 27, pp. 2603–2615, 2020.

[14] J. P. Ioannidis, “Coronavirus disease 2019: The harms of exaggerated information and non-evidence-based measures,” *European Journal of Clinical Investigation*, vol. 50, no. 4, p. e13222, 2020.

[15] I. Holmdahl and C. Buckee, “Wrong but useful—what COVID-19 epidemiologic models can and cannot tell us,” *New England Journal of Medicine*, vol. 383, no. 4, pp. 303–305, 2020.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT Press, 2016.

[17] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[18] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning*. Addison-Wesley, 1989.

[19] IMF, *World Economic Outlook, April 2020: The Great Lockdown*. International Monetary Fund, 2020.

[20] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[21] Z. Yang *et al.*, “Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions,” *Journal of Thoracic Disease*, vol. 12, no. 3, p. 165, 2020.

[22] S. Flaxman *et al.*, “Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe,” *Nature*, vol. 584, no. 7820, pp. 257–261, 2020.

[23] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. Springer, 2013.

[24] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[25] Y. Bengio, “Learning deep architectures for AI,” *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009, doi: 10.1561/2200000006.

[26] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, 2005, doi: 10.1016/j.neunet.2005.06.042.

[28] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” *Advances in neural information processing systems*, vol. 27, pp. 3104–3112, 2014.

6 APPENDIX

```
import pandas as pd

import numpy as np


# Read data

cases_2020 = pd.read_csv('China-total_amount.csv')

# Data preprocessing

cases_2020['Date_reported'] = pd.to_datetime(cases_2020['Date_reported'])

cases_2024 = cases_2020.sort_values(by='Date_reported')

from sklearn.linear_model import ElasticNet


# Prepare data features and targets

def prepare_data(data, train_end_date, predict_days):

    # Filter training data

    train_data = data[data['Date_reported'] <= train_end_date]

    X = (train_data['Date_reported'] -
train_data['Date_reported'].min()).dt.days.values.reshape(-1, 1)

    y = train_data['Cumulative_cases'].values

    # Build prediction time

    last_day = train_data['Date_reported'].max()
```

```
    predict_dates = [last_day + pd.Timedelta(days=i) for i in range(1, predict_days + 1)]
    X_pred = (pd.Series(predict_dates) - train_data['Date_reported'].min()).dt.days.values.reshape(-1, 1)
    return X, y, X_pred, predict_dates
# Establish an ElasticNet Model and make predictions
def train_and_predict(data, train_end_date, predict_days):
    X, y, X_pred, predict_dates = prepare_data(data, train_end_date, predict_days)
    model = ElasticNet(alpha=0.1, l1_ratio=0.5)
    model.fit(X, y)
    predictions = model.predict(X_pred)
    predictions = [round(pred) for pred in predictions]  # Round to the nearest integer.
    return predict_dates, predictions
def calculate_daily_errors(data, predict_dates, predictions):

    # Filter between true and predicted values
    true_values = data[data['Date_reported'].isin(predict_dates)]['Cumulative_cases'].values
    daily_absolute_errors = abs(true_values - predictions)
```

```
    daily_relative_errors = (daily_absolute_errors / true_values) * 100  # Percentage calculation.


    # Calculate the daily average error
    daily_relative_errors = np.round(daily_relative_errors, 5)
    # Calculate the average error
    avg_absolute_error = np.round(daily_absolute_errors.mean(), 5)
    avg_relative_error = np.round(daily_relative_errors.mean(), 5)
    return daily_absolute_errors, daily_relative_errors, avg_absolute_error, avg_relative_error


# Result output
results = []
for days, end_date in zip([3, 5, 7, 10, 14, 21, 30],
                          ['2024-09-27', '2024-09-25', '2024-09-23', '2024-09-20', '2024-09-16', '2024-09-09', '2024-08-31']):
    predict_dates, predictions = train_and_predict(cases_2020, end_date, days)
    daily_absolute_errors, daily_relative_errors, avg_absolute_error, avg_relative_error = calculate_daily_errors(
        cases_2020, predict_dates, predictions
    )
```

```
    results.append({
        'Predict the number of days': days,
        'Predict the number of cases':predictions,
        'Daily absolute error': daily_absolute_errors,
        'Daily relative error': daily_relative_errors,
        'Average absolute error': avg_absolute_error,
        'Average relative error': avg_relative_error
    })
    results_df = pd.DataFrame(results)
    results_df.to_csv(f"output2020.csv", index=False)
    print(f" The prediction results have been saved.")
    # Print preview of the results
    print("\n Preview of prediction results:")
    print(results_df.head())


# Result output
    #for date, abs_err, rel_err in zip(predict_dates, daily_absolute_errors, daily_relative_errors):
```

```
            #print(f"Date: {date}, Asolute error: {abs_err}, Relative error: {rel_err:.5f}%")

#for result in results:

    #print(f"Predict the number of days: {result['Predict the number of days']}")

    #print(f"Average absolute error: {result['Average absolute error']:.5f}")

    #print(f"Average relative error: {result['Average relative error']:.5f}%\n")
```

```
import pandas as pd

import matplotlib.pyplot as plt


file_path = "2020Data.csv"

data = pd.read_csv(file_path)

data['Date_reported'] = pd.to_datetime(data['Date_reported'])

last_month_start = data['Date_reported'].max() - pd.Timedelta(days=45)

last_month_data = data[data['Date_reported'] >= last_month_start]

plt.figure(figsize=(12, 6))


#plt.plot(data['Date_reported'], data['Cumulative_cases'],label='Actual')

#plt.plot(data['Date_reported'], data['3days predictions'], label='3 Days Predictions')

#plt.plot(data['Date_reported'], data['5days predictions'], label='5 Days Predictions')

#plt.plot(data['Date_reported'], data['7days predictions'], label='7 Days Predictions')

#plt.plot(data['Date_reported'], data['10days predictions'], label='10 Days Predictions')

#plt.plot(data['Date_reported'], data['14days predictions'], label='14 Days Predictions')

#plt.plot(data['Date_reported'], data['21days predictions'], label='21 Days Predictions')
```

```
#plt.plot(data['Date_reported'], data['30days predictions'], label='30 Days Predictions')


prediction_columns = ['Cumulative_cases',
    '3days predictions', '5days predictions', '7days predictions',
    '10days predictions', '14days predictions', '21days predictions',
    '30days predictions',]


for col in prediction_columns:
    prediction_values = last_month_data[col].dropna()
    prediction_dates = last_month_data['Date_reported'][:len(prediction_values)]
    plt.plot(prediction_dates, prediction_values, label=col)


plt.title('Cumulative Cases and Predictions Over Time', fontsize=16)
plt.xlabel('Date Reported', fontsize=12)
plt.ylabel('Cumulative Cases / Predictions', fontsize=12)
plt.legend(loc='upper left')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
```

```
plt.show()
```

7. ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank my supervisor, Professor Ievgen Meniailov, for his meticulous guidance and selfless assistance throughout my master's thesis research. You have not only been my academic mentor but also a great friend and teacher in life. During the entire research process, you taught me how to conduct rigorous scientific research and guided me in developing systematic research thinking. Your academic passion, your meticulous attitude towards scholarship, and your attention to detail have deeply inspired me and made me truly realize the responsibility and mission involved in scientific research. I am deeply grateful for the invaluable academic opportunity you provided me, and your teachings will be an invaluable asset for my future research endeavors.

I would like to thank Professor Stas Kachanov, my academic advisor, for providing me with many important suggestions and insights during my research. I am grateful for the precious time and effort you have given me despite your busy schedule, which helped me to continually improve and enhance my research. I know that my research achievements would not have been possible without your support and guidance.

I would also like to thank all my classmates in the MCS-54 group. Your friendship and support have been a great source of warmth throughout my research. In our group discussions, everyone actively shared their perspectives and suggestions, creating a harmonious atmosphere. Especially when I encountered difficulties, everyone was always willing to share their experiences and solutions, which provided me with a great deal of helpful guidance. It was your encouragement and insights that enabled me to overcome one challenge after another and successfully complete my

thesis. Thank you for your support and companionship; your selfless help has given me both warmth and strength.

I would like to express my sincere thanks to all the teachers and students in the faculty. The help and guidance you have provided me throughout my academic journey have been invaluable. The courses taught by each professor have greatly broadened my knowledge and provided crucial theoretical support for my research. In particular, the knowledge of mathematical modeling and machine learning covered in many courses has provided me with essential theoretical tools for better data analysis and model building in my study of COVID-19 epidemic prediction. I extend my heartfelt thanks to all the teachers and classmates who have helped me.

I would like to thank my family for their unwavering understanding, support, and encouragement. You have been the source of strength that drives me forward. During my pursuit of a master's degree, the care and support from my family have always been my strong support. Whether facing research bottlenecks or life pressures, the encouragement and understanding from my family have always given me warmth and strength. Especially during the writing of my thesis, the tolerance and support from my family allowed me to fully dedicate myself to my research. Your selfless dedication and solid support have been the foundation for my academic progress. I especially want to thank my parents for their hard work and limitless support over the years. Your dedication and selflessness have allowed me to pursue my dreams with peace of mind, and my gratitude is beyond words.

Finally, I would like to thank all my friends who have helped me during my research. Your care and support have warmed my heart, and I am grateful for your help and companionship at every stage of my research. Your encouragement and

advice always allow me to see more possibilities and have further strengthened my confidence and determination in pursuing academic endeavors.

In conclusion, I want to express my heartfelt thanks to everyone who has helped and supported me in my academic journey. It is because of your care and support that I have been able to successfully complete my master's thesis. In the future, I will continue to stay true to my original aspirations, explore relentlessly, and strive for excellence in the field of scientific research.