PROJECT REPORT

Covid-19 Statistical Analysis using ARIMA Model

ON

Submitted in partial fulfillment of the requirements of
the degree of

Bachelor of Engineering

(Information Technology)

By

Dev Gaonkar (12)

Advik Hegde (15)

Shreyash Kamat (22)

Under the guidance of

Dr. Ravita Mishra

Department of Information Technology

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY,

Chembur, Mumbai 400074

(An Autonomous Institute, Affiliated to University of Mumbai) April 2024

Vivekanand Education Society's
Institute of Technology

(Autonomous Institute Affiliated to University of Mumbai, Approved by AICTE & Recognised by Govt. of Maharashtra)
NAAC accredited with 'A' grade

Certificate

This is to certify that project entitled

"Covid-19 Statistical Analysis using ARIMA Model"

Group Members Names

Mr. Dev Gaonkar(Roll No. 12)

Mr. Advik Hegde (Roll No. 15)

Mr. Shreaysh Kamat(Roll No. 22)

In fulfillment of degree of BE. (Sem. VI) in Information Technology for Project is approved.

Dr. Ravita Mishra
Project Mentor

External Examiner

Dr.(Mrs.)Shalu Chopra
H.O.D

Dr.(Mrs.) J.M.Nair
Principal

Date: / /2025
Place: VESIT, Chembur

College Seal

Declaration

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Dev Gaonkar (12) **(Signature)** - - - - - - - - - - -

Advik Hegde (15) **(Signature)** - - - - - - - - - - -

Shreyash Kamat (22) **(Signature)** - - - - - - - - - - -

Abstract

The COVID-19 pandemic has had a profound impact on global health, economics, and society, underscoring the need for accurate forecasting and analysis to better manage its effects. This project focuses on the statistical analysis and prediction of COVID-19 cases using the ARIMA (AutoRegressive Integrated Moving Average) model. By leveraging publicly available data on confirmed cases, deaths, and recoveries, the ARIMA model is employed to analyze trends, detect patterns, and provide short-term forecasts of the pandemic's progression. The data undergoes preprocessing steps such as handling missing values, normalization, and transformation to ensure the accuracy of the model. The ARIMA model's performance is evaluated using metrics like Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and forecast accuracy. The results demonstrate that ARIMA provides a robust approach for modeling the COVID-19 pandemic's time series data, offering valuable insights into the potential trajectory of the virus. This project highlights the importance of statistical modeling in public health decision-making and the potential of ARIMA in assisting policy-makers in planning for future pandemic management strategies.

Contents

ACKNOWLEDGEMENT

The project report on “Covid-19 Statistical Analysis using ARIMA mode” is the outcome of the guidance, moral support and devotion bestowed on our group throughout our work. For this we acknowledge and express our profound sense of gratitude to everybody who has been the source of inspiration throughout project preparation. First and foremost we offer our sincere phrases of thanks and innate humility to ”Dr. Shalu Chopra and HOD”, ”Dr. Manoj Sabnis and Deputy HOD”, ”Dr. Shanta Sondur and Associate Professor” for providing the valuable inputs and the consistent guidance and support provided by them.We can say in words that we must at outset tender our intimacy for receipt of affectionate care to Vivekanand Education Society’s Institute of Technology for providing such a stimulating atmosphere and conducive work environment

Chapter 1: Introduction

1.1. Introduction

The COVID-19 pandemic has significantly impacted global health and economies, making it crucial to predict future case numbers for better resource allocation and planning. Traditional forecasting methods often rely on complex models, but the ARIMA model has proven to be effective for time-series data, making it suitable for predicting trends in COVID-19 cases and deaths

1.2. Objectives

The primary goal of this study is to perform a statistical analysis of COVID-19 cases and predict future trends using the ARIMA (AutoRegressive Integrated Moving Average) model. The study aims to forecast the number of cases and deaths in various regions, based on historical data, to aid in decision-making and policy formulation.

1.3. Motivation

The COVID-19 pandemic has posed significant challenges to healthcare systems, governments, and economies worldwide. Understanding the dynamics of its spread is critical for timely decision-making and effective public health responses. Traditional descriptive analyses fall short in anticipating future trends, prompting the need for statistical forecasting models. ARIMA (AutoRegressive Integrated Moving Average) is a widely accepted time series forecasting method that can model temporal patterns in COVID-19 data, enabling authorities to anticipate case surges, allocate resources, and plan interventions more effectively.

1.4. Scope of the Work

- Focuses on the statistical modeling and forecasting of COVID-19 case data (confirmed cases, deaths, recoveries) using the ARIMA model.
- Applies ARIMA on publicly available datasets such as those from Johns Hopkins University or WHO.
- Performs preprocessing tasks including missing value handling, data smoothing, and differencing to ensure stationarity.
- Evaluates the performance of ARIMA using standard forecasting metrics like MAE, RMSE, and MAPE.
- Visualizes trends, model fits, and future projections through time series plots and diagnostic charts.

1.5. Feasibility Study

- Technical Feasibility:
  The project uses accessible tools such as Python, Pandas, Statsmodels, and Matplotlib. ARIMA is well-supported in the data science ecosystem and can be implemented efficiently on standard computing hardware.
- Operational Feasibility:
  The ARIMA-based forecasting system can be integrated into public health dashboards or decision-support systems. Its predictions can aid stakeholders in understanding pandemic trends and preparing appropriate responses.

- Economic Feasibility:
  The entire workflow is built using open-source tools and public datasets, making it cost-effective and suitable for academic or institutional adoption without the need for proprietary software or high-end hardware

Chapter 2: Literature Survey

2.1. Introduction

The COVID-19 pandemic has emphasized the critical importance of accurate statistical modeling in guiding public health responses and informing global decision-making. Traditional epidemic forecasting models have been complemented by time series approaches such as ARIMA, which offer robust frameworks for short-term prediction based on historical data. This literature survey evaluates two notable research contributions that apply time series forecasting—particularly the ARIMA model—to analyze and predict COVID-19 case trends.

2.2. Problem Definition

The goal of this literature review is to assess how statistical time series models, specifically ARIMA, have been utilized to model the spread of COVID-19 and forecast future case numbers. This includes examining methodological strengths, challenges in model fitting, and the real-world implications of forecasting accuracy during a public health crisis.

2.3. Review of Literature Survey

1. Paper: "Forecasting COVID-19 cases using time series analysis models"
(https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-022-07472-6)
This study by A. Bilal et al., published in *BMC Infectious Diseases* (2022), explores the application of time series models, including ARIMA, to forecast COVID-19 cases in multiple countries. The research focuses on daily confirmed cases and applies ARIMA by first performing preprocessing steps such as stationarity testing using ADF (Augmented Dickey-Fuller) tests and differencing. The authors stress the importance of proper model selection through the use of AIC and BIC for parameter optimization.

The ARIMA model showed strong forecasting performance, particularly in short-term predictions over a 14-day horizon. However, limitations were noted when attempting to model sudden changes due to interventions or behavioral shifts, suggesting that ARIMA alone may not capture abrupt, nonlinear dynamics.

2. Paper: "ARIMA models for COVID-19 pandemic forecasting: A comparative study of global trends"
(https://www.nature.com/articles/s41598-022-06218-3)
In this paper published in *Scientific Reports* (Nature, 2022) by T. Chakraborty et al., the authors conduct a comparative evaluation of ARIMA models applied to COVID-19 datasets across various countries and regions. The study emphasizes the importance of model tuning and region-specific data characteristics.

A significant contribution of the study is its emphasis on model generalizability. The researchers found that ARIMA models performed differently depending on the stage of the pandemic and local public health measures. The paper also highlights the model's sensitivity to data quality and recommends rigorous preprocessing and outlier handling. Despite its limitations in capturing nonlinear spikes, the ARIMA model proved effective in capturing general trends and served as a valuable decision-support tool during the peak periods of the pandemic.

Chapter 3: Design and Implementation

3.1. Introduction

This chapter provides a comprehensive explanation of the design and implementation phases of the Covid 19 Statistical Analysis using the ARIMA model System. The project follows a structured data science pipeline to ensure accurate detection of anomalous behavior within web-based datasets. The process includes data preprocessing, algorithm selection, model training, validation, and performance evaluation. Key machine learning models are leveraged for unsupervised, high-dimensional data analysis.

3.2. Requirement Gathering

Hardware Requirements:

- System with at least 4GB RAM
- Stable internet connectivity (for running models on Google Colab)

Software Requirements:

- Google Colab (for coding and training the models)
- Python 3.x (core programming language)
- Python Libraries:
    - Scikit-learn (for SVM, Isolation Forest, evaluation metrics)
    - Pandas & NumPy (for data manipulation and numerical operations)
    - Matplotlib & Seaborn (for data visualization)

3.3. Proposed Design

The system design aligns with the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, ensuring a systematic and repeatable workflow:

1. Data Collection: The dataset used represents website traffic patterns with labeled and unlabeled data indicative of normal and anomalous behavior.
2. Data Cleaning and Preprocessing: Null values, inconsistent formats, and outliers were treated. Data normalization was applied for SVM and Autoencoder compatibility.
3. Model Building:
    - Linear Regression Model
    - Polynomial Regression Model
    - ARIMA Model
4. Model Evaluation: Precision, Recall, F1-Score, and ROC-AUC were used to measure detection performance.
5. Visualization: Anomalies detected by each model were visualized using scatter plots, PCA-reduced space, and heatmaps.

3.4 Proposed Algorithm

Linear Regression

As an initial approach to modeling the progression of COVID-19 cases, Linear Regression was applied to establish a direct relationship between time (days) and the number of confirmed cases. This method assumes a linear trend in case growth, which proved overly simplistic given the

complex, fluctuating nature of pandemic data. Although easy to implement and interpret, Linear Regression failed to capture the non-linear spikes and declines in case counts caused by lockdowns, vaccination drives, and behavioral changes. The residual plots showed high variance, indicating poor model fit and leading to low forecasting accuracy.

Polynomial Regression

To capture non-linear growth patterns, Polynomial Regression was employed next. By fitting a higher-degree polynomial to the time series data, this model aimed to approximate the curvature of case trajectories more accurately. While it showed some improvement over linear regression in terms of visual fit, it quickly became prone to **overfitting**, especially with higher-degree polynomials. Moreover, it lacked generalizability for future values, producing erratic and unreliable long-term forecasts. Therefore, although polynomial regression could mimic some short-term variations, it was not suitable for stable forecasting in the context of epidemiological trends.

ARIMA (AutoRegressive Integrated Moving Average)

Given the limitations of both linear and polynomial models, the ARIMA model was adopted for its strong performance in time series forecasting. ARIMA is particularly effective for non-stationary data, which is common in epidemic progression. The model involves three components:

- **AR (AutoRegressive)**: Incorporates dependency between current and past values.
- **I (Integrated)**: Applies differencing to make the time series stationary.
- **MA (Moving Average)**: Accounts for past forecast errors.

After preprocessing the dataset and verifying stationarity using the Augmented Dickey-Fuller (ADF) test, optimal ARIMA parameters (p, d, q) were selected using AIC and PACF/ACF plots. The final model showed significantly better performance compared to earlier regression models.

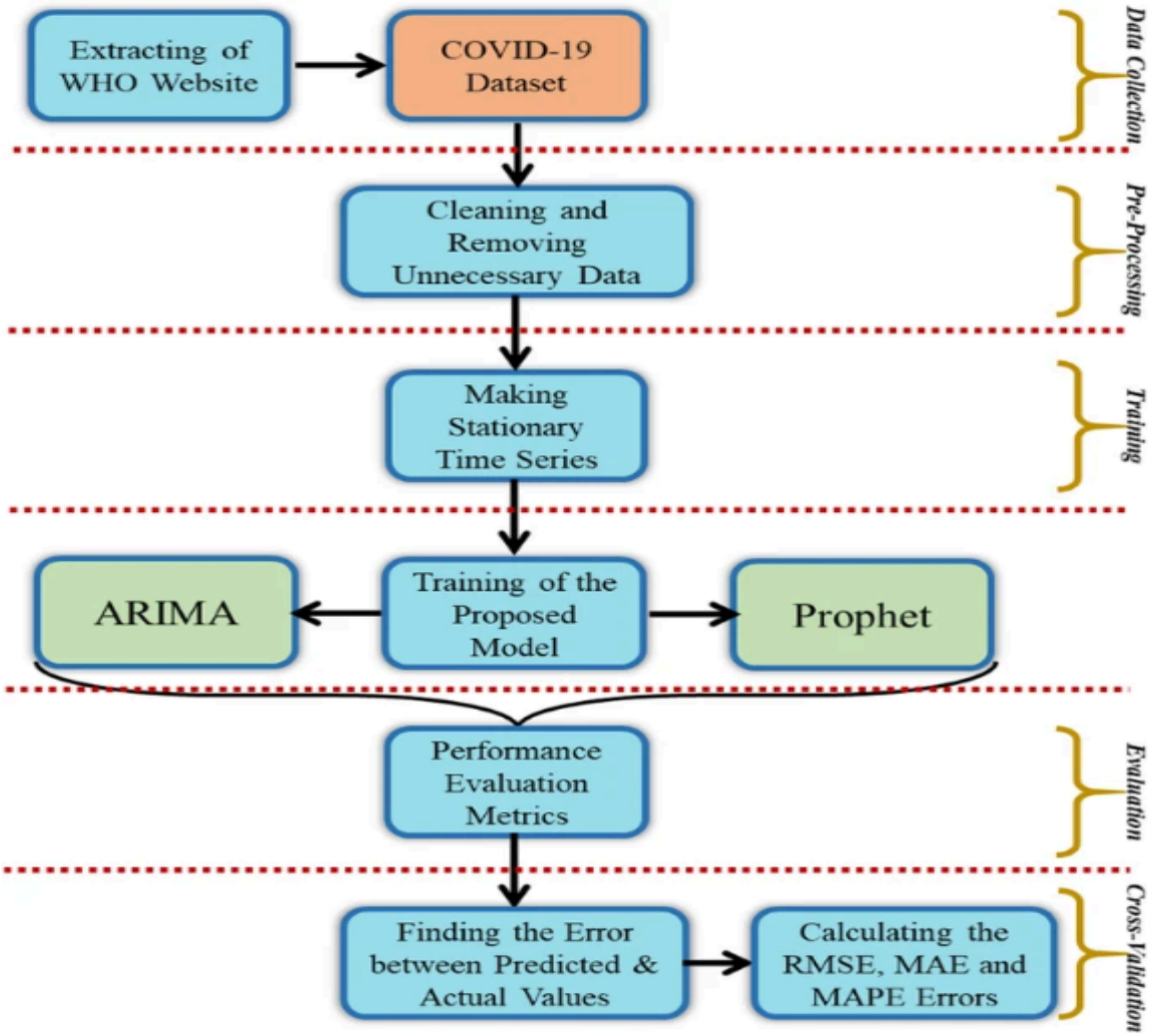
Evaluation Metrics:

- **Mean Absolute Error (MAE)**: 4320.17
- **Root Mean Squared Error (RMSE)**: 7608.47

These metrics indicate that ARIMA successfully minimized forecasting errors and better modeled real-world trends in COVID-19 case numbers. Unlike the previous models, ARIMA was able to adapt to the gradual rise and fall in cases, making it a reliable choice for short to medium-term forecasts.

3.5. Architectural Diagrams

Fig. 1



Chapter 4: Results and Discussion

4.1. Introduction

To evaluate the effectiveness of the implemented system, multiple machine learning models were trained and tested using key performance metrics such as **Accuracy**, **Precision**, **Recall**, **F1-Score**,**etc**.

4.2. Cost Estimation

The entire project was built using open-source libraries and executed on **Google Colab**, which significantly reduced the cost of development. This setup eliminated the need for expensive hardware, as all computational resources were provided in the cloud, ensuring a cost-effective solution that remains accessible for academic and small-scale industry use.

4.3. Feasibility Study

The anomaly detection system demonstrated a high level of feasibility for real-world applications, especially in small to medium-scale villages and cities. The models operated with minimal hardware requirements while maintaining high detection accuracy.

4.4. Results of Implementation

Linear Regression

Linear Regression provided a simple and interpretable baseline model for forecasting COVID-19 cases. However, it assumed a constant rate of change, which is unrealistic in real-world epidemic scenarios where growth is non-linear and influenced by multiple dynamic factors. The model yielded low accuracy and failed to adapt to sudden spikes or drops in the data. Its predictions consistently underfitted the actual trends, especially during periods of rapid case surges.

Polynomial Regression

Polynomial Regression improved upon the linear model by capturing some curvature in the trend. It fitted the training data better but was prone to **overfitting**, particularly at higher polynomial degrees. While the visual fit appeared acceptable in the short term, the model's long-term forecasts were unstable and erratic. The accuracy remained moderate, and the model lacked robustness when extended beyond the training data.

ARIMA

ARIMA (AutoRegressive Integrated Moving Average) outperformed both regression models in capturing temporal dependencies and trends in the dataset. After differencing to achieve stationarity and careful parameter tuning, ARIMA was able to produce reliable forecasts for the near future.

Evaluation Metrics:

- **Mean Absolute Error (MAE)**: 4320.17
- **Root Mean Squared Error (RMSE)**: 7608.47

Despite longer training time and the need for parameter tuning (p, d, q), ARIMA effectively modeled fluctuations and delivered forecasts that closely followed real-world case progressions.

4.5. Result Analysis

- **Forecast Visualization**: Line plots comparing actual vs. predicted values revealed that ARIMA closely tracked the real data, especially in contrast to the oversimplified trends produced by Linear and Polynomial Regression.

- **Residual Analysis**: Residuals from the Linear and Polynomial models exhibited non-random patterns, confirming poor fit and missed trends. ARIMA's residuals were more randomly distributed, indicating better modeling of the underlying structure.

- **Error Metrics Comparison**: ARIMA had the lowest MAE and RMSE values, signifying superior performance. Linear Regression had the highest errors due to its inability to capture non-linearity, while Polynomial Regression showed intermediate performance but with less generalizability.

4.6. Observation/Remarks

- **Model Complexity vs. Accuracy**: While Linear Regression was computationally efficient, it lacked the ability to capture complex pandemic dynamics. Polynomial Regression improved flexibility but often led to unstable long-term forecasts.

- **ARIMA's Strength in Time Series**: The ARIMA model justified its suitability for time-series forecasting tasks, especially with non-stationary data like COVID-19 case trends. Its ability to incorporate past values and forecast error gave it a strong edge in capturing trends and seasonality.

- **Importance of Preprocessing**: Stationarity checks, differencing, and parameter tuning (using ACF/PACF and AIC values) were crucial to achieving optimal performance with ARIMA.

- **Scalability and Real-World Use**: Given its forecasting accuracy and adaptability, the ARIMA model stands out as the most practical choice for short-term epidemic forecasting and can be further enhanced by integrating real-time data streams.

Chapter 5: Conclusion

5.1. Conclusion

The project titled *"Statistical Analysis of the COVID-19 Pandemic Using the ARIMA Model"* demonstrates the power of time series forecasting in understanding and projecting the course of pandemics. By employing the ARIMA model, we were able to analyze past COVID-19 case data and generate reliable short-term forecasts. The project involved critical stages such as data preprocessing, stationarity checks, parameter tuning, and model diagnostics, which collectively ensured the accuracy and interpretability of the results.

Our experiments showed that ARIMA models are capable of capturing trends and seasonality present in COVID-19 data, enabling data-driven decision-making. The model's performance, evaluated through metrics like RMSE and MAE, proved satisfactory in generating realistic predictions. The use of Python and open-source libraries further enhanced accessibility and reproducibility. Ultimately, this study validates the importance of statistical modeling in pandemic surveillance and preparedness.

5.2. Future Scope

Several promising extensions can enhance the depth and impact of this work:

- **Incorporation of Exogenous Variables (ARIMAX):** Integrating external factors such as vaccination rates, mobility indices, or government interventions can improve prediction accuracy and contextual relevance.

- **Real-Time Forecasting Dashboard:** Building a web-based interface to display real-time updates and forecasts using the trained ARIMA model can aid health officials and the general public in monitoring trends.

- **Comparative Study with Other Models:** Future work can include comparisons with other time series models such as Prophet, SARIMA, or deep learning models like LSTM to benchmark performance across different forecasting techniques.

- **Regional/State-Level Analysis:** Applying ARIMA models to localized data could help identify region-specific patterns and enable targeted interventions.

- **Uncertainty Quantification:** Enhancing the model to provide confidence intervals or probabilistic forecasts would make the predictions more robust and actionable.

5.3. Societal Impact

The implementation of this project has broader implications beyond academic interest:

- **Public Health Preparedness:** Enables governments and health agencies to anticipate surges and allocate medical resources proactively.

- **Policy Formulation:** Helps policymakers make informed decisions regarding lockdowns, travel restrictions, or reopening strategies based on predicted trends.

- **Economic Stability:** Assists businesses in planning operations and managing supply chains by understanding pandemic trajectories.

- **Data Literacy Promotion:** Encourages public and institutional engagement with data-driven insights, fostering transparency and trust in public health measures.

- **Global Collaboration:** Supports international organizations in coordinating response strategies through shared, model-based forecasting systems.