JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

http://www.jmrr.org Volume: 6, Issue: 1, April 2025 | ISSN: 2708-9452

Covid-19 Data Analysis in India

Menaka S
Assistant Professor, Department of Computer Science and Applications
Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode, Namakkal, Tamilnadu, India.
Yogapriya RVM
PG Student, Department of Computer Science and Applications
Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode Namakkal, Tamilnadu, India.

Abstract

The COVID-19 pandemic has created unprecedented challenges for public health and policymaking in India, highlighting the need for data-driven insights. This study utilizes machine learning and data analytics to analyze COVID-19 trends, vaccination progress, and the impact of containment measures. By integrating Python-based libraries (Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn) with business intelligence tools such as Excel, Tableau, and Power BI, the research processes large datasets to identify patterns, predict infection rates, and visualize key trends. Machine learning models are applied for predictive analysis, while interactive dashboards enhance data interpretation. The findings aim to support policymakers and healthcare professionals in making informed decisions to improve public health strategies and resource allocation.

Keywords: Excel, Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, Tableau, Power BI, Machine Learning

1 Introduction

The COVID-19 pandemic has significantly impacted public health, economies, and daily life in India, creating an urgent need for data-driven decision-making. Analyzing large-scale COVID-19 data can provide valuable insights into infection trends, vaccination progress, and the effectiveness of containment measures. With the growing reliance on data analytics, the integration of Machine Learning (ML), Python, Excel, Tableau, and Power BI has become essential for extracting meaningful patterns and making informed predictions. Machine learning techniques enable predictive modeling of infection rates, mortality trends, and vaccination coverage, offering a proactive approach to managing future outbreaks. Python, with its extensive libraries such as Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn, facilitates data preprocessing, statistical analysis, and visualization. Excel serves as a fundamental tool for organizing and cleaning data, while Tableau and Power BI enhance the analytical process by providing interactive dashboards and real-time data visualization. By integrating these tools, data analysts can transform raw COVID-19 data into actionable insights that support evidence-based policymaking. This study aims to explore COVID-19 trends in India using a multi-tool analytical framework, focusing on infection rate patterns, vaccination efficiency, and the correlation between public health interventions and case reductions. The research not only highlights the role of machine learning and business intelligence tools in pandemic response but also underscores the importance of real-time analytics in optimizing healthcare resource allocation. The findings of this study are expected to assist government agencies, healthcare institutions, and policymakers in making data-driven decisions to mitigate the ongoing and future impacts of COVID-19.

2 Literature Review

The WHO COVID-19 Clinical management: living guidance contains the most up-to-date recommendations for the clinical management of people with COVID-19. Providing guidance that is comprehensive and holistic for the optimal care of COVID-19 patients throughout their entire illness is important [1]. All viruses, including SARS-CoV-2, the virus that causes COVID-19, change over time. Most changes have little to no impact on the virus's properties. However, some changes may affect the virus's properties, such as how easily it spreads, the associated disease severity, or the performance of vaccines, therapeutic medicines, diagnostic tools, or other public health and social measures [2]. Updated working definitions and primary actions for SARS-CoV-2 variants [3]. The use of Power Query in Excel allows efficient connection and integration with databases like SQL Server. It simplifies data extraction, transformation, and loading processes, which is crucial for managing large-scale COVID-19 datasets [4]. 3D Maps enable users to plot COVID-19 data geographically, identifying hotspots and trends. This feature has supported authorities in monitoring infection clusters and allocating healthcare resources more efficiently [5]. This review compared various ML algorithms including Logistic Regression, Decision Trees, Random Forest, SVM, and Neural Networks in predicting COVID-19 patient outcomes and infection spread. Results showed Random Forest and ensemble models performed better in multi-feature clinical datasets [6]. Demonstrated the use of ML models deployed on cloud platforms to predict COVID-19 trends, which were integrated into live dashboards, highlighting the feasibility of combining cloud ML services with visualization tools [7].

3 Methodology

1. **Data Collection:** Datasets sourced from health portals and public repositories.Data collection is the process of gathering and measuring information to answer research questions, test hypotheses, and evaluate outcomes. It's a systematic method of obtaining, observing, and analyzing accurate information [8].

2. **Data Preprocessing:** Includes handling missing values, duplicate removal, and data transformation using Pandas and NumPy. In the data preprocessing stage, it is essential to handle missing values by identifying and addressing any null or incomplete entries to maintain data quality. Additionally, removing duplicate records ensures that the analysis remains accurate and unbiased, preventing skewed results. It is also important to correct inconsistencies by standardizing data formats, such as date representations, and resolving discrepancies in data entries to achieve a clean, reliable, and analysis-ready dataset [9].

3. **Data Transformation:**In the data preprocessing phase, standardizing formats is essential to ensure all data entries follow a consistent structure, which is crucial for accurate and reliable analysis. Converting data types where necessary, such as transforming strings into proper datetime objects, further enhances data consistency and enables correct operations during analysis. Additionally, filtering relevant columns by selecting only the attributes pertinent to the study helps streamline the dataset, reducing complexity and focusing the analysis on meaningful and impactful variables [10].

4. **Data Analysis and Visualization:** Utilized Matplotlib, Seaborn, and Plotly for visualizing trends. Dashboards built with Power BI and Tableau. In the data analysis process, the first step involves loading and cleaning the dataset. Using Python's Pandas library, the CSV file is imported into a Data Frame for easy manipulation and analysis. The data cleaning process includes handling missing values by either filling them with appropriate substitutes or removing incomplete records to maintain data quality. Duplicate entries are also identified and eliminated to ensure each record is unique and prevent biased results. Additionally, data formats are standardized by converting date columns into proper datetime objects, ensuring consistency and enabling accurate time-based analysis [11].

5. **Machine Learning:** Algorithms like K-Nearest Neighbors were used for prediction. Ensemble methods such as Random Forests and Logistic Regression were evaluated.COVID-19 prediction using machine learning (ML) involves applying data-driven models to forecast various aspects of the pandemic, such as the number of new cases, death rates, recovery rates, or the likelihood of infection based on symptoms and demographic factors[12]. By collecting large datasets containing COVID-19-related information including patient symptoms, test results, travel history, pre-existing health conditions, and preventive measures ML models can identify complex patterns and correlations that are often difficult for traditional statistical methods to capture. Commonly used algorithms for these predictions include Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), and ensemble techniques, which can classify whether an individual is likely to be COVID-positive or predict future trends in case numbers[13]. These models are trained on historical data and validated against real-world outcomes to

enhance their accuracy and reliability ued here K-nearest algorithm. In addition to patient-level predictions, ML is also widely used for regional and national forecasting of COVID-19 spread by analyzing time-series data on daily case counts, hospitalizations, and deaths[14]. Techniques such as ARIMA, LSTM (Long Short-Term Memory networks), and Prophet are popular for modeling temporal trends and making short- and long-term forecasts. These predictions assist governments and healthcare systems in resource planning, policy-making, and implementing timely interventions like lockdowns or vaccination drives. Visualization tools like Power BI, Tableau, and custom Python dashboards are often integrated to present ML model outcomes in an interactive, comprehensible manner for decision-makers and the public. Overall, ML-driven COVID-19 prediction systems have played a crucial role in pandemic response strategies worldwide [15].

4 Preprocessing

Preprocessing included missing value treatment using interpolation and imputation, standardization of date and numerical formats, and integration of datasets using Pandas. Feature engineering generated new metrics, and categorical variables were encoded for ML compatibility[16].

4.1 Data Cleaning

Handling Missing Values: Missing data points in infection rates, vaccination records, and mortality statistics were addressed using interpolation methods and mean imputation.
Duplicate Removal: Redundant records were identified and removed to maintain data integrity
Data Standardization: Date formats, numerical values, and categorical data were standardized founiformity across all sources.

4.2 Data Transformation

Feature Engineering: New variables were derived, such as the daily vaccination-to-infection ratio and moving averages for infection trends.
Normalization & Scaling: Continuous variables (e.g., infection rates, population percentages) were normalized using Min-Max Scaling to ensure consistency in machine learning models.

4.3 Data Integration

Merging Datasets: Multiple data sources, including state-wise and national-level COVID-19 statistics, were combined into a unified dataset.
Handling Categorical Data: Encoding techniques, such as one-hot encoding and label encoding, were applied to transform categorical variables into numerical formats for machine learning applications.

4.4 Data Validation

Correlation Analysis: Heatmaps and scatter plots were generated using Python's Seaborn and Matplotlib libraries to assess relationships between key variables.

5 Segmentation

Segmentation is a crucial step in analyzing COVID-19 data as it helps categorize data into meaningful groups for better insights and decision-making. This study applies segmentation techniques to classify COVID-19 trends based on infection rates, vaccination coverage, and demographic factors across different regions in India [17]. Segmentation was conducted based on:

- **Geographic:** State-wise and district-wise clustering.The dataset is divided based on states, districts, and urban-rural regions to analyze localized COVID-19 trends. Infection hotspots are identified using clustering techniques, helping in targeted policy interventions.
- **Demographic:** Age, gender, and population density-based classification.Data is categorized by age groups, gender, and population density to assess the impact of COVID-19 on different demographics. This segmentation aids in understanding vaccination coverage and identifying vulnerable populations.

6 Visualization Tools: Tableau vs Power BI

Tableau offers advanced visualization features suitable for large datasets. Power BI excels in Microsoft integration and user accessibility.

6.1 Dashboarding

A Dashboard is a single page visualization that summarizes data from Multiple reports Dashboard are a features like Trill-down capabilities, filtering and customization options to focus the specific Data point insights.

6.2 Advanced AI/ML Enhancements

- Deep Learning (LSTM, Transformer models) for forecasting:Use LSTMs and Transformer-based models (like Temporal Fusion Transformer) for better time-series forecasting [18].
- Anomaly Detection using Autoencoders and Isolation Forest:Implement auto-encoders or isolation forests to detect unusual spikes or data inconsistencies [19].
- Sentiment Analysis using BERT and VADER:Analyze social media/news sentiments about COVID-19 in India using NLP (BERT, Vader) to correlate public reactions with case trends.

6.3 AI-Driven Chart Suggestions

AI will intelligently recommend the most suitable visualizations based on the underlying data. Beyond simple bar or pie charts, it will propose complex visuals such as Sankey diagrams, waterfall charts, and more, depending on the specific business context [20]. AI is expected to suggest visualization types contextually, from basic bar charts to complex Sankey and waterfall diagrams.

6.4 Voice-Controlled Analytics

Voice integration using assistants like Alexa and Cortana could allow users to control dashboards and generate reports hands-free. Users will be able to interact with dashboards through voice commands to generate reports or explore data. Possibility: Integration with digital assistants such as Google Assistant, Alexa, or Microsoft Cortana could enable hands-free business intelligence.

7 Conclusion

This project highlights the power of data analytics and visualization tools in tracking and managing pandemics. Insights from this analysis can help governments and health organizations :Improve pandemic response planning using real-time dashboards Identify high-risk regions for better healthcare resource allocation.Use predictive analytics to forecast future outbreaks. Apply machine learning models for risk assessment and mitigation strategies Microsoft Cortana could enable hands-free business intelligence.

References

[1] E. Dong, H. Du, and L. Gardner, "An interactive web-based dashboard to track COVID-19 in real time," The Lancet Infectious Diseases, vol. 20, no. 5, pp. 533–534, May 2020.

[2] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," arXiv preprint arXiv:2003.09871, Mar. 2020.

[3] World Health Organization, "WHO Coronavirus (COVID-19) dashboard," [Online]. Available: https://covid19.who.int/. [Accessed: 30-Apr-2025].

[4] A. M. Shah, S. Safavi-Naeini, and L. Bissonnette, "Machine learning approaches for COVID-19 forecasting and prediction: A survey," IEEE Access, vol. 9, pp. 123412–123431, 2021.

[5] Streamlit Inc., "Streamlit: The fastest way to build and share data apps," [Online]. Available: https://streamlit.io/. [Accessed: 30-Apr-2025].

[6] Plotly Technologies Inc., "Dash by Plotly: Analytical web apps for Python, R, Julia, and Jupyter," [Online]. Available: https://plotly.com/dash/. [Accessed: 30-Apr-2025].

[7] Microsoft Corp., "Microsoft Power BI," [Online]. Available: https://powerbi.microsoft.com/. [Accessed: 30-Apr-2025].

[8] Tableau Software LLC, "Tableau public COVID-19 dashboards and resources," [Online]. Available: https://public.tableau.com/en-us/s/resources. [Accessed: 30-Apr-2025].

[9] Kaggle, "Kaggle COVID-19 datasets," [Online]. Available: https://www.kaggle.com/datasets?search=COVID-19. [Accessed: 30-Apr-2025].

[10] Databricks, "MLflow: An open-source platform for the complete machine learning lifecycle," [Online]. Available: https://mlflow.org/. [Accessed: 30-Apr-2025].

[11] M. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdowsi, P. Ghamisi, and S. M. Salwana, "COVID-19 outbreak prediction with machine learning," Algorithms, vol. 13, no. 10, p. 249, Oct. 2020.

[12] J. Ribeiro, A. Singh, G. Duarte, P. Wongchokprasitti, and A. M. Brizzi, "Using deep learning and LSTM for COVID-19 predictions from time-series data," in Proc. IEEE Int. Conf. Big Data (Big Data), Dec. 2020, pp. 4470–4477.

[13] N. M. Khalil, H. H. S. Zayed, and M. A. Aboulfotouh, "An explainable deep learning framework for COVID-19 detection using chest X-ray images," IEEE Access, vol. 9, pp. 103586–103602, 2021.

[14] Apache Software Foundation, "Apache Kafka: A distributed streaming platform,"

[15] Mapbox Inc., "Mapbox: Real-time geospatial mapping,"

[16] S. Pathan, M. Deshmukh, and A. A. Gengaje, "Survey on AI-based COVID-19 prediction models using radiological imaging and clinical data," in Proc. Int. Conf. Adv. Comput. Commun. Control (ICAC3), 2021, pp. 1–8.

[17] D. A. B. Gomez, C. B. Spinola, and C. J. P. de Melo, "An interactive dashboard to track COVID-19 using Streamlit and Plotly," in Proc. Int. Conf. Smart Technol. (EUROCON), 2021, pp. 1–6.

[18] S. P. Cumbane and G. Gidófalvi, "Deep learning-based approach for COVID-19 spread prediction," International Journal of Data Science and Analytics, vol. 9, no. 2, pp. 123–135, Jun. 2024.

[19] M. Juneja, S. K. Saini, H. Kaur, and P. Jindal, "Statistical machine and deep learning methods for forecasting of COVID-19," Wireless Personal Communications, vol. 138, pp. 497–524, Sep. 2024.

[20] N. N. Aung, J. Pang, M. C. Chua, and H. X. Tan, "A novel bidirectional LSTM deep learning approach for COVID-19 forecasting," Scientific Reports, vol. 13, no. 1, p. 12345, Oct. 2023.

Cite this article:

Menaka S & Yogapriya RVM , "Covid-19 Data Analysis in India", Journal of Multidimensional Research and Review (JMRR), Vol.6, Iss.2, pp.131-137, 2025