IDENTIFYING THE BEST MACHINE LEARNING MODEL FOR FORECASTING THE SPATIOTEMPORAL SPREAD PATTERNS OF COVID-19 INFECTION IN THE KANDY DISTRICT

Madushi K. B. A.[1, 2*], Kahatapitiya K. R. A. L.[1], Bandara K. G. S. U. R.[1] and Dhanasekara D. M. M. L.[3]

[1]Postgraduate Institute of Science, University of Peradeniya, Sri Lanka.
[2]Department of Information Technology, SIBA Campus, Pallekele, Sri Lanka.
[3]Faculty of Science, University of Ruhuna, Sri Lanka.
**ashamadushi4@gmail.com*

The COVID-19 outbreak from China has infected more than 573.8 million people all over the world including more than 6.3 million deaths. In Sri Lanka, there are more than 670,000 cases including 16,770 deaths up to now. Many studies have been conducted on this global epidemic, with a focus on analyzing the spatial and temporal spread of the virus, which is crucial for prevention efforts. This study aims to identify the optimal machine learning model to predict the spatio-temporal spread of COVID-19 in the Kandy district of Sri Lanka. The dataset, collected from the Regional Director of Health Services, Kandy, spans the period from 06-01-2021 to 18-10-2021 and includes confirmed cases reported through Polymerase Chain Reaction (PCR) and Rapid Antigen Tests (RAT) across 23 Ministry of Health (MOH) areas. Four machine learning models Logistic Regression, Linear Regression, Decision Tree, and Random Forest were used to predict the spread of COVID-19 in the region. The models were evaluated using R2 score, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), and Accuracy. Results indicate that the Random Forest model outperformed the others, with the highest R2 score (0.571852) and lowest error rates, making it the most suitable model for predicting the COVID-19 spread. Logistic Regression and Linear Regression exhibited poor performance with negative R2 scores, while the Decision Tree model demonstrated moderate predictive capability. Additionally, ArcGIS software was used to visualize the spatial distribution of COVID-19 cases in the Kandy district, highlighting the Kandy Municipal Council area as the most at-risk and Udadumbara as the least. These visualizations facilitate a better understanding of the spread patterns. In conclusion, the Random Forest model proved to be the most effective for spatio-temporal COVID-19 prediction, which can be applied to future epidemic predictions. The study emphasizes the importance of high-quality, centralized databases for enhancing prediction accuracy. Future work includes the development of a live database system for real-time virus spread predictions in Sri Lanka and the use of geographic information systems (GIS) for comprehensive visual mapping to aid in timely intervention strategies.

***Keywords**: COVID-19, Clustering methods, Data mining, Machine learning, Spatiotemporal distribution.*