Predicting disease outbreak using Data mining and Machine learning Techniques

Chaitanya Ranpara, Kurup Gokul, Tamboli Hayat | Dr. Jenicka S | SCOPE

Introduction

Early prediction of epidemics and pandemics helps officials and governments take preventive action, reduce panic, and save lives. Methods include data analysis, monitoring, and mathematical models. Advanced techniques like machine learning aid in identifying patterns and developing real-time monitoring systems for timely response.

Motivation

Developing a project to predict disease outbreaks is motivated by the goal of proactive public health response. By anticipating outbreaks early, we can save lives, allocate resources effectively, and minimize the impact on communities.

SCOPE of the Project

The scope of this project encompasses developing a predictive model for disease outbreaks based on data analysis and machine learning techniques.

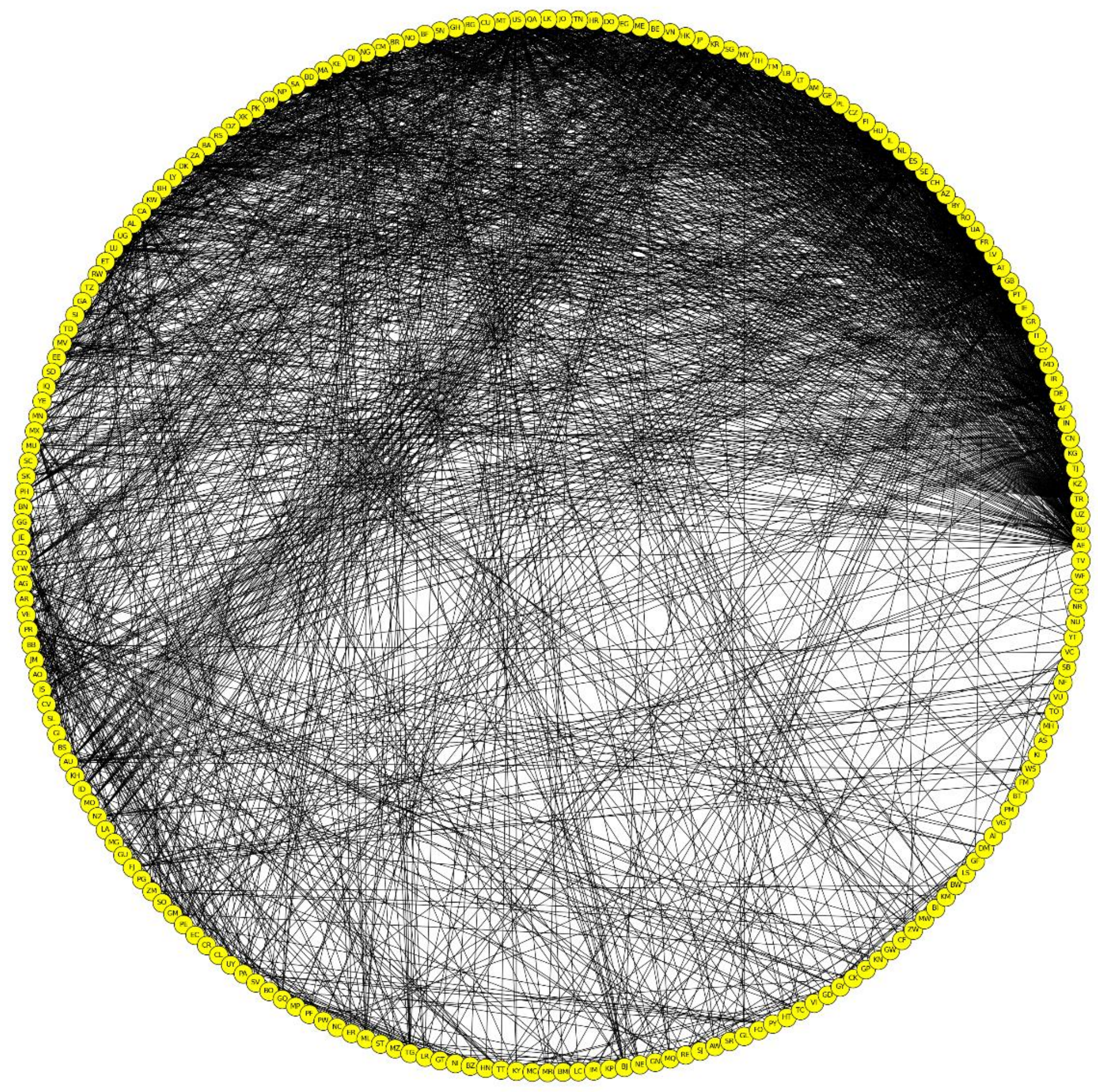
We will be using Data mining and Machine learning algorithms to predict disease outbreaks in several countries.

The scope is limited to the development and evaluation of the predictive model to aid in early detection and response to disease outbreaks.

Methodology

1. **Association Rule Mining using FP Growth** : In this project, we employ Association Rule Mining, specifically the FP Growth algorithm, to predict disease outbreaks. Association rules are If/Then statements that reveal relationships between different datasets. The FP Growth algorithm enhances the efficiency of the Apriori algorithm by organizing data into a tree structure. The four main steps of the FP Growth algorithm include counting individual item occurrences, filtering out non-frequent items, ordering item sets, and creating the tree. For training our model, we utilized a large and diverse dataset encompassing country codes, epidemics, and associated symptoms from 1996 to 2022. The dataset was pre-processed by combining disease and country as strings and generating a 2-D array of transactions. The data was trained using the FP Growth algorithm, with minimum support and confidence set. The model was evaluated by inputting disease-country strings, and the output provided relevant disease-country combinations.

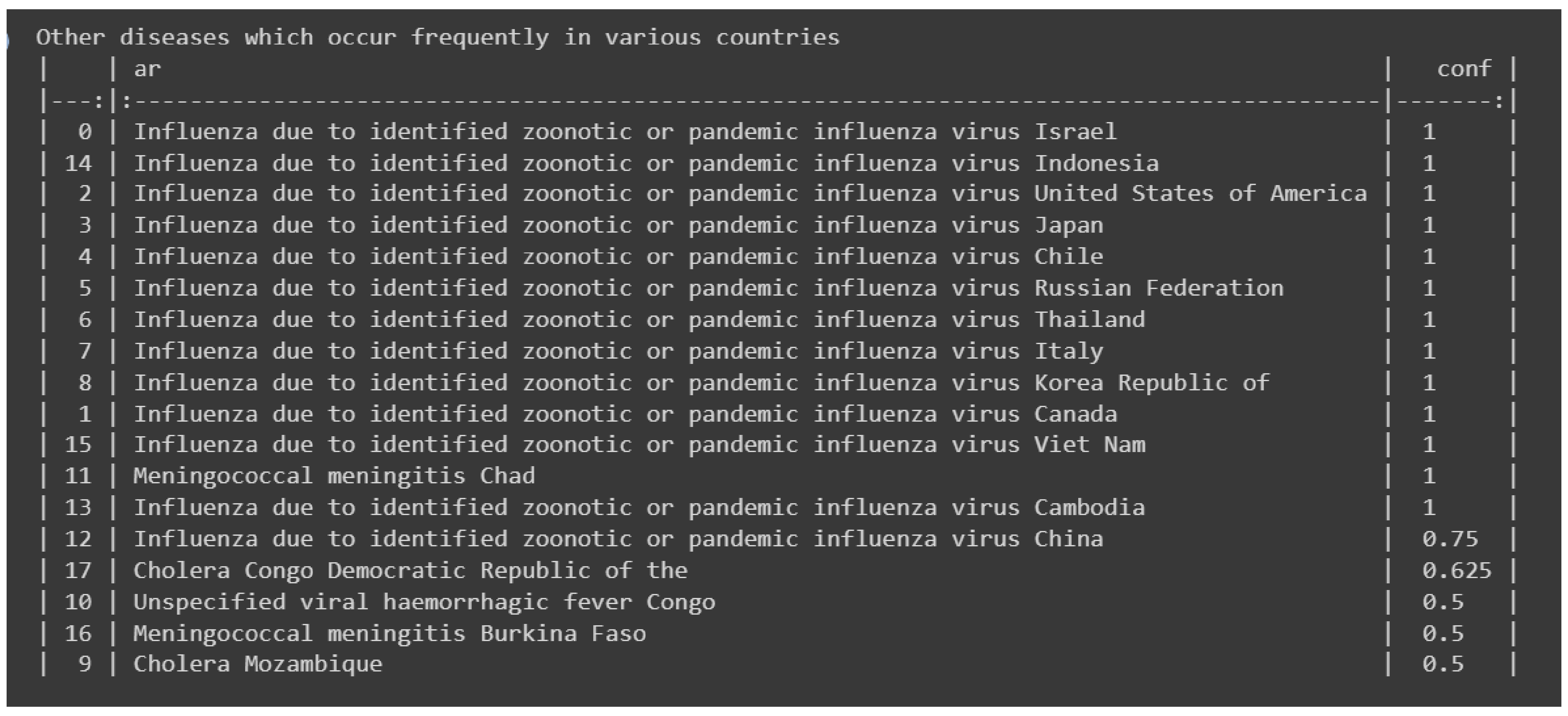
2. **Graph Node Classification using networkx and node2vec** : This approach utilizes data on airline routes between different countries worldwide to predict disease outbreaks. The dataset consists of two categories: diseases and airline routes. The disease dataset is the same as the one used in the Association Rule Mining method. The airline datasets contain information on flight routes and airline details. The data spans the time period of 1996-2022. The collected data is used to create a graph using the NetworkX Python library, which offers tools for analyzing and visualizing complex networks. In this graph, countries are represented as nodes, and airline routes are represented as weighted, unidirectional edges. Each node is associated with the latest disease occurrence in that country. Graph node embeddings are then generated, representing nodes as low-dimensional vector representations while preserving important structural information and relationships. Node embeddings capture the characteristics of nodes and can be used for tasks like node classification and link prediction. The node2vec method, an extension of DeepWalk, is used for graph node embedding. It incorporates biased random walks to balance exploration and exploitation and capture both homophily and structural equivalence in the resulting embeddings.



The graph represents the interconnectedness of countries through international air travel, with each country being a node and the air travel connections between them being the edges. The edges have weights which are the distances between these countries in kilometres.

Results

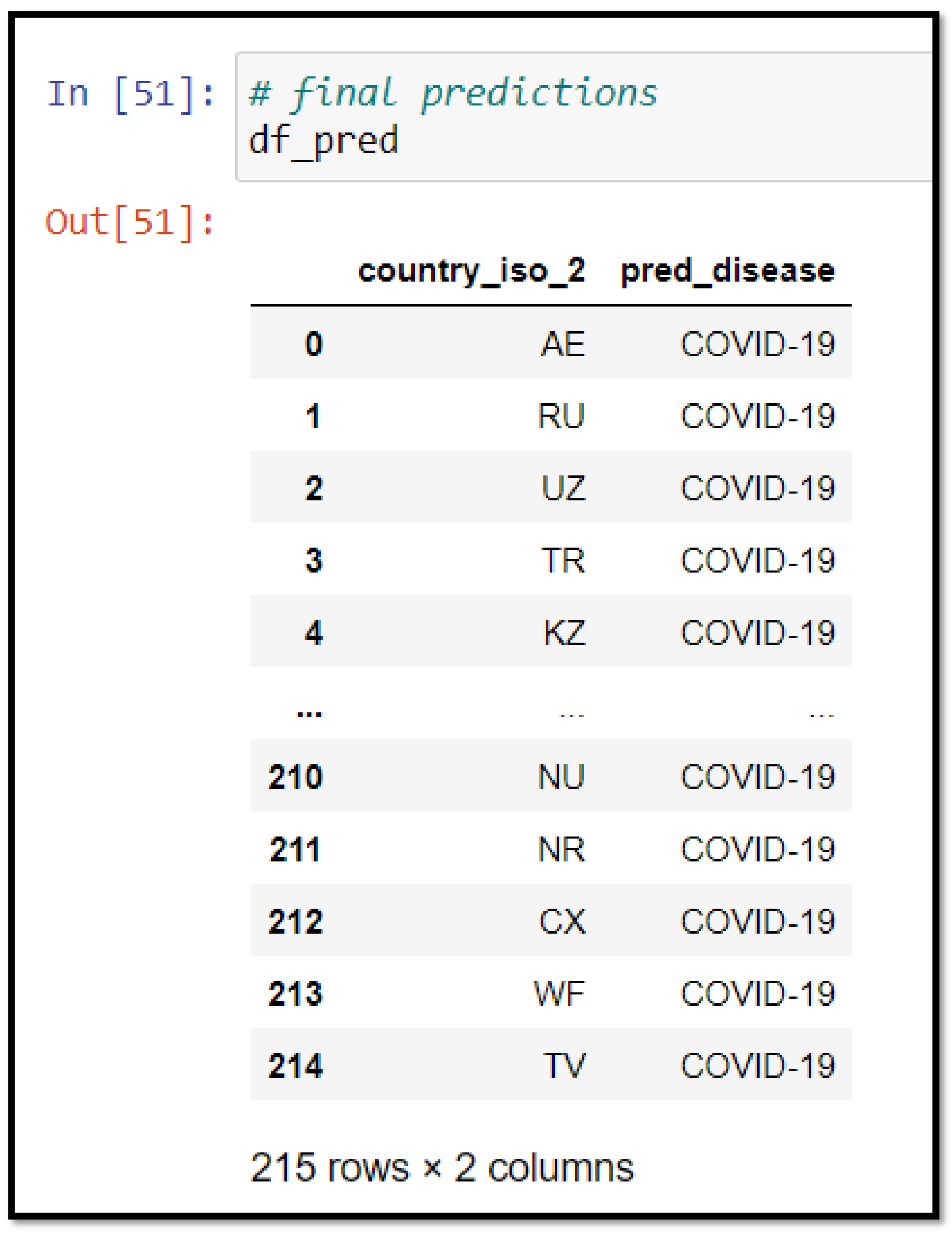
Prediction results from the Association rule mining method



```
Other diseases which occur frequently in various countries
|    | ar                                                                                        |   conf |
|---:|:------------------------------------------------------------------------------------------|-------:|
|  0 | Influenza due to identified zoonotic or pandemic influenza virus Israel                   | 1      |
| 14 | Influenza due to identified zoonotic or pandemic influenza virus Indonesia                | 1      |
|  2 | Influenza due to identified zoonotic or pandemic influenza virus United States of America | 1      |
|  3 | Influenza due to identified zoonotic or pandemic influenza virus Japan                    | 1      |
|  4 | Influenza due to identified zoonotic or pandemic influenza virus Chile                    | 1      |
|  5 | Influenza due to identified zoonotic or pandemic influenza virus Russian Federation       | 1      |
|  6 | Influenza due to identified zoonotic or pandemic influenza virus Thailand                 | 1      |
|  7 | Influenza due to identified zoonotic or pandemic influenza virus Italy                    | 1      |
|  8 | Influenza due to identified zoonotic or pandemic influenza virus Korea Republic of        | 1      |
|  1 | Influenza due to identified zoonotic or pandemic influenza virus Canada                   | 1      |
| 15 | Influenza due to identified zoonotic or pandemic influenza virus Viet Nam                 | 1      |
| 11 | Meningococcal meningitis Chad                                                             | 1      |
| 13 | Influenza due to identified zoonotic or pandemic influenza virus Cambodia                 | 1      |
| 12 | Influenza due to identified zoonotic or pandemic influenza virus China                    | 0.75   |
| 17 | Cholera Congo Democratic Republic of the                                                  | 0.625  |
| 10 | Unspecified viral haemorrhagic fever Congo                                                | 0.5    |
| 16 | Meningococcal meningitis Burkina Faso                                                     | 0.5    |
|  9 | Cholera Mozambique                                                                        | 0.5    |
```

Prediction of disease outbreaks which occur frequently with “Influenza due to identified zoonotic or pandemic influenza virus in Hong Kong” sorted by their confidence

Prediction results from the Graph node classification method



```
In [51]:  # final predictions
          df_pred

Out[51]:
               country_iso_2   pred_disease
           0              AE       COVID-19
           1              RU       COVID-19
           2              UZ       COVID-19
           3              TR       COVID-19
           4              KZ       COVID-19
          ...            ...            ...
         210              NU       COVID-19
         211              NR       COVID-19
         212              CX       COVID-19
         213              WF       COVID-19
         214              TV       COVID-19

         215 rows × 2 columns
```

Accuracy for this method when using data till 2022 is 92.5%

Prediction of disease outbreaks in every country in 2023 according to the latest disease epidemics occurred in all countries using graph node classification

Conclusion

Epidemics have had a significant impact on human existence throughout history, altering communities and leaving enduring legacies. These epidemics, ranging from historical plagues to contemporary outbreaks, have had an impact on the entire world and changed the path of human history. The ongoing COVID-19 outbreak, which is affecting the world on an unprecedented scale, is one of the most notable and recent occurrences.

With the use of these methods, we may learn more about how illnesses spread, predict their impacts, and create effective mitigation plans. For decision-makers, public health professionals, and researchers, the outcomes and forecasts produced by these methodologies can be quite useful. Comparing the outcomes and forecasts from other approaches might offer a wider viewpoint and support the conclusions. It enables us to spot recurring trends, confirm the correctness of our forecasts, and spot any ambiguities or anomalies that need more research.

Additionally, combining the results of many methodologies can produce even better outcomes. We may capitalise on each method's advantages and make up for its weaknesses by using the result of one as the input for another.

References

1. S. Raizada, S. Mala and A. Shankar, "Vector Borne Disease Outbreak Prediction by Machine Learning," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 213-218, doi: 10.1109/ICSTCEE49637.2020.9277286.
2. Khakharia, A., Shah, V., Jain, S. *et al.* Outbreak Prediction of COVID-19 for Dense and Populated Countries Using Machine Learning. *Ann. Data. Sci.* **8** , 1–19 (2021). https://doi.org/10.1007/s40745-020-00314-9
3. Salim, N.A.M., Wah, Y.B., Reeves, C. *et al.* Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques. *Sci Rep* **11**, 939 (2021). https://doi.org/10.1038/s41598-020-79193-2