**A Machine Learning Approach for Covid-19 Trend Analysis**

**Introduction**

As the Covid-19 outbreaks rapidly all around the globe daily and additionally influences the lives of millions, some of the countries declared complete lockdown to check its intensity. The Covid-19 outbreak was first reported in the Wuhan, China on 31st December 2019 when it was spreading rapidly all over the world (Tuli et al., 2020). On 11th March, 2020, WHO finally announced the Covid-19 outbreak as a pandemic observing its continuous intensity to spread. This Study presents a novel approach to analyze the global trend of the Covid-19 Pandemic. The several visualizations represent different findings during the data analysis. Finally, a machine learning model has been trained on the time series data of total Confirmed cases and the model achieved significant accuracy on the validation data.

**Literature Review**

A Covid-19 trend prediction model was introduced by Arora et al. (2020) for predicting the number of Covid-19 positive cases in different states of India. The authors focused on LSTM based prediction model and they tested different LSTM variants on the time-series data, and based on the absolute error the Bi-LSTM model achieved more accurate results over other LSTM models. A machine learning and cloud computing-based Covid-19 prediction model has been developed by Tuli et al. (2020) to predict the future trend of this pandemic. The authors used probabilistic distribution functions to predict the trend. Another study presents an approach for forecasting the trend of the Covid-19 pandemic in Canada using the LSTM model. The model has been trained on the Covid-19 dataset of Jhon Hopkins University and obtained 93.40% & 92.67% accuracy for short-term & long-term predictions respectively (Chimmula & Zhang, 2020).

**Data Analysis**

Initially, the Covid-19 dataset has been downloaded from Kaggle developed by Jhon Hopkins University. The dataset contains the date and county-wise confirmed, recovered, and death cases starting from 22nd January 2020 to 29th May 2021 throughout the world. The partial snapshot of the dataset is presented in Figure 1.

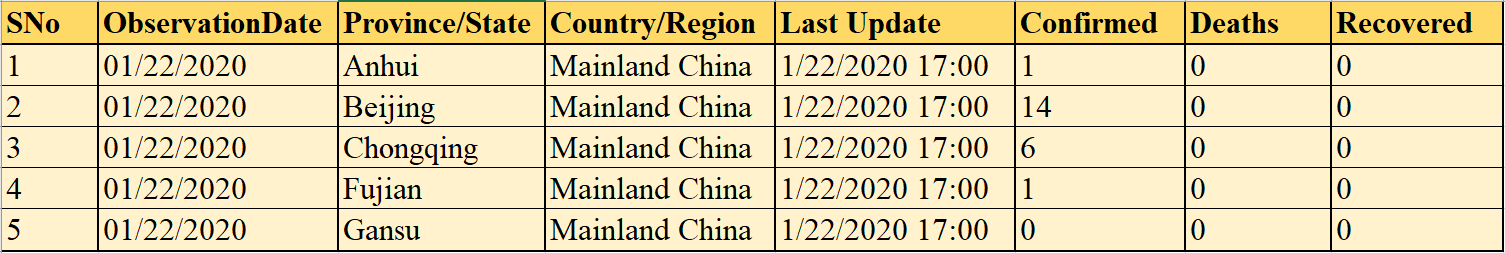


Fig. 1. The snapshot of the Covid-19 dataset.

In the data pre-processing part at first, the unnecessary columns have been dropped and the remaining columns have been renamed with an appropriate name. Then the date format has been changed to pandas DateTime format for more usability of the dataset. After pre-processing the country-wise overall trend in terms of daily confirmed, recovered and death cases has been extracted and the trend for each of the countries has been plotted using a line plot graph. Figure 2. represents the line plot graphs of the overall trend for some mostly affected countries.

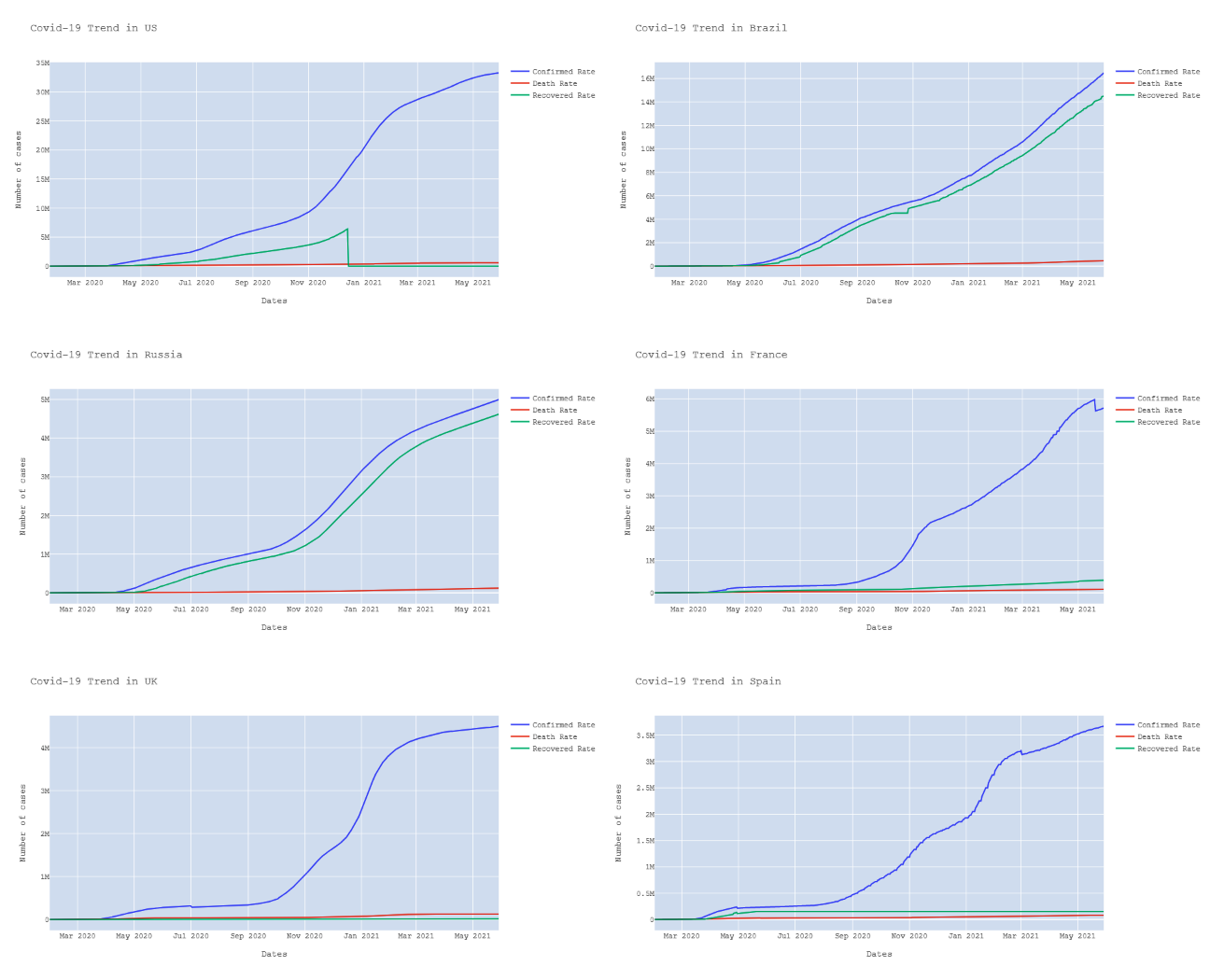


Fig. 2. Graphical representation of country-wise Covid-19 trend.

The worldwide trend of Covid-19 spread has been analyzed and the global trend of Covid-19 confirmed, recovered and death cases have been presented in Figure 3.

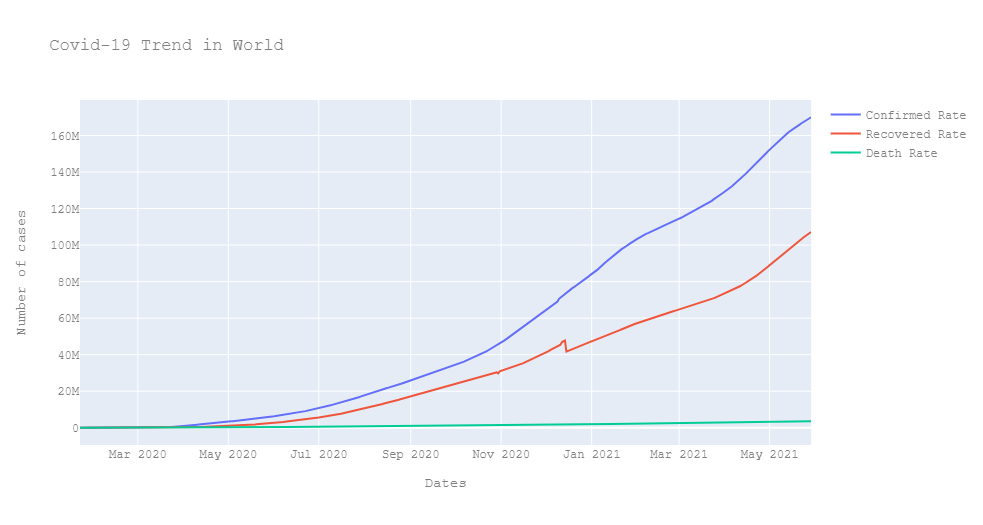


Fig. 3. Graphical representation of the Global Covid-19 trend.

Some most affected countries e.g., the USA, India, Brazil, Russia, France, UK, Turkey, Argentina, Colombia, Spain, Iran, and Italy has been considered for comparing the total Covid-19 trend of these countries to understand the difference in their overall trend. The bar plot chart has been used to plot the data of these countries and figure 4 represents the difference in the country-wise overall trend.

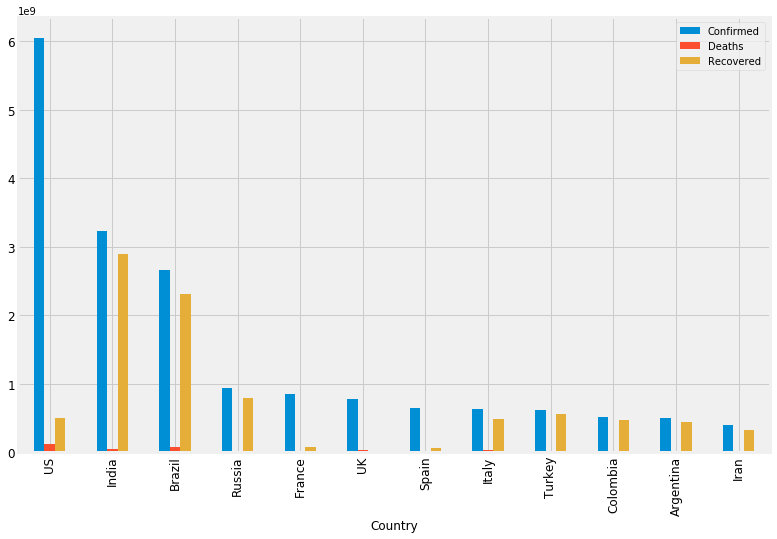


Fig. 4. Visualization of Covid-19 trend difference in most affected countries.

**Machine Learning Modeling**

Time series forecasting is an important area of machine learning. Predictions are made for new data when the actual outcome may not be known until some future date. The future is being predicted, but all prior observations are almost always treated equally (Weigend, 2018). Time series analysis involves developing models that best capture or describe an observed time series to understand the underlying causes. This often involves making assumptions about the form of the data and decomposing the time series into constitution components. The Support Vector Machine is a popular machine learning algorithm often used for time-series prediction tasks (Velásquez et al., 2010; Sapankevych & Sankar, 2009; Kim, 2003; Thissen et al., 2003). For this experiment, the Support Vector Machine algorithm is used to train the SVM classifier based on the date-wise global confirmed rate. The dataset has been divided into 80:20 ratio i.e., 80% for training purpose and 20% for validation purpose. After tainting the SVM classifier has been validated for the prediction of overall daily confirmed cases using the testing data. Finally, the proposed model achieved 74.88% of validation accuracy in the prediction of global confirmed cases. In figure 5, the line plot graph has been presented for the training, validation, and prediction data of the daily covid-19 confirmed cases.

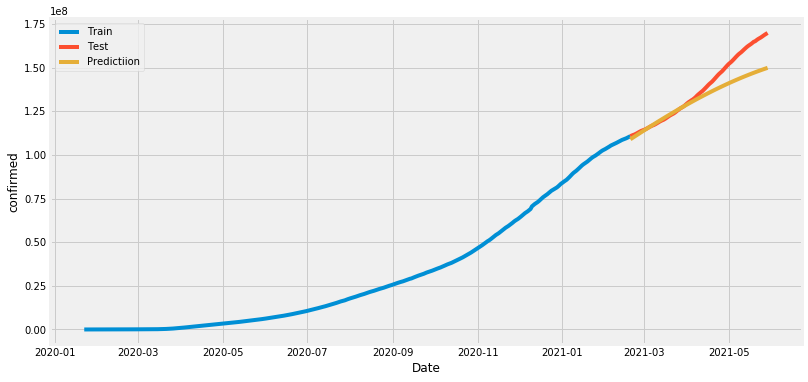


Fig. 5. Graphical representation of Covid-19 trend prediction using SVM model.

**Conclusion**

This experiment is mainly focused on Machine Learning Learning-based Trend Analysis of Covid-19 cases all over the world. During the data analysis phase, some important features have been extracted from the dataset, and using different types of plots, those extracted data have been visualized. After Data analysis a new dataset has been developed consisting of the date-wise global confirmed rate. Then the SVM Classifier has been trained based on the extracted training data and the model achieved 74.88% accuracy on the validation data. For future work, the proposed model can be improved by modifying the hyperparameters to ensure better prediction accuracy over the time-series data.

**Reference**

Arora, P., Kumar, H., & Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons & Fractals*, *139*, 110017.

Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons & Fractals*, *135*, 109864.

Kim, K. J. (2003). Financial time series forecasting using support vector machines. *Neurocomputing*, *55*(1-2), 307-319.

Sapankevych, N. I., & Sankar, R. (2009). Time series prediction using support vector machines: a survey. *IEEE Computational Intelligence Magazine*, *4*(2), 24-38.

Thissen, U. V. B. R., Van Brakel, R., De Weijer, A. P., Melssen, W. J., & Buydens, L. M. C. (2003). Using support vector machines for time series prediction. *Chemometrics and intelligent laboratory systems*, *69*(1-2), 35-49.

Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things*, *11*, 100222.

Velásquez, J. D., Olaya, Y., & Franco, C. J. (2010). Time series prediction using support vector machines. *Revista Chilena de Ingeniería*, *18*(1), 64-75.

Weigend, A. S. (2018). *Time series prediction: forecasting the future and understanding the past*. Routledge.