Vol. XVII & Issue No. 12 December - 2024

INDUSTRIAL ENGINEERING JOURNAL

COMPARING COMPARTMENTAL SIR MODELS AND STOCHASTIC PROCESSES IN MACHINE LEARNING FOR EPIDEMIOLOGY

Ankush A. Patil
Research Scholar, Department of Mathematics, LVH ASC College, Nashik, MS (India)
Email: ankush121patil@gmail.com

S. D. Manjarekar
Professor, Department of Mathematics, LVH ASC College, Nashik, Maharashtra (India)
Email: shrimathematics@gmail.com

Abstract

Compartmental SIR model and Stochastic processes in machine learning are two different approaches to model the spread of diseases or other phenomena that can be spread among individuals. Compartmental SIR model is a mathematical framework that divides the population into three compartments: Susceptible, Infected, Recovered. These models assume individuals transition from one compartment to another based on a set of well-defined rules and it can be used to estimate the spread of diseases and predict future trends. On other side Stochastic processes in machine learning involve randomness and uncertainty in modeling the spread of diseases or other phenomena. Stochastic processes can be used to model the spread of diseases in more complex and realistic ways, taking into account the variability of individuals and the environment and also be used for prediction, decision making and risk assessment. This research paper mainly focuses on analysing behaviour of Compartmental SIR model and Stochastic processes in machine learning of Epidemiology.

Keywords: Mathematical Modeling, Machine Learning, Epidemiology. AMS (2020) Classification: 00A71,68Q05,92C60

INTRODUCTION

The [2]SIR model, also known as the susceptible-infected-recovered model, is a mathematical framework employed in epidemiology for analyzing the transmission dynamics of infectious diseases among a population. In this model, the population is classified into three distinct groups: Susceptible (S), representing individuals who have not yet contracted the disease but are at risk of infection if they come into contact with the pathogen; Infected (I), referring to individuals currently carrying the disease and capable of transmitting it to susceptible individuals; and Recovered (R), denoting individuals who have recuperated from the disease and have developed immunity, rendering them no longer susceptible to infection. The SIR model assumes that the population is closed and that individuals move between compartments over time. The model also assumes that the rate at which individuals move from the susceptible compartment to the infected compartment is proportional to the number of susceptible individuals and the number of infected individuals, while the rate at which individuals move from the infected compartment to the recovered compartment is proportional to the number of infected individuals. Mathematically, The following system of differential equations can be utilized to represent the dynamics of the SIR model:

$$\frac{dS}{dt} = -\beta SI, \qquad \frac{dI}{dt} = \beta SI - \gamma I, \qquad \frac{dR}{dt} = \gamma I$$

In this representation, the transmission rate of the disease (β), which indicates how quickly susceptible individuals become infected, and the recovery rate (γ), which signifies the speed at which infected individuals recover and acquire immunity, are incorporated as essential parameters. The SIR model is a simple but powerful tool for understanding the spread of infectious diseases within a population. It can be used to estimate the size of an outbreak, predict the effectiveness of different control measures, and assess the impact of vaccination programs. However, model is based on a number of simplifying assumptions and may not always accurately reflect the complexity of real-world epidemics.to overcome this situation we added some randomness using stochastic processes using machine learning algorithms. The model assumes that the population is well-mixed and that the transmission of the disease occurs through direct contact between individuals. The basic reproduction number (R_0) serves as a metric to assess the average quantity of new infections that can be attributed to a single infected individual within a population that is entirely susceptible. [3]Stochastic processes are used in the SIR model to introduce randomness into the system, as the transmission of the disease is a probabilistic event. There are different ways to incorporate stochasticity into the SIR model, such as through discrete-time [5]Markov chains or stochastic differential equations. A frequently employed method involves utilizing Monte Carlo simulations, which entail iteratively sampling random variables to estimate the anticipated behaviour of the system. When applied to the SIR model, this entails [1]simulating the transmission of the disease within a population and monitoring the temporal changes in the numbers of susceptible, infected, and recovered individuals. By running multiple simulations, it is possible to estimate the distribution of outcomes and the probability of different scenarios, such as an outbreak or the effectiveness of interventions like vaccination or social distancing. Stochastic processes in machine learning can also be used to fit the SIR model to real-world data, by estimating the model parameters that best describe the observed dynamics of the disease. This involves using techniques like maximum likelihood estimation or Bayesian inference to find the parameter values that optimize the fit between the model and the data. Overall, stochastic processes in machine learning are a powerful tool for understanding the spread of infectious diseases and evaluating the impact of different control strategies.

DATASET

We will make use of the [7]SIRF model, a modified ODE model derived from the SIR model. Subsets of time series data from each nation will be subjected to parameter estimation of the SIR-F to assess the impact of measures. S-R trend analysis will be used to determine parameter change points. Datasets used for this analysis are 1) COVID-19 Data Hub: provides information on the case count, population statistics, and government responses, sourced from the Oxford Covid-19 Government Response Tracker. 2) Population Pyramid : offers insights into the age distribution of a population. 3) Our World In Data: presents data on COVID-19 testing, vaccinations, and the number of individuals who have received vaccinations. 4) COVID-19 dataset in Japan: contains information on the number of cases reported in the country. This dataset includes population values necessary for calculating the number of "susceptible" cases in the SIRF model, where "susceptible" is determined as the difference between the total population and the confirmed cases

METHODOLOGY

Modified [7]SIRF Model: While the "R" in the traditional SIR model represents individuals who have recovered and acquired immunity, in this modified version, I have redefined "R" as individuals who have either recovered or experienced a fatal outcome. This adjustment is made to acknowledge the significance of mortality rates in real COVID-19 data, as it cannot be disregarded.

S: Susceptible (= Population - Confirmed)
$S*$: Confirmed and un-categorized
I: Confirmed and categorized as Infected
R: Confirmed and categorized as Recovered
F: Confirmed and categorized as Fatal

Measurable variables
Confirmed = $I + R + F$
Recovered = R
Deaths = F

$$S \xrightarrow{\beta I} S* \xrightarrow{\alpha_1} F$$
$$S* \xrightarrow{1-\alpha_1} I \xrightarrow{\gamma} R$$
$$I \xrightarrow{\alpha_2} R$$

α_1: Represents the non-dimensional direct fatality probability of individuals in the susceptible population (S∗S∗).

α_2: Represents the mortality rate of infected cases, measured in units of inverse minutes [1/min].

β : Represents the mortality rate of infected cases, measured in units of inverse minutes [1/min].

γ : Renotes the recovery rate, measured in units of inverse minutes [1/min].

[7]Ordinary Differential Equation (ODE):

$$\frac{dS}{dT} = -N^{-1}\beta SI,$$
$$\frac{dI}{dT} = N^{-1}(1-\alpha_1)\beta SI - (\gamma + \alpha_2)I,$$
$$\frac{dR}{dT} = \gamma I,$$
$$\frac{dF}{dT} = N^{-1}\alpha_1 \beta SI + \alpha_2 I$$

Here, N represents the total population, which is the sum of individuals in the susceptible (S), infected (I), recovered (R), and fatal (F) compartments. T denotes the elapsed time from the initial start date.

S∗ refers to cases who are carriers of the disease, but their infection status is unknown to themselves and others. This group includes individuals who may pass away and are later confirmed as positive cases post-mortem. Additionally, some individuals from this group may be reclassified as infected after their infection status is confirmed.

[7]Non-dimensional SIRF model:

$$(S, I, R, F) = N \times (x, y, z, w)$$
$$(T, \alpha_1, \alpha_2, \beta, \gamma) = (\tau t, \theta, \tau^{-1}\kappa, \tau^{-1}\rho, \tau^{-1}\sigma)$$
$$\frac{dx}{dt} = -\rho xy$$
$$\frac{dy}{dt} = \rho(1-\theta)xy - (\sigma + \kappa)y$$
$$\frac{dz}{dt} = \sigma y$$
$$\frac{dw}{dt} = \rho\theta xy + \kappa y$$
$$0 \leq (x, y, z, w, \theta, \kappa, \rho, \sigma) \leq 1$$
$$1 \leq \tau \leq 1440$$

Reproduction number can be defined as

$$R_0 = \rho(1-\theta)(\sigma + \kappa)^{-1} = \beta(1-\alpha_1)(\gamma + \alpha_2)^{-1}$$

We applied this SIRF model for analysing the dataset of Covid -19 data related to Japan the SIRF model can help to estimate the potential spread of the disease in the population, how quickly it is spreading, and the effectiveness of measures implemented to slow its spread. By analysing the data using this model, it may be possible to identify trends and patterns in the spread of the disease that could inform public health interventions and policies. Figure 1 presents the chronological trend of cases in Japan, exhibiting a continuous upward trajectory of confirmed cases and fatal cases. The data for confirmed cases and fatal cases is supplemented with complementary information.

Figure 1

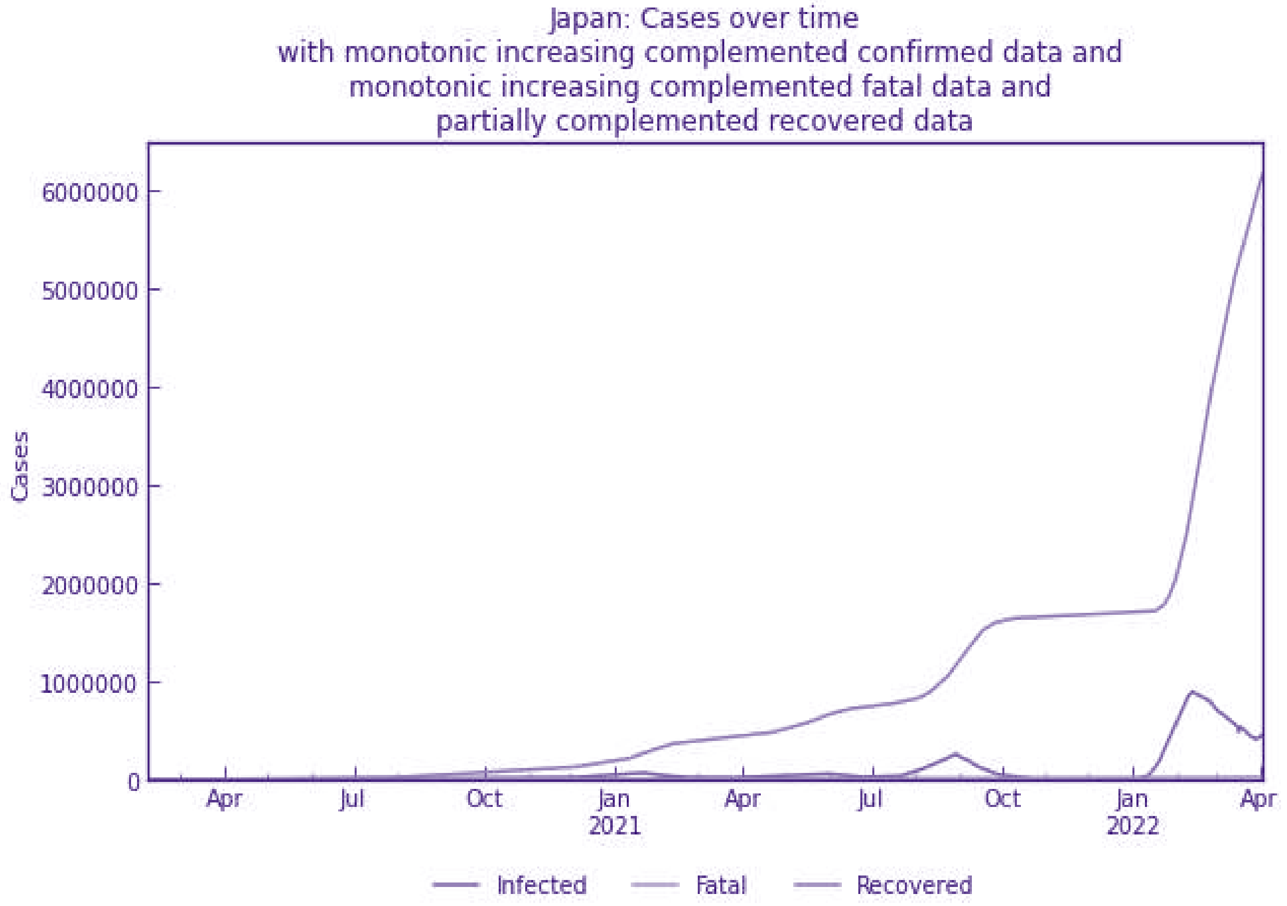
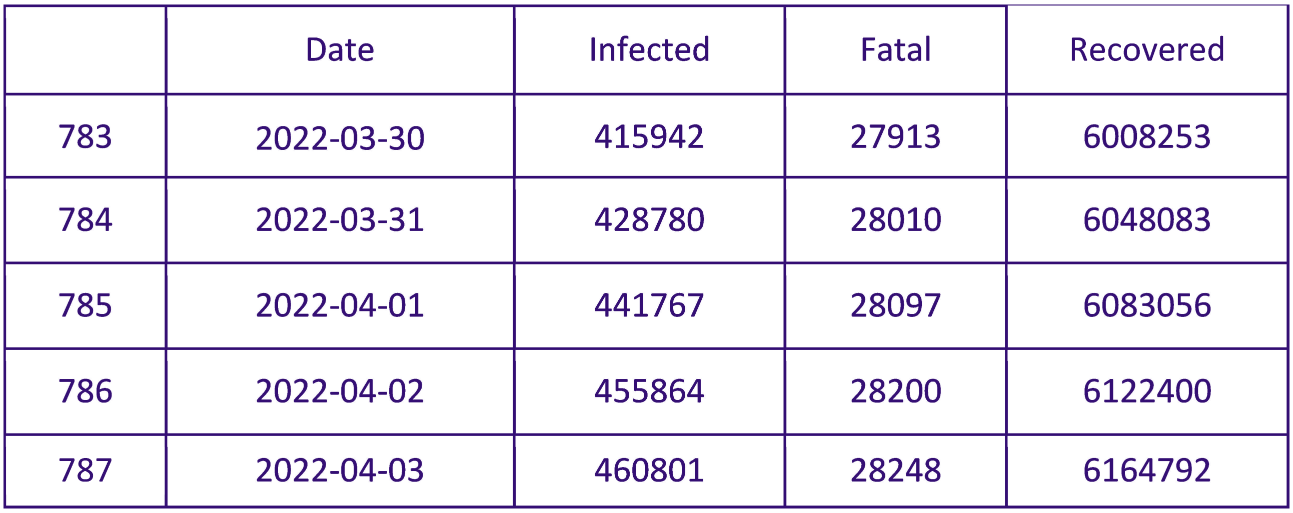


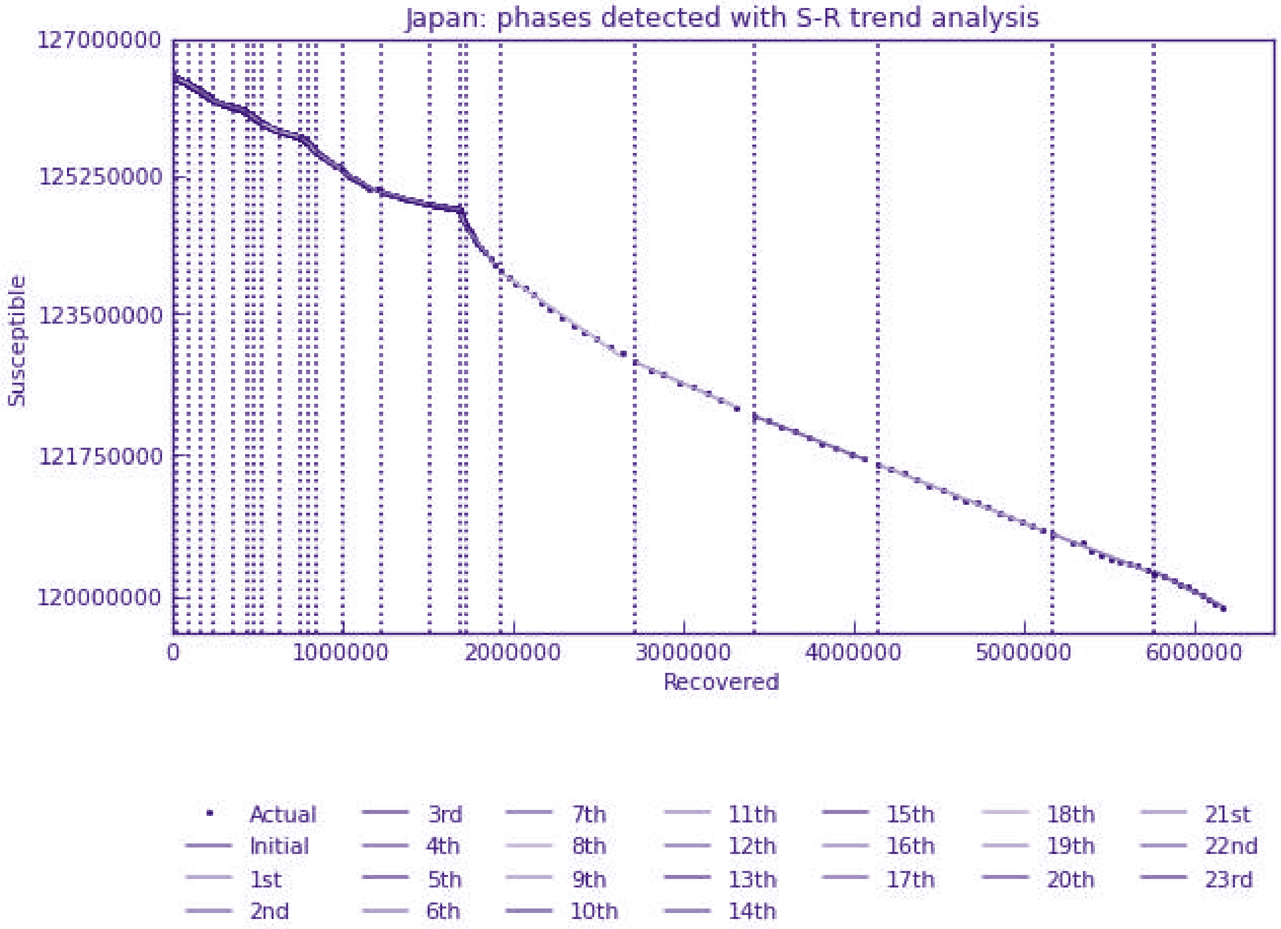
Table 1



	Date	Infected	Fatal	Recovered
783	2022-03-30	415942	27913	6008253
784	2022-03-31	428780	28010	6048083
785	2022-04-01	441767	28097	6083056
786	2022-04-02	455864	28200	6122400
787	2022-04-03	460801	28248	6164792

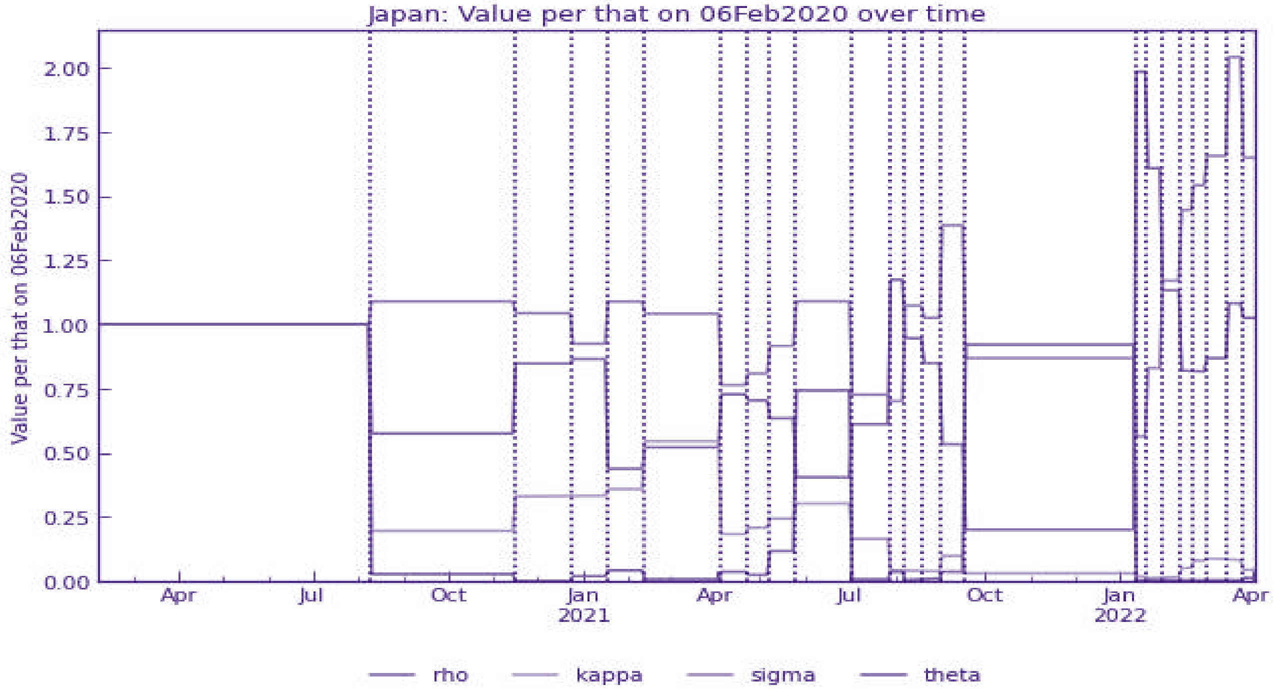
S-R Trend analysis gives brief introduction of phases of covid-19 in Japan. Figure 2 gives Brief idea about Susceptible verses Recovered in different phases. If the rate of new infections is increasing over time, the "S" curve may start to flatten out while the "R" curve remains steep. This could indicate that the disease is spreading more rapidly and that measures need to be taken to slow its spread. We also analysed change in different parameters of SIRF i.e. θ,κ,ρ,σ. Figure 3 depicts change in θ,κ,ρ,σ over time.

Figure 2



Estimating SIRF Parameters

Figure 3



RMSLE Table

Table 3

Start	End	Rt	θ	κ	ρ	σ	τ	α_1	$\frac{1}{\alpha_2}$ [day]	$\frac{1}{\beta}$[day]	$\frac{1}{\gamma}$ [day]	RMSLE
06-Feb-20	08-Aug-20	1.56	0.0271	0.0037	0.1173	0.0696	1440	0.027	267	8	14	1.192742
09-Aug-20	14-Nov-20	0.88	0.0007	0.0007	0.0674	0.0758	1440	0.001	1364	14	13	0.145371
15-Nov-20	23-Dec-20	1.34	0.0000	0.0012	0.0994	0.0727	1440	0.000	809	10	13	0.060244
24-Dec-20	16-Jan-21	1.54	0.0005	0.0012	0.1014	0.0644	1440	0.001	801	9	15	0.030658
17-Jan-21	10-Feb-21	0.66	0.0011	0.0013	0.0513	0.0758	1440	0.001	746	19	13	0.058724
11-Feb-21	03-Apr-21	0.82	0.0002	0.0020	0.0611	0.0725	1440	0.000	490	16	13	0.130384
04-Apr-21	21-Apr-21	1.58	0.0010	0.0007	0.0853	0.0532	1440	0.001	1460	11	18	0.012745
22-Apr-21	06-May-21	1.44	0.0007	0.0008	0.0825	0.0563	1440	0.001	1284	12	17	0.011666
07-May-21	24-May-21	1.15	0.0032	0.0009	0.0743	0.0638	1440	0.003	1099	13	15	0.034255
25-May-21	01-Jul-21	0.6	0.0201	0.0011	0.0473	0.0759	1440	0.020	884	21	13	0.032564
02-Jul-21	27-Jul-21	1.4	0.0002	0.0006	0.0717	0.0506	1440	0.000	1619	13	19	0.052941
28-Jul-21	06-Aug-21	2.81	0.0010	0.0002	0.1376	0.0488	1440	0.001	6251	7	20	0.020124
07-Aug-21	19-Aug-21	1.48	0.0001	0.0002	0.1110	0.0747	1440	0.000	5694	9	13	0.012027
20-Aug-21	31-Aug-21	1.39	0.0003	0.0001	0.0995	0.0714	1440	0.000	7165	10	14	0.028436
01-Sep-21	16-Sep-21	0.65	0.0010	0.0004	0.0625	0.0965	1440	0.001	2704	15	10	0.037157
17-Sep-21	10-Jan-22	0.38	0.0249	0.0001	0.0236	0.0604	1440	0.025	8738	42	16	0.54253
11-Jan-22	18-Jan-22	5.91	0.0001	0.0001	0.2326	0.0393	1440	0.000	14083	4	25	0.022941
19-Jan-22	28-Jan-22	3.26	0.0001	0.0000	0.1886	0.0578	1440	0.000	24211	5	17	0.038196
29-Jan-22	10-Feb-22	1.63	0.0002	0.0001	0.1328	0.0815	1440	0.000	15972	7	12	0.02268

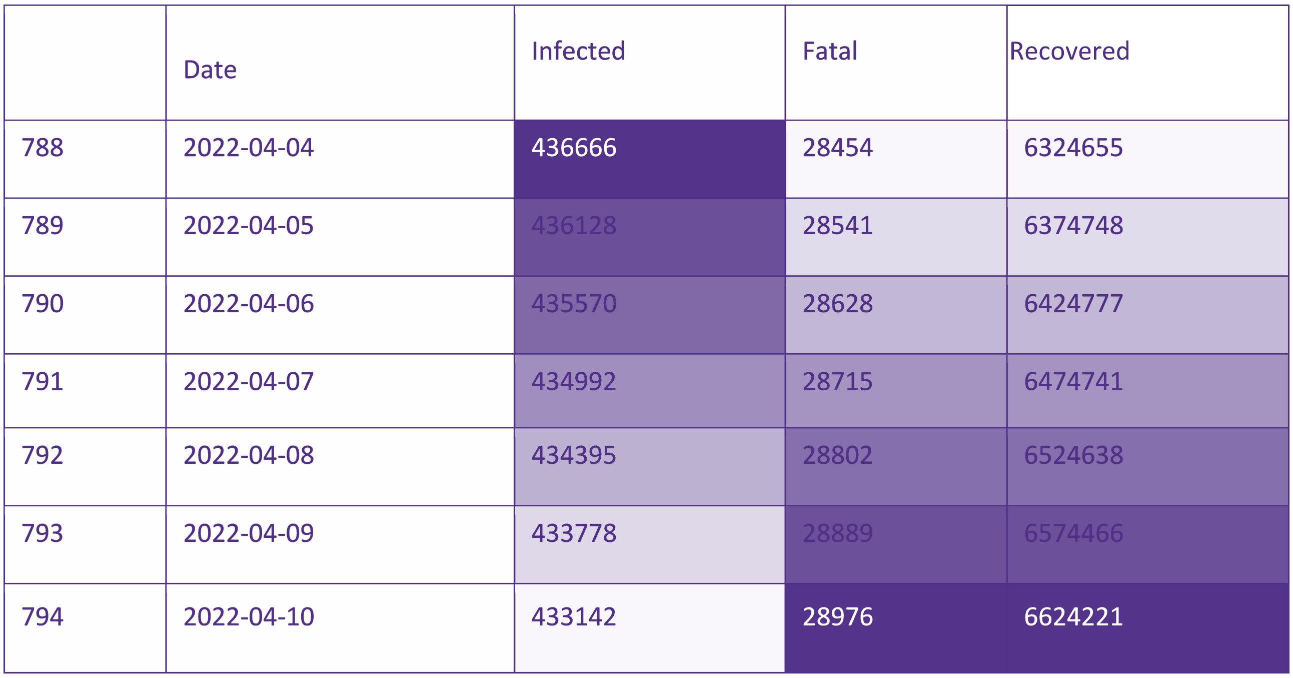
11-Feb-22	18-Feb-22	0.95	0.0001	0.0002	0.0962	0.1007	1440	0.000	5272	10	9	0.008513
19-Feb-22	27-Feb-22	0.89	0.0001	0.0003	0.0957	0.1074	1440	0.000	3320	10	9	0.009399
28-Feb-22	14-Mar-22	0.88	0.0001	0.0003	0.1018	0.1153	1440	0.000	3013	9	8	0.010579
15-Mar-22	24-Mar-22	0.89	0.0001	0.0003	0.1270	0.1421	1440	0.000	3249	7	7	0.02492
25-Mar-22	03-Apr-22	1.05	0.0003	0.0002	0.1202	0.1148	1440	0.000	6112	8	8	0.022018

RMSLE (Root Mean Squared Logarithmic Error) is a metric used to evaluate the performance of regression models. It measures the ratio between the actual and predicted values of a target variable, taking into account the logarithmic scale of the values. The RMSLE score is computed by taking the root mean squared error of the logarithmic differences between the actual and predicted values.:

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(x_i+1)-\log(y_i+1))^2}$$

where x_i and y_i are the predicted and true values of the target variable, respectively. The RMSLE score is useful when the target variable has a wide range of values and/or has a skewed distribution, as it puts less weight on large errors and more weight on small errors. A lower RMSLE score indicates better model performance. Table 3 gives RMSLE score of each phase starting from 6 Feb 2020 to 3 April 2022.

Table 4



	Date	Infected	Fatal	Recovered
788	2022-04-04	436666	28454	6324655
789	2022-04-05	436128	28541	6374748
790	2022-04-06	435570	28628	6424777
791	2022-04-07	434992	28715	6474741
792	2022-04-08	434395	28802	6524638
793	2022-04-09	433778	28889	6574466
794	2022-04-10	433142	28976	6624221

[3][6][7]Stochastic Machine Learning Model:

A stochastic machine learning model is a type of model that incorporates randomness or probability into the learning process. In stochastic models, the outcome is not deterministic, but rather, a probabilistic distribution over possible outcomes is generated. Stochastic models are commonly used in situations where the data is noisy or uncertain, or when the underlying process generating the data has some inherent randomness. Examples of stochastic models include logistic regression, [7]Gaussian mixture models, and Bayesian networks. Stochastic models are trained using stochastic optimization algorithms, which use random sampling techniques to estimate the model parameters. These algorithms are iterative and update the model parameters at each step, using a small subset of the data (called a mini-batch) to estimate the gradient of the loss function.

Here we imported dataset located in the specified path and loads it into a Pandas Data Frame. After that we completed pre-processing of Dataset that we using. A line plot of the three SIR variables (Confirmed, Recovered, and Fatal) is plotted over time using the plt. plot function from the matplotlib library Figure 4.After analysing the data following observation made that recovery rate is far much on higher side as compare to fatality rate and there is very less difference in rate of Infection and rate of Recovered. All attributes of SIRF model can be observed in Figure 5, which gives brief insights of simulation of SIRF model. Relationship between Infected and Recovered observed in Figure 6 which shows that Infected and Recovered are positively correlated.

Figure 4

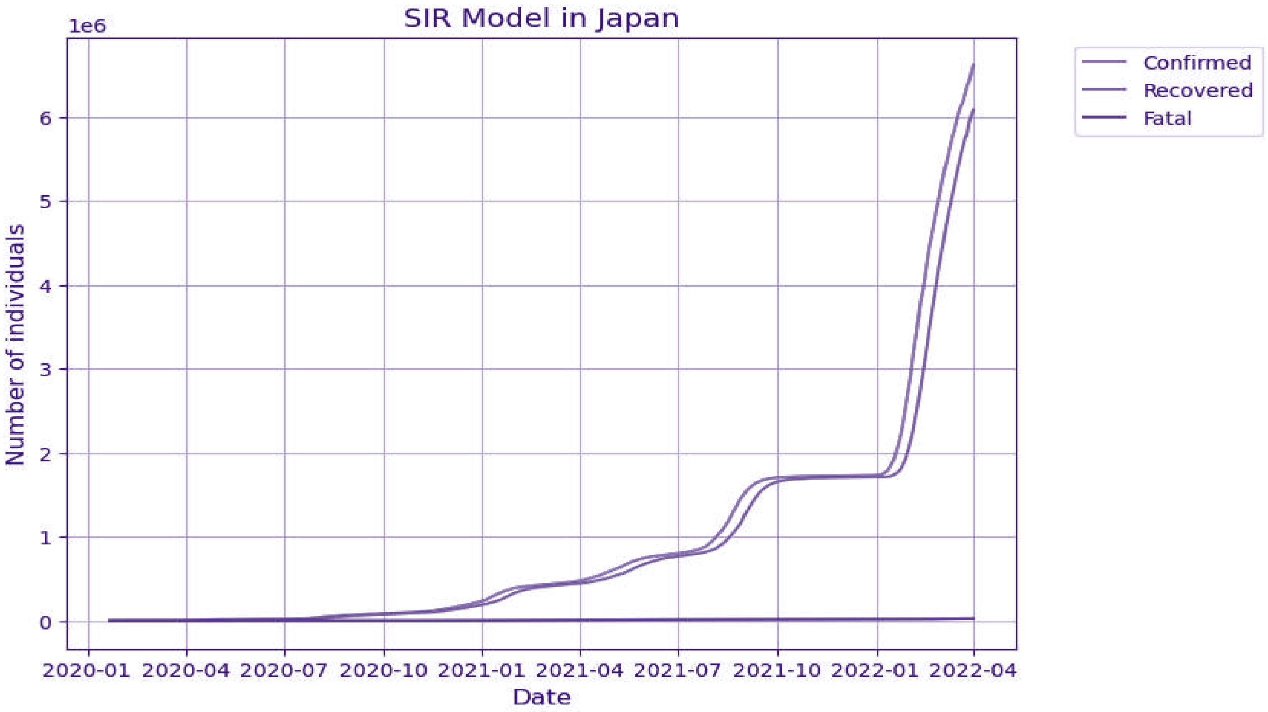
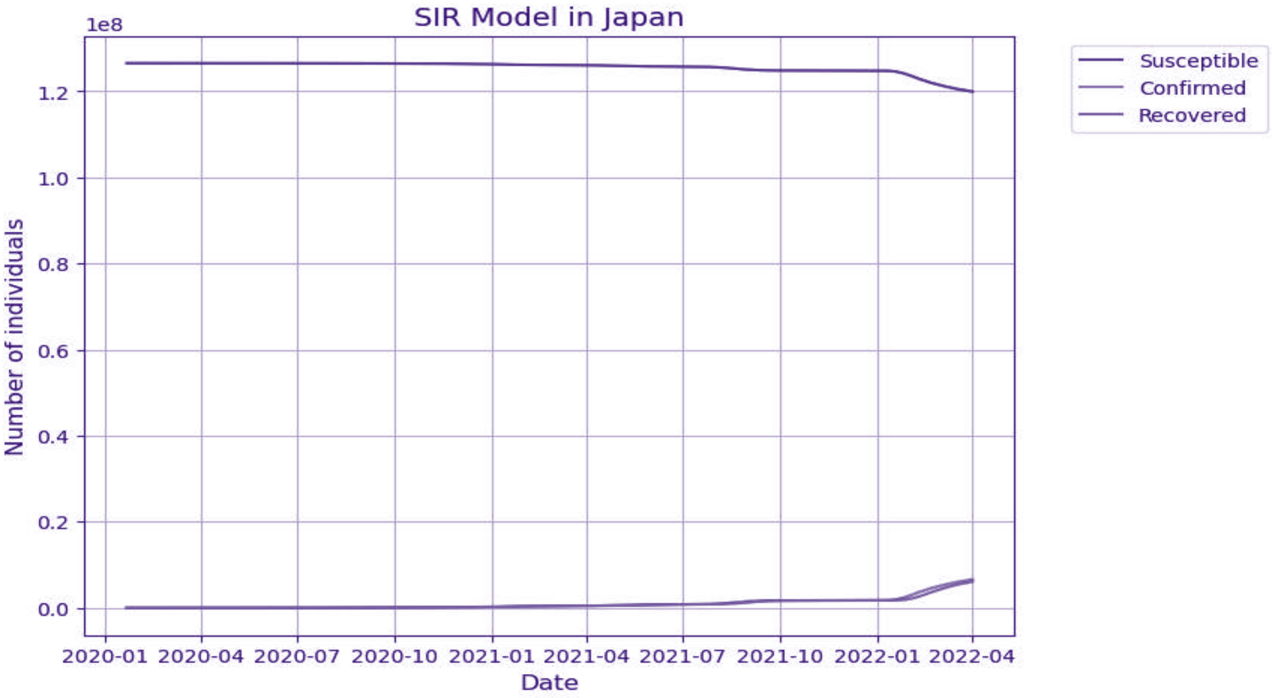
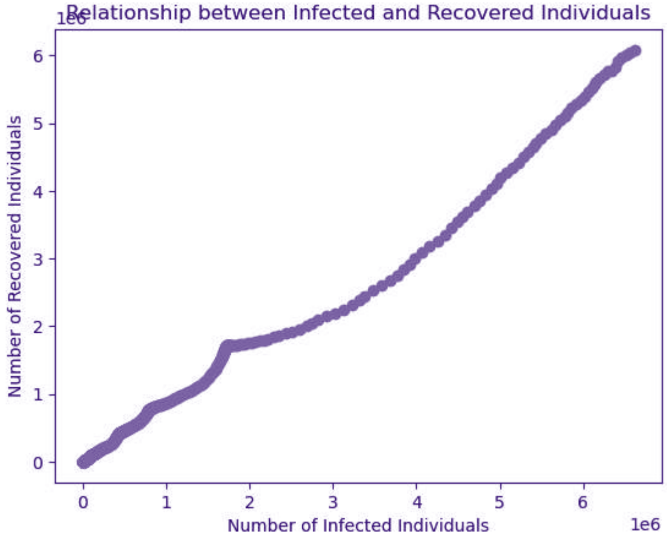


Figure 5

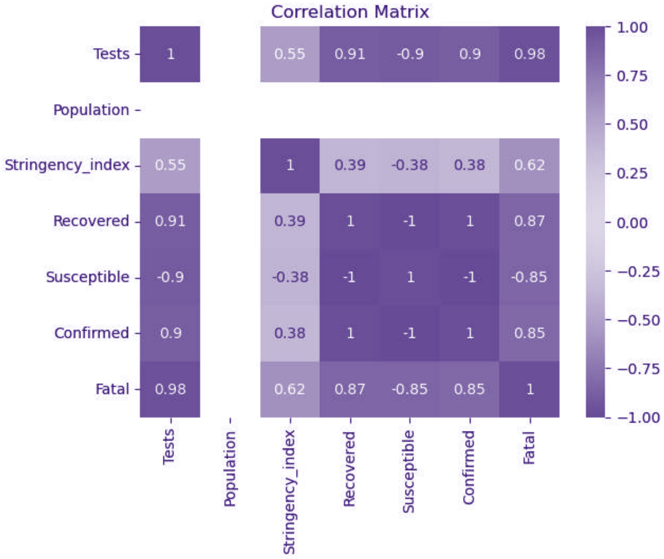


A scatter plot using the plt. scatter function from the matplotlib library. The plot shows the relationship between the number of confirmed cases (x-axis) and the number of recovered cases (y-axis)(Figure 6)

Figure 6



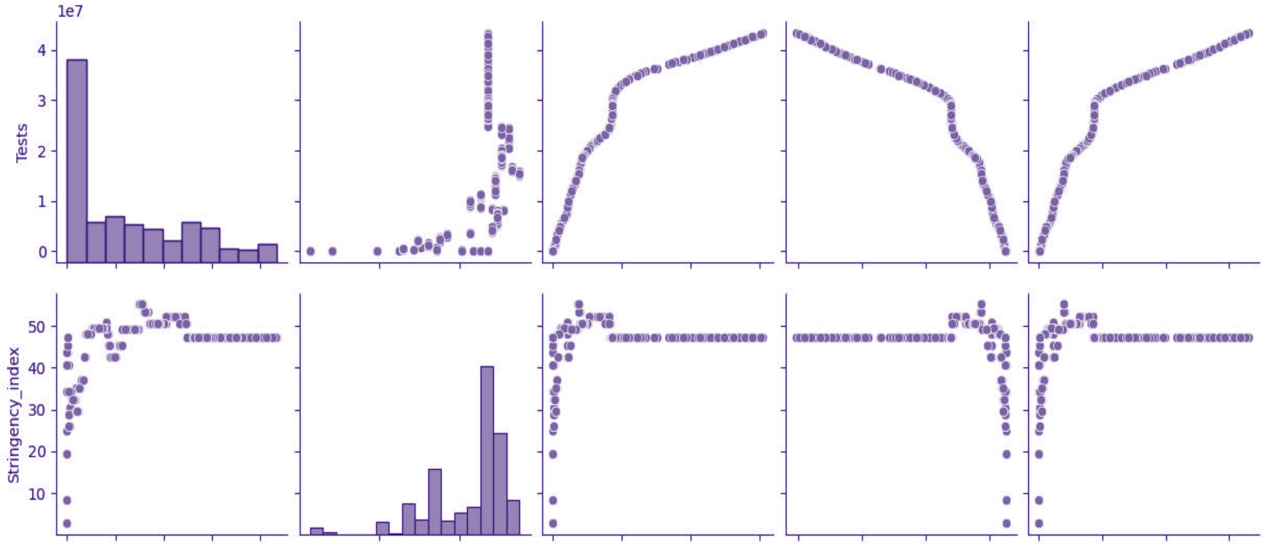
Generated a heatmap of the correlation matrix, you can utilize the sns.heatmap function from the seaborn library. The corr_matrix variable should store the correlation matrix, which is a square matrix with dimensions equal to the number of variables in the dataset.

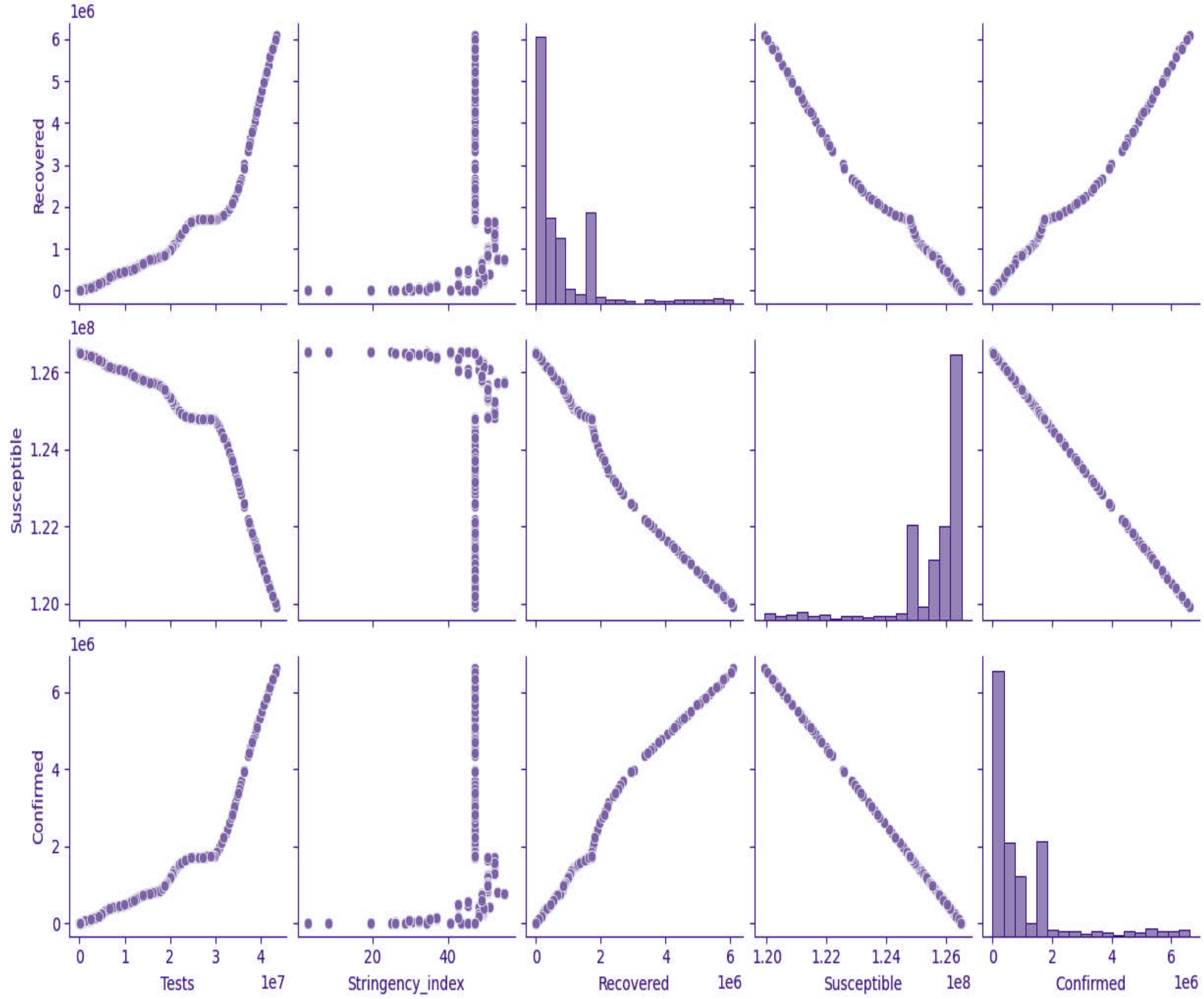


Next split the data into training and testing sets using the **train_test_split** function from the scikit-learn library. The independent variables (also called features) are defined as the columns **'Tests', 'Stringency_index', 'Recovered', 'Susceptible', 'Confirmed'** of the **df** Data Frame, and are assigned to the **X** variable. The dependent variable (also called target variable) is defined as the column **'Fatal'** of the **df** Data Frame, and is assigned to the **y** variable. The **test_size** parameter of the **train_test_split** function specifies the proportion of the data that should be allocated to the testing set, in this case, 30%. The **random_state** parameter ensures that the same split is obtained every time the code is run, which is useful for reproducibility purposes. After running this code, the training and testing sets are stored in the **X_train**, **X_test**, **y_train**, and **y_test** variables, respectively. The training sets are employed to train and fit the model, whereas the testing sets are utilized to assess the model's performance on unseen data.

After splitting Data created a pair plot using the **sns.pairplot** function from the seaborn library. The **X_train** variable should contain the training data in the form of a Pandas DataFrame, where each column represents a different feature and each row represents an instance or observation.(Figure7)

Figure 7

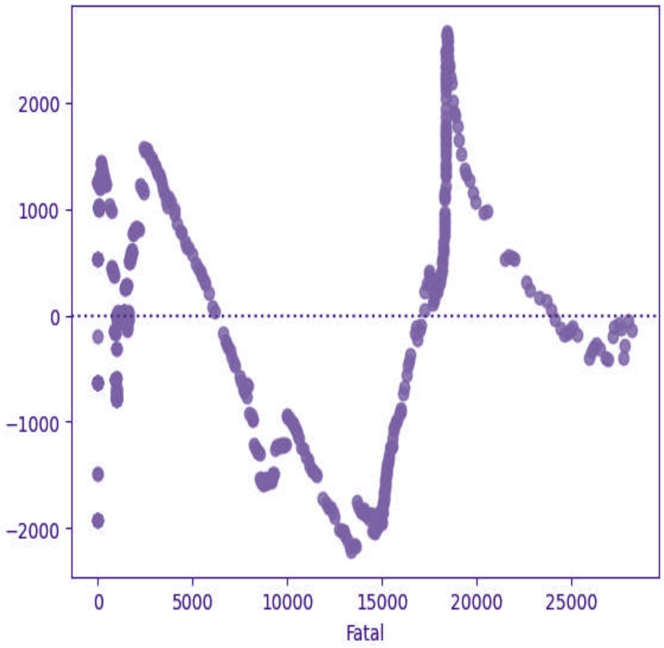




Once the data is split, a linear regression model is constructed using the Linear Regression class from the scikit-learn library. The model is then fitted to the training data using the fit method. The model variable is assigned as an instance of the Linear Regression class, representing the linear regression model. The fit method takes the training data, namely X_train and y_train, as input parameters. This process involves determining the optimal coefficients for the model that minimize the discrepancy between the predicted values and the actual values. After fitting data The score method of the model object takes the testing set (X_test and y_test) as arguments and returns the R-squared value, which is a measure of how well the model fits the data.

A residual plot using the Seaborn library in Python. A residual plot is a graph that shows the difference between the observed values of the dependent variable (y_train) and the predicted values (y_pred) on the vertical axis, against the independent variable(s) on the horizontal axis. Figure 8 is residual plot of Japan SIRF model. The resulting plot can help us determine if there is a pattern in the residuals, which can indicate if our model is missing something important or if it is overfitting.

Figure 8



OBSERVATION

Fall in ρ ,σ observed

Japan declared a state of emergency three times.

- 07 April 2020-15 April 2020: Only three metropolitan areas
- 16 April 2020 – 06 May 2020: Nation-wide
- 07 May 2020 – 31 May 2020: Nation-wide

The primary actions undertaken by Japan mainly consist of the following three measures.

- Avoiding enclosed settings, crowds, and close-up conversations through physical and social distance.
- Follow the links between patients and prioritise testing for those connected.
- Keep medical standards high to improve recovery rates and lower fatality rates.

After analysing the data following observation made

- Recovery rate is far much on higher side as compare to fatality rate and there is very less difference in rate of Infection and rate of Recovered.
- All attributes of SIRF model can be observed in Figure 5, which gives brief insights of simulation of SIRF model.
- Relationship between Infected and Recovered observed in Figure 6 which shows that Infected and Recovered are positively correlated.

RESULTS

RMSLE (Root Mean Squared Logarithmic Error) is a metric used to evaluate the performance of regression models. It measures the ratio between the actual and predicted values of a target variable, taking into account the logarithmic scale of the values. The RMSLE score is calculated as the root mean squared error of the logarithmic differences between the actual and predicted values. Table 3 shows all values of RMSLE obtain in simulation of SIRF model.

The R-squared value, which falls between 0 and 1, serves as a measure of the goodness of fit for a model. A higher R-squared value indicates a better fit. In our case, obtaining an R-squared value of 0.978 implies that the linear regression model explains approximately 97.8% of the variance in the dependent variable. This high value suggests that the model fits the data well and captures a significant portion of the variation.

The MAE measures the average absolute difference between the predicted and actual values. In this case, an **MAE** of **1013.54** indicates that, on average, the predicted values are approximately 1013.54 units away from the actual values. The units of the MAE are the same as the target variable.

The RMSE (Root Mean Squared Error) quantifies the square root of the average squared difference between the predicted and actual values. For this particular scenario, an RMSE value of 1205.39 suggests that, on average, the predicted values deviate by approximately 1205.39 units from the actual values. It's worth noting that the units of the RMSE are consistent with those of the target variable. Lower values for both MAE (Mean Absolute Error) and RMSE indicate improved performance of the linear regression model.

As the residuals are randomly scattered around the horizontal axis, this is a good indication that our model is a good fit for the data. (Figure 8)

CONCLUSION

SIRF models are designed specifically for modelling the spread of infectious diseases and can be highly predictive under specific assumptions, but they may not be as effective in predicting outcomes beyond the spread of the disease. Machine learning models are more versatile and can be used for a wide range of applications, but they can be more difficult to interpret and optimize.

REFERENCES

[1] *Al-Rahman, El-Nor Osman, M. Adu, I.K. Yang C. 2017, "A simple SEIR mathematical model of malaria transmission", Asian Res. J. Math. 7, pp. 1–22*

[2] *Chen, Y., Lu, P., Chang, C. 2020, "A time-dependent sir model for COVID-19"*

[3] *C. Castillo-Chavez, S. Fridman, X. Luo 1993, "Stochastic and deterministic models in epidemiology", Tech. Rep. BU 1192-M, Biometric Unit, Cornell University, USA*

[4] *Saad Awadh Alanazi, M. M. Kamruzzaman, Madallah Alruwaili, Nasser Alshammari, Salman Ali Alqahtani, Ali Karime, 2020, "Measuring and Preventing COVID-19 Using the SIR Model and Machine Learning in Smart Health Care" Hindawi, Article ID 8857346, 12 pages*

[5] *M. H. A. Davis, 1984 "Piecewise-deterministic Markov Processes: A General Class of Non-Diffusion Stochastic Models", r. R. Statist. Soc. B,46, No.3, pp. 353-388*

[6] *P. E. Greenwood, L. F. Gordillo, 2009, "Stochastic Epidemic Modeling", Springer Netherlands, Dordrecht, pp. 31-52.*

[7] *Roberto Vega, Leonardo Flores, Russell Greiner, 2022 "SIMLR: Machine Learning inside the SIR Model for COVID-19 Forecasting", Forecasting, 4, pp. 72–94.*