Chapter 9
Prediction of Corona Virus Disease Outcome and Future Forecast the Trend of COVID-19 in India: Observational Study

Amit Kumar Mishra, Ramakant Bhardwaj, and K. Sruthila Gopalakrishnan

Abstract This work is motivated by the disease caused by the novel corona virus Covid-19, rapid spread in India. An encyclopaedic search from India and worldwide social networking sites was performed between 1 March 2020 and 20 Jun 2020. Nowadays social network platform plays a vital role to track spreading behaviour of many diseases earlier then government agencies. Here we introduced the approach to predict and future forecast the disease outcome spread through corona virus in society to give earlier warning to save from life threats. We compiled daily data of Covid-19 incidence from all state regions in India. Five states (Maharashtra, Delhi, Gujarat, Rajasthan and Madhya-Pradesh) with higher incidence and other states considered for time series analysis to construct a predictive model based on daily incidence training data. In this study we have applied the predictive model building approaches like k-nearest neighbour technique, Random-Forest technique and stochastic gradient boosting technique in COVID-19 dataset and the simulated outcome compared with the observed outcome to validate model and measure the performance of model by accuracy (ACC) and Kappa measures. Further forecast the future trends in number of cases of corona virus deceased patients using the Holt Winters Method. Time series analysis is effective tool for predict the outcome of corona virus disease.

Keywords Corona virus trend · COVID-19 · Machine learning · Predictive modelling · K-nearest neighbour technique · Random-Forest technique and stochastic gradient boosting technique · Holt Winters Method · Mathematical modelling

A. K. Mishra (✉)
Department of Computer Science and Engineering, Amity School of Engineering and Technology (ASET), Amity University, Gwalior, Madhya Pradesh, India
e-mail: akmishra1@gwa.amity.edu

R. Bhardwaj · K. S. Gopalakrishnan
Department of Mathematics, Amity University, Kolkata, West Bengal, India
e-mail: rbhardwaj@kol.amity.edu

R. Anandan et al. (eds.), *How COVID-19 is Accelerating the Digital Revolution*,
https://doi.org/10.1007/978-3-030-98167-9_9

9.1 Introduction

The World Health Organization (WHO)-China office, informed the case of pneumonia detected in Wuhan city of China province on 31 December 2019 [1]. This novel virus further known as corona virus, COVID-19 (formally known as 2019-nCoV) [3]. This virus circulates in entire globe which affect all age groups and spared through physical contact of persons, resulting a serious health issue worldwide. From 11 January 2020 to till date i.e. 20 May 2020, total death worldwide due to COVID-19 noted 32,3156 and 4.86 million total confirmed cases [1]. First case of COVID-19 reported in India on 30 January 2020 and Ministry of Health and Family Welfare provided the data, total 11, 2196 corona cases and 3435 deaths in country [4, 2].

In this study, we use epidemiological data collected during COVID-19 outbreak in India in between 1 March 2020 and 20 May 2020 [5]. Here a model is developed to predict death or recovery from the unlabelled cases of COVID-19.

9.2 Data and Methods

9.2.1 Data Sources

Different sources were used to collect data for the Covid-19 trend identification. An application data is collected from internet [5] which has been developed to report live situation on corona virus pandemics in India. Validity and accuracy of reported data of this application is maintained by live updating of data based on state news bulletin, Official sites of Chief Ministers office /Health ministry, Press conferences, ANI reports and social networking platforms like Twitter, facebook etc. That data contains demographic details about the corona virus decease patients and deep analysis of Total cases, daily update of cases, cases by states, growth trends etc. Another source of data taken from social platform [3] used to show statistics of current trend in COVID-19 worldwide. This statistics and research was referenced to use data of World Health Organization and National Center for Health Statistics etc. This data contains worldwide statistics of total confirmed deaths, deaths trend on daily basis, cases growth rate with global comparison, testing pattern like testing pattern per thousand people, daily tests, test per case, confirmed case etc. Data used for this study validated by compare it with reports of Ministry of Health and Family Welfare [4]. They scheduled update the data periodically on situation of outbreaks of COVID-19 in India.

9.2.1.1 Data

In this study, we have considered the data of pandemic of COVID-19 from 1 March 2020 to 20 May 2020. Till date total 112,196 confirmed cases out of which 63,339 active cases, 45,422 recovered case and 3435 deaths reported. Age density of Death and Recover cases shown in Fig. 9.1.

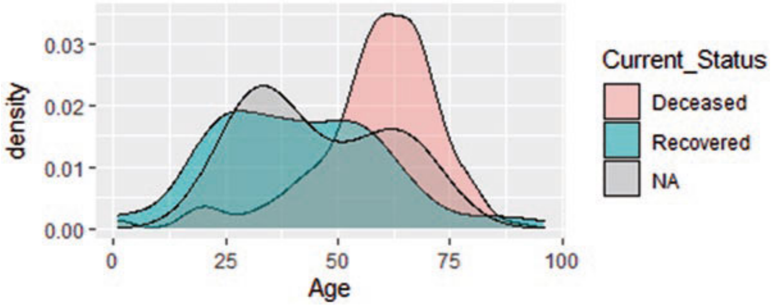

Fig. 9.1 Age density graph of death and recover cases

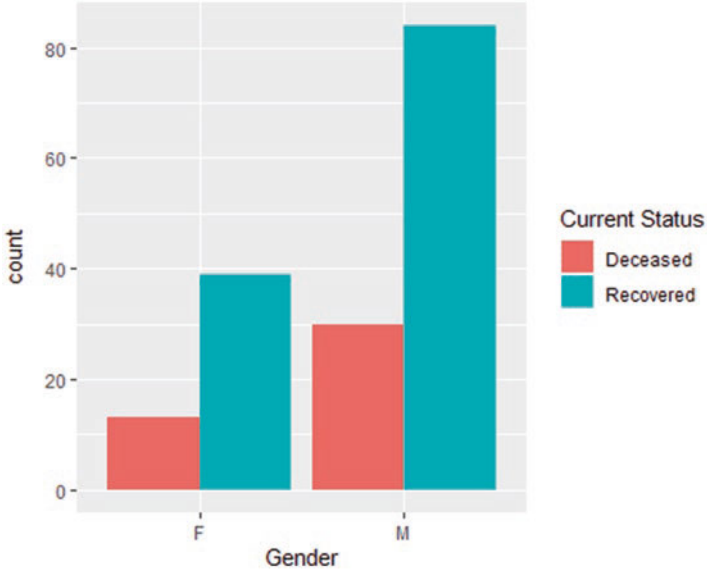

Fig. 9.2 Recovery and death ratio for male and female cases

Figure 9.2 show the recover case of Female patient and Male patients (based on data available till now and it is not sufficient and complete to provide accurate statistics).

Figure 9.3 shows the weekly time line data of recovered, deceased and hospitalized cases till 30 April 2020.

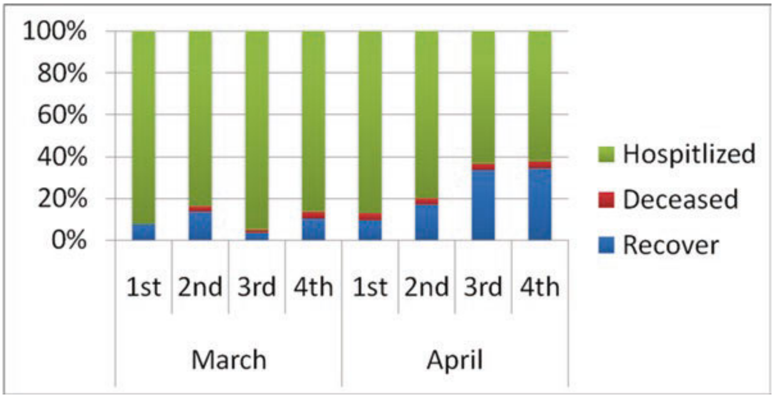

Fig. 9.3 Time line data of COVID-19 cases till 30 April 2020

9.2.2 Methods

Machine learning gives many ways to discover pattern and predict COVID-19 outcome from dataset of corona virus. In this study we used few machine learning techniques, such as Random Forest [6, 7], Stochastic gradient boosting [8, 9] and k-nearest neighbour [10], which have played an essential contribution to the field of data science.

This paper introduced a method of machine learning for predict the outbreak of corona virus outcome using web data. This framework of our introduced method is illustrated in Fig. 9.4.

9.2.2.1 Statistical Steps

In first step, we extract the features of our data to get important trend through machine learning model. This step is necessary to reduce the size of number of input features and to get important features out of that. Further checked the missing data values in data and imputed it with multivariate imputation by chained equation [11]. Most of the data mining algorithm cannot deal with missing data, so they exclude the data completely from model. If dataset is large and missing value is less then it is better to exclude them but if data set is small then better is to impute the missing values.

Consequently in step two, Fig. 9.5 shows the machine learning models will be used with different model parameters offered to train the model using training data.

In forth step, detection will get perform with proper feature subset and efficient model chooses through their accuracy and kappa value. The appropriate trained model used to predict the outcome of corona virus decease.

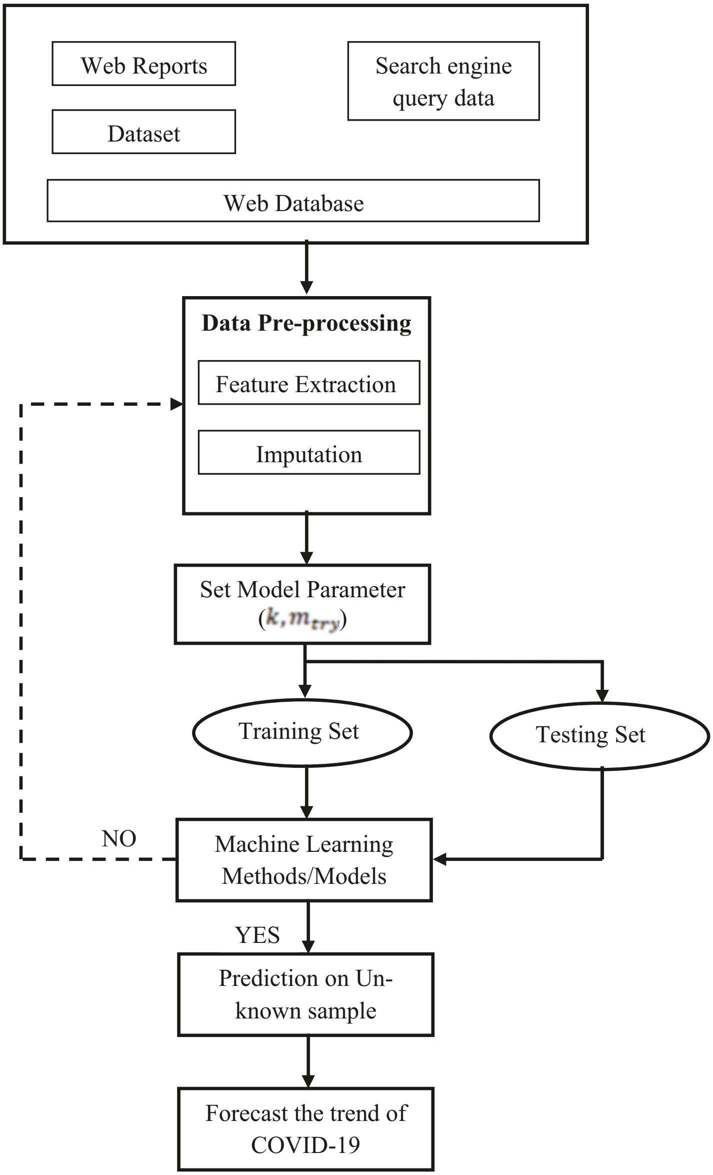

Fig. 9.4 Block diagram of proposed model

Additionally, the performance of prediction has been evaluated by the measures like Accuracy (ACC) [12], which is a measurement of closeness to specific value and another measure is Cohen kappa coefficient [13], which is a statistics to measure the inter-rater reliability for qualitative items:

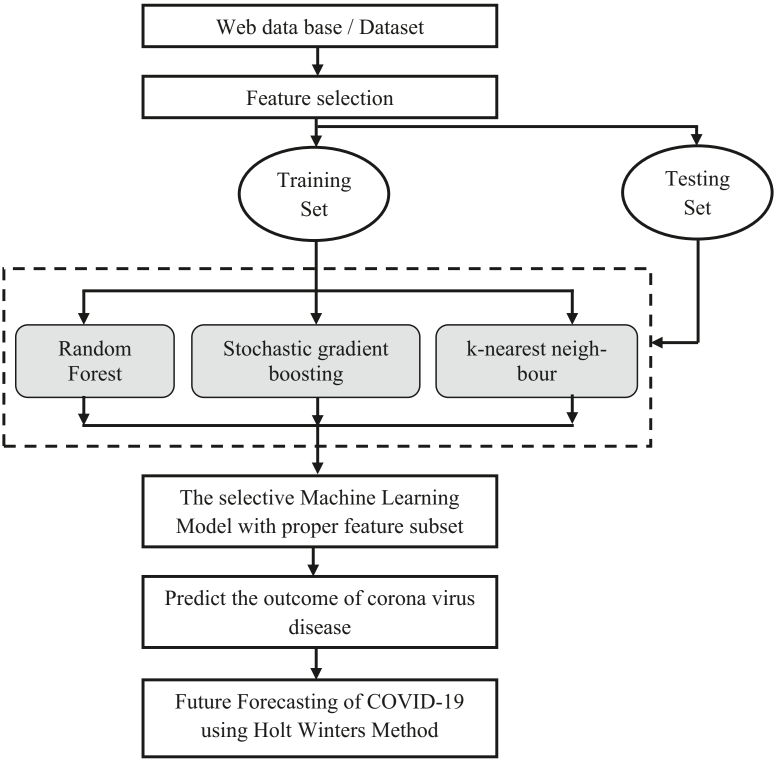
Web data base / Dataset
Feature selection
Training Set
Testing Set
Random Forest
Stochastic gradient boosting
k-nearest neighbour
The selective Machine Learning Model with proper feature subset
Predict the outcome of corona virus disease
Future Forecasting of COVID-19 using Holt Winters Method

Fig. 9.5 Machine learning model

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

TP = True positive; FP = False positive; TN = True negative; FN = False negative

Kappa statistics is to measure the agreement value K between two raters. Each raters classify N numbers of items into C classes of item categories. Here maximum value of K represents the complete agreement between raters.

$$K = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

Where,

p_o is observed agreement between raters.
p_e is hypothetical probability of chance agreement.

$$p_e = \frac{1}{N^2} \sum n_{k1} n_{k2}$$

Where, k is categories, N represent to observations and n_{ki} 0.Finally in fifth step, Future forecasting have been done by using Holt Winters Method [14], which is a triple exponential smoothing method. It is an appropriate technique only when trend and seasonality present in the time series dataset. It decomposes the data into three components: Level, Trend and Seasonality, which can be calculated as follows:

$$L_t = \alpha\left(\frac{D_t}{S_{t-s}}\right) + (1-\alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1}$$

$$S_t = \gamma\left(\frac{D_t}{L_t}\right) + (1-\gamma)S_{t-s}$$

Where,
L_t is level of series of period t.
$L_{t-1}$is smoothed value for period t-1.
D_t is actual value at period t.
T_t is trend at period t.
S_t is seasonality estimate at period t.
α is smoothing constant for Level ($0 \leq \alpha \leq 1$)
β is smoothing constant for Trend ($0 \leq \beta \leq 1$)
γ is smoothing constant for Trend ($0 \leq \gamma \leq 1$)
s is time period that select in seasonal cycle.

Next, forecast has been done by using following equation:

$$F_{t+m} = (L_t + mT_t) * S_{t+m-s}$$

Where,

m is number of period ahead to be forecast i.e 1 for this case.
F_{t+m} is Winter's forecast for m period ahead into future.

"Goodness of fit", which shows the accuracy of the forecast has been evaluated by four standard measures: Mean Absolute Deviation (MAD) [15], Mean Square Error (MSE) [16], Root Mean Square Error (RMSE) [17] and Mean Absolute Percentage Error (MAPE) [18].

$$MAD = \frac{\sum_{t=1}^{n} |D_t - F_t|}{n}$$

$$MSE = \frac{\sum_{t=1}^{n} (D_t - F_t)^2}{n}$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}\left(D_t - F_t\right)^2}{n}}$$

$$MAPE = \frac{\sum_{t=1}^{n}\frac{\left|D_t - F_t\right|}{D_t}}{n} \times 100\%$$

Where,

D_t is actual value at period t.
F_t is Winter’s forecast value for actual value.

9.3 Results

9.3.1 Overall Incidences

A time interval data were collected and verified with MoHFW. Figure 9.6 shows the average incidence from 30 January 2020 to 20 May 2020 in each region of India where corona virus deceased case were more than 1000.

In above histogram, we can see that most of the most of the regions in India fluctuate around the national average corona virus deceased incidence level. Among all the regions in the India, Maharashtra, Gujarat, Delhi, Rajasthan, Tamil Nadu and Uttar Pradesh regions have highest incidence, which is more than average of national average incidence. However, the incidences of corona virus decease in India, such as the Punjab, Haryana and Bihar regions, is low, less than one fourth of the national average.

Figure 9.7 show the nationwide trend of virus deceased cases from the beginning of COVID-19 incidences in India till 30 April 2020. The virus impact starts from mid march and mostly peaks in April month after second week.

9.3.2 Model Training

9.3.2.1 Random Forest, Stochastic Gradient Boosting and K-Nearest Neighbour

Here the training results evaluated using Random forest, stochastic gradient boosting and k-nearest neighbour technique respectively. Total 226 samples used to predict categorical outcome class (Deceased or recovered) based on 30 predictors. Here bootstrap sampling done using 226 samples with repeating 25 times each. Each table (Tables 9.1, 9.2 and 9.3) have different tuning parameters for different learning algorithms used here for training purpose. Random forest use single tuning

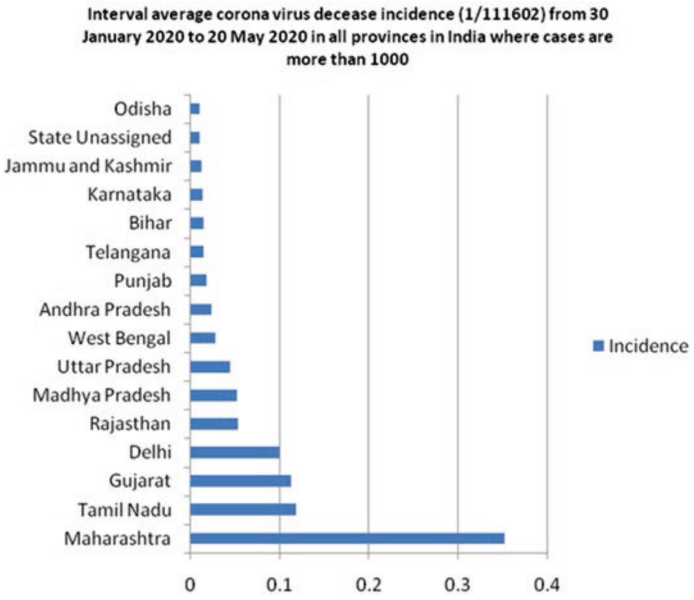

Fig. 9.6 Time interval average corona virus decease incidence (1/23,041) from 30 January 2020 to 20 May 2020 in all provinces in India where cases are more than 1000

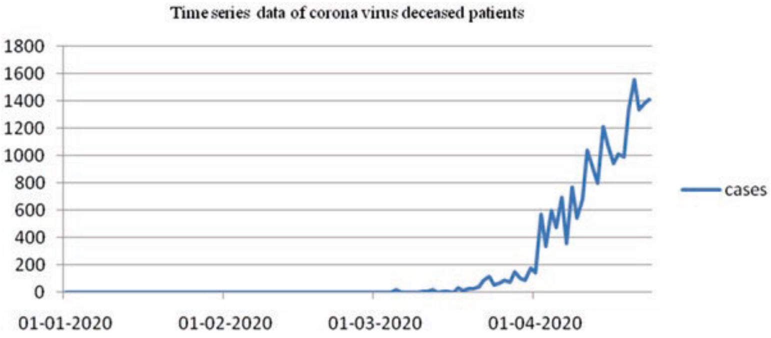

Fig. 9.7 Time series analysis of COVID 19 deceased patients

parameter m_{try} i.e. number of predictors which is used to split the classification. Stochastic gradient boosting uses two tuning parameters interaction depth and n trees along with constant shrinkage. K-nearest neighbour technique uses kmax parameter which means to maximum number of classes for classification of closed labelled neighbours.

The final value used for the model was mtry = 2.

The final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1.

Table 9.1 Tuning parameters of random forest technique

m_{try}	Accuracy	Kappa
2	0.92	0.80
3	0.91	0.78

Table 9.2 Tuning parameters of Stochastic gradient boosting

interaction.depth	n.trees	Accuracy	Kappa
1	50	0.8624	0.674
1	100	0.9064	0.782
1	150	0.9191	0.811
2	50	0.9142	0.799
2	100	0.9162	0.804
2	150	0.9178	0.808
3	50	0.9145	0.801
3	100	0.9163	0.805
3	150	0.9223	0.819

Table 9.3 Tuning parameters of k-nearest neighbour

kmax	Accuracy	Kappa
5	0.885	0.731
7	0.885	0.731
9	0.886	0.734

The final values used for the model were kmax = 9, distance = 2 and kernel = optimal.

9.3.2.2 Comparison of Model Training Algorithms

Table 9.4 shows the performance characteristics of all three types of training models. It seems that Random Forest model is best model as per performance among all.

Figure 9.8 visualize the boxplot diagram of comparison of performance characteristics.

9.3.3 Prediction

Table 9.5 shows the prediction of testing data, where actual outcome of patients are known and for the testing purpose we have changed it to unknown outcome and tried to predict it using trained model of all three learning algorithm. It has been

Table 9.4 Performance characteristics of training models

	Accuracy			Kappa		
	rf	gbm	knn	rf	gbm	knn
Min	0.86	0.87	0.84	068	0.73	0.63
1st Qu.	0.90	0.90	087	0.772	0.777	0.705
Median	0.92	0.91	0.88	0.82	0.80	0.73
Mean	0.92	0.91	0.88	0.82	0.80	0.73
3rd Qu.	0.94	0.92	0.90	0.86	0.84	0.76
Max	0.95	0.94	0.92	0.90	0.87	0.83
NA's	0	0	0	0	0	0

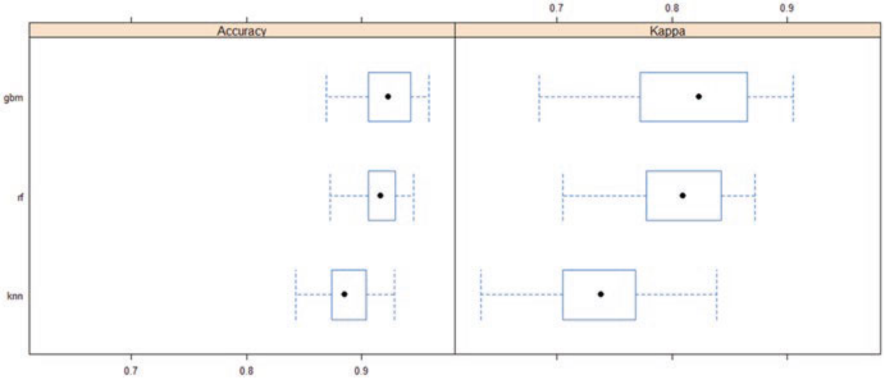

Fig. 9.8 Performance comparison of all three models used for training data

identified that Random forest and stochastic gradient boosting techniques predicated outcome with higher accuracy i.e. 90% in compare to k-nearest neighbour trained model.

Table 9.6 shows the predictive results of all three predictive models. Here we have 10 patients' details, which show current fitness condition of all patients along with their previous and current status of outcome.

Probability of predictions predicted by Random forest, stochastic gradient boosting and k-nearest neighbour techniques are shown in Table 9.7.

Table 9.5 Testing of outcome prediction of COVID-19 diseased patients

			Outcome prediction of COVID-19 diseased patients					
Case_ID	Actual status	Current Status (Testing)	Random Forest	Accuracy in %	Stochastic gradient boosting	Accuracy in %	k-nearest neighbour	Accuracy in %
36	Recovered	Unknown	Recovered	90%	Recovered	90%	Recovered	80%
57	Recovered	Unknown	Recovered		Recovered		Deceased	
135	Recovered	Unknown	Deceased		Deceased		Recovered	
430	Deceased	Unknown	Deceased		Deceased		Deceased	
558	Recovered	Unknown	Recovered		Recovered		Deceased	
855	Deceased	Unknown	Deceased		Deceased		Deceased	
860	Deceased	Unknown	Deceased		Deceased		Deceased	
863	Deceased	Unknown	Deceased		Deceased		Deceased	
1015	Recovered	Unknown	Recovered		Recovered		Recovered	
1293	Deceased	Unknown	Deceased		Deceased		Deceased	

Table 9.6 Outcome prediction of COVID-19 diseased patients

			Outcome prediction of COVID-19 deceased patients					
Case_ ID	Previous Status	Current status	Random Forest	Accuracy in %	Stochastic gradient boosting	Accuracy in %	k-nearest neighbour	Accuracy in %
4	Hospitalized	Deceased	Deceased	**70%**	Deceased	**70%**	Deceased	**70%**
136	Hospitalized	Recovered	Deceased		Deceased		Deceased	
165	Hospitalized	Deceased	Deceased		Deceased		Deceased	
285	Hospitalized	Recovered	Deceased		Deceased		Deceased	
1056	Hospitalized	Recovered	Recovered		Recovered		Recovered	
1289	Hospitalized	Deceased	Deceased		Deceased		Deceased	
1572	Hospitalized	Recovered	Deceased		Deceased		Deceased	
1879	Hospitalized	Deceased	Deceased		Deceased		Deceased	
3021	Hospitalized	Deceased	Deceased		Deceased		Deceased	
4534	Hospitalized	Deceased	Deceased		Deceased		Deceased	

Table 9.7 Probability of Outcome prediction of COVID-19 diseased patients

	Probability of Outcome prediction of COVID-19 diseased patients					
	Random Forest		Stochastic gradient boosting		k-nearest neighbour	
Case_ID	Deceased	Recovered	Deceased	Recovered	Deceased	Recovered
4	0.944	0.056	0.963	0.036	1	0
136	1.0	0.0	0.970	0.029	1	0
165	0.998	0.002	0.986	0.013	1	0
285	0.958	0.042	0.963	0.036	1	0
1056	0.016	0.984	0.116	0.883	0	1
1289	0.982	0.018	0.984	0.015	1	0
1572	0.960	0.040	0.965	0.034	1	0
1879	0.974	0.026	0.971	0.028	1	0
3021	0.998	0.002	0.984	0.015	1	0
4534	0.996	0.004	0.928	0.071	1	0

9.4 Forecasting Using Holt Winters Method

Initialisation

$$S_i = D_i\left[\left(\frac{1}{s}\right)\left(D_1 + D_2 + D_3 + \ldots + D_s\right)\right]$$

$$L_i = \frac{D_i}{S_1}$$

$$T_i = \frac{D_i}{S_1} - \frac{D_{i-1}}{S_s}$$

From the incidence trend in India, the corona virus incidence mostly starts increasing in March from 17th March 2020 onwards. Here we took India COVID-19 incidence data from 1 April 2020 to till 20 May 2020 for modelling and forecasting (Fig. 9.9).

Table 9.8 shows the smoothing parameters chooses for this forecast,

To validate the predictive capability of model used, we use the constructed model Holt Winters to predict the corona virus incidence from 1 April 2020 to till 20 May 2020 and then compared the actual deceased patients' numbers for that day with predicted values. The model fitted the data reasonably well (Fig. 9.10) with 15.17 MAPE. It shows that the Holt Winters Forecast model was Good [19] for forecasting COVID-19 deceased incidence in India.

We also construct a Holt Winters Future Forecast model (Fig. 9.11) for predicting next 11 days values for total cases of corona virus deceased patients from 21 May 2020 to 31 May 2020 (Table 9.9).

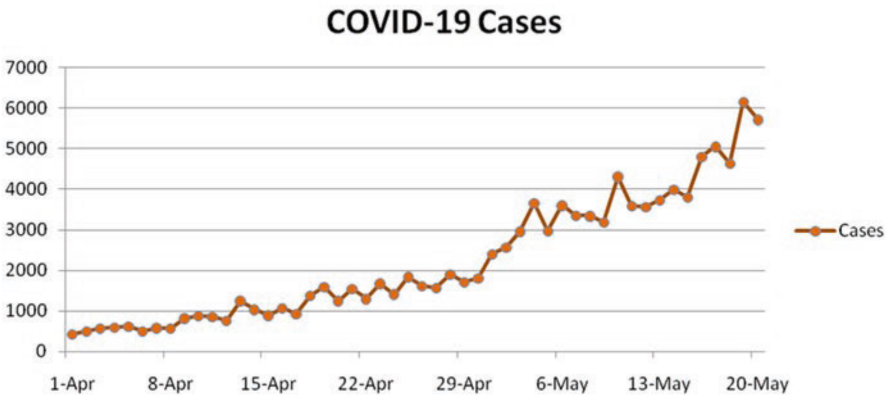

Fig. 9.9 Score of number of corona virus deceased patients increase pattern

Table 9.8 Smoothing parameters initial value

Parameters	
Alpha(α)	0.42
Beta(β)	0.42
Gama(γ)	0.99

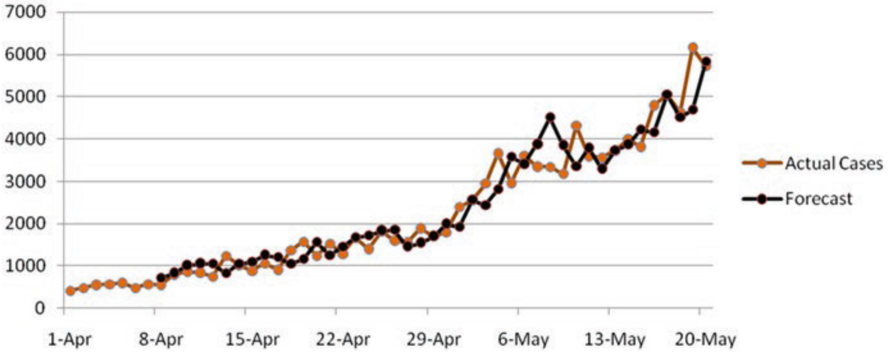

Fig. 9.10 Holt winters forecast model for COVID-19

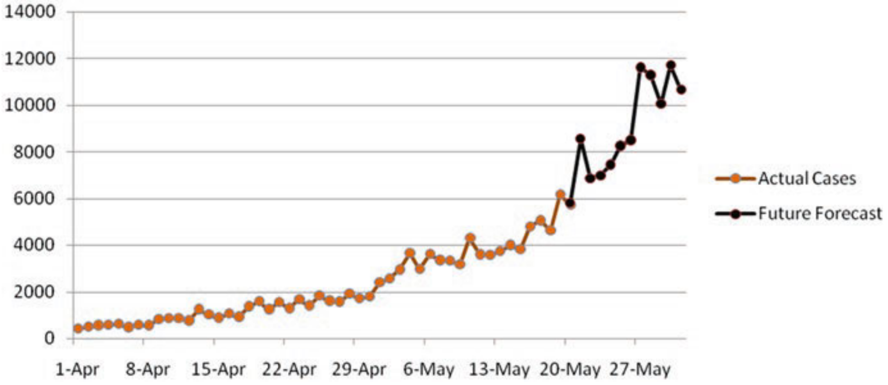

Fig. 9.11 Holt winters future forecast model for COVID-19

Table 9.9 Data table of future forecast of COVID-19

Sno	Date	Cases	Level	Trend	Seasonal	Forecast	Error	|Error|	Error^2	|%Error|
1	17-May	5049	5138.73	241.0228	0.98109	5048.511	−0.488568	0.48856	0.23869	0.009676
2	18-May	4628	5551.26	313.0570	0.83334	4511.104	−116.89527	116.895	13664.5	2.525826
3	19-May	6154	6623.54	631.932	0.92783	4688.935	−1465.0649	1465.06	2146415.	23.80671
4	20-May	5716	7068.89	553.565	0.80891	5819.891	103.89101	103.89	10793.3	1.817547
						Future forecast				
5	21-May					8552.1708				
6	22-May					6855.8701				
7	23-May					6979.9162				
8	24-May					7446.3633				
9	25-May					8255.0056				
10	26-May					8495.2111				
11	27-May					11596.051				
12	28-May					11280.093				
13	29-May					10042.598				
14	30-May					11694.996				
15	31-May					10643.848				

Table 9.10 Error measured by standard measures

Error measures	
MAD	325.7057
MSE	205281.2
RMSE	453.0797
MPSE	15.17755

9.5 Discussion

Corona virus decease put its impact directly or indirectly to all of us now days. The corona virus mutates according to environment easily as well as very quickly, which results a COVID-19 pandemic this year. This leads to social, economical and medical burden in country. We cannot repel the virus but we can minimize the impact of decease. In this behavioural study, we have introduced the statistical and forecasting model based on statistics for predicting number of cases of corona virus deceased patients. It will help to get the idea of research for COVID-19 prevention in future and also can guide for public health.

The Machine learning model used to train the data (Table 9.1, 9.2, 9.3 and 9.4) and the accuracy of the training model calculated (Fig. 9.8). Finally the outcome of COVID-19 decease tested (Table 9.5) and predicted (Table 9.6) with its probability of prediction (Table 9.7).

The Holt Winters method applied in different forecasting areas [20–21]. The national corona virus decease shows less seasonality and increasing fluctuation (Fig. 9.7). In this study, we select India as an example to construct a predictive model with corona virus decease incidence from 1 March 2020 to 20 May 2020 (Fig. 9.9). Then, we forecast the number of cases of COVID-19 and compared with actual number of cases (Fig. 9.10) and at last future forecast the total number of cases (Fig. 9.11) of corona virus deceased patients in India for next 9 days (Table 9.9) with reasonable accuracy (Table 9.10).

The accuracy of forecasting can increase by tuning the parameter α, β and γ more accurately. The Holt Winters method can forecast for short term with more accuracy. For this study, we have data with less time period, so the prediction and forecasting will deviate moderately.

Acknowledgments The author appreciates the covid19india.org: A volunteer-driven crowd sourced effort to track the corona virus in India, Our World in Data: a scientific online publication that focuses on large global problems such as poverty, disease, hunger, climate change, war, existential risks, and inequality and last but not the least, Ministry of Health and Family Welfare (MoHFW), Govt. of India for helping and supporting to get the COVID-19 incidence data for research and study.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Novel Coronavirus (2019-nCoV) situation reports. (2020). World Health Organization (WHO).
2. Novel coronavirus COVID-19 in China – Statistics & Facts. (2020). USA: Health & Pharmaceuticals: Statistics and facts on health and pharmaceuticals.
3. Roser, M., Ritchie, H., Ortiz-Ospina, E., & Hasell, J. (2020). *Coronavirus disease (COVID-19)*. Published online at OurWorldInData.org.
4. Home | Ministry of Health and Family Welfare | GOI. www.mohfw.gov.in. Retrieved 22 April 2020.
5. 19 Tracker: India. (n.d.). (2020), from https://www.covid19india.org
6. Ho, T. K. (1995). Random Decision Forests (PDF). In *Proceedings of the 3rd international conference on document analysis and recognition*, Montreal, QC, 14–16 August 1995 (pp. 278–282). Archived from the original (PDF) on 17 April 2016. Retrieved 5 June 2016.
7. Ho, T. K. (1998). The random subspace method for constructing decision forests (PDF). *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832–844. https://doi.org/10.1109/34.709601
8. Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis, 38*(4), 367–378.
9. Mason, L., Baxter, J., Bartlett, P. L., & Frean, M. (1999). Boosting algorithms as gradient descent (PDF). In S. A. Solla, T. K. Leen, & K. Müller (Eds.), *Advances in neural information processing systems 12* (pp. 512–518). MIT Press.
10. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression (PDF). *The American Statistician, 46*(3), 175–185. https://doi.org/10.1080/00031305.1992.10475879. hdl:1813/31637.
11. Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3).
12. BS ISO 5725-1. (1994). *Accuracy (Trueness and precision) of measurement methods and results – Part 1: General principles and definitions* (p. 1). ISO.
13. McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica., 22*(3), 276–282. https://doi.org/10.11613/bm.2012.031
14. Bermudez, J. D., Segura, J. V., & Vercher, E. (2007). Holt-Winters forecasting: An alternative formulation applied to UK air passenger data. *Journal of Applied Statistics, 34*, 1075–1090.
15. Pham-Gia, T., & Hung, T. L. (2001). The mean and median absolute deviations. *Mathematical and Computer Modelling, 34*(70), 921–936.
16. Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation* (2nd ed.). Springer. ISBN 978-0-387-98502-2.
17. Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688.
18. De Myttenaere, A., Golden, B., Le Grand, B., & Rossi, F. (2016). Mean absolute percentage error for regression models. *Neurocomputing, 192*, 38–48.
19. Klimberg, R. K., & Ratick, S. (2017). Development of a practical and effective forecasting performance measure. In *Advances in business and management forecasting* (pp. 103–118). https://doi.org/10.1108/s1477-407020170000012007