Prediction of COVID 19 Spread Technique using Dynamic Bayesian Network Model and Hidden Markov Model

Ruchi Sharma, Master of Technology (Software Systems), Birla Institute of Technology and Science (BITS) Pilani, INDIA

ABSTRACT

This paper aims to supply a modelling efforts to gauge the spread of COVID 19 pandemic specifically in INDIA. The information which has been taken into consideration for evaluating this infection is from 14th March 2020 to 19th April 2020. The evaluation is taken into account using Dynamic Bayesian Network Model and Hidden Markov Model only. As COVID-19 virus, has a time period of 14 days with people showing symptoms a mean of 5 days after catching it, consistent with research which backs up the findings of past studies on the new corona virus. During this paper, it's been assumed that the effect of current reproduction rate are going to be seen on day d, d+5. The sudden outbursts are captured using offset of values for all the states of INDIA. Accordingly, the paper prediction is for everywhere INDIA. Similarly, single states results also can be predicted using this system. Worldwide, many attempts are made to predict outburst of COVID-19 and within the model, described during this paper, turning point isn't predicted, as cases in India are still rising. The developed model is predicated cumulative infections and rationalization is administered for the population of varied regions, while predicting infections for whole India. Assigning a decay constant at this stage are going to be a premature exercise and keeping that in mind, exponential model predicts that in India, the reproduction rate from 14th March, 2020 to 19th April, 2020 is approximately 0.802716 per day. If we calculate the Prediction for spreading of COVID-19 then consistent with data, around 18,689 people will increase on every d+5 days. The figure of 8,41,005 Lakhs are going to be increased within the existing data by 31st July, 2020.

KEYWORDS

COVID-19, India, Modelling, Current Trend, Prediction, Infection, reproduction rate, incubation period.

1. Introduction :

Coronavirus disease (COVID-19) is a communicable disease caused by a newly discovered coronavirus. Most of the people infected with the COVID-19 virus will experience mild to moderate respiratory disease and recover without requiring special treatment. Older people, and people with underlying medical problems like disorder, diabetes, chronic respiratory illness, and cancer are more likely to develop serious illness. The simplest thanks to prevent and hamper transmission is to be informed about the COVID-19 virus, the disease it causes and the way it spreads. Protect yourself et al. from infection by washing your hands or using an alcohol based rub frequently and not touching your face. Although, many precautions are taken and advised to stay ourselves shielded from the COVID 19 but still the cases are increasing day by day. This paper is an effort to aware of the technique of calculating the numbers of patients which can increase. Similarly, during this paper, consistent with data the calculations are made up to 31st July 2020.

2. Dynamic Bayesian Network Model :

Bayesian networks are a kind of Probabilistic Graphical Model which can be used to build models from data and/or expert opinion.

They can be used for a good range of tasks including prediction, anomaly detection, diagnostics, automated insight, reasoning, statistic prediction and deciding under uncertainty.

Figure 1 below shows these capabilities in terms of the four major analytics disciplines, Descriptive analytics, Diagnostic analytics, Predictive analytics and Prescriptive analytics.

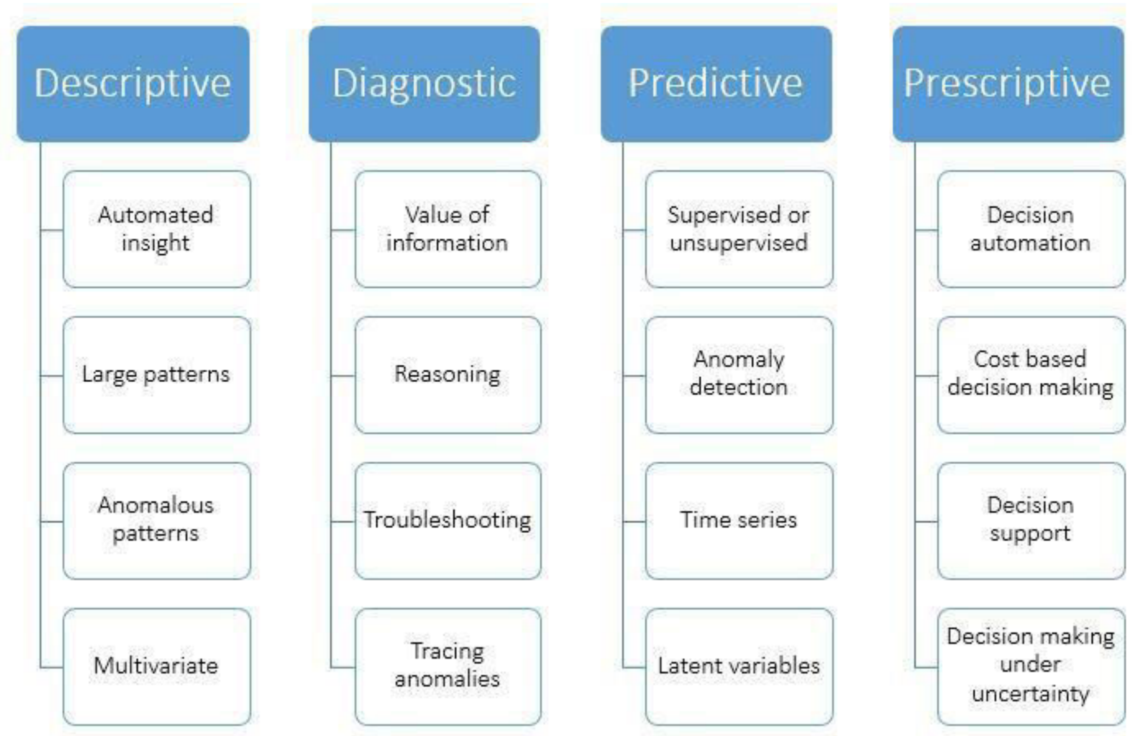


Figure 1

Prediction is that the process of calculating a probability distribution over one or more variables whose values we might wish to know, given information (evidence) we've about another variables.

3. Inputs and outputs :

The variables we are predicting are referred to as Output variables, while the variables whose information we are using to form the predictions are referred to as Input variables.

In statistics, Input variables are often called predictor, explanatory, or independent variables, while Output variables are often called Response or dependent variables.

Assuming State variables as

- Rd = Current reproduction rate

Evidence Variables as

- Nd = Total New Cases
- Sd = Total Active Cases

4. Prediction Strategy :

Assumptions as stated below:

- Nd -> Number of new patients who are tested positive for Corona.
- Sd -> Total number of active patients who are tested positive for Corona.
- Rd -> Number of patients recovered on day 'd'.
- Dd -> Number of deceased patients on day 'd'.

Reproduction Rate Rd – within a wait period of 5 days.

Reproduction rate calculation formula:

- Rd = (N(d+1)+N(d+2)+N(d+3)+N(d+4)+N(d+5))/Sd
- Rd[0.1:2.1] = round(Rd,1)
- P[Rd] = fq(Rd[0.1:2.1])/total observations [Note: total observations = 60 days from 14th march to 12th May]

Table 1: showing the calculations made on above assumptions

Date	Confirmed	Recovered	Deceased	Total New Case (Nd)	Total Active Cases (Sd)	Rd	P(Rd)	S(d+5)
14-Mar-20	81	9	2	70	140	0.864198	349.4571	102.7877
15-Mar-20	27	4	0	23	46	0.608696	164.2857	22.02
16-Mar-20	15	1	0	14	28	0.933333	62.14286	22.49333
17-Mar-20	11	1	1	9	18	0.818182	48.88889	12.38727

18-Mar-20	37	0	0	37	74	1	148	64.38
19-Mar-20	27	5	1	21	42	0.777778	123.4286	27.20667
20-Mar-20	58	3	0	55	110	0.948276	238.3273	90.01034
21-Mar-20	78	0	0	78	156	1	312	135.72
22-Mar-20	67	0	3	64	128	0.955224	274.2813	105.6287
23-Mar-20	102	12	2	88	176	0.862745	440.4545	128.9631
24-Mar-20	64	5	1	58	116	0.90625	269.2414	90.045
25-Mar-20	90	3	1	86	172	0.955556	368.3721	141.9956
26-Mar-20	65	7	5	53	106	0.815385	289.434	72.65077
27-Mar-20	164	26	3	135	270	0.823171	726.4593	187.1561
28-Mar-20	143	9	5	129	258	0.902098	603.0388	199.2013
29-Mar-20	110	5	3	102	204	0.927273	457.2549	162.6436
30-Mar-20	187	49	13	125	250	0.668449	933.504	134.6123
31-Mar-20	309	15	6	288	576	0.932039	1281.063	461.9744
01-Apr-20	424	16	9	399	798	0.941038	1749.133	647.2081
02-Apr-20	486	22	14	450	900	0.925926	2021.76	716.3333
03-Apr-20	560	38	14	508	1016	0.907143	2354.646	789.5771
04-Apr-20	579	56	13	510	1020	0.880829	2472.671	765.8456
05-Apr-20	609	43	22	544	1088	0.893268	2581.533	830.4352
06-Apr-20	484	65	16	403	806	0.832645	2130.561	566.3316
07-Apr-20	573	75	27	471	942	0.82199	2540.178	651.8541
08-Apr-20	565	96	20	449	898	0.79469	2551.938	596.8919
09-Apr-20	813	70	46	697	1394	0.857319	3522.611	1013.882
10-Apr-20	871	151	22	698	1396	0.801378	3915.756	937.2433
11-Apr-20	854	186	41	627	1254	0.734192	4034.367	757.6568
12-Apr-20	758	114	42	602	1204	0.794195	3424.85	799.6911
13-Apr-20	1243	112	27	1104	2208	0.888174	5285.002	1674.048
14-Apr-20	1031	167	37	827	1654	0.802134	4632.643	1111.709
15-Apr-20	886	144	27	715	1430	0.806998	3967.793	968.1068
16-Apr-20	1061	258	26	777	1554	0.732328	5019.609	936.0177
17-Apr-20	922	273	38	611	1222	0.66269	4626.599	650.9469
18-Apr-20	1371	426	35	910	1820	0.663749	6873.079	971.4233
19-Apr-20	1580	388	38	1154	2308	0.73038	7486.516	1385.676
Total	17305	2854	560	13891	27782	0.802716	77726.12	18689.4

On considering the information and assumptions stated above,

Assuming, that the effect of Rd of days d are going to be seen on day d + 5. Taking the recovery rate (patients getting cured) as 23%, and assume that they're discharged on day d + 21 is that if they were confirmed on day d. Similarly, assuming that 3% of all patients die on day d + 14 after confirmation on day d. Thus,

$$S_{d+5} = S_d + \sum_{k=0}^{k=4} R_{d-k} S_{d-k} - \sum_{k=0}^{k=13} 0.03 N_{d-k} - \sum_{k=0}^{k=20} 0.23 N_{d-k}.$$

Total new cases (Nd) = Total confirmed – Total Recovered – Total Deceased

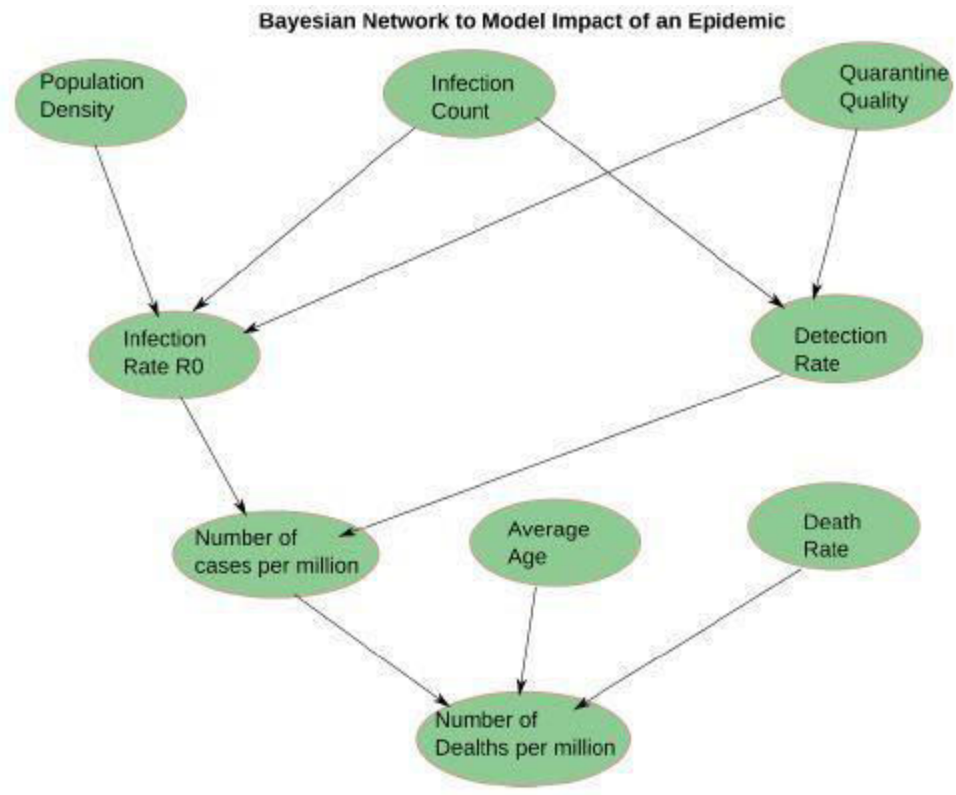
Total Active Cases (Sd) = Total New Cases + Total Confirmed – Total Recovered – Total Deceased

Reproduction Rate (Rd) = (Total Sd – Total Nd) / Total Confirmed S(d+5) =

Sd + Rd* Nd – 0.03 * Nd – 0.23* Nd

This is how the information has been calculated in Sheet.

5. Bayesian Network :



6. Time-Series Analysis :

A statistic may be a series of knowledge points indexed (or listed or graphed) in time order. Most ordinarily, a statistic may be a sequence taken at successive equally spaced points in time.

Time series analysis may be a statistical technique that deals with statistic data, or analysis. Statistic data means data is during a series of particular time periods or intervals. The information is taken into account in three types:

- Statistic data: a group of observations on the values that a variable takes at different times.
- Cross-sectional data: Data of 1 or more variables, collected at an equivalent point in time. Pooled data: a mixture of your time series data and cross-sectional data.

Values taken by a variable over time and tabulated or plotted as chronologically ordered numbers or data points. To yield valid statistical inferences, these values must be repeatedly measured, often over a four to 5 year period.

Time series contains four components:
(1) differences due to the season that repeat over a selected period like each day , week, month, season, etc.,
(2) Trend variations that move up or down during a reasonably predictable pattern,
(3) Random variations that don't fall into any of the above three classifications.

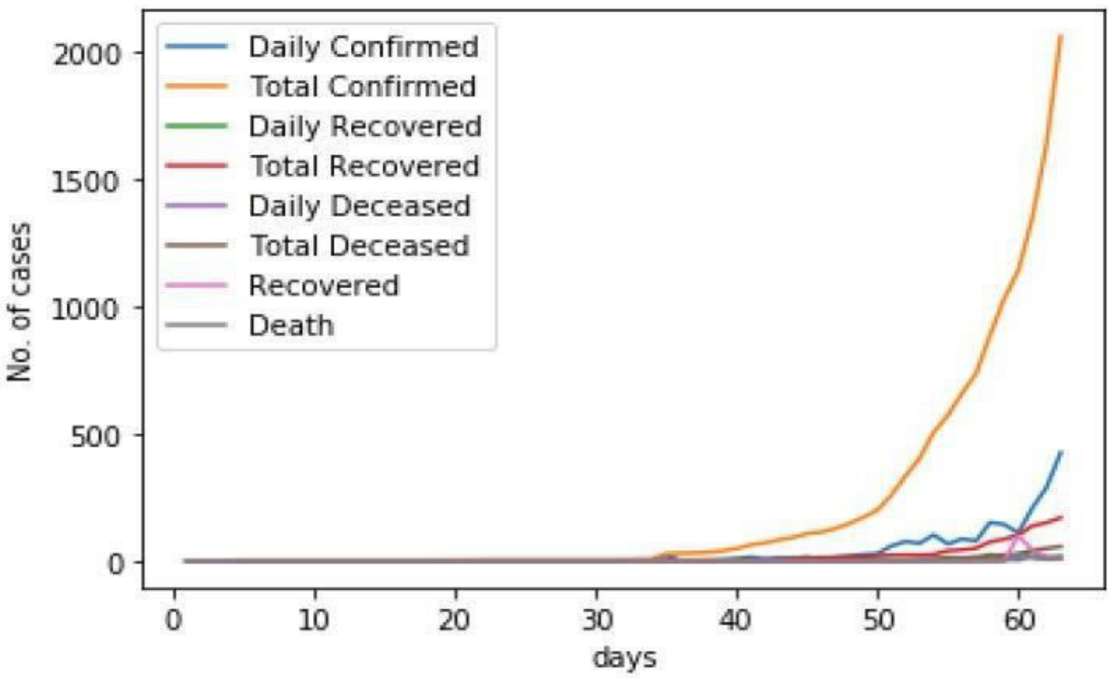
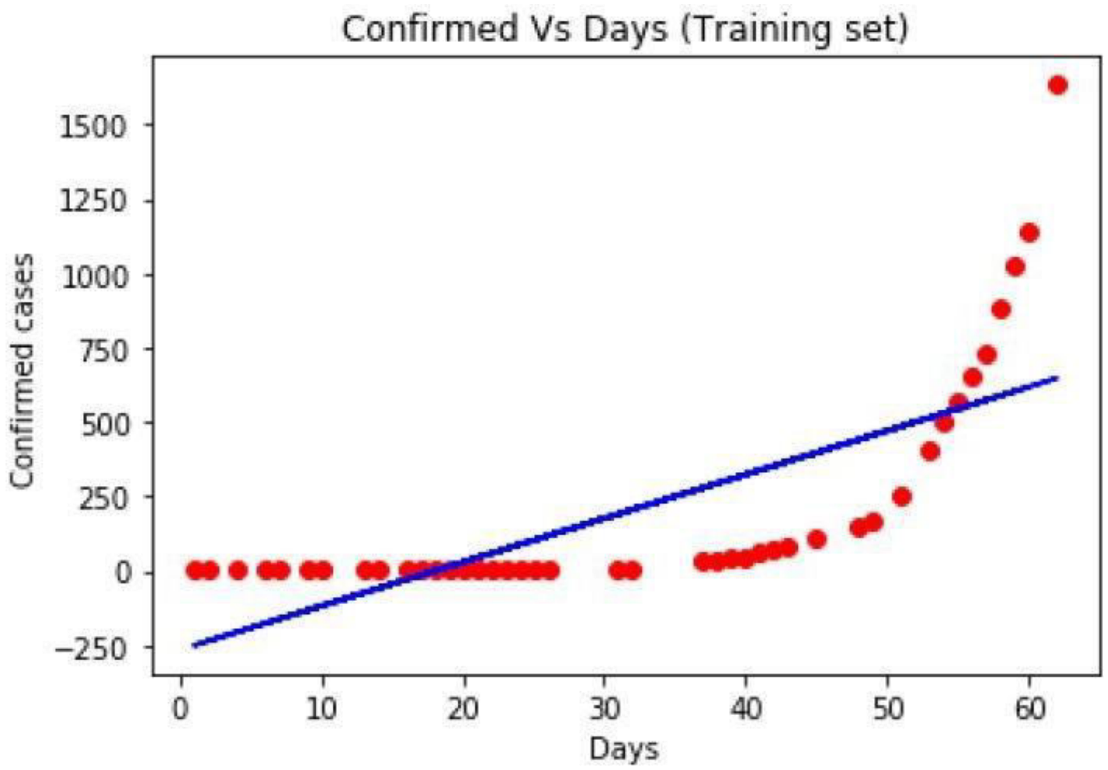
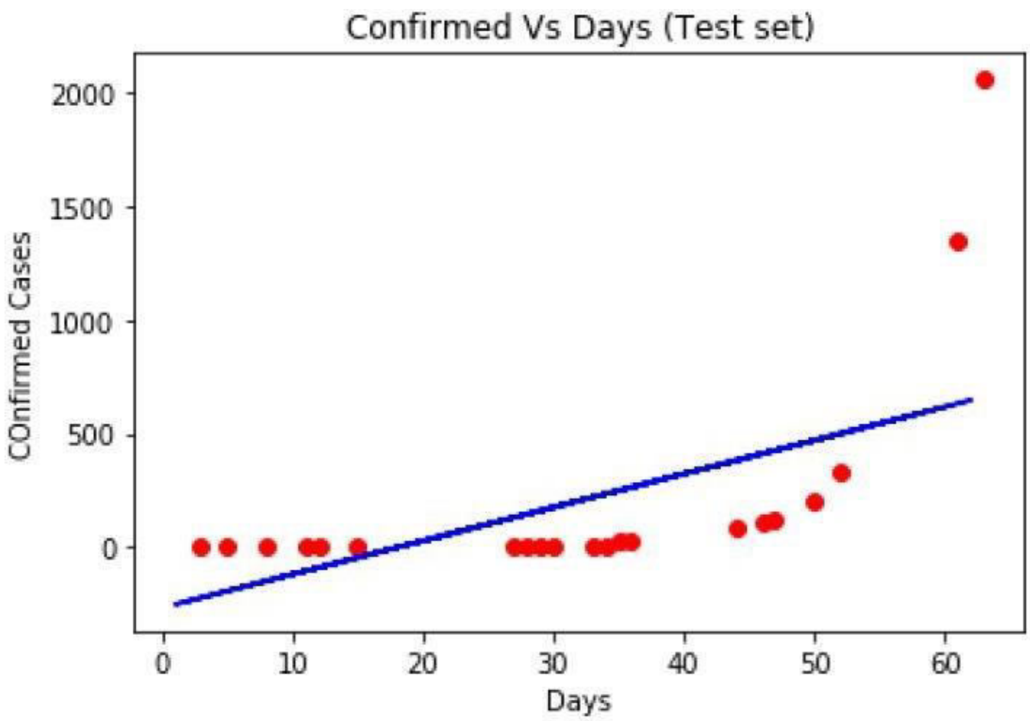


Fig : No. of cases confirmed, deceased and recovered


Confirmed Vs Days (Training set)
Confirmed cases
Days
1500
1250
1000
750
500
250
0
−250
0
10
20
30
40
50
60


Confirmed Vs Days (Test set)
COnfirmed Cases
Days
2000
1500
1000
500
0
0
10
20
30
40
50
60

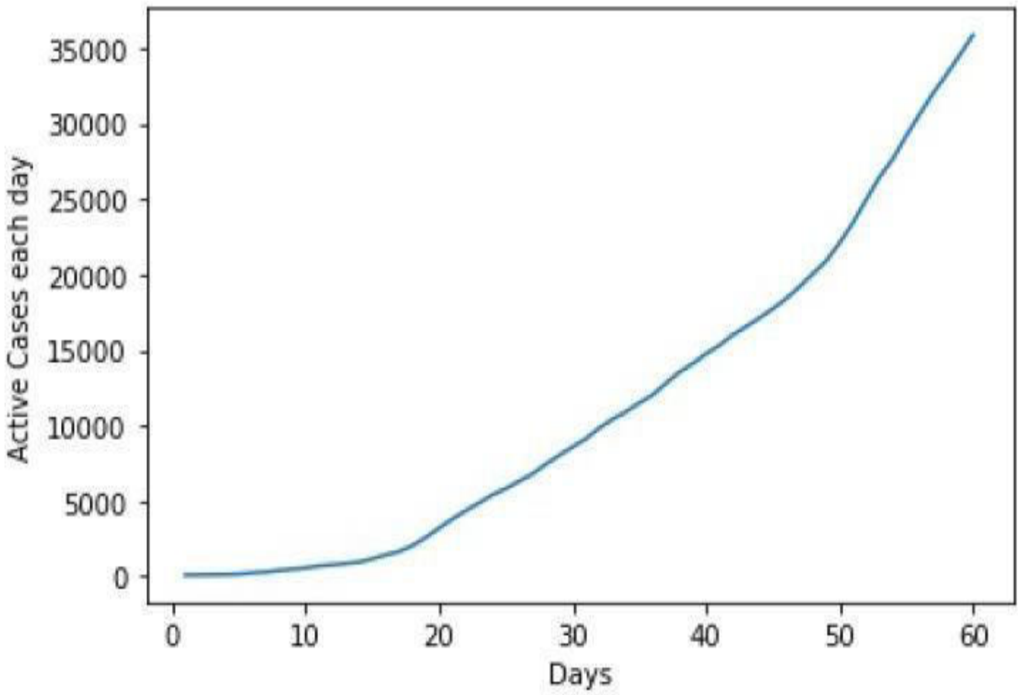


Fig : Active cases per day

7. ***Conclusion :***

The Bayesian network, a machine learning method, predicts and describes classification supported the Bayes theorem. Bayesian networks are widely utilized in medical decision support for his or her ability to intuitively encapsulate cause and effect relationships between factors that are stored in medical data.

A Hidden Markov Model (HMM) describes a process over time. This process is characterized by a sequence of states that change accidentally. The phase change is modelled with a variate, where the transition probability between the states only depends on the present state, not on the past states.

Combination of both the models are often wont to predict the analysis of wide spread of Corona Virus.

8. ***References:***

1. Stuart Russell and Peter Norvig, "Artificial Intelligence – A Modern Approach", Pearson Education, Third Edition
2. Elaine Rich and Kevin Knight, "Artificial Intelligence", Tata McGraw Hill Publishing Company, New Delhi, 2003
3. Ryszard S. Michalski, Jaime G. Carbonell and Tom M. Mitchell, "Machine Learning: An Artificial Intelligence Approach", Elsevier,2014
4. Dan W Patterson, "Introduction to AI and Expert Systems", Prentice Hall of India, New Delhi, 2010
5. A.M. Turing(1950) Computing Machinery and Intelligence Mind LIX (236): 433-460
6. Michael Skirpan, Micha Gorelick, The Authority of Fair in Machine Learning
7. Christoph Molnar, Interpretable Machine Learning
8. https://www.statisticssolutions.com/time-series-analysis/#:~:text=Time%20series%20analysis%20is%20a,particular%20time%20periods%20or%20intervals.
9. https://www.bayesserver.com/docs/techniques/prediction
10. https://towardsdatascience.com/implementation-of-naive-bayes- classifier-with-the-use-of-scikit-learn-and-ml-net-d1012e105fbc
11. https://www.who.int/bulletin/volumes/90/6/11-086009/en/
12. https://link.springer.com/article/10.1007/s00477-018-1547-8
13. https://www.bayesserver.com/docs/introduction/bayesian-networks

****