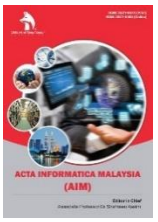
ZIBELINE INTERNATIONAL™ PUBLISHING

ISSN: 2521-0874 (Print)
ISSN: 2521-0505 (Online)
CODEN: AIMCCO

Acta Informatica Malaysia (AIM)

DOI: http://doi.org/10.26480/aim.01.2022.34.38



CrossMark

RESEARCH ARTICLE

CONVERGENCE OF MACHINE LEARNING AND STATISTICS TO PREDICT COVID-19 EVOLUTION

Ajay Singh[a], Madan Mohan Gupta[b]

[a] BBS College of Engineering and Technology, Prayagraj, India
[b] Department of Statistics, Meerut College, Meerut, India
** Corresponding Author Email: madangupta22@gmail.com*

ARTICLE DETAILS

Article History:

Received 13 July 2022
Accepted 16 August 2022
Available online 19 August 2022

ABSTRACT

The effect of the Covid pandemic is not restricted to sickness and death but also extends to socioeconomic concerns. The statistics-based assessment of covid data presented is the measure of the damages that happened to citizens of a country and the required actions taken towards those damages. This study aims to analyse the consequences of numerous considerations on the deaths due to the pandemic. The paper presents the statistics-based processing of COVID-19 data using logistic regression (LR) and decision tree (DT). Results are compared using the logistic regression algorithm (based on statistics) and the decision tree algorithm (based on machine learning). This study presented the predictive abilities of logistic regression and decision tree approaches and observed better results for the decision tree method. An accuracy of 94.10% for decision tree and 93.90% for logistic regression, respectively observed. It is also observed that highly populated countries are inclined to have more corona cases than those with low density. More females die than males, and a greater number of deaths are observed in cases of age greater than sixty-five years. The experimental data is gathered from the official website of the world health organization (WHO) between January 2020 and June 2020. The results presented are promising for the reported studies.

KEYWORDS

Covid-19, Machine Learning, Statistical Analysis, Decision Tree, Logistic Regression.

1. INTRODUCTION

The pandemic badly affected the economy, employment, tourism, hotels, restaurants, local and foreign travelling, etc. Consequently, the gross domestic product (GDP) just collapsed almost in all countries, especially the European countries. When the data is large, it becomes crucial to use prediction methods to observe the behaviour of the pattern for timely necessary action. Many useful artificial intelligence and statistical tools provide predictions with excellent accuracy (Hou, 2021). Advance prediction and analysis of pandemic data will help take on-time preventive measures that can minimize economic losses and save lives. The prediction and analysis are important ways to forecast future event values. Artificial Intelligence techniques are very helpful in tracking infectious diseases, forecasting the early signs of the infection, and speculating the peak level of covid pandemic spread. Machine learning, a subset of artificial intelligence, may remarkably predict the spread of the Covid virus using predictive analytics. A machine learning technique may help mine the data for better estimating and predicting the COVID-19 infection. Machine Learning uses self-learning from past experiences without explicitly coding the training models (Ahmed and Mahesh, 2021).

An ordinal classification or prediction is an essential alliance of real-time challenges, which are involved in keeping both the characteristics of a structure to be predicted and the group itself. To ensure this, the decision trees are flexible prediction models to add new features without additional parameters. The machine learning models may provide either a significant outcome from the input data or a numerically predicted outcome. The decision trees contain the number of nodes and branches. The decision tree predicts and computes the probability of the input features based on categorical classes. The decision tree predicts the target class by concluding the if-then rules. Logistic regression as a statistical method permits the effect of multiple independents on one binary dependent variable for some prediction or classification (Vershese et al., 2021). Logistic regression is used to evaluate the relation among the binary dependent variables. Every independent variable by controlling the other independent variables provides a prediction method.

2. LITERATURE REVIEW

The confirmation of the spread of COVID-19 was first observed in Italy by the end of January 2020, when a tourist from China visited the city of Rome and detected positive (https://www.corriere.it/cronache/20_gennaio_30/coronavirus-italia). After seven days of that, another male tourist returned to Italy from China, and he was admitted to the hospital and detected as the third case in Italy and on February 21, 2020, more cases were reported, with 16 confirmed cases in a region of Northern Italy (https://www.corriere.it/cronache/20_febbraio_22/coronavirus-italia-nuovi-contagi-lombardia-veneto; Anzolin and Amante, 2020). After that, 60 more cases with first death were reported due to the pandemic (https://www.corriere.it/cronache/20_febbraio_22/coronavirus-italia-nuovi-contagi-lombardia-veneto). During the first week of March 2020, COVID-19 started to spread across the country of Italy

Quick Response Code	Access this article online	
	Website: www.actainformaticamalaysia.com	**DOI:** 10.26480/aim.01.2022.34.38

Cite The Article: Ajay Singh, Madan Mohan Gupta (2022). Convergence of Machine Learning and Statistics To Predict Covid-19 Evolution. *Acta Informatica Malaysia, 6(1): 34-38*

(https://www.avvenire.it/attualita/pagine/coronavirus-aggiornamento-5-marzo-2020). As of July 19, 2020, 12440 active cases were reported in Italy. At the pandemic's peak, the observed majority of active cases were from Italy only, with a total of 244434 confirmed cases with 35045 deaths. Also, on average, 578 deaths per million citizens along with 196949 cases of survival were reported. During mid of July 2020, a total of 3741000 citizens of Italy tested positive (http://www.salute.gov.it/imgs/C_17_notizie_4623_0_file.pdf).

During January and February 2020, there were several cases when tourists were admitted to the hospital and could not survive. The pandemic spread and transmitted from one region to another (Coronavirus: First death confirmed in Europe, 2020). The main reason for the spread of COVID-19 across Europe was also due to the annual social meeting of nearly 2500 persons in the Christian Open-Door Church in mid of February 2020. Around half of the visitors were understood with corona infection (Zayet et al., 2020). As of June 2021, 29640 deaths, 160377 infected cases, and 74372 cases of recovered persons were reported in hospitals in France. With the impact of Covid-19, France faced a major recession, which affected the production volume, unstable supply chain management, and increased anxieties about the accessibility of raw materials.

The first Covid-19 positive case in Germany was reported on January 27, 2020 (Corona-Pandemie 900 Milliardengegen die Angst, 2020). After that, many instances of the Italian epidemic were reported in Baden-Wurttemberg by the end of February 2020. A huge gathering celebrating a festival in Heinsberg reported the first death in the middle of March 2020 (https://www.kreis-heinsberg.de/aktuelles/aktuelles/?pid=5136). Continuous involvement of other groups in social gatherings caused the wide spread of the pandemic (Koch, 2020). As of June 10, 2020, the Robert Koch Institute stated that the total active cases were 184861, of which 41.20% went in the age group of 35 to 59 years, and the count of deaths went 8729, in which 85.30% were the senior citizens ageing 70 to 100 years (Fisher, 2020). A post-Ad hoc genomic analysis presented that about fifteen different varieties of the coronavirus came into the knowledge and infected the population by the middle of February 2020.

3. METHODS

This study used data from the World Health Organisation database 0f 4countries of Europe (Italy, France, Germany, and Switzerland) from the period January 2020 to June 2020 with a correlation to attributes: the real growth in gross domestic progress, nation's health expenses per head, and expected senior citizen dependency ratio per hundred citizens. A count of total jobless youth cases (age group18 to 24 years) was estimated, and new cases (NewCases) were reported. The covid cases prediction analysis uses models and methods based on decision-tree and logistic regression (Wu et al., 2020; Alsulami, 2020a).

3.1 The Decision Tree Model

Machine learning has several applications in classification and prediction. As a machine learning method, the decision tree classifier is used as a supervised-learning technique as a machine learning paradigm (Almalki et al., 2021; Alsulami, 2020b). The model is developed by feeding input training data on accurate output descriptions. The model based on the decision tree is trained using the patterns from the training dataset (Abo-Dahad et al., 2020; Abdel-khalek et al., 2021). Then the remaining data is given as input for determining how the model is performing (Islam et al., 2020). This model includes three major components (nodes, edges, and leaves) as follows (Dharmapuri et al., 2020; Prkash et al., 2020):

(a) A node representing the decisions over the value of the specific characteristic

(b) An edge representing response from the nodes to create a connection to successive nodes

(c) The leaves representing the exit levels to collect results

3.2 The Logistic Regression Method

The Logistic Regression comprises the true regression equation contained by the sigmoid function (Alsulami, 2020; Elhag and Abu-Zinadah, 2020; Abu-Zinadah and Binkhamis, 2020; Alsulami, 2021). The Logistic Regression takes place using the formula:

$$f = \frac{1}{1 + \exp\left(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n)\right)}$$

The sigmoid function is used for mapping the values in the discrete range [0, 1].

3.3 The Data Analysis

3.3.1 Decision Tree

For decision tree-based analysis, the growth process with the dependent and independent variables of the method are used, which are presented in Table 1.

Table 1: Variables of the Proposed Approach

Conditions	Growing Technique	Chi Squared Automatic Interaction Detection (CHAID)
	Dependent Variable	Death
	Independent Variables	NewCases, national health expenses per citizen, Unemployment, young total age group 18-24) (as per International Labour Organization), GDP growth, projected senior citizen dependency ratio per 100 persons
	Validation	No
	Maximum Depth	Three
	Parent Nodes	Hundred
	Child Nodes	Fifty
Results	Independent Variables	New Cases, projected senior citizen dependency ratio per hundred persons
	Number of Nodes	Eight
	Number of Leaf Nodes	Five
	Tree Depth	Two

Table 2: Prediction of Accuracy

Observed	Cases Prediction Accuracy		
	Survived	Died	Percentage
Survived	133	15	89.90%
Died	14	330	95.90%
Overall (%)	29.90%	70.10%	94.10%

Table 2 presents the classification accuracy as 94.10% of the training input samples. The table presents the predicted and observed data. The analysis is done using the growing method of Chi-Squared Automatic Interaction Detection (CHAID), which is a useful method for summarizing the data and dependent variable as death.

Figure 1 presents the tree presentation having seven nodes with different categories of data predicting the percentage of no-death and deaths.

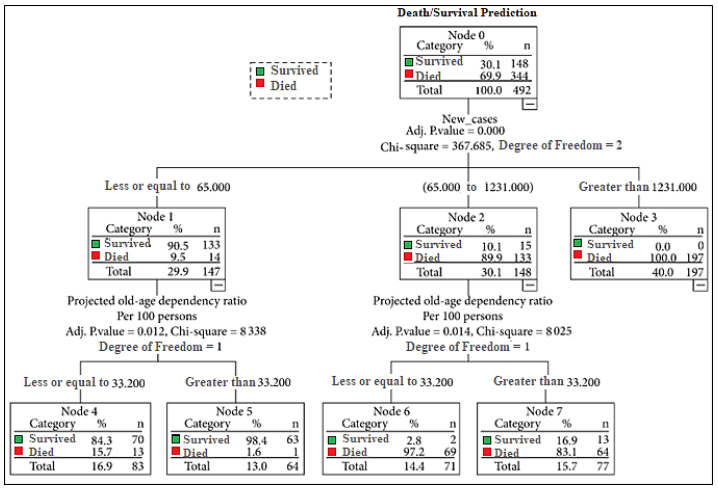


Figure 1: The decision tree diagram for death and survival prediction

3.3.2 Logistic Regression

Logistic Regression as a proposed statistical method for predicting deaths is based on the following five factors:

(a) F1: The estimated old-age dependency ratio

(b) F2: The GDP growth

(c) F3: The employment status

(d) F4: The health expenses of the country (per capita)

(e) F5: The NewCases reported

The following equation is used for Logistic Regression:

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

where P represents the probability, e is the logarithm base, *and* a and b are the model parameters. The value of a becomes equal to P with $X = 0$, and b fine-tunes on probability change on the update of X with a single unit.

The analysis based on 492 input samples with no excluded data is summarized in Table 3.

Table 3: Details of Case Processing Based on Included and Excluded Cases

	Unweighted Cases (as per Step 1 below)	Count	Percentage
Chosen cases	Inclusion in the analysis	492	100%
	Exclusion in the analysis	0	0%
	Total number of cases	492	100%
	Excluded cases	0	0%
	Total cases	492	100%

The process of analysing the data is as following:

Step 1. Input variables in NewCases

Step 2. Input variables in national health expenses per capita

Step 3. Input variables in Unemployment, young total age group 18-24 (as per International Labour Organization)

Step 4. Input variables in GDP growth

Table 4: Method Variables

Method Coefficients →		Bias (B)	S.E.	Wald	D. F.	Statistical Significance (p-values for each variable)	Exp(B) Exponential of B
Step 1 (Variables as per Step 1 above)	New Cases	0.008	0.001	70.065	1	0.00	1.008
	Constant	-1.699	0.213	63.896	1	0.00	0.183
Step 2 (Variables as per Step 2 above)	New Cases	0.009	0.001	65.844	1	0.00	1.009
	Health expenses by country per capita	-0.001	0.000	6.211	1	0.01	0.999
	Constant	0.130	0.736	0.031	1	0.86	1.139
Step 3 (Variables as per Step 3 above)	New Cases	0.009	0.001	65.992	1	0.00	1.009
	Unemployment, young total age group 18-24) (as per International Labour Organization)	-0.269	0.132	4.134	1	0.04	0.764
	Health expenses by country per capita	-0.002	0.001	7.218	1	0.01	0.998
	Constant	6.75	3.33	4.11	1	0.04	850.54
Step 4 (Variables as per Step 4 above)	New Cases	0.010	0.00	61.49	1	0.00	1.01
	Percentage GDP	1.372	0.49	7.91	1	0.01	3.945
	Unemployment, young total age group 18-24) (as per International Labour Organization)	-0.491	0.16	9.50	1	0.00	0.612
	Health expenses by country per capita	-0.004	0.001	13.50	1	0.00	0.996
	Constant	26.59	7.96	11.16	1	0.00	3.56×10^{11}

The method variables concerning their statistical significance are mentioned in Table 4. The following terminologies are used:

- B represents the Statistical bias, a systematic tendency to create distinctions between results and facts. The bias in statistical methods exists in several data analysis processes, selected estimation, and the circumstances when the data is analyzed.
- S.E. represents the standard error, which measures the accuracy of the predictions.
- Wald is a statistical parameter constraint about the weighted distance between unrestricted estimation and hypothesized value that corresponds to the null hypothesis.
- D. F. is the degree of freedom.
- Exp(B) is also known as the odds ratio. It represents the predicted shift in odds with a unit increase in the predictor. When the value of Exp(B) is less than 1, values of the variable corresponding to the decreasing odds of the event's occurrence are increased

Table 5 presents the compiled test results of the method constants based on the chi-square test. In case the *P*-value is<0.001, the null hypothesis is disallowed. The observation suggests that the method is the best fit for the data.

Table 5: Compilation of Method Constants

	Method Constants	Step	Block	Model
Step 1	Chi-square	391.93	391.93	392.94
	Degree of Freedom	1	1	1
	statistical significance	0.00	0.00	0.00
Step 2	Chi-square	6.56	400.01	400.01
	Degree of Freedom	1	2	2
	statistical significance	0.01	0.00	0.00
Step 3	Chi-square	4.27	402.78	402.78
	Degree of Freedom	1	3	3
	statistical significance	0.04	0.00	0.00
Step 4	Chi-square	9.01	413.12	413.12
	Degree of Freedom	1	4	4
	statistical significance	0.0028	0.00	0.00

Table 6 presents the accuracy of correct classification of death cases and survival as 93.90%.

Table 6: Classification Table for Observed and Predicted cases

Observed			Predicted		
			Deaths Cases		Percentage Correct
			No	Yes	
	Deaths Cases	No	143	5	96.50
		Yes	23	321	93.30
	Overall (%)				94.30
Step 2	Deaths Cases	No	144	4	97.22
		Yes	23	321	93.30
	Overall (%)				94.50
Step 3	Deaths Cases	No	143	5	96.50
		Yes	25	319	92.70
	Overall (%)				93.90
Step 4	Deaths Cases	No	144	4	97.22
		Yes	26	318	92.40
	Overall (%)				93.90

4. Conclusion

This study is targeted to evaluate the accuracy of the proposed machine learning model and statistical methods based on decision tree and logistic regression. The analysis presented the effects of different factors on the death cases because of the COVID pandemic. The study provided some statistics-based indicators for COVID-19. It is observed that the decision tree model produced better predictions than the logistic regression method. The overall prediction accuracies observed are 94.10% and 93.90%, corresponding to the decision tree and logistic regression methods. Additionally, the observations presented that the regions with dense populations are inclined to have a higher number of cases than those with low densely populated. It is also observed that the death count of female citizens was more than that of male citizens in the UK, which was predicted to be higher for persons70 years and more.

References

Abdel-khalek, S., Alhag, A., Ragab, M., Abo-Dahab, S.M., Algarni, A., and Ahmad, H., 2021. Atomic Fisher information and entanglement forecasting for quantum system based on artificial neural network and time series model, International Journal of Quantum Chemistry, 121.

Abo-Dahab, S.M., Ragab, M., Elhag, A.A., and Abdel-Khalek, S., 2020. Free convection effect on oscillatory flow using artificial neural networks and statistical techniques. Alexandria Engineering Journal, 59 (5), Pp. 3599–3608, 2020.

Abu-Zinadah, H., and Binkhamis, A., 2020. Goodness-of-fit tests for the beta Gompertz distribution. Thermal Science, 24 (1), Pp. 69–81, 2020.

Ahmed, M.Z., and Mahesh, C., 2021. A Comprehensive Review on Machine Learning Models for Medical Data Classification. *2021 2nd*

International Conference on Smart Electronics and Communication (ICOSEC), Pp. 1259-1264.

Almalki, S.J., Abushal, T.A., Alsulami, M.D., and Abd-Elmougod, G.A., 2021. Analysis of type-II censored competing risks' data under reduced new modified Weibull distribution, Complexity, Article ID 9932840, 13 pages.

Alsulami, M.D., 2020. Assorting faces by singular value decomposition. IOSR Journal of Mathematics, 16 (6), Pp. 01–05.

Alsulami, M.D., 2020. Stochastic modelling of infectious disease, International Journal of Innovation in Science and Mathematics, 8 (6), Pp. 262–269.

Alsulami, M.D., 2021. Computational mathematical techniques model for investment strategies. Applied Mathematical Sciences, 15 (1), Pp. 47–55.

Anzolin, E., and Amante, A., 2020. First Italian dies of coronavirus as outbreak flares in north, healthcare & pharma, https://www.reuters.com/article/us-china-health-italy-idUSKBN20F0UI.

Corona-Pandemie 900 Milliardengegen die Angst 2020, https://www.spiegel.de/consent.

Coronavirus: first death confirmed in Europe 2020, https://www.bbc.com/news/world-europe-51514837.

Dharmapuri, S.L., Dandamudi, P.K., Botcha, V.M., and Kolla, P.B., 2020. Detecting central nervous system disorder using machine learning technique (XGB classifier). International Journal of Emerging Trends in Engineering Research, 8 (4), Pp. 1142–1147.

Elhag, A.A., Abu-Zinadah, H., 2020. Forecasting under applying machine learning and statistical models. Thermal Science, 24 (1), Pp. 131–137.

Fischer, A., 2020. Covid-19 in Germany. Asia-Pacific BioTech News, Pp. 72-77.

Hou, L., 2021. Research on Artificial Intelligence Forecasting Method Integrating Data Mining and Statistical Analysis. *2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT)*, Pp. 505-508.

http://www.salute.gov.it/imgs/C_17_notizie_4623_0_file.pdf.

https://www.avvenire.it/attualita/pagine/coronavirus-aggiornamento-5-marzo-2020.

https://www.corriere.it/cronache/20_febbraio_22/coronavirus-italia-nuovi-contagi-lombardia-veneto.

https://www.corriere.it/cronache/20_gennaio_30/coronavirus-italia.

https://www.kreis-heinsberg.de/aktuelles/aktuelles/?pid=5136.

https://www.reuters.com/article/us-china-health-italy/coronavirus-outbreak-grows-in-northern-italy.

Islam, M.Z., Islam, M., and Asraf, A., 2020. A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, Informatics in Medicine Unlocked, vol. 20, article 100412.

Koch, R., 2020. Institute. COVID-19: Fallzahlen in Deutschland und weltweit.

Prakash, K.B., Imambi, S.S., Ismail, M., Kumar, T.P., and Pawan, Y.V.R.N., 2020. Analysis, prediction and evaluation of COVID-19 datasets using machine learning algorithms, International Journal of Emerging Trends in Engineering Research, 8 (5), 2020.

Verghese, A., Sudalaimuthu, T., Visalaxi, S., 2021. Analysis and Forecasting Covid-19 Spread in India Using Logistic Regression and Prophet Time Series. *International Conference on Computational Performance Evaluation (ComPE)*, Pp. 928-932.

Wu, J., Zhang, P., Zhang, L., 2020. Rapid and accurate identifcation of COVID-19 infection through machine learning based on clinical available blood test results, Tech. Rep., medRxiv.

Zayet, S., Klopfenstein, T., Kovàcs, R., Stancescu, S., Hagenkötter, B., 2020. Acute cerebral stroke with multiple infarctions and COVID-19, France, 2020. Emerging infectious diseases, 26 (9), Pp. 2258.