# In The Name of GOD

## به نام خدا

# Presented by:


## Mehdi Jabbari Nooghabi

Associate Professor,
Department of Statistics, Ferdowsi University of Mashhad, Iran
Department of Mathematical Sciences, University of Copenhagen, Copenhagen-Denmark,



**The12th International Scientific Conference, Iraqi Mathematical Society**
Mashhad, Iran, 2-3 July 2023

# Using data science methods to call online data, modeling and forecasting (Application in Covid 19 and Crypto Currency data)

Biomedical Signal Processing and Control 66 (2021) 102494

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

ELSEVIER

Check for updates

## Modeling and forecasting number of confirmed and death caused COVID-19 in IRAN: A comparison of time series forecasting methods

Nasrin Talkhi [a], Narges Akhavan Fatemi [b], Zahra Ataei [b], Mehdi Jabbari Nooghabi [b,*]

[a] Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran
[b] Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran

Cited by 39 Scopus papers till now.

# Abstract:


- **Monitoring of the COVID-19 pandemic is gradually discovering new cases every day.**
- **Forecasting the number of future patients and death cases helps the governments and health-policy makers to make the necessary decisions and impose restrictions to reduce prevalence.**
- **We applied nine models including NNETAR, ARIMA, Hybrid, Holt-Winter, BSTS, TBATS, Prophet, MLP, and ELM network models.**
- **The quality of forecasting models is evaluated by three performance metrics, RMSE, MAE, and MAPE.**
- **Forecasted for the 30 next days.**
- **The used data in this study is the absolute number of confirmed, death cases from February 20 to August 15, 2020.**

# Abstract:


## Results

- The suitable model with the lowest performance metrics for confirmed cases data obtained MLP network and the
- Holt-Winter model is the suitable model for forecasting death cases in the future.

## Conclusion

- We concluded that the MLP and Holt-Winter models had the lowest error in forecasting in comparison to other methods.
- Based on the trend of data and forecast results, the number of confirmed cases and death cases are almost constant and decreasing, respectively.

- There is a possibility of re-emerging this disease more seriously in Iran and this requires more preventive care.

# Introduction

In late December 2019, a novel virus appeared in Wuhan, China [1], which had an acute effect on the respiratory and it was spreading rapidly [1, 2]. The World Health Organization (WHO) introduced this novel virus as SARS-CoV-2 virus, which belongs to the coronavirus family [3].

Some researches and evidence indicate that the main origin of COVID-19 is bats, however, this is not confirmed definitely and needs more investigation and researches [1, 3].

One of the major problems with this virus is that its incubation period can last up to 14 days and during this period, it can transmit the infection without any symptoms [1, 6].

# Introduction

It should be noted that the lack of sufficient information in advance is one of the reasons for the difficulty of forecasting [6], however, it is still an effective policy and guidance for governments to avoid the spread of disease [2, 6, 8].

Therefore, because statistical and mathematical models that are used to forecast can play an effective role in informing the future trend of the disease [1], in this paper, we applied nine models including NNETAR, ARIMA, Hybrid, Holt-Winter, BSTS, TBATS, Prophet, MLP and ELM model to finding the best model for forecasting numbers of confirmed and death cases, separately, for the 30 next days in Iran.

# Material and Methods

## Neural Network Auto Regression Model (NNETAR)

- **A kind of statistical model is a neural network that it uses in machine learning problems.**
- **NNETAR Model is a kind of neural network and a parametric non-linear model which applied for forecasting problems [9].**
- **Forecasting is performed in two phases.**
  1. **For the desired time series, the order of the auto-regressive model is determined in the first phase.**
  2. **The neural network is trained by the training dataset by considering the order of auto-regressive. The number of input nodes or time series lags of the neural network is determined from the order of auto-regressive [9].**
- **The fitted model with a non-seasonal pattern consists of two components p and k, where p indicates the number of input lags and k indicates the number of hidden neurons (NNAR(p, k)).**
- **The fitted model for data with a seasonal pattern is presented as NNAR(p, P, k)[m].**
- **It is similar to ARIMA(p, 0, 0)(P, 0, 0)[m] with nonlinear functions [6].**

# Material and Methods

## Auto-Regressive Integrated Moving Average Model (ARIMA)

**The Box-Jenkins method was proposed by Box, Jenkins [7].**

**This method includes ARIMA models which are non-stationary time series but they are made stationary with differencing [7].**

**The auto-regressive integrated moving average (ARIMA) models are one of the most well-known and widely used models in forecasting time series [8].**

$$\phi_p(B)(1-B)^d y_t = e_0^{\cdot} + e_q^{\cdot}(B)e_t,$$

**where *p* denote orders of auto-regression, q is the order of moving average and d is the number of differencing times.**

# Material and Methods

## Holt-Winter (HW)

- **The Holt-Winter forecasting method is an extension of exponential smoothing and applied for univariate time series [8].**
- **This method doesn’t need a high data storage and is simple [11].**
- **The HW is suitable for short-term forecasting and uses the maximum likelihood function for estimating parameters [8, 11].**
- **There are two Holt-Winter models that use additive or multiplicative models based on the seasonal component [11].**
- **The additive model**

$$\hat{y}_{t+h/t} = a_t + h * b_t + s_{t-p+1+(h-1)mod(p)};$$

$$a_t = \alpha(y_t - s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1}),$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

$$s_t = \gamma(y_t - a_t) + (1-\gamma)s_{t-p}.$$

# Material and Methods

## Holt-Winter (HW)

- **The multiplicative model**

$$\hat{y}_{t+h/t} = (a_t + h*b_t)*s_{t-p+1+(h-1)mod(p)},$$

$$a_t = \alpha(y_t/s_{t-p}) + (1-\alpha)(a_{t-1} + b_{t-1})$$

$$b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1},$$

$$s_t = \gamma(y_t/a_t) + (1-\gamma)s_{t-p}$$

**where $a_t$, $b_t$ and $s_t$, are indicated level, slope, and seasonal of time series at time t, respectively. The p notation indicated the number of seasons in a year.**

- **Also, coefficients $\alpha$, $\beta$, and $\gamma$ are constant and smoothing parameters between zero and one interval. The end h is the forecast horizon [11].**

# Material and Methods

## Hybrid model

- There are appropriate functions for ensemble forecasts in R software.
- In the ‘forecastHybrid’ package, by default, Forecasts generated from auto.arima(), ets(), thetaf(), nnetar(), stlm(), tbats(), and snaive() can be combined with equal weights.
- The other weights are based on in-sample errors that introduced by Bates & Granger (1969), or cross-validated weights. Cross-validation is used to evaluate the accuracy of the model and is supported by user-defined models and forecasting functions.
- Two of the models used in the combination namely, NNETAR and auto.arima.

# Material and Methods

## Bayesian structural time-series (BSTS)

- The Bayesian approach based on prior experience and given data builds analytical models [12].
- Make the posterior distribution and this leads to the final Bayesian model [12].
- BSTS belong to the family of state-space models that are applied for time series data.

$$y_t = Z_t^T \alpha_t + \varepsilon_t$$

$$\alpha_{t+1} = T_t + \alpha_t + R_t \eta_t.$$

# Material and Methods

## TBATS model

- **The phrase BATS is abbreviated based on five features including Box-Cox transform, ARMA errors, Trend, and Seasonal components.**

- **It is supplemented by ($\omega$, $\emptyset$, $p$, $q$, $m1$,..., $mT$) to presenting the Box-Cox, damping, ARMA(p, q), and Seasonal periods ($m1$ ,..., $mT$) [8, 14].**
- **This model is a generalization of the traditional seasonal models with multiple seasonal periods [14].**

- **This class of model is called TBATS which the first T notation referred to "trigonometric". Considers any autocorrelation in the residuals and handles nonlinear attributes in real-time series [14].**

- **A large parameter space with the possibility of better forecasts and it is an efficient estimation procedure totally [8].**

# Material and Methods

## Prophet: Automatic Forecasting Procedure

- There is an available forecasting tool called Prophet in R and Python.
- The prophet is an additive regression that has a linear trend in piecewise or logistic growth curve trend.
- A yearly seasonal component modeled using the Fourier series and a weekly seasonal component modeled using dummy variables.
- 
- It is used for business tasks that we deal with on Facebook and has been optimized for this purpose [8].
- Decomposable time-series model consisting of trend, seasonality, and holiday components. The Prophet depends on the Fourier series to consider seasonality.
- Creates a more flexible model for periodic effects.

# Material and Methods

## Multilayer Perceptron (MLP)

- **MLP network is a kind of the main perceptron model [15].**
- **The network architecture is displayed in Fig. 1. MLPs include at least three layers.**
- **This model consists of inputs, weights, biases, and an activation function that yields the output [16].**
- **Each input *x_i* to a neuron, *j* is multiplied by an adaptive coefficient *w_ij*, called weight.**
- **Then with a nonlinear activation function ($\varphi$) such as sigmoid, hyperbolic tangent, etc.**

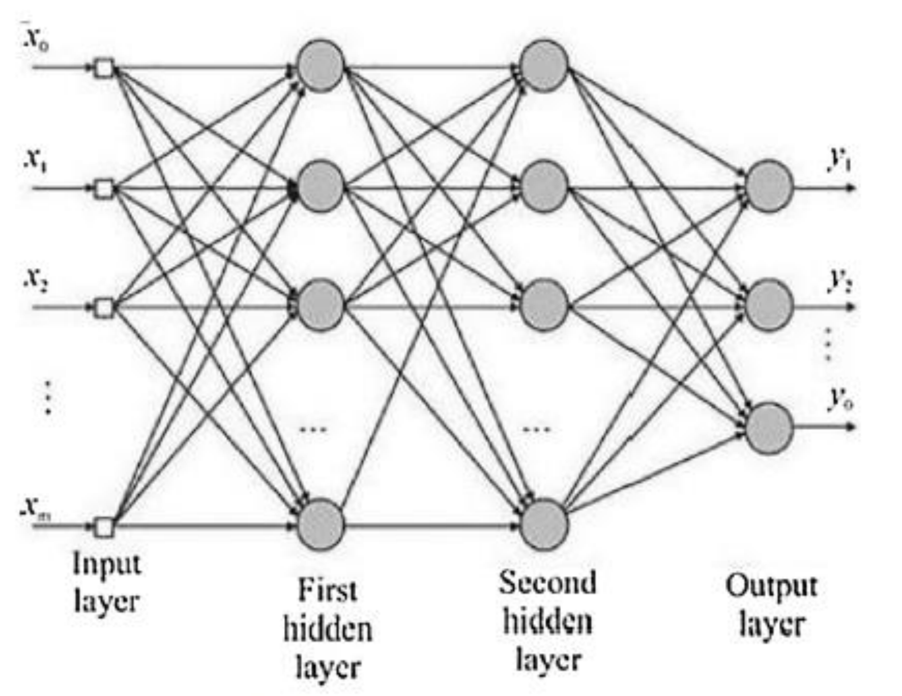


Fig. 1. The MLP network structure.

$$o_i = \varphi\left(\sum_{j=1}^{d}\left(x_j w_{ij} + b_j\right)\right)$$

# Material and Methods

## Extreme Learning Machines (ELM)

- **The ELM is a learning algorithm with high speed for the single hidden layer feed-forward neural networks (SLFN) [17] (Fig. 2).**
- **This method overcomes the debility of the traditional learning algorithms in the process of learning speed because ELM could be improving the generalization performance and reducing the training time [6].**
- **ELMs in comparison with traditional learning algorithms tend to reach the smallest training error [6].**

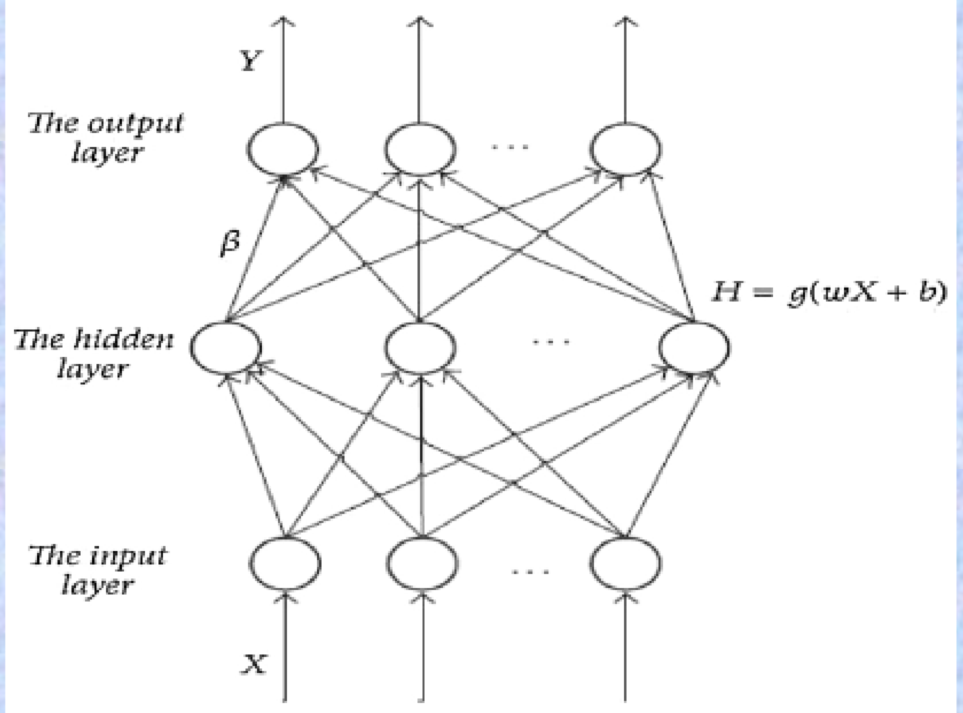


Fig. 2. The ELM network structure.

# Model Evaluation

- **To evaluate the quality or goodness of fit of the used methods three performance metrics, Root Mean Square Error (RMSE),**
- **Mean Absolute Error (MAE),**
- **Mean Absolute Percentage Error (MAPE)**
- **in the training and testing phases were applied.**

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2},$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|,$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} * 100\%$$

**where *y_i* is the actual value of time series at time *i*, and \hat{*y}*_*i* is the forecast value of the time series at time *i* [1].**

# Data Collection and Results

- **To forecast future behavior of COVID-19, dataset included**
  - **the absolute number of confirmed, death, and recovered cases caused by the new coronavirus in Iran.**
- **The dataset was available on the**
  - **https://www.worldometers.info/coronavirus/**
  - **Reported daily from February 20, 2020, on this site.**
  - **All data analysis was performed using R software version 4.0.2.**
  - **The trend of daily confirmed, death, and recovered cases in Iran from February 20 to August 15, 2020, is shown in Fig. 3.**
- **Nine different methods were fitted to the data of COVID-19 (confirmed and death cases).**
  - **We evaluated the performance of methods by training and testing dataset.**
  - **The first 70% of data are used as training and the next 30% data for testing the models.**
  - **Then, the forecasting quality of the models is evaluated by three metrics**
  - **RMSE, MAE, and MAPE.**

# Data Collection and Results

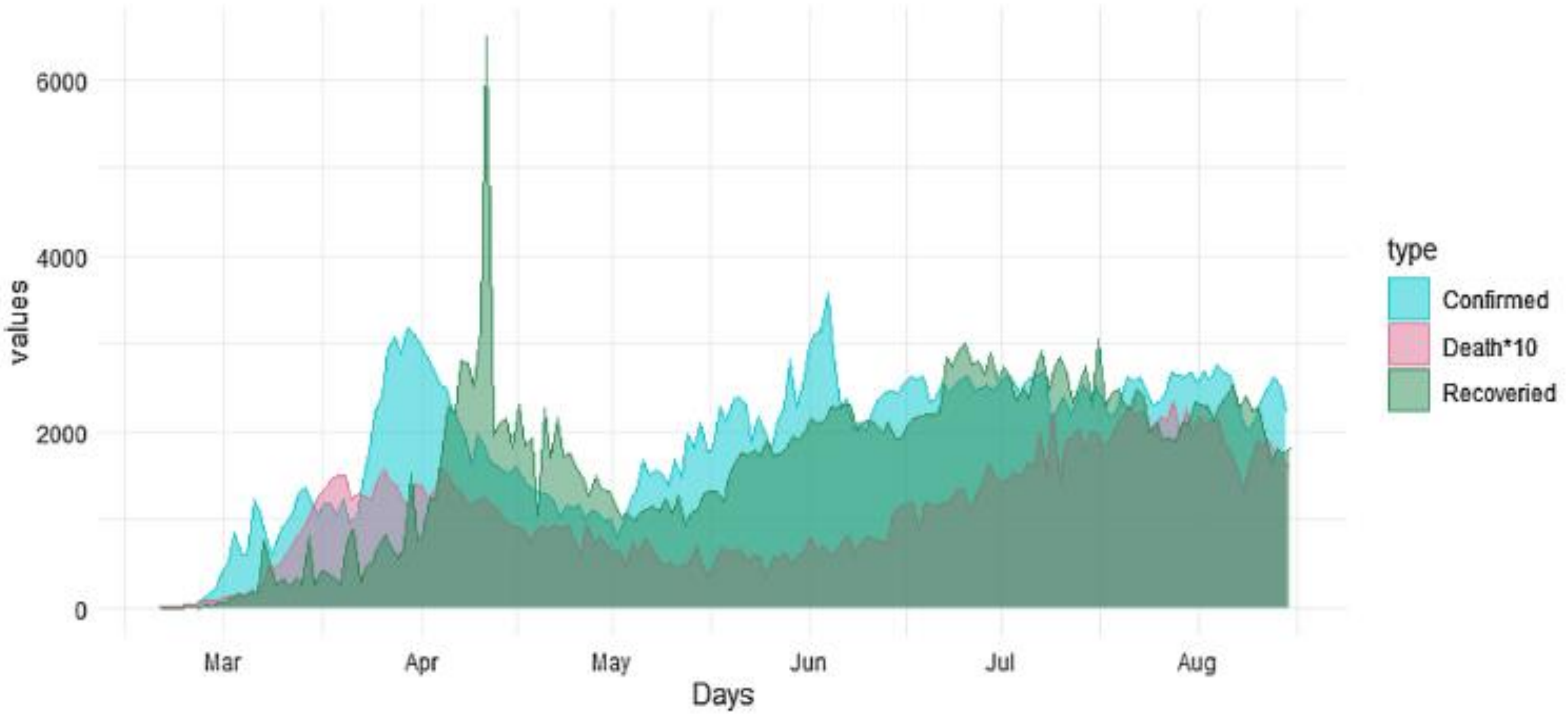



Fig. 3. The trend of Daily of Confirmed, death, and Recovered cases.

# Data Collection and Results

- **The performance metrics RMSE, MAE, and MAPE calculated for all of the models in the training and testing phases.**
- **These results are reported in Table 1 and Table 2.**
- **These results are shown in Fig. 4.**

**Table 1. The results of the models for confirmed cases.**

| Confirmed Cases | | | | | | |
|---|---|---|---|---|---|---|
| Models | Training Data | | | Testing Data | | |
| | RMSE | MAE | MAPE | RMSE | MAE | MAPE |
| NNETAR(1,1) | 255.7547 | 204.3763 | 39.566 | 291.4161 | 260.1861 | **10.22983** |
| ARIMA(1,0,0) | 231.6003 | 177.2125 | 82.10807 | 561.9214 | 501.4737 | **26.62457** |
| Hybrid-e | 227.5012 | 175.0365 | 21.23171 | 180.8860 | 151.9495 | **6.268913** |
| Hybrid-c | 227.4615 | 175.0335 | 21.34771 | 180.8883 | 151.9539 | **6.269047** |
| Holt-Winter | 233.5451 | 177.73 | 13.07673 | 299.6471 | 226.3595 | **9.735324** |
| BSTS | 254.8199 | 195.7948 | 16.58057 | 550.1058 | 455.7354 | **19.13969** |
| TBATS | 225.6698 | 170.7427 | 15.62544 | 217.2329 | 185.6827 | **7.394939** |
| Prophet | 608.2165 | 441.5421 | 311.6574 | 612.9864 | 537.7585 | **22.4437** |
| **MLP** | **224.4852** | **177.5885** | **24.95336** | **180.2759** | **142.8951** | **5.725628** |
| ELM | 237.8037 | 190.5021 | 39.43857 | 443.9748 | 405.2195 | **19.68961** |

# Data Collection and Results

**Table 2. The results of the models for death cases.**

| Death Cases | | | | | | |
|---|---|---|---|---|---|---|
| | Training Data | | | **Testing Data** | | |
| Models | RMSE | MAE | MAPE | RMSE | MAE | **MAPE** |
| NNETAR(1,1) | 14.14151 | 10.79158 | 24.94921 | 81.83506 | 75.38808 | **39.47772** |
| ARIMA(1,0,1) | 12.34115 | 9.318635 | 23.15612 | 89.47732 | 81.7967 | **84.53056** |
| Hybrid-e | 11.85159 | 8.795046 | 13.7387 | 65.13031 | 58.00313 | **29.9145** |
| Hybrid-c | 11.85194 | 8.795424 | 13.73874 | 65.13291 | 58.00584 | **29.91598** |
| **Holt-Winter** | **12.38061** | **9.435316** | **14.21699** | **35.4963** | **26.75278** | **15.10667** |
| BSTS | 12.86378 | 9.834921 | 15.14902 | 48.90122 | 41.58697 | **21.41159** |
| TBATS | 12.30943 | 9.057055 | 14.30562 | 42.37191 | 35.50072 | **18.09161** |
| Prophet | 37.13429 | 31.7645 | 175.111 | 101.7453 | 97.02142 | **51.92662** |
| MLP | 11.6038 | 8.513807 | 14.5441 | 60.86964 | 53.39749 | **27.38357** |
| ELM | 12.79517 | 10.33391 | 27.59607 | 87.46979 | 80.55371 | **42.1807** |

# Data Collection and Results


**Fig. 4.** The comparison of the performance metrics models for the confirmed and death in the test phase.

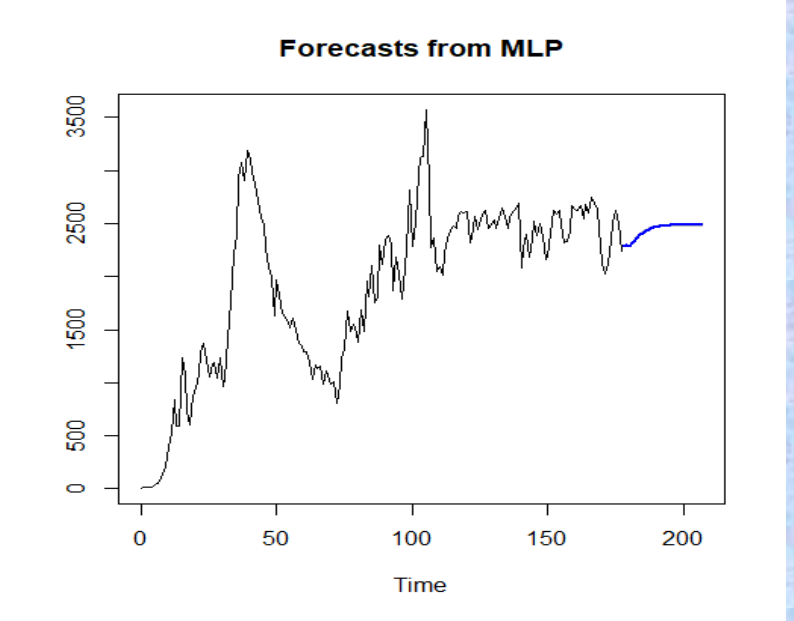
# Data Collection and Results

**By comparing performance metrics,**

- **We concluded that for confirmed cases, except for the Hybrid-e model, other models did not perform well in the test phase.**
- **The Holt-Winter model was the best model with the lowest performance metrics for death cases time series data.**
- **The Hybrid-e model is the best models with the lowest performance metrics to forecasting confirmed cases.**
- **The Holt-Winter model is the best models with the lowest performance metrics to forecasting death cases.**

# Forecasting

- The 30-days COVID-19 forecasting graphs of confirmed and death cases (Fig. 5) were plotted.
- The results of the forecast showed which on September 14, 2020, we will have 2,484 new confirmed and 114 new death cases of COVID-19.



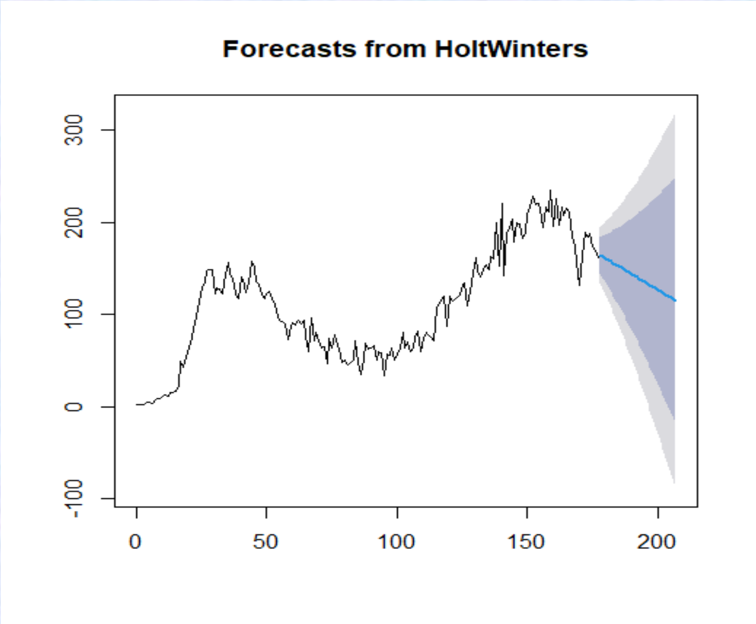




Fig. 5. Forecasting future of the time series for confirmed cases by MLP model (Left) and death cases by Holt-Winter model (Right).

# References:


1. Al-Qaness MAA, Ewees AA, Fan H, Abd El Aziz M. Optimization Method for Forecasting Confirmed Cases of COVID-19 in China. Journal of clinical medicine. 2020;9(3).
2. Ceylan Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. The Science of the total environment. 2020;729:138817.
3. Moftakhar L, Seif M, Safe MS. Exponentially Increasing Trend of Infected Patients with COVID-19 in Iran: A Comparison of Neural Network and ARIMA Forecasting Models. Iranian Journal of Public Health. 2020;49(Supple 1).
4. Yang Q, Wang J, Ma H, Wang X. Research on COVID-19 based on ARIMA modelΔ—Taking Hubei, China as an example to see the epidemic in Italy. Journal of Infection and Public Health. 2020.
5. Sahu KK, Mishra AK, Lal A. COVID-2019: update on epidemiology, disease spread and management. Monaldi Arch Chest Dis [Internet]. 2020 2020/04//; 90(1). Available from: http://europepmc.org/abstract/MED/32297723, https://doi.org/10.4081/monaldi.2020.1292.
6. Pontoh RS, Z S, Hidayat Y, Aldella R, Jiwani NM, Sukono. Covid-19 Modelling in South Korea using A Time Series Approach. International Journal of Advanced Science and Technology. 2020;29(7):1620 - 32.
7. Yonar H, Yonar A, Agah Tekindal M, Tekindal M. Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods. EJMO. 2020;4(2):160-5.
8. Papastefanopoulos V, Linardatos P, Kotsiantis S. COVID-19: A Comparison of Time Series Methods to Forecast Percentage of Active Cases per Population. Applied Sciences-Basel. 2020;10(11):3880.
9. Sena D, Nagwani NK. A neural network autoregression model to forecast per capita disposable income. ARPN Journal of Engineering and Applied Sciences. 2016;11:13123-8.
10. Almasarweh M, Alwadi S. ARIMA model in predicting banking stock market data. Modern Applied Science. 2018;12(11):4.

# References:


11. Awajan AM, Ismail MT, Al Wadi S. Improving forecasting accuracy for stock market data using EMD-HW bagging. PloS one. 2018;13(7):e0199582.
12. Jun S. Bayesian Structural Time Series and Regression Modeling for Sustainable Technology Management. Sustainability. 2019;11(18):4945.
13. Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL. Inferring causal impact using Bayesian structural time-series models. The Annals of Applied Statistics. 2015;9(1):247-74.
14. De Livera AM, Hyndman RJ, Snyder RD. Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. Journal of the American Statistical Association. 2011;106(496):1513-27.
15. Kaushik S, Choudhury A, Sheron PK, Dasgupta N, Natarajan S, Pickett LA, et al. AI in Healthcare: Time-Series Forecasting Using Statistical, Neural, and Ensemble Architectures. 2020;3(4).
16. Parhizkari L, Najafi A, Golshan M. Medium term electricity price forecasting using extreme learning machine. Journal of Energy Management and Technology. 2020;4(2):20-7.
17. Lai J, Wang X, Li R, Song Y, Lei L. BD-ELM: A Regularized Extreme Learning Machine Using Biased DropConnect and Biased Dropout. Mathematical Problems in Engineering. 2020:1-7.

# References:

18. Mounesan L, Eybpoosh S, Haghdoost A, Moradi G, Mostafavi E. Is reporting many cases of COVID-19 in Iran due to strength or weakness of Iran's health system? Iran J Microbiol. 2020;12(2):73-6.
19. Roosa K, Lee Y, Luo R, Kirpich A, Rothenberg R, Hyman JM, et al. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. Infect Dis Model. 2020;5:256-63.
20. Eubank S, Eckstrand I, Lewis B, Venkatramanan S, Marathe M, Barrett CL. Commentary on Ferguson, et al., "Impact of Non-pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand". Bulletin of mathematical biology. 2020;82(4):52.
21. https://cran.r-project.org/web/packages/forecastHybrid/index.html website.
22. Wang W, Tang J, Wei F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. Journal of medical virology. 2020;92(4):441-7.
23. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet (London, England). 2020;395(10223):497-506.
24. Tavakoli A, Vahdat K, Keshavarz M. Novel Coronavirus Disease 2019 (COVID-19): An Emerging Infectious Disease in the 21st Century. BPUMS. 2020;22(6):432-50.

# The other related researches:

A research plan awarded by FUM:

1. Estimate Daily Case Fatality Rate and Cure Rate of COVID-19 in IRAN and Compare to Other Countries: A Cluster Analysis

http://mathstat.um.ac.ir/index.php?option=com_content&view=article&id=1021:yhjk&catid=90:2015-10-17-05-28-08&Itemid=775&lang=fa

## انتشار مقاله طرح پژوهشی مرتبط با کووید 19 در دانشگاه فردوسی مشهد و انعقاد اولین قرارداد پژوهشی مرتبط با این بیماری تحت حمایت مرکز نوآوری دانشگاه توسط عضو هیات علمی دانشکده علوم ریاضی

مقاله مستخرج از طرح پژوهشی مرتبط با کووید 19 با عنوان: مدل سازی و پیش بینی تعداد مبتلابان و مرگ و میر ناشی از کووید19 در ایران-مقایسه مدل های مختلف پیش بینی سری زمانی، توسط دکتر مهدی جباری، دانشیار گروه آمار دانشکده علوم ریاضی در مجله معتبر Journal of Biomedical Signal Processing and Control دارای اعتبار JCR-Q2 به چاپ رسیده است که در لینک https://www.sciencedirect.com/science/article/pii/S1746809421000914 قابل دسترسی است.

هم چنین اولین قرارداد طرح پژوهشی مرتبط با کووید 19 تحت حمایت مرکز نوآوری دانشگاه فردوسی مشهد با عنوان: برآورد نرخ مرگ و میر و میزان بهبودی کووید19 در ایران و مقایسه آن با سایر کشورها-تحلیل خوشه ای، با ایشان منعقد شده است.

خروجی طرح های مذکور به صورت نرم افزار به آدرس http://shiny.um.ac.ir/jabbarinm-shiny/ طراحي شده است و مي تواند مورد استفاده عموم قرار بگیرد.

در زمینه این نرم افزار لازم به توضیح است که بایستي يك فايل حاوي داده هاي مرگ و مير يا تعداد بيماران مبتلا به كوويد19 (يا هر مشخصه ديگر) به صورت روزانه و به فرمت سي اس وي (csv) تهيه شده و در مسير تعبيه شده در تب (Data Import) دراين اپليكيشن بارگذاري گردد. سپس در تب كناري (Model_dev) مي توان مدل مناسب را انتخاب كرد و روي داده هاي بارگزاري شده و متغير منتخب برازش خواهد يافت. در ادامه نيز مي توان از تب Final Model (MLP) for confirmed cases or HoltWinters for death cases مدل نهايي براي مرگ و مير و يا تعداد مبتلايان را برازش و براساس آن پيش بيني براي 30 روز آينده انجام داد.

دانشكده علوم رياضی ضمن تبريك به اين همكار آمار كاربردي براي ايشان آرزوي سلامتي و موفقيت روز افزون دارد.

2. Talkhi N, Akhavan Fatemi N, Jabbari Nooghabi M. Revealing Behavior Patterns of SARS-CoV-2 using Clustering Analysis and XGBoost Error Forecasting Models. Iran J Med Microbiol. 2022; 16(3) :221-232.

Original Article | Iran J Med Microbiol. 2022; 16(3): 221-232

10.30699/ijmm.16.3.221

Iranian Journal of Medical Microbiology | ISSN:2345-4342

# Revealing Behavior Patterns of SARS-CoV-2 using Clustering Analysis and XGBoost Error Forecasting Models

Nasrin Talkhi[1], Narges Akhavan Fatemi[2], Mehdi Jabbari Nooghabi[2*]

1. D Department of Biostatistics, School of Health, Mashhad University of Medical Sciences, Mashhad, Iran
2. Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran

ABSTRACT

**Background and Aim:** COVID-19 is a highly contagious infectious disease, and it has affected people's daily life and has raised great concern for governments and public health officials. Forecasting its future behavior may be useful for allocating medical resources and defining effective strategies for disease control, etc.

**Materials and Methods:** The collected data was the cumulative and the absolute number of confirmed, death, and recovered cases of COVID-19 from February 20 to July 03, 2021. We used hierarchical cluster analysis. To forecast the future behavior of COVID-19, the Auto-Regressive Integrated Moving Average (ARIMA), Exponential Smoothing (ETS), Automatic Forecasting Procedure (Prophet), Naive, Seasonal Naive (s-Naive), boosted ARIMA, and boosted Prophet models were used.

**Results:** The results of clustering showed a similar behavior of coronavirus in Iran and other countries such as France, Russia, Turkey, United Kingdom (UK), Argentina, Colombia, Italy, Spain, Germany, Poland, Mexico, and Indonesia. It also revealed similar patterns of SARS-CoV-2 for the same countries in six groups. Results showed that XGBoost models' family had higher accuracy than other models.

**Conclusion:** In Iran, COVID-19 showed similar behavior patterns compared to the studied developed countries. The family of XGBoost models showed practical results and high precision in forecasting behavior patterns of the virus. Concerning the rapid spread of the virus worldwide, these models can be used to forecast the behavior patterns of SARS-CoV-2. Preventing the spread of the coronavirus, controlling the disease, and breaking down its chain necessitates community assistance, and in this mission, the role of statisticians cannot be neglected.

**Keywords:** ARIMA, Cluster Analysis, COVID-19, Forecasting, Naïve, XGBoost Error

Received: 2021/07/18; Accepted: 2022/01/30; Published Online: 2022/02/20

Corresponding Information: Mehdi Jabbari Nooghabi, *Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran*
Email: jabbarinm@um.ac.ir

Copyright © 2021, This is an original open-access article distributed under the terms of the Creative Commons Attribution-noncommercial 4.0 International License which permits copy and redistribution of the material just in noncommercial usages with proper citation.

Use your device to scan and read the article online

Talkhi N, Akhavan Fatemi N, Jabbari Nooghabi M. Revealing Behavior Patterns of SARS-CoV-2 using Clustering Analysis and XGBoost Error Forecasting Models. Iran J Med Microbiol. 2022; 16 (3) :221-232

Download citation: BibTeX | RIS | EndNote | Medlars | ProCite | Reference Manager | RefWorks
Send citation to: Mendeley Zotero RefWorks

## 1 Introduction

The novel coronavirus called SARS-CoV-2 was initially observed in Wuhan, China, in December 2019, and then it spread rapidly throughout the world and affected an enormous number of people (1-5). Because of its high contagiousness, particularly dynamic structure, unknown etiology, and hazardousness, the virus has raised great concern for the governments and public health officials (2, 6-8).

Some transmission routes are close personal contact (6) and respiratory droplets, when the infected person sneezes, coughs or even speaks (9, 10). The mortality rate in the elderly and high-risk groups such as those with cardiovascular diseases, diabetes, chronic respiratory disease, and hypertension is significantly higher than in the healthy ones (11). Being aware of the spread of COVID-19 can help governments and decision-makers make the right decisions and perform effective plans to prevent its spread (6, 7, 12). It is worth noting that the most critical factor in controlling and reducing the spread of the virus is its recognition, health care, and control measures (13). Also, to prevent the spread of disease

# Submitted Related Papers:

**3. Support vector machine as an appropriate model to forecast behavior of coronavirus disease 2019: A machine learning time series techniques.**

**4. Using Meta-Learning to Forecasting coronavirus disease 2019.**

**5. Investigation of the Status Active Case in Covid-19 Six Waves (days between 27th February 2020 and 19th March 2022) Using Statistical and Mathematical Models in Iran**

## Data Collecting by:

## Application Programming Interface (API)

Example:

https://pomber.github.io/covid19/timeseries.json
https://mahdisalehi.shinyapps.io/Covid19Dashboard/
https://coinmarketcap.com/
https://coin360.com/
….

# Example for Covid data:

```
library(RJSONIO)
url='https://pomber.github.io/covid19/timeseries.json'
x=fromJSON(url)
data=matrix(unlist(x$Iran),nc=4,byrow=T)
colnames(data)=c("date","confirmed","deaths","recover
ed")
head(data,30)
```

# Cumulative Data:

| | date | confirmed | deaths | recovered |
|---|---|---|---|---|
| [1,] | "2020-1-22" | "0" | "0" | "0" |
| [2,] | "2020-1-23" | "0" | "0" | "0" |
| [3,] | "2020-1-24" | "0" | "0" | "0" |
| [4,] | "2020-1-25" | "0" | "0" | "0" |
| [5,] | "2020-1-26" | "0" | "0" | "0" |
| [6,] | "2020-1-27" | "0" | "0" | "0" |
| [7,] | "2020-1-28" | "0" | "0" | "0" |
| [8,] | "2020-1-29" | "0" | "0" | "0" |
| [9,] | "2020-1-30" | "0" | "0" | "0" |
| [10,] | "2020-1-31" | "0" | "0" | "0" |
| [11,] | "2020-2-1" | "0" | "0" | "0" |
| [12,] | "2020-2-2" | "0" | "0" | "0" |
| [13,] | "2020-2-3" | "0" | "0" | "0" |
| [14,] | "2020-2-4" | "0" | "0" | "0" |
| [15,] | "2020-2-5" | "0" | "0" | "0" |
| [16,] | "2020-2-6" | "0" | "0" | "0" |
| [17,] | "2020-2-7" | "0" | "0" | "0" |
| [18,] | "2020-2-8" | "0" | "0" | "0" |
| [19,] | "2020-2-9" | "0" | "0" | "0" |
| [20,] | "2020-2-10" | "0" | "0" | "0" |
| [21,] | "2020-2-11" | "0" | "0" | "0" |
| [22,] | "2020-2-12" | "0" | "0" | "0" |
| [23,] | "2020-2-13" | "0" | "0" | "0" |
| [24,] | "2020-2-14" | "0" | "0" | "0" |
| [25,] | "2020-2-15" | "0" | "0" | "0" |
| [26,] | "2020-2-16" | "0" | "0" | "0" |
| [27,] | "2020-2-17" | "0" | "0" | "0" |
| [28,] | "2020-2-18" | "0" | "0" | "0" |
| [29,] | "2020-2-19" | "2" | "2" | "0" |
| [30,] | "2020-2-20" | "5" | "2" | "0" |

# Example to model by R:

```
# type : confirmed or death
type=data$Abs.cases
type=data$Abs.deaths
#### Normalization data
#data$type <- (type-min(type))/(max(type)-min(type))
head(data)
#### Creat Train and Test data
train_type1 <- head(type, round(length(type) * 0.70))
test_type1 <- tail(type, length(data$type) - length(train_type1))
#### Creat ts data
train_type <- ts(train_type1, frequency=1, start=c(20/02/2020,1))
test_type <- ts(test_type1, frequency=1, start=c(23/06/2020,1))
###### Forecasting data
ts_type <- ts(type, frequency=1, start=c(20/02/2020,1))
```

```
#++++++++++++++++++++++++++++++ NNETAR+++++++++++++++++++++++
set.seed(1234)
fitnnetar.train <- nnetar(train_type, decay=0.5,
maxit=150,lambda="auto",scale.inputs=T)
accuracy(fitnnetar.train)
forecast.nnetar.test=data.frame(forecast(fitnnetar.train, h=length(test_type)))
accuracy(forecast.nnetar.test[,1], test_type)
#++++++++++++++++++++++++++++++++ auto.arima
+++++++++++++++++++++++
fitarima.train <- auto.arima(train_type, trace=T, max.d=5, stationary = T,
seasonal = FALSE)
fitarima.train
summary(fitarima.train)
forecast.arima.test <- data.frame(forecast(fitarima.train, h=length(test_type)))
accuracy(forecast.arima.test[,1], test_type)

.....
```

# Forecasting:

```
#+++++++++++++++++++++++++++++ final stage +++++++++++++++++++++++++
#++++++++++++++++++++++++++++ forecasting ++++++++++++++++++++++++++

## confirmed cases
set.seed(123)
fit.rm.for <- mlp(ts_type,hd.auto.type="cv")
forecast.mlp.final <- data.frame(forecast(fit.rm.for, h=30))
plot(forecast(fit.rm.for,h=30))
write.xlsx(forecast.mlp.final, "Path to save\\mlpforecast.xlsx")



## death cases
set.seed(123)
fitHolt.train.for <- HoltWinters(ts_type, gamma = FALSE)
forecast.Holt.final <- data.frame(forecast(fitHolt.train.for,h=30))
plot(forecast(fitHolt.train.for, h=30))
```

# Modeling and Forecasting by using Shiny Apps:

http://shiny.um.ac.ir/jabbarinm/Covid19/

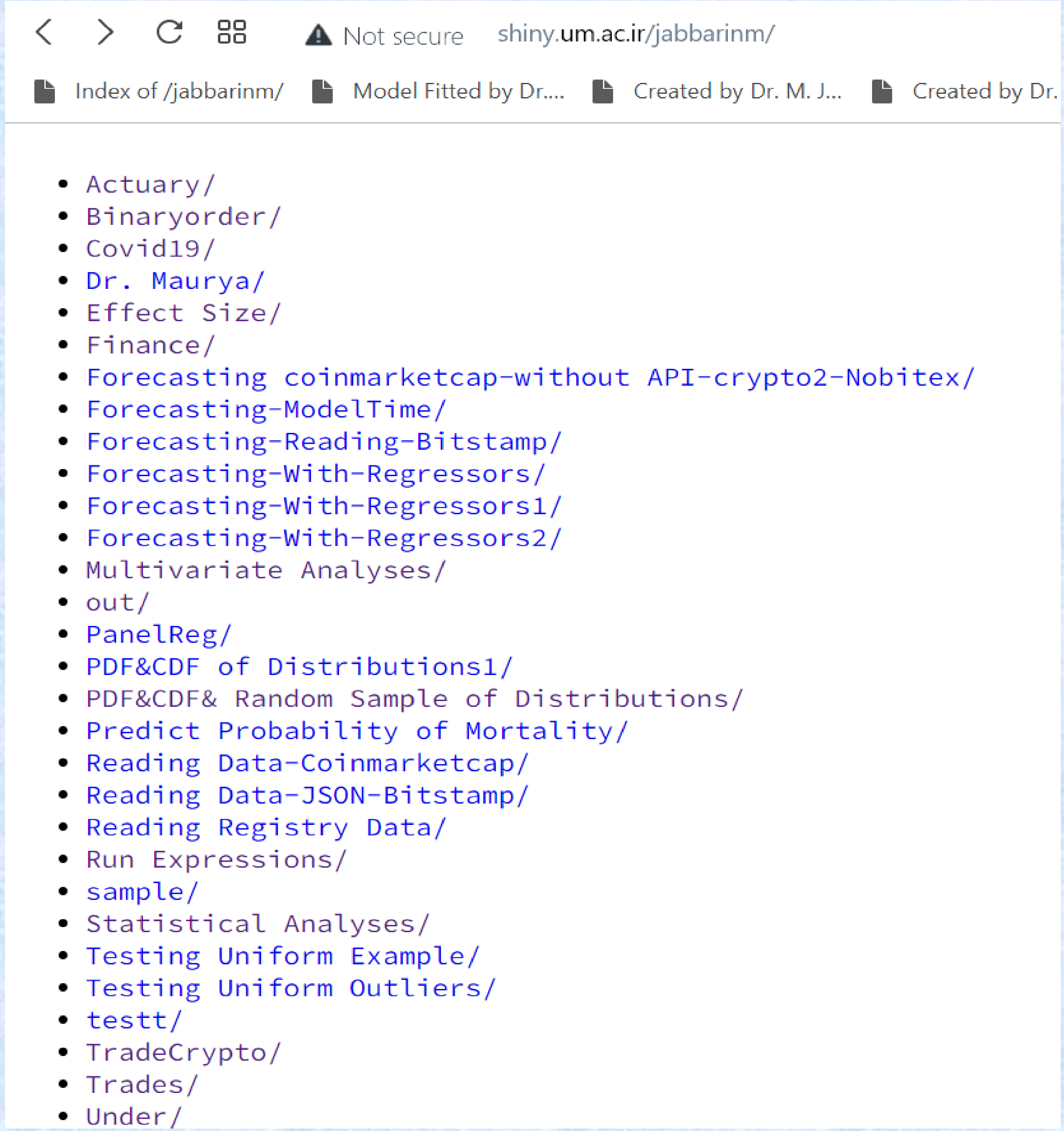
# The other Apps to Analyze Data by using Statistical Methods as well as Machine Learning and Trade CryptoCurrency:

http://shiny.um.ac.ir/jabbarinm/

http://shiny.um.ac.ir/jabbarinm/Statistical%20Analyses/

http://shiny.um.ac.ir/jabbarinm/Multivariate%20Analyses/

http://shiny.um.ac.ir/jabbarinm/TradeCrypto/



Not secure shiny.um.ac.ir/jabbarinm/

Index of /jabbarinm/ | Model Fitted by Dr.... | Created by Dr. M. J... | Created by Dr.

- Actuary/
- Binaryorder/
- Covid19/
- Dr. Maurya/
- Effect Size/
- Finance/
- Forecasting coinmarketcap-without API-crypto2-Nobitex/
- Forecasting-ModelTime/
- Forecasting-Reading-Bitstamp/
- Forecasting-With-Regressors/
- Forecasting-With-Regressors1/
- Forecasting-With-Regressors2/
- Multivariate Analyses/
- out/
- PanelReg/
- PDF&CDF of Distributions1/
- PDF&CDF& Random Sample of Distributions/
- Predict Probability of Mortality/
- Reading Data-Coinmarketcap/
- Reading Data-JSON-Bitstamp/
- Reading Registry Data/
- Run Expressions/
- sample/
- Statistical Analyses/
- Testing Uniform Example/
- Testing Uniform Outliers/
- testt/
- TradeCrypto/
- Trades/
- Under/

http://shiny.um.ac.ir/jabbarinm/TradeCrypto/

Not secure shiny.um.ac.ir/jabbarinm/TradeCrypto/

Index of /jabbarinm/ | Index of /jabbarinm/ | Model Fitted by Dr.... | Created by Dr. M. Ja...

# Index of /jabbarinm/TradeCrypto/

- Nobitex Robat-Navasan-Buy-V1/
- Nobitex Robat-Navasan-Buy-VN/
- Nobitex Robat-Navasan-Buy-VN7/
- Nobitex Robat-Navasan-Sell-VN7/
- Nobitex-Robat-Navasan-Buy-shiny-New-R6/
- Nobitex-Robat-Navasan-Buy-shiny-NewNew-R7/
- Nobitex-Robat-Navasan-Sell-shiny-New-R6/
- Nobitex-Robat-Navasan-Sell-shiny-NewNew-R7/

Not secure shiny.um.ac.ir/jabbarinm/TradeCrypto/Nobitex-Robat-Navasan-Buy-shiny-NewNew-R7/

Index of /jabbarinm/ | Index of /jabbarinm/ | Model Fitted by Dr.... | Created by Dr. M. Ja... | Created by Dr. M. Ja... | Moc

Robot-Nobitex NewNew by Dr. M. Jabbari Nooghabi (BUY)-R7 | Reading Priminarly Values | Reading Secondary Values

**Token:**

**curr1:**
dot

**curr2:**
usdt

**Order of Outliers:**
3

**Range of Boxplot:**
1

**Minumum wait time (Second):**
120

**Maximum wait time (Second):**
125

THANK YOU

باتشکر از شما