Callan Noak
Major in Computer Science
Mentor: Dr. Sujing Wang
Research in Data Science and Analysis
Department of Computer Science

Analysis of COVID-19 Data
Using Advanced Machine Learning Models

Abstract
The outbreak of COVID-19 has dramatically changed peoples' lives over the past two years. The goal of this project is to analyze COVID-19 data from John Hopkins University for the USA, Brazil, India, and Iran between January 23, 2020 and January 21, 2022, and to develop prediction models for the new daily cases and new daily deaths related to COVID-19 for these four countries. Two supervised learning techniques, i.e., Holt’s double exponential smoothing and Autoregressive Integrated Moving Averages (ARIMA), were developed to generate the prediction models for each country. Based on the results of this project, both Holt’s double exponential smoothing and ARIMA techniques can produce good models that fit the data well for all four countries and help to predict the patterns of the spread of COVID-19.

Introduction
In January 2020, the World Health Organization (WHO) declared COVID-19 as a Public Health Emergency of International Concern (PHEIC) [1]. This virus is infectious and spreads quickly across countries and continents through people, animals, and goods. People are still trying to combat the virus by wearing masks and social distancing as much as possible. The outbreak of COVID-19 has dramatically changed peoples' lives over the past two years.
Supervised machine learning is a subcategory of machine learning and artificial intelligence. Supervised machine learning techniques can build highly accurate models to solve a variety of real-world problems. In this project, we designed prediction models based on supervised machine learning techniques to predict and forecast the new daily cases and new daily deaths related to COVID-19 for the USA, Brazil, India, and Iran. These models can help to interpret patterns of the spread of COVID-19 and to assess political and economic influence of the spread of COVID-19. Two supervised machine learning techniques, i.e., Holt’s double exponential smoothing [2] and Autoregressive Integrated Moving Averages (ARIMA) [3], were utilized for this project. The data used for this project was downloaded from Kaggle [4]. The data was from John Hopkins University between January 23, 2020 and January 21, 2022.

**Literature Review**

There are several works in the literature focusing on analyzing and predicting COVID-19 data. In [5], four supervised machine learning models were used to perform 10-day predictions of COVID-19 data. The models used in this project were linear regression, least absolute shrinkage and selection operator (LASSO), support vector machine, and exponential smoothing. The project predicted data collected from John Hopkins University that was related to the number of cases, deaths, and recoveries for Australia, Canada, Algeria, and Afghan between January 22, 2020 to March 27, 2020. Doroshenko [6] used two clustering techniques, i.e., k-means and hierarchical clustering, to analyze COVID-19 data in Italy and provide more information about the impact of COVID-19. It was found that each cluster contained similar regions with similar levels of industrial development. Moreover, those regions with heavy industry were more affected by COVID-19. In [7], ARIMA models were developed to predict the number of COVID-19 total confirmed cases for 145 countries. The experimental results showed that those ARIMA models were able to accurately predict the data with a low RMSE average of 144.81. In [8], time series models such as Holt’s exponential smoothing and ARIMA were used to forecast the number of COVID-19 cases in Jakarta between March 1, 2020 and July 6, 2020. The results of this project showed that the ARIMA model provided the best prediction. Therefore, ARIMA was determined to be the optimal forecasting model.

Chakraborty et al. [9] proposed to use Granular box regression combined with linear regression and polynomial regression to predict the number of confirmed COVID-19 cases and
deaths in India between January 30, 2020 and May 15, 2020. The results showed that polynomial regression performed better than granular box regression when limiting the highest polynomial degree to two and the highest number of boxes to four. In [10], time series models such as exponential smoothing, ARIMA, Seasonal Autoregressive Integrated Moving Average (SARIMA), and Seasonal Autoregressive Integrated Moving Averages with eXogenous factors (SARIMAX) were used to predict the new COVID-19 cases in India between January 30, 2020 and December 23, 2020. The best model for prediction was found to be ARIMA with a root mean squared error (RMSE) of 2773.27. In [11], support vector machine and polynomial regression models were adopted to analyze COVID-19 data and trends in India between January 22, 2020 and June 24, 2020. The study found that the polynomial regression model was better at predicting the data with an accuracy of 93%.

In [12], different machine learning models, such as linear regression, polynomial regression, and support vector machine, were used to predict the number of global confirmed cases, deaths, and cardiovascular related deaths due to COVID-19. It was found that linear regression generated results close to the actual data. Mary et al. [13] proposed that classification models such as support vector machine, K-nearest neighbor, and Naïve Bayes could be used to predict the number of positive cases of COVID-19 in India. The project used both numerical and categorical variables for classification. The experimental results showed that support vector machine generated the best results with an accuracy of 85%.

In [14], a Susceptible-Exposed-Infectious-Recovered (SEIR) mathematics model in combination with the machine learning model of polynomial regression was implemented to predict and analyze the trend of COVID-19 confirmed cases in Xinjiang, China. The results of this project showed that the confirmed COVID-19 cases could be accurately predicted in a short term. Nurrahma et al. [15] utilized different machine learning classification techniques, i.e., support vector machine, decision tree, and neural network, to predict the different symptoms of COVID-19. It was found that the neural network model produced the best overall results for prediction of symptoms with an average accuracy of 97.10%. In [16], machine learning models such as linear regression, polynomial regression, and support vector regression were adopted to model the number of confirmed cases of COVID-19 in Saudi Arabia and Bahrain. The deep learning model of Long Short-Term Memory (LSTM) was also used to predict the number of cases, deaths, and
recoveries in both countries. The results of this project showed that support vector regression was the best model for Saudi Arabia and linear regression was the best model for Bahrain. In [17], the machine learning algorithms of ARIMA, artificial neural network, LSTM, and convolutional neural network were utilized to perform a 7-day prediction of COVID-19 cases for 189 different countries. It was found that convolutional neural network performed the best out of those four models.

Hossen et al. [18] proposed that three machine learning models, i.e., random forest, support vector machine, and k-nearest neighbors, could be utilized to predict the recovery rate of patients affected by COVID-19 based on their eating habits. The experimental results showed that certain foods were more likely to increase patients’ recovery rate related to COVID-19. In [19], an artificial neural network machine learning technique was developed to analyze the COVID-19 data in India for the cumulative confirmed cases, daily confirmed cases, and cumulative confirmed deaths. The case studies showed that each model was highly accurate at predicting the data with the mean absolute percentage error (MAPE) values of 3.981, 4.173, and 4.413 for cumulative confirmed cases, daily confirmed cases, and cumulative confirmed deaths respectively. In [20], support vector regression and polynomial regression were utilized to predict the global number of confirmed cases, confirmed deaths, recovered cases, and daily cases of COVID-19 between March 1, 2020 and April 30, 2020. The results of the study showed that the support vector regression models were able to predict the data more accurately for each category than the polynomial regression models. Podder et al. [21] used the machine learning techniques such as random forest, XGBoost, and logistic regression in a stacking ensemble to predict the number of COVID-19 patients and the intensive care unit required for a hospital in Brazil. The results of the project showed that the number of COVID-19 patients could be predicted with a 94.39% accuracy and the intensive care unit requirement could be predicted with a 98.13% accuracy.

All these works used machine learning techniques to analyze and predict the COVID-19 data for various countries and cities across different timeframes. While some works used the same techniques or studied the same countries as our project, most of these works used the COVID-19 data from before the vaccine was available. Vaccination can effectively change the spread of COVID-19 patterns. Our project focuses on analyzing four countries selected globally and the most recent data which includes COVID-19 data from after the vaccine was available.
Methodology

In this project we utilized two supervised machine learning techniques, i.e., Holt’s double exponential smoothing and ARIMA to predict the COVID-19 cases and deaths. Holt’s double exponential smoothing is a derivative of the exponential smoothing technique. Exponential smoothing assigns weights to previous data points that exponentially decrease over time. The value of the current data point called the level is calculated using a smoothing factor. The smoothing factor ranges from 0 to 1. If the smoothing factor is closer to 1, more weight is given to that data point. Typically, data close to the current level have a smoothing factor close to 1 and data further away from the current level have a smoothing factor close to 0. Holt’s double exponential smoothing uses a trend in addition to the level. The trend is expressed as the difference between the previous two levels and shows whether the data are increasing or decreasing. The trend is also calculated based on a smoothing factor that can range from 0 to 1. A trend smoothing factor close to 1 means that more weight will be put towards the recent trends of the data. The current level and trend are combined to calculate the forecast of the model which is the actual prediction value.

Equation (1) defines Holt’s double exponential smoothing formula for calculating the level of the model. The formula consists of \( x \) as the observation, \( t \) as the time, \( S \) as the level of the data, \( \alpha \) as the smoothing factor of the level, and \( B \) as the trend of the data [2]:

\[
S_t = \alpha x_t + (1 - \alpha)(S_{t-1} - B_{t-1}) \tag{1}
\]

Equation (2) defines Holt’s double exponential smoothing formula for calculating the trend of the model. Equation 2 uses of similar variables to Equation 1 with an additional smoothing factor of the trend as \( \beta \) [2]:

\[
B_t = \beta(S_t - S_{t-1}) + (1 - \beta)B_{t-1} \tag{2}
\]

Equation (3) defines Holt’s double exponential smoothing formula calculating the forecast of the model. The formula combines the results of Equations 1 and 2 to compute the final forecast value \( F \) at time \( t = (N+m) \) and \( m \) as the sum of the last forecast plus last trend [2]:
\[ F_{N+m} = S_t + mB_t \]  \hspace{1cm} (3)

ARIMA [3] is a forecasting model that has been widely used in many application domains. ARIMA integrates autoregressive (AR) model and moving average (MA) model. ARIMA model is generally denoted as ARIMA\((p,d,q)\) where, \(p\) is the order of auto-regression, \(d\) is the degree of difference and \(q\) is the order of moving average. When a dataset is stationary, the value of \(d\) is 0. Equation (4) is the formula for ARMA model [3].

\[ Y_t = \sum_{i=1}^{p} \Phi_i Y_{t-i} + a_t - \sum_{j=1}^{q} \theta_j a_{t-j} \]  \hspace{1cm} (4)

\(Y\) is the output variable, \(t\) is time, \(\Phi\) is the autocorrelation coefficient, \(a\) is the error residuals, \(\theta\) is the weight given to current and previous value, \(p\) is the order of autoregression, and \(q\) is the number of lagged values. Equation 4 can be simplified by introducing the Box-Jenkins backshift operator, \(B\). Equation (5) shows the definition of \(B\) [3].

\[ B^p X_t = X_{t-p} \]  \hspace{1cm} (5)

Equation (6) is the result of substituting this relationship into the ARMA formula. Equations (7) and (8) can be used to further simplify the model. Equation (9) is the final simplified ARMA model [3].

\[ (1 - \sum_{i=1}^{p} \Phi_i B^i)Y_t = (1 - \sum_{j=1}^{q} \theta_j B^j)a_t \]  \hspace{1cm} (6)

\[ \Phi_p(B) = (1 - \sum_{i=1}^{p} \Phi_i B^i) \]  \hspace{1cm} (7)

\[ \theta_q(B) = (1 - \sum_{j=1}^{q} \theta_j B^j) \]  \hspace{1cm} (8)

\[ \Phi_p(B) Y_t = \theta_q(B) a_t \]  \hspace{1cm} (9)

When a dataset is nonstationary, the difference between predicted values can be used to simulate stationary behavior. Equation (10) shows an order of differencing of 1. Equation (11) shows the formula for finding the difference for an order of differencing \(d\) [3].
\[ W_t = Y_t - Y_{t-1} = (1 - B)Y_t \]  
(10)

\[ W_t - \sum_{k=1}^{d} W_{t-k} = (1 - B)^dY_t \]  
(11)

This differencing relationship can be added to the ARMA model. The result of this is the final ARIMA\((p,d,q)\) model which is used to forecast the data. Equation (12) shows the formal definition of the ARIMA model [3].

\[ \Phi_p(B)(1 - B)^dY_t = \theta_q(B)a_t \]  
(12)

**Experimental Results**

**A. COVID-19 Global Data**

We collected the global COVID-19 data between January 23, 2020 and January 21, 2022 from Kaggle [4]. The data were reported by John Hopkins University with the number of new daily cases and number of new daily deaths related to COVID-19 for every country in the world. Figure 1 displays the top ten countries’ average daily cases by January 21, 2022 based on John Hopkins University data. The USA, Brazil, and India ranked as the top three countries with at least 30,000 new cases per day, and Iran ranked no. 10 with 10,000 new cases per day. Figure 2 shows top ten countries with the number of average daily deaths since January 2020 based on the John Hopkins University data. The USA, India, and Brazil ranked as the top three countries with at least 600 deaths per day. Iran is ranked no. 9 with 200 new deaths per day. For this project, we chose four countries to analyze, i.e., USA, Brazil, India, and Iran.

We used Python 3 in Jupyter Notebook for experimental study. The data was preprocessed by replacing all missing values with 0. We computed the 14-day rolling average of new daily cases and new daily death for each country. For Holt’s double exponential smoothing, the built-in Holt function was used to compute the smoothing factors for both the level and trend with the best fit for the data. For ARIMA, an auto-ARIMA function was used to determine the values of the parameters \(p\), \(d\), and \(q\). Similar to the Holt function, the auto-ARIMA function calculates the value of each parameter that will create a model with the best fit of the data. Those parameters were used in the built-in ARIMA function to create the prediction models.
The accuracy of each model was evaluated using several metrics, i.e., mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), $R^2$, and adjusted $R^2$. Both $R^2$ and adjusted $R^2$ measure how close the model is to the actual data. For $R^2$ and adjusted $R^2$, a value closer to one indicates a better fit of the model. For mean absolute error, mean squared error, and root mean squared error, a value closer to zero indicates a better fit.

In the following sections we discuss models for each selected country individually. Each graph shows three lines presenting the training data, the prediction data, and the model, respectively. The solid black line shown on each figure represents the training data used to create the model with dates ranging from January 23, 2020 to January 7, 2022. The solid red line shown on each figure represents the actual data between January 8, 2022 and January 21, 2022 that the model predicts. The dashed yellow line shown on each figure represents the predicted values of the model.

B. USA Models

Figure 3 shows the prediction model created using the Holt’s double exponential smoothing technique with a level smoothing factor of 0.9950 and a trend smoothing factor of 0.4027. Figure 4 shows the prediction model created using the ARIMA technique with a $(p, d, q)$ of $(3, 2, 0)$.
For the USA’s daily cases, both the Holt’s double exponential smoothing and the ARIMA models perform well with the R-squared values close to 1 and low root mean squared error values. Table I shows the results of our evaluation metrics for fitting the USA’s daily cases using Holt’s Double Exponential Smoothing (HDES) and ARIMA.

Table I: Evaluation Metrics for the USA’s Daily Cases

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>R²</th>
<th>R²_adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDES</td>
<td>3279.77</td>
<td>462231127</td>
<td>21499.56</td>
<td>0.9624</td>
<td>0.9624</td>
</tr>
<tr>
<td>ARIMA</td>
<td>3336.62</td>
<td>468454106</td>
<td>21643.80</td>
<td>0.9619</td>
<td>0.9619</td>
</tr>
</tbody>
</table>

Figure 5 shows the prediction model created using the Holt’s double exponential smoothing technique with a level smoothing factor of 0.9950 and a trend smoothing factor of 0.2842. Figure 6 shows the prediction model created using the ARIMA technique with a \((p, d, q)\) of \((2, 1, 1)\).

For the USA’s daily deaths, both the Holt’s double exponential smoothing and the ARIMA models produced good models with R-squared values close to 1 and low root mean squared error values. Table II shows the results of our evaluation metrics for fitting the USA’s daily deaths using HDES and ARIMA.
OFFICE OF UNDERGRADUATE RESEARCH – OUR GRANT RECIPIENTS 2021-22

Fig. 5. Holt’s Double Exponential Smoothing Model for the USA’s Daily Deaths

Fig. 6. ARIMA Model for the USA’s Daily Deaths

Table II. Evaluation Metrics for the USA's Daily Deaths

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDES</td>
<td>20.63</td>
<td>1547.08</td>
<td>39.33</td>
<td>0.9976</td>
<td>0.9976</td>
</tr>
<tr>
<td>ARIMA</td>
<td>19.81</td>
<td>1515.04</td>
<td>38.92</td>
<td>0.9976</td>
<td>0.9976</td>
</tr>
</tbody>
</table>

C. Brazil Models

Figure 7 shows the prediction model created using the Holt’s double exponential smoothing technique with a level smoothing factor of 0.9950 and a trend smoothing factor of 0.2369. Figure 8 shows the prediction model created using the ARIMA technique with a $(p, d, q)$ of $(1, 2, 1)$.

For Brazil’s daily cases, both the Holt’s double exponential smoothing and the ARIMA models produced good models with R-squared values close to 1 and low root mean squared error values. Table III shows the results of our evaluation metrics for fitting Brazil’s daily cases using HDES and ARIMA.
Table III. Evaluation Metrics for Brazil’s Daily Cases

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDES</td>
<td>869.17</td>
<td>13363196</td>
<td>3655.57</td>
<td>0.9725</td>
<td>0.9725</td>
</tr>
<tr>
<td>ARIMA</td>
<td>920.12</td>
<td>17819353</td>
<td>4221.30</td>
<td>0.9633</td>
<td>0.9633</td>
</tr>
</tbody>
</table>

Figure 9 shows the prediction model created using the Holt’s double exponential smoothing technique with a level smoothing factor of 0.9950 and a trend smoothing factor of 0.2606. Figure 10 shows the prediction model created using the ARIMA technique with a $(p, d, q)$ of $(2, 2, 2)$.

For Brazil’s daily deaths, both the Holt’s double exponential smoothing and the ARIMA models produced good models with R-squared values close to 1 and low root mean squared error values. Table IV shows the results of our evaluation metrics for fitting Brazil’s daily deaths using HDES and ARIMA.

Table IV. Evaluation Metrics for Brazil’s Daily Deaths

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDES</td>
<td>11.75</td>
<td>454.10</td>
<td>21.31</td>
<td>0.9991</td>
<td>0.9991</td>
</tr>
<tr>
<td>ARIMA</td>
<td>11.43</td>
<td>398.12</td>
<td>19.95</td>
<td>0.9992</td>
<td>0.9992</td>
</tr>
</tbody>
</table>

D. India Models

Figure 11 shows the prediction model created using the Holt’s double exponential smoothing technique with a level smoothing factor of 0.9950 and a trend smoothing factor of 0.9950. Figure 12 shows the prediction model created using the ARIMA technique with a $(p, d, q)$ of $(2, 1, 2)$. 
For India’s daily cases, both the Holt’s double exponential smoothing and the ARIMA models produced good models with R-squared values close to 1 and low root mean squared error values. Table V shows the results of our evaluation metrics for fitting India’s daily cases using HDES and ARIMA.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDES</td>
<td>897.63</td>
<td>36328714</td>
<td>6027.33</td>
<td>0.9936</td>
<td>0.9936</td>
</tr>
<tr>
<td>ARIMA</td>
<td>404.40</td>
<td>2294557.16</td>
<td>1514.78</td>
<td>0.9996</td>
<td>0.9996</td>
</tr>
</tbody>
</table>

Figure 13 shows the prediction model created using the Holt’s double exponential smoothing technique with a level smoothing factor of 1.0000 and a trend smoothing factor of 0.3178. Figure 14 shows the prediction model created using the ARIMA technique with a $(p, d, q)$ of (4, 1, 3).
For India’s daily deaths, both the Holt’s double exponential smoothing and the ARIMA models produced good models with R-squared values close to 1 and low root mean squared error values. Table VI shows the results of our evaluation metrics for fitting India’s daily deaths using HDES and ARIMA.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDES</td>
<td>12.67</td>
<td>1123.12</td>
<td>33.51</td>
<td>0.9986</td>
<td>0.9986</td>
</tr>
<tr>
<td>ARIMA</td>
<td>13.06</td>
<td>998.18</td>
<td>31.59</td>
<td>0.9987</td>
<td>0.9987</td>
</tr>
</tbody>
</table>

**E. Iran Models**

Figure 15 shows the prediction model created using the Holt’s double exponential smoothing technique with a level smoothing factor of 0.9950 and a trend smoothing factor of 0.7581. Figure 16 shows the prediction model created using the ARIMA technique with a $(p, d, q)$ of $(1, 1, 2)$.

For Iran’s daily cases, both the Holt’s double exponential smoothing and the ARIMA models produced good models with R-squared values close to 1 and low root mean squared error values for the size of the data. Table VII shows the results of our evaluation metrics for fitting Iran’s daily cases using HDES and ARIMA.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>$R^2$</th>
<th>$R^2_{adj}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARIMA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 17 shows the prediction model created using the Holt’s double exponential smoothing technique with a level smoothing factor of 0.9763 and a trend smoothing factor of 0.5415. Figure 18 shows the prediction model created using the ARIMA technique with a \((p, d, q)\) of \((1, 1, 1)\).
For Iran’s daily deaths, both the Holt’s double exponential smoothing and the ARIMA models produced good models with R-squared values close to 1 and low root mean squared error values for the size of the data. Table VIII shows the results of our evaluation metrics for fitting India’s daily deaths using HDES and ARIMA.

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE</th>
<th>MSE</th>
<th>RMSE</th>
<th>R²</th>
<th>R² adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDES</td>
<td>1.62</td>
<td>5.26</td>
<td>2.29</td>
<td>0.9997</td>
<td>0.9997</td>
</tr>
<tr>
<td>ARIMA</td>
<td>1.58</td>
<td>4.91</td>
<td>2.22</td>
<td>0.9998</td>
<td>0.9998</td>
</tr>
</tbody>
</table>

Based on our experiment results discussed in this section, the prediction models based on both Holt’s double exponential smoothing and ARIMA performed well. The lowest R-squared value was 0.9619 for the ARIMA model for the USA’s daily cases. This value also indicates a good fit. The small root mean squared error for each model also indicates a good fit. Overall, ARIMA models performed better than Holt’s double exponential smoothing models. However, this difference is not significant and both techniques were able to produce well fitted models.

**Conclusion**

COVID-19 has affected millions of lives around the world over the past two years and we are still struggling to combat the virus. Analyzing COVID-19 data can help us better understand the spread of the virus. In this project, we analyzed COVID-19 data reported by John Hopkins University for the USA, Brazil, India, and Iran, and developed prediction models for the new daily cases and new daily deaths for these four countries utilizing two machine learning techniques, i.e., Holt’s double exponential smoothing and ARIMA. Based on the experimental results, both techniques produced models that performed well. None of the models had a R-squared value of lower than 0.96. Both techniques produced similar performing models.

**Bibliography**


