ISSN 2278-3091

Volume 10, No.4, July - August 2021

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse211042021.pdf

https://doi.org/10.30534/ijatcse/2021/211042021

Data Analytics on the COVID-19 Outbreak in South Asia using Machine Learning Methods

Kajol Chandra Paul[1]*, Md. Ahsanul Hoque[2], Shubhra Mostafa Dhiman[3], Joynto Kumar Sen[4]

[1*]Jatiya Kabi Kazi Nazrul Islam University, Bangladesh, Email: kajolc.paul@gmail.com

[2]Jatiya Kabi Kazi Nazrul Islam University, Bangladesh, Email: ahsanshantanur@gmail.com

[3]Jatiya Kabi Kazi Nazrul Islam University, Bangladesh, Email: smdhiman9876@gmail.com

[4]University of Dhaka, Bangladesh, Email: joynto.eeedu@gmail.com

ABSTRACT

The spread of COVID-19 from the first case reported in China's Wuhan to a worldwide pandemic has been a tremendous topic of study among data analysts and scientists alike. In South Asia, this pandemic has brought about a disaster in the lives and livelihoods of most of its inhabitants. An exploratory analysis of the overall COVID-19 data provided by the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) is presented in this paper. The number of confirmed, death, recovered, and active cases as recent as July 18, 2021, have been explored with machine learning data analytics methods to draw crucial conclusions about the pandemic. The analytics and predictive modeling are performed in the context of the cases in the SAARC (South Asian Association of Regional Cooperation) countries. The calculated correlation coefficient demonstrates that the countries with higher GDP Per Capita have conducted more tests/1M population. To find and compare the severity of the pandemic, the countries are grouped based on the K-Means clustering algorithm. The confirmed and death cases are modeled with the polynomial regression technique and the future evolution of the pandemic is predicted with good accuracy. Based on the predictive model, the total cases estimate around 42.91 million confirmed and 0.58 million deaths till August 17, 2021.

Key words: COVID-19, data analytics, K-Means clustering, polynomial regression, predictive model, SAARC.

1. INTRODUCTION

Not long after the first documented case at Hubei province in China [1], the COVID-19 has been declared a global pandemic on 11th March 2020, by the World Health Organization (WHO) [2]. As of 18th July2021, the total global confirmed cases are around 191 million. The first reported case of infection among the SAARC countries was in Nepal, where the case was reported on 23rd January 2020. SAARC is a regional organization comprising of 8 neighboring countries, namely, Afghanistan, Bangladesh, Bhutan, India, Maldives, Nepal, Pakistan, and Sri Lanka. Often, the first confirmed cases in those countries have been found among the expatriates who returned from China, Italy, or Iran during the period January to March of 2020 [3], [4]. By July 2nd of that year, at least one case of COVID-19 was reported in every country. To analyze and compare the pandemic situation in SAARC countries, we cannot merely look at the sheer number of cases without considering the respective country's population. For example, India has the highest cumulative confirmed cases among all; still, it is 2.2% of its population. Contrarily, 13.8% population of Maldives has been infected with the virus, which ranks 7th in terms of confirmed cases.

The spread of the virus, number of deaths caused, number of tests conducted, cases of recovery depend on some underlying factors such as the population density, testing facilities, availability of hospital and ICU beds. Each of the countries has a different testing capacity and health care facility leading to a varying number of confirmed and death cases, which can essentially be related to its GDP Per Capita, health care index (HCI), the number of vaccination, and population age groups. With data analytics, it is possible to extract valuable information from real data and make a forecast on the future course of the pandemic. Several data-analytics-based studies have been carried out on the epidemiological outbreak of the coronavirus in China and some other countries [5]-[9]. Using K-Means clustering, countries or provinces of a country can be grouped to determine and contrast the severity of infections between the group members [10], [11]. Mathematical models such as the SIR model have been used to forecast the spread of the pandemic in countries like China [12], India [13], Sri Lanka [14], and the European countries [15]. Machine learning algorithms such as the linear, polynomial, and logistic regression techniques are also widely used to model the pandemic data and create a forecast on future cases [16]-[20]. In [21], fuzzy logic system has been used for the prediction of COVID-19 cases in South Africa and Egypt.

In this paper, we have presented a comprehensive and exploratory data analysis including visualization and forecast for the COVID-19 outbreak in SAARC countries. The analytical work examines both diagnostic and predictive features. It includes visualization concerning bar charts, scatter plots, bubble plots, and other types of graphs on the documented cases and other manipulated data. A statistical correlation is performed on the number of tests and death cases in terms of GDP Per Capita, HCI, the number of vaccinated people, population, and their age groups. Using K-Means clustering, the countries are segmented into groups of varying severity regarding confirmed, death, and active cases. We have also presented a prediction model based on polynomial regression for the number of infections and deaths till August 17, 2021. This way, we have calculated the future size of the Coronavirus pandemic in South Asia with a satisfactory level of evaluation metrics.

2. MATERIAL AND METHODS

The Python programming language is used for the visualization, analysis, and modeling of the data. The primary dataset employed in this research is collected from Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). The data is available on a public GitHub page: https://github.com/CSSEGISandData/COVID-19. JHU CSSE collects the data from primary sources such as the WHO, local and national health institutions, which are updated daily [22]. On their GitHub page, data of the confirmed, deaths, and recovered cases are time-series data given in separate CSV files. However, for ease of our work, we have brought all the data into the same CSV file datasheet. The dataset has been processed and cleaned using libraries like **NumPy** and **Pandas**. Manipulating these data, we have added more attributes to the dataset such as the active cases, daily new cases, daily new deaths, and daily new recovered cases. The secondary data source used here is https://www.worldometers.info/coronavirus/#countries, from where data such as the number of testing and the population of the countries are collected.

3. DATA-DRIVEN DIAGNOSIS

The diagnostic data analytics has been performed in terms of the reported cases, economic, and health factors. The distribution of the COVID-19 confirmed, deaths, recovered, and active cases for the SAARC countries along with their respective population are listed in Table 1. In this section, we have sought to identify the hidden factors behind the differences in the pandemic performance of the countries.

India constitutes 90.5% of the total confirmed cases. On the other hand, it shares only 74.2% of the total population of SAARC. In comparison, Pakistan and Bangladesh share 11.9% and 8.9% of the SAARC population, while their share of COVID-19 confirmed cases are 2.9% and 3.2%, respectively. The spread of Coronavirus has not been uniform or in proportion with the population distribution in the region. For example, Nepal has approximately 10 million fewer people than Afghanistan, but the confirmed cases in Nepal are nearly 5 times the cases in Afghanistan. It does not indicate that Afghanistan did exceptionally well to stop the spread rather; the lower number of confirmed cases can be attributed to the significantly lower number of testing conducted in Afghanistan.

There are several ways to calculate the case fatality rate (CFR) of a pandemic. Generally, CFR is the proportion of deaths among individuals diagnosed with a specific disease. In this approach of calculation, the outcomes of active cases are not considered. Since the pandemic is ongoing, the final CFR will change after the active cases are resolved. As a solution, the CFR can be calculated by dividing the total deaths by the total closed cases (deaths + recovered) [16]. Hence, the formulas to calculate the CFR are

$$CFR = \frac{D}{C} \quad (1)$$

$$FCFR = \frac{D}{D + R} \quad (2)$$

where *FCFR* stands for final case fatality rate. *C*, *D*, and *R* represent confirmed, death, and recovered cases respectively. The recovery rates are calculated the same way. Figure 1 shows the calculated fatality and recovery

Table 1: The country-wise distribution of COVID-19 cases in SAARC.

Country	Confirmed	Deaths	Recovered	Active	Population (mil.)
India	31144229	30308456	414108	421665	1394.12
Bangladesh	1103989	932008	17894	154087	166.39
Pakistan	991727	920066	22811	48850	225.34
Nepal	667109	632074	9550	25485	29.68
Sri Lanka	284932	256676	3779	24477	21.51
Afghanistan	137853	82586	5983	49284	39.83
Maldives	75879	73226	216	2437	0.55
Bhutan	2421	2042	2	377	0.78

rates. Initially, the highest recovery rate is observed in India which is 97.3% and the lowest recovery rate of 59.9% is seen in Afghanistan. On the other hand, Bhutan has the lowest fatality rate overall (approximately 0.1%) whereas Afghanistan tops the list. It is seen that Afghanistan's final case fatality rate is a staggering 6.8%, which is significantly higher than the rest of the SAARC countries. It is to be noted, the initial fatality and recovery rates do not add up to 100%, though the final rates do. Consequently, the final fatality and recovery rates can be interpreted probabilistically.

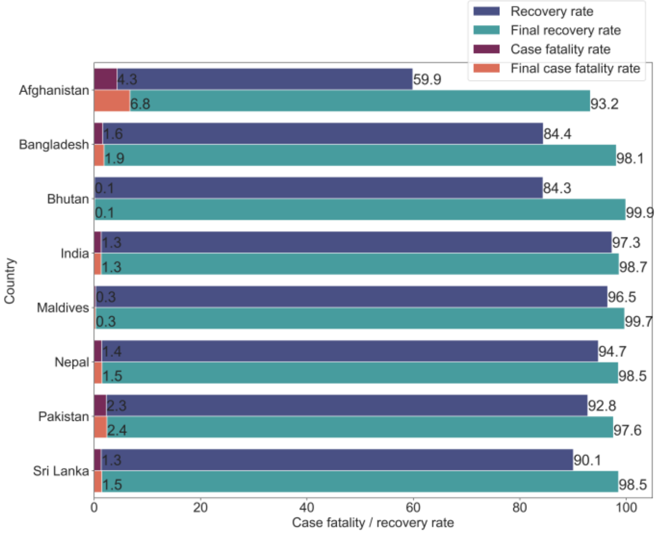


Figure 1: Initial recovery and case fatality rates are calculated without considering the active cases, whereas the final rates are calculated based on the resolved cases.

The underlying factors of the confirmed cases, often overlooked, are the number of tests conducted and the test positivity rate. The testing capability of a country depends on many aspects such as the number of testing labs, number of kits and technicians available, accessibility to the testing facility, etc. We have attempted to relate the number of tests conducted by each country to its GDP Per Capita. On the other hand, the factors such as age, vaccination, comorbidities, and HCI could have an impact on death cases. The HCI data [23] gives an estimation of the overall quality of the health care system in the country. In addition, the fatality rate from COVID-19 increases substantially for the population aged 65 and above [24]. The data for HCI, GDP Per Capita [25], number of people vaccinated with at least a single dose [26], and 65+ population [27] are collected and their relationship with the death cases are quantified based on the Pearson correlation coefficient.

The bubble plot in Figure 2 shows the tests/1M population conducted by each country. The correlation coefficient found between the number of tests and GDP Per Capita is 0.90, which suggests that the countries with higher GDP Per Capita have conducted more tests. The notable exception here is Sri Lanka, which ranks second on the GDP Per Capita. Despite that, it carried out fewer tests than that of Maldives, Bhutan, and India. On the other hand, the number of tests is found to be moderately negatively associated with the test positivity rate as the correlation coefficient between them is -0.50. This points to the fact that countries with higher positivity rates did not perform a sufficient amount of tests. For example, Afghanistan with the highest positivity rate i.e., 20.4% conducted the lowest number of tests i.e., 16986 tests/1M population, contributing to their lower confirmed cases.

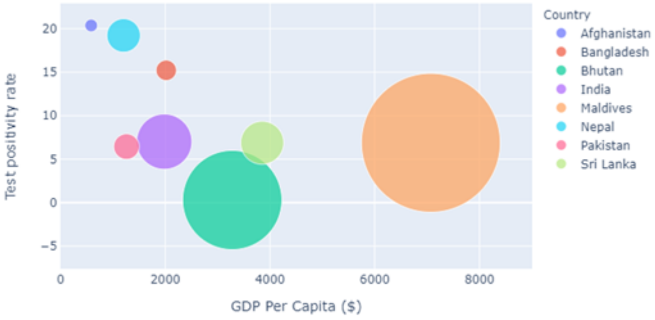


Figure 2: The number of tests conducted per 1M population (indicated by the bubble size) depends strongly on the GDP Per Capita.

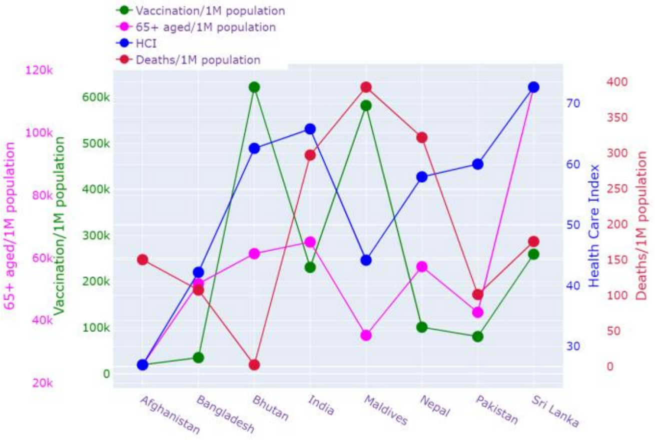


Figure 3: The number of deaths exhibits a positive linear association with population age and vaccination, whereas HCI has an unsubstantial negative linear correlation with death counts.

Similarly, the correlation coefficient found for the number of deaths concerning HCI, the number of vaccinated people, and population age are -0.04, 0.30, and 0.28 respectively. There is an absence of a perfect linear relationship between these factors. From Figure 3, it is evident that Maldives, despite having a small 65+ population size and high vaccination record has witnessed the highest count of nearly 392 deaths/1M population. The reason could be its inadequate health care facilities as it belongs to the bottom three countries on HCI rating. On the contrary, India is below Maldives and Nepal in deaths/1M population although it has the largest age group of 65+ years. The fact that it has a comparatively higher HCI and better vaccination record has contributed to a relatively lower death count.

4. K-MEANS CLUSTERING METHOD

An important way to compare the severity of the COVID-19 pandemic in different countries could be K-Means clustering based on various case criteria. The SAARC countries have been clustered based on the number of confirmed, death, and active cases per 1M population. The K-Means clustering is an unsupervised machine learning algorithm that divides the data into K clusters. There are several methods to find the optimal number of clusters such as the Elbow method, Silhouette method, and Gap statistic [11]. In this study, we have performed the K-Means clustering for a range of clusters K in iteration and calculated the optimal K value with the Elbow method. Figure 4 depicts the within-cluster sum of squares (WCSS) values for different cluster numbers. WCSS measures the sum of squared distances of each sample to their nearest cluster center.

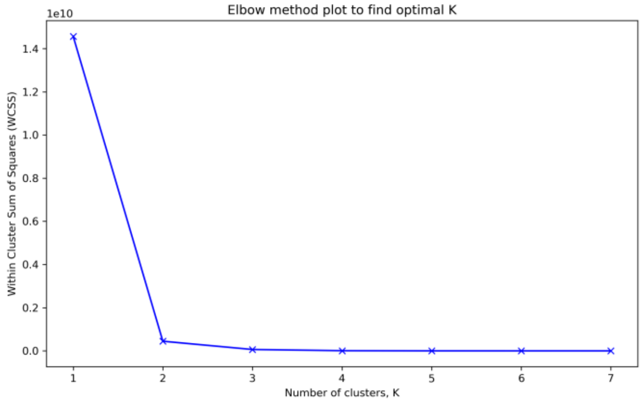
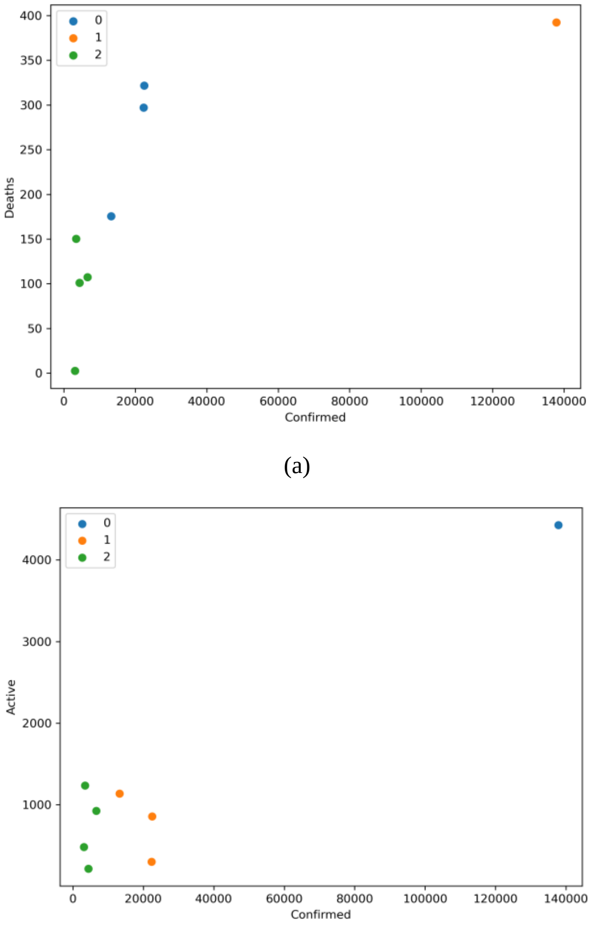


Figure 4: Elbow method plot of WCSS/inertia values to find the optimal number of clusters. This operation is performed based on the data of confirmed and death cases.

The value of K at the 'elbow' point is 3, as the inertia starts to decrease linearly from that point. Hence, the optimal value of K is 3 which means the data points can be grouped into 3 clusters. Consequently, the SAARC nations are segmented into 3 groups according to the number of confirmed and death cases. Furthermore, the Elbow method performed based on the data of confirmed and active cases determines the optimal value of K as 3. The graphical representation of the K-Means clustering is shown in Figure 5. In Figure 5(a), cluster 0 represents the countries with a high number of deaths but a low number of confirmed cases. India, Nepal, and Sri Lanka belong to this cluster. Maldives, which belongs to cluster 1 is the worst affected nation with a very high number of confirmed and death cases. Eventually, cluster 2 comprising of the rest of SAARC countries has a very low number of confirmed and death cases. Likewise, cluster 0 in Figure 5(b) depicts the worst affected country which is Maldives. The countries belonging to cluster 1 are moderately affected as in the case of India, Nepal, and Sri Lanka. Finally, cluster 2 is the set of countries comprising Afghanistan, Bangladesh, Pakistan, and Bhutan. These countries are less affected with a low number of confirmed and active cases.



(a)

(b)

Figure 5: SAARC nations grouped according to (a) confirmed-deaths and (b) confirmed-active cases using K-Means clustering.

5. PREDICTIVE MODELING OF THE PANDEMIC

To model the pandemic data, we have used regression analysis, which is a supervised machine learning method. The regression method attempts to model a continuous output variable based on one or several input variables [28]. It finds the relationship between the independent and dependent variables and builds a predictive model that can forecast future outcomes. There are different types of regression analysis techniques such as linear, polynomial, logistic growth model, and others. Linear regression is a simple algorithm and it cannot perform well when the data is complex and shows a curvilinear pattern. Therefore, applying linear regression to model the non-linear dataset does not yield accurate results.

5.1 Polynomial regression

In this paper, we have performed the data modeling based on polynomial regression, where a polynomial equation of n^th degree is used to model the non-linear dataset. The generic n^th degree polynomial equation is expressed as

$$Y = a + bX + cX^2 + dX^3 + \cdots\cdots nX^n \tag{3}$$

Where *a*, *b*, *c*, *d*, …., *n* are called the parameters of polynomial regression analysis. In our model, the number of days is the independent variable X, which we count from the starting date of our dataset (i.e., 22.01.2020) and the number of confirmed/death cases is the dependent variable or predictor *Y*. The degree of a polynomial is determined by the highest exponent of the independent variable. Some of the common polynomial functions are

1. Zero polynomial function $$Y = a \tag{4}$$
2. First-degree polynomial (Linear) $$Y = a + bX \tag{5}$$
3. Quadratic polynomial function $$Y = a + bX + cX^2 \tag{6}$$
4. Cubic polynomial function $$Y = a + bX + cX^2 + dX^3 \tag{7}$$

The first-degree polynomial is a straight line, known as a linear function where b is the slope and a is the Y-intercept of the line. In a similar fashion, the mathematical expression for higher-order polynomial regression can be found from Eq. (3). The higher order of the polynomial does not necessarily signify a better level of fitting i.e., the closeness between the observed sample values and the predicted values. The best possible fitted regression line can be of any order.

5.2 Evaluation metrics

Generally, the statistical measures used for evaluation of the regression model are: Sum of Squared Error (SSE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE). The finest regression line minimizes these values most. However, these metrics penalize for the big error, and consequently, Mean Absolute Error (MAE) could be used. Besides that, another metrics called R-squared (R^2) or the coefficient of determination is also used. It is an indication of the goodness of fit for the data points to the regression model. The formula used for the calculation of R^2 is [17]

$$R^2 = 1 - \frac{SS_{line}}{SS_{total}} \tag{8}$$

Where SS_{line} indicates the sum of squared error between the observed output values and the fitted regression line and SS_{total} indicates the total variation in *Y* values i.e., the sum of the squared differences between the observed output values and their mean. The value of R^2 ranges from 0 to 1. When SS_{line} is very small, R^2 is close to 1, which means the regression line perfectly fits the actual data points. For practical use, the modified version of R^2 which is called the adjusted R^2 is a better metric. While R^2 does not account for the additional input variables that are statistically insignificant, adjusted R^2 adjusts the statistical model based on the number of observations and variables. Thus, it is more useful to evaluate the predictive power of the model with adjusted R^2. At times, the model tries to fit every data points exactly, especially for higher degrees of the polynomial function. The situation is known by the name overfitting, where the model picks up too much noise and makes fatal errors when predicting unknown data. An effective way to solve the problem is to utilize cross-validation (CV), where the training dataset is divided into k-folds. At each iteration, the model is trained on k-1 folds, keeping the remaining fold for testing. The performance is measured as the mean of the values in each iteration, which is indicated by the cross-validation metric or CV score.

5.3 Experiment

With the regression process algorithm described in [16], the modeling experiment is performed in Python's **sklearn** machine learning library. The entire dataset is divided into two subsets of training and testing data. As we want to make the prediction modeling with polynomial regression, the input data points i.e., days are converted into polynomial features of different degrees. The model is fit on the training data and prediction is done on the testing data. The performance of the model changes slightly with the size of the training and testing data. Eventually, the accuracy of the predictions is evaluated in terms of the MAE, adjusted R^2, and CV score. The CV score is computed for 5 splits and their mean is taken. Table 2 illustrates the values of the evaluation metrics based on the testing size and degree of the polynomial function used to model the total deaths in SAARC.

Table 2: Evaluation metrics for the polynomial regression model of the total deaths in SAARC

Evaluation metrics	Polynomial model of degree 6		Polynomial model of degree 13	
	10% test size	20% test size	10% test size	20% test size
Adjusted R^2	0.9923	0.9914	0.9881	0.9863
MAE	9035.24	8976.31	11650.72	12294.90
CV score	0.9900	0.9885	0.9869	0.9864

It is seen that the adjusted R^2decreases slightly if the testing size is increased to 20%. In that case, there is less training data to train the model. It is also seen that a higher degree of the polynomial function does not necessarily yield better values of the evaluation metrics. Figure 6 and Figure 7 illustrate the predictive modeling of the cumulative confirmed and death cases in SAARC. In the figures, the solid green curve represents actual data, while the blue line represents the model that has been fit on the actual data points. The blue prediction curve is extended for the next 30 days from July 18, 2021 (depicted by the dashed vertical line).

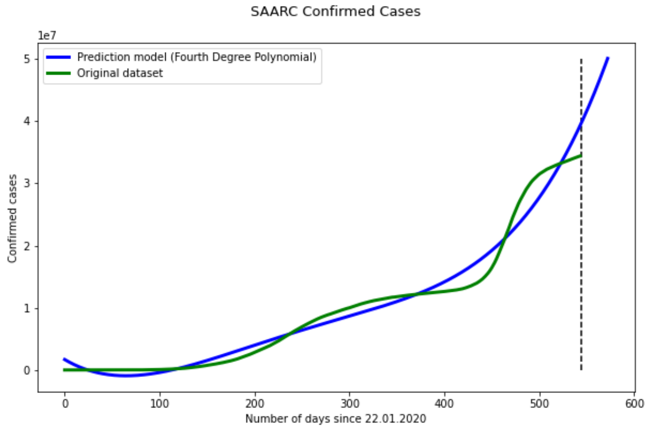


Figure 6: The total confirmed cases in SAARC are modeled with a fourth degree polynomial regression function.

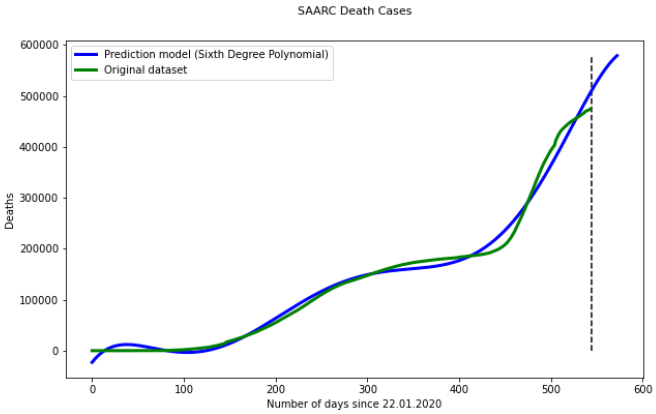


Figure 7: Predictive modeling of total deaths in SAARC exhibits a better coefficient of determination.

Assuming the current pattern sustains, the projection is made according to the best evaluation metrics achievable. We have performed similar modeling for the total confirmed and death cases in each of the countries of SAARC. The predicted cases as well as the performance metrics of the respective model are summarized in Table 3. In the next section, the important findings are presented along with a discussion on the merits and limits of the study.

Table 3: Predicted confirmed and death cases as of August 17, 2021.

Country	Cumulative confirmed cases				Cumulative deaths			
	Adjusted R^2	MAE	CV score	Predicted cases	Adjusted R^2	MAE	CV score	Predicted cases
India	0.9496	1681648.91	0.9539	38734943	0.9909	8413.28	0.9879	511467
Bangladesh	0.9956	16113.68	0.9945	1384232	0.9962	215.60	0.9964	21947
Pakistan	0.9926	21954.70	0.9925	1135729	0.9922	515.88	0.9925	27177
Nepal	0.9767	23113.02	0.9716	935114	0.9841	225.40	0.9759	12072
Sri Lanka	0.9829	7095.99	0.9846	382869	0.9922	51.38	0.9932	5011
Afghanistan	0.9832	2961.61	0.9857	252467	0.9878	118.35	0.9854	10867
Maldives	0.9848	1665.23	0.9788	89159	0.9742	5.56	0.9731	266
Bhutan	0.9920	42.68	0.9902	2750	0.9131	0.10	0.9039	3
Total cases				42917263				588810
SAARC	0.9713	1270958.53	0.9726	50512821	0.9923	9035.24	0.9900	580622

6. RESULTS AND DISCUSSION

Data analytics and modeling is a great tool to comprehend the spread of COVID-19 and its impact. The documented cases are diagnosed to unmask the underlying factors and the pattern of the pandemic is modeled with polynomial regression to predict the future course. The characteristics of the Coronavirus spread in South Asia and the related findings are encapsulated through the following points.

- Infections from COVID-19 spread all over South Asia in a non-uniform pattern. Despite being the least populated country in the region, Maldives has witnessed infections encompassing 13.8% of its population.
- Usually, the countries that rank higher up in GDP Per Capita have conducted more tests except for Sri Lanka. Afghanistan has conducted very few tests despite having a test positivity rate of over 20%, which effected its relatively low number of confirmed cases.
- The final case fatality rate calculated after considering resolved cases is found to be 6.8% in Afghanistan, which is the highest in South Asia. There is no definitive correlation between the number of deaths and

HCI, though a weak positive correlation is found with vaccination and 65+ age population data.

- As per confirmed, death, and active cases, the SAARC countries can be grouped into 3 clusters using the K-Means clustering algorithm. Maldives belongs to the severely affected cluster while India, Nepal, and Sri Lanka belong to the cluster which is moderately affected. The rest of the countries can be categorized as less affected.
- The polynomial regression model performed on the confirmed cases and deaths exhibit good accuracy overall. The adjusted R^2 value is over 0.9700 for all the instances of the model, except for Bhutan's death cases and India's confirmed cases. This can be interpreted as 97% of the total variations in the Y values are described by the fitted regression line.
- Bhutan has two cases of fatality from COVID-19, which occurred on January 8, and July 15 of 2021. This creates two abrupt steps in the curve and it becomes relatively imprecise to model these abrupt changes with polynomial regression. We observe many outlier data points on the modeling curve in the case of India and Nepal, which have resulted in a relatively lower adjusted R-squared value. The outlier data points often come from dumping previous-day data into the next day due to delays in data collection.
- If we sum up the predicted death cases for the individual country models, we get a number close to the total predicted cases in the SAARC model. The difference is only 1.4%. However, the difference of confirmed cases between individual country models and the SAARC model is 17.7%. The relatively lower adjusted R^2 scores have introduced this disagreement in the confirmed cases.

This study has enabled it to gather deep insights into the pandemic situation in South Asian nations. It can guide the research and shape the decision-making process in tackling the virus. According to the predictive model, the total estimated confirmed cases till August 17, 2021, are 42.91 million while total fatalities are 0.58 million.

7. CONCLUSION

The spread of the Coronavirus in the South Asian region may appear one-dimensional because of the sheer number of confirmed and death cases in India. However, the data analytics conducted in this study has shown diversified dynamics on the COVID-19 outbreak in SAARC countries. Be it the death rate, recovery rate, test positivity rate, or the number of testing done, each country has performed differently which can be correlated to their population, GDP Per Capita, vaccination number, and HCI. Countries such as Maldives are severely affected, according to the K-Means clustering. It is shown here that the number of confirmed or death cases can be modeled by polynomial regression techniques with good accuracy. The proposed regression models set forth a forecast for the future confirmed and death cases.

REFERENCES

1. J. Bryner, 1st known case of coronavirus traced back to November in China, Available at: https://www.Livescience.com/first-case-coronavirus-found.html, accessed May 2021.
2. Timeline: WHO's COVID-19 response, Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline, accessed May 2021.
3. N. Banka, Explained: How SAARC countries are fighting COVID-19, Available at: https://indianexpress.com/article/explained/explained-how-saarc-countries-are-fighting-covid-19-6331509/, accessed June 2021.
4. Reported Cases and Deaths by Country or Territory, Available at: https://www.worldometers.info/coronavirus/, accessed June 2021.
5. S. K. Dey, M. M. Rahman, U. R.Siddiqi, and A. Howlader. **Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach,** *Journal of Medical Virology*, vol. 92, no. 6, pp. 632–638, 2020.
6. H. Nishiura, S-m. Jung, N. M. Linton, R. Kinoshita, Y. Yang, K. Hayashi, T. Kobayashi, B. Yuan, and A. R. Akhmetzhanov. **The extent of transmission of novel Coronavirus in Wuhan, China**, *Journal of Clinical Medicine*, vol. 9, no. 2, pp. 330, 2020.
7. N. AL-Rousan, and H. AL-Najjar. **Data analysis of coronavirus COVID-19 epidemic in South Korea based on recovered and death cases**, *Journal of Medical Virology*, vol. 92, pp. 1603–1608, 2020.
8. T. Chakraborty, and I. Ghosh. **Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis**, *Chaos, Solitons & Fractals*, vol. 135, 109850, 2020.
9. S. K. Dey, M. M. Rahman, U. R. Siddiqi, and A. Howlader. **Exploring epidemiological behavior of novel coronavirus (COVID-19) outbreak in Bangladesh**, *SN Comprehensive Clinical Medicine*, vol. 2, pp. 1724–1732, 2020.
10. S. K. Saini, V. Dhull, S. Singh, and A. Sharma. **Visual Exploratory Data Analysis of COVID-19 Pandemic**, *5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, Jaipur, 2020, pp. 1-6.
11. D. Abdullah, S. Susilo, A. S. Ahmar, R. Rusli, and R. Hidayat. **The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data**. *Quality & Quantity*, 2021. DOI: https://doi.org/10.1007/s11135-021-01176-w.

12. C. Anastassopoulou, L. Russo, A. Tsakris, and C. Siettos. **Data-based analysis, modeling and forecasting of the COVID-19 outbreak**, *PLOS ONE*, vol. 15, no. 3, e0230405, 2020.

13. M. K. Kakkar, M. Sood, B. Sharma, and J. Bhatti. **Mathematical modeling and forecasting the spread of Covid-19 using Python**. *IJSTR*, vol. 9, no. 3, 2020.

14. S. Dharmaratne, S. Sudaraka, I. Abeyagunawardena, K. Manchanayake, M. Kothalawala, and W. Gunathunga. **Estimation of the basic reproduction number (R0) for the novel coronavirus disease in Sri Lanka**, *Virology Journal*, vol. 17, 144, 2020.

15. K. Linka, M. Peirlinck, and E. Kuhl. **The reproduction number of COVID-19 and its correlation with public health interventions**, *Computational mechanics*, vol. 66, pp. 1035–1050, 2020.

16. M. R. H. Mondal, S. Bharati, P. Podder, and P. Podder. **Data analytics for novel coronavirus disease**, *Informatics in Medicine Unlocked*, vol. 20, 100374, 2020.

17. R. S. Yadav. **Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India**, *International Journal of Information Technology*, vol. 12, pp. 1321–1330, 2020.

18. H. Singh, and S. Bawa. **Predicting COVID-19 statistics using machine learning regression model: Li-MuLi-Poly**, *Multimedia Systems*, pp. 1-8, 2021.https://doi.org/10.1007/s00530-021-00798-2.

19. N. Druss. **Covid-19 Data Analysis Using Linear Regression**, Available at: https://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1091&context=cday, accessed June 2021.

20. M. Batista. **Estimation of the final size of the coronavirus epidemic by the logistic model**, *medRxiv*, 2020. Available at: https://doi.org/10.1101/2020.02.16.20023606.

21. I. J. Eyoh, E. N. Udo, I. J. Umoeka, and J. E. Eyoh. **Prediction of COVID-19 time series – case studies of South Africa and Egypt using Interval Type-2 Fuzzy Logic system**, *Internation Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 2, pp. 627-635, 2021.

22. CSSEGISandData/COVID-19, Available at: https://github.com/CSSEGISandData/COVID-19, accessed July 2021.

23. Asia: Health Care Index by Country 2021 Mid-Year, Avaiable at: https://www.numbeo.com/health-care/rankings_by_country.jsp?title=2021-mid&region=142, accessed June 2021.

24. M. O'Driscoll, G. R. D. Santos, L. Wang, D. A. T. Cummings, A. S. Azman, J. Paireau, A. Fontanet, S. Cauchemez, and H. Salje. **Age-specific mortality and immunity patterns of SARS-CoV-2**, *Nature*, vol. 590, pp. 140-145, 2021.

25. SAARC - South Asian Association for Regional Cooperation, Available at: https://countryeconomy.com/countries/groups/southasian-association-regional-cooperation, accessed June 2021.

26. E. Mathieu, H. Ritchie, E. Ortiz-Ospina, M. Roser, J. Hasell, C. Appel, C. Giattino, and L. Rodés-Guirao. **A global database of COVID-19 vaccinations**, *Nature Human Behaviour*. 2021. DOI: https://doi.org/10.1038/s41562-021-01122-8.

27. Population ages 65 and above, total, Available at: https://data.worldbank.org/indicator/SP.POP.65UP.TO?end=2019&start=2019&view=bar, accessed June 2021.

28. A. Jain. Understanding Regression using COVID-19 Dataset - Detailed Analysis, Available at: https://towardsdatascience.com/understanding-regression-using-covid-19-dataset-detailed-analysis-be7e319e3a50, accessed June 2021.