ABSTRACT

Background The high prevalence of COVID-19 has made it a new pandemic. Predicting both its prevalence and incidence throughout the world is crucial to help health professionals make key decisions. In this study, we aim to predict the incidence of COVID-19 within a two-week period to better manage the disease. Methods The COVID-19 datasets provided by Johns Hopkins University, contain information on COVID-19 cases in different geographic regions since January 22, 2020 and are updated daily. Data from 252 such regions were analyzed as of March 29, 2020, with 17,136 records and 4 variables, namely latitude, longitude, date, and records. In order to design the incidence pattern for each geographic region, the information was utilized on the region and its neighboring areas gathered 2 weeks prior to the designing. Then, a model was developed to predict the incidence rate for the coming 2 weeks via a Least-Square Boosting Classification algorithm. Results The model was presented for three groups based on the incidence rate: less than 200, between 200 and 1000, and above 1000. The mean absolute error of model evaluation were 4.71, 8.54, and 6.13%, respectively. Also, comparing the forecast results with the actual values in the period in question showed that the proposed model predicted the number of globally confirmed cases of COVID-19 with a very high accuracy of 98.45%. Conclusion Using data from different geographical regions within a country and discovering the pattern of prevalence in a region and its neighboring areas, our boosting-based model was able to accurately predict the incidence of COVID-19 within a two-week period.

Table of Contents

List of Tables

List of Figures

1. INTRODUCTION

1.1 What Is Data Mining?

Data mining is the method of extracting valuable information from a large data set. In other words, it is the process of deduction to get relevant data from a vast database. We can use data mining in relational databases, data warehouses, & object-oriented databases.

1.2 Types Of Data Mining Models –

- Predictive Models
- Descriptive Models

1.3 What Is Predictive Model?

Predictive modeling is a commonly used statistical technique to predict future behavior. Predictive modeling solutions are a form of data-mining technology that works by analyzing historical and current data and generating a model to help predict future outcomes.

1.4 What Is Descriptive Model?

This technique is preferred to generate cross-tabulation, correlation, frequency, etc. These descriptive data mining techniques are used to obtain information on the data's regularity by using raw data as input and discovering important patterns.

2. STEPS INCLUDING IN PREDICTION OF COVID-19

- Step 1 Data Identification
- Step 2 Pre-processing the data
- Step 3 Implementing the selected data mining techniques
- Step 4 Evaluating and validating the performance of the models
- Step 5 Development

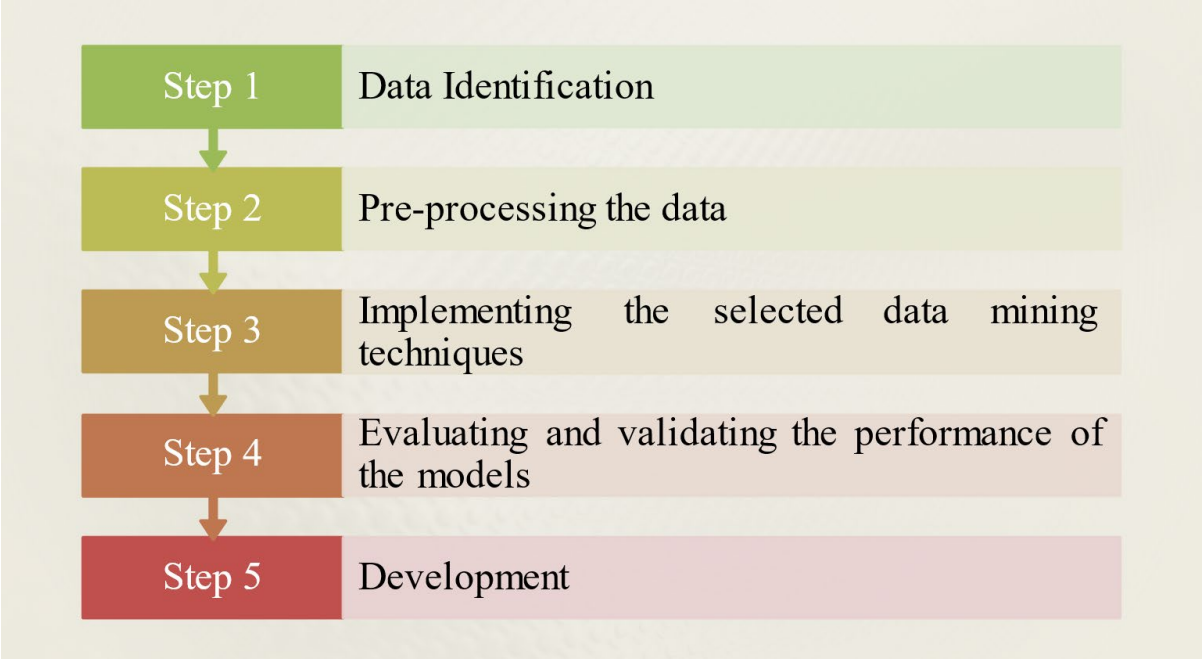

Fig.1 Steps Including In Prediction Of Covid-19

3. METHODS

The COVID-19 datasets provided by Johns Hopkins University, contain information on COVID-19 cases in different geographic regions since January 22, 2020 and are updated daily. Data from 252 such regions were analyzed as of March 29, 2020, with 17,136 records and 4 variables, namely latitude, longitude, date, and records. In order to design the incidence pattern for each geographic region, the information was utilized on the region and its neighboring areas gathered 2 weeks prior to the designing. Then, a model was developed to predict the incidence rate for the coming 2 weeks via a Least-Square Boosting Classification algorithm.

3.1 Dataset

COVID-19 epidemiological data have been compiled by the Johns Hopkins University Center for Systems Science and Engineering (JHU CCSE) [2]. The data have been provided in three separate datasets for confirmed, recovered, and death cases since January 22, 2020 and are updated daily. In each of these datasets, there is a record (row) for every geographic region. The variables in each dataset are province/state, country/region, latitude, longitude, and the incremental dates since January 22. For each region, the value for any date indicates the cumulative number of confirmed/recovered/death cases from January 22, 2020.

In this study, according to the input requirements of the proposed model, we changed the data representation so that instead of three separate datasets for three groups of confirmed, recovered, and death cases, only one dataset containing the information of all three groups was arranged. In this new dataset, each record (or row) of the dataset contains information about the number of confirmed, recovered, or deaths per day for each geographic region. As a result, the variables in this new dataset are: Province / State, Country / Region, Latitude (Lat), Longitude (Long), Date (specifying a certain date), Cases (indicating the number of confirmed, recovered, or death cases on the certain date), and Type (specifying the type of cases, i.e. confirmed, recovered, or death) as suggested by Rami Krispin [3].

In this study, the data were applied into the analysis by March 29, 2020, with 50,660 records and 7 variables. This period includes information about parts of winter and spring in the northern hemisphere and summer and autumn in the southern hemisphere. By March 29, the dataset consisted of cases from 177 countries and 252 different regions around the world. There were 720,139 confirmed, 33,925 death, and 149,082 recovered cases in the dataset.

3.2 Preprocessing Step

Pre-processing was carried out on the dataset before training the proposed model. Figure 2 shows the preprocessing steps. The dataset was first examined for noise, since the noise data were considered as having negative values in Cases variable. The dataset contained 42 negative values in this variable. After deleting these values, the number of records were reduced to 50,618.

Subsequently, the Date variable was written in numerical format and renamed into “Day” variable. To that effect, January 22, 2020 marked the beginning of the outbreak and the next days were calculated in terms of distance from the origin. As a result, January 22 and March 29 were considered as Day 1 and Day 68, respectively.

Since each region is uniquely identified by its latitude and longitude, the data for Province/State and Country/Region were excluded from the dataset. Moreover, as the study aimed at predicting the incidence in any geographical region, we considered only those records providing information on the confirmed cases (17,179 records), but not on the dead or the recovered. So, after preserving the records with “Confirmed” value in the Type variable, it was deleted from the dataset. In this study, the “Cases” is considered as the dependent variable.

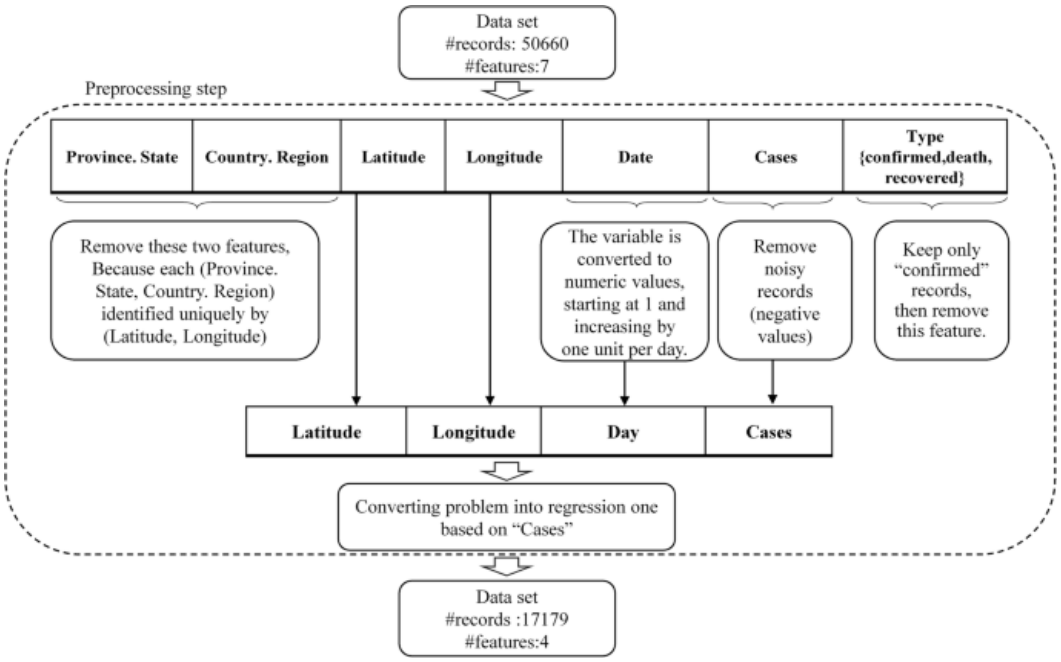

Fig.2 Preprocessing Step

3.3 Constructing The Prediction Model

An ensemble method of regression learners was utilized to predict the incidence of COVID-19 in different regions. The idea of ensemble learning is to build a prediction model by combining the strengths of a collection of simpler base models called weak learners. At every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners grown previously. One of the most commonly used loss functions is least-squares (LS) error [4].

In this study, the model employed a set of individual Least-squares boosting (LSBoost) learners trying to minimize the mean squared error (MSE). The output of the model in step m, Fm(x), was calculated using Eq. 1:

$$F_m(x) = F_{m-1}(x) + \rho_m \mathrm{h}(x; a_m) \text{-------Eq.1}$$

where x is input variable and h(x;a) is the parameterized function of x, characterized by parameters a [4]. The values of ρ and a were obtained from Eq. 2:

$$(\rho_m a_m) = argmin_{a,p} \sum_{i=1}^{N} [\tilde{y}_i - \rho \mathrm{h}(x_i; a)]^2 \text{ -------Eq.2}$$

Where N is the number of training data and y~i is the difference between the observed response and the aggregated prediction up to the previous step.

Due to the recent major changes in the incidence of COVID-19 worldwide over the past 2 weeks, we aimed to predict the number of new cases as an indicator of prevalence over the next 2 weeks. The structure of the proposed method is shown in Fig. 3.

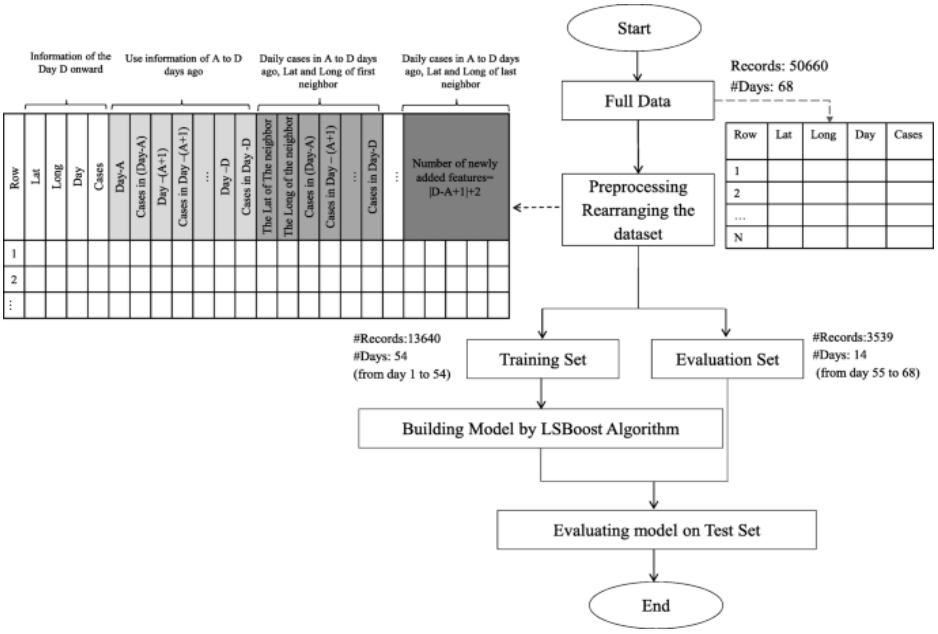

Fig.3 The Structure Of The Proposed Model

3.4 Evaluation The Actual Performance Of The Proposed Model

Given that the actual number of confirmed cases within March 30–April 12, 2020 period was available at the time of review, the performance of the proposed model was measured based on percent error between the predicted and the actual values. The percent error was calculated from Eq. 3:

$$\delta = \left(\frac{|v_A - v_E|}{v_A}\right) \times 100\text{-------Eq.3}$$

Where δ is percent error, v_A is the actual observed value and v_E is the expected (predicted) value. Furthermore, according to the predicted and actual confirmed cases in 252 geographical regions in the dataset, the continental incidence rate was calculated using Eq. 4:

$$\text{Continental incidence rate} = \left(\frac{I_C}{I_W}\right) \times 100\text{-------Eq.4}$$

where I_C is the incidence in each continent and I_W is the global incidence of COVID-19 from March 30 to April 12, 2020.

The experimentation platform is Intel® Core™ i7-8550U CPU @ 1.80GHz 1.99 GHz CPU and 12.0 GB of RAM running 64-bits OS of MS Windows. The pre-processing and model construction has been implemented in MATLAB.

4. RESULTS

4.1 Model Construction

The number of neighbors ranged from zero to 10. The value of 10 was obtained by trial and error. Euclidean distance based on latitude and longitude was used to calculate nearest neighbors. Given that the dataset contains data from January 22, 2020 to March 29, 2020 for the day we want to predict the incidence, the nearest and farthest days were selected as 14 and 54, respectively. Because the number of confirmed cases varies greatly from region to region, the proposed algorithm was implemented for 3 different groups of regions: for regions with less than 200 confirmed cases per day (16,825 records), those with 200 to 1000 cases per day (220 records), and those with over 1000 cases per day (152 records).

Table 1 shows the results of the best proposed model with regard to the different composition of the neighborhood and the days before. In order to predict the incidence of COVID-19 in regions with more than 1000 confirmed cases per day, the proposed model demonstrated the best performance with MAE of 6.13%, considering the information of the last 14 to 17 days of the region and its two neighboring areas. In the dataset, the number of cases records in these regions varied from 1019 to 19,821.

Maximum Number Of Confirmed Cases In A Day		Number Of Neighbours	Interval Of Days [Min, Max]	MSE		MAE	
				Value	Percent	Value	Percent
< 200	Train	–	[14,34]	1.86	0.005%	0.52	0.29%
	Test			407.47	1.04%	9.12	4.71%
[200,1000)	Train	9	[14, 20]	1.71	0.002%	0.62	0.07%
	Test			1.59e+ 04	1.87%	79.01	8.54%
≥1000	Train	2	[14, 17]	140.62	0.00003%	5.89	0.03%
	Test			7.14e+ 06	1.79%	1.2e+ 03	6.13%

Table 1 The Results Of The Best Models Evaluated On COVID-19 Dataset (January 22, 2020 To March 29, 2020)

4.2 Prediction Of Incidence By April 12, 2020

Figure 4 shows the prevalence of the COVID-19 from the first week to the tenth week in different regions, based on the information provided by the COVID-19 epidemiological dataset [2]. In this Figure, the diameter of the circles is proportional to the prevalence in those regions and the center of each circle matches the geographical coordinates of the region.

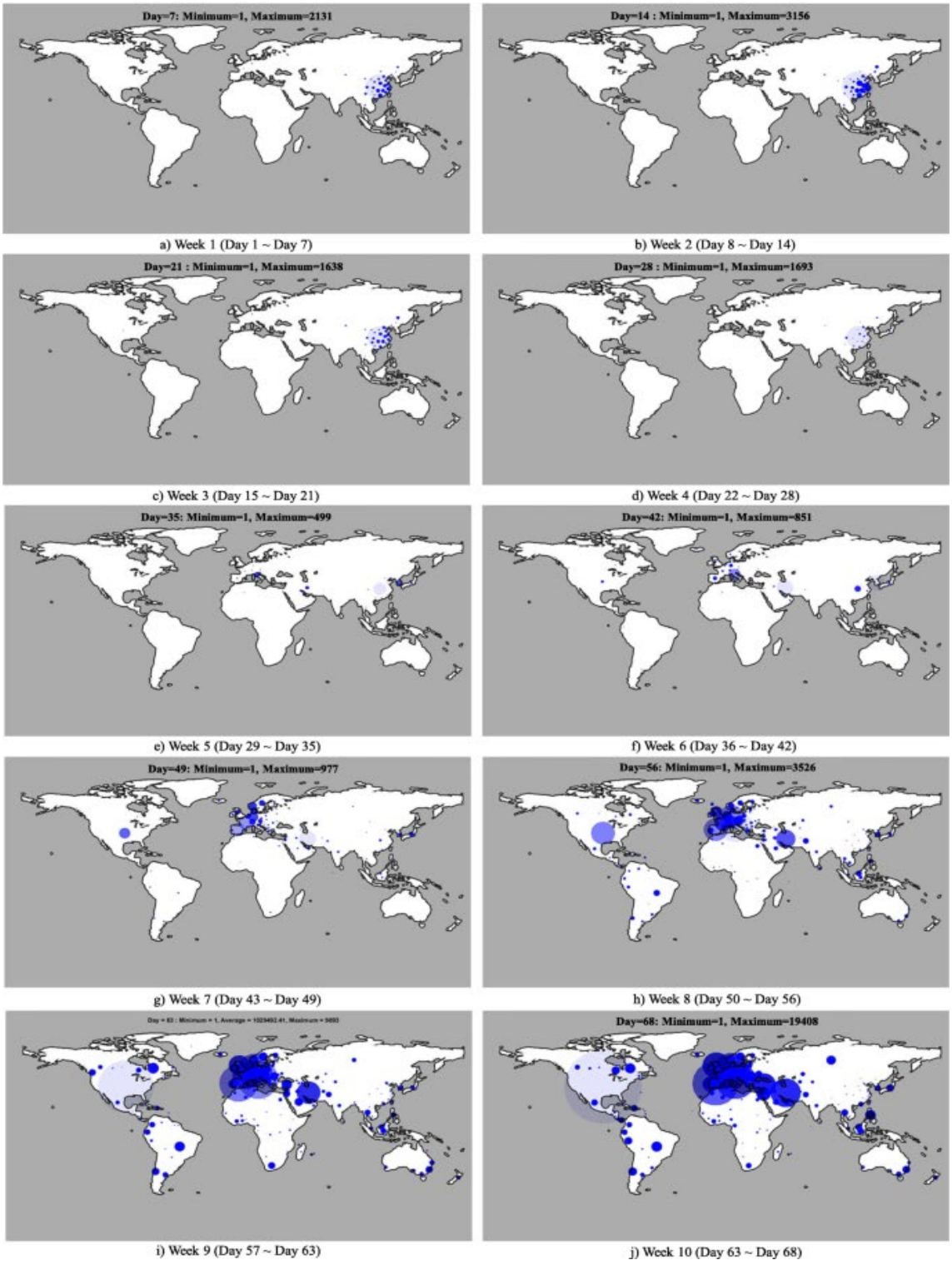

Fig.4 Visualize The Outbreak Over The Days (Created By Ourselves, Gimp Software, Open Source)

Table 2 shows the results of the forecast as to the number of new cases per day on different continents. According to the location of the continents in the northern and southern hemispheres, the period in question contains winter and early spring information in the continents of North America, Europe and almost entire parts of Asia. It includes summer and parts of autumn in Australian and approximately whole South America. Given that Africa lies in all four hemispheres, the data recorded for this continent in this period in the data set includes all seasons.

Date	***Continents***						***Total Number Of Confirmed Cases***
	Africa	**Asia**	**Australian**	**Europe**	**North America**	**South America**	
22 Jan ~ 29 Mar	4995	161,986	4522	385,097	150,877	11,740	719,217
30-Mar	635	7720	802	37,853	19,269	1906	68,185
31-Mar	820	7227	722	37,433	16,890	2000	65,092
1-Apr	472	7533	338	38,512	19,625	1508	67,988
2-Apr	1046	6438	981	44,047	18,435	1955	72,902
3-Apr	1047	6790	780	53,087	19,802	2359	83,865
4-Apr	1015	9739	872	51,954	19,302	2258	85,140
5-Apr	1014	10,563	1226	47,352	19,579	2490	82,224
6-Apr	1447	6867	1015	48,562	19,060	2530	79,481
7-Apr	1636	8027	1057	51,192	20,191	2768	84,871
8-Apr	2087	6786	1444	56,826	19,546	2550	89,239
9-Apr	2157	7749	1270	55,316	20,475	2685	89,652
10-Apr	1976	5818	1430	54,377	20,819	2573	86,993
11-Apr	1849	8962	1390	56,284	19,627	2351	90,463
12-Apr	1930	6781	1199	54,870	20,337	2806	87,923
Total	19,131	107,000	14,526	687,665	272,957	32,739	1,134,018
Prevalence Growth Rate	**283.00**	−33.94	221.23	78.57	80.91	178.87	57.67

Table 2 Forecast The Covid-19 New Cases For The Next 2 Weeks

4.3 Comparison Of Predicted And Actual Cases From March 30 To April 12, 2020

Table 3 shows the total number of daily cases in the 252 regions surveyed between March 30 and April 12, 2020. As shown, the daily percent error is below 20%. The best accuracy of the proposed model in predicting the incidence of COVID-19 was obtained on April 10 with 99.6%, and the worst on April 11 with 81.3%. Data analysis of the two-week continental incidence rates are also shown in Fig. 5. The best predicted continental incidence rates were found in South America and Asia with 18.15 and 21.04% percent error, respectively. The worst cases, still, were observed in Africa and Australian with more than 80% percent errors.

Date	Across All 252 Geographic Regions		Percent Error
	Predicted	Actual	
30-Mar	68,185	65,321	4.38%
31-Mar	65,092	76,799	15.24%
1-Apr	67,988	76,657	11.31%
2-Apr	72,902	81,340	10.37%
3-Apr	83,865	83,272	0.71%
4-Apr	85,140	80,392	5.91%
5-Apr	82,224	71,994	14.21%
6-Apr	79,481	73,285	8.45%
7-Apr	84,871	77,773	9.13%
8-Apr	89,239	84,275	5.89%
9-Apr	89,652	86,461	3.69%
10-Apr	86,993	87,520	0.60%
11-Apr	90,463	76,217	18.69%
12-Apr	87,923	95,353	7.79%
Total Number Of Confirmed Cases	1,134,018	1,116,659	1.55%

Table 3 Comparison Of Predicted And Actual Daily Incidence Of COVID-19

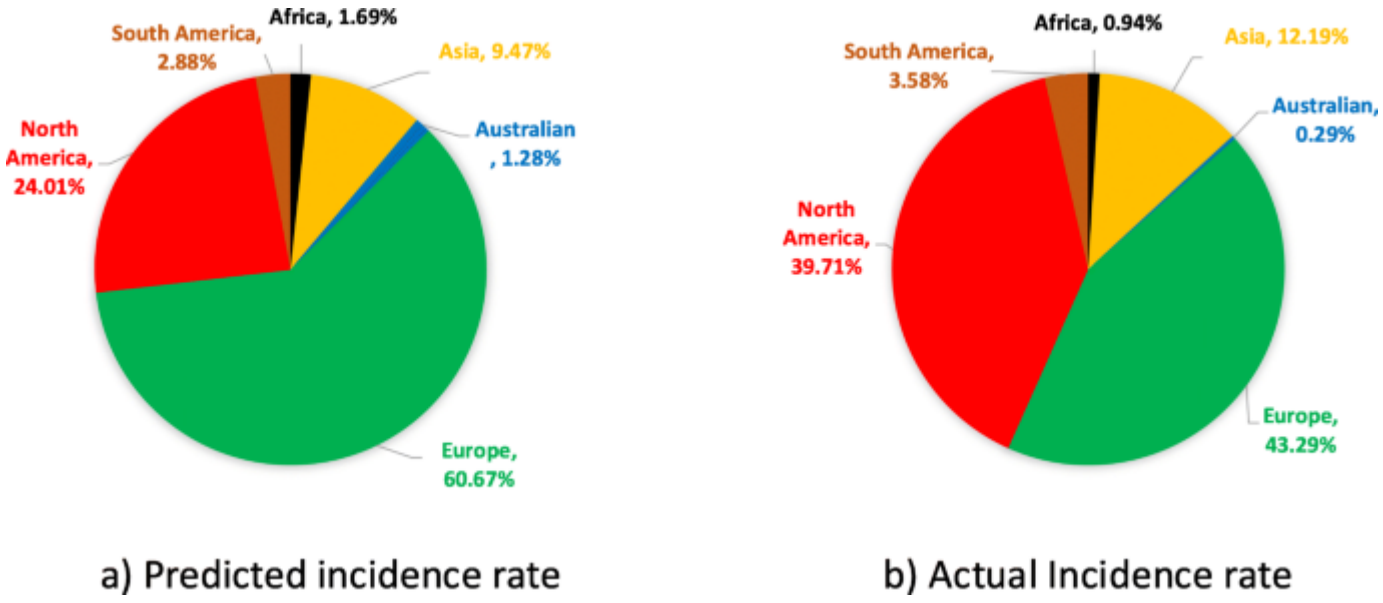

Fig.5 Comparison Of Predicted And Actual Continental Incidence Rates Between March 30 And April 12, 2020

5. AVAILABILITY OF DATA AND MATERIALS

The dataset analyzed during the current study is public and it is available in the [7],[8].

6. ADVANTAGES

- Fraud Detection

  We cannot overlook predictive analytics significance in recognizing abnormal patterns to avoid criminal behaviour. The rise in cybersecurity assists in uncovering cyber vulnerabilities and threats. Companies are peculiar about keeping their business safe from fraud and notorious threats, it helps them ensure a better procedure for smooth operational management.

- Optimizing Marketing Campaigns

  Market success is dependent on how well they know their customers and in order to get thorough know-how about shoppers out there, one really needs to observe customer responses and buying behaviour. Strategies are redefined based on customer likes and dislikes. Agencies use predictive analytics to categorize existing and potential customers. The secret lies in delivering the right message at the right time which ensures customer retention.

- Smart Decision Making

  Companies invest time and effort in making smart decisions that can change the course of business in all aspects. The art of decision-making encompasses all factors, one of which is analysis drawn from Ai-driven techniques. Machine learning models solely rely on data sets and they predict the outcome based on data fed into it. With advanced insights, firms make informed decisions for better growth and development. Businesses are harnessing the potential to acquire customer touchpoints and successfully utilize them for profitability.

- Operational Efficiency

  Drive operational excellence and boost your company’s growth rate with predictive analytics. Organizations are empowered by Ai-based solutions for charting a future map and revolutionizing uncertainty into a functioning procedure with high profitable prospects. Change is the only constant and how well a business keeps pace with changing demands and preferences of the customer base is the real game-changer. Actionable insights are provided by analytics for operational benefits.

7. CONCLUSION

Since epidemiological models such as SIR failed to accurately predict COVID-19 cases, as stated in [1, 5, 6], the current study relied on data from January 22 to March 29 provided by Johns Hopkins University and proposed a more complex model based on machine learning methods. The mean absolute error of the proposed model was 6.13% in predicting the incidence of COVID-19 in the two-week period of March 16–29 for regions with more than 1000 cases per day. The MAE was 8.45 and 4.71% for regions with a daily incidence rate between 200 and 1000 cases and less than 200 cases, respectively. An accuracy of more than 82% on the evaluation set confirms our perception that the pattern of incidence of a region is influenced by the pattern of disease in recent days in the same region and neighboring areas.

Last but not least, despite numerous limitations of the dataset, lack of knowledge about such an unknown disease and changes in disease control policies in different countries during the period under scrutiny, the proposed model proved effective in predicting the global incidence of COVID-19 in the two-week period of March 30 and April 12 with 98.45% accuracy. In addition, the accuracy of the proposed model in predicting daily cases in a worst-case scenario was 81.31%.

REFERENCES

1. Binti Hamzah FA, et al. CoronaTracker: world-wide COVID-19 outbreak data analysis and prediction. 2020.
2. (CCSE), J.H.U.C.f.S.S.a.E.J. Novel Coronavirus (COVID-19) Cases Data. 2020. Available from: https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases.
3. Krispin R. Coronavirus. 2020. Available from: https://github.com/RamiKrispin/coronavirus.
4. Friedman J. Greedy function approximation: a gradient boosting machine. Ann Stat. 2000;29:1189–232. https://doi.org/10.1214/aos/1013203451.
5. Postnikov EB. Estimation of COVID-19 dynamics “on a back-of-envelope”: Does the simplest SIR model provide quantitative parameters and predictions? Chaos, Solitons Fractals. 2020;135:109841. https://doi.org/10.1016/j.chaos.2020.109841.
6. Cooper I, Mondal A, Antonopoulos CG. A SIR model assumption for the spread of COVID-19 in different communities. Chaos, Solitons Fractals. 2020;139:110057.
7. [https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases]
8. [https://codeload.github.com/RamiKrispin/coronavirus-csv/zip/master]