

# **DATA MINING IN PREDICTION MODEL ON COVID-19**

- **Prepared By:**

- Hunaid Siamwala(12002040701067),Computer Engineering Department

# POINTS TO BE COVERED

- WHAT IS DATA MINING?
- TYPES OF DATA MINING MODELS?
- STEPS INCLUDING IN PREDICTION OF COVID-19?
- METHODS :
  1. DATASET
  2. PREPROCESSING STEP
  3. CONSTRUCTING THE PREDICTION MODEL
  4. EVALUATION THE ACTUAL PERFORMANCE OF THE PROPOSED MODEL
- RESULTS :
  - MODEL CONSTRUCTION
  - PREDICTION OF INCIDENCE BY APRIL 12, 2020
  - COMPARISON OF PREDICTED AND ACTUAL CASES FROM MARCH 30 TO APRIL 12, 2020
- CONCLUSION
- REFERENCES

# OBJECTIVE

- Data mining has been used in many industries to improve customer experience and satisfaction , and increase product safety and usability. Data mining in healthcare has proven effective in areas such as predictive medicine, customer relationship management, detection of fraud and abuse, management of healthcare especially in Covid-19 and measuring the effectiveness of certain treatments.
- Here is a short breakdown of two healthcare data mining applications with real-world examples of their use.
  - Measuring Treatment Effectiveness
  - Detecting Fraud and Abuse

# INTRODUCTION

WHAT IS DATA MINING?

---

# 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

# **TYPES OF DATA MINING MODELS –**

**PREDICTIVE  
MODELS**

**DESCRIPTIVE  
MODELS**

# 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

# WHAT IS DESCRIPTIVE MODEL?

---

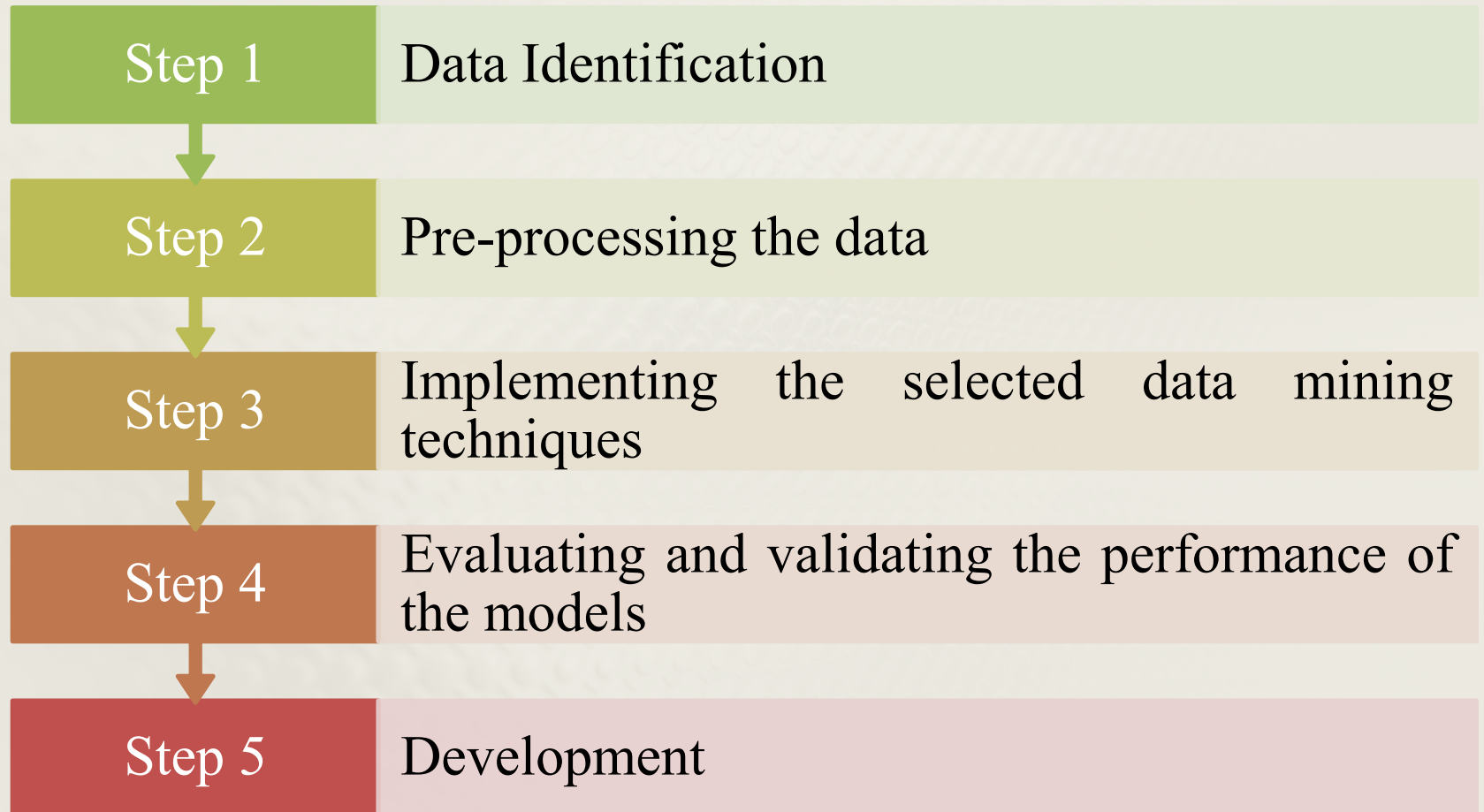
This technique is generally preferred to generate cross-tabulation, correlation, frequency, etc.

---

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



# STEPS INCLUDING IN PREDICTION OF COVID-19



# METHODS

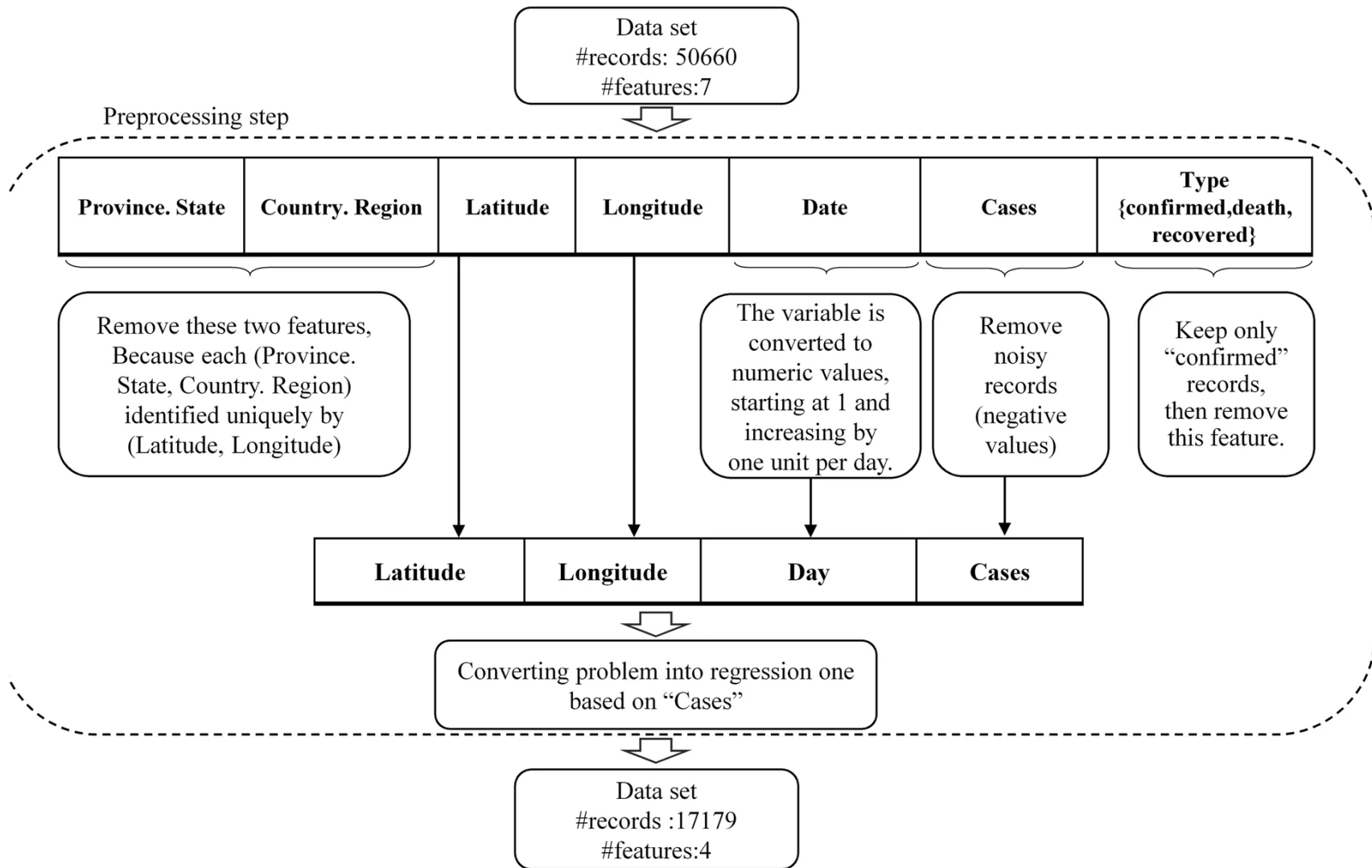
- The COVID-19 datasets provided by Johns Hopkins University, contain information on COVID-19 cases in different geographic regions since January 22, 2020 and are updated daily
- Data from 252 such regions were analyzed as of March 29, 2020, with 17,136 records and 4 variables, namely latitude, longitude, 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 which seek to improve the prediction power by training a sequence of weak models, each compensating the weaknesses of its predecessors

# 1. DATASET

- COVID-19 epidemiological data have been compiled by the Johns Hopkins University Center
- 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
- 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
- In this study, the data were applied into the analysis by March 29, 2020, with 50,660 records and 7 variables
- 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

## 2. PREPROCESSING STEP

- Pre-processing was carried out on the dataset before training the proposed model
- 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. 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.1 Preprocessing Step**

### 3. CONSTRUCTING THE PREDICTION MODEL

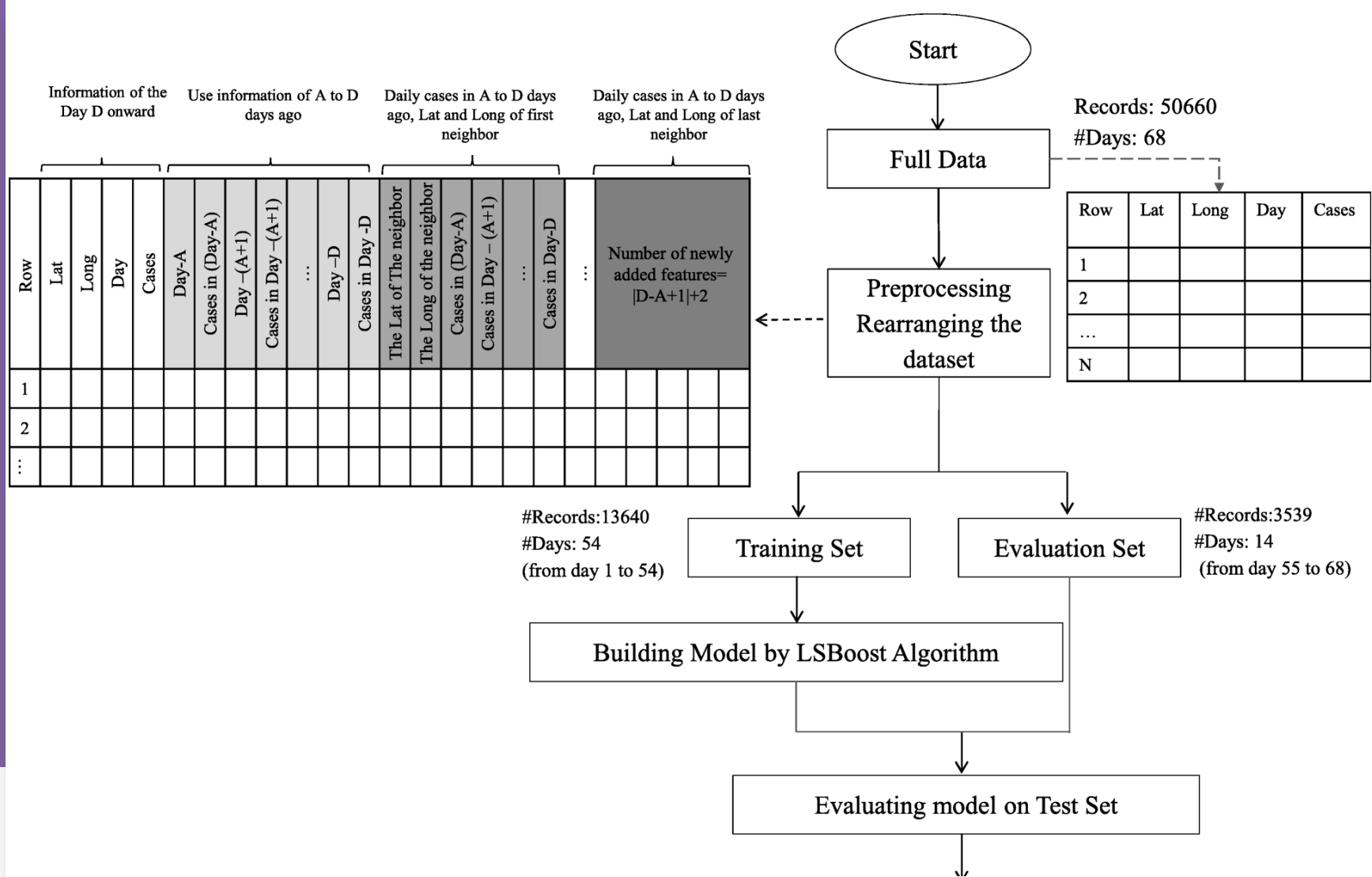
- An ensemble method of regression learners was utilized to predict the incidence of COVID-19 in different regions. The idea of ensemble learning is to build a prediction model by combining the strengths of a collection of simpler base models called weak learners
- At every step, the ensemble fits a new learner to the difference between the observed response and the aggregated prediction of all learners grown previously
- One of the most used loss functions is least-squares (LS) error
- 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  $m$ ,  $F_m(x)$ , was calculated using Eq. 1:

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

where  $x$  is input variable and  $h(x; a_m)$  is the parameterized function of  $x$ , characterized by parameters  $a$ . The values of  $\rho$  and  $a$  were obtained from Eq. 2:

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

- Where  $N$  is the number of training data and  $\tilde{y}_i$  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 the next figure



**Fig.2 The Structure Of The Proposed Model**



#### 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  $\delta$  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



# RESULTS

## 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).

---

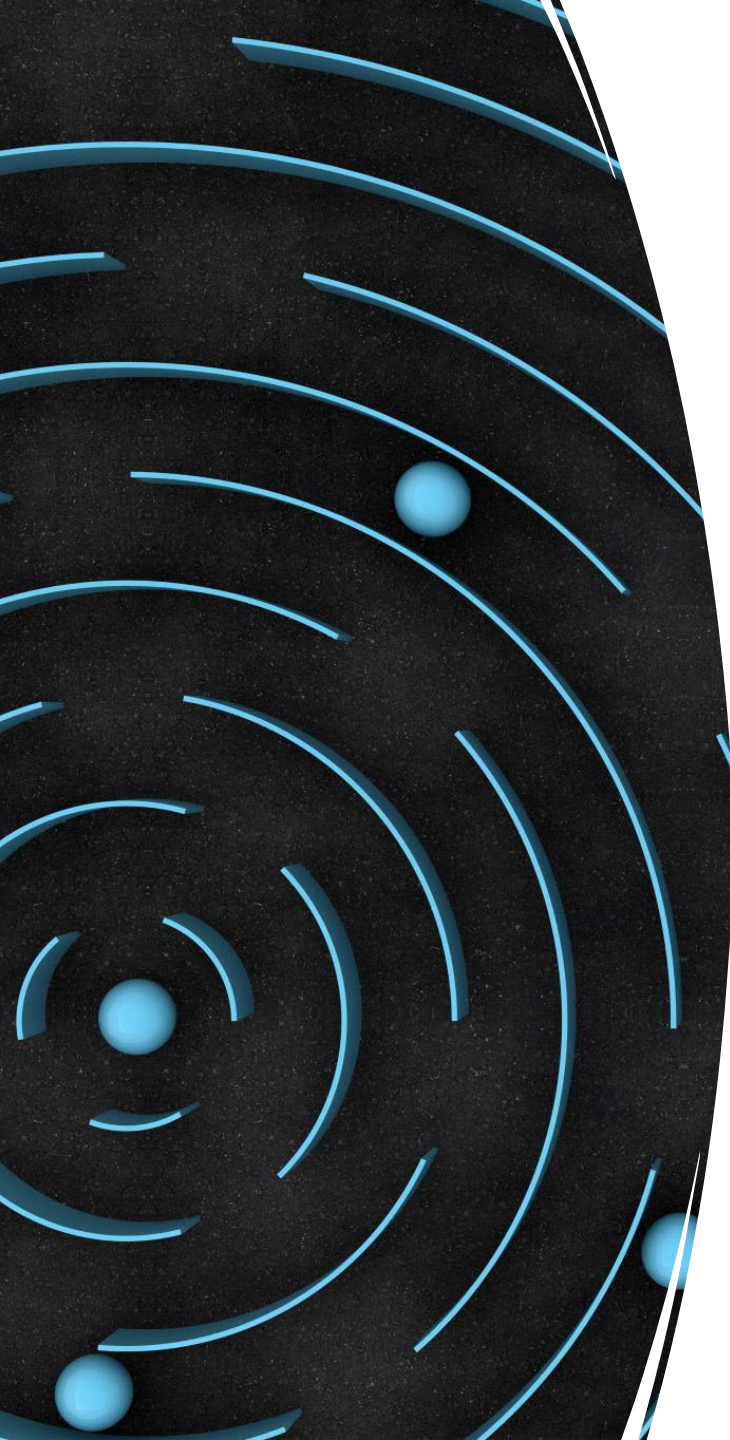
In order to predict the incidence of COVID-19 in regions with more than 1000 confirmed cases per day, the proposed model demonstrated the best performance with MAE of 6.13%, considering the information of the last 14 to 17 days of the region and its two neighboring areas. In the dataset, the number of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

---

For regions with 200 to 1000 cases per day, the proposed model performed best with respect to the 9 nearest neighboring areas and with data from the last 14 to 20 days, with MAE of 8.54% on the validation set. For regions with fewer than 200 cases per day, on the other hand, the proposed model performs best with MAE of 4.71%, considering the region data for the last 14 to 34 days.

Maximum number of confirmed cases in a day		Number of Neighbors	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%
$\geq 1000$	Train	2	[14, 17]	140.62	0.00003%	5.89	0.03%
	Test			7.14e+ 06	1.79%	1.2e+ 03	6.13%

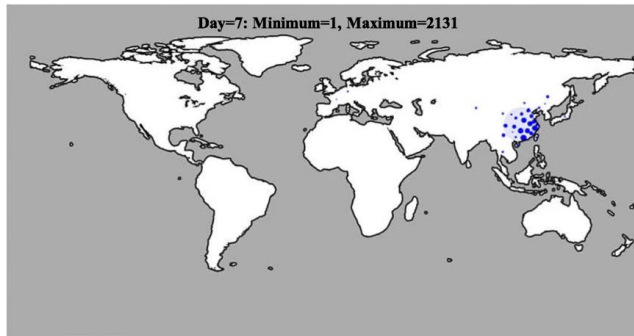
**Table 1 The Results Of The Best Models  
Evaluated On COVID-19 Dataset (January  
22, 2020 To March 29, 2020)**



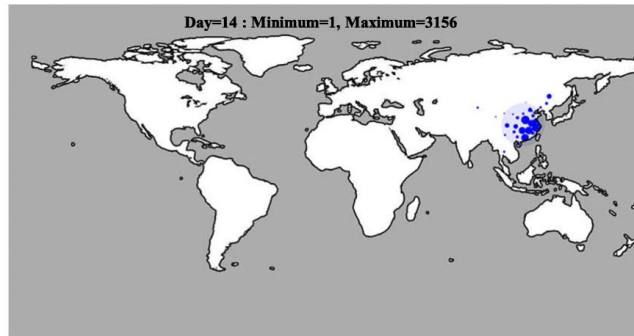
# PREDICTION OF INCIDENCE BY APRIL 12, 2020

---

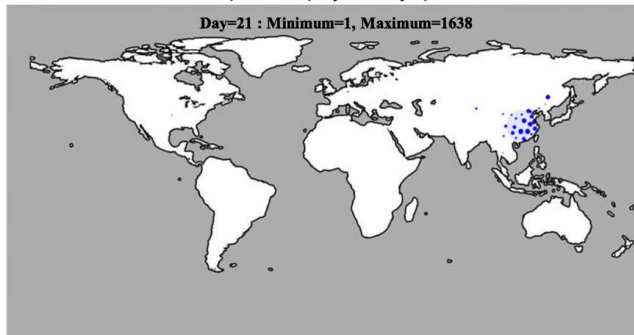
- Below Figure shows the prevalence of the COVID-19 from the first week to the tenth week in different regions, based on the information provided by the COVID-19 epidemiological dataset
- In this Figure, the diameter of the circles is proportional to the prevalence in those regions and the center of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



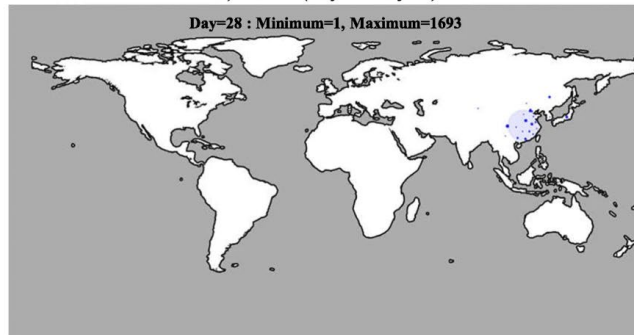
a) Week 1 (Day 1 ~ Day 7)



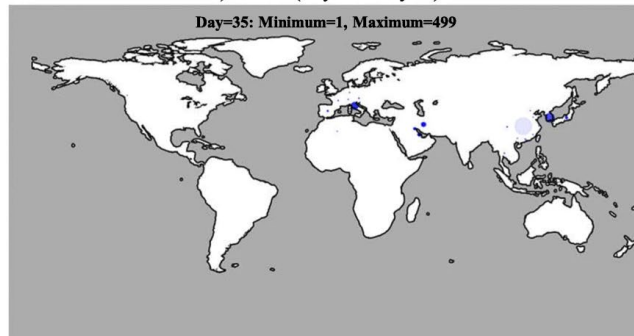
b) Week 2 (Day 8 ~ Day 14)



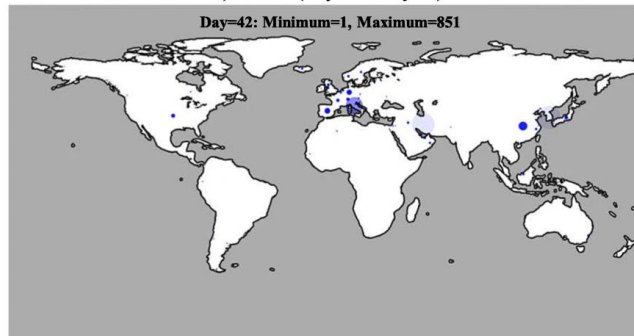
c) Week 3 (Day 15 ~ Day 21)



d) Week 4 (Day 22 ~ Day 28)

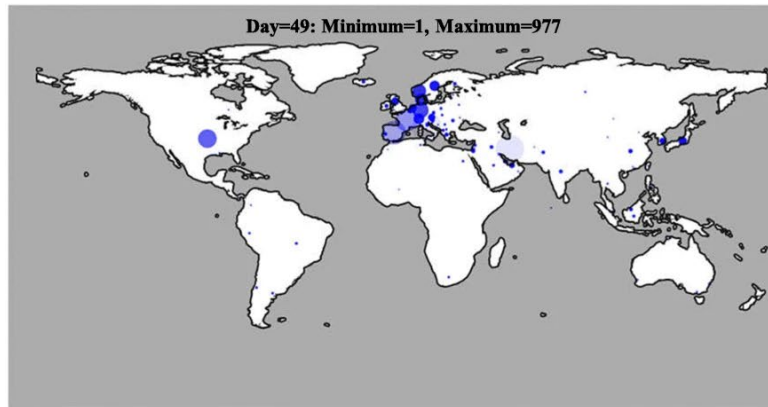


e) Week 5 (Day 29 ~ Day 35)

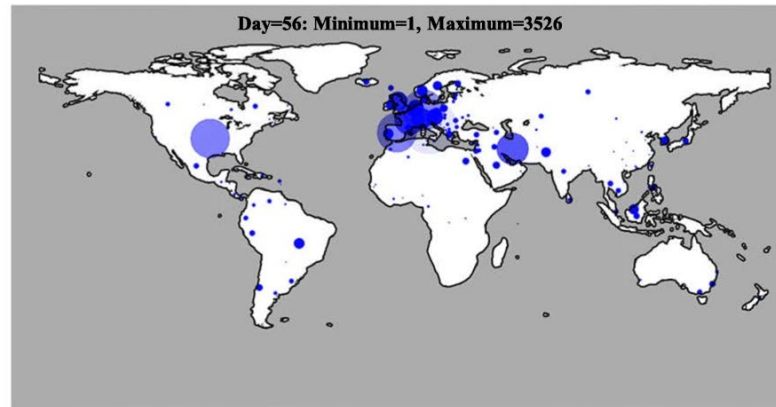


f) Week 6 (Day 36 ~ Day 42)

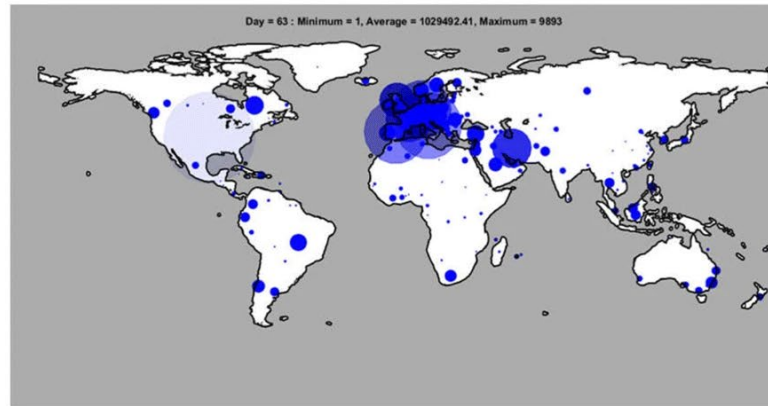
**Fig.3.1 Visualize The Outbreak Over The Days**



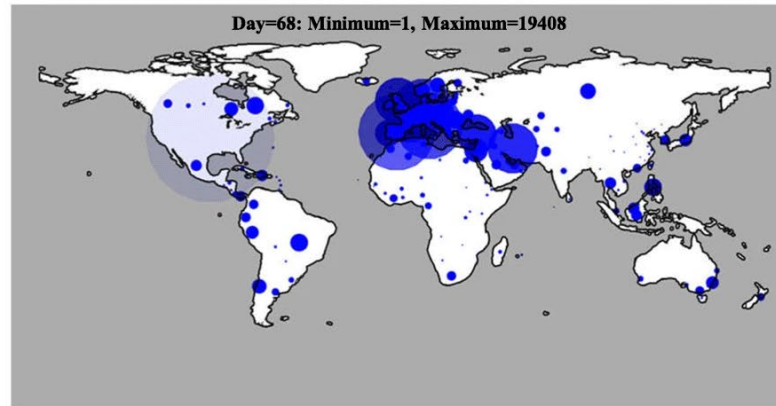
g) Week 7 (Day 43 ~ Day 49)



h) Week 8 (Day 50 ~ Day 56)



i) Week 9 (Day 57 ~ Day 63)



j) Week 10 (Day 63 ~ Day 68)

**Fig.3.2 Visualize The Outbreak Over The Days**



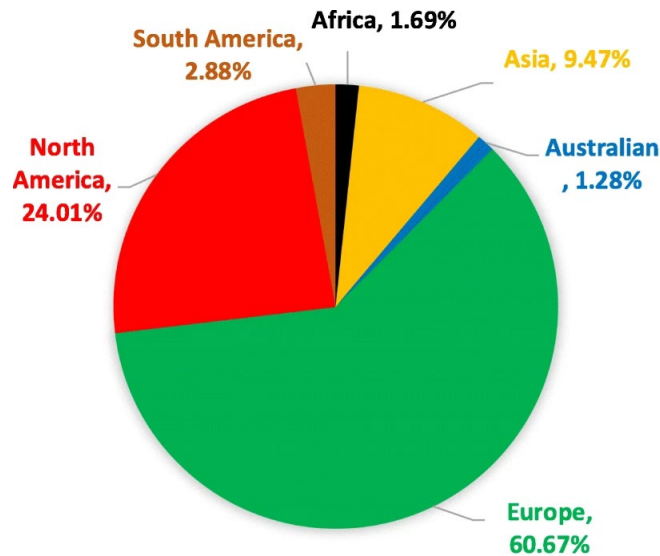
- Africa, Europe and South America had the highest rates of COVID-19 incidence, with 283, 221.23, and 178.87%, respectively
- Asia was the only continent that had slowed its growth with an incidence rate of – 34
- The prevalence rate of the disease is the proportion of the epidemiologic population with that disease at a point in time

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	78.57	221.23	80.91	178.87	57.67

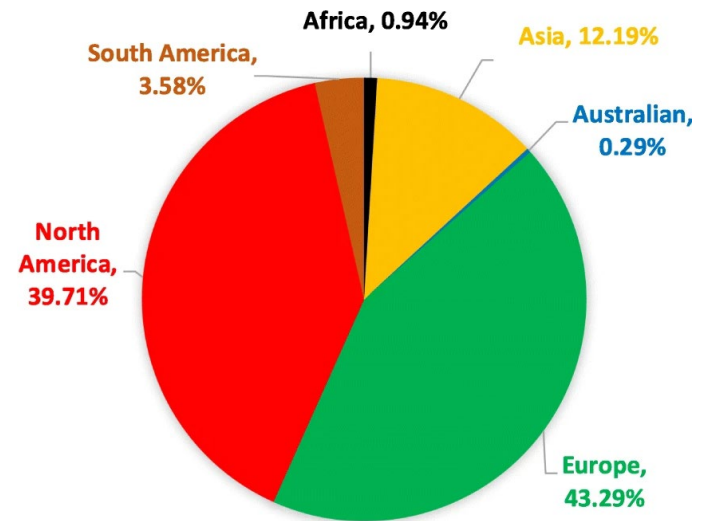
**Table 2 Shows The Results Of The Forecast As  
To The Number Of New Cases Per Day On Different Continents.**

## COMPARISON OF PREDICTED AND ACTUAL CASES FROM MARCH 30 TO APRIL 12, 2020

- ❖ 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



a) Predicted incidence rate



b) Actual Incidence rate

**Fig.4 Comparison Of Predicted And Actual Continental Incidence Rates Between March 30 And April 12, 2020**

# CONCLUSION

- ❑ 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. <https://www.usfhealthonline.com/resources/healthcare-analytics/data-mining-in-healthcare/>
2. <https://www.ijstr.org/final-print/oct2013/Data-Mining-Applications-In-Healthcare-Sector-A-Study.pdf>
3. [https://www.researchgate.net/publication/326022435\\_Data\\_Mining\\_Usage\\_and\\_Applications\\_in\\_Health\\_Services](https://www.researchgate.net/publication/326022435_Data_Mining_Usage_and_Applications_in_Health_Services)
4. <https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-11058-3>
5. <https://revolveai.com/predictive-analytics-advantages-and-disadvantages/>

# Thank you!

