GS International Conference on Computer Science and Engineering 2020 (GSICCSE), August 2, 2020

An intelligent analysis of COVID-19 patient's using Machine learning approach

Fatima Ijaz*[1]
Department of Computer Science
Riphah International University
Lahore,Pakistan
fatimaijaz084@gmail.com

Yasin Mubashar
Department of Computer Science
GC University
Faisalabad , Pakistan
Yaceenmubashar@gmail.com

Hira Naseer
Department of Computer Science
GC University
Faisalabad , Pakistan
hiramcs1210@gmail.com

Abstract

The virulent disease that is entirely over the world beating is coronavirus. All above the world patients are opposite unusual symptom. In this research, we analysis and predict the patient's death, confirmed and recovery rate by using the machine learning algorithms. Logistic regression algorithm analysis of the given dataset and find the root mean square error is 0.04 and mean absolute error is 0.01. The COVID-19 time series dataset is arranged by researchers from several health information of real time case in different countries that main focus is confirm, death and recovery patients analysis specific period of time. In this research, I will conduct Deep analysis, with different machine learning algorithms. The COVID-19 time series dataset given form the kaggle, Use rapid miner analysis tool for analysis and accuracy. Alongside patients who died the majority of the living in china, Italy, Hong Kong, Iran, India and America.

Keywords: COVID-19, Meta additive regression, linear regression, SMOreg, k means clustering, x means clustering.

1. INTRODUCTION

Corona virus is a worldwide outbreak in access of normal a rapid outbreak for which there is no natural immunity or immediately available treatments. We can say that it is a pandemic disease which further affected throughout the world. There is no treatment or vaccine found to cure or prohibit this disease. This disease secondary preventions being applied in which disease prevention strategy that focus on detection, diagnosis and treatment early in the disease process, to minimize the adverse and disabling effects. Corona is an infectious disease because it can be transmitted to other person from one person through contact. It is an infirmity appropriate to precise communicable manager or its poisonous harvest competent of creature in a straight line or circuitously transmit from persons to those from surroundings to human being.

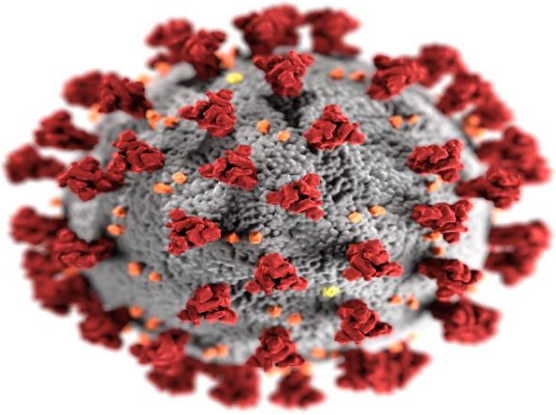

Figure 1: corona virus cell

Corona virus attacks the healthy cells rapidly from patients to other. It is very small in size and cannot be seen by naked eye. Corona virus hijacks our cells. It enters the main part throughout the nose, mouth or eyes then attach to cell in the airway that manufacture a protein called ACE2. The disease is supposed to have originated in bats someplace it possibly will have emotionally involved to a comparable protein.

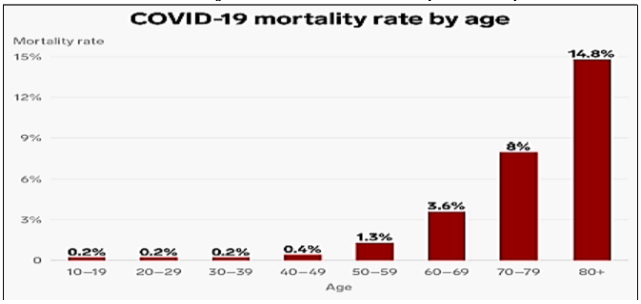

Figure2: mortality rate by age

According to Chinese center for disease control and preservation the mortality rate by age in corona patients. They prove that corona attacks to olds and child's as compare to younger's. Due to reason, the immunity system is strong in younger's as compare to olds and child's.

In this research I have discuss the new pandemic virus that is appeared in china city Wuhan in 2019 and spread across all over the world rapidly, lots of people are affected. The mainly frequent symptoms are passion, dry cough and fatigue and the serious symptoms are difficulty breathing or squatness of inhalation, chest twinge or pressure and defeat of vocalizations or association. Most infected people will develop mild to moderate illness and recover without hospitalization. I have given the one month day from COVID-19 master data set and perform analysis of confirmed, recovery and death rates.

The construction starts with the part1 with introduction of the content delivery system as Description. Segment 2 I will discuss the problems statements. Part 3 addresses previous researcher's latest novel with the descriptions of optimization techniques in use by different writers. Segment 4 outlines the implementation of the proposed plan of suggestions. The development of the current scheme is shown in Part 5, addresses system operation and outcomes with either the aid of the program's screenshot.

2. Problem statement

It is necessary to know the exact rate of confirmed, death and recovery cases of corona. So I take this step to analysis the corona disease. This virus spreads day by day and corona cases increased. Modification occurs in this data as time spends. So there is no limit of data.

3. Literature review

[1]In this research, numerous categorization models were tasted to construct the most excellent utilize of the clinical information provided online to be capable to forecast the gravity of the corona cases. SVM on 15 attributes of symptoms and supplementary pertinent in sequence of patient achieved smallest amount classification mistake of 0.21 which proves the probability of the projected method. We also proved that symptoms isolated cannot assist in deciding the gravity of the gear.

[2] in this work, Both models sampled exceeding formed comparatively comparable outcome by means of a small number of variations which might be based on statistical technique estimations in do extremely well model, the relevance on top of can be used to approximation rates for several given region with a the minority inputs and subsequent the excel model process you can produce similar outcome. The distinction flanked by the two models is ease of use, the submission can with no trouble be used by someone, even with fewer computational skills contrasting the excel model which may necessitate preceding specialist information with Microsoft excel, consequences predicted in cooperation models might be precise will less significant populations like for one city but can be in correct when a large illustration opening is taken outstanding to inconsistency in dynamics.

[3] In this study, the current expressive, tentative psychoanalysis of the initially 72.314 suitcases of COVID-19 report throughout February 11, 2020 offers significant original in sequence on the way to the worldwide collaboration happening the contagion appearing in china. In exacting this examination archives the tremendously prompt increase of the new coronavirus regardless of tremendous labors to hold it. The COVID-19 extend superficial Hubei region sometimes later than December 2019 and by February 11, 2020, 1.386 countries transversely the entire 31 provinces be pretentious. A whole of the 1.716 health staff have become contaminated and 5 have died (0.3%)

4. Methodology

4.1 Process flow diagram

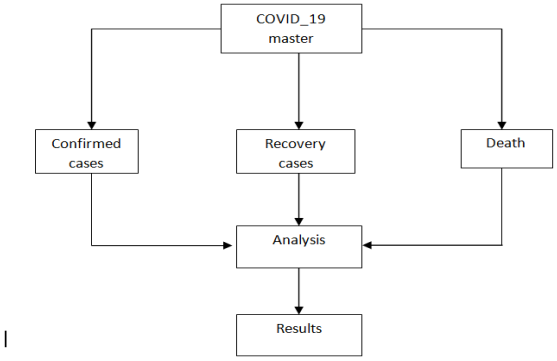

Figure 3: process flow diagram

The COVID_19 master dataset download from the kaggle. It has three data files that are confirmed, recovery and death cases in corona diseases. The dataset uploaded on the rapid miner tool for analysis and show the result arrive the variety of graphs and charts.

4.2 Dataset analysis and attributes

The COVID_19_master facts set download from the kaggle and select the archived_data files. The achieved time series contains three files that are time_series_2019_ncov_confirmed, time_series_2019_ncov_death and time_series_2019_ncov_recovery. The attributes these files are province/state, country/region, lat and long. I give the data of confirmed, death and recovery cases of different countries during 2/1/2020 to 3/9/2020.

4.3 Confirmed cases analysis

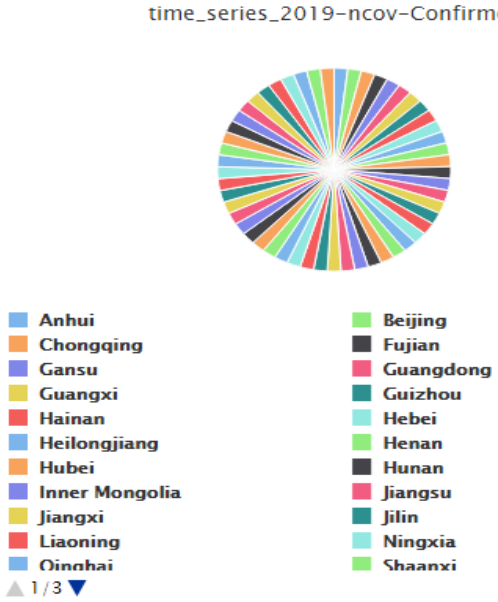

Figure 4: confirmed cases

The above figure shows the confirmed cases analysis of corona virus patients in different countries during the 2/1/2020 to 3/9/2020. The confirmed cases analysis conducts through the count (province/state) attribute. For this purpose 53 province states are selected and every state result shows the different colors.

4.4 Clustering

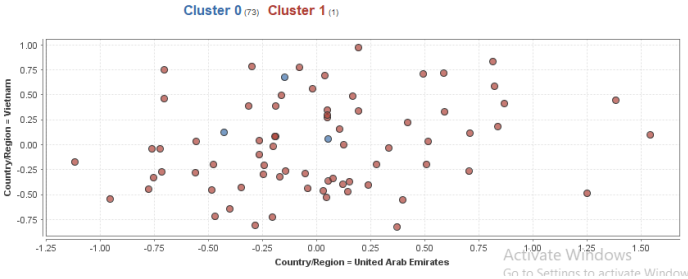

Figure 5: clustering of confirmed cases

The above figure shows the clustering of confirmed case analysis. K means and x means clustering is used to make clusters. It makes two clusters 0 and 1; the cluster 0 has 73 instances and cluster 1 and 1 instance. Theses clusters instances shows the different colored dots. These clustering perform on country region attribute.

4.5 Recovery cases analysis

4.5.1 Data

Figure 6: recovery cases data

The data given in excel form and then upload on the Rapid miner analysis tool. The following picture shows the recovery cases data that shows in dataset. The attributes are country/ region, state, lat, long and date during 2/1/2020 to 3/9/2020.

4.5.2 Statistics analysis of recovery cases

Statistics analysis performs on attributes and finds the distribution, maximum, minimum, average and standard derivation values.

Lat	Name	Values
	Maximum	61.924
	Average	30.267
	Standard deviation	19.531
Long		
	Maximum	153.025
	Average	59.999
	Standard deviation	87.749

Table 1: statistical analysis of recovery cases

The above table shows the statistical analysis of recovery cases, and the find the maximum, average, minimum and standard deviation values. Use the Lat and Long attributes and find the values. The late attribute average is 30.2 and standard deviation is 19.5. The long attribute average is 59.9 and the standard deviation is 87.7.

4.5.3 Correlation coefficient

Correlation coefficient is used to measure the strength of the relationship between two variables. A correlation coefficient of 1 means that for every positive increase of a fixed proportion in the other.

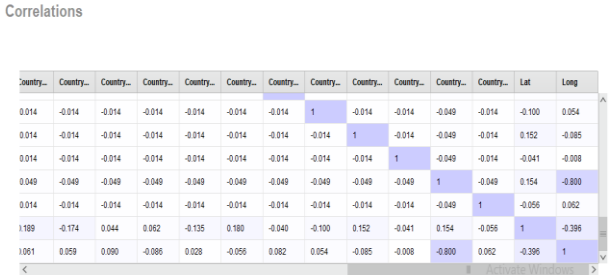

Figure 7: correlation coefficient of recovery

The above figure shows the correlation coefficient of recovery cases and finds the correlation coefficient 1 that is positive (increased) association instances.

4.5.4 Clustering

Clustering is apprehensive by means of confederacy matter collectively to facilitate be corresponding on the way to every additional and disparate in the direction of the items belong to additional clusters. K means uses kernel to approximation the detachment connecting substance and clusters. K means is a restricted clustering algorithm.

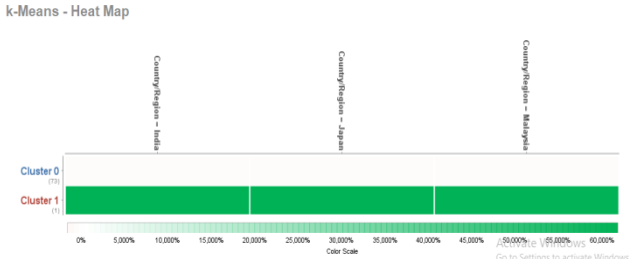

Figure 8: k means clustering of recovery

The above figure show clustering through heat map and make two clusters 0 and 1 for this purpose it select three countries India, Japan and Malaysia.

It makes two clusters 0 and 1. Cluster 0 has 73 instances and cluster 1 has 1 instances. It shows the recovery cases analysis during 2/1/2020 to 3/9/2020.

4.5.5 Recovery visualization

The recovery ceases visualization through a graph during the specific time period that 2/1/2020 to 3/9/2020 for these purpose 29 countries is selected.

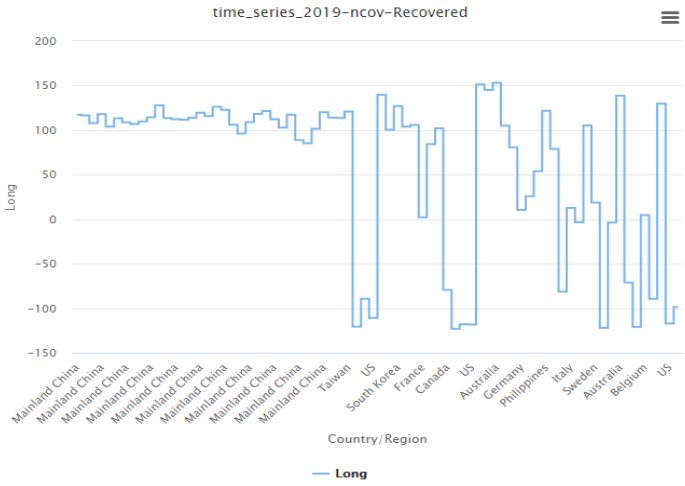

Figure 9: recovery cases visualization

The above figure shows the recovery cases analysis in COVID_19 master dataset. For this purpose many countries, states/region are used like US, mainland china, India Japan, Malaysia etc.

4.6 Death cases analysis

For analysis of death cases used the time series COVID_19_ncov death file from the COVID_19 master dataset. For this purpose 29 countries are selected and show the result through pie chart in different colors.

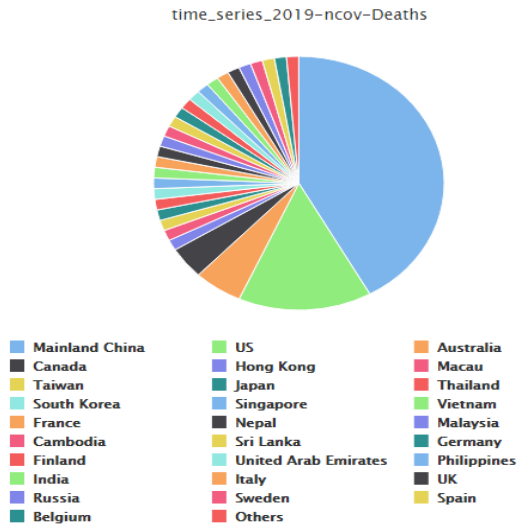

Figure 10: death cases analysis

The above figure shows the death rate in 29 countries from corona virus. The death rate Mainland china is 31%, US 11% and Australia 4%.

4.6.1 Clustering

Clustering is unsupervised machine learning technique involve the combination of statistics point. Particular a situate of data points, we be able to use a clustering algorithm to categorize apiece data spot into a precise cluster. Clustering is concerned by means of federation substance together that are comparable to each one other and disparate to the matter belonging to further clusters. Essential part k means uses kernels to approximation dissociate between substances and clusters.

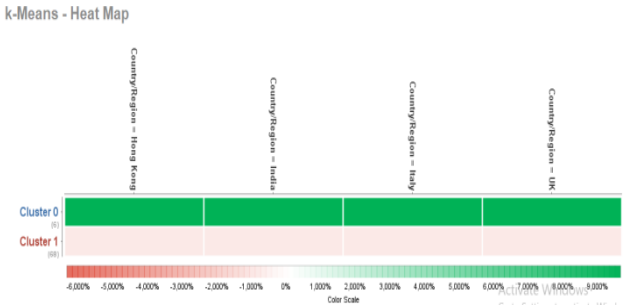

Figure 11: k means clustering

The above figure shows the k means clustering and x means clustering. K means shows the amount of clusters and x means shows the minimal and maximal number of clusters. In this image the country/region attribute is used to make clusters and four countries are selected for k means clustering that are Hong Kong, India, Italy and UK.

4.6.2 Death cases analysis through wordcloud

The wordcloud is an illustration unruffled of expressions used in a exacting manuscript or subject matter, in which the dimension of every word indicate its occurrence or significance. So the supplementary often a precise word appears in your manuscript, the superior and bolder it appears in your word cloud. The wordcloud shows the high frequency words very large.

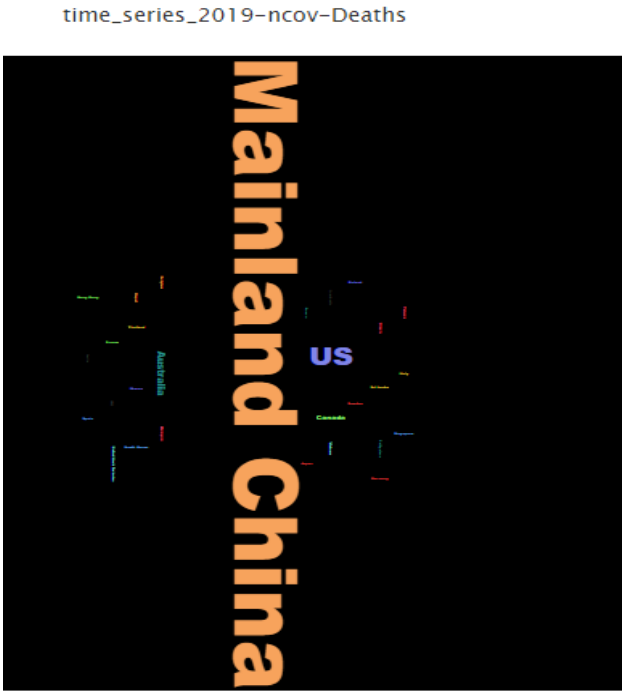

Figure 12: death cases analysis through wordcloud

In the above figure the death rate is show in wordcloud. The death rate due to corona virus is highest in mainland china that is 31%, USA 11% and Australia 4% till that 2/1/2020 to 3/9/2020.

5. Result and discussion

5.1 Experiment results

Three algorithms that are linear regression, Meta additive regression and SMOreg apply on COVID_19 master dataset and find correlation coefficient, mean absolute error, root mean squared error, relative absolute error, root relative squared error and total number of instances.

Algorithm	Correlation coefficient	Mean absolute error	Root mean square error	Relative absolute error	Root relative squared error	Total number of instances
Meta additive regression	0.9	6.8	11.3	4.8	2.4	74
Linear regression	1	0.01	0.04	0	0	74
SMOreg	1	4.0	4.3	2.8	0.9	74

Table 2: experimental results

The above table shows the result of the COVID_19 master dataset. In linear regression the time taken to build model is 0.04 seconds. Estimation arranged train situate. The period in use to exploration prototypical proceeding preparation records: 0.01 seconds.

In SMOreg the instance in use to construct replica is 0.02 seconds. Evaluation on training set. The number of kernel evaluation is: 2775. The cached is 86.281%. In Meta additive regression the time taken to build model is 0.second. The instant in use to test reproduction scheduled preparation facts: 0.07 seconds.

In function linear regression time taken to build model is 0.13 seconds. The evaluation on training set. And the phase in use to analysis prototype lying on training data is 0.01 seconds.

Conclusion

The world health organization constructs a dataset that named COVID-19 master (novel corona virus (2019-ncov)) they upload daily confirmed, recovery and death rate all

over the world due to corona virus. I took this dataset from kaggle between the times period 2/1/2020 to 3/9/2020. And I uploaded it rapid Miner for analysis. I used machine learning algorithms that are linear regression, Meta additive regression, SMOreg and clustering. The mean absolute error is 0.01 and root mean squared error is 0.04 that is less than previous researches it proves that it is more accurate than previous researches. For future succeed, if data set can be gather by researchers contain symptoms and conversant in sequence of suspect of corona to be to detect that innovative corona virus. Furthermore, data is altering and is additional each minute. As an outcome further report can be functional by our representation.

References

[1] A.Ella, A.Salama, and A. Darwish "Artificial Intelligence approach to predict the COVID-19 patient's recovery", EasyChair preprint, pp. 1-10, 2020.

[2] E.Dong, H.DU, L.Gardner, "An interative web based dashboard to track COVID-19 in real time", Lancet infect.Dis.3099,19-20(2020), pp. 1-2, 2020.

[3] A.Kyagulanyi, J.Tibabwetiza, D.Oscar, "risk analysis and prediction for COVID-19 demographics in low resources settings using a python desktop app and excel models", doi:https://doi.org/10.1101, pp.1-20, 2020.

[4] Z.Feng, Q. Li, Y.Zhang and Z.Wu, et al, "The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19) — China", Chinese Center for Disease Control and Prevention, pp.1-10, 2020

[5] Wang C, Hornby PW, Hayden FG, Gao GF, "A novel coronavirus outbreak of global health concern", http://dx.doi.org/10. 1016/S0140-6736(20)30185-9, 2020.

[6] Zhou, H.. "A clustering approach to free form surface reconstruction form multi-view range images" image and vision computing, 20090504

[7] Samrat K.Dey, Md. Mahbubur Rahman, Umme R. Siddiqi, Arpita Howlader. "Analyzing the epidemiological outbreak of COVID-19: A visual exploratory data analysis approach", joural of medical virology, 2020

[8] "Advances in computer science and ubiquitous computing", springer nature, 2018

[9] Conrad S. Tucker, Harrison M. Kim, Douglas E. Barker, Yuanhui Zhang."A relief attribute weighting and x-means clustering methodology for top dowm product family optimization", engineering optimization, 2010

[10] Coronavirus-awareness.weebly.com, rapidminer.com

[11] Hui DS, Azhar EI, Madani TA, Ntoumi F, Kock R, Dar O, "The continuing 2019-nCoV epidemic threat of novel corona viruses to global health - the latest 2019 novel coronavirus outbreak in Wuhan, China", Int J Infect Dis 2020;91(2020):264 – 6, 2020.