Prediction of Covid-19 Severity Level Using XGBoost Algorithm: a Machine Learning Approach Based on SIR Epidemical Model

Labeba Tahsin and Shaily Roy

July 16, 2021

Prediction Of Covid-19 Severity Level Using XGBoost Algorithm: A Machine Learning Approach Based On SIR Epidemical Model

Labeba Tahsin⋆[1] and Shaily Roy*[2]

[1] Computer Science and Engineering, Brac University, Bangladesh
labeba.tahsin@gmail.com
[2] Computer Science and Engineering, Brac University, Bangladesh
shaily.roy@bracu.ac.bd

Abstract. Covid-19 formally termed as "2019 novel coronavirus", is disrupting socio-economic conditions throughout the world. Due to the unavailability of efficient ways to predict the severity level of Covid-19, governmental officials and policymakers of different countries are facing difficulties to take precautionary measures for minimizing risks. This paper presents a model trained to predict Covid-19 situation severity level using XGBoost which is a gradient boosting algorithm. To categorize severity level, SIR epidemiological method has been employed which can express the current condition of any area affected by contagious diseases like Covid-19 analyzing the number of susceptible people which stands for S, the number of infected people stands for I, and the number of recovered people stands for R. By comparing the evaluation metrics of our model with other models based on different machine learning algorithms, it is deduced that the model performs better for less training time(speed), better accuracy rate, and has the ability to reduce over-fitting.

Keywords: Covid-19 · Machine Learning · XGBoost · SIR · Prediction.

1 Introduction

Covid-19 was first found in Wuhan, China and gradually it has spread all over the world. WHO termed it as a global pandemic on March 11, 2020 [1]. SARS-CoV-2 (Extreme Acute Respiratory Syndrome Corona Virus 2 virus) is responsible for Covid-19 disease. Aged people, patients with diabetes, obesity, and cardiovascular diseases are more likely to be affected by this virus [2]. As the Covid-19 pandemic is one of the most critical crises all over the world, without precautionary steps (like vaccination, lockdown, etc) it is not possible to reduce the death rate. This paper aims to provide the severity level of Covid-19 of a place so that necessary steps can be taken to reduce the severity. In this study, a model has been proposed which can predict the severity level of Covid-19 analyzing date, location, number of death, recovered and confirmed cases. For getting infection

⋆ Equal Contribution

rate and severity level, the SIR model has been used. SIR model is a basic transmission model designed for infectious diseases and using this epidemiological model, Covid-19 severity level types have been termed. The proposed model has been built using the XGBoost algorithm and its evaluation metrics are better compared to other models of different algorithms. The work will motivate other researchers for finding better solutions to forecast the severity level of pandemic situations.

2 Literature Review

According to ”WHO”, Covid-19 is an infectious disease caused by Corona Virus [3]. The word ”Covid” stands for 'corona(CO)', 'virus(VI)' and 'disease(D)'. Recent years, the disease is spreading epidemically all over the world causing huge economical and social disruptions. Several researches have been done on the spreading of Covid-19. To measure the mortality rate, G. Pinter et al. have worked on hybrid machine learning using the Covid-19 data from Hungary [4]. Those models predict that the statistics of spread and morality rate will decrease significantly. Though the machine learning approach doesn't assume the outbreak rate, it predicts the time statistics of the transmission rate along with the death rate. In order to evaluating, they have incorporated validation for nine days with a good model accuracy. F. Rustam et al. introduced forecasting mechanism of supervised machine learning to predict the future of Covid-19 [5]. They have basically tried to forecast the number of new infected patients, the number of deaths and also the number of recoveries per day. From their result analysis, it is deduced that among the four standard prediction models exponential smoothing (ES) performs the best and in contrast SVM performs the worst to find out the threatening factors of Covid-19 spread. For creating a national issue for many countries , Sina F. Ardabili et al. have proposed their work on Covid-19 outbreak using soft computing models [6]. They have chosen the soft computing models because epidemical models could not work better in long term prediction for the lack of accurate data with the restrictions provided by the authority. Their main task is to employ the potential machine learning models with a proper benchmarking. Moreover, Zoabi, Y et al. have focused on the Covid-19 diagnosis based on a machine learning model and they have gained an accuracy of 90%. Though the accuracy is pretty good, they have some flaws like their dataset have missing values for particular features, also some biasness which results in over-fitting later. They basically focus on the limitations of testing kits of Covid-19 tests, hence provide a solution to measure if a person need to test for Covid or not using machine learning models to limit the number of negative tests. However they scaled the dataset of Israeli Ministry of Health with a manual process to remove biasness, so the result they show might not be accurate for different areas or countries' cases. Yan L,et al. have shown the prediction or survivality of Covid patients using machine learning models [8]. They have colleced blood samples of 400+ infected patients in chine in order to finding out a common biomark for Covid-19 infection. From their research, it is decided that severe imbalance of lactic dehydrogenase (LDH) can be a major cause of patients' being in a serious condition due to Covid-19. This paper presets a quick process to find out the severity of an infected patient and suggest next step accordingly in advance which can help a patient to survive from the worst. The experiment is based on a small dataset with higher accuracy which may vary for different set of dataset and may result in a significant gap. In addition, a research is done on Covid-19 survival, death rate in India using machine learning models by V. K. Gupta et al. For finding out the regularity of the model's performance, they have applied k-fold cross validation for particular areas of India and among them random forest outperforms the all other methods. Our proposed work is inspired by the work of Iqbal, N.et al where they tried different type of machine learning models with prominent classifier to predict the situation of dengue outbreak [9].

3 Methodology

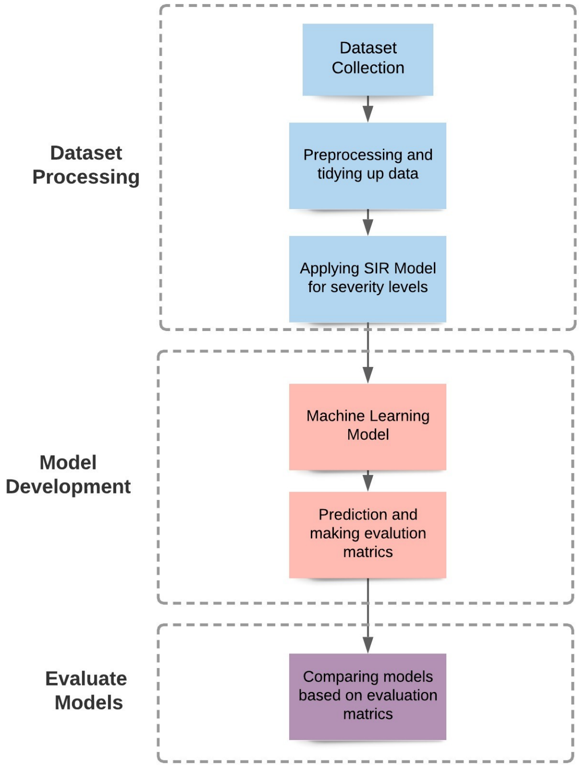
Dataset Processing
Dataset Collection
Preprocessing and tidying up data
Applying SIR Model for severity levels
Model Development
Machine Learning Model
Prediction and making evalution matrics
Evaluate Models
Comparing models based on evaluation matrics

Fig. 1: Proposed system model

3.1 System Model

According to the Fig 1 it is shown that the work is done with 3 steps : Data processing, Model Development and Model Evaluation.

Dataset and Pre-processing

Dataset: The dataset used in this paper has been collected from Kaggle and it can be accessed as “Covid-19 Coronavirus Dataset” [3]. The licence of the dataset is under the public domain. The author of the dataset has acknowledged John Hopkins CSSE as the source for live updates and data streaming on CoronaVirous.

Pre-processing: The dataset has some missing values. The rows having missing data have been filtered and separated from the main dataset. Because without the date, country and province the model can not be trained properly. The dataset also has some duplicate values in the same date, location and time of update. These duplications have been handled. The country and province columns have string type values and they have encoded with LabelEncoder. The dataset has the “date” column in Date format and year, month, day values are separated and turned into new columns. The processed dataset can be seen in Table 1 where Loc means location, Country means Country, Confm means Confirmed, Dt means Deaths, Rec means Recovered, Uh means Unhealed, InR means Infection-Rate, Y means Year, M means Months and D means Day.

Table 1: Preprocessed DataSet Sample

Serial	**Loc**	**Country**	**Confm**	**Dt**	**Rec**	**Uh**	**InR**	**Y**	**M**	**D**
0	8	5	1	0	0	1	1.000000	2020	1	22
1	15	5	14	0	0	14	1.000000	2020	1	22
2	35	5	6	0	0	6	1.000000	2020	1	22
3	74	5	1	0	0	1	1.000000	2020	1	22
...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...
3543	295	4	3	0	0	3	1.000000	2020	3	26
3544	296	5	178	2	172	4	-0.955056	2020	3	25
3545	297	5	1243	1	1222	20	-0.967820	2020	3	26

Model Development After processing and making the dataset tidy, SIR epidemiological model has been applied with a list of Machine Learning Models.

[3] The Dataset is downloaded from: https://www.kaggle.com/vignesh1694/Covid-19-coronavirus

SIR Epidemical Model: SIR is used to create new columns for categorizing severity levels of Covid-19 spread. To define the population difference using two parameters like β and γ. β means contact rate of N individual. γ represents the average healing rate. When $\frac{S}{N}$ is the probability of getting infected by the disease and $\frac{1}{\gamma}$ is the period of getting spread by the infected patients, the differential value of S, I and R can be calculated by-

$$\sqrt{(\frac{DS}{Dt})^2} = -\sqrt{(\frac{\beta.S.I}{N}))^2} \tag{1}$$

$$\sqrt{(\frac{DI}{Dt})^2} = \sqrt{(\frac{\beta.S.I}{N} - \gamma.I)^2} \tag{2}$$

$$\frac{DR}{Dt} = \gamma.I \tag{3}$$

The severity level has been named into the decision column and this column contains four types of value - Good, Hope, Danger, High Danger which are converted to numerical value like 0,1,2 and 3 later. Unhealed and infection rate columns were also generated by using the SIR model which can be seen from Table 2. The processed columns are compiled into a separated dataset which is used to train and test our model. As the decision column has String type values, LabelEncoder is used to encode the data.

Table 2: SIR epidemical Model

Ser.	**Country**	**Confm**	**Dt**	**Rec**	**Loc**	**Y**	**M**	**D**	**AC**	β	γ	δ	**InR**	**Decision**
0	5	1	0	0	8	2020	1	22	1	1	0	0	1	Hope
1	5	14	0	0	15	2020	1	22	14	1	0	0	1	Hope
3	5	1	0	0	74	2020	1	22	1	1	0	0	1	Hope
4	5	0	0	0	76	2020	1	22	0	0	0	0	0	Good
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
24532	27	17	0	0	276	2020	3	26	17	1	0	0	1	Hope
24535	4	3	0	0	295	2020	3	26	3	1	0	0	1	Hope
24536	5	178	2	172	296	2020	3	25	4	0.02	0.97	0.012	-0.96	Good

Decision Tree: For regression and classification decision tree is used to get a good result and for that entropy is calculate. Higher entropy means higher information contents [11]. If p_i is probability of class i and it has n classes the equation of entropy should be-

$$entropy = \sum_{i=1}^{n}(-p_i \log_2 P_i) \tag{4}$$

Random Forest: Random Forest is applied on the dataset following some steps. Firstly subset is generated bootstrapping the dataset of Covid-19. After that Gini Impurity for each features is calculated [12] . Let's say the database D has n classes and 2 splitted dataset on A d1 , d2, then gini impurity for D can be determined by

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2 \tag{5}$$

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2) \tag{6}$$

Based on the calculated gini impurity new sub trees are constructed from which bagging the ensemble, decision is taken based on the majority.

Logistic Regression: Logistic Regression can also be called as multi-linear-regression [13]. Firstly randomly co-efficient is determined for each row and analyzed the probability for sample instances of actual data from class 0 to class n. This process is applied repeatedly until getting a binary decision tree with proper prediction. The number of max iteration selected for this model is 300 as a parameter.

X Gradient Boost(XG Boost): The word XGBoost comes from eXtreme Gradient Boosting [14]. Scikit learn is used to implement XGBoos in this research with the parameter of ax_depth=1 and n_estimators=90. The library includes creating matrix, bagging and subsampling the data which results in better speed than any other algorithms in case of table value calculation.

Support Vector Machine(SVM): Support Vector Machine is one of the most popular classifier from machine learning algorithms. In order to applying it, firstly a line is created using epochs and spanning factors continuously unless it gives a correct classification [15]. It assures the decision boundary with largest margin from both classes.

4 Model evaluation

After tuning the parameters of all the models Decision tree, Random Forest, Logistic Regression , XGB and SVM, the study finds an accuracy of 95% from XGB. The performance of the models are determined by the metrices of Accuracy, Precision, F1-Score and Recall score . Additionally training time is also considered as a criteria of good performance where less time means the model is efficient and can be trained faster.

As it is shown in table 3 , given the F1 score and Recall score with accuracy , it can be clearly seen that the data was no where overfitted and still resulted in with good accuracy. The performance of Support Vector Machine(SVM) is very poor which gives an accuracy of 76% , on the other hand the performance of X gradient Boost(XGB) outstands with an accuracy of 95% and also very

Table 3: ML model performance result

Classifier	Accuracy	Precision	F1-Score	Recall	Time
DecisionTree	0.9169	0.86789	0.8226	0.7991	0.0028
RandomForest	0.9056	0.9513	0.7372	0.7214	0.2429
Logistic Regression	0.7647	0.9454	0.8981	0.8690	0.4350
XGB	0.9507	0.9454	0.8981	0.8690	0.3473
SVM	0.7619	0.7301	0.6545	0.6219	0.1275

impressive score in Precision, F1 , Recall and training time compared to other models applied.

The reason of XGBoost's impressive performance can be the ensamble methodology of having multiple hypothesis which creates multiple trees in order to deciding the outcome. Therefore it gains an advantage by repeating itself. Moreover, SVM is basically a linear separator, if data cannot be seperated linearly, A kernal needs to be occupied to manage the data into a point where it can separate it and it is the greatest advantage and also disadvantage means though it can identify a linear separation for almost any data but at the same time the model requires to occupy a Kernel and it cannot be guaranteed that the kernel will work for every set of data. Most importantly , SVM is normally designed for binary classification problems. For handling multi-class problems, either we perform 1-against-all strategy or do other tricks like optimizing classes together which cannot outperform other algorithms like random forest and boosting.

5 Conclusion

The study is solely focused on predicting the severity level of Covid situation for different areas analyzing the features of locations, number of death, recovered and confirmed cases with the highest accuracy of 95.07%. Though the work is done in a particular dataset from kaggle , this will also work very efficiently for any other available data provided with the same features. Because it is a very dynamic work with a popular epidemical model SIR and multiple machine learning motheds with modified parameters. In future, more machine learning algorithms can be tried to improve the accuracy and also this work can be extended by adding new feature of vaccination. Depending on the old features and the vaccination feature, the models can be trained to predict the next Covid situation of different locations after getting 1st and 2nd dose vaccination.

References

1. Gupta V. K., Gupta A., Kumar D. and Sardana A.(2021), Prediction of Covid-19 confirmed, death, and cured cases in India using random forest model, in Big Data Mining and Analytics, vol. 4, no. 2, pp. 116-123, doi: 10.26599/BDMA.2020.9020016

2. Kolla, B. (2021), Analysis, Prediction and Evaluation of Covid-19 Datasets using Machine Learning Algorithms, Volume 8. No. 5,doi: 10.30534/ijeter/2020/117852020
3. Coronavirus. (2021), Retrieved from https://www.who.int/health-topics/coronavirus#tab=tab_1
4. Pinter G, Felde I, Mosavi A, et al. (2020), Covid-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. Mathematics 8:890. doi: 10.3390/math8060890
5. Rustam F. et al.(2020), Covid-19 Future Forecasting Using Supervised Machine Learning Models, in IEEE Access, vol. 8, pp. 101489-101499, doi: 10.1109/ACCESS.2020.2997311
6. Ardabili, S.F.; Mosavi, A.; Ghamisi, P.; Ferdinand, F.; Varkonyi-Koczy, A.R.; Reuter, U.; Rabczuk, T.; Atkinson, P.M.(2020), Covid-19 Outbreak Prediction with Machine Learning. Algorithms. doi: org/10.3390/a13100249
7. Zoabi, Y., Deri-Rozov, S. & Shomron, N.(2021), Machine learning-based prediction of Covid-19 diagnosis based on symptoms. npj Digit.Med. 4, 3. doi: 10.1038/s41746-020-00372-6
8. Yan L, Zhang H, Goncalves J, et al.(2020), A machine learning-based model for survival prediction in patients with severe Covid-19 infection. medRxiv; DOI: 10.1101/2020.02.27.20028027
9. Iqbal, N.; Islam, M.(2019), Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers. Informatica, 43, 363–371, doi: 10.31449/inf.v43i3.1548
10. The SIR epidemic model. (2021). Retrieved from https://scipython.com/book/chapter-8
11. Rokach L., Maimon O. (2005), Decision Trees. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA. https://doi.org/10.1007/0-387-25465-X_9
12. Breiman, L., Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324
13. Sperandei S. (2014)., Understanding logistic regression analysis. Biochemia medica, 24(1), 12–18. https://doi.org/10.11613/BM.2014.003
14. Noh, B., Youm, C., Goh, E., Lee, M., Park, H., Jeon, H., & Kim, O. (2021)., XGBoost based machine learning approach to predict the risk of fall in older adults using gait outcomes. Scientific Reports, 11(1). doi: 10.1038/s41598-021-91797-w
15. Shmilovici, A., Support Vector Machines. Data Mining And Knowledge Discovery Handbook, 257-276. doi: 10.1007/0-387-25465-x_12