Multi-class classification and modeling of the hospitalization status of COVID-19 patients

Taiwo Olubunmi Adetiloye
BigCodeGen LLC,
Fort Worth, Texas, USA
Email: taiwo@bigcodegen.com
ORCID: 0000-0002-0172-7477

Emmanuel Jesuyon Dansu
Department of Mathematical Sciences,
Federal University of Technology,
Akure, Ondo State, Nigeria
Email: ejdansu@futa.edu.ng
ORCID: 0000-0002-7831-390X

Akinkunle Akinola
Department of Data Science,
Bowling Green State University,
Bowling Green, Ohio, USA
Email: akinola@bgsu.edu
ORCID: 0000-0002-2665-0496

***Abstract*—In recent times, the unprecedented surge in the Coronavirus disease 2019 (COVID-19) due to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has led to several attempts at understanding and containing the outbreak of the pandemic as well as to ultimately eradicate it. Steps taken so far include encouraging the wearing of face masks and shields, municipality restrictions such as work-from-home orders, the development of vaccines by health research institutions among others. It is widely believed that the main mode of transmission of the virus is from human to human. In this paper, we present the multi-class classification and modeling of the hospitalization status of COVID-19 patients by using both machine learning and compartmental mathematical models focusing on critical factors like hospital stay-days (SDs) and admission type based on severity of illness. The classification of hospitalization status of COVID-19 patients is necessary in order to know priority cases and give them prompt attention. Two key machine learning algorithms-the decision tree and random forest, are deployed in our analyses. The Levenberg–Marquardt (L-M) algorithm was used for parameter estimation for the mathematical model. From our results, it is easy to identify high risk patients in order to optimize treatment plans that would lower cost of treatments, reduce the chances of others getting infected and assist logistics teams to optimally allocate hospital resources. Hospital administrations can also be supported in deciding the number of staff and visitors per patient per day in a facility.**

I. Introduction

COVID-19 can be categorized as a black swan event of our time. As a result of its massive impact, it requires not only algorithm-based solutions but also detailed mathematical modeling and comprehensive analysis [2], [12]. [12] argues that black swan events can be very difficult to predict using standard tools of probability and prediction. This is due to the fact that such events, by definition, lack a large population and that past sample sizes are hardly available. Moreover, using statistics based extrapolation techniques on observations of past events may not be helpful in predicting them as there might be more chances of increasing human vulnerability to them.

Data practitioners have performed data mining to effectively model and predict the patterns of COVID-19 in order to generate meaningful insights to guide these efforts. For instance, [4] applied machine learning and natural language processing models on multiple data sets gotten from various sources such as PubMed and ArXiv to characterize the evolutionary trends of recent COVID-19 research themes. This was done by identifying the latent topics and analysing the publications' sentiments and similarities from January to May 2020, a duration widely considered as an early phase in the first-wave of the pandemic. The paper serves to create awareness around the level of attention given to high-risk groups, which comprise mostly of elderly people and people with underlying conditions. Similarly, a mathematical model of COVID-19 containing asymptomatic and symptomatic classes was put forward by [3]. This is an attempt to understand the pattern of the outbreak by analysing the data from the Nigeria Center for Disease Control (NCDC) and the World Health Organization (WHO).

A machine learning model for predicting severity prognosis in patients infected with COVID-19 was presented by [9] with evaluation of the chest CT of patients showing respiratory syndromes and positive epidemiological factors for COVID-19 infection. This is in order to find if correlations exist between the factors that cause the coronavirus disease. [1] analyzed COVID-19 related parameters and recommended various machine learning techniques to predict the pattern of its outbreak while utilizing real-time data from countries across the globe. A survey on deep learning applications for COVID-19 is seen in the work by [11] based on the evaluation of the current state of deep learning and its key limitations for COVID-19 treatment research.

In this paper, our objective is to classify the instances relevant to the critical factors into exactly one of the different classes vis-à-vis the SDs. This can be considered as a typical multi-class classification problem. We seek to validate our model using two machine learning algorithms: decision tree and random forest. As generally assumed for model predictions (PREDs), the decision tree generates a set of rules based on the widely known recursive partitioning (divide and conquer) approach. On the other hand, the random forest, a type of ensemble method, combines several decision trees in order to make its prediction based on majority votes. In addition, a compartmental mathematical model is formulated and analysed to study the dynamics of the hospitalization status of COVID-19 patients.

978-1-6654-7978-3/22/$31.00

In summary, our contributions are the following:

(i) multi-class classification of patients' SDs in hospitals using two key machine learning algorithms: decision tree and random forest;
(ii) formulation of a compartmental mathematical model to show points of focus in optimizing treatments for high risk patients;
(iii) parameter optimization using the L-M algorithm and simulation of the compartmental mathematical model.

The rest of this paper is organized as follows: Section II presents the materials and methods for the research. Section III bothers on the compartmental mathematical model. In Section IV, results from the models are discussed. In Section V, relevant conclusions are drawn.

II. Materials and methods

A. Data collection

To perform our experiment, we retrieved the Kaggle COVID-19 pandemic dataset[7] which is open source and publicly licensed for data scientists and practitioners. It is noteworthy that the dataset has been redacted in order to conceal the patients' identities, hence, making it useful for modeling while eliminating the underlying bias against any particular group or health facility.

B. Data preparation

Data preparation is essential before performing data analytics and modeling. This stage involves pre-processing and transformation of the raw data into forms that can be readily available for analysis. For the purpose of generating insights, we rely mainly on pyspark[8] since it supports high (parallel) distributed computing for fast data processing. Also, in our experimental setup, we use Google Collab Notebook with the runtime set to a GPU Hardware accelerator. In this dataset, the number of observations is 318438 with 15 critical factors as well as a multi-class label based on the SDs. Table I presents the summary statistics of the dataset.

TABLE I
Summary Statistics

Target variable	Class	Counts
SDs	0–10	23604
	11–20	78139
	21–30	87491
	31–40	55159
	41–50	11743
	51–60	35018
	61–70	2744
	71–80	10254
	81–90	4838
	91–100	2765
	More than 100 days	6683
	Total	318438

C. Critical factors

The following critical factors are considered: hospital (H), hospital-type (HT), hospital city (HC), hospital region (HR), available extra rooms in hospital (AERs), department (D), ward-type (WT), ward-facility (WF), bed-grade (BG), city code patient (CCP), type of admission (TA), illness severity (IS), patient-visitors (PVs), age (A), and admission deposit (AD).

III. Compartmental mathematical model

A. Assumptions

In order to obtain the model, the following are assumed.

(i) There are three types of hospital admission for COVID-19 patients, viz. *emergency*, *urgent* and *trauma*.
(ii) There is recruitment in each admission type proportional to the total hospital capacity.
(iii) Patients can transition from one type of admission to the other over time.
(iv) Patients are discharged from each type of admission when they recover or die.

B. Model

The model for the hospitalization status of COVID-19 patients is given as:

$$\begin{aligned}
\frac{dE}{dt} &= \bar{\lambda}_1 - \beta_{12}\frac{E}{H} - \beta_{13}\frac{E}{H} + \beta_{21}\frac{U}{H} + \beta_{31}\frac{T}{H} - \delta_1\frac{E}{H} \\
\frac{dU}{dt} &= \bar{\lambda}_2 - \beta_{21}\frac{U}{H} - \beta_{23}\frac{U}{H} + \beta_{12}\frac{E}{H} + \beta_{32}\frac{T}{H} - \delta_2\frac{U}{H} \qquad (1) \\
\frac{dT}{dt} &= \bar{\lambda}_3 - \beta_{31}\frac{T}{H} - \beta_{32}\frac{T}{H} + \beta_{13}\frac{E}{H} + \beta_{23}\frac{U}{H} - \delta_3\frac{E}{H}
\end{aligned}$$

with initial condition $(E(0), U(0), T(0)) = (E_0, U_0, T_0)$, where

- $E(t)$ is the population size of hospitalized individuals who have emergency status at time t;
- $U(t)$ is the population size of hospitalized individuals who have urgent status at time t;
- $T(t)$ is the population size of hospitalized individuals who have trauma status at time t;
- $H = E(t) + U(t) + T(t)$ is the total bed spaces in the hospital;
- the $\bar{\lambda}_i$s represent the hospitalization rates (i.e. number of patients admitted per day) in each admission type;
- the δ_is represent the discharge rates from each type of admission due to recovery or death;
- the β_{ij}s represent transitions from one type of admission to another.

C. Equilibrium state

We obtain the equilibrium state by setting $\frac{dE}{dt} = 0$, $\frac{dU}{dt} = 0$ and $\frac{dT}{dt} = 0$ such that

$$\begin{aligned} \lambda_1 - A_1E + \beta_{21}U + \beta_{31}T = 0; \\ \lambda_2 + \beta_{12}E - A_2U + \beta_{32}T = 0; \\ \lambda_3 + \beta_{13}E + \beta_{23}U - A_3T = 0; \end{aligned} \tag{2}$$

where $\lambda_i = \bar{\lambda}_i H, i = 1, 2, 3$, $A_1 = \beta_{12} + \beta_{13} + \delta_1$, $A_2 = \beta_{21} + \beta_{23} + \delta_2$ and $A_3 = \beta_{31} + \beta_{32} + \delta_3$.

The equilibrium values are obtained as

$$E = -\frac{B_1}{B}; \quad U = -\frac{B_2}{B}; \quad T = -\frac{B_3}{B}; \tag{3}$$

for

- $B_1 = \beta_{21}\beta_{32}\lambda_3 + \beta_{21}A_3\lambda_2 + \beta_{23}\beta_{31}\lambda_2 + \beta_{31}A_2\lambda_3 - \beta_{23}\beta_{32}\lambda_1 + A_2A_3\lambda_1$;
- $B_2 = \beta_{32}A_1\lambda_3 + A_1A_3\lambda_2 + \beta_{12}\beta_{31}\lambda_3 - \beta_{13}\beta_{31}\lambda_2 + \beta_{12}A_3\lambda_1 + \beta_{13}\beta_{32}\lambda_1$;
- $B_3 = \beta_{23}A_1\lambda_2 + A_1A_2\lambda_3 - \beta_{12}\beta_{21}\lambda_3 + \beta_{13}\beta_{21}\lambda_2 + \beta_{12}\beta_{23}\lambda_1 + \beta_{13}A_2\lambda_1$; and
- $B = \beta_{23}\beta_{32}A_1 - A_1A_2A_3 + \beta_{12}\beta_{21}A_3 + \beta_{13}\beta_{21}\beta_{32} + \beta_{12}\beta_{23}\beta_{31} + \beta_{13}\beta_{31}A_2$.

D. Stability analysis

In order to investigate the stability of our model, we set $\frac{dE}{dt} = E'$, $\frac{dU}{dt} = U'$, and $\frac{dT}{dt} = T'$ so that we obtain the Jacobian matrix as follows.

$$J(E,U,T) = \begin{pmatrix} \frac{\partial E'}{\partial E} & \frac{\partial E'}{\partial U} & \frac{\partial E'}{\partial T} \\ \frac{\partial U'}{\partial E} & \frac{\partial U'}{\partial U} & \frac{\partial U'}{\partial T} \\ \frac{\partial T'}{\partial E} & \frac{\partial T'}{\partial U} & \frac{\partial T'}{\partial T} \end{pmatrix}, \tag{4}$$

$$J(E,U,T) = \frac{1}{H}\begin{pmatrix} -A_1 & \beta_{21} & \beta_{31} \\ \beta_{12} & -A_2 & \beta_{32} \\ \beta_{13} & \beta_{23} & -A_3 \end{pmatrix} \tag{5}$$

with characteristic equation

$$\begin{aligned} &P(\lambda) = \\ &\lambda^3 + \frac{1}{H}(A_1 + A_2 + A_3)\lambda^2 \\ &+ \frac{1}{H}(A_1A_2 + A_1A_3 + A_2A_3 - \beta_{12}\beta_{21} \\ &- \beta_{13}\beta_{31} - \beta_{23}\beta_{32})\lambda \\ &- \frac{1}{H}(A_1\beta_{23}\beta_{32} - A_1A_2A_3 + \beta_{12}\beta_{21}A_3 + \beta_{13}\beta_{21}\beta_{32} \\ &+ \beta_{12}\beta_{23}\beta_{31} + \beta_{13}\beta_{31}A_2) \end{aligned} = 0. \tag{6}$$

Given that

- $c_{11} := (A_1 + A_2 + A_3)/H$,
- $c_{21} := (A_1A_2 + A_1A_3 + A_2A_3 - \beta_{12}\beta_{21} - \beta_{13}\beta_{31} - \beta_{23}\beta_{32})/H$,
- $c_{31} := -(A_1\beta_{23}\beta_{32} - A_1A_2A_3 + \beta_{12}\beta_{21}A_3 + \beta_{13}\beta_{21}\beta_{32} + \beta_{12}\beta_{23}\beta_{31} + \beta_{13}\beta_{31}A_2)/H$,
- $d_{11} := (c_{11}c_{21} - c_{31})/c_{11}$,

then the equilibrium state is stable if $c_{11} > 0$, $c_{21} > 0$, $c_{31} > 0$ and $d_{11} > 0$ going by the Routh-Hurwitz stability criterion.

IV. Results and Discussions

A. Data analyses

Our data analysis techniques are:

1) Bivariate analyses: Figures 1, 2 and 3 are the bivariate analyses of the decision tree and random forest, for the actual and the predicted results, respectively. The bivariate analyses entails the examination of two variables (SDs/PREDS and IL = Extreme, Moderate, Minor) in order to determine their empirical relationship. This allows us to establish the simple hypotheses of the connection between the variables under test and the closeness of actual to the predicted results in our models.

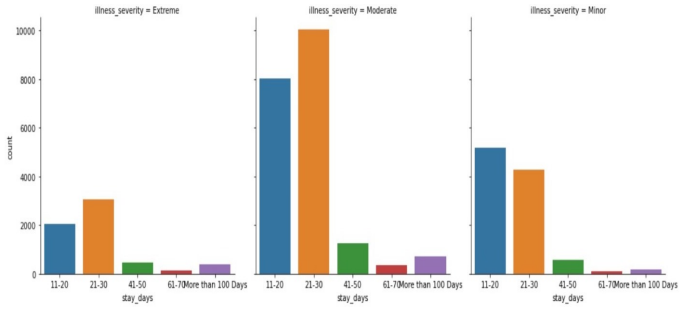

Fig. 1. Bivariate analysis of the actual data

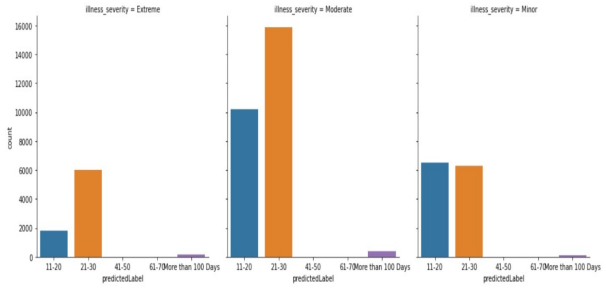

Fig. 2. Bivariate analysis of decision tree predictions

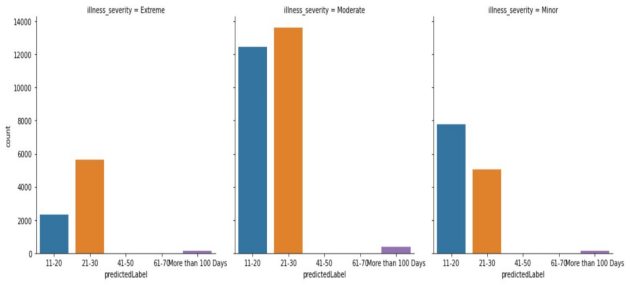

Fig. 3. Bivariate analysis of random forest predictions

2) Correlation matrices: Figures 4 and 5 are the correlation matrices based on the decision tree and random forest, respectively. The figures display the correlation coefficients which depict the correlation between critical factors that affect a patients' SDs in a hospital. Here, negative correlation coefficients indicate negative linear correlation between two factors; zero (0) indicates no linear correlation; and positive correlation coefficients indicate positive linear correlation between two factors. In the figures, we see that all the critical factors with the exception of the room availability and admission deposit are positively correlated with the estimated SDs.

	H	HT	HC	HR	AER	BG	CCP	PVs	AD	D	WF	WT	IS	TA	SDs	PREDs
H	1	0.04	0.13	-0.14	-0.06	-0.01	-0.02	-0.03	0.04	0.01	0.13	0.01	0	0	0.04	0
HT	0.04	1	-0.01	0.27	-0.1	0.03	0.06	-0.04	0.02	-0.03	0.26	0.22	0	-0.01	0.05	0.05
HC	0.13	-0.01	1	0.01	-0.04	-0.05	-0.03	0.02	-0.04	-0.02	0.1	-0.04	0	0.04	0.01	0.07
HR	-0.14	0.27	0.01	1	-0.01	-0.04	0.15	-0.02	0.07	-0.03	0.51	0.19	0.01	0.05	0.04	0.02
AER	-0.06	-0.1	-0.04	-0.01	1	-0.12	-0.01	0.09	-0.14	-0.09	-0.05	-0.05	-0.02	0.04	-0.1	-0.16
BG	-0.01	0.03	-0.05	-0.04	-0.12	1	-0.01	0.09	0.08	0.05	-0.08	-0.02	-0.05	-0.06	0.04	0.31
CCP	-0.02	0.06	-0.03	0.15	-0.01	-0.01	1	-0.01	0.03	-0.01	0.14	0.08	0.01	0.04	0.02	0
PVs	-0.03	-0.04	0.02	-0.02	0.09	0.09	-0.01	1	-0.15	0.01	-0.03	0.05	0.02	-0.02	0.4	0.66
AD	0.04	0.02	-0.04	0.07	-0.14	0.08	0.03	-0.15	1	0.08	0.02	-0.03	-0.02	-0.08	-0.09	-0.11
D	0.01	-0.03	-0.02	-0.03	-0.09	0.05	-0.01	0.01	0.08	1	-0.01	0.02	0	-0.02	0	0.05
WF	0.13	0.26	0.1	0.51	-0.05	-0.08	0.14	-0.03	0.02	-0.01	1	0.22	0.01	0.07	0.03	0.09
WT	0.01	0.22	-0.04	0.19	-0.05	-0.02	0.08	0.05	-0.03	0.02	0.22	1	0.01	0.03	0.08	0.04
IS	0	0	0	0.01	-0.02	-0.05	0.01	0.02	-0.02	0	0.01	0.01	1	0	0.03	-0.01
TA	0	-0.01	0.04	0.05	0.04	-0.06	0.04	-0.02	-0.08	-0.02	0.07	0.03	0	1	0	-0.02
SDs	0.04	0.05	0.01	0.04	-0.1	0.04	0.02	0.4	-0.09	0	0.03	0.08	0.03	0	1	0.36
PREDs	0	0.05	0.07	0.02	-0.16	0.31	0	0.66	-0.11	0.05	0.09	0.04	-0.01	-0.02	0.36	1

Fig. 4. Correlation matrix based on decision tree

	H	HT	HC	HR	AERs	BG	CCP	PVs	AD	D	WF	WT	IS	TA	SDs	PREDs
H	1	0.04	0.13	-0.14	-0.06	-0.01	-0.02	-0.03	0.04	0.01	0.13	0.01	0	0	0.04	0
HT	0.04	1	-0.01	0.27	-0.1	0.03	0.06	-0.04	0.02	-0.03	0.26	0.22	0	-0.01	0.05	0.06
HC	0.13	-0.01	1	0.01	-0.04	-0.05	-0.03	0.02	-0.04	-0.02	0.1	-0.04	0	0.04	0.01	0.06
HR	-0.14	0.27	0.01	1	-0.01	-0.04	0.15	-0.02	0.07	-0.03	0.51	0.19	0.01	0.05	0.04	0.04
AERs	-0.06	-0.1	-0.04	-0.01	1	-0.12	-0.01	0.09	-0.14	-0.09	-0.05	-0.05	-0.02	0.04	-0.1	-0.2
BG	-0.01	0.03	-0.05	-0.04	-0.12	1	-0.01	0.09	0.08	0.05	-0.08	-0.02	-0.05	-0.06	0.04	0.28
CCP	-0.02	0.06	-0.03	0.15	-0.01	-0.01	1	-0.01	0.03	-0.01	0.14	0.08	0.01	0.04	0.02	-0.02
PVs	-0.03	-0.04	0.02	-0.02	0.09	0.09	-0.01	1	-0.15	0.01	-0.03	0.05	0.02	-0.02	0.4	0.73
AD	0.04	0.02	-0.04	0.07	-0.14	0.08	0.03	-0.15	1	0.08	0.02	-0.03	-0.02	-0.08	-0.09	-0.1
D	0.01	-0.03	-0.02	-0.03	-0.09	0.05	-0.01	0.01	0.08	1	-0.01	0.02	0	-0.02	0	0.06
WF	0.13	0.26	0.1	0.51	-0.05	-0.08	0.14	-0.03	0.02	-0.01	1	0.22	0.01	0.07	0.03	0.1
WT	0.01	0.22	-0.04	0.19	-0.05	-0.02	0.08	0.05	-0.03	0.02	0.22	1	0.01	0.03	0.08	0.09
IS	0	0	0	0.01	-0.02	-0.05	0.01	0.02	-0.02	0	0.01	0.01	1	0	0.03	0
TA	0	-0.01	0.04	0.05	0.04	-0.06	0.04	-0.02	-0.08	-0.02	0.07	0.03	0	1	0	-0.03
SDs	0.04	0.05	0.01	0.04	-0.1	0.04	0.02	0.4	-0.09	0	0.03	0.08	0.03	0	1	0.4
PREDs	0	0.06	0.06	0.04	-0.2	0.28	-0.02	0.73	-0.1	0.06	0.1	0.09	0	-0.03	0.4	1

Fig. 5. Correlation matrix based on random forest

3) Compartmental mathematical modeling: In order to see the temporal variation of hospitalization by admission types, the optimal parameter values of the mathematical model were estimated from the actual data using the Levenberg–Marquardt (L-M) algorithm. The simulations of the model with the obtained values are illustrated in Figure 6 showing that patients who require urgent attention supersede others by the hundredth day and Figure 7 showing that patients who require emergency attention supersede others by the hundredth day.

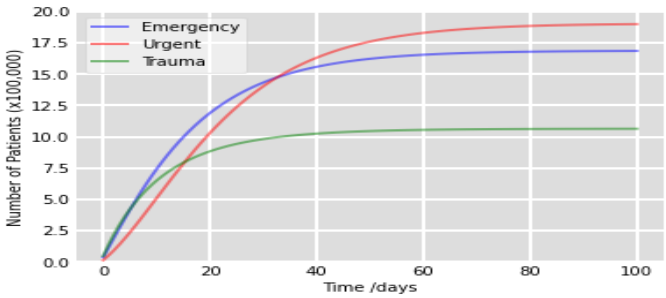

Fig. 6. Temporal variation of admission types with parameters estimated directly from actual data after optimization using L-M algorithm. $(E_0, U_0, T_0) = (0.37, 0.15, 0.48)$; $\lambda_1 = 0.77$, $\lambda_2 = 0.32$, $\lambda_3 = 1.00$; $\beta_{12} = 0.100$, $\beta_{13} = 0.010$, $\beta_{21} = 0.010$, $\beta_{23} = 0.001$, $\beta_{31} = 0.100$, $\beta_{32} = 0.010$, $\delta_1 = 0.01$, $\delta_2 = 0.100$, $\delta_3 = 0.001$.

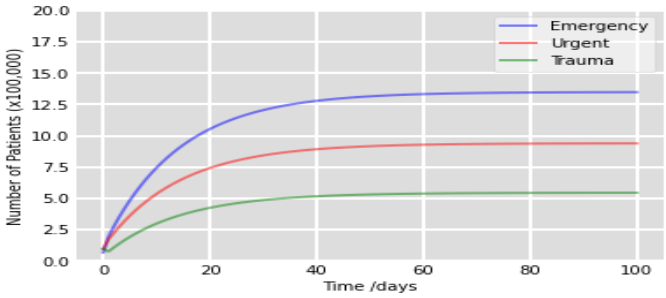

Fig. 7. Temporal variation of admission types with parameters estimated directly from actual data after optimization using L-M algorithm. $(E_0, U_0, T_0) = (0.68, 0.98, 0.97)$; $\lambda_1 = 0.24$, $\lambda_2 = 1.64$, $\lambda_3 = 0.13$; $\beta_{12} = 0.35$, $\beta_{13} = 0.97$, $\beta_{21} = 1.92$, $\beta_{23} = 0.32$, $\beta_{31} = 0.26$, $\beta_{32} = 2.71$, $\delta_1 = 0.14$, $\delta_2 = 0.01$, $\delta_3 = 0.002$.

B. Discussions

TABLE II
KEY CLASSIFICATION METRICS FOR MEASURING THE PERFORMANCES OF THE MACHINE LEARNING MODELS.

Metric	Decision Tree Classifier	Random Forest
Accuracy/Weighted Recall	0.35411	0.361112
F1/Weighted FMeasure	0.316467	0.311401
Hamming Loss	0.64589	0.638888
True Positive Rate/Recall	0.520559	0.632296
False Positive Rate	0.336413	0.401778
Precision	0.371852	0.375807
FMeasure	0.433816	0.471423
Weighted Precision	0.309697	0.290329
Log Loss	1.69783	1.68902
Runtime	0.27 secs	0.34 secs

Table II presents the key classification metrics for measuring the performances of the machine learning models. The results

indicate low precision and high recall for both algorithms with the random forest slightly better than the decision tree in the overall performance. However, the decision tree gives a better runtime. It is widely assumed that for an algorithm, the higher the precision the more the relevance of the results as a measure of the quality while higher recall implies that the algorithm returns more relevant results as a measure of the quantity. For partially low accuracy, we attribute this to the issue of imbalanced dataset which could be addressed using Synthetic Minority Oversampling Technique (SMOTE)[14]. This is done by duplicating the minority samples in order to improve the overall model performance. Imbalanced datasets occur regularly in multi-class classification problems. Dealing with class imbalance problems have continued to generate interests in academic research communities due to the difficult nature of classification caused by imbalanced class distributions[13], which can lead to poor model performance. However, we omit the SMOTE approach in order not to alter the objectivity of our results. For detailed definition of key metrics, readers can refer to [5].

As for the compartmental mathematical model, parameter optimization by the L-M algorithm gives the best values of the parameters for the model. In the early 1960s, the Levenberg-Marquardt method was created to tackle nonlinear least squares problems. Least squares difficulties arise when fitting a parameterized mathematical model to a set of data points by minimizing a goal stated as the sum of the squares of the model function and data point errors[6]. In our context, this algorithm helps to give parameter values that minimize the influx of COVID-19 patients into hospitals. It also results to the decrease of extreme cases represented by the trauma admission type.

V. Conclusion

This paper presents the multi-class classification and modelling of the hospitalization status of COVID-19 patients using the Kaggle COVID-19 dataset. Our findings corroborate the hypothesis that the attributes of black swan events can make accurate PREDs difficult. Apart from this phenomenon of the black swan, the limitations of the mathematical model, e.g. parameter estimation, is the inappropriateness for a NP-hard problem which yields optimal results only for a small number of decision variables [10], [15].

In addition, we observe the issue of imbalanced data which has the possibility of altering the performance of the algorithms measured based on the accuracy and precision without completely undermining the objectivity of our multi-class classification as seen in the validity of the recall.

In future works, our research would be directed towards new COVID-19 data with regards to geo-location assessment using spatial models. This is to estimate the likelihood of having COVID-19 spread with a root cause analysis. We would build machine learning pipelines to validate the model and measure the performance of our algorithms to support the decisions of relevant authorities and hospital managements in mitigating the impacts of COVID-19. We would examine the possibility of applying SMOTE to improve the model accuracy in the context of ongoing COVID-19 research efforts with emphasis on the complexity of the black swan event that COVID-19 represents.

References

[1] R. Agrawal and N. Gupta. *Analysis of COVID-19 data using machine learning techniques*. In: Khanna A., Gupta D., Pólkowski Z., Bhattacharyya S., Castillo O. (eds). Data Analytics and Management. Lecture Notes on Data Engineering and Communications Technologies, **54** (2021): Springer, Singapore. https://doi.org/10.1007/978-981-15-8335-3_45.

[2] I. Ahmad. *40 algorithms every programmer should know: Hone your problem-solving skills by learning different algorithms and their implementations in Python*. Packt Publishing Ltd.

[3] I. Ahmed, G.U. Modu, A. Yusuf, P. Kumam, I. Yusuf. *A mathematical model of Coronavirus disease (COVID-19) containing asymptomatic and symptomatic classes*. Results in Physics, 21 (2021) 103776.

[4] A. Ebadi, P. Xi, S. Tremblay, B. Spencer, R. Pall, A. Wong. *Understanding the temporal evolution of COVID-19 research through machine learning and natural language processing*. Scientometrics, (2020): 1–15. https://doi.org/10.1007/s11192-020-03744-7

[5] P.A. Flach. *The geometry of ROC space: Understanding machine learning metrics through ROC isometrics*. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC (2003).

[6] H.P. Gavin. *The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems*. Department of Civil and Environmental Engineering, Duke University, Durham, USA (2020).

[7] Kaggle. *COVID-19 hospital treatments plan. https://www.kaggle.com/arashnic/covid19-hospital-treatment*. Retrieved March 15, 2021.

[8] H. Karau. *Fast data processing with Spark: High speed distributed computing made easy with Spark*. Packt Publishing, Birmingham-Mumbai (2013).

[9] F.P.P.L. Lopes, F.C. Kitamura, G.F. Prado, P.E. de Aguiar Kuriki, and M.R.T. Garcia. *Machine learning model for predicting severity prognosis in patients infected with COVID-19: Study protocol from COVID-AI Brasil*. PLoS ONE, **16** (2021). https://doi.org/10.1371/journal.pone.0245384

[10] N. Olgac and R. Sipahi. *An exact method for the stability analysis of time-delayed linear time-invariant (LTI) systems*. IEEE Transactions on Automatic Control, **47**(5) (2002).

[11] C. Shorten, T.M. Khoshgoftaar and B. Furht. *Deep learning applications for COVID-19*. Journal of Big Data, **8**(18) (2021). https://doi.org/10.1186/s40537-020-00392-9

[12] N.N. Taleb. *Fooled by randomness: The hidden role of chance in the markets and in life*. Texere, New York (2001).

[13] S. Wang, X. Yao. *Multiclass imbalance problems: Analysis and potential solutions*. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics, **42**(4) (2012).

[14] T. Zhu, Y. Lin, and Y. Liu. *Synthetic minority oversampling technique (SMOTE) for multiclass imbalance problems*. Journal of Pattern Recognition, **72** (2020): 327–340.

[15] N. Zufferey, D.D. Molin, R. Glardon, and C. Tsagkalids. *A simulation-optimization approach for the production of components for a pharmaceutical company*. Handbook of Research on Applied Optimization Methodologies in Manufacturing Systems (2018): 269–283.