Predicting Severe Cases of COVID-19

Htrf88
Department of Computer Science
Durham University
Durham, England

Abstract — **Three classifiers were trained on epidemiological data from the COVID-19 outbreak to predict and gain insists into the severity of COVID-19 cases based on a patient's age, sex, pre-existing health conditions and location. A patient's location was the biggest factor in the determining the severity of their case. The classifiers showed a higher risk of a case becoming severe with age and pre-existing health conditions. Males were found to be at a higher risk of having a severe COVID-19 case compared to females.**

I. Introduction

Since the first case of COVID-19 was reported in December 2019, it has quickly spread worldwide causing the ongoing pandemic, and after 1.4 years and 153 million reported cases it is the cause of upwards of 3.2 million deaths [1]. An infection from Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the virus that causes COVID-19, can lead to respiratory tract damage and fatal lung failure damaging the respiratory system, central nervous system, and the digestive system [2]. In the severe cases of COVID-19, the patients develop pneumonia due to difficulty in breathing and need to be admitted into an intensive care unit for assisted breathing. The pandemic has put a further strain on medical infrastructure across the globe. Nearly half of hospitals in low- and middle-income countries have inconsistent or no supply of oxygen [3]. Even to the present day, patients in many parts of the world are unable to access ventilators and medical oxygen they critically need.

Several issues, such as problems in scaling the vaccine production, pose a challenge to ensuring global access to a COVID-19 vaccine and leave billions of individuals around the globe without access to a vaccine. This poses the threat of prolonging the pandemic, raising the risk of further mutations of the virus, and possibly undermining the efficacy of existing vaccines [4]. Thus, it is likely that the end of the pandemic is not close for most of the world population.

Therefore, being able to predict if a COVID-19 patient is at the risk of needing hospitalization for assisted breathing or other additional care can help hospitals and local/state governments anticipate the need for medical supplies, and prepare accordingly and will help better manage their current resources. Precautionary measures such as identifying and cautioning individuals most at risk of needing hospitalization if infected with COVID-19 to reduce strain on sparse resource can also be achieved with such a tool. Therefore, I have implemented three classifiers which can predict if a COVID-19 infection will become severe based on the available information about the patient. A severe case of COVID-19 is defined as the case which requires hospitalization and/or it leads to death.

II. Experimental Procedure

The nCoV2019 dataset [5], contains the information of cases at an individual level including the patient's age, sex, dates of case confirmation/death/discharge, case outcome, pre-existing health condition, geolocation, and other administrative data. After examining the attributes of the dataset, the ones relating to administrative information and the ones that did not contain any information were removed. Age, race/ethnicity, gender, some medical conditions, poverty, and crowding have been identified as potential risk factors for a severe illness or complications from COVID-19 [6]. Therefore, of the remaining attributes in the dataset the following were identified to be of importance — 'age', 'sex', 'chronic_disease_binary' which indicates the presence of the patient having a pre-existing medical condition, 'death_or_discharge' which contains the date of patient's discharge or death, 'outcome' which contains the last recorded health condition of the patient, 'date_admission_hospital' and geolocation data namely 'latitude' and 'longitude'.

The attributes 'chronic_disease_binary', 'latitude' and 'longitude' did not need any cleaning. Values in the 'age' attribute were a mix of ranges and discrete values with '35-59' and '15-34' being the most frequent values. I decided to convert this into a continuous attribute by taking the average where a range was recorded, the ranges with difference greater than 24 were considered erroneous and were removed from the dataset. Age attribute was standardized before training and testing the classifiers. The 'sex' attribute was encoded as 1 if 'male' and 0 if 'female'.

To determine if a case was severe, I checked for non-null entries in the 'date_admission_hospital' attribute and the 'date_death_or_discharge' attribute to identify cases where the patient was hospitalized or deceased. Similarly, I examined entries in the 'outcome' attribute to check for keywords such as "death", "severe", "critical", "intensive", etc. to identify the severe cases. This information was then combined into a single binary attribute 'is_severe' which will be used as the target attribute. Finally, rows with null entries were dropped after which the cleaned dataset had ~573.5k data points.

The ratio of severe cases to non-severe cases was 13:1. Ratio of male to females was 1.12:1 and ratio of cases with chronic diseases to cases without chronic diseases was 3394:1.

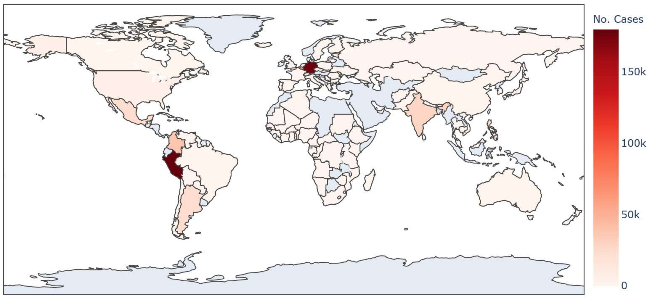

Figure 1 Geographic distribution of the cases in the cleaned dataset

Figure 1 shows the geographic distribution of cases in the cleaned dataset, in which the cases are highly concentrated in Central and South America, Germany, India, and the Philippines.

As seen in Figure 2, attributes in the cleaned dataset don't have any significant correlation (except latitude and longitude), therefore all of them were used as the input features.

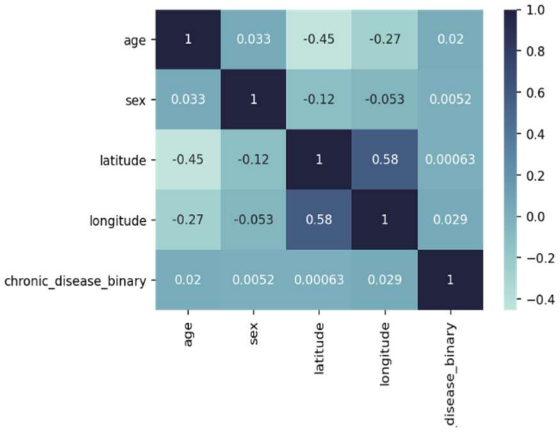

Figure 2 Feature correlation in the cleaned dataset

I wanted to observe the effects of hyperparameter tuning on the classification skills of the models, therefore I selected Random Forest (RF), K-Nearest Neighbors (KNN), and Logistic Regression (LR), all of which have various hyperparameters that can be tuned.

It is important that the classifiers represent both target classes well since poor representation of the positive class (i.e., a severe case) will result in an underestimation of the expected severe cases which will overwhelm the medical facilities and lead to a higher mortality rate. In contrast, poor representation of the negative class (i.e., a non-severe case) will result in an overestimation of the expected severe cases which can result in poor management of sparse medical resources also resulting in a higher mortality rate. Therefore, I used sensitivity, specificity, area under the ROC curve, F-1 score, and precision-recall curve to evaluate how well both classes are represented by the classifiers.

I first established baseline results by training the three classifiers on 70% of the cleaned and shuffled dataset, 30% was used for testing. Ratio of the target variable was preserved in the training and test data.

After training the baseline classifiers, the imbalanced dataset was balanced to avoid classifiers from overfitting to the majority class. Balancing was done by randomly under sampling the majority class to obtain a 1:1 ratio of severe to non-severe cases. The balanced dataset had 79,054 data points of which 70% were used for training and 30% used for testing. As with the baseline classifiers, the ratio of the target variable was preserved in the training and test data. I also, used stratified 3-fold cross validation while training the classifiers for hyperparameter tuning to further eliminate the possibility of the classifiers overfitting to a certain attribute. The list of hyperparameters used for tuning and their candidate values are detailed in Table 1.

Table 1. Hyperparameter and candidate values for the chosen classifiers

Classifier	Hyperparameter	Candidate Values
RF	max_features	sqrt, log2
	n_estimators	50, 100, 300
KNN	n_neighbours	7, 11, 19
LR	solver	newton-cg, lbfgs, sag
	C	0.5, 1, 1.5

III. RESULTS

Table 2. Results of the baseline classifiers.

	RF	KNN	LR
Accuracy	0.98	0.98	0.89
Sensitivity	0.81	0.81	0.04
Specificity	0.99	0.99	0.96
F1 Score	0.86	0.85	0.04

As seen in Table 2 and Figure 3, the baseline RF and KNN classifiers are very skilled and performed well on imbalanced data across all metrics. Both classifiers were better at predicting the negative class compared to the positive class but even with this imbalance positive class was predicted well. RF had an average precision (AP) of 93% whereas KNN had an AP of 89% indicating that RF is more skilled at accurately identifying the positive class compared to KNN.

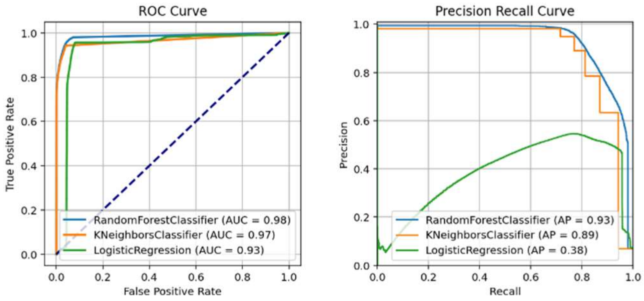

Figure 3 ROC and PR curve of baseline classifiers

The baseline LR was not a very skilled classifier, with just 4% sensitivity, F1 score of 0.04 and average precision of 38% it is evident that the classifier was overfitting to the negative class and was not very skilled at accurately predicting the positive class.

After balancing the dataset and retraining the classifiers, the optimal hyperparameters found are detailed in Table 3.

Table 3. Optimal hyperparameter values for chosen classifiers.

Classifier	Hyperparameter	Optimal Value
RF	max_features	sqrt
	n_estimators	100
KNN	n_neighbours	11
LR	solver	newton-cg
	C	1.5

Table 4. Results of the tuned classifiers.

	RF	KNN	LR
Accuracy	0.96	0.96	0.94
Sensitivity	0.97	0.98	0.96
Specificity	0.95	0.95	0.92
F1 Score	0.96	0.96	0.94

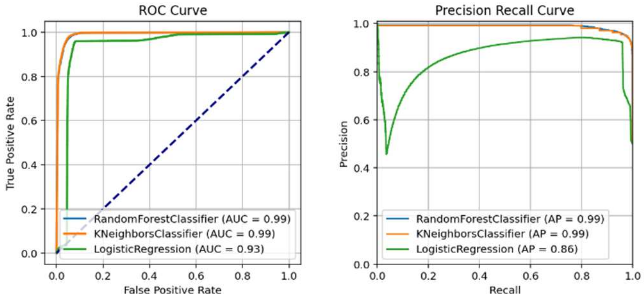

Figure 4 ROC and PR curve of tuned classifiers

As seen in Table 4 and Figure 4, amongst the tuned classifiers, LR saw the biggest improvement in classification skills across all metrics compared to it's baseline counterpart. The accuracy of LR improved from 89% to 94%, sensitivity improved from 4% to 96%, F1 score improved from 0.04 to 0.94 and AP increased from 38% to 86%.

RF and KNN also improved in their classification skills but saw trade-offs in the form of 2% reduction in accuracy and a 3% decrease in specificity. Though there were a significant increase in the sensitivity (improvement of 16 to 17%), F1 score (improvement of 11%), and average precision (improvement of 3 to 10%), indicating that both classifiers improved at predicting the positive cases.

Table 5. Feature importance of the tuned classifiers.

Classifier	Feature Importance (Higher to lower)
RF	longitude, latitude, age, chronic_disease_binary, sex
KNN	longitude, latitude, age, sex, chronic_disease_binary
LR	longitude, latitude, chronic_disease_binary, age, sex

As seen in Table 5, location of the patient was the most important feature for classification in all three classifiers. Generally, age was the second most important feature, followed by chronic_disease_binary. Sex was the least important feature.

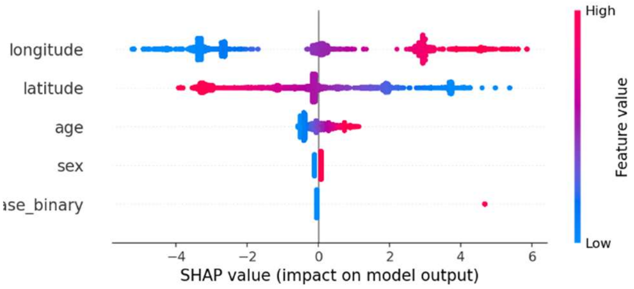

Figure 5 SHAP value impact of the tuned LR Classifier

The SHAP value impact (Figure 5) shows that the classifiers predicts that a case is severe if longitude values are positive (i.e., from Europe, and Asia), if latitude values are negative (i.e., around or below the equator), as the values of age increases, if the patient is a male and if the patient has a pre-existing health condition.

IV. Discussion of chosen classifiers, experimental procedure, and limitations

All classifiers saw significant improvements in their ability to accurately predict the severe COVID-19 cases after resampling the dataset and hyperparameter tuning. RF was the most skilled of the three classifiers having the highest accuracy, precision and F1 score for both balanced and imbalanced data. KNN has a very comparable performance to that of the RF classifier.

LR is the least skilled of the three classifiers. Even the baseline RF and KNN classifiers trained on imbalanced data performed significantly better compared to the baseline LR classifier. Whilst there were significant performance improvements after fitting the classifier on balanced data and hyperparameter tuning, its AP of 86% indicates that it was still more prone to false positive predictions when compared to its tuned RF and KNN counterparts with an almost perfect AP of 99%.

There is a wide variation between the severity of a COVID-19 infection for different demographic groups [6], due to which RF and KNN are better suited for this classification task compared to LR as evident from the results in Table 3 and Table 4.

The main limitations of this method were due to the limited information in the dataset. Of the identified risk factors of COIVD-19 [6], information about the patient's race and ethnicity, and details about the type of medical conditions the patients have, were not available but are vital for accurate predictions from a medical standpoint.

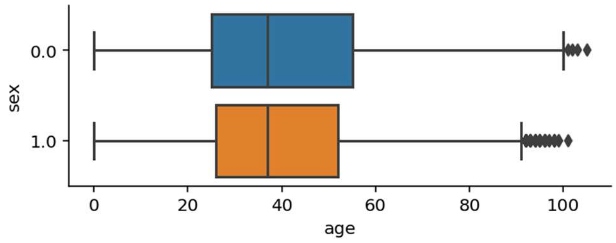

Figure 6 Distribution of severe CV-19 cases in the dataset by age

The interquartile range of the severe cases in the dataset (Figure 6) is between 26 and 53, but individuals aged 50 and over are the most vulnerable to severe illness [7], thus the classifiers would benefit from cases with a more diverse range of ages for practical applications.

Geographic distribution of the reported cases (Figure 1) has very significant impact on the classifiers' prediction (Table 5, Figure 5) but several countries which have seen or are seeing the highest infection rates of COVID-19 [1] e.g., the USA, Brazil, France, Turkey, Russia, the UK, etc. have very few to no usable entries in the dataset thus the classifiers could be prone to false predictions for cases from these countries.

Random under sampling when rebalancing the dataset adds to this disparity as a large amount of data is lost when under sampling the majority class. A more sophisticated method of sampling can be used while rebalancing the dataset, though I found these methods have significant cost in terms of running time due to the size of the dataset. Alternatively, since RF and KNN represented the positive class well even on imbalanced data, these classifiers can be trained on a dataset that isn't perfectly balanced to achieve comparable results to the ones detailed in Table 4 whilst reducing the loss of integrity or volume of the data when resampling.

V. Conclussion

Geolocation being the biggest determining factor in a severe case, the classifiers correctly identified the geo-clusters of cases in the cleaned dataset (Figure 1) which had the highest number reported cases. These observations are consistent with how the spread of the virus varied across the globe depending on the effectiveness of the type of non-pharmaceutical interventions (NPIs) the state/local governments adopted to mitigate the spread of COVID-19 [8]. The classifiers also correctly identified the correlation between older age groups and individuals with pre-existing health conditions (Figure 5) with the severity of a case which is consistent with the current studies. SHAP value impact (Figure 5) also shows an increased risk of a severe case for males compared to females, and this is consistent with other studies where male sex is associated with a significantly increased risk of ITU admission within COVID-19 patients [9].

In conclusion, the classifiers have correctly identified the factors that contribute towards the severity of a COVID-19 infection and can be used for proposed application but due to their limited exposure to certain demographic groups, their predictions may not be as reliable as the results.

VI. Lessons Learnt

Doing this assignment has given me valuable insights into the different stages of the machine learning workflow. I have learned how to extract information from a given dataset that can be useful in the context of the problem, various new concepts such as "curse of dimensionality" and "dummy variable trap" when I experimented with different ways of encoding information from the dataset, the importance of tuning hyperparameters, choosing the right evaluation metrics, and various other important concepts.

I feel confident in my ability to identify the applications of machine learning in my future projects and my skills to implement them.

References

[1] Wikipedia.org. 2021. *COVID-19 pandemic data*. [online] Available at:

<https://en.wikipedia.org/wiki/Template:COVID-19_pandemic_data> [Accessed 5 May 2021].

[2] Zhang, Y., Geng, X., Tan, Y., Li, Q., Xu, C., Xu, J., Hao, L., Zeng, Z., Luo, X., Liu, F. and Wang, H., 2020. *New understanding of the damage of SARS-CoV-2 infection outside the respiratory system.* Biomedicine & Pharmacotherapy, 127, p.110195.

[3] Craig Spencer MD MPH, Mar, 2021. *There's a Global Shortage of Medical Oxygen. Covid-19 Is Making It Worse* [online] <https://elemental.medium.com/theres-a-global-shortage-of-medical-oxygen-covid-19-is-making-it-worse-14553be5da09> [Accessed 4 May, 2021]

[4] Wouters, O., Shadlen, K., Salcher-Konrad, M., Pollard, A., Larson, H., Teerawattananon, Y. and Jit, M., 2021. *Challenges in ensuring global access to COVID-19 vaccines: production, affordability, allocation, and deployment.* The Lancet, 397(10278), pp.1023-1034.

[5] GitHub. 2021. *nCoV2019*. [online] Available at: <https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data> [Accessed 5 May 2021].

[6] Centers for Disease Control and Prevention. 2021. *Cases, Data, and Surveillance.* [online] Available at: <https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/assessing-risk-factors.html> [Accessed 5 May 2021].

[7] Centers for Disease Control and Prevention. 2021. *Risk for COVID-19 Infection, Hospitalization, and Death By Age Group.* [online] Available at: < https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html> [Accessed 5 May 2021].

[8] Haug, N., Geyrhofer, L., Londei, A. et al. *Ranking the effectiveness of worldwide COVID-19 government interventions.* Nat Hum Behav 4, 1303–1312 (2020). https://doi.org/10.1038/s41562-020-01009-0

[9] Peckham, H., de Gruijter, N.M., Raine, C. et al. *Male sex identified by global COVID-19 meta-analysis as a risk factor for death and ITU admission.* Nat Commun 11, 6317 (2020). https://doi.org/10.1038/s41467-020-19741-6