Using machine learning to create a decision tree model to predict outcomes of COVID-19 cases in the Philippines

Julius R. Migriño, Jr[a,b] and Ani Regina U. Batangan[a]

Correspondence to Julius R. Migriño, Jr (email: jrmjrmd-1@yahoo.com)

Objective: The aim of this study was to create a decision tree model with machine learning to predict the outcomes of COVID-19 cases from data publicly available in the Philippine Department of Health (DOH) COVID Data Drop.

Methods: The study design was a cross-sectional records review of the DOH COVID Data Drop for 25 August 2020. Resolved cases that had either recovered or died were used as the final data set. Machine learning processes were used to generate, train and validate a decision tree model.

Results: A list of 132 939 resolved COVID-19 cases was used. The notification rates and case fatality rates were higher among males (145.67 per 100 000 and 2.46%, respectively). Most COVID-19 cases were clustered among people of working age, and older cases had higher case fatality rates. The majority of cases were from the National Capital Region (590.20 per 100 000), and the highest case fatality rate (5.83%) was observed in Region VII. The decision tree model prioritized age and history of hospital admission as predictors of mortality. The model had high accuracy (81.42%), sensitivity (81.65%), specificity (81.41%) and area under the curve (0.876) but a poor F-score (16.74%).

Discussion: The model predicted higher case fatality rates among older people. For cases aged >51 years, a history of hospital admission increased the probability of COVID-19-related death. We recommend that more comprehensive primary COVID-19 data sets be used to create more robust prognostic models.

A novel coronavirus that causes severe respiratory symptoms was first detected in patients in Wuhan City, Hubei Province, China in December 2019. The World Health Organization declared the outbreak of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)[1] to be a public health emergency of international concern on 30 January 2020, and declared a pandemic on 11 March 2020. As of 22 August 2021, the Philippines' Department of Health (DOH) had tallied 1 839 635 total cases, 125 900 of which were active cases; there have been 1 681 925 recoveries and 31 810 deaths.[2] At that date, the virus had infected more than 209.9 million people and claimed more than 4.4 million lives worldwide.[3]

SARS-CoV-2 infection can cause a range of symptoms, from a common cold-like illness presenting with cough, dyspnoea, dysgeusia and fever to severe respiratory symptoms causing shock and multiorgan failure.[1] Cases in the Philippines are classified as mild, moderate, severe or critical.[4] According to Urwin, Kandola and Graziado (2020),[5] a community-friendly prognostic clinical risk prediction score for COVID-19 mortality, severity and complications and triage recommendations can be determined from current signs and symptoms, comorbidities, medical history and demographics. In that study, the demographic factors of risk included age, sex, country and partial postcode.

Predictive modelling could greatly help low- and middle-income countries such as the Philippines to address the COVID-19 pandemic by increasing the accuracy of diagnosis and prognosis of patients.[6,7] This may help in determining the outcomes and factors indicative of outcomes for patients with COVID-19. Predictive modelling may also aid policy-makers in determining which strategies are more effective, so that allocation of limited resources can be targeted to possible target populations more efficiently and cost-effectively, especially during a

[a] San Beda University College of Medicine, Manila, Philippines.
[b] School of Medicine and Public Health, Ateneo de Manila University, Pasig City, Philippines.
Published: 14 September 2021
doi: 10.5365/wpsar.2021.12.3.831

pandemic.[6] Predictive modelling may also help to inform patients about the possible course of their illness and help both patients and health-care workers to draw up diagnostic and therapeutic plans.[8]

Machine learning and artificial intelligence have been used to automate the detection of patterns in large data sets, especially in dealing with the massive amounts of data generated during a global event such as the current pandemic. Decision trees, a specific type of machine learning, are based on covariates to create a model for predicting outcomes.[9] Currently, artificial intelligence, including decision tree modelling, is being used in the COVID-19 pandemic for early detection and diagnosis, monitoring treatment, tracing contacts, developing drugs and vaccines, predicting cases and fatalities and even identifying the most vulnerable groups.[10–13] Machine learning has been used to identify demographic and clinical predictors of disease progression, which include age, sex, body temperature, associated signs and symptoms, minimum oxygen saturation, computed tomography scan features, C-reactive protein and lactic dehydrogenase levels and lymphocyte counts.[12,14]

In the Philippines, studies on COVID-19 modelling have been limited to compartmental models, such as "susceptible–infectious–recovered/removed" or "susceptible–exposed–infectious–recovered/removed" models,[15,16] usually for tracking epidemiological trajectories. Other types of models being used in the Philippines include regression analysis models to estimate case fatality rates[17] and to determine socioeconomic indicators of the number of cases.[18]

The aim of this study was to create a decision tree model with machine learning to predict outcomes (i.e. recovery or death) of COVID-19 cases based on publicly available data from the DOH COVID Data Drop.

METHODS

We used the publicly available DOH COVID Data Drop database for 25 August 2020.[19] This database is extracted from the COVID-19 information system by the DOH Epidemiology Bureau and is updated daily. The data are obtained from paper-based case investigation forms from all the regional epidemiology surveillance units in the country. The raw data set comprised 197 164 cases, which represented all reported COVID-19 cases with at least one positive reverse transcription-polymerase chain reaction test of a respiratory swab. The raw data set was filtered to include only resolved cases (i.e. cases with an entry under the attribute *RemovalType*), as unlabelled cases are still active. *RemovalType* was defined as the outcome for the patient and was coded as either "RECOVERED" or "DIED." Descriptive statistics, i.e. means, standard deviations, case fatality rates (CFR), t tests (for continuous variables) and Pearson's χ^2 tests (for nominal variables) were generated with StataCorp 2013 (Stata Statistical Software, Release 13; College Station, TX).

$$\text{CFR (\%)} = \frac{\textit{number of reported COVID-19 deaths}}{\textit{number of reported COVID-19 cases}} \text{ x } 100$$

We conducted an exploratory analysis to screen cases and attributes in the raw data set. The attribute *AgeGroup* was recoded to reclassify *Age* (age of patient, in years) into nine ranges according to the classification of the United States Centers for Disease and Control and Prevention.[20] *Pregnanttab* was defined as a binary variable representing whether a patient was pregnant at any time during COVID-19 infection, with male cases coded as missing values. Missing values for *CityMunRes* (patient's city of residence) and *ProvRes* (patient's province of residence) were recoded as "Repatriate" for all cases with *RegionRes* (patient's region of residence) = "Repatriate". The data set was then filtered to select only cases with no other missing values to generate the final data set. Details of the data pre-processing can be found in **Supplementary Information A**.

Attribute selection, random undersampling, hyperparameter optimizations, model generation, cross-validation and performance calculations were done in RapidMiner Studio 9.7.002 (rev. db1bb6, platform: WIN64) (see **Supplementary Information C**). The attribute *RemovalType* was labelled as the outcome in the data set. Attributes were selected with feature weights operators (*weightbyGiniIndex*, *weightbyInformationGain*, *weightbyInformationGainRatio*) to determine those appropriate for model generation. The subprocess *optimizeParameters(Grid)* was used to perform grid optimization of the hyperparameters for the decision tree operator *decisionTree* and the threshold operator *createThreshold*. The subprocess ran a fivefold cross-validation operator to train and validate the data set with the decision tree model and the optimized *decisionTree* and *createThreshold* hyperparameters generated for each fold. Random undersampling was done only on

the training data set for each fold in the cross-validation operator, with the sample operator to (i) select all cases with *RemovalType* = DIED and (ii) randomly select cases with *RemovalType* = RECOVERED using stratified sampling to achieve a 1:1 RECOVERED:DIED ratio. Stratified sampling generated two subsets from the modelling data set that ensured similar *RemovalType* case distribution (i.e. RECOVERED and DIED) between the two subsets by simple random sampling. All cases in the testing data set were used to validate the model for each fold in the cross-validation.

The decision tree model generated by the cross-validation training data set was also extracted. Performance metrics such as area under the curve (AUC), accuracy, F-score, sensitivity and specificity were extracted from the cross-validation with the positive class set as *RemovalType* = DIED. Similar cross-validation operators were used to train and validate a naïve Bayes model for comparison. Details of the model generation can be found in **Supplementary Information B**. The study adhered to the TRIPOD checklist for prediction model development.[21]

Ethics statement

The study was reviewed and approved on 19 August 2020 by the San Beda University Research Ethics Board under the study protocol code SBU-REB 2020–017. The study adhered to the TRIPOD checklist for prediction model development.

RESULTS

Description of cases

The final data set was a list of 132 939 resolved COVID-19 cases (98.16% of all resolved cases and 67.43% of total reported cases from the raw data set). Of the reported cases, 97.7% recovered and 2.3% died. There were more COVID-19 cases among males than females (145.67 per 100 000 vs 118.10 per 100 000; $P < 0.001$). CFRs were also higher among males than females (2.46% vs 1.97%; $P < 0.001$). The most resolved cases were among people aged 18–29 years. Cases aged ≥85 years had the highest CFR (22.57%), followed by those aged 75–84 years (17.99%) and 65–74 years (12.01%). The age group 18–29 years had the lowest CFR, at 0.27% (**Table 1**). Disaggregation of male and female cases showed similar patterns of cases by age group (**Table 2**).

The highest notification rates of COVID-19 cases were from the National Capital Region, followed by Regions VII and IV-A (590.20, 285.70, 121.08 per 100 000, respectively). The highest CFR was observed in Region VII (5.83%), followed by Regions I (4.09%) and IX (4.00%). The lowest CFR was observed among repatriates (0.23%), followed by Regions VIII (0.61%) and II (0.62%). Although the National Capital Region and Region IV-A had the most cases, they had low case fatality rates (1.88% and 1.45%, respectively) (**Table 1**).

Outcomes from machine learning models

Of the three feature weighting operators, only the attributes Age and Admitted were included in the final model. The decision tree model was trained and cross-validated with the following optimized hyperparameters: criterion = information_gain, maximal_depth = 8, minimal_gain = 0.0, minimal_leaf_size = 10, minimal_size_for_split = 100. The comparator naïve Bayes model used the same optimized hyperparameters in model training and cross-validation and had a higher AUC (0.881 ± 0.006), accuracy (81.68% ± 0.05%), F-score (16.75% ± 0.33%) and specificity (81.71% ± 0.52%) and a better receiver operating characteristic (ROC) curve. The decision tree model had greater sensitivity (81.65% ± 1.64%) (**Table 3**; **Fig. 1**).

The decision tree had seven levels, with each node splitting into two branches or leaves (**Fig. 2**; **Supplementary Information B5** provides other details of the decision tree, including the actual number of cases and outcomes per leaf). The primary (root) node was Age, with the split criterion being a cut-off of 51.5 years, based on the average of the split criterion of the values Age = 51 and Age = 52. The attribute Admitted split the lower branches further, with further splits according to Age. The majority of all cases in the model (53.54%) were <51.5 years and had no history of hospital admission, and most recovered (85.46%). Similarly, the majority of cases aged 51.5–57.5 years with no history of hospital admission recovered (55.14%). There were increasing proportions of deaths with increasing age, with the highest death rates among those >63.5 years (81.98%). A high proportion (93.33%) of people aged >51.5 years with a history of hospital admission died.

Table 1. **Demographic characteristics of resolved cases (recovered or died) from the Philippines COVID Data Drop from 25 August 2020**

	Recovered	Died	CFR (%)	$P < 0.001$
Sex, n = 135 434				
Male	73 919	1863	2.46	
Female	58 477	1175	1.97	
Age, n = 133 097				< 0.001
Mean age (years)	38.05 (± 15.93)	61.33 (± 16.73)		
Age group (years), n = 133 097				< 0.001
0–4	1830	32	1.72	
5–17	5563	26	0.47	
18–29	37 080	100	0.27	
30–39	32 632	147	0.45	
40–49	22 315	294	1.30	
50–64	21 907	995	4.34	
65–74	6148	839	12.01	
75–84	2092	459	17.99	
≥ 85	494	144	22.57	
Region, n = 131 614				< 0.001
BARMM	455	11	2.36	
CAR	370	8	2.12	
CARAGA	297	4	1.33	
NCR	74 572	1430	1.88	
Repatriate	6586	15	0.23	
Region I: Ilocos Region	609	26	4.09	
Region II: Cagayan Valley	483	3	0.62	
Region III: Central Luzon	3850	81	2.06	
Region IV-A: CALABARZON	17 201	253	1.45	
Region IV-B: MIMAROPA	396	7	1.74	
Region V: Bicol Region	773	22	2.77	
Region VI: Western Visayas	1865	48	2.51	
Region VII: Central Visayas	16 256	1006	5.83	
Region VIII: Eastern Visayas	1302	8	0.61	
Region IX: Zamboanga Peninsula	889	37	4.00	
Region X: Northern Mindanao	766	16	2.05	
Region XI: Davao Region	1480	56	3.65	
Region XII: SOCCSKSARGEN	429	4	0.92	

BARMM: Bangsamoro Autonomous Region in Muslim Mindanao; CAR: Cordillera Administrative Region; CARAGA: Caraga Administrative Region; NCR: National Capital Region; CALABARZON: Batangas, Cavite, Laguna, Quezon, Rizal and Lucena; MIMAROPA: Mindoro, Marinduque, Romblon and Palawan; SOCCSKSARGEN: South Cotabato, Cotabato, Sultan Kudarat, Sarangani and General Santos.

Table 2. **Resolved cases (recovered or died) from the Philippines COVID Data Drop from 25 August 2020 by age group and sex (n = 133 097)**

Age group (years)	Males (n = 74 395)			Females (n = 58 702)			CFR ratio[a]
	Recovered	Died	CFR (%)	Recovered	Died	CFR (%)	
0–4	978	18	1.81	852	14	1.62	1.12
5–17	2848	12	0.42	2715	14	0.51	0.82
18–29	19 967	61	0.30	17 113	39	0.23	1.30
30–39	18 995	97	0.51	13 637	50	0.37	1.38
40–49	13 397	191	1.41	8918	103	1.14	1.24
50–64	12 001	647	5.12	9906	348	3.39	1.51
65–74	3157	506	13.81	2991	333	10.02	1.38
75–84	1012	266	20.81	1080	193	15.16	1.37
≥ 85	178	64	26.45	316	80	20.20	1.31

[a] CFR ratio is computed as $\frac{CFR\ males}{CFR\ females}$

Table 3. **Performance metrics for the two machine learning models: decision tree and naïve Bayes using the modelling data set and optimized hyperparameters**

Model	AUC		Accuracy		F-score		Sensitivity		Specificity	
Decision tree	0.876	± 0.010	81.42%	± 1.01%	16.74%	± 0.55%	81.65%[a]	± 1.64%	81.41%	± 1.07%
Naïve Bayes	0.881[a]	± 0.006	81.68%[a]	± 0.05%	16.75%[a]	± 0.33%	80.63%	± 1.17%	81.71%[a]	± 0.52%

[a] Highest values for each metric across all models

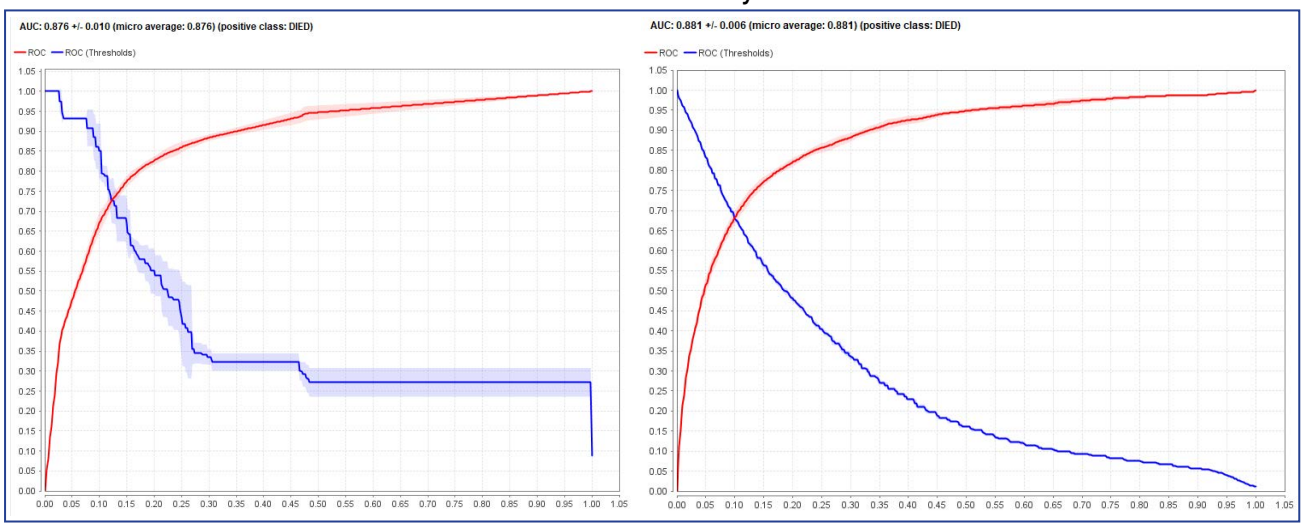
DISCUSSION

Using a decision tree model, we generated a simple seven-level, multinodal decision tree to predict COVID-19-related outcomes of reported cases in the Philippines on the basis of the attributes of age and hospital admission. Tree-based methods in the classification and regression tree paradigm are increasingly used and "have become one of the most flexible, intuitive, and powerful data analytic tools for exploring complex data structures."[22] Modern programming software and applications have enabled the use of machine learning algorithms such as decision trees to process large sets of data[23,24] and are widely used in health care, including as prediction models.[9,25–27] Decision trees are also easy to understand and interpret,[13,28] which is useful for both health workers and policy-makers, and are flexible enough to handle non-parametric class densities such as data from COVID-19 databases.[29]

Our decision tree model indicated age as the main predictor of clinical outcomes for COVID-19. In both the descriptive analysis and the decision tree model, younger cases had higher recovery rates, while older age groups had noticeably higher mortality rates, regardless of hospital admission status. These findings are consistent with the current international literature[20,30–32] as well as locally reported data.[2] The age cut-off of 51.5 years determined in the decision tree model was, however, lower than the current age cut-off used in most Philippine medical[4] or policy guidelines, suggesting that age cut-offs (both lower and upper bounds) for guidelines should be re-evaluated continually. The youngest age group in this study (0–17 years old) had higher case fatality rates than the baseline (18–29 years old), which was consistent even for cases <19 years of age with a history of hospital admission, who had a high death rate. This finding is inconsistent with the available literature but may be due to the relative paucity of confirmed cases and studies in younger COVID-19 cases. Alternatively, it may be due to the fact that age-differentiated studies have been conducted with data from developed countries, such as China, England and Wales, France, the Republic of Korea and Spain,[30,33] and may not be comparable to the situation in the Philippines.

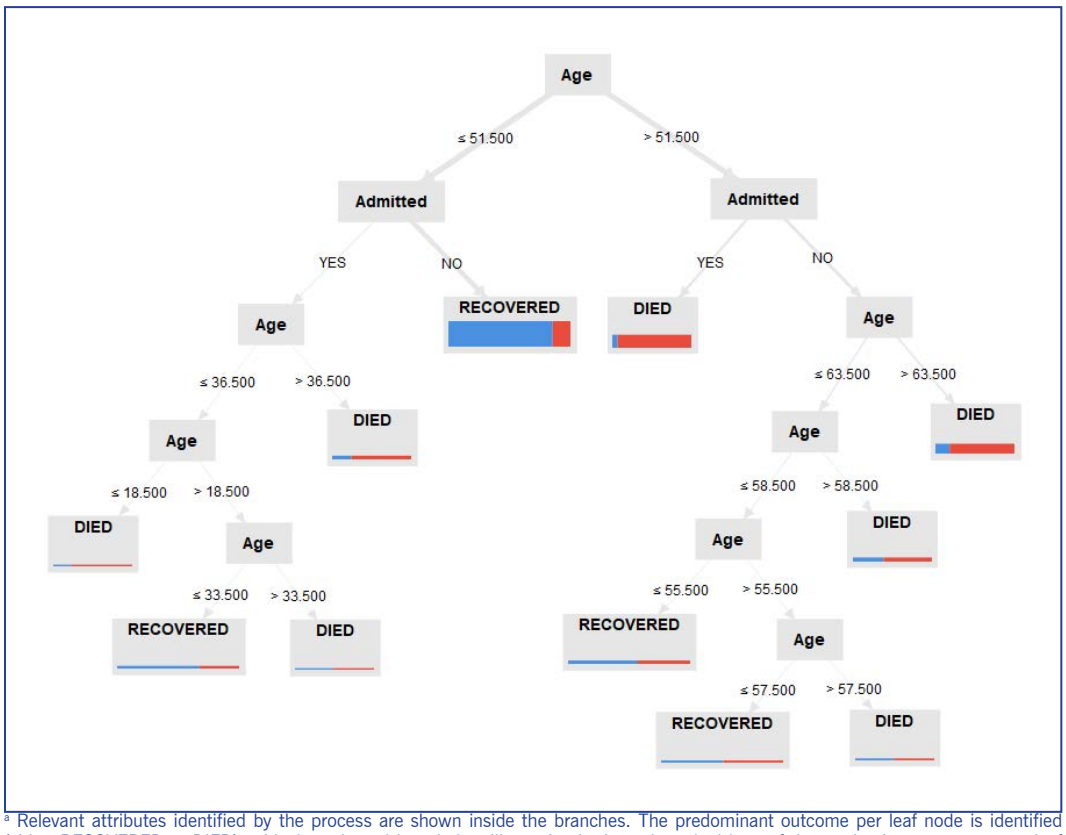
A history of hospital admission was another strong predictor of mortality from COVID-19, especially in cases

Fig. 1. **Receiver operating characteristic (ROC) curves for the two machine learning models: decision tree and naïve Bayes[a]**



[a] The ROC curve plots a model's sensitivity, or true positive rate, versus its false positive rate (one minus the specificity or true negative rate) as its discrimination threshold is varied. Generally, the closer the ROC curve (red curve) is to the top left corner of the graph, the better the model. The shaded regions represent the standard deviations.

Fig. 2. **Decision tree for predicted outcomes of resolved cases (recovered or died) from the Philippines COVID Data Drop from 25 August 2020[a]**



[a] Relevant attributes identified by the process are shown inside the branches. The predominant outcome per leaf node is identified (either RECOVERED or DIED), with the coloured bars below illustrating horizontal stacked bars of the predominant outcome per leaf (RECOVERED=blue, DIED=red). The width of the bars represents the relative number of cases in each leaf as compared with the total cases in the modelling dataset, while the thickness of each arrow illustrates the relative number of cases on each branch as compared with the total cases in the modelling dataset.

>51 years of age. Rationally, this is to be expected, as current national guidelines for hospitalization of COVID-19 patients are for those risk-stratified as moderate, severe or critical.[4] Our study affirms the use of the current guidelines in the Philippine setting.

Although sex was not a predictor in the decision tree model, males had a statistically significantly higher CFR fatality rate than females in the descriptive analysis. This difference in the adult population is consistent with that found in an international study by Bhopal and Bhopal using pooled data from multiple countries, with higher male:female CFRs for age groups ≥40 years (range: 1.65–2.6).[33] Sex differentials in COVID-19 mortality have been extensively documented in multiple studies,[32–34] and our study confirmed these findings in the Philippine setting. Proposed mechanisms for a sex differential include the fact that males generally have more pre-existing comorbidities like hypertension, cardiovascular disease and chronic obstructive pulmonary disease; poorer health behaviours (e.g. smoking and drinking alcoholic beverages); and even biological differences, such as specific receptor regulation, chromosomal variation and differences in interferon and hormone levels.[34]

Despite differences in responses to COVID-19, notification rates and CFRs in different geopolitical classifications (i.e. city/municipality, province or region) were not seen in the model. **Supplementary Information B2** provides details of the feature weights for city/municipality, province and region as attributes. This result suggests that guidelines can be national, albeit with a subgroup-targeted approach, for clinical and public health management, primarily based on disease interaction with age and sex,[1,33–35] and specifically focusing on the increased risk of males and older age groups for death from COVID-19.[33,35]

Our study has two types of limitations: the quality of the data set and model limitations. The quality of the data set was compromised mainly by availability and data points from the raw data set. As the DOH COVID Data Drop is publicly available, measures are in place to protect the privacy and confidentiality of sensitive patient information. Thus, some useful information potentially associated with COVID-19 mortality, such as presence of comorbidities, smoking history, travel history, exposure history to a confirmed COVID-19 patient, clinical signs and symptoms of disease processes and poor laboratory findings,[1,11] were not included in the initial data set and were therefore not included in our model. Furthermore, the DOH COVID Data Drop reported multiple instances of duplicate and missing entries[19] on different dates. Additionally, a potentially relevant attribute, HealthStatus (defined as “Asymptomatic”, “Mild”, “Severe”, “Critical”, “Recovered” or “Died”), was not included in the model, as its values change constantly, with no time stamp. We suggest that future studies use primary COVID-19 data sets that include these parameters for more robust prognosis modelling, such as case investigation forms from the DOH Epidemiology Bureau or from hospital records (e.g. PhilHealth claim form 4).

The model generally had high-performance metrics: AUC, accuracy, sensitivity and specificity were reasonably high; however, its calculated F-score was low due to poor model precision. In classification trees for disease diagnostics and prognostics, high sensitivity is preferred to accuracy,[25] especially in inherently imbalanced data sets such as COVID-19 prognosis databases. We tried to control for this imbalance by undersampling, which is more resistant to overestimation of predictive accuracy than oversampling techniques.[36,37] Other model limitations include the inherent propensity of decision trees for “over-fitting”, which often occurs in highly complex models for relatively simple data, which often capture too much noise from the data set.[25,38] We tried to reduce over-fitting with the following strategies: (i) conducting exploratory data set analysis to remove ambiguous, highly correlated or incompletely filled attributes; (ii) enabling pre-pruning and pruning during model training to limit the complexity of the model; and (iii) running decision tree grid optimization to determine which hyperparameter values would net the highest AUC. Another limitation of the model is sampling bias, in which active cases are excluded from the analysis. This limitation is related to the cross-sectional design of the study and the continuing evolution of COVID-19 in the Philippines. Comparable performance metrics in our study indicate that other classification models, such as naïve Bayes, random forest or deep learning, might be considered for future prognostic models.

In conclusion, our study showed that increasing age and history of hospital admission are important predictors of COVID-19 prognosis, consistent with the current literature. We were able to generate a sensitive, specific decision tree model with a high AUC and with Age and Admission as the main predictors of COVID-19 prognosis using a publicly available data set. We recom-

mend adaptation of our model with more comprehensive primary COVID-19 data sets to create robust COVID-19 prognostic models that could contribute to a review of clinical and public health guidelines.

Acknowledgements

We would like to acknowledge the following people who contributed to the study: TM for overall support and SG and MS of RapidMiner Community for methodology guidance.

Conflicts of interest

The authors have no conflicts of interest to declare.

Funding

A financial grant for the study was provided by the San Beda University Office of Research and Innovation under the study protocol code SBU-REB 2020-017. The manuscript was not developed through a Field Epidemiology Training Program (FETP) and/or during a WPSAR scientific writing workshop.

References

1. Beeching N, Fletcher T, Fowler R. Coronavirus disease 2019 (COVID-19). In: BMJ Best Practice [website]. London: BMJ; 2020. Available from: https://bestpractice.bmj.com/topics/en-gb/3000168, accessed 6 June 2021.
2. Cases information. Covid-19 Tracker Philippines. Manila: Department of Health; 2020. Available from: https://doh.gov.ph/covid19tracker, accessed 23 August 2021.
3. WHO coronavirus disease (COVID-19) dashboard. Geneva: World Health Organization; 2020. Available from: https://covid19.who.int, accessed 23 August 2021.
4. Interim guidance on the clinical management of adult patients with suspected or confirmed COVID-19 infection (version 3.1). Manila: Philippine Society for Microbiology and Infectious Diseases, Inc.; 2020. Available from: https://www.psmid.org/interim-management-guidelines-for-covid-19-version-3-1/, accessed 30 October 2020.
5. Urwin S, Kandola G, Graziadio S. What prognostic clinical risk prediction scores for COVID-19 are currently available for use in the community setting? Oxford: Centre for Evidence-based Medicine; 2020.
6. Why predictive modeling is critical in the fight against COVID-19? (Report No. 8). Washington DC: Pan American Health Organization; 2020. Available from: https://iris.paho.org/bitstream/handle/10665.2/52276/PAHOEIHISCOVID-19200007_eng.pdf?sequence=5&isAllowed=y, accessed 6 June 2021.
7. Deo RC. Machine learning in medicine. Circulation. 2015 Nov 17;132(20):1920–30. doi:10.1161/CIRCULATIONAHA.115.001593 pmid:26572668
8. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009 Feb 23;338;b375. doi:10.1136/bmj.b375 pmid:19237405
9. Venkatasubramaniam A, Wolfson J, Mitchell N, Barnes T, JaKa M, French S. Decision trees in epidemiological research. Emerg Themes Epidemiol. 2017 Sep 20;14(1):11. doi:10.1186/s12982-017-0064-4 pmid:28943885
10. Vaishya R, Javaid M, Khan IH, Haleem A. Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr. 2020 Jul - Aug;14(4):337–9. doi:10.1016/j.dsx.2020.04.012 pmid:32305024
11. Debnath S, Barnaby DP, Coppa K, Makhnevich A, Kim EJ, Chatterjee S et al. Northwell COVID-19 Research Consortium. Machine learning to assist clinical decision-making during the COVID-19 pandemic. Bioelectron Med. 2020 Jul 10;6(1):14. doi:10.1186/s42234-020-00050-8 pmid:32665967
12. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020 Apr 7;369:m1328. doi:10.1136/bmj.m1328 pmid:32265220
13. Yan L, Zhang HT, Goncalves J, Xiao Y, Wang M, Guo Y et al. An interpretable mortality prediction model for COVID-19 patients. Nat Mach Intell. 2020;2(5):283–8. doi:10.1038/s42256-020-0180-7
14. Yadaw AS, Li YC, Bose S, Iyengar R, Bunyavanich S, Pandey G. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Health. 2020 Oct;2(10):e516–25. doi:10.1016/S2589-7500(20)30217-X pmid:32984797
15. Abrigo MRM, Uy J, Haw NJ, Ulep VGT, Francisco-Abrigo K. Projected disease transmission, health system requirements, and macroeconomic impacts of the coronavirus disease 2019 (COVID-19) in the Philippines. Manila: Philippine Institute for Development Studies; 2020. Available from: https://pidswebs.pids.gov.ph/CDN/PUBLICATIONS/pidsdps2015.pdf, accessed 6 June 2021.
16. Bongolan VP, Minoza JMA, de Castro R, Sevilleja JE. Age-stratified infection probabilities combined with a quarantine-modified model for COVID-19 needs assessments: model development study. J Med Internet Res. 2021 May 31;23(5):e19544. doi:10.2196/19544
17. Medina MA. Preliminary estimate of COVID-19 case fatality rate in the Philippines using linear regression analysis (Report No.: ID 3569248). Rochester (NY): Social Science Research Network; 2020. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3569248, accessed 10 June 2020.
18. Alipio M. Do socio-economic indicators associate with COVID-2019 cases? Findings from a Philippine study. Germany: University Library of Munich; 2020. Available from: https://ideas.repec.org/p/pra/mprapa/99583.html, accessed 10 June 2020.
19. Department of Health COVID Data Drop. Manila: Department of Health; 2020. Available from: https://drive.google.com/drive/folders/10VkiUA8x7TS2jkibhSZK1gmWxFM-EoZP, accessed 30 August 2020.
20. COVID-19 hospitalization and death by age. Coronavirus disease 2019 (COVID-19). Atlanta (GA): Centers for Disease Control and Prevention; 2020. Available from: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-age.html, accessed 28 September 2020.
21. Collins G, Reitsma J, Altman D, Moons K. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. Oxford: EQUATOR Network; 2015. Available from: https://www.equator-network.org/reporting-guidelines/tripod-statement/, accessed 25 May 2020.

22. Banerjee M, Reynolds E, Andersson HB, Nallamothu BK. Tree-based analysis. Circ Cardiovasc Qual Outcomes. 2019 May;12(5):e004879. doi:10.1161/CIRCOUTCOMES.118.004879 pmid:31043064

23. Sharma A. Decision tree split methods | Decision tree machine learning. Gurgaon: Analytics Vidhya; 2020. Available from: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/, accessed 26 October 2020.

24. Khanday AMUD, Rabani ST, Khan QR, Rouf N, Mohi Ud Din M. Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol. 2020 Jun 30;12(3):1–9. doi:10.1007/s41870-020-00495-9 pmid:32838125

25. Serrano L. Grokking machine learning. MEAP edition (version 13). Manning Publications; 2021.

26. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019 Apr 4;380(14):1347–58. doi:10.1056/NEJMra1814259 pmid:30943338

27. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol. 2019 Mar 19;19(1):64. doi:10.1186/s12874-019-0681-4 pmid:30890124

28. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015 Apr 25;27(2):130–5. PMID:26120265

29. Sánchez-Montañés M, Rodríguez-Belenguer P, Serrano-López AJ, Soria-Olivas E, Alakhdar-Mohmara Y. Machine learning for mortality analysis in patients with COVID-19. Int J Environ Res Public Health. 2020 Nov 12;17(22):8386. doi:10.3390/ijerph17228386 pmid:33198392

30. Liu Y, Mao B, Liang S, Yang J, Lu H, Chai Y et al. Association between ages and clinical characteristics and outcomes of coronavirus disease 2019. Eur Respir J. 2020;55(5):2001112. doi:10.1183/13993003.01112-2020 pmid:32312864

31. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis. 2020 Jun;20(6):669–77. doi:10.1016/S1473-3099(20)30243-7 pmid:32240634

32. Jin JM, Bai P, He W, Wu F, Liu XF, Han DM et al. Gender differences in patients with COVID-19: Focus on severity and mortality. Front Public Health. 2020 Apr 29;8:152. doi:10.3389/fpubh.2020.00152 pmid:32411652

33. Bhopal SS, Bhopal R. Sex differential in COVID-19 mortality varies markedly by age. Lancet. 2020 Aug 22;396(10250):532–3. doi:10.1016/S0140-6736(20)31748-7 pmid:32798449

34. Haitao T, Vermunt JV, Abeykoon J, Ghamrawi R, Gunaratne M, Jayachandran M et al. COVID-19 and sex differences: Mechanisms and biomarkers. Mayo Clin Proc. 2020 Oct;95(10):2189–203. doi:10.1016/j.mayocp.2020.07.024 pmid:33012349

35. Griffith DM, Sharma G, Holliday CS, Enyia OK, Valliere M, Semlow AR et al. Men and COVID-19: A biopsychosocial approach to understanding sex differences in mortality and recommendations for practice and policy interventions. Prev Chronic Dis. 2020 Jul 16;17:E63. doi:10.5888/pcd17.200247 pmid:32678061

36. Blagus R, Lusa L. Joint use of over- and under-sampling techniques and cross-validation for the development and assessment of prediction models. BMC Bioinformatics. 2015 Nov 4;16(1):363. doi:10.1186/s12859-015-0784-9 pmid:26537827

37. Santos MS, Soares JP, Abreu PH, Araujo H, Santos J. Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research frontier]. IEEE Comput Intell Mag. 2018;13(4):59–76. doi:10.1109/MCI.2018.2866730

38. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Bossuyt P et al. Topic Group 'Evaluating diagnostic tests and prediction models' of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019 Dec 16;17(1):230. doi:10.1186/s12916-019-1466-7 pmid:31842878