Predictive analysis: Factors associated with mortality risk in confirmed COVID-19 patients

Student Name: Keyun Zhou

Student ID: 501147287

Supervisor: Dr. Zekiye Erdem

Date of Submission: October 24, 2022

Abstract

The coronavirus (COVID-19) is a pathogenic viral infection that has been negatively affecting human health on a global scale. Despite the effort to contain and delay the spread of the disease, the number of infected individuals continued to rise. With an increase in patient volume, hospital beds, medical equipment, and healthcare staffing became increasingly in demand. In addition, the lack of adequate response and immediate control measures has caused further stress with increased disease exposure and poor health outcomes.

COVID-19 is known to spread through respiratory droplets or direct contact with an infected individual, such as through saliva or sneezing, attacking the respiratory system of an individual (Han et al., 2020). However, the range of symptoms or the degree of severity after the contraction of COVID-19 varies from case to case. The underlying cause of such variation among individuals can be complex. Data showed that some population groups could be at greater risk than others, depending on several factors. For example, the study by Ma et al., 2020 suggested that more severe cases of COVID-19 were observed in males (59.7%) and patients of older age (54.5 vs 44.5 years of age) (Ma et al., 2020). In the current project, a predictive analytics theme will be used to examine the effect of demographic and clinical factors of patients on the mortality risk of COVID-19 patients, to facilitate medical decision-making, response planning, and mitigate stress on the healthcare system. This will help to prioritize and target patients to increase efficiency in patient care, and to raise public awareness of COVID-19 prevention targeting specific sub-populations. The questions under investigation are: what types of factors predict COVID-19 mortality best (demographic or clinic data), which factor best predicts mortality risk of COVID-19, which sub-populations of patients are more susceptible to COVID-19, and lastly the best machine learning algorithm predicting COVID-19 mortality risk. The dataset is obtained from the

Open Data Catalogue, by the Toronto Public Health (Toronto Government), providing confirmed cases of COVID-19 from 2020 to 2022. The project will use data-driven techniques to develop decision-making models by machine learning algorithms, specifically, logistic regression, decision trees, and XGBoost. The programs that will be used are Python and R.

COVID Data set link: COVID-19 Cases in Toronto - City of Toronto Open Data Portal

Census Data set link: Neighbourhood Profiles - Dataset - CKAN (prod-toronto.ca)

GitHub Repository: keyunzhou2/CIND820 (github.com)

Introduction

With the spread of the coronavirus disease in 2019 (COVID-19), it had caught many countries off-guard, causing great distress in the lives of many. Given the novelty of the disease, rapid mutation, and the lack of medical knowledge, it had resulted in more than 6 million deaths worldwide (WHO, 2022). During the many pandemic waves, hospital and intensive care units have long been overwhelmed and resources were exhausted. Healthcare organizations and professionals were unable to make accurate clinical decisions about resource allocation and tailoring individual treatment strategies.

COVID-19 is a virus that attacks the respiratory tract and its symptoms can vary on an individual basis (Mohapatra et al., 2020). While some patients experience little to no symptoms, some patients could have more severe onsets and eventual death (Alimohamadi et al., 2020; Han et al., 2020; Mohapatra et al., 2020). With its highly contagious characteristics, overwhelmed hospital capacity could lead to a further increase in the death rate with the lack of medical attention. As such, medical decision-making techniques could be vital in alleviating the stress of hospitals with limited capacity and time constrain. Prediction of the mortality risk could be crucial in helping the healthcare field in prioritizing care for individuals that are more likely to have severe conditions and to make unbiased risk assessments. The development of a predictive method would facilitate resource planning, reduce overcrowding, guide healthcare actions and mitigate the current healthcare burdens.

Among the predictive methods, artificial intelligence and machine learning algorithms are effective in predicting medical conditions and decision-making (Davenport, 2019; May, 2021). Predictive analytic algorithms based on machine learning can be used to predict the mortality risk

of patients with COVID-19 based on patient characteristics such as demographic information, medical histories, and symptoms. Logistic regressions are commonly used for predicting event probabilities where a linear effect is often assumed (Sperandei, 2014). This may be restrictive in some cases where non-linear relationship, collinearity, or complex interactive effect exists. To overcome complications, tree-based or ensemble-based algorithms such as XGBoost are often used for non-linear decisions. However, they are prone to over-fitting and limit generalizability (Feng et al., 2021).

In the current study, three machine learning algorithms will be used: logistic regression, decision tree, and XGBoost for predicting COVID-19 mortality risk.

The study has three aims:

1. Predictive model to guide healthcare professionals to prioritize patient care.
2. Types of factors and strongest predictor for COVID-19 mortality (demographic or clinic data),
3. To compare the three algorithms for the best performing algorithm for COVID-19 mortality risk, where the model accuracy will be evaluated with the receiver operating characteristic (ROC) curve.

The study would also determine whether there are health disparities in COVID-19. The personalized machine-learning predictive COVID-19 mortality model will benefit the medical field to improve the decision-making process.

Literature Reviews

Feng et al., (2021) Examined the mortality risk for COVID-19 using confirmed cases in Toronto, Canada, in 2020. The researchers used five techniques: random forest, extreme gradient boosting, logistic regression, generalized additive model, and linear discriminant analysis to

examine mortality patterns. The study was conducted on a sample of 49,216 positive cases while around 4% died from COVID-19. The researchers used patient data on demographic information such as age, and gender and combined it with hospitalization information to conduct predictive analyses. With the different methods, the researchers found that extreme gradient boosting was the best-performing model with the highest AUC score (AUC: 0.96) and Brier's score (0.025) for predicting COVID-19 mortality. Among the predictors, the researchers found age to be the strongest predictor under the XGBoost method, while gender had the least effect.

Ustebay et al., (2022) demonstrated that machine-learning techniques could be used in the prognosis of COVID-19. In the study, eight machine learning algorithms: support vector machines, logistic regression, random forest, XGBoost, multilayer perceptron, extra trees, CatBoost, and k-nearest neighbors' classifiers were used to predict patient needs for intensive care, intubation, and mortality risks. The study used confirmed positive cases on 13,351 patients using the information on demographics, clinic data, and blood tests. The researchers found that tree-based classifiers including XGBoost (Highest AUROC: 0.90), CatBoost (0.96), and extra trees (0.99) had higher AUROC for predicting COVID-19 than other algorithms. Ustebay et al., (2022) also found that lymphocyte counts are the most critical feature for predicting mortality for COVID patients.

Ottenhoff et al., (2021) conducted a retrospective study on predicting the mortality risk of COVID-19 patients in the Netherlands. The study used linear logistic regression and a non-linear tree-based gradient-boosting algorithm for predictive analysis on 2273 patients. The models were then compared with the age-based decision rule held in the Netherlands. Both machine learning algorithms showed better performance than age-based decisions (AUC: 0.57-0.74), while XBG had a better overall AUC score (0.79-0.85) than logistic regression (0.77-0.85). Age was identified as the most predictive feature in feature selection and SHAP analysis.

Jamshidi et al., (2021) conducted a retrospective study on 797 patients using five machine learning algorithms: random forest, logistic regression, gradient boosting classifier, support vector machine classifier, and artificial network algorithm to predict the mortality risk of COVID-19. The researchers have used patient demographic, past medical backgrounds, and laboratory biomarkers as predictor factors. A further 10-fold cross-validation was used for feature selection. The random forest was the best-performing machine learning algorithm (AUC: 0.79) in predicting the mortality outcome of COVID-19 among others. It was suggested that random forest outperformed other models through its capacity for non-linear correlations. The best predictor was hypoalbuminemia and renal functions, most were related to biomarkers and past medical history, however, demographic information such as age and gender was also shown to have an effect in influencing mortality risk.

Fernandes et al., (2021) studied a machine-learning approach to predict patients that are more likely to develop critical conditions. A total of 1040 patients in Brazil with confirmed COVID-19 cases were studied with demographic, laboratory, and clinical information collected. Five machine learning algorithms were applied: artificial neural networks, extra trees, random forest, CatBoost, and extreme gradient boosting. The researchers suggested that all models performed equally above the AUROC of 0.91 in the test set. It was further suggested that age was an important predictor responsible for most negative health outcomes.

Pourhomayoun & Shakibi (2022) aimed to use machine learning to study the mortality risk in COVID-19 patients. The research used six algorithms: artificial neural networks, random forest, decision tree, logistic regression, and K-Nearest neighbor to predict mortality rate. The research used more than 2,670,000 observations ranging from 146 countries for data analysis, where the predictor factors included symptoms, past medical history, and demographic information. Features

selection was done through a series of the wrapper and filter methods including correlation coefficient, Fisher score, and chi-square parameter. The neural network algorithm showed the best performance and accuracy among other models with 10-fold cross-validation of 89.98%.

Dataset & Methods

The dataset was obtained through the City of Toronto Open Data Portal, which contained COVID-19 cases from January 2020 to October 2022. The dataset contained basic demographics along with brief hospital-use status, a total of 17 variables in total. The variables were a mixture of categorical and continuous data: Assigned patient ID, Outbreak Associated (2 levels: Sporadic and Outbreak associated in healthcare institutions and other settings), Age Group (9 levels: from 19 and younger to 90 and older), Neighbourhood Names (140 distinct neighborhoods in Toronto), FSA (first three characters of postal code), Source of Infection (7 levels related to how patients acquired their COVID-19 infection), Classification (2 levels, whether the cases were probable or confirmed), Episode Date (best estimate of when the disease was obtained), Reported Date (the date that the case was reported to Toronto Public Health), Client Gender (A self-reported gender of patients), Outcome (3 levels of case outcomes), Binary variables with yes or no: Currently Hospitalized, Currently in ICU, Currently Intubated, Ever Hospitalized, Ever in ICU and Ever Intubated. Cases that have been reported as “Probable”, “Active” or are currently in the hospital were not included in the data analysis, and only confirmed COVID-19 cases were used. A descriptive analysis of the used variable could be found in Table 1. Another set of data was obtained through the City of Toronto Neighbourhood Profiles 2016 which contained general demographic information sorted by neighborhood names. Within the Neighbourhood Profiles 2016, two variables were joined with the COVID data to provide further demographic information on patients. The two continuous variables were household income after tax and population density

per kilometer squared of each Toronto neighborhood. A descriptive analysis of the two variables from the Neighborhood Profile was presented in Table 2. The overall approach to data preparation, analysis, and machine learning followed the tentative methodology in Figure 1.

Step 1: Data preprocessing, join dataset

The dataset was exported as a CSV file. The type and characteristics of each variable were examined through R Studio and Python. For the project dataset, the two datasets were joined using neighborhood names. In each categorical variable, the format of each case entry was verified and standardized to avoid data misinterpretation. In the current study, all non-confirmed or active COVID cases, and patients with missing variables were excluded from the study (n = 204227). The date variables were modified for system reading and only contain the month of the acquired disease. For skewed variables, the log-transformed value was used to standardize data. After data cleaning processing, a total of 183,045 cases were included in the final dataset.

Step 2. Data exploratory analysis

For descriptive analyses, the frequency and characteristics of each variable were sorted by case outcome (Fatal or Resolved) and its variable levels in R Studio (Tables 1-2). The visual interpretations were also sorted according to fatal outcome only to see the association between each variable and its relation and trend to patient death (Figures 2-7).

Descriptive Analysis

Categoric Variable	Levels	Total n = 183,045	%	Resolved n=179,560 (98%)	%	Fatal n=3485 (2%)	%
Outbreak Associated	Outbreak Associated	40370	22.05	38257	21.31	2113	60.63
	Sporadic	142675	77.95	141303	78.69	1372	39.37
Age Group	19 and younger	30355	16.58	30353	16.90	2	0.06
	20-29	31834	17.39	31828	17.73	6	0.17
	30-39	30090	16.44	30070	16.75	20	0.57
	40-49	25700	14.04	25655	14.29	45	1.29
	50-59	24238	13.24	24088	13.42	150	4.30
	60-69	15617	8.53	15236	8.49	381	10.93
	70-79	9274	5.07	8590	4.78	683	19.60
	80-89	9480	5.18	8275	4.61	1205	34.58
	90 and older	6457	3.53	5465	3.04	992	28.46
Source of Infection	close contact	19347	10.57	19205	10.70	142	4.07
	community	76729	41.92	75776	42.20	953	27.35
	household contact	41163	22.49	40936	22.80	227	6.51
	outbreak (congregated settings)	4658	2.54	4627	2.58	31	0.89
	outbreak (healthcare institute)	25771	14.08	23720	13.21	2051	58.85
	outbreak (other settings)	10976	6.00	10928	6.09	48	1.38
	Travel	4401	2.40	4368	2.43	33	0.95
Gender	female	97846	53.45	96174	53.56	1672	47.98
	male	85103	46.49	83290	46.39	1813	52.02
	non-binary	60	0.03	60	0.03	0	0.00
	trans man	7	0.00	7	0.00	0	0.00
	trans woman	12	0.01	12	0.01	0	0.00
	transgender	17	0.01	17	0.01	0	0.00
Month	January	22789	12.45	22317	12.43	472	13.54
	February	10513	5.74	10370	5.78	143	4.10
	March	25895	14.15	25584	14.25	311	8.92
	April	40632	22.20	39581	22.04	1051	30.16
	May	22304	12.18	21917	12.21	387	11.10
	June	5101	2.79	5010	2.79	91	2.61
	July	5318	2.91	5245	2.92	73	2.09
	August	6884	3.76	6834	3.81	50	1.43

	September	8736	4.77	8659	4.82	77	2.21
	October	5611	3.07	5476	3.05	135	3.87
	November	10090	5.51	9820	5.47	270	7.75
	December	19172	10.47	18747	10.44	425	12.20
Ever Hospitalized	yes	10294	5.62	8128	4.53	2166	62.15
	no	172751	94.38	171432	95.47	1319	37.85
Ever in ICU	yes	1886	1.03	178443	99.38	769	22.07
	no	181159	98.97	1117	0.62	2716	77.93
Ever Intubated	yes	1112	0.61	541	0.30	571	16.38
	no	181933	99.39	179019	99.70	2914	83.62

Table 1. Descriptive analysis of used variables in Toronto COVID-19 2020-2022 Dataset.

Numeric Variable	Population Density per km sq	Resolved	Fatal	Average after-tax income of households in 2016	Resolved	Fatal
n	183,045	179,560	3485	183,045	179,560	3485
Mean	6027	6033	5689	373647	374007	355085
SD	4939.72	4939.55	4937.297	205361.4	205597.6	191914.3
Median	4915	4915	4345	332776	332776	302358
Q1	3342	3342	3130	219209	219209	215135
Q3	7291	7291	6582	496958	496958	441052
Min	1040	1040	1040	102259	102259	102259
Max	44321	44321	44321	1413132	1413132	1413132

Table 2. Descriptive analysis of used variables in Toronto Neighbourhood Profile 2016 Dataset.

Step 3. Data preparation for machine learning algorithms

In the current study, three machine learning (ML) algorithms: logistic regression, decision tree, and XGBoost were applied to predict fatal outcomes in COVID-19 cases using the final dataset. The three ML algorithms were modeled and compared through ROC curves (AUC score), accuracy, recall, precision, and training time using R Studio and Python. To prepare categorical data for machine algorithms, dummy variables were used for each level within the variables (values of 0, 1). The class variable, outcome (0: fatal cases, 1: resolved cases), was also demonstrated to be imbalanced: 179,560 resolved cases and 3485 fatal cases. SMOTE technique (k=5) was used to balance data and see possible improvement in predicting fatal COVID-19 cases. Feature selection through the chi-square method (k=18) was also used to see whether the selection of key features would help with algorithm learning.

Step 4. Applying machine learning algorithms: logistic regression, decision tree, XGBoost

Logistic regression fits the data linearly and it was often used as a classification method for binary outcomes. In the current project, the binary outcome would be represented as the fatal or resolved COVID-19 outcomes. The regression can be expressed as $\text{logit}(\pi_i) = X_i\beta$, where π_i denotes the probability of a fatal outcome along with X_i of all variables with the coefficient β (Feng et al., 2021). A decision tree or classification tree is also another population method where it uses a binary tree graph. Each partition of the sample begins with a selected feature, and the selection is determined through the impurity score. The variable with the greatest reduction in impurity is selected first, and the partition continues until instructed criteria are met (Breiman et al., 1984). XGBoost is an ensemble machine-learning method operating under a gradient-boosting framework. It iteratively refits weak prediction models and combines them into one strong learner (Friedman and Popescu, 2003). XGBoost was suggested to improve the decision tree accuracy by decreasing

over-fitting problems. In the current project, each machine learning algorithm was applied under three scenarios: imbalanced data, balanced data, and data after feature selection. The effectiveness (Accuracy, precision, recall, AUC score), efficiency (time taken for the model to be trained), and stability (whether each model generates stable results) were tested and compared.

Results

Visual data interpretations

Based on visual analyses, there seemed to be an increase in patient death with the increase in patient age (Figure 2). The age group of 80 to 89 years (n = 1205, 34.6%) had the most fatal count among other age groups. By examining outbreak types and sources of infection, patients that acquired COVID-19 in compact settings such as healthcare institutes or communities had a higher chance of fatal outcomes than in other settings (Figure 3). For genders, the frequency of fatal cases was relatively equal for males and females, while there was insufficient data for other gender categories. With respect to episode months, an increasing trend of fatal cases was observed around August to April, with a peak in death count during April (Figure 5). The trend observed in episode month could have suggestive links with temperature. Through examining the medical history of the COVID-19 patients, a higher mortality count was observed in patients that have been hospitalized before (62% have been hospitalized in the past vs 38% not been hospitalized; Figure 6). For ICU and Intubation status, patients that have not had previous medical experience seemed to have higher fatal counts (Figure 6). Based on the neighborhood profiles, patients with fatal COVID-19 outcomes seemed to have lower household incomes than resolved cases, while density suggested the opposite trend with higher resolved cases for patients living in dense-populated communities.

Results for machine learning algorithms

Imbalanced data

The final data contained 10 variables. The sample was split into a training set (70% of data) and a test set (30%). In logistic regression, most variables were significant in predicting the COVID-19 outcome, with the exception of patient gender (Figure 8). In the decision tree algorithm, the parent node or strongest predictor was patient medical history relating to past hospitalization; while the second node was patients aged 70 and older (Figure 10). Using the XGBoost method, the past hospitalization status was also used as the parent node (from trees 0 – 2; Figure 11). The accuracy was roughly the same among the three algorithms while logistic regression had the highest precision rate as well as AUC score. Decision tree and XGBoost had similar recall rates which outperformed logistic regression (Table 3). Notably, all models had low precision, recall and AUC scores.

ML Algorithm	Accuracy	Precision	Recall	AUC score
Logistic Regression	0.98	0.20	0.54	0.61
Decision Tree	0.98	0.11	0.58	0.54
XGBoost	0.98	0.15	0.58	0.59

Table 3. Comparison of logistic regression, decision tree, XGBoost under imbalanced Toronto COVID-19 2020-2022 Dataset with all features.

Balanced data with all features

After balancing using SMOTE, the fatal cases were synthetically oversampled to match the resolved cases and modeled with the three ML algorithms. Each model was run three times

Before SMOTE:

```
1    179560
0      3485
```

After SMOTE:

```
1    125692
0    125692
```

to compare effectiveness, efficiency, and stability. Through observing the output of logistic regression, gender was not statistically significant in predicting the outcome of COVID-19 cases ($p>0.05$), while other predictors were significant ($p<0.001$) (Figure 13). In both tree graphs of the decision tree and XGBoost, hospitalization history was used as the strongest predictor related to COVID-19 fatality (Figure 14). With algorithm comparison, logistic regression had the highest accuracy as well as recall rate. Decision tree had the shortest training time among the three algorithms, and XGBoost had the longest training time of roughly 30s. For AUC scores, logistic regression and XGBoost had roughly the same scores as high as 0.97 which outperformed the decision tree. Precision was equivalent among the three ML algorithms (Table 4). All models demonstrated stability through consistent evaluation values and training times through each run. Interestingly, a low kappa score was observed among the three algorithms (Figure 15). Overall, significant precision and recall improvements were observed when compared to the imbalanced dataset.

Logistic Regression	**Accuracy**	**Precision**	**Recall**	**AUC Score**	**Train time**
First Run	0.91	0.998	0.91	0.97	7s
Second Run	0.92	0.998	0.92	0.96	7s
Third Run	0.91	0.998	0.91	0.97	7s

Decision Tree	**Accuracy**	**Precision**	**Recall**	**AUC Score**	**Train time**
First Run	0.86	0.999	0.85	0.94	2s
Second Run	0.86	0.999	0.86	0.93	2s
Third Run	0.86	0.999	0.86	0.94	3s

XGBoost	Accuracy	Precision	Recall	AUC Score	Train time
First Run	0.89	0.999	0.89	0.97	28s
Second Run	0.89	0.999	0.86	0.97	30s
Third Run	0.89	0.999	0.89	0.97	30s

Table 4. Comparison of logistic regression, decision tree, XGBoost under balanced Toronto COVID-19 2020-2022 Dataset with all features.

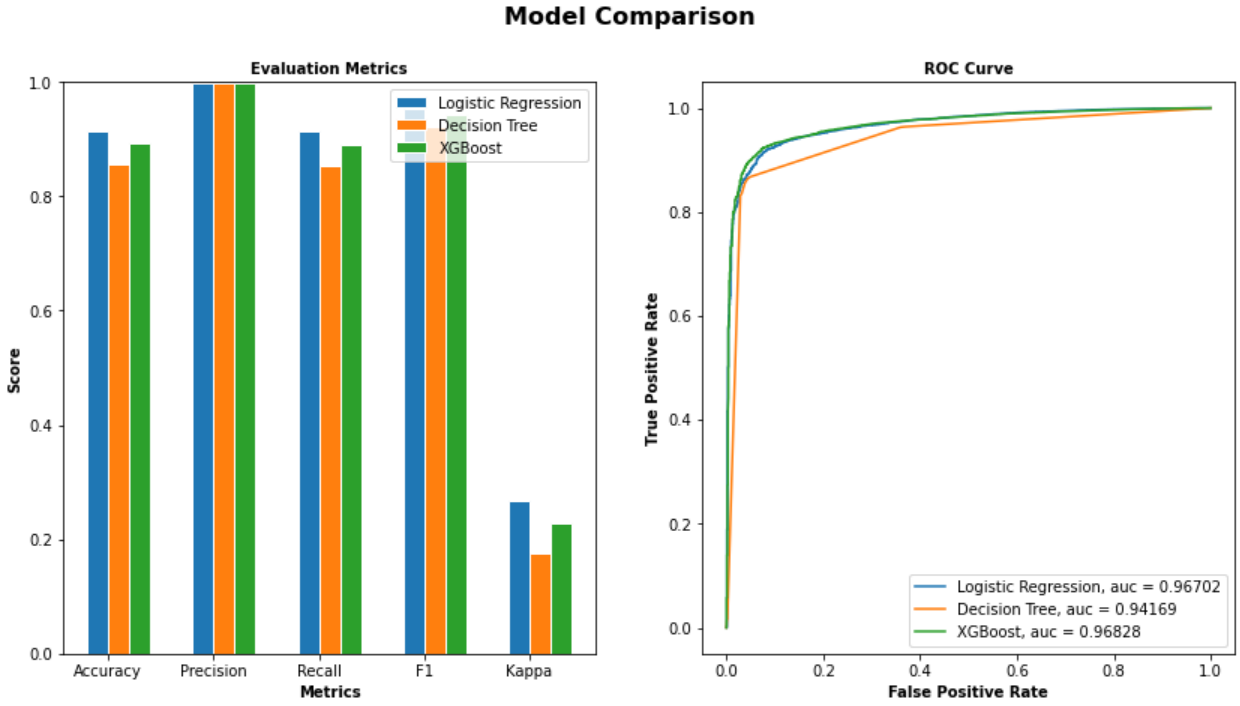

Figure 15. Model comparison for logistic regression, decision tree, XGBoost under balanced Toronto COVID-19 2020-2022 Dataset with all features.

Balanced data with selected features

Feature selection was achieved through the chi-square method (k=18) and 9 variables were selected to see possible improvements in algorithm performance (Figure 16). Among the predictors, patient medical history followed by age was ranked top compared to other variables suggesting strong predictive power. Among the three algorithms, logistic regression and XGBoost had the

highest accuracy (0.92) after feature selection, whereas XGBoost had slightly higher precision. AUC scores and recall were relatively equal among the algorithms, while logistic regression had the shortest training time. Notably, the training time of XGBoost was shortened by about one-third compared to the all-featured dataset. The three ML algorithms demonstrated stability through consistent effectiveness and training times throughout the three runs. After feature selection, the kappa scores drastically increased among the three models with XGBoost having the highest value of 0.84 (Figure 17).

	Score	P_Value
Ever.Hospitalized	20186.313857	0.0
Ever.in.ICU	15257.349183	0.0
Ever.Intubated	14556.374412	0.0
Age.Group.90.and.older	6262.922701	0.0
Age.Group.80.to.89.Years	5928.262025	0.0
Source.of.Infection.Outbreaks,.Healthcare.Institutions	5058.402537	0.0
Age.Group.70.to.79.Years	1486.592608	0.0
Outbreak.Associated	678.279570	0.0
Age.Group.20.to.29.Years	605.679563	0.0

Figure 16. Feature selection using the Chi-squared method.

Logistic Regression	Accuracy	Precision	Recall	AUC Score	Train time
First Run	0.91	0.93	0.91	0.96	2s
Second Run	0.92	0.93	0.91	0.96	1s
Third Run	0.92	0.92	0.91	0.96	1s

Decision Tree	Accuracy	Precision	Recall	AUC Score	Train time
First Run	0.91	0.92	0.89	0.95	2s
Second Run	0.91	0.92	0.89	0.95	2s

Third Run	0.90	0.91	0.89	0.95	3s

XGBoost	Accuracy	Precision	Recall	AUC Score	Train time
First Run	0.92	0.96	0.88	0.96	11s
Second Run	0.92	0.96	0.88	0.96	11s
Third Run	0. 92	0.95	0.88	0.96	11s

Table 5. Comparison of logistic regression, decision tree, XGBoost under balanced Toronto COVID-19 2020-2022 Dataset with selected features.

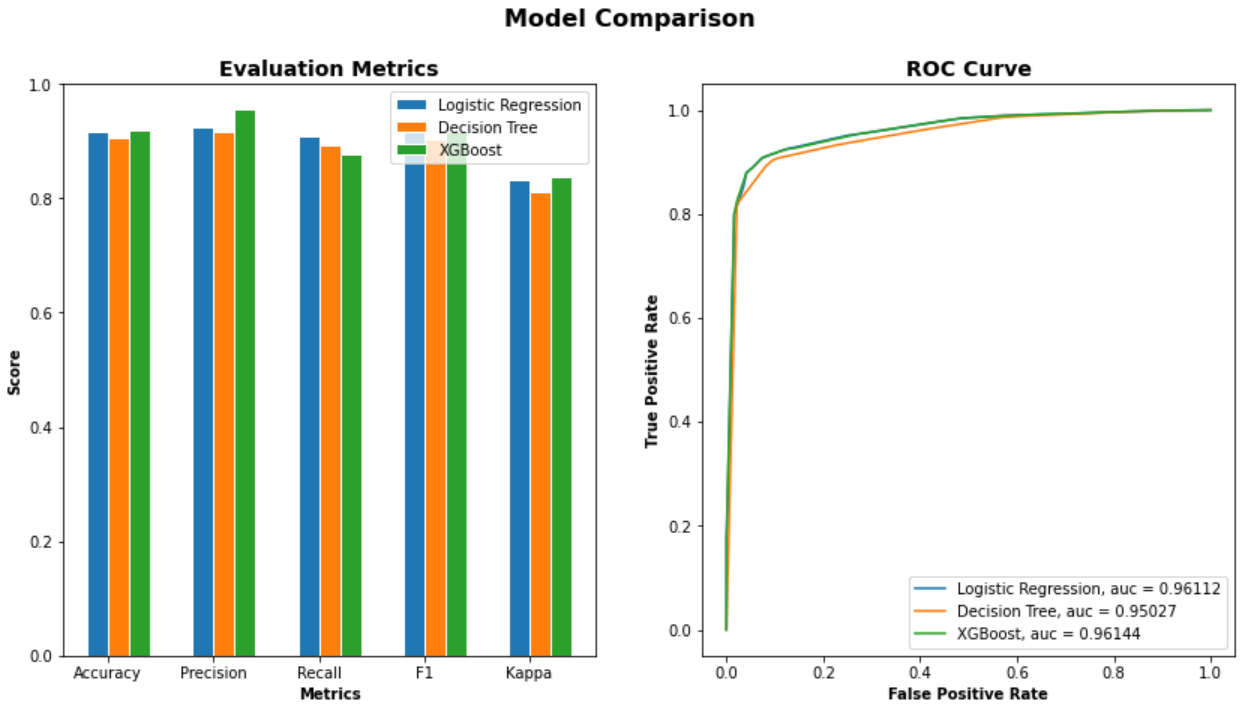

Figure 17. Model comparison for logistic regression, decision tree, XGBoost under balanced Toronto COVID-19 2020-2022 Dataset with selected features.

Discussion

The current study implemented machine learning algorithms: logistic regression, decision trees, and XGBoost in predicting COVID-19 mortality risk using imbalanced data, balanced data, and balanced data after feature selection obtained from COVID-19 patients in Toronto, Canada.

This study demonstrated that machine learning techniques could provide guidance to healthcare professionals in targeting patients with higher mortality risks. This would alleviate stress and burden from the current healthcare environment, allocate limited resources and prioritize healthcare goals.

Based on the findings, imbalanced data demonstrated low recall and precision in all algorithms whereas balanced data using SMOTE improved the case; however, low inter-rater reliability was suggested through the low kappa scores observed among all models. After feature selection, the kappa score drastically increased in all algorithms, suggesting the importance of feature selection. The accuracy in all models after feature selection remained relatively the same compared to the all-feature dataset; moreover, the efficiency improved, evident through the shortening of model training time. XGBoost and logistic regression had the best performance, whereas XGBoost had slightly better performance in terms of kappa score and precision. Thus, in the current study, XGBoost was the best-performing machine learning algorithm among the three. This was consistent with the findings of Feng et al., 2021, Ustebay et al., (2022), and Ottenhoff et al., 2011 where XGBoost outperformed all other algorithms. Given the performance in the current study, it is suggested that healthcare professionals could use XGBoost to predict patients that would be more prone to develop severe medical consequences after the acquisition of COVID-19. In all three datasets, hospitalization status was suggested to be the strongest predictor responsible for COVID-19 mortality, evident in the decision tree, XGBoost, and Chi-square method. This suggests that patients who have been previously admitted to the hospital have a higher mortality risk than others, as hospitalization is often linked with other medical histories. This provided insight into the need for medical screening for vulnerable populations and prioritizing populations that are immunocompromised or have had past medical conditions. Followed by hospitalization

was client age, in the current study an increase in the death count was observed with patient age, and it is often used as a second predictor following past hospitalizations status. This finding was consistent with the study of Ma et al., 2020 where a higher mortality rate was observed in patients 54.5 than 44.5 years of age.

There were limitations in the current study where many clinical characteristics were not publicly available. Characteristics such as the severity of symptoms, types of symptoms, and the number of symptoms could be great predictors of mortality risk for COVID-19. In the study of Pourhomayoun and Shakibi, 2021, the type of COVID symptoms experienced by the patients provided important information to the machine learning algorithm when predicting mortality risks. In addition, obesity was also suggested to link with increased mortality risk, suggested by the study by Guan et al., 2020. In the current study, an increase in fatality count with months was also evident, however, its underlying mechanism is unknown.

Conclusion

The current study demonstrated that machine learning can be used as an important tool to predict COVID-19 mortality risk with high accuracy. It also demonstrated the importance of data balancing and feature selection in machine learning algorithms as it would affect the predictive ability and power. Although the study only focused on open data, it was sufficient for the machine learning algorithm to establish stable results. Among the models, XGBoost had a slightly better performance in predicting the COVID-19 mortality risk which may assist in health planning under the current healthcare stress.

References

Alimohamadi, Y., Sepandi, M., Taghdir, M., & Hosamirudsari, H. (2020). Determine the most common clinical symptoms in COVID-19 patients: a systematic review and meta-analysis. *Journal of preventive medicine and hygiene*, *61*(3), E304–E312. https://doi.org/10.15167/2421-4248/jpmh2020.61.3.1530

Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees (The Wadsworth Statistics/probability Series). Belmont, California: Wadsworth International Group; 1984.

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future healthcare journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Feng, C., Kephart, G., & Juarez-Colunga, E. (2021). Predicting COVID-19 mortality risk in Toronto, Canada: a comparison of tree-based and regression-based machine learning methods. *BMC medical research methodology*, *21*(1), 267. https://doi.org/10.1186/s12874-021-01441-4

Friedman, J. H., & Popescu, B. E. (2003). Importance sampled learning ensembles. Department of Statistics, Stanford University, Stanford.

Fernandes, F. T., de Oliveira, T. A., Teixeira, C. E., Batista, A., Dalla Costa, G., & Chiavegatto Filho, A. (2021). A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil. *Scientific reports*, *11*(1), 3343. https://doi.org/10.1038/s41598-021-82885-y

Guan, W. J., Liang, W. H., Zhao, Y., Liang, H. R., Chen, Z. S., Li, Y. M., Liu, X. Q., Chen, R. C., Tang, C. L., Wang, T., Ou, C. Q., Li, L., Chen, P. Y., Sang, L., Wang, W., Li, J. F., Li, C. C., Ou, L. M., Cheng, B., Xiong, S., … China Medical Treatment Expert Group for COVID-19 (2020). Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. *The European respiratory journal*, *55*(5), 2000547. https://doi.org/10.1183/13993003.00547-2020

Jamshidi, E., Asgary, A., Tavakoli, N., Zali, A., Setareh, S., Esmaily, H., Jamaldini, S. H., Daaee, A., Babajani, A., Sendani Kashi, M. A., Jamshidi, M., Jamal Rahi, S., & Mansouri, N. (2022). Using Machine Learning to Predict Mortality for COVID-19 Patients on Day 0 in the ICU. *Frontiers in digital health*, *3*, 681608. https://doi.org/10.3389/fdgth.2021.681608

May M. (2021). Eight ways machine learning is assisting medicine. *Nature medicine*, *27*(1), 2–3. https://doi.org/10.1038/s41591-020-01197-2

Ottenhoff, M. C., Ramos, L. A., Potters, W., Janssen, M., Hubers, D., Hu, S., Fridgeirsson, E. A., Piña-Fuentes, D., Thomas, R., van der Horst, I., Herff, C., Kubben, P., Elbers, P., Marquering, H. A., Welling, M., Simsek, S., de Kruif, M. D., Dormans, T., Fleuren, L. M., Schinkel, M., … Dutch COVID-PREDICT research group (2021). Predicting mortality of individual patients with COVID-19: a multicentre Dutch cohort. *BMJ open*, *11*(7), e047347. https://doi.org/10.1136/bmjopen-2020-047347

Pourhomayoun, M., & Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart health (Amsterdam, Netherlands)*, *20*, 100178. https://doi.org/10.1016/j.smhl.2020.100178

Sperandei S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, *24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Ustebay, S., Sarmis, A., Kaya, G. K., & Sujan, M. (2022). A comparison of machine learning algorithms in predicting COVID-19 prognostics. *Internal and emergency medicine*, 1–11. Advance online publication. https://doi.org/10.1007/s11739-022-03101-x

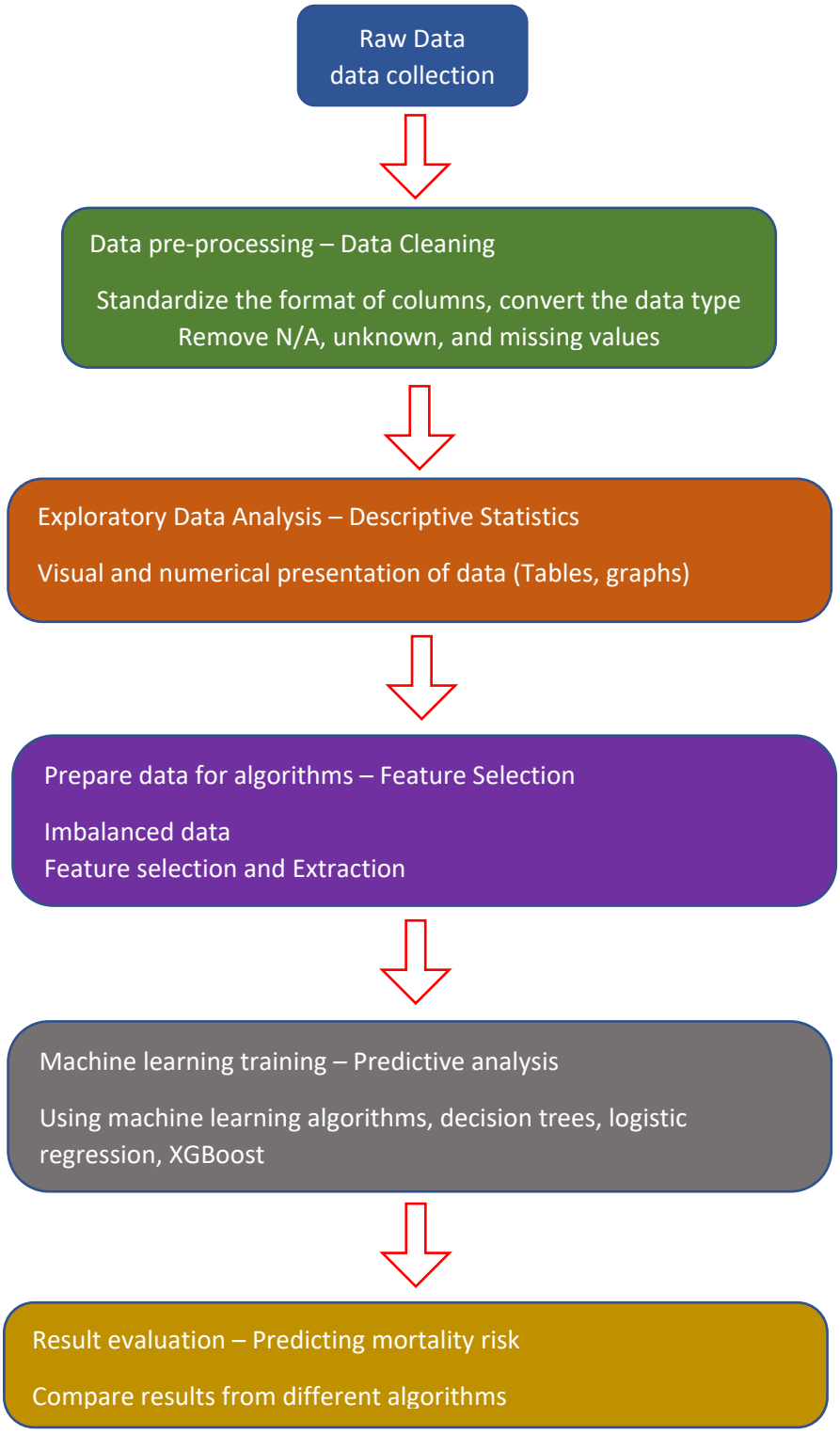

Figure 1. Tentative Overall Methodology

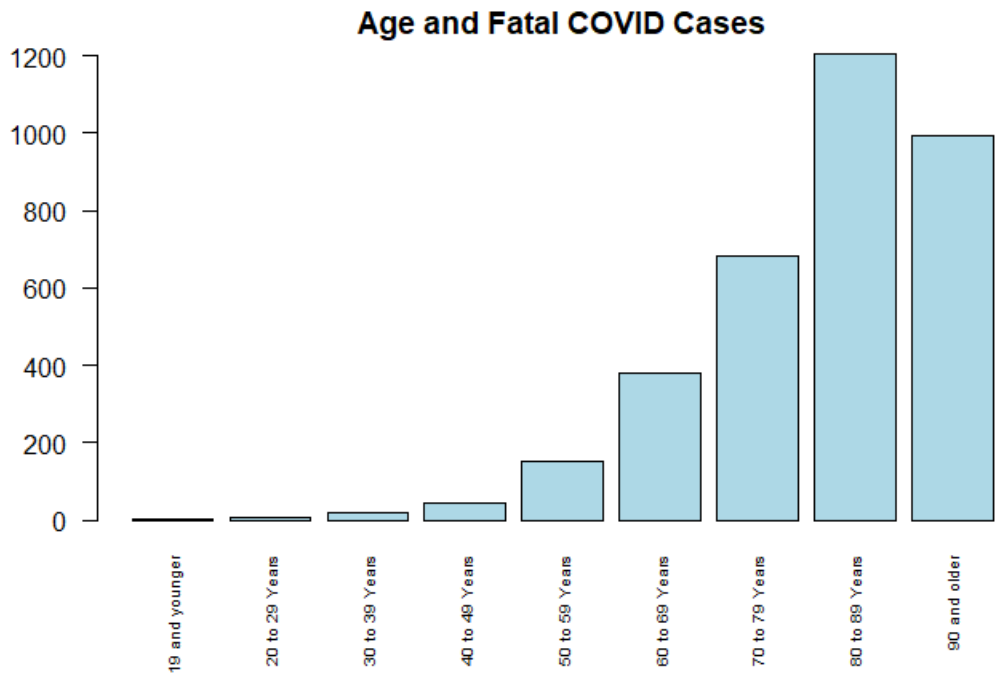

Figure 2. Bar graph of age group with the frequency of fatal COVID cases in Toronto.

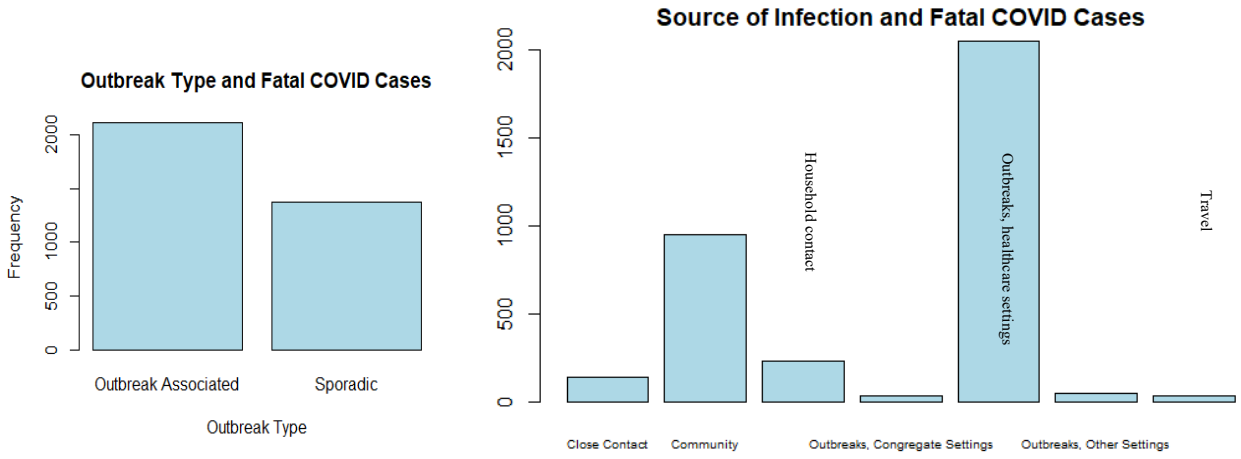

Figure 3. Bar graph of outbreak type and outbreak setting with the frequency of fatal COVID cases in Toronto.

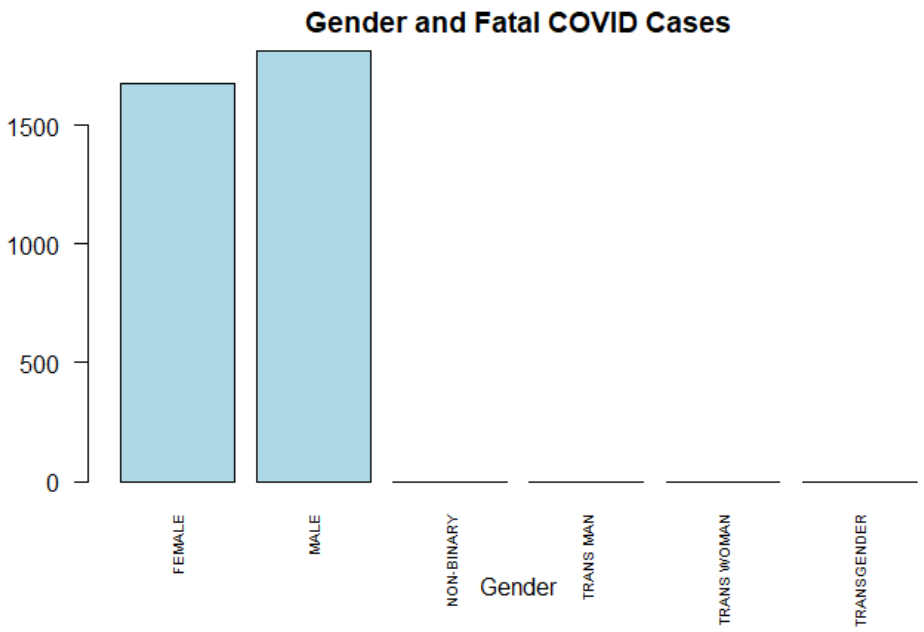

Figure 4. Bar graph of patient gender with the frequency of fatal COVID cases in Toronto.

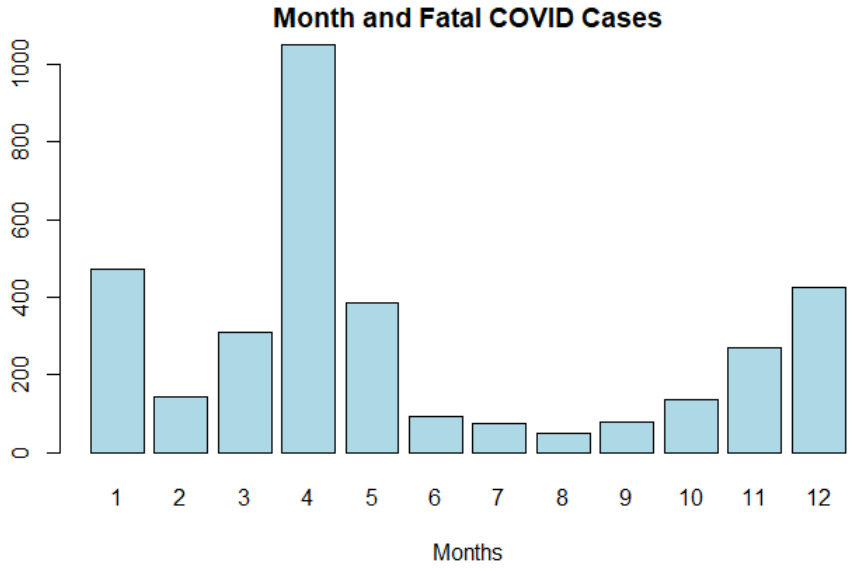

Figure 5. Bar graph of the month with the frequency of fatal COVID cases in Toronto.

Figure 6. Bar graph of past medical status with the frequency of fatal COVID cases in Toronto.

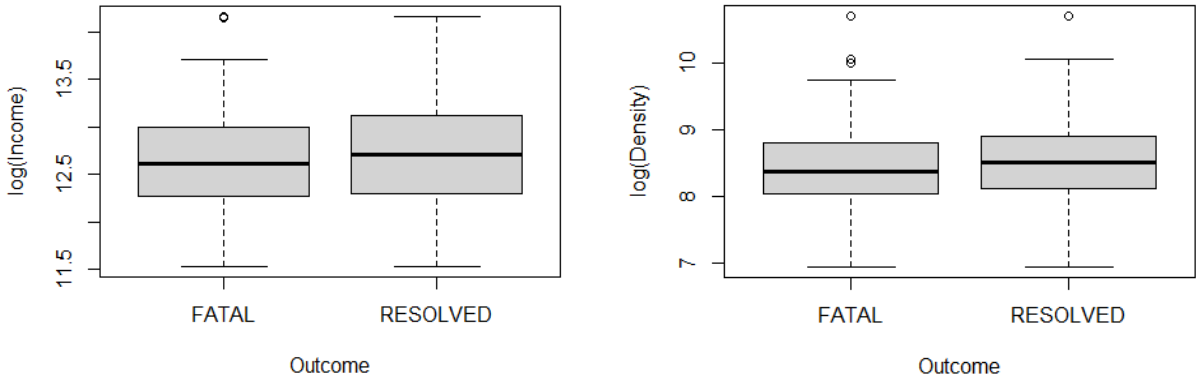

Figure 7. Boxplot of neighborhood household income and density per kilometer square with the frequency of COVID cases.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.5730   0.0164   0.0302   0.0718   2.4151

Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                         6.173937   1.088778   5.671 1.42e-08 ***
Outbreak.AssociatedSporadic                         0.071933   0.309205   0.233 0.816041
Age.Group20 to 29 Years                            -0.639842   0.839866  -0.762 0.446157
Age.Group30 to 39 Years                            -1.357841   0.767082  -1.770 0.076704 .
Age.Group40 to 49 Years                            -2.223396   0.732399  -3.036 0.002399 **
Age.Group50 to 59 Years                            -3.089894   0.717694  -4.305 1.67e-05 ***
Age.Group60 to 69 Years                            -4.031198   0.713775  -5.648 1.63e-08 ***
Age.Group70 to 79 Years                            -5.030746   0.712509  -7.061 1.66e-12 ***
Age.Group80 to 89 Years                            -5.864266   0.711724  -8.240  < 2e-16 ***
Age.Group90 and older                              -6.286101   0.712354  -8.824  < 2e-16 ***
Source.of.InfectionCommunity                        0.006610   0.130420   0.051 0.959581
Source.of.InfectionHousehold Contact                0.147197   0.154676   0.952 0.341275
Source.of.InfectionOutbreaks, Congregate Settings   0.002775   0.403315   0.007 0.994510
Source.of.InfectionOutbreaks, Healthcare Institutions -1.206761   0.322514  -3.742 0.000183 ***
Source.of.InfectionOutbreaks, Other Settings        0.002559   0.269170   0.010 0.992415
Source.of.InfectionTravel                          -0.066162   0.297992  -0.222 0.824295
Client.GenderMALE                                  -0.458580   0.050349  -9.108  < 2e-16 ***
Client.GenderNON-BINARY                             6.210140 127.353943   0.049 0.961108
Client.GenderTRANS MAN                              5.743843 339.145288   0.017 0.986487
Client.GenderTRANS WOMAN                            7.088680 368.680529   0.019 0.984660
Client.GenderTRANSGENDER                            7.899103 297.268259   0.027 0.978801
Ever.HospitalizedYes                               -2.206833   0.056351 -39.163  < 2e-16 ***
Ever.in.ICUYes                                     -1.523263   0.126600 -12.032  < 2e-16 ***
Ever.IntubatedYes                                  -1.476372   0.145836 -10.124  < 2e-16 ***
Month2                                             -0.050359   0.136768  -0.368 0.712720
Month3                                             -0.410255   0.107305  -3.823 0.000132 ***
Month4                                             -0.809911   0.078809 -10.277  < 2e-16 ***
Month5                                             -0.329038   0.096550  -3.408 0.000655 ***
Month6                                             -0.015324   0.153406  -0.100 0.920428
Month7                                              1.095290   0.163873   6.684 2.33e-11 ***
Month8                                              0.880178   0.183659   4.792 1.65e-06 ***
Month9                                              0.444005   0.167579   2.650 0.008060 **
Month10                                            -0.416463   0.137469  -3.029 0.002450 **
Month11                                            -0.502247   0.115031  -4.366 1.26e-05 ***
Month12                                            -0.550976   0.096738  -5.696 1.23e-08 ***
Income                                              0.158501   0.047908   3.308 0.000938 ***
Density                                             0.204982   0.041636   4.923 8.52e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8. Logistic regression summary for imbalanced data.

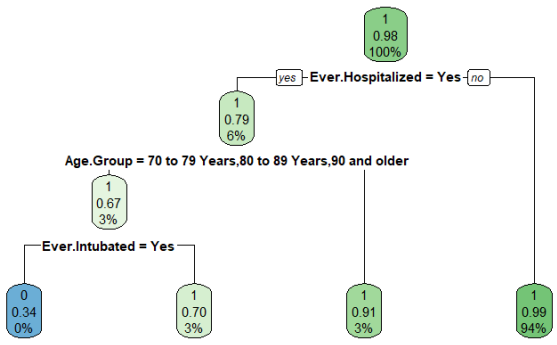

Figure 10. Decision tree for imbalanced data.

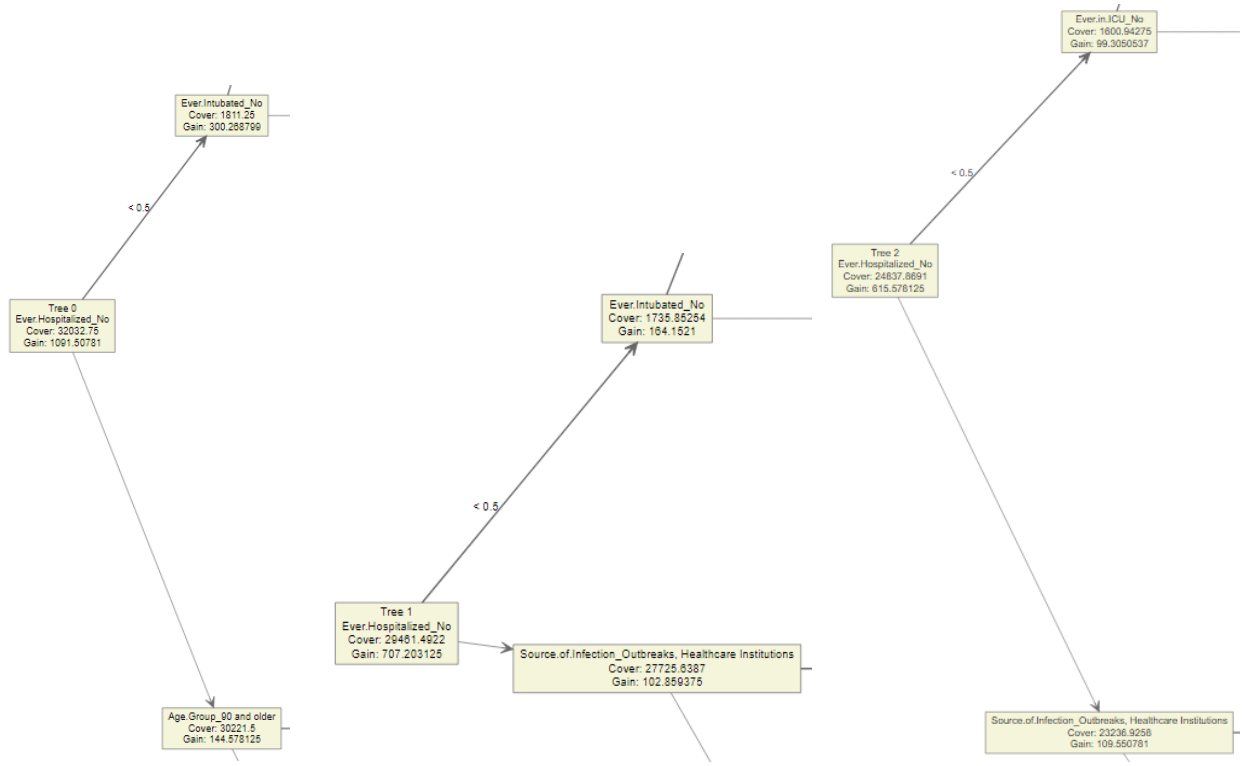

Figure 11. Snapshot of XGBoost tree 0-2 in imbalanced data.

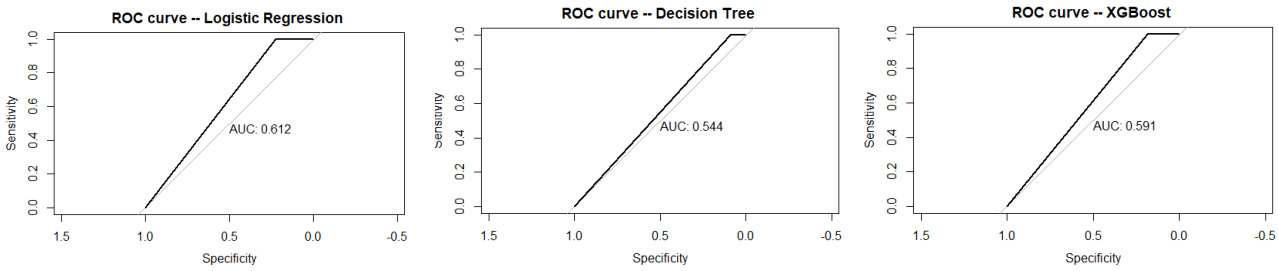

Figure 12. ROC curve for logistic regression, decision tree, and XGBoost in imbalanced data.

	coef	std err	z	P>|z|	[0.025	0.975]
Outbreak.Associated	-1.8481	0.133	-13.930	0.000	-2.108	-1.588
Ever.Hospitalized	-2.6426	0.025	-103.811	0.000	-2.693	-2.593
Ever.in.ICU	-2.1557	0.071	-30.548	0.000	-2.294	-2.017
Ever.Intubated	-1.3285	0.090	-14.716	0.000	-1.505	-1.152
Age.Group.19.and.younger	12.2728	1.006	12.194	0.000	10.300	14.245
Age.Group.20.to.29.Years	9.0883	0.169	53.828	0.000	8.757	9.419
Age.Group.30.to.39.Years	9.1126	0.158	57.774	0.000	8.803	9.422
Age.Group.40.to.49.Years	7.7905	0.108	71.804	0.000	7.578	8.003
Age.Group.50.to.59.Years	6.5399	0.093	69.964	0.000	6.357	6.723
Age.Group.60.to.69.Years	4.7533	0.088	54.232	0.000	4.581	4.925
Age.Group.70.to.79.Years	3.5122	0.086	40.618	0.000	3.343	3.682
Age.Group.80.to.89.Years	2.5602	0.086	29.836	0.000	2.392	2.728
Age.Group.90.and.older	1.9606	0.087	22.557	0.000	1.790	2.131
Source.of.Infection.Close.Contact	5.4761	0.141	38.871	0.000	5.200	5.752
Source.of.Infection.Community	4.5007	0.134	33.486	0.000	4.237	4.764
Source.of.Infection.Household.Contact	5.1777	0.139	37.354	0.000	4.906	5.449
Source.of.Infection.Outbreaks,.Congregate.Settings	4.7208	0.178	26.467	0.000	4.371	5.070
Source.of.Infection.Outbreaks,.Healthcare.Institutions	0.6936	0.108	6.441	0.000	0.483	0.905
Source.of.Infection.Outbreaks,.Other.Settings	5.3809	0.147	36.695	0.000	5.094	5.668
Source.of.Infection.Travel	7.9401	0.267	29.757	0.000	7.417	8.463
Client.Gender.FEMALE	0.5525	0.101	5.451	0.000	0.354	0.751
Client.Gender.MALE	-0.1028	0.101	-1.015	0.310	-0.301	0.096
Client.Gender.NON-BINARY	12.2326	1701.755	0.007	0.994	-3323.145	3347.610
Client.Gender.TRANS.MAN	12.6179	3346.062	0.004	0.997	-6545.543	6570.778
Client.Gender.TRANS.WOMAN	16.8757	1.56e+04	0.001	0.999	-3.05e+04	3.06e+04
Client.Gender.TRANSGENDER	14.2345	1440.441	0.010	0.992	-2808.978	2837.447
Month.1	5.2023	0.106	48.856	0.000	4.994	5.411
Month.2	5.0464	0.114	44.230	0.000	4.823	5.270
Month.3	4.4926	0.107	42.034	0.000	4.283	4.702
Month.4	4.1582	0.105	39.621	0.000	3.952	4.364
Month.5	4.5490	0.106	42.890	0.000	4.341	4.757
Month.6	5.6029	0.116	48.430	0.000	5.376	5.830
Month.7	6.8827	0.114	60.513	0.000	6.660	7.106
Month.8	7.1667	0.119	60.145	0.000	6.933	7.400
Month.9	6.7874	0.121	56.020	0.000	6.550	7.025
Month.10	5.3841	0.118	45.821	0.000	5.154	5.614
Month.11	4.5631	0.110	41.539	0.000	4.348	4.778
Month.12	4.7716	0.108	44.169	0.000	4.560	4.983
Income	-0.5639	0.014	-41.664	0.000	-0.590	-0.537
Density	-0.3498	0.015	-23.728	0.000	-0.379	-0.321

Figure 13. Logistic regression summary of balanced data with all features (first run)

Figure 14. Decision tree (left) and XGBoost (right) plot for balanced data with all features.