**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

**Revised 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 CONVID-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 in 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 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 predict mortality risk of COVID-19, which sub-populations of patients 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 tree and XGBoost. The programs that will be used are Python and R.

**Data set link:** [COVID-19 Cases in Toronto - City of Toronto Open Data Portal](https://open.toronto.ca/dataset/covid-19-cases-in-toronto/)

**Github Repository:** [keyunzhou2/CIND820 (github.com)](https://github.com/keyunzhou2/CIND820)

**Introduction**

With the spread of coronavirus disease in 2019 (COVID-19), it had caught many countries off-guard, causing a great distress in the lives of many. Given the novelty of the disease, rapid mutation and the lack of medical knowledge, it had resulted more than 6 million deaths world-wide (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 with regards to 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 further increase of death rate with the lack of medical attention. As such, medical decision-making techniques could be vital in alleviating the stress of hospitals with the 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 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 have shown to be 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 patient 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 and random forest is often used for non-linear decisions. However, they are prone to over-fitting and limits to 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 professions to prioritize patient care.   
2. Types of factors and strongest predictor for COVID-19 mortality (demographic or clinic data), 3. To compare among the three algorithms for the best performing algorithm for COVID-19 mortality risk, where the model accuracy will be evaluated with 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 decision-making process.

**Tentative Overall Methodology**

Raw Data  
data collection

Result evaluation – Predicting mortality risk

Compare results from different algorithms and apply them to answer research questions.

Machine learning training – Predictive analysis

Using machine learning algorithms, decision trees, logistic regression

Prepare data for algorithms – Feature Selection

Imbalanced data  
Feature selection and Extraction

Exploratory Data Analysis – Descriptive Statistics

Visual and numerical presentation of data (Tables, graphs)

Data pre-processing – Data Cleaning

Standardize the format of columns, convert the data type  
Remove N/A, unknown, and missing values

**Descriptive Analysis**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Categoric Variable** | **Levels** | **Total n = 176,432** | **%** | **Resolved n=170,233** | **%** | **Fatal n=3426** | **%** |
| Outbreak Associated | Outbreak Associated | 38703 | 21.9 | 36190 | 21.3 | 2090 | 61.0 |
| Sporadic | 137729 | 78.1 | 134043 | 78.7 | 1336 | 39.0 |
| Age Group | 19 and younger | 28564 | 16.2 | 28021 | 16.5 | 1 | 0.0 |
| 20-29 | 30664 | 17.4 | 30265 | 17.8 | 6 | 0.2 |
| 30-39 | 28980 | 16.4 | 28471 | 16.7 | 19 | 0.6 |
| 40-49 | 24842 | 14.1 | 24301 | 14.3 | 43 | 1.3 |
| 50-59 | 23458 | 13.3 | 22925 | 13.5 | 147 | 4.3 |
| 60-69 | 15162 | 8.6 | 14542 | 8.5 | 372 | 10.9 |
| 70-79 | 9080 | 5.1 | 8294 | 4.9 | 667 | 19.5 |
| 80-89 | 9326 | 5.3 | 8071 | 4.7 | 1186 | 34.6 |
| 90 and older | 6356 | 3.6 | 5343 | 3.1 | 985 | 28.8 |
| Source of Infection | close contact | 18531 | 10.5 | 18076 | 10.6 | 135 | 3.9 |
| community | 75478 | 42.8 | 73265 | 43.0 | 936 | 27.3 |
| household contact | 38562 | 21.9 | 37638 | 22.1 | 216 | 6.3 |
| outbreak (healthcare institute) | 3940 | 2.2 | 3882 | 2.3 | 30 | 0.9 |
| outbreak (congregated settings) | 25285 | 14.3 | 23052 | 13.5 | 20030 | 584.6 |
| outbreak (other settings) | 10475 | 5.9 | 10235 | 6.0 | 46 | 1.3 |
| Travel | 4161 | 2.4 | 4085 | 2.4 | 33 | 1.0 |
| Gender | female | 94267 | 53.4 | 91156 | 53.5 | 1640 | 47.9 |
| male | 82073 | 46.5 | 78988 | 46.4 | 1786 | 52.1 |
| non-binary | 57 | 0.0 | 55 | 0.0 | 0 | 0.0 |
| trans man | 7 | 0.0 | 7 | 0.0 | 0 | 0.0 |
| trans woman | 11 | 0.0 | 10 | 0.0 | 0 | 0.0 |
| transgender | 17 | 0.0 | 17 | 0.0 | 0 | 0.0 |
| Ever Hospitalized | yes | 10258 | 5.8 | 7960 | 4.7 | 2137 | 62.4 |
| no | 166174 | 94.2 | 162273 | 95.3 | 1289 | 37.6 |
| Ever in ICU | yes | 1880 | 1.1 | 1090 | 0.6 | 753 | 22.0 |
| no | 174552 | 98.9 | 169143 | 99.4 | 2673 | 78.0 |
| Ever Intubated | yes | 1109 | 0.6 | 531 | 0.3 | 559 | 16.3 |
| no | 175323 | 99.4 | 169702 | 99.7 | 2867 | 83.7 |
| Outcome | Resolved | 172979 | 98.0 | 170233 | 100.0 |  | 0.0 |
| fatal | 3453 | 2.0 |  | 0.0 | 3426 | 100.0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Numeric Variable** | **Days from onset**  **to report** | **Resolved** | **Fatal** | **Average after-tax income of households in 2015** | **Resolved** | **Fatal** |
| **n** | 176432 | 170233 | 3426 | 176432 | 170233 | 3426 |
| **Mean** | 4 | 4.05 | 3.8 | 375944 | 376340 | 356279 |
| **SD** | 5.1 | 4.98 | 4.9 | 205587 | 205835 | 191847.6 |
| **Median** | 3 | 3 | 2 | 332776 | 332776 | 302358 |
| **Q1** | 2 | 2 | 1 | 222404 | 222648 | 215135 |
| **Q3** | 5 | 5 | 5 | 496958 | 516575 | 441052 |
| **Min** | 0 | 0 | 0 | 102259 | 102259 | 102259 |
| **Max** | 383 | 383 | 90 | 1413132 | 1413132 | 1413132 |

**Literature Reviews & References**

Feng et al., (2021) Examined the mortality risk for COVID-19 using confirmed cases in Toronto, Canada, 2020. The researchers had 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, gender and combined 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 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 a 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 Netherland. Both machine learning algorithms showed better performance than age-based decision (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 mortality risk of COVID-19. The researchers have used patient demographic, past medical background and laboratory biomarker as predictors 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) 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 were also shown to have effect in influencing mortality risk.

Fernandes et al., (2021) studied for 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 AUROC of 0.91 in 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 series of wrapper and filter method 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%.

**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>

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>

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>

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>