COVID-19: Which Factors Impact Mortality?

Nick Wawee
Department of Computer Science
Bowling Green State University
Bowling Green, United States
Email: nwawee@bgsu.edu

Connor MacMillan
Department of Computer Science
Bowling Green State University
Bowling Green, United States
Email: cmacmil@bgsu.edu

Kshitij Saxena
Department of Computer Science
Bowling Green State University
Bowling Green, United States
Email: ksaxena@bgsu.edu

Grant Ferrell
Department of Computer Science
Bowling Green State University
Bowling Green, United States
Email: fgrant@bgsu.edu

Abstract— **Multiple models were used to determine which factors from the CDC's public-use dataset have the largest impact on patients' mortality rate. According to both the Logistic Regression and Multi-Layer Perceptron Classifier models, whether the patient is over the age of 80 has the largest impact, even more than pre-existing health conditions.**

Keywords—**Multi-Layer Perceptron Classifier (MLPClassifier), Artificial Neural Networks (ANN), Logistic Regression, Area Under the Curve (AUC), Receiver Operator Characteristic (ROC), Coronavirus Disease 2019 (COVID-19)**

I. Introduction

COVID-19 is a global, and current pandemic that is affecting all of our lives. With our research, we hope to contribute to the scientists that are helping to stop the spread of this virus and prioritize the limited supply of vaccines to the most at-risk populations. The dataset we used for this project, COVID-19 Case Surveillance Public Use Data, is freely available from the Center for Disease Control's website.

II. Background

Fabic et al. [1] in their paper discuss that using crude data, Case Fatality Rates (CRF) have decreased over time. However, adjusting the data by age reveals the differences in CRF by ethnicity. The study, using CDC's Case Surveillance Data, reveals that non-Hispanic (NH) whites have a lower CRF, while the CRF of NH Blacks, NH Asians, and Hispanics have seen a rise[1]. Fabic et al. [2] in their paper, using the CDC's Case Surveillance Data, discuss that while in late March, ages 50 and under accounted for 25% of weekly hospitalizations, by late June it was 32%.

Rumain et al. [3] discuss their findings that are contrary to other research that adolescents (10-19 years) and youth (15-24 years) are less susceptible to Covid-19 than older adults. Their research uses CDC's Case Surveillance Data to analyze patterns. The study determines that the prevalence of COVID-19 was much greater for adolescents and youth than it was for older adults with a p-value less than .00001.

III. Methodology

A. Data Cleaning & Description

COVID-19 surveillance data from the Centers for Disease Control and Prevention (CDC) was obtained [4]. The dates for all of the data were not included in this analysis because the main focus of this analysis is on the characteristics of the individuals. All records labeled unknown, missing, or other were removed, which lowered the data size from 1,048,576 records to just 328,977 records and eight columns to be analyzed. Cleaning the data involved dropping the dates columns, dropping any non-laboratory confirmed cases, keeping only "Male" and "Female" values in the "sex" column, removing rows where "age_group", "Race and ethnicity", "hosp_yn", "icu_yn", "death_yn", and "medcond_yn" were unknown.

The distributions of each variable, as well as visualizations for correlations within the data, are in the Describe Jupyter notebook. Pearson correlations were calculated between features as well as between each feature and response. The distribution of these correlations, as well as the correlation of each dummy variable to the response variable, is displayed within this notebook.

B. Logistic Regression

Logistic regression was performed to model the mortality rate based on the individual's demographics. K-fold cross-validation was employed with five folds in the data. All results reported are the mean and standard deviation of the five folds. Mortality was considered to occur if the predicted probability of the model was greater than or equal to 0.5. Statsmodels was utilized to perform the regression and SciKit-Learn was applied to determine the classification metrics [5] [6].

C. Multi-Layer Perceptron Classifier

A grid search was performed to find the best parameters for SciKit-Learn's MLPClassifier model. GridSearchCV iterated through multiple parameters to determine which is the best model. After finding the best version of the MLPClassifier, the model was re-run to perform cross-validation and store relevant metrics. These results were compared to the Logistic Regression to determine which model performed better. The Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) Curve showed that the MLPClassifier was slightly better with an AUC of 0.74 instead of 0.68. To relate the MLPClassifier model back to the question of which factors impact mortality, the weights for each feature were plotted. Since this is a single-layer neural network, the input weights should show us how each feature impacts the prediction of mortality, which is shown in Figure 5.

IV. Results & Discussion

Figure 1-2 visualizes the predictions made by the MLP Classifier and Logistic Regression models. Figure 3-4 displays the receiver operating characteristic curve, which shows the reliability of the predictions. Table 1 provides evaluation metrics based on these predictions. As shown, the MLP Classifier has a higher predictive capability because it has a larger and less variable F1 score and AUC. The Logistic Regression model shows that it has prediction results that vary more than the MLP based on the variation seen in the true positive rate (or recall).

FIGURE I. MLPClassifier Confusion Matrix

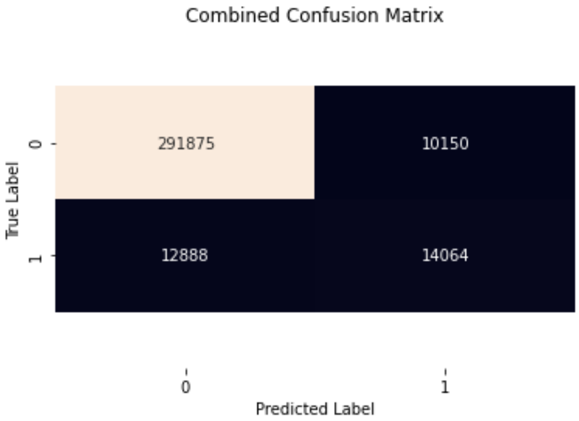


FIGURE II. Logistic Regression Confusion Matrix

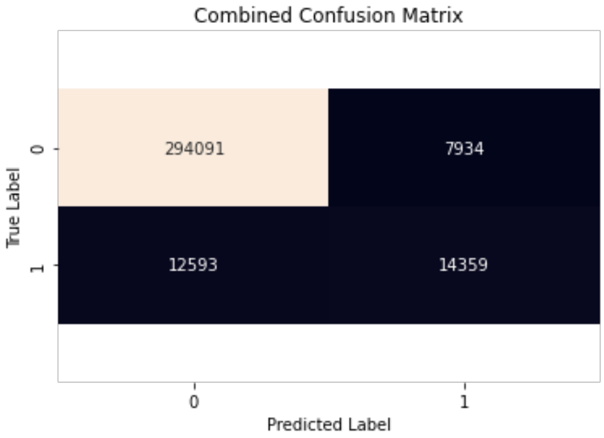


FIGURE III. MLP ROC Curve

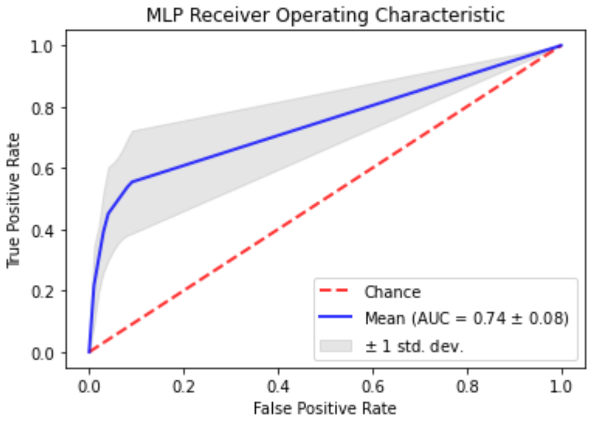
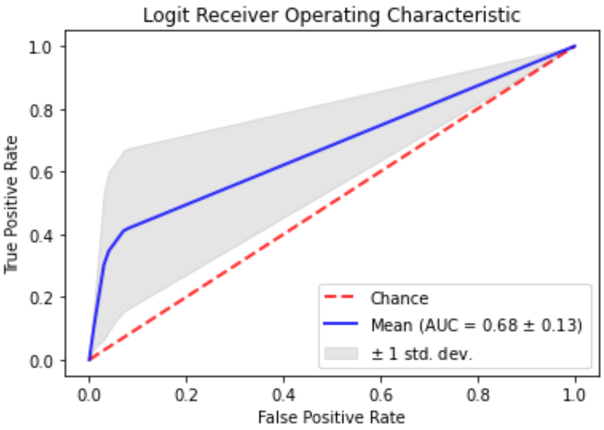


FIGURE IV. Logistic Regression ROC Curve



Comparing the Combined Confusion Matrices in Figure 1-2, the Logistic Regression performed slightly better than the MLP Classifier, but then reviewing the mean and standard deviations of those metrics shown in Table 1 shows that the MLP Classifier appears to perform better in most categories. The Logistic Regression model has a higher point estimate of precision, but it is statistically insignificant due to the standard deviation. The elevated point estimate makes sense because there are more false positives as shown in Figure 1-2. The F1 score and precision in both models exceed what was found in [7], but have a lower recall score ([7] did not provide standard deviation). It is interesting to note that if only the mean accuracy was used to pick the model, Logistic Regression would appear better. This illustrates a perfect example of how all metrics need to be evaluated before making a decision.

TABLE I. Model Comparison of Classification Metrics

Model	F1 Score	Precision	Recall	Accuracy
Logistic Regression	0.68 ± 0.13	0.83 ± 0.05	0.68 ± 0.13	0.94 ± 0.04
MLP Classifier	0.75 ± 0.06	0.82 ± 0.09	0.74 ± 0.09	0.93 ± 0.02

FIGURE V. MLPCLASSIFIER INPUT LAYER WEIGHTS

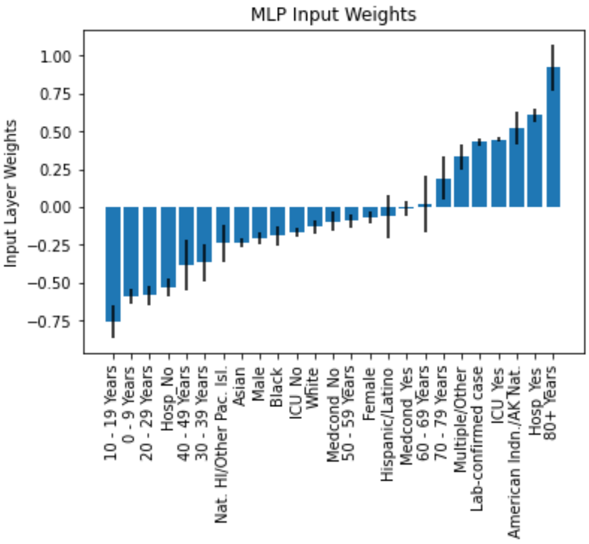


FIGURE VI. LOGISTIC REGRESSION COEFFICIENTS

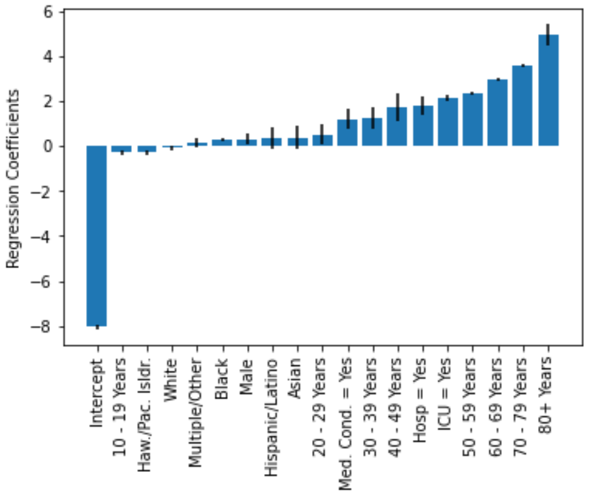
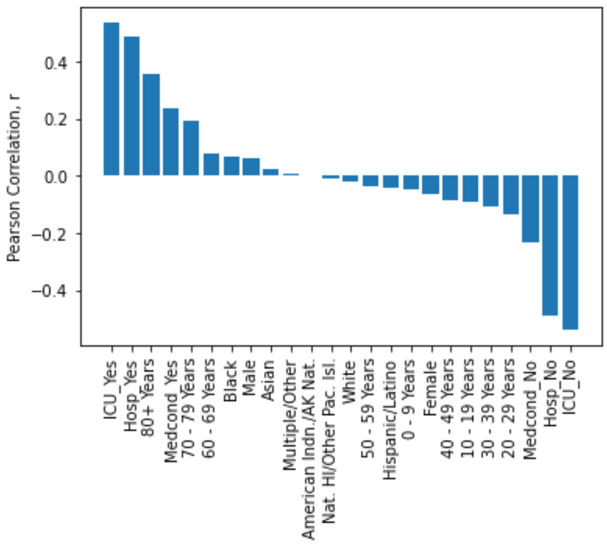


FIGURE VII. CORRELATIONS BETWEEN FEATURES AND MORTALITY



Using the input layer weights of the MLPClassifier is a good way to visualize each feature's impact on the mortality rate. Comparing the weights from the MLPClassifier to the coefficients of the Logistic Regression model shows that both found being over 80 years old as the most impactful feature. There are minor differences in the results, but the results are consistent with the largest and smallest values.

The age, hospitalization status, ICU status, and medical condition are the top four factors that impact mortality rate, as shown in Figure 5-7. Since the MLPClassifier shows large variation in the Input Layer Weights, the Logistic Regression is more reliable when it comes to finding which demographics to prioritize. Age in the logit model was determined to be a primary factor that impacted mortality, which parallels what was found in other COVID-19 survivability studies [8,9].

Ethnicity comparisons in the logit model showed to be primarily all statistically insignificant, with the exception of American Indian/Alaskan and Asian individuals. This contrasts with what was stated in [1] but coincides with what was found in [8]. The male sex of the individual in our logit model showed to have statistical significance by positively contributing to death unlike in [8], but had minimal impact. See the "Logit" notebook for specific statistics regarding the logistic regression model.

CONCLUSION

Comparing the two models used, the Logistic Regression and MLP Classifier, showed that they performed similarly, but the MLP Classifier was more consistent in its cross-validation. Both models were then used to find which features impacted mortality the most. Using this information can help prioritize who should get the first wave of vaccines. First the elderly (80+) should get treated, then 60-79-year-olds, focusing on American Indian/Alaska Native and Asian populations.

REFERENCES

[1] M. S. Fabic, Y. Choi, D. Bishai, "Deaths among COVID Cases in the United States: Racial and Ethnic Disparities Persist," medRxiv, 2020.

[2] M. S. Fabic and Y. Choi, "Trends in age distribution of COVID-19 cases, hospitalizations, and deaths by race in the United States", 13-Aug-2020. [Online]. Available: osf.io/preprints/socarxiv/7edgu.

[3] B. T. Rumain, M. Schneiderman, A. Geliebter, "Prevalence of Covid-19 in Adolescents and Youth Compared with Older Adults in States Experiencing Surges", medRxiv, Oct. 2020.

[4] Centers for Disease Control and Prevention, "COVID-19 Case Surveillance Public Use DataCase Surveillance," 11 November 2020. [Online]. Available: https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf.[Accessed 16 November 2020].

[5] Statsmodels,[Online].Available:https://www.statsmodels.org/stable/index.html. [Accessed 16 November 2020].

[6] scikit-learn, [Online]. Available: https://scikit-learn.org/stable/. [Accessed 16 November 2020].

[7] A. Patricio, R. S. Costa and R. Henriques, "COVID-19 in Portugal: predictability of hospitalization, ICU and respiratory-assistance needs," COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv, vol. Preprint,pp.https://www.medrxiv.org/content/10.1101/2020.09.29.20203141v1.article-info, 2020.

[8] A. S. Yadaw, Y.-c. Li, S. Bose, R. Iyengar, S. Bunyavanich and S. Pandey, "Clinical features of COVID-19 mortality: development and validation of a clinical prediction model," Lancet Digital Health, vol. 2, p. e516–25, 2020.

[9] Noam Barda, Dan Riesel, Amichay Akriv, et al., "Developing a COVID-19 mortality risk prediction model when individual-level data are not available," Nature Communications, Vols. 11, no. 4439, 2020