**Predicting the occurrence of seizures in a population of COVID-19 patients during prolonged hospital stays with Azithromycin and Hydroxychloroquine treatment options.**

Savannah Gilkes

Data Science Practicum

Regis University

August 22, 2021

**Research Question and Project Overview**

Throughout the COVID-19 pandemic, there has been a small population of individuals that have suffered neurological episodes, such as seizures, during their prolonged hospital stays due to the virus. The purpose of this project is to determine the factors that can be used to predict the occurrence of seizures. Understanding how COVID-19 may have long-term impacts on the nervous system is important in being able to help individuals with long-term symptoms receive proper care and treatment. Before establishing why this occurs, we need to focus on how it occurred. With the provided dataset, predictions can be done to determine what variables and attributes correlated with individuals that suffered a seizure during their hospital stay. R will be used to perform exploratory data analysis (EDA) and produce visualizations. R will also be used to build a logistic regression model and a random forest model to determine factors in the prediction of seizure occurrence in COVID-19 patients. Finally, ROC curves will be created to validate each model and 10-fold cross validation will be applied to compare each model.

**Data Description and Preparation**

This data was published in 2021 and was provided by the Chief Resident of the LSU Health Department of Neurology, Dr. David Chachkhiani. This data was cleared by the Institutional Review Board and made publicly available through Mendeley Data, where it was collected from. Before preparing this dataset, it had 32 variables with 250 observations. However, many of the variables were irrelevant to the statistical analysis and were removed. For example, variables that described a patient’s chief complaint were removed and any variables that only included one patient were removed because they are considered outliers. After removing the variables, missing values needed to be taken care of. In order to determine how many missing values are within the dataset, the function sum(is.na()) was used. To determine how many missing values are located in each variable, the function colSums(is.na()) was used. This is represented in Table 1.

Text

Description automatically generated with medium confidence

Table 1

It can be seen that there are 50 missing values in this dataset. In order to remove these missing values, the function na.omit() was used. The data was checked again to make sure there are no longer any missing values. This is represented in Table 2.

Table

Description automatically generated with medium confidence

Table 2

For this dataset, entire rows containing missing values were removed. This was because most of the missing values were located in the race variable, an important factor that cannot be replaced. The final cleaned data contained 19 variables with 208 observations.

**Exploratory Data Analysis (EDA) and Visualizations**

After the data had been cleaned, EDA was performed, and some visualizations were produced to better understand the data. Both the summary() and str() functions can give a quick overview of the data. From the summary, different attributes of the dataset can be determined, such as, the average (mean) age of the patients included in this dataset is just below 60 years old, the average (mean) hospital stay of the patients is just above 10 days, and the average BMI of the patients is 34.71.

Multiple visualizations were created to show certain aspects of the dataset. Figure 1 displays the range of ages included in the dataset. Figure 2 displays the range of the patients’ race. Figure 3 displays the male to female ratio of the dataset.

Chart, histogram

Description automatically generated

Figure 1

Chart, histogram

Description automatically generated

Figure 2

Chart, histogram

Description automatically generated

Figure 3

**Normalizing Data and Creating Train and Test Datasets**

After the data was cleaned and prepped, it needed to be normalized in order for the models to be successful. Normalizing the data was done with the following function:

Graphical user interface, text

Description automatically generated with medium confidence

For each model, training and testing datasets must be created. For this instance, a 70/30 ratio was used for the train and test datasets. The dim() function was then used to check both the train and test datasets. There are 146 objects in the train dataset and 62 in the test dataset.

**Logistic Regression**

The logistic regression model was created using the glm() function. The summary() function was then used to see the results of the model previously created. This is shown in Table 3 below.

Table

Description automatically generated

Table 3

From the above table, it was determined that race, ethnicity, and length of hospital stay all have small statistical significance regarding the occurrence of a seizure. The model was then run with the train and test datasets to compare results and determine the accuracy. The produced confusion matrix for the prediction against the test dataset can be seen in Table 4 below.

Text

Description automatically generated with medium confidence

Table 4

The accuracy of the model is strong at 0.9677, however, with the kappa value of 0.4833, there is a weak level of agreement. Next, a ROC curve was created to validate the model and the AUC value was found. The ROC curve can be seen in Figure 4 and the AUC value was 0.9083333.

Chart, box and whisker chart

Description automatically generated

Figure 4

Next, 10-fold cross validation was completed to further evaluate the model and ultimately compare this model to the random forest model. The results of performing 10-fold cross validation can be seen in Table 5 below. The accuracy is still high at 0.9568254 but has a very low level of agreement.

Text

Description automatically generated

Table 5

**Random Forest**

The random forest model was created using the randomForest() function. The results of the model are shown below in Table 6. Then, a ROC curve was created to validate the model and the AUC value was found. The ROC curve can be seen in Figure 5 and the AUC value was 0.591667.

Text

Description automatically generated

Table 6

Chart, histogram

Description automatically generated

Figure 5

Next, 10-fold cross validation was completed to further evaluate the model and ultimately compare this model to the linear regression model. The results of performing 10-fold cross validation can be seen in Table 7 below. The accuracy is still high at 0.9619048 but has no level of agreement, leading to the conclusion that the logistic regression model is the better model. In Figure 6 below, the accuracy against selected predictors for the cross validation can be seen.

**Text, letter

Description automatically generated**

Table 7

Chart, line chart

Description automatically generated

Figure 6

**Conclusion**

From the above analysis, it can be concluded that race, ethnicity, and length of stay at the hospital may have an impact on the occurrence of a seizure in a patient with COVID-19. However, the statistical significance of each was very small. The most accurate model created with the logistic regression model. These models can help provide healthcare works with information regarding who would be more likely to endure a seizure. They can be used as a preventative measure to assess what resources are needed and where they should be allocated. For future analysis, more data will need to be collected. With such a small data set from a small region of one country, the accuracy of the results is questionable.

Github:

<https://github.com/sdgilkesMSDS/Data_Science_Practicum>

**References**

Chachkhiani, D. (2021). Data for: Neurological Complications of COVID-19 Predict Worse

Outcomes During Hospitalization. *Mendeley Data*. doi: 10.17632/njhc957dfj.2

Lantz, B. (2015) *Machine Learning with R: Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R*. Packt Publishing.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

Narkhede, S. (2018, January 26). *Understanding AUC-ROC Curve.* Towards Data Science. https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Nwanganga, F., Chapple, M. (2020) *Practical Machine Learning in R*. Wiley.

Yu-Wei, C. (2015).  *Machine Learning with R Cookbook Explore over 110 Recipes to Analyze Data and Build Predictive Models with the Simple and Easy-to-Use R Code*. Packt Publishing.