**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 l

Regis University

**Purpose of Analysis 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 descriptive statistics, EDA, and produce visualizations. R will also be used to build a K-nearest-neighbor (KNN) classifier and logistic regression model to determine factors in the prediction of seizure occurrence in COVID-19 patients.

**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. Before importing the dataset into R and after removing certain variable while still using the excel file, the following variables we left.

*Age*

*Sex*

0 = Male

1 = Female

*Race*

0 = American Indian/Native Alaskan

1 = Asian

2 = Native Hawaiian or Pacific Islander

3 = African American

4 = Caucasian

*Ethnicity*

0 = Non-Hispanic

1 = Hispanic

*LOS (hospital length of stay)*

*BMI (body mass index)*

*Smoker*

0 = Never

1 = Former

2 = Current

*Asthma*

0 = No

1 = Yes

*Obesity*

0 = No

1 = Yes

*hx\_migraine (history of migraine)*

0 = No

1 = Yes

*hx\_epilepsy (history of epilepsy)*

0 = No

1 = Yes

*hx\_cva (history of stroke)*

0 = No

1 = Yes

*AMS\_comp (altered mental status during hospital stay)*

0 = No

1 = Yes

*Seizure\_comp (seizures during hospital stay)*

0 = No

1 = Yes

*Headache\_comp (headache during hospital stay)*

0 = No

1 = Yes

*Encephalitis\_comp (encephalitis during hospital stay)*

0 = No

1 = Yes

*Ageusia\_anosmia*\_*comp (ageusia or anosmia during hospital stay)*

0 = No

1 = Yes

*Azithromycin*

0 = No

1 = Yes

*Hydroxychloroquine*

0 = No

1 = Yes

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. Before moving onto visualizations and analysis, the following libraries were loaded into R.

Text

Description automatically generated

**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.

Summary:

Table

Description automatically generated

From the above summary, different attributes of the dataset can be determined. For example, 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.

Structure:

A picture containing calendar

Description automatically generated

From the above structure function, it can be seen that the variables are numeric which is needed for the knn classification model. Also, the seizure variable has been changed to factor for the logistic regression model.

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

Figure 1 shows that a majority of the patients in this data set are 50-60 years old, Figure 2 shows that most of the patients are African American, and Figure 3 shows that there are more female than male patients.

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

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

Graphical user interface, text

Description automatically generated with medium confidence

For both the KNN and the logistic regression models, training and testing datasets must be created. For this instance, 70/30 ratio was used for the train and test datasets. This was done with the following code:

A picture containing text

Description automatically generated

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.

**KNN Classification**

KNN is a type of supervised machine learning algorithm. KNN is a lazy, non-parametric algorithm. First, the correct k-value must be chosen. To begin, k = 1, k = 2, and k = 3 were chosen to see which was the most accurate and which had the lowest training error rate. The algorithm where k = 1 was produced with the following code:

k1 <- knn(train = train[1:18], test = test[1:18], cl = train$Seizure\_comp, k = 1)

Then, a cross table and confusion matrix using the test set for the algorithm were created. This can be seen in Figure 4 below.

Table

Description automatically generated

Figure 4: k = 1 Confusion Matrix

The above was done for k = 2 and the same results were produced. However, the results were less accurate with k = 3. This can be seen in Figure 5 below.

Table

Description automatically generated

Figure 5: k = 3 Confusion Matrix

Because k = 1 had an accuracy of 0.9839 with a kappa value of 0.6593, it was chosen as the optimal k-value. From Figure 4, it can be seen that the KNN model classified 61 objects out of 62 correctly and has a moderate level of agreement.

**Logistic Regression**

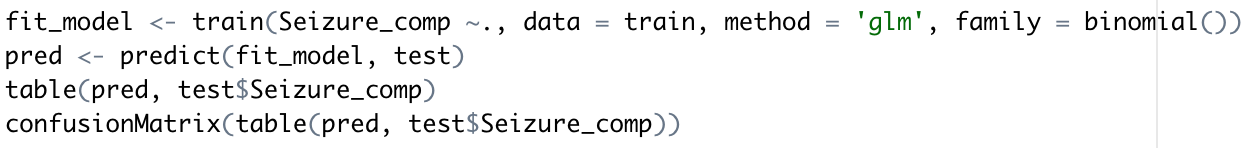
The logistic regression model was created using the glm() function with the following code: log <- glm(Seizure\_comp ~., data = df2, family = binomial()). The summary() function was then used to see the results of the model previously created. This is shown in Figure below.

Table

Description automatically generated

Figure 6

From the above table, it was determined that race, ethnicity, and length of hospital stay all have small statistical significance in regard to 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 Figure 7 below.



Text

Description automatically generated with medium confidence

Figure 7

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.

**Conclusion and Next Steps**

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 model can help provide healthcare works with information regarding who would be more likely to endure a seizure. It 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. Also, a decision tree would be a useful addition to this analysis to help healthcare workers in decision analysis regarding patient outlook.

**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.

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.