Final Report:

Diamond Data Analysis

**Introduction:**

According to the World Health Organization (WHO) strokes are the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Although strokes are such a common cause of death, according to the Center for Disease Control (CDC) 80% of strokes are preventable. Many strokes can be prevented through healthy lifestyle changes. We’ve obtained this dataset which includes patients and key information about them such as gender, age, various diseases, smoking status, and whether they’ve had a stroke or not. My goal for this dataset is to use the provided dataset to identify the key factors that are associated with individuals who have had a stroke and use this data to successfully predict which individuals are more likely to have a stroke. Potential stakeholders who could be interested in this model are Hospitals, Pharmaceutical companies, and at-risk individuals.

**Data:**

To conduct this analysis, obtained data around 5110 individuals. The data included a total of 12 columns with information around each individuals’ age, gender, medical information (BMI, average glucose level, etc.), personal information (work type, residence type, etc.), and whether they’ve has a stroke. 8 of the 12 features included were categorical features, while the remaining 4 were numerical features. Additionally noted that of the 5110 rows, 4.87% of the individuals have had a stroke.

**Data Wrangling:**

The raw dataset from Kaggle was fairly clear to begin with, however I did take some steps to get the data ready for the exploratory data analysis. I started with validating that the datatype of each column is accurate and does not need to be changed. As the data types appeared accurate, I moved on to check whether any columns have missing values. I determined that the only column with missing values is BMI, with a total of 201 missing values. Further noted that of the 201 missing values, 40 of them are patients who have had strokes. As there are only a total of 246 rows for patients who have strokes, dropping 40 of them may result in losing significant data, so I decided not to drop the data, instead I opted to replace the missing BMI values with the average BMI. I then checked all the numeric columns in the data to make sure that there are no errors, and all the numbers appears to be valid. There were 2 BMI data points that were outliers however as they did not appear to be due to any data errors, I decided to leave them in the data.Next, I filtered for all the categorical data and checked the values for this data to make sure there are no invalid inputs. Most columns appear to have valid inputs. I noted that there are 1544 Unknown values for 'smoking\_status'. As this is a significant number of columns, I cannot drop them, instead I decided to keep them as “unknown”.

**Exploratory Data Analysis:**

In order to further explore the data, I created charts to visualize each feature. For each categorical feature, I created bar charts to visualize the distribution with the overall data as well as just for the individuals who’ve had a stroke. Through conducting this analysis, I noted the following:

***Figure 1:*** *Overall distribution on top, distribution amongst individuals who’ve had strokes on the bottom.*

A graph of different types of diseases

Description automatically generated with medium confidence

**Figure 1:** From these charts, we noted that although the overall data has significantly more individuals with no heart disease and hypertension, of the individuals who’ve had strokes, the majority have heart disease and/or hypertension.

***Figure 2-4:*** *Overall distribution on top, distribution amongst individuals who’ve had strokes on the bottom.*

*A comparison of a bar graph

Description automatically generated* A screenshot of a graph

Description automatically generated A close-up of a graph

Description automatically generated

**Figure 2:** From these charts, we noted that although females are in the majority in the overall data, among the individuals who’ve had strokes, males are in the majority by a slight amount.

**Figure 3:** Additionally, noted that the most common job for individuals who’ve had a stroke is self- employed. As opposed to Private being the most common by far in the overall distribution.

**Figure 4:** Noted that, Never Smoked is in the majority in the overall distribution, however formerly smoked is in the majority among individuals who’ve had a stroke.

I then created a histogram and box plot for each numerical feature to see the overall distribution and compare the distribution for individuals who’ve had a stroke against the individuals who haven’t had a stroke. Through conducting this analysis, I noted the following:

***Figure 5-7:*** *Overall distribution on top, distribution of strokes vs no strokes on the bottom.*

A screenshot of a graph

Description automatically generated *A screenshot of a graph

Description automatically generated* *A graph of a number of objects

Description automatically generated with medium confidence*

**Figure 5:** Upon looking at these charts, noted that the average age for the individuals who’ve had strokes is higher than the average age for the individuals who haven’t had a stroke.

**Figure 6*:*** Per inspection of these charts, noted that the overall distribution appears to lean to the left. Additionally noted that the glucose level for individuals who’ve had a stroke tends to be higher than individuals who haven’t had a stroke.

**Figure 7:**The minimum BMI for individuals who’ve had a stroke is higher than for the individuals who have not had a stroke.

**Data Modeling:**

In order to prepare the data to be fit to a model, I first used the getdummies function on the categorical features in the data in order to create dummy variables. I then used the standard scalar function on the numerical features in the data in order to scale the data. Next, I created the X and y variables and then split them into test and train data sets, doing an 80/20 split.

I then performed the GridSearchCV function for 3 different types on models to see which hyperparameters would perform the best for each model type. The metric that I used to test the performance of each model is the Recall score. As the main purpose of this model is to accurately predict when people have strokes, we primarily want to avoid false negatives and do not care about false positives and true negatives as much.

The first model that I tested was the regression model. Following were hyperparameters that I compared using the Grid Search function: solver = lbfgs, liblinear, newton-Cholesky & C = 1, 0.5, 0.1, 0.01, 0.001. I plotted the results of the grid search in a line plot.

***Figure 8:***

A graph with numbers and lines

Description automatically generated

As a result of the Grid Search, noted that the liblinear solver with a C value of 0.001 had the best results. Which was a Recall score of 0.8561.

The second model that I tested was the Decision Tree model. Following were hyperparameters that I compared using the Grid Search function: criterion = gini, entropy, log\_loss & max\_depth = 2,3,4,5,6. I plotted the results of the grid search in a line plot.

***Figure 9:***

A graph with a line and a line

Description automatically generated with medium confidence

As a result of the Grid Search, noted that the gini criterion with a max depth value of 2 had the best results. Which was a Recall score of 0.8889.

The third model that I tested was the Random Forest model. Following were hyperparameters that I compared using the Grid Search function: criterion = gini, entropy, log\_loss, max\_depth = 2,3,4,5,6 & n\_estimators = 10,25,100,150. The results showed that the entropy criterion produced the best results. I plotted a chart comparing the various max\_depth and n\_estimator values for the entropy model.

***Figure 10:***

A graph with a line and a line

Description automatically generated with medium confidence

As a result of the Grid Search, noted that the entropy criterion with a max depth value of 2 and an n\_estimators value of 150 had the best results. Which was a Recall score of 0.8291.

Based on these results, noted that the model with the highest recall score is the Decision Tree model with a score of 0.8889. This model was able to predict 42 of the 51 instances of stroke accurately.

***Figure 11:***

A yellow blue and purple squares with white text

Description automatically generated

**Assumptions, Limitations, and Disclaimers:**

Upon conducting this analysis, I noted a few assumptions that we had to make about the data as well as limitations on what conclusions we could draw based on the data we had.

In order to judge performance of this model, We are assuming that the individual in the dataset who have not had a stroke, will not have a stroke in the future. In reality, there is a chance that individuals who have not had a stroke, may have one in the future, which could affect the performance of the model.

I also noted that there was a large number of unknown values for the smoking status column - it was the second most frequent response. As there was an extensive number of unknowns, we were unable to drop those values. Having more reliable data around the smoking status for each individual would improve the model.

Another limitation that I noted throughout my analysis is that the data around individuals who have had strokes is limited as there are only 51 lines of data as opposed to the 971 lines of data for individuals who have not had strokes in the testing data. A more balanced dataset would help improve the model.

**Recommendations:**

Following are some recommendations on ways that this model can be utilized:

* Hospitals can use this model to identify individuals who are at significant risk for having a stroke and can provide those patients with targeted preventative treatments.
* Pharmaceutical companies can also partner with hospitals and use this model to identify individuals who may be candidates for clinical trials for medications that prevent strokes.
* This model could be used by consumers directly to identify whether they are at a high risk for getting a stroke. If they are at a high risk, they can start implementing lifestyle changes to lower their chances of getting a stroke.

**Next Steps:**

There are ways that I noted that the model can be improved. Upon checking the features importance for the final model, we noted that this model is only using the age column to make predictions. This heavy reliance on the “age” column may be overfitting the model.

Potential Solutions:

* Remove the “age” column completely and train the model with the other features.
* Filter data to only include individuals that are the ages that people typically get a stroke. Then remove the age column and rely on other features to train the model.
* Remove the age column to train a separate model, then use the best parameters from that on this model that includes age.