Lukas Kania
Data 1030
Andras Zsom
October 10, 2021

Introduction:
In this project, I will investigate data collected from patients to determine whether a patient had a stroke or not. This dataset provides insights to what the World Health Organization calls the second highest cause of death in the world (Kaggle). This classification problem can be used to help anticipate a patient's health outcomes and inform health providers on certain indicators that can lead to a stroke.

This project will include multiple facets of a patient's health that all impact the stroke positivity or negativity. These features include metrics such as age and gender as well as more in-depth health related metrics such as presence of hypertension and heart disease. Other factors beyond bodily health are included with metrics such as residential type, marital status, and work type. All these metrics provide insights to a patient's overall health and will be used to determine their risk for a stroke. The risk will be a binary scale of 1 and 0, with 1 indicating a having/had a stroke and 0 indicating not having/had a stroke. The dataset provides these metrics for 5110 patients. As discussed previously, the target variable will be the result of a patient having or not having a stroke in the form of a classification of a binary 1 and 0.

This Kaggle dataset has been used in multiple studies, including the paper in the IEEE called "Predicting Stroke from Electronic Health Records." They used this dataset to do just what the title describes. They performed a principal component analysis of the features to see what datapoints contributed strongly towards a stroke. Another paper, "Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients" does not explicitly use the Kaggle set but provides insights to machine learning and the prediction of strokes. I will reference both papers throughout this project.

Feature Descriptions:
ID: Unique Integer identifier of a patient that provides no insight other than an identifier of a patient. Will not be used in analysis. Continuous column
Gender: String values that identify the gender of a patient with values of "Male", "Female", "Other". Categorical column
Age: Float values that identify the age of a patient. Continuous column
Hypertension: Integer values that describe whether a patient has hypertension or not. 1 = hypertension, 0 = no hypertension. Categorical column
Heart_Disease: Integer values that describe whether a patient has hypertension or not. 1 = heart disease, 0 = no heart disease. Categorical column
Ever_Married: String values that describe whether a patient has ever been married. "No" = never married, "Yes" = has been married. Categorical column
Work_type: String values that describe the type of work a patient does for a living. Values: "Children", "Govt_jov", "Never_worked", "Private", "Self-employed". Categorical column

Residence_type: String values that describe the type of residence that a patient resides in. Values: “Rural”, “Urban”. Categorical column
Avg_glucose_level: Float values that indicate the average glucose level of a patient. Continuous column
Bmi: Float values that indicate the body mass index of a patient. Continuous column
Smoking_status: String values that indicate the smoking habits of a patient. Values: “formerly smoked”, “never smoked”, “smokes” and “Unknown”. Categorical column

EDA:

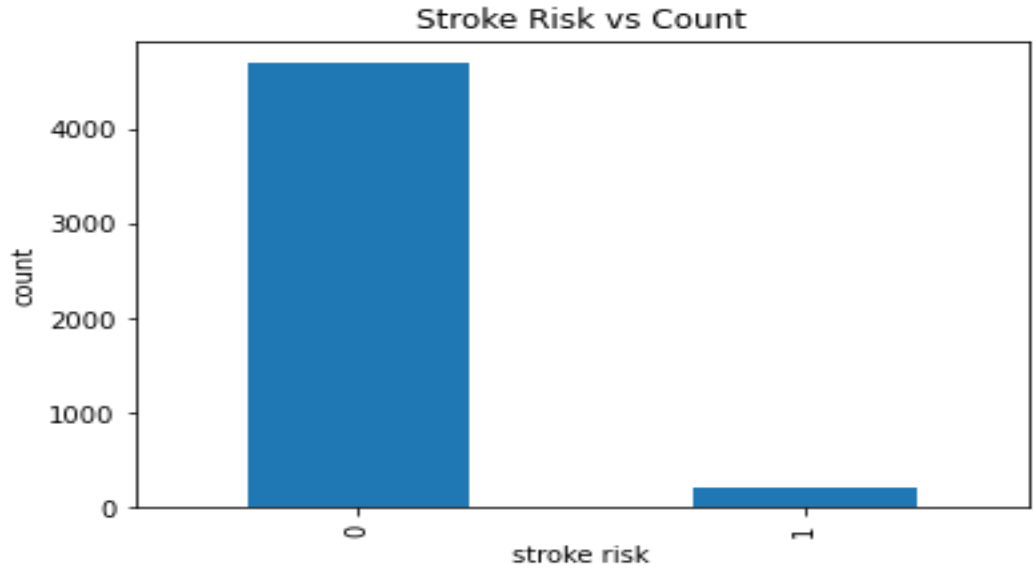

Figure 1: This bar plot compares our target variable outcomes. A patient either hasn’t had a stroke (“0”) or has had a stroke (“1”). This plot shows that our data is unbalanced and therefore the splitting of our data will need to be different from a standard “train_test_split().” This is discussed further in the “splitting the data” section.

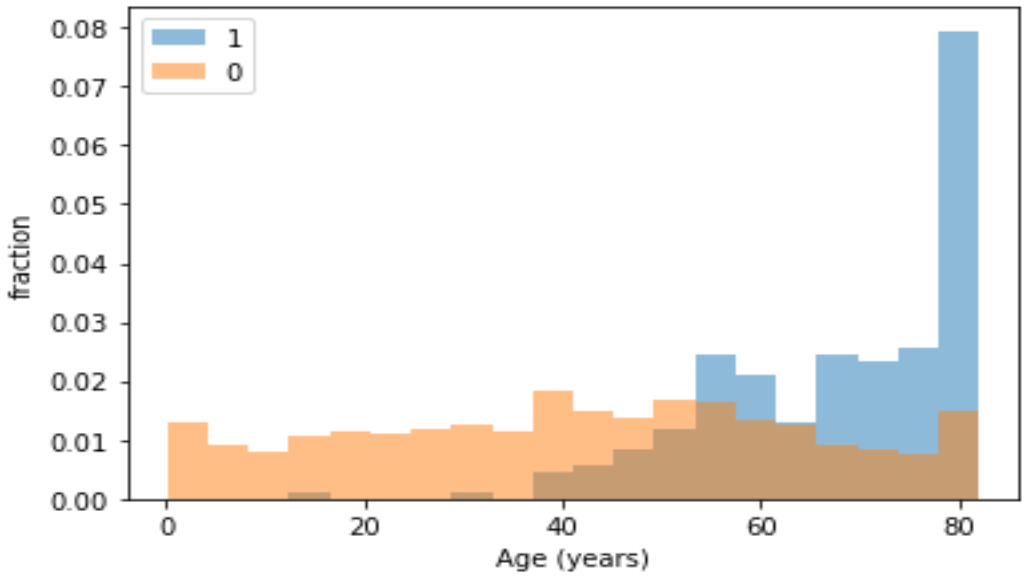

Figure 2: This histogram shows how the age of patient compares to their risk of having a stroke. We can see that there is an even spread throughout all ages if someone had not had a stroke. However, if a patient has had a stroke, they tend to be above 40 years old.

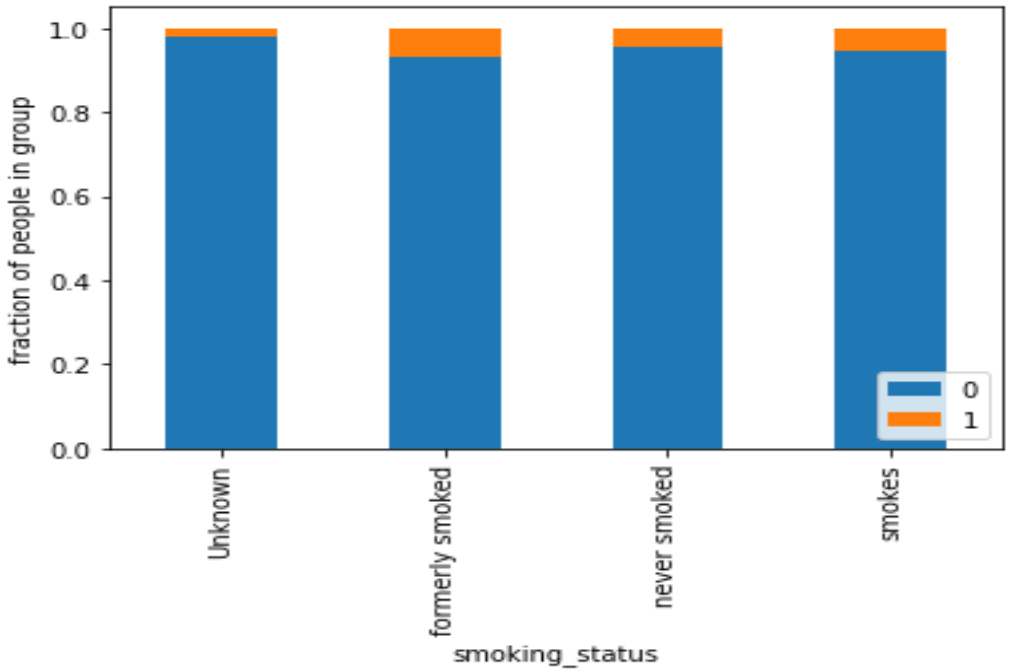

Figure 3: This stacked bar plot compares the risk of a stroke verses the smoking status. As described in the dataset description, “Unknown” means information about their smoking habits was not provided by the patient. I expected this to show more risk for former smokers and current smokers, but this is not as drastic as I anticipated.

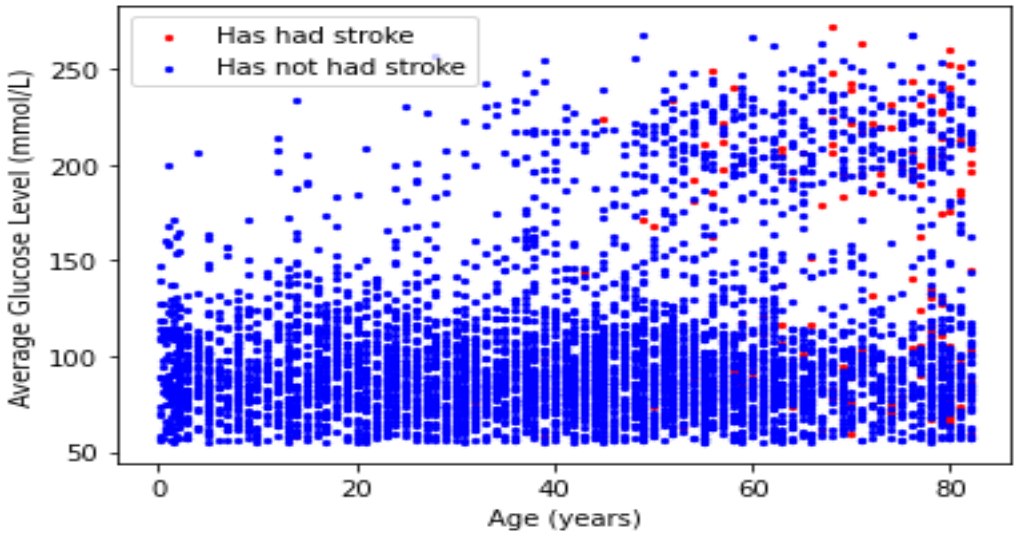

Figure 4: This plot shows the compares the continuous variables of age and average glucose level. It also makes the distinction between patients that have had a stroke and have not. We can see that being older in age has an impact of the risk of having a stroke. Average glucose level seems to be more spread-out regarding likelihood of a stroke.

Splitting the Data:

I decided to use a stratified kfold split on my data since it is quite unbalanced. Only 209 patients have been labeled a stroke risk. This is only 4.26% of the datapoints. I believe this data is IID because these are individual patients. This dataset is also not a timeseries as there is no information to tell when this data was collected. I had skepticisms about using stratified kfold split, but it seems to be the best option. I also utilized an 80/10/10 split due to the size of the dataset. There are only 5110 datapoints, 201 of which were eliminated because they did not provide a body mass index for the patient. The reason I chose to eliminate these datapoints is discussed below in the preprocessing section. A 98/1/1 split would not provide enough datapoints to the testing and validation sets.

Preprocessing:

After eliminating the ID feature column, 10 features will be used to predict the target variable. Since “id” is a given identifier for a patient and provides no information for predicting stroke. I used the StandardScaler preprocessor for the age column because this continuous feature roughly follows a normal distribution. I then used OneHotEncoder for the "work_type", "smoking_status", 'gender', 'ever_married', and 'Residence_type' features because these categorical columns needed to be given dummy variables so the machine learning pipeline would function. I felt that all these features could not be ordered so I moved away from the OrdinalEncoder. As I mentioned above, I decided to eliminate the 201 rows that were missing values for body mass index. I did this because this is data about the healthiness of a patient, and we are classifying their outcome of having a stroke. This is very sensitive data and imputing a value could make our classification less accurate. Since these datapoints only account for 3.93% of the dataset, it should not impact the results.

References:

C. S. Nwosu, S. Dev, P. Bhardwaj, B. Veeravalli and D. John, "Predicting Stroke from Electronic Health Records," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019, pp. 5704-5707, doi: 10.1109/EMBC.2019.8857234.

M. Monteiro et al., "Using Machine Learning to Improve the Prediction of Functional Outcome in Ischemic Stroke Patients," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 15, no. 6, pp. 1953-1959, 1 Nov.-Dec. 2018, doi: 10.1109/TCBB.2018.2811471.

Kaggle, https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

Github Repo: https://github.com/lkania1/data1030project