Stroke Prediction

Technical Data Analysis
by Nathan Billings

About the Data (Intro and Background)

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

The dataset provided is a training set for machine learning and as such, the original source remains confidential

The features included in this dataset are typically associated risk factors of a stroke (Age and Marital Status) or comorbid conditions (Heart Disease and Hypertension)

Initial Data Inspection

The dataset was fairly clean as-is

Inspecting the data showed several known comorbid conditions, like heart disease or hypertension, with strokes

Factors such as marriage and smoking were also suspected of being related to strokes

Data Cleaning

In [21]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5110 entries, 9046 to 44679
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   gender             5110 non-null   object 
 1   age                5110 non-null   float64
 2   hypertension       5110 non-null   int64  
 3   heart_disease      5110 non-null   int64  
 4   ever_married       5110 non-null   object 
 5   work_type          5110 non-null   object 
 6   Residence_type     5110 non-null   object 
 7   avg_glucose_level  5110 non-null   float64
 8   bmi                4909 non-null   float64
 9   smoking_status     5110 non-null   object 
 10  stroke             5110 non-null   int64  
dtypes: float64(3), int64(3), object(5)
memory usage: 479.1+ KB
```

Explanation of Cleaning Steps

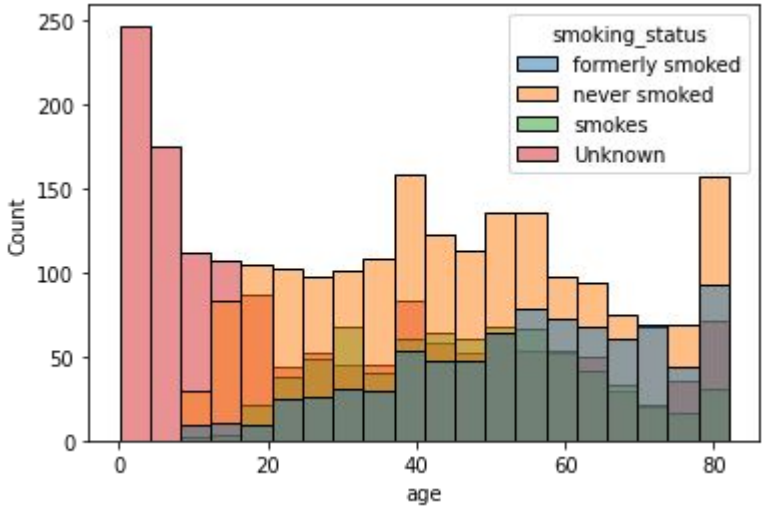
- No features were dropped
- No duplicate rows were found
- BMI had some null values
  - Decision to drop rows was made on the basis of it being minimal loss of 201 rows
- One row had Gender listed as Other and was dropped due to sample size
- Large quantity of Unknown smokers (1,483) were found but kept in the hope of a correlation to another feature, like Age, was found

Large Number of Unknown Smoking Status

The data source mentioned patients with an “Unknown” smoking status literally meant the data was unavailable for the patient

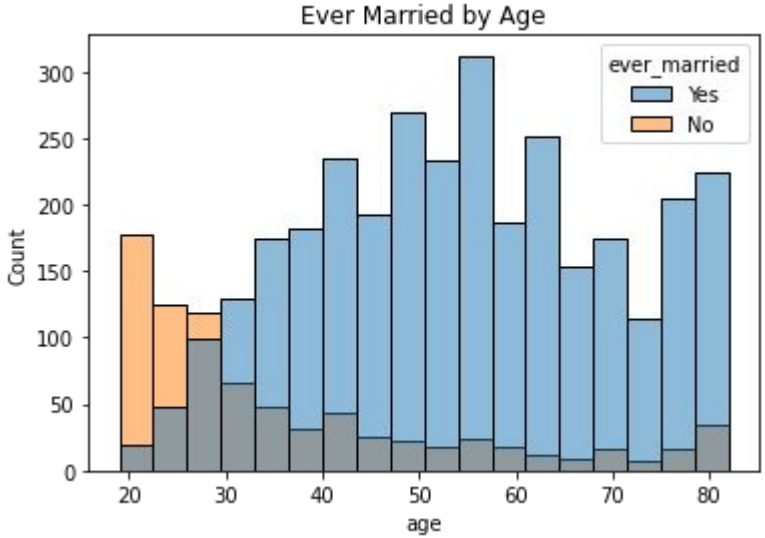
Plotting a histogram showed the majority of Unknowns in this feature were presumably too young to smoke

There is possible value for creating a new category for “too young” and selecting for anyone who is 16 or younger



Marriage Status

Binary classification less useful than hoped due to availability of more defined categories of Marital Status (widowed, divorced, remarried, etc) in medical literature about strokes



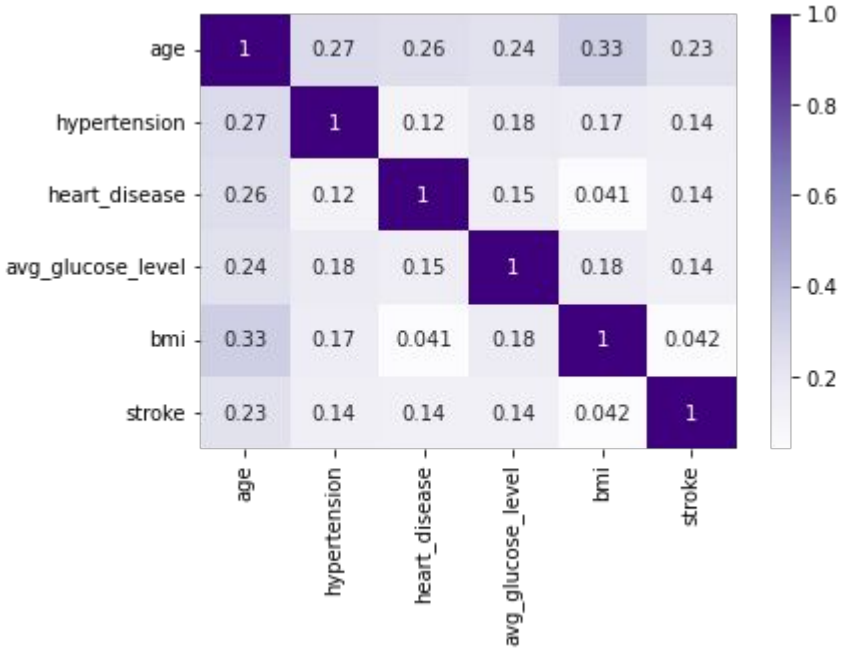
Plot filtered for 18+

Weak Correlations Across Features

Weak correlations most likely due to target imbalance

Age had the strongest correlation to strokes but was also correlated to other features just as easily

Oversampling or undersampling might be useful to draw stronger correlations



Interesting Challenges

A large quantity of Unknowns in Smoking Status prompted multivariate analysis to find a correlation with Age

Ever Married would have been more useful as Marital Status to include more than just a yes or no

Stroke target is unbalanced towards non-stroke patients and made multivariate analysis difficult

Undersampling or oversampling might show use before building machine learning models