Stroke Risk Prediction

Data Analysis and Machine Learning Model by Nathan A. Billings

About the Data

This data set was obtained from Kaggle and was a collection of diagnoses of various patients, with the intent to predict a stroke based on comorbidities such as hypertension or behaviors like smoking.

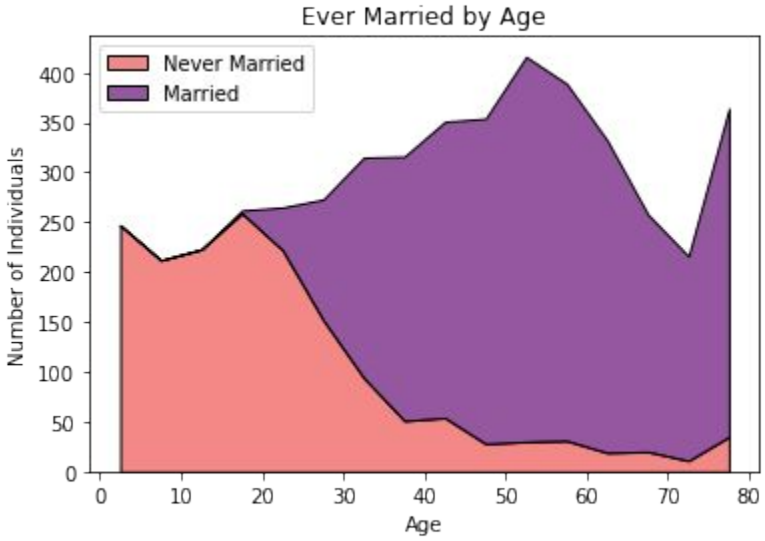
- Data features included age, gender, smoking status, if the patient lived in the city or rural area, if the patient has been married, BMI, average glucose level, hypertension, and heart disease
- Machine learning will hopefully be able to predict a stroke diagnosis
- There were some issues with the data

Marital Status Improvements

Marital status wasn't very useful as it was classified as "yes or no." Studies typically go into more granular detail when tying this feature to a stroke diagnosis.

- Better categories recommended (Unmarried, Married, Widowed, or Divorced)
- Marriage status at the time of stroke might also be useful

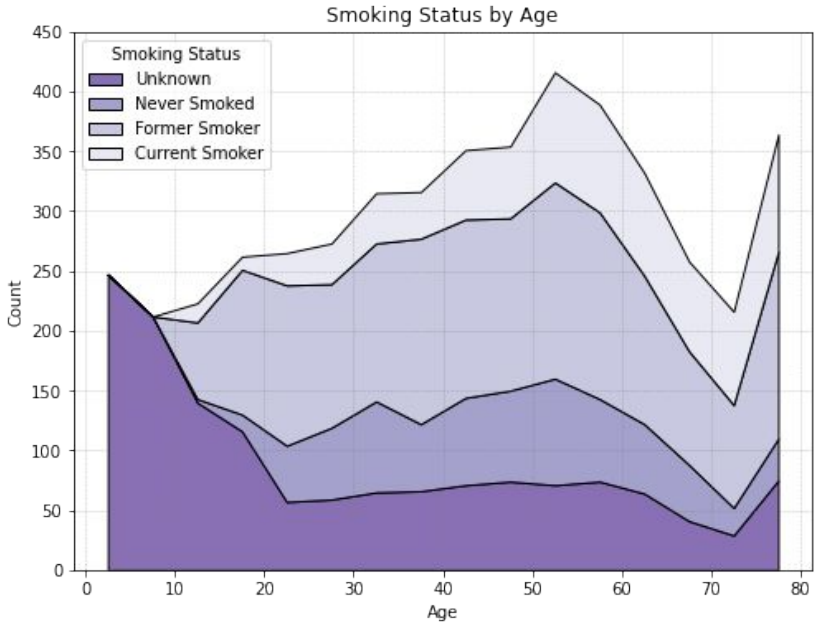
Study source:
https://heart.bmj.com/content/104/23/1937



Smoking Status: “Unknown”

There were some issues that came up while exploring the Smoking Status Feature

- A large quantity of “Unknown” smoking status
- Exploring the data showed a majority of “Unknown” were actually underage
- Better data collection might include “too young to smoke” or “prefer not to say”



Stroke Patients Under Represented

Challenges with predictions using this dataset are present because of the imbalance between stroke patients and non-stroke patients.

- More data is always better
- With current dataset, machine learning models are better equipped to predict if someone has NOT had a stroke

Logistic Regression Shows Promise

Logistic Regression was evaluated because of the relationship with between the risk factors in the data set and stroke diagnosis.

- The relationship is likely to be linear
- The diagnosis is categorical
- Logistic Regression model is not intensive to train or tune.

Model Evaluation

The model was evaluated using two metrics. Precision and Recall, or how well a model performs with avoiding false positives and how it performs avoiding false negatives, respectively.

- A Precision of 12 percent showed the selected model states a patient is likely to have a stroke when they have not had a stroke, 88 percent of the time.
- A Recall of 88 percent showed the selected model missed 12 percent of stroke patients in the predictions.

Final Recommendations

- Revisit data collection to get better categories of patients (Unknown smokers, Ever Married, etc.)
- Further experimentation with Neural Networks might yield better results
- More stroke patients would improve predictions