Prediction of Stroke Using Machine Learning

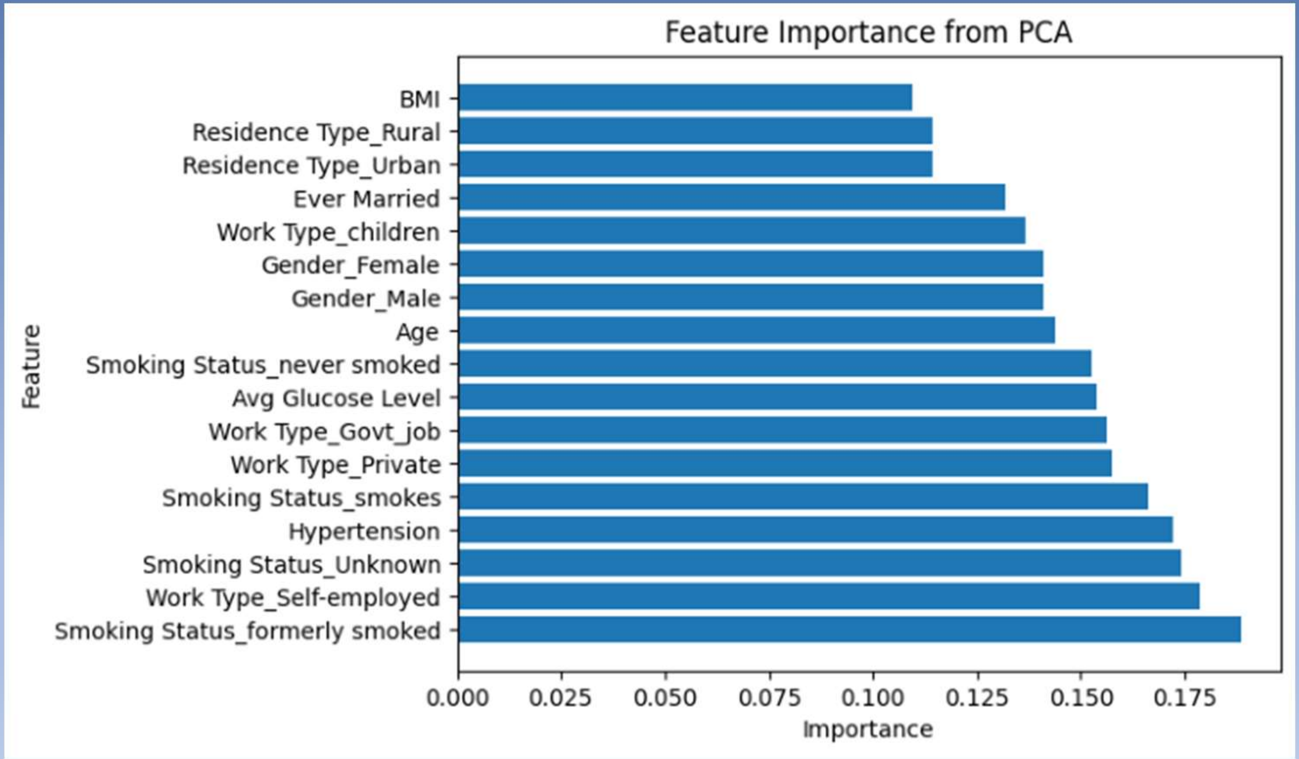
Author: Melissa Hoover

June 2023

Description of Problem

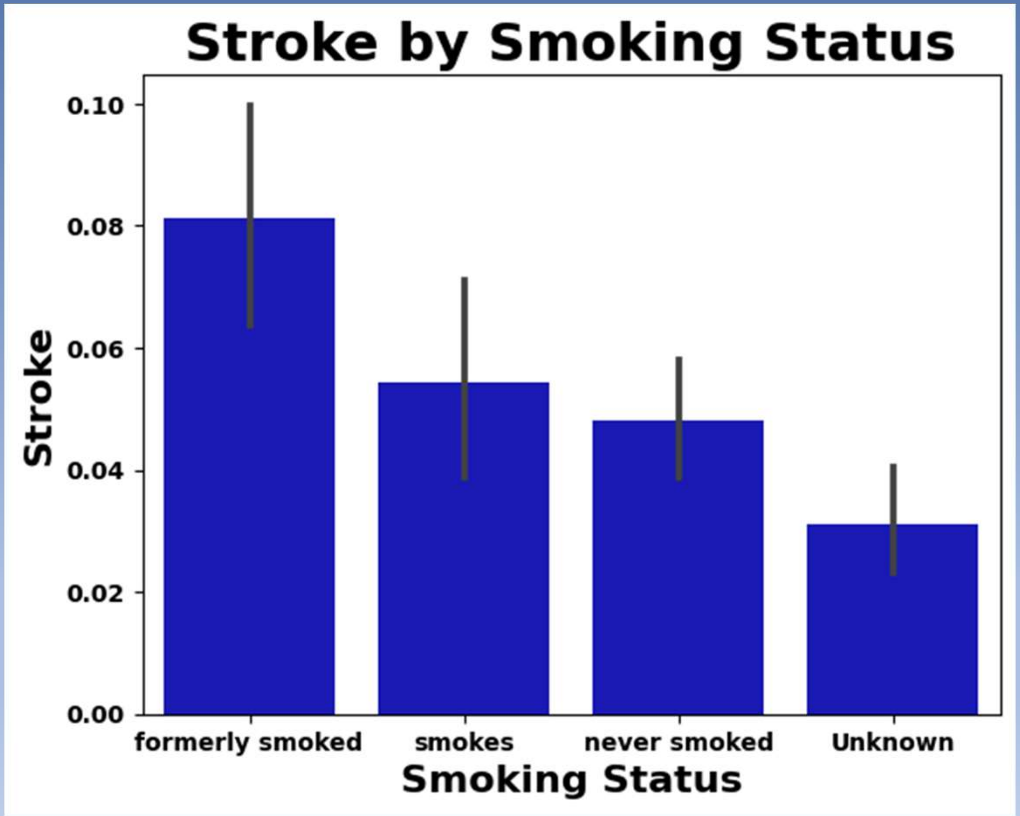
- This is a classification Machine Learning problem.
- We are predicting whether a patient is likely to have a stroke based on input parameters such as gender, age, hypertension, average glucose levels, BMI, work type, marital status, residence type, and smoking status.
- The output results will be either Yes for Stroke or No for No Stroke

Feature Importance from PCA



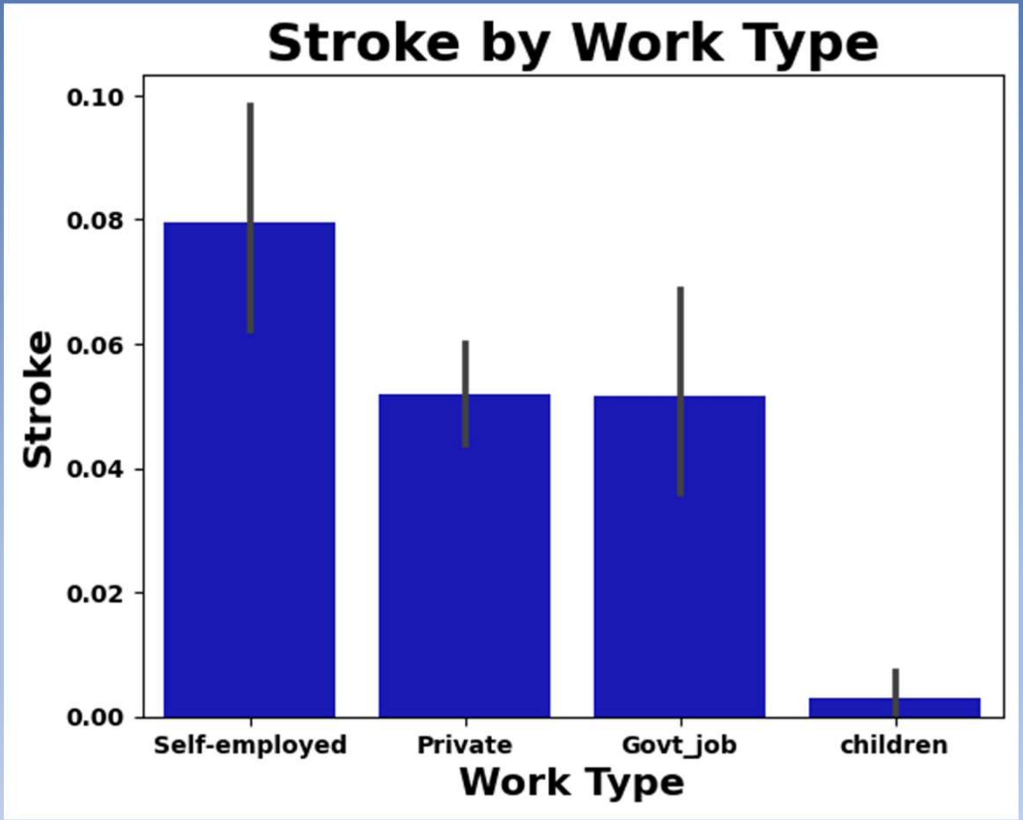
- BMI had the least importance
- Formerly Smoked had the highest feature importance

Stroke by Smoking Status



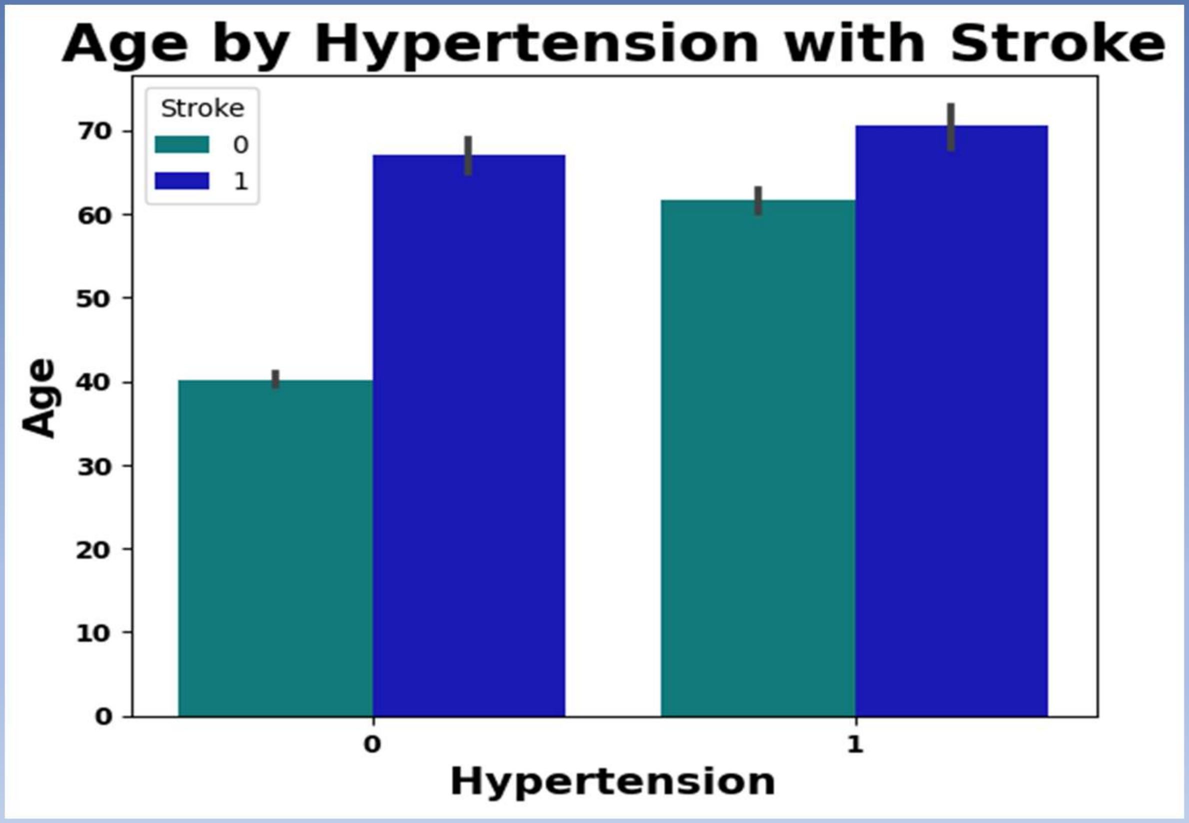
- Formerly smoked had higher incidence of stroke
- Unknown was the lowest
- Smoking can increase the risk of stroke by causing inflammation and damage to the blood vessels and can lead to a buildup of plaque in the arteries.

Stroke by Work Type



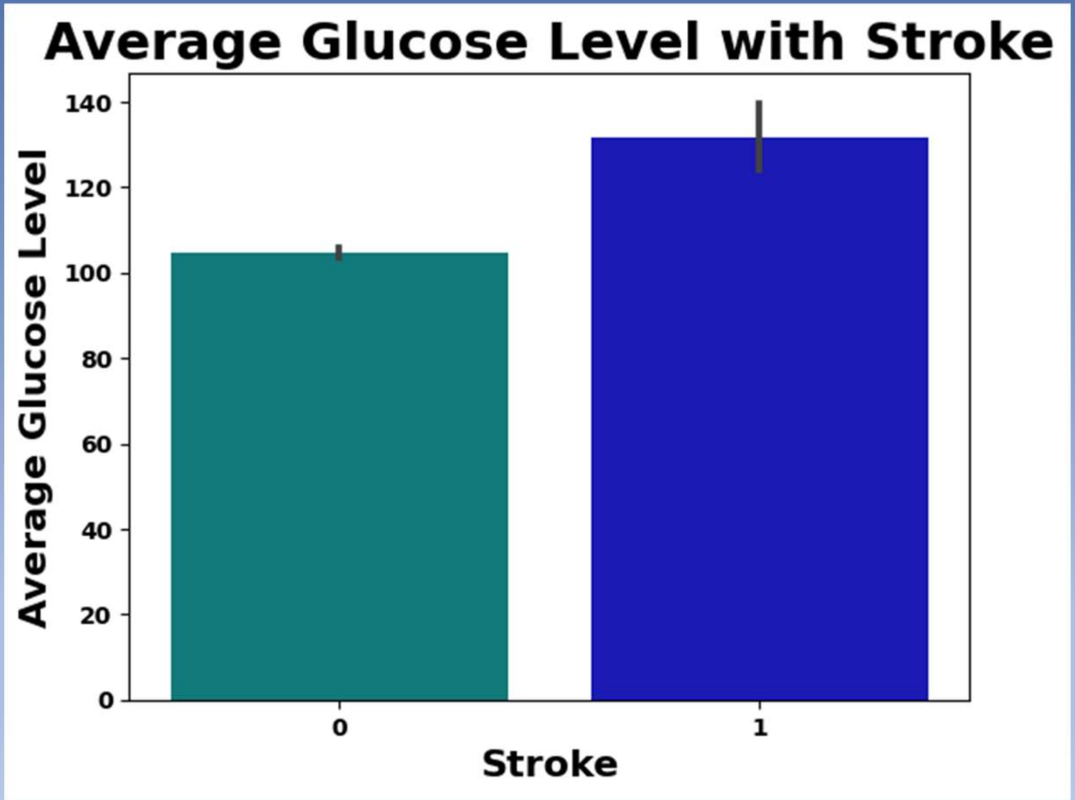
- Self employed work type had the most strokes and if they were stay at home parents that had the least strokes
- Self employed individuals could experience higher stress than other types of work. It could also be related to social support and health behaviors.

Age and Hypertension with Stroke



- Age had the highest correlation with stroke at 25%. Higher ages did result in strokes
- Hypertension was not highly correlated with only 13%. More strokes occurred in individuals with hypertension.

Average Glucose Level with Stroke



- Average Glucose Level is also not highly correlated with stroke at only 13% but there are slightly more strokes occurring with higher glucose levels.
- High average glucose levels can indicate diabetes or prediabetes.
- This can increase risk of stroke by damaging the blood vessels and preventing oxygen and nutrients from reaching the brain.

Machine Learning Models Used

- KNN
- Random Forest
- Logistic Regression

  - Principle Component Analysis and Feature Engineering were also applied and analyzed with each model

Model Limitations

- This is an extremely unbalanced dataset that affected our model's performance. I applied under sampling to account for the unbalanced dataset.

- Precision was low at 12%. This means that the model is not very accurate in predicting positive cases.

- False Positives 32%: The model incorrectly predicted that 32% of the patients had a stroke when in fact they did not have a stroke. This could lead to unnecessary treatment for patients who do not have a stroke.

- F1 score was 21%, although this is still low it is the best I was able to achieve. F1 is a combination of precision and recall.

Model Strengths

- Type II errors/False Negatives= 18%.
  - False negatives are the cases where the model predicted that the patient did not have a stroke when in fact they did have a stroke. This could mean that some patients who will have a stroke may not receive the appropriate treatment or precautions to prevent a stroke.

- Recall =82% of patients who were correctly identified by the model actually had a stroke.

Final Recommendations

- Model Recommended: Random Forest with Under Sampling gave the best results with the lowest False Negatives (Type II Errors) at 18%.