Prediction of Stroke Using Machine Learning

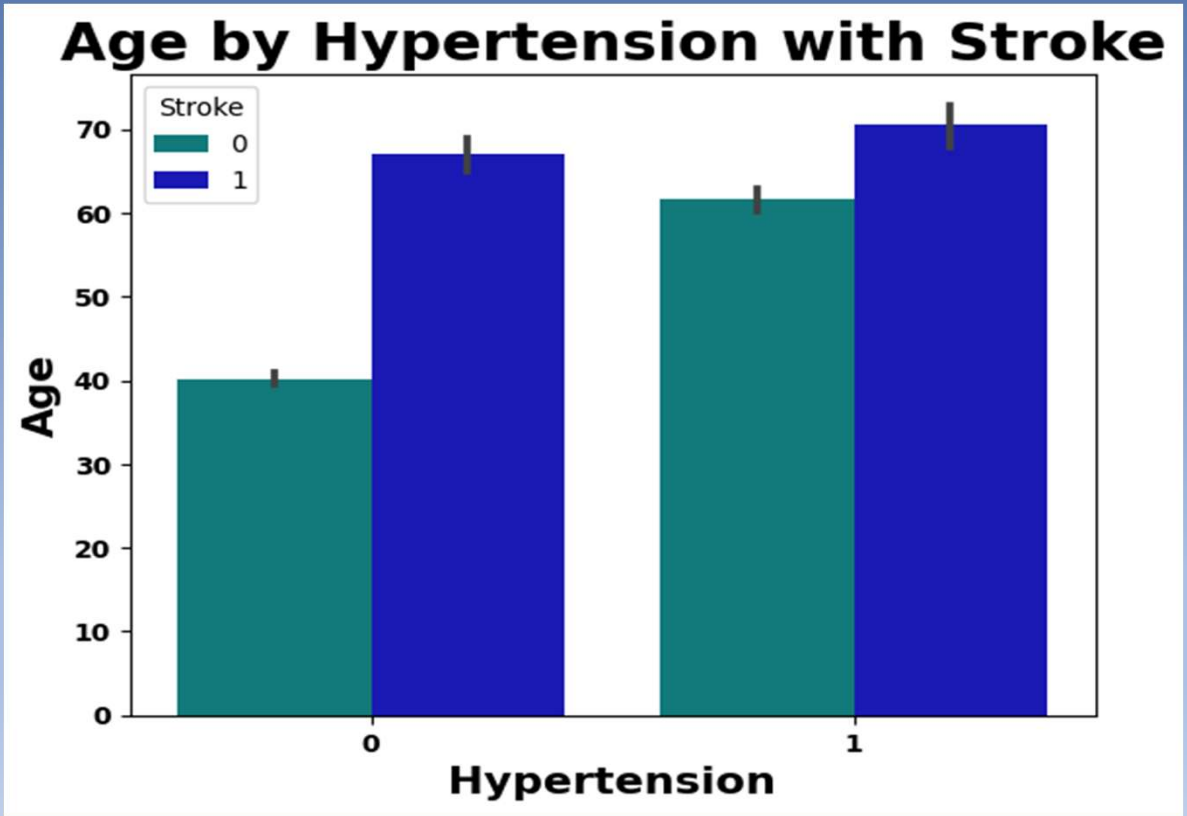
Author: Melissa Hoover

June 2023

Description of Problem

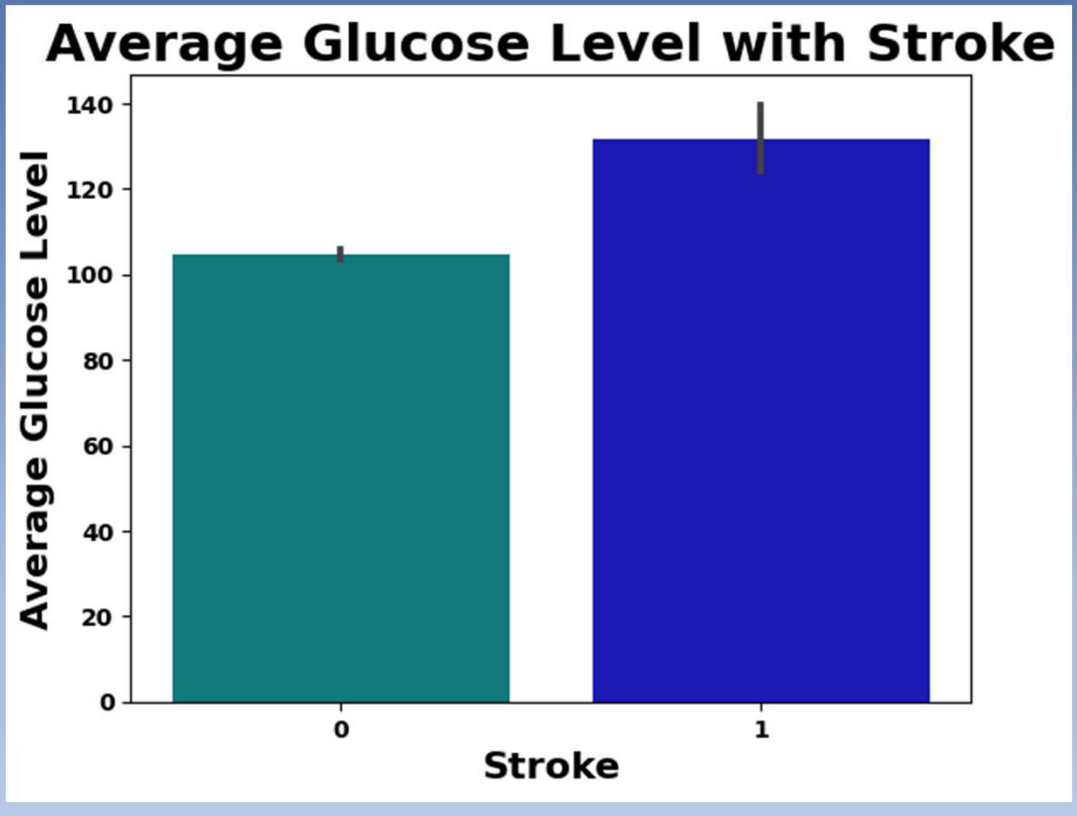
- This is a classification Machine Learning problem.
- We are predicting whether a patient is likely to have a stroke based on input parameters such as gender, age, hypertension, average glucose levels, bmi, work type, marriage status, residence type, and smoking status.
- The output results will be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

Age and Hypertension with Stroke



- Age had the highest correlation with stroke at 25%
- Hypertension was not highly correlated with only 13%

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.

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 14%. This means that the model is not very accurate in predicting positive cases.
- False Positives 26%: The model incorrectly predicted that 30% 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 24%, although this is still low it is the best I was able to achieve.

Model Strengths

- Type II errors/False Negatives= 19%.
    - 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 =81% of patients who were correctly identified by the model actually had a stroke.

Final Recommendations

- Model Recommended: Logistic Regression with Under Sampling gave the best results with the lowest False Negatives (Type II Errors) at 19%.