STROKE PREDICTION USING MACHINE LEARNING

BY FIYONA KRIEL

Image Source: https://medium.com/@jeffdm88/machine-learning-for-stroke-prediction-ab886a1f7828

Analyzing patient Medical and Lifestyle Data to Predict Likelihood of a Stroke

- Business Problem:
  - Use various individual medical and lifestyle data to predict whether a patient is likely to have a stroke
  - Accurate stroke prediction
    - Saves lives
    - Assist in recognizing contributing factors
    - Develop treatment plans and preventative measures
- Stakeholders:
  - Medical Team: Doctors, Surgeons (neurologists) and other support staff
    - Diagnosis, treatment and prevention plans
  - Medical Aid companies:
    - Identify at-risk clients
    - Tailor benefits and plans more effectively

Image Source: https://nutritionstudies.org/a-game-changing-solution-to-one-of-the-most-devastating-diseases-of-the-brain-stroke/

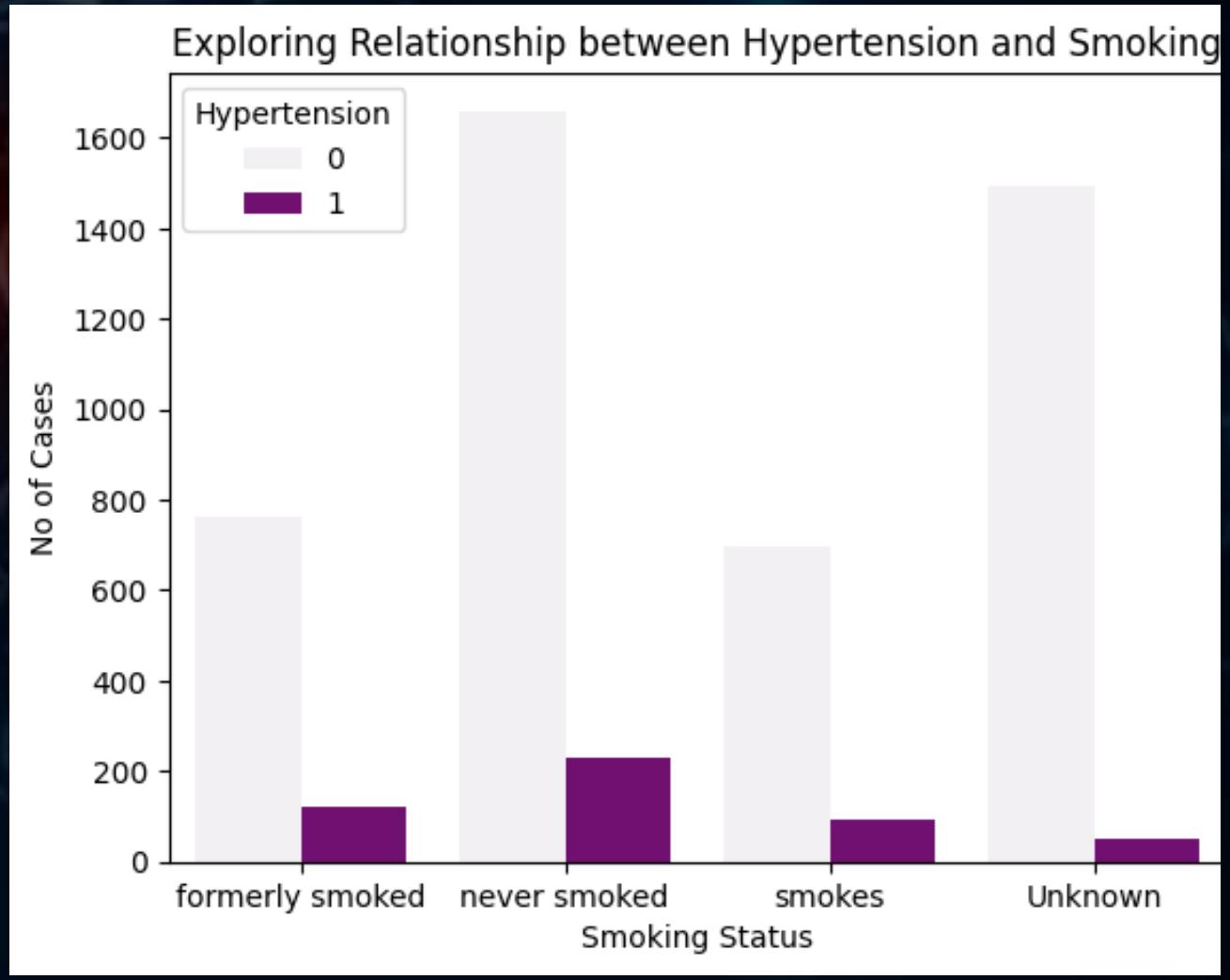

DATA

- According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally
  - Approximately 11% of total deaths worldwide
  - Predictions based on patients' medical and lifestyle factors
    - Gender
    - Age
    - Various diseases
    - Smoking status
- The target feature is stroke: 1 if the patient had a stroke or 0 if not
- Each row in the data provides relevant information about the patient.

ANALYTICAL INSIGHTS

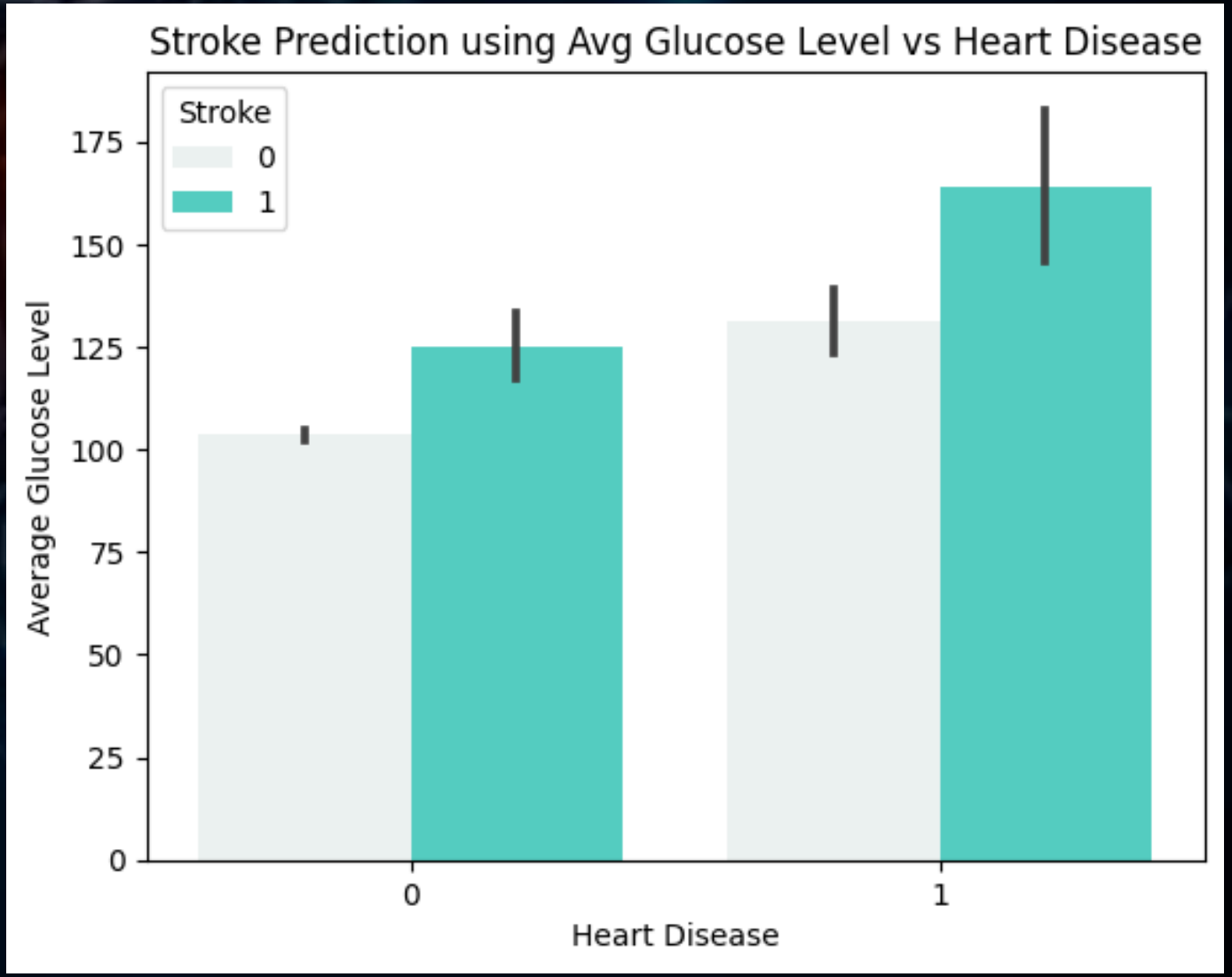
HYPERTENSION VS SMOKING

- Highest number of hypertensive cases presented in those who have never smoked
- Lower cases in current smokers and ex smokers.
- According to data, smoking does not play a big part in causing hypertension, if at all.



PREDICTING A STROKE USING GLUCOSE LEVEL VS HEART DISEASE

- Higher average glucose level seems to contribute to a stroke, regardless of presence of heart disease.
- Additionally, it should be noted that higher average glucose levels may also contribute to the development of heart disease in an individual.



DATA MODELING

- Model chosen based on 93% accuracy
  - 93% of predictions correct out of total number of predictions
- Highly imbalanced dataset
  - 95% - 5% split between no stroke vs stroke in initial data
  - Chose not to balance dataset at this point
- Key aim: fewer false negatives.
  - Patients who were predicted not likely to have a stroke, where in reality, their likelihood of having a stroke was quite high.
  - Detrimental to patient's diagnosis
  - Life threatening

SUMMARY AND RECOMMENDATIONS

- Various lifestyle factors within an individual's control that could be contributing factors
- Fewer false negative predictions to ensure patients get correct diagnosis and (life-saving) treatment
- Meet with stakeholders to find ways to improve data quality and class balance
- Good data is key – better data, better predictions
- These predictions will highlight patterns in health and lifestyle factors that could allow stakeholders to create effective prevention and treatment plans in order to reduce the overall number of stroke related incidents and/or deaths