PROJECT 2

STROKE PREDICTION

By Samar KRIMI

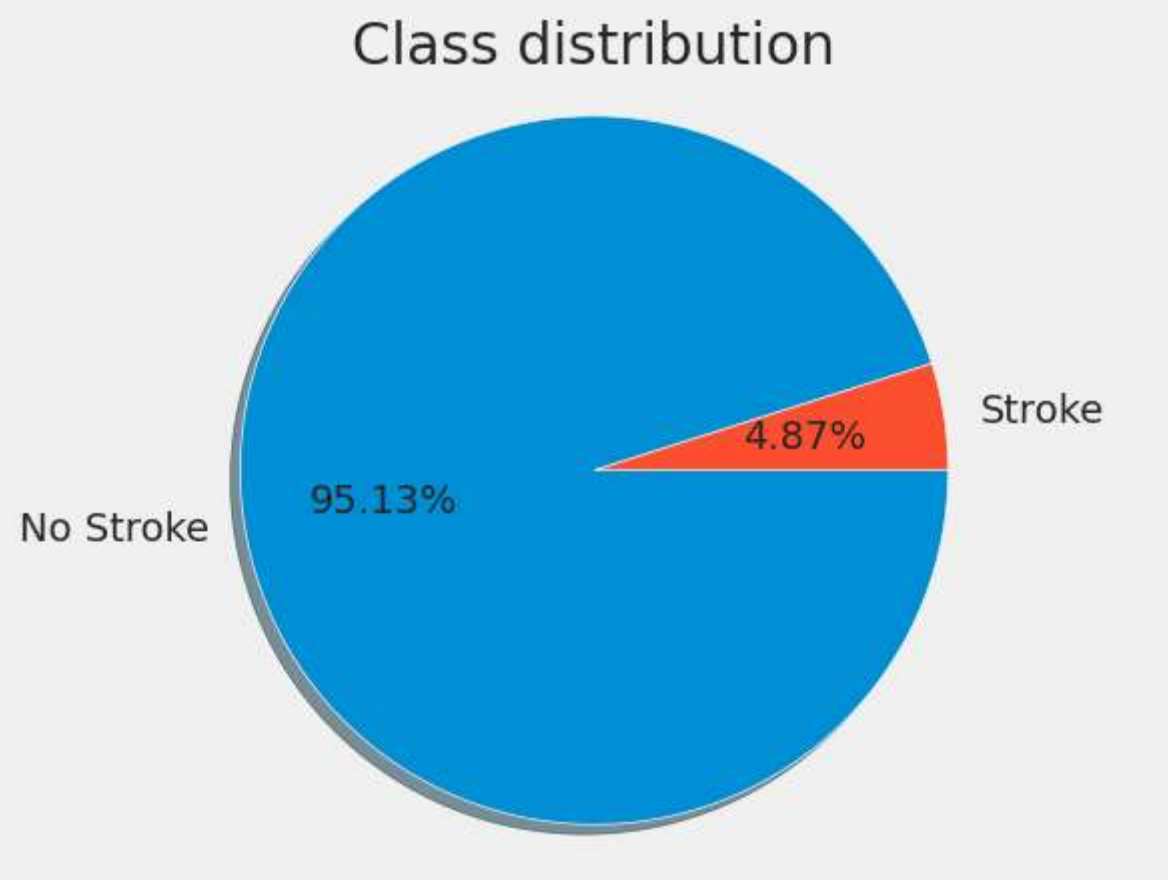
OVERVIEW

- This project is a healthcare prediction, stroke can be very hard to predict and therefore try to hinder, because it's the result of many different pathophysiologies.
- The task is to help doctors to predict a stroke for at-risk patients.
  - The stakeholders are the doctors that try to take care of the patients.
  - Their primary goal is to increase the stroke detection before it happens.
  - They plan to adjust their diagnosis.
  - They expect recommendations for which modifications they can make to increase the effectiveness of their diagnosis.

DATASET

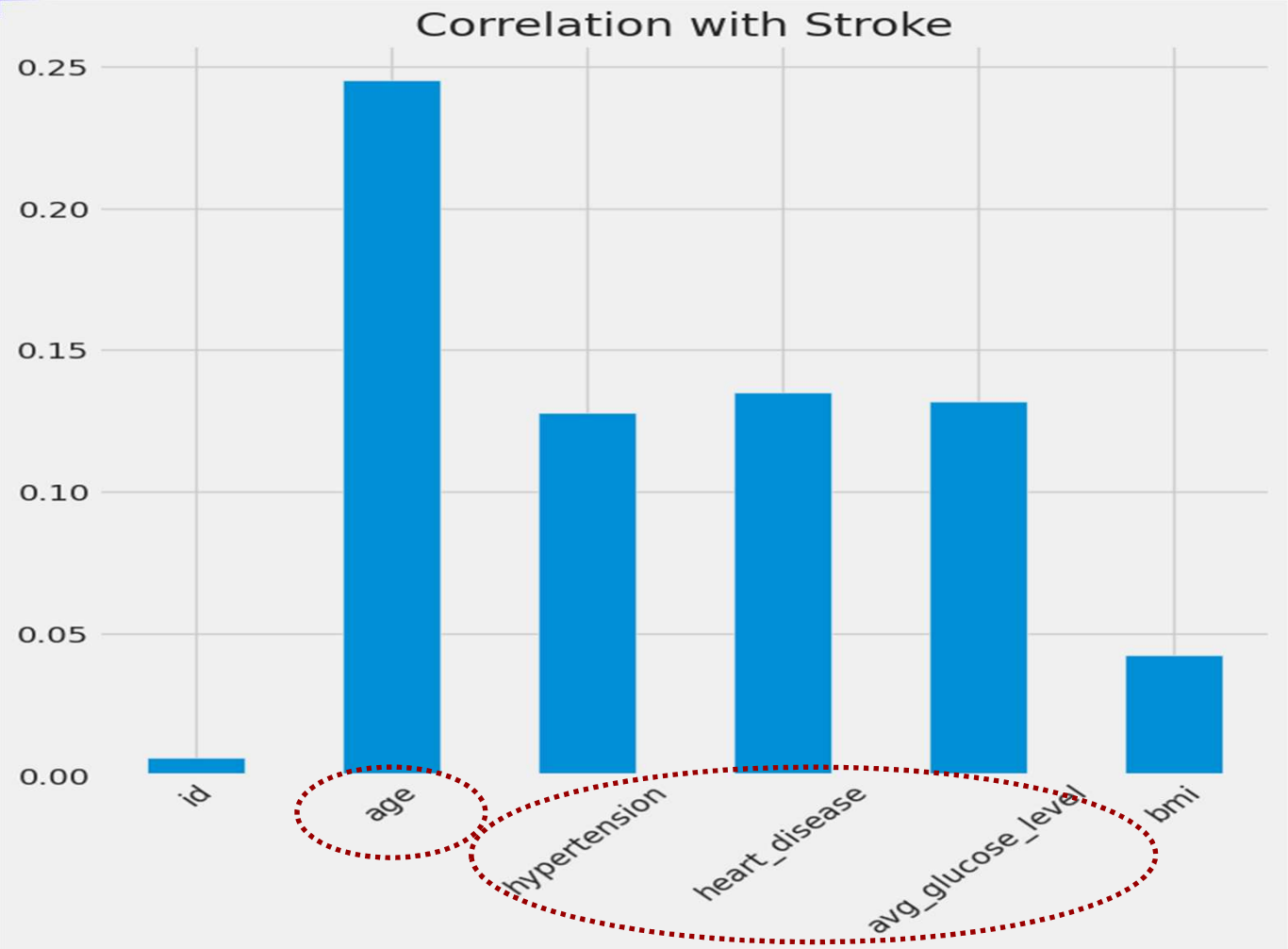
- This is a healthcare dataset used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status.
- The data is found from Kaggle (note that Kaggle is an online community platform for data scientists and machine learning passionates) : Let's understand what some columns tell us :
- bmi : body mass index, the normal index is between 18,5 and 25.
- avg_glucose_level : average glucose level in blood, the expected values for normal fasting blood glucose concentration are between 70 mg/dL and 100 mg/dL.
- hypertension : If the patient does not have hypertension, he has a great chance to avoid stroke.
- heart disease : If the patient doesn't have cardiovascular disease, he's more likely to avoid stroke.
- smoking status : patients how have never smoked are more likely to be spared from stroke although in some cases related to life quality they may develop stroke.

CLASS BALANCE

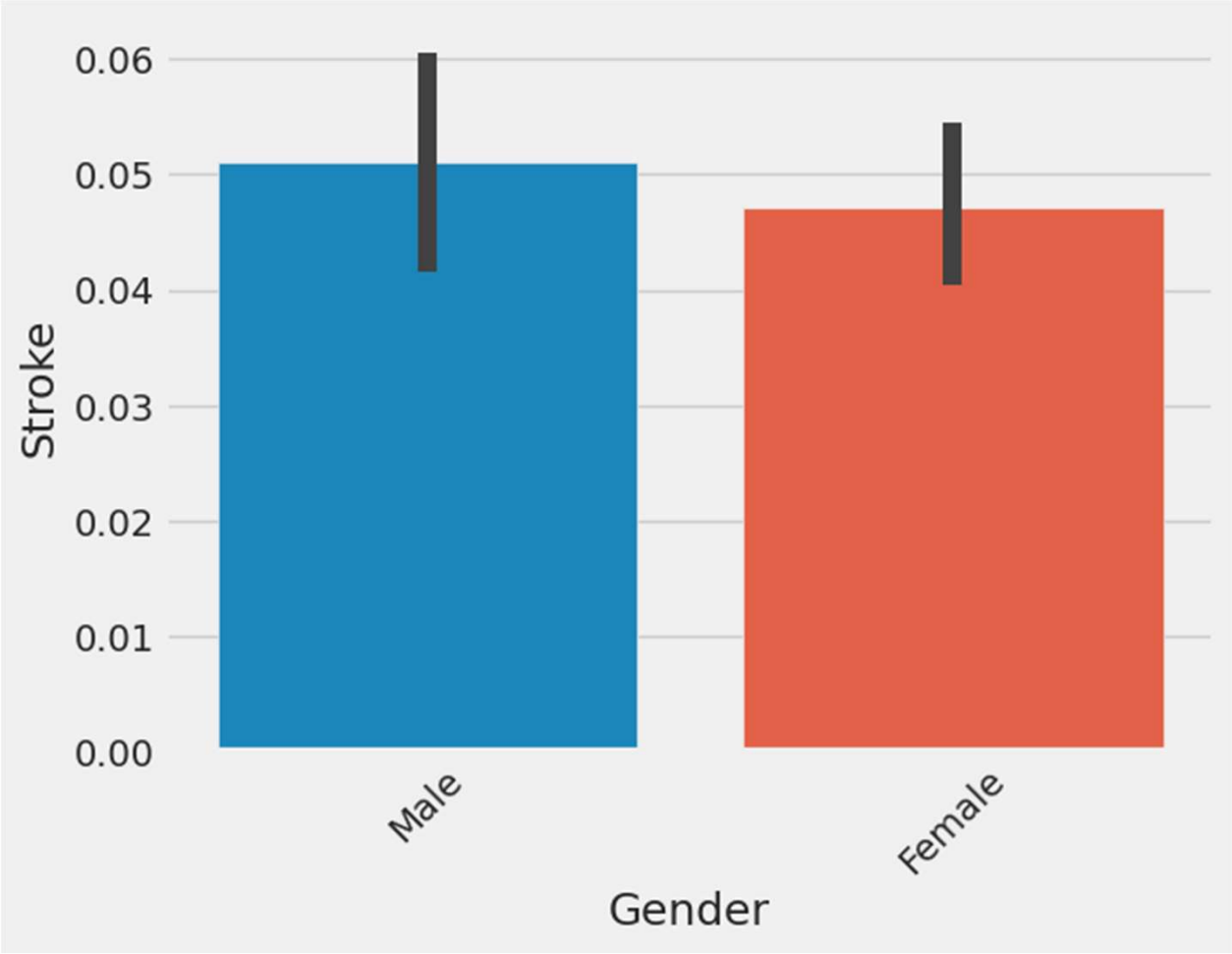


samples associated with no stroke: 4861
samples associated with stroke: 249

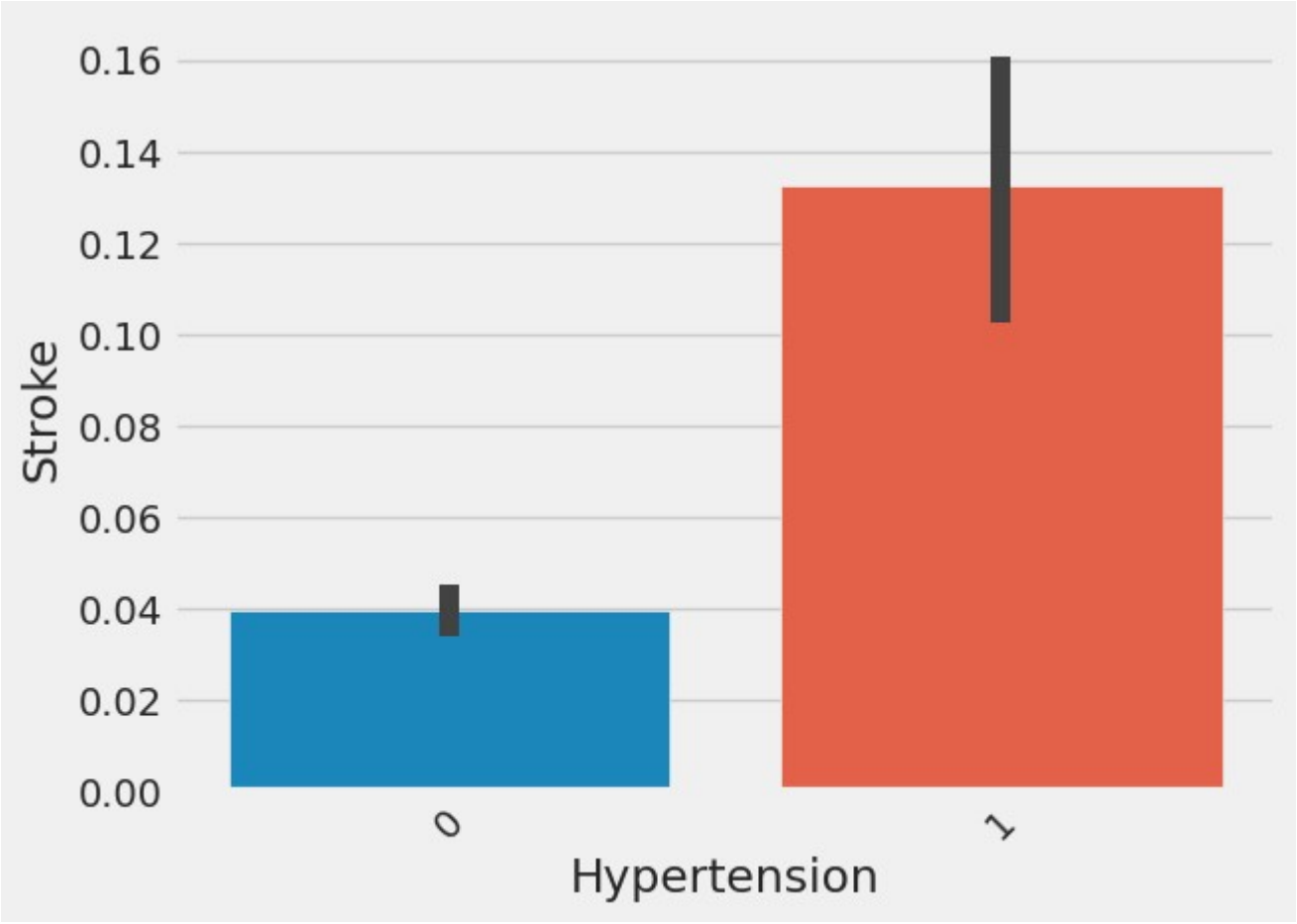
FEATURES CORRELATION



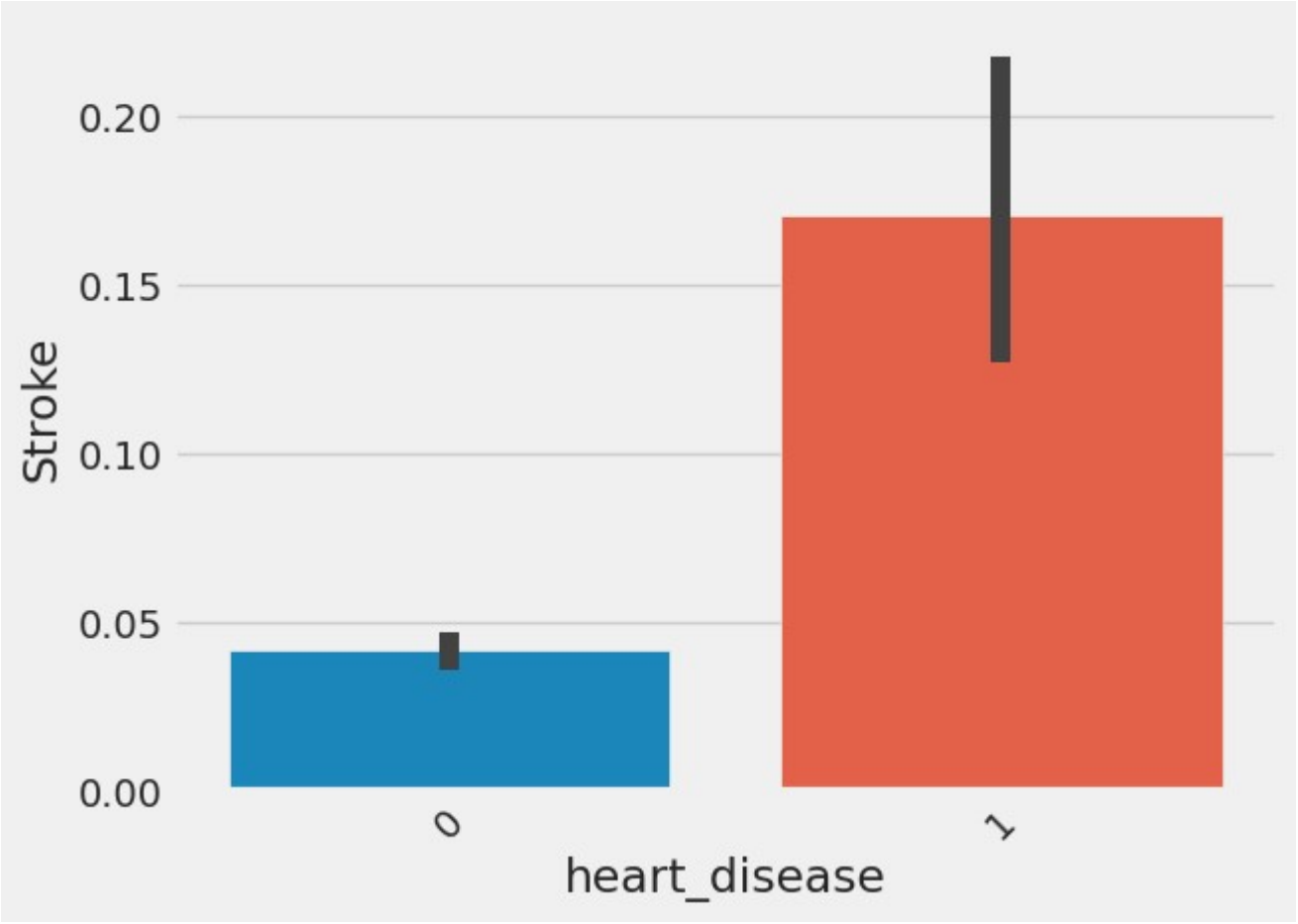
GENDER FEATURE



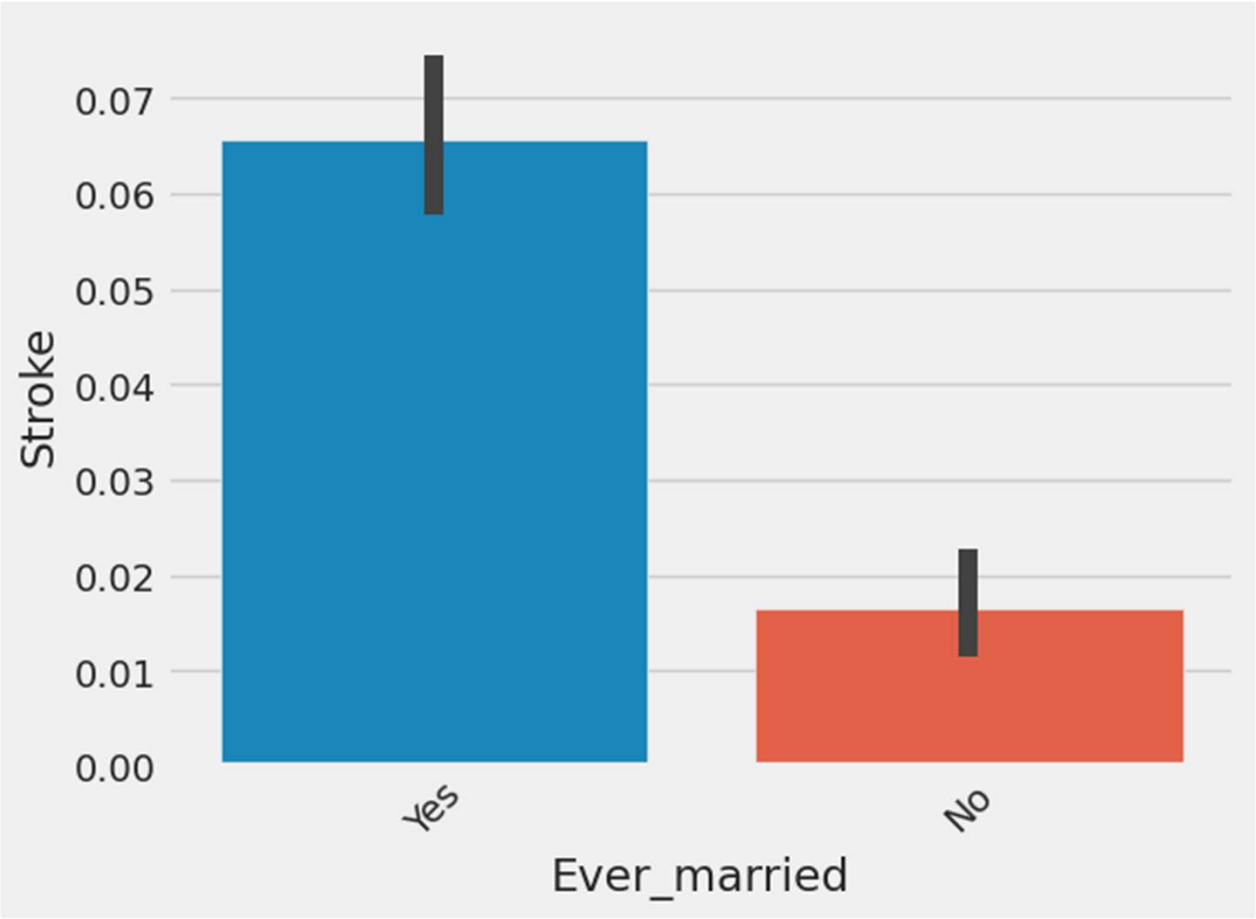
HYPERTENSION FEATURE



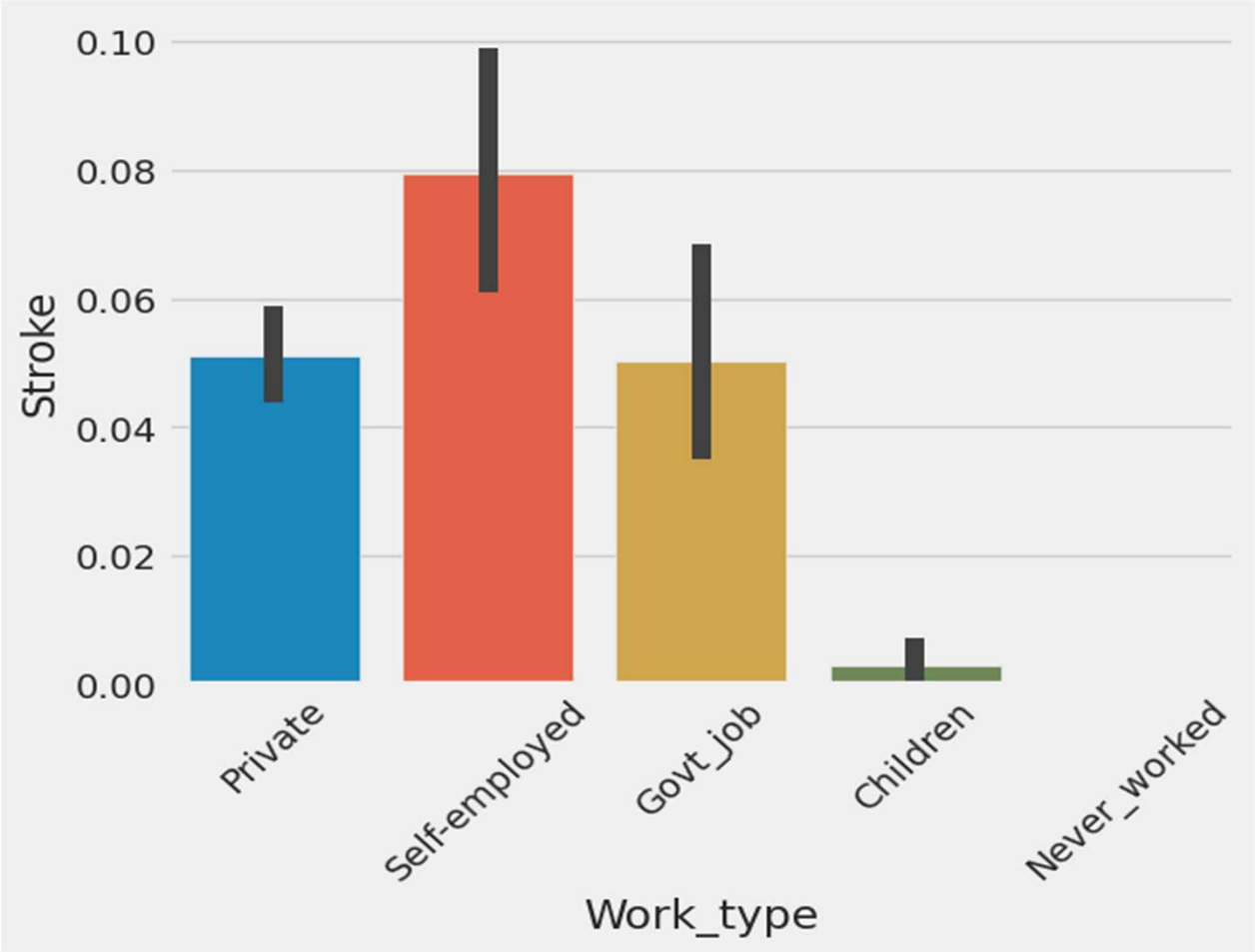
HEART DISEASE FEATURE



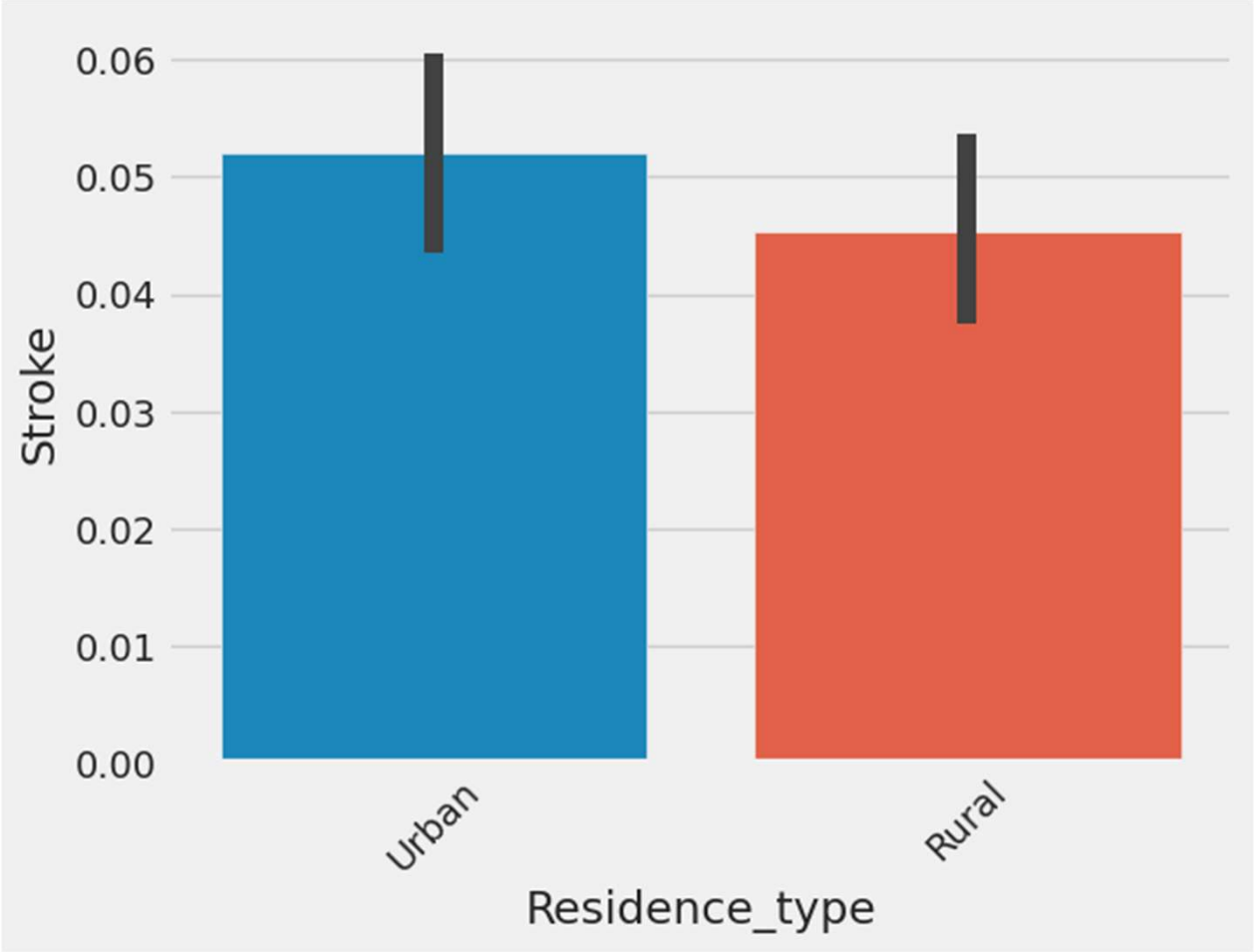
EVER MARRIED FEATURE



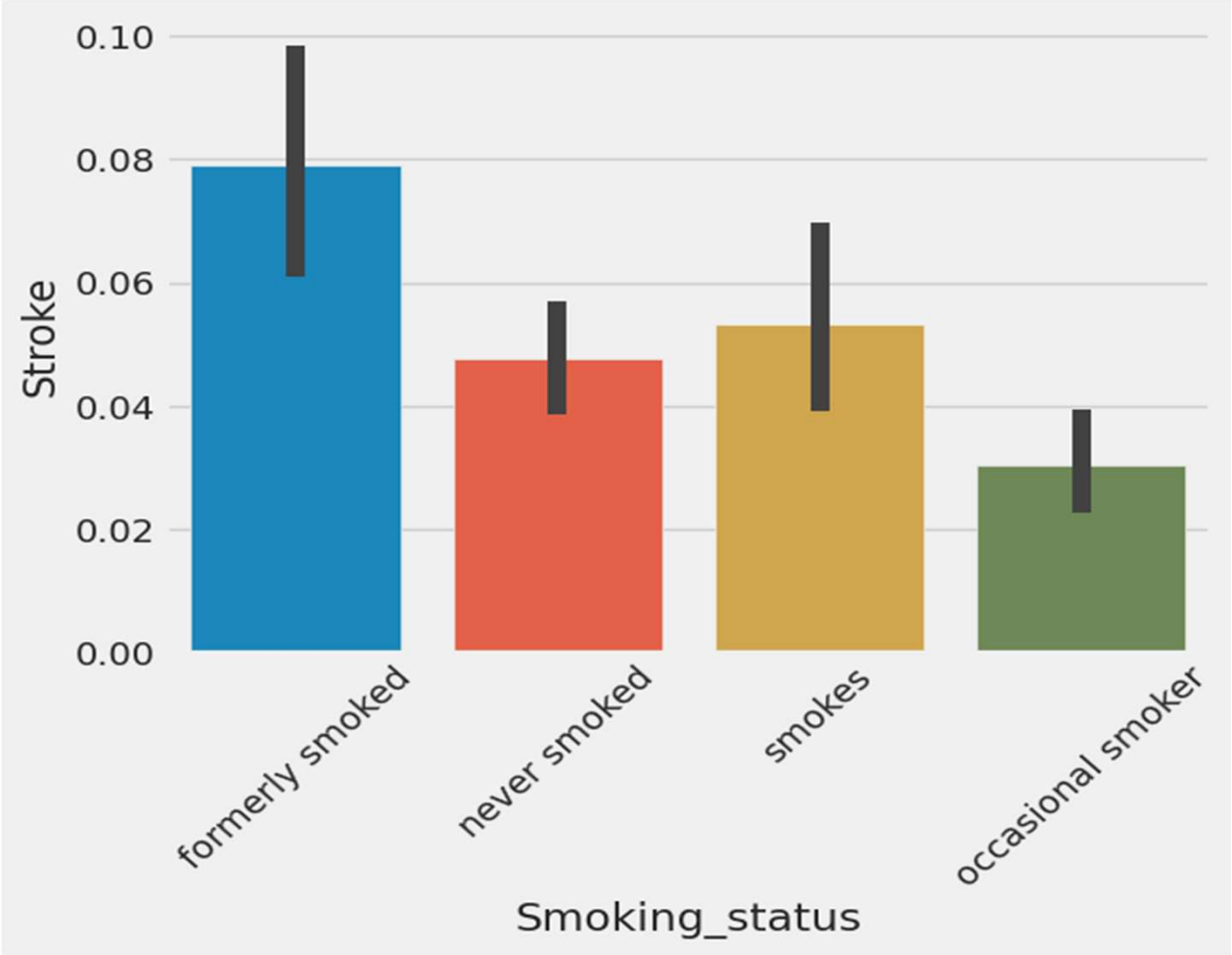
WORK TYPE FEATURE



RESIDENCE TYPE FEATURE



SMOKING STATUS FEATURE



MACHINE LEARNING MODELS

- This is a binary classification problem, there are 2 possible classes :
  - predict stroke (target): 1 if the patient had a stroke or 0 if not.
- Multiple Models Classifiers :

LGBM, XGB, AdaBoost, GradientBoosting, LogisticRegression, SGDC, Bagging, RandomForest.

  - LGBMClassifier performed the best : the overall accuracy is about 95%, False Negative (the most problematic) also known as type 2 errors are highly detected.

Light Gradient Boost Model :

- A fast higher performance model that increases efficiency of models and reduces memory usage.
- Prone to overfitting but can be regularized with hyperparameters.
- We must focus more on regularization to tune their hyperparameters and finding a good balance to reduce both types of errors I & II.

RECOMMENDATIONS

- Recommendations are more oriented towards men because stroke targets male patients more than females.
  - They must check frequently their high blood pressure.
  - Get treated early if they have cardiovascular disease or get tested regularly.
  - Stop smoking.
  - Avoid conflicts between spouses.
  - Avoid stressful jobs.
  - Explore rural life more often and have healthy life quality.
- Future directions may be :
  - Tune and regularize the model in order to improve the model classification performance.
  - Change model evaluation to have a better metric results.