Title of the Project: Stroke Prediction Using Machine Learning

Domain: Medical

Dataset Link: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Dataset Statistics: In the dataset total 5110 rows and 12 columns. Here each row corresponds to a unique patient.

For attacking stroke there are no big difference between males and females. The chances of having a stroke significantly increase with age. The patients with hypertension are more likely to experience a stroke. We infer that heart disease increases the chance of experiencing a stroke. Also, marriage significantly increases the chance of having a stroke. Private and Govt. job have a similar impact on strokes. Residence type does seem to affect strokes much. Also, we infer that there is a (positive) relation between strokes and being (pre)diabetic. Although smoking does not seem beneficial for your health, the relationship with strokes is not so clear from this plot due to the large confidence intervals. The distributions for stroke and no stroke are fairly similar. However, strokes seem to be relatively common in the BMI range of 26-32 kg/m^2. This dataset is extremely imbalanced. In fact, if our model would always predict 0, it would be correct 95% of the time.

Abstract : I have considered various machine learning algorithms to predict strokes in patients. Before modelling, I performed some exploratory data analyses and converted the categorical data into numerical data using either binary encoding or one-hot encoding. The missing BMI values were imputed from a linear regression model that was trained on the remaining data. To avoid any scale issues, I also standardized all features.

After having fully pre-processed the data, I applied cross validation to two different machine learning algorithms using the Scikit-learn default settings. We trained Support Vector Machine and Logistic Regression. To tackle the issue of significantly imbalanced class labels, I appealed to the SMOTE algorithm to oversample the stroke labels, creating a more balanced training set. After that I applied machine learning algorithm to get some inference from the data.