Towards Stroke Prediction

Codes can be found at https://github.com/kehindeowoeye/Predicting_stroke

1 Introduction (Option 1)

Early forecasting of stroke has important application for example in providing adequate care and intervention for patients with the risk of developing this condition. With more than a hundred thousand cases recorded in the United Kingdom each year, this is a medical condition that demands serious attention. In this work, I compare several models for predicting stroke at five levels of granularity and conclude by highlighting the set of features that are most influential with regards to this task using mutual information.

2 Data & Preprocessing

Dataset consists of 14 features including the attribute to be predicted. All missing values were replaced by the values corresponding to the immediate ID with respect to the feature of interest. I split the dataset into training: 90% and testing:10%.

Data	Class 0	Class 1	Class 2	Class 3	Class 4
Training	0.52767528	0.18081181	0.12546125	0.11808118	0.04797048
Test	0.64516129	0.19354839	0.06451613	0.09677419	0

Table 1: Distribution of the five classes in the target labels across training and test data.

3 Methods & Results

Models: Classical machine learning models have been used prdominantly as they can cope with little data and have also been compared to modern ones such as a simple two layer feed forward deep learning model. Crucially, i used Adaboost (100 estimators and no regularisation as it is robust to overfitting by combining weak learners) and compared this with logistic regression (LR), Support Vector Machine (SVM)and a two layer feed forward deep learning model with a dropout of 0.2 to avoid overfitting and a softmax at the output (DNN). All models were evaluated using accuracy and F1 scores as accuracy is not enough to reveal the true performance of a model.

Data	Adaboost	LR	SVM	DNN
accuracy	0.71	0.65	0.65	0.19
F1 scores	0.95 0.4 0 0 0	0.83 0.4 0 0 0	0.78 0 0 0 0	0 0.32 0 0 0

Table 2: Results show Adaboost performed best in terms of accuracy and the F1 scores. The deep learning model underperformed du to reasons bordering on the size of training data.

4 Influential features

To determine the most useful features relevant to the stroke prediction task, we compute the pairwise mutual information between each feature and the predicted attribute. To achieve this, we use the mutual information for discrete and continuous data proposed recently [**?**].The plot for this experiment can be seen in fig. The results suggest factors such as Age, resting blood pressure, serum cholesterol, maximum heart rate and Oldpeak (depression induced by exercise relative to rest) are the leading influential features relevant to the prediction task.

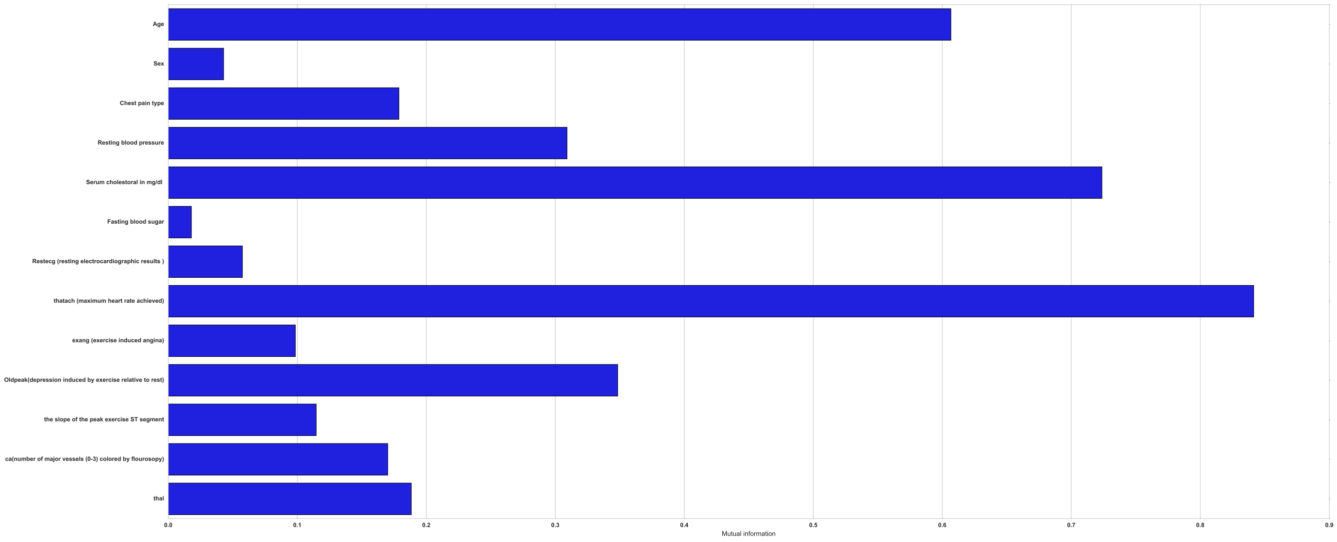

Figure 1: Mutual information between each of the 13 input features and the predicted attribute. See attached figure if labels are not clear here.