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 : We have considered various machine learning algorithms to predict strokes in patients. Before modelling, we 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, we also standardized all features, i.e., we enforced a zero mean and unit variance.

After having fully pre-processed the data, we applied cross validation to three different machine learning algorithms using the Scikit-learn default settings. We trained K-nearest neighbours, multilayer perceptron, and random forest classifiers. To tackle the issue of significantly imbalanced class labels, we appealed to the SMOTE algorithm to oversample the stroke labels, creating a more balanced training set.Subsequently, we performed grid searches to find

more optimal sets of hyperparameters. This time, instead of SMOTE, we tried out imbalanced-learns BalancedRandomForestClassifier to build a (balanced) random forest. Rather than oversampling the minority class, it undersamples the majority class. For the other two algorithms, we ran grid searches both with and without SMOTE. The (balanced) random forest model came out best from cross validation based on the F1-score. In general, all models that were trained with under- or oversampling performed significantly better than the ones that were trained without.

Finally, we investigated the obtained random forest model in more detail. In particular, we looked at the relative importance of the different features and visualized a random decision tree. As it turned out, age is by far the most important feature in predicting strokes. To assess the performance of this model in a more realistic setting, we had it predict on a hold-out test set that reflected the true class proportions. Unfortunately, the precision-recall curve does not show potential for improvement upon shifting the classification threshold. The ROC AUC score is pretty good, which means that the binary classes are sufficiently separated. To obtain a better performing model, more research is needed. Perhaps feature selection could benefit the model's performance. To arrive at a 'stronger' feature set, one could for example apply recursive feature elimination. Furthermore, one could try other machine learning algorithms. A promising choice would be XGBoost; over the last years, this algorithm has shown great performance in a wide variety of tasks.

Research/Implementation: We have optimized our model by performing grid searches and cross validation. Now that we have selected our final model based on the F1-score, we need to test its performance on the test set that we defined earlier. Note that the test set reflects the true, imbalanced situation where strokes are rare. From left to right, the first row of the confusion matrix contains the True Negatives (TNs) and the False Positives (FPs). The second row, again from left to right, contains the False Negatives (FNs) and the True Positives (TPs). As the diagonal contains the correct predictions, our goal has been to make this matrix as diagonal as possible. A high precision means that the FPs are suppressed, while a high recall corresponds to a low number of FNs. Ideally, both of these metrics are large. Both quantities are combined in the F1-score, which is the harmonic mean of precision and recall. For this reason, the F1-score is probably the most suitable metric to assess our model performance.