What was unique about the data?

The features was multi variant and it had many features for which we applied label encoding. The data was biased because it tried to predict people did not get stroke

Did you have to deal with imbalance? Yes, We did under sampling but model was over fitting so we shifted to over sampling and weight balancing

What data cleaning did you do? We handled missing and NaN values and unknown values.

Outlier treatment?

We did outlier treatment for BMI because there were few values above 60 and the count was insignificant.

We also removed one data entry for gender where it said Other.

Imputation?

We did that for BMI Values. For BMI we took mean for male and female seperately and replaced that.

Did you create any new additional features / variables?

We did not add any additional features but we did oversampling to make the model biased.

What was the process you used for evaluation? What was the best result?

We are computing accurace, precision and recall and we will display the confusion matrix.

What were the problems you faced? How did you solve them?

Initially the data was imbalanced so we fixed that with data undersampling and then the data was very less and model was overfitting so we did oversampling to fix that. We also tried various methods for supervision models

What future work would you like to do?

The accuracy for train for 95% and test was 90%. We would like to improve the accuracy using neural network. Currently, the data set is limited so we did not include neural network, but out future goal is to improve the data and add neural network.

Instructions for individuals that may want to use your work

Go with readme.