Project Report Part1

Zihan Wang 2022312876

1. Data preprocessing

The original data set has 299 features, however, not every feature is medically relevant to stroke onset, and some features are overlapped in terms of their meanings. Therefore, we first selected 131 meaningful features by carefully reading the feature description file and evaluating them by common sense. The selected features are listed in <Feature\_selected.csv>. Note that the feature “*DISPCODE*” describe whether the interview is completed. For those incomplete interview data, we delete them to prevent unexpected bias.

Since the stroke onset case only takes 3% of the total sample, to make the data set balance, we randomly selected samples without stroke with the same amount of stroke onset samples to build the dataset.

Before further feature selection, we removed features with more than 80% NaN values. For the rest features with less than 80% NaN value, we substitute the NaN values with a mean value of that feature. Then, we use the decision tree to further analyze the importance of different features to the stroke onset. It turns out, only a small number of features are significantly relevant to the stroke onset. So, we decide to use the top five largest features (Fig. 1) in the next stage of analysis.

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

Figure The feature analysis results

1. Train a model

We use logistic regression model to predict the stroke onset. Logistic regression is is a statistical model that models the probability of an event taking place, which have been used in clinical research.

The feature we used to train the Logistic regression model are:

X\_RFHYPE6 High Blood Pressure Calculated Variable

HeartDisease Ever had heart disease

X\_AGE80 Imputed Age value collapsed above 80

GENHLTH General Health

DROCDY3\_ Computed drink-occasions-per-day

The training curve of the model is plotted in Fig. 2. F1 score, sensitivity, precision, and AUC of stroke prediction are:

F1\_score 0.7290780141843971

sensitivity 0.7269220879403445

precision 0.731246766683911

AUC: 0.8

图表, 折线图

描述已自动生成

Figure Training Curve

图表, 折线图, 散点图

描述已自动生成

Figure The ROC plot