Proposal abstract:

According to the World Health Organization, stroke is the second leading cause of death globally. Thus, our group thinks that predicting the probability of getting a stroke in advance and interfering in time is crucial. In this project, we will train models with the dataset Stroke Prediction Dataset (https://www.kaggle.com/fedesoriano/stroke-prediction-dataset) using logistic regression, LDA, KNN and other possible models to predict whether a patient would get stroke.

Feedback:

Nice idea. One thing I'd suggest is being more explicit about which features you plan to use to predict p(stoke). Is your goal to make theoretical inferences or just maximize prediction? And how will you evaluate model performance in the latter, i.e., will you use cross-validation? Instead of comparing different classifiers, you could also compare different models, i.e., different feature-sets, which might be more informative.

Q1 Dataset

1 Point

In a short paragraph, describe the dataset you will use. You may provide a URL or reference for the dataset, but you should also describe in your own words the key characteristics of the dataset: What are the features/predictors? How many observations are there? Are these categorical (discrete) or continuous/quantitative variables?

Url: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset

The data contains 5110 observations, 12 columns with 10 predictors:

1. Gender: Binary categorical feature that indicates the gender of the patient.
2. Age: Ages of the patient.
3. Hypertension: Binary categorical feature that indicates whether the patient has hypertension.
4. Heart_disease: Binary categorical feature that indicates whether the patient has heart disease.
5. Ever_married: Binary categorical feature that indicates whether the patient has married.

6. Work_type: categorical variable indicating the patient's work status, such as "never employed", "self-employed", etc..
7. Residence_type: binary categorical variable indicating the patient's type of residence: urban or rural.
8. Average_clucose_level: Ratio variable that indicates the average glucose level in blood of the patient
9. BMI: continuous, quantitative variable. Body mass index.
10. Smoking_status: binary categorical variable: Categorical variable that indicates whether the patient smokes or not.

Q2 Modeling question

1 Point

Explain ONE key question, goal or hypothesis you hope to address with your project.

We would like to build a classifier that can predict the onset of stroke with high accuracy (hopefully near 90%). The takeaway of our project is that we hope the features which constitute our most accurate classifier can serve as precautionary factors in preventing stroke with some grain of truth.

Q3 Modeling method

1 Point

Explain ONE data analysis method that you will use to address your question/hypothesis. (For example: Multiple linear regression)

Since our data has clear labels (get stroke/ does not get stroke), we first narrow down the method to supervised methods. In addition, as the prediction of our prediction is categorical, we decided to choose between logistic regression and LDA.

We would like to build a classifier using logistic regression since our outcome consists of exactly two categories.

Q4 Model selection plan

1 Point

Explain how you will use model selection to compare at least TWO different models with different levels of complexity/flexibility. Note that the two models should not be completely different methods; they should both be instances of the same modeling method, but with different levels of complexity. For example, you may choose to use regularization with different values of lambda, or you may choose to use subset selection to select an appropriate set of features.

In order to evaluate the model performance and choose the best model, we would perform a cross-validation and split our data into training and testing sets at roughly 80 to 20 ratio.
For logistic regression, the complexity of the model is related to the number of independent variables used to fit the model. We plan on selecting the most accurate model through forward stepwise model selection.
Using the training data, we would first train our logistic model using a single predictor variable, training 10 models in total. After fitting the 10 models using training data, we would predict the result on the test data. Using the confusion matrix, we would decide the model with the best performance being the one with the highest prediction accuracy.
We then train a 2-predictor model on top of the model with the single best predictor, repeat the process of validation using the testing dataset like we did for 1-predictor model, then choose the best 2-predictor model.
We would iterate this procedure until eventually arriving at a model with 10 predictors. Ideally, the best performing 10-predictor model should be the best model. But if that's not the case, such as in the event of overfitting, we would ultimately decide on a single best model by choosing from the 10 models with the best testing accuracy at each stage (1-predictor, 2-predictor, etc.).