Firstly, I have to apologize that this is an individual project. The reason why I haven’t done it in a group is that: I’m a SPOC student, I haven't organized my time to finish all the courses gradually. So it remained 3 or 4 months for me to finish 8 courses, of which ML is one. As I didn’t have enough time to ask others and discuss with others, I did this project by myself. Thank you for your comprehension.

Then let’s come to the project topic.

## Data analysis and feature selection

After importing the data, It shows up the first 10 lines to take a glance of every column.

### NaN data

Using describe() and isna().sum(), I can see, fortunately, there’s only 201 cells missing for bmi feature.

How I decide whether to keep or drop the 201 lines is simple. I select all the lines where bmi is NaN, and describe it. I see that the mean of stroke feature is significantly higher than it is in the total dataset, as well as other features. But I finally decided to drop these 201 lines because I think there are many reasons which can affect the bmi level, I don’t want false information to pollute the data.

By using dtypes, I take a glance at the data types of each column. Then I analyze the relation between each feature with the target “stroke” mainly by using groupby(‘<feature>’)['stroke'].mean() and some plots.

| **Feature** | **Analysis** | **Action** |
| --- | --- | --- |
| Gender | After dropping the only line where the gender is “other”, I find that the mean of stroke in the female and male group is close, then this feature is not really important for the stroke, but I just decided to keep it. : ) | Kept |
| Age | From the plot and the mean, I can see clearly that older people have a stronger possibility of having a stroke. | Kept |
| Hypertension | It’s highly correlated to the stroke rate. | Kept |
| Heart\_disease | It’s highly correlated to the stroke rate. | Kept |
| Ever\_married | Strangely, it’s correlated to the stroke rate. : o | Kept |
| Work\_type | The “self-employed” has the highest mean of having a stroke. I can imagine. But as the dataset has a big bias on the “private” group and a very very few “never\_worked” group. I don’t think that this is good information to keep. So I dropped this feature. | Dropped |
| Residence\_type | The mean of this feature between “Rural” and “Urban” is close, so I drop this feature. | Dropped |
| Avg\_glucose\_level | It’s relevant to stroke rate. | Kept |
| Bmi | It’s relevant to stroke rate. | Kept |
| Smoking\_status | I can see there’s not much bias on the data, and people who smoke seem to have a higher rate of stroke. | Kept |

### Data processing

I changed the type of “gender”, “ever\_married” and “smoking\_status” into int. The final data is stored in “d2”. Then the heatmap plot shows the correlation between each feature.

## Model training and model evaluation

By using train\_test\_split(), I split the dataset into two sets where the test size is 0.2 of the data set.

Then I put all the splitted datasets into 4 models:

| **Model** | **Result** | **F1-score Accuracy** |
| --- | --- | --- |
| Logistic regression |  | 0.75 |
| MLP |  | 0.94 |
| Decision tree |  | 0.91 |
| Random forest |  | 0.96 |

In my first predicted result of LogisticRegression(), the model can barely predict the stroke, because in the data, the stroke=0 has a big bias, so I added the parameter class\_weight="balanced" to the LogisticRegression() in order to cancel the bias.

So from the result of the 4 models, I can see that the random forest has the best result.