Machine Learning with Python Labs
ZHOU Zhijian DS S21 SPOC

Due to a problem that I couldn't solve, I was unable to use Jupyter Notebook, and instead, I completed this project using Google Colaboratory.

1. Data

After using DataFrame to read the dataset, I noticed that there are numerical attributes and categorical attributes.

1.1 The "Other" gender label

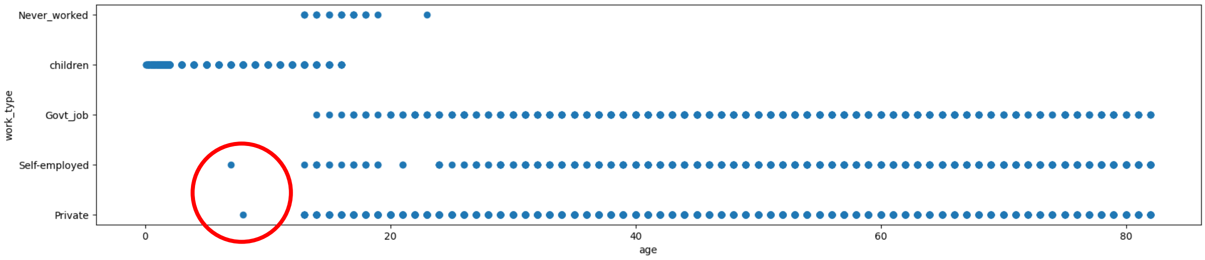
In this dataset, there are significantly more stroke labels than non-stroke labels, so the trained model inevitably has an advantage in predicting non-stroke cases. I have one non-stroke label with gender=Other", and I have decided to remove this row, even though removing data is not recommended. However, having one single row about "gender=Other" cannot contribute to train the model to predict having stroke or not for "gender=Other".

1.2 Missing values

In this dataset, there are 201 missing bmi data points. Considering the differences in body characteristics between males and females, all the missing bmi values are filled with the respective gender's average bmi.

1.3 Outliers

After plotting age against ever_married, work_type, and smoking_status separately, I noticed that among individuals under the age of 10, there were no records of marriage or clear indication of smoking status. However, there were two individuals under the age of 10 who were reported to be working. This seems suspicious, and I suspect that these two data points might be outliers. Leave them unprocessed for now.



1.4 Plotting and calculation

It appears that gender does not influence the occurrence of strokes, but stroke cases generally increase with age. The calculations indicate that individuals who have both hypertension and heart disease are more prone to stroke compared to those who only have hypertension or only have heart disease. The number of individuals who are already married in the dataset is twice as much as the number of unmarried individuals. Surprisingly, the proportion of married individuals who have stroke is significantly higher than that of unmarried individuals. The residence type of an individual does not appear to have any impact on stroke occurrence; therefore, I have decided to remove this column from the dataset. The average glucose levels are concentrated in two intervals: 50-120 and 180-220. The number of stroke cases is relatively higher in these two intervals. Similarly, the bmi is concentrated around 30, and this range also has a higher number of stroke cases. Stroke patients are evenly distributed across the four smoking status categories.

2. Features

2.1 Label encoder

To train the machine learning model, I replaced the different labels of categorical attributes with natural numbers.
gender: Female 0, Male 1
ever_married: No 0, Yes 1
work_type: children 0, Govt_job 1, Never_worked 2, Private 3, Self-employed 4
smoking_status: formerly smoked 0, never smoked 1, smokes 2, Unknown 3

2.2 Correlation between attributes

The top three correlated pairs are: age-ever_married, age-work_type, and age-smoking_status.

3. Model training

3.1 Train/test split

The dataset was split into training and testing sets using an 80%-20% ratio, and "stratify" was used to ensure that the proportion of stroke cases remains consistent between the training and testing sets. Throughout the model training phase, the random parameter was set to 42 for reproducibility.

Among the five models used, Logistic Regression, Decision Tree, AdaBoost, MLP, and Random Forest, overall, these models perform significantly better in predicting non-stroke cases than predicting stroke cases. Specifically, both Logistic Regression and AdaBoost failed to predict any stroke cases, regardless of whether the true condition was a stroke or not. What deserves more attention is the F1 score for predicting strokes. The Decision Tree and MLP models have the highest F1 scores for predicting strokes, with values of 0.15 and 0.11, respectively, which are currently the highest among the models tested.

4. Outliers suppression

As mentioned earlier, two data points were suspected to be outliers. However, after removing them from the dataset and re-splitting it into training and testing sets, the results show that it did not improve the performance of the five models.

5. Conclusion

This project simulated the process of using machine learning models to predict stroke, including data processing, feature engineering, model training, and comparison. Unfortunately, a high-performance model for predicting strokes was not obtained. This is likely due to the scarcity of stroke cases in the dataset, causing the trained models to perform more efficiently in predicting non-stroke cases. In the end, a requirements.txt file was generated, but since Google Colaboratory platform was used, it listed almost all the packages that were already installed on the platform.