Cardiovascular and Stroke Diseases Prediction and Relation

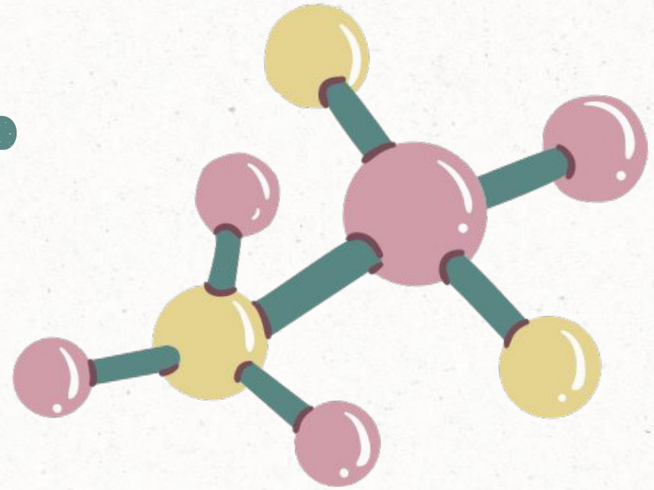
Zhao Hui, Pan Zhujing, Wei Yushan, Fu Siyao

Introduction

Machine learning has been used in medical fields to help medical practitioners identify diseases and make diagnosis more accurate and precise. This project focused on cardiovascular and stroke diseases. We used subjects' common characteristics that can be easily obtained to predict the probability of the above two diseases.

01

BASIC CONCEPTS



DEFINITIONS

Cardiovascular Disease

Cardiovascular disease (CVD) is a general term that describes a disease of the heart or blood vessels.
As for risk factors that may cause cardiovascular disease, most of them are listed below:

- High blood pressure (hypertension)
- Smoking (or other tobacco use)
- High blood cholesterol
- Diabetes
- Poor Diet
- Lack of exercise
- Being overweight
- Excessive alcohol consumption
- Stress

Stroke

A stroke is a serious medical condition that occurs when the blood supply to the brain is disturbed. It is a medical emergency that need to be treated as soon as possible. There are also kinds of risk factors that may lead to stroke:
The major risk factors for stroke include:

- High blood pressure
- Diabetes
- High LDL cholesterol levels
- Smoking
- Age
- Gender
- Anxiety, depression, and high stress levels
- Overweight and obesity or carrying extra weight around your waist and stomach.

INSIGHTS

From the list, we can see that there are many factors that may cause cardiovascular disease and stroke, while it is difficult to strictly follow the guidance on preventing the disease. Also, stroke is known as one type of cardiovascular disease, so the two diseases are strongly linked together.

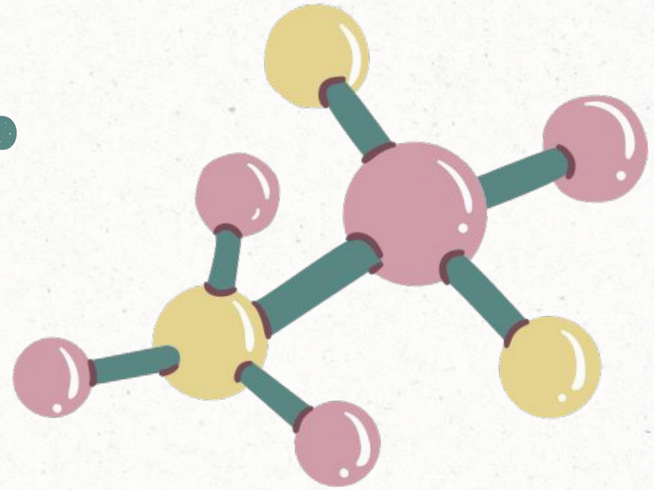
So, it is important to find out which are the dominant factors contributing to these diseases. Because it will enable better prediction of the probability of the diseases and mark at-risk groups to provide quantitative analysis and support for disease prevention.

MODELS

In the project, we used Linear Regression, Logistic Regression, Decision Tree, Svm, Cnn, Naive Bayes, Multi-Layer Perceptron, Ensemble Learning.

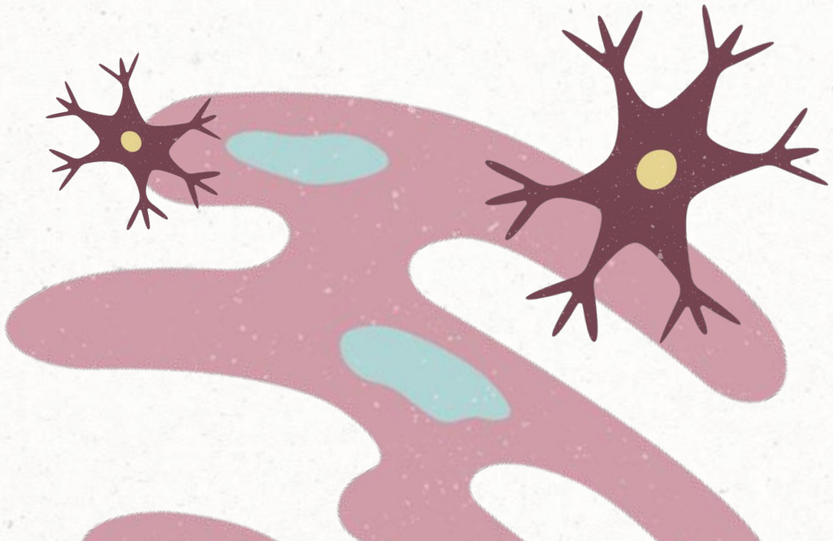
02

OBJECTIVES



01 Cardiovascular Disease Prediction

Our predictions are based on the “Cardiovascular Disease Dataset”. It includes 7000 cases of cardiovascular diseases and 13 attributes of subjects such as age, gender, cholesterol, smoke, etc.



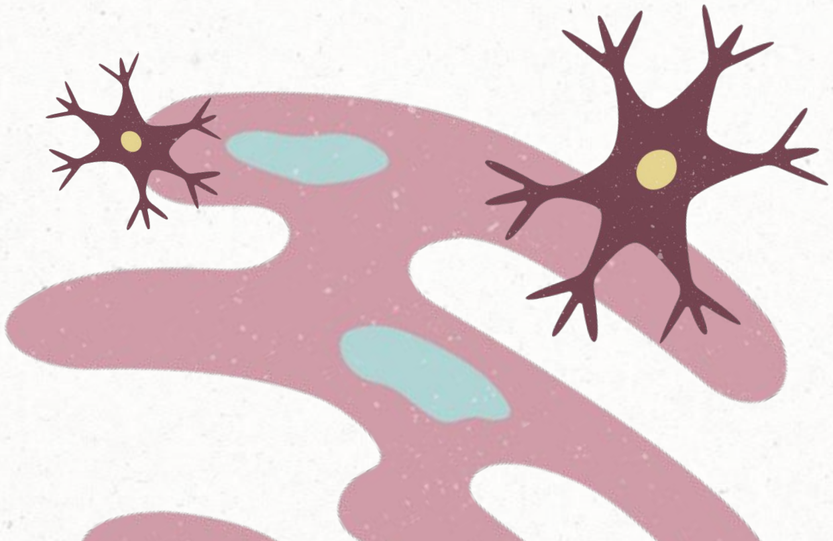
id	age	gender	height	weight	ap_hi	ap_lo	cholesterol	gluc	smoke	alco	active	cardio
0	18393	2	168	62.0	110	80	1	1	0	0	1	0
1	20228	1	156	85.0	140	90	3	1	0	0	1	1
2	18857	1	165	64.0	130	70	3	1	0	0	0	1
3	17623	2	169	82.0	150	100	1	1	0	0	1	1
4	17474	1	156	56.0	100	60	1	1	0	0	0	0
8	21914	1	151	67.0	120	80	2	2	0	0	0	0
9	22113	1	157	93.0	130	80	3	1	0	0	1	0
12	22584	2	178	95.0	130	90	3	3	0	0	1	1
13	17668	1	158	71.0	110	70	1	1	0	0	1	0
14	19834	1	164	68.0	110	60	1	1	0	0	0	0
15	22530	1	169	80.0	120	80	1	1	0	0	1	0
16	18815	2	173	60.0	120	80	1	1	0	0	1	0
18	14791	2	165	60.0	120	80	1	1	0	0	0	0
21	19809	1	158	78.0	110	70	1	1	0	0	1	0
23	14532	2	181	95.0	130	90	1	1	1	1	1	0
24	16782	2	172	112.0	120	80	1	1	0	0	0	1
25	21296	1	170	75.0	130	70	1	1	0	0	0	0

02 Stroke Disease Prediction

For the second dataset about stroke, it included 5110 cases with 12 attributes such as gender, age, hypertension, etc.

In the following steps, we will use the two datasets to do predictions.

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
13861	Female	52	1	0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1
68794	Female	79	0	0	Yes	Self-employed	Urban	228.7	26.6	never smoked	1



03

DATA PRE-PROCESSING

DATA CLEANING

Identify Duplicate Data

Replace Null

Drop Unnecessary Columns

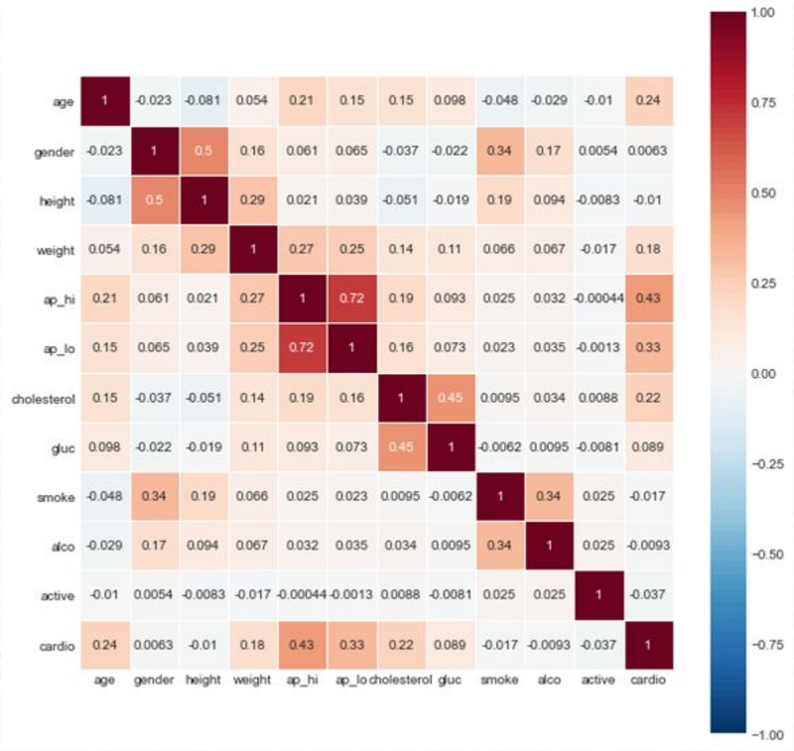
Unify Units (eg. days to years)

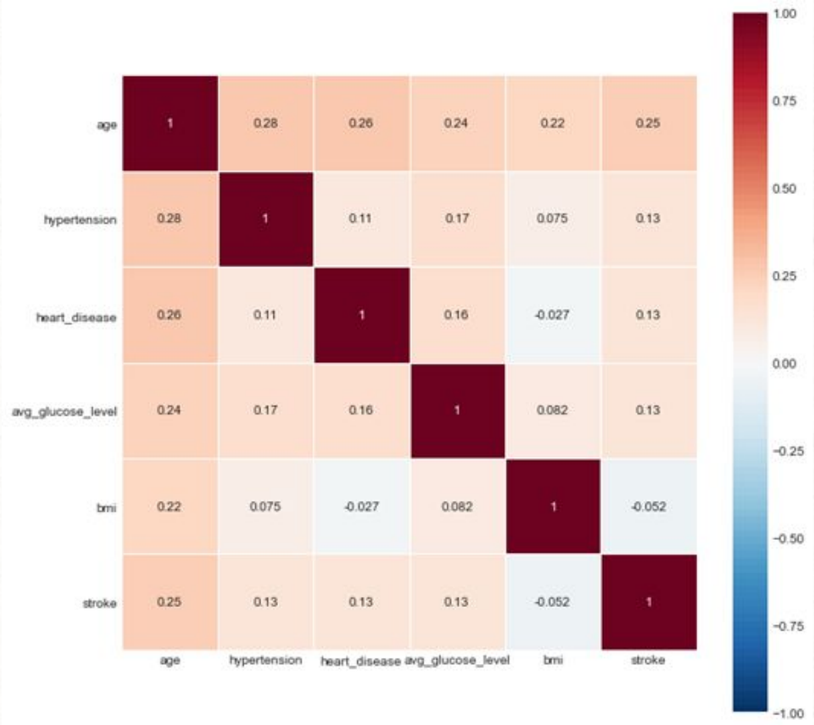
Text to Numbers

```
stroke['gender'] = stroke['gender'].replace(
    {
        'Male': 0,
        'Female': 1,
        'Other': 2
    }
)
```

FEATURE EXTRACTION

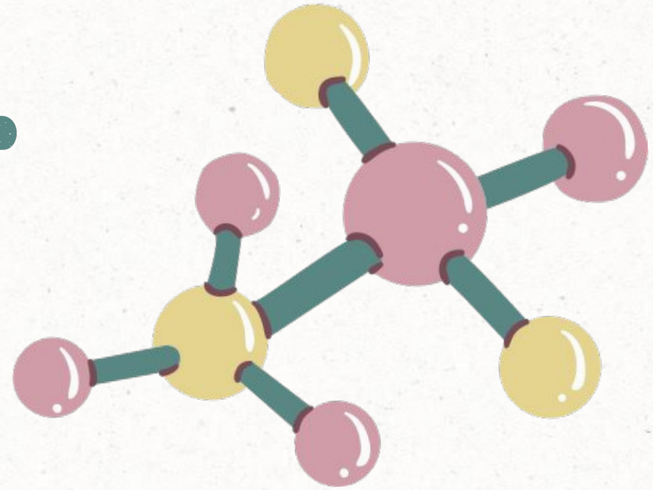
We plot the correlation between features in the cardio dataset to see which features need to be retained. Here is the correlation map for cardio.



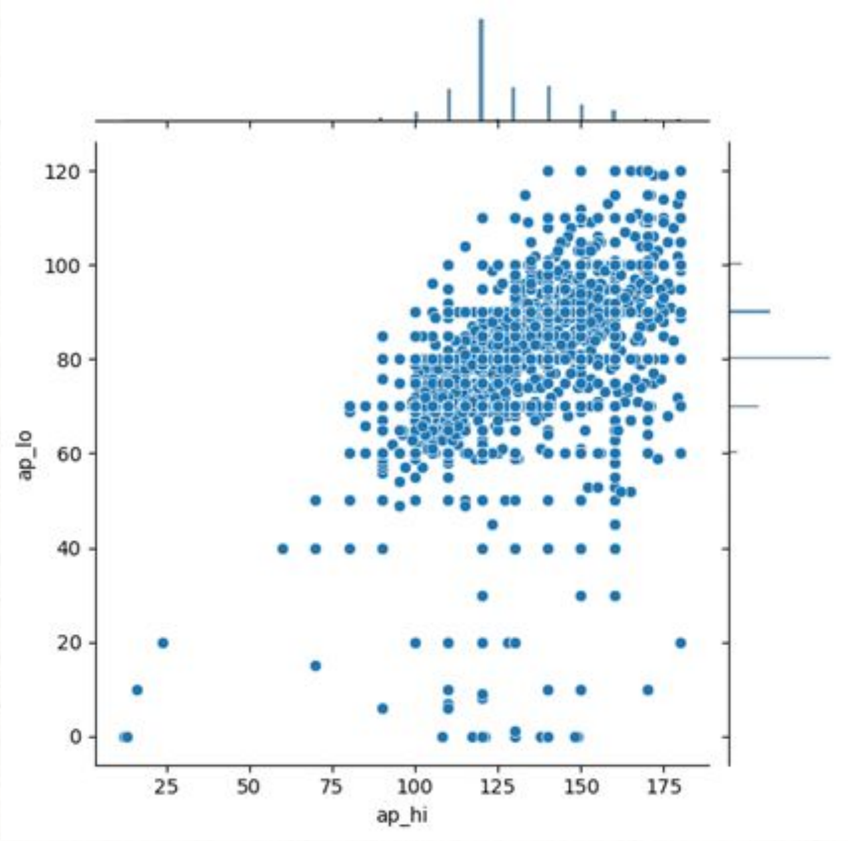


04

Exploratory Data Analysis

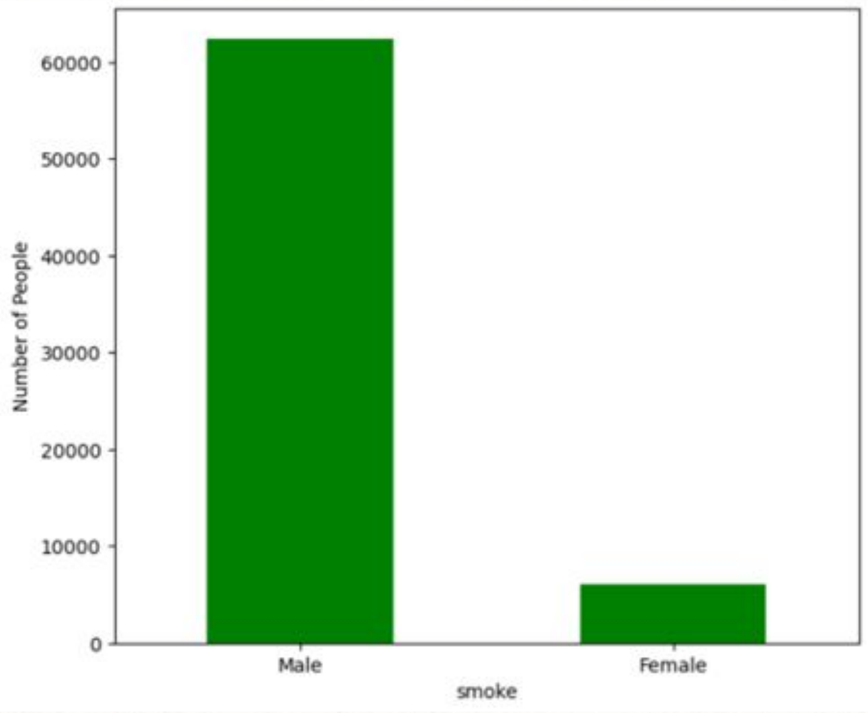


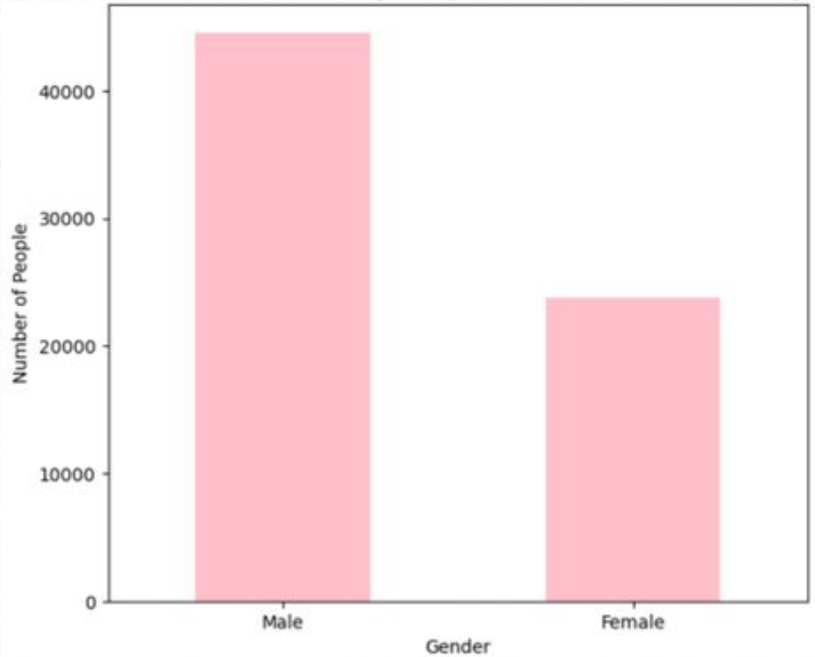
01 Visualization of Cardiovascular Dataset



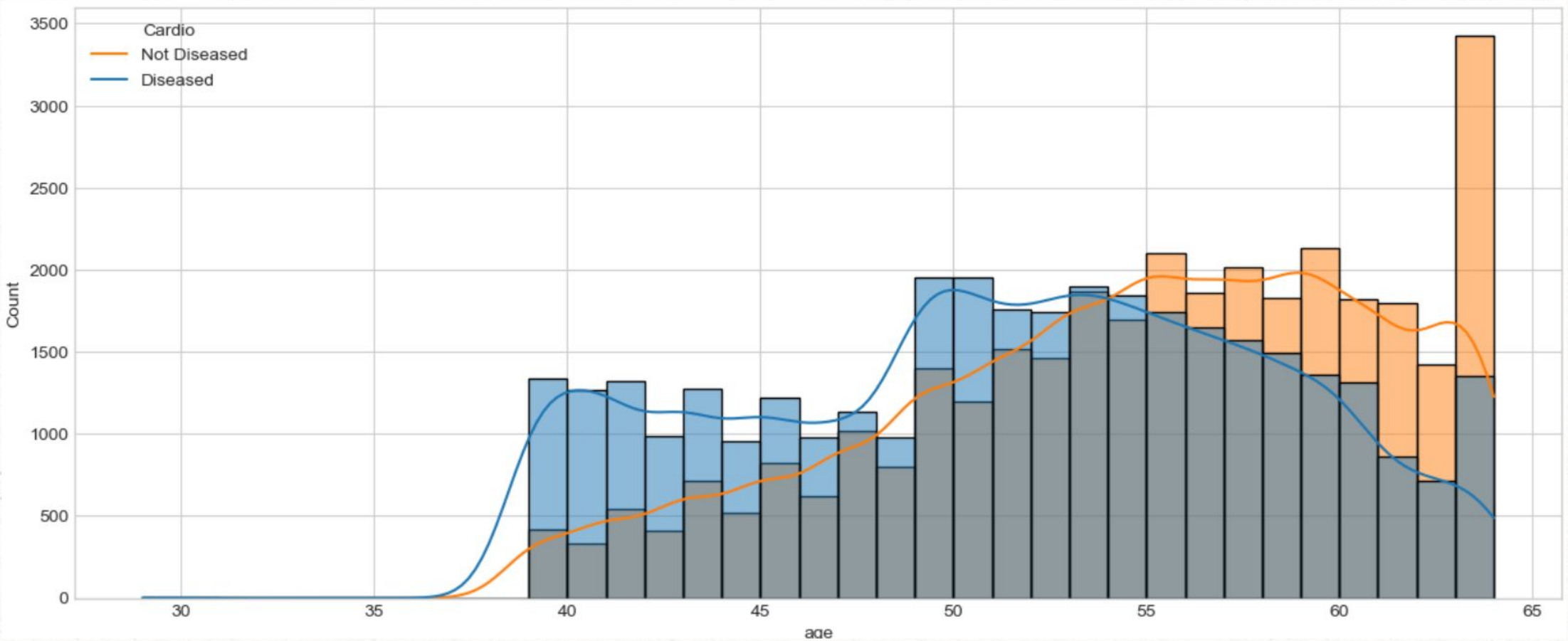
The left diagram is scatterplot of ap_hi and ap_lo which means systolic blood pressure and diastolic blood pressure distribution in the processed cardio dataset.

We can find that people with higher ap_hi values are not low in ap_lo. Although there are exceptions, they are in the minority.

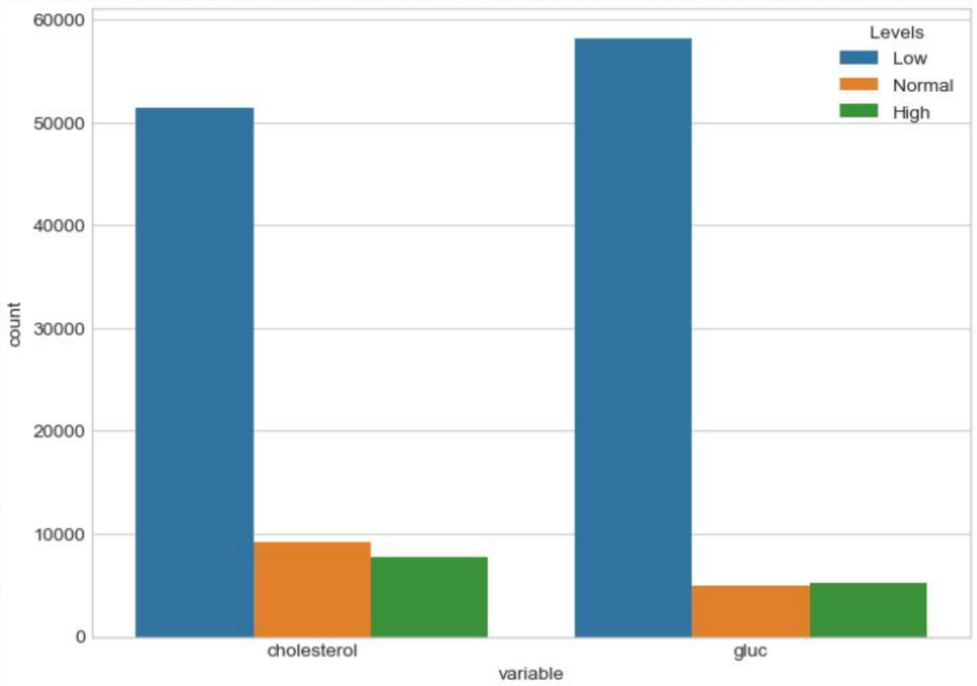


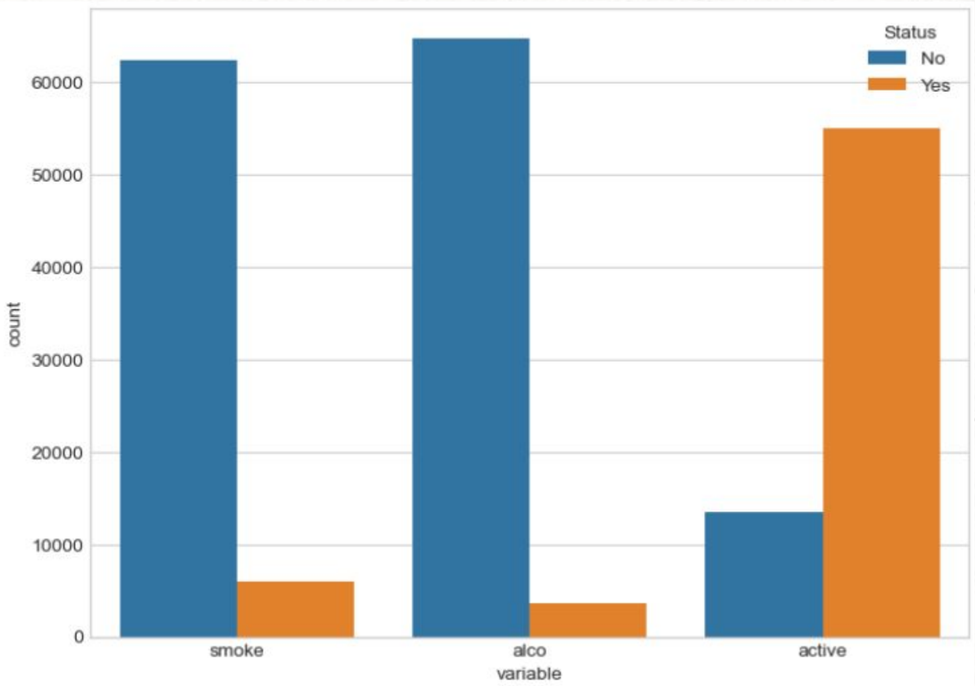


Through the figure above, we can find that men smoke far more than women and the number of male cardiovascular and cerebrovascular patients is also higher than that of women.

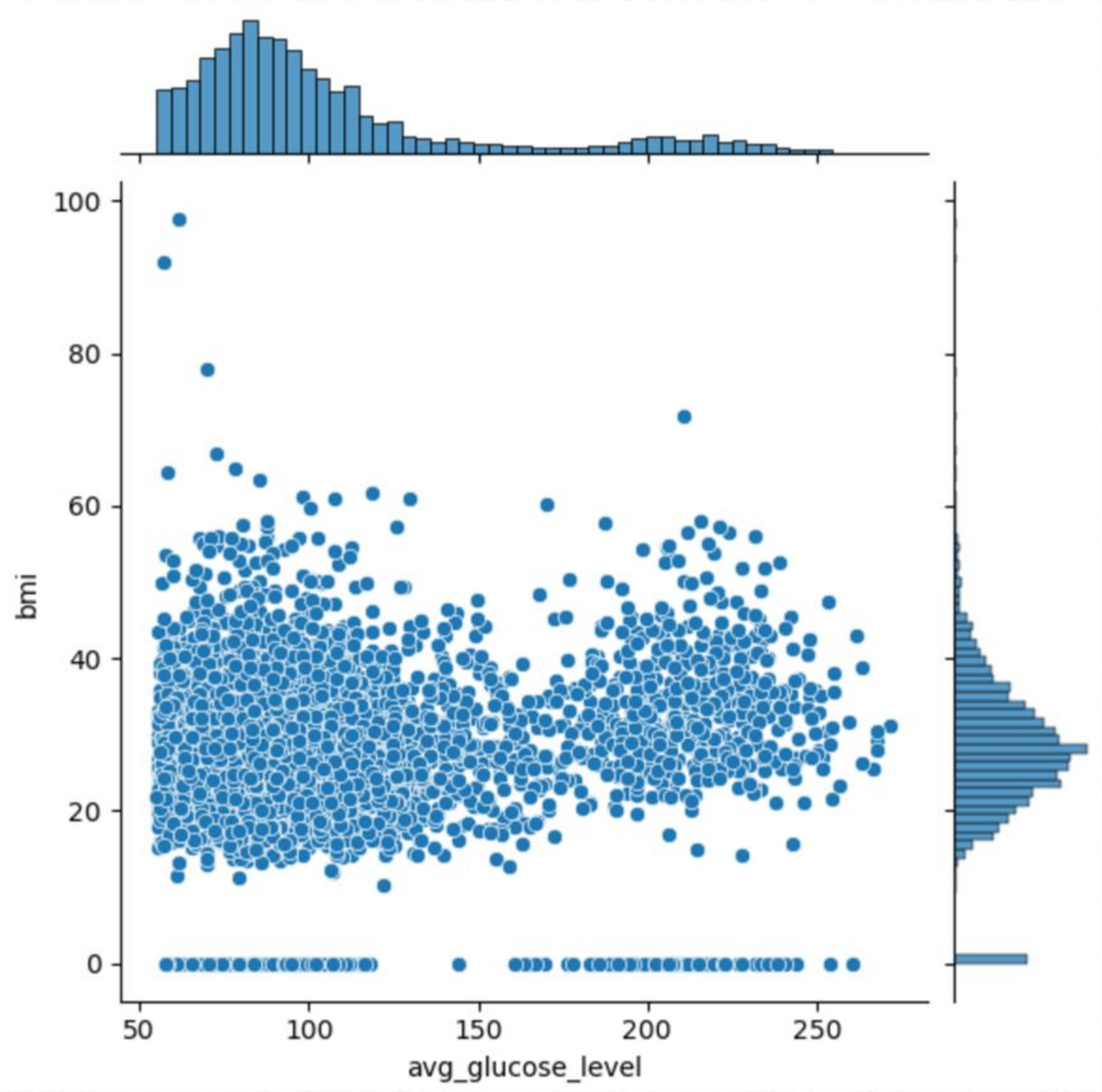


Through the survey of the age of people with and without cardiovascular disease, it seems that older people are more likely to be diagnosed with cardiovascular and cerebrovascular diseases.



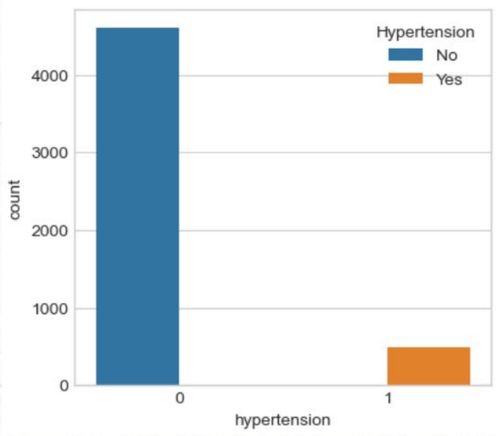


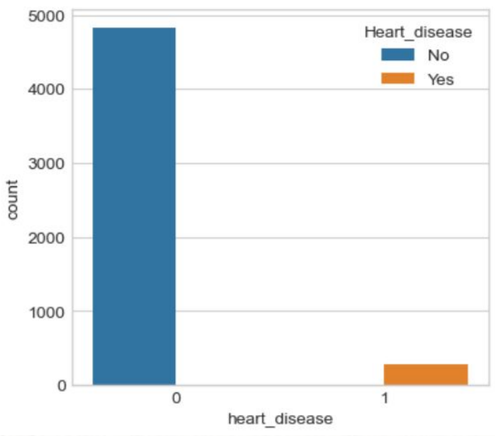
After counting the number of categories for all categorical variables, we can find that the number of different categories is very uneven.

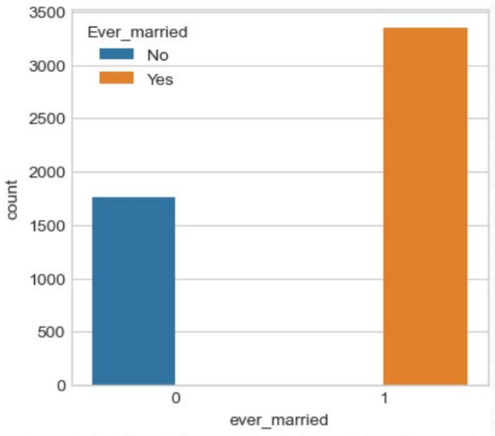


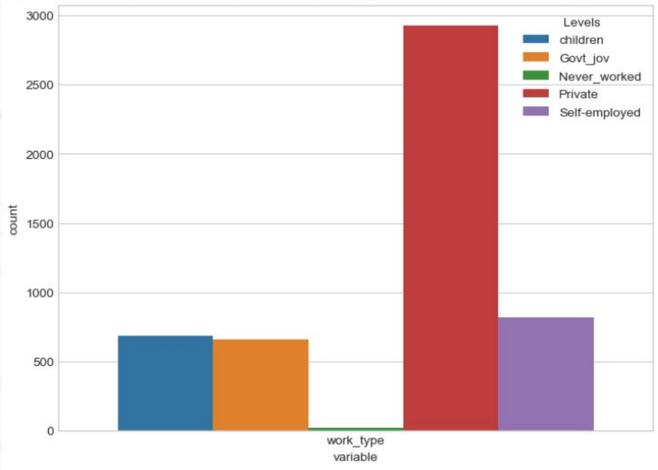
02 Visualization of Stroke Dataset

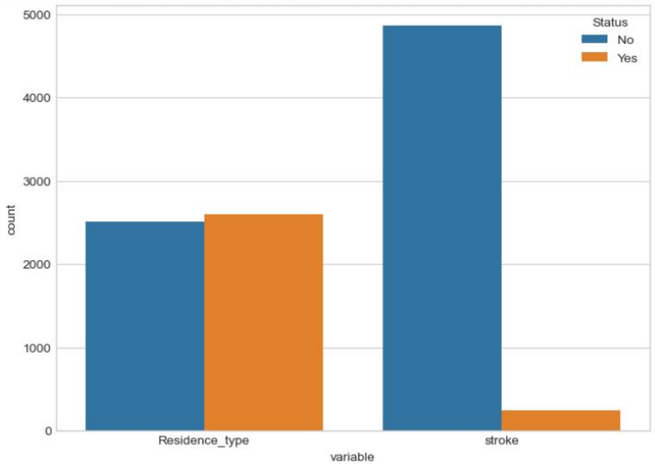
Here is scatterplots of avg_glucose_level and BMI in the processed stroke dataset. We can see that most of the avg_glucose_level is concentrated at 100 and most bmi is about 20 to 40.



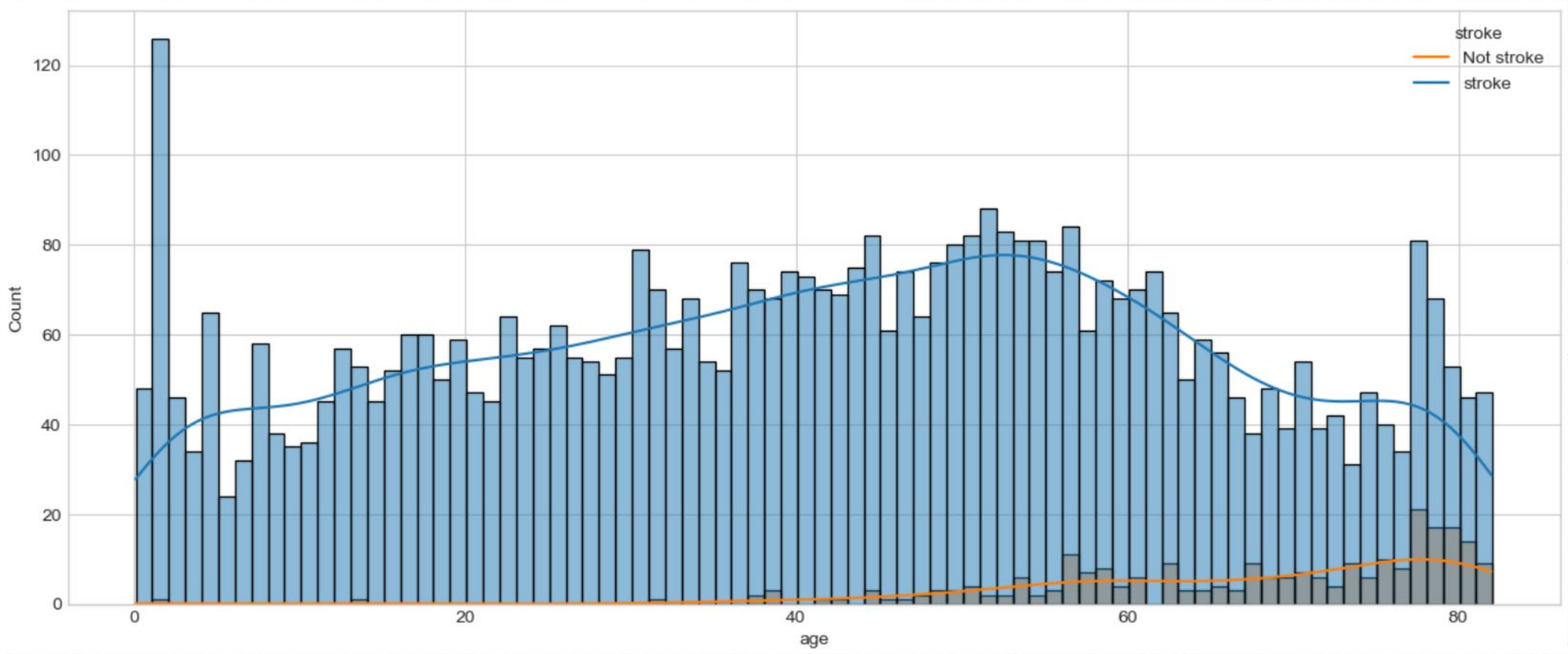




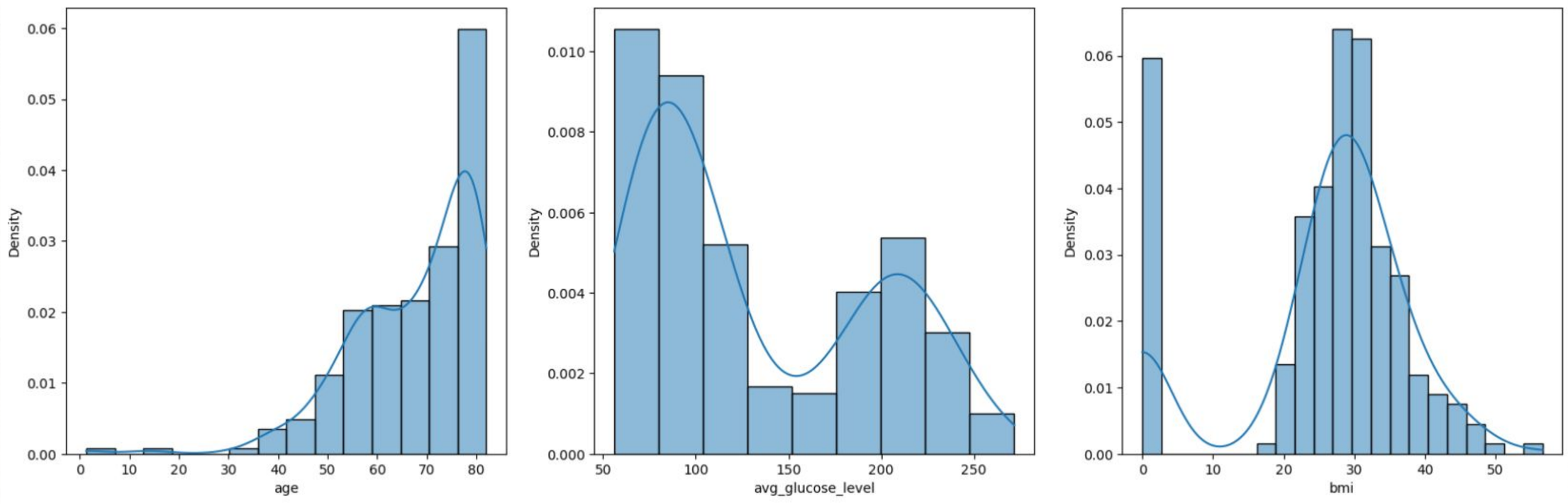




Similar to the cardiovascular and cerebrovascular datasets, the category distribution of the stroke dataset is also uneven.



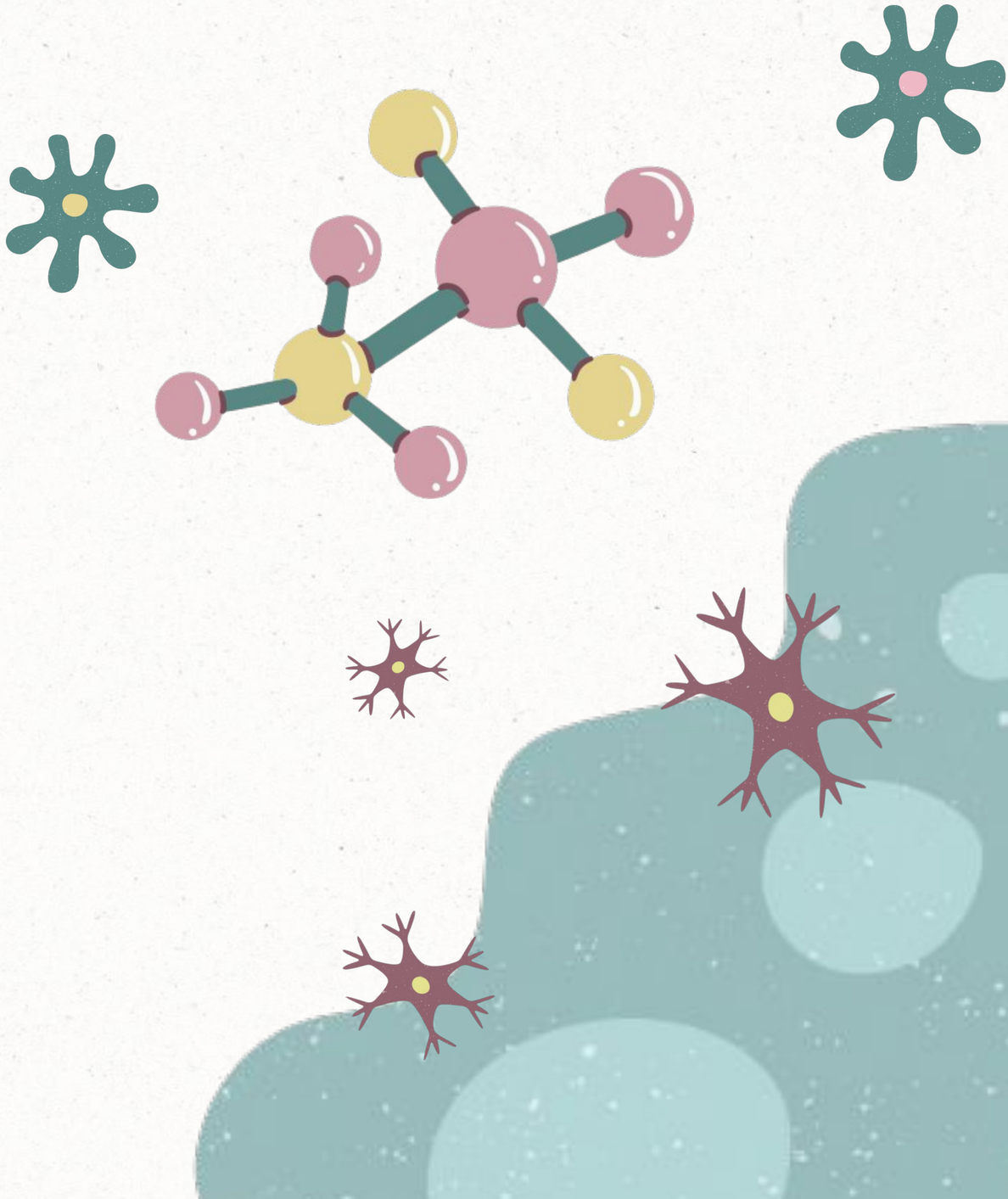
Slightly different from the cardio dataset, the proportion of strokes in different ages is much less than that of cardio. However, the probability of disease also increases positively with age.



By histogram statistics on the three columns of age, avg_glucose_level and BMI of the number of patients, we can find that the age of 80 years old is the largest, avg_glucose_level below 100 is the most. The number BMI is too low or around 30 are the largest.

05

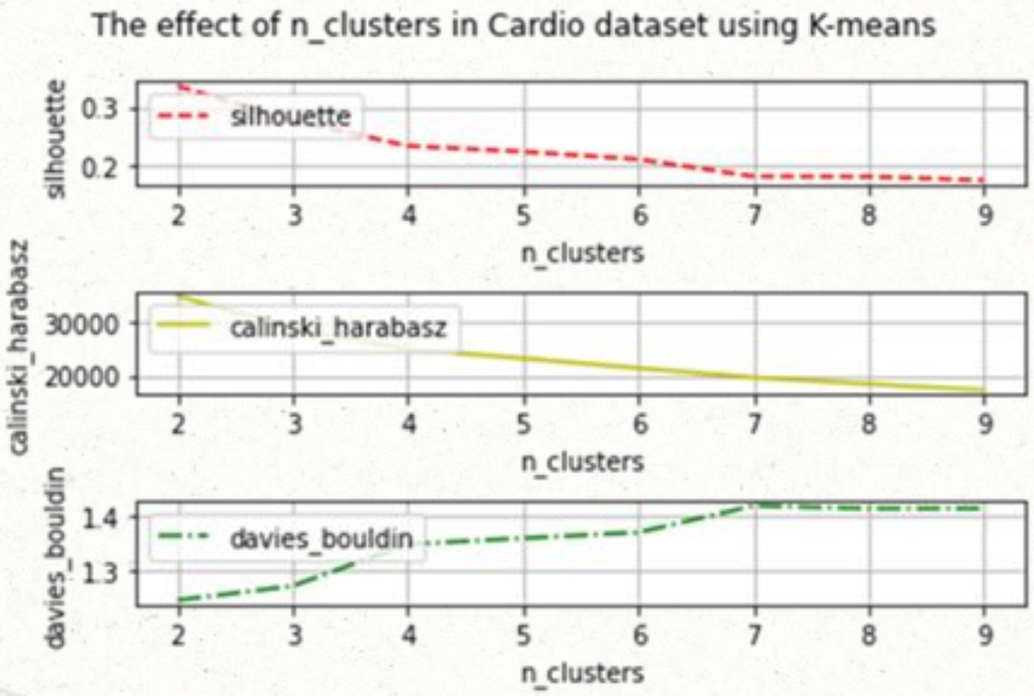
Clustering



METHODS

We used K-means++ and DBSCAN to do the clustering. And used five metrics to evaluate the results: silhouette coefficient, calinski-harabasz, davies-bouldin, adjusted Rand index and adjusted_mutual_info_score

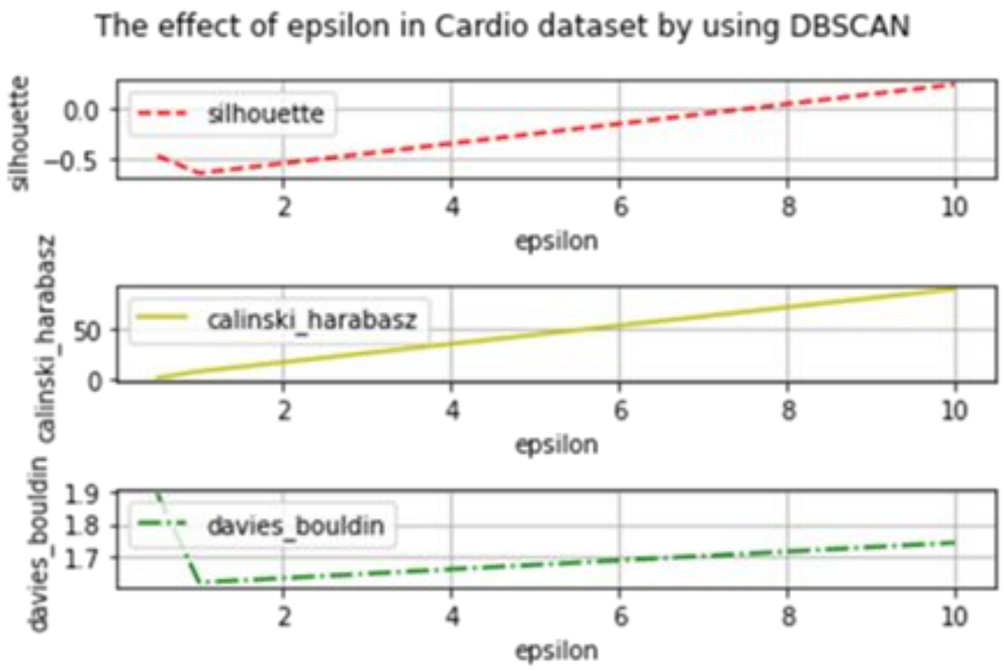
01 K-Means for Cardiovascular Dataset



These are the results of the change of the three indicators with n_clusters changes for K-means++.

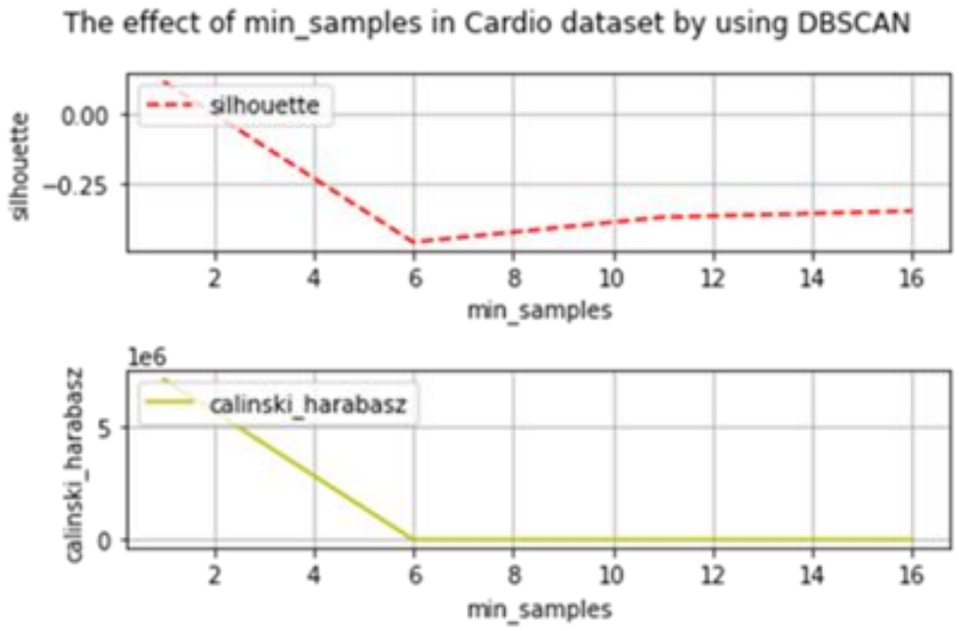
The larger the first two parameters, the better, and the smaller the last, the better.We can see that 2 is the best.

02 DBSCAN for Cardiovascular Dataset



For DBSCAN, we investigated the effect of epsilon and min_sampple on the results. The graph below shows the epsilon variation from 0.5 to 10.

02 DBSCAN for Cardiovascular Dataset



The graph below shows the change in min_samples from 1 to 20. We can see that 6 is a turning point, and before 6 the min_samples are as small as possible.

Cardiovascular Diseases

Any combination of the parameters of DBSCAN does not give as good results as k-means++. Therefore, by the above attempts, we chose k-means++ for the cardio set for aggregation.

Stroke Diseases

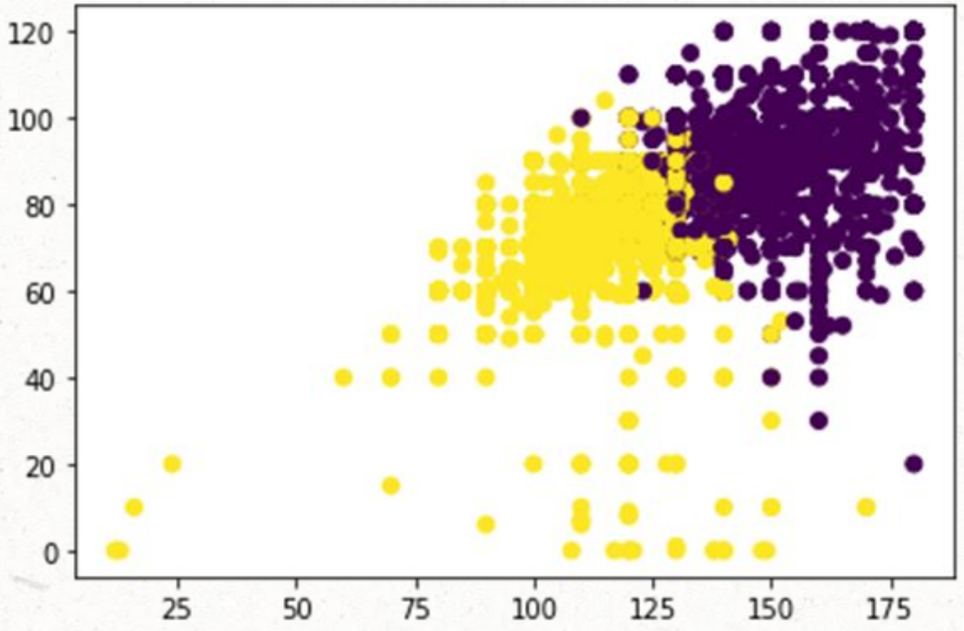
For the stroke dataset, we also tried different n_clusters k-means++ and same as cardio, 2 is the best. The following figure shows the results of the variation of its internal metrics with n_clusters.

Then we compare the clustering results of cardiovascular disease dataset and stroke dataset to analyze the similarities and differences between the two clusters.It can be found that the cluster center values of age for cardio and stroke are almost 55+.

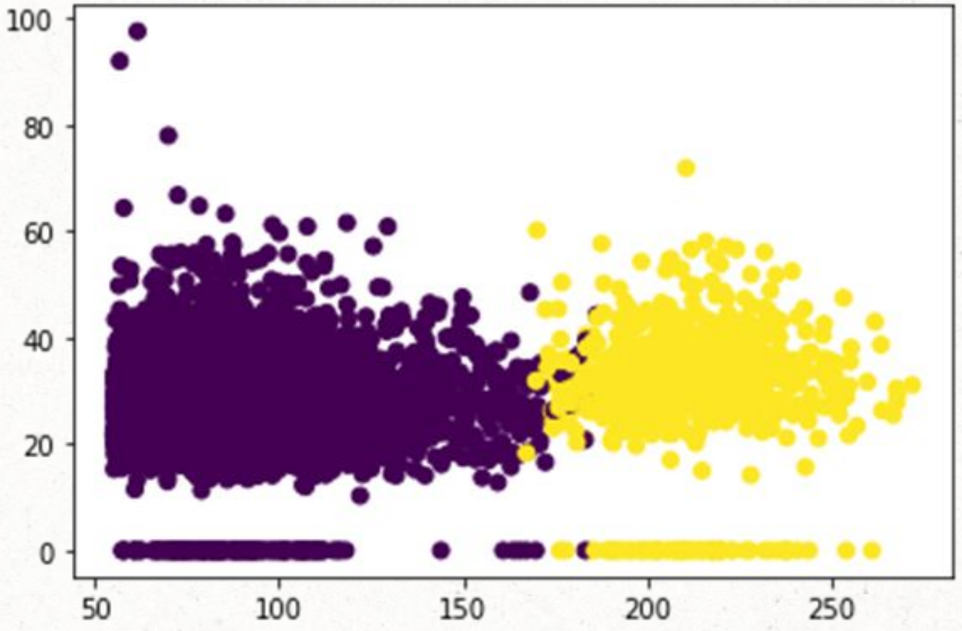
When we print out the clustering results of ap_hi and ap_lo as below, we can see that the areas of disease and not are obvious.

03 Results

When we print out the clustering results of ap_hi and ap_lo as below, we can see that the areas of disease and not are obvious.The same as the avg_glucose_level and BMI in stroke.



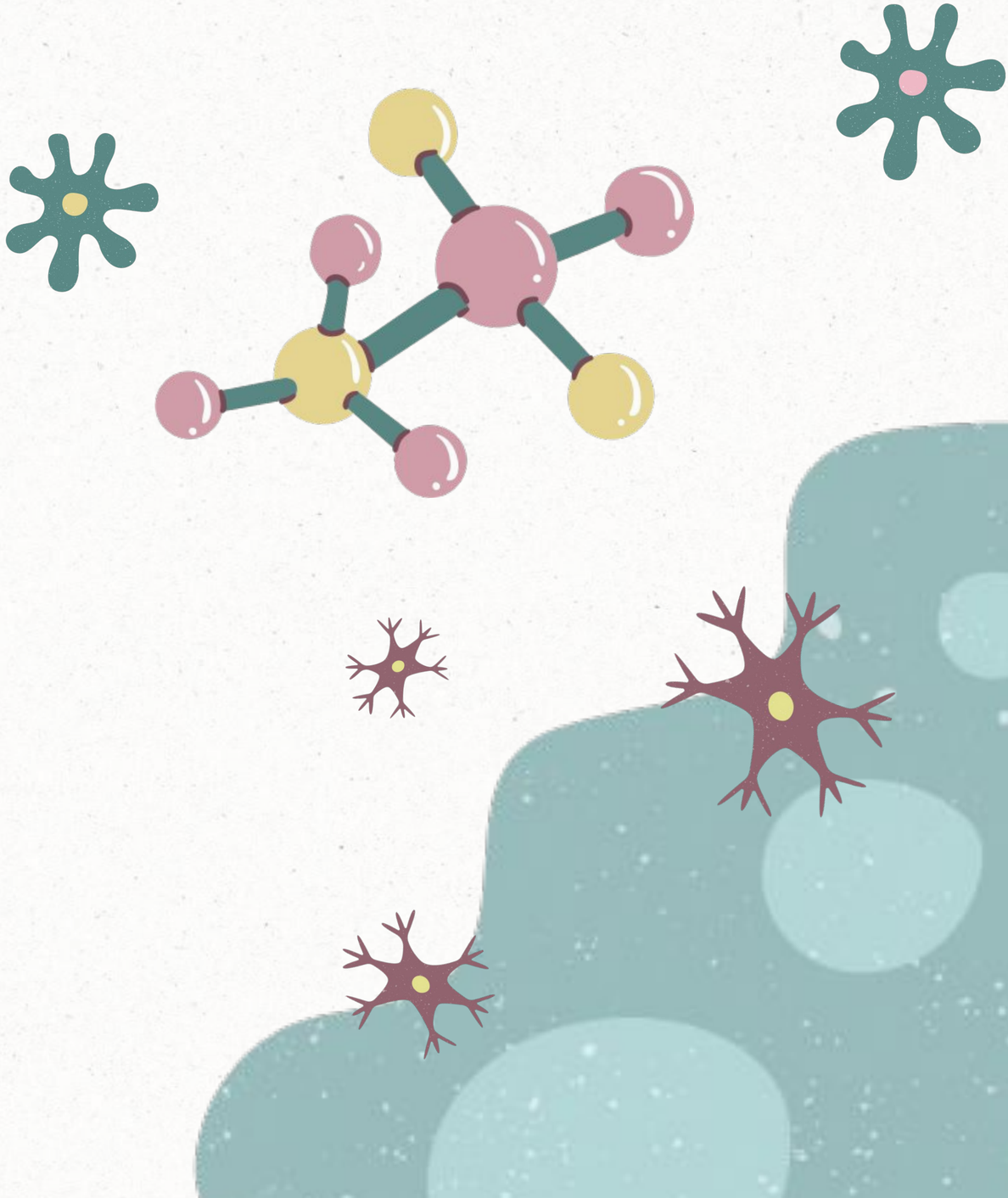
ap_hi and ap_lo



avg_glucose_level and BMI

05

Models



LOGISTIC REGRESSION

01 Cardiovascular Disease Prediction

max_iter	Accuracy
/	0.6796
1000	0.7258

When defined the Logistic Regression model, we first didn't set the max_iter, its accuracy is around 0.68. Then we set it as 1000, and found the accuracy of the model has raised. So we choose 1000 as its max_iter.

01 Cardiovascular Disease Prediction

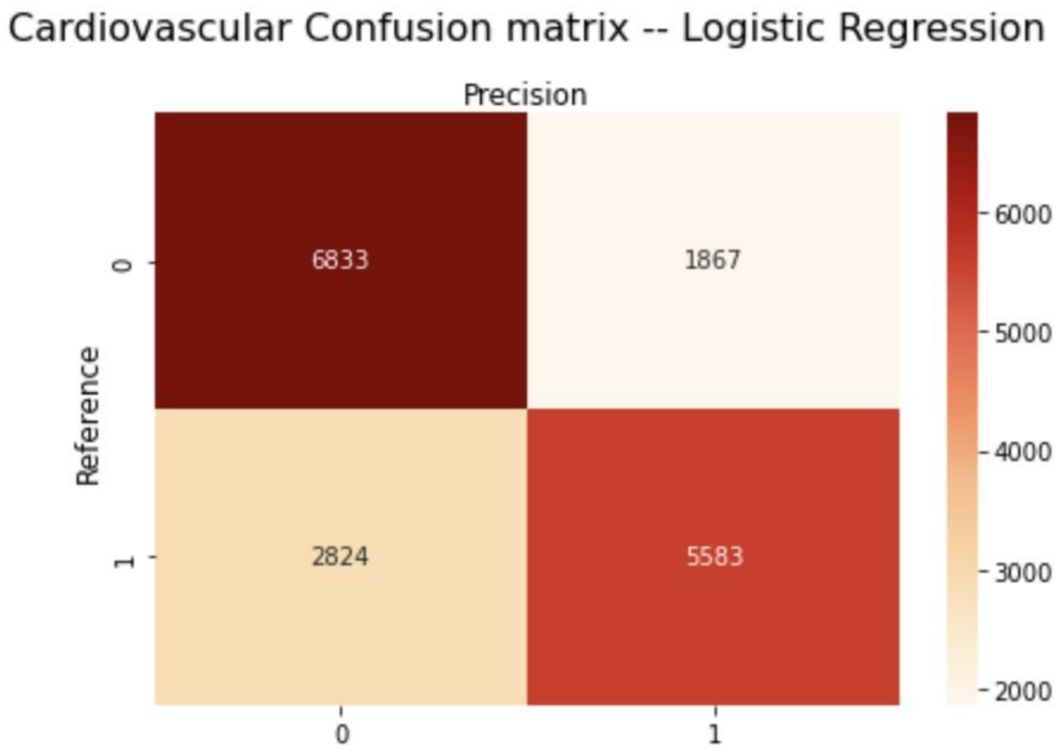
Result

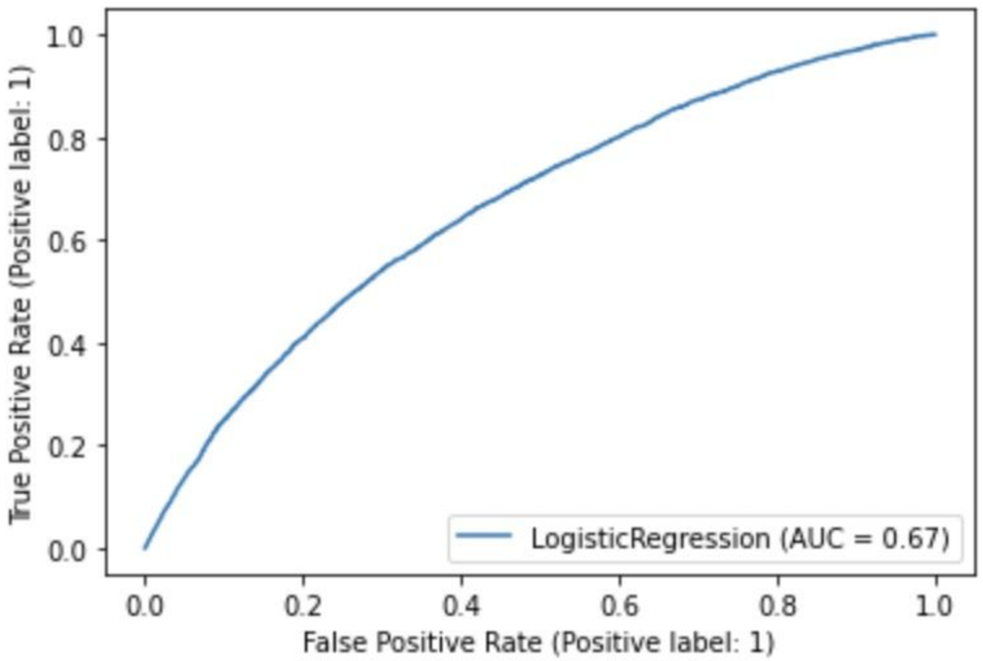
Accuracy	0.7258
Precision	0.7494
Recall	0.6641
F1 score	0.7042

Classification report

	precision	recall	f1-score	support
0	0.71	0.79	0.74	8700
1	0.75	0.66	0.70	8407
accuracy			0.73	17107
macro avg	0.73	0.72	0.72	17107
weighted avg	0.73	0.73	0.72	17107

01 Cardiovascular Disease Prediction





02 Stroke Disease Prediction

Result

Accuracy	0.9538
Precision	0.00
Recall	0.00
F1 score	0.00

Abnormal Result

Classification report

	precision	recall	f1-score	support
0	0.95	1.00	0.98	1220
1	0.00	0.00	0.00	58
accuracy			0.95	1278
macro avg	0.48	0.50	0.49	1278
weighted avg	0.91	0.95	0.93	1278

0.9538341158059468

02 Stroke Disease Prediction

Balance the stroke dataset

The reason of the meaningless result is that the dataset of stroke disease is imbalance. In the column “stroke”, the amount of “0” is much larger than “1”. Therefore, we have to balance the dataset to get more accurate classification result.

```
0    249
1    249
Name: stroke, dtype: int64
```

0.9538341158059468

02 Stroke Disease Prediction

Result

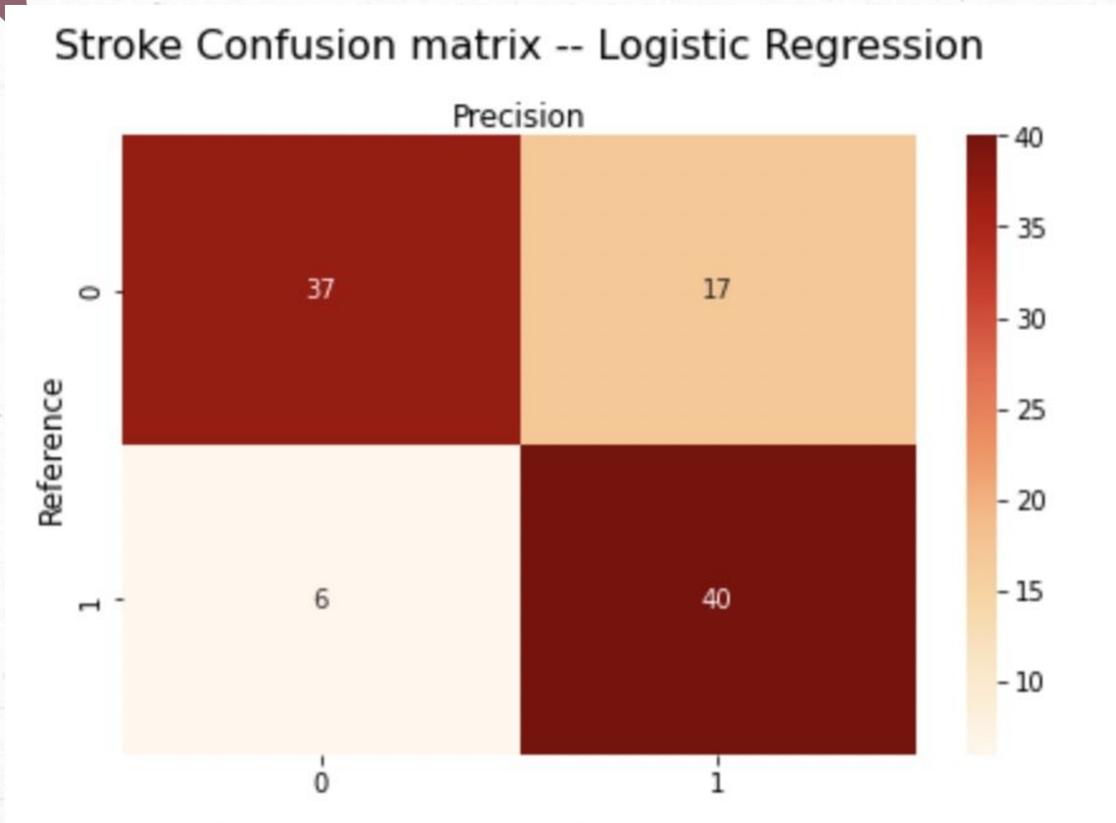
Accuracy	0.77
Precision	0.7018
Recall	0.8696
F1 score	0.7767

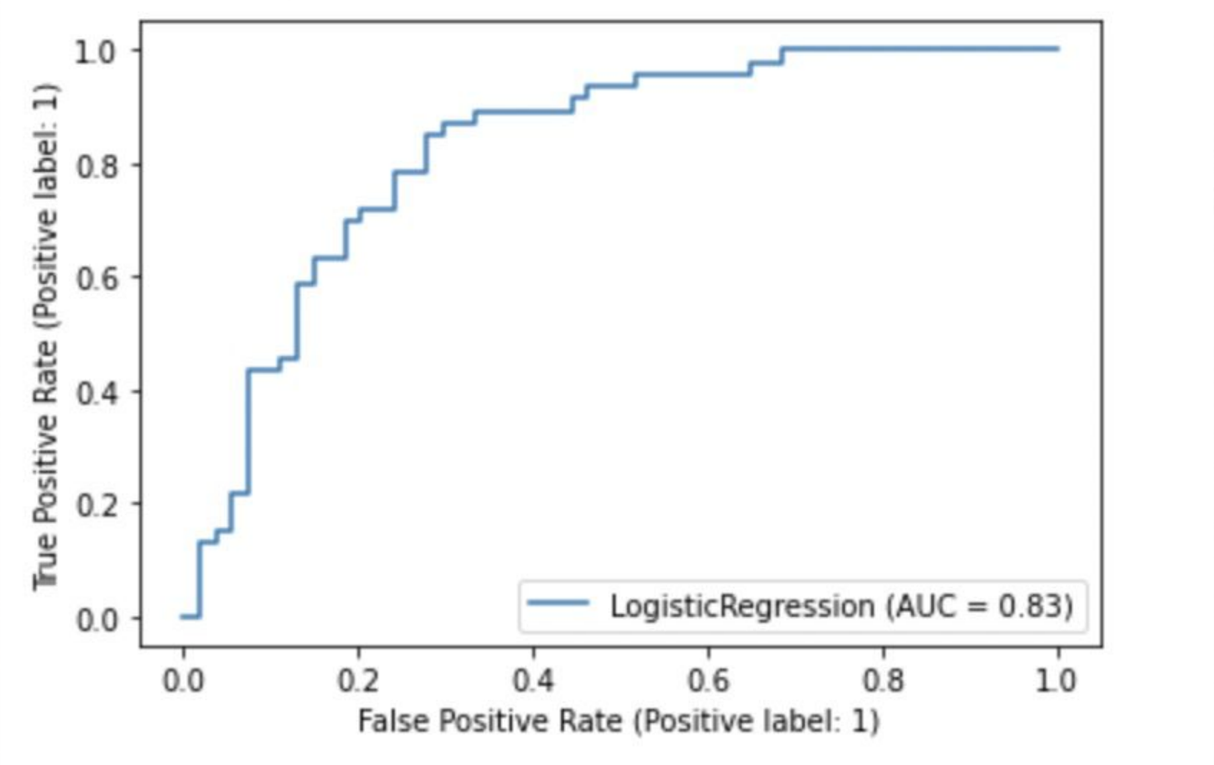
Classification report

	precision	recall	f1-score	support
0	0.86	0.69	0.76	54
1	0.70	0.87	0.78	46
accuracy			0.77	100
macro avg	0.78	0.78	0.77	100
weighted avg	0.79	0.77	0.77	100

0.9538341158059468

02 Stroke Disease Prediction





DECISION TREE

01 Cardiovascular Disease Prediction

max_depth	Accuracy
2	0.7124
3	0.7223
4	0.7262
5	0.7263
6	0.7270
7	0.7277
8	0.7263
9	0.7251

max_depth = 7
accuracy = 0.7277

01 Cardiovascular Disease Prediction

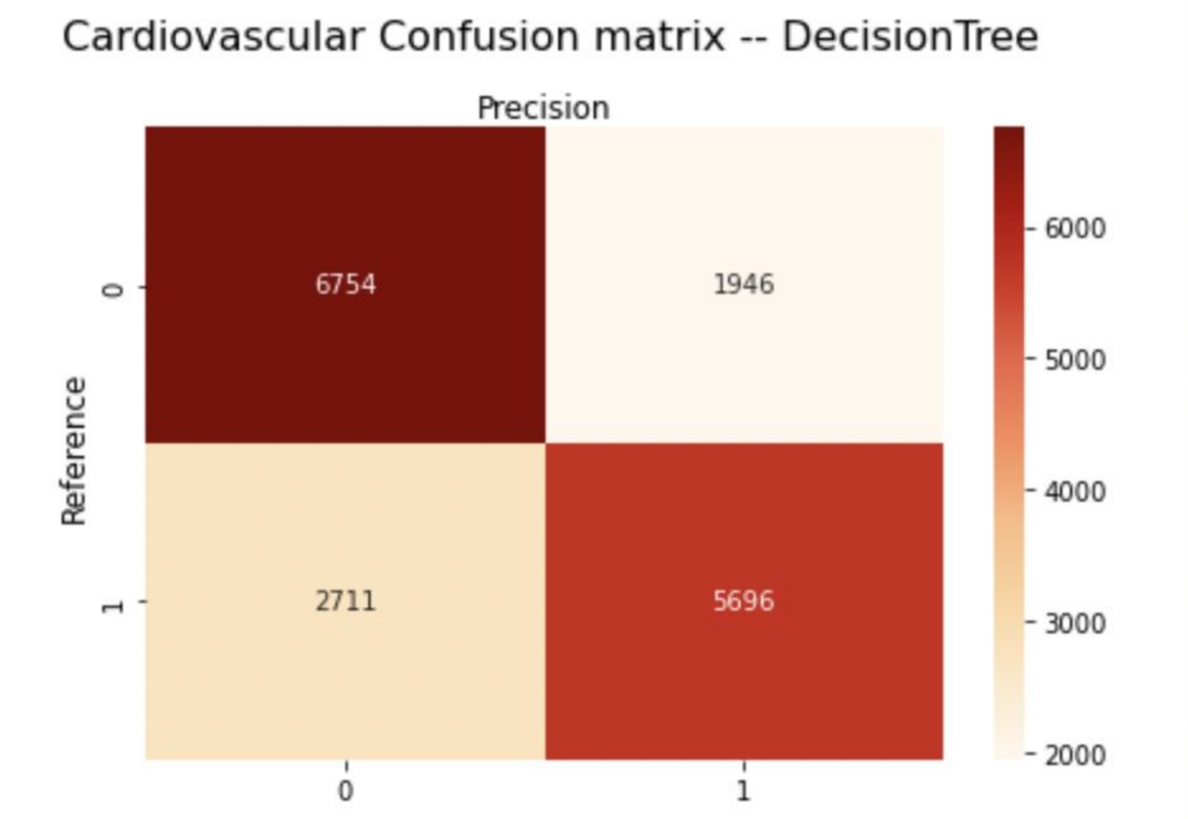
Result

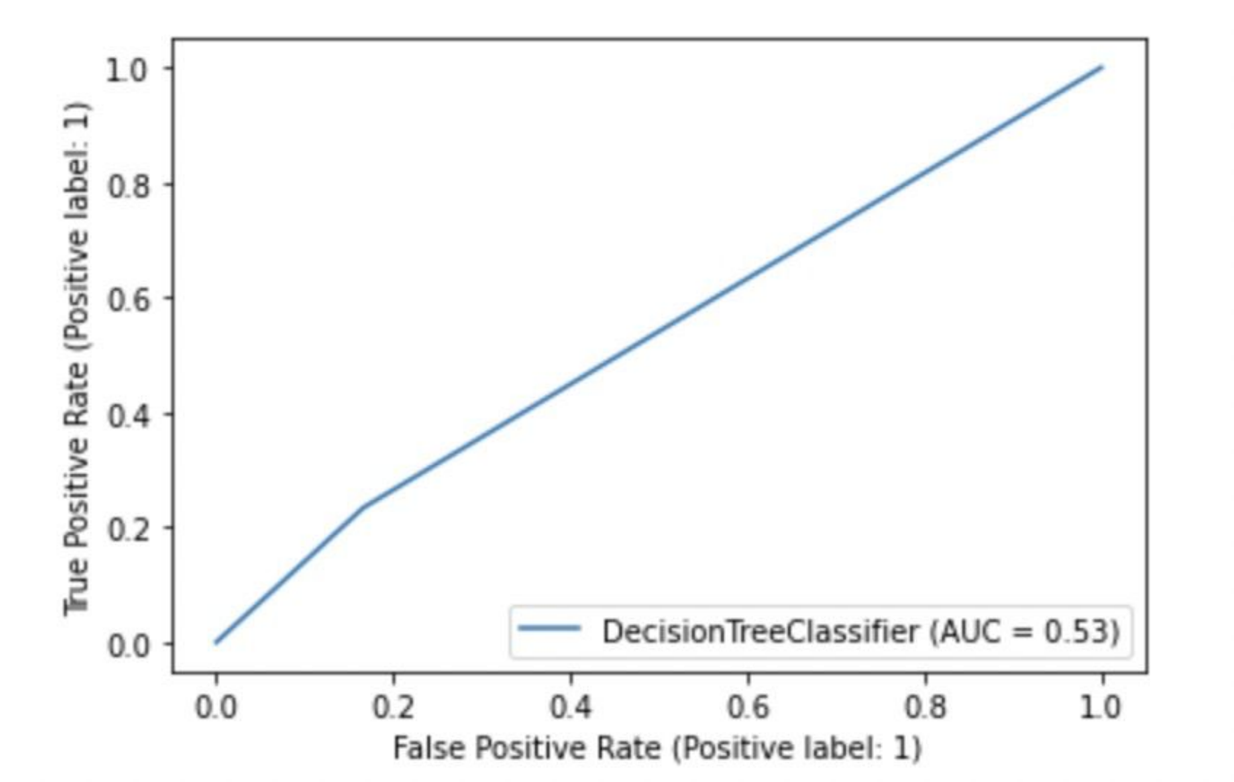
Accuracy	0.7277
Precision	0.7453
Recall	0.6775
F1 score	0.7098

Classification report

	precision	recall	f1-score	support
0	0.71	0.78	0.74	8700
1	0.75	0.68	0.71	8407
accuracy			0.73	17107
macro avg	0.73	0.73	0.73	17107
weighted avg	0.73	0.73	0.73	17107

01 Cardiovascular Disease Prediction





0.9538341158059468

02 Stroke Disease Prediction

Result

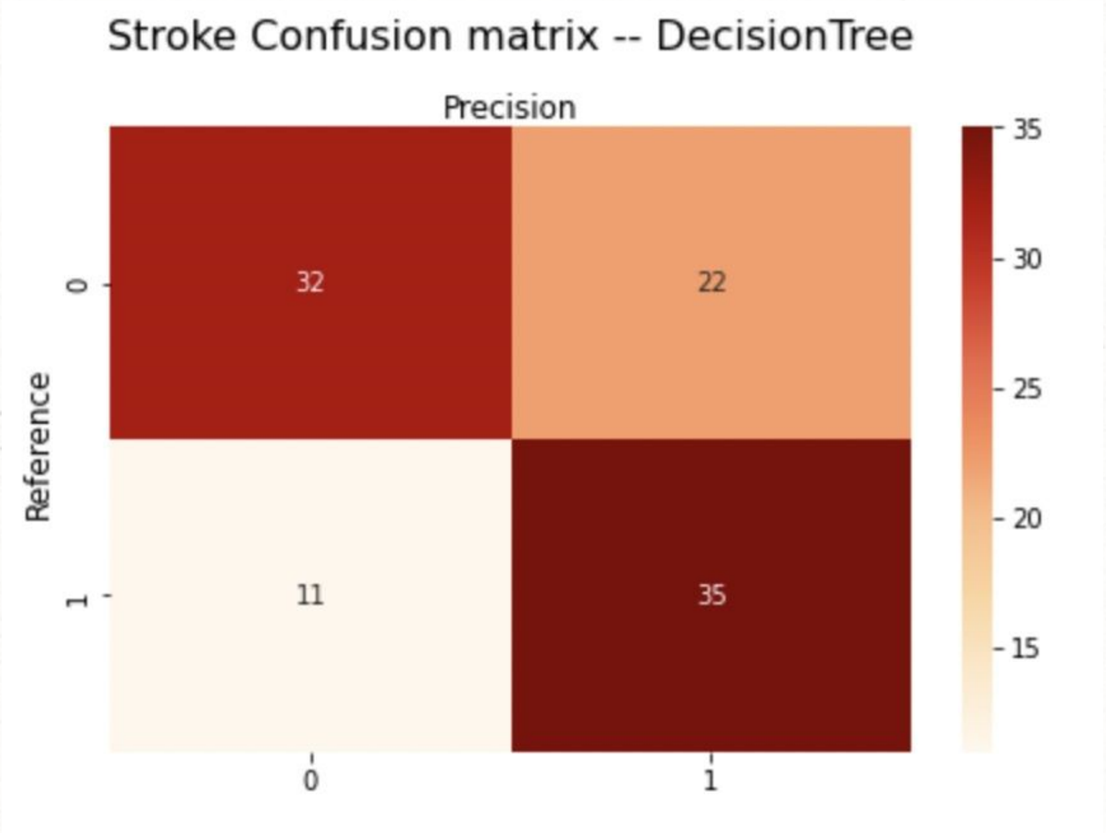
Accuracy	0.67
Precision	0.6140
Recall	0.7609
F1 score	0.6796

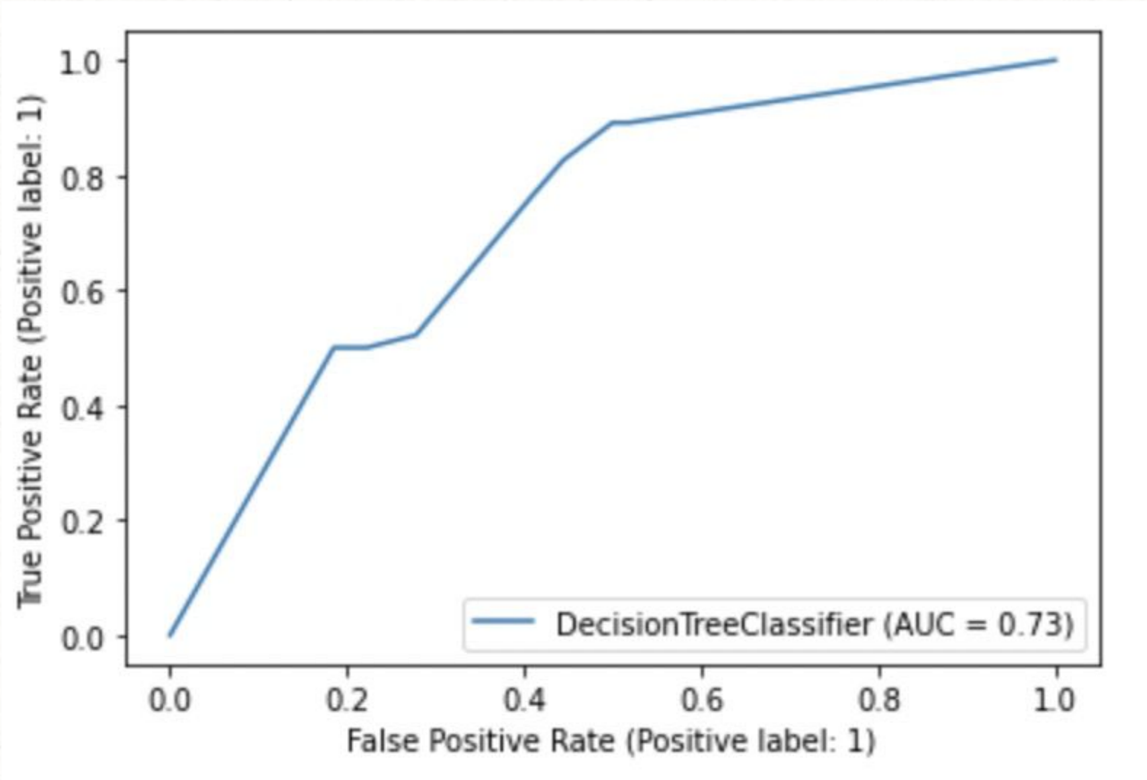
Classification report

	precision	recall	f1-score	support
0	0.74	0.59	0.66	54
1	0.61	0.76	0.68	46
accuracy			0.67	100
macro avg	0.68	0.68	0.67	100
weighted avg	0.68	0.67	0.67	100

0.9538341158059468

02 Stroke Disease Prediction





LINEAR REGRESSION

After data cleaning and feature extraction, we found that almost every attribute in each dataset may cause the corresponding disease. And as there are only numerical values in the datasets, we first tried the linear regression model to predict relations.

01 Cardiovascular Disease Relation

Parameter	Attribute	Value
w0		-1.7519
w1	age	0.0104
w2	gender	-0.0105
w3	weight	0.0020
w4	ap_hi	0.0105
w5	ap_lo	0.0020
w6	cholesterol	0.0968
w7	gluc	-0.0201
w8	smoke	-0.0307
w9	alco	-0.0336
w10	active	-0.0456

The parameter before "cholesterol" is the highest, followed by "active" and "alcohol". After normalizing by multiplying the parameters with the mean value of each column, the highest parameters lie in "ap_hi", "age" and "ap_lo".

And for qualitative analysis, the probability of getting cardiovascular disease is higher if the subject is an older female with heavier weight, higher blood pressure, cholesterol, abnormal glucose level, do not smoke and drink alcohol and do not exercise. Most of the results correspond to our expectations except for smoking and drinking.

02 Stroke Disease Relation

Parameter	Attribute	Value
w0		-0.1998
w1	age	0.0140
w2	gender	0.0147
w3	hypertension	0.0550
w4	heart_disease	0.0648
w5	ever_married	-0.0853
w6	work_type	-0.0267
w7	Residence_type	0.0716
w8	avg_glucose_level	0.0005
w9	BMI	-0.0034
w10	smoking_status	0.0357

the parameter before "ever_married", "Residence_type" and " heart_disease" are the highest. After normalization by multiplying the parameters with the mean value of each column, the highest parameters lie in "age", "ave_glucose_level" and "bmi".

Also, for qualitative analysis, the probability of getting stroke disease is higher if the subject is an older female with higher blood pressure and glucose level, and with heart disease, not married, living in urban areas, lower BMI and smoke. Except for BMI, other features correspond to our assumptions as some features are linked to each other.

SVM MODEL

SVC, Support Vector Classifier, is a supervised machine learning algorithm typically used for classification tasks. So it may be suitable for this task.

01 Cardiovascular Disease Prediction

Data Set	Linear SVM (C=1)	Linear SVM (C=0.5)	Non-Linear SVM
Cardiovascular	0.7222	0.7222	0.7157

```
confusion matrics:
 [[7096 1604]
 [3149 5258]]


classification report:
              precision    recall  f1-score   support

           0       0.69      0.82      0.75      8700
           1       0.77      0.63      0.69      8407

    accuracy                           0.72     17107
   macro avg       0.73      0.72      0.72     17107
weighted avg       0.73      0.72      0.72     17107
```

In this section, we used Linear SVM model to deal with the dataset.

We tuned the parameter of "C" with values of "0.5" and "1". However, we found it made little difference on the accuracy.

As the accuracy shows, the model did not perform well in this task. So we continued to try non-linear SVM models.

The accuracy did not improve a lot. So we continued to try other models in the following steps.

02 Stroke Disease Prediction

Data Set	Linear SVM (C=1)	Linear SVM (C=0.5)	Non-Linear SVM
Cardiovascular	0.76	0.77	0.6

```
confusion matrics:
 [[37 17]
 [ 6 40]]


classification report:
              precision    recall  f1-score   support

           0       0.86      0.69      0.76        54
           1       0.70      0.87      0.78        46

    accuracy                           0.77       100
   macro avg       0.78      0.78      0.77       100
weighted avg       0.79      0.77      0.77       100
```

We also tried the SVM for the stroke dataset. The model accuracy is 0.76 for Stroke Dataset.

Naive Bayes

01 Cardiovascular Disease Prediction

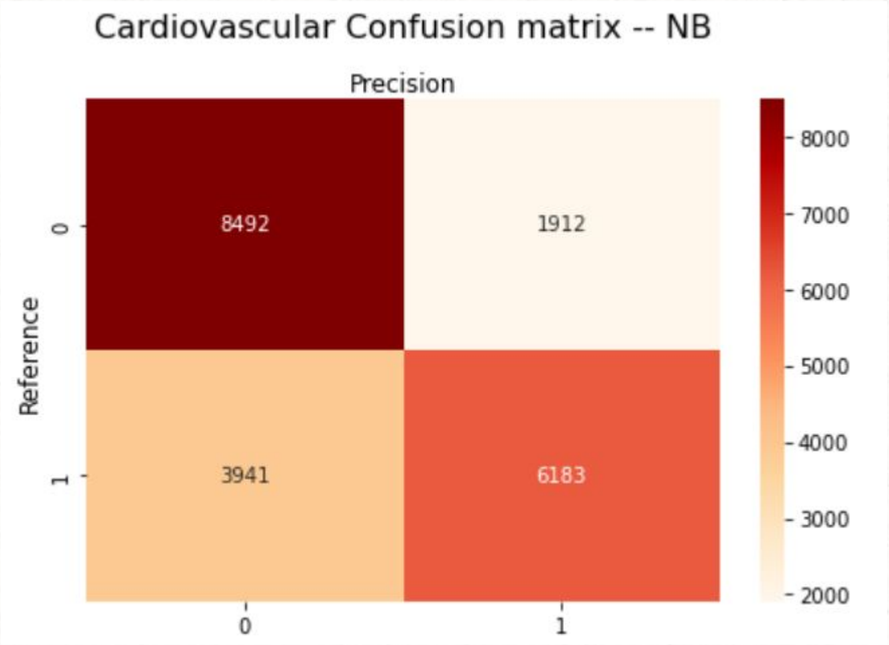
Results

Accuracy	0.7148
Precision	0.7638
Recall	0.6107
F1 score	0.6787

```
classification report:
              precision    recall  f1-score   support

           0       0.68      0.82      0.74     10404
           1       0.76      0.61      0.68     10124

    accuracy                           0.71     20528
   macro avg       0.72      0.71      0.71     20528
weighted avg       0.72      0.71      0.71     20528
```



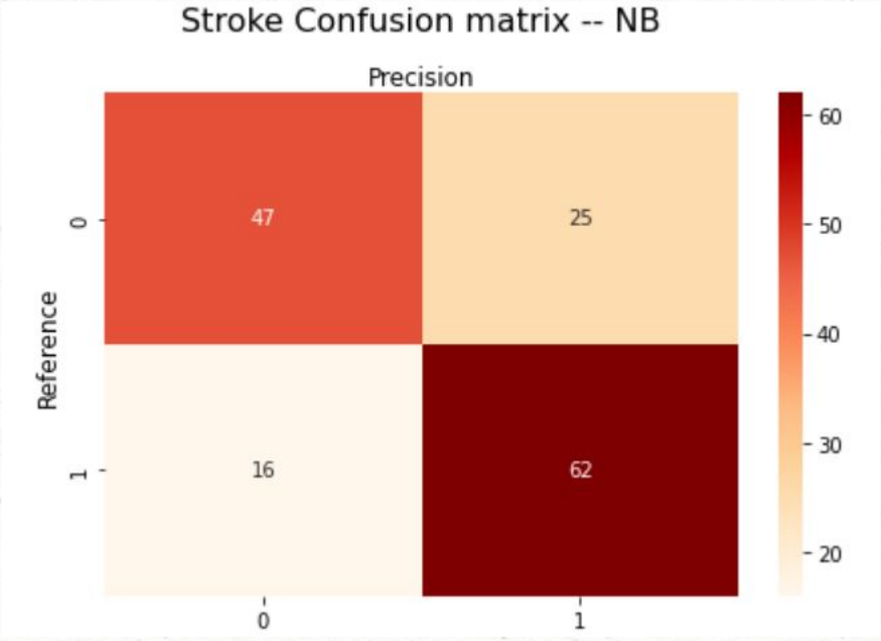
02 Stroke Disease Prediction

Accuracy	0.7267
Precision	0.7126
Recall	0.7948
F1 score	0.7515

```
classification report:
              precision    recall  f1-score   support

           0       0.75      0.65      0.70        72
           1       0.71      0.79      0.75        78

    accuracy                           0.73       150
   macro avg       0.73      0.72      0.72       150
weighted avg       0.73      0.73      0.73       150
```



Multi-layer Perceptron

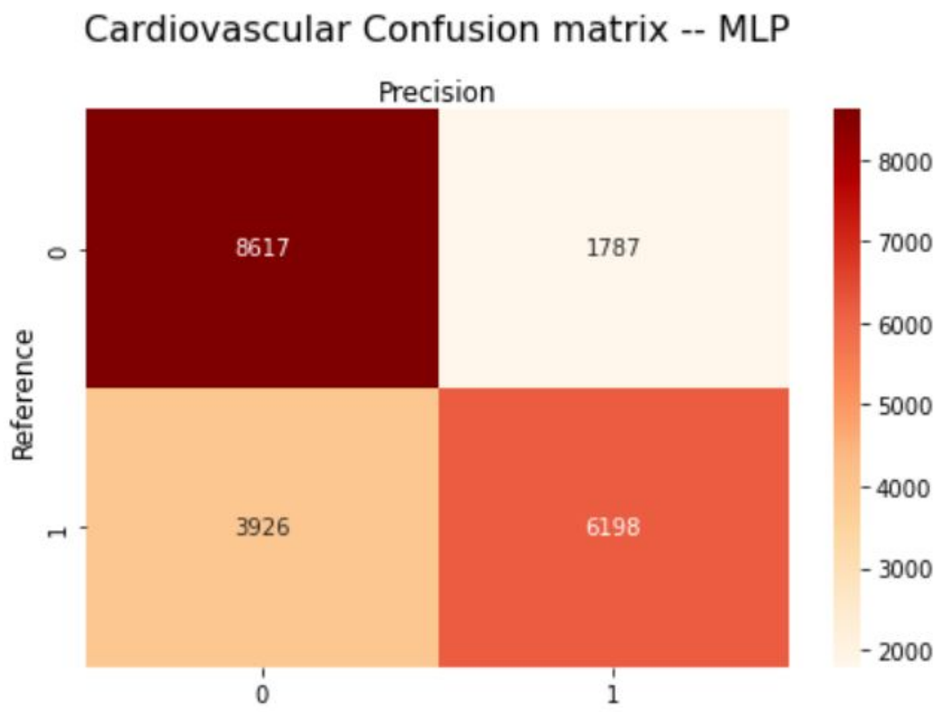
01 Cardiovascular Disease Prediction

Accuracy	0.7216
Precision	0.7762
Recall	0.6122
F1 score	0.6845

```
classification report:
              precision    recall  f1-score   support

           0       0.69      0.83      0.75     10404
           1       0.78      0.61      0.68     10124

    accuracy                           0.72     20528
   macro avg       0.73      0.72      0.72     20528
weighted avg       0.73      0.72      0.72     20528
```



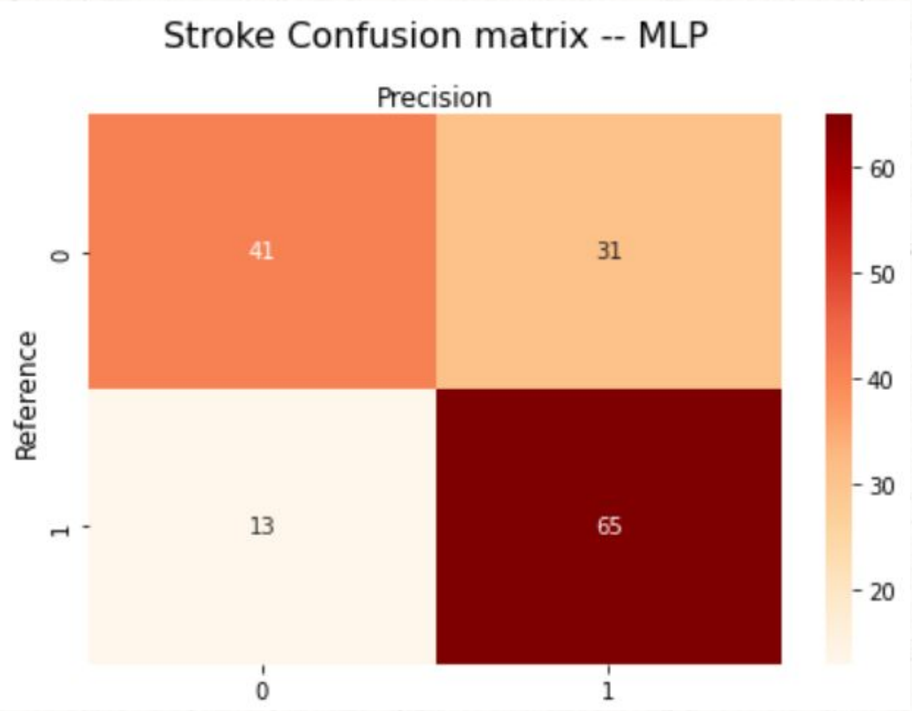
02 Stroke Disease Prediction

Accuracy	0.7066
Precision	0.6771
Recall	0.8333
F1 score	0.7471

```
classification report:
              precision    recall  f1-score   support

           0       0.76      0.57      0.65        72
           1       0.68      0.83      0.75        78

    accuracy                           0.71       150
   macro avg       0.72      0.70      0.70       150
weighted avg       0.72      0.71      0.70       150
```



CNN

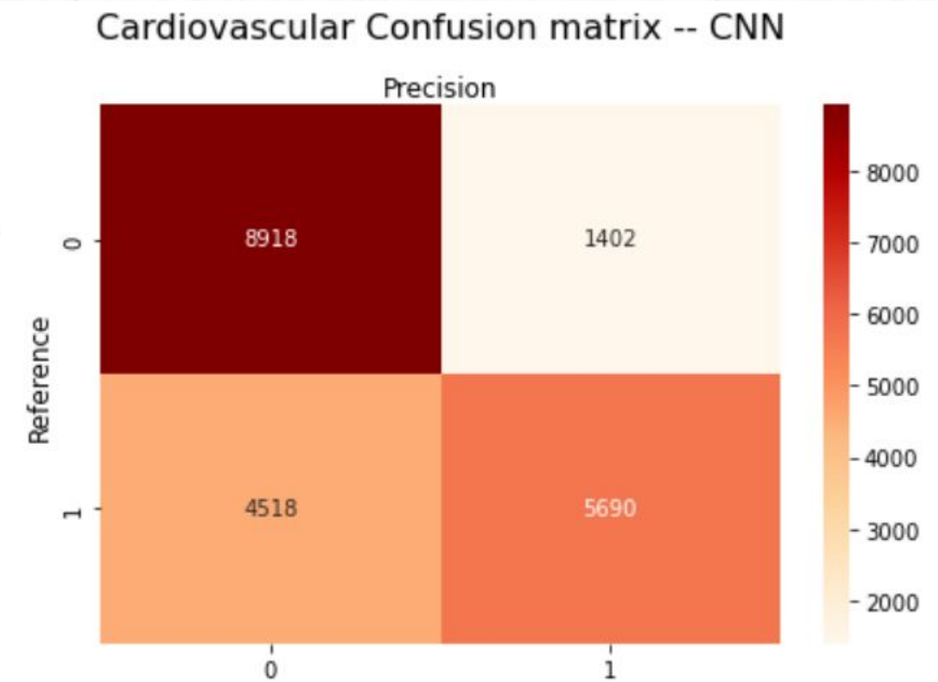
01 Cardiovascular Disease Prediction

Accuracy	0.7102
Precision	0.7966
Recall	0.5614
F1 score	0.6642

```
classification report:
              precision    recall  f1-score   support

         0.0       0.66      0.86      0.75     10320
         1.0       0.80      0.56      0.66     10208

    accuracy                           0.71     20528
   macro avg       0.73      0.71      0.70     20528
weighted avg       0.73      0.71      0.70     20528
```



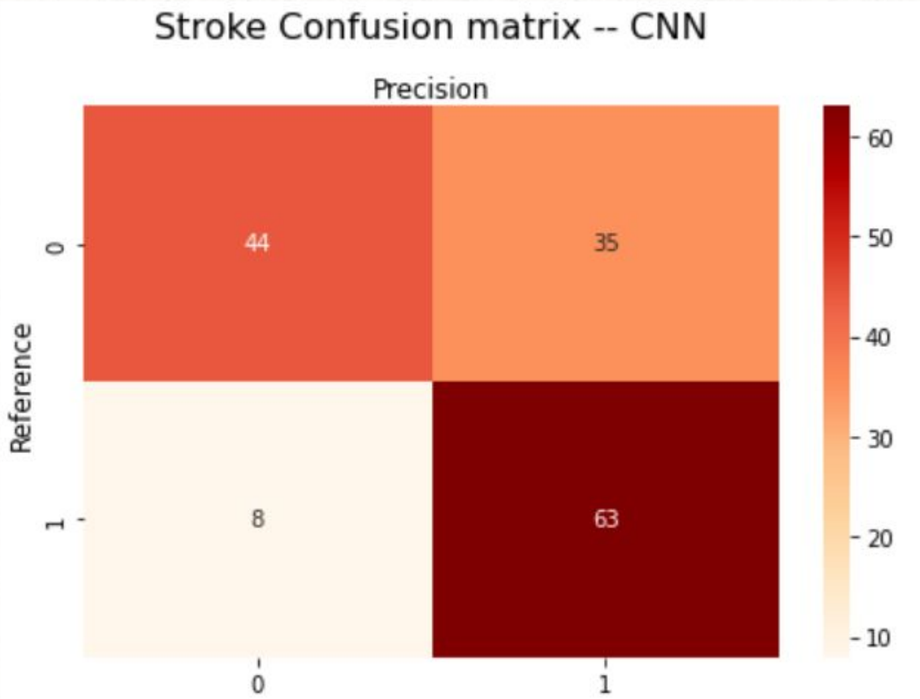
02 Stroke Disease Prediction

Accuracy	0.7133
Precision	0.6408
Recall	0.8904
F1 score	0.7522

```
classification report:
              precision    recall  f1-score   support

         0.0       0.85      0.56      0.67        79
         1.0       0.64      0.89      0.75        71

    accuracy                           0.71       150
   macro avg       0.74      0.72      0.71       150
weighted avg       0.75      0.71      0.71       150
```



Ensemble Learning

Methods

BAGGING

Reduce the variance in a model

Random forest

BOOSTING

Reduce the bias in a model

GBDT

STACKING

Increase the prediction accuracy of a model

Random forest &
Logistic Regression

Grid Search

01 Cardiovascular Disease

Parameters

Bagging

max_depth	10
min_samples_leaf	2
min_samples_split	10
n_estimators	200

Boosting

learning_rate	0.1
max_depth	5
n_estimators	50

Results

Bagging (Random forest)

accuracy	0.7316
f1	0.7083

Boosting (GBDT)

accuracy	0.7332
f1	0.7143

Stacking

accuracy	0.7308
f1	0.7091

Grid Search

02 Stroke Disease

Parameters

Bagging

max_depth	3
min_samples_leaf	4
min_samples_split	10
n_estimators	250

Boosting

learning_rate	0.1
max_depth	3
n_estimators	50

Results

Bagging (Random forest)

accuracy	0.7400
f1	0.7593

Boosting (GBDT)

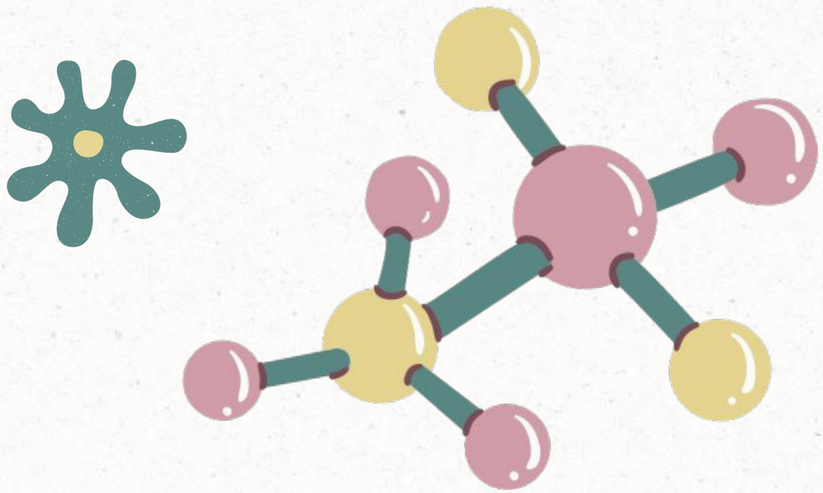
accuracy	0.7900
f1	0.7414

Stacking

accuracy	0.7500
f1	0.7788

06

Conclusion



01 Model Results for Cardiovascular Dataset

Model	Logistic Regression	Decision Tree	SVM
Accuracy	0.7258	0.7277	0.7222
F1	0.7042	0.7098	0.6888

Model	Naive Bayes	MLP	CNN	Ensemble Learning
Accuracy	0.7148	0.7145	0.7133	**0.7332**
F1	0.6787	0.6845	0.6842	0.7143

02 Model Results for Stroke Dataset

Model	Logistic Regression	Decision Tree	SVM
Accuracy	0.7700	0.6800	0.7700
F1	0.7767	0.6796	0.7767

Model	Naive Bayes	MLP	CNN	Ensemble Learning
Accuracy	0.7267	0.6933	0.7400	**0.7900**
F1	0.7515	0.7471	0.6022	0.7414

CONCLUSIONS #1

Cardiovascular Diseases

We found that the ap_hi which represents systolic blood pressure is the dominant factor that causes cardiovascular diseases. The conclusion has been supported by exploratory data analysis and linear regression model. Other factors such as age, ap_low (diastolic blood pressure) and, cholesterol are also important factors causing cardiovascular diseases.

Stroke Diseases

As for stroke disease, age is the dominant factor. To some degrees, other factors such as hypertension, heart_disease, avg_glucose_level also cause stroke.

CONCLUSIONS #2

Grid Search and Ensemble Learning

Compared with a single machine learning model, combining grid search and ensemble learning can further improve the accuracy and generalization ability of the model.

Through grid search, we can find the optimal combination of hyperparameters for each model, and through ensemble learning, we can combine the prediction results of multiple models to obtain more accurate prediction results.

Uneven Dataset

For a small and unbalanced dataset, it is easy to have a low recall, so in the case of a small dataset, the oversampling and undersampling method can be chosen to adjust.

THANK YOU!