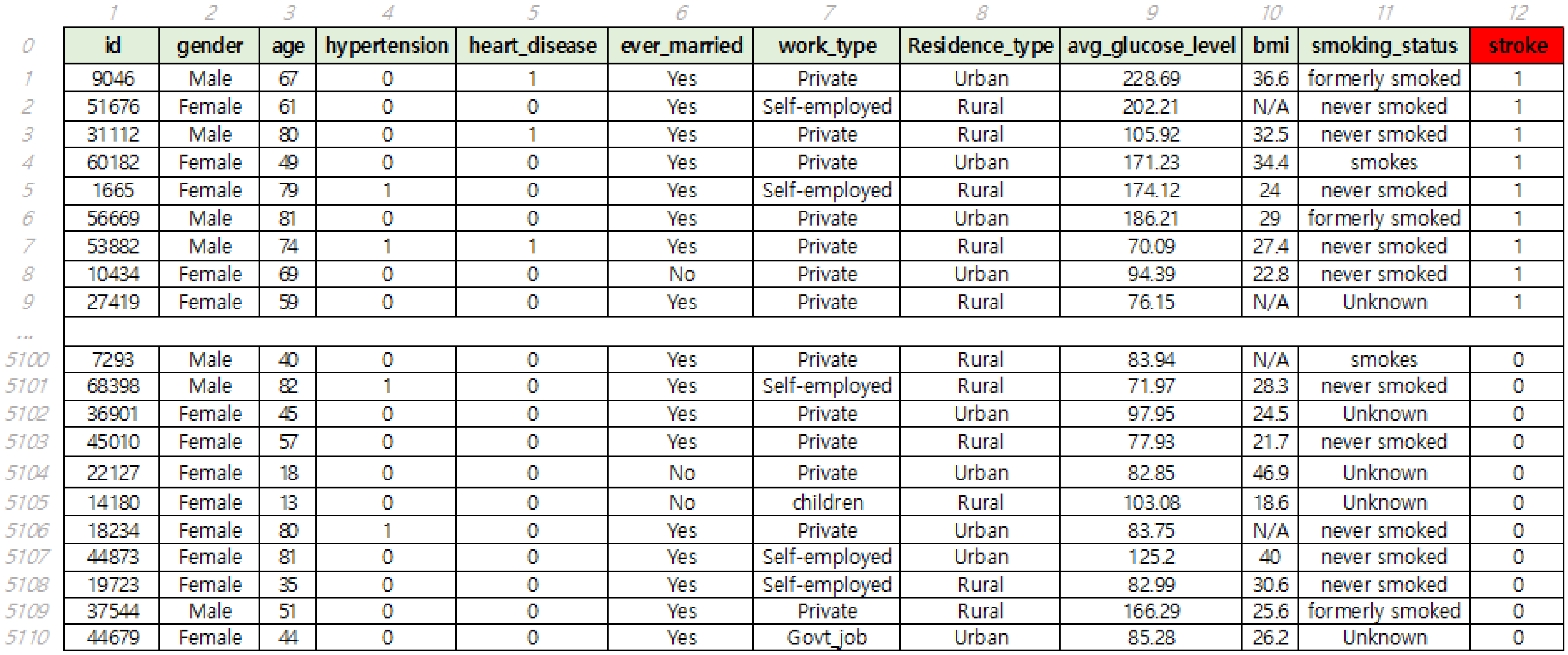
Stroke

Part 1 : Domain knowledge, EDA, Decision tree

Preview

- **To predict whether a patient is likely to get stroke** based on the input parameters like gender, age, various diseases, and smoking status.

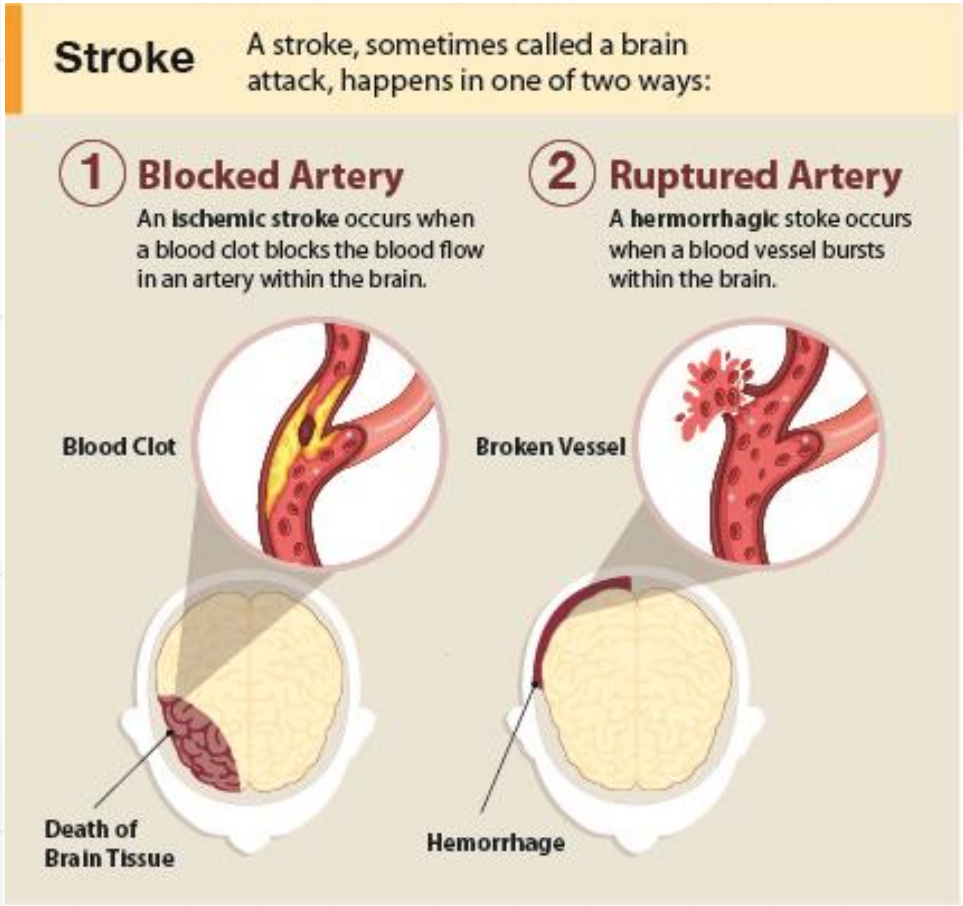
- 11 attributes, 1 stroke class, 5110 records



	1	2	3	4	5	6	7	8	9	10	11	12
0	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
2	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
3	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
4	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
5	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
6	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
7	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
8	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
9	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
...												
5100	7293	Male	40	0	0	Yes	Private	Rural	83.94	N/A	smokes	0
5101	68398	Male	82	1	0	Yes	Self-employed	Rural	71.97	28.3	never smoked	0
5102	36901	Female	45	0	0	Yes	Private	Urban	97.95	24.5	Unknown	0
5103	45010	Female	57	0	0	Yes	Private	Rural	77.93	21.7	never smoked	0
5104	22127	Female	18	0	0	No	Private	Urban	82.85	46.9	Unknown	0
5105	14180	Female	13	0	0	No	children	Rural	103.08	18.6	Unknown	0
5106	18234	Female	80	1	0	Yes	Private	Urban	83.75	N/A	never smoked	0
5107	44873	Female	81	0	0	Yes	Self-employed	Urban	125.2	40	never smoked	0
5108	19723	Female	35	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5109	37544	Male	51	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5110	44679	Female	44	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

Attribute

- **Stroke**
  - Sometimes called a **brain attack**, occurs when something blocks blood supply to part of the brain or when a blood vessel in the brain bursts.
  - A stroke can cause lasting brain damage, long-term disability, or even death
- **Hypertension**
  - Also called **High blood pressure**, is blood pressure that is higher than normal.



Stroke A stroke, sometimes called a brain attack, happens in one of two ways:

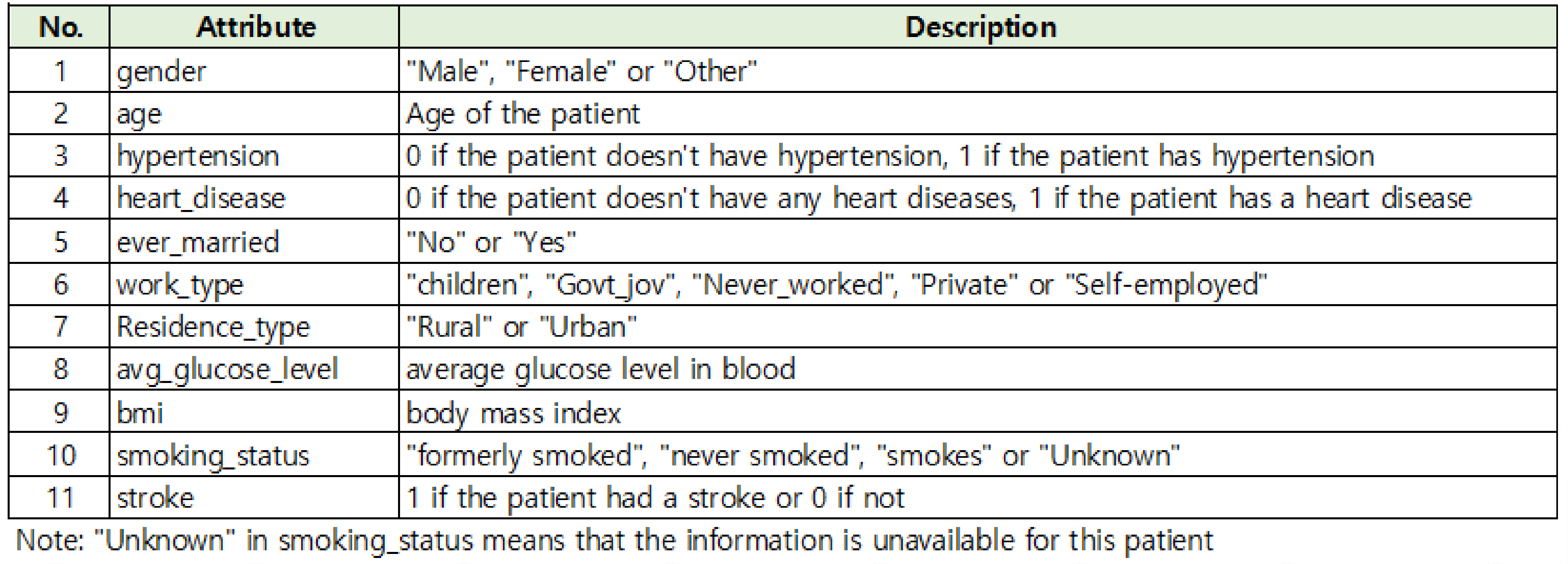
1 Blocked Artery

An **ischemic stroke** occurs when a blood clot blocks the blood flow in an artery within the brain.

2 Ruptured Artery

A **hermorrhagic** stoke occurs when a blood vessel bursts within the brain.

Attribute

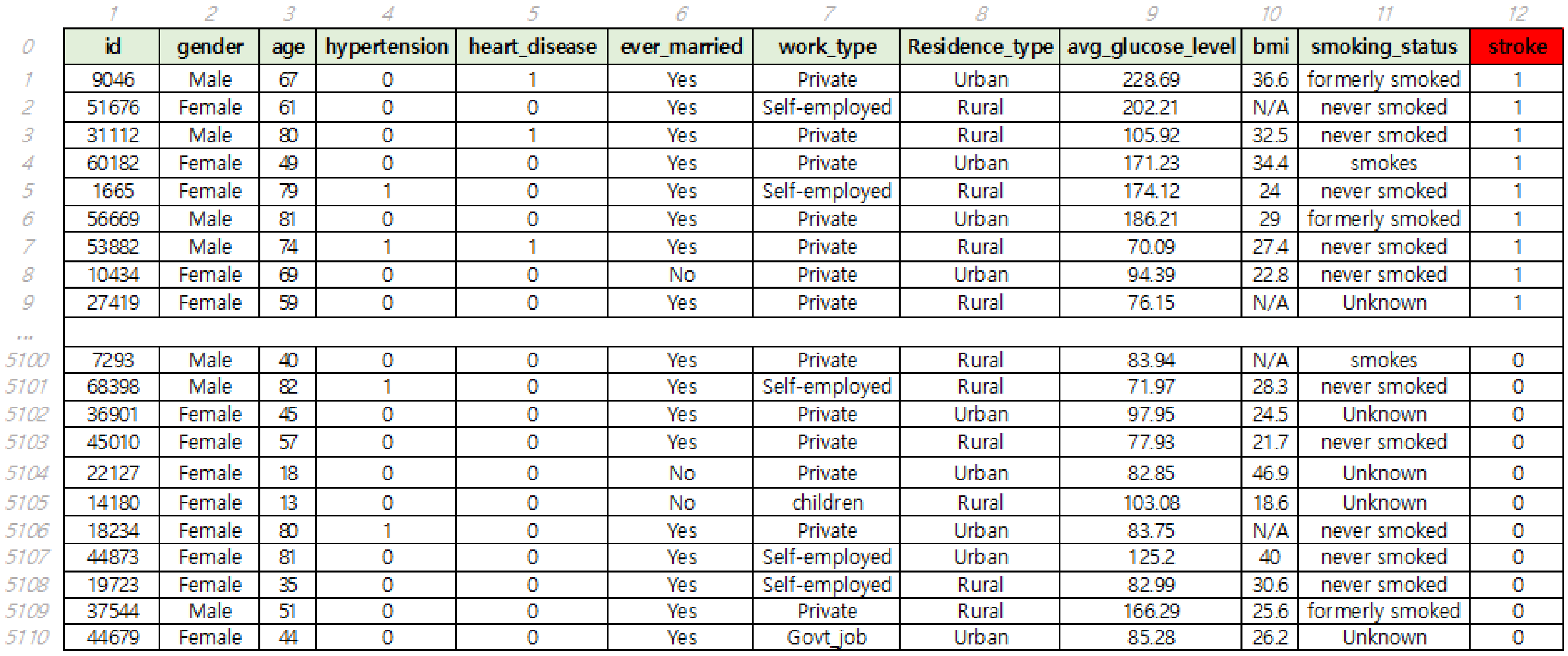


No.	Attribute	Description
1	gender	"Male", "Female" or "Other"
2	age	Age of the patient
3	hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
4	heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
5	ever_married	"No" or "Yes"
6	work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
7	Residence_type	"Rural" or "Urban"
8	avg_glucose_level	average glucose level in blood
9	bmi	body mass index
10	smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"
11	stroke	1 if the patient had a stroke or 0 if not

Note: "Unknown" in smoking_status means that the information is unavailable for this patient

Preprocessing

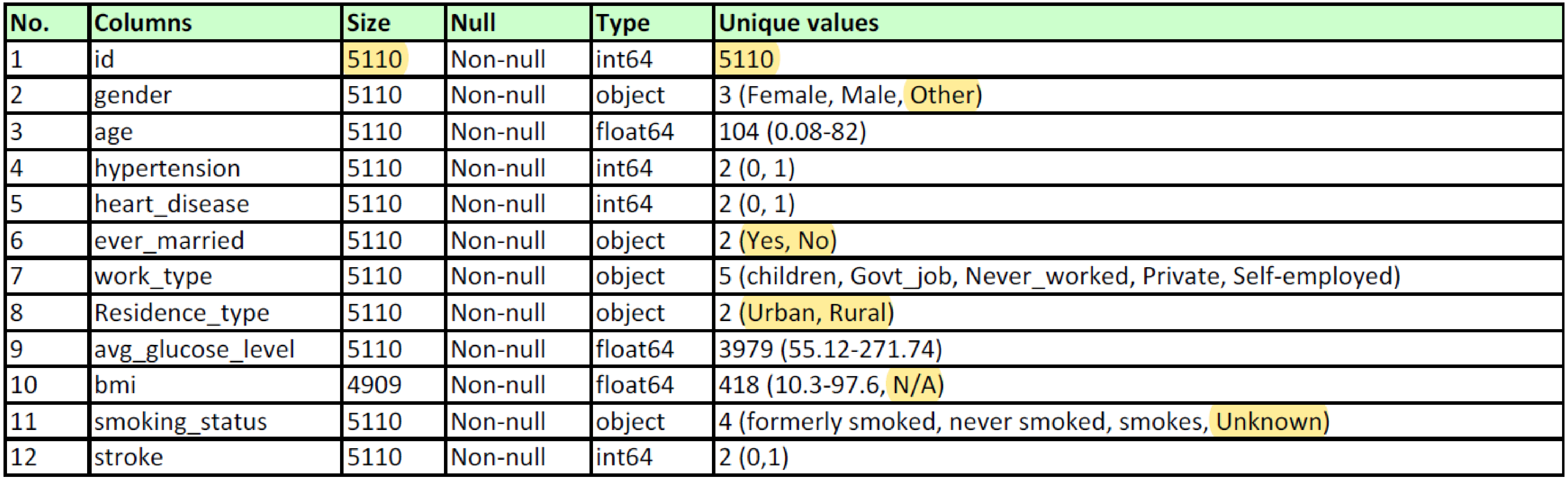
- Numerical variable (4): id, age, avg_glucose_level, bmi
- Categorical variable (7) : gender, hypertension, heart_disease, ever_married, work_type, Residence_type, smoking_status
- Stroke class: 0 = No stroke, 1 = Stroke



	1	2	3	4	5	6	7	8	9	10	11	12
0	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
2	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
3	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
4	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
5	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
6	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
7	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
8	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
9	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
...												
5100	7293	Male	40	0	0	Yes	Private	Rural	83.94	N/A	smokes	0
5101	68398	Male	82	1	0	Yes	Self-employed	Rural	71.97	28.3	never smoked	0
5102	36901	Female	45	0	0	Yes	Private	Urban	97.95	24.5	Unknown	0
5103	45010	Female	57	0	0	Yes	Private	Rural	77.93	21.7	never smoked	0
5104	22127	Female	18	0	0	No	Private	Urban	82.85	46.9	Unknown	0
5105	14180	Female	13	0	0	No	children	Rural	103.08	18.6	Unknown	0
5106	18234	Female	80	1	0	Yes	Private	Urban	83.75	N/A	never smoked	0
5107	44873	Female	81	0	0	Yes	Self-employed	Urban	125.2	40	never smoked	0
5108	19723	Female	35	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5109	37544	Male	51	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5110	44679	Female	44	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

Preprocessing

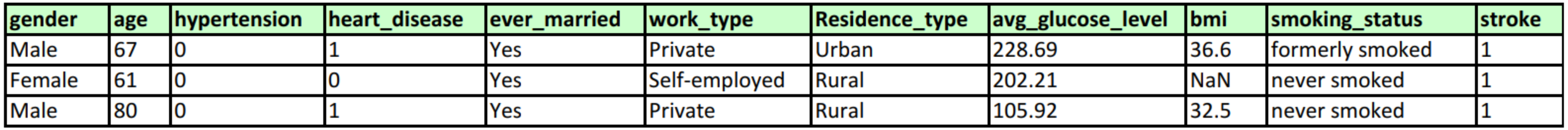
- Raw data show that ‘bmi’ variable has missing value = 201, No duplicate value in any column
- 3 variables have float64 type (age, avg_glucose_level, bmi)
- 4 variables have int64 (id, hypertension, heart_disease, stroke)
- 5 variables have object type (gender, ever_married, word_type, Residence_type, smoking_status)



No.	Columns	Size	Null	Type	Unique values
1	id	5110	Non-null	int64	5110
2	gender	5110	Non-null	object	3 (Female, Male, Other)
3	age	5110	Non-null	float64	104 (0.08-82)
4	hypertension	5110	Non-null	int64	2 (0, 1)
5	heart_disease	5110	Non-null	int64	2 (0, 1)
6	ever_married	5110	Non-null	object	2 (Yes, No)
7	work_type	5110	Non-null	object	5 (children, Govt_job, Never_worked, Private, Self-employed)
8	Residence_type	5110	Non-null	object	2 (Urban, Rural)
9	avg_glucose_level	5110	Non-null	float64	3979 (55.12-271.74)
10	bmi	4909	Non-null	float64	418 (10.3-97.6, N/A)
11	smoking_status	5110	Non-null	object	4 (formerly smoked, never smoked, smokes, Unknown)
12	stroke	5110	Non-null	int64	2 (0,1)

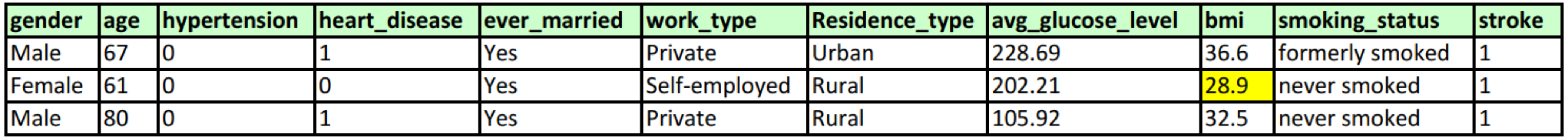
Preprocessing

- ‘ID’ variable has all unique value -> **remove ‘ID’ column**



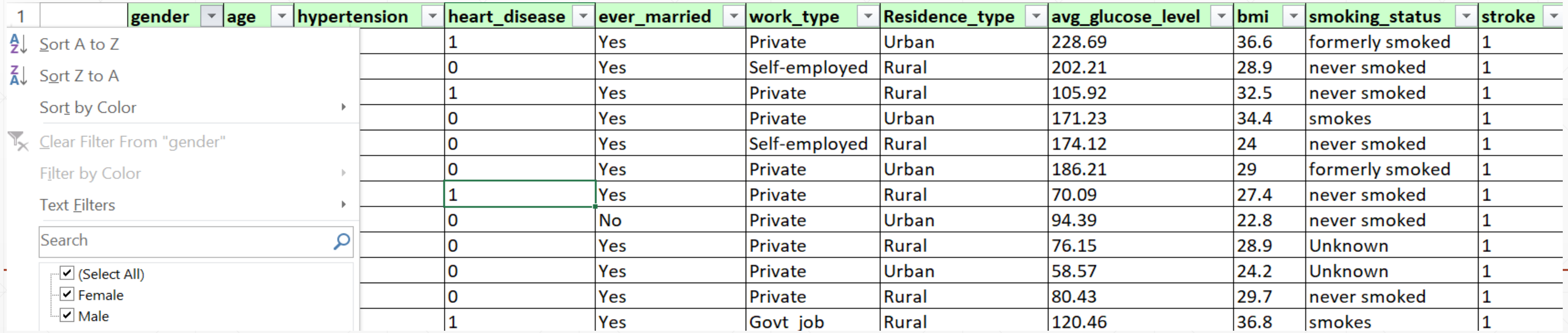
gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
Female	61	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1

- ‘bmi’ variable has missing value -> **replace with average value = 28.9**



gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
Female	61	0	0	Yes	Self-employed	Rural	202.21	28.9	never smoked	1
Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1

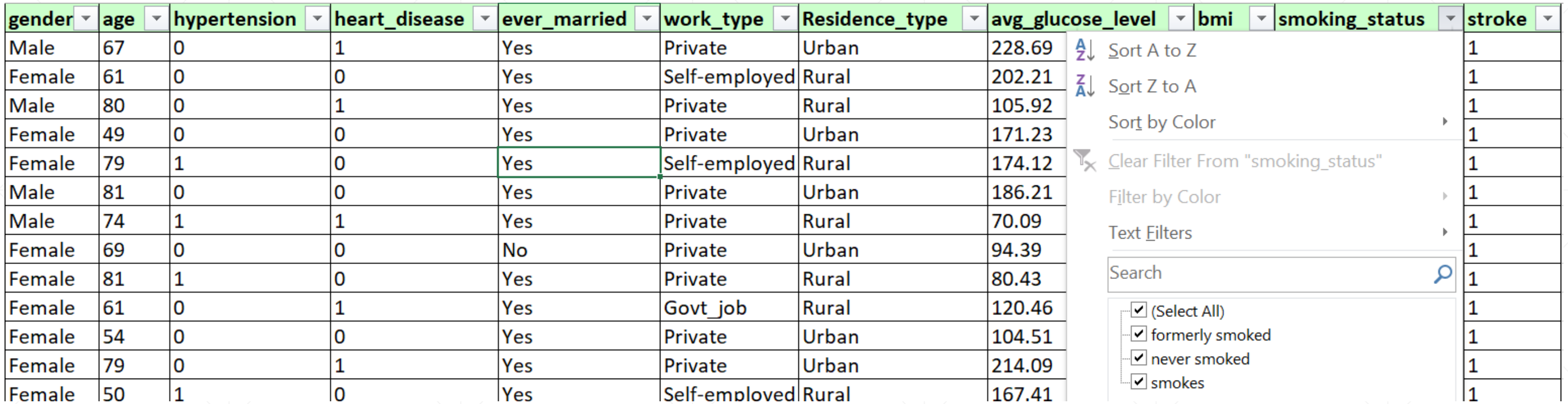
- ‘gender’ variable has only 1 ‘Other’ value -> **remove ‘Other’ value**



1	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
				1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
				0	Yes	Self-employed	Rural	202.21	28.9	never smoked	1
				1	Yes	Private	Rural	105.92	32.5	never smoked	1
				0	Yes	Private	Urban	171.23	34.4	smokes	1
				0	Yes	Self-employed	Rural	174.12	24	never smoked	1
				0	Yes	Private	Urban	186.21	29	formerly smoked	1
				1	Yes	Private	Rural	70.09	27.4	never smoked	1
				0	No	Private	Urban	94.39	22.8	never smoked	1
				0	Yes	Private	Rural	76.15	28.9	Unknown	1
				0	Yes	Private	Urban	58.57	24.2	Unknown	1
				0	Yes	Private	Rural	80.43	29.7	never smoked	1
				1	Yes	Govt_job	Rural	120.46	36.8	smokes	1

Preprocessing

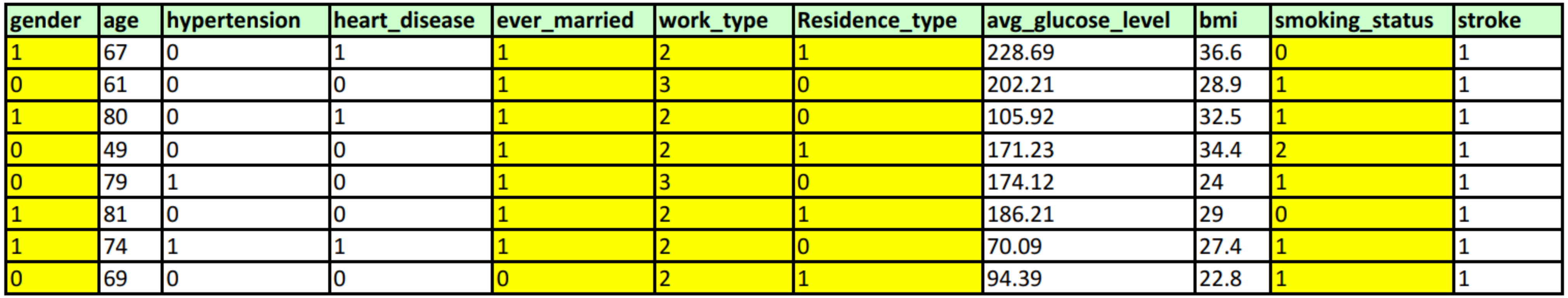
- ‘smoking_status’ variable has ‘Unknown’ value -> **remove ‘Unknown’ value**



gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	67	0	1	Yes	Private	Urban	228.69			1
Female	61	0	0	Yes	Self-employed	Rural	202.21			1
Male	80	0	1	Yes	Private	Rural	105.92			1
Female	49	0	0	Yes	Private	Urban	171.23			1
Female	79	1	0	Yes	Self-employed	Rural	174.12			1
Male	81	0	0	Yes	Private	Urban	186.21			1
Male	74	1	1	Yes	Private	Rural	70.09			1
Female	69	0	0	No	Private	Urban	94.39			1
Female	81	1	0	Yes	Private	Rural	80.43			1
Female	61	0	1	Yes	Govt_job	Rural	120.46			1
Female	54	0	0	Yes	Private	Urban	104.51			1
Female	79	0	1	Yes	Private	Urban	214.09			1
Female	50	1	0	Yes	Self-employed	Rural	167.41			1

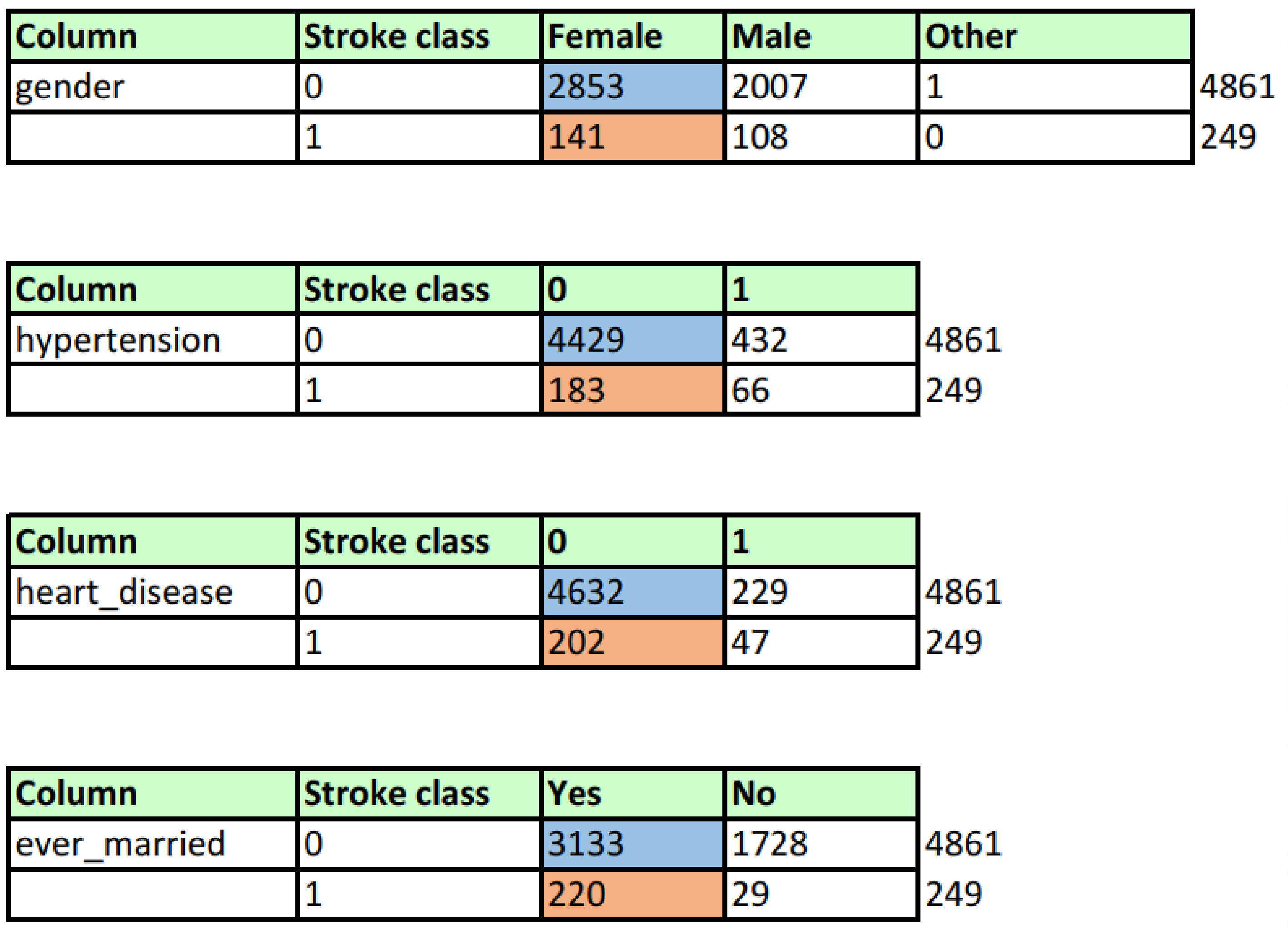
☑ (Select All)
☑ formerly smoked
☑ never smoked
☑ smokes

- 'ever_married', 'Residence_type', 'gender‘, 'work_type', 'smoking_status‘ -> **Convert categorical variable to numerical variable**



gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	67	0	1	1	2	1	228.69	36.6	0	1
0	61	0	0	1	3	0	202.21	28.9	1	1
1	80	0	1	1	2	0	105.92	32.5	1	1
0	49	0	0	1	2	1	171.23	34.4	2	1
0	79	1	0	1	3	0	174.12	24	1	1
1	81	0	0	1	2	1	186.21	29	0	1
1	74	1	1	1	2	0	70.09	27.4	1	1
0	69	0	0	0	2	1	94.39	22.8	1	1

Summary by stroke class



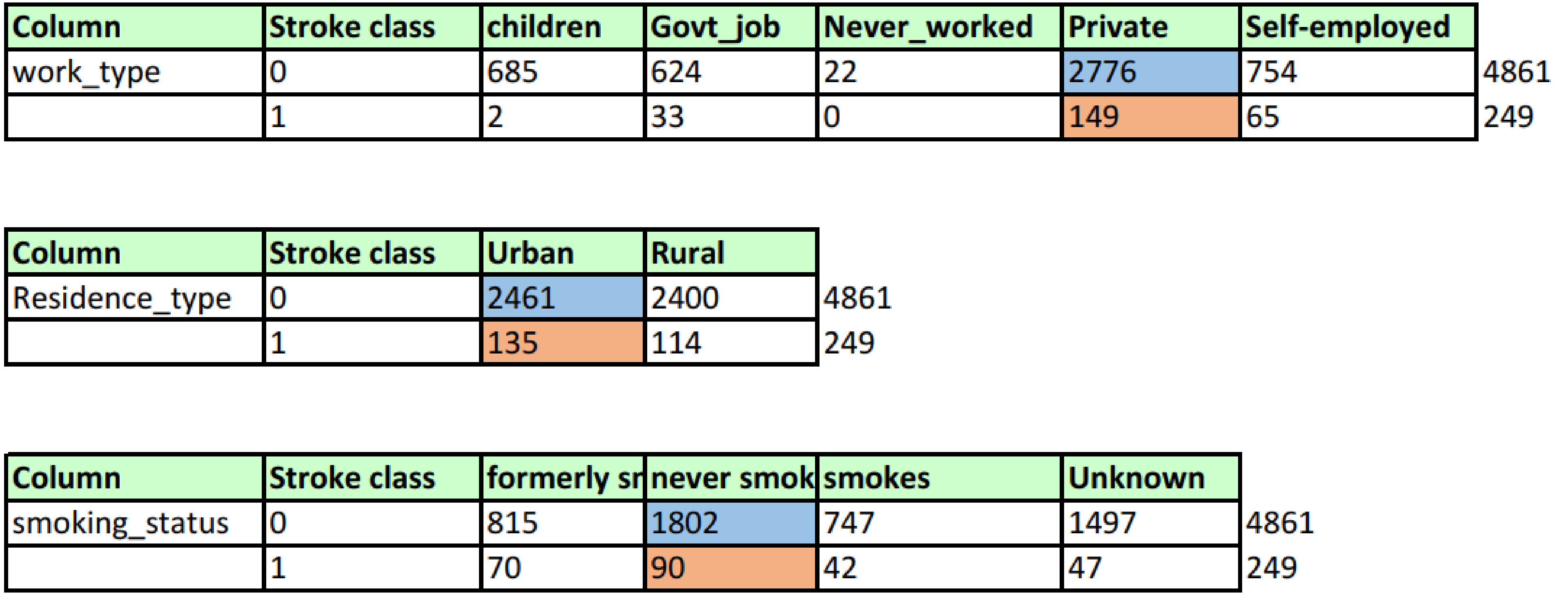
Column	Stroke class	Female	Male	Other	
gender	0	2853	2007	1	4861
	1	141	108	0	249

Column	Stroke class	0	1	
hypertension	0	4429	432	4861
	1	183	66	249

Column	Stroke class	0	1	
heart_disease	0	4632	229	4861
	1	202	47	249

Column	Stroke class	Yes	No	
ever_married	0	3133	1728	4861
	1	220	29	249

Summary by stroke class

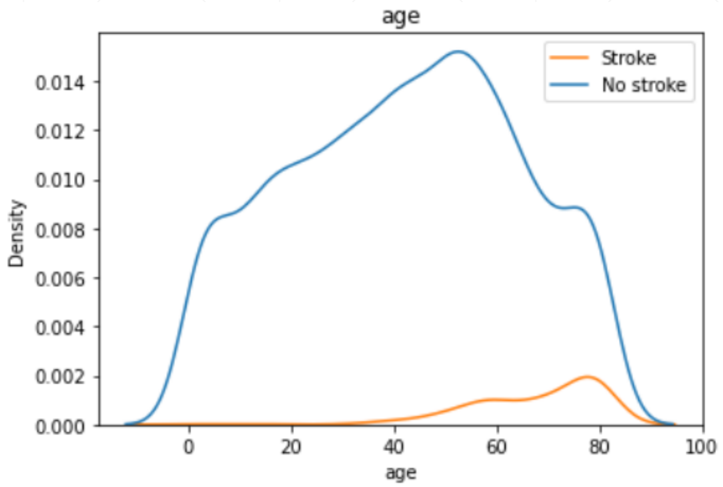


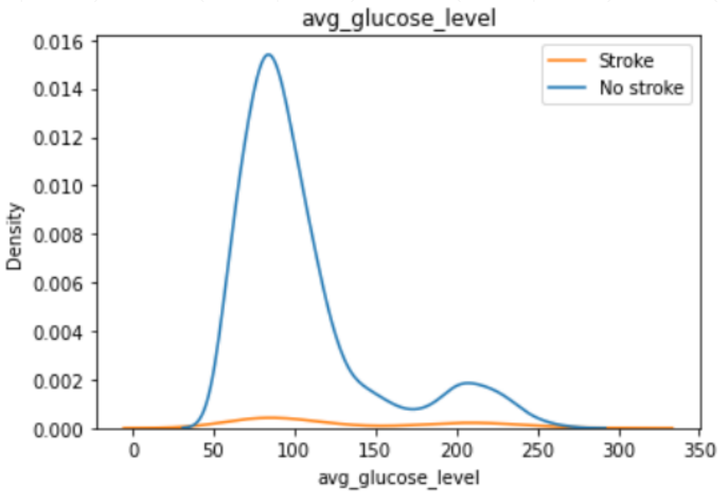
Column	Stroke class	children	Govt_job	Never_worked	Private	Self-employed	
work_type	0	685	624	22	2776	754	4861
	1	2	33	0	149	65	249

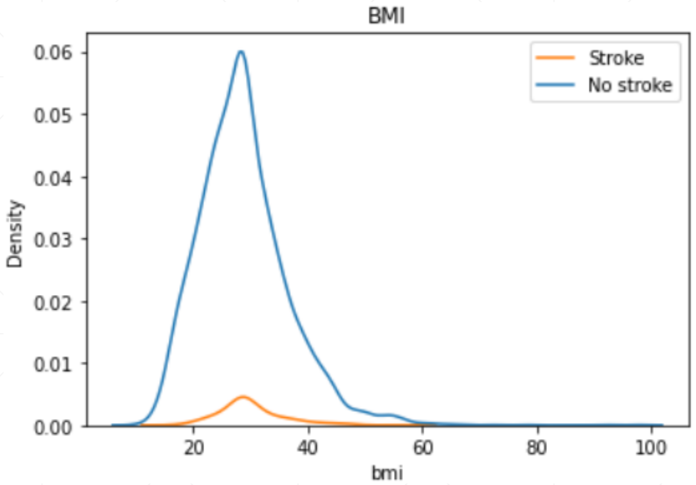
Column	Stroke class	Urban	Rural	
Residence_type	0	2461	2400	4861
	1	135	114	249

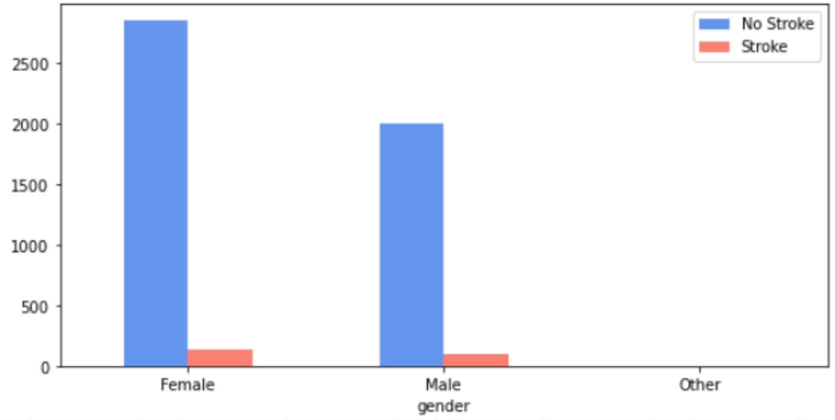
Column	Stroke class	formerly sm	never smok	smokes	Unknown	
smoking_status	0	815	1802	747	1497	4861
	1	70	90	42	47	249

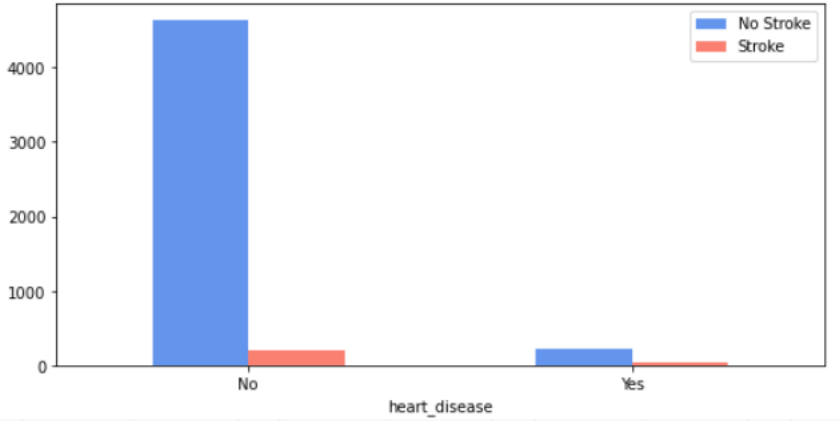
Univariate Analysis



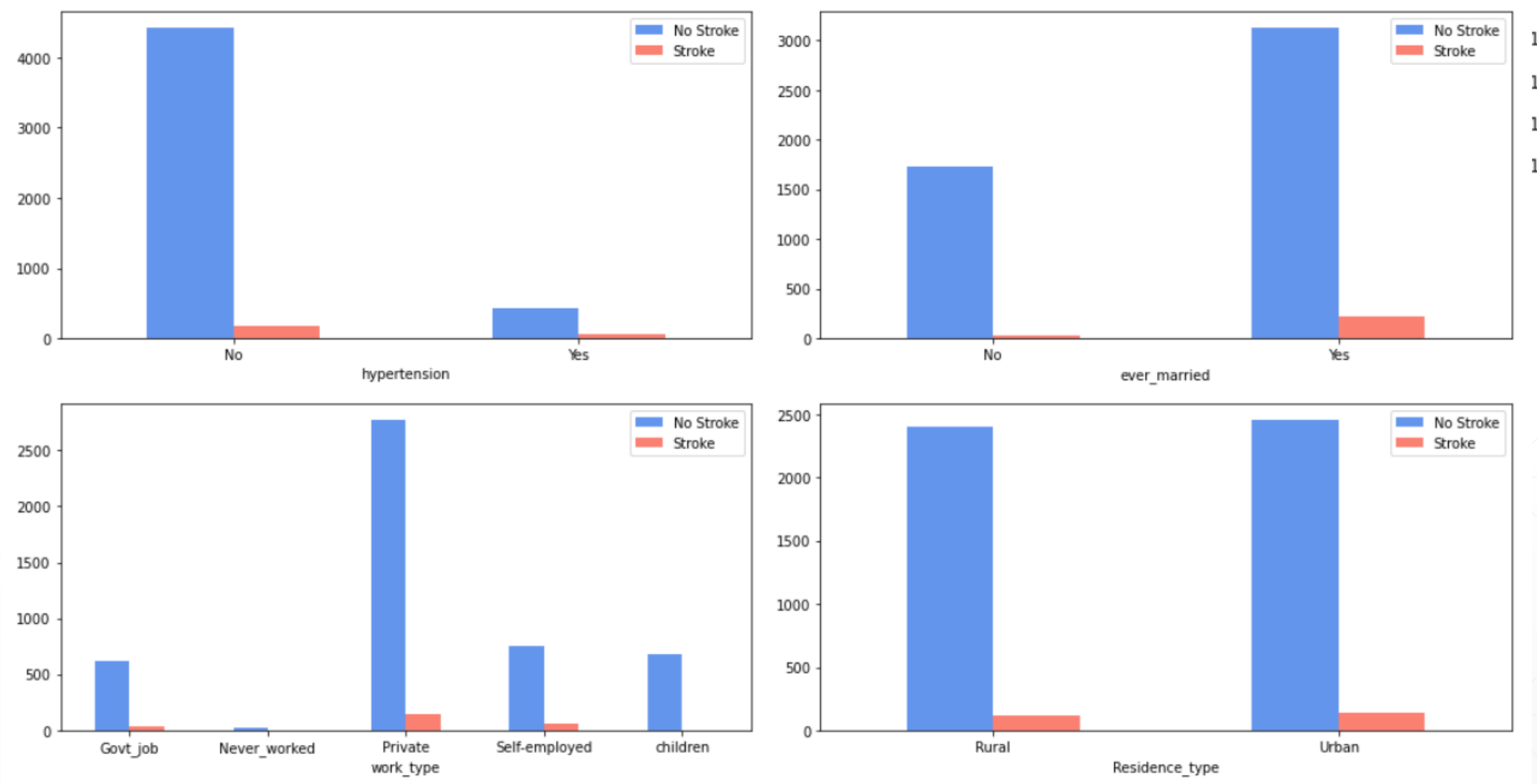


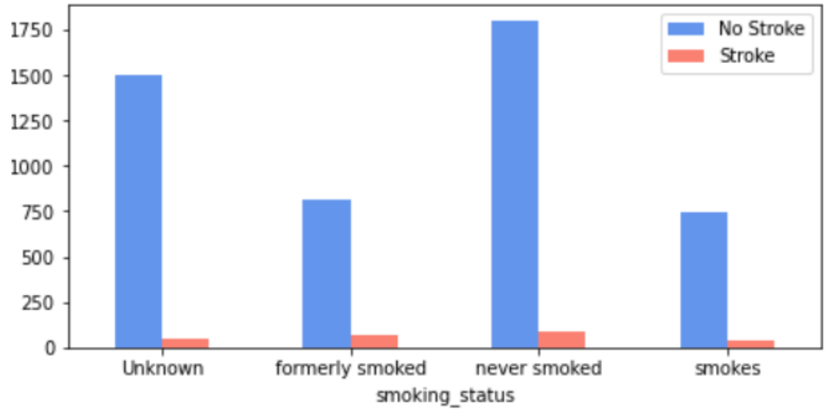




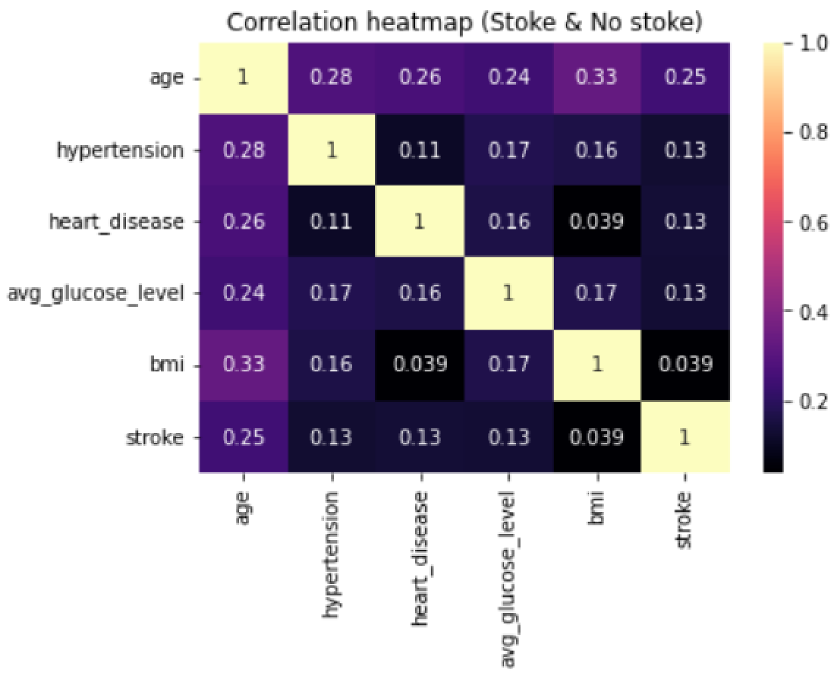


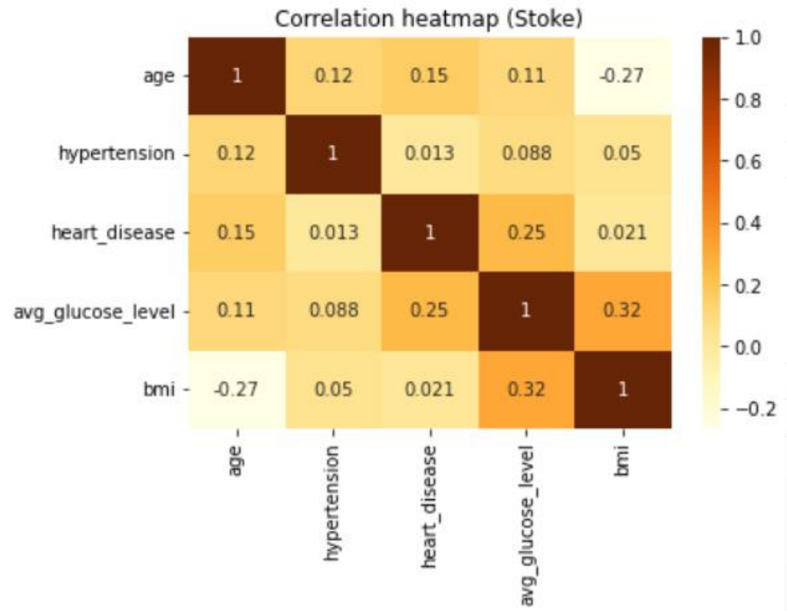
Univariate Analysis

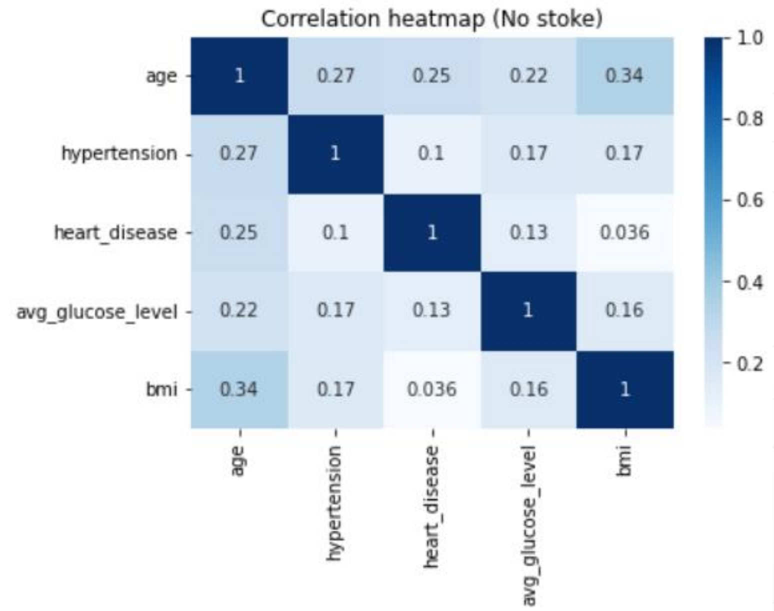




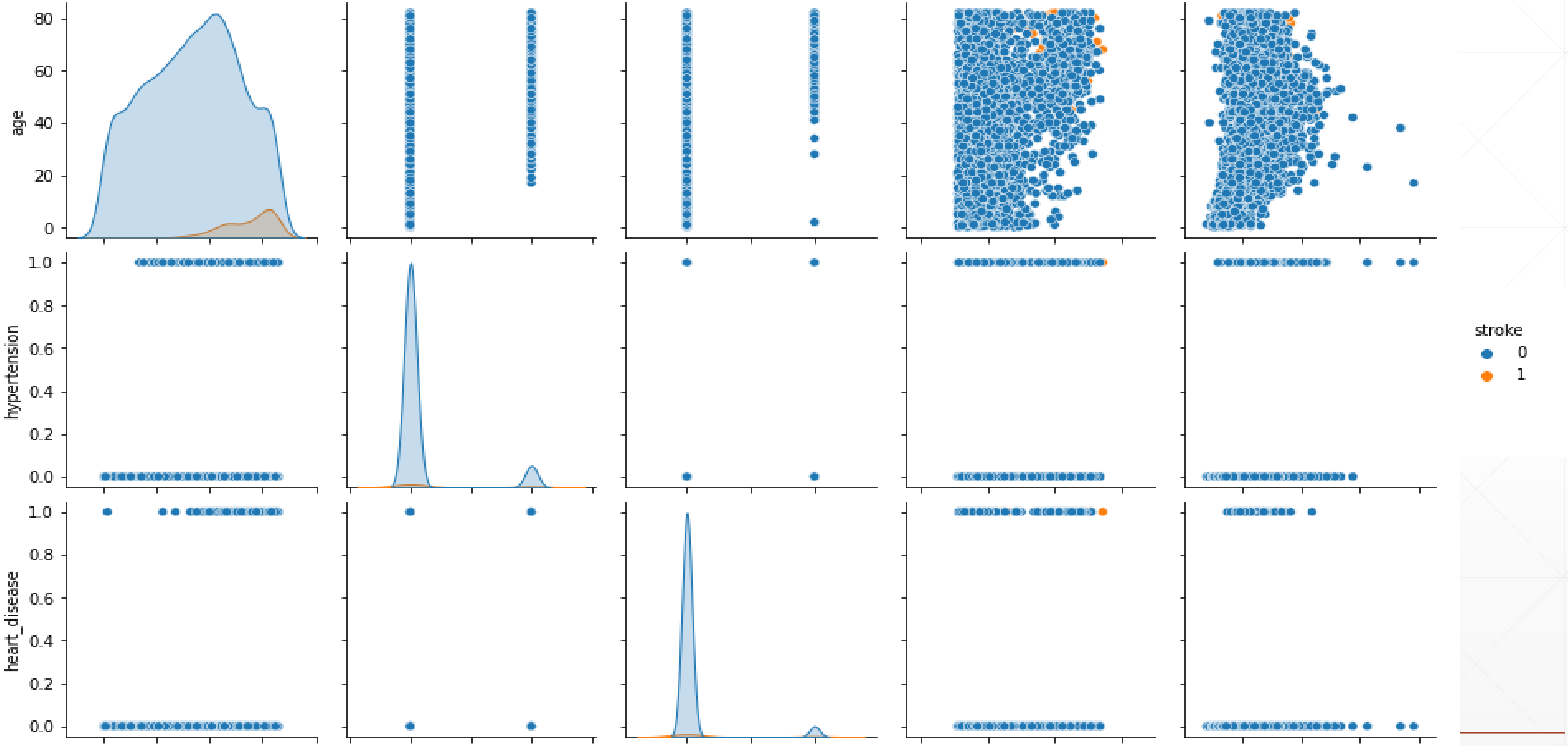
Correlation heatmap



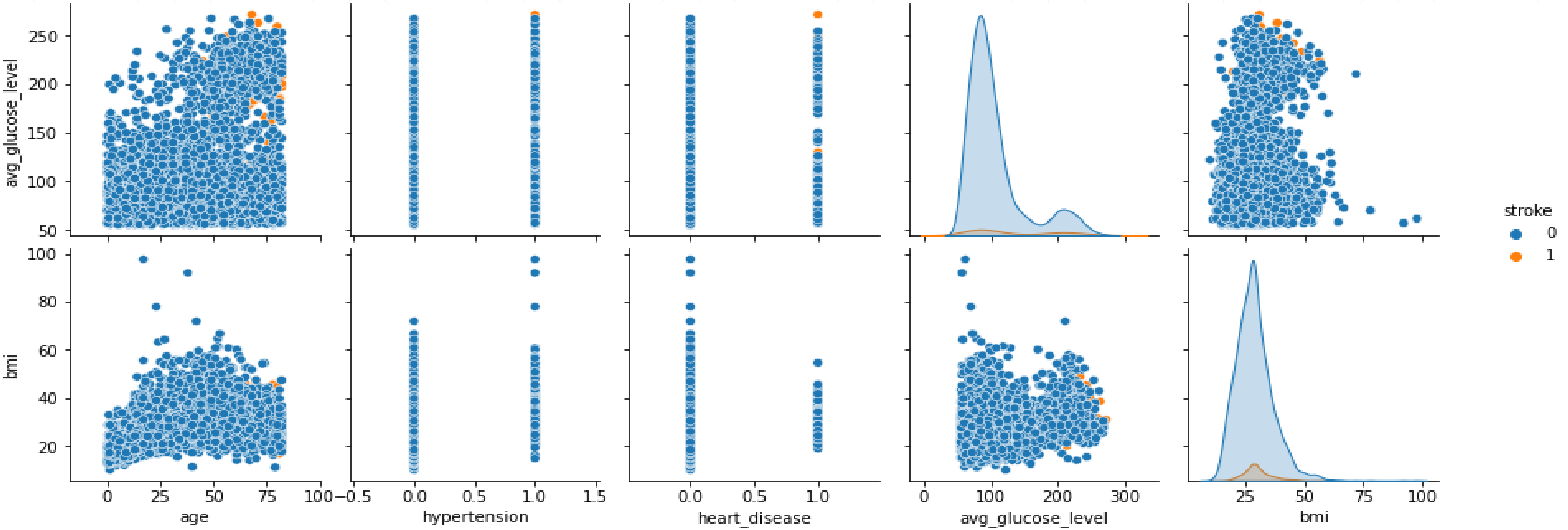




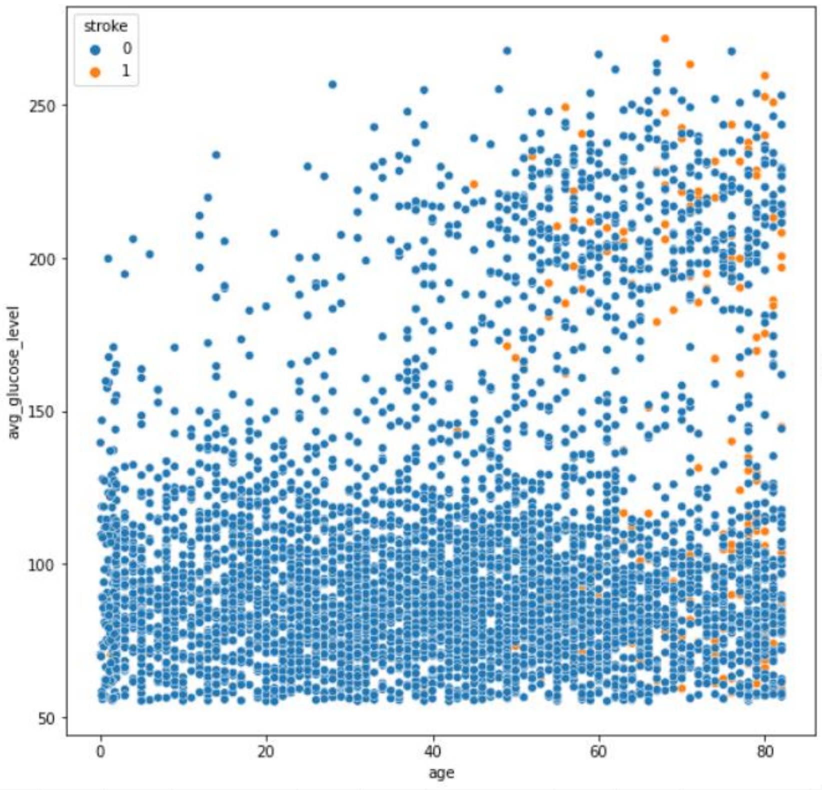
Bivariate Analysis

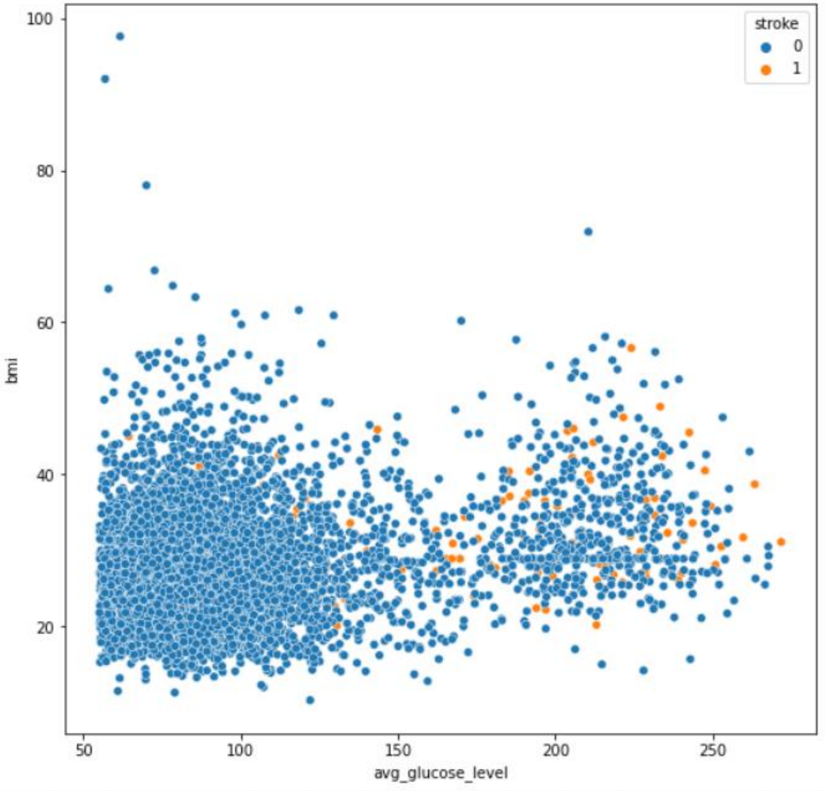


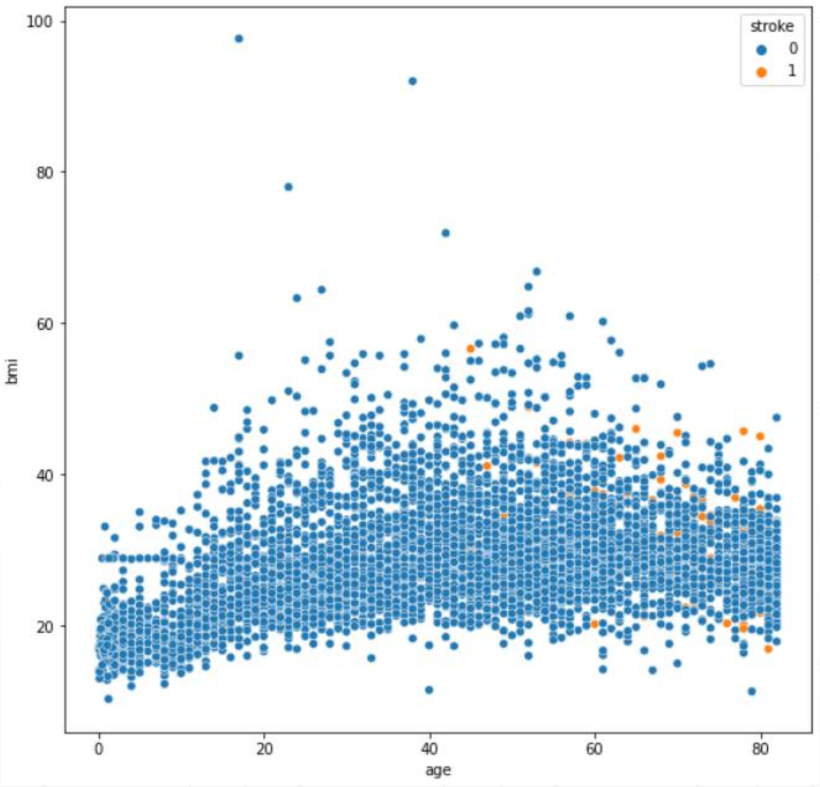
Bivariate Analysis



Bivariate Analysis

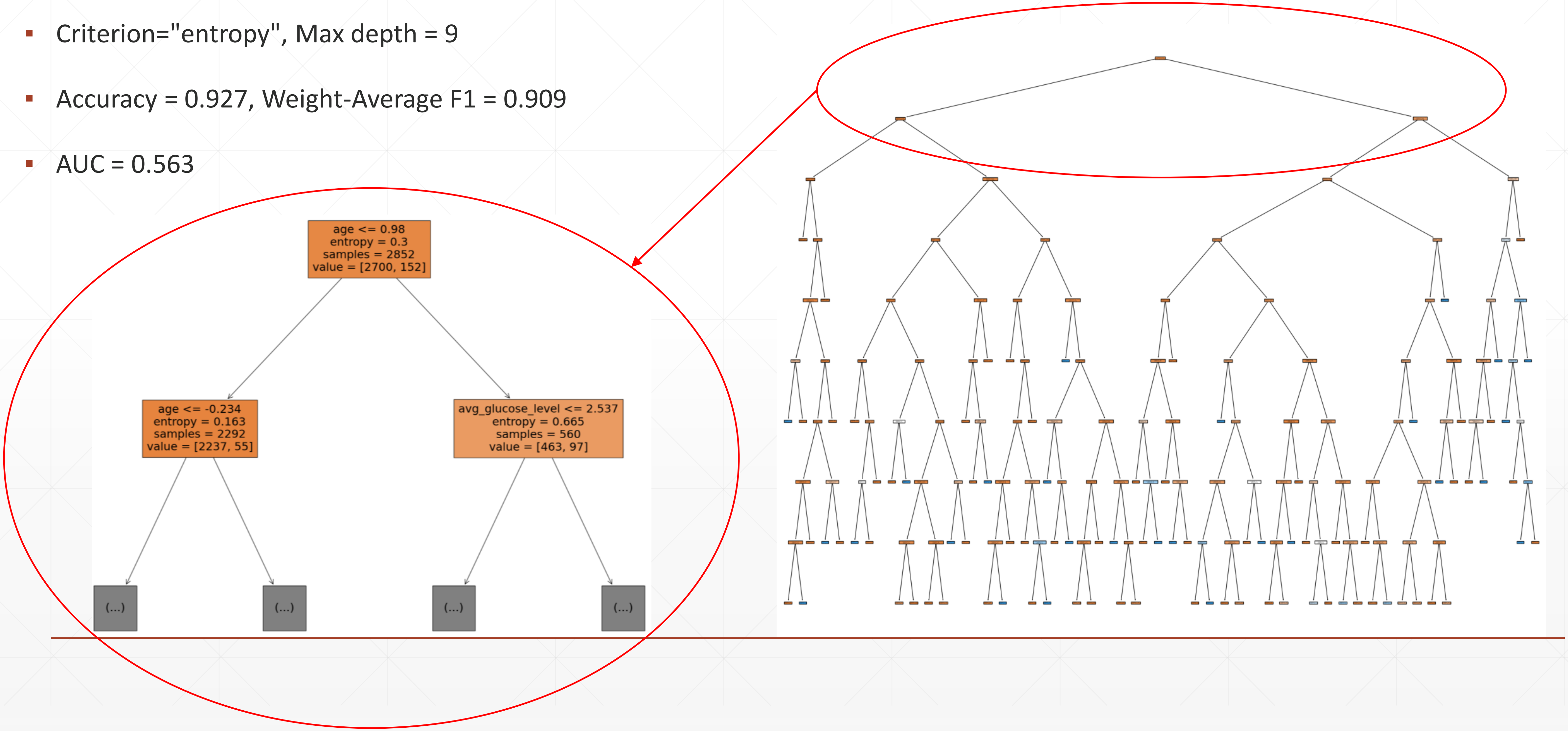






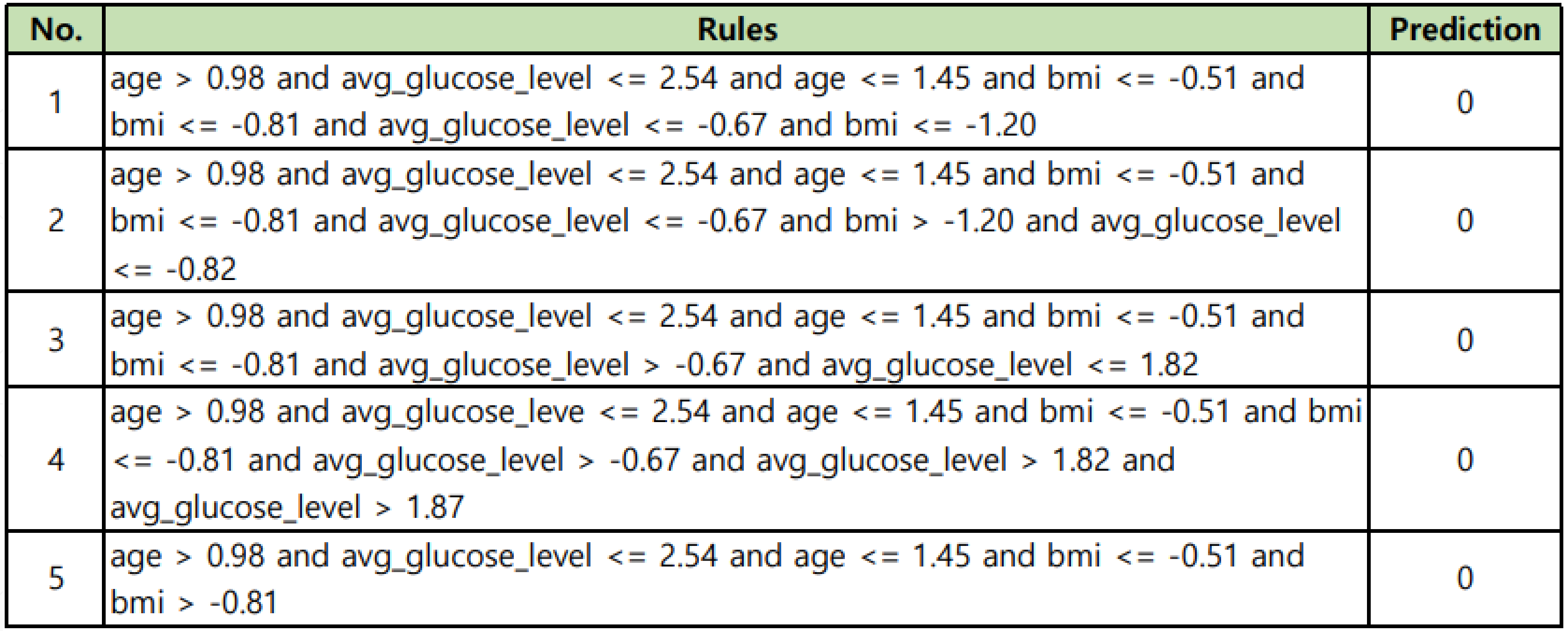
Decision Tree

- Training set : Testing set = 80 : 20, Random splitting = 42
- Criterion="entropy", Max depth = 9
- Accuracy = 0.927, Weight-Average F1 = 0.909
- AUC = 0.563



Decision Tree – Rule List (y = 0)

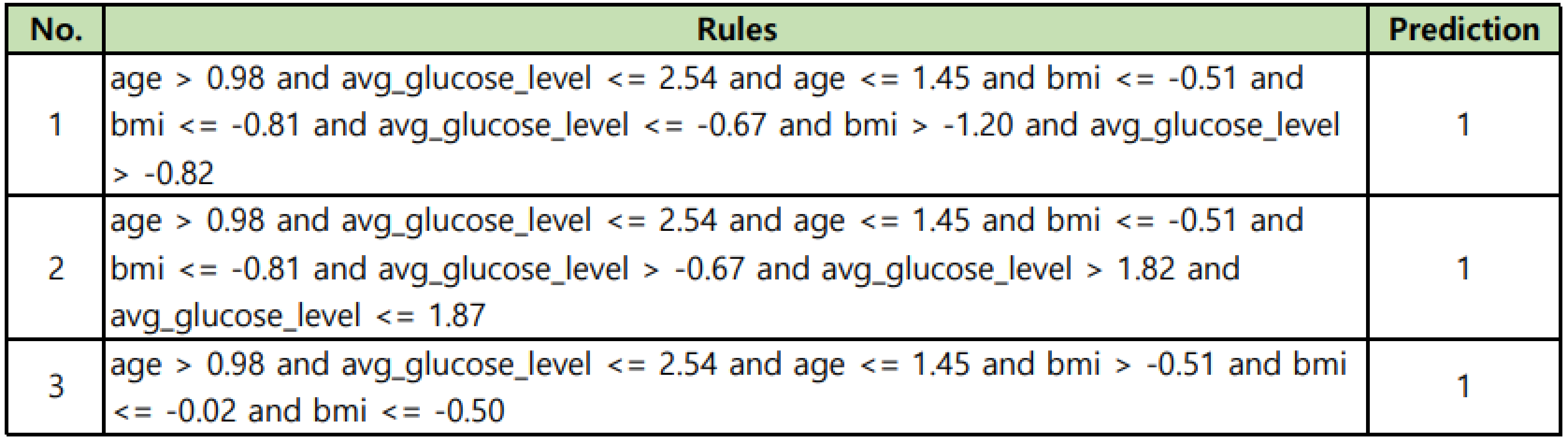
- Total rules = 60



No.	Rules	Prediction
1	age > 0.98 and avg_glucose_level <= 2.54 and age <= 1.45 and bmi <= -0.51 and bmi <= -0.81 and avg_glucose_level <= -0.67 and bmi <= -1.20	0
2	age > 0.98 and avg_glucose_level <= 2.54 and age <= 1.45 and bmi <= -0.51 and bmi <= -0.81 and avg_glucose_level <= -0.67 and bmi > -1.20 and avg_glucose_level <= -0.82	0
3	age > 0.98 and avg_glucose_level <= 2.54 and age <= 1.45 and bmi <= -0.51 and bmi <= -0.81 and avg_glucose_level > -0.67 and avg_glucose_level <= 1.82	0
4	age > 0.98 and avg_glucose_leve <= 2.54 and age <= 1.45 and bmi <= -0.51 and bmi <= -0.81 and avg_glucose_level > -0.67 and avg_glucose_level > 1.82 and avg_glucose_level > 1.87	0
5	age > 0.98 and avg_glucose_level <= 2.54 and age <= 1.45 and bmi <= -0.51 and bmi > -0.81	0

Decision Tree – Rule List (y = 1)

- Total rules = 30



No.	Rules	Prediction
1	age > 0.98 and avg_glucose_level <= 2.54 and age <= 1.45 and bmi <= -0.51 and bmi <= -0.81 and avg_glucose_level <= -0.67 and bmi > -1.20 and avg_glucose_level > -0.82	1
2	age > 0.98 and avg_glucose_level <= 2.54 and age <= 1.45 and bmi <= -0.51 and bmi <= -0.81 and avg_glucose_level > -0.67 and avg_glucose_level > 1.82 and avg_glucose_level <= 1.87	1
3	age > 0.98 and avg_glucose_level <= 2.54 and age <= 1.45 and bmi > -0.51 and bmi <= -0.02 and bmi <= -0.50	1