Stroke

Part 2 : Classification, Color DT

Review

- **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

Review: 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

Review: 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

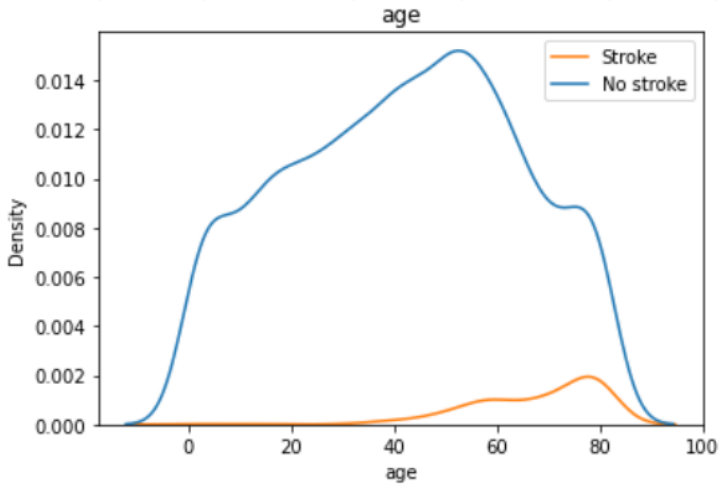
Review: Summary by stroke class

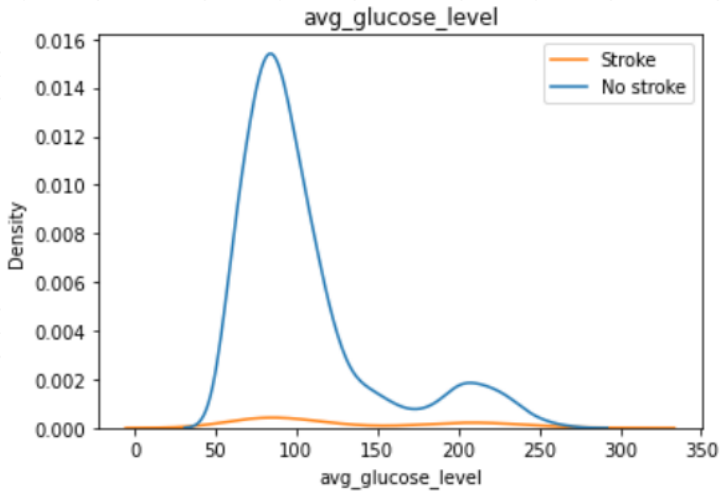
Column	Stroke class	children	Govt_job	Never_worked	Private	Self-employed	
work_type	0	685	624	22	2776	754	4861
	1	2	33	0	149	65	249

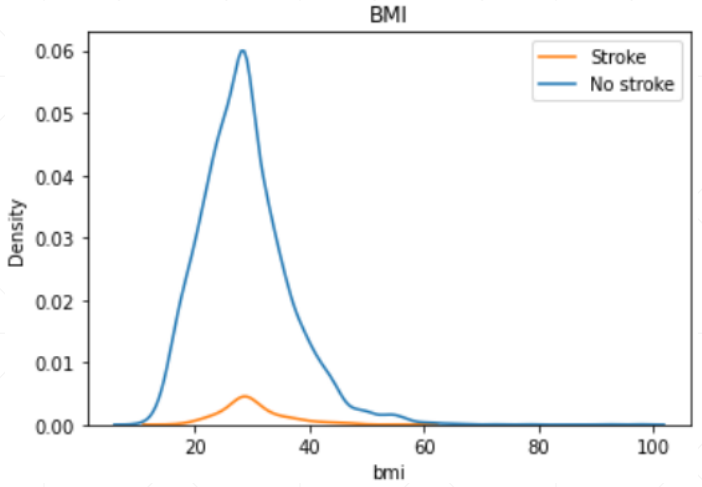
Column	Stroke class	Urban	Rural	
Residence_type	0	2461	2400	4861
	1	135	114	249

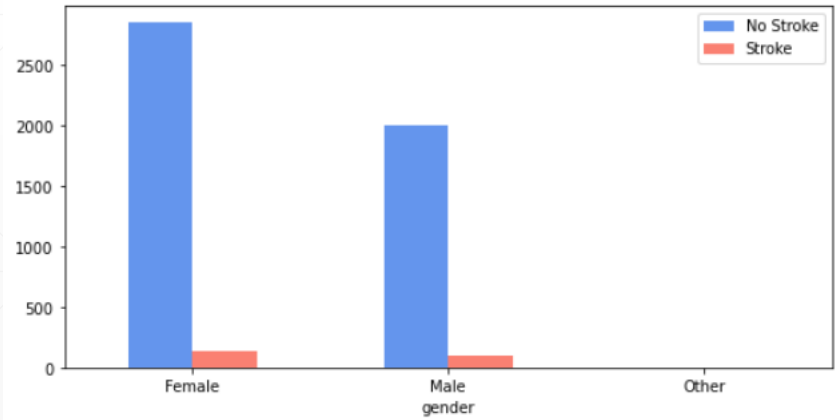
Column	Stroke class	formerly sr	never smok	smokes	Unknown	
smoking_status	0	815	1802	747	1497	4861
	1	70	90	42	47	249

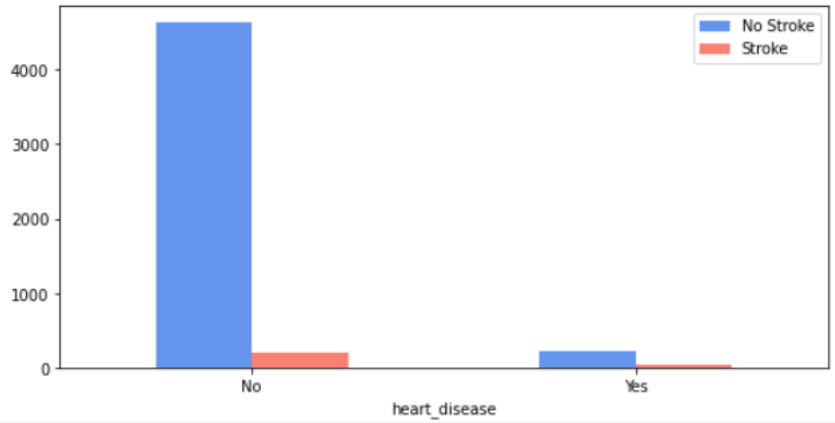
Review: Univariate Analysis



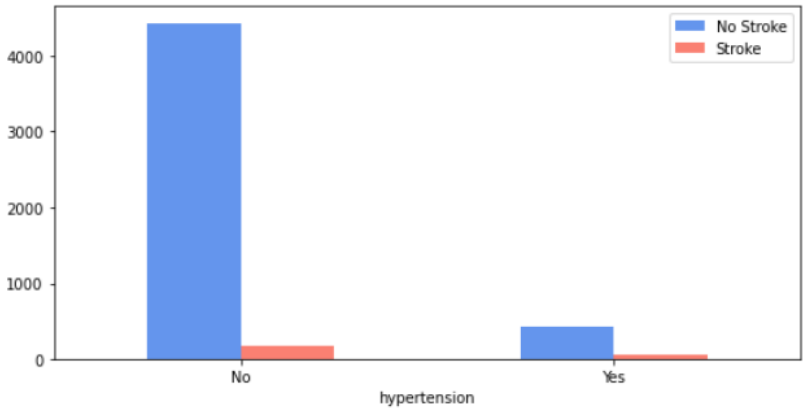


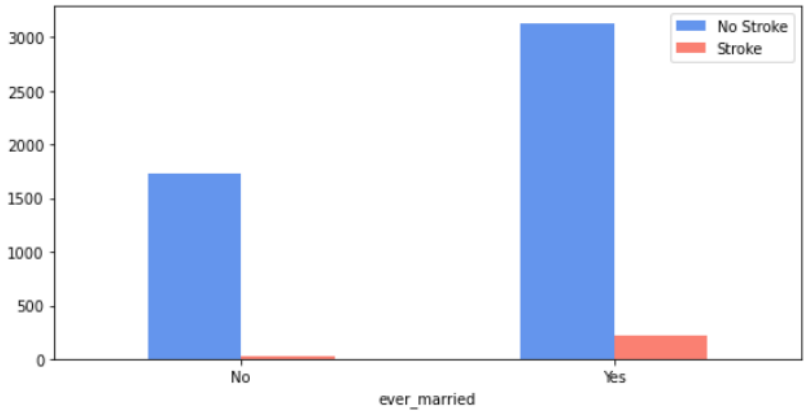


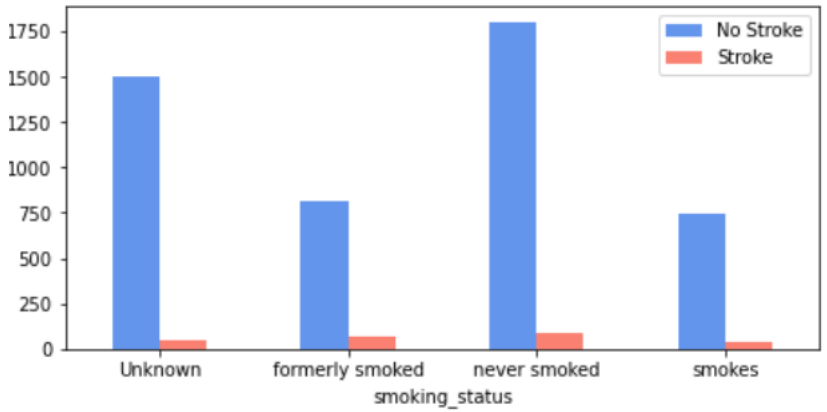


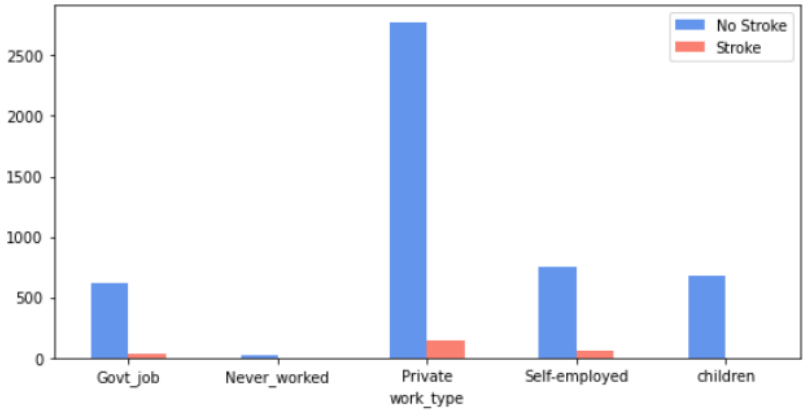


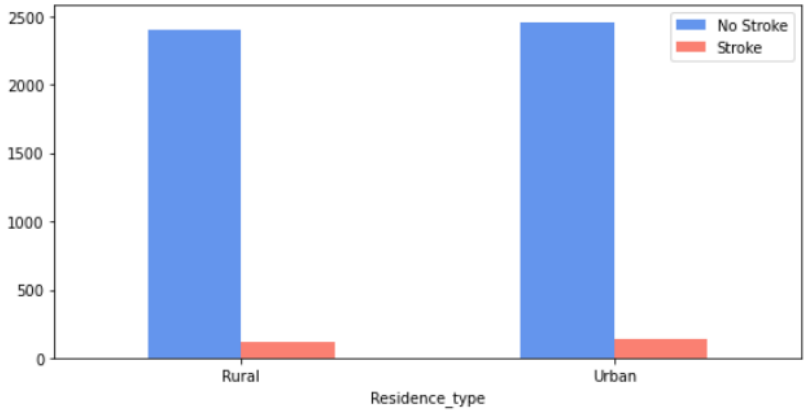
Review: Univariate Analysis



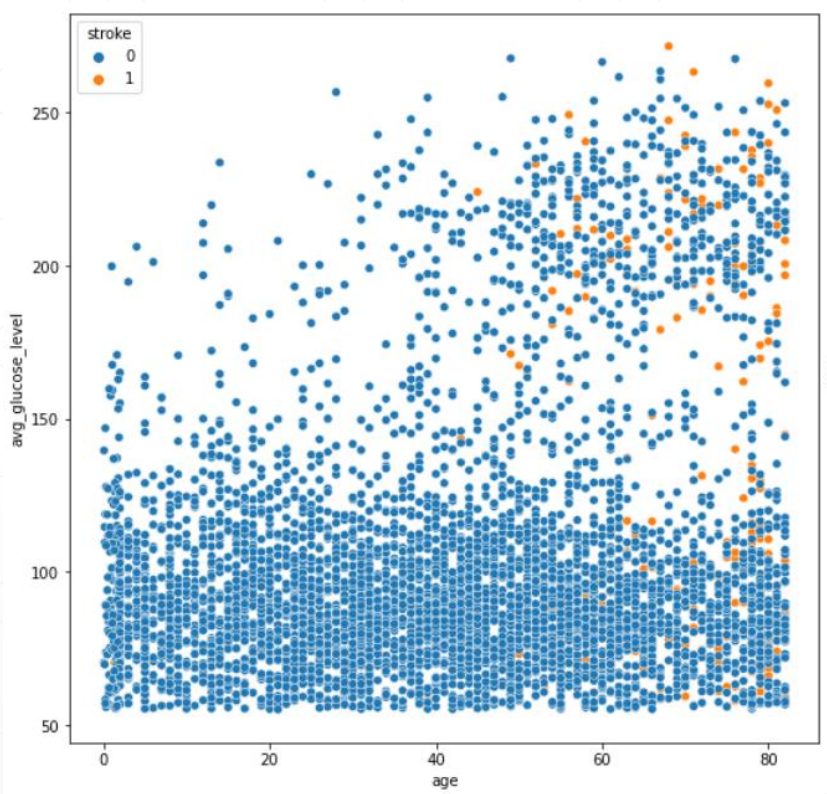


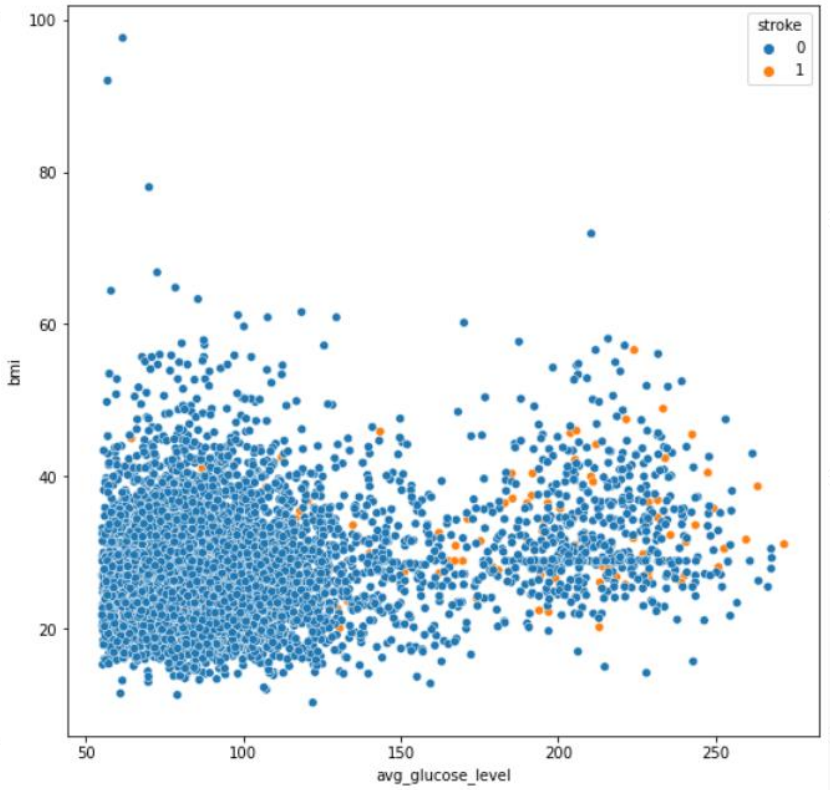


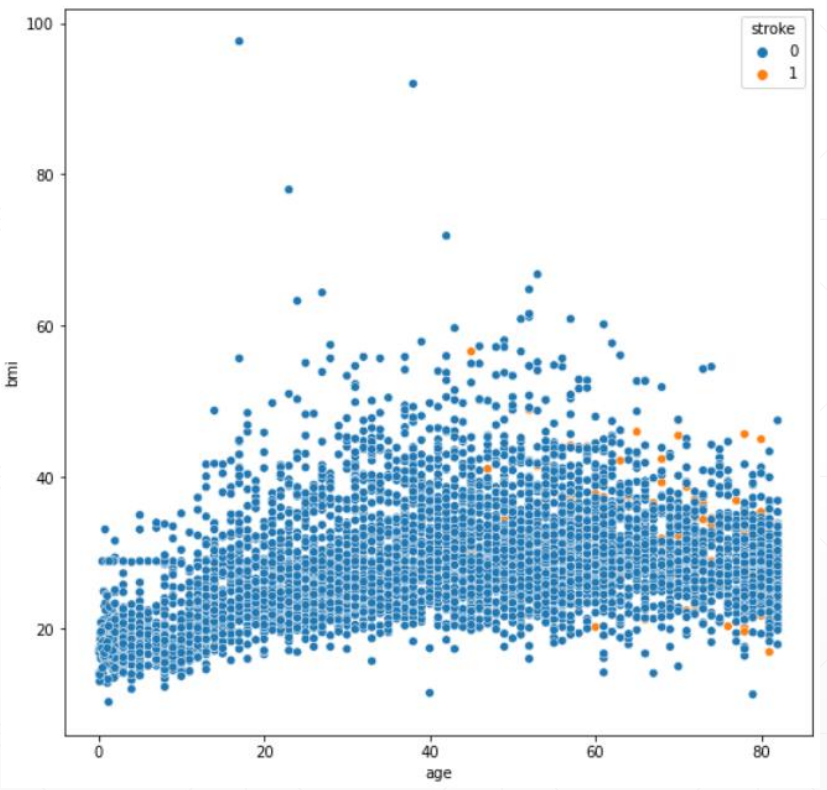




Review: Bivariate Analysis

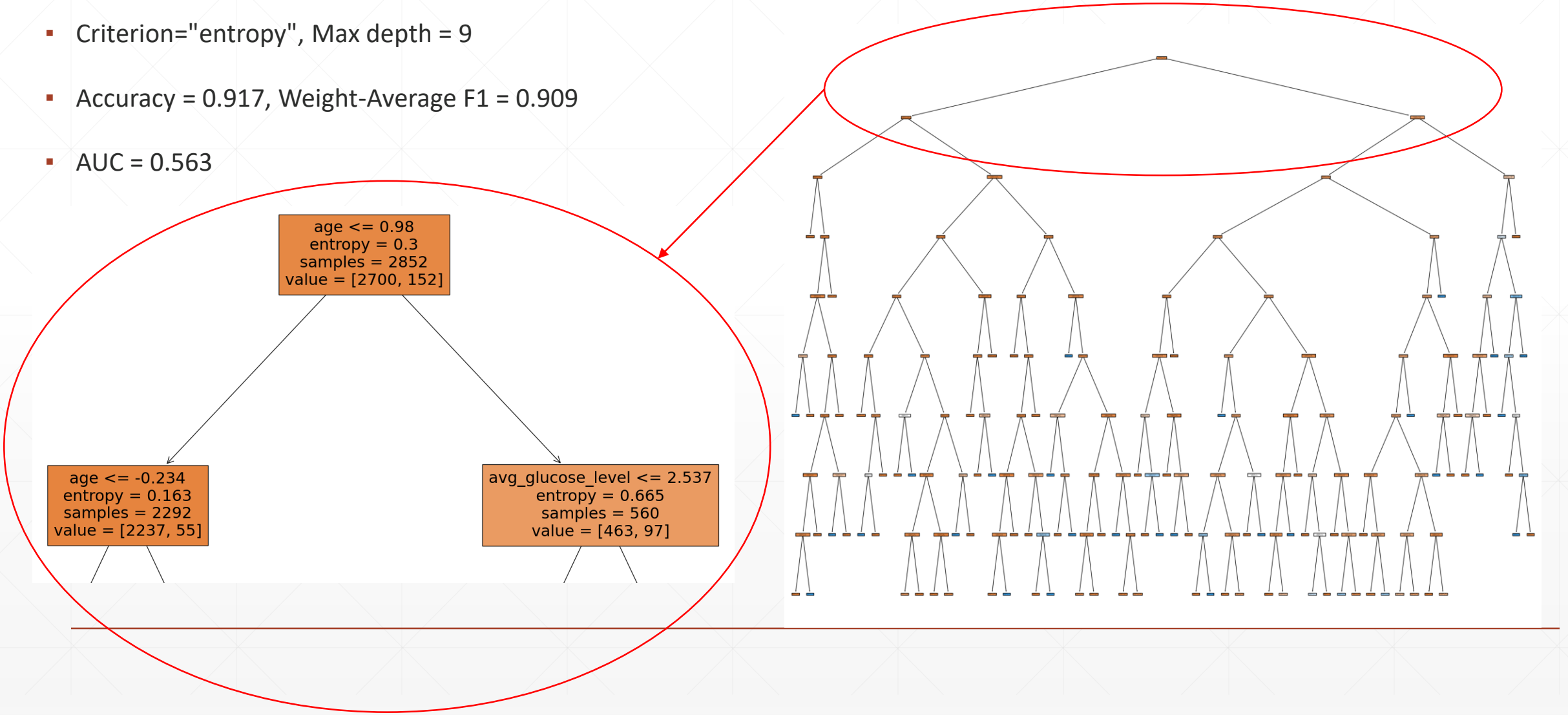






Review: Decision Tree

- Training set : Testing set = 80 : 20, Random splitting = 42
- Criterion="entropy", Max depth = 9
- Accuracy = 0.917, Weight-Average F1 = 0.909
- AUC = 0.563



Review: 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

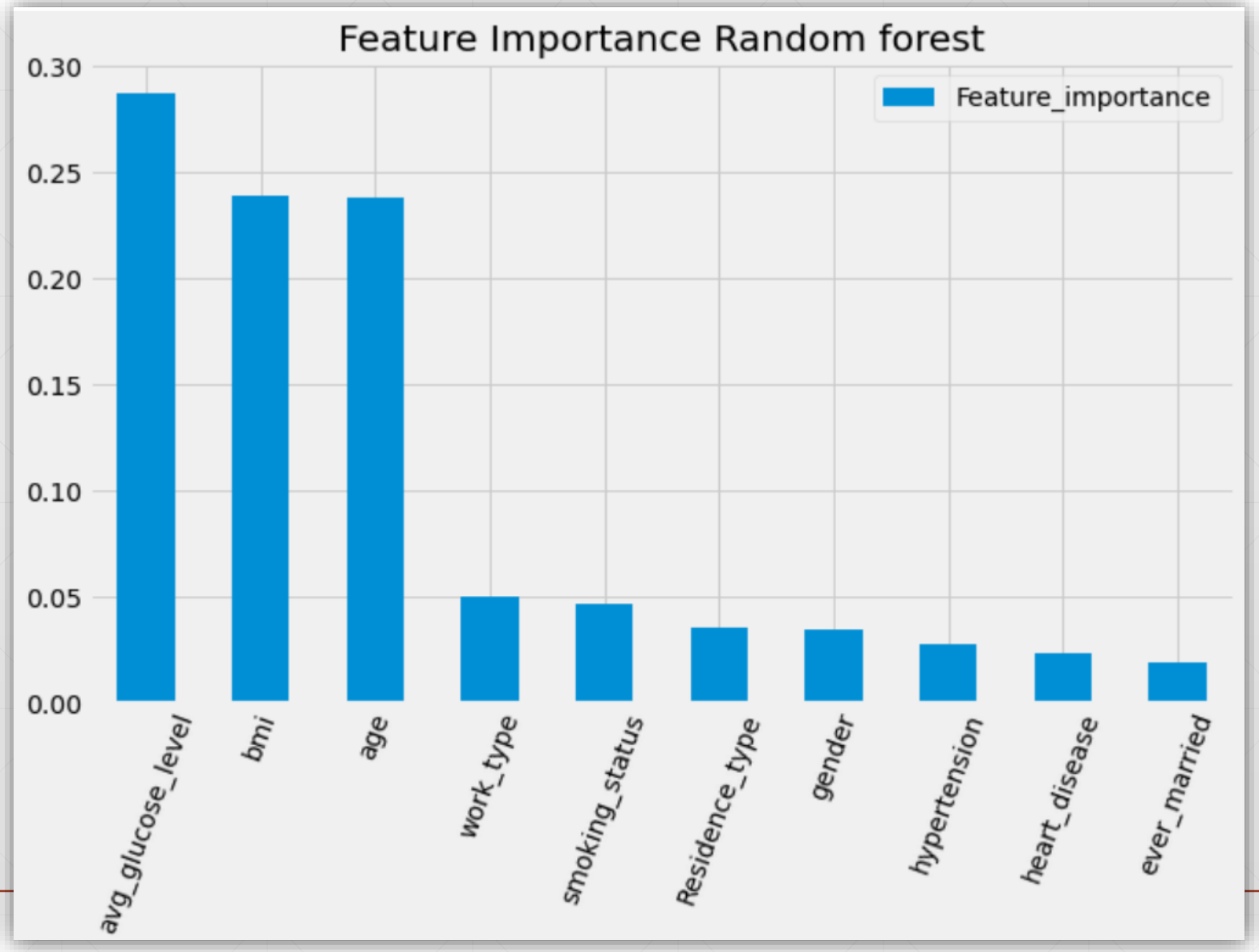
Review: Decision Tree – Rule List (y = 1)

- 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

Random Forest

- Training set : Testing set = 80 : 20
- No. of Tree = 100



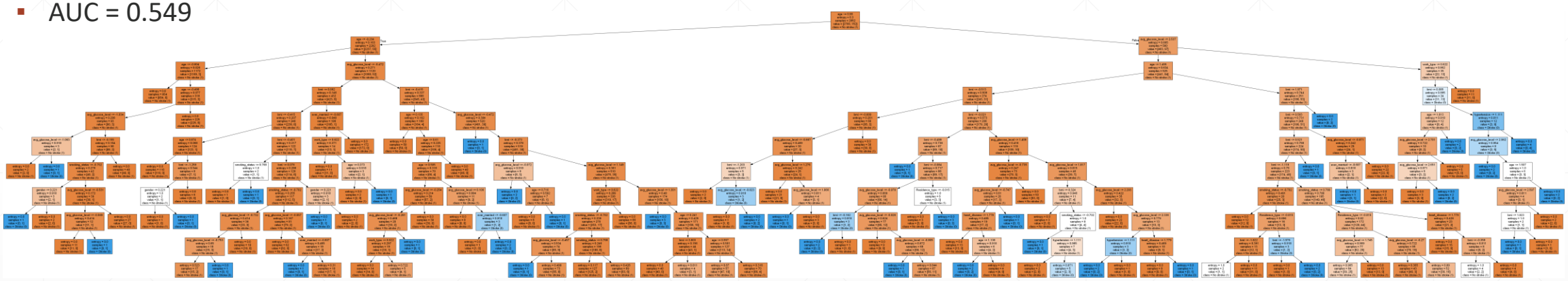
Performance

- Training set : Testing set = 80 : 20

	Decision Tree	Logistic Regression	Random Forest	Gradient Boosting	Naive Bayes	K-NN
ACC	0.917	0.929	0.928	0.929	0.872	0.928
Weight F1	0.901	0.896	0.895	0.898	0.884	0.895
AUROC	0.548	0.5	0.499	0.509	0.635	0.499

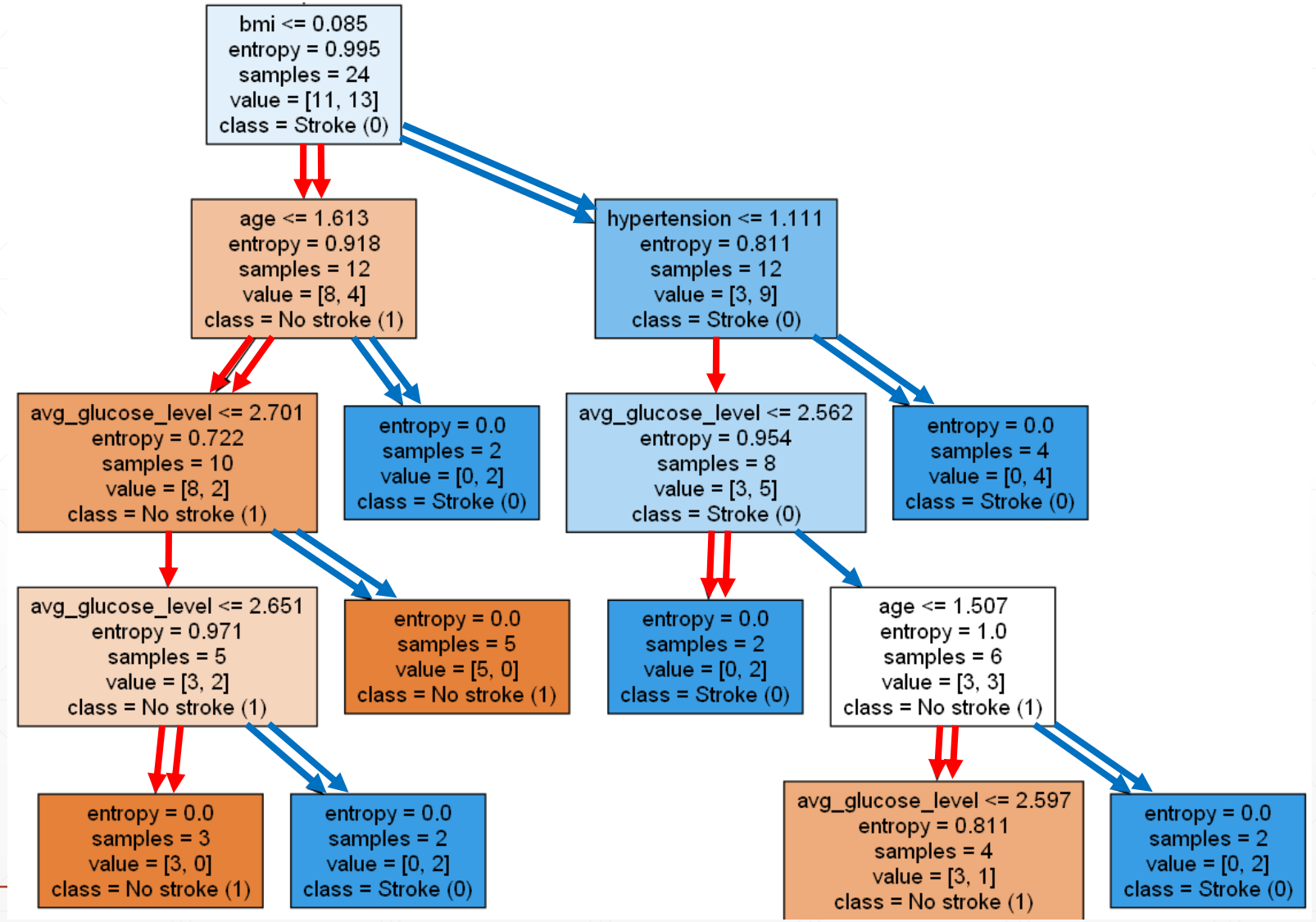
Color DT

- Training set : Testing set = 80 : 20
- Max depth = 10
- Accuracy = 0.922, , Weight-Average F1 = 0.902
- AUC = 0.549



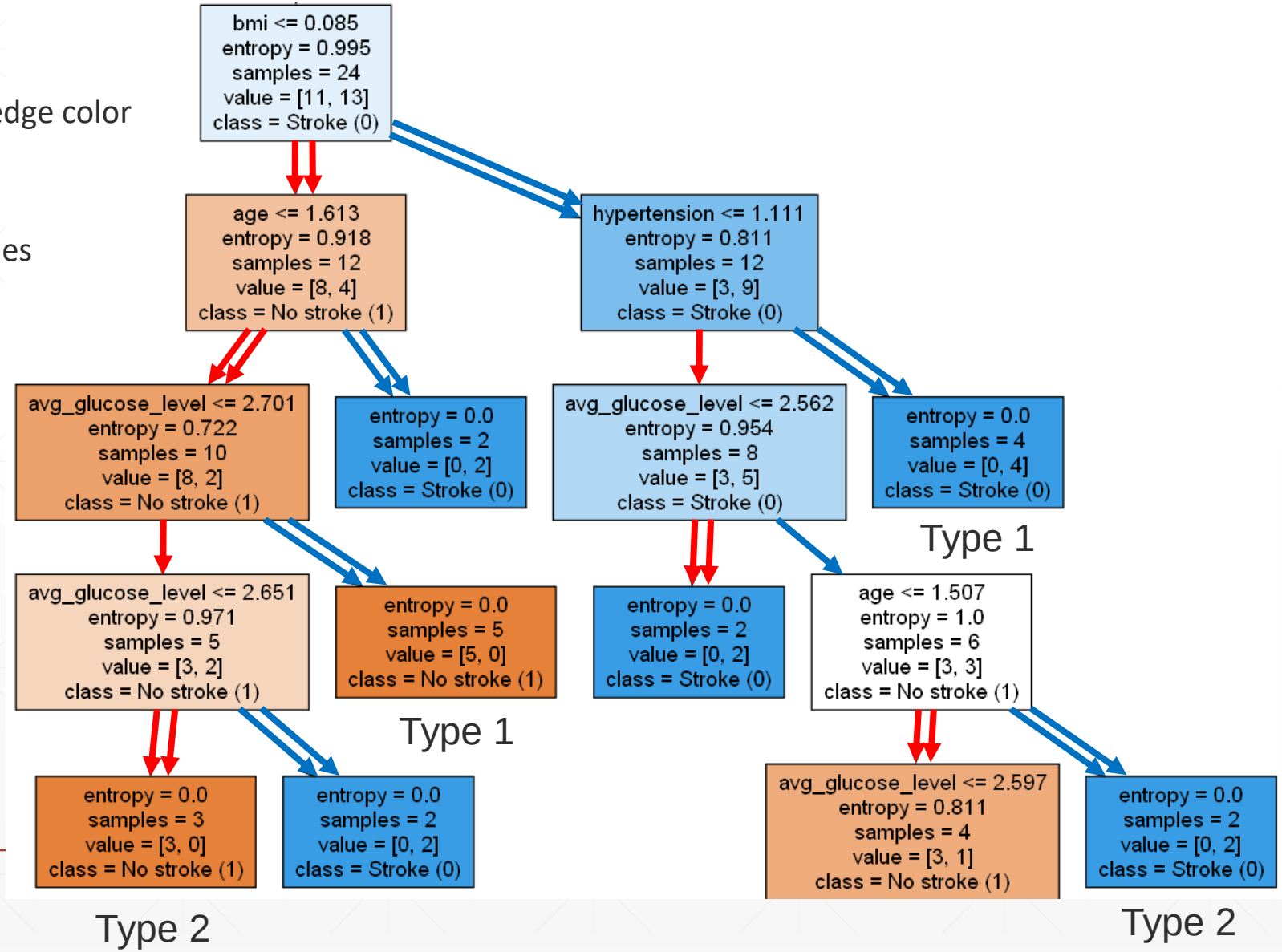
Color DT

- = Nodes with many class 1 (No stroke)
- = Nodes with many class 0 (Stroke)
- = Node with improved ratio of 1
- = Node with improved ratio of 0
- = Nodes with improved ratio of 1 and Gini or Entropy
- = Nodes with improved ratio of 0 and Gini or Entropy



Color DT: Irrelevant condition

- Type of leaf node
- Type 0: Node where leaf node color and last edge color do not match
- Type 1: Non-Type 0, none of the ancestor nodes have high purity
- Type 2: Non-Type 0, some of the ancestor nodes have high purity



Color DT: Underlying rules

Rule (Type 2)	Predicted class
bmi <= 0.085 and age <= 1.613 and avg_glucose_level <= 2.701 and avg_glucose_level <= 2.651	1 (No stroke)

Underlying Rule	Predicted class
bmi <= 0.085 and age <= 1.613 and ~~avg_glucose_level <= 2.701~~ and avg_glucose_level <= 2.651	1 (No stroke)

Rule (Type 2)	Predicted class
bmi > 0.085 and hypertension <= 1.111 and avg_glucose_level <= 2.562 and age > 1.507	0 (Stroke)

Underlying Rule	Predicted class
bmi > 0.085 and ~~hypertension <= 1.111~~ and avg_glucose_level <= 2.562 and age > 1.507	0 (Stroke)

Conclusion

- For EDA part, the most **feature importance** is **'Age'** = 0.25
- For random forest model, the most **feature importance** is **'avg_glucose_level'**
- The most **accuracy** model are **Logistic Regression and Gradient Boosting** = 0.929
- The most **AUROC** model is **Naïve Bayes** = 0.635
- The **best model** for the stroke prediction is **Naïve Bayes** (Because the dataset is imbalanced)