Heart disease prediction /심장질환 예측/

24510112 자르갈사이칸 나몽

Preview

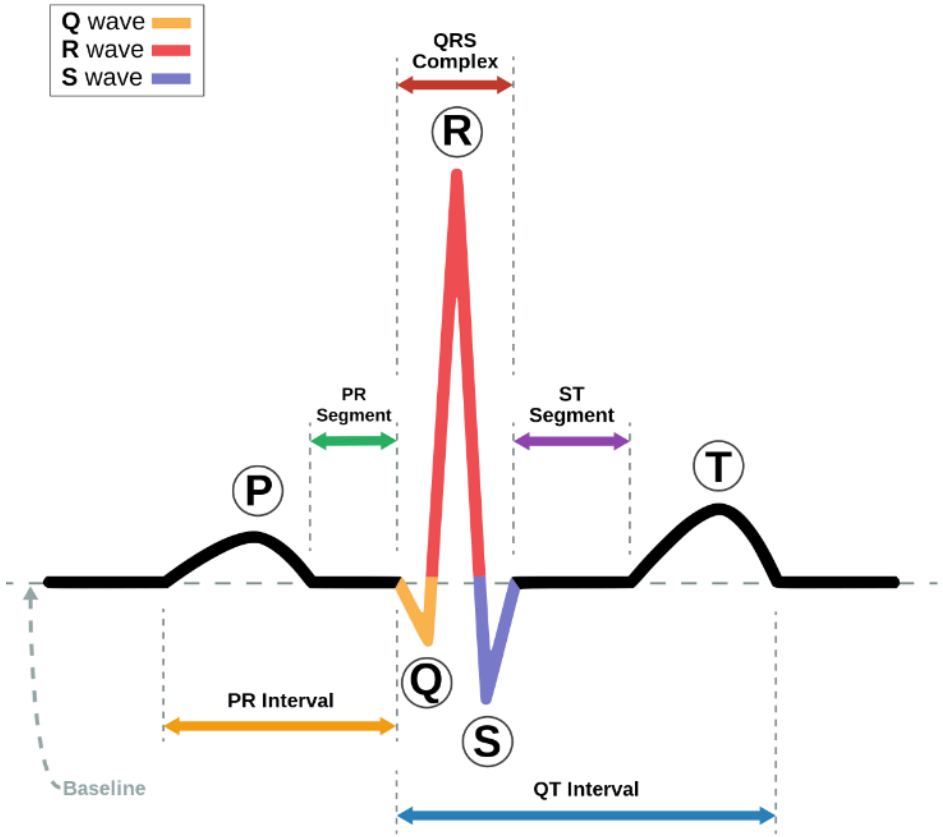
- 환자의 나이, 콜레스테롤 수치 등의 정보를 바탕으로 심장 질환 유무를 예측
- 11개의 설명변수와 1개의 목표변수, 총 918개의 환자 정보로 구성

No	Age	Sex	ChestPain Type	Resting BP	Cholesterol	FastingBS	Resting ECG	MaxHR	Exercise Angina	Oldpeak	ST_Slope	Heart Disease
1	40	M	ATA	140	289	0	Normal	172	N	0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0	Up	0
912	59	M	ASY	164	176	1	LVH	90	N	1	Flat	1
913	57	F	ASY	140	241	0	Normal	123	Y	0.2	Flat	1
914	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
915	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
916	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
917	57	F	ATA	130	236	0	LVH	174	N	0	Flat	1
918	38	M	NAP	138	175	0	Normal	173	N	0	Up	0

데이터 특성 정보

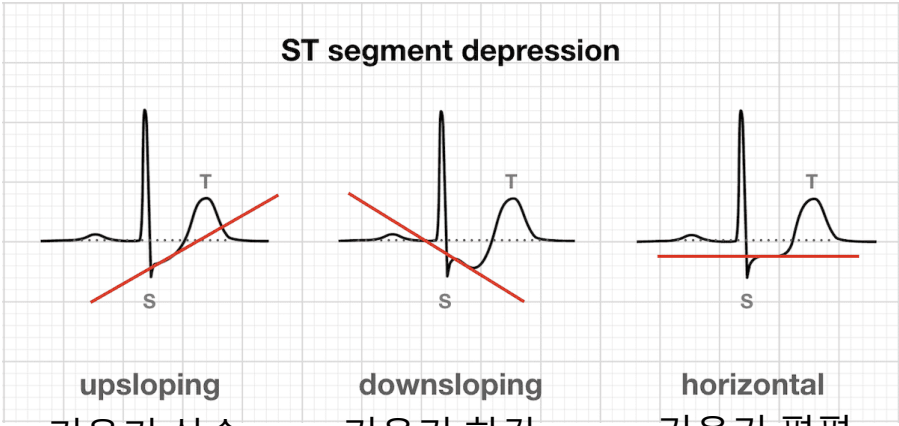
№	특성	설명
1	Age	나이
2	Sex	성별
3	ChestPainType	**가슴통증 타입** - TA (Typical Angina): 전형적 가슴통증 - ATA (Atypical Angina): 비전형적 가슴통증 - NAP (Non-Anginal Pain): 비심인성 가슴통증 - ASY (Asymptomatic): 무증상/가슴통증없음
4	RestingBP	휴식 혈압
5	Cholesterol	혈청 콜레스테롤
6	FastingBS	**공복 상태의 혈당** - 1: 120mg/dl보다 큰 경우, 비정상수치 - 0: 120mg/dl보다 작은 경우, 정상수치
7	RestingECG	**안정된 상태의 심전도** /심장을 박동하게 하는 전기 신호의 간격과 강도를 기록하는 검사/ - Normal: 정상 - ST: ST-T파 이상 상태 - LVH: 좌심실비대 상태
8	MaxHR	최대 심박수
9	ExerciseAngina	운동으로 발생된 협심증 - 협심증: 심장에 혈액을 공급하는 혈관인 관상 동맥이 동맥 경화증으로 좁아져서 생기는 질환
10	Oldpeak	노약 = **운동으로 발생하는 ST분절 저하**
11	ST_Slope	ST분절 기울기

특성 설명



ST_Slope /ST분절 기울기/:

- ST 분절이란 QRS파의 끝나는 점에서 T파의 시작점 사이의 간격을 나타내는 부분입니다.
- 정성적인 ST분절은 약간 위쪽으로 오목한 형태를 띈다. 평평하거나 기울기가 내리막인 ST분절은 심장질환을 의미할 수 있다고 합니다.
- 그런데 ST 분절 상승이 정상인을 뜻하는 것은 아니고, 상승 정도가 일정 수치 이상 상승하면 심근 경색을 뜻합니다.



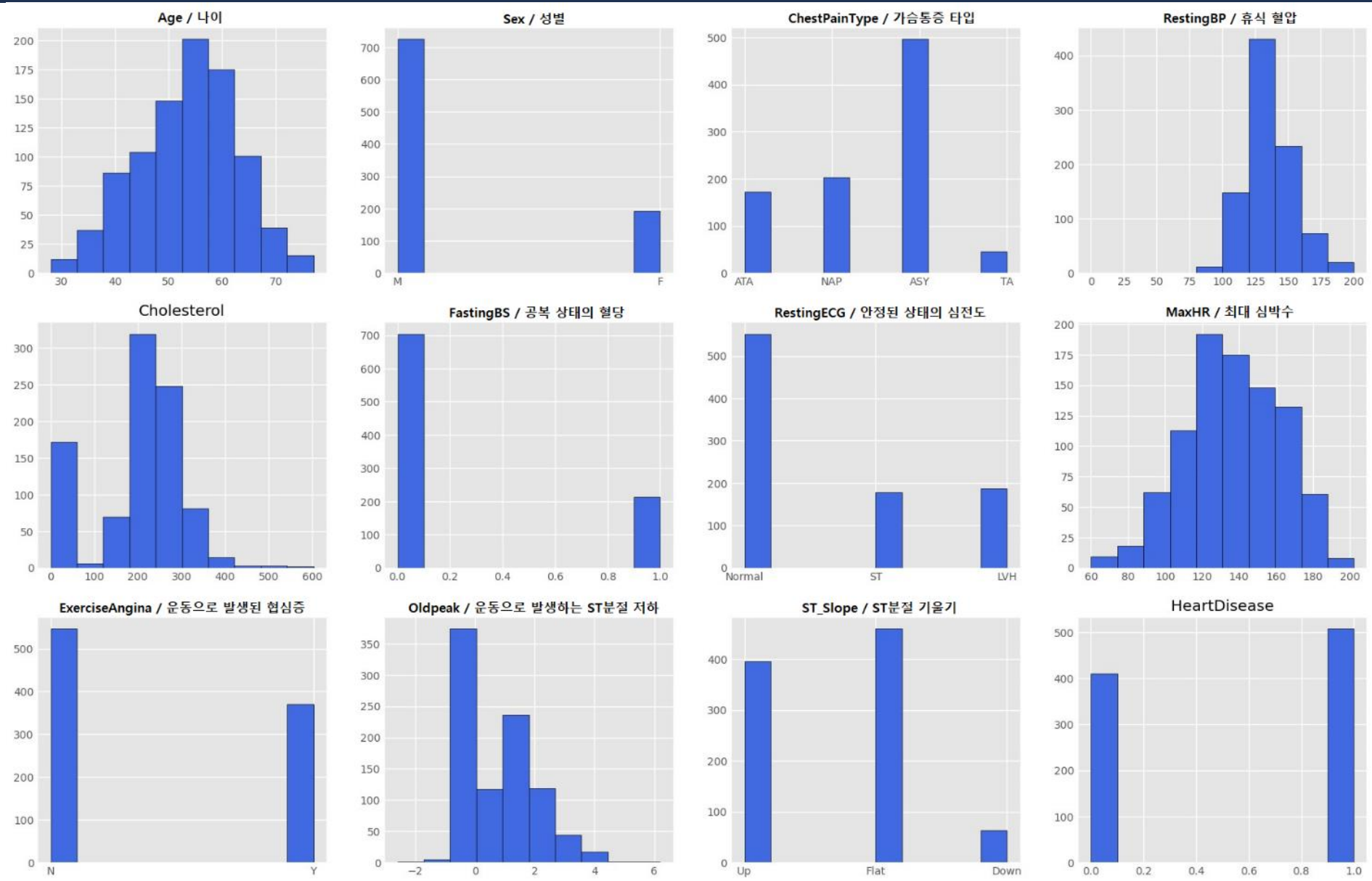
Data preprocessing

- Duplicated data 없음
- Missing value (결측치) 없음
- 5개의 범주형, 6개의 정수형, 1개의 실수형 특성으로 구성됨

№	Columns	Size	Null	Dtype	Unique values
1	Age	918	non-null	int64	50
2	Sex	918	non-null	object	2
3	ChestPainType	918	non-null	object	4
4	RestingBP	918	non-null	int64	67
5	Cholesterol	918	non-null	int64	222
6	FastingBS	918	non-null	int64	2
7	RestingECG	918	non-null	object	3
8	MaxHR	918	non-null	int64	119
9	ExerciseAngina	918	non-null	object	2
10	Oldpeak	918	non-null	float64	53
11	ST_Slope	918	non-null	object	3
12	HeartDisease	918	non-null	int64	2

EDA

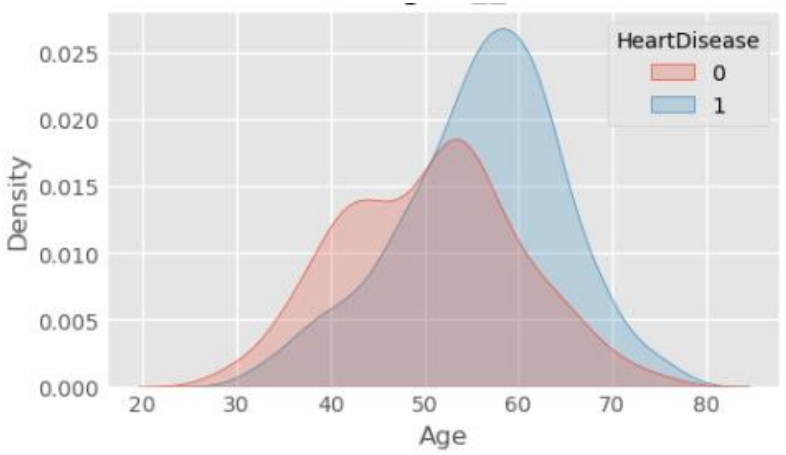
Histogram



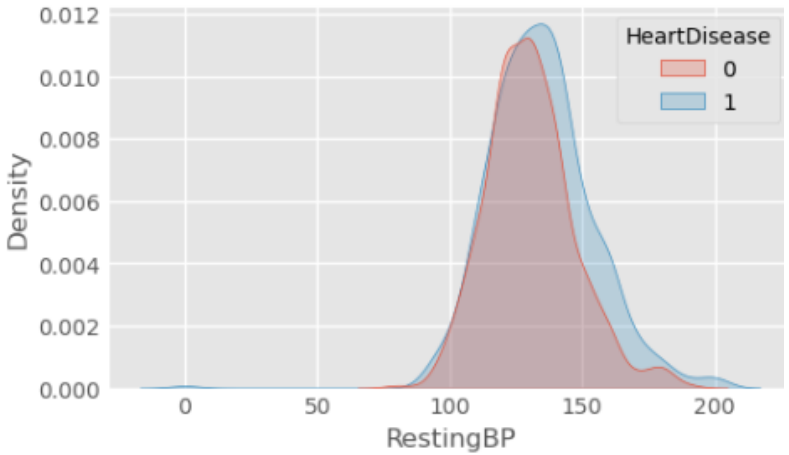
EDA

Relative frequency graph /수치형/

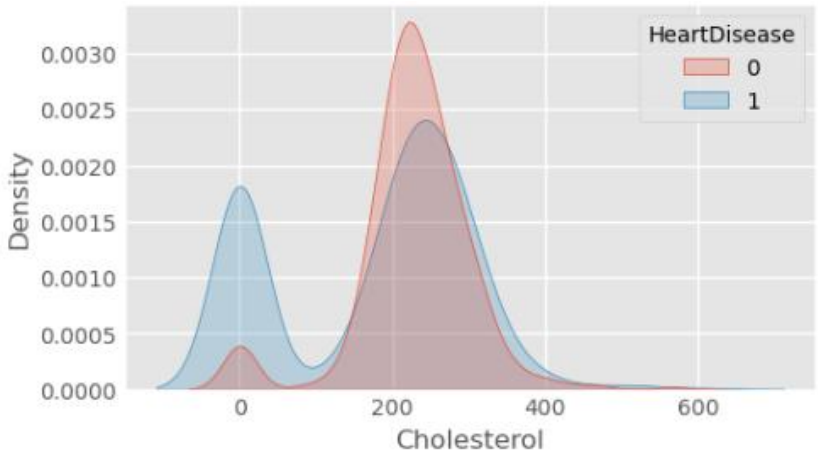
나이



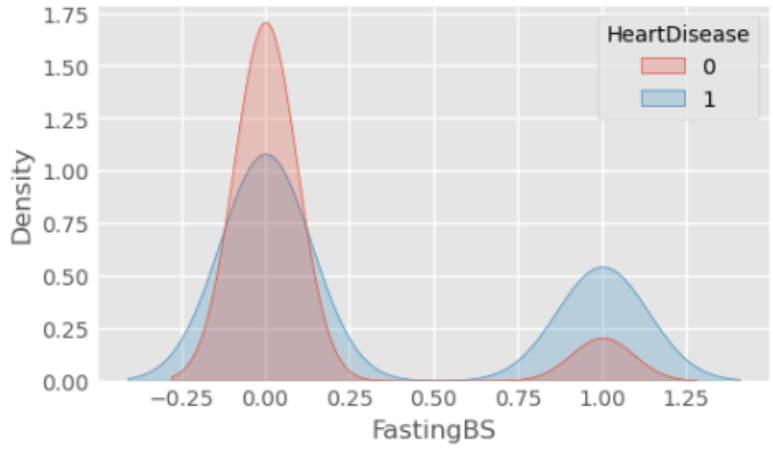
혈압



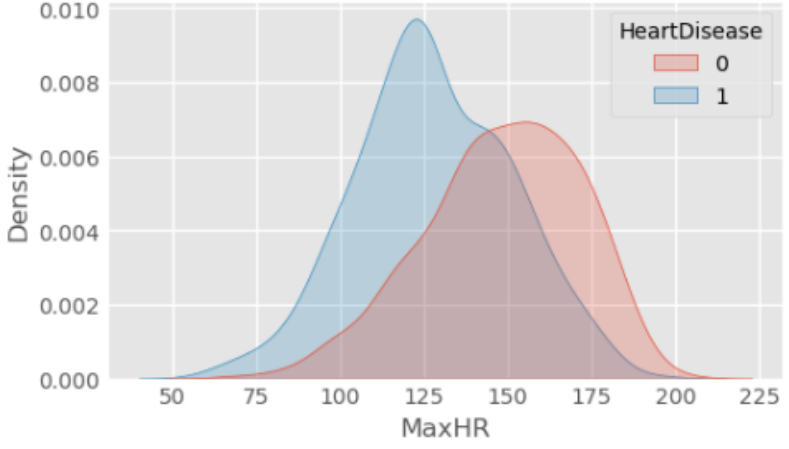
콜레스테롤



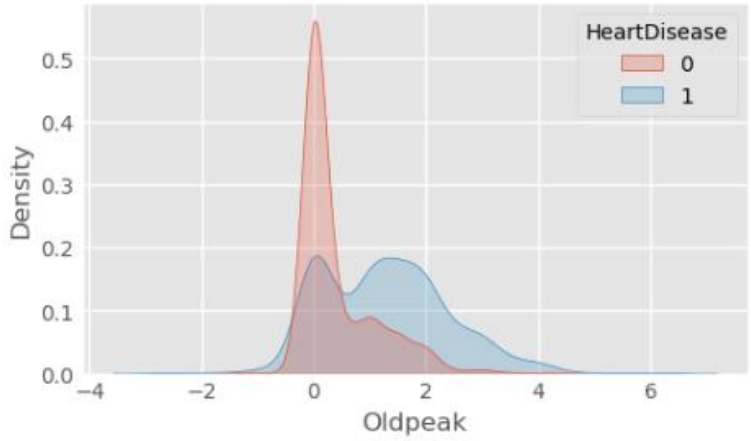
공복 상태의 혈당



최대 심박수



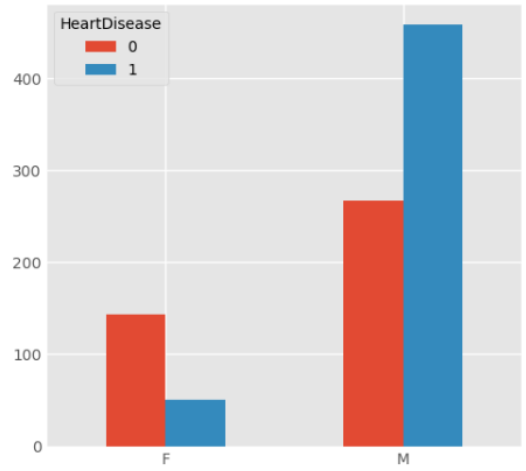
운동으로 발생하는 ST분절 저하



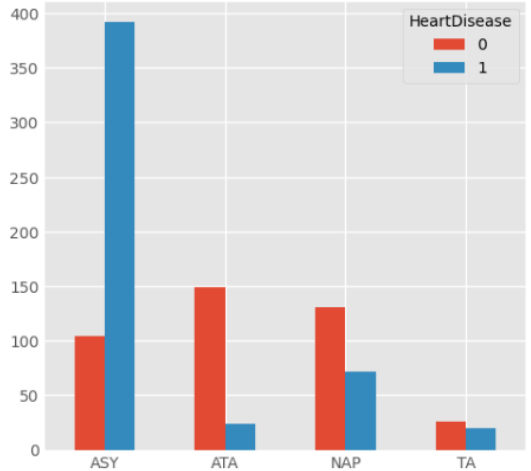
EDA

Relative frequency histogram /범주형/

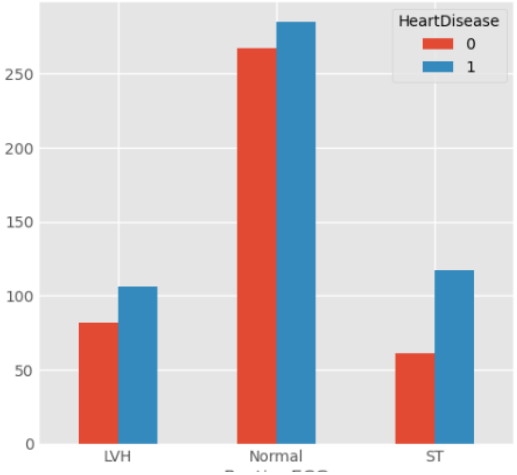
성별



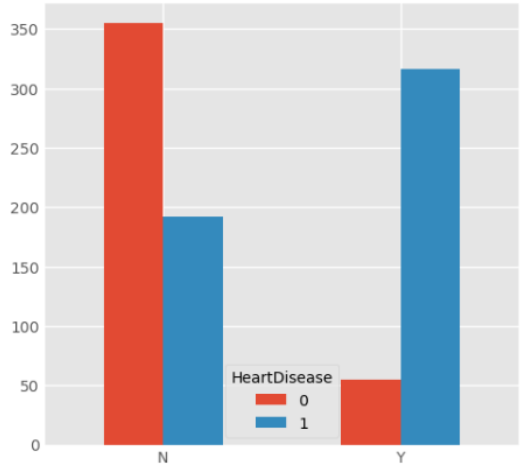
가슴통증 타입



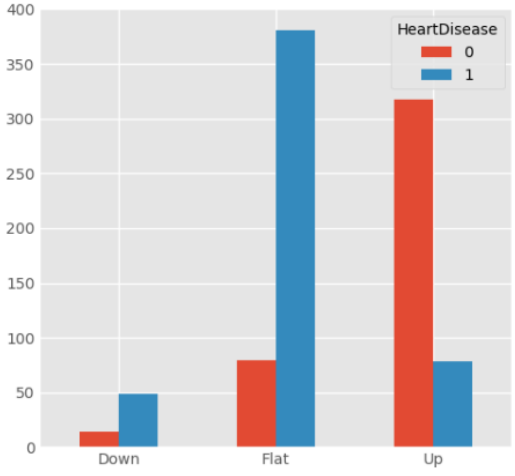
안정된 상태의 심전도



운동으로 발생된 협심증

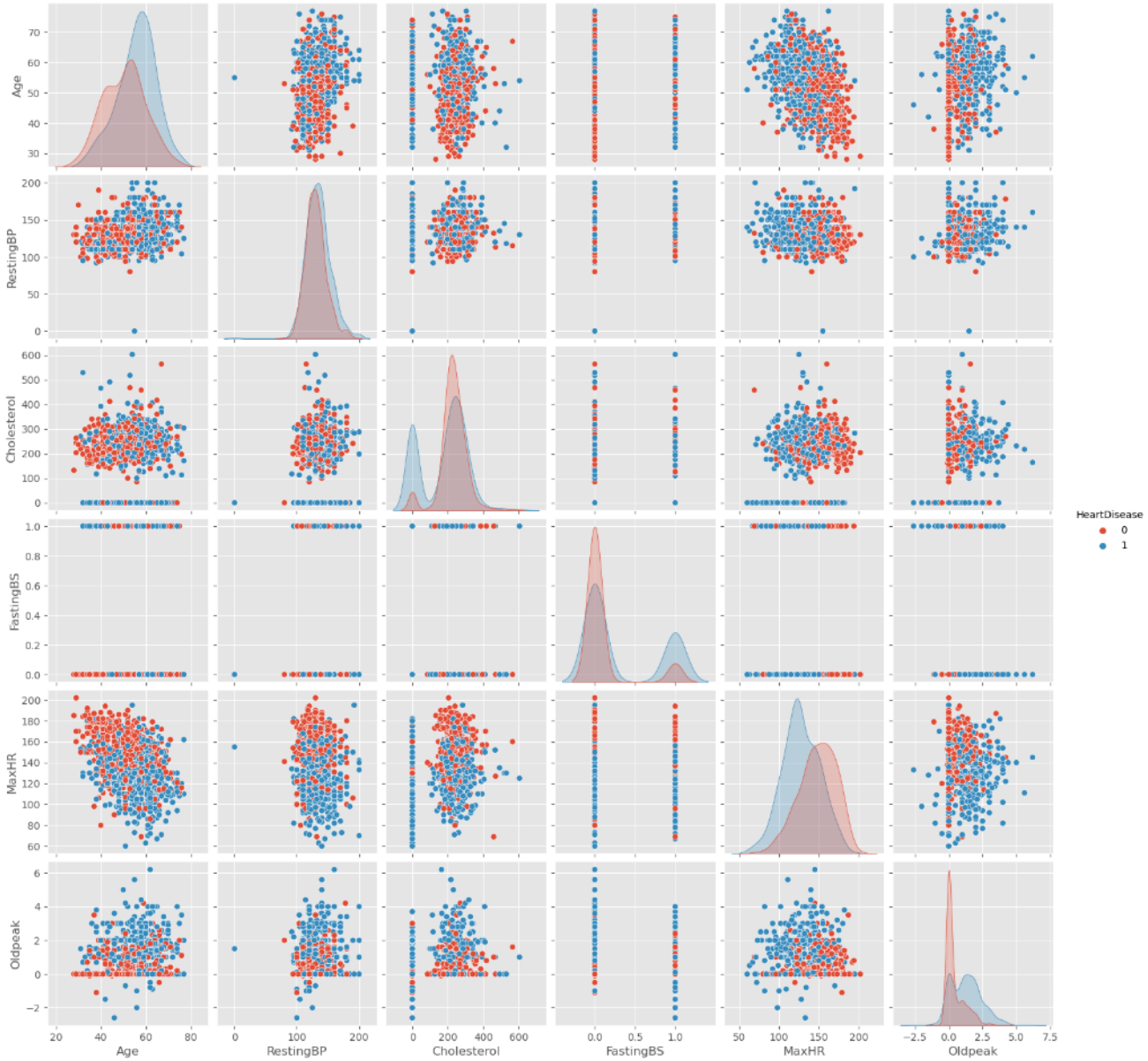


ST뷰절 기울기



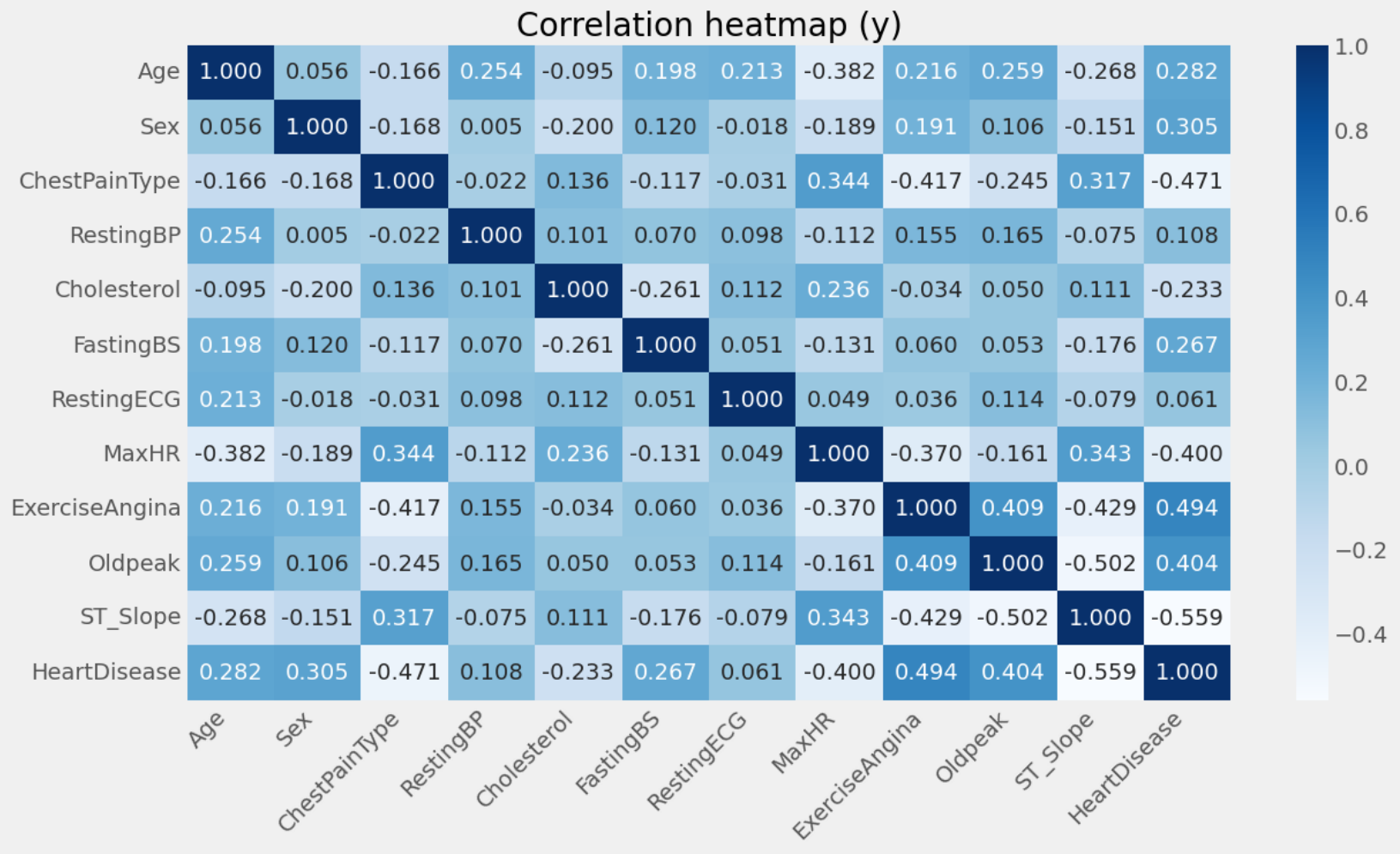
EDA

Pairplot



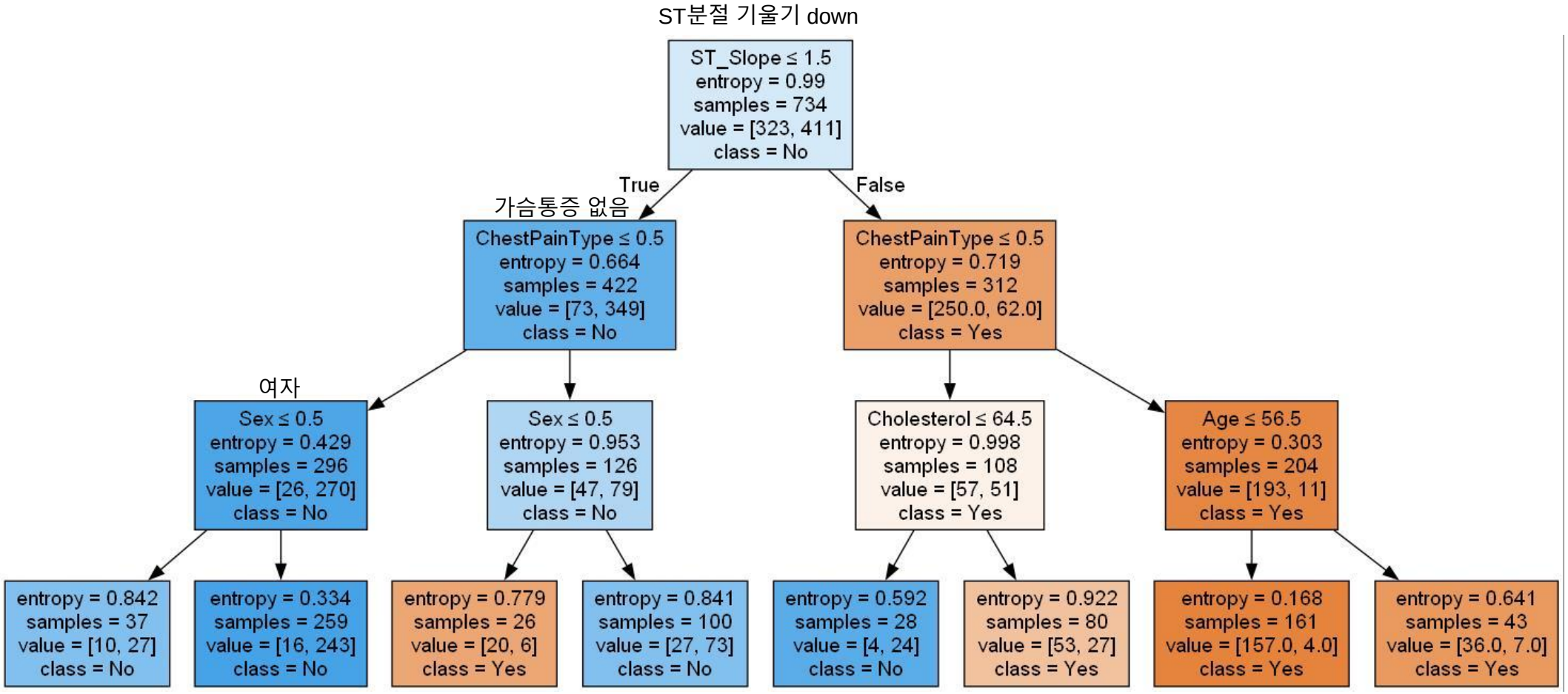
EDA

Correlation heatmap



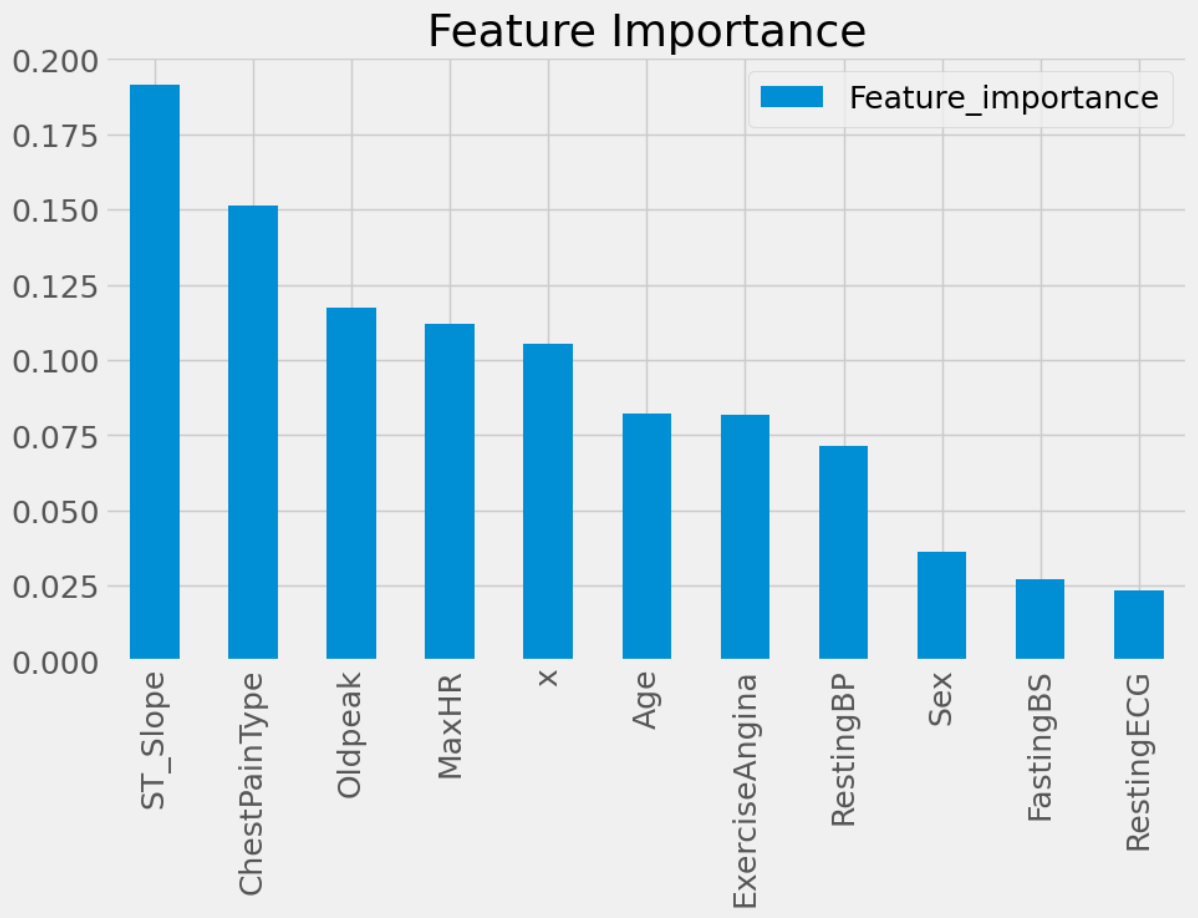
Decision tree

- Training data, test data를 80:20로 분리
- Accuracy score: 0.831, F1 score: 0.843, AUROC score: 0.830



Random forest

- Training data, test data를 80:20로 분리
- Accuracy score: 0.836, F1 score: 0.85, AUROC score: 0.835



Performance

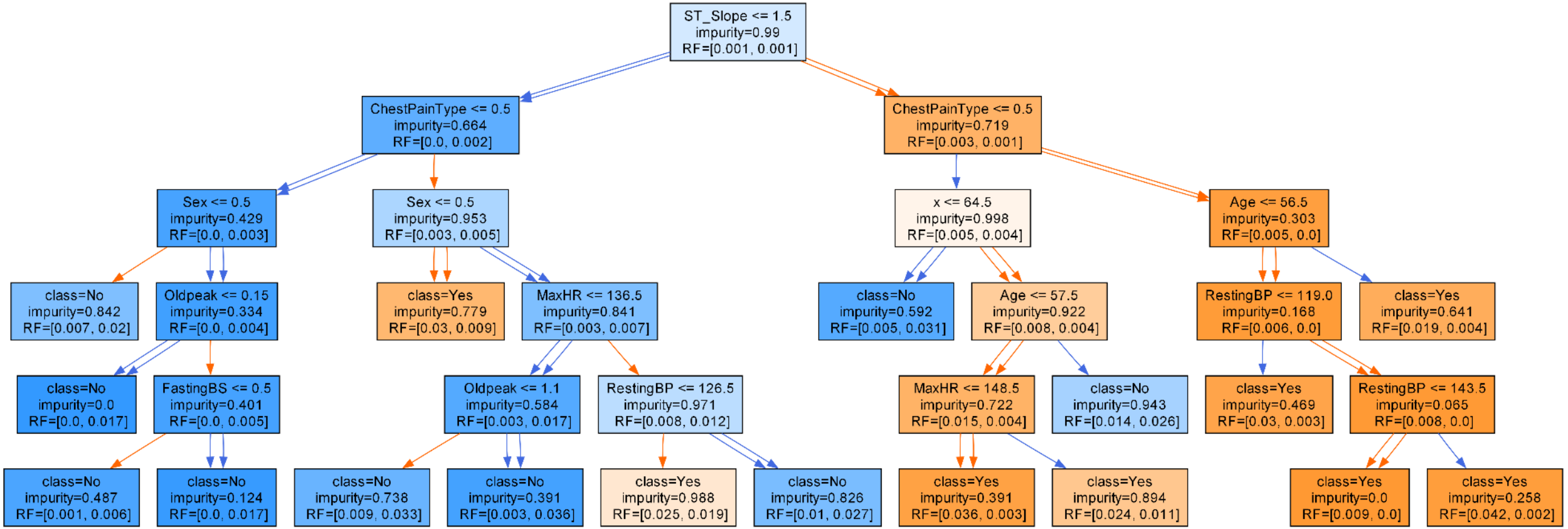
예측 성능 비교

	Decision tree	Logistic regression	Gradient boosting	Random forest	Naïve Bayes	K-NN
Accuracy score	0.831	0.826	0.842	0.836	0.826	0.674
F1 score	0.843	0.837	0.856	0.85	0.835	0.709
AUROC score	0.830	0.825	0.840	0.835	0.825	0.669

- 성능이 가장 높은 모델: Gradient boosting

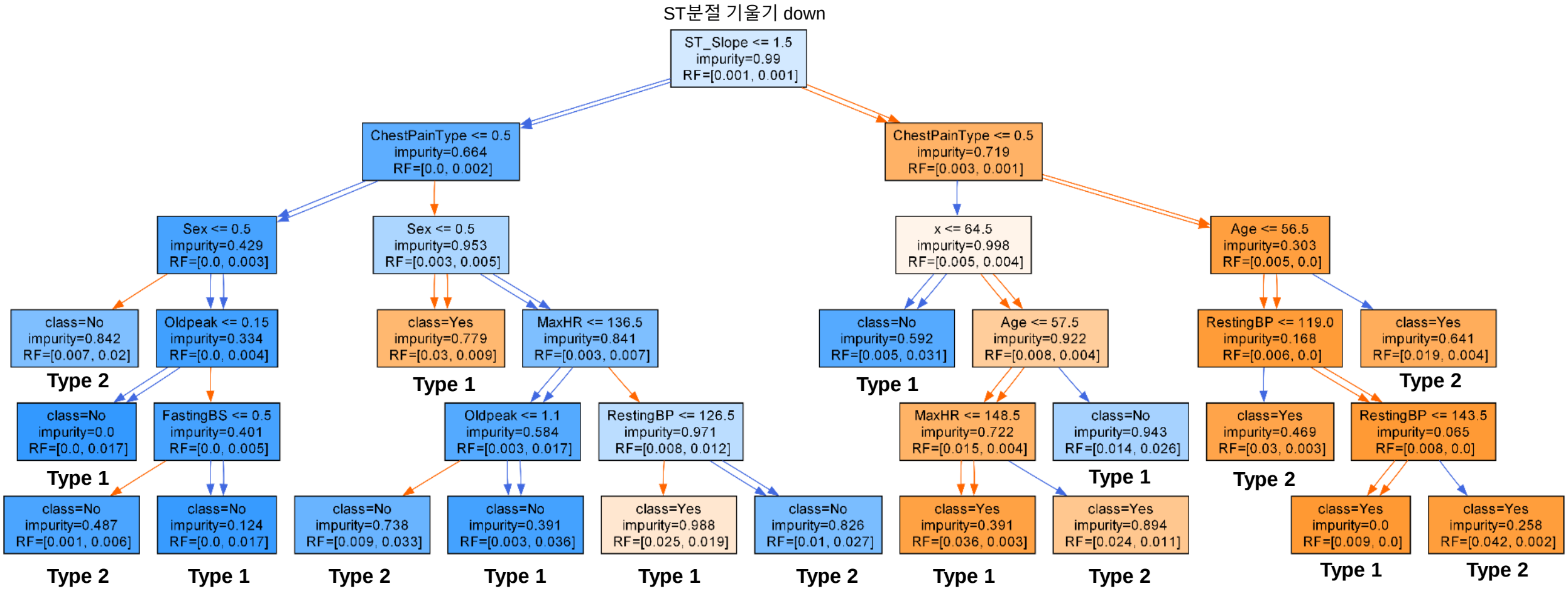
Color DT

Accuracy score: 0.946



Color DT

Type 1, type 2로 분리



Color DT

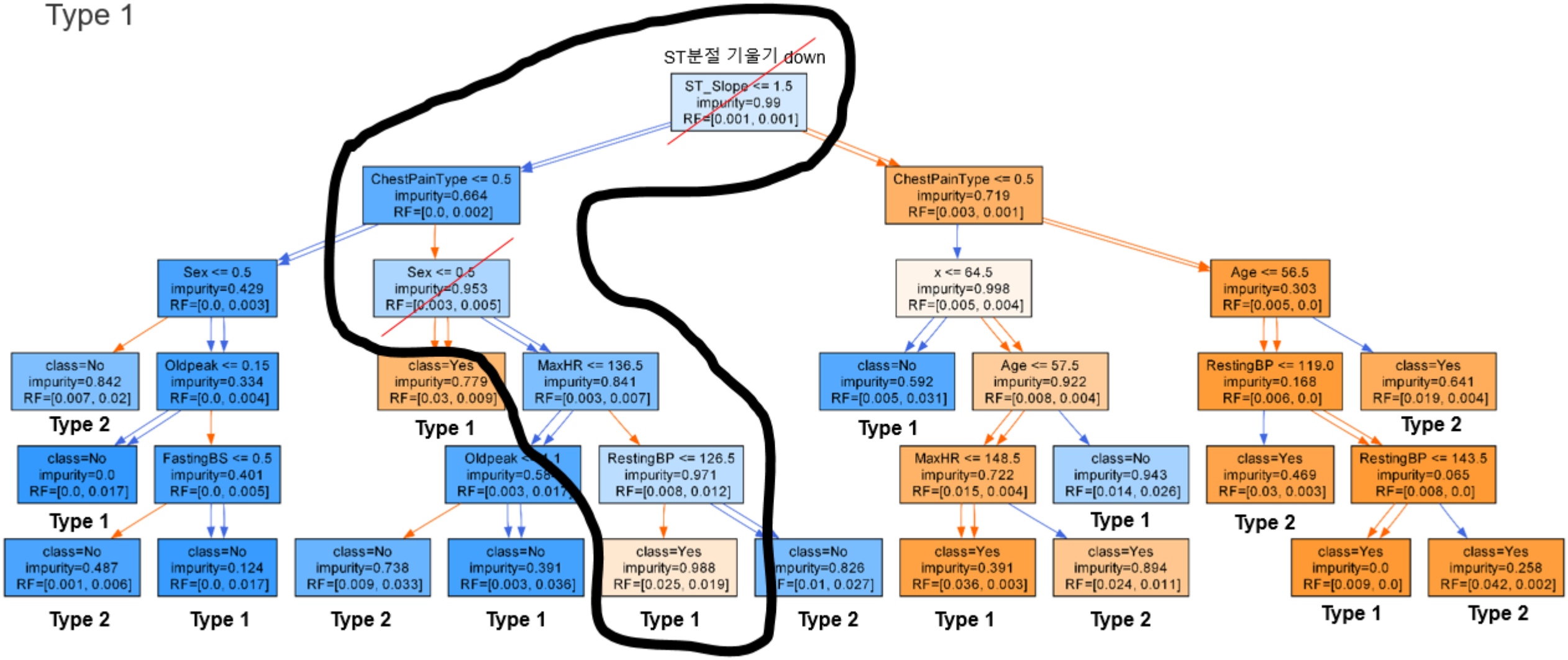
Type1인 rule에서 irrelevant condition 제거

Rule	Rule (Type 1)	Class	RF
R1	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex > 0.5 , Oldpeak <= 0.15	0	1.0
R2	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex > 0.5 , Oldpeak > 0.15 , FastingBS > 0.5	0	0.983
R3	ST_Slope <= 1.5 , ChestPainType > 0.5 , Sex <= 0.5	1	0.769
R4	ST_Slope <= 1.5 , ChestPainType > 0.5 , Sex > 0.5 , MaxHR <= 136.5 , Oldpeak > 1.10	0	0.923
R5	ST_Slope <= 1.5 , ChestPainType > 0.5 , Sex > 0.5 , MaxHR > 136.5 , RestingBP <= 126.5	1	0.565
R6	ST_Slope > 1.5 , ChestPainType <= 0.5 , Cholesterol <= 64.5	0	0.857
R7	ST_Slope > 1.5 , ChestPainType <= 0.5 , Cholesterol > 64.5 , Age <= 57.5 , MaxHR <= 148.5	1	0.923
R8	ST_Slope > 1.5 , ChestPainType <= 0.5 , Cholesterol > 64.5 , Age > 57.5	0	0.64
R9	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , RestingBP > 119.0 , RestingBP <= 143.5	1	1.0

Irrelevant condition을 제거한 rule

Rule	Compact rule (Type 1)	Class	RF
R1*	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex > 0.5 , Oldpeak <= 0.15	0	1.0
R2*	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex > 0.5 , ~~Oldpeak > 0.15~~ , FastingBS > 0.5	0	0.988
R3*	~~ST_Slope <= 1.5~~ , ChestPainType > 0.5 , Sex <= 0.5	1	0.921
R4*	ST_Slope <= 1.5 , ~~ChestPainType > 0.5~~ , Sex > 0.5 , MaxHR <= 136.5 , Oldpeak > 1.10	0	0.95
R5*	~~ST_Slope <= 1.5~~ , ChestPainType > 0.5 , ~~Sex > 0.5~~ , MaxHR > 136.5 , RestingBP <= 126.5	1	0.856
R6*	~~ST_Slope > 1.5~~ , ChestPainType <= 0.5 , Cholesterol <= 64.5	0	0.943
R7*	ST_Slope > 1.5 , ~~ChestPainType <= 0.5~~ , Cholesterol > 64.5 , Age <= 57.5 , MaxHR <= 148.5	1	0.964
R8*	~~ST_Slope > 1.5~~ , ChestPainType <= 0.5 , ~~Cholesterol > 64.5~~ , Age > 57.5	0	0.866
R9*	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , RestingBP > 119.0 , RestingBP <= 143.5	1	1.0

Color DT


Type 1
ST분절 기울기 down
ST_Slope <= 1.5
impurity=0.99
RF=[0.001, 0.001]
ChestPainType <= 0.5
impurity=0.664
RF=[0.0, 0.002]
ChestPainType <= 0.5
impurity=0.719
RF=[0.003, 0.001]
Sex <= 0.5
impurity=0.429
RF=[0.0, 0.003]
Sex <= 0.5
impurity=0.953
RF=[0.003, 0.005]
x <= 64.5
impurity=0.998
RF=[0.005, 0.004]
Age <= 56.5
impurity=0.303
RF=[0.005, 0.0]
class=No
impurity=0.842
RF=[0.007, 0.02]
Oldpeak <= 0.15
impurity=0.334
RF=[0.0, 0.004]
class=Yes
impurity=0.779
RF=[0.03, 0.009]
MaxHR <= 136.5
impurity=0.841
RF=[0.003, 0.007]
class=No
impurity=0.592
RF=[0.005, 0.031]
Age <= 57.5
impurity=0.922
RF=[0.008, 0.004]
RestingBP <= 119.0
impurity=0.168
RF=[0.006, 0.0]
class=Yes
impurity=0.641
RF=[0.019, 0.004]
Type 2
Type 1
Type 1
Type 2
class=No
impurity=0.0
RF=[0.0, 0.017]
FastingBS <= 0.5
impurity=0.401
RF=[0.0, 0.005]
Oldpeak <= 1.1
impurity=0.58
RF=[0.003, 0.017]
RestingBP <= 126.5
impurity=0.971
RF=[0.008, 0.012]
MaxHR <= 148.5
impurity=0.722
RF=[0.015, 0.004]
class=No
impurity=0.943
RF=[0.014, 0.026]
class=Yes
impurity=0.469
RF=[0.03, 0.003]
RestingBP <= 143.5
impurity=0.065
RF=[0.008, 0.0]
Type 1
Type 1
Type 2
Type 1
class=No
impurity=0.487
RF=[0.001, 0.006]
class=No
impurity=0.124
RF=[0.0, 0.017]
class=No
impurity=0.738
RF=[0.009, 0.033]
class=No
impurity=0.391
RF=[0.003, 0.036]
class=Yes
impurity=0.988
RF=[0.025, 0.019]
class=No
impurity=0.826
RF=[0.01, 0.027]
class=Yes
impurity=0.391
RF=[0.036, 0.003]
class=Yes
impurity=0.894
RF=[0.024, 0.011]
class=Yes
impurity=0.0
RF=[0.009, 0.0]
class=Yes
impurity=0.258
RF=[0.042, 0.002]
Type 2
Type 1
Type 1
Type 2
Type 1
Type 2
Type 1
Type 2

Color DT

Type 2 – method 1) 경로에서 irrelevant condition 제거

Rule	Rule (Type 2)	Class	RF
R1	ST_Slope <= 1.5 , ChestPainType <= 0.5, Sex <= 0.5	0	0.73
R2	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex > 0.5 , Oldpeak > 0.15 , FastingBS <= 0.5	0	0.894
R3	ST_Slope <= 1.5 , ChestPainType > 0.5 , Sex > 0.5 , MaxHR <= 136.5 , Oldpeak <= 1.10	0	0.792
R4	ST_Slope <= 1.5 , ChestPainType > 0.5 , Sex > 0.5 , MaxHR > 136.5 , RestingBP > 126.5	0	0.741
R5	ST_Slope > 1.5 , ChestPainType <= 0.5 , Cholesterol > 64.5 , Age <= 57.5 , MaxHR > 148.5	1	0.69
R6	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , RestingBP <= 119.0	1	0.9
R7	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , RestingBP > 119.0 , RestingBP > 143.5	1	0.957
R8	ST_Slope > 1.5 , ChestPainType > 0.5 , Age > 56.5	1	0.837

Irrelevant condition을 제거한 rule (method 1)

Rule	Compact rule (Type 2)	Class	RF
R1*	ST_Slope <= 1.5 , ChestPainType <= 0.5, ~~Sex <= 0.5~~	0	0.912
R2*	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex > 0.5 , ~~Oldpeak > 0.15 , FastingBS <= 0.5~~	0	0.938
R3*	ST_Slope <= 1.5 , ~~ChestPainType > 0.5~~ , Sex > 0.5 , MaxHR <= 136.5 , ~~Oldpeak <= 1.10~~	0	0.932
R4*	ST_Slope <= 1.5 , ~~ChestPainType > 0.5~~ , Sex > 0.5 , ~~MaxHR > 136.5~~ , RestingBP > 126.5	0	0.913
R5*	ST_Slope > 1.5 , ~~ChestPainType <= 0.5~~ , Cholesterol > 64.5 , Age <= 57.5 , ~~MaxHR > 148.5~~	1	0.931
R6*	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , ~~RestingBP <= 119.0~~	1	0.975
R7*	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , RestingBP > 119.0 , ~~RestingBP > 143.5~~	1	0.992
R8*	ST_Slope > 1.5 , ChestPainType > 0.5 , ~~Age > 56.5~~	1	0.946

Color DT

Type 2 – method 2) 경로에서 RF가 가장 높은 노드보다 위에 존재하는 irrelevant condition만을 제거

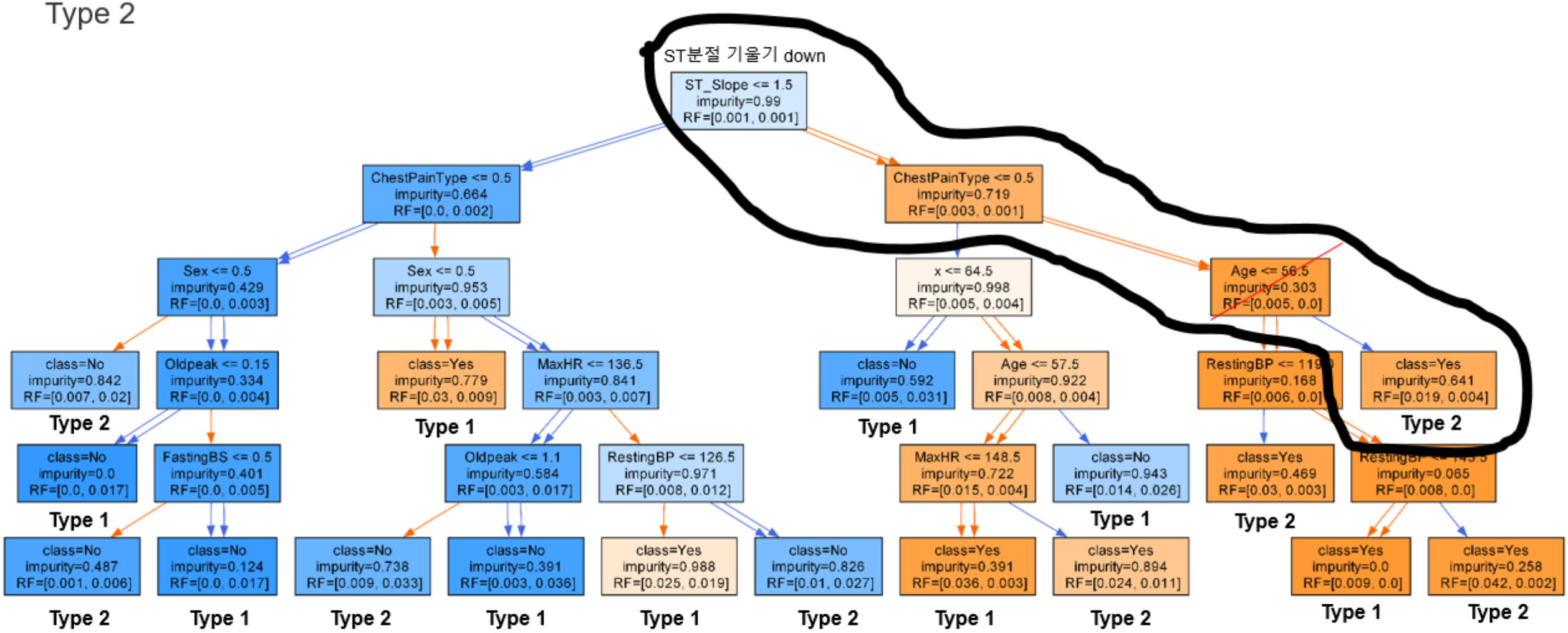
Rule	Rule (Type 2)	Class	RF
R1	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex <= 0.5	0	0.73
R2	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex > 0.5 , Oldpeak > 0.15000000223517418 , FastingBS <= 0.5	0	0.894
R3	ST_Slope <= 1.5 , ChestPainType > 0.5 , Sex > 0.5 , MaxHR <= 136.5 , Oldpeak <= 1.10	0	0.792
R4	ST_Slope <= 1.5 , ChestPainType > 0.5 , Sex > 0.5 , MaxHR > 136.5 , RestingBP > 126.5	0	0.741
R5	ST_Slope > 1.5 , ChestPainType <= 0.5 , Cholesterol > 64.5 , Age <= 57.5 , MaxHR > 148.5	1	0.69
R6	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , RestingBP <= 119.0	1	0.9
R7	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , RestingBP > 119.0 , RestingBP > 143.5	1	0.957
R8	ST_Slope > 1.5 , ChestPainType > 0.5 , Age > 56.5	1	0.837

Irrelevant condition을 제거한 rule (method 2)

Rule	Compact rule (Type 2)	Class	RF
R1*	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex <= 0.5	0	0.73
R2*	ST_Slope <= 1.5 , ChestPainType <= 0.5 , Sex > 0.5 , Oldpeak > 0.15000000223517418 , FastingBS <= 0.5	0	0.894
R3*	ST_Slope <= 1.5 , ~~ChestPainType > 0.5~~ , Sex > 0.5 , MaxHR <= 136.5 , Oldpeak <= 1.10	0	0.909
R4*	ST_Slope <= 1.5 , ChestPainType > 0.5 , Sex > 0.5 , MaxHR > 136.5 , RestingBP > 126.5	0	0.741
R5*	ST_Slope > 1.5 , ChestPainType <= 0.5 , Cholesterol > 64.5 , Age <= 57.5 , MaxHR > 148.5	1	0.69
R6*	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , RestingBP <= 119.0	1	0.9
R7*	ST_Slope > 1.5 , ChestPainType > 0.5 , Age <= 56.5 , RestingBP > 119.0 , RestingBP > 143.5	1	0.957
R8*	ST_Slope > 1.5 , ChestPainType > 0.5 , Age > 56.5	1	0.837

Color DT

Type 2



Color DT

Type별 RF table

Rule Type	Rule Count	RF Diff < Epsilon	Ratio
Type 1-n	2	2	1.00
Type 1-y	8	4	0.50
Type 2-m1	9	3	0.33
Type 2-m2-n	8	8	1.00
Type 2-m2-y	1	1	1.00

Type별 RF table의 평균

Rule Type	Rule Count	RF Diff < Epsilon	Ratio
Type 1-n	2.000	2.000	1.000
Type 1-y	7.867	4.800	0.605
Type 2-m1	10.833	3.167	0.285
Type 2-m2-n	7.933	7.933	1.000
Type 2-m2-y	2.900	1.933	0.673

최종 요약

- 6개의 모델을 사용하여 심장질환 예측을 수행
- Gradient boosting 모델의 예측 성능이 가장 높음
- 여러 개의 모델을 훈련했을때 가장 중요한 특성은 ST 분질 기울기
- Color DT 그리고 type 1, type 2로 분리하여 Irrelevant condition을 제거한 compact rule들을 만들었다
- 총 918명의 환자의 데이터를 가지고 모델을 훈련하였는데 더 많은 환자의 데이터를 모집해 훈련을 한면 더 성능이 좋은 모델을 만들 수 있을 것이라고 예상된다