Python EDA & Classification

Heart Diseases: What are the indicators?

Image source

“ ”

[ ]

* EDA *

Table of Contents

1. Exploratory Data Analysis

```
Shape of the data is  (303, 14)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

	0
age	int64
sex	int64
cp	int64
trestbps	int64
chol	int64
fbs	int64
restecg	int64
thalach	int64
exang	int64
oldpeak	float64
slope	int64
ca	int64
thal	int64
target	int64

```
int64/float64
```
/

1.1

1. age:
2. sex:
   - 1 =
   - 0 =
3. cp:
   - 0
   - 1
   - 2

- 3

4. trestbps: mm Hg
5. chol: mg/dl
6. fbs: > 120 mg/dl
   - 1 =
   - 0 =
7. restecg:
   - 0:
   - 1: ST-T T / ST > 0.05 mV
   - 2: Estes'
8. thalach:
9. exang:
   - 1 =
   - 0 =
10. oldpeak = ST
11. slope: ST
    - 0:
    - 1:
    - 2:
12. ca: 0-3
13. thal:
    - 0 = `0 NaN`
    - 1 =
    - 2 =
    - 3 =
14. target ( )
    - 0 =
    - 1 = :

      `0: < 50%` `1: > 50%`

- #93 159 164 165 252 `ca=4` NaNs
- #49 282 `thal = 0` NaNs

7

1.2

1.2.1

7

296 303

1.2.2

- / UCL
- 0 1 2 .. ‘ ’, ,
- Rob Harrand’s

	0
age	int64
sex	object
chest_pain_type	object
resting_blood_pressure	int64
cholesterol	int64
fasting_blood_sugar	object
resting_electrocardiogram	object
max_heart_rate_achieved	int64
exercise_induced_angina	object
st_depression	float64
st_slope	object
num_major_vessels	int64
thalassemia	object
target	int64

	age	sex	chest_pain_type	resting_blood_pressure	cholesterol	fasting_blood_sugar	resting
0	63	male	asymptomatic	145	233	greater than 120mg/ml	norma
1	37	male	non-anginal pain	130	250	lower than 120mg/ml	ST-T
2	41	female	atypical angina	130	204	lower than 120mg/ml	norma
3	56	male	atypical angina	120	236	lower than 120mg/ml	ST-T
4	57	female	typical angina	120	354	lower than 120mg/ml	ST-T

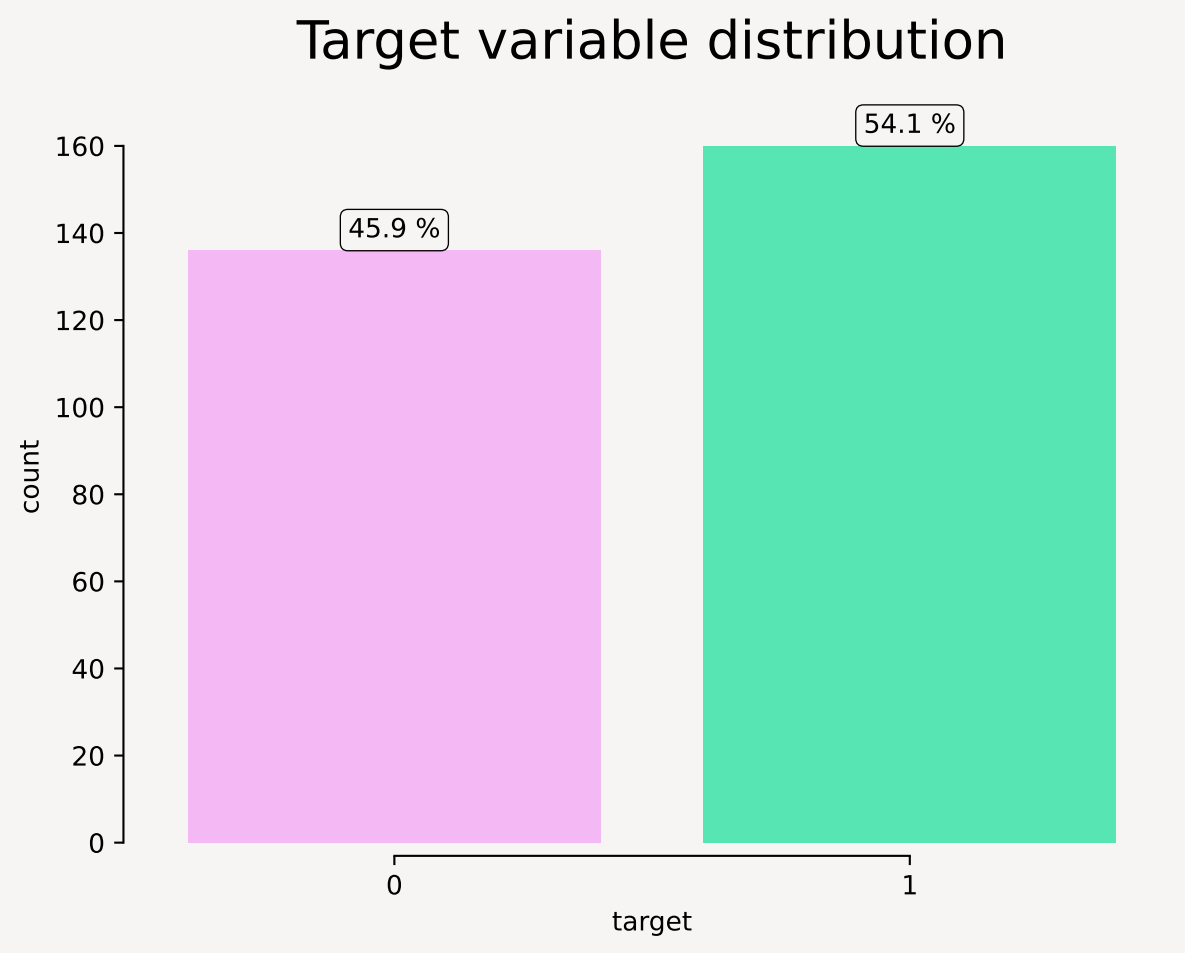
1.2.3

-

1.3 /

1.3.1

46% 54%



1.3.2

pandas `data.describe()`

	count	mean	std	min	25%	50%	75%	max
age	296.0	54.523649	9.059471	29.0	48.0	56.0	61.00	77.0
cholesterol	296.0	247.155405	51.977011	126.0	211.0	242.5	275.25	564.0
resting_blood_pressure	296.0	131.604730	17.726620	94.0	120.0	130.0	140.00	200.0
max_heart_rate_achieved	296.0	149.560811	22.970792	71.0	133.0	152.5	166.00	202.0
st_depression	296.0	1.059122	1.166474	0.0	0.0	0.8	1.65	6.2
num_major_vessels	296.0	0.679054	0.939726	0.0	0.0	0.0	1.00	3.0

:

54.5

77 29

247.15

564 126

[6] $< 200mg/dl$

131 200 94

149.5 bpm 202 71bpm

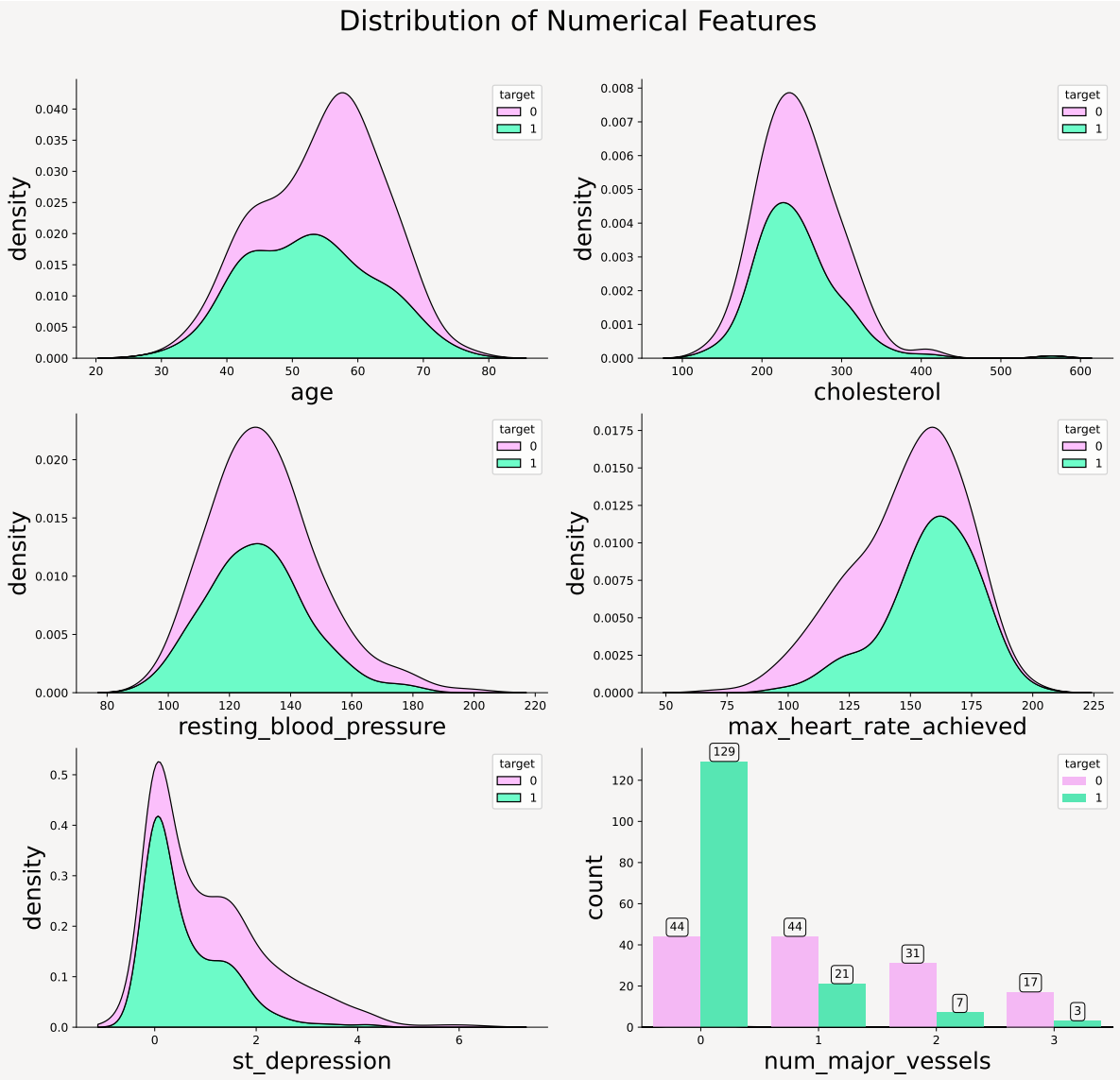
ST

st_dpression 1.06 6.2 0

3 0 0.68

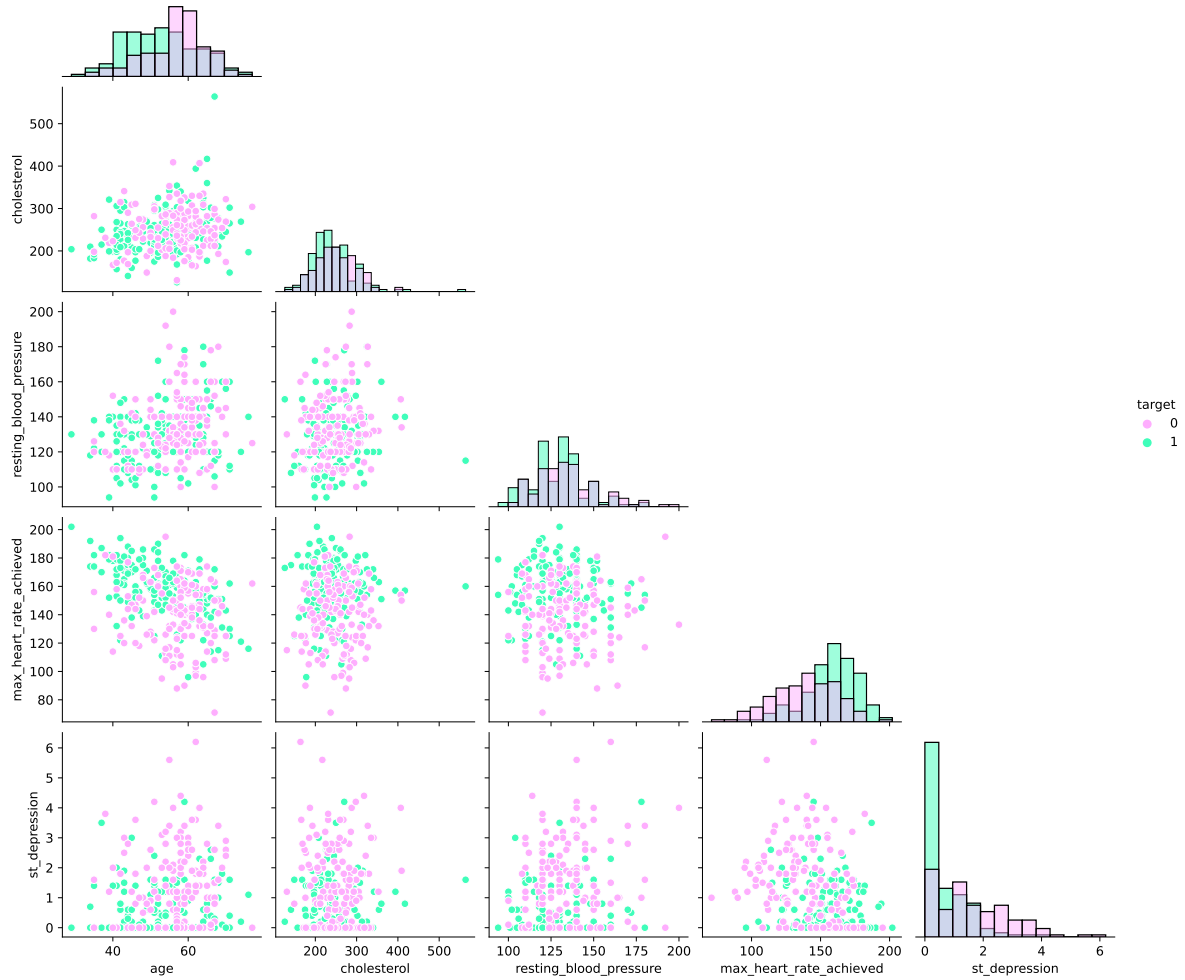
Distribution: Density plots

```
Text(0.5, 0.98, 'Distribution of Numerical Features')
```



Pair-plots

```
Text(0.5, 0.98, 'Pairplot: Numerical Features ')
```



reg-plots - maximum_heart_rate_achieved age st_depression - maximum_heart_rate_achieved - st_depression

```
<matplotlib.legend.Legend at 0x7f4e37020040>
```

1.3.3

=1 =0

75%

~1.4% ST-T REC 50-50

ST-T REC

ST

REC ST

ST

+

85% 120mg/ml

~54%

76% ~69%

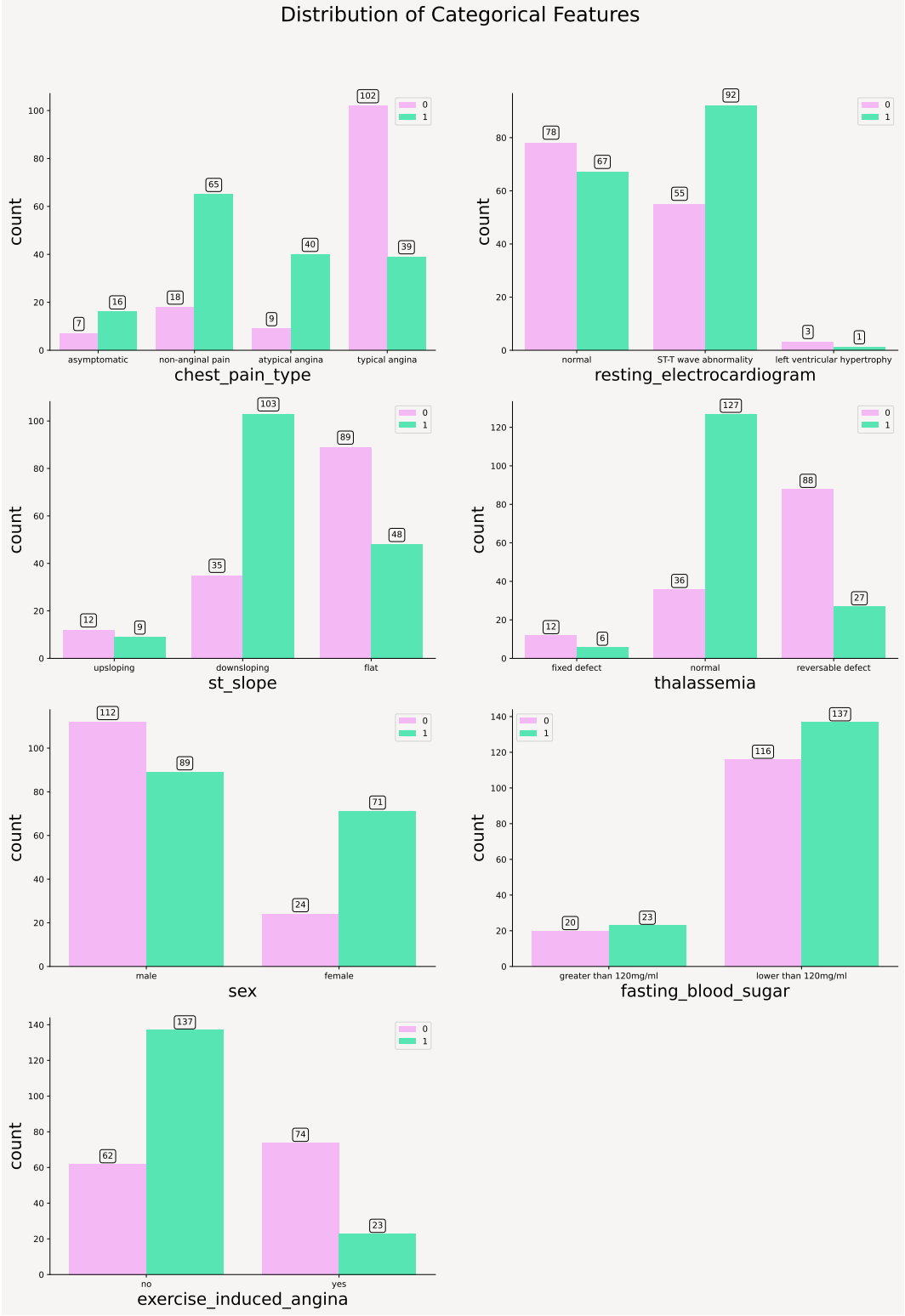
{''.join(['

' + dfs._repr_html_() + '

' for dfs in dataFrames_])}

Distribution: Count plots

0


Distribution of Categorical Features
count
chest_pain_type
asymptomatic
non-anginal pain
atypical angina
typical angina
7
16
18
65
9
40
102
39
resting_electrocardiogram
normal
ST-T wave abnormality
left ventricular hypertrophy
78
67
55
92
3
1
st_slope
upsloping
downsloping
flat
12
9
35
103
89
48
thalassemia
fixed defect
normal
reversable defect
12
6
36
127
88
27
sex
male
female
112
89
24
71
fasting_blood_sugar
greater than 120mg/ml
lower than 120mg/ml
20
23
116
137
exercise_induced_angina
no
yes
62
137
74
23
0
1

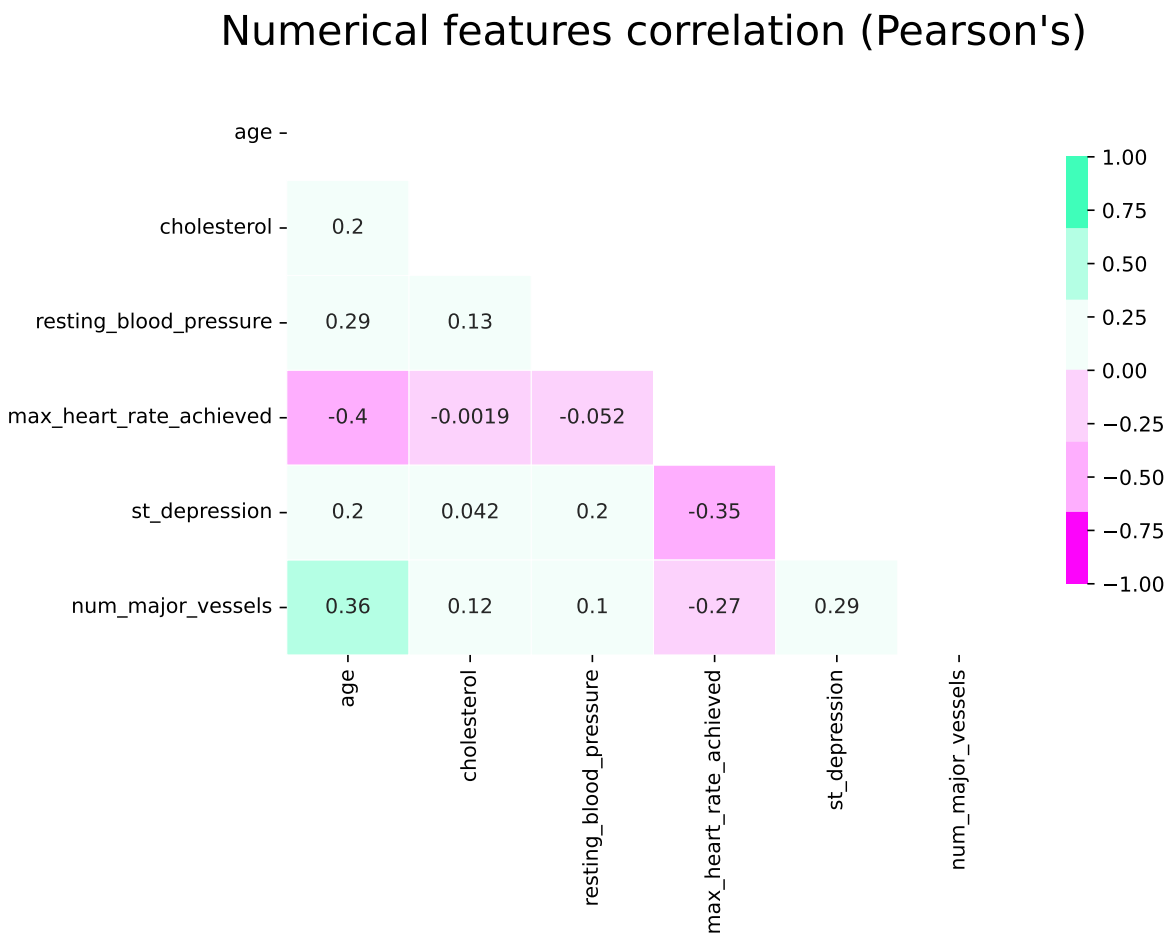
1.4

V

1.4.1

- -1 1

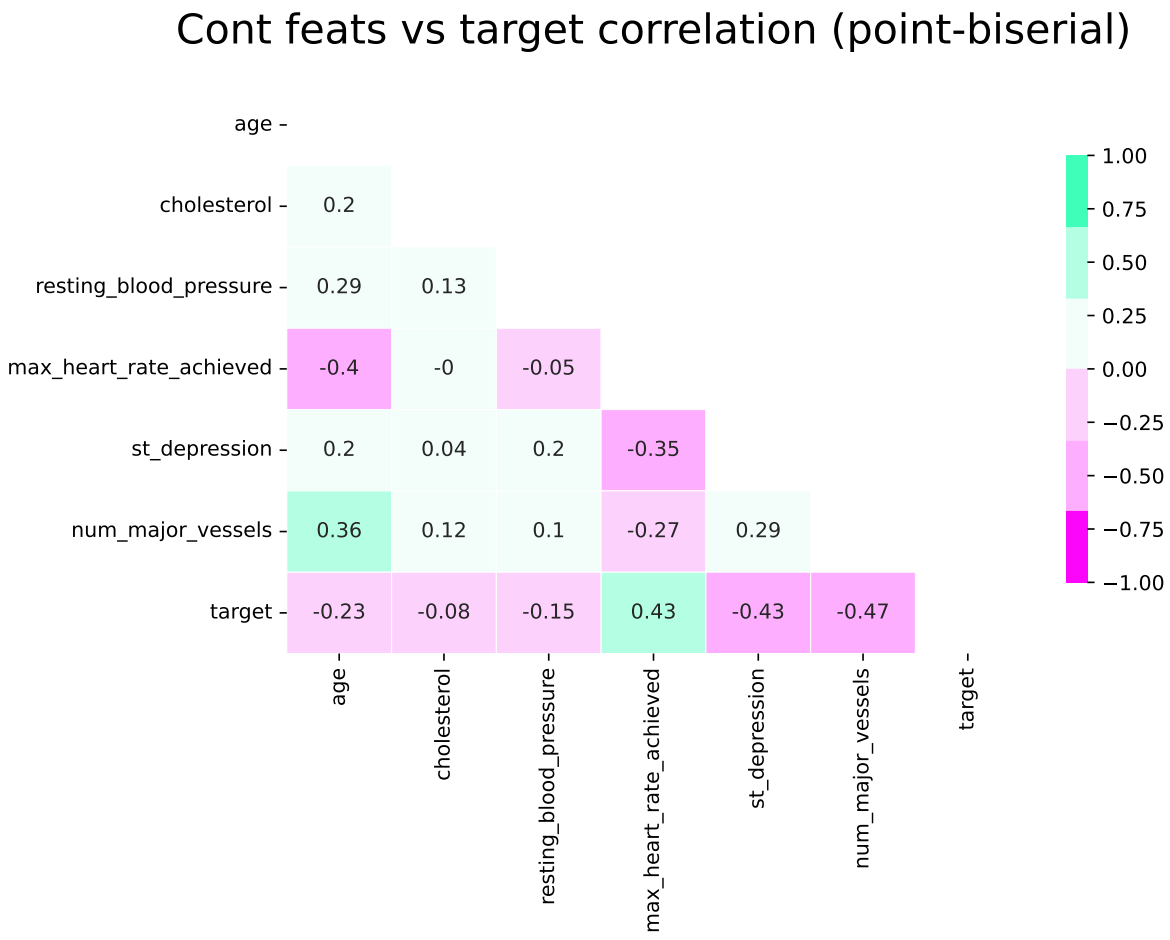
```
Text(0.5, 1.05, "Numerical features correlation (Pearson's)")
```



1.4.2 -

- - [ ]

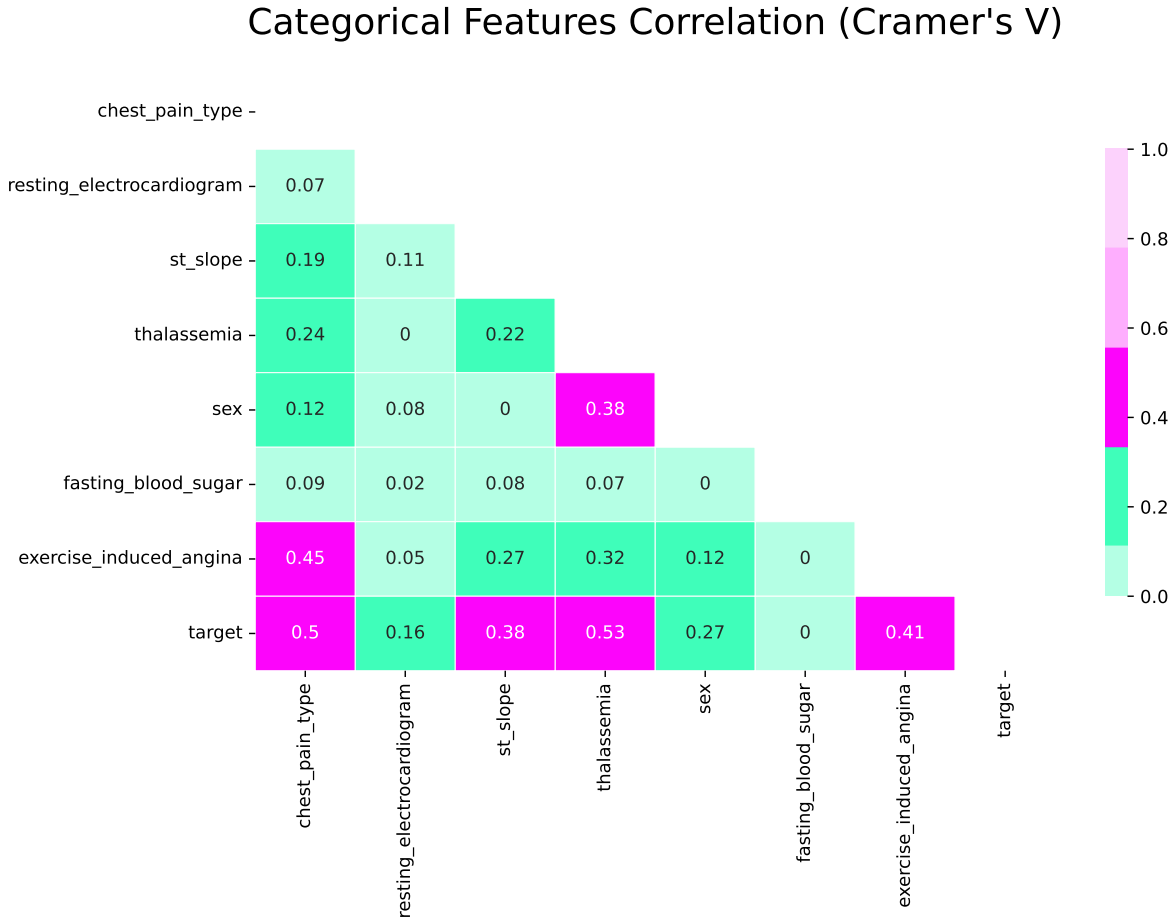
```
Text(0.5, 1.05, 'Cont feats vs target correlation (point-biserial)')
```



1.4.3 V

- V 0 +1 · 1946 [ ]

```
Text(0.5, 1.05, "Categorical Features Correlation (Cramer's V)")
```



Back to top

1.5 EDA

- 303 14 13 + > 296
-
-
  –
  –
- 54% 46%
-
  –
  – num_major_vessels max_heart_rate_achieved st_depression -
    0.47 0.43 -0.43

– chest_pain_type num_major_vessels thalassemia exercise_induced_angina thala
–

chest_pain_type num_major_vessels thalassemia exercise_induced_angina max_heart_rate_achieved st_depression

2.

: 297

2.1 Scikit Learn

Scikit learn / sklearn Nu SVC AdaBoost

2.1.1

, ,

[wiki] /

»

:

TP
FP
TN
FN

: /

$\frac{TP+TN}{TP+TN+FP+FN}$

: /

$\frac{TP}{(TP+FN)}$

:

$\frac{TP}{(TP+FP)}$

F1- :

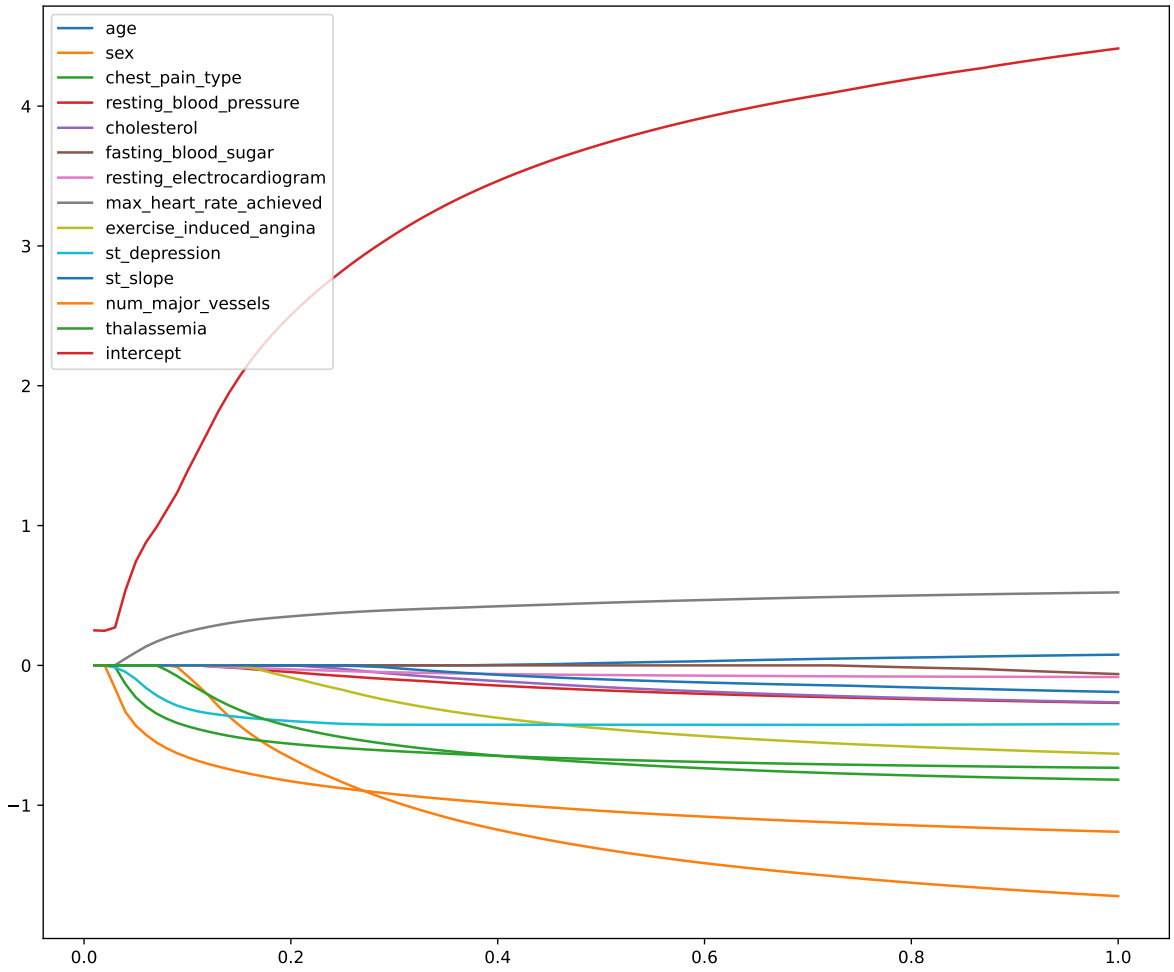
$2\frac{recall*precision}{recall+precision}$

ROC : ROC

2.1.2

$$\begin{aligned} ln\frac{p}{1-p} =& 4.50 - 0.16 * age - 1.57 * sex - 0.73cp - 0.33trestbps \\ & - 0.30chol - 0.12fbs - 0.08restecg + 0.55thalach \\ & - 0.70exang - 0.40oldpeak - 0.27slope - 1.24ca \\ & - 0.84thal \end{aligned}$$

```
[[ 0.         -1.11771927 -0.63421671 -0.14373253 -0.09988932  0.
  -0.0451463   0.40924687 -0.36128633 -0.42975961  0.         -0.96852612
  -0.61514065]]
```



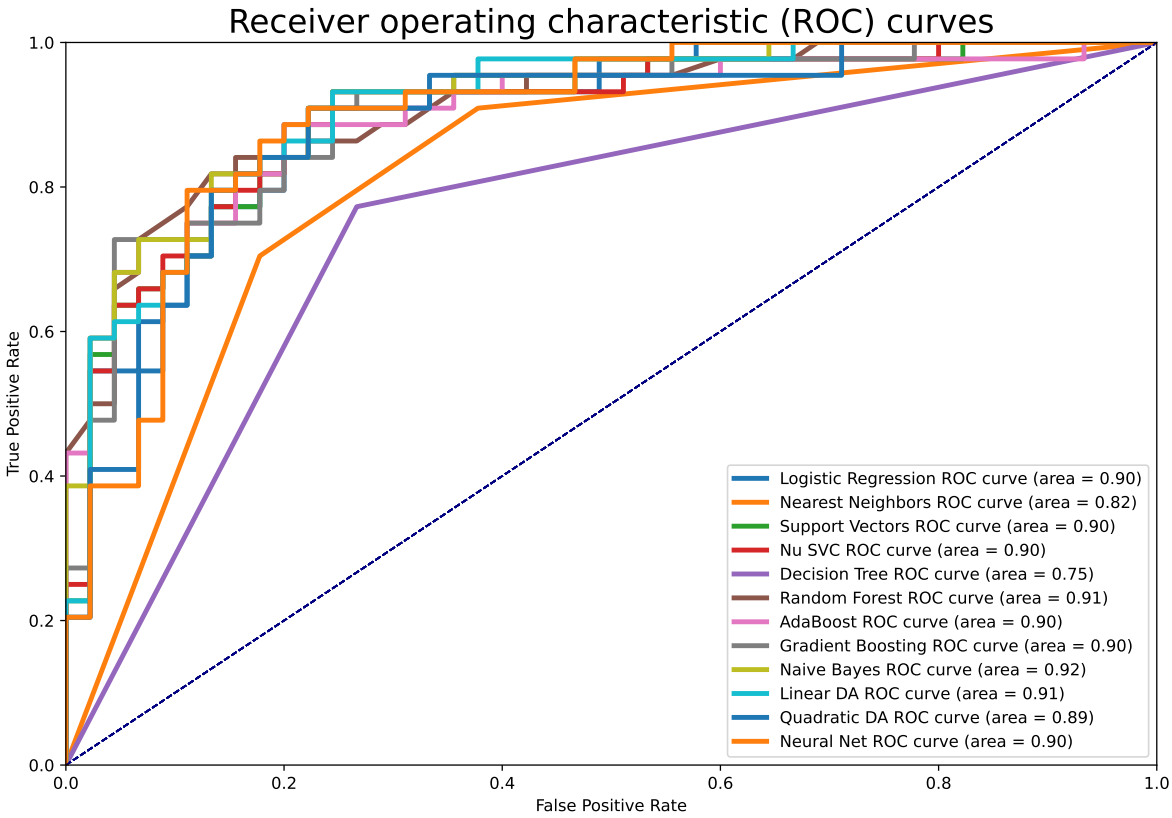
2.1.3 Performance metrics summary table

Table 2

	Classifier	Accuracy	ROC_AUC	Recall	Precision	F1
2	Support Vectors	84.270000	0.900000	0.930000	0.790000	0.850000
3	Nu SVC	84.270000	0.900000	0.930000	0.790000	0.850000
6	AdaBoost	84.270000	0.900000	0.890000	0.810000	0.850000
0	Logistic Regression	83.150000	0.900000	0.910000	0.780000	0.840000
8	Naive Bayes	83.150000	0.920000	0.890000	0.800000	0.840000
10	Quadratic DA	83.150000	0.890000	0.840000	0.820000	0.830000
11	Neural Net	83.150000	0.900000	0.910000	0.780000	0.840000
7	Gradient Boosting	82.020000	0.900000	0.890000	0.780000	0.830000

	Classifier	Accuracy	ROC_AUC	Recall	Precision	F1
9	Linear DA	82.020000	0.910000	0.860000	0.790000	0.830000
5	Random Forest	80.900000	0.910000	0.860000	0.780000	0.820000
1	Nearest Neighbors	76.400000	0.820000	0.700000	0.790000	0.750000
4	Decision Tree	75.280000	0.750000	0.770000	0.740000	0.760000

2.1.4 ROC curves



2.1.5

LR 86% 94% QDA 85% F1-

LR F1-

F1-

3.

EDA EDA ’ ’ LASSO