IBM PROFESSIONAL CERTIFICATE: Supervised Learning - Classification

Ibrahim Mohamed

July 2022

MAIN OBJECTIVE

- This analysis' primary goal is to predict the occurrence of cardiac muscle disease using various classification techniques
- This investigation tries train-test-split and cross-validation to get an idea of how these two strategies can influence model selection in different ways.
- Show the correlation between the features and the target predicted value and the most feature with impact on it.

ABOUT THE DISEASE

- Predicting and diagnosing heart disease is one of the biggest challenges in the medical industry because it depends on several factors such as the patient's physical examination and various symptoms and signs. Heart disease is considered one of the most deadly diseases in the world for the human body because the heart is unable to transport the amount of blood needed to perform the normal functions of the human body to other body organs. increase.

IBM

ABOUT THE DATA

- The dataset used in this analysis is a dedicated to the diagnoses of heart disease to different patients
- There are several factors that affect heart disease. Heart disease can be predicted based on a variety of symptoms such as age, gender, and heart rate, reducing mortality in heart disease patients. This report uses machine learning algorithms and the Python language to do this. .
- This data set has 303 records and 14 variables.

DATA EXPLORATION

- **Features:**

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.30	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.50	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.40	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.80	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.60	2	0	2	1

- **age:** patient age in years
- **sex:** patient sex (1 = male, 0 = female)
- **cp: chest pain type:**
  Value 0: asymptomatic
  Value 1: atypical angina
  Value 2: non-anginal pain
  Value 3: typical angina

- **trestbps:** resting blood pressure (mm Hg on admission to the hospital).
- **chol:** cholesterol measurement in mg/dl.
- **fbs:** fasting blood sugar (> 120 mg/dl) (1 = true; 0 = false).

IBM

DATA EXPLORATION

- **Features:**

- **restecg**: resting electrocardiographic results

  Value 0: showing probable or definite left ventricular hypertrophy by Estes' criteria
  Value 1: normal
  Value 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV).

- **thalach**: maximum heart rate achieved.
- **exang**: Exercise induced angina (1 = yes; 0 = no)
- **oldpeak**: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot)
- **slope**: the slope of the peak exercise ST segment (0: upsloping, 1: flat, 2: downsloping)
- **ca**: The number of major vessels (0–3)
- **thal**: A blood disorder called thalassemia

  Value 0: NULL (dropped from the dataset previously
  Value 1: fixed defect (no blood flow in some part of the heart)
  Value 2: normal blood flow
  Value 3: reversible defect (a blood flow is observed but it is not normal)

- **target**: Heart disease (0 = no, 1= yes)

DATA EXPLORATION

- Description:

Mean	Std	Max	Min
age: 54	age: 9	age: 77	age: 29
sex: 0.68	sex: 0.68	sex: 1	sex: 0
cp: 1	cp: 1	cp: 3	cp: 0
trestbps:131	trestbps: 17	trestbps: 200	trestbps: 94
chol: 246	chol: 51	chol: 564	chol: 126
fbs:0.15	fbs: 0.36	fbs: 1	fbs: 0
restecg: 0.53	restecg: 0.53	restecg: 2	restecg: 0
thalach: 149	thalach: 22	thalach: 202	thalach: 71
exang: 0.33	exang: 0.47	exang: 1	exang: 0
oldpeak:1	oldpeak:1	oldpeak: 6	oldpeak: 0
slope: 1.40	slope: 0.62	slope: 2	slope: 0
ca:0.73	ca:1	ca: 4	ca: 0
thal: 2	thal: 0.61	thal: 3	thal: 0
target: 0.5	target: 0.5	target: 1	target: 0

DATA EXPLORATION

- Description:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00	303.00
mean	54.37	0.68	0.97	131.62	246.26	0.15	0.53	149.65	0.33	1.04	1.40	0.73	2.31	0.54
std	9.08	0.47	1.03	17.54	51.83	0.36	0.53	22.91	0.47	1.16	0.62	1.02	0.61	0.50
min	29.00	0.00	0.00	94.00	126.00	0.00	0.00	71.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	47.50	0.00	0.00	120.00	211.00	0.00	0.00	133.50	0.00	0.00	1.00	0.00	2.00	0.00
50%	55.00	1.00	1.00	130.00	240.00	0.00	1.00	153.00	0.00	0.80	1.00	0.00	2.00	1.00
75%	61.00	1.00	2.00	140.00	274.50	0.00	1.00	166.00	1.00	1.60	2.00	1.00	3.00	1.00
max	77.00	1.00	3.00	200.00	564.00	1.00	2.00	202.00	1.00	6.20	2.00	4.00	3.00	1.00

DATA EXPLORATION

- **Data Types & Null Values**

- Our Data Types are the following
- Our data doesn’t have any missing values as it is already a small dataset so I guess it was already cleaned before.

	data
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

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

EXPLORATORY DATA ANALYSIS

Categorical features & Numerical features

- Our data contain both categorical features and numerical features
- For our luck, the categorical data is already transformed into numerical data.

- **Categorical Data:** sex, cp, fbs, restecg, exang, slope, ca, thal ,target
- **Numerical Data:** age, trestbps, thalach, oldpeak

EXPLORATORY DATA ANALYSIS

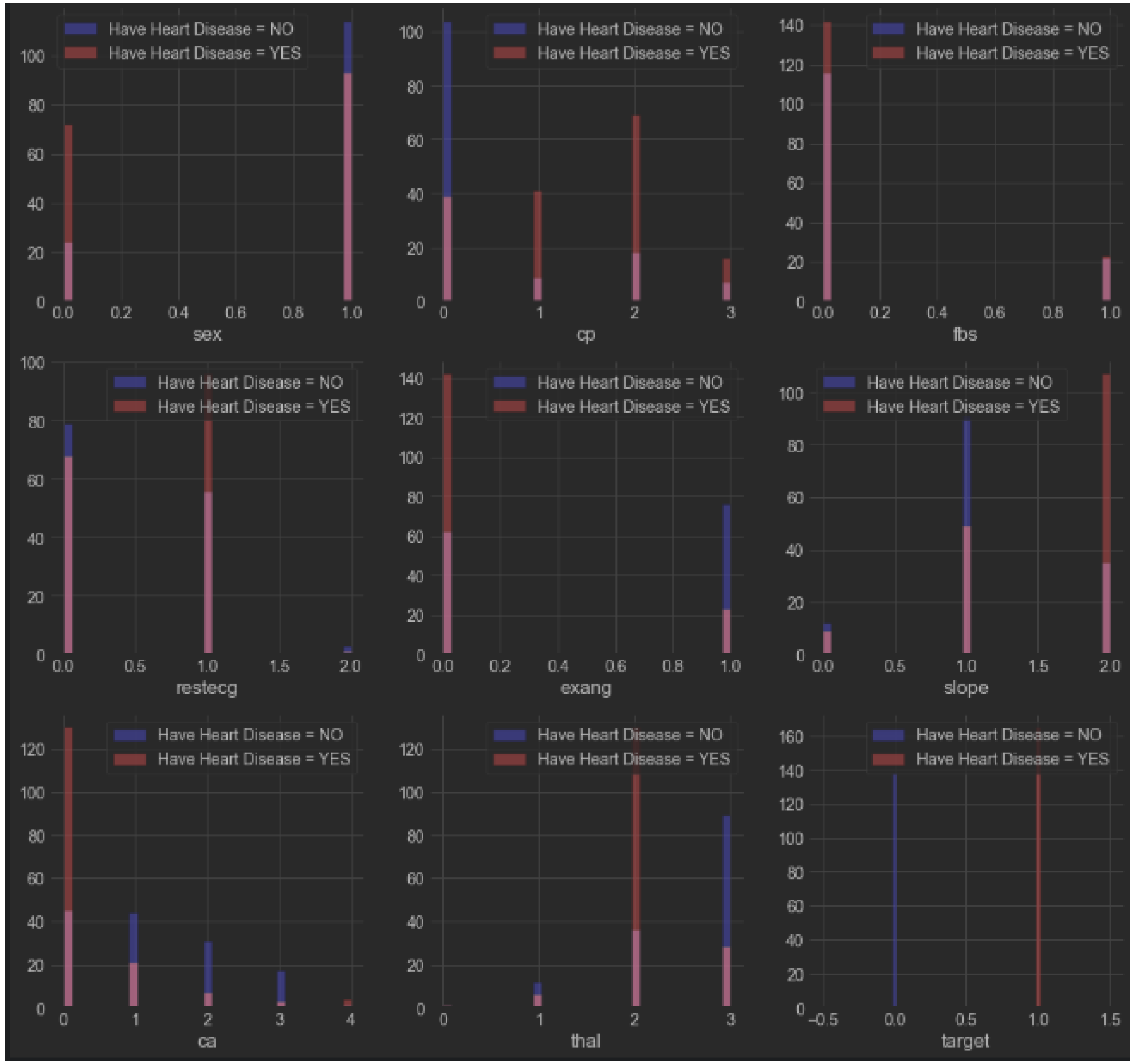
- Disease existence in the dataset

- Our data include 303 record as we said before for different 303 patient, 165 record have heart disease and 138 healthy record.

- The data is almost balanced with some plus unhealthy records

EXPLORATORY DATA ANALYSIS

- Categorical data correlation with target

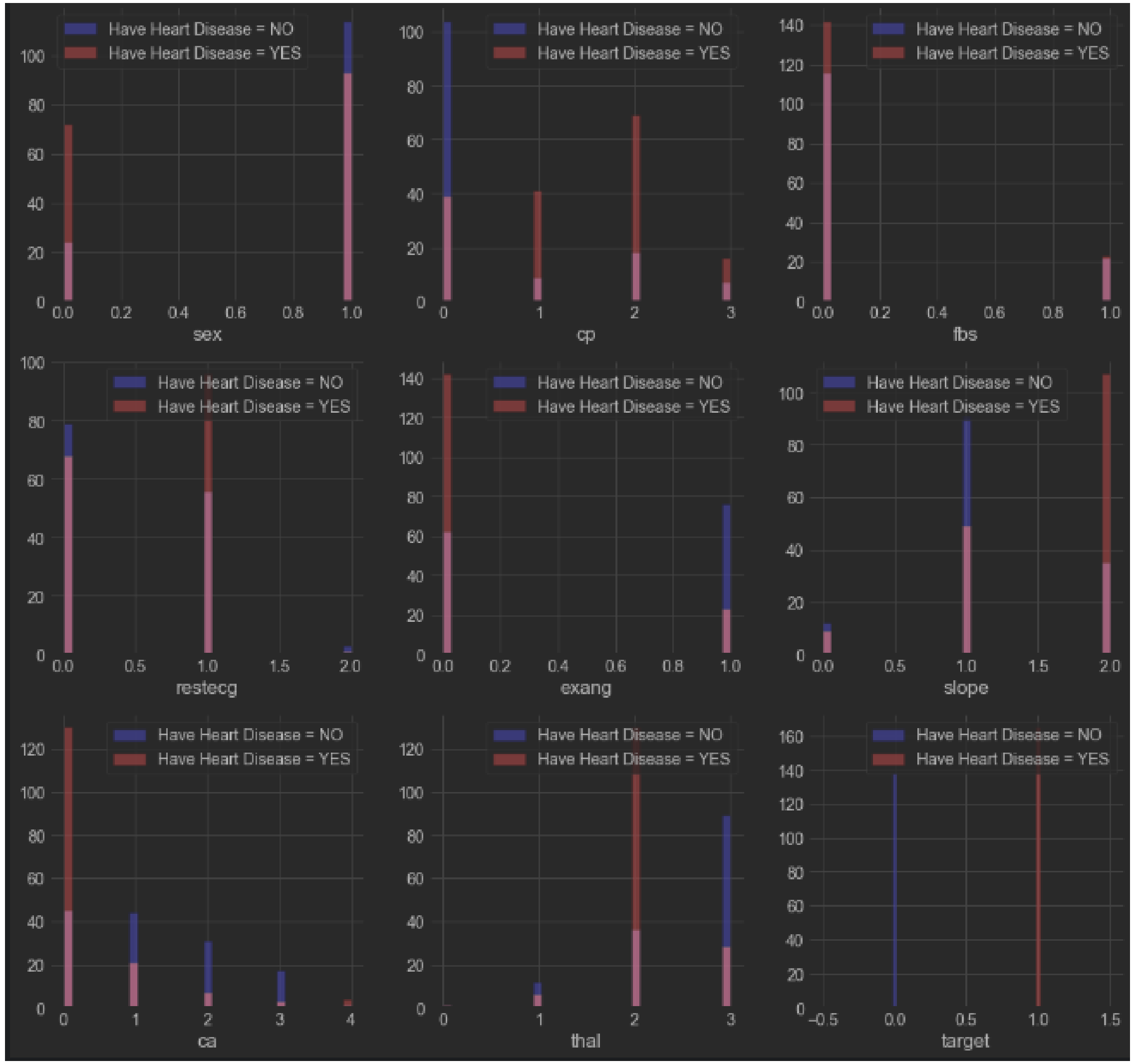


- **cp (chest pain):** patients with chest pain of the type: cp: [1, 2, 3] tend to have more heart disease than people without any chest pain cp: 0
- **restecg (resting ECG results):** patients with a value of 1 (having an abnormal heart rhythm, which can range from mild symptoms to severe problems) are more likely to develop heart disease.
- **exang (exercise-induced angina):** patients with non-exercise-induced angina who have a value of 0 are more likely to have heart disease than those who have exercise-induced angina with a value of 1.

EXPLORATORY DATA ANALYSIS

- Categorical data correlation with target

- **Slope (rectal slope for the ST segment of peak exercise):** patients with a downsloping slope of 2 have signs of an unhealthy heart therefore they more likely to have heart disease than people with an upsloping of 0 or a flat slope A value of 1: minimal change (typical healthy heart)).
- **ca (number of blood vessels (0-3)):** the more blood flow the better heart, so people with a vessel number ca equal to 0 are more likely to have heart disease.
- **thal (a blood disorder called thalassemia):** patietns with a thal value = 2 are more likely to have heart disease.



EXPLORATORY DATA ANALYSIS

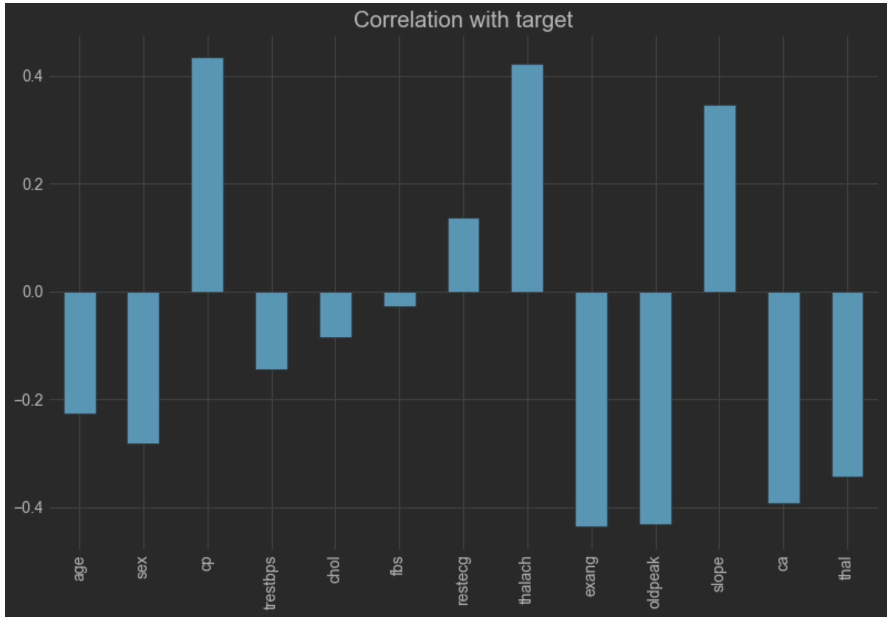
- Numerical data correlation with target

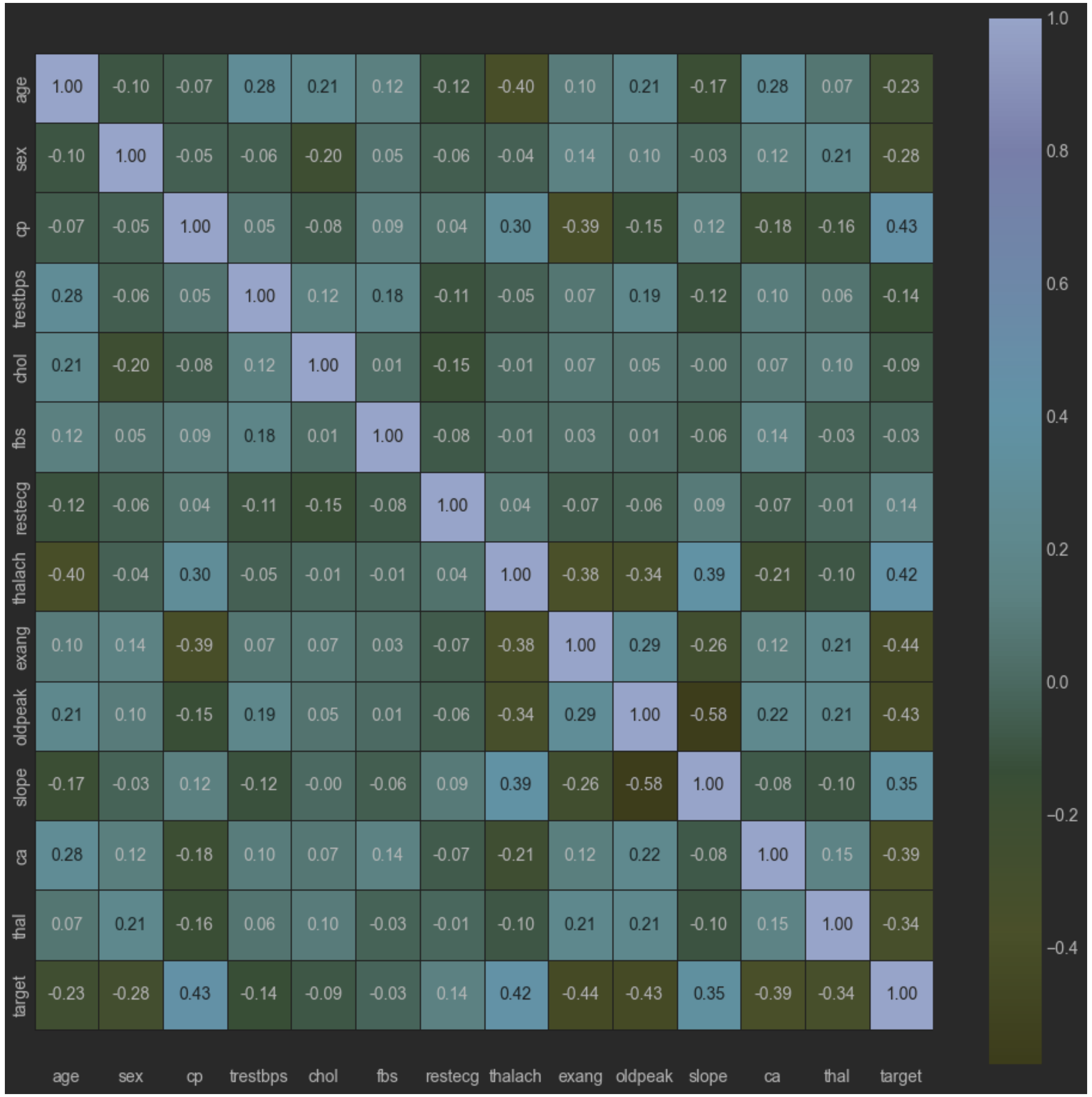
- **trestbps**: blood pressure higher than 130-140 mm Hg, causes concerns about having heart diseases.
- **chol**: cholesterol higher than 200 mg/dL, is a very dangerous indicator.
- **thalach**: People with a heart rate above 140 are more likely to have heart disease.

EXPLORATORY DATA ANALYSIS

- **Correlation between features**

- From the heat map we notice that fbs and chol are the least features impacting the target while the other features have high correlation with the target





EXPLORATORY DATA ANALYSIS

Feature Engineering

- Converting categorical data into numerical by splitting categorise into separate columns.

	age	trest	chol	tha	oldp	tar	sex	sex_1	cp_0	cp_1	cp_2	cp_3	fbs_0	fbs_1	restecg_0	restecg_1	restecg_2	exang_0	exang_1	slo
0	0.95	0.76	-0.26	0.02	1.09	1	0	1	0	0	0	1	0	1	1	0	0	1	0	
1	-1.92	-0.09	0.07	1.63	2.12	1	0	1	0	0	1	0	1	0	0	1	0	1	0	
2	-1.47	-0.09	-0.82	0.98	0.31	1	1	0	0	1	0	0	1	0	1	0	0	1	0	
3	0.18	-0.66	-0.20	1.24	-0.21	1	0	1	0	1	0	0	1	0	0	1	0	1	0	
4	0.29	-0.66	2.08	0.58	-0.38	1	1	0	1	0	0	0	1	0	0	1	0	0	1	
5	0.29	0.48	-1.05	-0.07	-0.55	1	0	1	1	0	0	0	1	0	0	1	0	1	0	
6	0.18	0.48	0.92	0.15	0.22	1	1	0	0	1	0	0	1	0	1	0	0	1	0	
7	-1.14	-0.66	0.32	1.02	-0.90	1	0	1	0	1	0	0	1	0	0	1	0	1	0	
8	-0.26	2.31	-0.91	0.54	-0.47	1	0	1	0	0	1	0	0	1	0	1	0	1	0	
9	0.29	1.05	-1.51	1.06	0.48	1	0	1	0	0	1	0	1	0	0	1	0	1	0	
10	-0.04	0.48	-0.14	0.45	0.14	1	0	1	1	0	0	0	1	0	0	1	0	1	0	

MACHINE LEARNING ANALYSIS

- Logistic Regression Model

Model Features and Parameters:

- Model = Logistic Regression()
- Solver = liblinear

	0	1	accuracy	macro avg	weighted avg
precision	0.78	0.80	0.79	0.79	0.79
recall	0.76	0.82	0.79	0.79	0.79
f1-score	0.77	0.81	0.79	0.79	0.79
support	41.00	50.00	0.79	91.00	91.00

MACHINE LEARNING ANALYSIS

- Logistic Regression with penalty = L1

Model Features and Parameters:

- Model = Logistic RegressionCV()
- Cs = 10
- cv: 4
- penalty = l1
- solver = liblinear

	0	1	accuracy	macro avg	weighted avg
precision	0.81	0.78	0.79	0.79	0.79
recall	0.71	0.86	0.79	0.78	0.79
f1-score	0.75	0.82	0.79	0.79	0.79
support	41.00	50.00	0.79	91.00	91.00

MACHINE LEARNING ANALYSIS

- Logistic Regression with penalty = L2

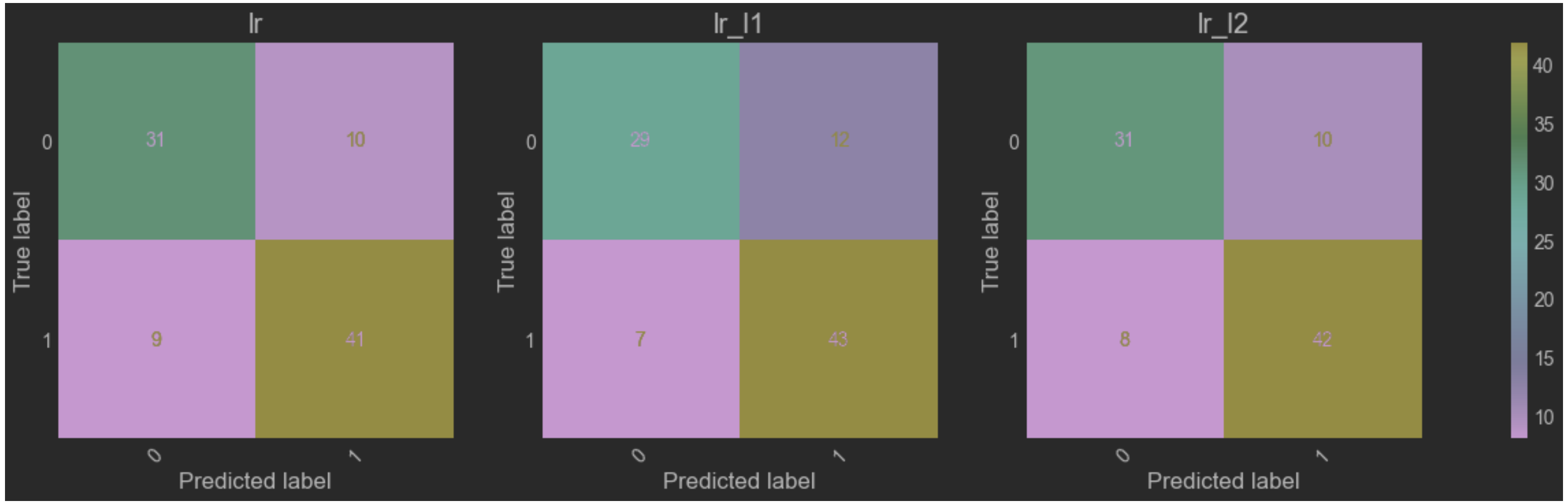
Model Features and Parameters:

- Model = Logistic RegressionCV()
- Cs = 10
- cv: 4
- penalty = l2
- solver = liblinear

	0	1	accuracy	macro avg	weighted avg
precision	0.79	0.81	0.80	0.80	0.80
recall	0.76	0.84	0.80	0.80	0.80
f1-score	0.77	0.82	0.80	0.80	0.80
support	41.00	50.00	0.80	91.00	91.00

MACHINE LEARNING ANALYSIS

- Logistic Regression models comparison



The best model in terms of prediction performance is Logistic Regression with penalty = 2

- Accuracy : 80%
- Precision : 80%
- Recall : 80%
- F1-score : 80%
- Support : 91%

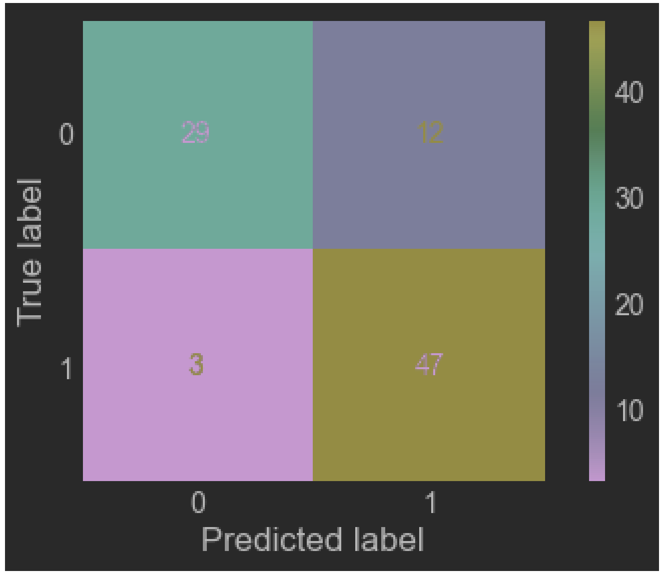
MACHINE LEARNING ANALYSIS

- K-Nearest Neighbors

Model Features and Parameters:

- Model = KNeighborsClassifier()
- n_neighbors=25
- weights=distance

	0	1	accuracy	macro avg	weighted avg
precision	0.91	0.80	0.84	0.85	0.85
recall	0.71	0.94	0.84	0.82	0.84
f1-score	0.79	0.86	0.84	0.83	0.83
support	41.00	50.00	0.84	91.00	91.00



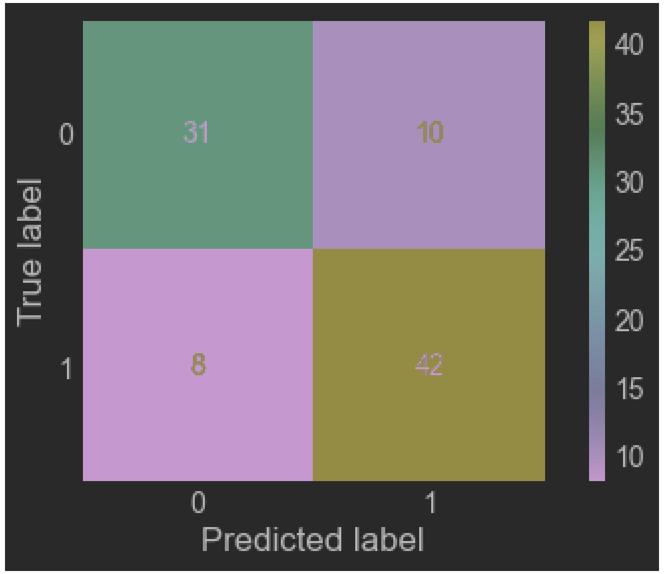
MACHINE LEARNING ANALYSIS

Support Vector Machine

Model Features and Parameters:

- Model = svc()
- Kernel: rbf

	0	1	accuracy	macro avg	weighted avg
precision	0.79	0.81	0.80	0.80	0.80
recall	0.76	0.84	0.80	0.80	0.80
f1-score	0.77	0.82	0.80	0.80	0.80
support	41.00	50.00	0.80	91.00	91.00

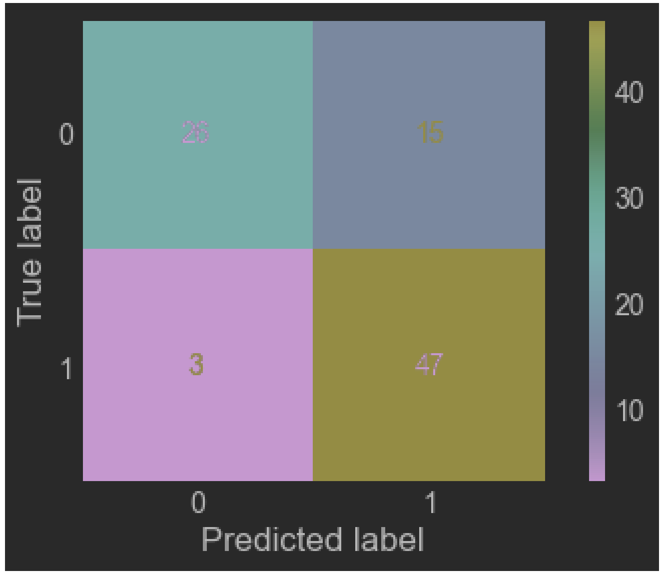


MACHINE LEARNING ANALYSIS

- XGBoost

- Model Features and Parameters:
- Model = xgb.XGBClassifer()
- Objective = binary:logistic

	0	1	accuracy	macro avg	weighted avg
precision	0.90	0.76	0.80	0.83	0.82
recall	0.63	0.94	0.80	0.79	0.80
f1-score	0.74	0.84	0.80	0.79	0.80
support	41.00	50.00	0.80	91.00	91.00



MACHINE LEARNING ANALYSIS

- Models Comparison

As shown in the previous analysis, all models gave very good predictions and these results are very close, but in the end to choose the best model of dataset that has the best results.

Here is the order according to the best four models:

1. KNN
2. XGBoost
3. Logistic Regression with L2
4. Support Vector Machine

	RMSE	R2	RMSE-SGD	R2-SGD
Linear	4496.560111	0.862103	4531.504262	0.859951
Lasso	4496.577652	0.862102	4570.227510	0.857548
Ridge	4494.682980	0.862218	4512.691171	0.861112
ElasticNet	4494.417701	0.862234	4528.496874	0.860137

ANALYSIS NEXT STEPS

- Models Flaws and Strength and further suggestions

- From a simplicity point of view, logistic regression yields high predictive results and at the same time is the simplest and fastest model in terms of parameters and training, but looking at other models like KNN, it provides the best results.

- However, it is time consuming from the perspective of the prediction process, as the distances between all the points in the dataset must be calculated to classify the individual points. XGBoost also performed very well, but unlike KNN, it uses a grid search technique to find the best parameters, which takes a long time in the training process. So, in the end, if you have a larger dataset, there is a trade-off. The performance of such models is high, but the training process is time consuming.