FINAL PROJECT - DS48

MAXIMILIAN KLIMKO

09.11.2021

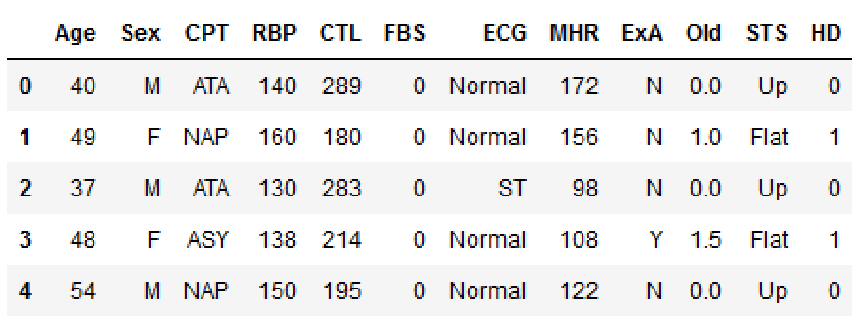
DATASET – HEART DISEASE

- 918 observations
- 5 different countries
- Sourced from University hospitals and Medical Institutes
- Taken from Kaggle
- Originally 5 datasets

https://www.kaggle.com/fedesoriano/heart-failure-prediction

WHAT DOES THE DATA SHOW?

- - Age
- - Sex
- - CPT: Angina Type
- - RBP: Resting Blood Pressure
- - CTL: Serum Cholesterol
- - FBS: Fasted Blood Sugar
- - ECG: Electrocardiogram Results
- - MHR: Max Heart Rate
- - ExA: Exercise-induced Angina
- - Old: ST [Numeric value measured in depression, implies restriction of bloodflow to tissue]
- - STS: ST segment slope during peak exercise



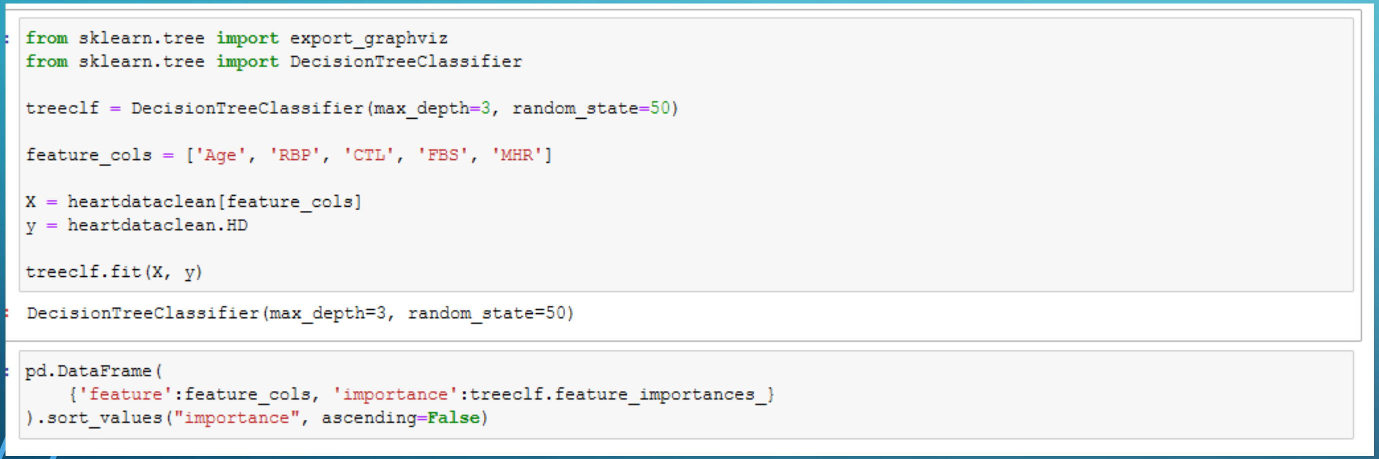
	Age	Sex	CPT	RBP	CTL	FBS	ECG	MHR	ExA	Old	STS	HD
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

Which of these can be used to predict whether a person will suffer heart failure?

FEATURE IMPORTANCE

- Simple decision tree

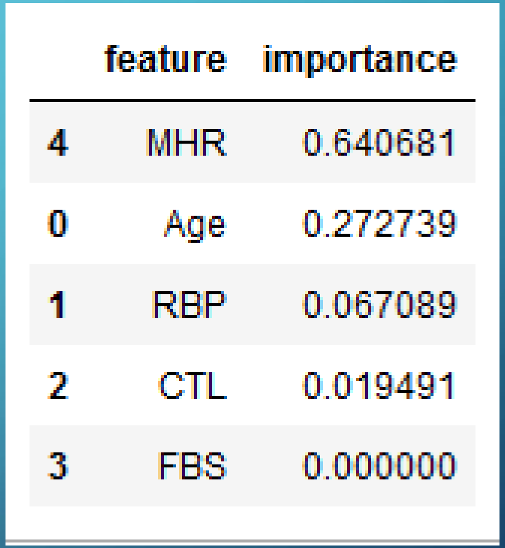
- Assessment of feature importance



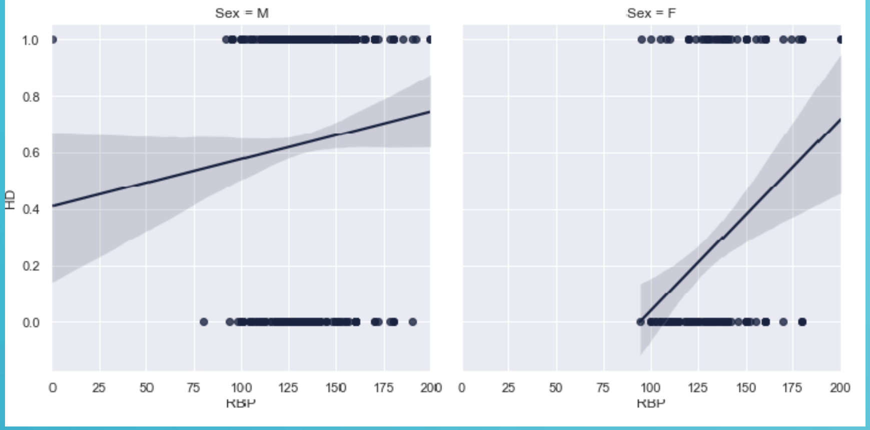
```
from sklearn.tree import export_graphviz
from sklearn.tree import DecisionTreeClassifier

treeclf = DecisionTreeClassifier(max_depth=3, random_state=50)

feature_cols = ['Age', 'RBP', 'CTL', 'FBS', 'MHR']

X = heartdataclean[feature_cols]
y = heartdataclean.HD

treeclf.fit(X, y)
```

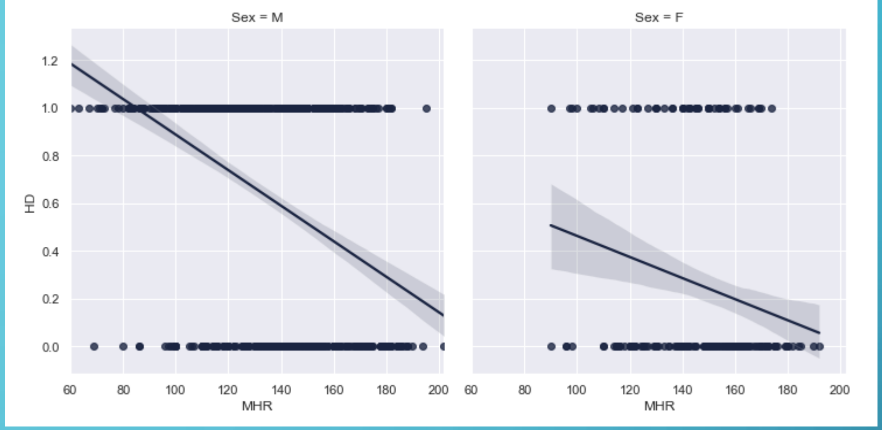
```
DecisionTreeClassifier(max_depth=3, random_state=50)
```

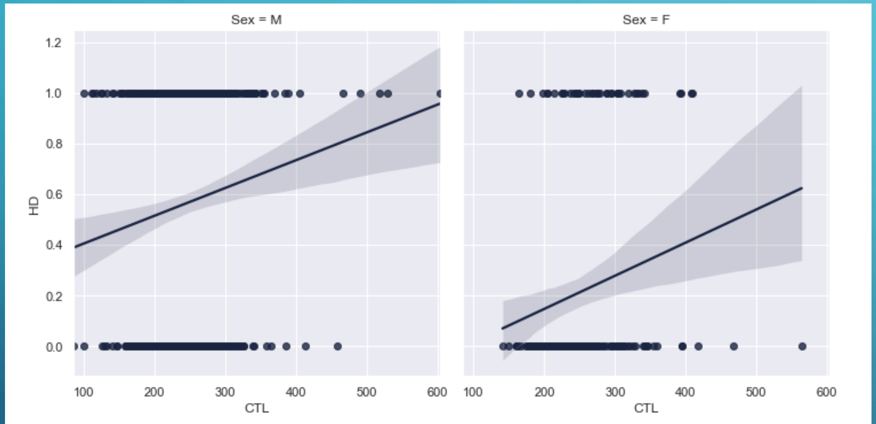
```
pd.DataFrame(
    {'feature':feature_cols, 'importance':treeclf.feature_importances_}
).sort_values("importance", ascending=False)
```

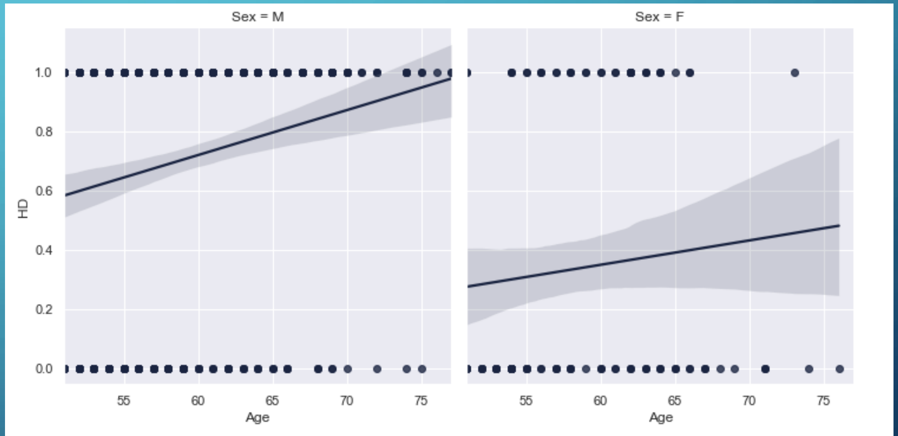


	feature	importance
4	MHR	0.640681
0	Age	0.272739
1	RBP	0.067089
2	CTL	0.019491
3	FBS	0.000000


Sex = M
Sex = F
HD
1.0
0.8
0.6
0.4
0.2
0.0
0
25
50
75
100
125
150
175
200
RBP
RBP


Sex = M
Sex = F
HD
1.2
1.0
0.8
0.6
0.4
0.2
0.0
60
80
100
120
140
160
180
200
MHR
MHR


Sex = M
Sex = F
HD
1.2
1.0
0.8
0.6
0.4
0.2
0.0
100
200
300
400
500
600
CTL
CTL

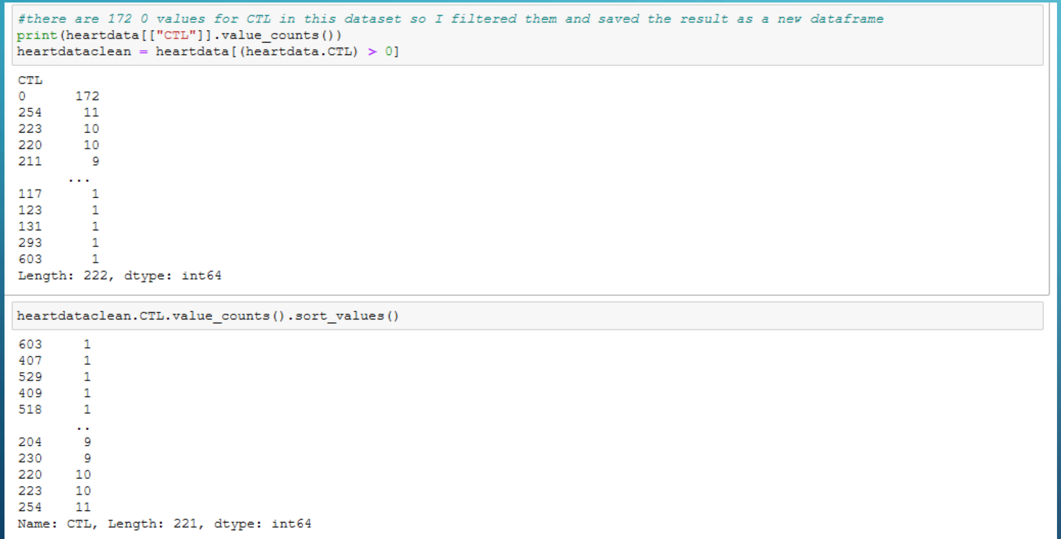

Sex = M
Sex = F
HD
1.0
0.8
0.6
0.4
0.2
0.0
55
60
65
70
75
Age
Age

LOGISTIC REGRESSION

- Small dataset

- Target is a class (0 or 1)

- CTL / MHR / Age

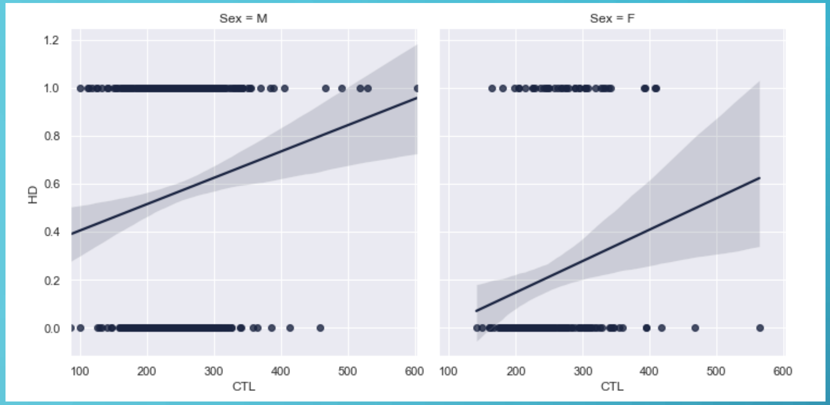


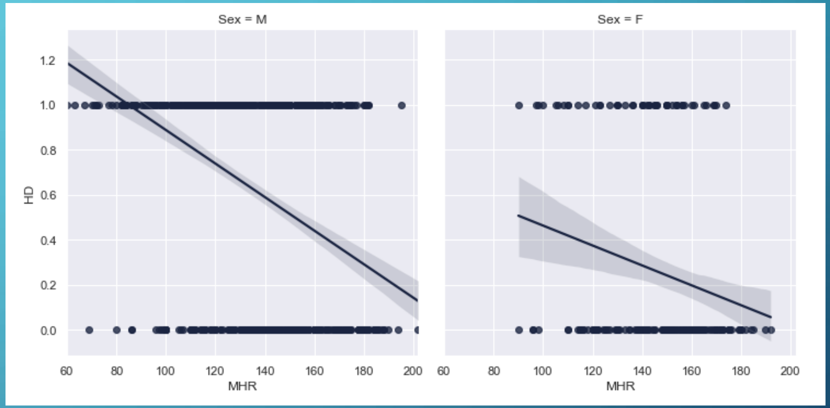
```
#there are 172 0 values for CTL in this dataset so I filtered them and saved the result as a new dataframe
print(heartdata[["CTL"]].value_counts())
heartdataclean = heartdata[(heartdata.CTL) > 0]
```

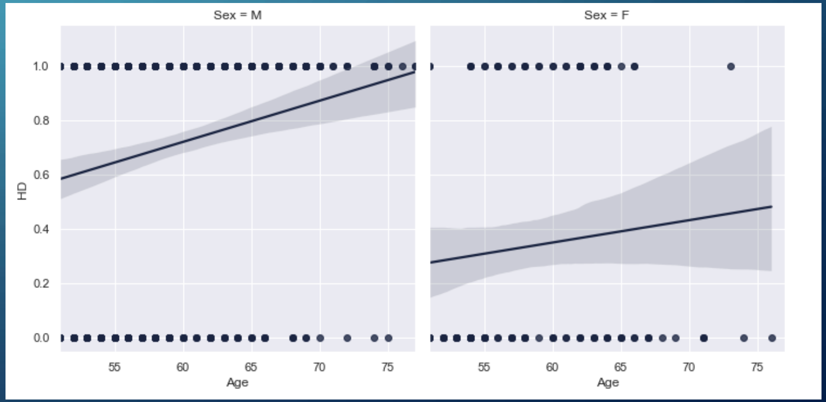
```
CTL
0      172
254     11
223     10
220     10
211      9
       ...
117      1
123      1
131      1
293      1
603      1
Length: 222, dtype: int64
```

```
heartdataclean.CTL.value_counts().sort_values()
```

```
603     1
407     1
529     1
409     1
518     1
       ..
204     9
230     9
220    10
223    10
254    11
Name: CTL, Length: 221, dtype: int64
```







CONFUSION MATRIX

TN: 63	FP: 23
FN: 40	TP: 61

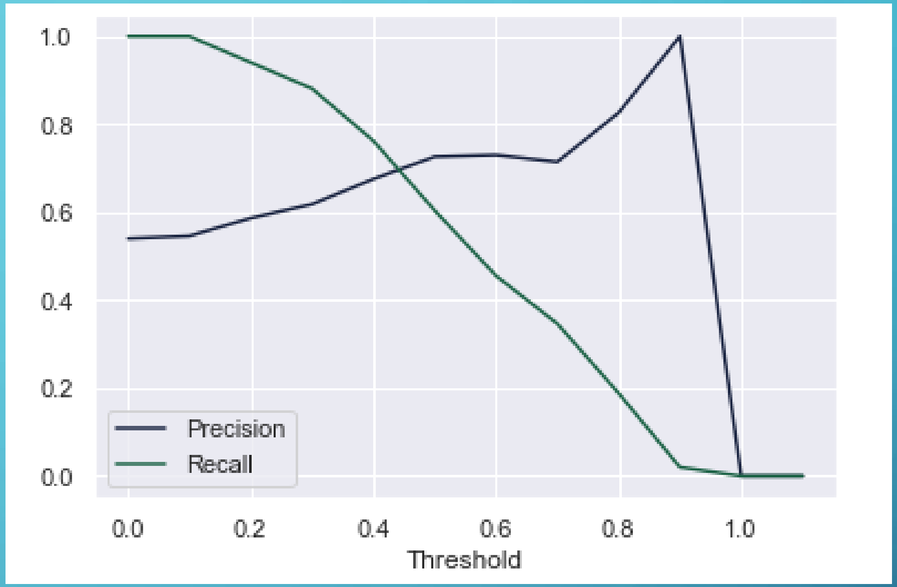
Accuracy – 0.69

Precision – 0.72

Recall – 0.60

False Positive rate – 0.25

F1 – 0.65



Lowering the threshold to 0.4 or even 0.3 would be justifiable as higher recall would mean less FN

PERSONAL TAKEAWAYS

- Learned about heart disease predictors

- Could have picked a larger dataset

- Or a different problem entirely

- I realized there is much more to learn about ML

- And that I wish to continue studying it