Supervised Machine Learning

Supervised Machine Learning Project #5

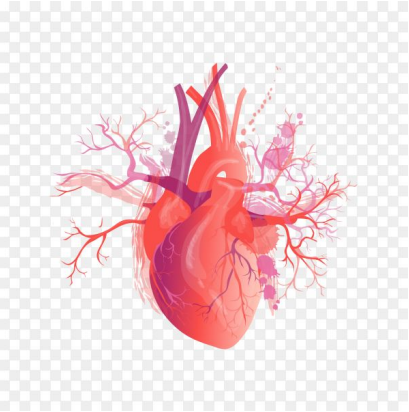
Oscar Lidheim

Question and Data

Can we predict Heart Failures accurately with classification machine learning?

How can we minimize the risk that we miss potential high risk heart attack patients (recall)?

How do we reduce costs of early warning (reduce features needed)?



Dataset:

299 Patients with 12 Features (Pakistan)

Lifestyle Habits

Blood measurements

Time until follow-up

Features and Target Variable

12 Features

Classification Problem (dead or not)

Good quality data, small sample

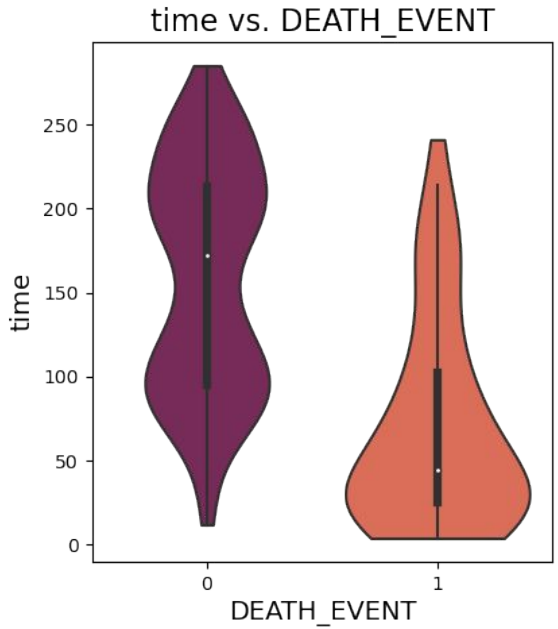
80/20% Train/Test Split

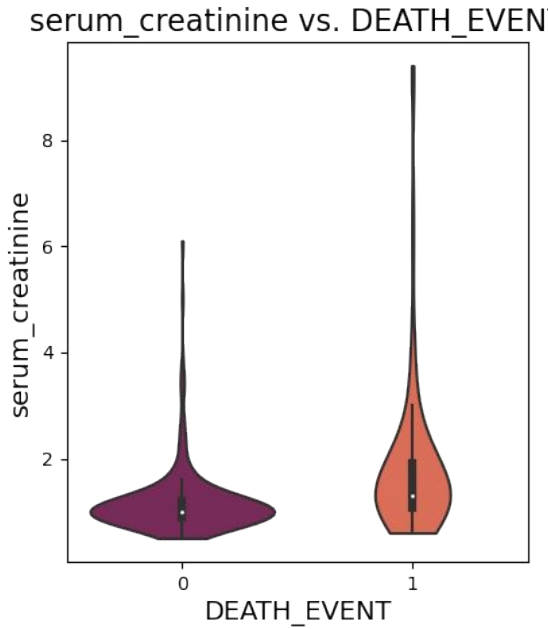
Variable Name	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer	Age	age of the patient	years	no
anaemia	Feature	Binary		decrease of red blood cells or hemoglobin		no
creatinine_phosphokinase	Feature	Integer		level of the CPK enzyme in the blood	mcg/L	no
diabetes	Feature	Binary		if the patient has diabetes		no
ejection_fraction	Feature	Integer		percentage of blood leaving the heart at each contraction	%	no
high_blood_pressure	Feature	Binary		if the patient has hypertension		no
platelets	Feature	Continuous		platelets in the blood	kiloplatelets/mL	no
serum_creatinine	Feature	Continuous		level of serum creatinine in the blood	mg/dL	no
serum_sodium	Feature	Integer		level of serum sodium in the blood	mEq/L	no
sex	Feature	Binary	Sex	woman or man		no
smoking	Feature	Binary		if the patient smokes or not		no
time	Feature	Integer		follow-up period	days	no
death_event	**Target**	**Binary**		**if the patient died during the follow-up period**		**no**

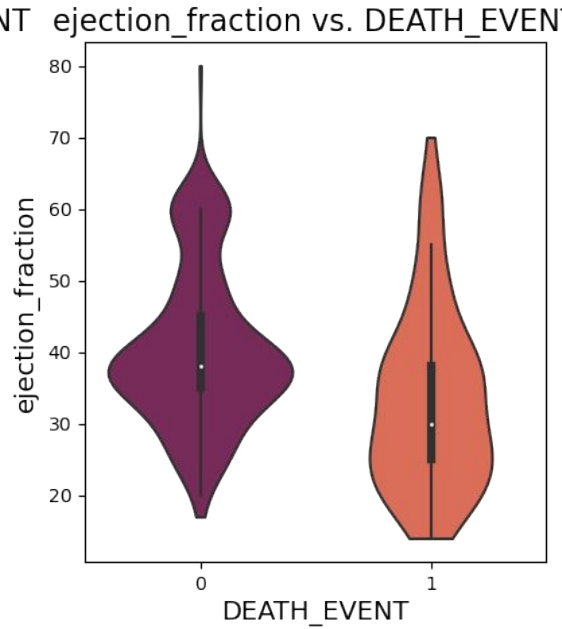
EDA - Evaluating Key Features

Correlation and Violin Plots

Three key features - No symmetry







Test 5 Models Standard Settings

5 Models evaluated

LogReg, DecTree, RandomForestC, SVC, KNC

Decision Tree Models best

Red Flag!

Both Decision Tree and RF had scores of 100% on Accuracy/Precision/Recall on full training data (not the folds)

```
Model: DecisionTreeClassifier()
Best Hyperparameters: {}
Train score for DecisionTreeClassifier() for recall 0.6209523809523809
For the full training data the scores are Accuracy: 1.0, Precision: 1.0, Recall: 1.0
```

```
Model: RandomForestClassifier()
Best Hyperparameters: {}
Train score for RandomForestClassifier() for recall 0.7885714285714286
For the full training data the scores are Accuracy: 1.0, Precision: 1.0, Recall: 1.0
```

Tuning Random Forest Model

Best Parameters:

- n_estimators: 10
- max_depth: 20
- min_samples_leaf: 4
- min_samples_split: 2
- CV = 10 (24 persons in each split)

Performance Metrics:

- Accuracy: 0.9497907949790795,
- Precision: 0.9538461538461539,
- Recall:
- 0.8732394366197183,
- F1 Score: 0.9117647058823529,
- AUC-ROC Score: 0.9882629107981221

Grid Parameters:

```
n_estimators': [10, 50, 100, 200],
max_depth': [None, 2, 5, 10, 20],
min_samples_split': [2, 3, 5, 10],
min_samples_leaf': [1, 2, 4, 10],
```

	Feature	Importance
1	time	0.415919
2	serum_creatinine	0.158207
3	ejection_fraction	0.117203
4	age	0.077161
5	creatinine_phosphokinase	0.075675
6	platelets	0.068311
7	serum_sodium	0.052898
8	anaemia	0.009492
9	smoking	0.007286
10	sex	0.006243
11	high_blood_pressure	0.006001
12	diabetes	0.005602

Final Evaluation on Test Data

The model score on test data are:

Accuracy: 0.7166666666666667,

Precision: 0.75,

Recall: 0.48,

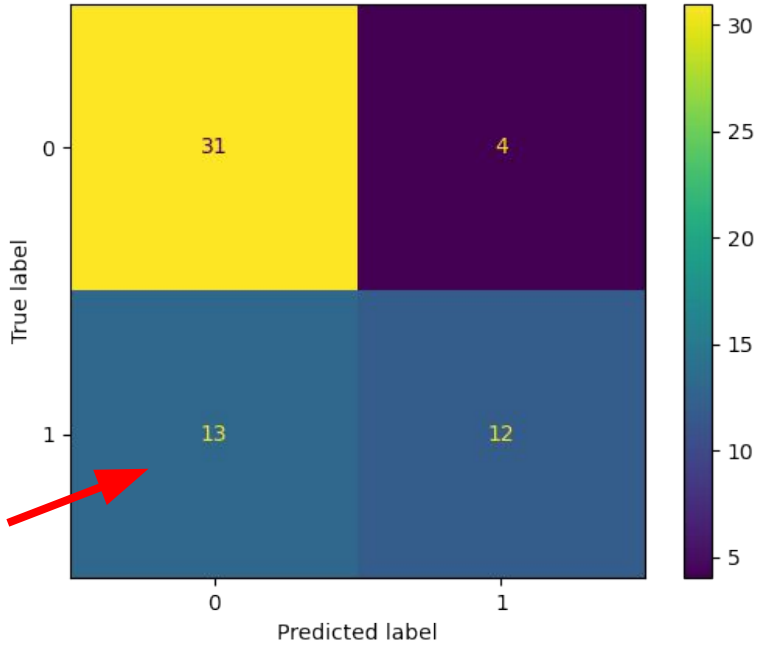
F1 Score: 0.5853658536585366,

AUC-ROC Score: 0.8228571428571428

Terrible recall Score!

Model is overfitted?

Sample size is too low and further fragmented by folds



Initial Questions

Can we predict Heart Failures accurately with classification machine learning?	**Not great**
How can we minimize the risk that we miss potential high risk heart attack patients (recall)?	**Recall score is weak**
How do we reduce costs of early warning (reduce features needed)?	**A few variables seems more important**

Key Learnings Model Testing

- Sample Size an issue
- Overfitting hard to see before testing
- Standard settings not always good

Key Learnings from Academic Study

- Used many models -> LogReg/RandomForest still best
- Manipulated test/train split to have same probability of death
- Did use many test score measures
- Tested suboptimal models to find most important features

Highest Results - Research Study 2020

Table 11 Survival prediction results including the follow-up time – mean of 100 executions

From: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone

Method	MCC	F_1 score	Accuracy	TP rate	TN rate	PR AUC	ROC AUC
Logistic regression	blue**+0.616***	blue0.719*	blue0.838*	blue0.785*	blue0.860*	blue0.617*	blue0.822*
(EF, SR, & FU)							
Logistic regression	**+0.607**	0.714	0.833	0.780	0.856	0.612	0.818
(all features)							

Thank You