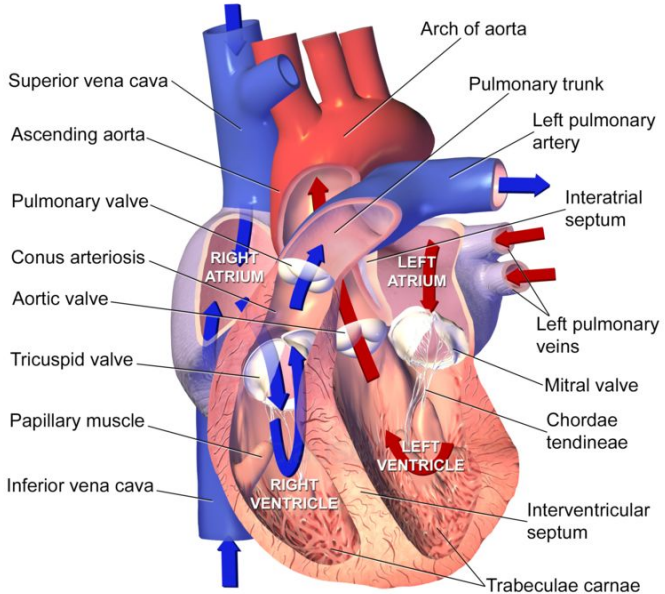
Heart Failure Prediction

Tawfik Abbas

What is heart failure?

- The heart’s function, simply put, is to circulate blood throughout the body
- Heart failure then, is a condition in which the heart struggles to perform this essential task



Sectional Anatomy of the Heart

Significance

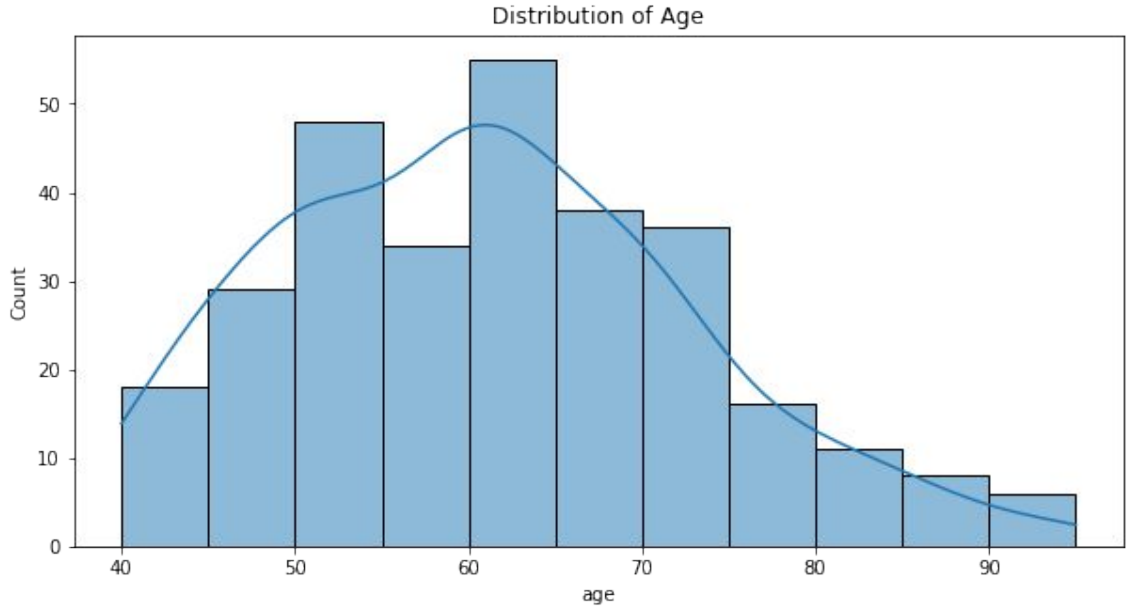
- More than 6 million adults in the United States alone have heart failure
- In 2012, heart failure cost the nation approximately $30.7 billion, including treatment costs and missed days of work
- As such, we may pose the following question: Given a patient’s collected medical data, are we able to predict the incidence of heart failure?

The Data

- Data was sourced from Kaggle
- Collected at the Faisalabad Institute of Cardiology and Allied hospital from April to December 2015
- Data shape: (299, 14)

age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
75	0	582	0	20	1	265000	1.9	130	1	0	4	1
55	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
65	0	146	0	20	0	162000	1.3	129	1	1	7	1
50	1	111	0	20	0	210000	1.9	137	1	0	7	1
65	1	160	1	20	0	327000	2.7	116	0	0	8	1
90	1	47	0	40	1	204000	2.1	132	1	1	8	1
75	1	246	0	15	0	127000	1.2	137	1	0	10	1
60	1	315	1	60	0	454000	1.1	131	1	1	10	1
65	0	157	0	65	0	263358.03	1.5	138	0	0	10	1
80	1	123	0	35	1	388000	9.4	133	1	1	10	1
75	1	81	0	38	1	368000	4	131	1	1	10	1
62	0	231	0	25	1	253000	0.9	140	1	1	10	1
45	1	981	0	30	0	136000	1.1	137	1	0	11	1
50	1	168	0	38	1	276000	1.1	137	1	0	11	1
49	1	80	0	30	1	427000	1	138	0	0	12	0
82	1	379	0	50	0	47000	1.3	136	1	0	13	1
87	1	149	0	38	0	262000	0.9	140	1	0	14	1
45	0	582	0	14	0	166000	0.8	127	1	0	14	1

The Data II: Demographics




anaemia
0
56.9%
43.1%
1
diabetes
0
58.2%
41.8%
1
high_blood_pressure
0
64.9%
35.1%
1
sex
1
64.9%
35.1%
0
smoking
0
67.9%
32.1%
1

Data Processing

- When exploring the dataset, we found that there was no missing data
- There were outliers/extreme values that required attention
- Class imbalance present in the target variable

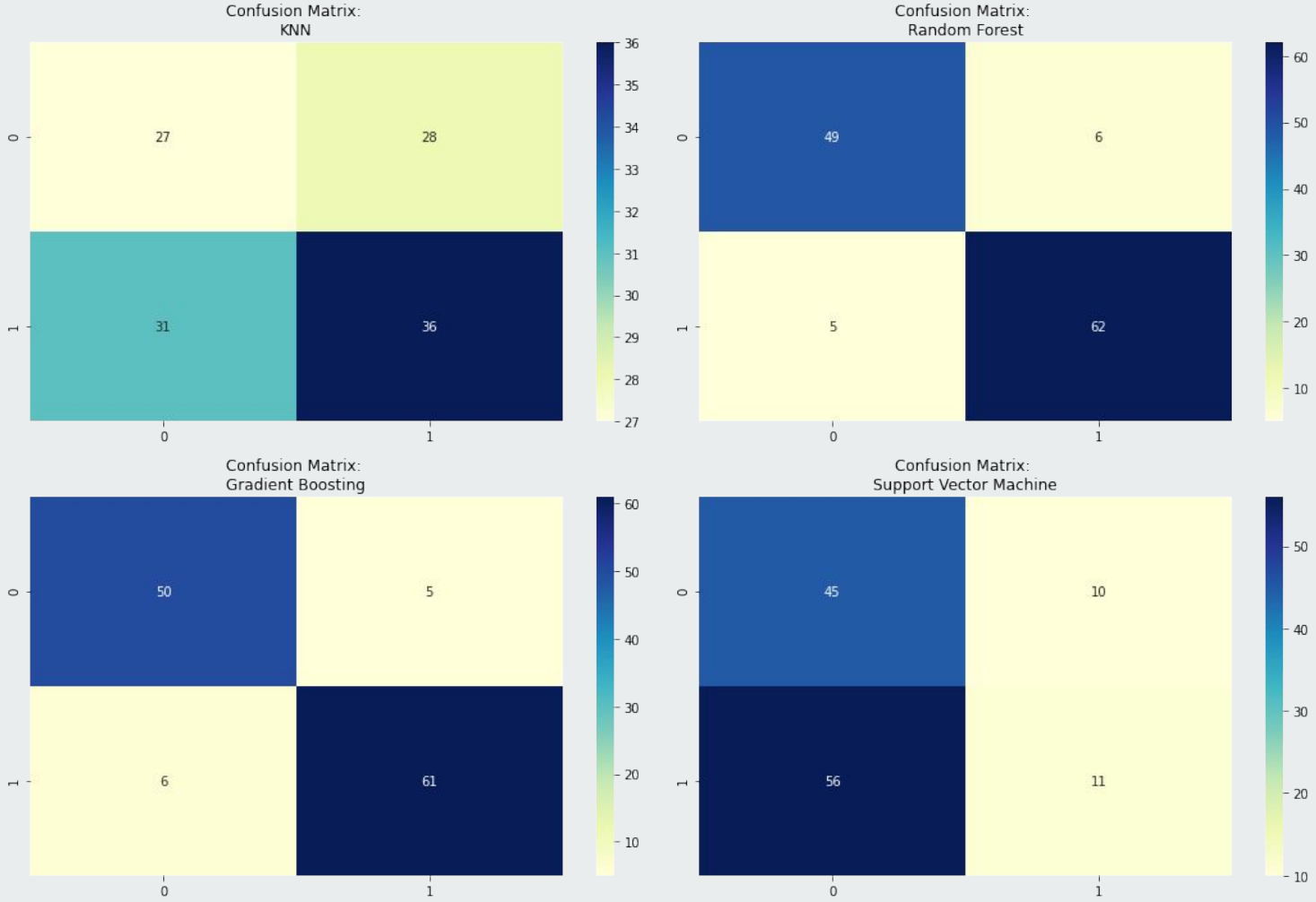
Model Building

- When deciding which models to train, the most significant factor was the nature of the problem, i.e. classification, or regression.
- Prior to training and fitting baseline models, the dataset was split into a training set and test set
- As the outcome variable is discrete, we selected the following classifiers to build baseline models:
    - K Nearest Neighbors
    - Random Forest
    - Support Vector Machine
    - Gradient Booster

Model Selection

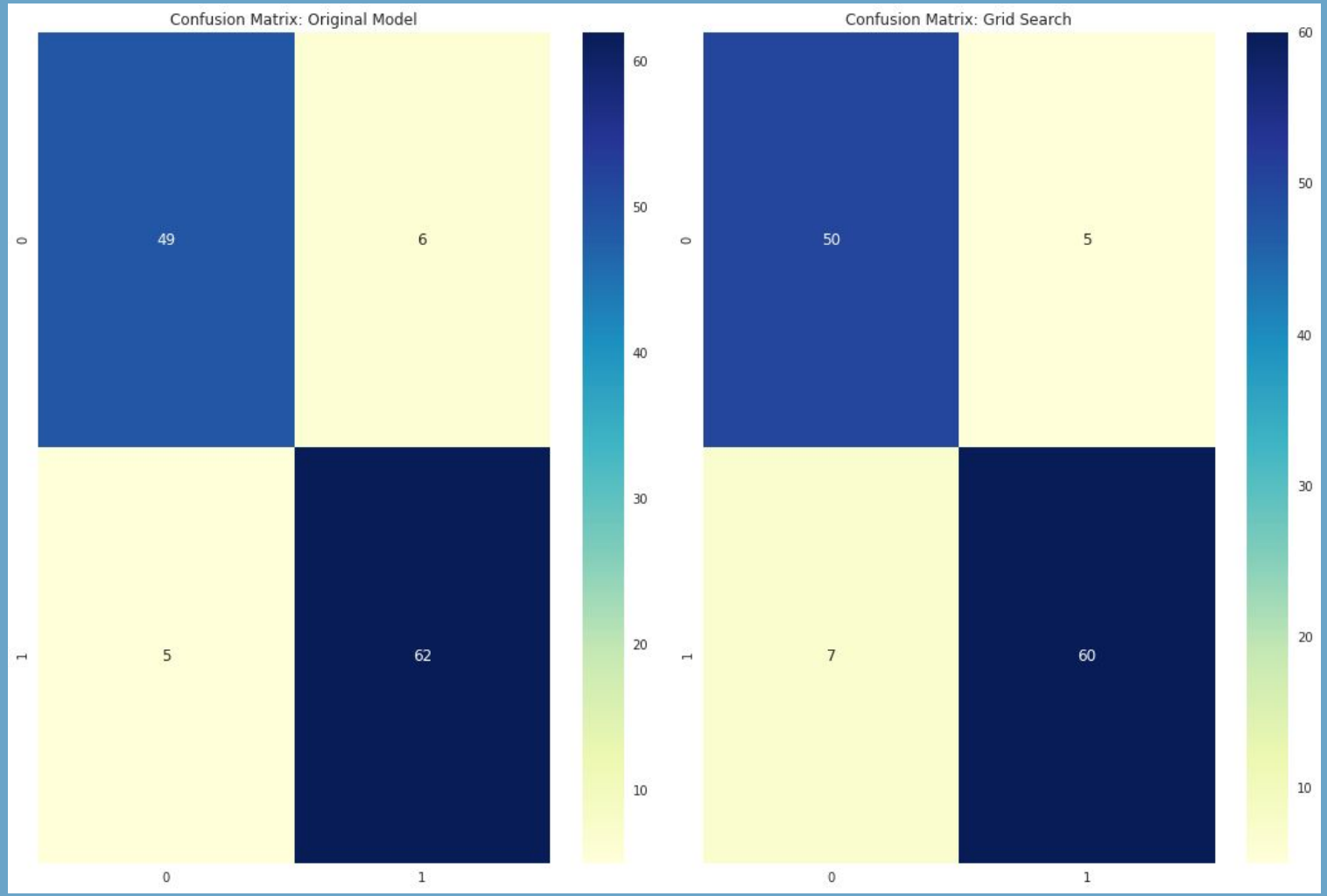
- After fitting the models to the training set and making predictions with each model, we utilized the following metrics for model comparison and selection:
  - Accuracy score
  - Precision score
  - Recall score
- In addition to the above metrics, we utilized confusion matrices

	Accuracy	Precision	Recall
K Nearest Neighbors	0.5164	0.5625	0.5373
Random Forest	0.9262	0.9394	0.9254
Support Vector Machine	0.459	0.5238	0.1642
Gradient Booster	0.9098	0.9242	0.9104


Confusion Matrix:
KNN
27
28
31
36
0
1
Confusion Matrix:
Random Forest
49
6
5
62
Confusion Matrix:
Gradient Boosting
50
5
6
61
Confusion Matrix:
Support Vector Machine
45
10
56
11

Results

- Following model selection, we tuned the hyperparameters of the Random Forest model further, making use of the Randomized Search and Grid Search methods
- Another confusion matrix was computed to compare the original baseline model to the GridSearch fit


Confusion Matrix: Original Model
0
1
49
6
5
62
0
1
60
50
40
30
20
10
Confusion Matrix: Grid Search
0
1
50
5
7
60
0
1
60
50
40
30
20
10

Potential Improvements

- Features within the dataset possess different scales
- The time variable within the dataset is the follow-up time with a patient

Future Goals

- Rescale the features prior to training the model
- Run the models without the time variable to see the effect on performance metrics
- Train and run other models (XGBoost, QLattice, etc.)

Any Questions?

Thank You!