Thoracic Surgery for the Treatment of Lung Cancer

Using Machine Learning to Predict Patient Outcomes

By Robert Lewis

The Problem

- Lung Cancer is the second most common cancer among US adults
- It leads all cancers in total number of deaths (accounting for 25% of all cancer deaths)
- Thoracic surgery is often the best treatment option
- Factors that lead to a patient's success post-surgery are still not well understood

Our Goal

A man uses a Spirometer to collect FVC and FEV1 values

- Forced Vital Capacity (FVC) and Forced Expiratory Volume in 1 second (FEV-1) remain the ‘gold standard’ metrics
- Our aim is to provide physicians with a tool that will better evaluate a lung cancer patient’s candidacy for surgery.
- This would both decrease the overall mortality of such surgeries, and allow poor candidates to seek alternative treatments

Overview of the Raw Data

Attribute	Description	Values	Data Type
DGN	ICD-10 codes for primary and secondary as well multiple tumours if any	DGN3,DGN4,DGN5,DGN6,DGN8	Nominal
PRE4	Forced Vital Capacity - FVC	Numeric	Numeric
PRE5	Forced Expiratory Volume 1s	Numeric	Numeric
PRE6	Performance status - Zubrod scale	PRZ0, PRZ1, PRZ2	Nominal
PRE7	Pain before surgery	T,F	Binary
PRE8	Haemoptysis before surgery	T,F	Binary
PRE9	Dyspnoea before surgery	T,F	Binary
PRE10	Cough before surgery	T,F	Binary
PRE11	Weakness before surgery	T,F	Binary
PRE14	Size of the original tumour	OC11,OC12,OC13,OC14 (sm - lg)	Nominal
PRE17	Type 2 Diabetes Mellitus	T,F	Binary
PRE19	MI up to 6 months	T,F	Binary
PRE25	Peripheral arterial disease (PAD)	T,F	Binary
PRE30	Smoking	T,F	Binary
PRE32	Asthma	T,F	Binary
AGE	Age at surgery	Numeric	Numeric
Risk1Yr	1 year survival period	True, if died	Binary

Link to Dataset : UCI Machine Learning Repository

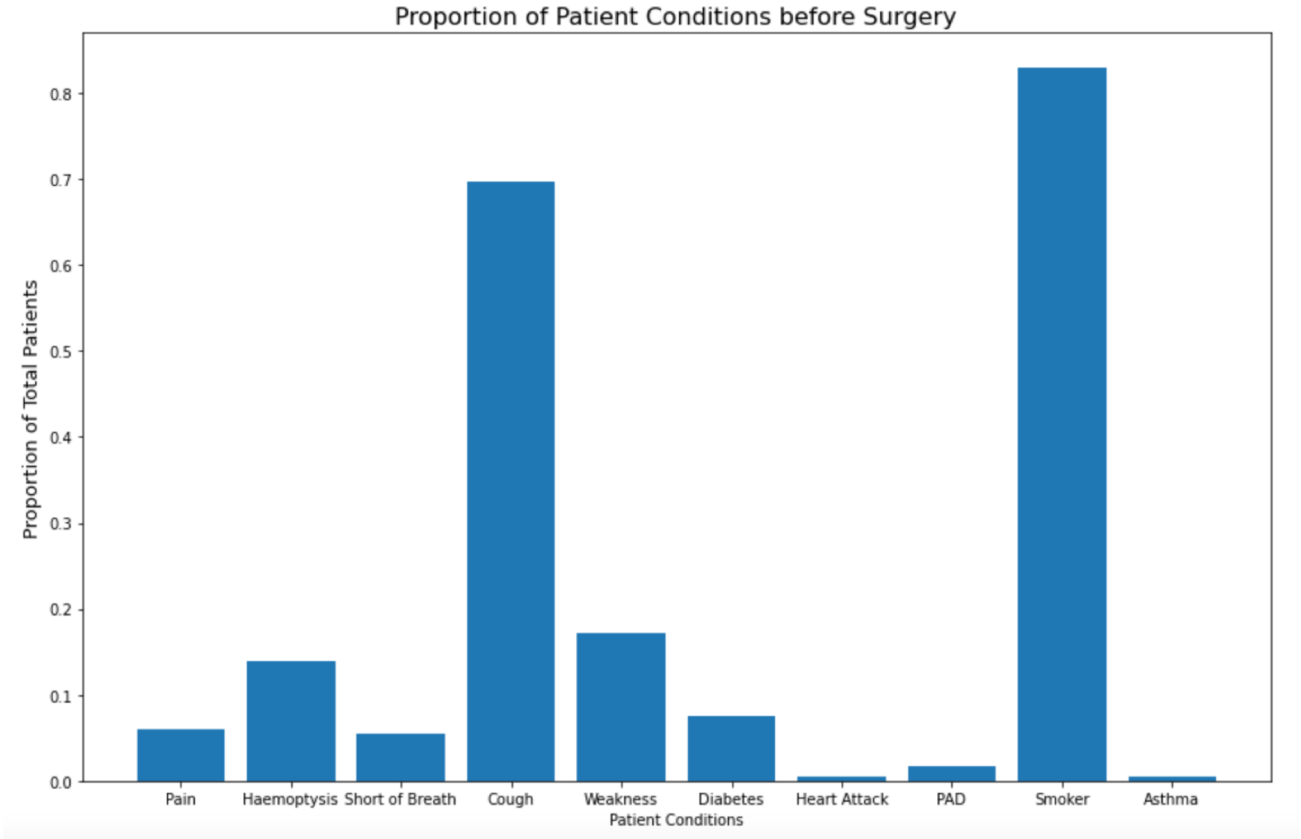
- 470 Total Observations
- 16 Predictors + Target Variable
- 15 Categorical Features
- 3 Numerical Features
- 400 Lived 70 Died

Exploratory Data Analysis:

How Are Patient Conditions Distributed?

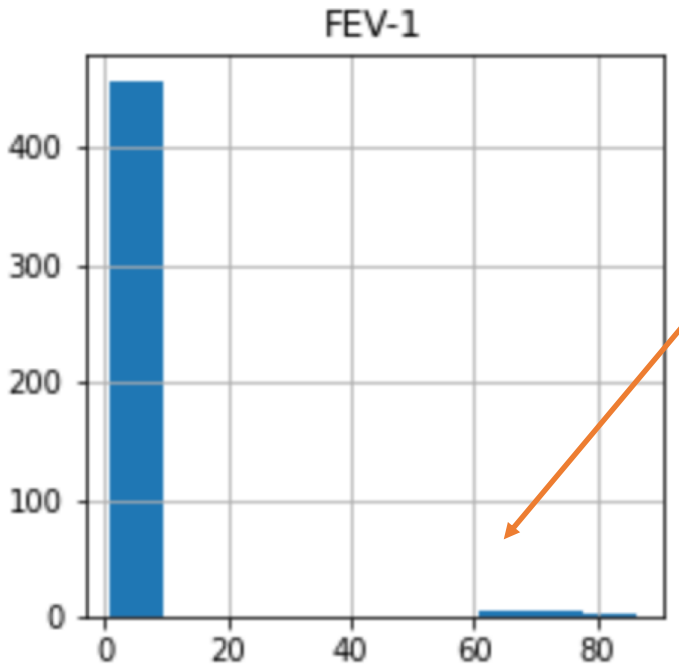
- An overwhelming majority of patients were Smokers and complained of a Persistent Cough

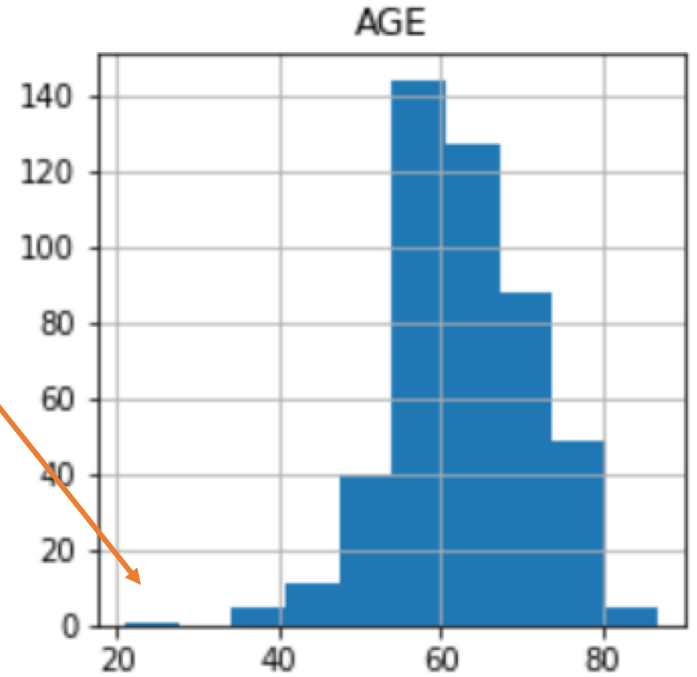
- Few reported a history of Asthma, Peripheral Artery Disease or a Heart Attack w/n last 6 mo.



Exploratory Data Analysis

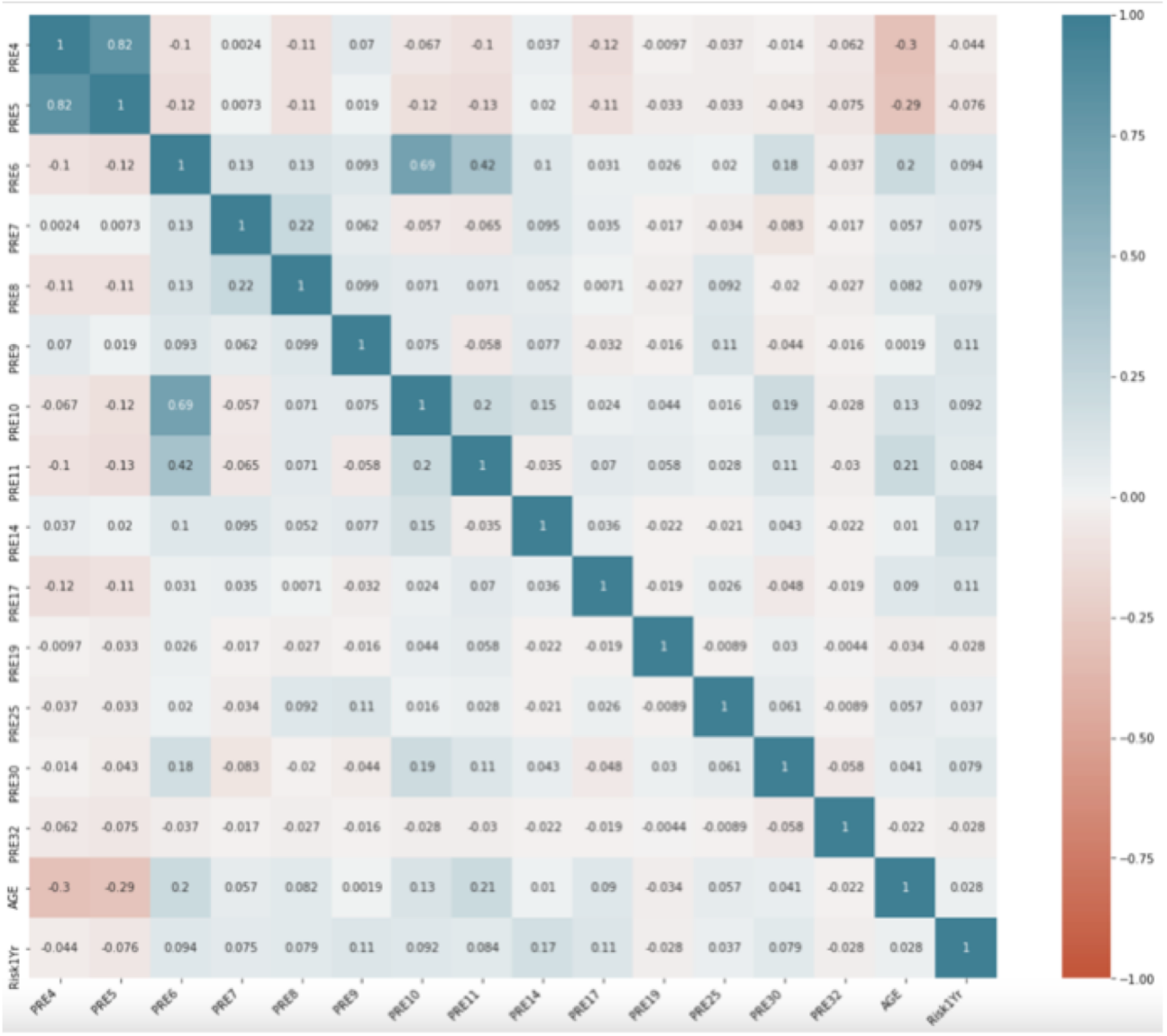
Of the 3 numerical features, 2 were found to contain outliers (Age & PRE5)





EDA Cont'd

- The figure to the right shows a heatmap of the correlations between all features
- Of note, high correlations (> 0.5) were found between FVC/FEV-1 ; Zubrod Score/Cough



Choosing an Appropriate Algorithm

Our model must meet 2 main criteria:

- ✓ Work for binary classification
- ✓ Handle continuous & categorical data types

3 Methods Were Tested

- Weighted Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

Weighted Logistic Regression

- Logistic Regression is a popular algorithm for binary classification that uses a Sigmoid cost function to map predicted values into probabilities between 0 and 1
- Weighting allows us to compensate for a low minority class in an imbalanced dataset

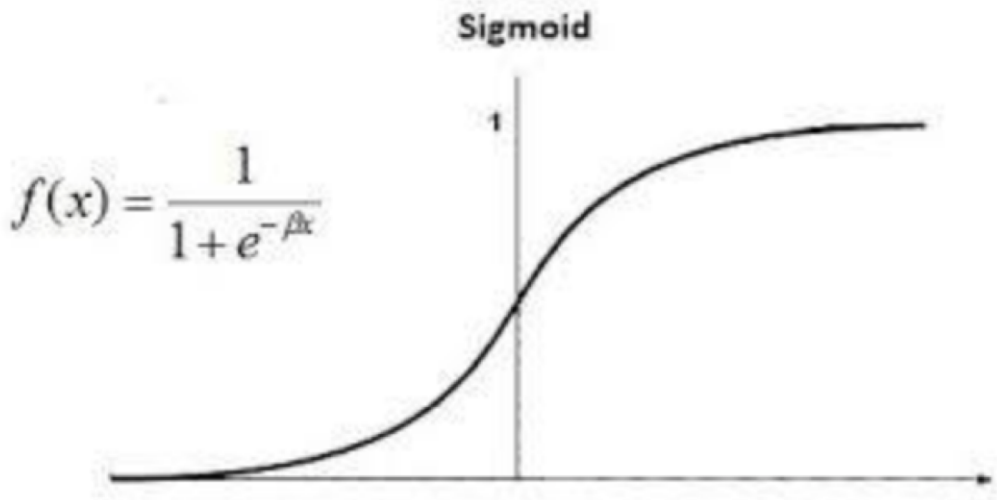


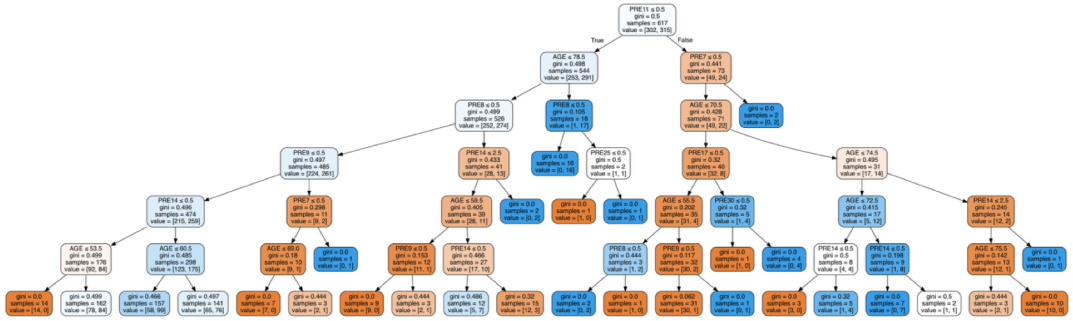
Image from GeeksForGeeks

OUTCOME

```
Accuracy Score: 0.7582417582417582
Confusion Matrix:
[[60 14]
 [ 8  9]]
Area Under Curve: 0.6701112877583465
Recall score: 0.5294117647058824
```

Decision Tree Classifier

- Decision trees apply a series of conditionals to split the data based on information gain
- Can use mixed data types
- Doesn't require scaling or normalization of features during pre-processing
- Easier to read than most algorithms



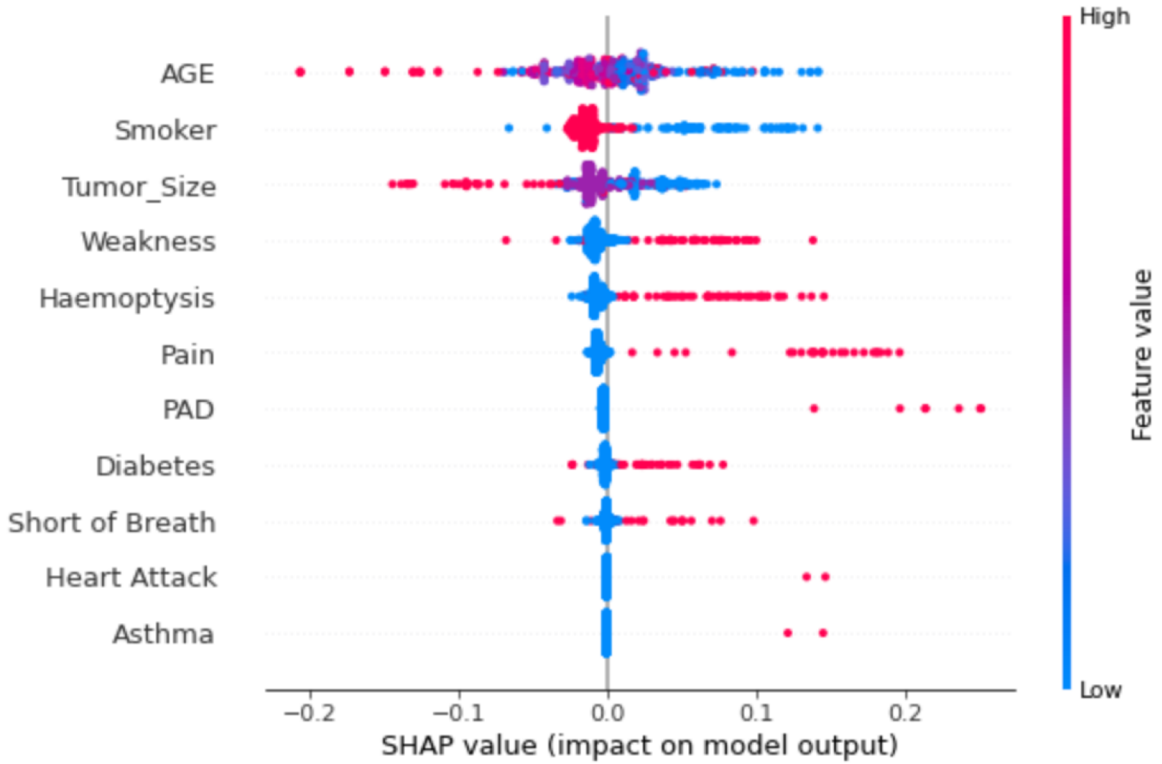
Decision Tree with max_depth = 6

<u>OUTCOME</u>

```
Gini impurity  model - max depth 6
Accuracy: 0.5741935483870968
Balanced accuracy: 0.5995137491616365
Precision score 0.5203252032520326
Recall score 0.9014084507042254
```

Random Forest Classifier

- An ensemble method of decision trees
- Builds multiple trees using a random subset of features
- Averages the predictions of all trees to make final prediction
- Many of the same advantages as Decision Trees



OUTCOME

```
Random Forest Model - max depth 6
Accuracy: 0.6580645161290323
Balanced accuracy: 0.6768947015425889
Precision score 0.5818181818181818
Recall score 0.9014084507042254
```

Conclusions

- Of the 3 models tested, Random Forest proved to give the best results with the highest level of Recall and improved Accuracy over the Decision Tree model
- Features that were important in deciding patient outcome included:
  - PRE8 → Coughing Blood
  - PRE11 → Weakness
  - PRE14 → Tumor Size
  - AGE
- Further hyperparameter optimization, in addition to testing other algorithms (such as SVM), could lead to more improvement and higher predictive capabilities