Thoracic Surgery

For Lung Cancer Patients

By: Seung Chi

Problem

- Patients who receive thoracic surgery for lung cancer do so with the expectation that their lives will be prolonged for a sufficient amount of time afterwards.
- The problem to solve is whether there is a way to determine postoperative 1 year survival of lung cancer patients utilizing the patient attributes in the data set.

Who benefits from answering this problem?

- Patients
- Families of Patients
- Physicians
- Hospitals
- Healthcare Organizations

Data Set

- Original from UCI Machine Learning Repository
  - Collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007-2011
  - 470 instances and no missing values
- This report consists of 454 patient data.
  - Excluding outliers from FEV1 and Age columns

Descriptions of Attributes (1)

Attribute	Description
Diagnosis	ICD-10 codes for primary and secondary as well multiple tumors if any
FVC	Amount of air which can be forcibly exhaled from the lungs after taking the deepest breath possible
FEV1	Volume that has been exhaled at the end of the first second of forced expiration
Performance	Performance status on Zubrod scale, Good (0) to Poor (2)
Pain	Pain, prior to surgery (T = 1, F = 0)
Haemoptysis	Coughing up blood, prior to surgery (T = 1, F = 0)
Dyspnoea	Difficult or labored breathing, prior to surgery (T = 1, F = 0)
Cough	Cough, prior to surgery (T = 1, F = 0)

Descriptions of Attributes (2)

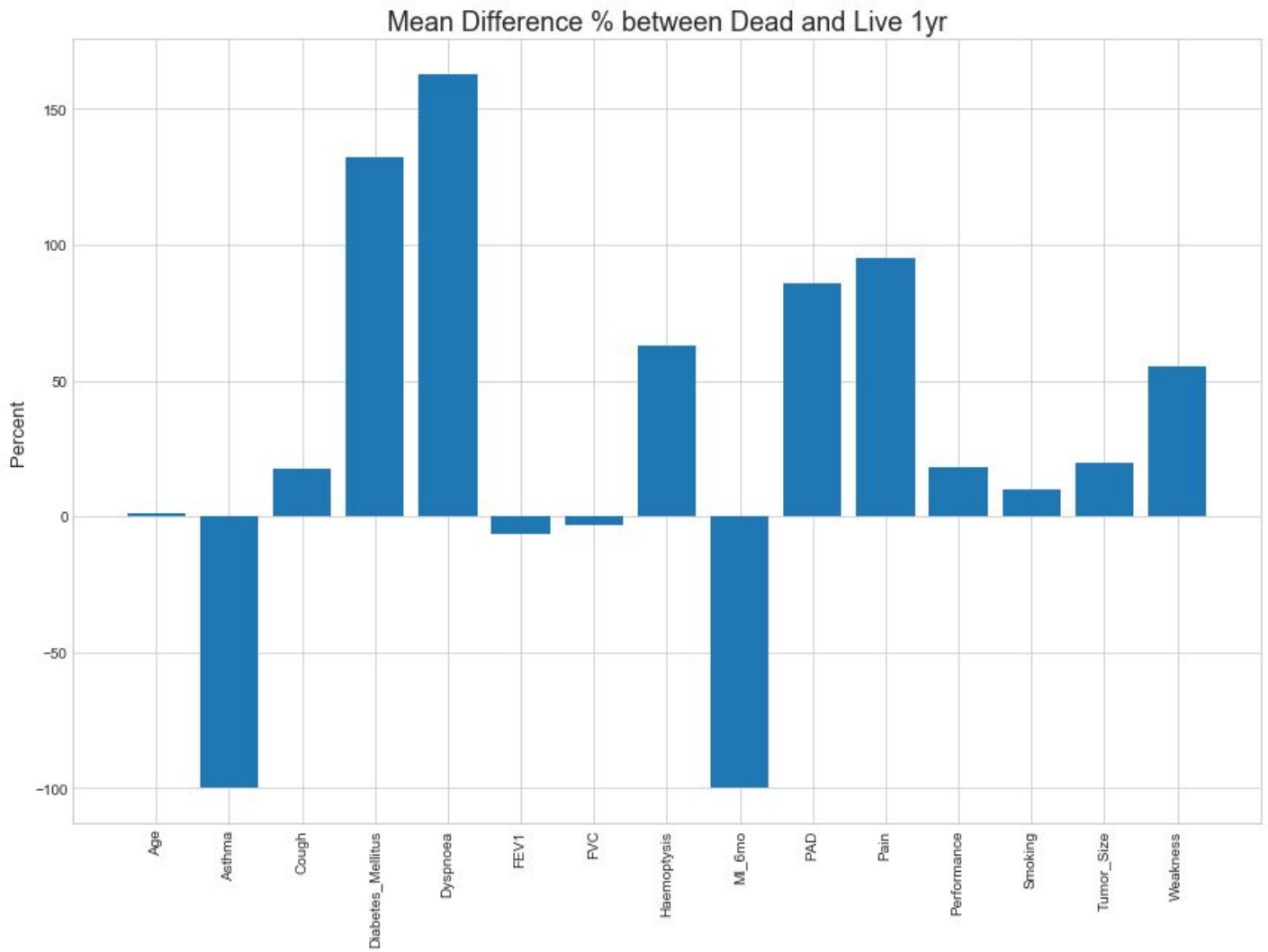
Attribute	Description
Weakness	Weakness, prior to surgery (T = 1, F = 0)
Tumor_Size	T in clinical TNM - size of the original tumor, 1 (smallest) to 4 (largest)
Diabetes_Mellitus	Type 2 diabetes mellitus (T = 1, F = 0)
MI_6mo	Myocardial Infarction (Heart Attack) up to 6 months prior (T = 1, F = 0)
PAD	Peripheral arterial diseases (T = 1, F = 0)
Smoking	Smoking (T = 1, F = 0)
Asthma	Asthma (T = 1, F = 0)
Age	Age at surgery
Death_1yr	1 year survival period - (T) value if died (T = 1, F = 0)

Difference between 1 year death and live patients

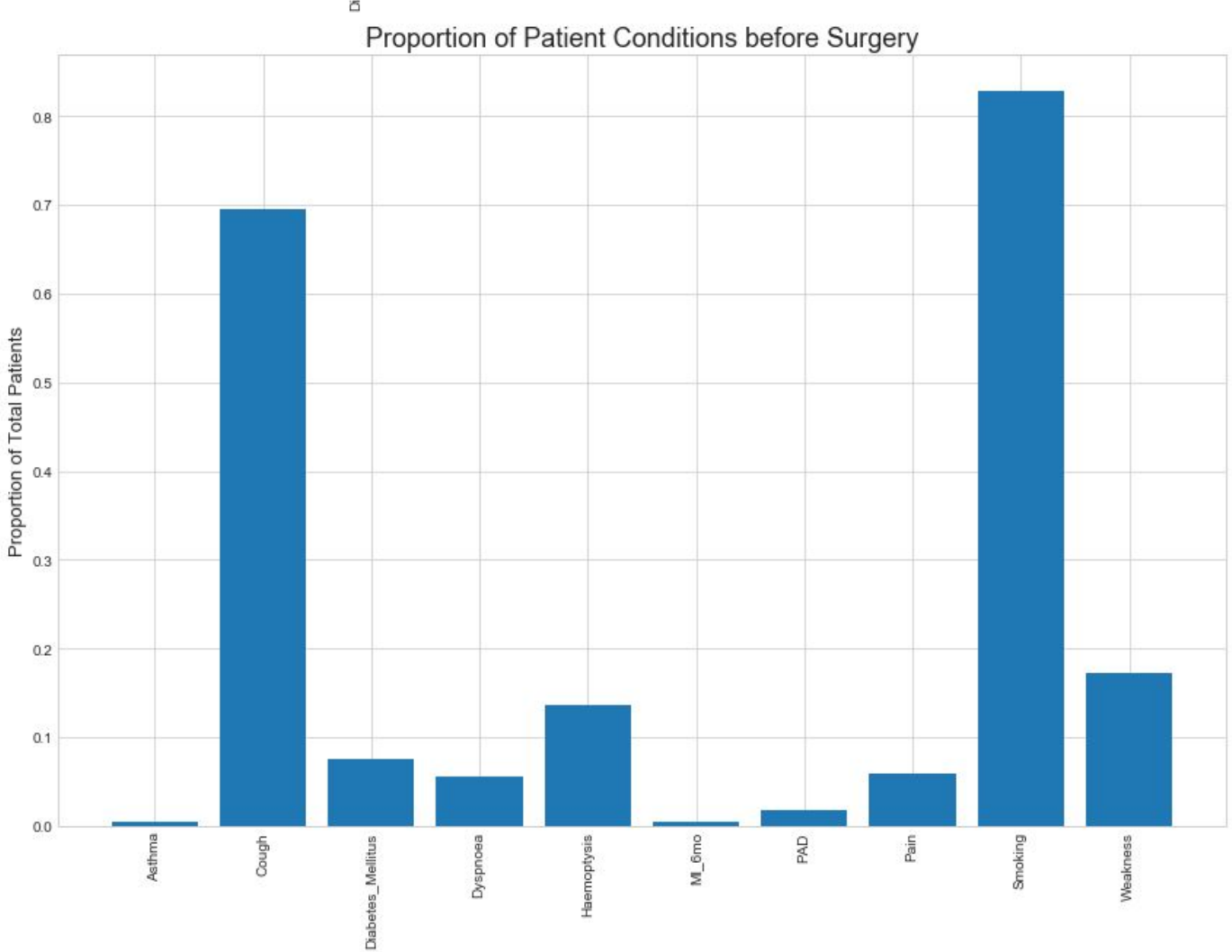
- 69 death out of 454; 15.20% death rate in 1 year post-op.

Attribute	Death in 1 year (Mean)	Live 1 year (Mean)
FVC	3.195072	**3.304597**
FEV1	2.383188	**2.540805**
Performance	**0.913043**	0.774026
Pain	**0.101449**	0.051948
Haemoptysis	**0.202899**	0.124675
Dyspnoea	**0.115942**	0.044156
Cough	**0.797101**	0.677922
Weakness	**0.246377**	0.158442

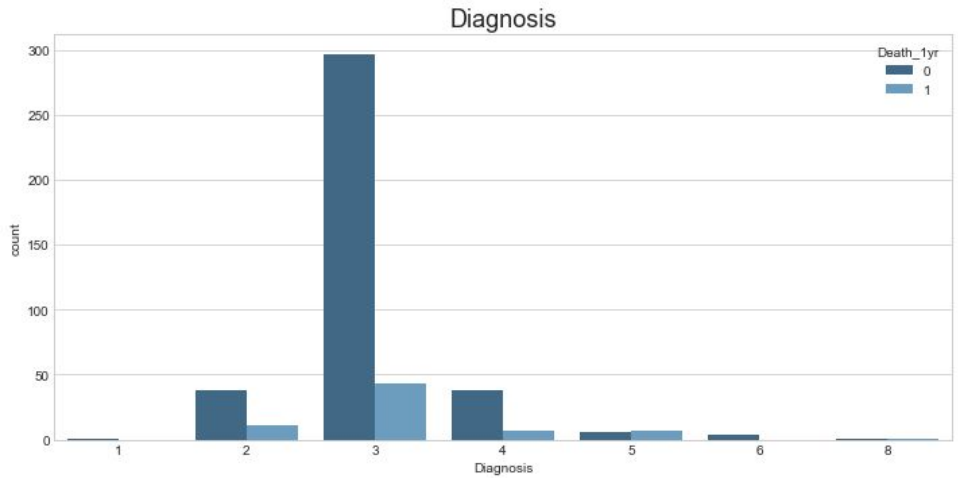
Attribute	Death in 1 year (Mean)	Live 1 year (Mean)
Tumor_Size	**2.014493**	1.683117
Diabetes_Mellitus	**0.144928**	0.062338
MI_6mo	0.000000	**0.005195**
PAD	**0.028986**	0.015584
Smoking	**0.898551**	0.815584
Asthma	0.000000	**0.005195**

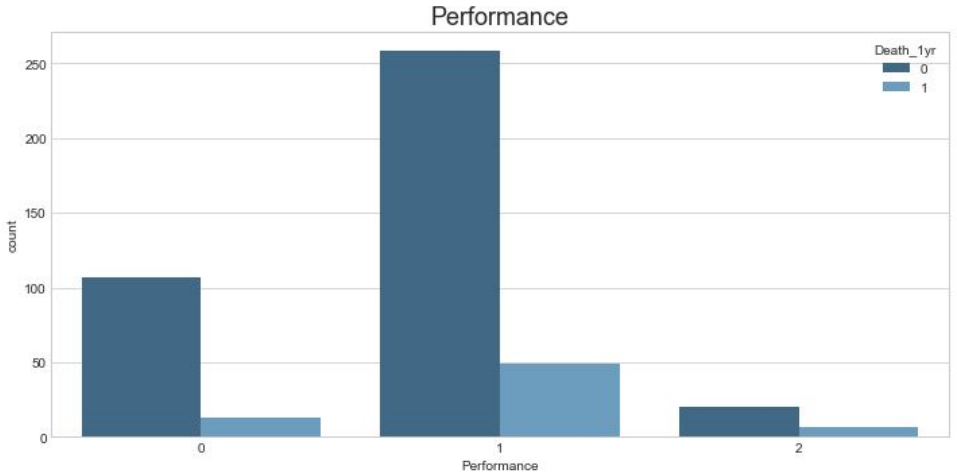


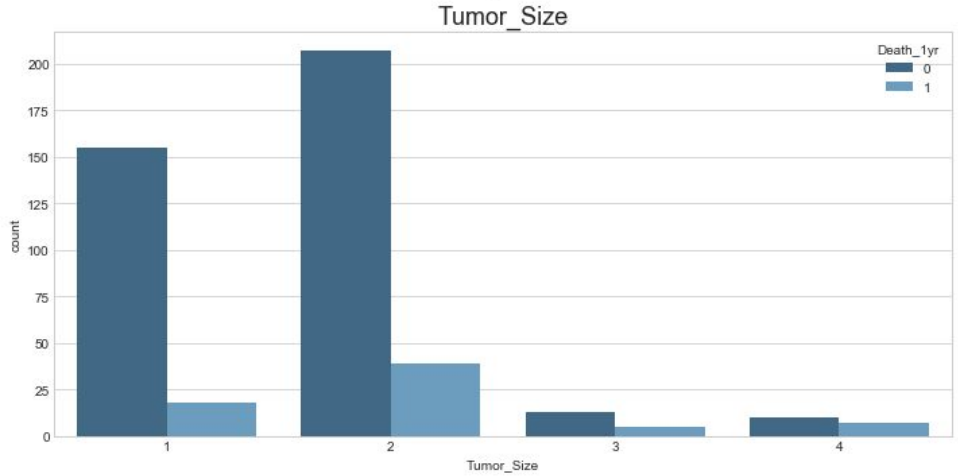
Mean Difference % of Dead and Live (1 yr)



Proportion Of Patient Conditions before Surgery







Categorical Data

Hypothesis Testing

- Null Hypothesis: The 1 year death and live patients have the same mean, tested for each attribute.
- Test Statistic: Mean difference between death and live patients
- Significance level: 0.05

Results of Hypothesis Testing

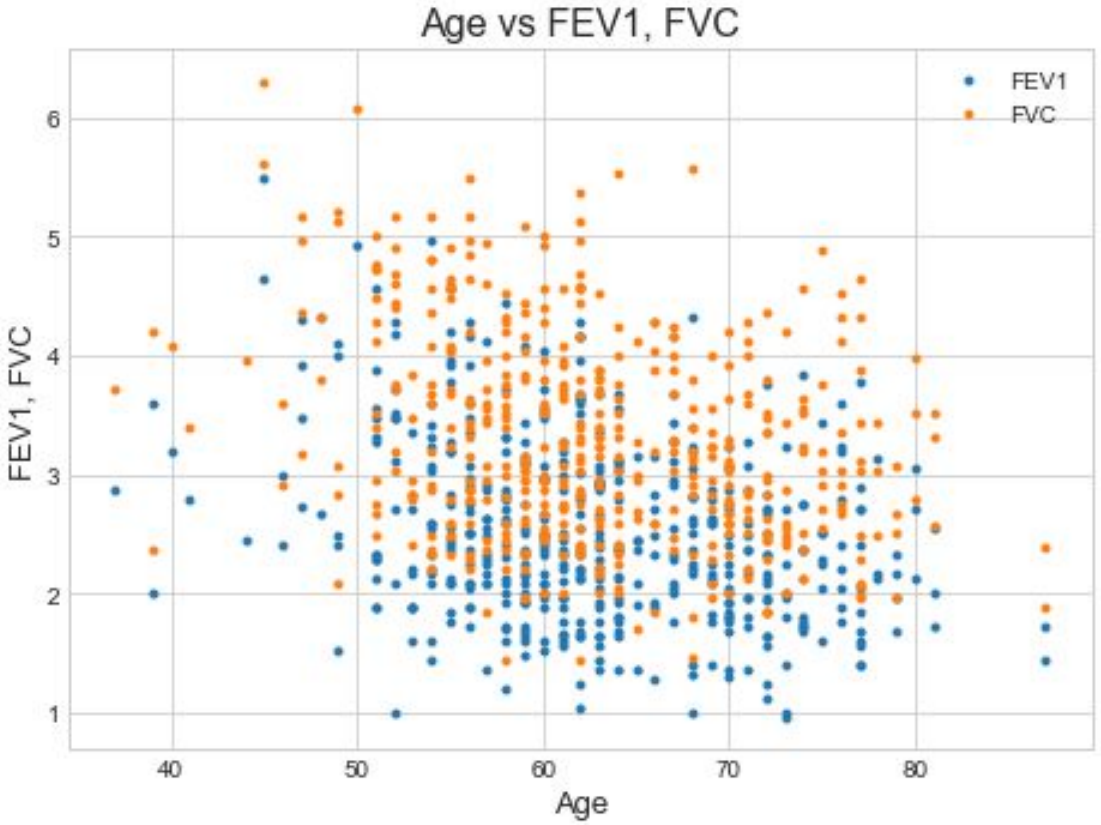
Attribute	**P value**
FVC	0.1706
FEV1	0.0588
Performance	**0.0300**
Pain	0.0964
Haemoptysis	0.0623
Dyspnoea	**0.0242**
Cough	**0.0320**

Weakness	0.0606
Tumor_Size	**0.0003**
Diabetes_Mellitus	**0.0209**
MI_6mo	0.7264
PAD	0.3498
Smoking	0.0581
Asthma	0.7178
Age	0.2714

Mean difference % for Attributes of Significance

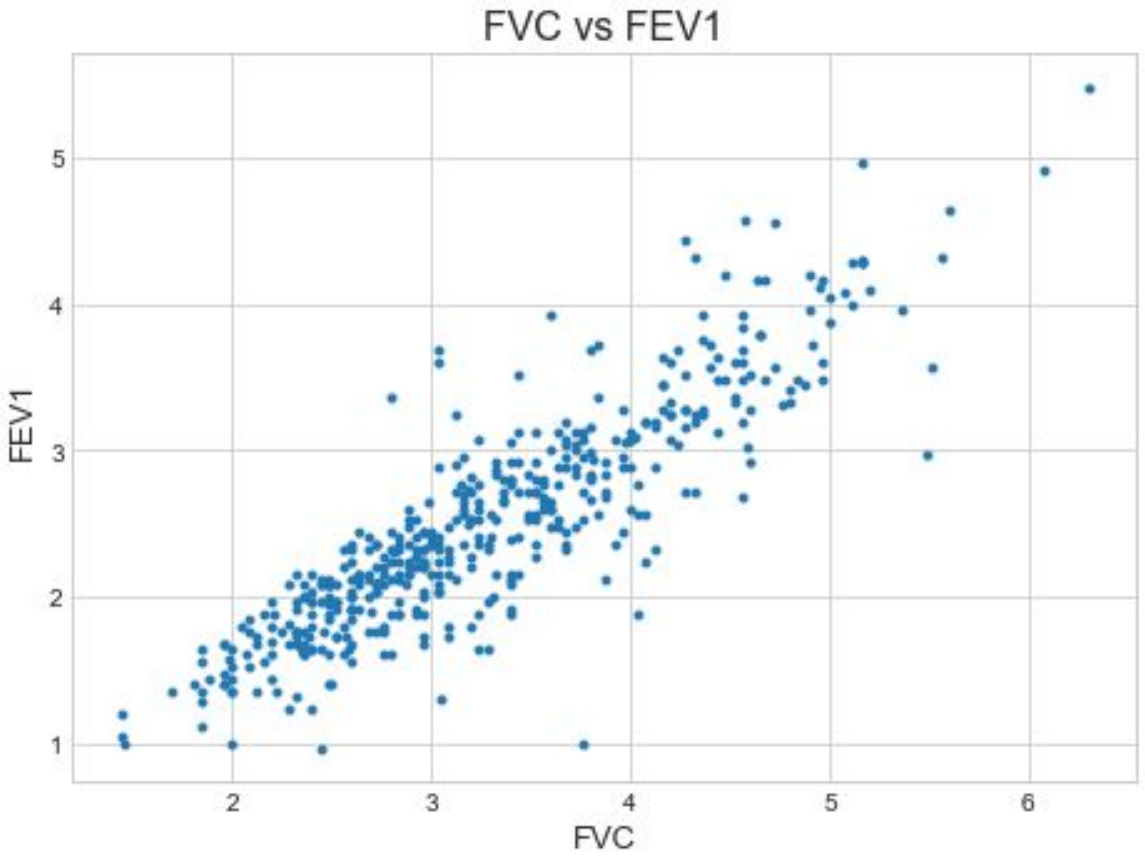
- Performance = 17.96%
- Dyspnoea = 162.57%
- Cough = 17.58%
- Tumor_Size = 19.69%
- Diabetes_Mellitus = 132.49%

Correlations of Numerical (Age, FVC, FEV1) Data



Correlation Coefficients:

- Age & FEV1
  - -0.2994
- Age & FVC
  - -0.3096



Correlation of FVC and FEV1

Correlation Coefficient:

- 0.8875

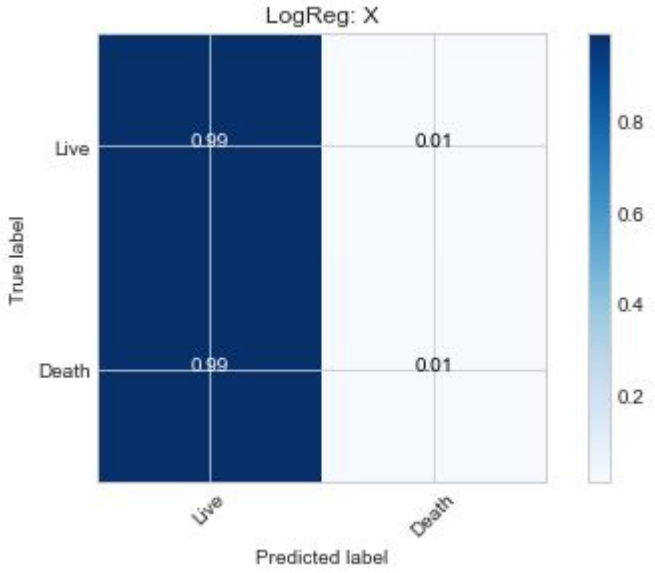
FEV1/FVC Ratio:

- Used in diagnosis of obstructive and restrictive lung disease

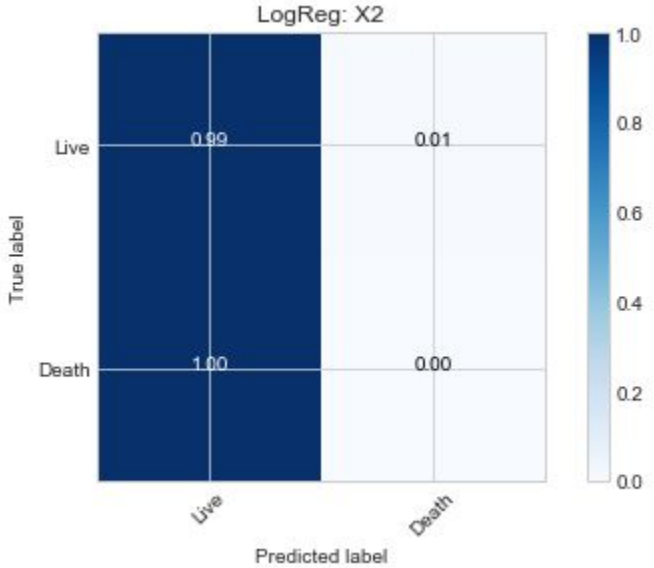
Machine Learning - Supervised Classification

- Target Variable: Death_1yr
- X
  - Drops target variable, MI_6mo, Asthma
- X2
  - Attributes of significance from Hypothesis testing
  - Performance, Dyspnoea, Cough, Tumor_Size, Diabetes_Mellitus

Logistic Regression

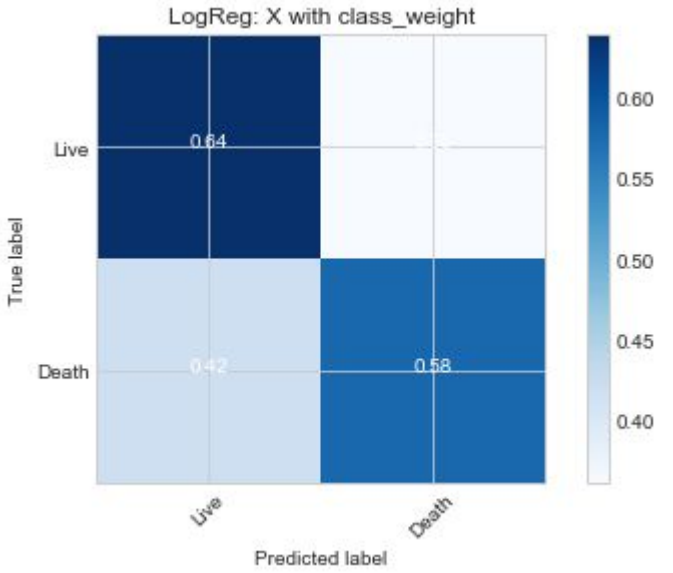


Accuracy score: 84.36%

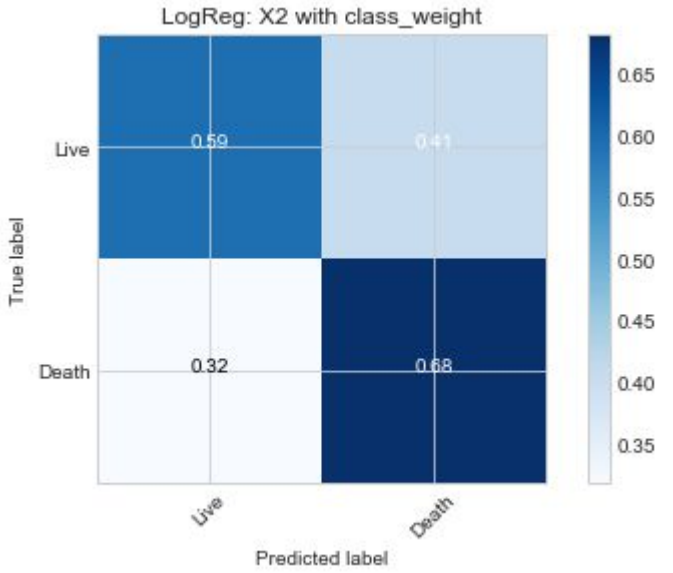


Accuracy score: 84.36%

Logistic Regression w/ Class Weight

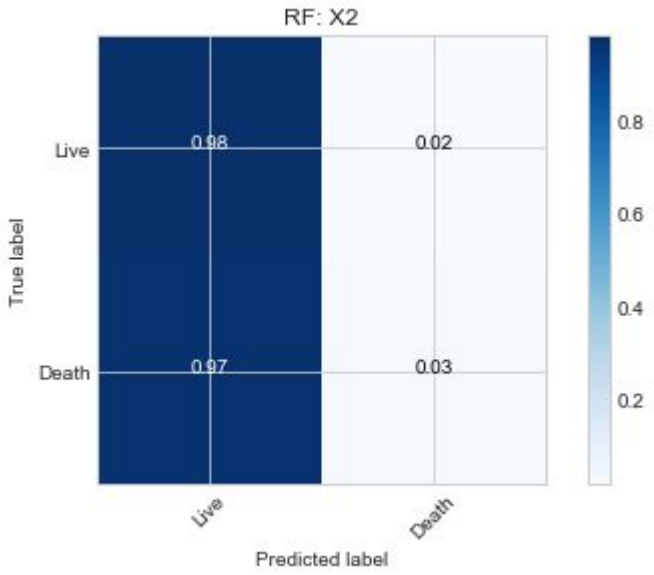


Accuracy score: 63.00%

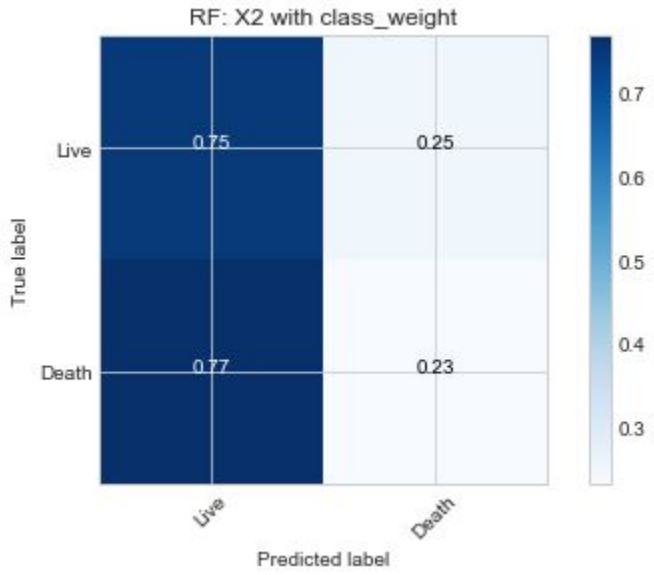


Accuracy score: 60.79%

Random Forest Classifier (X2 Data)



Accuracy score: 83.70%



Accuracy score: 66.96%

Accuracy vs Correct Death Predictions

Beyond this project...

- Decision on desired outcome considering false prediction costs
- More data
- Hyperparameter tuning
- Ensemble method with preferred model