Thoracic Surgery Data Set

Exploratory Data Analysis

Are there variables that are particularly significant in terms of explaining the answer to your project question?

- All the attributes were investigated for their means and relationships to the ‘Death_1yr’ column, the target variable, to help explain if there are attributes of note that relate to higher death rates, and perhaps predict those resulting in death within 1 year time.
- A normalized mean difference of the attributes reveal positive relationships to death for dyspnoea, diabetes mellitus, pain, PAD, haemoptysis, weakness, tumor size, cough, performance, and smoking (in descending value). Negative relationships to death are revealed to be asthma, MI of 6 months, FEV1, and FVC.
- Condition counts reveal most of the data points exhibit cough and smoking, being over 300 each, while the other attributes being under 80 data points.
- Hypothesis testing of mean differences reveal significant attributes to be performance, dyspnoea, cough, tumor size, and diabetes mellitus.
- Mean difference percentage for death in 1 year patients for these attributes are:
    - Performance = 17.96%
    - Dyspnoea = 162.57%
    - Cough = 17.58%
    - Tumor_Size = 19.69%
    - Diabetes_Mellitus = 132.49%

Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

- Among the numerical variables (Age, FVC, FEV1), there are notable correlations.
    - FVC and FEV1 exhibit a strong positive linear relationship, which can be seen on the scatter plot, with a computed Pearson correlation coefficient of 0.89.
    - Age exhibits a weaker, more spread out distribution and relationship with both FEV1 and FVC. The computed Pearson correlation coefficients are about -0.3 for both variables against Age.

- The relationships to be noted for independent variables against the dependent variable are stated in the answers of the previous question with the resulting attributes of significance.

What are the most appropriate tests to use to analyse these relationships?

- Hypothesis testing with permutations for mean differences between live and death patients of 1 year were performed to analysis significance of each attribute, with an alpha value of 0.05.
- For correlation, computing the Pearson correlation coefficient seemed sufficient for analyzing the relationships among the numerical variables Age, FVC, and FEV1.
- ECDF plots were made to look at the distribution of the numerical variables.