Thoracic Surgery Data Set

From UCI Machine Learning

https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data

Abstract: The data is dedicated to classification problem related to the post-operative life expectancy in the lung cancer patients: class 1 - death within one year after surgery, class 2 - survival.

What is the problem you want to solve?

- The problem to solve is whether there is a way to determine post-operative life expectancy in lung cancer patients from patient attributes in the data set.

Who is your client and why do they care about this problem?

- The data was collected retrospectively at Wroclaw Thoracic Surgery Centre. The research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland. However, this problem can concern any hospitals that performs thoracic surgery for lung cancer patients.
- If there is a solution to this problem, hospitals and medical professionals can realize higher death rates in lung cancer patients in specific attribute values and focus on whether they can save the patients who have higher likelihood for death. If there is a pattern, it may also be a point of research to see what the underlying message that the attribute is trying to convey and also why it correlates to higher mortality.

Briefly outline how you'll solve this problem.

- First, the data will need to be explored to find any missing values and if there is a need to tidy up the data for exploration.
- EDA will consist of exploring all attributes' values in terms of what they represent, how they relate to each other, and the quantitative aggregate values.
- Once data is ready, visualization in graphs will help quickly assess data patterns and overall trends.
- Hypothesis testing with the null being the attributes having no relations to death rate will be performed.
- Supervised ML will take into consideration the attributes of concern, which either were determined by EDA or will be determined tuning the ML model, to determine accuracy of the 1-year death rate.

What are your deliverables?

- Github code will be in jupyter notebook format.
- Either a powerpoint or paper will be made to fully report the results.