**Introduction**

The data is dedicated to classification problem related to the post-operative life expectancy in the lung cancer patients. In particular this dataset presents data of patients, attributes and whether they survive within one year of the thoracic operation. The goal is to understand if there is a way to determine the 1-year postoperative survival of lung cancer patients using patient attributes in the dataset. This could help patients and doctors assess the risks of the surgery and, therefore, decide whether to proceed or evaluate other alternatives.

**Dataset description**

The dataset is available at <http://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data>. According to the main repository site the data was collected retrospectively at Wroclaw Thoracic Surgery Centre for patients who underwent major lung resections for primary lung cancer in the years 2007-2011. The Centre is associated with the Department of Thoracic Surgery of the Medical University of Wroclaw and Lower-Silesian Centre for Pulmonary Diseases, Poland, while the research database constitutes a part of the National Lung Cancer Registry, administered by the Institute of Tuberculosis and Pulmonary Diseases in Warsaw, Poland.

The original dataset was in the form of a Weka ARFF file, so we decided to convert it to a CSV file.

**Attribute information**

There are 11 binary attibute containing object string for T and F value, 3 categorical attributes, containing data in the form of a combination of a string and an int value, and other 3 numerical attributes. The corresponding column descriptions found on the UCI machine learning repository site is shown below:

* **DGN**: Diagnosis - specific combination of ICD-10 codes for primary and secondary as well multiple tumours if any (DGN3,DGN2,DGN4,DGN6,DGN5,DGN8,DGN1)
* **PRE4**: Forced vital capacity - FVC (numeric). Amount of air which can be forcibly exhaled from the lungs after taking the deepest breath possible
* **PRE5**: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)
* **PRE6**: Performance status - Zubrod scale (from PRZ,PRZ1,PRZ0)
* **PRE7**: Pain before surgery (T,F)
* **PRE8**: Haemoptysis before surgery (T,F)
* **PRE9**: Dyspnoea before surgery (T,F)
* **PRE10**: Cough before surgery (T,F)
* **PRE11**: Weakness before surgery (T,F)
* **PRE14**: T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)
* **PRE17**: Type 2 DM - diabetes mellitus (T,F)
* **PRE19**: MI - Myocardial infarction (Heart Attack) up to 6 months before surgery (T,F)
* **PRE25**: PAD - peripheral arterial diseases (T,F)
* **PRE30**: Smoking (T,F)
* **PRE32**: Asthma (T,F)
* **AGE**: Age at surgery (numeric)
* **Risk1Y**: 1 year survival period - (T) value if died (T,F)

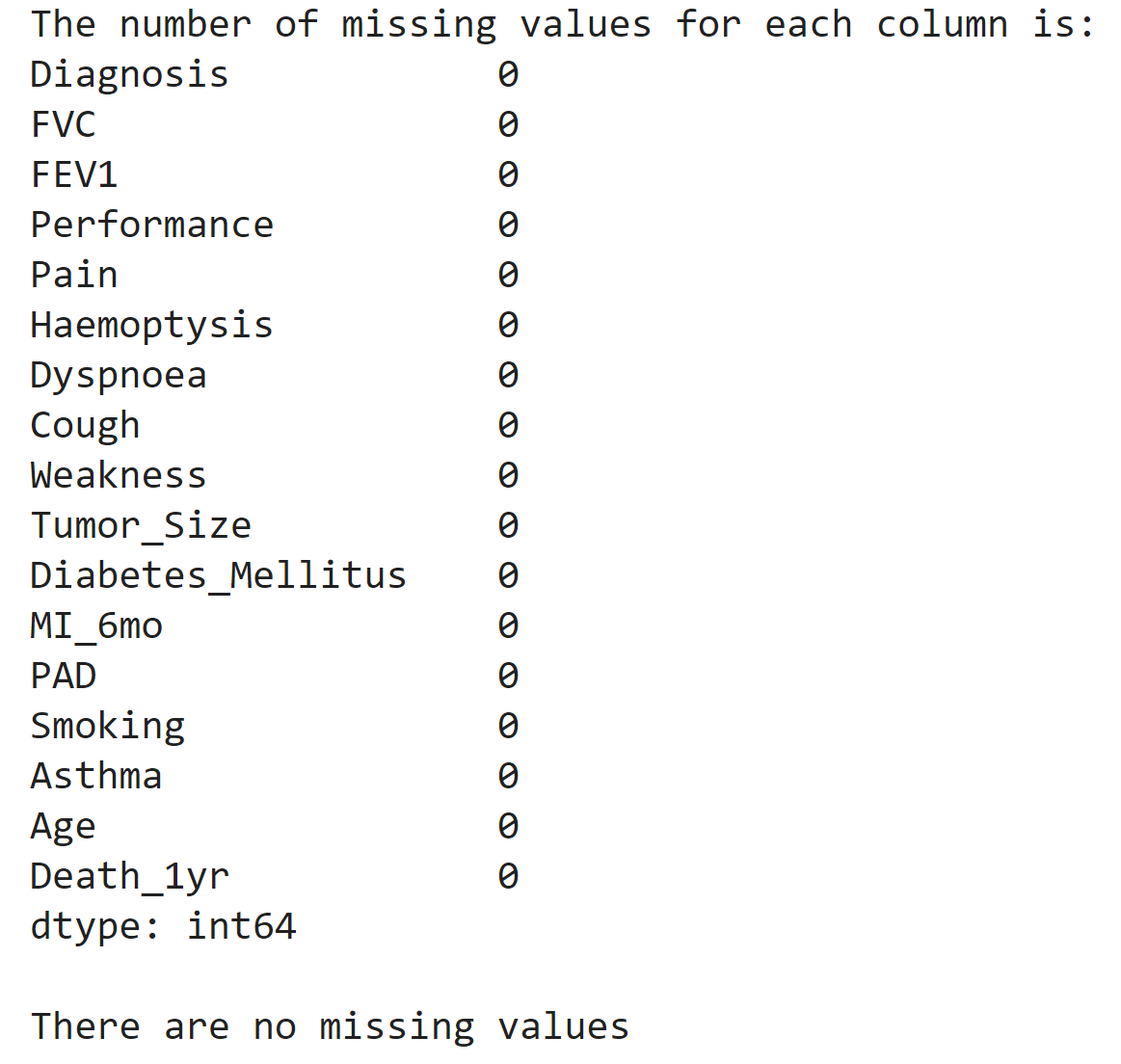
**Dataset cleaning**

We decided to make changes to the original dataset in order to make it more understandable and easier to use. In particular we mapped T-F values in 0-1 values, we removed the id column, becouse it was not necessary and useless in the description of the patients, and finally we renamed the attributes with more undestendable names. The mapping is shown below:

* 'DGN': 'Diagnosis',
* 'PRE4': 'FVC',
* 'PRE5': 'FEV1',
* 'PRE6': 'Performance',
* 'PRE7': 'Pain',
* 'PRE8': 'Haemoptysis',
* 'PRE9': 'Dyspnoea',
* 'PRE10': 'Cough',
* 'PRE11': 'Weakness',
* 'PRE14': 'Tumor\_Size',
* 'PRE17': 'Diabetes\_Mellitus',
* 'PRE19': 'MI\_6mo',
* 'PRE25': 'PAD',
* 'PRE30': 'Smoking',
* 'PRE32': 'Asthma',
* 'AGE': 'Age',
* 'Risk1Yr': 'Death\_1yr'

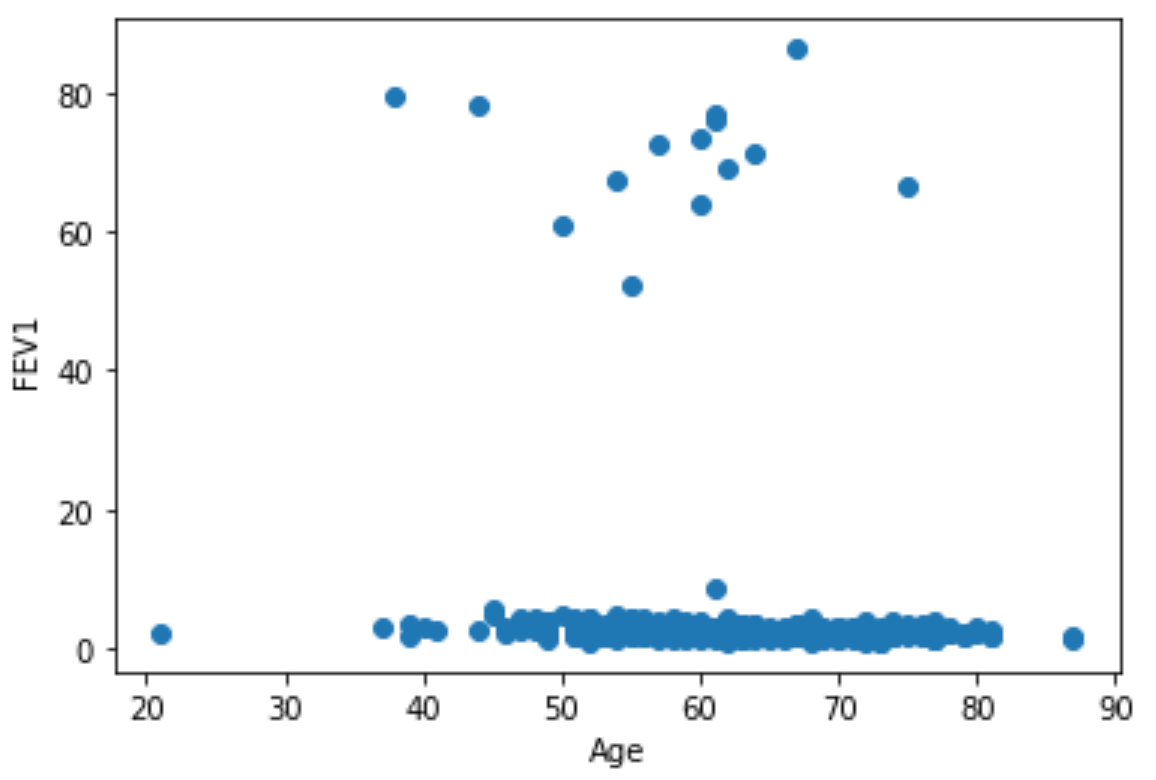
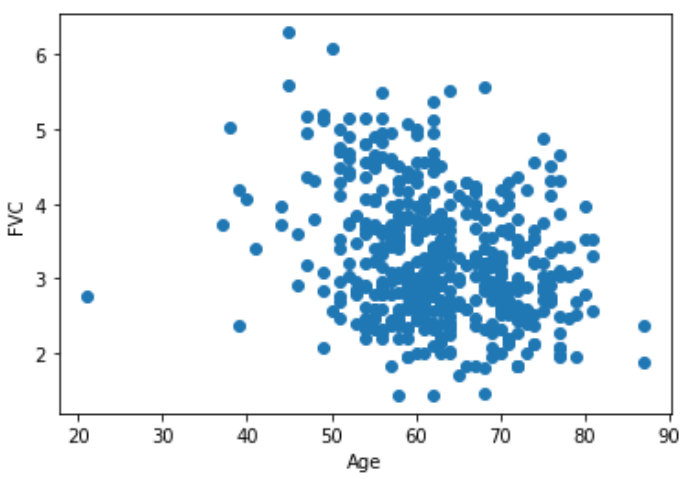
**Missing Values**

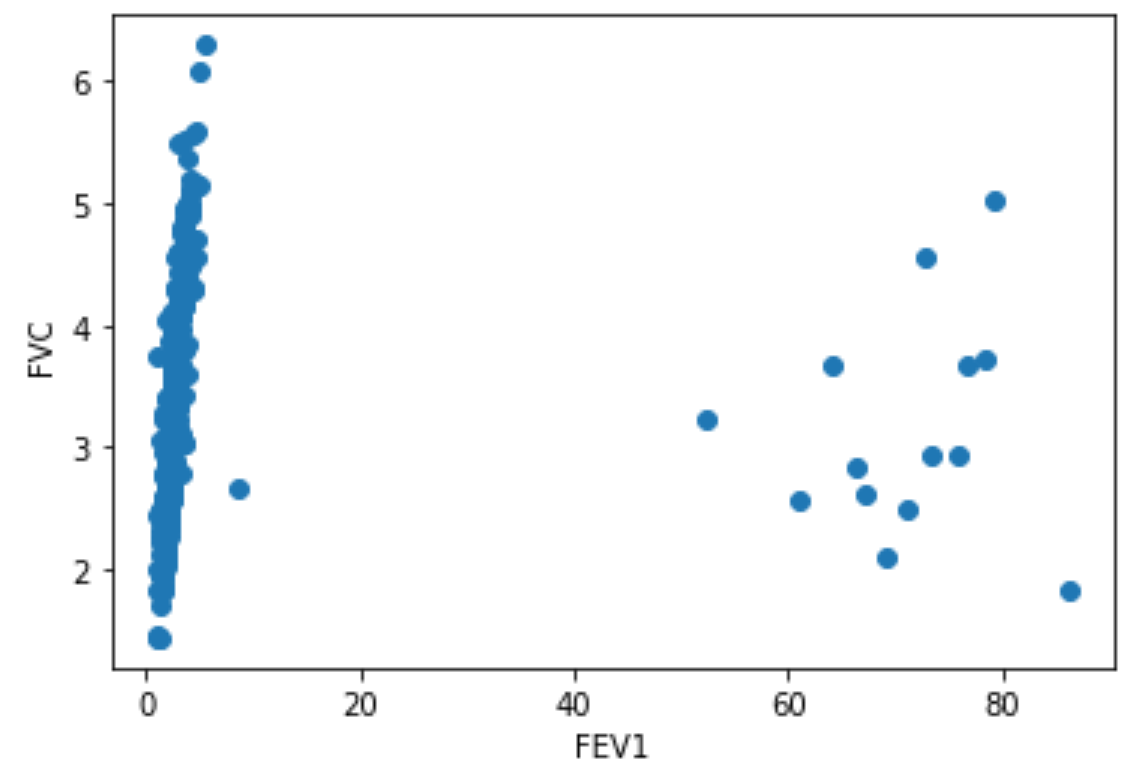
Verifying the presence of miss values, we found that there are none.



**Outliers detection**

For the outlier detection, there are only three numerical attributes to be considered: FVC, FEV1 and Age. The scatter plots of those attributes show the presence of some outliers:





So we decided to remove all data with a FEV1 value grater than 7. In this way we removed 15 outliers. Observing the Age attribute it is possible to notice the presence of a possible outlier with a value of about 20. However, we decided to keep it as we think it may contain important information for the purposes of classification.

After the outliers removal, the dataset contains 455 instances against the starting 470. This quantity seems sufficient compared to the original number of instances, without remove too much information from the original data.

The code for the above work can be found at:

<https://github.com/GiuseppeMoscarelli/Thoracic-Surgery/blob/main/src/0_clean_dataset.ipynb>