Thesis (Draft)

Topic: Prediction of Survival Rate using Advanced Data Analysis

Student: Anna Danielyan

Supervisor: Hrant Davtyan

Abstract

Nowadays Artificial Intelligence (AI) penetrates in every field of our lives. AI jumps in to predict and direct the best course of action, ideally predicting the incident way before it happens. In this concept healthcare is not an exception. Implementation of recent data modelling tools in the medical field brings it to a new stage within the scope of the advanced analytical framework and data processing. In this work, we are going to use data mining techniques to create effective diagnostic models and predict the survivability of an individual's performance within one year after thoracic surgery.

Introduction

Cancer is not one disease, but a collection of related complications that can occur almost anywhere in the body. In the past year, the global cancer burden is estimated to reach over 18.1 million new cases and 9.6 million deaths. In terms of incidence, the top three cancer types are cancers of the lung, female breast, and colorectum. Being ranked within the top five mortality causes, these three cancer types together are responsible for one-third of the cancer incidence worldwide.

Lung cancer is the most commonly diagnosed cancer (14.5% of the total cases in men and 8.4% in women) and the leading cause of cancer death in men (22.0%, i.e. about one in 5 of all cancer deaths). A general treatment used to diagnose or repair lungs affected by cancer is thoracic surgery. Later, however, refers not only to operations on lung but also on organs in the chest, including the heart, and esophagus.

Although early detection and treatment positively contribute to the survival rate of patients, the post-operative issues can hurt an individual's quality of life. Thus, while making decisions, not only the extremely inferior prognosis for patients should be accounted but also possible risks that surgery can bring both in short and long terms. Healthcare data analysis tends to resolve real-world issues in diagnosis providing more precise and accurate decision model to be used for an individual's treatment.

Doctors might use historical statistics that researchers have been collecting over a long period on individuals with the same type of diagnoses. Here, however, one must note that the structure of data might differ depending on its original source and used techniques. When doing predictive analysis, the abovementioned point should get a high priority to building univariate and usable models. For this purpose, the *National Institute of Cancer* suggests four commonly used statistics based on survival type:

• Cancer-specific survival

This is the percentage of patients with a specific type and stage of cancer who have not died from their cancer during a certain period after diagnosis. That can be 1 year, 2 years, 5 years, etc., with 5 years being the time period most often used. Cancer-specific survival is also

called disease-specific survival. In most cases, cancer-specific survival is based on causes of death listed in medical records.

• Relative survival

This statistic is another method used to estimate cancer-specific survival that does not use information about the cause of death. It is the percentage of cancer patients who have survived for a certain period after diagnosis compared to people who do not have cancer.

• Overall survival

This is the percentage of people with a specific type and stage of cancer who have not died from any cause during a certain period after diagnosis.

• Disease-free survival

This statistic is the percentage of patients who have no signs of cancer during a certain period after treatment. Other names for this statistic are recurrence-free or progression-free survival.

In this variety of statistics, however, there is no remark on the treatment method if any so it is hard to sort the data to specify its possible effects. To limit the scope of analysis and to avoid misunderstanding and restrictions, this work will not follow a particular definition.

Through this work, a certain dataset including 470 individual records will be under a detailed discussion. The main purpose is to showcase how data, analysed with advanced data mining tools, can support healthcare and research to facilitate collaboration between medical specialities for the sake of improved, more personalised and more targeted patient care.

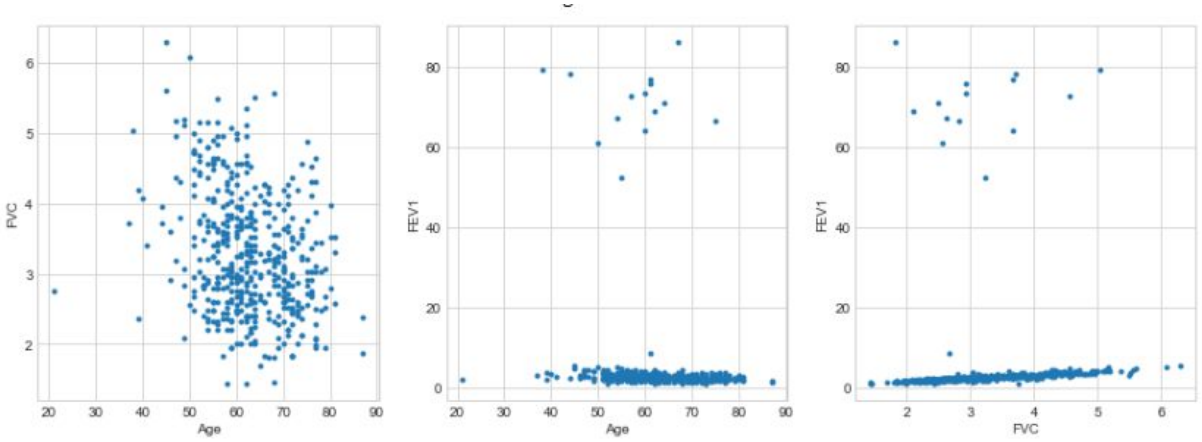
The work will be developed as follows: first, the used literature will be briefly reviewed. After the data will be presented in detail discussing included features to be used in the modelling stage. Further, we will examine a few classification models that might be consistent with the collected data and theory behind it. As a culmination, the final models will be structured and presented. The endpoint of the work will be the sum-up of the main findings of the analysis.

Data and Methodology

The main challenge when structuring classification models is the characteristics of the chosen dataset. Usually, raw data cannot be used directly without any changes due to various factors, such as missing values of some attributes, sequential nature of delivering data, or disproportions in class distribution.

The data collected for further investigation include 470 individual records for people undergone a thoracic surgery in the years between 2007 and 2011. It is compiled from patients with lung cancer registered in the Polish National Cancer Registry. Even though the selected data is relatively clean, before digging deep into analysis and modelling, yet are pre-processed and stored. The idea behind is to make the data qualified for clustering and upcoming analysis.

Graph 1: Finding outliers in the data



Given 17 variables only 3 of them numeric, namely those are Age, FEV1 and FVC. As shown in the plots above, there are outliers in Age and FEV1 variables. Latter are excluded from the preserved data considering the possible misleading effect they can bring into the final results.

After cleaning and disengage outliers from the initial dataset a total of 456 individual records are left. The table below compares the mean values of each feature for the existing classes.

Table 1: Mean values by class and feature

	survival	death
Variable		
FVC	3.306848	3.195072
FEV1	2.542248	2.383188
Performance	0.770026	0.913043
Pain	0.051680	0.101449
Haemoptysis	0.124031	0.202899
Dyspnoea	0.043928	0.115942
Cough	0.674419	0.797101
Weakness	0.157623	0.246377
Tumor_Size	1.684755	2.014493
Diabetes_Mellitus	0.062016	0.144928
MI_6mo	0.005168	0.000000
PAD	0.015504	0.028986
Smoking	0.813953	0.898551
Asthma	0.005168	0.000000
Age	62.540052	63.333333

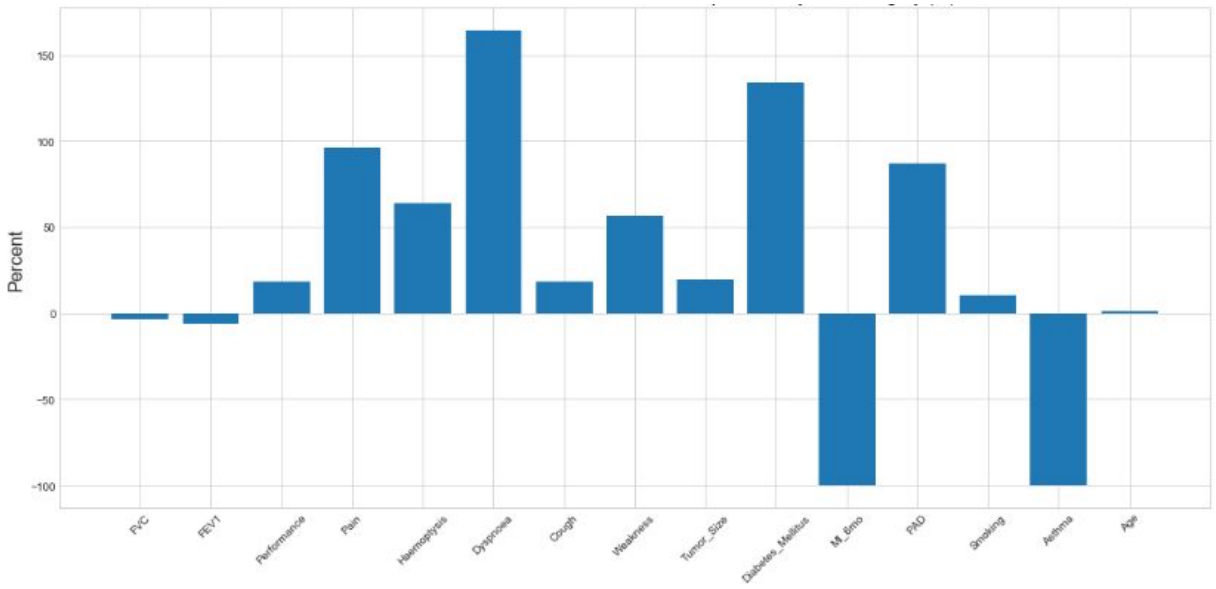
Out of those 456 patients, 69 did not survive in the first year of treatment. Thus, the overall picture of only one-year post-operational observation for each patient shows that received survival results are quite high. Such kind of disproportions in class distribution, mentioned in theory above, is also known as the imbalanced data problem and is the case in this analysis since we have only a 15.13% death rate for usable data.

Before moving forward, let's first examine survival rates per individual features.

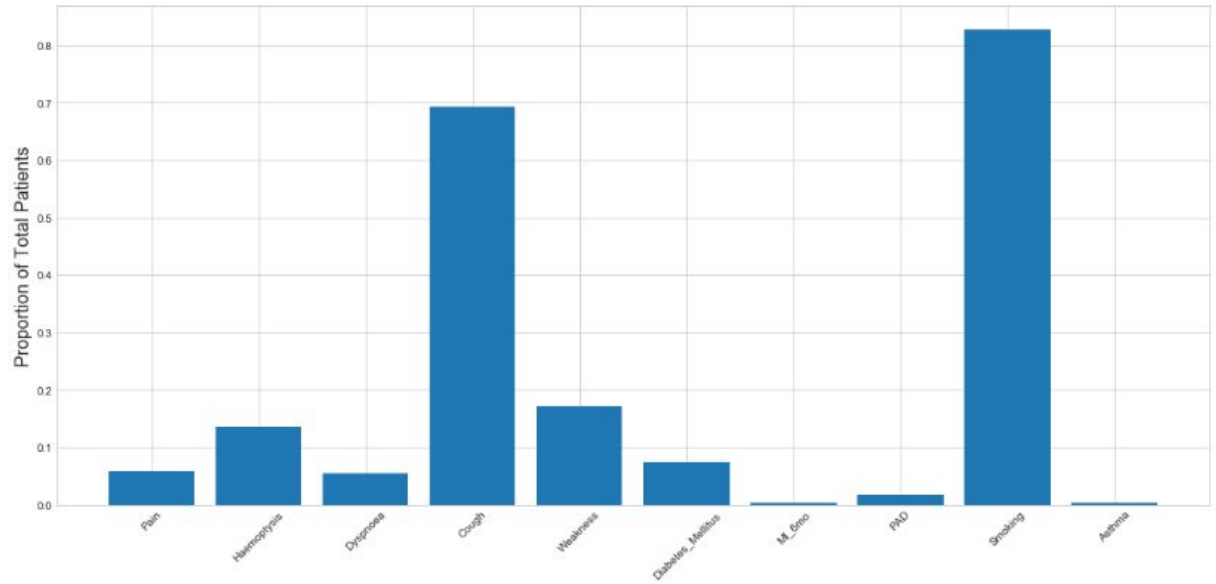
Looking at the means of the two different patient classes (Table 1), there are features with significant differences and those with a minor. However, just looking at the numbers without appropriately weighting them makes comparison difficult. Therefore, the mean

differences between died and survived patients within one year of surgery are observed together with the frequency of dummy conditions recorded within patients before the surgery.

Graph 2: Mean Difference between Death and Survival classes after 1 year of Surgery (%)



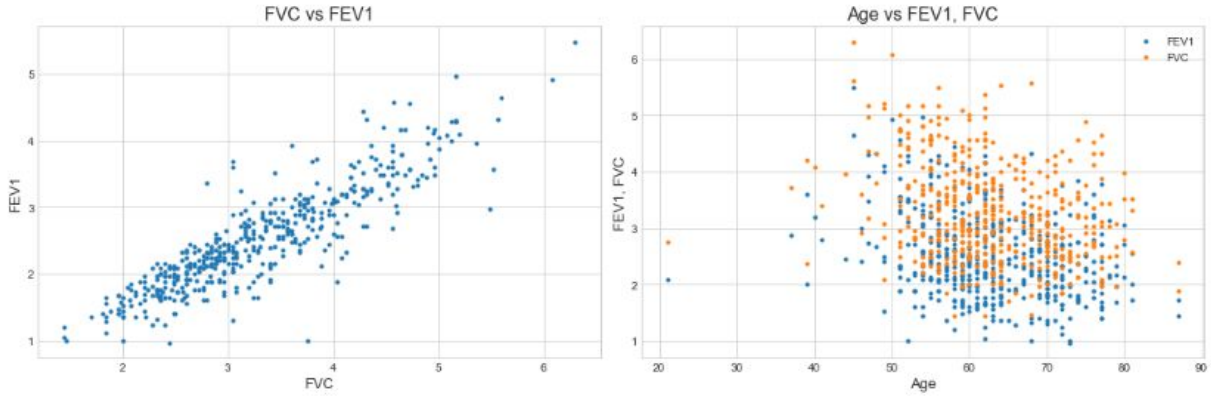
Graph 3: Proportion of Patient Conditions before Surgery



From Graph 2 Dyspnoea, Diabetes Mellitus, Pain, PAD, and Haemoptysis are highlighted to be strongly presented as notable attributes for those who died. Asthma or MI symptoms are shown only within the survived individuals which explains the negative 100% values of Asthma and MI of 6 months variables. The number of instances of each attribute in

combination with the mean differences, in fact, emphasises the importance of it, supporting the feature choosing a decision. In addition to this, the overall count should be considered when comparing mean differences because the lower count numbers will have larger fluctuations to small differences. The most noteworthy evidence of these are the proportional values of Cough and Smoking showing a strong correlation to those patients applicable for getting a thoracic surgery for lung cancer, but the mean differences are a small positive value indicating less representation in the dead patients.

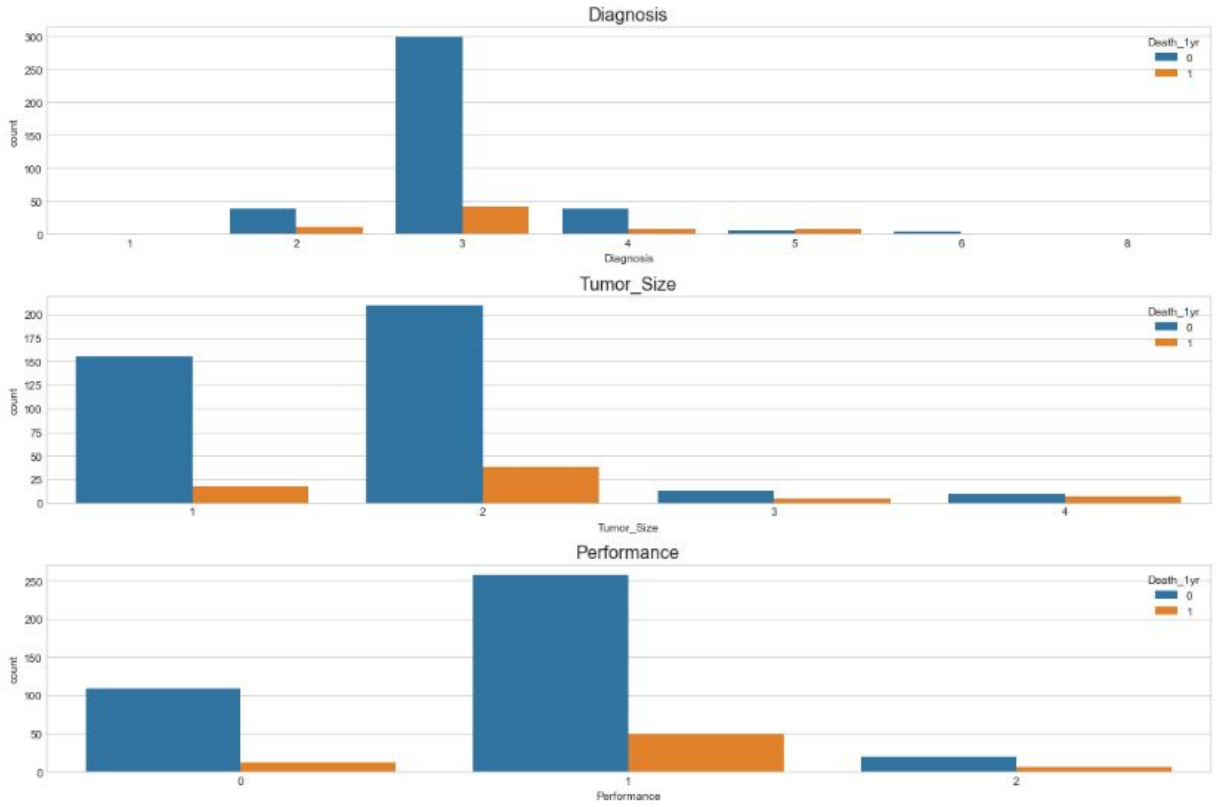
Graph 4: Correlation between numeric variables



Here one can see a strong positive correlation between FVC and FEV1, while Age has a slight negative trend in the graphs. The correlation coefficient calculated for FVC and FEV1 is 0.89, which is very strong on top of the fact that the data points are grouped together to show a visible linear trend. On the other hand, Age's correlation with FVC and FEV1 are about -0.3 for both, but the data points are more spread out. The mild negative trend for age against the other two features makes intuitive sense as it would be expected that as you get older, your lung capacity decreases.

Additionally, there are three categorical variables available in the data to be observed for intimating the number of deaths each one caused.

Graph 5: Distribution of categorical features by class



Summing up the data description: for further investigation, all attributes but target variable, Asthma and MI_6mo (heart attack up to 6 months prior the surgery) are excluded from features as they had -100% mean difference value.

The target to be followed is clustering the available data to get the survival probability within one year period after the surgery. Clustering is a process of assigning dataset to subgroups according to the unique features. Given a set of individual records, a modified clustering algorithm is used to classify each data point into two specific groups (in our case survival and death groups). What we want to accomplish is to have an algorithm, that will learn from our historical data the important features affecting the death evidence in one year of the surgery and use that information to predict turnover. This problem is called binary classification as the target variable is a dummy for it catches only two values (1 for those who died and 0 otherwise).

There are several classifiers (algorithms) implacable for such kind of data clustering among which Decision Tree and Logistic Regression Classifiers are highlighted to be used for finding out the frequent patterns of medical data.

The main concept for both models is splitting the data into two sets: train and test. The training component is used for model development, while the test set is for model validation. When the initial model is built and fitted on the training set it is followed by the estimation of model accuracy on both test set and predictions. Basically, the accuracy score of the predictions is calculated to see how well we did in the modelling.

Still, each clustering has a prediction error on the predictions and the general accuracy isn't enough to claim that the model is a good one, especially noting that the analysed data turned out to be imbalanced. Hence, general accuracy is not providing enough information on separate classes that is why, to understand what other metrics of evaluation are doing, a confusion matrix steps in.

We have two possible outcomes in reality (to survive or die), in general, meaning that there are four possible scenarios presented in this so-called confusion matrix.

Graph 6: Confusion Matrix

Confusion Matrix		Reality	
		0	1
Predicted	0	TN	FN
	1	FP	TP

When the prediction is 0, we call it *Negative*, and when it's 1, it is widely accepted to call it *Positive*. Similarly, when the prediction is correct, we say it is *True*, otherwise, it is *False*. Thus, if in reality someone died but was predicted to be a survivor, then we have *False Negative* (FN), as the prediction was both False and Negative.

General, accuracy score is a good choice only if classes in the dataset are balanced. However, class imbalance may lead to higher accuracy score, when in fact the model is failing to correctly predict death.

Based on four possibilities the confusion matrix provides, many different metrics are developed in analytics to measure the performance of the model.

Whenever the target of the prediction is mostly to focus on those who are dying, then probably less FN are prefered, that is, people who died in reality but your algorithm is not able to predict it. Hear **recall/sensitivity** score should be used. Higher values of recall correspond to lower values of FN. Likewise, if the target under the attention is those who survived, less FP is the target, which can be achieved with higher **specificity** score. To get the percentage of people who truly died within one year after the surgery among those who were predicted to die, then the **precision** score is handy to use.

Consequently, if target died patients, we would concentrate on recall, if survivors, then on specificity. However, if the endpoint objective is to have good predictions on both, then the best choice is to use the AUC score. AUC stands for Area Under Curve and is basically a compound measure that maximizes when both recall and specificity scores are maximized. To calculate the AUC score, one needs to place *Recall* on vertical, and *1-Specificity* on the horizontal axis and draw the curve in the graph, which is called ROC (Receiver Operator Characteristics).

While these other metrics are more robust and informative, they only partially solve the class imbalance problem. As in the very beginning, when starting the model development, there is no certain indicator about probabilities the original algorithm might be able to correctly predict 0s but not 1s. One possible solution towards the class imbalance argument is to change prior to probabilities assigning class weights which will make the probability of both

being 0 and 1 equal to 50%. This probably negatively affects the general accuracy, but AUC and especially recall should likely be improved, as now both classes are equally important.

Yet, a model which is accurate on one data, might not be that much accurate on the other. So another not less important objective is to achieve a model that is generalizable or in other words, works well not only on our current dataset but also in possible future datasets. Here is where the hyperparameters tunning and feature selection are implemented until the best possible model is developed.