Thesis

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, 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 more complex 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.

Yerevan 2019

Table of Contents

Introduction

Cancer is not one disease, but a collection of related complications that can occur almost anywhere in the body. In 2018, 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, 22%, cause of cancer death in men[1]. 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 a more precise and accurate decision model to be used for an individual's treatment. Here, however, one must note that the structure of data might differ depending on its source and used techniques. When doing predictive analysis, the above-mentioned points should get a high priority to building univariate and usable models. For this purpose, the *National Cancer Institute*[2] suggests four commonly used statistics:

• *Cancer-specific survival:* the percentage of patients with a specific type and stage of cancer who have not died from it during a certain period (commonly during the 5-years period) after diagnosis.

[1] www.who.int
[2] www.cancer.gov

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:* 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:* 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:* 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 471 individual records, will be under a detailed discussion. The main purpose is to showcase how data, which mainly includes secondary factors, analysed with complex 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.

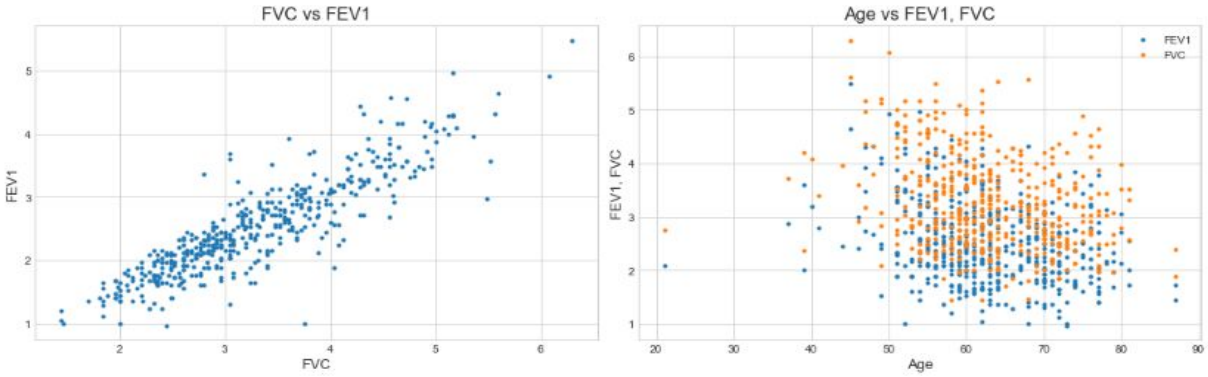
The work will be developed as follows: first, the data will be presented in detail, discussing features to be included in the modelling stage. Further, we will review the classification models together with the theory behind it and go through the methodology towards building the model. As a culmination, the final models will be structured and presented in Results section. The endpoint of the work will be the sum-up of the main findings of the analysis and suggestions for further expansion.

Exploratory Data Analysis

The main challenge when structuring classification models is the nature 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 includes 471 individual records for people who undergo 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 it was pre-processed and stored. The idea behind is to make the data qualified for classification and upcoming analysis.

Graph D1. Correlation between numeric variables



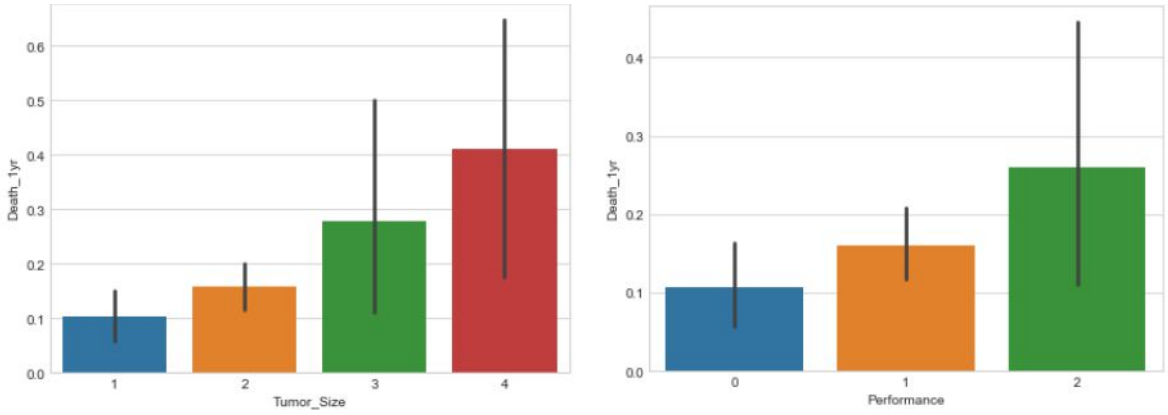
Given 17 variables only 3 of them are numeric, namely, those are the age of the patient at surgery (Age), the volume that has been exhaled at the end of the first second of forced expiration (FEV1) and the amount of air which can be forcibly exhaled from the lungs after taking the deepest breath possible (FVC). 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 to show a visible linear trend. On the other hand, correlation of Age with FVC and FEV1 is 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. 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 getting rid of outliers from the initial dataset, a total of 456 individual records are left. 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.

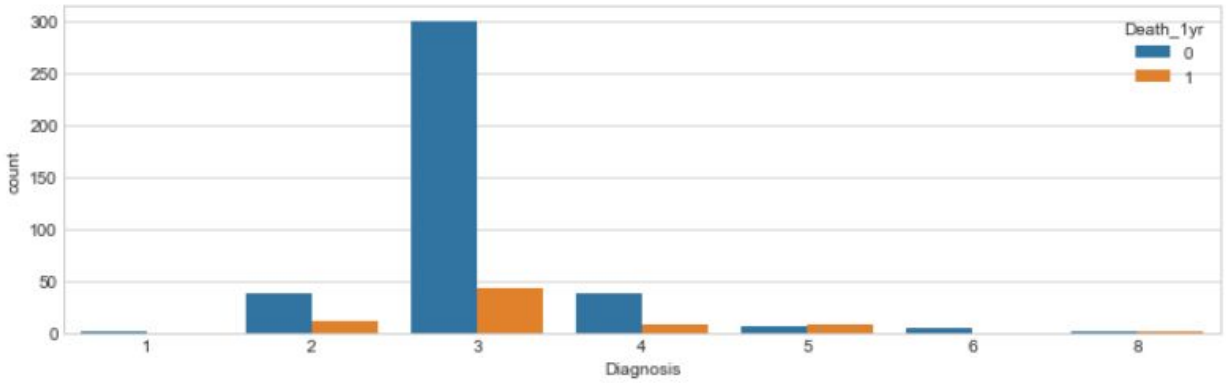
Going forward through the data, there are three categorical variables available to be observed for intimating the number of deaths each one caused. It can be seen that the bigger the tumor size, the higher is the probability the patient will not survive within a year after surgery. The probability distribution of performance among dead patients also has substantiated the same interpretation, noting that poor performance leads to a higher probability of death.

Graph D2. Death probability by Tumor_Size and performance



For Diagnosis, that is another categorical variable in our data, the large majority of patients are in category 3. The other categories are relatively small, while category 2 and 4 will be considered for their counts and included in the data as independent dummy variables.

Graph D3. Distribution of Diagnosis by class



The table below shows up the mean values of each feature for the generated risk groups/classes.

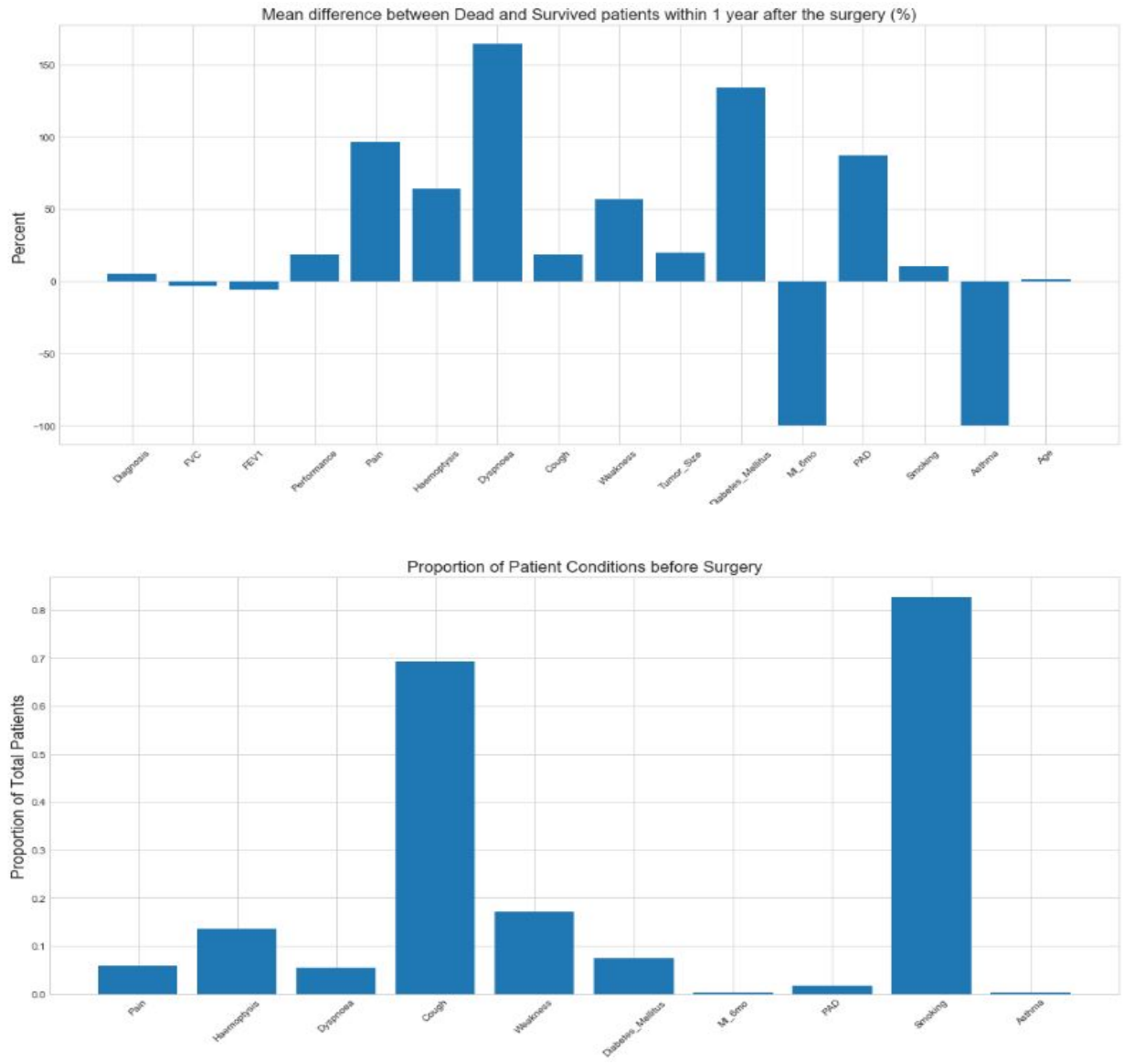
Table D1. Mean values by class and feature

Variable	Survival	Death	Variable	Survival	Death
Diagnosis	3.069767	3.217391	**Weakness**	0.157623	0.246377
FVC	3.306848	3.195072	**Tumor_Size**	1.684755	2.014493
FEV1	2.542248	2.383188	**Diabetes_Mellitus**	0.062016	0.144928
Performance	0.770026	0.913043	**MI_6mo**	0.005168	0
Pain	0.05168	0.101449	**PAD**	0.015504	0.028986
Haemoptysis	0.124031	0.202899	**Smoking**	0.813953	0.898551
Dyspnoea	0.043928	0.115942	**Asthma**	0.005168	0
Cough	0.674419	0.797101	**Age**	62.540052	63.333333

Looking at the calculated mean values, 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, to clarify the way of interpretation for each variable, the mean differences for two classes of possible scenarios (Survival and Death) are compared with the observed mean values for each of the conditions recorded within patients before the surgery.

Graph D4. Comparison between mean differences and proportions of conditions



From Graph D4 Dyspnoea, Diabetes Mellitus, Pain, PAD, and Haemoptysis are highlighted to be strongly presented as notable attributes for those who died. Asthma and heart attack up to 6 months prior

the surgery (MI_6mo) symptoms are shown only within the survived individuals, which explains the negative 100% values of these variables. The number of instances of each attribute in combination with the mean differences, in fact, emphasises the importance of it, supporting the feature selection 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.

Summing up the exploratory analysis: for further investigation, all attributes but target variable, Asthma and MI_6mo are excluded from features as they had -100% mean difference value. Other than these, feature importance will be discussed when building the effective classification models.

Methodology

The case that will be discussed further is known as binary classification problem for the target variable catches only two values: either 1 or 0 (in our case 1 for those who died and 0 otherwise). Hence, the work target to be followed is classifying the available data to get the survival probability within one year period after the surgery.

There are several classifiers (algorithms) implacable for dealing with the discrete outcome among which Logistic Regression, Decision Tree and Bayesian models are widely used for finding out the frequent patterns of medical data.

In a particular study, Decision Tree is used to support the early detection and diagnosis of myocardial infarction making parallels with the Logistic Regression models to predict the probability that a patient with chest pain is having a heart attack based solely upon data available at time of presentation to the ER (Tsien C., 1998). The work conducted on lung cancer diagnosis revealed that the most effective model for predicting the patients' disease appears to be Naive Bayes followed by Decision Tree and Neural Network as it fared better identifying all the significant medical predictors (Krishnaiah V., 2013).

To capture the intuition behind each of the studies, let's review the models first and then turn to their details and evaluation procedures to be applied.

Logistic Regression is a type of predictive analysis to conduct when the dependent variable is binary. The pivot of logistic regression is estimating the log odds of an event using a logistic function in this way estimating the probability of an event to happen. Mathematically speaking, the logistic model transforms the initial output z (1) into probabilities imputing them in the sigmoid function (2) thus

forcing the dependent variable to be either 0 or 1 and then estimates the probability of an event to happen.[3]

(1) $$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

(2) $$\hat{y} = \frac{1}{1 + e^{-z}}$$

In the final step of the logistic model, the predicted probabilities are being converted into decisions using a certain cut-off point (threshold) that in default is set to be 0.5. Based on the objectives, the threshold can be changed to assign the input to class 1 if the corresponding predicted probability is greater than the chosen threshold and to class 0 otherwise.

Decision Tree is another applicable tool used to conduct analysis to address the binary classification problem. Briefly saying, a certain tree algorithm is a representation of decision structure that is organized in the form of a tree that consists of nodes and branches. Based on the knowledge the algorithm learns from the data, it automatically assigns each observation to a certain node and continues the process until getting as pure leaves as possible.

Random Forest is a more complex alternative of the decision tree. It grows many trees inside and chooses the classification having the most "votes" overall the trees in the forest. Although this model is relevant to use for big datasets, it also will be included to check the existing theory.

The last model, **Naive Bayes**, is a subset of Bayesian decision theory. It is competent when dealing with a small amount of data and handles multiple classes. One of the commonly used extensions of the theory is the Gaussian Naive Bayes model.

(3) $$pdf(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} * e^{\frac{-((x-\mu)/\sigma)^2}{2}}$$

[3] Machine Learning in Action, P. Harrington, 2012

Model Performance Evaluation

Each of the discussed models has its unique parameters and attributes that are being used to make the model generalizable and usable not only for the observed data but also for the external inputs.

The main concept for such cases is splitting the data into two sets: train and test. The first set is used for building the model while the other is treated as outer data for performance evaluation and application of the model. After the initial model is built and fitted on the training set it is followed by the estimation of model scoring on both true and predicted values.

When making a prediction, there are two main objectives to follow: **model accuracy** (low bias) and **model consistency** (low variance). Whenever the accuracy scores between the sets differ a lot the overfitting comes to mind pointing on the model inconsistency. Nevertheless, the accuracy alone is not enough for precise evaluation especially noting the imbalanced nature of the data. Here sensitivity, specificity and other statistical metrics step in that can be defined using the confusion matrix.

Table 2: Confusion Matrix

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

Whenever the true and predicted values classify a patient into the death group, it falls into True Positive (TP) cell of the confusion matrix. Otherwise, the score counts for the True Negative (TN). Also, there are two types of prediction errors known as Type l and Type ll errors. If in reality someone survived but was predicted to be dead (FP), then we have **Type l error.** Likewise, if in reality someone died but was predicted to be a survivor (FN), then we have **Type ll error**, as the prediction was both False and Negative.

Sensitivity/recall score refers to the rate of dead patients that the model predicted to die (TPR).

Specificity score refers to the correctly predicted negatives. 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 (3) is handy to use.

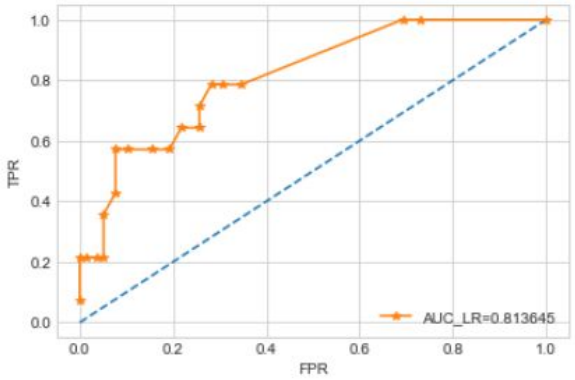
(4)

(5) $$Specificity = \frac{TN}{TN + FP}$$

(6) $$Precision = \frac{TP}{TP + FP}$$

To get better predictions accounting for all the thresholds for both classes, the **AUC** score is calculated. AUC stands for Area Under Curve and is a compound measure that maximizes when both recall and specificity scores are maximized. To visualize the obtained AUC score, one needs to place Recall (TPR) on vertical, and 1-Specificity (FPR) on the horizontal axis and draw the curve in the graph, which is called ROC (Receiver Operator Characteristics).

Graph 6: Receiver Operator Characteristics Example



As each model has its unique workspace for solving the classification problem, it is not necessary for all of them to include the same features. To get the most accurate model, the implementation of feature selection is used. This method is common for Decision Tree while for Logistic Regression p-values can lead to the selection of the needed bundle of independent variables.

Once the modified model returns high predictive accuracy, it can be proposed to be chosen for further analysis with similar datasets.

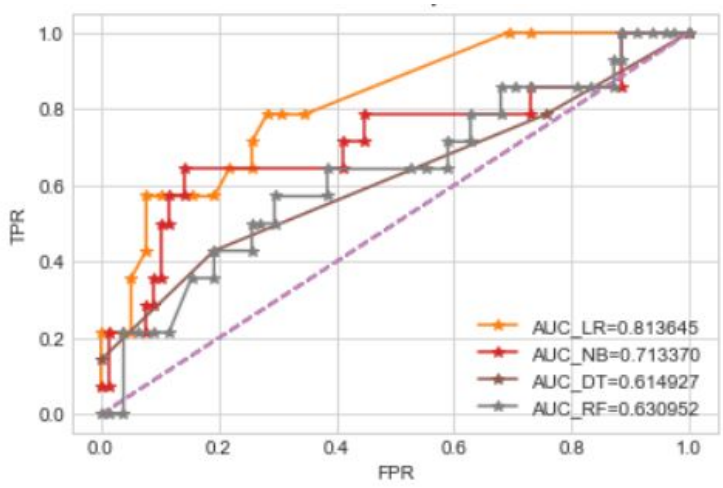
Results

Utilization of the steps discussed in the methodology section above brings us the final results that capture the likelihood and predictive power of a model to be applicable in real-world data. The numeric outputs of models are disclosed in Table R1 while the respective ROC curves can be found in Graph R1.

Table R1. Estimated scores on test set by the model

Score Type	Logistic Regression	Gaussian Naive Bayes	Decision Tree	Random Forest
Accuracy of train	0.848901	0.835165	0.71978	0.85989
Accuracy of test	0.858696	0.869565	0.75	0.728261
ROC_AUC	0.813645	0.71337	0.614927	0.630952
Recall	0.071429	0.214286	0.428571	0.428571
Precision	1	0.75	0.285714	0.26087

Graph R1. ROC Curves with respective AUC Scores by Models



The results above show a higher AUC score for Random Forest. However, the huge gap between train and test accuracies is an indicator of overfitting. Decision Tree also fails to solve the given problem of patient classification given the low scores in all aspects including only 28% of correct prediction among dead patients (precision). Regardless of the high accuracy score of the Bayesian model that also

cannot be labelled as a good one because of the lower AUC score and possible underfitting in the model predictions. Eventually, we are left with Logistic Regression that performed better predictive power in comparison with other models reporting relatively high scores for all measurements and is recommended to be used for further applications.

Now, when the picture of the overall results is clear, let's dig deeper into the details and address Logistic Regression separately.

Logistic Regression is finalized with only seven features in the bundle of usable independent variables including Dyspnoea, Weakness, Tumor_Size, Diabetes_Mellitus and three types of Diagnosis. To make the Logistic Regression output interpretable, the marginal effects are also estimated and presented below.

Table R2. Marginal effects with chosen features

	dy/dx	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Dyspnoea	0.1345	0.054	2.475	0.013	0.028	0.241
Weakness	0.0749	0.038	1.958	0.05	-7.49E-05	0.15
Tumor_Size	0.0684	0.021	3.305	0.001	0.028	0.109
Diabetes_Mellitus	0.1059	0.049	2.164	0.03	0.01	0.202
DGN2	-0.1429	0.071	-2.026	0.043	-0.281	-0.005
DGN3	-0.1951	0.057	-3.4	0.001	-0.308	-0.083
DGN4	-0.1556	0.073	-2.131	0.033	-0.299	-0.013

It can be claimed that with 95% of confidence, each of the features in the model is statistically significant and has its contribution to the performance of the final results. Generally speaking, the type 3 diagnosis (DGN3) and difficulty or laboured breathing (Dyspnoea) are the key factors influencing on the death rate for a patient.

The scoring results for Logistic Regression with weighted classes fetch up 85.86% of accuracy

for the test set while the AUC scores for 0.8136. Latter indicates that the built logistic model is eligible to classify the patient into a specific risk group with 81.36% of accuracy while taking into account all the thresholds.

Running a feature importance algorithm for Decision Tree revealed five outstanding features to be included in the model (Table R3). In this way, the Tree model can accurately classify the 75% of patients keeping the threshold 0.5 and only 61.49% when ignoring the limitations.

Table R3. Feature Importance for Decision Tree Classifier

Feature	**FVC**	**Cough**	**Tumor_Size**	**Smoking**	**Age**
Importance	0.12614	0.197279	0.234562	0.292723	0.149296

In Random Forest classifier the scope of used features was larger. As it is already mentioned, the obtained results didn't comply with the requirements.

Gaussian Naive Bayes model is built to emphasize the influence of numeric features on the final results. The initiation returned successful catching the accuracy of 86.96% and AUC score of 71.33%.

Conclusion

Recent achievements in advanced data mining techniques are capturing the field of medical data analysis. The results supposed to help and direct medical professionals when making decisions and increase the preciseness of the treatment.

Based on the observed results and remarking the importance of exploratory data analysis, there are several conclusions that should be addressed.

First, we have shown the data to be imbalanced and processed model selection accordingly to avoid misleading results. The idea behind the choice of the Logistic model held the intention to provide feature interpretation while the rest of the included models cover pure classification objectives.

In this work, models driven from different families have been analysed aiming to predict the severity of death within one year after the thoracic surgery. The used workflow covered model building and optimization towards the data. Then the performance of all models has been evaluated using different statistical metrics and ROC charts.

We have shown that despite the growing importance of sophisticated and more complex data mining techniques, yet they are not productive when dealing with a small data. In such cases, when they fail to return competent results, classical models like Logistic Regression are the role players. As it was expected, the Logistic model outperformed the other models. Hence, highlighting these results, we can suggest the model be used effectively in a real-life environment for the prediction of the patient's survivability.

A further expansion of analysis can be the collection of big historical data to better capture the feature influence to return a higher predictive power for a model. Another gap that should be filled is the time dependent variable inclusion that will allow conduction of survival analysis.

References

1. Machine Learning in Action, P. Harrington, 2012
2. Krishnaiah V., Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, 2013
3. Tsien, C., Kennedy, Using classification tree and logistic regression methods to diagnose myocardial infarction, 1998.
4. Podgorelec V., Kokol P., Decision Trees: An Overview and Their Use in Medicine, 2002
5. www.who.int
6. www.cancer.gov