

[**Thoracic Surgery\_Patient\_Survival**](https://github.com/sychi77/Thoracic_Surgery_Patient_Survival)

**REPORT**

Name : k. sai vivek reddy

Admn No: 11805765

Section : KM033

Subject: INT247

Roll No : 50

Email:ksaivivekreddy@gmil.com

Submitted to:Sanjay kumar singh

**Abstract**

The data is dedicated to the classification problrm related to the post-operative like expectancy in the lung cancet patients:class 1-death with one year after surgery,class,2=survival.

**Data Set Information**

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.

**How many patients died in 1 year?**

Out of the 454 patients, 69 did not survive 1 year after their operations, which is 15.20% of the total sample size.

Looking at the means of the two different patient classes, there are features with significant differences and those with minor. However, just looking at the numbers without scaling them appropriately to each other makes comparison difficult. So, let's do an approximate normalization of each value for convenient comparison

**Analysis of live and death patients for Diagnosis, Tumor\_Size, and Performance**

For Diagnosis, the large majority of patients are in category 3. The other categories are relatively small while category 4, 2, and 5 should be considered for their counts in that order. The proportion of live to dead at a glance seems to be similar for the diagnosis categories except for 5, where the death count is higher than the live count, which indicates this diagnosis is more fatal than the others even with surgery.

For Tumor Size, categories 1 and 2 are the majority. At a glance, the proportion of the dead to live generally increases with the tumor size ranging from 1 to 4, indicating the higher tumor size correlates to higher chance of death even with surgery. Category 4 tumor size is most even in its split between death and live patient data. Also looking at the dead to live mean difference graph, the dead had higher means indicating larger tumor sizes overall.

For Performance, categories are 1, 0, 2 in decreasing order of count. Performance 0 category reveals low death count and good proportion to live data, which makes sense since on the Zubrod scale 0 is good and 2 is poor. Category 1 and 2 display similar proportion to live and dead patients, but with category 1 having a majority of the count. Referring to the dead to live mean difference graph, the dead had higher means indicating the dead on average had poorer performance with a higher Zubrod score than the live.

**METADOLOGY**

### Hypothesis Test of Mean Differences between Live and Death Patients

All the observations above highlighted the trends and patterns in the attributes. However, to ascertain their significance, a hypothesis test will be useful to see if these patterns and trends are not just by chance.

* **Null Hypothesis:** The 1 year live and death patients have the same distribution and mean. (Tested for each attribute.)
  + **Test Statistic:** Mean difference between death and live patients.
  + **Significance Level:** 0.05

#### Results for Hypothesis Test

With significance level of **0.05**,

* **Cannot Reject Null Hypothesis:** FVC, FEV1, Pain, Haemoptysis, Weakness, MI\_6mo, PAD, Smoking, Asthma, Age
* **Reject Null Hypothesis:** Performance, Dyspnoea, Cough, Tumor\_Size, Diabetes\_Mellitus

The null hypothesis stated that the means of life and death patients were the same for the attributes tested. With the results above, the attributes of significance are those that rejected the null hypothesis. To highlight the trends for those that rejected the null hypothesis, the mean difference percentages are listed below.

**Mean difference % for death in 1 year patients for attributes of significance:**

* Performance = 17.96%
* Dyspnoea = 162.57%
* Cough = 17.58%
* Tumor\_Size = 19.69%.,,
* Diabetes\_Mellitus = 132.49%

### Numerical Data (Age, FVC, FEV1)

The mean difference graph reveals little difference in Age while there is a small negative difference for the dead compared to the live. So this indicates the dead patients on average performed worse for lung capacity compared to the live patients.

The plots below will further investigate the relationship between these three numerical data columns.

***Correlations of FVC, FEV1, and Age***

From looking at the graphs, 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.

### Predictive Modeling

So now that we have thoroughly explored the patterns and trends in the data set, the next step is to utilize machine learning models to see if how well we can predict the target variable, Death\_1yr, with the feature variables.

Since the data set is imbalanced and mostly live patients (85%), just predicting all live patients will give a high accuracy score ~85%. So for the model, accuracy will not be a good score method and instead I will look at average precision score, which summarizes the precision-recall curve. Also for the imbalance, there are couple options including downsampling, upsampling, or adjusting class weights to balance the classes. Since downsampling will create a small data set to work with and upsampling may complicate the data further, I will focus on adjusting the class weights.

**Logistic Regression**

Since the data set is imbalanced with only 15% patient death, the results of the model without any class weight to offset this imbalance favors the live column in the confusion matrix. As you can see above, the model predicts mostly all live patients to maximize the accuracy score to 85%, the size of the live patient data, in both the X and X2 data sets.

With the class weight parameter, the death prediction rate increases at the cost of live patient prediction, and also the accuracy. In order to see the effectiveness of the model for my purpose, the confusion matrix or classification report can be used to assess the death predictions. Also, the average precision score is a good summary of the precision-recall curve, which is useful in this case.

The class weight argument is set to 'balanced' to equalize the death to live ratio, which is 15 to 85. This argument can be altered with any ratio value and the effects can be seen in the graphs above. Although the correct death predictions increased with more class weight on the deaths, the false death predictions increased as well with decrease in correct live predictions. The influence of class weights can be seen in the graph above. Interesting to note that the score dips dramatically around the 5.67 value, which is the equalizing point for the ratio 15 to 85

Now.,

**Random Forest Classifier**

Since the data set is imbalanced with only 15% patient death, the results of the model without any class weight to offset this imbalance favors the live column in the confusion matrix. As you can see above, the model predicts mostly all live patients to maximize the accuracy score to 85%, the size of the live patient data, in both the X and X2 data sets.

With the class weight parameter, the death prediction rate increases at the cost of live patient prediction, and also the accuracy. In order to see the effectiveness of the model for my purpose, the confusion matrix or classification report can be used to assess the death predictions. Also, the average precision score is a good summary of the precision-recall curve, which is useful in this case.

The class weight argument is set to 'balanced' to equalize the death to live ratio, which is 15 to 85. This argument can be altered with any ratio value and the effects can be seen in the graphs above. Although the correct death predictions increased with more class weight on the deaths, the false death predictions increased as well with decrease in correct live predictions. The influence of class weights can be seen in the graph above. Interesting to note that the score dips dramatically around the 5.67 value, which is the equalizing point for the ratio 15 to 85

**Conclusion**

Firstly, more data will improve the scope of the data analysis and models. From analysis of this data set, it is clear that there are significant overlaps of attributes, so more patient data or perhaps creating a new data set with additional attributes could help better distinguish the differences and improve the model. If not new data recordings, there are probably similar data sets that have models that predict lung cancer deaths that could be of use in optimizing this model by using it in combination.

Another option is to optimize the models above with hyperparameter tuning. However, since the accuracy score is unreliable in determining positive death predictions, you would have to determine what score to maximize and minimize before proceeding forward in hyperparameter tuning. This scoring method could be the one used in this report, average precision score, or could be a custom scoring method created using the confusion matrix or the values in the classification report.

Finally, depending on the desired outcome considering false prediction costs, the models can be used in an ensemble method to maximize the outcome desired. The ensemble could be anything ranging from boosting methods or utilizing multiple different models. Again, the desired outcome will depend on the hospital or client and how they view the detriment of giving false positives and false negatives compared to the true predictions for live or death outcomes for patients; in other words, how they want to score the efficiency of the model.

Also by the end of this project, I have got in-depth knowledge about all the methods of the machine learning that I have covered in this project so far.

**Reference**

https://github.com/afonsoblneto/post\_operative\_ml

>>I took some of the methods as the reference to build the code ,from the above github link.

>>I took the data set from the UCI website as everyone.

>>I also visited the sites about the [Thoracic Surgery Patient Survival](https://github.com/sychi77/Thoracic_Surgery_Patient_Survival) , for preparing this report in best possible way.

**THANK YOU**