# Thoracic Surgery Data Set

Umut Deniz Zorgül – 201811070 – CENG 474

**Abstract**

Cancer. Nowadays a lot of people go through the winger about that. Also, lung cancer is the worst, I think. Because a lot of people using cigarette and day to day people who smokes cigarette, increasing. And the main factor of lung cancer is cigarette. After diagnose people trying to fight for life. In this article I worked about Lung Cancer people can survive after surgery.

Keywords—lung cancer, Parkinson's disease, surgery, data, method, analyze

**1-) Introduction:**

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.

**2-) Purpose of The Paper:**

With this article, I tried to interpret the data set of people who had lung cancer. Pre-lung cancer asthma status, age, etc. I'm using a dataset that shows whether they die within 1 year after surgery for lung cancer.

With this article, I want to show that I can interpret the methods I learned in the CENG-474 Introduction to Data Science course through Weka. First, that's my goal. My secondary goal is to examine this dataset, which arouses my curiosity.

**3-) Literature Review:**

I would like to give priority to the relevant article of the authors who had an impact on the formation of the data set.

ZiÄ™ba, M., Tomczak, J. M., Lubicz, M., & ÅšwiÄ…tek, J. (2013). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Computing.

- Results:

-- Boosted SVM for for imbalanced data gained the Gmean value equal 0.657,

-- Decision rules induced using Boosted SVM as an oracle gained the Gmean value equal 0.648.

They reach these results.

Some of the articles, the writers have used machine learning, deep learning or classification method. As an illustration ;

Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients

https://arxiv.org/abs/1504.04646

**4-) Theoretical Perceptive and Concepts**

My dataset contains these attributes.

1. 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)

2. PRE4: Forced vital capacity - FVC (numeric)

3. PRE5: Volume that has been exhaled at the end of the first second of forced expiration - FEV1 (numeric)

4. PRE6: Performance status - Zubrod scale (PRZ2,PRZ1,PRZ0)

5. PRE7: Pain before surgery (T,F)

6. PRE8: Haemoptysis before surgery (T,F)

7. PRE9: Dyspnoea before surgery (T,F)

8. PRE10: Cough before surgery (T,F)

9. PRE11: Weakness before surgery (T,F)

10. PRE14: T in clinical TNM - size of the original tumour, from OC11 (smallest) to OC14 (largest) (OC11,OC14,OC12,OC13)

11. PRE17: Type 2 DM - diabetes mellitus (T,F)

12. PRE19: MI up to 6 months (T,F)

13. PRE25: PAD - peripheral arterial diseases (T,F)

14. PRE30: Smoking (T,F)

15. PRE32: Asthma (T,F)

16. AGE: Age at surgery (numeric)

17. Risk1Y: 1 year survival period - (T)rue value if died (T,F)

Class Distribution: the class value (Risk1Y) is binary valued.

## A. WEKA Observations

### *a)* Methodlogy & Findings

#### 1-) Random Tree:

I use this algorithm and try to reach exact correct answer when depth is increasing, size also increasing. Dataset uses averaging to improve the predictive accuracy and control over-fitting. I used k values as 2,3 and 5.

```
Size of the tree : 12
Max depth of tree: 2

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         397               84.4681 %
Kappa statistic                          0.0435
Mean absolute error                      0.2468
Root mean squared error                  0.3605
Relative absolute error                 96.9095 %
Root relative squared error            101.2562 %
Total Number of Instances              470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Are
                 0,043    0,015    0,333      0,043    0,076      0,072    0,547     0,191
                 0,985    0,957    0,855      0,985    0,915      0,072    0,547     0,858
Weighted Avg.    0,845    0,817    0,777      0,845    0,790      0,072    0,547     0,759

=== Confusion Matrix ===

   a   b   <-- classified as
   3  67 |   a = T
   6 394 |   b = F
```

Figure 1-for k=2

```
Size of the tree : 30
Max depth of tree: 3

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         392               83.4043 %
Kappa statistic                          0.0398
Mean absolute error                      0.2433
Root mean squared error                  0.3713
Relative absolute error                 95.5328 %
Root relative squared error            104.2982 %
Total Number of Instances              470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Ar
                 0,057    0,030    0,250      0,057    0,093      0,053    0,570     0,207
                 0,970    0,943    0,855      0,970    0,909      0,053    0,570     0,870
Weighted Avg.    0,834    0,807    0,765      0,834    0,787      0,053    0,570     0,771

=== Confusion Matrix ===

   a   b   <-- classified as
   4  66 |   a = T
  12 388 |   b = F
```

Figure 2-for k=3

```
Size of the tree : 68
Max depth of tree: 5

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         378               80.4255 %
Kappa statistic                          0.0051
Mean absolute error                      0.2576
Root mean squared error                  0.4256
Relative absolute error                101.1592 %
Root relative squared error            119.5532 %
Total Number of Instances              470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Ar
                 0,071    0,068    0,156      0,071    0,098      0,006    0,462     0,147
                 0,933    0,929    0,852      0,933    0,890      0,006    0,462     0,831
Weighted Avg.    0,804    0,800    0,748      0,804    0,772      0,006    0,462     0,729

=== Confusion Matrix ===

   a   b   <-- classified as
   5  65 |   a = T
  27 373 |   b = F
```

Figure 3-for k=5

#### 2-) KNN:

KNN algorithm find the smallest distances between queries. I tried to find the most consistent results by changing the k's with the KNN method. I also wanted to examine the function of k in the KNN method. I have examined that the value of k=1 is rather insufficient. The value of k=1, which means the closest value, is completely dependent on chance and does not provide us with the correct data.

After k=15, my accuracy rate did not change further. I found that k values after k=15 are in such a way that they do not affect the results of my dataset and they are in the minority. I used for k=1, 3, 15.

```
IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         363               77.234  %
Kappa statistic                          0.0275
Mean absolute error                      0.2289
Root mean squared error                  0.476
Relative absolute error                 89.8978 %
Root relative squared error            133.7021 %
Total Number of Instances              470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,143    0,118    0,175      0,143    0,157      0,028    0,513     0,153     T
                 0,803    0,857    0,055      0,803    0,860      0,028    0,513     0,854     F
Weighted Avg.    0,772    0,747    0,754      0,772    0,763      0,028    0,513     0,750

=== Confusion Matrix ===

   a   b   <-- classified as
  10  60 |   a = T
  47 353 |   b = F
```

Figure 4-for k=1

```
IB1 instance-based classifier
using 3 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         388               82.5532 %
Kappa statistic                          0.0709
Mean absolute error                      0.2168
Root mean squared error                  0.3927
Relative absolute error                 85.114  %
Root relative squared error            110.3053 %
Total Number of Instances              470

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,100    0,048    0,269      0,100    0,146      0,082    0,556     0,174     T
                 0,953    0,900    0,858      0,953    0,903      0,082    0,556     0,866     F
Weighted Avg.    0,826    0,773    0,770      0,826    0,790      0,082    0,556     0,763

=== Confusion Matrix ===

   a   b   <-- classified as
   7  63 |   a = T
  19 381 |   b = F
```

Figure 5-for k=3

```
IB1 instance-based classifier
using 15 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         399               84.8936 %
Kappa statistic                          -0.0042
Mean absolute error                       0.2239
Root mean squared error                   0.3596
Relative absolute error                  87.9355 %
Root relative squared error             101.0044 %
Total Number of Instances               470

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC A
               0,000    0,003    0,000      0,000    0,000      -0,019   0,614     0,19
               0,998    1,000    0,851      0,998    0,918      -0,019   0,614     0,88
Weighted Avg.  0,849    0,851    0,724      0,849    0,782      -0,019   0,614     0,78

=== Confusion Matrix ===

   a   b   <-- classified as
   0  70 |   a = T
   1 399 |   b = F
```

Figure 6-for k=15

```
Correctly Classified Instances         375               79.7872 %
Kappa statistic                          0.0777
Mean absolute error                      0.2479
Root mean squared error                  0.4081
Relative absolute error                 97.3369 %
Root relative squared error            114.6275 %
Total Number of Instances              470

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,157    0,090    0,234      0,157    0,188      0,080    0,555     0,185     T
               0,910    0,843    0,861      0,910    0,885      0,080    0,555     0,869     F
Weighted Avg.  0,798    0,731    0,767      0,798    0,781      0,080    0,555     0,767

=== Confusion Matrix ===

   a   b   <-- classified as
  11  59 |   a = T
  36 364 |   b = F
```

Figure 8-learning rate 0.05, executed in 7.71 sec, 79.78 correctness.

**3-) Multilayer Perceptron:**

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to mean any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptron's (with threshold activation) Learning rate which is known as Turing Parameter, is used in statistics and machine learnings. That determines the step size at each iteration while moving toward a minimum of a loss function

I used learning rate 4 different values.

```
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,243    0,075    0,362      0,243    0,291      0,199    0,584     0,219     T
               0,925    0,757    0,875      0,925    0,899      0,199    0,584     0,878     F
Weighted Avg.  0,823    0,656    0,798      0,823    0,809      0,199    0,584     0,780

=== Confusion Matrix ===

   a   b   <-- classified as
  17  53 |   a = T
  30 370 |   b = F
```

Figure 7-learning rate 0.5, executed in 4.62 sec, 82.34 correctness.

*B.* Conclusion

As a conclusion all of dataset and used methots show that Thoracic Surgery can be ac chance to the lung cancers people. Methods which I used in this article, can change to change datasets. I used a lot of methods and according to number of my instances best is changed. We cannot say a method is the best but we can say some methods are faster the others according to my observation. Also, F1-Score important to comperison of accuracy. Because F1-Score shows harmonic average of precision and recall.

REFERENCES


[1] https://www.ieee.org/conferences/publishing/templates.html

[2] https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data#

[3] ZiÄ™ba, M., Tomczak, J. M., Lubicz, M., & ÅšwiÄ…tek, J. (2013). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. Applied Soft Computing.

[4] Performance Evaluation of Machine Learning Algorithms in Post-operative Life Expectancy in the Lung Cancer Patients(https://arxiv.org/abs/1504.04646)