APA- Practical Work 2017-2018

Albert Ribes Kerstin Winter

November 15, 2017

Contents

Todo list

1 Introduction

1.1 Desciption of the work and its goals

The goal of this project is to build a classification model to predict whether a lung cancer patient will die within one year after surgery or not. To do so we will study a dataset with real lung cancer patients.

As this is very sensitive information, our priority will be to minimize the amount of false negatives, i. e, avoid predicting a patient will not die within one year when it certainly does.

The data is taken from `https://archive.ics.uci.edu/ml/datasets/Thoracic+Surgery+Data#` [1]

1.2 Desciption of available data

The data we are working with is about patients who underwent major lung resections for primary lung cancer in the years from 2007 to 2011. For each patient we are given information about his diagnosis and effects produced by the cancer.

The dataset is very limited in the number of instances available: it only has 470. In addition, the distribution of the predicted class isn't quite balanced, since only 70 of the patients died in one year period. This may become a problem in some of the prediction models due to the fact that the results will be biased towards the biggest class. However, we can suppose that the data has been collected uniformly and that this proportion is similar to the real one.

For each patient we have 16 different atributes. 3 of them are numerical, and the rest are categorical. From those, 10 are binary. The response atribute is also binary.

2 Related Previous Work

3 Data exploration process

3.1 Pre-processing

3.1.1 Treatment of missing values

Our dataset do not have missing values, so there is no need to treat them.

3.1.2 Treatment of anomalous values

Quizá hay que quitar algunas personas por ser demasiado jóvenes comparadas con el resto

3.1.3 Treatment of incoherent values

Redactar que FEV1 está mal, referenciar algún artículo que hable sobre el tema y para justificar que está mal, decidir qué haremos con esos pacientes (eliminarlos, o inferir sus valores de FEV1 en función de sus vecinos) y si los inferimos poner el proceso como lo hemos hecho

El FEV1 tiene valores incoherentes. La mayoría están sobre 3, pero algunos están sobre 60

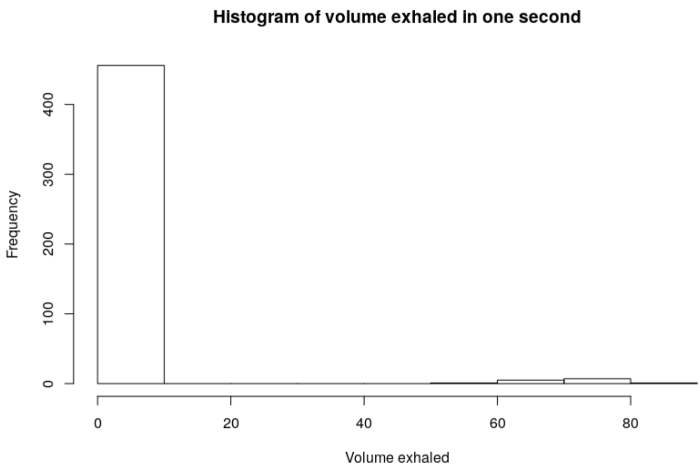

Figure 1: Show the ammount of people having each value

3.1.4 Coding of non-continuous or non-ordered variables

3.1.5 Possible elimination of irrelevant variables

Algunas variables están muy poco representadas: Haremos los experimentos dos veces, uno con todos los datos originales, y otro quitando el atributo de MI, ASHTMA, DGN1 y DGN8, y entonces veremos cual da mejores resultados

- DGN: Solo un paciente tiene DGN1, y solo 2 tienen DGN8
- Solo 2 pacientes tienen MI == True
- Solo 8 pacientes tienen PAD == True
- Solo 2 pacientes tienen ASHTMA == True

3.1.6 Creation of new useful variables (Feature extraction)

Entender cómo funciona MCA, y ver si podemos sacar una variable nueva (Quizá)Categorizar nuestras variables numéricas,

3.1.7 Normalization of the variables

Solo se pueden normalizar FVC, FEV1 y AGE. Miraremos cuales dan mejores resultados

3.1.8 Transformation of the variables

Skewness es la asimetría de los datos respecto a la media. Como la mayoría de nuestras variables son categóricas, no tiene mucho sentido medir el skewness, ni tampoco corregirlo Kurtosis igual que el skewness, no es necesario porque la mayoría con categóricas

El código en r para corregir la asimetría (skewness) de los datos

3.2 Clustering

hacer varios k-means con distintos valores de k (2,3,4,5,6) para ver si descubrimos algún cluster que nos permita crear una variable nueva

3.3 Visualization

Hacer MCA

4 Resampling protocol

5 Results obtained using linear/quadratic methods

5.1 LDA

Suponiendo que las varianzas de cada una de las clases son la misma, se usa este algoritmo, (que simplifica QLA) para ver la probabilidad de pertenencia a una clase

Mirar el vecino más cercano para precedir

Si suponemos que las variables son independientes: -Haces naive Bayes para ver la probabilidad de que pertenezca a cada una de las clases (habría que estudiar si las variables son independientes) - Logistic regression

6 Results obtained using non-linear methods

7 Desciption and justification of the final model chosen

7.1 Estimation of the generalization error

8 Self-assessment of successes, failures and doubts

9 Scientific and personal conclusions

10 Possible extensions and known limitations

References

[1] Maciej Zikeba et al. “Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients”. In: *Applied Soft Computing* (2013). DOI: [WebLink].