Tipologia i cicle de vida de les dades – Pràctica 2 – 2017/18

Carles Figuera (*cfiguerap*) – Màster en Ciències de Dades

1.- Descripció del dataset. Perquè es important i quina pregunta/problema pretén respondre?

Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic)

Per a la realització d'aquesta pràctica he escollit el dataset de viatgers del Titanic, el famós transatlàntic enfonsat l'any 1912 on van perdre la vida uns 1.500 passatgers.

El dataset està format per dos fitxers: un d'entrenament pel nostre model i un de testing. Els camps del dataset descarregat de *Kaggle* son els següents:

- **PassengerId**: número de passatger
- **Survived**: va sobreviure o no a l'accident (0, 1)
- **Pclass**: classe on viatjava el passatger: primera, segona o tercera (1, 2, 3)
- **Name**: nom del passatger
- **Sex**: sexe del passatger (male, female)
- **Age**: edat del passatger (anys)
- **SibSp**: nombre de germans o parelles sentimentals que viatjaven al vaixell
- **Parch**: nombre de pares o fills que viatjaven al vaixell
- **Ticket**: número de bitllet
- **Fare**: tarifa del bitllet
- **Cabin**: número de cabina
- **Embarked**: port d'embarcament (C: Cherbourg, Q: Queenstown, S: Southampton)

El problema que es pretén resoldre es trobar si existia alguna o més característiques entre els passatgers del vaixell que podrien predeterminar el fet de morir o no en l'accident que va tenir lloc l'any 1912.

2.- Integració i selecció de les dades d'interès a analitzar.

De les dades que conté el nostre dataset, creiem que no son d'interès pel nostre problema i seran eliminades les següents:

- **PassengerId**: identificador que no aporta informació per al nostre cas d'estudi.
- **Name**: identificador que no aporta informació per al nostre cas d'estudi.
- **Ticket**: identificador que no aporta informació per al nostre cas d'estudi.
- **Fare**: la tarifa pagada deu estar relacionat amb la classe del trajecte, així que prescindim d'aquest valor.
- **Embarked**: no crec rellevant en quin port es va pujar el passatger.

```
titanic$PassengerId <- NULL
titanic$Name <- NULL
titanic$Ticket <- NULL
titanic$Fare <- NULL
titanic$Embarked <- NULL
```

D'aquesta manera, les dades que farem servir en el nostre cas d'estudi serà el següent:

- **Survived**: va sobreciure o no a l'accident (0, 1)
- **Pclass**: classe on viatjava el passatger: primera, segona o tercera (1, 2, 3)
- **Sex**: sexe del passatger (male, female)
- **Age**: edat del passatger (anys)
- **SibSp**: nombre de germans o parelles sentimentals que viatjaven al vaixell
- **Parch**: nombre de pares o fills que viatjaven al vaixell
- **Cabin**: número de cabina

3.- Neteja de les dades

3.1.- Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

En el nostre cas, podem observar com el camp AGE, que representa l'edat dels viatgers, conté 177 dades buides. És a dir, dels 891 registres que tenim al dataset d'entrenament, no sabem l'edat de 177 passatgers.

```
> summary(titanic)
    Survived         Pclass          Sex           Age            SibSp           Parch                Cabin     Embarked
 Min.   :0.0000   Min.   :1.000   female:314   Min.   : 0.42   Min.   :0.000   Min.   :0.0000             :687    :  2
 1st Qu.:0.0000   1st Qu.:2.000   male  :577   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000   B96 B98    :  4   C:168
 Median :0.0000   Median :3.000                Median :28.00   Median :0.000   Median :0.0000   C23 C25 C27:  4   Q: 77
 Mean   :0.3838   Mean   :2.309                Mean   :29.70   Mean   :0.523   Mean   :0.3816   G6         :  4   S:644
 3rd Qu.:1.0000   3rd Qu.:3.000                3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000   C22 C26    :  3
 Max.   :1.0000   Max.   :3.000                Max.   :80.00   Max.   :8.000   Max.   :6.0000   D          :  3
                                               NA's   :177                                      (Other)    :186
```

Hem de decidir com gestionar aquests registres que contenen valors desconeguts. Una opció pot ser eliminar aquests registres però això suposaria desaprofitar informació. Si substituïm aquests valors per zeros estem afegint valors extrems al dataset que poden alterar el resultat de l'estudi. Així que crec millor afegir el valor mig d'aquesta columna al registres que no tinguin valor.

```
titanic$Age =
ifelse(is.na(titanic$Age),
ave(titanic$Age, FUN = function(x) mean(x, na.rm = TRUE)),
titanic$Age)
```

També s'observen faltes d'informació al camp CABIN, d'on falten 687 dades de les 891. Al ser un nombre tant elevat tampoc podem eliminar els registres ni posar un valor mitjà. En aquest cas, segurament el millor és prescindir d'aquesta columna.

```
titanic$Cabin <- NULL
```

3.2.- Identificació i tractament de valors extrems.

Si observem els rangs de valors dels camps numèrics del dataset, no s'observen valors mínims ni màxims impossibles.

```
> summary(titanic)
    Survived          Pclass          Age             SibSp           Parch
 Min.   :0.0000   Min.   :1.000   Min.   : 0.42   Min.   :0.000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:22.00   1st Qu.:0.000   1st Qu.:0.0000
 Median :0.0000   Median :3.000   Median :29.70   Median :0.000   Median :0.0000
 Mean   :0.3838   Mean   :2.309   Mean   :29.70   Mean   :0.523   Mean   :0.3816
 3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:35.00   3rd Qu.:1.000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :3.000   Max.   :80.00   Max.   :8.000   Max.   :6.0000
```

4.- Anàlisi de dades

4.1.- Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

En el nostre cas d'estudi, on volem esbrinar les característiques que van decidir la mortalitat dels passatgers del Titànic, podríem crear diferents subgrups d'estudis. Per exemple: estudiar la mortalitat per edat, per classe de viatger, per sexe.

En principi, pel nostre cas d'estudi, no farem distinció de casos a menys que els resultats obtinguts no siguin conclusius. En aquest cas, hauríem d'estudiar per subgrups buscant aquells factors determinants que van decidir la mortalitat dels passatgers.

4.2.- Comprovació de la normalitat i homogeneïtat de la variància.

Distribució normal:

Amb la prova d'Anderson-Darling podem mesurar què tan bé segueixen les dades una distribució específica. Per a un conjunt de dades i distribució en particular, mentre millor s'ajusti la distribució a les dades, menor serà aquest valor, normalment un valor acceptable és de 0.05.

```
library(nortest)
alpha = 0.05
cat("No normal distribution: ")
for (i in 1:ncol(titanic)) {
    if (is.integer(titanic[,i])) {
        p_val = ad.test(titanic[,i])$p.value
        if (p_val < alpha) {
            cat(colnames(titanic)[i])
            cat(", ")
        }
    }
}
```

```
No normal distribution: Survived, Pclass, SibSp, Parch,
```

Homogeneïtat de la variància

Per esbrinar si les dades numèriques son normalitzades apliquem el test de Shapiro Wilk:

```
> shapiro.test(titanic$Pclass)

        data:  titanic$Pclass
W = 0.71833, p-value < 2.2e-16

> shapiro.test(titanic$Age)

        data:  titanic$Age
W = 0.95882, p-value = 3.969e-15

> shapiro.test(titanic$SibSp)

        data:  titanic$SibSp
W = 0.51297, p-value < 2.2e-16

> shapiro.test(titanic$Parch)

        data:  titanic$Parch
W = 0.53281, p-value < 2.2e-16
```

Els valors resultants indiquen que les dades no son normalitzades ja que $p < 0.05$.

4.3.- Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc.

Donat el cas d'estudi, crec que el més convenient seria aplicar un model d'arbre de decisió. D'aquesta manera sabrem quines són les característiques que tenen més pes a l'hora de decidir la mortalitat.

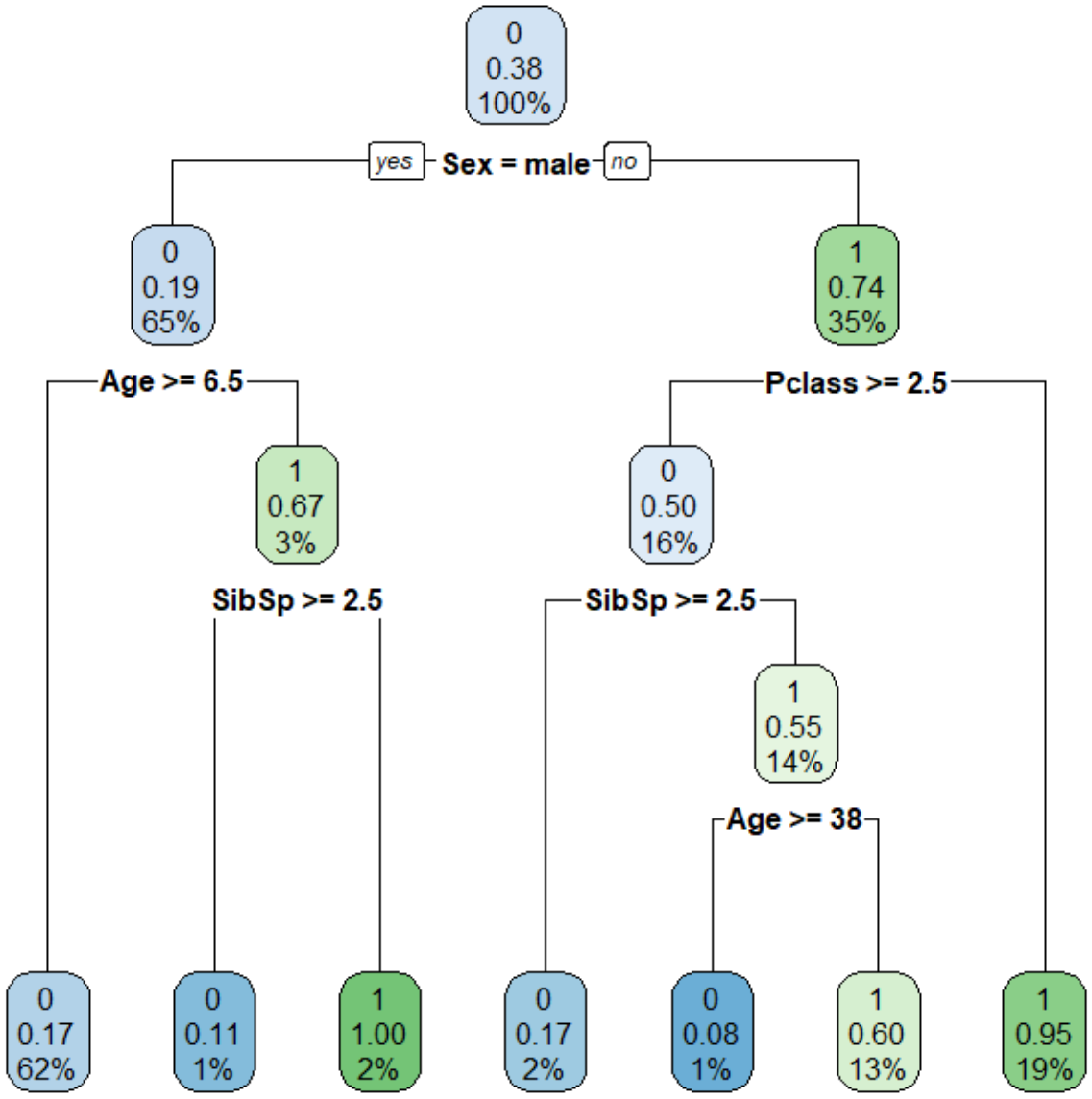
```
> arbre <- rpart(Survived ~ ., method="class", data = titanic)

> print(arbre)
n= 891

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 891 342 0 (0.61616162 0.38383838)
   2) Sex=male 577 109 0 (0.81109185 0.18890815)
     4) Age>=6.5 553  93 0 (0.83182640 0.16817360) *
     5) Age< 6.5 24   8 1 (0.33333333 0.66666667)
      10) SibSp>=2.5 9   1 0 (0.88888889 0.11111111) *
      11) SibSp< 2.5 15   0 1 (0.00000000 1.00000000) *
   3) Sex=female 314  81 1 (0.25796178 0.74203822)
     6) Pclass>=2.5 144  72 0 (0.50000000 0.50000000)
      12) SibSp>=2.5 18   3 0 (0.83333333 0.16666667) *
      13) SibSp< 2.5 126  57 1 (0.45238095 0.54761905)
        26) Age>=38.5 12   1 0 (0.91666667 0.08333333) *
        27) Age< 38.5 114  46 1 (0.40350877 0.59649123) *
     7) Pclass< 2.5 170   9 1 (0.05294118 0.94705882) *
```

Podem veure que el SEXE és el camp que més importància té en el nostre cas d'estudi.

Per part dels homes, els següents camps més importants respectivament son l'EDAT i l'SIBSP. Que volia dir el nombre de germans o parelles que viatjaven amb ells. Per part de les dones, és al revès dels homes, els següents camps més importants respectivament son l'SIBSP i l'EDAT.

5. Representació dels resultats a partir de taules i gràfiques.

Si representem gràficament el nostre arbre de decisió, seria el següent, on 0 representa la MORT i 1 la SUPERVIVENCIA.



6. Resolució del problema. A partir dels resultats obtinguts, quines son les conclusions? Els resultats permeten respondre al problema?

Podem observar veient l'arbre les següents argumentacions:

- Del 100% dels passatgers del dataset, van morir el 38%.
  - El 65% dels passatgers eren homes i van morir el 19%.
    - Dels homes morts, el 17% eren menors de 6 anys.
    - Dels homes supervivents, el 100% anaven acompanyats amb 3 o més germans.

- El 25% dels passatgers eren dones i van morir el 26%.
  - De les dones que van sobreviure, el 95% viatjaven en 3a classe.
  - De les dones mortes, el 60% eren majors de 38 anys.

És a dir, com a conclusió podem extreure que els viatgers amb més possibilitat de supervivència eren les dones que viatjaven en 3a classe.

Així doncs podem extreure les següents regles principals:

- Si (SEX = 'Home') llavors Sobreviu = NO
- Si (SEX = 'Dona') & (CLASS = 3a) llavors Sobreviu = SI
- Si (SEX = 'Dona') & (CLASS != 3a) llavors Sobreviu = NO

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.

https://github.com/referup-cfiguera/uoc-cleaning-data