

Pràctica2: Neteja i Validació de les Dades

Joaquim Dalmases i Juanjo Díez

5 de junio, 2019

Contents

1 Introducció.	2
1.1 Presentació.	2
1.2 Competències.	2
1.3 Objectius.	2
2 Resolució.	3
2.1 Descripció del dataset.	3
2.2 Integració i selecció de les dades d'interés a analitzar.	6
2.3 Neteja de les dades.	8
2.4 Anàlisi de les dades.	8
2.5 Representació dels resultats a partir de taules i gràfiques.	8
2.6. Resolució del problema i conclusions.	8
3 Recursos	9

1 Introducció.

1.1 Presentació.

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'integració, neteja, validació i anàlisi de les mateixes. Per fer aquesta pràctica haureu de treballar en grups de 2 persones. Haureu de lliurar un sol fitxer amb l'enllaç Github on es trobin les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius que corresponen a la vostra entrega. Cada membre de l'equip haurà de contribuir amb el seu usuari Github.

1.2 Competències.

En aquesta pràctica es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per abordar-lo i resoldre'l.
- Capacitat per aplicar les tècniques específiques de tractament de dades (integració, transformació, neteja i validació) per al seu posterior anàlisi. Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

1.3 Objectius.

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució de problemes en entorns nous o poc coneguts dintre de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants i els tractaments necessaris (integració, neteja i validació) per dur a terme un projecte analític.
- Aprendre a analitzar les dades adequadament per abordar la informació continguda en les dades.
- Identificar la millor representació dels resultats per tal d'aportar conclusions sobre el problema plantejat en el procés analític.
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar les habilitats d'aprenentatge que els permetin continuar estudiant d'una manera que haurà de ser en gran manera autodirigida o autònoma.

2 Resolució.

Aquesta pràctica s'ha desenvolupat seguint la bibliografia recomanada: (Calvo M 2019; Squire 2015; Jiawei Han 2012; Dalgaard 2008)

2.1 Descripció del dataset.

Per l'elaboració de la pràctica s'ha triat:

- el repositori de *Kaggle Red Wine Quality*
- que correspon amb el repositori de *UCI Wine Quality Data Set* i
- l'accés a les dades completes es pot trobar a *aquest enllaç*.

2.1.1 Càrrega de dades

```
# Fixem el directori de treball:
setwd("C:/Users/juanj/OneDrive/Documentos/GitHub/Practica2")

# Llegim els fitxers amb les dades de vins blancs i negres
# Ho ubiquem a dos datasets dsRed i dsWhite.
redFile <- "winequality-red.csv"
whiteFile <- "winequality-white.csv"
dsRed <- read.csv(file.path(getwd(), redFile), sep=";", encoding="UTF-8")
dsWhite <- read.csv(file.path(getwd(), whiteFile), sep=";", encoding="UTF-8")
# Observem que els fitxers originals tenen iguals capçaleres.

# Comprobació de la bona lectura/transferència de dades, mirem les dues primeres fileres
# de cada dataset i vegem la composició .
head(dsRed, 2)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1          7.4           0.70         0           1.9       0.076
## 2          7.8           0.88         0           2.6       0.098
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                11                34 0.9978 3.51      0.56      9.4
## 2                25                67 0.9968 3.20      0.68      9.8
##   quality
## 1        5
## 2        5
```

```
head(dsWhite, 2)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1          7.0           0.27         0.36          20.7       0.045
## 2          6.3           0.30         0.34           1.6       0.049
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                45                170 1.001 3.0      0.45      8.8
## 2                14                132 0.994 3.3      0.49      9.5
##   quality
## 1        6
## 2        6
```

summary(dsRed)

```
## fixed.acidity    volatile.acidity    citric.acid    residual.sugar
## Min.   : 4.60      Min.   :0.1200    Min.   :0.000    Min.   : 0.900
## 1st Qu.: 7.10      1st Qu.:0.3900    1st Qu.:0.090    1st Qu.: 1.900
## Median : 7.90      Median :0.5200    Median :0.260    Median : 2.200
## Mean   : 8.32      Mean   :0.5278    Mean   :0.271    Mean   : 2.539
## 3rd Qu.: 9.20      3rd Qu.:0.6400    3rd Qu.:0.420    3rd Qu.: 2.600
## Max.   :15.90      Max.   :1.5800    Max.   :1.000    Max.   :15.500
## chlorides        free.sulfur.dioxide    total.sulfur.dioxide
## Min.   :0.01200    Min.   : 1.00      Min.   : 6.00
## 1st Qu.:0.07000    1st Qu.: 7.00      1st Qu.:22.00
## Median :0.07900    Median :14.00      Median :38.00
## Mean   :0.08747    Mean   :15.87      Mean   :46.47
## 3rd Qu.:0.09000    3rd Qu.:21.00      3rd Qu.:62.00
## Max.   :0.61100    Max.   :72.00      Max.   :289.00
## density          pH          sulphates          alcohol
## Min.   :0.9901    Min.   :2.740    Min.   :0.3300    Min.   : 8.40
## 1st Qu.:0.9956    1st Qu.:3.210    1st Qu.:0.5500    1st Qu.: 9.50
## Median :0.9968    Median :3.310    Median :0.6200    Median :10.20
## Mean   :0.9967    Mean   :3.311    Mean   :0.6581    Mean   :10.42
## 3rd Qu.:0.9978    3rd Qu.:3.400    3rd Qu.:0.7300    3rd Qu.:11.10
## Max.   :1.0037    Max.   :4.010    Max.   :2.0000    Max.   :14.90
## quality
## Min.   :3.000
## 1st Qu.:5.000
## Median :6.000
## Mean   :5.636
## 3rd Qu.:6.000
## Max.   :8.000
```

summary(dsWhite)

```
## fixed.acidity    volatile.acidity    citric.acid    residual.sugar
## Min.   : 3.800    Min.   :0.0800    Min.   :0.0000    Min.   : 0.600
## 1st Qu.: 6.300    1st Qu.:0.2100    1st Qu.:0.2700    1st Qu.: 1.700
## Median : 6.800    Median :0.2600    Median :0.3200    Median : 5.200
## Mean   : 6.855    Mean   :0.2782    Mean   :0.3342    Mean   : 6.391
## 3rd Qu.: 7.300    3rd Qu.:0.3200    3rd Qu.:0.3900    3rd Qu.: 9.900
## Max.   :14.200    Max.   :1.1000    Max.   :1.6600    Max.   :65.800
## chlorides        free.sulfur.dioxide    total.sulfur.dioxide
## Min.   :0.00900    Min.   : 2.00      Min.   : 9.0
## 1st Qu.:0.03600    1st Qu.:23.00      1st Qu.:108.0
## Median :0.04300    Median :34.00      Median :134.0
## Mean   :0.04577    Mean   :35.31      Mean   :138.4
## 3rd Qu.:0.05000    3rd Qu.:46.00      3rd Qu.:167.0
## Max.   :0.34600    Max.   :289.00      Max.   :440.0
## density          pH          sulphates          alcohol
## Min.   :0.9871    Min.   :2.720    Min.   :0.2200    Min.   : 8.00
## 1st Qu.:0.9917    1st Qu.:3.090    1st Qu.:0.4100    1st Qu.: 9.50
## Median :0.9937    Median :3.180    Median :0.4700    Median :10.40
## Mean   :0.9940    Mean   :3.188    Mean   :0.4898    Mean   :10.51
```

```
## 3rd Qu.:0.9961 3rd Qu.:3.280 3rd Qu.:0.5500 3rd Qu.:11.40
## Max. :1.0390 Max. :3.820 Max. :1.0800 Max. :14.20
## quality
## Min. :3.000
## 1st Qu.:5.000
## Median :6.000
## Mean :5.878
## 3rd Qu.:6.000
## Max. :9.000
```

Perquè és important i quina pregunta/problema pretén respondre?

El dataset ‘Red Wine’ emmagatzema les característiques físico-químiques de les mostres de vi blanc i negre junt amb el ratio de la qualitat otorgada, en una escala de 0 a 10. Conté 1599 mostres de vi negre de la zona nord de Portugal.

Cada mostra de vi té assignada un valor de qualitat resultats de proves realitzades en la seva composició (tests de quantitat d'alcohol, nivell d'acidesa, contingut residual de sucres etc...). En total són 12 atributs descrivint característiques entre físico-químiques i la classificació de qualitat de la mostra.

Empreurem aquest dataset per respondre a la pregunta de quines característiques principals defineixen un vi de qualitat?, Varien si es tracta d'un vi negre o blanc?.

Descripció dels atributs o camps del datatset:

Atribut	Traducció	Descripció
fixed.acidity	<i>Acidesa fixe</i>	És la quantitat d'àcidesa que no s'evapora i per tant resta fixe al vi.
volatile.acidity	<i>Acidesa volàtil</i>	La quantitat en excès d'àcid acètic en vi, pot afegir sabor amarg o avinagrat, si les quantitats són altes.
citric.acid	<i>Àcid cítric</i>	Trobat en petites quantitats, l'àcid cítric pot afegir frescor i sabor als vins.
residual.sugar	<i>Sucre residual</i>	Quantitat de sucre derivada del procés de fermentació (normalment trobem més de 1 gr/litre i si supera els 45grm./litre considerem el vi dolç.
chlorides	<i>Clorurs</i>	La quantitat de sal del vi.
free.sulfur.dioxide	<i>Diòxid de sofre</i>	Prevé el creixement microbià i l'oxidació del vi (anti-oxidant). Els vins blancs mantenen millor l'aspecte de vi jove. La normativa de la Comunitat Europea obliga des de l'any 2005 que qualsevol aliment o beguda que contingui més de 10 mg/l de sulfits ha de portar-ho en l'etiqueta com advertència. El motiu és que aquest additiu té capacitat al·lèrgica, és a dir, un petit percentatge de la població pot ser sensible o al·lèrgic als sulfits.
total.sulfur.dioxide	<i>diòxid de sofre total</i>	Concentracions per sobre de 50 ppm (tant lliure com unit), Es detecta per olfacte i tast. Les quantitats excessives de SO2 poden inhibir la fermentació i causar efectes sensorials indesitjables.
density	<i>Densitat</i>	Serà propera a la de l'aigua (997 kg/m ³) i variaria segons les quantitats de sucre i alcohol, segons la qualitat de la fermentació.
pH	<i>pH</i>	Descrui com un vi àcid o bàsic és a una escala de 0 (molt àcida) a 14 (molt bàsica); la majoria dels vins tenen entre 3-4 a l'escala de pH.

Atribut	Traducció	Descripció
sulphates	<i>Sulfats</i>	Un additiu de vi que pot contribuir als nivells de diòxid de sofre (SO ₂), que actuen com a antimicrobians i antioxidants
alcohol	<i>Alcohol</i>	El percentatge de contingut alcohòlic del vi, és una variable de sortida (basada en dades sensorials)
quality	<i>Qualitat</i>	(escala 0-10) És la qualitat atorgada la vi.
color	<i>Color</i>	Determina si el vi és blanc o negre. Afegida per nosaltres a efectes de integrar les dades.

2.2 Integració i selecció de les dades d'interés a analitzar.

Disposem de dos fitxers de dades un que conté les característiques del vins blancs i l'altre dels vins negres, per tant ens interesera comprovar que tenen les mateixes capçalers i que els podem integrar en un sol dataset. A més per tal de no perdre informació en la integració afegirem una columna 'color' que identificara la font de les files o mostres emmagatzemant el color del vi amb valors (blanc)

```
# Volem analitzar el dataset de Red Wine tenint en compte el color del vi,
# afegim una columna 'color' i fusionem les dades tant dels vins blancs com
# dels negres, diferenciant-los per el camp color.

# Afegim el camp 'color' a cada datatset
dsRed["color"]<-"negre"
dsWhite["color"]<-"blanc"

# Tenim capçaleres iguals, si la suma de noms iguals és la suma total de camps.
a<-colnames(dsRed)
b<-colnames(dsWhite)
cat(paste0("El nombre de camps (",ncol(dsRed),") és igual al nombre de camps iguals ",
          sum(a==b),"\n"))
```

```
## El nombre de camps (13) és igual al nombre de camps iguals 13
```

```
cat("Files - Instàncies de vi negre:",nrow(dsRed),"\nColumnes-Atributs-Variables:",
    ncol(dsRed),"\n")
```

```
## Files - Instàncies de vi negre: 1599
## Columnes-Atributs-Variables: 13
```

```
cat("Files - Instàncies de vi blanc:",nrow(dsWhite),"\nColumnes-Atributs-Variables:",
    ncol(dsWhite),"\n")
```

```
## Files - Instàncies de vi blanc: 4898
## Columnes-Atributs-Variables: 13
```

```
# Combinem les mostres dels dos fitxers i factoritzem el camp color per determinar
# els valors que pren: 'blanc i 'negre'
d<-rbind(dsRed,dsWhite)
d$color<-factor(d$color)

# Dimensions del dataset:
cat("Files - Instàncies:",nrow(d),"\nColumnes-Atributs-Variables:",ncol(d),"\n")
```

```
## Files - Instàncies: 6497
## Columnes-Atributs-Variables: 13
```

```
# Revisem l'estructura de camps del dataset:
summary(d)
```

```
## fixed.acidity    volatile.acidity    citric.acid    residual.sugar
## Min.   : 3.800    Min.   :0.0800    Min.   :0.0000    Min.   : 0.600
## 1st Qu.: 6.400    1st Qu.:0.2300    1st Qu.:0.2500    1st Qu.: 1.800
## Median : 7.000    Median :0.2900    Median :0.3100    Median : 3.000
## Mean   : 7.215    Mean   :0.3397    Mean   :0.3186    Mean   : 5.443
## 3rd Qu.: 7.700    3rd Qu.:0.4000    3rd Qu.:0.3900    3rd Qu.: 8.100
## Max.   :15.900    Max.   :1.5800    Max.   :1.6600    Max.   :65.800
## chlorides        free.sulfur.dioxide    total.sulfur.dioxide
## Min.   :0.00900    Min.   : 1.00        Min.   : 6.0
## 1st Qu.:0.03800    1st Qu.: 17.00        1st Qu.: 77.0
## Median :0.04700    Median : 29.00        Median :118.0
## Mean   :0.05603    Mean   : 30.53        Mean   :115.7
## 3rd Qu.:0.06500    3rd Qu.: 41.00        3rd Qu.:156.0
## Max.   :0.61100    Max.   :289.00        Max.   :440.0
## density          pH          sulphates          alcohol
## Min.   :0.9871    Min.   :2.720    Min.   :0.2200    Min.   : 8.00
## 1st Qu.:0.9923    1st Qu.:3.110    1st Qu.:0.4300    1st Qu.: 9.50
## Median :0.9949    Median :3.210    Median :0.5100    Median :10.30
## Mean   :0.9947    Mean   :3.219    Mean   :0.5313    Mean   :10.49
## 3rd Qu.:0.9970    3rd Qu.:3.320    3rd Qu.:0.6000    3rd Qu.:11.30
## Max.   :1.0390    Max.   :4.010    Max.   :2.0000    Max.   :14.90
## quality          color
## Min.   :3.000    blanc:4898
## 1st Qu.:5.000    negre:1599
## Median :6.000
## Mean   :5.818
## 3rd Qu.:6.000
## Max.   :9.000
```

```
# write.csv(d,"Dataset_inicial.csv",row.names = FALSE)
```

Com es pot veure ens quedem amb tots els atributs i més tard en la fase d'anàlisi determinarem si es possible una reducció de camps. Ara per ara podem comptar amb tots els camps disponibles al dataset per esbrinar quins ens determinaran els vins de millor qualitat.

Si revisem les dades de color podem comprovar que els nombres quadren amb els elements dels datasets originals, pel que s'ens reafirma la correcta integració dels dos.

2.3 Neteja de les dades.

2.3.1 Zeros i elements buits.

2.3.2 Identificació i tractament de valors extrems.

2.4 Anàlisi de les dades.

2.4.1 Selecció dels grups de dades i planificació dels anàlisis.

2.4.2 Comprovació de la normalitat i homogeneïtat de la variància.

2.4.3 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. Aplicar almenys tres mètodes d'anàlisi diferents.

2.5 Representació dels resultats a partir de taules i gràfiques.

2.6. Resolució del problema i conclusions.

A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

3 Recursos

- Calvo M, Pérez D, Subirats L. 2019. *Introducción a La Limpieza Y Análisis de Los Datos*. Editorial UOC.
- Dalgaard, Peter. 2008. *Introductory Statistics with R*. Springer Science & Business Media.
- Jiawei Han, Jian Pei, Micheine Kamber. 2012. *Data Mining: Concepts and Techniques*. Morgan Kaufmann.
- Squire, Megan. 2015. *Clean Data*. Packt Publishing Ltd.