Martí Tuneu
Miquel Solé

Pràctica 2 (35% nota final)

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 (https://github.com) on es troben 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. Malgrat que no es tracta del mateix enunciat, els següents exemples d'edicions anteriors us poden servir com a guia: ● Exemple: https://github.com/Bengis/nba-gap-cleaning ● Exemple complex (fitxer adjunt).

Important: si escolliu un nou dataset és interessant que continga una àmplia varietat de dades numèriques i categòriques per poder fer una anàlisi més ric.

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.

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.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

Descripció de la Pràctica a realitzar

L'objectiu d'aquesta activitat serà el tractament d'un dataset, que pot ser el creat a la pràctica 1 o bé qualsevol dataset lliure disponible a Kaggle (https://www.kaggle.com). Alguns exemples de dataset amb els que podeu treballar són:

- Red Wine Quality (https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009).
- Titanic: Machine Learning from Disaster (https://www.kaggle.com/c/titanic).

L'últim exemple correspon a una competició activa a Kaggle de manera que, opcionalment, podeu aprofitar el treball realitzat durant la pràctica per entrar en aquesta competició.

Seguint les principals etapes d'un projecte analític, les diferents tasques a realitzar (i **justificar**) són les següents:

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

Hem escollit el Dataset del Red Wine Quality degut a que és indicat per problemes de classificació o regressió, tal com s'indica a la seva web de Kaggle. Com que alguns dels punts de la pràctica ens demanen de realitzar aquests tipus d'operacions ens ha semblat adient.

El problema a resoldre en aquest cas seria esbrinar quines característiques del vi tenen una major influencia en la qualitat final del vi. En altres paraules, ser capaç de predir la qualitat a partir dels atributs més rellevants.

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

Hem pres tot el conjunt de variables proporcionades pel Dataset, ja que pretenem descobrir les variables que major grau de correlació tenen amb la variable objectiu, Quality. Dit d'una altra manera, esbrinar quins són els atributs amb major poder de predicció respecte a la qualitat. Com

que a priori aquesta era la informació que desconeixíem, no tenia sentit en el nostre cas descartar-ne cap.

3. Neteja de les dades.

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

En el nostre cas no trobem la presencia d'elements buits o nuls en cap variable. En el cas que n'hi trobéssim els podríem omplir mitjançant el mètode dels veïns més propers, fent servir la distancia de Gower.

Pel que fa als 0s, en trobem en la variable "cítric.acid", on veiem que els valors van des de 0 a 1, representant el 0 l'absència d'àcid nítric i l'1 el valor màxim per a vins mol avinagrats. Per tant té sentit trobar tots els valors i no correspon cap tractament en les dades.

3.2. Identificació i tractament de valors extrems.

4. Anàlisi de les dades.

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

Comprovar quines són les variables que major incidència tenen respecte la variable objectiu: quality.

Es realitzarà en primer lloc un anàlisi de correlació entre les variables, per tenir un primer indicador de quines variables estan més correlacionades entre elles i en particular amb la qualitat. Es farà en segon lloc un anàlisi de regressió lineal amb les variables que més relació tenen amb la variable final qualitat, per comprovar quines són les que tenen més incidència. Per últim es pretén fer un anàlisi de predicció de qualitat del vi, dividint-lo en dues categories, "bo" i "dolent", aquest anàlisi es farà amb regressió logística aplicant com a predictors les variables de les quals n'hàgim extret conclusions que ens indiquin una forta relació amb la variable qualitat.

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

Per comprovar la normalitat duem a terme tres passos:

- Histograma per cada una de columnes del dataframe
- Gràfics de QQ per les variables que ens sembla que poden seguir una distribución normal a partir dels histogrames.
- Shapiro test a totes les columnes.

Els resultats d'aquestes proves ens indiquen que cap de les variables segueix una distribució normal.

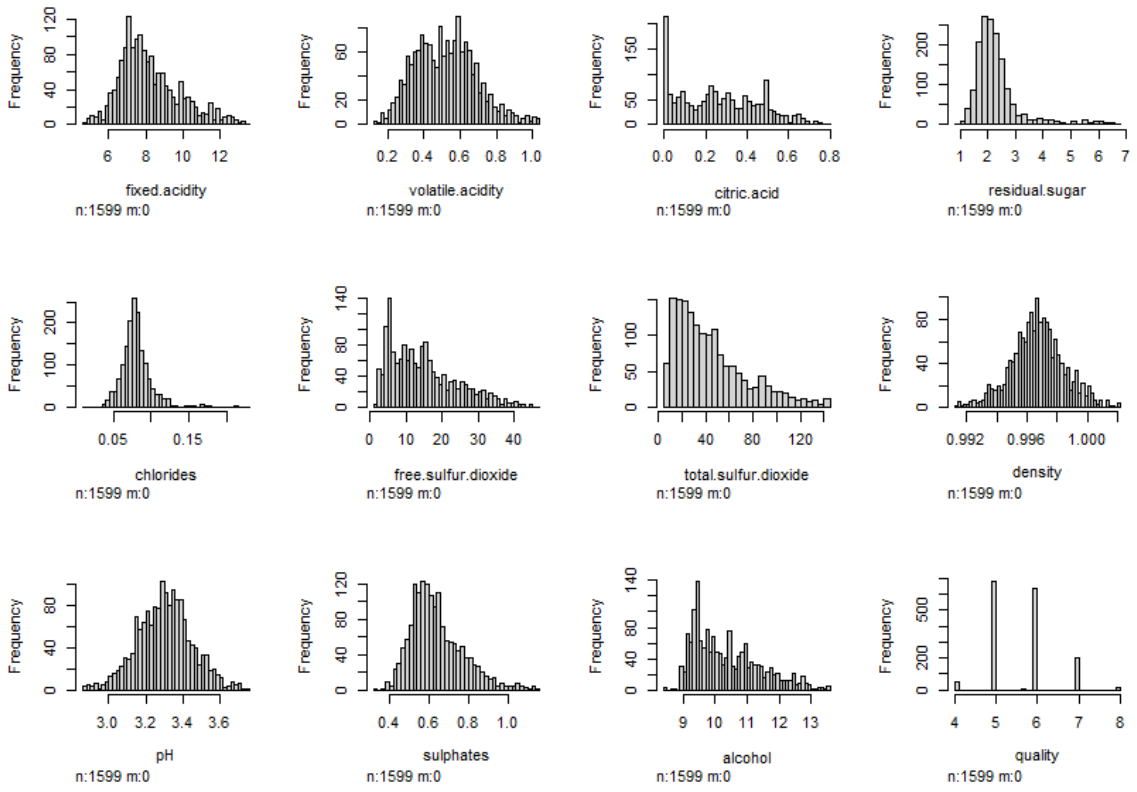
Hem utilitzat el test de Brown-Forsyth per comprovar la homogeneïtat de les variàncies entre les diferents variables a analitzar i tenint en compte els resultats no podem assegurar que les variàncies siguin homogènies.

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.

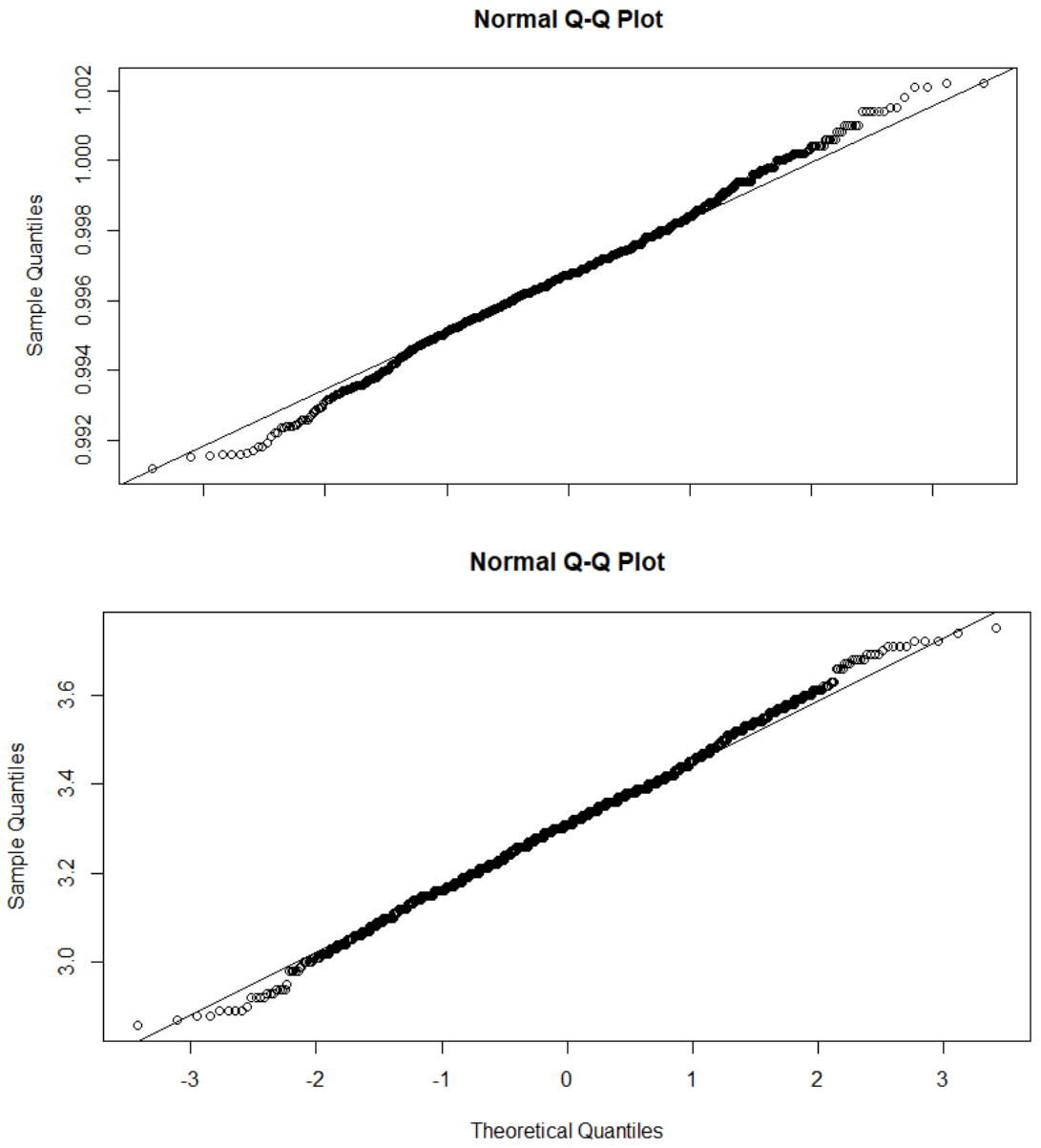
Com s'ha explicat anteriorment les proves aplicades han estat, la correlació entre variables, la regressió lineal i la regressió logística.

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

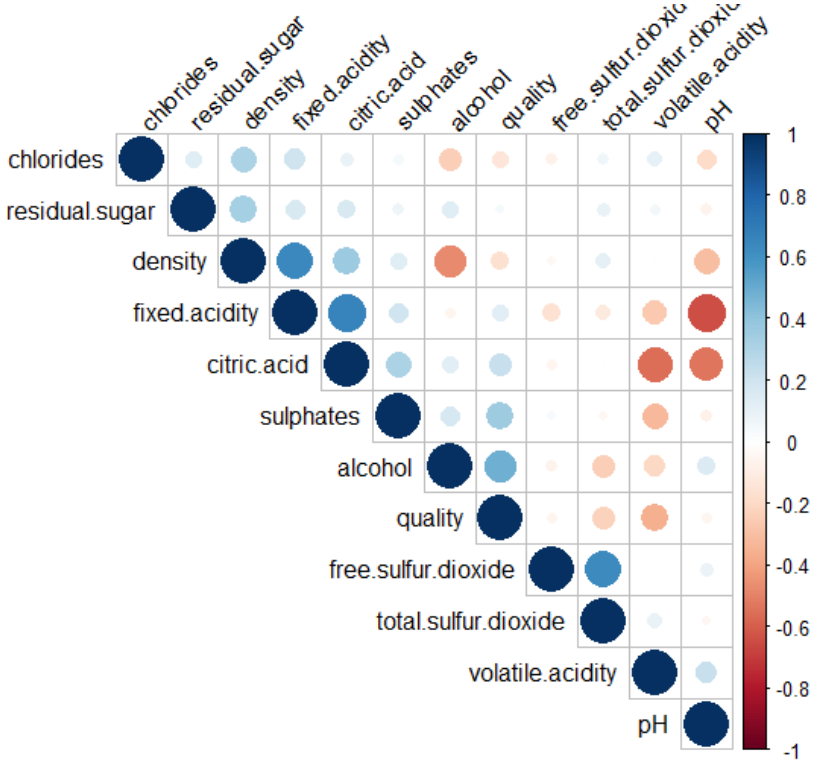
En primer lloc mostrem les gràfiques dels histogrames que ens indiquen que les variables no segueixen una distribució normal.



En segon lloc els gràfics QQ que ens mostren que les variables density i pH no segueixen una distribución normal tot i el que pugui semblar en els seus respectius histogrames. El primer gràfic QQ correspon a la density, mentres que el segon correspon al pH.



Els gràfics que es representen a continuació són els corresponents als anàlisis realitzats.
El primer dels gràfics correspon a la matriu de correlació de les variables:



La taula que es mostra a continuació mostra per cada un dels diferents models generats de regressió lineal els seus resultats en termes de coeficient de determinació.

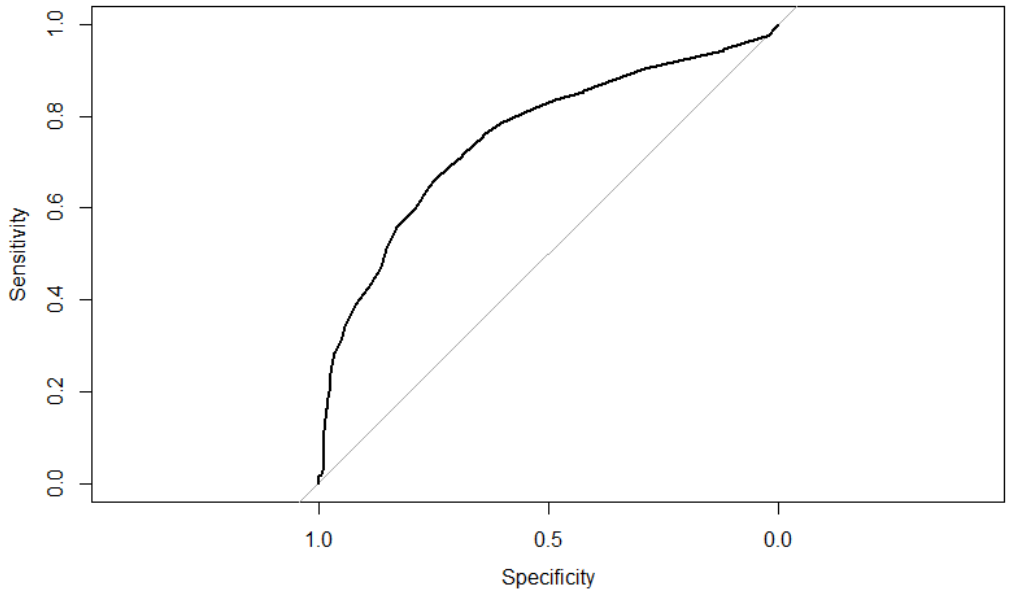
	Coeficient de determinació
Alcohol	0.23092960376205
Alcohol + Volatile	0.299568032040019
Alc + Vol + Sulphates	0.342134183994961
Alc + Vol + Sul + Citric	0.342145180971544

Com veiem en la taula dels coeficients de determinació, quan afegim l'àcid cítric el model no millora. Per veure més en profunditat perquè això passa mostrem els resultats del model:

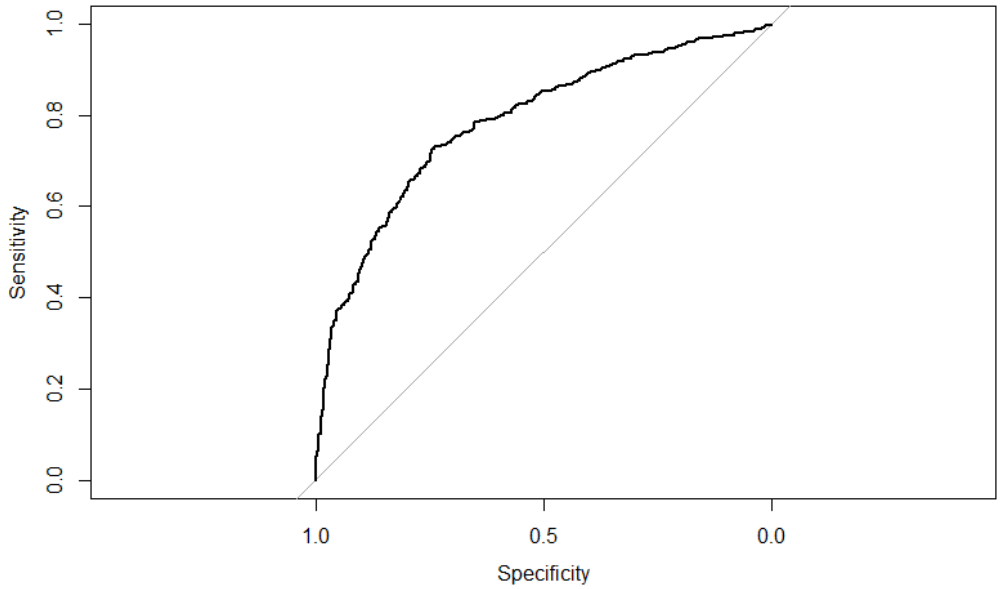
```
lm(formula = quality ~ alcohol + volatile.acidity + sulphates +
    citric.acid, data = WineQuality)

Residuals:
     Min       1Q   Median       3Q      Max
-2.36086 -0.38481 -0.07296  0.47362  2.12809

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.18446    0.20210  10.809  < 2e-16 ***
alcohol            0.30144    0.01580  19.074  < 2e-16 ***
volatile.acidity -0.94026    0.11576  -8.122 9.08e-16 ***
sulphates          1.28369    0.12797  10.031  < 2e-16 ***
citric.acid       -0.01634    0.10010  -0.163     0.87
---
```

Com podem veure, la variable citrc.acid té un valor p superior a 0.05 i per tant no té un nivell de significança prou alt com per establir una relació amb la variable qualitat.

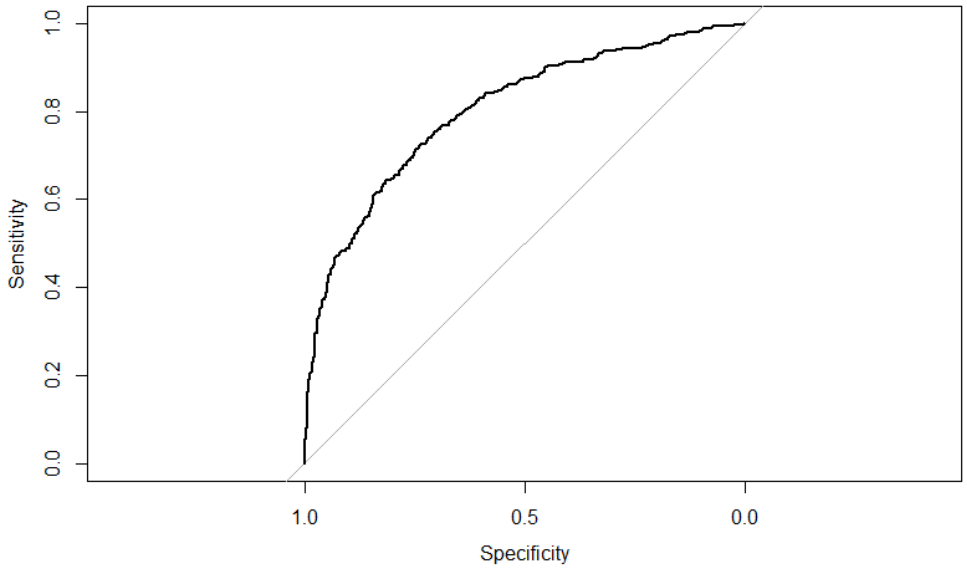
A continuació, el gràfic amb la corba ROC-AUC amb el primer dels models de regressió logística, que només té en compte l'alcohol com a variable predictora. L'AUC pren un valor de 0.754.



Seguidament, el gràfic que ens mostra la corba ROC-AUC per la predicció de la qualitat del vi a partir del model de regressió logística amb les variables predictores d'alcohol i volatile.acidity. L'AUC pren un valor de 0.7848.



Per últim, tenim el model que ens ha donat millors resultats en l'àrea sota la corba, aquest és el model de regressió logística on hem sumat, a més de les variables explicatives anteriors la variable sulphates, obtenint un AUC del 0,8015, el millor dels resultats.



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

Les conclusions que en podem extreure a partir dels diferents anàlisis és que són tres les variables que més influeixen en la qualitat final del vi. Aquestes tres variables són l'alcohol, l'aciditat volàtil i els sulfats. Aquestes tres variables, junt amb l'àcid cítric són que ens indicava la matriu de correlació com les de major impacte en la qualitat. Després de fer anàlisis de regressió lineal i regressió logística trobem que l'àcid cítric no té prou significança com per aportar una relació positiva en els anàlisis de regressió. Per concloure, hem comprovat que amb aquestes tres variables es pot generar un bon model predictor, a partir de la regressió logística, que classifiqui els vins entre "bons" i "dolents" amb una molt bona precisió.

Contribucions	Firma
Investigació prèvia	MSB, MTF
Redacció de les respostes	MSB, MTF
Desenvolupament codi	MSB, MTF