Universitat Politècnica de Catalunya

Facultat d'Informàtica de Barcelona

Estudi de la qualitat del vi blanc

Aprenentatge Automàtic I

Valèria Caro Via
Anna Ramon Hinojosa

Abril 2022

Índex

1 Introducció

L'objectiu d'aquest projecte és, mitjançant les tècniques de regressió estudiades a l'assignatura d'*Aprenentatge Automàtic I*, predir el nivell de qualitat del vi blanc, en una escala del 0 al 10 (sent 10 la màxima qualitat), donades les següents onze variables:

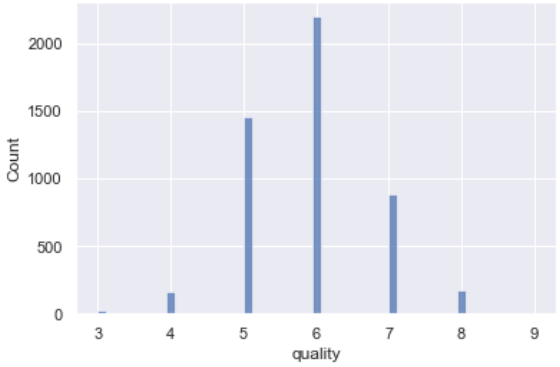
- Acidesa fixa: variable numèrica que ens informa de la quantitat d'àcid tartàric en g/dm^3 present en el vi.
- Acidesa volàtil: variable numèrica que ens informa de la quantitat d'àcid acètic en g/dm^3 present en el vi. En nivells massa alts pot donar lloc a gustos de vinagre desagradables.
- Àcid cítric: variable numèrica que ens informa de la quantitat d'àcid cítric en g/dm^3 present en el vi. En petites quantitats pot donar frescura i gust al vi.
- Sucre residual: variable numèrica que ens informa de la quantitat en g/dm^3 del sucre que queda després de la parada de la fermentació del vi. És estrany trobar vins amb menys d'$1g/L$ i els que en tenen més de $45g/L$ es consideren dolços.
- Clorurs: variable numèrica que ens informa de la quantitat de clorur de sodi (sal) en g/dm^3 present al vi.
- Diòxid de sofre lliure: variable numèrica que ens informa de la quantitat de SO_2 lliure en mg/dm^3. La forma lliure de SO_2 existeix en equilibri amb el SO_2 molecular (com a gas dissolt) i l'ió bisulfit. Impedeix el creixement de microbis i l'oxidació del vi.
- Diòxid de sofre total: variable numèrica que ens informa de la quantitat en mg/dm^3 de formes lliures i lligades de SO_2. En concentracions baixes és quasi bé indetectable al vi, però a partir de concentracions superiors a 50, es fa evident en el nas i el gust del vi.
- Densitat: variable numèrica que ens informa de la densitat del vi en g/cm^3.
- pH: variable numèrica que descriu com d'àcid és un vi en una escala de 0 (molt àcid) a 14 (molt bàsic).
- Sulfats: variable numèrica que ens informa de la quantitat de sulfat de potassi en g/dm^3. Aquest sulfat actua com a un additiu del vi que pot contribuïr als nivells de SO_2 fent que actuiï com a antimicrobià i antioxidant.
- Alcohol: variable numèrica que ens dona el percentatge d'alcohol al vi en tant per cent per volum.

El dataset utilitzat per a dur a terme aquest estudi ha sigut generat per Paulo Cortez, de la Universitat de Minho, Portugal. Les mostres de vi preses provenen del nord de Portugal. En el següent enllaç es troba l'explicació original relativa a la base de dades: `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`
En les següents pàgines es troba l'estudi realitzat complet, amb el preprocessat de les dades inclòs.

2 Exploració de les dades

2.1 Visualització de les dades

Per començar el nostre estudi sobre la qualitat del vi blanc fem una primera exploració de les dades. Observem que el nostre dataset està composat per 4898 instàncies i 11 atributs més l'atribut objectiu que indica la qualitat del vi. Tots els atributs són variables numèriques a excepció de la variable objectiu que és categòrica de 10 nivells diferents. Explorant la variable objectiu ens adonem que les dades estan altament desbalancejades.

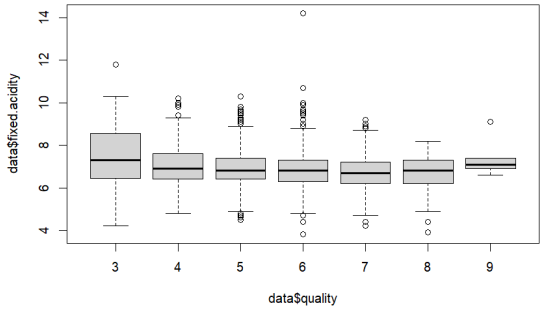


Observem que els dos primers nivels, 1 i 2 no són assignats a cap instància i el nivell de màxima qualitat, el 10, tampoc és assolit en cap moment. D'altra banda, cal destacar que la majoria de dades es concentren en els nivell 5 (30% de les dades), 6 (45% de les dades) i 7 (18% de les dades). A continuació es mostra el summary de les dades que permet fer-nos una idea dels valors que pren cada variable així com possibles valors atípics:

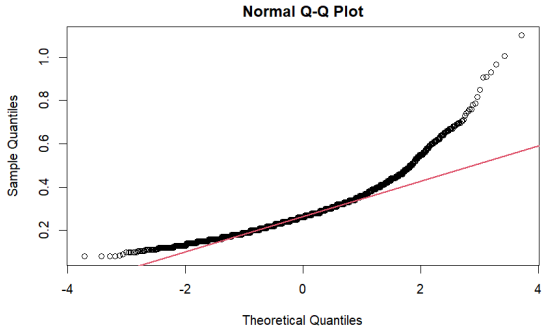
```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   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    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH           sulphates         alcohol          quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

A partir del resum del dataset, els histogrames de cada variable, els qq-plots i els boxplots som capaces d'identificar els següents outliers:

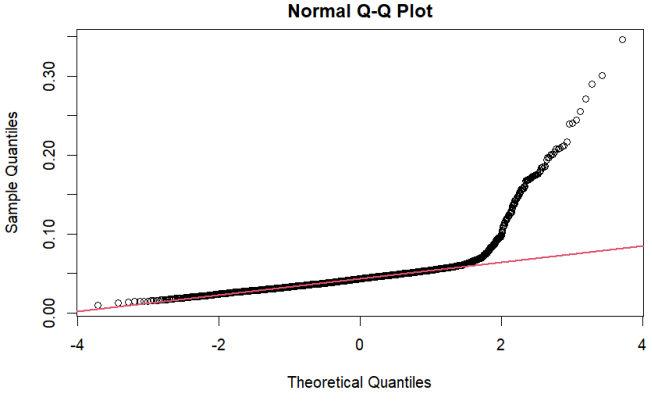
- Per començar, hem observat que tots els valors superiors a 10 g/dm^3 de la variable *acidesa fixa* són inusuals. Mitjançant el boxcox ens adonem de la presència de molts outliers, en concret en el nivell 6 i analitzant el qqplot deduïm el llindar de banalització de les observacions.



- Pel que fa a la variable *acidesa volàtil*, considerem que els valors superiors a 0.8 g/dm^3 també són atípics. El qq-plot ens permet veure que són instàncies puntuals les que prenen valors per sobre de 0.8.



- Altrament, pel que fa a l'*àcid cítric*, els vins que en tenen una presència superior a 1 g/dm^3 considerem que són outliers. A més, també treurem aquells que en tinguin una quantitat menor a 0.1 g/dm^3, ja que valors tan baixos ens poden donar lloc a problemes computacionals més endavant o provocar valors que tendeixen a infinit al fer transformacions de tipus boxcox.
- Per una banda, sembla que hi ha un valor atípic en l'observació amb 65.8 g/dm^3 de *sucre residual* quan la mitjana per a aquesta variable és 6.391 g/dm^3; de fet, considerem que tots els que tenen un valor superior o igual a 30 g/dm^3 són inusuals.
- Per altra banda, per a la variable *clorurs*, els valors superiors a 0.25 g/dm^3 són atípics.



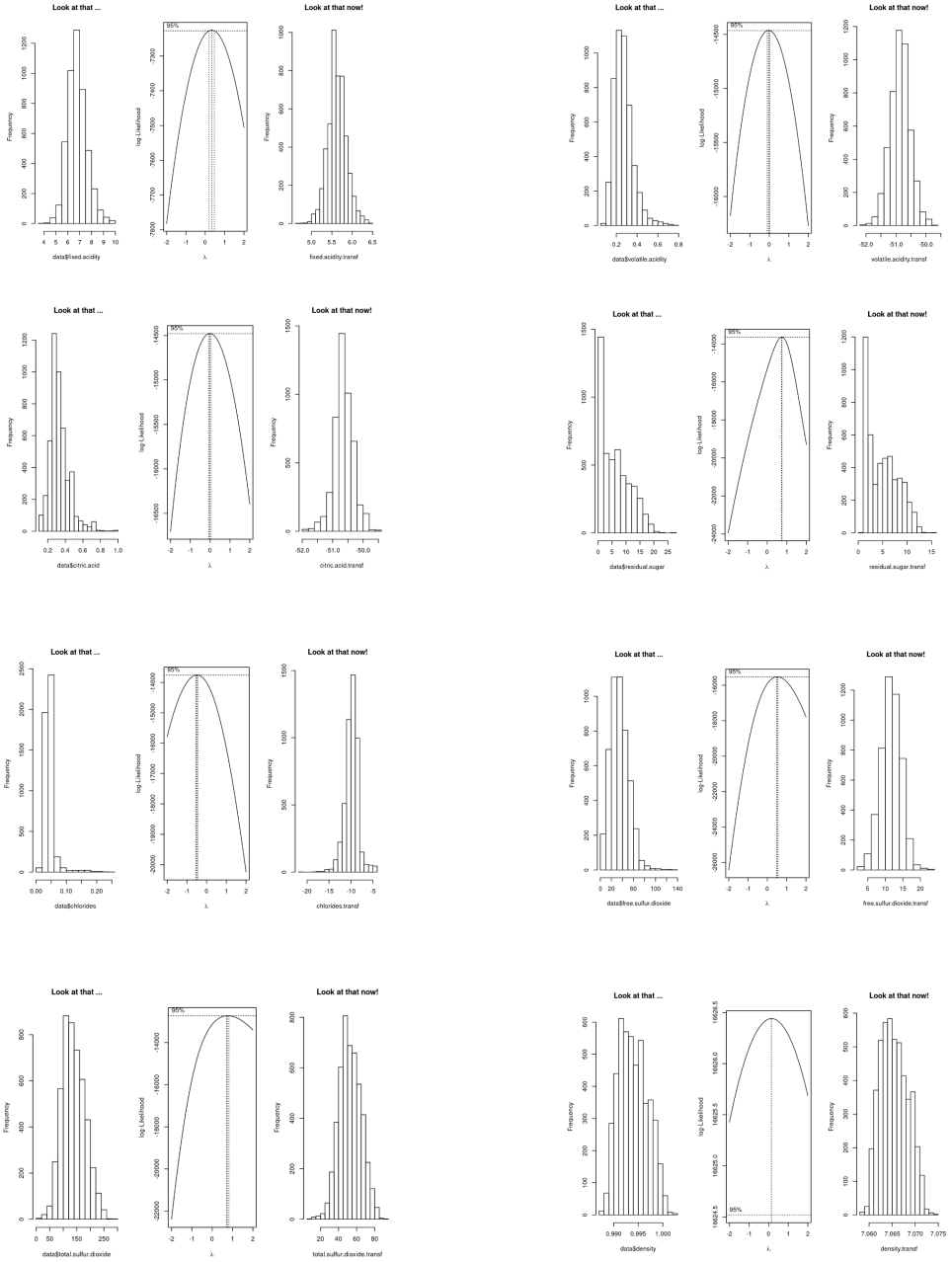
- Ens fixem en la variable *diòxid de sofre lliure*: el valor màxim és 289 i el tercer quartil és 46; sembla ser un valor màxim massa alt. Treurem tots els que en tenen una quantitat igual o superior a 200 mg/dm^3.
- Si ens centrem amb l'atribut *diòxid de sofre total* veiem que el valor màxim, 440, s'allunya molt del mitjà, 138, i el del tercer quartil, 167. Per tant, ens fa pensar que l'instància que aporta aquest valor pot no ser correcta. Treiem tots els valors superiors a 300 g/dm^3.
- Finalment, per a la variable *sulfats* considerarem que valors superiors a 1 g/dm^3 també són poc usuals.

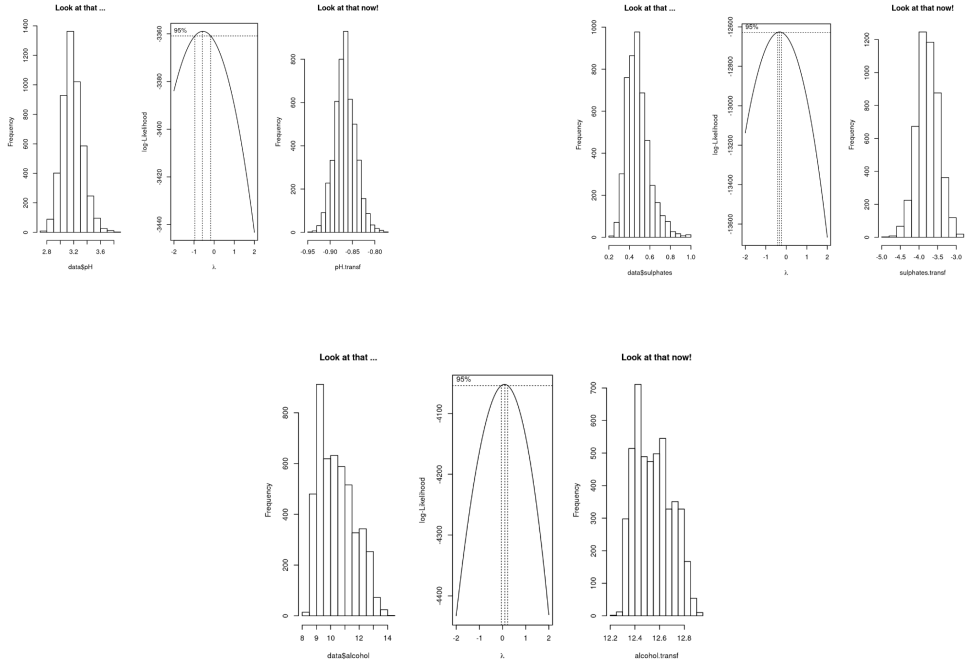
Hem plasmat els qqplots i boxcox només d'algunes variables per poder apreciar gràficament el nostre criteri a l'hora de determinar els outliers. En el codi d'R adjunt es troben els qqplots de totes les variables, si més no, no em considerat oportú incloure tots els gràfics.

Les instàncies que proporcionen els valors determinats com a atípics hem decidit directament eliminar-les, ja que el còmput total d'instàncies que contenien valors atípics eren 120, que representen un 2.45% del total de les instàncies. Per tant, al ser un percentatge tan ínfim ens ha semblat una bona decisió treure-les. Després d'haver fet aquesta extracció, ens adonem que en els qq-plots de les variables les dades sembla que segueixin una mica més fidelment la recta que defineix la normalitat.

2.2 Gaussianitat

Per tal de fer un bon preprocessat de les dades cal tenir en compte els missing values, en el nostre dataset no n'hi ha i per tant, podem saltar-nos aquesta part. Un altre aspecte que cal destacar és que tots els nostres atributs són numèrics, conseqüentment, no cal que fem una distinció pel tractament dels atributs.
Un altre pas a dur a terme és assegurar la gaussianitat de les dades. Per a aquest motiu, hem analitzat els histogrames i els qqplots de totes les variables i hem estudiat quina transformació boxcox caldria aplicar a les dades. Hem arribat a les següents conclusions:



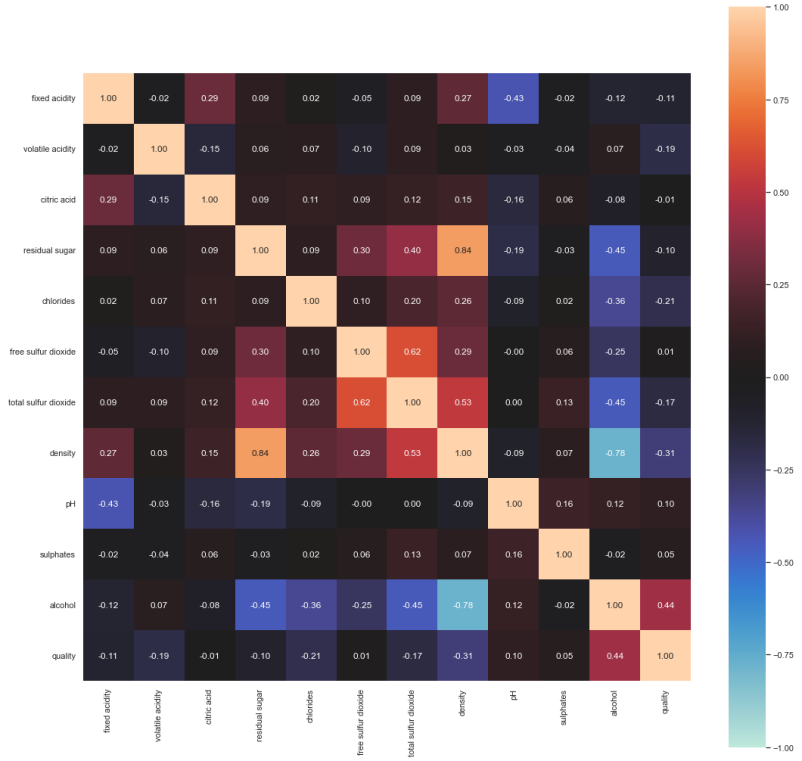


Variable	Transformació
Acidesa fixa	$Y^{0.34}$
Acidesa volàtil	$Y^{-0.020}$
Àcid cítric	$Y^{-0.020}$
Sucre residual	$Y^{0.75}$
Clorurs	$Y^{-0.5}$
Diòxid de sofre lliure	$Y^{0.5}$
Diòxid de sofre total	$Y^{0.75}$
Densitat	$Y^{0.14}$
pH	$Y^{-0.59}$
Sulfats	$Y^{-0.34}$
Alcohol	$Y^{0.10}$

Per assolir el nostre objectiu de ser capaces de predir la qualitat del vi utilitzarem diferents models, entre ells farem ús de la regressió amb regularització, Lasso i Ridge. En aplicar la regularització, penalització dels coeficients, volem que s'apliqui de manera equitativa en totes les variables, per tant, optem per escalar les dades.

2.3 Correlació de variables

Finalment per concloure l'exploració de les dades, estudiem la correlació entre els atributs per veure si podem prescindir d'algun d'ells i, d'entrada, ja simplificar els models:



Observem que la correlació entre les variables *densitat* i *alcohol* i *densitat* i *sucre residual* és molt alta. En el primer cas, d'un 84% i en el segon cas del -78%. Això es deu que, en funció del volum d'alcohol i la quantitat de sucre residual presents en el vi, aquest tindrà una densitat o una altra. Realment, com que la variable *densitat* es pot explicar per a aquestes altres dues, podem prescindir d'ella. A més a més, destaquem que la variable densitat presenta una correlació força alta amb la resta de variables, per exemple, d'un 53% amb la variable *total diòxid de sofre*. Optem per treure-la del dataset.
Després de tots aquests passos ja tenim el dataset llest per a poder aplicar les tècniques de regressió escollides. Recordem que ara el nostre objectiu és predir la qualitat del vi blanc en una escala del 0 al 10 basant-nos en 10 variables numèriques. Comencem separant el dataset en un conjunt de dades d'entrenament, que utilitzarem per a entrenar els models de regressió, i un altre de test, que utilitzarem per a validar els models obtinguts. Per fer aquesta partició hem tingut en compte el fet que les dades estan altament desbalancejades, és important tenir-ho en compte ja que sinó podriem obtenir prediccions molt dolentes, errors de training massa optimistes i errors de test molt dolents. Decidim destinar un 70% de les dades en la mostra de training i un 30% en el test.

3 Regressió lineal

Per tal d'assolir l'objectiu del nostre projecte, predir la qualitat dels vins, ens proposem seguir l'estratègia usual:

- Dividir el nostre conjunt de dades en train/set
- Partint de la mostra de train: trobar l'hiperparàmetre òptim per la regressió ridge fent servir GCV, trobar l'hiperparàmetre òptim per la regressió lasso fent ús de la 10-fold cross validation, trobar el model de regresió linial estàndar que més s'ajusta a les dades.
- Dels 3 models obtinguts, seleccionar el model que millor prediu mitjançant la 10x10 cross validation, és a dir, quedar-nos amb el model que tingui un training error més baix.
- Tornar a entrenar el model seleccionat com a millor utilitzant les dades del train set

3.1 Escalat de les dades

Abans de començar a aplicar les tècniques de regressió, fem un escalat de les dades per tal d'evitar els canvis de dimensió i fer una comparació directa entre variables. Com bé hem comentat anteriorment, per fer una bona regressió amb penalitzacions segons la complexitat del model és necessari fer una estandarització de les dades, ja que no volem que el nostre model es centri en minimitzar variables que prenen un llindar de valors alts i descuidi aquelles que prenen valors petits. Volem que totes les variables siguin tractades per igual.

3.2 Regressió estàndard

El primer model que proposem és una regressió estàndar. Per fer-ho, utilitzem el conjunt de train i anem provant diferents models, extraient variables no significatives. Primerament partim d'un model format amb totes les variables. Per poder analitzar les seves característiques fem un summary del model i obtenim el següent output:

```
Call:
lm(formula = quality ~ ., data = training_set)

Coefficients:
             (Intercept)        fixed.acidity.transf      volatile.acidity.transf        citric.acid.transf      residual.sugar.transf
                 5.89813                   -0.04518                   -0.17538                    0.01696                    0.11965
       chlorides.transf   free.sulfur.dioxide.transf  total.sulfur.dioxide.transf                pH.transf          sulphates.transf
                -0.04390                    0.14443                   -0.06383                    0.01260                    0.04770
         alcohol.transf
                 0.42390


Call:
lm(formula = quality ~ ., data = training_set)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2627 -0.4926 -0.0248  0.4577  3.1427

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.89813    0.01294 455.809  < 2e-16 ***
fixed.acidity.transf   -0.04518    0.01504  -3.005 0.002677 **
volatile.acidity.transf -0.17538    0.01383 -12.682  < 2e-16 ***
citric.acid.transf      0.01696    0.01373   1.235 0.216802
residual.sugar.transf   0.11965    0.01585   7.551 5.56e-14 ***
chlorides.transf       -0.04390    0.01593  -2.756 0.005881 **
free.sulfur.dioxide.transf  0.14443    0.01748   8.261  < 2e-16 ***
total.sulfur.dioxide.transf -0.06383    0.01941  -3.288 0.001020 **
pH.transf               0.01260    0.01524   0.827 0.408378
sulphates.transf        0.04770    0.01336   3.570 0.000363 ***
alcohol.transf          0.42390    0.01815  23.352  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.749 on 3342 degrees of freedom
Multiple R-squared:  0.276,     Adjusted R-squared:  0.2739
F-statistic: 127.4 on 10 and 3342 DF,  p-value: < 2.2e-16
```

Ens adonem que es tracta d'un model que explica poca variabilitat de les dades, la R_{adjust} és molt petita. A més a més, observem que no totes les variables del model són significatives, caldria treure'n algunes. Finalment, després de treure les variables no significatives,

acabem obtenint com a millor model de regressió lineal el que ens proporciona el AIC més baix. El model de regressió estàndar que millor descriu el nostre dataset és el següent:

$$\text{Qualitat} = 5.88 - 0.045 * \text{Acidesa fixa} - 0.18 * \text{Acidesa volàtil} + 0.12 * \text{Sucres residuals} + 0.14 * \text{Diòxid de sofre lliure} - 0.06 * \text{Diòxid de sofre total} + 0.05 * \text{Sulfats} + 0.43 * \text{Alcohol} \quad (1)$$

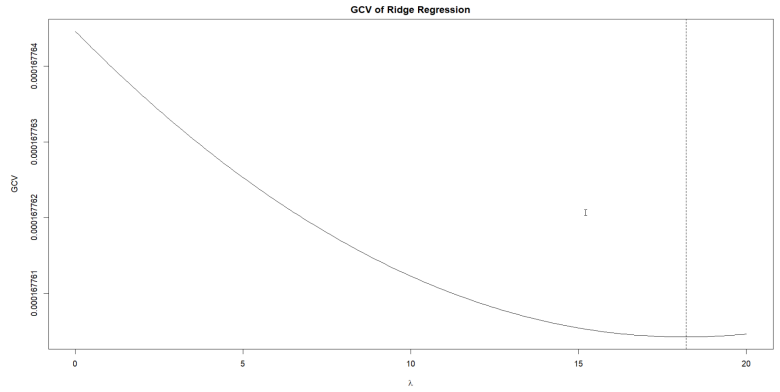
Veiem que variables com l'*acidesa fixa*, l'*acidesa volàtil* i el *diòxid de sofre total* disminueixen la qualitat del vi; sobretot, com més acidesa volàtil té el vi, menys qualitat se li pot otorgar (per cada unitat que augmenta aquesta variable, la qualitat disminueix en 0.18). Per altra banda, les variables *sucres residuals*, *diòxid de sofre lliure*, *sulfats* i, sobretot, *alcohol* aporten qualitat al vi. Segons el nostre model, l'*àcid cítric*, els *clorurs* i el *pH* no tenen cap mena d'efecte en el seu carisme.

3.3 Regressió de Ridge

El segon model de regressió que proposem es basa en la Ridge regression, també coneguda com a Regularized least squares. L'objectiu d'aquesta regressió és penalitzar l'excés de paràmetres en un model i minimitzar el seu error quadràtic mitjà. La seva fórmula general és:

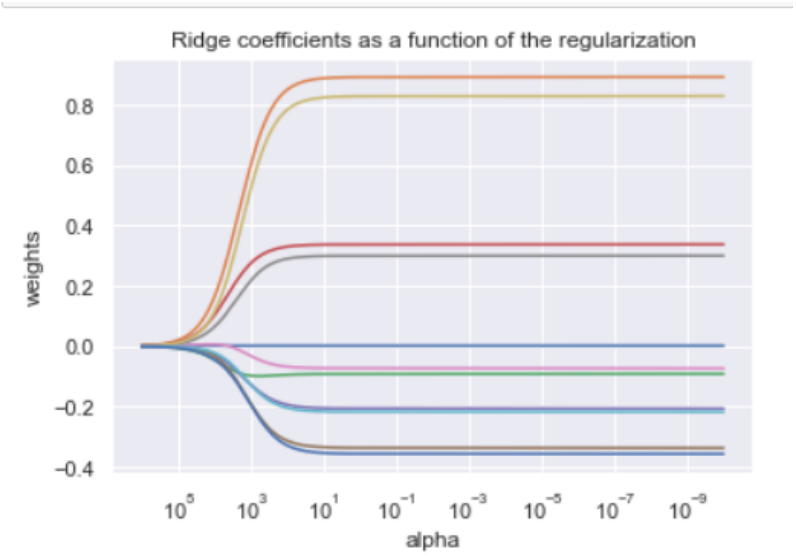
$$y(X; w) = ||t - Xw||^2 + \lambda * ||w||^2 \quad (2)$$

Per trobar el valors de l'hiperparàmetre λ podem utilitzar LOOCV o el mètode GCV, que és el que utilitzarem nosaltres i el qual és més estable. Creem un model mitjançant la regressió Ridge amb el conjunt d'entrenament i apliquem GCV per trobar el millor valor de λ, a continuació grafiquem com influeix la λ en el valor del GCV:



Mitjançant aquest plot observem com disminueix el valor de la Generalized Cross-Validation a mesura que augmentem λ fins obtenir un valor mínim. La funció és una paràbola, ens quedem amb el valor de λ que minimitza la GCV; en el nostre cas, després de vàries execucions, trobem que λ es troba en l'interval $(17, 21)$.
Alpha, o lambda, és l'hiperparàmetre regulador de la complexitat del model, com més proper a 0 és menys importància se li dona a la complexitat del model, és a dir, es prioritza la mimimització de l'error quadràtic mitjà a la simplicitat del model. Per tant, estaríem davant d'una regresió estàndar com en l'apartat anterior. I com més gran és l'hiperparàmetre regulador, més importància se li dona a la complexitat del model, buscant models senzills abans que models amb un MSE mínim. En el següent plot obtenim el pes

dels coeficients de cada variable que minimitzen la funció objectiu segons l'alpha que estem agafant en cada moment. En el nostre cas hem decidit agafar 200 alphes diferents entre 10^{-10} i 10^{6} i ens quedem amb aquella que minimitza la funció objectiu:



L'expressió resultant del model obtingut aplicant la Ridge regression és la següent:

$$\begin{aligned}\text{Qualitat} = 5.89 - 0.0258 * \text{Acidesa fixa} - 0.1856 * \text{Acidesa volàtil} - 0.0031 * \text{Àcid cítric} \\ +0.1087 * \text{Sucre residual} - 0.0797 * \text{Clorurs} + 0.1431 * \text{Diòxid de sofre lliure} \\ -0.0553 * \text{Diòxid de sofre total} + 0.0385 * \text{pH} + 0.0239 * \text{Sulfats} + 0.3893 * \text{Alcohol}\end{aligned} \tag{3}$$

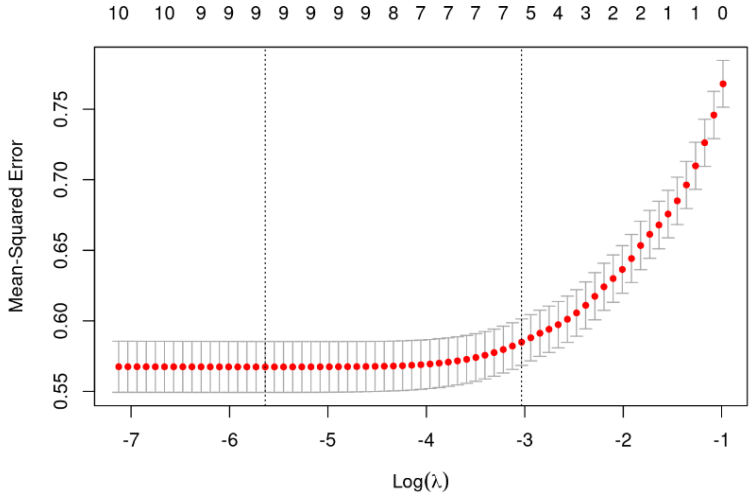
La característica més rellevant, i pròpia dels models de regressió Ridge, és que el model té en compte totes les variables, encara que n'hi hagi que tinguin un pes molt petit; per cada unitat que augmenta l'*àcid cítric*, per exemple, només fa disminuir la qualitat del vi en 0.0031. Igual que en la regressió estàndard, les variables que resten valor a la beguda són l'*acidesa fixa*, l'*acidesa volàtil* i el *diòxid de sofre total*; en aquest cas, s'hi sumen també els *clorurs*, però tenen un pes negatiu poc significatiu. Per altra banda, seguim observant que el *sucre residual*, el *diòxid de sofre lliure*, els *sulfats* i l'*alcohol* segueixen aportant qualitat al vi. També observem que com més bàsica sigui la beguda, en termes de *pH*, més amunt en l'escala del 0 al 10 es trobarà.

3.4 Regressió Lasso

Finalment, proposem un tercer model de regressió: la Lasso, un model lineal que penalitza el vector de coeficients afegint la seva norma L1 (basada en la distància Manhattan) a la funció de cost:

$$RSS_{lasso} = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4}$$

La regressió Lasso tendeix a generar coeficients dispersos: vectors de coeficients en els que la majoria prenen valor zero. Per tant, el model ignorarà algunes variables predictives per a intentar fer el model més simple d'interpretar, i de manera que només es posin en manifest les característiques més importants del conjunt de dades.
En el següent gràfic veiem que la λ mínima ens surt propera a 0.004; és el valor més gran pel qual l'error no se'ns comença a disparar:



En cada execució hem obtingut valors de λ semblants a aquest.
Així doncs, el model resultant és el següent:

$$\begin{aligned}\text{Qualitat} = 5.89 - 0.0039 * \text{Acidesa fixa} - 0.1253 * \text{Acidesa volàtil} + 0.0179 * \text{Sucre residual} \\ -0.0182 * \text{Clorurs} + 0.0774 * \text{Diòxid de sofre lliure} \\ +0.3511 * \text{Alcohol}\end{aligned} \tag{5}$$

Observem que la regressió Lasso ens dona un model similar a la regressió estàndard, en quant a la presència de variables; ha considerat que l'*àcid cítric* i el *pH* tampoc tenen cap mena d'influència en la qualitat del vi i, en addició al model de regressió estàndard, ha apuntat que el pes del *diòxid de sofre total* i dels *sulfats* també és nul. Per altra banda, a diferència del primer model, la regressió Lasso sí que ha donat importància als *clorurs*, considerant que per a cada unitat en que augmenta la seva quantitat, la qualitat del vi disminueix en 0.0182. Finalment, seguim veient com l'*acidesa fixa* i l'*acidesa volàtil* segueixen restant carisme a la beguda, mentre que el *sucre residual*, el *diòxid de sofre lliure* i l'*alcohol* n'hi sumen.

3.5 Comparació de models

Per tal de comparar de manera justa els tres models, hem d'utilitzar la mateixa *cross-validation* per cada tècnica emprada. Per aquest motiu, utilitzarem una 10x10-CV, ja que tot i que les regressions Ridge i estàndard permetin calcular la LOOCV ràpid, amb la regressió LASSO això no és possible. Per a les regressions LASSO i Ridge usarem els valors de λ trobats anteriorment.
Per a poder fer aquesta tasca, fem 10 divisions del conjunt de dades d'entrenament, de manera que una d'elles sempre sigui utilitzada com a conjunt de validació. Per 10 repeticions, ajustem els tres models amb 9 d'aquests subconjunts i, amb el desè, en fem prediccions. Finalment, calculem la diferència entre els valors reals de les dades de validació i les prediccions obtingudes: aquest és l'error mitjà de la validació creuada.
Obtenim els següents resultats:

Tècnica	λ	Error mitjà
Regressió estàndard	-	0.73523
Regressió Ridge	17.2	0.73521
Regressió LASSO	0.0008	0.73522

Observem que el model amb menor error mitjà és el construït mitjançant la regressió Ridge, tot i que valor pels tres models és molt proper. Per aquest motiu, escollim, d'entre aquests tres, el segon. Cal destacar però, que els errors obtinguts no són gaire bons.

3.6 Entrenament del model final

A continuació, entrenem el model final: una regressió Ridge amb $\lambda = 17.2$. Obtenim el següent:

$$\begin{aligned}\text{Qualitat} = 5.89 - 0.0578 * \text{Acidesa fixa} - 0.1732 * \text{Acidesa volàtil} + 0.0195 * \text{Àcid cítric} \\ +0.1275 * \text{Sucre residual} - 0.0580 * \text{Clorurs} + 0.1163 * \text{Diòxid de sofre lliure} \\ -0.0395 * \text{Diòxid de sofre total} + 0.0216 * \text{pH} + 0.0391 * \text{Sulfats} + 0.4285 * \text{Alcohol}\end{aligned} \tag{6}$$

Finalment, calculem l'error de generalització d'aquest últim model utilitzant el conjunt de test; obtenim un valor de, aproximadament, 0.75.
Aquest error és molt elevat, per la qual cosa ens sembla que utilitzar un model de regressió lineal pel nostre problema no és el més adient. Cal tenir en compte que a la nostra base de dades la qualitat de totes les observacions prenen valors discrets en una escala del 0 al 10, mentre que la regressió ens dona valors contínus amb decimals, la qual cosa ens aporta una diferència significativa respecte el valor real de la variable subjecte a estudi; això podria ser un dels motius pels quals obtenim un error tan gran. Per això, en les properes pàgines s'intenten dur a terme altres tècniques per a predir la qualitat del vi blanc, per tal de trobar una solució que s'ajusti més al problema.

4 Suport Vector Regression

En aquest apartat resoldrem el nostre problema mitjançant l'extensió de les Suport Vector Machines en problemes de regressió. L'objectiu és minimitzar la suma dels errors de classificació ponderada a la condició que les observacions que es troben més aprop de l'hiperpla classificador estiguin el més lluny possible (fora del *epsilon*-tube):

$$C\sum_{n=1}^{N} E_\epsilon(y(x)-t) + \frac{1}{2}||w||^2 \tag{7}$$

On:

$$E_\epsilon(y(x)-t) = \begin{cases} 0 & \text{si } |y(x)-t| < \epsilon, \\ |y(x)-t| - \epsilon & \text{altrament} \end{cases} \tag{8}$$

Per tant, el problema a resoldre és:

$$min\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\overline{\xi_i} + \underline{\xi_i}) \tag{9}$$

subject to

$$\begin{cases} (w^T\phi(x_i) + w_o) - y_i \leq \epsilon + \overline{\xi_i} \\ y_i - (w^T\phi(x_i) + w_o) \leq \epsilon + \underline{\xi_i} \\ \overline{\xi_i}, \underline{\xi_i} \geq 0 \qquad\qquad i = 1...n \end{cases} \tag{10}$$

La idea bàsica de les SVR és utilitzar un nombre limitat de mostres d'entrenament (els vectors de suport) per fer prediccions que seran una combinació lineal dels vectors de suport, similar a la regressió lineal. La gran ventatja d'utilitzar les SVR és l'ús de Kernels que permeten projectar les dades a un altre espai on tractar-les sigui més fàcil. Mitjançant la feature mapping function podem enviar les nostres dades de l'espai original a un espai de dimensions superiors, on separar les dades sigui més senzill. En aquest apartat compararem una SVR lineal, una SVR gaussiana i l'algorisme de KNN. Per crear la SVR, primer cal determinar els dos hyperparametres que conté: epsilon (els errors més petits que l'epsilon es consideren 0) i el parametre regularitzador (evita l'ús exessiu del model de vectors suport).
Per trobar ambdós hyperparàmetres decidim realitzar un 5-CV i calculem diferents mesures d'ajust combinant l'ús de direfents kernels amb l'objectiu de quedar-nos amb el millor mètode kernel i els valors de C i epsilon que minimitzen la funció objectiu. Per decidir quin és el millor model i els seus hiperparàmetres farem ús de les mètriques de regressió següents:

- MSE:

$$MSE(t,y) = \frac{1}{D}\sum_{i=1}^{D}(t - y(x;w))^2 \tag{11}$$

  Sabent que t és el valor real del target i y(x;w) és el valor de la nostra predicció. El MSE s'utilitza freqüentment en la fase d'entrenament dels model. L'inconvenient d'aquesta mètrica és que no som capaços d'analitzar-la de manera generalitzada, és a dir, podem comparar-la amb la dels altres models que entrenem i quedar-nos amb el model que presenti un MSE més petit, però depenent del rang de valors de la variable target el MSE prendrà valors molt grans o molt petits.

- R2:

$$normMSE(t, y) = \frac{MSE(t, y)}{s^2(t)} \tag{12}$$

  Sabent que t és el valor real del target i y(x;w) és el valor de la nostra predicció i s^2 és la variància esbiaixada. Aquesta mètrica ens proporciona la variabilitat de les dades que explica el model, sent $R^2 = 1$ el resultat d'un model que explica el 100% de la variabilitat de les dades.

- Median Absolute Error:

$$MedAE(t, y) = median(|t - y(x; w)_1|, ..., |t - y(x; w)_D|) \tag{13}$$

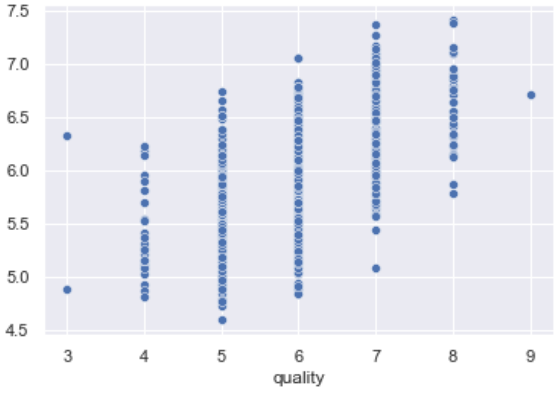
- Mean Absolute Error:

$$normMSE(t, y) = \frac{1}{D} \sum_{i=1}^{D} |t - y(x; w)|^2 \tag{14}$$

Modelem les nostres dades de training fent ús dels models comentats anteriorment. Comencem agafant com a baseline el model KNN de regressió amb els paràmetres per defecte. Després ens centrem amb la SVR on utilitzarem dos kernels diferents i mitjançant una CV obtenim els millors hiperparàmetres. Finalment, obtenim els resultats següents:

	Kernel	C	epsilon	R2	MSE	median_absolute_error	mean_absolute_error
KNN	-	-	-	0.319387	0.512844	0.4	0.53936
LinearSVR-default	linear	1	0	0.283061	0.540215	0.455322	0.573581
LinearSVR-best	linear	30	0	0.284339	0.539252	0.459396	0.573571
RBF-SVR-default	RBF	1	0	0.393943	0.456666	0.400224	0.509728
RBFSVR-best	RBF	10	0.001	0.348574	0.490851	0.380934	0.514311

Per decidir quins hiperparàmetres eren els més bons en funció de cada kernel hem fet servir com a score el R^2. Per tant, comparant els models acabaríem quedant-nos amb la SVR amb un kernel RBF, c=1 i epsilon=0 ja que el R^2 obtingut és el major. Tot i així, cal destacar que un $R^2 = 0.4$ és molt petit, això ens estaria dient que el millor dels nostres models només explica un 40% de la variabilitat de les dades. Fem el plot dels valors predits per la millor SVR i els valors reals:



Observem que les prediccions de la SVR no s'ajusten gairebé als valors reals de la qualitat. Obtenint aquests resultats que no són gaire bons ens vam plantejar que el preprocessat podria contenir errors greus, ja que un mal preprocessat de les dades origina mals resultats

de predicció. No obstant, després de revisar la gaussianitat de les dades, l'escalat, la possible presència de valors atípics no enregistrats... ens adonem que les dades no semblen tenir cap tipus de mancança. D'aquesta manera, concloem que l'analisi de la qualitat del vi blanc mitjançant tèciques de regressió de Machine Learning no sembla un bon mètode. Per tant, descartem la SVR com un bon modelatge de les dades i optem per buscar una nova via. Una possible alternatiuva seria afrontar el problema com un problema de classificació en comptes de regressió.

5 Classificació

Prenent consciència de la mala qualitat dels models obtinguts mitjançant regressió hem arribat a la conclusió que el problema de predir la qualitat del vi no pot tractar-se com una regressió. Com a alternativa vam considerar que podria ser una bona idea tractar-lo com un problema de classificació.

5.1 Classificació 1-vs-tots

La classificació 1-vs-tots és basada en la classificació binària, doncs en aquesta es crea un model binari per cadascuna de les diferents classes de sortida. Llavors, el resultat final de la classificació 1-vs-tots donat un objecte concret és aquella classe que, donades les sortides binària que s'han provat per cada classe, té una probabilitat superior de contenir l'objecte.
Implementant en Python el codi que ens permet entrenar un model de classificació amb les dades d'entrenament, obtenim que la seva *training accuracy* és del 44.83% i, al provar el mateix classificador per predir les dades de test, trobem que la *test accuracy* és del 43.48%.
Seguim observant valors molt baixos per a l'*accuracy*, la qual cosa ens fa pensar que podria deure's a que hi ha classes diferents amb observacions molt pròximes en valors i que, per tant, es confonguin entre elles. Aquest problema de distingibilitat entre classes podria deure's a que la seva representació en la base de dades està desproporcionada, per la qual cosa pot ser que les característiques d'una classe amb molta presència quedin molt clares mentre que les d'unes altra menys representades es barregin entre elles. Tenint en compte la nostra hipòtesi, decidim canviar la manera de mesurar la qualitat del vi blanc; en lloc de considerar una escala del 0 al 10, considerem tres nivells: vins de baixa qualitat (aquells que abans es trobaven entre el nivell 0 i el 4, inclosos), vins de qualitat mitjana (els que es corresponen als nivells 5, 6 i 7) i vins de qualitat alta (corresponents a les classes 8, 9 i 10).
Tornant a aplicar l'algorisme de classificació 1-vs-tots amb les dades separades en conjunt d'entrenament i de test, ara observem una *training accuracy* del 92.95% i una *test accuracy* del 92.96%. Aquest resultat ja és més agradable, doncs sembla que al separar en tres classes, en comptes de en 10, el classificador és capaç de predir millors resultats. Cal tenir en compte, però, que aquest resultat no acaba de ser del tot real i que ens hem aprofitat de que la majoria d'instàncies de la base de dades pertanyen al nivell mitjà definit; per tant, si quasi bé totes les instàncies s'envien al nou nivell mig (classes 5, 6 i 7), el classificador estarà molt entrenat per classificar les observacions a aquesta classe, però quan arribin observacions molt diferents és molt probable que cometem errors rellevants. Per a poder fiar-nos del tot d'aquesta nova manera de classificar i del classificador muntat, hauríem de disposar de moltes més observacions en la nostra base de dades que es corresponguessin als nivells de més baixa i alta qualitat, fent que les representacions de totes les classes fossin

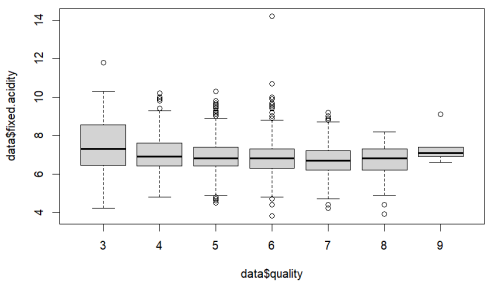
proporcionals entre elles. Decidim provar un altre mètode de classificació per veure si també pateix la mateixa dificultat a l'hora de classificar les instàncies que prenen qualitats en els nivells 5, 6 i 7.

5.2 LDA

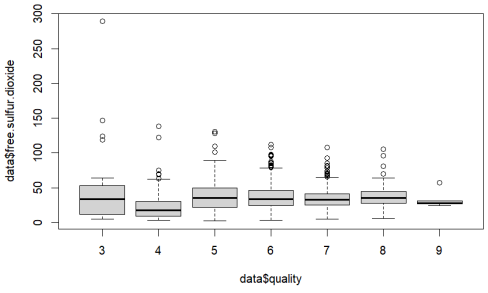
L'anàlisi discriminant linial del dataset ens sembla força acurada tenint en compte la gaussianitat de les nostres dades. Per poder apreciar la qualitat de les prediccions del model analitzem la confusion matrix que ens proporciona la funció d'R i per calcular l'accuracy del conjunt de training realitzem un 10-fold-CV.

```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    3    4    5    6    7    8    9
         3  0.0  0.0  0.1  0.0  0.0  0.0  0.0
         4  0.0  0.4  0.6  0.3  0.0  0.0  0.0
         5  0.1  1.3 14.4  7.9  0.8  0.0  0.0
         6  0.2  1.2 14.1 32.9 12.2  2.3  0.0
         7  0.0  0.1  0.3  4.1  5.2  1.3  0.1
         8  0.0  0.0  0.0  0.0  0.0  0.0  0.0
         9  0.0  0.0  0.0  0.0  0.0  0.0  0.0

 Accuracy (average) : 0.5292
```

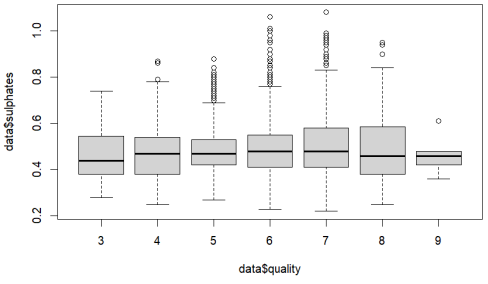
El model prediu correctament amb una fidelitat del 53%. Comparant-ho amb els resultats obtinguts mitjançant regressió observem un augment en la qualitat de les prediccions, tot i així, equivocar-se en un 47% del casos segueix sent un llindar força alt. Si analitzem la confusion matrix propiciada pel model observem que les prediccions pels nivells 5, 6 i 7 són les problemàtiques, tenim el mateix problema que en el cas del modelat de les dades mitjançant 1-vs-all. Arrel d'aquest fet hem decidit analitzar el dataset i ens hem adonat que les instàncies que tenen nivells de qualitat 5, 6 o 7 tenen valors per la resta de variables molt similars. Fent el box plot de cada variable contra la variable objectiu, hem confirmat que en totes les variables la distribuició de valors de les instàncies amb qualitat 5,6 i 7 és molt similar, així com el seu rang interquartílic.



(a) Boxplot *acidesa fixa*



(b) Boxplot *diòxid de sofre lliure*



(c) Boxplot *sulfats*

Per tant, sembla raonable que el model confongui aquests nivells i, en força ocasions predigui malament. Decidim aplicar una nova distinció segons la qualitat dels vins, agrupant les instàncies amb nivells de qualitat propers. Definim 3 nivells: vins de baixa qualitat -qualitat per sota de 5, L- vins de qualitat mitjana -qualitat de 5 i 6, M- vins d'alta qualitat -qualitat superior a 6 H. Els resultats del model en aquest cas són força més bons, l'accuracy ascendeix fins al 76.6%

```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    H    L    M
         H  6.7  0.1  4.8
         L  0.0  0.2  0.2
         M 15.3  3.0 69.7

 Accuracy (average) : 0.7664
```

En última instància decidim aplicar la mateixa distinció que en el model 1-vs-all, agrupant els nivells de qualitat amb característiques molt similars, els del nivell 5, 6 i 7 (els anomenem vins de qualitat mitjana M) i els vins amb una qualitat inferior són els de baixa qualitat (L low) i els de qualitat superior són els de bona qualitat (H):

```
Cross-Validated (10 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    H    L    M
         H  0.0  0.0  0.0
         L  0.0  0.2  0.3
         M  3.7  3.1 92.7

 Accuracy (average) : 0.9293
```

Aquesta última agrupació d'instàncies dona un accuracy molt alt, del 92% tot i que no acaba de ser representativa. El model prediu el 98% de les instàncies com a classe mitjana. És cert que fent aquesta nova agrupació les 3 classes estan molt desbalancejades i gairbé totes les instàncies es troben en el grup de vins de mitjana qualitat. Per tant, no acabaria de ser una agrupació justa.
Conlcusió, els vins de qualitat 5, 6 i 7 comparteixen molts dels seus trets característics, si més no, les variables que estudiem per classificar els vins són molt similars en els tres grups. Aquest fet dificulta la distinsió entre els tres nivells donant peu a errors de classificació entre les 3 classes. Una idea ha estat agrupar els tres nivells degut a la seva similitud i proximitat en l'escala d'1 a 10.
Hem repetit les mateixes agrupacion i hem fet un anàlisi discriminant quadràtic. Els resultats obtinguts en el primer cas, és a dir, tenint en compte els 10 nivells, ens dona un accuracy del 58%, un 5% superior al del LDA. En els altres dos casos l'accuracy obtingut és més dolent que en el LDA, per aquest motiu hem decidit que no era rellevant fer èmfasi dels models obtinguts fent QDA.

6 Conclusions

Després de l'exploració de diverses tècniques de regressió i classificació per tal de resoldre el problema que ens havíem plantejat des de bon principi, hem arribat a les següents conclusions:
En primer lloc, les tècniques de regressió no acaben de ser del tot apropiades al nostre problema, doncs ens havíem proposat predir la qualitat del vi blanc dins una escala discreta del 0 al 10 i, en canvi, els models construïts produeixen errors força grans degut al fet que predeixen la qualitat com a un valor numèric continu. D'aquesta manera, no acabem

d'englobar cada observació dins el grup que li pertany, sinó que el situem enmig de dos valors dins l'escala i, per tant, no acabem d'aclarar si pertany a la classe afitada per sota o a la de sobre. Vam provar de fer ús de la funció round() però el problema seguia vigent, ens trobàvem davant d'un model regressiu de mala qualitat. En vistes d'aquest problema, hem decidit recórrer a tècniques de classificació.

A l'implementar la classificació 1-vs-tots i LDA ens hem adonat que classificar en tants grups com teníem previst és difícil degut a com estan repartides les classes en les instàncies de la nostra base de dades: hi ha classes amb molta representació i altres on aquesta és mínima. Això fa que l'*accuracy* dels classificadors amb prou feines arribi al 50%; doncs resulta difícil distingir entre nivells pròxims ja que hem corrovorat que estan molt correlats. Nivells propers comparteixen característiques molt similars, per tant, extreure un perfil de cada nivell és gairebé impossible. Com a solució, hem intentat agrupar les classes de manera que en tinguéssim menys i quedessin més separades entre elles. Hem definit tres nivells: vins de baixa qualitat (corresponents a l'escala del 0 al 4, inclosos), vins de qualitat mitjana (els que pertanyen als grups 5, 6 i 7) i vins d'alta qualitat (els que se situen entre el 8 i el 10, inclosos). Amb aquesta agrupació hem aconseguit millorar molt l'*accuracy* dels classificadors, però també som conscients que els resultats als que hem arribat són poc reals: la presència de vins de qualitat mitjana dins la base de dades és molt superior a la resta, per no dir que pràcticament la inmensa majoria d'instàncies pertanyen a aquest grup. Per tant, per a poder assegurar-nos que els classificadors fan una bona feina predint la qualitat de la beguda ens caldria disposar de més observacions de baixa i alta qualitat i entrenar de nou amb les noves dades els models. Un altre aspecte que cal remarcar és la subjectivitat en la classificació de la qualitat dels vins. La distinsió d'un vi bo respecte un vi dolent és subjectiva, però és cert que es pot arribar a un consens força generalitzat. En aquest cas estaríem davant d'una classificació binària i la caracterització de cada classe segurament seria bastant diferent. En canvi, fer una classificació de vins puntuant-los de 0 a 10 no és una tasca que pugui generalitzar-se facilment. Partim del fet que aquesta diferenciació és totalment subjectiva i no té una base en funció dels nivells del vi, sinó del paladar de cada humà. És cert que per arribar a aquesta classificació s'ha arribat a un acord, però com de diferents són un vi de qualitat 5 i un de qualitat 6? Doncs basant-nos en els paràmetres estudiats hem vist a nivell qualitatiu que no hi ha gairebé diferència. Aquest fet és el que ha complicat la modelització de les dades.

En conclusió, podem dir que ens hem vist, en certa mesura, limitades en recursos pel que fa a la base de dades; ens hagués agradat disposar de més observacions, i que aquestes estiguessin repartides de manera més proporcional entre classes, per tal de fer un estudi més precís i extens sobre quin tipus de models són més adients per resoldre el nostre problema. Tot i així, podem dir que hem conclòs que una classificació sembla més adequada que una regressió si el que volem és situar cada vi en una escala del 0 al 10 pel que fa a la seva qualitat.

Com a últim incís volem remarcar que vam acabar trobant un model que semblava predir de manera força rigurosa, una xarxa neuronal. Al llarg d'aquest curs no hem tractat aquest algorisme i, per tant, no en farem molta èmfasi però ens ha semblat interessant nombrar-ho. El model que hem utilitzat és el Multi-layer Perceptron d'una hidden-layer que donat un conjunt de variables i un target és capaç de trobar una funció que millor aproximi el comportament propi de cada classe. Aquest mètode és capaç de fer prediccions molt bones. Observem les mètriques obtingudes entrenant un model mitjançant el MLP i fent prediccions sobre les dades de test:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	3
4	0.92	1.00	0.96	34
5	1.00	1.00	1.00	348
6	1.00	1.00	1.00	535
7	1.00	1.00	1.00	217
8	0.97	0.90	0.94	42
9	0.00	0.00	0.00	1
accuracy			0.99	1180
macro avg	0.70	0.70	0.70	1180
weighted avg	0.99	0.99	0.99	1180

La millora de les prediccions és tan notòria respecte a les dels altres mètodes emprats degut a la gran capacitat d'aprenentatge que presenten les xarxes neuronals i a que els paràmetres que passem com a input són els coeficients PC's. El fet de realitzar el PCA de les dades ens assegura la no correlació entre les variables. L'input és propagat cap a la hidden layer combinat amb diferents pesos i aplicant-hi la funció logística com a funció d'activació. Finalment, l'última capa només conté una neurona que combina els resultats prèvis de la hidden-layer i els aplica a la funció d'activació. Per calcular el nombre de neurones és bo realitzar una CV i per evitar overfitting de la xarxa cal regularitzar el model.

7 Bibliografia i referències

Per a dur a terme aquest projecte hem consultat les següents fonts:

- Paulo Cortez, University of Minho, Guimarães, Portugal. A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal. UCI Machine Learning Repository - `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`
- Ukutu's blog, One-vs-All Classification Using Logistic Regression - `https://utkuufuk.com/2018/06/03/one-vs-all-classification/`
- Sckit-learn, scikit-learn developers, Multiclass One-vs-All Classifier - `https://scikit-learn.org/stable/modules/generated/sklearn.multiclass.OneVsRestClassifier.html`
- Sckit-learn, scikit-learn developers, SVR - `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html/`
- Sckit-learn, Metrics and scoring: quantifying the quality of prediction - `https://scikit-learn.org/stable/modules/model_evaluation.html`
- S. Abirami, P. Chitra, in Advances in Computers, 2020. ScienceDirect, Multilayer-Perceptron - `https://sciencedirect.com/topics/computer-science/multilayer-perceptron`