APA. Predicció de la qualitat del vi

Carlos Escolano
Bernat Huguet

Contents

1 Introducció

L'objectiu d'aquest treball és estudiar com es relaciona la qualitat dels vins amb la concentració de certs components en ells. Per dur a terme aquesta tasca fem servir un dataset de vins portuguesos que inclou observacions de vins blancs i negres i la seva qualitat graduada en una escala de l'1 al 9.

En primer lloc analitzem el comportament les característiques de les observacions, detectant els valors anòmals (Outlayers) a les dades i tractant els possibles casos de skewness o kurtosis. Seguidament busquem possibles relacions entre aquestes característiques i la qualitat amb què estan graduades a la mostra.

Un factor important en el tractament del conjunt de dades és la gran diferència en volum de les diferents qualitats. Per intentar remeiar-ho proposem una reducció de dimensionalitat, la qual intentem predir amb les mateixes tècniques per veure si suposa una diferència important en els resultats.

Amb les dades ja tractades intentarem predir la qualitat dels vins segóns la graduació original (del 1 al 9) i la reduïda (1 a 3) a partir d'un subconjunt de la mostra (Training set) fent servir un conjunt de mètodes lineals (KNN LDA) i no lineals (Random Forest, SVM amb kernel RBF, MLP) per veure amb quina certesa podem predir la qualitat del vi a partir dels seus atributs. Seguidament compararem com són de precisos els models amb el test set i seleccionarem el millor candidat.

Finalment analitzem els resultats del model escollit (Random Forest, amb una precisió de predicció del 73,73%), avaluarem i discutirem el procés seguit i els problemes enfrontats i plantejarem altres problemes interessants a resoldre en aquest àmbit.

2 Estudis Previs

L'anàlisi de la qualitat del vi és un tema que s'ha tractat ja en multitud d'estudis, però potser els més controvertit de tots ells un estudi realitzat a finals dels 80. Orley Ashenfelter, que era professor d'economia a Princeton amb un fort interès en l'econometria, considerava que la metodologia d'avaluació i predicció de la qualitat del vi estava desfasada respecte al temps. Això el va portar a fer recerca sobre el tema, recollir dades i publicar un model predictiu lineal per a la qualitat del vi d'una temporada. Concretament, la formula que va desenvolupar va ser:

quality = 12.145 + 0.00117 winter rainfall + 0.0614 avg growing season temperature - 0.00386 harvest rainfall

Malauradament no va tardar a sorgir resistència d'avant d'aquest enfoc. Un dels opositors més destacats va ser Robert M. Parker, un critic de vi de fama mundial, que considerava un art més que una ciència la valoració del vi. Aquesta situació va portar a una competència entre els crítics de vi tradicionals i els models de predicció matemàtics, en la qual els primers es van veure superats en la precisió de les seves prediccions.

Un altre estudi a destacar és el que va tractar aquestes mateixes dades realitzat per Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Mato i José Reis a la universitat de Minho. Aquest equip divers, combinant el coneixement de domini d'un Enòleg, el domini de l'estadística d'un matemàtic, un enginyer en biologia i dos informàtics, van desenvolupar una SVM aconseguint una precisió entre el 63.3% i el 85.5% segons la classe concreta a predir.

3 Anàlisi de les dades

Per aquest projecte tenim dos datasets sobre vins portuguesos, un primer dataset per vins blancs i un segon per vins negres. Per la naturalesa de les nostres variables no hem de tractar possibles variables contínues (tret de la qualitat que volem predir), ja que totes les variables són variables contínues que presenten la presència de components fisicoquímics a les observacions. Podem observar els indicadors de centre i dispersió dels dos tipus:

Indicadors de centre i dispersió dels vins blancs

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide
Min. : 3.800	Min. :0.0800	Min. :0.0000	Min. : 0.600	Min. :0.00900	Min. : 2.00
1st Qu.: 6.300	1st Qu.:0.2100	1st Qu.:0.2700	1st Qu.: 1.700	1st Qu.:0.03600	1st Qu.: 23.00
Median : 6.800	Median :0.2600	Median :0.3200	Median : 5.200	Median :0.04300	Median : 34.00
Mean : 6.855	Mean :0.2782	Mean :0.3342	Mean : 6.391	Mean :0.04577	Mean : 35.31
3rd Qu.: 7.300	3rd Qu.:0.3200	3rd Qu.:0.3900	3rd Qu.: 9.900	3rd Qu.:0.05000	3rd Qu.: 46.00
Max. :14.200	Max. :1.1000	Max. :1.6600	Max. :65.800	Max. :0.34600	Max. :289.00

total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
Min. : 9.0	Min. :0.9871	Min. :2.720	Min. :0.2200	Min. : 8.00	Min. :3.000
1st Qu.:108.0	1st Qu.:0.9917	1st Qu.:3.090	1st Qu.:0.4100	1st Qu.: 9.50	1st Qu.:5.000
Median :134.0	Median :0.9937	Median :3.180	Median :0.4700	Median :10.40	Median :6.000
Mean :138.4	Mean :0.9940	Mean :3.188	Mean :0.4898	Mean :10.51	Mean :5.878
3rd Qu.:167.0	3rd Qu.:0.9961	3rd Qu.:3.280	3rd Qu.:0.5500	3rd Qu.:11.40	3rd Qu.:6.000
Max. :440.0	Max. :1.0390	Max. :3.820	Max. :1.0800	Max. :14.20	Max. :9.000

Indicadors de centre i dispersió dels vins negres

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide
Min. : 4.60	Min. :0.1200	Min. :0.000	Min. : 0.900	Min. :0.01200	Min. : 1.00
1st Qu.: 7.10	1st Qu.:0.3900	1st Qu.:0.090	1st Qu.: 1.900	1st Qu.:0.07000	1st Qu.: 7.00
Median : 7.90	Median :0.5200	Median :0.260	Median : 2.200	Median :0.07900	Median :14.00
Mean : 8.32	Mean :0.5278	Mean :0.271	Mean : 2.539	Mean :0.08747	Mean :15.87
3rd Qu.: 9.20	3rd Qu.:0.6400	3rd Qu.:0.420	3rd Qu.: 2.600	3rd Qu.:0.09000	3rd Qu.:21.00
Max. :15.90	Max. :1.5800	Max. :1.000	Max. :15.500	Max. :0.61100	Max. :72.00

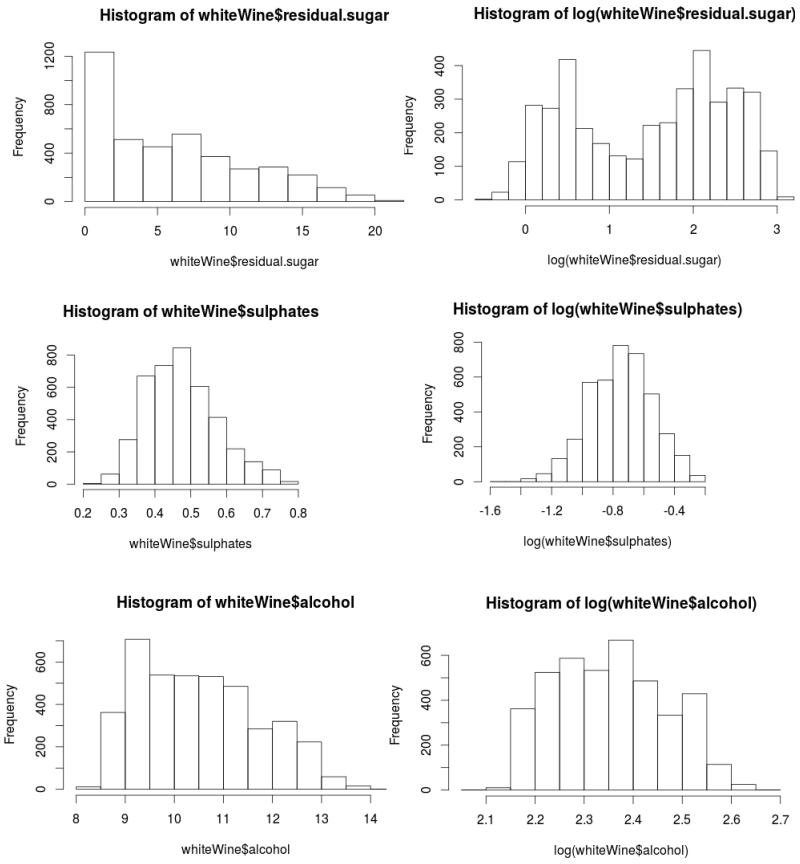
total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
Min. : 6.00	Min. :0.9901	Min. :2.740	Min. :0.3300	Min. : 8.40	Min. :3.000
1st Qu.: 22.00	1st Qu.:0.9956	1st Qu.:3.210	1st Qu.:0.5500	1st Qu.: 9.50	1st Qu.:5.000
Median : 38.00	Median :0.9968	Median :3.310	Median :0.6200	Median :10.20	Median :6.000
Mean : 46.47	Mean :0.9967	Mean :3.311	Mean :0.6581	Mean :10.42	Mean :5.636
3rd Qu.: 62.00	3rd Qu.:0.9978	3rd Qu.:3.400	3rd Qu.:0.7300	3rd Qu.:11.10	3rd Qu.:6.000
Max. :289.00	Max. :1.0037	Max. :4.010	Max. :2.0000	Max. :14.90	Max. :8.000

En analitzar les dades veiem dos factors importants:

- La diferència entre la concentració de components fisicoquímics varia molt entre vi blanc i vi negre. Exemples d'aquesta situació són les variables sucre residual (residual sugar) amb una diferència del 256%, acidesa volàtil(valite acidity) amb el 192%, o la quantitat total de diòxid de sulfur(total sulphure dioxide) amb el 300%. Aquesta situació pot reflectir diferències subjectives a l'hora de qualificar el gust d'un tipus de vi, però en ordre de no proporcionar als models informació contradictòria decidim tractar els dos tipus de vins per separat. Els models per aquest motiu els aplicarem sobre el conjunt de dades de vins blancs, ja que aquest dataset suposa el 68% del total de les dades i evitem mesclar informació possiblement contradictòria.

- La distribució de les característiques de les dades no sempre és la més adequada, a vegades amb considerable skewness i/o kurtosis. Per comprovar-ho hem realitzat histogrames de cada variable

individualment i hem calculat la seva skewness i kurtosis. En realitzar aquest anàlisi hem trobat que per alguns atributs calia corregir la seva skewness. Aquestes variables són alcohol(0,39 a 0,21), residual.sugar(0,748 a 0,17) i sulphates(0,45 a -0.05). Als següents histogrames podem veure'n la diferencia:



Comparativa entre els histogrames de les 3 variables abans i després de fer el logaritme de les dades.

Un cop tractades les dades podem comprovar quina és la correlació de cadascuna d'elles amb la qualitat que volem predir i observem que les dues variables amb major correlació amb la qualitat són la quantitat d'alcohol i la densitat de l'observació amb un 0.41 i -0.29 respectivament, mentre que les altres mostren una correlació molt baixa no superant el 0.12.

```
"fixed.acidity : -0.0524192864513668"
"volatile.acidity : -0.117125850001902"
"citric.acid : 0.0358270876981967"
"residual.sugar : -0.066432600066059"
"chlorides : -0.279538296180308"
"free.sulfur.dioxide : 0.0170360242390806"
"total.sulfur.dioxide : -0.165039356154331'
"density : -0.298326809705554"
"pH : 0.0791526348942642"
"sulphates : 0.00396455892167523"
"alcohol : 0.417528980577889"
```

Finalment si observem la distribució del volum de les observacions segons l'atribut qualitat, ens adonem que la seva distribució és extremadament desigual. Hi ha moltes més observacions de qualitat 6 que de cap de les altres, les qualitats 1 i 2 no estan representades i de la qualitat 9 només tenim 4. Per intentar solucionar aquesta heterogeneïtat en el volum segons la qualitat de les dades proposem una classificació diferent que balancegi la quantitat d'observacions entre les diferents qualitats:

- Les observacions amb qualitat inferior a 6 les classificarem com a baixa qualitat("bad").
- Les observacions amb qualitat 6 les classificarem com a qualitat regular("normal")
- Les observacions amb qualitat superior a 6 les classificarem com a bona qualitat ("good")

Table 1: Distribució de classes original (1-9)

3	4	5	6	7	8	9
20	163	1457	2198	880	175	5

Table 2: Distribució de classes proposada(bad-normal-good)

bad	normal	good
1236	1883	955

Com podem veure la distribució en tres qualitats presenta un balanç de representació molt més uniforme que les nou categories originals. A la fase d'experimentació compararem els resultats obtinguts en aplicar els diferents algoritmes sobre les dues classificacions.

4 Resampling Protocol

En aquesta secció tractarem els mètodes que hem escollit per distribuir les dades en els nostres experiments. Com a conjunt de comprovació (Test set) hem escollit aleatòriament un 10% de les dades (407 obs.), i la resta l'hem designat com a conjunt d'entrenament (training set) (3667 obs.). Una opció que vam considerar va ser dedicar una major proporció al conjunt de comprovació (Entorn a un terç de les observacions) però aquestes opcions reduïen molt la mida del conjunt d'entrenament, podent provocar una disminució en la precisió dels models construïts.

Un altre punt important és realitzar cross-validation per tal d'ajustar els paràmetres dels nostres models sense caure en overfitting. En aquest sentit hem de diferenciar entre els models lineals i els models no lineals. Per als models lineals ens hem decidit per realitzar leave one out cross validation, perquè són models molt ràpids d'entrenar. Pel que fa als models no lineals hem decidit fer servir 10-fold cross validation perquè implica menys re-entrenaments que l'escollida per als models lineals, reduint així el temps total d'execució.

Durant la resta del treball prendrem com a mesura de l'error la formula:

$$error = 1 - (obs.correctes/totalobservacions)$$

5 Predicció amb Models Lineals o Quadràtics

En primer lloc discutirem els models lineals que hem emprat en aquest projecte: K nearest neighbors i LDA.

5.1 KNN

En primer lloc per KNN hem d'ajustar el paràmetre de quants veïns ha de tenir en compte l'algorisme per això fem servir LOOCV per provar quin valor entre 1 i 10 ens dóna un error menor. A l'esquerra mostrem els errors obtinguts per al model de 9 categories de qualitat i a la dreta per al model amb 3 categories de qualitat:

k	LOOCV error
1	0.4150532
2	0.5031361
3	0.5216798
4	0.5374966
5	0.5306790
6	0.5402236
7	0.5374966
8	0.5407690
9	0.5478593
10	0.5424052

k	LOOCV error
1	0.3959640
2	0.4895010
3	0.4941369
4	0.5085901
5	0.5096809
6	0.5102263
7	0.5126807
8	0.5074993
9	0.5189528
10	0.5154077

Com podem veure en els dos casos el valor que ens dóna un error menor és quan el valor de k és 1.

Un cop ajustat el valor de k entremen els models i comprovem el seu error amb el conjunt de test. Els resultats obtinguts són d'un error de 0.4226 per la clasificació amb 9 classes i 0,3686 per 3 classes. Tot i que no hi ha una gran diferencia a l'error si observem les matrius de confussió:

	whiteWine.test.classes						
myknn	3	4	5	6	7	8	9
3	0	0	0	1	0	0	0
4	1	2	1	4	2	0	0
5	0	2	61	31	7	1	0
6	1	1	36	117	25	3	1
7	0	2	8	31	45	3	0
8	0	1	3	4	3	10	0
9	0	0	0	0	0	0	0

	whiteWine.3.test.classes		
myknn	bad	normal	good
bad	77	42	9
normal	30	123	19
good	19	31	57

Matrius de confusió. Model amb 9 classes (dreta) i model amb 3 classes (esquerra)

A la matriu de confusió del model de 9 classes veiem que els resultats tendeixen a acumular-se a les classes majoritàries (5 i 6). En canvi, al model amb 3 classes es veu com tot i que tenim un error alt les prediccions es reparteixen millor entre les classes.

5.2 LDA

En aquesta secció discutirem els resultats obtinguts d'entrenar un model LDA amb les nostres dues classificacions. L'objectiu original d'aquest apartat era realitzar una comparativa entre LDA i QDA però vam trobar un problema amb QDA, pel fet que hi ha classes amb menys observacions que variables al nostre model pel cas de 9 classes, fet que fa impossible entrenar amb QDA. Tot i això per a la classificació de 3

classes vam entrenar models LDA i QDA i vam obtenir errors de LOOCV (leave one out cross validation) de 44.39596 i 47.15026. Per aquests motius hem decidim comparar els resultats obtinguts amb LDA.

Hem entrenat models LDA per les dues classificacions i hem obtingut errors de predicció per al conjunt de test de 48.64 i 43.73 per 9 i 3 classes respectivament. Amb les següents matrius de confusió:

	3	4	5	6	7	8	9
3	0	0	1	1	0	0	0
4	0	0	1	6	1	0	0
5	0	0	55	53	1	0	0
6	0	0	33	140	15	0	0
7	0	0	5	63	14	0	0
8	0	0	1	11	5	0	0
9	0	0	0	0	1	0	0

	bad	normal	good
bad	63	61	2
normal	46	130	20
good	6	43	36

Matrius de confusió. Model amb 9 classes (dreta) i model amb 3 classes (esquerra)

A les matrius de confusió podem veure com els models tendeixen a concentrar les serves prediccions a les classes majoritàries no ajustant bé les altres.

6 Predicció amb Models no Lineals

6.1 SVM amb kernel RBF

En aquesta secció discutirem els resultats obtinguts amb un model SVM amb kernel RBF sobre els dos mètodes de classificació. En primer lloc hem realitzat 10-fold cross validation per obtenir els valors més adients per als paràmetres cost i gamma. Per a 3 classes hem obtingut que els millors valores eren cost = 21 i gamma = 0.8 per als dos models amb un menor error de cross validation de 0.285 per 3 classes i 0.305 per 9 classes.

Un cop hem definit aquests paràmetres entrenem els models i computem el seu error de predicció per al conjunt de test i obtenim un error de 0,4064 pel model amb 9 classes i 0,348 per al model amb 3 classes. Si observem les seves matrius de confusió:

	truth						
pred	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0
5	0	1	52	27	3	0	0
6	2	5	51	139	35	4	1
7	0	1	4	20	42	3	0
8	0	0	2	2	2	10	0
9	0	0	0	0	0	0	0

	truth		
pred	bad	normal	good
bad	59	20	3
normal	57	162	38
good	10	14	44

Matrius de confusió. Model amb 9 classes (dreta) i model amb 3 classes (esquerra)

Veiem com els models continuen predint erròniament les classes més representades però ja es pot observar com la majoria de les prediccions ja es troben a la diagonal de la matriu, que són les prediccions correctes.

6.2 Random Forest

En el cas de random forest hem utilitzat dos processos per ajustar els paràmetres dels models. En primer lloc hem realitzat 10-fold cross validation per decidir quin nombre de variables era el més adient per als models, ja que amb random forest podem mesurar la importància que tenen els models cadascuna de les variables. En tots dos casos el nombre de variables que ha donat com a resultat un error OOB menor ha sigut 11, és a dir utilitzant totes les variables.

Per al model per classificar 3 classes hem realitzat un tractament addicional, afegint un sampling a les classes a 800 cadascuna, ja que la classe menys representada té 855 observacions, per igualar la

Amb els paràmetres ja fixats entrenem els dos models i calculem els seus errors de predicció amb el conjunt de test i hem obtingut per al model amb 9 classes un error de 0.317 i per al model amb 3 classes un error de 0.2629. Si observem les seves matrius de confusió:

Posteriorment hem entrenat el nostre model per diferent nombre d'arbres per seleccionar el valor amb l'error OOB menor i hem obtingut que els millors valors eren 631 arbres pel model amb 9 classes i 398 per al model amb 3 classes.

Truth \ Pred	3	4	5	6	7	8	9
3	0	0	1	1	0	0	0
4	0	1	2	5	0	0	0
5	0	1	66	41	1	0	0
6	0	0	21	154	13	0	0
7	0	0	2	33	47	0	0
8	0	0	0	2	5	10	0
9	0	0	0	0	1	0	0

Truth \ Pred	bad	normal	good
bad	102	21	3
normal	39	127	30
good	3	11	71

Matrius de confusió. Model amb 9 classes (dreta) i model amb 3 classes (esquerra)

Veiem com les prediccions es concentren més a la diagonal principal de la matriu, significant que són prediccions correctes. Aquest comportament és present especialment a la matriu de confusió del model amb 3 classes.

6.3 MLP

El procés que hem seguit per entrenar els nostres models MLP ha sigut seleccionar un nombre de neurones força alt (50 per tots dos casos) i per 10-fold cross-validation trobar quin és el millor valor per al paràmetre decay de la xarxa entre 0.0001 i 0. Els resultats que hem obtingut és que per totes dues classificacions el paràmetre que millors resultats donava era 0.01 amb valor Kappa de 0,3219 per 3 classes i 0.2805 per 9 classes.

Fixant aquests paràmetres entrenem els models per les dues classificacions amb nombre màxim d'iteracions 500 i obtenim com a errors de validació amb test 0.4864 per al model amb 9 classes i 0.6584 per al model amb 3 classes. Com podem observar els models no ajusten bé les classes, especialment el model amb 3 classes.

7 Model Definitiu

De tots els models amb els quals hem experimentat el que ha donat uns millors resultats ha sigut random forest per a totes dues classificacions. Podem comprovar la importància de les variables al nostre model:

	MeanDecreaseGini
fixed.acidity	183.1544
volatile.acidity	230.5649
citric.acid	196.3709
residual.sugar	220.8167
chlorides	204.8876
free.sulfur.dioxide	227.7912
total.sulfur.dioxide	230.2147
density	262.5802
pH	215.8987
sulphates	199.0939
alcohol	275.6273

	MeanDecreaseGini
fixed.acidity	113.8076
volatile.acidity	151.1679
citric.acid	123.6317
residual.sugar	136.6441
chlorides	139.5474
free.sulfur.dioxide	147.4535
total.sulfur.dioxide	141.5541
density	175.6131
pH	136.8657
sulphates	112.5183
alcohol	221.0309

Importància de les variables. Model amb 9 classes (dreta) i model amb 3 classes (esquerra)

Veiem com per tots dos models les classes que tenen més impacte en el model són la proporció d'alcohol i la densitat de l'observació. Les quals coincideixen amb les classes que individualment tenien una major correlació amb la qualitat. Les classes que menys impacte tenen en la predicció són la acidesa fixa(fixed.acidity) i els sulfats(sulphates). També podem destacar que hi ha una major diferencia entre els valors de les variables alcohol i densitat amb la resta al model de 9 classes que en el model de 3.

A l'apartat anterior hem vist que per cross validation obteniem errors menors fent servir totes les variables, però podem comprovar com en tots dos casos l'error augmenta fins a 0.288 per a 3 classes i 0.342 per al model amb 9 classes. És a dir, tot i que no suposen un gran augment de l'error sí que afecten negativament al percentatge d'encert a l'hora de predir el conjunt de test.

8 Conclusions

En enfrontar la complexitat de la predicció de la qualitat del vi ens hem trobat que el tipus del vi influencia substancialment les seves característiques. Suposem que fins i tot la qualitat es veu afectada, donat que la valoració del vi és subjectiva i el que es valora en un vi blanc pot ser diferent del que es busca en un vi negre. També hem observat que les característiques de les quals disposem que més ajuden a determinar la qualitat del vi blanc són: la seva densitat i tassa d'alcohol. A més, de totes les tècniques utilitzades els models no lineals han estat els que més precisió han aconseguit, força per sobre dels lineals. Accessòriament, la recerca d'estudis previs ens ha ensenyat detalls històrics interessants sobre el conflicte que origina enfrontar expertes en un tema i noves tecnologies. També ha estat interessant observar metodologies seguides per doctorats per solucionar el mateix problema.

Hi han algunes decisions que podríem haver pres de manera diferent. Per exemple, podríem haver juntat els datasets de vi blanc i negre a l'hora de fer les prediccions e intentar predir tant la qualitat com el seu tipus. De la mateixa manera podríem haver decidit diferents classes de qualitat que les escollides. (¡6,6, 6¡)

També ens hem plantejat un seguit de possibles extensions del treball:

- L'ampliació més simple seria provar altres tècniques no explicades en aquest curs, per exemple utilitzar deep learning.
- Una extensió natural també seria intentar classificar els vins d'aquest dataset en blanc i negre.
- Una opció extra interessant seria recollir dades de diverses tipologies de vins i fer un estudi relacionant les característiques desitjades per a cada classe, descobrint les característiques desitjades segons el tipus de vi.
- Una altra opció extra seria recollir més característiques per cada vi e intentar afinar la predicció. Per exemple, les característiques climatològiques a la que es va exposar al cultiu (Utilitzades per Ashenfelteren el seu estudi) i l'any de producció.
- També resultaria interessant avaluar el comportament de la qualitat. Podríem agafar dades de vins valorats en diferents anys, fer patter matching i relacionar-los amb vins d'altres anys per les seves característiques. Seguidament podríem comparar les diferents valoracions de vins similars en diferents anys i comprovar si la escassetat de vins desitjables altera la valoració i de quina manera ho fa. Aquesta millor comprensió de com varia la qualitat segons factors externs del mercat (Per exemple) ens permetria fer-li un pretractament per poder millorar la precisió de les prediccions en millorar la coherència de les valoracions de qualitat de vins de diferents anys.

La valoració final és que aquest treball ha estat un bon exercici d'interiorització i assentament dels conceptes tractats durant el curs. Anàlisis de dades i avaluació de les seves característiques, aplicació i avaluació de models de predicció lineals i no lineals i avaluació de prediccions mitjançant mètodes de càlcul d'error i cross validation.

9 Bibliografia

- Cortez, Paulo; Cerdei, António; Almeida, Fernando; Matos, Telmos ;Reis, José. Modeling wine preference by data mining from physicochemical properties. University of Miho. 2009.
- Bertsimas, Dimitris; O'Hair, Allison; Silberholz, John; Dunning, Iain; King, Angie; Misic, Velibor; Youssef, Nataly; Weinstein, Alex ; Kung, Jerry. EdX courses: The Analytics Edge.
- Russell, Stuart; Norvig, Peter. Machine Learning, a modern approach 3rd edition. 2009.
- Bishop, Christopher. Pattern Recognition and Machine Learning. 2006.