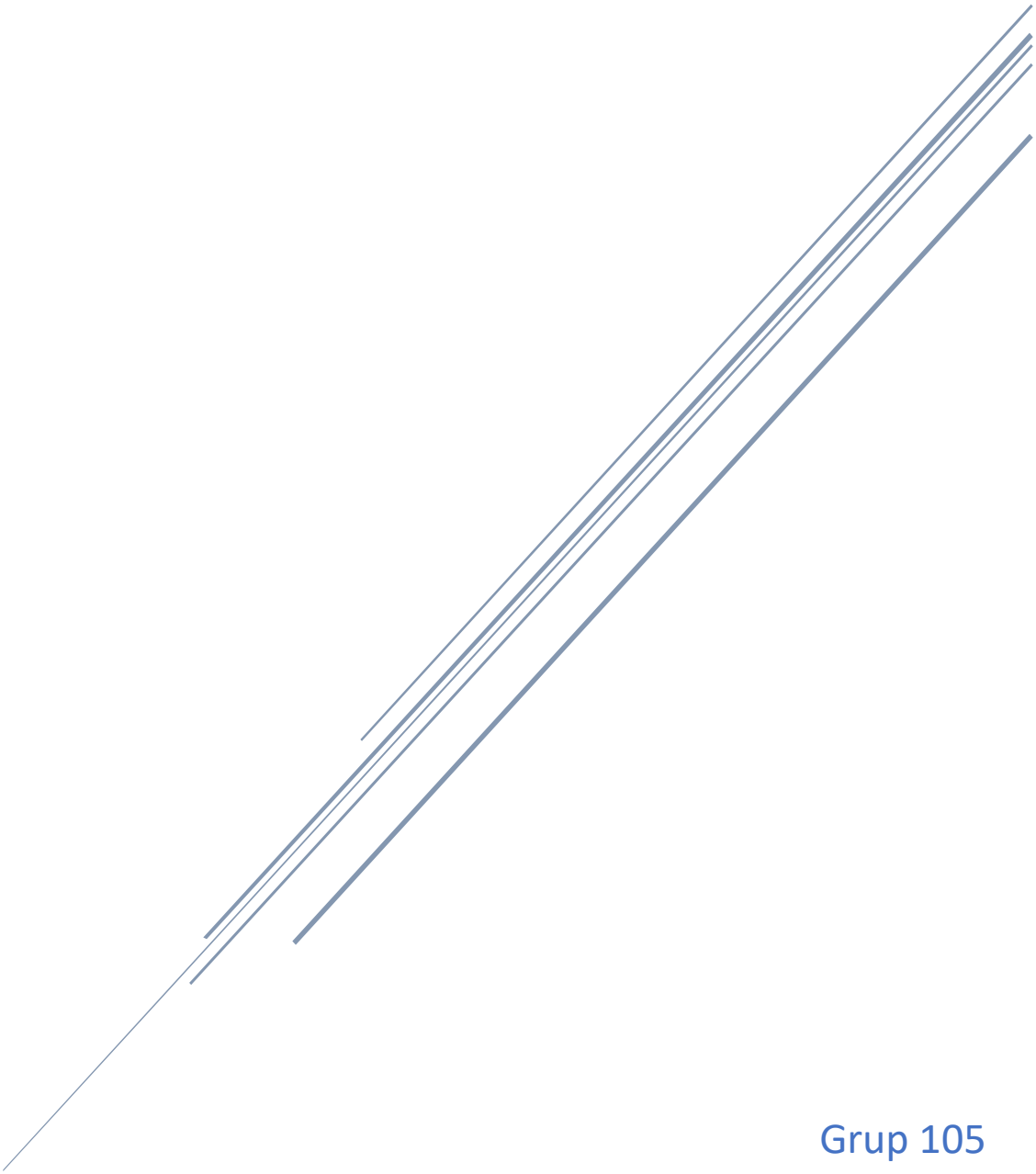
PRÀCTICA 1: REGRESSIÓ

Red wine quality dataset



Grup 105
Laia Rubio – NIU:1600830
Erik Villarreal – NIU:1599119
Raúl Villar – NIU:1596830

Índex

Introducció

L'objectiu de la pràctica és analitzar una base de dades sobre la qualitat del vi vermell amb l'objectiu de poder aplicar-li models de regressió. Per a aconseguir-ho, al llarg de la memòria s'explicarà el procés aconseguit i es resoldran les preguntes de l'enunciat en forma de apartats i subapartats.

Les preguntes a resoldre son:

- Apartat C:
    - Quin és el tipus de cada atribut?
    - Quins atributs tenen una distribució Gaussiana?
    - Quin és l'atribut objectiu? Per què?
- Apartat B:
    - Quin són els atributs més importants per fer una bona predicció?
    - Amb quin atribut s'assoleix un MSE menor?
    - Quina correlació hi ha entre els atributs de la vostra base de dades?
    - Com influeix la normalització en la regressió?
    - Com millora la regressió quan es filtren aquells atributs de les mostres que no contenen informació?
    - Si s'aplica un PCA, a quants components es redueix l'espai? Per què?
- Apartat A:
    - Com influeixen tots els paràmetres en el procés de descens? Quins valors de learning rate convergeixen més ràpid a la solució òptima? Com influeix la inicialització del model en el resultat final?
    - Quines funcions polinomials (de diferent grau, de diferents combinacions d'atributs, ...) heu escollit per ser apreses amb el vostre descens del gradient? quina ha donat el millor resultat (en error i rapidesa en convergència)?
    - Utilitzeu el regularitzador en la fòrmula de funció de cost i descens del gradient i proveu polinomis de diferent grau. Com afecta el valor del regularitzador?
    - Quina diferència (quantitativa i qualitativa) hi ha entre el vostre regressor i el de la llibreria ?
    - Té sentit el model (polinomial) trobat quan es visualitza sobre les dades?
    - Ajuda la visualització a identificar aquelles mostres per a les que el regressor obté els pitjors resultats de predicció?

La base de dades sobre la que operarem tracta sobre unes 1600 mostres aproximadament, i aporta informació sobre la composició de cada vi tractat, i la seva qualificació.

Plantejament de dades

Inicialitzar, visualitzar i preparar les dades

La base de dades bé donada per l'arxiu "winequality-red.csv" i ens aporta un total de 12 variables i 1599 mostres. Imprimint les primeres 5 mostres, i amb la funció 'describe()' podem veure i aprendre tota la informació que ens donarà la base de dades al llarg de la pràctica.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.400	0.700	0.000	1.900	0.076	11.000	34.000	0.998	3.510	0.560	9.400	5
1	7.800	0.880	0.000	2.600	0.098	25.000	67.000	0.997	3.200	0.680	9.800	5
2	7.800	0.760	0.040	2.300	0.092	15.000	54.000	0.997	3.260	0.650	9.800	5
3	11.200	0.280	0.560	1.900	0.075	17.000	60.000	0.998	3.160	0.580	9.800	6
4	7.400	0.700	0.000	1.900	0.076	11.000	34.000	0.998	3.510	0.560	9.400	5

Taula de les primeres 5 mostres de la BD i els seus respectius valors

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000	1599.000
mean	8.320	0.528	0.271	2.539	0.087	15.875	46.468	0.997	3.311	0.658	10.423	5.636
std	1.741	0.179	0.195	1.410	0.047	10.460	32.895	0.002	0.154	0.170	1.066	0.808
min	4.600	0.120	0.000	0.900	0.012	1.000	6.000	0.990	2.740	0.330	8.400	3.000
25%	7.100	0.390	0.090	1.900	0.070	7.000	22.000	0.996	3.210	0.550	9.500	5.000
50%	7.900	0.520	0.260	2.200	0.079	14.000	38.000	0.997	3.310	0.620	10.200	6.000
75%	9.200	0.640	0.420	2.600	0.090	21.000	62.000	0.998	3.400	0.730	11.100	6.000
max	15.900	1.580	1.000	15.500	0.611	72.000	289.000	1.004	4.010	2.000	14.900	8.000

Taula obtinguda amb la funció describe() de la llibreria pandas

Amb aquestes mostres podem crear una taula d'informació sobre las variables com aquesta:

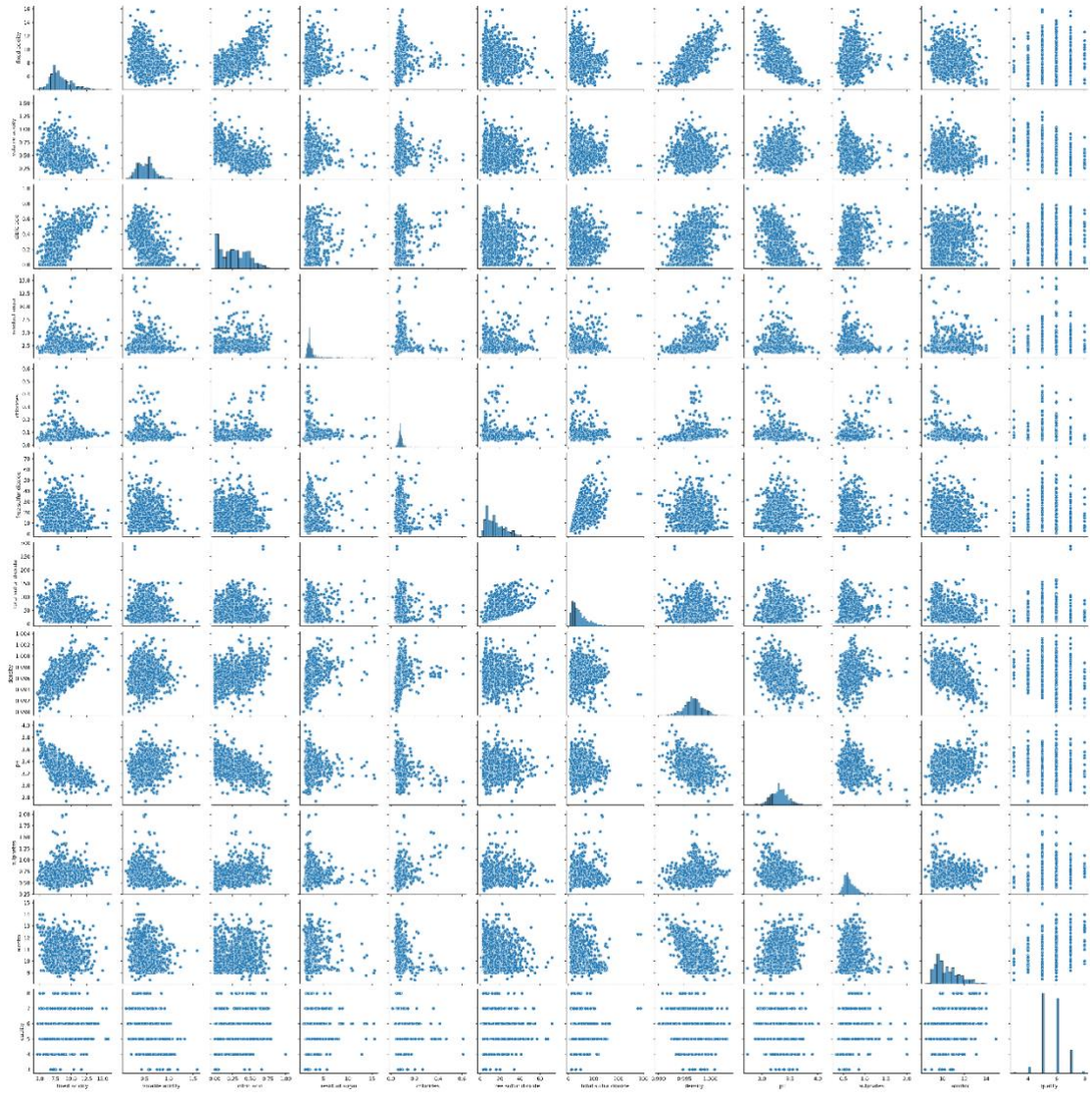
Nom Variable	Tipus de dada	Rang	Tipus de variable
fixed acidity	Float64	(4,600 – 15,900)	Continua
volatile acidity	Float64	(0,120 – 1,580)	Continua
citric acid	Float64	(0 – 1)	Continua
residual sugar	Float64	(0,900 – 15,500)	Continua
chlorides	Float64	(0,012 – 0,611)	Continua
free sulfur dioxide	Int64	(1 – 72)	Discreta
total sulfur dioxide	Int64	(6 – 289)	Discreta
density	Float64	(0,990 – 1,004)	Continua
pH	Float64	(2,740 – 4,010)	Continua
sulphates	Float64	(0,330 – 2,000)	Continua
alcohol	Float64	(8,400 – 14,900)	Continua
quality	Int64	(3 – 8)	Discreta

Com no tenim ninguna variable categòrica, no haurem de fer ningun tractament específic a alguna de les variables. Abans de continuar amb l'anàlisi, ens hem assegurat de que totes les variables contenen valors i no hi ha cap mostra que pugui tenir NULLS en algun apartat de la taula.

Una vegada ja sabem de que tracta la nostra base de dades, i ja no pot tenir valors que puguin donar resultats il·lògics, ja podem començar a analitzar els atributs que la composen i la seva relació entre ells.

Comprendre els atributs

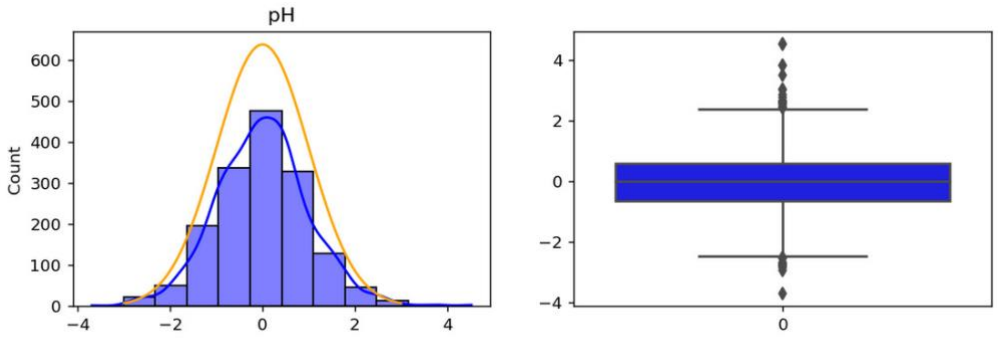
Per a poder veure com es comporten cada atribut i com es relacionen entre ells, la millor opció és utilitzar la funció 'plairplot()' de la llibreria seaborn, per a crear gràfiques entre variables.



Gràfica generada amb la funció 'plairplot()' de la llibreria seaborn

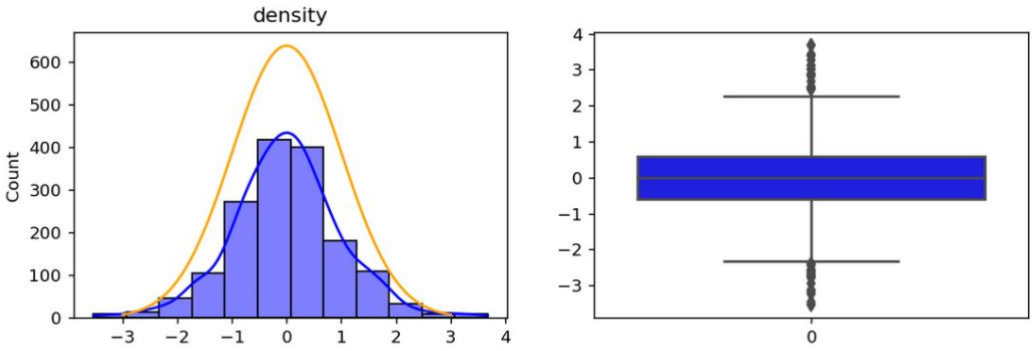
Aquesta gràfica ens dona informació molt important per a escollir quins atributs necessitarem per a fer la regressió lineal, però de moment ens centrarem en els histogrames generats en la diagonal de l'imatge. Aquests histogrames ens ajudaran a veure les distribucions de cada variable, i poder veure quines tenen distribucions Gaussianes. De entre totes, hem intuït que només els atributs 'density' i 'pH' tenen aquesta distribució.

Per a comprovar aquesta hipòtesi hem utilitzat la funció 'normaltest()' de la llibreria SciPy, la qual ens pot dir de manera estadística si de veritat son distribucions Gaussianes o no. Una vegada implementada la funció, ens ha donat uns resultats inesperats: No troba ninguna distribució Gaussiana. Amb ajuda gràfica podem veure millor que és el que esta passant.




```
[pH] : No es distribucio normal
```

Gràfica de la distribució de l'atribut 'pH'.



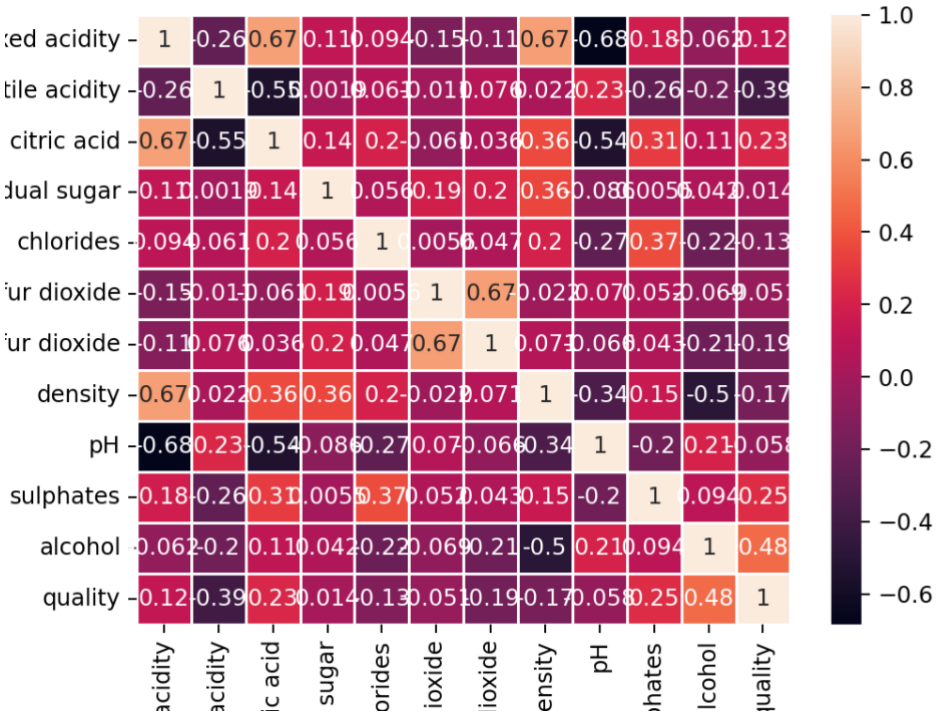

```
[density] : No es distribucio normal
```

Gràfica de la distribució de l'atribut 'density'.

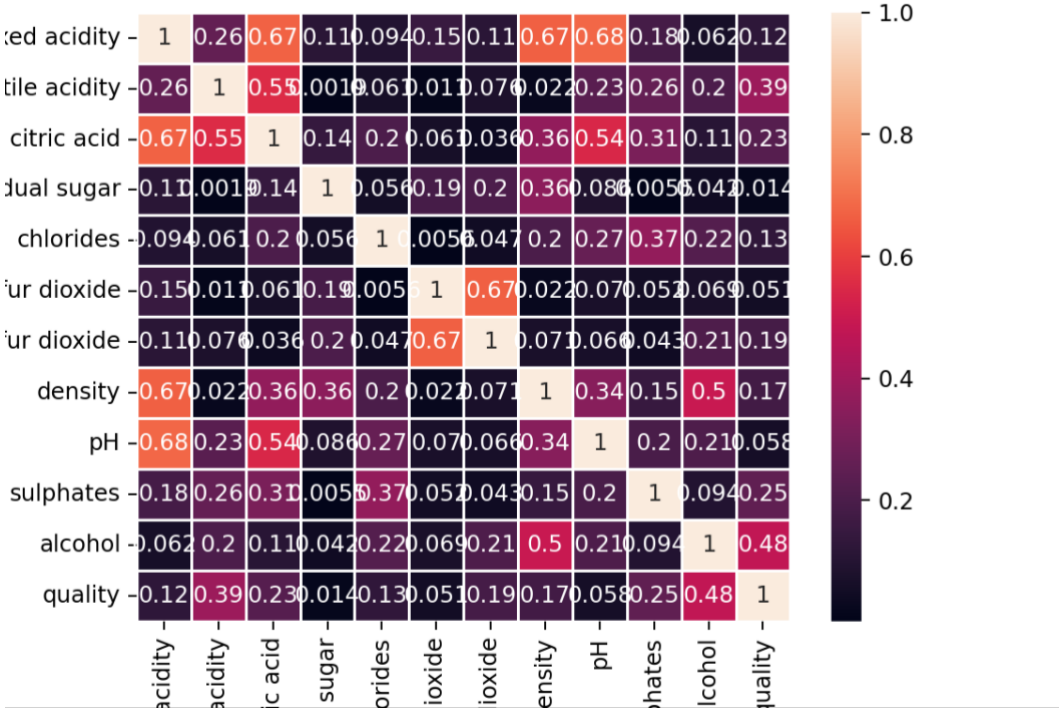
A les gràfiques anteriors podem visualitzar una distribució Gaussiana idònia (línia groga) i les distribucions que tenim per als atributs 'density' i 'pH' (línia blava). Podem observar com la causa de que no es considerin distribucions Gaussianes son la quantitat de mostres que hi ha als outliers. Si apliquem un tractament de dades com la regla del 3σ, podem eliminar totes les mostres amb valors molt distants, i si tornem a executar les funcions, ara els atributs 'density' i 'pH' si que passen el test de normalització i podem confirmar que tenen distribucions Gaussianes.

Per a la creació de la regressió lineal, farem els tests de dues formes degut a aquest canvi de les dades. Continuarem utilitzant les dades sense tractar els outliers, i també utilitzarem les dades tractades amb la regla del 3 σ per a comparar quines de les dues pot obtenir millors resultats.

A més de les gràfiques, per a poder comprendre del tot la relació entre cada atribut, hem creat dues matrius de correlació. Una, la ja explicada a classe amb els valors per defecte dels atributs, i una altre, amb tot valors absoluts per a veure relacions inversament proporcionals.



Matriu de correlació generada amb la funció 'corr()' de la llibreria pandes i 'heatmap()' de la llibreria seaborn



Matriu de correlació amb valors absoluts generada amb la funció 'corr()' de la llibreria pandes i 'heatmap()' de la llibreria seaborn

Construcció del regressor lineal

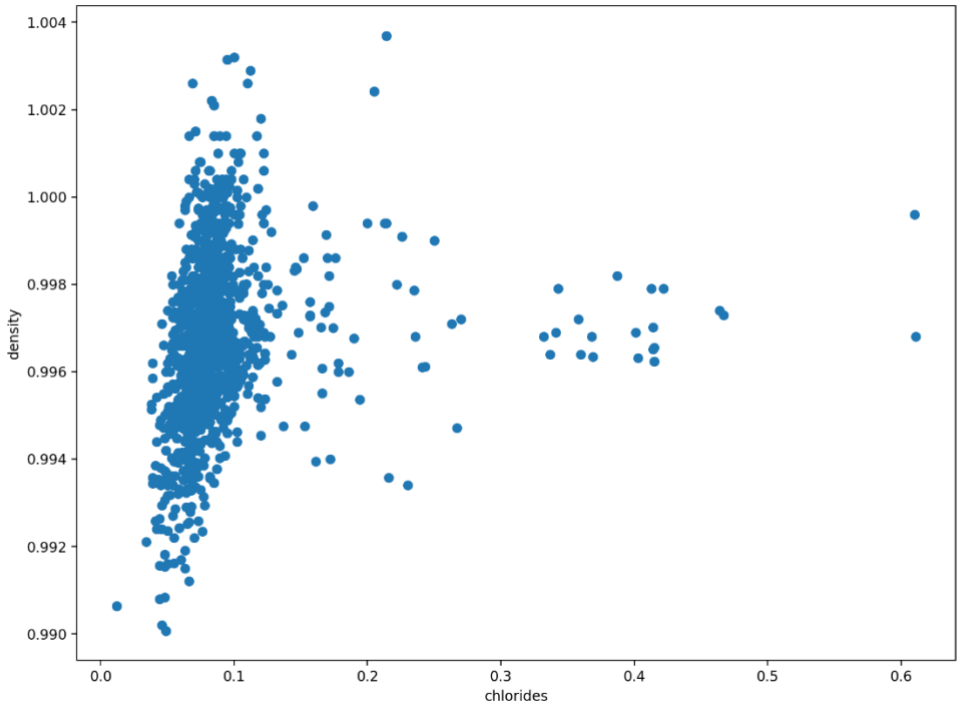
Selecció d'atributs

Amb tota aquesta informació, ja podem començar a decidir quin serà l'atribut objectiu de la nostra regressió lineal.

Degut a que els atributs 'quality', 'free sulfur dioxide' i 'total sulfur dioxide' son variables discretes, les eliminarem dels atributs candidats ja que son menys òptims que la resta per a obtenir una regressió correcta.

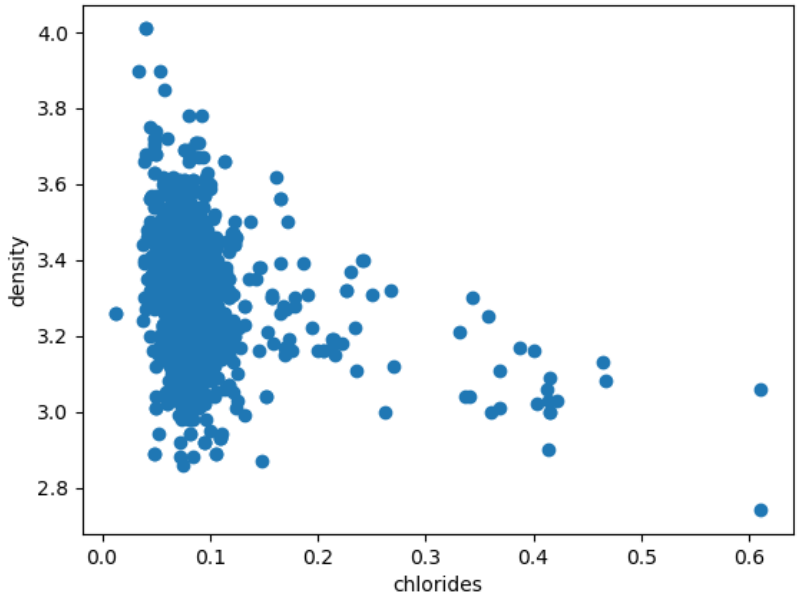
Per a aquesta pràctica hem decidit que és bona idea veure com es comporta una distribució Gaussiana en la nostra regressió, per tant ens decidirem entre 'pH' i 'density0'.

Observant les matrius de correlació, ambdues tenen molt bona correlació amb els altres atributs, però hem decidit com a atribut objectiu 'density' ja que ens sembla el més equilibrat entre tots els atributs.



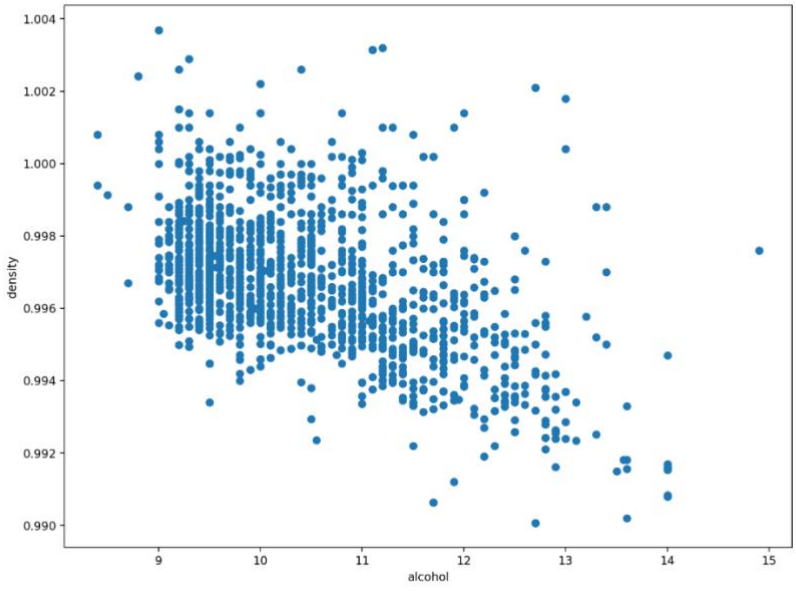
Gràfica de correlació entre les variables 'density' i 'chlorides'

Aquest atribut sembla que segueix una distribució, però trobem moltes dades que es troben amb valors molt més alts que la resta de les mostres, sembla que no ens pugui servir, però amb un tractament de dades podria cambiar les dades al nostre favor. Aplicant la regla del 3 σ obtenem la següent gràfica:



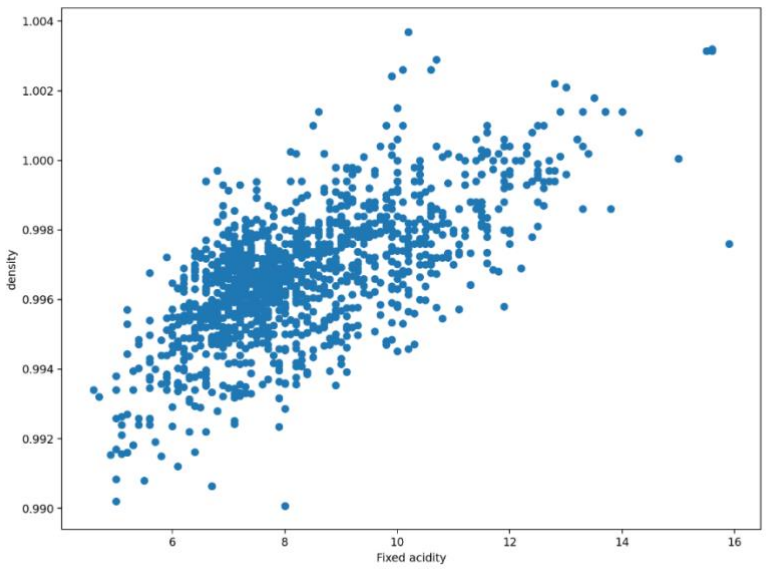
Gràfica de correlació entre les variables 'density' i 'chlorides' amb correcció 3 σ

Com podem veure, ens ha eliminat algunes dades sobre les densitats més grans i petites, però no ha cambiat els valors outliners de l'atribut 'chlorides', per tant el descartarem de la regressió ja que no té una correlació molt bona.



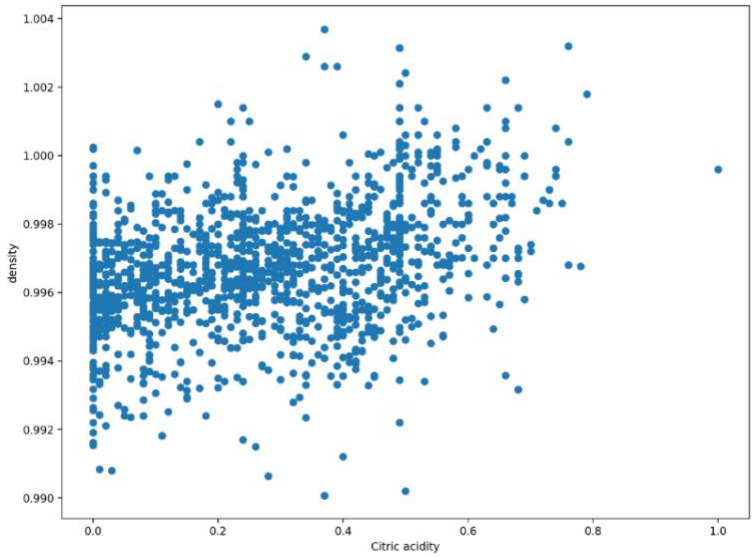
Gràfica de correlació entre les variables 'density' i 'alcohol'

L'atribut 'alcohol', sembla seguir una distribució més coherent, encara que potser més ampli del que podria ser un atribut molt bó, malgrat això, l'introduirem a la llista de candidats per a veure si pot generar bons resultats.



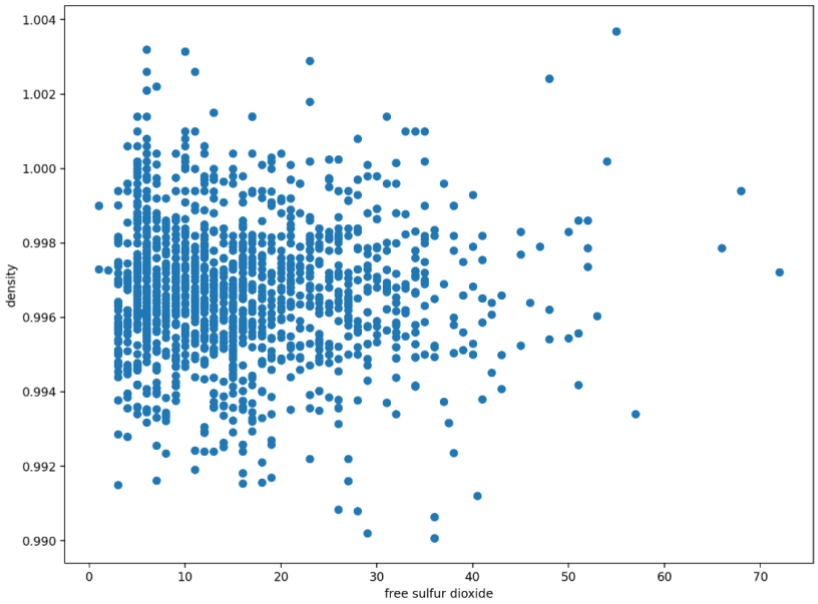
Gràfica de correlació entre les variables 'density' i 'fixed acidity'

L'atribut 'fixed acidity' sembla seguir una distribució bastant bona i coherent per a una regressió lineal, per tant estem segurs que serà un bon atribut per al nostre regressor.



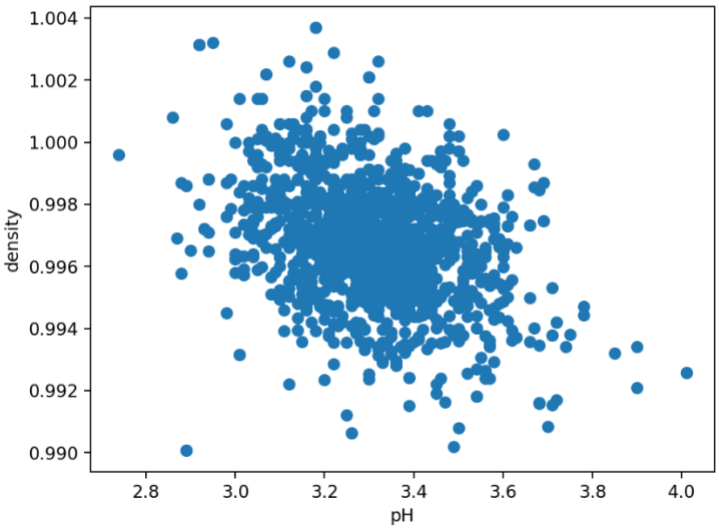
Gràfica de correlació entre les variables 'density' i 'citric acidity'

L'atribut 'citric acidity' conté moltes mostres als valors 0 amb resultats molt variats de valors respecte l'atribut 'density', i conté bastantes mostres que s'allunyen del grup general de mostres, per tant no considerarem aquest atribut part del nostre regressor.



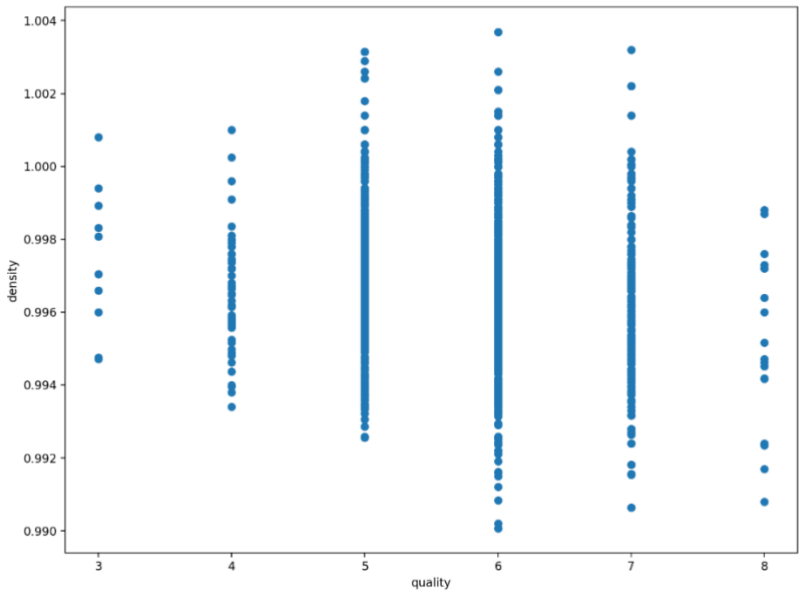
Gràfica de correlació entre les variables 'density' i 'free sulfur dioxide'

Igual que l'atribut anterior, la distribució que tenim de l'atribut 'free sulfur dioxide' és molt dispers i no sembla que ens pugui ajudar al nostre regressor. Per tant no l'utilitzarem.



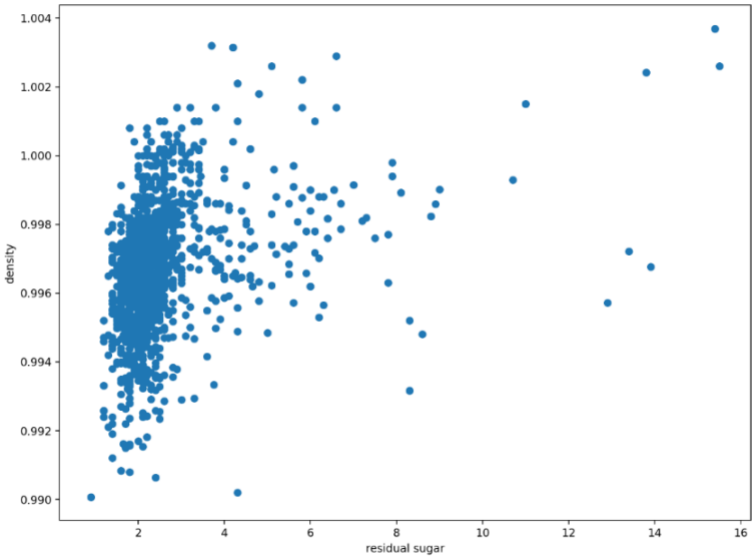
Gràfica de correlació entre les variables 'density' i 'pH'

L'atribut 'pH' no té correlació amb 'density' i forma un punt a la gràfica, per tant tampoc l'utilitzarem per a la regressió lineal.



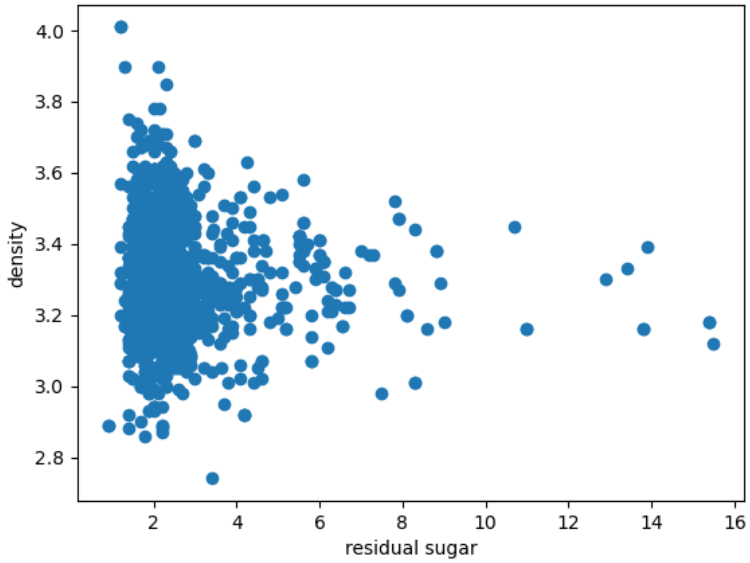
Gràfica de correlació entre les variables 'density' i 'quality'

L'atribut 'quality' és una variable discreta amb un rang de valors molt baix, per tant no ens aporta la suficient informació com a per a que sigui d'utilitat.



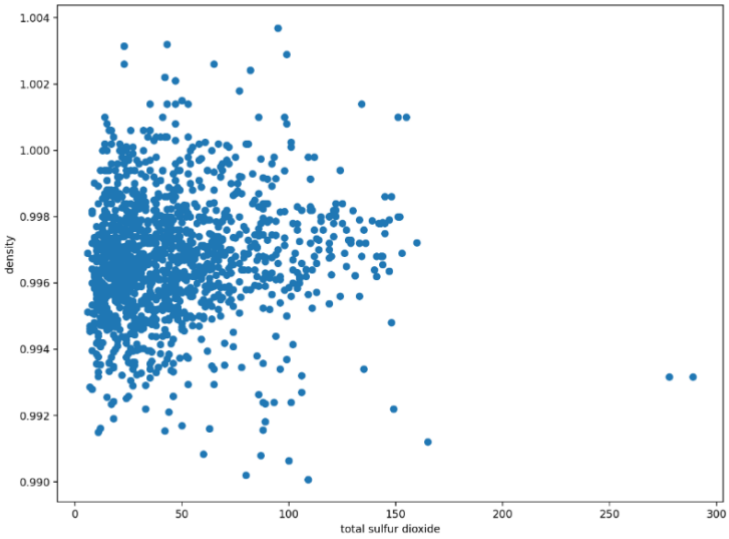
Gràfica de correlació entre les variables 'density' i 'residual sugar'

Igual que amb l'atribut 'chlorides', sembla que podria tenir una distribució útil per al nostre cas, però conté alguns valors molt distants que poden desestabilitzar el regressor, per tant intentarem aplicar la regla del 3 σ per a veure si podem aprofitar les dades:



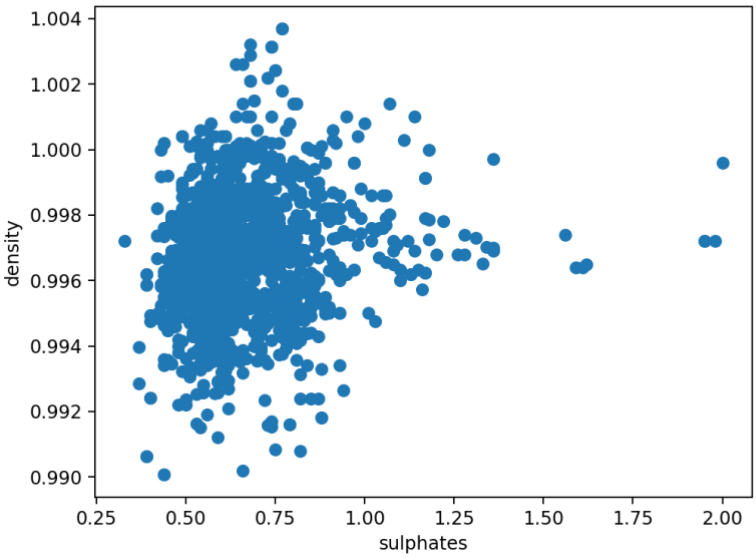
Gràfica de correlació entre les variables 'density' i 'residual sugar' amb correcció 3 σ

La regla del 3 σ ha ajudat molt a desfer-nos de les dades que no ens servien, però ara que podem observar millor com es comporta la distribució de l'atribut, podem descartar-lo ja que la majoria de mostres son totes dels mateixos valors.



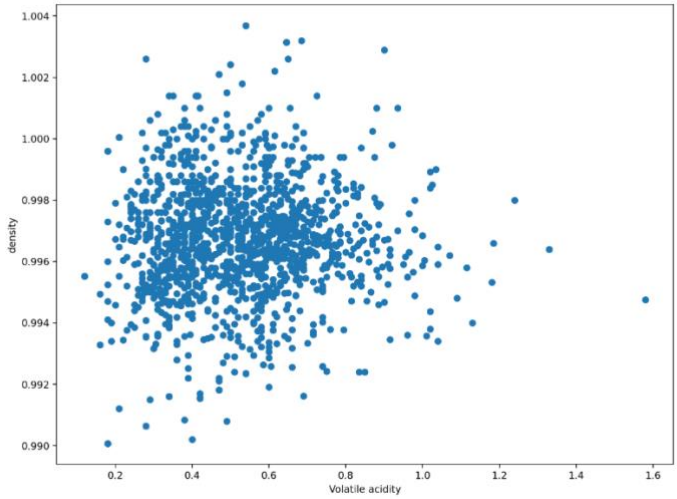
Gràfica de correlació entre les variables 'density' i 'total sulfur dioxide'

L'atribut 'total sulfur dioxide' no és un bon candidat per al regressor ja que té les mostres molt disperses.



Gràfica de correlació entre les variables 'density' i 'sulphates'

L'atribut 'sulphates' no sembla tenir una correlació important amb el nostre atribut objectiu, ja que la seva distribució forma un punt, i per tant el descartarem.



Gràfica de correlació entre les variables 'density' i 'volatile acidity'

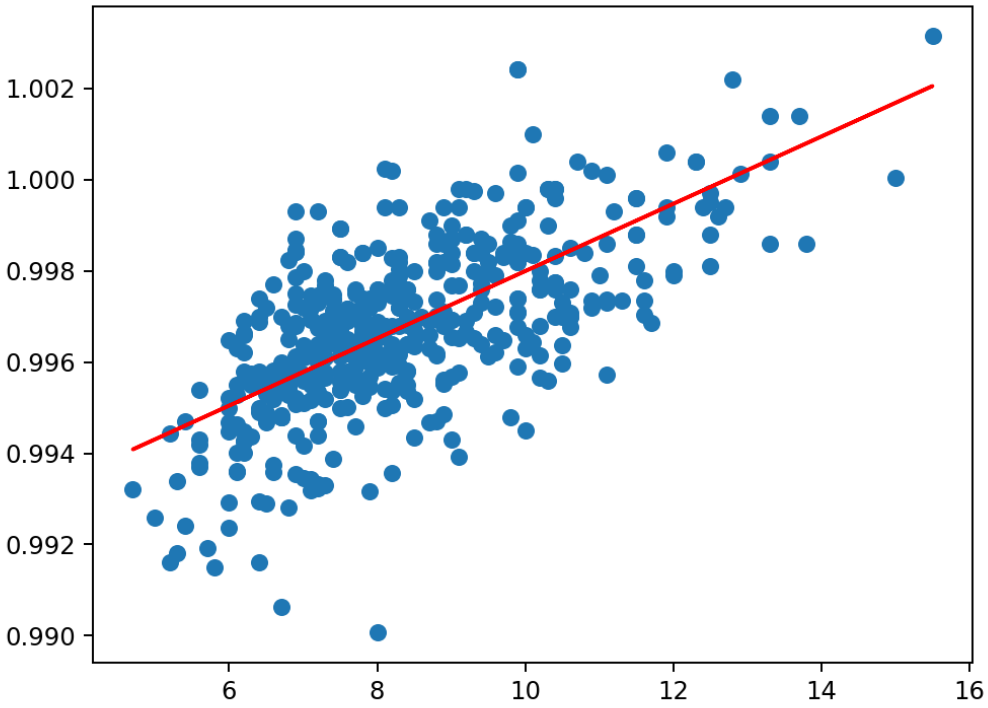
Per últim, l'atribut 'volatile acidity' té una distribució semblant a l'anterior però encara més dispers, per tant tampoc l'utilitzarem al nostre regressor.

Una vegada analitzat tots els atributs, ens hem quedat amb que els atributs més importants són: 'fixed acidity' i 'alcohol'.

Primeres regressions

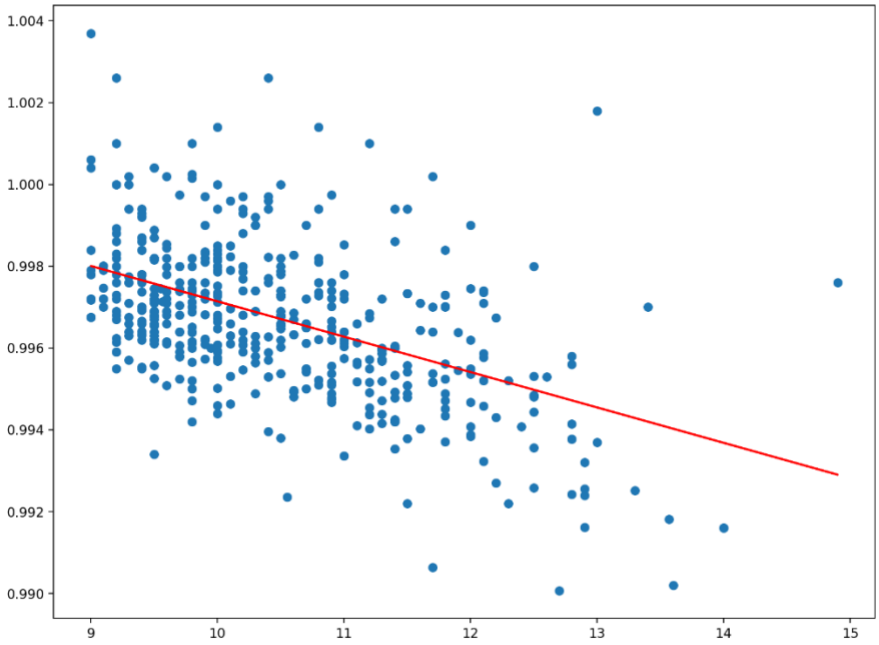
Començarem les nostres regressions amb les dades sense tractar, i utilitzarem l'error quadràtic mitjà com a mesura de la precisió del nostre regressor.

Primer visualitzarem com es comporta la regressió en els dos atributs que pensem que són els més idonis per a l'anàlisi, i els compararem amb altres dos atributs que segurament no ens serveixin.



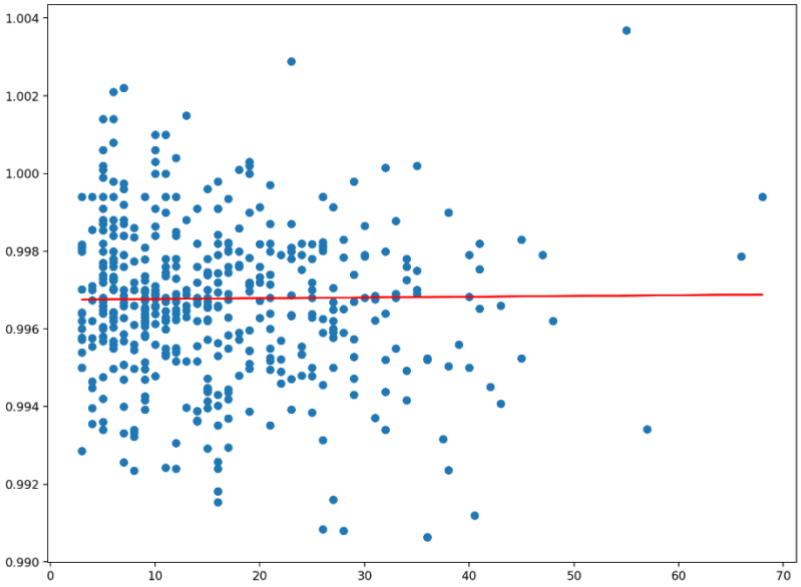
Regressió lineal de l'atribut 'fixed acidity'

El primer atribut que trobàvem com candidat era 'fixed acidity', el qual ens ha donat un error quadràtic mitjà de "1.962230173625623e-06". Sembla un bon valor i encara que per als valors més baixos de l'eix X, no és molt precís, sembla que para la resta de valors és un bon predictor.

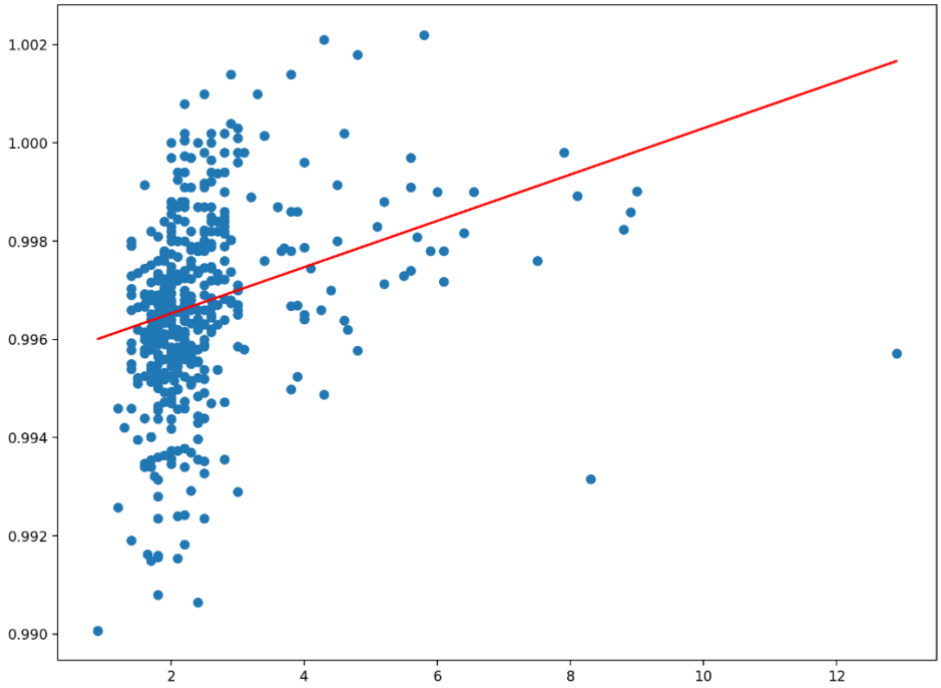


Regressió lineal de l'atribut 'alcohol'

L'atribut alcohol, semblava a primera vista que no funcionaria tan bé perquè els valors eren més dispersos, però el seu error quadràtic es semblant a 'fixed acidity', amb un valor de: "2.556753647863408e-06".

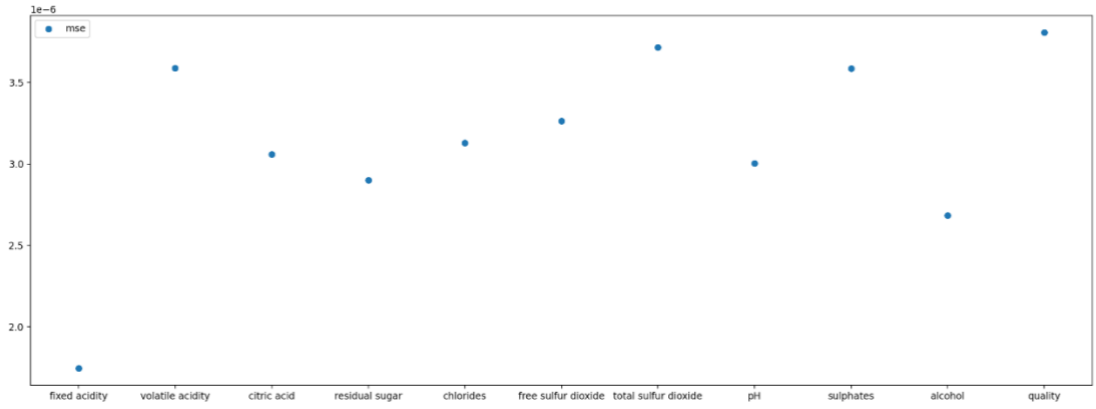


Regressió lineal de l'atribut 'free sulfur dioxide'



Regressió lineal de l'atribut 'residual sugar'

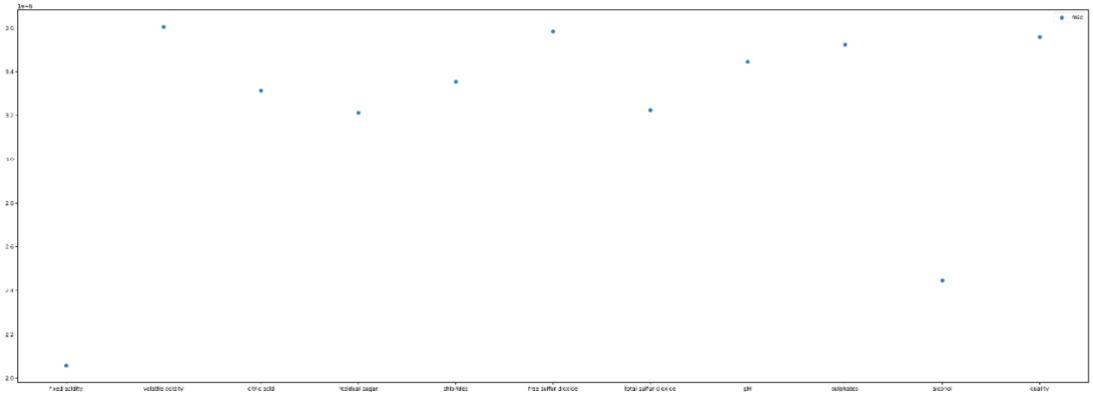
Els atributs 'residual sugar' i 'free sulfur dioxide' han donat els valors de "3.42128846525714e-06" i "4.066611982540099e-06" respectivament. Son valors que dupliquen el MSE comparat amb l'atribut 'fixed acidity', però que tampoc semblen molt llunyans, per a fer una millor comparativa, hem creat una gràfica amb els valors MSE de cada atribut:



Gràfica amb l'error quadràtic mitjà respecte a l'atribut 'density'

Podem observar a la gràfica com l'únic atribut realment diferenciant és 'fixed acidity', i que 'alcohol' sembla una mica millor que la resta, però segueix molt proper al grup principal. Per a assegurar-nos de que els valors que veiem són realment el que està passant, hem de estandarditzar les dades, ja que si no és el cas, els atributs amb rangs molt alts poden tenir un MSE molt més gran encara que siguin més precisos.

Una vegada estandarditzat les dades, hem tornar a fer una gràfica amb tots els atribut, per a veure si ara hi ha una diferència més clara:

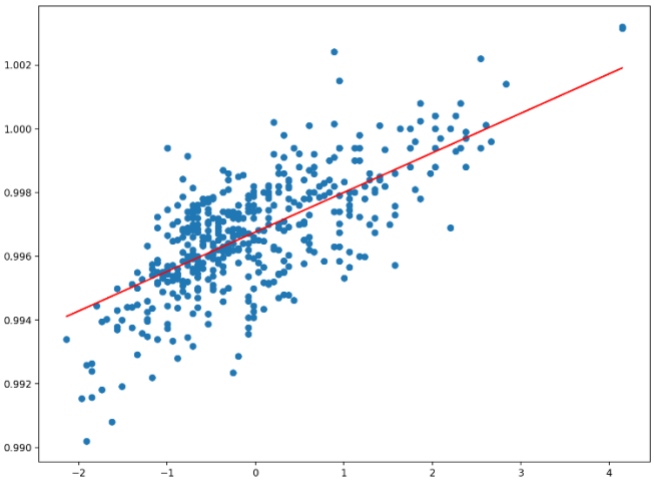


Gràfica amb l'error quadràtic mitjà amb dades estandaritzades

Amb els valors normalitzats podem observar com la diferencia entre els dos atributs òptims ('fixed acidity' i 'alcohol') i la resta dels atributs ha crescut molt i podem diferenciar els dos grups d'una forma molt més clara.

Resultats

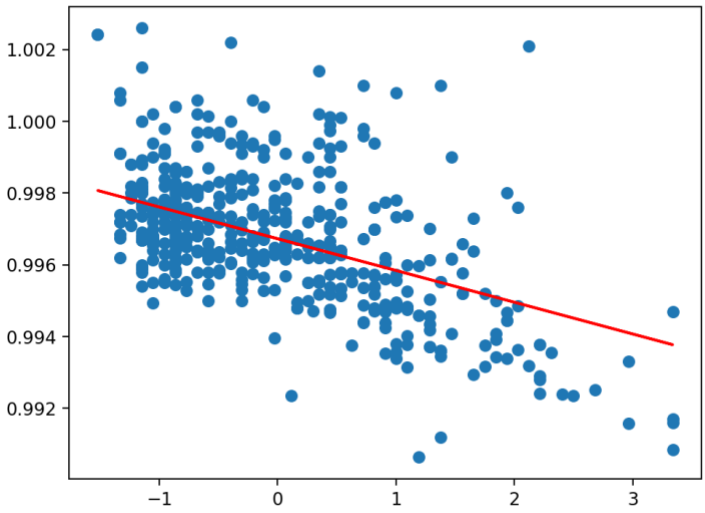
Com a regressions lineals definitives, utilitzarem les regressions anteriors però amb les dades normalitzades. Tenim tres models diferents que poden ser els millors: dues regressions lineals dels atributs 'alcohol' i 'fixed acidity' per separat, i una altre de tres dimensions amb els atributs en conjunt.



Regressió lineal de l'atribut 'fixed acidity' amb valors normalitzats

Mean squeared error: 1.7038353473252258e-06

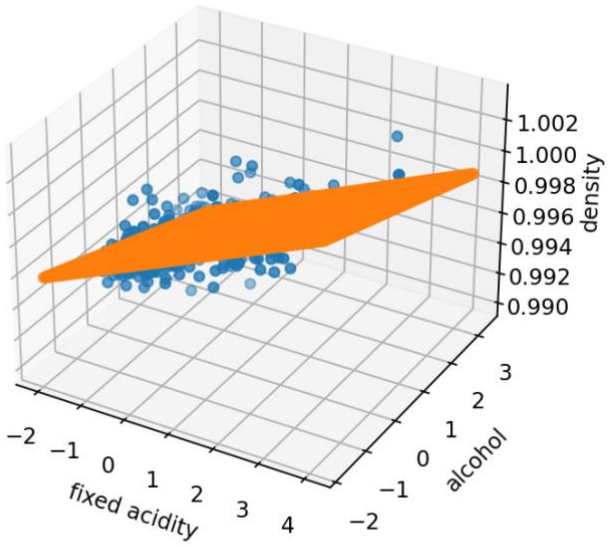
R2 score: 0.49999090261740686



Regressió lineal de l'atribut 'alcohol' amb valors normalitzats

Mean squeared error: 2.572149757827434e-06

R2 score: 0.2953867445535745



Regressió lineal conjunta dels atributs 'alcohol' i 'fixed acidity' amb valors normalitzats

Mean squeared error: 1.1575664915633185e-06

R2 score: 0.6733369024339324

Com podem observar dels tres models, l'atribut amb més correlació és 'fixed acidity', el qual obté molt bons resultats. Encara que sembli que l'atribut 'alcohol' no pot ajudar a la nostra regressió ja que de forma individual, no té tan bona correlació, a la part pràctica ens demostra que és essencial per al nostre regressor. El model en conjunt dels dos atributs ha aconseguit un resultat amb un 47,19% més de millora respecte als models individuals.

Conclusions

Aquesta pràctica ha estat molt útil per diverses raons.

Hem analitzat un dataset que no tenia dades molt correlacionades entre tots els atributs (com a màxim 0.68) per tant ens ha forçat a analitzar aquestes dades per saber quines estan relacionades i quines s'haurien de modificar per estar relacionades.

Hem hagut d'utilitzar 2 mètodes per modificar aquestes dades: l'estandarditzación i la regla del 3σ.

Aquests mètodes tenen finalitats distintes: estandarditzar modifica les dades perquè el seu rang no influeixi en l'anàlisi, i la regla del 3σ elimina valors outliers en una distribució estàndard.

Després d'escollir un atribut objectiu s'ha observat que després de realitzar aquestes modificacions hi ha hagut diverses variables que han millorat la seva correlació i d'altres que han empitjorat.

Es poden fer regressions lineals amb tots els atributs que hi havia al dataset, però no és gaire eficient i no es poden visualitzar de forma fàcil els resultats.

Una manera recomanable de realitzar regressions lineals és utilitzar un atribut escollit. D'aquesta manera es pot representar en una gràfica en 2D.

També es pot fer una regressió amb 2 atributs escollits. Es representaria amb una gràfica en 3D. Aquesta regressió té molta més precisió que l'anterior amb només un atribut.

Fer servir massa atributs pot fer que hi hagi overfitting, per tant, és important utilitzar només el nombre òptim d'atributs per obtenir la precisió desitjada.