Pràctica 2: Classificació

Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740

Introducció

En aquesta pràctica analitzarem una base de dades sobre la potabilitat de l'aigua i les diferents concentracions d'algunes substàncies. Veurem quines distribucions tenen els atributs, la relació entre ells i intentarem determinar si donada una mostra, aquesta és potable o no.

Apartat B

Exploratory Data Analysis

Els atributs que tenim a la base de dades són els següents:

0. pH [float]: pH de l'aigua.
1. Hardness [float]:concentració de calci i magnesi.
2. Solids [float]: edat del jugador mesurada en anys i dies.
3. Chloramines [float]: concentració de cloramines.
4. Sulfate [float]: concentració de sulfats.
5. Conductivity [float]: conductivitat de l'aigua.
6. Organic_carbon [float]: concentració de compostos orgànics.
7. Trihalomethanes [float]: concentració de trihalometans.
8. Turbidity [float]: Terbolesa de l'aigua.
9. Potability [integer]: 0 si no és potable, 1 si és potable.

Tenim 10 atributs a la base de dades, tots de tipus float menys l'últim que és binari, per tant l'atribut categòric només pot prendre dos valors.

Pel que fa a les correlacions amb l'atribut categòric, totes són poc significatives com es pot veure en la figura 1.

A continuació veurem la distribució de la potabilitat 2 i les distribucions dels altres atributs segons la potabilitat 3.

A simple vista podem veure en la figura 3 que els atributs tenen una distribució gaussiana o pràcticament gaussiana. Les etiquetes de potabilitat estan balancejades (proporció 3:2).

Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740

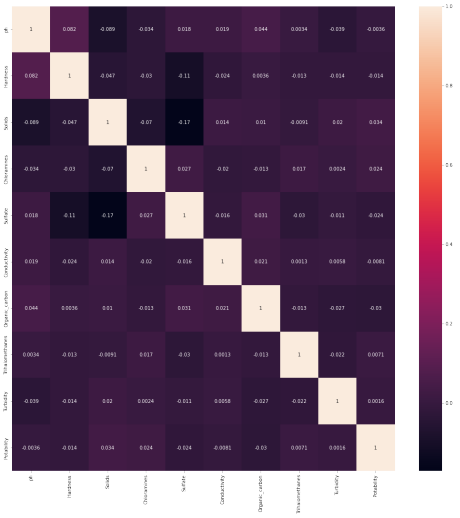


Figura 1: Correlació entre els atributs

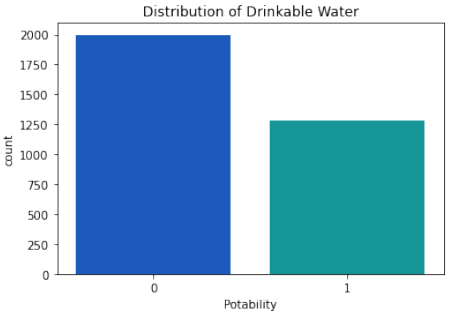


Figura 2: Distribucio de la potabilitat

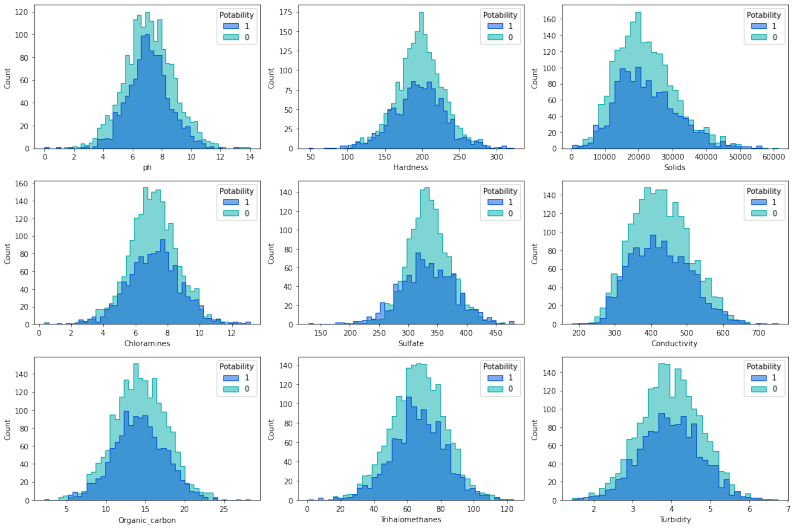


Figura 3: Distribucions dels atributs segons la potabilitat

Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740

Preprocessing

En primer lloc mirem quines variables tenen NaNs, que són pH, Sulfate i Trihalomethanes. Els valors mitjans dels atributs, tant quan l'aigua és potable com no, són molt semblants, per tant, substituim els NaNs per la mitja dels atributs.

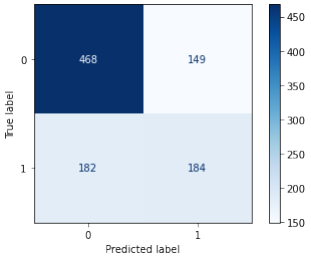
La variable categòrica pren només dos valors, per tant tècniques com Ordinal Encoder o One-HotEncoder no tenen gaire sentit en aquest context.

Hi ha rangs de variables molt aixos en comparació amb uns altres. L'atribut solids té el rang més alt de l'ordre de $6 \cdot 10^6$, mentre que el rang del pH és de 14 i el de Turbidity és de 3. És per això que cal normalitzar les dades. La normalització que hem fet servir és la estàndar.

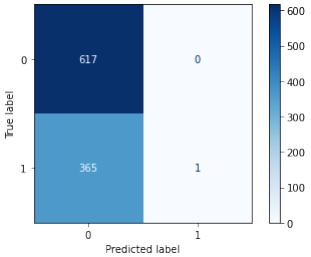
No tenim files repetides a la base de dades.

Model Selection

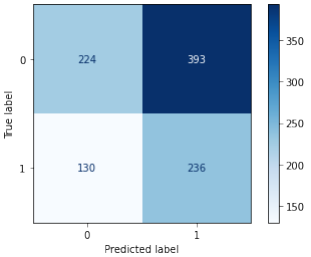
Els models que hem fet són el knn amb $k = 1..4$, regressió logística, SVM (lineal, polinomial, sigmoid, rbf), perceptró, random forest, bagging en decision trees i svc i, finalment, boosting.



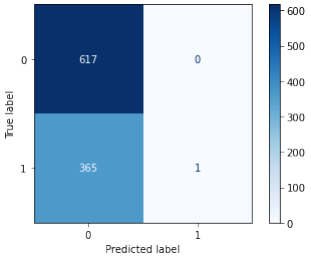
Knn per $k = 1..4$



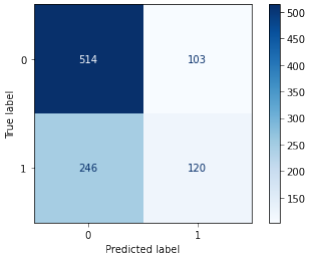
Regressió logística



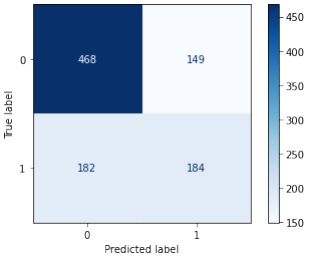
Perceptró



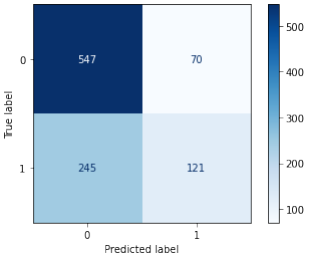
SVM linear kernel



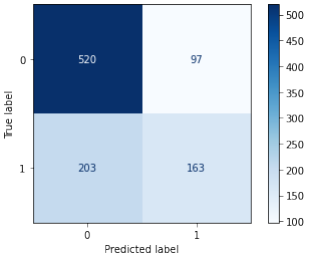
SVM polynomial kernel



SVM sigmoid or rbf kernel

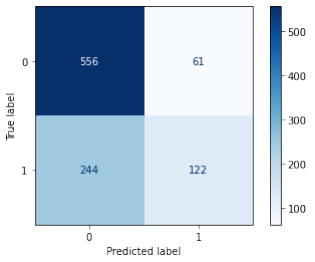


Random forest

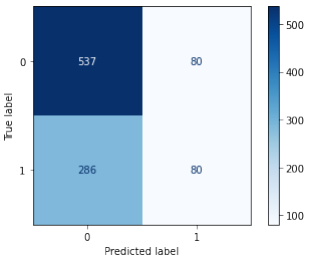


Bagging en decision trees

Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740



Bagging en SVC



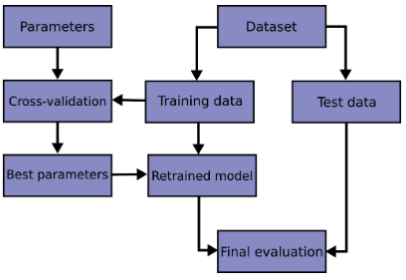
Boosting

F1 scores obtinguts:

	SENSE PCA	PCA	PCA+POLYNOMIAL
LOGISTIC REGRESSION	0	0.005	0.46
SVM LINEAR	0.005	0	0.31
SVM POLY	0.4	0.326	0.22
SVM RBF	0.52	0.41	0.41
SVM SIGMOID	0.52	0.31	0.38
KNN	0.35	0.31	0.38
PERCEPTRO	0.47	0.47	0.47
RF	0.458	0.44	0.44
BAGGING TREE	0.52	0.52	0.48
BAGGING: SVC	0	0.44	0.43
BOOSTING	0.3	0.3	0.47

Crossvalidation

És molt important separar el conjunt d'entrenament del conjunt de test, ja que un model que sencillament repeteix les mateixes dades tindria un resultat perfecte, però no podrà predir res útil. Això es l'overfitting, per evitar-ho cross-validem els resultats per crear conjunts d'entrenament i conjunts de test diferents sobre un mateix model.



Els resultats seran millors amb un conjunt d'entrenament més gran fins a cert punt, amb poques dades d'entrenament l'accuracy serà menor i el volem el més alt possible, però si ens passem amb el tamany del conjunt d'entrenament podem arribar a comprometre els resultats, si les categories estan desbalancejades, o el rendiment.

Hem fet servir el k-fold normal, en comparació amb el punt anterior hem aconseguit augmentar l'accuracy utilitzat k=5, aquest augmenta ja que hem entrenat el model amb 5 conjunts d'entrenament i test diferents

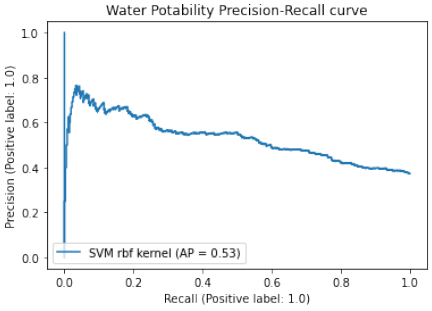
No és eficient aplicar Leave-One-Out ja que prioritza que els conjunts d'entrenament siguin els més grans possibles, i en models amb moltes dades (+3000 en el nostre cas) compromet el rendiment consumint més temps del necessari.

Hem calculat els accuracys per SVM, knn, regressió logística i random forest que són respectivament 0.61, 0.56, 0.61, 0.63.

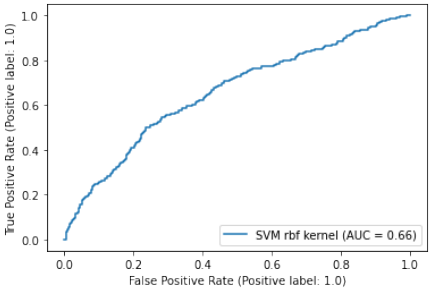
Metric Analtyics

Vegem les mètriques pel SVM de rbf kernel, que és el que millor F1 score ens ha donat.

	precision	recall	f1-score	support
Non-potable water	0.72	0.76	0.74	617
Potable water	0.55	0.50	0.53	366
accuracy			0.66	983
macro avg	0.64	0.63	0.63	983
weighted avg	0.66	0.66	0.66	983



PR curve

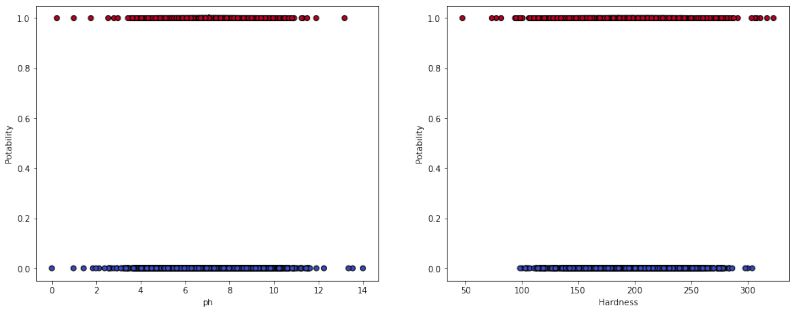


ROC curve

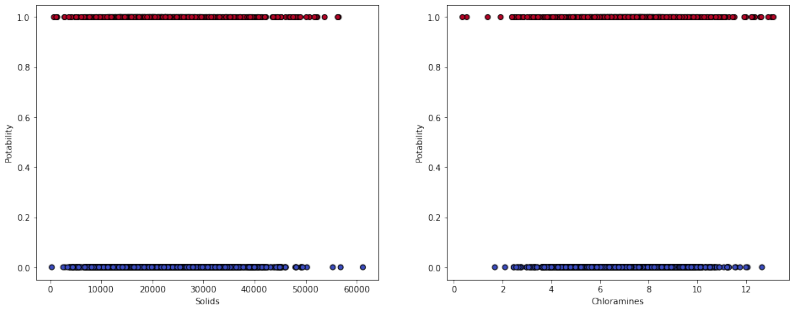
Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740

Apartat A

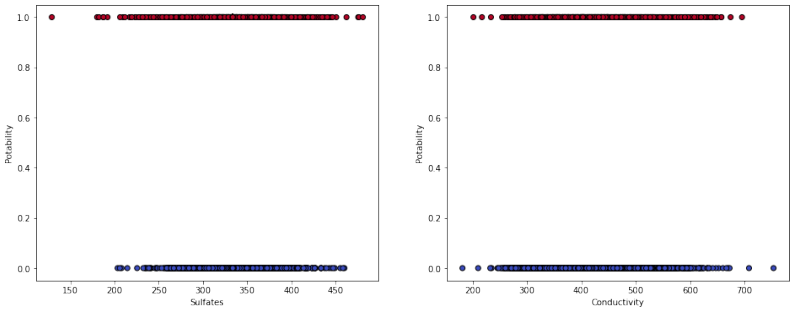
A continuació veurem la distribucio de la potabilitat segons els atributs. És molt difícil que puguem tenir un bon model, ja que tots els punts estan molt solapats i els atributs tenen poca correlació entre ells.



pH i Hardness

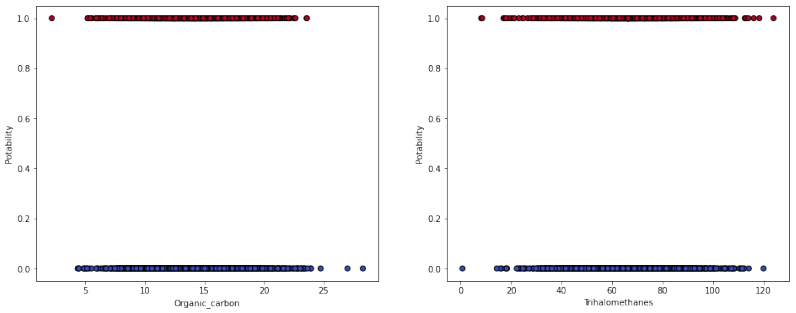


Solids i Cloramines

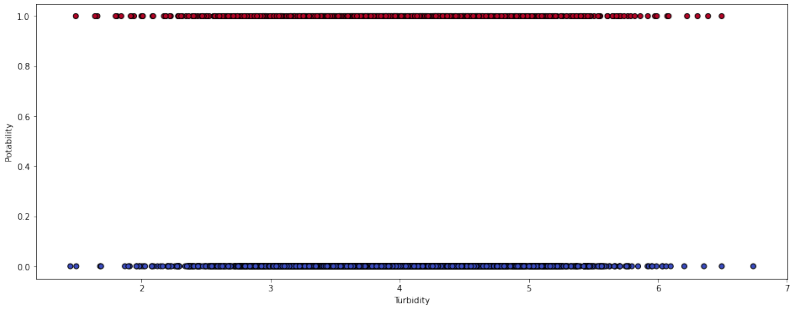


Sulfats i Conductivitat

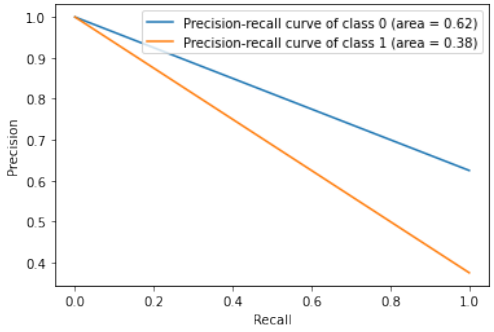
Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740



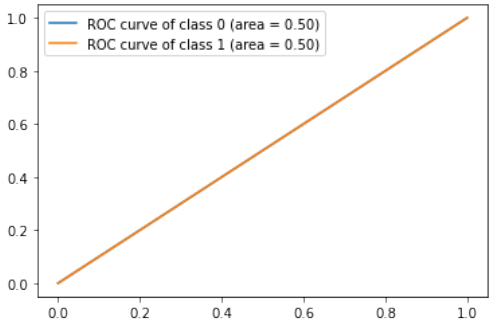
Organic Carbon i Trihalomethanes



Turbidity



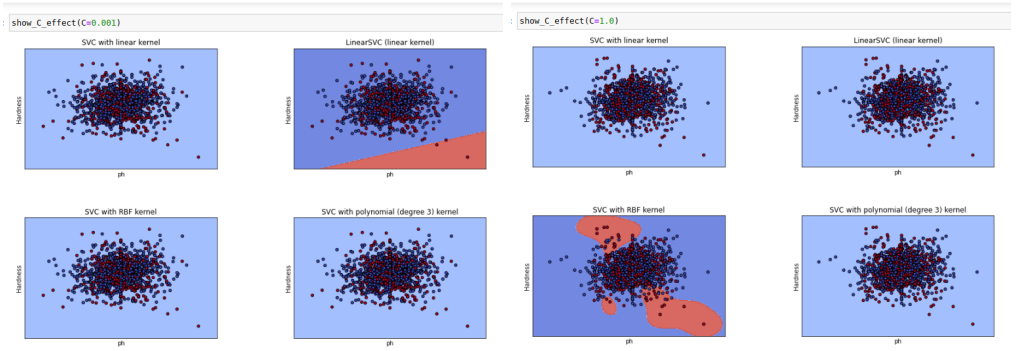
PR curve amb SVM rbf



ROC curve amv SVM rbf

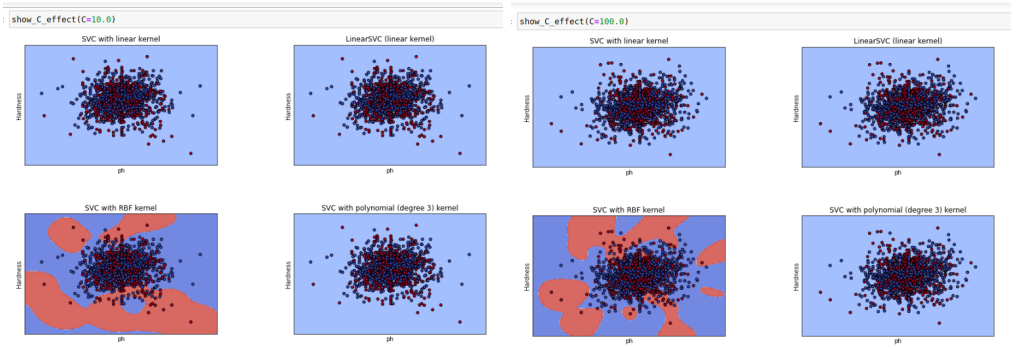
Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740

També hem vist els efectes de variar els valors de regularització (C, degree i γ):



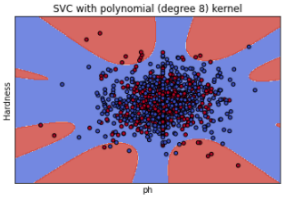
C=0.001

C=1

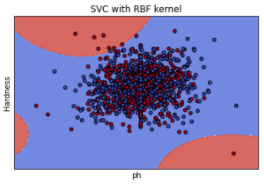


C=10

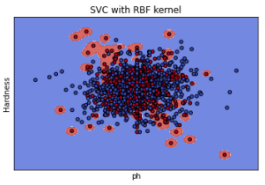
C=100



degree=8



γ=0.1



gamma=10