Rapport Projet Science des données

1. Présentation de la problématique du projet et des données

On dispose d'un jeu de données correspondant à des propriétés chimiques de vins, provenant d'un jeu de données disponible librement sur le dépôt de données de l'*University of California Irvine 1*. L'objectif de ce projet est de pouvoir prédire la qualité d'un vin à partir des données mises à disposition.

Les données sont définies dans un fichier ".csv", possédant les valeurs de 12 propriétés chimiques différentes caractérisant chaque vin, et séparées par le caractère ','.

Ces propriétés chimiques sont respectivement :

- "fixed acidity"
- "volatile acidity"
- "citric acid"
- "residual sugar"
- "chlorides"
- "free sulfur dioxide"
- "total sulfur dioxide"
- "density"
- "pH"
- "sulphates"
- "alcohol"
- "quality"

2. Analyse des données du problème

Tout d'abord, étant donné que nous cherchons à prédire la qualité d'un vin à partir des données mises à disposition, la classe à prédire est donc la "quality" du vin. Elle ne peut prendre que deux valeurs possibles (-1 ou 1), ce qui nous permet donc d'affirmer qu'il s'agit d'un problème de classification binaire.

Ensuite, il est important d'analyser les données pour identifier et corriger les données manquantes ou aberrantes. Une donnée manquante est représentée dans le fichier CSV par deux guillemets consécutifs, et par la valeur NaN dans un langage informatique comme python dans notre cas, comme le montre l'exemple ci-dessous :

```
54 6.6,"0.5","","2.1","0.068","6.0","14.0","0.9955","3.39","0.64","9.4","1"
```

```
fixed acidity           6.6000
volatile acidity        0.5000
citric acid                NaN
residual sugar          2.1000
chlorides               0.0680
free sulfur dioxide     6.0000
total sulfur dioxide   14.0000
density                 0.9955
pH                      3.3900
sulphates               0.6400
alcohol                 9.4000
quality                 1.0000
```

En ce qui concerne les données aberrantes, nous n'avons pas les connaissances nécessaires en oenologie pour affirmer qu'une valeur, supérieure à 3 écarts-types de la moyenne, soit réellement une valeur aberrante ou une valeur réelle mais hors-normes.

Dans le cadre de ce projet, nous avons pris le choix de conserver ces valeurs. En revanche, nous savons que le pH est une propriété chimique prenant des valeurs de 0 à 14, c'est pourquoi nous pouvons éliminer toutes données dont le pH n'est pas contenu dans cet interval.

Les histogrammes par attributs, disponibles en Annexe 1, permettent de visualiser les distributions des données restantes et ainsi de repérer quelles valeurs sont les plus exprimées par attribut.

On sait que deux attributs sont significativement corrélés si la valeur absolue du coefficient de corrélation > 0.5. Or cela peut poser problème chez certains algorithmes de classification qui peuvent être sensible à la corrélation, et donc produire un modèle moins performant. On choisit d'afficher les coefficients de corrélation significatifs entre deux attributs pour avoir une idée de quels attributs pourraient être supprimés pour possiblement obtenir de meilleurs résultats :

Corrélations absolues significatives entre les différents attributs

	Fixed acidity	Volatile acidity	Citric acidity	Free sulfur dioxide	Total sulfur dioxide	Density	pH
Fixed acidity			0.671739			0.667610	0.684950
Volatile acidity			0.554347				
Citric acidity	0.671739	0.554347					0.540597
Free sulfur dioxide					0.668295		
Total sulfur dioxide				0.668295			
Density	0.667610						
pH	0.684950		0.540597				

Pour avoir un meilleur aperçu du type de corrélation qu'il existe entre ces attributs, vous pouvez trouver en Annexe 2 les graphiques de chacune de ces corrélations testées. On remarque alors que ces corrélations sont de types linéaires.

Avec ces résultats, on peut se demander si supprimer l'attribut "*fixed acidity*" pourrait potentiellement améliorer le modèle produit pour certains classifieurs parce que cet attribut est fortement corrélé à certains autres attributs. On pourrait également essayer de supprimer l'attribut "*free sulfur dioxide*" ou "*total sulfur dioxide*" car un seul suffirait à être pertinent. Nous verrons si cela est vrai par la suite lors de l'analyse des résultats et de l'influence des différents facteurs.

Certains classifieurs sont sensibles à la proportion d'individus de chaque classe et peuvent produire un modèle de mauvaise qualité si celle-ci est déséquilibrée. Après préparation des données, c'est-à-dire suppression des données manquantes et aberrantes, on obtient les proportions suivantes : proportion de 1 = 54% et proportion de -1 = 46%. On peut donc en conclure que les classes sont assez bien équilibrées.

Enfin, la normalisation des données peuvent être nécessaire chez certains classifieurs pour réaliser des modèles performant à partir de leurs modèles mathématiques sous-jacent. Les données restantes ont donc subit une normalisation.

Finalement, on se retrouve avec 1566 individus différents.

3. Description du protocole expérimental mis en place

On dispose, après préparation des données, de données normalisées. On souhaite réaliser une cross-validation afin de permettre à nos classifieurs de pouvoir s'entraîner et produire un modèle à partir de données d'entrainements, puis de les évaluer à l'aide de métriques grâce aux données de tests restantes. Il est donc primordiale pour une bonne comparaison d'utiliser les mêmes données d'entrainements et de tests entre classifieurs.

Pour effectuer la cross-validation, nous avons tout d'abord divisé nos données en différents jeux (ou plis) de tests et d'entraînements en utilisant la méthode du Stratified KFold à 10 itérations. Cela permet d'éviter un sur-apprentissage des classifieurs et de respecter la stratification, c'est à dire d'assurer une distribution des classes dans les plis fidèle au jeu de données complet. On se retrouve alors avec les données suivantes :

- X_train : la liste de chaque plis d'entraînements, utilisé par le classifieur pour créer son modèle.
- y_train : la liste des valeurs de qualité réelle correspondant respectivement à chacun des plis d'entraînements, utilisé par le classifieur pour créer son modèle.
- X_test : la liste de chaque plis de tests, utilisée par le classifieur pour prédire une qualité à partir du modèle créé.
- y_test : la liste des valeurs de qualité réelle correspondant aux plis de tests, utilisée pour tester et évaluer le classifieur par comparaison avec les qualités prédites.

Dans le cadre de ce projet, nous allons tester les classifieurs suivants :

- Régression logistique (Logistic regression)
- Machines a vecteurs supports (SVC avec max_iter = 10000)
- Analyse discriminante linéaire (Linear discriminant analysis)
- Analyse discriminante quadratique (Quadratic discriminant analysis)
- K-plus proches voisins (K-nearest neighbors classification)
- Arbres de décision (Decision trees)
- Perceptron (avec epochs=500, learning_rate=0.01)

Comme expliqué précédemment, on commence par entraîner chaque classifieur avec chacun des plis de X_train et y_train créés pour produire un modèle par cross-validation. Ce modèle est ensuite testé à l'aide des plis de X_test afin d'obtenir des qualités prédites. Ces qualités sont comparées aux vraies qualités correspondantes (y_test) en calculant les métriques suivantes :

- Précision
- Recall
- F1 score
- Accuracy

Puisqu'on calcule ces métriques pour chaque jeux de tests, on réalise la moyenne des résultats obtenus pour avoir des résultats de métriques plus précis.

4. Résultats expérimentaux obtenus

Les résultats obtenus seront présentés sous forme de tableaux regroupant chaque classifieur et chaque métrique correspondante. Le plus grand résultat pour chaque métrique sera mis en valeur par la coloration de la cellule correspondante en jaune. Chaque tableau représente une situation différente. On retrouvera donc, dans l'ordre :

- Résultats obtenus avec les données aberrantes retirées et sans normalisation
- Résultats obtenus avec les données aberrantes retirées et normalisation
- Résultats obtenus avec les données aberrantes et les valeurs supérieures à 3 écarts-types de la moyenne retirées et normalisation
- Résultats obtenus avec les données aberrantes, "fixed acidity" retirés et normalisation
- Résultats obtenus avec les données aberrantes, "fixed acidity" et "free sulfur dioxide" retirés et normalisation

1/Résultats obtenus avec les données aberrantes retirées et sans normalisation

	Précision	Recall	F1 score	Accuracy
Régression logistique	0.7646076819335 1	0.7362324929971988	0.7435216439962973	0.7343867385268659
SVC	Ne converge pas			
Analyse discriminante linéaire	0.7667390341090918	0.7362184873949 58	0.7434827934040067	0.7356606238771844
Analyse discriminante quadratique	0.7266725079039789	0.7718767507002802	0.7419875345615331	0.7145394414502695
K-plus proches voisins	0.6226537528756054	0.6234033613445378	0.6158784436956447	0.5918544830965212
Arbres de décision	0.6666112473590428	0.6602521008403361	0.6593960898145832	0.6411481300016331
Perceptron	0.7291199066392877	0.7042296918767508	0.6878133543931 97	0.6832965866405357

2/Résultats obtenus avec les données aberrantes retirées et normalisation

	Précision	Recall	F1 score	Accuracy
Régression logistique	0.7628833490083109	0.7445518207282913	0.7467058950318171	0.7363016495 1821
SVC	0.7650980065491467	0.7397899159663865	0.7451417640417775	0.7363016495182101
Analyse discriminante linéaire	0.7667390341090918	0.7362184873949958	0.7434827934040067	0.7356606238771844
Analyse discriminante quadratique	0.7266725079039789	0.7718767507002802	0.7419875345615331	0.7145394414502695
K-plus proches voisins	0.6913392065964394	0.7375490196078431	0.7057607952812984	0.6769108280254778
Arbres de décision	0.6626357675071856	0.6578711484593838	0.6562996855041525	0.6373183080189448
Perceptron	0.7745714237049759	0.5878011204481793	0.6585383895881509	0.6864486362894007

3/Résultats obtenus avec les données aberrantes et les valeurs supérieures à 3 écarts-types de la moyenne retirées et normalisation

	Précision	Recall	F1 score	Accuracy
Régression logistique	0.7580127574840508	0.7526973026973028	0.7448892916260922	0.7311385797301291
SVC	0.7674885810117255	0.7475024975024975	0.7468603521991227	0.7374667585935192
Analyse discriminante linéaire	0.7716441973252365	0.7462037962037963	0.7468219749537753	0.7388702846449325
Analyse discriminante quadratique	0.7399146900846458	0.7292873792873793	0.7208234362753386	0.7072638628976657
K-plus proches voisins	0.7004351339855133	0.7268231768231769	0.7028727252771974	0.6784644932532257
Arbres de décision	0.6627616748921992	0.6478354978354978	0.6484154032523388	0.6320890377228406
Perceptron	0.7056197128633885	0.6076090576090575	0.6444065729152629	0.651024327784891

4/Résultats obtenus avec les données aberrantes, “fixed acidity” retirées et normalisation

	Précision	Recall	F1 score	Accuracy
Régression logistique	0.7666594716759618	0.7481232492997199	0.7505360840401764	0.7407806630736566
SVC	0.7721819935588243	0.7421848739495798	0.7499231127184574	0.7420463824922423
Analyse discriminante linéaire	0.7772999579673613	0.7421848739495798	0.7512452606945953	0.7445941531928792
Analyse discriminante quadratique	0.7225452422902949	0.7813865546218486	0.7416866425666147	0.7126449452882574
K-plus proches voisins	0.7020306852450935	0.7482212885154061	0.7169071486512034	0.6877429364690512
Arbres de décision	0.6712661355613829	0.6756862745098039	0.6695580979225966	0.6494937122325657
Perceptron	0.7633286288631962	0.6103781512605042	0.6705920186533906	0.6909276498448473

5/Résultats obtenus avec les données aberrantes, “fixed acidity” et “free sulfur dioxide” retirées et normalisation

	Précision	Recall	F1 score	Accuracy
Régression logistique	0.7568072348358513	0.7480812324929971	0.7470547411784442	0.7356565409113 18
SVC	0.7616161503361452	0.7433333333333333	0.7467342320051349	0.7369304262616365
Analyse discriminante linéaire	0.7657768930161184	0.7397619047619047	0.7469189154172445	0.7382083945778213
Analyse discriminante quadratique	0.7285692678334504	0.7826050420168067	0.7489257792440236	0.7222154172791115
K-plus proches voisins	0.6885506097231298	0.7362605042016807	0.7039290820097283	0.6749591703413 36
Arbres de décision	0.6708648604406326	0.6685434173669468	0.6659924754084171	0.6482116609505144
Perceptron	0.7940808379254957	0.6067226890756302	0.6794062767817362	0.7075167401600522

5. Conclusion et analyse critique des résultats obtenus

Après analyse des résultats obtenus, on remarque tout d'abord que la normalisation des données est importante pour un meilleur fonctionnement des classifieurs comme le SVC notamment.

En revanche, la suppression des valeurs à plus de 3 écarts-types apportent des légères améliorations de métriques pour certains classifieurs, mais ce n'est pas le cas pour tous.

Selon la préparation des données et la métrique privilégiée, le choix du classifieur ne sera pas forcément le même, comme présenté dans le tableau ci-dessous.

Meilleurs classifieurs selon les différentes situations et métriques

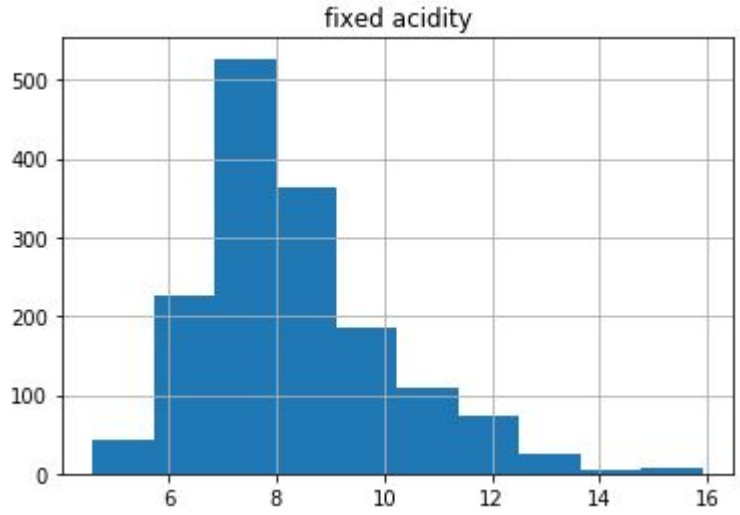
	Précision	Recall	F1 score	Accuracy
Situation 1	Analyse discriminante linéaire	Régression logistique	Régression logistique	Analyse discriminante linéaire
Situation 2	Perceptron	Analyse discriminante quadratique	Régression logistique	SVC
Situation 3	Analyse discriminante linéaire	Régression logistique	SVC	Analyse discriminante linéaire
Situation 4	Analyse discriminante linéaire	Analyse discriminante quadratique	Analyse discriminante linéaire	Analyse discriminante linéaire
Situation 5	Perceptron	Analyse discriminante quadratique	Analyse discriminante quadratique	Analyse discriminante linéaire

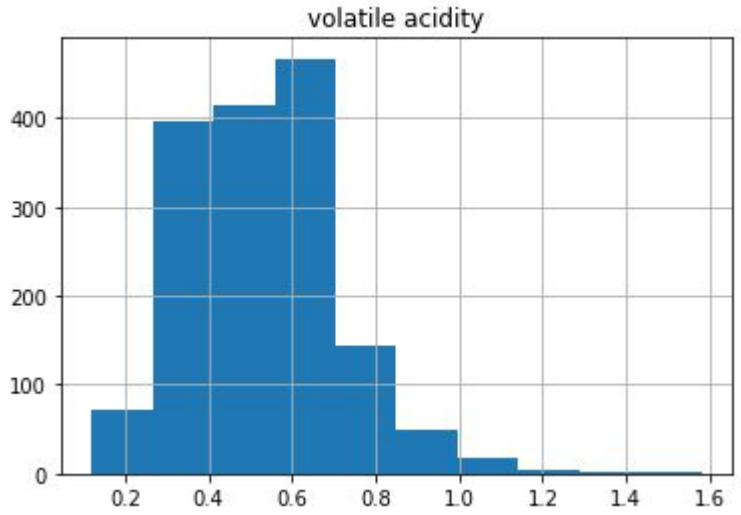
En général, on remarque que le classifieur à "*Analyse discriminante linéaire*" est le plus performant dans la majorité des situations, mais la situation 4 semble être préférable aux autres car ce classifieur atteint le maximum de F1 score et accuracy, et la deuxième meilleure précision toutes situations confondues.

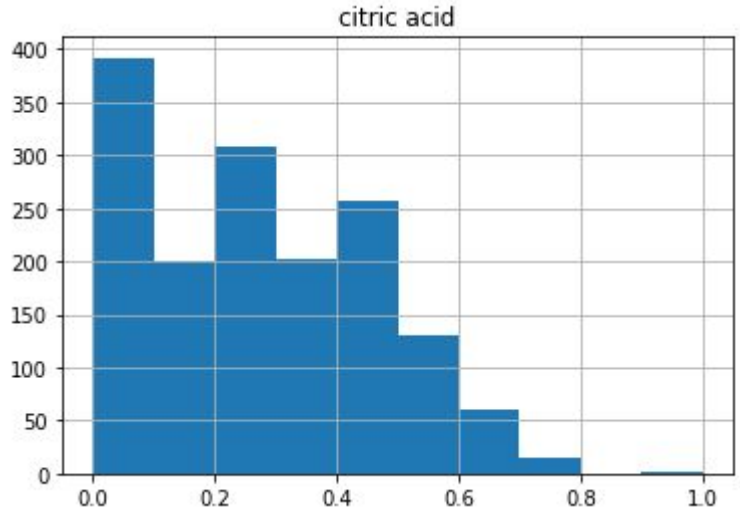
L' accuracy obtenue à la situation 4 est 0.7445941531928792. Si nous voulons avoir des classifieurs avec une accuracy accrue, il nous faudrait donc un échantillon plus large, ainsi les classifieurs pourront avoir plus de données d'entraînement et de test pour produire des modèles plus fiables.

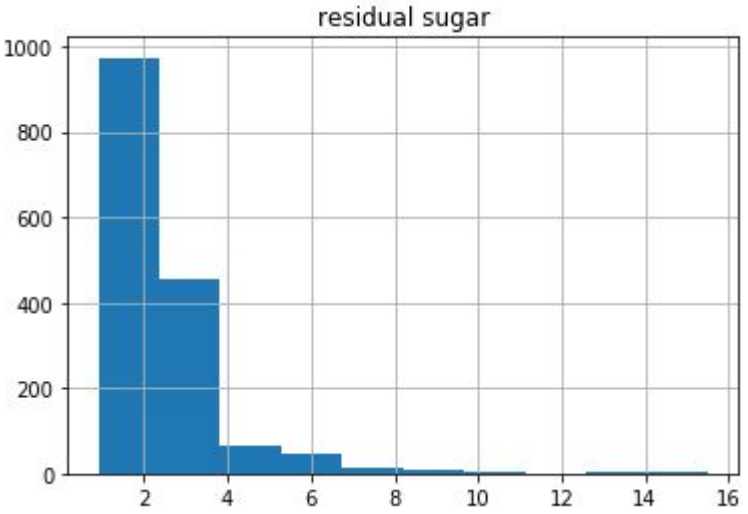
ANNEXES

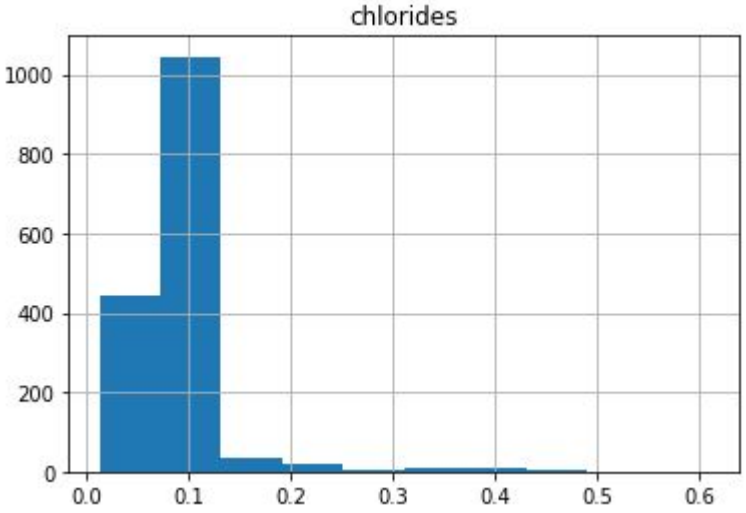
1/Histogrammes représentant la distribution des attributs

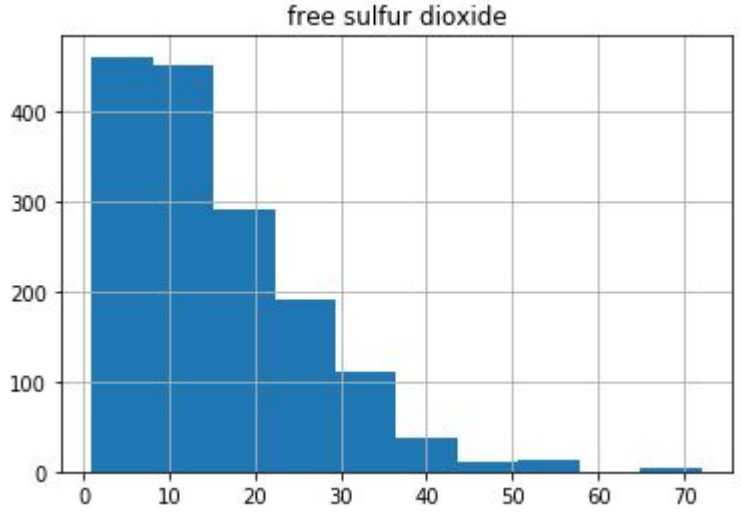


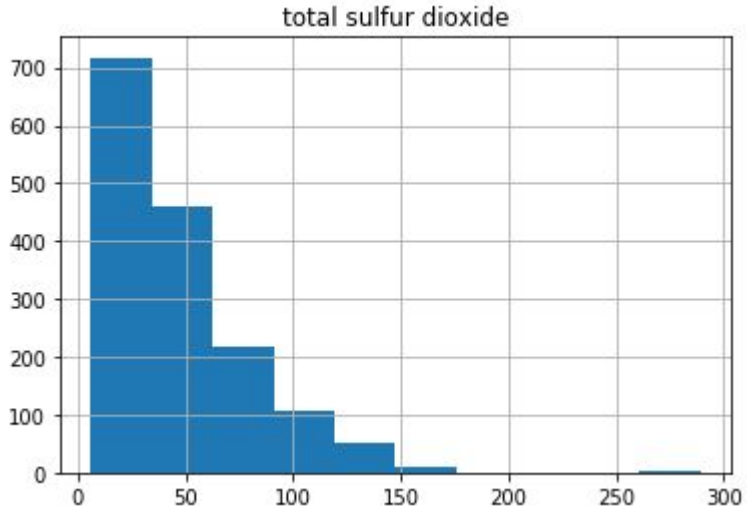


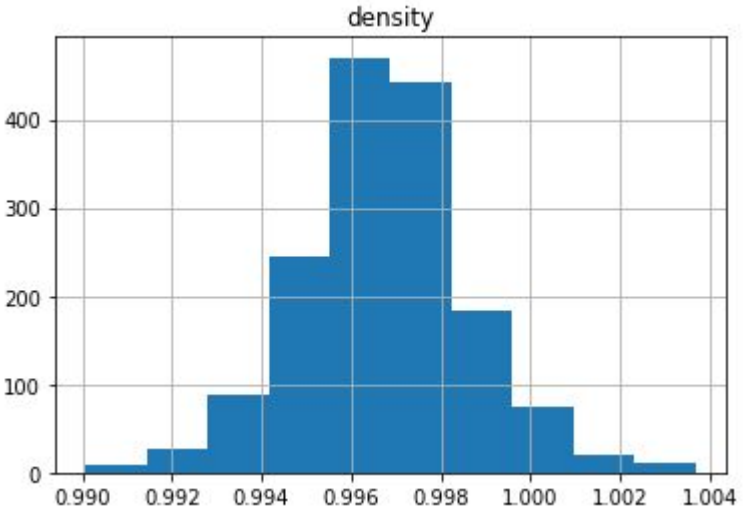


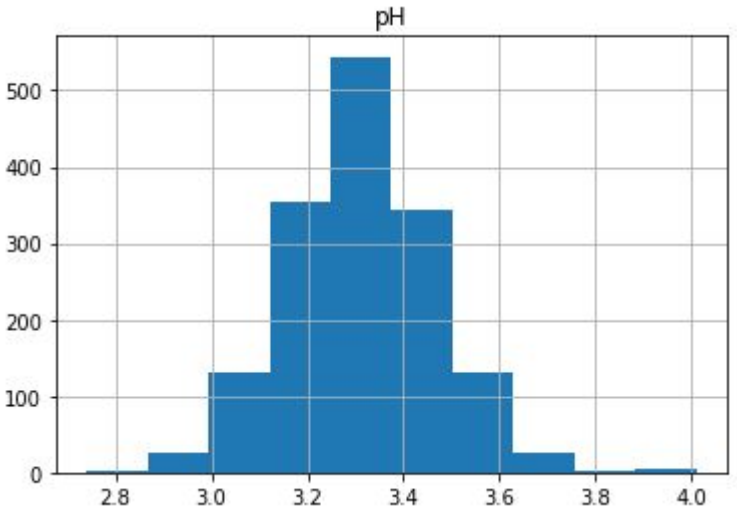


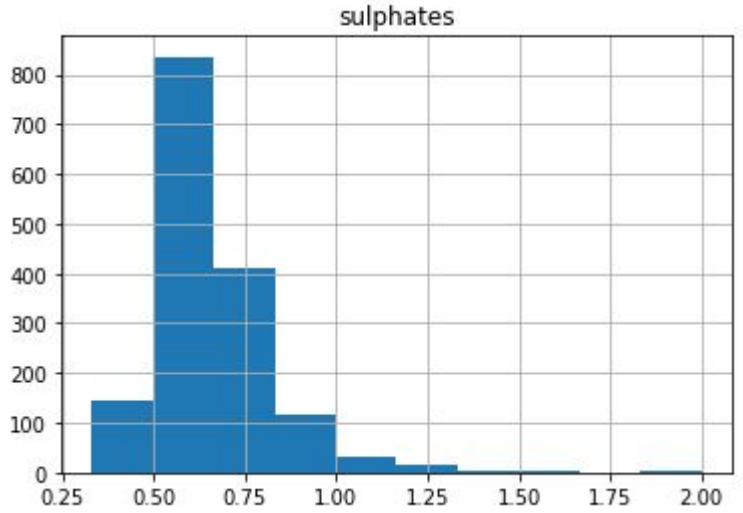


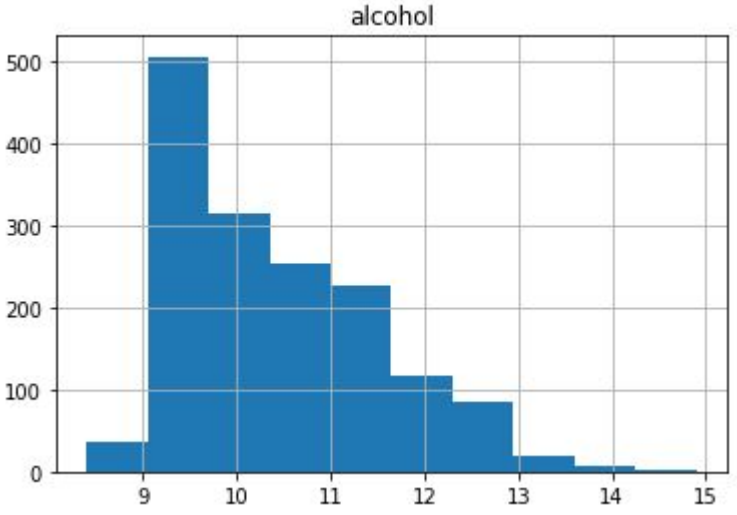


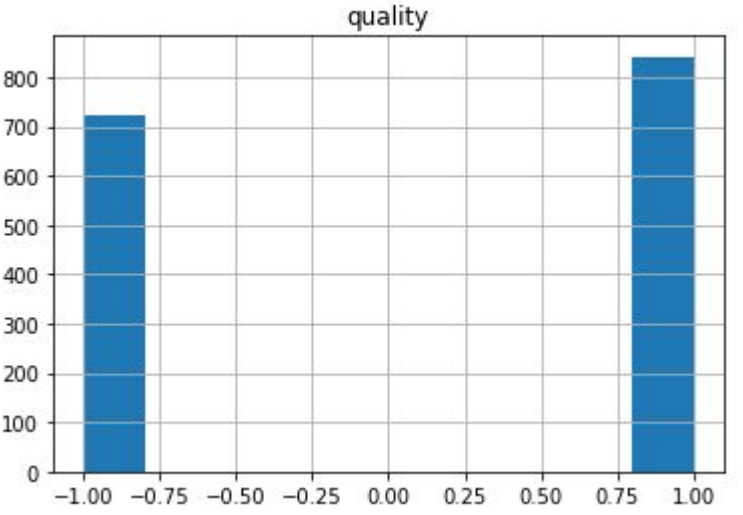

total sulfur dioxide
0
100
200
300
400
500
600
700
0
50
100
150
200
250
300


density
0
100
200
300
400
0.990
0.992
0.994
0.996
0.998
1.000
1.002
1.004

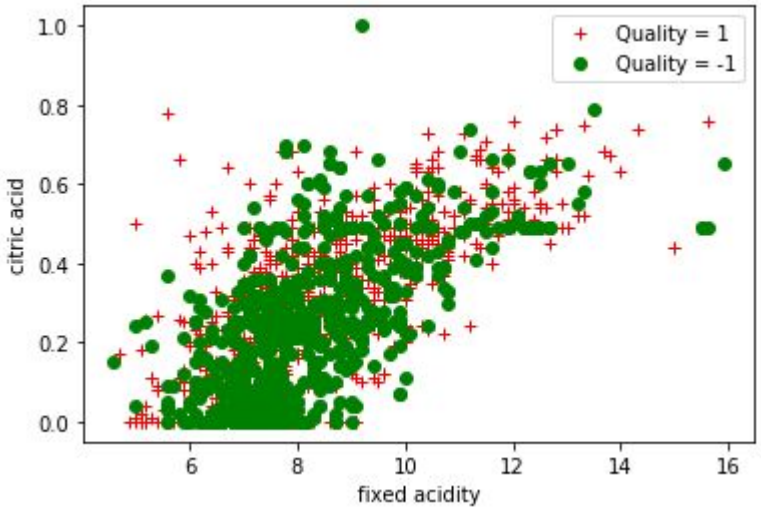

pH
0
100
200
300
400
500
2.8
3.0
3.2
3.4
3.6
3.8
4.0

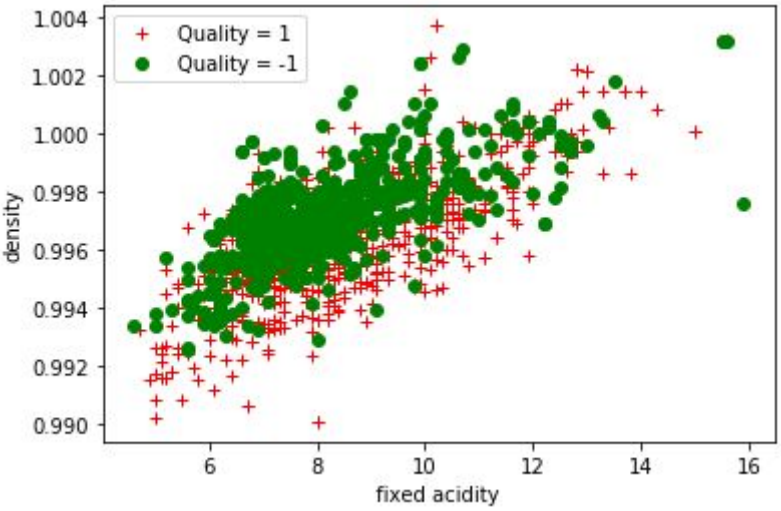

sulphates
0
100
200
300
400
500
600
700
800
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

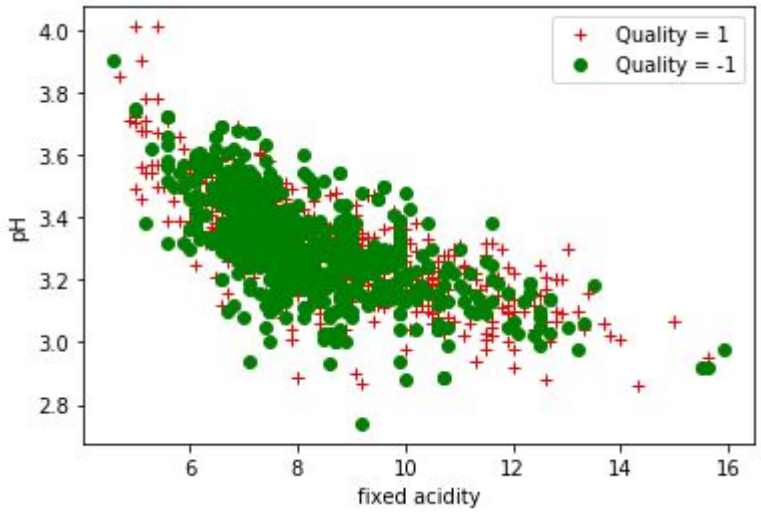

alcohol
0
100
200
300
400
500
9
10
11
12
13
14
15

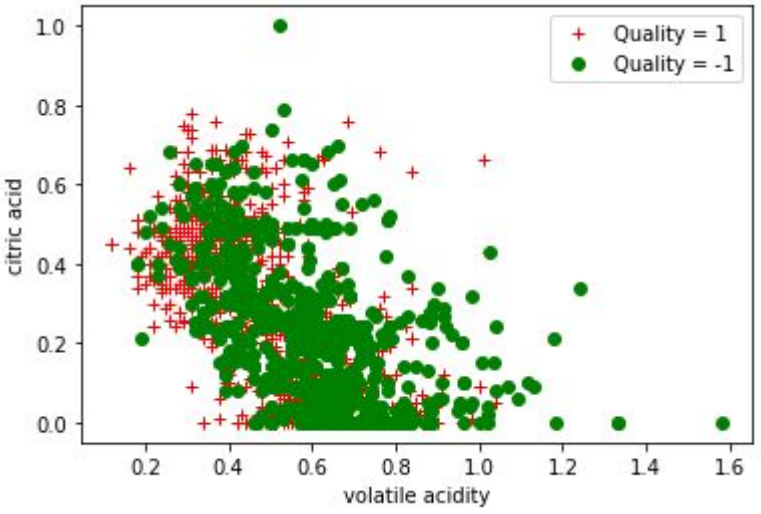

quality
0
100
200
300
400
500
600
700
800
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

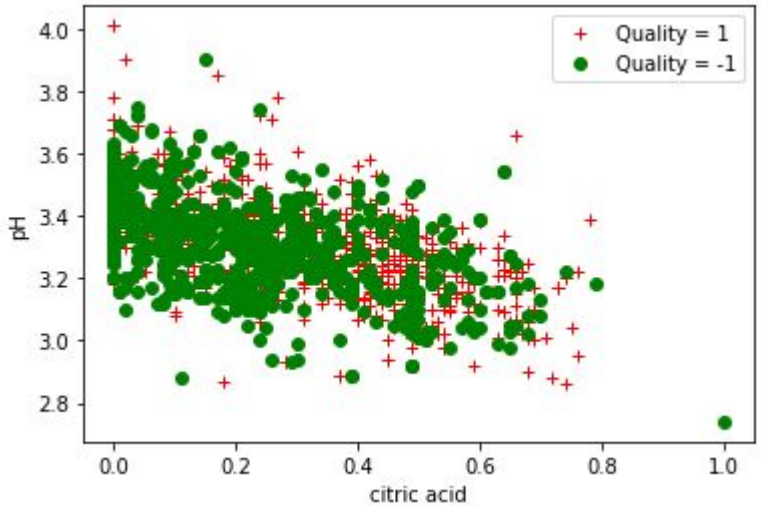
2/Graphiques des corrélations entre les différents attributs

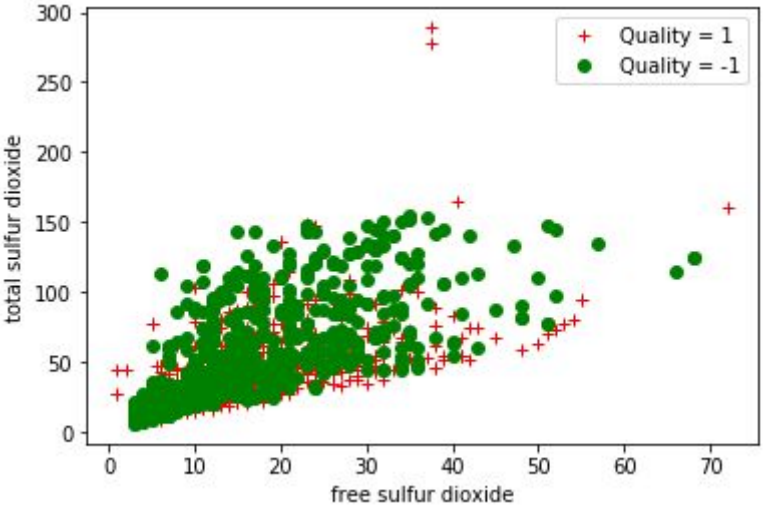


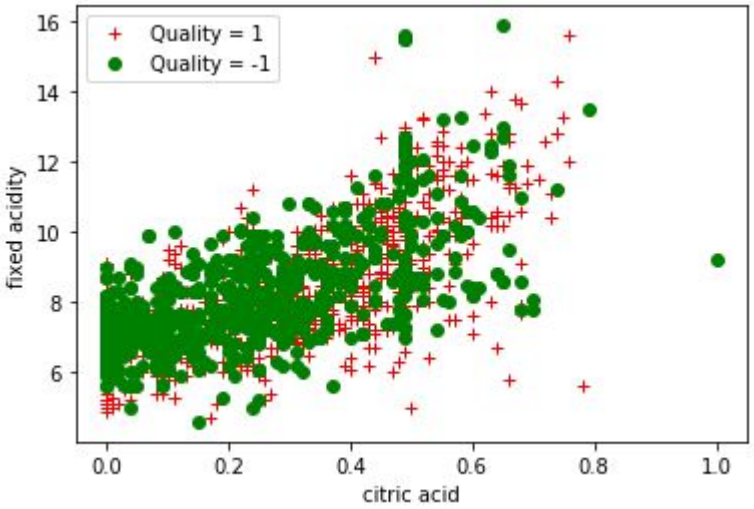


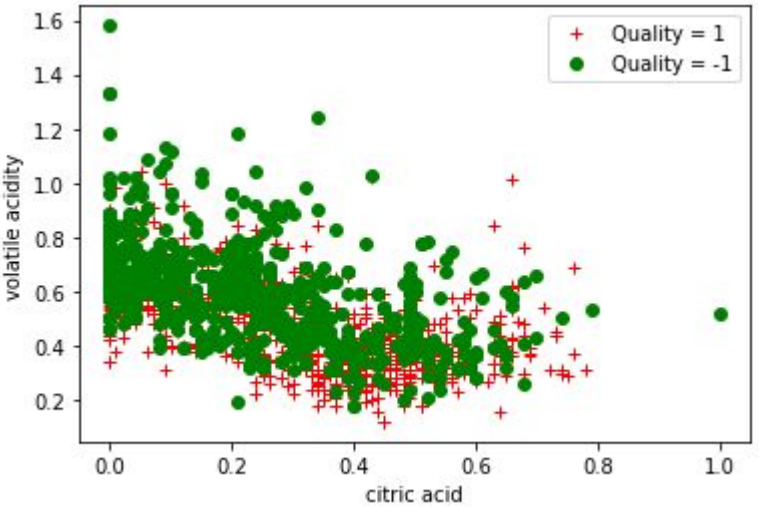


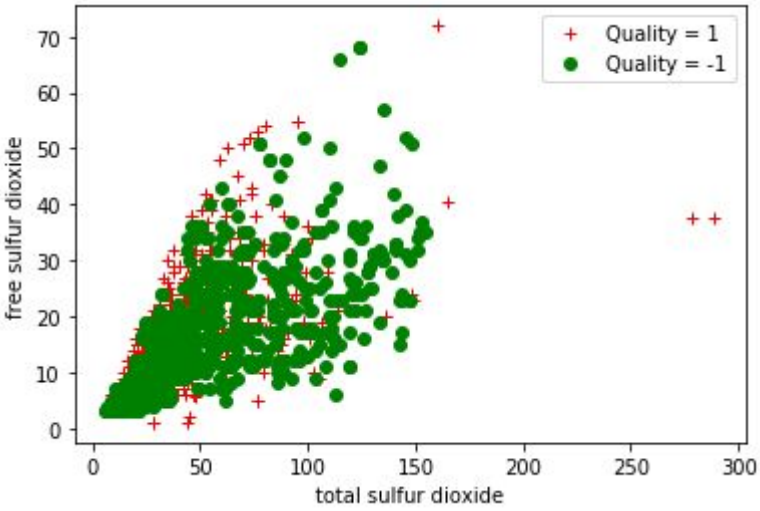


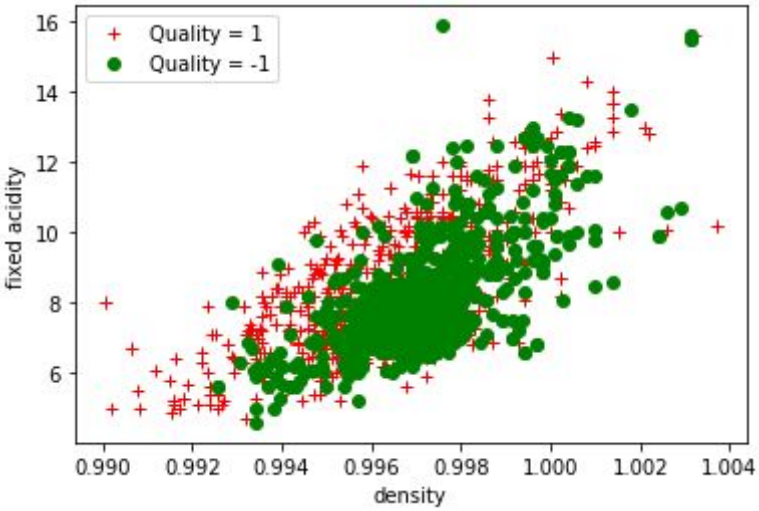


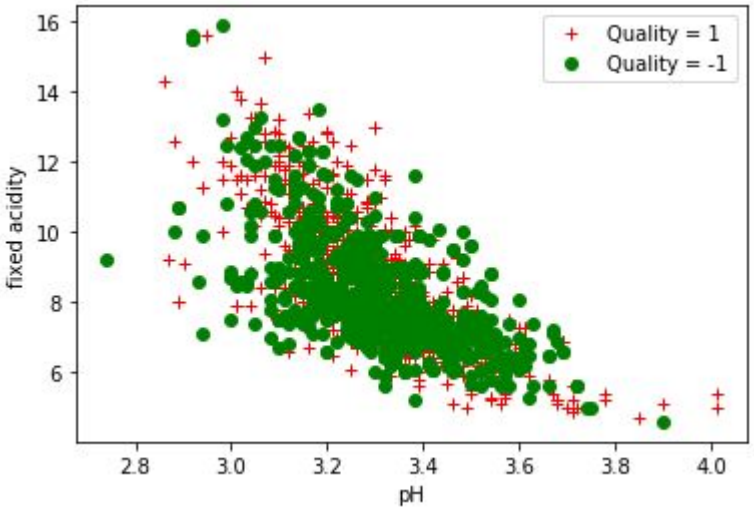


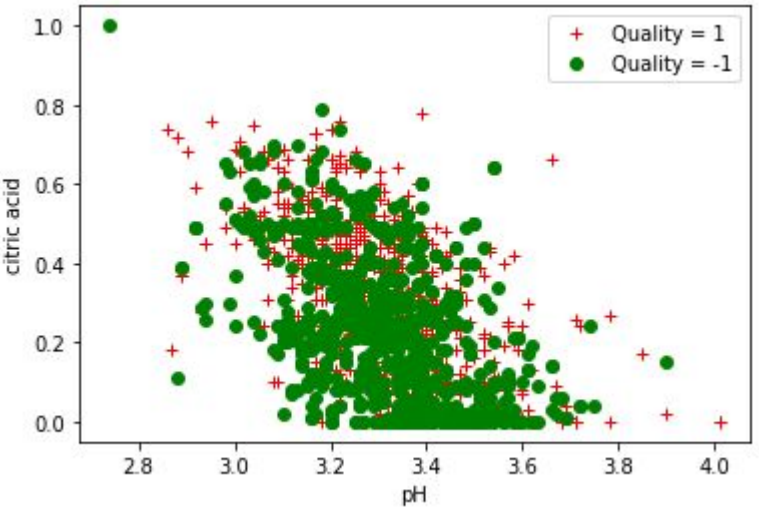

Quality = 1
Quality = -1
fixed acidity
citric acid


Quality = 1
Quality = -1
volatile acidity
citric acid


Quality = 1
Quality = -1
free sulfur dioxide
total sulfur dioxide


Quality = 1
Quality = -1
fixed acidity
density


Quality = 1
Quality = -1
fixed acidity
pH


Quality = 1
Quality = -1
citric acid
pH