Wine quality

Le Corre maxime

TABLE DES MATIÈRES

INTRODUCTION

1) Téléchargement des datasets
2) Exploration des données
3) Préparation des données
4) Construction d’un modèle
5) Tuning du modèle

Le Corre maxime

INTRODUCTION

Ce projet a pour but de prédire la qualité du vin à partir des features suivantes :

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

La valeur à prédire est entre 0 et 10.

Le projet a été réalisé entièrement en Python

Le Corre maxime

Téléchargement des datasets

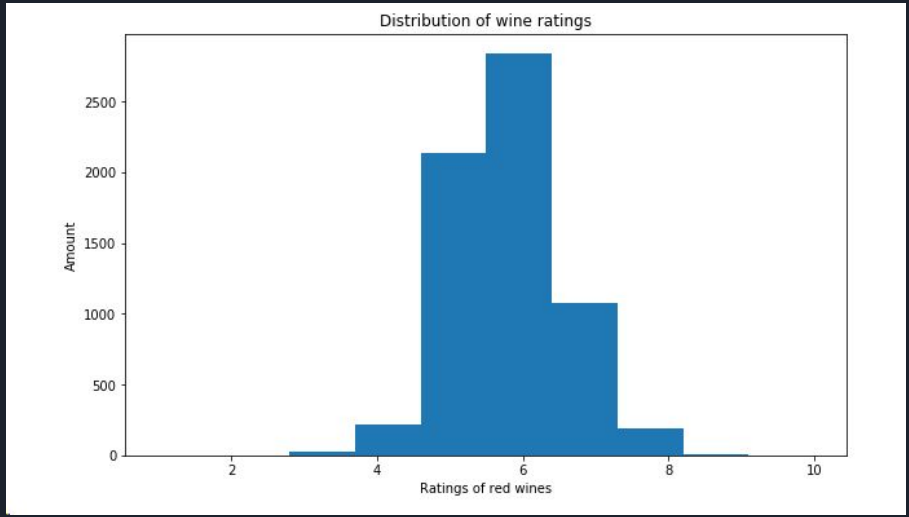
Le scraping des liens de téléchargement à été réalisé grâce à selenium.

Les fichiers récupérés sont :

- winequality.names, Une description du dataset
- winequality-red.csv, Un premier dataset contenant le vin rouge
- winequality-white.csv, Un second dataset contenant le vin blanc

Exploration des données

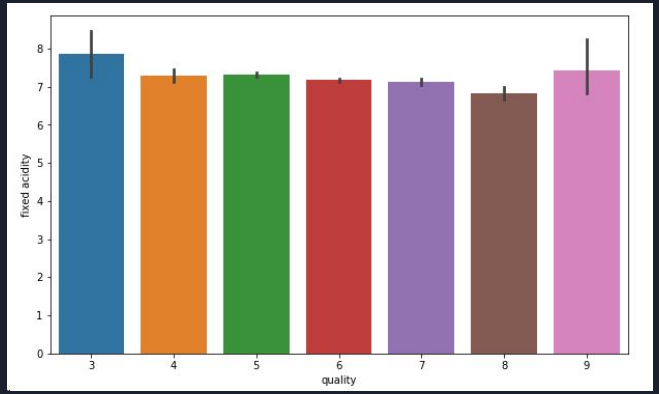
- J’ai fusionné les deux datasets
  - 6497 lignes
  - 13 colonnes

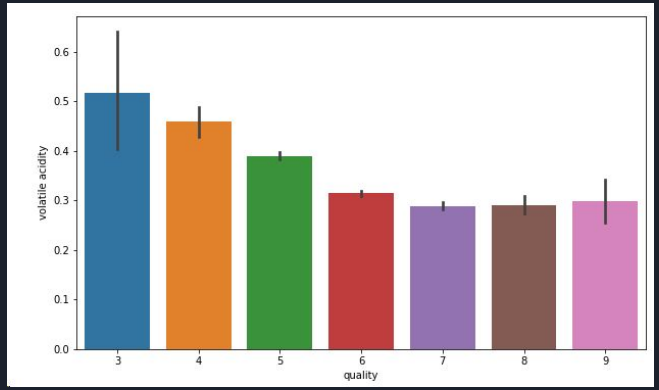


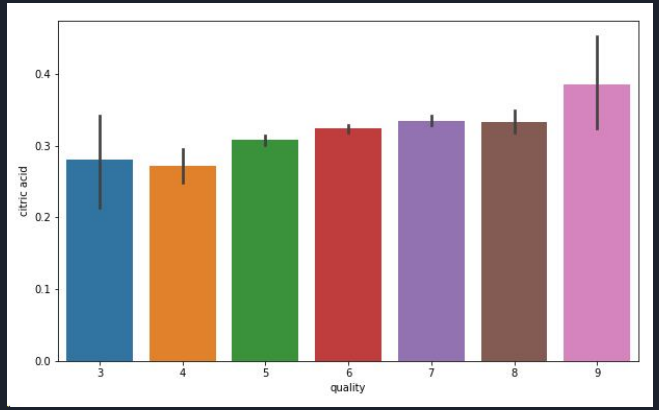
Le Corre maxime

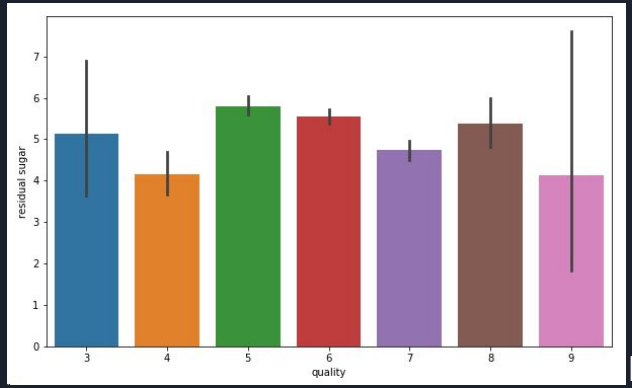
Exploration des données

Features par rapport à la qualité





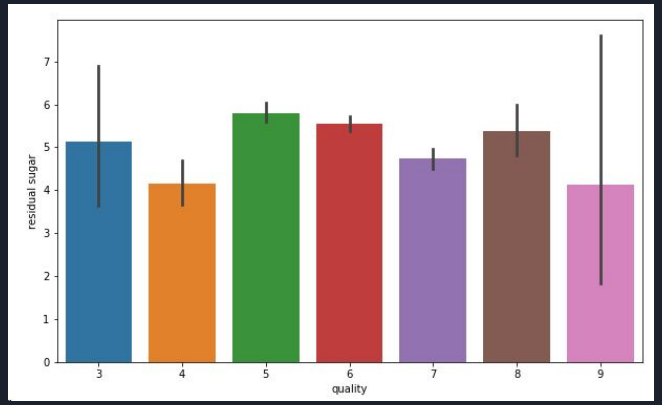


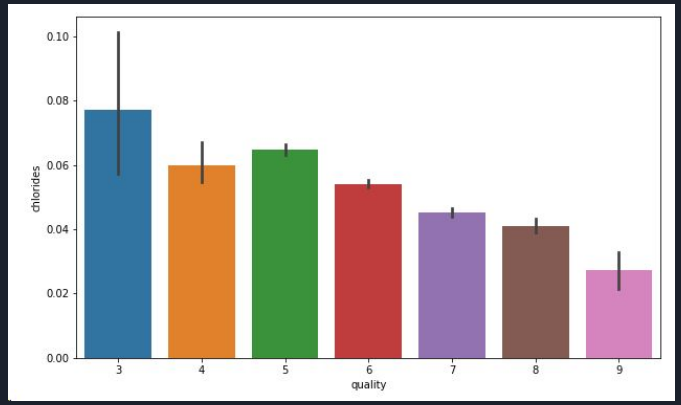


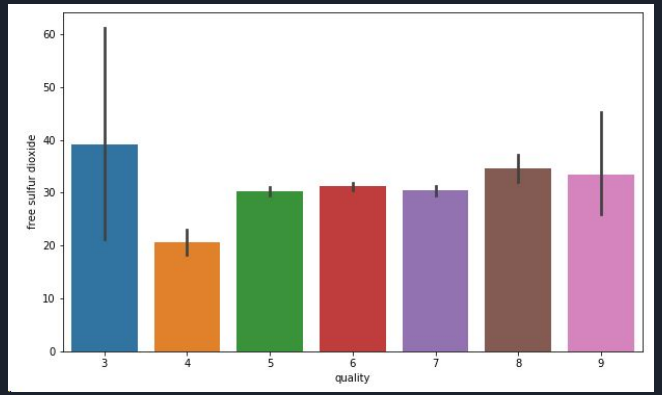
Le Corre maxime

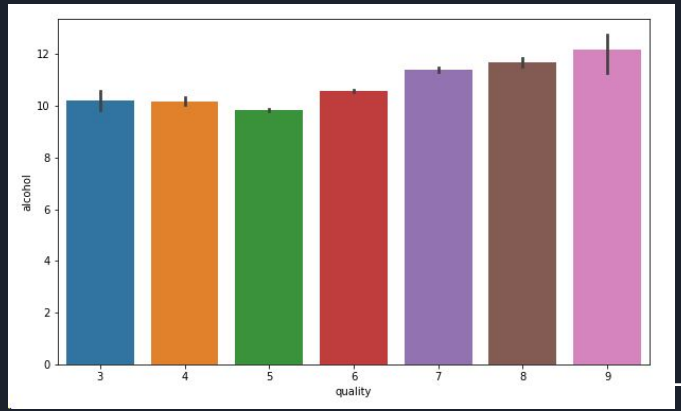
Exploration des données

Features par rapport à la qualité









Le Corre maxime

Préparation des données

- Les deux dataset vin rouge et vin blanc on été concaténé en ajoutant une colonne pour les différencier.

- Découpage des données en un jeu de train (⅔ des données) et un jeu de test (⅓ des données).

Le Corre maxime

Construction d’un modèle

- Choix de l’algorithme de classification : SVM

- En laissant les paramètre de l’algorithme par défault : `Accuracy: 0.28 (+/- 0.03)`

- Test sur une classification binaire, note >= 6 étant un bon vin et note <6 étant un mauvais vin. Résultat avec les paramètres par défaut de l'algorithme : `Accuracy: 0.96 (+/- 0.00)`

Tuning du modèle

GridSearch :

```
hyperparametres = {  'gamma' : [0.1, 0.5, 0.8, 1]          ,
                     "kernel" : ['linear', 'rbf', 'sigmoid',  ],
                     "C" : [10, 50]}
```

Pour la classification sur les notes de 0 à 10 :

```
0.33042154849148875 SVR(C=10, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma=1,
  kernel='rbf', max_iter=-1, shrinking=True, tol=0.001, verbose=False)
```

Gain de 5% sur l'accuracy

Pour la classification binaire (vin bon/mauvais) :

```
(0.9572231325550087,
 SVR(C=10, cache_size=200, coef0=0.0, degree=3, epsilon=0.1, gamma=0.1,
   kernel='sigmoid', max_iter=-1, shrinking=True, tol=0.001, verbose=False))
```

Aucun gain

Le Corre maxime