

***Data Science Potabilité de l’eau***

Groupe :

David Mamou

Quentin Joubert

Sébastien Salien

Matthieu Poirier

Tuteur : Samy Kerboua-Benlarbi

Session : 2020 – 2021

Sommaire :

1. Introduction
2. Quelle problématique globale avez-vous identifié/choisi ?
3. Quels problèmes avez-vous rencontré avec les données ? Quelles explorations en conséquence ?
4. Quel(s) modèle(s) avez-vous choisi ? Pourquoi ?
5. Analyse de vos résultats en fonction de la problématique et conclusion(s).
6. Améliorations proposées et le cas échéant, propositions pour poursuivre.

INTRODUCTION

Durant notre année de formation au siens du CFA INSTA, nous avons été amenés à réaliser un projet en python dans le cadre du cours de Data Science. Celui-ci à pour but de nous initier au monde du machine learning, de l’intelligence artificiel afin d’avoir une vision du métier de data scientiste. Pour cela il nous a été demandé de résoudre un problème, qui nous a été soumis ou créer par nos soins. Nous avons utilisé une base de données fournit par Kaggle pour répondre au mieux à la problématique posé. Nous allons donc vous expliquer nos méthodes utilisées ainsi que l’analyse de leurs données pour résoudre ce problème.

Problématique

Nous avons choisi la problématique suivante :

L’eau est elle potable et dans quelles conditions ?

Cette problématique est accompagnée d’une base de données dont les champs nous permettaient de répondre au mieux à cette problématique.

Grâce aux recherche établie par des chercheurs, nous pouvons établir une analyse des données afin de dire si l’eau échantillonner est potable ou non.

Problèmes rencontrés et conséquences

Les problèmes rencontrés sont les suivants :

* Le jeu de données n’était pas complet
* Nous avons été contraints par le temps
* L’adaptation au langage
* Le choix de la méthode afin d’analyser le jeu de données
* Le travail et la répartition sur un seul fichier.

Les résolutions des problèmes sont les suivantes :

* Des informations manquaient ce qui a eu pour conséquence d’erronée notre précisions et nos résultats.
* Nous avons donc décidé de travailler au mieux en répartissant à chacun les tâches par groupe de 2, afin d’optimiser au mieux le code.
* Nous avons tous déjà travailler avec du python. Cependant nous n’avons pas les connaissances pour travailler au mieux avec les bibliothèques fournit pour répondre à la problématique.
* Des données manquaient nous avons choisi de les remplacer par la moyenne des colonnes concerné afin de limiter les données biaisées et ainsi réduire une marge d’erreurs liés à un champs nulle ou arbitrairement remplie.
* Nous avons donc dû travailler sur plusieurs branches et utilisé les merges de celles-ci, ce qui n’est pas habituelle pour tout le monde.

Les modèles

Nous avons choisi de travailler avec plusieurs modèles afin d’optimiser nos résolutions de problème :

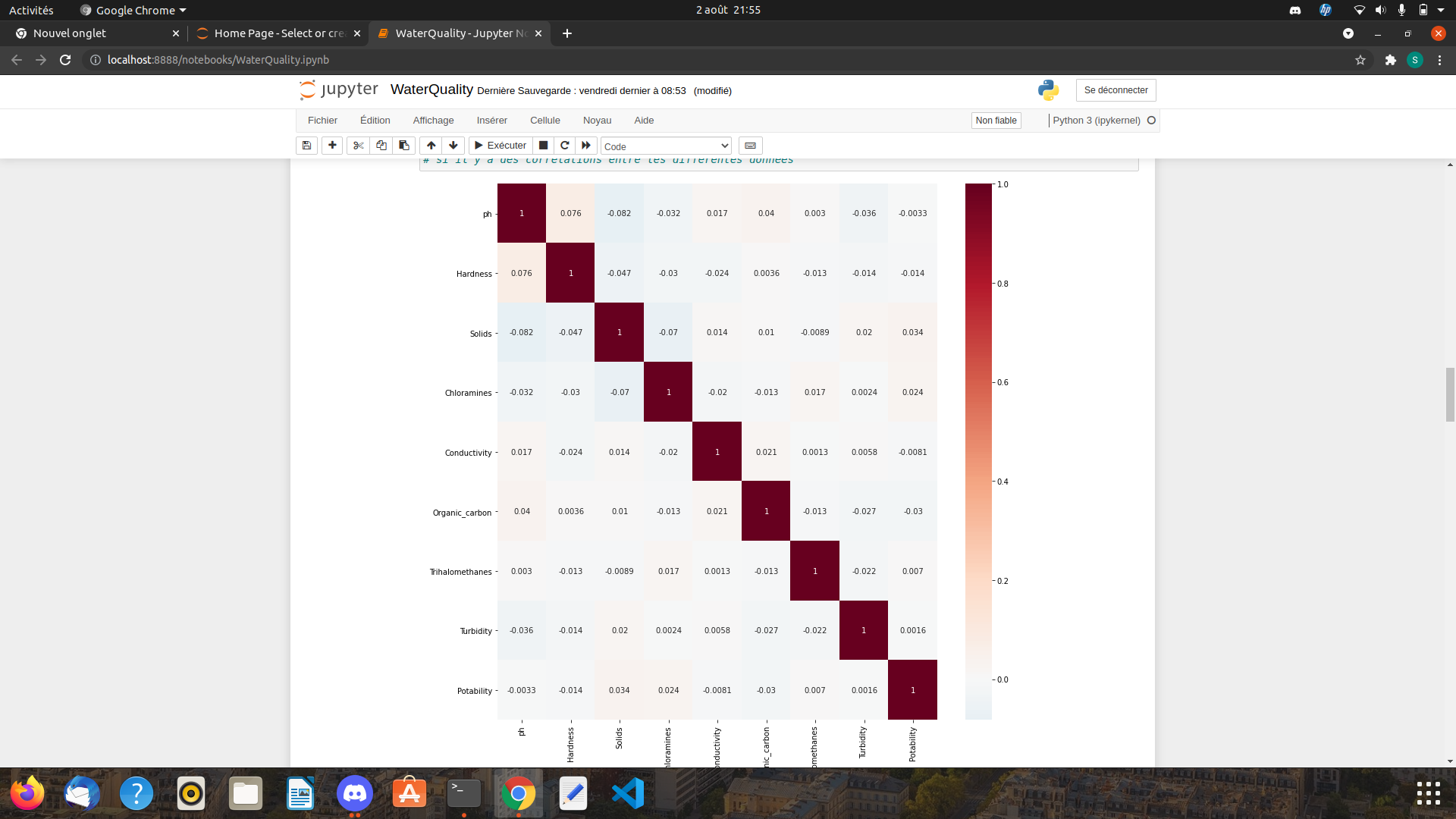
* La régression logistique
* L'arbre de décision
* La random forest

Voici le lien du site que l’on a utilisé pour retranscrire les définitions de chaque modèle utilisé : <https://dataanalyticspost.com/>

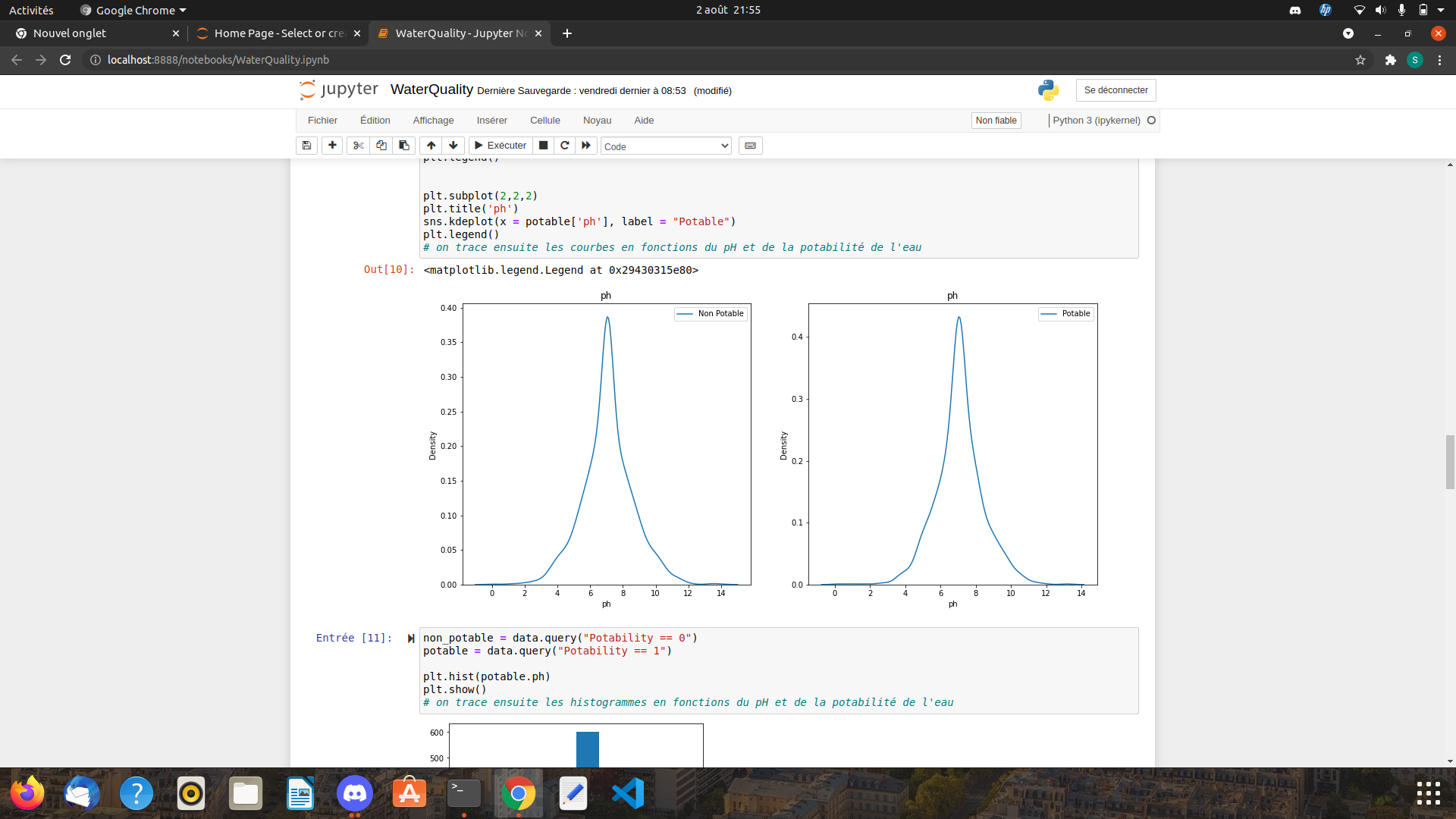
La régression logistique est un modèle statistique permettant d'étudier les relations entre un ensemble de variables qualitatives

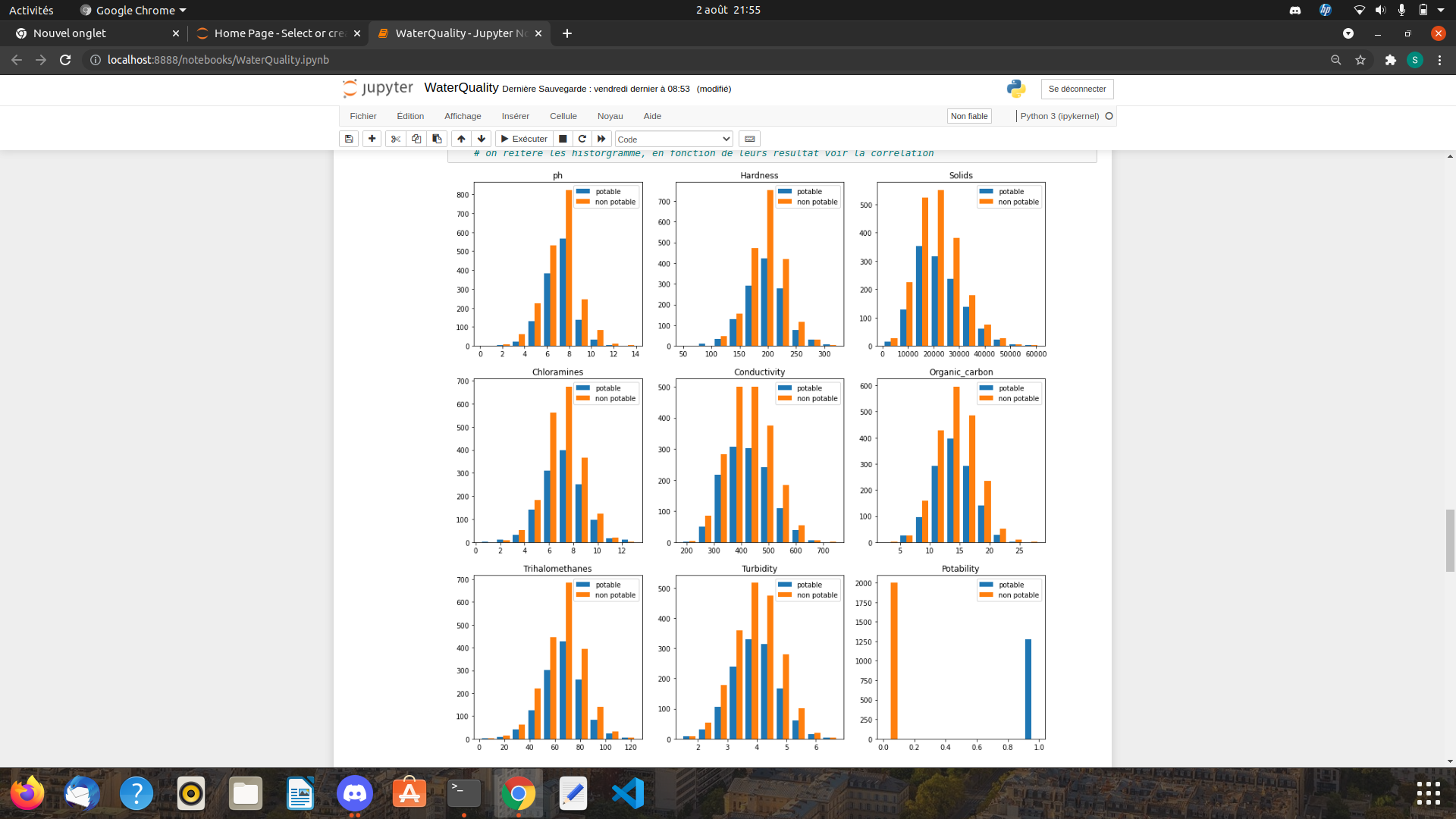
L’arbre de décision d'aide à la décision ou d'exploration de données permet de représenter un ensemble de choix sous la forme graphique d'un arbre. C'est une des méthodes d'apprentissage supervisé les plus populaires pour les problèmes de classification de données.

L’algorithme des « forêts aléatoires » (ou Random Forest parfois aussi traduit par forêt d’arbres décisionnels) est un algorithme de classification qui réduit la variance des prévisions d’un arbre de décision seul, améliorant ainsi leurs performances. Pour cela, il combine de nombreux arbres de décisions dans une approche de type bagging.

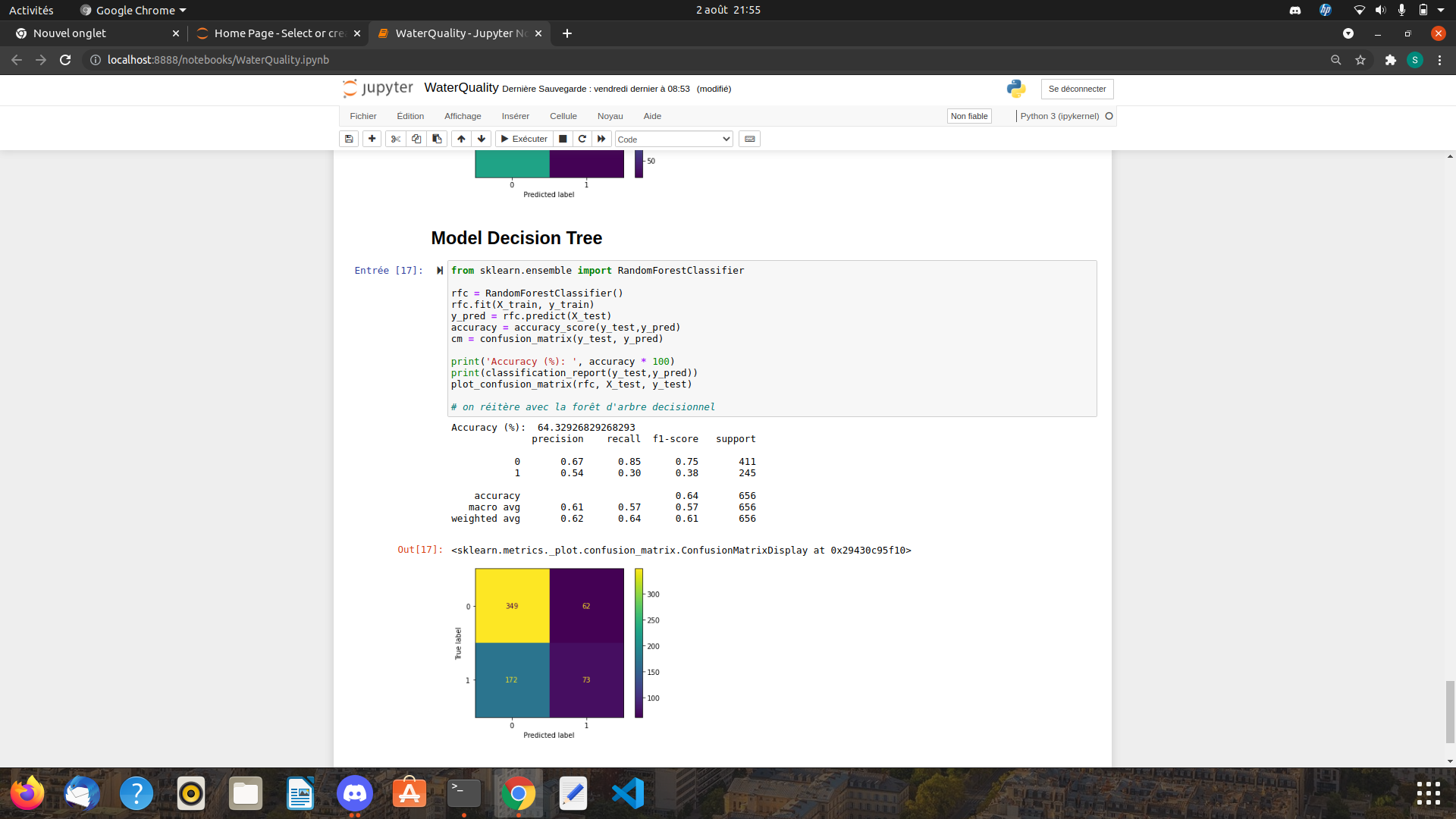


Nous avons aussi utilisé un tableau de type map afin de voir une potentielle corrélation entre les données.





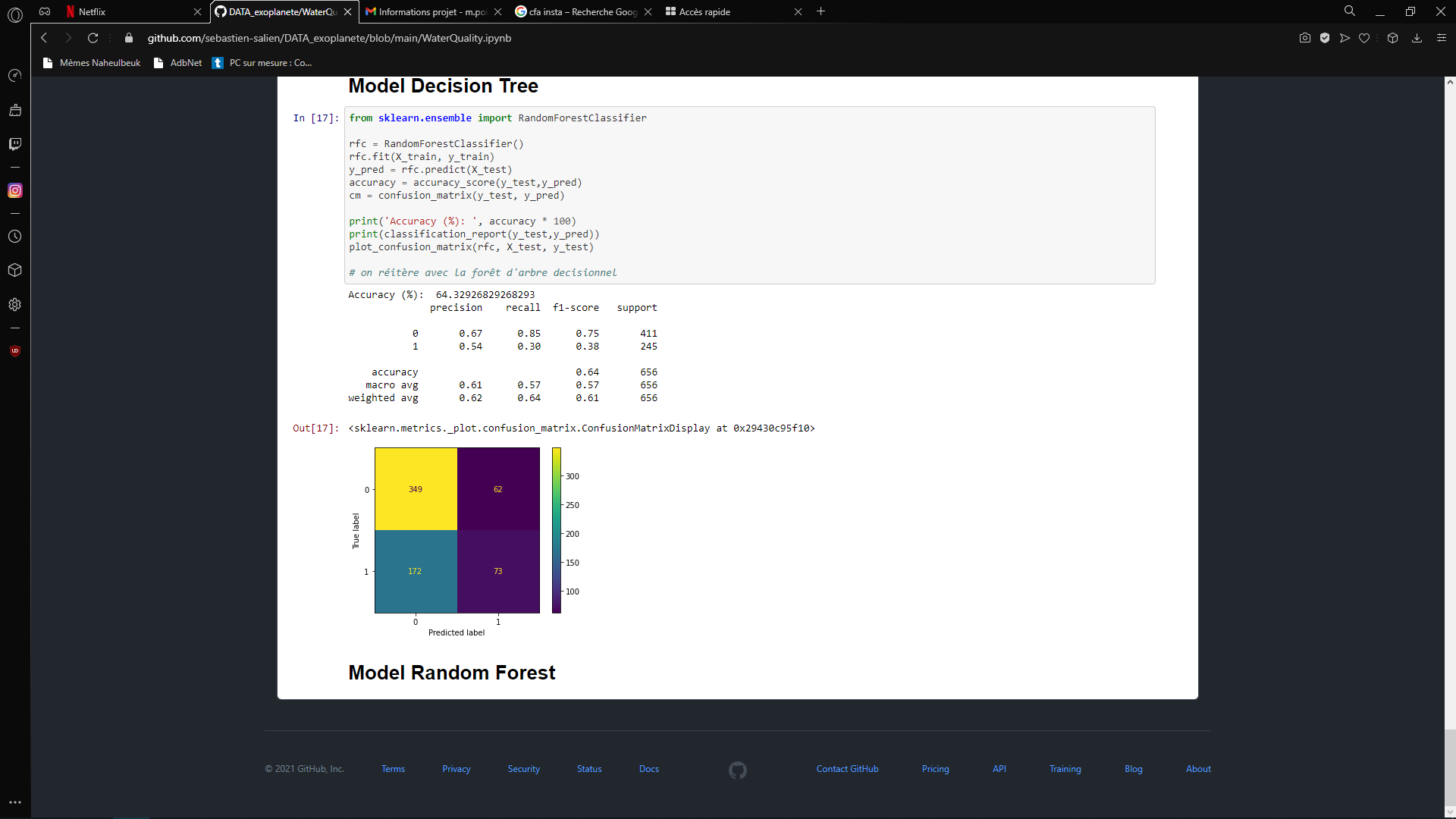
Nous avons choisi d’illustrer cela par différentes courbes ainsi que des histogrammes.



Pour le modèle de décision nous avons choisi d’ajouter les métriques liées à la matrice.

Analyse des données

Nous avons donc une précision qui au mieux atteint les 50 % et ce grâce au modèle de décision.



1. Critique

Cela n’est pas suffisant car afin d’avoir un meilleur résultat nous devons avoir au moins 80% de précision de nos données. Cela peut être dut à nos données qui sont incomplètes ou que nos calculs ne soient pas suffisamment précis.

Améliorations possibles du projet

A ce jour nous ne possédons pas toutes les données nécessaires par les chercheurs (données de compositions chimiques) ce qui fait que cela nous donne à ce jour une large marge d’erreurs. Avec ces nouvelles données nous pourrions avoir des résultats plus précis.