Python for Data Analysis

Statlog (Landsat Satellite) Data Set

Ibrahim-Ba DIA

Contexte du dataset

Le dataset contient les valeurs multi-spectrales de carrés de dimension 3x3 pixels, issus d'images satellites.

Le pixel central de chacun de ces carrés est classifié selon le type de sol dans la zone où l'image a été prise.

L'objectif est de pouvoir prédire cette classification.

Data Set Characteristics:	Multivariate	**Number of Instances:**	6435	**Area:**	Physical
Attribute Characteristics:	Integer	**Number of Attributes:**	36	**Date Donated**	1993-02-13
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	109984

Différentes features du dataset

Le dataset est composé de 37 colonnes, dont :

- 36 colonnes représentant les 4 valeurs spectrales de chacun des 9 pixels du carré 3x3;

- 1 colonne représentant la classification du pixel central.

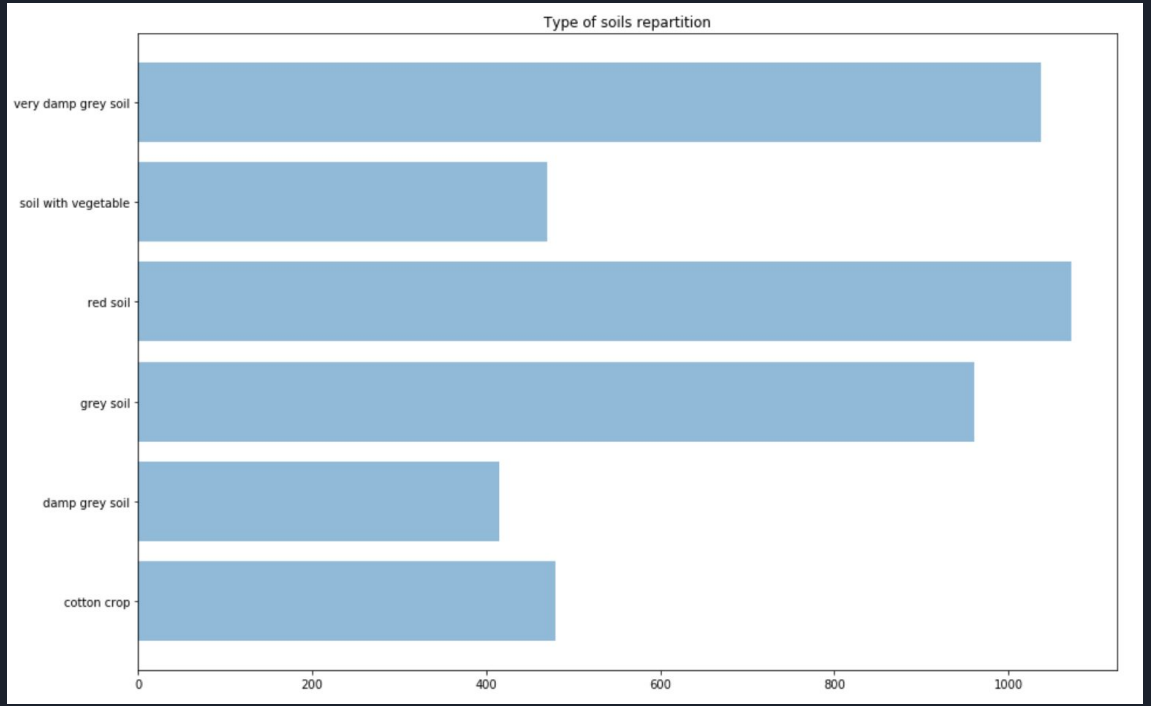
Classification des pixels centraux

Les différentes classifications possibles pour le pixel central de chacune des lignes du dataset sont les suivantes :

- “Red soil”
- “Cotton crop”
- “Grey soil”
- “Damp grey soil”
- “Soil with vegetable stubble”
- “Very damp grey soil”

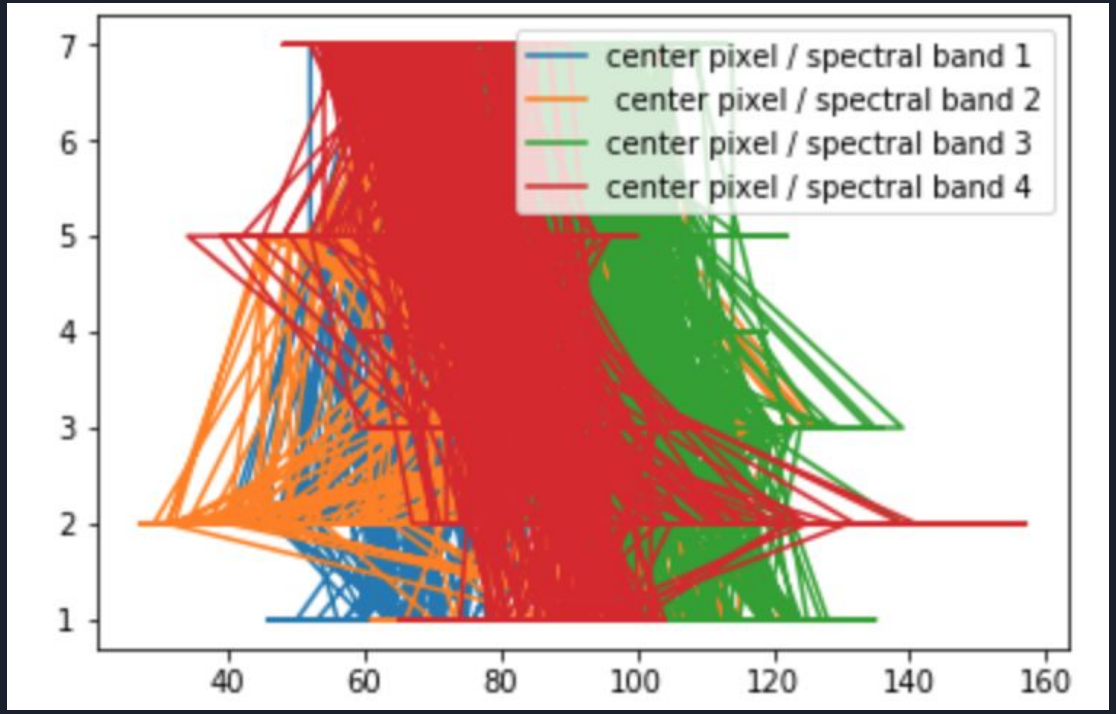
Visualisation des données

On voit que la majorité des pixels centraux proviennent d'une zone de type "Red soil".



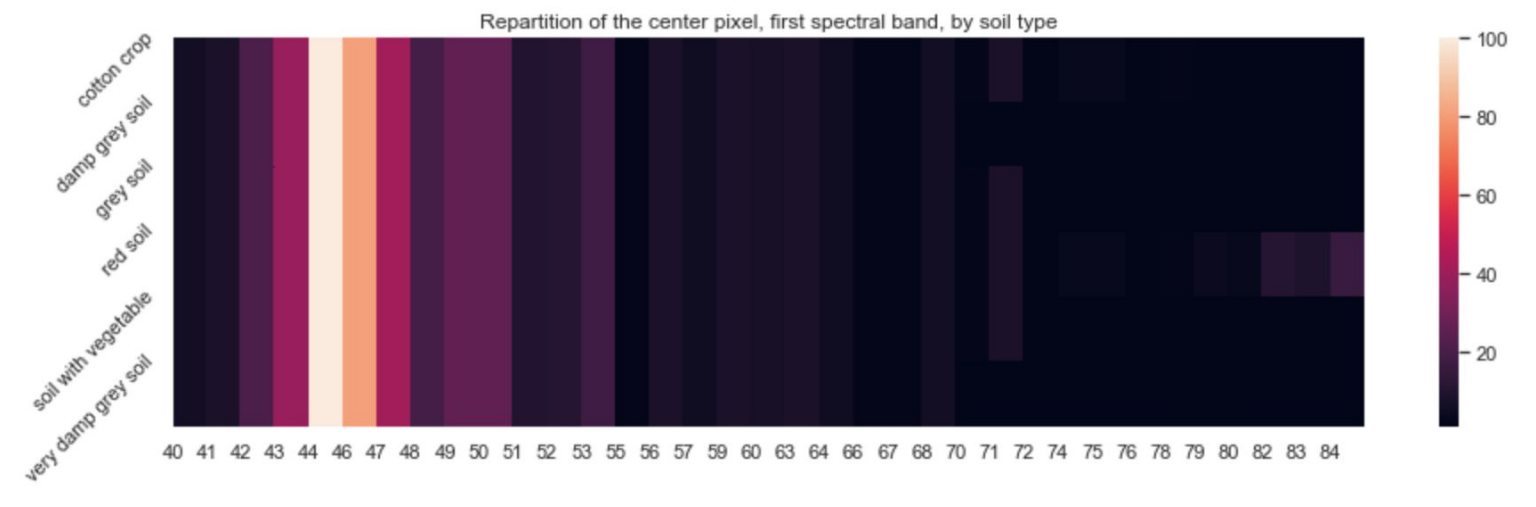
Visualisation des données

On peut voir la répartition de chacune des bandes spectrale des pixels centraux.



Visualisation des données

On note la dominance d'une valeur de couleur pour une bande spectrale donnée sur chacun des pixels centraux.



Modélisation des données

Je fais une classification des données pour la colonne “Class” , et j’utilise les 36 autres colonnes pour créer le modèle permettant de faire la prédiction pour cette colonne.

J’ai utilisé les algorithmes “Random Forest Regressor”, “SVR”, “Gaussian NB” et “Decision Tree Classifier”.

L’algorithme donnant le meilleur score est le Random Forest Regressor avec 0,89.

Conclusion du projet

Le dataset semble peu adapté aux différentes possibilités de visualisation.

Je suis spécialisé dans le domaine des réseaux, c'était donc mon premier projet (et mon premier module) de Data Science et j'ai trouvé cela intéressant, même si l'ensemble des actions à faire pour la modélisation était un peu dur à assimiler.