Projet Python For Data Analysis A5

Rayan AL QARAOUI
Elies ADJAL

Le Dataset

Dataset : Statlog (Landsat Satellite) Data Set

Le dataset nous provient du département « statistics and data modeling » de l'université de Strathclyde à Glasgow (Ecosse). Ce dataset est en fait une fraction d'un jeu de données tenu par un centre de la NASA basé en Australie.

Il est composé de 6435 fractions d'images satellites (appelées Neighbourhoods (NBH)) décomposées en 9 pixels (3x3) chacune.

Ici un découpage a été effectué par Alistair Sutherland :

- **training set 4435** NBH
- **Test set 2000** NBH

Le Dataset

Essayons de clarifier un peu plus les données que nous avons :

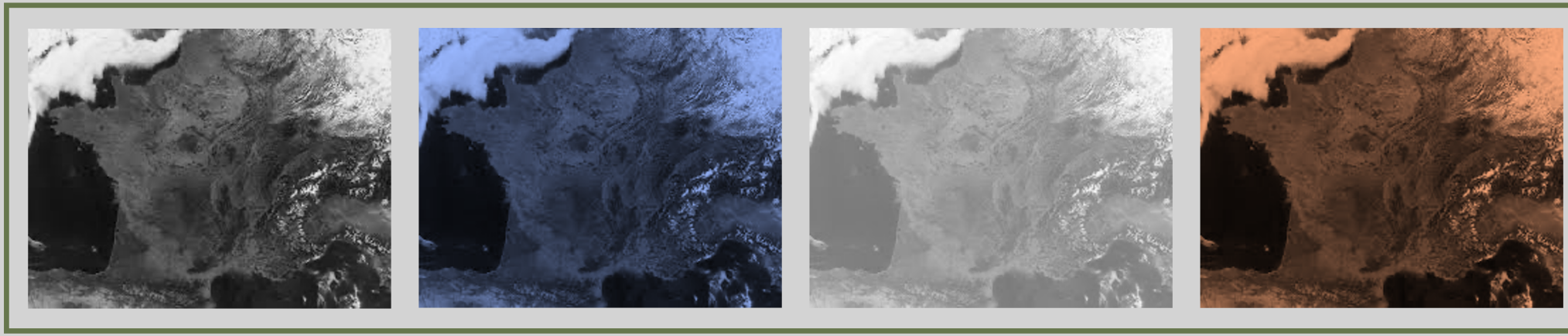
- Chaque **LIGNE** de notre dataset représente un Neighbourhood (ensemble de 9 pixels accolés comme ci-dessous)

Dans chaque pixel de cette image satellite est capturée une parcelle de 80m x 80m sur terre

Le Dataset

Les longueurs d'ondes avec lesquelles les récepteurs des satellites travaillent sont (pour la plupart) hors du domaine visible.

L'information de la couleur de l'image capturée est décomposée en différents canaux ou « bandes spectrales ». Une image est donc composée de quatre images numériques de la même scène respectivement dans 4 bandes spectrales différentes.



(Couleurs fictives)

Le Dataset

Les couleurs étant « erronées » par rapport à ce qui est visible par l'œil sur Terre, l'image va passer par un processus de superpositions des images obtenues (en ajoutant des filtres, mais cela ne sera pas important pour le projet).

De la nécessité d'avoir l'image dans les 4 bandes spectrales du satellite découle les **COLONNES** du dataset**.**

- Chaque **COLONNE** du dataset représente les 9 pixels de chaque neighbourhood, dans chacune des 4 bandes spectrales.

  Il y aura donc **9x4 = 36 COLONNES** dans le dataset

À ces dernières s'ajoutent une 37ème **COLONNE**

Le Dataset

La 37ème **COLONNE** est appelée « Classification Label ». Elle représente en fait la classe du pixel central.

Très concrètement, cette classe est le type de sol qui peut être observé sur la portion d'image contenu dans le **pixel central** de chaque **Neighbourhood**. Il y a 6 types de sols observables selon notre dataset :

1 : Terre rouge

2 : Culture du coton

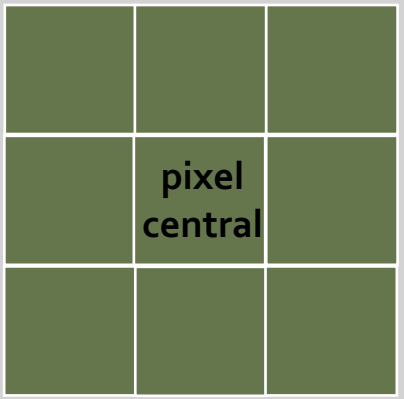
3 : Sol gris

4 : Sol gris humide

5 : Sol avec des chaumes de végétation

6 : Classe de mélange (tous les types présents)

7 : Sol gris très humide



Neighbourhood

L’Objectif

L’objectif de ce projet réside dans la 37ème colonne que nous venons d’introduire. En effet, c’est un travail de classification qui nous est suggéré ici.

Concrètement nous voulons, en nous basant sur les « valeurs » des pixel d’un même neighbourhood dans chacune des 4 bande spectrale, prédire quel type de sol est capturé dans pixel central. C’est à dire prédire le « classification Label ».

Visualisation des données

Nous allons utiliser les données contenues dans le fichier sat.trn pour visualiser les données du problème et entrainer des modèles.

Tout d'abord nous allons attribuer à chaque pixel (3x3) une lettre pour faciliter la lisibilité :

A	B	C
D	E	F
G	H	I

nous ajoutons un indice à chaque pour exprimer la bande spectrale du pixel en question (par exemple, le top-middle pixel dans la bande spectrale 3 est appelé : B3)

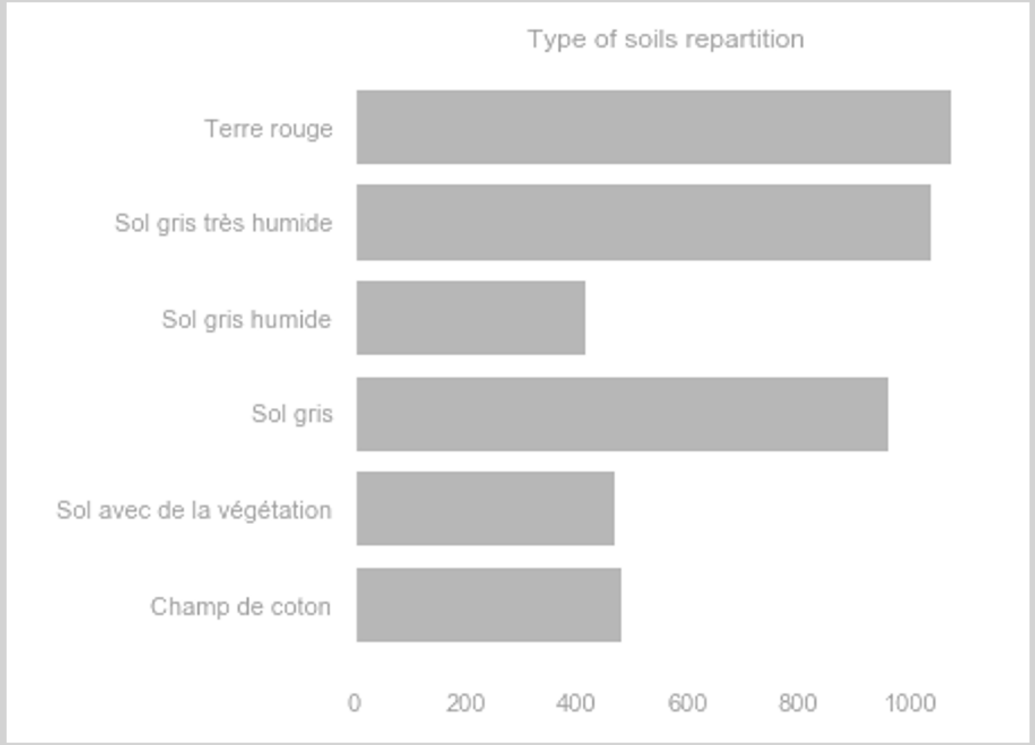
Visualisation des données

	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	...	G4	H1	H2	H3	H4	I1	I2	I3	I4	Class
0	92	115	120	94	84	102	106	79	84	102	...	104	88	121	128	100	84	107	113	87	3
1	84	102	106	79	84	102	102	83	80	102	...	100	84	107	113	87	84	99	104	79	3
2	84	102	102	83	80	102	102	79	84	94	...	87	84	99	104	79	84	99	104	79	3

Nos données se présentent maintenant ainsi. Il reste maintenant à ajouter une colonne qui traduit le numéro de la classe du pixel central en le type de sol qu'elle représente :

	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	...	H1	H2	H3	H4	I1	I2	I3	I4	Class	soil
0	92	115	120	94	84	102	106	79	84	102	...	88	121	128	100	84	107	113	87	3	grey soil
1	84	102	106	79	84	102	102	83	80	102	...	84	107	113	87	84	99	104	79	3	grey soil
2	84	102	102	83	80	102	102	79	84	94	...	84	99	104	79	84	99	104	79	3	grey soil

Visualisation des données



On remarque que les type de sol apparaissant le plus souvent sont le sol gris très humide et la Terre rouge. Le sol gris très humide représente très surement un bitume ou de la pierre d'une couleur foncée (soit due à l'humidité justement, ou tout simplement dû à la teinte qui lui a été donnée avant sa pose sur les routes). La terre rouge, elle, arrive très logiquement en tête.

Le sol gris humide est en dernière position. Nous pensons que cela est dû au fait que la frontière entre sol gris humide et les deux autres sols gris présents dans les données est assez fine. C'est une sorte d'entre-deux.

Modélisation des données

Nous allons dans cette partie essayer d'atteindre l'objectif de ce projet, à savoir essayer de prédire le type de sol apparaissant dans les pixels centraux de chaque Neighbourhood.

Pour cela, nous allons tester plusieurs algorithmes connus et voir lequel obtient le score le plus élevé.

Nous avons découpé notre training set en deux parties (70/30). Nous aurions pu aussi utiliser les données contenues dans le fichier sat.tst en tant que test set.

Nous compilerons nos résultat dans un tableau récapitulatif.

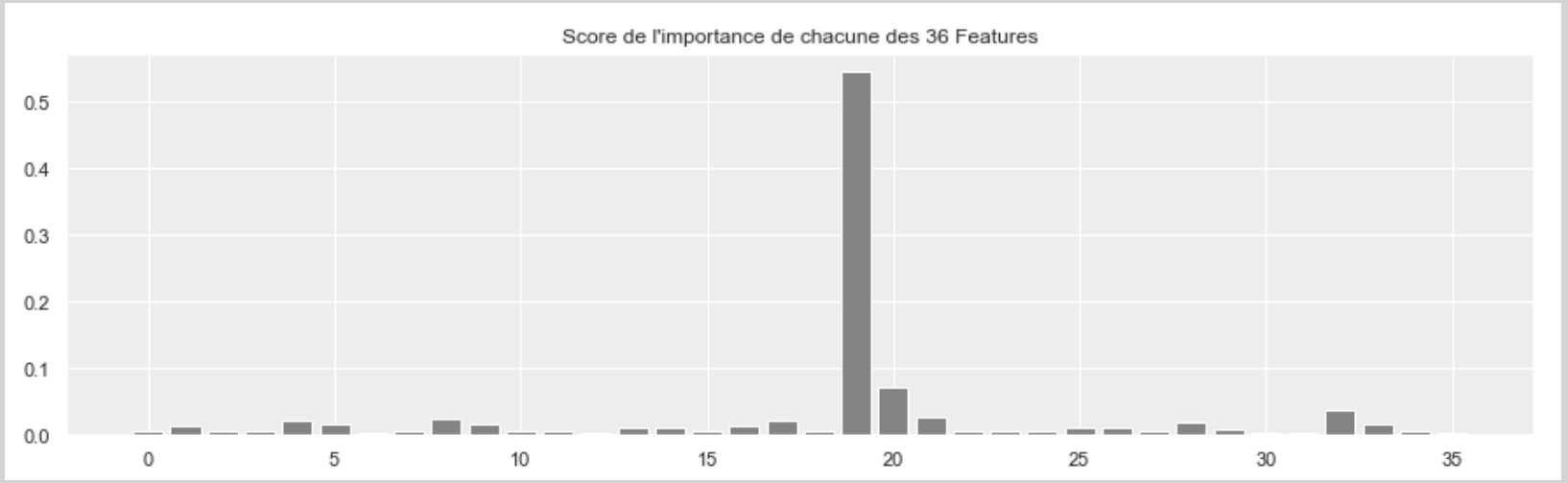
Modélisation des données

	RandomForest Regressor	Support Vector Regression (SVR)	Gaussian Naive Bayes	Decision Tree Classifier
SCORE	88.27%	69.20%	79.56%	83.77%

L'algorithme présentant le score le plus élevé est le Random Forest Regressor. Il est donc l'algorithme le plus concluant.
L'algorithme le moins efficace est le SVR.

Modélisation des données

Mettons finalement en surbrillance une donnée remarquable, si ce n'est la plus remarquable. Voyons laquelle des 36 features a la plus grande incidence sur le résultat final.



On peut voir que la Feature la plus importante dans ce jeu de donnée est la 19ème , qui représente la valeur du pixel central (comprise en 0 et 255) dans la bande spectrale 4 (variable appelée E4 dans le Dataset). Disons que c'est elle qui a le plus d'incidence sur le résultat de « Soil ».

Conclusion

Ce Dataset était assez complexe et nous a demandé un travail de recherche pour sa compréhension. Cependant il était intéressant à étudier. Faire parler ses données n'était pas forcément évident car nous ne savions pas instinctivement quelle analyse était vraiment pertinente en regard de notre problème. Nous avons quand même pu obtenir des résultats assez concluants.

Pour ce projet, il nous était demandé de transformer notre modèle en API Django, chose que nous n'avons malheureusement pas accomplie dans le temps imparti.