EY Open Science Data Challenge 2023

Finale France

Jack Leckert

EY Open Science Data Challenge 2023

Ma solution en 180 secondes

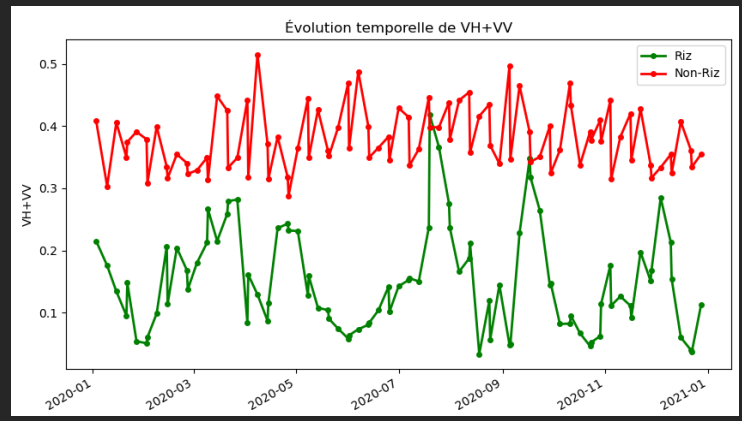
1 Recueil des données du satellite Sentinel 1

2 Moyennage des valeurs VH et VV sur une fenêtre de pixels 10x10

3	7	9	2
1	8	0	3
3	6	5	7
4	0	8	2

→ 4

3 Choix d'une combinaison unique de VH et VV permettant de distinguer le riz du non-riz



$VH + VV$?

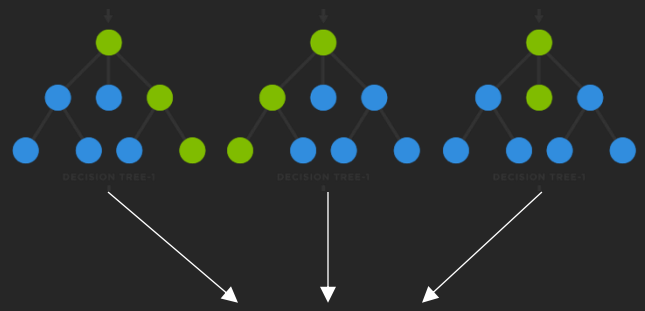
$\frac{VH - VV}{VH + VV}$?

$\frac{4 * VH}{VH + VV}$?

4 Extraction de caractéristiques des signaux

Moyenne	Écart-type	Médiane	…
0.37	0.08	0.39	…
0.16	0.15	0.17	…

5 Entrainement d'un algorithme d'IA, le Random Forest



Riz ou Non-Riz ?

6 Évaluation du modèle entrainé

Score: 0,99

Sommaire

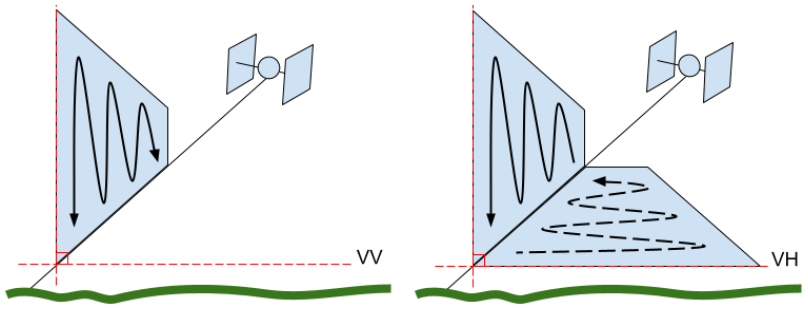
- Présentation du problème
- Méthodologie
- Pistes d’amélioration
- Conclusion: Business case

Présentation du problème

Objectif: prédire la présence de cultures de riz à 250 géolocalisations (latitude et longitude) dans la province d'An Giang au Vietnam en utilisant des données satellitaires.

Données: évolution temporelle de 2 bandes (signaux) radio à 5.4 GHz du satellite Sentinel-1, l'une étant polarisée horizontalement (VH) et l'autre verticalement (VV).

Au total, 600 données d'entrainement pour 250 données de prédiction.



Problème en science des données: classification de géolocalisations en 'Riz' (1) ou 'Non Riz' (0).

Méthodologie

Prétraitement des données → **Extraction de caractéristiques** → **Entrainement du modèle** → **Évaluation**

- **Choix d'un indice de végétation**
- Choix d'une fenêtre de pixel
- Sélection d'un intervalle de temps

- **Passage de données temporelles à des données caractéristiques**

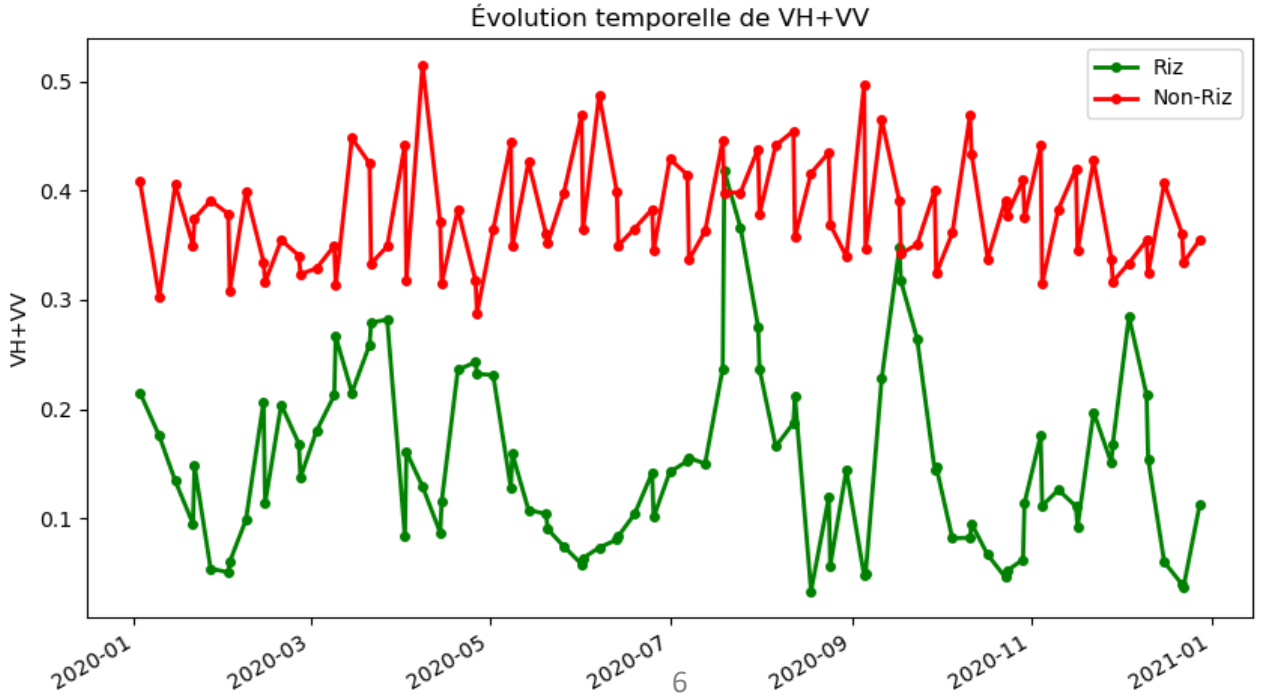
- **Sélection et entrainement d'un modèle d'apprentissage supervisé**

- **Évaluation du modèle selon une métrique de classification**

Prétraitement des données

- Indice de végétation: combiner les deux bandes temporelles pour obtenir une seule évolution temporelle, plus facile à classifier.

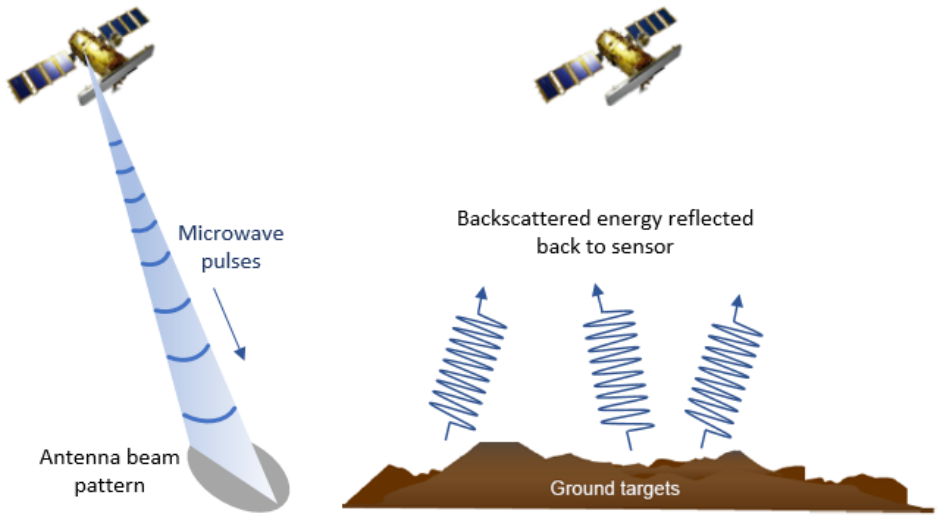
$VH + VV$? $VH - VV$? $\frac{VH}{VV}$? $SNI = \frac{VH - VV}{VH + VV}$? $\frac{VH + VV}{VH - VV}$? $RVI = \frac{4 * VH}{VH + VV}$?



Prétraitement des données

- Fenêtre de pixel: éviter de prendre en compte des variations inhérentes du signal à l'échelle du pixel en moyennant sur une zone plus large.

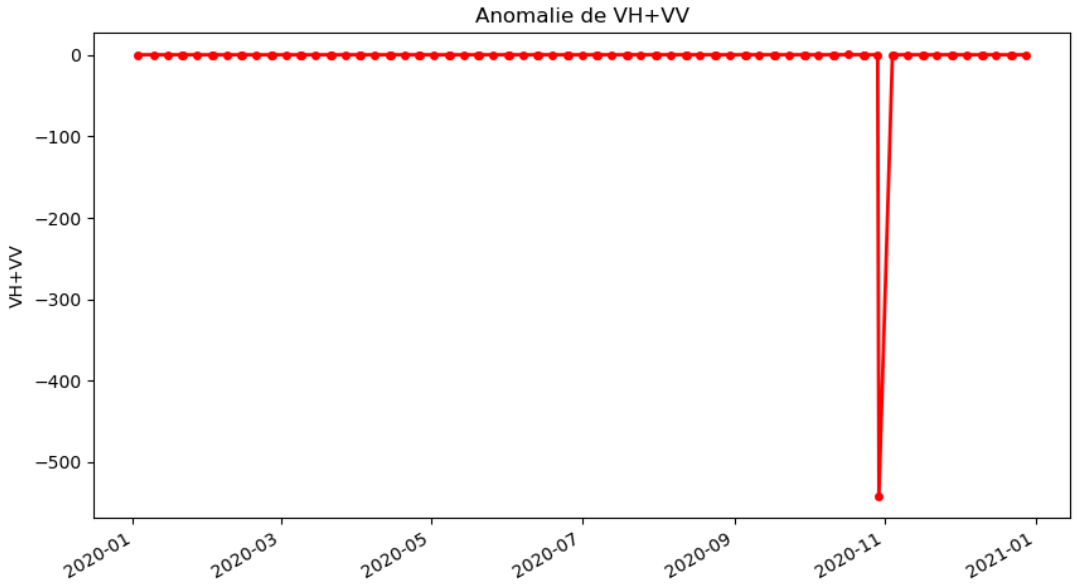
Pixels	Précision
3x3	0,96
5x5	0,97
7x7	0,98
8x8	0,98
9x9	0,99
10x10	0,99
11x11	0,98



Prétraitement des données

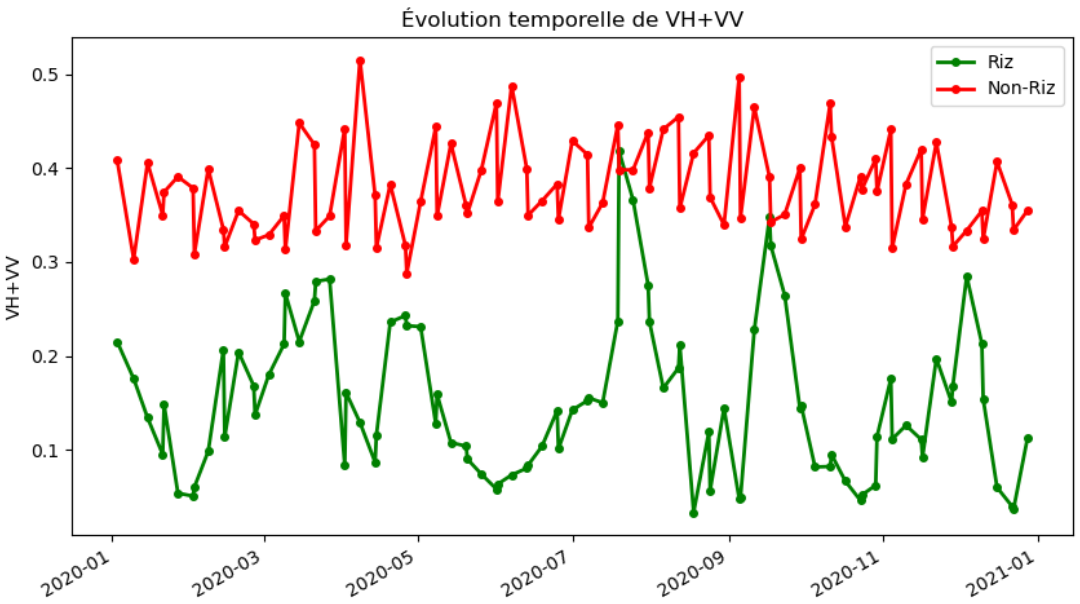
Essai sans succès:

- Suppression des données temporelles présentant des anomalies.



- Choix d’un intervalle de temps différent ou plus long.

Extraction de caractéristiques



Moyenne	Écart-type	Max	Médiane	...
0.37	0.08	0.52	0.39	
0.16	0.15	0.41	0.17	

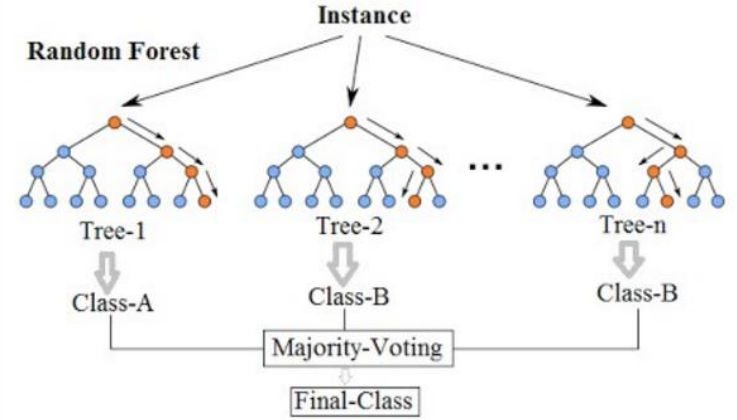
<u>Essai sans succès:</u>

- Automatiser l'extraction et la sélection de meilleures caractéristiques à l'aide du module open-source « tsfresh ».
- Choix des caractéristiques en fonction de la correlation entre elles.

Entrainement et évaluation d'un modèle

- Normalisation des données pour faciliter leur analyse par l'algorithme d'apprentissage.

- Modèle le plus performant: **Random Forest**
  - ✓ Modèle d'apprentissage supervisé pour classification
  - ✓ Limite le sur-ajustement (overfitting) du modèle
  - ✓ N'est pas sensible aux valeurs aberrantes



- Métrique: $F1 = 2 * \frac{précision \bullet rappel}{précision + rappel}$, $précision = \frac{VP}{VP+FP}$, $rappel = \frac{VP}{VP+FN}$

Score final: F1=0,99

- *Essais sans succès:* entrainement d'un petit réseau de neurones (TensorFlow) mais score légèrement inférieur à celui de Random Forest.

Pistes d’amélioration

- Augmenter le nombre de géolocalisations fournies (1000 par exemple).
- Utiliser d’autres indices / combinaisons d’indice VV et VH.
- Choisir de meilleures caractéristiques de signaux avec une méthode de sélection plus rigoureuse.

Conclusion: Business case

- À quoi pourrait ressembler ma solution dans la vie réelle ? À qui est-ce qu'on la vend et pour quel usage ?

Objectif du challenge niveau 1: identifier les cultures de riz à des géolocalisations précises.

Utilité / argument de vente: estimer la production locale puis globale de riz chaque année.

Demande potentielle de cette information:

- Les Nations Unies
- Des gouvernements nationaux
- Des entreprises d'agroalimentaires
- Des agriculteurs locaux