MACHINE LEARNING

CLASSIFICATION ARBRES / FLEURS

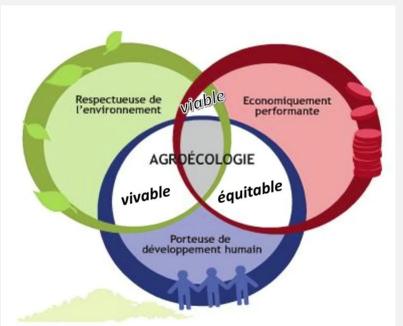


Gundelwein Marion Joigneau Marie Lajeunesse Armelle Ren Laurence



CONTEXT

<u>Problématique</u>: Comment différencier les arbres et fleurs pour qu'un drone soit capable de donner le bon engrais en Agroécologie?







SOMMAIRE

- Pré-traitement des images
 - Même format
 - ☐ Même nombre de pixels
 - Conversion en noirs et blancs
 - Zoom

2

- Méthodes de classification classiques
 - ☐ Arbres de décisions
 - Random Forest
 - ☐ KNN
 - ☐ Régression logistique

3

- Deep Learning
 - ☐ Réseau de neurones (CNN)

PRÉ-TRAITEMENT DES IMAGES

- ☐ Même format
- ☐ Même nombre de pixels
- ☐ Conversion en noirs et blancs
- □ Zoom

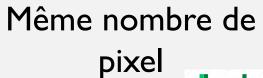


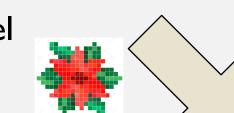
518 images au même format



Conversion en noir et blanc















Liste avec largeurs et longueurs de toutes les images -> Obtention de la médiane -> Images redimensionnées en carré de la médiane







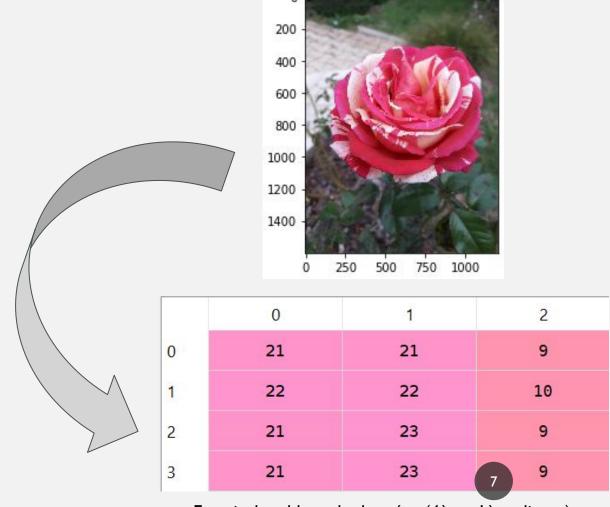
MÉTHODES DE CLASSIFICATION CLASSIQUES

- ☐ Préparation du jeu de données
- □ K plus proches voisins
- ☐ Régression logistique
- ☐ Arbres de décisions
- ☐ Random Forest

PRÉPARATION JEUX DE DONNÉES

- Etape I : Conversion des images en un dataframe de valeurs
- Etape 2 : Données mélangées
- Etape 3 : Séparation du jeux de données en train (70%) et test (30%)





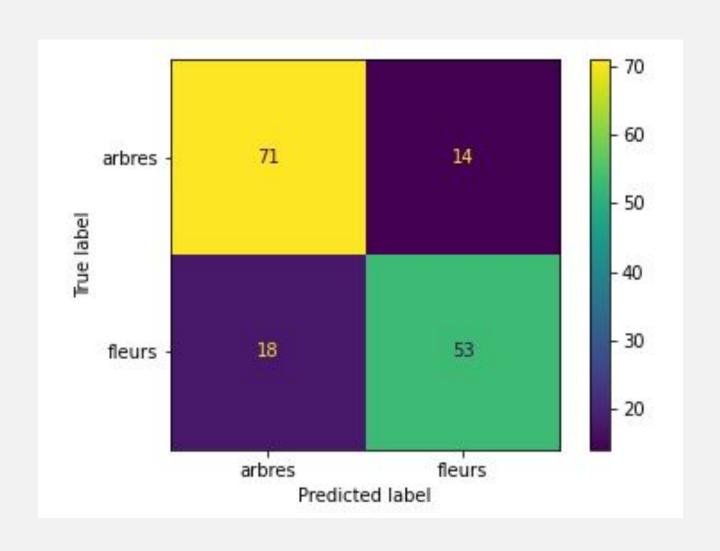
Extrait du tableau de données (4ème 1ères lignes)

COMPARAISON DES MODÈLES

Accuracy	Image originale	Noir et blanc	Zoom
knn	0.59615 (k=2)	0.51923 (k=2)	0.69231 (k=62)
Régression logistique	0.60897	0.56410	0.57692
Arbres de décision	0.62821	0.60897	0.69872
Random forest	0.70513	0.71154	0.79487

☐ Meilleur modèle = Random Forest sur photos zoomées

MATRICE DE CONFUSION DU MEILLEUR MODÈLE



DEEP LEARNING

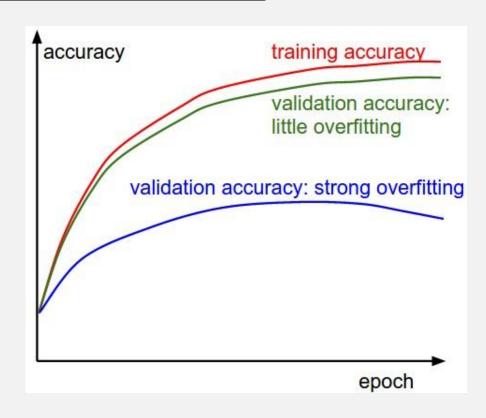
3

☐ Réseau de neurones

(Convolutionnal Neural Network)

RÉSEAUX DE NEURONES

- Le jeu de donnée est divisé en 3, 20% test, 80% train (dont 20% de données de validation utilisé dans la construction du modèle)
- Objectifs:
 - Minimiser la fonction de perte (arriver à 0 pour la jeu de données de validation est l'idéal).
 - Minimiser au maximum le sur-apprentissage
 - Avoir des courbes régulières
- ⇒ jouer sur les paramètres (nombre de couches de convolution, batch size, etc.)

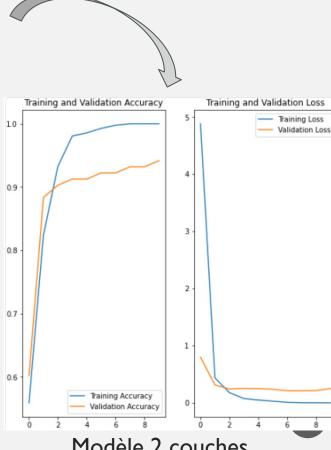


Recherche du meilleur modèle

- Modifier des paramètres :
 - nombre de couches de convolution,
 - nouvelles couches (drop),
 - taille du batch size,
 - fonction d'activation,
 - optimizer, etc.
- → efficacité du modèle peut être mesuré par l'accuracy sur les données test.



Modèle 3 couches

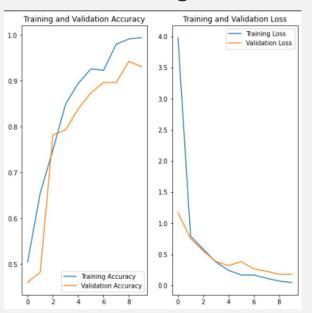


Modèle 2 couches

Modèle de CNN basé sur différents prétraitement

Meilleur modèle → modèle à 2 couches

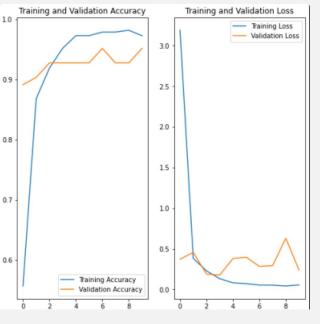
Modèle avec images colorées



Modèle avec images noir/blanc



Modèle avec image zoomés



Test sur de nouvelles images (couleurs)

- Jeu de données test issu des images initiales :
 - **Accuracy** ~ 0.8
- Nouveau jeu de données test (2422 images):
 - Images plus hétérogènes (plus d'espèces d'arbres et de fleurs différentes, toutes les saisons, etc.)
 - **Accuracy** = 0.63
- → spécifique à notre jeu de donnée et non généralisable





CONCLUSION

- Meilleures performances avec la méthode de deep learning, mais très spécifique à notre jeu de donnée et non généralisable
- Les méthodes sont meilleures avec les images zoomées, mais la différence avec les autres modèles est minime
- Utiliser un modèle deeplearning sur les photos zoomées pour le drone (avec des améliorations)
- Pistes d'amélioration :
 - Augmenter le jeu de données en tournant les images
 - Enrichir notre modèle grâce à des images plus variées (ex internet)