📘 Initiation à l'Explicabilité (XAI)

Parcours complet en 4 Notebooks Jupyter (NBJ1 → NBJ4)

🎯 Objectif général

Comprendre comment une IA analyse une image et apprendre à interpréter ses décisions à l'aide de quatre activités pratiques (NBJ). L'élève apprend à :

- Lire une carte d'attention (heatmap) - Détecter un biais
- Comparer plusieurs méthodes XAI - Expliquer et critiquer une décision d'IA

1️⃣ Installation de l'environnement

A) Installer Python & Jupyter

B) Installer les bibliothèques nécessaires

Dans un terminal :

```
pip install tensorflow torch torchvision tf-explain captum lime shap
scikit-image matplotlib numpy pillow
```

C) Organisation des fichiers

Dans un même dossier, mettre :

- Les 4 notebooks : `NBJ1`, `NBJ2`, `NBJ3`, `NBJ4`
- Le pack d'images fourni et ce guide

2️⃣ Présentation des 4 Notebooks

Notebook	Méthodes principales	Objectif pédagogique
NBJ1	Occlusion, Saliency, SmoothGrad	Comprendre comment une IA « regarde » une image

NBJ2	Grad-CAM (tf-explain)	Visualiser les zones importantes & identifier un biais
NBJ3	Integrated Gradients, Saliency, Grad-CAM (Captum)	Obtenir une explication fine et stable
NBJ4	LIME, SHAP	Expliquer un modèle sans le connaître & corriger un biais

---

3️⃣ Description détaillée des 4 TP

🔵 NBJ1 — Occlusion & Saliency (sans outil XAI)

- Méthodes : occlusion par patch, saliency brute, SmoothGrad
- Approche "code maison" pour comprendre les bases
- Premier contact avec la notion de “zone importante”

👉 Objectif : comprendre *comment une IA regarde une image*
👉 Résultat attendu : l’élève sait lire une carte d’attention simple

---

🟢 NBJ2 — Grad-CAM (tf-explain)

- Découverte des couches convolutionnelles
- Visualisation des zones sémantiques
- Exemple pédagogique du **biais neige → loup**
- Comparaison chien / loup dans la neige

👉 Objectif : comprendre les biais visuels
👉 Résultat attendu : l’élève sait dire si une IA regarde la bonne zone

---

🟣 NBJ3 — Integrated Gradients (Captum)

- Passage à PyTorch + Captum
- Méthode fiable : baseline → intégration des gradients
- Explication fine, précise, pixel-level
- Exemple : **biais médical** (la règle chirurgicale)

👉 Objectif : découvrir une explication rigoureuse
👉 Résultat attendu : l’élève comprend la notion de biais non visible

---

🟠 NBJ4 — LIME & SHAP (méthodes agnostiques)

- Explication d'un modèle sans connaître son architecture
- LIME : segmentation en superpixels
- SHAP : valeurs positives/négatives
- Correction simple d'un biais (recadrage, données)

👉 Objectif : analyser un modèle comme une « boîte noire »
👉 Résultat attendu : l'élève sait expliquer et corriger un biais

---

4️⃣ Méthodologie de travail (très important)

✔ Toujours afficher l'image originale

Avant d'observer une heatmap.

✔ Lire les cartes comme des indices

Une explication n'est **pas une vérité**, mais un signal.

✔ Croiser plusieurs méthodes

Aucune méthode XAI ne dit tout seule la vérité. Comparer toujours :

- Occlusion - Saliency - Grad-CAM - IG - LIME - SHAP

✔ Chercher les biais

Exemples classiques :

- Neige - Main - Arrière-plan - Texte / règle - Texture d'objet non pertinente

✔ Comparer prédiction ↔ explication

Une prédiction correcte peut être **incorrectement justifiée**.

5️⃣ Objectif final

Devenir capable de :

- Lire une carte d'explication - Détecter un biais - Interpréter la logique d'un modèle
- Critiquer les décisions d'une IA - Proposer une correction ou une amélioration

C’est la compétence centrale du programme :
💡 *penser comme un scientifique face à une IA*.

6️⃣ Contenu des 4 TPs (checklist pour l’élève)

✔ **NBJ1 :** Charger un modèle - Faire une prédiction - Calculer occlusion, saliency, smoothgrad - Répondre aux questions intégrées

✔ **NBJ2 :** Utiliser tf-explain - Générer un Grad-CAM - Comparer 2 images biaisées - Identifier un biais

✔ **NBJ3 :** Passer à PyTorch - Calculer Integrated Gradients - Comprendre baseline & intégration - Détecter le biais médical

✔ **NBJ4 :** Utiliser LIME (superpixels) - Utiliser SHAP (importance positive/négative) - Proposer une correction de biais - Faire une synthèse des 6 méthodes XAI

7️⃣ Bibliothèques utilisées

- **TensorFlow / Keras** pour NBJ1
- **tf-explain** pour NBJ2
- **PyTorch + Captum** pour NBJ3
- **LIME + SHAP** pour NBJ4

8️⃣ Pack d'images nécessaires

- `image_chien.jpg - loup_neige.jpg - chien_neige.jpg - melanome_regle.jpg - oiseau_branche.jpg`

9️⃣ Conseils pour réussir

- Lire toutes les questions dans les notebooks - Tester plusieurs images
- Varier les méthodes - Interpréter les résultats qualitativement
- Travailler en binôme (recommandé)

🔟 Conclusion

L’explicabilité est un outil essentiel pour comprendre :

- comment une IA prend une décision
- quand elle se trompe
- pourquoi elle se trompe
- comment on peut corriger ou améliorer un modèle

À la fin de ce parcours, tu auras manipulé les méthodes les plus utilisées au monde en XAI.
