**C du Propre**

## Détails des étapes

**Séance 1 - Cadrage**

Cette première séance a pour objectif de bien comprendre l’intérêt du POC, les parties prenantes et qui sera amené à l’utiliser en cas de déploiement dans une collectivité.

Pour le groupe concerné, il s’agit d’abord de structurer l’équipe pour déterminer les rôles de chacun, de s’imprégner du POC et de son contexte, ainsi que documenter la réalisation de la solution au cours des prochaines séances, jusqu’à la démonstration.

**Activités prévus**

*Les activités prévues indiquent ce que tu vas faire lors de la séance*.

* Comprendre et maîtriser le jeu de données, notamment en identifiant en amont les limites des données.
* Répartir les différents rôles au sein de l’équipe.
* Concevoir un diagramme Gantt/plan d’action pour chaque séance à partir du sommaire.

**Consignes**

*Les consignes te permettent de suivre pas-à-pas le déroulé de la séance*.

* Inspecter le jeu de données TACO sur le site officiel pour comprendre de quoi il est constitué et quels types d’images sont disponibles.
* Rédiger un document où l'équipe est présentée avec les différents rôles, ainsi que le diagramme Gantt.

**Ressources à disposition**

*Les ressources à ta disposition sont disponibles dans l'onglet « Données » de la page du projet. Pour cette séance, tu auras besoin des ressources suivantes.*

* **Jeu de données** : le dépôt GitHub contenant les données brutes est disponible ici, ou facilement manipulable avec Python sur Kaggle.
* **Site officiel des données** : le site officiel des données Taco Dataset.

**Livrables**

*Les livrables indiquent ce que tu dois être en mesure de fournir à la fin de la séance.*

* **Synthèse du cadrage** : document (type Word) qui permet de recenser les premières informations essentielles sur le jeu de données (de quoi est-il constitué, quelles sont les prises de vues), de documenter l’approche de réalisation de la solution à partir du sommaire et répartir les différents rôles au sein de l’équipe.

**Séance 2 - Étude de faisabilité**

Pour cette deuxième séance, tu dois inspecter en détail le jeu de données pour comprendre la structure de chaque image, des annotations présentes, tout en étant capable de déceler en amont les potentiels problèmes.

Il est attendu des graphes, statistiques ou toutes autres données permettant de mettre à plat les choix stratégiques qui seront utilisés tout au long du projet.

**Activités prévus**

*Les activités prévues indiquent ce que tu vas faire lors de la séance*.

* Calculer des statistiques sur les images (nombre d’images, d’annotations, nombre d’annotations par catégorie).
* Construire des graphiques de Data Visualization pour illustrer certaines statistiques.
* Produire un document de synthèse pour centraliser toutes les informations précédentes.

**Consignes**

*Les consignes te permettent de suivre pas-à-pas le déroulé de la séance*.

* Télécharger le Notebook « Analyse exploratoire des données » et le compléter afin de faire une étude descriptive sur les données.

**Ressources à disposition**

*Les ressources à ta disposition sont disponibles dans l'onglet « Données » de la page du projet. Pour cette séance, tu auras besoin des ressources suivantes.*

* **Notebook** : le Notebook « Analyse exploratoire des données » est à récupérer.
* **AWS SageMaker** : tu pourras utiliser AWS SageMaker pour cette séance.

**Livrables**

*Les livrables indiquent ce que tu dois être en mesure de fournir à la fin de la séance.*

* **Synthèse de l'étude descriptive** : document (type Word) qui recense l’ensemble des choix qui ont été décidés pendant l’étude avec des graphiques permettant de justifier les choix précédents et des exemples d’images avec des annotations dessinées dessus.

**Séance 3 - Préparation des données**

En accord avec les choix qui ont été décidés lors de la séance précédente, les étudiants vont maintenant devoir préparer les données en vue d'un entraînement par un modèle de Deep Learning.

En particulier, il est nécessaire de traiter les données de sorte à ce qu’elle puisse être compatible avec le modèle YoloV5 qui sera utilisé.

**Activités prévus**

*Les activités prévues indiquent ce que tu vas faire lors de la séance*.

* Comprendre le format des fichiers qui recense les annotations pour chaque image.
* Récupérer toutes les informations techniques des catégories (identifiants).
* Construire la base d’apprentissage en créant les fichiers textes des annotations pour chaque image.

**Consignes**

*Les consignes te permettent de suivre pas-à-pas le déroulé de la séance*.

* Télécharger le Notebook « Préparation des données » et le compléter afin de construire les fichiers d’annotations qui seront utilisés pour entraîner le modèle Yolo de la prochaine séance.
* Sauvegarder en local l’ensemble des fichiers d’annotations pour les réutiliser plus tard.

**Ressources à disposition**

*Les ressources à ta disposition sont disponibles dans l'onglet « Données » de la page du projet. Pour cette séance, tu auras besoin des ressources suivantes.*

* **Notebook** : le Notebook « Préparation des données » est à récupérer.
* **AWS SageMaker** : tu pourras utiliser AWS SageMaker pour cette séance.

**Livrables**

*Les livrables indiquent ce que tu dois être en mesure de fournir à la fin de la séance.*

* **Données d'entraînement** : données à utiliser pour entraîner le modèle Yolo par la suite.
* **Code source de la séance** : code source permettant de créer la base d’apprentissage avec les annotations à partir des données brutes (téléchargées depuis Kaggle).

**Séance 4 - Première modélisation**

Lorsque les données sont prêtes pour l’entraînement du modèle, une première phase de modélisation entre en jeu afin de pouvoir s’approprier correctement le modèle utilisé tout en s’assurant qu’il n’y a pas eu d’erreur lors de la préparation.

La première modélisation va permettre aux étudiants de se familiariser avec le modèle YoloV5, notamment dans le choix des hyperparamètres à apporter (taille de batch, nombre d’itérations) mais également fournir des métriques sur un modèle de base peu robuste qui pourra être utilisé par la suite en tant que modèle par défaut lors des tests de performances et les tests de déploiement.

**Activités prévus**

*Les activités prévues indiquent ce que tu vas faire lors de la séance*.

* Entraîner une première version du modèle sur les 60 classes du Taco Dataset.
* Récupérer les fichiers de poids du modèle entraîné.

**Consignes**

*Les consignes te permettent de suivre pas-à-pas le déroulé de la séance*.

* Télécharger le Notebook « Entraînement » et l’exécuter afin d’entraîner une première version du modèle Yolo.
* Sauvegarder en local les fichiers de poids qui pourront être réutilisés par la suite lors de l’inférence.

**Ressources à disposition**

*Les ressources à ta disposition sont disponibles dans l'onglet « Données » de la page du projet. Pour cette séance, tu auras besoin des ressources suivantes.*

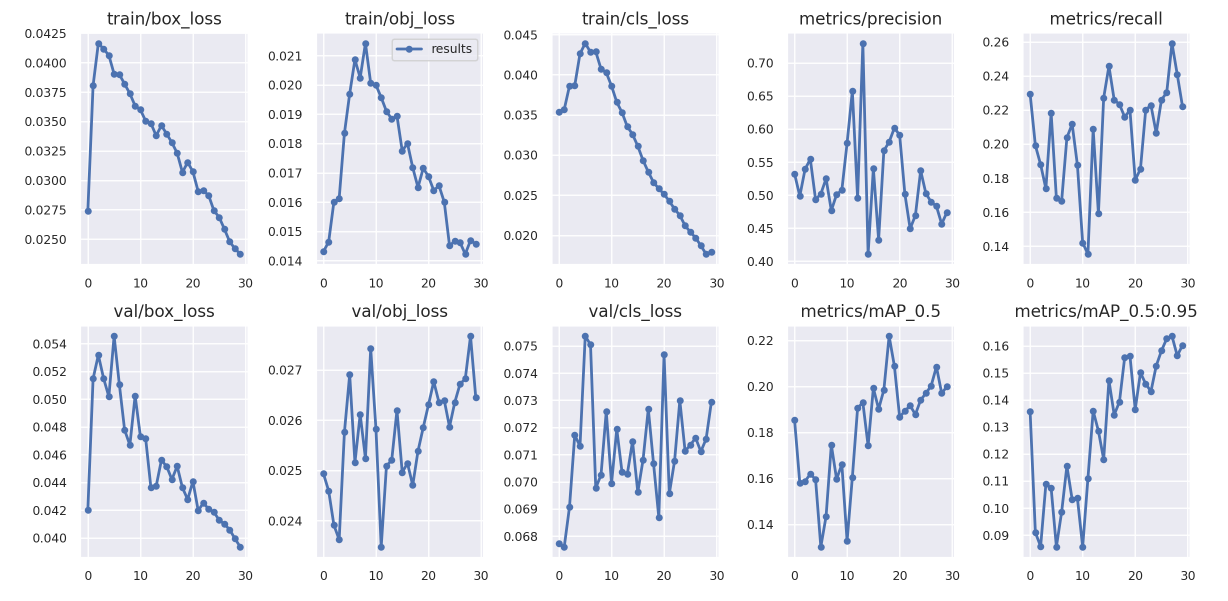
* **Notebook** : le Notebook « Entraînement » est à récupérer.
* **AWS SageMaker** : tu pourras utiliser AWS SageMaker pour cette séance.

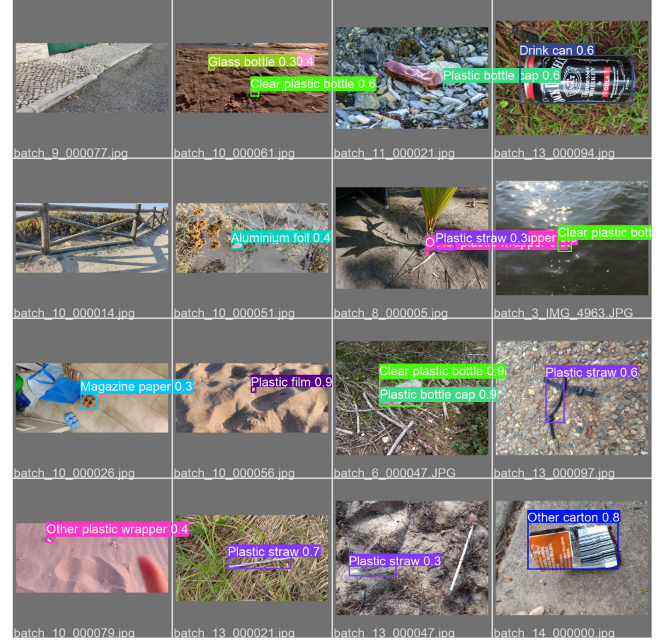
**Livrables**

*Les livrables indiquent ce que tu dois être en mesure de fournir à la fin de la séance.*

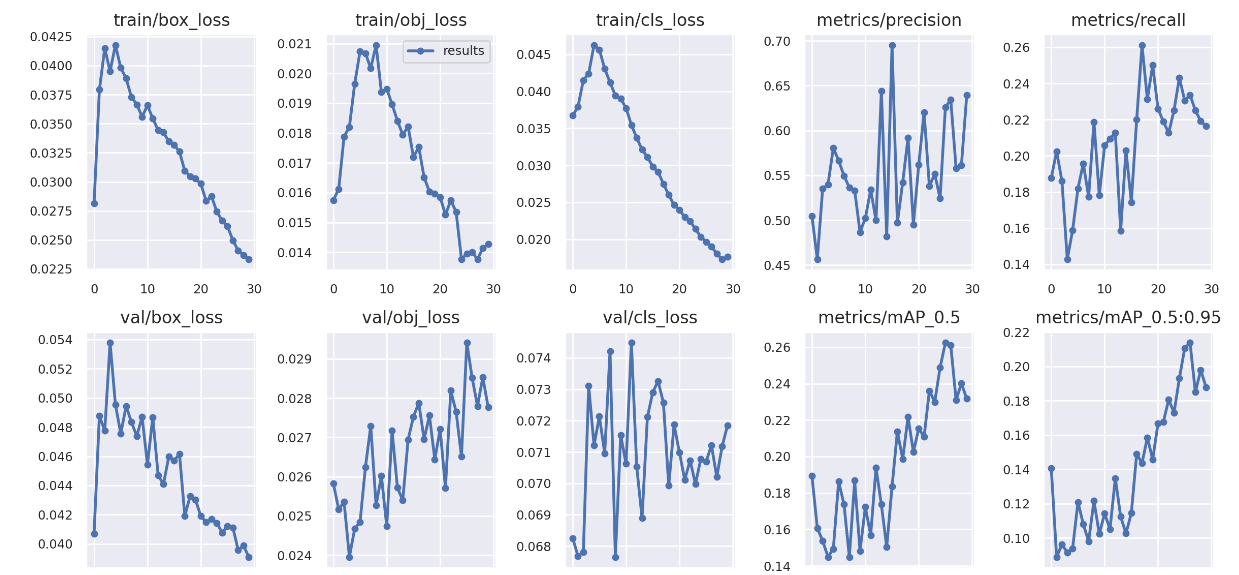
* **Fichiers de poids** : fichiers de poids (extension \*.pt) du modèle entraîné.
* **Code source de la séance** : code source permettant d’entraîner le modèle YoloV5 sur un GPU avec PyTorch.
* **Graphiques d'évaluation** : graphiques d’évaluation (courbes PR, F1, métriques au cours des itérations).

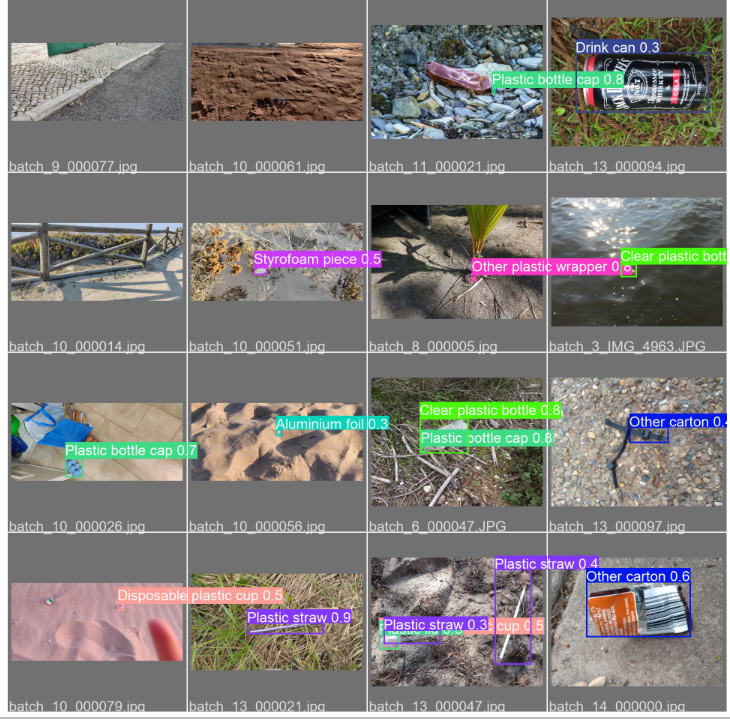
**Courbe de modélisation avant l’entrainement :**



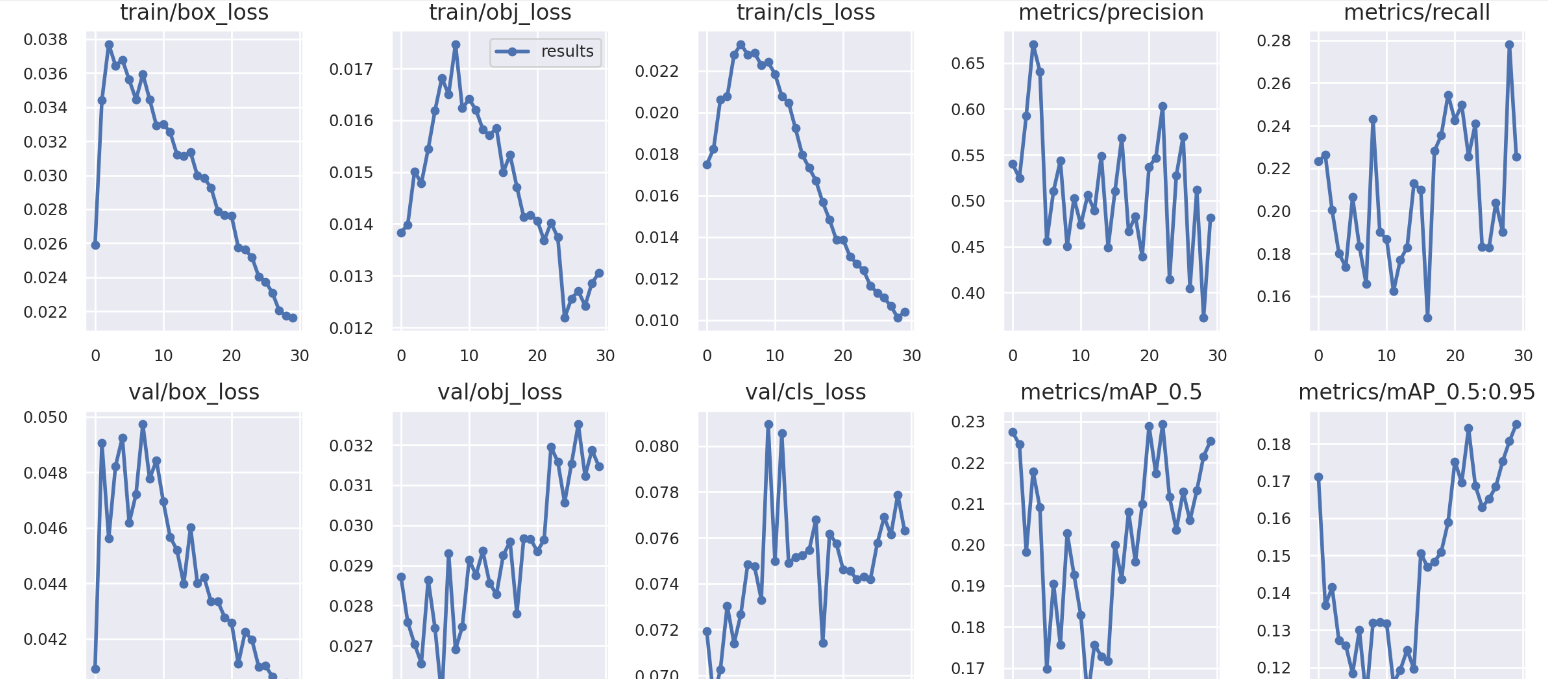


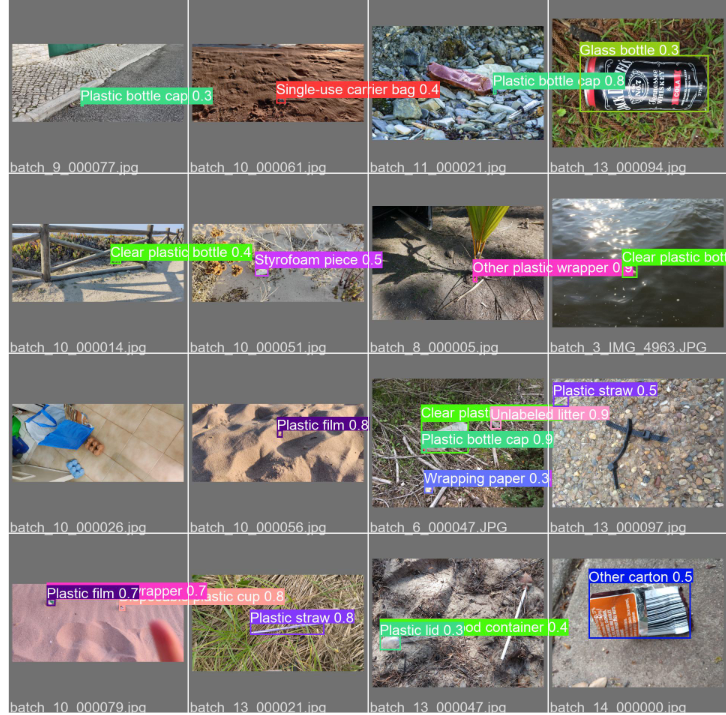
**Courbe de modélisation apres le premier entrainement :**

****



## Courbe de modélisation apres le second entrainement





**Séance 5 - Enrichissement des données**

Afin de proposer un POC personnalisé pour chaque équipe et adapté aux collectivités, l’objectif de cette séance est de récolter des données supplémentaires, notamment en identifiant et en prenant des photos de déchets que l’on peut trouver dans la rue.

Cette étape va permettre non seulement d’enrichir certaines catégories avec de nouveaux exemples, mais aussi de démontrer la force du POC avec des exemples concrets lors du Demo Day. Cela permet également d’ajouter des informations de géolocalisation sur les images que l’on pourra utiliser pour tester le POC lors de son déploiement.

**Activités prévus**

*Les activités prévues indiquent ce que tu vas faire lors de la séance*.

* Enrichir le jeu de données avec les images supplémentaires (en intérieur par exemple).
* Annoter les images avec un outil d’annotation assistée.

**Consignes**

*Les consignes te permettent de suivre pas-à-pas le déroulé de la séance*.

* Prendre des photos pour enrichir les catégories existantes et prendre les paramètres de géolocalisation des photos.

**Ressources à disposition**

*Les ressources à ta disposition sont disponibles dans l'onglet « Données » de la page du projet. Pour cette séance, tu auras besoin des ressources suivantes.*

* **Outil V7** : l’enrichissement peut être optimisée en utilisant des outils d’annotation assistée d’images avec V7.

**Livrables**

*Les livrables indiquent ce que tu dois être en mesure de fournir à la fin de la séance.*

* **Données enrichies** : photos des déchets qui ont été trouvés dans les rues.
* **Dataset enrichie** : annotations et catégorisation réalisées manuellement sur ces déchets.

**Photos prises pour l’enrichissement :**

****

****

****

**Etc…**

**Séance 6 - Annotations et entraînement sur GPU**

En utilisant des données supplémentaires et annotées récoltées par les étudiants, un modèle plus performant avec de nombreuses itérations sera entraîné. L’objectif ici est d’obtenir des performances beaucoup plus intéressantes que lors de la première modélisation tout en incluant des déchets trouvées dans les rues de la collectivité avec des métadonnées (coordonnées GPS par exemple).

**Activités prévus**

*Les activités prévues indiquent ce que tu vas faire lors de la séance*.

* Entraîner une première version du modèle sur les 60 classes du Taco Dataset avec les données enrichies.
* Récupérer les fichiers de poids du modèle entraîné.

**Consignes**

*Les consignes te permettent de suivre pas-à-pas le déroulé de la séance*.

* Télécharger le Notebook « Entraînement » et l’exécuter afin d’entraîner une nouvelle version du modèle Yolo avec les données enrichies.
* Sauvegarder en local les fichiers de poids qui pourront être réutilisés par la suite lors de l’inférence.

**Ressources à disposition**

*Les ressources à ta disposition sont disponibles dans l'onglet « Données » de la page du projet. Pour cette séance, tu auras besoin des ressources suivantes.*

* **Notebook** : le Notebook « Entraînement » est à récupérer.
* **AWS SageMaker** : tu pourras utiliser AWS SageMaker pour cette séance.

**Livrables**

*Les livrables indiquent ce que tu dois être en mesure de fournir à la fin de la séance.*

* **Fichiers de poids** : fichiers de poids (extension \*.pt) du modèle entraîné.
* **Code source de la séance** : code source permettant d’entraîner le modèle YoloV5 sur un GPU avec PyTorch.
* **Graphiques d'évaluation** : graphiques d’évaluation (courbes PR, F1, métriques au cours des itérations).