



Challenge Machine Learning

Identification automatique de réactions chimiques

École d'Été sur l'Intelligence Artificielle, EEIA 2026

Contexte

Dans de nombreuses industries, qu'il s'agisse de la pharmacie, de l'agroalimentaire, du traitement de l'eau ou de la chimie fine, il est capital de **surveiller les réactions chimiques** qui se déroulent dans les réacteurs et de **prédire leur comportement**. Pour cela, des capteurs mesurent à intervalle régulier la **concentration d'une molécule d'intérêt** au cours de la réaction.

Selon les conditions opératoires (température, catalyseur, réactifs en présence), la réaction observée appartient à l'une de **4 grandes familles cinétiques**, notées **A, B, C et D**. Identifier automatiquement la famille d'une réaction en cours permettrait de :

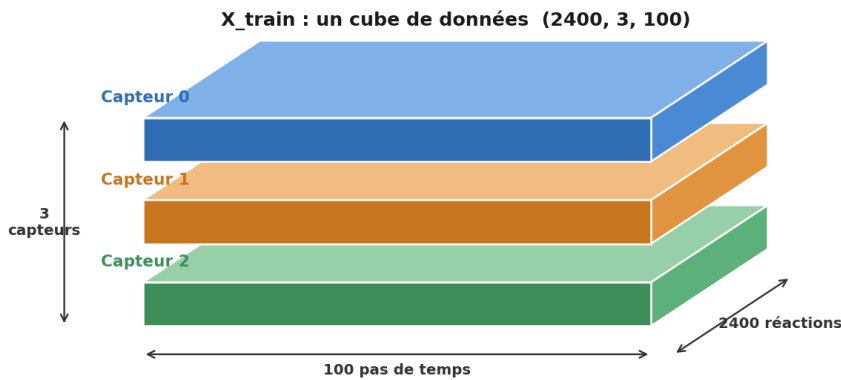
- **Piloter le procédé** : adapter en temps réel la température ou l'ajout de réactifs ;
- **Détecter des anomalies** : une réaction classée dans une famille inattendue peut signaler une contamination ou une dérive du procédé ;
- **Contrôler la qualité** : en pharmaceutique, vérifier que chaque lot suit bien la cinétique attendue ;
- **Gagner du temps** : éviter des analyses manuelles longues et coûteuses en laboratoire.

Votre mission : construire un modèle de classification qui, à partir des mesures des capteurs, prédit la famille (A, B, C ou D) de chaque réaction.

Les données

Fichier	Contenu
X_train.npy	Mesures des capteurs pour les réactions d' entraînement
y_train.csv	La famille (A, B, C ou D) de chacune de ces réactions
X_test.npy	Mesures pour les réactions dont vous devez prédire la famille

Chaque réaction est décrite par les enregistrements de **3 capteurs**, chacun échantillonné sur **100 pas de temps**. Les données d'entraînement forment donc un cube à trois dimensions :



Concrètement, $X_train[i, c, t]$ est la valeur mesurée par le capteur c au pas de temps t pour la réaction i . À vous d'explorer ce que mesure chaque capteur et de déterminer lesquels sont utiles.

Évaluation

Les soumissions sont évaluées avec l'**accuracy** : la proportion de réactions dont la famille est correctement prédite.

$$\text{accuracy} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(\hat{y}_i = y_i) = \frac{\text{nombre de prédictions correctes}}{\text{nombre total de prédictions}}$$

L'accuracy varie entre 0 (aucune bonne réponse) et 1 (toutes les réponses correctes). Les 4 familles étant équilibrées dans le jeu de données, un modèle répondant au hasard obtiendrait environ 0,25.

Format de soumission

Votre fichier de soumission est un CSV contenant **une seule colonne** nommée `label`, avec une ligne par réaction de X_test , **dans le même ordre que le fichier** et sans colonne d'index :

```
label
A
C
B
...
```

Les labels doivent être les **lettres** A, B, C ou D. Si vous avez encodé la variable cible en nombres pendant l'entraînement, pensez à la reconvertir avant de soumettre.

Règles du challenge

Plateforme du challenge :

<https://eeia-ml-challenge.beninexcellence.org/>

C'est sur cette plateforme que vous déposez vos soumissions et que vous consultez le classement.

- **5 soumissions par jour et par participant** au maximum. Utilisez-les avec discernement : construisez d'abord un jeu de validation fiable pour tester vos idées localement, sans consommer vos soumissions.

- **Clôture du challenge** : dimanche 9 août 2026 à 00h00. Aucune soumission ne sera acceptée au-delà de cette échéance.
- Le classement public est établi sur l'accuracy obtenue sur le jeu de test.

Phase finale : le classement ne fait pas tout

À la clôture, les **5 meilleurs participants** au classement seront invités à **présenter leur travail** devant le jury : leur démarche, les pistes explorées, celles qui ont échoué, et le raisonnement qui les a menés à leur résultat final.

Rien n'est joué d'avance. Le vainqueur du challenge est désigné à l'issue de ces présentations. Un participant dont le score est légèrement inférieur à ce stade, mais dont la démarche est plus brillante ou mieux expliquée, peut tout à fait terminer premier.

Le jury sera notamment attentif :

- à la **qualité de l'exploration des données** et à la compréhension du problème ;
- à la **justification des choix** de prétraitement, de modèle et de validation ;
- à la **rigueur méthodologique** (protocole de validation, absence de fuite de données) ;
- à la **clarté de l'exposé** et à la reproductibilité du travail.

Documentez donc vos expériences au fil de l'eau : gardez une trace de vos essais, de leurs scores et de ce que vous en avez conclu. Ce carnet de bord fera la différence le jour de la présentation.

Pour démarrer

Un **notebook starter** vous est fourni. Il contient une chaîne complète de bout en bout : chargement des données, exploration, prétraitement, validation, modèles de référence et génération du fichier de soumission ; il propose également des pistes de travail. Les scores qu'il obtient sont volontairement modestes : ils constituent le point de départ que vous devrez dépasser.

Rendez-vous sur <https://eeia-ml-challenge.beninexcellence.org/> pour déposer votre première soumission. Bon challenge !