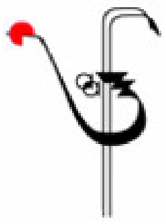
AgroParisTech

université PARIS-SACLAY

Analyse de données pour la **prédiction des valeurs nutritionnelles des fourrages** pour ruminants par **machine learning** et **Large Language Models**

Un projet de l'**Association Française de Zootechnie (AFZ)** encadré par Valérie Heuzé et Gilles Tran


Réalisé par
Aristide Lauront, Matéo Petitet, Raphaël Genin et Raphaël Rubrice




But : utiliser un **LLM** afin de prédire plus précisément des valeurs nutritionnelles de fourrages

---

Pourquoi ce besoin ?

- calcul exact coûteux
- description du fourrage complexe
- fortes variation

Pour quel usage ?

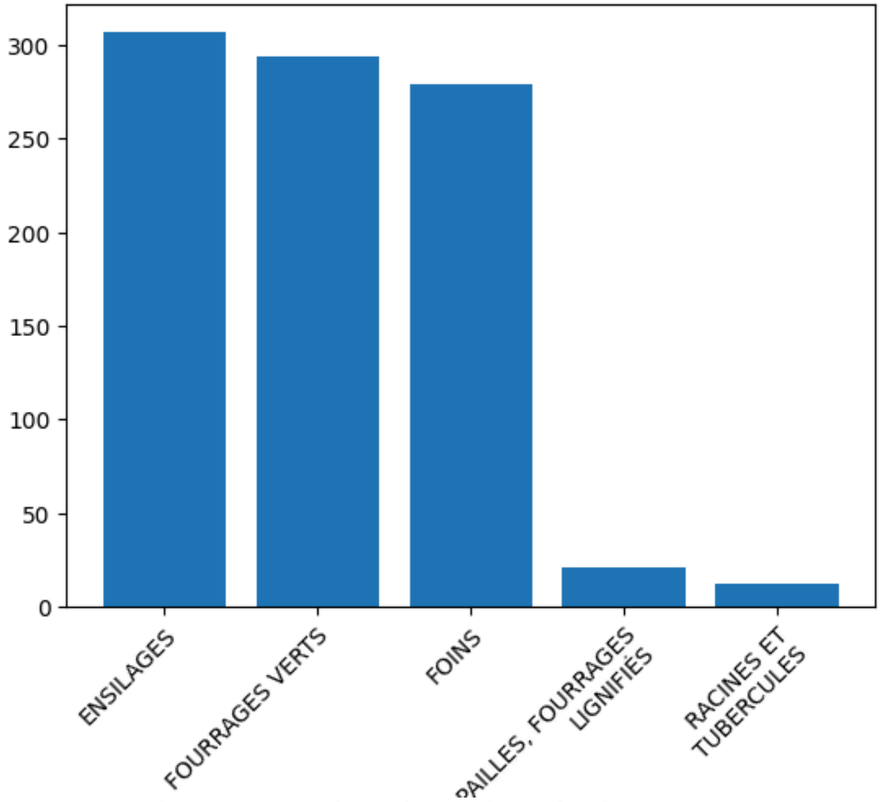
- adaptation en direct des rations
- suivi précis des apports

Type et forme des **données**

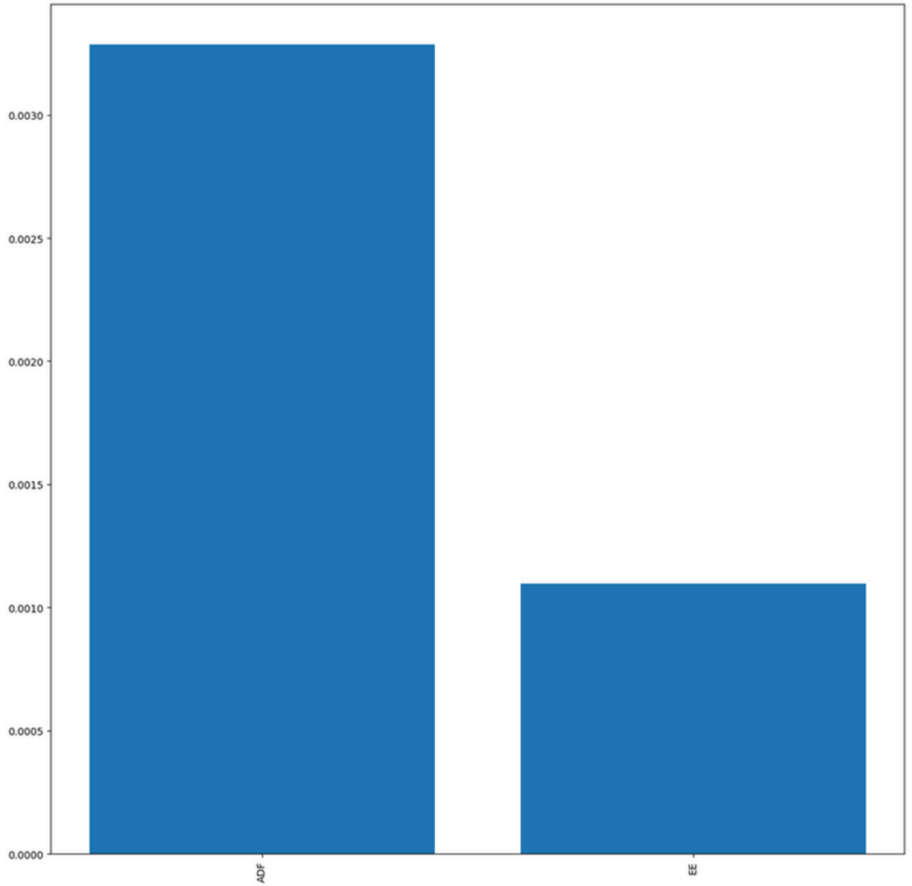
Deux fichiers excel décrivant des caractéristiques de fourrages et de concentrés ont été fournis.
Chaque aliment est caractérisé par :

- un numéro de ligne
- son **identifiant** INRAe unique
- **5 libellés** de plus en plus précis (du libellé 0, systématique, au libellé 4, facultatif) ; ces libellés décrivent qualitativement le fourrage
- **92 valeurs chimiques** mesurées en laboratoire ; nous cherchons à prédire 5 d'entre elles

Type et forme des **données**



Répartition des données de fourrages



Proportion de valeurs manquantes dans les valeurs utilisées (feature non présentes : pas de valeur manquante)

Focus sur les valeurs **recherchées**

Valeurs définies par l'INRAe

- **UFL** : Unité Fourragère Lait, quantité d'énergie nette absorbable pendant la lactation ou l'entretien du ruminant (1 UFL = 1700 kcal)
- **UFV** : Unité Fourragère Viande, quantité d'énergie nette absorbable lors de l'engraissement d'un ruminant (1 UFL = 1820 kcal)
- **PDI** : Protéines Digestibles dans l'Intestin, valeurs nutritives en azote (protéines métabolisables) chez les ruminants
- **PDIA** : PDI d'origine Alimentaire, non dégradées dans le rumen
- **BPR** : Bilan Protéique du Rumen, différence entre les protéines ingérées et celles passant au duodénum

Focus sur les valeurs **utilisées**

Valeurs mesurables facilement et à faible coût

- **MS** : Matière Sèche, le restant après retrait de toute l'eau du produit
- **MM** : Matière Minérale, portion non-organique du produit
- **MAT** : Matière Azotée Totale, protéines brutes
- **CB** : Cellulose Brute
- **NDF** : Neutral Detergent Fiber, fibres totales présentes dans un aliment
- **ADF** : Acid Detergent Fibers, fibres totales présentes dans un aliment sauf hémicellulose
- **EE** : EthylE

Focus sur les **libellés**

5 niveaux de précision sont donnés par les libellés

- **Niveau 0** : catégorie générale de l'aliment considéré (ex : FOURRAGES VERT) ; **5 modalités**
- **Niveau 1** : sous-catégorie de l'aliment (ex : PRAIRIES PERMANENTES, PLAINE (NORMANDIE)) ; **53 modalités**
- **Niveau 2** : précisions sur les conditions de culture et/ou récolte de l'aliments (ex : 1er cycle) ; **42 modalités**, présence non-systématique
- **Niveau 3** : précisions supplémentaires sur les conditions de culture et/ou récolte de l'aliments (ex : 15-25 avril, déprimage, ST = 172°C) ; **113 modalités**, présence non-systématique, informations parfois de même nature que le niveau 2
- **Niveau 4** : informations complémentaires sur l'aliment (ex : Épiaison du dactyle) ; **55 modalités**, présence non-systématique

Les libellés sont standards du niveau 0 à 2, et deviennent ensuite plus imprécis et inconsistants.

Données **additionnelle**

Fichiers excel d'entraînement des modèles insuffisamments dotés en vocabulaire dans les libellés pour entraîner un LLM ; ajout de la base **Feedipedia** (© INRAE CIRAD AFZ FAO) intégrale en français et anglais, fournissant un corpus détaillé relatif à l'alimentation animale.

Ce corpus se présente sous forme d'un fichier excel contenant des **noms et descriptions d'aliments** en français et en anglais. Au total, **16 186** lignes de données y sont présentées.

Approche Machine Learning

Objectif : Utiliser des algorithmes de Machine Learning afin d'apprendre le lien entre les variables d'entrée et celles de sortie.

- Comment tirer parti des libellés ?
- Quel niveau de performance de base peut on atteindre ?

Approche Machine Learning : **Pré-traitemements**

Différents prétraitements à considérer :

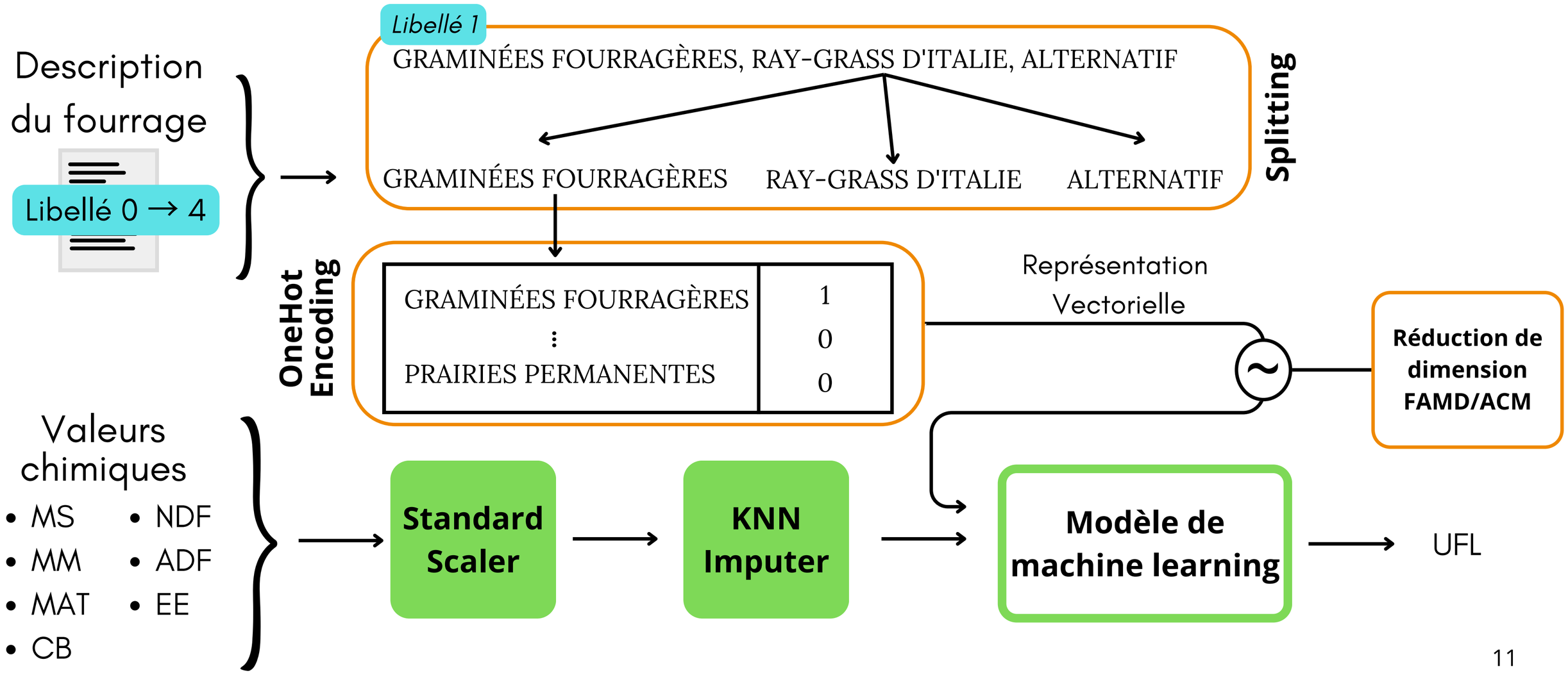
> Pour les valeurs numériques :

- Normalisation
- Gestion des valeurs manquantes

> Pour les libellés

- Vectorisation : encodage disjonctif complet ‘OneHot’
- Réduction eventuelle de dimension :
  - Analyse en Composante Multiple (ACM) ;
  - Analyse de facteurs mixtes (FAMD).

Approche Machine Learning : **Pré-traitemements (2)**



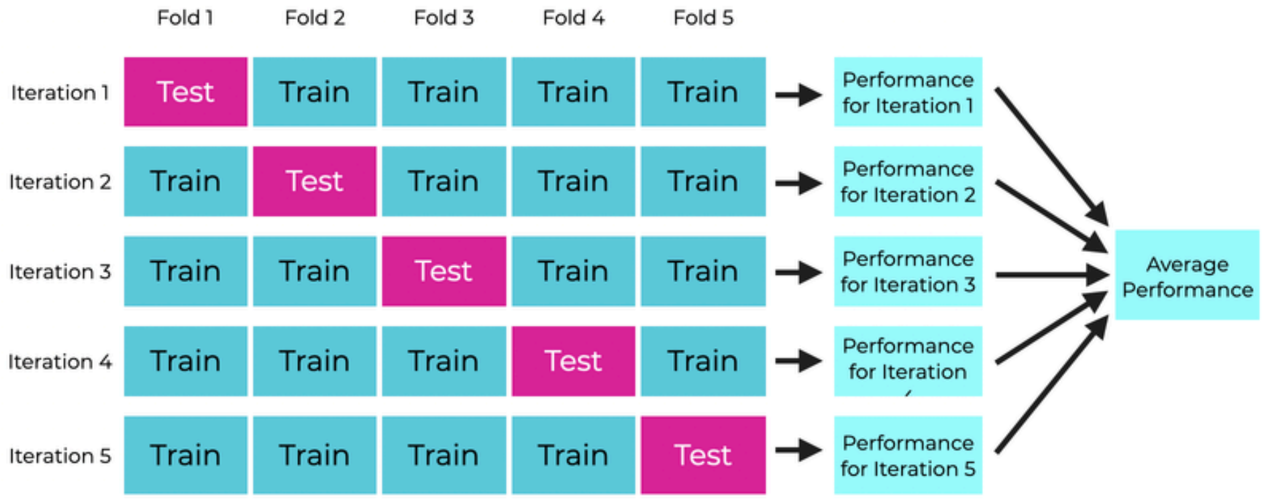
Approche Machine Learning : **Protocole**

1) Standardisation $$z = \frac{x_i - \mu}{\sigma}$$

2) Remplacement des valeurs manquantes (KNN Imputer)

3) Validation croisée en 5 plis

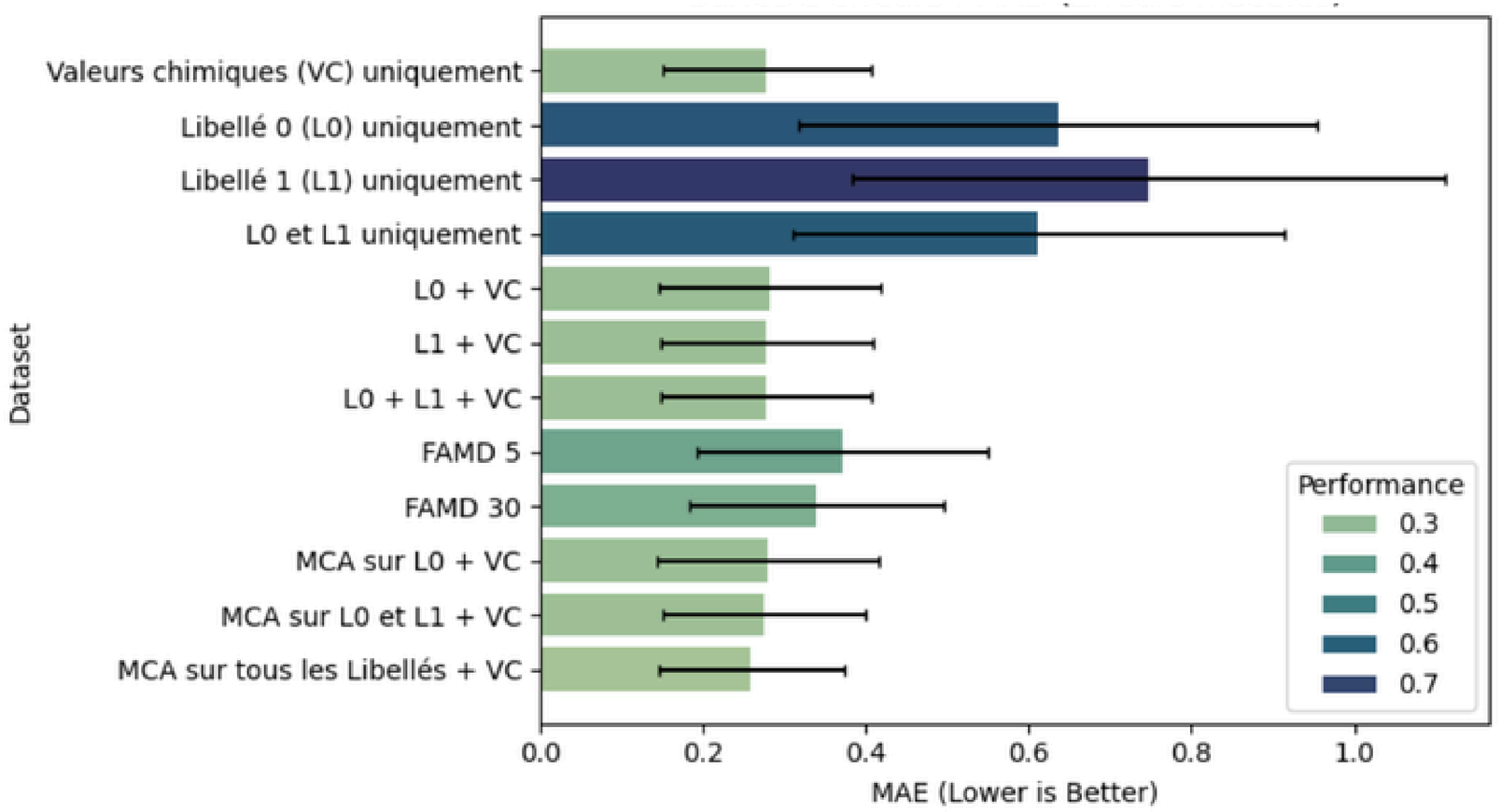
4) Réalisé pour plusieurs algorithmes.



Approche Machine Learning : **Résultats**

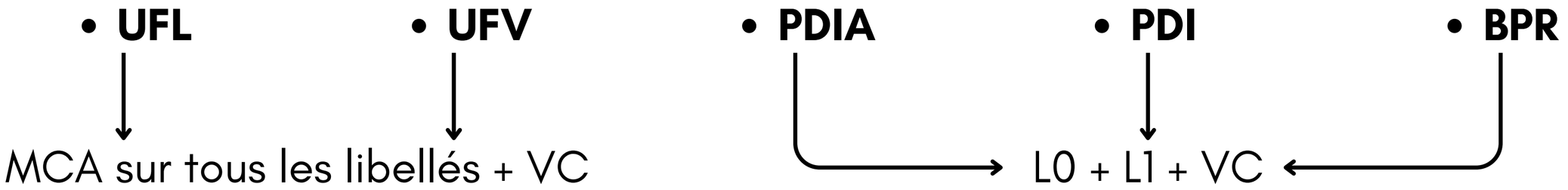
Quel pré-traitement apporte les meilleurs résultats ?

Exemple pour les valeurs UFL :



Approche Machine Learning : **Résultats**

Quel pré-traitement apporte les meilleurs résultats ?

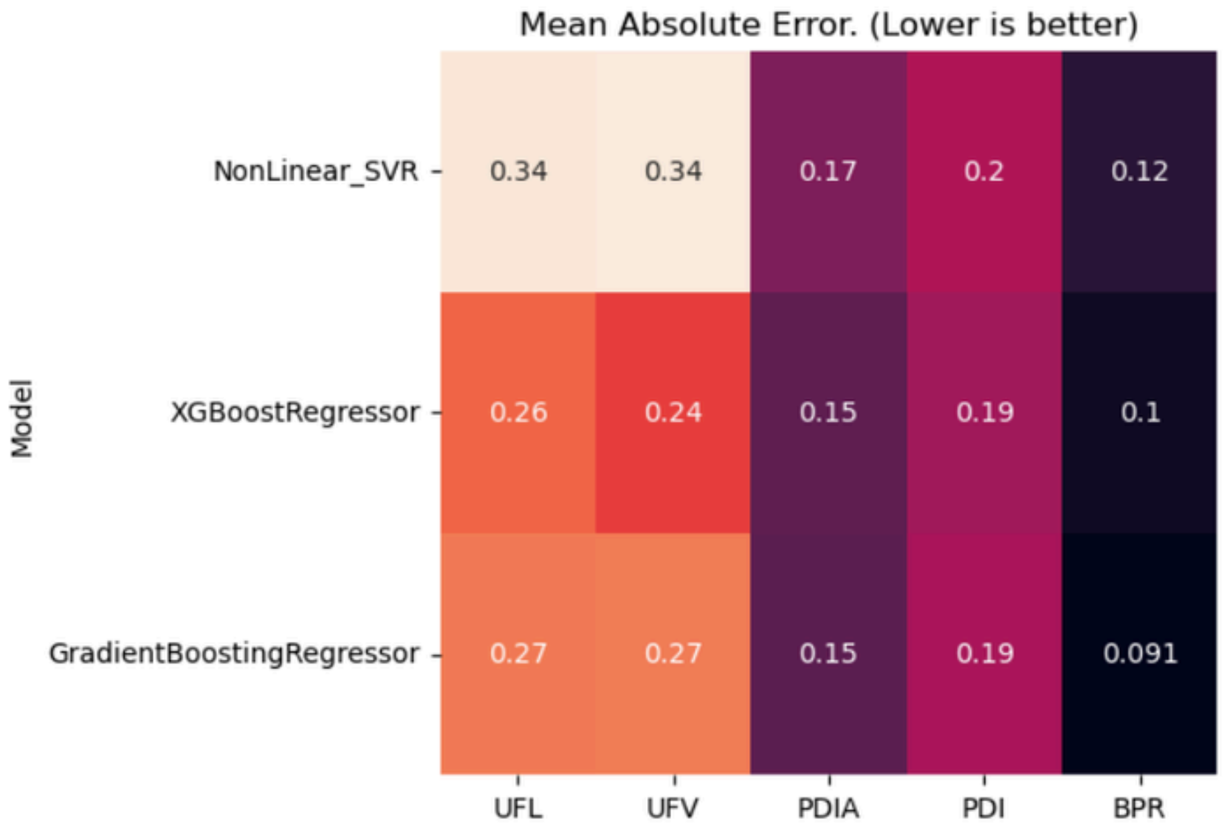


2 algorithmes performent mieux que les autres :

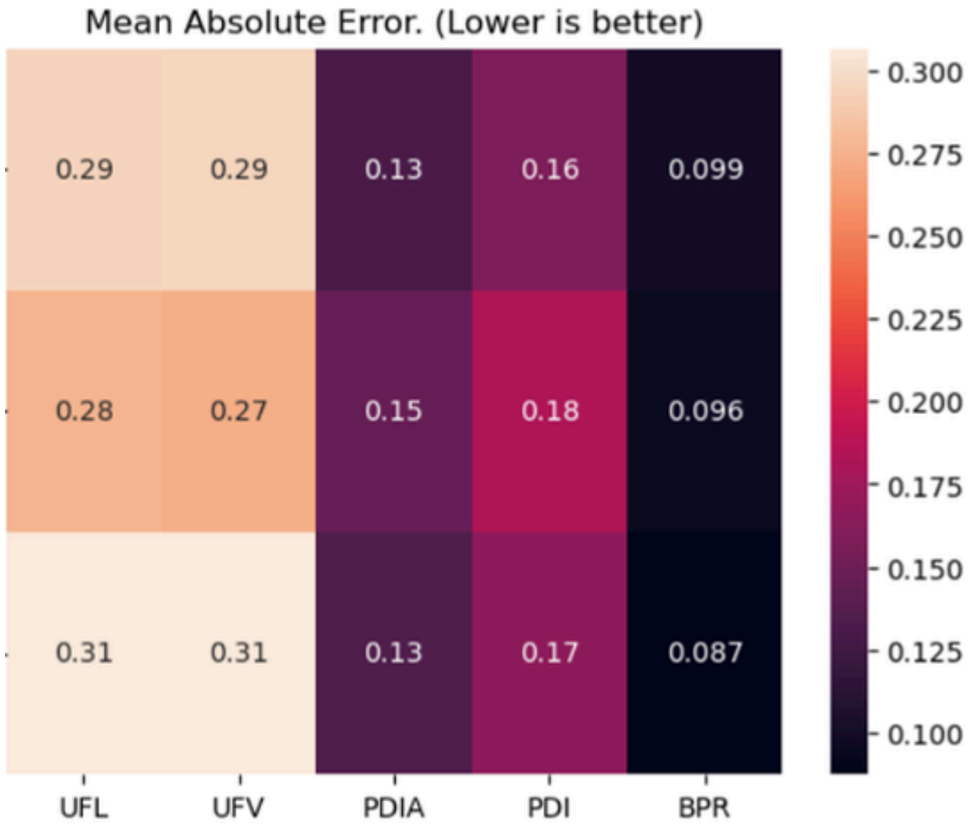
- Séparateur à Vaste Marge à noyau Non Linéaire (rbf)
- Ensemble d'Arbre de régression avec Boosting de Gradient

Approche Machine Learning : **Résultats**

MCA sur tous les libellés + VC



L0 + L1 + VC



Très similaires. XGBoostRegressor.

Approche Deep Learning (en cours)

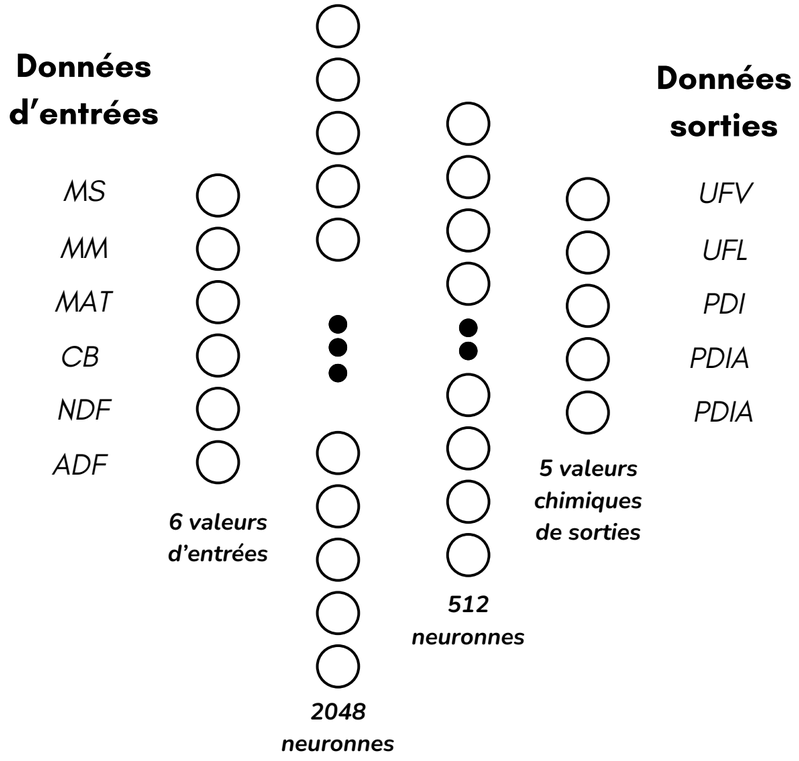
Objectif : Possibilité pour l'exploitant de décrire son fourrage en language naturel.

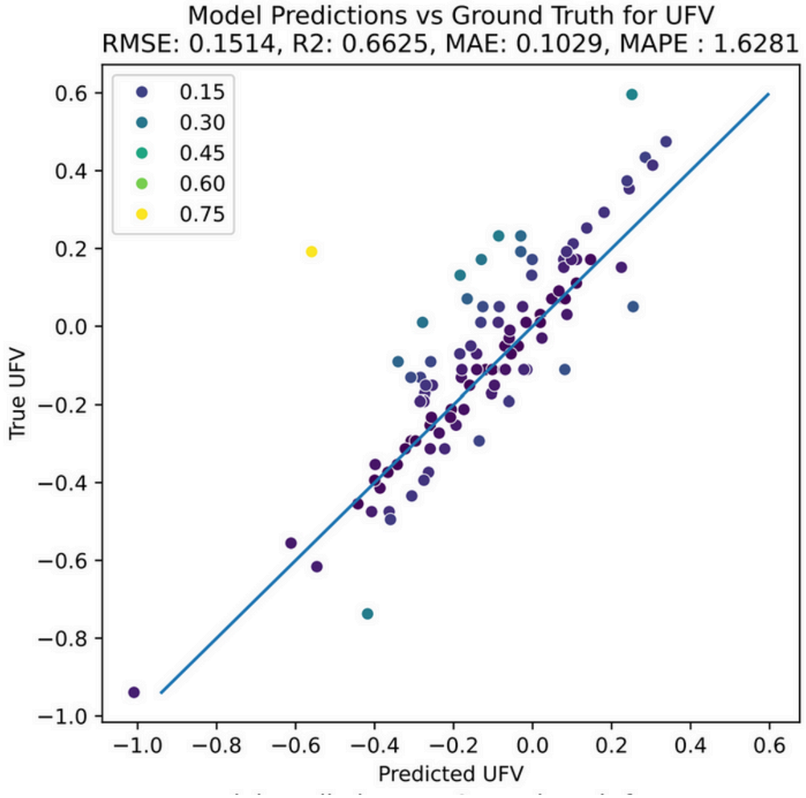
- Les approches Machine Learning nous limitent à des listes déroulantes
- Le Deep Learning est désormais bien connu pour traiter le Language Naturel

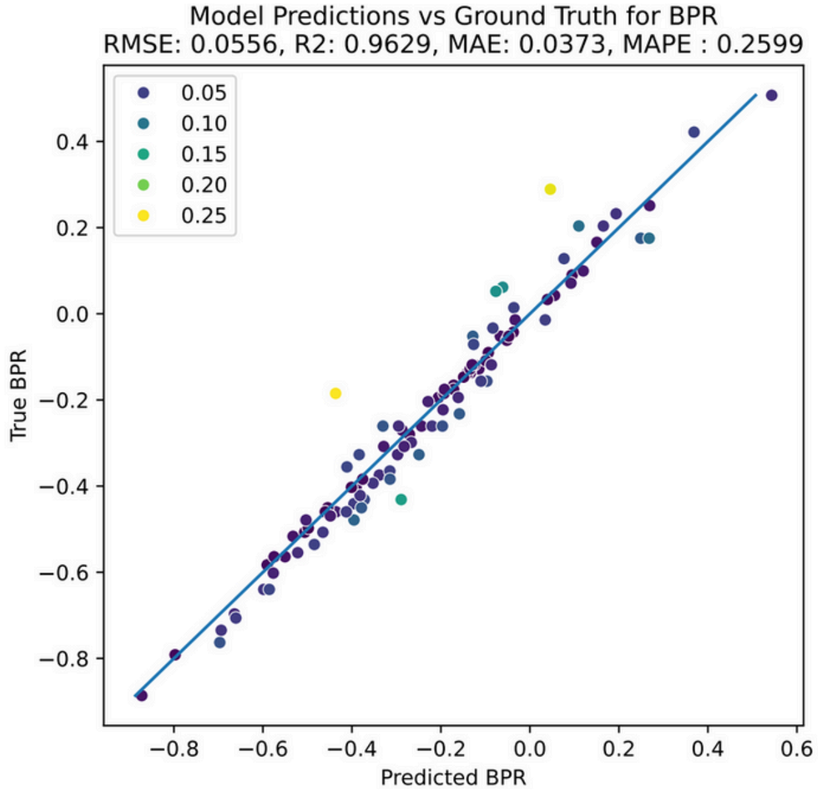
=> Développer un système capable de prédictions à partir des valeurs infrarouges ET d'une description textuelle du fourrage.

Approche Deep Learning : **MLP simple**

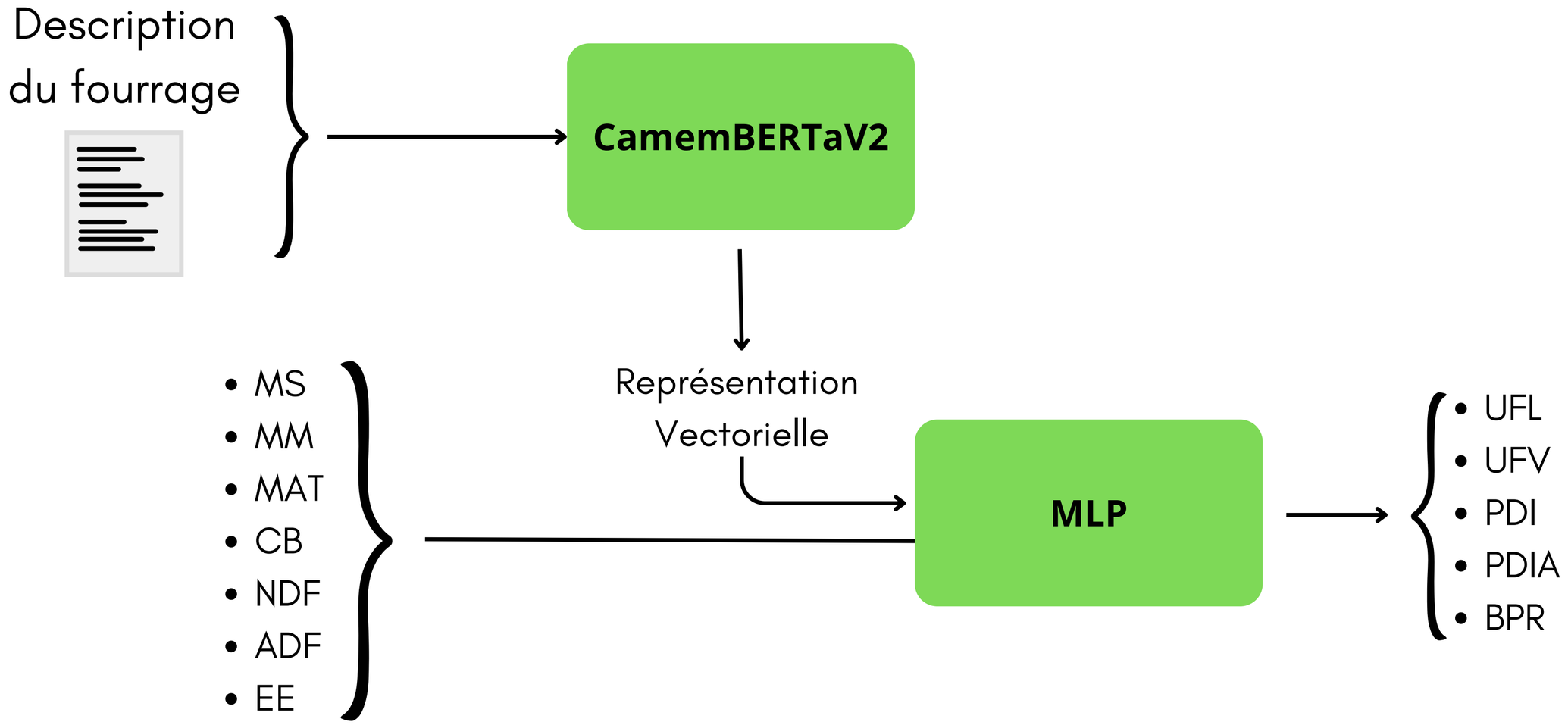
Objectif : réaliser un perceptron multicouche pour approcher les valeurs chimiques à partir des données d'entrées numériques





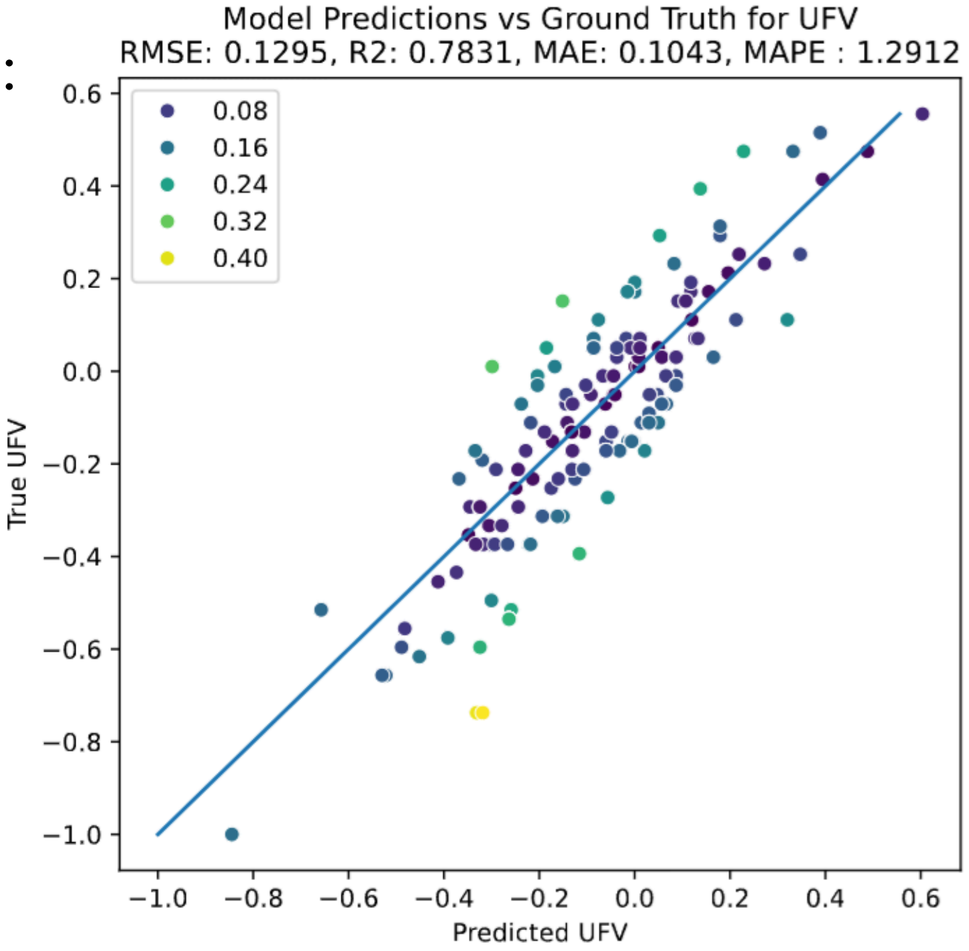


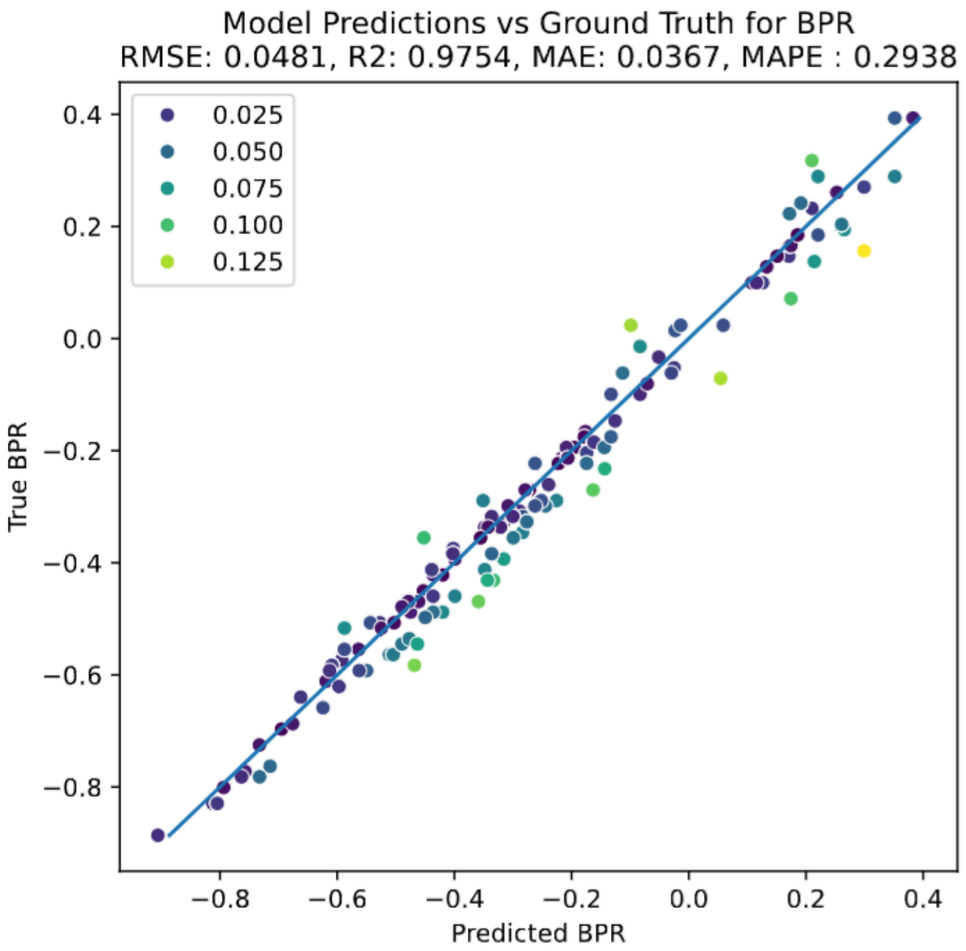
Approche Deep Learning : **CamemBERTaV2**



Approche Deep Learning : **CamemBERTaV2**

Exemples :





=> Plus globalement, les performances sont améliorées

Démonstration d’utilisation du projet (en cours)

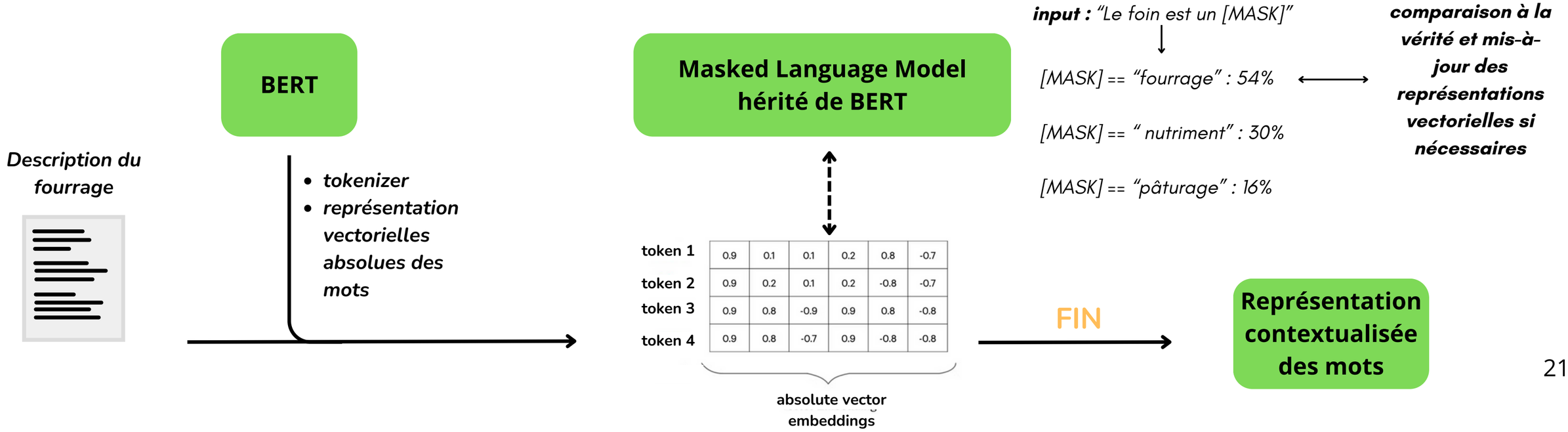
Démo web

Futures approches

Objectif : réaliser une adaptation des représentations vectorielles des mots au contexte agricole

Résultats attendus :

- des meilleurs résultats à partir des combinaisons des descriptions dispoibles des données
- des meilleurs résultats en “zero-shots” (ie. : pour des contextes et descriptions inconnues)



Pour la suite

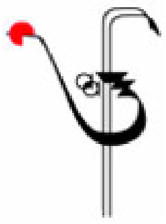
- Identifier la meilleure solution au problème

(LLM pré entrainé brut ? Fine-tuné ? Construire un nouveau modèle entièrement ?)

- Améliorer la solution trouvée
- Finir la démo
- Rapport et soutenance

Merci pour votre attention !

Un projet de l'**Association Française de Zootechnie (AFZ)** encadré par Valérie Heuzé et Gilles Tran

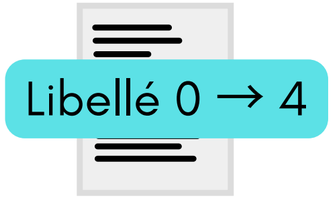




Réalisé par Aristide Lauront, Matéo Petitet, Raphaël Genin et Raphaël Rubrice




Approche Machine Learning : **Pré-traitemements (3)**

Description du fourrage

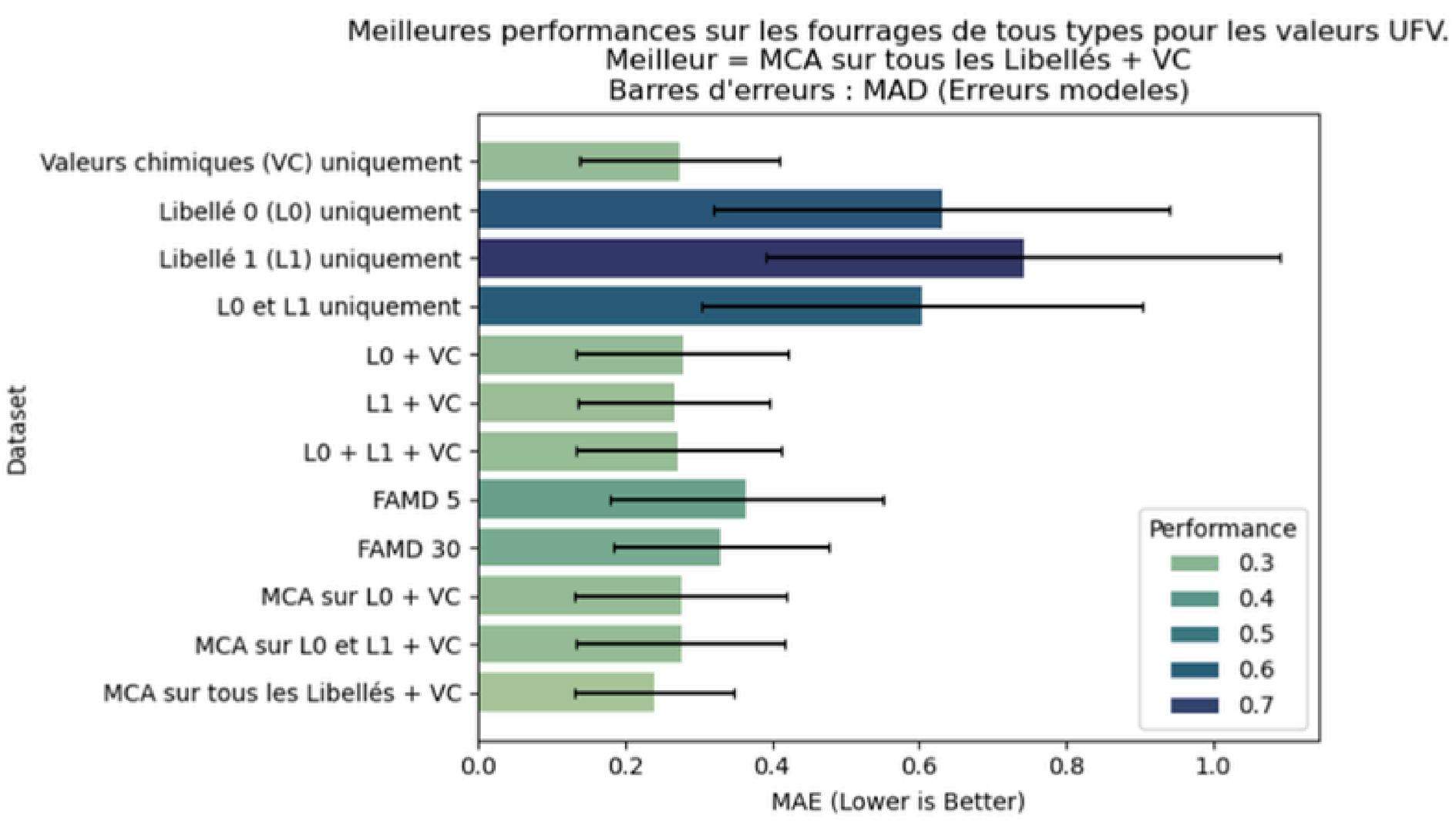


Valeurs chimiques

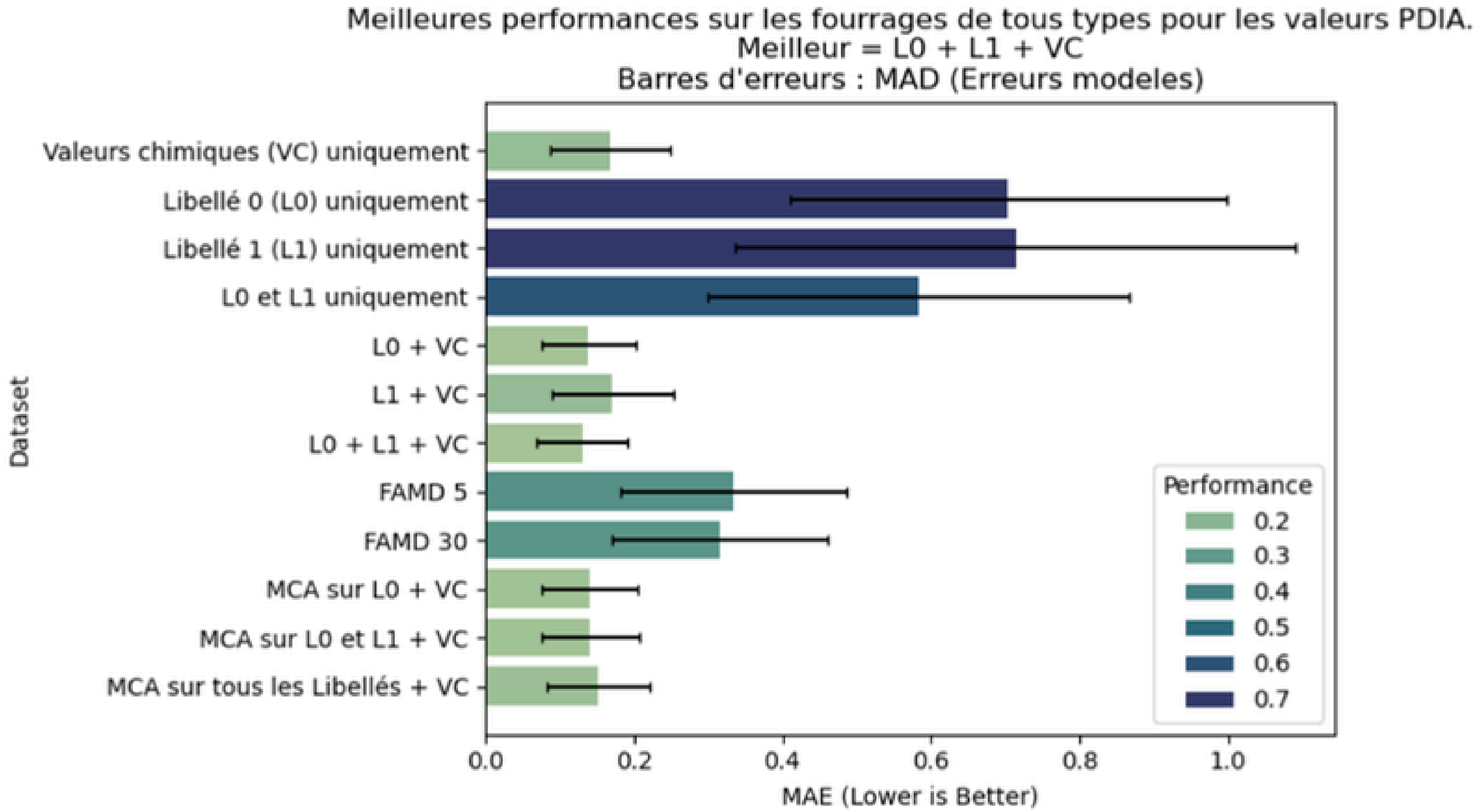
- MS
- MM
- MAT
- CB
- NDF
- ADF
- EE

Valeurs chimiques (VC)	×		
Libellé 0 (L0)			×
Libellé 1 (L1)			×
Libellé 0 & 1 (L0+L1)			×
L0 +VC			
L1 +VC			
L0 + L1 +VC			
FAMD 5			
FAMD 30			
ACM + VC			

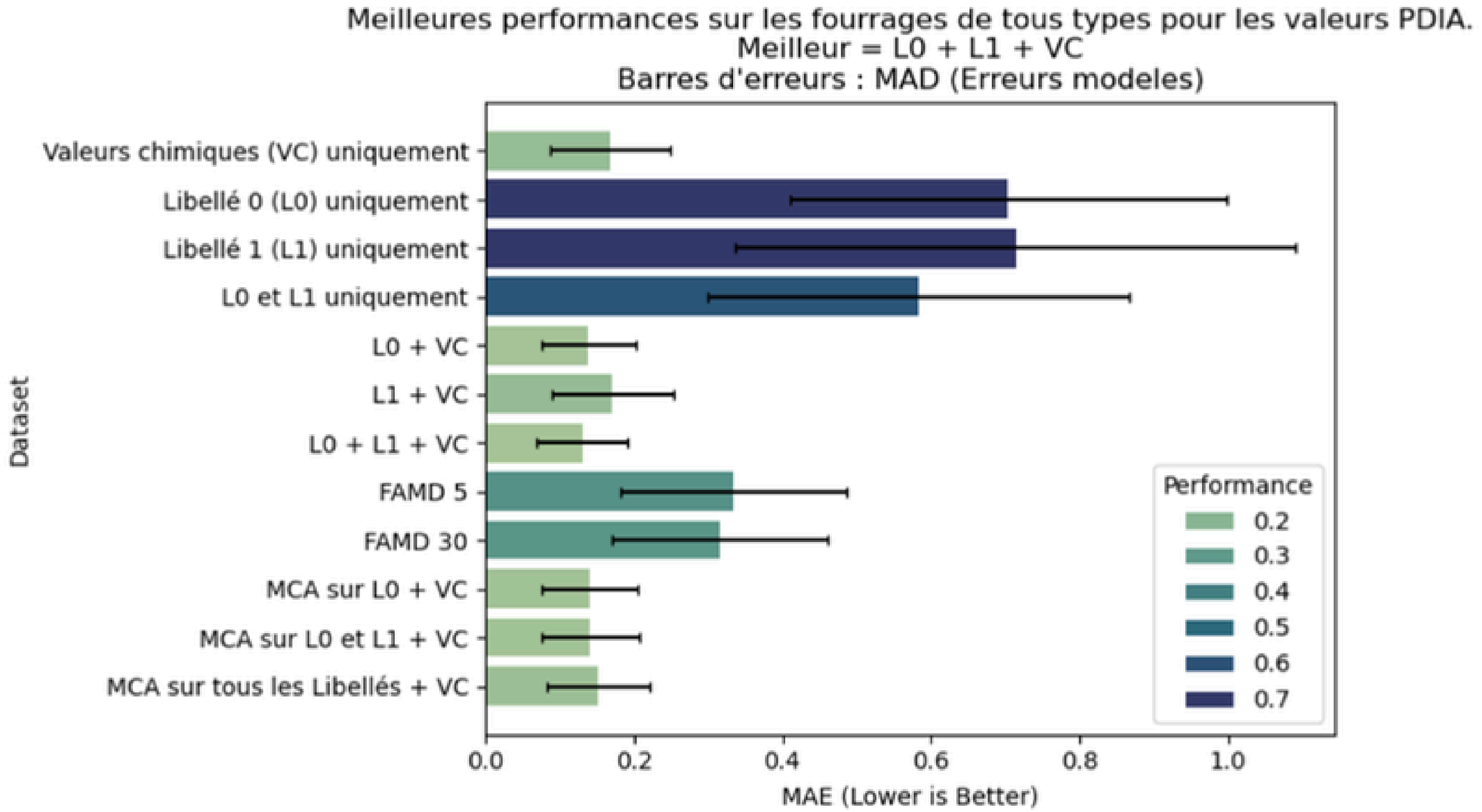
Meilleurs pré-traitements UFV



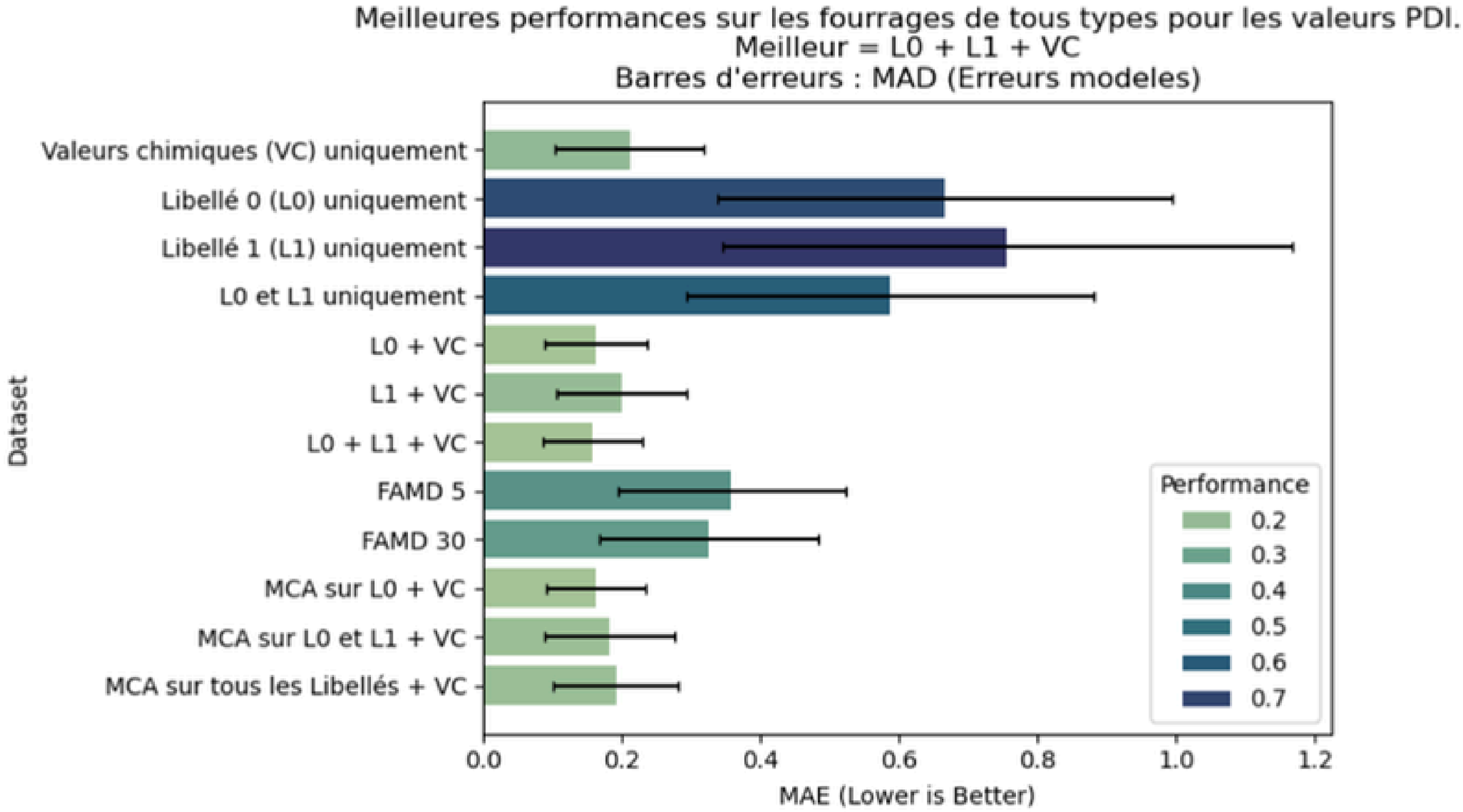
Meilleurs pré-traitements PDIA



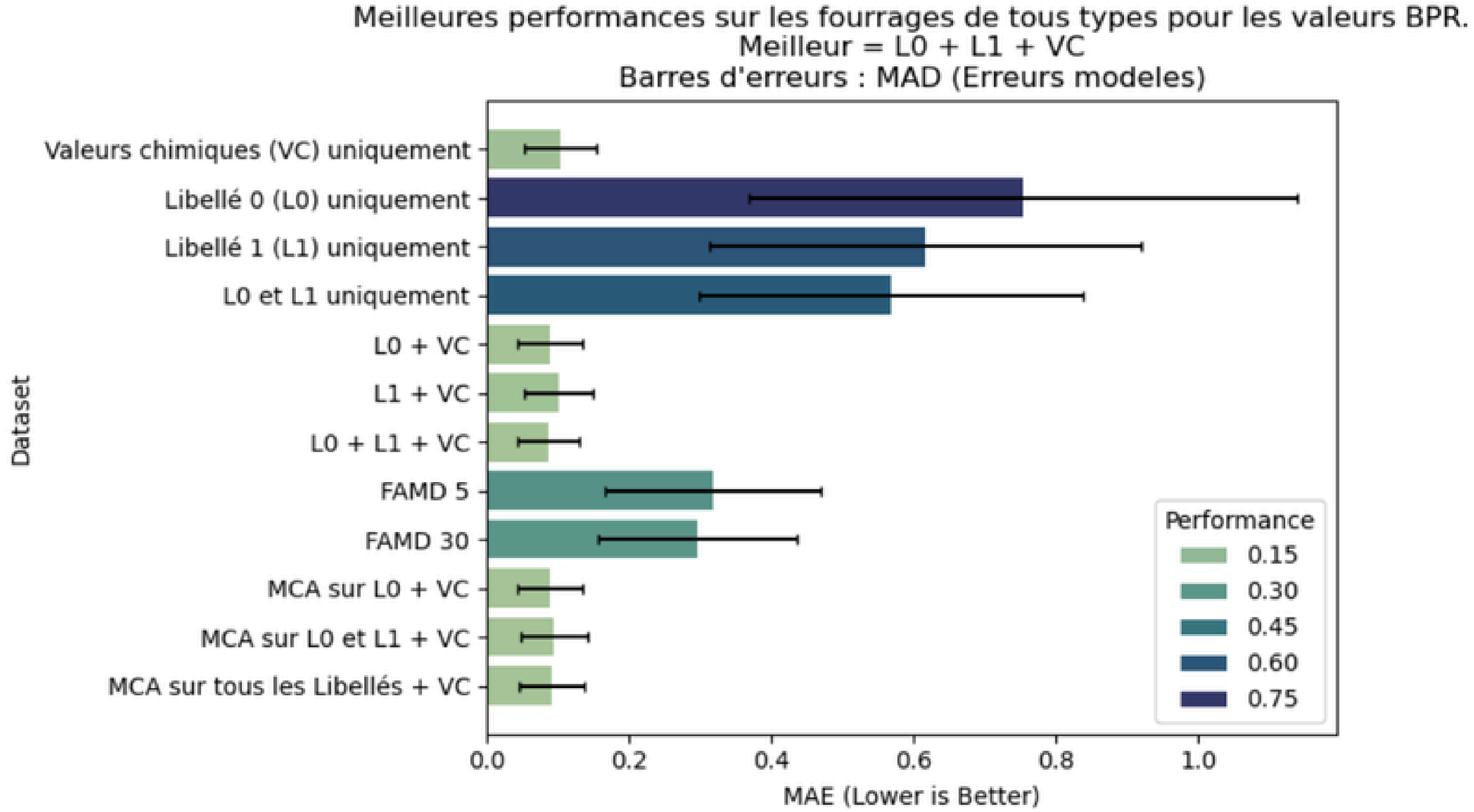
Meilleurs pré-traitements PDIA



Meilleurs pré-traitements PDI

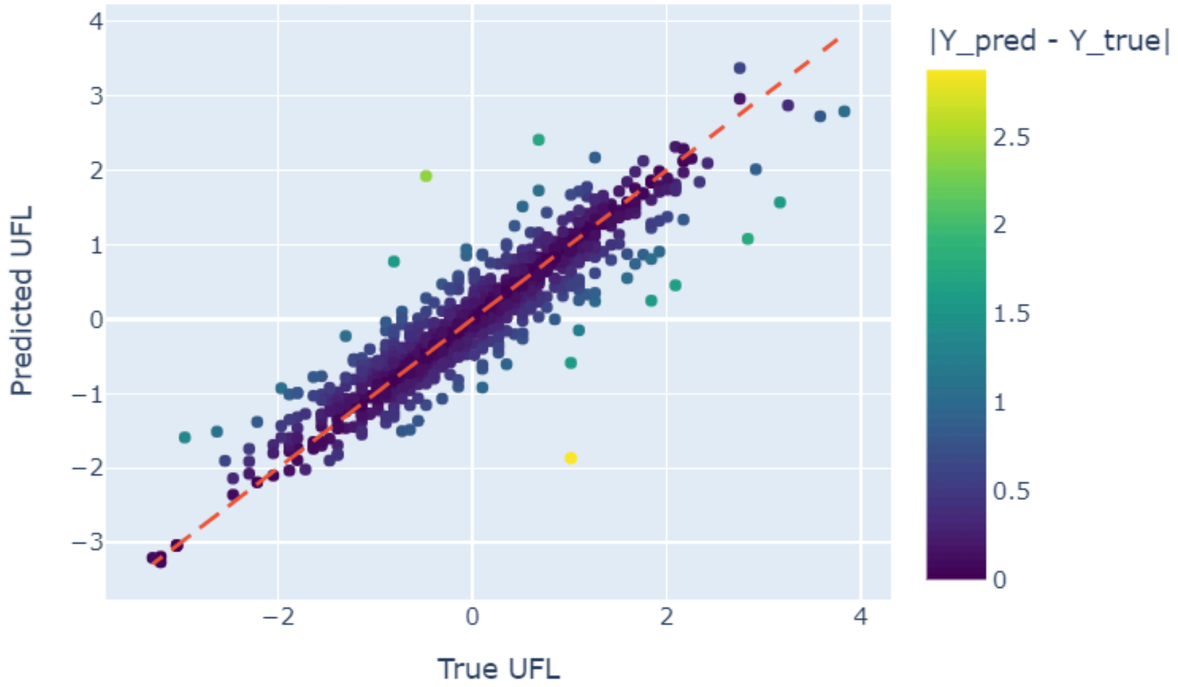


Meilleurs pré-traitements BPR

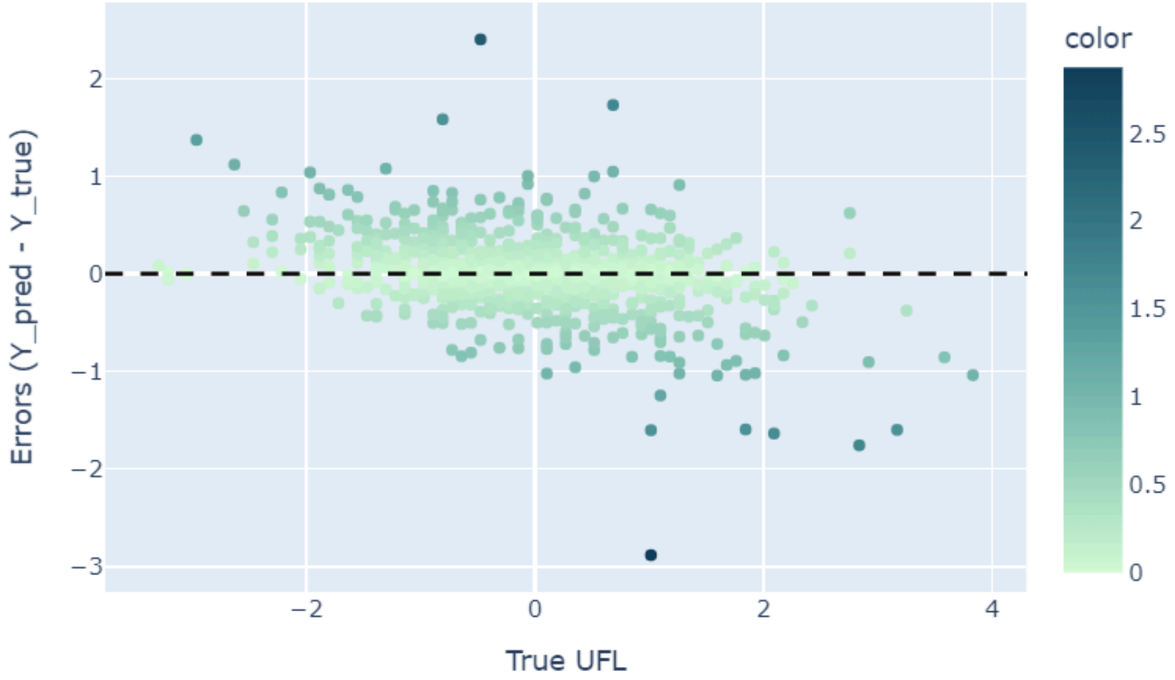


Meilleur modèle pour UFL

Meilleur Modèle pour prédire UFL: XGBoostRegressor
Dataset:MCA sur tous les Libellés + VC
MAE score: 0.2591, R² score: 0.8521

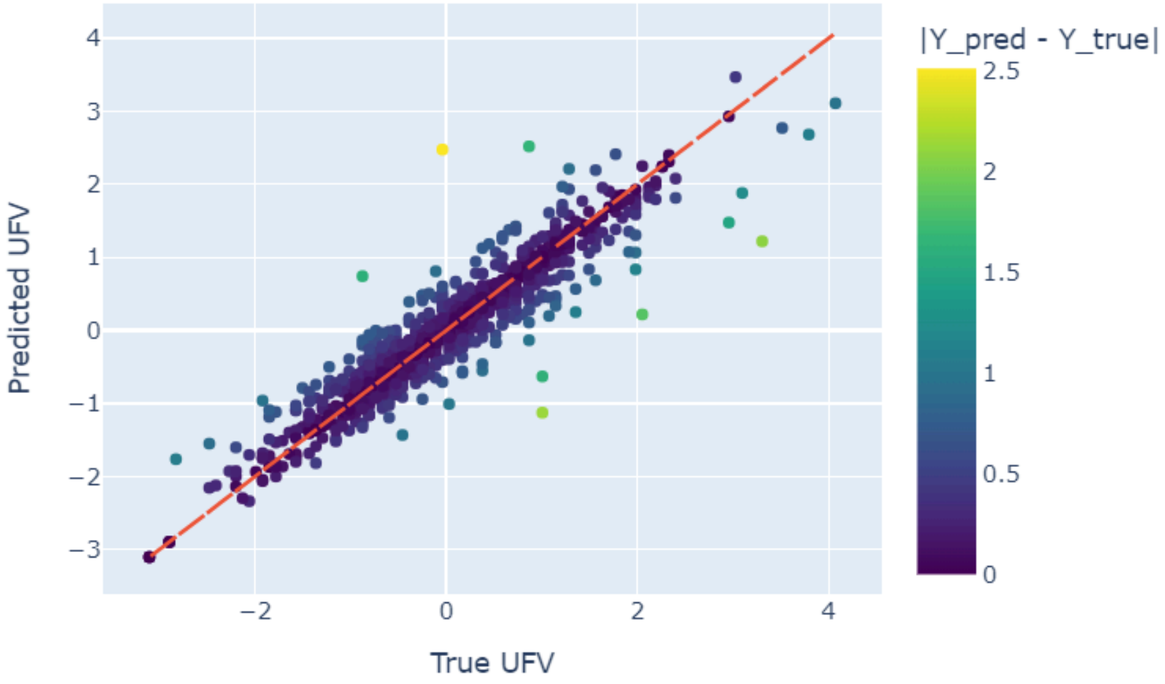


Plot des Résidus.
Ecart-type = 0.2842, Erreur Absolue Moyenne = 0.2591

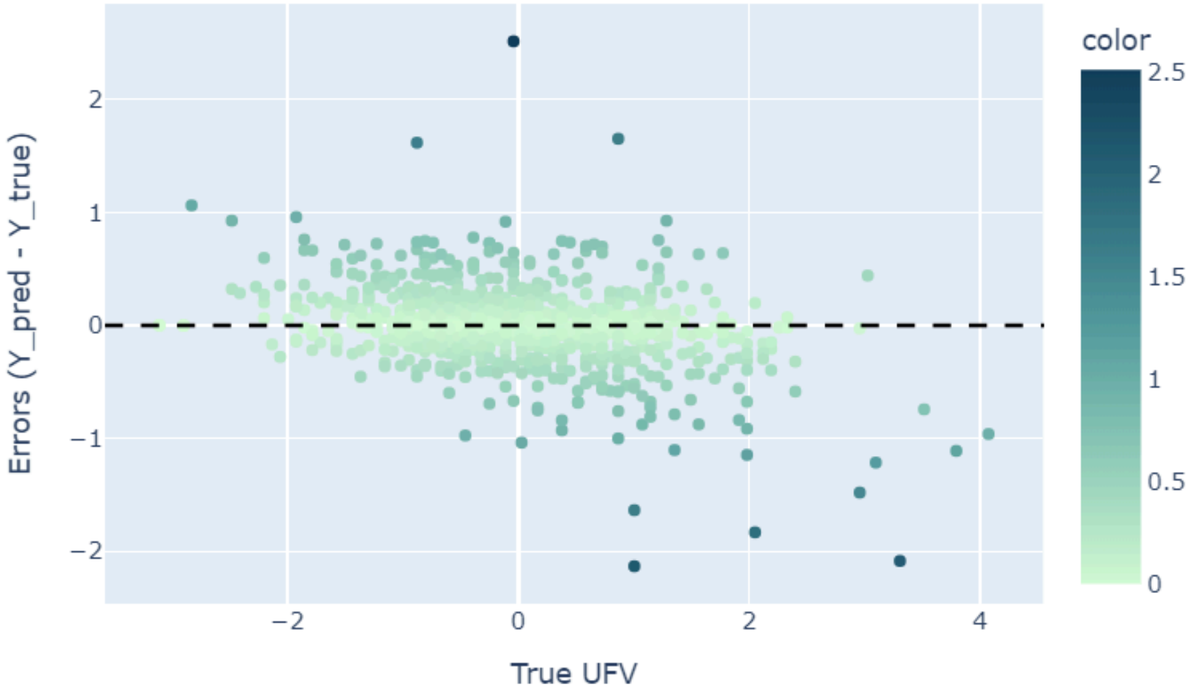


Meilleur modèle pour UFV

Meilleur Modèle pour prédire UFV: XGBoostRegressor
Dataset:MCA sur tous les Libellés + VC
MAE score: 0.2399, R² score: 0.8736

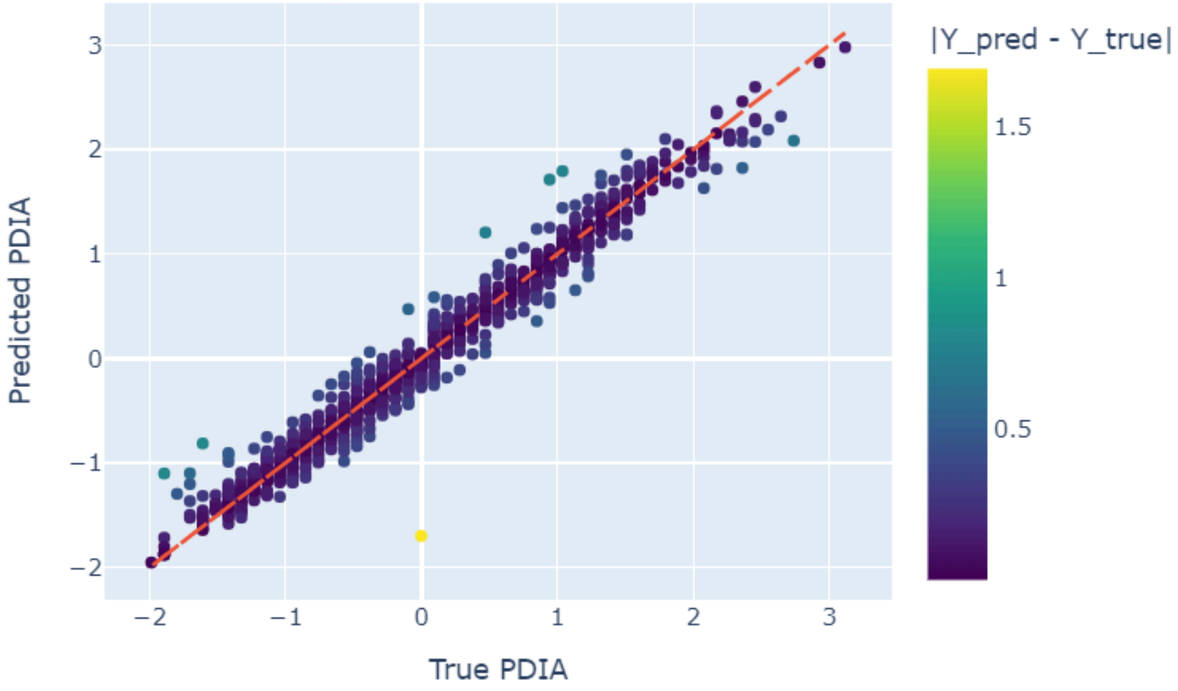


Plot des Résidus.
Ecart-type = 0.2624, Erreur Absolue Moyenne = 0.2399



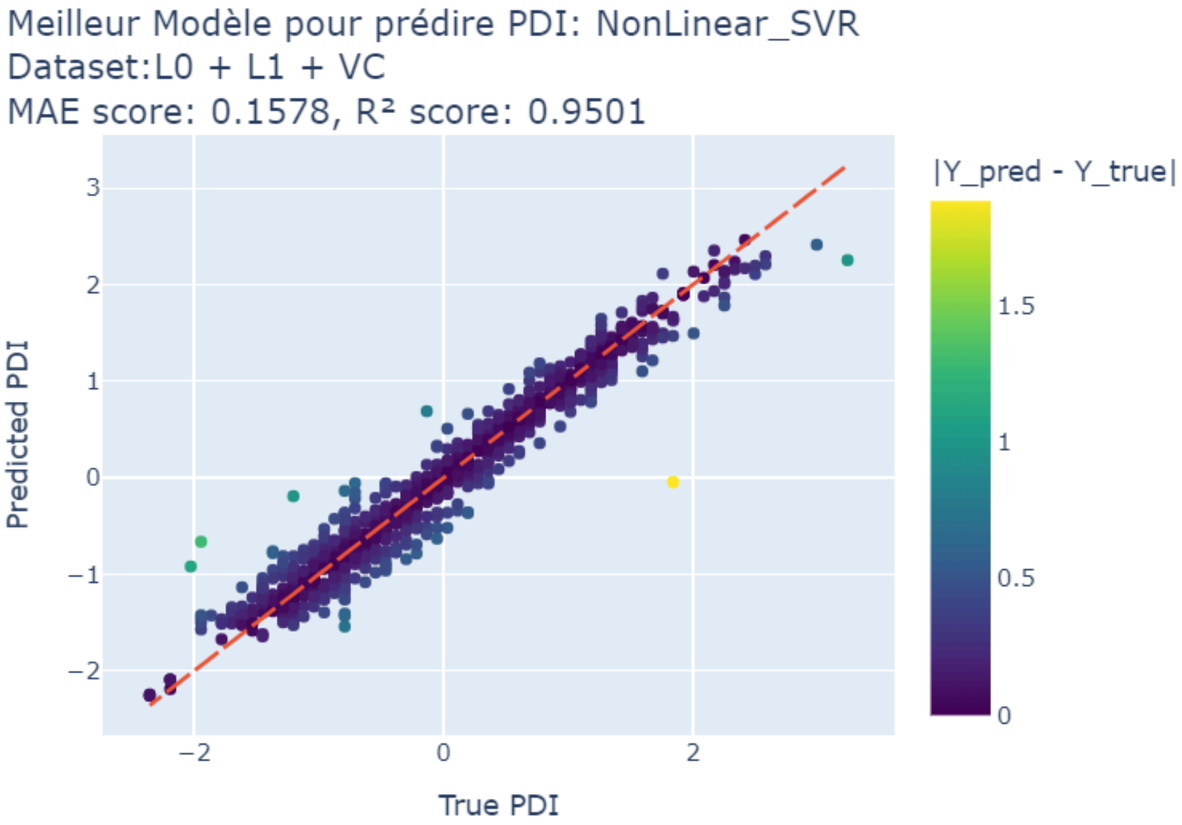
Meilleur modèle pour PDIA

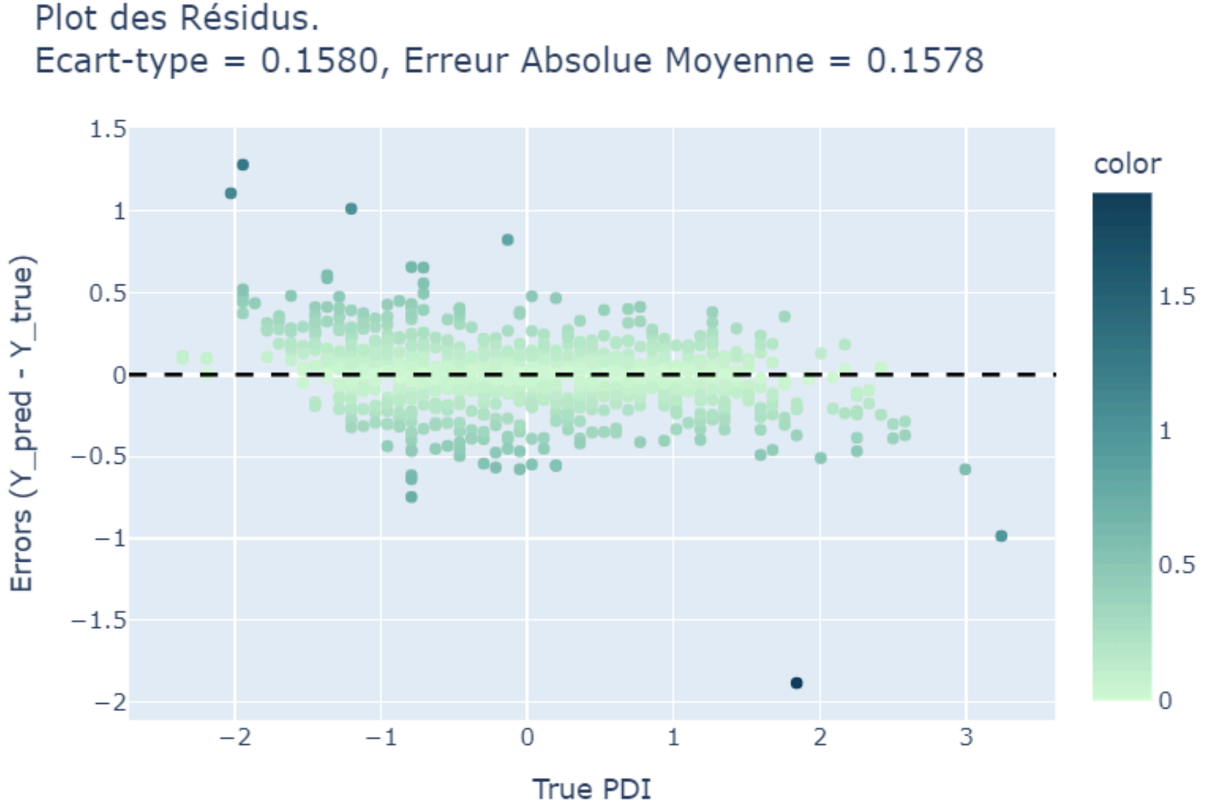
Meilleur Modèle pour prédire PDIA: MLPRegressor
Dataset:L0 + L1 + VC
MAE score: 0.1301, R² score: 0.9673



Plot des Résidus.
Ecart-type = 0.1256, Erreur Absolue Moyenne = 0.1301

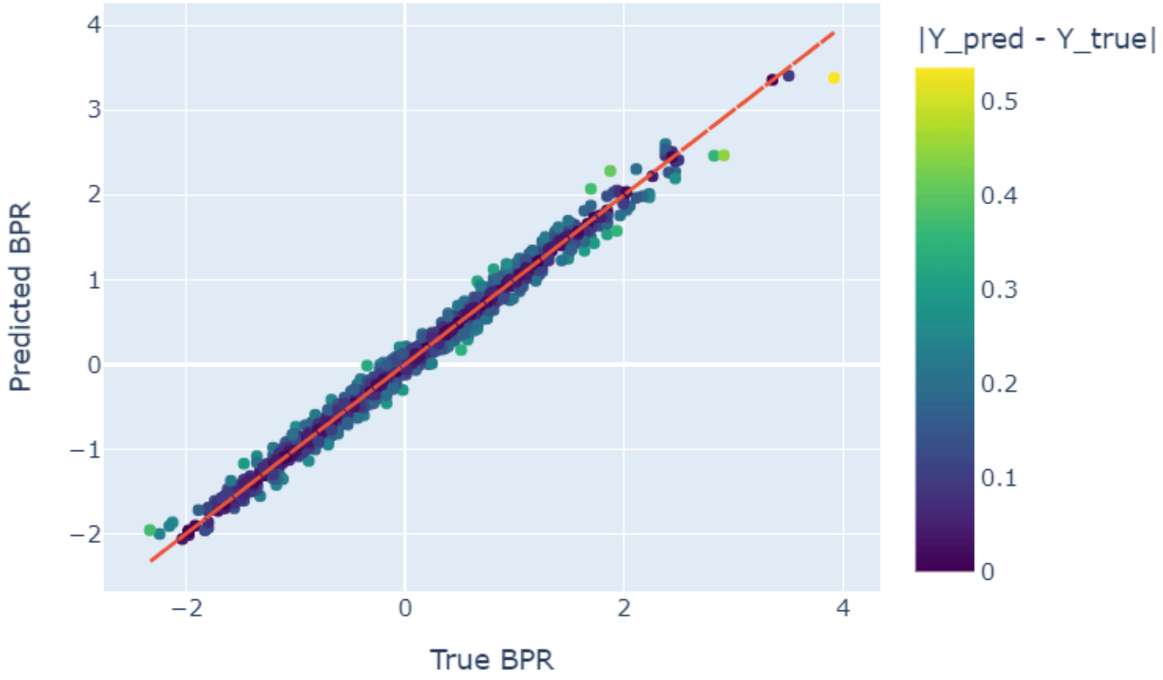
Meilleur modèle pour PDI





Meilleur modèle pour BPR

Meilleur Modèle pour prédire BPR: GradientBoostingRegressor
Dataset:L0 + L1 + VC
MAE score: 0.0872, R² score: 0.9870



Plot des Résidus.
Ecart-type = 0.0731, Erreur Absolue Moyenne = 0.0872