Université Paris Cité

Projet Kaggle

Parkinson's Disease Progression Prediction

DESVILLES Roxane, DUFOUR Laura, IANCU Nicoleta et NABI Appoline

24/10/2025

Introduction

La maladie de Parkinson est une pathologie neurodégénérative chronique, incurable.

Objectif du projet :

Prédire l'évolution clinique à travers 4 scores UPDRS d'un patient Parkinsonien à partir de dosages protéiques par spectrométrie de masse

- **2 méthodes d'imputation des données protéiques :**
  - imputation hiérarchique en 3 étapes
  - imputation par méthode unique : médiane globale des protéines

- **2 modèles de prédiction testés : MLP, RF**
  - Multi-Layer Perceptron (MLP) → capacité à modéliser les relations non linéaires complexes
  - Random Forest (RF) → modèle plus robuste face aux données bruitées

I. Formatage de 4 fichiers bruts CSV (Pandas) :

1. Concaténation des 2 fichiers avec données cliniques : **patient_id + visit_month + scores UPDRS + ON/OFF medication**

2. Merge des scores avec le fichier de données biologiques : **patient_id + visit_month + colonne UniProt + colonne NPX (dosage protéiques)**

3. Pivot du tableau : chaque protéine = en colonne

4. Tri par patient_id et visit_month

Dataframe final :

	patient_id	visit_month	code_Uniprot [1]	...	code_Uniprot [227]	updrs_1	...	updrs_4	on/off medication
chaque visite →									

II. Présentation des données Patients et Protéiques :

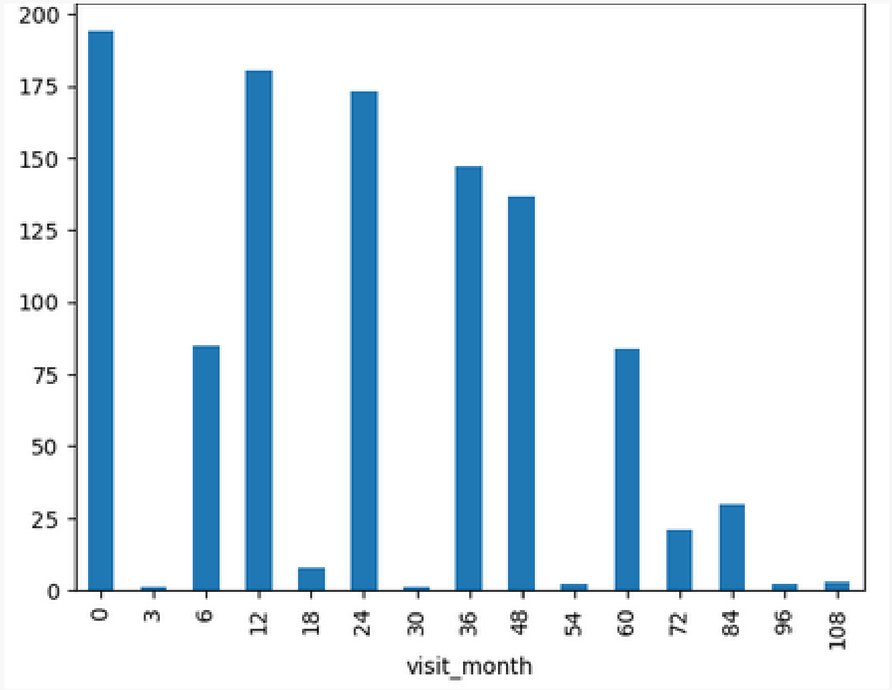


Figure 1 : Répartition des patients par mois de visite

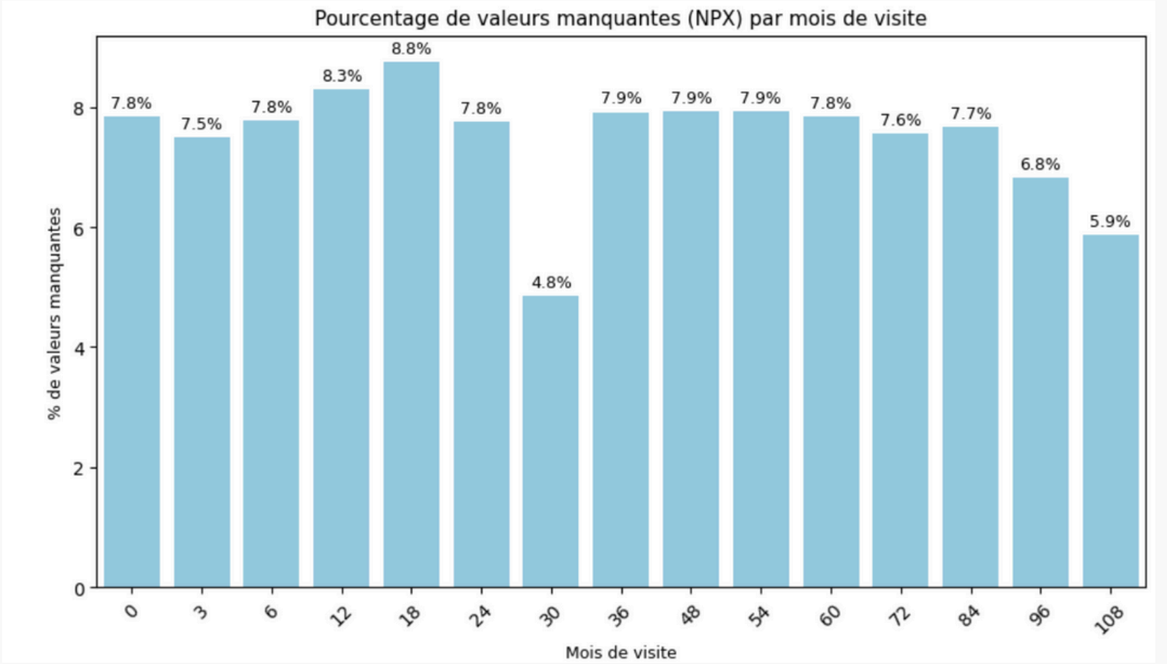
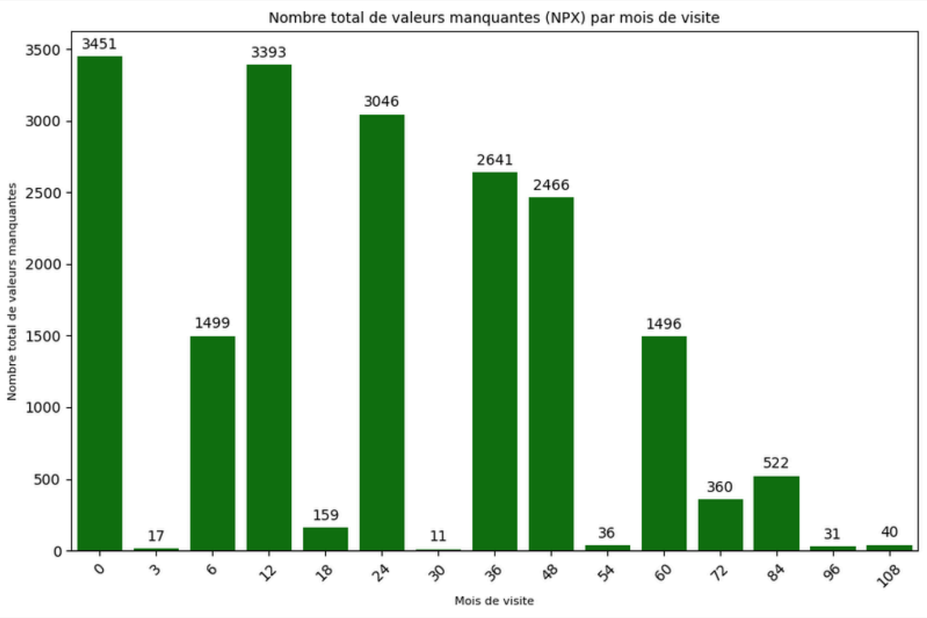


Figure 2 : Représentation du % de valeurs manquantes par mois de visite

- **248 patients** : répartition hétérogène **malades et témoins confondus**
- durée max du suivi : 108 mois de visite
- 4 visites en moyenne /patient

- **227 protéines dosées** (NPX)
- 4.8 à 8.8 % de NaN / mois de visite→ stabilité temporelle
- faible sous-ensemble fortement laculaire

III. Imputation des données protéiques manquantes :



NaN avant imputation : **19168**

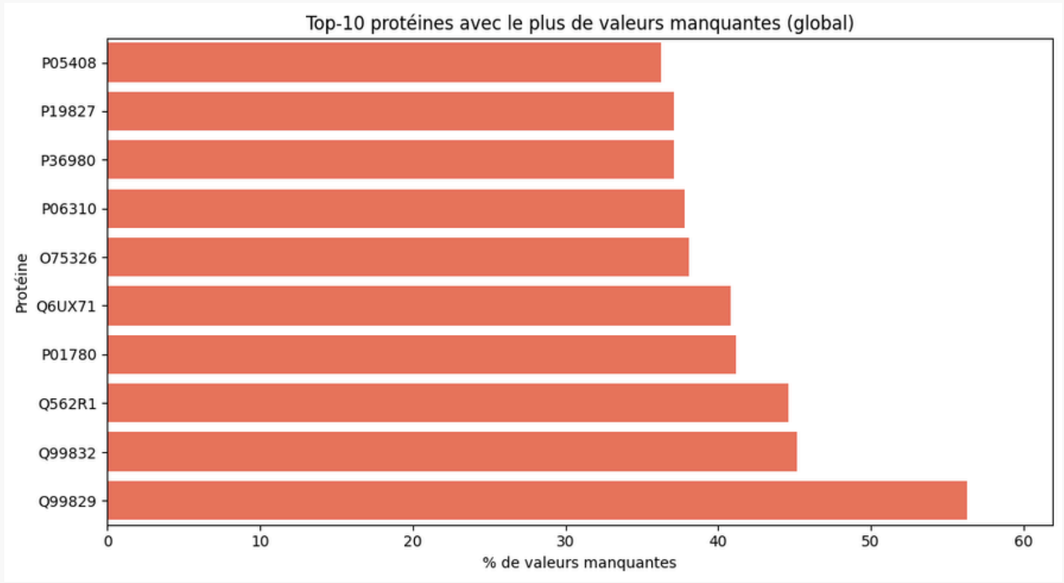


Figure Annexe : % valeurs manquantes - Top10 des protéines

2 méthodes d'imputation utilisées :

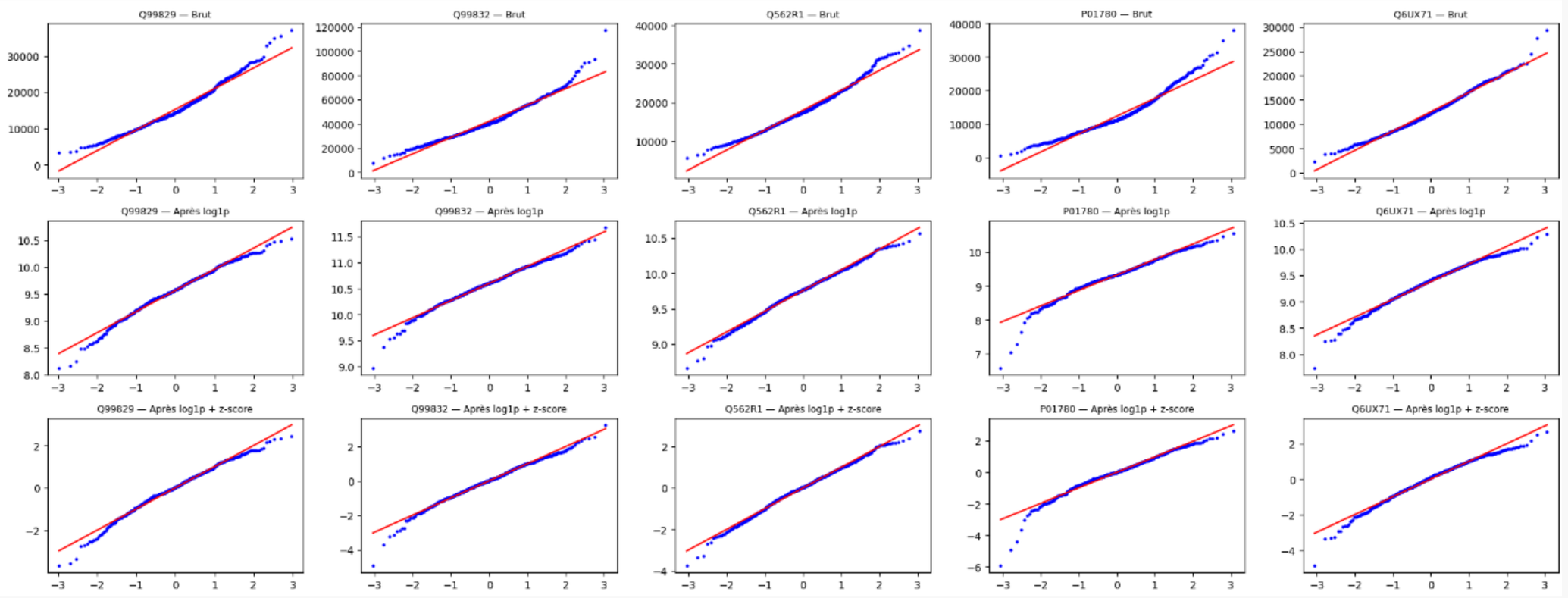
- **Imputation hiérarchique par protéine - 3 étapes : → conserver la dimension longitudinale des données**

1) imputation via une interpolation linéaire (intra-patient)
2) imputation par la valeur de dosage médiane calculée au même mois de visite (inter-patient)
3) imputation par la médiane globale protéine (inter-patient)

- **Imputation par la médiane globale des protéines** - étape unique : → **ne conserve pas la dimension longitudinale des données**

IV. Transformation logarithmique, normalisation/scaling

- QQplots avant et après imputation hiérarchique : top 5 des protéines aux valeurs manquantes



V. Analyse univariée

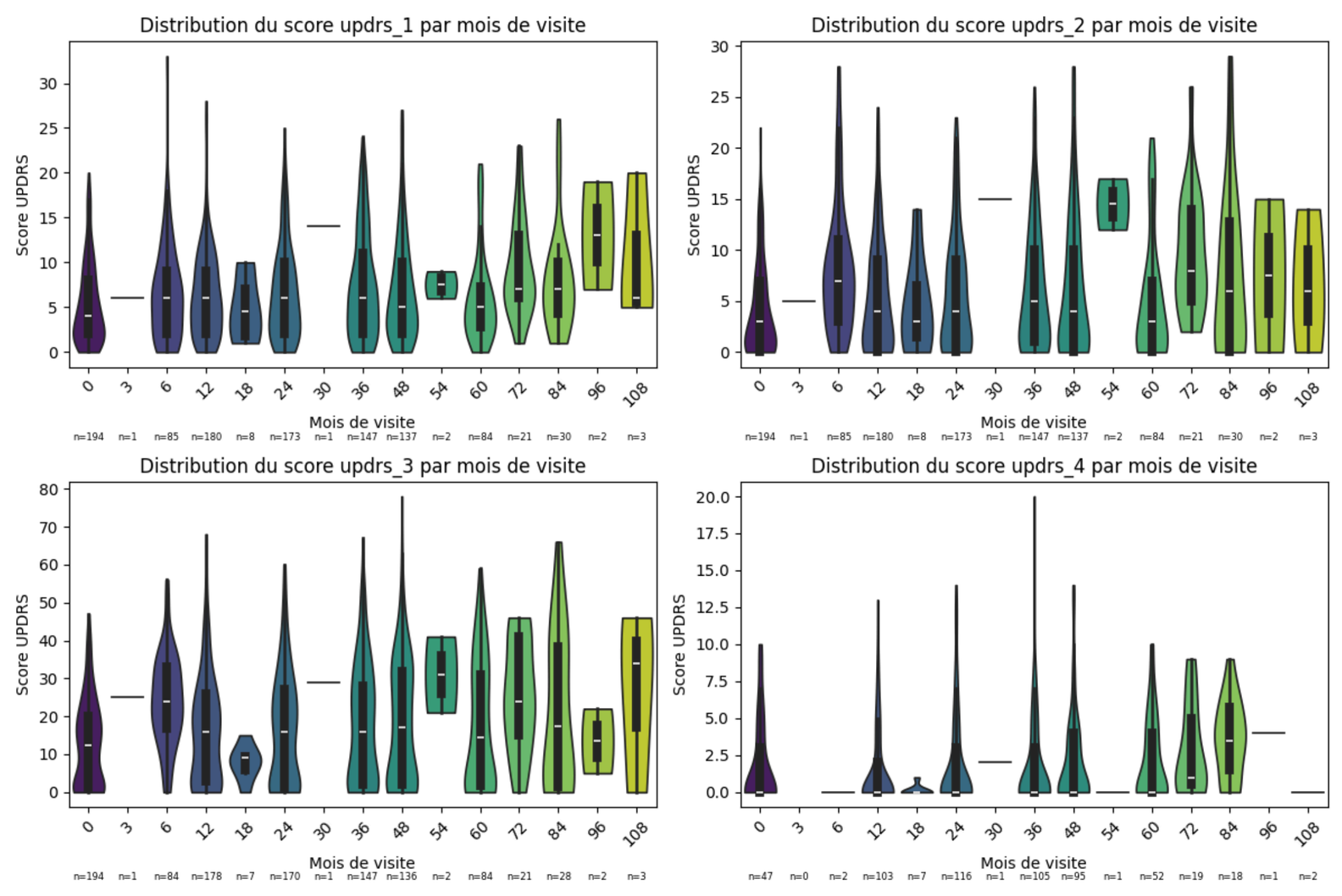
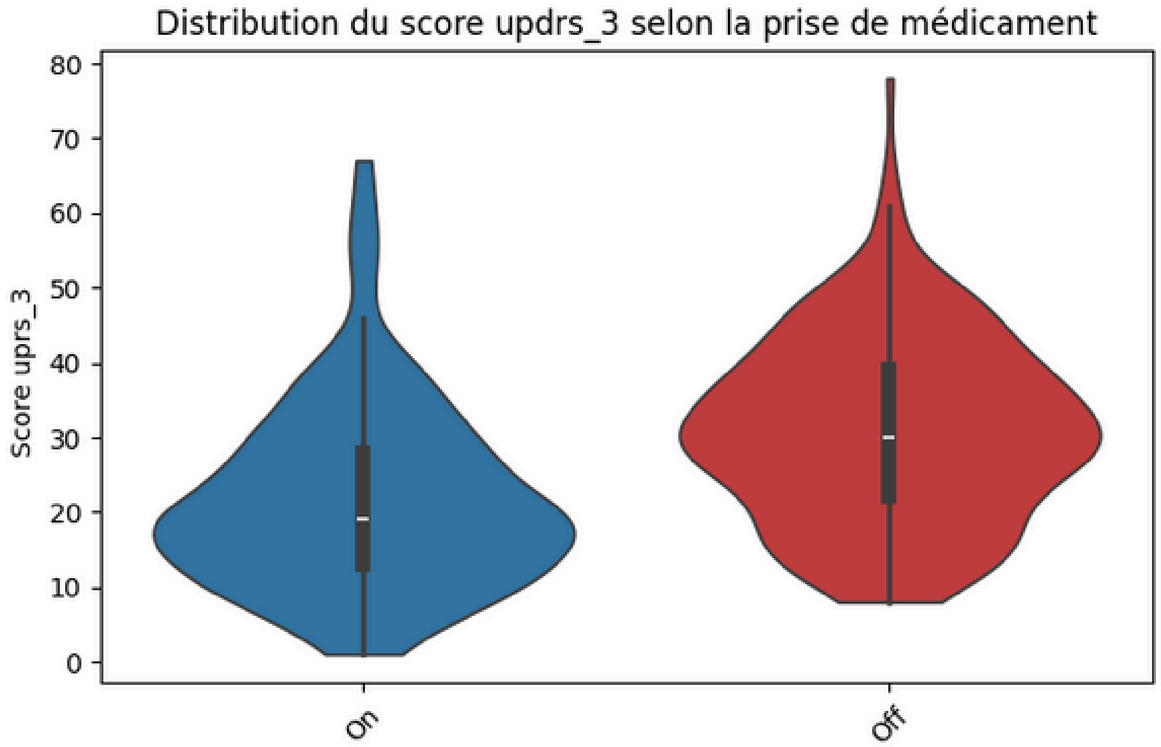


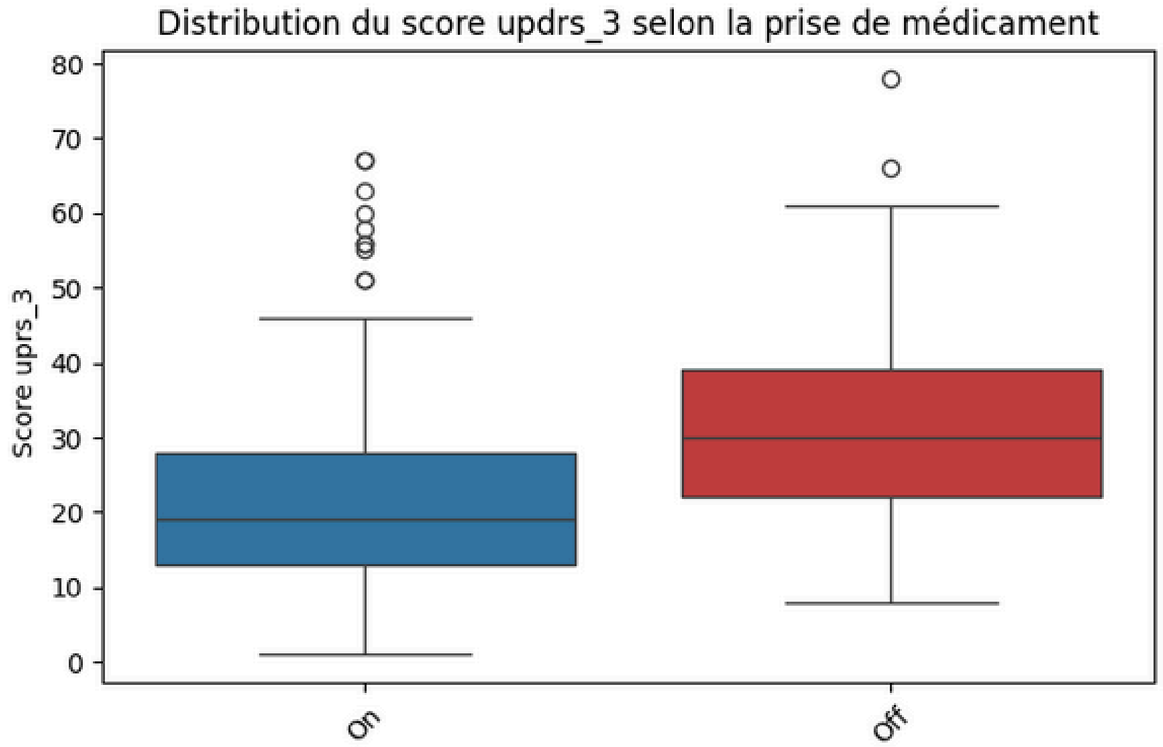
Figure 4 *: Violin plot des distributions des scores UPDRS, par mois de visite*

- *distribution homogène au cours du temps,*
  malgré un nombre élevé de NaN pour updrs_4

→ *absence de structure temporelle à priori*

V. Analyse univariée





- test non-paramétrique de Whitney : différence non-significative (0.44 >> 0.05)

V. Analyse univariée

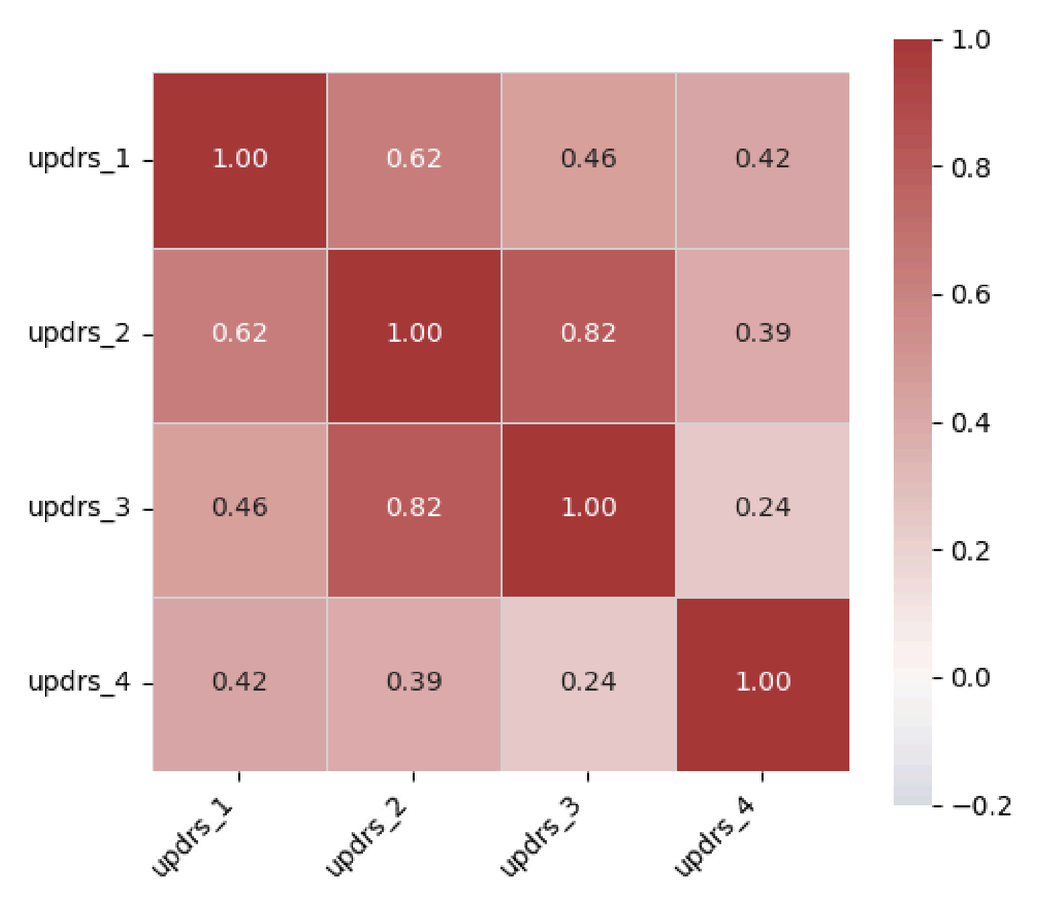


***Figure 6** : Matrice de corrélation de Spearman sur les scores UPDRS*

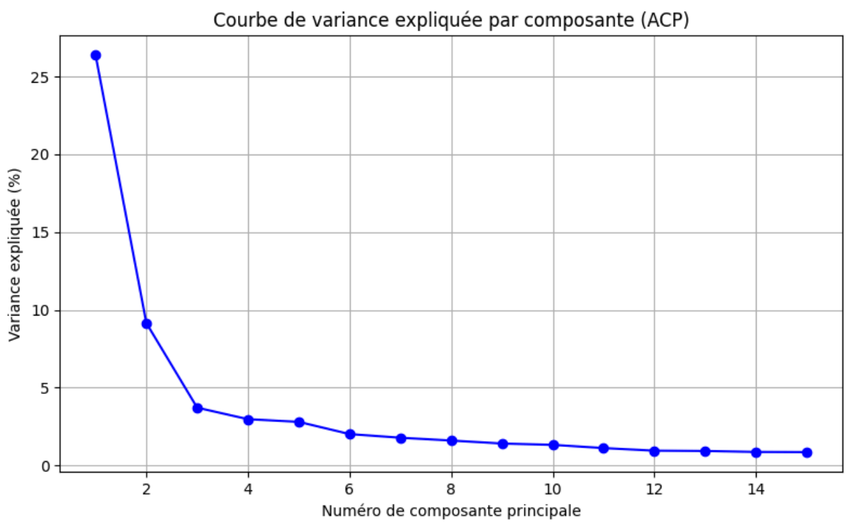
- forte corrélation entre updrs_2 (autonomie quotidienne) et updrs_3 (fonctions motrices)

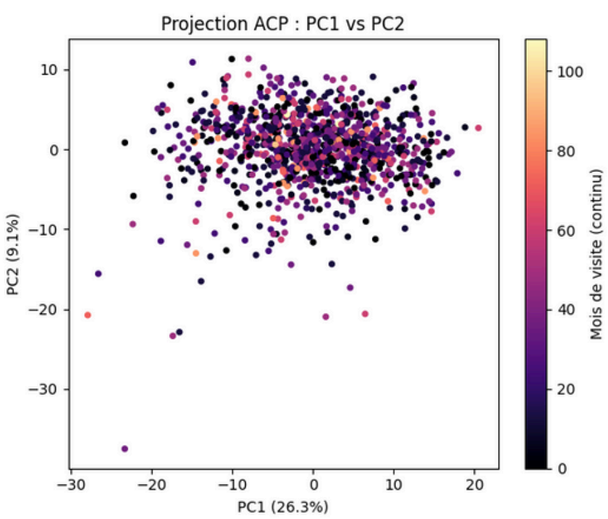
→ dégradation simultanée

VI. Analyse multivariée

Objectif d'identification par réduction de dimension :

- sous-groupes de patients
- atteints/ sains
- trajectoires temporelles sur 108 mois
- protéines les plus contributives

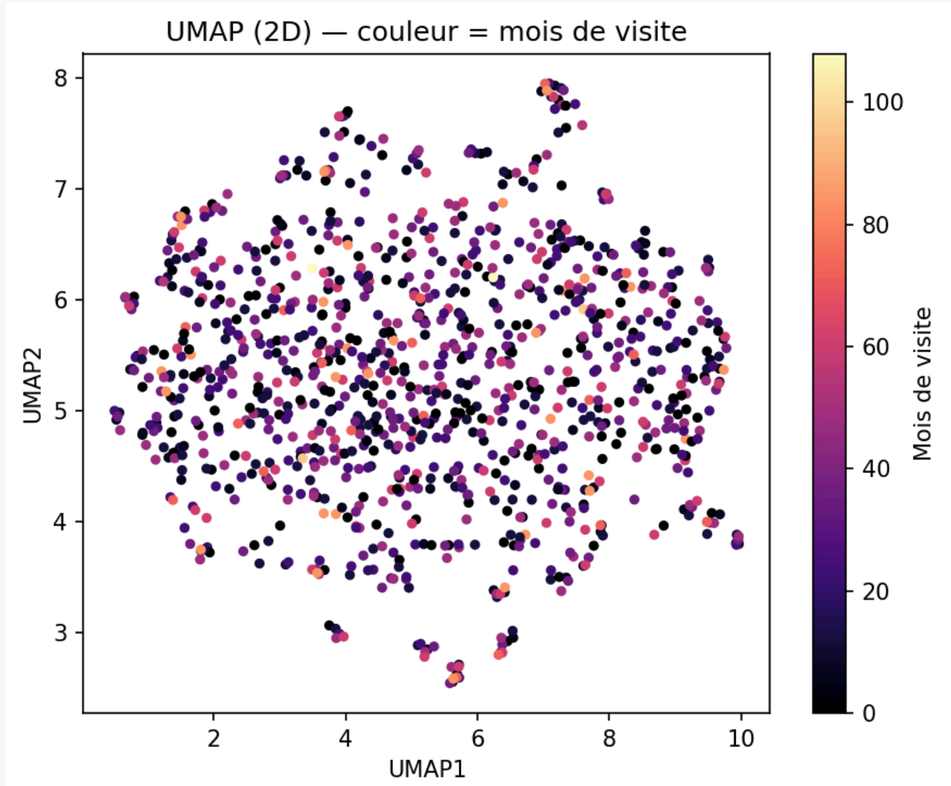




→ 6 composantes principales capturent 46.88% de la variance totale

→ homogénéité des profils

→ glycoprotéine CD59 (PC1) et Vitamine D (PC2) non spécifiques à Parkinson



→ absence d'organisation temporelle

VII. Modèles de prédictions

- **2 modèles :**

→ Multi Layer Perceptron (MLP) : relations non linéaires complexes
→ Random Forest (RF) : robuste aux données bruitées et petits effectifs

- **MLP :**

→ 2 couches cachées + ReLU + dropout 0.2 puis 0.3 + sortie linéaire 4 neurones
→ hyperparamètres : validation split 0.2, 100 époques, batch size 16, optimisation RMSProp (learning rate 0.001), perte MSE

- **RF :**

→ prédiction d'une valeur continue : régression
→ combinaison de plusieurs arbres de décision indépendants entraînés sur des sous-échantillons aléatoires
→ prédiction finale : moyenne des prédictions des arbres individuels
→ hyperparamètres : n_estimators = 100 (nombre d'arbres), max_depth = None (profondeur illimitée pour chaque arbre), random_state = 42 (pour la reproductibilité), les autres paramètres sont laissés par défaut

VIII. Résultats (1) - MLP

- Données utilisées : imputation sur la médiane globale des NPX, choix d'un target month unique = mois 24
- **Split train/test par patient : train → 0.8 et test → 0.2**
- Avec **epochs = 100** + **batch size = 16** → **validation loss autour de 57**, augmenter le nombre d'époques ou faire varier le batch ne change rien
- Métriques comparables à celles d'une baseline naïve :

UPDRS	MAE	RMSE
updrs_1	4.35	5.55
updrs_2	4.57	5.65
updrs_3	11.04	13.88
updrs_4	2.57	3.34

→ Le MLP n'apprend rien de significatif, probablement en raison du faible nombre d'exemples et du bruit élevé dans les données

Résultats (2) - RF sur données imputées à la médiane

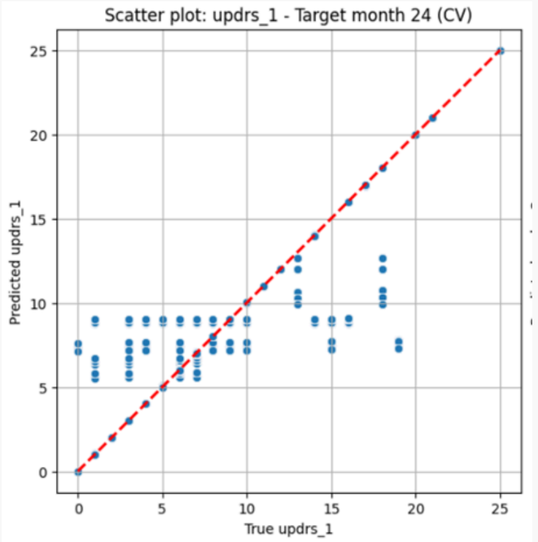
- Données utilisées : imputation sur la médiane globale des NPX, **choix d'un target month unique = mois 24**
- Split train/test par patient : train → 0.8 et test → 0.2

UPDRS	MAE	RMSE	R^2	SMAPE
updrs_1	4.44	5.82	0.10	61.35
updrs_2	3.32	4.42	0.16	59.40
updrs_3	11.99	14.83	-0.10	68.03
updrs_4	2.73	3.11	-0.13	159.56

→ Le modèle ne trouve pas de vrai lien prédictif entre NPX et UPDRS
→ Valeurs de SMAPE : prédictions raisonnables, mais pas discriminantes

- Données utilisées : imputation sur la médiane globale des NPX, choix d'un **target month unique = mois 24** (gestion de RAM)
- **5-fold cross validation**

UPDRS	MAE	RMSE	R^2	SMAPE
updrs_1	1.11	2.53	0.79	15.63
updrs_2	0.63	1.52	0.93	15.44
updrs_3	0.36	1.48	0.99	4.45
updrs_4	0.00	0.00	1.00	1.78

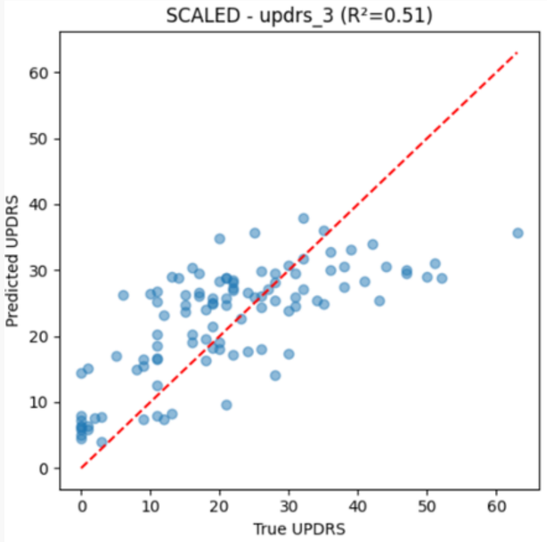


→ Data leakage

Résultats (3) - RF sur données interpolées et scalées

- Données utilisées : interpolation linéaire + scaling des NPX, toutes visites combinées
- Split train/test par patient : train → 0.8 et test → 0.2

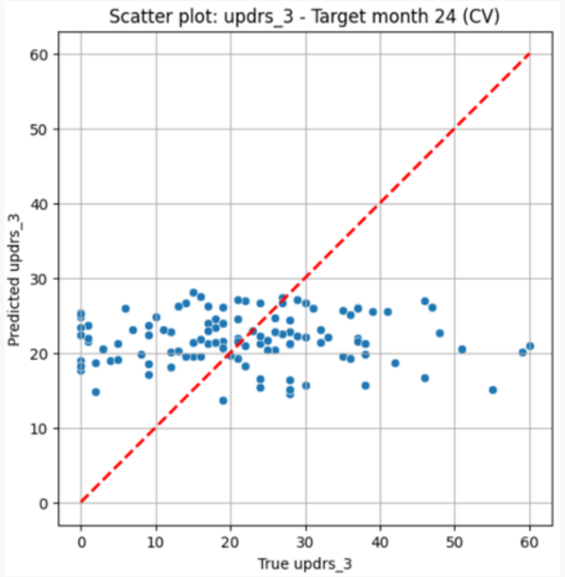
UPDRS	MAE_SCALED	RMSE_SCALED	R²_SCALED	SMAPE_SCALED
updrs_1	3.07	3.98	0.41	43.7
updrs_2	3.27	4.10	0.52	65.6
updrs_3	7.62	9.45	0.51	51.3
updrs_4	2.52	3.64	0.06	158.0



→ Métriques correctes
→ Coup de chance ?

- Données utilisées : interpolation linéaire + scaling des NPX, choix d'un target month unique = mois 24 (gestion de RAM)
- 5-fold cross validation

UPDRS	MAE_SCALED	RMSE_SCALED	R²_SCALED	SMAPE_SCALED
updrs_1	4.41	5.57	-0.016	60.74
updrs_2	4.31	5.41	0.052	66.09
updrs_3	11.27	14.03	-0.094	60.68
updrs_4	2.55	3.28	-0.048	156.88

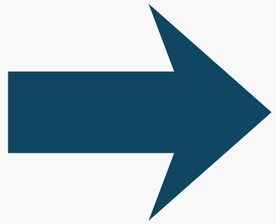


→ Le modèle ne généralise pas
→ Signal temporel très faible voire inexistant
→ Profils artificiellement lissés

IX. Discussion et Conclusion

Convergence des résultats: de l'exploration à la prédiction

- Exploration non supervisée (ACP, UMAP): absence de structure, pas de gradient temporel
- Modélisation supervisée (MLP, RF): incapacité à généraliser



Signal protéomique seul insuffisant

IX. Discussion et Conclusion

Contraintes des données et complexité biologique

Limitation du jeu de données

- Volume de valeurs manquantes
- absence de labels
- hétérogénéité de la cohorte
- déséquilibre temporel

Maladie de Parkinson

- hétérogénéité phénotypique, clinique et pathologique (*Seppi 2023, Berg 2021)*
- Profils protéomiques insuffisants → nécessité intégration données cliniques, génétiques, environnementales

IX. Discussion et Conclusion

Pistes d'amélioration

Notre approche	Approche gagnante
Longitudinale: prédiction des scores futurs	Transversale ("snapshot"): prédiction au même mois
Imputation globale/interpolation temporelle	Imputation locale par visite
Protéines seules	Protéines + peptides fusionnés
Adaptée aux données longitudinales de qualité	Adaptée aux données lacunaires

Adaptation essentielle de la méthode de modélisation au jeu de données