Projet ML

Parkinson Telemonitoring

**Clément Peillon**

Github : [ML\_Parkinson\_Telemonitoring](https://github.com/DataPyPeil/ML_Parkinson_Telemonitoring)

SIGMA Clermont

Mastère Spécialisé Expert en Science des Données

# Introduction

La maladie de Parkinson touchait en 2019 plus de 8.5 millions de personnes dans le monde selon l’OMS. Cette maladie se manifeste progressivement chez les individus et peut représenter un réel handicap. Le suivi des symptômes est onéreux pour les hôpitaux, nécessite une mobilisation importante de personnel soignant spécialisé et exige la présence du patient pour réaliser des tests souvent chronophages. Pour faire face à ces problèmes, des chercheurs de l’université d’Oxford (Tsanas Athanasios, 2009) ont proposé une méthode de suivi non-invasive basée sur l’analyse de tests de langage et ont mis en ligne une partie de la base de données anonimisée qu’ils ont constituée. Dans ce rapport, on se propose d’étudier différents algorithmes de machine learning pour tenter de prédire l’évolution de la maladie chez les patients.

# Présentation du dataset

Accessible sur le site de [UC Irvine Machine Learning Repository](https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring), ce dataset regroupe les données de 42 patients recrutés pour une campagne de test sur six mois afin de mesurer l’évolution de leur maladie de Parkinson. Ces tests présentent deux avantages majeurs. D’abord, ils sont non-invasifs donc le patient n’est pas affecté physiquement et sera moins enclin au stress. De plus, comme il s’agit d’enregistrement de la voix, ces tests peuvent être simplement réalisés à la maison sans avoir besoin de se déplacer dans un hôpital ou une clinique. Cela bénéficie à la fois au patient mais aussi au service hospitalier qui gagnent du temps pour traiter d’autres patients.

Les tests effectués environ tous les deux/trois jours sur la période ont permis de récolter 5875 données. Chaque test consiste en un enregistrement vocal duquel les chercheurs ont extrait 16 caractéristiques détaillées ci-dessous :

* **Jitter (%), Jitter (Abs), Jitter:RAP, Jitter:PPQ5, Jitter:DDP:** Mesures de la variabilité de la fréquence fondamentale (F0) de la voix, qui est souvent altérée chez les patients atteints de la maladie de Parkinson. Un *jitter* élevé indique une instabilité vocale.
* **Shimmer, Shimmer (dB), Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA:** Mesures de la variabilité de l'amplitude de la voix. Un *shimmer* élevé est associé à une voix plus rauque ou instable, un symptôme courant chez les patients parkinsoniens.
* **NHR, HNR**: rapport harmonique-bruit. Mesure de la proportion d'énergie harmonique par rapport au bruit dans le signal vocal. Un HNR plus faible indique une voix plus bruyante et moins harmonieuse.
* **RPDE, DFA, PPE**: mesures non linéaires de la dynamique vocale.

En plus de ces caractéristiques extraites de la piste audio, ont été relevés **l’âge** et le **sexe** du patient ainsi qu’un indicateur de l’avancement dans l’étude **time\_test**.

Enfin, pour conclure avec la présentation du dataset, il convient de présenter quelles sont les variables à prédire. D’abord, **motor\_UPDRS** est une évaluation de la fonction motrice des patient atteints de Parkinson. L'échelle UPDRS (Unified Parkinson's Disease Rating Scale) est utilisée pour évaluer la progression de la maladie. La partie motrice de cette échelle évalue des aspects tels que la tremblote, la rigidité, la bradykinésie, etc… La variable **totale\_UPDRS** représentela somme des scores de toutes les sections de l'échelle UPDRS, comprenant les aspects moteurs, non moteurs et les complications liées au traitement. Ce score global reflète la sévérité globale de la maladie. On comprend alors qu’il sera sûrement intéressant de n’étudier qu’une des deux variables car celles-ci sont fortement. On vérifiera cette hypothèse en mesurant leur corrélation.

# EDA – Exploration des données

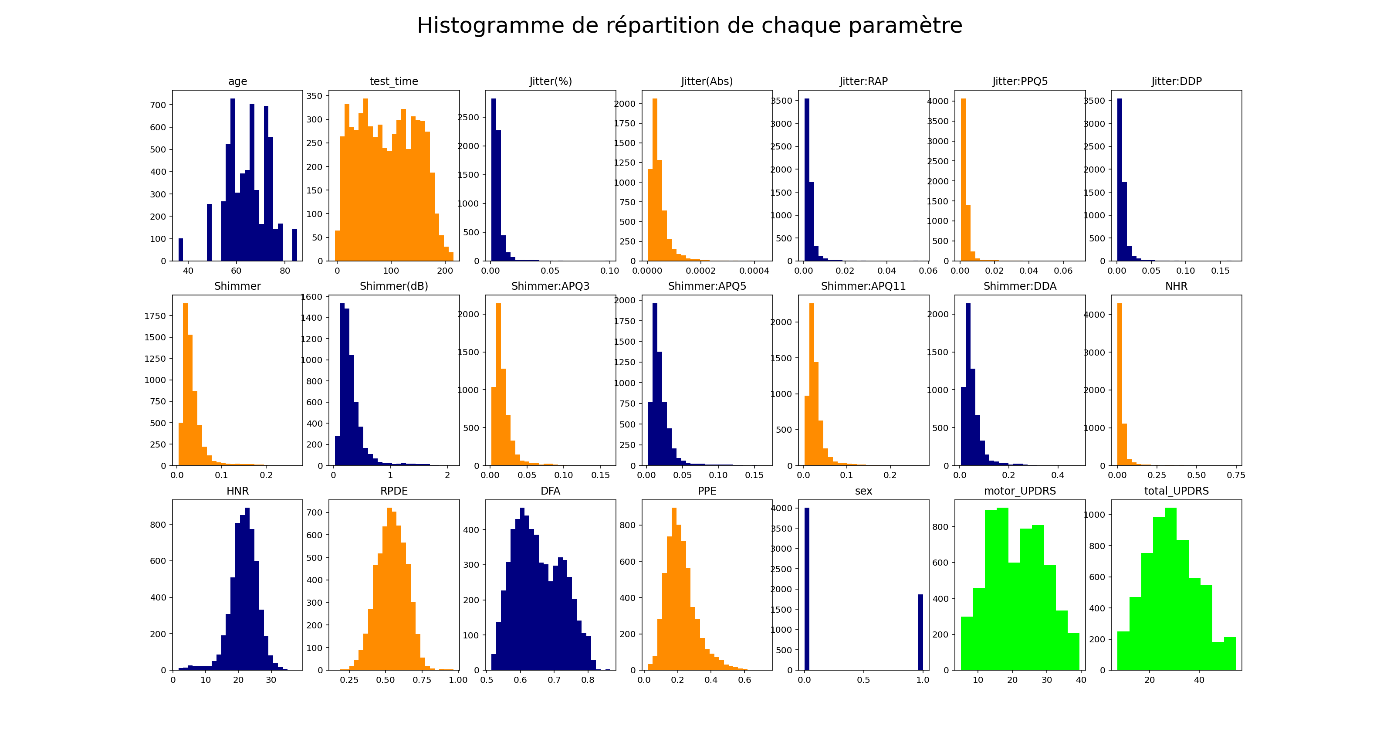


Figure 1: Distribution des variables

Tel que partagé par les chercheurs, le jeu de données nécessite quelques ajustements avant de pouvoir être utilisé pour l’ajustement d’un modèle de machine learning. En effet, les variables **Jitter** et **Shimmer** semblent être distribuées selon des lois exponentielles (*Figure 1*). Pour faciliter l’apprentissage, on ramènera ces variables à des distributions normales en utilisant la fonction logarithme népérien (*Figure 2*). Cette modification est nécessaire car certains modèles présupposent que les variables ou les résidus sont distribuées selon une loi gaussienne, c’est le cas pour la régression linéaire. Pour d’autres algorithmes, comme celui des k-plus proches voisins, avoir des variables normales permet d’utiliser la distance euclidienne qui a plus de sens dans un espace normalisé.

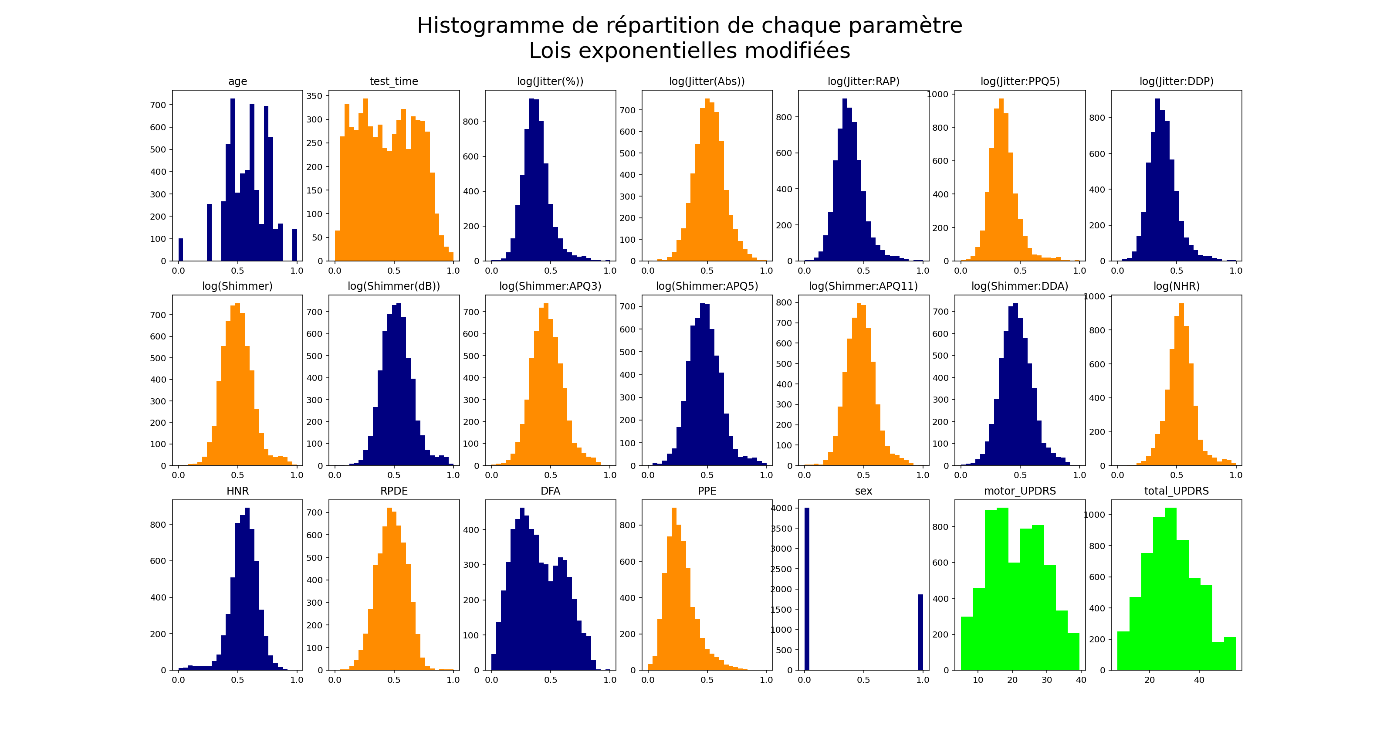


Figure 2: Distribution des variables après ajustement

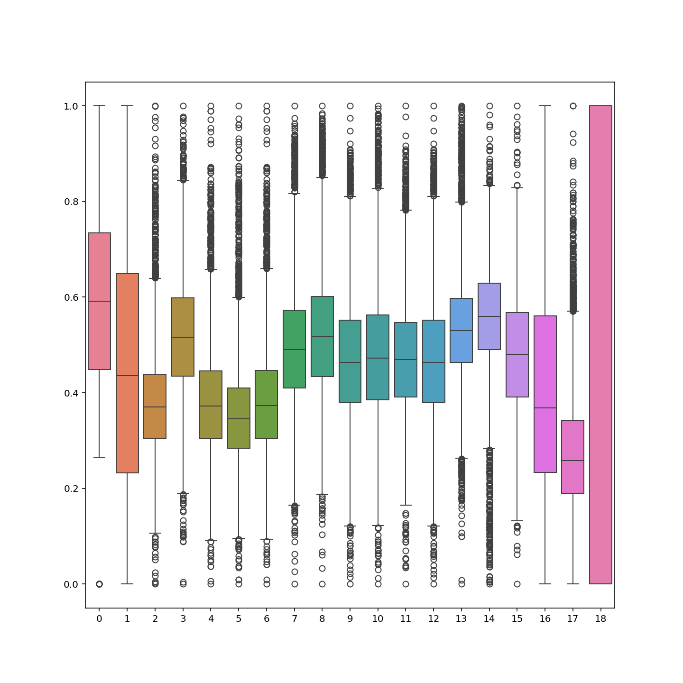


Figure 3: Boxplot pour détection des outliers

## Ensuite, il peut être intéressant de détecter des *outliers*, des valeurs aberrantes. La *Figure 3* permet de mettre en avant ces *outliers*. En revanche, tous ceux identifiés sur cette figure ne sont pas pertinent au regard des histogrammes en *Figure 2*. De ce fait, une seule valeur aberrante sera supprimée du dataset, il s’agit de l’ensemble des données associées à l’individu le plus jeune. D’un point de vue médical, ceci peut avoir du sens car cette maladie est généralement développée à un stade plus avancé de la vie.

Enfin, pour conclure l’exploration du dataset, on peut s’intéresser à la corrélation entre les variables explicatives et celles à prédire (*Figure 4*). Cette figure nous permet de confirmer l’hypothèse sur la forte corrélation entre les deux variables à prédire. Ainsi, on fait le choix, arbitraire, de ne conserver que la variable **total\_UPDRS**. Prédire les deux n’aurait que peu d’intérêt vis-à-vis du surcoût computationnel que ça représenterait.

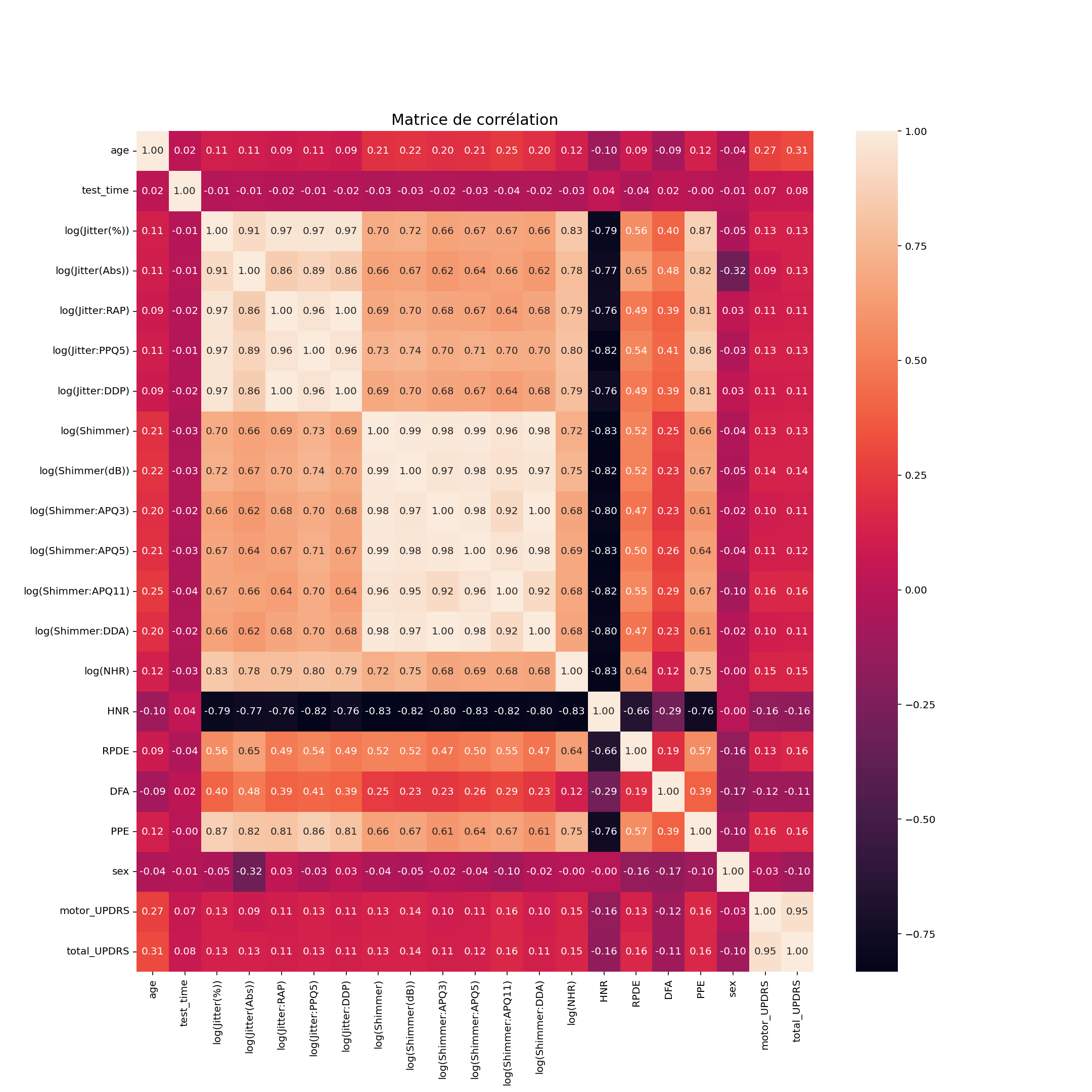


Figure 4: Corrélation des variables explicatives avec les variables à expliquer

# Modèle de prédiction

Afin de choisir le meilleur modèle pour cette tâche de régression, on va entrainer plusieurs modèles sur un jeu de données d’entrainement et les tester sur un autre jeu de données inconnues du point de vue du modèle.

## Méthodes linéaires

Dans cette partie, on cherche à trouver une fonction polynomiale en les variables explicatives pour prédire **total\_UPDRS**. La classe de fonctions recherchées couvre ici les polynômes de degré 1, 3 et 5.

### Régression linéaire / polynomiale

Dans cette partie, on cherche une relation purement linéaire. A chaque variable est associé un poids correspondant à l’influence de la variable sur la prédiction.

|  |  |  |
| --- | --- | --- |
|  |  |  |

Comme on pouvait le prédire en consultant la matrice de corrélation, les résultats observés ne sont pas probants pour le polynôme de degré 1. En revanche, les résultats sont meilleurs pour les polynômes de degré supérieurs.

### Régression Lasso

En intégrant une pénalité à la fonction coût avec la norme L1 sur les coefficients du modèle, on entraine un modèle de régression LASSO. En théorie, ce type de modèle permet de faire de la sélection de variables en réduisant à zéro celles qui n’ont pas de lien avec la prédiction. En revanche, dans le cas où les variables d’entrées sont corrélées entres elles comme dans notre cas (Annexe 1), ce modèle éprouve des difficultés. Les résultats, non satisfaisant, sont présentés en annexe.

### Régression Ridge

La régression Ridge quant à elle impose une pénalité avec la norme L2 des poids du modèle. Cette méthode s’est révélée être plus efficace que la régression lineéaire non-pénalisée.

## Méthode non-paramétrique

### K-plus-proches-voisins (KNN)

## Support Vector Machines (SVM)

### SVM noyau linéaire

### SVM noyau RBF

## Méthodes ensemblistes

### Random Forest

### Agrégation simple (moyenne arithmétique)

### Agrégation par stacking

### AdaBoost

### Gradient Boosting

# Meilleur modèle optimisé

# Conclusion

# Annexe

Annexe 1: Matrice de corrélation des variables