RYTHM

COHEN Sarah
sareden.cohen@gmail.com

DAHAN Ilan
ilandahan1@hotmail.fr

DESCHAMPS-PEUGEOT Théo
theo.dp@zerlo.fr

GHEZAIL Elisheva
g.elisheva@gmail.com

Contents

1 Introduction

Rythm est une jeune entreprise de neurotechnologie qui se fonde sur les neurosciences pour améliorer les performances de l'Homme, par la recherche, le suivi et la compréhension du cerveau humain. Leur produit Dreem est le premier objet connecté actif dédié au sommeil. Il surveille l'activité cérébrale par électroencéphalographie (EEG) et détecte ainsi les différentes phases de sommeil. Basé sur des années de recherche en neuroscience, cet objet a été élaboré pour ensuite transmettre des sons en synchronisation avec l'activité du cerveau afin d'améliorer la qualité du sommeil profond.

Le suivi de l'activité cérébrale présente un grand intérêt pour étudier le sommeil et les maladies du sommeil. Nous proposons de creuser la question de la dépendance entre l'âge et les caractéristiques du sommeil à travers un problème de régression sur l'âge d'un sujet.

Nous devons donc prédire l'âge du sujet en fonction de son activité cérébrale pendant le sommeil profond (EEG) et de la séquence de ses stades de sommeil sur une nuit (Hypnogramme).

2 Description des données

Nous cherchons à prédire l'âge d'un sujet à partir de 5 minutes d'un enregistrement EEG (75 000 données) pendant son sommeil profond et de l'hypnogramme associé à la nuit en question. Des règles d'identification qui distinguent les différentes phases de sommeil ont été élaborées (voir tableau ci-dessous). L'échantillon d'entraînement a 581 individus. L'échantillon de test en a 249 .

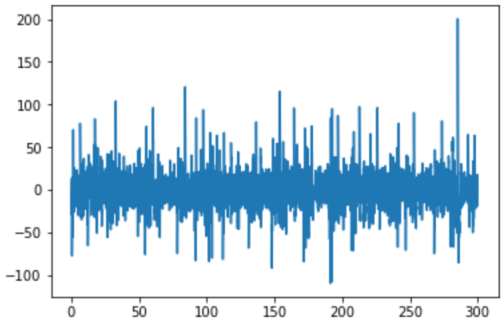

Figure 1: *Signal EEG de l'ID 1*

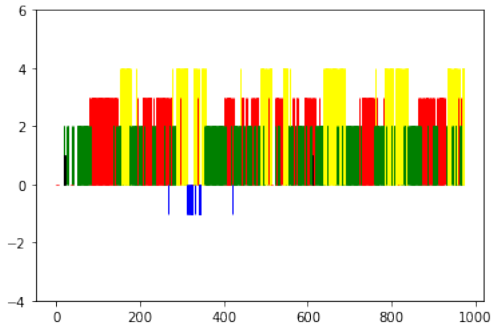

Figure 2: *Hyppnogramme de l'ID 1*

Nous avons entrepris des recherches sur les différentes variables proposées afin d'en comprendre leur signification, leur utilité et leur importance. Voici un tableau qui récapitule l'ensemble de nos variables:

ID	Identifiant de l'échantillon. Les échantillons sont enregistrés sur différent sujets ce qui peut induire un certaine variabilité dans les données (notamment les signaux EEG)
DEVICE	Identifiant du dispositif utilisé pour enregistrer le signal EEG. Les signaux EEG proviennent de deux dispositifs d'enregistrement (0 ou 1). Les signaux sont enregistrés à partir des mêmes cannaux EEG mais il peut exister un changement de gain (de l'amplificateur) en train les deux dispositif.
EEG_0 … EEG_74999	Ces entrées correspondent à 5 minutes de signal EEG, enregistré pendant le sommeil profond (dérivation Fp2 - A2), échantillonné à 250Hz.
HYPNOGRAM	Correspond à l'hypnogramme associé à l'échantillon en question. Il s'agit d'une liste de taille variable correspondant aux différents stades de sommeil. Les valeurs sont: 0 (pour le stade Wake), 1 (pour N1), 2 (N2), 3 (Sommeil profond - N3), 4 (REM), -1 (quand le scoring n'était pas possible).

3 Travail sur les données: features

3.1 Features sur l'hypnogramme

1. **Pourcentages des types (type={-1,0,1,2,3,4})**

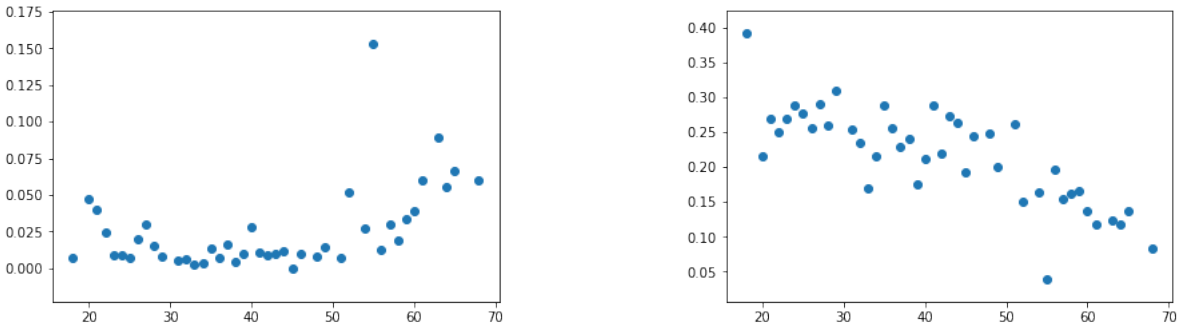

Figure 3: *Pourcentages moyens par âge de sommeil de type 1 (gauche) et de type 3 (droite)*

Nous pouvons observer à travers ces représentations graphiques que le pourcentage moyen de sommeil de type 1 a tendance à croître selon l'âge tandis que celui de type 3 tend à décroître. Ceci confirme une premiere idée qui est que le sommeil léger s'intensifie avec l'âge à l'inverse du sommeil profond.

2. **Rapport de phases**

La littérature nous apprend que le rapport entre le nombre de phases 1 et 3 est discriminant: ce rapport croît avec l'âge. Nous extrayons donc les 15 rapports de phases possibles.

3. **Matrice de transitions des types**

A l'image des matrices de transition des chaines de Markov, nous calculons la matrice de transition de l'hypnogramme.

4. **Violence du sommeil**

Nous souhaitons exploiter le sommeil "hachuré" des personnes plus âgées.

Nous caractérisons la violence, comprise entre 0 et 5, du passage d'une phase à une autre comme la valeur absolue de la différence des phases. Un passage d'une phase 1 à 4 a une "violence" de 3. Nous additionons le nombre de sauts par violence (les passages de 1 à 4 avec les passages de 0 à 3...) que nous exprimons en pourcentage du total de phase.

5. **Hypnogramme par bloc**

La réalisation de l'hypnogramme se fait en mesurant la fréquence de l'électroencéphalogramme, ainsi une phase 3 correspond à une fréquence des ondes inférieure à 3,5 Hz mais un point de mesure à 3,51 ferait apparaitre un passage en phase 2 au milieu d'une phase 3. Nous avons donc construit un hypnogramme "par bloc" permettant d'extraire des informations sur la durée des phases et leur nombre.

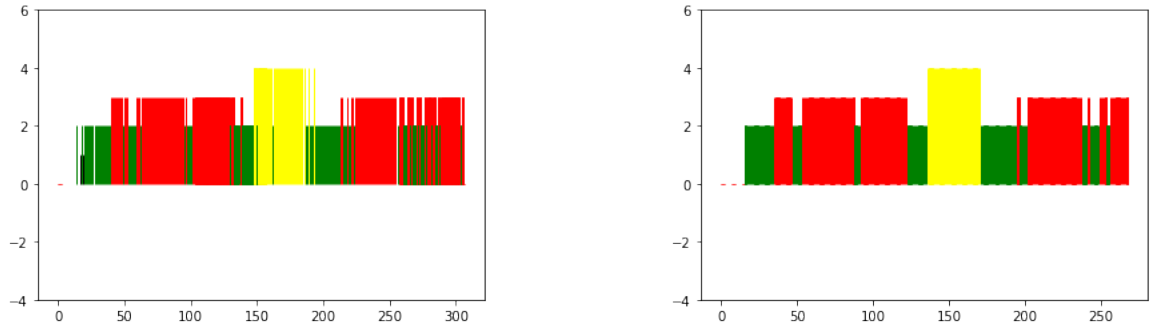

Figure 4: *Hypnogramme et hypnogramme par bloc d'un individu*

6. **Nombre de phases par type et nombre de phases par hypnogramme**

   En se basant sur l'hypnogramme par bloc, nous récupérons le nombre de phases de chaque type que nous exprimons en pourcentage du nombre de phases total de l'hypnogramme. Nous gardons aussi ce total comme feature pour décrire un individu.

7. **Longueur des phases pour chaque type et longueur moyennes des phases par type**

   Nous calculons la longueur de chacune des phases de l'hypnogramme par bloc de l'individu que nous classons par type de phase. Nous utilisons comme feature la longueur moyenne par type.

8. **Les k phases dominantes (k fixé)**

   Nous récupérons la nature des k phases les plus longues que nous avons extrait précédemment (Longueur des phases pour chaque type). En pratique nous avons gardé la nature des 5 phases dominantes.

9. **Temps d'endormissement**

   Ce feature a été extrait des hypnogrammes par bloc. Afin de déterminer les profils aberrants, nous déterminons le temps d'endormissement en additionnant les phases de 0 successives au début de l'hypnogramme que nous divisons par la longueur totale de l'hypnogramme. La plupart des individus s'endorment quasiment immédiatement:

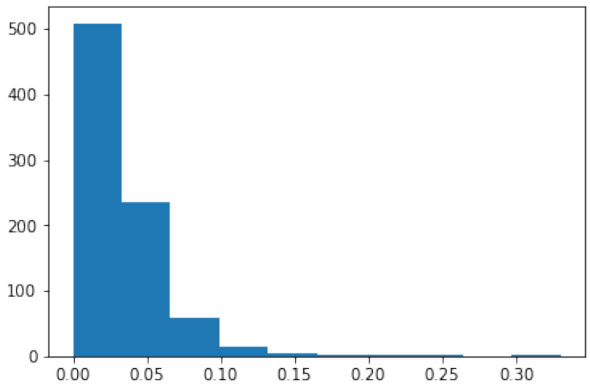

Figure 5: *Distribution du temps d'endormissement*

10. **Durée du sommeil (court ou long dormeur, seuil fixé)**

    La longueur de l'hypnogramme n'est pas discriminante: la mesure de l'hypnogramme peut avoir été effectuée pendant une "bonne" ou une "mauvaise" nuit. Cependant une mauvaise nuit correspond à de mauvaises informations sur l'individu: au cours d'une mauvaise nuit un jeune va avoir plus de phases 1 et moins de phases 3 qu'en temps normal. Nous construisons donc un feature qualitatif correspondant à une courte ou longue nuit pour signaler les individus ayant des attributs aberrants.

3.2 Features sur les EEG

1. **Visualisation du signal EEG**

   L'électroencéphalogramme est un tracé représentant l'activité électrique du cerveau.

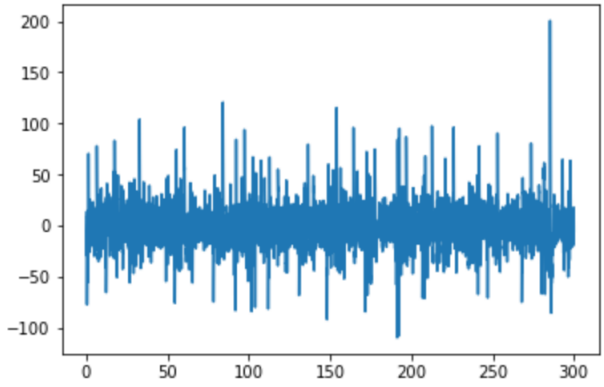

Figure 6: *Signal EEG de l'ID n1*

L'utilisation brutale des EEGs n'étant pas pertinente, nous avons transformé nos 75 000 variables en données utilisables.

2. **Calcul de la moyenne et de la variance des EEGs**

   Suite à nos recherches, nous avons débuté par une approche statistique. En effet, nous avons calculé la moyenne, la variance ainsi que la fréquence des EEGs.

3. **Transformée de Fourrier rapide, moyenne, variance, skew et kurtosis**

   Ensuite, nous avons appliqué la transformée de Fourier rapide (FFT) aux EEGs permettant de déterminer les principales bandes de fréquences. Nous avons effectué une étude statistique de ces transformées de Fourier, à savoir : la moyenne, la variance, le skew et le kurtosis.

4. **Puissance spectrale des Fourrier**

   Nous avons également ressorti la densité de la puissance spectrale des FFT, ce qui nous a permis d'observer les variations fréquentielles en fonction du temps. Nous avons en particulier choisi la méthode de Welch qui fournit un estimateur de cette densité.

5. **Transformation en ondelettes, coefficients, moyenne, variance, skew et kurtosis**

   Par ailleurs, nous avons déterminé les ondelettes des EEGs. Nous avons obtenu deux types de coefficients : les coefficients d'approximation et ceux de détails. Pour chacun des coefficients, nous avons calculé la moyenne, la variance, le skew et le kurtosis.

6. **Filtrage des EEGs par FiltFilt**

   Enfin, nous avons filtré les EEGs par un filtre forward-backward et recréé toutes les variables décrites précédemment sur les EEGs filtrés.

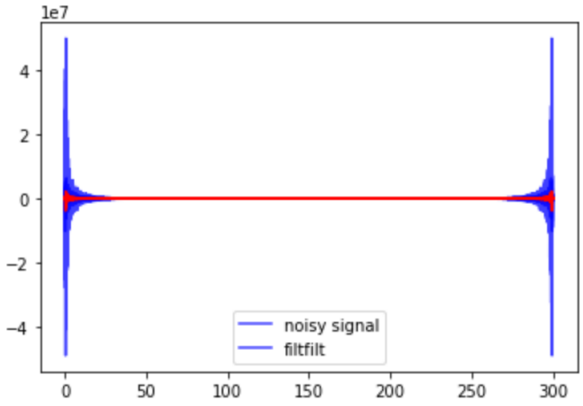

Figure 7: *Fourrier Filtré*

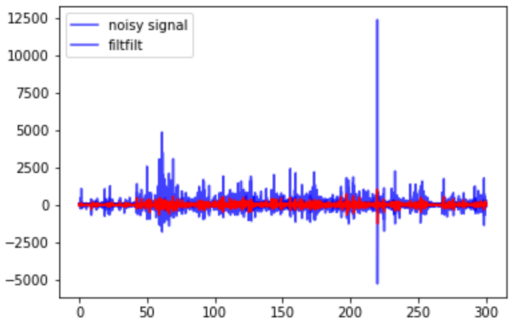

Figure 8: *Ondelettes: Coeff Details Filtrés*

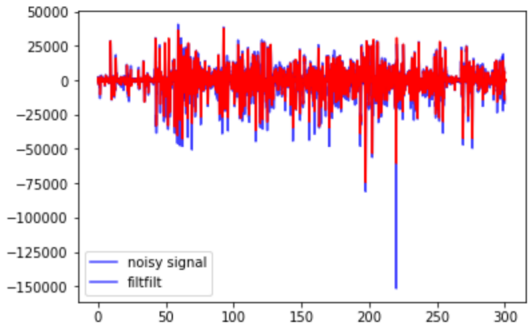

Figure 9: *Ondelettes: Coeff App Filtrés*

Toutes ces nouvelles features ont permis de perfectionner notre modèle, nous faisant gagner en précision.

4 Transformation sur les données : Valeurs aberrantes

Après exploration des données et création des features, nous nous sommes aperçus que pour un âge donné, certains individus se distinguent. Ces derniers peuvent présenter des complications à la prédiction et nous avons décidé de les traiter comme des cas extraordinaires, ou plutôt comme des valeurs aberrantes. Ainsi, nous les avons repérés puis écartés de notre échantillons. De plus, pour conserver un échantillon d'individus assez important, nous avons décidé d'appliquer ce raisonnemment seulement aux individus dont l'âge se répète plus de 15 fois. En effet, nous avons considéré que pour un âge présent plus de 15 fois, il était possible de répérer plus aisément et plus précisément les individus ayant des features particuliers.

1. **Liste des âges par indice**

   Nous avons tout d'abord créer une liste de 51 éléments, soit le nombre d'âges différents que nous avons dans l'échantillon. A chaque élément (correspondant ainsi à un âge), nous avons créé la liste des individus ayant cet âge. Chaque individu est caractérisé dans cette liste par son indice. Ceci nous a permis d'agir directement sur les individus dont l'âge se répéter plus de 15 fois.

2. **Boites à moustache : Région de confiance à 95%**

   Afin de supprimer les valeurs aberrantes, nous avons procédé de la façon suivante. Pour les individus d'un même âge, nous avons déterminé la moyenne pour chaque feature. Ensuite, nous avons calculé, par individu, la distance (différence en valeur absolue) de chaque feature par rapport à sa moyenne. Ceci nous a permis de répérer les individus dont les features s'éloignent le plus de leurs moyennes. Pour chaque individu, nous avons attribué un "0" ou un "1" pour chaque feature selon que sa distance à la moyenne est dans les 95% plus proche ou non.

3. **Suppression des valeurs aberrantes (seuil à 0.15)**

   Par le suite, nous avons déterminé le pourcentage de "1" par individu, c'est-à-dire le nombre de fois que la valeur d'un feature est trop éloignée de sa moyenne. Ainsi, chaque individu ayant un pourcentage de "1" supérieur à 15% a été enlevé de notre échantillon.

5 Implémentation de modèles et techniques d'apprentissage

5.1 Définition de la métrique

Notre métrique est celle de la valeur absolue de la moyenne en pourcentage d'erreur (MAPE) et est définie comme suit :

$$\text{MAPE} = \sum_{i=0}^{n} \frac{|y_i - f(x_i)|}{y_i}$$

avec n, le nombre d'individus dont l'âge est à prédire, y_i, l'âge réel de l'individu i et $f(x_i)$, l'âge prédit à partir des features x_i de l'individu i, pour $i \in \{0, ..., n\}$.

5.2 Random Forest

Les forêts d'arbres décisionnels ont été formellement proposées en 2001 par Leo Breiman et Adèle Cutler. Elles font partie des techniques d'apprentissage automatique. Cet algorithme combine les concepts de sous-espaces aléatoires et de « bagging ». L'algorithme des forêts d'arbres décisionnels effectue un apprentissage sur de multiples arbres de décision entraînés sur des sous-ensembles de données légèrement différents.

Le random forest construit donc plusieurs arbres de décision en parallèle de la manière suivante :

Random forest = tree bagging + feature sampling

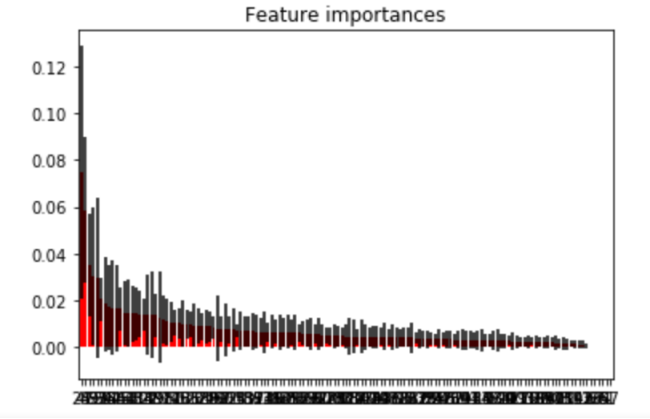

Figure 10: *Features Importance Random Forest*

Le score obtenu par le Random Forest des **21.02**

5.3 Gradient Boosting

Alors que le random forest construit plusieurs arbres en parallèle, le Gradient Boosting construit lui aussi k arbres mais en serie. L'arbre k+1 connait l'erreur de son prédécesseur. En conséquence, il peut concentrer son effort sur la correction de ces erreurs désormais connues.

On doit le gradient boosting à Jerome Friedman, statisticien américain, un des co-auteurs de "The elements of statistical learning". Le gradient boosting reprend les principes d'Adaboost, mais alors que Adaboost n'utilise qu'une fonction de cout, le Gradient Boosting les généralise à plusieurs fonctions de coût rendue possible par l'utilisation de la descente de gradient dans la construction itérative des algorithmes faibles.

Dans le Random Forest, les algorithmes faibles sont des arbres de décision unitaires, construits de façon totalement indépendante. Chaque algorithme dispose de la même voix pour le vote final. Le Boosting, lui réalise une somme pondérée pour la décision finale.

Gradient Boosting = Descente de Gradient + Boosting

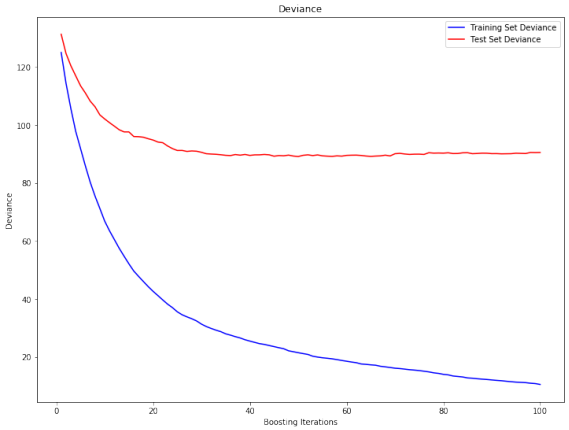

Figure 11: *Deviance of the training and test sets*

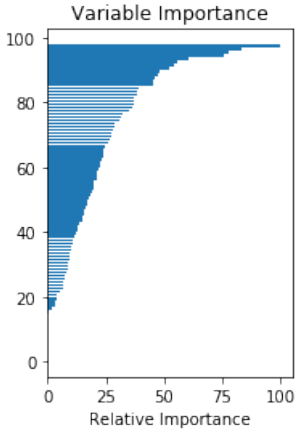

Figure 12: *Features Importance Gradient Boosting*

Le graphique de convergence des divergences sur le train et le test sets montre la bonne paramétrisation de notre algorithme: nous n'overfittons pas et l'erreur converge vers un minimum. Cette paramétrisation sera utilisée pour initialiser notre exploration du XGBoost.

5.4 XGBoost

Le XGBoost (pour extreme gradient boosting) apporte deux éléments très intéressants par rapport au Gradient Boosting.

• Son implémentation est parallèle, ce qui permet de l'entraîner beaucoup plus vite

• Alors que le Gradient Boosting n'implémente que des arbres de régression comme algorithmes faibles, XGBoost permet l'utilisation d'autres algorithmes sous-jacents, comme des modèles linéaires.

Donc le XGBoost crée des arbres de décision en série, en parallèle. Nous utilisons les paramètres du Gradient Boosting pour entrainer initialement notre XGBoost, puis nous avons créé une fonction d'exploration des différentes paramétrisations possibles afin de trouver la meilleure. Cet algorithme nous donne notre meilleur résultat.

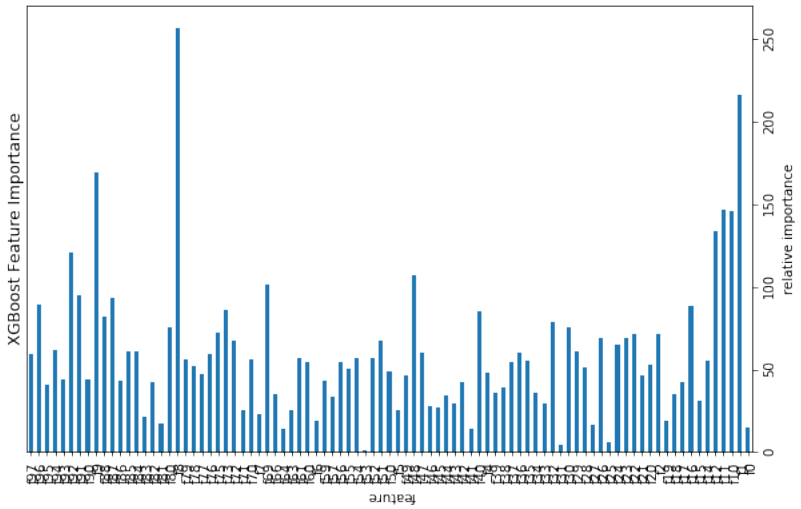

Figure 13: *Features Importance XGBoost (l'axe des abscisses représente chaque feature dans leur ordre de création, du feature 0 au feature 97)*

Parmi les features les plus importants, nous retrouvons bien ce qui était indiqué dans la littérature, c'est-à-dire le pourcentage de phases de sommeil de type 1 et 3 ansi que le rapport des phases. Nous trouvons aussi des résultats discriminants concernant la filtration des signaux EEG après avoir calculé la moyenne, la variance, la skewness et le kurtosis.

Nous avons essayé plusieurs soumissions en utilisant le XGBoost. Nous avons tout d'abord effectué plusieurs tests à partir de résultats obtenus sur un seul échantillon d'entrainement. Ensuite, nous avons effectué plusieurs tests sur une trentaine d'échantillon et nous avons fait la moyenne de toutes nos prédictions. Puis nous avons aussi fait une moyenne pondérée selon nos résultats sur l'échantillon d'entrainement.

6 Conclusion

Scores Obtenus

Random Forest	21.02
Grandient Boosting	20.05
XGBoost	18.79

Le challenge RYTHM nous a permis de comprendre les problématiques quotidiennes d'un Data scientist. L'exploration d'une littérature qui nous était inconnue, afin de construire les features les plus discriminants, a rendu la découverte des différents algorithmes classiques de Machine Learning passionnante. Ce projet est d'autant plus une réussite que nous nous classons (au 23/03 à 16h50 sous le pseudonyme Ilan Dhn) à la cinquième place du classement du challenge sur un ensemble de 100 participants.

Références

[EEG] Methods of EEG Signal Features Extraction Using Linear Analysis in Frequency and Time-Frequency Domains https://www.hindawi.com/journals/isrn/2014/730218/

[Comprehenstion des EEG] These: INSOMNIES ET HYPNOTIQUES: Etude épidémiologique à l'officine de l'Universite de Limosge par Carole DEMONPION

[1] Leo Breiman and Adele Cutler. "random forests - classification description". Department of Statistics Homepage, 28 Apr. 2010.

[2] Lutz, Michel. «Data science : fondamentaux et études de cas.»

[3] https://tel.archives-ouvertes.fr/tel-00944790/document