Tentative de classification d’essais de tâche de Simon chez le rat en fonction de lésions des sous-régions du striatum dorsal.

Résumé :
Les données utilisées pour ce projet proviennent d’une expérience de tâche de Simon chez des rats ayant subits une lésion bilatérale d’une des sous-régions du striatum dorsal (région médiane ou latérale et groupe contrôle). Pour ce travail, j’ai émit l’hypothèse que ces lésions entraînent des modifications des performances des animaux à la tâche, suffisamment différentes pour permettre à un modèle de machine learning d’en classifier les essais. J’ai tout d’abord entraîné un arbre de décision et un random forest classifier pour la catégorisation des essais avec réponse. Les résultats, bien que passables, semblent indiquer des différences dans les caractéristiques des réponses entre les différents groupes. J’ai ensuite entraîné un classifier XGBoost à la catégorisation des essais sans réponse, mais de trop gros déséquilibres dans les données ont conduits à un mauvais apprentissage du modèle et à l’impossibilité de conclure pour ces essais. Finalement, il semble que des lésions des sous-régions du striatum dorsal ne génèrent pas de différences suffisamment majeur dans les performances à une tâche de Simon pour pouvoir être classifier de manière efficace par un programme de machine learning.

Introduction :

Les données utilisées pour ce devoir proviennent d'une expérience de comportement chez le rat visant à étudier l'implication des deux sous-régions striatales, le striatum dorso-médian (DMS) et le striatum dorso-latéral (DLS), dans des processus de contrôle de l'action révélés (notamment) dans lors de la résolution d'une tâche de Simon (Simon 1969).
Brièvement, cette tâche chez le rat nécessite pour l'animal de répondre par un « nose poke » droit ou gauche à l'intensité d'un stimulus lumineux (faible ou forte), présenté lui même latéralisé, mais en ignorant sa position. Le côté de présentation du stimulus, même si non pertinent pour répondre à la tâche, défini 2 types d'essais : compatibles, pour lesquelles le stimulus est présenté du même côté que la réponse attendue ; incompatibles, quand les deux sont opposés.
L'effet Simon décrit une altération des performances (en temps de réaction et taux d'erreur), spécifique aux essais incompatibles et expliquée par un traitement automatique de la position du stimulus par le sujet, qui viendrait interférer avec la sélection de la réponse correcte donnée par l'intensité lumineuse. Interférence qu'un certain nombre de mécanismes de contrôle de l'action tendront à minimiser lorsqu'elle survient.
Les connaissances concernant l'implémentation neuronale de ces mécanismes sont, à l'heure actuelle, principalement corticale. Toutefois un certain nombre d'études suggèrent l'implication des ganglions de la base et notamment du striatum dorsal (Sebastian et al. 2013).
Notre travail visaient donc à étudier les modifications de performances de rats effectuant la tâche, après lésion bilatérale spécifique d'une des deux sous-régions du striatum dorsal.
Du point de vue du machine learning, la question que l'on peut poser est, si c'est lésions ont entraîné des modifications de comportement suffisamment importantes et spécifiques lors de la réalisation de cette tâche pour pouvoir être retrouvées à l'échelle de l'essai par un programme de classification.
Si cela se révèle être le cas, on peut alors se demander quelles sont les variables les plus pertinentes pour permettre la distinction des différents types de lésion.

Matériel et méthodes :

Procédure expérimentale :
42 rats mâles de la souche long-evans ont été utilisés pour cette tâche. Les animaux ont d'abord été entraînés sans l'interférence de position du stimulus pour faciliter l'apprentissage de la règle de réponse (association intensité lumineuse – poke latéral). Ils ont ensuite subit une opération de lésion excitotoxique visant à endommager spécifiquement la partie médiane (DMS) ou latérale (DLS) du striatum dorsal, bilatéralement (12 DMS, 12 DLS, 18 SHAM (e.g contrôles)). Après une période de ré-entrainement, les animaux ont finalement été présentés à la tâche dans sa version complète, avec interférence de position du stimulus pour la sélection de la réponse correcte. Ils ont

alors été soumis à 57 sessions de comportement, durant lesquelles leurs performances ont été enregistrées. Lors de cette phase, un essai correct été comptabilisé pour chaque réponse dans le poke latéral correspondant à l'intensité lumineuse présentée et donnait lieu à une récompense alimentaire. Toute réponse avec un temps de réaction inférieur à 50 millisecondes ou un temps de mouvement supérieur à 800 millisecondes été comptabilisées respectivement comme une anticipation ou une omission.

Système d'acquisition :

8 cages opérantes Med-associates ont été utilisées, équipées de 3 nose pokes sur le mur avant, permettant de gérer le déroulement des essais, et d'une mangeoire reliée à un distributeur de pellets de sucre sur le mur arrière pour délivrer la récompense. Chacun été équipé d'un émetteur infra-rouge associé à une cellule photovoltaïque, permettant de comptabiliser les entrées de l'animal dans chaque élément de la cage et d'en calculer la durée. Tous les évènements de la procédure comportementale étaient ainsi enregistrés en temps réel au format MPC.

Données :

A la fin de la procédure expérimentale, les données ont été formatées en .csv pour pouvoir être analysées en python. Le tableau de données final comprend 410037 lignes et 27 colonnes décrivant pour chaque ligne un essai effectuer par un animal, en 27 variables différentes.

Analyses :

- Pré-traitement :

Les données ont d'abord été pré-traitées pour pouvoir les rendre compatibles avec l'utilisation de programmes de classification. Les erreurs d'acquisition ont été supprimés, 0.08 % des essais, ainsi que les variables temps d'entrée dans les différents éléments de la cage puisqu'il s'agit de temps absolus rendant compte de l'avancement de la session, indépendantes des animaux, et à partir desquels des variables plus pertinentes avaient déjà été calculées comme le temps de réaction et le temps de mouvement. Les variables liées à la procédure expérimentale comme l'effectif des animaux ou leur règle de réponse, qui peuvent être déséquilibrées entre les groupes et donc faciliter l'apprentissage du modèle, mais en faussant sa capacité de généralisation, ont également été supprimées. Une nouvelle variable a été créée à partir des données existantes, la variable Previous_trial, décrivant le type d'essai à l'essai n-1. La littérature humaine sur les processus de contrôle de l'action décrivent en effet, une variation des performances des sujets en fonction du type d'essai précédemment rencontré et de la vitesse de réponse, l'effet Gratton (Gratton et al. 1988). Bien que n'ayant jamais été étudié chez le rat, cet effet pourrait exister et varier différemment en fonction des lésions. Les variables de durée, de temps de réaction et de temps de mouvement ont été standardisées par transformation logarithmique, permettant à la fois de les normaliser et de les mettre à la même échelle. Une transformation nécessaire pour éviter

que les programmes de classification n'attribuent de manière artéfactuelle un poids plus important aux variables dont les mesures sont plus grandes. Les temps de réaction aberrants ont été déterminés comme 2.5 fois supérieurs ou inférieurs à la déviation absolue à la médiane (Leys et al. 2013) et les essais correspondant ont également été supprimés (3.7 % des essais). Les variables catégorielles ont été dummifiées et finalement, 6 rats SHAM on été aléatoirement supprimés pour équilibrer le nombre d'essais par groupe afin de ne pas être en situation de classes non balancées.

Des valeurs manquantes sont présentent dans les variables décrivant la réponse de l'animal (e.g. temps de réaction, durée de la réponse etc) parce qu'un certain nombre d'essais n'ont pas conduit à une réponse (les essais avortés avant le stimulus, les anticipations de réponse et les omissions). Imputer des valeurs dans ces variables n'est donc pas pertinent selon moi puisqu'il ne s'agit pas de mesure manquante. Traiter ces essais particulier peut toutefois être intéressant. On pourrait par exemple émettre l'hypothèse que les lésions entraînent différents patterns d'anticipations ou d'omissions, ce qui permettrait une classification correcte de ces essais. Les valeurs non attribuées ne pouvant pas être traitées par la plupart des programmes de classification, l'analyse a été conduite en deux phases : traitement des essais ayant conduits à une réponse (52 % des essais) ; traitement des autres types d'essais.

- Classification des essais ayant conduits à une réponse :

Par souci d'interprétabilité et d'explicabilité du modèle de classification, j'ai commencé par utilisé un arbre de décision pour cette analyse (un modèle simple et directement interprétable). Le vecteur « y » a été défini comme la colonne « groupe » du tableau, le reste des variables formant la matrice explicative « X ». Les données ont été séparées en un jeu d'entraînement et un jeu de test (méthode hold-out 80/20). Le modèle a d'abord été fité avec les paramètres par défaut pour avoir une base de comparaison. Les prédictions sur les données de test ont été évalués avec plusieurs indicateurs moyens (moyennage entre les classes) : la précision (la proportion de vrais positifs parmi les positifs prédits pour cette classe), le recall (la proportion de prédiction correcte pour cette classe parmi tous les éléments de cette classe), le score f1 (une moyenne entre les deux premières mesure résumant leur évolution), la top k accuracy (la capacité à bien dissocier une classe des autres dans un problème non-binaire (n classes > 2)). Après quoi les hyper-paramètres du modèle ont été optimisés avec la méthode de GridSearch associé à une cross-validation (10 folds) avec maximisation du score f1 (pour rendre compte à la fois de la précision et du recall). Les scores ont à nouveau été calculés, ainsi que l'importance des différentes variables dans la prédiction. Finalement, l'apprentissage du modèle a été évalué grâce à des courbes d'apprentissage (avec cross-validation (5 folds) : sous-ensemble d'entrainement (80 % du jeu initial) ; sous-ensemble de validation (20 % du jeu initial)), révélant un underfitting et montrant ainsi qu'il serait intéressant de tester des modèles de classification plus complexes pour améliorer l'apprentissage des données et les performances de prédiction. Les performances (avec les

paramètres par défaut) des modèles SVM, K-Neighbors, Random Forest, AdaBoost et XGBoost ont donc été comparées. Le modèle random forest ayant obtenus les meilleurs performances, il a été sélectionné pour la suite de l'analyse. Les hyper-paramètres du modèle ont été optimisés de la même manière que précédemment. Seul le nombre d'arbre de décisions à intégrer au modèle (e.i. n-estimators) a été optimisé différemment, en calculant l'erreur « out-of-bag » pour un nombre croissant d'arbre de décision, jusqu'à atteindre un plateau dans la minimisation de cette erreur. Finalement, la librairie SHAP a été utilisée pour évaluer l'importance des différentes variables dans ses prédictions.

- Classification des essais n'ayant pas conduit à une réponse :

Pour cette analyse, seule les essais sans temps de réaction (e.i. sans réponse enregistrée) ont été conservée. Comme précédemment, ce nouveau jeu de donnée à été séparé en jeu d'entrainement et de test (80/20). J'ai ensuite testé les performances de classification de trois modèles : un arbre de décision pour son interprétabilité directe et les modèles Random Forest et XGBoost qui avaient eu les meilleures performances à l'analyse précédente. Le modèle XGBoost a obtenu les meilleures performances avec les paramètres par défaut. Le nombre d'arbre de décisions intégrés au modèle a été optimisé par minimisation de la fonction de coût. Les autres paramètres ont été optimisés avec la méthode GridSearch associée à une cross-validation (3 folds). L'importance des différentes variables a été calculées avec la librairie SHAP comme pour le modèle de Random Forest précédant.

Résultats :

<u>Analyse des essais ayant conduits à une réponse :</u>

- Arbre de décision :

Le modèle d'arbre de décision a eu comme performance 45 % pour la précision, le recall et le score f1 et 72 % pour la top k accuracy avant optimisation puis 52 % et 81 % après optimisation des hyper-paramètres. La matrice de confusion (figure 1A) montre un score de classification équivalent pour les classes SHAM et DLS mais des performances moindre

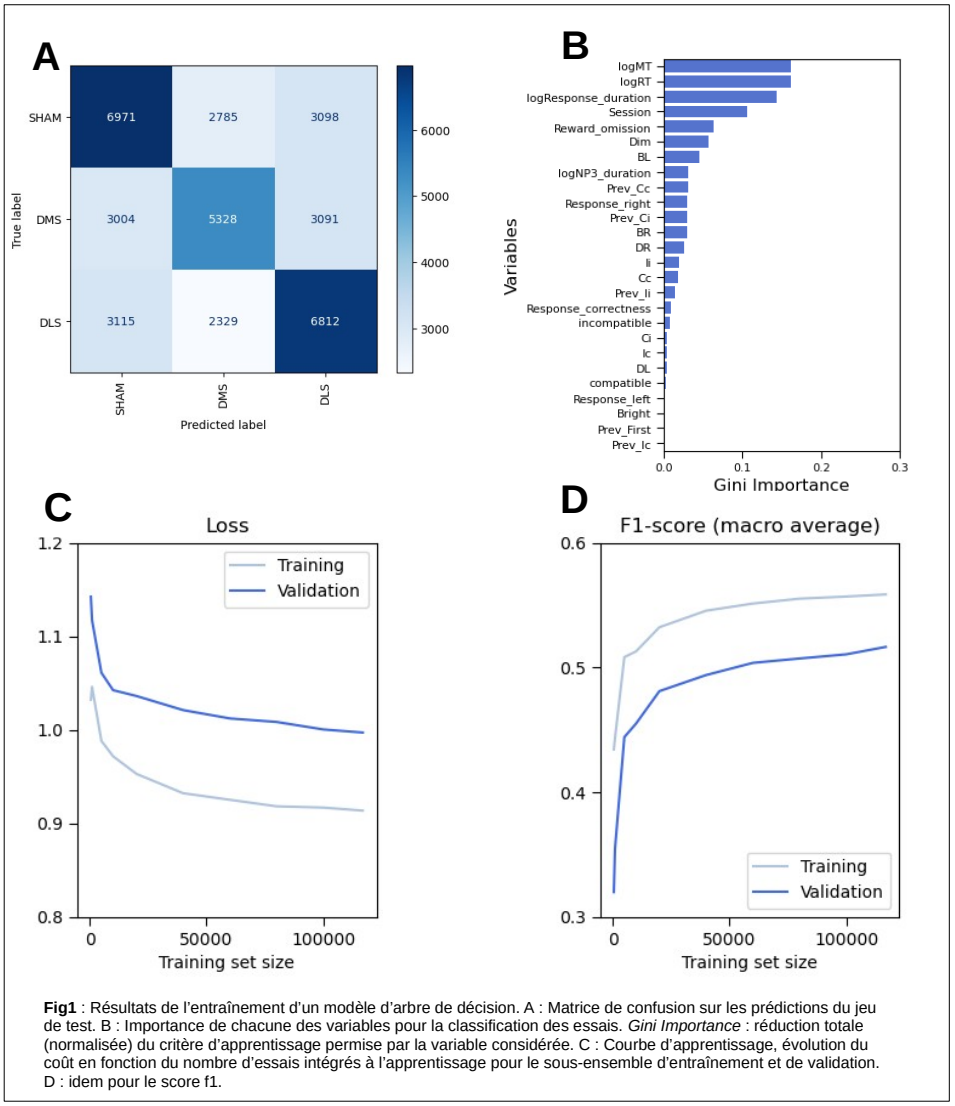

Fig1 : Résultats de l'entraînement d'un modèle d'arbre de décision. A : Matrice de confusion sur les prédictions du jeu de test. B : Importance de chacune des variables pour la classification des essais. *Gini Importance* : réduction totale (normalisée) du critère d'apprentissage permise par la variable considérée. C : Courbe d'apprentissage, évolution du coût en fonction du nombre d'essais intégrés à l'apprentissage pour le sous-ensemble d'entraînement et de validation. D : idem pour le score f1.

pour la classe DMS. L'importance des différentes variables dans l'apprentissage du modèle d'arbre de décision est présentée figure 1B et montre que l'attribution d'un essais à une classe par ce modèle est principalement permise par les caractéristiques de la réponse (temps de mouvement, temps de réaction, durée du poke dans le nose poke latéral) et les sessions d'apprentissage des animaux.Les courbes d'apprentissage du modèle sont représentées en figure 1C et D. Le panneau C montre la diminution de la fonction de coût avec l'augmentation du nombre d'essais considérés durant l'apprentissage du modèle en bleu clair pour le sous-ensemble d'entraînement et en bleu foncé pour le sous-ensemble de validation. Le panneau B montre l'évolution du score f1 calculé pour les sous-ensembles d'entraînement (bleu clair) et de validation (bleu foncé) en fonction du nombre d'essais utilisés pour l'apprentissage.
La différence entre les courbes d'entraînement et de validation ainsi que la pente des courbes qui n'est pas nulle sur les deux graphiques semble indiquer un underfitting. Le modèle d'arbre de décision ne semble pas assez complexe pour modéliser correctement les données.

- Comparaison des modèles :

Les performances des différents modèles entraînés avec les paramètres par défaut sont présentées en table 1. Le Random Forest Classifier qui obtient les meilleurs résultats est conservé pour la suite de l'analyse.

Mesures (%)	DTCopti	SVC	KNC	RFC	ADA	XGB
Top K Accuracy :	81.8	72.6	79.8	83.1	74.1	83.6
Précision :	52.2	41.5	50.4	54.1	40.0	54.0
Recall :	52.1	39.2	49.8	54.1	39.9	53.8
score F1 :	52.1	34.9	49.7	54.1	38.7	53.7

Table 1 : Comparaison des performances de différents modèles de classification pour les essais ayant conduits à une réponse. *DTCopti* : Decision Tree Classifier optimisé ; *SVC* : Support Vector Classification ; *KNC* : K-Neighbors Classifier ; *RFC* : Random Forest Classifier ; *ADA* : AdaBoost Classifier ; *XGB* : XGBoost Classifier

- Random Forest Classifier :

Le nombre d'arbres intégrés au modèle a été déterminé par minimisation de l'erreur « Out-of-Bag » présenté en figure 2A. Cette figure montre que l'erreur se stabilise autour de 250 arbres de décisions, rendant inutile l'ajout d'arbres supplémentaires. Les autres hyper-paramètres du modèle ont été optimisés par GridSearch avec cross-validation et minimisation du score f1 tel que décrit précédemment. Le modèle, après optimisation, a eu comme performance 55 % pour la précision, le recall et le score f1 et 84 % pour la top k accuracy.
La matrice de confusion des prédictions faites sur le jeu de données test est présentée en figure 2B. Celle-ci montre une meilleure performance de classification des classes SHAM et DLS que DMS.

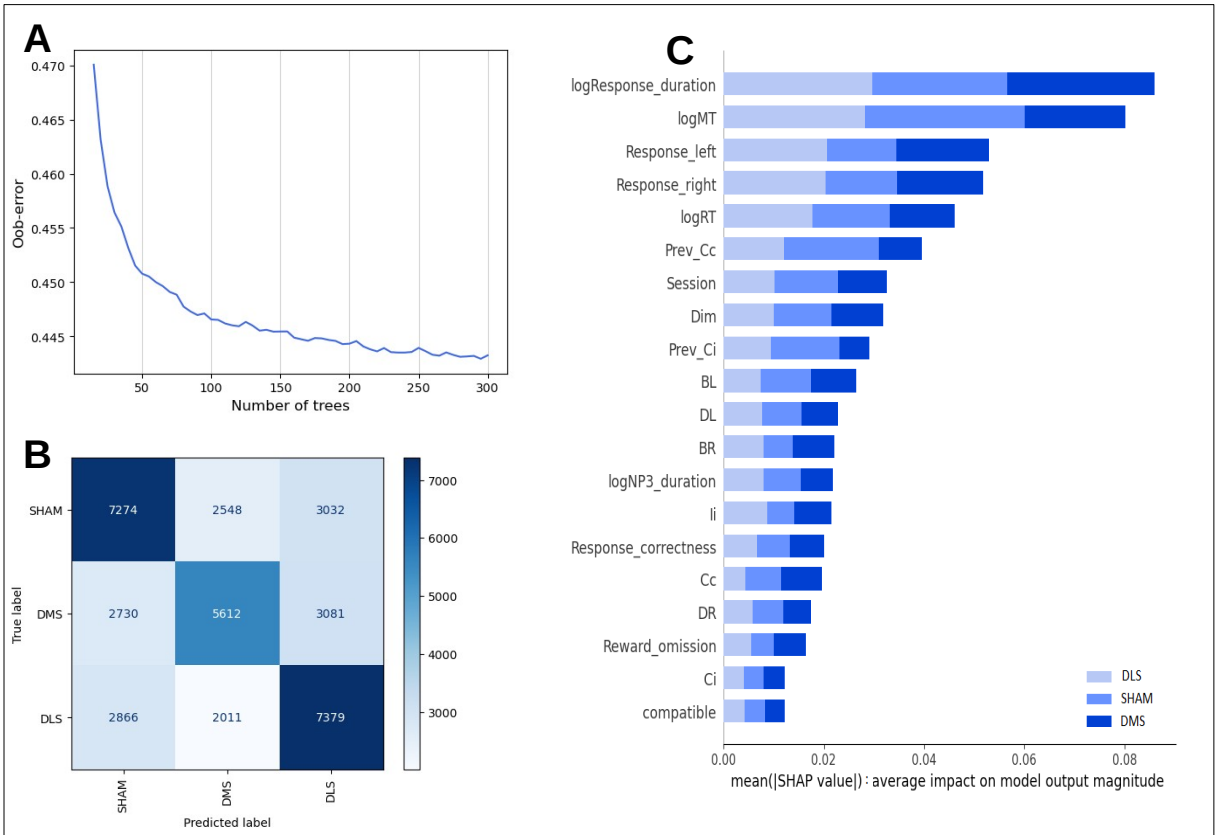

Fig2 : Résultats de l'entraînement d'un modèle de Random Forest Classifier. A : Evolution de l'erreur « Out-of-Bag » à mesure que des arbres de décision sont ajoutés au modèle. B : Matrice de confusion sur les prédictions du jeu test. C : Importance des différentes variables dans l'apprentissage du modèle pour chacune des classes.

L'importance des différentes variables dans l'apprentissage du modèle est présentée figure 1C et montre que les variables ayant le plus d'impact pour l'attribution d'un essais à une classe par ce modèle sont la durée de réponse et le temps de mouvement. Contrairement au modèle d'arbre de décision simple, le temps de réaction et les sessions d'apprentissage n'apparaissent que plus bas dans la liste, précédées par le côté de réponse.

Analyse des essais n'ayant pas conduits à une réponse :

- Comparaison des modèles :

Les performances, avec les paramètres par défaut, des trois modèles comparés pour la classification des essais n'ayant pas conduit à une réponse est présenté en table 2. Le modèle XGBoost ayant eu les meilleures performances, il est conservé pour la suite de l'analyse.

Mesures (%)	DTC	RFC	XGB
Top K Accuracy :	70.8	70.7	74.4
Précision :	36.1	36.7	41.5
Recall :	36.1	36.7	39.8
score F1 :	36.0	36.6	38.2

Table 2 : Comparaison des performances de différents modèles de classification pour les essais n'ayant pas conduit à une réponse. *DTCopti* : Decision Tree Classifier optimisé ; *SVC* : Support Vector Classification ; *KNC* : K-Neighbors Classifier ; *RFC* : Random Forest Classifier ; *ADA* : AdaBoost Classifier ; *XGB* : XGBoost Classifier

- XGBoost :

La vitesse d'apprentissage du modèle a été fixée à 0.3. Le nombre d'arbres intégrés au modèle a été optimisé par minimisation d'une fonction de coût présenté en figure 3A. On observe sur cette figure que les performances du modèle sur le jeu de test sont à leur maximum à 40 arbres, après quoi le modèle overfit. Les autres paramètres du modèle ont été optimisé avec la méthode de GridSearch. Les performances du modèle après optimisation sont

de 74 % pour la top k accuracy, 40 % pour la précision, 39 % pour le recall et 38 % pour le score f1. La matrice de confusion des prédictions faites sur le jeu de test est présentée en figure 3B. Elle montre que le modèle prédit principalement la classe DLS, ce qui explique les mauvais score de précision et recall. Cela pourrait indiquer un déséquilibre dans le jeu de données en faveur de la classe DLS. C'est effectivement le cas, puisque la classe DLS apparaît 25 % plus souvent que la classe SHAM et 15 % plus souvent que la classe DMS.

L'importance des variables pour les prédictions de ce modèle sont présentées en figure 1C. Les trois variables les plus impactantes sont la durée du poke central (poke qui précède l'apparition du stimulus), les sessions d'apprentissage des animaux et le type d'essai compatible.

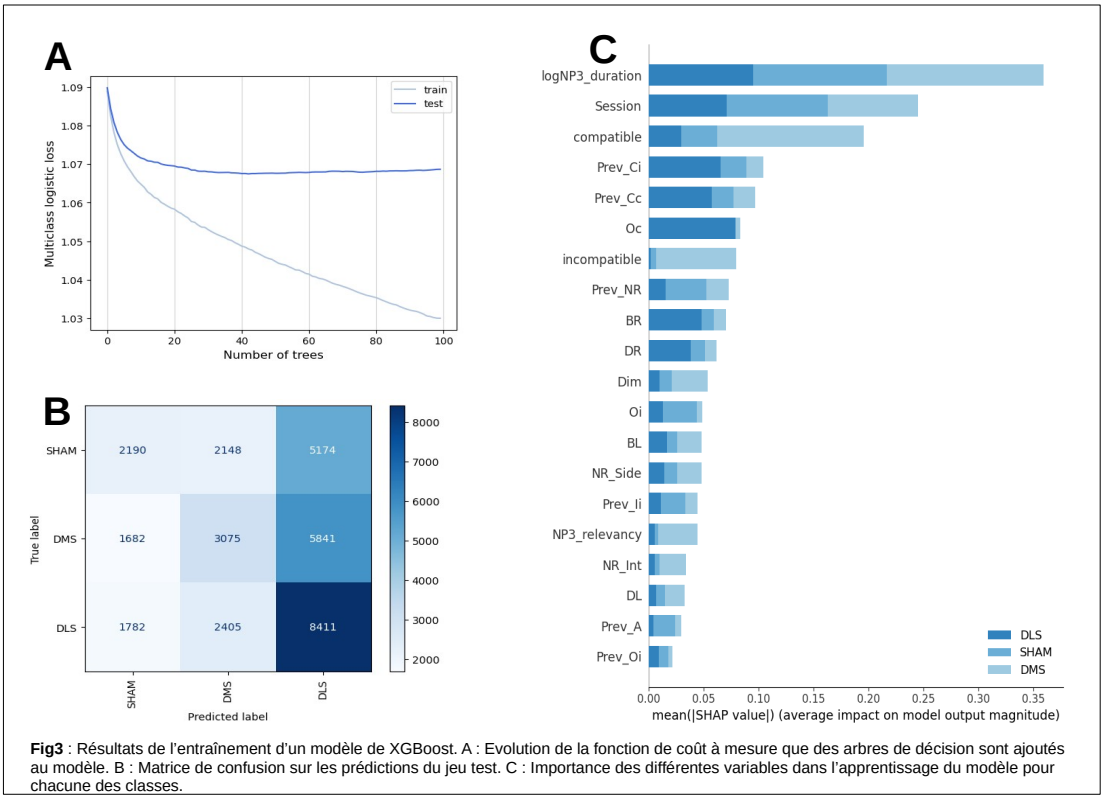

Fig3 : Résultats de l'entraînement d'un modèle de XGBoost. A : Evolution de la fonction de coût à mesure que des arbres de décision sont ajoutés au modèle. B : Matrice de confusion sur les prédictions du jeu test. C : Importance des différentes variables dans l'apprentissage du modèle pour chacune des classes.

Discussion :

Le but de ce projet a été de tester si des lésions chez le rat des deux sous-régions du striatum dorsal entraînent des modifications de comportement suffisamment importantes, lors de la réalisation d'une tâche de Simon, pour permettre la classification correcte de ses essais par un programme de machine learning. L'analyse a été faite en deux phases : une classification des essais ayant conduits à une réponse ; une classification des essais pour lesquels ça n'a pas été le cas (essais avortés, anticipations, omissions).

Tout d'abord, la classification des essais avec réponse montre des résultats mitigés. Un arbre de décision entraîné avec ces essais obtient un score d'accuracy correct (> 80%) mais les scores de précision, recall et f1 restent faibles (≈ 50%) indiquant un nombre d'erreurs important. La complexification du modèle utilisé, suggéré par l'interprétation des courbes d'apprentissage de l'arbre de décision, n'est finalement pas très utile, avec des scores obtenus par le Random Forest Classifier seulement 3 % supérieurs aux scores de l'arbre de décision. Cette faible amélioration est

également complètement aspécifique, puisque le pattern de performance révélé par les matrices de confusion des deux modèles est le même, avec une moins bonne classification des essais DMS que DLS et SHAM. L'importance des différentes variables dans les performances de classification des deux modèles montrent aussi un pattern assez équivalent. On observe ainsi que celui-ci se base principalement sur les caractéristiques de la réponse (temps de mouvement, temps de réaction et durée du poke de réponse), les sessions et le côté de réponse.
La variable sessions pourraient indiquer un apprentissage différent de la tâche par les animaux en fonction de la lésion, ce qui est effectivement le cas. La variable côté de réponse pourrait s'expliquer de plusieurs manières. La plus simple serait un déséquilibre dans les données entre les groupes, issu du tirage aléatoire du côté de réponse durant la tâche. Toutefois il est intéressant de remarquer que ces variables ont plus d'impact sur la classification des essais en DLS et DMS qu'en SHAM. Le striatum dorsal jouant un rôle important dans le contrôle moteur et la sélection de l'action, sa lésion pourrait entraîner un pattern de réponse plus stéréotypé chez ces animaux, expliquant ainsi l'importance de cette variable dans la classification des essais. Finalement, les caractéristiques de la réponse comme facteurs principaux permettant la dissociation entre les classes est assez cohérent au vue des fonctions connues des structures lésées. Une lésion dorso-striatale entraîne, lors d'une tâche de Simon, un ralentissement globale du temps de mouvement et du temps de réaction. Cette augmentation est d'autant plus importante pour les rats DLS en situation d'essais incompatibles.
Malgré tout, les performances globales de ces modèles restent passables, ce qui laisse à penser que les variables utilisées ne sont pas suffisamment clivantes entre les groupes pour permettre une classification correctes par un modèle de machine learning.
Ensuite concernant la classification des essais n'ayant pas conduits à une réponse, les résultats sont assez médiocres. Le modèle XGBoost entraîné pour cette tâche n'a en effet obtenu que 38 % de score f1. La matrice de confusion de ses prédictions montre que le modèle prédit majoritairement la classe DLS. Cela doit pouvoir s'expliquer par le déséquilibre important entre les classes dans le jeu de données, la classe DLS apparaissant 15 % plus souvent que la classe DMS et près de 25 % de plus que la classe SHAM. Ce déséquilibre en lui même tant à montrer l'impact de cette lésion sur le comportement des animaux durant la tâche, qui font donc plus d'essais avortés, d'anticipations et d'omissions que les autres. Je n'ai pas trouver de solution pour équilibrer proprement le jeu de données. Dans ces conditions je ne pense pas que l'interprétation des variables importantes aux prédictions du modèle soit pertinente.

Dans ce projet j'ai donc voulu tester la classification d'essais de tâche de Simon chez le rat en fonction de la lésion dorso-striatale qu'avaient subits les animaux. La difficulté principale rencontrée pour entraîner correctement les modèles a été le déséquilibre du nombre de chaque type d'essais par animaux et type de lésion. Déséquilibre trop important pour pourvoir classifier les essais sans réponse. Pour les essais avec réponse, les résultats ont possiblement mis en

évidence des différences dans les caractéristiques de la réponse entre les groupes, mais rien de suffisamment important pour permettre une classification avec de bonnes performances.

Bibliographie :

- Gratton, G., Coles, M. G., Sirevaag, E. J., Eriksen, C. W., & Donchin, E. (1988). Pre-and poststimulus activation of response channels: a psychophysiological analysis. *Journal of Experimental Psychology: Human perception and performance*, *14*(3), 331.

- Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, *49*(4), 764-766.

- Sebastian, A., Pohl, M. F., Klöppel, S., Feige, B., Lange, T., Stahl, C., ... & Tüscher, O. (2013). Disentangling common and specific neural subprocesses of response inhibition. *Neuroimage*, *64*, 601-615.

- Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of experimental psychology*, *81*(1), 174.