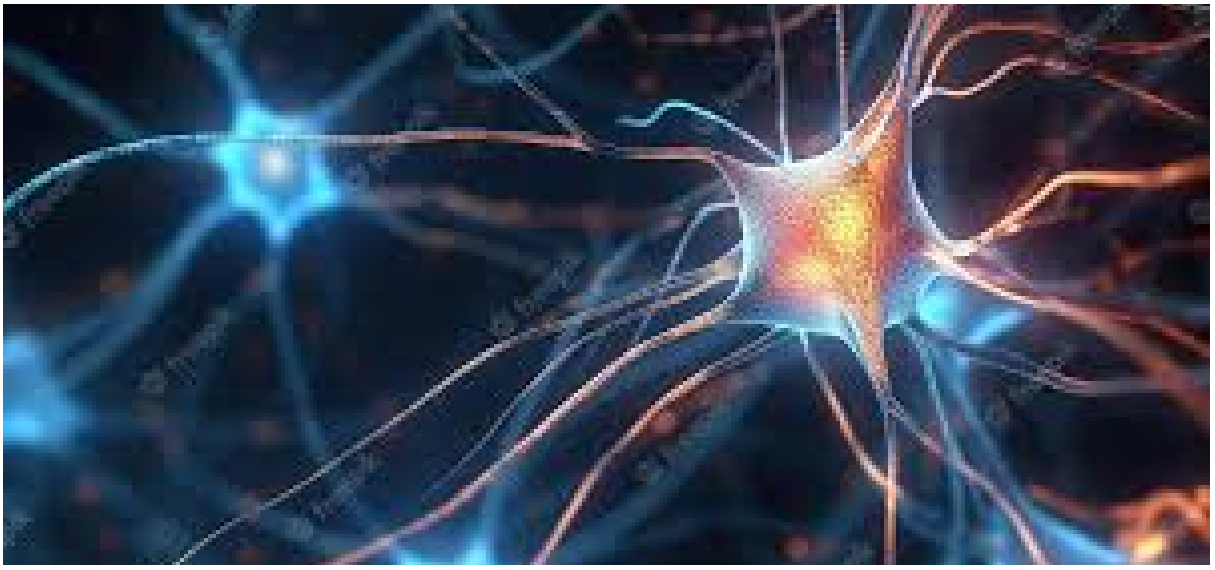

Réunions Scrum Masters

Compte-rendu S1

Réunion Scrum Masters : Sprint S1

Compte-rendu S2

Réunion Scrum - Sprint 2

Compte-rendu S3

- Le dataset "heart diseases" contient une ligne en doublon dont toutes les valeurs sont strictement identiques à la ligne qui la précède. Comment gérer ce doublon (suppression etc.) ? On doit le supprimer.
- Sur le dataset "cancer breast", 13 lignes ont les variables relatives à la concavité égales à 0. Que souhaitez-vous que nous fassions avec ces valeurs ? Les laisser comme telles / les supprimer / les modifier (le cas échéant, comment ?) Deux façons de faire : enlever toutes les lignes soit utiliser moyenne ou médiane pour la Nan (va nous faire un message sur Slack pour nous dire comment procéder) : pour variable non-skewed = médiane ou moyenne / pour les skewed = utiliser la moyenne. Soufiane revient vers nous avec plus d'informations.
- Les outliers peuvent nuire au modèle.

- Dans la dataset “liver diseases”, bcp plus de patient appartenant à une catégorie plutôt qu’une autre (416 VS 167). Est-ce que, pour l’entraînement du modèle, ça ne risque pas de poser un pb (mieux entraîner à détecter un patient d’une catégorie plutôt qu’une autre) ? Plusieurs méthodes à creuser pour résoudre le soucis. 1ère façon = créer et simuler des nouvelles données pour équilibrer les groupes / 2ème façon = gérer la distribution tout seul càd pdt l’entraînement, couper la df en part égale avec les deux catégories (pdt le train, test, split). Il faut à peu près une proportion de 80 pour 100.
- Les stat’ permettent de faire une justification scientifique, pas la tendance donc impératif de faire des statistiques pour valider les tendances. Utiliser la méthode Kruskal Wallis ou Test Student de la librairie Statannot pour comparer les différence entre deux variables pour deux classes différentes et valider qu’une tendance est vraie avec le score P par exemple.
- CKD, dataset avec le plus de challenge en termes de traitement. Si 30% de valeurs Nan ou à 0 sur une ligne, on exclue le patient. A partir de 3 valeurs manquantes pour un patient dans ce dataset, on supprime la ligne. Pas de suppression de lignes sur les autres dataset.
- Calculer écart interquartiles ? Quel avantage par rapport à la boîte à moustache ? boxplot = approche graphique (parler à un artiste) / si on veut du très précis, écart interquartile pour avoir les chiffres (parler à un statisticien).
- Pb des 0 = construire le modèle en conservant les 0 ou les remplacer par la moyenne-médiane (avec la skewness) et voir lequel des deux modèles est le plus performant.
- Quand on a des valeurs 0, 1, 2, 3, 4, 5 = ce sont toujours des classes ou des niveaux.

Réunion Scrum 4

Compte-rendu dernière semaine

Choix des variables, méthodes, comment le modèle répond aux attentes du client.

A quel point on simule la réalité ? Faisons-nous des calculs ou du ML ?

Présenter cela sous forme de rapport dans les livrables.

Cancer du sein, score à 99. 2 ou 3 variables suffisent pour dire si cellules malades ou non.
supprimer les colonnes correllées entre elles (rayon et périmètre)

Expliquer pourquoi modèle fonctionne très très bien sur K sein et moins sur une autre patho par ex.

Rein: scores très élevés. lié à la petite taille du jeu de données ? NON

comment avons-nous stocké nos modèles

joblib ou pickle: bibliothèque qui permet de faire des pipelines

Pour les applis, technos utilisées: docker. snowflake.
3 outils nécessaires dbt, snowflake et airflow sont nécessaires pour être valuable sur le marché.

Condenser le code notebook. faire des fonctions

Utiliser une classe réutilisable dans n'importe quelle appli. **Voir ce qu'est une classe.**

Attentes démo day

Prez comprise par tout le monde ⇒ vulgariser au maximum
Introduction / contexte général du projet 3
Problématique
Orientation que nous avons prise vs pbtique
approche utilisée: ex nettoyage des données: qu'avons-nous fait exactement?
présenter les limites de ce que nous avons fait.

données test: 1 patient par dataset mercredi soir.