# Rapport Projet Apprentissage Automatique & Réseaux de Neurones

Aghiles DJEBARA
171731096628

Aymen Rayane KHOUAS
161731063111

# Table des matières

# Chapitre 1

# État de l'art

## 1.1 Introduction

Depuis que l'OMS [1] a déclaré une pandémie mondiale, la COVID-19 est devenu l'un des sujets de recherches principaux des scientifiques. Différents chercheurs de différents domaines essayent désespérément de comprendre ce virus afin de cerner le problème et de limiter les dégâts, Notamment les data scientistes qui, grâce aux statistiques récoltées en masse durant cette période, ont tenté de d'extraire des informations pertinentes et d'établir et modèles de prédiction qui contribueraient à la gestion de cette catastrophe.

## 1.2 Problématique

Durant la pandémie, plusieurs théories ont été avancées sur le possible impact du régime alimentaire d'un individu sur la COVID-19 *cf.* 1.3, l'objet de se travaille est de vérifier ces hypothèses en exploitant les données récoltées durant et avant la pandémie sur les cas de COVID-19 [1], et les apports alimentaires par pays [2].

## 1.3 Travaux liés

### 1.3.1 Impact of Nutrition and Diet on COVID-19 Infection and Implications for Kidney Health and Kidney Disease Management

Ce papier [4] parle de l'impact de la nutrition sur la COVID-19 et ses implications sur le fonctionnement des reins. Ils résument les différents travaux liés expliquant la potentielle corrélation existante.

### 1.3.2 Nutrition advice for adults during the COVID-19 outbreak

Cet article [5] recense les conseils de l'OMS concernant le régime alimentaire à adopter durant la période de pandémie. Il est à noté que le régime recommandé durant la pandémie est différent de celui recommandé en temps normal.

### 1.3.3 The Role of Nutrition in the COVID-19 Pandemic

Cet Article [6] publier en 2021 dans la revue Nutrients "vise à résumer la relation complexe entre l'infection par le SRAS-CoV-2 et l'état nutritionnel et les effets de la malnutrition en termes de gravité de la maladie, de temps de récupération des patients, d'incidence des complications et de taux de mortalité." [6]. l'article confirme l'existence d'une relation cruciale entre le régime alimentaire et la manifestation de COVID-19, particulièrement chez les personnes a risque.

1. Organisation Mondiale de la Santé

# Chapitre 2

# Conception & Implémentation

## 2.1 Approche

Afin de déterminer s'il existe une quelconque relation entre la nutrition et la COVID-19, nous allons entraîner un modèle de réseau de neurones qui prendra en input la moyenne des différentes doses de nutriments par groupe de personnes, et en output, le tôt de mortalité au sein du même groupe de personnes. Si l'apprentissage converge, alors nous pourront affirmer qu'il existe un lien entre le régime alimentaire et la COVID-19.

## 2.2 Description des données

Les données utiliser provienne de deux sources différentes, nous avons dans un premier lieu des données par le GDD [1] [2] qui décrivent les taux d'alimentation moyens de certains produits par pays, tranches d'ages, etc, et dans un second lieu des données sur le taux de cas et e décès lier à la COVID-19 par pays par l'ECDC [2] [1].

### 2.2.1 Données des taux de Covid19 par l'ECDC

Le dataset de l'ECDC [1] décrit les cas et morts journaliers par COVID-19 dans 210 pays différents. chaque ligne du dataset représente les données d'un jour dans un pays donné. Pour chaque jour le dataset renseigne la date avec le format JJ/MM/AAAA ainsi que dans les colonnes "day", "month" et "year" (pour facilité la manipulation), on va utiliser les dates pour enlever la période du début de la campagne de vaccination si nécessaire pour ne pas fausser les résultats. Figure aussi dans le dataset le nom des 212 pays traiter sous trois formats diffèrent (un code du pays sur 2 et 3 lettres et le nom complet du pays), le continent ainsi que la taille de la population du pays, et enfin le dataset contient les données lier à la COVID-19, à savoir le nombre de cas et décès journaliers, mais aussi le cumul des cas sur les 14 derniers jours en fonction de la population, donnée qu'on ne va pas utiliser, car cherchant à faire un cumul des cas par pays, il est plus intéressant et simple d'utiliser les cas journaliers.

### 2.2.2 Données d'apport alimentaire global par le GDD

Nous avons utilisé les données du GDD les plus récentes [2] (07/06/2021) qui récapitulent des données alimentaire allant jusqu'à 2018, ces dernières contiennent plusieurs datasets représentant les données moyenne de l'apport alimentaire de certains produits par pays (ou région en fonction du dataset) sur plusieurs tranches d'âges, niveaux d'éducations.. Etc, et enfin de l'année du recensement (1990, 2000, 2015, 2018, etc.).

les données sont représentées dans trois datasets différents, dans le premier les données ne sont pas séparé par région ou pays, mais globale a toute la population mondiale, le deuxième séparent les données

1. Global Dietary Database
2. European Centre for Disease Prevention and Control

par super-régions et le troisième par pays (185 pays), ça sera ce dernier qui vas nous intéresser étant donnée que les données à disposition pour la COVID-19 sont répertorié par pays.

Notre Dataset est divisé en plusieurs fichiers CSV ou chacun représente les données des aliments spécifiques qui sont listé dans la première page de la documentation fournie avec les données [3] (36 au total). Chaque ligne du datasets représente les données alimentaires d'un groupe de personnes avec les paramètres donné plus hauts (âge, sexe, pays, année de l'étude, etc.).

Les données sont séparées en plusieurs valeurs dont le pays (avec un encodage sur 3 lettres), l'année de la récolte de données (7 ans différente de 1990 à 2018), mais aussi d'autres paramètres qui sont codifié sur des entiers (par exemple les zones d'habitat sont codifié sur 2 entiers, 1 pour une zone urbaine et 0 pour une zone rurale), le paramètre est ignoré avec la valeur 999, donc à valeur 999 on prend en compte toute la population sans tenir compte du paramètre (pour le même exemple sur la zone d'habitat la valeur 999 représente l'union des gens habitant une zone rurale et urbaine. Les autres paramètres brièvement sont le sexe, la tranche d'âge, les niveaux d'éducation et la zone (urbaine et rurale) [4]

Pour chaque entrer dans le dataset, est disponible la moyenne de consommation journalière de produits traiter dans le fichier dans la colonne "median" [5] avec comme unité une unité spéciale utiliser par GDD appeler sobrement GDD units [6], le upperci_95 (respectivement lowerci_95) qui représente la moyenne de la consommation moyenne des 95% consommant le plus (respectivement le moins), est disponible aussi serving, s_lowerci_95 et s_upperci_95, qui représente la même chose avec des unités différentes (remarque qu'elle n'est pas disponible pour tous les produits).

## 2.3 Pré-traitement des données

Avant de pouvoir utiliser nos données pour entraîner notre modèle, nous devons d'abord traiter nos données puis fusionner nos datasets.

### 2.3.1 Pré-traitement des données des taux de Covid19

pour ce qui est du prétraitement du dataset des taux de COVID-19 le but et d'avoir à la fin un datasets constituer des pays leur nombre de cas total de COVID-19 et le nombre de décès dû à la COVID-19 durant la période de temps choisit. Nous avons considéré la période de vaccination, mais étant donnée qu'elle a commencé début 2021 elle était de base en dehors de la période couverte par notre dataset. Comme discuter plus hauts nous avons fait la somme du nombre journalier de décès et de cas de COVID-19 pour chaque pays que nous avons ensuite divisé sur la population du pays pour avoir un résultat qui ne sera pas influencé par la taille de population, néanmoins les résultats obtenus étaient trop petits, la valeur max du nombre de morts / population étant de 0.0015 et celle du nombre de cas / population de 0.09 et ce par ce que la taille de la population reste bien supérieur au nombre de personnes ayant contracté la COVID-19, pour avoir des targets allons de 0 à 1 et des données plus distribuer nous avons multiplié le nombre de décès par 100 et le nombre de cas par 10.

### 2.3.2 Pré-traitement des données d'apport alimentaire

Comme expliquer plus haut (2.2.2) les données d'apport alimentaire sont constituées de 3 datasets différents desquels on va prendre celui divisant nos données en pays. pour l'année de l'étude on a pris l'année 2018 étant la dernière étude à cette date et l'année la plus proche de la pandémie. Pour les autres paramètres(Les tranches d'âge, sexe, zone et niveaux d'éducation) nous avons décidé de les ignorer (autrement dit de prendre la valeur 999 qui ignore la séparation de la population par ces paramètres) car

3. Le fichier "GDD 2018 Codebook_Jun 7 2021.xlsx"

4. le sexe et la zone d'habitat sont codifier sur deux entiers, le niveaux d'éducation sur 3 et les tranches d'âge sur 22 entiers (plus pour chacun d'entre eux la valeur 999 qui ignore le paramètre)

5. Malgré le nom "median" il est confirmer dans la documentation qu'il s'agit bien de la la moyenne

6. L'unité dépend de chaque produit par exemple "grams per day pour du jus de fruits", les valleurs sont donner intégralement dans la documentation

ces derniers ne sont pas représenté dans le dataset de COVID-19, donc séparer nos données par l'âge par exemple fausserais nos résultats en associant les données alimentaires de pays par tranche d'âge à celles des cas/décès de COVID-19 en rapport a la population entière d'un pays. Pour ce qui est des données de test nous avons pris uniquement la moyenne donc la colonne "median" de chaque produit,

Après avoir répété ce traitement pour chaque fichier du dataset [7] nous avons fusionné ces derniers pour en créer un seul avec 37 colonnes, à savoir le code à trois lettres du pays et les 36 moyenne de consommations pour chaque produit.

### 2.3.3 Dataset final

Notre Dataset final est une fusion des datasets résultant du prétraitement effectuer sur les deux datasets (taux de COVID-19 et apport alimentaire), nous avons effectué une jointure sur le codage a 3 chiffres des pays (iso3), les données résultantes sont un dataset de 39 colonnes (pays, 36 colonnes pour les moyennes des apports alimentaire et 2 colonnes pour le nombre de cas/décès) et 180 lignes [8], une pour chaque pays. Noter qu'on a supprimé la colonne "pays" avant de sauvegarder le dataset dans un fichier CSV car inutile pour faire le train.

**Problème avec la taille des données :** Il faut noter que 180 inputs la plupart du temps n'est d'être suffisants pour faire un train, nous avons essayé de régler ou contourner le problème, mais nous avons toujours étais ralenti par le fait que les dataset de taux de COVID-19 et apport alimentaire ont seulement en commun les pays, donc utiliser les tranches d'âges, sexes, etc. Ce n'est pas une solution acceptable. Nous avons discuté deux autres solutions, l'une qui consiste à prendre les données sur différentes années, mais le problème avec cette approche étant non seulement que nous aurons des données biaiser (étant donnée que le target sera le même pour toutes les années prise en compte) mais on devra prendre aussi des données qui ne sont plus a jours et potentiellement très loin de la pandémie donc pas pertinente (2018 et 2015 sont les seules années assez proches de notre intervalle de temps sur COVID-19). La deuxième solution consistait à prendre une paire de pays/produits comme input, donc un input contiendra la moyenne d'un seul produit au lieu de tous les produits par pays ce qui nous permettrais d'avoir 36*180 inputs, mais cette solution est aussi problématique, principalement par ce qu'elle est aussi soumise au problème des target qui seront les mêmes pour tous les inputs d'un pays, mais aussi, car elle s'éloigne du problème poser étant donné que nous testerons l'impact des quelques produits séparément sur la COVID-19 sans pouvoir démontrer l'impacte de l'alimentation en général. Donc ayant préféré avoir des données de qualité et fiable plutôt que des données nombreuses nous avons décidé de travailler avec 180 lignes (en ayant conscience que cela pourrait impacter négativement les résultats).

## 2.4 Conception du réseau de neurones

Afin de déterminer l'architecture la plus adéquate à ce problème, nous allons faire plusieurs entraînements avec diverses architectures et fonctions d'apprentissages. La sélection de l'architecture que nous utiliserons dépend des critères suivants :

- La performance sur les données d'entraînement et de test.
- Présence ou absence du problème de sur-apprentissage.

Une fois l'architecture sélectionnée, nous effectuerons des réglages sur les hyper-paramètres que nous évaluerons sur les données de validation afin d'obtenir les meilleurs résultats possibles.

### 2.4.1 Détermination du target

En ce qui concerne le target, nous avons jugé plus adéquat d'étudier la causalité entre la nutrition et le taux de mortalité, car le nombre de cas dans un pays dépend d'autres paramètres [3] qui ne figurent pas

7. Un fichier du dataset représente les données sur un produit en particulier

8. il y avait 179 pays résultant de la jointure, certain pays ne figurait pas dans un dataset ou l'autre, mais après vérification manuelle il c'est avérer que l'un d'entre eux était présent avec un codage diffèrent (taiwan) nous l'avons donc ajouter au dataset (plus précisément nous avons modifier son codage dans la phase de pré-traitement)

dans nos données. Dans le cadre de ce travail, nous nous sommes concentrés sur l'impact de la nutrition sur la COVID-19 afin de confirmer les hypothèses des biologistes.

## 2.5 Organisation & utilisation du code

Nous avons utilisé python avec la bibliothèque pandas pour faire le prétraitement des données et MATLAB pour notre réseau de neurones. l'intégralité du code de notre projet est disponible sur ce répertoire GitHub `https://github.com/Ghiles1010/Nutrition-Impact-On-Covid-19.git` mais aussi sur le CD transmit avec ce rapport, nous allons maintenant brièvement expliquer le script du prétraitement et comment le faire marcher avant de passer vers les scriptes MATLAB.

### Pré-traitement

Comme citer plus hauts nous avons fait le prétraitement avec python en utilisant pandas, nous avons travaillé sur un notebook Google collab en utilisant dont vous pouvez accéder et exécuter sur ce lien `https://colab.research.google.com/drive/18VbEG7i65iFUS2dSK2DspqhJbKGvBcEW?usp=sharing`, nous avons aussi transmis un script python générer à partir de Google collab (avec quelques modifications pour qu'ils puissent fonctionner), il faudra néanmoins avoir une version de Python et installer pandas [9], nous avons aussi transmis les résultats obtenus après prétraitement dans un fichier csv. Le script commence d'abord par le prétraitement des données des taux de COVID-19 puis ceux d'apport alimentaire et enfin joins les 2 data set.

### Code de traitement

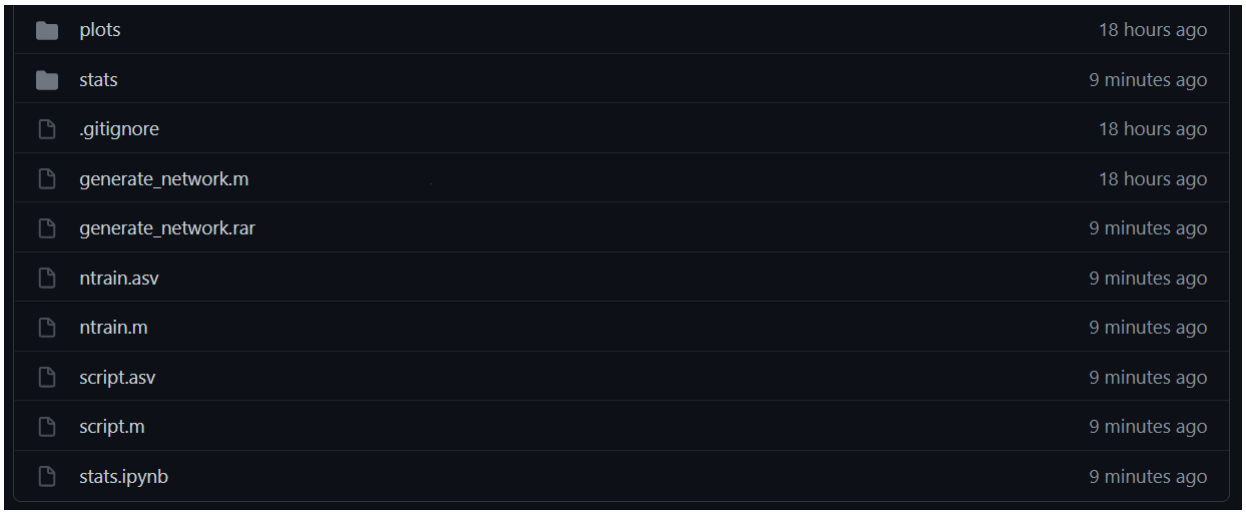


FIGURE 2.1 – Code de traitement

- **script.m** : Ce script est le fichier principal, il utilise les deux fonctions **ntrain** et **generate_network.m**.
- **generate_network.m** Ce script est une fonction qui génère une architecture de façon aléatoire.
- **ntrain.m** Ce script est une fonction qui entraine une architecture passée en entrée.
- **stats.ipynb** Ce script est un notebook qui permet de générer des statistiques.

9. pour installer python il suffit de suivre les instructions sur ce lien `https://www.python.org/downloads/` et pour pandas executer la commande "pip install pandas"

# Chapitre 3

# Évaluation & Discussion

## 3.1 Évaluation des architectures

Après avoir généré plusieurs architectures comme expliquée plus précédemment, nous avons effectué des statistiques sur leurs performances. En voici quelques-unes :

- La figure 3.1 représente les 10 meilleures configurations :

| | num_layers | layer_1 | layer_2 | layer_3 | layer_4 | best_train_performance | best_test_performance | train_fct |
|---|---|---|---|---|---|---|---|---|
| 323 | 4 | 6 | 17 | 12 | 11 | 0.000695 | 0.000445 | trainlm |
| 117 | 2 | 12 | 15 | 0 | 0 | 0.000588 | 0.000468 | trainscg |
| 184 | 3 | 16 | 12 | 3 | 0 | 0.000160 | 0.000515 | trainrp |
| 290 | 1 | 17 | 0 | 0 | 0 | 0.000665 | 0.000518 | traincgp |
| 387 | 3 | 15 | 5 | 4 | 0 | 0.000638 | 0.000539 | trainlm |
| 89 | 1 | 3 | 0 | 0 | 0 | 0.000574 | 0.000541 | trainscg |
| 239 | 2 | 6 | 4 | 0 | 0 | 0.000371 | 0.000548 | trainlm |
| 176 | 1 | 19 | 0 | 0 | 0 | 0.000403 | 0.000550 | trainrp |
| 385 | 3 | 15 | 5 | 4 | 0 | 0.000661 | 0.000551 | trainscg |
| 187 | 3 | 16 | 12 | 3 | 0 | 0.000503 | 0.000566 | trainlm |

FIGURE 3.1 – Les meilleures configurations

On remarque que 9 des 10 meilleurs résultats ont moins de 3 couches seulement où le nombre de neurones de la première couche est relativement élevé (12 en moyenne) par rapport au nombre de neurones de la deuxième couche et troisième (6 et 2 en moyenne respectivement). On peut donc en conclure qu'une complexité élevée d'un réseau de neurones n'améliore pas forcément les performances.

- La figure 3.2 représente la moyenne des performances sur les données de test par fonction d'entrainement :

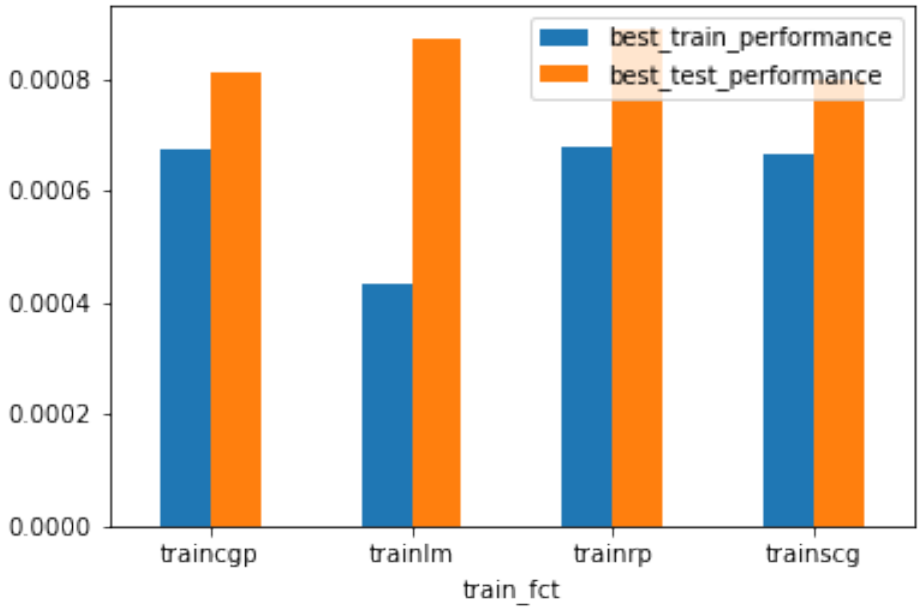


FIGURE 3.2 – Moyenne de performances par fonctions

On remarque que la fonction trainlm engendre une bonne performance sur les données de tests, mais de mauvaises performances sur les données de tests, ceci est nous dit que la fonction trainlm souffre du problème d'overfitting. Par contre, la fonction traincgp présente des résultats sur les données de tests et performances assez bons et sans grande différence.

## 3.2 Choix de l'architecture

Étant donné les résultats des tests, nous avons choisi d'adopter l'architecture suivante :

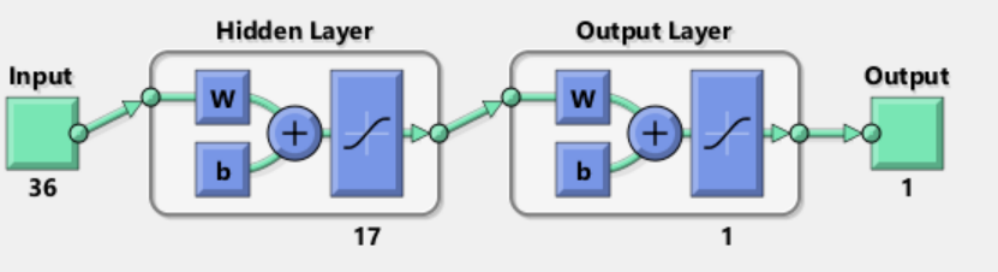


FIGURE 3.3 – Architecture adoptée

Avec la fonction Traincgp car cette configuration a donné de bons résultats sans sur-apprentissage.

## 3.3 Résultats finaux

Après avoir ajusté les hyper-paramètres, voici les résultats finaux.

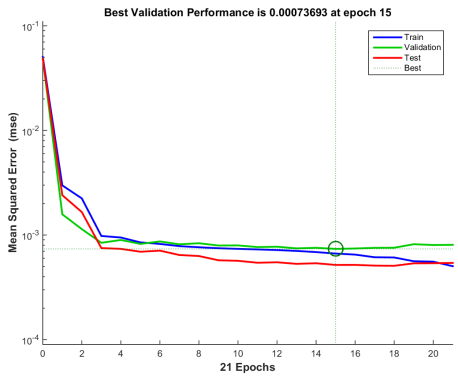


FIGURE 3.4 – Graphe de performance

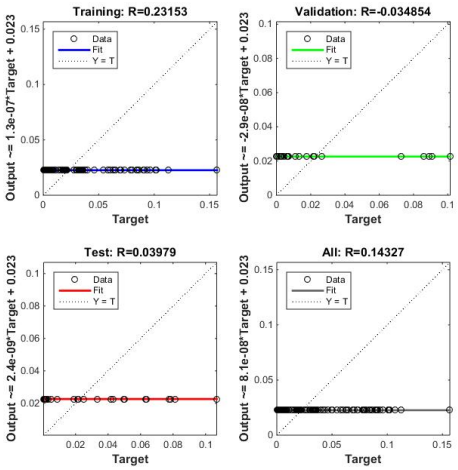


FIGURE 3.5 – Graphe régression

## 3.4 Discussion

Ayant choisi et évaluer la meilleure configuration, il est temps maintenant de répondre à notre question de départ, à savoir est ce que le régime alimentaire influe sur les contaminations/décès de COVID-19, d'après les résultats des performances obtenu dans la plupart de notre architecture on peut en déduire que oui le régime alimentaire influe sur les décès de COVID-19.

Néanmoins, les résultats de régression on était plutôt mauvaises ce qui est surement du a la faible taille des données comme citer dans la section 2.3.3, il faut aussi noter que les résultats peuvent être biaisé, étant donnée que des données complètes ne sont pas encore disponible pour COVID-19, et surtout en considérant que les données d'apport alimentaire date de 2018 il pourrait avoir changés entre cette dernière et le début de la pandémie.

## 3.5 Conclusion

On a pu concevoir un modèle qui converge sur les données de nutrition afin de prédire le taux de mortalité (nombre de morts dû à la covid par habitant). Ce qui nous permet d'affirmer qu'il existe une relation entre les deux et de confirmer les hypothèses des biologistes.

# Répartition des taches

### Partage des taches pour le script matlab et pretraitement

| Tache | Répartition |
|---|---|
| Comprendre le problème et les données | Ensemble |
| **Pre-traitement des données** | |
| Brainstorming et décider de l'approche à prendre | Ensemble |
| Prétraitement des données avec python | Khouas |
| Jointure des Datasets | Khouas |
| **Script Matlab** | |
| Brainstorming et décider de l'approche à prendre | Ensemble |
| Script du choix des architectures | Djebara |
| Script pour entrainer et générer les réseaux de neurones | Djebara |
| Exécution et évaluation des résultats (graphes, etc) | Djebara |

TABLE 3.1 – Partage des tâches pour le prétraitement et script MATLAB

### Répartition des taches pour la rédaction du rapport

| Tache | Répartition |
|---|---|
| État de l'art | Ensemble |
| **Conception & Implémentation** | |
| Description de l'approche | Djebara |
| Description des données | Khouas |
| Pré-traitement des données | Khouas |
| Conception du réseau de neurones | Djebara |
| Organisation & utilisation du code | Ensemble |
| **Évaluation & Discussion** | |
| Évaluation des architectures | Djebara |
| Choix de l'architecture | Djebara |
| Discussion | Khouas |

TABLE 3.2 – Partage des tâches pour la rédaction du rapport

noter que malgré la division du travail, nous avons adopté une méthode de travail flexible ou chacun continuait à donner des feedbacks et aider dans le travail de l'autre.

# Bibliographie


[1] European Centre for Disease Prevention and Control. "Daily number of new reported COVID-19 cases and deaths worldwide". https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide, 2020. consulté le 12-06-2020.

[2] GDD. "Set of dietary factors in GDD 2018". https://www.globaldietarydatabase.org/, 2021. consulté le 12-06-2020.

[3] Coronavirus disease (COVID-19) : How is it transmitted ? https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-how-is-it-transmitted consulté le 12-06-2020.

[4] Kalantar-Zadeh, Kamyar, and Linda W. Moore. "Impact of nutrition and diet on COVID-19 infection and implications for kidney health and kidney disease management." Journal of renal nutrition 30.3 (2020) : 179-181.

[5] Nutrition advice for adults during the COVID-19 outbreak [en ligne] http://www.emro.who.int/nutrition/news/nutrition-advice-for-adults-during-the-covid-19-outbreak.html, consulté le 24 Juin 2021

[6] Mentella, Maria Chiara, et al. "The Role of Nutrition in the COVID-19 Pandemic." Nutrients 13.4 (2021) : 1093.