**USTHB, Faculté d'Electronique et d'Informatique**
**Département d'Informatique**
**Master Systèmes Informatiques Intelligents**

**Année 2020 – 2021**
**Module Apprentissage Automatique et Réseaux de Neurones**

# Mini-projet

## Date de Remise du Mini-projet : Jeudi 24 juin 2021 (13h00)

**But du mini-projet:**

Le but de ce projet est de vous permettre d'approfondir votre compréhension des Réseaux de Neurones et de l'apprentissage automatique en mettant un accent plus particulier sur les concepts vus en cours et en TP. Et pour bien maitriser cela, il vous est demandé d'utiliser (si nécessaire après amélioration) le script que vous avez développé pour le Lab 7 afin de résoudre un problème lié à la COVID-19.

**Le problème et les données :**

Durant une discussion scientifique entre amis, vous vous êtes posé la question de savoir si le régime alimentaire (Diet en Anglais) influe sur les nombres cumulés des contaminations et des décès dus à la COVID-19. La question vous intrigant, vous avez fouiné un peu plus et, Google aidant, vous avez trouvé qu'il existe des données sur les régimes alimentaires par pays ainsi que des données sur les contaminations et décès dus à la COVID-19. Plus particulièrement, vous avez trouvé les datasets suivants :

1. La base « Global Dietary Database » développée par Tufts University aux Etats Unis ui donne énormément de données sur le régime alimentaire par pays, par sous-région, et globalement. En fait, c'est un ensemble de bases de données que vous pouvez télécharger[1] après inscription sur le site https://www.globaldietarydatabase.org/ (Sélectionnez à partir du menu « Global Dietary Data », le choix « Download Dietary Intake Estimates »). Parmi les fichiers que vous téléchargerez (dans le fichier .rar), le fichier Excel « GDD 2015 Codebook_Feb 3 2020.xlsx » vous expliquera tous les codes utilisés dans les fichiers de données (format csv). Vous utiliserez dans le répertoire « Country data » le fichier de données « all_cnty_yr_2015.csv ». Nous appellerons cette base « Base Diet ».

2. Sur le site[2] de la « European Centre for Disease Prevention and Control » vous avez trouvé un dataset qui vous donne beaucoup d'informations sur les contaminations et décès quotidiens dus à la COVID-19 ainsi que les chiffres cumulés par pays. Les informations disponibles ont commencé à différentes dates selon le pays, mais vous constatez que pour tous les pays inclus dans l'étude, les données couvrent la période allant de Mars à Décembre 2020 (pratiquement avant la vaccination[3]). Nous appellerons cette base « Base COVID ».

Bien sûr, ayant vu la disponibilité des ces données, vous vous êtes dit que vous pourrez joindre les deux bases pour voir si vous pouvez prédire, sur la base du régime alimentaire, le cumul de décès ainsi que le cumul de contaminations pendant la période d'étude comme expliqué dans le point 2 ci-dessus.

Ce mini-projet vous demande de faire le travail que vous vous êtes mis en tête[4] de réaliser comme expliqué ci-dessus. Votre travail consiste donc à faire ce qui suit :

1. Télécharger la « Base Diet » et la « Base COVID »
2. Considérer les pays qui sont communs entre les deux bases pour le travail que vous allez faire.
3. Vous avez 2 targets : les cumuls des contaminations et les cumuls décès dus à la COVID-19 pendant la période d'étude.
4. Vous allez réfléchir à quelles lignes et quelles colonnes vous allez utiliser de chaque base et comment.

---

[1] J'ai pu télécharger ces datasets mais j'ai accepté les règles de Tufts University d'utilisation pour des besoins personnels (recherche) et ai accepté de ne pas les partager/distribuer. C'est pourquoi votre binôme devra télécharger les données pour votre mini-projet.

[2] https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

[3] Nul besoin de vous dire que votre travail devrait exclure la période où la vaccination a commencé pour que ce facteur n'influe pas sur l'étude.

[4] Quand-bien même vous n'y avez pas pensé, c'est ce qui vous est demandé de faire ! 😉

5. A vous de voir si vous allez faire des encodages sur les données ou pas. Par exemple, allez vous faire une régression linéaire (i.e. vous essaierez d'approximer chaque cumul) ou bien allez vous diviser ces cumuls en intervalles ?
6. Après avoir traité vos données et éliminé tous les attributs et/ou lignes non pertinents à votre étude (bien sûr, des choix que vous expliquerez dans votre rapport), vous allez utiliser votre script, i.e. celui développé pour le Lab 7, ou éventuellement amélioré pour ce mini-projet, avec les données résultantes (inputs et targets, que vous aurez extraits, et peut-être traités) à partir des datasets pour trouver le meilleur réseau de neurones. Ceci devra vous permettre de répondre aux questions que vous vous êtes posées plus haut. Vous pourriez décider d'avoir deux réseaux de neurones séparés, un pour chaque cumul, ou bien un seul réseau. Mais tous ces choix doivent être expliqués dans votre rapport.

**Aspects pratiques :**

Le travail DOIT être fait en binômes, ce qui veut dire exactement deux personnes pourront travailler ensemble sur ce devoir, ni plus ni moins. **Toute autre formation sera pénalisée**.

Il vous est demandé ce qui suit :

1. Envoyez-moi un email au plus tard le Vendredi 04/06/2020 à 23h59 pour me donner la composante de votre équipe, à savoir les Nom, Prénom, Matricule, email et Groupe de chaque membre du binôme.

2. Si les deux membres d'un binôme sont du même groupe, ce binôme remettra éventuellement son rapport et CD à l'enseignant(e) de TP. Si un binôme est formé de deux étudiant(e)s qui sont dans les groupes 1 et 2 (respectivement), ou 2 et 3 (respectivement), je leur dirai à qui remettre leur rapport et CD.

3. Faites la conception de votre solution et implémentez-la en utilisant un script en MATLAB qui vous permet de faire l'apprentissage automatique de réseaux de neurones, sauvegardera dans un fichier **résultats.txt** (qui sera clairement mentionné dans le rapport) les résultats intermédiaires (comme expliqué pour le Lab 7), et sauvegardera dans le même fichier **résultats.txt** pour conclure les détails de l'architecture du meilleur réseau de neurones avec la performance et la régression sur les données de validation.

   Evidemment, votre script devra être conçu de telle sorte à retourner (dans le fichier **résultats.txt**) le meilleur modèle possible. Ceci consistera par exemple, selon le cas, à faire varier la configuration du RNs (architecture, fonctions d'apprentissage, et tous les hyperparamètres que vous voudrez/pourrez.) et de sauvegarder les résultats obtenus dans chacun des cas, en vous assurant que vous sauvegarderez l'architecture/le modèle qui vous aura donné les meilleurs résultats. Ce réseau de neurones sera nommé **covid_net** et sera remis avec les autres fichiers.

**Livrable :**

Vous remettrez un rapport sous forme de document imprimé ainsi qu'un CD.

1. Le rapport doit :
   a. Expliquer le problème traité, les données choisies, le format de ces données, le traitement qui a été fait sur les données, la représentation des données (si elle est modifiée pour être plus appropriée à une utilisation avec les réseaux de neurones), etc. Vous devrez vous assurer que votre rapport définit clairement le problème.
   b. Expliquer avec tous les détails nécessaires vos choix de conception de votre solution, et les résultats, y compris des tableaux récapitulatifs/comparatifs (le cas échéant).
   c. Inclure les lignes directrices à suivre pour utiliser votre solution.
   d. Dire qui a fait quoi de façon très précise (**des points seront déduits si la distribution des responsabilités n'est pas précisée dans le rapport**.)

   Respectez toutes les règles mentionnées dans ce document, y compris les noms de fichiers, de réseau, etc.

2. Le CD doit contenir :
   - votre rapport en format pdf ;
   - les 2 bases de données « Base Diet » et « Base COVID » d'origine ainsi que les données après traitement (le cas échéant). Placez les dans deux répertoires différents avec des noms parlants pour ces derniers ;
   - le script/programme pour implémenter votre solution et retourner le meilleur modèle ;
   - les figures (performance et régression) sauvegardées dans des répertoires bien structurés. Les noms des (fichiers des) figures et des répertoires doivent être aussi intuitifs et parlants que possible ; et
   - la « Déclaration sur le plagiat et la malhonnêteté intellectuelle » renseignée au stylo et signée par les deux étudiants. J'ai horreur du plagiat que ce soit dans le rapport ou dans le script : apprenez à compter sur vos propres efforts, développez vos compétences, et évitez de sérieux problèmes !!

**Date de remise du rapport avec le CD : Jeudi 24 Juin 2021 (13h00)**

**Pour chaque 24h de retard à partir de cette échéance, 25% de la note seront déduits (Vendredi et Samedi comptant pour 1 jour).**
**Aucun rapport ne sera donc accepté après Mardi 29 Juin 2021 à 13h00.**

**Remarque importante:**

Un bon rapport n'est pas nécessairement long ! A vous de juger comment avoir un rapport suffisamment complet qui reste aussi court que possible. En d'autres termes, il ne s'agit pas de faire du remplissage mais bien de faire une présentation écrite, scientifique et aussi claire que possible. Pour vous aider, je limite la longueur du rapport à 10 pages avec une police Times New Roman, taille 11. Les captures écrans, si vous en avez, doivent être incluses sur CD comme expliqué plus haut ; par contre, des tableaux de résultats et/ou comparatifs doivent être inclus dans le rapport. Sur le rapport sur CD, donnez avec chaque entrée/résultat dans le tableau un lien vers la capture écran correspondante sur le CD.

La note de votre binôme au projet sera relative à celle des autres binômes !!

**Notation du mini-projet :**

| Partie du mini-projet | Grille de notation du mini-projet |
|---|---|
| Rapport | 30 % |
| Création des données | 20 % |
| Qualité du script | 25 % |
| Résultats de l'exécution du script | 15 % |
| Analyse et discussion des résultats en termes du problème posé (relation entre le régime alimentaire et la COVID-19). | 10 % |

**N.B. :**
Je n'accepterai aucune remise du projet ou parties du projet par email ; vous devrez tout remettre dans ma boite postale ou celle de Mme Belhadi (selon le cas) au niveau du département. (Il se pourrait que je vous demande de le faire sur Moodle plutôt que sur CD.)

**Bonus** :
Toute innovation intéressante (non requise dans cet énoncé) sera récompensée. A vous de la mettre en exergue et de l'expliquer.

***Bon courage!***