EPU GB5 **Introduction à l'IA pour la biologie** novembre 2022

Feuille de route pour le projet

Ce document fournit quelques lignes directrices pour le projet d'*Intelligence artificielle* à faire par groupe de 2 (ou 3). Notez que le projet représente une partie importante de votre note finale et que vous devez donc y consacrer un temps plus que raisonnable. Il est cependant difficile de quantifier avec précision le temps total que vous devriez consacrer au projet et à sa rédaction...

Quelques exemples de sources de données :
- données de production solaire :
  `https://www.kaggle.com/c/ams-2014-solar-energy-prediction-contest`
- données de consommation/production électrique en europe :
  `https://opendata.reseaux-energies.fr/pages/accueil/`
- données macro-économiques/sociales :
  `https://data.oecd.org/fr/energie.htm`, `https://www.insee.fr/`
- données de fréquentation de station vélib. à Paris & à Lyon :
  `https://maxhalford.github.io/blog/a-short-introduction-and-conclusion-to-the-openbikes-2016-challenge/`
- données météo :
  `https://www.ncdc.noaa.gov/cdo-web/`
- données de pollution atmosphérique :
  `https://www.epa.gov/outdoor-air-quality-data`,
  `http://www.openair-project.org`
  `https://data-airparif-asso.opendata.arcgis.com/`
- open data gouvernemental :
  `http://www.data.gouv.fr/fr`
- données de transport :
  `http://www.data.gouv.fr/fr/datasets/trafic-annuel-entrant-par-station-2013/`
- données financières :
  `https://cran.r-project.org/web/packages/tidyquant/index.html`
- plateforme de compétition de data science :
  `https://www.kaggle.com`,
  `https://challengedata.ens.fr/`,
  `https://www.datascience.net/fr/home/`,
  `https://www.crowdanalytix.com/community`
- données de consommation de bâtiments tertiaires :
  `https://github.com/buds-lab/the-building-data-genome-project`

N'hésitez pas à sortir des sentiers battus et à aller chercher d'autres données sur d'autres sites web (ces pointeurs pourront venir compléter la liste précédente).

Proposition de projet, à rendre avant les vacances de Noël. Votre proposition de projet doit être dactylographiée et faire 1 page de longueur au maximum, en interligne simple. L'objectif de la proposition est de s'assurer que vous êtes sur la bonne voie et de donner suffisamment d'informations pour éventuellement corriger le tir.

Dans votre proposition, vous devez couvrir les éléments suivants :
- Titre préliminaire et liste des étudiants travaillant sur le projet (2 ou 3 étudiants par projet).
- Résumé qui devra apparaître dans votre compte-rendu de projet. Il devrait faire un paragraphe de long, pour le moment peut-être seulement 5-15 lignes. Il doit fournir une description rapide de votre projet et définir vos objectifs principaux.
- Description succincte de ce que vous envisagez de faire.
  - Quel problème tentez-vous de résoudre ?
  - Comment formulez-vous le problème en tant que problème de machine learning (par exemple, s'agit-il de classification supervisé, de classification non supervisée, de régression, de prédiction, etc.) ? Que voulez-vous exactement prévoir (pour les tâches de prédiction) et comment évaluerez-vous vos résultats ? Comment saurez-vous si vos résultats sont bons ? À quoi pouvez-vous les comparer ? Il est essentiel que votre problème soit bien défini.
  - Quels ensembles de données comptez-vous utiliser ? Si vous devez effectuer un travail important pour obtenir les données ou les convertir au format approprié, décrivez le processus et l'effort approximatif requis. Combien y a-t-il d'exemples dans l'ensemble de données ? Quels

outils d'apprentissage prévoyez-vous d'utiliser et quels algorithmes prévoyez-vous d'utiliser (par exemple, arbres de décision, réseaux de neurones, etc.) ?
— Si nécessaire, donnez quelques articles scientifiques relatifs.

Idéalement, vous devriez essayer de faire quelque chose d'original, comme étudier un ensemble de données qui n'a pas été complètement évalué ou utiliser sur ces mêmes données une approche différente en montrant par exemple les différences dans les résultats obtenus. Vous devez vous assurer que votre analyse n'est pas une application directe du cours. Par exemple, exécuter une ou deux méthodes sur un ensemble de données en utilisant `scikit learn` et passer une heure à analyser puis à rédiger rapidement un résumé serait considéré comme non intéressant. Vous devez étudier l'ensemble de données, déterminer les problèmes, résoudre tous les problèmes de prétraitement, essayer plusieurs techniques de machine learning et peut-être envisager des adaptations pour améliorer les performances prédictives.

Description complète du projet (document à rendre, une semaine avant les soutenances) Le rapport que vous devez me rendre devra tenir sur 3 à 7 pages (jusqu'à 8 si vous êtes 3 sur le projet), en simple interligne. La présentation orale de vos projets en janvier (date pas encore fixée) complètera le document écrit.

Le document n'a pas besoin d'être organisé exactement comme décrit ci-dessous, mais il devrait être assez similaire, étant donné que le schéma ci-dessous est assez standard pour un document en machine learning.

1. Résumé : résume le document et les objectifs du travail (obligatoire, environ 300 mots).
2. Introduction : présente le projet et ce que vous essayez de faire. Cette partie peut si nécessaire inclure une description du contexte scientifique.
3. Présentation du domaine : en fonction du projet, il peut être important de donner les connaissances spécifiques au domaine. Par exemple, cette section peut fournir des informations sur comment la régulation de la charge électrique sur le réseau d'EDF est actuellement assurée. Si le domaine est bien maitrisé par un large public, cette section peut être omise.
4. Méthodologie expérimentale : cette partie décrit les expériences et les ensembles de données, les métriques d'évaluation, les algorithmes d'exploration de données utilisés, la méthodologie précise relative à la configuration des expériences et tout autre détail relatif aux expériences.

   Il y aura généralement une sous-section pour chacun des sous-thèmes susmentionnés.
5. Résultats : présente les résultats des expériences en machine learning, ainsi qu'une discussion et une analyse des résultats. Normalement, une section de discussion séparée n'est pas nécessaire.
6. Travaux connexes : Le travail peut être complété par une description des travaux connexes, avec citations des articles pertinents. Si vous rédigez un document sur l'application de méthodes de machine learning, cette partie n'est peut être pas pertinente, néanmoins, il devrait presque toujours y avoir des travaux connexes cités. S'il n'y en a pas beaucoup, il peut être possible d'inclure ces travaux connexes dans l'introduction (étant donné que ces travaux peuvent fournir une motivation et un contexte pour le projet). S'il y a beaucoup de travaux connexes, cette section peut être utile.
7. Conclusion : un résumé de vos principaux résultats est attendu ainsi qu'une discution sur les limites et pistes d'amélioration pour des travaux futurs.

Présentations orales : 15 min (ou 22 min 30) de présentations (chaque personne du groupe prendra la parole pour 7 mn 30) et une démonstration du code développé pour cette étude sera faite. Quelques questions seront alors posées.

Le code `scikit-learn` du projet (ainsi que les données utilisées) sera de même annoté et rendu pour tests.