Exercices et projets guidés

LES BASES DE LA SCIENCE DES DONNÉES

Entre les sessions

Session 1 à Session 2

- terminez les exercices de la session 1
- téléchargez les ensembles de données depuis le site web
- lisez le Programming Primer (sections 1 - 4)
- installez R / RStudio (Posit)
- installez les librairies R suivantes : dplyr, xts, knitr, tidyverse, ggplot2, pastecs, Hmisc, e1071, psych, quantmod, ggm, kerndwd, MASS, DMwR, ROCR, car, forcats, corrplot

De la session 2 à la session 3

- terminez les exercices de la session 2

De la session 3 à Session 4

- terminez les exercices de la session 3

Après la session 4

- terminez les exercices de la session 4
- essayez les projets guidés

Projet guidé I

Sélectionnez un projet de données qui vous intéresse et fournissez une ébauche de planification pour celui-ci, en abordant les sujets abordés dans ce cours. Les questions suivantes peuvent vous aider :

1. Quelles sont les questions associées au projet ?
2. Quel est le modèle conceptuel de la situation sous-jacente ?
3. Quel type d'ensemble(s) de données existe(nt) qui pourrai(en) vous aider à répondre à ces questions ?
4. Y a-t-il des limites aux données ou à l'analyse ?
5. Devez-vous collecter de nouvelles données pour traiter ces questions ?
6. Comment les données sont-elles stockées/accédées ? Quelles sont les exigences en matière d'infrastructure ?
7. À quoi ressemblent les produits livrables ?
8. Comment les succès seraient-ils quantifiés/qualifiés ?
9. Quels sont vos délais et votre disponibilité ?
10. Quelles sont les compétences requises pour travailler sur ce projet ?
11. Travailleriez-vous sur ce projet seul ou au sein d'une équipe ?
12. Quel serait le coût du lancement et de la réalisation de ce projet ?
13. À quoi ressemble le pipeline d'analyse des données ?
14. Quels logiciels et méthodes d'analyse seront utilisés ?

Projet guidé II

Rédigez un article discutant de certaines des questions éthiques entourant l'utilisation de l'intelligence artificielle, de la science des données et des algorithmes d'apprentissage automatique.

Établissez une liste des 3 principes éthiques les plus importants auxquels l'utilisation de tels algorithmes devrait se conformer. Expliquez pourquoi vous avez choisi chacun de ces principes.

Décrivez (au moins) 2 cas réels d'utilisation de l'I.A./S.D./A.A. dans le secteur public, le secteur privé, ou le milieu universitaire, lorsque les principes éthiques que vous avez choisis ont été violés. Discutez de la manière dont le non-respect de vos principes éthiques a occasionné (ou pourrait occasionner) des dommages à des personnes, des organisations, des pays, etc.

Suggérez comment les projets discutés ci-dessus auraient pu être modifiés afin que leur utilisation des algorithmes I.A./S.D./A.A. respecte les principes éthiques que vous avez choisis.

Projet guidé III

Ce projet utilise l'outil Gapminder (il y a aussi une version hors-ligne)

1. Prenez le temps d'explorer l'outil. Dans la version en ligne, le point de départ par défaut est un graphique à bulles montrant l'espérance de vie en 2020, ainsi que le revenue par personne, par pays (la taille des bulles étant associée à la population totale). Dans la version hors ligne, sélectionnez l'option "Bubbles".
2. Pouvez-vous identifier les catégories de variables disponibles, ainsi que certaines des variables? [Vous devrez peut-être fouiller un peu].
3. Pourquoi pensez-vous que Gapminder ait choisi l'espérance de vie et le revenu par personne comme variables par défaut ?
4. Remplacez l'espérance de vie par le nombre de bébés par femme. Observez et discutez des changements par rapport au graphique par défaut.
5. Formulez quelques questions auxquelles vous pourriez répondre avec les données par défaut.
6. Formulez quelques questions auxquelles vous pourriez répondre en utilisant certaines des autres variables.
7. À quel moment du "flux de travail de la science des données" pensez-vous que des visualisations de cette nature pourraient être utiles ?
8. Ces visualisations permettent-elles de bien comprendre le système étudié (la Terre géopolitique) ?

Projet guidé III (suite)

9. Quelles sont, selon vous, les sources de données de l'ensemble de données sous-jacent? [Vous devrez peut-être fouiller sur Internet pour y répondre].
10. Toutes les variables et mesures sont-elles dignes de confiance? Comment pouvez-vous le déterminer?
11. L'ensemble de données sous-jacent est-il structuré ou non structuré?
12. Fournissez un modèle de données (“data model”) potentiel pour l'ensemble de données sous-jacent.
13. Quels sont les types des 4 variables par défaut (espérance de vie, revenu, population, régions)?
14. Jouez un peu avec les graphiques. Pouvez-vous trouver des paires de variables qui sont positivement corrélées? Négativement corrélées? Non corrélées?
15. Parmi les variables qui sont corrélées, certaines vous semblent-elles présenter une relation dépendante-indépendante? Comment pouvez-vous identifier de telles paires?
16. Pouvez-vous fournir une estimation visuelle de la moyenne, de la médiane, et de l'étendue de diverses variables numériques?
17. Pouvez-vous estimer à vue d'œil le mode des variables catégorielles?
18. Pouvez-vous identifier des moments spéciaux (points temporels particuliers) dans les données, où un changement à longue haleine se produit, par exemple?
19. L'outil et son jeu de données sous-jacent sont-ils utilisables ? De quels facteurs dépend votre réponse ?

Projet guidé III (suite)

20. Pensez-vous qu'il pourrait y avoir des problèmes avec les valeurs rapportées ? Par exemple, sélectionnez la Suède et les États-Unis dans le menu de cases à cocher à droite et suivez leur parcours de 1799 à 2018/2020. À partir de quel moment les valeurs sont-elles raisonnables ? À votre avis, que se passe-t-il au début de la série chronologique ?
21. Suivez l'Érythrée pendant la même durée. Recherchez la date d'indépendance de ce pays (vis-à-vis de l'Éthiopie). A votre avis, que représentent les mesures antérieures à cette date ?
22. Suivez l'Autriche pendant la même durée. Recherchez la chronologie historique des frontières du pays (Autriche-Hongrie, Anschluss, frontières modernes, etc.). Qu'est-ce que cela implique pour les mesures rapportées ?
23. Suivez la Finlande pendant la même durée. Que se passe-t-il en 1809 ? Cela vous apprend-il quelque chose sur la façon dont les données sont codées dans l'ensemble de données ?
24. Désélectionnez tous les pays et laissez la simulation se dérouler de 1799 à 2018/2020. Pouvez-vous identifier des cas où un grand sous-ensemble d'observations se comporte de manière inattendue ? Si oui, pensez-vous que cela est dû à des problèmes de nettoyage/de traitement des données ?
25. Continuez à explorer l'ensemble de données. Vous pouvez modifier les variables affichées ou utiliser d'autres méthodes de visualisation. Globalement, pensez-vous que l'ensemble de données est fiables ? L'utiliseriez-vous pour effectuer des analyses ? Quelles sont ses forces et ses faiblesses ?

Projet guidé IV

Sélectionnez un ensemble de données dans la liste ci-dessous (ou tout autre ensemble qui vous intéresse) :

- GlobalCitiesPBI.csv
- 2016collisionsfinal.csv
- sondages_us_election_2016.csv
- HR_2016_Census_simple.xlsx

Pour votre/vos ensemble(s) de données :

1. Créez un "dictionnaire de données" pour expliquer les différents champs et variables. Pouvez-vous trouver une source pour ces ensembles de données ?
2. Dressez une liste des questions auxquelles vous aimeriez obtenir des réponses sur ces données.
3. Étudiez les variables individuelles (au moyen de graphiques simples, de statistiques univariées, etc.)
4. Répétez le processus avec des paires de variables (par le biais de graphiques simples, de distributions conjointes, d'interactions entre variables, etc.)
5. Faites-vous confiance à l'ensemble de données, ou non ? Justifiez votre réponse. Si vous ne faites pas confiance à l'ensemble de données, signalez les entrées potentiellement invalides, les observations anormales, les valeurs manquantes ou les valeurs aberrantes. Comment ces entrées doivent-elles être traitées ?
6. Votre analyse suggère-t-elle que certaines des variables devraient être transformées ? L'une des questions que vous avez élaborées à l'étape 2 soutient-elle de telles transformations ? Si c'est le cas, transformez les données de manière appropriée.