SDS 3786 – Laboratoires

Chaque rapport de laboratoire doit comporter **au maximum 6 pages**. La contribution de chacun des membres doit être donnée de **manière explicite** pour chaque rapport de laboratoire.

Labo 1: Les aspects non-techniques de l'analyse des données (9–23 sept)

1. Présentez en une page les membres de l'équipe et leurs intérêts analytiques. Identifiez les complémentarités et les lacunes.
2. En équipe, rédigez une déclaration d'éthique d'une page liée à votre rôle dans l'utilisation de l'intelligence artificielle, de la science des données et/ou des algorithmes d'apprentissage automatique. Établissez une liste d'au moins 3 principes éthiques auxquels votre utilisation de ces algorithmes devraient se conformer. Expliquez pourquoi vous avez sélectionné chacun de ces principes.
3. Téléchargez les ensembles de données `PIMENTO_CASES.xlsx` et `PIMENTO_PROGRAMS.xlsx` dans R (ou dans un autre logiciel de votre choix); vous aurez peut-être à fouiller en ligne afin de trouver une librairie vous permettant de le faire.
4. Décrivez la structure de l'ensemble de données. Créez un dictionnaire de données qui explique les différentes variables (que devrait contenir le dictionnaire?).
5. Sur la base de votre exploration, dressez une liste de questions auxquelles il pourrait être intéressant d'obtenir des réponses sur les ensembles de données.
6. En prévision des laboratoires 2 à 5, préparez une analyse/un plan de gestion de projet pour votre équipe (en gardant à l'esprit qu'un projet indépendant sera introduit après le congé d'octobre).

Diapositives: `DSE-1-Non Technical Aspects of Quantitative and Data Work-fr`
`DSE-2-Data Science Basics-fr`

Vidéos: Les aspects non-techniques de la sciences des données

- 1ière partie (24:07) – `youtu.be/KgfdS-LdRLQ`
- 2ième partie (14:01) – `youtu.be/UgRJE6GyYVM`
- 3ième partie (17:56) – `youtu.be/V5s5NSBlG5A`

Les bases de la science des données

- 1ière partie: Les préliminaires (25:04) – `youtu.be/qY8DAJF424A`
- 2ième partie: Les préliminaires (17:33) – `youtu.be/XZnp0BCbRrk`
- 3ième partie: Les cadres conceptuels (19:01) – `hyoutu.be/Zn0umEzn0o8`
- 4ième partie: Les cadres conceptuels (23:46) – `youtu.be/cVFoe-Uh6hc`
- 5ième partie: L'éthique de la science des données (10:52) – `youtu.be/LD3XCDGnM4w`
- 6ième partie: L'éthique de la science des données (16:58) – `youtu.be/scaoiQEORxc`
- 7ième partie: Le flux de travail analytique (12:20) – `youtu.be/0Rrx7sdu3pw`
- 8ième partie: Le flux de travail analytique (12:55) – `youtu.be/9ZdMpS6GTpo`
- 9ième partie: Les données et les renseignments (22:26) – `youtu.be/jmNB1m7ZydU`
- 10ième partie: Les données et les renseignments (14:16) – `youtu.be/9tZiNcNc_2s`

DUDADS: *Fundamentals of Data Insight*, chapitres 13 et 14

Labo 2: La visualization des données (24–30 sept)

Pour chacun des ensembles de données PIMENTO_CASES.xlsx et PIMENTO_PROGRAMS.xlsx:

1. Établissez une liste de variables (7 au maximum) que vous jugez essentielles à une bonne compréhension de l'ensemble de données. Justifiez vos choix.
2. Créez 2 visualisations "ridicules" et atroces; en mettant en évidence les informations inutiles et/ou les "idées" trompeuses qui en découlent.
3. Créez 4 visualisations multivariées pour l'ensemble de données en utilisant (certaines) des variables énumérées à l'étape 1.

Suggestions: d'une part, vous devriez utiliser votre compréhension du contexte afin de créer des visualizations, mais vous devriez aussi envisager la creation d'un nombre raisonnablement élevé de graphiques en utilisant une sélection aléatoire de variables afin de minimiser les chances de manquer des informations utiles ou des renseignements important. C'est une approche particulièrement importante quand on fait affaire à des ensembles comptant un nombre élevé d'attributs.

Diapositives: DVD-1-Data-Visualization-Concepts-fr.pdf
Vidéos: Les concepts de la visualisation des données

- 1ière partie: L'analyse exploratoire des données (18:47) – youtu.be/v6ukC8CHVm8
- 2ième partie: L'analyse exploratoire des données (19:33) – youtu.be/bWu5gfmeqM8
- 3ième partie: La communication et la visualisation des données (29:09) – youtu.be/6nlzkKogIkk
- 4ième partie: La communication et la visualisation des données (11:00) – youtu.be/0g6U47pcYJs
- 7ième partie: L'esthétique des graphiques (25:34) – youtu.be/YO-bkJe3rgY
- 8ième partie: L'esthétique des graphiques: Les tableaux de bord (19:46) – youtu.be/4jfAbuYiI9w
- 9ième partie (19:02) – youtu.be/GKPkcZnpWAk

DUDADS: *The Practice of Data Visualization*, chapitres 1, 2, 3, 4, 5, 6, and 9

Labo 3: Le traitement des données (1–7 oct)

Nettoyez chacun des ensembles de données PIMENTO_CASES.xlsx et PIMENTO_PROGRAMS.xlsx, en gardant en tête les concepts introduits en classe: entrées non-valides, valeurs manquantes, observations anormales, création de nouvelles variables, réduction de la dimension. Ensuite, créez des ensembles de données dérivées (en agrégeant au niveau du consulat/ambassade et de la région) ainsi que les dictionnaires de données correspondants.

Diapositives: DSE-3-Data Science Basics-fr
Vidéos: La préparation des données

- 1ière partie: La qualité et le traitement des données (21:28) – youtu.be/JJeWMm1rKH0
- 2ième partie: La qualité et le traitement des données (14:52) – youtu.be/ILE87pZpepY
- 3ième partie: Les valeurs manquantes (10:30) – youtu.be/D5aHhKVJKtE
- 4ième partie: Les valeurs manquantes (11:18) – youtu.be/gGl2WUCd-X4
- 5ième partie: Les observations anormales (19:25) – youtu.be/ydc-GKlyFAc
- 6ième partie: Les observations anormales (08:44) – youtu.be/xAYtjJ2WCMU
- 7ième partie: La dimensionnalité et les transformations de données (10:54) – youtu.be/5rqxkLixlYo
- 8ième partie: La dimensionnalité et les transformations de données (18:08) – youtu.be/gCaXNzOzOLg

DUDADS: *Fundamentals of Data Insight*, chapitre 15

Labo 4: La grammaire des graphiques et ggplot2 (22–28 oct)

Produisez (au moins) 2 visualisations ggplot2 "définitives" (à expliquer en classe) pour chacun des ensembles de données PIMENTO_CASES.xlsx et PIMENTO_PROGRAMS.xlsx, ainsi que pour vos ensembles dérivés. Vous devez mettre l'accent sur le contenu ET sur la presentation.

Suggestions: vous devriez utiliser votre compréhension du contexte et le travail effectué lors des 3 premiers laboratoires afin de créer les visualizations.

Diapositives: DVD-3-Data-Visualization-with-ggplot2-fr.pdf
Vidéos: La visualisation des données avec ggplot2

- 1ière partie: La grammaire des graphiques (08:54) – youtu.be/r52r5GRvl6Q
- 2ième partie: La grammaire des graphiques (16:06) – youtu.be/Vv1bipiJUZc
- 3ième partie: Les bases de ggplot2 (19:14) – youtu.be/jnStV-RN0D0
- 4ième partie: Les bases de ggplot2 (14:02) – youtu.be/o0CPFi8zpIs

DUDADS: *The Practice of Data Visualization*, chapitre 12

Labo 5: La mise en récit de données (29 oct–4 nov)

Raconter "l'histoire" du contexte représenté par les ensembles de données PIMENTO_CASES.xlsx et PIMENTO_PROGRAMS.xlsx et par vos ensembles dérivés, par l'entremise des analyses effectuées lors des 4 premiers laboratoires. Cette histoire sera transmise à une partie prenante de votre choix (patron, client, ministre, etc.). Présentez des explications plausibles et des informations exploitables; soutenez votre récit à l'aide de visualizations.

Suggestions: c'est surtout un travail de "vulgarization" – il est rare que les parties prenantes savent s'y prendre à un niveau technique: comment peut-on alors leur transmettre des renseignements utiles ?

Diapositives: DVD-2-Data-Visualization-Concepts-fr.pdf
Vidéos: La mise en récit de données

- 1ière partie: Les éléments de la narration (16:36) – youtu.be/v6ukC8CHVm8
- 2ième partie: Les éléments de la narration (27:38) – youtu.be/bWu5gfmeqM8
- 3ième partie: Histoires et illustrations (13:19) – youtu.be/6nlzkKogIkk
- 4ième partie: Histoires et illustrations (08:23) – youtu.be/0g6U47pcYJs
- 5ième partie: L'évolution d'une mise en récit de données (19:07) – youtu.be/YO-bkJe3rgY
- 6ième partie: L'évolution d'une mise en récit de données (13:34) – youtu.be/2FYxDULWvr4
- 7ième partie: L'anatomie des tableaux de bord narratifs (12:18) – youtu.be/_rfwJlAC74o

DUDADS: *The Practice of Data Visualization*, chapitres 7 et 8

Labo 6: L'apprentissage automatique et les règles d'association (5–18 nov)

Effectuez une analyse de règles d'association des ensembles de données PIMENTO_CASES.xlsx et PIMENTO_PROGRAMS.xlsx ou des ensembles dérivés (vous devrez "catégorifier" les variables numériques). En utilisant soit la méthode de la force brute, soit l'algorithme apriori, déterminez 10 à 20 règles d'association fortes. Visualisez-les et interprétez leurs résultats. Certaines d'entre elles sont-elles exploitables ?

Suggestions: la taille de vos ensembles de données affectera le temps de calcul – à garder en tête! Utilisez la mise en récit de données.

Diapositives: IML-1-Statistical Learning-fr, IML-2-Association-Rules-Mining-fr.pdf IML-5-Issues and Challenges-fr.pdf
Vidéos: *L'apprentissage statistique

- 1ière partie (12:26) – youtu.be/vjwtpjjjGE4
- 2ième partie (20:08) – youtu.be/t2MhwYS850U
- 3ième partie (18:57) – youtu.be/xWkbivjva2g

L'extraction de règles d'associations

- 1ière partie: Aperçu (18:17) – youtu.be/jDOqWQwFhuk
- 2ième partie: Étude de cas (11:55) – youtu.be/TUEeCFkQb-0
- 3ième partie: Concepts de règles d'associations (21:33) – youtu.be/SyY0ys7PNW4
- 4ième partie: Concepts de règles d'associations (17:33) – youtu.be/UPZWxk99aVA

DUDADS: *Spotlight on Machine Learning*, chapitre 19 (sections 19.3 et 19.7.1)

Labo 7: La classification et la régression (19 nov–2 déc)

Dans ce laboratoire, vous utiliserez un ensemble dérivé des données PIMENTO où les rangées représentent les consulats et/ou les ambassades, et où la variable réponse/cible est le nombre d'employé.e.s moyen.ne.s par année (que vous aurez à calculer à même les données originales): vous pouvez aborder le problème sous l'angle de la classification (en "catégorifiant" la variable cible) ou sous l'angle de la régression.

1. Mettez de côté 20% des observations dans un ensemble de validation.
2. Créez une paire formation/test sur les 80% d'observations restantes et formez un arbre de décision; un classificateur de Bayes naïf (pour la classification); un réseau neuronal artificiel; un modèle de régression (pour la régression), et une machine à vecteurs de support afin de prédire la variable cible. Évaluez les performances de chaque modèle. Quels sont les modèles les plus performants sur votre paire formation/test ?
3. Répétez l'étape précédente sur au moins 20 nouvelles paires formation/test. Évaluez les performances de chaque modèle.
4. Combinez (comment ?) tous les modèles obtenus à l'étape précédente afin d'établir une prédiction pour les observations de l'ensemble de validation.

Suggestions: présentez vos résultats à l'aide d'une mise en récit de données.

Diapositives: IML-3-Classification-fr
Vidéos: La classification

- 1ière partie: Aperçu (16:43) – youtu.be/rFlUkTl8B90

- 2ième partie: Étude de cas (20:13) – youtu.be/rFlUkTl8B90
- 3ième partie: Arbres de décision et autres algorithmes (12:20) – youtu.be/zDcXLaa3cLg
- 4ième partie: Arbres de décision et autres algorithmes (18:55) – youtu.be/aK7chdM5-4Y
- 5ième partie: Évaluation de la performance (12:46) – youtu.be/r-VD3N5P5bM

DUDADS: *Spotlight on Machine Learning*, chapitres 19 (sections 19.4 et 19.7.2); *20, et *21

Labo 8: Le regroupement (19 nov–2 déc)

Dans ce laboratoire, vous utiliserez un ensemble dérivé des données `PIMENTO` où les rangées représentent les missions (consulats et/ou ambassades) afin de trouver des groupes naturels dans ces dernières.

1. Mettez à l'échelle/transformez les données (il y a plus à faire qu'il n'y paraît, nous en discuterons en classe).
2. Exécutez l'algorithme des $k-$moyennes sur les données mises à l'échelle, en utilisant TOUTES les caractéristiques, pour k= 3, ..., 10. Utilisez l'indice de Davies-Bouldin et l'indice Within-SS pour déterminer le nombre optimal de grappes. Le schéma de regroupement optimal est-il plausible ?
3. Réduisez la dimension de l'ensemble de données (en effectuant une analyse en composantes principales et en conservant les composantes principales qui expliquent jusqu'à 80% de la variabilité des données). Répétez l'étape précédente. Les résultats sont-ils significativement différents ?
4. Écrivez une routine qui sélectionne un certain nombre de caractéristiques, un ensemble de caractéristiques et un algorithme de regroupement de manière aléatoire ($k-$moyennes, DBSCAN, regroupement hiérarchique, regroupement spectral, etc., avec des choix de paramètres aléatoires), et qui produit et enregistre une affectation de regroupement pour chaque mission, ainsi que certaines métriques de regroupement internes (de votre choix).
5. Exécutez votre routine un certain nombre de fois (50 ? 100 ? 200 ? 500 ?) sur l'ensemble de données mis à l'échelle (mais non réduit par l'ACP). Produisez une matrice de similarité qui mesure le pourcentage de fois où chaque paire de missions se trouve dans le même groupe. Y a-t-il des missions qui semblent se retrouver dans le même groupe plus de 95
6. Sur la base de vos résultats, dressez une liste des missions dont les données semblent indiquer qu'elles sont de véritables pairs.

Suggestions: présentez vos résultats à l'aide d'une mise en récit de données.

Diapositives: IML-4-Clustering-fr
Vidéos: Le regroupement

- 1ière partie: Aperçu (19:23) – youtu.be/jXr-7h_Im0g
- 2ième partie: Étude de cas (11:32) – youtu.be/2xtrwjaisVE
- 3ième partie: $k-$moyennes et autres algorithmes (16:04) – https://youtu.be/8qvBJs5fS_A
- 4ième partie: Validation et commentaires (10:06) – https://youtu.be/yaMItI2C3dI
- 5ième partie: Validation et commentaires (08:37) – https://youtu.be/YTBmLAZAeA4

DUDADS: *Spotlight on Machine Learning*, chapitres 19 (sections 19.5 et 19.7.3); *22

Projet final: tout au long du semestre, vous travaillerez (en groupe) sur un projet de science des données de votre choix (vous devez d'abord obtenir mon approbation, cependant). Nous discuterons des modalités et des détails en classe et sur Slack lors des 2 premières semaine.

- Choix d'équipe: 20 sept
- Proposition: 21 oct [5]
- Rapport de progrès: 11 nov [5]
- Présentation finale: à déterminer [10]
- Rapport final et portefolio: à déterminer [10]