Mini-projet
Exploration de données et Apprentissage

Julien Blanchard

Novembre 2023

Consignes

- **Le travail est à réaliser en groupes de 2**
- **Les noms des étudiants du groupe et les données choisies sont à indiquer dans ce tableau avant le mercredi 13 décembre.**
- **Toute œuvre de plagiat, depuis internet ou entre étudiants, sera sanctionnée par une note nulle.**
- **Date limite de remise des travaux sur madoc dans la section "Mini-projet Exploration de données et Apprentissage" : mardi 9 janvier 2024**

Dans ce projet, vous allez explorer et analyser les données de votre choix issues de l'open data français ou d'une compétition Kaggle en cours. Vous réaliserez en particulier des modèles de scoring à partir des données en prenant soin d'évaluer les algorithmes d'apprentissage comme il se doit. Vous retranscrirez votre démarche d'analyse et ses résultats au sein d'une narration.

Vous pouvez réaliser le projet en R ou Python. Les notebooks sont préférés mais pas obligatoires. Des plateformes comme Deepnote, colab, cocalc et RStudio Cloud vous fourniront des ressources pour vos traitements et vous aideront à collaborer au sein du groupe.

1 Choix du jeu de données

- Vous devez étudier un jeu de données qui contient au moins 3000 lignes et 8 variables.
- Vous pouvez croiser plusieurs jeux de données pour produire des informations plus intéressantes.
- Chaque binôme travaille sur son propre jeu de données. Vérifiez dans ce tableau que le jeu de données n'a pas déjà été choisi.
- Pour choisir les données à étudier :
  - Utilisez l'open data **français** :
    - Santé Publique France
    - Ministère de la Santé
    - L'Assurance Maladie
    - Pôle emploi
    - Urssaf
    - SNCF, RATP
    - Ministère des Transports
    - La Poste
    - Marques et brevets à l'INPI
    - Tribunaux de Commerce
    - Ministère des Finances
    - Education Nationale
    - Ministère de l'Enseignement Supérieur et de la Recherche
    - Ministère de la Culture
    - Ministère de l'Intérieur
    - RTE, ENEDIS
    - Ministère de la Transition écologique
    - Ministère de l'Agriculture et de l'Alimentation
    - IFREMER
    - Open Data Paris
    - Open Data Nantes Métropole, Loire Atlantique, et Pays de la Loire
    - ...
  - Ou participez à une compétition **en cours ou très récente** de Kaggle. Les compétitions des catégories "Getting started" et "Playground" ne sont pas éligibles.

2 Travail demandé

1. Présentez vos données : source, conditions de recueil, nombre de fichiers, nombres de variables et d'individus, expliquer l'activité derrière les données...Donnez la signification des individus et des variables.
   Énoncez les questions auxquelles vous souhaitez répondre durant le projet (ou l'objectif de la compétition si vous travaillez avec Kaggle). Ces questions pourront être affinées par la suite. Les questions peuvent porter par exemple sur :
   - l'identification de liens entre variables,
   - la possibilité d'expliquer une variable en fonction d'autres,
   - la prédiction d'un évènement,
   - la mise en évidence de groupes homogènes dans les données pour leur donner du sens.

2. Préparez les données (sélection des individus et variables, transformation de variables, création de nouvelles variables...). Cette phase est plus ou moins conséquente suivant les données étudiées. N'hésitez pas à échantillonner si vous constatez dans les phases d'analyse des temps de calcul rédhibitoires, surtout pour débuter votre projet. Vous pouvez aussi cibler les données selon un critère géographique ou temporel par exemple.

Remarque : la préparation des données peut aussi intervenir au fur et à mesure de l'étude. Adaptez-vous aux données et aux besoins de l'analyse.

3. Faites l'analyse statistique exploratoire 1D et 2D à l'aide des représentations dédiées. Repérez les phénomènes remarquables, atypiques, voire anormaux dans les données (distributions multimodales ? asymétriques ? valeurs extrêmes ? manquantes ? classe déséquilibrée ? variables dépendantes ? ...). C'est une démarche exhaustive, mais dans votre rendu vous ciblez les variables que vous jugez intéressantes ou remarquables.

4. Explorez et analysez le jeu de données à l'aide des techniques d'apprentissage de votre choix. Commencez par une phase **exploratoire** (réduction de dimension, visualisation, clustering...) avant d'aller vers des méthodes **supervisées** pour prédire/expliquer les variables qui vous intéressent (y compris la figure imposée ci-dessous). N'oubliez pas d'évaluer la **qualité des résultats**. Retranscrivez les informations intéressantes dégagées dans une narration, et **répondez aux questions** posées à l'étape 1. S'il s'agit d'une compétition Kaggle, décrivez votre démarche pour atteindre l'objectif.

   Figure imposée :
   A partir des données, réalisez plusieurs modèles de scoring sur les variables [1] cibles de votre choix. Essayez différents algorithmes et/ou différents paramétrages. Evaluez les modèles et interprétez vos résultats.

3 A remettre sur Madoc

Déposez de préférence un notebook structuré intégrant le code source, ses résultats et le commentaire sur le travail, avec en particulier tous les points 2.1 à 2.4. Le notebook doit aussi contenir :
- une analyse réflexive tirant le bilan du travail réalisé, des difficultés rencontrées et des compétences acquises ;
- la proportion de travail réalisé par les personnes au sein du groupe (à décider collégialement et à décliner par tâche éventuellement).

A défaut de notebook, vous pouvez déposer :
- votre code source commenté,
- un rapport pdf structuré qui reprend les éléments ci-dessus.

Fournissez le notebook ou le code source dans une archive avec l'arborescence appropriée et les fichiers nécessaires à l'exécution, **en particulier les données** (ou une URL si trop volumineux). Ne purgez pas les sorties du notebook (résultats produits, graphiques générés...).

Si votre notebook est en ligne, fournissez l'URL mais déposez quand même le notebook sur Madoc.

4 Principaux critères d'évaluation

- La rédaction est claire, les choix sont justifiés.
- Les données sont présentées, les étudiants se sont documentés sur le domaine des données.
- Les outils d'analyse, techniques d'apprentissage et méthodes d'évaluation sont bien choisis.
- La mise en oeuvre en R/Python est juste et commentée.
- Les tendances remarquables sont commentées, les résultats sont interprétés et discutés.
- Qualité de la narration (data storytelling) : les objectifs sont indiqués, le cheminement du processus d'analyse est détaillé, les résultats sont remis dans le contexte des données.

1. Si vous n'avez pas de classe binaire, créez-en une en binarisant une variable catégorique, ou en discrétisant une variable numérique en deux classes.

Cependant, étant donné que les tâches de scoring sont imposées (voir 2.4), vous pouvez les sortir du récit si elles ne s'intègrent pas naturellement.

Les caractéristiques des données (volume, complexité, nécessité de prétraiter les données) ainsi que l'étendue de l'analyse menée seront également prises en compte.