

Lancement miprojet

MiProjet

Alice HELIOU and Vincent THOUVENOT

Laboratoire de Data Science de ThereSIS



Content

1. Déroulement

2. Cas d'usage

3. Ce qui est attendu

Déroulement

Paris-Saclay University – Lancement projet

Déroulement

- Sessions dédiées: semaines du 19/01, et du 09/02
- rendu des notebooks individuels avant le cours de la semaine du 16/02 donc avant le 18/02 pour le groupe 1 et avant le 20/02 pour le groupe 2
- Rendu note de mi-projet: semaine du 09/03

Déroulement

- Sessions dédiées: semaines du 19/01, et du 09/02
- rendu des notebooks individuels avant le cours de la semaine du 16/02 donc avant le 18/02 pour le groupe 1 et avant le 20/02 pour le groupe 2
- Rendu note de mi-projet: semaine du 09/03
- Début du projet final avec création des trinomes la semaine du 09/03, sessions dédiées : 09/03 et 23/03
- Rendu des notebooks finaux avant le cours de la semaine du 30/03 donc avant le pour le 01/04 groupe 1 et avant le 03/04 pour le groupe 2
- Soutenance: sur les créneaux de cours de la semaine du 7 avril

Déroulement

- Sessions dédiées: semaines du 19/01, et du 09/02
- rendu des notebooks individuels avant le cours de la semaine du 16/02 donc avant le 18/02 pour le groupe 1 et avant le 20/02 pour le groupe 2
- Rendu note de mi-projet: semaine du 09/03
- Début du projet final avec création des trinomes la semaine du 09/03, sessions dédiées : 09/03 et 23/03
- Rendu des notebooks finaux avant le cours de la semaine du 30/03 donc avant le pour le 01/04 groupe 1 et avant le 03/04 pour le groupe 2
- Soutenance: sur les créneaux de cours de la semaine du 7 avril
- note finale =
$$\frac{miprojet + notebookfinal + 2xsoutenance}{4}$$

Cas d'usage

Cas d'usage

Chest X ray NIH 14 Dataset

<https://www.kaggle.com/datasets/nih-chest-xrays/data>

Dataset d'image de rayons X, fourni avec de nombreuses informations sur les personnes (notamment age, genre, maladies)



Ce qui est attendu

Mi-parcours

Analyse des données et pre-processing

Pour le mi-parcours nous ne considérerons pas les images et vous vous concentrerez sur les métadonnées. Nous avons attribué à chacun de vous un sous-ensemble des métadonnées de 15000 individus, à vous d'en faire l'analyse. Ensuite vous devrez appliquer une méthode de pré-processing pour y réduire les biais identifiés.

Evaluation

Un notebook à rendre, décrivant votre analyse et construisant votre dataset transformé pour réduire les biais. Le notebook devra être clairement documenté, une attention particulière sera apportée à la rédaction et aux explications du code.

Mi-parcours

Barême

- 2/20 : Introduction (explication du cas d'usage et de l'objectif)
- 2/20 : Préparation de la donnée (gestion des valeurs abhorrantes, transformation des colonnes etc.)
- 7/20 : Analyse descriptive et observation des biais
- 3/20 : Méthode de mitigation des biais par pré-processing
- 1/20 : Conclusion
- 5/20 : Soins apportés à la présentation, à l'explication et aux commentaires tout au long du notebook

Projet complet

Cas d'usage complet (sauf le in-processing)

Pour le projet final nous vous fournirons du code pour apprendre un model sur les images qui prédit si la personne est malade ou non. Vous devrez analysez l'impact du pre-processing sur le modèle et utiliser les approches de post-processing pour réduire autant que possibles les biais identifiés.

Evaluation

- Un notebook (ou plusieurs, ou autre forme de code) décrivant votre analyse complète, les différentes approches et résultats obtenus.
- Une soutenance, expliquant les difficultés, observations, et résultats.

Notation

- 1/4: Note d'étape intermédiaire (individuelle):
 - Introduction de la problématique
 - Analyse descriptive des métadonnées (avec l'interprétation)
 - Pre-processing pour réduire les biais
 - Premières conclusions sur le pré-processing
- 1/4: Rapport final:
 - Application de méthodes de in/post processing
 - Apprentissage des premiers modèles de classification
 - Analyses des approches utilisées
 - Conclusion
- 1/2: Soutenance de 15 minutes, (10' présentation, 5' de questions)