Projets DataScientest

Étapes importantes d’un projet Data

- État de l’art
  - se renseigner sur les modèles qui existent dans et hors de la recherche
- Exploration des données
  - comprendre les données
- Modélisation des données
  - quel niveau d’agrégation?
  - quelles variables?

Étapes importantes d’un projet Data

- Visualisations
  - constitution d’un jeu de graphiques pour des présentations futures
  - compréhension et communication sur les données
- Premier modèle
  - modèle simple, naïf
  - doit constituer un socle pour la suite
- Itérations et évaluations
- Présentation et ouvertures

Objectifs des projets DataScientest

- Confrontation à un vrai problème de Data Science / Data Analysis
  - Choisir son projet et sa problématique
  - Identifier les données nécessaires
  - Apporter une expertise métier
  - DA : Analyser de manière approfondie une problématique en s'appuyant de les données et des arguments statistiques ou ML.
  - DS : Modéliser le problème de ML de manière approfondie (données, choix des algorithmes, …)
  - Apporter un regard critique sur des résultats

- DA : 9 semaines pour mettre en application les connaissances apprises. Volume d'effort attendu : 50 H.

- DS : 11 semaines pour mettre en application les connaissances apprises Volume d'effort attendu : 120 H.

Choix de sujets

- Option A : Choisir un sujet dans le catalogue (Nouveau !)
- Option B : Proposition libre de sujet
  - en rapport avec une activité que vous connaissez un minimum
  - qui représente un enjeu
  - basé sur des données (!) disponibles (!!)

Conseils pour ceux qui souhaitent proposer un sujet

- Données disponibles facilement
- Connaissance du domaine (plus facile pour expliquer les résultats et les choix de modélisation)
- Volumétrie des données raisonnable
  - >5K observations sur des données tabulaires
  - > 10K observations pour des images
  - Grands corpus pour des données textes
- Intérêt pour le sujet

Présentation de quelques sites (liens sur slack)

Création des équipes

- Trinômes (exceptionnellement binômes)
- Formulation de 4 voeux ordonnés à partir des fiches du catalogue
  - *Alloué d'office en cas de non-réponse.*
- Allocation des groupes projets réalisée par l'équipe DataScientest en fonction des préférences de chacun

Livrables attendus

- Remise d’une fiche d’exploration des données
  - Template guide fourni
  - À remplir pour chacune des Tables x Variables
  - Pour bien prendre en main et comprendre ses données

- Présentation de visualisations intermédiaires
  - 3 à 5 graphiques qui apportent des informations sur le jeu de données
  - Les visualisations doivent être analysées et approfondies

Livrables attendus

- Remise d’un rapport de fin projet et codes correspondants
- Soutenance Finale et démonstration
  - *Sous réserve de validation du rapport*

Livrables - Aide

- Des documents templates génériques pourront vous guider
  - Attentions aux projets sur données non structurées
- 1 Mentor DataScientest dédié à chaque projet qui pourra adapter les attendus de chaque livrable en accord avec la direction pédagogique.
  - cas des données non structurés, ou projets atypiques.

Encadrement

- Support, channel slack dédié
- Mentor DataScientest
- Alumni DST (éventuels)
  - Membre externe qui pourra apporter une expertise métier (communauté alumni, réseau professionnel etc…)
- Git - vous serez aidés à maintenir une version stable de votre projet

Critères d’évaluation

- Remise des différents travaux
  - Rapport d’exploration des données (ou équivalent)
  - Data visualisation (ou équivalent)
  - Rapport de fin de projet et codes
  - Soutenance finale
  - Démo

- Rayonnement (obligatoire) :
  - GitHub : repo à héberger sur le compte organization DataScientest dédié à l’avancement du projet
  - WebApp illustrative Streamlit (ou équivalent) : à héberger sur DataScientest

Condition de validation

- Les travaux sont évalués selon plusieurs critères :
  - Qualité des rendus : pas d'obligation de résultats (!), mais une obligation d'effort
  - Gestion de votre temps et de votre effort
  - Preuve de travail en équipe
  - Communication et suivi avec votre mentor
  - Prise de recul sur le sujet

En cas de non-validation

- Si votre projet n’est pas validé:
  - Vous êtes informés avant la fin de votre formation
  - Pas de soutenances en première session
  - Médiation avec votre mentor, le chef de cohorte, et le responsable pédagogique

- Le deuxième (et dernier) jury de validation a lieu deux semaines après la fin de votre formation

DA - Calendrier

- Séance 2
  - Allocation projet
  - Installer anaconda : https://datascientest.com/environnement-python-installer-anaconda-pour-bien-demarrer
- Séance 3 (8 Octobre) : Remise de la fiche d'exploration des données
- Séance 4 (15 Octobre) : Remise des DataViz' & Présentation StreamLit et GIT

DA - Calendrier

- Séance 5 & 6 (22 & 29 Octobre) : Itérations - (Validation statistique ou modélisation ML)
- **Avant le Mercredi 4 Novembre :** Remise du rapport fin de projet et des codes
- **Avant le 10 Novembre : Validation du rapport par le jury DataScientest**

Si le rapport est validé:

- **10 Novembre : Soutenances et présentation des webapps**

Si le rapport n'est pas validé:

- **10-25 Novembre : rattrapages projets à définir en accord avec votre mentor, le chef de cohorte et le directeur pédagogique**

DS - Calendrier

- Séance 2
  - Allocation projet
  - Installer anaconda : https://datascientest.com/environnement-python-installer-anaconda-pour-bien-demarrer
- Séance 3 (8 ou 9 Octobre) : Remise de la fiche d'exploration des données
- Séance 4 (15 Octobre) : Remise des DataViz' & Présentation StreamLit et GIT

DS - Calendrier

- Séance 5-6-7 (~22/29 Octobre &) : Itérations
  - 1 - Essai d'une grande variété de modèles et optimisation d' hyperparamètres.
  - 2 - Traitement plus fin de la donnée (Features), Calibrage plus fin des modèles : analyse d'erreurs, interprétabilité de résultats, combinaison de modèles.

- **Avant le Mercredi 18 Novembre :** Remise du rapport fin de projet et des codes

- **Avant le 26 Novembre : Validation du rapport par le jury DataScientest**

DS - Calendrier

Si le rapport est validé

- **26 Novembre : Soutenances et présentation des webapps**

Si le rapport n'est pas validé

- **25 Novembre - 10 Décembre : rattrapages projets à définir en accord avec votre mentor, le chef de cohorte et le directeur pédagogique.**