Projet (40%)
Vue d'ensemble

- **Groupes de 2 à 4 personnes**, dans votre groupe de TD, ou pas. Si vous travaillez avec des étudiants d'un autre groupe de TD, alors **choisissez un chargé de TD "référent"** (il doit y avoir ~25 étudiants, c.a.d. ~8 groupes par chargé de TD).
- **Sujet : libre !!**
  Cherchez un **sujet qui vous passionne** + formez un groupe.
  Trouvez des **données** (voir liens plus loin).
  Définissez une **tâche** à accomplir assez précisément.
  Validez ce **sujet** avec votre chargé de TD **référent**. Et ensuite, go !
  (il y a un énoncé détaillé, projet-énoncé.pdf*, qui vous guide étape par étape)*
- **Rendus**:
  - un micro rapport d'étape (1 page)
  - le **code** (standardisé pour test facile)
  - un bref **rapport** (4 à 6 pages)
  - les **slides** (envoyées >2 jours avant l'oral)
- Présentations **orales**: (majorité de la note)
  - **4 min/**personne **présentation** (**slides**)
  - **1 min/**personne de questions
  (10, 15 ou 20 minutes par groupe (de 2,3 ou 4 personnes).

Planning détaillé

- Choisissez **un jeu de données** et définissez **une tâche** à accomplir, et formez un **groupe de 2 à 4 personnes.**
- **Rapport d'étape**: consignez votre groupe + lien vers les données + description de la tâche (l'objectif à accomplir) en **~1 page**. Concrètement: décrivez rapidement en quoi consistent les données, puis expliquez ce que vous comptez en faire.
  - A**vant le 7 mars**: envoyez le rapport d'étape à votre chargé de TP référent.
  - Le sujet du mail doit contenir : "[IAS-projet-validation]" (+tout le groupe en cc)
  - **Si vous avez besoin d'aide**, envoyez un **brouillon incomplet avant le 1er mars**, afin d'avoir un échange avec votre référent (ce n'est *pas* pénalisant !)
- Une fois votre rapport **validé** par le référent, **inscrivez vous:**
  https://docs.google.com/spreadsheets/d/1xClW8k6QZU7x-pgjcS2X34nVOF3lXvL4rtZbwecR1gs/edit?usp=sharing
- Séance 9 (23-24/03) = 2h de **TP-projet** : **approfondissement** (le sujet est *déjà* défini).
  - Affiner la tâche à résoudre (choix de mesures de score, etc)
  - Exploration de modèles envisageables.
  - Travail sur le code, le pipeline de traitement des données.
- Travail personnel ([février]-**mars**-avril):
  travail sur le fond= **code**, **figures**, **rapport**
- 15 avril : date limite de rendu **rapport (6 pages max) + code utilisable**
- 19 avril : date limite de rendu des **slides**
- Jeudi 21 avril : **soutenances** – présence obligatoire.

Objectifs de l'enseignement (rappel)

[**Bleu**=Algos, maths, code] [**Vert**=savoir-faire propre au ML]

- Avoir des **bases** solides en aprentissage statistique, bien maîtriser les **fondamentaux** (avec un accent sur les **méthodes probabilistes**). C'est ce qui vous permettra **d'approfondir**.
- Découvrir le **traitement automatique des langues** (TAL, ou *NLP* en anglais)
- Bien se repérer dans le **vocabulaire** du ML
- Avoir quelques **réflexes** de base, une connaissance basique du **pipeline**

Plus concrètement:

1. Connaître à fond quelques **algorithmes** (savoir en écrire le pseudo-code, les expliquer)
2. À partir d'une description intuitive d'un algo, formaliser son expression mathématique, et logicielle. En **lisant la doc**. d'un algo, savoir le coder
3. Pour un problème donné (**tâche**), trouver quelle classe de méthodes est pertinente
4. Face à des situations classiques (résultats d'une expérience), **analyser la situation** de façon **critique**, et savoir faire les bons choix

Objectifs du Projet

Objectifs :

- Approfondir la maîtrise ou découvrir de nouveaux **algorithmes** / pre-processings
- à partir de (données+problème), définir la ***tâche***, et trouver le *pipeline* adapté
- Face à des résultats d'expérience, **analyser la situation** de façon **critique**, savoir faire les bons choix

Données
quelques liens

Pour rechercher un problème un peu original:

- Une longue liste (on peut chercher dedans) de data sets déjà bien formatés https://www.kaggle.com/datasets?sort=votes
- Un moteur de recherche de data-sets: https://datasetsearch.research.google.com/
- De nombreux data sets, orientés économie: https://data.worldbank.org/
- Et beaucoup d'autres… ! (Paris open data, data.gouv, etc)

Si on veut plus travailler sur le fond des algos (dataset perçus comme "cheaps" car déjà très étudiés):

- La solution de facilité: https://scikit-learn.org/stable/datasets/index.html
- Des variantes à MNIST (un peu original, mais en restant raisonablement difficile pour de la reconnaissance d'images) : https://lionbridge.ai/datasets/mnist-datasets-for-machine-learning/
- Des régressions: https://lionbridge.ai/datasets/10-open-datasets-for-linear-regression/
- Encore une méta-liste (dont bcp de pbm trop durs) https://lionbridge.ai/datasets/ultimate-dataset-aggregator-for-machine-learning/

Autre projet: Créer son dataset ? Possible, mais ardu. À discuter au cas par cas.

Chaque groupe à un projet différent ! Certains dataset très riches peuvent être utilisés par plusieurs groupes. Dans ce cas, **il faut que les différents groupes utilisant le même dataset aient le même référent**.

Conseils d’organisation

- **Démarrer tôt**
- Profiter au maximum de votre chargé de TP (définition du sujet: discutez en pendant les TD/TP)
- Privilégier la **qualité** à la quantité,
  Privilégier la **bonne compréhension des outils** à la recherche de la “performance”
- Lire ces slides (slides-projet.pdf) et surtout **lire guide-projet.pdf**
- Quand vous êtes en peine avec la librairie **sklearn**, allez voir les corrigés des TD/TP.
  Idem pour la librairie **pandas**: aller voir les notebooks (corrigés) liés au TD *Exercices de type “prise en main d’un problème”*

  Le lien entre TDs/TPs et projet est assez fort:
- Pour démarrer / faire le rapport d’étape:
  consultez le TD6, *Exercices de type “prise en main d’un problème”*
- Lors du TP5, on verra 2 choses importantes pour le projet:
  - Un exemple de pré-traitement (*pre-processing*), la PCA
  - Un exemple d’optimisation d’hyper-paramètre
  Vous pouvez suivre la trame de ce TP pour faire votre projet.
- En TD7, *Analyse de résultats d’expérience*, on analysera des résultats, c’est le genre de graphes et d’analyse qui sont attendus dans le projet.

  Barème approximatif disponible sur le gitlab.
  Rapport d’étape: pénalité (-2 sur 20 !) si rendu en retard, +0 si tout est ok, bonus (+1,+2) si très bon