



# Description des projets

Initiation au Machine Learning

Présenté par : Mously Diaw

## Informations utiles

**Projet:** individuel ou en groupe (2 à 4 personnes)

**Consignes:** pas plus de 2 groupes sur le même sujet

**Livrables:** Les livrables sont à envoyer au plus tard **48h avant la date des soutenances**

**Mentorat:** vous pouvez me contacter pour des sessions de mentorat (difficultés rencontrées, choix méthodologique, modélisation, MLOps, ...)

**Compétences évaluées:**

-

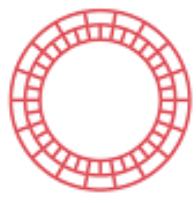


# Régression

10 propositions de projet

**REG01**

House prices prediction



# Missions et objectifs

Vous êtes data scientist chez Laplace Immo, un réseau national d'agences immobilières.

Demandez à un acheteur de décrire la maison de ses rêves et il ne commencera probablement pas par la hauteur du plafond du sous-sol ou la proximité d'une voie ferrée est-ouest. Mais il se trouve que les négociations de prix sont influencées par bien d'autres éléments que le nombre de chambres à coucher ou une clôture à piquets blancs.

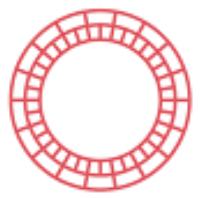
Laplace Immo souhaite que vous fournissez à ses équipes un algorithme de prédiction des prix des maisons.

Avec 79 variables explicatives décrivant (presque) tous les aspects des maisons résidentielles à Ames (Iowa, US), ce projet a pour objectif de construire un modèle de prédiction du prix final de chaque maison.

## Votre mission

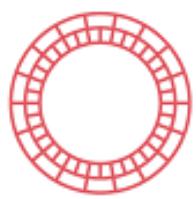
Votre mission est d'aider les équipes de Laplace Immo à avoir un simulateur des prix des maisons à vendre. Voici un récapitulatif de votre mission :

- Réaliser une analyse exploratoire.
- Tester différents modèles de prédiction afin de répondre au mieux à la problématique.
- Il faut utiliser Github Actions pour le **déploiement continu** et l'interface web 'UI MLflow" ou un autre outil d'affichage des résultats du **tracking**, concevoir des **tests unitaires** avec Pytest (ou Unitest) et les exécuter de manière automatisée lors du build réalisé par Github Actions et respecter les **conventions PEP**.



- Lien github contenant
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifiez clairement le modèle final choisi.
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
  - Un fichier README.md détaillant le projet ainsi l'organisation du répertoire
- Un support de présentation pour la soutenance (25 slides maximum)

Pour faciliter votre passage au jury, déposez sur Github tous les livrables du projet. La structure du répertoire github est à définir mais il doit contenir au minimum un dossier appelé “notebooks” contenant les notebooks créés pour répondre à cette mission. A titre illustratif, vous pouvez nommer les fichiers selon l'ordre dans lequel il apparaît, par exemple “house\_price\_01\_analyse.ipynb”, “house\_price\_02\_essais.ipynb”, et ainsi de suite.



- 15 à 20 min de Présentation
  - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.
  - Présentation du cleaning effectué, du feature engineering et de l'exploration.
  - Présentation des différentes pistes de modélisation effectuées.
  - Présentation du modèle final sélectionné ainsi que des améliorations effectuées
  - Tous les scripts Python
  - Démo en live si API ou dashboard (optionnel)
- 5 à 10 min de questions-réponses

**REG02**

## Prediction of Building energy

## Missions et objectifs



La ville de Seattle s'intéresse de près aux émissions des bâtiments non destinés à l'habitation: Prédiction de la consommation d'énergie

Vous travaillez pour la ville de Seattle. Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, votre équipe s'intéresse de près à la consommation totale d'énergie des bâtiments non destinés à l'habitation.

Votre prédiction se basera sur les données déclaratives du permis d'exploitation commerciale (taille et usage des bâtiments, mention de travaux récents, date de construction..)

Vous cherchez également à évaluer l'intérêt de l'ENERGY STAR Score pour la prédiction de consommation d'énergie, qui est fastidieux à calculer avec l'approche utilisée actuellement par votre équipe.

### Votre mission

Voici un récapitulatif de votre mission :

1. Réaliser une analyse exploratoire.
2. Tester différents modèles de prédiction afin de répondre au mieux à la problématique.
3. API ou dashboard pour prédire la consommation d'un bâtiment
4. Réaliser les tâches MLOps pour ce projet

Faites bien attention au traitement des différentes variables, à la fois pour trouver de nouvelles informations (peut-on déduire des choses intéressantes d'une simple adresse ?) et optimiser les performances en appliquant des transformations simples aux variables (normalisation, passage au log, etc.).

PJ: Données à utiliser, description-données

L'objectif est de te passer des relevés de consommation annuels futurs (attention à la fuite de données: data leakage en anglais). Nous ferons de toute façon pour tout nouveau bâtiment un premier relevé de référence la première année, donc rien ne t'interdit d'en déduire des variables structurelles aux bâtiments, par exemple la nature et proportions des sources d'énergie utilisées.

Mettez en place une évaluation rigoureuse des performances, et optimisez les hyperparamètres et le choix d'algorithmes de ML à l'aide d'une validation croisée.

- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
  - Un fichier README.md détaillant le projet ainsi l'organisation du répertoire
- Un support de présentation pour la soutenance (25 slides maximum)

Pour faciliter votre passage au jury, déposez sur Github tous les livrables du projet. La structure du répertoire github est à définir mais il doit contenir au minimum un dossier appelé “notebooks” contenant les notebooks créés pour répondre à cette mission. A titre illustratif, vous pouvez nommer les fichiers selon l'ordre dans lequel il apparaît, par exemple “energy\_01\_analyse.ipynb”, “energy\_02\_essais.ipynb”, et ainsi de suite.

- 15 à 20 min de Présentation
  - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.
  - Présentation du cleaning effectué, du feature engineering et de l'exploration.
  - Présentation des différentes pistes de modélisation effectuées.
  - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
  - Démo de l'API de prédiction et/ou le dashboard (si besoin)
- 5 à 10 min de questions-réponses

**REG03**

**Prédiction de l'émission du CO<sub>2</sub>**

# Missions et objectifs



La ville de Seattle s'intéresse de près aux émissions des bâtiments non destinés à l'habitation: Prédiction de l'émission du CO2  
Vous travaillez pour la ville de Seattle. Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, votre équipe s'intéresse de près à l'émission du CO2 des bâtiments non destinés à l'habitation.

Votre prédiction se basera sur les données déclaratives du permis d'exploitation commerciale (taille et usage des bâtiments, mention de travaux récents, date de construction..)

Vous cherchez également à évaluer l'intérêt de l'ENERGY STAR Score pour la prédiction de consommation d'énergie, qui est fastidieux à calculer avec l'approche utilisée actuellement par votre équipe.

## Votre mission

Voici un récapitulatif de votre mission :

1. Réaliser une analyse exploratoire.
2. Tester différents modèles de prédiction afin de répondre au mieux à la problématique.
3. API ou dashboard pour prédire l'émission du CO2
4. Réaliser les tâches MLOps pour ce projet

Faites bien attention au traitement des différentes variables, à la fois pour trouver de nouvelles informations (peut-on déduire des choses intéressantes d'une simple adresse ?) et optimiser les performances en appliquant des transformations simples aux variables (normalisation, passage au log, etc.).

PJ: Données à utiliser, description-données

L'objectif est de te passer des relevés de consommation annuels futurs (attention à la fuite de données: data leakage en anglais). Nous ferons de toute façon pour tout nouveau bâtiment un premier relevé de référence la première année, donc rien ne t'interdit d'en déduire des variables structurelles aux bâtiments, par exemple la nature et proportions des sources d'énergie utilisées.

Mettez en place une évaluation rigoureuse des performances, et optimisez les hyperparamètres et le choix d'algorithmes de ML à l'aide d'une validation croisée.

- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifiez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
  - Un fichier README.md détaillant le projet ainsi l'organisation du répertoire
- Un support de présentation pour la soutenance (25 slides maximum)

Pour faciliter votre passage au jury, déposez sur Github tous les livrables du projet. La structure du répertoire github est à définir mais il doit contenir au minimum un dossier appelé “notebooks” contenant les notebooks créés pour répondre à cette mission. A titre illustratif, vous pouvez nommer les fichiers selon l'ordre dans lequel il apparaît, par exemple “energy\_01\_analyse.ipynb”, “energy\_02\_essais.ipynb”, et ainsi de suite.

- 15 à 20 min de Présentation
  - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.
  - Présentation du cleaning effectué, du feature engineering et de l'exploration.
  - Présentation des différentes pistes de modélisation effectuées.
  - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
  - Démo de l'API de prédiction et/ou le dashboard (si besoin)
- 5 à 10 min de questions-réponses

**REG04**

## Prédiction de la demande de vélos

## Missions et objectifs

Vous êtes Data Scientist chez Capital Bikeshare, un service de vélos en libre-service qui permet aux utilisateurs de louer des vélos pour des trajets à court terme dans une ville. Capital Bikeshare vise à analyser et prévoir l'utilisation des vélos.

Les systèmes de vélos en libre-service sont un moyen de louer des vélos où le processus d'adhésion, de location et de restitution du vélo est automatisé via un réseau de kiosques répartis dans la ville. Ces systèmes permettent de louer un vélo à un endroit donné et de le rendre à un autre endroit en fonction des besoins.

La demande de vélos peut varier en fonction de plusieurs facteurs, tels que les conditions météorologiques, l'heure de la journée, le jour de la semaine, et des événements spécifiques dans la ville.

Les données générées par ces systèmes sont la durée du trajet, le lieu de départ, le lieu d'arrivée et le temps écoulé. Les systèmes de partage de vélos fonctionnent donc comme un réseau de capteurs, qui peut être utilisé pour étudier la mobilité dans une ville.

L'objectif principal est de **prédir le nombre de vélos qui seront empruntés** dans le futur en fonction des caractéristiques d'entrée, afin d'aider Capital Bikeshare à prévoir la demande de location de vélos dans le cadre du programme Capital Bikeshare à Washington, D.C.

Les données sont disponibles [ici](#)

# Livrables



- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
  - Un fichier README.md détaillant le projet ainsi l'organisation du répertoire
- Un support de présentation pour la soutenance (25 slides maximum)

# Soutenance



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**REG05**

**Prix des voitures d'occasion**

## Missions et objectifs



Vous êtes consultant Data Scientist chez WardsAuto, une plateforme dédiée à la vente et à l'achat de voitures d'occasion. WardsAuto souhaite optimiser ses opérations et améliorer l'expérience utilisateur en fournissant des estimations précises des prix des véhicules.

Le marché des voitures d'occasion est en constante évolution, et les prix peuvent varier considérablement en fonction de nombreux facteurs tels que la marque, le modèle, l'année, le kilométrage et d'autres caractéristiques techniques. Comprendre les déterminants du prix des voitures d'occasion peut aider à mieux évaluer les transactions et à informer les acheteurs et les vendeurs.

L'objectif principal de votre mission est de créer un modèle pour prédire le prix des voitures d'occasion en fonction de caractéristiques comme la marque, le modèle, l'année, le kilométrage et d'autres attributs afin de permettre aux vendeurs de fixer des prix compétitifs.

Les données sont disponibles [ici](#)

- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
  - Un fichier README.md détaillant le projet ainsi l'organisation du répertoire
- Un support de présentation pour la soutenance (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**REG06**

## **Student Performance**

## Missions et objectifs



Vous êtes Data Scientist au sein d'une école, nommée "School", qui propose des cours.

Dans un environnement éducatif de plus en plus axé sur les données, il est essentiel pour les établissements scolaires de comprendre les facteurs qui influencent la performance académique des étudiants.

Les performances académiques des étudiants peuvent être influencées par divers facteurs, notamment leur milieu socio-économique, leur engagement, et les ressources disponibles. Analyser ces facteurs peut fournir des insights précieux pour les éducateurs et les décideurs afin d'améliorer les résultats scolaires.

L'objectif principal est de **prédir les résultats scolaires des étudiants** en fonction de leurs caractéristiques, afin d'identifier les étudiants à risque et d'aider les éducateurs à intervenir de manière proactive.

Les données sont disponibles [ici](#)

# Livrables



- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
  - Un fichier README.md détaillant le projet ainsi l'organisation du répertoire
- Un support de présentation pour la soutenance (25 slides maximum)

# Soutenance



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**REG07**

## **Prévision des coûts médicaux**

## Missions et objectifs



Vous êtes Machine Learning Engineer au sein d'une entreprise, nommée "Health Cost", qui propose un outil d'estimation de prix pour les soins médicaux aux compagnies d'assurance santé.

La prévision des coûts médicaux est un enjeu majeur pour les compagnies d'assurance, qui cherchent à gérer les risques financiers tout en offrant des tarifs compétitifs à leurs clients. Avec la hausse constante des dépenses de santé, il devient essentiel pour les assureurs d'anticiper les coûts associés aux soins médicaux. Cela leur permet de mieux planifier leurs budgets, d'optimiser leurs politiques de couverture et d'améliorer la gestion des sinistres.

Ce projet utilise un ensemble de données qui contient des informations personnelles et médicales pour prédire les coûts des soins de santé.

**Objectif:** prédire le coût des soins médicaux en fonction de caractéristiques personnelles telles que l'âge, le sexe, le statut d'obésité, et d'autres facteurs de santé

Les données sont disponibles [ici](#)

# Livrables



- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
  - Un fichier README.md détaillant le projet ainsi l'organisation du répertoire
- Un support de présentation pour la soutenance (25 slides maximum)

# Soutenance



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**REG08**

**Taxi: Prédiction des prix des trajets**

## Missions et objectifs



Vous êtes Data Scientist chez NYC Taxi, entreprise de transport, dans le cadre d'un projet de prédition de prix des trajets en milieu urbain.

Les taxis à New York génèrent une énorme quantité de données au quotidien, incluant les trajets, les lieux de prise en charge et de dépôt, et les tarifs associés. Le montant d'une course de taxi dépend de divers facteurs, tels que la distance parcourue, l'heure de la journée, la circulation, la météo, ainsi que d'autres éléments comme les péages et les tarifs minimums.

Avec l'essor des services de transport à la demande et l'augmentation de la concurrence, il devient crucial pour les compagnies de taxi et les conducteurs d'optimiser les tarifs pour rester compétitifs tout en garantissant un service de qualité aux passagers.

L'objectif principal du projet, pour NYC Taxi, est de construire un modèle prédictif capable d'estimer avec précision le prix d'une course de taxi à partir de différentes caractéristiques disponibles avant le début de la course.

Le jeu de données contient des informations détaillées sur chaque course, telles que la date et l'heure de prise en charge et de dépôse, le point de départ et d'arrivée, la distance parcourue, le montant total de la course, etc.

Les données sont disponibles [ici](#)

- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
  - Un fichier README.md détaillant le projet ainsi l'organisation du répertoire
- Un support de présentation pour la soutenance (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**REG09**

**Prévision des ventes**

# Missions et objectifs



Vous êtes Data Scientist chez Favorita, supermarché offrant une large gamme de produits alimentaires et non alimentaires, dans le cadre d'un projet visant à améliorer la gestion des stocks et à optimiser la chaîne d'approvisionnement .

Dans un environnement de vente au détail de plus en plus compétitif, la prévision des ventes est cruciale pour optimiser la gestion des stocks, améliorer le service client et maximiser les revenus. C'est dans ce cadre que Favorita a fait appel à vos services pour un projet de prévision des ventes.

Les fluctuations des ventes peuvent être influencées par divers facteurs tels que les tendances saisonnières, les promotions, les jours fériés, et les événements spéciaux. Ainsi, une prévision précise des ventes permettrait à Favorita d'anticiper la demande, d'ajuster les niveaux de stock, et de planifier efficacement les opérations.

Les données sont disponibles [ici](#)

L'objectif principal de ce projet est de développer un modèle de prévision des ventes pour les différentes catégories de produits vendus par Favorita.

- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
- Un support de présentation pour la soutenance (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

# **REG10**

## **Prédiction des revenus des films**

## Missions et objectifs



Vous êtes Data Scientist chez Netflix, plateforme de streaming.

Avec l'essor des plateformes de streaming, Netflix a considérablement élargi son catalogue de films et de séries originales. Cependant, la concurrence croissante dans l'industrie du divertissement nécessite que Netflix prenne des décisions stratégiques éclairées concernant ses investissements dans la production de contenu. La capacité à prédire avec précision les revenus potentiels au box office d'un film avant sa sortie est essentielle pour optimiser le budget de production, le marketing et la distribution.

Objectif: fournir une estimation précise des revenus au box-office pour de futurs films, en tenant compte de divers facteurs influents.

Les données sont disponibles [ici](#)

## Livrables

- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
  - Un fichier README.md détaillant le projet ainsi l'organisation du répertoire
- Un support de présentation pour la soutenance (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)



# Classification

13 propositions de projet

**CLF01**

**Churn score**

## Missions et objectifs



Vous êtes Data Scientist au sein de "fortuneo banque", une banque en ligne qui propose une gamme variée de services financiers et bancaires pour des personnes ayant peu ou pas du tout d'historique de prêt.

L'entreprise souhaite mettre en œuvre un outil de "churn scoring" pour calculer la probabilité qu'un client quitte la banque, puis classifie le client en 'churn' ou pas. Elle souhaite donc développer un algorithme de classification en s'appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).

**Objectif :** Identifier les clients susceptibles de quitter la banque.

**Impact :** Réduire le churn en prenant des mesures préventives ciblées pour retenir les clients à haut risque.

**Les données:** Voici les données dont vous aurez besoin pour réaliser l'algorithme de classification. L'ensemble des données contient des informations sur les clients des banques qui ont quitté la banque ou qui restent clients. La description des attributs est disponible [ici](#)

- Lien github contenant:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre, dans lequel vous identifierez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code pour les tests unitaires, tracking, déploiement continu, ...
- Un support de présentation pour la soutenance (25 slides maximum)

Pour faciliter votre passage au jury, déposez sur Github tous les livrables du projet. La structure du répertoire github est à définir mais il doit contenir au minimum un dossier appelé “notebooks” contenant les notebooks créés pour répondre à cette mission. A titre illustratif, vous pouvez nommer les fichiers selon l'ordre dans lequel il apparaît, par exemple “*churn\_01\_analyse.ipynb*”, “*churn\_02\_model.ipynb*”, et ainsi de suite.

- 15 à 20 min de Présentation
  - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.
  - Présentation du cleaning effectué, du feature engineering et de l'exploration.
  - Présentation des différentes pistes de modélisation effectuées.
  - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
  - Démo de l'API de prédiction et/ou le dashboard (si besoin)
- 5 à 10 min de questions-réponses

**CLF02**

**Crédit scoring**

# Missions et objectifs



Vous êtes Data Scientist au sein d'une société financière, nommée "Prêt à dépenser", qui propose des crédits à la consommation pour des personnes ayant peu ou pas du tout d'historique de prêt.

L'entreprise souhaite mettre en œuvre un outil de "scoring crédit" pour calculer la probabilité qu'un client rembourse son crédit, puis classifie la demande en crédit accordé ou refusé. Elle souhaite donc développer un algorithme de classification en s'appuyant sur des sources de données variées (données comportementales, données provenant d'autres institutions financières, etc.).

Vous aurez sûrement besoin de joindre les différentes tables entre elles.

De plus, les chargés de relation client ont fait remonter le fait que les clients sont de plus en plus demandeurs de transparence vis-à-vis des décisions d'octroi de crédit.

## Les données

Voici les données dont vous aurez besoin pour réaliser l'algorithme de classification.

# Missions et objectifs



## Votre mission

- Construire un modèle de scoring qui donnera une prédition sur la probabilité de faillite d'un client de façon automatique.
- Réaliser les tâches MLOps pour ce projet
- Visualiser le score et l'interprétation de ce score pour chaque client de façon intelligible pour une personne non experte en data science.
- Construire un dashboard interactif à destination des gestionnaires de la relation client permettant d'interpréter les prédictions faites par le modèle, et d'améliorer la connaissance client des chargés de relation client (optionnel).

Vous pourrez ainsi vous focaliser sur l'élaboration du modèle, son optimisation et sa compréhension. A cet effet, je vous incite à sélectionner un kernel Kaggle pour vous faciliter la préparation des données nécessaires à l'élaboration du modèle de scoring. Vous analyserez ce kernel et l'adapterezez pour vous assurer qu'il répond aux besoins de votre mission.

Il faut utiliser Github Actions pour le **déploiement continu** et l'interface web 'UI MLFlow" ou un autre outil d'affichage des résultats du **tracking** et concevoir des **tests unitaires** avec Pytest (ou Unittest) et les exécuter de manière automatisée lors du build réalisé par Github Actions.

# Livrables



- Un repository GITHUB contenant les éléments suivants:
  - Un notebook de l'analyse exploratoire mis au propre et annoté.
  - Le code (ou un notebook) des différents tests de modèles mis au propre (fonction cout métier, optimisation, métrique d'évaluation, interprétabilité globale/locale du modèle final), dans lequel vous identifiez clairement le modèle final choisi.
  - Tous les scripts Python
  - Le code de l'API déployé dans le cloud ainsi que le dashboard (streamlit, gradio, ...)
  - un fichier README.md détaillant le projet ainsi l'organisation du répertoire
  - tous autres documents/codes relatifs à la réalisation des missions
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

Pour faciliter votre passage au jury, déposez sur Github tous les livrables du projet. La structure du répertoire github est à définir mais il doit contenir au minimum un dossier appelé “notebooks” contenant les notebooks créés pour répondre à cette mission. A titre illustratif, vous pouvez nommer les fichiers selon l'ordre dans lequel il apparaît, par exemple “scoring\_01\_analyse.ipynb”, “scoring\_02\_essais.ipynb”, et ainsi de suite.

# Soutenance



La soutenance se déroulera comme suit:

- 15 à 20 min de Présentation
  - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.
  - Présentation du cleaning effectué, du feature engineering et de l'exploration.
  - Présentation des différentes pistes de modélisation effectuées.
  - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
  - Démo de l'API et/ou dashboard (si besoin)
- 5 à 10 min de questions-réponses

**CLF03**

**Détection tweets de catastrophes**

## Missions et objectifs



Vous êtes consultant pour Twitter, réseau social. Twitter est devenu un important canal de communication en cas d'urgence.

L'omniprésence des smartphones permet aux gens d'annoncer une urgence qu'ils observent en temps réel. C'est pourquoi de plus en plus d'organismes s'intéressent à la surveillance programmatique de Twitter (par exemple, les organisations de secours en cas de catastrophe et les agences de presse).

Mais il n'est pas toujours évident de savoir si les mots d'une personne annoncent réellement une catastrophe. A cet effet, Twitter souhaite avoir un modèle de détection des tweets qui annoncent des catastrophes.

Les données sont disponibles ici

### Votre mission

- Vous devez construire un modèle d'apprentissage automatique qui prédit quels tweets sont liés à des catastrophes réelles et lesquels ne le sont pas.
- Réaliser les tâches MLOps pour ce projet
- Construire une API de prédiction

## Missions et objectifs



- Un notebook de l'analyse exploratoire mis au propre et annoté.
- Le code (ou un notebook) des différents tests de modèles mis au propre (optimisation, métrique d'évaluation, interprétabilité du modèle final), dans lequel vous identifierez clairement le modèle final choisi.
- Tous les scripts Python
- Le code de l'API, des tests, ...
- Un support de présentation pour la soutenance, détaillant le travail réalisé (canva, google slides, ...) (25 slides maximum)

Pour faciliter votre passage au jury, déposez sur Github tous les livrables du projet. La structure du répertoire github est à définir mais il doit contenir au minimum un dossier appelé “notebooks” contenant les notebooks créés pour répondre à cette mission. A titre illustratif, vous pouvez nommer les fichiers selon l'ordre dans lequel il apparaît, par exemple “disastertweets\_01\_analyse.ipynb”, “sdisastertweets\_02\_essais.ipynb”, et ainsi de suite.

- 15 à 20 min de Présentation
  - Présentation de la problématique, de son interprétation et des pistes de recherche envisagées.
  - Présentation du cleaning effectué, du feature engineering et de l'exploration.
  - Présentation des différentes pistes de modélisation effectuées.
  - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
  - Démo de l'API de prédiction (si besoin)
- 5 à 10 min de questions-réponses

**CLF04**

## **Classifiez des biens de consommation**

# Missions et objectifs



Vous êtes Data Scientist au sein de l'entreprise "Place de marché", qui souhaite lancer une marketplace e-commerce.

Sur cette place de marché anglophone, des vendeurs proposent des articles à des acheteurs en postant une photo et une description.

Pour l'instant, l'attribution de la catégorie d'un article est effectuée manuellement par les vendeurs, et est donc peu fiable. De plus, le volume des articles est pour l'instant très petit.

Pour optimiser l'expérience utilisateur des vendeurs (faciliter la mise en ligne de nouveaux articles) et des acheteurs (faciliter la recherche de produits), et dans l'optique d'un passage à l'échelle, il devient nécessaire d'automatiser cette tâche.

Les données sont disponibles [ici](#)

## Objectifs:

- étudier la faisabilité d'un moteur de classification des articles en différentes catégories, à partir du texte (en anglais) ou de l'image, avec un niveau de précision suffisant.
- Réaliser les tâches MLOps
- Mettre en place un interface déployé sur le cloud pour tester la catégorisation des articles (streamlit, gradio, ...) à partir d'une image ou d'un texte descriptif

# Livrables



- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des données textes ou images ainsi que les résultats de l'étude de faisabilité
  - Un dossier Python contenant le code de l'API
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

# Soutenance



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Explication des prétraitements, des extractions de features et des résultats de l'étude de faisabilité
  - Résultats de la classification supervisée
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CLF05**

## **Analyse des commentaires des films**

## Missions et objectifs



Vous êtes Data Scientist chez Netflix, plateforme de streaming.

Netflix souhaite comprendre en profondeur les perceptions et les réactions de son audience face aux films et séries disponibles sur sa plateforme. À l'ère des réseaux sociaux et des critiques en ligne, les avis des utilisateurs influencent grandement l'engagement et la fidélisation. En analysant les sentiments des critiques de films, Netflix peut mieux cerner les éléments qui plaisent ou déplaisent au public, ajuster ses stratégies de contenu et optimiser les recommandations personnalisées.

Les données sont disponibles [ici](#)

**Objectif:** classifier des critiques de films en utilisant le traitement du langage naturel (NLP) pour mieux comprendre les préférences des utilisateurs.

## Livrables

- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Explication des prétraitements, des extractions de features et des résultats de l'étude
  - Présentation du test de l'API ou du dashboard (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CLF06**

## Classification de produits

Le Groupe Otto est l'une des plus grandes entreprises de commerce électronique au monde, avec des filiales dans plus de 20 pays, dont Crate & Barrel (États-Unis), Otto.de (Allemagne) et 3 Suisses (France). Nous vendons chaque jour des millions de produits dans le monde entier, et plusieurs milliers de produits viennent s'ajouter à notre gamme.

Une analyse cohérente de la performance de nos produits est cruciale. Cependant, en raison de la diversité de notre infrastructure mondiale, de nombreux produits identiques sont classés différemment. Par conséquent, la qualité de l'analyse de nos produits dépend fortement de notre capacité à regrouper avec précision les produits similaires. Plus la classification est bonne, plus nous pouvons obtenir d'informations sur notre gamme de produits.

Le processus de classification manuelle est long et sujet à des erreurs, surtout avec des milliers de nouvelles références ajoutées régulièrement.

**Objectif:** ce projet de classification des produits vise donc à automatiser la catégorisation en utilisant des techniques d'apprentissage automatique, ce qui permettra d'optimiser l'organisation et le traitement des données produits.

Les données sont disponibles [ici](#)

- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Explication des prétraitements, des extractions de features et des résultats de l'étude
  - Présentation du test de l'API ou du dashboard (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CLF07**

**Satisfaction client**

## Missions et objectifs



Dans un marché bancaire compétitif, la satisfaction client est cruciale pour la rétention des clients et la croissance des activités. Santander, une institution bancaire mondiale, cherche à mieux comprendre et prédire la satisfaction de ses clients afin d'améliorer leurs expériences.

Ce projet vise à développer un modèle capable de classifier les clients insatisfaits, permettant aux équipes de support, ainsi qu'aux niveaux de direction (C-suite), de prendre des décisions proactives pour réduire les taux d'attrition et améliorer les services.

Objectif: aider à identifier les clients insatisfaits dès le début de leur relation

Les données sont disponibles [ici](#)

- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Explication des prétraitements, des extractions de features et des résultats de l'étude
  - Présentation du test de l'API ou du dashboard (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CLF08**

## **Classification du premier pays de destination**

## Missions et objectifs



En tant que plateforme de réservation d'hébergements et d'expériences, Airbnb vise à proposer des séjours personnalisés pour attirer et fidéliser ses utilisateurs. Une partie essentielle de cette stratégie consiste à prédire les premiers choix de destinations des nouveaux utilisateurs, ce qui permet à Airbnb de mieux orienter son contenu, de proposer des recommandations personnalisées et d'améliorer le taux de conversion.

Les nouveaux utilisateurs d'Airbnb peuvent réserver un logement dans plus de 34 000 villes réparties dans plus de 190 pays. En prédisant avec précision où un nouvel utilisateur réservera sa première expérience de voyage, Airbnb peut partager un contenu plus personnalisé avec sa communauté, réduire le délai moyen de la première réservation et mieux prévoir la demande.

**Objectif:** Dans le cadre de ce projet, Airbnb vous met au défi de prédire dans quel pays un nouvel utilisateur effectuera sa première réservation en fonction de ses données de profil et de comportement initial.

Les données sont disponibles [ici](#)

- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

# Soutenance



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Explication des prétraitements, des extractions de features et des résultats de l'étude
  - Présentation du test de l'API ou du dashboard (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CLF09**

# **Classification des logiciels malveillants**

La sécurité informatique est un enjeu majeur dans l'écosystème numérique actuel, où le nombre et la sophistication des malwares (logiciels malveillants) ne cessent d'augmenter. Pour une entreprise comme Microsoft, garantir la sécurité des utilisateurs et des systèmes est une priorité.

Ce projet vise à améliorer les techniques de détection et de classification des malwares en se basant sur l'analyse des signatures et des comportements des logiciels malveillants.

L'objectif est de classifier différents types de malwares pour permettre à Microsoft d'identifier et de contrer plus efficacement ces menaces en développant des systèmes de protection plus réactifs et précis.

Les données sont disponibles [ici](#)

- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Explication des prétraitements, des extractions de features et des résultats de l'étude
  - Présentation du test de l'API ou du dashboard (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CLF10**

## **Classification des maladies cardiaques**

## Missions et objectifs



Les médiums spiritualistes, engagés dans une pratique axée sur le bien-être holistique et l'équilibre énergétique, cherchent à intégrer des approches scientifiques modernes pour améliorer la santé physique de leurs clients. Comprendre et prédire les risques de maladies cardiaques permettrait d'offrir un accompagnement plus complet, avec des recommandations plus personnalisées.

Ce projet de classification des maladies cardiaques a pour but de fournir un outil prédictif qui aide les médiums et praticiens de bien-être à évaluer le risque cardiovasculaire de leurs clients, en tenant compte de leurs antécédents médicaux et de divers facteurs de santé.

**Objectif:** prédire de manière fiable le risque de maladies cardiaques sur la base des données médicales du client

Les données sont disponibles [ici](#)

# Livrables



- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

# Soutenance



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Explication des prétraitements, des extractions de features et des résultats de l'étude
  - Présentation du test de l'API ou du dashboard (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CLF11**

**Détection de fraude**

## Missions et objectifs



Vesta Corporation, un leader dans la gestion de transactions en ligne, cherche à renforcer sa capacité à détecter les fraudes dans les transactions électroniques. La fraude en ligne est une préoccupation majeure pour toutes les entreprises qui gèrent des paiements numériques, car elle entraîne des pertes financières importantes et diminue la confiance des clients.

Dans le cadre de ce projet, Vesta souhaite développer des modèles de machine learning fiables qui identifient les transactions frauduleuses à partir de vastes ensembles de données transactionnelles (montant, heure, type de transaction) et comportementales et des informations sur les appareils utilisés.

Le défi est de concevoir un modèle de classification performant, capable de distinguer les transactions légitimes des transactions frauduleuses.

Les données sont disponibles [ici](#)

- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Explication des prétraitements, des extractions de features et des résultats de l'étude
  - Présentation du test de l'API ou du dashboard (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CLF12**

**Détection de la race du chien sur une photo**



Agir pour  
les animaux

## Missions et objectifs

Vous êtes bénévole pour l'association de protection des animaux, Agir pour les animaux, de votre quartier.

Vous vous demandez donc ce que vous pouvez faire en retour pour aider l'association.

Vous apprenez, en discutant avec un bénévole, que leur base de données de pensionnaires commence à s'agrandir et qu'ils n'ont pas toujours le temps de référencer les images des animaux qu'ils ont accumulées depuis plusieurs années. Ils aimeraient donc obtenir un algorithme capable de classer les images en fonction de la race du chien présent sur l'image.

Les données sont disponibles [ici](#)

### Votre mission:

L'association vous demande de réaliser un algorithme de détection de la race du chien sur une photo, afin d'accélérer leur travail d'indexation. Il s'agit également de mettre en place un dashboard permettant de tester le modèle choisi et suivre les bonnes pratiques MLOps pour ce projet.

## Recommandations



Attention sur le fait que l'entraînement (même partiel) d'un réseau de neurones convolutionnels est très gourmand en ressources. Si le processeur de votre ordinateur ne suffit pas, voici plusieurs solutions :

- Limitez le jeu de données, en ne sélectionnant que quelques classes (races de chiens), ce qui permettra déjà de tester la démarche et la conception des modèles, avant une éventuelle généralisation.
- Utilisez la carte graphique de l'ordinateur en tant que GPU (l'installation est un peu fastidieuse, et l'ordinateur est inutilisable le temps du calcul).
- cloud computing, qui permet d'avoir temporairement accès à des machines très puissantes, en étant facturé seulement durant le temps d'utilisation. Le plus connu est AWS, mais d'autres existent (Google, Microsoft...). Vous pouvez tester également [Google Colaboratory](#) ou [Kaggle](#) qui permet de mettre en œuvre gratuitement des réseaux CNN utilisant de la GPU.

# Livrables



- Un repository GITHUB contenant les éléments suivants:
  - Le code/notebook Python
  - Un programme Python qui prend une image en entrée et retourne la race la plus probable du chien présent sur l'image via un interface (si besoin)
  - Tous les scripts Python
- Votre support de présentation (25 slides maximum).

# Soutenance



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CLF13**

## **Classification des ordures**

# Missions et objectifs



Vous êtes Data Scientist au sein de l'entreprise "WasteNet", pionnière dans le domaine du recyclage des ordures.

WasteNet souhaite proposer un objet de tri intelligent afin de recycler certains objets et de diminuer les impacts néfastes sur l'environnement.

L'objectif de ce projet est de proposer un modèle prédictif capable de détecter le type d'ordure à partir d'une image. Il est aussi demander un suivi des bonnes pratiques MLOps.

Les données sont disponibles [ici](#)

Voici quelques exemples de types d'ordures: électronique, carton, plastique, verre, métal, organique ...



## Livrables



- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Un dossier Python contenant le code de l'API (si besoin)
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard, ...
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

# Soutenance



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Explication des prétraitements, des extractions de features et des résultats de l'étude
  - Présentation du test de l'API ou du dashboard (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

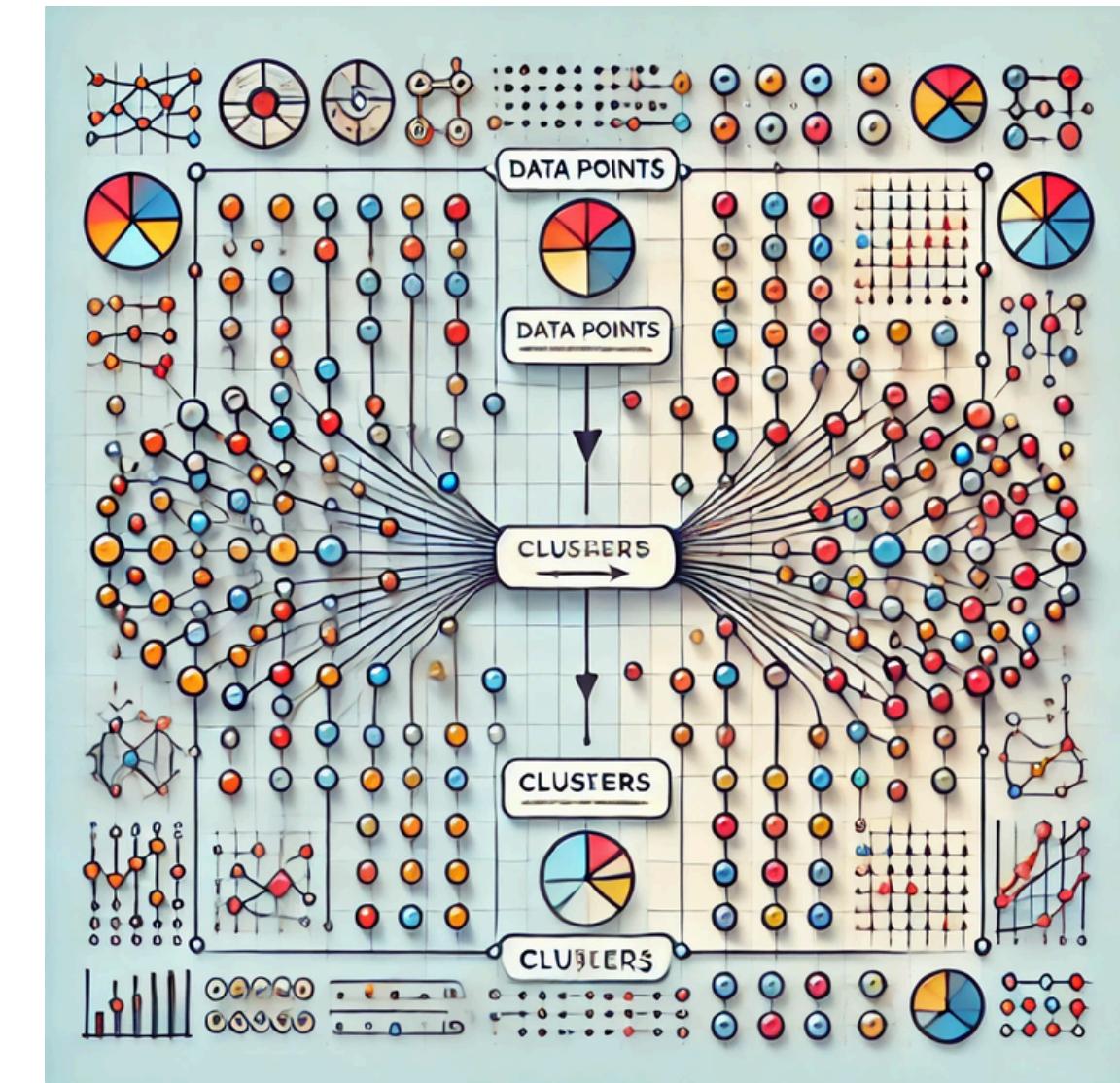


# Clustering

Quelques propositions de projet

## Notes: Sujets précédemment présentés dans la section Classification

Vous avez la possibilité de sélectionner un sujet de classification et de le traiter en utilisant des techniques de clustering, sans tenir compte des étiquettes lors du processus d'apprentissage.



**CTR01**

## Olist: Segmentation des clients

## Missions et objectifs



Vous êtes consultant pour **Olist**, une entreprise brésilienne qui propose une solution de vente sur les marketplaces en ligne. Olist souhaite que vous **fournissiez à ses équipes d'e-commerce une segmentation des clients** qu'elles pourront utiliser au quotidien pour leurs **campagnes de communication**. Votre objectif est de **comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles**. Vous devrez fournir à l'équipe marketing **une description actionable de votre segmentation** et de sa logique sous-jacente pour une **utilisation optimale**, ainsi qu'une **proposition de contrat de maintenance (optionnel)** basée sur une analyse de la stabilité des segments au cours du temps.

### Votre mission

- Votre mission est d'aider les équipes d'Olist à **comprendre les différents types d'utilisateurs**. Vous utiliserez donc des **méthodes non supervisées pour regrouper des clients de profils similaires**. Ces catégories pourront être utilisées par l'équipe Marketing pour mieux communiquer.
- Il faut utiliser Github Actions pour le **déploiement continu** et l'interface web 'UI MLFlow" ou un autre outil d'affichage des résultats du **tracking**, concevoir des **tests unitaires** avec Pytest (ou Unittest) et les exécuter de manière automatisée lors du build réalisé par Github Actions et **respecter les conventions PEP**.

- Un repository GITHUB contenant les éléments suivants:
  - Un notebook (ou code commenté au choix) de l'analyse exploratoire & d'essais des différentes approches de modélisation.
  - Un notebook de simulation pour déterminer la fréquence nécessaire de mise à jour du modèle de segmentation.
  - un fichier README.md détaillant le projet ainsi l'organisation du répertoire
  - Tous autres documents/codes relatifs à la réalisation des missions
- Un support de présentation pour présenter votre travail à un collègue (25 slides maximum)

Pour faciliter votre passage devant le jury, déposez sur la plateforme Github tous les livrables du projet. La structure du répertoire github est à définir mais il doit contenir au minimum un dossier appelé “notebooks” contenant les notebooks créés pour répondre à cette mission. A titre illustratif, vous pouvez nommer les fichiers selon l'ordre dans lequel il apparaît, par exemple “segmentation\_01\_analyse.ipynb”, “segmentation\_02\_essais.ipynb”, et ainsi de suite.

La soutenance se déroulera comme suit:

- 15 à 20 min de Présentation
  - Présentation de la problématique, du cleaning effectué, du feature engineering et de l'exploration
  - Présentation des différentes pistes de modélisation effectuées, des tests unitaires et des expériences sur MLFlow
  - Présentation du modèle final sélectionné ainsi que des améliorations effectuées.
  - Présentation de la simulation pour définir le délai de maintenance du modèle (si besoin)
- 5 à 10 min de questions-réponses

**CTR02**

## **Dubai Mall: Segmentation de la clientèle**



## Missions et objectifs

Le secteur du commerce de détail est de plus en plus axé sur la compréhension des comportements des clients afin d'optimiser les stratégies de marketing et d'améliorer l'expérience client. À Dubaï, un des centres commerciaux les plus fréquentés, le Dubai Mall, souhaite mieux segmenter sa clientèle pour adapter ses services et offres commerciales

L'analyse des données des clients peut révéler des patterns de comportement et des préférences qui, une fois identifiés, peuvent aider les gestionnaires du centre commercial à mieux cibler leurs campagnes et à améliorer la satisfaction des clients.

**Objectif:** développer un modèle de segmentation client qui permettra de mieux comprendre la diversité de la clientèle du Dubai Mall et de répondre à ses besoins spécifiques.

Les données sont disponibles [ici](#)

## Livrables



- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CTR03**

## **Segmentation de voitures**

## Missions et objectifs

La Centrale est une entreprise spécialisée dans la vente de véhicules neufs et d'occasion. Dans un marché concurrentiel, il est crucial de comprendre les préférences des clients et d'optimiser l'inventaire en fonction des tendances du marché. Pour cela, l'analyse des caractéristiques des voitures et leur évaluation par les consommateurs peuvent fournir des informations précieuses.

Le jeu de données contient divers attributs de voitures, tels que le type, la taille, le prix, la sécurité, et d'autres caractéristiques.

Les données sont disponibles [ici](#)

## Livrables

- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

## Soutenance

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CTR04**

## **Engagement dans les médias sociaux**



## Missions et objectifs

Avec l'augmentation exponentielle de l'utilisation des réseaux sociaux, comprendre le comportement des utilisateurs sur des plateformes comme Facebook est essentiel pour améliorer l'engagement, la satisfaction des utilisateurs et, en fin de compte, la rentabilité de l'entreprise. Facebook, en tant que l'une des plus grandes plateformes de médias sociaux au monde, collecte une immense quantité de données sur les interactions des utilisateurs, y compris les likes, les commentaires, les partages et les abonnements. Ce projet vise à analyser ces données pour identifier les facteurs qui influencent l'engagement des utilisateurs.

Chaque entrée fournit des mesures détaillées sur la façon dont les messages sont reçus par le public, ce qui permet d'obtenir des informations fondées sur des données concernant la performance du contenu.

**Objectif:** comprendre le Comportement des Utilisateurs : Déterminer les comportements d'engagement des utilisateurs afin de mieux cibler les campagnes de marketing et les contenus proposés.

Les données sont disponibles [ici](#)



- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CTR05**

**Analyse de la qualité de l'air**

# Missions et objectifs



Cet ensemble de données, qui couvre 170 pays et plus de 300 villes, offre une vue d'ensemble de la dynamique de la qualité de l'air au niveau mondial. La pollution de l'air est un problème de santé publique majeur qui affecte des millions de personnes dans le monde. L'Organisation Mondiale de la Santé (OMS) s'engage à surveiller et à réduire les risques liés à la pollution de l'air afin d'améliorer la santé publique et de promouvoir un environnement sain.

Axé sur des polluants essentiels tels que le monoxyde de carbone, l'ozone, le dioxyde d'azote et les particules (PM2.5), les informations tirées de cet ensemble de données permettent aux utilisateurs d'analyser les tendances en matière de qualité de l'air, de formuler des politiques efficaces et de contribuer à la promotion d'une planète plus saine.

Avec des colonnes essentielles telles que le nom du pays, le nom de la ville, les valeurs globales de l'indice de qualité de l'air (IQA) et les concentrations de polluants spécifiques, cet ensemble de données permet des analyses approfondies et des études de corrélation.

**Objectif:** Utiliser des techniques de clustering pour identifier des modèles dans les données de pollution de l'air, permettant de segmenter les régions en fonction de leurs niveaux de pollution et de leurs caractéristiques.

Les données sont disponibles [ici](#)

- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

**CTR06**

**Clustering de pays**

## Missions et objectifs



L'entreprise HELP, une ONG humanitaire internationale, s'engage à lutter contre la pauvreté et à améliorer le bien-être des populations à travers le monde en fournissant des équipements de base et des secours en cas de catastrophes et de calamités naturelles. Dans le cadre de cette mission, il est essentiel de comprendre les différentes dynamiques socio-économiques et environnementales des pays.

Ce projet vise à exploiter les techniques d'apprentissage non supervisé pour identifier des segments de pays en fonction de divers indicateurs.

**Objectif:** regrouper des pays en fonction de divers indicateurs économiques (PIB, taux d'alphabétisation, etc.) pour analyser les similitudes et les différences entre les économies mondiales.

# Livrables



- Un repository GITHUB contenant les éléments suivants:
  - Un ou des notebooks (ou des fichiers .py) contenant les fonctions permettant le prétraitement et la feature extraction des images ainsi que les résultats de la modélisation
  - Tous les scripts Python
  - Un fichier README.md décrivant la structure du repository Github
  - Le code de l'API ou/et du dashboard (si besoin)
- Un support de présentation pour la soutenance, détaillant le travail réalisé (25 slides maximum)

# Soutenance



La soutenance se déroulera comme suit:

- Présentation (15 à 20 minutes)
  - Rappel de la problématique et présentation du jeu de données
  - Présentation des différentes pistes de modélisations effectuées et le choix du modèle final
  - Présentation du test de l'API (si besoin)
- Discussion (10 minutes)
- Débriefing (5 minutes)

# Annexes

# Exemple: 100 projets Data Science

## 100 DATA SCIENCE PROJECT

1. Housing Price Predictor
2. Customer Segmentation
3. Fraud Detection
4. Movie Recommender
5. Stock Price Predictor
6. Churn Predictor
7. Sentiment Analysis
8. Image Classifier
9. Credit Risk Analysis
10. Email Spam Detector
11. Disease Diagnosis
12. E-commerce Recommender
13. Text Summarizer
14. Weather Forecaster
15. Visitor Predictor
16. Flight Delay Predictor
17. Handwritten Digit Recognizer
18. Lifetime Value Predictor
19. Loan Default Predictor
20. Autonomous Navigation
21. Employee Attrition Predictor
22. Traffic Flow Predictor
23. Music Genre Classifier
24. Object Detector
25. Fashion Recognizer
26. Employee Performance Predictor
27. Language Translator
28. Personality Predictor
29. Product Demand Predictor
30. Image Segmenter
31. Topic Modeler
32. Movie Genre Classifier
33. Speech Emotion Recognizer
34. E-commerce Sales Forecaster
35. Intrusion Detector
36. Restaurant Recommender
37. Movie Sentiment Analyzer
38. Emotion Detector
39. Social Media Analyzer
40. Species Identifier
41. License Plate Recognizer
42. Salary Predictor
43. Customer Satisfaction Analyzer
44. Video Classifier
45. Student Performance Predictor
46. Face Recognizer
47. Energy Predictor
48. Fake News Classifier
49. Traffic Sign Recognizer
50. Threat Detector
51. Complaint Analyzer
52. Hospital Readmission Predictor
53. Equipment Fault Detector
54. News Feed Recommender
55. Outbreak Predictor
56. Network Analyzer
57. Hotel Booking Predictor
58. Activity Recognizer
59. Ticket Classifier
60. Crop Yield Predictor
61. Mask Detector
62. Reputation Analyzer
63. Social Media Event Detector
64. Image Describer
65. User Engagement Predictor
66. Event Recommender
67. Defect Detector
68. Purchase Behavior Analyzer
69. Store Sales Predictor
70. Conservation Monitor
71. Air Quality Predictor
72. Response Time Predictor
73. Market Trend Analyzer
74. Document Classifier
75. Dropout Predictor
76. Fake Account Detector
77. Food Recognizer
78. Solar Power Predictor
79. Product Review Analyzer
80. Traffic Anomaly Detector
81. Shelf Space Optimizer
82. Service Time Predictor
83. Credit Card Fraud Predictor
84. Ticket Sales Predictor
85. Text-to-Speech Converter
86. Demographic Segmentation
87. Price Optimizer
88. Inventory Demand Predictor
89. Congestion Predictor
90. Turnover Predictor
91. Disease Spread Predictor
92. Speech-to-Text Converter
93. Road Condition Monitor
94. Taxi Demand Predictor
95. Passenger Demand Forecaster
96. Playlist Generator
97. Burnout Predictor
98. Poaching Detector
99. Energy Consumption Predictor
100. Cyberbullying Detector