

Projets Encadrées

❖ **Livrable 1 - Analyse exploratoire des données et préparation :**

Dans cette première étape, les étudiants effectueront une analyse exploratoire des données (EDA) pour comprendre les caractéristiques et les tendances des données. Ils nettoieront et prépareront les données pour l'entraînement, en effectuant des opérations telles que la gestion des valeurs manquantes, la normalisation et l'encodage des variables catégorielles.

❖ **Livrable 2 - Modélisation et optimisation :**

Les étudiants entraîneront et évalueront plusieurs modèles en fonction des objectifs du projet (classification, régression, détection d'anomalies, etc.). Ils sélectionneront le meilleur modèle en fonction de métriques pertinentes et optimiseront les hyperparamètres pour améliorer les performances.

❖ **Livrable 3 - Visualisation des résultats et interprétation :**

Les étudiants visualiseront les résultats de leurs analyses et modèles pour faciliter la compréhension et l'interprétation des résultats. Ils utiliseront des graphiques, des tableaux et d'autres techniques de visualisation pour mettre en évidence les tendances et les caractéristiques importantes.

❖ **Livrable 4 - Documentation et validation des résultats :**

Les étudiants documenteront leurs résultats, justifieront leurs choix de modèles et d'algorithmes, et valideront leurs conclusions en utilisant des techniques de validation croisée ou des ensembles de données de test. Cette étape garantira la robustesse et la généralisation du modèle.

❖ **Livrable final - Rapport, présentation, démo vidéo et code source :**

Le livrable final comprendra un rapport détaillé décrivant les objectifs, les méthodologies, les résultats et les conclusions du projet. Les étudiants créeront également une présentation PowerPoint pour résumer et communiquer les résultats de manière concise. Ils réaliseront une démo sous forme de vidéo pour montrer le fonctionnement de leur solution, en expliquant les différentes étapes et fonctionnalités. Enfin, ils partageront le code source du projet sur GitHub, en veillant à ce qu'il soit bien organisé et documenté, afin que d'autres puissent facilement comprendre et réutiliser leur travail.

NB : des sessions d'encadrement peuvent être programmées à la convenance des parties prenante.

LISTE DES PROJETS ENCADRES EN DATA SCIENCE

1. Analyse de sentiments sur les critiques de films

Description : Analyser les critiques de films pour déterminer si elles sont positives ou négatives en utilisant des techniques de traitement du langage naturel (NLP).

Dataset : IMDB Movie Reviews (<https://ai.stanford.edu/~amaas/data/sentiment/>)

Étapes à suivre :

- a. Nettoyer et préparer les données.
- b. Explorer les données et visualiser les résultats.
- c. Appliquer des techniques de NLP pour extraire des caractéristiques.
- d. Entraîner un modèle de classification pour prédire la polarité des critiques.
- e. Évaluer et optimiser le modèle.

2. Prédiction du prix des logements

Description : Construire un modèle de régression pour prédire les prix des logements en fonction de diverses caractéristiques.

Dataset : Boston Housing Dataset (<https://www.kaggle.com/c/boston-housing>)

Étapes à suivre :

- a. Charger et explorer les données.
- b. Nettoyer et préparer les données pour l'entraînement.
- c. Entraîner différents modèles de régression.
- d. Comparer les performances des modèles et choisir le meilleur.
- e. Optimiser le modèle sélectionné et évaluer ses performances.

3. Prédiction des retards de vols

Description : Prédire les retards de vols en fonction des informations météorologiques et des données historiques des vols.

Dataset : Flight Delay Dataset (<https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018>)

Étapes à suivre :

- a. Charger et explorer les données.
- b. Nettoyer et préparer les données pour l'entraînement.
- c. Entraîner un modèle de classification pour prédire les retards de vols.
- d. Évaluer et optimiser le modèle.
- e. Visualiser les résultats.

4. Système de recommandation de films

Description : Développer un système de recommandation de films en utilisant des techniques de filtrage collaboratif.

Dataset : MovieLens Dataset (<https://grouplens.org/datasets/movielens/>)

Étapes à suivre :

- a. Charger et explorer les données.
- b. Préparer les données pour l'entraînement.
- c. Implémenter des techniques de filtrage collaboratif pour générer des recommandations.
- d. Évaluer les performances du système de recommandation.
- e. Visualiser et analyser les résultats.

5. Détection de fraude par carte de crédit

Description : Identifier les transactions frauduleuses en utilisant des techniques de classification et d'apprentissage non supervisé.

Dataset : Credit Card Fraud Detection Dataset (<https://www.kaggle.com/mlg-ulb/creditcardfraud>)

Étapes à suivre :

- a. Charger et explorer les données.
- b. Préparer les données pour l'entraînement (normalisation, réduction de la dimensionnalité, etc.).
- c. Entraîner un modèle de classification pour détecter les fraudes.
- d. Utiliser des techniques d'apprentissage non supervisé pour identifier les anomalies.
- e. Évaluer et optimiser les modèles.

6. Analyse des tendances YouTube

Description : Analyser les tendances des vidéos YouTube pour comprendre les facteurs influençant leur popularité.

Dataset : YouTube Trending Videos Dataset (<https://www.kaggle.com/datasnaek/youtube-new>)

Étapes à suivre :

- Charger et explorer les données.
- Nettoyer et préparer les données pour l'analyse.
- Analyser les facteurs influençant la popularité des vidéos.
- Visualiser les résultats et tirer des conclusions.
- Proposer des recommandations pour créer du contenu populaire.

7. Prédiction de l'espérance de vie

Description : Estimer l'espérance de vie en fonction de divers facteurs socio-économiques et de santé.

Dataset : Life Expectancy Dataset (<https://www.kaggle.com/kumarajarshi/life-expectancy-who>)

Étapes à suivre :

- Charger et explorer les données.
- Nettoyer et préparer les données pour l'entraînement.
- Entraîner un modèle de régression pour prédire l'espérance de vie.
- Évaluer et optimiser le modèle.
- Visualiser les résultats et interpréter les facteurs importants.

8. Reconnaissance de chiffres manuscrits

Description : Identifier les chiffres manuscrits en utilisant des techniques d'apprentissage profond.

Dataset : MNIST Dataset (<http://yann.lecun.com/exdb/mnist/>)

Étapes à suivre :

- Charger et préparer les données.
- Créer un réseau de neurones convolutif (CNN) pour la classification d'images.
- Entraîner et optimiser le modèle CNN.
- Évaluer les performances du modèle sur un ensemble de test.
- Visualiser les résultats.

9. Prédiction de l'énergie éolienne

Description : Estimer la production d'énergie éolienne en fonction de conditions météorologiques.

Dataset : Wind Energy Dataset (<https://www.kaggle.com/berkerisen/wind-turbine-scada-dataset>)

Étapes à suivre :

- Charger et explorer les données.
- Nettoyer et préparer les données pour l'entraînement.
- Entraîner un modèle de régression pour prédire la production d'énergie éolienne.
- Évaluer et optimiser le modèle.
- Visualiser les prédictions et les comparer avec les données réelles.

Cordialement