# Projet Data IA Analyse de données et prédictions

**Description de l'entreprise**

**Qui :** Client SNCF
**Quoi :** Exploitation des données d'utilisation du site data.sncf.com
**Où :** l'apprenant n'a pas besoin de se déplacer, sa domiciliation peut se faire sur toute la périphérie marseillaise. Le client peut-être contacté à distance.
**Quand :** 6 mois à partir d'octobre 2020.
**Comment :** Mettre en place un cahier des charges répondant au besoin client. Exploitation des données d'utilisation du site data.sncf.com. En partant des données connues, réaliser des prédictions sur les données d'utilisation du site data.sncf.com pour l'année 2021. Mise en place des outils de visualisation adaptés. Mettre les données accessibles sous forme de BBD.
Vous devrez mettre en place des outils gestion de projet (git, planification). Vous devrez présenter vos objectifs et réalisation chaque fin de semaine au travers de rapport et de stands-up. Vous présenterez chaque semaine la veille réalisée sur les techno que vous pensez implémenter sous la forme qui vous apparaît la plus adaptée (vidéo, présentation, rapport etc.).

Le travail peut-être réalisé en groupe de 2 ou 3 personnes ou individuellement.

Ce jeu de données présente les indicateurs de fréquentation du site data.sncf.com, et d'utilisation des jeux de données.
Ces données proviennent du Back office d'OpenDataSoft et de Google Analytics.
Le bilan Google Analytics débute en février 2020, date à laquelle ce dispositif a été mis en place.

https://data.sncf.com/explore/dataset/statistiques-dutilisation-datasncfcom/

Vous allez réaliser un modèle de prédiction des données provenant du site data.sncf.com Vous devez faire ressortir quelles sont les statistiques d'utilisation du site puis réaliser des prédictions d'utilisation du site :
1. Dans un premier temps vous allez récupérer les données correspondantes.
2. Après avoir effectué une analyse globale des données et des pré-traitements, vous déploierez les données dans une base de données (relationnelle?).
3 . Ensuite vous allez réaliser une étude statistique de ces données, puis mettre en place des outils de visualisation adaptés.
4. Puis, vous développerez un outil de prédiction d'utilisation du site data.sncf.com par le biais de réseaux neuronaux. (algorithme Deep?)
5. Vous mettrez ensuite en place, des outils de visualisation adaptés.

**Objectif d'apprentissage visé :**

**A3. Gestion de projet et qualité - Niveau 3**
C8. Analyser et formaliser la demande ou le besoin en développement de base de données C9.
Auto-contrôler, tout au long du processus de développement, la cohérence des données et la conformité à la demande.
C10. Suivre, adapter et rendre compte de la réalisation du projet à partir du planning projet validé. C12.
Rechercher des solutions pour la résolution de problèmes techniques rencontrés au moyen des ressources disponibles.

**A4.Exploiter l'intelligence artificielle dans le développement d'applications - Niveau 2**

C13. Constituer un jeu de données exploitable de manière à entraîner un modèle d'apprentissage en utilisant la méthodologie et/ou l'outil approprié en fonction des standards de l'écosystème C14. Interpréter les données grâce à des outils de visualisation de données en vue d'expliquer les caractéristiques du jeu de données
C15. Exploiter un modèle d'apprentissage supervisé ou non supervisé permettant la classification ou la prédiction d'une variable en fonction des données disponibles et des outils sélectionnés C16. Améliorer les performances d'un modèle d'apprentissage à l'aide d'une évaluation de la qualité des données et de la technique de modélisation afin de réduire les biais et les anomalies de résultats C17. Concevoir un modèle d'apprentissage efficient en exploitant les méthodes standards d'apprentissage profond pour répondre à une problématique identifiée

**A1. Développement d'une base de données – Niveau 2**
C1. Concevoir et structurer physiquement une base de données relationnelle ou non, à partir des besoins, contraintes et données du commanditaire.
C2. Acquérir des données, les combiner et les structurer en données propres en vue de leur intégration dans la structure de la base de données.
C3. Intégrer des données propres et préparées dans la base de données finale, en utilisant des langages informatiques, logiciels ou outils.
C4. Optimiser une base de données afin d'en maintenir la fiabilité et la qualité des données. Nettoyer et améliorer les performances.

**A2. Exploitation d'une base de données – Niveau 3**
C5. Interroger et traiter, simultanément et au niveau approprié, des données afin de les stocker en sécurité, brutes ou traitées, provisoirement ou durablement, en fonction du résultat recherché. C6. Concevoir et réaliser un rendu visuel des données issues du processus d'extraction, à l'aide d'un (des) support(s) adapté(s) répondant aux attentes du commanditaire.
C7. Mettre à disposition les rendus visuels simples des données en accès libre ou contrôlé.

**Tâches mobilisatrices**
- Identification du type de base de données approprié à la demande.
- Conception du modèle de données en respectant les standards.
- Création d'une base de données relationnelles et/ou NoSQL.
- Recensement des données à utiliser, leurs formats, leurs sources, leurs structures ainsi que leurs détenteurs.
- Collecte des données.
- Nettoyage des données à importer, à l'aide de scripts ou de logiciels spécifiques appropriés.
- Manipulation des données sous divers formats de fichier plats (XML, JSON, CSV)
- Gestion des fichiers de métadonnées associés aux fichiers : création, mise à jour ou suppression.
- Choix de la méthode d'import.
- Intégration, à partir de fichiers plats, de tables ou d'une interface de programmation, automatiquement ou manuellement, les données dans la base.

- Analyse de la demande client.
- Identification, à partir du cahier des charges, les utilisateurs et leurs profils, les différents besoins, les contraintes techniques et réglementaires ainsi que les données du commanditaire. - Le cas échéant, formalisation d'un cahier des charges du projet à partir de la demande client. - Suivi, adaptation et communication de la réalisation du projet à partir du planning projet validé. - Suivi du projet, dans un objectif d'optimisation, en utilisant une méthodologie adaptée. - Adaptation du projet aux contraintes et problématiques rencontrées
- Animation des réunions de travail ou d'ajustement du projet.
- Documentation et analyse des informations sur les technologies informatiques récentes pour répondre à

un besoin de compréhension ou de recherche d'information
- Recherche de solutions pertinentes pour la résolution de problèmes techniques à partir de :
    - sites spécialisés
    - communautés de spécialistes des données accessibles par internet.
    - autres
- Sélection de l'outil d'analyse de données en fonction des standards de l'écosystème technique du projet - Détection des valeurs anormales dans le jeu de données / Validation des données par la détection de valeurs anormales
- Nettoyage et traitement des données exploitables à l'aide d'une bibliothèque logicielle - Constitution d'un jeu de donnée au format de donnée préalablement identifié/sélectionné - Encodage des données au format adapté à l'aide de l'outil préalablement sélectionné - Génération de données pour augmenter la quantité de données exploitables
- Réduction de la dimensionnalité des données
- Visualisation des données à l'aide d'outils
- Identification du modèle d'apprentissage optimal en fonction du problème à résoudre, des données disponibles et de leurs natures
- Sélection de l'outil (langage, bibliothèque, framework, plateformes)
- Entraînement et exploitation d'un modèle d'apprentissage supervisé* à l'aide d'outils préalablement sélectionnés
- Classification ou prédiction d'une variable à partir d'un modèle d'apprentissage supervisé
- Réalisation de divers traitements à l'aide d'un modèle d'apprentissage : séries temporelles
- Utilisation de l'apprentissage non supervisé pour créer des catégories
- Evaluation de la performance d'un modèle d'apprentissage avec les métriques standards et spécifiques
- Identification des hyper-paramètres du modèle
- Amélioration de données d'apprentissage d'après une analyse des métriques de performance - Sélection d'une architecture d'apprentissage profond standard en fonction des données disponibles -Implémentation d'un modèle d'apprentissage profond préalablement sélectionné à l'aide d'une bibliothèque
- Utilisation d'un modèle d'apprentissage profond pré-entraîné (apprentissage par transfert )
- Versionnage du code source
- Partage des différentes sources à l'aide du système de versionnage
- Identification de la cible auprès de laquelle communiquer (interne/externe, équipe projet ou direction opérationnelle, tout public ou restreint..)
- Sélection des moyens de diffusion des résultats auprès de la cible
- Réalisation des visualisations afin de communiquer ses résultats

**Obstacles identifiés dont le dépassement permet d'atteindre l'objectif**

La qualité des données va demander une attention particulière pour être traitable
Quelles sont les données pertinentes, quelles sont les données non nécessaires.
La quantité de données est-elle suffisante pour utiliser un modèle d'apprentissage profond ?
Quel choix de modèle(s) prédictif(s) s'adaptent le mieux à notre problématique.

**Ressources qui permettent de construire les savoirs**

- Données clients (sous forme de fichier .csv)
- Mise en place / Exploitation d'une base de données
- Algorithmes de prédiction
- Langage Python (librairies de data Science : Pandas, Numpy)
- https://keras.io/
- https://www.tensorflow.org/

**Autres contextes d'utilisation de l'objectif :**

Ce projet est déployable à l'échelle d'autre sites en ligne de réservation.