

# Objectif :

Réaliser un projet de data science de bout en bout à partir de données de cas d'usage réel.

L'objectif sera alors d'appliquer un algorithme d'apprentissage automatique et/ou de réseau neuronal sur un ensemble de données réelles.



# Data :

Vous avez 2 sujets donc 2 jeux de données à votre disposition. Vous devez choisir un ensemble de données parmi les 2.

## 1. Supervision de lignes de bus :

Les transports publics font face tous les jours à des flux de passagers différents, il est important de mettre en place un outil leur permettant de **prédirer l'affluence** sur les 3 prochains jours.

L'objectif est de prédirer l'affluence des passagers sur des lignes de bus, par jour, pour les 3 prochains jours.

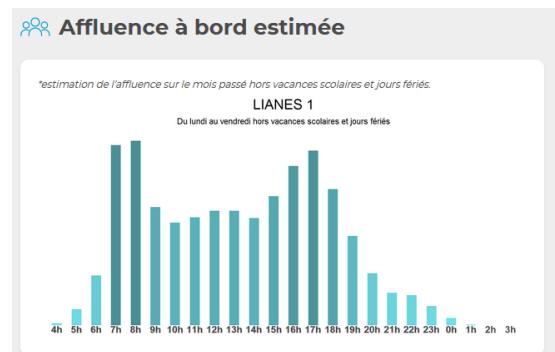
Vous aurez à disposition un historique du cumul journalier de voyageurs allant du **05 avril 2019 au 08 mars 2023**, pour toutes les lignes de bus de la compagnie.



Le jeu de données est composé de 36 901 observations et 3 variables explicatives originales qui sont, la date, la ligne de bus et le type de ligne de bus (jour ou nuit). La variable quantitative à prédirer est le nombre de passagers/usagers.

### Variable de sortie :

Nombre de passagers (variable numérique continue)



## 2. Le technicien en mobilité

Pour une compagnie de maintenance de réseau électrique, la sécurité du technicien est un enjeu fort. Ainsi, ceux-ci sont équipés d'EPI (Equipements de Protection Individuelle), qu'ils doivent porter durant leurs interventions (casque, gant, visière, tapis, etc.). HUPI travaille depuis plusieurs mois afin de développer des EPI "connectés". Ces EPI connectés prennent en compte les caractéristiques d'intervention du technicien, ainsi que son profil et habitude de travail, et doivent l'alerter automatiquement mais judicieusement, en cas de port non conforme des EPI durant intervention



L'objectif est de travailler sur de modèles auto-apprenants permettant de prendre en compte les caractéristiques propres de chaque individu/intervention afin de générer des alertes personnalisées et adaptées à chaque contexte d'intervention.

Vous aurez à votre disposition un ensemble de données issues des capteurs mesurant le port ou non des équipements. Les relevés des données des capteurs sont toutes les 5 secondes, donc vous aurez une observation toutes les 5 secondes. Toutes les variables binaires explicatives sont caractérisées par la manière suivante :

- 0 : équipement non porté
- 1 : équipement porté

La variable à prédire est la variable "danger".

### **Variable de sortie :**

Outcome (0 or 1)

## Méthodologie :

### Prétraitement et chargement des données :

Chargez l'ensemble de données avec lequel vous avez choisi de travailler. L'une des premières étapes du processus d'exploration des données, lorsque l'objectif final est de prédire le résultat, consiste à créer des visualisations qui aident à connaître le résultat et à découvrir les relations entre les attributs et le résultat.

Vous pouvez utiliser plusieurs outils de visualisation des données tels que les diagrammes à barres, les histogrammes, les diagrammes en boîte, les matrices de corrélation, les diagrammes à paires, l'analyse PCA.

Choisissez les variables que vous pensez utiles pour votre prédiction.

Veillez à éliminer les caractéristiques corrélées.

N'hésitez pas à créer de nouvelles caractéristiques en combinant d'autres caractéristiques.

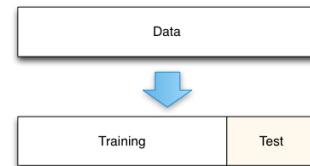
Si vous pensez qu'elles ont un impact négatif sur vos prédictions, supprimez certaines valeurs aberrantes.



### Ensemble de données d'ingénierie:

Divisez votre ensemble de données en 2 sous ensembles :

- training set
- test set



Si vous pensez que c'est utile, appliquez la normalisation sur votre ensemble de données. Veillez à appliquer la transformation inverse sur vos prédictions à la fin pour comparer votre résultat avec la vraie sortie.

### Définition du modèle :

Choisissez un algorithme d'apprentissage automatique et adaptez-le aux données, et/ou adaptez un réseau neuronal aux données.

Pour les deux modèles, essayez d'utiliser la recherche sur grille pour optimiser les hyperparamètres.

### Évaluer les modèles :

Évaluez chaque modèle sur l'ensemble de test et conservez le meilleur modèle.

Faites une visualisation de certaines prédictions et comparez-la aux véritables prédictions avec le meilleur modèle.

## Comment obtenir une bonne note ?



- Commentez chaque bloc de code et justifiez chaque choix.  
Une visualisation qui n'est pas commentée est considérée comme non faite.  
ex : Le box plot montre 3 valeurs aberrantes, nous les supprimons de notre jeu de données pour le reste de l'étude.
- Il est plus important d'avoir de mauvaises prédictions et d'avoir une bonne méthodologie et une bonne justification que le contraire.

Conseils :

- Utiliser la bibliothèque python sklearn pour l'apprentissage automatique :  
<https://scikit-learn.org/stable/>

## Environnement de travail :

Vous travaillerez avec les notebooks google colab par exemple.

Vous restituerez votre travail à partir d'un rapport écrit contenant l'ensemble de la documentation du projet. A partir de la documentation des spécifications du projet, des analyses des données, de la modélisation, jusqu'au résultats et l'analyse des résultats, c'est-à-dire l'ensemble des étapes constituant un projet de data science.

La restitution se fera aussi avec une soutenance qui aura lieu **31/03/2023 de 14h à 17h**.  
Nous planifierons en avance le planning des passages des groupes.

**La date limite de restitution du travail est donc le 31/03/2023 avec le rapport écrit et la soutenance.**

Si vous avez des questions, n'hésitez pas à me contacter à l'adresse suivante :  
[kattin.dassance@hupi.fr](mailto:kattin.dassance@hupi.fr)

Good luck ! :)