WILD CODE SCHOOL

Projets finaux Data 2023

Projets finaux Data

Votre projet final est professionnalisant, vous devez le réaliser dans les mêmes conditions et avec la même qualité que si vous étiez en poste… avec l'avantage d'être accompagné et de pouvoir demander de l'aide si besoin.

Objectifs
Ce projet vous permettra de continuer à monter en compétences, de travailler en groupe avec des méthodes agiles, et de disposer d'un projet dans votre portfolio, dont vous pourrez parler pendant les entretiens de recrutement.

Contenu
Les projets sont fournis par des partenaires de la Wild Code School. Les partenaires ont déjà réalisé ces projets en interne, et ils nous les partagent car ils les trouvent représentatifs de ce qu'ils attendent d'un junior dans le domaine. Cela permet de disposer d'un portfolio représentatif de vos compétences sur des cas d'usage réels.

ADN Tourisme

Projet Machine Learning

ADN Tourisme - l'organisme

ADN Tourisme est née le 11 mars 2020 du **regroupement des trois fédérations** historiques des acteurs institutionnels du tourisme, **Offices de Tourisme de France, Tourisme & Territoires et Destination Régions.**

En associant ainsi les représentants des trois échelons territoriaux métropolitains et ultramarins (offices de tourisme, comités départementaux et régionaux du tourisme), ADN Tourisme représente, au niveau national, les forces conjuguées de 1200 structures et 13 500 salariés.

Tout en tenant compte des compétences partagées et des activités propres à chaque échelon territorial et dans le respect du code du tourisme, ADN Tourisme a pour objectif de proposer à ses adhérents une offre de services innovante et une expertise de qualité. Elle a également pour ambition de développer des partenariats forts avec l'Etat et ses opérateurs, ainsi qu'avec les acteurs privés, dans la perspective d'inscrire son action dans une vision partagée d'un tourisme responsable et de qualité.

ADN Tourisme - l’objectif

ADN Tourisme recense environ 450k établissements en France. Les plus grandes catégories sont les hôtels et les restaurants. La base est mise à jour par les 1200 offices de tourisme. ADN Tourisme souhaiterait un outil pour auditer la qualité de données, notamment sur les 2 catégories Hôtels et Restaurants. Le but est de vérifier si ces établissements sont bien catégorisés.

ADN Tourisme - le projet

A partir des données des points d'intérêts de tourisme :

- Proposer un tableau de bord synthétique permettant d'accéder aux informations touristiques, notamment pour les catégories **Accommodation** et **FoodEstablishment**
- Créer un modèle de classification (Machine Learning) pour prédire la catégorie en fonction de la description de l'établissement
    - Par simplification, vous partirez sur les catégories Accommodation et FoodEstablishment
    - Vous commencerez par la région de votre choix
    - Vous indiquerez les métriques utilisées, et l'importance de chaque variable. L'interprétation des variables explicatives (monogrammes, bigrammes, nettoyage, etc…) sera plus importante que la précision dans ce projet.
    - Vous mettrez en évidence si des établissements devraient changer de catégorie d'après votre modèle, ou avoir une double catégorie.
    - Vous classerez les établissements que l'algorithme prédit comme mal catégorisé par leur probabilité
    - Une fois ce travail réalisé, vous pourrez l'étendre aux autres régions, puis éventuellement à d'autres catégories.

ADN Tourisme - défis techniques

- Problématiques big data : la base est découpée en fichier JSON. Chaque fichier JSON représente un seul établissement. Il y a donc un travail d'automatisation et de retraitement préalable
- Machine Learning : la classification ne pose pas de réel problème. Il faudra par contre bien travailler sur les métriques, et les probabilités de prédiction, afin de répondre au besoin de nettoyage.

Astuce : Vous pouvez volontairement mal classer certains établissements pour vérifier qu'une alerte remonte.

- Pour accéder aux données brutes, vous devrez vous inscrire sur la plateforme datatourisme.fr, c'est gratuit et immédiat.
- Les données CSV sur data.gouv.fr sont incomplètes, il vaut mieux partir de la plateforme.
- Vous pouvez obtenir les fichiers JSON en cliquant sur "créer un flux".

Transports publics

Projet Business Intelligence

Transports publics - l’organisation

La Wild Code School a effectué des missions en partenariat avec l’agglomération nantaise. Mais ce projet est réplicable dans toutes les grandes villes. Vous êtes donc libres de choisir une ville, qui sera plus adaptée à votre portfolio, afin de pouvoir communiquer sur les réseaux sociaux vers votre écosystème local.

Transports publics - le projet

La plupart des sociétés de transports (publiques ou privées) diffusent les données sur les horaires en open data. Le plus souvent, ces données sont “instantanées”. Par exemple :

- Vous pouvez savoir dans combien de temps va passer le prochain bus de la ligne 141 à l’arrêt Victor Hugo.
- Mais vous ne pouvez pas savoir à quelle heure sont passés tous les bus de la ligne 141 hier ou la semaine dernière.

→ **Il va donc falloir automatiser la collecte de données afin de se constituer un historique de données.**

Transports publics - étapes techniques

Voici une vision globale des étapes techniques pour un premier moyen de transport (par exemple les bornes de vélos en libre-service) :

- Récupération par fichier plat des données unitaires et affichage d'une carte
- Collecte via une API des données unitaires et mise à jour de la carte
- Automatisation toute les 15 minutes de la collecte via API et stockage de cet historique dans une base de données

Suivant l'heure du jour, vous verrez une sorte de "respiration", en fonction des lieux de sorties, des quartiers de bureaux ou des zones résidentielles.

Puis, vous pourrez ajouter un second moyen de transport, etc...

Transports publics - limites et outils

- Suivant les villes il y a des bornes de comptage sur les voies cyclables, et de trafic sur les voies rapides
- Concernant les transports en commun, vous obtiendrez la fréquence de desserte, mais vous n'avez généralement pas accès aux données sur la fréquentation (nombre d'usagers)

Pour automatiser la collecte, votre formateur vous donnera accès à une quête nommée "schedule". Dans un process réel en entreprise, vous pourriez payer pour un serveur cloud. Ici, un ordinateur d'un membre du groupe pourra jouer le rôle de serveur, il faudra le faire tourner pendant minimum 48h pour avoir des résultats représentatifs.

Enedis - scraping

Projet Business Intelligence

Enedis - l’entreprise

Enedis est une entreprise de service public. Elle exploite, modernise et dépanne le réseau d'électricité en France 24h/24 et 7j/7. Enedis est organisée par région.

Enedis - le projet

- Problématique principale : Réaliser une veille sur la marque Enedis
- Problématique secondaire (au choix) :
  - produire un dashboard synthétique sur la consommation/distribution électrique basée sur l'open-data fournie par Enedis, afin d'encourager l'utilisation et la communication de ces données.
  - effectuer du Sentiment Analysis sur les articles collectés pour la veille, en établissant les temps forts positifs et négatifs dans l'historique de la marque

Enedis - Problématique principale

Enedis souhaite disposer d'un outil afin de suivre la fréquence à laquelle l'entreprise est citée en ligne. Pour cela, vous allez devoir développer un script de scraping ou via API afin de collecter l'ensemble des titres, journaux et dates, pour lesquels la marque Enedis est citée.

Vous effectuerez ensuite un tableau de bord présentant les médias qui en parle le plus, les saisonnalités (pic de citation), etc… Vous pourrez compléter par du Sentiment analysis (positif/négatif) ou de l'affichage de mots clés (wordcloud), ou tout outil ou visualisation qui vous semble pertinent.

N'hésitez pas à compléter par d'autres sources (comme google trends), comme des réseaux sociaux (twitter par exemple).

Enedis

Projet Machine Learning

Le projet

A partir des données de consommation électrique par région, pour les sites de soutirage inférieurs à 36kVA :

- Produire des indicateurs macro des usages régionaux dans le temps
- Créer un modèle de régression (Machine Learning) pour prédire la consommation en fonction de paramètres météorologiques et de calendrier
- Créer un outil permettant à un utilisateur d'ajuster des moyennes mensuelles pour au moins 2 variables, et d'évaluer l'impact sur la consommation électrique.
    - Les variables seront "température moyenne" et "quantité de pluie". Vous pouvez ajouter toute variable qui a un impact sur votre modèle.

Enedis - défis techniques

Vos variables explicatives seront

- la météo
  - Par simplification, vous partirez des données météo des préfectures des départements des régions. Vous pouvez les agréger, ou les conserver en variables multiples.
  - Vous commencerez par les régions Centre-Val de Loire et Hauts-de-France
  - Vous indiquerez les marges d'erreur, les métriques utilisées, et l'importance de chaque variable. L'interprétation des variables explicatives sera plus importante que la précision dans ce projet.
- et le calendrier
  - week-end
  - vacances scolaires de la région zone B
  - jours fériés
  - confinements