Fiche n°7 – Annexe n°3 - Contenu de la formation de data science

I - Module d'introduction à la data science (18 heures)

1. Introduction : quel est l'intérêt de centraliser la data science à la DGFIP ?

- Présentation de la DTNum, du pôle donnée, et de l'équipe data science
- Histoire et enjeux dans le champ de la data science
- Description de quelques projets de data science à la DGFIP
- Les prérecquis pour suivre la formation (python, éventuellement une familiarité avec les notebook Jupyter)
- Exemple d'une régression simple, de ses limites et de l'intérêt de suivre la formation de data scientist au sein de la DGFiP

2. Apprentissage supervisé pour les régressions : comment estimer un bien immobilier ?

- L'apprentissage supervisé / non supervisé : les données sont-elles étiquetées ou non ?
- Qu'est-ce qu'un modèle de machine learning, comment entraîner un modèle ?
- Pourquoi partager les données en un set d'entraînement et un set d'évaluation ?
- Présentation des modèles de régressions linéaires et de ses variantes
- Rapide revue des modèles à base d'arbres

3. apprentissage supervisé pour les classifications : cette tumeur est-elle ou non cancéreuse ?

- Les modèles de classification les plus courants
- Comment mesurer la qualité d'une classification ?
- Comment rééquilibrer un jeu de données ?
- Comment intégrer la problématique métier à un problème de classification ?

4. clustering : comment classifier des données qui ne sont pas déjà étiquetées ?

- Présentation des principaux modèles de clustering
- Evaluer la qualité d'un clustering ? Difficile et nécessaire !
- Les applications du clustering : de la détection des fraudeurs à la segmentation d'images

5. construction et mise en production d'un modèle : de l'exploration des données à la mise en production

- Les nettoyages des données : pratiques courantes
- La gestion des données manquantes
- Création d'un pipeline : unifier l'ensemble du processus de traitement
- Déploiement d'un modèle, scalabilité et mise en production

II - Module approfondissement en data science (12 heures)

6. interprétabilité et explicabilité d'un modèle : comment le modèle fonctionne, et pourquoi dans tel cas particulier prend-il telle décision ?

- Mesurer l'importance des variables dans les prises de décisions
- Représentation des arbres de décision
- Interprétation de modèles avec les algorithmes LIME et SHAP
- Sélectionner les variables pertinentes et réduction de la dimension

7. nlp ou traitement automatique du langage : lorsque la machine tente de comprendre les mots

- Les étapes de nettoyage du texte : lemmatisation et racinisation
- Topic modelling ou comment classifier des textes non labellisés
- Les plongements : faire correspondre à un mot ou un texte un vecteur de réels
- Utiliser les plongements pour classifier, faire de l'analyse de sentiment ou du paraphrase mining

8. deep learning : comment un véhicule autonome identifie-t-il les piétons visibles ?

- Du machine learning au deep learning : quelles différences ?
- Le principe de rétropropagation du gradient ou comment apprendre de ses erreurs
- Les architectures classiques de réseau de neurones
- Le transfert learning : modèle pré-entraîné ou fine-tuning ?