**Data Science Data Analytics Engineering – Période bloquée du 24 septembre au 5 octobre**

**I-           Organisation**

**Organisation du cours**: les cours du matin (9h – 12h15) sont structurés autour d’un modèle,  les fondements sous-jacents et des pratiques de manipulation des données en data science. Un atelier (LAB) est prévu chaque après-midi (13h45 – 17h) pour mettre en pratique les méthodes vues en cours pour l’analyse de données réelles. Chaque journée débute par un récapitulatif des éléments vus la veille pour permettre aux étudiants de revenir sur les points qui restent à éclaircir. En plus des travaux pratiques de l'après-midi, nous présenterons des cas de projections issus de l’industrie.

**Objectifs :** proposer une solution pour un problème “Data Science” en respectant les étapes clés : inspecter les données, choisir un modèle, définir des mesures de qualité, apprendre le modèle qui donne la meilleure qualité, visualiser et interpréter les solutions.

**Outils**: nous utiliserons python 3 comme langage de programmation sur des notebooks Jupyter avec des librairies tel que scikit-learn, pandas, matplotlib, NumPy,  SciPy, .... Les premières séances serviront pour la prise en main de ces outils. La plateforme logicielle sera installée dans la salle de TP. Toutefois, les recommandations pour installer les outils nécessaires pour travailler sur une machine personnelle seront données et expliquées lors des TPs. Les étudiants seront libres de travailler sur une machine de la salle ou sur une machine personnelle mais nous recommandons les machines de la salle pour des raisons de pratiques. Un environnement de travail de notebook Jupyter sera mis à disposition des étudiants et sera accessible via un simple navigateur sur leur machine personnelle.

**Projet**: un projet final basé sur une compétition Kaggle fera l'objet de présentations. Les étudiants travaillent en binôme. Ce projet leur permettra d’utiliser la plupart des notions vues dans le cours et de les appliquer sur un jeu de données non trivial.

**Modalité d’évaluation**: Quiz de 15 minutes à la fin de chaque cours (QCM), le premier projet sera exposé le jour 10 et le deuxième projet sera exposé le jour 20 (voir plan détaillé  ci-dessous).  Le code bien documenté est à rendre la veille du jour de chaque exposé avec un petit rapport de trois pages minimum pour expliquer les grandes lignes du projet. Le code doit être executable dans un environnement semblable à celui de la salle de TP.

**Note finale**:  exposé 40%, code projet 40%, quiz 20%

**Prérequis**

Algèbre linéaire, analyse (différentiation, fonctions de plusieurs variables, séries), probabilités (bases du calcul de probabilités, conditionnement, principales distributions), bases de statistiques inférentielles (modèles statistiques, estimation, test)

En informatique, connaissance de base (variables, fonctions, ...) avec une pratique de python, des notebooks Jupyter ou d’un langage scripté.

**Livres recommandés**

**Think python** d’Allen Downey est un excellent livre d’introduction au langage python. Il est disponible gratuitement sur :

<https://greenteapress.com/wp/think-python/>.

Ses autres livres de stats sont aussi excellent.

Plus facile que “Elements of statistical learning”, le livre “**An Introduction to Statistical Learning**” de D. Witten, G. James, R. Tibshirani, et T. Hastie est disponible en pdf a l’adresse:

[https://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf](https://www-bcf.usc.edu/%7Egareth/ISL/ISLR%20First%20Printing.pdf)

**Titre : Pattern Recognition and Machine Learning**

Auteur : Christopher Bishop

Editeur : Springer

ISBN : 978-0387310732

**Titre : Python Machine Learning**

Auteur : Sebastian Raschka

Editeur : Packt Publishing

ISBN : 978-1783555130

**Titre : Building Machine Learning Systems with Python**

Auteur : Willi Richert, Luis Pedro Coelho

Editeur : Packt Publishing

ISBN : 978-1782161400

**II- Programme**

Partie 1 : Programme 24 septembre au 5 octobre

Jour 1 : Intro, data science et régression linéaire

* Après-midi :
  + Présentation du programme, de la compétition Kaggle, du travail attendu
  + Mise en place de l'environnement de travail (Jupyter notebook) et des outils (python, scikit-learn, numpy, pandas, ...),
  + Pratiques python, et fonctionnalités pandas
  + Qu'est-ce que la data science ?
    - Data Science, Machine Learning, Intelligence Artificielle, Deep Learning,
    - Finalités et étapes d'un projet,
    - Supervisée ou non-supervisée,
    - Régression ou classification.
  + Jeux de données classiques: Titanic, iris, ...
  + LAB : Régression(s) sur des datasets simples
    - Régression linéaire,
    - Régression logistique,
    - Conditions sur les données, cas d'application, p-value, R^2, intervalles de confiance, ...

Jour 2 : Modèles ou Algorithmes

* Matin :
  + Pratique du maximum de vraisemblance
  + Importance de la linéarité : définition et tests statistiques associés
  + Impact de la multi-colinéarité
  + Régression polynomiale
* Après-midi :
  + Présentation du projet final : compétition Kaggle.
  + LAB : On continue sur les différents types de régression avec des datasets non linéaires

Jour 3 : Data visualisation et python

* Matin :
  + Visualisation des données, principales visualisation, matplotlib, plot.ly
  + Best practice en python pour la data science; structure de données avec pandas.
* Après-midi :
  + LAB: travaux pratiques de visualisation d’un dataset, exploration et identification des aberrations

Jour 4 : Gradient stochastique

* Matin :
  + Recap : l'analyse prédictive supervisée : Modèle ou Algorithme, minimisation d'une fonction de coût, score, prédicteurs et outcome, ...
  + Découpage en test, train validation ; validation croisée
  + Algorithme du gradient stochastique, théorie, pratique et applications
* Après-midi :
  + LAB : Gradient stochastique, convergence, paramétrisation, visualisation, scikit-learn, batch mode, conditions
  + Cas : L’utilisation du gradient stochastique dans le service AWS ML

Jour 5 : Arbres et Forêts

* Matin :
  + Arbre de décisions, forêts aléatoires, XGBoost
  + Ensembling et Boosting
  + Score de classification : matrice de confusion, AUC, F1, ...
* Après-midi :
  + Présentation d’un cas : classification d’échantillons de musique électronique pour Splice.com
  + LAB : forêts aléatoires et XGBoost

Jour 6 : Support Vector Machines et overfitting

* Matin :
  + Support Vector Machines avec différents kernels
  + Overfitting, compromis biais-variance, régularisations L1 et L2, Ridge et Lasso
  + Courbes d’apprentissage : détecter et corriger l'overfitting
* Après-midi :
  + LAB : SVM
  + LAB : Régularisation

Jour 7 : Création et sélection des variables

* Matin :
  + Modèle bayésien naïf
  + Feature engineering approche manuelle, brute ou approche bayésienne
  + Améliorer les variables : Box-Cox, valeurs manquantes ou aberrantes
  + Bootstrapping
  + Sélection des variables
  + Curse of dimensionality
* Après-midi :
  + LAB : nettoyage des données, feature engineering et sélection

Jour 8 : Classification déséquilibrée

* Matin :
  + Le paradoxe de la précision
  + Méthodes de sous-échantillonnage et de sur-échantillonnage
  + SMOTE
* Après-midi :
  + Cas : prédiction de résiliation dans l’assurance
  + LAB : sur dataset caravan
  + Point projet

Jour 9 :Séries temporelles

* Matin :
  + Exponential smoothing
  + Modèles linéaires AR, MA, ARIMA
  + Décomposition en tendance, saisonnalité et résidus
* Après-midi :
  + Cas : Prédiction de la demande pour un producteur d'électroménager
  + LAB sur séries temporelles

Jour 10 : Présentations des projets

* Matin : session ouverte à l’initiative des étudiants, questions / réponses, intro aux outils cloud AWS ou Google.
* Après-midi : Présentations des projets étudiants

Partie 2 : programme du 08 au 19 octobre

Jour 11 : réduction de dimension

* Matin :
  + Rappels et Introduction
  + Problèmes liés à la grande dimension
  + Analyse en composante principale
  + Décomposition en valeur singulière
  + D’autres techniques (LDA, ICA, PLS) et exemples d’application
* Après-midi :
  + Cas : volume de stockage et temps de calcul
  + LAB : classification avant et après la réduction de dimension

Jour 12 : évaluation des modèles et optimisation des hyper-paramètres

* Matin :
  + Rappels  et introduction
  + La nécessité de diviser les données
  + La validation  croisée
  + Optimisation des paramètres
  + Mesure de la performance pour la régression ou la classification
* Après-midi :
  + Cas : comment choisir et apprendre  un bon modèle
  + LAB :

·       Classification multi-classes avec interprétation des performances en fonction des paramètres

·        Estimateurs, k-fold, courbes de validation.

Jour 13 : Clustering hiérarchique et clustering non hiérarchique

* Matin :
  + Rappels et introduction
  + Principe
  + Le choix des distances
  + Le critère d’agrégation
  + Le critère de partitionnement
  + k-means, mean-shift, k-medoids et d’autres variantes
* Après-midi :
  + Cas : initialisation et regroupement stable
  + LAB : clustering avec DBSCAN.

Jour 14 : modèles graphiques : apprentissage et inférence

* Matin :
  + Rappels et introduction
  + Notion d’indépendance de variables
  + Formulation
  + Méthodes pour trouver un choix optimal
  + Exemple d’application : segmentation
* Après-midi :
  + Cas : regroupement d’individus
  + LAB : Segmentation d’une image 2D (individu=pixel): application d’un modèle et visualisation des résultats

Jour 15 : Application pas à pas  et démo  pour la détection de visages 2D

* Matin :
  + Rappels et introduction
  + Formulation du problème
  + Le choix des modèles
  + Apprentissage du modèle
  + Optimisation des paramètres
  + résultats et interprétation
  + discussion autour des choix et des performances
* Après-midi :
  + Cas : détection de visages 2D
  + LAB : extraction des patterns, évaluation d’un modèle vu en cours .

Jour 16 : Manifold Learning pour la transformation des données

* Matin :
  + Rappels et introduction
  + Pourquoi transformer les données ?
  + Isomap
  + Multi-dimensional Scaling (MDS)
  + Transformations basées sur les noyaux
  + Transformations locales et adaptative
* Après-midi :

·      Cas :

·        classification des données avant transformation

·       visualisation et classification des données après transformation

Jour 17 : mixture de modèles gaussiens

* Matin :
  + Rappels et introduction
  + Les inconvénients  de k-means
  + Combinaison de plusieurs modèles
  + Sélection des composantes
  + Estimateurs de modèles
* Après-midi :

·      Cas : classification avec sélection de modèle par

·       BIC

·        AIC

·      visualisation avec différents noyaux

Jour 18 : clustering de données de grande tailles (Big Data)

* Matin :
  + Rappels et introduction
  + Echantillonnage de données en sous-ensembles
  + PAM sur les sous-ensembles
  + Choisir une stratégie pour l’ensemble complet
  + Algorithme CLARA (Clustering LARge Applications)
  + CLARANS, CURE, etc.
* Après-midi :

·      Cas : clustering du big data avec CLARA

Jour 19 : sélection de patterns

* Matin :
  + Rappels et introduction
  + Sélection des patterns utiles
  + Méthodes basées sur la variance
  + Méthodes récursives (RFE)
  + Méthodes récursives avec validation croisée (RFECV)
  + Autres variantes
* Après-midi :

·      Cas : augmenter la performance et réduire la dimension

·      LAB : RFE et RFECV pour la classification de digits

Jour 20 : exposés et discussion

* Matin :
  + Rappels des méthodes vues
  + session ouverte à l’initiative des étudiants autour  d’une application complète en IoT  pas à pas
* Après-midi :

·      LAB  : présentations  des projets étudiants.