Master 2 MASERATI

Méthodes Appliquées de la Statistique et de L'Econométrie à la Recherche, l'Analyse et le Traitement de l'Information

Parcours Data Analyst (DA) et Data Science (DS)

Livret de l'Apprenti 2020 - 2021

FACULTÉ DE SCIENCES ÉCONOMIQUES ET DE GESTION
Département d'Économie
Mail des Mèches - 61 av. du Général de Gaulle
94010 Créteil Cedex

Responsables

Pierre Blanchard (Bureau 218 – 2ème étage) - pierre.blanchard@u-pec.fr
Zineb Abidi (Bureau 225 - 2ième étage) – zineb.abidi@upec.fr

Gestionnaire Pédagogique du Master Economie Appliquée parcours MASERATI

Chantal Kuadjovi (Bureau 6 - RDC) – chantal.kuadjovi@u-pec.fr

Le M2 Maserati est classé 6ième en 2020 au classement SMBG

TABLE DES MATIERES

1) Présentation générale de la formation

OBJECTIFS

La formation a pour but de préparer au métier de chargé d'études statistiques (data analyst et data scientist):

- par l'organisation des enseignements qui visent à faire acquérir aux étudiants les techniques et outils fondamentaux et à leur en permettre l'application sous la forme de cas pratiques proposés par les enseignants issus de l'université et du monde professionnel (qui constituent plus de la moitié du corps enseignant),
- par la mise en situation que constitue l'activité professionnelle en apprentissage et le mémoire professionnel qui clôture le cursus.

DEBOUCHES

Le Master 2 MASERATI permet d'acquérir un profil hautement qualifié, grâce à des enseignements interdisciplinaires portant sur des méthodes économétriques et statistiques appliquées à plusieurs domaines et à tous types d'entreprises. En effet, tous les métiers de l'entreprise se trouvent confrontés à la croissance constante et massive des données. Cette abondance des données, qui a connu une croissance particulière ces dernières années, a fait émerger de nouveaux enjeux et besoins tant pour les start-ups que pour les grandes entreprises. Par conséquent, les profils de data analyst et data scientist sont les plus recherchés sur le marché du travail en France et à l'international.

- Data analyst :

Avec des enseignements transversaux portant sur les méthodes économétriques et statistiques appliquées à l'économie, à la finance, au marketing..., le Master 2 MASÉRATI parcours Data Analyst permet aux étudiants d'accéder à une large palette de métiers de chargés d'études, aussi bien dans les entreprises privées que publiques, opérant dans tous les secteurs d'activité (banque, assurance, études et conseil, télécommunication, distribution, automobile, aéronautique, administrations centrales, organismes de recherche privés et publiques…). Le data analyst est un chargé d'études en statistique / marketing / économie / techniques quantitatives / finance quantitative. Il utilise les sources d'informations statistiques (internes et externes à l'entreprise) pour répondre à des questions relevant de l'analyse de la politique commerciale, des ressources humaines, de la finance, du marketing, de l'analyse des politiques publiques et privées...

- Data scientist :

Les données massives « big data » ont connu une évolution spectaculaire ces dernières années. Ces données sont sous formes structurées et non structurées (numériques, textes, vidéo, sons…), elles proviennent de sources variées (données classiques d'entreprise, page web, réseaux sociaux, capteurs, open data…) et sont de qualités diverses. Ayant une formation à l'interface entre modélisation mathématique, statistique et informatique, le data scientist est chargé de collecter ces données, de les mettre en forme et de valoriser l'information contenue dans ces données. La variété des données se retrouve dans les applications métiers : les data scientists travaillent dans tous les secteurs d'activité et répondent à tous types de questions (gestion de la relation client, maintenance prédictive, ressources humaines, prévisions financières, objets connectés, détection de fraude, cyber-sécurité, gestion des risques, exploitation des données clients des banques et assurances ou encore des données sur internet (Google, Twitter, etc…), ainsi que l'évaluation des politiques publiques (INSEE, OCDE, Banque de France…)).

CONDITION D'ADMISSION

Le Master 2 MASERATI parcours Data Analyst et Data Science est ouvert aux titulaires d'un diplôme de niveau Bac + 4 (1ère année de Master) en économétrie, techniques quantitatives, statistique, économie, finance, gestion... ainsi qu'aux diplômés des Écoles d'Ingénieurs et des Écoles de Commerce souhaitant acquérir une formation complémentaire en méthodes quantitatives. Ce Master est accessible en formation par alternance (contrat d'apprentissage en priorité mais aussi en contrat de professionnalisation). Le Master 2 n'est pas proposé en formation initiale.

Les étudiants du M1 Maserati de la FSEG-UPEC ayant validé leur première année de Master sont admis de droit.

Pour les candidatures extérieures au M1 Maserati, une présélection est effectuée sur dossier (relevés de notes des années antérieures, lettre de motivation, CV…). Les candidats présélectionnés sont reçus pour un entretien, destiné à vérifier leurs motivations pour le métier de chargé d'études, ainsi que la qualité de leur expression orale.

La non-admission n'est pas susceptible de recours sauf erreur matérielle.

Après la sélection universitaire, les candidats sont sélectionnés par les entreprises selon leurs propres critères. Ils doivent signer un contrat d'apprentissage ou de professionnalisation au plus tard 3 mois après le début de la formation. La non-obtention d'un contrat d'alternance par un étudiant présélectionné se traduira automatiquement par une non admission au M2 MASERATI.

DUREE ET RYTHME DE LA FORMATION

La formation est d'une durée d'un an et se fait dans le cadre de l'apprentissage sur un rythme de 2/3 jours à l'université et de 3/2 jours en entreprise entre septembre et mi-mai (cf. le calendrier ci-après pour plus de précision). Elle comprend 450 heures d'enseignement (y compris des heures de travail personnel).

Les enseignements sont organisés en deux semestres. Le 1er semestre comporte trois modules : un module d'enseignements fondamentaux, un module outils de l'entreprise et un module d'enseignements spécialisés. Le 2ième semestre est un module d'activités professionnelles en entreprise. Ce dernier débouche sur la rédaction et la soutenance du mémoire professionnel de fin d'études.

COMPETENCES VISEES

A l'interface de plusieurs services, le data analyst et le data scientist mobilisent à la fois des compétences informatiques et statistiques. Le Master 2 Maserati met l'accent sur la valorisation de l'information, et forme ainsi des experts analystes quantitatifs ayant des bases solides en langage de programmation et en informatique, notamment en big data pour le parcours DS.

Le master MASERATI Data Analyst et Data Science permet d'acquérir les compétences suivantes :

- Conduire en collaboration avec les services compétents une étude économique, financière ou marketing dans toutes ses dimensions (conception, traitement statistique, rédaction du rapport, présentation des résultats…).
- Produire du reporting automatisé de qualité professionnelle.
- Élaborer, adapter et estimer des modèles statistiques et économétriques aussi bien sur données individuelles (scoring, segmentation…) que sur séries temporelles (prévision…),
- Programmer la mise en œuvre de méthodes statistiques et économétriques complexes,
- Collecter et analyser des données sur internet.
- Maitrise des techniques quantitatives (analyse de données, économétrie des données individuelles, de panel, sur séries temporelles…),
- Maitrise en logiciels (SAS, SQL, R, Python…),
- Avoir des compétences métiers (scoring, datamining, Text mining…),

Plus spécifiquement, pour le parcours DS :

- Visualiser des données pour faciliter la prise de décision,
- Manipuler des bases de données volumineuses et complexes (Hadoop, Spark…),
- Elaborer, adapter et estimer des modèles prédictifs ou explicatifs à l'aide d'algorithmes de machine learning.

Le Master 2 MASERATI met l'accent sur la valorisation de l'information, et forme ainsi des experts analystes quantitatifs ayant des compétences techniques et des méthodologies appliquées solides. Le Master 2 insiste également sur les langages de programmation informatique pour les deux parcours (Data Analyst (DA) et Data Scientist (DS)) avec un approfondissement en big data pour le parcours DS).

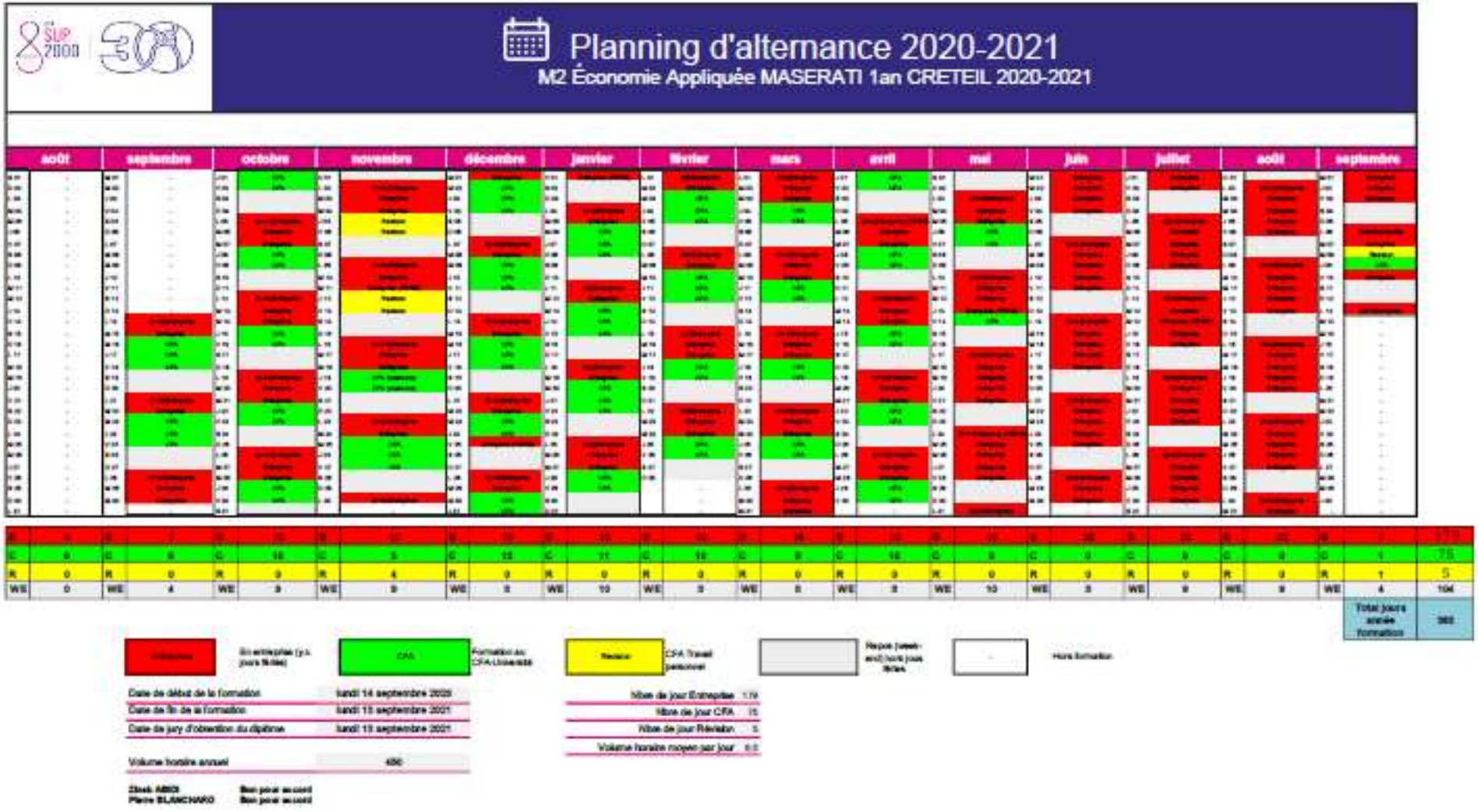
PEDAGOGIE

La formation est adossée au laboratoire de recherche ERUDITE (EA 437) et 8 enseignants-chercheurs de l'ERUDITE enseignent dans le Master 2 un total de 10 cours.

Les enseignements sont dispensés sous forme de cours et d'exercices sur ordinateur (tous les cours ont lieu dans une salle informatique, dédiée au M2). Les professionnels représentent plus de 50% du corps enseignant.

Les étudiants passent durant l'année de formation la Certification SAS « Programmation de Base » organisée par SAS Institute.

2) Calendrier du M2 MASERATI *Parcours DA et DS* 2020-2021



COURS ET ALTERNANCE

La présence aux cours (en vert) est obligatoire. Une feuille de présence est établie à chaque cours (par matinée et après-midi) avec l'indication des heures qui est signée par les apprentis et l'enseignant. Elle est transmise par le secrétariat du diplôme, au CFA qui se chargera de la transmettre à l'entreprise. Au bout de 2 absences non justifiées, l'apprenti est convoqué dans le bureau du responsable du diplôme. Si les absences non justifiées sont excessives, cela peut conduire à l'exclusion de la formation de l'apprenti. La liste des absences justifiées est donnée aux apprentis au début de l'année et l'apprenti doit fournir un justificatif écrit pour que son absence soit considérée comme justifiée. Les retards sont également notés sur la feuille de présence. Un retard supérieur à 15 minutes sera considéré comme une absence sauf justificatif valide.

La présence en entreprise (en rouge) est obligatoire : les étudiants sont salariés de l'entreprise d'apprentissage et doivent respecter les règles en vigueur dans leur entreprise.

Les zones en jaune représentent journées de révisions (5 jours par an).

RELATION UNIVERSITE-ENTREPRISE-CFA SUP 2000

La fiche de liaison CFA-Université / Entreprise est établie par l'entreprise ; elle est signée par le responsable du Master 2 et par l'apprenti et elle est transmise à l'entreprise.

Le maître d'apprentissage (ou maître de stage) est désigné sur la fiche de liaison. Il est salarié de l'entreprise qui a embauché l'alternant en apprentissage. Il a un diplôme ou titre équivalent à la qualification visée par l'apprenti.

Le tuteur pédagogique universitaire est un enseignant du Master 2 désigné au sein de l'équipe pédagogique. Il encadre en liaison avec le maitre d'apprentissage l'activité de l'apprenti en entreprises et il effectue les deux visites obligatoires en entreprise.

Deux visites en entreprise sont programmées en janvier/février et en septembre durant lesquelles une appréciation est donnée sur l'attitude et la méthode de travail ainsi que sur les activités menées durant la période. La deuxième visite donne lieu à la soutenance du rapport d'activité en entreprise de l'apprenti devant le maitre d'apprentissage et le tuteur universitaire. La soutenance a lieu dans l'entreprise d'accueil.

Une réunion de rentrée est prévue dans la première quinzaine de novembre entre les maîtres d'apprentissage, les apprentis, le CFA et l'équipe pédagogique. Elle permet la rencontre et l'échange entre les différents partenaires de la formation.

Deux réunions de bilan pédagogique ont lieu en janvier et mai qui réunissent les maîtres d'apprentissage, les délégués et l'équipe pédagogique. Un bilan est transmis aux participants à ces deux réunions.

3) Maquette du M2 MASERATI *Parcours DA + DS*

Intitulé de l'UE et du cours	Durée	ECTS	DA/DS	Contrôle
UE1: Cours fondamentaux	**120**	**11**		
Certification SAS	24	3	DA+DS	Examen
Introduction à R	24	2	DA+DS	Examen
Introduction à SAS	24	2	DA+DS	Examen
Rappels d'économétrie	24	2	DA+DS	Examen
SAS avancé	24	2	DA+DS	Examen
UE2 : Outils de l'entreprise	**174**	**6**		
Anglais	24	6	DA+DS	Projet
Aspects juridiques et protection des données	12	Projet	DA+DS	Projet
Data visualisation	12	Projet	DA+DS	Projet
Introduction au Datamining	24	Projet	DA+DS	Projet
Introduction au Web Scraping	12	Projet	DA+DS	Projet
Logiciel SGBD MySQL	24	Projet	DA+DS	Projet
Machine/Deep learning	30	Projet	DA+DS	Projet
Rappels de Python	12	Projet	DA+DS	Projet
Scoring	24	Projet	DA+DS	Projet
UE3 : Cours de spécialité	**138 (DA)/144 (DS)**			
Cointégration et Modèles VAR	18	Projet	DA	Projet
Econométrie des études d'impact	18	Projet	DA	Projet
Marchés financiers et risque	24	Projet	DA+DS	Projet
Modèles de durée	18	Projet	DA	Projet
Econométrie des variables qualitatives	24	Projet	DA	Projet
Panel	18	Projet	DA	Projet
Python avancé	12	Projet	DS	Projet
SAS pour le big data	24	Projet	DS	Projet
Technologie big data	24	Projet	DS	Projet
Text mining	18	Projet	DA	Projet
Text mining	24	Projet	DS	Projet
Web mining et Web Analytics	24	Projet	DS	Projet
Python avancé : Traitement géospatial et intro aux graphes	12	Projet	DS	Projet
UE4 : Rapport d'activité, projet, IIGP	**18 (DA)/12 (DS)**			
Module IIGP		3	DA+DS	Cf. ci-après
Projet en anglais	18 (DA) - 12 (DS)	10	DA+DS	Projet

Rapport d'activité en entreprise		30	DA+DS	Rapport
Total heures d'enseignement	450	60	DA+DS	

NB1 : UE signifie Unité d'enseignements qui regroupe plusieurs Enseignements Constitutifs d'une Unité d'Enseignement ECUE sous une dénomination commune. ECTS est l'acronyme de **"**European Credit Transfer System". Les ECTS servent à pondérer les notes obtenues.

NB2 : Modalités particulières du contrôle des connaissances : la note de « Initiative, Implication et Gestion de Projet » (IIGP)

Outre les notes relatives

- Aux examens sur les matières obligatoires (4 x 2 ECTS),
- A l'examen de certification SAS (3 ECTS),
- Aux interrogations écrites et orales en anglais (4 ECTS),
- Au projet à soutenir et à rédiger en anglais (12 ECTS dont 2 pour l'anglais),

on définit dans le contrôle des connaissances une matière composite « Initiative, Implication et Gestion de Projet », IIGP, faisant l'objet d'une note bonus d'au maximum de 1 sur 20 (comptant pour 3 ECTS) élaborée en fonction des différentes rubriques suivantes:

- Participation au salon SMBG
- Participation aux salons de l'UPEC
- Exercice du rôle de délégué
- Présence en cours et ponctualité
- Note de qualité de la participation en cours
- Réalisation du trombinoscope
- Réalisation des deux bilans pédagogiques
- Participation à la présentation du M2 aux réunions d'information pour les L2 et L3
- Gestion de la liste des anciens étudiants et animation du groupe LinkedIn du M2 Maserati
- Développer la présence du M2 Maserati sur les réseaux sociaux –Twitter, Facebook…
- Projets libres selon proposition
- Publication des offres d'emplois sur un petit site internet
- Organisation de conférences "métiers"/ parcours d'anciens
- Note sur 20 relative à l'évaluation des enseignements ne faisant pas l'objet d'un examen spécifique (matières obligatoires, certification SAS, Anglais). Les enseignants de ces matières pourront, s'ils le souhaitent, mettre en place un mode d'évaluation des acquis des étudiants sur leur enseignement sous la forme qu'ils choisiront : mini-projet, relevé d'exercice à faire en cours, QCM…

4) Modalités de contrôle des connaissances

Les modalités de contrôle des connaissances sont différentes selon les matières. Les acquis en SAS 1, SAS avancé, rappels d'économétrie, introduction à R, et certification SAS sont évalués par un examen terminal de 2 ou 3 heures, début novembre et au semestre 2 pour la certification SAS. Les acquis en anglais sont évalués en contrôle continu à l'oral et/ou à l'écrit et lors de la soutenance du projet qui se déroule intégralement en anglais en présence de deux enseignants. Les acquis des autres matières sont évalués par la rédaction et la soutenance d'un projet sous la direction d'un tuteur universitaire. Les examens et les soutenances du projet font l'objet de deux sessions.

Les matières (ou ECUE) se compensent à l'intérieur des UE. De plus, les UE se compensent au sein du semestre. L'étudiant qui a la moyenne globale au semestre, même s'il ne l'a pas dans toutes les UE, acquiert le semestre. Toutes les UE du semestre sont alors validées, soit directement avec une moyenne au moins égale à 10, soit par compensation. Une note inférieure à 6/20 à la première session à un examen ou au projet implique que la matière ou le projet doit être repassé en seconde session. Une note inférieure à 6/20 à la seconde session à un examen ou au projet implique l'ajournement. Il n'y a pas de deuxième session pour la soutenance du rapport d'activités en entreprise.

De plus, les semestres se compensent entre eux au sein de la même année : le semestre 1 se compense avec le semestre 2, à l'exception du bloc "activités en entreprise". Une note inférieure à 10/20 au rapport de stage et à la soutenance est éliminatoire. Le redoublement n'est autorisé qu'en cas de situation exceptionnelle (maladie…).

Une seconde session d'examen est organisée au mois de septembre. Tous les étudiants qui n'ont pas validé certaines de leurs matières ont le droit à participer à cette seconde session s'ils n'ont pas obtenu la moyenne à l'UE.

Sous ces conditions, les étudiants choisissent les matières qu'ils souhaitent repasser parmi les matières qu'ils n'ont pas validées. La meilleure note sera retenue entre la première et la seconde session.

LE JURY D'EXAMEN

Le jury est composé des enseignants-chercheurs qui interviennent au cours de l'année. Il fait l'objet d'un arrêté désignant le Président et prévoyant une composition minimum. Ce jury se réunit après chaque session d'examens et de soutenance, en mai et en septembre.

LE PROJET ET RAPPORT D'ACTIVITE

Le projet est une analyse approfondie, argumentée et documentée sur une problématique ayant, de préférence, un lien direct ou indirect avec la thématique d'alternance ou de stage, les activités réalisées, ou les champs d'activités abordés. Il donne lieu à un rapport encadré par un tuteur à l'université. Un rapport d'activité en entreprise est aussi demandé. Il rend compte de la progression du travail de l'apprenti au sein de l'entreprise et des différentes missions réalisées par l'apprenti. La soutenance a lieu en entreprise en présence du tuteur en entreprise et du tuteur pédagogique universitaire.

ROLE DU TUTEUR PEDAGOGIQUE

Le tuteur pédagogique du projet aide l'étudiant à définir son sujet. Le rôle du tuteur consiste à le guider. Plus précisément, il aide l'étudiant durant sa réalisation pour l'amélioration de l'expression de la problématique, le choix des hypothèses, le choix de la méthodologie la plus adaptée au problème posé, éventuellement, sur des pistes de lectures. Il est donc nécessaire que l'étudiant entretienne avec le tuteur une relation régulière. Les premières semaines sont souvent déterminantes pour le succès du mémoire. La fréquence des contacts est à déterminer d'un commun accord mais elle reste à l'initiative de l'étudiant ; elle peut être modulée en fonction de son degré d'autonomie, de la vitesse d'avancement du travail et de la phase de mémoire dans laquelle il se trouve.

En ce qui concerne son rapport d'activité en entreprise, l'étudiant peut solliciter son tuteur pédagogique, en accord avec son tuteur en entreprise, sur tous les sujets qui portent aussi bien sur sa conception que sur les aspects techniques qu'il contient.

5) Plans de cours

RAPPELS D'ECONOMETRIE

Zineb ABIDI (Université Paris Est Créteil - ERUDITE)
Durée : 24 H
Parcours : DA et DS

Objectif du cours :

L'objectif du cours est de présenter des rappels d'économétrie sur les données individuelles en les illustrant avec des applications réalisées avec le logiciel SAS.

Une évaluation des connaissances acquises est effectuée en fin de formation sous la forme d'un examen sur ordinateur.

Prérequis :

- Statistiques (espérance, variance, distribution…) et économétrie.
- Connaissance du langage de programmation SAS.

Plan du cours :

Chapitre I) Rappels sur le modèle linéaire multiple et sur la méthode des moindres carrés ordinaires. Problèmes de spécification. Hétéroscédasticité.

Chapitre II) La méthode des variables instrumentales.

Bibliographie :

- L'ouvrage de référence est Verbeek, M., A Guide to Modern Econometrics, 2012, 4ième ed. Wiley.
- Bourbonnais R. (2018), Econométrie, 10éme édition – DUNOD.

SCORING

Zineb ABIDI (Université Paris Est Créteil - ERUDITE)
Durée : 24 H
Parcours : DA/ DS

Objectif du cours :

Ce cours fournit une compréhension générale des méthodes de scoring. Les étudiants seront plus précisément initiés aux notions, aux concepts et aux domaines d'utilisation du scoring.

Pour ce faire, le cours est à la fois théorique (principes et méthodes du scoring) et appliqué en utilisant les logiciels R et SAS.
Chaque chapitre du cours est structuré comme suit : rappels théoriques ou présentation théorique d'un concept du scoring, exemple d'implémentation de la procédure sur le logiciel R et/ou SAS dans le cours. Une évaluation des connaissances/compétences acquises est effectuée en fin de formation.

Prérequis :

- Statistiques (espérance, variance, distribution…) et économétrie.
- Connaissance des langages de programmation R et SAS.

Plan du cours :

Introduction

- Objectifs du scoring
- Quelques domaines d'application
- Différentes étapes de réalisation d'un score

1. Analyse discriminante et théorie de la décision

- L'analyse discriminante géométrique
- L'analyse discriminante probabiliste
  - Analyse discriminante linéaire (ADL)
  - Analyse quadratique discriminante (AQD)

2. La régression logistique

- Sélection pas à pas et sélection globale des variables
- Grille de scores
- Courbe de ROC et courbe de lift

2. Arbres de décision

- Principe de construction d'un arbre de décision
- Critères d'impureté
- Erreur de classification
- Rendre un arbre robuste (bagging et boosting)

3. Introduction aux réseaux de neurones artificiels

Bibliographie :

- Analyse discriminante – Application au risque et scoring financier. M. Bardos. DUNOD, 2001.
- Credit scoring and its applications. L.C. Thomas, D.B. Edelman and J.N. Crook, SIAM, 2002.

- Data Mining et statistique décisionnelle. S. Tufféry,Technip, 2007.
- Statistical learning from a regression perspective. Berk, R. A. Volume 14. Springer, 2008.
- An Introduction to Statistical Learning with Application in R. G. James, D. Witten, T. Hastie, R. Tibshrani, 2013.

INTRODUCTION A R

Christophe Barat (Editions Francis Lefebvre/UPEC)
Durée : 24 H
Parcours : DA/DS

Objectif du cours :

Ce cours vise une première approche des concepts de base du langage R, outil open source de traitement et d'analyse de données. A partir d'une description des bonnes pratiques, des objets et des possibilités offertes par le langage, l'objectif est l'acquisition d'une culture suffisante pour pouvoir évoluer en toute autonomie à partir de la documentation accessible en ligne via le CRAN ou les communautés d'utilisateurs. L'environnement RStudio sera mis en œuvre durant le cours.

Prérequis :

Aucun

Plan du cours :

1) Introduction
   - Un peu d'histoire
   - Contexte
   - Les bonnes pratiques

2) Démarrage rapide
   - Installation
   - Choix d'un environnement de développement : R Studio
   - Espace de travail
   - Aide en ligne

3) Les types de données
   - Tout type est objet
   - Types de base (Nombre, Chaîne de caractères, Booléen)
   - Autres types : Vecteur (vector), Facteur (factor), Tableau (array), Liste (list), Matrice (matrix), Data frame (data.frame), Formule (formula), Package

4) Manipulation des objets
   - Import / export de données
   - Constructeurs
   - Sélection d'éléments
   - Utilitaires

5) Script
   - Fonctions (créations, fonctions natives, fonctions vectorielles de type apply)
   - Exemple de création et d'utilisation de fonction sur la factorielle, les MCO…
   - Structures de contrôle (exécution conditionnelle, Boucles)

6) Graphiques avec ggplot2
   - Principe de superposition de couches
   - Grammaire graphique et éléments de structure
   - Paramètres esthétiques, géométriques
   - Fonctions statistiques et d'échelle
   - Positions, facettes et coordonnées
   - Exemples : Histogramme, Boxplot, Courbe
   - Exemples 3D (hors ggplot)

7) Introduction au tidyverse : manipulation de données avec dplyr
   - Opérateur pipe %>% pour enchaîner les opérations
   - Fonctions de traitement et de synthèse

8) Manipulations avancées avec data.table
   - Syntaxe des crochets pour sélection d'observations, de variables, synthèse de résultat (grouper), ajout, modification, suppression de variables
   - Empiler les crochets pour enchaîner les opérations

9) Traitement des expressions régulières (chaines de caractères)
   - Concaténer, Convertir, Découper
   - Extraire des sous-chaînes
   - Détecter, extraire, remplacer des motifs
   - Insérer une variable dans une chaîne

10) Traitement des dates avec lubridate
   - Manipulation de dates
   - Opérations sur les dates

11) Reporting dynamique avec R Markdown
   - Structure du code
   - Exemples de sorties pdf, html

Bibliographie :

- The R Ecosystem : Chris van Hasselt, 2016
- R, Bonnes pratiques, Christophe Genolini
- An Introduction to R, R Foundation for Statistical Computing, Venables, W. N., D. M. Smith et R Core Team. 2013,
- R for data science, Hadley Wickham, Garrett Grolemund, 2016
- Ggplot2 : elegant graphics for data analysis, Hadley Wickham, 2010

INTRODUCTION AU DATA MINING

Erwann Bargain (Celtis Conseil/ UPEC)
Durée : 24 H
Parcours : DA/DS

Objectif du cours :

L'objectif du cours est de donner aux élèves une vision d'ensemble du Data Mining, à la fois en terme technique et opérationnel.
Les sujets de gestion de projets Data Mining et de qualité des données sont donc abordés dans un premier temps, suivis d'une revue de différentes techniques de Machine Learning avec des applications sur des données classiques et textuelles.

Prérequis :

Notion de Python

Plan du cours :

1) Introduction au Data Mining
- Définitions, domaines d'application, méthodes, logiciels, potentiel et limites

2) Gestion de projets Data Mining
- Présentation de la méthodologie CRISP : objectifs métiers, qualité des données, enrichissement de données, éthique, déploiement et suivi des modèles

3) Présentation de la librairie Pandas pour la manipulation des données

4) Présentation de techniques de Machine Learning
- Présentation des K plus proches voisins, du classifieur naïf Bayésien et de la régression logistique (ridge et lasso)
- Présentation des arbres de décisions, Bagging, Boosting et Random Forest
- Présentation succincte des réseaux de neurones

5) Application des méthodes sur des données textuelles
- Rapide introduction au Text Mining (NLP) : tokenization, stopwords, lemmatisation
- Prise en main des librairies scikit-learn et xgboost
- Mise en œuvre des techniques précédentes de Machine Learning pour classer un texte dans l'une des catégories attendues (détection de la langue, un produit vers sa catégorie)

4) Compléments scikit-learn
- Pipeline, GridSearch, Transformers personnalisés, stockage, …

4) Application sur une compétition ouverte sur le site challengedata.ens.fr ou Kaggle

Bibliographie :

- Stéphane Tuffery, Data Mining et statistique décisionnelle, l'intelligence des données, éditions Technip (2010).
- Stéphane Tuffery, cours en ligne (2012), disponible sur http://data.mining.free.fr/
- Ricco Rakotomalala : cours : http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html
- Ricco Rakotomalala : Blog pour les tutoriaux sur R, Python et la Data Science http://tutoriels-data-mining.blogspot.fr/
- Aurélien Géron, Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd edition, O'Reilly (2019)
- Andreas C. Müller, Sarah Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, Editions O'Reilly (2016)
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (2009)

- Livre téléchargeable sur : http://www-stat.stanford.edu/~tibs/ElemStatLearn/
- Lebart L., Morineau A., Piron M., Statistique exploratoire multidimensionnelle (1995)
- Livre téléchargeable sur : http://horizon.documentation.ird.fr/exl-doc/pleins_textes/divers11-10/010007837.pdf

SAS AVANCE

Erwann Bargain (Celtis Conseil/ UPEC)
Durée : 24 H
Parcours : DA/DS

Objectif du cours :

A l'issue de ce cours, les étudiants devront acquérir une aisance sur le logiciel SAS, en matière de manipulation de données, de macro langage et de reporting, qui leur permettra d'être opérationnels sur des domaines variés :

- Informatique décisionnelle : alimentation d'un datawarehouse, automatisation de reporting (graphiques ou tables)

- Etudes statistiques ou économétriques (manipulation de données + industrialisation et mise en production des résultats de l'étude)

- Travail sur la qualité des données

De manière plus précise, les objectifs techniques suivants devront être atteints :

- Montrer une grande aisance et une capacité à tirer parti de plusieurs fonctions du logiciel pour résoudre des problèmes complexes de manipulations de données.

- Etre capable d'automatiser des programmes répétitifs en utilisant le macro langage.

- Etre capable d'optimiser les programmes, en termes de performances, mais aussi en termes de clarté, et de facilité de maintenance.

- Etre capable de produire des rapports sous forme tabulaire ou graphique.

Prérequis :

Avoir suivi un cours d'introduction à SAS

Plan du cours :

1) Positionnement de SAS sur le marché de l'analytics

2) Fonctionnement détaillé de l'étape data

2) Procédures les plus usuelles et options avancées

3) Création de macros

4) ODS et Proc SQL

Bibliographie :

- Sébastien Ringuedé, Introduction au décisionnel : du data management au reporting, 4ième édition, avril 2019, Edition Eyrolles
- Site avec des milliers de pdf du niveau débutant à avancer : https://www.lexjansen.com/

CERTIFICATION SAS

Erwann Bargain (Celtis Conseil/UPEC)
Durée : 24 H
Parcours : DA/DS

Objectif du cours :

Ce cours a pour objectif de préparer les étudiants à la Certification SAS Base qui couvre le programme suivant :

- importation et exportation de tables de données
- manipulation et transformation des données
- combinaison de bases de données SAS
- création de rapports
- identification et correction des erreurs dans les données
- identification et correction d'erreurs de syntaxe et de logique de programmation

Pour obtenir cette certification, les étudiants doivent réaliser un score d'au moins 725 points/ 1000 sur des questions en anglais qui se présentent sous la forme d'un QCM et de questions nécessitant de coder en langage SAS.

Le support sera en anglais de manière à familiariser les étudiants avec les termes techniques et afin qu'ils puissent par conséquent lire et comprendre rapidement les questions.
Des applications seront également effectuées pendant chaque cours dans le but de comprendre le fonctionnement de SAS et de déjouer les pièges classiques.
A la fin de chaque partie, un QCM sera proposé aux étudiants pour les aider à juger leur niveau.

Prérequis :

- Cours d'introduction à SAS

Plan du cours :

1) Présentation de la nouvelle certification
2) Basic Concepts
3) Accessing Your Data
4) Creating SAS Data Sets
5) Identifying and Correcting SAS Language Errors
6) Creating Reports
7) Understanding DATA Step Processing
8) BY-Group Processing
9) Creating and Managing Variables
10) Combining SAS Data Sets
11) Processing Data with DO Loops
12) SAS Formats and Informats
13) SAS Date, Time, and Datetime Values
14) Using Functions to Manipulate Data
15) Producing Descriptive Statistics
16) Creating Output

Bibliographie :

- SAS Certified Specialist Prep Guide: Base Programming Using SAS® 9.4, février 2019, SAS Institute Inc Edition

- Sébastien Ringuedé, Introduction au décisionnel : du data management au reporting, 4ième édition, avril 2019, Edition

TEXT MINING

Sofia Belcadi (EasyVista)
Durée : 18 H
Parcours : DA

Objectif du cours :

L'objectif du cours est d'appréhender les notions élémentaires du Text Mining. A travers la théorie et la pratique, nous parcourrons toutes les étapes nécessaires à la mise en œuvre d'un projet de NLP (Natural Language Processing). Les étudiants seront en mesure à l'issue du cours d'appliquer un algorithme de Machine Learning sur du texte. Ils pourront également apprendre à utiliser une des API les plus répandues du marché.

Prérequis :

- Python (importer une librairie, exécuter des fonctions)
- Manipulation des chaines de caractères en python

Plan du cours :

1) Introduction
   - Intérêt et cas d'usage du NLP.
   - Exemples d'applications courantes.
   - Différence entre Deep Learning et Machine Learning,
   - Présentation du programme.

2) Manipulation des chaines de caractères
   - Rappels de fonctions de la librairie str de python
   - manipulation de textes.

3) Expressions régulières ou Regex
   - Trouver automatiquement des entités définies dans un texte : adresse email, téléphone, numéro de facture.
   - Rechercher une Regex dans un corpus,
   - Trouver la première occurrence d'un terme, le nombre d'occurrences, etc

4) Structuration des données
   - Vocabulaire du NLP
   - Transformation d'un texte en matrice
   - Création et Manipulation des tokens
   - Fréquence et occurrence des mots dans un corpus
   - Savoir identifier les mots et n-grammes caractéristiques d'un corpus

5) Réduction de dimensionnalité
   - Introduction au stopwords
   - Stemming
   - Lemmatisation
   - Calcul de fréquences
   - Correction de fautes d'orthographe
   - Indicateurs TF IDF, TF et TF binaire
   - Transformation d'un texte en matrice exploitable par un algorithme de Machine Learning

6) Application d'un algorithme de Machine Learning
- Introduction à la CAH
- Diagnostics des corpus de texte 20NewsGroup et Reuters
- Mise en œuvre d'une CAH sur un corpus homogène
- Mise en œuvre d'une CAH sur un corpus hétérogène
- Comparatif

7) Utilisation d'une API de traitement du texte
- Introduction aux outils du marché (Google, Amazon, etc)
- Utilisation de l'API Google sentiment pour détecter le sentiment d'un texte.
- Avantages et inconvénients des API

Bibliographie :

- Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning (Benjamin Bengfort, Tony Ojeda)

- Natural Language Processing with Python (Steven Bird, Ewan Klein)

COINTEGRATION ET MODELES VAR

Vincent Bouvatier (Université Paris Est Créteil)
Durée : 18 H
Parcours : DA

Objectif du cours :

Le cours propose aux étudiants de développer leurs compétences et connaissances en économétrie des séries temporelles. Les modèles multivariés seront abordés et une attention particulière sera portée aux traitement des séries non stationnaires.
Le cours sera orienté dans une perspective d'économétrie appliquée. Le logiciel SAS sera utilisé pour illustrer la mise en œuvre des différents modèles et tests abordés.

Prérequis :

Prérequis en économétrie :
- modèle de régression classique
- modèles ARMA et ARIMA
- tests de racine unitaire et de stationnarité

Prérequis sous SAS :
- étapes data
- proc reg
- proc arima

Plan du cours :

1) Chapitre introductif
- tests de racine unitaire
- tests de stationnarité
- tests de cointégration

2) Modèles VAR
- spécification du modèle
- éléments d'analyse structurelle
- estimation du modèle

3) Cointégration
- approche univariée : Engle-Granger
- approche multivariée : VECM

Bibliographie :

- H. Lütkepohl, New Introduction to Multiple Time Series Analysis, 2007, Springer.
- J. Hamilton, Time Series Analysis, 1994, Princeton University Press.
- S. Lardic et V. Mignon, Econométrie des Séries Temporelles Macroéconomiques et Financières, - 2002, Economica.

LOGICIEL DE CREATION ET DE GESTION DE BASES DE DONNEES

(Logiciel SGBD MySQL)

Olivier Bretel (Disney Corp.)
Durée : 24 H
Parcours : DA/DS

Objectif du cours :

Les étudiants apprendront à installer un serveur MySQL et ses outils graphiques, le configurer, l'administrer pour les fonctions basiques, créer des bases de données, manipuler et interroger les données par SQL.
Le cours mettra principalement l'accent sur la mise en pratique du langage SQL.

Prérequis :

Avoir eu une introduction à SQL.

Plan du cours :

1. Présentation des bases de données du marché :
   - Grand acteurs
   - Utilisation dans les entreprises
   - MYSQL

2. Fonctionnement d'un SGBDR

- Architecture physique
- Index
- Grands principes

3. Mise en Pratique : Installation du moteur MySQL et des outils
   - Téléchargement et installation
   - Configuration
   - Découverte des outils

4. Mise en Pratique : langage SQL ANSI (SQL III) → Requêtes : Select simples

5. Mise en Pratique : langage SQL ANSI (SQL III)
   → Requêtes : Select avec jointures, agrégations

6. Mise en Pratique : Exemple d'interaction avec d'autres outils : Excel

7. : Mise en Pratique : TP1
   - Chargement d'un fichier de données dans une table de travail
   - Stockage des données dans une structure en étoile
   - Manipulation simples : filtres, regroupements, fonctions SQL

8. Mise en Pratique : TP2
   - Agrégats
   - Requêtes imbriquées
   - Recodage de variables
   - Requêtes Analytiques

Bibliographie :

- Soutou, Christian, Apprendre SQL avec MYSQL : avec 40 exercices corrigés, Eyrolles, 2006.
- Dinimant, Antoine, MYSQL 5, Micro application. DL 2006 - Le guide complet.

ECONOMETRIE DES ETUDES D'IMPACT

Thibault BRODATY (Université Paris Est Créteil)

Durée : 18 H
Parcours : DA

Objectif du cours :

L'objectif de ce cours est de présenter un ensemble de méthodes économétriques qui permettent de réaliser des études d'impact. Ainsi, ce cours a pour but plus général de présenter des méthodes permettant d'évaluer l'impact d'un « événement » sur un « critère de résultat ». Les applications possibles vont donc de l'évaluation de l'effet de la baisse des charges sur l'emploi à l'évaluation d'une action de formation dans une entreprise, en passant par exemple par l'évaluation de l'effet d'une campagne de publicité sur les ventes d'une entreprise. Pour chaque méthode, on présentera dans un premier temps les fondements théoriques, puis un article scientifique d'application.

Prérequis :

Econométrie linéaire.

Plan du cours :

1) Le modèle causal de Rubin

2) Différences de différences

3) Contrôle synthétique

4) Variables instrumentales avancées

5) Regression discontinuity design

6) Matching

Bibliographie :

- Angrist, Pischke, 2009, Mostly Harmless Econometrics – An Empiricist Companion.
- Wooldridge, 2010, Econometric Analysis of Cross Section and Panel Data.

ECONOMETRIE DES VARIABLES QUALITATIVES

Thibault BRODATY (Université Paris Est Créteil)

Durée : 24 H
Parcours : DA

Objectif du cours :

L'objectif de ce cours est de présenter le spectre des méthodes économétriques permettant d'analyser des données qualitatives. Ces données sont en effet très courantes et il est donc important d'être en mesure de les traiter. Elles peuvent être binaires (à deux modalités) comme la décision d'un consommateur d'acheter ou non un produit. Elles peuvent être à au moins trois modalités non ordonnées comme la décision d'un consommateur de choisir entre au moins trois marques. Elles peuvent être à au moins trois modalités et ordonnées comme la satisfaction d'un consommateur ou la santé perçue d'un individu. Elles peuvent enfin traduire un décompte (données de comptage) comme le nombre d'appels au service client d'une entreprise ou le nombre de visites d'un patient chez son médecin. Ces méthodes seront également appliquées pour traiter les problèmes de censure, de troncature et de sélection des variables continues. Toutes ces méthodes se basent sur le maximum de vraisemblance. Au sein de chaque chapitre, les données de panel seront présentées après les méthodes concernant les données en coupe. La mise en pratique de ces méthodes sera effectuée sous le logiciel SAS sur des données issues de l'enquête « Santé et

Itinéraire Professionnel ». Un accent particulier sera mis sur la procédure NLMIXED qui permet de programmer sa propre vraisemblance, en facilitant la prise en compte d'effets aléatoires gaussiens.

Prérequis :

Maximum de vraisemblance, économétrie linéaire.

Plan du cours :

1) Rappels sur la méthode du maximum de vraisemblance

2) Modèles qualitatifs binaires

3) Modèles multinomiaux ordonnés

4) Modèles multinomiaux non ordonnés

5) Données de comptage

6) Censure, troncature, sélection

Bibliographie :

Wooldridge, 2010, Econometric Analysis of Cross Section and Panel Data.

INTRODUCTION A SAS

Arnaud DAUCHY
Durée : 24 H
Parcours : DA/DS

Objectif du cours :

L'objectif est de proposer une initiation au logiciel SAS : il s'agit de permettre une prise en main par la pratique.
Le déroulement du cours est axé sur une présentation des différents concepts : Importer des données, les transformer, les traiter et générer des sorties.
Chaque notion sera illustrée au travers d'exercices pratiques sur SAS Windows.

Prérequis :

Notions de statistique descriptive et inférentielle
IL est important d'avoir installé le logiciel sur son PC avant le cours

Plan du cours :

1) Prise en main
    - L'Environnement
    - Librairies, tables, programmes
    - Charger des données
    - Mode liste / colonne / format
    - PROC IMPORT
2) Manipulation de données
    - Etape DATA
    - DROP, KEEP, RENAME
    - Instructions SET et MERGE
    - Gestion des variables
    - Variables numériques et caractères
    - Attributs : longueur, label, formats
    - Gestion des observations
    - La clause WHERE
    - Fonctions et opérateurs
    - Boucles
    - Instructions IF - THEN, SELECT
3) Traitement des données
    - Les procédures
    - PROC SORT, PRINT, CONTENTS
    - PROC FORMAT
    - PROC FREQ, MEANS, TRANSPOSE
4) Reporting
    - L'ODS
    - PROC REPORT
    - PROC EXPORT
    - Les Graphes

Webographie :

http://www.sasreference.com/
https://od-datamining.com
https://lexjansen.com

SAS POUR LE BIG DATA

Grégoire de Lassence (SAS INSTITUTE)
Durée : 24H
Parcours : DS

Objectif du cours :

Conceptualiser l'industrialisation de l'analytique sur des projets de Big Data sur SAS Viya

Prérequis :

Notions élémentaires de programmation SAS
Notion de base des algorithmes de Machine Learning

Plan du cours :

Introduction générale au Big Data - Rappels historiques – évolution de la plateforme SAS

Panorama de l'analytique afin de faire le lien interdisciplinaire : présentation des différences et complémentarités entre prévision et prédiction, entre Statistique et Data Mining, entre statistiques textuelles et Text Mining, entre Data Mining et Machine Learning ; tout cela avec l'économétrie, la recherche opérationnelle, l'analyse des réseaux sociaux, le Deep Learning, etc.

Cycle d'un projet analytique. Pourquoi la majorité des projets de Big Data sont des échecs ? C'est-à-dire, qu'elles sont les bonnes pratiques pour que cela fonctionne.

Pris en main de SAS Visual Analytics pour la Data Viz.
Reporting et présentation de modélisation de Machine Learning avec SAS Visual Analytics – Communication sur le RoI des modèles.

À la suite de cette introduction, la partie principale porte sur la modélisation de Machine Learning en Pipeline dans SAS Model Studio. Préparation des données – stratégie d'optimisation de la recherche des hyperparamètres des modèles de Machine Learning. Comparaison et mises en production automatique de modèles.

Pour finir, introduction la programmation spécifique à SAS Viya. Integration de code open-sources.

Bibliographie :

- Tamming the Big Data Tidal Wave – Bill Francks – The Wiley and SAS Business Series
- Analytics at Work - Thomas H. Davenport – The Wiley and SAS Business Series
- Business Analytics for Managers - Thomas H. Davenport – The Wiley and SAS Business Series
- Competing on Analytics - Thomas H. Davenport – The Wiley and SAS Business Series

ANGLAIS - ENGLISH

Anne-Pierre DE PEYRONNET (Consultant)
Durée : 24 H
Parcours : DA/DS

Objectif du cours :

-Students will prepare for their final presentation: doing a 20 minute presentation and writing a report.
-In doing so, students will improve their expression ability. In class, students will discuss industry topics, and articulate their thoughts. Doing so, students will study:

-The language of the industry.

-Key grammar points.

Prérequis :

-Ability to read with ease a Wall Street Journal article.
-Ability to listen to a 5-minute radio show and discuss the topics raised during the show.
-Create an account at: https://be-in-charge.fr

Plan du cours :

1) “Spotting Global Risks Through Global Collaboration”
   -Prasad Ananthakrishnan, Head of Strategy and Planning, Monetary and Capital Markets department. International Monetary Fund
   -Listening comprehension: be ready to answer questions.

2) Slide design principles
   -In teams of 4 – 5, students will prepare and present a 5-minute PowerPoint-presentation answering a question relating to the IMF interview of Prasad Ananthakrishnan.

3) "Behind an Effort to Fact-Check Live News With Speed and Accuracy"
   -By Laine Higgins - The Wall Street Journal, Nov. 23, 2018 11:00 a.m.
   -Reading comprehension: be ready to answer questions.

4) Aplying Slide Design Principles
   -In teams of 4 – 5, students will prepare and present a 5-minute PowerPoint-presentation answering a question from Laine Higgins’ WSJ article.

5) "Improving taxi fleet efficiency in Singapore"
   -By Associate Professor Cheng Shih-Fen, Singapore Management University
   -Listening comprehension: be ready to discuss the podcast.

6) Slide design principles
   -In teams of 4 – 5, students will prepare and present a 5-minute PowerPoint-presentation answering a question relating to the Singapore Management University podcast.

7) "Making Sound Decisions Through Data Analytics"
   -By Associate Professor Manoj Thulasidas, Singapore Management University. Listening comprehension: be ready to answer questions.

8) FINAL EXAM
   -Bring your computers!

This will be a LISTENING COMPREHENSION online quiz. The quiz will include both: an MCQ and an essay.

Bibliographie :

- The Wall Street Journal: http://wsj.com
- National Public Radio podcasts: https://www.npr.org
- The International Monetary Fund podcasts: https://www.imf.org/en/News/Podcasts/All-Podcasts/
- The Singapore Management University: https://soundcloud.com/sgsmu

MODELES DE DUREE

Emmanuel DUGUET (UPEC)
Durée : 18 H
Parcours : DA

Objectif du cours :

L'objectif de ce cours est de familiariser les étudiants avec l'économétrie spécifique aux variables positives. Nous examinerons d'abord les problèmes de censure des données avant de passer à l'estimation de modèles économétriques standard (hasards proportionnels, durée accélérée) et moins standard (avec survivants) sur données de durées censurées. Tous les chapitres font l'objet d'application sur SAS ou sur R.

Prérequis :

Très bonne maitrise de l'économétrie linéaire. Bonnes connaissances de la théorie des probabilités. Connaissances de base en SAS et de R.

Plan du cours :

1 - Introduction

Les variables de durée, présentation des concepts pertinents. Définition des quantités à calculer lors d'une étude appliquée. Les modélisations disponibles avec variables explicatives

2 - Statistiques descriptives sur données censurées

Statistiques descriptives sur variables censurées et tests de comparaison. Estimation paramétrique, avec et sans survivants (i.e., des individus qui ne sortent jamais de l'état étudié).

3 - Modèles à durée accélérée

Modélisation en logarithmes. Estimation paramétrique (avec proc lifereg sous SAS et le package « survival » de R).

4 – Modèles à hasards proportionnels

Estimation semi-paramétrique du modèle de Cox (avec proc phreg sous SAS et le package « eha » de R)

5- Modèles avec survivants

Ces modèles ne sont ni à durée accélérée ni à hasard proportionnels. Estimation paramétrique d'un modèle Logit/Weibull (directement par le maximum de vraisemblance)

Bibliographie :

- Cox D.R., Oakes D., 1984, Analysis of survival data. Monographs on Statistics and Applied Probability n°21, Chapman & Hall/CRC. ISBN 041224490X.
- Duguet E., 2018, Econométrie appliquée aux variables de durée, coll. « Economie et Statistique Avancée ». *Economica*. ISBN 978-2-7178-7045-9.

ASPECTS JURIDIQUES & PROTECTION DES DONNEES

Philippe Guilbert (WAGRAM Consulting)
Durée : 12 H
Parcours : DA/DS

Objectif du cours :

Sensibiliser à la déontologie et réglementation de protection des données personnelles et donner les clés pour agir en entreprise en tant que professionnel de la data.

Prérequis :

Aucun, initiation

Plan du cours :

1) Jour 1 Enjeux, déontologie
   - Enjeux de la data
   - Enjeux de la protection
   - Déontologie, qualité, norme
   - Code déontologique
   - Réglementation

2) Jour 2 Réglementation, pratique
   - Réglementation
   - La data en entreprise
   - Sécurité
   - Mise en conformité
   - A votre arrivée

Bibliographie :

- Guide Sécurité des données personnelles, CNIL 2018 : https://www.cnil.fr/sites/default/files/atoms/files/cnil_guide_securite_personnelle.pdf

- Charte Ethique et big data, Alliance Big Data 2013 : http://wiki.ethique-big-data.org/chartes/CharteEthiqueBigDatav8.pdf

- Ethique & Numérique, Syntec Numérique / Cigref 2018 : https://syntec-numerique.fr/sites/default/files/Brochure_Cigref_-_Syntec_PDF_0.pdf

ECONOMETRIE DES DONNEES DE PANEL

Guillaume Horny (Banque de France)
Durée : 18 H
Parcours : DA

Objectif du cours :

Ce cours a pour objet de fournir aux étudiants les bases théoriques de l'économétrie des données de panel, de leur présenter des applications et de les accompagner dans leurs réalisations sous SAS et R. A l'issue du cours, les étudiants sauront pourquoi le suivi longitudinal des individus fourni de l'information utile à l'économètre, de quelle manière cette information peut affecter les résultats et comment se prémunir d'éventuels effets indésirables. Les étudiants sauront également mettre en forme des données pour constituer un panel, comment utiliser les estimateurs et tests standards de la littérature, ainsi qu'interpréter leurs résultats

Prérequis :

Econométrie linéaire.

Les diaporamas du cours dispensé l'année précédente sont disponibles à : http://horny.economics.free.fr

Plan du cours :

Introduction

Chapitre 1 : Modèle empilé

Chapitre 2 : Modèle à effet fixe

Chapitre 3 : Effet fixe et prévisions

Chapitre 4 : Modèle à effet aléatoire

Chapitre 5 : Effet fixe ou aléatoire ?

Conclusion

Bibliographie :

- P. Sevestre (2002), Econométrie des données de panel, Dunod.
- Jeffrey Wooldridge (2010), Econometric Analysis of Cross Section and Panel Data, MIT Press.
- Colin Cameron et Pravin Trivedi (2005), Microeconometrics - Methods and Applications, Cambridge University Press.

Machine/Deep learning
Sophie LARUELLE (LAMA-UPEC)
Durée : 30 H
Parcours : DS/DA

Objectif du cours : Ce cours présente un ensemble de techniques d'apprentissage automatique ainsi que leur implémentation en Python à l'aide des librairies Scikit-Learn et TensorFlow pour l'apprentissage profond. Le but est donc de se familiariser avec ces outils et de comprendre quand et comment les utiliser et les implémenter.

Prérequis : programmation Python avec les librairies NumPy, Scipy, Matplotlib et Pandas ; statistique descriptives, analyse de données, séries temporelles.

Plan du cours :

1) Introduction à l'apprentissage automatique
   - définition et typologies
   - un exemple détaillé d'application

2) Modèles linéaires et extensions
   - régression linéaire, polynomiale, descente de gradient
   - modèles linéaires régularisés (Ridge, Lasso, Elastic Net)
   - régression logistique et softmax

3) Machines à vecteurs de support (SVM)
   - classification linéaire et non linéaire

- régression

4) Arbres de décision
- visualisation, entraînement, prédiction, classification
- algorithme CART et régularisation

5) Apprentissage d'ensembles et forêts aléatoires
- bagging et pasting
- boosting et stacking

6) Apprentissage non supervisé
- réduction de dimension : ACP, ACP à noyau, LLE
- partitionnement : Kmeans et extensions, DBSCAN
- mélanges gaussiens

7) Introduction à l'apprentissage profond
- perceptron multi-couche (PMC) et rétropropagation
- régression et classification avec un PMC
- implémentation à l'aide de TensorFlow : initialisation, fonctions d'activation, optimiseurs, régularisation

8) Réseaux de neurones convolutifs (RNC)
- couches convolutives et couches de pooling
- architectures RNC
- implémentation en Keras

9) Réseaux de neurones récurrents (RNR)
- neurones et couches récurrents, cellule mémoire
- entraînement et application aux prévisions de séries temporelles

Bibliographie :

- C. Albon, *Machine learning with Python cookbook: practical solutions from preprocessing to deep learning*, O'Reilly, 2018.
- A. Géron. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*, O'Reilly, 2e édition, 2019.
- S. Raschka, V. Mirjalili. *Python Machine Learning: machine learning and deep learning with Python, scikit-learn, and tensorflow 2*, Packt Birmingham-Mumbai, 3e édition, 2019

WEB MINING ET WEB ANALYTICS

François Legendre – Université Paris-Est Créteil

Durée : 24 H
Parcours : DS

Objectif du cours :

Le *Web mining* est le *data mining* appliqué aux traces laissées par les visiteurs d'un site web, en général à l'aide de méthodes de *machine learning* non supervisées. Le W*eb analytics* est moins ambitieux : il a pour objet principal de mesurer le trafic d'un site web.

L'objectif du cours est double : en premier lieu, comprendre de quelles manières on peut obtenir de l'information sur les internautes ; en second lieu, mettre en œuvre les outils standards de *Web mining* et de *Web analytics*.

Prérequis :

Niveau intermédiaire en Sas et en Python.

Plan du cours :

Le cours débute par une demi-journée consacrée à l'Internet : commutation de paquets, couche réseau (protocole IP), protocoles TCP et UDP, protocole HTTP, architecture client/serveur, HTML et JavaScript.

Ensuite, chaque étudiant aura à développer un mini site web pour, dans un premier temps, utiliser les outils qui permettent d'analyser, en Sas et en Python, les logs du serveur HTTP. Le site sera développé en local puis déployé sur l'internet au moyen d'un *virtual private server* (VPS).

Les outils qui reposent sur l'utilisation d'un serveur tiers, dont notamment la solution proposée par Google, seront ensuite présentés et discutés. Chaque étudiant aura à mettre en place une solution en Python basée sur l'utilisation de *cookies* pour son mini site web.

Enfin, un aperçu des méthodes de *Web mining* sera présenté ; les applications seront réalisées en utilisant les bibliothèques spécialisées disponibles sur le *Python Package Index*.

Bibliographie :

- Brian Clifton, 2015, *Successful Analytics: Gain Business Insights by Managing Google Analytics*, Advanced Web Metrics Ltd.

MARCHES ET PRODUITS FINANCIERS

Daniel Szpiro (UPEC)
Durée : 24 H
Parcours : DA/ DS

Objectif du cours :

L'objectif de ce cours est de présenter les produits financiers et la façon dont les institutions financières les utilisent et doivent récolter et utiliser l'information pour une gestion qui répond aux normes réglementaires étendues dans ce domaine. À cette fin, les différents risques sont définis tels qu'ils sont décomposés par la réglementation du Comité de Bâle sur la Supervision Bancaire, ce qui permet de présenter dans un deuxième temps les concepts et les données qui sont utilisés dans

les exigences minimal de capital. Les grandes lignes du fonctionnement des marchés financiers sont ensuite exposées, puis les spécificités des produits dérivés, contrats à terme ou options. Ce cours offre un large panorama de ce qui est à savoir sur le système financier lorsque l'on est amené à travailler dans une banque ou une assurance.

Prérequis :

Statistiques simples (espérance, variance, covariance, quantiles), économétrie, dérivation mathématique. Aucun prérequis n'est demandé en comptabilité ou en finance.

Plan du cours :

1) Les risques financiers
   - le risque de liquidité
   - le risque de contrepartie
   - le risque de marché
   - le risque opérationnel

2) La réglementation bancaire
   - Définition de la banque, les spécificités systémiques, la séparation des activités
   - Les normes comptables IFRS
   - La réglementation prudentielle de Bâle, les trois piliers
   - Les exigences en capital, méthode standard ou modèles internes, Value at Risk, Espérance conditionnelle de pertes.
   - Les ratios de liquidité

3) Les marchés
   - Marché obligataire et marché des actions
   - Les chambres de compensation et le système de règlement-livraison
   - Produits dérivés, garanties et appels de marge

4) Les contrats à terme
   - Définitions, exemples standards ou exotiques
   - Les trois usages du produit
   - Les déterminants du prix

5) Les options et les produits structurés

- Définitions et usages
- Les déterminants du prix
- Les swaps et autres produits structurés

Bibliographie :

- Hull J., 2013, *Gestion des risques & institutions financières*, Pearson Education, 598 p.
- Szpiro D., 1997, *Introduction à la Finance de Marché*, Economica, 148 p.
- Szpiro D., 2021, *Produits financiers et gestion de portefeuille,* Ellipses, 500 p, à paraître.

TECHNOLOGIE BIG DATA

Kevin TRAN-DAI (WIDE)
Durée : 24 H
Parcours : DS

Objectif du cours :

Présentation et manipulation des technologies BigData construites autour de la solution HADOOP.

- histoire et principe du calcul distribué (slides)
- commandes shell linux / shell HADOOP
- gestion de données structurées et Objet avec Hive (syntaxe HQL)
- présentation et utilisation de la base NOSQL HBASE
- présentation et utilisation du moteur SPARK avec python et HQL
- Cas Pratiques :
    - Utilisation de données Google Map avec Hive
    - modélisation d'un random Forrest avec SPARK
    - Régression linéaire simple en paralléle avec SPARK

Prérequis :

Connaissance du shell linux, du language python et du SQL

Plan du cours :

1) Concepts Hadoop (0.25J)
    - histoire / map Reduce
    - architecture Hadoop / presentation de la machine virtuelle
2) Atelier commandes SHELL (0.25J)
    - lancement et gestion des outils hadoop
    - manipulation des fichiers sous hadoop
3) Atelier Hive sur données structurées (0.5J)
4) Présentation HBASE et commandes Hbase (0.5J)

- Shell Hbase
- API REST

5) Présentation et manipulation PYPSPARK (2.5J)
- Pratique SparkSQL : régression linéaire simple sur données de séquences
- Pratique SparkMLLIB : classification avec une Random Forest

Bibliographie :

- Hadoop the definitive guide, Tom White, O'REILLY
- Learning Spark, Holden Karau, O'REILLY

TEXT MINING ET NLP

Arnold Vialfont (Université Paris Est Créteil - ERUDITE)
Durée : 24 H
Parcours : DS

Objectif du cours :

Ce cours présente les opérations classiques du text mining et du traitement automatique du langage naturel (NLP en anglais) et inclus la méthodologie de création d'un corpus à partir de fichiers txt et/ou html. L'exploitation d'un corpus est réalisée de façon intégrée afin de présenter les tâches habituelles réalisées sur les textes : classification, clustering, topic modeling et analyse de sentiments.

Nous utiliserons principalement la librairie Spacy sous Python qui est assez simple d'utilisation et plus rapide que la librairie « historique » NLTK. Nous intégrerons l'ensemble des étapes de traitement des textes dans des classes que nous construirons et les placerons ensuite dans des Pipelines de la librairie Scikit-Learn en vue de comparer facilement les différents algorithmes utilisés.

Prérequis :

- Connaissance du langage de programmation Python (notions des méthodes et classes).
- Avoir intégré les notions du cours de datamining (dont l'utilisation de la libraire Scikit-Learn).
- Connaître les principaux algorithmes de Machine Learning pour la deuxième partie du cours.

Plan du cours :

1) Traitement de textes et constitution d'un corpus
   - Tokenisation en fonction de la langue du texte
   - Normalisation (correction orthographique, stopwords, lemmatisation/stemmisation)
   - Vectorisation (encodage one-hot, vecteur de fréquence, tf-idf, doc2vec)
   - Extraction des entités nommées et phrases clés
   - Stockage du corpus traité (paragraphes, phrases, mots, fonctions grammaticales, IOB)

2) Exploitation d'un corpus de textes
   - Description du corpus : statistiques descriptives et visualisations sommaires

- Classification de textes labellisés (comparaison d'algorithmes et étude des erreurs)
- Clustering de textes et topic modeling (constitution et caractérisation des clusters)
- L'analyse de sentiments (applications des réseaux de neurones MLP et LSTM)

Bibliographie :

Livres :

- Bengfort B., Bilbro B. et Ojeda T. (2018), « Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning », Ed. O'Reilly.
- Lane H., Howard C. et Hapke H. (2019), « Natural Language Processing in Action: Understanding, analyzing, and generating text with Python », Ed. Manning.

- Russell M. et Klassen M. (2019), « Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Instagram, GitHub, and More », Ed. O'Reilly.
Liens utiles :
- Cours de R. Rakotomalala : http://eric.univ-lyon2.fr/~ricco/cours/cours_text_mining.html
- Github du livre « Mining Social Web » : https://github.com/mikhailklassen/Mining-the-Social-Web-3rd-Edition

- Documentation de Spacy : https://spacy.io/usage/spacy-101Ref1

PYTHON AVANCE : TRAITEMENT GEOSPATIAL ET INTRO AUX GRAPHES

Aric Wizenberg (Groupe Setec)
Durée : 12 H
Parcours : DS

Objectif du cours :

Dans ce cours de Python avancé, l'étudiant(e) apprendra à manier une structure de données omniprésente, mais généralement peu connues : les « graphes ». A ne pas confondre avec l'abréviation de « graphiques », il s'agit du nom que porte les structures de type « réseau », c'est-à dire des structures ou les points de données sont reliées les unes aux autres.

Réseaux sociaux (au sens moderne ou originel), réseaux informationnels, infrastructures de communication et de transport, réseaux économiques (commerce international, relations géopolitiques), les données pouvant être structurées en utilisant des graphes sont nombreuses.

Ce sujet, qui peut paraitre théorique au premier abord, a en réalité des applications concrètes dans de très nombreux domaines et permet d'analyser des problématiques très actuelles : propagation de fausses rumeurs, bulles informationnelles et communautés, marketing viral…

Dans une première partie, les fondements théoriques seront abordés, dans une deuxième partie, les outils de Python seront présentés, et enfin l'exemple des réseaux de transport servira à mettre en pratique ces savoirs et ces outils.

Prérequis :

De bonnes bases en Python (équivalent au contenu traité en cours de M1)
Cours de « Rappels de Python » du M2.

Plan du cours :

1) Introduction aux graphes : tour d'horizon et théorie
    - Les graphes : pourquoi et quand les utiliser
    - Bases théoriques et mathématiques
2) Outils pour le traitement, l'analyse et la visualisation de graphes
    - Le module NetworkX
    - Le programme Gephi
3) Application à un cas d'étude : les réseaux de transport

Bibliographie :

Graph Analysis and Visualization, David Jonker, Richard Brath, 2015, Wiley

Complex Network Analysis in Python, Dmitry Zinoviev, 2018, The Pragmatic Programmers

RAPPELS DE PYTHON : ANALYSE ET EXPLORATION DE DONNEES

Aric Wizenberg (Groupe Setec)
Durée : 12 H
Parcours : DA/DS

Objectif du cours :

Dans ce cours, l'étudiant(e) approfondira ses connaissances en Python selon 3 axes.
Le premier consistera à présenter la structuration d'un projet en Python. En effet, il est très différent de savoir coder du Python dans un script ou un notebook unique, et de savoir construire un projet plus complexe, reposant sur un ensemble de notebooks ou de scripts devant interagir entre eux.

Le second consistera à présenter tout un ensemble d'outils d'analyse de données avancés (dataviz, mise en forme de dataframes, widgets) qui faciliteront la pratique de cette activité préalable à toute activité en lien avec des données.

Le troisième consistera en une présentation des outils permettant de travailler avec des données géolocalisables ou géolocalisées, forme de données extrêmement courante qui permettra d'offrir à l'étudiant(e) de nouvelles possibilités en termes d'analyse, de traitement et de représentation des données.

Prérequis :

De bonnes bases en Python (équivalent au contenu traité en cours de M1).

Plan du cours :

1) Passer à l'étape supérieure : du simple Notebook/script au projet en Python
    - Bien structurer un projet Python, arborescence et organisation
    - Interactions entre Notebooks, échanges de données au sein d'un projet
    - Création d'un template pour les projets
2) Outils d'analyse de données
    - Data Visualisation en Python : vue d'ensemble des outils
    - Styling de DataFrame avec Pandas
    - Outils interactifs : les Widgets
3) Données géolocalisées et cartographie
    - Introduction aux données géolocalisées
    - Les modules de Python : GeoPandas et Shapely
    - Présentation rapide du logiciel QGIS

Bibliographie :

Learning IPython for Interactive Computing and Data Visualization, 2nd Edition, Cyrille Rossant, 2015, Packt

Python Geospatial Analysis Essentials, Erik Westra, 2015, Packt

INTRO AU WEBSCRAPING : EXTRACTION AUTOMATISEE APPLIQUEE AUX APIS

Aric Wizenberg (Groupe Setec)
Durée : 12 H
Parcours : DA/DS

Objectif du cours :

Le webscraping est un ensemble de méthodes et de pratiques permettant d'extraire, de manière automatisée tous types d'information depuis des sites web.

Dans ce cours, l'étudiant(e) apprendra comment communiquer avec un site web avec Python en utilisant le module Requests, puis il/elle se familiarisera avec les structures en arbre, omniprésentes sur Internet, et dans bien d'autres domaines et secteurs. Enfin il/elle apprendra à coder en pratique un scraper.

Dans le cadre de ce cours d'intro, seules les extractions au travers d'interfaces structurés (les APIs) seront abordées, l'extraction de données hors API sera abordée dans le cours de webscraping avancé.

A la fin de ce cours l'étudiant(e) sera en capacité de coder un scraper adapté à tous types d'API, lui permettant de récupérer et traiter facilement des données depuis Internet de manière automatisée.

Prérequis :

De bonnes bases en Python (équivalent au contenu traité en cours de M1)
Cours de « Rappels de Python » du M2.

Plan du cours :

1) Communiquer avec un site web
    - Principe de fonctionnement : requêtes, réponses, headers, cookies…
    - Bases du fonctionnement du module Requests
2) Structures de données de type arbre
    - Structures en arbres : principe de fonctionnement
    - Le module JSON
3) Le webscraping en pratique
    - Structure d'un scraper
    - Application à des exemples d'APIs

Bibliographie :

Practical Web Scraping for Data Science, Seppe vanden Broucke & Bart Baesens, 2018, Apress

WEBSCRAPING AVANCE : METHODES ET OUTILS POUR ALLER PLUS LOIN

Aric Wizenberg (Groupe Setec)
Durée : 12 H
Parcours : DS

Objectif du cours :

Suite au cours d'Intro au webscraping, l'étudiant aura acquis une compréhension des échanges simples entre un client (l'ordinateur de l'étudiant) et un server (le site web), et des moyens d'automatiser ces échanges en utilisant des interfaces proprement organisées par les concepteurs de site (les APIs).

Dans ce cours de webscraping avancé, l'étudiant(e) apprendra à extraire des données d'un site web même lorsque cela n'a pas été prévu par le concepteur du site. En effet, le nombre de sites web proposant une API est faible par rapport à l'ensemble des sites du web. Or, la règle est que lorsqu'une information (chiffre, texte, image…) apparait sur un site web, elle peut être, plus ou moins facilement, extraite de manière automatisée.

A la fin de ce cours l'étudiant(e) sera en capacité de coder un scraper adapté à tous types de sites web, lui permettant de récupérer facilement des données depuis Internet, même lorsque le site ne propose pas d'API.

Prérequis :

De bonnes bases en Python (équivalent au contenu traité en cours de M1)
Cours de « Rappels de Python » et cours d'« Intro au webscraping » du M2.

Plan du cours :

1) La console développeur du navigateur

- La console développeur (Firefox+Chrome) : présentation complète
- Comprendre en détail les échanges entre votre ordinateur et un site web

2) Structures de données de type arbre : le format Markup Language (ML)
- La structure Markup Language (HTML, XML, SVG, KML...)
- Le module BS4 (Beautiful Soup 4)

3) Le webscraping en pratique
- Structuration d'un programme de scraping complexe
- Application : exemples de scrapers

4) Outils de webscraping supplémentaires
- Le module Selenium
- Le framework Scrapy

Bibliographie :

- Web Scraping with Python, 2nd Edition, Ryan Mitchell, 2018, O'Reilly