Syllabus de la formation

Le **Certificat Toulouse Tech "Sciences des données et Big Data : outils et introduction"** est un module complémentaire au diplôme d'ingénieur qui s'adresse aux étudiants qui n'ont pas choisi le Big Data comme spécialité mais qui souhaitent pouvoir comprendre les enjeux liés aux données qu'ils seront susceptibles de rencontrer dans leur parcours professionnel.

DURÉE DE LA FORMATION : septembre à mars

VOLUME HORAIRE : 90h *(24 heures d'enseignements en e-learning, 36 heures de cours et travaux pratiques (TP), 20 heures de projets, 10 heures de conférences)*

PROGRAMME : statistiques, algorithmique et informatique

LES OBJECTIFS : cette formation offre la possibilité aux étudiants de :

- découvrir et maîtriser les aspects fondamentaux du Big Data,
- être en capacité de dialoguer avec des data scientists,
- découvrir les outils et méthodes applicables aux problèmes de données.

LE PUBLIC: la formation est ouverte à tous les étudiants en ingénierie des établissements de Toulouse Tech [1], non spécialisés en sciences des données, intégrant à partir de la rentrée scolaire un **niveau M1 (Bac+4) ou un niveau M2 (Bac+5)**[2].

LES PRÉREQUIS : les étudiants doivent avoir des connaissances de base en statistiques, probabilités, calcul différentiel et algèbre linéaire numérique. Il est également demandé des connaissances de base de l'environnement Linux et de la programmation (R, python), ainsi que la gestion de structures de données dans ces langages.

LES UTILITAIRES : pour pouvoir suivre la formation les étudiants doivent avoir un ordinateur portable, avec un browser internet.

L'ÉVALUATION : pour obtenir le certificat, les étudiants devront réussir les évaluations incluses dans la formation. A la fin de chaque module, est proposé un projet à réaliser en binôme ou en équipe, évalué avec des notes allant de E à A. La validation du certificat requiert au moins deux notes de C dans l'un des deux projets, et aucune note en dessous de D. L'assiduité est également prise en compte.

[1] ENAC, Icam - site de Toulouse, IMT Mines Albi, INSA Toulouse, ISAE- SUPAERO, Institut National Universitaire Champollion – ISIS, Toulouse INP - EI Purpan, ENIT, ENM, ENSAT, ENSEEIHT, ENSIACET, Université Toulouse III Paul Sabatier - Cursus Master Ingénierie et UPSSITECH

[2] Équivalent aux 4ème et 5ème années des écoles en 5 ans, ou aux 2ème et 3ème année des écoles en 3 ans

LES COMPÉTENCES ACQUISES : à l'issue de la formation « Sciences des données et Big Data : outils et introduction », les apprenants auront les compétences suivantes[3] :

Langage de programmation R	
- Notions de calcul statistique avec outil informatique **R** (statistiques élémentaires, estimation et test, analyse descriptive uni et bidimensionnelle).	●●○○○
- Utilisation des outils de classification en **R**.	●●○○○
Langage de programmation Python	
- Compétences de mise en œuvre d'une analyse exploratoire de données multidimensionnelle et des premiers pas de la modélisation statistique à l'aide de la librairie Scikit Learn de **Python**.	●●●●○
- Compétences de mise en œuvre d'une analyse exploratoire de données complexe et de diverses méthodes de classification supervisée de ces données (données fonctionnelles ou en grande dimension) à l'aide de la librairie Scikit Learn de **Python**.	●●●●○
- Compétences de mise en œuvre d'algorithmes standards en **optimisation** numérique (sans contrainte, avec contraintes d'égalité) ainsi que d'algorithmes utilisés pour l'apprentissage.	●●●○○
- Connaissance théorique des principales méthodes d'apprentissage et l'utilisation pratique de ces méthodes en **Python**.	●●●●○
Les infrastructures logicielles pour le big data	
- Connaissance des bases du traitement de données massives dans des infrastructures de calcul.	●●●○○
- Capacité de concevoir une application de traitement avec **Spark**.	●●○○○
- Connaissance des techniques de **virtualisation** des infrastructures et capacité d'utiliser une infrastructure virtualisée.	●●○○○

L'ORGANISATION : la formation s'articule autour de **trois modules** :

- un **module de sensibilisation**, pour être initié aux fondamentaux des sciences des données et comprendre ce qui est le big data ;
- un **module immersion**, pour aller un peu plus loin sur les méthodes et les outils utilisés dans ce domaine, sans pour autant devenir autonome ;
- un **module d'information**, comportant un cycle de conférence.

[3] ●○○○○ = notions ; ●●●●● = expert

Programme détaillé du certificat

Module 1 : MODULE DE SENSIBILISATION (septembre-novembre)

Ce module compte :

- 24h de statistique de base, analyse statistique de données, modélisation statistique et apprentissage statistique ;
- 10h d'infrastructures logicielles pour le big data ;
- 4h d'optimisation ;
- 10h de projet de fin module.

OBJECTIFS

L'objectif de ce module est d'acquérir les connaissances théoriques et pratiques de base pour pouvoir travailler activement en science des données et big data.

D'une part le module prévoit l'introduction des outils logiciel indispensables pour la science des données (**R, Python**) et pour le big data (**Hadoop, Spark, Mapreduce)** pour le traitement pratique des données. D'autre part le module donne aux étudiants les connaissances de base en statistique pour comprendre la théorie derrière les méthodes d'apprentissage présentes. En particulier les suivantes sont les objectifs de ce module :

- Comprendre l'analyse et exploration de masses de données d'un point de vu statistique ;
- Apprendre les notions de base d'optimisation pour l'entraînement de modèles en apprentissage statistique ;
- Apprendre à utiliser une infrastructure virtualisée ;
- Apprendre le traitement de données massives dans des infrastructures de calcul ;
- Mettre en œuvre les notions apprises par des travaux pratiques.

Cours	Programme
MODULE 1 : SENSIBILISATION (septembre-novembre)	
1	**- Introduction à R** **- Eléments d'analyse statistique de données :** statistique élémentaire, descriptive unidimensionnelle et bidimensionnelle, estimation et tests statistiques **- Modélisation statistique :** modèle gaussien-régression linéaire multiple, modèle binomial-régression logistique **FORMAT :** 10h en e-learning **PREREQUIS :** Notions d'analyse statistique pour l'ingénieur & notions de programmation informatique et calcul numérique pour l'ingénieur **UTILITAIRES :** Langage de programmation R (et RStudio comme interface) **EVALUATION :** Exercices de calcul et modélisation statistique en binôme. Rendu obligatoire pour obtention du certificat. Pas de note. **COMPÉTENCES ACQUISES :** Notions de calcul statistique avec outil informatique R (statistiques élémentaires, estimation et test, analyse descriptive uni et bidimensionnelle)
2	**Séance de discussion questions/réponses** sur les contenus de la première phase d'e-learning **FORMAT :** 2h en présentiel **PREREQUIS :** Avoir rendu les exercices de calcul et modélisation statistique de la séance d'e-learning

	précédente
3	- **Introduction à Python** - **Exploration multidimensionnelle :** data mining, ACP-AFD, clustering avec k-means - **Principes de l'apprentissage statistique** **FORMAT :** 6h en e-learning **PREREQUIS :** Notions de base en probabilités et statistique. Compétences élémentaire en algorithmique et programmation. **UTILITAIRES :** Python et jupyter notebook **COMPÉTENCES ACQUISES :** Mise en œuvre d'une analyse exploratoire de données multidimensionnelle et des premiers pas de la modélisation statistique à l'aide de la librairie Scikit Learn de Python.
4	**Travaux pratiques sur l'exploration multidimensionnelle** **FORMAT :** 3h en présentiel, TP en groupe de 25 personnes **PREREQUIS :** Connaissances et compétences acquises dans la partie « e- learning » qui précède **MOTS-CLÉS :** Visualisation des données, réduction de dimension, Analyse en composantes principales, analyse factorielle discriminante, classification non supervisée **UTILITAIRES :** Python et jupyter notebook **COMPÉTENCES ACQUISES :** Mise en œuvre d'une analyse exploratoire de données complexe (données fonctionnelles ou en grande dimension) à l'aide de la librairie Scikit Learn de Python.
5	**Travaux pratiques sur l'apprentissage statistique** **FORMAT :** 3h en présentiel, TP en groupe de 25 personnes **PREREQUIS :** Connaissances et compétences acquises dans la partie « e- learning » qui précède **MOTS-CLÉS :** Apprentissage supervisé à l'aide de modèles linéaires ou de modèles linéaires généralisés (régression logistique) et analyse discriminante. **UTILITAIRES :** Python et jupyter notebook **COMPÉTENCES ACQUISES :** Mise en œuvre de diverses méthodes de classification supervisée sur des données complexes (données fonctionnelles ou en grande dimension) à l'aide de la librairie Scikit Learn de Python
6	**Les algorithmes classique d'optimisation** (introduction sur l'optimisation, algorithmes classiques avec gradient) **FORMAT :** 2h en présentiel, cours magistral **PREREQUIS :** Notions d'algèbre linéaire et de calcul différentiel

	COMPÉTENCES ACQUISES : Analyser, modéliser et résoudre théoriquement et/ou numériquement un problème d'optimisation sans contrainte, ou avec contrainte d'égalité.
7	**Travaux pratiques programmation d'un algorithme classique d'optimisation** **FORMAT :** 2h en présentiel, TP en groupe de 25 personnes **PREREQUIS :** Maîtrise de python, cours d'optimisation **UTILITAIRES :** jupyter notebook, Python **COMPÉTENCES ACQUISES :** Mise en œuvre de méthodes d'optimisation basée sur des directions de descentes. Analyser les performances et résultats de méthodes d'optimisation.
8	**Les techniques de virtualisation et containerisation ; les plateformes cloud** **FORMAT :** 2h en présentiel, cours magistral **COMPÉTENCES ACQUISES :** Connaissance des techniques de virtualisation des infrastructures
9	**Travaux pratiques Déploiement d'une infrastructure dans le cloud** **FORMAT :** 2h en présentiel, TP en groupe de 25 personnes **PREREQUIS :** Connaissances de base de l'utilisation d'un environnement Linux (commande shell de base) ; Connaissances en algorithmique et programmation en Java ou Python **COMPÉTENCES ACQUISES :** Capacité d'utiliser une infrastructure virtualisée
10	**Les infrastructures logicielles pour le big data :** modèle MapReduce ; Démystification Hadoop, Spark **FORMAT :** 2h en ligne, cours magistral **PREREQUIS :** Base en programmation Java ou Python, gestion de structures de données dans ces langages **COMPÉTENCES ACQUISES :** Connaissance des bases du traitement de données massives dans des infrastructures de calcul
11	**Travaux pratiques avec Spark** **FORMAT :** 4h en présentiel, TP en groupe de 25 personnes **PREREQUIS :** Connaissances de base de l'utilisation d'un environnement Linux (commande shell de base) **UTILITAIRES :** Eclipse pour programmation Java ou Pyhton. $Spark **COMPÉTENCES ACQUISES :** Capacité de concevoir une application de traitement avec Spark

PROJET DE FIN MODULE

Projet de fin de module "Développer un algorithme sur une infrastructure"

FORMAT : 10h en distanciel – **travail en binôme ou trinôme**

PREREQUIS : Cours d'infrastructure logicielle pour le big data et d'optimisation

UTILITAIRES : jupyter notebook , Python, Python $Spark, $Spark

EVALUATION : Evaluation sur retour d'un rapport et d'une archive du projet réalisé.
Obligatoire pour la validation de la formation

COMPÉTENCES ACQUISES : Capacité de modélisation et de mise en œuvre d'une application issue des sciences des données dans un environnement $Spark. Capacité à analyser, de manière rudimentaire, les performances d'un algorithme dans un environnement de calculs distribués.

BIBLIOGRAPHIE SUGGEREE

1. https://www.my-mooc.com/fr/categorie/statistiques-et-probabilites
2. http://onlinestatbook.com/
3. http://wikistat.fr/
4. An Introduction to Statistical Learning: With Applications in R
5. Data science : fondamentaux et études de cas : Machine Learning avec Python et R

Module 2 : MODULE IMMERSION (novembre-mars)

Ce module compte :

- 18h de modélisation statistique et apprentissage statistique ;
- 4h d'optimisation ;
- 10h de projet de fin module.

OBJECTIFS

Ce module a pour l'objectif d'utiliser les compétences acquises dans le premier module pour comprendre et mettre en œuvre les principales méthodes de modélisation statistiques des données, les méthodes d'optimisation pour l'adaptation efficace de ces modèles aux données et l'utilisation d'infrastructures virtualisées pour le traitement pratique de grande masses des données. De manière plus détaillées, les objectifs de ce module sont :

- Comprendre la modélisation de masses de données d'un point de vu statistique ;
- Comprendre les principes et mettre en œuvre un certain nombre d'algorithmes d'optimisation pour l'entraînement de modèles en apprentissage statistique ;
- Mettre en œuvre les notions apprises par des travaux pratiques.

Cours	Programme
MODULE 2 : IMMERSION (novembre-mars)	
1	**L'introduction aux méthodes d'apprentissage :** arbre binaire de décision, réseaux de neurones, support vector machines, agrégation d'arbres **FORMAT :** 4h en présentiel, cours magistral **PREREQUIS :** Compétences de base en statistique, acquises par le module de sensibilisation **COMPÉTENCES ACQUISES :** Connaissance théorique de principales méthodes d'apprentissage
2	**Travaux pratiques sur l'arbre binaire de décision** **FORMAT :** 3h en présentiel, TP en groupe de 25 personnes **PREREQUIS :** Connaissance des principales méthodes d'apprentissage acquise par le matériel fourni en cours précèdent **UTILITAIRES :** Langage de programmation Python **COMPÉTENCES ACQUISES :** Utilisation pratique des méthodes d'apprentissage en Python
3	**Les méthodes d'apprentissage :** arbre binaire de décision, réseaux de neurones, support vector machines, agrégation d'arbres **FORMAT :** 8h en e-learning **PREREQUIS :** Compétences de base en statistique, acquises par le module de sensibilisation **COMPÉTENCES ACQUISES :** Connaissance théorique de principales méthodes d'apprentissage

4	**Travaux pratiques sur spam** **FORMAT :** 3h en présentiel, TP en groupe de 25 personnes **PREREQUIS :** Connaissance des principales méthodes d'apprentissage acquise par le matériel fourni en cours précèdent **UTILITAIRES :** Langage de programmation R **COMPÉTENCES ACQUISES :** Utilisation des outils de classification en R
5	**Les algorithmes stochastiques plus sophistiqués :** optimisation parcimonieuse, factorisation non négative de matrice **FORMAT :** 2h en présentiel, cours magistral **PREREQUIS :** Connaissance de techniques de base en optimisation (ex. méthode du gradient) **COMPÉTENCES ACQUISES :** Connaissance de principales techniques d'optimisation pour l'apprentissage statistique
6	**Travaux pratiques programmation d'un algorithme de chaque type** **FORMAT :** 2h en présentiel, TP en groupe de 25 personnes **PREREQUIS :** Notions en optimisation pour l'apprentissage statistique acquises en cours **UTILITAIRES :** Langage de programmation Python **COMPÉTENCES ACQUISES :** Capacité d'implémenter les principales méthodes d'optimisation pour l'apprentissage stochastique
PROJET DE FIN MODULE	**Projet de fin de module "Challenge Kaggle"** **FORMAT :** 10h en distanciel – **travail en groupe de 4-5 personnes** **PREREQUIS :** Connaissance des principales méthodes d'apprentissage vues dans le module immersion **UTILITAIRES :** Connaissance d'un langage de programmation au choix (R, Python conseillés) **EVALUATION :** Note de A à E. **Obligatoire pour la validation de la formation** **COMPÉTENCES ACQUISES :** Capacité de construire un modèle prédictif sur des données réels.

BIBLIOGRAPHIE SUGGEREE POUR « ALLER PLUS LOIN »

1. **Elements of Statistical Learning**: data mining, inference, and prediction

Module 3 : MODULE D'INFORMATION

Ce module d'information est décliné en **cycle de conférences**. Ce cycle comprend cinq conférences :

- **trois sur des retours d'expériences d'applications industrielles** pour illustrer les cas d'application de manière concrète et
- **deux d'ouverture** visant à mettre en lumière les enjeux éthiques et de respect de la vie privée associés au traitement de données.