Diplôme Universitaire
Data Scientist

Programme année 2022-2023

Université Clermont Auvergne

UE 1		Data Analysis with Python
Présentation	Description et objectifs	Cours de programmation dans le langage Python, orienté pour l'analyse de données.
	Compétences visées	- programmer à l'aide du langage Python. - analyser des données efficacement avec numpy - représenter des données avec matplotlib
	Pré-requis	- connaissance de bases d'un langage informatique - connaissance de base de mathématiques et statistiques
	Formation	Lieu : UCA, Campus Cézeaux Durée : 25h Date indicative: Septembre 2022 Formation continue : oui
Contenu	Programme détaillé	1) Introduction to data analysis tools - generalities on Python language - basics on Unix and shell - data analysis with Anaconda and Notebooks 2) Practical introduction - object, collections, functions - loops and few pythonics syntax - basic file manipulation 3) Numpy introduction - numpy arrays vs python list - vectorization, (fancy) indexing, broadcasting 4) Three important tools to know - data representation: matplotlib (histo, 1D plot, 2D scatter) - import/manipulate data as numpy array: pandas (loading csv, accessing columns, plotting) - mathematics, physics and engineering: scipy (studied example: model fitting) 5) Basic of image processing - loading/plotting, colors, grey scale - image filters: kernel, blocks, sliding windows
	Volume horaire	TP : 25h
	Équipe enseignante	Chercheur CNRS du Laboratoire de Physique de Clermont.
	Ressources et moyens	Salle informatique

UE 2		Advanced Statistics
Présentation	Description et objectifs	Cours et exercices pratiques permettant d'appréhender les concepts statistiques essentiels au 'data scientist'.
	Compétences visées	- Maîtriser des concepts statistiques avancés. - Analyser des données à l'aide de méthodes numériques. - Appliquer des méthodes d'analyse prédictive. - Construire des modèles statistiques.
	Pré-requis	Notions de probabilités. Connaissances mathématiques (niveau Licence).
	Formation	Lieu : UCA, Campus Cézeaux Durée : 20h Date indicative: Octobre 2022 Formation continue : oui
Contenu	Programme détaillé	1) Basics on theory of probabilities - Properties of distributions - Random variables distributions - Samples and statistical properties of samples 2) Likelihood - Parameter estimation - Hypothesis testing 3) Confidence level and confidence intervals 4) Bayesian inference
	Volume horaire	CM : 15h TD : 5h
	Équipe enseignante	Enseignant-chercheur UCA du Laboratoire de Physique de Clermont.
	Ressources et moyens	

UE 3		Machine Learning
Présentation	Description et objectifs	Introduction aux techniques multi-variables et aux méthodes d'apprentissage supervisé (Machine Learning).
	Compétences visées	- Appliquer des méthodes de classification et de clustering. - Implémentation d'algorithmes d'apprentissage supervisé. - Concevoir des méthodes d'analyse prédictive et d'optimisation. - Utiliser un algorithme de Deep Learning.
	Pré-requis	Notions de probabilités et de statistiques. Connaissances mathématiques (niveau Licence).
	Formation	Lieu : UCA, Campus Cézeaux Durée : 20h Date indicative : Décembre 2022 Formation continue : oui
Contenu	Programme détaillé	1) Basic concepts - Types of learning: supervised, unsupervised - Regression, Classification, Clustering - Training and testing, cross validation - Bias-variance decomposition - Curse of dimensionality 2) Regression with linear models - Simple exemple: polynomial curve fitting - Linear basis function models - Regularization - Likelihood and regression 3) Classification - Linear discriminant analysis - Logistic regression and Gradient descent - Perceptron's algorithm - Towards Neural Networks 4) Introduction to neural networks - building blocks and architecture - Introduction to Pytorch - Deep learning practice with Pytorch
	Volume horaire	CM : 10h TP : 10h
	Équipe enseignante	Enseignant-chercheur UCA du Laboratoire de Physique de Clermont.
	Ressources et moyens	Salle informatique

UE 4		Data mining
Présentation	Description et objectifs	Introduction aux méthodes de fouille de données et aux techniques d'apprentissage non supervisé.
	Compétences visées	- Savoir évaluer les méthodes de fouille de données. - Savoir utiliser les techniques de fouille de données. - Maîtriser une technique de classification non supervisée. - Maîtriser une technique de recherche de règles d'association. - Connaître les différentes approches de fouille de données complexes.
	Pré-requis	Notions de probabilités et de statistiques. Connaissances mathématiques (niveau Licence).
	Formation	Lieu : UCA, Campus Cézeaux Durée : 20h Date indicative : Novembre 2022 Formation continue : oui
Contenu	Programme détaillé	1) Introduction to big data and data mining - data mining vs machine learning - challenges and criticisms - competence needed 2) Preprocessing data - data type and visualisation - data preprocessing scheme - cleaning methods - space transformation - dimensionality reduction - instance reduction 3) Clustering - hierarchical clustering - partitional clustering (k-means, dbscan) - clustering evaluation 4) Feature reduction and extraction - PCA, CA, MCA, MDS - kernel PCA - locally Linear Embedding - t-SNE
	Volume horaire	TD : 25h
	Équipe enseignante	Eneignant-chercheur UCA du Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes.
	Ressources et moyens	Salle informatique

UE 5		Data Engineering
Présentation	Description et objectifs	Formation aux techniques modernes de calcul scientifique et de traitement des données dispensée par le Centre de Calcul de l'IN2P3 (cc.in2p3.fr).
	Compétences visées	- Comprendre les problèmatiques liées au calcul (architectures modernes, calcul sur la grille et dans le cloud) - Savoir utiliser une grille de calcul. - Savoir choisir le stockage le plus adapté : système de fichiers, stockage dans le cloud et bases de données (SQL et NoSQL). - Programmation avec le langage Python,. - Utilisation d'un notebook de type Jupyter pour mener une analyse de données.
	Pré-requis	Connaissance minimale d'un système d'exploitation (Linux de préférence) et de la ligne de commande. Connaissance minimale d'un langage de programmation (écriture du code, compilation et exécution)
	Formation	Lieu : Centre de Calcul de l'IN2P3 (Villeurbanne) Durée : 30h Date indicative : Mars 2023 Formation continue : oui
Contenu	Programme détaillé	1) Distributed data processing - introduction to Python language and notebooks - Using a Jupyter notebook 2) Computing - computation architectures - CPU and GPU computing, parallel computing - virtualization and distributed computing: cloud and grid 3) Storage - object storage in the cloud - relational and non-relational databases - HADOOP & SPARK ecosystem 4) Visit of the CCIN2P3 computer centre
	Volume horaire	CM : 15h TP : 15h
	Équipe enseignante	Ingénieurs CNRS du Centre de Calcul de l'IN2P3 (CCIN2P3).
	Ressources et moyens	Ressources informatiques du CCIN2P3.

UE 6		Statistical tools : R language
Présentation	Description et objectifs	Ce cours vise à se familiariser avec les fonctionnalités de programmation matricielles et objet du logiciel R.
	Compétences visées	- Se familiariser avec le logiciel R - Savoir programmer sous R(fonctions et objets). - Outils de visualisation graphique - Connaître les interfaçages Python, Shiny, ggplots. - Utiliser des routines C ou Fortran.
	Pré-requis	Connaissance en statistique
	Formation	Lieu : UCA, Campus Cézeaux Durée : 20h Date indicative : Avril 2023 Formation continue : oui
Contenu	Programme détaillé	1) R software and packages for data analysis - objects, structure, packages, functions and graphics 2) Statistics with R - basics, tests, regression, classification 3) Use case studies 4) Links R/python and R/C
	Volume horaire	CM:10h TP: 10h
	Équipe enseignante	Enseignant-chercheurs UCA du laboratoire de mathématiques Blaise Pascal.
	Ressources et moyens	Salle informatique

UE 7		Time series
Présentation	Description et objectifs	Il s'agit d'une introduction à la modélisation et la prévision de séries temporelles.
	Compétences visées	-Connaître les concepts de base des séries temporelles. -Connaître les principaux outils statistiques d'analyse d'une série temporelle. -Savoir modéliser des séries temporelles par les modèles classiques fondés sur des processus à temps discrets usuels et savoir en tirer des prévisions.
	Pré-requis	Connaissance des probabilités, de la statistique avancée, de régression linéaire.
	Formation	Lieu : UCA, Campus Cézeaux Durée : 20h Date indicative : Mai 2023 Formation continue : oui
Contenu	Programme détaillé	1) Trend modelling and seasonality, seasonal adjustment. 2) Main properties of Fourier analysis, periodogram 3) Modeling and forecasting by ARMA process. 4) Elements on the ARIMA and SARIMA models 5) Applications with the R software.
	Volume horaire	CM:8h TD: 4h TP : 8h
	Équipe enseignante	Enseignant-chercheurs UCA du laboratoire de mathématiques Blaise Pascal.
	Ressources et moyens	Salle informatique

UE 8		**Deep Learning**
Présentation	Description et objectifs	Introduction aux méthodes d'apprentissage profond pour la classification, la régression, la réduction de dimensionalité et l'apprentissage de métriques. Introduction aux réseaux de neurones récurrents
	Compétences visées	- Concevoir et entrainer un réseau de neurones profond supervisé. Perceptron multi-couches, DCNN, autoencodeur - Savoir implémenter un réseau de neurones sur une architecture GPUs. - Savoir analyser et évaluer les performances d'un entrainement.
	Pré-requis	Connaissances mathématiques (niveau Licence). Notions de probabilités, de statistiques et d'optimisation. Notions de base en Machine Learning et réseaux de neurones. Programmation Python.
	Formation	Lieu : UCA, Campus Cézeaux Durée : 20h Date indicative : Janvier/Février 2023 Formation continue : oui
Contenu	Programme détaillé	1) Introduction aux réseaux de neurones profonds - architecture générale, fonctions d'activation (Relu, …), softmax. - réseaux convolutifs (DCNN) 2) Entrainement d'un réseau - fonctions de coûts pour la classification et la régression - techniques d'optimisation et de régularisation, - stratégies d'entrainement pour limiter le surapprentissage 3) Architectures de réseaux pour l'analyse d'images : classification, détection, segmentation, réseaux siamois 4) Autoencodeurs - Principe et exemples d'utilisations : réduction de dimension, débruitage, génération 5) Réseaux de neurones récurrents - Structure générale et entrainement - Architectures GRU, LSTM <u>Deep learning practice with Pytorch</u> 1) Réseaux de neurones pour la classification, influence des hyperparamètres et régularisation 2) Auto-encodeurs et application au débruitage d'image 3) Utilisation d'un réseau existant et transfert d'apprentissage 4) Réseau de neurones récurrent et prédiction de séquence
	Volume horaire	CM : 10h TP : 10h