

PROGRAMME « DU Big Data et Statistique pour l'Ingénieur » 2025 - 2026

Cette formation vise, à l'issue du cursus :

- A sensibiliser les stagiaires de la formation aux problématiques et aux outils actuels et futurs du **Big Data et de l'Intelligence Artificielle (IA)** ;
- A maîtriser les outils de la **Statistique, du traitement des données et de l'IA**, et à les mettre en œuvre sur des applications concrètes.

La présentation théorique des différents méthodologies de la Statistique et de l'IA sera illustrée par de nombreux cas pratiques. Des outils actuels du Big Data et leurs mises en œuvre seront présentés par des start-up spécialistes du domaine, hébergées à l'ENSC. La maîtrise de nombreuses méthodes sera proposée à partir d'un logiciel libre R destiné à la Statistique et à la Science des Données, logiciel permettant aux participants de poursuivre éventuellement leurs activités avec ce même outil. D'autres logiciels (**Python, Matlab**) et outils numériques (**IBM Watson, ...**) seront également manipulés lors de la formation.

Session 1 :

Enjeux socio-économiques du Big Data.

Introduction à la Statistique.

Statistique descriptive univariée et bivariée.

Initiation au logiciel statistique R.

Classification non supervisée de variables (approche ClustOfVar)

Session 2 :

Statistique inférentielle : estimateurs, intervalles de confiance et tests d'hypothèses paramétriques et non paramétriques.

Illustration à l'aide du logiciel R.

Modèle linéaire

Modèle statistique

Session 3 :

Statistique multidimensionnelle : analyse en composantes principales, analyse factorielle des correspondances, analyse des données mixtes, analyse discriminante, classification automatique (clustering), classification de variables, réseaux de neurones.

Illustration à l'aide du logiciel R.

Session 4 :

Big Data : stockage de très grands volumes de données dans des architectures passant à l'échelle, calcul parallèle, méthodologie adaptée, etc.

Mise en œuvre pratique.



Session 5 :

Modélisation statistique : régression linéaire simple et régression linéaire multiple, analyse de la variance (ANOVA), analyse de la covariance (ANCOVA).

Illustration à l'aide du logiciel R.

Session 6 :

Modélisation statistique : forêts aléatoires, régression non paramétrique, régression semi paramétrique.

Illustration à l'aide du logiciel R.

Session 7 :

Etudes de cas.

Deep learning.

Analyse des données du web et gestion de la qualité des données.