

## INITIATION A LA DATA SCIENCE

Initiation aux langages, aux outils et aux méthodologies de la Data Science permettant via des modèles de Machine Learning de valoriser vos données.

### Bénéfice attendu pour le client :

Grâce à un syllabus couvrant tous les aspects d'un projet de Machine Learning, vous aurez toutes les clefs en main pour réaliser avec autonomie des projets de Data Science. Par une meilleure compréhension des processus associés, cette formation facilitera également vos interactions avec les Data Scientists.

### Description du module :

Les différentes phases d'un projet de Data Science sont présentées et expérimentées sur 2 jours (Cadrage, Data Prep, Modélisation, Industrialisation). Par ailleurs, et si nécessaire, deux journées optionnelles de formation sont proposées au préalable. Elles sont indépendantes et traitent des sujets suivants : prise en main du langage Python ; présentation et formation à l'utilisation / configuration des outils indispensables à la Data Science. Tout au long de la formation, guidés par nos experts, vous mettrez en pratique outils et théories au service de problématiques concrètes.

### Certification à l'issue de la formation :

Chaque participant passe à la fin de la seconde journée un quiz lui permettant de recevoir la certification Saegus.

### Durée

- 2 à 4 jours (2 jours optionnels)

### Nombre de participants

- Minimum 3 participants
- Maximum 12 participants par session

### Pré requis

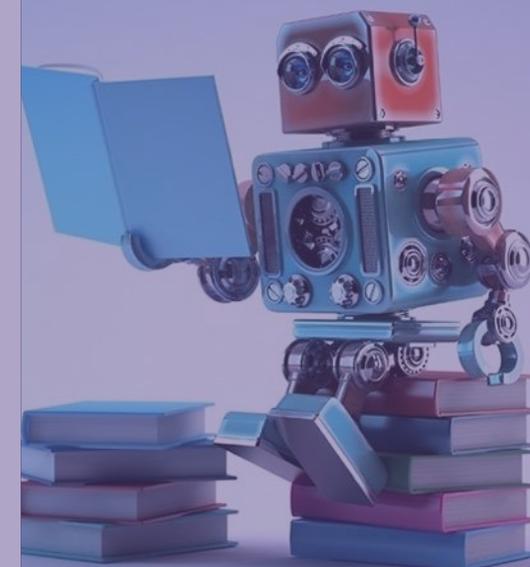
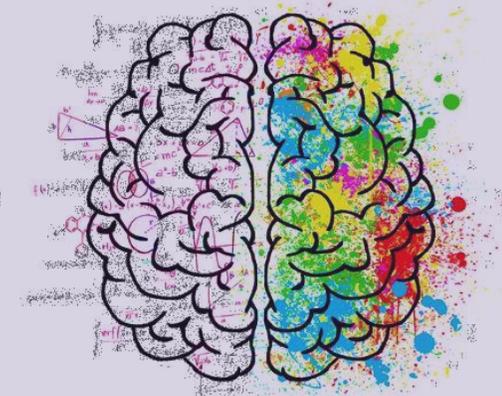
- Connaissances de base en algèbre linéaire et en statistiques
- En l'absence de la journée de formation optionnelle dédiée, une bonne maîtrise de Python et de la programmation bash

### Prestations incluses

- Animation de la formation
- Accès aux notebooks d'entraînement
- Certification Saegus

### Tarif

- Dans les locaux Saegus :
  - 2200€ HT / participant pour 2 jours
  - 3200€ HT / participant pour 3 jours
  - 4000€ HT / participant pour 4 jours
- Dans les locaux du commanditaire :
  - 2000€ HT / participant pour 2 jours
  - 2900€ HT / participant pour 3 jours
  - 3500€ HT / participant pour 4 jours



# INITIATION A LA DATA SCIENCE: DÉROULÉ DE FORMATION

## J1 (Optionnel) : Programmation pour la data science

MATIN	<b>Introduction Python</b>	<i>Introduction générale à Python (historique, intérêts, usages, etc.) et séance d'installation.</i>
	<b>Types natifs, opérateurs logiques et boucles</b>	<i>Les briques élémentaires de la programmation sont présentées : les types natifs, leur manipulation grâce aux opérateurs et aux structures logiques.</i>
	<b>Les fonctions</b>	<i>La programmation fonctionnelle est détaillée : architecture des fonctions, normes d'implémentation (type hinting, docstring, ...), structures récursives, etc.</i>
	<b>Objets, classes et méthodes</b>	<i>La programmation orientée objet est expliquée avec ses instances (objets, classes et méthodes), ses règles et ses bonnes pratiques. Les notions de librairies et de packages sont également abordées.</i>

APRÈS-MIDI	<b>Structure de données avancées : Numpy et Pandas</b>	<i>Introduction aux outils de manipulation de données structurées : array, dataframe, opérations élémentaires (descriptives, par ligne, par colonne, ...) et avancées (pivot, agrégation, fusion, ...), et bonnes pratiques.</i>
	<b>Visualisation de données</b>	<i>La bibliothèque de référence de dataviz matplotlib est présentée : structure globale, principaux graphiques, options de style, bonnes pratiques et alternatives.</i>
	<b>Pause</b>	<i>Pause !</i>
	<b>Programmation Bash</b>	<i>Les rudiments socles de la programmation bash sont transmises : commandes les plus utiles, scripts et structures logiques, expressions régulières.</i>
	<b>Les bases de SQL</b>	<i>Présentation générale des langages SQL, requêtes élémentaires, opérations (filtres, agrégats, jointure, etc.), création de tables, gestion des utilisateurs</i>
	<b>Révisions</b>	<i>Des révisions sont effectuées sur l'ensemble du programme de la journée.</i>
	<b>Passage de la certification</b>	<i>Certification Initiation Python (Optionnel).</i>

## J2 (Optionnel) : Les Outils de la Data Science

MATIN	<b>Jupyter Notebooks</b>	<i>Manipuler et optimiser l'usage des jupyter notebooks : installation, extensions, raccourcis, bonnes pratiques, ... Jupyter lab sera en particulier détaillé.</i>
	<b>Du bon usage d'un IDE</b>	<i>La configuration des IDE (au choix : VS Code, PyCharm) est présentée afin de maximiser la productivité des utilisateurs, et réalisée en séance par les participants.</i>
	<b>Templating de projets et CookieCutter</b>	<i>L'usage de CookieCutter est démontré pour créer des templates afin de générer de nouveaux projets aux architectures répliquables et intégrant tous les outils facilitant les développements.</i>
	<b>Les bases de Git</b>	<i>Toutes les notions élémentaires de Git sont abordées : commit, branches, DAG, collaboration, principales commandes, intro aux plateformes Git (Gitlab, Github), avec une mise en situation des participants.</i>

APRÈS-MIDI	<b>Les fondamentaux des Statistiques</b>	<i>Les connaissances socles des statistiques sont parcourues : stats descriptives, théorie des probabilités, échantillonnage, statistiques inférentielles et tests, rôle des graphiques, mises en pratique.</i>
	<b>Pause</b>	<i>Pause !</i>
	<b>Introduction à l'usage de containers</b>	<i>La technologie Docker est présentée, les concepts d'infrastructures associés, les usages et commandes correspondantes, avec un atelier pratique.</i>
	<b>Révisions</b>	<i>Des révisions sont effectuées sur l'ensemble du programme de la journée.</i>
	<b>Passage de la certification</b>	<i>Certification Outils Data Science (Optionnel).</i>

## Compétences acquises à l'issue de la formation

- Apprentissage du socle de base de la programmation Python, en particulier dans le cadre du traitements de données (optionnel)
- Montée en compétences sur d'autres langage de programmation utiles à la Data Science : bash, SQL (optionnel)
- Familiarisation aux principaux outils facilitant le travail des Data Scientists : notebooks, IDE, templates et CookieCutter, usages de Git et conteneurisation (optionnel).
- Utilisation des statistiques et des probabilités dans l'analyse de données (optionnel)
- Acquisition d'un bagage culturel sur les enjeux de la Data Science et de l'IA
- Cadrage et gestion de projet Data Science
- Techniques de préparation des données
- Méthodes de modélisation du Machine Learning
- Découvertes des solutions d'industrialisation des solutions de Machine Learning

## Moyens pédagogiques & techniques d'encadrement

- 2 formateurs Data Scientists cumulant de nombreuses années d'expériences.
- Mise à disposition de ressources informatiques permettant la bonne réalisation des ateliers pratiques.
- Les supports de présentation permettant l'acquisition des bases théoriques et les supports d'exercices nécessaires à la réalisation des cas pratiques.
- Accès à notre application de validation des connaissances afin de délivrer les Certifications Saegus correspondantes.



# INITIATION A LA DATA SCIENCE: DÉROULÉ DE FORMATION

## J3 : Cadrage Projet et Data Préparation

MATIN	<b>Data Science et IA : Notions Générales</b>	<i>Les réalités et les mythes de l'IA sont abordées, avec une description des cas d'usage emblématiques et la stratégie de mise en œuvre associée. Le cas pratique sur lequel s'appuiera l'ensemble des modules et exercices des jours 3 et 4 est présenté.</i>
	<b>Cadrage de projet Data et Data Thinking</b>	<i>Les méthodes de gestion de projet Data Science sont présentées ainsi que la méthodologie inclusive de Data Thinking qui permet de faire de l'idéation de de la priorisation d'opportunité Data. Une mise en situation permet d'intégrer ces notions par la pratique.</i>
APRÈS-MIDI	<b>Data Prep : nettoyage des données</b>	<i>Les différentes opérations de nettoyage de la donnée sont présentées : sélection des données, valeurs manquantes et imputation, gestion des valeurs aberrantes, typage, ... Un aspect opérationnel sera également abordé, avec mise en situation.</i>
	<b>Data prep : analyse exploratoire des données</b>	<i>Les différentes méthodologies exploratoires des données sont présentées : études des distributions, des corrélations, des différences entre groupes de variables, etc. Un accent particulier sera donné à la visualisation de ces effets lors des ateliers pratiques.</i>
	<b>Pause</b>	<i>Pause !</i>
	<b>Data Prep : Features Engineering</b>	<i>La manipulation et la création de features sont abordées : encodage des valeurs, techniques de standardisation et de normalisation, combinaison,... Ces méthodes seront appliquées à des cas pratiques.</i>
	<b>Révisions</b>	<i>Des révisions sont effectuées sur l'ensemble du programme de la journée.</i>
	<b>Passage de la certification</b>	<i>Certification Cadrage Projet et Data Préparation (Optionnel).</i>

## J4 : Modélisation et Industrialisation

MATIN	<b>Introduction à Scikit-Learn</b>	<i>La bibliothèque scikit-learn est présentée : structures des objets correspondants, périmètre d'application, exemples ... Les alternatives sont également présentées.</i>
	<b>Classification et Régression</b>	<i>Les modèles supervisés classiques sont présentés : GLM, régression logistique, méthodes ensemblistes,... Une zoologie des modèles est présentée avec des éléments facilitant la prise de décision sur les types de modèles à envisager selon les situations. Certains de ces modèles sont testés sur le cas pratique de la formation.</i>
	<b>Sélection de modèles</b>	<i>La théorie de la sélection de modèles est présentée : optimisation, validation croisée, structuration des jeux de données, critères de décisions, underfitting/overfitting,...</i>
	<b>Modèles non-supervisés</b>	<i>Les modèles non-supervisés sont présentés, principalement de type clustering dont K-means et clustering hiérarchique (+ dendrogramme)</i>
APRÈS-MIDI	<b>Réduction de dimension</b>	<i>Les techniques de réduction de dimension sont présentées ainsi que leur finalité dont l'ACP/ACI et le manifold learning.</i>
	<b>Projet Data Science</b>	<i>L'ensemble des notions vues lors des jours 3 et 4 (notamment Data Prep et Modélisation) sont employées dans le cadre d'un mini-projet de Data Science</i>
	<b>Pause</b>	<i>Pause !</i>
	<b>Industrialisation des solutions de ML</b>	<i>Une présentation des enjeux de l'industrialisation des modèles de Machine Learning est effectuée (model serving, model monitoring, model registry, etc.).</i>
	<b>Révisions</b>	<i>Des révisions sont effectuées sur l'ensemble du programme de la journée.</i>
	<b>Passage de la certification</b>	<i>Certification Modélisation et Industrialisation (Optionnel).</i>

## Compétences acquises à l'issue de la formation

- Apprentissage du socle de base de la programmation Python, en particulier dans le cadre du traitements de données (optionnel)
- Montée en compétences sur d'autres langage de programmation utiles à la Data Science : bash, SQL (optionnel)
- Familiarisation aux principaux outils facilitant le travail des Data Scientists : notebooks, IDE, templates et CookieCutter, usages de Git et conteneurisation (optionnel).
- Utilisation des statistiques et des probabilités dans l'analyse de données (optionnel)
- Acquisition d'un bagage culturel sur les enjeux de la Data Science et de l'IA
- Cadrage et gestion de projet Data Science
- Techniques de préparation des données
- Méthodes de modélisation du Machine Learning
- Découvertes des solutions d'industrialisation des solutions de Machine Learning

## Moyens pédagogiques & techniques d'encadrement

- 2 formateurs Data Scientists cumulant de nombreuses années d'expériences.
- Mise à disposition de ressources informatiques permettant la bonne réalisation des ateliers pratiques.
- Les supports de présentation permettant l'acquisition des bases théoriques et les supports d'exercices nécessaires à la réalisation des cas pratiques.
- Accès à notre application de validation des connaissances afin de délivrer les Certifications Saegus correspondantes.



