

Data



# Data Scientist

Analysez des données pour identifier des tendances et faire des prédictions...  
Maîtrisez la Data Science !



Durée de la  
formation :  
10 mois



Diplôme niveau  
7 (Bac+5)\*

**OPENCLASSROOMS**

## Quel est le rôle d'un Data Scientist ?

Les entreprises produisent une quantité astronomique de données. **Être capable de les analyser et les valoriser représente un enjeu crucial et un avantage compétitif indéniable.**

En tant que Data Scientist, votre rôle sera de **traduire un besoin métier en une problématique de data science**, puis de **la résoudre grâce à vos algorithmes**.

Vous réaliserez par exemple des **moteurs de recommandations**, des **prédictions** pour améliorer les ventes de votre entreprise, ou encore des **intelligences artificielles pour des applications mobiles**.

Au contact avec les équipes métiers, vous mènerez à bien un projet data, de la collecte des données à la mise en production de vos algorithmes.

## Ce que vous saurez faire

- Collecter et préparer les données en vue de l'analyse
- Programmer des algorithmes de Machine Learning à l'aide du langage Python
- Déployer des algorithmes dans le cloud avec les outils du Big Data
- Communiquer les résultats à des spécialistes ou des néophytes



# CentraleSupélec

Parcours en partenariat avec CentraleSupélec

## Votre rémunération

Rémunérations moyennes pour le métier de Data Scientist :

- Débutant : 35 000 € à 45 000 € annuels bruts
- Expérimenté : 45 000 € à 65 000 € annuels bruts

(Source : <http://datarecrutement.fr/etude-salaire/>)

Ces profils sont très recherchés, majoritairement en CDI. En freelance, les Data Scientist facturent jusqu'à 1000€ par jour de travail.

## Prérequis

**Niveau de langue** : Pour les apprenants étrangers, un niveau de français B1-B2 (utilisateur indépendant) est conseillé pour la réussite de la formation.

**Matériel** : Accès à un ordinateur (PC ou Mac), muni d'un microphone, une webcam et une bonne connexion internet (3.2 Mbps en envoi et 1.8 Mbps en réception de données). Pour tester la qualité de votre connexion, cliquez sur ce [lien](#).

**Niveau requis** : Prépa scientifique ou Bac + 2 en mathématiques.

### Prérequis techniques :

- Math (analyse réelle, algèbre, proba, stat).
- Notions d'informatique (algorithmique, base de données, terminal).

## Votre orientation

Ce parcours donne accès aux métiers suivants :

- Data Scientist
- Data Analyst, Business Analyst, BI Analyst

## Quel parcours Data est fait pour vous ?

[Data Analyst](#) : Vous débuterez dans la data en analysant des données et en réalisant des reportings et des dashboards.

*Data Scientist* : Vous avez un bagage mathématique, et vous souhaitez réaliser des analyses poussées à l'aide d'algorithmes.

[Ingénieur Machine Learning](#) : Vous avez un solide bagage mathématique et vous souhaitez développer des algorithmes de machine learning avancés.

[Data Architect](#) : Vous avez un bagage informatique et vous souhaitez mettre en place l'architecture et les outils de traitement des données.

## Admission

Complétez notre [test de positionnement](#) pour évaluer votre niveau. Notez que ce test ne constitue pas une validation.

Pour vous inscrire à ce parcours, vous devrez obligatoirement remplir ce [formulaire d'admission](#).

Si vous ne possédez pas le niveau de prérequis attendu et/ou que vous êtes déjà en activité, la durée de votre formation sera allongée.

## Votre diplôme

OpenClassrooms est un établissement privé d'enseignement à distance déclaré au rectorat de l'Académie de Paris, délivrant ses propres diplômes ainsi que ceux d'autres partenaires académiques prestigieux.

A l'issue de votre formation et de la validation de vos compétences par un jury organisé par OpenClassrooms, vous pourrez obtenir le certificat « Data scientist ».

Vous pourrez également obtenir ce [titre enregistré au Répertoire National des Certifications Professionnelles](#)\*, de niveau 7 (Bac+5) sur les cadres français et européen des certifications (European Qualifications Framework), à la condition que vous validiez les pré-requis nécessaires pour accéder à la certification.

Si vous avez des questions à propos de son équivalence pour poursuivre vos études, contactez l'université ou école dans laquelle vous voulez continuer après le diplôme.

\* Fiche accessible à l'adresse suivante :

<https://certificationprofessionnelle.fr/recherche/rncp/34545>

## Projet 1

# Définissez votre stratégie d'apprentissage !

Vous embarquez sur un grand parcours d'apprentissage ! Équipez-vous des outils et des bonnes pratiques dont vous aurez besoin tout au long de vos cours et de vos projets.

## Compétences cibles

- Construire pas à pas son projet professionnel

## Cours associés



### Apprenez à apprendre



Facile



6 heures

Être capable d'apprendre vite et bien est une compétence clé qui vous ouvrira les portes de n'importe quel domaine, tout au long de votre vie. Suivez ce cours pour améliorer votre capacité d'apprentissage !

## Projet 2

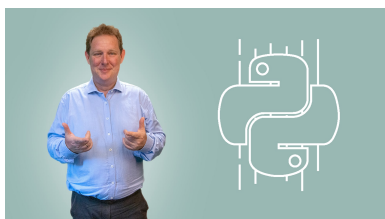
# Analysez des données de systèmes éducatifs

L'entreprise "Academy" cherche à s'étendre à l'international. Dans ce premier projet, vous ferez des recommandations stratégiques à partir de données de systèmes éducatifs.

## Compétences cibles

- Utiliser un notebook Jupyter pour faciliter la rédaction du code et la collaboration
- Mettre en place un environnement Python
- Manipuler des données avec des librairies Python spécialisées
- Effectuer une représentation graphique à l'aide d'une librairie Python adaptée
- Maîtriser les opérations fondamentales du langage Python pour la Data Science

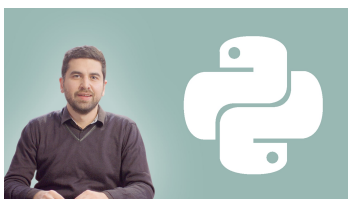
## Cours associés



### Initiez-vous à Python pour l'analyse de données

 Facile  12 heures

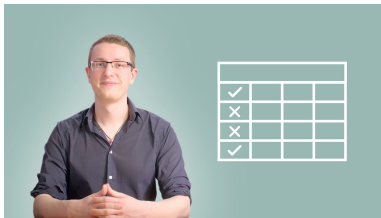
Dans ce cours, vous apprendrez un langage de programmation indispensable pour l'analyse de données : Python. Nous aborderons ensemble les notions fondamentales de la programmation Python, à l'aide d'exemples simples et d'exercices pratiques.



### Découvrez les librairies Python pour la Data Science

 Moyenne  10 heures

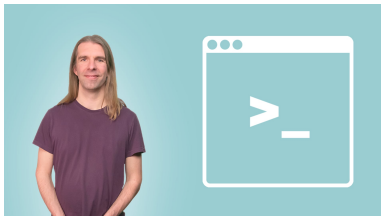
Python s'est imposé comme le langage incontournable pour la Data Science et le Machine Learning, avec de nombreuses librairies spécialisées. Découvrez les notebooks Jupyter et les librairies incontournables que sont Numpy, Matplotlib et Pandas.



## Décrivez et nettoyez votre jeu de données

 Moyenne  15 heures

Prêt à entrer dans l'univers de la statistique descriptive ? Avec ce cours, vous découvrirez comment se compose un jeu de données. Vous serez capable de le nettoyer et le décrire en vue de l'analyser.



## Apprenez à utiliser la ligne de commande dans un terminal

 Facile  6 heures

Bienvenue dans ce cours d'introduction pour apprendre à utiliser la ligne de commande ! Apprenez à écrire des lignes de commandes qui vous permettront de communiquer avec votre ordinateur.

## Projet 3

# Concevez une application au service de la santé publique

L'agence "Santé publique France" a lancé un appel à projet autour des problématiques alimentaires. Vous proposerez une application basée sur des données nutritionnelles.

## Compétences cibles

- Communiquer ses résultats à l'aide de représentations graphiques lisibles et pertinentes
- Effectuer une analyse statistique univariée
- Effectuer des opérations de nettoyage sur des données structurées
- Effectuer une analyse statistique multivariée

## Cours associés



### Initiez-vous au Machine Learning

 Moyenne  10 heures

Découvrez le Machine Learning et ses différentes techniques (régression linéaire, classification non supervisée...). Vous verrez comment un algorithme apprend pour résoudre un problème de Data Science, et vous entraînerez votre premier modèle !



# Anticipez les besoins en consommation électrique de bâtiments

Pour atteindre son objectif de ville neutre en émissions de carbone en 2050, la ville de Seattle a besoin de vous. Votre mission ? Prédire la consommation électrique des bâtiments municipaux.

## Compétences cibles

- Adapter les hyperparamètres d'un algorithme d'apprentissage supervisé afin de l'améliorer
- Évaluer les performances d'un modèle d'apprentissage supervisé
- Mettre en place le modèle d'apprentissage supervisé adapté au problème métier
- Transformer les variables pertinentes d'un modèle d'apprentissage supervisé

## Cours associés



### Évaluez les performances d'un modèle de machine learning

 Moyenne  10 heures

Apprenez à évaluer un algorithme de machine learning, évitez le sur-apprentissage, et choisissez le meilleur modèle pour votre problème, à l'aide de la validation croisée et la grid-search.



## Entraînez un modèle prédictif linéaire

■ Moyenne ⌚ 10 heures

Découvrez les algorithmes d'apprentissage supervisés. Appliquez une régression linéaire ou logistique et appréhendez les méthodes à large marge (SVM).



## Utilisez des modèles supervisés non linéaires

■ Moyenne ⌚ 12 heures

Etendons les méthodes linéaires à la modélisation de relations non linéaires entre les données, notamment à l'aide du SVM et du perceptron. Vous découvrirez aussi une famille d'algorithme très populaire... les réseaux de neurones !



## Modélisez vos données avec les méthodes ensemblistes

■ Moyenne ⌚ 15 heures

Décuplez la robustesse et l'efficacité de vos algorithmes à l'aide des méthodes ensemblistes, le bagging et le boosting. Vous découvrirez aussi les forêts aléatoires et le très prisé XGBoost.

# Segmentez des clients d'un site e-commerce

Vous êtes consultant pour Olist, un site e-commerce brésilien. Les équipes marketing ont besoin de segmenter leurs clients pour optimiser les campagnes de communication.

## Compétences cibles

- Mettre en place le modèle d'apprentissage non supervisé adapté au problème métier
- Transformer les variables pertinentes d'un modèle d'apprentissage non supervisé
- Adapter les hyperparamètres d'un algorithme non supervisé afin de l'améliorer
- Évaluer les performances d'un modèle d'apprentissage non supervisé

## Cours associés



### Explorez vos données avec des algorithmes non supervisés



Difficile



15 heures

Comment faire parler vos données, sans les étiquetter ? Apprenez à mettre en œuvre le clustering (k-means, DBSCAN, clustering hiérarchique) et la réduction dimensionnelle (ACP, MDS, t-SNE)

## Projet 6

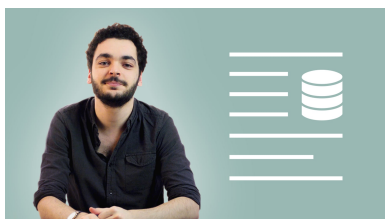
# Classifiez automatiquement des biens de consommation

Votre entreprise cherche à lancer une place de marché e-commerce. Vous devrez tester la faisabilité d'un moteur de classification de biens de consommation.

## Compétences cibles

- Collecter des données répondant à des critères définis via une API
- Prétraiter des données image pour obtenir un jeu de données exploitable
- Prétraiter des données texte pour obtenir un jeu de données exploitable
- Représenter graphiquement des données à grandes dimensions
- Mettre en œuvre des techniques de réduction de dimension

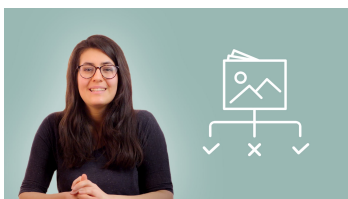
## Cours associés



### Analysez vos données textuelles

■ Moyenne ⌚ 8 heures

Les données textuelles, non structurées, sont omniprésentes dans vos fils d'actualité, ou encore sur les réseaux sociaux. Transformez et modélisez vos données textes grâce aux bag of words, aux word embedding et même aux réseaux de neurones !



### Classez et segmentez des données visuelles

■ Difficile ⌚ 15 heures

Enrichissez votre palette de Data Scientist en classant des données visuelles. Dans ce cours, vous allez prétraiter des images et les modéliser grâce au SIFT et au Deep Learning (CNN).

## Projet 7

# Implémentez un modèle de scoring

Au sein d'une société financière, vous allez développer et implémenter un modèle de scoring pour aider les équipes métiers à accorder un crédit à un client.

## Compétences cibles

- Déployer un modèle via une API dans le Web
- Réaliser un dashboard pour présenter son travail de modélisation
- Rédiger une note méthodologique afin de communiquer sa démarche de modélisation
- Utiliser un logiciel de version de code pour assurer l'intégration du modèle
- Présenter son travail de modélisation à l'oral

## Cours associés



### Utilisez Git et GitHub pour vos projets de développement



Facile



12 heures

Grâce à Git et GitHub, gérez votre code source et suivez les modifications apportées à vos fichiers au fur et à mesure que vos projets de programmation se construisent !

## Projet 8

# Déployez un modèle dans le cloud

Votre startup AgriTech souhaite développer une application mobile permettant de détecter des fruits sur une photo. A vous d'industrialiser le modèle à grande échelle grâce aux outils du big data !

## Compétences cibles

- Utiliser les outils du cloud pour manipuler des données dans un environnement Big Data
- Identifier les outils du cloud permettant de mettre en place un environnement Big Data
- Paralléliser des opérations de calcul avec Pyspark

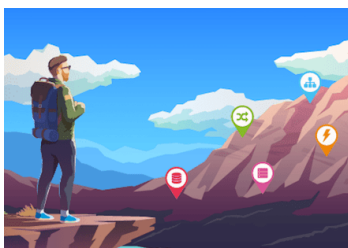
## Cours associés




### Découvrez le cloud avec Amazon Web Services

 Moyenne  20 heures

Vous avez entendu parler du cloud et notamment d'Amazon Web Services (AWS), le leader du cloud ? Venez découvrir comment l'utiliser dans ce cours d'introduction !



### Concevez des architectures Big Data

 Moyenne  6 heures

Nous sommes à l'âge d'or du Big Data et les Data Architects disposent de tous les outils dont ils ont besoin pour gérer des données massives. Mais comment les assembler ? Familiarisez-vous avec une vision d'ensemble pour la conception d'architectures Big Data complètes.



## Réalisez des calculs distribués sur des données massives

 Moyenne  20 heures

Dans ce cours, vous apprendrez à réaliser des analyses de données massives sur des centaines de machines dans le cloud grâce à Hadoop MapReduce, Spark et Amazon Web Services.