

Data



# Data Scientist

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



Master's-level diploma \*

\*Data scientist - code NSF 326 - Master's-level diploma - certification professionnelle enregistrée au Répertoire National des Certifications Professionnelles (RNCP) par décision de France compétences publiée le 20/03/2020

**OPENCLASSROOMS**

## Projet 1 - 10 heures

# Définissez votre stratégie d'apprentissage

Mettez-vous dans les meilleures conditions pour réussir votre parcours : projetez-vous dans votre formation, définissez votre planning et appropriiez-vous les outils essentiels pour apprendre.

## Compétences cibles

- Définir le cadre de votre formation

## Cours associés



### Engagez-vous dans votre formation OpenClassrooms

 Facile  2 heures

Prenez en main votre parcours OpenClassrooms et réalisez votre premier projet en suivant ce cours conçu pour vous accompagner dans ces premières étapes de formation.

## Projet 2 - 60 heures

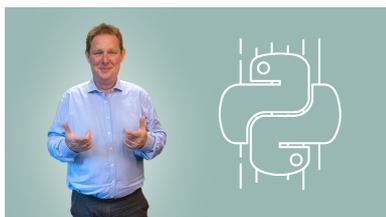
# Analysez des données de systèmes éducatifs

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

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

## Cours associés



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

 Facile  6 heures

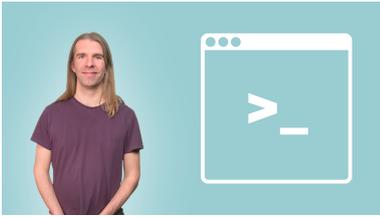
Dans ce cours, vous apprendrez un langage de programmation indispensable pour l'analyse de données : Python. Nous aborderons ensemble les notions fondamentales de la programmation Python, à l'aide d'exemples simples et d'exercices pratiques.



### Découvrez les librairies Python pour la Data Science

 Moyenne  8 heures

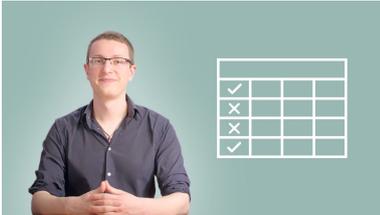
Découvrez les librairies Python spécialisées pour la Data Science. Maîtrisez NumPy, Matplotlib, Pandas et Seaborn pour créer des Data Frames et des Data Visualisations.



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

 Facile  6 heures

Apprenez à écrire des lignes de commandes qui vous permettront de communiquer avec votre ordinateur.



## Nettoyez et analysez votre jeu de données

 Facile  10 heures

Prêt à entrer dans l'univers de la statistique descriptive ? Après ce cours, vous serez capable de nettoyer et décrire un jeu de données.

## Projet 3 - 70 heures

# 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 des opérations de nettoyage sur des données structurées
- Effectuer une analyse statistique multivariée
- Effectuer une analyse statistique univariée

## 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 !



### Réalisez une analyse exploratoire de données

 Moyenne  15 heures

Comprenez les tendances de votre jeu de données avec la puissante ACP (Analyse en Composantes Principales) et d'autres méthodes de classification automatique comme le k-means.

## Projet 4 - 100 heures

# Anticipez les besoins en consommation 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'algorithmes 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

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

## Cours associés



### Explorez vos données avec des algorithmes non supervisés

 Difficile  15 heures

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

## Projet 6 - 100 heures

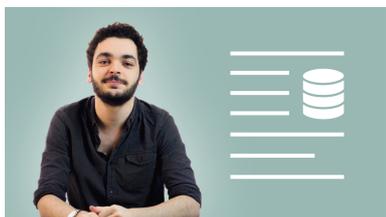
# Classifiez automatiquement des biens de consommation

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

- Définir la stratégie de collecte de données en recensant les API disponibles
- Définir la stratégie d'élaboration d'un modèle d'apprentissage profond
- Évaluer la performance des modèles d'apprentissage profond selon différents critères
- Utiliser des techniques d'augmentation des données
- Prétraiter des données texte pour obtenir un jeu de données exploitable
- Représenter graphiquement des données à grandes dimensions
- Prétraiter des données image pour obtenir un jeu de données exploitable
- Mettre en œuvre des techniques de réduction de dimension

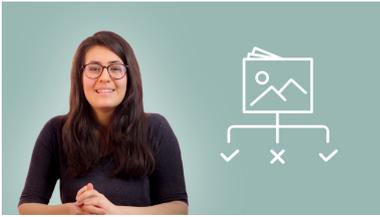
## Cours associés



### 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).

# 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
- Définir et mettre en œuvre une stratégie de suivi de la performance d'un modèle
- Définir et mettre en œuvre un pipeline d'entraînement des modèles
- Définir la stratégie d'élaboration d'un modèle d'apprentissage supervisé
- Évaluer les performances des modèles d'apprentissage supervisé

## Cours associés



### Gérez du code avec Git et GitHub

 Facile  6 heures

Installez Git et GitHub et apprenez les commandes de base de Git pour gérer votre code et déployer vos projets de développement.

## Projet 8 - 70 heures

# 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 ⌚ 12 heures

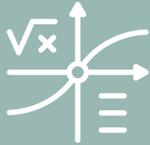
Grâce à ce cours, prenez en main le cloud Amazon Web Services (AWS), montez un serveur web EC2, installez une base de données sur RDS et stockez des fichiers sur S3.



### 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

 Difficile  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.