Big Data - Python pour l'analyse de données

Cours Pratique de 3 jours - 21h
Réf : PBD - Prix 2024 : 2 280€ HT

Le langage Python dispose d'un écosystème scientifique, permettant entre autres, les traitements statistiques : de la construction de modèles d'analyse, à leur évaluation jusqu'à leur représentation. Ce cours vous permet d'analyser des données d'horizon divers avec les bibliothèques Python.

OBJECTIFS PÉDAGOGIQUES

À l'issue de la formation l'apprenant sera en mesure de :

Comprendre le principe de la modélisation statistique

Savoir utiliser les principaux outils de traitement et d'analyse de données pour Python

Savoir appliquer les pratiques optimales en matière de nettoyage et de préparation des données avant l'analyse

Choisir entre la régression et la classification en fonction du type de données

Apprendre à mettre en place un modèle d'apprentissage simple

Être capable d'extraire des données d'un fichier

Développement/réalisation d'analyses avec Python, utilisations des modules pandas, NumPy, SciPy.

LE PROGRAMME

dernière mise à jour : 11/2022

1) Présentation de l'écosystème Python scientifique

- Panorama de l'écosystème scientifique de Python : les librairies incontournables.
- Savoir ou trouver de nouvelles librairies et juger de leur pérennité.
- Les principaux outils et logiciels open source pour la data science.

Travaux pratiques : Installation de Python 3, d'Anaconda et de Jupiter Notebook.

2) Travailler les données avec Python

- Le socle scientifique Python : la SciPy Stack.
- Les bonnes pratiques pour bien démarrer votre projet de data science avec Python.
- Les formats de fichiers scientifiques et les librairies pour les manipuler.
- Pandas : l'analyse de données tabulaires (fichiers csv, excel...), statistiques, pivots, filtres, recherche…
- Numpy : calcul numérique et algèbre linéaire (les vecteurs, matrices, images).
- L'extraction des données,la préparation , le nettoyage.

Travaux pratiques : Ecrire des scripts Python permettant de travailler avec des données issues de fichiers, afin d'appliquer des filtres, des traitements de formatage, de nettoyage.

3) Introduction à la modélisation

- Les étapes de construction d'un modèle.
- Les algorithmes supervisés et non supervisés.

PARTICIPANTS

Développeurs en Python, responsables infocentre, développeurs de logiciels, programmeurs, data analysts, data scientists.

PRÉREQUIS

Maîtrise de la programmation Python. Connaissances de base en statistiques ou avoir suivi le stage "Statistiques, maîtriser les fondamentaux" (Réf. STA).

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les stages pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque stage ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le stagiaire a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Vous avez un besoin spécifique d'accessibilité ? Contactez Mme FOSSE, référente handicap, à l'adresse suivante psh-accueil@orsys.fr pour étudier au mieux votre demande et sa faisabilité.

- Le choix entre la régression et la classification.
Travaux pratiques : Intégration dans l'environnement installé de scripts Python, pour analyse.

4) Procédures d'évaluation de modèles
- Les techniques de ré-échantillonnage en jeu d'apprentissage, de validation et de test.
- Test de représentativité des données d'apprentissage.
- Mesures de performance des modèles prédictifs.
- Matrice de confusion, de coût et la courbe ROC et AUC.
Travaux pratiques : Mise en place d'échantillonnage de jeux de données. Effectuer des tests d'évaluations sur plusieurs modèles fournis.

5) Les algorithmes supervisés
- Le principe de régression linéaire univariée.
- La régression multivariée.
- La régression polynomiale.
- La régression régularisée.
- Le Naive Bayes.
- La régression logistique.
Travaux pratiques : Mise en œuvre des régressions et des classifications sur plusieurs types de données.

6) Les algorithmes non supervisés
- Le clustering hiérarchique.
- Le clustering non hiérarchique.
- Les approches mixtes.
Travaux pratiques : Traitements de clustering non supervisés sur plusieurs jeux de données.

LES DATES

CLASSE À DISTANCE
2024 : 27 mai, 17 juil., 28 oct.

PARIS
2024 : 10 juil., 21 oct.