Stage pratique
Réf : BDA spé

Participants

Responsables Infocentre (Datamining, Marketing, Qualité...), utilisateurs et gestionnaires métiers de bases de données.

Pré-requis

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

Modalités d'évaluation

L'évaluation des acquis se fait tout au long de la session au travers des multiples exercices à réaliser (50 à 70% du temps).

Big Data Analytics avec Python (4 jours)
modélisation et représentation des données

Le Big Bata Analytics repose sur la maîtrise des techniques d'exploration de données fondamentales : statistiques descriptives, prédictives ou exploratoires. Ce stage pratique vous présentera des méthodes telles que les régressions et les ACP et vous apprendra à les mettre en œuvre avec le logiciel Python.

OBJECTIFS PEDAGOGIQUES

Comprendre le principe de la modélisation statistique
Choisir entre la régression et la classification en fonction du type de données
Évaluer les performances prédictives d'un algorithme
Créer des sélections et des classements dans de grands volumes de données pour dégager des tendances
Identifier l'utilité des algorithmes présentés pour l'apprentissage automatique

Travaux pratiques

Développement/réalisation d'analyses sur le logiciel Python, avec les modules pandas, NumPy, SciPy, MatPlotLib, seaborn, scikit-learn et statsmodels.

PARTICULARITES :

L'accent devra être placé sur l'utilité des algorithmes présentés pour
l'Apprentissage Automatique.

1) Introduction à la programmation en Python sur 1 jour

- Syntaxe du langage Python.
- Programmation Objet en Python.
- Débogage.
- Recherche, téléchargement, utilisation des modules.
- Manipulation de data frame (concaténation, ré-indexation,…) avec le module pandas.

Travaux pratiques
Installation de Python 3, écriture de programmes de base, utilisation du module pandas, débogage.

2) Introduction à la modélisation

- Les étapes de construction d'un modèle.
- Les algorithmes supervisés et non supervisés.
- Le choix entre la régression et la classification.

Travaux pratiques
Mise en place d'échantillonnage de jeux de donnes. Effectuer des tests d'évaluations sur plusieurs modèles fournis.

2) 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 donnes. Effectuer des tests d'évaluations sur plusieurs modèles fournis.

3) Les algorithmes supervisés

- Le principe de régression linéaire univariée.
- La régression multivariée.
- La régression polynomiale.
- La régression régularisée.
- La régression logistique.

- Apprentissage par arbre de décision. L'algorithme Random Forest.
- Les réseaux de neurones.
- Les réseaux de neurones récurrents et les transformeurs pour l'analyse de séries temporelles.

Travaux pratiques
Mise en œuvre des régressions sur plusieurs types de données. Mise en œuvre d'un apprentissage supervisé à l'aide de l'algorithme Random Forest puis à l'aide d'un réseau de neurones. Mise en œuvre d'un apprentissage de réseaux de neurones récurrents.

4) 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.

5) Analyse en composantes

- Analyse en composantes principales.

Travaux pratiques
Mise en œuvre de la diminution du nombre des variables et identification des facteurs sous-jacents des dimensions associées à une variabilité importante.