Introduction

Cette formation de trois jours offre une introduction aux fondamentaux de l'Analyse de Données sur l'environnement KNIME®. La formation est destinée aux personnes ayant une connaissance de KNIME® mais sans expérience préalable dans l'analyse de données. Les thèmes abordés au travers du cours sont la prise en main de l'outil KNIME®, la collecte de données, l'exploration, le traitement, la visualisation, ainsi que la programmation des plus importants modèles de données par diagrammes de nœuds.

À qui s'adresse cette formation ?

Cette formation s'adresse aux scientifiques qui souhaitent acquérir une connaissance solide de l'analyse de données sur l'environnement KNIME®.

Vous allez apprendre à :

•Collecter vos données depuis différentes sources, explorer vos données et comprendre leur nature, décorréler, filtrer les échantillons et variables inutiles.

•Choisir un modèle adapté à vos données et le tester, ainsi que visualiser et mettre en valeur vos résultats dans un environnement ouvert, puissant et polyvalent.

Quels sont les prérequis ?

•Avoir une première expérience prouvée dans l'utilisation de l'environnement KNIME® à un niveau équivalent à la formation PIKAÏROS « *Initiation à l'Exploration de Données à travers KNIME* » ou avoir passé la certification L1 de KNIME.

•Comprendre l'anglais informatique élémentaire (outil en langue anglaise.)

Modalités, délais d'accès et adaptation aux personnes en situation de handicap

Cette formation est entièrement basée sur le logiciel pédagogique KNIME avec la mise en place de workflows comme méthode d'apprentissage. Une évaluation continue par QCM est effectuée sur les exercices proposés lors de la formation. La formation est proposée uniquement en présentiel en intra-entreprise. Possibilité d'adapter le temps et mode de formation aux personnes en situation d'handicap. Délais d'accès maximum d'un mois entre la commande et le début de la formation.

Programme complet

Premier jour

Présentation de l'environnement d'analyse Knime

Introduction à l'Analyse de Données

- Définition d'Analyse de Données
- Quelle est la relation entre l'analyse de données et autres disciplines scientifiques : exploration de données, apprentissage automatique, reconnaissance de formes, intelligence artificielle, statistiques.
- Les étapes essentielles de l'Analyse de Données : définition de besoins, collecte, exploration, traitement et modélisation des données, visualisation et communication des résultats.

Problématiques résolues à l'aide de l'Analyse de Données : régression, inférence, classification et partitionnement.

La nature cachée des données :

- Les qualités d'un bon ensemble de données : L'importance de la dimension des données, le nombre d'échantillons, le nombre de variables.
- Le traitement et filtrage des données : Sélection de variables, corrélation, le domaine d'existence des données, la course à la dimensionnalité.
- Pourquoi et comment diviser un ensemble d'apprentissage : Échantillonnage des données, techniques de validation croisée

Exercices : Collecte, traitement et filtrage de données dans l'environnement Knime

- Lecture de données, échantillonnage de données et génération des ensembles d'apprentissage, Validation Croisée et Test
- Evaluation de la corrélation entre variables
- Mise en pratique de la sélection de variables

Deuxième jour

Nature des Modèles de Données :

- **Modèles classiques**
- Modèle linéaire
- Modèle Bayésien
- Analyse de Données traditionnelle :
  - Analyse en Composantes Principales
  - Arbres de Décision et
  - Regroupement Hiérarchique

Exercices: La Modélisation de Données en pratique

- Application des méthodes présentées durant le cours sur des exemples réels
  - Classificateur Linéaire
  - Analyse en Composantes Principales (ACP)
  - Modèle Naïf Bayésien
  - Arbres de Décision
  - Un exemple pratique de surapprentissage
  - Les effets de la course à la dimensionnalité
  - Ecriture et Lecture des Modèles de Données dans KNIME®

Troisième jour

La nature des Modèles de Données :

- Méthodes d'Apprentissage Automatique
- Définition de l'Apprentissage Automatique :
  - Notion d'Apprentissage et Fonction de Coût
  - Degrés de liberté d'un modèle
  - Surapprentissage et méthodes de régularisation

- Apprentissage supervisé :
  - les méthodes à Noyaux (RBF et SVM)
  - Forêts d'Arbres Décisionnelles
  - Ensembles de Classification

- Apprentissage non-supervisé :
  - la méthode des K-Moyennes
  - Espérance-Maximisation
  - T-SNE (t-distributed stochastic neighbour embedding)
  - Méthodes de Regroupement Hiérarchique

Exercices : La Modélisation de Données en pratique

- Application des méthodes présentées durant le cours sur des exemples réels:
- les méthodes à Noyaux (RBF et SVM)
- Forêts d'Arbres Décisionnelles, Ensembles de Classification
- K-Moyennes, T-SNE