Qualité et préparation des données pour l'IA

INFORMATIONS GÉNÉRALES

Type de formation : Formation continue

Domaine : IA, Big Data et Bases de données

Filière : IA

Rubrique : Certification ATLAS : CISIA (Actions co.)

Éligible au CPF : Non

Action collective : Oui
Code ACO : CISIA

PRÉSENTATION

Objectifs & compétences

Identifier et valider les sources de données pertinentes pour l'IA.
Détecter et corriger les erreurs et biais présents dans les jeux de données.
Appliquer les techniques de nettoyage et de normalisation des données.
Améliorer la qualité des jeux de données pour garantir des modèles IA performants.

Public visé

Professionnels de l'IT, Data scientists, ingénieurs en données, et professionnels de l'IA impliqués dans la préparation et le traitement des données.

Pré-requis

Connaissances de base en gestion de données et en analyse de données
Connaissance de base en gestion de données et en statistiques. Expérience préalable avec des outils de traitement de données est un plus.

PROGRAMME

Identifier et Valider les Sources de Données Pertinentes – 2H00 – 2
Évaluer et Nettoyer des Données – 12H00 – 1, 2

Outils utilisés?: Python, Scikit-learn, Jupyter Notebook, Optuna.

Mots clés?: Nettoyage des données, Data wrangling, Filtrage des anomalies, Gestion des valeurs manquantes, Détection des «?outliers?», Balance des classes, Intégrité des données, Validation des jeux de donnée, Biais et éthique des données.

CISIA-PDD01 - Identifier et Valider les Sources de Données Pertinentes - 2h

Introduction (15 min)

- **Objectif :** Présenter les objectifs de la formation et l'importance de choisir des sources de données pertinentes et accessibles.
- **Contenu :** Vue d'ensemble des critères de sélection des sources de données.

Session 1 : Identifier les Critères de Pertinence et d'Accessibilité des Sources de Données (45 min)

- **Objectif :** Identifier des sources de données pertinentes pour les besoins métiers et les cas d'usage (Compétence C1)
- **Contenu :**
  - Critères de pertinence : alignement avec les besoins métiers, qualité des données

Code de formation : CISIA-PDD

€ Tarifs

Prix public : 2000 €

Tarif & financement :
Financement possilble via les Actions Collectives ATLAS ou le Plan de Formation.

Lieux & Horaires

Durée : 12 heures

Rythme : 9h30-12h30 et 14h-17h

Délai d'accès :
Jusqu'a 8 jours avant le début de la formation, sous condition d'un dossier d'insciption complet

Prochaines sessions

Consultez-nous pour les prochaines sessions.

- Critères d'accessibilité : disponibilité, droits d'accès, facilité d'intégration
  - **Activités :**
- **Présentation théorique :** Critères de sélection des données
- **Exercice pratique :** Étude de cas pour identifier des sources de données pertinentes

Session 2 : Évaluer les Risques Éthiques et Sociétaux des Sources de Données (30 min)

- **Objectif :** Identifier les risques éthiques et sociétaux associés aux sources de données choisies (Compétence C2)
- **Contenu :**

- Évaluation des risques : biais potentiels, respect de la confidentialité, conformité réglementaire

- **Activités :**
- **Discussion en groupe :** Identifier les risques éthiques dans des scénarios de données réels
- **Étude de cas :** Analyse des implications éthiques des données choisies

Session 3 : Utiliser les Techniques et Outils pour Valider les Sources de Données (30 min)

- **Objectif :** Préparer les données pour renforcer leur intégrité et leur pertinence (Compétence C3)
- **Contenu :**
- Techniques pour évaluer la qualité des données disponibles
- Outils pour la recherche et la validation des sources de données
- **Activités :**
- **Démo pratique :** Utilisation d'outils pour trouver et valider des sources de données
- **Exercice pratique :** Recherche de sources de données en ligne et validation de leur pertinence

Conclusion et Évaluation (30 min)

- **Objectif :** Réviser les concepts clés abordés et évaluer la compréhension des participants
- **Contenu :** Résumé des points importants, évaluation par QCM

CISIA-PDD02 - Évaluer et Nettoyer des Données - 12h

Introduction (1h)

- Présentation des objectifs de la formation
- Contexte et importance de la qualité des données

Session 1 : Évaluation de la qualité et de la pertinence des données (4h)

- **Concepts clés** : Visualisation des données, indicateurs de cohérence, distribution des données
- **Activités :**
- **Présentation théorique :** Méthodes pour évaluer la qualité des données
- **Atelier pratique :** Analyse de jeux de données réels à l'aide d'outils de visualisation
- **Lien avec les compétences :**
- C1 : Identifier un jeu de données pour répondre aux besoins métiers et aux cas d'usage en tenant compte des enjeux de pertinence et de cohérence.

Session 2 : Identification des biais et évaluation des risques résiduels (4h)

- **Concepts clés** : Types de biais courants, techniques d'atténuation, évaluation des risques résiduels
- **Activités :**

- **Présentation théorique :** Détection des biais et méthodes d'atténuation
- **Étude de cas :** Identification des biais dans des jeux de données et évaluation des risques
  - **Lien avec les compétences :**
- C2 : Identifier les risques éthiques et sociétaux à prendre en compte dans le cadre de l'exploitation de la solution d'IA pour prévenir les dérives éventuelles, en tenant compte du cadre règlementaire.

Session 3 : Data-Cleaning : Méthodes et applications (4h)

- **Concepts clés** : Traitement des données manquantes, identification et traitement des données aberrantes
  - **Activités :**
- **Présentation théorique :** Techniques de nettoyage des données
- **Atelier pratique :** Application des techniques de data-cleaning sur des jeux de données
  - **Lien avec les compétences :**
- C3 : Préparer les données pour renforcer leur intégrité et leur pertinence en vue du développement de la solution IA, en mobilisant les techniques de traitement adaptées et en tenant compte des attendus (besoins métiers, cas d'usage, etc.) identifiés en phase de cadrage du projet.

Conclusion et Évaluation (1h)

- **Synthèse des acquis :** Récapitulatif des principales compétences acquises
- **Évaluation :** QCM et études de cas pour évaluer la compréhension des objectifs pédagogiques
- **Feedback :** Retour sur la formation et discussion des points à améliorer

MODALITÉS

Modalités

L'ensemble du parcours est accessible en présentiel, à distance ou mode hybride.

Présentation théorique en présentiel.
Atelier pratique avec exercices en ligne et en présentiel.
Études de Cas : Analyse d'applications réelles des techniques de génération et d'augmentation.
Discussion Interactive : Échange sur les meilleures pratiques, les défis rencontrés et les retours d'expérience.

CERTIFICATIONS

A l'issus du parcours (10 modules), les candidats pourrons passer le jury de certifcation ATLAS :
Concevoir et implémenter une solution d'IA