CONCEPTEUR DEVELOPPEUR EN SCIENCE DES DONNEES

REFERENTIEL D'ACTIVITES	REFERENTIEL DE COMPETENCES	REFERENTIEL D'EVALUATION	
décrit les situations de travail et les activités exercées, les métiers ou emplois	*identifie les compétences et les connaissances, y compris transversales, qui découlent du référentiel*	*définit les critères et les modalités d'évaluation des acquis*	
	Bloc n°1 - Construction et alimentation d'une infrastructure de gestion de données	**MODALITÉS D'ÉVALUATION**	**CRITÈRES D'ÉVALUATION**
A1. Construction d'une infrastructure de gestion de données adaptée à l'organisation	C1.1 - Concevoir une architecture de données robuste et adaptée en créant des lacs de données (*Data Lake* en anglais) et des entrepôts de données (*Data Warehouse* en anglais) afin de répondre aux besoins de stockage, d'utilisation, de sécurité et de protection de l'organisation définie par un cahier des charges C1.2 - Intégrer la dimension de stockage et de calcul distribuée à l'infrastructure de données via l'utilisation d'outils comme *Spark* ou *AWS Redshift* afin de l'adapter à des besoins de gestion de données massives (*Big Data* en anglais)	**Type d'évaluation :** Une étude de cas sur des données réelles **Thème d'évaluation :** Construction d'une infrastructure *Cloud* accueillant des données Big Data (collecte de données web, intégration des données dans un *Data Lake* , nettoyage et chargement des données dans une base de données type AWS Redshift par traitement parallélisé si nécessaire via la construction d'un processus ETL) **Contexte:** En centre de formation ou à distance, étalé sur 10 heures. **Livrable :** Une étude de 1 page décrivant schématiquement l'infrastructure conceptualisée et le code source permettant de construire l'infrastructure, évalués par le jury de certification	L'infrastructure de stockage, collecte et mise à disposition des données proposée répond aux besoins définis dans le cahier des charges et elle est efficace : - Simplicité du schéma d'infrastructure proposé - Capacité de stockage du schéma d'infrastructure proposé (peut recevoir des données *Big Data*) - Optimisation des coûts de construction du schéma d'infrastructure proposé - Qualité des données extraites depuis le web vers le *Data lake* - Accessibilité des données disponibles dans le *Data Warehouse* - Robustesse et efficacité du processus ETL construit - Conformité du processus de collection de la donnée avec les normes de protection des données utilisateur définies dans le RGPD
A2. Collecte de données	C1.3 - Collecter des données provenant de différentes sources (Web, Logiciels internes de type *Sage* / *Excel* ou externes de type *Google Analytics*) via des librairies de programmation de type *Scrapy* ou *Beautifulsoup* dans le respect des normes de protection des données utilisateurs définies dans le RGPD pour alimenter le Data Lake afin d'affiner le résultat d'analyses futures.		
A3. Gestion d'entrepôts de données (*Data Warehouse*)	C1.4 - Nettoyer et organiser les données dans l'entrepôt de données (*Data Warehouse* en anglais) en écrivant des processus d'extraction, transformation et chargements (ETL en anglais) afin de rendre ces données disponibles et compréhensibles pour les autres équipes métiers.		
	Bloc n°2 - Analyse exploratoire, descriptive et inférentielle de données		
A4. Analyse exploratoire de données	C2.1 - Traiter des bases de données grâce à des analyses statistiques descriptives et inférentielles via des librairies de programmation comme *Numpy* ou *Pandas* , pour les organiser et les nettoyer afin de les normaliser par rapport à la population étudiée. C2.2 - Effectuer des analyses univariées et multivariées sur des bases de données structurées afin de préciser des relations entre plusieurs variables et d'établir des liens statistiques entre elles. C2.3 - Optimiser les analyses statistiques grâce au traitement parallélisé via l'utilisation d'outils comme *Spark* pour accélérer le temps de calcul d'un ordinateur afin de pouvoir analyser des volumes de données massifs (*Big Data*)	**Type d'évaluation :** Deux études de cas sur des données réelles **Thème d'évaluation :** - Gestion de valeurs manquantes et aberrantes d'une base de données non-massives puis analyse pour déterminer et présenter des tendances par le biais de graphiques. - Analyse d'une base de données massives déstructurées (Utilisation de *Spark*) adaptée à une problématique définie. **Contexte :** En centre de formation ou à distance, étalé sur 20 heures. **Livrable :** Deux codes sources décrivant l'analyse de chacune des bases de données, incluant la construction de graphiques, évalués par le jury de certification	Le jeu de données choisi est capable de répondre à la problématique fixée : - Pertinence des éléments et de la méthodologie de nettoyage de données (gestion des valeurs manquantes ou incohérentes, etc.) - Clarté de la base de données - Disponibilité des données pour l'analyse Les analyses définies sont efficaces pour répondre à la problématique fixée : - Pertinence du choix des analyses effectuées (Analyses univariées de moyennes & variances effectuées, détection d'anomalies par analyse de distributions effectuée, analyse de corrélations (Matrice de Pearson) effectuée, etc.) - Efficacité des analyses effectuées - Efficacité du traitement parallélisé appliqué Les recommandations qui ressortent des résultats de ces analyses sont pertinentes : - Clarté et simplicité des graphiques construits et présentés. - Clarté de l'exposé
A5. Visualisation et présentation de données	C2.4 - Présenter le résultat d'une analyse statistique de données structurées, massives ou non, grâce à des librairies de programmation comme *Plotly* ou *Matplotlib* pour synthétiser ce résultat devant un public profane afin de faciliter la prise de décisions et appuyer leurs déclinaisons opérationnelles.		
	Bloc n°3 - Analyse prédictive de données structurées par l'intelligence artificielle		
A6. Mise en place d'un algorithme d'apprentissage automatique	C3.1 - Traiter des données structurées en créant un pipeline de traitement grâce à des librairies de programmation comme *Scikit-Learn* pour encoder, normaliser et découper des données afin de les rendre interprétables par un algorithme d'apprentissage automatique (*Machine Learning* en anglais) C3.2 - Effectuer des analyses prédictives sur un jeu de données structurées grâce à des algorithmes d'apprentissage automatique supervisés adaptés afin d'automatiser des tâches liées aux résultats des prédictions de ces algorithmes	**Type d'évaluation :** trois études de cas pratiques tirées de cas réels **Thème d'évaluation :** - Optimisation des processus marketing de qualification de prospect par le biais d'algorithmes d'apprentissage supervisés - Optimisation d'algorithmes d'apprentissage automatique supervisé sur des bases de données déséquilibrées - Localisation de zones de densité géographique par l'élaboration d'algorithmes d'apprentissage automatique non-supervisé **Contexte :** En centre de formation ou à distance, étalé sur 30 heures. **Livrable :** Trois codes sources incluant la conception et l'optimisation de trois algorithmes adaptés à la problématique ainsi que des recommandations sur les prédictions obtenues, évalués par le jury de certification	Les process d'analyses prédictives définis sont efficaces : - Pertinence du choix du type d'algorithme utilisé (apprentissage automatique supervisé ou non-supervisé) - Conception d'un pipeline de préparation des données pour entrainement d'algorithme - Qualité de l'optimisation du regroupement (méthode de choix de l'indicateur K pour *K-Moyenne* ou Epsilon pour *DBScan*) - Propreté du code (respect des normes *PEP8* en Python) - Performance de l'algorithme programmé (déterminé par l'analyse de *R2* pour de la régression, *F1_Score* pour de la classification, utilisation de la *K-Fold Crossvalidation* et tests de vérification de sur-entrainement et sous-entrainement) Les performances des algorithmes de Machine Learning sont optimisées : - Pertinence du choix des critères d'évaluation de la performance d'un algorithme (choix d'un *F1_Score* ou de sensibilité pour de la classification en fonction du type de problématique, par exemple) - Indicateurs mesurés meilleurs que dans la version précédente (comparaison du nouveau modèle par rapport à ce qu'il y a déjà en place) Les analyses prédictives sont justes et les recommandations formulées sont pertinentes.
A7. Segmentation et réduction de base de données	C3.3 - Elaborer un algorithme d'apprentissage automatique non-supervisé pour segmenter une base de données en différents groupes homogènes ou réduire la dimension de cette dernière afin de pouvoir comprendre des observations de manière granulaire et de permettre leur visualisation.		
A8. Optimisation des performances des algorithmes d'apprentissage automatique	C3.4 - Evaluer la performance prédictive des algorithmes d'apprentissage automatique en déterminant l'influence des différentes variables pour pouvoir l'améliorer afin de démontrer son utilité aux directions métiers, par rapport aux processus déjà établis dans l'organisation		
	Bloc n°4 - Analyse prédictive de données non-structurées par l'intelligence artificielle		
A9. Mise en place d'un apprentissage automatique profond	C4.1 - Traiter des données non-structurées (image, texte, audio) par la création de fonction de traitements via l'utilisation de librairies de programmation comme *TensorFlow* ou *Numpy* pour les transformer en matrices afin de les rendre interprétables par un algorithme d'apprentissage automatique profond (*Deep learning* en anglais) C4.2 - Elaborer des réseaux de neurones adaptés (classiques, convolutifs ou recursifs) en superposant des couches neuronales via des librairies de programmation comme *TensorFlow* pour analyser des données non-structurées afin de détecter des signaux sur ces dernières C4.3 - Créer un algorithme robuste et précis en configurant un réseau de neurones pré-entrainé profond afin de répondre à des problématiques de prédiction sur des volumes de données massifs C4.4 - Créer des données non-structurées en élaborant des réseaux de neurones adverses afin de construire de nouvelles bases d'entrainement pour des applications d'intelligence artificielle	**Type d'évaluation :** une étude de cas pratique sur des données non-structurées **Thème d'évaluation :** Analyse de sentiment, par l'élaboration d'un algorithme permettant de déterminer le sentiment d'un utilisateur à l'égard d'un produit (avec possibilité de créer de la nouvelle donnée pour agrémenter la base). **Contexte :** En centre de formation ou à distance, étalé sur 20 heures **Livrable :** Un code source incluant la conception de l'algorithme et les métriques de performances sur des données de validation, évalué par le jury de certification	Les process d'analyses prédictives par des algorithmes d'apprentissage automatique profond définis sont efficaces et répondent à la problématique : - Qualité de préparation de la donnée (création de tenseurs, utilisation de techniques d'augmentation de données) - Pertinence du choix du type de réseau de neurones (classiques, convolutifs ou récursifs) - Efficacité des réseaux de neurones construits (vérifiée par des tests de généralisation des performances, sur des données de test et de validation) - Clarté de la détection des signaux provenant des données non-structurées - Propreté du code (respect des normes *PEP8* en *Python*) - Pertinence du choix des critères d'évaluation de la performance d'un algorithme (élaboration d'une fonction de coût - *Cross Entropy* , *Mean Squared Error*) - Qualité des données non-structurées générées
A10. Optimisation des performances des algorithmes d'apprentissage automatique profond pour industrialisation	C4.5 - Evaluer la performance d'un algorithme d'apprentissage automatique profond en évaluant des indicateurs sur des données d'entrainement et de validation afin d'industrialiser son utilisation		
	Bloc n°5 - Industrialisation d'un algorithme d'apprentissage automatique et automatisation des processus de décision		
A11. Industrialisation d'algorithmes d'apprentissage automatique	C5.1 - Standardiser la construction et l'environnement informatique d'un algorithme d'apprentissage automatique grâce des outils de production comme *MLflow* et *Docker* afin de faciliter la mise en production de projets d'intelligence artificielle sur tous types de plateformes C5.2 - Créer une interface de programmation applicative grâce à des outil comme *AWS sagemaker* afin de donner un accès à échelle aux prédictions des algorithmes d'apprentissage automatique à l'ensemble des équipes métiers concernées	**Type d'évaluation :** Etude de cas pratique sur le déploiement d'un algorithme d'apprentissage automatique **Thème d'évaluation :** *Web dashboard* , construction et mise en production d'une application web d'intelligence artificielle **Contexte :** En centre de formation ou à distance, étalé sur 10 heures. **Livrable :** Un code source contenant la création de l'environnement standardisé, le déploiement de l'algorithme et l'application web ainsi qu'un lien URL vers l'application déployée, évalué par le jury de certification	L'utilisation des algorithmes d'apprentissage automatique est automatisée et accessible par une interface web : - Accessibilité du processus de construction d'algorithme d'apprentissage automatique dans un environnement (Containerisation via *Docker* et normalisation via *MLFlow*) - Qualité des données retournées par l'interface de programmation applicative (API) incorporant les prédictions de l'algorithme mise en place - Propreté du code (respect des normes *PEP8* en Python) - Praticité de l'interface web incluant l'utilisation de l'interface de programmation applicative - Utilité de l'interface web incluant l'utilisation de l'interface de programmation applicative - Accessibilité de l'interface web incluant l'utilisation de l'interface de programmation applicative
A12. Production d'applications d'intelligence artificielle utilisables par toutes les équipes métier	C5.3 - Déployer une application web intégrant des algorithmes de statistiques prédictives (*Machine Learning* et *Deep Learning*) grâce à des outils comme *Flask* , *Heroku* ou *AWS sagemaker* pour les rendre utilisables par l'ensemble des équipes métiers afin d'automatiser leurs processus de décision		
	Bloc n°6 - Direction de projets de gestion de données		
A13. Définition d'une modélisation statistique qui permette de répondre aux problématiques des directions métiers	C6.1 - Traduire les enjeux métiers en problématiques mathématiques/data grâce à une compréhension des besoins propres à chaque projet data afin de pouvoir répondre aux objectifs de l'organisation C6.2 - Maîtriser les technologies les plus récentes et adaptées du marché grâce à de la veille technologique et de la pratique constante pour développer une expertise afin d'être à même de proposer aux directions métiers les solutions les plus adaptées actuellement à une problématique et l'amélioration constante des process de gestion de données déjà en place C6.3 - Définir un cahier des charges, un retroplanning et un budget afin de défendre et détailler aux directions métier un projet data répondant aux besoins de l'organisation	**Type d'évaluation :** projet data conçu de A à Z. **Thème d'évaluation :** libre. Les apprenants peuvent préparer le projet data de leur choix. Celui-ci peut être personnel, développé par le candidat dans le cadre de son activité professionnelle, ou défini par une entreprise partenaire. **Contexte :** En centre de formation (possibilité de l'effectuer à distance si les conditions sanitaires l'imposent uniquement), sur 50 heures (2 semaines). **Livrable :** Le code source correspondant au projet data développé et une soutenance orale de 10 minutes suivie de 5 minutes de questions et éventuellement de 5 minutes d'entretien.	Le candidat a démontré une maîtrise totale de tous les aspects du projet développé et une expertise sur l'ensemble des technologies et processus utilisés : - Pertinence de la traduction d'une problématique métier en problématique data - Réalisme du jalonnage du projet - Pertinence du choix d'un jeu de données permettant de répondre à la problématique - Pertinence du choix des technologies utilisées pour mener à bien ce projet - Conformité de l'ensemble des processus définis dans le projet avec les normes de protection des données utilisateur définies dans le RGPD - Efficacité de la construction d'un jeu de données permettant de répondre à la problématique - Pertinence et compréhension des variables et de l'algorithme choisi - Pertinence du choix des indicateurs de performance de ce modèle (R2, F1_Score, Précision,etc) - Performance du modèle d'apprentissage automatique ou d'apprentissage automatique profond choisi (amélioration des indicateurs définis par rapport au modèle en place) - Efficacité de ce modèle pour répondre à la problématique définie - Accessibilité par un public profane de la présentation des processus et des technologies utilisées - Respect des étapes de réalisation du projet L'exposé est clair, pertinent et convaincant : - Pertinence des recommandations - Clarté et accessibilité de l'exposé
A14. Construction d'un système de gestion et de suivi de projet d'analyse et de gestion de données	C6.4 - Gérer un projet d'analyse et de gestion de données (analyse statistique descriptive, Machine Learning, Deep Learning, Big Data ou non) grâce à l'élaboration d'indicateurs adaptés et de tableaux de bords, afin de faire le suivi et le bilan de l'action, ainsi que de la déclinaison opérationnelle de ses résultats, le tout dans le respect des normes de protection des données utilisateurs définies dans le RGPD		
A15. Direction de projets de gestion de données	C6.5 - Transmettre aux directions-métiers le process d'extraction d'informations et d'analyse de données en le vulgarisant afin de soutenir la mise en place d'une stratégie et d'actions futures. C6.6 - Diriger un projet de gestion de données, allant de sa conception à la mise en place de solutions, afin de le mener jusqu'à son terme, d'être la personne clé disposant de toutes les informations sur le projet à tout moment, et d'accompagner d'autres services de l'organisation dans l'ensemble des activités relatives à celui-ci		