



## REFERENTIEL D'ÉVALUATION ET DE COMPETENCES DE LA CERTIFICATION CONCEVOIR ET GERER DES SOLUTIONS DE MACHINE LEARNING POUR LE CLOUD AWS

*(Titre en anglais : AWS Certified Machine Learning - Specialty)*

### MODALITES D'ÉVALUATION

Pour obtenir cette certification, il est proposé au candidat d'évaluer leurs compétences à travers un (1) examen en ligne, supervisé par l'organisme Pearson VUE ou PSI Services, délivrés dans un centre d'examen accrédité (ou via de la surveillance à distance).

L'examen dure environ trois heures (3h00) – livres fermés - et comprend une variété de questions appelant différentes formes de réponse\* : Questions à choix multiples, Questions à réponses multiples.

\*Détaillées à cette URL : [https://d1.awsstatic.com/training-and-certification/docs-ml/AWS-Certified-Machine-Learning-Specialty\\_Exam-Guide.pdf](https://d1.awsstatic.com/training-and-certification/docs-ml/AWS-Certified-Machine-Learning-Specialty_Exam-Guide.pdf)

Le seuil de réussite est fixé à environ 75% de bonnes réponses, qui correspond à un score de passage de 750 points (score à l'échelle). Le pourcentage réel varie d'un examen à l'autre. La note de passage est basée sur l'apport d'experts en la matière, le niveau de compétence requis pour être considéré comme compétent dans le domaine du contenu, et la difficulté des questions livrées pendant l'examen. Les pourcentages dans le tableau des compétences évaluées indiquent le poids relatif de chaque sujet principal de l'examen. Plus le pourcentage est élevé, plus les candidats devront répondre à des questions sur cette zone de contenu. La liste des tâches évaluées n'est pas exhaustive et peut couvrir d'autres tâches dans le cadre des compétences évaluées.

L'examen compte en tout 65 questions mais seulement 50 sont évaluées. L'examen comprend 15 questions non notées qui n'affectent pas le score du candidat. AWS collecte des informations sur la performance des candidats sur ces questions non notées afin d'évaluer ces questions en vue d'une utilisation future. Ces questions non notées ne sont pas identifiées lors de l'examen. Dans le résumé des compétences ci-dessous le pourcentage représente uniquement le contenu scoré. Il en est de même pour le nombre de questions par domaine évalué.

Le contenu des tests est réévalué régulièrement par les équipes Amazon Web Services pour refléter les dernières évolutions des services et de la plate-forme AWS.



RESUME DES ACTIVITES PRINCIPALES :

Compétences mobilisables évaluées		Nature des tâches évaluées permettant de valider la compétence	Evaluation		
			% de l'évaluation globale	Modalités d'évaluation	Critères
<b>Effectuer l'ingénierie des données pour préparation avant exploitation de ces dernières</b>					
	<b>Créer des référentiels de données pour le Machine Learning.</b>	<ul style="list-style-type: none"> <li>▶ Créer une instance (Jupyter) Notebook et configurer le rôle IAM attribué.</li> <li>▶ Utiliser les SageMaker Jupyter notebooks pour charger des sets de données depuis Scikit-learn.</li> <li>▶ Identifier les sources de données (par exemple, le contenu et l'emplacement, les sources principales telles que les données utilisateur)</li> <li>▶ Déterminer les supports de stockage (p. ex. DB, Data Lake, S3, EFS, EBS)</li> <li>▶ Gérer et décider de la meilleure approche en cas de données manquantes : ne rien faire, supprimer le set entier de données, prendre le mode/la moyenne/la médiane en donnée de remplacement, la valeur la plus fréquente, utiliser une imputation basée sur le modèle de Machine Learning (K-nearest neighbors, régression linéaire, deep learning), extrapoler, forward ou backward filling.</li> </ul>	<b>20% de l'évaluation globale de l'examen</b>	Examen en ligne avec une variété de questions (cf. détail plus haut) ** <b>Environ 36 minutes</b> (pour 9 à 11 questions évaluées) sont consacrées à cette compétence	Examen compensatoire, le taux de bonnes réponses doit être au <b>global de 75% minimum</b>



	<b>Identifier et mettre en œuvre une solution d'ingestion de données.</b>	<ul style="list-style-type: none"><li>▶ Comprendre la « malédiction » de la dimensionnalité et le traiter : gérer des ratios importants d'input variables / observation et choisir entre les sets de data multidimensionnels et les sets de data a 2 ou 3 dimensions.</li><li>▶ Utiliser des filtres pour écarter les données redondantes ou non pertinentes.</li><li>▶ Utiliser l'algorithme d'analyse des composantes principales (PCA) pour créer des nouveaux inputs orthogonaux en combinant linéairement les inputs originels.</li><li>▶ Utiliser le PCA pour aider à la visualisation des données et à l'accélération du Machine Learning.</li></ul>			
	<b>Identifier et mettre en œuvre une solution de transformation des données.</b>	<ul style="list-style-type: none"><li>▶ Utiliser selon le cas les différentes approches d'encodage des données : encodage binaire, encodage par label/étiquette, encodage par transformation de valeurs nominales en valeurs numériques avec création de nouvelles colonnes.</li><li>▶ Transformer les données pour permettre une meilleure analyse par les solutions de Machine Learning via normalisation et standardisation.</li><li>▶ Utiliser les différentes approches de répartition des données (binning) : répartition par catégorie, par quantile, par valeur numérique.</li><li>▶ Transformer du texte en données pour une meilleure analyse par les solutions de Machine Learning via création de sac de mots (« bag-of-words »), N-Gram, fréquence d'apparition de termes dans des documents.</li><li>▶ Utiliser selon le cas les différents services de migration de données AWS et savoir dans quels cas les utiliser : Amazon Data Pipeline, AWS Database Migration Service (DMS), AWS Glue, Amazon SageMaker, Amazon Athena.</li><li>▶ Transformer le transit des données (ETL : Glue, EMR, AWS Batch)</li><li>▶ Gérer les données spécifiques à ML à l'aide de la réduction de la carte (Hadoop, Spark, Hive)</li></ul>			



Effectuer une analyse exploratoire des données pour préparer leur traitement				
	<p><b>Assainir et préparer les données en vue de la modélisation.</b></p>	<ul style="list-style-type: none"> <li>▶ Comprendre les termes de Data Producer, Data Consumer, Shard et Data Stream.</li> <li>▶ Identifier et gérer les données manquantes, les données corrompues, les mots d'arrêt, etc.</li> <li>▶ Formater, normaliser, augmenter et mettre à l'échelle des données</li> <li>▶ Utiliser Kinesis Data Streams pour enregistrer en temps réel des données provenant de multiples ressources et les analyser en direct.</li> <li>▶ Utiliser Amazon Kinesis Producer Library et Amazon Kinesis Agent pour regrouper les données.</li> <li>▶ Utiliser Kinesis Data Firehose pour capturer, transformer et télécharger les données dans S3, Redshift, Elasticsearch et Splunk ainsi que pour compresser, transformer et crypter les données.</li> <li>▶ Suivre les métriques sur les données (quantité de données envoyée à destination, temps de l'opération, pourcentage de succès, ...).</li> <li>▶ Au niveau des données étiquetées / taguées, reconnaître quand vous disposez de suffisamment de données étiquetées et identificate0072 des stratégies d'atténuation [Outils d'étiquetage des données (Mechanical Turk, main-d'œuvre manuelle)].</li> </ul>	<p><b>24% de l'évaluation globale de l'examen</b></p>	<p>Examen en ligne avec une variété de questions (cf. détail plus haut) **</p> <p><b>Environ 43 minutes</b> (pour 11 à 13 questions évaluées) sont consacrées à cette compétence</p>
	<p><b>Effectuer l'ingénierie des données.</b></p>	<ul style="list-style-type: none"> <li>▶ Créer des données : identifier les données dans le set de données qui sont pertinentes pour le problème attribué.</li> <li>▶ Transformer des données : gérer le remplacement des données manquantes ou invalides.</li> <li>▶ Extraire des données : créer de nouvelles données a partie d'autres existants déjà principalement dans le but de réduire la dimensionnalité/complexité de ces données.</li> <li>▶ Sélectionner des données : filtrer les données non pertinentes ou redondantes dans le set de données (binning, tokenisation, valeurs aberrantes, entités synthétiques, 1</li> </ul>		



		encodage à chaud, réduction de la dimensionnalité des données).			
	<b>Analyser et visualiser les données pour le Machine Learning.</b>	<ul style="list-style-type: none"><li>▶ Etudier les données avant de choisir l’algorithme de Machine Learning : identifier des modèles, identifier les données corrompues, identifier les valeurs extrêmes (outliers), identifier les déséquilibres dans les données, explorer les liens importants entre les données et leur force.</li><li>▶ Démontrer les informations suivantes avec les outils de Business Intelligence (BI tools) : indicateurs clés de performances (KPIs), relations, comparaisons, distributions, compositions.</li><li>▶ Prouver des relations entre deux ou plusieurs variables avec des diagrammes de dispersion et des graphiques à bulles.</li><li>▶ Comparer des variables sur le temps avec des diagrammes en bâtons, des graphiques à colonnes, des graphiques linéaires ou des tableaux.</li><li>▶ Montrer la distribution des données avec des histogrammes à colonnes ou des diagrammes de dispersion.</li><li>▶ Montrer les éléments composant un ensemble de données statiques ou variables avec des diagrammes circulaires, des tree maps et des graphiques empilés (stacked charts).</li><li>▶ Maîtriser l’utilisation des différents outils BI : Amazon Quicksight (Machine Learning insights), TensorFlow avec TensorBoard, Tableau.</li></ul>			



Créer les modélisations pour la mise en place du Machine Learning					
	<b>Traiter les problèmes d'entreprise comme des problèmes d'apprentissage automatique.</b>	<ul style="list-style-type: none"><li>▶ Définir les critères établissant une réussite du projet de Machine Learning.</li><li>▶ Etablir des métriques de performance observables et quantifiables pour le projet tels que la précision, la latence de prédiction ou la minimisation de la valeur de l'inventaire.</li><li>▶ Formuler la question de Machine Learning en terme d'inputs, d'outputs désirés et de métriques de performance à optimiser.</li><li>▶ Estimer si une approche Machine Learning est réalisable et si cela est l'approche appropriée.</li><li>▶ Créer un objectif de data sourcing et une stratégie pour y parvenir.</li><li>▶ Commencer avec un modèle de Machine Learning simple à interpréter et qui rend le debugging plus facilement gérable.</li></ul>	<b>36% de l'évaluation globale de l'examen</b>	Examen en ligne avec une variété de questions (cf. détail plus haut) ** <b>Environ 65 minutes</b> (pour 17 à 19 questions évaluées) sont consacrées à cette compétence	Examen compensatoire, le taux de bonnes réponses doit être au <b>global de 75% minimum</b>
	<b>Sélectionner le(s) modèle(s) approprié(s) pour un problème de Machine Learning donné.</b>	<ul style="list-style-type: none"><li>▶ Utiliser selon le cas un des 3 types de modèles de Machine Learning : classification binaire (problème de classification a deux sorties), classification multiclasse (problème de classification a plus de 2 sorties) et régression (problème de prédiction de valeur numérique).</li><li>▶ Utiliser selon le cas un des 3 types d'algorithmes de Machine Learning : supervisé, non-supervisé, renforcé. Savoir quelle approche utiliser en fonction des situations : données discrètes ou continues variables cibles ou non.</li></ul>			



	<b>Concevoir les modèles de Machine Learning.</b>	<ul style="list-style-type: none"><li>▶ Créer la tâche d'entraînement dans SageMaker pour entraîner le modèle de Machine Learning et sauvegarder les résultats dans S3.</li><li>▶ Utiliser les instances et les algorithmes intégrés à SageMaker pour l'entraînement des modèles de Machine Learning.</li><li>▶ Suivre les différentes étapes pour l'entraînement : regrouper des données dans le set de données, randomiser le set de données, diviser le set de données entre données test et données d'entraînement choisir le meilleur algorithme, gérer la capacité de computation, créer une instance, définir les hyperparamètres du modèle, entraîner le modèle.</li></ul>			
	<b>Effectuer l'optimisation des hyperparamètres.</b>	<ul style="list-style-type: none"><li>▶ Différencier hyperparamètres de modèle et hyperparamètres d'algorithme. Choisir l'approche appropriée : aléatoire ou bayésienne.</li><li>▶ Effectuer les tâches de réglages sur les hyperparamètres en effectuant beaucoup de tâches d'entraînement sur le set de données avec des hyperparamètres différents. Choisir les valeurs d'hyperparamètres pour lesquelles le modèle fonctionne le mieux.</li><li>▶ Utiliser le réglage automatique des hyperparamètres grâce aux algorithmes intégrés et aux frameworks de Machine Learning disponibles.</li><li>▶ Mettre à l'échelle les hyperparamètres en choisissant l'échelle de réglage : automatique, linéaire, logarithmique, logarithmique inversée.</li></ul>			
	<b>Évaluer les modèles de Machine Learning.</b>	<ul style="list-style-type: none"><li>▶ Créer la tâche d'entraînement dans SageMaker pour entraîner le modèle de Machine Learning et sauvegarder les résultats dans S3.</li><li>▶ Utiliser SageMaker Model Monitor pour monitorer en continu la qualité du modèle de Machine Learning : créer des alertes en cas de déviation.</li></ul>			



<b>Mettre en œuvre et opérer les solutions de Machine Learning</b>					
	<b>Créer des solutions d'apprentissage automatique qui sont performantes, disponibles, évolutives, résilientes et tolérantes aux pannes.</b>	<ul style="list-style-type: none"><li>▶ Déployer rapidement des modèles de grande qualité à l'échelle avec SageMaker.</li><li>▶ Optimiser la performance des solutions de Machine Learning avec SageMaker Autopilot (créer et entraîner des modèles avec une visibilité et un contrôle total).</li><li>▶ Organiser vos artefacts, suivre les métriques et évaluer les training runs en utilisant SageMaker Experiments.</li><li>▶ Gérer les itérations de modèles en capturant les paramètres d'entrée, les configurations et les résultats.</li><li>▶ Analyser, identifier et créer des alertes pour les problèmes de Machine Learning en temps réel avec SageMaker Debugger.</li><li>▶ Détecter et corriger la dérive conceptuelle lorsque les patrons utilisés pour former le modèle de Machine Learning ont changé au fil du temps avec SageMaker Monitor.</li></ul>	<b>20% de l'évaluation globale de l'examen</b>	Examen en ligne avec une variété de questions (cf. détail plus haut) ** <b>Environ 36 minutes</b> (pour 9 à 11 questions évaluées) sont consacrées à cette compétence	Examen compensatoire, le taux de bonnes réponses doit être au <b>global de 75% minimum</b>
	<b>Recommander et mettre en œuvre les services et fonctionnalités de Machine Learning appropriés pour un problème donné.</b>	<ul style="list-style-type: none"><li>▶ Choisir parmi différents services AWS lesquels implémenter pour choisir la solution de Machine Learning correspondant à la problématique business.</li><li>▶ Utiliser Lex (chatbots) pour construire des interfaces de conversation texte/voix dans les applications.</li><li>▶ Créer un service d'analyse d'images et de vidéos avec Rekognition</li><li>▶ Utiliser Transcribe pour créer des services de speech-to-text (dictée de textes)</li><li>▶ Utiliser Translate pour ajouter un service de traduction utilisant le deep learning aux applications</li><li>▶ Ajouter un service de text-to-speech aux applications avec Polly.</li><li>▶ Mettre au point un service d'analyse de texte avec Comprehend pour trouver les relations et les points importants dans un texte.</li></ul>			



	<b>Appliquer les pratiques de sécurité AWS de base aux solutions de Machine Learning.</b>	<ul style="list-style-type: none"><li>▶ Contrôler l'accès aux notebooks Jupyter et aux modèles de Machine Learning stockés avec IAM.</li><li>▶ Relier les ressources SageMaker à un VPC pour leur appliquer les contrôles de sécurité du réseau.</li><li>▶ Crypter les données au repos et en transit entre les sets de données stockées dans S3 et les notebooks Jupyter.</li><li>▶ Utiliser les cycles de vie pour renforcer la sécurité des OS utilisés par SageMaker ou installer des agents de sécurité.</li><li>▶ Utiliser CloudWatch et CloudTrail pour surveiller les interactions API.</li></ul>			
	<b>Déployer et opérationnaliser des solutions de Machine Learning.</b>	<ul style="list-style-type: none"><li>▶ Suivre les trois étapes dans le cycle de vie des données : création de la donnée, production du modèle, exécution du modèle.</li><li>▶ Monitorer le modèle de production pour détecter des déviations dans la qualité de la donnée fournie par rapport à des bases modèles. Analyser et ré-entraîner le modèle de Machine Learning au besoin.</li></ul>			