

Data scientist

RNCP39590 · RNCP · Publiée

L'essentiel

Niveau de qualification	Niveau 7
Code(s) NSF	326
Formacode(s)	31026, 31052, 31067
Code(s) ROME	M1805
Type d'enregistrement	Enregistrement sur demande
Date d'enregistrement	01/10/2024
Date d'échéance	01/10/2026

Certificateur(s)

ASSOCIATION DE GESTION ICAM STRASBOURG-EUROPE · SIRET : 51773337400022

Partenaires habilités

1 organisme habilité, d'après la liste publiée par France compétences : 1 à former et organiser l'évaluation.

Organisme	SIRET	Habilitation
UNIVERSITE DE LORRAINE	13001550600012	Former et organiser l'évaluation

Résumé de la certification

OBJECTIFS ET CONTEXTE

La révolution numérique a fait rentrer nombre d'entreprises dans l'ère du big data. Le big data joue un rôle de plus en plus stratégique pour les entreprises : l'exploitation des données permettant notamment d'assurer un meilleur suivi de leur performance, d'analyser les comportements des consommateurs, d'identifier des opportunités de marché, etc. Les entreprises ont besoin de mobiliser des compétences spécifiques pour tirer profit au mieux des données. Ces compétences relèvent de trois champs distincts : celui de la collecte et du stockage des données, celui du traitement et de la modélisation des données, et celui de la protection des données.

La certification concerne l'exploitation des données numériques de l'entreprise à des fins de décisions stratégiques, managériales, marketing, techniques. Elle vise à qualifier des Data Scientists disposant des compétences expertes et transverses nécessaires.

ACTIVITÉS VISÉES

Analyse des besoins métiers de la structure en matière d'accompagnement décisionnel pour l'établissement d'une stratégie globale de transformation de la donnée prenant en compte les contraintes sociales, économiques et environnementales

Pilotage de la remontée et l'intégrité des données en collaboration avec les équipes techniques en vue de leur exploitation

Exploration et analyse des ensembles de données pour fournir des connaissances ("insights") exploitables et soutenir la prise de décision

Développement des modèles par l'utilisation des techniques avancées de statistiques et d'apprentissage automatique pour fournir des informations et accompagner le processus de décision

Pilotage des projets en communiquant avec les parties prenantes et en guidant les membres de l'équipe pour atteindre des objectifs spécifiques

COMPÉTENCES ATTESTÉES

Traduire de manière analytique le besoin métier en rédigeant un cahier des charges en collaboration avec les parties prenantes pour fixer les objectifs attendus

Proposer une conception technique et technologique grâce à un état de l'art prenant en compte l'ensemble des contraintes (temporelles, budgétaires, techniques, managériales, juridiques, etc.) afin de répondre au cahier des charges

Cartographier les flux de données et leur structure par des modèles logiques afin de les préparer aux processus d'intégration

Superviser le développement de systèmes d'acquisition, de transfert et de stockage des données en accord avec les besoins techniques et juridiques (RGPD) en collaborant avec des équipes techniques de façon transparente et inclusive pour pérenniser les flux de données

Assurer l'intégrité des données tout au long de leur cycle de vie par le développement de pipelines ETL (Extract, Transform, Load) afin de réduire les erreurs potentielles et d'assurer une gestion pérenne des flux de données

Analyser des ensembles de données par l'application de techniques statistiques afin de mettre en valeur des tendances, des corrélations et des modèles significatifs

Transmettre des informations issues des données transformées par le développement d'outils de visualisation tels que des tableaux de bords, des infographies et des KPI pour accompagner la prise de décision de l'entreprise

Déterminer une représentation des données à l'aide de techniques d'ingénierie de caractéristiques ("feature engineering") pour entraîner un modèle répondant à la problématique

Entraîner un modèle d'apprentissage supervisé ou non-supervisé à l'aide d'un ensemble de données représentées et en ajustant ses hyper-paramètres pour améliorer ses performances et sa capacité à répondre de manière généralisée à une problématique

Établir une stratégie pour la création d'un modèle d'apprentissage en concevant ou réutilisant des modèles pré-entraînés (utilisant le transfert d'apprentissage), et entraîner ces modèles dans le but de résoudre une problématique métier spécifique

Évaluer les modèles d'apprentissage établis en utilisant des techniques de validations croisées et des métriques de performances afin de déterminer le modèle le plus adapté à la problématique métier

Communiquer les résultats des modèles en détaillant les liens avec les objectifs et en faisant preuve d'éthique et de transparence sur leurs limites afin d'accompagner les processus décisionnels

Assurer une veille technologique par des stratégies de surveillance des tendances et des innovations dans le domaine des sciences de la donnée afin de développer des idées novatrices et accompagner la résolution de problèmes

Collaborer avec des équipes pluridisciplinaires en appliquant les outils du management dans le respect des valeurs éthiques et inclusives afin d'améliorer la performance individuelle et collective

Organiser les activités et les ressources par la mise en place d'une stratégie de gestion afin d'assurer la conduite du projet dans les critères coût-qualité-délais du cahier des charges

Blocs de compétences

RNCP39590BC01 — Analyser les besoins métiers de la structure en matière d'accompagnement décisionnel afin d'établir une stratégie globale de transformation de la donnée prenant en compte les contraintes sociales, économiques et environnementales

LISTE DE COMPÉTENCES

Traduire de manière analytique le besoin métier en rédigeant un cahier des charges en collaboration avec les parties prenantes pour fixer les objectifs attendus

Proposer une conception technique et technologique grâce à un état de l'art prenant en compte l'ensemble des contraintes (temporelles, budgétaires, techniques, managériales, juridiques, etc.) afin de répondre au cahier des charges

MODALITÉS D'ÉVALUATION

Présentation de la mission en entreprise - Partie définition du besoin :

Mission en situation réelle - Individuel. Le candidat doit présenter le cahier des charges issue de la problématique soumise par l'entreprise. Ce cahier des charges établit des objectifs chiffrés et les contraintes liées au projet (techniques, technologiques, budgétaires, juridiques, délais, politiques...). Au travers de sa présentation, le candidat démontre sa compréhension du besoin métier et du contexte dans lequel il s'inscrit, de ces interlocuteurs, ainsi que les objectifs fixés. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de

l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Partie Etat de l'Art : Mission en situation réelle - Individuel. Le candidat présente l'existant en entreprise et son historique, en matière d'acquisition, d'intégration et d'exploitation des données. Il identifie les caractéristiques en lien avec la problématique pour pouvoir y répondre. Si l'infrastructure initiale ne permet pas de répondre à la problématique, il propose des solutions techniques et/ou méthodologiques pour pallier à ce manque. Il justifie ses choix par un argumentaire technique mais aussi temporel, budgétaire, managérial, juridique, etc. Lors de sa soutenance, le candidat doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière les éléments du contexte, de la problématique et des éléments de justification. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Data Science réalisé en Centre - Partie choix de la solution technique :

En autonomie - Projet individuel. Face à une problématique imposée par l'évaluateur et des données fournies, le candidat doit présenter une liste de solutions potentielles, leurs caractéristiques spécifiques et leur capacité à répondre au besoin. Il doit exposer les limites de chacune des solutions et justifier le choix de l'une d'entre-elles par un argumentaire technique mais aussi temporel, budgétaire, managériale, juridique, etc. Lors de sa soutenance, le candidat doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière les éléments du contexte, de la problématique et des éléments de justification. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

RNCP39590BC02 — Piloter la remontée et l'intégrité des données en collaborant avec les équipes techniques en vue de leur exploitation

LISTE DE COMPÉTENCES

Cartographier les flux de données et leur structure par des modèles logiques afin de les préparer aux processus d'intégration

Superviser le développement de systèmes d'acquisition, de transfert et de stockage des données en accord avec les besoins techniques et juridiques (RGPD) en collaborant avec des équipes techniques de façon transparente et inclusive pour pérenniser les flux de données

Assurer l'intégrité des données tout au long de leur cycle de vie par le développement de pipelines ETL (Extract, Transform, Load) afin de réduire les erreurs potentielles et d'assurer une gestion pérenne des flux de données

MODALITÉS D'ÉVALUATION

Présentation de la mission en entreprise - Partie cartographie : Mission en situation réelle - Présentation individuelle. Le candidat présente l'ensemble des flux de données ainsi que leur structure, leurs sources, leurs caractéristiques, leur conformité aux normes RGPD, etc., dans la structure d'accueil à l'aide d'outils standards et des modèles logiques. Il s'appuie sur sa production pour identifier les éléments permettant de répondre à la problématique et, le cas échéant, justifie une transformation de ces données afin qu'elles puissent être exploitées dans ce cadre. Lors de sa soutenance, le candidat doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Intégration des données - Partie cartographie : En autonomie - En groupe de deux à trois candidats - Soutenances individuelles. Le groupe se voit proposer un jeu de données brutes (différents formats, structures...etc.) dans une problématique d'intégration de celles-ci. Ils doivent exposer les éléments permettant leur mise en relation et justifiant un jeu de données unifié et standardisé, à l'aide d'outils standards et de modèles logiques. Lors de sa soutenance, le candidat doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Partie supervision technique : Mission en situation réelle - Individuel. Le candidat, en collaboration avec des équipes technique, suit et participe à la mise en place des systèmes d'acquisition, de transfert et de stockage des données (data lake) afin de répondre à la problématique d'intégration, tout en assurant le respect des normes de sécurité et juridiques telles que les normes RGPD. Il s'appuie sur une expression écrite de ses besoins, des outils de gestion de projet et une posture professionnelle, ouverte et transparente pour

mener à bien le projet. Il supervise les tests de ces infrastructures et produit une synthèse en conséquence. Lors de sa soutenance, le candidat présente sa démarche, le lien avec la problématique, les différentes phases du projet, les résultats des tests, et expose la synthèse de ses travaux. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Partie ETL : Mission en situation réelle - Individuel. Le candidat, à partir d'un ensemble de données brutes (ex: data lake) fourni par la structure d'accueil, et s'appuyant sur une conception technique vis-à-vis d'une problématique donnée, met en place un pipeline ETL permettant d'extraire des données les caractéristiques répondant à la problématique et de les transformer en vue d'une exploitation. Il utilise pour cela des outils standards (Talend, Alteryx...etc.). Lors de sa soutenance, le candidat présente le pipeline en question avec les différentes étapes d'extraction et de transformation, et leurs justifications. Il présente le jeu de données en sortie. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Intégration des données - Partie ETL :

Mission en situation réelle - Groupe de deux à trois candidats - Soutenance individuelle. Le candidat, à partir d'un ensemble de données brutes (ex: data lake) qu'il a constitué, et s'appuyant sur une conception technique vis-à-vis d'une problématique donnée, met en place un pipeline ETL permettant d'extraire des données les caractéristiques répondant à la problématique et de les transformer en vue d'une exploitation. Il utilise pour cela des outils standards (Talend, Alteryx...etc.). Lors de sa soutenance, le candidat présente le pipeline en question avec les différentes étapes d'extraction et de transformation, et leurs justifications. Il présente le jeu de données en sortie. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet DataScience - Partie ETL : En autonomie - Individuel. Le candidat, à partir d'un ensemble de données brutes (ex: data lake) qu'il a constitué, et s'appuyant sur une conception technique vis-à-vis d'une problématique donnée, met en place un pipeline ETL permettant d'extraire des données les caractéristiques répondant à la problématique et de les transformer en vue d'une exploitation. Il utilise pour cela des outils standards (Talend, Alteryx...etc.). Lors de sa soutenance, le candidat présente le pipeline en question avec les différentes étapes d'extraction et de transformation, et leurs justifications. Il présente le jeu de données en sortie. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

RNCP39590BC03 — Explorer et analyser des ensembles de données afin de fournir des connaissances ("insights") exploitables et de soutenir la prise de décision

LISTE DE COMPÉTENCES

Analyser des ensembles de données par l'application de techniques statistiques afin de mettre en valeur des tendances, des corrélations et des modèles significatifs

Transmettre des informations issues des données transformées par le développement d'outils de visualisation tels que des tableaux de bords, des infographies et des KPI pour accompagner la prise de décision de l'entreprise

MODALITÉS D'ÉVALUATION

Présentation de la mission en entreprise - Partie analyse des données :

Mission en situation réelle - Individuel. Le candidat, à partir des données transformées, applique, pour mettre en valeur des informations en lien avec la problématique, un ensemble de techniques statistiques telles que : Matrices de corrélation,

ACP (Analyse par Composante Principale), Test ANOVA (Analyse de la Variance), Analyse univariée, Analyse multivariée. Il en présente les informations extraites et les analyses qui peuvent en découler. À travers ses travaux, il démontre sa maîtrise des différentes techniques et de leur contexte d'application, et est capable de présenter leurs limites. Lors de sa soutenance, le candidat présente et justifie l'ensemble des techniques qu'il a appliqué et les résultats obtenus. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en

lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Analyse des données : En autonomie - Groupe de deux à trois candidats - Soutenance individuelle. Le candidat, à partir des données transformées, applique, pour mettre en valeur des informations en lien avec la problématique, un ensemble de techniques statistiques telles que : Matrices de corrélation, ACP (Analyse par Composante Principale), Test ANOVA (Analyse de la Variance), Analyse univariée, Analyse multivariée. Il en présente les informations extraites et les analyses qui peuvent en découler. À travers ses travaux, il démontre sa maîtrise des différentes techniques et de leur contexte d'application, et est capable de présenter leurs limites. Lors de sa soutenance, le candidat présente et justifie l'ensemble des techniques qu'il a appliqué et les résultats obtenus. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Data Science - Partie analyse des données : En autonomie - Individuel. Le candidat, à partir des données transformées, applique, pour mettre en valeur des informations en lien avec la problématique, un ensemble de techniques statistiques telles que : Matrices de corrélation, ACP (Analyse par Composante Principale), Test ANOVA (Analyse de la Variance), Analyse univariée, Analyse multivariée. Il en présente les informations extraites et les analyses qui peuvent en découler. À travers ses travaux, il démontre sa maîtrise des différentes techniques et de leur contexte d'application, et est capable de présenter leurs limites. Lors de sa soutenance, le candidat présente et justifie l'ensemble des techniques qu'il a appliqué et les résultats obtenus. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Partie visualisation des données : Mission en situation réelle - Individuel. Le candidat à l'aide d'outils standards, doit concevoir et déployer des tableaux de bord permettant une visualisation d'informations, de KPI, afin de mettre en place une démarche de "data storytelling" (une interprétation vulgarisée) auprès des décideurs. Ces informations sont issues des données transformées et/ou des analyses qui en ont été faites. Les outils utilisés servent la création de rapports et de représentation de données (Power BI, Tableaux...etc.). L'ensemble des productions respecte les principes de l'accessibilité numérique pour favoriser l'inclusivité, conformément au référentiel RG2A (Référentiel Général d'Amélioration de l'Accessibilité). Lors de sa soutenance, le candidat présente ses travaux, les informations issues de ces tableaux de bord, et les exploite dans un data storytelling en lien avec la problématique de la mission. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Visualisation des données : En autonomie - Groupe de deux à trois candidats - Soutenance individuelle. Le candidat à l'aide d'outils standards, doit concevoir et déployer des tableaux de bord permettant une visualisation d'informations, de KPI, afin de mettre en place une démarche de "data storytelling" (une interprétation vulgarisée) auprès des décideurs. Ces informations sont fournies par l'encadrant du projet. Le contexte dont celles-ci sont issues, soient les données transformées et/ou des analyses, est précisé. Les outils utilisés servent la création de rapports et de représentation de données (Power BI, Tableaux...etc.). L'ensemble des productions respecte les principes de l'accessibilité numérique pour favoriser l'inclusivité, conformément au référentiel RG2A (Référentiel Général d'Amélioration de l'Accessibilité). Lors de sa soutenance, le candidat présente ses travaux, les informations issues de ces tableaux de bord, et les exploite dans un data storytelling en lien avec la problématique du projet. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Data Science - Partie visualisation des données : En autonomie - Individuel. Le candidat à l'aide d'outils standards, doit concevoir et déployer des tableaux de bord permettant une visualisation d'informations, de KPI, afin de mettre en place une démarche de "data storytelling" (une interprétation vulgarisée) auprès des décideurs. Ces informations sont issues des données transformées et/ou des analyses qui en ont été faites. Les outils utilisés servent la création de rapports et de représentation de données (Power BI, Tableaux...etc.). L'ensemble des productions respecte les principes de l'accessibilité numérique pour favoriser l'inclusivité, conformément au

référentiel RG2A (Référentiel Général d'Amélioration de l'Accessibilité). Lors de sa soutenance, le candidat présente ses travaux, les informations issues de ces tableaux de bord, et les exploite dans un data storytelling en lien avec la problématique du projet. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

RNCP39590BC04 — Développer des modèles en utilisant des techniques avancées de statistiques et d'apprentissage automatique pour fournir des informations et accompagner le processus décisionnel

LISTE DE COMPÉTENCES

Déterminer une représentation des données à l'aide de techniques d'ingénierie de caractéristiques (feature engineering) pour entraîner un modèle répondant à la problématique

Entraîner un modèle d'apprentissage supervisé ou non-supervisé à l'aide d'un ensemble de données représentées et en ajustant ses hyperparamètres pour améliorer ses performances et sa capacité à répondre de manière généralisée à une problématique

Établir une stratégie pour la création d'un modèle d'apprentissage en concevant ou réutilisant des modèles pré-entraînés (utilisant le transfert d'apprentissage), et entraîner ces modèles dans le but de résoudre une problématique métier spécifique

Évaluer les modèles d'apprentissage établis en utilisant des techniques de validations croisées et des métriques de performances afin de déterminer le modèle le plus adapté à la problématique métier

Communiquer les résultats des modèles en détaillant les liens avec les objectifs et en faisant preuve d'éthique et de transparence sur leurs limites afin d'accompagner les processus décisionnels

Assurer une veille technologique par des stratégies de surveillance des tendances et des innovations dans le domaine des sciences de la donnée afin de développer des idées novatrices et accompagner la résolution de problèmes

MODALITÉS D'ÉVALUATION

Présentation de la mission en entreprise - Partie Ingénierie de caractéristiques (Feature Engineering) : Mission en situation réelle - Individuel. Le candidat, à partir des données transformées après ETL, applique des techniques dites de feature engineering, pour venir enrichir les données existantes de nouvelles variables ne figurant pas dans l'ensemble d'apprentissage, afin de les rendre exploitables par un modèle d'apprentissage qui viendra répondre à la problématique fixée. Les techniques utilisables sont : encodage (ex: One-Hot Encoding), mathématiques (ex: Transformations logarithmiques, scaling, normalisation...etc.), gestion des données manquantes, gestion des valeurs aberrantes (outliers), extraction de caractéristiques. Lors de sa soutenance, le candidat présente et justifie l'ensemble des techniques qu'il a appliqué et les résultats obtenus. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Analyse des données - Partie Ingénierie de caractéristiques (Feature Engineering) : En autonomie - Groupe de deux à trois candidats - Soutenance individuelle. Le candidat, à partir des données transformées fournies par le référent, applique des techniques dites de "feature engineering", pour venir enrichir les données existantes de nouvelles variables ne figurant pas dans l'ensemble d'apprentissage, afin de les rendre exploitables par un modèle d'apprentissage qui viendra répondre à la problématique fixée. Les techniques utilisables sont : encodage (ex: One-Hot Encoding), mathématiques (ex: Transformations logarithmiques, scaling, normalisation...etc.), gestion des données manquantes, gestion des valeurs aberrantes (outliers), extraction de caractéristiques. Lors de sa soutenance, le candidat présente et justifie l'ensemble des techniques qu'il a appliqué et les résultats obtenus. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Data Science - Partie Ingénierie de caractéristiques (Feature Engineering) : En autonomie - Individuel. Le candidat, à partir des données transformées fournies par le référent, applique des techniques dites de "feature engineering", pour venir enrichir les données existantes de nouvelles variables ne figurant pas dans l'ensemble d'apprentissage, afin de les rendre exploitables par un modèle d'apprentissage qui viendra répondre à la problématique fixée. Les techniques utilisables sont : encodage (ex: One-Hot Encoding), mathématiques (ex: Transformations logarithmiques, scaling, normalisation...etc.), gestion des données manquantes, gestion des valeurs

aberrantes (outliers), extraction de caractéristiques. Lors de sa soutenance, le candidat présente et justifie l'ensemble des techniques qu'il a appliqué et les résultats obtenus. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications.

Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Partie Sélection et entraînement des Modèles d'apprentissage : Mission en situation réelle - Individuel. Le candidat, à partir de données transformées par ETL et enrichies par un processus de "feature engineering", identifie l'ensemble des modèles d'apprentissage (supervisé ou non) permettant de répondre à une problématique d'aide à la décision, et les entraîne pour qu'ils puissent fournir des informations à partir de données généralisées. Le candidat ajuste les hyper-paramètres des modèles en fonction d'évaluation de leurs performances, dans un processus itératif. Les modèles d'apprentissage possibles : Clustering, arbre de décision, régression (logistique, linéaire...etc.), Deep Learning, etc. Les techniques d'optimisation des hyperparamètres peuvent être : Grid Search, Random Search, Optimisation Bayésienne, etc. Lors de sa soutenance, le candidat justifie le choix des modèles et présente leur processus d'entraînement. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Data Science - Partie Sélection et entraînement des Modèles d'apprentissage :

En autonomie - Individuel. Le candidat, à partir de données transformées par ETL et enrichies par un processus de "feature engineering", identifie l'ensemble des modèles d'apprentissage (supervisé ou non) permettant de répondre à une problématique d'aide à la décision, et les entraîne pour qu'ils puissent fournir des informations à partir de données généralisées. Le candidat ajuste les hyper-paramètres des modèles en fonction d'évaluation de leurs performances, dans un processus itératif. Les modèles d'apprentissage possibles : Clustering, arbre de décision, régression (logistique, linéaire...etc.), Deep Learning, etc. Les techniques d'optimisation des hyperparamètres peuvent être : Grid Search, Random Search, Optimisation Bayésienne, etc. Lors de sa soutenance, le candidat justifie le choix des modèles et présente leur processus d'entraînement. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Partie Création des Modèles d'apprentissage : Mission en situation réelle - Individuel. Le candidat établit une stratégie pour la création d'un modèle d'apprentissage en concevant ou réutilisant des modèles pré-entraînés (utilisant le transfert d'apprentissage). Lors de sa soutenance, le candidat devra présenter sa stratégie de création de modèle, en mettant en avant les choix de conception et de réutilisation des modèles pré-entraînés. Il devra également exposer les performances obtenues par ses modèles sur les données de test, ainsi qu'une analyse des résultats et leur pertinence par rapport à la problématique métier. Une justification claire de ses décisions et de ses approches sera également attendue. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Data Science - Partie création des Modèles d'apprentissage : En autonomie - Individuel. Le candidat établit une stratégie pour la création d'un modèle d'apprentissage en concevant ou réutilisant des modèles pré-entraînés (utilisant le transfert d'apprentissage). Lors de sa soutenance, le candidat devra présenter sa stratégie de création de modèle, en mettant en avant les choix de conception et de réutilisation des modèles pré-entraînés. Il devra également exposer les performances obtenues par ses modèles sur les données de test, ainsi qu'une analyse des résultats et leur pertinence par rapport à la problématique métier. Une justification claire de ses décisions et de ses approches sera également attendue. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieures à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Partie Evaluation des Modèles d'apprentissage :

Mission en situation réelle - Individuel. Le candidat quantifie la précision de modèles d'apprentissage et la représentativité des données d'entrées à l'aide de métriques et des techniques d'évaluation de leurs performances respectives. Il identifie des situations précises concernant l'apprentissage des modèles : Overfitting (Sur-apprentissage), Underfitting (Sous-apprentissage), Goodfitting (Apprentissage correct), Unrepresentative Dataset (Données non représentatives). Finalement, il préconise des modifications à apporter, que ce soit au niveau des choix de modèles, des choix de données, des paramétrages, dans un processus itératif d'entraînement. Les métriques/techniques d'évaluation employées : Courbes d'apprentissage, Cross-Validation, courbe ROC (Receiver

Operating Characteristic), AUC (Area Under the Curve),

F1-Score, Score de silhouettes, etc. Lors de sa soutenance, le candidat présente les métriques/techniques exploitées pour justifier la performance de ces modèles tout au long du processus itératif d'entraînement. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet DataScience - Partie Evaluation de Modèles d'apprentissage : En autonomie - Individuel. Le candidat quantifie la précision de modèles d'apprentissage et la représentativité des données d'entrées à l'aide de métriques et des techniques d'évaluation de leurs performances respectives. Il identifie des situations précises concernant l'apprentissage des modèles : Overfitting (Sur-apprentissage), Underfitting (Sous-apprentissage), Goodfitting (Apprentissage correct), Unrepresentative Dataset (Données non représentatives). Finalement, il préconise des modifications à apporter, que ce soit au niveau des choix de modèles, des choix de données, des paramétrages, dans un processus itératif d'entraînement. Les métriques/techniques d'évaluation employées : Courbes d'apprentissage, Cross-Validation, courbe ROC (Receiver Operating Characteristic), AUC (Area Under the Curve), F1-Score, Score de silhouettes, etc. Lors de sa soutenance, le candidat présente les métriques/techniques exploitées pour justifier la performance de ces modèles tout au long du processus itératif d'entraînement. Il doit adapter une posture professionnelle et exploiter un support de présentation mettant en lumière ses productions et leurs justifications. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Partie extraction et présentation des résultats : Mission en situation réelle - Individuel. Le candidat replace des résultats produits par des modèles d'apprentissages entraînés dans le contexte d'une problématique d'aide à la décision. Il les mets en lien et fournit une explication compréhensible par un corpus décisionnaire qui ne possède pas de notion de datascience. Pour cela, il vulgarise son discours, et s'appuie sur des faits, des rapports, des outils visuels (graphiques, KPI...), et en adoptant le vocabulaire du métier. Il fait preuve d'éthique et de transparence face à son auditoire, pour s'assurer de la compréhension par tous.

La soutenance est partie intégrante de l'évaluation de cette compétence. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation du projet Data Science - Partie extraction et présentation des résultats : En autonomie - Individuel. Le candidat replace des résultats produits par des modèles d'apprentissages entraînés dans le contexte d'une problématique d'aide à la décision. Il les mets en lien et fournit une explication compréhensible par un corpus décisionnaire qui ne possède pas de notion de datascience. Pour cela, il vulgarise son discours, et s'appuie sur des faits, des rapports, des outils visuels (graphiques, KPI...), et en adoptant le vocabulaire du métier. Il fait preuve d'éthique et de transparence face à son auditoire, pour s'assurer de la compréhension par tous. La soutenance est partie intégrante de l'évaluation de cette compétence. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Echange avec le jury sur des sujets d'actualités : Mission en situation réelle - Individuel. Lors d'un échange avec le jury, le candidat se voit poser des questions sur des sujets d'actualités autour de la data science et en lien avec la mission effectuée. Il doit y répondre en démontrant sa connaissance du sujet et des contextes qui y sont liés. Il présente en outre les outils et méthodes qu'il exploite pour assurer sa veille technologique (suivi des actualités, journaux, forums, conférences, recherche...etc.). Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

RNCP39590BC05 — Piloter des projets en communiquant avec les parties prenantes et en guidant les membres de l'équipe pour atteindre des objectifs spécifiques

LISTE DE COMPÉTENCES

Collaborer avec des équipes pluridisciplinaires en appliquant les outils du management dans le respect des valeurs éthiques et inclusives afin d'améliorer la performance individuelle et collective

Organiser les activités et les ressources par la mise en place d'une stratégie de gestion afin d'assurer la conduite du projet dans les critères coût-qualité-délais du cahier des charges

MODALITÉS D'ÉVALUATION

Présentation de la mission en entreprise - Echange avec le jury sur la collaboration avec les équipes : Mission en situation réelle -

Individuel. Lors d'un échange avec le jury, le candidat se voit poser des questions sur les échanges et les interactions qu'il a pu avoir avec des interlocuteurs dans l'entreprise et/ou externes à celle-ci. Il doit y répondre en présentant le contexte de ces interactions, et en démontrant sa capacité à s'adapter à ses interlocuteurs, en adoptant une posture qui favorise la transmission des informations et l'engagement dans le projet. Il explique les outils et les méthodes employées lors de la communication de ses travaux et résultats. Les difficultés rencontrées dans les échanges et leur résolution sont exposées. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Présentation de la mission en entreprise - Partie définition de la gestion du projet : Mission en situation réelle - Individuel. Le candidat doit présenter, à partir du cahier des charges validé par l'entreprise, la stratégie et les outils de développement du projet. Il doit présenter : le planning (ex: Gantt) du développement du projet, les responsabilités de chacun des interlocuteurs du projet, les indicateurs clés et des outils de gestion pour suivre l'avancement du projet et le respect des contraintes, les outils de communication et d'échanges avec le client et les interlocuteurs pour anticiper les problématiques liées au projet. Les candidats ont pour contraintes : d'assurer le respect des budgets, des niveaux de qualités, et des délais associés au projet, de favoriser le travail collaboratif, d'adopter une posture professionnelle et inclusive dans les échanges qu'ils peuvent avoir, et de démontrer leur capacité à exploiter des outils standards de gestion de projet. Évalué par un jury d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieurs à la structure d'accueil et de la structure de certification.

Secteur d'activité et type d'emploi

SECTEURS D'ACTIVITÉS

Ce métier relève des enjeux à la fois fonctionnels et stratégiques. Bien qu'il s'agisse d'un métier récent, on retrouve les Data Scientists dans de nombreux secteurs d'activité. Finance, informatique, assurance, e-commerce ou encore grande distribution ; tant de domaines dans lesquels le Data Scientist peut exercer dans le cadre d'agence digitale, de conseil, une ESN, un annonceur, dans la distribution, le Retail, le E-commerce, chez un éditeur de logiciels, pour des plateformes digitales, d'intermédiation, de Pure player, ou encore au sien d'une Start-up.

TYPE D'EMPLOIS ACCESSIBLES

DataScientist, Consultant Data Senior, Chef de projet Data, Consultant Data, Chief Data scientist

En fonction de sa montée en compétences, le Data Scientist

pourra évoluer vers un poste de Chef de projet data ou Data Scientist senior après une première expérience significative.

Marché du travail

Indicateurs au niveau des familles professionnelles (FAP) des métiers visés, agrégés au niveau national — France Travail / DARES (open data).

Tension de recrutement (DARES 2024)	Très élevée (5/5) Moyenne pondérée par l'emploi ; les familles ci-dessous s'échelonnent de 4 à 5 sur 5.
Facteurs principaux	Manque de main-d'œuvre disponible, Lien formation-emploi
Salaire médian (mensuel)	2 864 €
Diplôme le plus fréquent	Diplôme de niveau bac+5 ou plus (doctorat, master ou équivalent) en communication et information (41 % des actifs en emploi)
Familles professionnelles (FAP)	M1X80 — Techniciens d'étude et de développement en informatique · tension 4/5 (Élevée) M2X90 — Ingénieurs et cadres d'étude, recherche et développement en informatique et télécom · tension 5/5 (Très élevée)

Projets de recrutement (BMO 2026)	27 458
Jugés difficiles à recruter	53 %
Emplois saisonniers	1 %
Familles professionnelles (FAP)	M1X80 — Techniciens d'étude et de développement en informatique M2X90 — Ingénieurs et cadres d'étude, recherche et développement en informatique et télécom

Voies d'accès

MODALITÉS D'ACCÈS

Formation continue, Contrat de professionnalisation, VAE

PRÉREQUIS À L'ENTRÉE EN FORMATION

Pour les publics demandeurs d'emplois et hors VAE : une formation scientifique menant à un niveau 6 en informatique avec de bonnes connaissances en mathématiques et statistiques, ou à un niveau 7 (type master/diplôme d'ingénieur scientifique), complété de la réussite préalable à un test de recrutement écrit de 2 heures et à un entretien de motivation professionnelle.

PRÉREQUIS À LA VALIDATION

La validation totale de la certification s'obtient par la validation des 5 blocs qui la compose ; hors VAE, avoir suivi l'intégralité de la formation.

COMPOSITION DU JURY — FORMATION INITIALE, FORMATION CONTINUE, APPRENTISSAGE, CONTRAT DE PROFESSIONNALISATION, VAE

Le jury est composé d'au moins trois professionnels issus du domaine de l'informatique et/ou de l'industrie, dont au moins 50% de Data Scientist professionnels, et extérieures à la structure d'accueil et de la structure de certification. Une représentation équilibrée hommes/femmes sera recherchée

Base légale

Date de décision	01/10/2024
Durée de l'enregistrement	2 ans
Date d'échéance de l'enregistrement	01/10/2026
Date de dernière délivrance possible	01/10/2030

Statistiques

Année	Certifiés	Dont VAE	Insertion 6 mois	Métier 2 ans
2023	11	0	64 %	—
2022	21	0	60 %	—
2021	16	0	64 %	—

Liens avec d'autres certifications

ANCIENNES VERSIONS

RS2858

Pour plus d'informations

LIEN VERS LE DESCRIPTIF

<https://www.icam.fr/formations-professionnelles/formation-data-scientist/>

