5 - REFERENTIELS

Article L6113-1 En savoir plus sur cet article... Créé par LOI n°2018-771 du 5 septembre 2018 - art. 31 (V)
« *Les certifications professionnelles enregistrées au répertoire national des certifications professionnelles permettent une validation des compétences et des connaissances acquises nécessaires à l'exercice d'activités professionnelles. Elles sont définies notamment par un* ***référentiel d'activités*** *qui décrit les situations de travail et les activités exercées, les métiers ou emplois visés, un* ***référentiel de compétences*** *qui identifie les compétences et les connaissances, y compris transversales, qui en découlent et un* ***référentiel d'évaluation*** *qui définit les critères et les modalités d'évaluation des acquis.* »

Chaque bloc de compétences est évalué sous la forme de projet, accompagné et validé par un mentor (professionnel du métier), puis présenté au jury du titre.

REFERENTIEL D'ACTIVITES *décrit les situations de travail et les activités exercées, les métiers ou emplois visés*	**REFERENTIEL DE COMPETENCES** *identifie les compétences et les connaissances, y compris transversales, qui découlent du référentiel d'activités*	**REFERENTIEL D'EVALUATION** *définit les critères et les modalités d'évaluation des acquis*	
		MODALITÉS D'ÉVALUATION	**CRITÈRES D'ÉVALUATION**
BLOC 1 – Prétraiter et analyser des données structurées pour répondre à un problème métier.			
1.1- Cadrage et problématisation d'une analyse de données.	Déterminer les objectifs d'une analyse de données à partir d'un problème métier.	Mise en situation professionnelle (projet) Et Soutenance orale	La demande du client interne/ externe est prise en compte. La problématique métier est correctement identifiée. Les objectifs de l'analyse de données à mener sont pertinents et permettent de répondre à la problématique métier.
1.2- Nettoyage de données (data cleaning).	Effectuer des opérations de nettoyage sur des données structurées.		Les données manquantes sont correctement identifiées. Des actions correctives sont proposées.

			Les duplicats de variables et d'enregistrement son traités. Les variables non pertinentes pour la problématique sont éliminées.
1.3- Analyse exploratoire de données structurées.	Effectuer une analyse statistique univariée à partir de données structurées et nettoyées.		Les métriques adaptées sont utilisées. Les valeurs aberrantes sont identifiées. Des actions correctives sont proposées. Les distributions observées sont correctement caractérisées. Les quantiles sont correctement définis.
	Effectuer une analyse statistique multivariée à partir de données structurées et nettoyées.		L'analyse effectuée est pertinente au regard de la problématique. Les résultats de l'analyse sont corrects.
	Représenter des données grâce à des graphiques clairs et pertinents.		Les graphiques produits comprennent au moins un graphique de chaque parmi les suivants : box-plot (boîte à moustaches), barplot (graphique en barres), pie chart (diagramme circulaire),

			histogramme, scatter plot (nuage de points). Les graphiques sont lisibles.
1.4- Formulation de préconisations pour une analyse automatisée des données.	Formuler des préconisations pour un traitement des données permettant au client interne/ externe d'en automatiser l'analyse.		Des propositions sont formulées pour un traitement potentiel des données. Ces propositions sont cohérentes et pertinentes.
BLOC 2 – Entraîner un modèle d'apprentissage automatique supervisé pour réaliser une analyse prédictive.			
2.1- Mise en place d'un modèle d'apprentissage automatique supervisé.	Sélectionner et/ ou transformer les variables pertinentes pour la modélisation (*feature engineering*).	Mise en situation professionnelle (projet) Et Soutenance orale	Les variables catégorielles sont identifiées et transformées de façon adaptée. Si nécessaire, les variables sont normalisées. Si nécessaires, des transformations mathématiques sont utilisées.
	Sélectionner et mettre en place un modèle d'apprentissage supervisé adapté à une problématique métier.		La (les) variable(s) cibles sont correctement choisies et corrélées. Au moins un algorithme de régression linéaire et un algorithme de régression non linéaire sont testés.

2.2- Évaluation et amélioration d'un modèle d'apprentissage automatique supervisé.	Évaluer les performances d'un modèle d'apprentissage supervisé.		La métrique choisie est adaptée (RMSE, écart moyen). D'autres indicateurs de performance que le score sont utilisés pour analyser le résultat (coefficients des variables en fonction de la pénalisation, visualisation des erreurs en fonction des variables du modèle, temps de calcul...) Les données sont séparées en train/test pour les évaluer de façon pertinente et détecter le surapprentissage (*overfitting*). Une *baseline data* est mise en place pour évaluer le pouvoir prédictif du modèle choisi (variable qualitative – *dummy regressor*). L'ensemble des résultats sont présentés en allant des modèles les plus simples aux plus complexes

	Adapter les paramètres d'un modèle d'apprentissage supervisé afin de l'améliorer.		Une validation croisée est mise en place pour optimiser les hyperparamètres. Les hyperparamètres pertinents sont optimisés dans les différents algorithmes. Le choix final d'algorithme et d'hyperparamètres est justifié.
BLOC 3 – Entraîner un modèle d'apprentissage non supervisé adapté à une problématique de segmentation ou de réduction de données.			
3.1- Mise en place d'un modèle d'apprentissage automatique non supervisé.	Sélectionner, transformer et créer les variables pertinentes pour la modélisation (*feature engineering*).	Mise en situation professionnelle (projet) Et Soutenance orale	Les variables pertinentes sont transformées pour permettre leur exploitation (variables catégorielles en particulier). Une ou plusieurs variables pertinentes pour l'amélioration de la solution son créées.

	Sélectionner et mettre en place un modèle d'apprentissage non supervisé adapté une problématique métier.		Le nombre de segments et la répartition bien choisis. La stratégie d'ajout de nouveaux segments est explicitée. La nature des variables d'entrée est prise en compte dans le choix de l'algorithme et de la distance.
3.2- Évaluation et amélioration d'un modèle d'apprentissage automatique non supervisé.	Évaluer les performances d'un modèle d'apprentissage non supervisé.		La forme des clusters est évaluée. La stabilité des clusters est évaluée. La compatibilité avec des connaissances extérieures ("test set" et/ou a priori) est évaluée.
	Adapter les paramètres d'un modèle d'apprentissage non supervisé afin de l'améliorer.		Les étapes d'évaluation sont automatisées et permettent de pouvoir tester facilement plusieurs combinaisons de paramètres. Le choix des valeurs de paramètres testés est pertinent.

BLOC 4 – Prétraiter et analyser des données non structurées (texte, images).			
4.1- Collecte et prétraitement de données non structurées.	Collecter des données répondant à des critères définis *via* une interface de programmation (API).	Mise en situation professionnelle (projet) Et Soutenance orale	Les conditions d'utilisation de l'API dans la documentation (limite du nombre d'appels et limitations d'utilisation) sont correctement identifiées. La structure de l'API et les champs contenant les données d'intérêt *via* la documentation sont correctement identifiés. La requête permettant de récupérer les données d'intérêt *via* l'API est écrite.
	Prétraiter des données textuelles non structurées pour obtenir un jeu de données exploitable.		Les champs de texte sont nettoyés (retirer la ponctuation, les mots de liaison, mettre tout en minuscule, ...) Une fonction permettant de tokeniser une phrase a été écrite et fonctionne correctement. Une fonction permettant de *stemmer* (raciniser) une phrase a été écrite et fonctionne

			correctement. Une fonction permettant de lemmatiser une phrase a été écrite et fonctionne correctement. Des variables (*features*) *bag-of-word* sont construites (avec étapes de nettoyage supplémentaires : seuil de fréquence des mots, normalisation des mots (racines, utilisation du package NLTK). Une phrase (ou un court texte) d'exemple permet d'illustrer et de tester la bonne réalisation des 5 étapes précédentes.
	Prétraiter des données sous forme d'images non structurées pour obtenir un jeu de données exploitable.		Des librairies spécialisées sont utilisées pour un premier traitement du contraste (ex : openCV). Le bruit est filtré. L'histogramme est égalisé. Une fonction permettant d'extraire des *features* (*via* l'un des algorithmes suivants : ORB, SIFT, SURF ou réseaux de

			neurones comme par exemple CNN) a été écrite et fonctionne correctement. Une fonction permettant de construire des *features* de type "bag-of-image" a été écrite et fonctionne correctement.
4.2- Analyse exploratoire de données de grande dimension.	Réduire la dimension de données de grande dimension afin d'optimiser les temps de calcul.		La nécessité de réduction de dimension (en particulier dans le cas de données image et texte) est justifiée (coût d'acquisition, de stockage, complexité de calcul…) Une méthode de méthode de réduction de dimension (ex : ACP) pour des données image et texte est appliquée. Le choix des valeurs des paramètres dans la méthode de réduction de dimension retenue (ex : le nombre de dimensions conservées pour l'ACP) est justifié.
	Représenter graphiquement des données de grande dimension afin d'en réaliser une analyse exploratoire.		Au moins un graphique représentant les informations contenues dans des données à plus de 2D a été réalisé.

			Le graphique réalisé est lisible et compréhensible. Les différents éléments graphiques sont expliqués (variables représentées sur les axes, par les couleurs, la taille, ...)
BLOC 5 – Présenter et déployer un modèle d'apprentissage automatique auprès de ses utilisateurs finaux.			
5.1- Mise en production d'un modèle d'apprentissage automatique.	Déployer un modèle *via* une interface de programmation (API) dans le web.	Mise en situation professionnelle (projet) Et Soutenance orale	Un fichier (par exemple format pickle) contenant un modèle de machine learning sérialisé a été créé et le modèle chargé depuis le fichier fonctionne. Le modèle d'apprentissage automatique (machine learning) est déployé sous forme d'API (*via* Flask par exemple). L'API renvoie effectivement la prédiction correspondante à un client en réponse à un identifiant client. L'API est déployée dans le web *via* un outil gratuit et disponible plusieurs mois (en vue du jury).

5.2- Formalisation et présentation d'une démarche de modélisation et de ses résultats.	Réaliser un tableau de bord (*dashboard*) pour présenter son travail de modélisation.		Au moins un parcours utilisateur simple permettant de répondre aux besoins des utilisateurs (les différentes actions/clics sur les différents graphiques permettant de répondre à une question que se pose l'utilisateur) a été décrit et conçu. Au moins deux graphiques interactifs permettant aux utilisateurs d'explorer les données clients (permettant de répondre à des questions de type "quel est le client avec le plus de transactions ?") ont été développés. Les graphiques réalisés sont lisibles, clairs et pertinents. Le *dashboard* est accessible pour d'autres utilisateurs sur leurs postes de travail (déploiement dans le web).
	Réaliser la présentation orale d'une démarche de modélisation à un client interne/ externe.		La démarche de modélisation est détaillée avec : - La méthode d'évaluation de la performance du modèle de

			machine learning. - La façon d'interpréter les résultats du modèle. - La façon d'interpréter l'importance des variables du modèle. Les explications sont claires et fournies avec assurance. Les explications sont compréhensibles par un public non technique.
	Rédiger une note méthodologique afin de communiquer sa démarche de modélisation.		La démarche de modélisation est présentée de manière synthétique dans la note (2 pages max). La fonction coût, l'algorithme d'optimisation et la métrique d'évaluation sont explicités (1 page max). L'interprétabilité du modèle est explicitée (la façon d'interpréter l'importance des variables n'est pas la même

			pour une régression logistique que pour un random forest) et les limites éventuelles sont précisées (1 page max). Le limites et les améliorations envisageables pour gagner en performance et en interprétabilité de l'approche de modélisation sont décrites (1 page max).
	Assurer l'intégration du modèle auprès de collaborateurs en utilisant un logiciel de version de code.		Un dossier contenant tous les scripts du projet a été créé dans un logiciel de version de code (ex : Github). L'historique des modifications du projet affiche au moins trois versions distinctes et l'on peut accéder à ces anciennes versions. La liste des packages utilisés ainsi que leur numéro de version est disponible et tenu à jour. Le travail est réutilisable par d'autres personnes et la collaboration est facilitée : - un fichier introductif permet de comprendre l'objectif du

			projet et le découpage des dossiers - les scripts et les fonctions sont commentés.
BLOC 6 – Déployer un modèle d'apprentissage automatique à l'échelle en utilisant les technologies du *Big data*.			
6.1- Analyse et modélisation de données dans un environnement *Big data*.	Sélectionner les outils du *Cloud* permettant de disposer d'un environnement *Big Data*.	Mise en situation professionnelle (projet) Et Soutenance orale	Les différentes briques d'architecture nécessaires pour la mise en place d'un environnement *Big data* sont identifiées. Les outils du cloud permettant de mettre en place l'environnement *Big Data* sont identifiés.
6.2- Réalisation d'opérations de calcul de grands volumes de données.	Prétraiter, analyser et modéliser des donnés dans un environnement *Big data* en utilisant les outils du *Cloud*.		Les fichiers (de départ et ceux après transformation) sont chargés dans un espace de stockage cloud. Les scripts ont été exécutés en utilisant des machines dans le cloud. Un des scripts permet d'écrire les sorties du programme directement dans l'espace de stockage cloud.

	Réaliser des calculs distribués sur des données massives en utilisant les outils adaptés.		Les traitements critiques lors d'un passage à l'échelle en termes de volume de données sont identifiés. Les scripts sont développés en Pyspark (API).