8IAR403 : Apprentissage Machine pour la SD

Devoir #1
Compréhension et préparation des données d'apprentissage

Le travail est en individuel
Date de remise le 16 février 2026

1. But

- Valider les 4 premières étapes de la méthodologie de conduite de projets en Machine Learning vue en cours sur une étude de cas de vente en ligne;
- Se familiariser davantage avec la problématique de la compréhension et la préparation des données d'apprentissage;
- Utilisation du concept de pipeline et plusieurs sources de données.

2. Étude de cas

Le contexte métier utilisé est celui d'un site d'e-commerce pour lequel on souhaite prédire les revenus que vont générer de nouveaux clients. Par conséquent, chercher le profil du bon client ! L'ensemble des données utilisées sur les clients, comme le montre la figure suivante, possède 10 000 lignes et **disponible sur le site du cours dans un fichier (Customer.csv) en format CSV**.

	age	pages	first_item_prize	gender	ReBuy	News_click	country	revenue
0	41	6	28.00000	Fem	False	4	China	113
1	34	4	15.50000	Fem	True	2	China	36
2	38	5	40.43173	Fem	False	7	China	111
3	20	1	44.00000	Fem	False	2	China	71
4	39	10	10.00000	Fem	True	4	China	80

Les variables (caractéristiques) décrivant les clients sont :

1. **age** : âge du client;
2. **pages** : le nombre de pages du site visité;

3. **first_item_prize** : le prix du premier article acheté;
4. **gender** : masculin ou féminin;
5. **ReBuy** : le client a-t-il acheté le premier article plus d'une fois ?
6. **New_click** : nombre de fois où le client a cliqué sur une compagne de publicité du site;
7. **country** : le pays dont provient l'adresse IP;
8. **revenue** : revenu généré par le client sur le site.

Deux « datasets » additionnels CountryGDP et CountryPopulation, sont aussi disponibles en format CSV sur le site du cours, indiquent respectivement le PIB (produit intérieur brut) et la population du pays correspondant à l'adresse IP de la requête du client.

3. Travail à faire

Il s'agit de reproduire **les étapes (2-4)** de la démarche de conduite de projet ML vue en cours[1] dans le but de préparer les données nécessaires à l'entrainement du modèle prédictif visé dans le prochain **devoir#2**. Plus précisément, ce travail servira à préparer vos données d'apprentissage pour la suite des travaux à faire.

Pour cela, vous devez conduire deux opérations principales de nettoyage et d'enrichissement des données à travers l'écriture d'un **pipeline pour automatiser**, le plus possible, ces opérations en question :

- **Nettoyage** : repérer les données manquantes, aberrantes puis remédier selon les techniques appropriées (Cf. sous-section 3.2);

- **Enrichissement** : transformer le dataset de base (Customer.csv) en rajoutant les informations liées à la population (CountryPopulation.csv) et au PIB (CountryGDP.csv) du pays de provenance de la requête (Cf. sous-section 3.3).

3.1 Reproduire les étapes 2-4 de la méthodologie de conduite de projet ML du chapitre 2. Vous devez inclure uniquement les opérations utiles (2 points).

3.2 Nettoyage des données du dataset de base Customer.csv

3.2.1 Remplacement des données manquantes (6 points)

Utilisez les fonctions info(), describe(),…, et la visualisation pour avoir une idée sur les anomalies au sein des données afin de les repérer. Vous devez écrire des fonctions en python et/ou recourir à des outils intégrés dans la libraire scikit-learn, pour localiser et corriger ces anomalies. Il est important d'intégrer la localisation et la transformation dans un pipeline en vue d'automatiser le processus de nettoyage de toutes les données y compris celles qui vont être manipulées par la suite, comme le jeu de test ou autre jeu de données.

[1] A consulter le tutoriel du chapitre 2 du cours.

3.2.2 Remplacement des données aberrantes (extrêmes, outliers) (6 points)

Utilisez la méthode « **boite à moustaches** » permettant de visualiser graphiquement le bruit, s'il y en a, dans vos données relativement à chaque variable. Pour y remédier, écrivez une fonction (un transformateur sur mesure) permettant de remplacer ou éliminer les données en question. Par exemple, implanter la technique de « amplitude interquartile » basée sur les quartiles permettant de remplacer le bruit par une valeur estimée (interpolation) moyennant le premier quartile (25%), la médiane (deuxième quartile) et le troisième quartile (75%). Vous devez inclure cette fonction dans le pipeline global.

3.3 Enrichissement des données (6 points)

L'étape d'enrichissement ressemble à l'étape 3.3 du tutoriel sur la méthodologie de conduite de projets ML (Cf. chapitre2) sur l'expérimentation avec la combinaison de variables (ou variables dérivées). Dans ce devoir, au lieu de dériver de nouvelles variables, on vous demande de rajouter les deux datasets CountryPopulation et CountryGDP au dataset de base Customer.

Écrivez une fonction qui jouera le rôle de transformateur afin de former un dataset fusionné via une **jointure** et qui contiendra les données sur la Population et éventuellement le PIB. On peut supposer que le PIB sera l'argument de ce transformateur. Pour cela, ce transformateur doit:

a) Nettoyer les deux dataset (CountryGDP et CountryPopulation) qu'on souhaite ajouter au dataset de base (Customer.csv).

b) Utiliser la fonction **merge()** fournit par Pandas par exemple ou un autre outil intégré à scikit-learn, pour faire la première jointure entre le dataframe de base (Customer) et le dataframe (CountryPopulation). Pour faire une jointure, il faut une clé commune entre ces deux dataframes. La clé commune est la variable **country** (de type Object) qui doit être écrite de la même manière dans les deux datasets en **respectant les majuscules et les minuscules**. Renommer s'il y a lieu pour égaliser.

c) Si l'option de l'argument « PIB » est à True, cela signifie qu'on doit ajouter les données de CountryGDP. Par conséquent, faire une deuxième jointure entre le **résultat de la première jointure** et le dataframe CountryGDP. Avant de faire cette jointure s'assurer aussi que la clé « country » de cette dernière est conforme, comme le premier cas.

d) Retourner le résultat de la jointure comme nouveau dataset fusionné.

Vous devez inclure cette étape d'enrichissement dans le pipeline global.

4. Livrable

4.1 Fournir un seul fichier **ipynb** comprenant les étapes de **2-4 de la méthodologie de conduite de projet ML** Je vous recommande de travailler avec jupyterLab.

4.2 Le résultat d'exécution (sortie) de chaque cellule du notebook doit figurer. Le but est de m'éviter la réexécution de votre code et de me faciliter ainsi la correction de votre devoir.

4.3 Commenter chacune des étapes en utilisant Markdown et en faisant référence aux questions posées. Même s'l n'y a pas de réponse à une question, il faut l'indiquer dans une cellule du notebook pour me faciliter la correction. Les réponses doivent respecter l'ordre des questions.

4.4 Indiquer dans l'introduction du fichier **ipynb, votre nom et prénom et** si vous avez ajouté de nouvelles fonctionnalités non-demandées. Il faut les commenter.

4.5 Transmettez un fichier compressé en .zip. Veuillez le déposer dans le dossier « dépôt » prévu sur le site moodle du cours, dans la section « devoir ». A Éviter de déposer sur d'autre site de téléchargement ! AUCUN RENDU PAR COURRIEL NE SERA ACCEPTÉ

4.6 Le nommage du rendu compressé doit être comme suit: nom_ prénom_tp1_h2026.zip

Critères de correction :

Critères	Points
Q3.1 : Reproduction des étapes (2 à 4)	/2
Q3.2.1 : Localisation des données manquantes	/1
Q3.2.1 Nettoyage DM	/3
Q3.2.1 Utilisation d'un pipeline	/2
Q3.2.2 : Localisation de données aberrantes	/1
Q3.2.2 : Nettoyage DA	/3
Q3.2.2 Utilisation d'un pipeline	/2
Q3.3 Enrichissement des données: Nettoyage des deux datasets (PIB, Population)	/1
Q3.3 Enrichissement des données : Jointure paramétrable	/3
Q3.3 Utilisation d'un pipeline	/2

Total : /20

Bonne continuation !