Université de Montréal
Science des données

IFT-6758 – Examen Intra

2023/10/30

Nom Complet (Prénom Nom)

Numéro de matricule (UdeM) | | | | | | | | |

Instructions

- Aucune ressource externe n'est autorisée.
- Donnez de brèves explications pour vos réponses. Nous accorderons un crédit partiel pour un bon raisonnement.
- Pour les questions avec une boîte a remplir, indiquer votre réponse dans la boîte.
- Il y a du papier supplémentaire à l'avant de la pièce. Assurez-vous de mettre votre nom, votre identifiant et votre numéro de question pour toutes les pages supplémentaires que vous utilisez.

Notations and rappels:

- $\|x_i - x_j\|_2^2 = \sum_{k=1}^{d}([x_i]_k - [x_j]_k)^2$ = Euclidean distance between points x_i and x_j
- $A \setminus B$ = Complémentaire d'un ensemble: éléments dans A mais pas dans B
- $H(X) = -\sum_{i=1}^{d} p(X = x_i)\log(X = x_i)$, entropie d'un variable aléatoire discrète X.
- $I(X,Y) = \sum_{i,j}^{n} p(X = x_i, Y = y_i)\log\left(\frac{p(X=x_i,Y=y_i)}{p(X=x_i)p(Y=y_i)}\right)$ information mutuelle (ou gain d'information) entre deux variables aléatoires X et Y.
- Taux de vrai positifs : $=\frac{\#(\text{prédiction}=+|\text{vrai valeur}=+)}{\#(\text{vrai valeur}=+)}$, Taux de vrai négatifs: $=\frac{\#(\text{prédiction}=-|\text{vrai valeur}=-)}{\#(\text{vrai valeur}=-)}$

Distribution des points

Question	Points	Score
1	2	
2	17	
3	9	
4	13	
5	15	
6	10	
7	13	
8	6	
9	0	
Total:	85	

1. **Échauffement**
   (a) (1 point) Écrire correctement et lisiblement votre nom complet et numéro de matricule (voir Figure 1).
   (b) (1 point) (À la fin) Mettre sa copie dans la bonne pile (marquée par la **première lettre** de votre **prénom**)

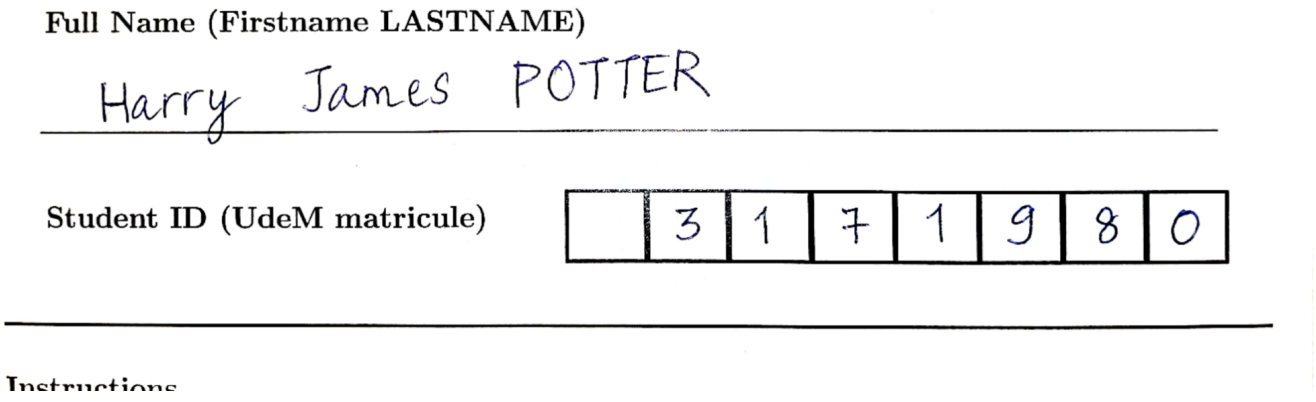


Figure 1: ou comment marquer un point facilement.

2. **Ingénierie des caractéristiques**

Votre équipe de science des données a acquis un ensemble de données relatives à la location de vélos dans une ville métropolitaine. On souhaite utiliser l'ensemble de données pour prédire le nombre de locations de vélos sur une période horaire en fonction de divers facteurs.

- `heure_journée` : Représente l'heure de la journée sour la forme `HH:MM:SS` avec `HH` un entier entre 0 et 23, `MM` un entier entre 0 et 59 et `SS` $\in [0, 60[$ un nombre réel entre 0-60.
- `condition_météo` : Conditions catégorisées comme 'Clair', 'Nuageux', 'Pluie', etc.
- `temperature` : Température enregistrée en degrés Celsius.
- `humidité` : Représentation en pourcentage de l'humidité.
- `vitesse_vent` : Vitesse du vent mesurée en km/h.
- `jour_semaine` : Spécifie le jour de la semaine ('Lundi', 'Mardi', etc.).
- `saison` : Saison correspondante de l'année ('Printemps', 'Été', 'Automne', 'Hiver').
- `locations_heure_précédente` : Nombre de locations dans l'heure précédente.
- `vacances` : Une valeur binaire (0 ou 1) signifiant s'il s'agit d'un jour férié.
- `nombre_de_locations` : Variable cible - nombre de vélos loués dans l'heure spécifiée.

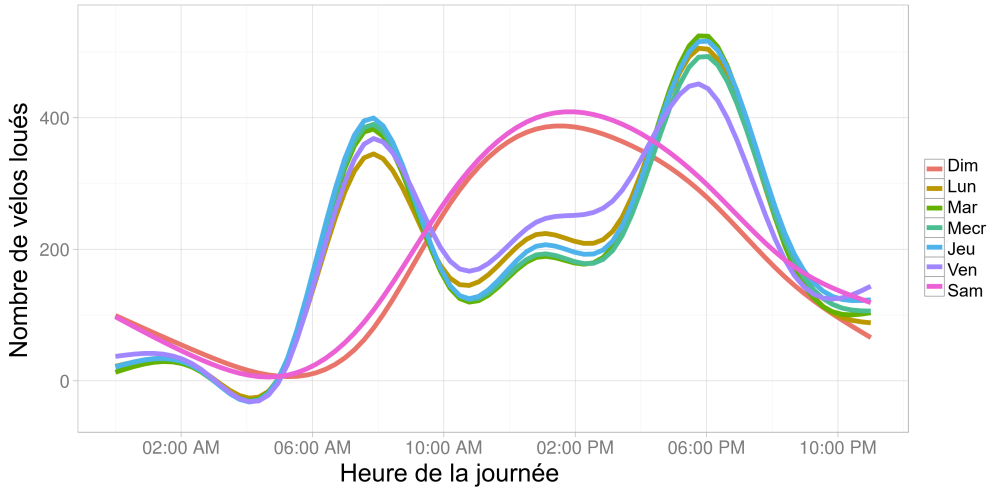


Figure 2: Visualisation de la relation entre `heure_journée` et `nombre_de_locations` pour différentes valeurs de `jour_semaine`.

(a) (3 points) Expliquez comment vous géreriez mieux les valeurs manquantes dans les caractéristiques `temperature` et `conditions_météo` de manière à ne pas polluer les données? Fournissez un **raisonnement** pour votre méthode choisie dans chaque cas.

Solution: `température` : On pourrait remplacer les valeurs manquantes par la température moyenne ou médiane **de la saison respective** ou **d'une plage temporelle valide** (jour précédent et suivant, etc.). Le choix entre la moyenne et la médiane dépendrait de la distribution des données de température. Si les données sont normalement distribuées sans valeurs aberrantes, la moyenne est un choix approprié. Si il y a des valeurs aberrantes, la médiane serait un meilleur choix car elle est plus robuste aux v aleurs aberrantes.

Raison : La moyenne est sensible aux valeurs aberrantes, ce qui peut fausser la représentation de la tendance centrale, tandis que la médiane fournit une meilleure valeur centrale dans de tels cas.

`conditions_météoro` : Une imputation par le mode est appropriée pour les données catégorielles. On pourrait remplacer les valeurs manquantes par la condition météorologique la plus fréquente **de la saison respective** ou **d'une plage temporelle valide** (heure précédente et suivante, etc.).

Raison : Les conditions météorologiques sont catégorielles et n'ont pas de moyenne significative. Le mode, étant l'observation la plus fréquente, est une estimation raisonnable des conditions météorologiques typiques et maintient la nature catégorielle des données.

(b) (3 points) En utilisant Fig. 2 ci-dessus, que pouvez-vous conclure sur l'impact des caractéristiques `heure_journée` et `jour_semaine` sur la variable cible `nombre_de_locations` ? Suggérez un autre type de visualisation que vous utiliseriez à la place pour étudier l'impact de ces deux caractéristiques sur la variable cible. **Justifiez** votre réponse.

Solution: D'après la Fig. 2, il existe un schéma clair dans les locations de vélos en fonction de l'heure de la journée et du jour de la semaine. Les locations atteignent un pic pendant certaines heures, en particulier autour des heures de trajet typiques, et varient en fonction du jour, les week-ends présentant un schéma différent par rapport aux jours de semaine.

Un autre type de graphique qui pourrait être utilisé pour étudier l'impact de ces deux caractéristiques est une carte de chaleur (heatmap), qui peut représenter visuellement la concentration des locations à différents moments et jours.

Justification : Une carte de chaleur permet une comparaison facile des densités, rendant possible l'identification non seulement des pics mais aussi des variations à travers tout le spectre des combinaisons de temps et de jours. (Alternativement, tout autre type de graphique qui visualise les comptages de locations de vélos par rapport à un continuum des heures sur les jours est acceptable tant qu'il est clairement décrit)

(c) (3 points) `jour_semaine` et `vacances` ont tous deux des implications sur le nombre de locations. Suggérez une méthode pour combiner ces caractéristiques qui permet de mieux prédire les augmentations de location le weekend (fin de semaine) ou les vacances.

Solution: Pour combiner `jour_semaine` et `vacances` dans une seule caractéristique, on pourrait créer une nouvelle caractéristique qui capture les week-ends normaux, les week-ends adjacents aux jours fériés (prolongant potentiellement les jours fériés) et les jours fériés autonomes. Il peut s'agir d'une caractéristique catégorielle comportant plusieurs niveaux, indiquant les jours normaux, les week-ends, les week-ends/périodes de vacances prolongés et les jours fériés.

(d) (3 points) La caractéristique `heure_journée` est une représentation continue des heures d'une journée. En quoi cela pourrait-il être problématique en termes de représentation de sa vraie nature, en particulier si l'on considère les heures du matin et du soir ? Proposez une transformation qui puisse résoudre ce problème.

Solution: La caractéristique `heure_journée` peut être problématique car elle ne tient pas compte de la nature cyclique du temps. Les heures du matin et du soir sont proches l'une de l'autre en termes d'activités quotidiennes, mais éloignées numériquement.

Une transformation qui peut résoudre ce problème consiste à convertir `heure_journée` en caractéristiques cycliques à l'aide de transformations sinus et cosinus. De cette façon, les heures comme 23h00 et 00h00 sont plus rapprochées dans l'espace de caractéristiques transformé.

`heure_journée_sin` $= \sin\left(\frac{2\pi \times \text{hour}}{24}\right)$

`heure_journée_cos` $= \cos\left(\frac{2\pi \times \text{hour}}{24}\right)$

(e) (3 points) La caractéristique `locations_heure_précédente` peut être vue comme une représentation temporelle. Comment cette caractéristique pourrait-elle aider à prédire le `nombre_de_locations`? Serait-elle mieux représentée comme une caractéristique catégorielle ou continue? Expliquez.

Solution:

La caractéristique `locations_heure_précédente` est indicative des tendances récentes et peut fournir un contexte pour les locations de l'heure en cours. Il devrait s'agir d'un élément continu, car il représente un décompte qui peut varier dans une large mesure et n'est pas de nature catégorique. La représentation continue permet de capturer la variabilité et l'évolution des locations d'une heure à l'autre, ce qui peut être très prédictif du futur immédiat (locations de l'heure en cours).

(f) (2 points) Si vous deviez représenter `conditions_météo` comme une caractéristique numérique continue plutôt que catégorique, quelle méthode d' encodage de caractéristique utiliseriez-vous? À quoi ressembleraient les valeurs de la caractéristique? **Justifiez**.

Solution:

Si nous représentons `conditions_météo` comme une caractéristique numérique continue, on pourrait utiliser un codage de fréquence où les conditions sont codées par leurs fréquences dans les données.

Les valeurs des caractéristiques peuvent ressembler à :

- Claire = 0,4
- Nuageux = 0,25
- Pluvieux = 0,3
- Orageux = 0,15

Justification: Le codage de fréquence est justifié car il pourrait y avoir une différence naturelle dans l'ampleur de l'impact de ces catégories sur la location de vélos, par exemple, le temps clair étant plus courant et idéal pour la location de vélos, tandis que le temps orageux étant rare et pouvant les réduire considérablement.

3. **Selection de caractéristique**

Considérez le même scénario détaillé dans **Question 2** pour répondre aux questions suivantes.

(a) (2 points) Votre équipe envisage l'application de méthodes d'encapsulage pour la sélection de caractéristiques, mais n'est pas sûre qu'il s'agisse de la meilleure option. Indiquez **deux** situations dans lesquelles cette méthode pourrait être la plus adaptée et avantageuse pour cet ensemble de données.

Solution: Les méthodes de type wrapper seraient les plus appropriées pour le jeu de données de location de vélos dans les deux situations suivantes :

1. Lorsque les effets d'interaction entre les caractéristiques sont significatifs pour la prédiction de la variable cible, les méthodes de type wrapper peuvent être avantageuses car elles prennent en compte le sous-ensemble de caractéristiques qui fonctionnent le mieux ensemble dans le contexte du modèle prédictif choisi.

2. Si le jeu de données n'est pas trop grand, les méthodes de type wrapper, qui sont intensives en termes de calcul, peuvent être utilisées sans coûts de temps prohibitifs. Elles fournissent une évaluation plus précise de l'importance des caractéristiques en évaluant les sous-ensembles de caractéristiques basés sur la performance du modèle.

(b) (2 points) Vous souhaitez utiliser une méthode de filtrage pour capturer la relation entre les caractéristiques de manière indépendante et la variable cible. Quelle méthode recommanderiez-vous? **Justifiez** votre réponse.

Solution: Pour capturer la relation entre les caractéristiques de manière indépendante et la variable cible, une méthode de filtrage comme l'information mutuelle est recommandée.

Justification : Ces méthodes évaluent l'importance de chaque caractéristique en mesurant la dépendance statistique entre les caractéristiques et la variable cible. L'information mutuelle peut saisir tout type de dépendance statistique (pas seulement linéaire), et les coefficients de corrélation sont efficaces pour identifier les relations linéaires, ce qui peut être particulièrement utile pour des variables continues comme la température et l'humidité.

Un membre X de votre équipe Data Science propose la solution suivante pour la sélection de caractéristiques dans l'ensemble de données de location de vélos :

- Commencez par construire un modèle d'arbre de décision simple en utilisant l'ensemble des caractéristiques de l'ensemble de données. Entraînez le modèle sur les données pour apprendre à prédire la variable cible `number_of_rentals`.
- Une fois entraîné, le modèle d'arbre de décision peut fournir une liste classée des importances des caractéristiques. Ce classement est basé sur le nombre de fois qu'une caractéristique est utilisée pour diviser les données et sur l'amélioration qu'elle apporte à chaque utilisation. L'importance d'une caractéristique est calculée comme la réduction totale (normalisée) du critère (par exemple, l'impureté de Gini) apportée par cette caractéristique.
- Une fois que vous avez l'importance des caractéristiques, définissez une valeur seuil. Les caractéristiques dont les valeurs d'importance sont supérieures à ce seuil sont sélectionnées, tandis que les autres sont ignorées. Par exemple, si les valeurs d'importance sont comprises entre 0 et 1, vous pouvez définir un seuil à 0,05. Cela signifierait que vous n'êtes intéressé que par les caractéristiques qui contribuent au moins 5% au processus de prise de décision du modèle.

(c) (2 points) Énoncez les **deux** avantages de la méthode ci-dessus proposée par le membre.

Solution: Deux avantages de la méthode de sélection de caractéristiques proposée par le membre X sont :

1. Cette méthode fournit une interprétation directe de l'importance des caractéristiques car elle est directement liée au processus de prise de décision du modèle. Elle est claire et intuitive, aidant à comprendre les divisions de l'arbre de décision.

2. Elle prend intrinsèquement en compte l'interaction entre les caractéristiques puisqu'un arbre de décision évalue la contribution de chaque caractéristique dans le contexte du branchement de l'arbre. Cela peut révéler des relations et des dépendances complexes qui pourraient ne pas être capturées par des évaluations de caractéristiques individuelles.

(d) (3 points) Parmi les méthodes suivantes, laquelle catégoriseriez-vous la méthode ci-dessus proposée par le membre X? **Encerclez** le bon choix. **Justifiez** votre choix ci-dessous.

`filtrage` `encapsulage` `embarquée`

Solution: La méthode proposée par le membre X serait catégorisée comme **embarquée**.

Justification : Les méthodes embarquées impliquent la sélection de caractéristiques comme partie intégrante du processus de formation du modèle. Puisque l'arbre de décision effectue intrinsèquement la sélection de caractéristiques pendant sa construction en évaluant les importances des caractéristiques, cette méthode n'est pas séparée de la formation du modèle (contrairement aux méthodes de filtre) et ne s'enroule pas autour (contrairement aux méthodes de wrapper). Elle est intégrée dans le processus d'apprentissage du modèle.

4. **Détection d'exemples abérants**

   (a) (2 points) Considérons les données brut (Fig 3a), dans quelle.s queue.s de la distribution sont les exemples abérants:

   Droite : [x] ; Gauche : [ ] ; Les deux : [ ] ; Aucune des Deux : [ ]

   Expliquer:

Solution: Cette distribution ne contient que des échantillons positifs (la population ne peut pas être négative). Avec cette visualisation, les valeurs extrêmes que `population` peut prendre sont de grandes valeurs positives correspondant au côté droit de la distribution.

   (b) (4 points) Considerons maintenant les données transformées (Fig.3b-3e) à quelle transformation de l'entrée correspondent chaque visualisation:

   $\sqrt{x}$: [c] ; $\log_{10}(1+x)$: [e] ; $1/(1+x)$: [b] ; $\log_{10}(.001+x)$: [d]

   Expliquer:

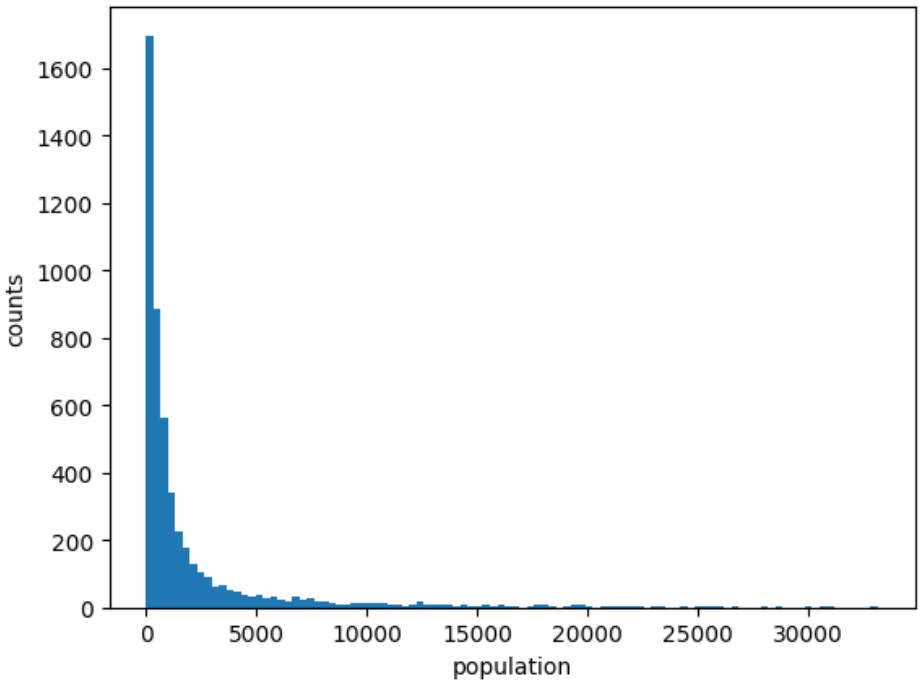
Solution: Il existe de nombreuses façons d'expliquer cela.

- $T(x) = \frac{1}{1+x}$ correspond à la transformation 1 (Fig 2b) car la distribution de sortie est bornée entre 0 et 1.
- $T(x) = \sqrt{x}$ correspond à la transformation 2 (Fig 2c), une façon de le voir est que l'étendue de la distribution est maintenant approximativement $[0, 175]$ ce qui correspond à la racine carrée de l'étendue de la distribution originale $[0, 30000]$.
- $T(x) = \log(1+x)$ correspond à la transformation 4 (Fig 2e), une façon de le voir est que le minimum est 0 ce qui correspond à $\log(1+0)$. Les transformations 3 et 4 sont celles qui restreignent le plus l'étendue comme on pourrait s'y attendre d'une transformation logarithmique.
- $T(x) = \log(.001+x)$ correspond à la transformation 3 (Fig 2d), une façon de le voir est que le minimum est -6 ce qui correspond à $\log_e(.001)$ (erratum : c'était $\log_e(0.001+x)$). Les transformations 3 et 4 sont celles qui restreignent le plus l'étendue comme on pourrait s'y attendre d'une transformation logarithmique.

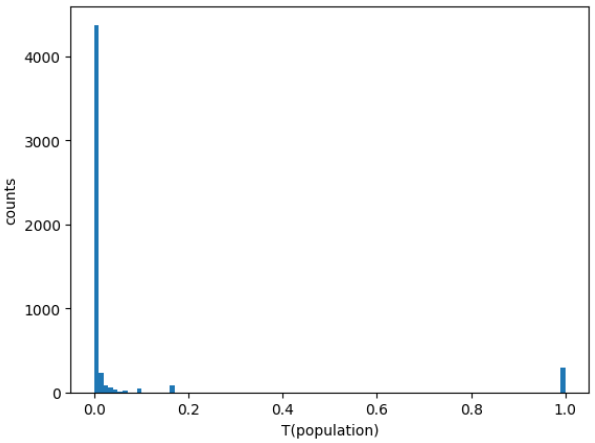
   (c) (2 points) Si l'on considére la transformation 3 (Fig. 3d), dans quelle.s queue.s de la distribution sont les exemples abérants:

   Droite : [ ] ; Gauche : [x] ; Les deux : [ ] ; Aucune des Deux : [ ]

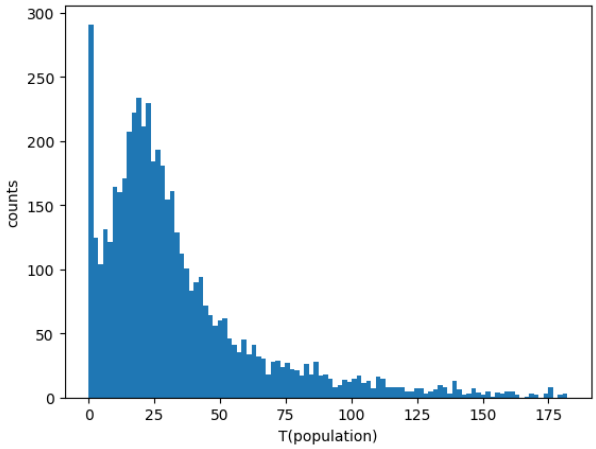
   Expliquer:



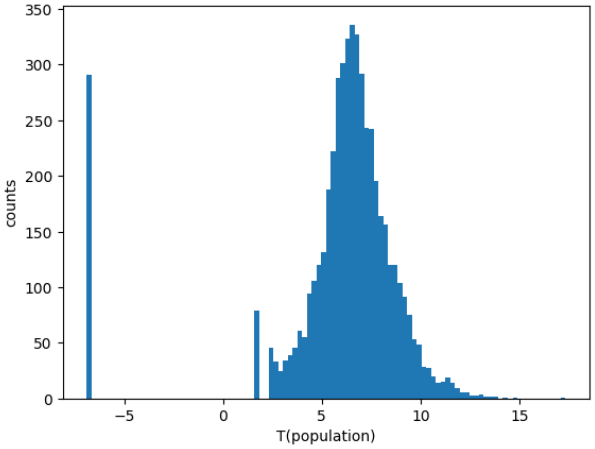
(a) Données brut



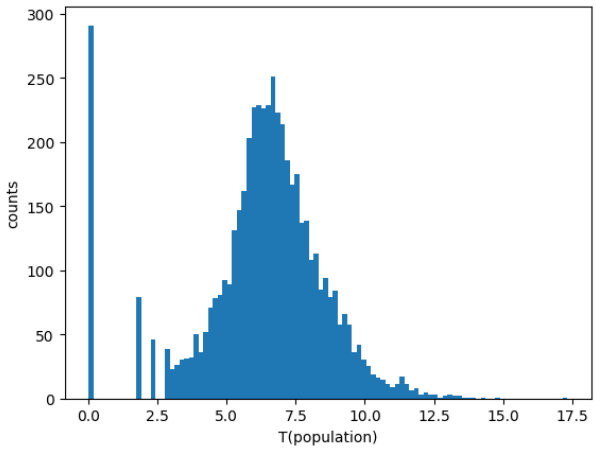
(b) Transformation 1



(c) Transformation 2



(d) Transformation 3



(e) Transformation 4

Figure 3: Histograme de la population des communes Canadiennes avant (Fig a) et apres transformation the l'entrée (i.e. la population). Chaque histogramme est composé de 100 fenêtres, la hauteur de chaque rectangle correspond au nombre de communes avec une population dans la fenêtre associée.

Solution: Après la transformation logarithmique, il semble qu'il n'y ait plus de grandes valeurs aberrantes (comme on pouvait s'y attendre d'une transformation logarithmique). D'autre part, il y a quelques valeurs très négatives (correspondant à des endroits où la population est égale à 0), ce qui implique que les valeurs aberrantes se trouvent maintenant sur la gauche.

(d) (2 points) Si l'on considére la transformation 3 (Fig. 3d), quelle methode est-il préférable d'utiliser pour détecter les exemples abbérants):

IQR : [x] ; T-test : [ ] ; Z-test : [ ] ; STD-test : [ ]

Expliquer:

Solution: Le nombre de valeurs aberrantes est important, donc l'estimation de la moyenne sera affectée par ces valeurs aberrantes. D'autre part, l'estimation de la médiane et des quartiles ne sera pas affectée par ces valeurs aberrantes, il est donc préférable d'utiliser l'IQR (Intervalle Interquartile).

(e) (3 points) Dans le contexte de ce jeux de données, doit-on supprimer les exemples abérants identifiés dans:

Fig 3a : [ ] ; Fig 3d : [ ] ; Les deux : [ ] ; Aucune des Deux : [ ]

Expliquer:

Solution: Plusieurs réponses étaient correctes tant qu'elles étaient correctement justifiées. Selon la tâche, les deux extrêmes (pas de population ou population très importante) pourraient être considérés comme des valeurs aberrantes et devraient être supprimés.

5. **Test d'hypothèse** Vous analysez des traces de données provenant d'ascensions de ballons météo. Chaque échantillon $x_i \in \mathbb{R}^p$ donne des mesures du carbone atmosphérique à différentes altitudes. Une partie du capteur a été remplacée pendant le processus de collecte, et vous voulez tester si ce changement a affecté les lectures. Formellement, $x_i \sim F_A$ avant le changement, $x_i \sim F_B$ après le changement, et l'hypothèse nulle est $H_0 : F_A = F_B$.

   Étant donné que les données sont hautement non normales, on considère le test suivant: Calculez le ℓ^1-distance d_{ij} pour toutes les paires d'échantillons i, j. Pour chaque échantillon, on place une arête $e_{i,N(i)}$ entre l'échantillon et son plus proche voisin; ce qui nous donne au total n arêtes. On dit d'une arête qu'elle est "pure" si elle se lie a un échantillon du même lot (*Batch*). Le graphique produit par ces étapes est illustré par la figure 4.

   (a) (3 points) Considérez la quantité:

$$T(x_1, \ldots, x_n) = \frac{\text{nombres d'arêtes pures}}{\text{nombre d'arrête}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{e_{i,N(i)} \text{ est pure}\right\}.$$

   Décrivez une stratégie pour approximer la distribution de T sous l'hypothèse nulle.

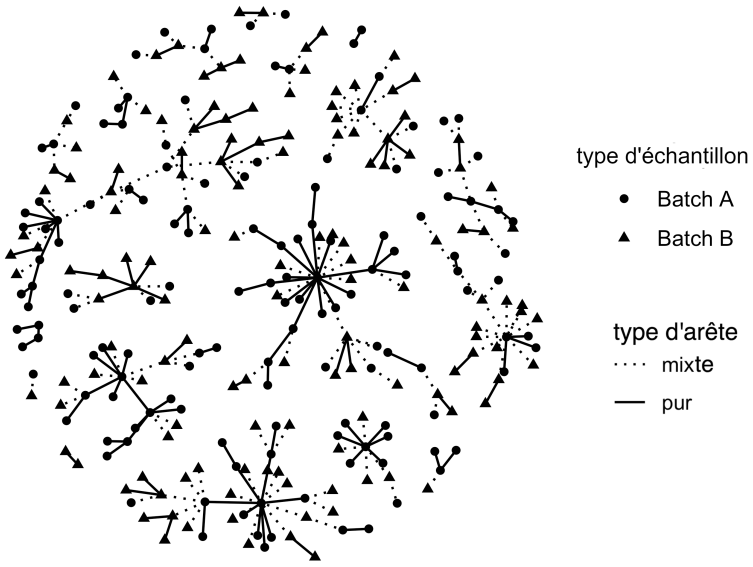


Figure 4: Graphe de voisinage utilisé pour tester si les lots peuvent être analysés ensemble.

Solution: L'hypothèse nulle correspond au fait que les lots A et B proviennent de la même distribution. Ainsi, sous l'hypothèse nulle, les étiquettes A et B sont indépendantes de l'entrée (elles sont simplement attribuées de manière aléatoire). Par conséquent, la distribution de T sous l'hypothèse nulle peut être approximée par un *mélange des étiquettes*, ce qui signifie que, étant donné n points de données, nous tirons σ uniformément sur l'ensemble des permutations de $\{1, \ldots, n\}$ et attribuons à x_i l'étiquette (A ou B) de $x_{\sigma(i)}$.

(b) (3 points) Décrivez ce que vous traceriez comme visualisation ainsi que ce que vous vous attendreriez à voir.

Solution: Pour chaque permutation des étiquettes, nous pouvons calculer T et ainsi construire un histogramme qui approximerait la distribution de T sous l'hypothèse nulle. À partir de cet histogramme (qui devrait approximativement ressembler à une gaussienne), nous pouvons estimer la probabilité d'observer un événement *au moins aussi extrême que* $T(x_1, x_2, \ldots, x_n)$ sous l'hypothèse nulle.

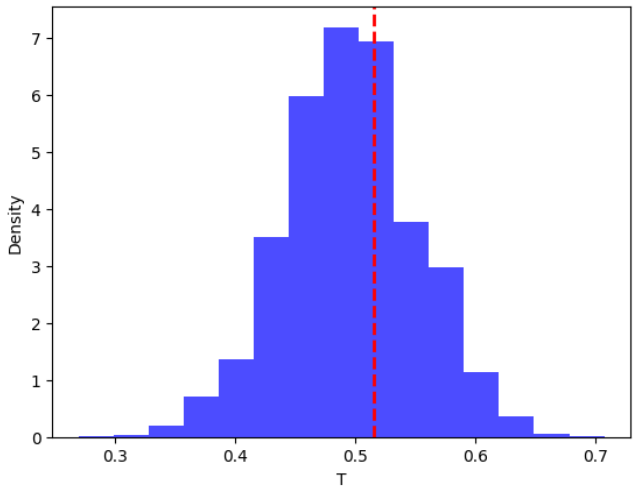


Figure 5: Exemple illustratif de ce que l'on pourrait observer. Dans ce cas, on rejetterait l'hypothèse nulle car l'événement observé a une probabilité très faible.

Vous pouvez trouver le code qui a produit la figure ici : https://colab.research.google.com/drive/1wc7ht52N9l
JypL9rEup53H?usp=sharing[colab]

(c) (3 points) Sous l'hypothèse nulle, quelle serait la valeur moyenne de $T(x_1, \ldots, x_n)$?

Valeur moyenne : 1/2

Solution:

Réponse attendue: sous l'hypothèse nulle, il est tout aussi probable qu'une arête soit pure ou mixte. Par conséquent, la valeur attendue de T serait de 1/2.

Réponse plus subtile (qui pourrait rapporter des points bonus) :

- Nous devons supposer que les lots A et B correspondent respectivement à n_A et n_B échantillons. Sous l'hypothèse nulle, étant donné un point x_i appartenant au lot A, la probabilité que ses voisins les plus proches soient du même lot (c'est-à-dire que l'arête soit pure) que x_i dépendra de $n_A - 1$ et n_B.
  Plus précisément, étant donné x_1 fixe et $X_2, \ldots, X_{n_A+n_B} \sim p_d$, nous définissons $\varphi(x) := \mathbb{P}_{X_i \sim p_d}(\|x_1 - X_i\| \geq x)$, $i \geq 2$. Par l'indépendance de $X_2, \ldots, X_{n_A}$, nous avons que $\mathbb{P}_{X_i \sim p_d}(\min_{2 \leq i \leq n_A} \|x_1 - X_i\| \geq x) = \mathbb{P}_{X_i \sim p_d}\|x_1 - X_i\| \geq x\,, \forall 1 \leq i \leq n_A) = \varphi(x)^{n_A - 1}$. Ainsi, la densité de la variable aléatoire $Y := \min_{n_A+1 \leq i \leq n_B+n_A} \|x_1 - X_i\|$ est $(\mathbf{P}(Y < x))'$ qui corre-

spond à $p(x) = (\mathbb{P}(\min_{n_A+1\leq i\leq n_B+n_A} \|x_1 - X_i\| < x))' = (1-\varphi(x)^{n_B})' = -n_B\varphi'(x)\varphi(x)^{n_b-1}$.

$$\mathbb{E}[T] := \mathbb{P}_{X_i\sim p_d}(\min_{2\leq i\leq n_A} \|x_1 - X_i\| \geq \min_{n_A+1\leq i\leq n_B+n_A} \|x_1 - X_i\|) \quad (1)$$

$$= \int_{x=0}^{\infty} p(x)\varphi(x)^{n_A-1}dx \quad (2)$$

$$= -\int_{x} n_B\varphi'(x)\varphi(x)^{n_A-1+n_B-1}dx \quad (3)$$

$$= -\frac{n_B}{n_A+n_B-1}[\varphi(x)^{n_A+n_B-1}]_0^{\infty} = \frac{n_B}{n_A+n_B-1} \quad (4)$$

où nous utilisons le fait que $\varphi(\infty) = 0$ et $\varphi(0) = 1$.
En particulier, pour $n_A = n_B = n$ nous obtenons que $\mathbf{E}[T] = \frac{1}{2-\frac{1}{n}} \approx \frac{1}{2}$.

(d) (3 points) Pour quelles valeurs de cette statistique rejetteriez-vous l'hypothèse nulle ? (vous pouvez dessiner un archétype de la visualisation obtenue en question (b) pour illustrer votre réponse)

Solution: Cela dépend

(e) (3 points) Si vous rejetez l'hypothèse nulle, F_A et F_B doivent avoir des moyennes différentes. (**Cochez** la bonne case)

VRAI : ☐ FAUX : ☐

Si vrai, donnez un **argument conceptuel**, si faux, **fournissez un contre-exemple**.

6. **Diagnostique des courbes d'apprentissage** Dans cette question, l'étudiant doit sélectionner le problème le plus probable qui est à l'origine du comportement sur le graphique d'apprentissage.

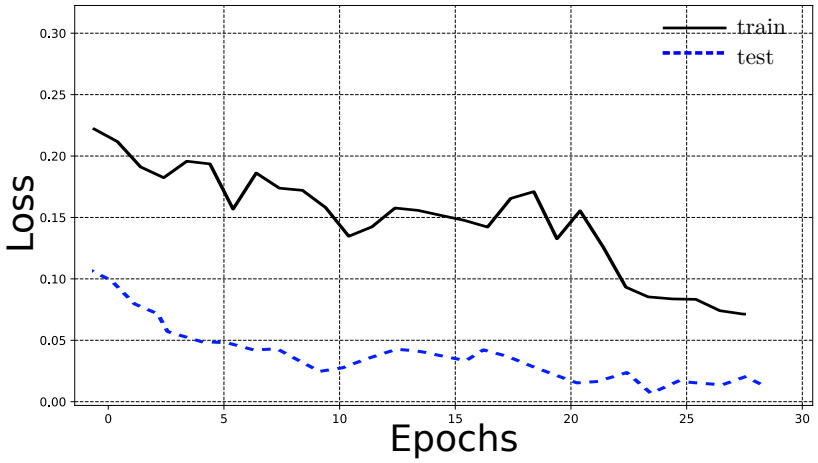


Figure 6: Sous-apprentissage : [ ] ; Sur-apprentissage : [ ] ; Ensemble de test trop petit : [ ] ; Ensemble d'entrainement trop petit : [ ] Ensembles d'entraînement et de test différents : [ ] Ensemble de test trop facile : [x]

(a) (2 points) Justifier la case cochée pour Figure 6:
La perte pour l'ensemble de test est constamment inférieure à celle de l'ensemble d'entraînement. Cela indique que l'ensemble de test est plus simple, plus petit que l'ensemble d'entraînement et très probable.

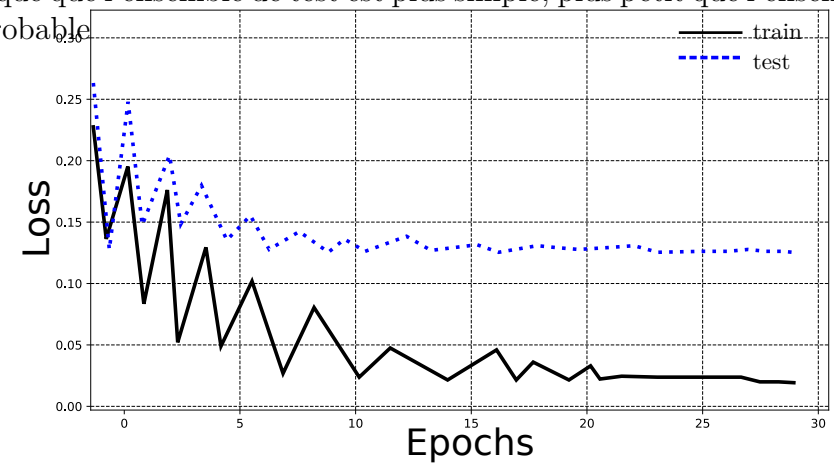


Figure 7: Sous-apprentissage : [ ] ; Sur-apprentissage : [x] ; Ensemble de test trop petit : [x] ; Ensemble d'entrainement trop petit : [ ] Ensembles d'entraînement et de test différents : [ ] Ensemble de test trop facile : [ ]

(b) (2 points) Justifiez la case cochée pour la Figure 7:
La perte sur l'ensemble d'entraînement est facile à optimiser et bruyante. De plus, la performance sur l'ensemble d'entraînement ne se traduit pas par des améliorations sur l'ensemble de test.

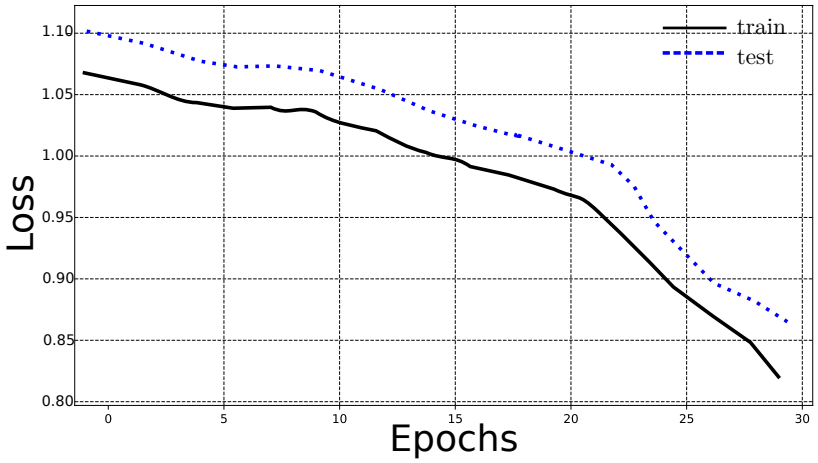


Figure 8: Sous-apprentissage : [x] ; Sur-apprentissage : [ ] ; Ensemble de test trop petit : [ ] ; Ensemble d'entrainement trop petit : [ ] Ensembles d'entraînement et de test différents : [ ] Ensemble de test trop facile : [ ]

(c) (2 points) Justifier la case cochée pour la Figure 8: La perte pour les ensembles d'entraînement et de test continue de diminuer.

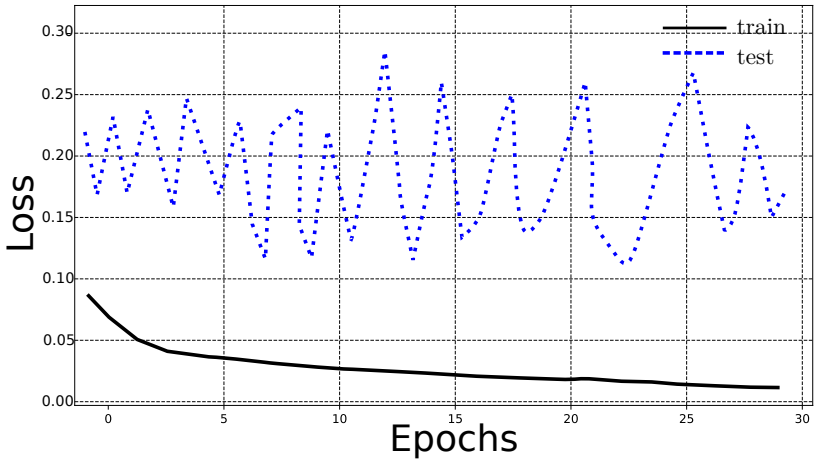


Figure 9: Sous-apprentissage : [ ] ; Sur-apprentissage : [ ] ; Ensemble de test trop petit : [ ] ; Ensemble d'entrainement trop petit : [ ] Ensembles d'entraînement et de test différents : [x] Ensemble de test trop facile : [ ]

(d) (2 points) Justifier la case coché pour la Figure 9:
La perte d'entraînement converge, mais elle n'a pas de corrélation avec la performance du test.

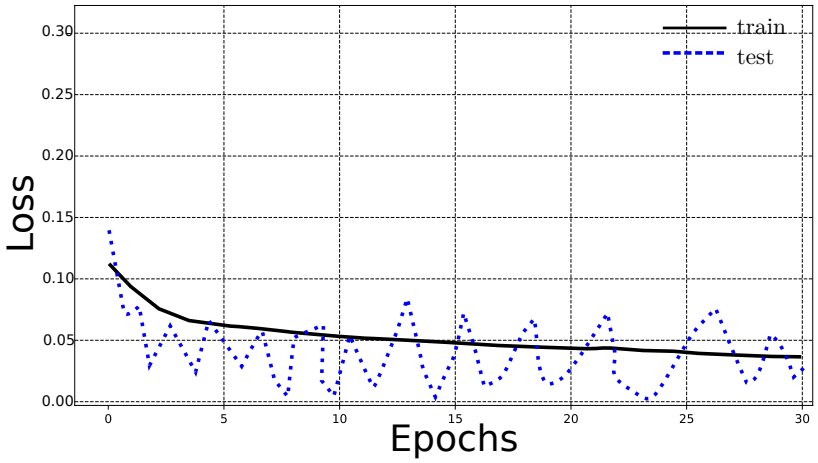


Figure 10: Sous-apprentissage : ☐ ; Sur-apprentissage : ☐ ; Ensemble de test trop petit : ☒ ; Ensemble d'entrainement trop petit : ☐ Ensembles d'entraînement et de test différents : ☐ Ensemble de test trop facile : ☐

(e) (2 points) Justifier la case cochée pour la Figure 10:

La perte de test est corrélée avec la perte d'entraînement, mais la perte de test change radicalement entre les époques, indiquant que le modèle alterne entre l'identification correcte de classes spécifiques.

7. **Hyperparametres vs parametres**

(a) (2 points) Definir les paramètres.

Solution: Les paramètres du modèle sont des quantités qui sont *apprises* à partir des données. Exemples : les poids d'un réseau de neurones, les vecteurs de support dans une SVM ou les coefficients d'un modèle de régression.

(b) (2 points) Definir les hyper-paramètres.

Solution: Les hyperparamètres déterminent la manière dont l'apprentissage progresse. Exemples : complexité du modèle (nombre de couches, taille des filtres, etc.), fonction de perte utilisée, taux d'apprentissage, méthode de régularisation.

(c) (2 points) Si nous avons un modèle avec d hyperparamètres, chacun avec m configurations différents, combien de combinaisons différentes d'hyperparamètres existe-t-il ? Mettez votre réponse finale dans cette case $\boxed{m^d}$.

Solution: L'espace de recherche augmente de manière exponentielle avec le nombre d'hyperparamètres. Cela augmentera le temps et les ressources nécessaires pour trouver l'ensemble optimal d'hyperparamètres.

(d) (3 points) Expliquez le concept de validation croisée et son rôle dans la sélection des hyperparamètres. Comment la validation croisée aide-t-elle à sélectionner les hyperparamètres optimaux ?

Solution: La validation croisée est une méthode qui utilise différentes parties des données d'entraînement pour entraîner et évaluer un modèle. Notez que différentes méthodes existent pour déterminer comment réaliser un tel partage. Par exemple, dans la validation croisée k-Fold, l'ensemble d'entraînement est divisé en k sous-ensembles, et un modèle est entraîné sur $k-1$ sous-ensembles et évalué sur le dernier sous-ensemble. La procédure se répète pour chacun des k plis.

La validation croisée est importante car elle empêche d'ajuster les hyper-paramètres pour qu'ils fonctionnent de manière optimale sur un seul ensemble de validation, ce qui risque de surajuster.

(e) (4 points) Vous trouverez ci-dessous un tableau contenant des algorithmes de sélection des hyperparamètres ainsi que des propriétés potentielles de ces algorithmes. Cochez les cases qui sont correctes.

Type	Peut trouver l'optimum	parallélisable	Utilise les itérations précédentes
Tableaux orthogonaux		X	
Recherche par grille	X	X	
Échantillonnage bayésien	X		X
Hypercube latin		X	

Table 1: Algorithmes de sélection des hyperparamètres et propriétés potentielles.

Solution:
L'échantillonnage bayésien et la parallélisation – il est possible de paralléliser (partiellement), mais c'est plus difficile que les autres méthodes.

Autres méthodes et parallélisation – ne dépendent pas des exécutions précédentes, peuvent donc être parallélisées.

Les tableaux orthogonaux restreignent l'espace de recherche à un nombre fini.

Les méthodes qui utilisent l'échantillonnage peuvent ne pas trouver la solution optimale si l'échantillonnage de l'espace n'est pas optimal (souvent le cas avec la recherche en grille).

8. **Apprentissage non-supervisé**

(a) (3 points) Sélectionnez dans la liste des exemples d'apprentissage non supervisé.

Clustering : [X] ; Réglage des hyperparamètres : [ ] ; Classification : [ ] ;

Réduction de dimensionnalité : [X] Régression : [ ] Standardisation : [ ]

Expliquer:

Solution:
Les méthodes de regroupement (clustering) n'utilisent pas d'étiquettes tandis que les tâches de classification nécessitent des étiquettes.

La régression est une méthode supervisée qui prédit des valeurs réelles à partir de données.

Le réglage des hyperparamètres n'est pas un algorithme d'apprentissage, c'est une méthode permettant de déterminer comment l'apprentissage progresse.

La réduction de la dimensionnalité vise à trouver la représentation latente de faible dimension des données.

La standardisation est une méthode de traitement / mise à l'échelle des données par laquelle les données sont centrées autour d'une moyenne de 0 avec une variance unitaire.

(b) (3 points) L'apprentissage non-supervisé:

A toujours pour ojectif d'obtenir de meilleurs représentations : [X] ; N'utilise jamais les étiquettes des données : [X^*] ; Ne necéssite pas de prétraitement : [ ] ;

Permet de trouver des structures interprétables : [ ]

Expliquer:

Solution: L'apprentissage non supervisé forme un modèle lorsque nous avons seulement les entrées x_i, en comparaison à l'apprentissage supervisé qui utilise des paires entrée-sortie (x_i, y_i).

Tout type de données peut bénéficier d'un prétraitement, tel que le nettoyage des données, la normalisation, la standardisation, etc.

$*$ On pourrait argumenter que les étiquettes de données sont utilisées pour évaluer des méthodes

telles que le regroupement.

9. **Programmation et Débogage**

   Vous avez été approché par un.e collègue qui a des difficultés à déboguer du code Python qu'il/elle a écrit pour un projet de tâche en science des données. Le but est d'utiliser un modèle d'arbre de décision pour calculer l'importance des caractéristiques et ensuite entraîner un modèle de régression linéaire pour estimer la variable cible. Identifiez les bogues dans le code ci-dessous, en expliquant chaque bogue et comment le corriger.

   Vous obtiendrez +1 point pour chaque bogue correctement identifié, et +1 pour avoir identifié la correction appropriée. Si une ligne a plusieurs erreurs, veuillez indiquer chaque erreur séparément.

   Supposez que tous les modules nécessaires ont été importés correctement. Vous ne devez pas introduire de nouvelles variables, méthodes ou lignes de code. Il est possible de déplacer des lignes de code à un autre numéro de ligne et cela est compté comme un seul bogue.

```
# Charger les données
data = pandas.read_csv(data_path)

# Extraire les caractéristiques et la variable cible
data_target = data['nombre_de_locations']
data_attrs = data.drop(columns=['nombre_de_locations'])

# Division des données en ensemble d'entraînement et de test sans préciser la taille du test
X_train, X_test, y_train, y_test = train_test_split(data_attrs, data)

# Utilisation d'un arbre de décision pour calculer l'importance des caractéristiques
rf = DecisionTreeRegressor()
rf.fit(X_train, y_train)
feature_importances = rf.feature_importances_

# Sélection des 5 meilleures caractéristiques en utilisant np.argsort() qui trie un tableau par ordre croissant et renvoie les indices dans l'ordre trié
top_features = data.columns[np.argsort(feature_importances)[:5]]

# Mise à l'échelle des caractéristiques
X_train_scaled = pd.DataFrame(StandardScaler().fit(X_train), columns=data.columns)

# Utilisation uniquement des meilleures caractéristiques pour entraîner le modèle
X_train_selected = X_train_scaled[top_features]
X_test_selected = X_test[top_features]

# Entraînement d'un modèle de régression linéaire
model = LinearRegression()
model.fit(X_train_selected, y_test)  # Utilisation de y_test au lieu de y_train
predictions = model.predict(X_test)

# Calculer la RMSE
rmse = numpy.sqrt(mean_squared_error(y_test, predictions))
print(rmse)

```

Écrivez votre solution ici. Indiquez le numéro de ligne pour chaque bogue.

Solution:

Ligne 9 :

Erreur : La fonction `train_test_split` est appelée avec `data` comme second argument, qui contient à la fois les caractéristiques et la variable cible.

Correction : Changez le second argument en `data_target` pour représenter correctement la variable cible.

```
X_train, X_test, y_train, y_test = train_test_split(data_attrs, data_target)
```

Ligne 17 :

Erreur : La mise à l'échelle des caractéristiques est effectuée après le calcul de l'importance des caractéristiques, ce qui devrait être fait avant l'ajustement de l'arbre de décision.

Correction : Déplacez le bloc de code pour la mise à l'échelle des caractéristiques (Lignes 17-18) avant l'ajustement de l'arbre de décision (Ligne 13).

Ligne 18 :

Erreur : Création incorrecte d'un DataFrame après l'ajustement du StandardScaler. La méthode `fit` est utilisée sans la méthode `transform`.

Correction : Utilisez la méthode `fit_transform` pour mettre à l'échelle `X_train` et corrigez les colonnes du DataFrame pour utiliser `data_attrs.columns` au lieu de `data.columns`.

```
X_train_scaled = pd.DataFrame(StandardScaler().fit_transform(X_train),
↪  columns=data_attrs.columns)
```

Ligne 23

Erreur : Les données `X_test` ne sont pas mises à l'échelle en utilisant le même scaleur ajusté sur `X_train`.

Correction : Mettez à l'échelle `X_test` en utilisant le scaleur déjà ajusté avant de sélectionner les meilleures caractéristiques.

```
X_test_scaled = pd.DataFrame(scaler.transform(X_test), columns=data_attrs.columns)
X_test_selected = X_test_scaled[top_features]
```

Ligne 27 :

Erreur : `model.fit` est appelé avec `y_test` au lieu de `y_train`.

Correction : Changez `y_test` en `y_train` pour entraîner correctement le modèle de régression linéaire.

```
model.fit(X_train_selected, y_train)
```

Ligne 28 :

Erreur : Le modèle fait des prédictions sur `X_test`, qui inclut toutes les caractéristiques, pas seulement celles sélectionnées.

Correction : Changez `X_test` en `X_test_selected` pour faire des prédictions en utilisant seulement les caractéristiques sélectionnées.

```
predictions = model.predict(X_test_selected)
```