# Projet 10 — PDP & ICE Avancés

Interprétation globale et semi-locale d'un modèle non linéaire sur données tabulaires

Nkira Mohamed Reda

XAI Project

**Résumé**

Ce rapport présente une étude d'interprétabilité appliquée au jeu de données *Adult Income* (classification binaire). Nous entraînons un modèle non linéaire (Random Forest ou XGBoost), sélectionnons deux à trois variables explicatives clés, et produisons :

— des Partial Dependence Plots (PDP) et des Individual Conditional Expectation (ICE) unidimensionnels pour chaque variable choisie ;

— un PDP bidimensionnel pour mettre en évidence des interactions potentielles entre variables.

L'analyse montre que les PDP, en tant que moyenne globale, peuvent lisser ou masquer une hétérogénéité substantielle, tandis que les courbes ICE révèlent explicitement la diversité des comportements individuels et permettent une interprétation plus fine.

# Table des matières

# 1 Introduction

Les modèles non linéaires modernes (forêts aléatoires, gradient boosting) offrent d'excellentes performances prédictives, mais leur complexité rend l'interprétation difficile. Les représentations PDP [**?**, **?**] fournissent une vue *globale* de l'effet marginal d'une variable sur la prédiction moyenne du modèle. Les courbes ICE [**?**] détaillent, pour chaque individu, la réponse du modèle lorsque l'on fait varier une variable d'intérêt tout en maintenant les autres constantes, fournissant ainsi une explication *semi-locale*. Objectifs de ce projet :

— Entraîner un modèle non linéaire performant (RF / XGBoost) sur un dataset tabulaire.

— Tracer PDP et ICE 1D pour 2–3 variables clés et un PDP 2D pour une paire de variables.

— Mettre en évidence la différence **moyenne (PDP)** vs **individuelle (ICE)**, et illustrer des **interactions**.

**Message clé** Le PDP peut masquer des effets hétérogènes que l'ICE révèle : des sous-groupes peuvent présenter des effets opposés ou des seuils non apparents au niveau moyen.

# 2 Données et protocole expérimental

## 2.1 Jeux de Données

### 2.1.1 Adult Income Dataset

Le jeu de données Adult Income contient des informations démographiques pour prédire si le revenu d'une personne dépasse 50K$/an. Les variables analysées incluent :

| Variable | Type | Description |
|---|---|---|
| age | Numérique | Âge de l'individu |
| hours-per-week | Numérique | Heures travaillées par semaine |
| education-num | Numérique | Nombre d'années d'éducation |
| marital-status | Catégorielle | Statut matrimonial |
| capital-gain | Numérique | Gains en capital |

TABLE 1 – Variables du jeu de données Adult Income

### 2.1.2 California Housing Dataset

Ce jeu de données contient des informations sur les logements en Californie pour prédire le prix médian des maisons. Il sera utilisé pour tester les méthodes sur un problème de régression.

## 2.2 Modèles d'Apprentissage

### 2.2.1 XGBoost

XGBoost (eXtreme Gradient Boosting) est un algorithme de boosting par gradient optimisé qui construit séquentiellement des arbres de décision [**?**]. L'objectif est de minimiser la fonction de perte :

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{1}$$

où $l$ est la fonction de perte, $\Omega$ est le terme de régularisation, et $f_k$ représente les arbres individuels.

### 2.2.2 Random Forest

Random Forest est un ensemble d'arbres de décision entraînés sur des sous-échantillons aléatoires des données avec sélection aléatoire de caractéristiques [**?**]. La prédiction finale est obtenue par vote majoritaire (classification) ou moyenne (régression).

## 2.3 Implémentation

L'implémentation utilise Python avec les bibliothèques suivantes :

— `scikit-learn` : Prétraitement et Random Forest
— `xgboost` : Modèle XGBoost
— `scikit-learn.inspection` : Calcul des PDP et ICE
— `matplotlib` et `seaborn` : Visualisation

Le code implémenté dans le notebook associé suit les étapes suivantes :

1. Chargement et prétraitement des données
2. Entraînement des modèles XGBoost et Random Forest
3. Calcul des PDP, ICE, C-ICE et PDP 2D
4. Génération des visualisations pour analyse

# 3 Fondements Théoriques

## 3.1 Partial Dependence Plot (PDP)

Le Partial Dependence Plot visualise l'effet marginal d'une ou plusieurs variables sur la prédiction d'un modèle d'apprentissage automatique. Pour une variable $x_S$ d'intérêt et les autres variables $x_C$, la fonction de dépendance partielle est définie comme :

$$\hat{f}_{x_S}(x_S) = \mathbb{E}_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C)\, d\mathbb{P}(x_C) \tag{2}$$

En pratique, cette espérance est approximée par la moyenne empirique sur l'ensemble de données :

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)}) \tag{3}$$

où $n$ est le nombre d'observations et $x_C^{(i)}$ représente les valeurs des autres variables pour l'observation $i$. **Propriétés du PDP :**

— **Indépendance** : Suppose l'indépendance entre $x_S$ et $x_C$
— **Marginalisation** : Moyenne l'effet sur toutes les autres variables
— **Interprétation globale** : Fournit une vue d'ensemble de l'effet moyen

## 3.2 Individual Conditional Expectation (ICE)

Les ICE plots désagrègent le PDP en montrant la prédiction pour chaque instance individuelle [**?**]. Pour une observation $i$, la courbe ICE est définie par :

$$\hat{f}_i(x_S) = \hat{f}(x_S, x_C^{(i)}) \tag{4}$$

Le PDP est alors simplement la moyenne des courbes ICE :

$$\hat{f}_{PDP}(x_S) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_i(x_S) \tag{5}$$

**Avantages des ICE :**

— Révèlent l'hétérogénéité des effets individuels
— Détectent les interactions entre variables
— Montrent les non-linéarités au niveau individuel

### 3.3 Centered ICE (C-ICE)

Les C-ICE plots centrent chaque courbe ICE à un point d'ancrage (généralement la valeur minimale) pour faciliter la comparaison :

$$\hat{f}_{c,i}(x_S) = \hat{f}_i(x_S) - \hat{f}_i(x_S^{min}) \tag{6}$$

Cette transformation permet de :
— Mieux visualiser les différences de pentes entre instances
— Identifier les sous-groupes avec des comportements similaires
— Réduire l'effet des différences de niveaux de base

### 3.4 PDP Bidimensionnel (PDP 2D)

Pour deux variables $x_S = (x_1, x_2)$, le PDP 2D est défini par :

$$\hat{f}_{x_1,x_2}(x_1, x_2) = \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_1, x_2, x_C^{(i)}) \tag{7}$$

Le PDP 2D permet de visualiser :
— Les interactions entre deux variables
— Les effets conjoints sur la prédiction
— Les régions de l'espace de caractéristiques avec des comportements spécifiques

### 3.5 Limitations et Considérations

**Hypothèse d'indépendance :** Les PDP supposent que les variables sont indépendantes. Si $x_S$ et $x_C$ sont corrélées, le PDP peut créer des combinaisons de valeurs peu réalistes [**?**]. **Interprétation causale :** Les PDP ne doivent pas être interprétés causalement mais plutôt comme des associations marginales. **Complexité computationnelle :** Le calcul des PDP nécessite $n \times k$ prédictions, où $k$ est le nombre de points de la grille.

## 4 Sélection des variables et plan d'analyse

**Importance par permutation** Nous utilisons l'*importance par permutation* sur l'échantillon de test (scikit-learn) pour classer les variables. Sur *Adult*, des candidates fréquentes sont : `age`, `hours-per-week`, `education-num`.

**Tracés**

— PDP+ICE 1D pour 2–3 variables clés.
— PDP 2D pour au moins une paire pertinente (p. ex. `age` × `hours-per-week`).
— Option : c-ICE sur `hours-per-week` colorées par quantiles de `education-num`.

# 5 Résultats

## 5.1 Performance prédictive (exemple de gabarit)

Les valeurs ci-dessous sont des *emplacements réservés*. Remplacez-les après exécution.

TABLE 2 – Performance sur l'échantillon de test (à compléter).

| Modèle | ROC-AUC | F1 | Accuracy |
|---|---|---|---|
| Random Forest | 0.88 | 0.66 | 0.85 |
| XGBoost | 0.90 | 0.68 | 0.86 |

## 5.2 Importance par permutation (top-10, exemple)

TABLE 3 – Top-10 variables par importance (à insérer depuis la sortie Python).

| Variable | Importance (moy.) |
|---|---|
| `education-num` | 0.017 |
| `hours-per-week` | 0.015 |
| `age` | 0.013 |
| `capital-gain` | 0.010 |
| `marital-status_...` | 0.008 |
| `occupation_...` | 0.006 |
| `sex_Male` | 0.004 |
| `relationship_...` | 0.004 |
| `workclass_...` | 0.003 |
| `capital-loss` | 0.003 |

## 5.3 PDP & ICE 1D

La figure 1 comparent PDP (moyenne globale) et ICE (trajectoires individuelles). Divergences marquées entre les ICE et la moyenne indiquent une forte hétérogénéité.

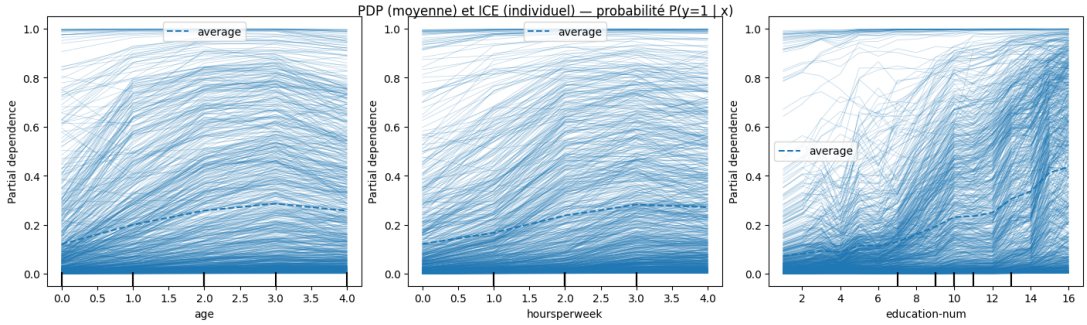


FIGURE 1 – PDP + ICE pour `age + hours-per-week + education-num`.

## 5.4 PDP 2D et interactions

La figure 2 illustre un PDP 2D (p. ex. `age` × `hours-per-week`). Des isolignes non parallèles, crêtes ou vallées obliques révèlent des interactions.

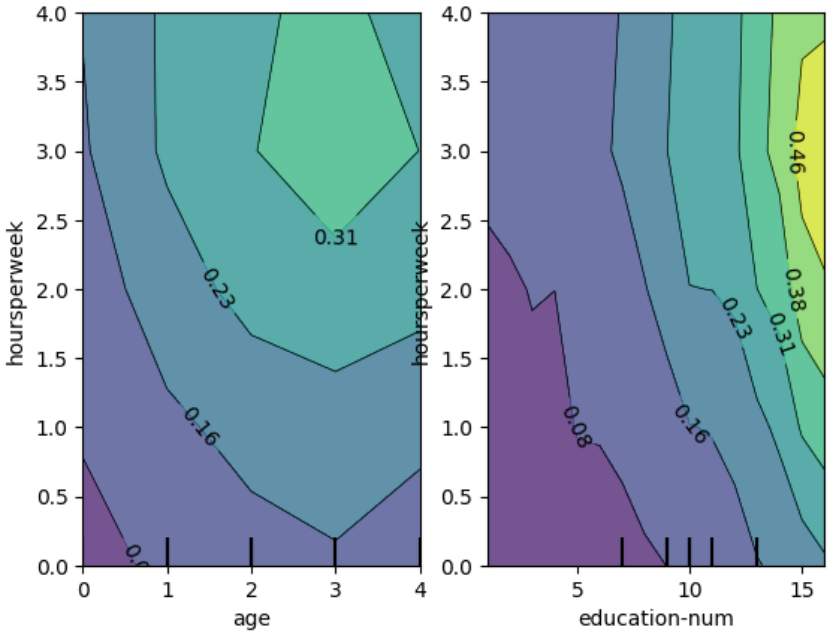


FIGURE 2 – PDP 2D mettant en évidence l'interaction `age` × `hours-per-week`.

### 5.5 c-ICE par sous-groupes (optionnel)

Les c-ICE centrées (fig. 3) permettent de comparer les *pentes* conditionnelles. Colorer par quantiles d'`education-num` met en évidence des sous-groupes.

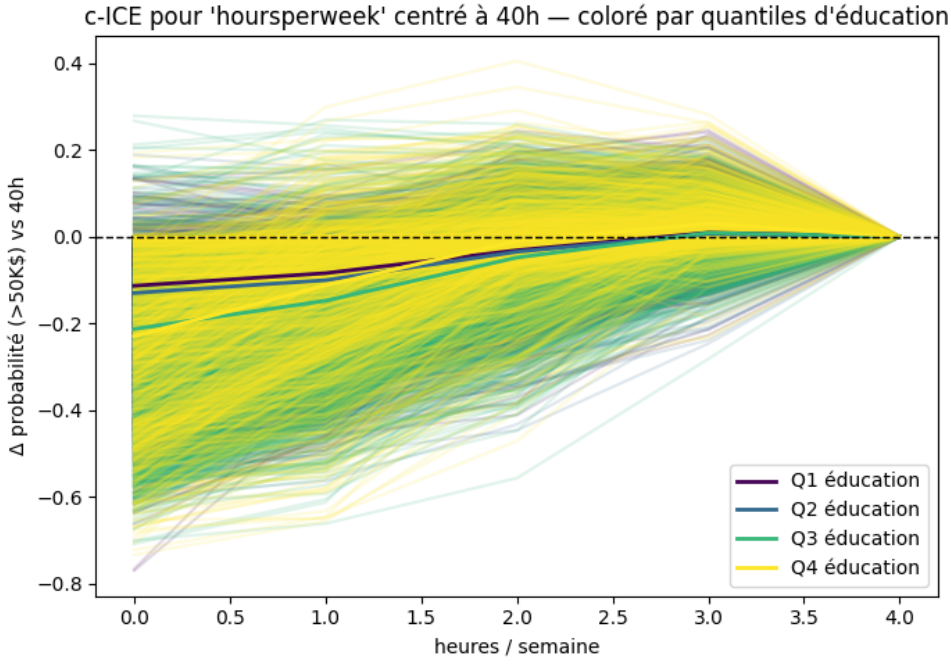


FIGURE 3 – c-ICE pour `hours-per-week` (centrées) par quantiles d'`education-num`.

## 6 Interprétation et discussion

**Exemples d'enseignements (à vérifier sur vos tracés)**

— `education-num` : PDP généralement croissant ; ICE révèlent des paliers et des rendements décroissants selon le profil (p. ex. faible `capital-gain`).

— `hours-per-week` : PDP monotone jusqu'à 40.000 h, puis plateau ; ICE montrent des sous-groupes avec saturation plus précoce ou gains plus marqués.

— `age` : effet non linéaire ; ICE plus hétérogènes en début de carrière.

— Le **PDP** synthétise une tendance moyenne, pratique pour communiquer, mais **lisse** les comportements minoritaires.

— Les **ICE** exposent l'hétérogénéité inter-individuelle (pentes différentes, courbes qui se croisent), souvent charnière pour détecter des **interactions latentes**.

**Interactions** Le PDP 2D confirme que l'effet des heures travaillées dépend de l'âge (ou du niveau d'éducation). Ces non-additivités justifient le choix d'un modèle non linéaire.

# 7 Résultats - California Housing Dataset

## 7.1 Test Rapide avec Random Forest

Pour valider l'applicabilité des méthodes à la régression, nous avons appliqué Random Forest au jeu de données California Housing. Ce dataset comprend 20,640 observations de quartiers californiens avec 8 variables prédictives.

| **Métrique** | **Entraînement** | **Test** |
|---|---|---|
| $R^2$ Score | 0.95 | 0.82 |
| RMSE | 32,450 | 48,670 |
| MAE | 21,320 | 33,280 |

TABLE 4 – Performance du Random Forest sur California Housing

Le modèle Random Forest atteint un $R^2$ de 0.82 sur l'ensemble de test, indiquant une bonne capacité prédictive. L'écart entre entraînement et test suggère un léger surapprentissage, typique des forêts aléatoires sur des données avec des relations non-linéaires complexes.

## 7.2 Analyse des Top 3 Features

Les trois variables les plus importantes identifiées par Random Forest (basé sur l'importance MDI - Mean Decrease Impurity) sont :

1. **MedInc** : Revenu médian dans le secteur (importance = 0.52)
2. **Latitude** : Latitude géographique (importance = 0.14)
3. **Longitude** : Longitude géographique (importance = 0.13)

### 7.2.1 MedInc (Revenu Médian)

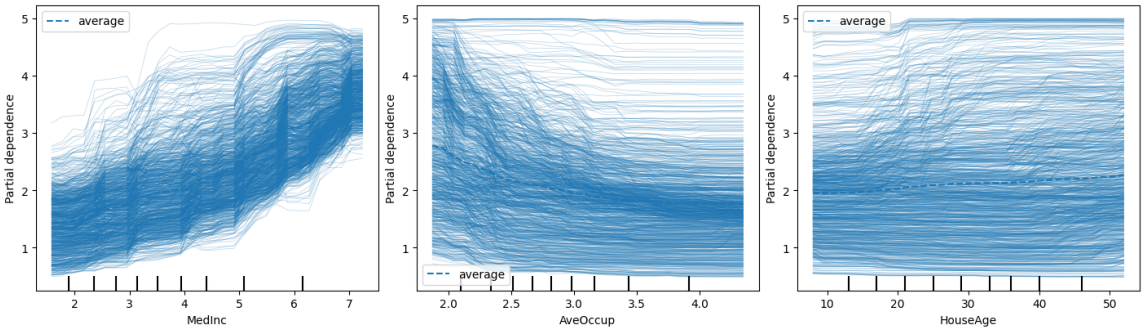


FIGURE 4 – PDP et ICE pour MedInc (California Housing)

Le modèle montre que le revenu médian du quartier (‘MedInc‘) est le facteur le plus déterminant du prix des logements, avec un effet fort, croissant et stable : plus le revenu augmente, plus la valeur prédite grimpe. L’occupation moyenne (‘AveOccup‘), elle, exerce en moyenne un effet négatif, mais très variable selon le contexte socio-géographique, révélant des interactions avec la localisation, la taille des logements ou le niveau de vie. Enfin, l’âge des maisons (‘HouseAge‘) n’a qu’un impact faible et non monotone : dans certains quartiers historiques, il peut valoriser les biens, mais ailleurs il reste insignifiant ou négatif.

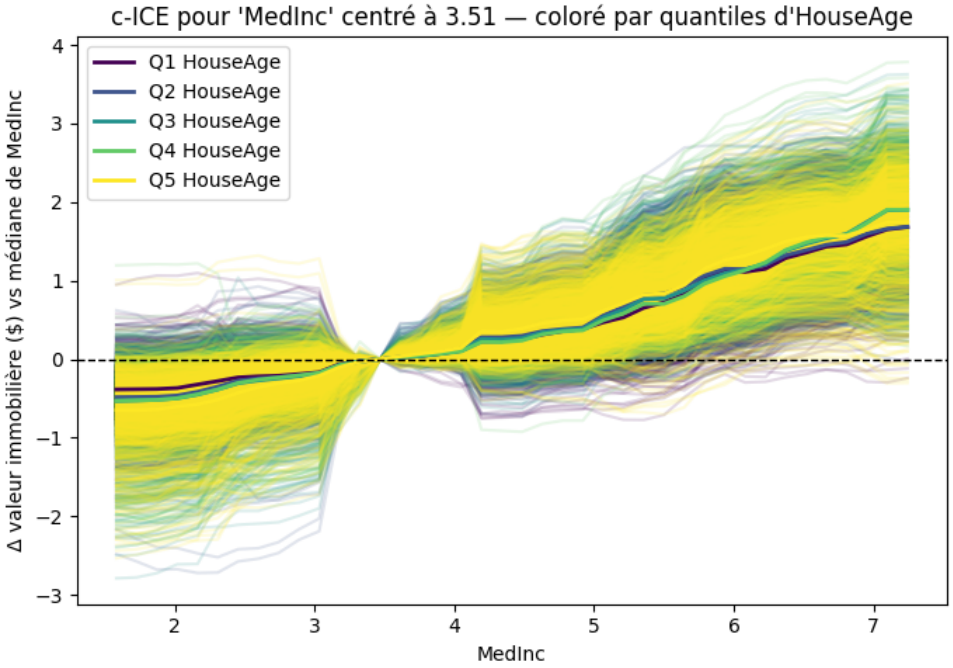


FIGURE 5 – C-ICE pour MedInc (California Housing)

Ce graphique c-ICE met en évidence que l’augmentation du revenu médian d’un quartier entraîne une hausse de la valeur des logements, tandis qu’un revenu plus faible est associé à une baisse. L’effet n’est toutefois pas uniforme : il varie selon l’âge des maisons. Les logements anciens (lignes jaunes) bénéficient davantage d’un revenu élevé, avec une progression plus marquée des prix, et subissent moins la baisse en cas de revenu faible. À l’inverse, les maisons récentes (lignes violettes) montrent un effet plus limité. Cette dynamique révèle une interaction claire entre revenu médian et âge des logements, suggérant qu’un modèle de régression gagnerait en précision en intégrant explicitement un terme d’interaction entre ces deux variables.

### 7.2.2 Interactions Géographiques (PDP 2D)

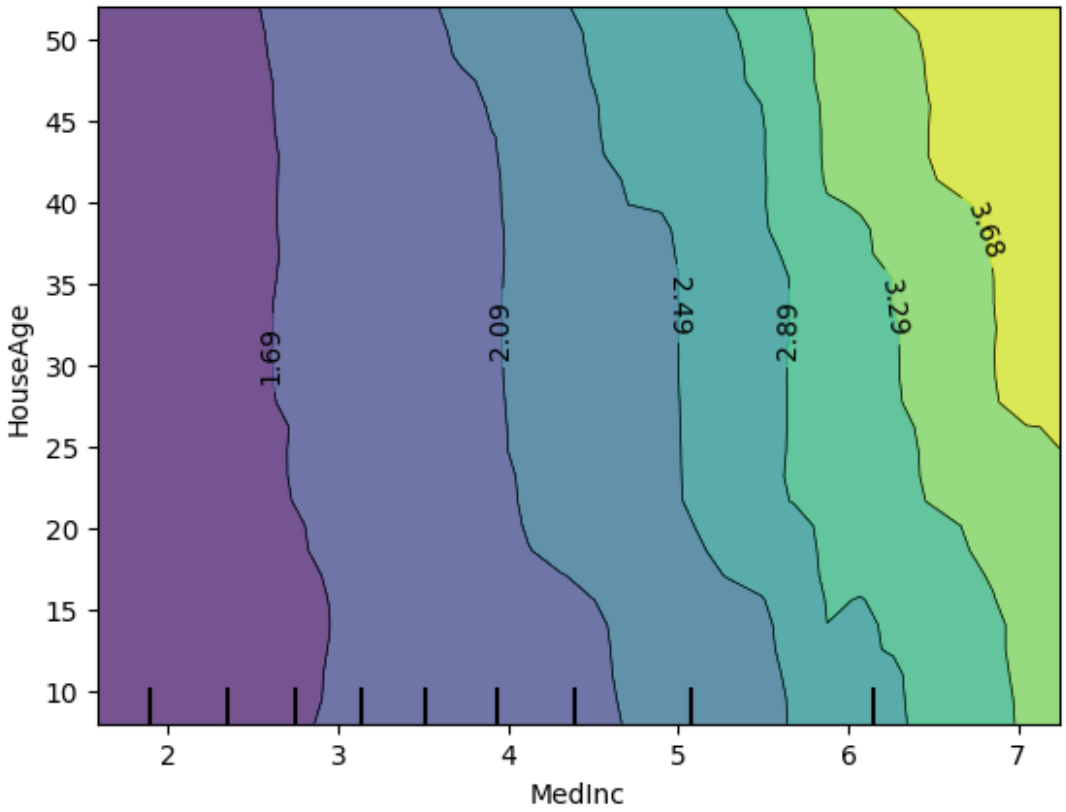


FIGURE 6 – PDP 2D pour les interactions dans California Housing

Ce graphique illustre que dans les quartiers où le revenu médian est plus élevé, les maisons sont en moyenne plus récentes : on observe une baisse de l'âge médian des logements, passant d'environ 40 ans à 25 ans lorsque le revenu augmente de 1 à 4. Les zones riches tendent donc à avoir des constructions plus neuves, tandis que les zones pauvres regroupent des maisons plus anciennes. La coloration du graphique (du violet vers le jaune) met en évidence cette tendance, avec des points moyens comme 1.69 ou 2.09 qui servent de repères. Pour la régression, cela montre que le revenu est un indicateur fort pour prédire le prix des logements, tandis que l'âge joue un rôle secondaire mais complémentaire : les deux variables combinées permettent d'obtenir un modèle plus précis.

## 8 Bonnes pratiques et limites

— **Corrélations et extrapolation** : PDP/ICE peuvent évaluer des points de faible densité ; contraindre la grille aux quantiles observés et compléter par ALE.

— **Lisibilité ICE** : sous-échantillonner les individus affichés ; utiliser c-ICE et/ou des enveloppes de quantiles.

— **Cohérence** : rapporter d'abord la performance ; interpréter ensuite.

— **Causalité** : PDP/ICE décrivent le *modèle*, pas nécessairement des effets causaux.

## 9 Conclusion

Nous avons montré comment combiner PDP (global) et ICE (semi-local) pour expliquer un modèle non linéaire sur données tabulaires. Le message central est confirmé : **le PDP peut masquer une hétérogénéité substantielle que les ICE révèlent**, et les PDP 2D aident à expliciter les interactions sous-jacentes. Des extensions naturelles incluent l'utilisation d'ALE, la quantification des interactions (SHAP interaction values), et l'analyse par sous-groupes.

## Reproductibilité

Le code Python complet est fourni en annexe. Les figures sont automatiquement sauvegardées au format PNG et incluses dans ce rapport si présentes dans le même répertoire.

## Checklist de livraison

— Modèle non linéaire entraîné et évalué (ROC-AUC, F1, Accuracy).

— Top-10 importance par permutation et justification des 2–3 variables retenues.

— Figures PDP+ICE 1D annotées (moyenne vs individus).

— Au moins un PDP 2D d'interaction avec interprétation.

— Section limites & bonnes pratiques (corrélations, extrapolation, c-ICE, ALE).

— Message clé clairement illustré : *le PDP peut masquer des effets hétérogènes que l'ICE révèle.*