INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Apprentissage Supervisé

Hassouna Mohamed Amine

Laforge Mateo

Plan :

1. Analyse et préparation des jeux de données:
   - Analyse des attributs
   - Répartition des classes
   - préparation des données pour l'apprentissage automatique
2. Expérimentation 1 : comparaison des modèles par défaut
3. Expérimentation 2 : comparaison des modèles après optimisation
4. Expérimentation 3 : comparaison des meilleurs modèles
5. Expérimentation 4 : inférence sur un autre jeux de données (Colorado et Nevada)
6. Expérimentation 5 : impact de la taille du de données
7. Explicabilité des prédictions
8. Explicabilité avec LIME et SHAP
9. Explication Contrefactuelle

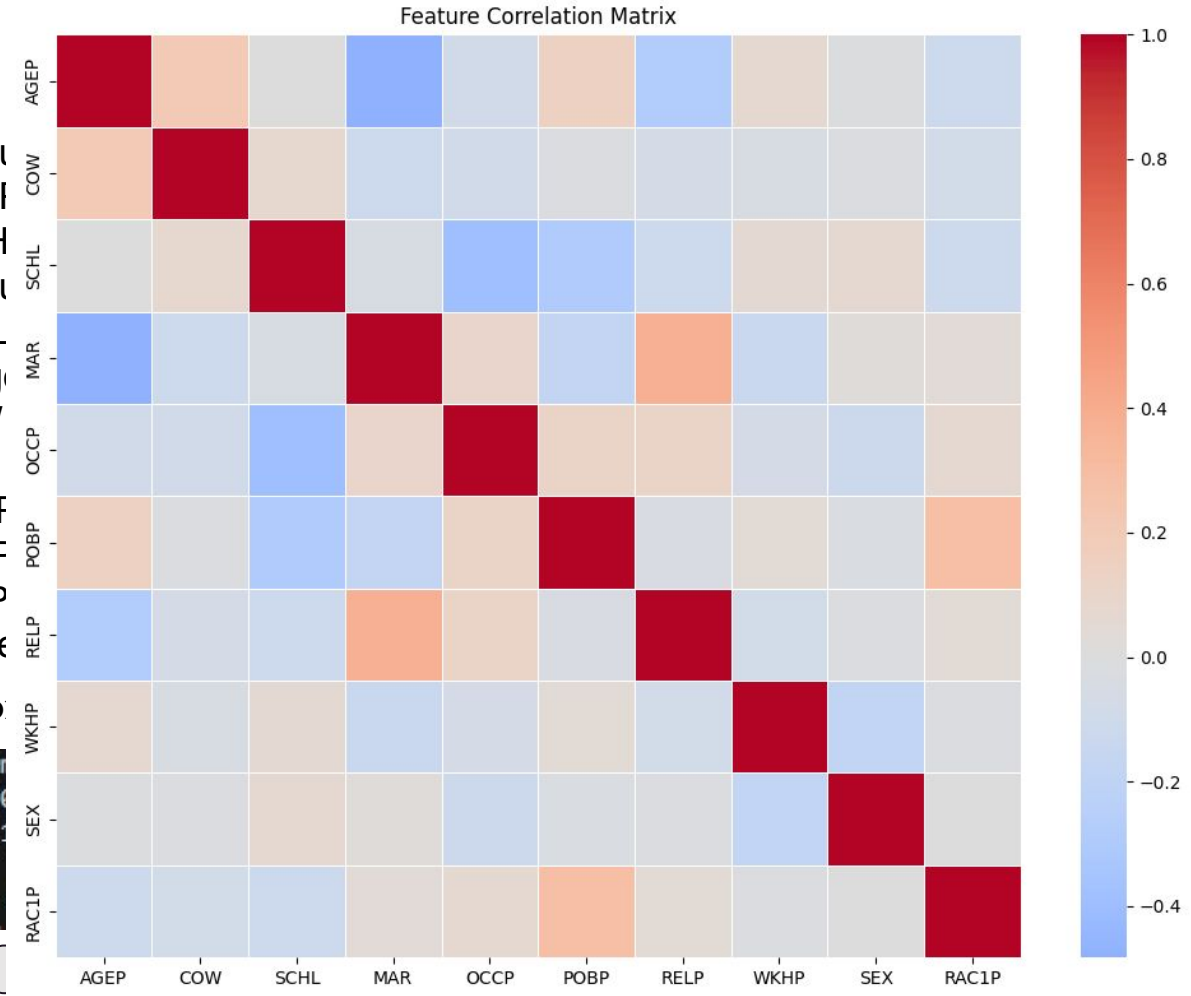
Analyse et préparation des jeux de données:

Analyse

- Attribu
  AGEI
  WKH
- Attribu
  SCHL
- catég
  COW
  MAR
  OCCI
  POBI
  RELP
  SEX e

Répartitio

```
y_norr
False (
True  (
Total
Saved:
```



es : (one-hot encoding)
t des attributs

nées :
5 %

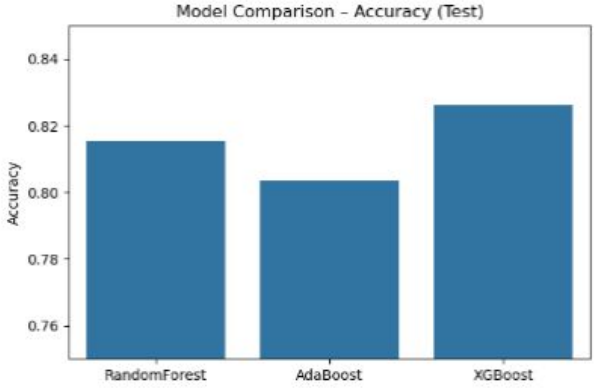
Expérimentation 1 : comparaison des modèles par défaut

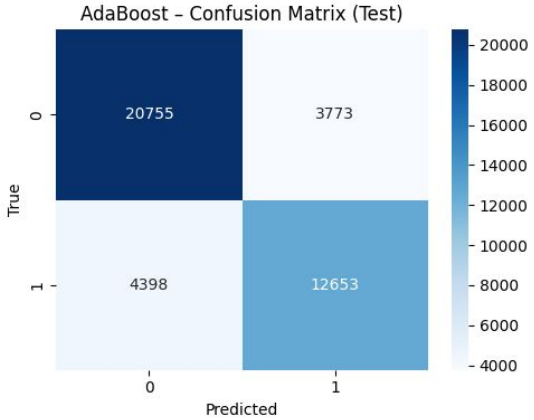
Performances sur l'ensemble d'entraînement

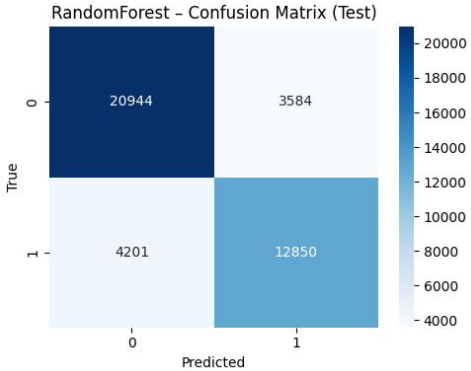
Métrique	Random Forest	AdaBoost	XGBoost
Accuracy	≈ 0.83	≈ 0.81	≈ 0.84
Temps de calcul (s)	14.47	2.37	0.92

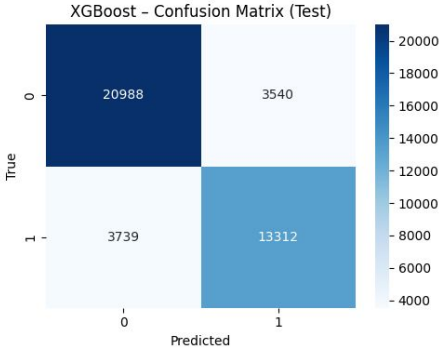
Performances sur l'ensemble du test

Métrique	Random Forest	AdaBoost	XGBoost
Accuracy	0.8155	0.8034	**0.8261**
Temps de calcul (s)	14.47	2.37	**0.92**









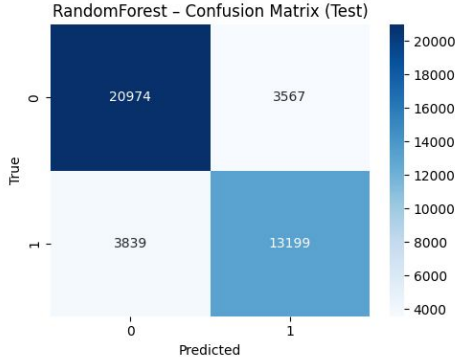
Expérimentation 2 : comparaison des modèles après optimisation

Random-Forest

Hyperparamètres testés
- `n_estimators` $\in \{100, 400\}$
- `max_depth` $\in \{\text{None}, 10, 20\}$
- `min_samples_leaf` $\in \{1, 3, 5, 10\}$

Validation croisée
- Nombre de plis : 3
- Nombre de combinaisons testées : $2 \times 4 \times 4 = 24$
- Nombre total d'entraînements : $24 \times 3 = 72$

Performances
- Accuracy entraînement : 0.9581
- Accuracy test : 0.8210
- Temps de calcul total : 1101.96 s



AdaBoost

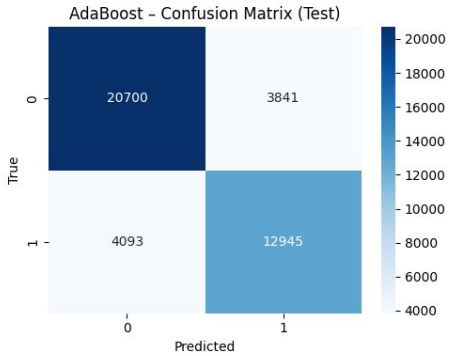
Hyperparamètres testés
- `n_estimators` $\in \{50, 200\}$
- `learning_rate` $\in \{0.5, 1.0, 2.0\}$

Validation croisée
- Nombre de plis : 3
- Nombre de candidats : $2 \times 3 = 6$
- Nombre total d'entraînements : $6 \times 3 = 18$

Meilleurs hyperparamètres

```
n_estimators = 200
learning_rate = 1.0
```

Performances
- Accuracy entraînement : 0.8100
- Accuracy test : 0.8092
- Temps de calcul total : 173.58 s



XGBoost

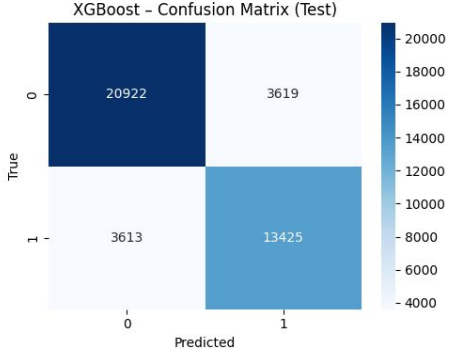
Hyperparamètres testés et rôle
- `max_depth` $\in \{3, 6\}$
- `n_estimators` $\in \{100, 300\}$
- `learning_rate` $\in \{0.05, 0.1\}$:

Validation croisée
- Nombre de plis : 3
- Nombre de candidats : $2 \times 2 \times 2 = 8$
- Nombre total d'entraînements : $8 \times 3 = 24$

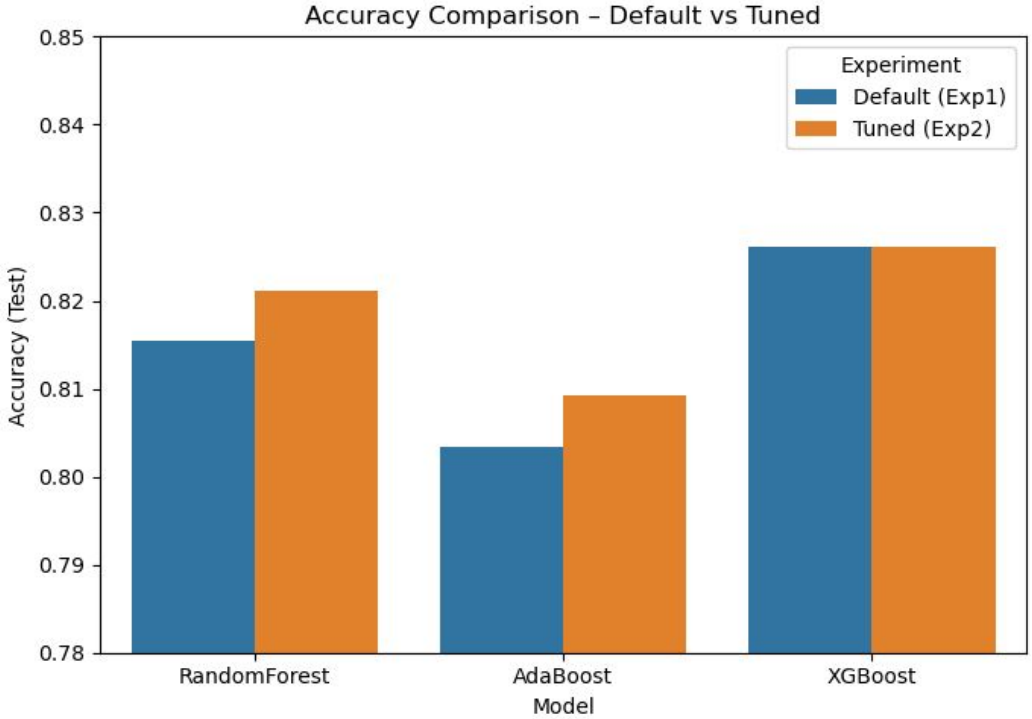
Meilleurs hyperparamètres

```
n_estimators = 300
max_depth = 6
learning_rate = 0.1
```

Performances
- Accuracy entraînement : 0.8440
- Accuracy test : 0.8261
- Temps de calcul total : 27.18 s



Expérimentation 2 : comparaison des modèles après optimisation



Expérimentation 3 : comparaison des meilleurs modèles

Résultat sur l'ensemble du test

Métrique	Random Forest	AdaBoost	XGBoost
Accuracy	0.8210	0.8092	**0.8261**
Rappel	0.7762	0.7598	**0.7879**
Temps de calcul (s)	113.50	28.76	**12.04**

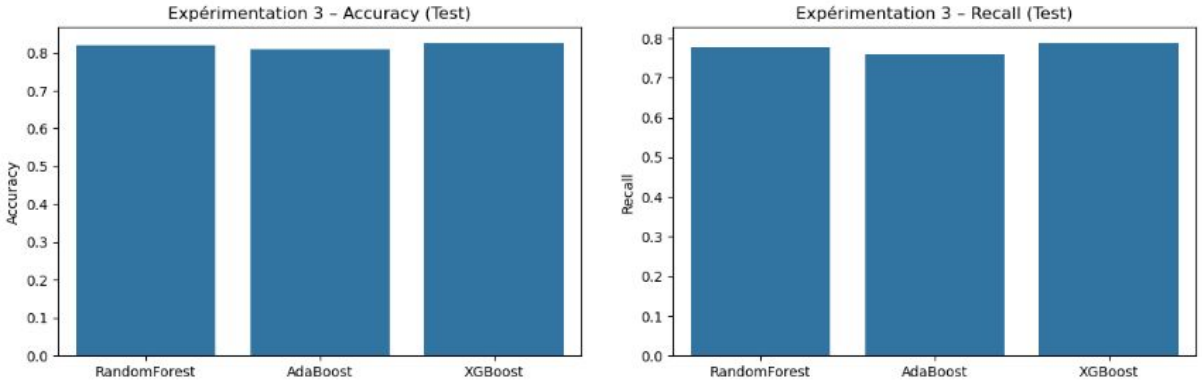
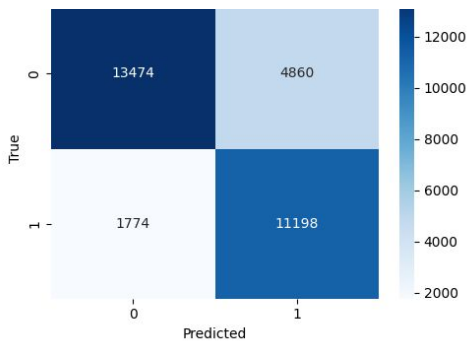


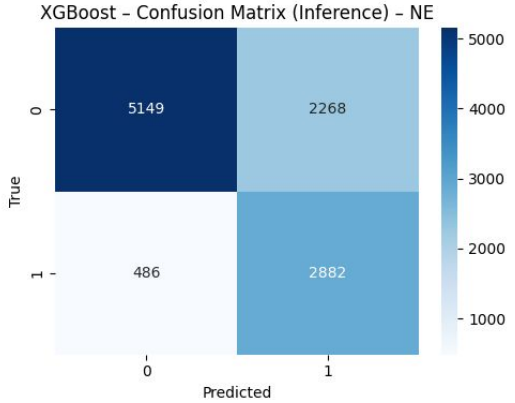
FIGURE 4.2 – Comparaison des métriques (accuracy , rappel) sur l'ensemble de test

Expérimentation 4 : inférence sur un autre jeu de données (optionnel)

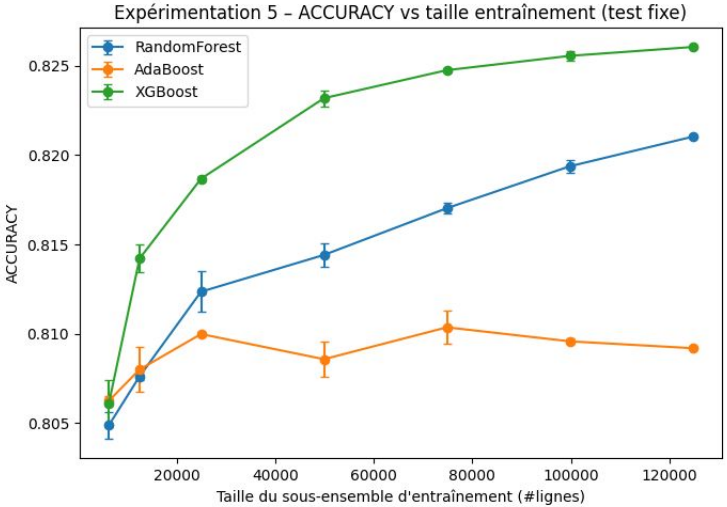
Modèle	Accuracy	Rappel
Random Forest	0.7805	0.8455
AdaBoost	0.7766	0.8384
XGBoost	**0.7881**	**0.8632**

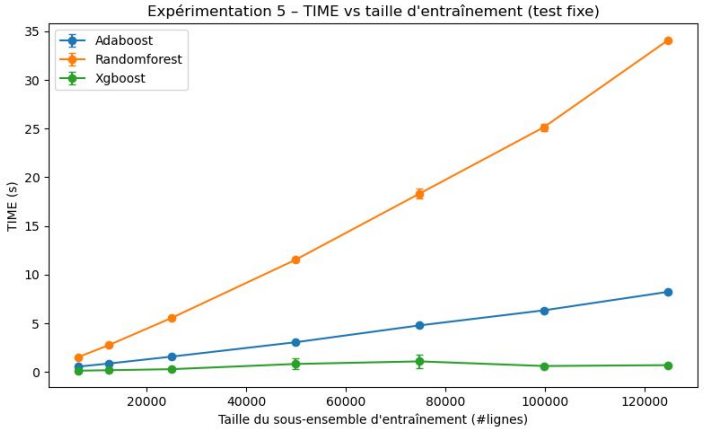


Modèle	Accuracy	Rappel
Random Forest	0.7371	**0.8560**
AdaBoost	0.7317	0.8281
XGBoost	**0.7446**	0.8557



Expérimentation 5 : impact de la taille du jeu de données





Sur la taille maximale ($n_{train} \approx 124736$), les performances finales observées sont :

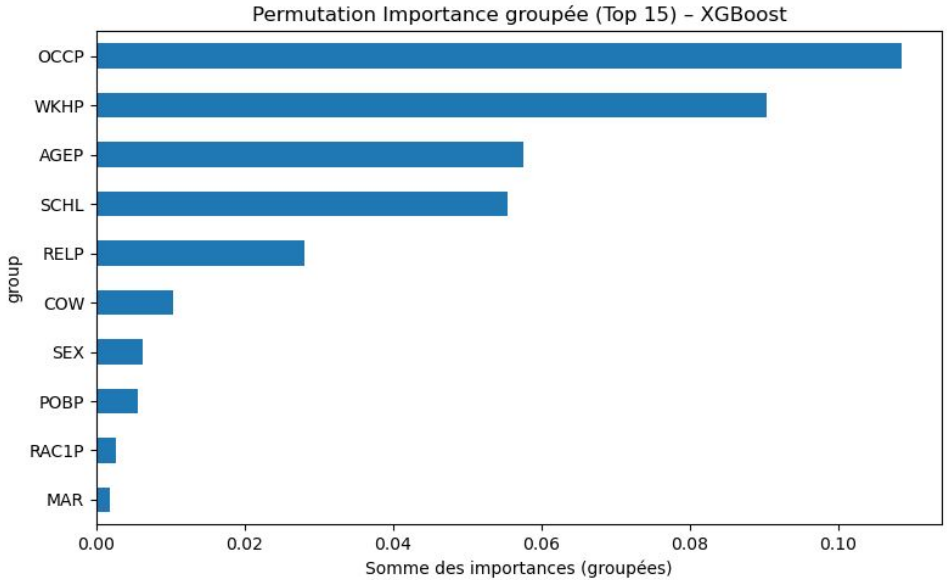
— **Accuracy :** XGBoost (0.8261) > RandomForest (0.8210) > AdaBoost (0.8092)

— **Rappel :** XGBoost (0.7879) > RandomForest (0.7762) > AdaBoost (0.7598)

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Explicabilité

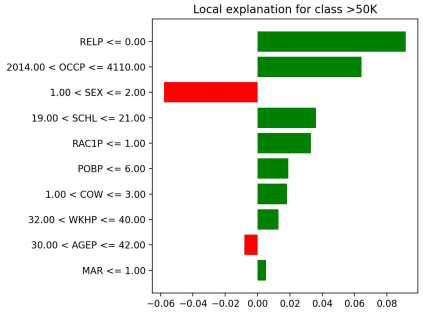
Explicabilité des prédictions

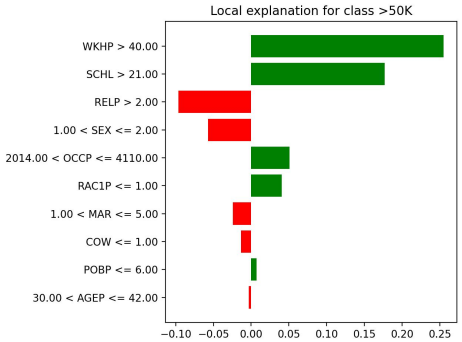


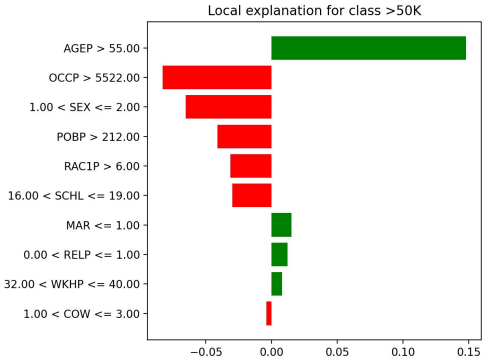
⇒ **la profession (OCCP)** est le facteur le plus déterminan

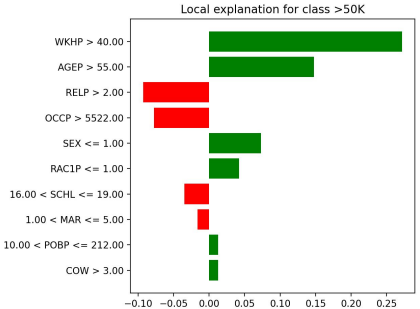
⇒ le modèle repose principalement sur des **facteurs socio-professionnels cohérents avec les connaissances économiques**.

Explicabilité avec LIME et SHAP

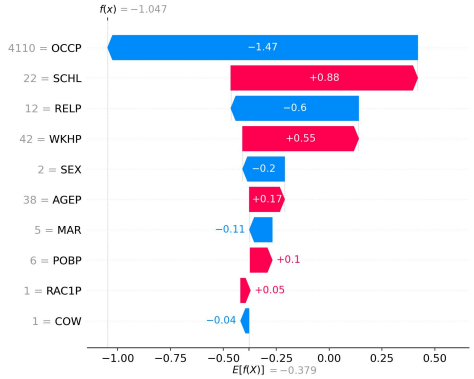


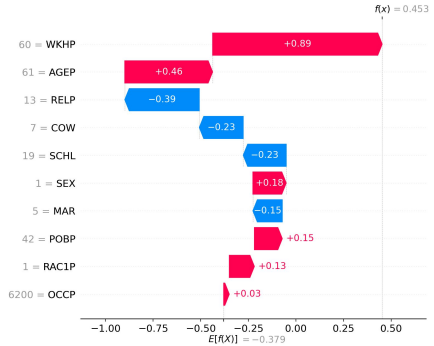


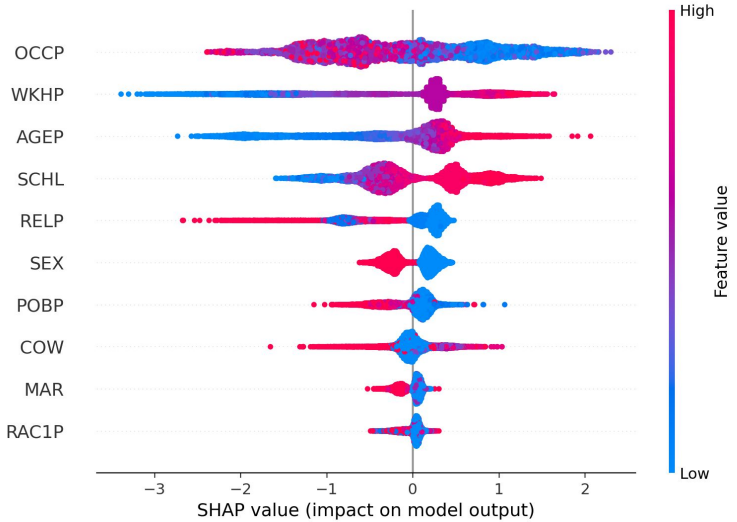




Explicabilité avec LIME et SHAP

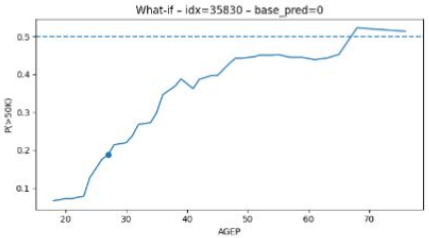


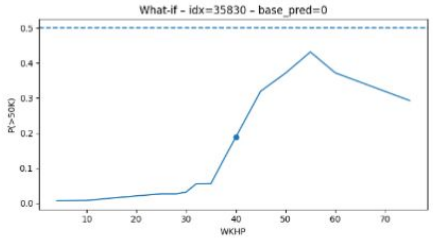


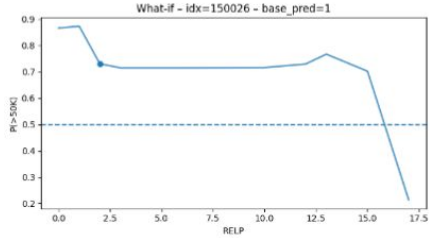


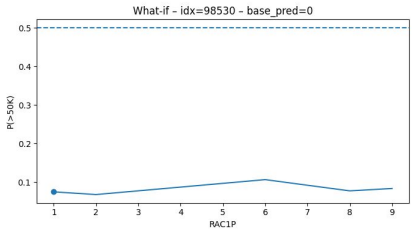
1. On entraîne le modèle
2. On prédit sur le jeu de test
3. On sépare les observations en TP / TN / FP / FN
4. On calcule les valeurs SHAP pour chaque groupe
5. On génère un summary plot par groupe

Explicabilité ContreFactuelle : quelles modifications minimales des attributs pourraient inverser la prédiction ?









```
TRAIN_FEATURES = "../Dataset/alt_acsincome_ca_features_85.csv"
TRAIN_LABELS   = "../Dataset/alt_acsincome_ca_labels_85.csv"

TEST_FEATURES  = "../Dataset/features_split_5.csv"
TEST_LABELS    = "../Dataset/labels_split_5.csv"
```

```
PS C:\Users\medam\OneDrive - insa-toulouse.fr\INSA

=== Evaluation on NEW test dataset (split 5) ===
Accuracy = 0.8270
Confusion matrix:
 [[2430  426]
 [ 420 1615]]
```