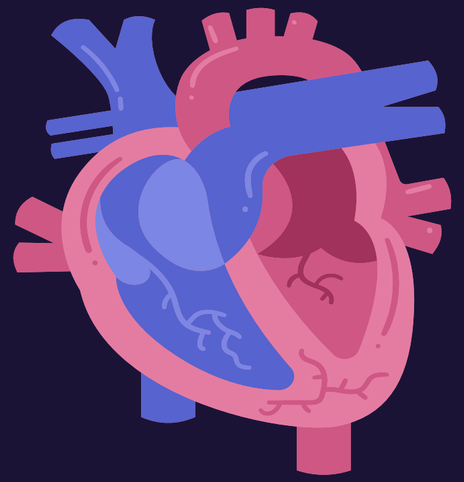
13 Mai 2025

PROJET MACHINE LEARNING

PRÉDICTION DES MALADIES CARDIOVASCULAIRES PAR MACHINE LEARNING



Présenté par :

Aïssatou NDIAYE

Encadré par :

Mr Jean Marc GIRAULT

Mme Sarra MARGI

GRANDE ÉCOLE D'INGÉNIEURS

Table des matières

Contexte et Base de données

Cardiovascular Disease dataset

The dataset consists of 70 000 records of patients data, 11 features + target.

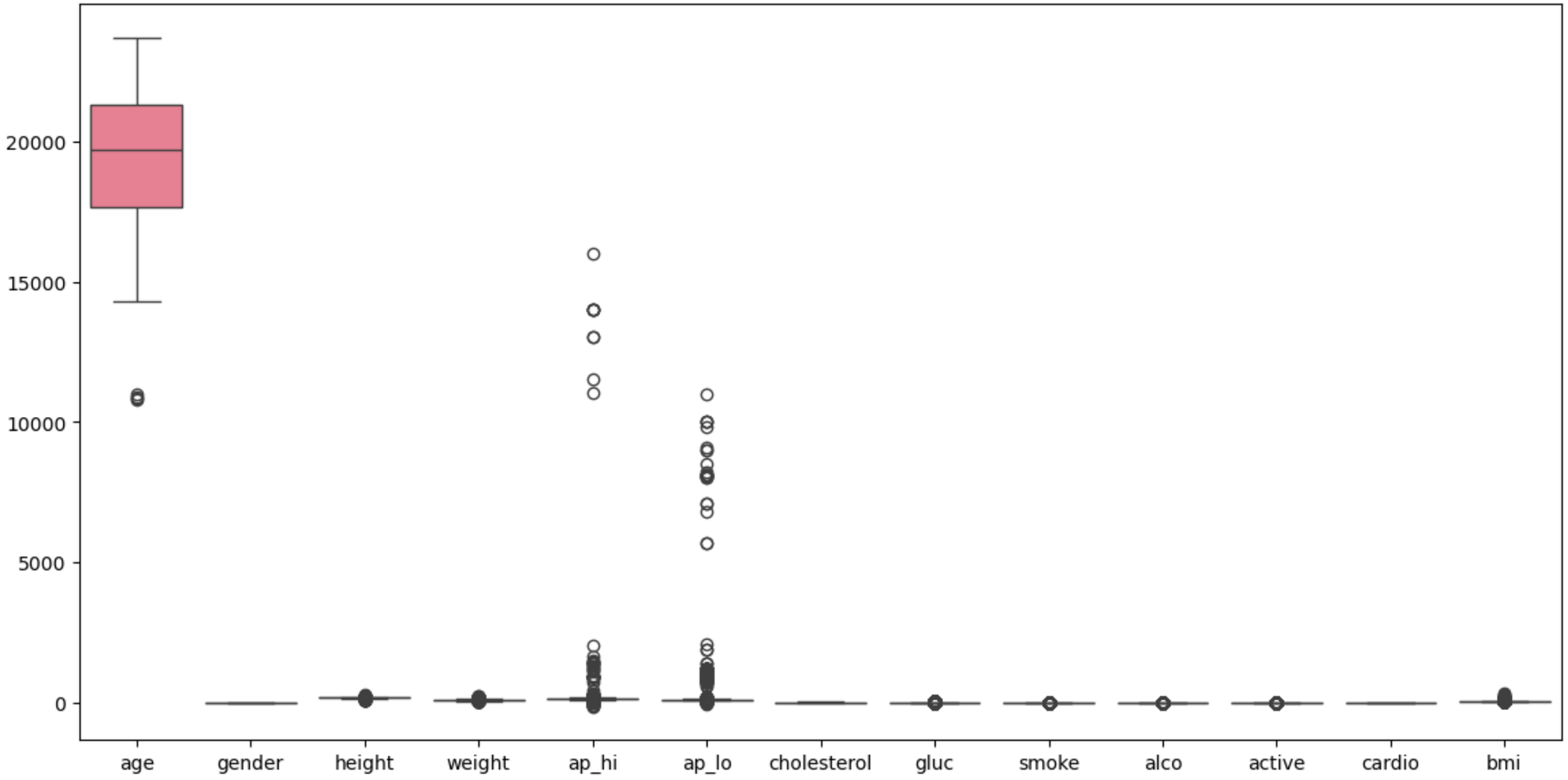
1. Age | Objective Feature | age | int (days)
2. Height | Objective Feature | height | int (cm) |
3. Weight | Objective Feature | weight | float (kg) |
4. Gender | Objective Feature | gender | categorical code |
5. Systolic blood pressure | Examination Feature | ap_hi | int |
6. Diastolic blood pressure | Examination Feature | ap_lo | int |
7. Cholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |
8. Glucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | Subjective Feature | smoke | binary |
10. Alcohol intake | Subjective Feature | alco | binary |
11. Physical activity | Subjective Feature | active | binary |
12. Presence or absence of cardiovascular disease | Target Variable | cardio | binary |

Traitement et Nettoyage des données

Etapes :

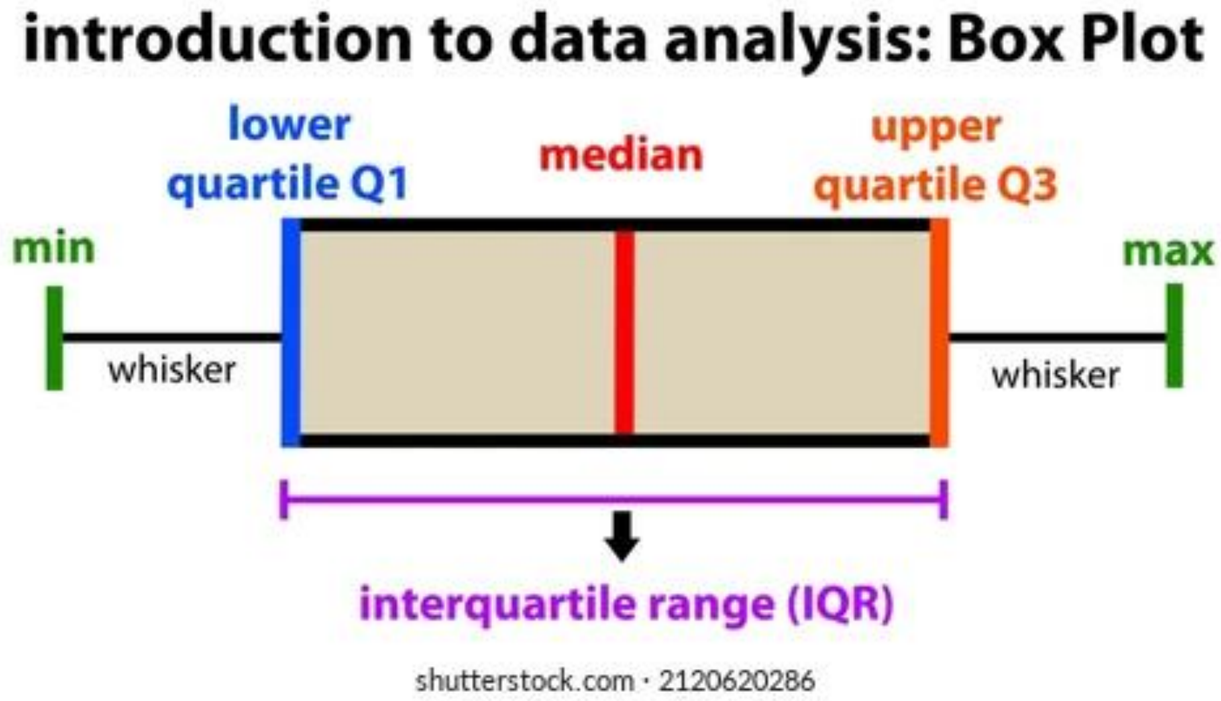
- Gestion des valeurs manquantes,
- Outliers,
- Encodage,
- Standardisation.

Traitement et Nettoyage des données



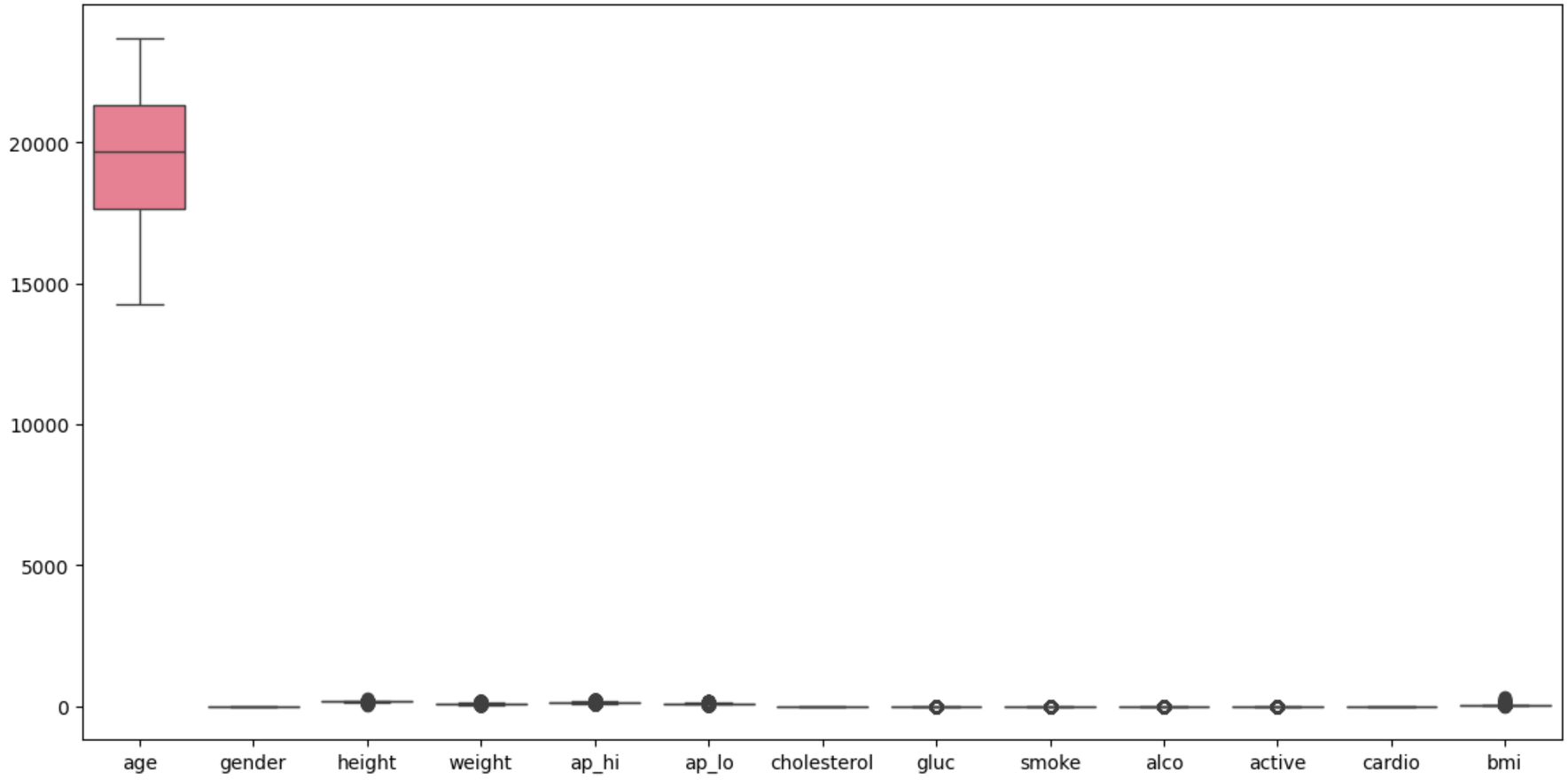
AVANT NETTOYAGE

Traitement et Nettoyage des données



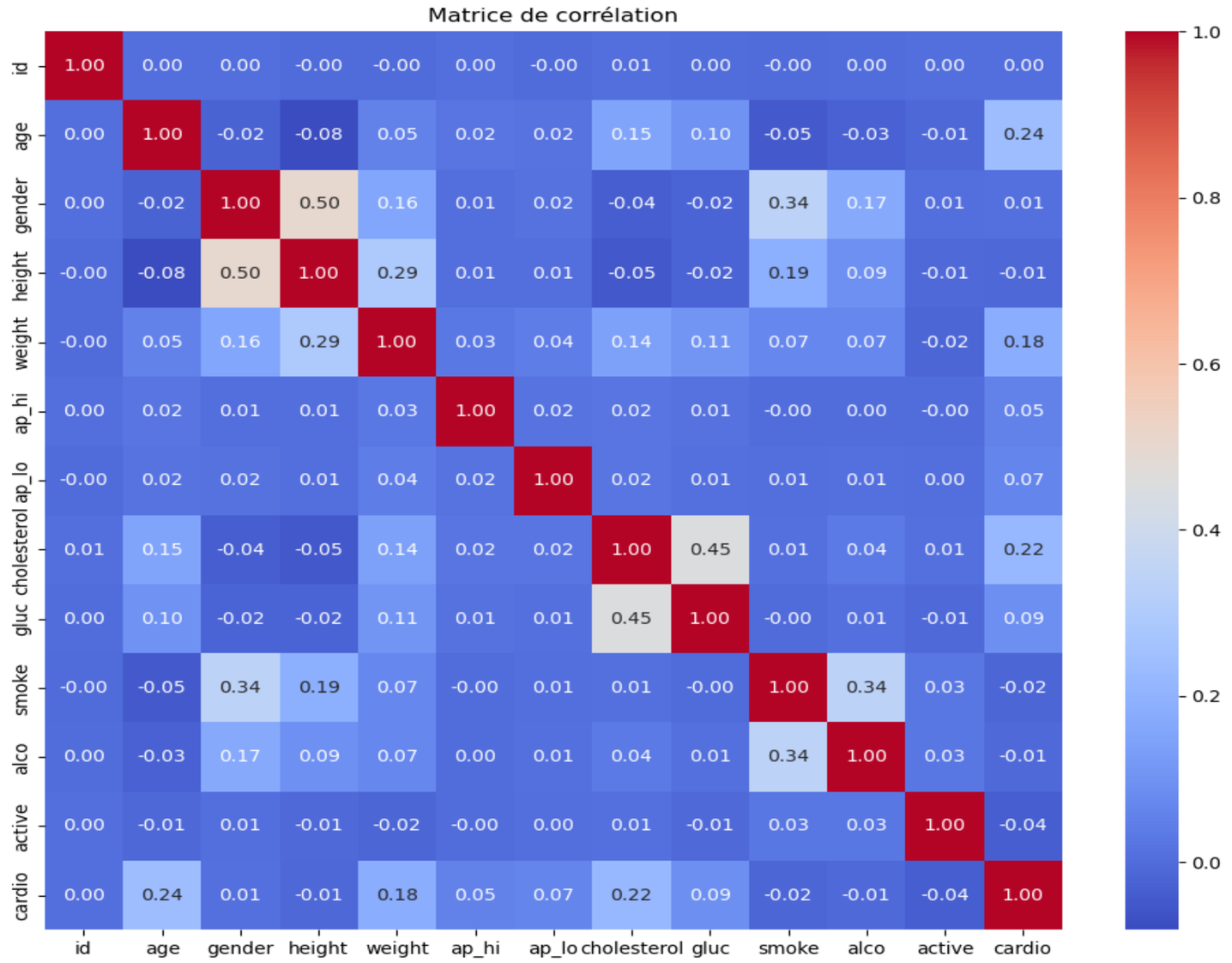

```
df_clean1 = df[(df['ap_hi'] > 70) & (df['ap_hi'] < 250) & (df['ap_lo'] > 40) & (df['ap_lo'] < 180)]
```

Traitement et Nettoyage des données



APRES NETTOYAGE

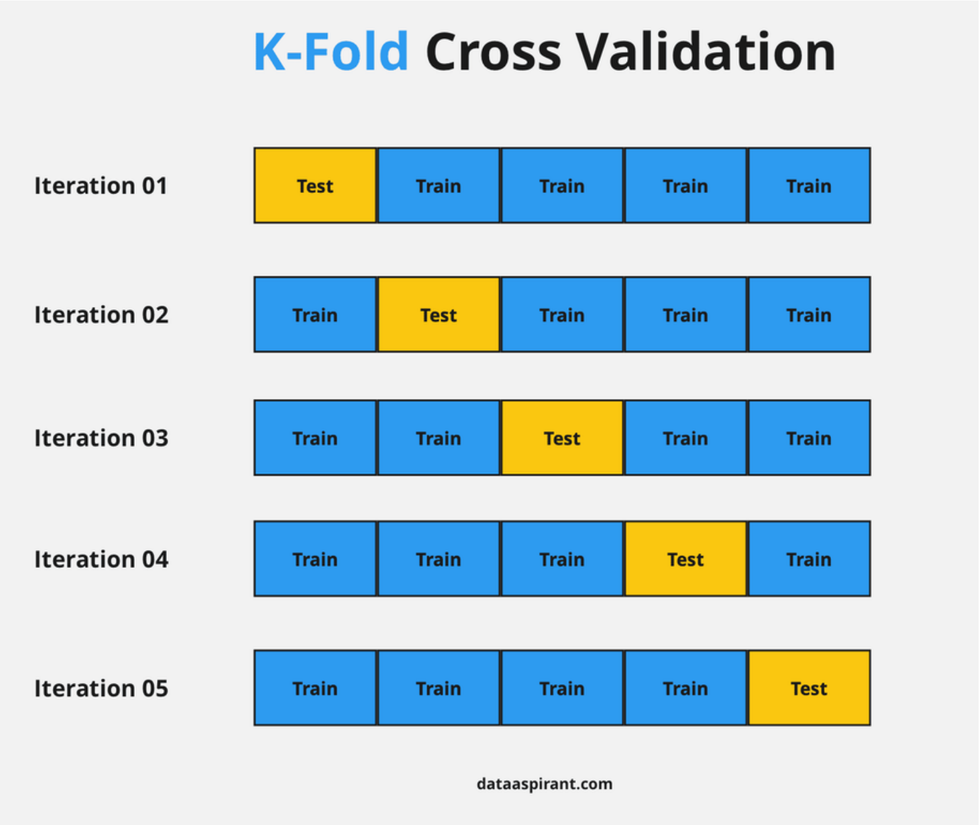
Extraction et sélection des features

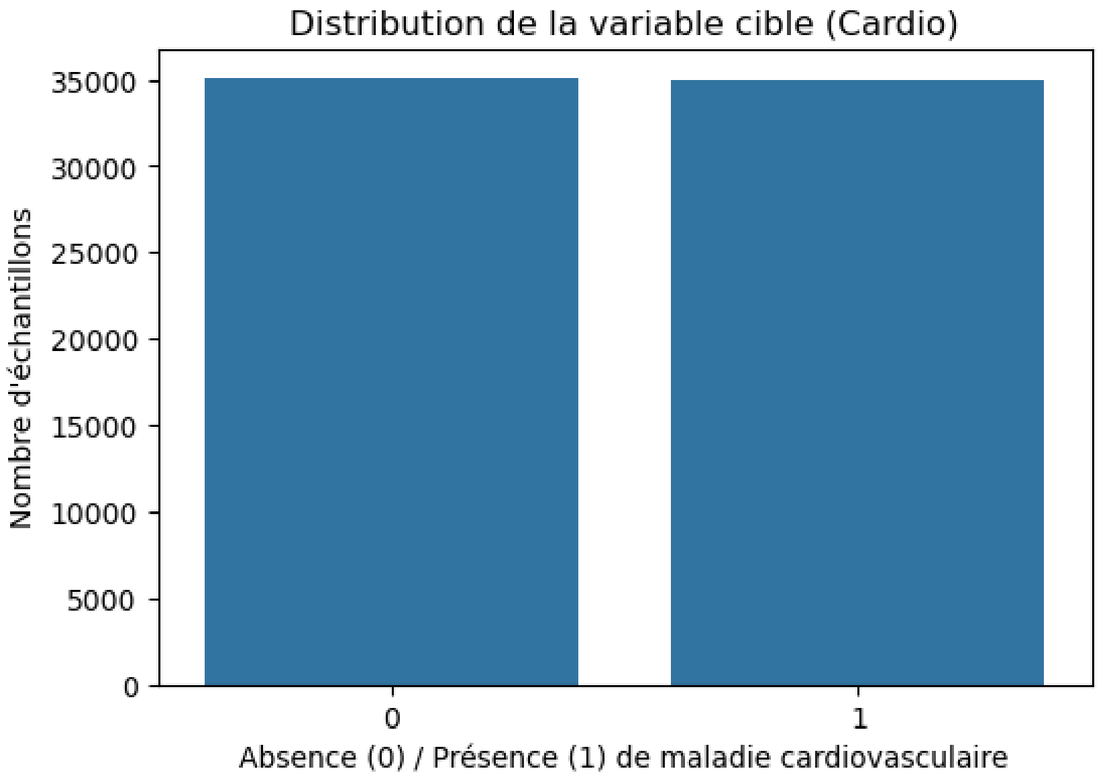


Extraction et sélection des features


```
numerical_features = ['age', 'height', 'weight', 'ap_hi', 'ap_lo', 'imc', 'ap_ratio']
categorical_features = ['gender', 'cholesterol', 'gluc', 'active', 'chol_gluc']
```

Cross Validation et Equilibrage





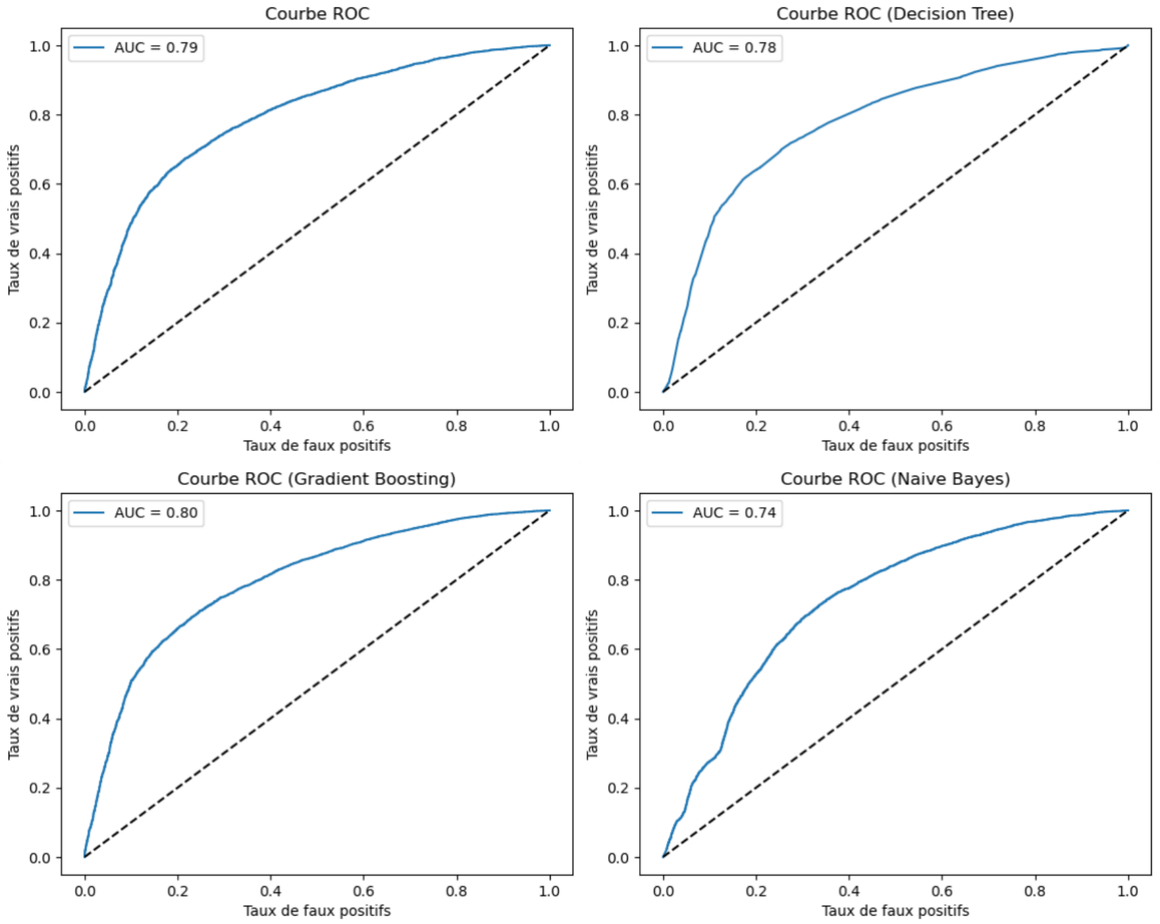
50,03% / 49,97%

```
grid_search = GridSearchCV(pipeline_rf, param_grid, cv=5, scoring='accuracy', n_jobs=-1)
```

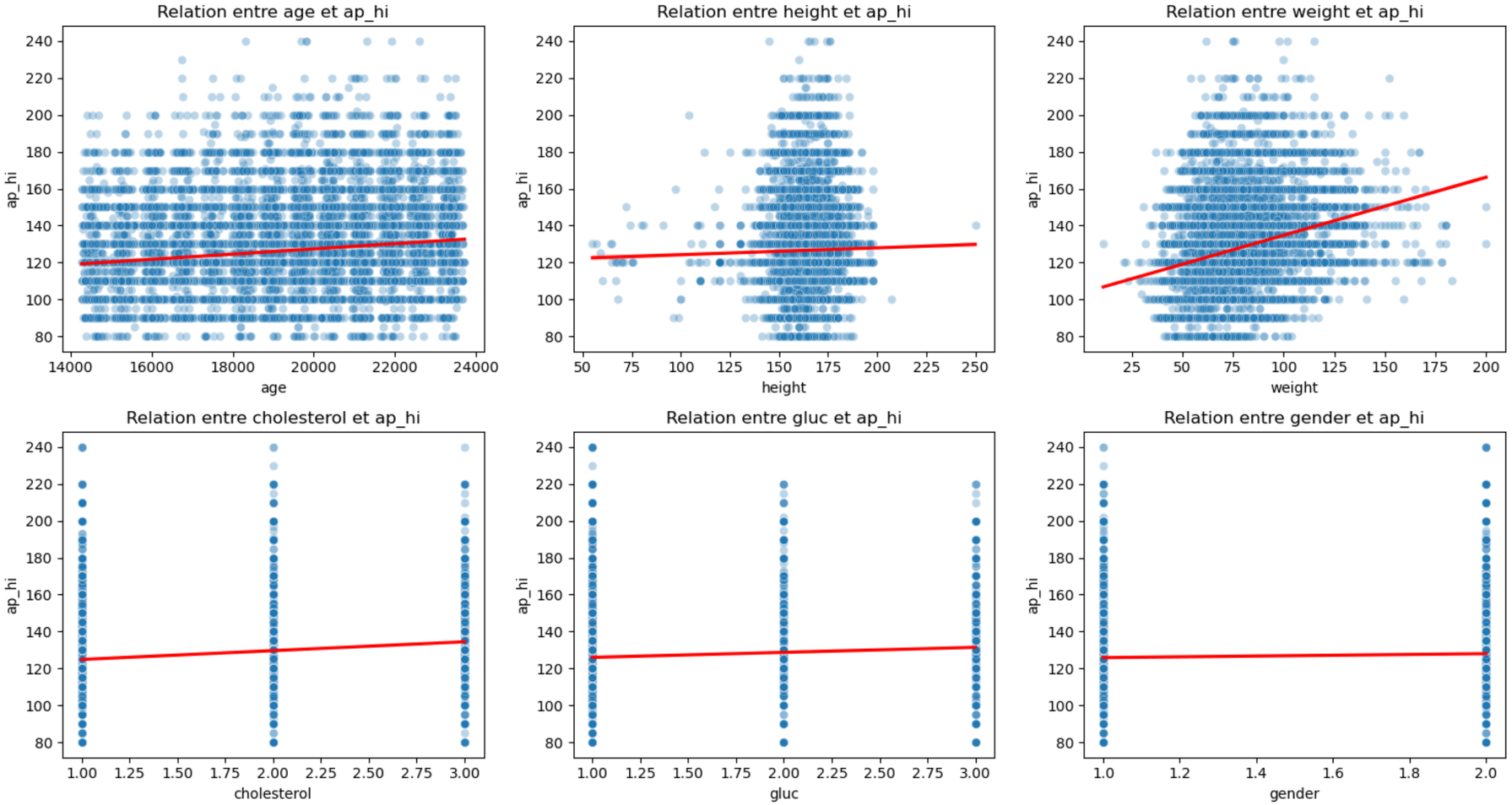
Résultats Classification : Matrice de confusion et Courbe ROC

Pour les matrices de confusion

- **Random Forest**
  - Vrais Négatifs (TN) : 5507 (40.05 %)
  - Faux Positifs (FP) : 1436 (10.45 %)
  - Faux Négatifs (FN) : 2252 (16.39 %)
  - Vrais Positifs (TP) : 4552 (33.11 %)
- **Decision Tree**
  - Vrais Négatifs (TN) : 5184 (37.71 %)
  - Faux Positifs (FP) : 1759 (12.80 %)
  - Faux Négatifs (FN) : 2064 (15.02 %)
  - Vrais Positifs (TP) : 4740 (34.47 %)
- **Gradient Boosting**
  - Vrais Négatifs (TN) : 5376 (39.12 %)
  - Faux Positifs (FP) : 1567 (11.40 %)
  - Faux Négatifs (FN) : 2133 (15.52 %)
  - Vrais Positifs (TP) : 4671 (33.96 %)
- **Naive Bayes**
  - Vrais Négatifs (TN) : 5759 (41.91 %)
  - Faux Positifs (FP) : 1184 (8.62 %)
  - Faux Négatifs (FN) : 3619 (26.34 %)
  - Vrais Positifs (TP) : 3185 (23.14 %)

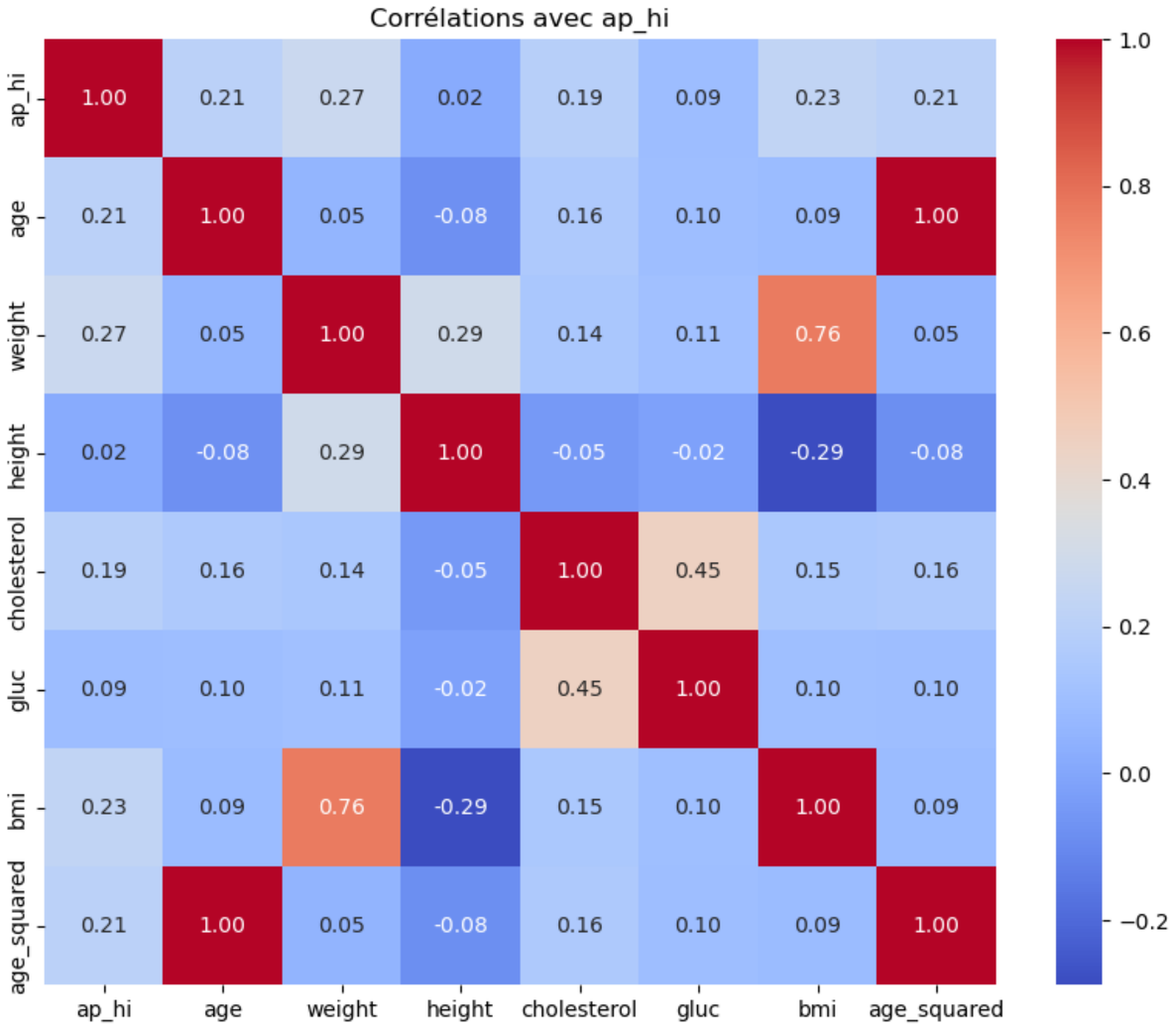


Résultats Régression : EQM / MSE et Scatter Plot



Régression linéaire entre chaque variable et la cible

Résultats Régression : EQM / MSE et Scatter Plot




```
=== Random Forest ===
R²: -0.10354926598548797
MSE: 33737.82540312458
RMSE: 183.6785926642639
MAE: 17.917680788265308
```

```
=== Gradient Boosting ===
R²: -0.1829177771316144
MSE: 36164.28795816303
RMSE: 190.16910358458082
MAE: 17.032929563035466
```

```
=== Linear Regression ===
R²: 0.0009506007982974518
MSE: 30543.044373523047
RMSE: 174.76568419893835
MAE: 15.017121064147442
```

```
=== SGD Regressor ===
R²: 0.0008157514276443401
MSE: 30547.167002809343
RMSE: 174.77747853430472
MAE: 15.499494024214252
```

Discussion et Limites

- **Base de données assez volumineux : manque de temps pour bien tester les modèles de classification.**
- **Variables importantes absentes (génétique, VES, mode de vie, stress, rigidité artérielle, antécédents, marqueurs biologiques…).**
- **Données simplifiées (ex : cholestérol codé sur 3 niveaux)**
- **Données statiques, pas d’évolution dans le temps**

Conclusion

01

CONCLUSION

En résumé, la classification donne d’assez bons résultats pour détecter les patients à risque, mais la régression reste un défi.

02

PERSPECTIVES

Pour aller plus loin, il faudrait enrichir la base de données avec des variables plus fines, et explorer de nouveaux modèles ou de nouveaux types de données.


eseo
GRANDE ÉCOLE D'INGÉNIEURS
Angers • Dijon • Paris-Vélizy