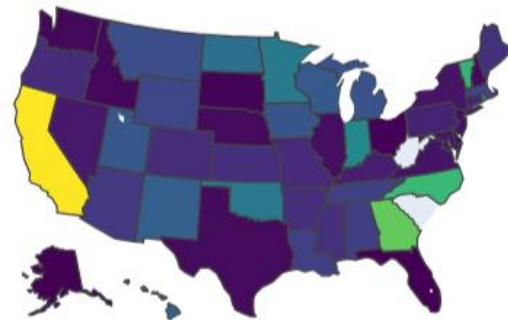




# Détection de Fraude des Prestataires de Santé aux États-Unis

Projet Machine Learning en assurance  
M2 Actuariat 2024-2025

*Groupe ADMW*  
AMIEL Florian  
DJIBRIL OMAR Emma  
MOREL-LE GUYADER Julie  
WENDLING Solène



# Plan

---

- 1 | Contexte et problématique
- 2 | Données
- 3 | Modèles
- 4 | Optimisation des hyperparamètres
- 5 | Conclusion



# 1 | Contexte du projet

- Medicare aux Etats-Unis
- Types de Fraudes et Sanctions
- Visualisation et chiffres principaux
- Problématique



# Medicare

- Programme gouvernemental américain d'assurance santé
- Créé en 1965 sous Lyndon B. Johnson
- 4 parties (A, B, C, D) couvrant hospitalisations, soins médicaux, médicaments
- Éligibilité : 65+ ans, résidents américains, cotisations sociales

## Force et Défis de Medicare :

- Croissance démographique massive
- Coûts de santé en augmentation
- Risques de fraudes et abus
- Coordination avec d'autres programmes



# Types de Fraude en santé



## Fraudes par Prestataires de Santé

- Double facturation : Multiplier les réclamations pour un service
- Facturation fictive : Facturer des services jamais réalisés
- Upcoding : Facturer un service plus cher que le réel
- Pots-de-vin : Rémunérations contre recommandations

Impact : Augmentation artificielle des coûts médicaux

## Fraude par patients/individus

- Vol d'identité : Utiliser l'assurance d'autrui
- Marketing frauduleux : Collecter des informations personnelles
- Usurpation d'identité professionnelle : Services sans licence

Risque : Atteinte à la vie privée, système de santé

## Fraudes liées aux prescriptions

- Falsification d'ordonnances
- Détournement : Revente de médicaments
- Doctor shopping : Multiplier les consultations pour obtenir des substances

Danger : Trafic de médicaments, risques sanitaires



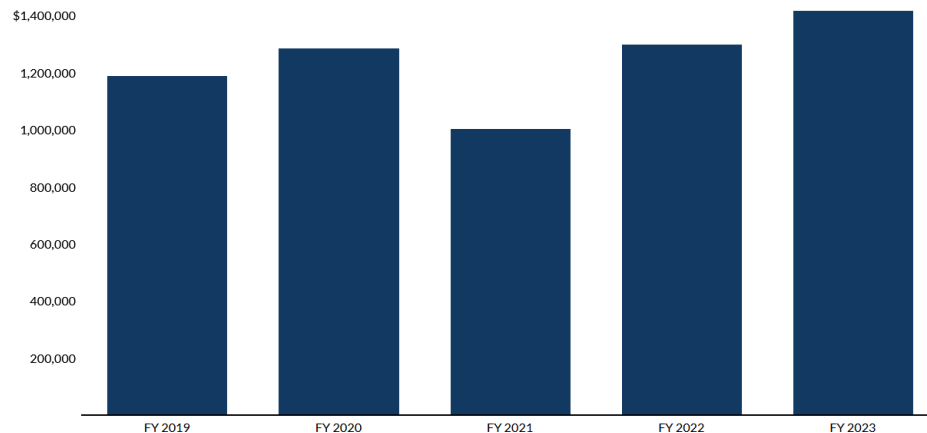
### SANCTIONS

- Amendes
  - Peines de prison
  - Exclusion des programmes publics
- Nécessité : Collaboration public-privé

# Visualisation de la fraude

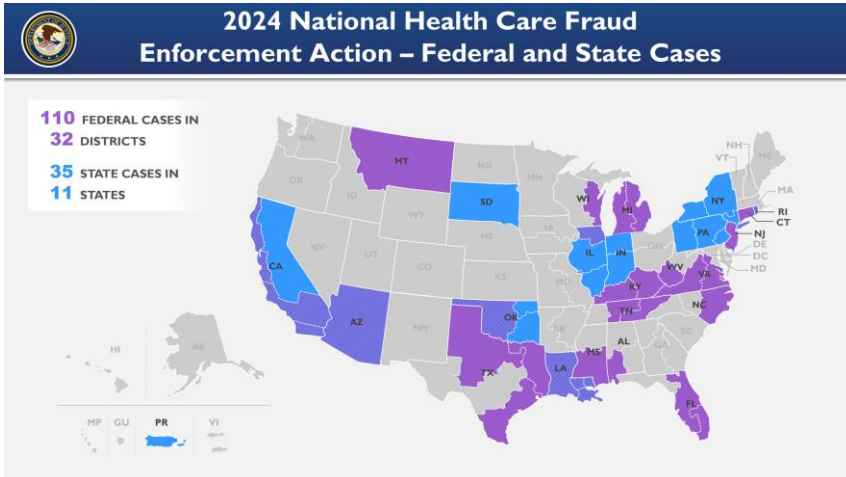
- Fraude répartie sur l'ensemble des Etats-Unis
- Perte moyenne de plus de 1 M\$
- Plus de 300 personnes condamnées chaque année

Median Loss for Individuals Sentenced for Health Care Fraud

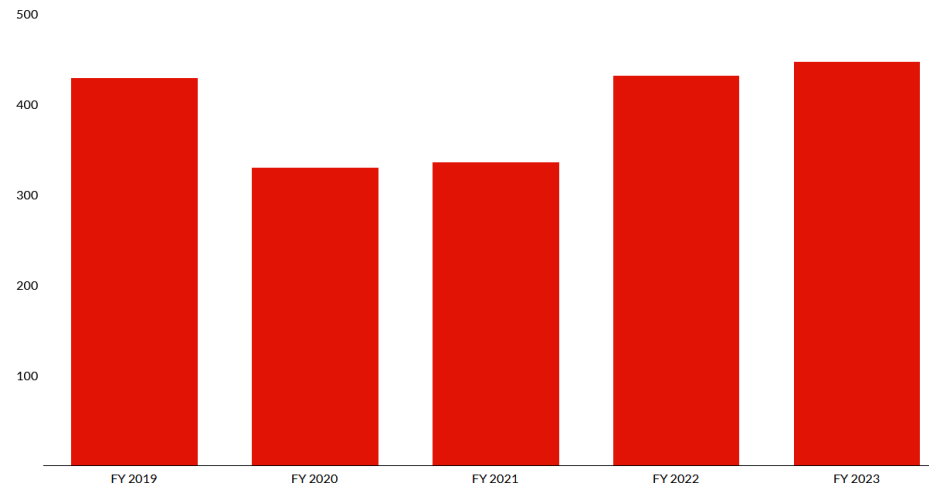


Cases with incomplete sentencing information were excluded from the analysis.

Source: [United States Sentencing Commission, FY 2019 through FY 2023 Datafiles, USSCFY19-USSCFY23](#). • [Get the data](#) • [Download PDF](#)



Number of Individuals Sentenced for Health Care Fraud



Source: [United States Sentencing Commission, FY 2019 through FY 2023 Datafiles, USSCFY19-USSCFY23](#). • [Get the data](#) • [Download PDF](#)

# Problématique

**Comment peut-on utiliser les données des déclarations médicales pour détecter de manière fiable les comportements frauduleux parmi les prestataires de soins de santé ?**



# 2 | Données



- **Présentation des données et Caractéristiques principales**
- **Data Preprocessing**
- **Gestion des données manquantes et autres traitements appliqués aux données**
- **Exploration des données**



# Données

- **Données réelles : Kaggle**
- **Plusieurs fichiers distincts, regroupés en quatre grandes catégories** : les données sur les prestataires, les bénéficiaires, et les réclamations médicales pour soins hospitaliers ou ambulatoires

## **Données des Prestataires (Providers) :**

- **Identifiant unique du prestataire**
- **Variable cible : Indicateur de fraude**
- **Permet l'entraînement des modèles prédictifs**

## **Données des Bénéficiaires (Beneficiaries) :**

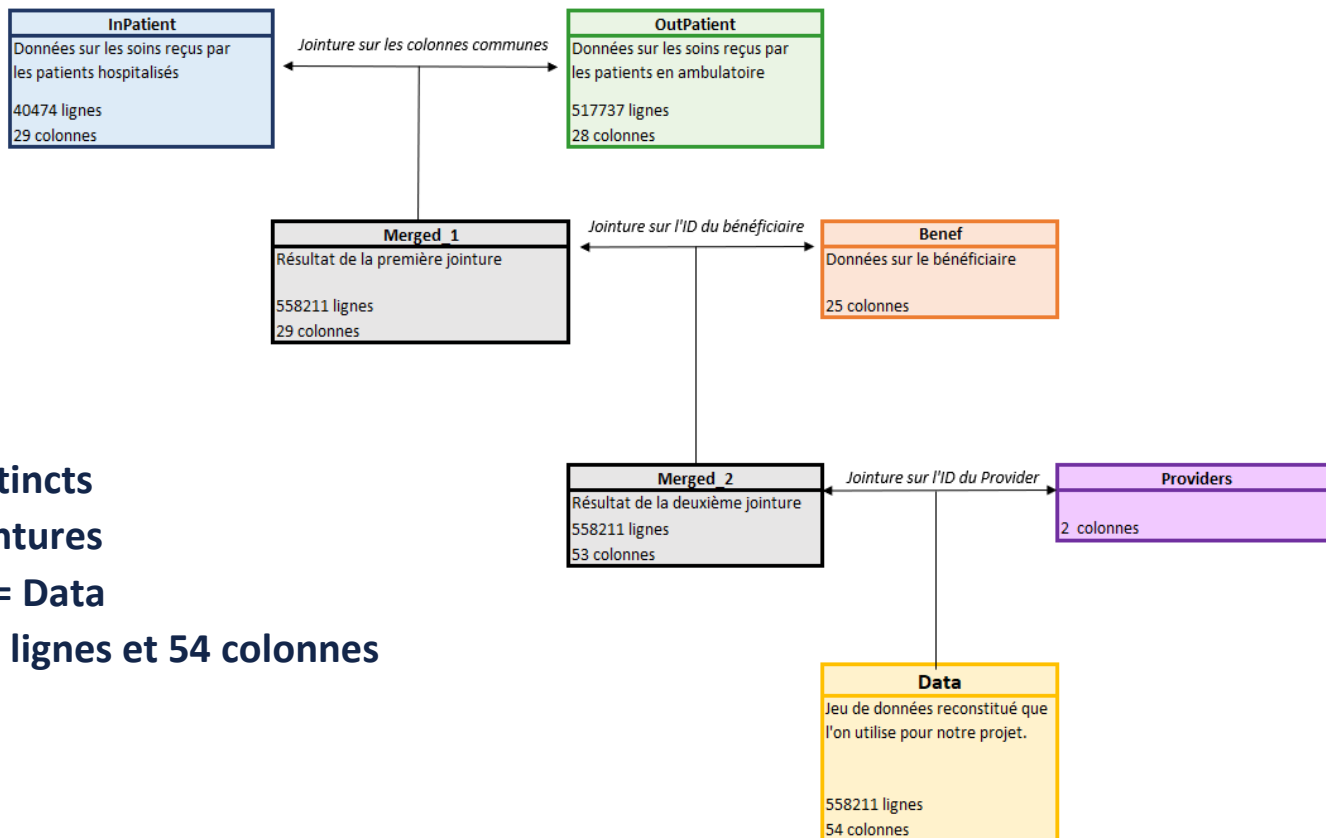
- **Informations démographiques**
- **Contexte médical (âge, sexe, conditions chroniques)**
- **Identifiant du bénéficiaire**

## **Données Hospitalières (Inpatient) ou Ambulatoires (Outpatient) :**

- **Réclamations patients hospitalisés ou non hospitalisés**
- **Trois catégories :**
  - Informations générales
  - Données cliniques
  - Procédures réalisées



# Data Preprocessing



- 4 fichiers distincts
- Plusieurs jointures
- Table finale = Data
- Avec 558211 lignes et 54 colonnes

# Gestion des Données manquantes

- Le traitement des valeurs manquantes = crucial en machine learning
- ⇒ Evite les biais + préserve la capacité de généralisation des modèles.

- Les stratégies clés :

- imputation
- suppression
- transformation

des colonnes en fonction des motifs spécifiques de ces valeurs manquantes.



## Diagnostics (CImDiagnosisCode) :

- 10 colonnes avec taux de valeurs manquantes > 70%
- Action : Suppression des colonnes
- Création d'une nouvelle colonne NbDiagnosis

## Procédures (CImProcedureCode) :

- 5 colonnes avec taux de valeurs manquantes 95-100%
- Action : Suppression totale des colonnes

## Montant Déductible (DeductibleAmtPaid) :

- 0,16% de valeurs manquantes
- Méthode : Imputation par la médiane
- Avantages : robuste face aux valeurs extrêmes (outliers) + distribution de la variable reste cohérente

# Autres traitements des données

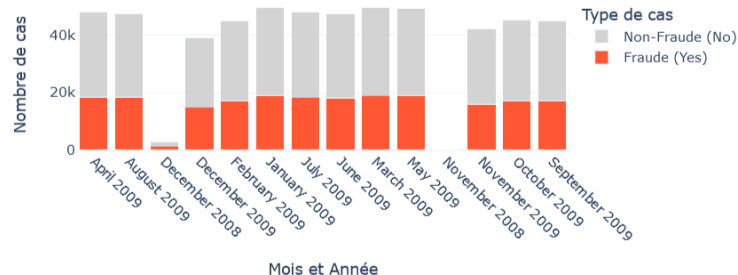
## Nouvelles Colonnes :

- **PatientType** : Différenciation des soins hospitaliers
- **DureeClaim** : Calcul de la durée d'hospitalisation
- **StateRisk** : Classification des États en 4 niveaux de risque :
  - Formule :  $\text{Risk\_Score} = \text{Fraud\_Percentage} \times \log(\text{Total\_Count} + 1)$
  - Niveaux : Faible, Modéré, Élevé, Très élevé
- **CountyRisk** : Classification des comtés par risque
- **CodeProvider** : Regroupement des identifiants de prestataires

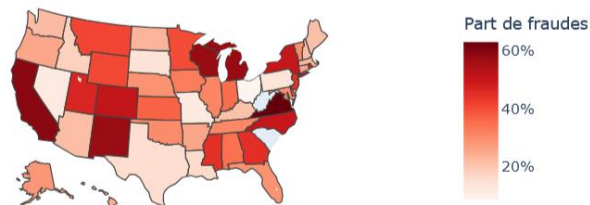
## Transformations de Colonnes

- **PotentialFraud** : Conversion Yes/No en 1/0
- **RenalDiseaseIndicator** : Conversion 0/Y en 0/1

Nombre de fraudes potentielles par mois et année



Part des fraudes potentielles par état (%)



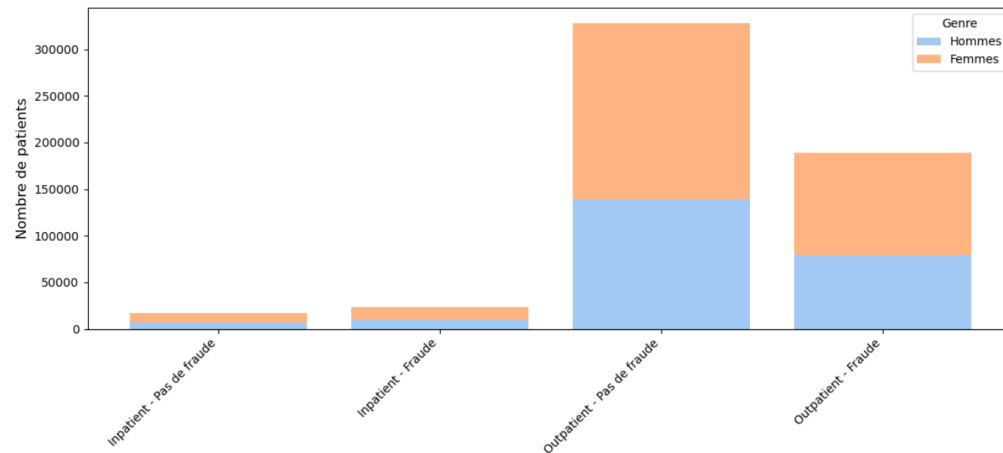
Nombre de cas par état



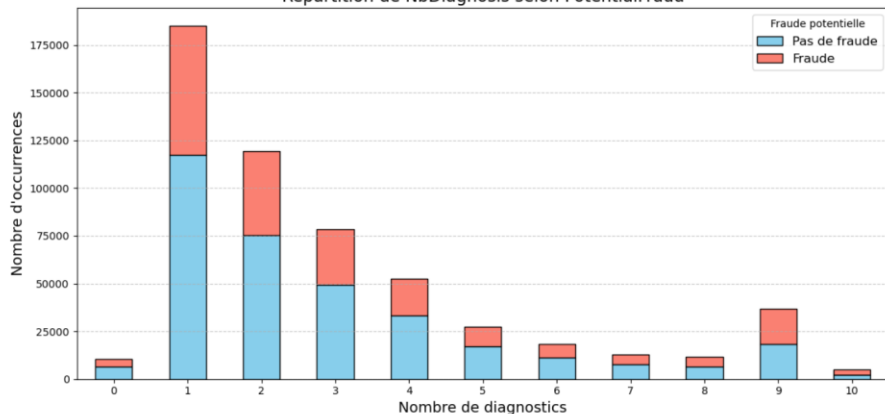
# Exploration des données

- Majorité des cas observés dans nos données : **Outpatient**
- Proportion plus élevée de **femmes**

Répartition des fraudes par type de patient et genre



Répartition de NbDiagnosis selon PotentialFraud



- Majorité des réclamations pour un faible nombre de diagnostics : non-fraude
- Réclamations avec nombre élevé de diagnostics corrélé avec risque de fraude

# Exploration des données

Répartition des valeurs pour les 6 variables relatives aux montants payés et remboursés

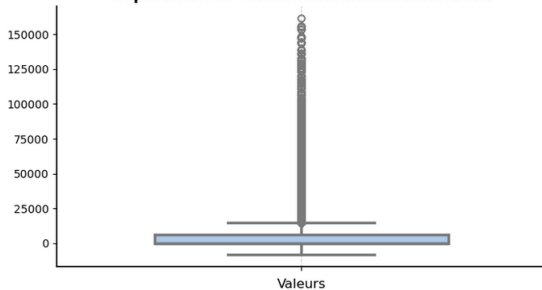


## Répartition des niveaux de risque à travers les Etats-Unis

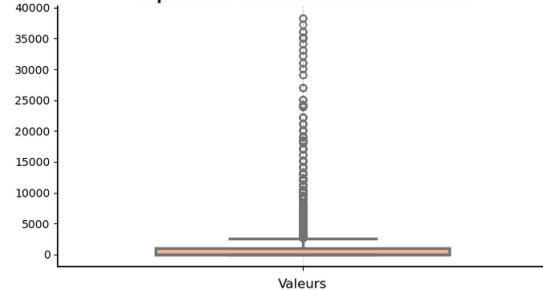
Niveau de risque par état (StateRisk)



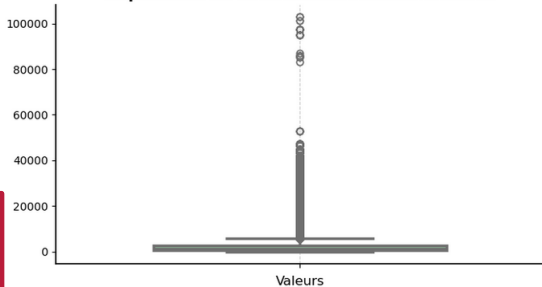
Répartition de IPAnnualReimbursementAmt



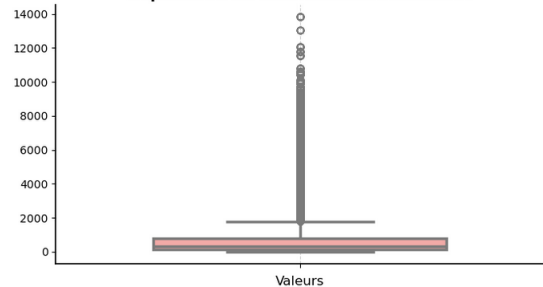
Répartition de IPAnnualDeductibleAmt



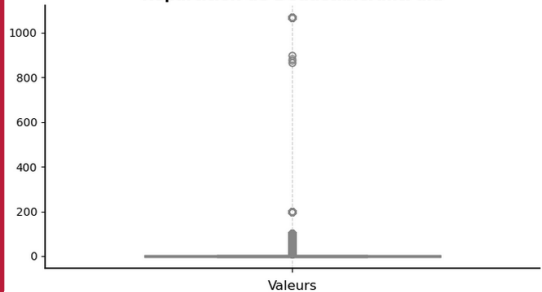
Répartition de OPAnnualReimbursementAmt



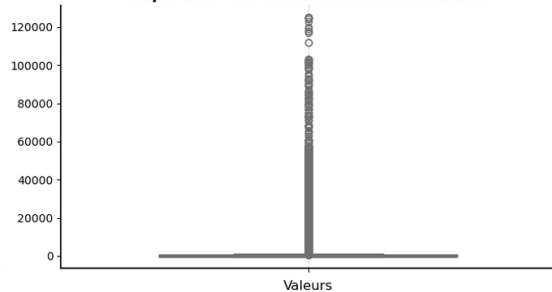
Répartition de OPAnnualDeductibleAmt



Répartition de DeductibleAmtPaid



Répartition de InscClaimAmtReimbursed



# 3 | Modèle

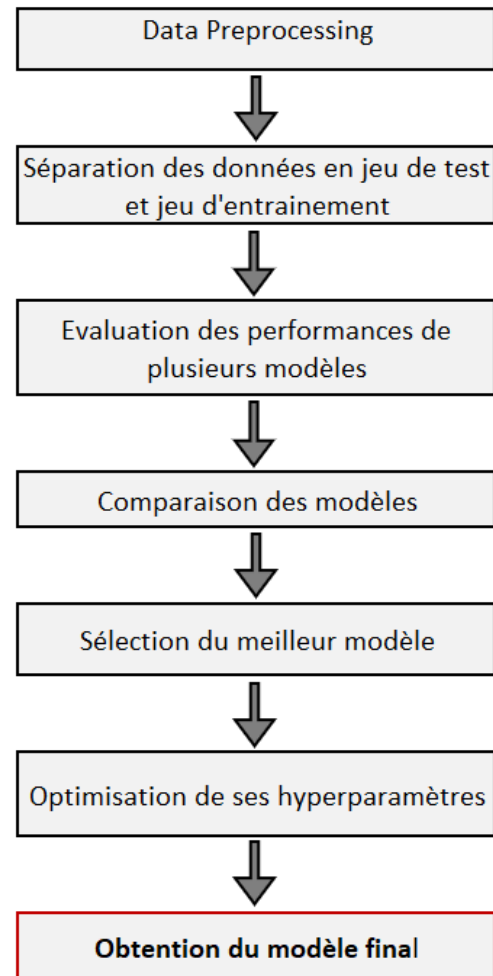
- Démarche
- Présentation des modèles
- Comparaison des modèles



# Démarche

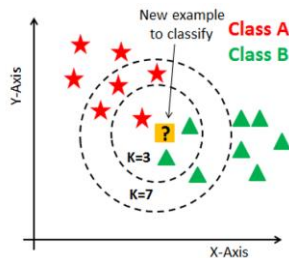


- Rappelons les différentes étapes du projet
- Utilisation de la bibliothèque Scikit-learn





# Modèles

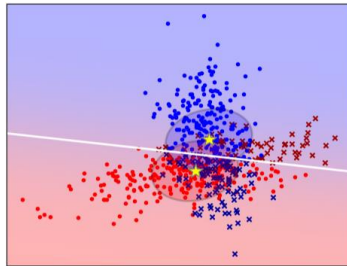
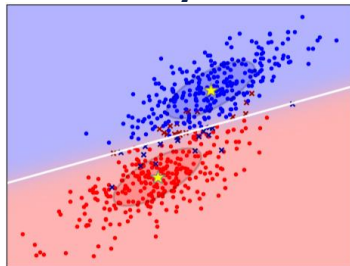


K-Nearest Neighbors

Régression linéaire

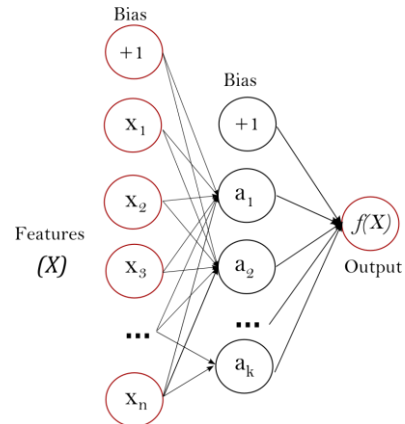
Arbre de décision

Linear Discriminant  
Analysis LDA



7 Modèles

Neural Network



Random Forest

XGBoost

# Comparaison des modèles

## Métriques utilisées :

- Précision globale (Accuracy)
- Précision et rappel spécifiques
- Aire sous la courbe ROC accompagnée des valeurs de l'AUC (Aire sous la courbe)



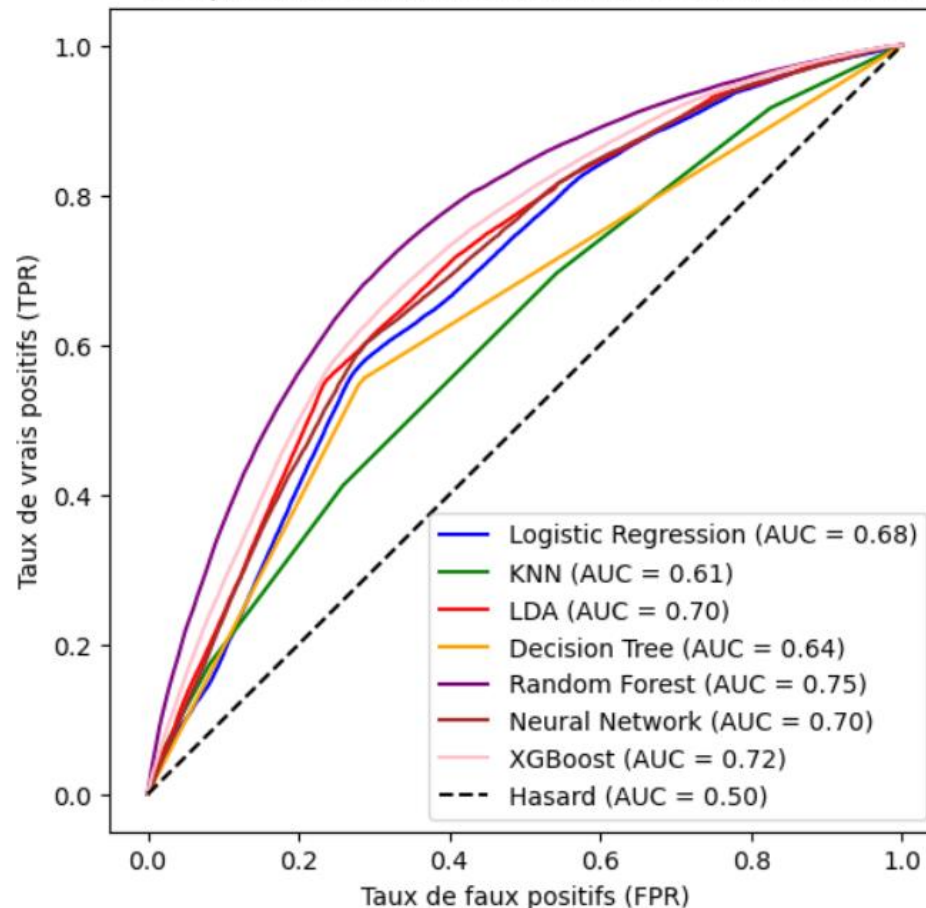
Random Forest

XGBoost

LDA et Neural Network

KNN et Decision Tree

Comparaison des courbes ROC de tous les modèles



# Comparaison des modèles

	AUC	AUC PR	Précision	Rappel	F1-score	Temps d'inférence moyen par échantillon
KNeighborClassifieur	0.61	0.50	0.495	0.412	0.450	1038.31 $\mu$ s
LDA	0.70	0.55	0.592	0.538	0.564	0.22 $\mu$ s
Régression Linéaire	0.68	0.52	0.527	0.153	0.238	0.19 $\mu$ s
Decison Tree	0.64	0.63	0.546	0.547	0.547	0.88 $\mu$ s
Random Forest	0.75	0.64	0.635	0.562	0.596	70.45 $\mu$ s
Neural Network	0.70	0.54	0.561	0.603	0.581	1.74 $\mu$ s
XGBoost	0.72	0.58	0.599	0.544	0.570	1.42 $\mu$ s

## Analyse des performances des modèles :

### ■ AUC et AUC PR

- Random Forest : meilleures AUC et AUC PR
- XGBoost et LDA : performances élevées aussi
- KNeighborClassifieur : AUC la plus faible

### ■ Précision et Rappel

- Random Forest : modèle le plus équilibré suivi par :
- XGBoost et LDA : performances proches

### ■ F1-Score

- Random Forest : score le plus élevé
- Neural Network : 2ème
- XGBoost : 3ème

### ■ Temps d'Inférence

- LDA et Régression Linéaire : les plus rapides
- KNeighborClassifieur : le plus lent



**Random Forest le plus performant**

# 4 | Optimisation des hyperparamètres

- Objectif
- Approche adoptée
- Résultats de l'optimisation



# Hyperparamètres

## Définition des Hyperparamètres :

- Paramètres définis avant l'entraînement du modèle
- Influencent directement la performance du modèle
- Impactent la capacité de généralisation

## Hyperparamètres Clés de Random Forest :

- Nombre d'arbres (n\_estimators)
- Profondeur maximale (max\_depth)
- Échantillons pour diviser un nœud (min\_samples\_split)
- Échantillons dans une feuille (min\_samples\_leaf)
- Variables par division (max\_features)
- Pondération des classes (class\_weight)



**Objectif  
maximiser la  
performance  
prédictive**



# Optimisation



## Méthodes d'Optimisation :

- Recherche par grille (Grid Search)
- Recherche aléatoire (Random Search)
- Exploration systématique de l'espace des hyperparamètres

## Validation Croisée :

- Division des données en sous-ensembles
- Évaluation sur des partitions indépendantes
- Prévention du surapprentissage



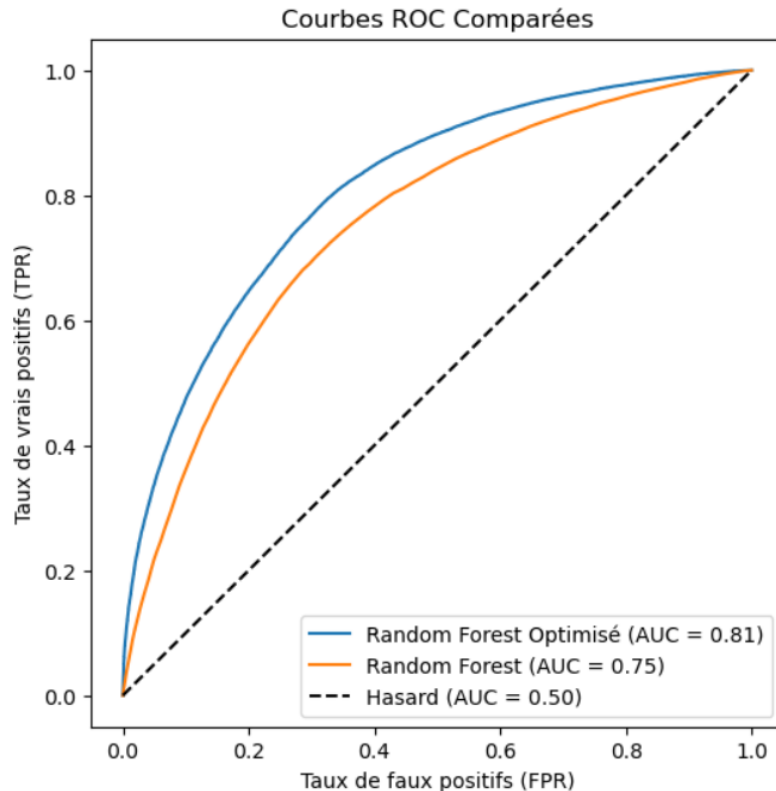
**Généralisation  
des  
performances  
observées**

# Résultats de l'optimisation

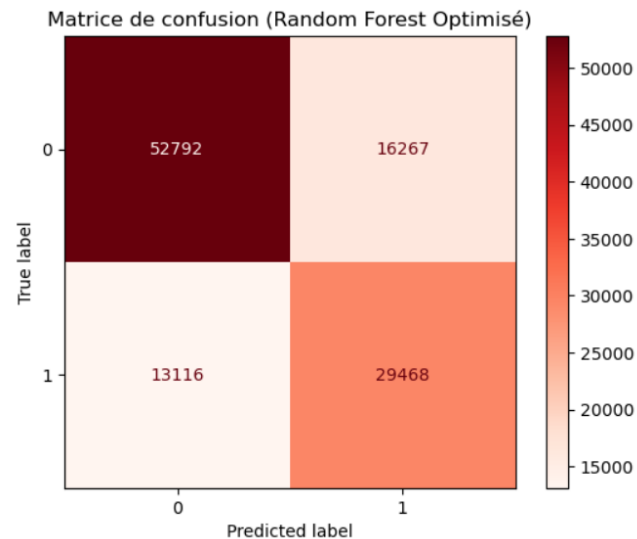
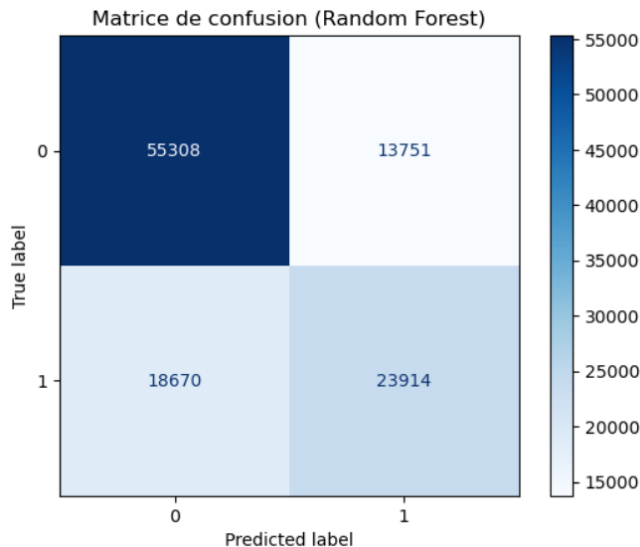
- Recherche par grille + validation croisée à 3 plis
- 432 combinaisons d'hyperparamètres

## Meilleurs hyperparamètres trouvés :

- class\_weight : balanced
- max\_depth : None
- max\_features : sqrt
- min\_samples\_leaf : 4
- min\_samples\_split : 10
- n\_estimators : 200



# Résultats de l'optimisation

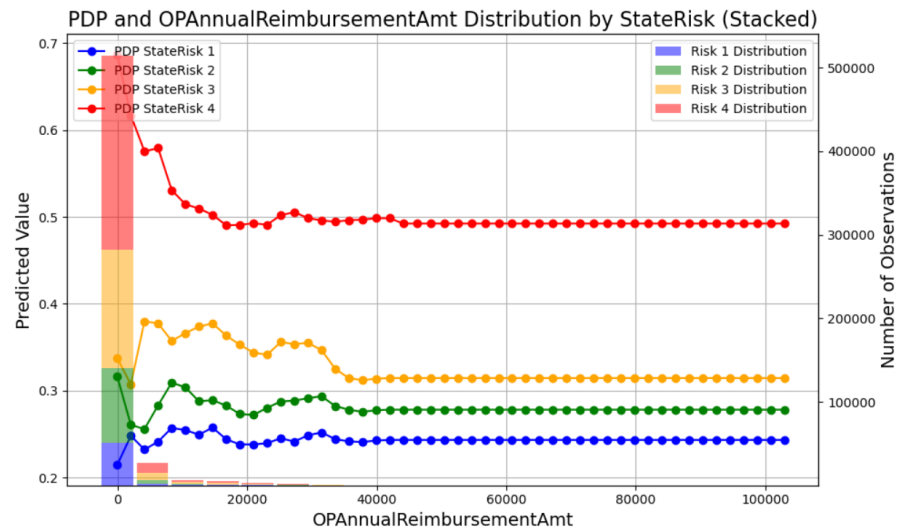
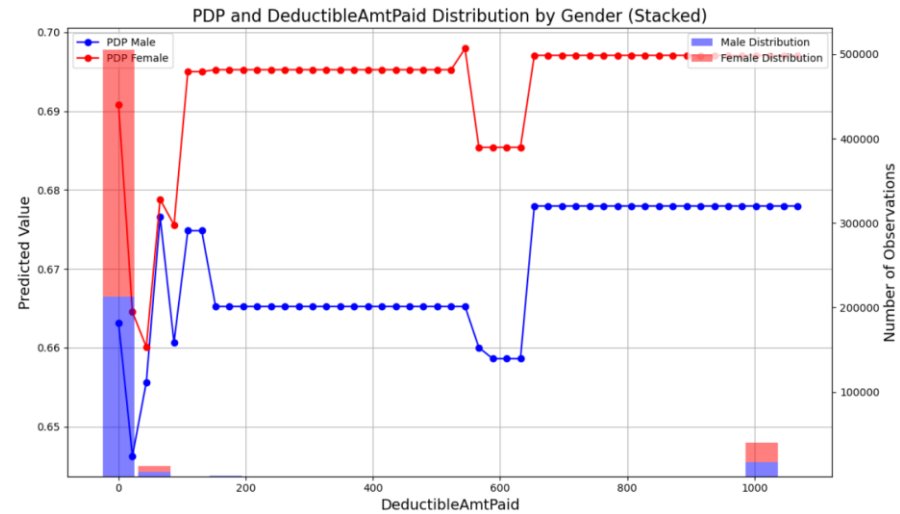
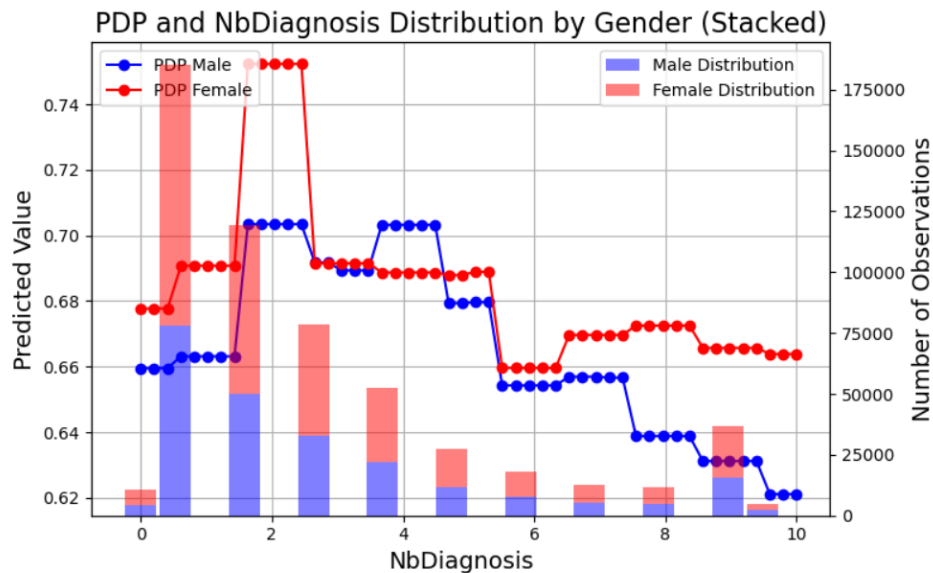


- Amélioration de la précision
- Amélioration de la capture des cas positifs
- Meilleur équilibre précision/rappel

	Précision	Rappel	F1-score
Random Forest	0.635	0.562	0.596
Random Forest Optimisé	0.644	0.692	0.667



# Résultats de l'optimisation



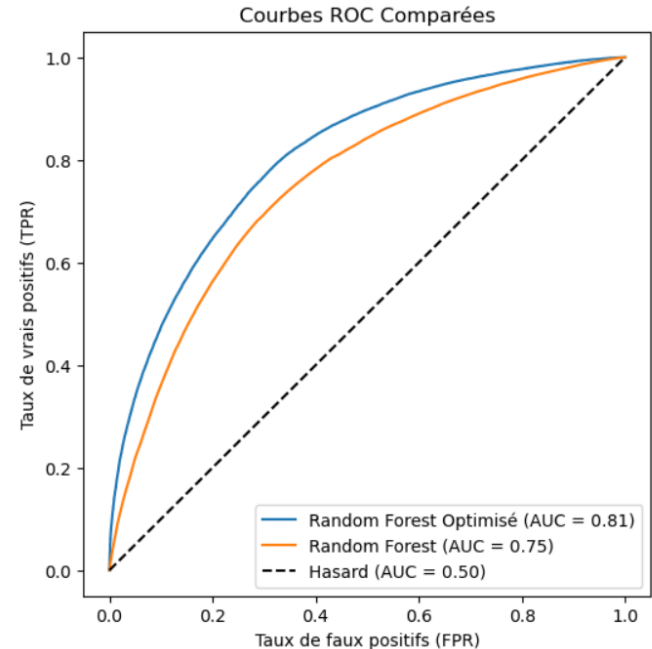
# 5 | Conclusions

## Bénéfices majeurs :

- Réduction significative des pertes financières
- Protection du système de santé
- Identification proactive des comportements frauduleux

## Modèle retenu :

- Random Forest optimisé
- AUC de 0.81
- Meilleur équilibre précision/rappel





# Merci pour votre attention

