

# Détection de la fraude à l'assurance

*Régression et classification en grande dimension*

Réalisé par :

- KHABIR Safia
- BOUATMANE Kaouthar
- EL MENOVAR Adam
- CAIDI Yassine

Encadré par :

Mr. JANATI Hicham

**Année universitaire 2024–2025**

# Plan de la présentation

- 1 Introduction
- 2 Données
- 3 Prétraitement et EDA
- 4 Analyse exploratoire des données
  - Distribution de la variable cible
  - Distribution des variables numériques
  - Analyse des corrélations
- 5 Préparation des données
- 6 Modélisation
- 7 Résultats et interprétation
- 8 Conclusion

## Contexte général

- La fraude à l'assurance représente un enjeu économique majeur.
- Les méthodes traditionnelles sont coûteuses et peu efficaces à grande échelle.
- L'augmentation du volume des sinistres complexifie la détection.
- Les données historiques offrent un fort potentiel d'exploitation.

## Question centrale

Comment exploiter efficacement les données de sinistres pour détecter les comportements frauduleux, dans un contexte de données hétérogènes, de grande dimension et fortement déséquilibrées, tout en garantissant des décisions interprétables pour les experts métier ?

# Objectifs du projet

- Préparer et analyser les données de sinistres.
- Comparer plusieurs modèles de classification supervisée.
- Gérer le déséquilibre des classes.
- Sélectionner le modèle le plus performant et le plus stable.
- Interpréter les résultats et analyser le seuil de décision.

# Présentation des données

- Source : Mendeley Data (Insurance Claims Fraud Dataset).
- 1000 sinistres automobiles.
- Variable cible : `fraud_reported`.

# Variables clés pour la détection de la fraude

- **Profil de l'assuré** : âge, ancienneté du client (`months_as_customer`), relation avec le souscripteur.
- **Caractéristiques du sinistre** : type et gravité de l'incident, heure de survenance, nombre de véhicules impliqués, présence de témoins et rapport de police.
- **Aspects financiers** : montants des indemnisations (corporelles, matérielles et véhicule), gains et pertes en capital.
- **Contrat et garanties** : prime annuelle, franchise (`policy_deductible`), limite de garantie (`umbrella_limit`).
- **Véhicule et dommages** : type de collision, dommages matériels, marque du véhicule.

# Nettoyage des données

- Suppression des variables non informatives ou purement identifiantes (numéro de police, localisation exacte, code postal).
- Traitement des valeurs manquantes pour certaines variables catégorielles en introduisant la modalité "Unknown".
- Vérification de la cohérence globale des données et absence de valeurs manquantes après nettoyage.

**Résultat :** un jeu de données propre et exploitable pour l'analyse et la modélisation.

# Analyse exploratoire des données

Cette section vise à explorer la structure des données, à identifier les caractéristiques principales des variables et à mettre en évidence les premiers signaux potentiellement liés à la fraude à l'assurance. Une attention particulière est portée à la distribution de la variable cible, aux variables numériques clés et aux relations entre variables.

La variable cible `fraud_reported` indique si un sinistre est frauduleux (Y) ou non (N). La Figure 1 met en évidence un fort déséquilibre de classes, avec une majorité de sinistres non frauduleux.

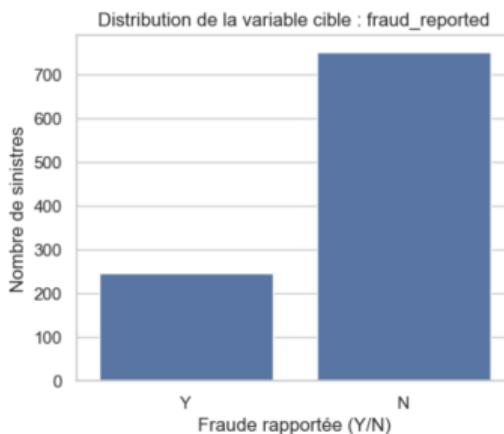


Figure – Distribution de la variable cible `fraud_reported`

Ce déséquilibre justifie l'utilisation de techniques spécifiques telles que le rééchantillonnage (SMOTE) et l'évaluation via des métriques adaptées (Recall, F1-score, ROC-AUC).

Les Figures 2 et 3 présentent les distributions de quelques variables numériques clés.

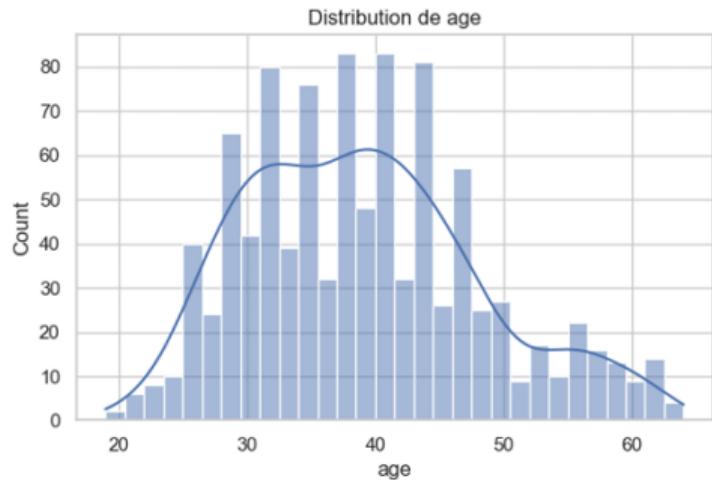
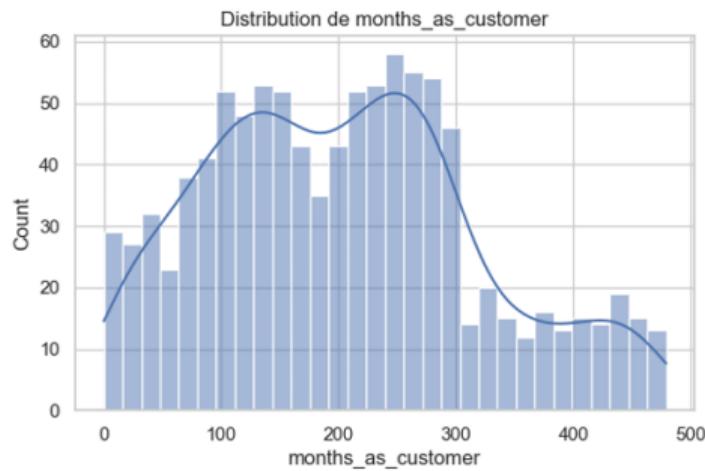


Figure – Distribution de months\_as\_customer et de l'âge de l'assuré

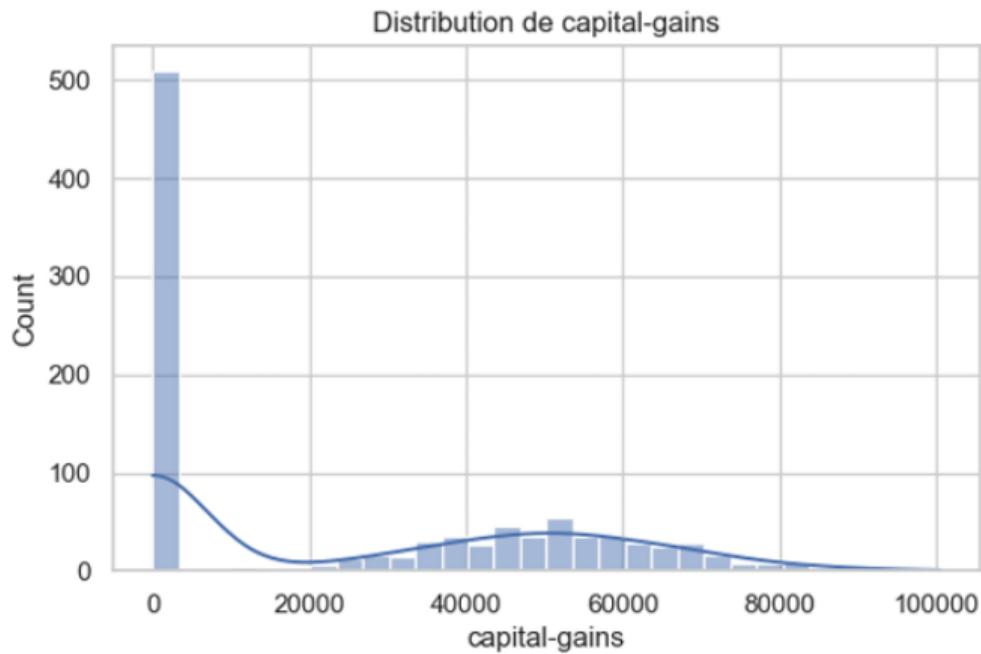
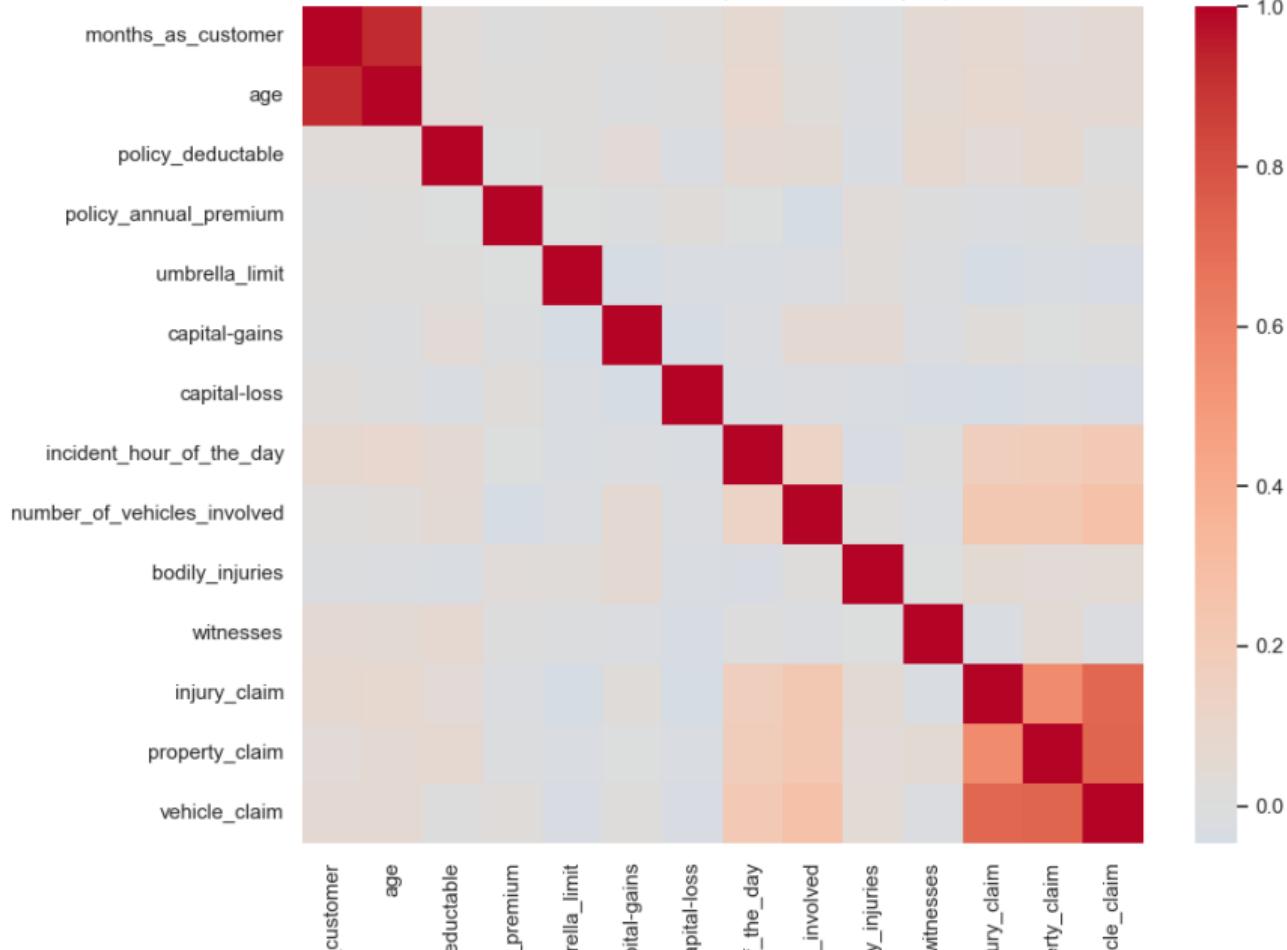


Figure – Distribution de la variable capital\_gains

On observe des distributions asymétriques et la présence de valeurs extrêmes, notamment pour les variables financières, ce qui justifie un traitement spécifique des outliers.  
La Figure 4 présente la matrice de corrélation des principales variables numériques.

Matrice de corrélation (variables numériques)



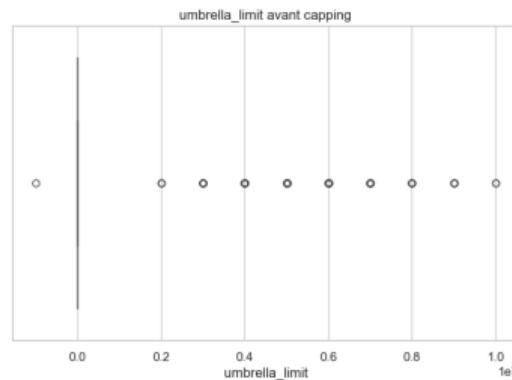
## Analyse des corrélations

Les résultats montrent des corrélations globalement faibles, à l'exception des variables liées aux montants des sinistres (`injury_claim`, `property_claim`, `vehicle_claim`), ce qui limite les risques de multicolinéarité.

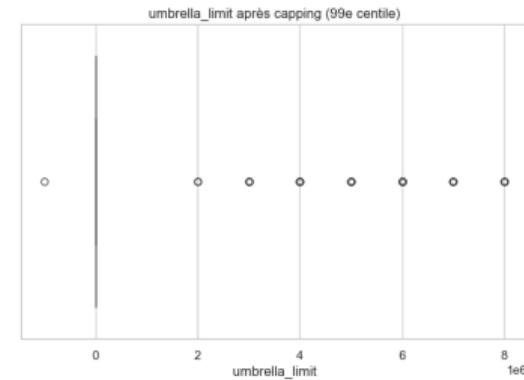
# Préparation des données

## Traitement des valeurs aberrantes

Les valeurs extrêmes de certaines variables numériques, notamment `umbrella_limit`, peuvent biaiser l'apprentissage des modèles. Une technique de *capping* au 99<sup>e</sup> centile a été appliquée afin de limiter leur influence sans supprimer d'observations.



Avant capping



Après capping (99<sup>e</sup> centile)

# Préparation des données

## Encodage et mise à l'échelle

- Les variables catégorielles ont été transformées par *one-hot encoding*.
- Les variables numériques ont été standardisées afin d'assurer une échelle homogène pour les algorithmes sensibles à la distance (régression logistique, SVM).

## Rééquilibrage des classes (SMOTE)

- La variable cible présente un fort déséquilibre entre fraude et non-fraude.
- La méthode **SMOTE** a été appliquée uniquement sur le jeu de entraînement.
- Objectif : améliorer la capacité des modèles à détecter les sinistres frauduleux.

## Conclusion

La préparation finale garantit des données équilibrées, normalisées et adaptées à une modélisation robuste.

## Modèles étudiés

Quatre modèles de classification supervisée ont été évalués pour la détection de la fraude :

- **Régression logistique** : modèle de référence, simple et interprétable.
- **SVM** : modèle non linéaire adapté aux données de grande dimension.
- **Random Forest** : méthode densemble capturant des relations complexes.
- **Voting Classifier** : combinaison de modèles pour améliorer la performance.

**Objectif** : identifier le modèle le plus performant pour la détection de la fraude.

# Comparaison des performances des modèles

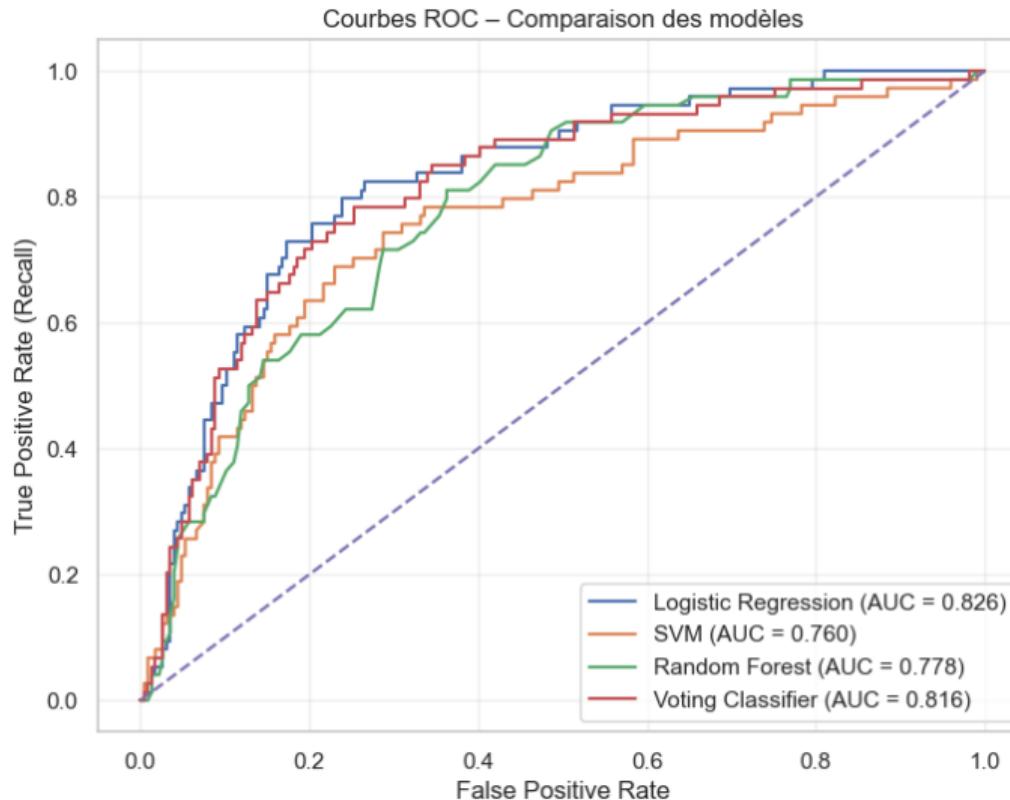
Modèle	Accuracy	Précision	Rappel	F1-score	ROC-AUC
Régression logistique	0.800	0.621	0.486	<b>0.545</b>	<b>0.826</b>
Voting Classifier	0.790	<b>0.634</b>	0.351	0.452	0.816
Random Forest	0.770	0.568	0.284	0.378	0.778
SVM	0.757	0.526	0.135	0.215	0.760

Figure – Comparaison des performances des modèles sur le jeu de test

## Lecture du tableau :

- La **régression logistique** présente le meilleur F1-score.
- Le **Voting Classifier** offre un bon compromis précision–rappel.
- Le **Random Forest** et le **SVM** sont moins efficaces sans ajustement de seuil.

# Courbes ROC Comparaison des modèles



# Validation croisée multi-métriques

Table – Résultats moyens de la validation croisée (5-fold)

Modèle	Accuracy	Précision	Rappel	F1-score	ROC-AUC
Régression logistique	0.888	0.910	0.863	0.885	0.959
SVM	0.917	<b>0.963</b>	0.869	0.913	0.978
Random Forest	<b>0.929</b>	0.951	0.905	<b>0.927</b>	<b>0.980</b>
Voting Classifier	0.922	0.936	<b>0.907</b>	0.921	0.979

## Conclusion :

- Le **Random Forest** obtient les meilleures performances globales.
- Il présente le meilleur compromis entre **rappel**, **F1-score** et **ROC-AUC**.
- Ce modèle est retenu pour l'évaluation finale sur les données réelles.

# Analyse du seuil de décision Random Forest

Seuil	Précision	Rappel	F1-score
0.5	0.553	0.284	0.375
0.4	0.494	0.581	0.534
0.3	0.401	0.824	0.540

## Choix du seuil de décision

Seuil retenu : 0.3

Le seuil de décision a été abaissé afin de maximiser la capacité de détection des fraudes.

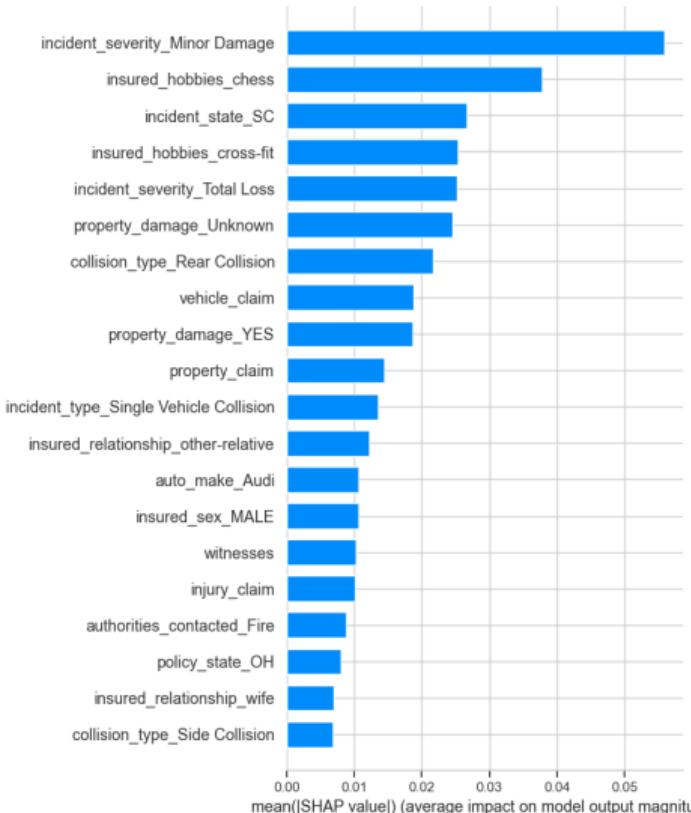
- Le rappel atteint **82.4%**, ce qui réduit fortement le risque de fraude non détectée.
- Laugmentation des faux positifs est jugée acceptable dans un contexte de contrôle.
- Le modèle est utilisé comme **outil daide à la décision** et non de décision automatique.

# Interprétabilité du modèle SHAP

Les modèles de type **Random Forest** sont performants mais peu interprétables. Afin de comprendre les décisions du modèle retenu, nous avons utilisé la méthode **SHAP (SHapley Additive exPlanations)**.

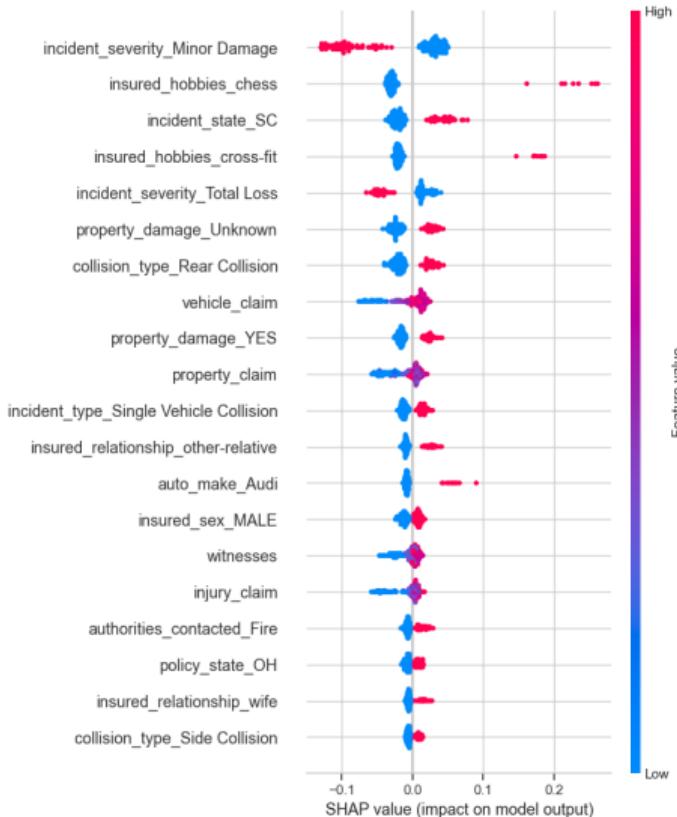
- SHAP mesure la contribution de chaque variable à la prédiction finale.
- Les contributions peuvent être positives ou négatives.
- L'interprétation est cohérente avec les principes de la théorie des jeux.

# Variables les plus influentes (SHAP)



- La gravité du sinistre est le facteur le plus déterminant.
- Les dommages matériels et les montants réclamés ont un impact fort.
- Certaines caractéristiques assurantielles jouent un rôle secondaire.

# Analyse fine des contributions (SHAP)



- Chaque point représente l'impact d'une observation individuelle.
- Les valeurs élevées (rouge) peuvent augmenter ou diminuer la probabilité de fraude.
- La variabilité des contributions souligne l'hétérogénéité des sinistres.

## Conclusion générale

- La fraude à l'assurance constitue un enjeu majeur pour les compagnies d'assurance.
- Une démarche complète a été mise en œuvre :
  - préparation et nettoyage des données,
  - analyse exploratoire approfondie,
  - comparaison de plusieurs modèles supervisés.
- La validation croisée multi-métriques a permis d'identifier le **Random Forest** comme modèle le plus performant.
- L'analyse de seuil montre qu'un seuil de **0,3** permet de mieux détecter les sinistres frauduleux en privilégiant le rappel.
- L'interprétation SHAP renforce la transparence du modèle et facilite son exploitation métier.

**Ce travail illustre l'apport des méthodes de machine learning pour une détection plus efficace et interprétable de la fraude.**

# Merci pour votre attention

---

*Nous restons à votre disposition pour vos questions*