Projet de Machine Learning

Détection de Fraude des Prestataires de Santé aux États-Unis

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

M2 Actuariat – 2024/2025

ISUP

À rendre le 07 Janvier 2025

Table des matières

1 Contexte

1.1 Medicare

Medicare est un système d'assurance santé géré par le gouvernement fédéral des États-Unis, conçu pour offrir une couverture médicale aux personnes âgées de 65 ans et plus, ainsi qu'à certains groupes spécifiques. Le nom « Medicare » est issu d'une contraction des mots anglais *medical* (médical) et *care* (soin).

Origines et histoire
Medicare a été officiellement créé le 30 juillet 1965 sous la présidence de Lyndon B. Johnson, dans le cadre de sa « guerre contre la pauvreté ». Ce programme a été introduit par un amendement à la Loi sur la Sécurité sociale. Le président Johnson a signé la loi dans la bibliothèque présidentielle de Harry S. Truman, située à Independence, dans le Missouri. À cette occasion, l'ancien président Harry S. Truman et son épouse Bess Truman sont devenus les premiers bénéficiaires du programme Medicare, recevant la première carte Medicare.

Avant cette création, le terme « Medicare » désignait initialement un programme de soins médicaux destiné aux familles des militaires, mis en place par la Loi sur les soins médicaux des personnes à charge adoptée en 1956. Ce n'est qu'en 1965 que Medicare a pris sa forme actuelle, en vertu du titre XVIII de la Loi sur la Sécurité sociale.

Critères d'éligibilité
En règle générale, Medicare est accessible aux individus remplissant les conditions suivantes :
- Avoir 65 ans ou plus ;
- Résider de manière permanente aux États-Unis ;
- Avoir cotisé à la sécurité sociale ou au fonds de financement des retraites des cheminots pendant au moins dix ans.

Certaines exceptions permettent également aux individus de moins de 65 ans d'être éligibles, notamment s'ils souffrent d'un handicap reconnu, d'insuffisance rénale terminale (IRT) ou de sclérose latérale amyotrophique (maladie de Lou Gehrig).

Structure et parties de Medicare
Medicare est divisé en plusieurs parties, chacune répondant à des besoins spécifiques :
- **Partie A (assurance hospitalisation)** : Couvre les frais liés aux hospitalisations, aux soins en hospice, aux soins à domicile et aux séjours en établissements de soins qualifiés. La plupart des bénéficiaires ne paient pas de prime pour cette partie s'ils ou leurs conjoints ont suffisamment cotisé.
- **Partie B (assurance médicale)** : Couvre les services médicaux nécessaires tels que les consultations, les soins ambulatoires, les équipements médicaux durables (comme les fauteuils roulants) et les services préventifs. Une prime mensuelle standard est généralement appliquée.
- **Partie C (Medicare Advantage)** : Plans offerts par des compagnies privées qui regroupent les prestations des parties A et B, et souvent la partie D. Ces plans incluent parfois des services supplémentaires, tels que les soins dentaires, visuels et auditifs, non couverts par l'Original Medicare.
- **Partie D (couverture des médicaments)** : Aide à couvrir les coûts des médicaments prescrits, y compris certains vaccins. Cette partie est administrée par des compagnies privées suivant les règles fixées par Medicare.

Financement de Medicare
Le financement de Medicare provient principalement des cotisations sociales payées par les travailleurs et leurs employeurs, des primes versées par les bénéficiaires, et des fonds généraux du gouvernement fédéral.

— **Partie A** est financée principalement par une taxe dédiée sur les revenus des salariés (2,9%, partagée entre l'employé et l'employeur), avec un supplément pour les revenus élevés.
— **Parties B et D** sont financées en partie par des primes payées par les bénéficiaires et en partie par le Trésor public.
— **Partie C** est financée par des paiements combinés des autres parties et des primes supplémentaires pour les prestations additionnelles.

Forces et défis

Medicare est l'un des piliers du système de santé américain, assurant la couverture de millions de personnes âgées ou vulnérables. Cependant, plusieurs défis menacent sa pérennité :

— **Croissance démographique et baby-boomers** : Avec l'augmentation du nombre de retraités, la pression sur le système augmente. D'ici 2030, environ 77 millions d'Américains pourraient être couverts par Medicare, ce qui représente un défi pour le financement.
— **Coûts croissants des soins de santé** : L'augmentation des dépenses liées aux soins hospitaliers, aux médicaments et aux technologies médicales constitue une charge importante pour Medicare.
— **Fraude et abus** : Des pratiques telles que la facturation de services non fournis ou inutiles, le "upcoding", et les pots-de-vin alourdissent les coûts de Medicare et compromettent l'intégrité du programme.
— **Coordination avec d'autres programmes** : Medicare travaille en partenariat avec Medicaid pour couvrir les populations les plus vulnérables. Les chevauchements et les écarts dans les couvertures nécessitent une meilleure intégration.

Importance sociale et économique

Medicare offre une sécurité essentielle pour les populations âgées et vulnérables, réduisant les inégalités en matière de santé. En permettant un accès universel aux soins hospitaliers et médicaux, il joue un rôle clé dans l'amélioration de la qualité de vie des Américains tout en stimulant le secteur des soins de santé.

Malgré ses défis, Medicare reste une réussite emblématique de la politique sociale américaine, symbolisant l'engagement envers les générations les plus âgées et les plus fragiles de la société.

1.2 Les différents types de fraudes dans le domaine de la santé

La fraude dans le secteur de la santé constitue un enjeu majeur, particulièrement dans des programmes publics tels que Medicare et Medicaid. Ces pratiques frauduleuses, souvent orchestrées par une minorité de prestataires malhonnêtes ou par des groupes criminels organisés, entraînent des coûts financiers considérables et des risques pour les patients. Voici les principaux types de fraudes observés dans le domaine de la santé :

Fraudes commises par les prestataires de santé

Les prestataires de santé malhonnêtes peuvent recourir à diverses pratiques frauduleuses, notamment :

— **Double facturation** : Soumission de plusieurs réclamations pour un même service.
— **Facturation fictive ("phantom billing")** : Facturation de visites, de services ou de fournitures que le patient n'a jamais reçus.
— **Dégroupage ("unbundling")** : Facturation séparée de services qui devraient être regroupés en un forfait unique, afin d'augmenter les remboursements.
— **"Upcoding"** : Facturation d'un service plus coûteux que celui effectivement fourni.
— **Facturation de services ou d'articles non fournis** : Présentation de factures pour des examens, des tests ou des fournitures inexistants.
— **Pots-de-vin** : Offre ou réception d'une rémunération en échange de recommandations de patients ou de services.

— **Diagnostics falsifiés** : Manipulation des diagnostics pour justifier des traitements, tests ou chirurgies inutiles.

Fraudes commises par les patients ou d'autres individus

Les patients ou d'autres individus peuvent également être à l'origine de fraudes, par exemple :

— **Marketing frauduleux** : Convaincre des individus de fournir leur numéro d'identification d'assurance maladie et d'autres informations personnelles pour facturer des services non fournis, voler leur identité ou les inscrire à des plans fictifs.
— **Vol ou usurpation d'identité** : Utilisation de l'assurance maladie d'une autre personne ou autorisation donnée à autrui d'utiliser son assurance.
— **Usurpation d'identité professionnelle** : Fourniture de services ou équipements de santé sans licence appropriée.

Fraudes liées aux prescriptions

Les prescriptions médicales sont également une cible fréquente des fraudes :

— **Falsification ("forgery")** : Création ou utilisation d'ordonnances falsifiées.
— **Détournement ("diversion")** : Réutilisation illégale de prescriptions, par exemple pour revendre des médicaments obtenus légalement.
— **"Doctor shopping"** : Consultation de multiples médecins pour obtenir des prescriptions de substances contrôlées ou recours à des cliniques peu scrupuleuses pour obtenir ces substances.

Conséquences

Les sanctions prévues par la loi incluent des amendes, des peines de prison, et l'exclusion des programmes publics. Une coopération étroite entre les secteurs privé et public est essentielle pour détecter, prévenir et sanctionner ces actes frauduleux.

1.3 Les enjeux de la fraude

La fraude dans le domaine de la santé affecte à la fois les institutions publiques et les assureurs santé privés, générant des coûts énormes et compromettant l'efficacité des systèmes de santé. Ces impacts sont ressentis non seulement financièrement, mais également en termes de confiance et de qualité des soins.

1. Coûts financiers et systémiques

En 2018, les dépenses en santé aux États-Unis ont atteint 3,6 trillions de dollars, incluant des milliards de réclamations d'assurance maladie. Bien que seules une petite fraction de ces réclamations soient frauduleuses, les pertes associées sont considérables.

Le National Health Care Anti-Fraud Association (NHCAA) estime que les pertes financières dues à la fraude dans la santé s'élèvent à des dizaines de milliards de dollars chaque année. Une estimation conservatrice indique que la fraude représente 3% des dépenses totales, tandis que d'autres agences placent ce chiffre entre 700 et 800 milliards de dollars par an, soit jusqu'à 10% des dépenses annuelles en santé.

2. Impact sur les primes et les coûts des consommateurs

La fraude entraîne une augmentation directe des primes et des dépenses personnelles pour les consommateurs, ainsi qu'une réduction des prestations ou des couvertures. Les employeurs, qu'ils soient privés ou publics, supportent également ces coûts accrus pour fournir des prestations d'assurance à leurs employés, ce qui augmente les coûts d'exploitation.

Pour de nombreux Américains, cette hausse des coûts due à la fraude peut signifier la différence entre pouvoir ou non se permettre une assurance santé.

3. Conséquences humaines

La fraude en matière de santé n'est pas un crime sans victime. Les individus directement touchés incluent :

- Les patients soumis à des traitements inutiles ou dangereux.
- Ceux dont les dossiers médicaux sont compromis ou dont les informations d'assurance sont utilisées à des fins frauduleuses.

Ces situations peuvent avoir des conséquences graves sur leur santé, leur vie privée et leur accès futur aux soins.

4. Solutions technologiques pour prévenir la fraude

L'intelligence artificielle (IA) est un outil clé dans la lutte contre la fraude. Des plateformes comme Integr8 AI offrent des capacités avancées, notamment :

- L'analyse contextuelle des réclamations en prenant en compte les comportements des prestataires, les relations entre prestataires, les données patient, et d'autres variables clés.
- Une détection accrue des paiements frauduleux, doublant voire décéuplant les capacités des technologies classiques.
- Une réduction significative des pertes dues à la fraude, aux abus et aux gaspillages.

Ces outils permettent de prévenir des pertes massives et de renforcer l'intégrité du système.

5. Données chiffrées et visualisations

Les données ci-dessous illustrent l'étendue de la fraude en matière de soins de santé aux États-Unis :

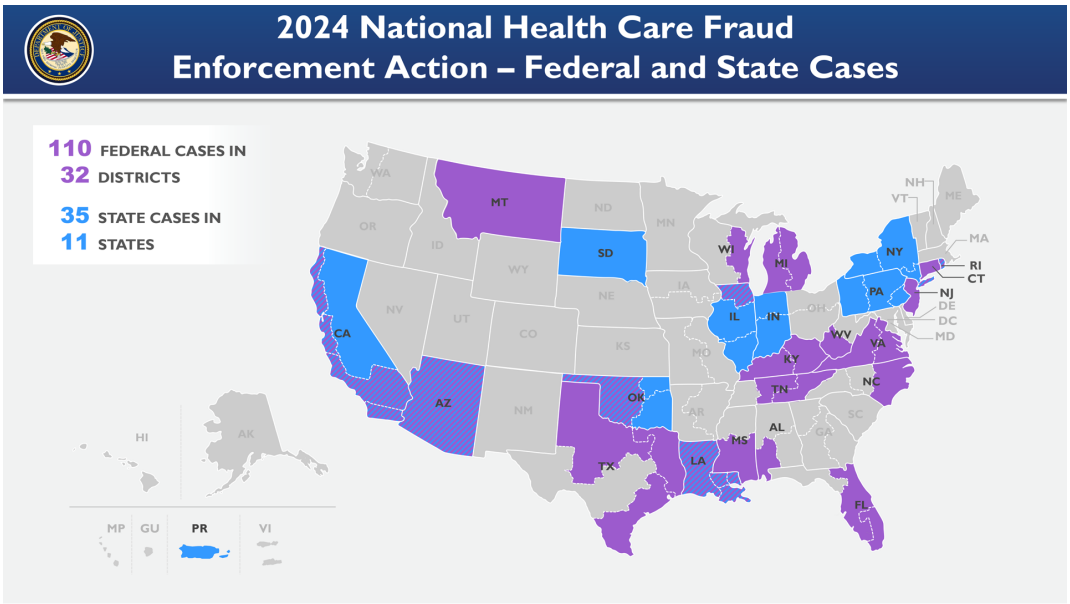


FIGURE 1 – Carte de la distribution géographique des cas de fraude en 2024

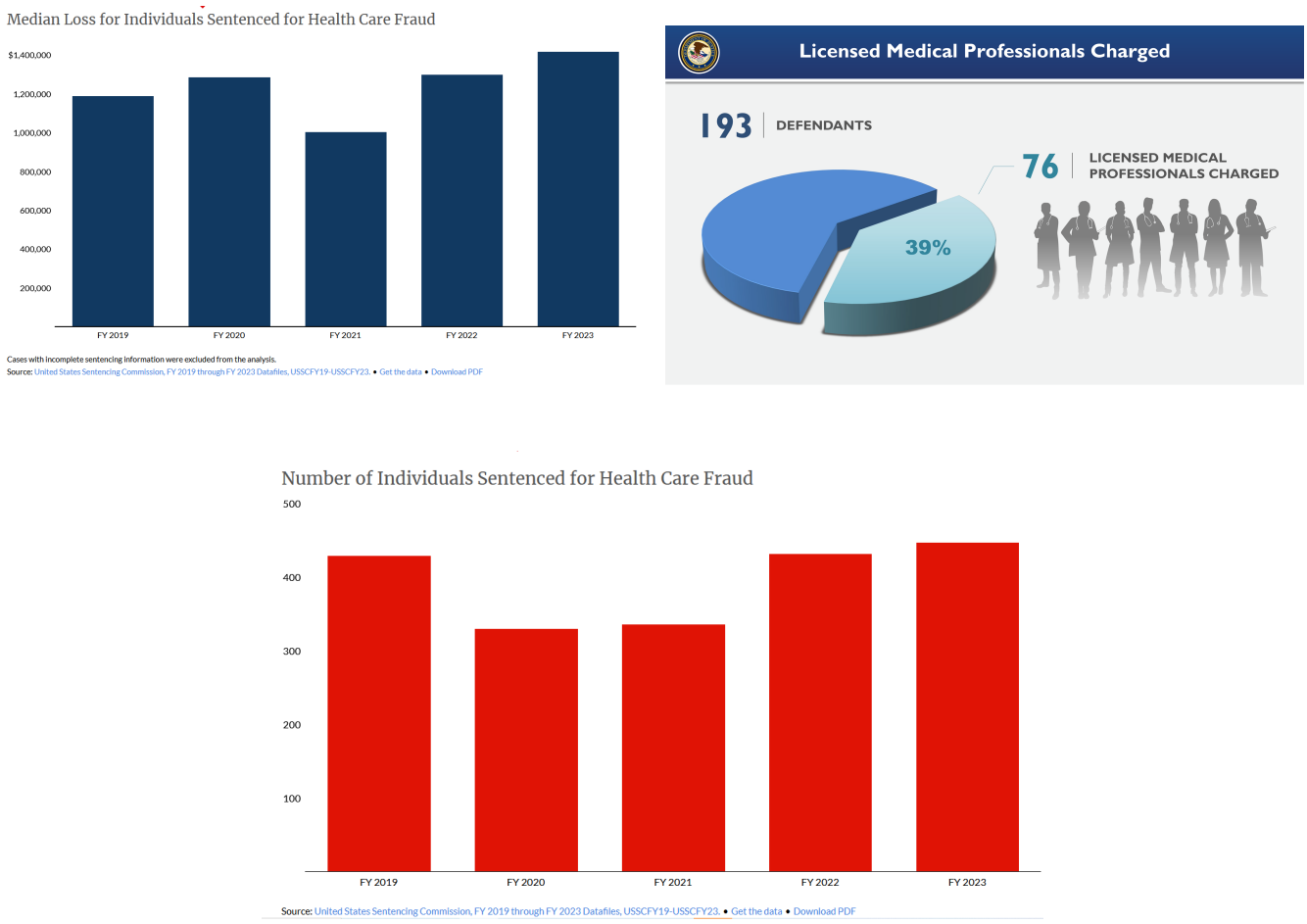


FIGURE 2

Ces graphiques soulignent l'ampleur du problème et la nécessité d'une action coordonnée pour réduire les pertes et améliorer la confiance dans le système de santé.

2 Problématique

Le système de santé américain est lourdement impacté par des fraudes organisées impliquant des prestataires, des médecins et des bénéficiaires. Ces fraudes génèrent des réclamations fictives qui augmentent les dépenses de Medicare, affectant directement les compagnies d'assurance et les assurés à travers des primes élevées et une augmentation du coût global des soins de santé. Dans ce contexte, il est essentiel de développer des modèles prédictifs capables d'identifier efficacement les prestataires potentiellement frauduleux. La question principale est donc : **comment peut-on utiliser les données des réclamations médicales pour détecter de manière fiable les comportements frauduleux parmi les prestataires de soins de santé ?**

3 Données

3.1 Présentation des données

Les jeux de données utilisés dans ce projet sont des données réelles, librement accessibles sur Kaggle. Elles se composent de plusieurs fichiers distincts, regroupés en quatre grandes catégories : les données sur les prestataires (**Providers**), les bénéficiaires (**Beneficiaries**), et les réclamations médicales pour soins hospitaliers (**Inpatient**) ou ambulatoires (**Outpatient**). Chaque fichier contient des informations spécifiques permettant d'analyser les schémas de réclamations et de détecter les comportements frauduleux. Ces fichiers seront fusionnés pour constituer une base de données complète, offrant une vision cohérente et centralisée des informations nécessaires au projet de détection de fraude.

1. Données des prestataires (Providers)

Ces données contiennent des informations essentielles sur les prestataires de soins de santé, notamment une étiquette indiquant si le prestataire est impliqué dans une fraude. Cela constitue la variable cible du projet, essentielle pour entraîner et évaluer les modèles prédictifs.

Colonnes principales :

- Identifiant du prestataire.
- Indicateur de fraude (1 = fraude, 0 = non-fraude).

2. Données des bénéficiaires (Beneficiaries)

Ces données regroupent des informations sur les patients bénéficiant des soins, fournissant un contexte démographique et médical qui peut aider à comprendre les réclamations et les tendances de fraude.

Colonnes principales :

- Identifiant du bénéficiaire.
- Informations démographiques (sexe, région).
- Informations médicales (présence de conditions chroniques comme le diabète ou les maladies cardiaques).

3. Données hospitalières (Inpatient Data)

Les données hospitalières (**Inpatient Data**) concernent les réclamations déposées pour les patients hospitalisés. Elles peuvent être regroupées en trois catégories principales :

- Informations générales sur la réclamation.
- Informations cliniques.
- Procédures réalisées.

4. Données ambulatoires (Outpatient Data)

Les données ambulatoires (**Outpatient Data**) concernent les réclamations déposées pour les patients traités sans hospitalisation. Elles suivent une structure similaire aux données hospitalières et peuvent également être regroupées en trois catégories principales :

- Informations générales sur la réclamation.
- Informations cliniques.
- Procédures réalisées.

Intégration des données

Les fichiers des prestataires, bénéficiaires, soins hospitaliers et ambulatoires seront fusionnés sur les colonnes communes (telles que `BeneID`, `ClaimID` et `Provider`). Cela permettra d'obtenir une base de données complète et centralisée, intégrant des informations démographiques, cliniques et financières, essentielles pour détecter les comportements frauduleux.
Cette intégration facilitera une analyse approfondie et la création de modèles prédictifs robustes pour identifier les prestataires à risque.

3.2 Data Preprocessing

Le data preprocessing est une étape cruciale dans tout projet de machine learning. La qualité des données utilisées influence directement la performance des modèles. Avant toute phase d'entraînement, il est essentiel de transformer, nettoyer et structurer les données afin de garantir leur cohérence, leur complétude et leur pertinence. Cette étape permet également d'assurer que toutes les informations nécessaires sont correctement intégrées dans une base de données exploitable par les algorithmes de machine learning.

3.2.1 Construction de la base de données

Pour ce projet, la construction de la base de données a consisté à regrouper les informations issues de plusieurs fichiers distincts afin d'obtenir une vue complète et centralisée des données. Les fichiers suivants ont été utilisés :

- **InPatient** : Données sur les soins reçus par les patients hospitalisés.
- **OutPatient** : Données sur les soins reçus en ambulatoire.
- **Benef** : Données médicales des bénéficiaires.
- **Providers** : Informations sur les prestataires, incluant l'indicateur de fraude.

Pour combiner ces fichiers, plusieurs jointures ont été effectuées sur des colonnes communes comme les identifiants des bénéficiaires et des prestataires. Ce processus a permis d'unifier les données en une seule table.
Le schéma ci-dessous illustre le processus de construction de la base de données :

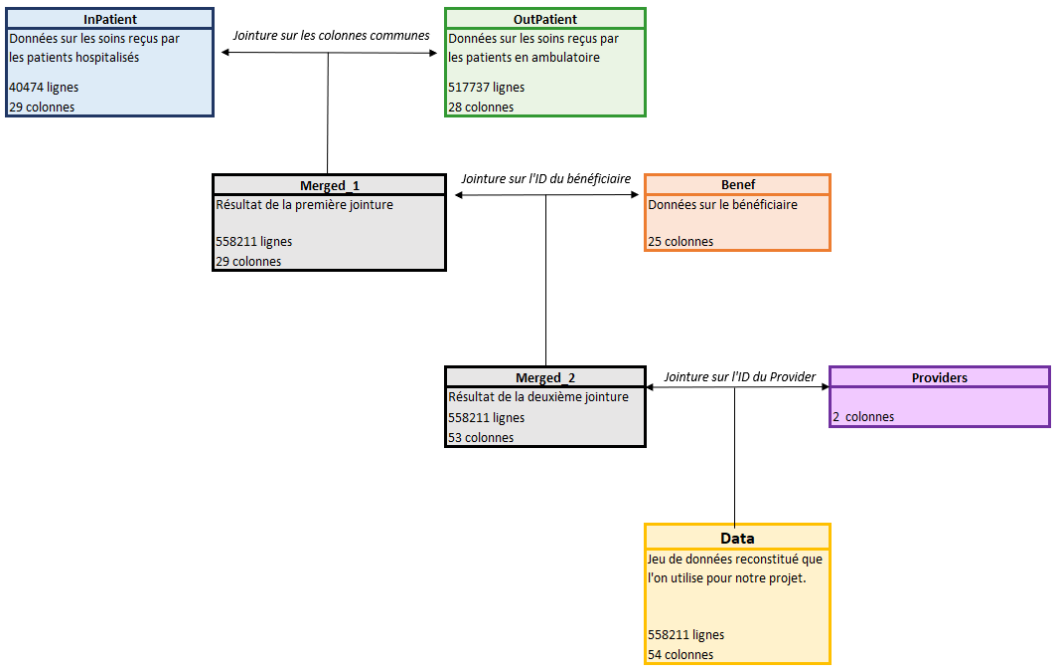


FIGURE 3 – Construction de la base de données

La table finale, nommée **Data**, contient 558211 lignes et 54 colonnes. Elle inclut des informations détaillées sur les diagnostics, les procédures, les montants facturés et les caractéristiques des prestataires, offrant ainsi une base solide pour l'entraînement des modèles de détection de fraude.

3.2.2 Traitement des valeurs manquantes

En machine learning, le traitement des valeurs manquantes est une étape essentielle du prétraitement des données. Les valeurs manquantes peuvent biaiser les modèles ou réduire leur capacité à généraliser, surtout si ces absences sont fréquentes ou concentrées sur des variables clés. Il est donc crucial d'analyser les motifs des valeurs manquantes et de prendre des décisions appropriées, telles que leur imputation, leur suppression, ou encore la transformation des colonnes.

Colonnes `ClmDiagnosisCode`

Dans notre jeu de données, 10 colonnes (`ClmDiagnosisCode_1` à `ClmDiagnosisCode_10`) contiennent les codes des diagnostics posés lors des hospitalisations des patients. Ces colonnes reflètent les diagnostics enregistrés pour un patient pendant son séjour à l'hôpital :
- Les colonnes (`ClmDiagnosisCode_1` à `ClmDiagnosisCode_10`) sont remplies séquentiellement, de la première jusqu'à la dernière, en fonction du nombre de diagnostics posés pendant l'hospitalisation. Par exemple, si 3 diagnostics ont été posés, seules les colonnes `ClmDiagnosisCode_1`, `ClmDiagnosisCode_2` et `ClmDiagnosisCode_3` seront remplies, tandis que les colonnes suivantes (`ClmDiagnosisCode_4` à `ClmDiagnosisCode_10` resteront vides. - Ainsi, le nombre de colonnes remplies correspond exactement au nombre de diagnostics posés pour cette hospitalisation.
Nous avons constaté que le taux de valeurs manquantes dépasse 70% à partir de `ClmDiagnosisCode_4`, rendant ces colonnes peu fiables et peu utiles. Pour résoudre ce problème, nous avons décidé de :

- Supprimer les 10 colonnes `ClmDiagnosisCode`.
- Créer une nouvelle colonne `NbDiagnosis` qui indique directement le nombre de diagnostics posés pour chaque hospitalisation.

Cette transformation permet de conserver l'information essentielle sur le nombre de diagnostics tout en réduisant la complexité des données et en traitant le problème des valeurs manquantes pour les diagnostiques.

Colonnes `ClmProcedureCode`

Les 5 colonnes `ClmProcedureCode_1` à `ClmProcedureCode_5` contiennent des codes spécifiques identifiant les procédures chirurgicales, médicales ou diagnostiques réalisées pendant l'hospitalisation d'un patient. Cependant, nous avons observé des taux de valeurs manquantes très élevés (entre 95% et 100%) dans ces colonnes, ce qui rend leur utilisation impraticable. Contrairement aux colonnes des diagnostics, il est impossible de reconstituer ces informations manquantes avec un degré de certitude suffisant.
Ainsi, nous avons décidé de supprimer entièrement ces 5 colonnes, car elles n'apportent pas d'informations exploitables pour la modélisation.

Colonne `DeductibleAmtPaid`

La colonne `DeductibleAmtPaid` représente le montant que le patient doit payer pour les services de santé couverts avant que le plan d'assurance ne commence à rembourser les frais. Ce champ contient 0,16% de valeurs manquantes, un taux faible mais non négligeable. Afin de garantir l'intégrité des données et de ne pas perdre d'échantillons utiles, nous avons choisi de combler ces valeurs manquantes par une méthode d'imputation.
L'imputation a été réalisée en remplaçant les valeurs manquantes par la médiane de la colonne `DeductibleAmtPaid`. Cette approche est robuste face aux valeurs extrêmes (outliers) et garantit que la distribution de la va-

riable reste cohérente. Cette stratégie permet également d'éviter d'introduire un biais potentiel qui pourrait affecter les performances du modèle.

Conclusion

Ces décisions de traitement des valeurs manquantes nous ont permis de simplifier et de structurer le jeu de données, tout en préservant les informations essentielles sur les diagnostics réalisés pendant les hospitalisations des patients.

3.2.3 Autres traitements appliqués aux colonnes

Dans cette étape, nous avons réalisé plusieurs transformations pour améliorer la qualité et la pertinence des données utilisées dans notre modèle. Ces transformations visent à réduire le bruit dans les données et à conserver uniquement les informations les plus utiles.

Ajout de la colonne `Patient_Type`
Nous avons ajouté une colonne `Patient_Type` afin de différencier les soins en hospitalisation et les soins ambulatoires. Cela permet de conserver cette information importante dans notre modèle.

Création de la colonne `DureeClaim`
Les colonnes `ClaimStartDt` (date de début de l'hospitalisation) et `ClaimEndDt` (date de fin de l'hospitalisation) fournissent des informations temporelles. Cependant, pour rendre ces données plus pertinentes, nous avons calculé la durée de l'hospitalisation (`DureeClaim`) à l'aide de la formule suivante :

$$\texttt{DureeClaim} = \texttt{ClaimEndDt} - \texttt{ClaimStartDt} + 1$$

Avant de supprimer les colonnes `ClaimStartDt` et `ClaimEndDt`, nous avons vérifié que la proportion de fraudes était relativement stable au cours des mois. Cette stabilité a été confirmée, comme illustré dans l'image suivante :

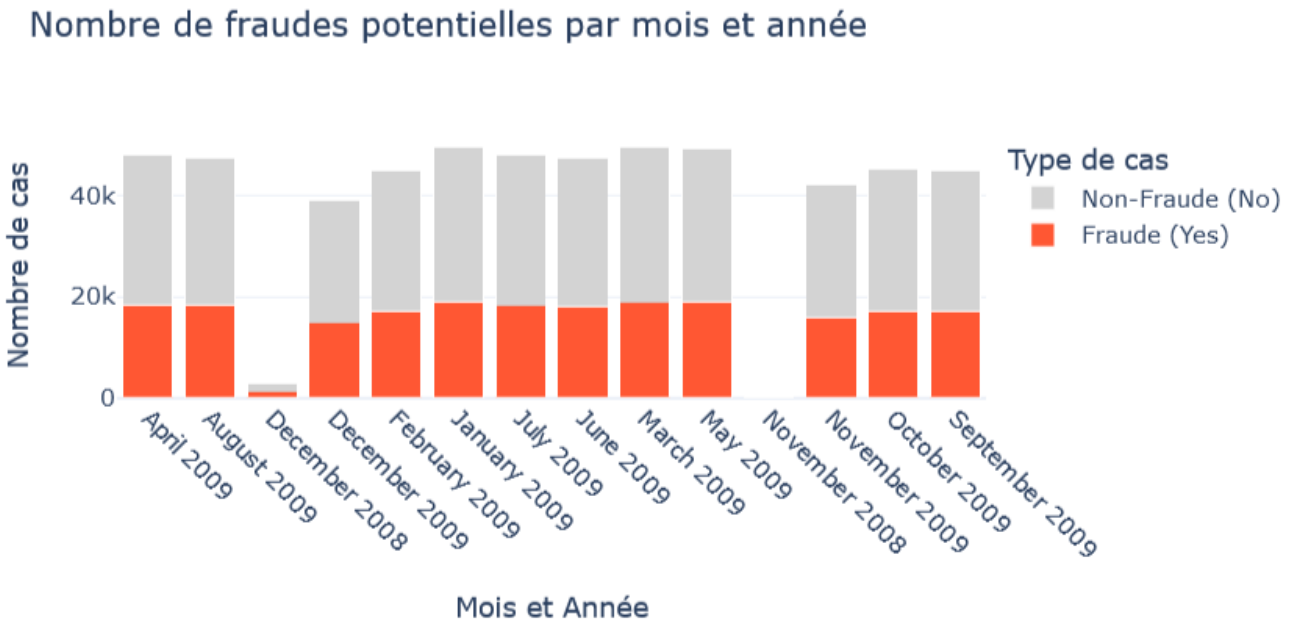


FIGURE 4 – Répartition des fraudes par mois.

Création de la colonne `StateRisk`
La colonne `State` contient 54 États différents. Pour simplifier le modèle et réduire le nombre de classes, nous avons regroupé les États en niveaux de risque (`StateRisk`). Ce regroupement est basé sur un score de risque (`Risk_Score`) calculé comme suit :

$$\texttt{Risk\_Score} = \texttt{Fraud\_Percentage} \times \log(\texttt{Total\_Count} + 1)$$

Les États sont ensuite classés en 4 niveaux de risque (`Risk_Level`) :
- 1 : Faible risque.
- 2 : Risque modéré.
- 3 : Risque élevé.
- 4 : Risque très élevé.

Cette classification est fondée à la fois sur le pourcentage de fraudes (`Fraud_Percentage`) et le nombre total de cas (`Total_Count`) ; Les graphes ci-dessous illustrent l'importance de considérer à la fois la part des fraudes et la quantité de données disponibles par États :

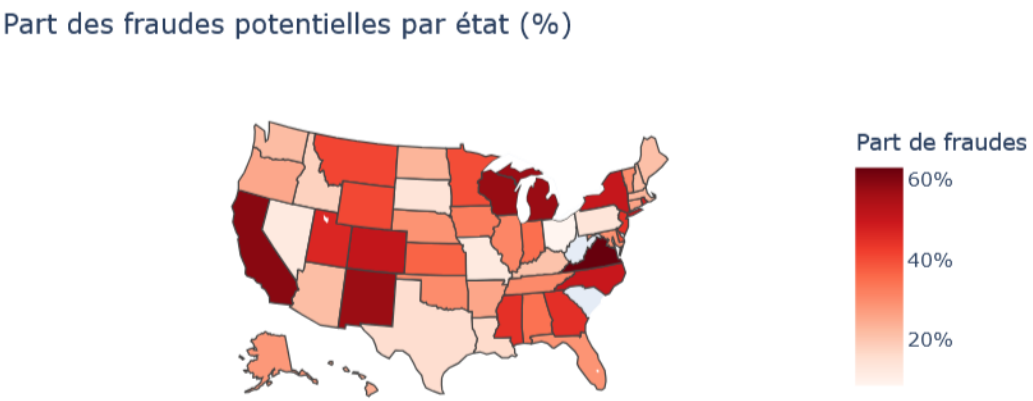


FIGURE 5 – Part des fraudes potentielles par État (%).

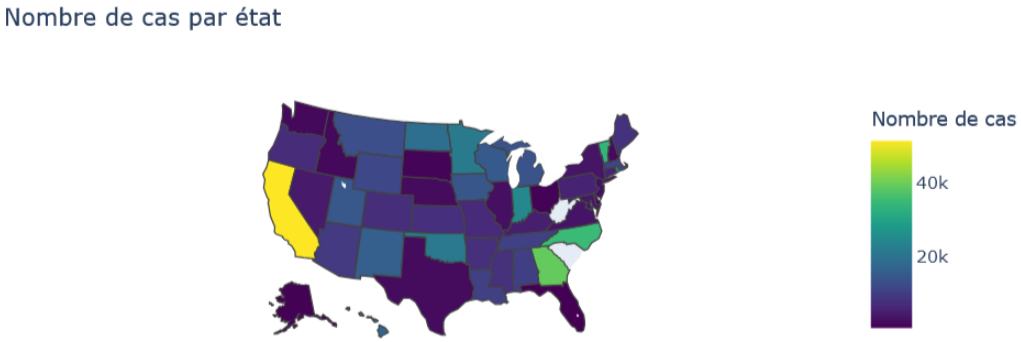


FIGURE 6 – Nombre de cas par État.

Nous avons supprimé la colonne `State` après la création de `StateRisk`.

Création des colonnes `CountyRisk` et `CodeProvider`
De manière similaire, nous avons créé :

— `CountyRisk` : Une classification des comtés basée sur le risque calculé à partir de la colonne `County`.

Nous avons supprimé les lignes contenant des valeurs dépassant le 85ème percentile pour les colonnes `InscClaimAmtReimbursed`, `DeductibleAmtPaid`, `IPAnnualReimbursementAmt`, et `IPAnnualDeductibleAmt`, afin de limiter l'influence des valeurs extrêmes et d'améliorer la robustesse du modèle.

Transformation des colonnes `PotentialFraud` et `RenalDiseaseIndicator`

— `PotentialFraud` : Les valeurs `Yes` et `No` ont été transformées en 1 et 0 respectivement, pour les rendre exploitables par les modèles.

— `RenalDiseaseIndicator` : Cette colonne, contenant les valeurs `"0"` et `"Y"`, a également été transformée en 0 et 1.

Ces transformations ont permis de simplifier les données tout en augmentant leur pertinence pour le modèle.

3.3 Exploration des données

L'exploration des données est une étape essentielle dans un projet de machine learning, car elle permet de comprendre les caractéristiques des variables, leurs distributions et leurs relations. Cette analyse approfondie aide à identifier les anomalies, les valeurs manquantes et les tendances clés, ce qui est crucial pour guider la préparation des données et le choix des modèles.

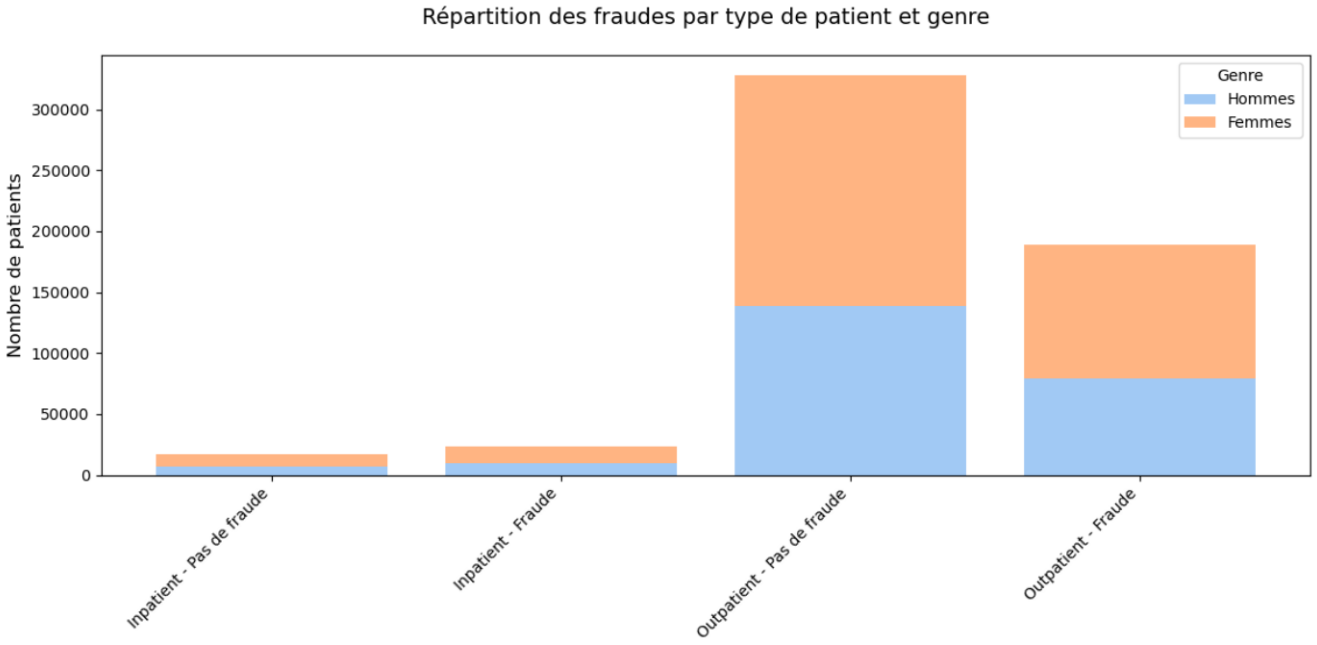


FIGURE 7

Le graphique ci-dessus illustre la répartition des fraudes potentielles en fonction du type de patient (*Inpatient* ou *Outpatient*) et du genre (*Hommes* ou *Femmes*). Il montre que la majorité des cas de fraude, ainsi que des non-fraudes, concernent des patients *Outpatient*, avec une proportion plus élevée de femmes dans chaque catégorie. Les patients *Inpatient* représentent une part significativement plus faible, indépendamment du statut de fraude.

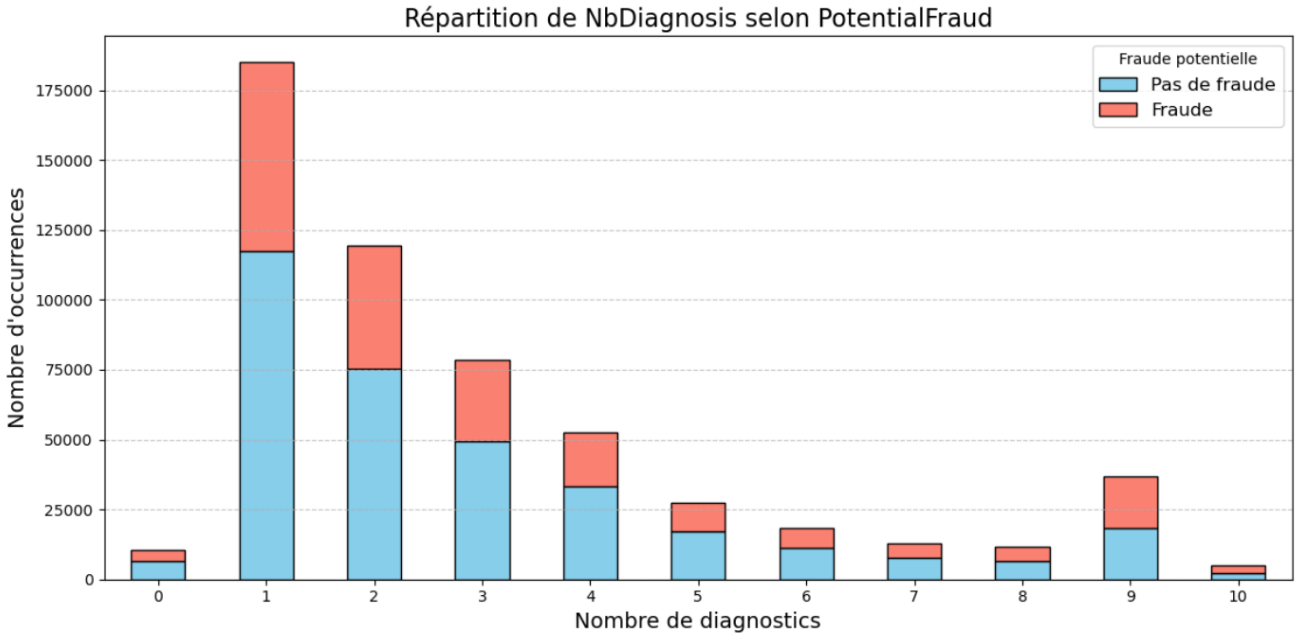


FIGURE 8

Ce graphique montre la répartition du nombre de diagnostics (**NbDiagnosis**) en fonction de la présence ou non d'une fraude potentielle (**PotentialFraud**). On observe que la majorité des réclamations avec un faible nombre de diagnostics (1 ou 2) sont associées à des cas sans fraude, tandis que les fraudes sont plus fréquentes proportionnellement pour les réclamations avec un nombre élevé de diagnostics. Cette distribution met en évidence une possible corrélation entre le nombre de diagnostics et le risque de fraude.

Les graphiques suivants présentent la répartition des valeurs pour six variables relatives aux montants payés et remboursés : **IPAnnualReimbursementAmt**, **IPAnnualDeductibleAmt**, **OPAnnualReimbursementAmt**, **OPAnnualDeductibleAmt**, **DeductibleAmtPaid** et **InscClaimAmtReimbursed**. Malgré la présence de valeurs extrêmes, ces variables ont été conservées dans leur intégralité, car elles représentent des informations essentielles sur les montants impliqués, cruciales pour l'apprentissage du modèle.

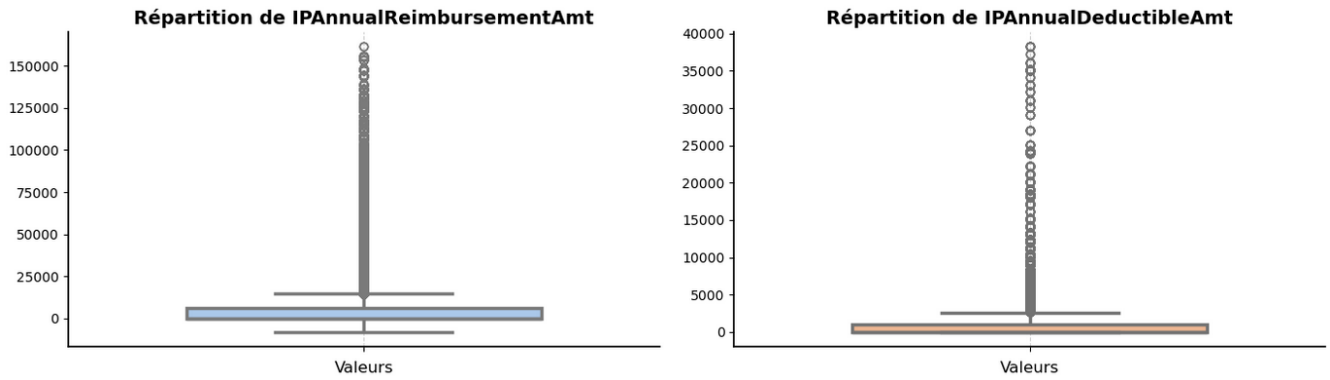


FIGURE 9

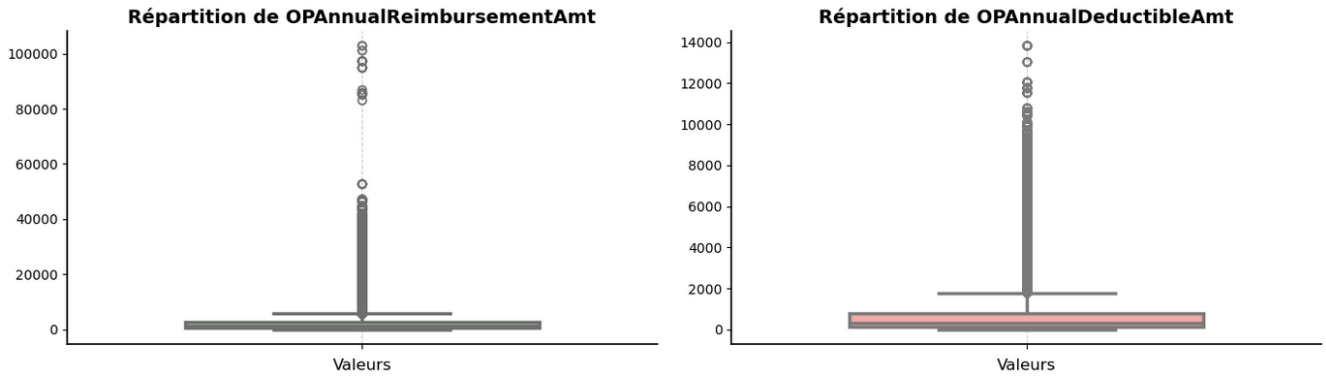


FIGURE 10

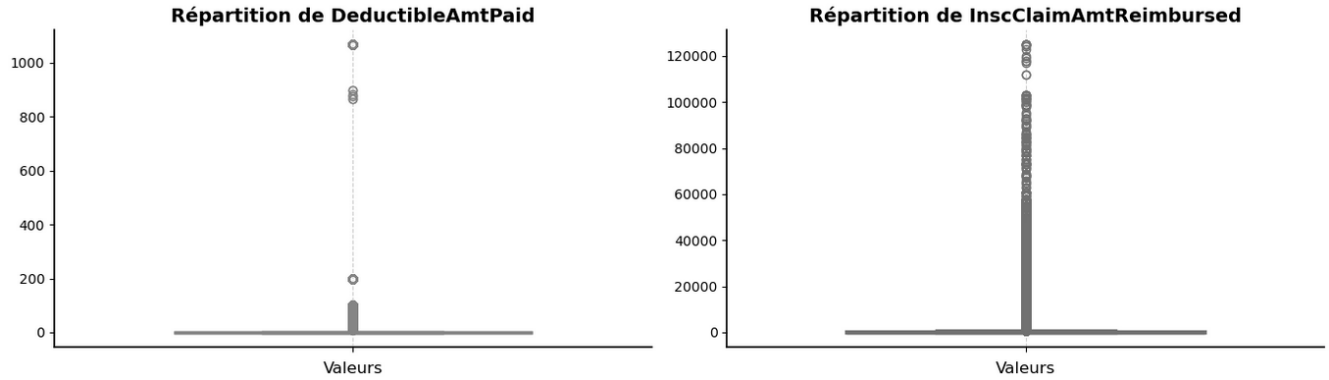


FIGURE 11

Enfin, le graphique qui suit illustre la répartition des niveaux de risque (**StateRisk**) à travers les différents États des États-Unis. Les couleurs représentent les niveaux de risque, allant de 1 (*faible risque*) à 4 (*risque critique*). Cette visualisation permet de comparer rapidement les États en fonction de leur niveau de risque, fournissant une vue d'ensemble des zones à faible ou à haut risque dans le contexte analysé. Ce type d'analyse est fortement utile pour identifier les régions nécessitant une attention particulière.

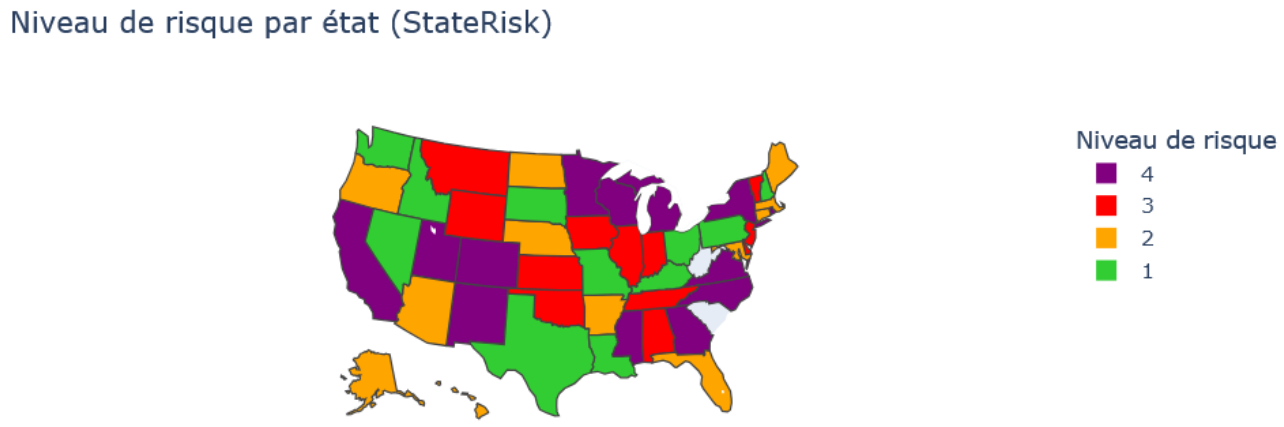


FIGURE 12

4 Modèles

4.1 Démarche

Pour mener à bien notre projet, nous avons suivi la démarche suivante :

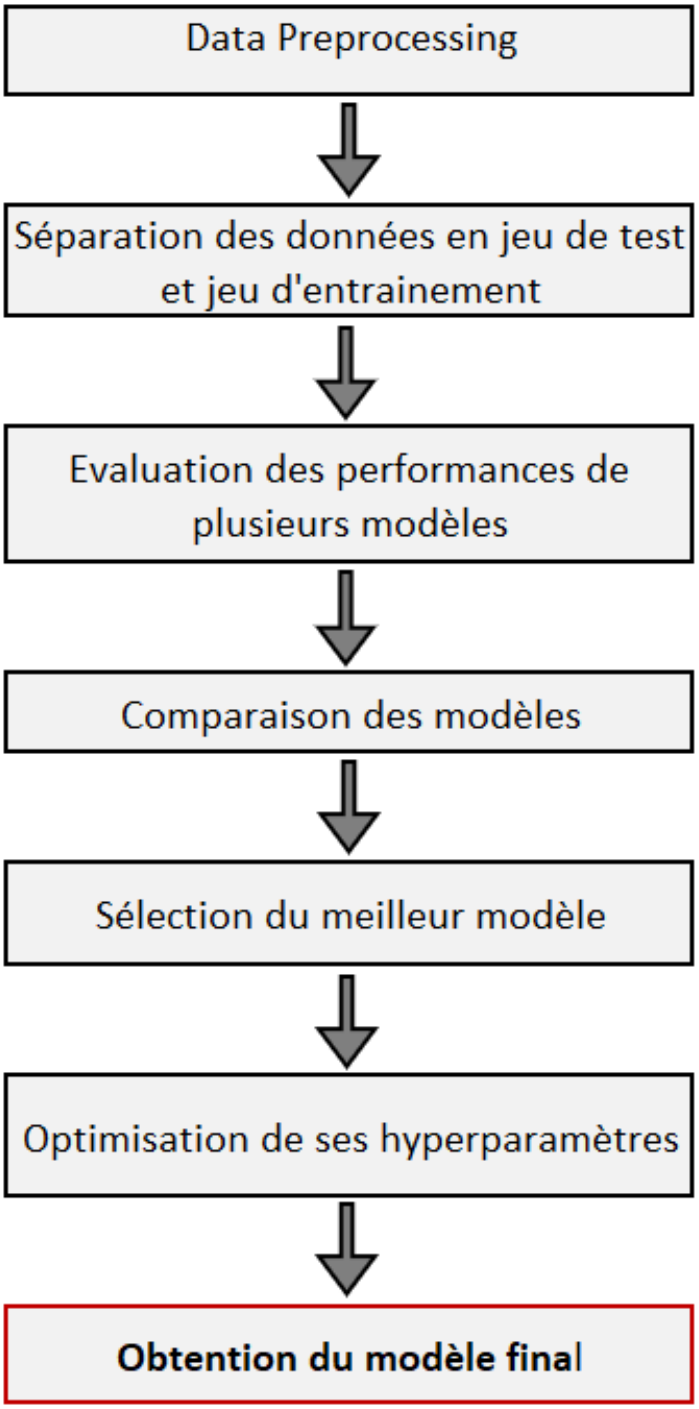


FIGURE 13 – Démarche adoptée

4.2 Séparation des données en ensembles d'entraînement et de test

Pour construire et évaluer le modèle de machine learning, les données ont été divisées en deux ensembles : un ensemble d'entraînement (*train set*) et un ensemble de test (*test set*). Cette séparation permet d'entraîner le modèle sur un sous-ensemble des données tout en évaluant ses performances sur des données indépendantes, garantissant ainsi une évaluation fiable.

4.3 Présentation de `scikit-learn`

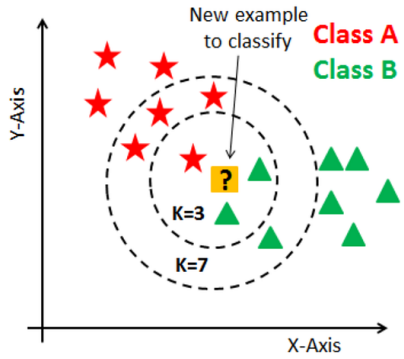
Pour ce projet, nous utilisons `scikit-learn`, une bibliothèque populaire de machine learning en Python. Elle offre une large gamme d'algorithmes supervisés et non supervisés, ainsi que des outils pour le prétraitement des données, l'évaluation des modèles et l'optimisation des hyperparamètres. Sa simplicité, sa cohérence et son intégration avec d'autres librairies comme `pandas` et `NumPy` en font un choix idéal pour notre projet.

4.4 Présentation des modèles

Dans le cadre de notre projet, nous avons comparé les performances de différents modèles de classification.

4.4.1 K-Nearest Neighbors (KNeighborClassifier)

Le modèle `K-Nearest Neighbors` (KNN) est un algorithme basé sur la proximité. Il classe un échantillon en fonction de la majorité des classes de ses voisins les plus proches, selon une distance métrique.



4.4.2 Linear Discriminant Analysis (LDA)

L'analyse discriminante linéaire est un modèle probabiliste qui vise à trouver une combinaison linéaire des caractéristiques permettant de séparer au mieux les classes cibles dans les données.

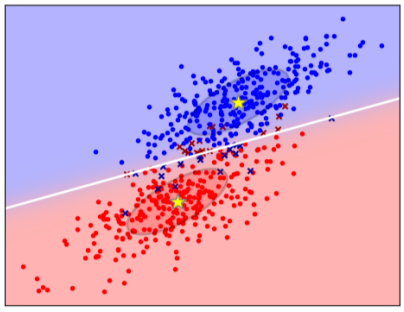


FIGURE 14 – LDA avec des variances identiques

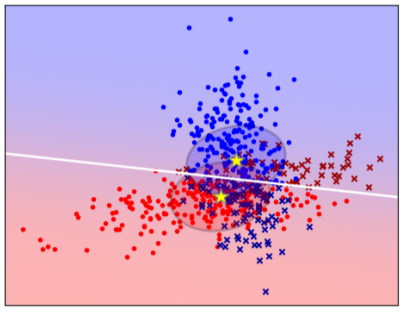


FIGURE 15 – LDA avec des variances distinctes

4.4.3 Régression Linéaire

Bien que principalement utilisé pour les problèmes de régression, le modèle de régression linéaire peut être adapté pour résoudre des problèmes de classification binaire via des approches spécifiques.

4.4.4 Decision Tree

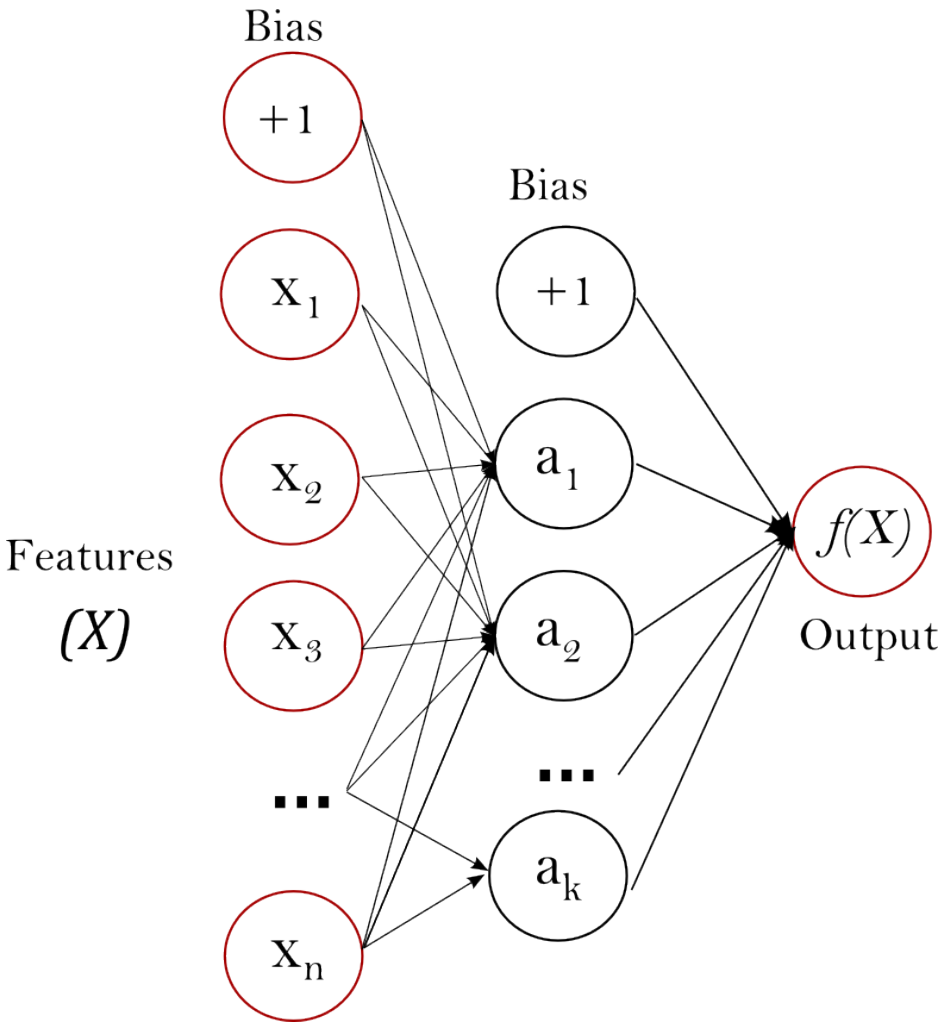
L'arbre de décision est un algorithme qui partitionne récursivement l'espace des caractéristiques pour prédire la classe cible en fonction de règles simples dérivées des données.

4.4.5 Random Forest

Le `Random Forest` est un modèle basé sur des ensembles qui combine plusieurs arbres de décision pour améliorer la précision et réduire la variance. Il s'agit d'une méthode robuste et performante.

4.4.6 Neural Network

Le réseau de neurones artificiels (ANN) est un modèle inspiré du fonctionnement biologique des neurones. Il est composé de couches interconnectées qui apprennent des représentations complexes des données.



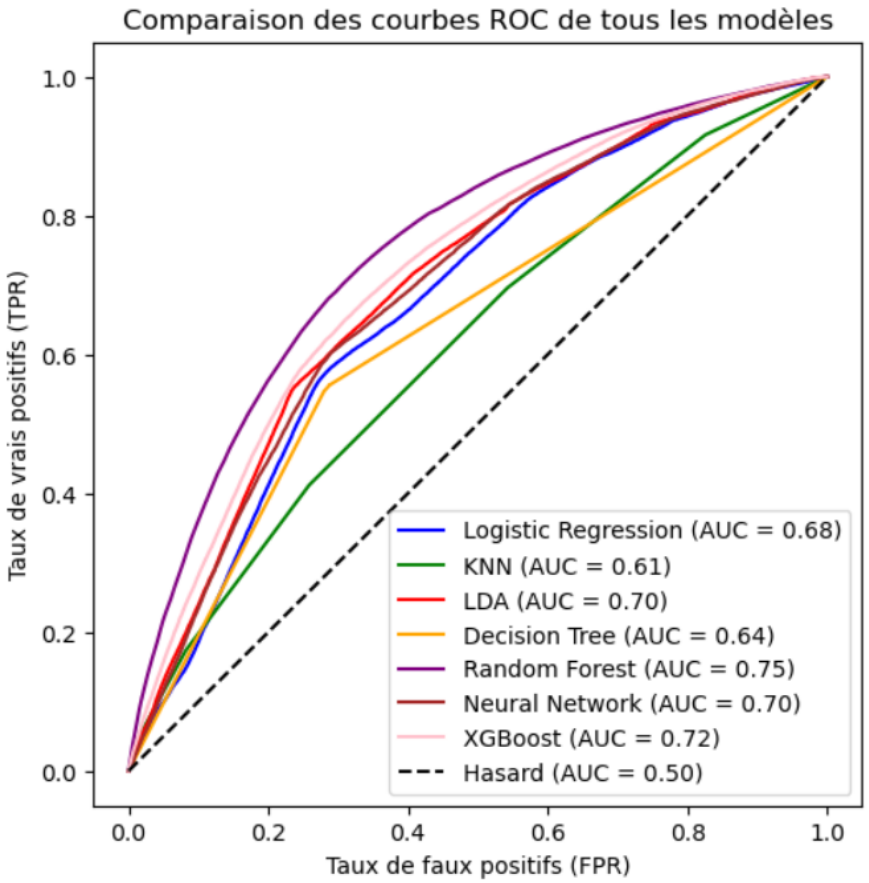
4.4.7 XGBoost

`XGBoost` (Extreme Gradient Boosting) est un algorithme d'ensemble basé sur les arbres de décision. Il utilise une méthode de boosting par gradient pour optimiser les prédictions de manière itérative.

4.5 Comparaison des modèles

Métriques utilisées pour évaluer les modèles
Afin de comparer les performances des modèles de machine learning, plusieurs métriques ont été utilisées pour évaluer leur capacité à prédire correctement les classes cibles. Tout d'abord, la **précision globale** (*Accuracy*) mesure le pourcentage de prédictions correctes parmi l'ensemble des observations, offrant une vision globale des performances. Ensuite, des métriques spécifiques comme la **précision** (*Precision*) et le **rappel** (*Recall*) ont été analysées pour évaluer respectivement la capacité à minimiser les faux positifs et à capturer les vrais positifs. Pour mieux équilibrer ces deux aspects, le **score F1**, qui combine précision et rappel, a été utilisé. Enfin, l'**aire sous la courbe ROC** (*AUC-ROC*) a permis de mesurer la capacité des modèles à discriminer entre les classes positives et négatives sur différents seuils, fournissant une évaluation robuste des performances globales des modèles.

Comparaison des courbes ROC des modèles
La figure ci-dessous présente une comparaison des courbes ROC pour les différents modèles testés dans ce projet, accompagnées des valeurs de l'AUC (Area Under the Curve). La courbe ROC mesure la performance d'un modèle en traçant le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR) pour différents seuils de classification.



— Le modèle **Random Forest** affiche la meilleure performance avec une AUC de 0.75, indiquant une excellente capacité à discriminer entre les classes positives et négatives.
— **XGBoost** suit avec une AUC de 0.72, démontrant également de bonnes performances.
— Les modèles **LDA** et **Neural Network** obtiennent des AUC similaires, autour de 0.70, représentant des performances satisfaisantes.
— En revanche, les modèles **KNN** (AUC = 0.61) et **Decision Tree** (AUC = 0.64) montrent des résultats moins robustes, avec une capacité de discrimination plus faible.
— La courbe *Hasard* (AUC = 0.50), représentée par une ligne en pointillés, illustre une performance équivalente à une classification aléatoire.

Analyse des performances des modèles

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

Le tableau ci-dessus présente les performances des différents modèles testés à l'aide de plusieurs métriques d'évaluation : **AUC**, **AUC PR**, **précision**, **rappel**, **F1-score**, et le **temps d'inférence moyen par échantillon**. Ces résultats permettent d'évaluer les modèles en termes de précision, de robustesse et d'efficacité. Voici les principales observations :

— **AUC et AUC PR :** Le modèle **Random Forest** se distingue avec la meilleure AUC (0.75) et AUC PR (0.64), montrant une excellente capacité à discriminer entre les classes. **XGBoost** (AUC = 0.72, AUC PR = 0.58) et **LDA** (AUC = 0.70, AUC PR = 0.55) suivent également avec des performances élevées. À l'opposé, le **KNeighborClassifier** a l'AUC la plus faible (0.61).
— **Précision et Rappel :** Le **Random Forest** reste le modèle le plus équilibré, avec une précision de 0.635 et un rappel de 0.562. Ces performances sont suivies de près par **XGBoost** et **LDA**. Les modèles **Régression Linéaire** et **KNeighborClassifier**, bien qu'efficaces dans certains contextes, affichent des performances moins robustes sur ces métriques.
— **F1-score :** Le score F1, qui équilibre précision et rappel, est le plus élevé pour le **Random Forest** (0.596), suivi par **Neural Network** (0.581) et **XGBoost** (0.570), ce qui confirme leur capacité à bien performer dans des scénarios équilibrés.
— **Temps d'inférence :** En termes de rapidité, **LDA** (0.22 µs) et **Régression Linéaire** (0.19 µs) sont les plus rapides, ce qui les rend idéaux pour des applications nécessitant une faible latence. En revanche, le **KNeighborClassifier** est de loin le plus lent (1038.31 µs), ce qui peut le rendre moins adapté dans des environnements où la vitesse est un facteur critique.

Conclusion : Le **Random Forest** apparaît comme le modèle le plus performant globalement, offrant un excellent équilibre entre AUC, F1-score et AUC PR, tout en ayant un temps d'inférence raisonnable. Cependant, pour des scénarios où la rapidité est prioritaire, des modèles comme **LDA** ou **Régression Linéaire** sont à privilégier.

On peut retrouver les graphiques relatifs au **Random Forest** ci-dessous :

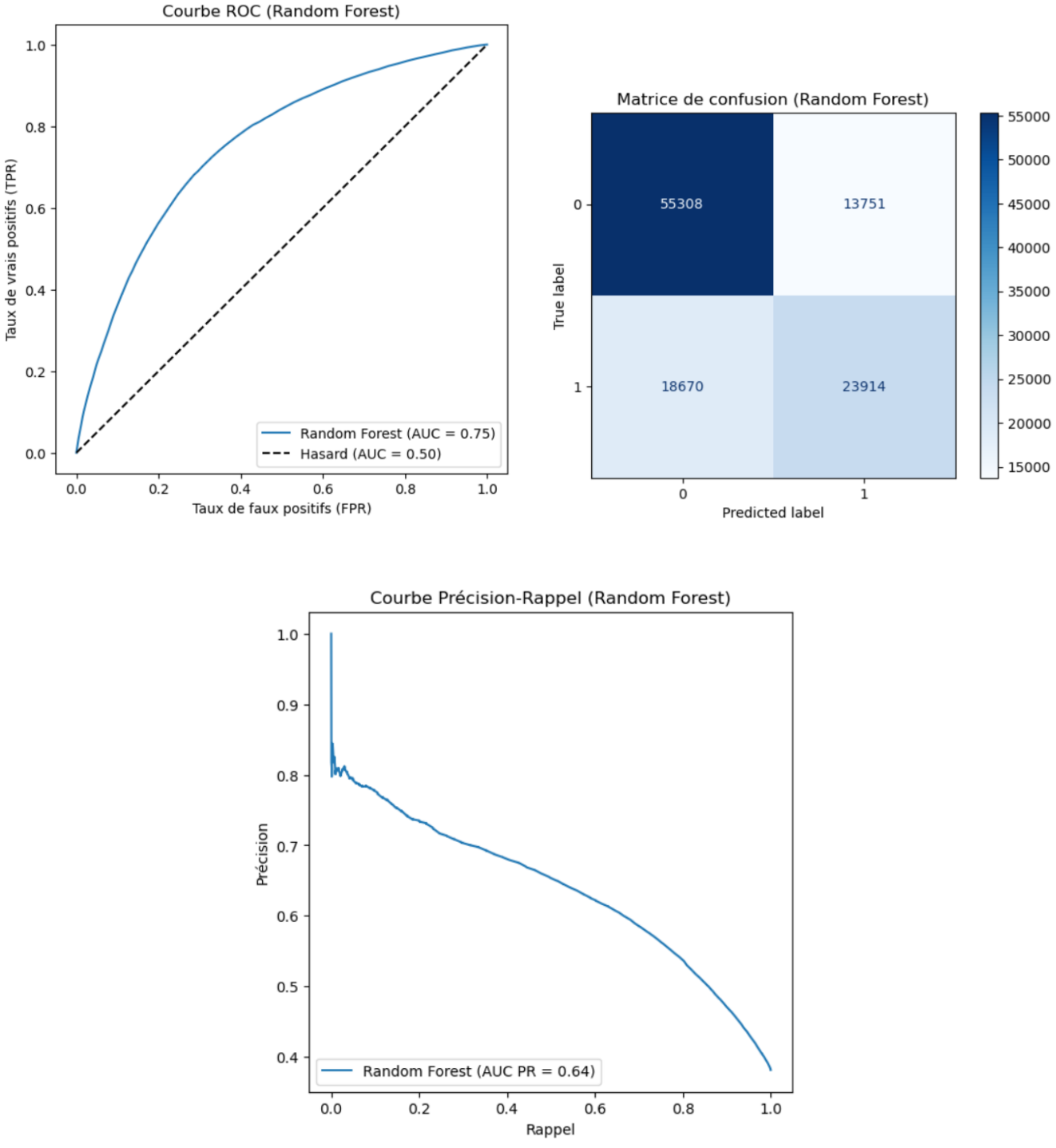


FIGURE 16 – Description globale pour la méthode Random Forest

5 Optimisation des hyperparamètres du modèle

L'optimisation des hyperparamètres constitue une étape cruciale dans le cadre du développement de modèles de machine learning. Pour le modèle que nous avons retenu, **Random Forest**, cette phase est particulièrement déterminante pour maximiser les performances prédictives tout en minimisant les risques de surapprentissage ou de sous-apprentissage.

5.1 Pourquoi optimiser les hyperparamètres ?

Contrairement aux paramètres appris directement par le modèle pendant l'entraînement, les hyperparamètres sont définis a priori et influencent la manière dont le modèle détermine les relations entre les données. Dans le cas de **Random Forest**, les hyperparamètres tels que :

- Le nombre d'arbres dans la forêt (`n_estimators`),
- La profondeur maximale des arbres (`max_depth`),
- Le nombre minimal d'échantillons requis pour diviser un nœud (`min_samples_split`),
- Le nombre minimal d'échantillons dans une feuille terminale (`min_samples_leaf`),
- Le nombre de variables à considérer pour chaque division (`max_features`),
- La pondération des classes pour gérer les déséquilibres dans les données (`class_weight`),

ont un impact direct sur la capacité du modèle à capturer les tendances dans les données tout en généralisant efficacement sur des données inconnues.

5.2 Approche adoptée pour l'optimisation

Pour garantir une performance optimale, nous avons utilisé des méthodes systématiques d'optimisation des hyperparamètres, telles que la recherche par grille (*Grid Search*) et la recherche aléatoire (*Random Search*). Ces approches permettent d'explorer un espace d'hyperparamètres défini et d'identifier les combinaisons qui minimisent une métrique de performance choisie, comme l'AUC, la précision, ou la F1-score.
De plus, afin de limiter les risques d'overfitting lors de l'optimisation, une validation croisée (*cross-validation*) a été intégrée. Cette technique consiste à diviser les données en plusieurs sous-ensembles pour évaluer le modèle sur des partitions indépendantes et ainsi garantir que les performances observées sont généralisables.

5.3 Résultats de l'optimisation

Dans cette étape, nous avons réalisé une recherche approfondie des hyperparamètres pour optimiser les performances du modèle **Random Forest** sur notre jeu de données. Pour cela, une recherche sur grille combinée à une validation croisée à 3 plis (*3-fold cross-validation*) a été effectuée. Au total, **432 combinaisons** d'hyperparamètres ont été explorées, représentant **1296 entraînements** (3×432).
Les meilleurs hyperparamètres trouvés sont les suivants :

- **class_weight** : balanced
- **max_depth** : None
- **max_features** : sqrt
- **min_samples_leaf** : 4
- **min_samples_split** : 10
- **n_estimators** : 200

Ces hyperparamètres permettent de maximiser la performance du modèle en équilibrant les classes dans les données, en limitant la profondeur des arbres, et en ajustant les critères de division et d'échantillonnage. Grâce à cette optimisation, le modèle est à la fois plus robuste et plus adapté à notre problème spécifique.

Analyse des courbes ROC

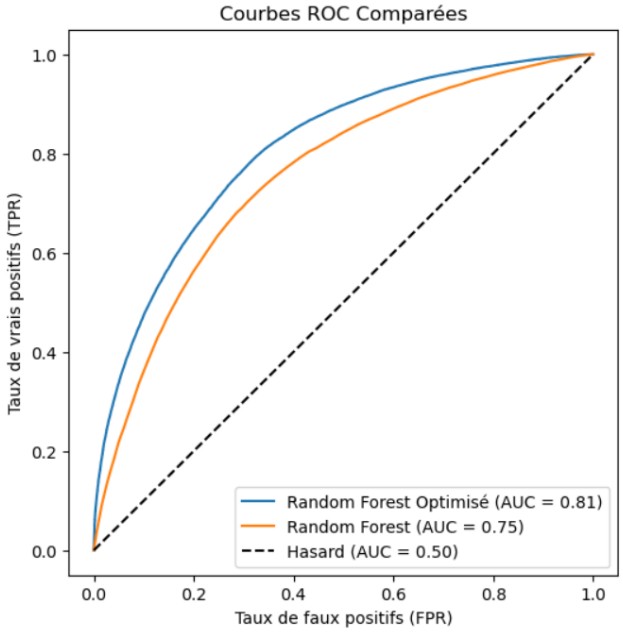
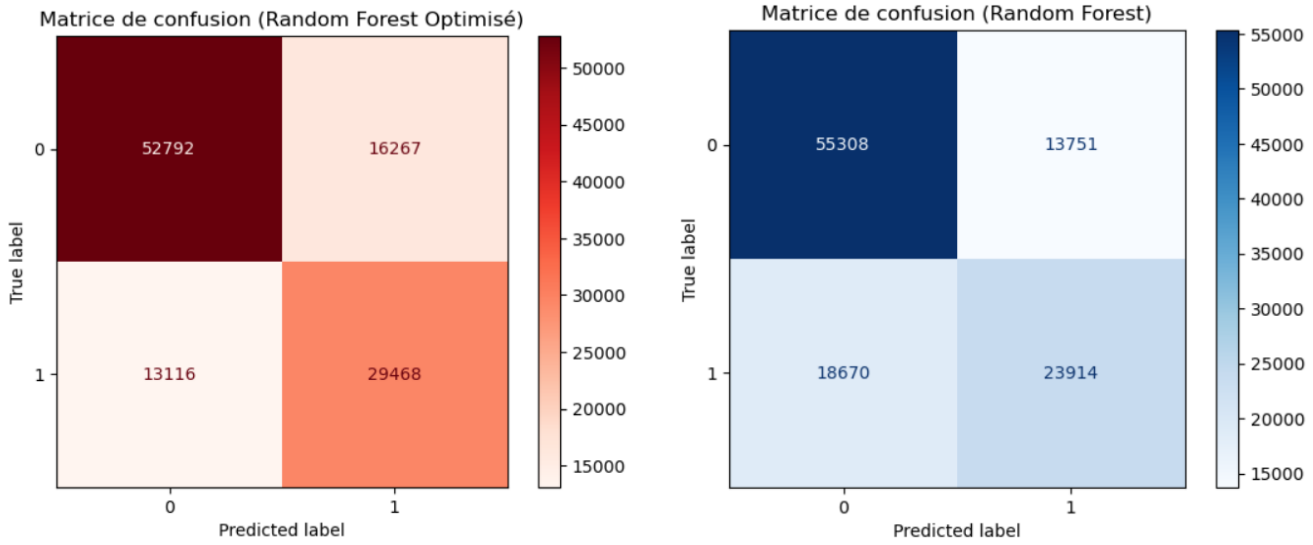


FIGURE 17 – Comparaison de la courbe ROC entre le modèle optimisé et Random Forest

On peut noter que l'AUC de notre modèle optimisé (0.81) est nettement supérieure à celui de base (0.75)

Matrices de confusion



(a) Matrice de confusion du modèle optimisé.

(b) Matrice de confusion de Random Forest.

FIGURE 18 – Comparaison des matrices de confusion entre le modèle optimisé et Random Forest.

Importance des variables

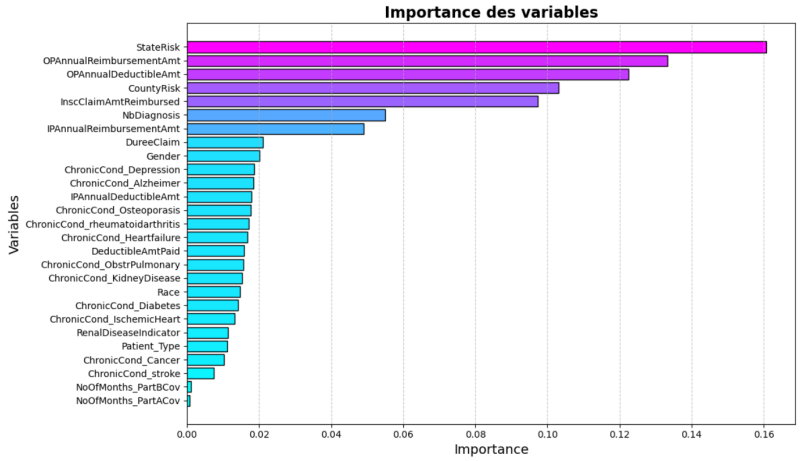
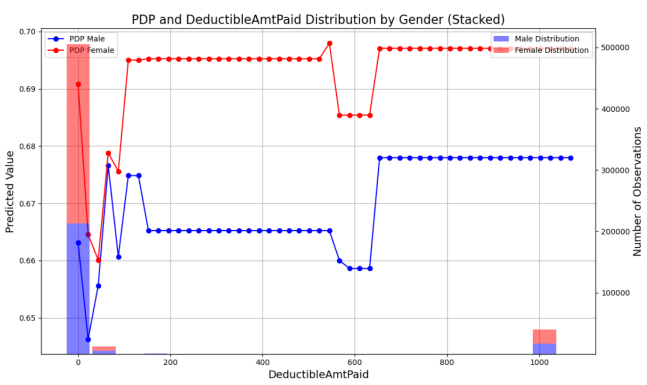
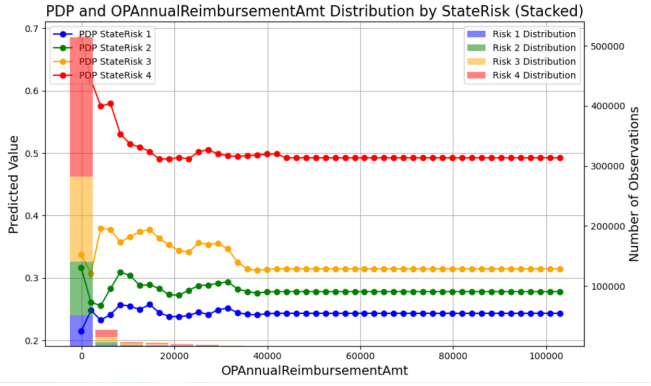


FIGURE 19 – Importance des variables dans le modèle optimisé.

Visualisations complémentaires





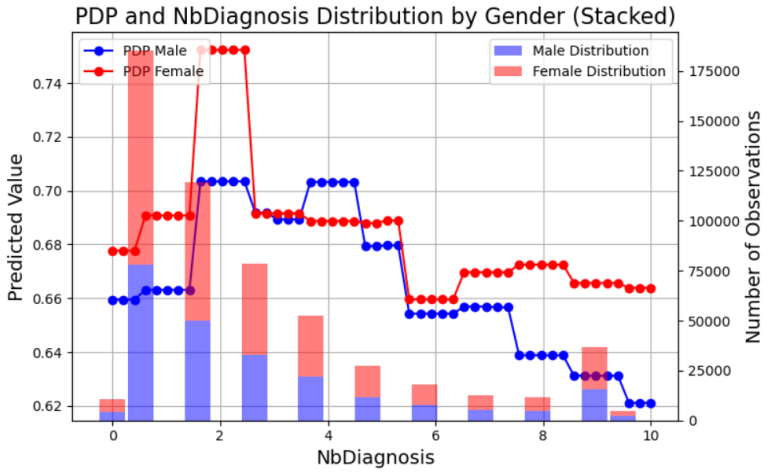


FIGURE 20 – Visualisations du modèle optimisé.

Comaparaison des performances par type de patients

Group	Precision	Recall	F1-score	ROC AUC
Patient_Type=1	0,595455	0,970576	0,738088	0,517851
Patient_Type=2	0,609	0,626598	0,617674	0,697433

FIGURE 21 – Comparaison des métriques PatientType=1 :InPatient vs PatientType=2 : OutPatient

Les métriques de performance ont été calculées séparément pour les groupes **Patient_Type=1 (InPatient)** et **Patient_Type=2 (OutPatient)**.
Le modèle présente une précision légèrement supérieure pour les OutPatients (**60.9%**) par rapport aux InPatients (**59.5%**). Cependant, le rappel est nettement meilleur pour les InPatients (**97.1%**) que pour les OutPatients (**62.7%**), indiquant que le modèle détecte plus efficacement les InPatients.
Le **F1-score**, une mesure combinée, est également supérieur pour les InPatients (**73.8%**) que pour les OutPatients (**61.8%**), reflétant une performance globale plus équilibrée pour ce groupe.
En revanche, l'aire sous la courbe ROC (**ROC AUC**) est plus élevée pour les OutPatients (**69.7%**) que pour les InPatients (**51.8%**), suggérant que le modèle est globalement meilleur pour distinguer les classes dans le groupe OutPatient.

Comparaison des métriques

Enfin, le tableau ci-dessous compare les résultats des 2 modèles pour les autres métriques.

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

FIGURE 22 – Comparaison des métriques

Comparons chaque paramètre :

- **Précision** : Le modèle optimisé présente une légère amélioration de la précision, passant de 0.635 à 0.644. Cela indique que notre modèle optimisé fait un peu moins d'erreurs parmi les prédictions positives.
- **Rappel** : Une amélioration significative du rappel est observée, passant de 0.562 à 0.692. Le modèle optimisé capture donc mieux les cas positifs réels.
- **F1-score** : Le F1-score, qui représente un équilibre entre la précision et le rappel, montre une amélioration notable, passant de 0.596 à 0.667. Cela démontre que le modèle optimisé est globalement plus performant.

Ainsi, nous pouvons conclure que le Random Forest Optimisé surpasse le modèle de base, et à fortiori, tous les modèles testés précédemment. Nous conservons donc ce modèle prédictif.

6 Conclusion

Dans ce projet, nous nous sommes intéressés à une problématique cruciale pour le système de santé américain : **la détection des fraudes parmi les prestataires de soins de santé**. Les fraudes organisées dans ce secteur entraînent des réclamations fictives qui alourdissent les dépenses de Medicare, affectant directement les compagnies d'assurance et les assurés par des primes élevées et une hausse des coûts des soins. L'objectif principal était de développer un modèle prédictif efficace pour identifier ces comportements frauduleux à partir des données des réclamations médicales.
Après une phase d'exploration approfondie des données et d'évaluation de plusieurs modèles de machine learning, nous avons retenu un **modèle Random Forest optimisé** pour résoudre notre problème de classification. Ce modèle, ajusté avec les paramètres suivants :

- `class_weight` : balanced
- `max_depth` : None
- `max_features` : sqrt
- `min_samples_leaf` : 4
- `min_samples_split` : 10
- `n_estimators` : 200

a obtenu des performances satisfaisantes pour la prédiction des fraudes. En conclusion, ce modèle pourrait être intégré dans un cadre opérationnel, tout en étant enrichi par de nouvelles données et ajusté en fonction des évolutions des schémas de fraude. Ce projet montre l'importance et la puissance des outils de machine learning pour relever les défis complexes du secteur de la santé.

7 Bibliographie

https://www.cms.gov/files/document/overviewfwacommonfraudtypesfactsheet072616pdf
https://www.nhcaa.org/tools-insights/about-health-care-fraud/the-challenge-of-health-care-fraud/
https://www.medicare.gov/publications/11306-F-Medicare-Medicaid.pdf
https://www.fbi.gov/investigate/white-collar-crime/health-care-fraud
https://www.ussc.gov/research/quick-facts/health-care-fraud
https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis?select=Train_Inpatientdata-1542865627584.csv

8 Annexes

8.1 Opérations sur les colonnes

	Train_In	Train_Out
Supprimée pour la prédiction	~~BeneID~~	~~BeneID~~
Supprimée pour la prédiction	~~ClaimID~~	~~ClaimID~~
Transformées	ClaimStartDt	ClaimStartDt
	ClaimEndDt	ClaimEndDt
Supprimée pour la prédiction	~~Provider~~	~~Provider~~
	InscClaimAmtReimbursed	InscClaimAmtReimbursed
	AttendingPhysician	AttendingPhysician
Supprimée à cause des NA	~~OperatingPhysician~~	~~OperatingPhysician~~
Supprimée à cause des NA	~~OtherPhysician~~	~~OtherPhysician~~
Supprimée pour la prédiction	~~AdmissionDt~~	
Supprimée à cause des NA	~~ClmAdmitDiagnosisCode~~	~~ClmAdmitDiagnosisCode~~
	DeductibleAmtPaid	DeductibleAmtPaid
Supprimée pour la prédiction	~~DischargeDt~~	
Supprimée à cause des NA	~~DiagnosisGroupCode~~	
Transformées	ClmDiagnosisCode_1	ClmDiagnosisCode_1
	ClmDiagnosisCode_2	ClmDiagnosisCode_2
	ClmDiagnosisCode_3	ClmDiagnosisCode_3
	ClmDiagnosisCode_4	ClmDiagnosisCode_4
	ClmDiagnosisCode_5	ClmDiagnosisCode_5
	ClmDiagnosisCode_6	ClmDiagnosisCode_6
	ClmDiagnosisCode_7	ClmDiagnosisCode_7
	ClmDiagnosisCode_8	ClmDiagnosisCode_8
	ClmDiagnosisCode_9	ClmDiagnosisCode_9
	ClmDiagnosisCode_10	ClmDiagnosisCode_10
Supprimée à cause des NA	~~ClmProcedureCode_1~~	~~ClmProcedureCode_1~~
Supprimée à cause des NA	~~ClmProcedureCode_2~~	~~ClmProcedureCode_2~~
Supprimée à cause des NA	~~ClmProcedureCode_3~~	~~ClmProcedureCode_3~~
Supprimée à cause des NA	~~ClmProcedureCode_4~~	~~ClmProcedureCode_4~~
Supprimée à cause des NA	~~ClmProcedureCode_5~~	~~ClmProcedureCode_5~~
Supprimée à cause des NA	~~ClmProcedureCode_6~~	~~ClmProcedureCode_6~~

FIGURE 23 – Colonnes sur les hospitalisations

	Train_Benef
Supprimée pour la prédiction	~~BeneID~~
Supprimée pour la prédiction	~~DOB~~
Supprimée à cause des NA	~~DOD~~
	Gender
	Race
Transformée	RenalDiseaseIndicator
Transformée	State
Transformée	County
	NoOfMonths_PartACov
	NoOfMonths_PartBCov
	ChronicCond_Alzheimer
	ChronicCond_Heartfailure
	ChronicCond_KidneyDisease
	ChronicCond_Cancer
	ChronicCond_ObstrPulmonary
	ChronicCond_Depression
	ChronicCond_Diabetes
	ChronicCond_IschemicHeart
	ChronicCond_Osteoporasis
	ChronicCond_rheumatoidarthritis
	ChronicCond_stroke
	IPAnnualReimbursementAmt
	IPAnnualDeductibleAmt
	OPAnnualReimbursementAmt
	OPAnnualDeductibleAmt

	Train_Providers
Supprimée pour la prédiction	~~Provider~~
Transformée	PotentialFraud

FIGURE 24 – Autres Colonnes