UNIVERSITÉ
CÔTE D'AZUR

2023-2024
Master 1 Ingénierie mathématique

Rapport du projet analyse de données

Résumé :

Ce document examine de près, les données des personnes assurées fournies par une société d'assurance. Cette étude vise à analyser, et à construire un modèle de prédiction des déclarations frauduleuses des clients assurés par la société. Dans le cadre de cette investigation, nous explorons et visualisons les données afin de les traiter et par une méthode efficace d'évaluation des classifieurs, nous ferons un choix du classifieur le plus performant afin de prédire au mieux avec un minimum d'erreur les fraudes.

Rapporteurs :

YANG Quiyan

KPATCHA Essowaza Celestin

Introduction :

Les principaux objectifs des compagnies d'assurances sont de générer des profits, de gérer les risques, de conserver et d'attirer de nouveaux clients… Ainsi pour atteindre ces objectifs, il est important de détecter toutes actions frauduleuses des assurés. C'est donc pour cette raison qu'il est nécessaire de mettre en place un modèle de prédiction performant capable de non seulement détecter tout incident frauduleux mais aussi de minimiser le risque ceci afin de maximiser le profit tout en offrant une satisfaction à la clientèle.
De fait, pour mener à bien notre travail, nous allons d'abord partir d'une grande partie d'un ensemble de données passées (750 assurés soit 68,18% de notre matrice de donnée) que nous devons analyser pour poser de différents modèles de prédiction. Ensuite on testera ces modèles sur les 31,82% restant de la matrice de donnée pour vérifier l'efficacité des modèles et pour finir on appliquera le modèle le plus efficace à la nouvelle matrice de donnée ceci afin de détecter toutes fraudes de la part des assurés.

Nous utiliserons dans un premier temps le logiciel Rstudio pour élaborer notre travail.

I. Problématique et objectif de l'étude :

1. Contexte et problématique :

Dans le contexte dynamique du secteur de l'assurance, la détection de fraudes émerge comme un enjeu stratégique majeur. Les compagnies d'assurance sont confrontées à une diversité croissante de comportements frauduleux de la part des assurés, allant des déclarations trompeuses aux activités malveillantes plus sophistiquées. Cette complexité pose des défis significatifs, non seulement en termes financiers, mais aussi en ce qui concerne la préservation de la confiance des clients. Les fraudes ont le potentiel de compromettre la stabilité économique des compagnies et de générer des répercussions négatives sur l'intégrité de tout le secteur. Face à ces défis, la nécessité d'adopter des approches de détection de fraudes avancées et adaptatives devient impérative, motivant ainsi cette étude à explorer des solutions innovantes pour renforcer les mécanismes de sécurité dans le domaine de l'assurance.

2. Objectif de l'Étude :

L'objectif principal de cette étude est d'évaluer et de mettre en œuvre divers classifieurs dans le but spécifique de détecter les comportements frauduleux chez les assurés. À travers une analyse approfondie des données clients fournies par la compagnie d'assurance, cette recherche vise à définir les caractéristiques de fraude, à identifier les classifieurs les plus pertinents, et à évaluer leur efficacité respective. Les questions clés à résoudre incluent la définition de critères de fraude, la sélection judicieuse des variables influentes, et l'exploration des algorithmes de classification pour optimiser la précision de la détection. En alignant ces objectifs, cette étude aspire à fournir des recommandations pratiques pour renforcer les capacités de détection de fraudes dans le secteur de l'assurance, contribuant ainsi à une gestion plus proactive des risques et à la préservation de l'intégrité financière des compagnies

II. Exploration de données

1. Identification des variables
   - Variable cible :

FRAUDULENT est la variable de la classe prédictive. Elle est égale à Yes si le client fraude et No sinon.

- Variables prédictives :

Age, Gender, Incident_cause, Date_to_incident, Claim_area, Police_report, Claim_type, Claim_amount, Total_policy_claims sont les caractéristiques à tester afin d'affecter une valeur Yes ou No à Fraudulent.

- Variable à ignorer :

Claim_ID et Customer_ID sont respectivement le numéro unique d'identication de la déclaration de l'incident et le numéro d'identification de l'assuré. Ces deux variables ne renvoient donc aucune information commune permettant de prédire une valeur Yes ou No à la variable cible

```
apprenti <- subset(apprenti, select = -claim_id)

apprenti <- subset(apprenti, select = -customer_id)
```

2. **Définition des ensembles :**

Ensemble d'apprentissage : 750

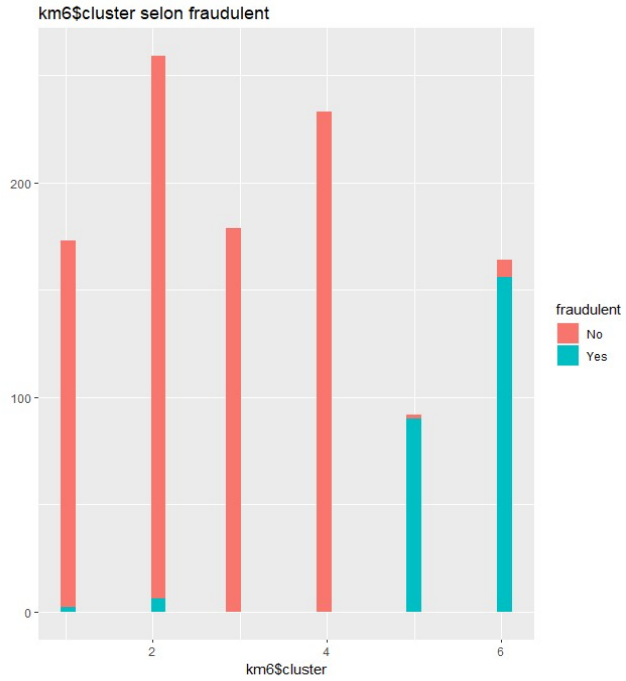
Ensemble Test : 350

III. Méthodologie :

A. Clustering des données :

❖ Clustering par partitionnement :

Approche algorithmique : Algorithme des K-means
Le code que nous avons généré évalue la performance du clustering pour différentes valeurs de k (de 4 à 7) en examinant les tables de contingence pour chaque k. Nous avons ainsi visualisé et analysé les différents graphes obtenu afin de connaitre la performance pour chaque valeur de k, ceci facilitant le choix du nombre optimal de clusters. On constate d'après les résultats obtenus qu'il y a un déséquilibre entre les classes fraudulent Yes et fraudulent No montrent ainsi que la valeur optimale pour obtenir des clusters pertinents pour la détection de fraudes est k=6. En effet dans ce cas nous avons moins de classe multiple et plus de classe unique dans l'ensemble des graphes que nous avons réalisés.



❖ **Clustering hiérarchique :**

De même comme dans la méthode de clustering précédente, nous avons trouvé ainsi que la valeur optimale pour obtenir des clusters pertinents pour la détection de fraudes est : k=8 pour le clustering hiérarchique par agglomération et k=3 pour le clustering hiérarchique par division.

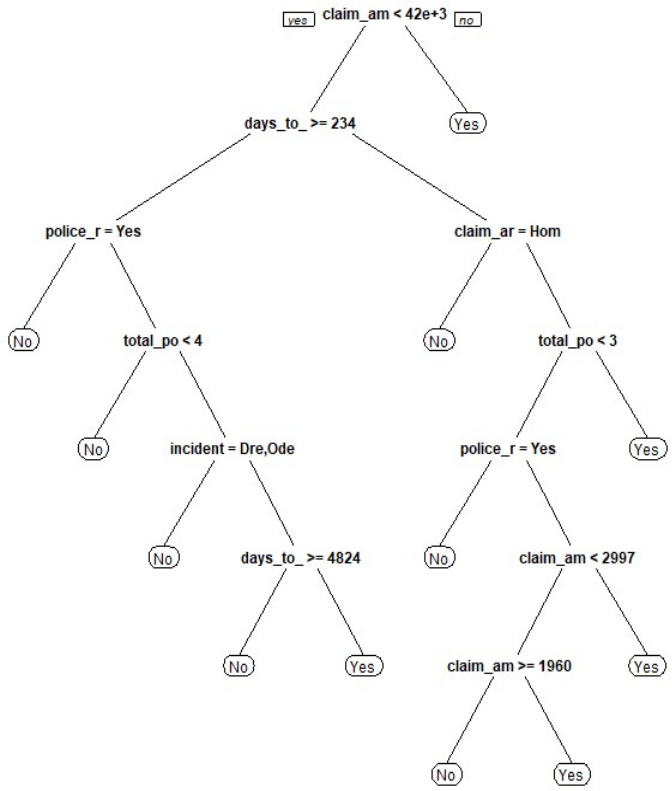
❖ **Clustering basé sur la densité :**

Nous avons dans notre code utilisé trois fonctions à savoir : dbscan et hdbscan avec une distance de voisinage (eps) de 0.125 et une taille minimale de voisinage (minPts) de 3. Et comme précédemment nous avons utilisé la boucle for pour un k allant de 3 à 8 ce qui nous a permis de conclure que k=7 est le meilleur pour la fonction dbscan et k=4 le meilleur pour la fonction hdbscan.

B. <u>Classification supervisée</u> :

Cette section détaille la méthodologie utilisée, incluant le chargement des données, la préparation des ensembles d'apprentissage et de test, ainsi que l'application de trois types d'arbres de décision (rpart, C5.0, tree) pour la classification des comportements frauduleux. Ces méthodes prédictives nous permettrons par une approche par arbres de décision de partir d'un ensemble d'apprentissage à un ensemble de prédiction en passant par un ensemble de test.

1. **Le classifieur Rpart() :**
   - **Graphe :**

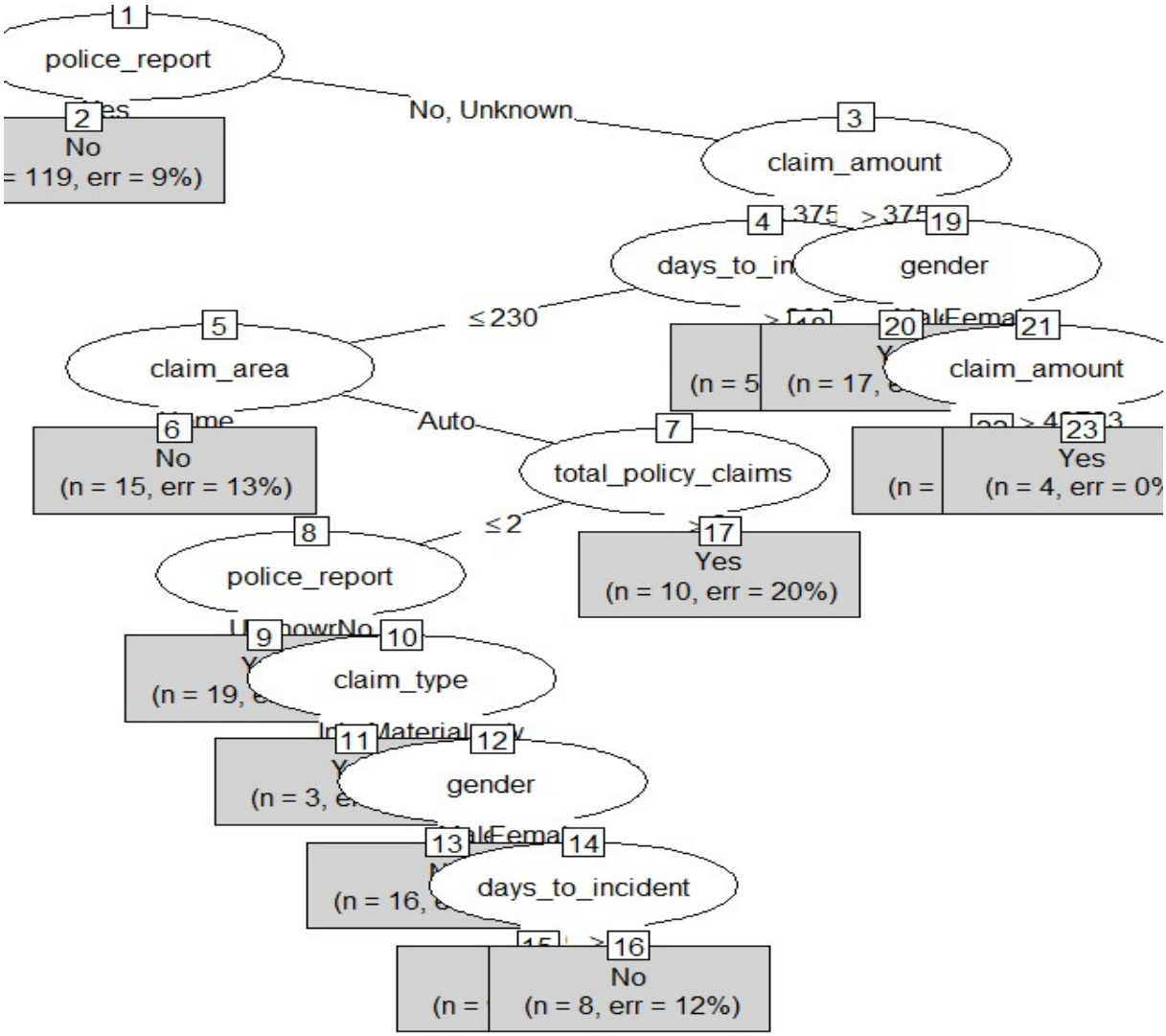


   - **Résultats et évaluation du classifieur :**

L'application de la prédiction, avec le classifieur Rpart(), nous donne un taux de succès égale de 76,28 %. On construit la matrice de confusion qui permet de quantifier numériquement l'importance des différents types de succès et erreurs et détaille le nombre de prédictions correctes et incorrectes pour chaque classe prédite et chaque réelle. Ainsi, on observe qu'en utilisant les mesures d'évaluations des succès et échecs, on a :

- Mesure du rappel (sensibilité) = 30.48%
- Mesure de spécificité = 90.29%
- Mesure de précision = 49.01%
- Mesure du taux de vrais négatifs =80.83%

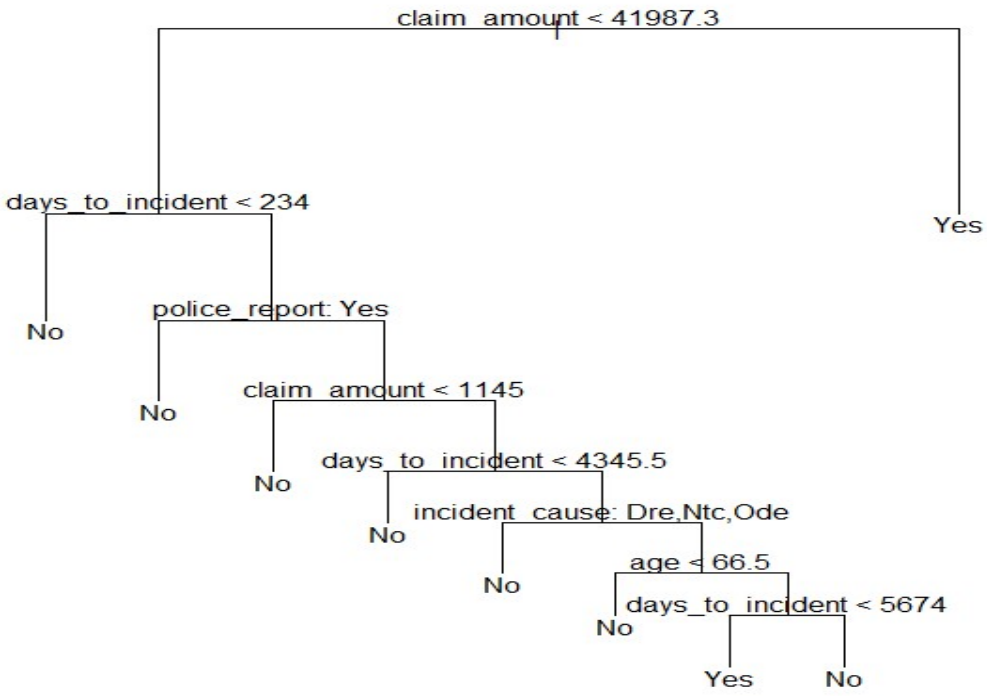
2. **<u>Le classifieur C5.0 ()</u> :**
   - ❖ **<u>Graphe</u> :**



L'application de la prédiction, avec le classifieur C5.0(), nous donne un taux de succès égale de 77.14 %. On construit la matrice de confusion qui permet de quantifier numériquement l'importance des différents types de succès et erreurs et détaille le nombre de prédictions correctes et incorrectes pour chaque classe prédite et chaque réelle. Ainsi, on observe qu'en utilisant les mesures d'évaluations des succès et échecs, on a :

- Mesure du rappel (sensibilité) = 24.39%
- Mesure de spécificité = 93.28%
- Mesure de précision = 52.63%
- Mesure du taux de vrais négatifs =80.12%

3. **Le classifieur tree :**
   - **Graphe :**

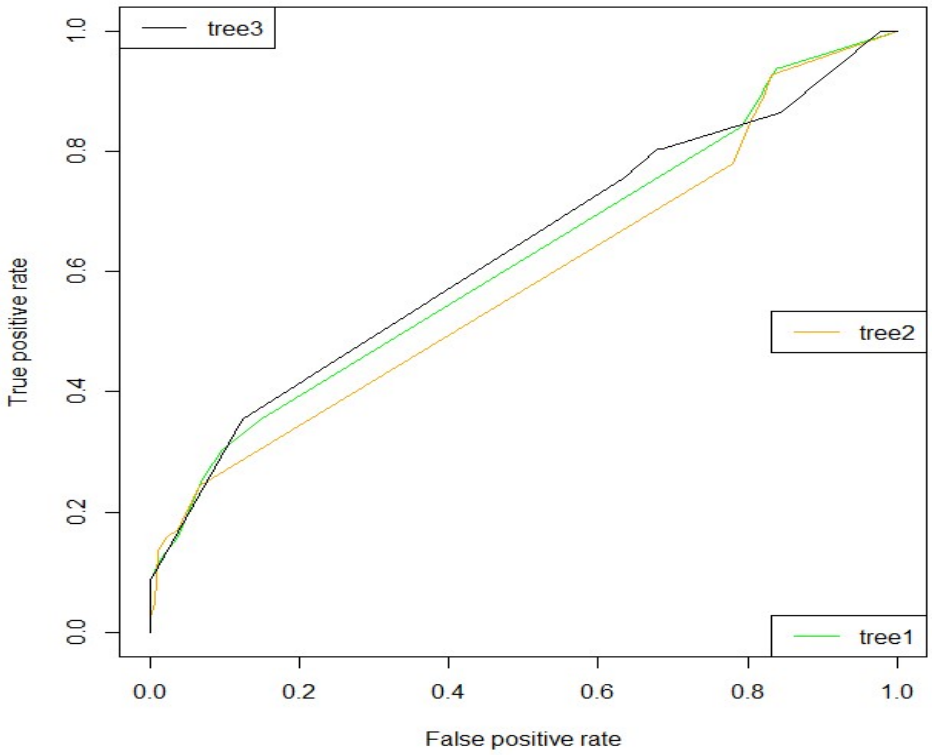


L'application de la prédiction, avec le classifieur tree(), nous donne un taux de succès égale de 78.57 %. On construit la matrice de confusion qui permet de quantifier numériquement l'importance des différents types de succès et erreurs et détaille le nombre de prédictions correctes et incorrectes pour chaque classe prédite et chaque réelle. Ainsi, on observe qu'en utilisant les mesures d'évaluations des succès et échecs, on a :

- Mesure du rappel (sensibilité) = 8.53%
- Mesure de spécificité = 100%
- Mesure de précision = 100%
- Mesure du taux de vrais négatifs =78.13%

C. Courbes ROC et Calcul des indices AUC :

1. **Courbes ROC :**



2. Calcul des indices AUC (Area Under the Curve) :

En calculant les indicateurs AUC(Area Under the Curve), on trouve :

- ❖ pour l'arbre Rpart(), l'indice AUC vaut 61.82%
- ❖ pour l'arbre C5.0(), l'indice AUC vaut 58.22%
- ❖ pour l'arbre Tree(), l'indice AUC vaut 63.02%

IV. Analyse et comparaison des différents classifieurs :

Dans cette partie nous utiliserons l'ensembles des résultats obtenus avec la compilation de notre code.

- **Taux de succès :** En comparant les trois taux de succès obtenus précédemment avec les trois arbres de décisions (rpart, C5.0 et tree), nous observons que le taux de succès du classifieur tree() qui est égale à 78.57 % est le plus élevé donc nous pouvons conclure que ce classifieur est le plus performant.
- **Précision des prédictions :**

La moyenne des probabilités des prédiction pour la classe fraudulent = Yes
La moyenne des probabilités pour Rpart() est 71.69% et sa médiane 68.18%
La moyenne des probabilités pour C5.0() est 68.03% et sa médiane 67.62%
La moyenne des probabilités pour tree() est 100% et sa médiane 100%
Conclusion : En comparant ces trois résultats nous concluons que le classifieur le plus performant est le classifieur tree().

- **Matrice de confusion :**

D’après les résultats des matrice de confusion mis plus haut nous pouvons dire que :

- le classifieur le plus performant parmi les trois arbres de décision selon le critère d'évaluation du Rappel est Rpart()
- le classifieur le plus performant parmi les trois arbres de décision selon le critère d'évaluation de la Spécificité est tree()
- le classifieur le plus performant parmi les trois arbres de décision selon le critère d'évaluation de la Précision est tree()
- le classifieur le plus performant parmi les trois arbres de décision selon le critère d'évaluation du Taux de Vrais Négatifs est Rpart()

- **Comparaison des courbes ROC :** la courbe en noire (courbe tree) est très souvent au-dessus des deux autres courbes donc le classifieur tree() est le plus performant des trois.
- **indices AUC :** Avec le calcul des indices AUC en utilisant les courbes ROC, on remarque que l’indice AUC correspondant à la courbe ROC en noire est plus grand donc le classifieur tree() est le plus performant des trois.

V. Choix du classifieur le plus performant :

Sur la base de nos analyses comparatives approfondies, le classifieur tree(), se distingue comme la méthode la plus robuste pour la détection de fraudes dans cet ensemble de données d'assurance. Il offre un équilibre optimal entre précision, rappel et spécificité. Par conséquent, nous recommandons l'adoption de tree() comme principal outil de détection de fraudes pour garantir des résultats fiables et efficaces.

VI. Application du classifieur aux données à prédire :

Nous avons dans cette partie appliquée le classifieur sélectionné (tree) aux données à prédire et nous obtenons une répartition des classe de fraudulent avec 5 Yes et 195 No.

De plus nous avons calculé la probabilité minimale, maximale et moyenne associée à la classe de fraudulent Yes et nous avons trouvé respectivement 0 ; 1 ; 0.2209.

De même pour la classe de fraudulent No nous avons calculé la probabilité minimale, maximale et moyenne associée, qui sont respectivement égale à 0 ; 1 ; 0.7790

Conclusion :

En conclusion, cette étude approfondie sur la détection de fraudes dans le secteur de l'assurance souligne le rôle crucial du classifieur tree() dans l'identification fiable des comportements frauduleux des assurés. Les résultats obtenus, appuyés par des indicateurs de performance solides, indiquent que ce classifieur offre une solution proactive et efficace pour renforcer la sécurité financière des compagnies d'assurance. Cependant, la dynamique en constante évolution des fraudes nécessite une approche agile, incitant à la recherche future sur

des techniques plus avancées. En adoptant judicieusement le classifieur tree(), la compagnie d'assurance peut non seulement optimiser leur lutte contre la fraude, mais aussi instaurer un environnement plus sûr et plus fiable, renforçant ainsi la confiance de leurs assurés.