PROJET : ANALYSE

**I/Introduction :**

Dans ce projet, nous allons essayer de prédire les fraudes faites par ses usagers au sein d’une assurance. Grâce à une base de données des différentes déclarations qu’ils ont eu, nous allons pouvoir en tirer des informations pour essayer de prédire si les futures déclarations sont des fraudes ou non. Pour cela nous allons nous aider de Rstudios et de ses librairies telles que rpart, C5.0, Tree et bien d’autres. Dans un premier temp nous allons traiter la base de données pour en garder que les informations qui nous sont utiles puis une fois ce travail effectué, nous allons prédire si les futures déclarations sont des fraudes ou non.

**II/Analyse des données :**

Tout d’abord, j’ai augmenté le nombre de données pour équilibrer le nombre de fraude =oui et le nombre de fraude = non et de ce fait pouvoir voir plus visiblement des corrélations par la suite. Ensuite j’ai coupé en deux ma base de données entre les données qui m’aideront à prédire et celles où je testerais mes prédictions. A l’aide de la librairie ggplot2 j’ai tracé toutes les combinaisons des variables entre elles en ajoutant un marqueur pour reconnaitre les fraudes ou non. Cela m’a permis de voir si des corrélations entre les variables par rapport aux fraudes étaient visibles et de ce fait, si elles étaient utiles pour la prédiction. Dans la globalité, on en retient que deux variables n’ont presque pas d’utilité pour la prédiction, il s’agit du genre ainsi que claim area. De plus, grâce au test X-2, j’ai pu constater que la variable incident\_cause n’a pas de corrélations non plus avec fraudulent et mec onfirmer que le genre et claim area sont inutiles.

Une image contenant capture d’écran, diagramme, Tracé, Logiciel de graphisme

Description générée automatiquement

Par exemple, on peut voir que 2/3 des incidents traités le premier jour sont des fraudes.

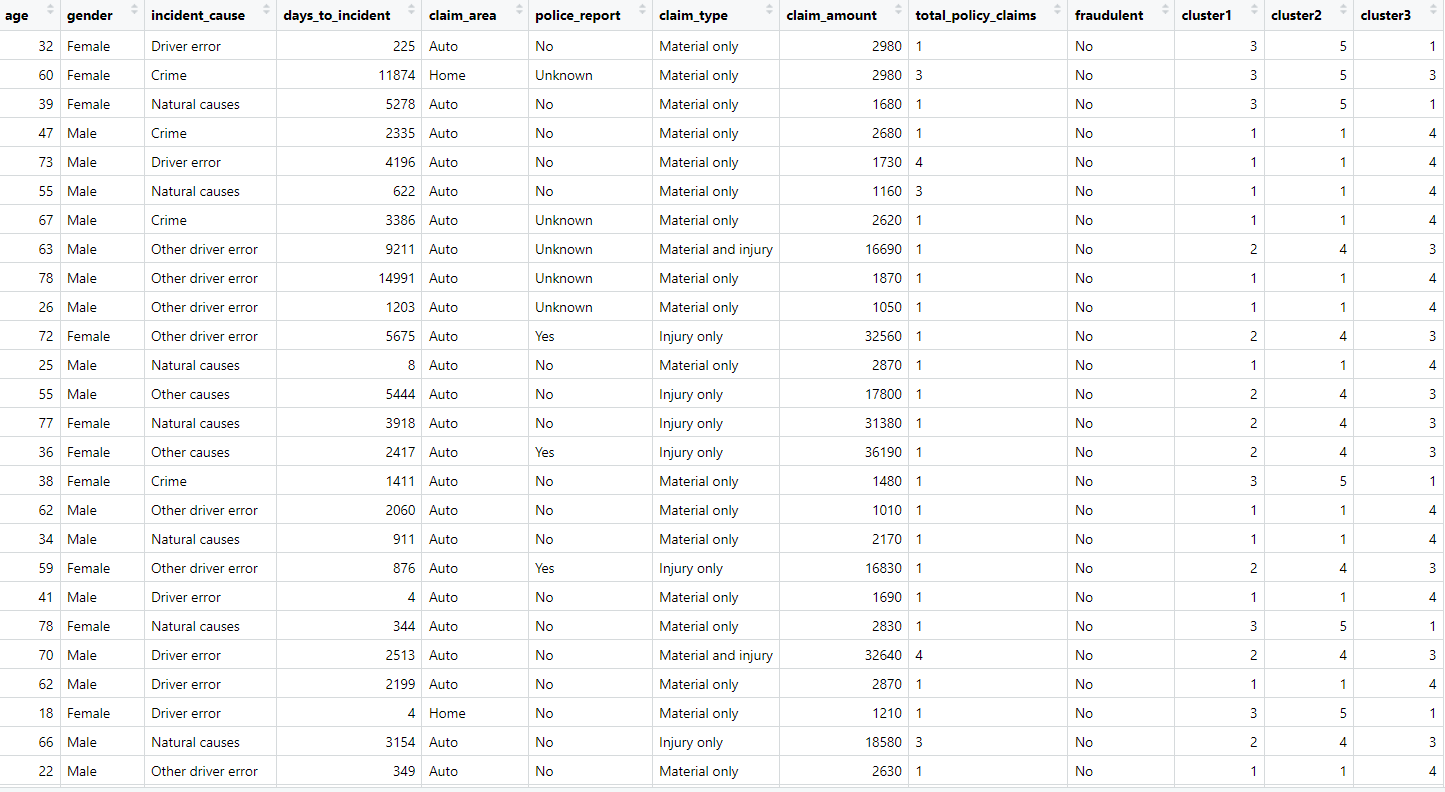
Une image contenant capture d’écran, Rectangle, carré

Description générée automatiquementEt on peut voir que dans cet exemple-là, la répartition de fraudes est la même selon le genre.

**III/Cluster :**

J’ai par la suite effectué des clusters avec les fonctions Kmeans, Agnes et diana. Mais cela n’a pas abouti à des clusters satisfaisants, seule la fonction Kmeans donne des résultats intéressant mais de façon aléatoire. Pour minimiser l’aléa j’ai décidé d’effectuer plusieurs fois la fonction kmeans et d’ajouter les clusters crées dans une nouvelle base de données.

Une image contenant capture d’écran, texte, Tracé, diagramme

Description générée automatiquement

**III/ Prédiction :**

Pour effectuer les prédictions, je me suis servi des trois arbres de décisions C5.0, Tree et rpart. Lorsque je me sers de ma base de données avec les clusters de kmeans ajouté, la prediction est de 1, mais cette méthode étant aléatoire, la fiabilité n’est pas maximale. Et ma méthode pour réduire l’aléatoire n’est pas fonctionnelle, car lorsque je fais une prédiction, la corrélation entre les clusters généré aléatoirement qui ont partitionné parfaitement les fraudes ou non est tellement forte que seule celle-ci est prise en compte pour la prédiction.

En revanche, sans passer par les clusters je trouve le tableau suivant :

|  |  |
| --- | --- |
| Arbre de décision | Taux de succès |
| rpart | 0.27 |
| tree | 0.45 |
| C5.0 | 0.66 |

**IV/Conclusion :**

Les deux arbres de décision rpart et tree sont, dans ce cas-là pas fonctionnel car la prédiction à plus de chances d’être fausse que vrai. En ce qui concerne C5.0 le taux de succès est supérieur a 50%, il est d’environ 2/3 ce qui peut nous donner une indication sur les futures déclarations mais néanmoins largement pas assez suffisant pour ne se baser que sur ces prédictions.