Les accidents de la Route en France

Projet Fil Rouge - **Partie I**

Exploration, Data Visualisation et Pré-Processing des données

Sommaire

MINISTÈRE
DE L'INTÉRIEUR
ET DES OUTRE-MER
Liberté
Égalité
Fraternité

DataScientest

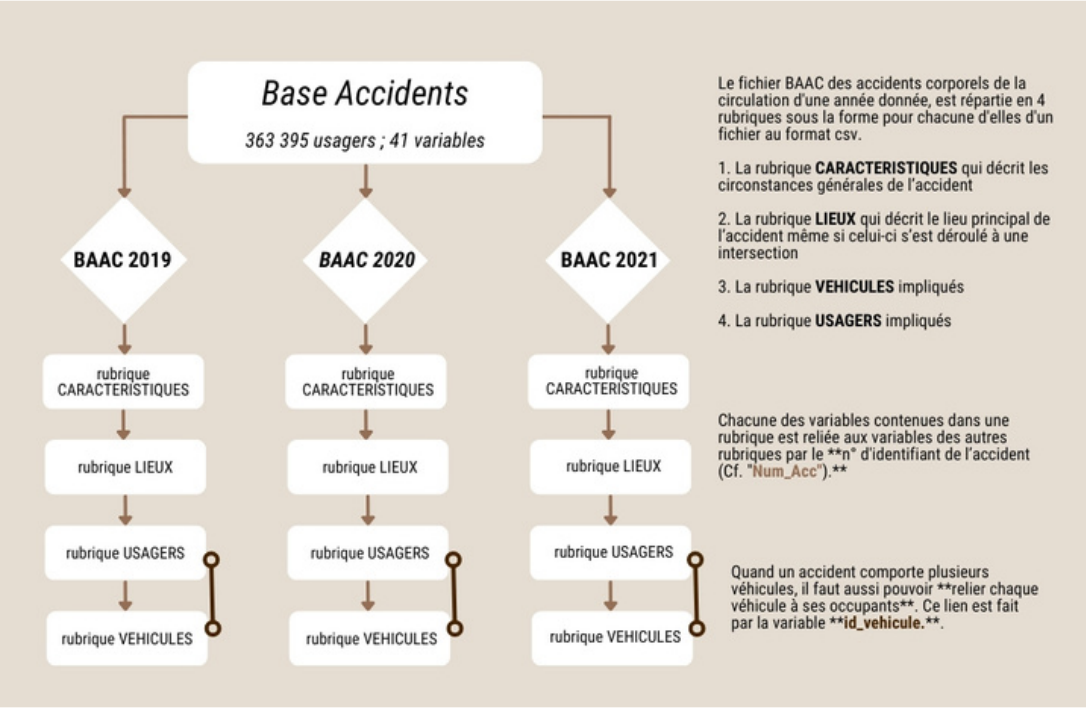
01. Contexte et Objectifs

L' Observatoire national interministériel de la sécurité routière met à disposition chaque année depuis 2005, des bases de données des accidents corporels de la circulation routière.

Pour chaque accident corporel (soit un accident survenu sur une voie ouverte à la circulation publique, impliquant au moins un véhicule et ayant fait au moins une victime ayant nécessité des soins), des saisies d'information décrivant l'accident sont effectuées par l'unité des forces de l'ordre (police, gendarmerie, etc.) qui est intervenue sur le lieu de l'accident.
Ces saisies sont rassemblées dans une fiche intitulée bulletin d'analyse des accidents corporels. L'ensemble de ces fiches constitue le fichier national des accidents corporels de la circulation dit « Fichier BAAC » administré par l'Observatoire national interministériel de la sécurité routière "ONISR".

Les bases de données, extraites du fichier BAAC, répertorient l'***intégralité des accidents corporels de la circulation, intervenus durant une année précise en France métropolitaine, dans les départements d'Outre-mer (Guadeloupe, Guyane, Martinique, La Réunion et Mayotte depuis 2012) et dans les autres territoires d'outre-mer*** (Saint-Pierre-et-Miquelon, Saint-Barthélemy, Saint-Martin, Wallis-et-Futuna, Polynésie française et Nouvelle-Calédonie; disponible qu'à partir de 2019 dans l'open data) avec une description simplifiée.

Cela comprend des informations de localisation de l'accident, telles que renseignées ainsi que des informations concernant les caractéristiques de l'accident et son lieu, les véhicules impliqués et leurs victimes.



Objectifs

L’ objectif de ce projet est de prédire la gravité de l'ensemble des accidents routiers en France intervenus entre 2019 et 2021.

Une **première étape** est d’étudier et appliquer des méthodes pour **nettoyer le jeu de données.**

Une fois le jeu de données propre, une **deuxième étape** est d’**extraire les caractéristiques qui semblent être pertinentes pour estimer la gravité des accidents**.

Ensuite, à partir de ses résultats, l’objectif est de travailler un scoring des zones à risque en fonction des informations météorologiques, l’emplacement géographique (coordonnées GPS, images satellite, …) …

Une fois l’entraînement du modèle effectué, nous allons comparer notre modèle avec les données historiques.

Objectif de la partie I

L'objectif de ce premier rapport est d'effectuer l'exploration, la visualisation et la pré-traitement de ces bases de données.

02. Exploration des données

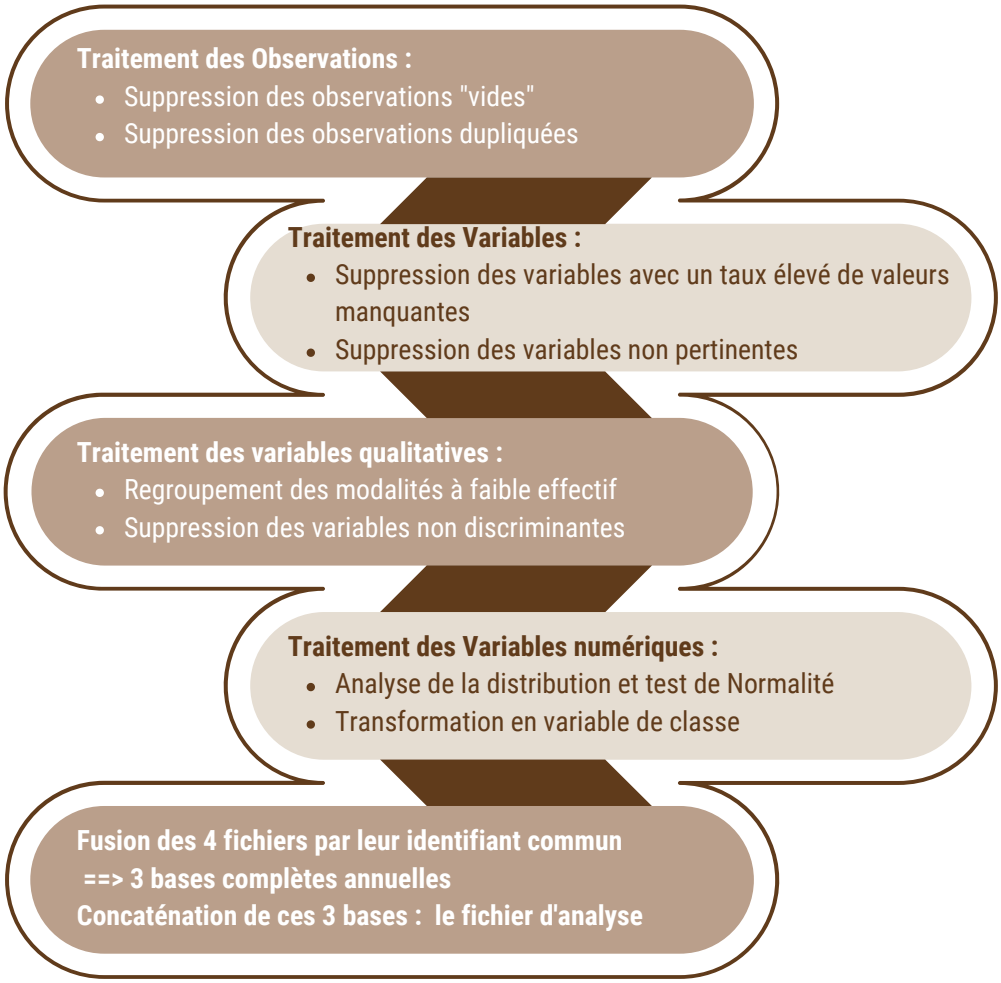
Bien que l'on puisse penser qu'il suffit d'un grand nombre de données pour avoir un algorithme performant, les données dont nous disposons sont souvent non adaptées. Il faut donc les comprendre et les traiter préalablement pour pouvoir ensuite les utiliser : **c'est l'étape d'exploration et de visualisation des données.**

En effet, des erreurs d'acquisition liées à des fautes humaines ou techniques peuvent corrompre notre dataset et biaiser l'entraînement. Parmi ces erreurs, nous pouvons citer des informations incomplètes, des valeurs manquantes ou erronées ou encore des bruits parasites liés à l'acquisition de la donnée. Il est donc souvent indispensable d'établir une stratégie de pré-traitement des données à partir de nos données brutes pour arriver à des données exploitables qui nous donneront un modèle plus performant.

La particularité de notre jeu de données est que nous avons 4 fichiers par année avec les mêmes variables.

Afin d'effectuer les mêmes traitements pour les 3 années, nous avons décidé d'analyser minutieusement chaque variable de chaque rubrique de l'année 2021, puis d'appliquer exactement les mêmes transformations pour les rubriques des années 2019 et 2020.
En utilisant cette méthodologie de travail, nous nous sommes assurés d'obtenir un jeu de données par année avec les mêmes variables et les mêmes transformations.

Traitement effectué sur chacune des 4 rubriques de l'année 2021

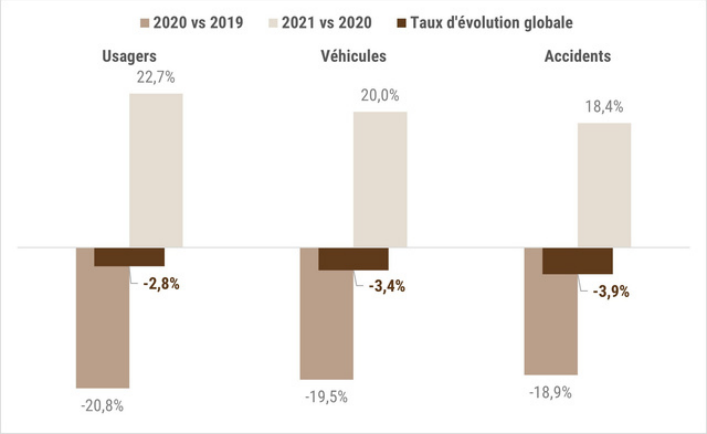


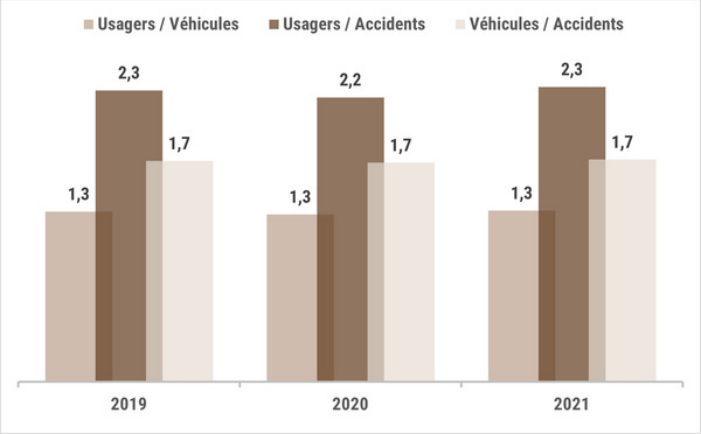
Les grandes tendances :

En **2019**, il y a eu 132 876 usagers et 100 710 véhicules impliqués dans 58 840 accidents.

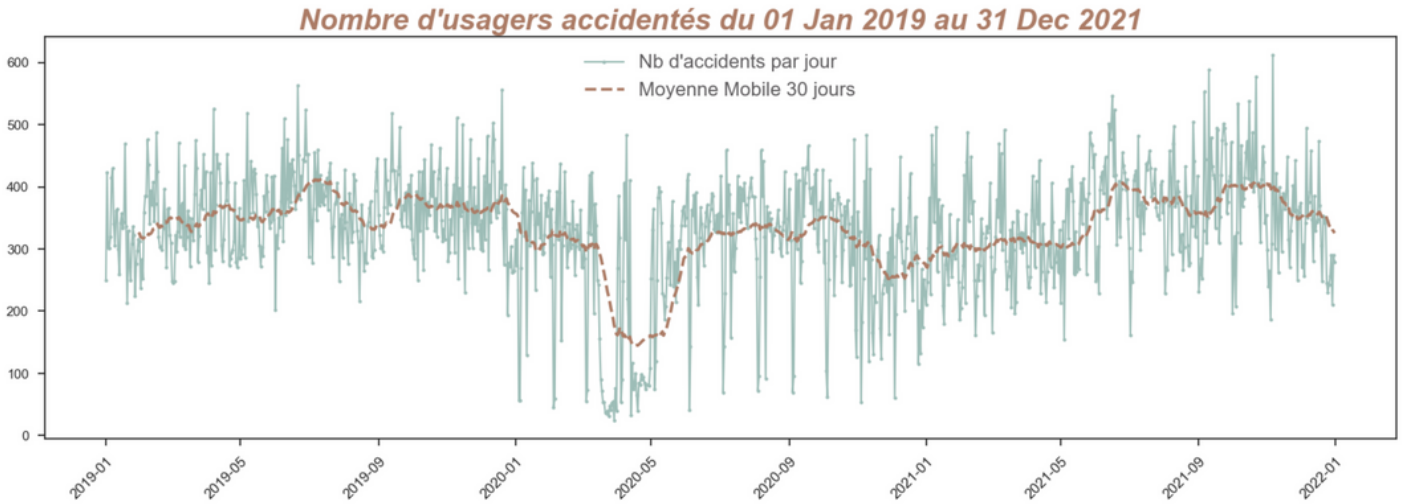
En **2020**, il y a eu 105 232 usagers et 81 066 véhicules impliqués dans 47 744 accidents.

En **2021**, il y a eu 129 153 usagers et 97 315 véhicules impliqués dans 56 518 accidents.

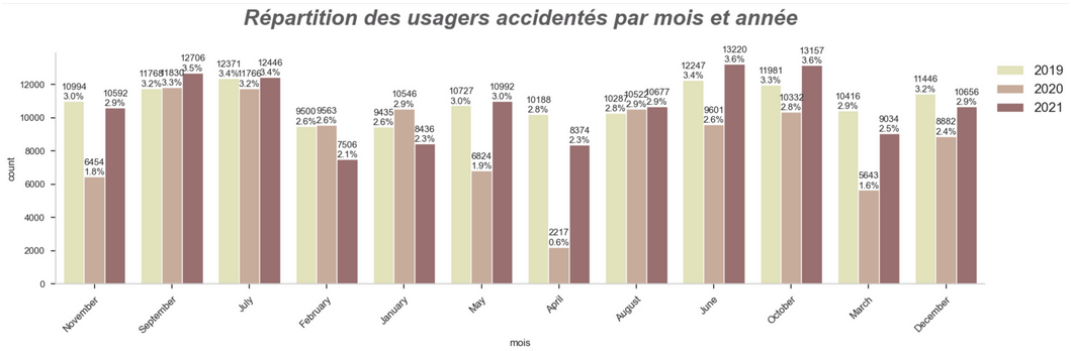




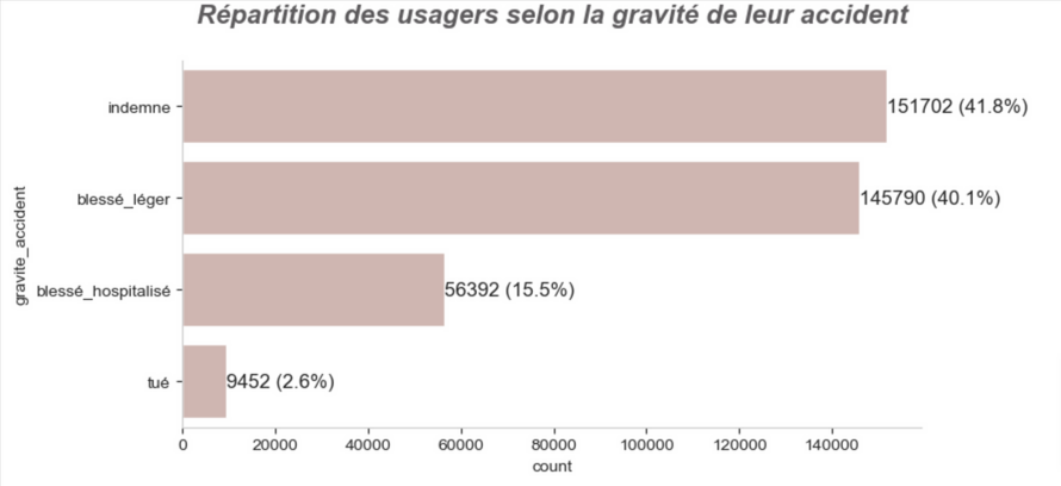
Même si les évolutions annuelles sont très erratiques, on constate une légère diminution du taux de croissance global de 4% tandis que le nombre moyen d'usagers par accident et le nombre moyen de véhicules par accident, restent stables.



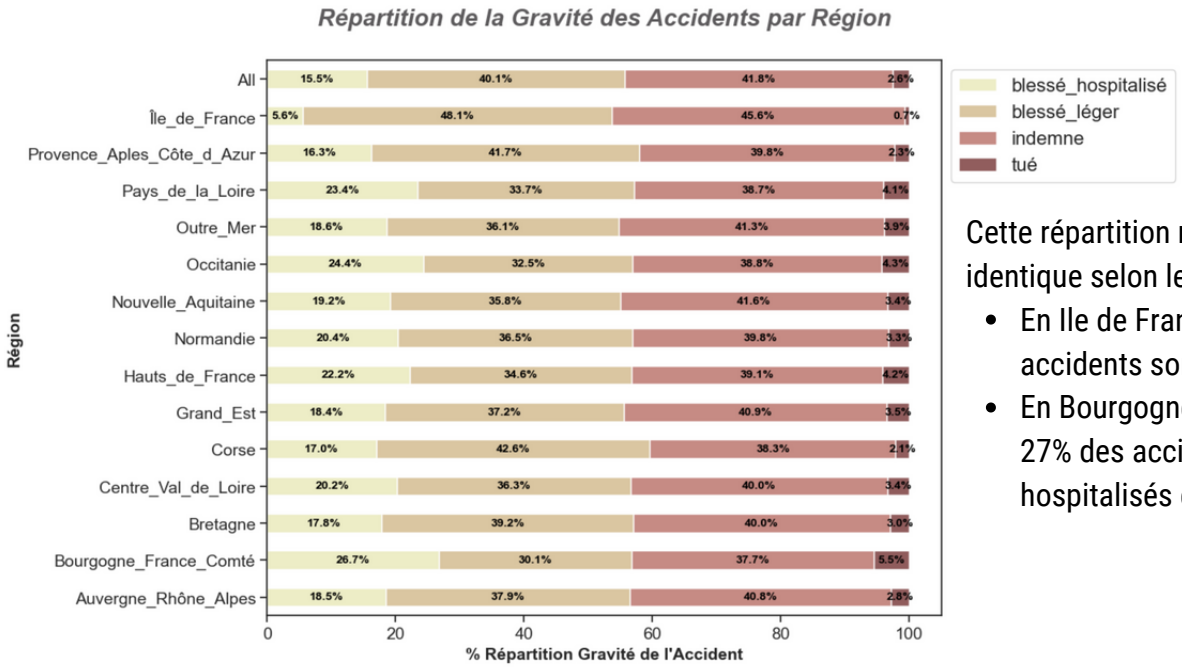
L'année 2020 est en effet très différente des 2 autres années. Alors qu'en 2019 et 2021, le nombre d'accidents est réparti de manière homogène sur les 12 mois de l'année, on constate que ce n'est pas le cas en 2020 ; il y a eu nettement moins d'accidents en avril, mai et novembre,en raison des confinements liés au Covid-19.



03. La variable cible : la gravité des accident

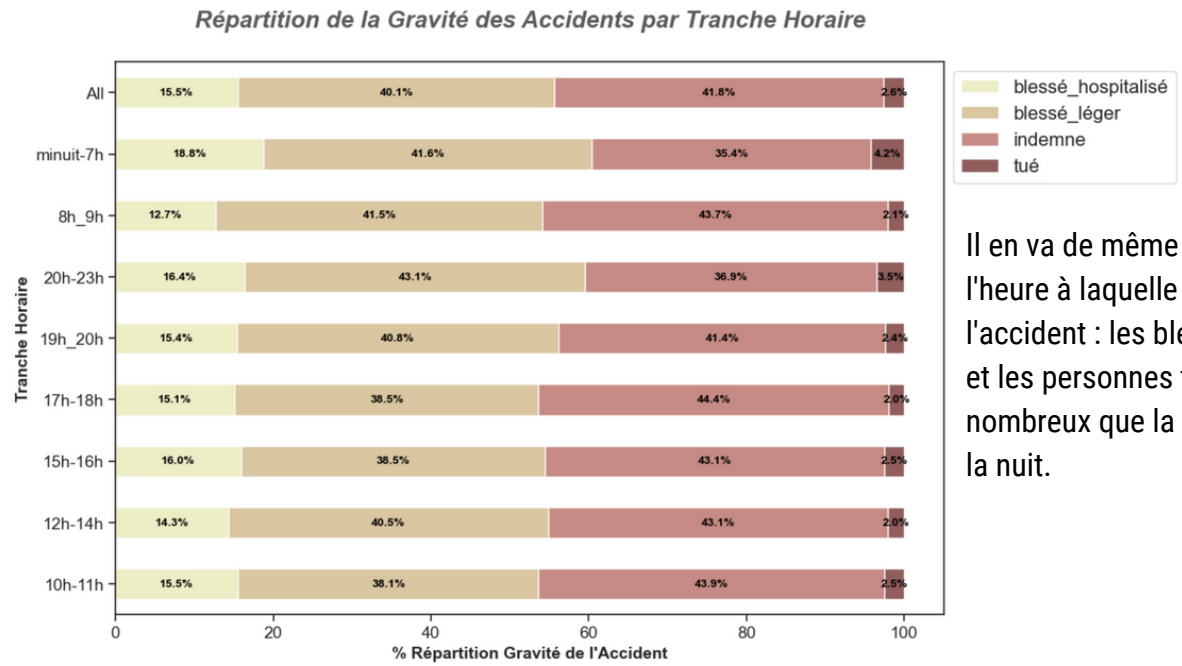


Fort heureusement, Près de 82% des usagers impliqués dans un accident routier sont indemnes ou ont de légères blessures. En revanche, près de 3% d'entre eux sont tués (environ 3 150 personnes par an)!



Cette répartition nationale n'est pas identique selon les régions :

- En Ile de France 94% des accidents sont sans gravité
- En Bourgogne-Franche-Comté, 27% des accidentés doivent être hospitalisés et 5,5% sont tués.



Il en va de même au regard de l'heure à laquelle s'est passé l'accident : les blessés hospitalisés et les personnes tuées sont plus nombreux que la moyenne nationale la nuit.

04. Pre-Processing

Après avoir nettoyé et décrit nos données brutes, puis déterminé notre variable cible, nous obtenons un dataset de 363 336 usagers impliqués dans un accident routier caractérisés par 40 variables qualitatives.

Nous devons maintenant les transformer en données utilisables par les différents algorithmes de Machine Learning pour les 3 raisons principales suivantes :

- La majorité des algorithmes ont des contraintes sur le format des données en entrée. Cela peut concerner leur type, par exemple uniquement des variables numériques, ou des contraintes sur leur format, comme des réels entre 0 et 1.
- Préparer les données permet de grandement améliorer les résultats des algorithmes, en extrayant ou en créant des colonnes plus adaptées au problème
- Ce n'est pas parce que la fonctionnalité existe dans notre jeu de données qu'elle est pertinente pour notre modèle et que nous devons l'utiliser. En effet, ***"Les bons sous-ensembles de caractéristiques contiennent des caractéristiques hautement corrélées avec la cible, mais non corrélées les unes aux autres".***

04.1. Quelles sont les caractéristiques corrélées à la cible ?

L'objectif ici est de sélectionner les variables qui ont un lien statistique avec la variable cible.
Pour cela, nous avons utilisé le ***Test du χ2 d'indépendance.***

Le test d'indépendance du χ2 est utilisé pour **déterminer s'il existe une relation significative entre deux variables catégorielles (nominales)**.
Cela signifie que le test d'indépendance du χ2 est un test d'hypothèse avec 2 hypothèses présentes ; l'hypothèse nulle et l'hypothèse alternative.

- ***Hypothèse nulle (H0) : Il n'y a pas de relation entre les variables***
- ***Hypothèse alternative (H1) : Il existe une relation entre les variables***

Comme tout test statistique, nous le testons par rapport à notre valeur p choisie (souvent elle est de 0,05).
Si la valeur de p est significative (p ≤ 0,05), nous pouvons rejeter l'hypothèse nulle et affirmer que les résultats soutiennent l'hypothèse alternative.

Or, toutes nos p-values sont nulles : Il existe une relation significative entre la variable cible et les 40 variables qualitatives explicatives dont nous disposons.

Cependant, comme nous avons plusieurs modalités dans chacune de nos variables catégorielles, nous ne sommes pas en mesure de dire facilement quelle est celle qui est responsable de la relation (si la table χ2 est supérieure à 2 × 2).
Pour identifier quelle(s) modalité(s) est(sont) responsable(s), nous avons besoin d'un test post hoc.
Mais, il faut d'abord quantifier nos variables catégorielles.

04. Pre-Processing

04.2. Transformation des caractéristiques

Cette étape de prétraitement regroupe les changements effectués sur la structure même de la donnée. Ces transformations sont liées aux définitions mathématiques des algorithmes et à la manière dont ceux-ci traitent les données, de manière à optimiser les performances.

Nous avons essentiellement utilisé :

- La **discrétisation de variables continues** (à l'aide du découpage en intervalles) qui permet d'abaisser le nombre de modalités d'une variable et d'en supprimer les éventuelles valeurs aberrantes.
- La **"dummification" des variables catégorielles nominales** ou la variable initiale est remplacée par un ensemble de nouvelles variables dites fictives (ou indicatrices) prenant les valeurs 0 ou 1 pour indiquer l'absence ou la présence de la modalité.

Base de données après Transformation
Nombre d'usagers ayant eu un accident de la route entre 2019 et 2021 : 363 336
Nombre de variables explicatives : **126**

04.3. Réduction du nombre de caractéristiques

Lors de la construction d'un modèle prédictif, nous avons souvent de nombreuses fonctionnalités ou variables dans notre ensemble de données qui peuvent être utilisées pour former notre modèle.

Cependant, ce n'est pas parce que la fonctionnalité existe dans notre jeu de données qu'elle est pertinente pour notre modèle et que nous devons l'utiliser. En effet, **"Les bons sous-ensembles de caractéristiques contiennent des caractéristiques hautement corrélées avec la cible, mais non corrélées les unes aux autres".**

Alors, comment savons nous quelles fonctionnalités utiliser pour notre modèle ?

C'est là qu'intervient la **sélection des caractéristiques**.

La sélection des caractéristiques est simplement un processus qui réduit le nombre de variables d'entrée, afin de ne conserver que les plus importantes.

Il existe ***trois catégories de méthodes de sélection de fonctionnalités***, en fonction de la manière dont elles interagissent avec la variable à prédire, à savoir les ***méthodes de filtrage, d'encapsulation et intégrées.*** Nous les avons utilisées toutes les 3.

04.3.1. Test du χ2 d'indépendance avec les valeurs p ajustées de Bonferroni

La correction de Bonferroni consiste à diviser le niveau alpha souhaité par le nombre de comparaisons pour utiliser le nombre ainsi calculé comme valeur p afin de déterminer la signification. Ainsi, par exemple, avec alpha fixé à 0,05 et trois comparaisons, la valeur p-ajusté requise pour la signification serait de 0,05/3 = 0,0167.

04. Pre-Processing

En utilisant la valeur P ajustée, nous pouvons alors tester tous les résultats précédemment significatifs pour voir **quelle modalité est responsable de la création d'une relation significative entre la variable cible et nos 126 variables indicatrices.**

Or, toutes nos p-values sont nulles : Il existe une relation significative entre la variable cible et les 126 modalités des 40 variables qualitatives explicatives dont nous disposons.

04.3.2. Analyse des corrélations

L'analyse des corrélations évalue des sous-ensembles de caractéristiques sur la base de l'hypothèse suivante : **"Les bons sous-ensembles de caractéristiques contiennent des caractéristiques hautement corrélées avec la cible, mais non corrélées les unes aux autres".**

Nous avons déjà constaté que toutes nos variables indicatrices sont hautement corrélées avec notre variable cible grâce au Test du $\chi 2$ d'indépendance et les valeurs p ajustées de Bonferroni.

Nous allons maintenant effectuer des **tests d'indépendance entre les variables explicatives par pair afin d'exclure celles potentiellement non porteuses d'information.**

Pour cela nous allons analyser les **coefficients de corrélations de Spearman** pour ne pas évaluer que les relations linéaires.

La corrélation de Spearman entre deux variables est égale à la corrélation de Pearson entre les valeurs de rang de ces deux variables ; alors que la corrélation de Pearson évalue les relations linéaires, la corrélation de Spearman évalue les relations monotones (qu'elles soient linéaires ou non). Le coefficient de corrélation sur les rangs (Rho de Spearman) s'interprète de la même manière qu'un coefficient de corrélation de Pearson : une valeur positive (maximum = +1) indique une variation simultanée dans le même sens, une valeur négative (minimum = -1) une variation simultanée en sens inverse.

Mais nous obtenons un tableau de coefficients de Spearman de 101 x 101 (puisque nous avons pu éliminer 25 caractéristiques quasi-constantes qui montrent la même valeur pour la grande majorité des observations de l'ensemble de données).

Après avoir supprimer les p_value des variables avec elles-mêmes et les pairs identiques, il reste 4207 pairs de variables non indépendantes (p-value < 0.05) et 842 pairs de variables indépendantes !

On ne peut donc pas utiliser cette méthode pour sélectionner les variables pertinentes pour la réalisation de nos objectifs, mais elle nous montre le potentiel de réduction de nos variables explicatives puisqu'il n'y a que 842 pairs de variables indépendantes !

04. Pre-Processing

04.3.3. Les méthodes Embedded

Les méthodes Embedded intègrent la sélection de fonctionnalités à la construction de l'algorithme d'apprentissage automatique.
Les méthodes intégrées **effectuent la sélection des fonctionnalités lors de l'apprentissage du classifieur** ou des régresseurs**.** Ces méthodes sont donc embarquées dans l'algorithme.

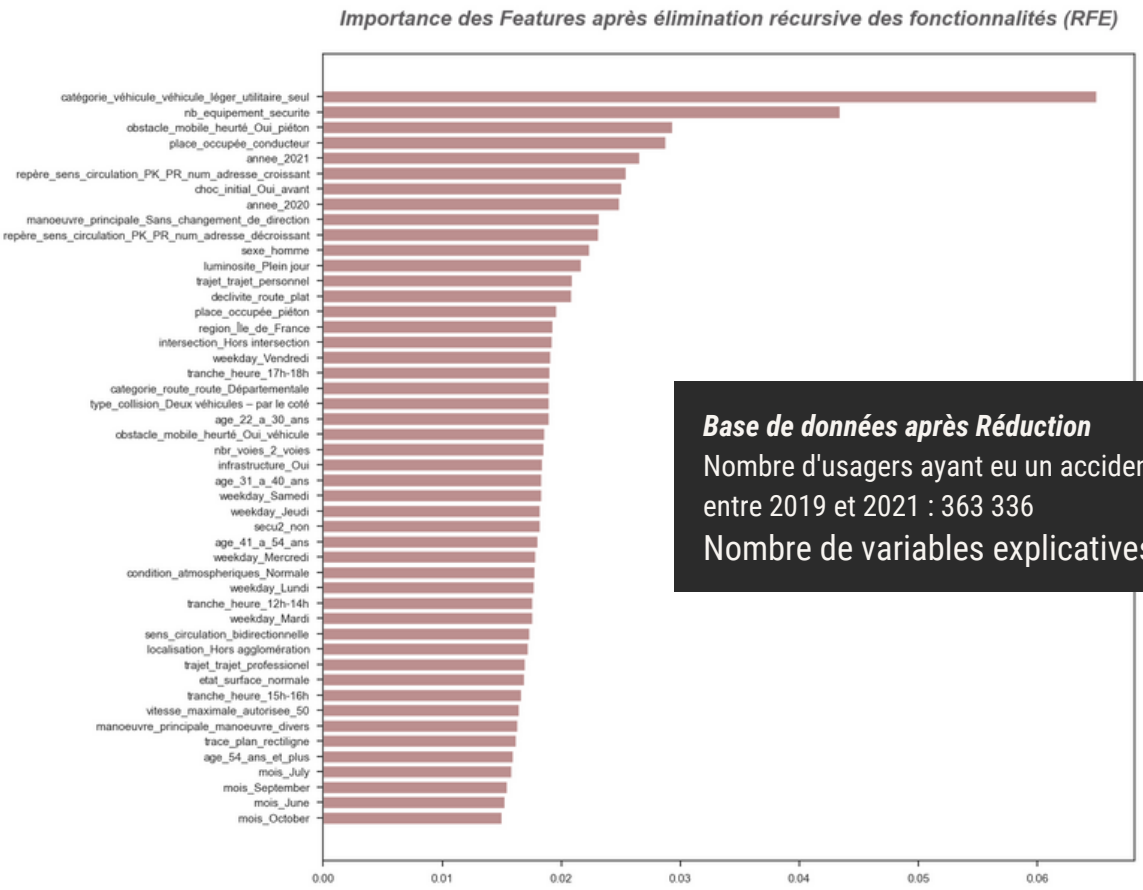
Les méthodes embarquées ont les avantages à la fois des Méthodes Wrapper (que nous n'utiliserons pas ici car elles demandent beaucoup de temps et de ressources) et Filter. Elles incluent l'interaction des caractéristiques avec le modèle de classifieur ou le régresseur modèle tout comme les méthodes Wrapper, et comme les méthodes de filtrage, elles sont beaucoup moins gourmandes en calcul. De plus, les méthodes embarquées sont capables de détecter l'interaction entre les variables car elles évaluent l'ensemble des données en même temps et trouvent le sous-ensemble de fonctionnalités qui convient à l'algorithme en cours de formation.

Cela signifie que les méthodes intégrées sont généralement la méthode de choix au moment de la sélection des fonctionnalités.

Les méthodes embarquées font partie de la formation de l'algorithme d'apprentissage automatique. Ainsi, les étapes typiques impliquent la formation d'un algorithme d'apprentissage automatique en utilisant toutes les fonctionnalités, puis en dérivant l'importance de ces caractéristiques selon l'algorithme utilisé, et enfin en supprimant les fonctionnalités non importantes en suivant certains critères qui dépendront de l'algorithme lui-même.
Un exemple de méthodes embarquées est la régularisation LASSO (pour la régression). Et dans une certaine mesure, nous pourrions également utiliser les coefficients de régression des modèles linéaires pour évaluer l'importance des différentes caractéristiques de l'ensemble de données et décider lesquelles conserver et lesquelles exclure.
Les arbres et les forêts aléatoires sont également très utiles pour sélectionner des fonctionnalités.

En résumé, les méthodes embarquées donnent une **meilleure précision prédictive** que les méthodes de filtrage car elles sont rapides et moins coûteuses en calcul que les méthodes wrapper. Elles ont également **tendance à trouver un très bon sous-ensemble de fonctionnalités** pour l'algorithme donné. En revanche, nous devons garder à l'esprit les caractéristiques des différents algorithmes pour comprendre l'importance dérivée de ce que nous pouvons faire avec eux.
Le processus d'ingénierie des caractéristiques consiste à **sélectionner les caractéristiques minimales requises pour produire un modèle valide, car plus un modèle contient de caractéristiques, plus il est complexe** (et plus les données sont rares), **donc plus le modèle est sensible aux erreurs dues à la variance.**

Une approche courante pour éliminer les caractéristiques consiste à décrire leur importance relative pour un modèle, puis à éliminer les caractéristiques faibles ou les combinaisons de caractéristiques et à réévaluer pour voir si le modèle se comporte mieux lors de la validation croisée, c'est la **méthode RFE pour élimination récursive des fonctionnalités.**



Base de données après Réduction

Nombre d'usagers ayant eu un accident de la route entre 2019 et 2021 : 363 336

Nombre de variables explicatives : **48**

05. Conclusion

Notre jeu de données est maintenant prêt pour la phase suivante de modélisation.

Nous l'avons décomposé en 2 sous-ensembles : 70% pour entraîner nos modèles de prédiction et 30% pour les tester.

Nous n'avons pas réalisé de standardisation des données puisque nous ne disposons que de variables explicatives dichotomiques.

Nous avons choisi, dans un premier temps, d'entraîner des modèles de Classification binaire en transformant la variable cible :

- "blessés légers", "blessés hospitalisés", "tués"
- Indemnes

Dans une seconde étape, nous utiliserons des modèles de Classification muticlasse avec la variable cible en 4 classes.