Rapport de Projet : Analyse et Prédiction de la Gravité des Accidents de la Route (2019)

Partie 1 : Introduction et Nettoyage des Données (Data Cleaning)

1.1. Contexte et Choix du Dataset

Pour ce projet, nous avons choisi de travailler sur les données des accidents de la route en France pour l'année 2019, issues de *data.gouv.fr*. Ce sujet nous a semblé particulièrement intéressant car il traite d'un enjeu de santé publique majeur. L'objectif est de prédire la **gravité de l'accident** pour un usager (variable grav) en fonction de facteurs environnementaux et personnels.

Le jeu de données est complexe car il est divisé en quatre fichiers (Caractérisques, Lieux, Véhicules, Usagers). Après avoir fusionné ces fichiers, nous disposions d'un échantillon initial massif de plus de **250 000** lignes et **56 variables**.

1.2. Stratégie de Nettoyage

Le nettoyage a été l'étape cruciale de notre travail pour transformer ces données brutes en un format exploitable par nos modèles de Machine Learning :

- **Fusion (Merge) :** Nous avons regroupé les fichiers autour de l'identifiant technique de l'accident (Num_Acc).
- **Gestion de l'âge :** Nous avons calculé l'âge à partir de l'année de naissance. Nous avons filtré les données pour supprimer les valeurs aberrantes (âges négatifs ou supérieurs à 120 ans).
- **Sélection des variables :** Nous avons décidé de nous concentrer sur 7 variables qui nous semblaient les plus impactantes : age, sexe, la catégorie du véhicule (catv), la luminosité (lum), les conditions météo (atm), le type de route (catr) et la localisation en agglomération (agg).
- **Valeurs manquantes :** Pour la variable numérique (age), nous avons utilisé la **médiane** pour ne pas être influencés par les extrêmes (même si c'était inutile dans notre jeu de données).
- Pour les variables catégorielles, nous avons complété les trous par le **mode** (la valeur la plus fréquente).
- **Réduction de volume :** Après suppression des doublons et des lignes incomplètes, nous avons stabilisé notre dataset à **68 500 entrées**, ce qui offre un bon compromis entre représentativité et temps de calcul.

Partie 2 : Analyse Exploratoire et Modélisation

2.1. Analyse Exploratoire (EDA)

Avant de lancer l'apprentissage, nous avons cherché à comprendre la structure de nos données.

- **Profil type :** L'âge moyen des personnes impliquées est de 41 ans.
- **Corrélations :** Notre matrice de corrélation a révélé des liens logiques, notamment entre le type de route (catr) et le fait d'être en agglomération (agg).
- **Déséquilibre des classes :** Nous avons remarqué que la variable cible grav n'est pas uniformément répartie (certains types de gravité sont beaucoup plus fréquents que d'autres), ce qui rend la prédiction difficile pour les classes minoritaires.

2.2. Préparation des données (Preprocessing)

Pour que les algorithmes fonctionnent correctement, nous avons mis en place un **Pipeline** de traitement :

- **Standardisation :** L'âge a été centré et réduit (StandardScaler) pour éviter qu'il n'écrase les autres variables par sa magnitude.
- **Encodage :** Les variables catégorielles (météo, sexe, etc.) ont été transformées via un OneHotEncoder pour être lisibles par les modèles mathématiques.
- **Découpage :** Nous avons séparé les données en un ensemble d'entraînement (80%) et un ensemble de test (20%).

2.3. Construction des Modèles

Nous avons testé deux approches différentes pour notre tâche de classification :

- **La Régression Logistique :** Utilisée comme modèle de référence (baseline). Elle est simple et permet de voir facilement l'influence de chaque variable.
- **Le Random Forest (Forêt Aléatoire) :** Un modèle plus robuste capable de capturer des relations non-linéaires entre les variables (par exemple, l'interaction entre la pluie et le type de route).

Partie 3 : Évaluation, Interprétation et Conclusion

3.1. Résultats et Comparaison

Nos deux modèles ont obtenu des performances plutôt faibles, avec une **Accuracy (précision globale)** d'environ **30% pour la régression logistique** et d'environ **23% pour la Random Forest**.

Ce score peut sembler bas, mais il s'explique par le fait que la gravité d'un accident dépend de facteurs imprévisibles (vitesse au moment du choc, port de la ceinture, état mécanique) qui ne sont pas présents dans le dataset de base.

3.2. Interprétation des résultats (Feature Importance)

L'analyse des coefficients de notre modèle de régression nous a permis de tirer des conclusions intéressantes :

- **L'Âge :** Le coefficient est négatif (-0.18) pour la classe "Sévère". Cela suggère que, dans ce dataset, les personnes plus âgées sont statistiquement moins impliquées dans des accidents très graves, peut-être en raison d'une conduite plus prudente ou de trajets moins risqués.
- **Le Sexe :** Les hommes ont une probabilité plus élevée d'être associés à une gravité forte (classe 4).
- **L'Environnement :** Le facteur "Hors Agglomération" est un prédicteur fort de gravité élevée, sans doute à cause de la vitesse autorisée plus importante.

3.3. Recommandations et Conclusion

Ce projet nous a permis de valider toute la chaîne de traitement de la donnée. Pour améliorer ces résultats à l'avenir, nous préconisons :

- **L'équilibrage des données :** Utiliser des techniques comme le SMOTE pour donner plus de poids aux accidents graves, souvent moins nombreux dans les fichiers.
- **Nouvelles Features :** Intégrer des données sur la vitesse réelle et l'alcoolémie, qui sont des facteurs majeurs de gravité.

En conclusion, bien que la prédiction exacte soit complexe, notre modèle permet déjà d'identifier des profils à risque (jeunes conducteurs, zones hors agglomération).

Ce travail montre que le nettoyage et la préparation des données représentent une grande partie du travail d'un Data Scientist.