Projet SD201 : Analyse de la gravité d'accidents routiers

10 décembre 2021

Ariel GABAI

Sami CONTESENNE

TABLE DES MATIERES

Ariel GABAI

Sami CONTESENNE

INTRODUCTION

Les accidents de la route ont énormément diminué au cours des dernières années notamment grâce à des études d'accidents établies par des experts permettant par exemple de créer des designs de voiture optimisant la protection des usagers. Nous voulons nous aussi étudier ces accidents en utilisant les sciences des données.

Nous avons donc choisi d'étudier des bases de données répertoriant des accidents routiers, contenant de multiples informations à propos de ces derniers, par exemple sur l'environnement, le véhicule, et le conducteur.

Nous souhaitons à travers cette étude répondre aux problématiques suivantes :

Quels sont les facteurs qui influent sur la dangerosité d'un accident ?

Peut- on prédire la dangerosité d'un accident à partir d'informations le concernant ?

Connaitre ces facteurs permettrait de mieux prévenir et éviter les comportements qui augmentent les chances d'avoir un accident grave.

Prédire la gravité de l'accident permettrait une réponse plus rapide et efficace des secours ainsi que des témoins. En effet, une description de l'accident plus ou moins précise, pourrait faire office d'examen préliminaire et permettrait d'avoir une première idée des ressources à employer.

Dans cette optique, nous avons recherché une base de données contenant la description d'accidents de la route.

Ariel GABAI

Sami CONTESENNE

PARTIE 1 : Recherche, étude et nettoyage des données

1.1 Recherche de la base de données appropriée

La première étape de notre travail a été de trouver une base de données assez grande et détaillée pour être représentative de la réalité. Cette étape nous a demandé plus de temps et de travail que prévu. En effet, une première tentative a été d'étudier une base de données répertoriant les accidents en France, cependant, celui-ci ne contenait pas d'informations concernant le conducteur comme nous le souhaitions. Par ailleurs, plusieurs champs possédaient la valeur nulle, ce qui le rendait inexploitable. C'est pourquoi, il a fallu poursuivre les recherches jusqu'à trouver une base de données plus convaincante. Nous avons finalement décidé de nous appuyer sur trois bases de données rassemblant les différentes caractéristiques d'accidents qui ont eu lieu aux Etats-Unis, au cours de l'année 2019. Bien que ces bases de données n'étaient pas prêtes à être analysées, nous avons estimé que le temps nécessaire pour les préparer était suffisamment raisonnable.

Ces trois bases de données représentent respectivement les caractéristiques du conducteur, du véhicule ainsi que de l'accident. Elles peuvent facilement être jointe grâce à un identifiant dépendant de l'accident commun.

Cliquez ici pour visualiser les bases de données et avoir plus d'informations sur celles-ci.

1.2 Etude exhaustive des caractéristiques renseignées dans les bases de données

En amont de l'analyse des données, il a fallu étudier en détail les différentes caractéristiques renseignées par cette étude. Il a fallu dans un premier temps, se familiariser avec chacune des codifications utilisées pour les évaluer. Ainsi, on a pu identifier les attributs ayant un intérêt dans le cadre de notre étude. Cette étude nous a également été d'une grande aide pour l'exploitation des résultats fournis par les algorithmes de prédiction. Cependant, on s'est rendu compte de la nécessité de pouvoir faire appel à un expert du sujet étudié, ne serait-ce que pour éliminer les attributs inutiles.

1.3 Nettoyage et jointure des bases de données

La dernière étape avant de pouvoir réellement exploiter les données a été le retrait des colonnes et des lignes inutiles ou impertinentes des différentes bases de données.
Par exemple, on a retiré l'année pendant laquelle a eu lieu l'accident car c'est une donnée qui nous a paru impertinente, d'autant plus que la totalité des accidents répertoriés ont eu lieu la même année. En utilisant cette même logique, nous avons pu retirer d'autres colonnes de la base de données. Nous avons également fait le choix de ne garder uniquement les caractéristiques des conducteurs pour simplifier l'étude.
Après avoir « nettoyé » les trois bases de données on a fait une jointure de celles-ci, notamment grâce à l'identifiant commun qu'elles possèdent. On a donc conservé une unique base de données avec une cinquantaine de caractéristiques concernant, le véhicule, le conducteur ainsi que l'environnement dans lequel a eu lieu l'accident.
Le but étant d'appliquer différents algorithmes de machine learning à cette base de données pour faire une classification ainsi qu'une prédiction, nous avons choisi comme cible la blessure maximale qu'il y a eu lors de l'accident. Celle-ci étant quantifiée par un nombre entre 0 et 4 où 0 signifie « pas de blessure » et 4 signifie « mort ».
Dans ce modèle la dangerosité de l'accident correspond donc à la plus haute gravité de blessure dans l'accident. Ceci peut être critiqué, une métrique plus complexe faisant par exemple également intervenir le nombre de blessés aurait pu être pertinente, là encore nous avons pu constater que l'avis d'un expert du domaine était important pour faire ces choix.
Avant de procéder à l'application de ces algorithmes, nous avons retiré les colonnes dans lesquelles nous avions une explication de la signification des chiffres de la colonne précédente.
Une fois toutes ces manipulations effectuées, la base de données était prête à être analysée par les algorithmes de *machine learning*.

PARTIE 2 : Analyse et interprétation des données

2.1 Résultats obtenus pour chacun des algorithmes

Nous avons décidé de nous baser sur trois algorithmes pour prédire la gravité de l'accident en fonction des caractéristiques de celui-ci.

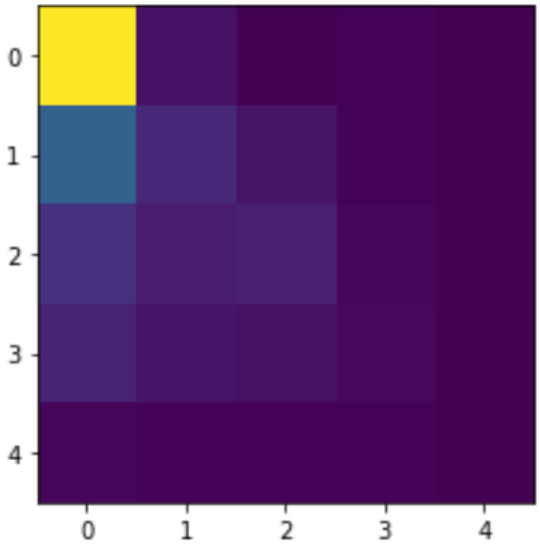
Dans un premier temps, nous avons choisi d'utiliser un **Decision Tree** qui construit un arbre de décision en choisissant pour décision les caractéristiques qui maximisent le plus l'entropie. Cet algorithme a donné sur le test set qu'on lui a fourni une précision d'environ 0.57. C'est une précision qui nous a un peu déçu, cependant, on peut se rendre compte que l'algorithme reste bien meilleur que de l'aléatoire. En effet, l'aléatoire aurait donné une précision d'environ 0.2 sur un très grand test set. De plus, la matrice de confusion nous montre que l'algorithme, lorsqu'il se trompe, l'erreur n'est pas grossière (1 au lieu de 0 ou bien 2 au lieu de 3…).

```
Predicted values:
[1 0 0 ... 0 0 0]
Accuracy :  57.18762636473919
Report :                precision    recall  f1-score   support

           0       0.64      0.94      0.76      4924
           1       0.36      0.23      0.28      2235
           2       0.45      0.28      0.34      1512
           3       0.34      0.10      0.15      1021
           4       0.00      0.00      0.00       200

    accuracy                           0.57      9892
   macro avg       0.36      0.31      0.31      9892
weighted avg       0.50      0.57      0.51      9892
```

Résultats de l'algorithme de Decision Tree

Ariel GABAI

Sami CONTESENNE



Matrice de confusion algorithme Decision Tree

On remarque dans la partie report qu'on a une précision de 0 lorsqu'il s'agit de prédire un accident de dangerosité 4. On verra dans la prochaine section dédiée aux limites du modèle les possibles raisons pour lesquels nous obtenons ce résultat.

On a ensuite utilisé l'algorithme de **Random Forest** pour voir s'il donnait une meilleure précision, mais on s'est rendu compte que celle-ci ne changeait pas et qu'elle valait également environ 0.57.

Accuracy : 57.0966437525273
Report :

	precision	recall	f1-score	
0	0.65	0.92	0.76	4924
1	0.38	0.23	0.29	2235
2	0.40	0.30	0.34	1512
3	0.34	0.13	0.19	1021
4	0.18	0.01	0.02	200
accuracy			0.57	9892
macro avg	0.39	0.32	0.32	9892
weighted avg	0.51	0.57	0.52	9892

Résultats de l'algorithme de Random Forest

Ariel GABAI

Sami CONTESENNE

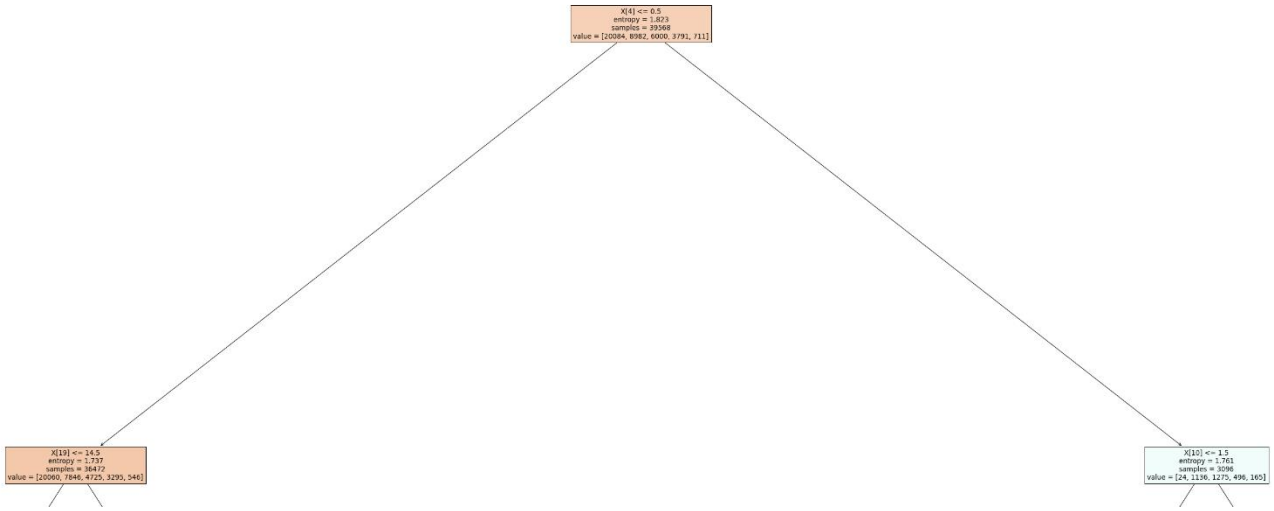
La précision n'a pas changé globalement, cependant on remarque qu'une très petite amélioration pour la prédiction de la classe 4.

Enfin, on a essayé d'utiliser un algorithme basé sur un autre principe pour essayer d'obtenir une meilleure précision. Pour cela, on a utilisé l'algorithme Gradient Boosting, mais là encore on a trouvé une précision quasiment identique.

2.2 Interprétation des résultats de l'algorithme Decision Tree

Nous avons décidé d'utiliser un Decision Tree car les résultats de cet algorithme sont très simples à interpréter. En effet nous pouvons voir quelles caractéristiques séparent le mieux les données et leur niveau d'influence sur la dangerosité des accidents.

La forme du Decision Tree ayant donné les résultats de la section 2.1 permet donc de faire des interprétations sur nos données.

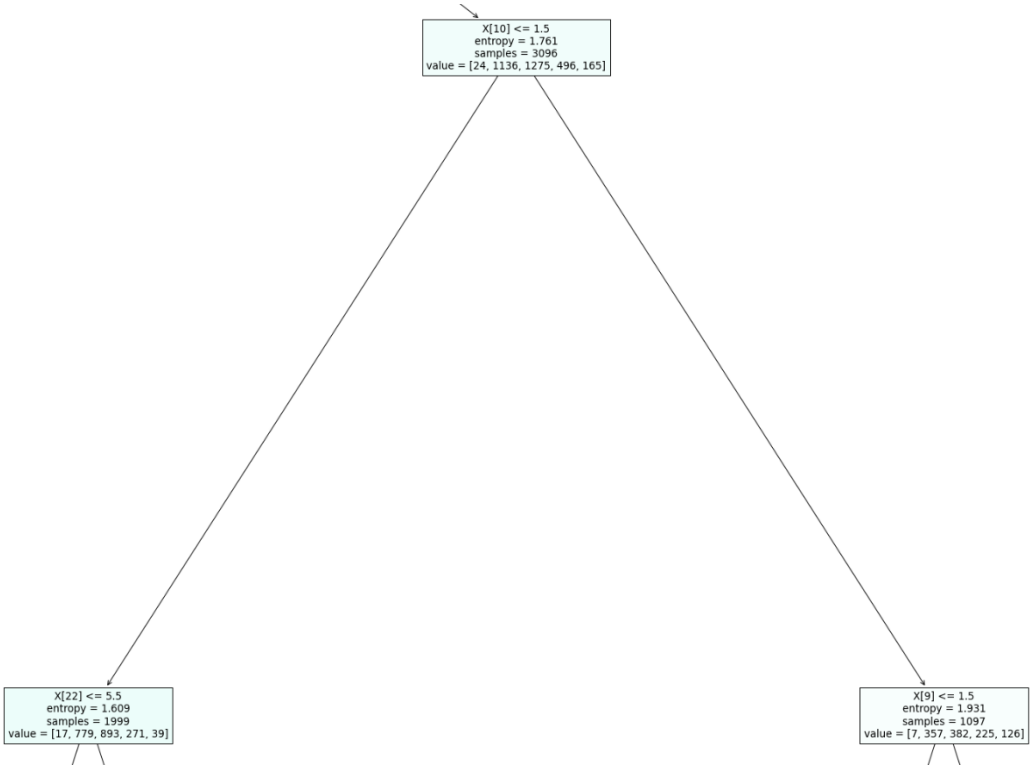


Racine et premier niveau de l'arbre

L'arbre possédant une profondeur de 4 il est difficile de pouvoir lire les valeurs sur les nœuds en l'affichant complètement, c'est pour cette raison que l'on affichera seulement les parties que nous souhaitons commenter, l'arbre complet est disponible en pièce jointe.

Nous voyons que la première caractéristique pour séparer les données est la 4ème correspondant à la présence (1) ou non (0) de piétons lors de l'accident. On observe

que les accidents de niveau de dangerosité 0 n'existe presque plus dans le fils droit de l'arbre (correspondant donc à la présence de piéton lors de l'accident). Cela paraît cohérent, en effet les piétons ne sont pas protégés contrairement aux conducteurs qui le sont par la carrosserie de leurs véhicules.



Niveaux 1(fils droit de la racine) et 2 de l'arbre

Ici nous avons déjà l'hypothèse que des piétons sont impliqués dans l'accident. La seconde séparation choisie par l'algorithme est sur l'éclairage de la route. Du côté gauche les accidents où la route est éclairée par le jour, à droite le reste des accidents (c'est-à-dire des routes de nuit éclairées ou non). On remarque qu'après cette séparation, la majorité des accidents ayant entraînés une blessure fatale sont sur les routes de nuits. Cela est aussi cohérent, en effet un conducteur avec une vision réduite ne pourra réagir que très tard à la présence d'un piéton sur la route et ralentira donc beaucoup moins que si l'accident se produisait de plein jour.

De tels résultats nous ont permis de voir que ce travail était réellement utile dans l'interprétation des causes des accidents et qu'ils pouvaient également être très utiles aux experts du domaine afin de réduire le risque d'avoir des accidents graves.

PARTIE 3 : Limites et défauts de cette étude

3.1 Exploitation insuffisante des données du conducteur

Dans beaucoup d'accidents les données concernant le conducteur n'étaient pas répertoriées. En effet, une caractéristique décrivait si celui-ci était sous l'effet d'une drogue, mais la plupart des champs était mis à « not reported », nous voulions supprimer les lignes où c'était le cas mais cette suppression enlevait énormément de données et nous avons jugé qu'il était mieux de supprimer complètement cette caractéristique. Nous étions certains que cette caractéristique aurait pu être crucial dans la prédiction d'accident de classe 4, et que nos modèles auraient pu bénéficier d'une meilleure précision.

3.2 *Skewed Data*

Les accidents de classe 4 sont très peu nombreux dans notre base de données (seulement environ 1.5% de la base de données). Nous faisons face à un problème connu du machine learning, celui de « *skewed data* » qui correspond au fait que nous avons une classe dont le nombre de données est trop peu nombreux face aux autres classes. Nous pensons donc que c'est ce qui explique les très mauvaises précisions que nos algorithmes donnent dans le cas de prédiction d'accidents de classe 4.

Conclusion

Nos analyses ont permis de conclure sur les caractéristiques rendant un accident plus ou moins dangereux, et également d'avoir un model prédisant sa dangerosité.

Nous avons donc pu avoir des réponses à nos problématiques, cependant elles ne sont pas aussi bonnes que nous l'espérions. En effet d'une part, certaines caractéristiques ne sont pas prises en compte et qui pourraient être importante à cause d'un manque de données et d'autre part les modèles de classification donnent des précisions assez basses.

Ce projet nous a permis de nous rendre compte que la tâche de recherche et de nettoyage des données est très répétitive et chronophage, nous aurions aimé passer moins de temps sur ces étapes, et plus sur des améliorations des réponses que nous avons apportées.

Ariel GABAI

Sami CONTESENNE