Le 21 avril 2024 Exploration de données industrielles IND 6212

Polytechnique Montréal **Yassine Azrou** 2265579

**David De Blas** 2003038

**Yannis Redjah** 1748777

# PREDICTION OF THE NUMBER OF ROAD ACCIDENT VICTIMS IN QUEBEC

**INTRODUCTION**

Both summer and winter present challenges for drivers in Quebec, requiring heightened caution on the roads. Despite an overall reduction in road accidents in 2022 compared to the previous year, there has been a 13.2% increase in fatal accidents compared to the average from 2017 to 2021.

Using datasets available on the Données Québec website, we selected a dataset provided by the Société de l'assurance automobile du Québec (SAAQ), which is openly accessible. The most recent dataset, corresponding to accident reports from 2022, serves as the basis for this analysis.

This report aims to analyze the dataset and explore the variables it contains. Following this, we will employ a statistical model to predict the number of victims based on explanatory variables. Finally, the results of the model will be evaluated to assess its predictive capabilities, and we will discuss the limitations of the approach.

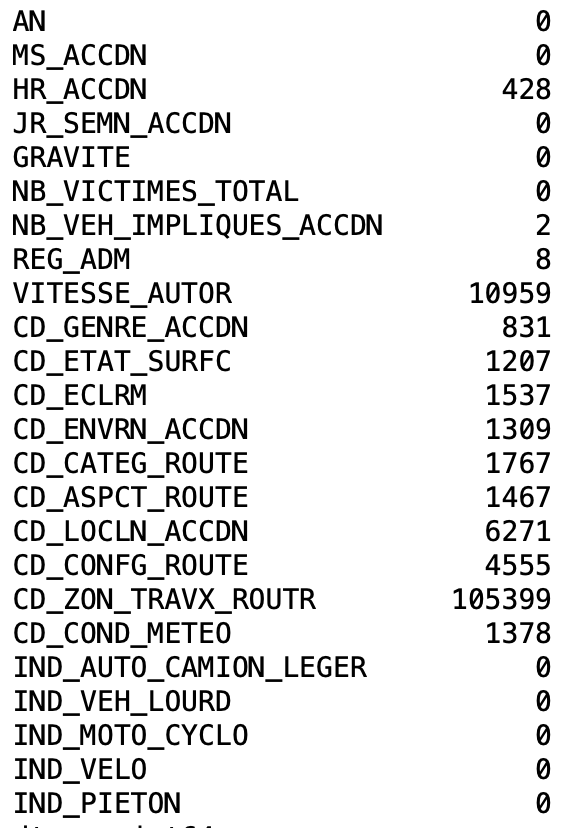
# DESCRIPTION DES DONNÉES

The dataset contains 108,185 accident records, each with 25 variables describing the incident, such as the year of occurrence, time of day, location details, severity, and the number of victims involved.

***Table 1: Variables Description***

|  |  |  |
| --- | --- | --- |
| **Nom de la variable** | **Description** | **Type** |
| AN | Year of the accident | Numeric |
| NO\_SEQ\_COLL | Unique accident identifier | Categorical |
| MS\_ACCDN | Month of the accident (coded 1 to 12) | Categorical |
| HR\_ACCDN | Four-hour interval of accident occurrence | Categorical |
| JR\_SEMN\_ACCDN | Day of the week (weekday or weekend) | Categorical |
| GRAVITE | Severity of the accident (fatal, serious, or property damage) | Categorical |
| NB\_VICTIMES\_TOTAL | Total number of victims | Numeric |
| NB\_VEH\_IMPLIQUES\_ACCDN | Number of vehicles involved | Numeric |
| REG\_ADM | Administrative region of Quebec | Categorical |
| VITESSE\_AUTOR | Speed limit at the accident location | Categorical |
| CD\_GENRE\_ACCDN | Collision type (vehicle, pedestrian, cyclist, etc.) | Categorical |
| CD\_ETAT\_SURFC | Road surface condition (dry, wet, etc.) | Categorical |
| CD\_ECLRM | Lighting conditions during the accident | Categorical |
| CD\_ENVRN\_ACCDN | Surrounding environment type (school, residential, etc.) | Categorical |
| CD\_CATEG\_ROUTE | Road aspect at the site (straight, curved, etc.) | Categorical |
| CD\_ASPCT\_ROUTE | Longitudinal position of the accident | Categorical |
| CD\_LOCLN\_ACCDN | Position longitudinale de l'accident (intersection, entre intersections, etc.) | Categorical |
| CD\_CONFG\_ROUTE | Road configuration | Categorical |
| CD\_ZON\_TRAVX\_ROUTR | Construction zone presence | Categorical |
| CD\_COND\_METEO | Weather conditions | Categorical |
| IND\_AUTO\_CAMION\_LEGER | Presence of a light vehicle or truck (Yes/No) | Binary |
| IND\_VEH\_LOURD | Presence of a heavy vehicle (Yes/No) | Binary |
| IND\_MOTO\_CYCLO | Presence of a motorcycle or moped (Yes/No) | Binary |
| IND\_VELO | Presence of a bicycle (Yes/No) | Binary |
| IND\_PIETON | Presence of a pedestrian (Yes/No) | Binary |

# Data Preparation and Formatting



Avant de débuter l’analyse des données, il est important de préparer et formater notre base de données pour faciliter son traitement. Nous avons tout d’abord supprimé toutes les informations contenues dans les colonnes **AN** et **NO\_SEQ\_COLL** puisqu’elles n'apportent aucune valeur ajoutée à notre modèle prédictif. Nous avons également constaté que pour la variable **CD\_ZON\_TRAVX\_ROUTR**, il y a énormément de données manquantes. Bien que les zones de travaux routiers aient un impact sur la fluidité du trafic routier, elles peuvent également influencer d’autres variables présentes dans notre base de données telle que **VITESSE\_AUTOR**. Puisque plusieurs éléments entrent en jeu et par souci de simplicité, nous avons décidé de retirer cette variable de l’équation. D’ailleurs, certaines variables comme **VITESSE\_AUTOR** présentent des données manquantes pour certains accidents. Par souci de précision, nous avons pris la décision de supprimer toutes les lignes contenant des données manquantes. Cette opération a fait en sorte de baisser la population de notre échantillon de 18%. Bien que cette baisse soit significative, nous disposons d’un volume de données suffisamment grand pour construire notre modèle.

Afin que notre base de données soit uniforme, il est important que toutes nos variables suivent un même format. Nous avons donc transformé toutes nos variables catégorielles en variable numérique de manière à ce que notre modèle puisse les interpréter correctement. Pour ce faire, nous regroupons toutes nos variables catégorielles et nous appliquons sur chacun d’eux l’encodage *one-hot encoding*, qui génère une colonne binaire pour chaque valeur unique d'une variable donnée. Par exemple, la variable **CD\_GENRE\_ACCDN**, qui catégorise le type d'accident, illustre parfaitement la nécessité de cette transformation. Par exemple, une collision avec un objet inerte n’a pas le même effet sur la perte humaine ou matérielle qu’une collision multiple avec d’autres véhicules. L'encodage *one-hot* va nous permettre de traiter ces catégories de manière indépendante et équitable dans notre analyse. Pour les variables binaires, nous effectuons une transformation en variable booléenne afin que notre modèle les traite correctement (au lieu de représenter les états par des strings ‘oui’ ou ‘non’, on le représente maintenant par 1 ou 0).

Après avoir uniformisé le format de nos données, nous allons établir un ensemble d'entraînement et un ensemble de tests. Nous utilisons une répartition classique, soit 80% des données à l'entraînement et 20% pour les tests. L'ensemble d'entraînement servira à l’apprentissage, soit identifier les tendances et les corrélations inhérentes à nos données. L'ensemble test, quant à lui, servira à évaluer le résultat de notre modèle par rapport aux données. Cette étape est cruciale, car elle permet de voir si notre modèle est adéquat et nous assure que les performances ne sont pas le fruit d’un surapprentissage. En bref, nous nous assurerons d'identifier avec fiabilité la valeur prédictive du modèle et, si nécessaire, procéder aux ajustements pour améliorer sa précision.

# MODÈLE PRÉDICTIF : LES FORÊTS ALÉATOIRES

Nous avons opté pour le modèle prédictif des forêts aléatoires pour prédire le nombre de victimes par accidents de la route au Québec. Le choix du modèle a été basé sur la nature et le type de données présents dans notre jeu de données, soit des variables numériques et catégorielles. Les forêts aléatoires sont un excellent modèle pour traiter cette structure de données. De plus, cette méthode est particulièrement efficace pour analyser des jeux de données de grande taille, tant en nombre de variables qu'en volume d'observations, ce qui correspond à la configuration de notre ensemble de données. Les avantages des forêts aléatoires ne s'arrêtent pas à leur capacité à gérer divers types de données. Elles sont également appréciées pour leur rapidité d'entraînement et leur efficacité à exécuter des calculs en parallèle, ce qui est d'autant plus avantageux lorsqu'il s'agit d'une recherche exhaustive d'hyperparamètres qui peut être potentiellement coûteuse en temps et ressource. De plus, les forêts aléatoires sont robustes face aux variables non pertinentes, au problème causé par la multicolinéarité et sont également moins sujettes au surajustement, même avec un grand nombre d'enregistrements. En somme, ces caractéristiques justifient notre choix de modèle afin de répondre à la problématique.

Afin de raffiner notre modèle, d’améliorer les indicateurs de performance et de réduire la probabilité de surajustement, nous avons eu recours à une procédure nommée *Grid Search*. Cette technique consiste à explorer les combinaisons d’une gamme étendue d'hyperparamètres afin de déterminer celle qui offre la meilleure précision. Cependant, cette procédure pourrait conduire à des conclusions trompeuses si elle est basée sur une seule séparation des données d'entraînement et de validation. Ainsi, nous avons décidé de l’associer à une *validation croisée*. Ceci permet un partitionnement des données en plusieurs sous-ensembles. Dans notre cas, on divise les données en 3 sous-ensembles (*fold)*. Le modèle est ensuite entraîné sur chaque combinaison possible de ces sous-ensembles puis testé. Les hyperparamètres étudiés correspondent à la profondeur maximale choisie par l’arbre (*max\_depth*), au nombre minimal d’échantillons requis avant de diviser un nœud (*min\_samples\_split*) et au nombre minimum d’échantillons requis dans une feuille (*min\_samples\_leaf*). Ici on essaie de trouver une combinaison minimisant le surajustement du modèle tout en améliorant au maximum les performances.

# ANALYSE DES RÉSULTATS

L'évaluation de notre modèle *Random Forest* sur les ensembles de données d'entraînement et de test a révélé des résultats encourageants.

Voici une synthèse des performances observées :

* **Erreur quadratique moyenne (MSE) sur l'entraînement** : 0.0543
* **Erreur quadratique moyenne (MSE) sur le test** : 0.0589
* **Score R² sur l'entraînement** : 0.8439
* **Score R² sur le test** : 0.8275

Ces indicateurs suggèrent que le modèle présente de bonnes prédictions, car les *MSE* sont relativement faibles. Ce constat est renforcé par les valeurs de R², ce qui indique que nos variables explicatives expliquent environ 80% de la variation de la variable de réponse.

Si l’on se penche sur les indicateurs de performance pour l’ensemble de tests, alors nous obtenons les résultats ci-dessous :

* **Erreur quadratique moyenne (MSE)** : 0.0589
* **Racine de l’erreur quadratique moyenne (RMSE)** : 0.2428
* **Erreur absolue moyenne (MAE)** : 0.0803
* **Score R²** : 0.8275

L'erreur RMSE, nous renseigne sur la déviation standard des résidus, soit la précision des prédictions du nombre de victimes par accident. Le RMSE est d’environ 0.25 sachant que les valeurs prédites peuvent varier de 0 à 3, cela indique que les prédictions du modèle peuvent varier en moyenne d’un quart de l’intervalle possible de valeurs. Le score R² de notre modèle est d'environ 83%, ce qui signifie que nos prédictions correspondent assez bien à ce qui se passe réellement. En clair, 83% des variations que nous observons dans les résultats réels peuvent être expliquées par notre modèle. Ces résultats sont satisfaisants quant à la précision générale du modèle utilisé pour prédire le nombre de victimes sur les routes du Québec.

Comparons maintenant ces résultats à ceux d'un modèle de base qui prédit systématiquement la moyenne des victimes observées dans les données d'entraînement. Nous obtenons les résultats ci-dessous :

* **Valeur moyenne prédite** : 0.2898
* **Erreur quadratique moyenne (MSE)** : 0.3418
* **Racine de l’erreur quadratique moyenne (RMSE)** : 0.5847
* **Score R²** : 0.00034

L'analyse de ces indicateurs révèle que notre modèle *Random Forest* surpasse largement le modèle de base. Alors que le modèle de base ne fait guère mieux qu'une prédiction aléatoire. Notre modèle prédictif montre des valeurs de MSE et RMSE bien inférieures, suggérant des prédictions beaucoup plus précises et fiables. De plus, bien que les valeurs R² se situent entre 0 et 1, nous obtenons une valeur négative. Étant donné que la valeur de ce coefficient est proche de 0, nous pouvons affirmer que la performance du modèle de base est nettement inférieure à celle des forêts aléatoires.

# CRITIQUE DES RÉSULTATS ET LIMITATIONS

Les résultats de la modélisation sont prometteurs, avec un score R² avoisinant les 80%, ce qui signifie que le modèle parvient à expliquer une grande partie de la variabilité du nombre de victimes par les différents facteurs définis plus haut. Un RMSE proche de zéro suggère également que les prédictions du modèle sont généralement précises, avec peu d'écarts par rapport aux valeurs réelles. Cependant, bien que les résultats soient excellents, il est possible d’améliorer davantage le modèle. Actuellement, les hyperparamètres sélectionnés pour la recherche par grille ne constituent qu'une fraction des nombreuses possibilités existantes. L'élargissement de cette exploration d’hyperparamètre pourrait potentiellement renforcer la précision du modèle.

De plus, bien que le modèle des forêts aléatoires ait été choisi en raison de sa polyvalence et de sa robustesse, il faut reconnaître que ce modèle présente quelques désavantages. En effet, les forêts aléatoires peuvent être complexes et très difficiles à interpréter. D’ailleurs, cette complexité a un effet direct sur le temps de calcul pour obtenir des résultats. Il est aussi nécessaire de traiter l’information a priori et donc d’éliminer au maximum les données manquantes avant de travailler sur un modèle de forêt aléatoire. Les forêts aléatoires sont aussi très sensibles aux paramètres tels que le nombre d’arbres et la profondeur. Trouver un juste équilibre pourrait se révéler critique dans la performance de ce modèle.

D’autre part, il faut aussi noter qu’il se peut que l’utilisation d'autres méthodes comme les réseaux de neurones ou les modèles bayésiens aient pu générer de meilleurs résultats que ce qu’on a actuellement. Bien que ces solutions puissent nécessiter des temps d'entraînement plus longs ou un traitement différent de la base de données, elles auraient pu conduire à des améliorations de performance. Par exemple, concernant le traitement différent de la base de données, l’un des critères pour avoir des résultats satisfaisants avec ces 2 méthodes est une atténuation importante de la multicolinéarité entre les variables, car ces méthodes sont plus sensibles aux corrélations internes qui peuvent perturber le processus d'apprentissage et la validité des prédictions. En revanche, les forêts aléatoires se montrent bien plus robustes face à ce problème, principalement grâce à leur mécanisme de sélection aléatoire de variables lors de la création des arbres individuels.

D’autre part, l'ajout de variables explicatives supplémentaires, telles que l'âge des conducteurs, aurait pu enrichir notre modèle prédictif, car l'âge est un facteur souvent corrélé au risque d'accidents de la route. Les statistiques montrent que les jeunes conducteurs sont généralement plus impliqués dans les accidents que les autres groupes d'âge. Par conséquent, en intégrant l'âge comme variable, nous aurions pu potentiellement affiner nos prévisions et mieux comprendre les dynamiques sous-jacentes des accidents de la route.

# CONCLUSION

La modélisation du nombre de victimes sur les routes du Québec à l'aide des forêts aléatoires a permis d'obtenir des résultats prometteurs. Le modèle propose une précision satisfaisante avec une valeur de R² près de 80% et un RMSE assez peu élevé. En fin de compte, bien que les forêts aléatoires soient un excellent choix pour modéliser notre problématique, d'autres approches auraient pu se révéler plus pertinentes. Une recherche plus approfondie des hyperparamètres et l'incorporation de données supplémentaires pourraient également contribuer à améliorer la précision du modèle et à mieux comprendre les facteurs influençant le nombre de victimes sur les routes du Québec.