Polytechnique Montréal

**Projet de session**

*Prévision du nombre d’accidents sur les routes du Québec*

Travail présenté à

Monsieur Bruno Agard

Dans le cadre du cours de

Exploration de données industrielles IND 6212

Par

Yassine Azrou XXXXXXX

David De Blas 2003038

Yannis Redjah 1748777

Le 14 avril 2024

**INTRODUCTION**

Dans un contexte de sécurité routière, nous constatons que les conditions météorologiques sont de plus en plus extrêmes, notamment sur les routes du Québec. Été comme hiver, les usagers de la route doivent redoubler de prudence lors de leurs déplacements (Alam, 2022). D’ailleurs, bien que l’on ait enregistré une baisse des accidents sur la route en 2022 par rapport à l’année dernière, on déplore une hausse de 13.2% d’accidents mortels par rapport à la moyenne entre 2017 et 2021 (SAAQ, 2022).

À travers les données récoltées par les différents services de polices en 2022, nous avons une base de données qui répertorie la gravité des accidents et certains paramètres dont les conditions routières, les conditions météorologiques et l’aspect de la route. Cette base de données gouvernementale, publiée sur le site de *Données Québec,* sera un point de départ pour analyser et comprendre comment nous pouvons prédire le nombre de victimes selon différents paramètres externes que nous allons détailler dans les paragraphes à suivre.

À travers cette analyse, nous allons discuter de la nature des éléments présents dans notre base de données et de la façon dont nous les avons traités. Par la suite, nous serons en mesure d’utiliser un modèle statistique pour exploiter ces données et prédire la gravité de ces accidents selon différentes variables explicatives. Le choix des variables explicatives sera bien entendu détaillé et justifié. Puis, nous examinerons les résultats produits par notre modèle et effectuerons une analyse des résidus pour évaluer la qualité d’ajustement de notre modèle par rapport aux données recensées. Finalement, nous traiterons des limites de ce modèle par rapport à la problématique et aux résultats enregistrés.

**SOURCE DES DONNÉES**

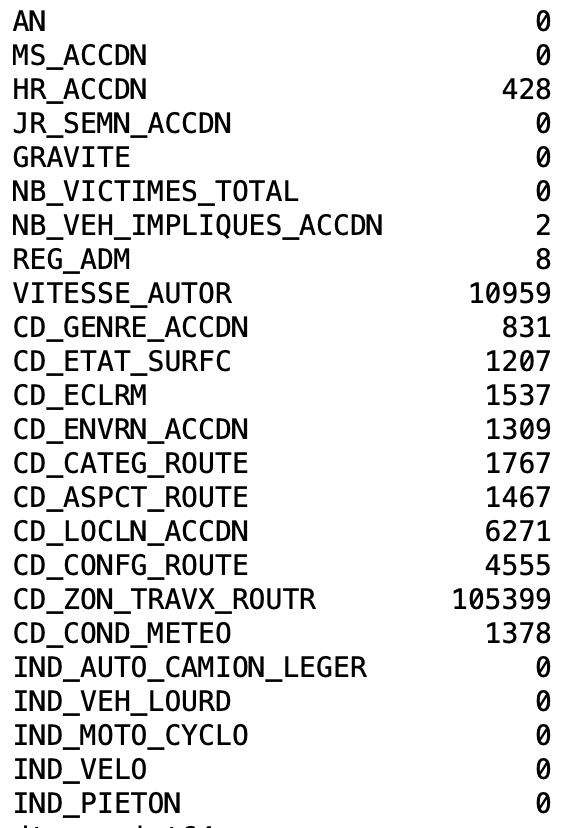
Pour mener à bien notre étude des incidents routiers, nous nous sommes appuyés sur les ressources mises à disposition par Données Québec. En particulier, nous avons sélectionné un ensemble de données fourni par la Société de l'assurance automobile du Québec (SAAQ), disponible en accès libre. Pour explorer plus avant ces informations, vous pouvez visiter la page dédiée : [Données Québec - Rapports d'accident](https://www.donneesquebec.ca/recherche/dataset/rapports-d-accident). Cette ressource regroupe des données exhaustives sur les accidents routiers survenus à travers le Québec. Nous avons extrait le fichier le plus actualisé, correspondant aux rapports d'accidents de l'année 2022. Ce fichier offre une vue détaillée et complète des accidents de la route recensés au Québec, ce qui fournit ainsi une base solide pour notre analyse.

**DESCRIPTION DES DONNÉES**

Notre jeu de données se compose de 108 185 enregistrements d’accidents, chacun détaillé par 25 variables explicatives. Ces variables capturent un large éventail d'aspects cruciaux relatifs à chaque incident, incluant l'année où cela s’est produit, le moment précis de la journée, ainsi que des informations spécifiques sur la localisation, la sévérité de l'événement et le nombre de personnes affectées. Pour une compréhension approfondie, les caractéristiques de chaque variable sont détaillées dans le tableau suivant :

Tableau 1: Description des variables explicatives de la base de données sur les accidents aux québec

| Nom de la variable | Description | Type |
| --- | --- | --- |
| AN | Année de l'accident | Numérique |
| NO\_SEQ\_COLL | Identifiant unique pour l'accident | Catégoriel |
| MS\_ACCDN | Mois de l'accident codé de 1 à 12 | Catégoriel |
| HR\_ACCDN | Intervalle de quatre heures contenant l'heure de l’accident | Catégoriel |
| JR\_SEMN\_ACCDN | Jour de la semaine de l'accident (SEM ou FDS) | Catégoriel |
| GRAVITE | Gravité de l'accident, de mortel ou grave à dommages matériels seulement | Catégoriel |
| NB\_VICTIMES\_TOTAL | Nombre total de victimes dans l'accident | Numérique |
| NB\_VEH\_IMPLIQUES\_ACCDN | Nombre de véhicules impliqués dans l'accident | Numérique |
| REG\_ADM | Région administrative du Québec où l'accident a eu lieu | Catégoriel |
| VITESSE\_AUTOR | Vitesse autorisée sur la portion de route où l'accident a eu lieu | Catégoriel |
| CD\_GENRE\_ACCDN | Type de collision ou d'accident (véhicule, piéton, cycliste, etc.) | Catégoriel |
| CD\_ETAT\_SURFC | État de la surface de roulement lors de l'accident (sèche, mouillée, etc.) | Catégoriel |
| CD\_ECLRM | Niveau de luminosité au moment de l'accident | Catégoriel |
| CD\_ENVRN\_ACCDN | Type de zone environnante où l'accident s'est produit (scolaire, résidentiel, commercial, etc.) | Catégoriel |
| CD\_CATEG\_ROUTE | Type de route où l'accident s'est produit (chemin public, hors chemin public, etc.) | Catégoriel |
| CD\_ASPCT\_ROUTE | Aspect de la route au lieu de l'accident (droit, courbe, etc.) | Catégoriel |
| CD\_LOCLN\_ACCDN | Position longitudinale de l'accident (intersection, entre intersections, etc.) | Catégoriel |
| CD\_CONFG\_ROUTE | Caractéristiques des voies de circulation où l'accident a eu lieu | Catégoriel |
| CD\_ZON\_TRAVX\_ROUTR | Présence d'une zone de travaux au moment de l'accident | Catégoriel |
| CD\_COND\_METEO | Conditions météorologiques au moment de l'accident | Catégoriel |
| IND\_AUTO\_CAMION\_LEGER | Présence d'un véhicule léger ou d'un camion (Oui ou Non) | Binaire |
| IND\_VEH\_LOURD | Présence d'un véhicule lourd (Oui ou Non) | Binaire |
| IND\_MOTO\_CYCLO | Présence d'une moto ou d'un cyclomoteur (Oui ou Non) | Binaire |
| IND\_VELO | Présence d'un vélo (Oui ou Non) | Binaire |
| IND\_PIETON | Présence d'un piéton (Oui ou Non) | Binaire |

**TRAITEMENT DES DONNÉES/PRÉPARATION DES DONNÉES/FORMATAGE DES DONNÉES**

Avant d’aller plus loin dans l’analyse des données, il est important de préparer et formater notre base de données pour faciliter son traitement. Comme indiqué plus tôt, nous avons environ 25 variables explicatives pour plus de 108 000 enregistrements. Parmi ces nombreuses caractéristiques, plusieurs ne seront pas utiles pour déterminer le nombre de victimes. Dans un premier temps, **nous avons omis d’inclure les colonnes AN et NO\_SEQ\_COLL** puisqu’elles n'apportent aucune valeur ajoutée dans le modèle que nous tentons de construire. Nous avons également observé qu’au niveau de la variable CD\_ZON\_TRAVX\_ROUTR, il y a énormément de données manquantes. Plusieurs hypothèses peuvent être établies à ce sujet, comme le fait que ces zones de travaux peuvent varier dans le temps et dans l’espace. Les zones de travaux peuvent également avoir un impact temporaire sur la vitesse autorisée dans les alentours. Puisque plusieurs éléments entrent en jeu et par soucis de simplicité, nous avons donc décidé de retirer cette variable de l’équation. De plus, certaines variables comme VITESSE\_AUTOR présentent des données pour lesquelles nous n’avons aucune information. Pour que notre modèle donne de meilleurs résultats, nous avons pris la décision de **supprimer tous les enregistrements contenant des données manquantes**. Ainsi, le nombre d’enregistrements de notre base de données se voit diminuer d’environ 18%. Bien que cette baisse soit significative, nous disposons d’un grand volume de données ce qui nous permet tout de même de préparer notre modèle adéquatement.

Maintenant pour bien faire notre prédiction, il faut faire en sorte que toutes nos variables soit dans le même format numérique. En effet, un arbre de décision ne peut pas traiter directement des données textuelles ou des catégories pour effectuer ces calculs ou ces comparaisons. Par conséquent, il est important de transformer toutes nos variables catégorielles en nombre de manière à 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 crée une nouvelle colonne binaire pour chaque catégorie unique d'une variable. Par exemple, la variable CD\_GENRE\_ACCDN, qui catégorise le type d'accident, illustre parfaitement la nécessité de l'encodage one-hot. Chaque catégorie, allant de collisions avec des véhicules routiers à d'autres objets fixes ou à des incidents sans collision, reflète une nature différente d'accident avec un impact potentiellement différent sur les résultats, comme la gravité des dommages ou le nombre de victimes. L'encodage one-hot va nous permettre de traiter ces catégories de manière indépendante et équitable dans notre analyse.

Pour nos variables binaires, il faut juste **transformer les indicateurs binaires ‘O’ et ‘N’ en 1 et 0**, de cette manière le modèle devrait bien comprendre

Après avoir uniformisé le format de nos données, l'étape suivante consiste à les répartir en deux ensembles distincts : l'ensemble d'entraînement et l'ensemble de test. Ici nous utilisons une répartition classique, nous affectons 80% des données à l'entraînement et réservons les 20% restants pour les tests. L'ensemble d'entraînement sert de matière première sur laquelle le modèle apprendra à identifier les tendances et les corrélations inhérentes à nos données. L'ensemble de test, quant à lui, permet de confirmer si le modèle est efficace en évaluant l'efficacité de notre modèle sur des données qu'il n'a pas eu l'occasion de voir durant sa phase d'apprentissage. Cette évaluation est très importante, car elle permet de voir si notre modèle possède une bonne capacité de généralisation et cela nous permet aussi de nous assurer que ses performances ne sont pas simplement le résultat d'une mémorisation spécifique des données d'entraînement. Grâce à ce processus, nous pouvons identifier avec fiabilité la valeur prédictive réelle du modèle et, si nécessaire, procéder aux ajustements pour améliorer sa précision.

**CHOIX DE L’OUTILS DE PRÉDICTION, DES PARAMÈTRES OUTILS ET JUSTIFICATION**

Nous avons opté pour l'approche des forêts aléatoires dans notre objectif de prédiction du nombre de victimes résultant des accidents de la route au Québec. Ce choix a été guidé principalement par la diversité des types de données présentes dans notre jeu de données, combinant des variables numériques et catégorielles. Les forêts aléatoires excellent dans le traitement de telles structures de données variées. De plus, cette méthode est particulièrement efficace pour analyser des jeux de données de grande envergure, tant en nombre de variables qu'en volume d'observations, ce qui correspond à la configuration de notre ensemble de données. Les atouts des forêts aléatoires ne se limitent pas à leur rapidité d'entraînement et à leur capacité à être parallélisées, ce qui est d'autant plus avantageux lorsqu'il s'agit d'une recherche exhaustive d'hyperparamètres qui peut être potentiellement coûteux en temps. De plus, les forêts aléatoires sont robustes face aux variables non pertinentes et sont également moins sujettes au surajustement, même avec un grand nombre d'enregistrements, ce qui renforce ainsi la capacité de généralisation du modèle.. En somme, ces caractéristiques positionnent les forêts aléatoires comme un choix privilégié pour atteindre de bonnes performances générales dans notre étude.

Pour affiner notre modèle prédictif, améliorer les indicateurs de performance et réduire la probabilité de surajustement, nous avons recours à une procédure nommée **Grid Search**. Cette technique consiste à explorer en toutes les combinaisons d’une gamme étendue d'hyperparamètres, en testant chaque combinaison possible afin de déterminer celle qui confère à notre modèle la meilleure précision. Le Grid Search seule, bien qu'utile, pourrait conduire à des conclusions trompeuses s' il est basé sur une seule séparation des données d'entraînement et de test. Ainsi pour renforcer la fiabilité de notre évaluation, nous avons décidé de l’associer à une **validation croisée**. Ceci permet un partitionnement des données en plusieurs sous-ensembles (ici on divise les données en 3 sous-ensemble, ou autrement dit 3 folds). Le modèle est ensuite entraîné sur chaque combinaison possible de ces sous-ensembles et testé sur le sous-ensemble restant, ce qui permet de s'assurer que toutes les données contribuent à la fois à l'entraînement et à la validation du modèle. En combinant ces 2 méthodes, nous pouvons juger de manière plus fiable des performances du modèle car on élimine d’une part les biais associés à entraîner notre modèle que sur un ensemble de test, ce qui permet de ne pas choisir les hyperparamètres qui marchent seulement sur cette partie des données et on choisit d’autre part la meilleur combinaison d’hyperparamètres que l’on veut explorer. Les hyperparamètres étudiés concerne la profondeur maximale choisies par l’arbre (max\_depth), le nombre minimal d’échantillon requis avant de diviser un noeud (min\_samples\_split) et le nombre minimum d’échantillons requis dans une feuille (min\_samples\_leaf). Ici on essaie de trouver une combinaison qui apporte un équilibre en surajustement et sous-ajustement.

**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 une bonne capacité de généralisation, les scores MSE étant relativement similaires sur les deux ensembles. Ce constat est renforcé par les valeurs de R², avec des valeurs tout aussi similaires et élevées, ce qui indique une précision élevée du modèle dans ses prédictions.

Si l’on s’intéresse plus en détail sur les indicateurs de performance pour l’ensemble de test, alors on obtient ceci:

* **Erreur quadratique moyenne (MSE)** : 0.0589
* **Erreur quadratique moyenne racine (RMSE)** : 0.2428
* **Erreur absolue moyenne (MAE)** : 0.0803
* **Score R²** : 0.8275

L'erreur RMSE, nous renseigne sur la déviation standard des résidus, offrant une idée de la précision des prédictions en termes de victimes par accident. Ici le RMSE est d’environ 0.25 sachant que les valeurs prédites peuvent varier de 0 à 3, cela indique que les prédictions du modèles 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, 80% des variations que nous observons dans les résultats réels peuvent être expliquées par notre modèle. C'est un bon signe que notre modèle donne des prédictions fiables.

Comparons maintenant ces résultats à ceux d'un modèle de base, qui prédit systématiquement la moyenne des victimes observées dans les données d'entraînement. Le modèle de base montre les performances suivantes sur l'ensemble de test :

* **Valeur moyenne prédite** : 0.2898
* **Erreur quadratique moyenne (MSE)** : 0.3418
* **Erreur quadratique moyenne racine (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 (R² étant pratiquement nul), notre modèle prédictif montre des valeurs de MSE et RMSE bien inférieures, suggérant des prédictions beaucoup plus précises et fiables.

**CRITIQUE RÉSULTATS *VS* OUTILS *VS* PROBLÉMATIQUE + 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 des données. 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, ces indicateurs, bien que positifs, laissent entrevoir une opportunité d'affiner davantage le modèle. Actuellement, les hyperparamètres sélectionnés pour la recherche par grille ne constituent qu'une fraction des nombreuses possibilités. L'élargissement de cette exploration hyper paramétrique pourrait notablement renforcer la précision du modèle.

De plus, bien que les forêts aléatoires aient été le choix de prédilection pour ce modèle en raison de leur adaptabilité et de leur robustesse, il faut reconnaître que d'autres méthodes comme les réseaux de neurones ou les modèles bayésiens auraient pu être considérées. Bien que ces alternatives puissent nécessiter des temps d'entraînement plus longs, elles auraient pu potentiellement conduire à des améliorations de performance.

Aussi on peut se faire une réflexion sur la pertinence des variables explicatives utilisées. Par exemple, si l’on avait aussi eu des variables explicatives telles que l'âge des conducteurs, cela aurait pu se révéler être des prédicteurs significatifs des résultats des accidents, étant donné l'association reconnue entre l'âge et la probabilité d'accidents de la route (il est connu que les jeunes conducteurs font plus d’accidents que les autres catégories de conducteurs).

Malgré le fait que des améliorations sont envisageables, il ne faut pas oublier de souligner que notre modèle actuel performe de manière satisfaisante dans la prédiction du nombre de victimes d'accidents de la route selon les différents paramètres de la route. Dans l'ensemble, les résultats obtenus avec le modèle actuel sont solides, les prédictions apportées constituent dans la grande majorité une base fiable, que l’on pourrait utiliser afin de prévoir de telles situations et pouvoir agir en conséquence.

**CONCLUSION**