UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

ANALYSE ET PRÉDICTION D'ACCIDENTS DE LA ROUTE DANS LA VILLE D'OTTAWA

MÉMOIRE PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DU PROGRAMME DE MAÎTRISE EN SCIENCES ET TECHNOLOGIES DE

L'INFORMATION

PAR

Aboubacar Sékou Traoré

Mai 2018

Jury d'évaluation

Président du Jury :	Dr. Marek Zaremba
Membre du Jury :	Dr. Rokia Missaoui
Directeur de recherche :	Dr. Ana-Maria Cretu

Dédicace

Je dédie ce mémoire à mes très chers parents pour avoir façonné ma personnalité dans la pure tradition de Dia* et pour leur appui multiforme qui me donne le privilège d'être devant vous ici et maintenant :

- Mon très cher père, Monsieur Sékou TRAORE, qui a toujours cru en moi et a mis à ma disposition tous les moyens nécessaires pour que je réussisse dans mes études.

- Ma très chère mère, Aïssata TRAORE, pour m'avoir tout donné. Elle a fait de moi l'homme comblé que je suis aujourd'hui.

Que ce modeste travail soit le témoignage de ma profonde affection et de mon estime la plus sincère. Que Dieu le Tout Puissant leur accorde santé et longue vie. Puisse ce travail intellectuel contribuer à préserver des vies humaines sur les routes d'ici et d'ailleurs.

**** Dia : Mon village d'origine où sont nés mes parents. Situé au centre du Mali, c'est une cité historique de savoir et de culture.***

Remerciements

Mes sincères remerciements vont tout d'abord au Tout Puissant, qui, par sa grâce, m'a donné l'opportunité de mener à bien ce travail.

Mes vifs remerciements vont ensuite à l'endroit de mes professeurs et encadreurs de l'UQO pour la qualité de l'enseignement et du suivi dont j'ai bénéficié durant toutes ces années académiques et pour la passion des TIC qu'ils m'ont transmise.

À la Direction de l'Université pour les facilités multiformes qui m'ont été accordées.

À ma Directrice de recherche Ana-Maria Cretu qui, par son ouverture d'esprit, son sens du leadership et son expertise dans les STI a su m'accompagner dans la réalisation de cette œuvre scientifique.

À mon père Sékou TRAORE et ma mère Aïssata Traoré pour tous les sacrifices consentis durant tout mon cheminement académique.

À mon tonton Youssouf Coulibaly pour le temps consacré à la relecture du présent document.

À la famille de Patrice et Solange Tadonki pour son soutien inestimable dès mon premier jour au Canada.

Au Service de police d'Ottawa pour m'avoir donné accès aux données dont il dispose.

À tous mes proches, mes frères et sœurs, mes amis et toutes les personnes qui ont contribué de près ou de loin à la réalisation de ma réussite scolaire.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations, sigles et acronymes

ABRÉVIATIONS, SIGLES ET ACRONYMES	DENOMINATIONS
AADT	Moyenne annuelle de la circulation journalière (*AverageAnnual Daily Traffic*)
AUC	Zone sous la courbe (*Area Under the Curve*)
CMF	Facteurs de modification d'accident (*Crash Modification Factor*)
DSi	Morts et blessures graves (*Death and Severeinjury*)
EPDO	Equivalence seulement aux dommages matériels(*Equivalent Property Damage Only*)
HSM	Manuel de sécurité sur les autoroutes (*HighwaySafetyManual*)
HRI	Intersection à haut risque (*High-Risk Intersections*)
LOSS	Niveau de service de sécurité (*Level of Safety Service*)
OMS	Organisation Mondiale de la Santé
RRI	Indice de risques routiers (*Road Risk Indices*)
SIIG	Guide d'information sur les intersections signalisées (*Signalized Intersections Informational Guide*)
SPF	Fonction de performance de sécurité (*Safety Performance Functions*)
STI	Scienceset Technologies de l'Information
SVM	Machine à vecteurs de support (*Support Vector Machine*)
TIC	Technologiesde l'Information et de la Communication

Résumé

Le phénomène des accidents routiers est une problématique de portée mondiale. Nous pouvons dénombrer des millions de victimes à travers le monde. Parmi ces victimes, en termes de vies humaines, nous constatons des décès, des blessures graves, des traumatismes psychologiques de divers degrés au point que certains en ressortent handicapés à vie. De par les dommages aux personnes physiques, les accidents de la route peuvent avoir aussi un impact socio-économique très important.

Le but de ce mémoire est d'analyser les données sur les accidents de la route survenus dans la ville d'Ottawa pendant les années 2013 à 2016 afin de pouvoir les prédire. Les relations entre les variables caractérisant les accidents seront analysées et visualisées afin d'identifier leurs possibles liens et corrélations. Divers modèles seront ensuite construits permettant d'analyser et de prendre des décisions valables et fournir des prédictions sur les accidents. Dans ce contexte, nous nous sommes attachés en premier lieu à prédire le risque des accidents dans des conditions données (par exemple accident/pas d'accident), et en deuxième lieu à prédire le risque en termes de type d'accidents (par exemple accidents fatals, accidents avec des blessures ou accidents avec des dommages matériels). Pour arriver à un tel résultat, nous nous proposons d'analyser et de classifier les accidents de la route en utilisant une série de techniques existantes, à savoir les arbres de décision, les machines à vecteurs de support, les réseaux de neurones, l'algorithme AdaBoost, les arbres de décision « gradient boosted tree », l'algorithme naïf bayésien et la méthode du k-voisins le plus proche (*k-nearest neighbors*).

Étant donné le fait que les deux bases de données utilisées dans ce mémoire sont déséquilibrées, c'est-à-dire qu'on a par exemple dans une des bases de données seulement 71 cas d'accidents fatals contre 43 000 cas d'accidents non fatals, nous avons fait appel à la technique de sur-échantillonnage synthétique de la minorité (SMOTE). Celle-ci ajoute des échantillons synthétiques à la classe minoritaire en tenant compte des données voisines, nous offrant ainsi une solution à ce problème de déséquilibre.

Une évaluation de la performance de chaque modèle sur diverses combinaisons des variables (par exemple données météorologiques, jour de la semaine, heure, lieu de l'accident, etc.) sera proposée afin d'identifier les modèles les plus prometteurs pour la prédiction des accidents. Une analyse de ces variables et de leur importance sera également incluse dans le présent mémoire. Les travaux de recherche pour l'analyse et la classification des données sont basés sur l'outil R et plus précisément la librairie « Rattle », ainsi que sur le logiciel Rapid Miner.

Le présent mémoire apporte donc une contribution dans le contexte des sciences et technologies de l'information afin de rendre la circulation routière plus sécurisée, notamment dans la ville d'Ottawa.

Abstract

The phenomenon of road accidents is a worldwide problem. We can count millions of victims around the world. Among those victims, in terms of human lives, we can observe deaths, serious injuries, and psychological trauma of varying degrees to the point that some victims become disabled for life. By the damages incurred to persons, road accidents can also have a very important socio-economic impact.

The purpose of this thesis is to analyze traffic accident data for the City of Ottawa during the years 2013 to 2016 in order to enable the prediction of accidents in given conditions. The relationships between the variables characterizing the accidents are first analyzed and visualized to identify their possible links and correlations. Various models are then constructed to analyze, make valid decisions and provide predictions on accident occurrence. In this context, we focused first of all on predicting the risk of accidents under given conditions (for example accident/no accident), and secondly to predict the risk in terms of accident types (for example fatal accidents, accidents with injuries or accidents with property damage). For this purpose, we propose to analyze and classify road accidents using a series of existing intelligent techniques, namely decision trees, support vector machines, neural networks, the AdaBoost algorithm, gradient boosted trees, the Naive Bayes algorithm and the k-nearest neighbors respectively.

Given the fact that the two datasets used for testing are unbalanced, i.e. only 71 fatal accidents cases and 43 000 non-fatal accident cases, we use the Synthetic Minority Oversampling Technique (SMOTE). This technique adds synthetic samples to the minority class by taking into account the neighborhood of existing data and thus allow us to address the issue of unbalanced data.

An evaluation of the performance of each model using various combinations of variables (e.g. weather, day of the week, time of day, accident location, etc.) is executed in this work. An analysis of these variables and their significance is also included. The research work on the analysis and the classification of the data is based on R language, and more precisely on its library "Rattle", as well as on the Rapid Miner software.

This thesis brings a contribution in the context of information science and technology to make road traffic safer, particularly in the City of Ottawa.

CHAPITRE 1 : INTRODUCTION

1.1. Problématique de la sécurité routière

De nos jours, on assiste à une croissance fulgurante du nombre de véhicules automobiles en circulation, et comme conséquence, à une augmentation continuelle du nombre d'accidents de la route, malgré les efforts consentis pour la réalisation d'infrastructures modernes répondant aux normes internationales. Ils représentent maintenant l'une des premières causes de mortalité dans le monde [1]. Le problème n'est pas particulier à un pays donné, mais ce sont les pays du monde entier qui subissent ce phénomène qui ne saurait être une fatalité.

Au cours de la dernière décennie, la problématique des accidents de la route a pris un caractère mondial. Cela est attesté dans l'aide-mémoire de l'Organisation Mondiale de la Santé (OMS) sur les accidents de la route en mai 2017 [1]:

- Environ 1,25 million de décès par an ;
- Première cause de décès chez les jeunes âgés de 15 à 29 ans ;
- 90% des décès sur les routes surviennent dans les pays à revenus faible ou intermédiaire qui possèdent environ 54% du parc mondial de véhicules ;
- Près de la moitié des personnes tuées sur les routes sont des «usagers vulnérables» (piétons, cyclistes et motocyclistes) ;
- Sans une action soutenue, les accidents de la route deviendront, selon les projections, la septième cause de mortalité d'ici à 2030 ;
- Le Programme de développement durable à l'horizon 2030 des Nations Unies a fixé un objectif ambitieux pour la sécurité routière, à savoir diminuer de moitié le nombre total des morts et des blessés dus aux accidents de la route d'ici à 2020.

La conséquence des accidents de la route ne se résume pas seulement à la perte de vies humaines, elle a aussi un impact économique considérable pour les proches des victimes et les pays concernés. Le traitement des victimes nécessite très souvent des montants importants et le coût de réparation des deniers publics endommagés peut coûter à certains gouvernements jusqu'à 5% de leur produit national brut [1].

La ville d'Ottawa, dont les données de collisions enregistrées de 2011 à 2016 sont présentées à la figure 1, n'est pas en marge de ce fléau. On peut constater sur la figure que le nombre de collisions est assez variable d'une année à l'autre, 2016 comptant le nombre le moins élevé.

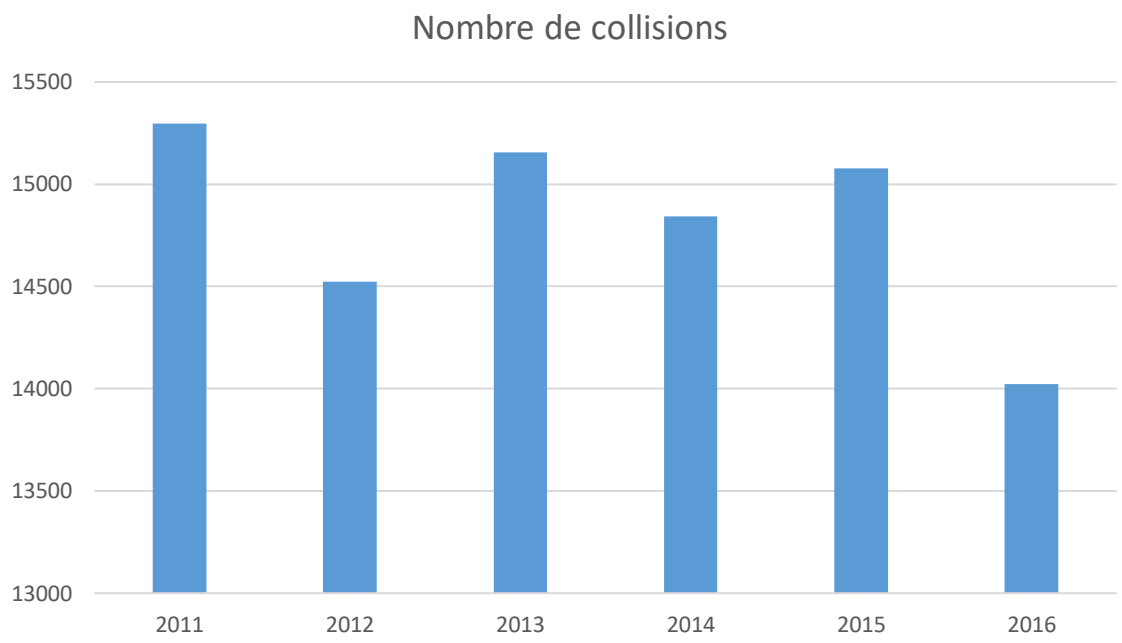


Fig.1 : Collisions à Ottawa de 2011 à 2016 (source de données [2])

Les recherches récentes sur les accidents de la circulation et leurs causes tendent à identifier les collisions de véhicules comme des «événements prévisibles et évitables» [3] pour lesquels nous pouvons identifier les causes et prendre des mesures pour les éviter» [4]. Les systèmes experts dédiés à la prédiction des accidents de la route devraient d'abord être en mesure de traiter des ensembles de données provenant de diverses sources et de divers types (par exemple données météorologiques, jour de la semaine, heure de la journée, les intersections et sous-divisions géographiques de la ville où l'accident a eu lieu, etc.). Les variables appropriées devront être sélectionnées pour faciliter cette capacité. Au-delà de la détection des tendances spécifiques des données, ces systèmes devront également produire des prédictions en dépit de données manquantes, corrompues et bruyantes.

À cause de leur capacité à traiter ces types de données, on s'attend à ce que des techniques d'apprentissage (par exemple arbres de décision, réseaux de neurones, etc.) puissent offrir de bonnes performances sur des sous-ensembles de variables.

Le fait de voir ce fléau prospérer est donc la source principale de motivation pour la mise en place d'une solution informatique qui sera pour tous porteuse d'un nouvel espoir dans la sécurité routière. D'ici dérive l'intérêt dans la conception et le développement des techniques intelligentes capables d'identifier, sur la base de données existantes, les risques de collisions.

C'est dans cette logique de production scientifique que le thème d'analyse et prédiction d'accidents de la route dans la ville d'Ottawa a retenu notre attention.

Elle apportera une contribution dans le domaine des sciences et technologies de l'information qui s'invitent dans tous les domaines de la vie, surtout, lorsque les mesures et les lois des gouvernements ne donnent pas toujours un impact suffisamment satisfaisant dans les régulations des accidents routiers.

Les principaux domaines traités et les résultats obtenus dans le cadre de la présente recherche sont présentés de sorte à donner un aperçu global et spécifique de l'évolution de la sécurité routière dans la ville d'Ottawa en s'articulant autour de sept parties distinctes et complémentaires. Le chapitre 1 est une partie introductive qui relate la problématique de la sécurité routière, les causes, les conséquences et les défis à relever. Il présente aussi les objectifs de ce mémoire de recherche visant à explorer les liens entre les données de trafic et l'occurrence d'accidents de la route dans la ville d'Ottawa au moyen d'outils de fouille de données. Le chapitre 2 porte sur l'état de l'art actuel en matière d'accidents de la route et des travaux effectués dans la prédiction de ceux-ci en intégrant diverses méthodologies, dont l'arbre de décision, le réseau de neurones, l'algorithme AdaBoost, les arbres de décision « *gradient boosted tree* », l'algorithme naïf bayésien, la méthode desk-voisins le plus proches (KNN) et les machines à vecteurs de support (*support vector machines, SVM*). Le chapitre 3 décrit les méthodologies utilisées, à savoir les différents algorithmes pour l'atteinte des objectifs fixés de la recherche. Le chapitre 4 présente les résultats obtenus. Le chapitre 5 porte sur les conclusions tirées des résultats obtenus.

Les travaux de recherche pour l'analyse et la classification des données sont basés sur l'outil R et plus précisément la librairie « Rattle », ainsi que sur le logiciel Rapid Miner.

1.2. Objectifs de la recherche

1.2.1. Objectif global de la recherche

L'objectif global de ce mémoire est d'explorer les liens entre les données de trafic, les données environnementales et l'occurrence d'accidents de la route au moyen d'outils de fouille de données. Ces liens devraient aider à analyser les principales causes des accidents de la route et apporter des facteurs d'amélioration aux mesures déjà en place. Une approche proactive devra donc être adoptée pour la gestion des accidents par une détection automatique des lieux et des conditions propices aux accidents.

1.2.2. Objectifs spécifiques de la recherche

Les objectifs spécifiques fixés par la présente recherche sont :

- Étudier l'impact de diverses variables sur l'occurrence d'accidents en utilisant des données fournies par le Service de police d'Ottawa pour l'année 2013 et les données de la ville d'Ottawa pour les années 2014 à 2016;
- Analyser un ensemble de données fournies par le Service de police d'Ottawa pour l'année 2013 et prédire s'il y'a un accident ou pas dans des conditions données ;
- Prédire les divers types d'accidents pour l'année 2013 ainsi que pour les années 2014 à 2016 (à savoir : accident fatal, accident avec blessure grave et accident incluant un dommage matériel) ;
- Trouver et implémenter une solution afin de résoudre le problème de déséquilibre de données ;
- Effectuer une évaluation comparative des résultats obtenus par les divers algorithmes, ainsi qu'avec les solutions semblables proposées dans la littérature ;
- Produire une liste des intersections dangereuses à Ottawa en utilisant des données sur le volume de trafic fournies par la ville d'Ottawa pour les années 2014, 2015 et 2016.

1.2.3. Cibles de la recherche

Ce mémoire de recherche s'adresse à toute personne ou organisme soucieux de la sécurité routière. Elle est destinée :

- à servir aux Services de police de la ville d'Ottawa dans la répartition des agents de la circulation dans les lieux stratégiques ;
- à aider les autorités routières de la ville d'Ottawa dans la mise en œuvre des règles de circulation ;
- à prévenir les usagers de la route sur les risques liés à la circulation dans certaines conditions ;
- à servir de base pour les futures recherches dans la prédiction des accidents de la route.

CHAPITRE 2 : ÉTAT DE L'ART

Ce chapitre porte essentiellement sur la revue de différents travaux de recherche axés sur l'analyse et la modélisation de la circulation routière, incluant les sources de données utilisées, les mesures standards utilisées pour évaluer la sécurité routière, les types, les impacts et les causes d'accidents. Il inclut aussi les techniques existantes dans la littérature pour la modélisation et la prédiction des accidents de la route.

2.1. Mesures d'évaluation des risques d'accidents

Le Ministère des Transports et les différents Conseils de différents pays utilisent une série de douze(12) mesures et méthodes pour évaluer la sécurité routière ([5], [6]).Ces mesures sont : la fréquence de collision, le taux de collision, la fréquence et le taux de collision combinés, la gravité de collision critique, le taux de collision, la méthode de gravité de collision, la méthode d'analyse de risque, les fonctions de performance de sécurité, la méthode Bayes empirique, le risque personnel et collectif, le niveau de sécurité et les indices de risque pour la sécurité routière. Celles-ci seront brièvement présentées dans les sections qui suivent.

2.1.1. La fréquence de collisions

La fréquence de collisions est une méthode pour identifier et évaluer la sécurité d'un site. Les fréquences de collisions observées dans le passé peuvent être utilisées pour comparer et classer le site avec des fréquences de collision dans un groupe d'emplacements similaires. Bien que simple en tant que concept, il existe plusieurs inconvénients d'utiliser cette mesure pour évaluer la sécurité [5]. Comme les collisions ne sont pas des événements fréquents, une fréquence de collisions élevée dans une année donnée à une intersection particulière pourrait simplement représenter une fluctuation aléatoire autour d'une moyenne à long terme beaucoup plus faible sur le site. Ce problème s'appelle le *problème de la régression vers la moyenne* (*regression to the meanproblem*). En outre, les sites avec des volumes plus élevés auront toujours une fréquence de collision plus élevée que les sites avec des volumes plus faibles. Enfin, cette méthode ne traite pas de la gravité des collisions: elle ne permet pas d'identifier les sites où le public risque davantage de blessures ou des décès.

2.1.2. Le taux de collisions

Le taux de collisions constitue une amélioration par rapport à la fréquence, car il considère l'exposition, qui représente également une mesure de risque auquel les utilisateurs font face sur une route spécifique.

Il est calculé en divisant la fréquence de collisions pendant une période de temps par le trafic annuel moyen estimé (*AADT*) des véhicules dans cette période.

$$R = C * \frac{10^8}{\sum AADT * 365.25} \quad (1)$$

où*C* représente la fréquence de collisions.

Tout comme la fréquence de collisions, le taux de collisions d'un site soumis à une évaluation de sécurité peut être comparé à des intersections similaires (par exemple, signalisées ou non, avec le même nombre de branches, ou avec la même quantité de trafic).

Le principal avantage de connaître le taux de collisions est le fait qu'il prend en compte l'effet qu'a le volume de collisions sur la fréquence.

Le principal inconvénient est le fait que l'utilisation d'un taux de collisions pour classer les sites qui ont un volume de trafic différent implique que la fréquence de collisions et le volume ont une relation linéaire, mais la recherche suggère que cela n'est pas généralement le cas [5]. En outre, le taux de collisions, comme la fréquence de collisions, ne tiennent pas compte de la gravité de la collision.

2.1.3. La fréquence et le taux de collisions combinés

Cette mesure capitalise les deux précédentes pour tenter de surmonter certains de leurs inconvénients. Les intersections avec une fréquence de collisions élevée et un taux de collisions élevé peuvent ensuite être des candidats pour des diagnostics de sécurité plus détaillés.

2.1.4. La mesure de la gravité d'une collision

Une autre méthode largement utilisée pour le dépistage de la sécurité des routes est la mesure de la gravité d'une collision (*collision severity method*) ou la méthode du taux critique pondéré. Elle équivaut à la fréquence de collisions en termes d'équivalence seulement aux dommages matériels (*EPDO*).

L'indice *EPDO* attache une plus grande importance, ou un poids, aux collisions causant une blessure grave ou une fatalité, et la moins grande importance pour les collisions liées à la propriété et aux dommages matériels. De cette façon, le problème de ne pas tenir compte de la gravité (la fréquence de collisions, le taux de collisions et leur combinaison) est abordé.

Le Département des Transports des États-Unis et l'Association américaine des administrateurs de véhicules à moteur, dans le document intitulé « *L'évaluation statistique dans les études de sécurité routière* », identifient les facteurs de pondération indiqués dans le tableau 1 à utiliser pour le calcul de la méthode de gravité de collision [5].

En utilisant ces poids, l'index *EPDO* peut être calculé comme :

$$EPDO = 9.5F + 9.5\ MAJ + 3.5\ MIN + PDO \quad (2)$$

où F représente le nombre de collisions mortelles, MAJ le nombre de collisions avec blessures majeures, MIN le nombre de collisions avec blessures mineures et le PDO le nombre de dommages matériels dans les collisions.

Tableau 1 : Valeurs du poids dans la méthode de gravité de collision proposée dans [5]

Gravité	Poids
Collisions fatales	9.5
Blessures d'infirmité (blessure de type A) -> Toutes blessures non fatales mais handicapantes à vie (rendant la victime incapable de marcher, de conduire ou de mener d'autres activités).	9.5
Blessures de non infirmité (blessure de type B) -> Toute blessure qui n'est pas fatale ou ne rend pas infirme la victime.	3.5
Possibilité de blessure (blessure de type C) - Blessure non visible mais empreinte de douleur.	3.5
Collision PDO -> uniquement biens touchés.	1.0

Outre les valeurs des poids dans l'équation, l'utilisation d'autres schémas de pondération (par exemple 100/100/10/1, 40/40/3/1) est analysée dans [7].

On suggère dans [5] que, selon les considérations locales, l'équation (2) pourrait également être modifiée pour tenir compte des valeurs réelles en termes de coût. Une fois que l'*EPDO* est calculé, le risque (*Rw*) peut être calculé en remplaçant la fréquence de collisions par l'indice *EPDO* dans l'équation (1), [7]:

$$Rw = EPDO * \frac{10^8}{\sum AADT * 365.25} \quad (3)$$

Le principal avantage de cette mesure est le fait qu'elle considère la gravité, alors qu'elle est désavantageuse puisqu'elle tend à mettre en évidence les emplacements avec des vitesses plus élevées.

Plus précisément, les intersections signalées sur les routes avec une vitesse d'exploitation plus élevée, comme dans une zone rurale, auront probablement un indice *EPDO* plus élevé que dans une zone urbaine.

2.1.5. Le taux critique de collisions

Le taux critique de collisions représente le taux de collisions prévu des emplacements ayant des caractéristiques similaires (par exemple, le même dispositif de contrôle de la circulation). Il permet une comparaison avec d'autres sites similaires et incorpore un test statistique simple pour déterminer si le taux de collisions est significativement plus élevé que prévu. Le taux de collisions critique (*Rc*) peut être calculé en fonction du taux de collisions moyen pour toute intersection *Ra* (qui peut être calculé en utilisant soit les équations (1) ou (3)), *m* est le nombre de millions de véhicules entrant dans l'intersection et *k* est une constante, dont la valeur est fixée à 1.282 pour un niveau de confiance de 90% [7]:

$$Rc = Ra + k\sqrt{Ra/m} + \frac{1}{2m} \qquad (4)$$

Si le taux de collisions réel est supérieur au taux critique calculé dans l'équation (4), l'écart est probablement dû aux caractéristiques défavorables de l'intersection ou de la section routière [5]. Cette mesure est plus robuste que l'utilisation de la fréquence de collisions ou du taux de collisions seul, car elle fournit un moyen de tester statistiquement la différence de taux de collisions sur un site par rapport à un groupe de sites similaires.

Cependant, le principal inconvénient de la méthode est le fait qu'elle assure que le volume de trafic et les collisions ont une relation linéaire et ne considère pas le problème de la régression vers la moyenne.

2.1.6. Les méthodes d'analyse de risques

L'idée principale de l'analyse des risques est de déterminer le risque de collisions en utilisant des données de collisions et de volume dans des endroits spécifiques (risque local), dans un groupe spécifique de localisation (risque de zone) ou dans l'ensemble de la juridiction (risque global) [5].

Les collisions de différentes gravités peuvent être pondérées dans l'analyse selon l'indice *EPDO* (comme dans la section 2.5).

Les emplacements peuvent ensuite être comparés en fonction de leur risque relatif en combinant les résultats des calculs de risques locaux, régionaux et globaux.

Cette catégorie de méthodes est robuste, car elle tient compte de l'exposition (volume) et de la gravité des collisions. Cependant, elle suppose toujours que la relation entre la fréquence de collisions et le volume est linéaire et ne considère pas le problème de la régression vers la moyenne.

2.1.7. La fonction de performance de sécurité

Cette méthode met l'accent sur l'utilisation de modèles statistiques pour traiter le caractère aléatoire inhérent des accidents. Une fonction de performance de sécurité (*Safety Performance Function ouSPF*) est un modèle de régression pour estimer la fréquence de collisions moyenne prédite des segments ou intersections de routes individuelles. Les SPF sont développés en utilisant des données historiques de collisions recueillies au cours d'un certain nombre d'années, dans des emplacements donnés présentant des caractéristiques similaires. Les paramètres de régression sont déterminés en supposant que les fréquences de collisions suivent une distribution binomiale négative, qui est une extension de la distribution de Poisson [6]. Graphiquement, un SPF représente une courbe qui est la meilleure possible grâce aux différents points. Généralement, les SPF démontrent que le nombre attendu de collisions augmente à mesure que le volume de trafic augmente (les sites à volume supérieur ont un taux de collisions inférieur à celui des volumes plus bas).

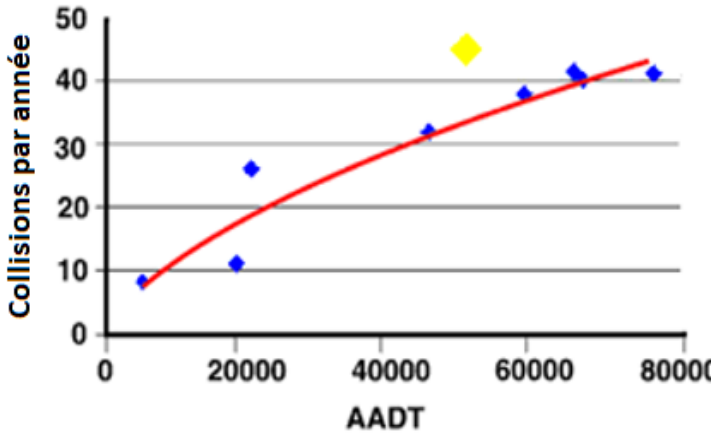


Fig.2 Un exemple de fonction de performance de sécurité (adapté de [5])

La figure 2 montre un exemple de fonction de performance de sécurité où les points bleus représentent les intersections individuelles avec leur fréquence de collision moyenne annuelle et le trafic quotidien moyen annuel (*AADT*). Le point jaune est un exemple d'intersection qui présente un rendement pire que prévu.

Les modèles multivariés peuvent également incorporer d'autres variables en plus du volume de trafic, y compris les éléments géométriques de la chaussée, la densité d'accès ou la météo [6].

Les SPF peuvent être développés pour la fréquence totale des accidents, y compris tous les niveaux de gravité des collisions. Les SPF peuvent également être développés spécifiquement pour un type de collision donné.

Les avantages de l'utilisation d'une telle méthode sont que le potentiel d'amélioration de la sécurité est plus précisément calculé et qu'elle reconnaît que la relation entre la fréquence et le volume de collision n'est pas linéaire. Le principal inconvénient est sa complexité relative.

2.1.8. La méthode Empirical Bayes (EB)

Tout comme beaucoup de juridictions aux États-Unis, le Canada utilise déjà la méthode Empirical Bayes (EB). Elle calcule les fréquences de collisions prévues grâce à une combinaison de fréquences de collisions observées et estimées (prédites). Les fréquences de collisions estimées proviennent du développement d'une courbe SPF. Dans ce cas, la courbe SPF rapporte le niveau de sécurité d'une intersection au volume de trafic et aussi à d'autres facteurs géométriques pertinents.

La méthode EB utilise un facteur pondéré (w), voir l'équation (5), qui est une fonction du paramètre de dispersion excessive (*overdispersion*) du SPF (*k*, dans l'équation (6)) pour combiner les deux estimations. Lorsque la valeur du paramètre de dispersion excessive augmente, le facteur d'ajustement pondéré diminue. De cette façon, on met davantage l'accent sur les collisions observées / déclarées plutôt que sur la fréquence de collisions prédite par SPF. Cette estimation dépend des caractéristiques des données (dispersées par rapport à la dispersion excessive) utilisées pour développer les modèles de prédiction.

La fréquence de collisions prévue peut être calculée comme [6]:

$$N_{prévue} = w * N_{prédit} + (1 - w) * N_{observé} \quad (5)$$

Où w est l'ajustement pondéré à placer sur l'estimation du modèle de prédiction qui peut être calculé comme suit :

$$w = \frac{1}{1+k*\sum_{Touslesannéesd'étude} N_{prédit}} \quad (6)$$

Le paramètre *k* est le paramètre de dispersion excessif du SPF associé utilisé pour estimer $N_{prédit}$. La valeur de ce paramètre *k* peut être ajustée conformément au tableau 2.

Tableau 2: Paramètres de dispersion excessive (extrait de [6])

Type de facilité	Paramètres de dispersion excessive (k)
Rural à deux voies, Segment de route à deux voies	0,236 par longueur du tronçon routier
Intersection contrôlée des trois côtés	0,54
Intersection contrôlée des quatre côtés	0,24
Interception signalée des quatre côtés	0,11

La fréquence d'accidents prédite dans des conditions réelles peut être calculée selon [6]en utilisant l'équation 7:

$$N_{prédit} = (N_{spf\,x} * (CMF_{1x} * CMF_{2x} * CMF_{3x} * \ldots\ldots * CMF_{yx}) + N_{pedx} + N_{bikex}) * C_x \qquad (7)$$

Où$N_{prédit}$ est la fréquence moyenne prédite de collision pour une année spécifique sur le type de site *x*, $N_{spf\,x}$ représente les conditions de base pour la fréquence de collisions moyenne prédites pour le type de site *x*, et qui est calculé pour un segment $N_{spf\,x}$ comme $N_{spf\,x} = N_{bmv\,x} + N_{bsv\,x} + N_{bdwy\,x}$, et pour une intersection en utilisant la formule$N_{spf\,x} = N_{bmv\,x} + N_{bsvx}$,et$N_{bmvx}$ représente les conditions de base de prédiction des collisions deplusieurs véhicules sur des voies qui ne sont pas des chaussées pour le type de site *x*. Les conditions de base $N_{bsv\,x}$ représente la fréquence moyenne de collision impliquant un seul véhicule pour le type de site *x*, les conditions de base de $N_{bdwy\,x}$ représente la fréquence moyenne de collisions impliquant plusieurs véhicules pour le type de site *x*, $N_{ped\,x}$ représente la fréquence moyenne de collision entre les véhicules et les piétons par année pour le type de site x, $N_{bike\,x}$ représente la fréquence d'accident moyenne prédite des collisions entre des véhicules et des vélos par an pour le type de site *x*, CMF_{yx} sont des facteurs de modification de collision spécifiques au type de site *x* et des fonctions de conception géométrique et de contrôle de trafic spécifiées *y*. Finalement, C_x est un facteur d'étalonnage pour ajuster la fonction de performance de sécurité pour les conditions locales pour le type de site *x*[6].

CMF est donc le facteur de modification de collision et représente le changement relatif de la fréquence de collisions moyenne estimée en raison de différences pour chaque condition spécifique. Il fournit une estimation de l'efficacité de la mise en œuvre d'une contre-mesure particulière, par exemple, le pavage des épaules de gravier, l'ajout d'une voie de virage à gauche ou l'augmentation du rayon d'une courbe horizontale. Ces coefficients sont disponibles dans [6].

Tableau 3 : Résumé de la démarche adoptée dans l'état de l'art (extrait de [5])

Méthodes	Avantages	Inconvénients
1. Fréquence de collisions	o Facile à utiliser o Facile à comprendre	⇒ Biaisée pour les sites à forte circulation ⇒ Ne prend pas en compte l'exposition ⇒ Ne tient pas compte de la gravité d'accidents ⇒ Régression à la moyenne non adressée
2. Taux de collisions	o Facile à utiliser o Prend compte de l'exposition	⇒ Biaisée pour les sites à faible circulation ⇒ Requiert des données de taille ⇒ Assure que les collisions et le volume ont une relation linéaire ⇒ La gravité n'est pas prise en compte ⇒ Régression à la moyenne non adressée
3. Taux de collisions critique	o Relativement simple o Prend compte de l'exposition o Se base sur des méthodes statistiques bien établies	⇒ Requiert des données de taille ⇒ Assure que les collisions et le volume ont une relation linéaire ⇒ La gravité n'est pas prise en compte ⇒ Régression à la moyenne non adressée
4. Méthode d'analyse de collisions	o Relativement simple o Considère l'exposition	⇒ Biaisée dans des sites de haute vitesse ⇒ Assure que les collisions et le volume ont une relation linéaire ⇒ Régression à la moyenne non adressée
5. Méthodes d'analyse de risque	o Précise o Considère la gravité et l'exposition o Considère la variation du niveau de sécurité localement parmi un groupe de localisations similaires	⇒ Requiert des données de taille ⇒ Assure que les collisions et le volume ont une relation linéaire ⇒ Régression à la moyenne non adressée
6. Fonctions de performance de sécurité	o Plus précise o Considère l'exposition o Reconnait que les collisions et le volume ont une relation non linéaire	⇒ Requiert des données de taille ⇒ Régression à la moyenne non adressée ⇒ Calculs intensifs ⇒ Difficile à comprendre pour le public
7. Méthode EB	o Plus précise o Considère l'exposition o Reconnait que les collisions et le volume ont une relation non linéaire o Aborde le problème de la régression à la moyenne	⇒ Requiert des données de taille ⇒ Plus difficile à comprendre

À l'aide de ce modèle, les sites peuvent être classés pour déterminer le nombre le plus élevé de collisions en fonction des comptes de collisions réels.

Parmi les avantages de cette méthode, nous pouvons citer le fait qu'elle est exacte et produit des normes de sécurité plus stables et plus précises. Elle aborde aussi le problème de la régression vers la moyenne, permet des estimations au fil du temps de la collision prévue et reconnaît la relation non linéaire entre les collisions et le volume de trafic[6]. Le tableau 3 présente un résumé des avantages et inconvénients des différentes méthodes utilisées dans l'état de l'art.

2.1.9. Le risque collectif

Le Guide d'intersections à haut risque de l'Agence de Transport de la Nouvelle-Zélande classe le statut des intersections individuelles en fonction de trois mesures de risque, à savoir le risque collectif, le risque personnel et le niveau de sécurité [8, 9].

Il existe deux définitions acceptées pour le risque collectif [8]. La première indique que le risque collectif est égal à la densité de collisions, ou plus spécifiquement au nombre de blessures graves ou de décès (*Death or severeinjury*, DSi) qui se sont produites à une intersection dans une période de temps (normalement 5 ou 10 ans). Dans ce cas, le risque collectif peut être considéré comme l'équivalent des accidents DSi réels. La deuxième définition implique la multiplication de chaque accident de type blessure à une intersection donnée par le taux d'indice de gravité correspondant afin de tenir compte de la gravité de l'accident. Dans ce cas, le risque collectif devient la prédiction du nombre de collisions DSi en fonction de toutes les collisions avec blessures qui se sont produites à une intersection et qui équivaut à des collisions DSi estimés.

Selon le tableau 4, les intersections classées avec un risque moyen-élevé ou élevé, ou avec un risque prédit supérieur à 1.2, sont considérées comme des intersections à haut risque.

Tableau 4 : Critères d'identification des intersections à haut risque basés sur le risque collectif (extrait de [8])

Niveau de risque collectif	Collisions Dsi estimées (5 ans) pour accidents de type blessures
Elevé	> 1.6
Moyen Elevé	1.2 – 1.6
Moyen	0.85 – 1.2
Faible Moyen	0.5 – 0.85
Faible	< 0.5

2.1.10. Le risque personnel

Le risque personnel mesure le risque pour chaque personne qui utilise une certaine intersection. En particulier, il identifie le site présentant le risque le plus élevé par véhicule.

Il est calculé comme le risque collectif divisé par une mesure de l'exposition au volume de trafic [9]:

$$risquepersonnel = \frac{\max(CollisionsF\&Sreportés * 0.5, equivautauDSIestimé) * 10^8}{(moy(Q_{major1,}Q_{major2}) * moy(Q_{minor1,}Q_{minor2}))^{0.4} * 5\ ans * 365\ jours * 1.7} \quad (8)$$

où les *Collisions F&S reportées* représentent des collisions fatales et sérieuses qui ont été signalées, les DSi estimés sont les collisions DSi estimés, Q_{major1} et Q_{major2} représentent le volume de liaison bidirectionnelle (AADT) sur chaque branche d'une route majeure, et Q_{minor1} et Q_{minor2} le volume de liaison bidirectionnelle (AADT) sur chaque branche d'une route mineure. Pour l'intersection en forme de T, Q_{minor1} prend la valeur de ADDT sur la route secondaire et Q_{minor2} est égal à 0 [9].

Afin de générer une estimation fiable du risque personnel, cette mesure n'est calculée que pour les intersections avec quatre accidents de blessures enregistrés ou plus au cours des cinq dernières années. Cela remonte à des conclusions potentiellement trompeuses sur le risque d'intersections avec de faibles volumes de trafic, qui sont particulièrement sensibles aux variations dans les nombres d'accidents. Selon cette formule et les critères du tableau 5, les intersections classées avec un risque moyen-élevé ou élevé, ou avec un risque prédit de plus de 100, sont considérées comme des intersections à haut risque.

Tableau 5: Critères d'identification des intersections à hauts risques basés sur le risque personnel (extrait de [8])

Niveau de risque personnel	Métriques de risque
Elevé	>130
Moyen Elevé	100 – 130
Moyen	70 – 100
Faible Moyen	40 – 70
Faible	< 40

2.1.11. Le niveau de service de sécurité

Le niveau de service de sécurité (*Level of safety service*, LoSS) est une mesure des performances de sécurité historiques d'une intersection par rapport à la performance attendue calculée sur la base d'une analyse statistique d'une intersection [10].

L'objectif est d'identifier les intersections qui sont peu performantes par rapport aux intersections similaires de la même configuration, en tenant compte de facteurs tels que : l'environnement de vitesse, la forme d'intersection et la quantité de trafic traversant l'intersection.

Les calculs LoSS ne nécessitent aucune information supplémentaire au-delà de celle utilisée pour calculer les niveaux de risque personnel. La performance de collisions de blessure d'une intersection a été séparée en cinq bandes LoSS pour aider à prioriser les intersections problématiques pour le traitement, comme le montre le tableau 6.

Tableau 6 : La bande LoSS (extrait de [10])

Niveau de service de sécurité	Performance de la sécurité	Définition
LoSS V	90 - 100ème percentile	Le taux de collisions de blessure observé est dans le pire 10% - plus élevé (pire) que celui attendu des 90% d'intersections similaires.
LoSS IV	70 - 90ème percentile	Le taux de collisions de blessure observé est dans le pire 30% - moins (meilleur) que celle prévu des 90% d'intersections similaires et plus élevé (pire) que celui de 70%.
LoSS III	50 – 70ème percentile	Le taux de collisions de blessure observé est faible (positivement) que celui prévu des 70% d'intersections similaires et plus élevé (pire) que celui des 50%
LoSS II	30 – 50èmepercentile	Le taux de collisions de blessure observé est faible (meilleur) que celui prévu des 50% d'intersections similaires et plus élevé que celui des 30%
LoSS I	0 – 30ème percentile	Le taux de collisions de blessure observé est faible (meilleur) que celui prévu des 30% d'intersections similaires.

2.1.12. Indices de risque routier

Un cadre d'indices de risque routier (*road risk indices*, RRI), basé sur la géométrie de la route, les conditions de circulation et les données historiques des collisions afin d'évaluer les risques de l'infrastructure routière existante pour les usagers et les agences de la route, est proposé dans [10,11].

Deux RRI sont proposés : 1) l'indice fournissant des informations de sécurité pour les conducteurs individuels (RRI pour l'exposition individuelle) et 2) l'indice reflétant les performances de sécurité des sections de la route et des intersections, à utiliser par les agences de la route (RRI pour une section de route). Ceux-ci sont quelque peu semblables au risque personnel et collectif décrit ci-dessus.

Cependant, la principale différence réside dans le fait que le RRI prend en compte le niveau relatif des coûts engagés dans l'accident.

Le RRI pour l'exposition individuelle mesure le risque individuel d'un conducteur sur un segment de chaussée homogène ou une intersection. Il est formulé en fonction de deux facteurs : le taux de collisions et les contraintes de collisions et il est défini sur une échelle de 0 à 10, avec 0 représentant le risque le plus bas. Il peut être calculé comme [11]:

$$RRI_{ind}(i) = f\left(\sum_j N_{ij} * AS_j\right) \tag{9}$$

Où i est le segment de la route ou l'intersection ; j = 1 à 3, indique les trois types de collisions :(1) accident mortel, (2) blessures et (3) dommages matériels uniquement ; N_{ij} est le nombre prévu d'accidents de type j sur le segment routier i (par million de véhicules-mile), ou à l'intersection i (par million de véhicules);*ASj* est le niveau relatif des coûts correspondant au type de collisions j; et f (x) est une fonction de transformation qui contraint l'entrée à une plage souhaitée. En particulier, la fonction logistique généralisée est généralement utilisée :

$$f(x) = A + \frac{C}{(1 + T * \exp(-B * (x - M)))^{1/T}} \tag{10}$$

Où *A* est l'asymptote inférieure (par exemple 0), *C* représente l'asymptote supérieure (par exemple 10), *M* est le temps de croissance maximale, *B* la vitesse de croissance et *T* est un facteur apparent près duquel se produit la croissance maximale de l'asymptote.

Le RRI pour une section de route représente le potentiel de risque accumulé d'un lien ou d'un nœud en fonction de son influence sur la fiabilité du service qu'un segment de la route ou une intersection est supposée fournir.

Pour un segment de route homogène, le RRI est défini comme suit :

$$RRI_{acu}(i) = g(RRI_{ind}(i) * EXPO_i) \tag{11}$$

Où $EXPO_i$, l'exposition (en millions de miles de parcours) des véhicules par année sur la chaussée section i, est calculée comme suit:

$$ADT_i * L_i * 10^{-6} * 365$$

Où bien $EXPO_i$, l'exposition (en millions de miles de parcours) des véhicules par année à l'intersection i, est calculé comme suit:

$$(ADT_{i,1} + ADT_{i,2}) * 10^{-6} * 365$$

ADT_i représente la circulation journalière moyenne (*ADT*) sur le segment homogène i, $ADT_{i,1}$ est le trafic quotidien moyen sur la route principale à l'intersection i, $ADT_{i,2}$ le trafic quotidien moyen sur la route mineure à l'intersection i, L_i est la longueur du segment homogène *i* et *g* est une fonction de transformation qui contraint l'entrée à une plage souhaitée.

2.2. Types d'accidents

Selon les mesures présentées à la section 2.1, trois principaux types d'accidents sont généralement identifiables : accidents mortels, accidents de blessures et accidents accompagnés par des dommages matériels [7]. Un accident mortel est une collision qui entraîne au moins un mort. Le décès survient soit sur la scène, soit dans un certain laps de temps à partir de la date de la collision (souvent 30 jours). Un accident de blessures est une collision qui entraîne au moins une blessure pour toutes les personnes impliquées dans la collision. Le plus souvent, la blessure doit être évidente pour que le personnel de réponse à l'urgence la classifie comme une collision de blessure. Dans certains cas, l'accident sera enregistré comme un accident de blessures si la personne impliquée dans la collision indique qu'elle croit qu'elle est blessée.

Enfin, nous nous référons à un accident en tant qu'un accident de dommages matériels (DOP) lorsque la collision ne concerne que des dommages matériels aux véhicules et/ou à la propriété en raison de la collision, et aucune blessure apparente (ou déclarée) ou décès n'est survenu(e).

2.3. Tendances temporelles des accidents

Plusieurs rapports ont analysé la fréquence des accidents au cours de la semaine, au cours de la journée et au cours des mois.

Quelques études canadiennes [12, 13] identifient le vendredi comme le jour de la semaine le plus propice à la collision. Près de 17% des collisions déclarées par la police au Canada semblent avoir lieu le vendredi [14]. Le vendredi est théoriquement suivi, en ordre, par le mercredi, le jeudi, le mardi et le lundi. Le samedi est le deuxième plus bas et le dimanche, le plus bas.

Une étude de la répartition des accidents au cours des jours a identifié l'heure de pointe du soir (à partir de 15 heures et se terminant à 18 heures) comme le moment où survient la plupart des collisions, environ 25% de la totalité des collisions qui se produisent au cours de la journée [14].

En termes d'accidents mortels, la majorité semble se produire entre 6 heures et 9 heures, tandis que le deuxième moment le plus meurtrier de la journée était l'intervalle compris entre 15h et 18h [15].

Enfin, une certaine tendance peut être identifiée au cours des mois dans [13, 16]. En général, pendant les mois d'automne et d'hiver (d'octobre à janvier) et le mois de juin, un nombre plus élevé d'accidents semblent

se produire. En termes d'accidents mortels, les mois d'été semblent être les plus mortels, le mois de juin se trouvant au sommet [17].

Une réévaluation périodique de ces tendances et leur personnalisation pour des régions spécifiques pourraient être incluses dans le système expert afin d'améliorer les prédictions des collisions en les considérant comme variables d'entrée ou en ajoutant une poids plus significatif pour certains jours et mois [18].

2.4. Causes des accidents

Les causes des accidents de la route sont multiples. Les statistiques démontrent que les causes humaines sont les principales causes des accidents de la route [19]. Les causes humaines arrivent loin devant les autres causes comme les causes climatiques, environnementales et autres causes.

Dans plus de 90% des accidents de la route on peut constater un facteur humain [19], ces facteurs sont entre autres : l'alcool, la vitesse et la somnolence qui se retrouvent respectivement dans 31%, 25% et 8% des accidents fatals. On constate aussi que le non port de la ceinture de sécurité dans 21% des accidents fatals et 6% des motocyclistes ayant succombé à un accident de la route ne portaient pas de casque [19].

Parmi les causes des accidents nous avons aussi les causes environnementales et climatiques. Dans [20] nous pouvons constater que plus de 50% des accidents se produisent sur une surface sèche, 60% en plein jour et 75% par temps clair. D'autres facteurs comme la pluie, la neige et le vent font partie des causes environnementales les plus pertinentes dans les accidents de la route. Dans ce projet de mémoire nous proposons d'ajouter des paramètres comme les conditions climatiques et environnementales aux données que nous avons à notre disposition pour une étude de leur impact sur les résultats de prédictions.

2.5. Méthodes pour la modélisation et la prédiction d'accidents

En raison de son importance, le sujet de la modélisation et de la prédiction des accidents de la circulation a suscité l'intérêt de nombreux chercheurs. Dans ce chapitre, nous discutons une variété de techniques et méthodologies sous les termes parapluies «intelligence informatique» et «extraction de données», qui ont été utilisées dans la littérature à cet effet, à savoir: les réseaux de neurones, les machines de vecteurs de support, la régression, les arbres de décision, les réseaux bayésiens, les règles d'association, les techniques de regroupement, le raisonnement basé sur des cas (*case base reasoning*) et ontologies.

La figure 3 montre la répartition des méthodes sur les documents identifiés dans la littérature.

Alors que certaines solutions sont plus utilisées que d'autres, en particulier les réseaux de neurones, la régression, et les arbres de décision, aucune d'entre elles n'est une méthode de choix clair au sein de la communauté de recherche.

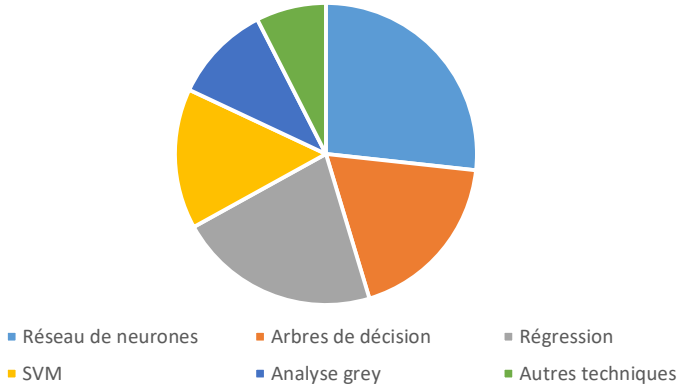


Fig.3 : Distribution de quelques méthodes utilisées dans la prédiction et modélisation d'accidents.

Ces techniques et leur utilisation pour la modélisation et la prédiction des accidents sont brièvement décrites dans les sous-sections qui suivent.

2.5.1. Les réseaux de neurones

Parmi les solutions les plus fréquemment employées figurent les réseaux de neurones. La force des réseaux neuronaux découle de leur non-linéarité intrinsèque, de leur simplicité de calcul et de leur capacité à apprendre, et donc à prédire. Un réseau neuronal a la capacité d'apprendre des associations redondantes à haute dimension à partir d'ensembles de données mesurées sans faire appel à un modèle mathématique. Bien qu'elle puisse prendre un temps relativement long pour apprendre, la phase de rappel se déroule en temps réel. Une estimation de la sortie peut être fournie instantanément pour les valeurs d'entrée ne faisant pas partie du jeu d'entraînement. Des formes diverses ont été utilisées dans la littérature pour la prédiction des accidents de la route : les réseaux de rétro-propagation (*backpropagation*) [21, 22], les réseaux de neurones probabilistes [23, 24], et les réseaux de neurones en ondelettes [25].

Les machines à vecteurs de support (SVM), ont également été utilisées dans le contexte des accidents de la circulation dans plusieurs publications [26, 27, 28, 29]. Des solutions basées sur des réseaux de neurones hybrides, telles que des techniques neuro-floues adaptatives (par exemple ANFIS), sont proposées dans [30, 18] et les hybrides de réseaux de neurones Grey dans [31, 32].

Un réseau neuronal de rétro-propagation [22] est entraîné sur des données représentant 102 intersections signalées et 3441 enregistrements d'accidents de 1999 à 2004. Le réseau a 34 entrées représentant la largeur de la route, le pourcentage de motocyclettes, la complexité du scénario de synchronisation des signaux, le

volume global de trafic, le pourcentage de véhicules tournant à gauche, le temps de cycle du signal, le volume de trafic directionnel, l'existence de panneaux d'avertissement, le type médian entre les voies de circulation rapides et lentes, les obstacles dans les voies de circulation et le pourcentage de véhicules tournant à droite et prédit le nombre d'accidents. On a constaté que le modèle de réseau neuronal peut donner des prédictions plus fiables que le modèle de régression binomiale négative. Un autre réseau de neurones de rétro-propagation dans [21] classe le type d'accident, comme un accident du véhicule unique, un accident de collision arrière, un accident de collision avant, un accident de collision latérale ou un accident d'égratignure en utilisant comme entrées le volume de trafic à l'intersection, l'emplacement et le type d'intersection, le niveau des routes en intersection et le mode de contrôle de la circulation.

Une solution est proposée dans [25] pour la prédiction de la perte de vie causée par des accidents de la circulation basée sur les réseaux de neurones en ondelettes et qui tient compte du nombre d'accidents, du nombre de décès et du nombre de blessures.

Li et al. [18] utilisent un modèle de réseau neuronal hybride pour prédire l'indice de risque routier discuté dans la section 2.1.12. Les données brutes sont divisées en 3 classificateurs utilisant un classificateur de type C-moyen flou et pour chaque classificateur, un réseau neuronal distinct est développé. Chacun de ces trois réseaux a comme entrées la rugosité de la route, la limite de vitesse, la longueur du segment, le nombre de voies, l'ADDT par voie, la largeur par voie, la courbe et le grade ; et les trois sorties correspondant aux trois types d'accidents (mortel, blessures, et dommages matériels).

Pour calculer la sortie du réseau hybride, la distance euclidienne est calculée entre les entrées et les centres des clusters. Le réseau neuronal correspondant au classificateur de distance minimale est appliqué pour prédire le type d'accident et l'indice de risque routier. Le modèle est également augmenté avec une couche dynamique pour inclure les influences du temps, l'heure du jour et le jour de la semaine en employant des facteurs de correction multiplicatifs.

Polat et Durduran [33] proposent une pondération d'attributs basée sur l'algorithme K-moyen pour un réseau neuronal, afin d'augmenter les performances de classification de l'algorithme de classification et de transformer l'ensemble de données sur les accidents de la route linéairement non séparables en un ensemble linéairement séparable.

Les variables d'entrée considérées sont : le jour, la température, l'humidité, les conditions météorologiques et le mois de l'accident de la circulation et la sortie est accident/pas d'accident. La performance de leur réseau de neurones est d'environ 74,15%.

Zu [31] prédit les accidents de la circulation en formant un réseau de rétro-propagation Grey sur les données de la criminalité en Chine de 1889 à 2006. L'analyse de l'entropie de la relation Grey est également utilisée pour la prédiction des accidents dans [23, 34].

2.5.2. Les arbres de décision

Une autre parmi les solutions les plus utilisées pour l'analyse des données de trafic routier et pour l'identification des causes des accidents est l'arbre de décision.

Zhang et Fan [35] obtiennent la répartition probabiliste des facteurs qui causeraient divers types d'accidents de la route basés sur des données d'une période d'un an fournies par l'Autorité routière de la Saskatchewan et en utilisant un arbre décisionnel. Les facteurs considérés comprennent : l'attention, l'alcool, l'état physique (état), l'inexpérience, la règle (le chauffeur enfreint les règles de circulation), les erreurs, la vitesse, le véhicule, la route, la météo et la visibilité. Les expériences se déroulent selon trois aspects différents en ce qui concerne l'âge, la saison et le genre. En particulier, en ce qui concerne l'âge, les principaux facteurs contributifs pour le groupe junior étaient «boire», «règle» et «erreur», pour le groupe des adultes : «règle», «boisson» et «attention» et pour la «règle» du groupe des personnes âgées. Pour les saisons, causes majeures des accidents hivernaux identifiés ont été «boire» et «règle», alors que pour une condition non hivernale : «boire», «erreur» et «règle». Enfin, selon le sexe, pour le groupe masculin : «boire», «erreur» et «règle» et le groupe féminin : «règle», «temps» et «boisson». La performance obtenue sur les trois aspects varie entre 75,9% et 82%.

Les auteurs de [36] classent les accidents comme mortels, blessés et non blessés en utilisant des arbres de décision CART (*Classification and RegressionTrees*), TreeNet et la forêt aléatoire (*Randomforest*).

Une série de 32 attributs sont utilisés sur un ensemble de données de 4 ans, y compris: type de collisions, arrondissement, âge et profession des victimes, type de véhicule, état de santé des victimes, cause immédiate de l'accident, catégorie de victimes, heure de l'accident, niveau de permis de conduire, expérience de conduite, jour d'accident, état de la lumière, catégorie de plaque de véhicule, séparation de route, âge du conducteur, semaine spécifique d'un mois, année de service du véhicule, propriété du véhicule, niveau de conduite du conducteur, type de jonction entre routes. Cela prend aussi en compte le genre de conducteur, l'orientation routière, le mouvement des piétons pendant l'accident, le statut technique du véhicule, l'état de la route, les conditions météorologiques et le type de surface de la route. Pour la prédiction des accidents mortels, la performance varie entre 64,2% et 77,4%, pour les accidents de type blessures entre 55,3% et 77,9%, tandis que pour les accidents sans-blessure entre 99,9% et 100%. La forêt aléatoire prévoit mieux les accidents mortels, tandis que TreeNet est mieux pour un accident de blessure. Dans l'ensemble, TreeNet

semble fonctionner mieux, mais avec seulement 0,98% par rapport à CART et avec 3,75% par rapport aux forêts aléatoires.

Un arbre de décision CART pour prédire le niveau de blessure (mortel, blessure et sans blessure) en utilisant 22 variables est proposé dans [37]. Les variables comprennent : le mois, l'heure, le jour de la semaine, les conditions météorologiques, l'état de la lumière, la surface de la route, l'obstruction de la route, l'emplacement de l'accident, le type de contrôle, l'autoroute divisée, la limite de vitesse, le sexe, l'âge, la qualification, le système de retenue, l'état de sobriété, l'action du conducteur / véhicule / piéton, le type de collisions et les circonstances contributives et leur description et type. Le modèle obtient une performance de 96,4% pour les blessures et 88,5% pour les non-blessures, mais n'est pas capable de prédire les accidents mortels (0% de performance).

2.5.3. La régression

La dernière parmi les techniques souvent utilisées dans la prédiction d'accident de la route est la régression, sous ses diverses formes : la régression de la probabilité conjointe [38], la régression linéaire généralisée [39], la régression logistique [40], logit ou régression logistique [41], le probit ordonné, le logit ordonné, le logit multinomial [42], l'indicateur local I de Moran de l'association spatiale [43, 44], Poisson-lognormal multivarié [45], pour n'en citer que quelques-uns. La régression de probabilité conjointe est utilisée dans [38] pour modéliser la fréquence et la gravité des collisions.

Les variables considérées sont la densité de la route, la densité à la jonction, la vitesse du véhicule, le temps et la géométrie locale de la route. Les auteurs de [39] proposent un modèle de régression linéaire généralisé pour produire une estimation de la fréquence de collisions pour un emplacement en fonction des caractéristiques propres au site en utilisant des données de trafic en Colombie-Britannique. Deux principales sources de données sont utilisées : (1) les segments routiers décrits par l'utilisation du sol (urbain / rural), la classe routière (artère, voie rapide, autoroute), médiane (divisée / non divisée) et le nombre de voies (2 ou 4) et (2) les intersections routières décrites comme contrôlées par signal ou par un stop et 4-jambes ou 3-jambes, respectivement. Les données sur l'histoire des collisions de cinq ans (2001-2005) et les volumes de trafic pour chaque segment d'autoroute sont utilisées pour créer des modèles distincts par segment d'autoroute et par intersection pour chaque type de collision, y compris les collisions mortelles et blessures (combinées en "sévères") et les collisions DOP (dommages matériels uniquement). L'estimation des paramètres du modèle se fait à l'aide de la régression linéaire généralisée.

Un modèle de régression logistique pour la gravité des accidents est proposé dans [40] en fonction de la date de l'accident (jour ouvrable, vacances), conditions météorologiques, type de chaussée, section

transversale de la route, emplacement de l'accident, alignement de la route, type de route, contrôle de la circulation et conditions d'éclairage.

L'heure des collisions impliquant de gros camions est analysée à l'aide d'un modèle de régression logit dans [41]. Zhan et al. [45] utilisent une régression Poisson-lognormal multivariée pour modéliser les dommages matériels, les blessures possibles et les accidents de blessures évidents sur l'ensemble de données sur les autoroutes de l'Etat de Washington en fonction de 8 variables qui sont : la longueur du segment de l'autoroute, la moyenne annuelle du trafic journalier par voie, la différence de grade maximale dans le segment, le nombre de courbes horizontales par mille dans le segment, le pourcentage de camions dans le trafic, l'indicateur de faible précipitation (≤ 12 po par an), l'indicateur de neige lourde (≥ 18 po par an) et l'indicateur local routier.

2.5.4. Autres approches pour la prédiction et la modélisation d'accidents de route

Certaines autres solutions proposées pour la modélisation et la prédiction des accidents de la circulation sont des techniques de regroupement telles que la méthode du k-voisins le plus proche [46, 47], le regroupement C-moyen [46], ou DBSCAN [48], le raisonnement basé sur des cas [49, LiW05] et les ontologies [48].

Le travail de [46] utilise des techniques de regroupement sur les données représentant l'écart moyen et standard du volume de trafic, de la vitesse et du temps de progression afin de prédire les accidents dans un modèle de trafic autoroutier en temps réel simulé.

L'étude de [47] compare trois méthodes, à savoir les arbres de décision, l'algorithme naïf bayésien et la méthode du k-voisins le plus proche pour classer la gravité de l'accident (par exemple, une blessure grave, une légère blessure et dommages matériels). Les variables d'entrée sont la ville (où l'accident s'est produit), la zone particulière (école, marché), la séparation des routes, l'orientation routière, le type de jonction entre les routes, le type de surface de la route, les conditions de la route, les conditions météorologiques et les conditions de luminosité.

Les trois méthodes ont abouti à des résultats très similaires (dans une différence de 0,8%, environ 80% de performance) en termes d'exactitude, la méthode du k-voisins le plus proche étant la première et L'algorithme naïf bayésien la dernière. Jagannathan et al. [49] utilise un système de raisonnement basé sur des cas pour prédire l'issue des conditions du trafic en fonction des cas historiques qui ont entraîné des accidents et pour différencier les conditions d'accident et de non-accident/ pas accident.

Quatre groupes d'attributs sont utilisés dans le processus de prise de décision, ce sont: (1) les attributs nominaux: heure, intervalle de temps / distance ; (2) attributs de point unique: nombre moyen de véhicules (débit) sur toutes les voies, vitesse moyenne de véhicules sur toutes les voies, progression (distance)

moyenne entre les véhicules sur toutes les voies, l'occupation moyenne mesurée par le capteur, l'écart type dans le nombre moyen de véhicules (flux) entre les voies, l'écart type de la vitesse moyenne des véhicules entre les voies, l'avancement des véhicules sur toutes les voies et l'écart type dans l'occupation des véhicules entre toutes les voies; (3) les attributs temporels: variation de la vitesse moyenne sur toutes les voies et variation du débit moyen sur toutes les voies; et (4) les attributs spatiaux: variation de la vitesse moyenne sur toutes les voies et variation du débit moyen sur toutes les voies. La similitude est calculée à l'aide de l'algorithme du k-voisins le plus proche et est basée sur des combinaisons de similarités d'attributs dans la mesure de similarité. La mesure de similarité est donc calculée comme la moyenne de la similitude de la catégorie horaire, de la similitude basée sur les attributs spatio-temporels d'un seul point et de la similarité basée sur la variation de la circulation sur la distance ou l'intervalle de temps.

Une version améliorée du regroupement spatial basée sur la densité des applications avec du bruit (*density-based spatial clustering of applications with noise*(DBSCAN)) qui tient compte du nombre d'accidents et de leur niveau de gravité est proposée dans [48].

Les auteurs proposent également un cadre de cartographie des risques liés aux accidents basé sur l'ontologie, dans lequel l'ontologie représente les connaissances du domaine liées aux accidents de la route et prend en charge l'extraction des données en fonction des besoins des utilisateurs.

2.6. Sélection de variables pour l'apprentissage

Comme indiqué dans l'introduction, en raison du fait qu'un grand nombre de variables caractérisent les accidents de la circulation, des variables appropriées devront être sélectionnées afin d'identifier les plus remarquables, et donc permettre l'apprentissage et la prédiction des systèmes experts pour la détection des collisions et prédiction.

Des solutions diverses ont été utilisées dans la littérature dans ce contexte, y compris : le critère d'information d'Akaike, le critère d'information bayésien, les arbres de décision, à savoir l'arbre de décision CART [27] et les forêts aléatoires [50] et le sélecteur des caractéristiques basé sur la corrélation (*Correlation-basedFeatureSelector*) qui propose une mesure heuristique du mérite du sous-ensemble de fonctionnalités [28].

Hossein pour et al. [30] utilisent trois critères, à savoir le critère d'information d'Akaike, le critère d'information Bayésien et le Cp de Mallows pour sélectionner le meilleur sous-ensemble de variables (celui qui obtient les valeurs les plus petites pour ces critères). En particulier, les valeurs sont calculées sur une série de 7 variables et leurs différentes combinaisons.

Un sélecteur des caractéristiques basé sur la corrélation simplifiée est également le coefficient d'information maximale. Ce coefficient capture une large gamme d'associations non limitées à des types de fonctions spécifiques (par exemple, linéaires, exponentielles ou périodiques) ou même à toutes les relations fonctionnelles [51].

Divers arbres de décision ont également été employés afin de sélectionner des variables appropriées. Dans [27], un arbre de décision CART est utilisé pour sélectionner les variables contributives les plus importantes avant l'entraînement d'un SVM pour l'évaluation du risque de collision en temps réel. L'importance d'une variable est calculée en fonction du nombre de fois où cette variable est apparue et sa position relative dans l'arbre.

2.7. Conclusion sur l'état de l'art

En conclusion, on peut constater qu'en raison de son importance, le thème de l'analyse et de la prédiction des accidents de la circulation a suscité l'intérêt de nombreux chercheurs. Des méthodologies et des techniques intéressantes ont été développées, dont les 12 mesures et méthodes utilisées pour évaluer la sécurité routière. Bien que ces mesures comptent quelques inconvénients, leur avantage principal est qu'elles sont précises et assez simples à utiliser.

D'autres mesures comme le calcul du risque collectif, du risque personnel, du niveau de service de sécurité et de l'indice de risque routier s'avèrent très importantes dans la détermination des accidents DSi (*Death and Severeinjury*), ou les accidents mortels et à blessures graves.

Des algorithmes tels que les réseaux de neurones, les machines à vecteur de support (SVM), les arbres de décision et la régression ont fait leur preuve dans la prédiction des accidents de la route dans plusieurs publications.

Enfin l'analyse de la tendance des accidents, des types et des causes des accidents nous a permis de déterminer qu'en plus de la cause humaine les accidents de la route sont la résultante de plusieurs facteurs dont les facteurs climatiques et environnementaux.

CHAPITRE 3 : METHODOLOGIE

Conformément aux objectifs que nous nous sommes fixés, dans ce chapitre, nous allons décrire les différentes méthodologies que nous allons utiliser pour les atteindre. Nous allons tout d'abord aborder l'une des méthodologies de fouille de données les plus utilisées, appelée la méthode CRISP-DM. Nous allons ensuite décrire les données que nous allons utiliser dans ce travail et les diverses techniques adoptées, à savoir le réseau de neurones, l'arbre de décision, la machine à vecteur de support (SVM), l'arbre de décision « *gradient boosted tree* », l'algorithme naïf bayésien, l'algorithme du k-voisins le plus proche et l'algorithme AdaBoost.

3.1. CRISP-DM

Le modèle CRISP-DM de son acronyme *Cross Industry Standard for Data Mining* [52], est un ensemble d'étapes à suivre afin de permettre aux chercheurs de résoudre le problème de fouille de données. Ce modèle forme un cycle composé de 6 phases comme le montre la figure 4.

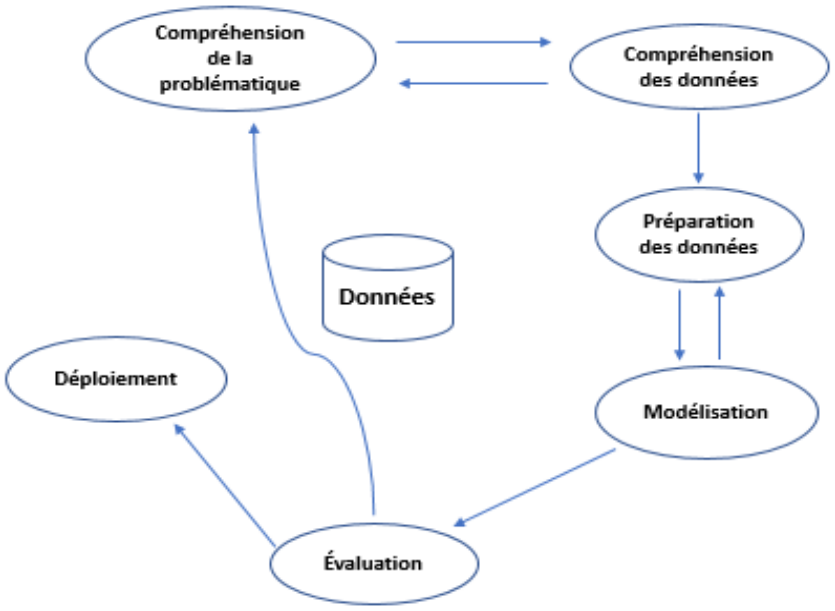


Fig.4: Modèle CRISP-DM (adapté de [53]).

Dans ce schéma, on peut constater la présence de six blocs fonctionnels. Leur interprétation peut être expliquée comme il suit :

- La compréhension de la problématique : cette étape se base sur la formation du problème de fouille de données. Pour le chercheur, il s'agira de convertir le projet d'une idée de recherche à une problématique de fouille de données et de comprendre les besoins du problème identifié (au besoin en collaboration avec le bénéficiaire).

- La compréhension des données : dans cette étape, il s'agit de chercher ou de collecter les données nécessaires. Ces données doivent être ensuite traitées de manière qualitative et quantitative afin de formuler les premières hypothèses et d'identifier des corrélations. Cela permettra de sélectionner les données et des variables appropriées pour la modélisation.
- La préparation des données : cette étape constitue une étape cruciale dans le processus d'analyse de données. Elle consiste à mettre les données constituées à l'étape antérieure dans le format idéal qui correspondrait aux différents outils qui seront utilisés dans le processus de fouille de données.
- La modélisation des données : il s'agit dans cette étape d'appliquer aux données recueillies et formatées différents modèles de fouille de données. Il se peut qu'à cette étape on soit obligé de revenir à l'étape précédente pour reformater les données en fonction des spécificités du modèle utilisé.
- L'évaluation : dans cette étape il s'agit d'évaluer les performances du modèle afin d'identifier celui qui donne un résultat avoisinant le plus possible les objectifs du projet. Le(s) modèle(s) sera (seront) appliqué(s) à de nouvelles données de test afin de s'assurer que le modèle est capable de généraliser.
- Le déploiement : a lieu une fois qu'un modèle suffisamment performant est identifié.

Un sondage réalisé en 2014montre que 43% des entreprises utilisent la méthodologie CRISP-DM pour leur problématique de fouille de données [54]. Dans ce même sondage, la méthodologie la plus citée après CRISP-DM est la méthode SEMMA (de son acronyme *Sample, Explore, Modify, Model, Assess*) avec 8.5% [54].

Nous pouvons utiliser SEMMA pour mieux comprendre les activités spécifiques qu'un analyste de données exécute à chaque étape de son travail de recherche, tel que montré à la figure 5.Nous distinguons ainsi 5 étapes avec SEMMA :

- L'échantillonnage permet d'extraire de l'ensemble des données une partie à utiliser. Cette partie doit être assez grande et significative pour permettre l'identification des motifs dans des données. Comme le montre la figure 5 l'échantillonnage correspond à l'étape de la compréhension de données dans le modèle CRISP-DM.
- L'exploration des données consiste en une visualisation de celles-ci ainsi qu'à voir leur distribution. Comme le montre la figure 5, l'exploration des données correspond à l'étape de la compréhension de données dans le modèle CRISP-DM.

- La modification de données consiste à les formater, à structurer l'information de manière à ce qu'on puisse en tirer profit. La modification des données correspond à l'étape de la préparation de données dans le modèle CRISP-DM.
- La modélisation consiste à appliquer les différents modèles de fouille de données, comme les réseaux de neurones, les arbres de décision, et autres modèles sur les données formatées et structurées. Cette étape correspond à la modélisation des données dans le modèle CRISP-DM.
- L'étape d'évaluation consiste à évaluer les performances du modèle utilisé. Comme le montre la figure 5 cette étape correspond à l'étape d'évaluation des données dans le modèle CRISP-DM.

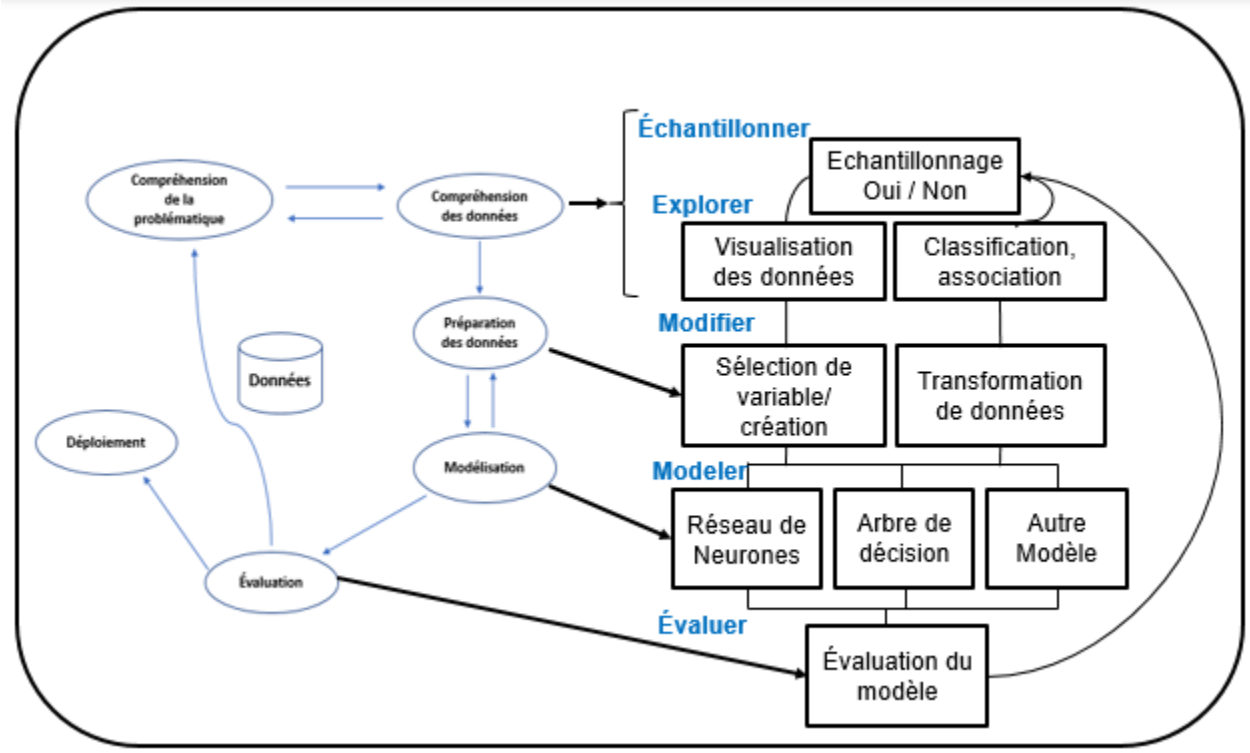


Fig.5 : CRISP-DM avec SEMMA

Ce projet de recherche est réalisé en suivant les différentes étapes du CRISP-DM et en utilisant les activités précisées par SEMMA. Ce sont la compréhension de la problématique se trouvant dans l'introduction et l'état de l'art, la compréhension des données et la préparation des données dans la section 3.2, et l'étape de la modélisation et de l'évaluation dans la section 3.3 et le chapitre 5.

3.2. Compréhension et préparation des données

3.2.1. La compréhension des données

Dans ce travail de recherche nous allons utiliser deux bases de données, à savoir une base de données fournie par le Service de police de la ville d'Ottawa portant sur l'année 2013 et une autre base de données fournie par la ville d'Ottawa portant sur les années 2014, 2015 et 2016[76].

3.2.1.1. Les données de 2013

La base de données de 2013 est un enregistrement de 26 723 lignes et 39 attributs portant sur les accidents qui se sont produits dans la ville d'Ottawa au courant de l'année 2013. Dans le tableau 7 se trouve une description des différentes variables de cette base de données.

Tableau 7: Les variables de la base de données 2013

Nom de la variable	Description	Nature de la variable	Nombre de valeurs uniques
Occ_date	Date à laquelle l'accident s'est produit	Catégorique	365
H	Heure à laquelle l'accident s'est produit	Numérique	24
M	Minute à laquelle l'accident s'est produit	Numérique	60
Em	Seconde à laquelle l'accident a eu lieu	Numérique	60
Location	L'adresse où l'accident s'est produit	Catégorique	10948
Roadway1c	Le nom de la rue.	Catégorique	2711
Place_name	Nom de la place publique si toutefois l'accident s'est produit dans une place publique ou à un établissement connu	Catégorique	1577
TranslationProperAll	Cette variable porte sur le type de l'accident et les caractéristiques de l'accent comme par exemple chauffeur ayant bu ou non, etc.	Catégorique	1689
AccidentType	Contient les différents types d'accident à savoir : accident fatal, accident avec blessure, accident avec dommage matériel ou autre	Catégorique	4
AccFatal	Permet de savoir l'accident a été fatal. Prend 1 si oui et 0 si non.	Numérique	2

Nom de la variable	Description	Nature de la variable	Nombre de valeurs uniques
AccInjuriesTra	Permet de savoir si l'accident compte un blessé. Prend 1 si oui et 0 si non.	Numérique	2
AccPropDamage	Permet de savoir s'il y'a eu un dommage matériel dans l'accident. Prend 1 si oui et 0 si non.	Numérique	2
AccNonReport	Permet de savoir si l'accident a été reporté. Prend 1 si oui et 0 si non.	Numérique	2
ImpairedOver.0.8	Conducteur ayant un taux d'alcool supérieur à 0.8.	Numérique	2
FailToRemain	Détermine si le conducteur est resté surplace. Prend 1 si oui et 0 si non.	Numérique	2
VehicleTowed	Détermine si le véhicule a été remorqué. Prend 1 si oui et 0 si non.	Numérique	2
VehicleAbandon	Détermine si le véhicule impliqué dans l'accident a été abandonné. Prend 1 si oui et 0 si non.	Numérique	2
TrafficComplaint	Détermine s'il y'a eu des plaintes concernant la circulation. Prend 1 si oui et 0 si non.	Numérique	2
ProvOffense	Précise si l'accident fait état d'une offense provinciale. Prend 1 si oui et 0 si non.	Numérique	2
Susp90	Cas où le chauffeur a été suspendu pendant 90 jours. Prend 1 si oui et 0 si non.	Numérique	2
Careless Driving	Façon dont le chauffeur conduisait : excès de vitesse, etc…	Numérique	2
Poss Cannabis	Cas ou le chauffeur possédait du cannabis. Prend 1 si oui et 0 si non.	Numérique	2
MotorTheft	Détermine si le véhicule était volé ou pas. Prend 1 si oui et 0 si non.	Numérique	2
XCoordinate	Représente la coordonnée géographique X du lieu de l'accident.	Numérique	7296
YCoordinate	Représente la coordonnée géographique Y du lieu de l'accident.	Numérique	7296

Nom de la variable	Description	Nature de la variable	Nombre de valeurs uniques
District	Le district dans lequel l'accident a eu lieu.	Numérique	6
Atom	La division du district dans laquelle l'accident s'est produit.	Numérique	582
Location	Description du lieu de l'accident, comme par exemple si c'est un espace de stationnement ou sur une route, etc.	Catégorique	26
VehicleCount	Détermine le nombre de véhicules impliqués dans l'accident	Numérique	12
AccidentCount	Détermine s'il y'a un accident.	Numérique	2
Jour_char	Détermine le jour de la semaine pendant laquelle l'accident a eu lieu.	Catégorique	7
Mois_char	Détermine le mois de l'année pendant lequel l'accident a eu lieu.	Catégorique	12
MinTemp	Détermine la température minimale le jour où l'accident a eu lieu.	Numérique	219
MeanTemp	Détermine la température moyenne le jour où l'accident a eu lieu.	Numérique	244
MaxTemp	Détermine la température maximale du jour où l'accident a eu lieu.	Numérique	229
Rain	Détermine la quantité de pluie le jour de l'accident.	Numérique	69
Snow	Détermine la quantité de neige le jour de l'accident.	Numérique	31
Rainc	Détermine s'il a plu le jour de l'accident. Prend la valeur 1 si oui et 0 si non.	Numérique	2
Snowc	Détermine s'il a neigé le jour de l'accident. Prend la valeur 1 si oui et 0 si non.	Numérique	2

Les données météorologiques ont été ajoutées à partir d'une base de données téléchargée du site du gouvernement[56]. Ce processus est décrit plus en détail dans la section 3.2.2.

Un graphe représentant la pertinence des variables selon la classification accident/pas accident se trouve à la figure 6. Notons que ce graphique a été obtenu en utilisant l'option importance de l'algorithme forêt aléatoire dans l'onglet « Model » de la librairie Rattle. La dernière est décrite en détails à la section 3.3.

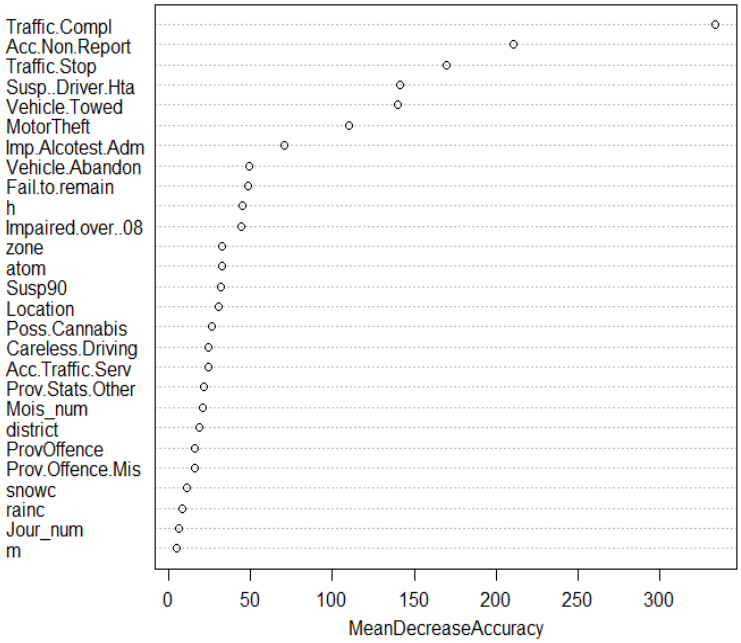


Fig.6. Graphe d'importance des variables selon la classification accident/pas accident

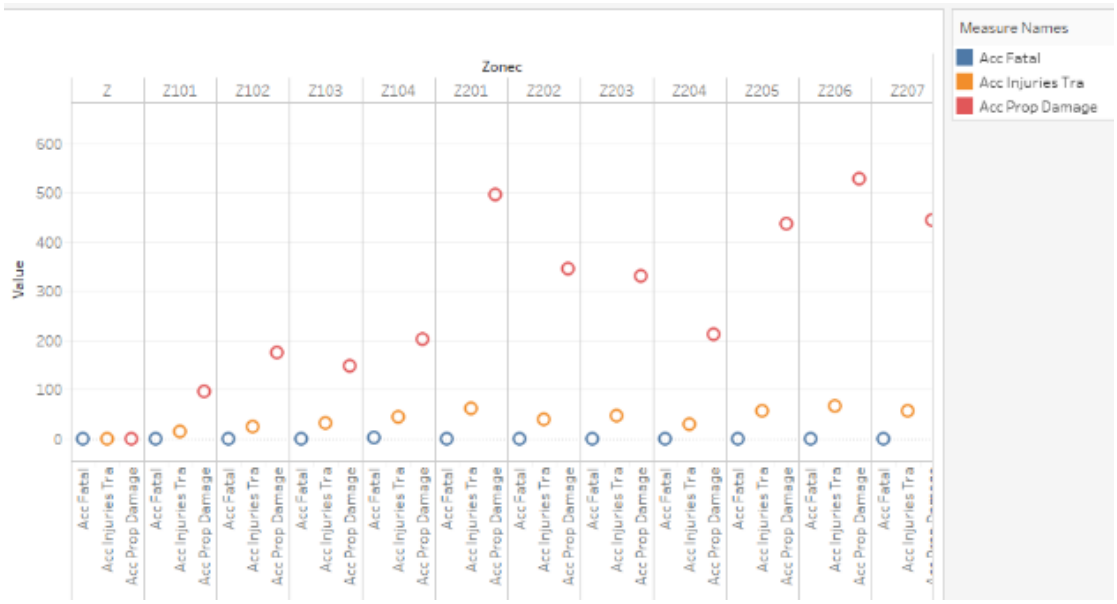
On peut remarquer sur la figure 6 que la variable qui a la plus grande importance dans ce modèle sur la précision de la prédiction est la variable "Traffic_Compl" qui représente le dépôt d'une plainte au sujet de la circulation. Les variables sont donc classées dans l'ordre décroissant, selon leur importance, de la variable "Traffic_Compl" à la variable "m" qui représente la minute à laquelle l'accident s'est produit. Les variables les plus importantes sont telles que lorsqu'on les enlève du jeu de données le risque d'erreur de prédiction augmente.

Certaines de ces variables sont connues lors de l'accident, telles que la densité de la circulation, l'heure, la zone dans laquelle l'accident a lieu etc., tandis que d'autres variables ne sont connues qu'après l'accident telles que le taux d'alcoolisme, les caractéristiques de l'accident, le dépôt d'une plainte au sujet de la circulation etc.

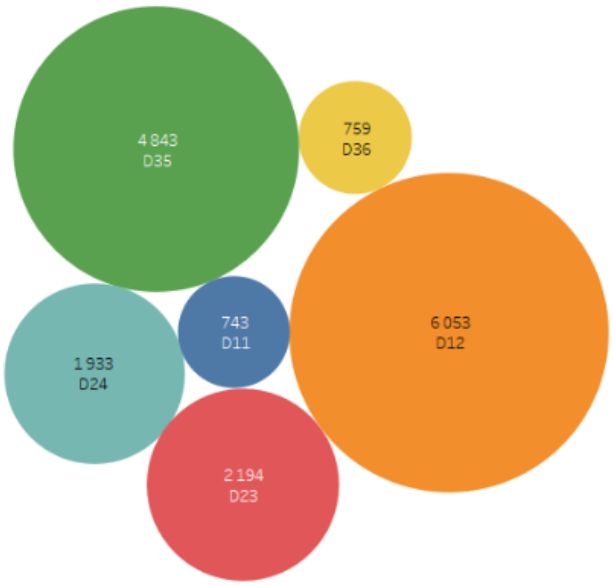
On peut ainsi constater que parmi les variables qui sont connues lors d'un accident les plus importantes sont en ordre : l'heure, la zone de l'accident, l'atome (division du district dans lequel l'accident a eu lieu), la localisation, ou 'location' à la figure (la rue dans laquelle l'accident s'est produit), le mois, le district dans lequel l'accident s'est produit, le fait qu'il y ait de la pluie ou pas, le fait qu'il est neigé ou pas, le jour, et la

minute. Ces dernières pourront être ainsi utilisées pour la prédiction des accidents. Les variables connues après l'accident ne seront pas utilisées dans la prédiction.

Afin de mieux comprendre les données, nous faisons aussi appel à un outil de visualisation. À l'aide du logiciel Tableau [55], nous arrivons à obtenir diverses représentations graphiques de la base de données pour une meilleure compréhension sur la nature de données. Dans la figure 7 par exemple, nous pouvons constater la distribution des accidents par zone et par district.



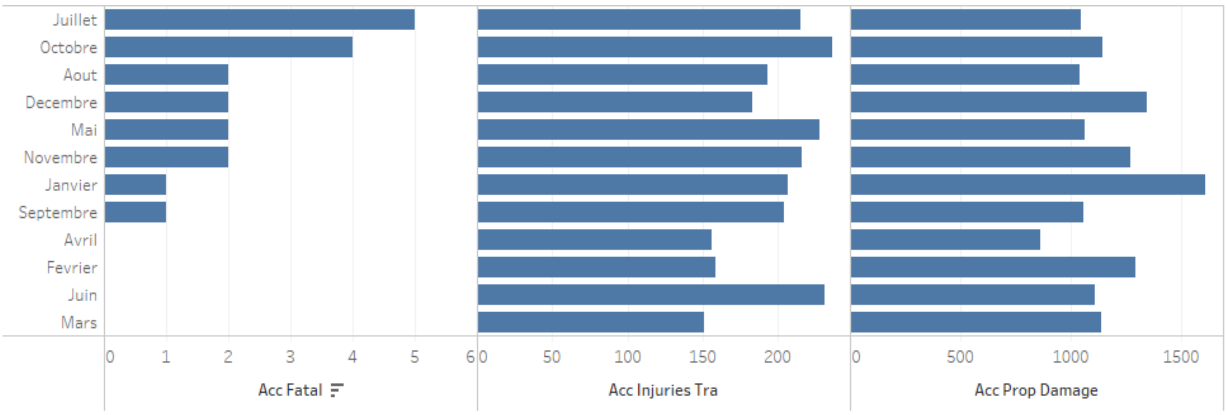
(a)



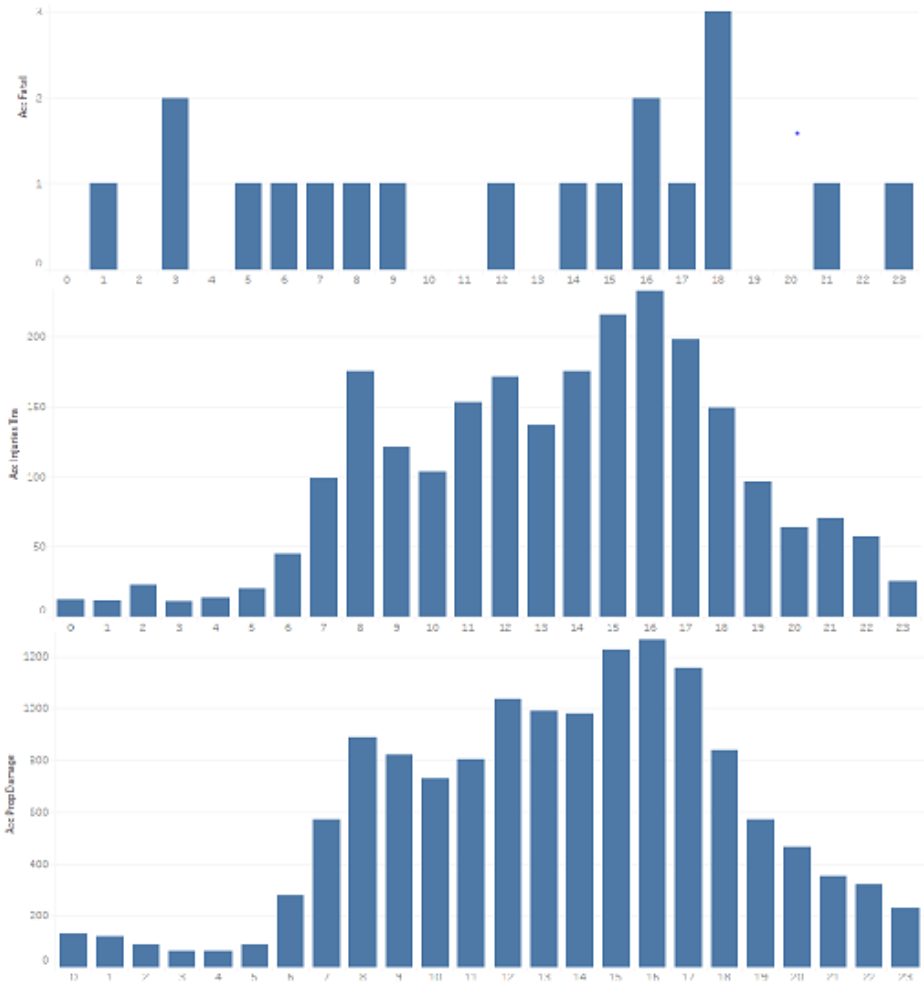
(b)

Fig.7 (a) Distribution des accidents par zone, (b) Distribution des accidents par district

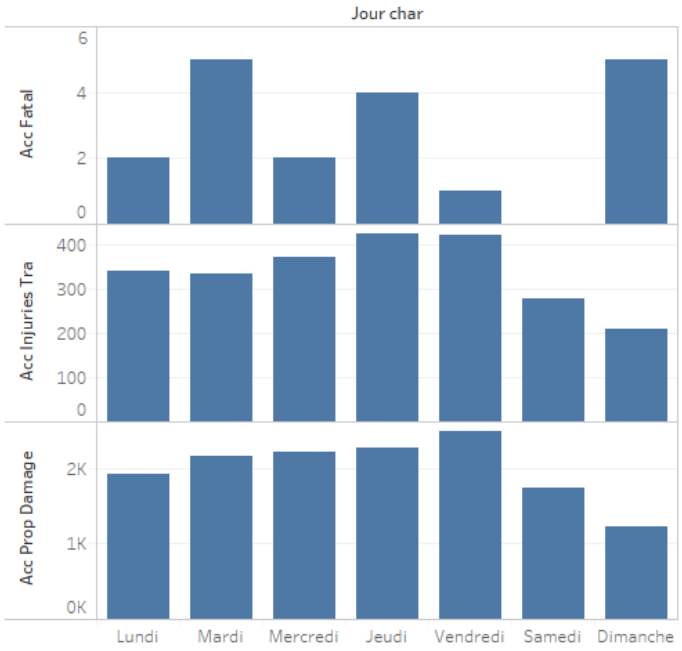
Nous pouvons constater sur la figure 7b que le district D12 compte le plus grand nombre d'accidents, suivi respectivement par les districts D35, D23, D24, D36 et D11.



(a)



(b)



(c)

Fig.8 Distribution des accidents : (a) par mois, (b) selon l'heure, et (c) selon le jour de la semaine.

Dans un premier temps, nous pouvons remarquer sur la figure 8a, que conformément aux tendances évoquées dans la section 2.4 un nombre plus important d'accidents semble se dérouler pendant les périodes d'automne et d'hiver, la période de janvier comptant le plus grand nombre d'accidents, suivie successivement par décembre, novembre et octobre. Cependant le mois de juillet compte le plus grand nombre d'accidents meurtriers, suivi par le mois d'octobre.

En second lieu, nous constatons sur la figure 8b qu'un très grand nombre d'accidents se produisent entre 15h et 17h. Toutefois, les accidents les plus meurtriers se produisent vers 18h. Les informations obtenues de la Figure 88b viennent donc confirmer les tendances évoquées dans la section 2.4.

Enfin la figure 8c nous démontre que le vendredi est le jour de la semaine qui compte le plus grand nombre d'accidents routiers, suivi en ordre par le jeudi, mercredi, le mardi, le lundi, le samedi et le dimanche. Le mardi et le dimanche se trouvent être les jours les plus meurtriers.

Ces informations viennent confirmer en grande partie les tendances évoquées dans la section 2.4 sauf que dans nos données le jeudi vient avant le mercredi contrairement aux tendances évoqué**e**s dans la section 2.4. Cela peut être à cause du fait qu'on ne regarde que les données de l'année 2013, tant que la section 2.4 analyse des tendances plus générales sur les accidents routiers.

La figure 9 est une illustration de la distribution des accidents en fonction du taux d'alcoolisme (chauffeur avec un taux d'alcoolisme plus haut que 0.08) selon les jours de la semaine et le mois.

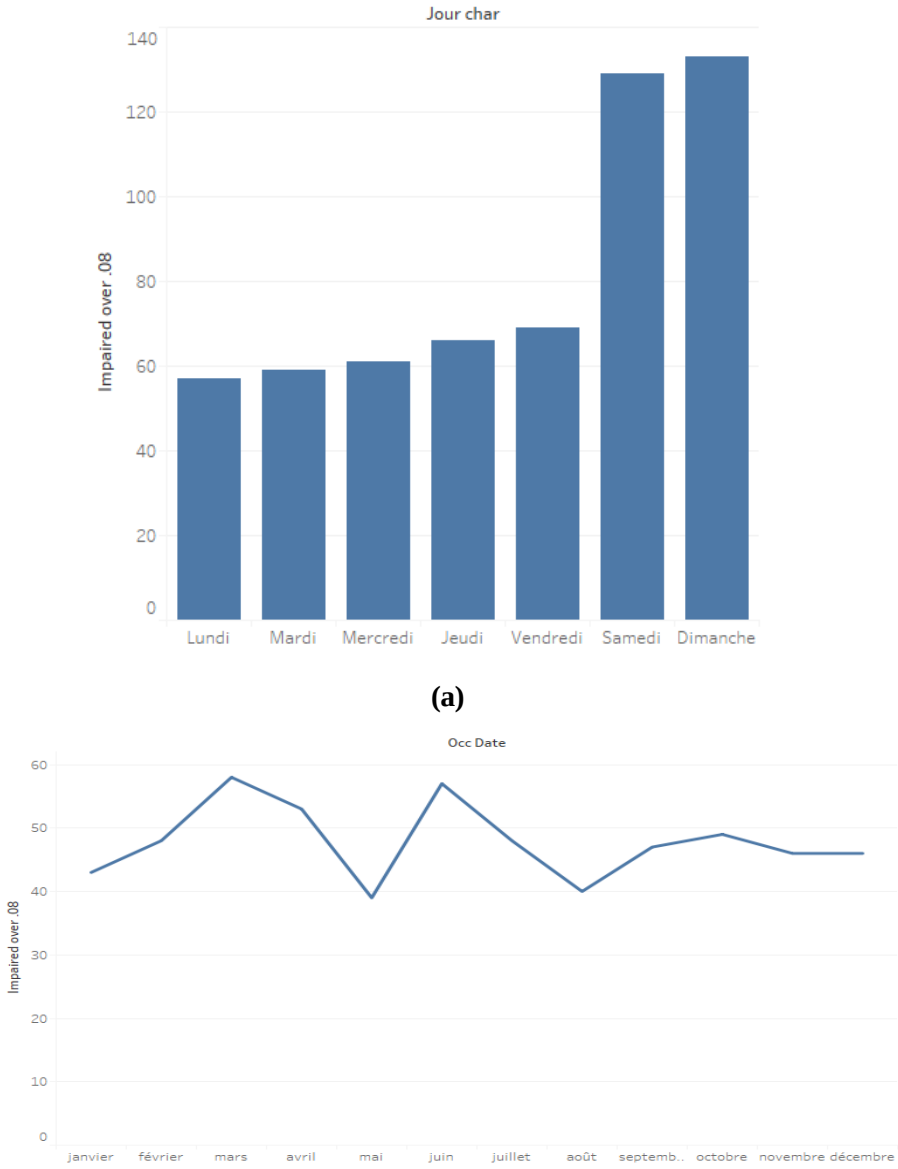
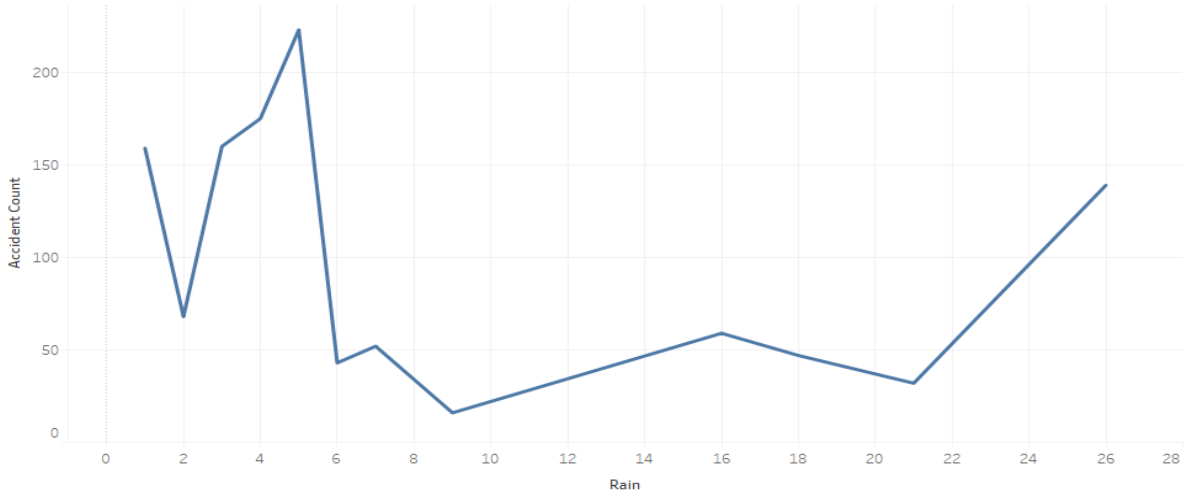
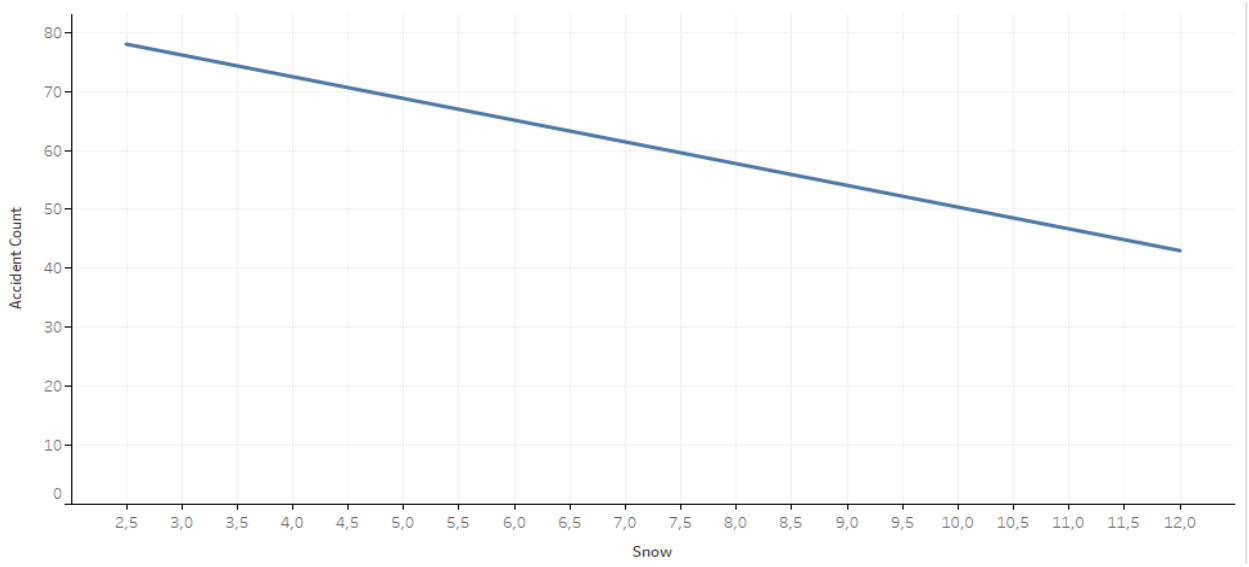


Fig.9 Distribution des accidents en fonction du taux d'alcoolisme selon : (a) le jour de la semaine, et (b) le mois

Dans la figure 9a on constate un nombre plus élevé d'accidents les fins de semaine, ce qui est très normal vu que la majorité du monde sort plus en fin de semaine pour des activités sociales. On peut également remarquer sur la même figure que ce nombre d'accidents de la route est d'ailleurs décroissant du dimanche au lundi. Cela nous permet donc de tirer la conclusion que plus on tend vers la fin de semaine plus le nombre d'accidents dus au taux d'alcoolisme augmente.



(a)



(b)

Fig.10. Distribution du nombre d'accidents : (a) selon la quantité de pluie, et (b) selon la quantité de neige

Le graphe de la Figure 9b nous permet d'observer les variations du nombre d'accidents selon le taux d'alcoolisme des conducteurs à travers les mois de l'année. Nous pouvons remarquer sur le graphe un pic durant les mois de mars et de juin. L'hypothèse expliquant cela peut être le changement de climat. En effet le mois de mars fait état de transition entre l'hiver (où il fait très froid au Canada en général) et le printemps (où la température est un peu plus douce).

Cette transition pousse beaucoup plus de monde à sortir, et donc à boire, ce qui pourrait expliquer ce pic pendant ce mois-là. Le taux d'accidents baisse ensuite entre le mois de mars et celui de mai, puis recommence à monter pour atteindre un second pic durant le mois de juin, qui se trouve dans la période de l'été. Selon nos données, le second mois où il y a moins d'accidents dû au taux d'alcoolisme est le mois d'août. Au mois d'août le taux baisse et remonte par la suite, pour ensuite s'équilibrer entre les périodes de septembre à janvier.

Les conditions climatiques sont donc un facteur déterminant dans l'occurrence des accidents de la route. La figure 10 est une illustration de la distribution des accidents de la route en fonction des conditions météorologiques. La figure 10a illustre la courbe du volume d'accidents en fonction de la quantité de pluie. On constate que le sommet de la courbe se situe entre 4 et 5 mm de pluie. Par contre dans la figure 10b on constate une courbe décroissante en fonction de la quantité de neige. D'après les données dont nous disposons, plus la quantité de neige est abondante, moins il y a d'accidents. Cela peut être expliqué par le fait que moins des véhicules se trouvent sur les routes lors des tempêtes hivernales et que les chauffeurs tendent à être beaucoup plus prudents dans ces conditions.

En conclusion nous pouvons affirmer que nos données correspondent aux tendances générales évoquées dans la section 2.4. Ces tendances sont reliées à plusieurs facteurs, comme le nombre de voitures en circulation (un plus grand nombre d'accidents se produit pendant les heures de pointe), les conditions climatiques et l'état psychologique des conducteurs. Les Figures 9a et 9b illustrent en effet à quel point l'état psychologique du conducteur est un facteur déterminant dans les accidents de la route comme évoqué dans la section 2.5.

3.2.1.2. Les données de 2014 à 2016

La base de données de 2014 à 2016 [76] est un enregistrement de 43 944 lignes et 19 attributs. Dans le tableau 8 se trouve une description des différentes variables de cette base de données. Sur la figure 11, nous pouvons remarquer le graphe d'importance des variables pour la classification selon les types d'accidents. La variable cible est la variable *collision classification*. Toutes les variables présentées sur la figure sont des variables disponibles avant l'accident, qui peuvent donc être utilisées pour la prédiction. Nous pouvons remarquer que les variables les plus importantes sont celles qui ont le poids (*weight*) le plus élevé.

Tableau 8: Les variables de la base de données 2014 à 2016

Nom de la variable	Description	Nature de la variable	Nombre de valeurs uniques
Location	Lieu ou intersection où à laquelle l'accident a eu lieu	Catégorique	10345
CoordX	Coordonnées X du lieu de l'accident	Numérique	38101
CoordY	Coordonnées Y du lieu de l'accident	Numérique	38975
Date	Date à laquelle l'accident a eu lieu	Catégorique	1096
Time	Heure et minute à laquelle l'accident a eu lieu	Catégorique	1424
Environment	Conditions environnementales dans lesquelles l'accident a eu lieu. Elle prend 9 valeurs distinctes à savoir : dry, fog, clear, drifting snow, freezing rain, rain, snow strong wind, unknown	Catégorique	9
Road_Surface	Les conditions de la route quand l'accident s'est produit. Elle comprend 11 valeurs distinctes à savoir : unknown, dry, wet, loose snow, slush, packed snow, ice, mud, loose sand or gravel, spilled liquid, Other	Catégorique	11
Traffic_control	Porte sur les signalisations de la route. Elle comprend 9 valeurs distinctes, à savoir : traffic signal, stop sign, yield sign, pedestrian crossing, school bus, traffic gate, traffic controller, no control, round about	Catégorique	9
Collision_Location	Type d'endroit auquel l'accident a eu lieu. Elle comporte 9 valeurs distinctes, à savoir: at intersection, at railway crossing, at private drive, intersection related, non-intersection, overpass or bridge, underpass or tunnel.	Catégorique	9
Light	Lumière de jour quand l'accident s'est produit. Elle comprend 5 valeurs distinctes, à savoir : dark, dawn, daylight, dusk, unknown.	Catégorique	5
Collision_classification	Type de collision. Elle comprend 3 valeurs distinctes à savoir : accident	Catégorique	3

Nom de la variable	Description	Nature de la variable	Nombre de valeurs uniques
	fatal, accident avec blessures et accident avec dommage matériel.		
Impact_type	Type d'impact	Catégorique	8
Acc_fatal	Dit si l'accident est de type fatal ou pas. Elle prend 2 valeurs uniques, à savoir : oui ou non	Catégorique	2
Acc_injury	Dit si l'accident est de type accident avec blessures ou pas. Elle prend 2 valeurs uniques : oui et non.	Catégorique	2
Acc_propDamage	Dit si l'accident est de type accident avec dommages matériels ou pas. Elle prend 2 valeurs uniques : oui et non.	Catégorique	2
H	Heure à laquelle l'accident s'est produit	Numérique	24
M	Minute à laquelle l'accident s'est produit	Numérique	60
Jour_char	Détermine le jour de la semaine pendant lequel l'accident a eu lieu.	Catégorique	7
Mois_char	Détermine le mois de l'année pendant lequel l'accident a eu lieu.	Catégorique	12

Comme on peut constater dans la figure, les variables les plus importantes sont donc les variables *coordX* et *coordY* (*X* et *Y* dans la figure) représentant les coordonnées de l'accident suivies par les variables liées à la date (*Date* dans la figure) et le temps (*Time* dans la figure). Ce graphe a été obtenu grâce à l'option importance de l'algorithme forêt aléatoire. Le processus Rapid Miner utilisé pour l'obtenir est représenté à la figure 12. Rapid Miner est un logiciel de fouille de données servant à la préparation des données, à l'apprentissage automatique et aux déploiements de modèles prédictifs [77] qui sera présente à la section 3.3.1.

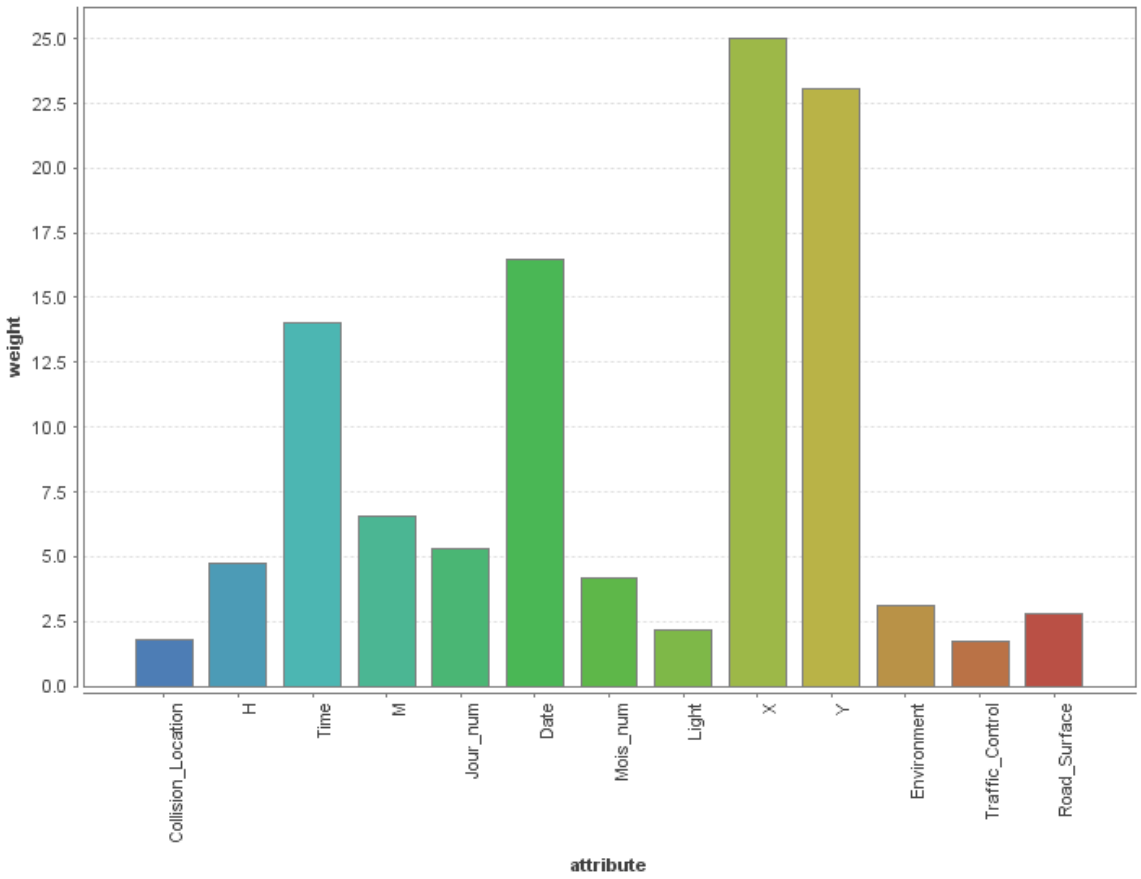


Fig.11 : Importance des variables selon la classification des types d'accidents

Dans la figure 12, le rôle des différents blocs se définit comme suit : Le bloc "Retrieve" est le bloc d'entrée qui contient le fichier de la base de données portant sur les accidents dans le format *csv*. Les attributs à utiliser dans le jeu d'entraînement ont été sélectionnés en utilisant le bloc "Select Attributes". Dans le bloc "Set Role", on choisit la variable cible.

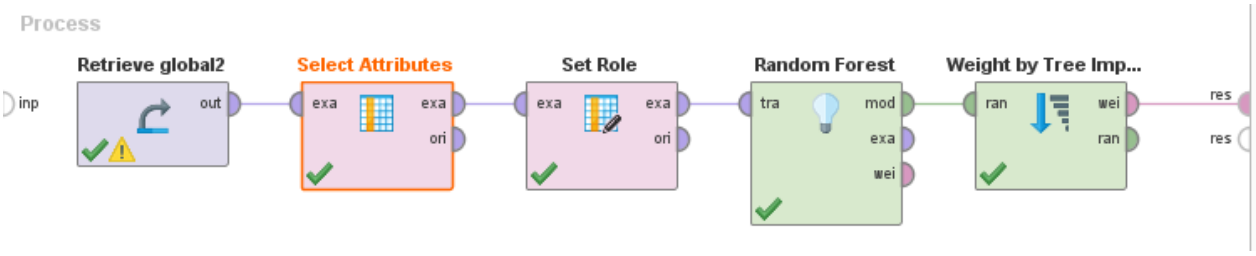
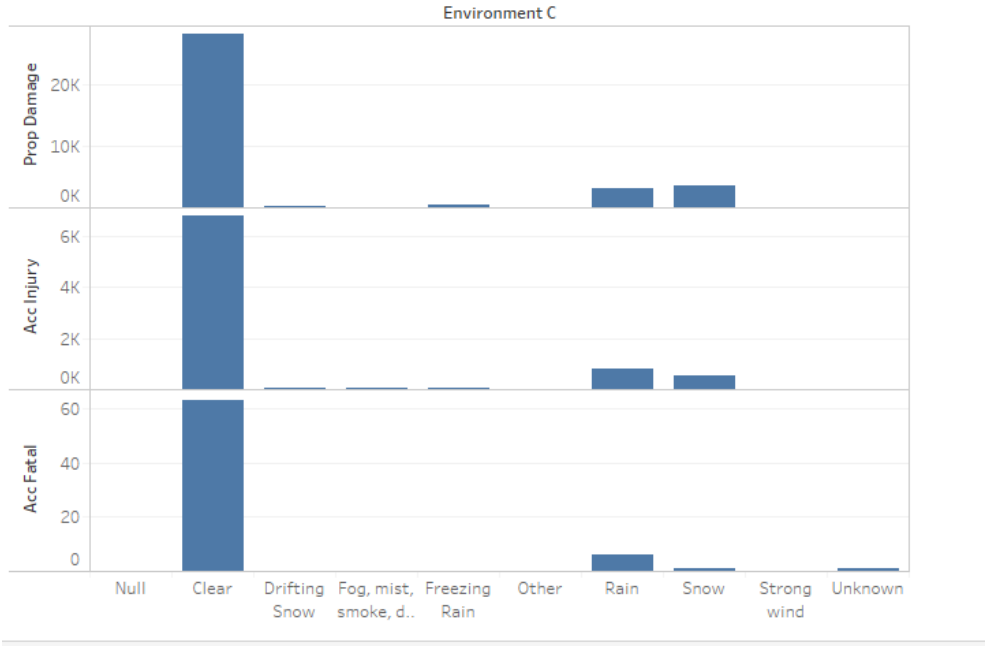


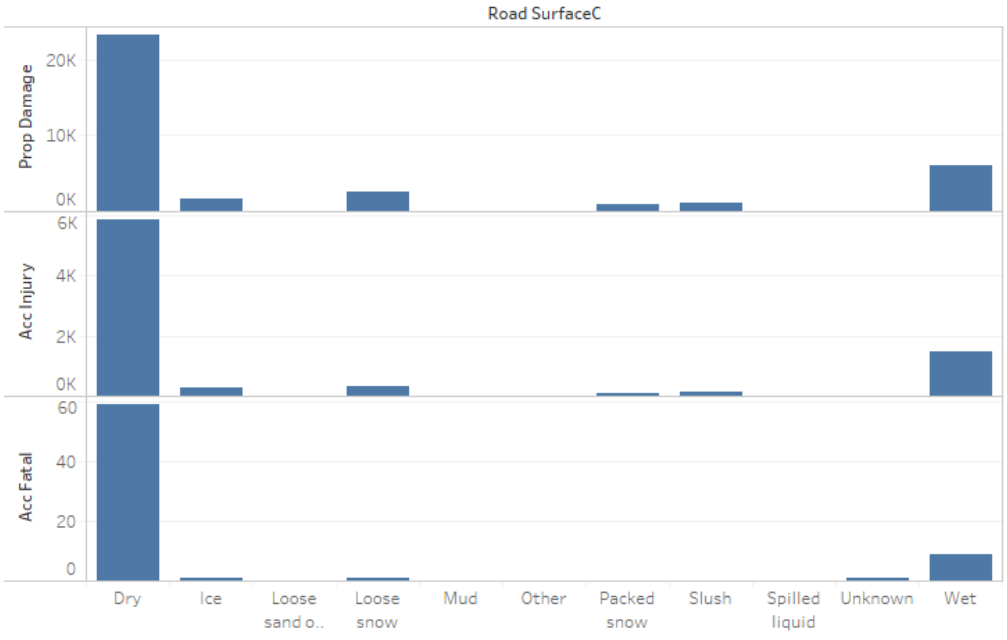
Fig.12 : Processus RapidMiner pour le calcul de l'importance des variables

Le bloc "Random Forest" montre que nous utilisons l'algorithme forêt aléatoire. Et enfin le bloc "Weight by tree importance" associe un poids à chaque variable ; plus le poids est élevé plus la variable est importante.

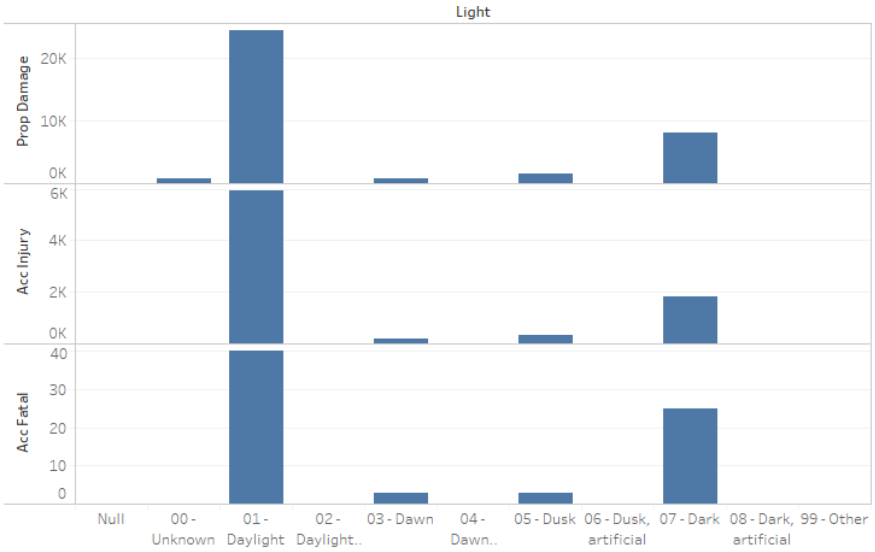
Nous faisons encore appel au logiciel Tableau [55], afin de pouvoir obtenir des représentations graphiques de la base de données pour une meilleure compréhension de celle-ci. Nous pouvons remarquer sur la figure 13 que la plupart des accidents de la route surviennent en pleine journée sur des routes sèches et dans de bonnes conditions météorologiques avec une vue claire et dégagée. Quelques cas d'accidents sont aussi observés dans la nuit en pleine obscurité ou lorsque la route est humide à cause de la pluie.



(a)



(b)



(c)

Fig.13. Distribution du nombre d'accidents selon : (a) l'environnement, (b) la surface de la route, (c) la lumière du jour

Tel qu'illustré à la figure 14, tout comme pour les données de 2013, nous constatons que la plupart des accidents de la route surviennent à l'heure de pointe, entre 15h et 17h. Seize heures est l'heure à laquelle on constate le plus grand nombre d'accidents. Ceux-ci correspondent donc aux tendances évoquées à la section 2.4.

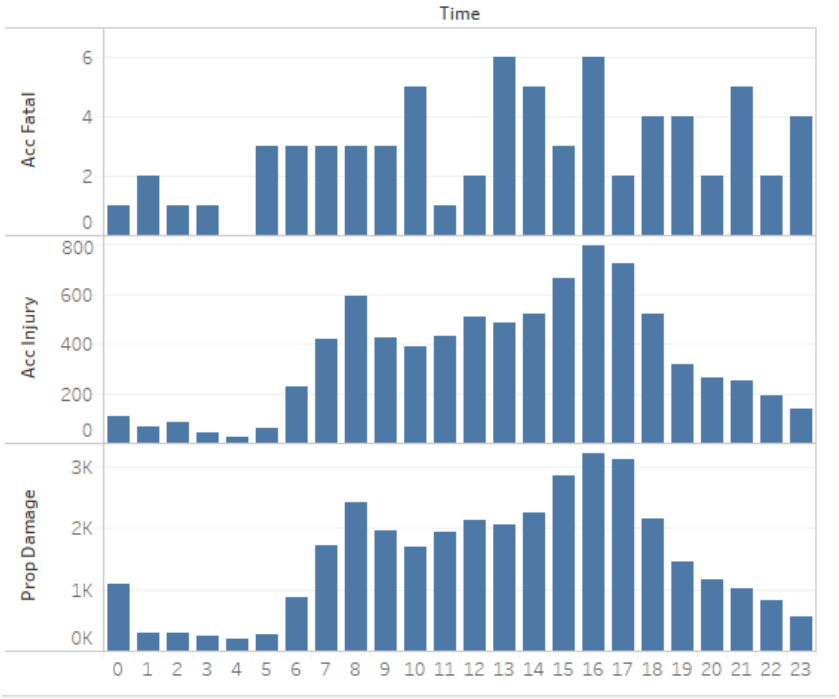


Fig.14 : Distribution des accidents selon l'heure

La figure 15 nous montre une distribution des accidents selon les signalisations de la route. On constate que le plus grand nombre d'accidents surviennent lorsqu'il n'y a aucune signalisation de la route.

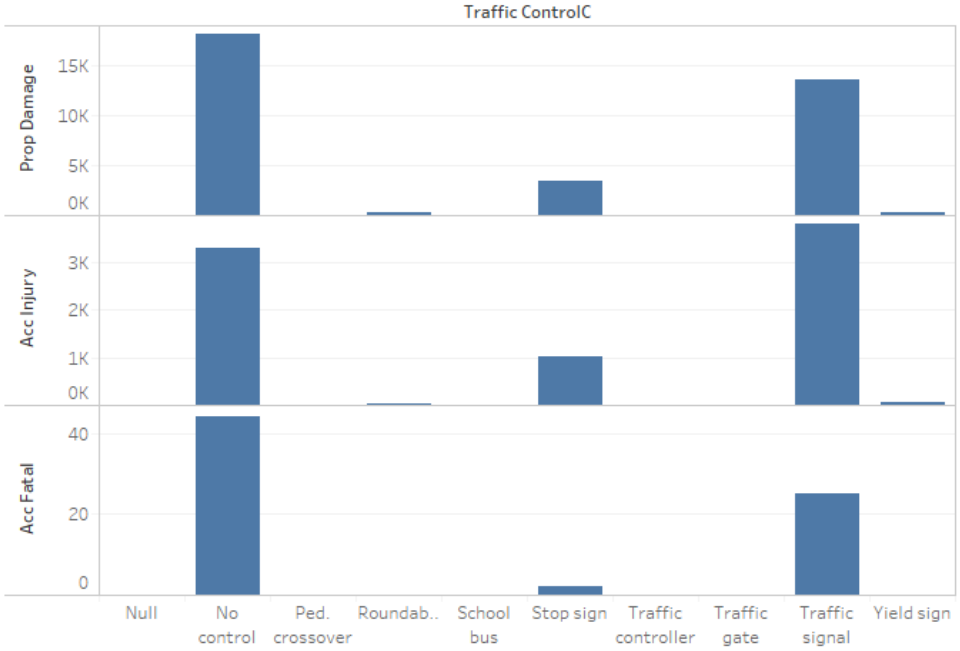


Fig.15 : Distribution des accidents selon les signalisations de la route

3.2.2. La préparation des données

Comme expliqué dans la section 3.2.1, les données utilisées dans ce travail de recherche sont des données recueillies de la base de données du Service de police d'Ottawa portant sur les accidents de la route dans la ville d'Ottawa en 2013 et des données recueillies par la ville d'Ottawa portant sur les accidents de la route dans la ville d'Ottawa de 2014 à 2016.

Dans la base de données de 2013, une vingtaine d'informations datant des années 2010, 2011, 2012 et 2014 se trouvaient. La première étape était donc de supprimer ces données. Ensuite on a constaté des cas dans lesquels la variable "accident_count", variable binaire définissant s'il y a accident (variable égale 1) ou pas accident (variable égale 0), prenait la valeur 0 dans des enregistrements annotés comme accident. Après un échange avec le Service de police, nous avons pu comprendre que cette anomalie était due à une panne technique. L'un des prétraitements les plus importants a donc été de créer un script pour corriger la variable "accident_count" dans ces cas contradictoires (annexe A).

Ensuite comme deuxième opération dans les deux bases de données, un script Excel a été créé pour ajouter des variables comme le jour de la semaine, le mois en fonction de la date à laquelle l'accident a eu lieu. Une description de ce script se retrouve dans l'annexe B.

Les données concernant la météo de l'année 2013 ont aussi été téléchargées sur le site du gouvernement [56] et ajoutées à la base de données de 2013 via un script Excel décrit à l'annexe C.

D'autres prétraitements ont été effectués aussi comme :

- La correction des heures dans un format heure ;
- La décomposition des heures en variables telles qu'heure, minute et seconde ;
- L'élimination de l'espace vide devant certaines valeurs de variables, notammentRoadway1
- La conversion de certaines variables numériques en variables catégoriques et vice versa. Ces variables sont: *atom* (par exemple110127 en numérique et A110127 en catégorique), zone (par exemple 101 en numérique et Z101 en catégorique), *district* (par exemple 11 en numérique et D11 en catégorique), jour (égale par exemple à 1 en numérique et lundi en catégorique), mois (par exemple 1 en numérique et janvier en catégorique), *accident_count* (par exemple1 en numérique et « oui » en catégorique), *rain* (par exemple1 en numérique et « oui » en catégorique), *snow* (par exemple1 en numérique et « oui » en catégorique), *environment* (prend par exemple les valeurs « *clear* » en catégorique et « 1 » en numérique), *road_surface* (prend par exemple les valeurs « Dry » en catégorique et « 1 » en numérique), *traffic_control* (prend par exemple les valeurs « *stop sign* » en catégorique et « 2 » en numérique) et *light* (prend par exemple les valeurs « *dark* » en catégorique et « 7 » en numérique). Cette conversion est nécessaire parce que certains algorithmes fonctionnent mieux avec le type catégorique que numérique et, dans certains cas, un algorithme peut fonctionner uniquement avec une catégorie.

Et enfin, à cause du fait que les deux bases de données sont déséquilibrées, c'est-à-dire qu'on dénombre par exemple 71cas d'accidents fatals contre 43 874 cas d'accidents non fatals dans la base de données 2014-2016, l'algorithme SMOTE a été utilisé pour prétraiter les données tel que décrit dans la section 3.4.

3.3. Modélisation

3.3.1. Les outils de modélisation R-Rattle et RapidMiner

L'outil utilisé pour la prédiction est l'outil R et plus précisément la librairie 'Rattle' [57]. La librairie Rattle offre une interface simple et intuitive qui permet à un utilisateur de charger rapidement des données à partir d'un fichier csv (ou via ODBC), de transformer et d'explorer les données, de construire et d'évaluer des modèles et d'exporter des modèles en PMML (*PredictiveModellingMarkupLanguage*) ou comme des scores. La figure 16 représente une image de l'espace de travail de l'outil R et de la librairie Rattle.

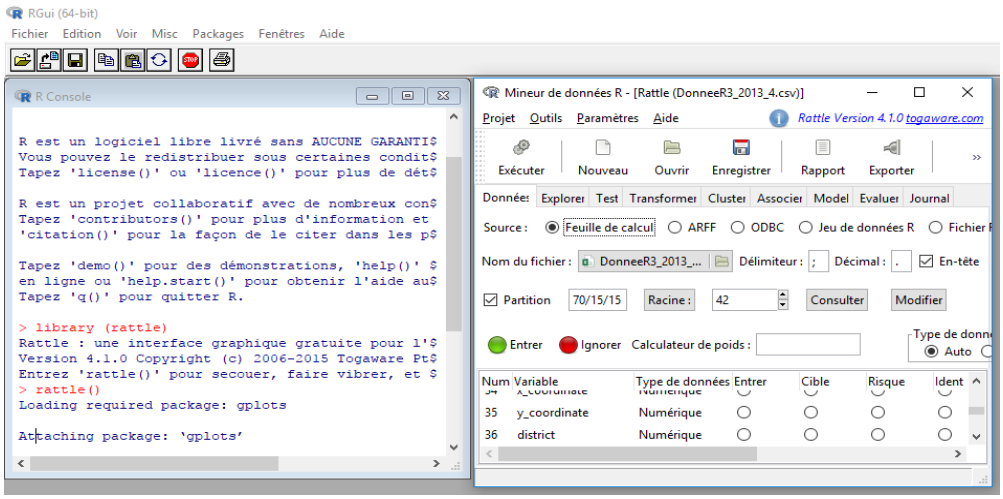


Fig.16 : Interface de l'outil R et la librairie 'Rattle'

Notons que dans ce travail de recherche nous avons à faire face à un problème de classification puisque le but est de prédire les données en classes « accident » ou « pas d'accident » ou selon les types d'accidents. Nous allons donc utiliser 7 algorithmes dans notre processus de modélisation/prédiction. Dans Rattle, nous utiliserons l'arbre de décision, le réseau de neurones, les machines de vecteurs de support (SVM) et le modèle AdaBoost, et dans Rapid Miner nous utiliseronsla méthode de k-voisins le plus proche (KNN), l'arbre de décision « *gradient boosted tree* » et l'algorithme naïf bayésien.

Ces méthodes vont donc servir de classificateurs et sont celles qui ont donné les résultats les plus probants. Un descriptif de chaque modèle se trouve dans les sections qui suivent.

Pour utiliser les méthodes choisies parmi les différentes méthodes dans Rattle, il faut cliquer sur l'onglet Model, sélectionner le modèle, ajuster les paramètres et cliquer sur le bouton 'Exécuter'. Un extrait de l'onglet se trouve à la figure 17.

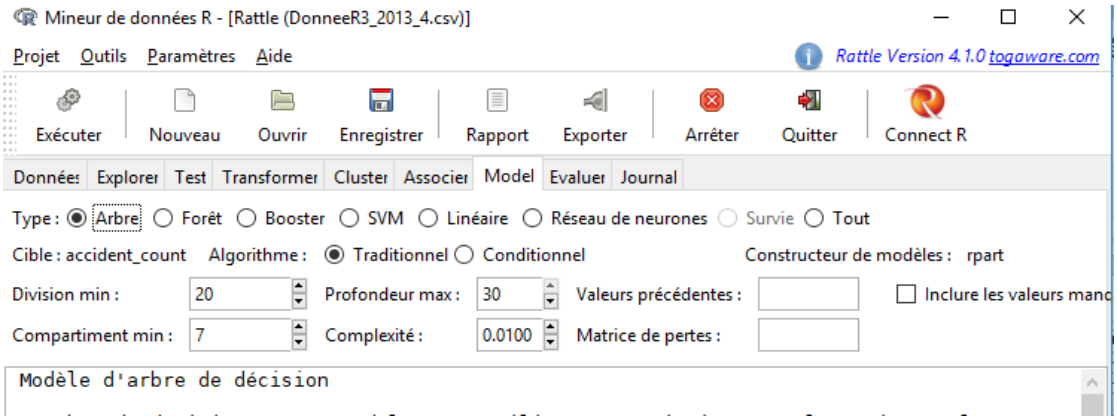


Fig.17 : Extrait de l'onglet "Model" de Rattle

Nous allons aussi utiliser l'outil Rapid Miner pour prédire et évaluer la performance des modèles. Rapid Miner est un logiciel de fouille de données servant à la préparation des données, à l'apprentissage automatique et au déploiement de modèles prédictifs[77]. L'interface du logiciel se divise en plusieurs blocs. La figure 18 illustre ces différents blocs. Comme nous pouvons le remarquer sur la figure 18, l'interface du logiciel Rapid Miner se divise en plusieurs blocs. Le bloc 1 appelé « Opérateur » est utilisé pour créer des processus RapidMiner. Le bloc 2, appelé « entrepôt », est le lieu de stockage des données et processus RapidMiner dans RapidMiner Studio. Le bloc 3, appelé « panneau de processus », est l'endroit qui sert à la construction des processus RapidMiner. Le bloc 4, appelé « vues », nous permet d'accéder aux différentes zones de travail. Il a deux modes : le mode « Design » qui nous permet de construire nos processus et le mode « Result » qui nous permet d'afficher le résultat du processus. Le bloc 5, appelé « port », est le mécanisme d'entrée et de sortie pour les opérateurs et les processus. Le bloc 6, appelé « paramètre », permet de modifier le comportement d'un opérateur en agissant sur ses paramètres. Finalement, le bloc 7, appelé « aide », nous permet de recevoir de l'aide sur l'opérateur sélectionné.

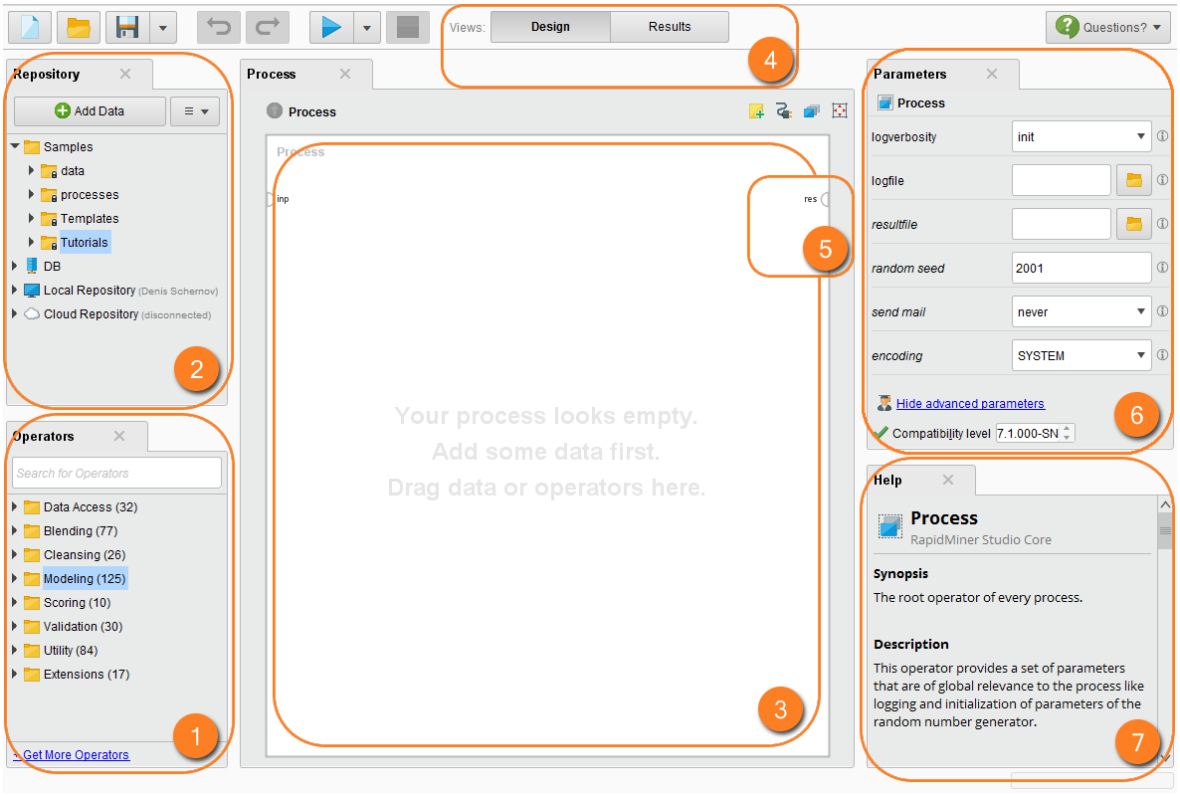


Fig.18 : Interface de Rapid Miner (extrait de [81])

La figure 19 illustre le processus que nous utiliserons dans RapidMiner pour effectuer de la prédiction et évaluer la performance des modèles.

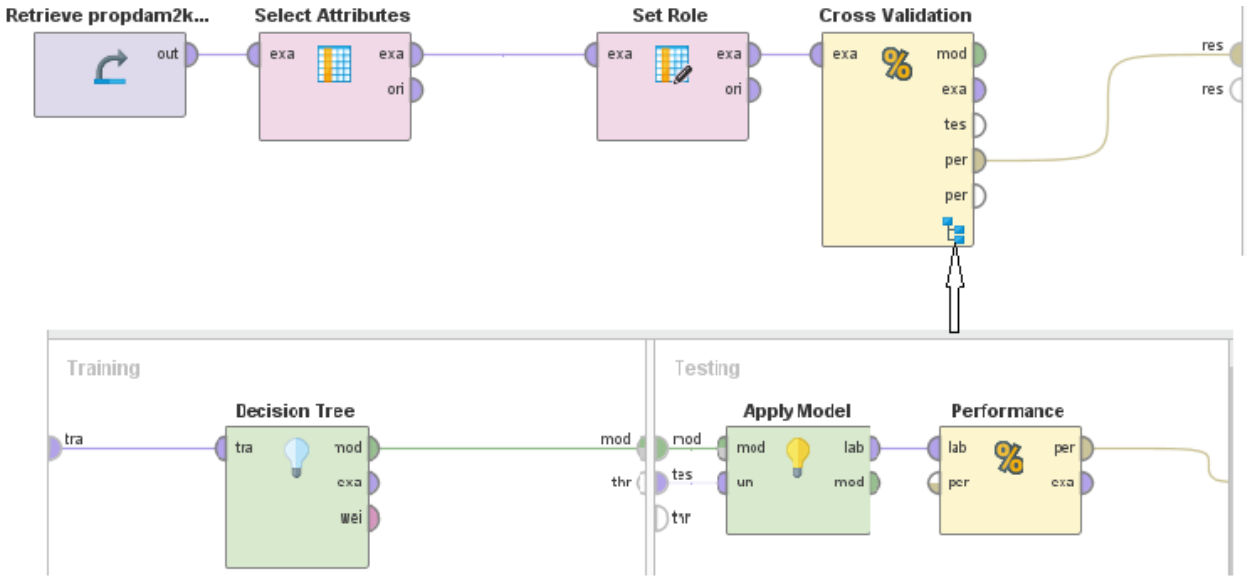


Fig.19 : Processus utilisé dans RapidMiner pour l'apprentissage et l'évaluation de performance

Le bloc "Retrieve" est le bloc d'entrée qui contient le fichier de la base de données portant sur les accidents dans un format *csv*. Les attributs à utiliser dans le jeu d'entraînement ont été sélectionnés en utilisant le bloc "Select Attributes". Dans le bloc "Set Role", on choisit la variable cible. Le bloc "Cross Validation" représente la validation croisée (k=10 plis) utilisée pour former et tester le modèle. Les blocs "Apply Model" et "Performance" ont été utilisés pour appliquer le modèle formé aux données de test et évaluer la performance en termes de précision et de rappel. Le même diagramme a été utilisé pour les autres modèles, en remplaçant le bloc "Decision Tree" par d'autres algorithmes d'apprentissage automatique.

Les prochaines sections sont dédiées à la description des divers algorithmes de classification employés dans ce mémoire.

3.3.2. Les arbres de décision

Comme son nom l'indique, un arbre de décision est une représentation schématique d'une décision et des différentes branches qui mènent à cette décision. C'est une méthode très utilisée dans l'apprentissage automatique et dans la fouille de données. Il décrit comment répartir un ensemble de choix en différents groupements homogènes selon des variables bien définies et en fonction d'un objectif fixé. Plus une variable est discriminante, plus elle est haute dans l'arbre de décision.

Lors de la construction de l'arbre deux métriques sont les plus utilisées pour évaluer la qualité d'une classe ou d'une sous classe [73].

- soit l'indice de Gini (G),

$$G = \sum_{k=1}^{n} P_{mk}\,(1 - P_{mk}) \tag{12}$$

- ou l'entropie croisée (E)

$$E = -\sum_{k=1}^{n} P_{mk} log P_{mk} \tag{13}$$

où k=1, 2, 3, …, *n* est l'ensemble des valeurs de la classe et P_{mk} représente l'ensemble des éléments *m* dans la classe *k*.

La figure 20 illustre un exemple simplifié d'un arbre de décision.

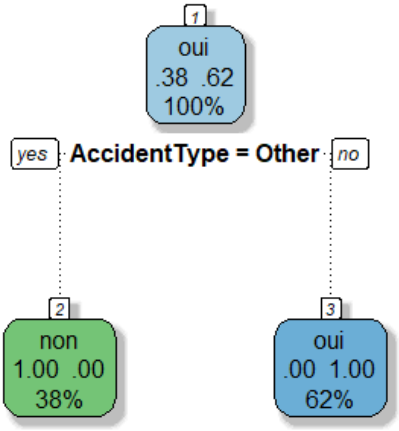


Fig.20 : Exemple d'arbre de décision

À travers l'outil R et à partir des données dont nous disposons, nous avons pu dessiner un exemple simplifié d'un arbre de décision, tel qu'illustré à la Figure 20.

L'objectif dans cet exemple est de déterminer s'il y'a un accident ou pas en fonction de la variable type d'accident (*AccidentType* sur la figure). La variable type d'accident a pour valeur les différents types d'accident (accident fatal, accident avec blessure, dommages matériels, etc.). La variable 'Accident Type' est donc soit égale à l'un des types d'accidents sinon elle est égale à autre (la valeur *other* sur la figure).

La condition ici est que si la variable ‘AccidentType’ est égale à ‘Other’ alors il n’y a pas d’accident, sinon il y a accident.

On peut noter sur la figure que dans 62% des cas il y a accident (62% de oui), et dans 38% il n’y a pas d’accident (38% de non). Dans Rattle, l’arbre de décision prend tout type de données, qu’il soit catégorique ou numérique.

3.3.3. Les réseaux de neurones

Le réseau de neurones est une méthodologie dont l’algorithme est inspiré du fonctionnement des neurones du cerveau. Pour comprendre cette méthodologie, une brève connaissance du fonctionnement du cerveau humain s’impose. Physiologiquement le cerveau est constitué de 10^{11} neurones interconnectés par 10^{15} connexions [58] à travers les axones. Les neurones reçoivent des signaux (sur la forme d’impulsions électriques) et envoient l’information par les axones.

Le cerveau humain est d’une extrême complexité ; il permet à l’être humain d’apprendre, de raisonner, de parler. L’ensemble de ces processus mentaux s’appelle la cognition. Le connexionnisme est le fait de vouloir rendre compte de la cognition humaine par les réseaux de neurones.

La notion de réseau de neurones formel fut évoquée pour la première fois par les neurologues Warren McCulloch et Walter Pitts [59]. Un réseau de neurones formel est constitué de plusieurs cellules interconnectées. Une cellule peut manipuler des valeurs binaires ou réelles. Plusieurs fonctions différentes peuvent être utilisées pour générer la sortie. La figure 21 est un exemple de schéma simplifié d’un réseau de neurones.

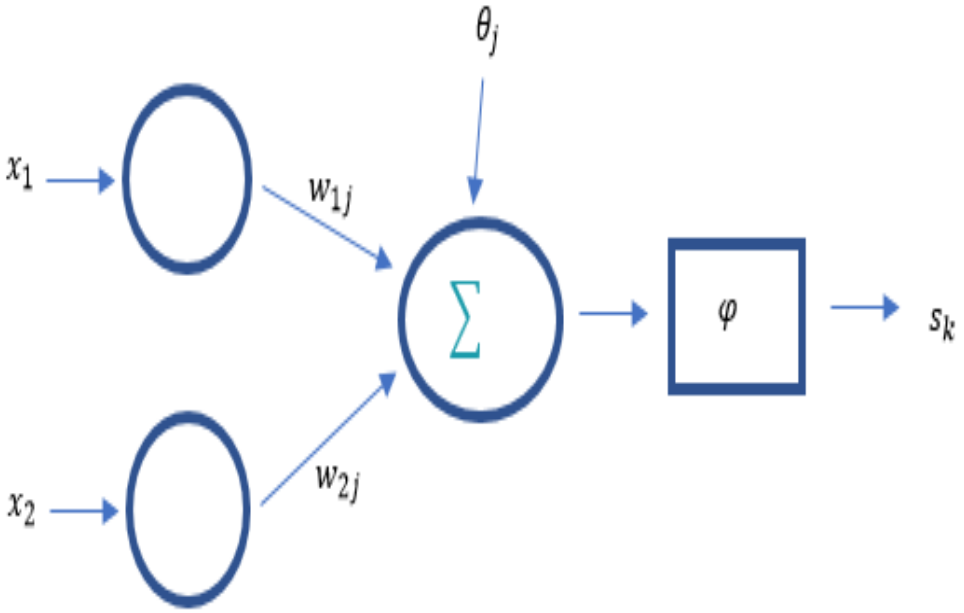


Fig.21 : Exemple d'un réseau de neurones.

Mathématiquement un réseau de neurones prend la forme suivante :

$$s_k = \varphi(\sum_{j=1}^{m} x_j w_{kj} + \theta_j) \qquad (14)$$

où x_j est le signal x associé à l'entrée j, w_{kj} est le poids synaptique associé à l'entrée j, θ_j est le seuil et φ est la fonction d'activation ou encore appelé fonction de sortie.

Il est à noter que l'algorithme du réseau de neurones prend en charge les données de type numériques et catégoriques. C'est grâce à la fonction *nnet* de R que Rattle arrive à faire de la prédiction avec le modèle du réseau de neurones.

3.3.4. Les SVM

Une machine à vecteur de support ou SVM (*support vector machines*) ou encore souvent appelé « séparateur à vastes marges » a pour objectif de représenter les données sous forme de points dans l'espace. Elle a été initialement définie pour la discrimination c'est-à-dire la prévision d'une variable qualitative binaire. Elle a ensuite été généralisée pour la prévision de variables qualitatives.

La machine à vecteur de support fut introduite durant les années 90 par les scientifiques Vladimir Vapnik (dans sa théorie appelé la théorie de Vapnik-Chervonenkis), Bernhard Boser et Isabelle Guyon [60]. Elle se base sur deux principes [61]:

- La définition de l'hyperplan comme solution d'un problème d'optimisation sous contraintes dont la fonction-objectif ne s'exprime qu'à l'aide de produits scalaires entre vecteurs et dans lequel le nombre de contraintes "actives" ou vecteurs supports contrôle la complexité du modèle ;
- La recherche de surfaces séparatrices non linéaires obtenues par l'introduction d'une fonction noyau (kernel) dans le produit scalaire induisant implicitement une transformation non linéaire des données vers un espace intermédiaire de plus grande dimension. De là provient l'appellation de machine à noyau ou « kernel machine ».

Le kernel que nous allons utiliser dans Rattle est le noyau Gaussien. Il s'écrit sous la forme suivante :

$$k(x_i, x_j) = e^{\frac{(\|x_i, x_j\|)^2}{2\gamma^2}} \tag{15}$$

où γ représente la largeur du kernel, $x_i et x_j$sont les entités.

La figure 22illustre deux exemples de problèmes de classification à deux classes ou deux groupes. Dans la figure 22(a) les données sont facilement séparées par une droite linéaire appelée« hyperplan ». Chaque groupe se trouve d'un côté de cette droite, on dit donc que le problème est linéairement séparable. La figure 22(b) montre l'avantage de l'utilisation des noyaux qui assurent la capacité de l'algorithme de résoudre des problèmes de classification non linéaires.

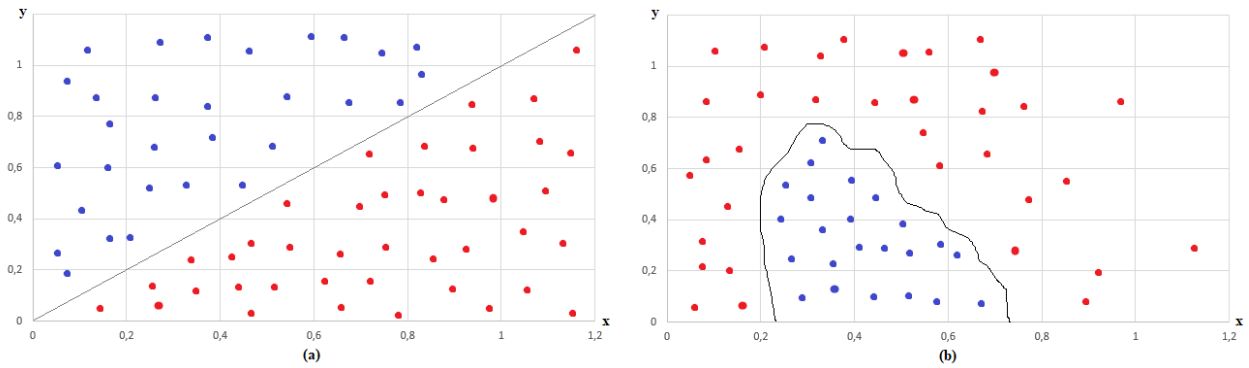


Fig.22 : (a) Exemple d'un problème à 2 classes avec un séparateur linéaire, (b) Exemple d'un problème à 2 classes avec séparateur non linéaire de type noyau(adapté de [71])

Dans l'implémentation Rattle, l'algorithme du SVM fonctionne avec des données numériques et avec des données catégoriques.

3.3.5. L'algorithme AdaBoost

La méthode AdaBoost (ou *Adaptative Boosting*) a été le premier algorithme d'amplification (*boosting*) développé pour la classification binaire. Elle fut introduite par Yoav Freund et Robert Schapire vers la fin des années 90[62]. Leurs motivations étaient de combiner le résultat de plusieurs classificateurs pour produire un ensemble plus puissant.

Adaboost peut donc être définie comme une méthode de sélection itérative de classificateurs faibles pour en créer des classificateurs forts. Cela se fait en construisant un modèle à partir des données d'apprentissage, puis en créant un deuxième modèle qui tente de corriger les erreurs du premier modèle. Les modèles sont ajoutés jusqu'à ce que l'ensemble des données d'apprentissage soit prédit parfaitement ou qu'un nombre maximum de modèles soit ajouté.

La fonction de base d'augmentation s'écrit sous la forme :

$$f(x) = \sum_{m=1}^{M} \beta_m b(x, \gamma_m) \qquad (16)$$

où les β_m sont des coefficients d'augmentation avec m = 1, 2,…, M et $b(x, \gamma_m) \in R$ sont de simples fonctions de l'argument multivarié x caractérisé par un ensemble de paramètresγ.

L'algorithme AdaBoost prend en entrée des données numériques ou catégoriques.

3.3.6. L'arbre de décision «Gradient boosted tree»

Tout comme AdaBoost, la méthode « *gradient boosted trees* » est un algorithme d'amplification (*boosting)*. Son but est de faire appel à plusieurs classificateurs « faibles » pour créer des classificateurs forts. L'algorithme construit une série de plusieurs petits arbres de décision. Chaque arbre tente de corriger les erreurs de l'étape précédente. La méthode se base sur trois éléments[78] :

- **La fonction de perte** : la fonction de perte utilisée dépend du type de problème à résoudre. Elle doit être différentiable, mais de nombreuses fonctions de perte standard sont prises en charge.
- **Le classificateur faible pour faire des prédictions** : les arbres de décision sont utilisés en tant que classificateur faible. Les arbres sont construits d'une manière optimale, en choisissant les meilleurs

points de partage basés sur des scores de pureté comme Gini ou pour minimiser la perte. Il est courant de contraindre les classificateurs faibles de manière spécifique, par exemple en utilisant un nombre maximum de couches, de nœuds, de divisions ou de nœuds feuilles.

- **Le modèle additif** : un modèle additif pour ajouter des classificateurs faibles afin de minimiser la fonction de perte. Les arbres sont ajoutés un à la fois et les arbres existant dans le modèle ne sont pas modifiés. Une procédure de descente de gradient (*gradient descent*) est utilisée pour minimiser la perte lors de l'ajout d'arbres.

3.3.7. L'algorithme naïf bayésien

L'algorithme naïf bayésien est un algorithme de classification simple et très puissant. En dépit de sa simplicité, le classificateur fait souvent étonnamment bien sa tâche et est largement utilisé parce qu'il surpasse souvent les méthodes de classification plus sophistiquées [73].

L'algorithme du classificateur est basé sur le théorème de Bayes avec les hypothèses d'indépendance entre les prédicateurs [79]. Le théorème de Bayes fonctionne sur la base de la probabilité conditionnelle.

La probabilité conditionnelle est la probabilité que quelque chose se produise, étant donné que quelque chose d'autre s'est déjà produit.

En utilisant la probabilité conditionnelle, nous pouvons calculer la probabilité d'un événement en utilisant ses connaissances antérieures.

La formule pour calculer la probabilité conditionnelle est la suivante :

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)} \qquad (17)$$

Où :

- *P(H)* est la probabilité que l'hypothèse *H* soit vraie
- *P(E)* est la probabilité d'évidence
- *P(E|H)* est la probabilité d'évidence en supposant que l'hypothèse est vraie
- *P(H|E)* est la probabilité de l'hypothèse en supposant qu'il y ait une évidence

3.3.8. L'algorithme des k-voisins les plus proches (KNN)

KNN est un algorithme d'apprentissage non paramétrique [80]. Cela signifie qu'il ne fait aucune hypothèse sur la distribution de données sous-jacentes. En d'autres termes, la structure du modèle est déterminée à partir des données.

L'algorithme KNN commence par un jeu de données d'apprentissage composé d'exemples classés en plusieurs catégories, étiquetés par une variable nominale. Supposons que nous ayons un jeu de données de test contenant des exemples sans étiquette qui ont les mêmes caractéristiques que les données d'apprentissage. Pour chaque enregistrement de l'ensemble de données de test, KNN identifie *k* enregistrements dans les données d'apprentissage qui sont les «plus proches» dans la similarité, où *k* est un nombre entier spécifié à l'avance. Une instance de test non marquée est assignée par l'algorithme à la classe de la majorité des *k* voisins les plus proches.

Pour illustrer le principe de fonctionnement de cette méthode, prenons un exemple très simple. Dans la figure 23(a) se trouvent deux classes distinctes « accident » et « pas accident ». Nous avons l'intention de prédire la classe de la donnée (*x)*. Dans ce cas, nous assumons que la valeur de *k* est fixée à 3. Dans la figure 23(b), le cercle en jaune couvre les 3 éléments les plus proches de *x*. Les 3 points les plus proches de *x* sont tous de la classe « accident ». Nous conclurons que *x* sera donc de la classe « accident ».

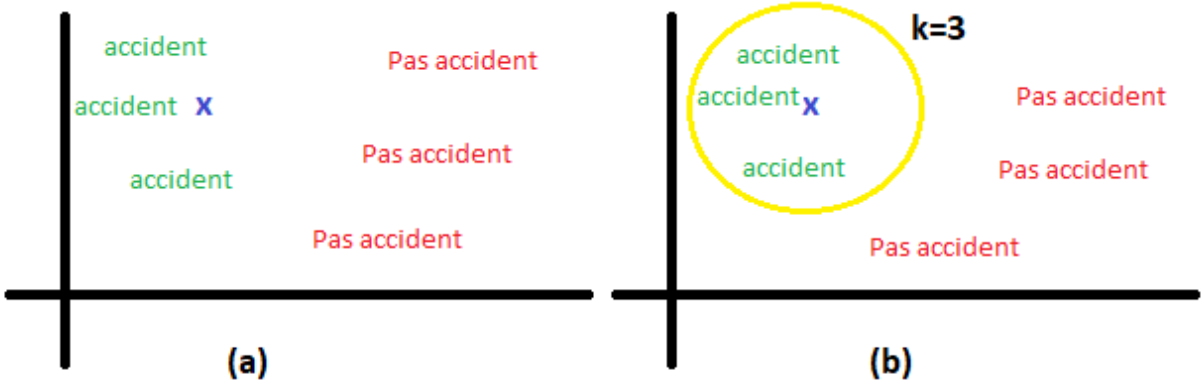


Fig.23 : Exemple d'un problème de classification KNN avec k=3 (adapté de [72])

3.3.9. L'évaluation de performance

L'évaluation de la performance d'un modèle est une étape très importante dans le processus de fouille de données. L'une des approches les plus populaires dans la mesure de la performance est le calcul du taux d'erreur en tant que nombre proportionnel de cas que le modèle classe incorrectement.

Dans ce travail de recherche, nous allons utiliser l'onglet « Évaluer»de Rattle pour évaluer la performance de nos modèles et les comparer. La figure 24 illustre l'interface de l'onglet « Évaluer »dans Rattle.

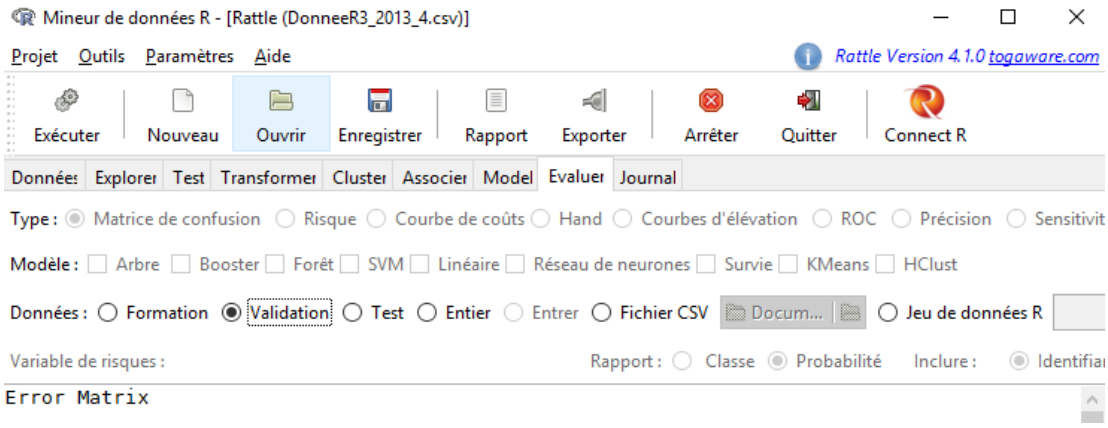


Fig.24 : Extrait de l'onglet « Évaluer » dans Rattle

Nous avons principalement trois lignes pour effectuer l'évaluation.

- La première ligne**:** porte sur le type d'évaluation que nous voulons effectuer, commençant de la génération d'une matrice de confusion à la représentation d'une courbe de sensitivité.
- La seconde ligne : permet de sélectionner le (s) modèle(s) voulu(s). Le modèle doit avoir été construit en premier lieu pour ensuite pouvoir en faire une évaluation.
- La troisième ligne : porte sur le jeu de données. La première option n'est pas très recommandée car elle a tendance à donner une estimation optimiste du jeu de données. La meilleure option est donc de choisir un jeu de données de test ou de validation ou même de sélectionner un fichier *csv* ou un fichier R existant pour tester sur des nouvelles données.

La courbe ROC est une des approches que nous allons utiliser pour évaluer la performance des modèles. Un exemple de courbe ROC se trouve à la figure 25. Notons tout d'abord sur la figure l'AUC (*Area Under the Curve*) ou la partie sous la courbe est une mesure de la précision du modèle. L'AUC prend des valeurs dans l'intervalle [0.5, 1]. Plus la valeur du AUC est proche de 1, plus le modèle est bon. Le taux de "vrai positifs" (*True Positive*) représente dans notre cas le taux d'accidents prédits en tant qu'accidents alors que ce sont de vrais accidents et le taux de "vrai négatifs" (*False Positive*) est le taux d'accidents prédits en tant qu'accidents alors que ce ne sont pas des accidents en réalité.

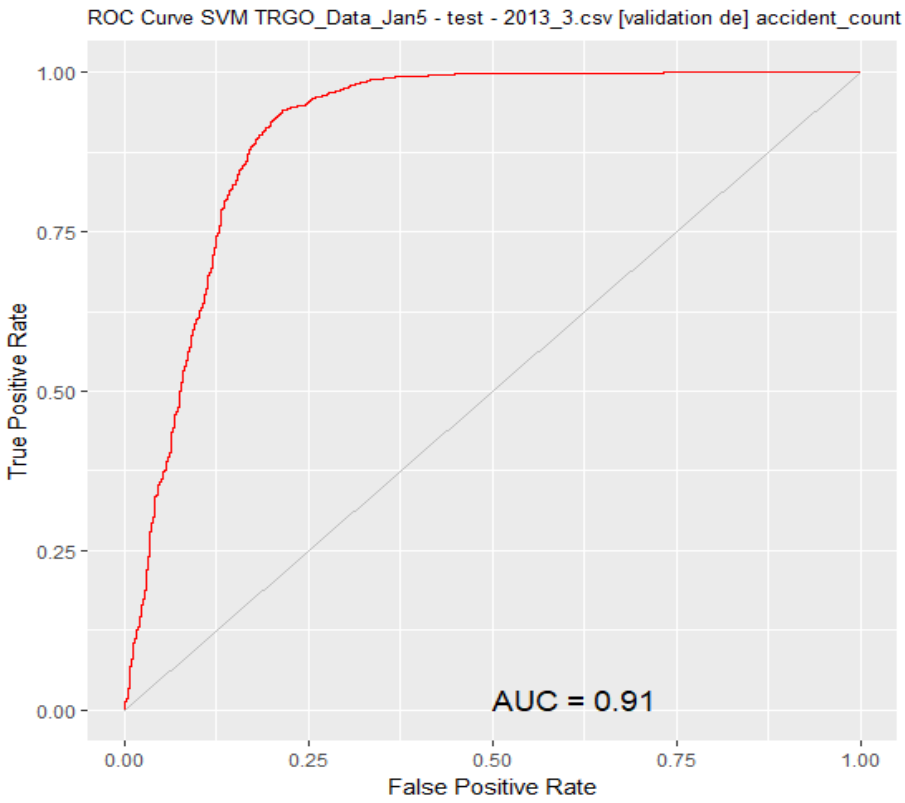


Fig.25 : Exemple de courbe ROC.

Pour entrainer et tester la précision des modèles nous allons utiliser deux méthodes connues pour les techniques d'apprentissage :

- Le test utilisant des données de validation (« Tests et validation »): Dans ce cas, on divise l'échantillon de taille *n* en deux sous-échantillons, le premier dit d'apprentissage (normalement supérieur à 60 % de la taille de l'échantillon) et le second dit de test. Le modèle est bâti sur l'échantillon d'apprentissage et validé sur l'échantillon de test. Nous utilisons cette technique dans la librairie Rattle. En effet dans Rattle nous utilisons 70% des données en données d'apprentissage, 15% en donnée de test et 15% en donnée de validation.
- La validation croisée à *k*-plis (*k-fold cross validation*) : Désigne le processus qui permet d'estimer la fiabilité d'un modèle fondé sur une technique d'échantillonnage [82]: on divise l'échantillon original en *k* échantillons, puis on sélectionne un des *k* échantillons comme ensemble de validation et les $k - 1$ autres échantillons constitueront l'ensemble d'apprentissage. On calcule le score de

performance, puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les $k - 1$ échantillons qui n'ont pas encore été utilisés pour la validation du modèle.

L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La performance est ensuite calculée comme la moyenne des scores de performance sur chaque échantillon. Nous utiliserons cette technique dans RapidMiner avec un nombre de plis k=10.

Pour évaluer la performance du modèle dans RapidMiner nous allons utiliser des matrices de confusion. La figure 26 illustre un exemple de matrice de confusion.

accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

	true non	true oui	class precision
pred. non	120688	0	100.00%
pred. oui	12	120771	99.99%
class recall	99.99%	100.00%	

Fig.26 : Exemple de matrice de confusion

Comme nous pouvons le remarquer sur la figure 26, dans ce cas, la précision du modèle (*accuracy*) est de 100%. Sur la totalité des cas de « pas accident », 120 688 événements ont été prédits comme pas accident alors que ce ne sont pas des accidents. Il n'y a pas d'événement prédit en tant que « accident » alors que ce sont des « pas accidents ».

Douze événements ont été prédits en tant que « pas accident » alors que ce sont des « accidents » et 120 771 événements ont été prédits en tant que « accident » alors que ce sont vraiment des « accidents ». Sur les 120 700 cas de « pas accident » prédits, 12 cas sont en fait des cas « accident ».

3.4. SMOTE (SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE)

Dans les deux bases de données utilisées dans ce mémoire, certaines variables cibles n'ont pas de classes égales. Par exemple, la variable « accident fatal », dans la base de données de 2014 à 2016, on dénombre 71 cas d'accident contre 43 874 cas de non accident. Ce déséquilibre de données conduit en général à une mauvaise performance pour la classification en raison du fait que le classificateur n'est pas suffisamment entrainé avec d'échantillons de la classe minoritaire pour lui permettre de faire de bonnes prédictions sur les nouvelles données. La classe minoritaire est la classe qui est sous-représentée (dans notre exemple, c'est la

classe contenant les 71 cas d'accidents), tandis que la classe majoritaire est la classe qui est surreprésentée (dans notre exemple c'est la classe contenant les 43 874 cas qui ne représentent pas des accidents).

SMOTE [74], de son acronyme *Synthetic Minority Oversampling Technique* (technique de sur-échantillonnage synthétique de la minorité), est une technique utilisée pour résoudre les problèmes de déséquilibre de données. On dit qu'un ensemble de données est déséquilibré si les classes de la variable cible ne sont pas approximativement égales. L'échantillonnage de données est l'une des solutions les plus largement utilisées pour traiter les déséquilibres de données. L'idée principale est de créer des instances équilibrées à travers les classes en ajoutant des données, en supprimant des données ou en ajoutant et en supprimant des données [74-75] de sorte que la précision de prédiction des classes minoritaires s'améliore. L'échantillonnage peut prendre deux formes principales, à savoir le sous-échantillonnage (suppression de données de la classe majoritaire) et le sur-échantillonnage (ajout de données de la classe minoritaire). Chacune de ces formes sera utilisée dans ce travail à travers la technique SMOTE.

Pour utiliser cette technique, dans notre travail de recherche nous utiliserons la librairie « DMwR » de R. Comme décrit dans [80], la librairie DMwR est constituée d'un ensemble de fonctions et d'algorithmes de fouille de données. L'une de ces fonctions est la fonction « SMOTE » que nous allons utiliser pour équilibrer nos deux bases de données. Le code R utilisé pour appliquer cette technique à notre base de données est décrit dans l'annexe D.

CHAPITRE 4 : RÉSULTATS

Pour la prédiction des accidents, nous avons procédé par plusieurs séries de tests. Dans chaque test, nous avons choisi diverses combinaisons de variables en entrée avec divers paramètres et différentes méthodologies, dont l'arbre de décision, le réseau de neurones, l'algorithme AdaBoost, SVM, l'algorithme naïf bayésien, les arbres de décision «*gradient boosted trees*» et l'algorithme des k-voisins les plus proches (KNN). Dans ce chapitre**,** nous allons exposer et discuter les résultats les plus probants. Pour rappel, dans nos tests avec Rattle, 70% des données sont utilisées comme données d'entrainement, 15% comme données de test et 15% comme données de validation. Pour l'apprentissage et l'évaluation avec Rapid Miner nous avons utilisé la validation croisée à 10 plis.

Dans le présent mémoire, les abréviations ADA, AD, RN, SVM, NB, GB et KNN sont utilisées respectivement pour l'algorithme AdaBoost, l'arbre de décision, le réseau de neurones, la machine à vecteurs de support, l'algorithme naïf bayésien, l'arbre de décision « *gradient boosted trees*» et l'algorithme dek-voisins le plus proches.

Pour les quatre algorithmes utilisés dans Rattle, c.à.d. l'arbre de décision, l'algorithme AdaBoost, le réseau de neurones et la machine à vecteurs de support, les paramètres utilisés sont les suivants :

- AD : Division min =20, profondeur max = 30, valeurs précédentes = 0, compartiment min=7, complexité=0.0100, matrice de perte=0
- ADA : nombre d'arbres=50, profondeur max=30, Division min=20, complexité=0.0100, valeur X=10
- RN : nombre de couches masquées=10
- SVM : kernel= base du radial

Pour les algorithmes de Rapid Miner nous avons utilisé les paramètres par défaut, comme il suit :

- KNN : valeur de *k*=1, type de mesure= « Mixedmeasure », mesure mixé=« MixedEuclideanDistance »
- *Gradient boosted tree* : nombre d'arbres=20, profondeur maximale=5, minimum de lignes=10, minimum de la division d'accroissement=0, nombre de boîtes=20, taux d'apprentissage=0.1, taux d'échantillonnage=1, distribution=auto.
- Algorithme naïf bayésien :*k*=1

Dans les sections suivantes, nous exposons un sommaire des meilleurs résultats obtenus. La matrice de confusion de la meilleure performance est illustrée dans chaque section. Dans chaque cas, les résultats complets se trouvent à l'annexe.

4.1. Classification « ACCIDENT » / « PAS D'ACCIDENT »

Une première catégorie de tests vise l'entrainement des classificateurs pour la prédiction du fait qu'un accident a lieu étant donné une combinaison des variables d'entrée. Le but est d'identifier les variables qui sont les plus sensées à identifier les conditions propices pour l'occurrence d'un accident. Dans ce but, nous avons effectué plusieurs tests (voir annexe E) à l'aide de Rattle et Rapid Miner. Ces tests sont effectués sur la base de données de 2013 en deux temps. Une première fois sur la base de données originale et une autre fois sur la base de données après l'application de l'algorithme SMOTE.

Le tableau 9 résume les résultats sur la base de données originale. Lorsque le champ « Paramètres » est vide dans le tableau, cela signifie qu'on utilise les paramètres par défaut, dans le cas contraire les paramètres sont présentés dans l'ordre dans lequel ils ont été décrits dans la section 4.

Tableau 9 : Prédiction accident/pas accident sur la base de données 2013 originale

N°	Entrée	Modèle	AUC	Paramètres
1	h, m, district, Location, jour_num, mois_num	ADA	63	
2	h, m, jour_char, mois_char, rainc2, snowc2	ADA	62	
3	h, m, jour_num, mois_num	ADA	61	
4	h, m, jour_num, mois_num, rainc, snowc	RN	61	5
5	h, m, district, Location, jour_num, mois_num, tempmin, tempmax, temp mean, rainc, snowc	SVM	61	
6	h, m, jour_num, mois_num	RN	60	
7	h, m, district, Location, jour_num, mois_num, tempmin, tempmax, temp mean, rainc, snowc	ADA	60	
8	occ_date, h, m	ADA	59	
9	Place_name	SVM	58	
10	Place_name	RN	58	3
11	h,m	ADA	58	
12	h,m	RN	58	n=5, 10, 25, 50, 75, 100
13	occ_date, h, m	SVM	58	
14	H	ADA	57	
15	Place_name	RN	57	5
16	district, zone, roadway1c,atom, Location	ADA	57	
17	Roadway1c	AD	56	
18	mois_num	RN	56	
19	Occ_time	ADA	55	
20	Occ_time	RN	55	

On peut constater que la performance est assez basse, il s'agit d'un AUC maximal de 63% avec l'algorithme Adaboost. Cette mauvaise performance est liée au déséquilibre dans la base de données.

Les tableaux10 et 11 résument les résultats obtenus de la base de données2013 après l'application de l'algorithme SMOTE pour l'équilibrer. Les résultats dans le tableau 10 représentent les méthodes testées en Rattle et le tableau 11 celles testées à l'aide de Rapid Miner. Dans ces tableaux, '*All*' dénote l'utilisation de toutes les variables d'entrée.

Tableau 10 : Résultats pour la prédiction accident / pas d'accidentsur la base de données traitée avec SMOTE(AUC)

N°	Entrée	Modèle	AUC
1	location, occDate	ADA	95
2	All	ADA	95
3	All	AD	93
4	location, occDate	AD	91
5	coordx, coordy	ADA	90
6	Location	ADA	89

N°	Entrée	Modèle	AUC
7	Location	AD	88
8	coordx, coordy	AD	88
9	roadway1c	ADA	78
10	H	ADA	77

Tableau 11 : Résultats pour la prédiction accident / pas d'accident sur la base de données traitées avec SMOTE (précision du modèle)

N°	Entrée	Modèle	Précision
1	All	KNN	92,40
2	All	GB	90,35
3	All	NB	88,69
4	coordx, coordy	GB	82,62
5	coordx, coordy	NB	82,42
6	Location	NB	79,99
7	Location	KNN	78,85
8	Location	GB	78,76
9	coordx, coordy	KNN	75,01
10	roadway1c	NB	71,24
11	H	NB	68,74
12	Atomc	NB	66,91
13	occDate	NB	66,2
14	Zonec	NB	64,95
15	Mois	NB	64,22
16	District	NB	63,95
17	roadway1c	GB	63,83
18	H	GB	63,32
19	occDate	GB	62,41
20	Atomc	GB	61,52

Comme nous pouvons le remarquer en comparant les résultats des 3 tableaux, les tests sur la base de données après l'application de l'algorithme SMOTE nous donnent de meilleurs résultats que sur la base de données originale. Dans les tableaux 10 et 11, les variables qui obtiennent les meilleures performances sont les variables reliées à la localisation telle que la variable '*Location*' avec une performance de 89% pour Adaboost (tableau 10) et les coordonnées *x* et *y* avec une performance de 82.62% en utilisant l'algorithme *gradient boosted tree*(tableau 11). Les variables reliées aux noms de la rue (*Roadway1c*) et l'heure nous donnent aussi de très bons résultats avec une performance respective de 78% et 77%. Les plus grandes

valeurs de l'AUC et de la précision sont obtenues en combinant les variables '*Location*' et'*occDate*' et une combinaison de toutes les variables (95% dans les deux cas avec Adaboost).

accuracy: 92.40% +/- 0.29% (mikro: 92.40%)

	true oui	true non	class precision
pred. oui	29312	4384	86.99%
pred. non	988	36016	97.33%
class recall	96.74%	89.15%	

Fig.27 : Matrice de confusion pour le meilleur résultat accident / pas accident

La figure 27 représente la meilleure performance de la précision obtenue en Rapid Miner qui est de 92.40% avec l'algorithme KNN et en utilisant toutes les variables. Nous pouvons aussi remarquer sur la figure 27 que sur 30 300 cas d'accidents, 988 ont été prédits comme « pas d'accident» et sur 40 400 cas de non accident 4 384 ont été prédits comme des accidents.

4.2. Classification selon les différents types d'accidents

Pour procéder à la classification selon les différents types d'accidents, nous avons utilisé deux techniques différentes. Dans un premier temps, nous avons procédé à une classification binaire, en considérant chaque type d'accidents, à savoir accident avec blessures, accident avec dommages matériels et accident fatal contre les deux autres types d'accidents.

Les résultats sont présentés à la section 4.2.1 à 4.2.3. Dans un deuxième temps, nous avons considéré ce problème de classification comme un problème de classification multi-classes. Les résultats sont présentés dans ce cas à la section 4.2.4.

Nous allons utiliser dans cette partie la base de données 2014 à 2016 après le prétraitement avec SMOTE afin de l'équilibrer, tel que décrit à la section 3.4. Nous résumons dans différents tableaux les valeurs de l'AUC à travers la courbe ROC ainsi que les valeurs obtenues pour la précision.

4.2.1. Le type d'accident avec blessures

Les tableaux 12 et 13résument les résultats des tests de classification des accidents en tant qu'accident avec blessures ou accident sans blessures. Comme dans la section précédente, le premier tableau montre la performance en termes de l'AUC (avec Rattle) et le deuxième la précision pour la validation croisée à 10 plis effectuée en Rapid Miner.

Tableau 12 : Prédiction d'accidents avec blessures (AUC)

N°	Entrée	Model	AUC
1	All	ADA	100
2	CoordX, CoordY	AD	99
3	All	AD	99
4	CoordX, CoordY	ADA	98
5	h, jour, mois	ADA	84
6	h, jour, mois	RN	83
7	jour, mois	ADA	78
8	jour, mois	RN	78
9	h, jour, mois	AD	78
10	h, m	ADA	76

Tableau 13 : Prédiction accident avec blessures (précision du modèle)

N°	Entrée	Modèle	Précision
1	date, jourChar, moisChar, h, coordx, coordy	NB	99,49
2	coordX, coordY	NB	98,87
3	date, jourChar, moisChar, h, location	KNN	98,82
4	date, jourChar, moisChar, h, coordx, coordy	GB	98,44
5	coordX, coordY	GB	98,43
6	coordX, coordY	KNN	96,23
7	date, light, location, environment, jourChar, moisChar, h	NB	93,61
8	date, jourChar, moisChar, h, location	NB	93,48
9	date, light, location, environment, jourChar, moisChar, h	GB	90,43
10	date, jourChar, moisChar, h, location	GB	89,81
11	date, jourChar, moisChar, h	NB	88,56
12	date, light, location	NB	87,57
13	date, light, location	GB	85,99
14	date, jourChar, moisChar, h	GB	85,46
15	date, jourChar, moisChar	GB	82,94
16	date, jourChar, moisChar	NB	81,65
17	Location	NB	79,35
18	H	NB	74

N°	Entrée	Modèle	Précision
19	Date	NB	73,24
20	Location	GB	71,07

On peut observer dans les tableaux 12 et 13, que les variables qui obtiennent les meilleures performances sont encore une fois les variables reliées à la localisation, telles que les coordonnées *x* et *y*, avec une performance de 99% pour l'arbre de décision (tableau 12) et 98,87% pour l'algorithme naïf bayésien (tableau 13). Les variables reliées à la date et l'heure nous donnent de très bons résultats aussi avec une performance respective de 73.24% et 74%. Les plus grandes valeurs de l'AUC et de la précision sont obtenues en utilisant une combinaison des variables : coordonnées *x* et *y*, *date*, *h*, *jour*, *mois* ou une combinaison de toutes les variables. La figure 28 représente la meilleure performance de la précision qui est de 99.49% en utilisant le jour, l'heure, les coordonnées *x* et *y*, et la date comme variables d'entrée et en utilisant l'algorithme naïf bayésien.

accuracy: 99.49% +/- 0.05% (mikro: 99.49%)

	true non	true oui	class precision
pred. non	164184	953	99.42%
pred. oui	776	172255	99.55%
class recall	99.53%	99.45%	

Fig.28 : Matrice de confusion de la meilleure performance pour les accidents avec blessures.

Nous pouvons aussi remarquer sur la figure 28 que sur 164 960 cas d'accidents avec blessures, 776 ont été prédits comme sans blessures et sur 173208 cas d'accidents sans blessures953 ont été prédits comme avec blessures.

4.2.2. Le type accident avec dommages matériels

Les tableaux 14 et 15 résument les résultats des tests de classification en accident avec dommages matériels et sans dommages matériels.

Tableau 14 : Prédiction d'accidents avec dommages matériels (AUC)

N°	Entrée	Modèle	AUC
1	CoordX, CoordY	AD	100
2	All	AD	100
3	All	ADA	100
4	h, jour, mois	ADA	73

N°	Entrée	Modèle	AUC
5	h, m	ADA	71
6	h, jour, mois	RN	71
7	h, m	RN	68
8	h, m	AD	65
9	jour, mois	RN	63
10	road_surface	ADA	62

Tableau 15 : Prédiction d'accidents avec dommages matériels(précision du modèle)

N°	Entrée	Modèle	Précision
1	date, jourChar, moisChar, h, coordx, coordy	KNN	99,76
2	date, jourChar, moisChar, h, coordx, coordy	NB	99.48
3	coordX, coordY	NB	98,98
4	date, jourChar, moisChar, h, coordx, coordy	GB	98,74
5	coordX, coordY	GB	98,73
6	date, light, location, environment, jourChar, moisChar, h	NB	91,59
7	date, jourChar, moisChar, h, location	NB	91,4
8	date, light, location	NB	88,65
9	date, jourChar, moisChar	GB	86,25
10	date, jourChar, moisChar, h	NB	86,01
11	date, jourChar, moisChar, h	GB	83,34
12	Location	NB	83
13	date, light, location, environment, jourChar, moisChar, h	GB	82,6
14	date, jourChar, moisChar, h, location	GB	82,34
15	date, light, location	GB	82,21
16	H	NB	79,88
17	Date	NB	79,54
18	date, jourChar, moisChar	NB	79,34
19	Location	GB	75,03
20	H	GB	74,68

accuracy: 99.76% +/- 0.03% (mikro: 99.76%)

	true oui	true non	class precision
pred. oui	332759	1242	99.63%
pred. non	1	173457	100.00%
class recall	100.00%	99.29%	

Fig.29 Matrice de confusion de la meilleure performance pour les accidents avec dommages matériels.

Les variables qui nous donnent les meilleures performances dans les tableaux 14 et 15 sont les variables reliées à la localisation telles que les coordonnées *x* et *y* avec une performance est de 100%. Les variables reliées à la date et l'heure nous donnent de très bons résultats aussi avec une performance respective de 79.54% et 79.88%. Les plus grandes valeurs de l'AUC et de la précision sont obtenues en utilisant une combinaison des variables coordonnées *x* et *y*, *date*, *h*, *jour*, *mois* et une combinaison de toutes les variables. La meilleure performance de la précision est de 99.76% avec la méthode KNN comme le démontre la figure 29. On peut constater dans cette figure que sur 332 760 cas d'accidents avec dommages matériels, un seul a été prédit comme sans dommages matériels et sur 174 699 cas de sans dommages matériels 1 242ont été prédits comme avec dommages matériels.

4.2.3. Le type accident fatal

Les résultats des tests de classification en accident fatal et accident non fatal sont présentés aux tableaux 16 et 17 respectivement.

Tableau 16: Prédiction d'accidents fatals (AUC)

N°	Entrée	Modèle	AUC
1	CoordX, CoordY	AD	100
2	CoordX, CoordY	ADA	100
3	All	AD	100
4	All	ADA	100
5	h, m, jour, mois	ADA	96
6	road_surface, traffic_control, environment, light, h, m	ADA	95
7	h,m	ADA	94
8	h, m, jour, mois	RN	88
9	road_surface, traffic_control, environment, light, h, m	RN	87
10	h,m	AD	86

Tableau 17 : Prédiction d'accidents fatals (précision du modèle)

N°	Entrée	Méthode	Précision
1	date, jourChar, moisChar, h, coordx, coordy	KNN	100
2	date, jourChar, moisChar, h, location	KNN	100
3	coordX, coordY	NB	100
4	date, jourChar, moisChar, h, coordx, coordy	NB	99,99
5	date, jourChar, moisChar, h, location	NB	99,99
6	date, light, location, environment, jourChar, moisChar, h	NB	99,99
7	date, light, location	NB	99,87

N°	Entrée	Méthode	Précision
8	date, light, location, environment, jourChar, moisChar, h	KNN	99,83
9	date, jourChar, moisChar, h	NB	99,81
10	date, jourChar, moisChar, h, coordx, coordy	GB	99,8
11	coordX, coordY	GB	99,78
12	date, jourChar, moisChar, h	KNN	99,73
13	date, light, location	KNN	99,72
14	coordX, coordY	KNN	99,48
15	Location	NB	98,09
16	H	NB	97,03
17	Date	NB	96,94
18	date, jourChar, moisChar	NB	96,77
19	Date	GB	95,88
20	date, jourChar, moisChar, h	GB	95,88

Pour la prédiction accident fatal versus accident non fatal, les variables portant sur les coordonnées *x* et *y* obtiennent une performance de 100%.

Elles sont suivies d'autres variables comme la date et l'heure avec une performance respective de 96.94% et 97.03%. Les meilleures performances sont obtenues en combinant des variables coordonnées *x* et *y*, *date*, *h*, *jour*, *mois* ou une combinaison de toutes les variables.

accuracy: 100.00% +/- 0.00% (mikro: 100.00%)

	true non	true oui	class precision
pred. non	120695	0	100.00%
pred. oui	5	120771	100.00%
class recall	100.00%	100.00%	

Fig.30 Matrice de confusion de la meilleure performance accident fatal.

Nous pouvons remarquer sur la figure 30qui montre la matrice de confusion basée sur l'algorithme KNN et les variables *date, jourChar, moisChar, h, coord x* et *coord y* que sur 120 700 cas d'accidents fatals, seulement 5 ont été prédits comme non fatals.

4.2.4. La classification multi-classes

Pour mieux comparer la performance, nous avons aussi procédé à des tests pour la classification multi-classes pour les trois types d'accidents. La variable utilisée comme variable cible est la variable « *collision classification* » dans la base de données 2014-2016. Elle prend comme valeur « 1 » pour accident fatal, « 2 »

pour accident avec blessure et « 3 » pour accident avec dommages matériels. Ces tests ont été effectués en Rapid Miner seulement et en utilisant les algorithmes *gradient boosted tree*, le k-voisin le plus prôche et l'algorithme naïf bayésien. Le tableau 17 résume les résultats obtenus dans ce cas.

Tableau 18 : Prédiction selon les 3 types ensemble (Précision du modèle)

Entrées	Précision (%)		
	GB	***KNN***	***NB***
light	48.40	33.02	48.40
Date	50.10	42.30	49.97
Road_surface	44.49	36.09	44.48
Traffic_control	45.76	30.56	45.76
Location	51.63	41.00	54.25
Coordx	97.11	37.86	38.50
Coordy	97.84	37.54	37.86
H	46.40	37.42	46.46
collision_location	42.71	33.01	42.71
Environment	39.95	34.56	39.95
coordx,coordy, location	98.81	45.20	55.60
location, light, date	65.01	49.91	63.96
coordx, coordy, location, light, date	98.68	52.68	65.28
location, light, date, h, traffic control	98.71	56.94	69.81
All	98.70	60.19	72.36

Comme nous pouvons le remarquer dans le tableau 18, les résultats sont moins bons dans ce cas que lorsque les différents types sont pris à part comme variables cibles. Nous avons tout de même de très bons résultats. Il est important de noter que cette solution est favorable à cause du fait qu'elle demande seulement l'apprentissage d'un seul algorithme au lieu des trois pour la classification séparée de chaque type d'accident. La meilleure précision obtenue est de 98.81% comme nous pouvons le remarquer à la figure 31. Si une erreur de 1.19% est tolérable, cette solution peut être considérée comme meilleure par rapport aux trois classificateurs dans les sections4.2.1 à 4.2.3.

accuracy: 98.81% +/- 0.04% (mikro: 98.81%)

	true 1	true 2	true 3	class precision
pred. 1	35398	5	22	99.92%
pred. 2	45	35247	627	98.13%
pred. 3	190	381	34984	98.39%
class recall	99.34%	98.92%	98.18%	

Fig.31 : Matrice de confusion de la meilleure performance selon la classification multi-classes de 3 types d'accidents

Ce résultat est obtenu avec une combinaison des variables coordonnées *x* et *y* et '*Location*' et en utilisant la méthode *gradient boosted tree*.

On peut constater aussi que, de nouveau, l'utilisation des variables reliées à la localisation nous donne les meilleurs résultats, tout comme dans le cas de l'utilisation des classificateurs binaires.

4.3. Importance des variables

Nous pouvons remarquer dans nos résultats que certaines variables sont plus pertinentes que d'autres pour la prédiction d'accidents. La figure 32 illustre le graphique d'importance des variables selon le pourcentage de la précision dans tous nos résultats. Les variables de la figure 32 sont celles qui contribuent le plus à l'obtention de meilleures performances quant à la prédiction des accidents de la route. Afin de produire ce graphique, nous avons calculé une moyenne des meilleures performances obtenues dans toutes les prédictions avec ces variables.

La variable portant sur les coordonnées *x* et *y* vient en tête avec une performance moyenne de 99.28%. Elle est suivie par les variables '*Location*' et heure ('h') avec des performances respectives de 86.81% et 83.63%. Ces résultats confirment le graphique d'importance obtenu par l'algorithme forêt aléatoire présenté à la section 3.2.1.2.

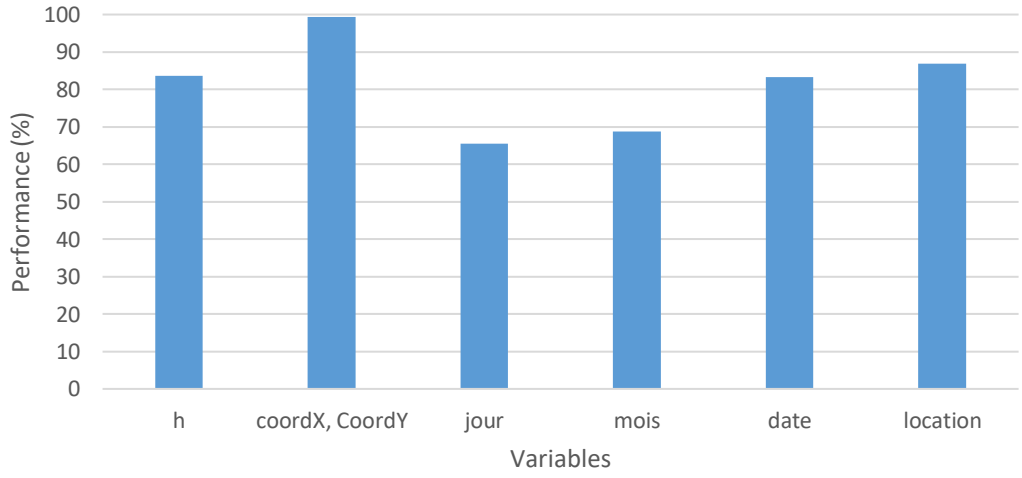


Fig.32 Graphe d'importance des variables selon les résultats obtenus

4.4. Performance des modèles selon l'ensemble des classifications

Nous avons aussi analysé la performance de l'ensemble des classificateurs. La figure 33montre la performance obtenue sur la base de l'AUC ainsi que sur la base de la précision pour les méthodes

d'apprentissage testées dans ce mémoire. Notons que ces résultats se basent pour ce graphique sur les tests effectués sur les deux bases de données après l'application de l'algorithme SMOTE.

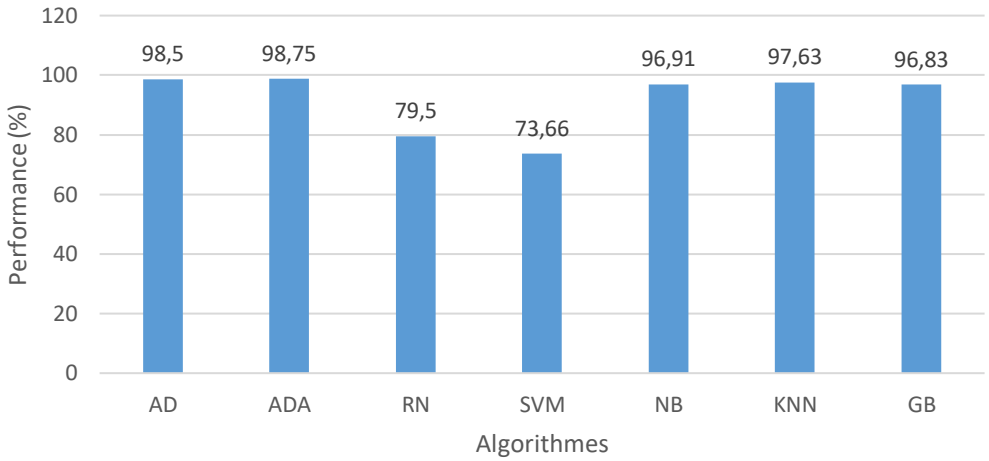


Fig.33 Comparaison de performance pour les algorithmes évalués

Si nous évaluons d'une manière générale la performance moyenne des différents algorithmes, selon la meilleure performance dans chaque type de prédiction, AdaBoost s'avère être l'algorithme qui offre la plus grande performance avec une performance moyenne de 98.75%. Il est suivi en ordre par l'arbre de décision avec 98.5%, et KNN avec 97.63%, L'algorithme naïf bayésien avec 96.91%,l'arbre de décision «*gradient boosted tree*» avec 96.83%,le réseau de neurones avec 79.50%, et la machine à vecteur de support avec 73.66%. Les valeurs de la figure 33 sont basées sur la moyenne de la performance des algorithmes selon les meilleures performances obtenues dans les tableaux10 à 17.

Chacun de ces algorithmes prend un temps d'exécution assez considérable. La figure 34 représente le temps d'exécution moyen de chaque algorithme. Le temps d'exécution est très dépendant de la taille de la base de données, de la quantité de variables utilisées en entrée et de la puissance de l'ordinateur.

Pour nos tests de prédictions nous avons utilisé des bases de données de taille variant entre 8 000 ko et 136 000 ko sur un système ayant 4 Go de RAM avec un processeur Intel core i3.Le nombre de lignes varie de 26 723 à 507 460. Nous pouvons constater que KNN prend le temps le plus long avec 25 minutes par test effectué. Il est suivi par SVM et AdaBoost.

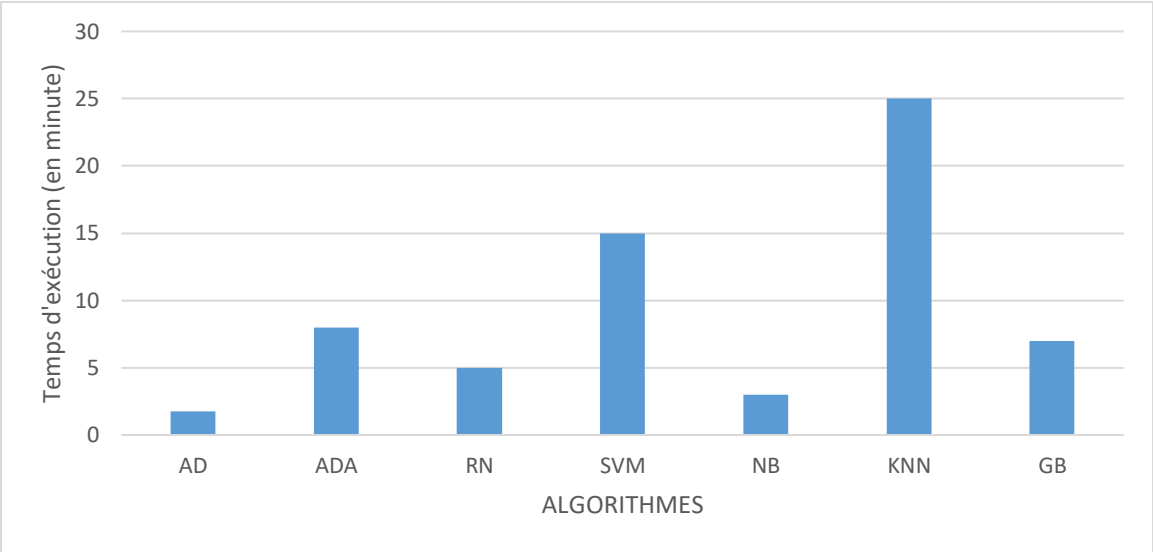


Fig.34 Temps d'exécution moyen des algorithmes

4.5. Comparaison avec la littérature

Finalement, nous avons comparé les meilleurs résultats obtenus dans ce mémoire avec ceux obtenus par des travaux semblables publiés dans la littérature. On vise spécifiquement les trois types d'accidents pour lesquels des résultats sont publiés dans [82, 83], puisque ces auteurs ont publié leurs résultats pour chacun des types étudiés dans ce travail.

Le tableau 19 montre les résultats de cette comparaison.

Pour rappel, pour présenter les résultats de la classification selon les types d'accidents au tableau 19 nous utilisons la base de données de 2014 à 2016 équilibrée à l'aide de SMOTE.

Tableau 19 : Comparaison des résultats avec la littérature

	Modèle(s)	Précision (%)			Nombre de variables
		Accident fatal	Accident avec blessures	Accident dommages matériels	
Littérature	Meilleur entre : CART decision trees, TreeNet, Random Forests [83]	77.4	77.9	100	32
	CART decision tree [84]	0	88.5	96.4	22
	Arbre de décision	100	99	100	19

Travail actuel	AdaBoost	100	100	100	19
	Réseau de neurones	88	83	71	19
	Algorithme naïf bayésien	100	99.49	99.48	19
	Gradient boosted tree	99.80	98.44	98.74	19
	KNN	100	98.82	99.66	19
	SVM	-	69	81	39

Dans le cas de la prédiction des accidents fatals et des accidents avec blessures, nous dépassons les résultats de la littérature qui sont respectivement de 77.4% et 77.9%. En prétraitant les données avec l'algorithme SMOTE et en utilisant une des méthodes : arbre de décision, Adaboost, l'algorithme naïf bayésien (KNN) nous obtenons une performance de 100%.Pour la prédiction d'accidents avec dommages matériels les résultats obtenus dans ce mémoire sont aussi bons que ceux publiés dans la littérature. À noter que pour les résultats présentés au tableau 19, la base de données 2014-2016 de 19 attributs a été utilisée pour tous les algorithmes à l'exception de SVM, pour lequel on a utilisé la base de données de 2013 avec 39 variables. À cause de la grandeur de la base de données 2014-2016, nous n'avons pas réussi à entrainer l'algorithme SVM sur l'ordinateur utilisé pour effectuer les tests.

4.6. Les intersections les plus dangereuses

Pour atteindre le dernier objectif de ce mémoire, nous avons dressé la liste des intersections les plus dangereuses dans la ville d'Ottawa. Nous avons vu dans l'état de l'art, dans la section 2.1, comment évaluer le risque de collision. Nous allons donc nous servir de ces techniques pour évaluer le risque d'accident aux différentes intersections. L'objectif est de pouvoir élaborer une liste des intersections les plus dangereuses dans la ville d'Ottawa et mettre en évidence le risque routier relié à ces intersections.

Le résultat se trouve dans le tableau 19. Les données portent sur les années 2014 à 2016. Le nombre de collisions sur les intersections est obtenu grâce à la création d'un tableau croisé dynamique dans Excel. Dans le logiciel Microsoft Excel on utilise donc l'option « tableau croisé dynamique » de l'onglet insertion pour obtenir ce résultat. Nous calculons ensuite le taux de collision selon son équation, tel que décrit dans la section 2.1.4. Ce taux est obtenu en fonction de la fréquence de collision selon l'équivalence de dommages matériels uniquement (*EPDO*) et la somme de la moyenne du trafic annuel journalier (*AADT*).

Tableau 20 : Top 10 des intersections les plus dangereuses

Intersections	**Collisions de 2014 à 2016**	**$\sum$*AADT* de 2015 et 2016**	***EPDO* de 2015 à 2016**	**Taux de collision *Rw***

ST. JOSEPH BLVD @ JEANNE D'ARC BLVD	173	39101	185	1295
HUNT CLUB RD @ RIVERSIDE DR	143	72358	319	1207
INNES RD @ TENTH LINE RD	110	28676	282	2692
PRINCE OF WALES DR @ WEST HUNT CLUB RD	109	66761	204	837
HIGHWAY 417 btwn HWY417 IC117 RAMP51 & HWY417 IC117 RAMP35	100	-	-	-
BASELINE RD @ WOODROFFE AVE	99	61375	203	906
WEST HUNT CLUB RD @ WOODROFFE AVE	99	57022	165	792
HIGHWAY 417 btwn HWY417 IC118 RAMP57 & HWY417 IC118 RAMP35	92	-	-	-
INDUSTRIAL AVE @ RIVERSIDE DR	92	70374	175	681
HIGHWAY 417 btwn HWY417 IC126 RAMP61 & HWY417 IC124 RAMP76	89	-	-	-

Les intersections sont classées dans ce tableau selon le nombre de collisions le plus élevé. Nous constatons que l'intersection « ST. JOSEPH BLVD @ JEANNE D'ARC BLVD » compte le nombre de collisions le plus élevé de 2014 à 2016 et l'intersection « HIGHWAY 417 btwn HWY417 IC126 RAMP61 & HWY417 IC124 RAMP76 » compte le nombre de collisions le moins élevé.

Nous calculons ensuite le taux de collision (*Rw*) en fonction de l'indice *EPDO,* comme décrit dans les équations (2) et (3) dans la section 2.1.

D'après ce calcul, l'intersection avec le plus élevé taux de collisions est « INNES RD @ TENTH LINE RD » avec 2692 collisions par cent millions de véhicules entrant dans cette intersection. Elle est suivie par l'intersection « ST. JOSEPH BLVD @ JEANNE D'ARC BLVD », et en troisième position vient l'intersection « HUNT CLUB RD @ RIVERSIDE DR ».

CHAPITRE 5 : CONCLUSION

Ce chapitre présente un sommaire de travaux effectués dans le cadre de ce mémoire, les conclusions tirées, les contributions ainsi que des pistes pour des travaux futurs dans le domaine.

5.1. Sommaire des résultats

Ce travail de recherche sur l'analyse et la prédiction des accidents de la route dans la ville d'Ottawa révèle que les sciences et technologies de l'information ont un rôle de premier plan à jouer dans l'amélioration de la sécurité routière. L'étude de l'état de l'art nous a démontré que ce thème a suscité l'intérêt de nombreux chercheurs qui ont développé des techniques et méthodologies intéressantes comme l'analyse service du risque routier, du risque personnel et la détermination du niveau de sécurité.

Afin de mener à bien ce travail de recherche nous avons utilisé deux bases de données, l'une provenant des services de police portant uniquement sur l'année 2013 et l'autre de la ville d'Ottawa en regroupant les données des années 2014, 2015 et 2016. Ces bases de données comportaient toutefois un problème d'équilibre des données : nous pouvions constater dans la base de données de la ville par exemple seulement 71 cas d'accidents fatals sur 44 015 collisions. Nous avions donc dû utiliser la technique de génération de données synthétiques (SMOTE) pour résoudre ce problème. D'après les résultats obtenus nous pouvons conclure que l'usage de SMOTE augmente la performance de classification.

Une analyse pertinente des données a été réalisée à l'aide de l'outil Tableau pour confirmer ou infirmer certaines tendances évoquées dans l'état de l'art.

La prédiction des accidents en accident / pas accident en utilisant les données de 2013 nous a démontré que les variables les plus importantes sont les coordonnées *x* et *y* avec une performance de 90%. Elles sont suivies par la variable '*Location'* avec une performance de 89%. D'autres variables telles que l'heure et le nom de la rue ou de l'intersection (*Roadway1c*) nous donnent une bonne performance avec des AUC respectifs de 77% et 78%.

Concernant la prédiction selon les types d'accidents, nous avons obtenu une meilleure performance en utilisant toutes les variables dont nous disposons.

Toutefois certaines variables se démarquent, telles que les coordonnées *x* et *y* qui offrent une performance au-delà de 90%, atteignant même 100% dans certains cas. Nous pouvons remarquer cela dans la classification selon les 3 différents types d'accident**s**. D'autres variables comme la variable '*Location'* et les variables concernant le jour et l'heure nous donnent de très bons résultats aussi, le meilleur étant de98.09%.

Les résultats obtenus dans ce travail de recherche viennent ajouter un pas aux études qui ont été réalisées sur le sujet. Nous avons remarqué que nos résultats sont plus performants que ceux disponibles dans la littérature actuelle.

AdaBoost est l'algorithme qui nous a globalement donné les meilleures performances, avec une performance moyenne de 98.75%. Il est suivi en ordre par l'arbre de décision avec 98.5%, la méthode de k-voisins le plus proche avec 97.63%, l'algorithme naïf bayésien avec 96.91%, l'arbre de décision «*gradient boosted tree*» avec 96.83%, le réseau de neurones avec 79.50%, et la machine à vecteur de support avec 73.66% respectivement.

Nous pouvons constater que l'analyse approfondie des accidents survenant sur le réseau routier de la ville d'Ottawa montre qu'un accident est la résultante d'une ou plusieurs défaillances dans un système complexe incluant les conducteurs, les véhicules, la route et le climat. Nous retenons toutefois que ces accidents de la route sont des événements prévisibles et nous osons espérer qu'à travers ce travail de recherche nous arriverons à une meilleure compréhension de ceux-ci et pour mieux contribuer à leur prévention.

Pour finir, ce fut une très grande joie d'effectuer ce travail de recherche. Les quelques mois passés à utiliser différents systèmes experts tels que Tableau, Rapid miner et R m'ont permis de comprendre et d'effectuer la fouille de données sur un ensemble de données relativement complexe.

5.2. Contributions

Ce mémoire apporte les contributions suivantes :

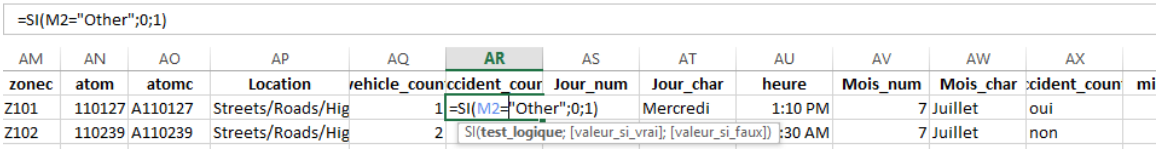
- L'étude des diverses variables qui caractérisent l'occurrence d'accidents rapportés dans les données 2013 à 2016 du Service de police de la ville d'Ottawa et de la ville d'Ottawa ;
- La prédiction des accidents dans des conditions données par des méthodes d'apprentissage pour l'année 2013 ;
- L'exploration et l'évaluation de divers algorithmes d'apprentissage pour la prédiction de types spécifiques d'accidents (à savoir : accident fatal, accident avec blessure grave et accident avec dommages matériels) pour les années 2013 à 2016;
- L'implémentation d'une solution pour résoudre le problème de déséquilibre des données ;
- L'évaluation comparative des résultats obtenus par les divers algorithmes et comparaison avec des solutions proposées dans la littérature ;
- La détermination d'une liste d'intersections dangereuses à Ottawa à partir des données sur le volume de trafic fournies par la ville d'Ottawa pour les années 2013, 2014, 2015 et 2016.

5.3. Travaux futurs

Certains aspects qui n'ont pas été considérés à ce stade, mais qui pourraient être exploités sont : la dynamique des accidents, l'analyse des scénarios de collisions [63], les trajectoires de mouvement du véhicule [64], la géométrie de la route, telles que : les courbes [65] ou les modèles de rue [38], la chronologie des accidents (découverte, réponse, suppression, récupération) [66], les modèles de comportement de conduite [67], la communication sur les réseaux de véhicules des données d'accident [68] et les données de trafic en temps réel [69] et flux vidéo [70]. D'autres algorithmes d'apprentissage et techniques d'équilibration des données peuvent être aussi utilisés tels que : l'approche GRSOM [85], la sélection pas à pas [86] et l'approche mixte [87].

Annexe A

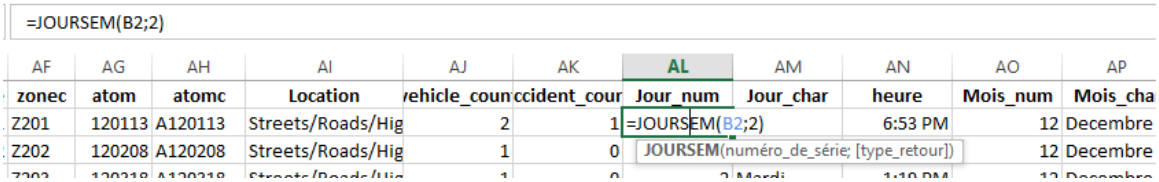
CORRECTION DE LA VARIABLE ACCIDENT_COUNT



Cette ligne stipule que si la valeur du champ “AccidentType” est égale à “Other”, alors le champ accident_count prend 0, ce qui veut dire qu’il n’y a pas d’accident. Par contre si le champ “ AccidentType” est égale à tout autre chose, alors accident_count prend la valeur 1, ce qui veut dire qu’un accident a lieu.

Annexe B

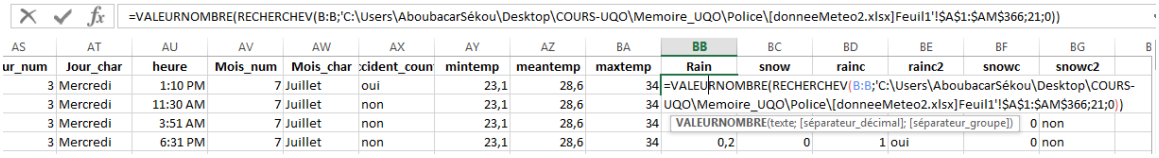
AJOUT DES VARIABLES JOUR ET MOIS



La fonction JOURSEM d'Excel nous permet de retrouver le jour de la semaine auquel correspond une date donnée. La colonne B est la colonne des dates et le paramètre 2 représente le type de retour, c'est à dire que 1 représente lundi et 7 dimanche.

Annexe C

AJOUT DES DONNÉES MÉTÉOROLOGIQUES



L'ajout des données de la météo correspondant à chaque jour qu'un accident a lieu est fait à travers la fonction RECHERCHEV. Cette fonction nous permet dechercher les données météorologiques de chaque jour dans lequel un accident a lieu dans un autre fichier Excel dans lequel les données météorologiques se trouvent.

Annexe D

APPLICATION DE L'ALGORITHME SMOTE SUR LA BASE DE DONNEES

```
##Importer la base de donnée
>data<-read.csv(file="path", sep="")

##appliquer l'algorithme Smote sur la base

>library(DMwR)

>newData <- SMOTE(Species ~ ., data, perc.over = 100,perc.under=100)
```

Annexe E

TABLEAUX COMPLETS DES RESULTATS DE PREDICTION

Prédiction d'accidents avec blessures (précision)

N° Teste	Entrée	Méthode	Précision
1	date, jourChar, moisChar, h, coordx, coordy	NB	99,49
2	coordX, coordY	NB	98,87
3	date, jourChar, moisChar, h, location	KNN	98,82
4	date, jourChar, moisChar, h, coordx, coordy	GB	98,44
5	coordX, coordY	GB	98,43
6	coordX, coordY	KNN	96,23
7	date, light, location, environment, jourChar, moisChar, h	NB	93,61
8	date, jourChar, moisChar, h, location	NB	93,48
9	date, light, location, environment, jourChar, moisChar, h	GB	90,43
10	date, jourChar, moisChar, h, location	GB	89,81
11	date, jourChar, moisChar, h	NB	88,56
12	date, light, location	NB	87,57
13	date, light, location	GB	85,99
14	date, jourChar, moisChar, h	GB	85,46
15	date, jourChar, moisChar	GB	82,94
16	date, jourChar, moisChar	NB	81,65
17	Location	NB	79,35
18	H	NB	74
19	Date	NB	73,24
20	Location	GB	71,07
21	moisChar	NB	69,48
22	H	GB	66,8
23	environment, light, roadSuface, trafficControl	NB	66,77
24	jourChar	NB	65,67
25	roadSurface	NB	64,71
26	environment, light, roadSuface, trafficControl	GB	64,46
27	moisChar	GB	61,82
28	Date	GB	60,26
29	Light	NB	58,92
30	roadSurface	GB	54,77
31	colliisonLocation	NB	53,86
32	colliisonLocation	GB	53,78

N° Teste	Entrée	Méthode	Précision
33	Light	GB	53,66
34	Environment	NB	53,37
35	trafficControl	NB	53,2
36	jourChar	GB	52,93
37	Environment	GB	51,69
38	trafficControl	GB	51,64
39	colliisonLocation	AD	51,22
40	coordX, coordY	AD	51,22
41	Date	AD	51,22
42	Environment	AD	51,22
43	environment, light, roadSuface, trafficControl	AD	51,22
44	jourChar	AD	51,22
45	date, jourChar, moisChar	AD	51,22
46	date, light, location	AD	51,22
47	date, jourChar, moisChar, h	AD	51,22
48	date, light, location, environment, jourChar, moisChar, h	AD	51,22
49	date, jourChar, moisChar, h, location	AD	51,22
50	date, jourChar, moisChar, h, coordx, coordy	AD	51,22
51	Light	AD	51,22
52	Location	AD	51,22
53	moisChar	AD	51,22
54	roadSurface	AD	51,22
56	H	AD	51,22
57	trafficControl	AD	51,22
58	colliisonLocation	KNN	48,82
59	H	KNN	48,82

Prédiction d'accidents avec blessures (AUC)

N°	Entrée	Modèle	ROC
1	road_surface	AD	67
2	road_surface	ADA	68
3	road_surface	RN	68
4	Environment	AD	52
5	Environment	ADA	53
6	Environment	RN	53
7	traffic_control	AD	53
8	traffic_control	ADA	55
9	traffic_control	RN	55
10	Light	AD	58
11	Light	ADA	58
12	Light	RN	59

N°	Entrée	Modèle	ROC
13	h, m	AD	74
14	h, m	ADA	76
15	h, m	RN	74
16	CoordX, CoordY	AD	99
17	CoordX, CoordY	ADA	98
18	jour, mois	AD	76
19	jour, mois	ADA	78
20	jour, mois	RN	78
21	h, jour, mois	AD	78
22	h, jour, mois	ADA	84
23	h, jour, mois	RN	83
24	All	AD	99
25	All	ADA	100

Prédiction d'accidents avec dommages matériels (précision)

N°	Entrée	Méthode	Précision
1	colliisonLocation	GB	60
2	Date	GB	73,11
3	Environment	GB	40,03
4	jourChar	GB	36,21
5	Light	GB	49,2
6	Location	GB	75,03
7	moisChar	GB	49,84
8	roadSurface	GB	47,84
9	H	GB	74,68
10	trafficControl	GB	44,69
11	coordX, coordY	GB	98,73
12	date, jourChar, moisChar	GB	86,25
13	date, jourChar, moisChar, h	GB	83,34
14	date, jourChar, moisChar, h, coordx, coordy	GB	98,74
15	environment, light, roadSuface, trafficControl	GB	67,81
16	date, jourChar, moisChar, h, location	GB	82,34
17	date, light, location	GB	82,21
18	date, light, location, environment, jourChar, moisChar, h	GB	82,6
19	colliisonLocation	NB	67,01
20	Date	NB	79,54
21	Environment	NB	68,59
22	jourChar	NB	65,57
23	Light	NB	65,57
24	Location	NB	83
25	moisChar	NB	65,57
26	roadSurface	NB	66,92
27	H	NB	79,88
28	trafficControl	NB	65,57
29	coordX, coordY	NB	98,98
30	date, jourChar, moisChar	NB	79,34
31	date, jourChar, moisChar, h	NB	86,01
32	date, jourChar, moisChar, h, coordx, coordy	NB	99,48
33	environment, light, roadSuface, trafficControl	NB	70,24
34	date, jourChar, moisChar, h, location	NB	91,4
35	date, light, location	NB	88,65
36	date, light, location, environment, jourChar, moisChar, h	NB	91,59
37	colliisonLocation	AD	65,57
38	Date	AD	65,57
39	Environment	AD	65,57

N°	Entrée	Méthode	Précision
40	jourChar	AD	65,57
41	Light	AD	65,57
42	Location	AD	65,57
43	moisChar	AD	65,57
44	roadSurface	AD	65,57
45	H	AD	65,57
46	trafficControl	AD	65,57
47	coordX, coordY	AD	65,57
48	date, jourChar, moisChar	AD	65,57
49	date, jourChar, moisChar, h	AD	65,57
50	date, jourChar, moisChar, h, coordx, coordy	AD	65,57
51	environment, light, roadSuface, trafficControl	AD	65,57
52	date, jourChar, moisChar, h, location	AD	65,57
53	date, light, location	AD	65,57
54	date, light, location, environment, jourChar, moisChar, h	AD	65,57

Prédiction d'accidents avec dommages matériels (AUC)

N°	Entrée	Modèle	AUC
1	road_surface	AD	51
2	road_surface	ADA	62
3	road_surface	RN	62
4	Environment	AD	58
5	Environment	ADA	60
6	Environment	RN	60
7	traffic_control	AD	50
8	traffic_control	ADA	55
9	traffic_control	RN	55
10	Light	AD	50
11	Light	ADA	60
12	Light	RN	60
13	h, m	AD	65
14	h, m	ADA	71
15	h, m	RN	68
16	coordX, coordY	AD	100
17	jour, mois	AD	50
18	jour, mois	ADA	62
19	jour, mois	RN	63
20	h, jour, mois	AD	60
21	h, jour, mois	ADA	73
22	h, jour, mois	RN	71
23	All	AD	100
24	All	ADA	100

Prédiction d'accidents fatals (précision)

N°	Entrée	Méthode	Précision
1	collisionLocation	GB	62,37
2	Date	GB	95,88
3	Environment	GB	53,98
4	jourChar	GB	64,84
5	Light	GB	53,54
6	Location	GB	90,32
7	moisChar	GB	67,84
8	roadSurface	GB	55,35
9	H	GB	94,44
10	trafficControl	GB	57,46
11	coordX, coordY	GB	99,78
12	date, jourChar, moisChar	GB	91,39
13	date, jourChar, moisChar, h	GB	95,88
14	date, jourChar, moisChar, h, coordx, coordy	GB	99,8
15	environment, light, roadSuface, trafficControl	GB	67,81
16	date, jourChar, moisChar, h, location	GB	94,3
17	date, light, location	GB	95,5
18	date, light, location, environment, jourChar, moisChar, h	GB	93,27
19	colliisonLocation	KNN	49,99
20	Date	KNN	49,99
21	environment	KNN	49,99
22	jourChar	KNN	49,99
23	Light	KNN	49,99
24	location	KNN	52,39
25	moisChar	KNN	49,99
26	coordX, coordY	KNN	99,48
27	date, jourChar, moisChar	KNN	84,05
28	date, jourChar, moisChar, h	KNN	99,73
29	date, jourChar, moisChar, h, coordx, coordy	KNN	100
30	clenv,environment, light, roadSuface, trafficControl	KNN	51,16
31	date, jourChar, moisChar, h, location	KNN	100
32	date, light, location	KNN	99,72
33	date, light, location, environment, jourChar, moisChar, h	KNN	99,83
34	collisionLocation	NB	64,23
35	Date	NB	96,94
36	environment	NB	57,72
37	jourChar	NB	65,2
38	Light	NB	56,38
39	location	NB	98,09

N°	Entrée	Méthode	Précision
40	moisChar	NB	71,14
41	roadSurface	NB	62,95
42	H	NB	97,03
43	trafficControl	NB	61,11
44	coordX, coordY	NB	100
45	date, jourChar, moisChar	NB	96,77
46	date, jourChar, moisChar, h	NB	99,81
47	date, jourChar, moisChar, h, coordx, coordy	NB	99,99
48	environment, light, roadSuface, trafficControl	NB	68,59
49	date, jourChar, moisChar, h, location	NB	99,99
50	date, light, location	NB	99,87
51	date, light, location, environment, jourChar, moisChar, h	NB	99,99
52	colliisonLocation	AD	50,01
53	Date	AD	50,01
54	environment	AD	50,01
55	jourChar	AD	50,01
56	Light	AD	50,01
57	location	AD	50,01
58	moisChar	AD	50,01
59	roadSurface	AD	50,01
60	H	AD	50,01
61	trafficControl	AD	50,01
62	coordX, coordY	AD	50,01
63	date, jourChar, moisChar	AD	50,01
64	date, jourChar, moisChar, h	AD	50,01
65	date, jourChar, moisChar, h, coordx, coordy	AD	50,01
66	environment, light, roadSuface, trafficControl	AD	50,01
67	date, jourChar, moisChar, h, location	AD	50,01
68	date, light, location	AD	50,01
69	date, light, location, environment, jourChar, moisChar, h	AD	50,01

Prédiction d'accidents fatals (AUC)

N°	Entrée	Model	ROC
1	road_surface	AD	63
2	road_surface	ADA	63
3	road_surface	RN	63
4	Environment	AD	58
5	Environment	ADA	58
6	Environment	RN	58
7	traffic_control	AD	61
8	traffic_control	ADA	63
9	traffic_control	RN	63
10	Light	AD	58
11	Light	ADA	58
12	Light	RN	58
13	h,m	AD	86
14	h,m	ADA	94
15	h,m	RN	68
16	jour, mois	AD	75
17	jour, mois	ADA	82
18	jour, mois	RN	83
19	h, m, jour, mois	AD	76
20	h, m, jour, mois	ADA	96
21	h, m, jour, mois	RN	88
22	coordX, coordY	AD	100
23	coordX, coordY	ADA	100
24	road_surface, traffic_control, environment, light, h, m	AD	79
25	road_surface, traffic_control, environment, light, h, m	ADA	95
26	road_surface, traffic_control, environment, light, h, m	RN	87
27	All	AD	100
28	All	ADA	100

Prédiction accident / pas accident (précision)

N° Teste	Entrée	Modèle	Précision
1	All	KNN	92,06
2	All	GB	90,35
3	All	NB	88,69
4	coordx, coordy	GB	82,62
5	coordx, coordy	NB	82,42
6	Location	NB	79,99
7	Location	KNN	78,85
8	Location	GB	78,76
9	coordx, coordy	KNN	75,01
10	roadway1c	NB	71,24
11	Heure	NB	68,74
12	Atomc	NB	66,91
13	occDate	NB	66,2
14	Zonec	NB	64,95
15	Mois	NB	64,22
16	District	NB	63,95
17	roadway1c	GB	63,83
18	Heure	GB	63,32
19	occDate	GB	62,41
20	Atomc	GB	61,52
21	place_name	NB	61,41
22	place_name	GB	61,07
23	Mois	GB	60,27
24	Jour	NB	58,48
25	District	GB	58,31
26	Zonec	GB	57,73
27	Jour	GB	57
28	Atomc	KNN	51,76
29	roadway1c	KNN	51,34
30	place_name	KNN	44,46
31	Heure	KNN	44,04
32	Zonec	KNN	42,87
33	Jour	KNN	42,86
34	Mois	KNN	42,86
35	District	KNN	42,86
36	occDate	KNN	42,86

Prédiction accident / pas accident (AUC)

N° Teste	Entrée	Modèle	AUC
1	occDate	AD	74
2	occDate	ADA	75
3	occDate	RN	75
4	Location	AD	88
5	location	ADA	89
6	roadway1c	AD	75
7	roadway1c	ADA	78
8	place_name	AD	56
9	place_name	ADA	55
10	place_name	RN	76
11	district	AD	68
12	district	ADA	70
13	district	RN	70
14	Zone	AD	69
15	Zone	ADA	72
16	Zone	RN	72
17	Atom	AD	73
18	Atom	ADA	75
19	Atom	RN	75
20	Jour	AD	54
21	Jour	ADA	55
22	Jour	RN	55
23	heure	AD	76
24	heure	ADA	77
25	Mois	AD	67
26	Mois	ADA	69
27	Mois	RN	69
28	Rain	AD	50
29	Rain	ADA	56
30	Rain	RN	56
31	Snow	AD	55
32	Snow	ADA	55
33	Snow	RN	55
34	coordx, coordy	AD	88
35	coordx, coordy	ADA	90
36	location, occDate	AD	91
37	location, occDate	ADA	95
38	All	AD	93
39	All	ADA	95

Prédiction accident / pas accident avec la base de données originale

NumTest	Entrée (s)	Modèle	AUC (en %)	Paramètres
1	Occ_time	ADA	55	
2	Occ_time	RN	55	
3	H	AD	53	
4	H	AD	54	12, 20, 4, 0.0100
5	H	ADA	57	
6	H	RN	56	
7	H	SVM	55	
8	Roadway1c	AD	56	
9	Roadway1c	ADA	55	
10	Place_name	SVM	58	
11	Place_name	RN	57	5
12	Place_name	RN	58	3
13	Vehicle.towed	AD	58	
14	Vehicle.towed	ADA	58	
15	Vehicle.towed	SVM	58	
16	Vehicle.towed	RN	58	
17	traffic.Compl	AD	71	
18	traffic.Compl	ADA	71	
19	traffic.Compl	RN	71	
20	traffic.Compl	SVM	71	
21	ProvOffense	ADA	59	
22	ProvOffense	SVM	59	Polynomial
23	ProvOffense	SVM	59	Lineaire
24	ProvOffense	SVM	59	Tangentehyperbolique
25	ProvOffense	RN	59	
26	mois_num	ADA	55	
27	mois_num	RN	56	
28	minTemp	ADA	55	
29	minTemp	RN	55	
30	h,m	ADA	58	
31	h,m	SVM	56	
32	h,m	RN	58	5, 10, 25, 50, 75, 100
33	occ_date, h, m	AD	55	
34	occ_date, h, m	ADA	59	
35	occ_date, h, m	SVM	58	
36	occ_date, h, m	RN	55	
37	h, m, jour_num, mois_num	ADA	61	
38	h, m, jour_num, mois_num	SVM	57	
39	h, m, jour_num, mois_num	RN	60	

NumTest	Entrée (s)	Modèle	AUC (en %)	Paramètres
40	h, m, jour_num, mois_num, rainc, snowc	ADA	61	
41	h, m, jour_num, mois_num, rainc, snowc	SVM	58	
42	h, m, jour_num, mois_num, rainc, snowc	RN	61	n=5
43	h, m, jour_num, mois_num, rainc, snowc	RN	60	n=10
44	h, m, jour_num, mois_num, rainc, snowc	RN	61	n=25
45	h, m, jour_num, mois_num, rainc, snowc	RN	59	n=75
46	h, m, jour_num, mois_num, rainc, snowc	RN	60	n=100
47	h, m, jour_char, mois_char, rainc2, snowc2	ADA	62	
48	h, m, jour_char, mois_char, rainc2, snowc3	SVM	59	
49	h, m, jour_char, mois_char, rainc2, snowc4	SVM	55	Polynomial
50	h, m, jour_char, mois_char, rainc2, snowc5	RN	58	
51	h, m, jour_char, mois_char, rainc2, snowc6	RN	59	30
52	h, m, jour_char, mois_char, rainc2, snowc7	RN	61	50
53	h, m, jour_char, mois_char, rainc2, snowc8	RN	50	75
54	h, m, jour_char, mois_char, rainc2, snowc9	RN	60	90
55	district, zone, roadway1c,atom, Location	AD	55	
56	lieu, district, zone, roadway1c,atom, Location	ADA	57	
57	lieu, district, zone, roadway1c,atom, Location	SVM	0	
58	district, zone, roadway1c,atom, Location	RN	0	
59	district, zone, roadway1c,atom, Location, place_name	AD	55	
60	district, zone, roadway1c,atom, Location, place_name	ADA	57	
61	h, m, district, Location, jour_num, mois_num	AD	56	
62	h, m, district, Location, jour_num, mois_num	ADA	63	
63	h, m, district, Location, jour_num, mois_num	SVM	59	

NumTest	Entrée (s)	Modèle	AUC (en %)	Paramètres
64	h, m, district, Location, jour_num, mois_num, tempmin, tempmax, temp mean, rainc, snowc	AD	57	
65	h, m, district, Location, jour_num, mois_num, tempmin, tempmax, temp mean, rainc, snowc	ADA	60	
66	h, m, district, Location, jour_num, mois_num, tempmin, tempmax, temp mean, rainc, snowc	SVM	61	
67	h, m, district, Location, jour_num, mois_num, tempmin, tempmax, temp mean, rainc, snowc	RN	58	
68	h,m,traffic.compl, jour_num,mois_num	AD	75	
69	h,m,traffic.compl, jour_num,mois_num	ADA	80	
70	h,m,traffic.compl, jour_num,mois_num	RN	78	
71	h,m,traffic.compl, jour_num,mois_num, zone, district, location	AD	77	
72	h,m,traffic.compl, jour_num,mois_num, zone, district, location	ADA	82	
73	h,m,traffic.compl, jour_num,mois_num, zone, district, location	RN	80	
74	h,m,traffic.compl, jour_num,mois_num, zone, district, location	RN	78	
75	h,m,traffic.compl, jour_num,mois_num, zone, district, location	SVM	80	5
76	h,m,traffic.compl, jour_num,mois_num, zone, district, location	SVM	76	
77	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	AD	84	Polynomial
78	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	ADA	89	

NumTest	Entrée (s)	Modèle	AUC (en %)	Paramètres
79	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	SVM	89	
80	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	SVM	84	Linéaire
81	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	SVM	88	Polynomial
82	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	RN	88	5
83	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	RN	88	2
84	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	RN	50	20
85	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	RN	90	
86	h,m,Acc.Non.Report, Acc.Traffic.Serv, Impaired.over.08, Fail.To.Remain, Vehicle.Abandon, traffic.compl, Provoffence, Susp90 jour_num,mois_num, zone, district, location, rainc, snowc	RN	90	25

Bibliographie

[1] : World Health Organization, “Global Status report on road safety 2015”, WHO Library Cataloguing-in-Publication Data, pp 9 – 13, 2015, http://www.who.int/mediacentre/factsheets/fs358/fr/.

[2] :Ottawa annual safety report, http://ottawa.ca/en/residents/transportation-and-parking/road-safety/annual-safety-reports#2015-ottawa-road-safety-report

[3]: S. Park S.Kim, Y. Ha, “Highway traffic accident prediction using VDS big data analysis”, Springer, 2016.

[4]: J. Hourdos, V. Garg, P. Michalopoulos, “Accident Prevention Based on Automatic Detection ofAccident Prone Traffic Conditions: Phase I”, Department of Civil Engineering University of Minnesota,pp. 10 -11, 2008.

[5]:L.A. Rodegerdts,B.Nevers, B.Robinson, “Signalized intersections: informational guide”, U.S. Department of transportation, Federal Highway Administration, 2004.

[6]:“Highway Safety Manual User Guide”, National Cooperative Highway Research Program, 2014, http://onlinepubs.trb.org/onlinepubs/nchrp/docs/NCHRP17-50_UserGuide.pdf

[7]: R. Duckworth, M. Imran, J. Chan, “Combined Ranking Method for Screening Collision Monitoring Locations along Alberta Highways”, Alberta Transportation, pp 4 – 12, 2011.

[8]: L. Bates, P. Durdin, “Intersection – Determining the Good, the Bad and the Ugly”, IPENZ Transportation Group Conference, Dunedin, pp. 1-11, 2013.

[9]: C. Brodie, P. Durdin, F. Tate, H. Mackie, “Targeting High Risk Intersections”, 2013 Australasian College of Road Safety Conference – “A Safe System: The Road Safety Discussion”, pp. 2-8, 2013.

[10]: NZ Transport Agency (NZTA), « High-risk intersections guide», 2012, http://www.nzta.govt.nz/assets/consultation/high-risk-intersections-guide/docs/high-risk-intersections-guide.pdf

[11]: H. Wu, “A Framework for Developing Road Risk Indices Using Quantile Regression Based Crash Prediction Model”, Ph.D. Thesis, University of Texas at Austin, 2011.

[12]: J.Hourdos, V.Garg, P.Michalopoulos, "Accident Prevention Based on Automatic Detection ofAccident Prone Traffic Conditions: Phase I", Department of Civil Engineering University of Minnesota, pp. 13 -15, 2008.

[13]:Allstate, "Allstate 2015 Safe Driving Study Results", 2015, http://www.citynews.ca/2015/11/26/car-accidents-on-the-rise-in-canada-toronto-study/.

[14]:5 Times when you're more likely to get in a car accident, https://www.insurancehotline.com/five-times-when-youre-more-likely-to-get-in-a-car-accident.

[15]: Fatal car accident statistics, http://www.hg.org/article.asp?id=29836.

[16]:OPP Statistics Reveal Deadliest Month On Ontario Road, https://www.yd.com/blog/june-most-deadliest-month-on-ontario-roads/.

[17]:"Canadian MotorVehicle TrafficCollision Statistics2013", Minister of Transport, 2015, http://www.tc.gc.ca/media/documents/roadsafety/cmvtcs2013_eng.pdf.

[18]: Z. Li, I. Kolmanovsky, E. Atkins, J. Lu, D. Filev and J. Michelini, "Road Risk Modeling and Cloud-Aided Safety-Based Route Planning", IEEE Trans. On Cybernetics, 2016.

[19]: G. Delashmit, H Bédard, "Accidents: Causes, Analysis and Prevention", Nova Science Publishers, New York, 2009.

[20]: "2009 Saskatchewan Traffic Accident Facts", 2009, https://www.sgi.sk.ca/documents/625510/627017/TAIS_2009_Annual_Report.pdf/45fa424a-492c-40e0-8726-1a700203e90c.

[21]: L. Yuejing, Z. Xing-lin, Z. Haixia, L. Ming, L. Jie, "Research on Accident Prediction of Intersection and Identification Method of Prominent Accident Form Based on Back Propagation Neural Network", International Conference on Computer Application and System Modeling (ICCASM 2010), pp. VI-434-VI-438, 2010.

[22]: P. Liu, S.-H. Chen, and M.-D. Yang, "Study of Signalized Intersection Crashes Using Artificial Intelligence Methods", A. Gelbukh and E.F. Morales (Eds.): MICAI 2008, LNAI 5317, pp. 987–997, 2008.

[23] W. Huiying, L. Jun, C. Xiaolong, G. Xiaohui, “Real-time Highway Accident Prediction Based on Grey Relation Entropy Analysis and Probabilistic Neural Network”, pp. 1420-1423, 2011.

[24]: J.-W. Hwang, Y.-S. Lee, and S.-B. Cho, “Hierarchical Probabilistic Network-based System for Traffic Accident Detection at Intersections”, Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, pp. 211-216, 2010.

[25]: S. Li, D. Zhao, “Prediction of Road Traffic Accidents Loss Using Improved Wavelet Neural Network”, IEEE Conf. Computers, Communications, Control and Power Engineering, pp. 1526-1529, 2002.

[26]: Y. Lv, S. Tang, H. Zhao, and S. Li, “Real-time Highway Accident Prediction based on Support Vector Machines”, Chinese Control and Decision Conference. pp. 4403-4407, 2009.

[27]: R. Yu, M. Abdel-Aty, “Investigating the different characteristics of weekday and weekend crashes”, Journal of Safety Research,vol.46,pp. 91–97, 2013.

[28] N. Dong, H. Huang, L. Zheng, “Support vector machine in crash prediction at the level of traffic analysis zones: Assessing the spatial proximity effects”, Accident Analysis and Prevention 82,pp.192–198, 2015.

[29]: R. Gang, Z. Zhuping, “Traffic safety forecasting method by particle swarm optimization and support vector machine”, Expert Systems with Applications, vol. 38,pp.10420–10424, 2011.

[30] M. Hosseinpour, A. S. Yahaya, S. M. Ghadiri, and J. Prasetijo, “Application of Adaptive Neuro-Fuzzy Inference System for Road Accident Prediction”, KSCE Journal of Civil Engineering 17(7), pp. 1761-1772, 2013.

[31]: X. Zhu, “Application of Composite Grey BP Neural Network Forecasting Model to Motor Vehicle Fatality Risk”, 2010 Second International Conference on Computer Modeling and Simulation, pp. 236-240, 2010.

[32]: X. Xu, B. Chen and F. Gan, “Traffic Safety Evaluations Based on Grey Systems Theory and Neural Network”, 2009 World Congress on Computer Science and Information Engineering, pp. 603-607, 2009.

[33]: K. Polat, and S. S. Durduran, “Automatic determination of traffic accidents based on KMC-based attribute weighting“, Neural Comput. &Applic. 21, pp. 1271–1279, 2012.

[34]:Q. Wuyong, D. Yaoguo, M. Sen, L. Xuemei, "The Intelligent Optimization of GM(1,1) Power Model and its Application in the Forecast of Traffic Accident", IEEE International Conference on Grey Systems and Intelligent Services (GSIS), pp. 385-389, 2011.

[35]: X.-F. Zhang, L. Fan, "A Decision Tree Approach for Traffic Accident Analysis of Saskatchewan Highways", 26th IEEE Canadian Conference Electrical and Computer Engineering, pp. 1-4, 2013.

[36]: T. Beshah, D. Ejigu, A. Abraham, V. Snasel, P. Kromer, "Knowledge Discovery from Road Traffic Accident Data in Ethiopia: Data Quality, Ensembling and Trend Analysis for Improving Road Safety", Neural Network World, 22(3), pp. 215-244, 2012.

[37]: L.-Y. Chang, H.-W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques", Accident Analysis and Prevention 38, pp. 1019–1027, 2006.

[38]: [GPY15]: Q. Guo, X. Pei, D. Yao, S. Wong, "Role of street patterns in zone-based traffic safety analysis," J. Cent. South Univ. 22, pp. 2416−2422, 2015.

[39]: T. Sayed, P. de Leur, "Collision Prediction Models for British Columbia", Technical Report for BC Ministry of Transportation & Infrastructure, 2008.

[40]: Z. Ma, C. Shao, H. Yue, S. Ma, "Analysis of the Logistic Model for Accident Severity on Urban Road Environment", pp. 983-987, 2009.

[41]: J. Pahukula, S. Hernandez, A. Unnikrishnan, "A time of day analysis of crashes involving large trucks in urban areas", Accident Analysis and Prevention 75, pp. 155–163, 2015.

[42]: S. Park, K. Jang, S. H. Park, D.-K. Kim, and K. S. Chon, "Analysis of Injury Severity in Traffic Crashes: A Case Study of Korean Expressways", KSCE Journal of Civil Engineering, 16(7):1280-1288, 2012.

[43]: E. Moons, T. Brijs, and G. Wets, "Improving Moran's Index to Identify Hot Spots in Traffic Safety", B. Murgante, G. Borruso, A. Lapucci (Eds.): Geocomputation& Urban Planning, SCI 176, pp. 117–132, 2009.

[44]: C. Xu, P. Liu, W. Wang, and X. Jiang, "Development of a Crash Risk Index to Identify Real Time Crash Risks on Freeways", KSCE Journal of Civil Engineering (2013) 17(7):1788-1797, 2013.

[45]: X. Zhan, H.M.A. Aziz, S. V. Ukkusuri, “An efficient parallel sampling technique for Multivariate Poisson-Lognormal model: Analysis with two crash count datasets”, Analytic Methods in Accident Research 8 (2015) 45–60.

[46]: Y. Lv, S. Tang, H. Zhao, “Real-time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method”, 2009 International Conference on Measuring Technology and Mechatronics Automation. pp. 547-550, 2009.

[47]: T. Beshah, S. Hill, “Mining Road Traffic Accident Data to Improve Safety: Role of Road-related Factors on Accident Severity in Ethiopia”, AAAI, 2010.

[48]: J. Wang and X. Wang, “An Ontology-Based Traffic Accident Risk Mapping Framework”, D. Pfoser et al. (Eds.): SSTD 2011, LNCS 6849, pp. 21–38, 2011.

[49]: R. Jagannathan, S. Petrovic, G. Powell, and M. Roberts, “Predicting Road Accidents Based on Current and Historical Spatio-temporal Traffic Flow Data”, D. Pacino, S. Voß, and R.M. Jensen (Eds.): ICCL 2013, LNCS 8197, pp. 83–97, 2013.

[50] J. Wua, M. Abdel-Aty, R. Yu, Z. Gao, “A novel visible network approach for freeway crash analysis”, Transportation Research Part C 36 (2013) 72–77.

[51]: D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, and P.C. Sabeti, “Detecting novel associations in large data sets”, Science 334, 1518–1524, 2011.

[52]:P.Chapman, J.Clinton, R. Kerber, T. Khabaza, T.Reinartz, C. Shearer and R.Wirth, “CRISP-DM 1.0, step-by-step data mining guide”, pp. 11 – 12, 2000.

[53]: “IBM SPSS Modeler CRISP-DM Guide”, IBM corporation, 2011, ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf

[54]:Data Mining Methodology,https://www.kdnuggets.com/polls/2004/data_mining_methodology.htm, April 2004.

[55]: Tableau Software: Business Intelligence and Analytics, https://www.tableau.com/.

[56]: http://climat.meteo.gc.ca/historical_data/search_historic_data_f.html.

[57] : J. Ledolter, “Data Mining And Business Analytics With R”, Wiley, pp 193 – 195, 2013.

[58]:S. Nissen, “Création d'un réseau de neurones – c'est facile”, http://fann.sourceforge.net/fann_fr.pdf.

[59]:G. Petitjean, “Introduction aux réseaux de neurones”, https://www.lrde.epita.fr/~sigoure/cours_ReseauxNeurones.pdf.

[60]: B. E. Boser, I. M. Guyon, V. N. Vapnik, “A Training Algorithm for Optimal Margin Classifiers in Fifth Annual Workshop on Computational Learning Theory”, pages 144--152, Pittsburgh, ACM, 1992.

[61]:“Machines à vecteurs supports”, wikistat, pp 2 – 3.

62]: R. E. Schapire, “Explaining AdaBoost”, http://rob.schapire.net/papers/explaining-adaboost.pdf

[63]: A. Bujari and C. E. Palazzi, «Intersection Collision: Causes and Avoidance Techniques», R. Naja (ed.), Wireless Vehicular Networks for Car Collision Avoidance, pp. 189-227.

[64]: W. Hu, X. Xiao, D. Xie, T. Tan and S. Maybank, “Traffic Accident Prediction Using 3-D Model-Based Vehicle Tracking”, IEEE Trans. Vehicular Technology, vol.53, no.3, pp. 677- 694, May 2004.

[65]: Z. Yingxue, “Analysis the Relation between Highway Horizontal Curve and Traffic Safety”, 2009 International Conference on Measuring Technology and Mechatronics Automation, pp. 479-481,2009.

[66]: Z.W. Chang, J.-J. Wang, « Discussion on emergency traffic organization programs of expressway traffic accident under circumstance of road network», 2009 Second International Conference on Intelligent Computation Technology and Automation, pp. 571-574, 2009.

[67]: A. Hassen, A. Godesso, L. Abebe and E. Girma, “Risky driving behaviors for road traffic accident among drivers in Mekele city, Northern Ethiopia”, Hassen et al. BMC Research Notes 2011, 4:535, pp. 1-6, 2011.

[68]: M. Fogue, P. Garrido, F. J. Martinez, J.-C. Cano, T. Calafate, and P. Manzoni, «Using Data Mining and Vehicular Networks to Estimate the Severity of Traffic Accidents», J. Casillas et al. (Eds.): Management Intelligent Systems, AISC 171, pp. 37–46, 2012.

[69]: X. Binglei, H. Zheng, M. Hongwei, « Fuzzy-Logic-Based Traffic Incident Detection Algorithm for Freeway», Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, pp. 1254-1259, 2008.

[70]: U. Er, S. Yüksel, O. Aköz, M. E. Karslıgil, “Traffic Accident Risk Analysis Based on Relation of Common Route Models”, 21st International Conference on Pattern Recognition (ICPR 2012), Japan, pp. 2561-2564, 2012.

[71]:A. Cornuéjols, "Les séparateurs à vastes marges (SVM) et les méthodes à noyaux", AgraParisTech - INRA MIA 518, https://www.lri.fr/~antoine/Courses/Master-ISI/ISI-10/Tr-cours-SVM_2014_2x2.pdf.

[72]: B. Lantz, “Machine Learning with R”, Packt Publishing, Birmingham, pp.69, 2013.

[73]: G. James, D. Witten, T. Hastie, and R. Tibshirani “An Introduction to statistical learning with R”, Springer, pp 311-312, 2013.

[74]: N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-Sampling Technique,” Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[75]: P. Sobhani, H. Viktor, and S. Matwin, “Learning from Imbalanced Data Using Ensemble Methods and Cluster-based Undersampling,” International Workshop on New Frontiers in Mining Complex Patterns. Springer, Cham, pp. 69–83, 2014.

[76] : Transportation Services – Organizations, http://data.ottawa.ca/organization/transportationservices.

[77]: A. Chisholm, “Exploring Data with Rapid Miner: Explore, Understand, and Prepare Real Data Using Rapid Miner's Practical Tips and Tricks”, Packt Publishing, Birmingham, 2013.

[78]:“A gentle introduction to the gradient boosting algorithm for machine learning”, https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/.

[79]: N. Ye, “Data Minig: Theories, Algorithms and Examples”, CRC Press (Taylor et Francis group), pp 31-34, 2014.

[80]: L. Torgo, “Data Mining with R, Learning with case studies”, CRC Press (Taylor et Francis group), pp 255-256, 2011.

[81]: RapidMiner, “RapidMiner Studio Manual”, rapidMiner, 2014, https://docs.rapidminer.com/downloads/RapidMiner-v6-user-manual.pdf

[82]: G. James, D. Witten, T. Hastie, and R. Tibshirani “An Introduction to statistical learning with R”, Springer, pp 176-178, 2013.

[83]: L.-Y. Chang, H.-W. Wang, “Analysis of traffic injury severity: An application of non-parametric classification tree techniques”, Accident Analysis and Prevention 38 (2006) 1019–1027.

[84]: D. Saha, P. Alluri, A. Gan, ”Prioritizing Highway Safety Manual’s crash prediction variables using boosted regression trees”, Accident Analysis and Prevention 79 (2015) 133–144.

[85]: D. Chetchotsak, S. Pattanapairoj, B. Arnonkijpanich, “Integrating new data balancing technique with committee networks for imbalanced data: GRSOM approach”, Springer, 2015.

[86]: Y. Zhang, “Severity analysis in motor vehicle crashes in the state of Iowa using multiple machine learning and data balancing techniques”, ProQuest Dissertation publishing, pp 36-38, 2017

[87]: R. O. Mujalli, G. Lopez, L. Garah, “Bayes classifiers for imbalanced traffic accidents datasets”, Elsevier, 2016.