Modélisation et l'analyse dans le domaine de l'assurance en utilisant les techniques de Datamining

Modeling and analysis in the field of insurance using Datamining techniques

SOUAKRI Roufaida* [1], SAHRAOUI Abdelaziz [2]

[1] USTHB (Algérie), rsouakri@usthb.dz

[2] université de Msila (Algérie), abdelaziz.sahraoui@univ-msila.dz

Reçu : .../.../2022 *Accepté : .../.../2022* *Publié : .../.../2022*

Résumé :

Dans cet article, on s'intéresse aux éléments qui expliquent le nombre d'accidents déclarés. En règle générale, des méthodes de classification de Data Mining (la technique des arbres de décision) sont utilisés dans la modélisation de la fréquence des sinistres et la détermination du coût de ces derniers Cette technique nous l'avons appliqué sur les données recueillies auprès de la Société Algérienne d'Assurance (SAA), après une étude théorique de l'assurance et la branche automobile, et la présentation de quelques notions de base de Datamining.

Les mots clés : Data Mining ; modélisation ; assurance ; sinistres ; automobile

Abstract:

In this article, we are interested in the elements that explain the number of declared accidents. Typically, data mining classification methods (the decision tree technique) are used in claims frequency modeling and cost determination of claims, we applied this technique to the data collected from the Algerian Insurance Company (SAA), after a theoretical study of insurance and the automobile branch, and the presentation of some basic notions of Datamining.

Keywords: Data Mining; modeling; Insurance; claims; automobile.

Auteur correspondant.

1. INTRODUCTION

Le monde est aujourd'hui en proie à d'énormes changements. La volonté de se protéger contre les aléas de l'existence nous pousse à appréhender le futur et à avoir un réflexe conservateur, nous avons peur du lendemain car nous évoluons dans l'incertitude, il importe donc que tout agent économique mette en œuvre des moyens lui permettant de s'adapter à un tel environnement. A cet effet, par mesure de précaution, ces agents se prémunissent auprès des sociétés ou des associations contre des risques éventuels dont la réalisation serait ainsi de répondre à un besoin impérieux de protection des personnes et de leurs activités, Ces sociétés ou associations sont appelées les Assurances.

L'assurance trouve son principe dans l'esprit de solidarité, ce besoin d'association qui pousse les individus à s'unir et à se grouper pour mieux défendre un péril commun, contre lequel chacun, pris isolément s'avère impuissant. Pour cela l'individu doit apporter la contribution de son effort personnel à la communauté des assurés qui, en échange et par réciprocité, assure sa protection. L'assurance joue un rôle non négligeable et important dans la contribution de la croissance économique.

Les réformes économiques engagées en Algérie ont pour principal objectif la libéralisation de toutes les activités. A l'instar des autres secteurs, celui des assurances est en train de vivre des changements tant dans sa structure que dans son organisation.

L'assurance automobile constitue la principale branche du marché des assurances algérien, ce secteur voit malgré tout un ralentissement de sa croissance, ce qui est dû à la chute des importations algériennes de véhicules

La sinistralité en assurance automobile est un problème important pour les pays industrialisés. Pour les assureurs, elle se mesure en termes de fréquence des accidents et du montant de ces accidents.

Aujourd'hui, les compagnies d'assurances ont engrangé des masses de données importantes. En effet, les faibles coûts des machines en termes de stockage et de puissance ont encouragé les compagnies à accumuler toujours plus d'informations. Selon les compagnies d'assurances, l'estimation de la quantité de données collectées dans le monde double tous les 20 mois, alors que les informations à valeur ajoutée soutirées de ces données n'augmentent que très peu.

Dans cette optique, la constitution d'une base de données, regroupant sous une forme homogène et cohérente toutes les données de la compagnie d'assurance, offre des prospectives nouvelles aux utilisateurs, notamment en termes d'extraction de connaissances grâce aux outils de data Mining (fouille de données). Le développement récent et l'analyse des outils actuariels soulèvent des problématiques auxquelles il est tentant d'appliquer les méthodes et techniques de fouille de données.

Dans le marché fortement concurrentiel de l'assurance automobile, qui représente la branche la plus importante de l'assurance non-vie, les sociétés d'assurances cherchent à déterminer des facteurs qui contribuent à expliquer la

sinistralité. Ces facteurs lui permettent, en construisant des classes de risque, de segmenter son portefeuille et de hiérarchiser ces classes à l'aide d'indicateurs de sinistralité, comme la prime pure. Cette démarche vise à obtenir une bonne adéquation entre la sinistralité et les primes payées par les assurés. Afin de répondre à cet objectif, on s'intéresse à la sélection des principaux facteurs qui contribuent à expliquer la sinistralité en termes de fréquence et de coût en appliquant les méthodes de Data Mining dans un cadre d'apprentissage supervisé.

Dans cet article, on s'intéresse aux éléments qui expliquent le nombre d'accidents déclarés. En règle générale, des méthodes de classification de Data Mining (**la technique des arbres de décision**) sont utilisés dans la modélisation de la fréquence des sinistres. Cette technique nous l'avons appliqué sur les données recueillies auprès de la Société Algérienne d'Assurance (SAA).

De ce fait, l'objective de la présente étude est de répondre à la problématique suivante :

« Quelles sont les variables les plus déterminantes susceptibles d'influencer sur l'augmentation de la fréquence de sinistres et la détermination du coût de ces derniers ?»

Pour pouvoir répondre à notre problématique, nous avons défini quelques hypothèses qui nous aiderons dans notre démarche :

- Le système de règlement des sinistres par la méthode Bonus/Malus est capable de responsabiliser les assurés.
- Les critères liés aux conducteurs l'âge du conducteur et sa profession, les critères liés aux véhicules tels que la puissance, type d'énergie et usage, les garanties souscrites.
- Parmi ces critères, le plus influant la puissance du véhicule.
- Les facteurs qui rentrent en considération lors du règlement de sinistres sont :

  Les garanties souscrites ; par exemple si l'assuré achète la garantie bris de glace, et si le sinistre est survenu, l'assureur va indemniser l'assuré.

2. l'aspect théorique sur l'assurance et la branche intéressant l'automobile

2.1 Aspects généraux sur l'assurance

Le souci du lendemain et le dessein de l'avenir sont le propre de l'homme, et c'est cela qui fait naître en chaque homme le besoin de sécurité optimum et de s'assurer contre les aléas du sort.

Afin de mieux cerner le domaine des assurances, et de bien comprendre son fonctionnement, ce qui vient répondre aux questions élémentaires relatives à l'assurance.

2.1.1 Définition de l'assurance

L'assurance peut être définie selon plusieurs optiques, on proposera ci-dessous quelques définitions juridiques et techniques.

Définition juridique

Selon l'article 2 de l'ordonnance n°95/07 du 25 janvier 1995 relative aux assurances modifiée et complétée par la loi n°06/04 du 20 février 2006, selon l'article 619 du code civile : « l 'assurance est un contrat par lequel l'assureur s'oblige, moyennant le paiement des primes ou autres versements pécuniaire, à fournir à l 'assuré ou au tiers bénéficiaire au profit duquel l'assurance est souscrite, une somme d 'argent, une rente ou une autre prestation pécuniaire, en cas de réalisation du risque prévu au contrat ».

Définition technique

Il existe une nuée de définitions techniques, nous citons dans ce qui suit les plus substantielles, techniquement l'assurance peut être définie comme étant :

« Un système de transfert de risques qui s'effectue selon les normes juridiques légales et conventionnelles moyennant paiement d'une certaine somme (la prime ou cotisation) ».

« L'opération par laquelle un assureur, organisant en mutualité une multitude d'assurés exposés à la réalisation de certains risques, ceux d'entre eux qui subissent un sinistre grâce à la masse commune des primes collectées ».

2.1.2 Eléments constitutifs de l'assurance

Les deux définitions de l'assurance ont l'avantage de faire ressortir les éléments qui caractérisent l'opération de l'assurance.

Mutualité : La mutualité est « le principe de base de l'assurance selon lequel les cotisations modiques versées par chacun des membres d'un groupe de personnes (les assurés) sont utilisées et suffisent théoriquement à l'indemnisation de quelques-unes d'entre elles qui s'avèrent victimes de l'événement assuré »

À cet effet, le rôle de l'assureur est de mutualiser les risques : les mettre en commun, les répartir et les compenser en s'appuyant sur des lois mathématiques appliquées sur les statistiques collectées.

Assureur : L'assureur est celui qui s'engage dans le contrat et qui accepte le risque et détermine la cotisation à verser. Pour le public l'assureur est souvent l'intermédiaire qui représente la société d'assurance.

Assuré : Plusieurs définitions peuvent être avancées selon la nature de l'assurance en cause :

En assurance de personne, l'assuré est l'individu, désigné au contrat, sur la tête du quelle repose e risque de décès ou de vie, appelé couramment « tête assuré ».

En assurance de choses, il s'agit de la personne physique ou morale, désigné au contrat, dont les biens font l'objet de la garantie, ou qui, par les faits d'une assurance pour compte de qui il atteindra, se trouve tête bénéficiaire de cette garantie.

Cependant, dans le domaine de l'assurance automobile, l'assuré est le propriétaire du véhicule. En assurance habitation, l'assuré est le propriétaire ou l'habitation, l'assuré est le propriétaire ou le locataire immobilier. En assurance de

personne, l'assuré est la personne sur laquelle repose le risque (décès, maladie, invalidité).

Souscripteur Le souscripteur est la personne qui souscrit un contrat d'assurance, signe les documents contractuels et qui se trouve seule engagées envers l'assureur pour le paiement des primes. Il peut le souscrire pour son propre compte, en tant qu'assuré ou pour le compte de personnes faisant partie d'un groupe en tant que contractant.

Bénéficiaire : Représente la personne qui reçoit l'indemnisation par l'assureur en cas de réalisation du risque. Une même personne peut être à la fois souscripteur, assuré et bénéficiaire de la garantie, c'est le cas en assurance de dommages alors qu'en assurance de personnes, le souscripteur diffère souvent de l'assuré et du bénéficiaire.

Contrat d'assurance : Le contrat de l'assurance est la convention par laquelle une partie appelée « l'assuré » se fait promettre une prestation pécuniaire fournie par une autre partie appelée « l'assureur » en cas de réalisation d'un risque moyennant le paiement d'une prime ou d'une cotisation.

Risque : Le risque est un événement futur et aléatoire dont la survenance ne dépend pas exclusivement de la volonté de l'assuré.

Dans le sens courant, les risques sont les coups du sort contre lesquels on désire se prémunir. Dans le vocabulaire des assureurs, il représente l'ensemble des périls couverts par l'assurance et classés dans une même catégorie. Il est aléatoire et incertain dont la survenance ne dépend pas exclusivement de la volonté de l'assuré.

Prime (ou cotisation) : Elle peut être définie comme étant :

« La somme due par le souscripteur à l'assureur en contrepartie du risque pris en charge par celui-ci, payable d'avance par versements généralement périodique ».

Prestation de l'assureur :

L'assureur s'engage à indemniser l'assuré en cas de sinistre, il s'agit d'une somme d'argent versée au souscripteur et assuré, ou bien au tiers ou au bénéficiaire comme dans l'assurance vie en cas de décès.

Sinistre : Le sinistre est la réalisation d'un risque entrant dans l'objet du contrat d'assurance. Le sinistre fait naître l'obligation pour une entreprise d'assurance d'exécuter la garantie prévue dans un contrat d'assurance.

2.1.3 Importance de l'assurance dans les économies modernes :

Aujourd'hui, l'assurance est souvent perçue comme un frein au développement de l'activité économique, elle est devenue une branche majeure de l'économie, le chiffre d'affaires des assureurs ne cesse jamais de progresser et représente un pourcentage très croissant d'année en année du Produit Intérieur Brut

(PIB) de chaque pays. La part de l'assurance dans le PIB peut atteindre jusqu'au 15% et généralement d'autant plus élevée que le pays considéré a atteint un niveau de développement économique important.

L'assurance est un moteur de développement économique et social.

2.2 l'Assurance automobile en Algérie :

L'assurance automobile appartient aux groupes des opérations d'assurance qui non pas pour objet la vie de l'assuré, elle représente d'un côté, une part très importante du patrimoine des individus, d'un autre côté, elle est exposée aux accidents qui causent la mortalité, la chose qui a poussé les états de rendre ce genre d'assurance obligatoire pour sa partie de responsabilité civile.

Pour cela le produit d'assurance le plus familier du grand public est l'assurance automobile, c'est grâce à ce contrat que la victime d'un accident automobile est indemnisée par la compagnie assurant le responsable de l'accident.

2.2.1 L'assurance automobile est obligatoire :

Selon l'article premier de l'ordonnance 74/15 du 30 janvier 1974 relative à l'obligation d'assurance des véhicules automobiles et au régime d'indemnisation des dommages : « Tout propriétaire d'un véhicule doit, avant de le mettre en circulation, souscrire une assurance couvrant les dommages causés aux tiers par ce véhicule ».

Quelles sont les personnes soumises à l'obligation d'assurance automobile ?

Généralement c'est le propriétaire du véhicule ou toute autre personne ayant l'obligation de garde et le droit d'usage du véhicule.

2.2.2 Le contrat d'assurance automobile

Le contrat d'assurance automobile est en général un contrat « multirisques » destiné à couvrir des risques aussi divers.

3. Datamining concepts et techniques

3.1 Définition

Le datamining est l'art d'extraire des connaissances à partir de données. Les données peuvent être stockées dans des entrepôts (data Warehouse), dans des bases de données distribuées ou sur internet (web mining). Le datamining ne se limite pas au traitement des données structurées sous formes de tables numériques ; il offre des moyens pour aborder les corpus en langage naturel (text mining), les images (image mining), le son (Sound mining) ou la vidéo et dans ce cas, on parle alors plus généralement de (MultiMedia mining**).**

D'après Kantardzic Mehmed dans son livre **Datamining, concepts, models, methods and algorithms** " Le data mining, ou fouille de données, est l'ensemble des méthodes et techniques destinées à l'exploration et l'analyse de bases de données informatiques (souvent grandes), de façon automatique ou semi-automatique, en vue de détecter dans ces données des règles, des associations, des tendances inconnues ou cachées, des structures particulières restituant l'essentiel de l'information utile tout en réduisant la quantité de données"

3.2 Les arbres de décisions

De façon plus précise, le datamining nécessite l'application de méthodes statistiques telles que l'ACP[1], la CAH[2], les K-plus proche voisins… ou des méthodes algorithmiques telles que l'arbre de décision, réseaux neurones… etc.

3.2.1 Notion des arbres de décision

Les arbres de décisions sont des outils d'aide à la décision qui permettent selon des variables discriminantes de répartir une population d'individus en groupes homogènes en fonction d'un objectif connu. Les arbres de décision sont des outils puissants et populaires pour la classification et la prédiction. Un arbre de décision permet à partir des données connues sur le problème de donner des prédictions par réduction, niveau par niveau, du domaine des solutions.

Un arbre de décision est un enchainement hiérarchique de règles logiques construites automatiquement à partir d'une base d'exemples. Un exemple est constitué d'une liste d'attributs dont la valeur détermine l'appartenance à une classe donnée. La construction de l'arbre de décision consiste à utiliser les attributs pour subdiviser progressivement l'ensemble d'exemples en sous-ensembles de plus en plus fin.

3.2.2 Principe de l'arbre de décision

La technique de l'arbre de décision est employée en classement pour détecter des critères permettant de répartir les individus d'une population en n classes (souvent n=2) prédéfinies. On commence par choisir la variable qui, par ses modalités, sépare le mieux les individus de chaque classe, de façon à avoir des sous-populations, que l'on appelle nœuds, contenant chacune la plus forte proportion possible d'individus d'une seule classe, puis on réitère la même opération sur chaque nouveau nœud obtenu jusqu'à ce que la séparation des individus ne soit plus possible ou plus souhaitable (au vu de certains critères dépendant du type d'arbre).

Par construction les nœuds terminaux (les feuilles) sont tous majoritairement constitués d'individus d'une seule classe. Un individu est affecté à une feuille, et donc à une certaine classe avec une assez forte probabilité, quand il satisfait l'ensemble des règles permettant l'arriver à cette feuille. L'ensemble des règles de toutes les feuilles constitue le modèle de classement.

3.2.3 Construction de l'arbre de décision :

L'arbre est construit d'un ensemble de règles de classification basant leur décision sur des tests associés aux attributs, organisés de manière arborescente.

Un attribut peut être de nature qualitative ou quantitative en fonction de l'ensemble des valeurs qu'il peut prendre. Un attribut est qualitatif si on ne peut pas en faire une moyenne ; sa valeur est d'un type défini en extension (une couleur, une marque de voiture, ...). Attribut quantitatif Sinon, l'attribut est de nature

[1] Analyse en composantes principales.

[2] Classification ascendante hiérarchique.

quantitative : un entier, un réel, ... ; il peut représenter un salaire, une surface, un nombre d'habitants. Un attribut peut également être un enregistrement (une date par exemple), donc composé lui-même de sous-attributs (jour, mois, année dans le cas d'une date).

L'arbre débute par la racine qui se divise en banches, Les branches conduisent à des nœuds qui peuvent à leur tour se diviser en branches, ce qui donne naissance, pour chaque nœud, à deus nœuds-fils ou plus, chaque nœud-fils donne à son tour naissance à deux nœuds ou plus. Un nœud terminal est le résultat final d'un chemin de décision et appelé aussi une feuille.

Les nœuds de l'arbre testent les attributs. Un nœud est défini par le choix conjoint d'une variable test parmi les explicatives et d'une division qui induit une partition en deux classes. Il y a une branche pour chaque valeur de l'attribut testé, Les feuilles spécifient les catégories et contiennent les décisions de classement final.

Le processus d'apprentissage de l'arbre de décision passe par deux phases importantes :

- **Phase d'apprentissage :** qui consiste à prélever un premier échantillon, appelé échantillon d'apprentissage (training set) dont le classement est connu pour construction le modèle de prédiction.
- **Phase de validation :** a pour but de mesurer la performance du modèle sur un autre échantillon non étiqueté, appelée échantillon test (test set). Le modèle qui en découle est utilisé pour classer de nouvelles données.

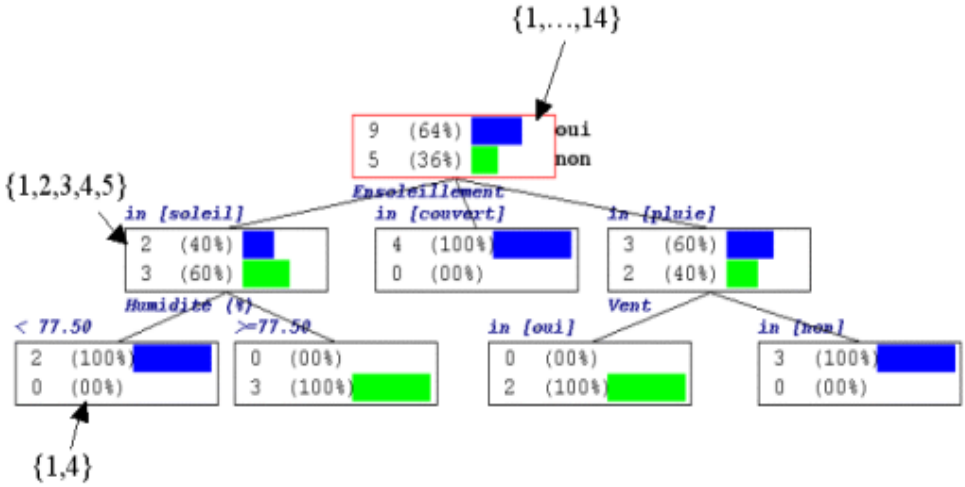
3.2.4 L'algorithme de l'arbre de décision nécessite :

- La définition d'un critère permettant de sélectionner la "meilleure" division parmi toutes celles admissibles pour les différentes variables,
- Une règle permettant de décider qu'un nœud est terminal : il devient ainsi une feuille.
- L'affectation de chaque feuille à l'une des classes ou à une valeur de la variable à expliquer**.**

Exemple d'arbre de décision :

Un exemple qui est présenté dans l'ouvrage de Quinlan (1993). Il s'agit d'expliquer le comportement des individus par rapport à un jeu {jouer, ne pas jouer} à partir des prévisions météorologiques.

Figure 1 : Exemple d'arbre de décision



4. Partie pratique : Extraction des connaissances et interprétation des résultats

4.1 Prétraitement

La phase de prétraitement des données est souvent la plus laborieuse et celle qui demande le plus de temps. Cette première phase est cruciale car du choix des descripteurs et de la connaissance précise de la population que va dépendre la mise au point des modèles de prédiction. L'information nécessaire à la construction d'un bon modèle de prévision peut être disponible dans les données mais un choix inapproprié de variables ou de l'échantillon d'apprentissage peut faire échouer l'opération.

4.1.1 Origine des données

Les données de notre étude proviennent d'un portefeuille d'assurance automobile de la société algérienne d'assurance (Agence N'gaous de la wilaya de Batna) de l'année 2017. Il s'agit d'un certain nombre d'informations dont l'assureur a besoin lors d'une souscription à une Garantie donnée.

Nous disposons d'une base de données initiale constituée de n = 538 lignes et p = 55 colonnes.

Nous avons effectué un travail préliminaire et très déterminant pour la suite de l'étude. Il s'agit du formatage de la base de données. Toutefois, notons qu'à l'issue de ce travail de vérification des données, notre jeu de données est désormais constitué de 488 observations et 24 variables.

Nous avons particulièrement utilisé le logiciel Excel à travers des tableaux croisés dynamiques pour l'exploration de notre base de données.

4.1.2 Choix et présentation des variables

Les variables sont classées en quatre groupes : celles qui caractérisent le conducteur, celles qui caractérisent le véhicule assuré, celles qui caractérisent le

contrat d'assurance, et enfin celles qui caractérisent la sinistralité du véhicule assuré.

Les variables qui caractérisent le conducteur :

- ✓ **SEXE :** variable qualitative binaire qui indique le sexe du preneur d'assurance. Ses modalités sont : Homme et Femme.
- ✓ **AGE de l'assuré :** variable à valeurs entières qui donne l'âge du preneur d'assurance à la date de souscription. Cette valeur est obtenue à partir de sa date de naissance et de l'année de souscription.
- ✓ **AGE du PERMIS** : variable qualitative binaire qui indique l'âge du permis de conduire du preneur d'assurance, à la date de souscription. Cette valeur est obtenue à partir de sa date de délivrance et de l'année de souscription à la garantie. Les modalités sont : plus 1 an et moins 1 an.
- ✓ **Commune :** dans notre base de données, on a 22 communes mais après analyse statistique, nous avons retenu les communes qui représentent le plus grand nombre d'individus, les autres ont été intégrées à la commune la plus proche. Nous obtenons ainsi quatre modalités pour cette variable :

  5600 N'gaous) ,05610(Soufiane) ,05680(OSS) ,05660(Taxlent).

Les variables qui caractérisent le véhicule assuré :

Il existe plusieurs variables qui caractérisent le véhicule, parmi lesquelles ont cité : le genre, la marque, le type, l'usage que l'on fait, le poids, l'énergie, la puissance…ETC.

Dans le cadre de cette étude, nous avons décidé de garder les variables suivantes :

- ✓ **Genre :** au départ nous avions 9 genres :

  00 : véhicules particuliers sans remorque
  30 : véhicule dont le poids total excède 3,5 T
  31 : Remorques dont le poids excède 3,5 T
  32 : TPM dont CU excède 2T sans mat.inf
  33 : Remorques TPM (quintaux)
  34 : transport public des voyageurs (TPV)
  36 : Tracteurs routiers seuls (puissance)
  45 : Engins chantiers TP utilises/voie publique
  50 : Tracteur pneumatique avec remorque

Nous nous intéressons dans cette étude, uniquement aux véhicules particuliers sans remorque.

- ✓ **Marque :** Initialement**,** notre base de données comportait 42 marques, nous avons décidé de sélectionner les marques qui représentent le plus grand nombre d'individus. De plus, nous avons créé une nouvelle modalité dite « autre » pour intégrer toutes les autres marques marquées par un faible nombre d'individus.

  Une fois cette opération réalisée, nous obtenons 4 modalités pour cette variable : HYUNDAI, PEUGEOT, RENAULT, Autre.

- ✓ **Usage :** est une variable qualitative décrivant l'usage ou la catégorie d'usage du véhicule. Les véhicules sont classés en 6 catégories d'usage *: Affaire, Commerce, Fonctionnaire, Commerce C. Bis, Location, Taxi.*

Nous avons opéré comme pour les autres variables : fusion des modalités de faible effectif. Ainsi les deux modalités *Commerce et Commerce C. Bis* ont été regroupée sous une nouvelle modalité dite *Commerce* et les trois modalités *fonctionnaire, location et Taxi* sous une nouvelle modalité dite *Fonctionnaire*.
Ce qui nous donne, pour la variable Usage trois modalités, à savoir : *Affaire, Commerce, Fonctionnaire.*

- ✓ **Poids :** est variable qualitative à trois modalités (Leger, Lourd, Autre), elle indique le poids du véhicule assuré. Les deux modalités Lourd et Autre représentent un faible effectif, nous avons dû les fusionner dans la modalité Autre.
- ✓ **Energie :** est une variable qualitative à deux modalités qui indique la source d'énergie du moteur, Essence (ES ou Diesel (DS).
- ✓ **Puissance :** la puissance en chevaux d'un véhicule désigne par défaut la puissance maximum fournie par le moteur, variable qualitative à 13 modalités, nous avons choisi d'en garder 4 qui sont : $\leq 4CV$, 5CV, 6CV, $\geq 7CV$.
- ✓ **Nombre de place :** est **variable** qualitative initialement à 7 modalités, que nous avons réduites à 3 modalités : 2Plac, 3Plac, 4Plac.

Les variables qui caractérisent le contrat d'assurance :

Ces variables sont : N° police, Durée, Les garanties, Bonus, Malus, Pénalité pour âge, Pénalité pour permis, Prime à payer.

Nous retenons les variables suivantes :

- ✓ **La Durée :** variable quantitative binaire qui désigne la durée du contrat d'assurance, ses modalités sont : 6 mois et 1 an.
- ✓ **Les garanties** : Les garanties existantes dans notre base de données sont listées dans le tableau suivant :

Tableau 1 : Garanties de l'assurance automobile

Les garanties	Abréviation
Responsabilité Civile	RC
Responsabilité Civile interarabe	RC interarabe
Tout risque simple	TR simple
Dommage Collusion	DC
Dommage Collusion valeur Vénale	DC VV
Vole autoradio de véhicule	VAV
Vole incendie de véhicule	VIV
Bris de Glaces	BDG

Bris de Glaces panoramique	BDG panoramique
Défense et recours	DR
Personnes Transportées Assurées	PTA
Assistance au véhicule	Assist véhicule
Perte d'exploitation et jouissance	PEA

Source : élaboré par les auteurs

Chaque garantie est marquée par les modalités oui et non.

- ✓ **Etat :** est une variable qualitative à trois modalités : Bonus, Malus, Nul.
- ✓ **La prime à payer :** Variable quantitative. Elle représente la somme effectivement payée par le souscripteur.

Nous avons éliminé les deux variables pénalité pour âge et pénalité pour permis parce qu'elles sont corrélées avec les deux variables âge de l'assuré et âge du permis successivement.

Les variables caractérisant la sinistralité du véhicule assuré :

Les Variables décrivant la dernière sinistralité sont :

- ✓ **Nature dommage :** variable qualitative binaire à deux modalités : Matériel, Mixte.
- ✓ **Type de lieu :** variable qualitative à deux modalités : Durant le voyage, En stationnement.
- ✓ **Taux de vétusté (**pourcentage) : Ce taux est déduit du prix d'achat d'un bien (dans notre cas le véhicule) afin d'obtenir sa valeur réelle. C'est un facteur que les compagnies d'assurance prennent en compte pour décider du montant d'indemnisation. Le taux de vétusté est calculé différemment selon les assureurs, mais il comprend en général toujours les caractéristiques suivantes : La durée de vie, l'ancienneté, l'entretien. Ses modalités sont : 0%, 5%, 10%, 15%, 20%, 25%.
- ✓ **Catégorie de véhicule :** le tableau suivant présente toutes les catégories de véhicule.

Tableau 2 : Les différentes catégories de véhicule

Code	Catégorie de véhicule
1	Location
2	TPV-long trajet-
3	TPV - urbain -
4	TPV - transport de personnel-
5	Véhicule agricole
6	Véhicule tourisme/Autoécole/ Taxi
7	Véhicule Utilitaire léger inférieur à 3,5
8	Véhicule Utilitaire lourd supérieur à 3,5

Source : élaboré par les auteurs

Dans notre base de données on trouve six catégories, il s'agit de catégories : 1, 2, 4, 6, 7 et 8.

- ✓ **La date de sinistre :** Correspond à la date de survenance du dernier sinistre.

Par ailleurs, nous tenons à préciser que les variables suivantes ont été éliminées car jugées inutiles pour notre étude : Nombre de choc, type de choc, Cause, Dégâts constatés, Montant des fournitures, montant de la peinture, Nombre des jours d'immobilisations, Montants de la main d'œuvre, taux de vétusté, type de lieu, catégorie de véhicule, la date de sinistre.

Et pour Finir, les deux variables à expliquer sont :

- ✓ **Nbr des sinistres-4ans- :** nombre total de sinistres déclarés par l'assuré à la compagnie sur 4 ans, ses modalités : zéro, un, deux, trois et plus.
- ✓ **Montant du règlement :** c'est la charge totale du sinistre déclaré au cours de la période du contrat par l'assuré, c'est-à-dire le coût total mis par l'assuré à la charge de la compagnie pour le règlement de ses sinistres.

4.1.3 Nettoyage et transformation des données

Traitement des valeurs manquantes

Le traitement des valeurs manquantes est l'une des étapes les plus importantes de la phase du prétraitement des données. En ce qui concerne le remplacement statistique des valeurs manquantes, il faut être précautionneux. Le plus simple est de remplacer la valeur manquante par la valeur la plus fréquente ou la moyenne ou la médiane. En revanche, nous pouvons procéder à la suppression des observations touchées par cet événement incertain si ses contributions à l'étude ne paraient pas essentielle. Notre base de données ne présente aucune donnée manquante car lors de la saisie des données nous avons pris le soin de bien contrôler nos données.

Transformation des variables

L'introduction de la variable **« *Etat* »** a été déduite de la variable Bonus et de la variable Malus.

Discrétisation des variables quantitatives

La discrétisation est l'opération qui permet de découper en classes les variables continues, elle est satisfaisante lorsqu'elle permet la création de classes homogènes et distinctes entre elles. Nous avons effectué la discrétisation à l'aide de l'algorithme de Fisher sous **xlstat,** nous avons choisis cet algorithme parce qu'il nous donne une meilleure discrétisation pour nos variables, **les** résultats sont présentés dans le tableau 3 ci-dessous.

Tableau 3 : Codification des variables continues discrétisées

Variables	Classe	Codage
Age	[18 ; 31[=1 [31 ; 40[=2 [40 ; 49[=3 [49 ; 58[=4 [58 ; 69[=5 [69 ; 88] =6	Age1 Age2 Age3 Age4 Age5 Age6
La prime à payer	[101 ; 9051.74 [=1 [9051.74 ; 20444.69 [=2 [20444.69 ; et plus=3	Prime1 Prime2 Prime3
Montant de règlement	[1352.39 ; 27292.65 [=1 [27292.65 ; 70799.6 [=2 [70799.6 ; et plus=3	MR1 MR2 MR3

Source : élaboré par les auteurs

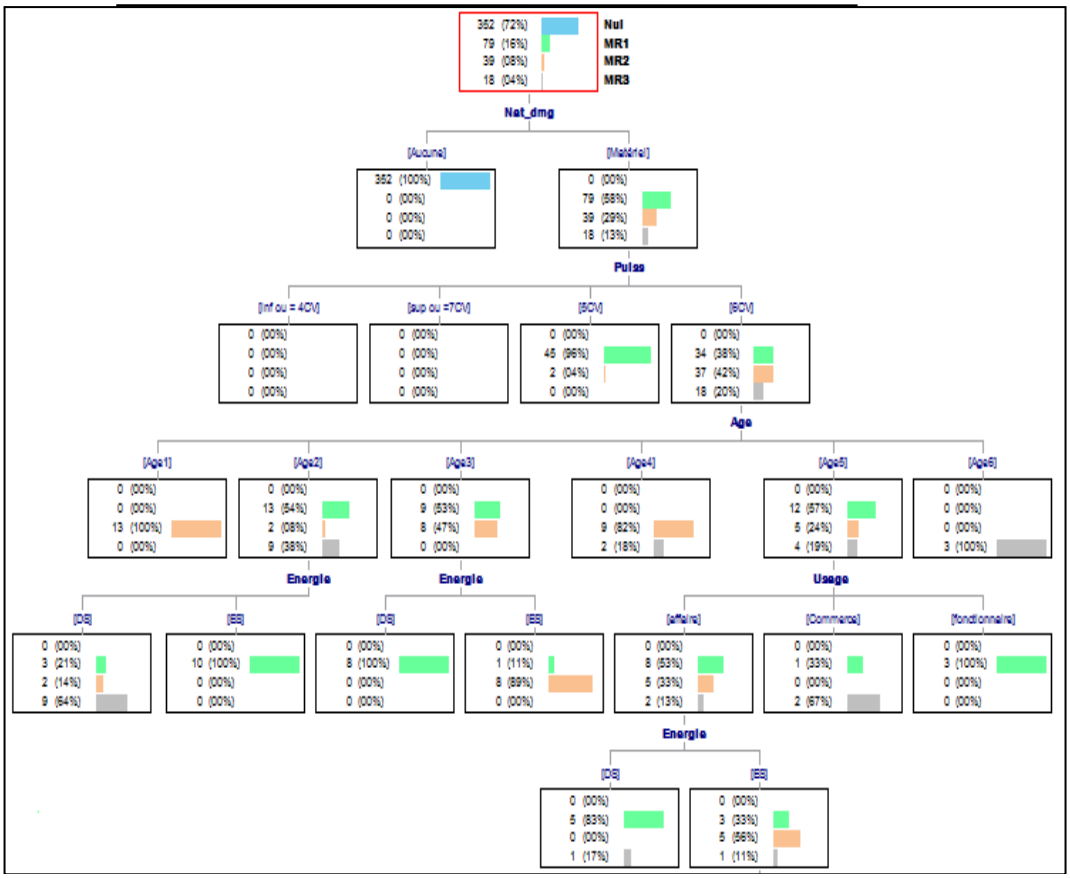
4.1.3 Les arbres de décisions

Les arbres de décision sont une présentation commode de fonction de classification, ils permettent de classer un objet. Pour la construction de l'arbre de décision, nous avons opté pour la méthode d'apprentissage C4.5, basée sur le taux d'erreur comme mesure d'évaluation de partitionnement. Nous appliquons le modèle à deux reprises. D'abord sur la variable dépendante *Coût de sinistre,* puis par rapport à la variable *fréquence de sinistres.*

4.1.3.1 La variable à prédire « *Coût de sinistre* » :

L'application de L'algorithme C4.5 a permis de construire l'arbre de décision illustré en Figure 2.

Figure 2 : L'arbre de décision C4.5 pour la variable à prédire *« coût de sinistre »*



Source : élaboré par les auteurs en utilisant le logiciel SIPINA

Nous avons élagué certains nœuds non pertinents manuellement, ce qui donne les règles suivantes (Chaque chemin de l'arbre, de la racine à une feuille représente une règle) :

Si Nat_dmg = Aucune **Alors** Mont_reglmen = **Nul** à 100% (352 individus)

Si Nat_dmg = Matériel et Puiss = 5CV **Alors** Mont_reglmen = **MR1** à 96% (46 individus)

Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age3 et Energie = DS **Alors** Mont_reglmen = **MR1** à 100% (8 individus)

Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age2 et Energie = ES **Alors** Mont_reglmen = **MR1** à 100% (10 individus)

Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age5 et Usage = fonctionnaire **Alors** Mont_reglmen = **MR1** à 100% (3 individus)

Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age5 et Usage = affaire et Energie = DS **Alors** Mont_reglm = MR**1** à 83% (6 individus)

Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age1 **Alors** Mont_reglmen = **MR2** à 100% (13 individus)

Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age4 **Alors** Mont_reglmen = **MR2** à 82 % (11 individus)
Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age3 et Energie = ES **Alors** Mont_reglmen = **MR2** à 89% (9 individus)
Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age6 **Alors** Mont_reglmen = **MR3** à 100% (3 individus)

- **Matrice de confusion :**

Tableau 4 : Matrice de confusion-1-

	Nul	MR1	MR2	MR3
Nul	352	0	0	0
MR1	0	76	2	1
MR2	0	3	34	2
MR3	0	1	2	15

Cost : 0.0225

Source : élaboré par les auteurs en utilisant le logiciel SIPINA[3]

Le taux d'erreur en test est égal à 2,3%, nous pouvons donc dire qu'en classant un individu pris au hasard dans la population, nous avons 2,3 chances sur 100 de réaliser une mauvaise affectation. Ce taux assez faible, nous pouvons avancer que notre modèle a un bon pouvoir prédictif.

De manière à bien comprendre les résultats découlant de l'arbre, nous expliquons deux niveaux de l'arbre.

Le premier sommet est appelé la « **racine** » de l'arbre. Il est situé sur le premier niveau. Nous y observons la distribution de fréquence de la variable à prédire « ***Coût de sinistre* ».**

Nous constatons sur la racine de l'arbre que 72% des observations ont été annotées « **Nul** », 16% pour la classe « **MR1** », 8% pour la classe « **MR2** » et 4% pour la classe « **MR3** ». La première variable de segmentation choisie par l'algorithme est « **Nature de dommage** »(Nat_dmg) : sur le sommet de droite (2ième niveau), 45 individus ont une nature de dommage Matériel, la proportion de **Montant de Règlement (MR1) = [1352.39 ; 27292.65 [** est de 58%, le nœud de gauche du même niveau avec une nature de dommage = Aucune, concerne la classe **Nul**, avec un pourcentage de 100%, Ce sommet n'ayant plus de sommets enfants ce qui est normal puisqu'il est « pur » du point de vue de la variable à prédire.

D'après l'arbre produit par SIPINA au moyen de la méthode C4.5, les variables les plus significatives au sens de la variable cible le ***Montant de règlement* sont** :

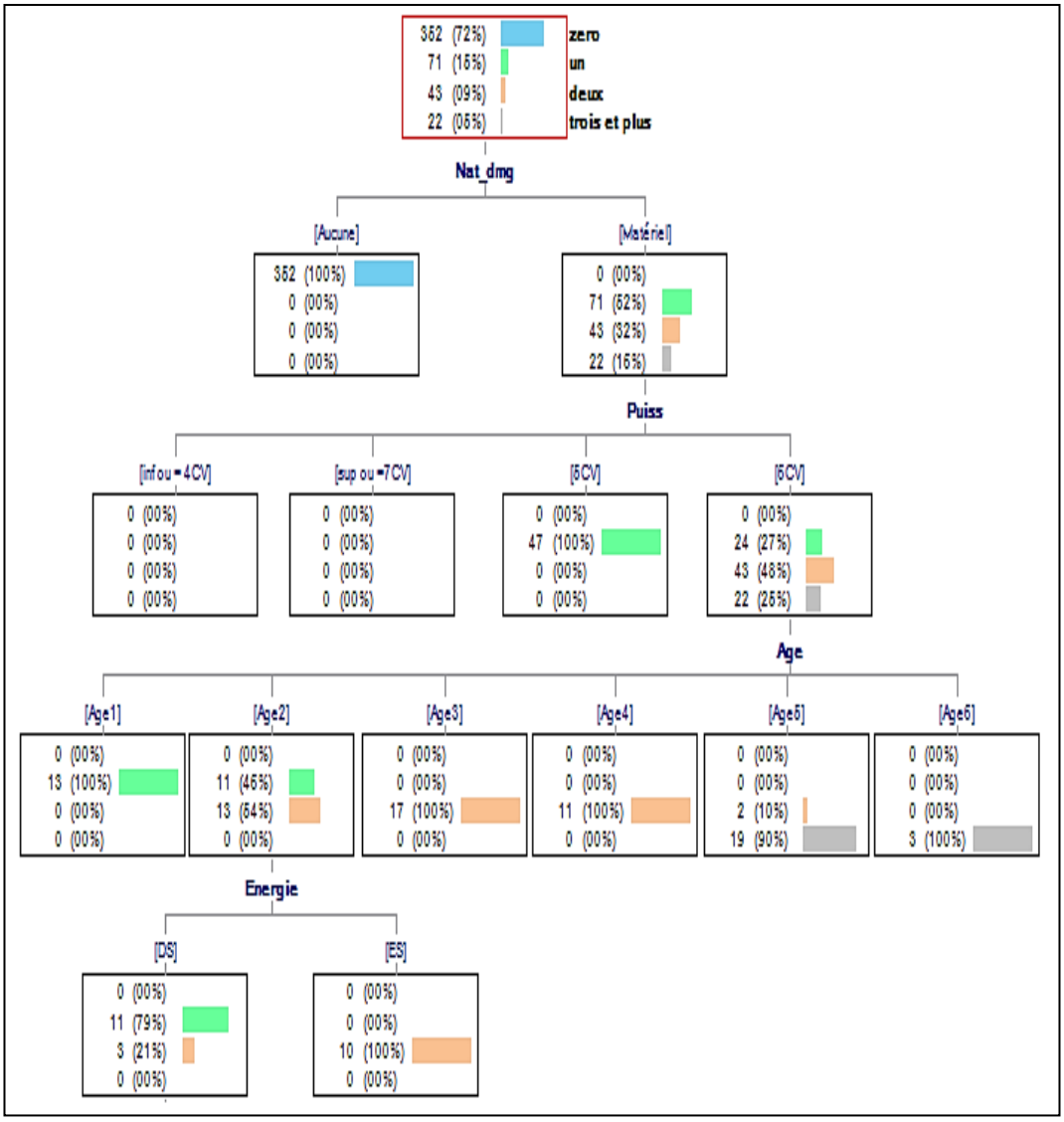
Nature de dommage, Puissance du véhicule, Age du conducteur, Usage et Energie du véhicule.

[3] Système interactif pour l'induction Non-Arborescente, la version que nous utilisons est la version 3.12

4.1.3.2 La variable à prédire « *Fréquence des sinistres* »

L'application de C4.5 sur la variable à expliquer ***Nombre des sinistres a*** permis de construire l'arbre de décision illustré en Figure 3.

Figure 3 : L'arbre de décision C4.5 pour la variable à prédire « *le Nombre des sinistres* »



Source : élaboré par les auteurs en utilisant le logiciel SIPINA

Après avoir élagué certains nœuds non pertinents manuellement, nous pouvons extraire les règles de décision suivantes :

Si Nat_dmg = Aucune **Alors** Nbr_sinistr = **Zéro** à 100% (352 individus)

Si Nat_dmg = Matériel et Puiss = 5CV **Alors** Nbr_sinistr = **Un** à 100% (47 individus)
Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age1 **Alors** Nbr_sinistr = **Un** à 100% (13 individus)
Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age3 **Alors** Nbr_sinistr = **deux** à 100% (17 individus)
Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age4 **Alors** Nbr_sinistr = **deux** à 100% (11 individus)
Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age5 **Alors** Nbr_sinistr = **trois et plus** à 90% (21 individus)
Si Nat_dmg = Matériel et Puiss = 6CV et Age = Age6 **Alors** Nbr_sinistr = **trois et plus** à 100% (3 individus)

- **Matrice de confusion :**

Tableau 5 : Matrice de confusion-2-

	zero	un	deux	trois et plus
zero	352	0	0	0
un	0	71	0	0
deux	0	0	41	2
trois et plus	0	0	0	22

Cost : 0.0041

Source : élaboré par les auteurs en utilisant le logiciel SIPINA

Le taux d'erreur en test est estimé à 0,41%, donc le modèle a classé parfaitement presque toutes les observations de l'échantillon (on a seulement deux observations mal classées).

D'après l'arbre produit par SIPINA au moyen de la méthode C4.5, les variables les plus significatives au sens de la variable cible le ***Nombre des sinistres*** sont :

Nature de dommage, Puissance du véhicule, Age du conducteur et Energie

5. CONCLUSION

Pour mener cette étude nous avons appliqué à notre jeu de données **les arbres de décision** implémentés dans logiciel SIPINA.

D'après le traitement et l'analyse des données, on a pu tirer les conclusions suivantes :

- ✓ Les variables les plus significatives au sens de la variable cible **coût de sinistre** sont :

Nature de dommage, Puissance du véhicule, Age du conducteur, Usage et Energie du véhicule.

- ✓ Les variables les plus significatives au sens de la variable cible **Fréquence de sinistres sont** :

Nature de dommage, Puissance du véhicule, Age du conducteur et Energie.

Ces variables contribuent à expliquer la sinistralité et Permettant à l'assureur d'obtenir une bonne adéquation entre la sinistralité et les primes payées par les assurés, en construisant des classes de risque, de segmenter son portefeuille et de hiérarchiser ces classes à l'aide d'indicateurs de sinistralité.

Vu la situation grave de l'accidentologie en Algérie et l'instar de ce travail, nous recommandons de prendre en compte les variables que nous avons obtenu à fin d'augmenter la valeur de la prime et cela pour inciter les assurés à respecter le code de la route pour diminuer les ravages que font ces accidents.

Comme pour toute recherche, nous avons rencontré un certain nombre de difficultés, depuis la phase documentaire jusqu'à la fin de nos travaux sur tous ce qui concerne la base de données.

5. Liste bibliographique :

1. **Livres :**

   ALAIN TOSETI et d'autres (2000), comptabilité-Réglementation-Actuariat, Ed Economica, Paris ;

   Jean Luc Besson et Christian Patrat (2005), Assurance non vie, Modélisation et simulation, Ed Economica, Paris ;

   JM Rousseau, T. Blayac, N. Oulmane (2001), Introduction à la théorie des assurances, Manuel et Exercices Corrigés, Edition Dunod, 3éme Edition ;

   Lambert Denis-Claire (1996), Economie des assurances, Edition Armando Colin, Paris ;

   Lambert Yvonne (1997), Droit des assurances, Ed Dalloz, 9éme Edition, France ;

   Lefébure (R), Venturi (G) (2001), Data mining : gestion de la relation client, Ed Evrolles.

   Michel Laterasse et d'autres (2002), Les grandes principes de l'assurance, Edition Argus, 5éme Edition, Paris ;

   Stéphane tuffery (2012), Datamining et statistique décisionnelle, l'intelligence des données, Edition TECHNIP 4iémeédition, Paris ;

2. **Thèses :**

   BRAHIMI Blegacem (2011), mémoire magister intitulé par Extraction de connaissances à partie de données incomplètes et imprécises, ENSSEA, Algérie ;

   Frédric Pennerath (2009), une thèse en vue de l'obtention de doctorat en informatique intitulée par Méthodes d'extraction de connaissances à partir de données modélisables par des graphes, Université Henri Poincaré-Nancy 1, France ;

3. **Articles de revue :**

Ricco RAKOTOMALALA (2005), Arbres de décision, MODULAD, France, 25 pages Nm 33 ;

4. **Articles de loi :**
   Ordonnance N° 95-07 du 23 chaabane 1415 correspondant au 25 janvier 1995 relatives aux assurances se ses textes.
5. **Sites Internet :**
   WWW.SAA.dz
   WWW.cnc.dz