ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

La mise en œuvre des algorithmes d'apprentissage automatique en assurance non-vie : une revue de littérature

The Implementation of Machine Learning Algorithms in Non-Life Insurance : A Literature Review

Safaa MOKDAR, *(Doctorante)*
Laboratoire de Modélisation Appliquée à l'Economie et à la Gestion (MAEGE)
Faculté des Sciences Juridiques, Economiques et Sociales Ain Sebaa
Université Hassan II de Casablanca, Maroc

Tarek ZARI, *(professeur de l'enseignement supérieur)*
Laboratoire de Modélisation Appliquée à l'Economie et à la Gestion (MAEGE)
Faculté des Sciences Juridiques, Economiques et Sociales Ain Sebaa Université Hassan II de Casablanca, Maroc

Adresse de correspondance :	Faculté des sciences Juridiques, Economiques et Sociales Ain Sebaa Université Hassan II de Casablanca- Maroc Téléphone : 0522766984 Maroc (Casablanca)
Déclaration de divulgation :	Ce travail a été réalisé avec le soutien du Centre National de la recherche scientifique et technique (CNRST) dans le cadre de la Bourse des Doctorants Moniteurs « PhD-ASsociate Scholarship – PASS ». Ce financement n'impacte par l'objectivité de cette étude. Les auteurs sont responsables de tout plagiat dans cet article.
Conflit d'intérêts :	Les auteurs ne signalent aucun conflit d'intérêts.
Citer cet article	MOKDAR, S., & ZARI, T. (2024). La mise en œuvre des algorithmes d'apprentissage automatique en assurance non-vie : une revue de littérature. *International Journal of Accounting, Finance, Auditing, Management and Economics*, *5*(12), 270-288. https://doi.org/10.5281/zenodo.14285939
Licence	**Cet article est publié en open Access sous licence CC BY-NC-ND**

Received: October 07, 2024 *Accepted: November 25, 2024*

La mise en œuvre des algorithmes d'apprentissage automatique en assurance non-vie : une revue de littérature

Résumé :

Le secteur de l'assurance non-vie est un secteur extrêmement concurrentiel, il est nécessaire pour les compagnies d'assurances de créer une tarification ajustée aux exigences du marché. Généralement, les assureurs utilisent les modèles linéaires généralisés pour la modélisation de la prime pure. Ces modèles traditionnels présentent des limites qui imposent des contraintes sur la structure du risque modélisé et sur les interactions entre les variables explicatives du risque, ce qui peut conduire à une estimation biaisée de la prime d'assurance. Pour surmonter ces limites, les actuaires optent pour des algorithmes efficaces, appelés modèles d'apprentissage automatique qui ont démontré leurs capacités à extraire les structures de dépendances et des particularités entre les données.
La mise en œuvre de ces méthodes en assurance non-vie a connu une croissance significative ces dernières années, en raison de leur capacité à estimer la prime pure d'une manière précise et adaptée à chaque assuré selon son profil de risque. De même, les algorithmes d'apprentissage automatique offrent des avantages en termes de la gestion des risques, la détection de la fraude et l'automatisation de la souscription des contrats.
Dans cet article, nous examinons le développement des algorithmes d'apprentissage automatique en assurance non-vie, en identifiant les limitations des modèles linéaires généralisés ainsi que les divers outils utilisés pour évaluer la performance des modèles adoptés.

Mots clés : la tarification, GLM, prime pure, apprentissage automatique, apprentissage statistique.
JEL Classification : C60, G22
Type du papier : Recherche Théorique

Abstract:

The non-life insurance sector is extremely competitive, and insurance companies need to create pricing that is adjusted to market requirements. Typically, insurers use generalized linear models to model pure premium. These traditional models have limitations that impose constraints on the structure of the modeled risk and on the interactions between risk explanatory variables, which can lead to a biased estimate of the insurance premium. To overcome these limitations, actuaries opt for efficient algorithms, known as machine learning models, which have demonstrated their ability to extract dependency structures and peculiarities between data.
The use of these methods in non-life insurance has grown significantly in recent years, due to their ability to estimate the pure premium in a way that is accurate and tailored to each policyholder's risk profile. Similarly, machine learning algorithms offer advantages in terms of risk management, fraud detection and policy underwriting automation.
In this article, we examine the development of machine learning algorithms in non-life insurance, identifying the limitations of generalized linear models as well as the various tools used to evaluate the performance of the models adopted.

Keywords: underwriting, GLM, pure premium, machine learning, statistical learning.
Classification JEL: C60, G22
Paper type: Theoretical Research

ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

1. Introduction

L'assurance non-vie est une opération par laquelle une compagnie d'assurance s'engage à couvrir les dommages matériels et corporels subis par l'assuré moyennant le paiement d'une prime. Elle couvre des risques variés tels que l'automobile, l'habitation, l'incendie et la responsabilité civile. Le secteur de l'assurance non-vie est considéré comme étant l'un des secteurs les plus dynamiques de l'économie nationale en vue de sa croissance significative et son important chiffre d'affaires. L'intégration des méthodes de Machine Learning dans le domaine de l'assurance non-vie marque une évolution significative ces dernières années, cela a transformé la manière dont les assureurs gèrent leurs risques. Les actuaires utilisent généralement les modèles linéaires généralisés pour estimer la prime pure, ces méthodes permettent de quantifier la relation linéaire entre la variable à expliquer et les variables explicatives. Cette méthode classique est particulièrement populaire au sein de la profession actuarielle puisqu'elle permet de donner des données facilement interprétables, ainsi qu'elle permet de modéliser à la fois les comportements non linéaires et les distributions des résidus non-gaussiens. Elle permet ainsi d'améliorer la qualité de prédiction des modèles (Paglia et al., 2011).

Cependant, ces modèles imposent des contraintes sur la structure des risques et les interactions entre les variables explicatives du risque (P.McCullagh & Nelder, 1989). De même que ces méthodes comportent des hypothèses statistiques lourdes qui ne peuvent pas modéliser les données non structurées (Kévin, 2017). De plus, les GLM sont moins performants face à la croissance des données qui ne sont pas homogènes et volumineux issues de sources telles que les capteurs télématiques, les antécédents des sinistres effectuées par l'assuré et son comportement.

À cet effet, ces différentes limites peuvent entraîner une estimation biaisée de la prime pure, ce qui a poussé les actuaires à chercher des algorithmes efficaces que sont les méthodes d'apprentissage automatique. Ces méthodes offrent de nouvelles perspectives, permettant aux compagnies d'assurance d'améliorer le processus de tarification en analysant de données volumineux et complexes, leur permettant de prendre des décisions plus éclairées en proposant des primes plus personnalisées et adaptés à leurs risques. Elles ont ainsi une capacité à capturer automatiquement les interactions non linéaires afin de construire un modèle plus précis.

De même, pour la gestion des sinistres, ces algorithmes permettent ainsi la détection de la fraude, une souscription automatisée des contrats et l'adjudication des sinistres (Somani, 2021). En effet, les arbres de classification et de régression (CART), les forêts aléatoires, le Gradient Boosting et le Bagging représentent les algorithmes les plus connus qui offrent un nouveau contexte permettant de réaliser une appréhension des risques sous-jacents au sein du portefeuille. La méthode CART est une technique d'apprentissage supervisé développée par Léo Breiman, Friedman, Olshen et Stone en 1984, qui permet de segmenter un ensemble des assurés en groupes homogènes à partir des variables prédictives et en accordant à chaque classe de risque une valeur moyenne (Paglia, 2011). Random Forest est une technique d'apprentissage supervisé appliquée aux problèmes de prédiction, elle est généralement reconnue par sa simplicité, sa précision et sa capacité à traiter les données de grande taille (Zhang, 2023). La méthode Bagging permet d'améliorer les schémas d'estimation ou de classification instables, elle adopte le vote pour la classification et la moyenne pour la régression (Terefe, 2023). Elle est généralement très utile pour les modèles prédictifs qui ne sont pas stables en vue de réduire la variance (Grari, 2015). Pesantez-Narvaez et al. (2019) trouvent que le modèle XG-Boost est caractérisé par sa réputation en raison de sa capacité à fournir des prédictions précises, il est capable de mesurer l'effet des variables explicatives sur la variable à expliquer à chaque itération.

Les algorithmes d'apprentissage automatique ont montré d'après des études effectuées qu'ils

ont une capacité à extraire des structures de dépendances entre les données, ce qui n'existe pas dans les GLM.
Par ce propos, nous cherchons à examiner une revue de littérature récente sur le développement des algorithmes d'apprentissage automatique dans l'assurance non-vie et les divers outils utilisés pour évaluer la performance des modèles adoptés tels que la précision, la sensibilité, la spécificité, rappel, score F1, L'erreur absolue moyenne (MAE), Erreur Quadratique Moyenne (MSE)…
La première section de cet article sera consacrée à la présentation d'une revue conceptuelle, regroupant l'ensemble des définitions clés mentionnées dans les revues de littérature. Une seconde section, intitulée revue théorique, présente les grandes théories tels que l'apprentissage automatique en assurance, les indicateurs de performance et la présentation du GLM et ses limites en assurance. La troisième section est sous forme d'une revue empirique sur la mise en œuvre des méthodes d'apprentissage automatique en assurance non-vie, où des hypothèses sont formulées, en s'appuyant sur des études existantes. La dernière section sera consacrée à la comparaison des deux approches classique GLM et moderne telle que l'apprentissage automatique en assurance non-vie.

2. Revue conceptuelle

2.1. L'assurance non-vie

L'assurance sur le plan technique est une opération par laquelle l'assureur s'engage à couvrir les dommages causés aux assurés en cas de réalisation de risques, moyennant le paiement d'une prime (Carlot, 2018). On distingue deux grandes catégories d'assurance : celles qui couvrent les personnes sont appelées assurances vie et celles qui couvrent les biens et les responsabilités sont appelées assurances non-vie. Ces deux catégories sont complètement différentes en ce qui concerne la structure des risques et leur modélisation.
L'assurance non-vie couvre les dommages matériels et corporels subis par l'assuré ou causés aux tiers, elle permet de protéger le patrimoine de l'assuré contre les pertes matérielles et immatérielles subies par un sinistre. Elle est régie par le principe indemnitaire selon l'article 39 du Code des Assurances : « *L'assurance relative aux biens est un contrat d'indemnité. L'indemnité due par l'assureur à l'assuré ne peut dépasser le montant de la valeur de la chose assurée au moment du sinistre. Il peut être stipulé que l'assuré restera obligatoirement son propre assureur pour une somme ou quotité déterminée ou qu'il supportera une déduction fixée d'avance sur l'indemnité du sinistre* ». Par ce principe indemnitaire, le législateur s'oppose à ce que l'Assurance non-vie puisse devenir une source d'enrichissement pour l'Assuré.

2.2. L'intelligence artificielle

L'intelligence artificielle a été utilisée pour la première fois en 1956 au cours d'un colloque appelé « *Dartmouth Summer Research Project on Artificial Intelligence* » organisé dans l'université Américaine de Dartmouth par plusieurs chercheurs dont John McCarthy, Marvin Minsky, Nathaniel Rochester, Alan Newell, Arthur Samuel, Herbert Simon et Claude Shannon. Ce colloque a été considéré comme le point de départ de l'intelligence artificielle en tant que discipline de recherche. L'intelligence artificielle selon (Minsky, 1967) permet de réaliser des *« programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisantes par des êtres humains car elles demandent des processus mentaux de haut niveau tels que l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique.* ». Roder, 2019 *considère que « l'intelligence artificielle rassemble toutes les techniques permettant à des ordinateurs de simuler et de reproduire l'intelligence humaine. La notion d'intelligence artificielle fait son apparition dans le langage courant ces dernières années, mais on peut considérer qu'elle existe depuis que l'ordinateur fait tourner*

ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

des algorithmes qui ne sont que des reproductions du raisonnement humain. » L'intelligence artificielle englobe divers sous-domaines qui permettent aux machines de faire des prédictions et d'interpréter le langage tels que l'apprentissage automatique et le traitement du langage naturel (Djellaba & Benamara, 2023).
Elle a connu une évolution importante avec le développement de nouvelles techniques du Machine Learning qui ont permis aux systèmes d'intelligence artificielle de s'améliorer dans l'analyse et l'interprétation des données de grandes quantités.

2.3. Apprentissage automatique

2.3.1. Définitions de l'apprentissage automatique

L'apprentissage automatique (Machine Learning en anglais) est une branche de l'intelligence artificielle qui permet de donner aux ordinateurs la capacité d'apprendre à partir des données pour résoudre des problèmes complexes sans être explicitement programmé (Joshi, 2020).
Il est basé sur la mise en œuvre d'algorithmes et de modèles statistiques permettant aux systèmes informatiques grâce à l'exploitation des données et à des processus d'apprentissage itératifs de faire des prédictions (Dong & Quan, 2024).
Les algorithmes d'apprentissage automatique permettent de traiter les problèmes de classification, de régression et de clustering. On distingue dans le tableau ci-dessous les quatre catégories d'apprentissage automatique.

Tableau 1 : Les types d'apprentissage automatique

L'apprentissage supervisé	C'est une méthode d'apprentissage statistique qui se caractérise par la présence de la variable à expliquer Y, le modèle est formé à partir des données étiquetées dont l'objectif est de permettre au modèle d'apprendre la relation entre les entrées X et les sorties Y afin de faire des prédictions sur de nouvelles données non étiquetées.
L'apprentissage non supervisé	Ce type d'apprentissage se caractérise par l'absence de la variable à expliquer Y seules les entrées sont fournies, le problème central permet d'effectuer un partitionnement des données. Il est nécessaire de regrouper les observations en classes homogènes à partir des caractéristiques communes (Laloë, 2010).
L'apprentissage semi-supervisé	Ce type d'apprentissage combine l'apprentissage supervisé et l'apprentissage non supervisé. On dispose d'un modèle formé à partir d'un ensemble des données étiquetées et non étiquetées afin de construire un modèle performant. L'objectif est d'utiliser les données étiquetées pour orienter l'algorithme dans la classification ou la prédiction des données non étiquetées ce qui permet d'améliorer la précision du modèle par rapport à une approche strictement non supervisée.
L'apprentissage par renforcement	Ce type d'apprentissage permet d'apprendre un modèle à partir de l'interaction directe avec un environnement complexe où l'agent doit accomplir ses actions selon une stratégie efficace afin d'obtenir systématiquement une meilleure récompense possible. Il est généralement utilisé dans les jeux et la robotique (Azencott, 2022).

Source : développée par les auteurs

2.3.2. Les algorithmes d'apprentissage supervisé

2.4. La méthode CART

La méthode CART a été introduite en 1984 par Leo Breiman et Richard Olshen pour traiter les problèmes de classification et de régression, cet algorithme est prédictif puisqu'on cherche à expliquer et à prédire une variable endogène par un ensemble de variables exogènes. Elle comprend deux types d'arbres : les arbres de classification et les arbres de régression qui sont considérés comme deux méthodes d'estimation non paramétriques qui ne nécessitent pas d'hypothèse à priori sur la distribution des données, elles se distinguent par la nature de la variable d'intérêt. Nous parlions d'arbre de classification lorsqu'il s'agissait de variables discrètes, ou d'arbre de régression lorsqu'il s'agissait de variables continues (Breiman et al., 1984).

2.5. Random Forest

Les forêts aléatoires sont des méthodes d'apprentissage statistique très populaires, ces méthodes introduites par (Breiman, 2001) sont très répandues pour les tâches de prédiction, elles consistent à agréger un nombre important d'arbres de décision afin de réduire la variance. Leur principe réside dans la construction d'un vaste ensemble d'arbres de décision qui ne sont pas corrélés, suivis de la prise de la moyenne de leurs prédictions. La prédiction de l'algorithme Random Forest repose sur les prédictions de chaque arbre de décision individuel (Madhumita & Gor, 2022).

2.6. Bagging

La méthode Bagging a été introduite par (Breiman, 1996) pour générer plusieurs versions d'un prédicteur afin de les utiliser pour obtenir un modèle de prédiction agrégé en utilisant des échantillons Bootstrap de l'ensemble des données d'apprentissage pour les utiliser comme nouveaux ensembles d'apprentissage.

2.7. Gradient Boosting

L'algorithme Gradient Boosting est une méthode d'apprentissage statistique qui a été introduite dans la littérature de (Schapire, 1990) et développée ultérieurement dans les travaux (Freund, 1995). Le gradient Boosting est un algorithme qui appartient à la famille des algorithmes de Boosting, permet de combiner plusieurs modèles faibles en un ensemble caractérisé par sa performance prédictive capable de fournir des prédictions précises (Ahlgren, 2018). Il réside à noter qu'il peut être généralement appliqué à de nombreux modèles statistiques d'ensemble pour la régression ou la classification (Terefe, 2023), la particularité de cet algorithme est que l'arbre de décision est ajusté de manière séquentielle d'où à chaque itération, un modèle tente de corriger les erreurs commises par le modèle précédent dans le but d'améliorer les prédictions (Tadjuidje, 2023).

3. Revue théorique

3.1. Apprentissage automatique en assurance

L'apprentissage automatique est une discipline qui a connu une évolution importante ces dernières années dans divers domaines, son utilisation en assurance est très récente, il permet aux compagnies d'assurance de réaliser une tarification précise des risques à chaque client (Hopkins, 2023).

Les algorithmes d'apprentissage automatique permettent de diminuer le temps nécessaire pour le traitement des demandes d'indemnisation relatives aux sinistres afin d'améliorer la satisfaction des clients, ils contribuent aussi à la détermination de la prime pure pour chaque assuré selon son risque en identifiant les facteurs qui influencent la sinistralité. Ces algorithmes permettent d'accroître l'efficacité et l'efficience du processus de souscription. Il est utilisé pour la détection des réclamations frauduleuses afin d'obtenir des solutions rapides et de réduire les pertes (Adedotun et al., 2023), ainsi dédié pour la personnalisation de l'offre des produits d'assurance aux assurés en analysant les données selon leur profil et leurs besoins. Rustam et Ariantari 2018 ont utilisé les algorithmes d'apprentissage automatique pour la prédiction de la survenance de sinistres d'assurance automobile en se basant sur les antécédents de la sinistralité. L'apprentissage automatique est largement utilisé par les compagnies d'assurance automobile afin de décider le renouvellement de l'assurance par les assureurs (Wang, 2020).

Selon (Jones & Sah, 2023), les assureurs utilisent les algorithmes de machine Learning pour la prédiction des réclamations, la prévision des primes et des pertes pour les polices d'assurance. Somani (2021) prévoit que l'application de l'apprentissage automatique en assurance offre plusieurs avantages tels que le traitement et l'adjudication des sinistres, la détection de la fraude,

ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

la prévision des sinistres de grande taille, une meilleure tarification et gestion des risques, une segmentation adaptée des assurés et aussi une souscription automatisée des contrats. De plus, les algorithmes d'apprentissage automatique peuvent être utilisés pour la prédiction du tarif de prime et la probabilité de réclamation d'un client en utilisant des données historiques des anciens assurés (Bau & Hanif, 2024).

S'ajoute la comparaison de (Pesantez-Narvaez et al., 2019) par rapport à la performance des méthodes de Machine Learning dans la détection des événements rares de l'assurance responsabilité civile automobile afin de trouver que la régression RiskLogitboost assure une meilleure performance par rapport à d'autres méthodes. Virgilis et al. (2022) estime que les algorithmes d'apprentissage automatique ont une capacité à capturer automatiquement les interactions non linéaires des données afin de construire un modèle plus précis. Ils sont très efficaces en termes de traitement des données de grande taille (Terefe, 2023), ils permettent ainsi de capturer la structure sous-jacente des données afin de construire un meilleur modèle. Ding et al. (2020) ont montré que ces méthodes ont une grande capacité à améliorer les estimations comptables des sinistres par rapport aux estimations traditionnelles qui figurent dans les états financiers.

Au-delà, les méthodes d'apprentissage automatique permettent de prédire le nombre de sinistres qui ne sont pas déclarés de la garantie RC automobile sur la base de données télématiques (Qazvini, 2019), elles permettent aussi d'améliorer la prédiction des sinistres de l'assurance non-vie (Gabrielli et al., 2020).

(Panda et al., 2022) trouvent que l'intervention humaine dans le processus de calcul des coûts de l'assurance santé peut parfois entrainer des résultats incorrects, c'est pour cela qu'ils ont développé un système de prédiction des coûts d'assurance nommé ML Health Insurance Prediction System (MLHIPS), ce système a aidé les assureurs à déterminer facilement et rapidement les valeurs des primes et réduire ainsi les dépenses de santé. Les algorithmes d'apprentissage automatique présentent plusieurs avantages pour la tarification en assurance santé, ils permettent d'analyser de données volumineuses, d'évaluer plus précisément le profil de risque pour chaque assuré et de créer des tarifs personnalisés d'assurance (Bau & Hanif, 2024). Ces algorithmes sont utilisés dans plusieurs taches dans le domaine de l'assurance, y compris l'analyse du taux de fidélisation, la prévision des réclamations, le ciblage des inspections et la détermination des meilleures stratégies de tarification (Behrend & S, 2022). (Dwivedi et al., 2024) présentent que l'intégration des algorithmes d'apprentissage automatique avec la technologie Blockchain a démontré un grand potentiel dans la détection de la fraude en assurance santé, les algorithmes de Machine Learning permettent de distinguer d'une manière très efficace et avec précision les réclamations frauduleuses à partir de données historiques, tandis que la blockchain garantit la sécurité et la transparence. L'article de (Chandravanshi et al., 2024) a examiné la contribution de l'apprentissage automatique à l'amélioration de l'efficacité opérationnelle, la précision et l'identification des demandes d'indemnisation frauduleuses dans le secteur des assurances, ce qui permet de renforcer la confiance des clients et de protéger les intérêts financiers des assureurs.

3.2. Les indicateurs de performance

Il existe différents critères de performances qui permettent de mesurer la performance du modèle adopté tels que la précision, la sensibilité, la spécificité, rappel, score F1, L'erreur absolue moyenne (MAE), Erreur Quadratique Moyenne Racine (RMSE), Erreur Quadratique Moyenne (MSE), Erreur Absolue Moyenne en Pourcentage (MAPE), ROC et le coefficient de détermination (R^2).

On peut citer plusieurs formules mathématiques de mesure de performance telles que :

Tableau 2 : La matrice de confusion

		Prédiction	
		Positive	Négative
ACTUAL	Positive	True positive (TP)	False négative (FN)
	Négative	False positive (FN)	True positive (TN)

Source : développée par les auteurs

Tableau 3 : Formules mathématiques de mesure de performance

Mesure de performance	**Formules mathématiques**
Précision (Précision)	$\frac{TP + TN}{TP + TN + FP + FN}$
Précision (Accuracy)	$\frac{TN}{TN + FP}$
Spécificité (Specificity)	$\frac{TP}{TP + FP}$
Rappel (Recall)	$\frac{TP}{TP + FN}$
MSE	$\frac{\sum(y_i - \hat{y}_i)^2}{n}$
RMSE	$\sqrt{MSE}$
F1-Score	$2 \times \frac{Précision \times Rappel}{Précision + Rappel}$

Source : développée par les auteurs

3.3. Les modèles linéaires généralisées

3.3.1. GLM en assurance

Les modèles linéaires généralisés sont introduits par John Nelder et Robert Wedderburn en 1972 et développés par McCullagh & Nelder en 1989, ils permettent d'étudier et quantifier la relation entre une variable à expliquer Y et des variables explicatives $X = x_1 \ldots . x_n$. Ils sont considérés comme l'outil de référence pour modéliser la prime pure en assurance automobile qui est traditionnellement estimée par les actuaires en se basant sur un modèle collectif "fréquence × coût moyen". Ces méthodes permettent à la fois de modéliser les comportements non-linéaires et des distributions de résidus non-gaussiens (P.McCullagh & Nelder, 1989). Ils se caractérisent par leur facilité d'utilisation dans le processus de tarification en assurance non-vie et plus précisément en assurance automobile d'où ces modèles permettent de fixer une prime pure plus juste basée sur le risque en fonction des critères tarifaires tels que l'âge du conducteur, l'âge du véhicule, la puissance fiscale, le type du carburant et les antécédents de la sinistralité. Ces méthodes sont utilisées aussi dans la prévision des coûts futurs des sinistres en prenant en considération les variables explicatives qui influencent la sinistralité ce qui est important pour conserver la solvabilité financière des compagnies d'assurance (Clark, 2023).
Le GLM est très courant en assurance non-vie pour l'élaboration du tarif d'où la loi de poisson est utilisée pour la modélisation de la fréquence des sinistres et la loi Gamma est utilisée pour la modélisation du coût des sinistres (Bouchta, 2017). Il permet d'étudier la relation entre la variable cible Y et les variables explicatives X afin de fixer une prime pure plus juste basée sur le risque en fonction des critères tarifaires.

3.3.2. Les limites de GLM

Parmi les limites du GLM, nous citons que la capacité de GLM à capter les interactions entre les variables quantitatives n'est pas fiable, il est nécessaire de faire une discrétisation des variables continues pour implémenter le modèle. De plus, le statisticien doit spécifier à priori les interactions entre les variables quantitatives dont il faut tester la significativité (Guillot, 2015). Ainsi ce modèle impose à la variable à expliquer de suivre une certaine loi, cette hypothèse de distribution n'est pas toujours vérifiée par des tests statistiques. II n'est pas adapté

ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

aussi aux problèmes de plus grande taille ou à des jeux de données plus complexes.
De plus, sa capacité a estimé les sinistres rares n'est pas robuste selon les représentations des histogrammes des données réelles et estimées puisque les sinistres rares disposent des valeurs extrêmes qui influencent les paramètres estimés du modèle pour tous les sinistres ce qui affecte la robustesse et la qualité des prédictions du modèle (Fofana, 2017).
Les GLM reposent sur le fait que la relation linéaire entre les prédicteurs et la fonction de lien de la variable à expliquer cela ne permet pas de capturer les interactions non linéaires complexes dans l'assurance (Dissard, 2013).
Il est important de tenir en compte une autre limitation qui réside dans la durée d'exécution de ces modèles qui consomment beaucoup de temps pour l'entraînement et la prédiction, cela peut conduire à ne pas détecter toutes les interactions entre les variables quantitatives ou qualitatives pour conserver le meilleur modèle.
Ces différentes limites entravent la détermination de la bonne tarification ce qui nous poussent à chercher de nouveaux modèles non paramétriques c'est-à-dire les méthodes d'apprentissage automatique qui ont démontré leurs capacités à extraire des structures de dépendances et des particularités entre les données qui restaient jusque-là non détectées par les modèles linéaires généralisés.

4. Revue empirique

Les algorithmes d'apprentissage automatique ont exploré leur efficacité dans divers domaines de l'assurance non-vie. L'objectif de cette revue empirique est de compléter la revue théorique en analysant des travaux antérieurs sur l'utilisation des algorithmes d'apprentissage automatique en assurance non-vie, en formulant des hypothèses de recherche en fonction des études présentées.
H1 : L'utilisation des algorithmes d'apprentissage automatique améliore significativement la précision des prédictions en assurance non-vie et plus précisément l'assurance santé.
- L'étude de (Prova, 2024) a exploré l'utilisation des techniques d'apprentissage automatique pour l'analyse prédictive de l'assurance maladie. Cette étude a examiné l'utilisation de ces méthodes pour l'analyse des données historiques, la prédiction de l'utilisation de l'assurance santé et l'évaluation des risques. Elle vise à améliorer la gestion des risques et l'efficacité globale des systèmes d'assurance santé en utilisant les techniques d'apprentissage automatique. L'auteur a utilisé le coefficient de détermination R^2 pour évaluer la performance de ces méthodes de régression : Arbre de décision, Random Forest, XG-Boost et KNN. Il a démontré que l'arbre de décision a montré son efficacité à capter la variation de la variable à expliquer avec une valeur R^2 de 0.85, suivi par la méthode Random Forest avec une valeur R^2 de 0.83 et une bonne capacité de prédiction. L'étude a révélé que ces deux algorithmes surpassaient les autres algorithmes en termes de précision prédictive sur la base de coefficient de détermination R^2.
- L'article de (Sazzad, 2023) a proposé de faire une prédiction du prix des assurances santé à l'aide des algorithmes d'apprentissage automatique. Il a utilisé dans ce travail les données de santé individuelles et locales pour la prédiction du coût d'assurance pour les différentes catégories de personnes. L'auteur a comparé l'efficacité de neuf algorithmes d'apprentissage automatique que sont : la régression linéaire, la régression Lasso, la régression Ridge, la régression par vecteur de support, l'arbre de décision, Random Forest, le modèle KNN, le Gradient Boosting et la régression par XG-Boost. Ces modèles ont été comparés selon le coefficient de détermination R^2, l'erreur absolue moyenne (MAE) et la racine de l'erreur quadratique moyenne (RMSE) afin d'évaluer la précision des modèles d'apprentissage dans la prédiction des coûts. Donc le meilleur modèle est XG-Boost, suivi par le Gradient Boosting et Random Forest.

- l'étude de (Hu, 2024) a comparé et analysé l'efficacité des modèles d'apprentissage automatique pour la prévision des primes en assurance santé. Il a comparé la performance prédictive des trois modèles que sont la régression linéaire, la forêt aléatoire et l'arbre de décision sur la base de coefficient de détermination R^2. Parmi ces trois modèles, le modèle Random Forest est le modèle le plus performant en termes de prédiction avec un coefficient de détermination de 0.8564, suivie par le modèle régression linéaire avec un coefficient de 0.7584, tandis que l'arbre de décision a obtenu le score le plus faible. L'auteur a trouvé que le modèle Random Forest est très efficace pour la prédiction en assurance santé en raison de sa capacité à traiter les données volumineux et sa capacité à éviter le problème de sur-apprentissage. Il a montré que le modèle de régression linéaire est capable de capturer les relations linéaires.
- Les chercheurs (Billa & Nagpal, 2024) ont exploré l'utilisation des algorithmes d'apprentissage automatique pour la prédiction des primes en assurance santé. Les chercheurs ont effectué une analyse complète de divers algorithmes d'apprentissage automatique tels que les arbres de décision, les forêts aléatoires, Gradient Boosting, Régression Ridge, K-Nearest Neighbours, SVR et le modèle linéaire, et ils ont évalué leurs performances sur un ensemble de données en assurance santé du monde réel par le coefficient de détermination R^2, l'erreur absolue moyenne (MAE) et la racine de l'erreur quadratique moyenne (RMSE). Les résultats ont montré que ces modèles d'apprentissage automatique prédisent avec précision les variations des prix en assurance santé, d'où le Gradient Boosting assure une meilleure prédiction de prime. Les chercheurs ont discuté les implications des résultats pour les assureurs et les assurés, d'où ces techniques peuvent optimiser les techniques de tarification et améliorer l'efficacité et l'efficience de la prestation en assurance santé.
- Les chercheurs (Rishmawi & Abu-Hassan, 2024) ont étudié le rôle de l'apprentissage automatique dans le secteur de l'assurance maladie selon une étude de cas. L'objectif de cette étude est de prédire avec précision le niveau de risque des souscripteurs d'assurance maladie en utilisant les algorithmes d'apprentissage automatique afin d'aider les compagnies d'assurance à déterminer un taux de prime précis pour les nouveaux clients et à minimiser les risques. Les chercheurs ont trouvé que ces algorithmes sont efficaces en termes de souscription et de tarification ce qui aidera l'entreprise d'assurance à obtenir un avantage concurrentiel et à augmenter son bénéfice. L'étude a porté sur 10000 souscripteurs qui ont été classés selon trois niveaux de risque d'où les chercheurs ont comparé l'ensembles des algorithmes de classification : Random Forest, Machine à Vecteurs de Support (SVM), Arbre de décision, Réseau neuronal et la régression logistique multinomiale à l'aide du langage de programmation statistique R. Ils ont trouvé que le modèle Random Forest est le meilleur modèle pour la prédiction du niveau de risque avec une précision de 94%, une sensibilité de 81% pour le niveau de risque moyen et 72% pour le niveau de risque élevé, suivi par SVM et l'arbre de décision avec une précision de 92%. Les chercheurs ont recommandé d'encourager les assureurs à utiliser les algorithmes prédictifs dans leurs processus de tarification et souscription et exécuter les algorithmes d'apprentissage automatique qui nécessitent une grande puissance de calcul afin d'améliorer l'efficacité.

H2 : Les algorithmes d'apprentissage automatique surpassent les modèles linéaires généralisés dans la tarification des risques.

- Les chercheurs (Ben Hamida et al., 2024) ont proposé la tarification de l'assurance maladie individuelle par les algorithmes d'apprentissage automatique. Les chercheurs ont comparé la performance de ces algorithmes non paramétriques tels que l'arbre de décision (CART), Random Forest, le gradient boosting extrême, la régression par vecteurs de support et les réseaux de neurones artificiels par rapport au modèle linéaire généralisé (GLM) en termes de performance prédictive. Le résultat de cette comparaison a montré que les algorithmes d'apprentissage automatique surpassaient le GLM en termes de puissance prédictive. Le modèle XG-Boost offre une meilleure performance dans la modélisation de la fréquence et la gravité

ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

des sinistres sur la base de MSE et RMSE, suivi par l'algorithme Random Forest. Cette étude montre l'importance de l'intégration des méthodes d'apprentissage automatique dans la tarification de l'assurance maladie d'où ces modèles prédictifs permettent d'améliorer la segmentation des clients et offrir une tarification adaptée et personnalisée.

- L'étude de (Bellina, 2014) a exploré l'application des méthodes d'apprentissage automatique dans la tarification non-vie et plus précisément en assurance automobile. L'auteur a comparé la performance des algorithmes d'apprentissage automatique tels que les forets aléatoires, CART, Bagging et le Gradient Boosting par rapport aux modèles classiques tels que les modèles linéaires généralisés. Les résultats ont montré que les algorithmes de Machine Learning sont performants par rapport aux modèles linéaires généralisés dans la tarification non-vie. Dans cette étude, les méthodes d'agrégation sont les meilleures. Les arbres de régression se caractérisent par leurs interprétations faciles dans le processus de tarification. Les résultats ont montré que les algorithmes améliorent la segmentation des clients et offrent une vision synthétique du portefeuille des assurés.
- L'étude de (Paglia, 2011) a présenté une approche innovante de la tarification en assurance non-vie par les méthodes d'apprentissage statistique. L'auteur a comparé les résultats produits par les principaux algorithmes d'apprentissage et le GLM sur la base de l'erreur quadratique moyenne MSE, les résultats ont montré que les algorithmes d'apprentissage statistique sont performants dans la modélisation de la prime pure d'où les arbres de décision améliorent les mesures d'erreur entre le risque mesuré et le risque modélisé et optimisent les résultats en utilisant plusieurs bases de données ce qui rend cet algorithme robuste aux fluctuations de sinistralité. L'étude a proposé ainsi une modification de l'algorithme CART afin de prendre en compte la période d'exposition de l'assuré au risque de risque.
- l'étude de (Ejjami, 2024) a exploré les techniques d'apprentissage automatique dans la tarification de l'assurance responsabilité civile au Maroc. Cette étude vise à évaluer l'efficacité des algorithmes d'apprentissage automatique, notamment les arbres de décision, Random Forest, le gradient boosting et les modèles polynomiaux dans la tarification en assurance responsabilité civile automobile. L'auteur a évalué la qualité de précision à l'aide du coefficient de détermination R^2, l'erreur quadratique moyenne (MSE) et la racine de l'erreur quadratique moyenne (RMSE) sur la base d'entrainement et la base test. Les résultats montrent que les modèles d'apprentissage automatique améliorent la précision des prédictions de primes d'assurance par rapport aux méthodes traditionnelles, et plus précisément la régression par modèle polynomial d'où les valeurs de MSE et RMSE sont réduites sur la base d'apprentissage et la base test. Donc ces méthodes non paramétriques permettent d'améliorer la précision et l'efficacité dans l'évaluation des primes d'assurance. L'auteur a recommandé d'examiner l'utilisation de ces algorithmes dans différents types d'assurance dans des recherches complémentaires, afin de valider les résultats et élargir leur portée.

H3 : Les méthodes d'apprentissage automatique améliorent d'une manière efficace la détection de la fraude en assurance non-vie grâce à leur capacité à identifier les schémas complexes.

- L'étude de (Adedotun et al., 2023) vise à créer un modèle prédictif qui permet de déterminer avec précision les réclamations frauduleuses en assurance automobile. Les chercheurs ont exploré les techniques de Machine Learning pour la détection des réclamations frauduleuses en assurance automobile. Ils ont évalué la performance des algorithmes tels que l'arbre de décision, SVM, Random Forest, XG-Boost, AdaBoost, la régression logistique et KNN dans la détection de la fraude en assurance automobile à travers des métriques d'évaluation des performances, telles que la précision, le rappel et le score F1. Les résultats montrent que ces algorithmes sont efficaces pour la détection des réclamations frauduleuses, permettant aux assureurs d'améliorer les stratégies de gestion des risques.
- L'article de (Bouzgarne et al., 2019) a présenté une comparaison des dix algorithmes d'apprentissage automatique en évaluant leurs performances, afin de déterminer le meilleur

modèle pour la prédiction de la fraude à partir de données réelles. Cette comparaison a montré que l'algorithme Random Forest offre une meilleure performance selon K-measure, RMSE, rappel et précision, il atteint ainsi le score de précision le plus élevé. Cet algorithme fonctionne mieux que les autres algorithmes comparés dans la prédiction de la fraude.

- L'étude de (Cherkaoui et al., 2024) a montré la performance des méthodes d'apprentissage automatique supervisé dans la détection de la fraude en assurance avec la disponibilité des grandes quantités de données et la puissance de calcul. Elle vise à offrir une évaluation précise des algorithmes d'apprentissage automatique dans la détection de la fraude en utilisant des données réelles en assurance santé. Les chercheurs ont comparé la performance des algorithmes d'apprentissage de classification supervisée tels que Random Forest, la régression logistique, les réseaux neuronaux, XG-Boost. Les résultats montrent que l'algorithme XG-Boost obtient des meilleurs résultats, ce qui signifie que cet algorithme contribue à l'amélioration de la détection de la fraude en assurance santé.
- Les chercheurs (Nabraw & Alanazi, 2023) ont développé un modèle qui permet de détecter automatiquement la fraude en assurance santé dans l'Arabie Saoudite. Ils ont appliqué dans cette étude trois techniques d'apprentissage automatique telles que la régression logistique, Random Forest et les réseaux de neurones artificiels pour détecter la fraude, chaque méthode a élaboré des résultats acceptables selon des métriques de performance telles que la précision (accuracy), la précision (precision), la spécificité, le rappel, le score F1 et l'AUC.

Le tableau ci-dessous représente les différentes études effectuées par les auteurs sur la mise en œuvre des algorithmes d'apprentissage automatique dans la branche de l'assurance non-vie en se concentrant sur des mesures de performance spécifiques.

Tableau 4 : Différentes études sur la mise en œuvre des méthodes de Machine Learning en assurance non-vie

Auteurs	Titre	Algorithmes testés	Les outils d'évaluation	Le meilleur modèle
(Rawat et al., 2021)	Application of machine Learning and data visualization techniques for decision support in the insurance sector	Logistic Regression Random Forest Decision Tree Support Vector Machine Gaussian Naïve Bayes Bernoulli Naïve Bayes Mixed Naïve Bayes K-Nearest Neighbors	Accuracy Precision Recall F1 score	Random Forest Decision Tree KNN
(Poufinas et al., 2023)	Machine Learning in Forecasting Motor Insurance Claims	Support Vector Machines Arbres de Décision Random Forest Boosting	Mean Absolute Percentage Error (MAPE)	Random Forest Boosting
(Al-Ataby et al., 2023)	Insurance Risk Prediction Using Machine Learning	Decision Tree Random Forest XGBoost	Accuracy Precision Recall Sensitivity Specificity F1-Score Kappa Statistics AUC	XGBoost
(Abdelhadi et al., 2020)	A proposed model to predict auto insurance claims using machine learning techniques	XGBoost J48 ANN Naïve Bayes	Accuracy ROC	XGBoost J48
(Pesantez-Narvaez et al., 2019)	Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression	Logistic Regression XGBoost (tree booster) XGBoost (linear booster)	Accuracy Sensitivity Specificity	XGBoost
(Zhang, 2023)	Prediction for Insurance Premiums Based on Random Forest and Multiple Linear Regression	Random Forest Multiple Linear Regression	MSE	Random Forest

ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

(Selvakumar et al., 2021)	Predictive Modeling of Insurance Claims Using Machine Learning Approach for Different Types of Motor Vehicles	Régression linéaire Exponential Smoothing ARIMA Hybrid ARIMA- ANN Artificial neural network (ANN)	RMSE MAPE	Artificial neural network (ANN)
(Hanafy & Ming, 2021)	Machine Learning Approaches for Auto Insurance Big Data	La régression logistique XGBoost Random Forest arbres de décision naïve Bayes K-NN	Sensitivity Specificity Precision Recall The F- Measure	Random Forest
(Orji & Ukwandu, 2024)	Machine learning for an explainable cost prediction of medical insurance	Extreme Gradient Boosting (XGBoost) Gradient-boosting Machine (GBM) Random Forest	Le coefficient de détermination (R^2) MAE RMSE MAPE	XGBoost
(Severino & Peng, 2021)	Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata	Logistic Regression Penalized Logistic Regression Naive Bayes KNN Polynomial Kernel SVM Gaussian Kernel SVM Deep Neural Network Random Forest GBM	Accuracy, precision, F1 Score, Kappa MCC Recall	Random Forest Deep Neural Network Gradient Boosting
(Bau & Hanif, 2024)	Comparative Analysis of Machine Learning Algorithms for Health Insurance Pricing	Random Forest XGBoost Multiple Linear Regression Ridge regression	MAE MSE RMSE R^2	Random Forest

Source : développée par les auteurs

5. Comparaison GLM & Méthodes d'apprentissage automatique en assurance

Les méthodes d'apprentissage automatique ont démontré leurs capacités à extraire des structures de dépendances et des particularités entre les données qui restaient jusque-là non détectées par les GLM. Nous allons commencer par la méthode CART qui est considérée comme une méthode d'estimation non paramétrique qui ne nécessite pas d'hypothèse sur la distribution des données, de même, cette méthode permet de modéliser les multiples interactions entre les variables explicatives contrairement aux modèles linéaires généralisés (Bellina, 2014). L'utilisation de cet algorithme est simple en raison du faible nombre de paramètres à optimiser, en particulier par rapport à d'autres méthodes (Rudelle, 2018).

A la lumière de ce qui précède, la méthode CART est l'une des meilleures méthodes qui permettent de segmenter le portefeuille des assurés en classes de risque homogènes à travers des variables tarifaires, elle a une capacité à traiter à la fois des données catégorielles et numériques, ce qui rend cet algorithme adapté à la modélisation prédictive (Saed & Jaharadak, 2022). Cependant, nous pouvons déduire que la méthode CART est un outil performant qui permet d'améliorer la segmentation d'un portefeuille des assurés afin de classer les assurés en classes de risques homogènes et ensuite accorder à chaque classe de risque la valeur moyenne selon le coût et la fréquence de sinistre (Paglia, 2011). CART est l'une des méthodes les plus populaires en apprentissage statistique en raison de sa rapidité d'exécution et sa capacité à sélectionner automatiquement les variables discriminantes (Grari, 2015).

Bellina (2014) approuve que cette méthode permet d'offrir une vision synthétique de la base de données et leurs interprétations aisées peuvent enrichir la réflexion autour de la stratégie tarifaire. Elle est aussi considérée comme robuste dans le traitement des valeurs (Murphy,

2012). D'après une étude effectuée par (Dissard, 2013) qui a comparé la performance de GLM par rapport à CART en tarification pour l'assurance automobile, il a trouvé que l'algorithme CART offre des meilleurs résultats en termes de prédiction. Il réalise aussi un traitement très efficace aux valeurs manquantes par la méthode "Surrogate Split" et elle se caractérise ainsi par sa facilité d'interprétation par rapport aux GLM.
Les travaux de (Paglia, 2011) ont déjà exploré l'utilisation des méthodes de l'apprentissage statistique en tarification non-vie d'où il a déterminé la fréquence de sinistres graves à l'aide d'un arbre de décision et il a comparé sa performance par rapport à GLM et il a trouvé que l'algorithme CART est très performant.
En termes de pouvoir prédictif, le modèle CART obtient de meilleurs résultats sur la base de MSE et MAE que ceux obtenus avec les modèles linéaires généralisés puisque le pouvoir prédictif de GLM est limité (Clement, 2022).
L'un des défis pratiques lors de la prédiction réside dans la présence des valeurs manquantes, CART propose un traitement efficace pour bien gérer les valeurs manquantes en prédiction (Genuer & Poggi, 2017). Par contre, la méthode CART est instable ce qui entraîne d'importantes variations dans les estimations et les prévisions selon (Bellina, 2014), une autre limitation réside dans son manque de robustesse.
Pour pallier ce problème, la technique Bagging vise à réduire la variance des prédictions en réalisant un rééchantillonnage des données d'entraînement afin d'obtenir une prédiction globale plus robuste (Sutton, 2005). Le bagging semble particulièrement performant pour les procédures à forte variance et faible biais comme les arbres (Friedman et al., 2008). Elle est généralement très utile pour les modèles prédictifs qui ne sont pas stables en vue de réduire la variance (Grari, 2015). Par contre le Bagging se caractérise par sa difficulté d'interprétation par rapport à GLM et il est opérationnellement difficile à implémenter et très couteux en calcul (Chatelain, 2024)). De plus, cette méthode met en œuvre la construction d'un ensemble d'arbres différents, ce qui permet d'augmenter le temps de calcul de cet algorithme (Bouchta, 2017).
D'autre part, la méthode Random Forest permet ainsi d'améliorer la robustesse, elle est très populaire, très performant en termes de prédiction et il se caractérise par peu de paramètres hypercomplexes (Hanafy & Ming, 2021).
D'après une étude effectuée par (Fofana, 2017), l'algorithme Random Forest présente l'erreur quadratique moyenne le plus faible par rapport à GLM, cela explique que cet algorithme d'apprentissage automatique dispose d'une meilleure performance prédictive.
Le GLM repose sur le principe de l'absence d'interactions entre les variables. En d'autres termes, l'algorithme XG-Boost permet de révéler plusieurs effets hétérogènes, il permet de modéliser les interactions complexes entre les variables explicatives à travers des outils tels que les statistiques H, et les graphiques PDP qui permettent d'examiner la combinaison de deux variables (Delcaillau, 2019).
(Bakacha & Mehdaoui, 2021) estiment que l'algorithme Random Forest présente des meilleurs résultats dans la modélisation du coût et du fréquence des sinistres, Il permet ainsi d'éviter le problème de sur-apprentissage du modèle (Fofana, 2017). D'après (Franquet, 2018), Le GLM est moins performant par rapport aux méthodes d'apprentissage statistique utilisées, particulièrement la forêt aléatoire et le Gradient Boosting, alors que ces deux algorithmes sont difficiles à interpréter que le GLM et l'arbre de décision.

6. Conclusion

Le secteur Marocain d'assurance non-vie est un secteur très concurrentiel d'où la nécessité d'élaborer un tarif adapté pour chaque assuré selon son risque c'est-à-dire que la prime pure doit couvrir le risque garanti par l'assureur.
Cet article compare la méthode classique GLM qui est largement utilisée par les actuaires avec les méthodes d'apprentissage automatique telles que CART, Random Forest, Bagging et XG-

ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

BOOST d'où on a trouvé que les GLM présentent des limitations. Parmi ces limites, on trouve que le GLM ne permet pas de capter les interactions complexes entre la variable à expliquer et les variables explicatives en assurance non-vie.
En plus, ces modèles ne sont pas adaptés pour modéliser les problèmes de grande taille, ils consomment beaucoup de temps pour la prédiction ce qui ne permet pas de détecter toutes les interactions entre les variables quantitatives ou qualitatives pour conserver le meilleur modèle. Les méthodes d'apprentissage automatique sont devenues ces dernières années très populaires en raison de leur capacité à améliorer la segmentation des risques et à fournir des prédictions plus précises pour la modélisation de la prime pure. Elles permettent ainsi la détection des réclamations frauduleuses, le traitement des sinistres, la prévision des sinistres de grande taille, une souscription automatisée des contrats et une meilleure tarification et gestion des risques.
L'algorithme CART est facile à interpréter, il offre des bons résultats de prédiction par rapport à GLM selon le critère MSE dans la tarification en assurance non-vie d'après une étude effectuée par (Paglia et al., 2011).
La méthode Random Forest présente moins de risques de sur-apprentissage que la méthode CART, elle permet aussi d'améliorer le manque de robustesse. La méthode XG-Boost est largement utilisée en raison de sa réputation pour fournir des prédictions précises, elle est capable d'estimer les coefficients dans certaines situations, ce qui permet d'évaluer l'ampleur des effets (Pesantez-Narvaez et al., 2019). Finalement, l'algorithme Bagging est particulièrement performant pour les procédures à forte variance et faible biais comme les arbres (Friedman et al., 2008).
Pour conclure, nous avons mis en évidence les limites des modèles linéaires généralisés qui présupposent une forme spécifique des risques et de leurs interactions, c'est pourquoi on a utilisé les méthodes de machine Learning qui ont démontré leur performance par rapport à d'autres méthodes.

Références

(1). Abdelhadi, S., Elbahnasy, K., & Abdelsalam, M. (2020). A proposed model to predict auto insurance claims using Machine Learning techniques. *Journal of Theoretical and Applied Information Technology*, 1-10.
(2). Adedotun, A. F., Olumide, A., Odusanya, O., Odusanya, O., & Okagbue. (2023). Prediction of automobile insurance fraud claims using machine learning. *The Scientific Temper*.
(3). Ahlgren, M. (2018). Claims Reserving using Gradient Boosting and Generalized Linear Models. *KTH ROYAL INSTITUTE OF TECHNOLOGY SCHOOL OF ENGINEERING SCIENCES*, 2-58.
(4). Al-Ataby, A., Jayabalan, M. A., & Hamid, A. A. (2023). Insurance Risk Prediction Using Machine Learning. *ResearchGate*, 1-16.
(5). Amal, B. H., Manel, K., Christian, D. P., & Lotfi, B. (2024). Machine learning based methods for ratemaking health care insurance. *International Journal of Market Research*.
(6). Azencott, C.-A. (2022). *Introduction au Machine Learning.* Paris: Dunod 2ème édition.
(7). BAKACHA, A., & MEHDAOUI, M. (2021). *Méthodes d'apprentissage automatique appliquées à la tarification automobile à usage touristique (A1).* INSEA.
(8). Bau, Y.-T., & Hanif, S. A. (2024). Comparative Analysis of Machine Learning Algorithms for Health Insurance Pricing. *INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION*, 481-491.

(9). Bellina, R. (2014). *Méthodes d'apprentissage appliquées à la tarification non-vie.* Lyon.
(10). Billa, M. M., & Nagpal, T. (2024). Medical Insurance Price Prediction Using Machine Learning. *J.Electrical Systems 20-7s*, 2270-2279.
(11). Bouchta, G. (2017). *Mise en œuvre de méthodes innovantes de tarification.* Paris: l'Université Paris Dauphine.
(12). Bouzgarne, I., Youssfi, M., Qbadou, M., & Bouattane, O. (2019). Performance comparative study of machine learning algorithms for automobile insurance fraud detection. *Third International Conference on Intelligent Computing in Data Sciences (ICDS)*.
(13). Breiman, L. (1996). *Bagging Predictors. Machine Learning.* Boston. Manufactured in The Netherlands.: Kluwer Academic Publishers.
(14). Breiman, L. (2001). *Random Forest Machine Learning.* California.
(15). Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1984). *Classification and Regression Trees.* United States of America: Chapman and Hall/ CRC.
(16). Carlot, J.-F. (2018). *Droit des assurances.* Dalloz, 7 e édition.
(17). Chandravanshi, S., Kotdiya, G., Kotdiya, Y., & Patidar, Y. (2024). Study the Insurance Claims Detection Using Machine Learning. *International Journal of Engineering Applied Science and Management*, 1-6.
(18). Chatelain, P. (2024, Mai 15). *Tarification à l'adresse en assurance habitation individuelle.* LYON : L'UNIVERSITÉ CLAUDE BERNARD.
(19). Cherkaoui, O., Anoun, H., & Maizate, A. (2024). A benchmark of health insurance fraud detection using machine learning techniques. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 1925~1934.
(20). Clark, D. R. (2023). Glm for Dummies (and actuaries). *CAS E-Forum Summer*.
(21). Clement, M. (2022). *Utilisation d'arbres de régression pour la prédiction de couts automobiles.* Paris: l'Université de Paris-Dauphine.
(22). Dangeti, P. (2017). *Statistics for machine learning: build supervised, unsupervised, and reinforcement learning models using both Python and R.* Packt Publishing.
(23). Delcaillau, D. (2019). *Contrôle et Transparence des modèles complexes en actuariat.* l'Institut des Actuaires.
(24). Ding, K., Lev, B., Peng, X., Sun, T., & Vasarhelyi, M. A. (2020). Machine learning improves accounting estimates: evidence from insurance payments . *Review of Accounting Studies* , 1098–1134.
(25). Dissard, G. (2013). *Evolutions tarifaires des produits d'assurance automobile et IRD.* Strasbourg: l'Institut des Actuaires.
(26). Djellaba, A., & Benamara, M. (2023). Intégration de l'IA dans l'audit : l'effet sur le processus d'audit. 1-14.
(27). Dong, P., & Quan, Z. (2024). Automated Machine Learning in Insurance. 2-46.
(28). Dwivedi, V., Kundu, D., & Sharma, V. (2024). Integrating Blockchain with Machine Learning for Fraud Detection in Health Insurance Claims Management. *International Journal of INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING*, 128–139.
(29). Ejjami, R. (2024). Machine Learning Approaches for Insurance Pricing: A Case Study of Public Liability Coverage in Morocco. *International Journal for Multidisciplinary Research (IJFMR)*, 1-23.
(30). F, A. A., Olumide, A., Odusanya, O., J.A, A., & Okagbue. (2023). Prediction of automobile insurance fraud claims using machine learning. *The Scientific Temper Vol. 14 (3)*, 756-762.

ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

(31). Fofana, C. H. (2017). *Approche tarifaire des contrats collectifs Frais de Santé à l'aide des méthodes d'apprentissage.* Lyon: l'Institut des Actuaires.
(32). Franquet, S. (2018). *Modélisation de la fréquence des sinistres graves en assurance automobile : apports et interprétabilité des méthodes d'apprentissage statistique.* Brest: Institut des Actuaires.
(33). Freund, Y. (1995). Boosting a Weak Learning Algorithm By Majority. Information and Computation. *Information and Computation*, 256–285.
(34). Friedman, J., Hastie, T., & Tibshirani, R. (2008). *The Elements of Statistical Learning Data Minning, Inference and Prediction. Stanford, California: Springer.* California: Springer.
(35). Gabrielli, A., Richman, R., & Wüthrich, M. V. (2020). Neural network embedding of the over-dispersed Poisson reserving model. *Scandinavian Actuarial Journal* , 1-29.
(36). Genuer, R., & Poggi, J.-M. (2017). Arbres CART et Forêts aléatoires, Importance et sélection de variables. *Hal Open Science*, 1-46.
(37). Grari, V. (2015). *Impact des données exogènes sur la tarification en santé.* Lyon : l'Institut des Actuaires.
(38). Guillot, A. (2015). *Apprentissage statistique en tarification non-vie : quel avantage opérationnel ?* Paris: l'Institut des Actuaires.
(39). Hamida, A. B., Kacem, M., Peretti, C. D., & Belkacem, L. (2024). Machine learning based methods for ratemaking health care insurance. *International Journal of Market Research*, 1-22.
(40). Hanafy, M., & Ming, R. (2021). Machine Learning Approaches for Auto Insurance Big Data. *Risks*, 1-23.
(41). Hopkins, L. (2023, Aout 7). *The Role of Machine Learning in Insurance Pricing.* Récupéré sur EFBPublic: https://www.efbpublic.org/insurance-pricing
(42). Hu, Y. (2024). Comparison and Analysis of the Effectiveness of Linear Regression, Decision Tree, and Random Forest Models for Health Insurance Premium Forecasting. *Proceedings of the 3rd International Conference on Business and Policy Studies*, 1-7.
(43). Jones, K. I., & Sah, S. (2023). The Implementation of Machine Learning In The Insurance Industry With Big Data Analytics. *International Journal of Data Informatics and Intelligent computing (IJDIIC),*, 21-38.
(44). Joshi, A. V. (2020). *Machine Learning and Artificial Intelligence.* USA: Springer International Publishing.
(45). Kévin, A. K. (2017). *TARIFICATION AUTOMOBILE : GLM vs RÉSEAUX DE NEURONES.* Lyon: l'Institut des Actuaires.
(46). KOUACH, Y., ATTAR, A. E., & HACHLOUFI, M. E. (2022). Auto insurance fraud detection using unsupervised learning. *Alternatives Managériales Economiques*, 96-113.
(47). Laloë, T. (2010). Sur quelques problèmes d'apprentissage supervisé et non supervisé. *Hal Open science*.
(48). Madhumita, G., & Gor, R. (2022). Health insurance premium prediction using Blockchain technology and Random Forest regression algorithms. *International Journal of Engineering Science Technologies,*, 74-82.
(49). Minsky, M. L. (1967). *Computation: Finite and Infinite Machines.* London: Prentice Hall.
(50). Murphy, K. P. (2012). *Murphy, K. P. (2012). Machine Learning : A Probabilistic Perspective.* London, England: The MIT Press.
(51). Nabraw, E., & Alanazi, A. (2023). Fraud Detection in Healthcare Insurance Claims Using Machine Learning. *Risks* , 1-11.

(52). Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications*, 1-20.
(53). P.McCullagh, & Nelder, J. A. (1989). *Generalized Linear Models.* New York: Chapman and Hall ,2ème édition.
(54). Paglia, A. (2011). *Tarification des risques en assurance non-vie, une approche par modèle d'apprentissage statistique.* Cedex: Institut d'actuariat.
(55). Paglia, A., Phélippé-Guinvarc'H, M., & Lenca, P. (2011). *Adaptation de l'algorithme CART pour la tarification des risques en assurance non-vie.* Post-print, HAL.
(56). Panda, S., Purkayastha, B., Das, D., Chakraborty, M., & Biswas, S. K. (2022). Health Insurance Cost Prediction Using Regression Models . *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*.
(57). Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression. *Risks*, 1-16.
(58). Poufinas, T., Gogas, P., Papadimitriou, T., & Zaganidis, E. (2023). Machine Learning in Forecasting Motor Insurance Claims. . *Risks*, 1-19.
(59). Prova, N. N. (2024). A Study of Machine Learning Techniques for Predictive Analysis of Health Insurance.
(60). Qazvini, M. (2019). On the validation of claims with excess zeros in liability insurance: A comparative study . *Department of Actuarial Mathematics and Statistics, School of Mathematical and Computer Sciences, Risks*.
(61). Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights*, 1-15.
(62). Rishmawi, R., & Abu-Hassan, H. (2024). The Role of Machine Learning in Health Insurance Industry-a case study.
(63). Roder, S. (2019). *Guide pratique de l'intelligence artificielle dans l'entreprise : Anticiper les transformations, mettre en place des solutions.* Paris: Eyrolles.
(64). Rudelle, D. F. (2018). *Apport des méthodes d'apprentissage statistique pour le provisionnement individuel en assurance non-vie.* Paris: Institut des actuaires.
(65). Rustam, Z., & Ariantari., N. P. (2018). Support Vector Machines for classifying policyholders satisfactorily in automobile insurance. *Journal of Physics: Conference Series, Volume 1028, 2nd International Conference on Statistics, Mathematics,*.
(66). Saed, A. A., & Jaharadak, A. A. (2022). Implementation with performance evaluation of decision tree classifier for uncertain data: Literature review. *International Journal of Multidisciplinary Research and Publications*, 1-8.
(67). Sazzad, H. (2023). Medical Insurance Cost Prediction Using Machine Learning.
(68). Schapire, R. E. (1990). *The strength of weak learnability. Machine Learning, 197–227.* Boston Manufactured in The Netherlands. : Kluwer Academic Publishers.
(69). Selvakumar, V., Kumar Satpathi, D., Praveen Kumar, P., & Haragopal, V. (2021). Predictive Modeling of Insurance Claims Using Machine Learning Approach for Different Types of Motor Vehicles. *Universal Journal of Accounting and Finance*, 1-14.
(70). Severino, M. K., & Peng, Y. (2021). Severino, M. K., & Peng, Y. Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, 1-14.
(71). Somani, S. (2021, Septembre 8). *17 Disruptive AI and Machine Learning Use Cases in Insurance World-AI and ML in Insurance Industry.* Récupéré sur birlasoft: https://www.birlasoft.com/articles/17-ai-and-ml-use-cases-insurance
(72). Sutton, C. D. (2005). Classification and Regression Trees, Bagging, and Boosting . *ScienceDirect*, 303-329.

ISSN: 2658-8455
Volume 5, Issue 12 (2024), pp. 270-288.

IJAFAME
International Journal of Accounting, Finance, Auditing, Management & Economics

(73). Tadjuidje, J. P. (2023). *Un modèle multi-catégoriel avec gradient boosting et extrêmes pour projeter à l'ultime les paiements en assurance automobile.* Québec: Université Laval.

(74). Terefe, E. M. (2023). Tree-Based Machine Learning Methods For Vehicle Insurance Claims Size Predicti. 1-23.

(75). Wang, H. D. (2020). Research on the Features of Car Insurance Data Based on Machine Learning. *3rd International Conference on Mechatronics and Intelligent Robotics (ICMIR-2019)*, 582–587.

(76). Zhang, T. (2023). Prediction for Insurance Premiums Based on Random Forest and Multiple Linear Regression . *BCP Business & Management* , 2315-2321.