Approche hybride pour la prédiction des coûts de réparation automobile : intégration du raisonnement d'ontologie avec des modèles de régression

Hamid Ahaggach[*,**], Lylia Abrouk[*,***], Eric Lebon[**]

[*]Laboratoire d'informatique de Bourgogne
Université de Bourgogne Franche-Comté, France
prénom.nom@u-bourgogne.fr
[**]Syartec, Aix en Provence, France
elebon@syartec.com
[***]MISTEA, Université de Montpellier
INRAE & Institut Agro, France

Résumé. L'estimation des coûts de réparation des dommages automobiles est essentielle pour les assureurs et les ateliers. Les méthodes traditionnelles, souvent manuelles et lentes, peuvent entraîner des erreurs. Cet article, publié dans le journal *Intelligent Systems with Applications* (Ahaggach et al., 2024a), présente une approche combinant des modèles de régression et une ontologie des dommages automobiles (OCD). Peuplée d'informations extraites de données non structurées grâce à des techniques de reconnaissance d'entités nommées et d'extraction de relations, et enrichie par des règles SWRL, cette ontologie permet de générer de nouvelles variables influençant les coûts. La méthode hybride développée, testée sur 300 000 enregistrements, améliore la précision des prédictions.

1 Introduction

La gestion des données non structurées est devenue essentielle pour faciliter une prise de décision éclairée. Dans l'industrie automobile, ce besoin est essentiel lorsqu'il s'agit d'évaluer avec précision les dommages aux véhicules et de prédire les coûts de réparation en utilisant des rapports de dommages. Ces rapports, souvent non structurés, nécessitent une saisie manuelle des données, un processus chronophage et sujet aux erreurs. De plus, les agents ne disposent pas des moyens nécessaires pour estimer avec précision les coûts de réparation en temps réel. Pour répondre à ces défis, nous présentons dans cet article une approche intégrant des règles *Semantic Web Rule Language* (SWRL) pour générer de nouvelles caractéristiques et optimiser les modèles de régression utilisés dans la prédiction des coûts de réparation automobile. En combinant ces règles avec des modèles de régression, notre méthodologie améliore la précision des prédictions, proposant une approche plus efficace et fiable pour l'évaluation des dommages à partir des rapports non structurés. Cet article est organisé comme suit : nous commençons par une présentation de l'état de l'art dans la section 2. La section 3 détaille la méthodologie adoptée, en soulignant les étapes clés de notre démarche. Dans la section 4, nous présentons

l'expérimentation et la comparaison des modèles de régression, avec et sans intégration d'ontologies, afin d'évaluer leur efficacité. Nous concluons avec la section 5, où nous résumons nos principales contributions et les différentes perspectives.

2 État de l'art

Ces dernières années, l'utilisation des algorithmes d'apprentissage automatique, notamment les modèles de régression, a considérablement progressé dans l'estimation des coûts. Ces techniques peuvent analyser de vastes ensembles de données pour identifier des motifs et des tendances. Cependant, ces modèles se heurtent souvent à des défis importants. Les schémas de dommages ne sont pas toujours évidents, présentant parfois une complexité qui rend difficile la prédiction précise à partir des seules caractéristiques disponibles. En complément de ces techniques, l'ontologie offre un potentiel significatif pour la prédiction. Cet outil puissant permet de modéliser les connaissances du domaine et de capturer les relations sémantiques entre diverses caractéristiques. Le raisonnement à l'aide de règles SWRL dans l'ontologie peut considérablement améliorer les capacités prédictives d'un système.

Approche basée sur les modèles de régression : Les modèles de régression sont utilisés pour prédire les valeurs futures de la variable dépendante, identifier les prédicteurs importants et comprendre les relations entre les variables. Cependant, dans certains cas, la relation entre les variables prédictrices et la variable prédite peut être plus complexe et non linéaire, et dans ces cas, même les modèles de régression les plus avancés rencontrent des difficultés. Ces modèles incluent la régression à vecteurs de support, et la régression par réseaux de neurones (Smola et Schölkopf, 2004). Les modèles de régression ont été appliqués à un large éventail de domaines, tels que la finance (Cook et al., 2008), la construction (Jung et al., 2020) et la santé (Stone et al., 2022). Dans l'industrie automobile, l'analyse de régression a été appliquée pour résoudre différents problèmes, tels que l'estimation des valeurs de revente des véhicules (Lessmann et Voß, 2017; Gegic et al., 2019), la prévision du temps de vente d'une voiture (Ahaggach et al., 2023), et la prédiction de la maintenance (Chen et al., 2019).

Approche basée sur l'ontologie : L'ontologie est une représentation formelle des concepts et des relations au sein d'un domaine spécifique. Elle fournit un moyen structuré et standardisé de représenter les connaissances, ce qui facilite leur partage et leur réutilisation dans différentes applications et systèmes. L'ontologie peut être utilisée pour soutenir le développement de modèles prédictifs en fournissant un vocabulaire commun pour décrire les variables et les relations impliquées. L'utilisation des ontologies dans les tâches de prédiction a été explorée dans divers domaines. Dans le domaine de la construction, les ontologies ont été utilisées pour automatiser le processus d'estimation des coûts de construction. Niknam (2015) a proposé une approche basée sur la sémantique pour l'estimation des coûts de construction. Lee et al. (2014) ont proposé une approche pour l'estimation des coûts de construction qui utilise les données de modélisation des informations du bâtiment et le raisonnement d'ontologie. Dans le secteur de la santé, Thirugnanam et al. (2013) ont développé une ontologie pour offrir des informations précises et pertinentes sur les maladies humaines et leurs symptômes. Ils ont implémenté des règles SWRL pour la prédiction des maladies. Dans une autre étude Chandra et al. (2023), ont

développé une ontologie pour les maladies et des règles SWRL ont été intégrées à des fins de diagnostic et de classification.

Approche hybride : De nombreux chercheurs ont exploré l'intégration de l'ontologie et des techniques d'apprentissage automatique pour diverses tâches de prédiction. Cao et al. (2019) ont employé une combinaison de techniques de regroupement flou et d'ontologie pour classifier les défaillances de produits. Tang et al. (2018) ont introduit un système de détection de fraude dans les états financiers qui utilise une ontologie et un algorithme d'arbre de décision pour la détection de fraude. Le système combine les règles de l'arbre de décision pour acquérir des règles SWRL et les utiliser pour permettre au moteur d'inférence de tirer parti des connaissances existantes et d'explorer de nouvelles connaissances. Dans le même contexte, Jabardi et Hadi (2021) utilisent l'apprentissage automatique et l'ontologie pour apprendre des règles sémantiques dans la classification des fraudes. Le modèle proposé utilise une ontologie pour représenter des connaissances spécifiques et des arbres de décision comme méthode d'apprentissage de règles axée sur les données. Ils discutent également des règles SWRL, qui sont utilisées pour extraire des connaissances cachées à partir des ontologies.

Dans le secteur de la santé, El Massari et al. (2022) ont comparé les approches basées sur les règles SWRL de l'ontologie avec les approches basées les algorithmes d'apprentissage automatique pour prédire les maladies cardiovasculaires. Ils concluent que les ontologies donnent de meilleurs résultats par rapport aux algorithmes de l'arbre de décision. Cette conclusion semble contradictoire, puisque les deux méthodes reposent sur les mêmes règles.

Discussion : Les modèles de régression sont très utilisés pour les tâches de prédiction. Cependant, malgré leur potentiel, il y a des défis à relever, en particulier lorsqu'il s'agit de données de faible qualité, d'informations manquantes ou incomplètes, et de motifs complexes qui compliquent la tâche d'extraction de la relation entre les caractéristiques et la variable dépendante. Les ontologies existantes trouvent des applications dans divers domaines. Cependant, il convient de noter que dans le domaine automobile, aucune ontologie spécifique n'a été définie pour l'évaluation des dommages aux voitures. De plus, il y a un manque de travaux sur la prédiction et l'évaluation des dommages aux voitures dans ce contexte.
L'intégration des techniques d'ontologie et d'apprentissage automatique a été largement étudiée. Il existe deux approches principales observées dans la littérature lors de l'intégration des techniques d'ontologie et d'apprentissage automatique. D'une part, certaines études utilisent les résultats de l'apprentissage automatique pour construire des règles, visant à améliorer la précision des prédictions. Ces règles sont dérivées des motifs et des relations identifiés par les modèles d'apprentissage automatique et sont conçues pour capturer des informations précieuses, améliorant ainsi les performances globales des tâches de prédiction. D'autre part, dans d'autres approches, les ontologies, sont utilisées pour valider les prédictions des modèles d'apprentissage automatique et fournir des explications sémantiquement riches.

Notre approche combine l'ontologie avec des modèles de régression. En enrichissant l'entrée du modèle avec des caractéristiques supplémentaires dérivées de l'ontologie. Le raisonnement d'ontologie est considéré comme une couche qui contribue à améliorer les résultats de prédiction. La section suivante fournit une description détaillée de notre approche.

3 Méthodologie

La Figure 1 illustre la méthodologie adoptée pour estimer les coûts de réparation des dommages automobiles. Elle se compose de trois étapes principales : l'extraction d'informations pour le peuplement d'ontologies, l'intégration de règles SWRL, et la prédiction des coûts.

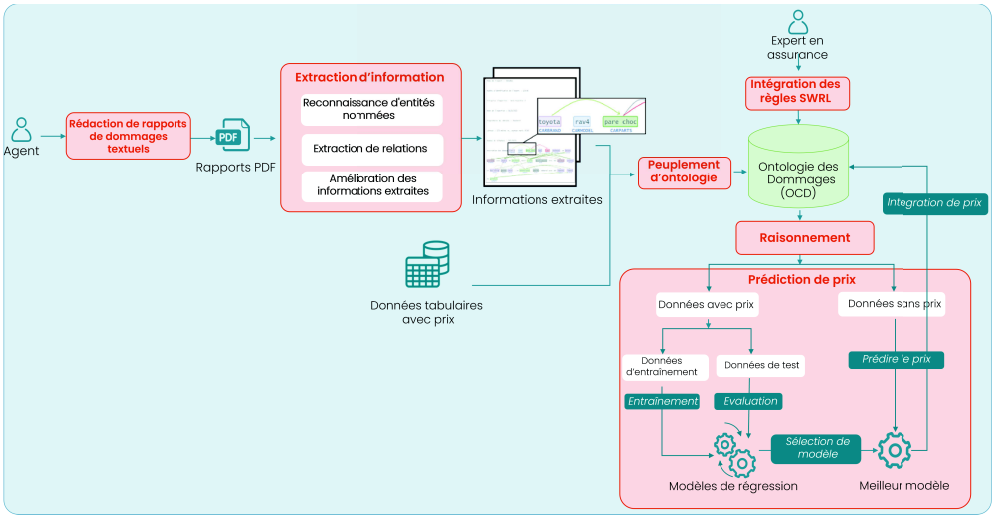

FIG. 1 – *Méthodologie pour la prédiction de prix de réparation des dommages.*

Extraction d'informations et peuplement d'ontologie : Au cours de cette première étape, des informations sont extraites de rapports de dommages non structurés à l'aide des techniques NER et de RE. Ces informations sont ensuite intégrées dans l'ontologie OCD. De plus, des données tabulaires récupérées auprès d'ateliers de réparation sont également utilisées pour peupler l'ontologie. Cette étape a été évaluée dans ce travail (Ahaggach et al., 2024b).

Intégration de règles SWRL : Dans cette étape, des règles SWRL sont intégrées dans l'ontologie pour permettre un raisonnement enrichissant les informations recueillies, en identifiant des liens logiques et en inférant de nouvelles connaissances non explicitement mentionnées dans les rapports.

Prédiction des coûts de réparation : Enfin, les informations sont exploitées pour prédire les coûts de réparation des différents types de dommages. Cette prédiction est effectuée à l'aide de modèles de régression, qui analysent les motifs et tendances au sein des données pour fournir des estimations précises des coûts de réparation.

3.1 Intégration des règles SWRL dans l'ontologie

L'objectif de cette approche est de réduire la dépendance aux experts humains en intégrant des informations pertinentes non présentes dans les rapports, telles que la décision de remplacer

ou de réparer une pièce de voiture endommagée selon des critères spécifiques établis par des experts.
Exemple : Si la pièce est endommagée en raison d'une perte, d'une bosse ou d'une fracture, et qu'il s'agit d'une roue de véhicule, alors la pièce doit être remplacée.

$$\texttt{CarParts}(?\textit{part}) \wedge \texttt{PartName}(\texttt{?part}, \texttt{"Wheel"}) \wedge \texttt{IsDamaged}(?\textit{part}, \texttt{True}) \wedge \texttt{Damage}(\texttt{?damage}) \wedge \texttt{hasDamage}(\texttt{?part, ?damage}) \wedge (\texttt{DamageType}(\texttt{?damage, "Dent"}) \vee \texttt{DamageType}(\texttt{?damage, "Breakage"})) \Rightarrow \texttt{RepairAction}(\texttt{?damage, "Replace"})$$

3.2 Prédiction des coûts de réparation

La prédiction des coûts de réparation, après le raisonnement d'ontologie, nous permet de prétraiter à la fois les données de rapport non structurées et les données tabulaires. Nous exécutons des requêtes SPARQL pour extraire les données contenant des prix des ateliers (Tableau 1), utilisées pour l'entraînement et l'évaluation de nos modèles de régression. En outre, nous extrayons des données sans prix afin de prédire les coûts en utilisant le meilleur modèle de régression. Dans les sections suivantes, nous détaillons chaque étape de ce processus pour estimer les coûts de réparation.

TAB. 1 – *Données tabulaires contenant des cas de réparation avec des informations sur les prix.*

Réf. Rapport	Marque Voiture	Modèle Voiture	Pièce Voiture	···	Dommage	Gravité	Action	Coût
vehicle_1171348_rep	Honda	Civic	porte	···	rayure	Mineure	Réparer	53 €
vehicle_1171348_rep	Honda	Civic	Pare-chocs	···	Éraflure	Mineure	Réparer	100 €
···	···	···	···	···	···	···	···	···
vehicle_1181441_rep	BMW	Série 3	Pare-brise	···	Éclat	Mineur	Réparer	75 €

3.2.1 Analyse et prétraitement des données

Le prétraitement des données est important pour assurer la qualité et l'adéquation du jeu de données. Les étapes principales incluent :

— Gestion des données et remplacement des valeurs manquantes (ex : *Severity* par *Minor*).
— Nettoyage des données et correction des anomalies et des incohérences (ex : suppression du symbole € dans la colonne *Coût*).
— Encodage des données et conversion des variables catégorielles en format numérique via encodage *one-hot* et par étiquettes.
— Normalisation et mise à l'échelle *Min-Max* pour standardiser les caractéristiques.
— Sélection des caractéristiques en utilisant l'Élimination Récursive des Caractéristiques pour identifier les attributs les plus pertinents influençant le coût de réparation.

3.2.2 Modèles de régression

Cette section présente divers modèles de régression utilisés pour estimer le coût de réparation des dommages automobiles. L'objectif est d'apprendre une fonction f qui associe les caractéristiques d'entrée X au coût de réparation Y. **Régression linéaire multiple :** Modélise

la relation linéaire entre plusieurs variables indépendantes et le coût de réparation. **Arbres de décision :** Modèle en forme d'arbre où chaque nœud représente une décision basée sur une caractéristique. **Forêts aléatoires :** Ensemble d'arbres de décision combinés pour améliorer la précision des prédictions. **boosting de gradient :** Combine itérativement des arbres de décision pour minimiser l'erreur de prédiction. **Régresseur XGB :** Implémentation avancée du boosting de gradient, optimisée pour les performances et l'efficacité. **Machines à vecteurs de support (SVR) :** Trouve un hyperplan optimal dans un espace de caractéristiques de dimension supérieure. **Perceptron multicouche :** Réseau de neurones qui permet une transformation non linéaire des caractéristiques d'entrée.

4 Expérimentation et résultats

Pour nos expérimentations, nous utilisons un jeu de données qui comprend 300 000 cas de réparation sous format tabulaire. Le jeu de données est divisé en plusieurs ensembles pour l'apprentissage et l'évaluation des modèles : 70% des données pour l'entraînement, 10% à l'ensemble de validation pour permettre des ajustements fins des hyperparamètres, et 20% à l'ensemble de test pour évaluer les performances des modèles. Pour garantir des performances optimales des modèles de régression, il est nécessaire de choisir des hyperparamètres appropriés en utilisant la technique de recherche en grille (*Grid Search*).

Le tableau 2 présente une comparaison complète de divers modèles de régression, évalués à la fois avec et sans l'utilisation d'ontologie. Les modèles sont évalués à l'aide de quatre métriques : *MSE* (Erreur Quadratique Moyenne), MAE (Erreur Absolue Moyenne), *RMSE* (Racine de l'Erreur Quadratique Moyenne) et R^2 (Coefficient de détermination). Dans le scénario sans ontologie, le modèle *MLPRegressor* montre la meilleure performance avec un score R^2 de 91%, indiquant un ajustement acceptable aux données. Le modèle *SVR* montre la performance la plus faible avec un score R^2 de 61%. Cela suggère que, sans ontologie, *MLPRegressor* est un choix plus fiable pour ce jeu de données, car il capture mieux les motifs sous-jacents des données. Cette capacité est due à la présence de nombreuses transformations de caractéristiques au sein de ses couches, lui permettant de faire des prédictions plus précises.

TAB. 2 – *Comparaison des modèles de régression avec et sans intégration de l'ontologie.*

Modèle	**Sans Ontologie**				**Avec Ontologie**			
	MSE	**MAE**	**RMSE**	**R2**	**MSE**	**MAE**	**RMSE**	**R2**
LinearRegression	973.208	22.720	31.196	0.712	918.140	22.035	30.300	0.727
DecisionTreeRegressor	769.082	8.600	27.732	0.772	138.071	2.133	11.750	0.959
RandomForestRegressor	692.443	8.517	26.314	0.795	**114.126**	**2.043**	**10.683**	**0.966**
GradientBoostingRegressor	723.964	18.313	26.906	0.786	367.824	13.482	19.178	0.890
XGBRegressor	936.508	21.616	30.602	0.723	662.352	18.088	25.736	0.803
SVR	1330.991	21.927	36.482	0.607	1214.532	20.198	34.850	0.640
MLPRegressor	285.921	10.334	16.909	0.915	206.529	7.482	14.371	0.938

L'intégration de l'ontologie dans les modèles de régression a entraîné des améliorations significatives de la précision des prédictions pour tous les modèles. Les modèles *RandomForestRegressor* et *DecisionTreeRegressor*, ayant les meilleures performances, fournissant les estimations de coût les plus précises pour la réparation des dégâts de voiture suite à l'inté-

gration de l'ontologie. La performance du modèle *RandomForestRegressor* s'est considérablement améliorée lorsque l'ontologie a été intégrée, atteignant une erreur quadratique moyenne faible de 114 et une erreur absolue moyenne faible de 2, indiquant qu'en moyenne, les coûts de réparation prédits s'écartent des coûts réels de seulement 2 euros. Il a surpassé tous les autres modèles, y compris ceux sans ontologie. La valeur élevée de R^2 de 97% indique que les estimations des coûts prédits sont étroitement alignées avec les valeurs réelles. La capacité du modèle à capturer des relations complexes et des motifs au sein des données en fait un candidat solide pour l'estimation précise des coûts. Bien que les modèles *MLPRegressor* et *GradientBoostingRegressor* aient affiché des valeurs de *MSE* relativement plus élevées par rapport aux modèles *RandomForestRegressor* et *DecisionTreeRegressor*, ils ont tout de même fourni des estimations de coûts fiables. Le modèle *MLPRegressor* a atteint un *MSE* de 206 et une valeur R^2 de 94%, indiquant une forte capacité prédictive. Le modèle *GradientBoostingRegressor* a atteint un *MSE* de 368 et une valeur R^2 de 89%.

5 Conclusion

Nous avons présenté dans cet article, nos travaux publiés dans le journal *international Intelligent Systems with Applications (SJR Q1)* sur la prédiction de coûts de réparation basée sur les ontologies. L'intégration de l'ontologie dans les modèles de régression améliore significativement leur précision de prédiction en offrant une compréhension sémantique plus riche des données. Cependant, cette intégration augmente également le temps de prédiction. Les modèles comme le *RandomForestRegressor* et le *DecisionTreeRegressor* excellent avec des données enrichies en raison de leur capacité à gérer la complexité et la non-linéarité, ce qui les rend particulièrement adaptés aux tâches telles que l'estimation des coûts de réparation des dommages automobiles. La variation des performances entre les modèles souligne l'importance de sélectionner le modèle approprié en fonction des caractéristiques spécifiques des données et de la tâche à accomplir.

Références

Ahaggach, H., L. Abrouk, S. Foufou, et E. Lebon (2023). Predicting car sale time with data analytics and machine learning. In *Product Lifecycle Management. PLM in Transition Times : The Place of Humans and Transformative Technologies : 19th IFIP WG 5.1 International Conference, PLM 2022, Grenoble, France, July 10–13, 2022, Revised Selected Papers*, pp. 399–409. Springer.

Ahaggach, H., L. Abrouk, et E. Lebon (2024a). Enhancing car damage repair cost prediction : Integrating ontology reasoning with regression models. *Intelligent Systems with Applications*, 200411.

Ahaggach, H., L. Abrouk, et E. Lebon (2024b). Information extraction from automotive reports for ontology population. *Applied Ontology* (19), 1–30.

Cao, Q., A. Samet, C. Zanni-Merk, F. d. B. de Beuvron, et C. Reich (2019). An ontology-based approach for failure classification in predictive maintenance using fuzzy c-means and swrl rules. *Procedia Computer Science 159*, 630–639.

Chandra, R., S. Tiwari, S. Agarwal, et N. Singh (2023). Semantic rule web-based diagnosis and treatment of vector-borne diseases using swrl rules. *arXiv preprint arXiv :2301.03013*.

Chen, C., Y. Liu, X. Sun, C. Di Cairano-Gilfedder, et S. Titmus (2019). Automobile maintenance prediction using deep learning with gis data. *Procedia CIRP 81*, 447–452.

Cook, D. O., R. Kieschnick, et B. D. McCullough (2008). Regression analysis of proportions in finance with self selection. *Journal of empirical finance 15*(5), 860–867.

El Massari, H., N. Gherabi, S. Mhammedi, H. Ghandi, M. Bahaj, et M. R. Naqvi (2022). The impact of ontology on the prediction of cardiovascular disease compared to machine learning algorithms. *International Journal of Online & Biomedical Engineering 18*(11).

Gegic, E., B. Isakovic, D. Keco, Z. Masetic, et J. Kevric (2019). Car price prediction using machine learning techniques. *TEM Journal 8*(1), 113.

Jabardi, M. H. et A. S. Hadi (2021). Using machine learning to inductively learn semantic rules. In *Journal of Physics : Conference Series*, Volume 1804, pp. 012099. IOP Publishing.

Jung, S., J.-H. Pyeon, H.-S. Lee, M. Park, I. Yoon, et J. Rho (2020). Construction cost estimation using a case-based reasoning hybrid genetic algorithm based on local search method. *Sustainability 12*(19), 7920.

Lee, S.-K., K.-R. Kim, et J.-H. Yu (2014). Bim and ontology-based approach for building cost estimation. *Automation in construction 41*, 96–105.

Lessmann, S. et S. Voß (2017). Car resale price forecasting : The impact of regression method, private information, and heterogeneity on forecast accuracy. *International Journal of Forecasting 33*(4), 864–877.

Niknam, M. (2015). *A semantics-based approach to construction cost estimating*. Ph. D. thesis, Marquette University.

Smola, A. J. et B. Schölkopf (2004). A tutorial on support vector regression. *Statistics and computing 14*, 199–222.

Stone, K., R. Zwiggelaar, P. Jones, et N. Mac Parthaláin (2022). A systematic review of the prediction of hospital length of stay : Towards a unified framework. *PLOS Digital Health 1*(4), e0000017.

Tang, X.-B., G.-C. Liu, J. Yang, et W. Wei (2018). Knowledge-based financial statement fraud detection system : based on an ontology and a decision tree. *Knowledge Organization 45*(3), 205–219.

Thirugnanam, M., T. Thirugnanam, et R. Mangayarkarasi (2013). An ontology-based system for predicting disease using swrl rules. *IJCSBI 7*(1).