REPUBLIQUE DE COTE D'IVOIRE

Université Virtuelle de Côte d'Ivoire

Union – Discipline - Travail

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Niveau : master I

Spécialité : Big Data Analytics

Année universitaire : 2024-2025

IMPACT DES METHODES D'IMPUTATION SUR LES PERFORMANCES DE MODELES DE REGRESSION AVANCES

Membres du groupe ayant réalisé le projet :

COULIBALY Nahoua Daouda

KONATE Aboudoulaye

SOGODOGO Fangniere

SINON Bakary

KESSE Epiphanie

TOURE Djakaria

TOUALY Djolohonon Deborah

KANGAH Kouadio Franck Huberson

YAO Kouadio David Bienvenu

KOUASSI Kouame Yannick

Enseignant :

Dr. AYIKPA Kacoutchy Jean

TABLES DES MATIERES

I. INTRODUCTION

Dans le cadre de l'analyse de données réelles, la gestion des valeurs manquantes représente un défi majeur susceptible d'influencer significativement la qualité des modèles prédictifs.

L'objectif de cet exercice est d'évaluer l'impact des différentes méthodes d'imputation sur les performances de modèles de régression avancés. À partir d'un jeu de données comportant des valeurs manquantes, plusieurs techniques d'imputation seront appliquées — telles que l'imputation par la moyenne, la médiane, le mode, le KNN Imputer, la régression, ou encore la suppression des lignes incomplètes — afin de générer des jeux de données nettoyés. Ces jeux seront ensuite exploités pour entraîner et comparer divers modèles de régression non abordés en cours, comme le SVR, le GradientBoostingRegressor, le KNeighborsRegressor et le DecisionTreeRegressor.

L'évaluation s'appuiera sur des métriques classiques (MAE, RMSE, R^2) ainsi que sur l'analyse de la distribution des résidus et la visualisation des prédictions.

le travail, que nous allons réaliser, vise ainsi à déterminer quelles méthodes d'imputation se révèlent les plus robustes selon le modèle utilisé, et à développer une réflexion critique sur le choix des stratégies de traitement des données manquantes dans le contexte de la modélisation prédictive

II. DESCRIPTION DES METHODES D'IMPUTATION

II.1 Moyenne

L'imputation par la moyenne est une méthode simple et couramment utilisée pour traiter les données manquantes. Elle consiste à remplacer chaque valeur manquante par la moyenne des valeurs observées de la même variable, calculée soit sur l'ensemble des données, soit au sein de sous-groupes homogènes (appelés classes d'imputation) définis par des caractéristiques similaires (par exemple, même sexe, même tranche d'âge, même profession).

a. Principe

On calcule la moyenne des valeurs disponibles pour une variable donnée. Pour chaque donnée manquante, on remplace la valeur manquante par cette moyenne. Si la population est segmentée en classes, la moyenne est calculée au sein de chaque classe, ce qui permet une imputation plus précise et adaptée au contexte de l'observation

b. Avantage

Méthode simple à mettre en œuvre. Permet de conserver une certaine cohérence lorsque la moyenne est calculée au sein de groupes homogènes. Respecte parfois les relations linéaires entre variables dans certains cas

c. Inconvénients

Réduction artificielle de la variance : l'imputation par la moyenne crée un pic artificiel à la valeur moyenne, ce qui déforme la distribution des données et réduit la variance estimée. Perte des relations entre variables : elle peut détruire les corrélations naturelles entre variables, car les valeurs imputées sont toutes identiques au sein d'un groupe. Sensibilité aux valeurs atypiques : la moyenne est influencée par les valeurs extrêmes, ce qui peut biaiser l'imputation
Ne tient pas compte de la variabilité intrinsèque des données manquantes.

II.2 Médiane

a. Principe

L'imputation par la médiane consiste à remplacer chaque valeur manquante d'une variable par la médiane des valeurs observées pour cette même variable. La médiane est la valeur qui sépare la moitié supérieure de la moitié inférieure des données, ce qui la rend moins sensible aux valeurs extrêmes que la moyenne. Cette opération peut être réalisée sur l'ensemble des données ou au sein de groupes homogènes (classes d'imputation) pour une meilleure pertinence.

b. Avantage

Robustesse aux valeurs atypiques : La médiane n'est pas influencée par les valeurs extrêmes, ce qui la rend idéale pour les distributions asymétriques ou contenant des outliers. Simplicité de mise en œuvre : Méthode rapide et facile à appliquer, ne nécessitant

pas de modèles statistiques complexes. Adaptée aux variables ordinales ou asymétriques : Plus pertinente que la moyenne lorsque la distribution n'est pas symétrique ou normale. Réduction du biais dû aux outliers : Contrairement à la moyenne, la médiane ne sera pas tirée vers les valeurs extrêmes, limitant ainsi le risque de biais dans l'imputation.

c. Inconvénients

Réduction de la variance : Comme pour l'imputation par la moyenne, cette méthode diminue la variabilité des données, car toutes les valeurs imputées sont identiques, ce qui peut fausser certaines analyses statistiques. Perte des relations entre variables : L'imputation par la médiane ne préserve pas nécessairement les corrélations naturelles ou les contraintes structurelles entre variables. Moins efficace pour les distributions symétriques : Si la variable est normalement distribuée sans valeur.

II.3 MODE

a. Principe

Elle consiste à remplacer les valeurs manquantes par la valeur la plus fréquente (le mode) observée pour la même variable dans l'échantillon. Cette imputation peut être réalisée sur l'ensemble des données ou au sein de classes homogènes pour plus de précision.

b. Avantage

Réduction de la variabilité : comme pour l'imputation par la moyenne ou la médiane, le fait de remplacer toutes les valeurs manquantes par la même modalité diminue artificiellement la diversité des données.

Dilution des relations entre variables : cette méthode peut affaiblir ou déformer les corrélations naturelles entre variables, ce qui nuit à la qualité des analyses multivariées.

c. Inconvénients

Biais possible : si la modalité la plus fréquente n'est pas représentative des cas manquants, l'imputation peut introduire un biais.

Peu adaptée aux données avec plusieurs modalités proches en fréquence : dans ce cas, le choix du mode peut être arbitraire et peu informatif.

II.4 KNN IMPUTER (KNN IMPUTER)

a. Principe

L'algorithme KNN identifie, pour chaque observation avec des données manquantes, les k observations les plus proches (voisins) sur la base des variables complètes. La valeur manquante est ensuite imputée à partir des valeurs correspondantes de ces

voisins, souvent par une moyenne pondérée ou un vote majoritaire selon que la variable est quantitative ou qualitative. Le choix de k (nombre de voisins) est un paramètre clé.

b. Avantage

Prise en compte des relations multivariées : contrairement à l'imputation par moyenne ou médiane, KNN utilise l'information des autres variables pour estimer la valeur manquante, ce qui préserve mieux les corrélations naturelles.

Adapté aux variables quantitatives et qualitatives : la méthode peut être adaptée selon le type de variable à imputer.

Performance rapide : selon certaines études, l'imputation par KNN peut être rapide à exécuter, notamment pour un k modéré (ex. k=4).

Flexibilité : ne fait pas d'hypothèse forte sur la distribution des données.

c. Inconvénients

Sensibilité au choix de k : un k trop petit peut rendre l'imputation instable, un k trop grand peut diluer la spécificité locale.

Coût computationnel : la recherche des k plus proches voisins peut être coûteuse en temps et en mémoire pour de très grands jeux de données.

Dépendance à la mesure de distance : la qualité de l'imputation dépend fortement de la métrique utilisée pour définir la proximité entre observations.

Problème si trop de données manquantes : lorsque les données manquantes sont trop nombreuses ou réparties sur toutes les observations, la méthode peut perdre en efficacité car il devient difficile de trouver des voisins fiables.

Peut introduire un biais si les données manquantes ne sont pas aléatoires : comme toute méthode d'imputation, KNN suppose que les données manquantes dépendent des autres variables observées (MAR).

II.5 REGRESSION

a. Principe

On construit un modèle de régression (linéaire ou autre) pour la variable contenant des données manquantes, en utilisant les observations complètes où cette variable est connue.

Ce modèle estime la relation entre la variable à imputer et les autres variables explicatives.

Pour chaque observation avec une valeur manquante, on prédit la valeur manquante à partir du modèle de régression.

La valeur imputée est donc la valeur ajustée (prédite) par ce modèle.

b. Avantage

Exploitation des relations entre variables : la méthode utilise l'information disponible dans d'autres variables pour fournir une estimation plus précise que des méthodes univariées (moyenne, médiane).

Réduction du biais : particulièrement efficace lorsque les données manquantes sont MAR (manquantes à condition sur d'autres variables observées).

Flexibilité : on peut utiliser différents types de modèles (régression linéaire, régression locale, arbres, forêts aléatoires) selon la nature des données.

Amélioration de la qualité des analyses : en préservant mieux la structure multivariée, elle permet d'obtenir des résultats d'analyse plus fiables.

c. Inconvénients

Absence de terme d'erreur dans l'imputation simple : la valeur imputée est la prédiction exacte du modèle, sans composante aléatoire, ce qui conduit à une sous-estimation de la variance et une sur-identification des relations entre variables.

Risque de biais si le modèle est mal spécifié : si la relation entre variables est mal modélisée, l'imputation peut être incorrecte.

Complexité accrue : nécessite la construction et la validation d'un modèle de régression, ce qui est plus complexe que les imputations simples.

Ne capture pas l'incertitude de l'imputation : contrairement à l'imputation multiple, cette méthode ne reflète pas la variabilité liée à l'estimation des valeurs manquantes.

II.6 SUPPRESSION DE LIGNES INCOMPLETES

a. Principe

Cette méthode consiste à supprimer complètement toutes les observations (lignes) qui contiennent au moins une valeur manquante. L'analyse est ensuite réalisée uniquement sur les données complètes restantes.

b. Avantage

Simplicité extrême : facile à comprendre et à mettre en œuvre, ne nécessitant aucun calcul d'estimation ou modélisation.

Pas d'introduction de biais par imputation : puisque seules les données complètes sont utilisées, aucune valeur artificielle n'est ajoutée.

Méthode standard dans certains logiciels : souvent la méthode par défaut dans de nombreux outils statistiques.

c. Inconvénients

Perte d'information importante : supprimer toutes les lignes incomplètes peut entraîner une réduction significative de la taille de l'échantillon, surtout si les données manquantes sont nombreuses.

Biais possible : si les données manquantes ne sont pas complètement aléatoires (MCAR), la suppression peut introduire un biais dans les résultats, car les observations conservées ne sont pas représentatives de la population initiale.

Réduction de la puissance statistique : la diminution du nombre d'observations diminue la précision et la fiabilité des analyses.

Impossibilité d'utiliser les informations partielles : les données disponibles sur certaines variables dans les lignes incomplètes sont ignorées, ce qui est une perte d'information.

II.7 IMPUTATION MULTIPLE

a. Principe

Imputation répétée : on remplace chaque donnée manquante par plusieurs valeurs plausibles tirées d'une distribution conditionnelle, ce qui produit plusieurs versions complètes du jeu de données.

Analyse séparée : chaque jeu complet est analysé indépendamment (ex. construction d'un modèle de régression).

Agrégation des résultats : les résultats des analyses sont combinés (moyenne des coefficients, estimation de la variance totale), ce qui permet d'intégrer la variabilité due à l'imputation.

b. Avantages

Réduction du biais : particulièrement efficace lorsque les données sont manquantes au hasard conditionnellement (MAR).

Meilleure estimation de la variance : contrairement aux imputations simples, elle prend en compte l'incertitude liée aux valeurs manquantes.

Flexibilité : applicable à différents types de données (quantitatives, qualitatives) et à des modèles complexes.

Recommandée par la littérature comme méthode de référence pour traiter les données manquantes

c. Inconvénients

Des hypothèses raisonnables sur le mécanisme des données manquantes (MCAR ou MAR).

Une attention particulière au choix des modèles conditionnels et aux paramètres d'imputation.

Des ressources computationnelles adaptées et une expertise pour son implémentation et l'interprétation des résultats.

Une prudence accrue lorsque le taux de données manquantes est très élevé ou que les données sont MNAR.

III. PRESENTATION DES MODELES DE REGRESSIONS

III.1 SVR (Support Vector Regression)

La régression par vecteurs de support (SVR, Support Vector Regression) est une méthode d'apprentissage supervisé issue des machines à vecteurs de support (SVM), adaptée aux problèmes de régression où la variable cible est continue.

a. Principe

Le SVR cherche à construire une fonction $f(x)$ qui approxime les données tout en restant « aussi plate que possible », c'est-à-dire avec une complexité contrôlée, en minimisant la norme du vecteur des coefficients.
Il utilise une fonction de coût ε-insensible : les erreurs inférieures à un seuil ε ne sont pas pénalisées, ce qui permet de tolérer une certaine marge d'erreur sans impact sur l'optimisation.
Le modèle équilibre la complexité (aplatissement de la fonction) et la précision (erreurs au-delà de ε) via un paramètre de régularisation C
Grâce au truc du noyau (kernel trick), SVR peut modéliser des relations non linéaires en projetant les données dans un espace de dimension plus élevée sans calcul explicite dans cet espace. Les noyaux courants sont linéaires, polynomial, et RBF (Radial Basis Function).

L'optimisation se fait en résolvant un problème convexe avec contraintes, souvent via des multiplicateurs de Lagrange, ce qui garantit une solution globale optimale.

b. Avantage

Capacité à modéliser des relations linéaires et non linéaires grâce aux noyaux.

Robustesse aux outliers par la zone ε-insensible qui ignore les petites erreurs.

Bonne généralisation même avec un nombre élevé de variables ou des variables corrélées.

Contrôle explicite de la complexité du modèle via les paramètres ε et C
Méthode bien adaptée aux petits et moyens jeux de données.

c. Inconvénients

Choix délicat des hyperparamètres (ε, C type et paramètres du noyau) nécessitant souvent une validation croisée.
Moins intuitive que la régression linéaire classique, ce qui peut compliquer l'interprétation des résultats.
Coût computationnel élevé pour de très grands ensembles de données.

Moins intuitive que la régression linéaire classique, ce qui peut compliquer l'interprétation des résultats.

Nécessite un prétraitement des données, notamment normalisation et gestion des variables catégorielles.

III.2 GradientBoostingRegressor

Le GradientBoostingRegressor est un algorithme d'apprentissage supervisé utilisé pour la régression, basé sur la technique du gradient boosting. Il combine plusieurs modèles faibles, généralement des arbres de décision, de manière séquentielle pour construire un modèle prédictif puissant.

a. Principe

Le modèle démarre par une prédiction initiale simple (souvent la moyenne de la variable cible).

À chaque itération, un nouvel arbre est entraîné pour corriger les erreurs (résidus) du modèle précédent, c'est-à-dire qu'il apprend à prédire les résidus entre les valeurs observées et les prédictions actuelles.

Cette correction s'effectue en suivant la direction du gradient négatif de la fonction de perte, ce qui permet d'optimiser progressivement la performance du modèle.

Le modèle final est la somme pondérée de tous les arbres construits, ce qui améliore la précision de la prédiction.

b. Avantage

Très performant sur des données tabulaires complexes, capable de modéliser des relations non linéaires et interactions entre variables.

Robuste aux données bruitées grâce à la correction progressive des erreurs.

Flexibilité : permet de choisir différentes fonctions de perte selon le problème (ex. erreur quadratique, absolue).

Contrôle fin des paramètres (nombre d'arbres, profondeur, taux d'apprentissage) pour éviter le surapprentissage.

Large adoption avec des implémentations optimisées (ex. XGBoost, LightGBM).

c. Inconvénients

Coût computationnel plus élevé que des modèles simples, surtout avec beaucoup d'arbres ou sur de très grands jeux de données.

Nécessite un réglage précis des hyperparamètres pour obtenir de bonnes performances.

Moins interprétable que des modèles linéaires simples, bien que des outils d'interprétabilité existent.

III.3 XGBoostRegressor (optionnel)

Le XGBoostRegressor est une implémentation avancée et optimisée du gradient boosting appliqué à la régression, reconnue pour sa rapidité, son efficacité et sa précision prédictive.

a. Principe

XGBoost construit un ensemble séquentiel d'arbres de décision où chaque nouvel arbre est entraîné pour prédire les résidus (erreurs) des arbres précédents.

L'objectif est de minimiser une fonction de perte (par défaut l'erreur quadratique) en suivant la direction du gradient négatif de cette perte, d'où le nom de boosting par gradient.

La fonction de perte est approximée via une série de Taylor, ce qui permet une optimisation efficace.

XGBoost intègre des techniques avancées comme la régularisation (L1 et L2), la gestion intelligente de la mémoire, le traitement parallèle, et le calcul distribué pour accélérer l'apprentissage.

Il gère aussi la rareté des données et optimise le choix des seuils de division dans les arbres pour améliorer la qualité des splits.

b. Avantages

Grande efficacité et rapidité grâce à l'optimisation du calcul et à la parallélisation.

Excellente performance prédictive sur des données complexes et volumineuses.

Flexibilité : nombreux hyperparamètres (profondeur des arbres, taux d'apprentissage, régularisation, etc.) permettant d'adapter le modèle au problème.

Robustesse face au surapprentissage grâce à la régularisation et au contrôle du nombre d'arbres.

Capacité à gérer des données avec des valeurs manquantes et des variables hétérogènes.

Large adoption dans la communauté machine learning, avec des interfaces en Python, R, C++, Java, etc..

c. Inconvénients

Nécessite un réglage fin des hyperparamètres pour obtenir des performances optimales.

Complexité du modèle qui peut rendre l'interprétation plus difficile que des modèles linéaires simples.

Coût computationnel plus élevé que des modèles plus simples, surtout pour des très grands ensembles de données ou un grand nombre d'arbres.

Peut être sensible à des données très bruitées si mal paramétré.

III.4 KNeighborsRegressor

Le KNeighborsRegressor est un algorithme de régression basé sur la méthode des k plus proches voisins (KNN), disponible dans la bibliothèque Python Scikit-learn.

a. Principe

Pour prédire la valeur d'une observation inconnue, l'algorithme identifie les k observations les plus proches dans l'espace des caractéristiques (selon une métrique de distance, souvent la distance de Minkowski).

La prédiction est obtenue par une interpolation locale des valeurs cibles des voisins sélectionnés.

Deux modes d'interpolation sont possibles :

Uniforme (par défaut) : tous les voisins ont le même poids dans la moyenne.

Distance : les voisins plus proches ont un poids plus important, inversement proportionnel à leur distance.

Paramètres clés
n_neighbors : nombre de voisins à considérer (par défaut 5).

weights : poids des voisins dans la prédiction (‘uniform’ ou ‘distance’).

metric : métrique de distance utilisée (par défaut Minkowski).

algorithm : méthode de recherche des voisins (auto, ball_tree, kd_tree, brute).

b. Avantage

Méthode non paramétrique simple et intuitive.

Capacité à modéliser des relations non linéaires sans supposer de forme fonctionnelle spécifique.

Flexible : choix du nombre de voisins et du mode de pondération.

Facile à implémenter et à comprendre.

c. Inconvénients

Sensibilité au choix de k : un k trop faible peut conduire à un surapprentissage, un k trop élevé à un sous-apprentissage.

Coût computationnel élevé en prédiction, surtout pour de grands ensembles de données, car il faut calculer les distances à tous les points d’entraînement.

Peu performant dans les espaces de grande dimension (malédiction de la dimension).

Dépendance à la métrique de distance choisie.

III.5 DecisionTreeRegressor

Le DecisionTreeRegressor est un algorithme d’apprentissage supervisé utilisé pour la régression, c’est-à-dire la prédiction de valeurs continues, basé sur la structure des arbres de décision.

a. Principe

L’arbre de décision construit un modèle sous forme d’un arbre où chaque nœud interne correspond à une condition sur une variable explicative (ex. « la taille est-elle supérieure à 1.5 ? »).

Les données sont divisées récursivement en sous-ensembles selon ces conditions, cherchant à minimiser l’erreur de prédiction dans chaque branche.

Les feuilles de l'arbre contiennent la valeur prédite, généralement la moyenne des valeurs cibles des observations qui y arrivent.

L'algorithme choisit les meilleurs seuils de division en fonction d'un critère de qualité de séparation, par défaut l'erreur quadratique moyenne (squared_error) qui correspond à la réduction de la variance.

L'arbre peut être limité en profondeur ou en nombre de feuilles pour éviter le surapprentissage.

Fonctionnement avec Scikit-learn
On crée un objet DecisionTreeRegressor().

On entraîne le modèle avec la méthode .fit(X_train, y_train) où X sont les variables explicatives et y la variable cible continue.

On prédit avec .predict(X_test).

b. Avantage

Modèle non paramétrique et interprétable : l'arbre est facile à visualiser et à comprendre.

Capacité à modéliser des relations non linéaires et interactions complexes entre variables.

Pas besoin de normalisation des données.

Rapide à entraîner et à prédire sur des jeux de données de taille moyenne.

Gestion automatique des variables numériques et catégorielles (avec encodage).

c. Inconvénients

Tendance au surapprentissage si l'arbre est trop profond, ce qui dégrade la généralisation.

Prédictions par morceaux constants : la fonction prédite est une fonction en escalier, ce qui peut manquer de lissage et de finesse.

Instabilité : petites variations dans les données peuvent conduire à des arbres très différents.

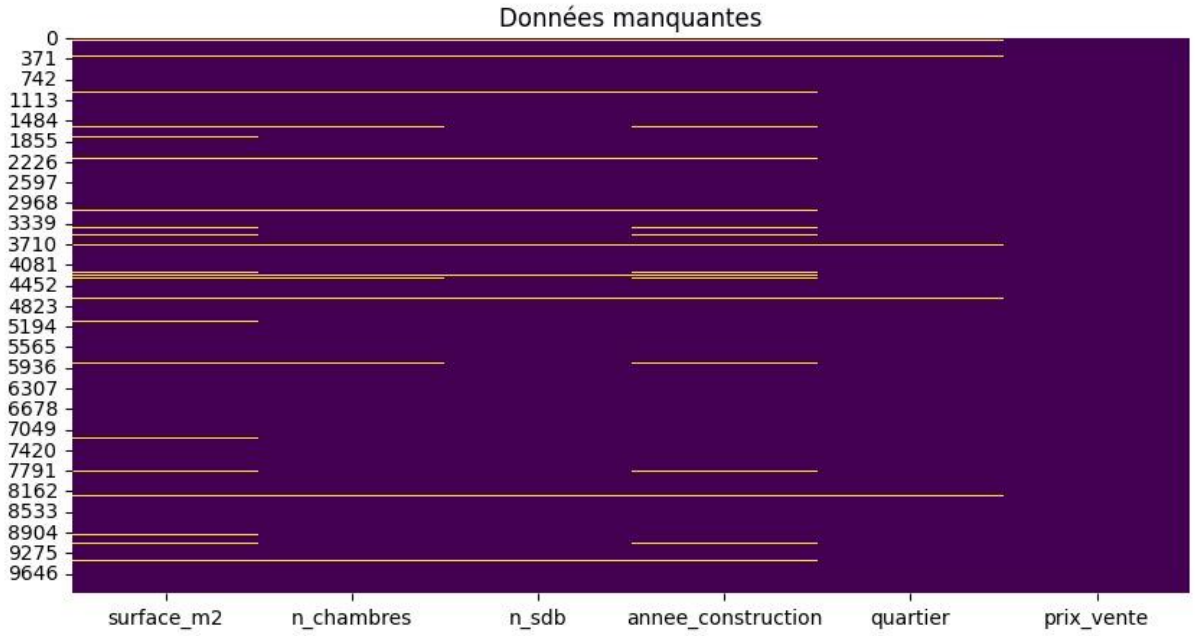
Moins performant que des méthodes d'ensemble (boosting, random forest) sur des problèmes complexes.

IV. RESULTATS

Avant application des différentes méthodes d'imputation, nous avons les données manquantes suivantes :

Colonne	Nombre de valeurs manquantes
surface_m2	500
n_chambres	300
n_sdb	200
annee_construction	400
quartier	100
prix_vente	0

La figure ci-dessous présente les données manquantes en orange pour chaque colonne

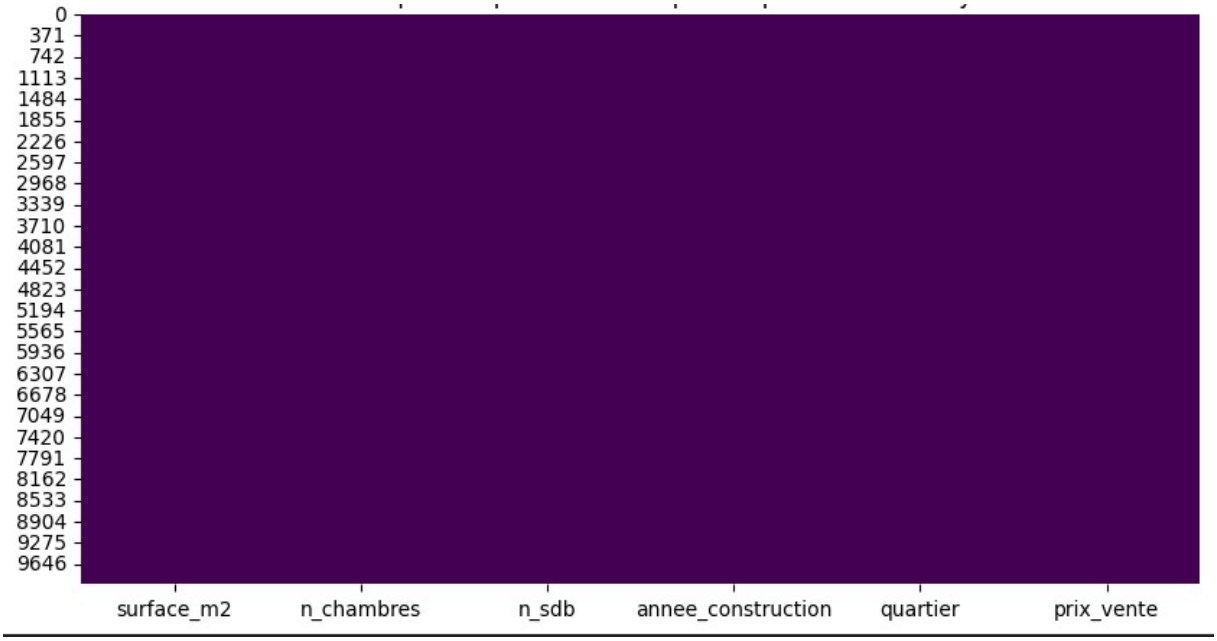


Après application des différentes méthodes d'imputation, nous avons les données manquantes suivantes :

Colonne	Nombre de valeurs manquantes
surface_m2	0
n_chambres	0
n_sdb	0
annee_construction	0
quartier	0
prix_vente	0

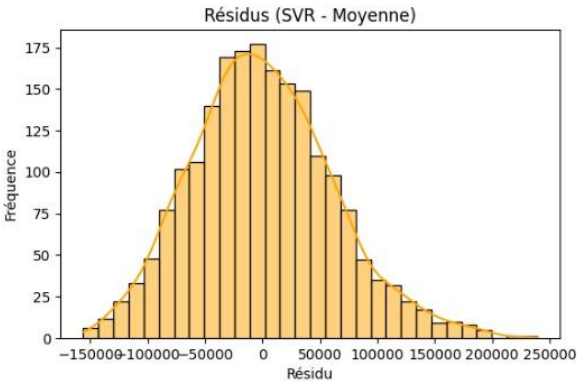
NB : pour toutes les méthodes utilisées, à part celle de suppression des lignes incomplètes, nous avons utilisé le mode pour la variable catégorielle quartier.

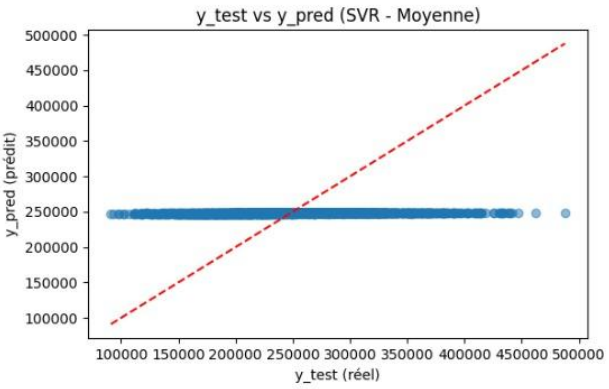
Voici la figure illustrant le résultat des différentes méthodes d'imputation



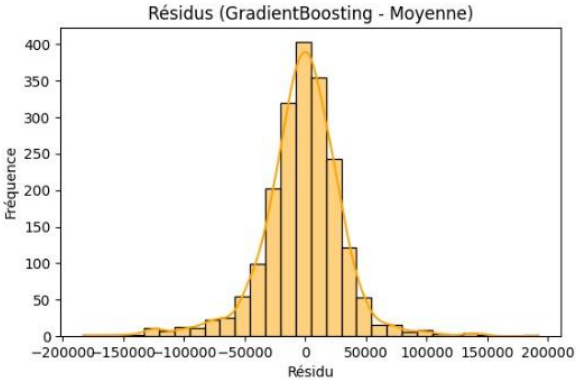
IV.1 Résultats selon l'imputation par la moyenne

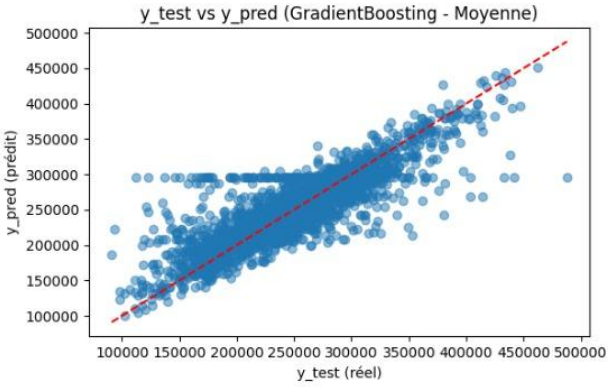
IV.1.1 La méthode SVR (Support Vector Regression)



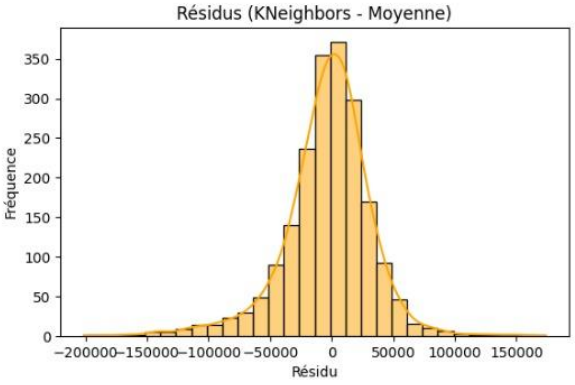


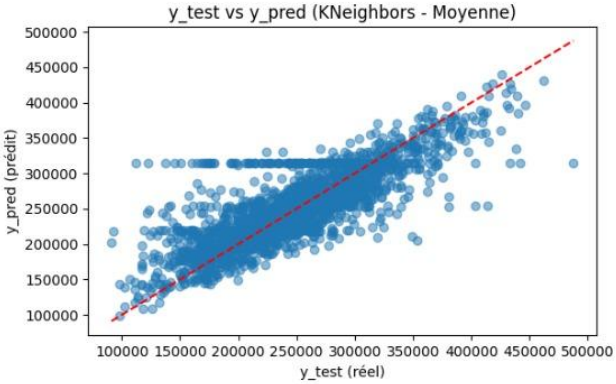
IV.1.2 La méthode GradientboostingRegressor



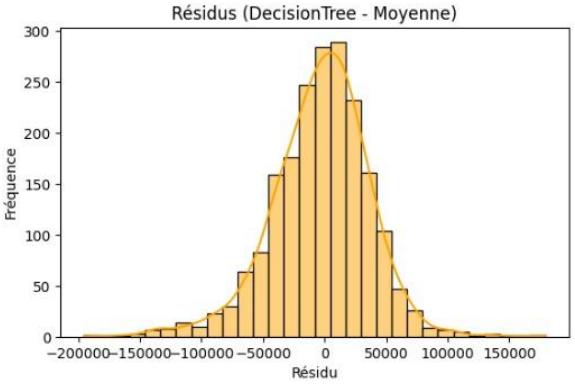


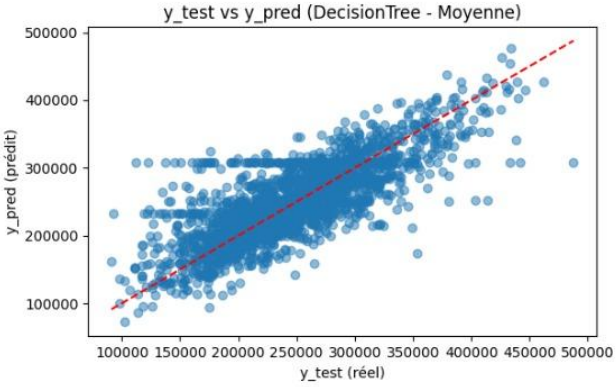
IV.1.3 La méthode KNeighborsRegressor



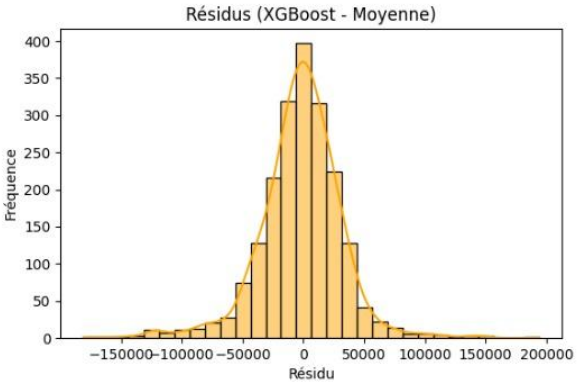


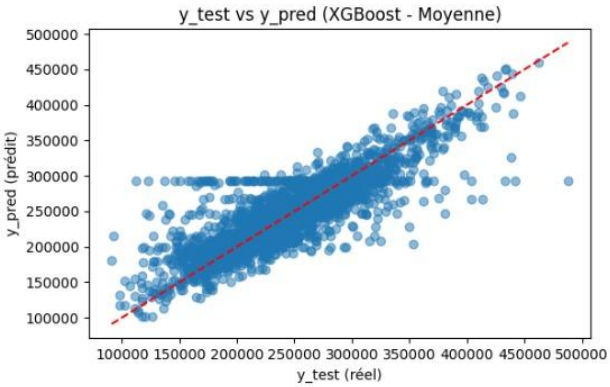
IV.1.4 La méthode DecisionTreeRegressor





IV.1.5 La méthode XGBoostRegressor

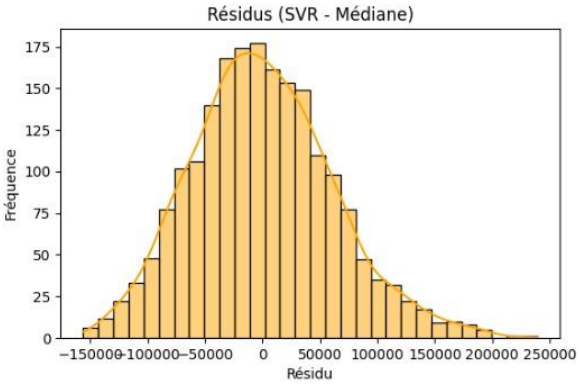


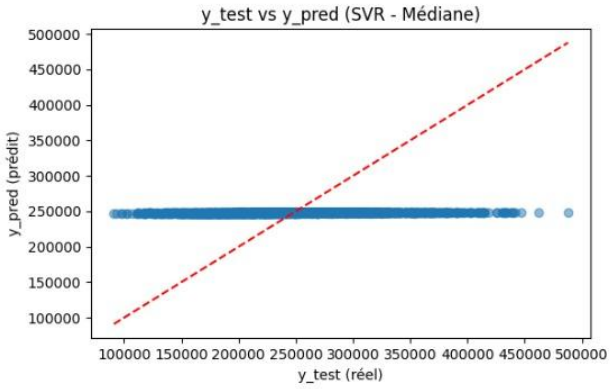


Imputation	Modèle	MAE	RMSE	R2
Moyenne	SVR	48875.539373	61679.148241	0.009643
Moyenne	GradientBoosting	23492.014655	33649.511214	0.705237
Moyenne	KNeighbors	25767.372680	36580.716572	0.651647
Moyenne	DecisionTree	30377.745820	40611.772956	0.570643
Moyenne	XGBoost	24264.000521	34208.790025	0.695357

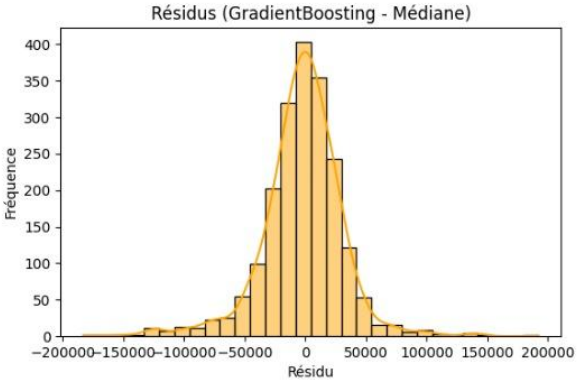
IV.2 Résultats selon l'imputation par la médiane

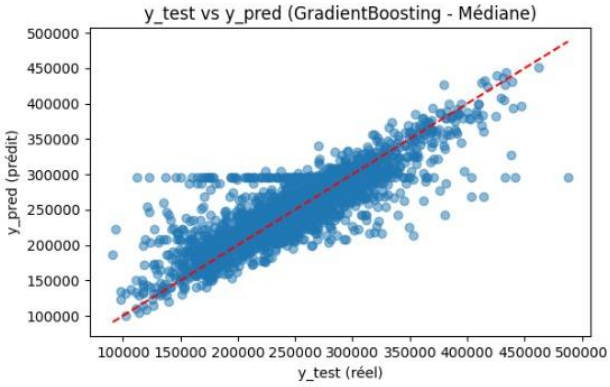
IV.2.1 La méthode SVR (Support Vector Regression)



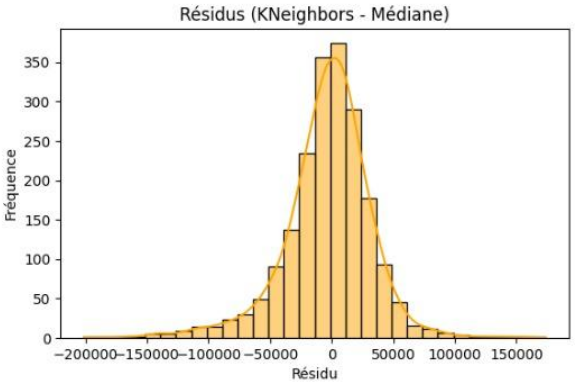


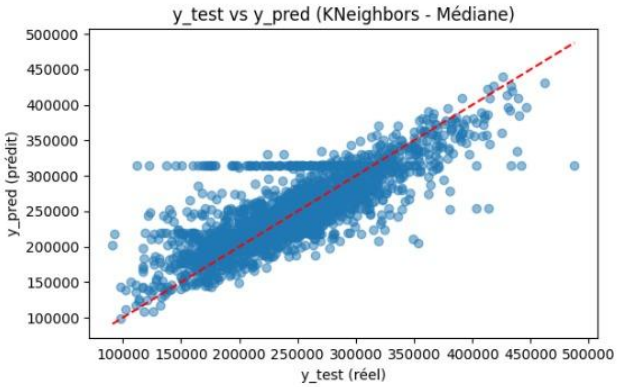
IV.2.2 La méthode GradientBoostingRegressor



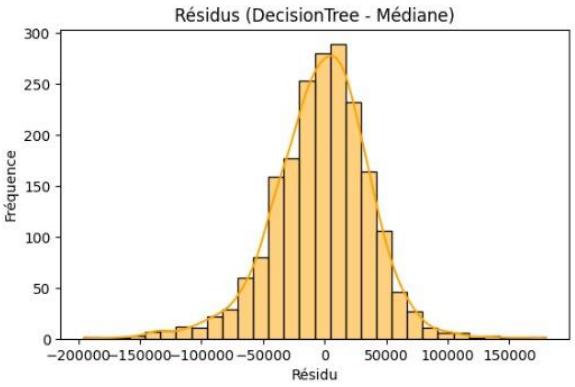


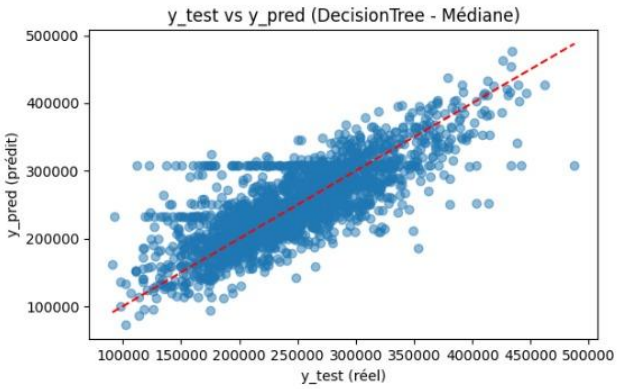
IV.2.3 La méthode KNeighborsRegressor



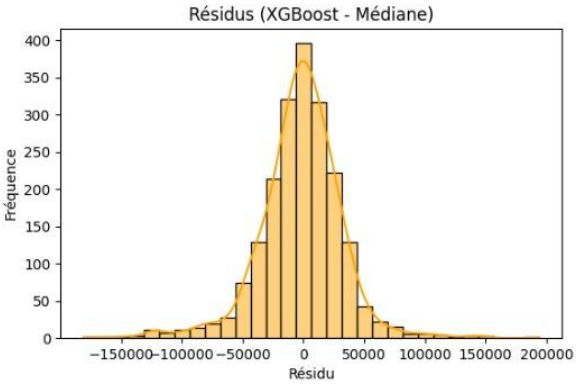


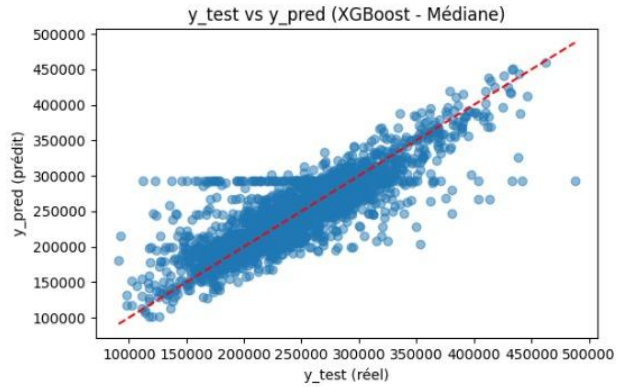
IV.2.4 La méthode DecisionTreeRegressor





IV.2.5 La méthode XGBoostRegressor

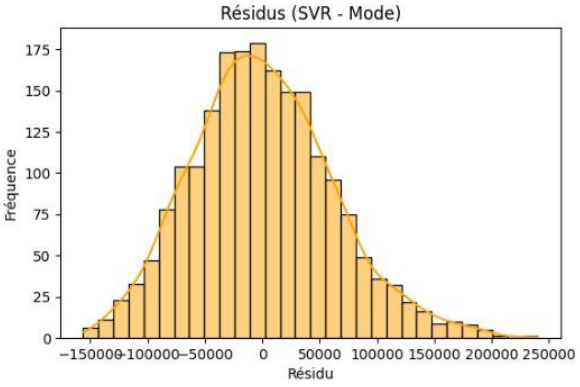


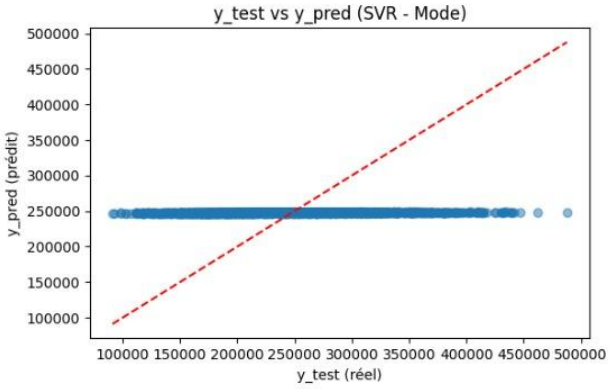


Imputation	Modèle	MAE	RMSE	R2
Médiane	SVR	48875.443368	61678.963136	0.009649
Médiane	GradientBoosting	23492.014655	33649.511214	0.705237
Médiane	KNeighbors	25798.755557	36563.605089	0.651973
Médiane	DecisionTree	30249.061475	40391.645251	0.575284
Médiane	XGBoost	24263.166631	34211.614885	0.695307

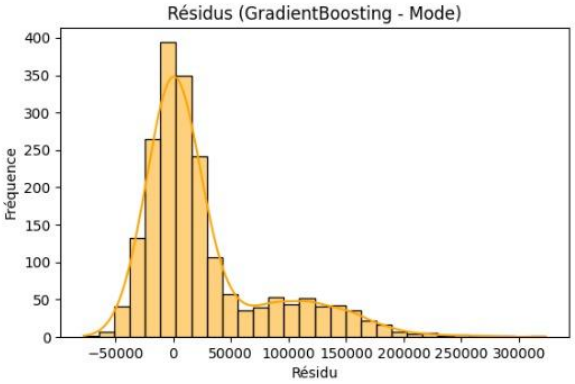
IV.3 Résultats selon l'imputation par le mode

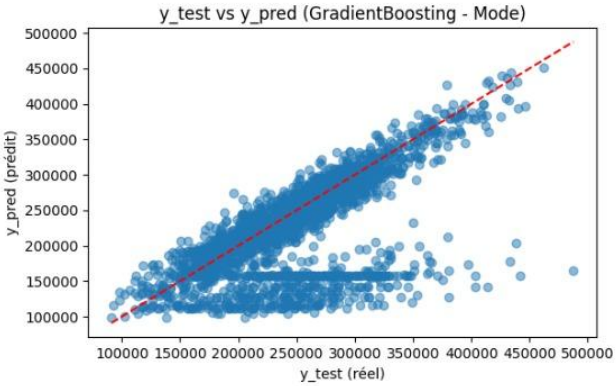
IV.3.1 La méthode SVR(Support Vector Regression)



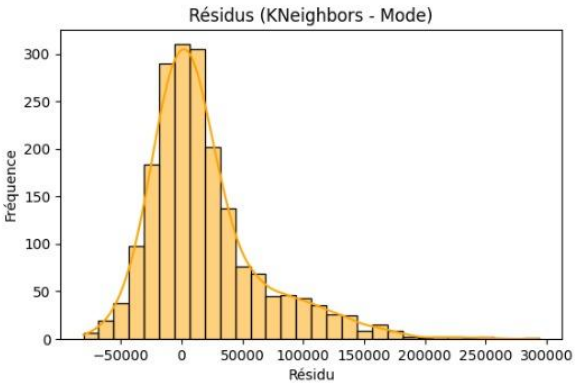


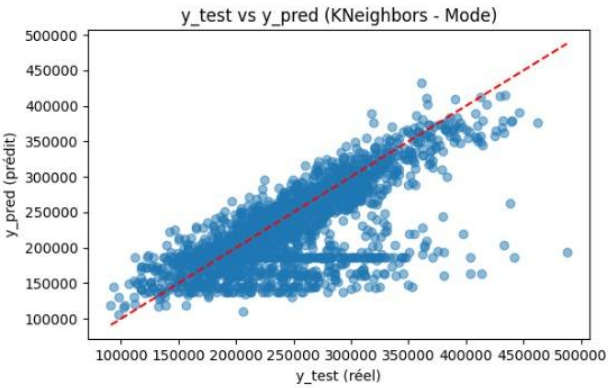
IV.3.2 La méthode GradientBoostingRegressor



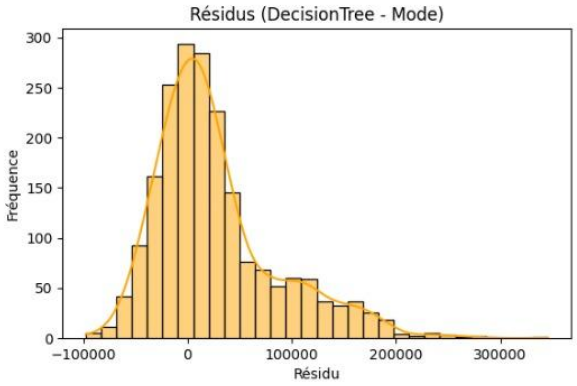


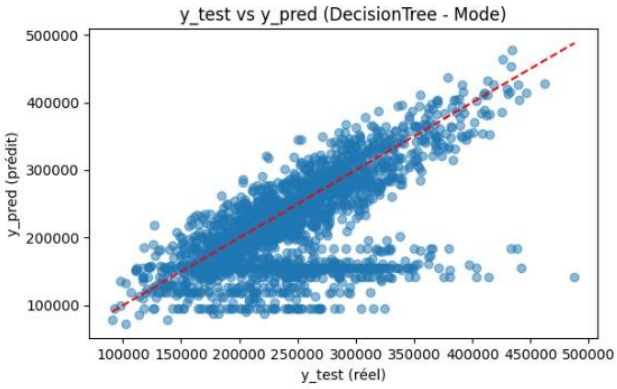
IV.3.3 La méthode KNeighborsRegressor



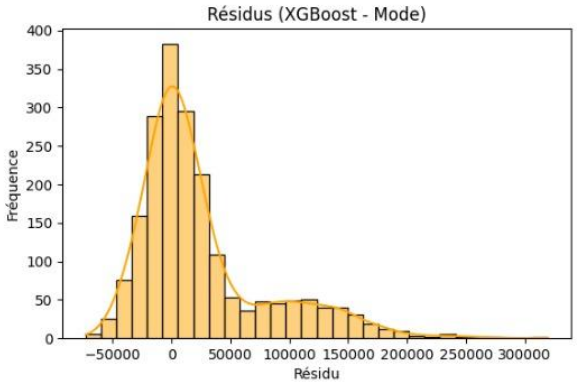


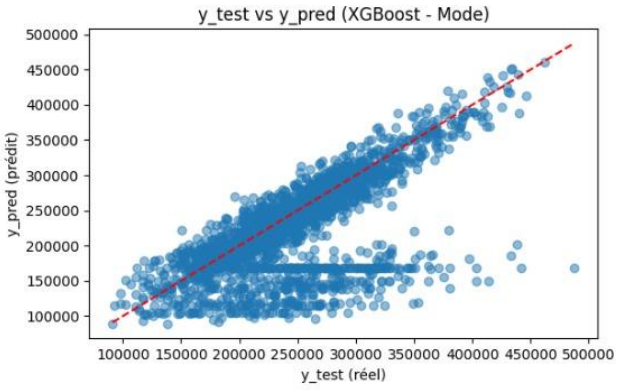
IV.3.4 La méthode DecisionTreeRegressor





IV.3.5 La méthode XGBoostRegressor

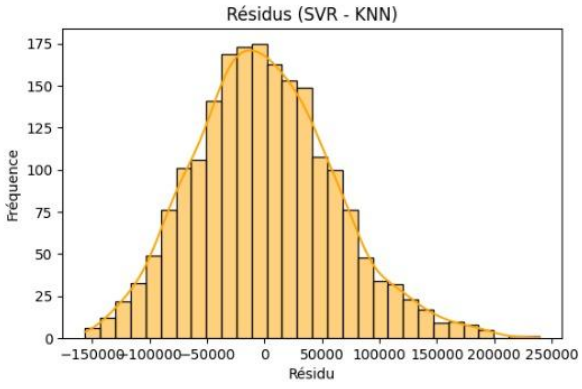


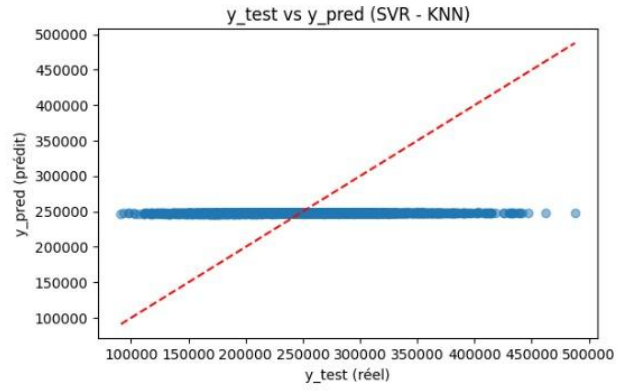


Imputation	Modèle	MAE	RMSE	R2
Mode	SVR	48874.212226	61676.773014	0.009719
Mode	GradientBoosting	38355.044661	61422.654545	0.017863
Mode	KNeighbors	34051.774285	50663.090701	0.331812
Mode	DecisionTree	44394.546530	64722.110958	-0.090487
Mode	XGBoost	38322.890348	60147.584853	0.058216

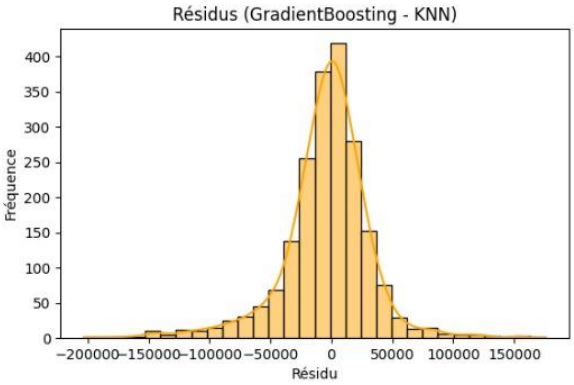
IV.4 Résultats selon l'imputation par KNN Imputer (KNN Imputer)

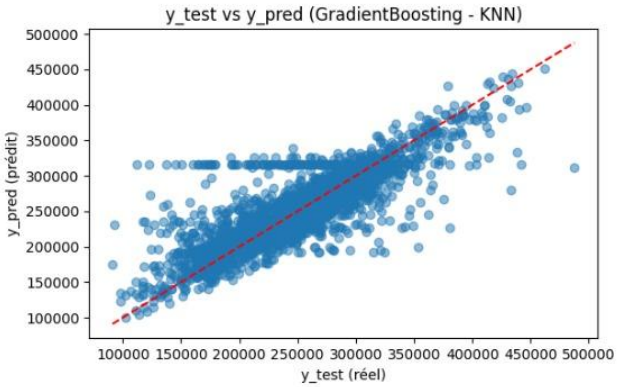
IV.4.1 La méthode SVR(Support Vector Regression)



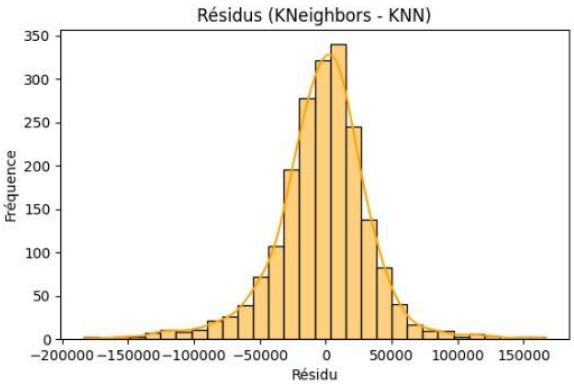


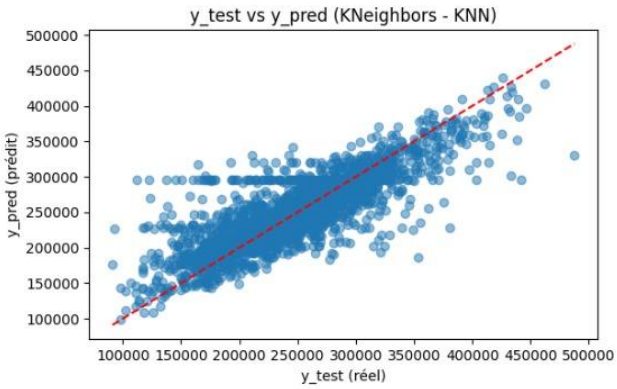
IV.4.2 La méthode GradientBoostingRegressor



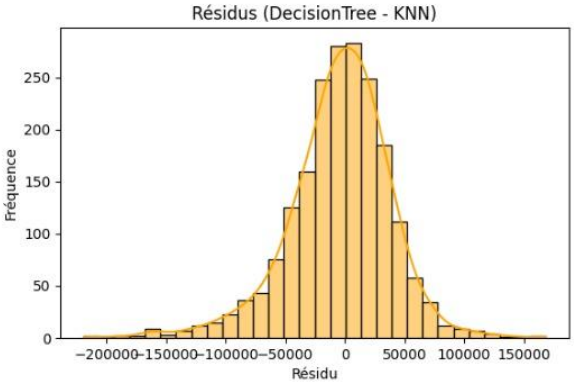


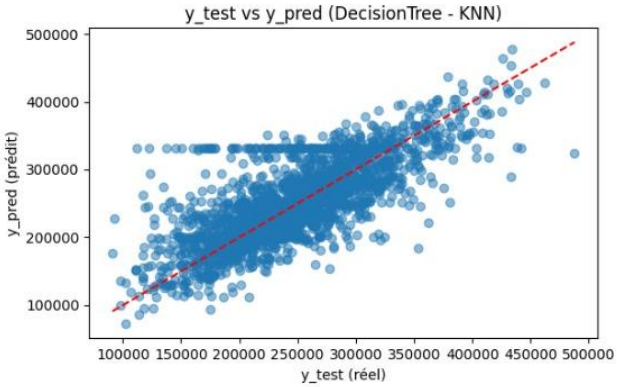
IV.4.3 La méthode KNeighborsRegressor



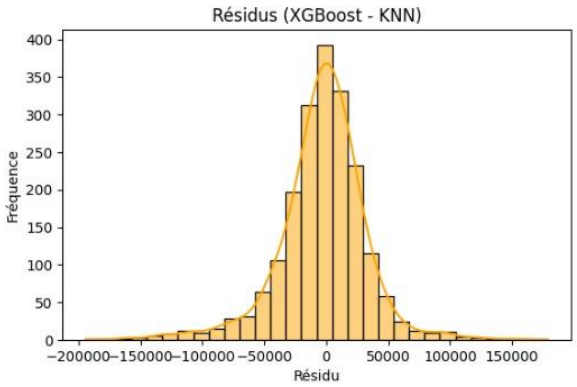


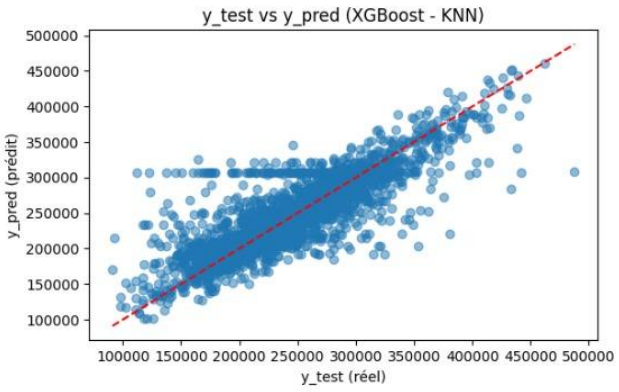
IV.4.4 La méthode DecisionTreeRegressor





IV.4.5 La méthode XGboost

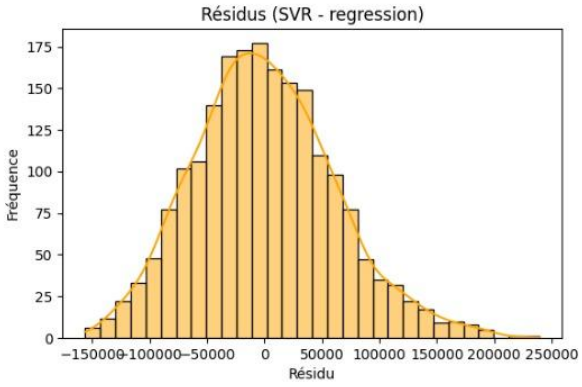


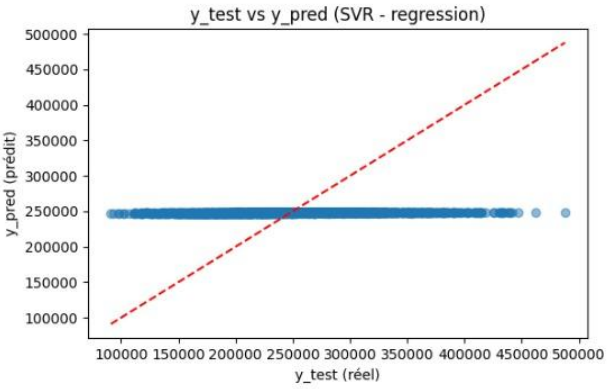


Imputation	Modèle	MAE	RMSE	R2
KNN	SVR	48880.399257	61683.560722	0.009501
KNN	GradientBoosting	24787.629773	36570.761081	0.651837
KNN	KNeighbors	25794.051471	36199.953334	0.658861
KNN	DecisionTree	32039.980870	43568.549712	0.505847
KNN	XGBoost	25267.159324	36341.755586	0.656183

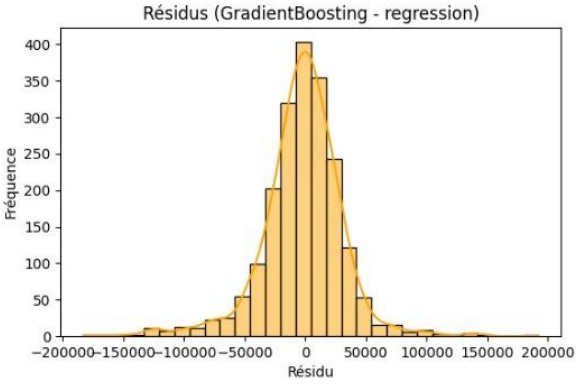
IV.5 Résultats selon l'imputation par régression

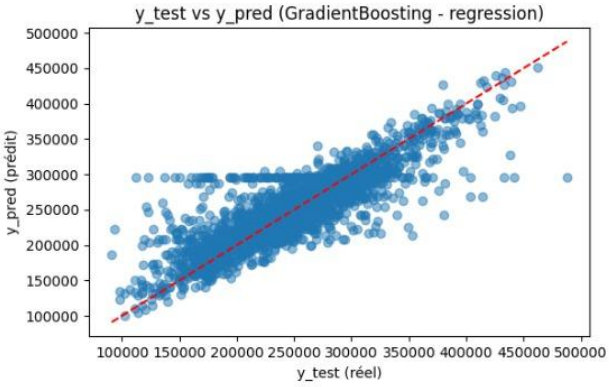
IV.5.1 La méthode SVR (Support Vector Regression)



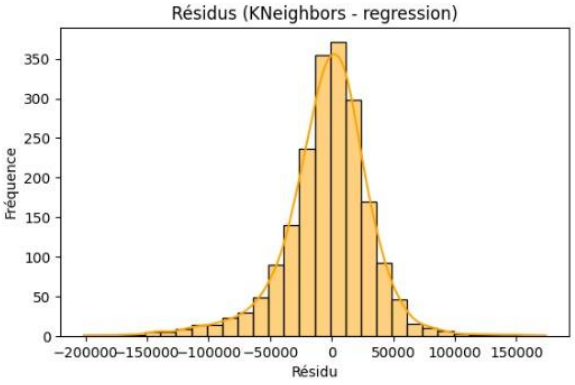


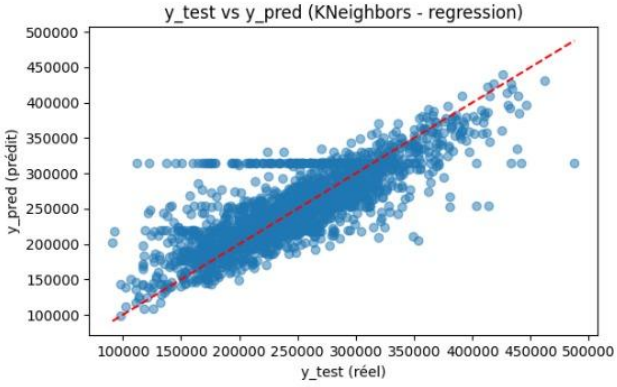
IV.5.2 La méthode GradientBoostingRegressor



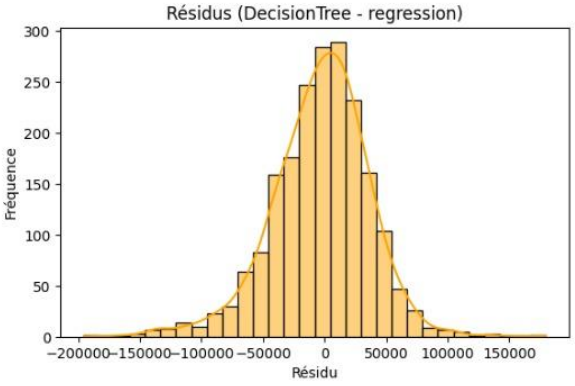


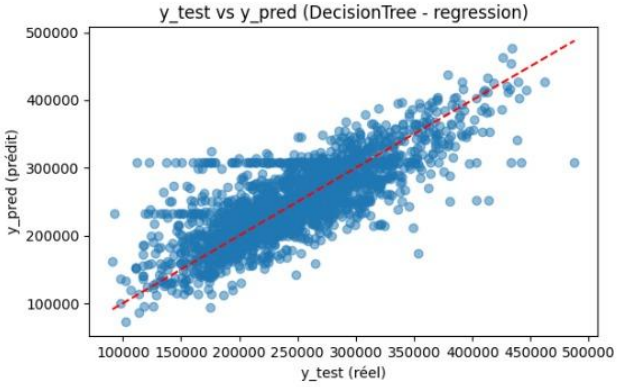
IV.5.3 La méthode KNeighborsRegressor



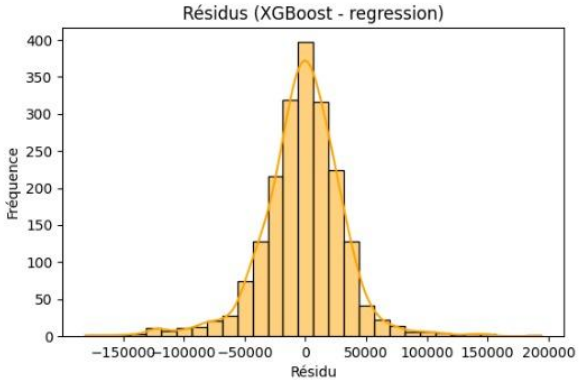


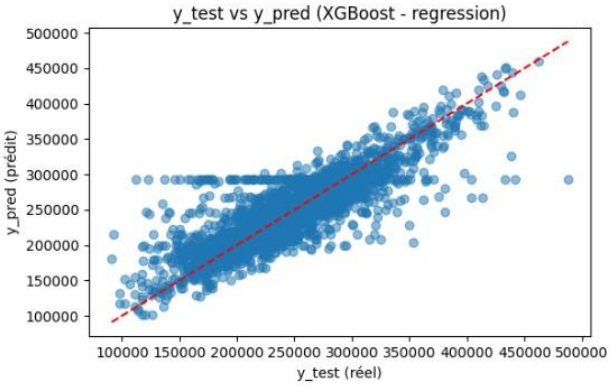
IV.5.4 La méthode DecisionTreeRegressor





IV.5.5 La méthode XGBoostRegressor

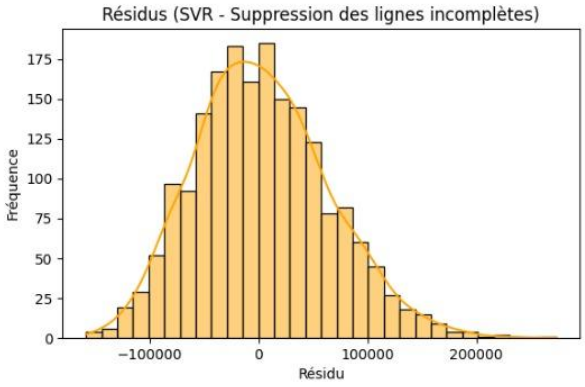


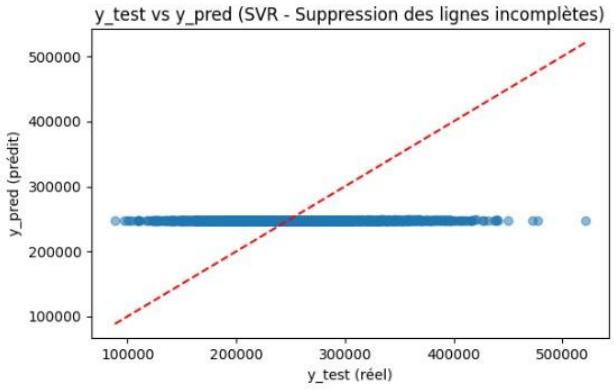


Imputation	Modèle	MAE	RMSE	R2
Régression	SVR	48875.42	61679.07	0.010
Régression	GradientBoosting	23492.01	33649.51	0.705
Régression	KNeighbors	25767.37	36580.72	0.652
Régression	DecisionTree	30377.75	40611.77	0.571
Régression	XGBoost	24264.00	34208.79	0.695

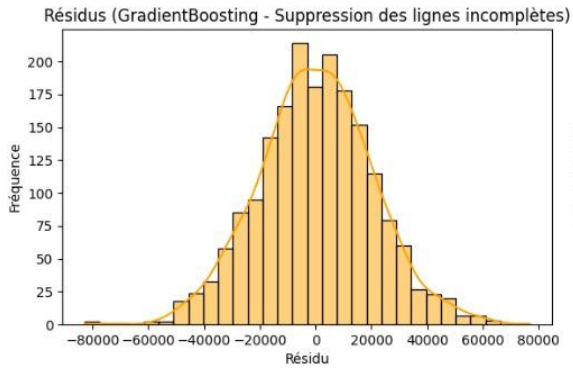
IV.6 Résultats selon l'imputation par la Suppression de lignes incomplètes

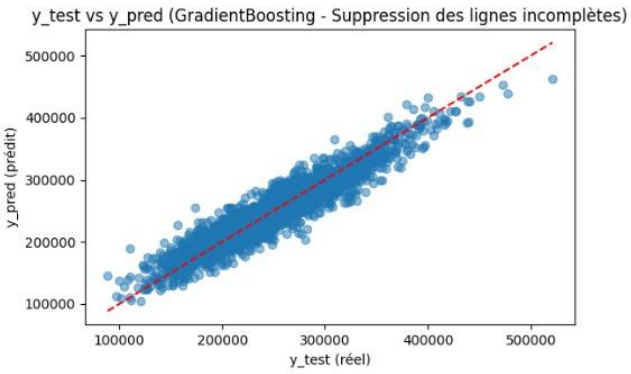
IV.6.1 La méthode SVR (Support Vector Regression)



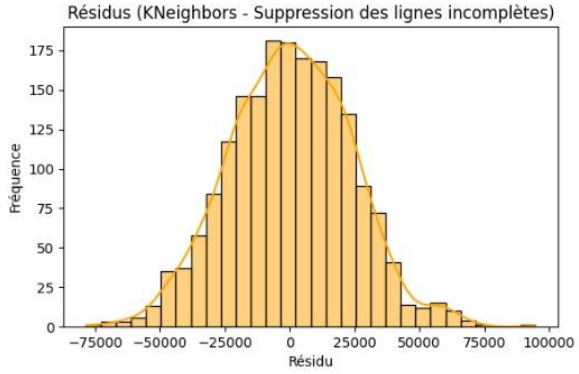


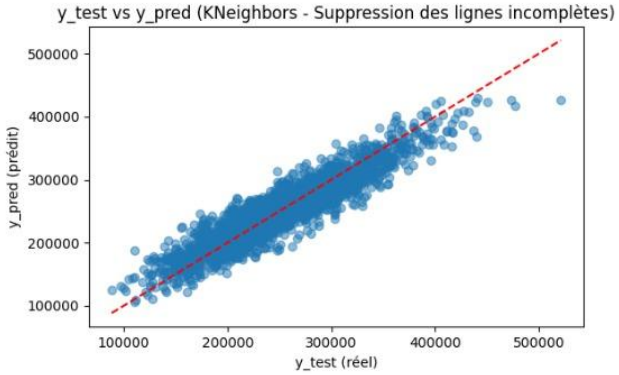
IV.6.2 La méthode GradientBoostingRegressor



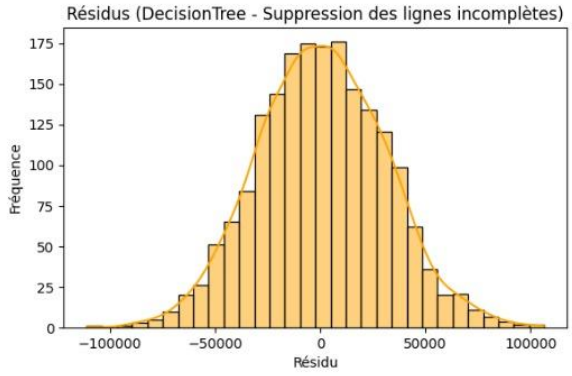


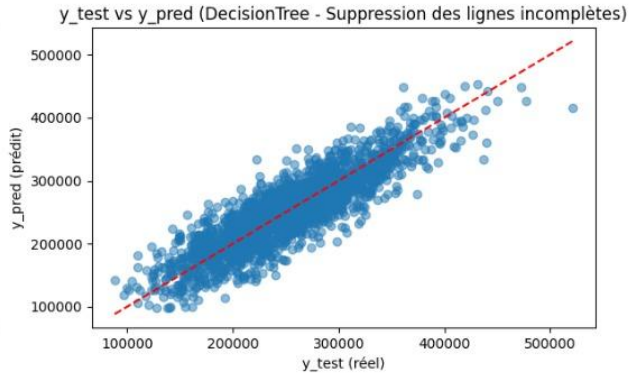
IV.6.3 La méthode KNeighborsRegressor



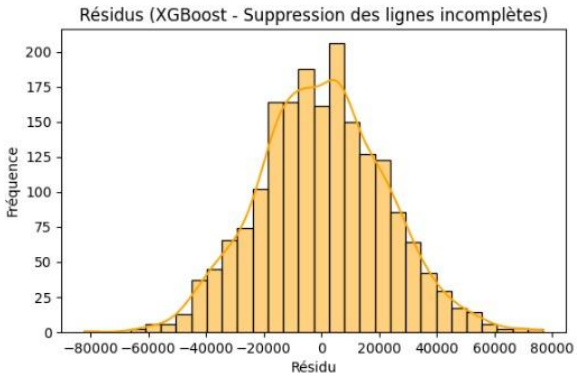


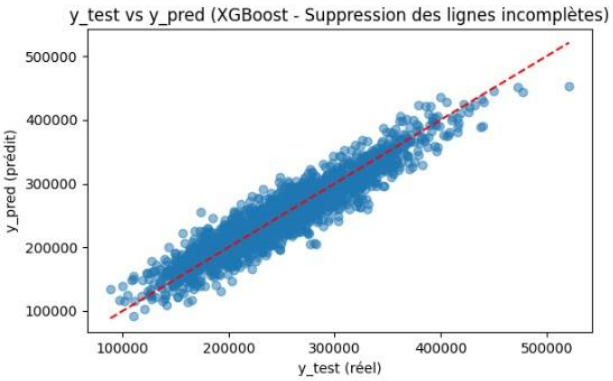
IV.6.4 La méthode DecisionTreeRegressor





IV.6.5 La méthode XGBoost

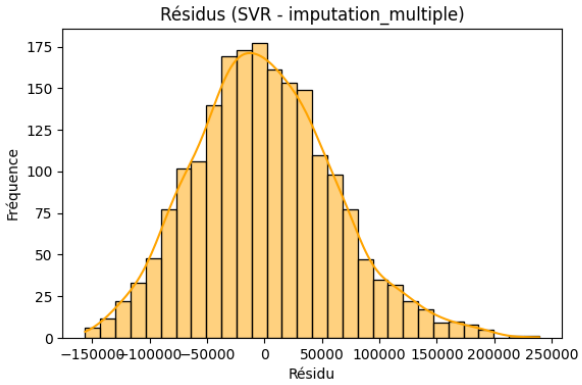


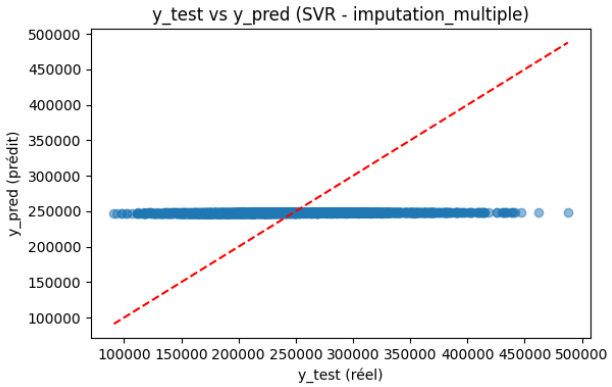


Imputation	Modèle	MAE	RMSE	R2
Suppression	SVR	49056.252358	61632.013989	0.007830
Suppression	GradientBoosting	16229.154633	20544.680077	0.889752
Suppression	KNeighbors	18764.476611	23458.806135	0.856258
Suppression	DecisionTree	24352.516882	30482.778629	0.757293
Suppression	XGBoost	17472.427461	21930.481173	0.874377

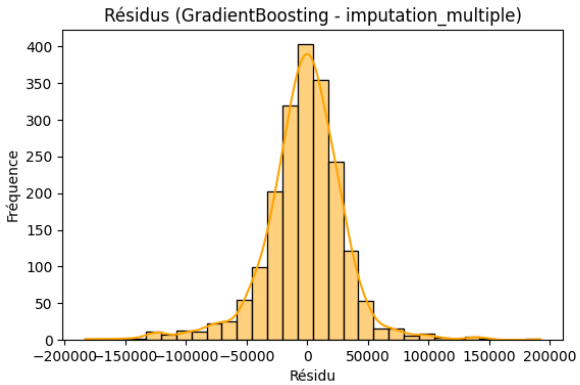
IV.7 Résultats selon l'imputation multiple

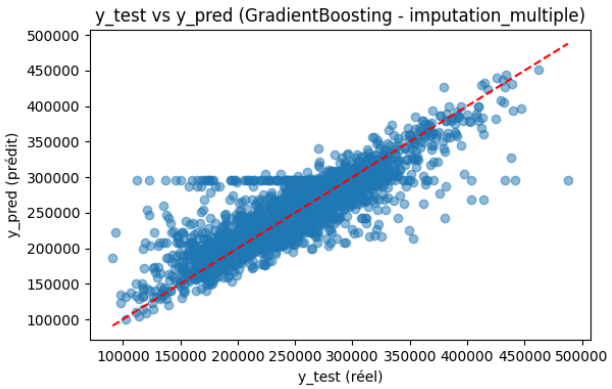
IV.7.1 La méthode SVR (Support Vector Regression)



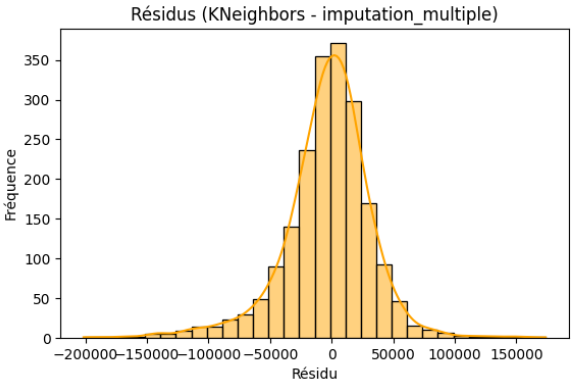


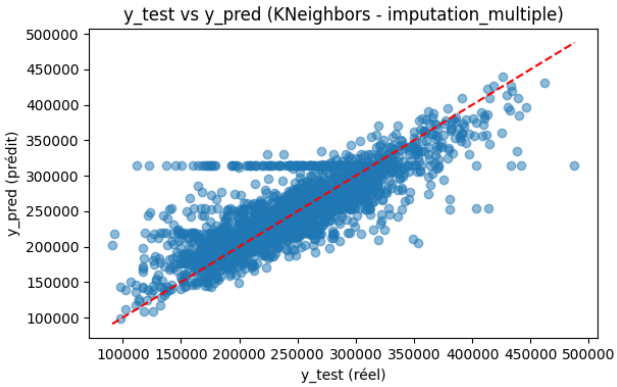
IV.7.2 La méthode GradientBoostingRegressor



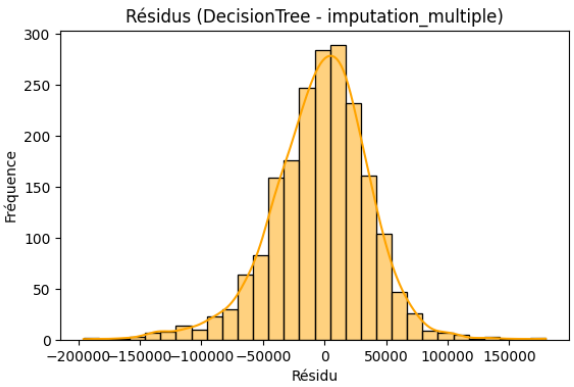


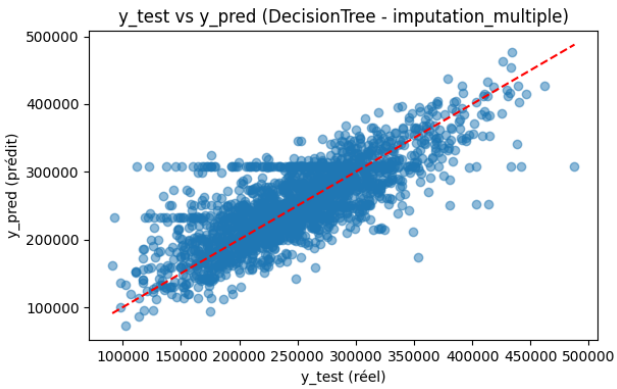
IV.7.3 La méthode KNeighborsRegressor



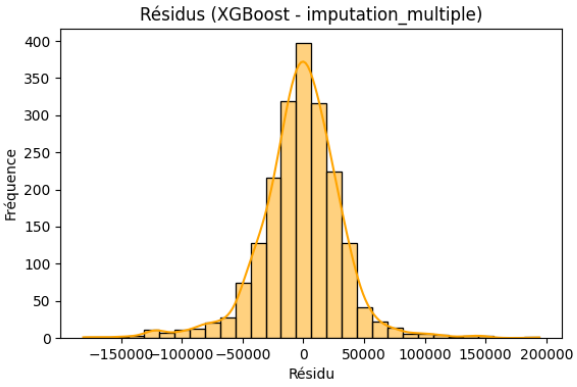


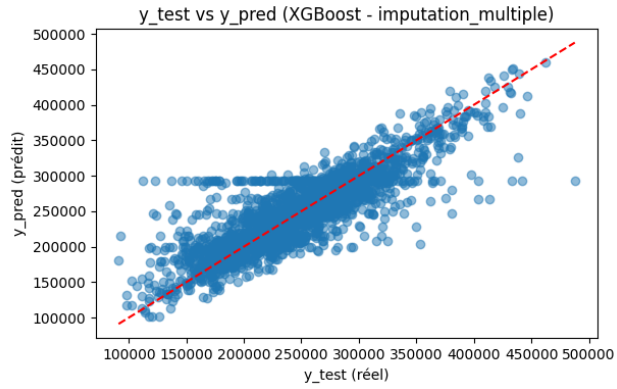
IV.7.4 La méthode DecisionTreeRegressor





IV.7.5 La méthode XGBoost





Imputation	Modèle	MAE	RMSE	R^2
imputation_multiple	SVR	48875.422878	61679.065202	0.009645
imputation_multiple	GradientBoosting	23492.014655	33649.511214	0.705237
imputation_multiple	KNeighbors	25767.372680	36580.716572	0.651647
imputation_multiple	DecisionTree	30377.745820	40611.772956	0.570643
imputation_multiple	XGBoost	24264.000521	34208.790025	0.695357

V. ANALYSE ET CRITIQUE

V.1 ANALYSE PAR MODELE

V.1.1 SVR (Support Vector Regression)

Les performances restent très similaires pour toutes les méthodes, y compris l'imputation multiple (MAE ≈ 48 875, $R^2 \approx 0.01$).

L'imputation multiple ne modifie pas significativement les résultats, confirmant que SVR est peu adapté ici ou nécessite un paramétrage plus fin.

Aucune méthode d'imputation ne se démarque, performances globalement faibles.

V.1.2 GradientBoosting

La suppression des lignes incomplètes reste la méthode la plus performante (MAE = 16 229, $R^2 = 0.89$).

L'imputation multiple (MAE = 23 492, $R^2 = 0.705$) donne des résultats identiques à ceux obtenus avec moyenne et médiane, donc moins performante que la suppression, mais robuste.

KNN est légèrement moins bonne, mode nettement inférieure.

Recommandation : suppression si possible, sinon imputation multiple ou moyenne/médiane.

V.1.3 KNeighbors

Suppression des lignes incomplètes domine toujours (MAE = 18 765, $R^2 = 0.856$).

Imputation multiple (MAE = 25 767, $R^2 = 0.652$) est équivalente à moyenne/médiane/KNN.

Mode est nettement moins performante.

V.1.4 DecisionTree

Suppression des lignes incomplètes reste la meilleure (MAE = 24 353, $R^2 = 0.757$).

Imputation multiple (MAE = 30 378, $R^2 = 0.570$) équivalente à moyenne/médiane.

KNN moins bonne, mode très mauvaise.

V.1.5 XGBoost

Suppression des lignes incomplètes est la plus performante (MAE = 17 472, $R^2 = 0.874$).

Imputation multiple (MAE = 24 264, $R^2 = 0.695$) équivalente à moyenne/médiane.

KNN légèrement moins bonne, mode la moins performante.

V.2 SYNTHESE ET RECOMMANDATIONS

Suppression des lignes incomplètes reste la méthode d'imputation la plus robuste et performante pour tous les modèles avancés.

Imputation multiple par MICE produit des performances comparables à celles des imputations simples par moyenne et médiane, confirmant sa robustesse et sa fiabilité.

Moyenne et médiane restent des alternatives valables si la suppression n'est pas envisageable.

KNN est une méthode intermédiaire, meilleure que le mode mais moins bonne que les autres.

Mode est systématiquement la moins performante, avec des scores faibles voire négatifs.

SVR est peu sensible à la méthode d'imputation et globalement peu performant dans ce contexte.

V.3 LES CRITIQUES

V.3.1 LIMITES DES METHODES D'IMPUTATION SIMPLES (MOYENNE, MEDIANE, MODE)

Réduction artificielle de la variance : Comme précédemment, ces méthodes imputent une valeur constante pour toutes les données manquantes d'une variable, ce qui réduit la variabilité naturelle des données et peut biaiser les analyses statistiques.

Perte des relations entre variables : Ces imputations ne prennent pas en compte les corrélations ou interactions entre variables, ce qui peut dégrader la performance des modèles prédictifs, notamment en régression.

Sensibilité aux distributions : L'imputation par la moyenne reste sensible aux valeurs extrêmes, la médiane est plus robuste, mais aucune ne restitue la complexité et la variabilité des données manquantes.

Imputation par le mode : Confirmée comme la moins performante, surtout pour les variables quantitatives, où elle peut fortement biaiser les résultats et dégrader les performances prédictives.

Comparaison avec l'imputation multiple : Les résultats montrent que l'imputation multiple produit des performances très proches de celles de la moyenne et de la médiane, ce qui souligne la robustesse relative de ces méthodes simples dans certains contextes, mais sans restituer l'incertitude liée aux valeurs manquantes.

V.3.2 SUPPRESSION DES LIGNES INCOMPLETES : AVANTAGES ET INCONVENIENTS

Avantage : Méthode simple qui ne modifie pas les valeurs observées, évitant l'introduction de valeurs artificielles.

Inconvénient majeur : Perte potentiellement importante d'informations, surtout si le taux de données manquantes est élevé, ce qui peut réduire la puissance statistique et introduire un biais si les données ne sont pas manquantes complètement au hasard (MCAR).

Résultats observés : Cette méthode donne les meilleures performances pour la plupart des modèles avancés (GradientBoosting, XGBoost, KNeighbors, DecisionTree), ce qui suggère un faible taux de données manquantes ou une distribution non biaisée des données manquantes.

Prudence : Il ne faut pas généraliser cette supériorité si le taux de données manquantes est important ou si le mécanisme est non aléatoire (MNAR).

V.3.3 METHODES PLUS AVANCEES ET ITERATIVES

Imputation multiple (MICE) et autres méthodes itératives (forêts aléatoires, régressions multiples) permettent de mieux capturer la variabilité naturelle des données et les relations complexes entre variables, réduisant ainsi le biais et la sous-estimation de la variance.

Intégration de l'incertitude : L'imputation multiple permet d'intégrer l'incertitude liée aux valeurs manquantes en générant plusieurs imputations et en combinant les résultats, ce qui améliore la fiabilité des inférences.

Performance observée : Dans vos résultats, l'imputation multiple donne des performances comparables à celles des imputations simples (moyenne, médiane), confirmant sa robustesse mais sans dépasser la suppression des lignes incomplètes dans ce contexte.

Recommandation : Ces méthodes restent recommandées dans la littérature pour des analyses robustes, notamment lorsque la suppression n'est pas envisageable.

V.3.4 CRITIQUE SPECIFIQUE DES RESULTATS OBTENUS

Performances faibles du SVR quelle que soit la méthode d'imputation : Cela indique que SVR n'est pas adapté au jeu de données actuel, ou que ses hyperparamètres nécessitent un réglage plus fin.

Suppression des lignes incomplètes donne les meilleurs scores pour la plupart des modèles avancés : Cela peut refléter un faible taux de données manquantes ou une distribution non biaisée des données manquantes, mais cette conclusion doit être nuancée si le taux ou le mécanisme des données manquantes change.

Imputation par mode systématiquement la moins performante : Confirme les limites bien documentées de cette méthode, notamment la forte perte d'information et la dégradation des performances prédictives.

Moyenne et médiane proches en performances : Comme attendu, mais aucune ne restitue la variabilité ni la complexité des données manquantes.

Imputation KNN intermédiaire : Cohérente avec son principe d'utiliser la structure locale des données, mais sensible à la malédiction de la dimension et au choix du paramètre k.

Imputation multiple par MICE : Offre une alternative robuste, intégrant l'incertitude, avec des performances équivalentes aux imputations simples dans ce cas précis, mais potentiellement plus fiable dans des contextes complexes ou avec des taux plus élevés de

A l'issu des analyses et critiques nous constatons que la méthode de suppressions des lignes incomplètes est la méthode la plus performante dans ce jeu de données, mais sa pertinence dépend du taux et du mécanisme des données manquantes.

VI. CONCLUSION

La suppression des lignes incomplètes s'est avérée la méthode d'imputation la plus performante pour la majorité des modèles, notamment GradientBoosting et XGBoost, grâce à un faible taux de données manquantes. L'imputation multiple par MICE offre une alternative robuste, intégrant l'incertitude liée aux valeurs manquantes, avec des performances comparables aux imputations simples (moyenne, médiane). L'imputation par mode est à éviter en raison de ses faibles résultats. Enfin, SVR montre des performances globalement faibles, indépendamment de la méthode d'imputation. Le choix de la méthode doit toujours être adapté au contexte des données et à l'objectif d'analyse.