REGRESSION– CONSOMMATION DE CARBURANTS

Objectif: Prédire la consommation de carburant des voitures en fonction de caractéristiques techniques

Données: AutoMPG - 398 observations, 8 variables

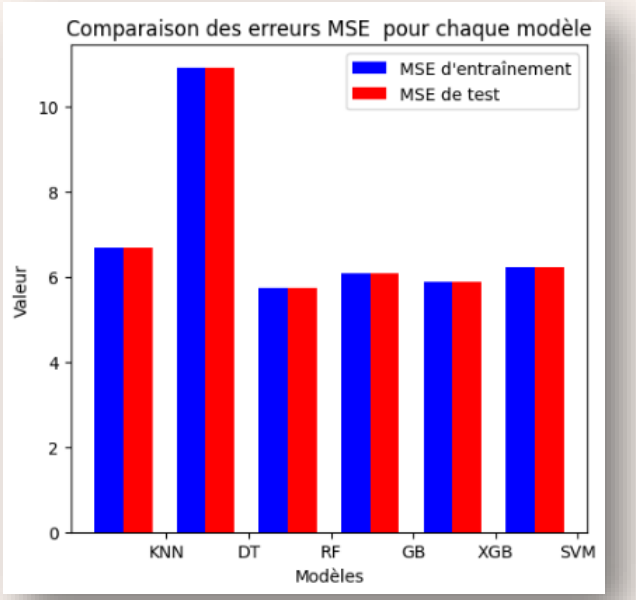
Variables quantitatives
Displacement - Cylindrée
Cylinders - Cylindres
Horsepower - Puissance
Weight - Poids
Acceleration - Accélaration
Model_year - Année du modèle
Origin - Origine
Mpg - Consommation

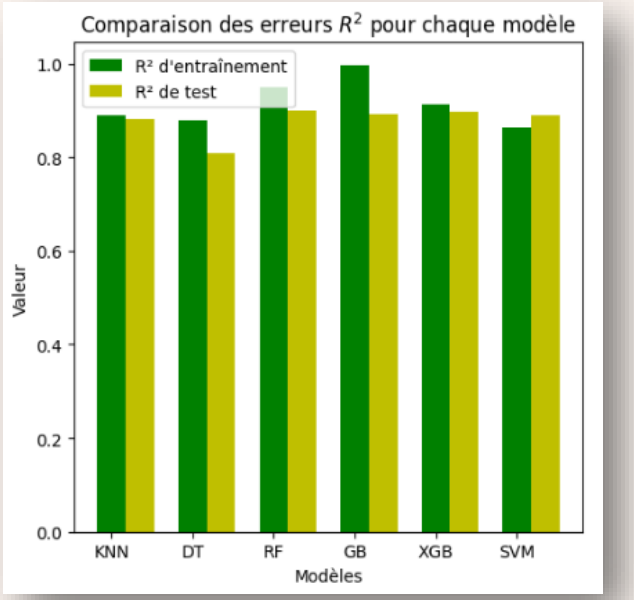
Modèles de régression
KNN
Arbres de décision
Forêt aléatoire
Gradient Boosting
XGBoost

Evaluation/Métriques utilisées
R^2
MSE
RMSE
MAE

RÉSULTATS - COMPARAISON DES MODÈLES

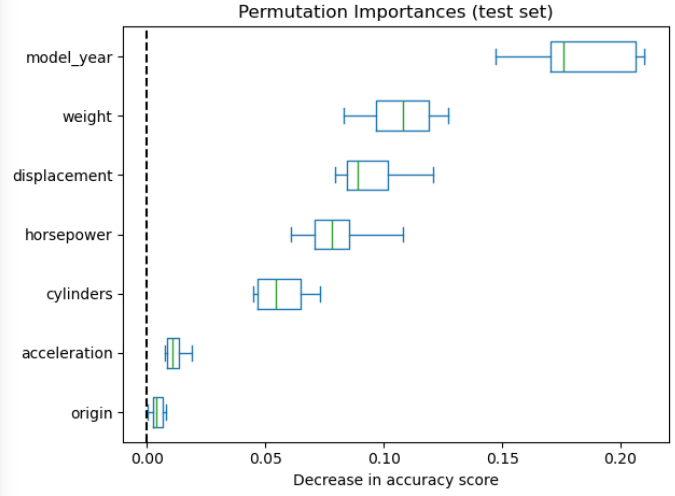
Index	Modèle	MSE	RMSE	MAE	R2
0	KNN	6.68	2.59	1.94	0.88
1	DecisionTree	10.89	3.30	2.45	0.80
2	RandomForest	5.73	2.39	1.77	0.90
3	GradientBoosting	6.07	2.46	1.79	0.89
4	XGBoost	5.87	2.42	1.80	0.90
5	SVM	6.22	2.49	1.81	0.89



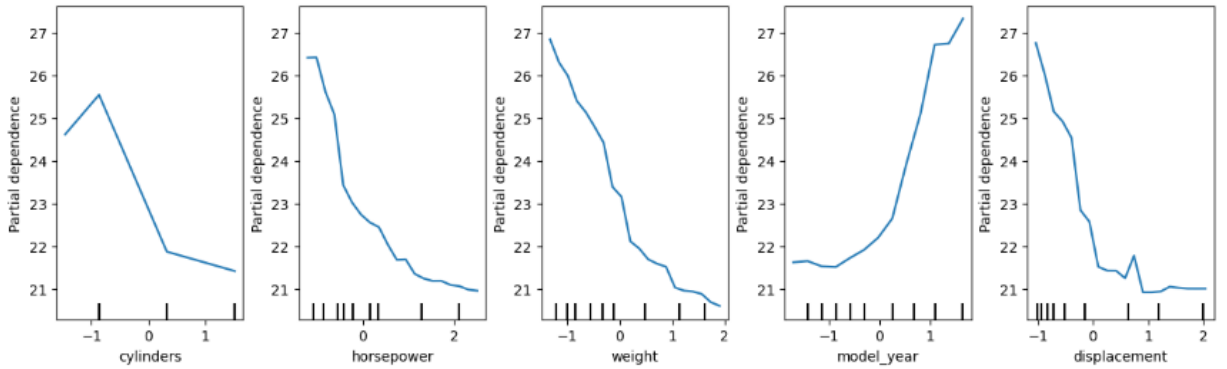


CONCLUSION ET AMELIORATIONS

IMPORTANCE DES VARIABLES PAR PERMUTATION



GRAPHES DE DEPENDANCE PARTIELLE DES VARIABLES



AMELIORATIONS A ENVISAGER

- INTÉGRER PLUS DE DONNÉES (OBSERVATIONS)
- INTEGRER DE NOUVELLES VARIABLES
- TESTER D'AUTRES MODELES
- TESTER D'AUTRES OPTIMISATIONS DE PARAMETRES DES MODELES

CLASSIFICATION BINAIRE – SURVIVANTS AU NAUFRAGE DU TITANIC

Objectif: Savoir si un passager a survécu en fonction de caractéristiques personnelles

Données: Titanic - 1309 observations et 14 variables

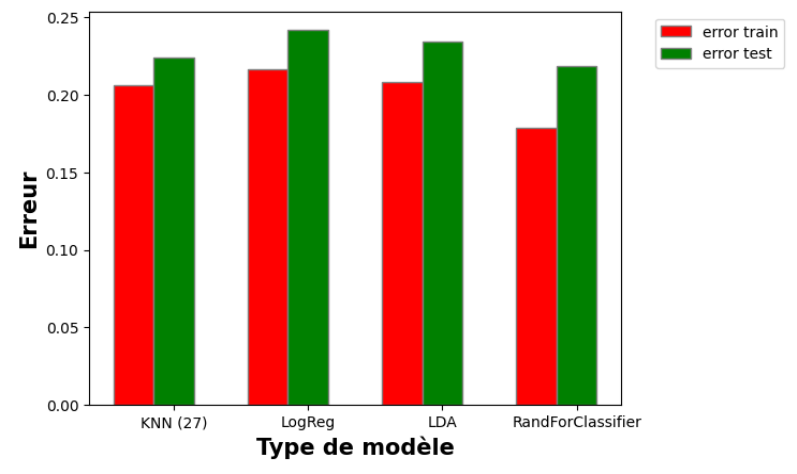
Variables quantitatives	Variables catégorielles
age	pclass
sibsp	survived
parch	name
fare	sex
body	cabin
	embarked
	boat
	Home.dest

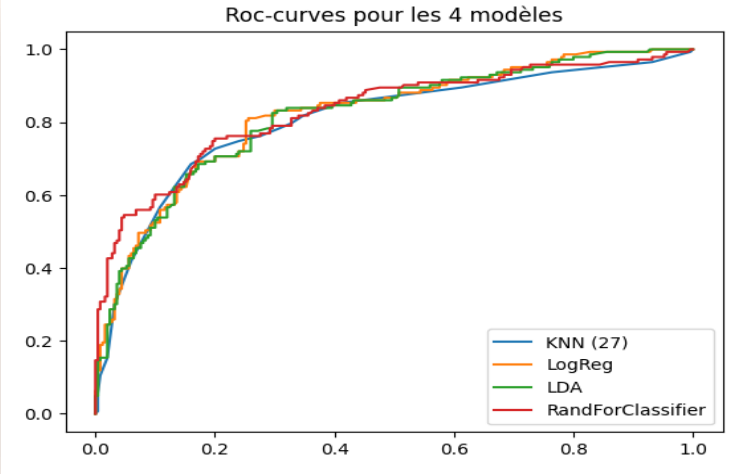
Modèles de régression
KNN
Régression logistique
Analyse discriminante linéaire
Random Forest

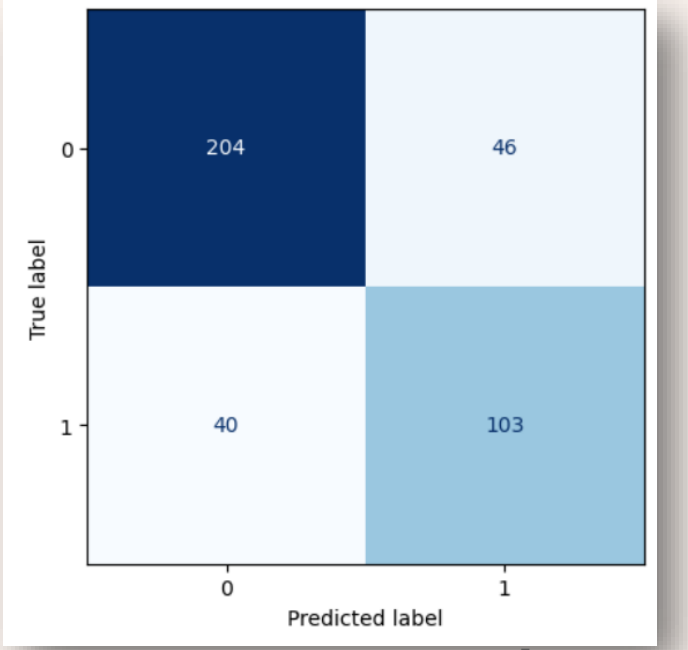
Evaluation/Métriques utilisées
Accuracy et F1-Score
Courbes ROC
Matrice de confusion

RÉSULTATS - COMPARAISON DES MODÈLES

Index	Modèle	Accuracy	AUC	Error_train	Error_test	F1-score (classe0)	F1-score (classe1)
0	KNN	0.78	0.82	0.206	0.224	0.822	0.699
1	Régression logistique	0.76	0.82	2.216	0.242	0.81	0.68
2	LDA	0.77	0.82	0.209	0.234	0.81	0.68
3	RandomForest	0.78	0.84	0.179	0.219	0.83	0.71

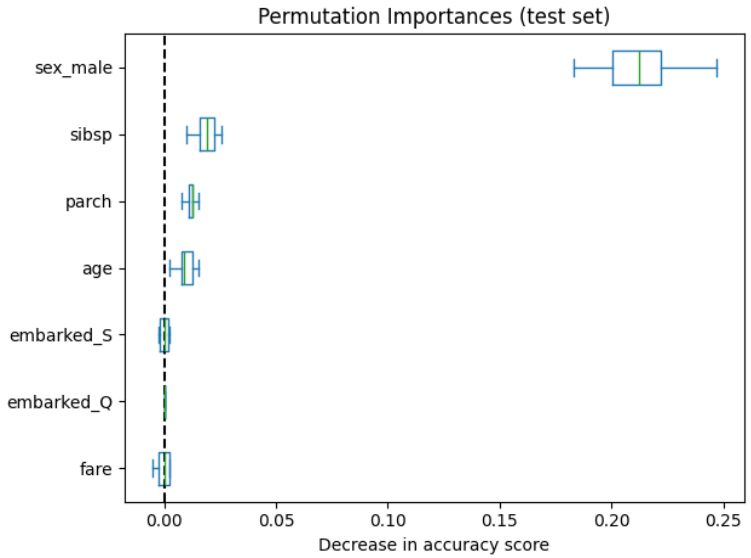




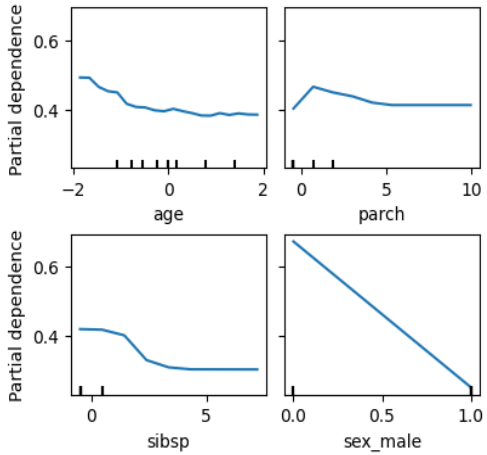


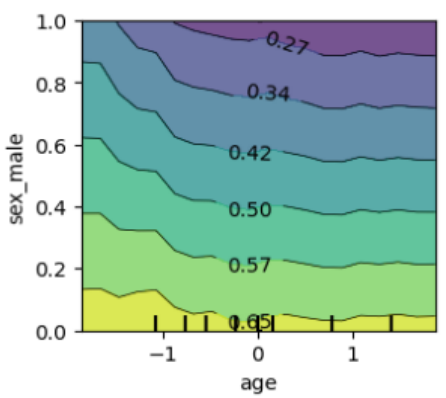
CONCLUSION ET AMELIORATIONS

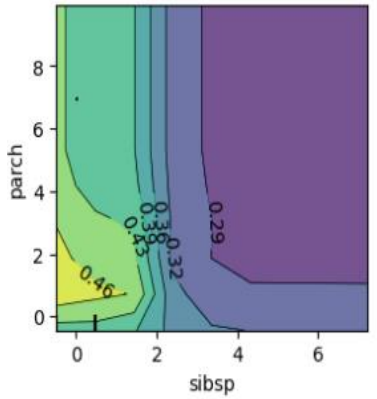
IMPORTANCE DES VARIABLES PAR PERMUTATION



GRAPHES DE DEPENDANCE PARTIELLE DES VARIABLES







AMELIORATIONS A ENVISAGER

- APPROFONDIR LES DISPARITES DANS LES CLASSES DESEQUILIBREES

=> 62% DE NON-SURVIVANTS ET 32% DE SURVIVANTS

=> DES POIDS DE CLASSES class_weight='balnced'

- UTILISER DES COURNES PRECISION-RECALL
- TESTER D'AUTRES MODELES
- TESTER D'AUTRES OPTIMISATIONS DE PARAMETRES DES MODELES