

Exercice 8 : Étude de modélisation avec MARS et réseaux de neurones

8a)

Figure 62: Model coefficients

Model coefficients (BostonHousing in 2023.MTH8302-Devoirs-data (2))
NOTE: Highlighted cells indicate basis functions of type
max(0, independent-knot), otherwise max(0, knot-independent)
Include condition: GROUP="M"

Coefficients	Y_MV	Knots	Knots	Knots	Knots	Knots	Knots	Knots	Knots	Knots	Knots	Knots
		X1_CRIM	X2_NOX	X3_AGE	X4_DIS	X5_RM	X6_LSTAT	X7_RAD	X8_CHAS	X9_INDUS	X10_TAX	X11_PT
Intercept	21.977											
Term 1	-0.553						6.72000					
Term 2	11.591					6.434000						
Term 3	-2.552					6.434000						
Term 4	-196.290		0.631000			6.434000						
Term 5	0.016						6.72000				304.0000	
Term 6	0.025						6.72000				304.0000	
Term 7	-0.136	15.28800										
Term 8	0.162	15.28800										
Term 9	7.606				1.858900	6.434000						
Term 10	-1.038		0.614000				6.72000					
Term 11	2.517		0.614000				6.72000					
Term 12	-0.174						6.72000					19.20000
Term 13	0.688					6.434000	26.82000					
Term 14	-0.490				3.262800							
Term 15	1.977				3.262800							

Figure 63: Model summary

Model specifications	Value
Independents	11
Dependents	1
Number of terms	16
Number of basis functions	23
Order of interactions	2
Penalty	2,000000
Threshold	0,000500
GCV error	12,36635
Prune	Yes

Figure 64 : Regression Statistics

Regression statistics	Y_MV
Mean (observed)	22,64568
Standard deviation (observed)	9,31405
Mean (predicted)	22,64568
Standard deviation (predicted)	8,72810
Mean (residual)	-0,00000
Standard deviation (residual)	3,25143
R-square	0,87814
R-square adjusted	0,87311

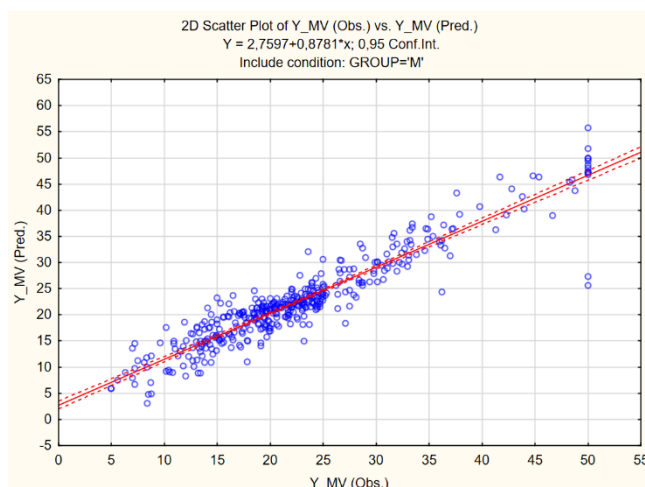
La Figure 62 (Model coefficients) présente les coefficients significatifs du modèle MARS. À l'aide des valeurs des coefficients significatifs, nous pouvons identifier les nœuds et comprendre l'importance des différentes variables dans le modèle. Grâce aux valeurs des coefficients de ce tableau, on peut constater que la variable X6_LSTAT est la plus importante car elle est référencée 7 fois dans le tableau. Cela suggère que cette variable joue un rôle clé dans la prédiction de la valeur médiane des maisons (Y_MV). La Figure 63 (Model summary) fournit la valeur du GVC et montre que le modèle MARS tient compte des termes d'interaction et effectue un élagage (pruning) pour éviter la surcomplexité du modèle. Le GVC de 12,36 suggère une performance prédictive globale acceptable. Enfin, Dans la Figure 64 (Regression Statistics), on peut observer un coefficient de détermination (R2) de 0,88, ce qui indique que le modèle MARS explique 88% de la variance des données. Cela montre que le modèle a une performance relativement bonne pour prédire la valeur médiane des maisons en fonction des 11 indicateurs de qualité de vie.

Nous pouvons établir l'équation suivante du modèle.

Equation du modèle :

$$Y_MV = 21.977 - 0.553\max(0, X6_LSTAT-6.72) + 11.591\max(0, X5_RM-6.434) - 2.552\max(0, 6.434-X5_RM) - 196.290\max(0, X2_NOX-0.631)\max(0, X5_RM-6.434) + 0.016\max(0, 6.72-X6_LSTAT)\max(0, X10_TAX-304) + 0.025\max(0, 6.72-X6_LSTAT)\max(0, 304-X10_TAX) - 0.136\max(0, X1_CRIM-15.288) + 0.162\max(0, 15.288-X1_CRIM) + 7.606\max(0, 1.859-X4_DIS)\max(0, 6.434-X5_RM) - 1.038\max(0, X2_NOX-0.614)\max(0, X6_LSTAT-6.72) + 2.517\max(0, 0.614-X2_NOX)\max(0, X6_LSTAT-6.72) - 0.174\max(0, X6_LSTAT-6.72)\max(0, X11_PT-19.2) + 0.688\max(0, 6.434-X5_RM)\max(0, X6_LSTAT-26.82) - 0.490\max(0, X4_DIS-3.263) + 1.977*\max(0, 3.263-X4_DIS)$$

Figure 65: ScatterPlot Y_MV(Obs) vs Y_MV(Pred)



La Figure 65 (ScatterPlot Y_MV(Obs) vs Y_MV(Pred)) montre l'ajustement du modèle en comparant les valeurs prédites (Y_Pred) et les valeurs observées (Y_Obs) de la valeur médiane des maisons. On peut constater qu'il y a quelques données aberrantes qui se situent en dehors de l'intervalle de confiance. Cela peut indiquer la présence de certaines observations pour lesquelles le modèle MARS n'a pas été en mesure de faire des prédictions précises. Cependant, malgré la présence de quelques données aberrantes, il semble que globalement, le modèle MARS présente un bon ajustement de Y_MV. En effet, comme mentionné précédemment, le modèle MARS présente un coefficient de détermination (R2) de 0,878 et un R2 ajusté de 0,873. Ces valeurs élevées confirment que

le modèle est capable d'expliquer une grande partie de la variance des données et qu'il possède une bon ajustement.

Enfin, en examinant les graphiques qui identifient les nœuds, nous pouvons mieux comprendre les relations non linéaires entre les variables indépendantes et la variable dépendante.

Graphique des nœuds au modèle

Figure 66: X1

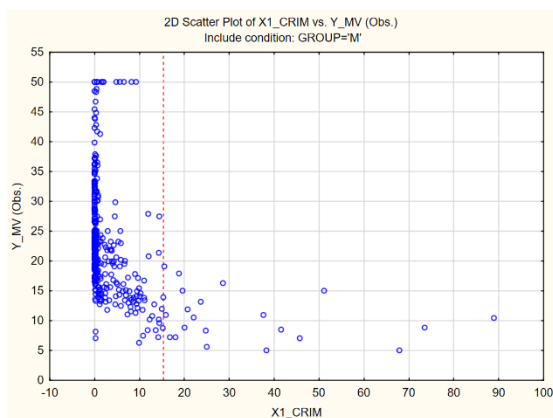


Figure 67 : X2

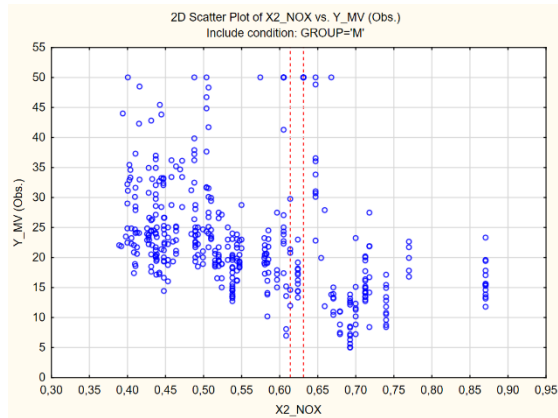


Figure 68: X4

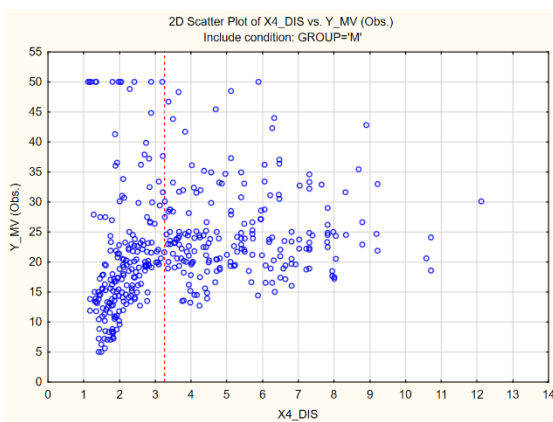


Figure 69: X5

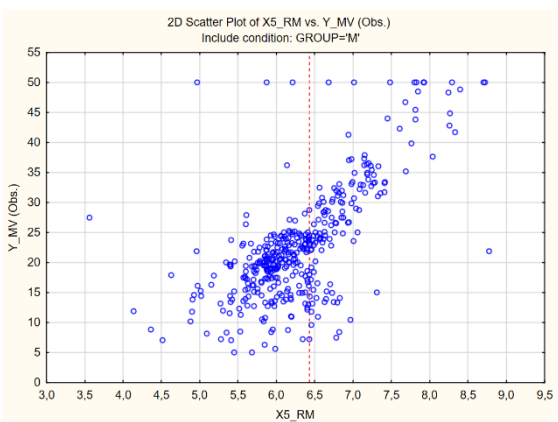


Figure 70 : X6

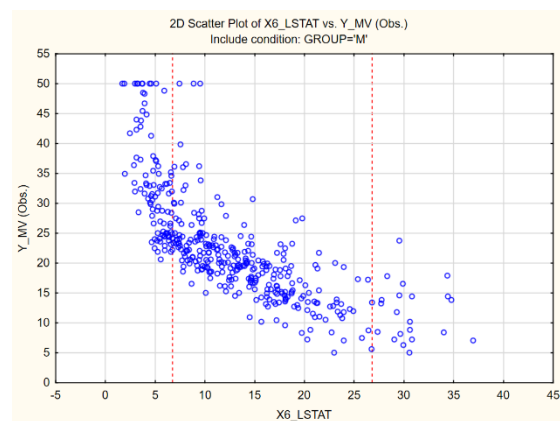


Figure 71: X10

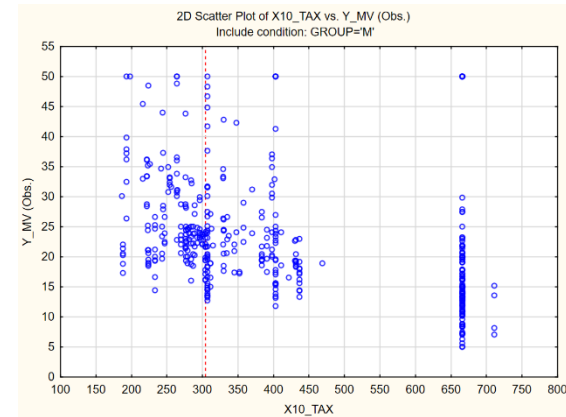
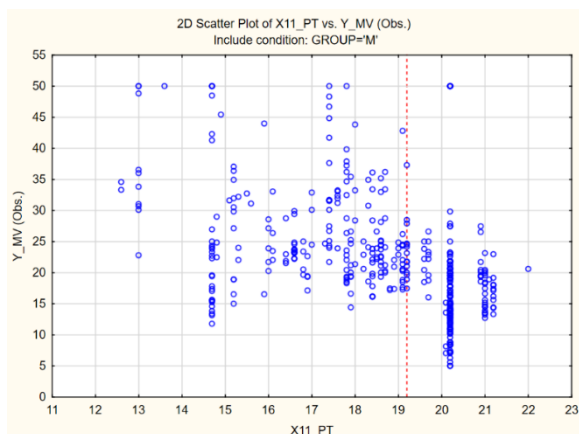


Figure 72 : X11

Lorsqu'une variable a plus de nœuds (ou coefficients significatifs) dans un modèle MARS, cela signifie généralement que la relation entre cette variable et la variable réponse (Y_MV) est plus complexe et non linéaire. Le fait d'avoir plus de nœuds permet au modèle de mieux capturer ces relations non linéaires et d'améliorer la précision des prédictions. En revanche, il faut aussi veiller à ne pas introduire trop de complexité dans le modèle, car cela peut conduire à un surajustement. Ainsi, dans notre cas, la présence de plusieurs nœuds pour la variable X6_LSTAT indique que le modèle MARS capture avec succès les relations complexes entre cette variable et la valeur médiane des maisons (Y_MV).

8b)

On a retenue les deux meilleurs reseaux de neurones suivant:

Figure 73 : Summary of active networks

Summary of active networks (BostonHousing in 2023-MTH8302-Devoirs-data (2))

Include condition: GROUP=M'

Index	Net. name	Training perf.	Test perf.	Validation perf.	Training error	Test error	Validation error	Training algorithm	Error function	Hidden activation	Output activation
1	MLP 11-5-1	0,960339	0,961256	0,917960	3,305116	3,893407	6,303604	BFGS 54	SOS	Logistic	Exponential
2	MLP 11-9-1	0,935732	0,937157	0,895034	5,375874	6,211639	7,836235	BFGS 23	SOS	Exponential	Tanh

Figure 74: Sensitivity Analysis

Sensitivity analysis (BostonHousing in 2023-MTH8302-Devoirs-data (2))

Samples: Train

Include condition: GROUP=M'

Networks	X5_RM	X6_LSTAT	X4_DIS	X7_RAD	X2_NOX	X10_TAX	X3_AGE	X11_PT	X1_CRIM	X9_NDUS	X8_CHAS
1.MLP 11-5-1	5,041721	4,083146	2,751858	1,854776	1,927073	1,596801	1,644785	1,405797	1,308549	1,099789	1,145373
2.MLP 11-9-1	2,548216	3,197287	1,730690	1,785809	1,677607	1,678760	1,347506	1,333907	1,143797	1,150650	1,051476
Average	3,794969	3,640217	2,241274	1,820293	1,802340	1,637781	1,496145	1,369852	1,226173	1,125219	1,098424

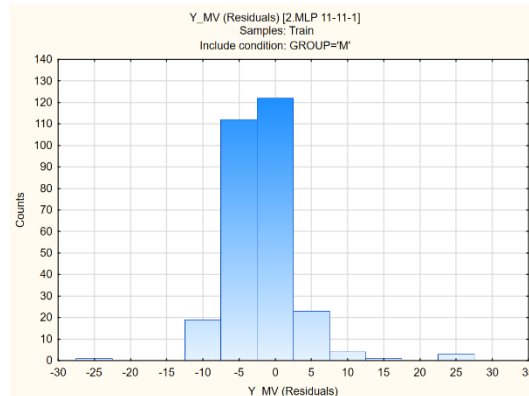
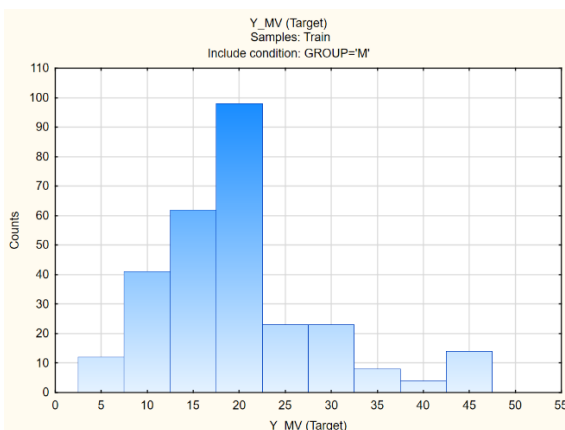
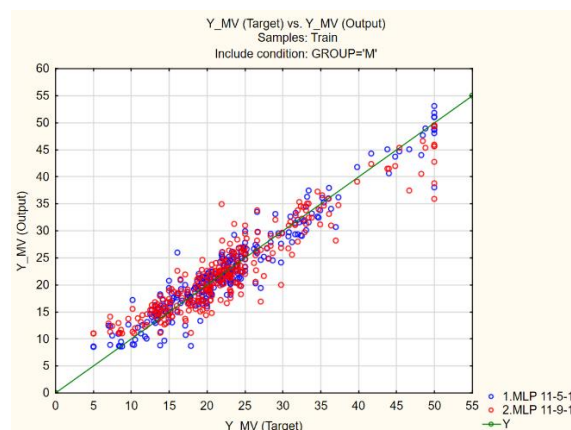
La figure 73 présente un résumé des différents réseaux de neurones activés, incluant les informations sur le type d'algorithme utilisé, leurs paramètres et leurs performances. On remarque que le premier réseau de neurones, MLP 11-5-1, présente une meilleure performance globale, avec un taux de 96% pour les performances d'entraînement et de test, comparativement à 93,5% pour le deuxième réseau de neurones. De plus, l'erreur est significativement plus faible dans le premier réseau de neurones pour les activités d'entraînement et de test. Ainsi, le réseau MLP 11-5-1 semble être un choix plus performant pour cette étude en raison de ses meilleures performances et de ses erreurs plus faibles par rapport au deuxième réseau de neurones. L'analyse de sensibilité, présentée dans la Figure 74, montre l'importance relative des variables d'entrée pour chaque réseau de neurones. Cette figure permet de comprendre l'influence des différentes variables sur la prédiction de la valeur médiane des maisons (Y_MV) et de sélectionner les réseaux de neurones qui capturent les relations les plus pertinentes entre les variables d'entrée et la variable cible. Le tableau montre que la variable X5_RM a la valeur la plus élevée parmi toutes les variables d'entrée pour les deux réseaux de neurones retenus. Cela indique que X5 a une importance plus élevée dans la prédiction (Y_MV). Le fait que X5 soit une variable clé dans les deux réseaux de neurones suggère que cette variable joue un rôle crucial dans la détermination de la valeur des maisons et que les réseaux de neurones sont capables de capturer cette relation de manière efficace.

Figure 75: Predictions Statistics

Predictions statistics (BostonHous)		
Target: Y_MV		
Include condition: GROUP='M'		
Statistics	1.MLP 11-5-1	2.MLP 11-9-1
Minimum prediction (Train)	8,51245	11,0062
Maximum prediction (Train)	53,06027	49,5335
Minimum prediction (Test)	8,27722	10,6786
Maximum prediction (Test)	54,14231	47,7513
Minimum prediction (Validation)	8,85026	11,0715
Maximum prediction (Validation)	51,41650	44,1153
Minimum prediction (Missing)		
Maximum prediction (Missing)		
Minimum residual (Train)	-9,84315	-13,0195
Maximum residual (Train)	11,95854	14,0822
Minimum residual (Test)	-5,14179	-11,0577
Maximum residual (Test)	10,60036	9,3383
Minimum residual (Validation)	-8,50362	-13,1782
Maximum residual (Validation)	12,69822	11,3715
Minimum standard residual (Train)	-5,41428	-5,6152
Maximum standard residual (Train)	6,57786	6,0736
Minimum standard residual (Test)	-2,60585	-4,4367
Maximum standard residual (Test)	5,37224	3,7469
Minimum standard residual (Validation)	-3,38695	-4,7076
Maximum standard residual (Validation)	5,05764	4,0622

Figure 76 : Network Weights

Network weights (BostonHousing in 2023-MTH8302-Devoirs-data (2))				
Include condition: GROUP='M'				
Weight ID	Connections		Connections	
	1.MLP 11-5-1	Weight values	2.MLP 11-9-1	Weight values
1	X1_CRIM -> hidden neuron 1	0,60429	X1_CRIM -> hidden neuron 1	-2,27091
2	X2_NOX -> hidden neuron 1	-1,49844	X2_NOX -> hidden neuron 1	-1,18033
3	X3_AGE -> hidden neuron 1	1,88160	X3_AGE -> hidden neuron 1	-0,10331
4	X4_DIS -> hidden neuron 1	-1,66835	X4_DIS -> hidden neuron 1	-0,45985
5	X5_RM -> hidden neuron 1	-4,29634	X5_RM -> hidden neuron 1	2,52188
6	X6_LSTAT -> hidden neuron 1	-0,27167	X6_LSTAT -> hidden neuron 1	-1,99153

Figure 77: Histogramme des résidus**Figure 78: Histogramme****Figure 79: Y_MV(Target) vs Y_MV(Output)**

La Figure 77 présente un histogramme des résidus pour les deux réseaux de neurones retenus. On observe que les résidus sont distribués autour de 0, ce qui indique que les erreurs de prédiction sont réparties de manière homogène. Cela suggère également une bonne constance de la variance et un bon ajustement.

La Figure 79 illustre la relation entre les valeurs observées de Y_MV (Target) et les valeurs prédites de Y_MV (Output) pour les deux réseaux de neurones sélectionnés. On constate que les deux réseaux de neurones présentent des performances de prédiction très similaires et correspondent à un bon ajustement. Les points sont proches de la ligne diagonale, ce qui signifie que les valeurs prédites sont en accord avec les valeurs réelles de Y_MV.

8c) MODELE T

Figure 80 : Model coefficients

Coefficients, knots and basis functions											
	Coefficients	Knots	Knots	Knots	Knots	Knots	Knots	Knots	Knots	Knots	Knots
	Y_MV	X1_CRIM	X2_NOX	X3_AGE	X4_DIS	X5_RM	X6_LSTAT	X7_RAD	X8_CHAS	X9_INDUS	X10_TAX
Intercept	25,40372										
Term.1	-0,79551						6,360000				
Term.2	11,52364					6,431000	6,360000				
Term.3	-2,85796					6,431000					
Term.4	-0,93411										
Term.5	-2,23975					6,750000					
Term.6	0,05722	7,672020					6,360000				
Term.7	9,97471				2,021800	6,431000					
Term.8	-0,18696				3,102500		6,360000				
Term.9	0,03571										
Term.10	0,01113									287,0000	
Term.11	0,00777			95,40000			6,360000				
Term.12										287,0000	19,20000

Figure 81: Model summary

Model specifications	Value
Independents	11
Dependents	1
Number of terms	13
Number of basis functions	18
Order of interactions	2
Penalty	2,000000
Threshold	0,000500
GCV error	7,510701
Prune	Yes

Figure 82 : Regression statistics

Regression statistics	Y_MV
Mean (observed)	22,08020
Standard deviation (observed)	8,74232
Mean (predicted)	22,08020
Standard deviation (predicted)	8,49311
Mean (residual)	-0,00000
Standard deviation (residual)	2,07249
R-square	0,94380
R-square adjusted	0,93540

La Figure 80 présente les coefficients du modèle de régression MARS sur l'ensemble de test T. On observe que X6_LSTAT est la variable la plus fréquemment utilisée dans le modèle, suivie des variables X4, X5, X6, X10 et X11, qui sont toutes considérées comme significatives. La Figure 81 montre le résumé du modèle MARS, où l'on constate que le critère de validation croisée généralisé (GCV) est de 7.15, ce qui est une amélioration par rapport au modèle précédent. Cela suggère que ce modèle MARS est plus performant en termes de précision des prédictions. La Figure 82 présente les statistiques de régression pour le modèle MARS sur l'ensemble de test T. Le coefficient de détermination R2 est de 0.943, ce qui est plus élevé que précédemment. Cela indique que le modèle MARS est capable d'expliquer 94,3% de la variance dans les données, ce qui est une performance satisfaisante.

Equation du modèle T :

$$Y_MV = 25.404 - 0.796\max(0, X6_LSTAT-6.36) + 1.423\max(0, 6.36-X6_LSTAT) + 11.534\max(0, X5_RM-6.431) - 2.858\max(0, 6.431-X5_RM) - 0.934\max(0, X11_PT-15.2) - 2.24\max(0, X5_RM-6.75)\max(0, X11_PT-15.2) + 0.057\max(0, 7.67202-X1_CRIM)\max(0, X6_LSTAT-6.36) + 9.875\max(0, 2.0218-X4_DIS)\max(0, 6.431-X5_RM) - 0.187\max(0, X4_DIS-3.1025)\max(0, X6_LSTAT-6.36) + 0.036\max(0, 287-X10_TAX) + 0.011*\max(0, 95.4-X3_AGE)\max(0, X6_LSTAT-6.36) + 0.008\max(0, X10_TAX-287)*\max(0, X11_PT-19.2)$$

Figure 83 : ScatterPlot Y MV(Obs) Vs Y MV(Pred)

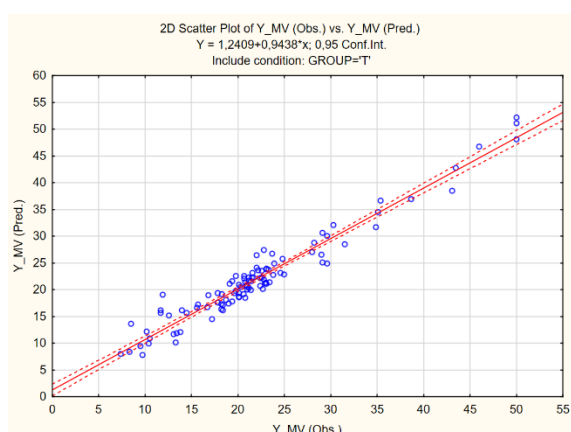


Figure 84 : Histogramme des residus

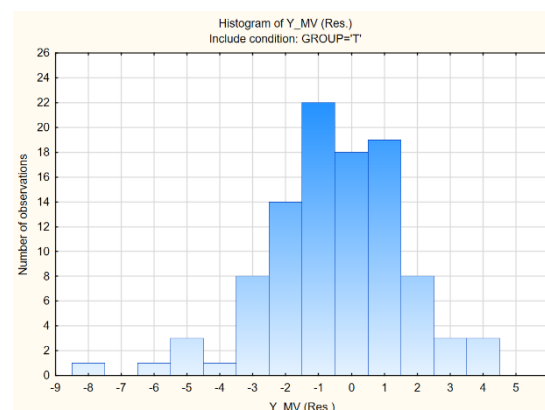


Figure 85 : Summary of active networks

Summary of active networks (BostonHousing in 2023-MTH8302-Devoirs-data (1))
Include condition: GROUP='T'

Index	Net. name	Training perf.	Test perf.	Validation perf.	Training error	Test error	Validation error
1	MLP 11-8-1	0.985207	0.654081	0.946256	1,172001	4,462624	5,628586
2	MLP 11-14-1	0.969986	0.671434	0.942993	2,357336	3,860944	5,975040

Figure 86: Sensitivity Analysis

Sensitivity analysis (BostonHousing in 2023-MTH8302-Devoirs-data (1))
Samples: Train
Include condition: GROUP='T'

Networks	X5 RM	X1 CRIM	X6 LSTAT	X7 RAD	X11 PT	X3 AGE
1.MLP 11-8-1	18,75888	5,265644	3,963826	3,713668	3,345396	3,912921
2.MLP 11-14-1	6,66960	3,586912	2,662306	2,021253	2,386858	1,318117
Average	12,71424	4,426278	3,313066	2,867460	2,866127	2,615519

Figure 87: Y MV(Target) vs Y MV(Output)

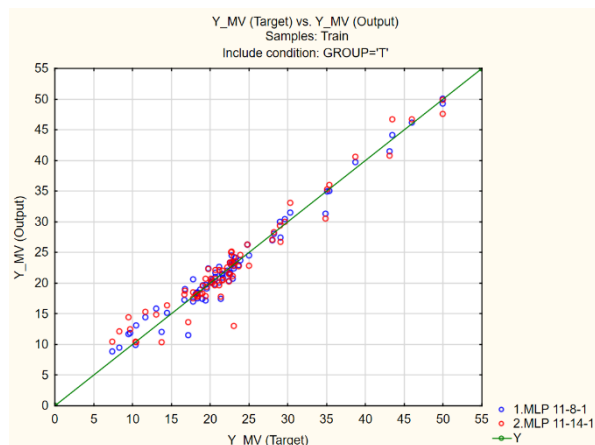
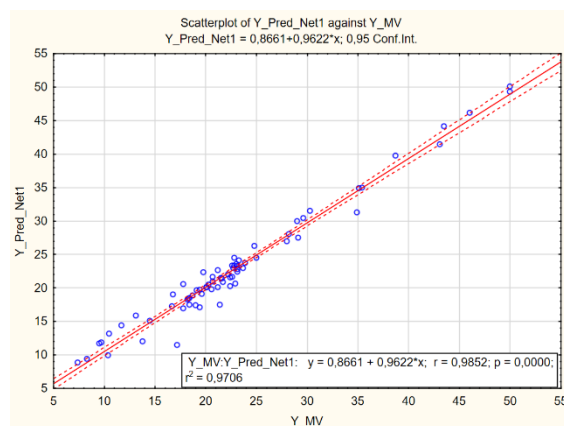


Figure 88: Y MV Network1 vs Y MV



Dans la Figure 85, le MSE (Mean Squared Error) correspond à l'erreur de test (Test Error) pour les réseaux de neurones.

Tableau de comparaison

Modele	R ²	R ² adjusted	MSE	SS_resid	AICc
Modele 8a	0.943	0.935	4.25	429.5	457.7
Modele 8b	0,97	0,97	5.6	166.42	571.5

SS Res calculer pur le premier reseaux de neurones

	0,04005	0,20012	71
SS_Res	166,42415	-0,16996	

Le modèle 8b (réseau de neurones) a un R² et un R² ajusté légèrement supérieurs à ceux du modèle 8a (MARS), ce qui suggère qu'il explique une plus grande partie de la variance dans les données. Cependant, le modèle 8b a également un MSE plus élevé et un SS_res inférieur, indiquant que les erreurs de prédiction sont plus importantes pour ce modèle que pour le modèle 8a. Le modèle 8a a un AICc plus faible que le modèle 8b, ce qui suggère que le modèle 8a est un meilleur compromis entre la complexité du modèle et la capacité à expliquer les données. En d'autres termes, le modèle MARS pourrait être préféré au réseau de neurones pour ce problème particulier, car il offre une meilleure performance en termes de critère d'information et de précision des prédictions. Le AICc est plus élevé que pour le modèle 8b. Cela est principalement dû au fait que le réseau de neurones nécessite un plus grand nombre de paramètres pour être ajusté, ce qui entraîne une complexité accrue du modèle.

8d)

Figure 89: Tableau de comparaison des 3 modèles

Modele	R ²	R ² adjusted	MSE	SS_resid	AICc
Modele 8a Mars	0.943	0.935	4.25	429.5	457.7
Modele 8b ANN	0,97	0,97	5.6	166.42	571.5
Modele 5e MRF	0,72	0,713	16.5	1522	566.8

En comparant les trois modèles, il est évident que le modèle MARS (Modèle 8a) est le meilleur en termes de performances par rapport au modèle MRF (Modèle 5e) et au réseau de neurones (Modèle 8b). Le modèle MARS présente des valeurs de R² et de R² ajusté plus élevées, ainsi qu'une erreur quadratique moyenne (MSE) et une somme des carrés des résidus (SS_res) plus faibles par rapport au modèle MRF. Bien que le réseau de neurones ait un R² et un R² ajusté légèrement plus élevés que le modèle MARS, il présente un AICc plus élevé, ce qui indique une moins bonne performance globale en tenant compte du nombre de paramètres. Ainsi, en prenant en compte les différents critères de performance, le modèle MARS est le meilleur choix parmi les trois modèles étudiés.

8e) Tableau des forces et faiblesses

Méthode	Forces	Faiblesses
MARS	<ul style="list-style-type: none"> -Ce modèle permet d'améliorer la précision en utilisant les techniques de backward et forward stepwise pour sélectionner uniquement les variables significatives. -Il permet d'introduire des effets d'interaction sans surajustement. -Il est capable de prendre en compte différents types de variables d'entrée et de sortie, qu'elles soient catégoriques ou continues. -Il est adapté pour traiter de grands échantillons de données. 	<ul style="list-style-type: none"> -Les données aberrantes (outliers) peuvent avoir un impact négatif sur les performances prédictives du modèle. -Il peut être limité dans sa capacité à capturer des relations complexes et non-linéaires entre les variables, en particulier lorsqu'il y a des interactions de haut ordre. -Il peut être coûteux en termes de temps et de ressources de calcul, en particulier pour les grands ensembles de données avec un grand nombre de variables d'entrée et des interactions complexes.
ANN	<ul style="list-style-type: none"> -Il est capable d'identifier les relations non-linéaires complexes et les interactions entre les variables, ce qui en fait un outil puissant pour la modélisation de relations complexes. -Il permet de régler facilement les paramètres pour optimiser la précision du modèle. - Il est capable de gérer de grandes quantités de données, ce qui le rend utile pour les données de grande taille. - Il peut compenser si certaines données sont manquantes ce qui est très pratique et montre sa capacité adaptative. 	<ul style="list-style-type: none"> -Il peut être sensible au (overfitting) si le modèle est trop complexe ou s'il n'y a pas assez de données d'entraînement ou trop de données aberrantes, ce qui peut rendre les prédictions moins précises avec de nouvelles données. -Il peut nécessiter une grande quantité de données pour être fiable et performant, ce qui peut rendre l'apprentissage coûteux en termes de temps et de ressources de calcul. En raison de sa complexité, le modèle ANN peut être difficile à interpréter et à comprendre, ce qui peut rendre difficile son analyse. -Si le nombre de neurones et de couches trop grand, peut impliquer une surinterprétation des données.
MRO	<ul style="list-style-type: none"> -Le MRO est facile à utiliser et permet une analyse simplifiée des résultats avec une interprétation claire des résultats. -Il peut être utilisé pour des données continues et catégoriques. -Peut éliminer les variables non significatives du modèle pour améliorer la précision des prédictions. -Bon pour les modèles de régression linéaire. 	<ul style="list-style-type: none"> -Peut ne pas prendre en compte les interactions entre les variables explicatives, ce qui peut conduire à négliger certains effets d'interaction qui peuvent influencer la variable réponse. -Il peut être affecté par un problème de multicollinéarité causé par une forte corrélation entre les variables, ce qui peut affecter la précision des prédictions du modèle. -Il est également sensible aux données aberrantes, ce qui peut introduire un biais dans les prédictions. -Il ne permet pas de traiter les données manquantes de manière optimale et peut ne pas être adapté à un grand nombre de données.
MRF	<ul style="list-style-type: none"> -Permet de réduire la complexité du modèle en sélectionnant uniquement les variables importantes, éliminant ainsi les variables non influentes. -Facilite l'interprétation des résultats en utilisant des outils tels que le diagramme de Pareto et les tableaux ANOVA et de régression pour mettre en évidence l'influence des variables significatives sur la variable dépendante. - Démontre étape par étape ce que le modèle 	<ul style="list-style-type: none"> -Peut ne pas prendre en compte les interactions entre les variables, ce qui peut limiter sa capacité à expliquer les relations complexes entre les variables. -Il y a un risque d'éliminer des variables importantes pour la prédiction de Y, ce qui peut entraîner une perte d'informations importantes pour le modèle. - Il peut être sensible à l'overfitting si un grand nombre de variables est pris en

	effectue et les variables qu'il exclut du modèle. - -Très rapide et efficace dans la sélection des variables importantes.	compte dans le modèle, ce qui peut réduire sa capacité à généraliser et à fournir des prévisions précises avec de nouvelles données.
MRB	-Permet de réduire la complexité du modèle en sélectionnant uniquement les variables importantes, éliminant ainsi les variables non influentes. -Facilite l'interprétation des résultats en utilisant des outils tels que le diagramme de Pareto et les tableaux ANOVA et de régression pour mettre en évidence l'influence des variables significatives sur la variable dépendante. - Démontre étape par étape ce que le modèle effectue et les variables qu'il exclut du modèle. - -Très rapide et efficace dans la sélection des variables importantes.	-Peut ne pas prendre en compte les interactions entre les variables, ce qui peut limiter sa capacité à expliquer les relations complexes entre les variables. -Il y a un risque d'éliminer des variables importantes pour la prédiction de Y, ce qui peut entraîner une perte d'informations importantes pour le modèle. - Il peut être sensible à l'overfitting si un grand nombre de variables est pris en compte dans le modèle, ce qui peut réduire sa capacité à généraliser et à fournir des prévisions précises avec de nouvelles données.

8f)

Dans l'ensemble, les modèles de régression, y compris les réseaux de neurones, présentent des avantages et des inconvénients distincts. Le choix du modèle statistique de régression dépend de divers facteurs, tels que la quantité de données disponibles, la fiabilité des données, la complexité du modèle souhaité, les buts de modélisation spécifiques, ainsi que la résilience et la précision requises. Il est primordial de comprendre le contexte du problème pour interpréter et valider les résultats avec justesse.

Après avoir examiné les avantages et les inconvénients de chaque méthode de modélisation de régression, il est important de noter que chaque méthode a sa propre utilité en fonction de la nature des données et des objectifs de modélisation spécifiques. Par exemple, le modèle de régression MRO est facile à utiliser et permet une analyse simplifiée des résultats avec une interprétation claire des résultats, ce qui en fait une méthode utile pour les modèles de régression linéaire simples. Les méthodes de régression backward stepwise (MRB) et forward stepwise (MRF) ont des avantages similaires à la méthode de régression ordinaire (MRO) en termes de sélection de variables importantes et de facilitation de l'interprétation des résultats. Cependant, elles peuvent être plus adaptées à des données plus complexes et à un grand nombre de variables.

la méthode de régression MARS est une méthode efficace pour modéliser des relations non linéaires avec une grande précision, tout en étant adaptable vis-à-vis des données aberrantes et manquantes. En revanche, les réseaux de neurones sont très utiles pour modéliser des relations non linéaires complexes, mais leur complexité peut les rendre difficiles à interpréter et à comprendre. En conclusion, dans le contexte de cet exercice, la méthode MARS a été identifiée comme la méthode de modélisation la plus adaptée pour les données non linéaires et complexes. Cela est dû à sa capacité à identifier les relations non linéaires complexes et les interactions entre les variables, ce qui en fait un outil puissant pour la modélisation de relations complexes.