Compte Rendu :

Linear Regression

Parkinson Speech Dataset with Multiple Types of Sound Recordings Data Set

Encadré par :

Marie SZAFRANSKI

Réalisé par :

Toufik GUENANE

1-Introduction :

La maladie de Parkinson est une maladie une maladie neurodégénérative qui peut affecter la parole de plusieurs manières. De nombreuses personnes atteintes de cette maladie parlent doucement et d'un seul ton ; ils ne transmettent pas beaucoup d'émotions. Parfois, la parole semble soufflée ou rauque. Les personnes atteintes de la maladie de Parkinson peuvent injurier des mots, marmonner ou s'arrêter à la fin d'une phrase. La plupart des gens parlent lentement, mais certains parlent rapidement, même en bégayant ou sans bégayer. Mais, afin de prévenir les risques de ces maladies et mieux se protéger, on doit comprendre et apprendre comment les détecter d'une manière efficace et sécurisé. Cette recherche vise à identifier les facteurs de risque de maladie parkinson les plus pertinents ainsi qu'à prédire le risque global à l'aide de Régression Linéaire (RL).

2-Analyse de Données : (à regarder l'annexeRL)

Figure 1 –Matrice de corrélation du dataset et Figure 2 – Valeurs de corrélation de UPDRS avec les variables indépendantes : les figures suivantes représentent en premier une matrice de corrélation des valeurs du dataset, et en deuxième on a une précision détaillée des valeurs de corrélation avec notre variable dépendante (Y = UPDRS). D'après notre matrice de corrélation on peut déduire directement que UPDRS n'est pas en forte corrélation avec quelconque des variables indépendantes. Et la deuxième figure le prouve, et cela à partir de la matrice de corrélation sur et de la projection sur notre label Y=UPDRS, on s'apperçoit que la variable la plus corélée est class information, mais dans notre cas on est pas entrain de faire une etude sur la classification mais c'est sur la regression, donc on vas retirer par la suite cette variable automatiquement. Donc y'a aucune corrélation de UPDRS avec les autres variables.

3-Exploration unidimensionnelle : (à regarder l'annexeRL)

Figure 3 – Exploration unidimensionnelle avant transformation : Cette figure represente les differentes distrubutions des varaibles, à partir de cette exploration unidimensionnelle, on peut voir que y'a plusieurs colonnes ou leur parition ne sont pas repartis en loi normal ce qui nous invite à gerer çà en transformant nos données utilisant SQRT et LOG.

Figure 4 – Exploration unidimensionnelle après transformation : Cette figure représente la nouvelle redistribution unidimensionnelle des variables après les transformations exécutes.

4-Linear Regression :

La régression linéaire est le modèle d'apprentissage automatique supervisé dans lequel le modèle trouve la ligne linéaire la mieux ajustée entre la variable indépendante et dépendante, c'est-à-dire qu'il trouve la relation linéaire entre la variable dépendante et indépendante. La régression linéaire est de deux types : simple et multiple. La régression linéaire simple est là où une seule variable indépendante est présente et le modèle doit trouver la relation linéaire de celle-ci avec la variable dépendante. Alors que, dans la régression linéaire multiple, il existe plusieurs variables indépendantes pour que le modèle trouve la relation.

4-1 Régression Linéaire Simple :

La régression linéaire simple est un modèle de régression linéaire avec une seule variable explicative, qui concerne des points d'échantillonnage bidimensionnels avec une variable indépendante et une variable dépendante.

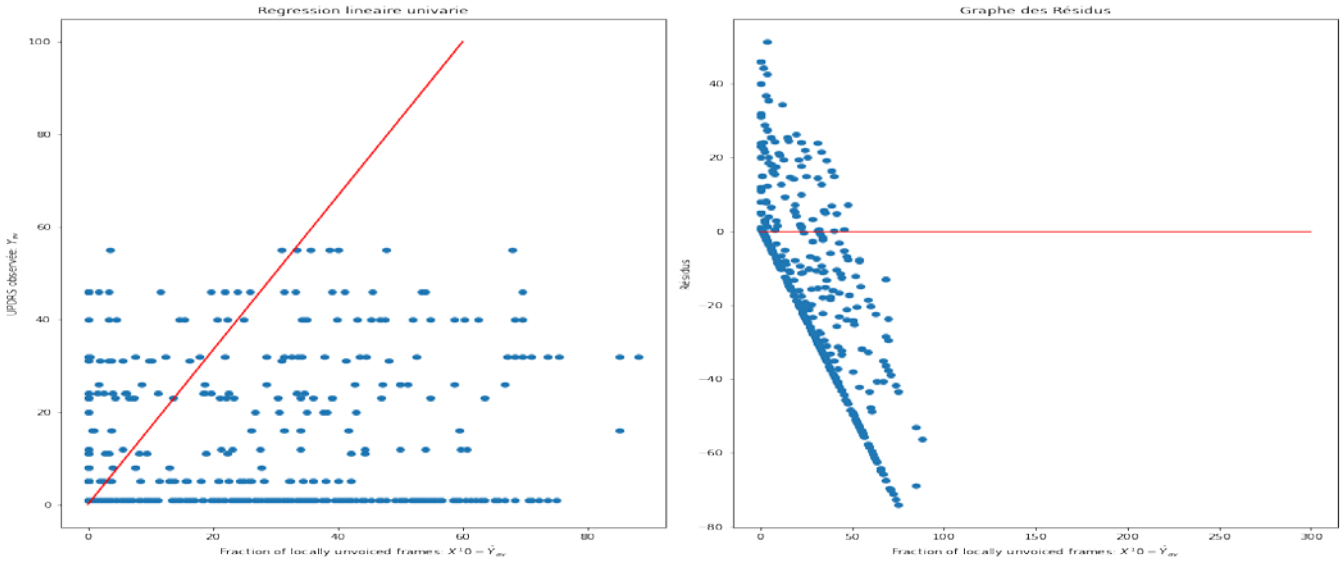


Figure 1 –Regression lineaire univarie X10 et UPDRS Observe avec le Graphe des Résidus–

- On peut voirque le R carre dans cette etude est négatif qui indique que le nombre de predictions est trop petit, et que le model est absolument pas bon, on peu meme compter à l'oeil nu le nombre de predictions.

- Sur le deuxieme plot on peut voir que le niveau de residus est tres élevé dans trop de valeurs "Y_av - x_10" sont loin d'etre egale à 0, avec une difference qui arrice jusqu'à -70, ce qui fait l'augmentation des residus et aussi la baisse de R carré de Coefficient de détermination : -2.4077049293570143 et Erreur quadratique : 859.09

4-2 Régression Linéaire Ordinaire :

Il s'agit d'ajuster un nuage de points, selon une relation linéaire, prenant la forme de la relation matricielle Y=XB+ep ou ep

est un terme d'erreur.

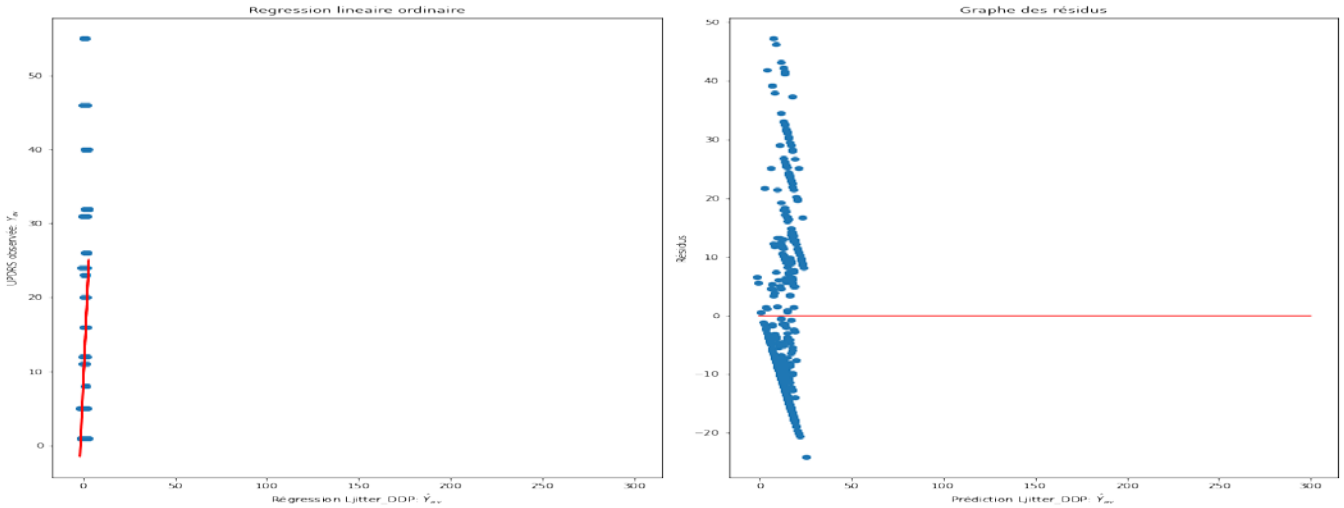


Figure 2 – Regression lineaire ordinaire de et UPDRS Observe avec le Graphe des Résidus

-On peut voir que le R carre dans cette etude est trop petit 0.07 (Coefficient de détermination : 0.07), indiquant aisni que le model est pas bon et manque de précision, on peu meme compter à l'oeil nu le nombre de predictions.

- Sur le deuxieme plot on peut voir que le niveau de residus est tres élevé dans trop de valeurs "Y_av - Y_hat_av" sont loin d'etre egale à 0, avec une difference qui arrice jusqu'à 48, ce qui fait l'augmentation des residus et aussi la baisse de R carré (Erreur quadratique : 233.63)

4-3 Régression Linéaire Multivaluée:

Comme le modèle de régression linéaire simple, on définit le modèle de régression linéaire multiple comme tout modèle de régression linéaire avec au moins deux variables explicatives.

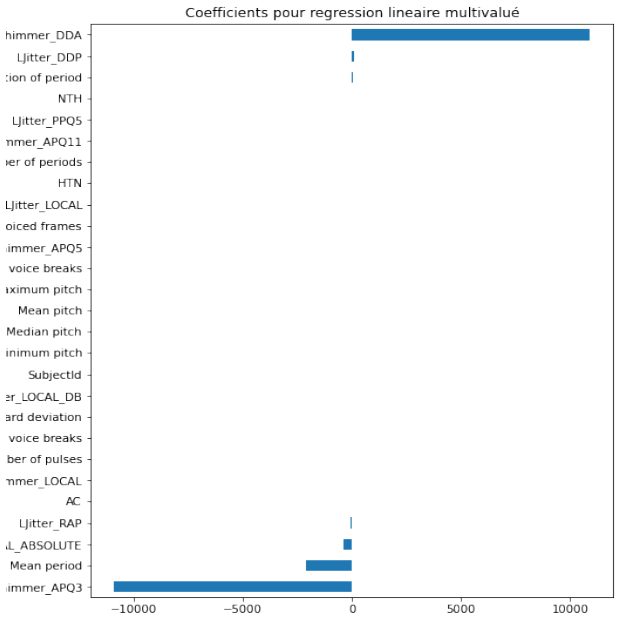


Figure 3 – Coefficients de la régression linéaire multivalue

A partir de ce plot, on aperçoit que la variable la plus impactante dans cette regression lineaire multivalué sur ce dataset est celle ci: LShimmer_DDA. Et la valeur la moins impactante est LShimmer_APQ3

```
Coffecients de regression [[-7.95817184e-01 -1.15433437e+00 -1.49887570e+01  1.51676906e+01
   1.29060727e+00 -2.42221065e-02 -2.29100561e-02 -2.80142836e-02
   1.00254067e-02 -2.12723141e+03  1.29074119e-01  4.59167508e-02
   7.97737723e-01 -3.58737628e+02 -7.86356968e+01  6.37863286e+00
   7.79247688e+01 -6.13817952e+00 -1.08874083e+04  1.11861207e-01
   4.50365650e+00  1.08921588e+04 -2.14650978e+00 -3.71029533e+00
   3.62519898e+00  3.04400897e+01 -2.67703376e+00]]

**********Y_av **************************************
Erreur quadratique (learning set) Y_av: 141.37
Coefficient de détermination (learning set) de Y_av: 0.44
************************************************************

**********Y_t **************************************
Erreur quadratique (testing set) Y_t: 148.07
Coefficient de détermination (testing set) de Y_t: 0.39
************************************************************
```

A partir de ces resultats on peu voir que coeffecient de determination de training set est egale à 0.44 et il est superieur à celui de test set, et que l'erreur quadaratique de training set est inferieur à celui de test set, et leurs valeurs ne sont pas loins non plus. (modele pas vraiment satisfaisant pour la prediction avec 44% de coef de determination)

4-4 Régression Linéaire Pénalisé : RIDGE :

La régression ridge consiste à minimiser le critère des moindres carrés pénalisé par la norme 2 des coefficients. Parmis les comportements attendu du Ridge on a: Si λ (alpha de Ridge) =0, on retrouve le biais et la variance de l'estimateur des Moindres Carrés Ordinaire (y'aura pas d'estimateurs Ridge). si λ "grand" alors $\beta_R \approx 0$ (estimateurs de Ridge seront nulles) Si λ est positif(augemente) ⇒ le biais augemente et la variance diminue & et réciproquement lorsque λ negatif(diminue). Et on trouve ça sur cette exemple:

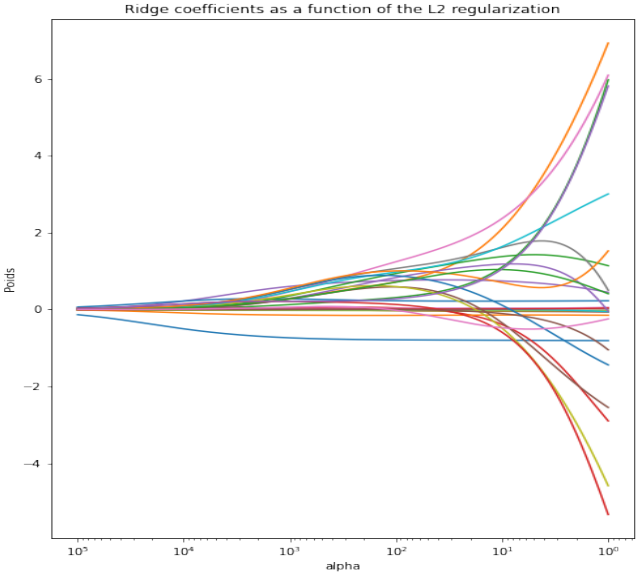


Figure 4 – Variation des coefficients de Ridge Linear Regression par rapport aux valeurs d'alpha

On voit que plus alpha est grand plus les poids des coefficients convergent vers zéro, et quand alpha est petit les poids divergent de zéro. Et cela nous invite à chercher le mielleur alpha et cela en utilisant **RidgeCV qui est une Ridge regression construite à partir de cross-validation.**

On a eu ces resultats avec RidgeCV :

```
Le score 'R carre' de la cross validation de Ridge Linear Regression model 0.4201095415442607
Les coeffecients du model RidgeCV Linear Regression sont: [[-7.74150930e-01  2.60538329e-01 -9.68
   8.70354731e-01 -1.05351777e-02  2.58398174e-02 -1.46942722e-02
   7.31566357e-03 -4.09662929e-02  1.47296648e-01  2.79194932e-03
  -1.00228949e+00 -1.81177888e-02  5.85734456e-01  4.45007830e+00
   5.83893556e-01 -1.69612932e+00  4.92603388e-01  4.56792559e-01
   2.93459988e+00  4.95408001e-01 -9.78715782e-01 -1.81612476e+00
   1.84740529e+00 -9.13823110e-02 -2.11024367e+00]]

Le parametre alpha optimal pour la regression ridge est alpha =  10.0
L'Erreur quadratique (learning set) : 146.19
```

On a eu un coeffecient quadratique de 0.42 et erreur quadratique de 146.19 pour un alpha optimal de RidgeCV de 10, afin de verifier l' éfficacité de modele on a ploté les Y originales de UPDRS et Y predite par ce modele de alpha=10.

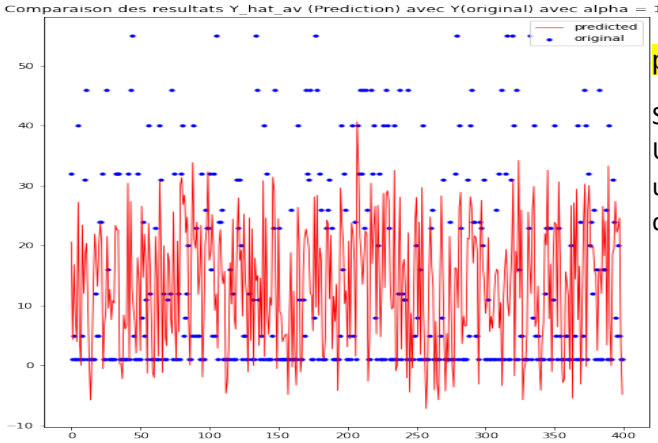


Figure 5 – Ridge Comparasion des valeurs existantes et valeurs predites de UPDRS

Sur le plot suivant on peut voir la comparaison entre les valeurs initiales de UPDRS celles predites et celles non predites, ou chaque ligne rouge qui passe sur un point bleu indique une prediction et sinon cette valeur n'est pas predite. Et que la plupart des valeurs originales ne sont pas predites.

4-5 Régression Linéaire Pénalisé : LASSO :

La régression lasso consiste à minimiser le critère des moindres carrés pénalisé par la norme 1 des coefficients. Parmis les comportements attendu du Lasso on a: Si λ est positif(augemente) ⇒ le biais augemente et la variance diminue & et réciproquement lorsque λ negatif(diminue). Contrairement à ridge : λ est grand(augmente) ⇒ le nombre de coefficients nuls augmente,et c'est le cas si deux variables sont corrélées. L'une sera sélectionnée par le Lasso, l'autre supprimée. C'est aussi son avantage par rapport à une régression ridge qui ne fera pas de sélection de variables.

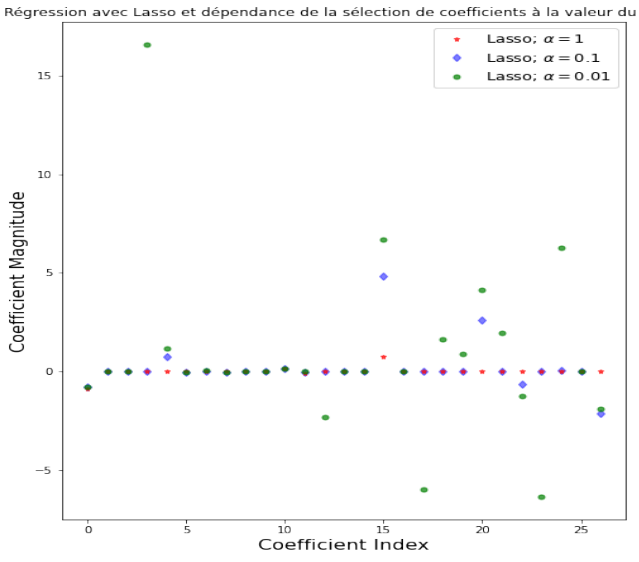


Figure 6 – Variation des coefficients de Lasso Linear Regression par rapport aux valeurs d'alpha

On voit que plus alpha est grand plus les poids des coefficients convergent vers zéro, et quand alpha est petit les poids divergent de zéro. Et cela nous invite à chercher le mielleur alpha et cela en utilisant **LassoCV qui est une Lasso regression construite à partir de cross-validation.**

On a eu ces resultats avec LassoCV :

```
Le score 'R carre' de la cross validation de Lasso Linear Regression model 0.41567981777976326
Les coeffecients du model LassoCV Linear Regression sont: [-7.78104198e-01  0.00000000e+00 -0.00000
  7.73245590e-01 -4.41737691e-03  1.09823169e-02 -4.36203528e-03
  5.13689481e-03 -0.00000000e+00  1.51281049e-01 -1.08064033e-02
 -0.00000000e+00 -0.00000000e+00  0.00000000e+00  4.82087002e+00
  0.00000000e+00 -0.00000000e+00  0.00000000e+00  0.00000000e+00
  2.60168459e+00  0.00000000e+00 -6.46421385e-01 -0.00000000e+00
  8.42661437e-02 -0.00000000e+00 -2.10626551e+00]

Le parametre alpha optimal pour la regression Lasso est alpha =  0.1
Erreur quadratique (learning set) : 147.31
```

On a eu un coeffecient quadratique de 0.415 et erreur quadratique de 147.31 pour un alpha optimal de LassoCV de 0.1, afin de verifier l' éfficacité de modele on a ploté les Y originales de UPDRS et Y predite par ce modele de alpha=0.1.

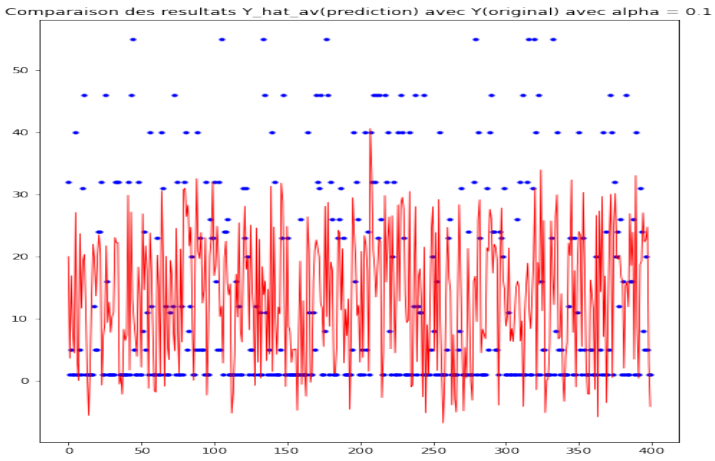
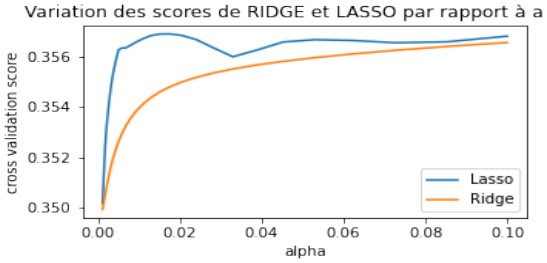


Figure 7 – Lasso Comparasion des valeurs existantes et valeurs predites de UPDRS

Sur le plot suivant on peut voir la comparaison entre les valeurs initiales de UPDRS celles predites et celles non predites, ou chaque ligne rouge qui passe sur un point bleu indique une prediction et sinon cette valeur n'est pas predite. Et que la plupart des valeurs originales ne sont pas predites, ce qui éxplliques la mauvise valeur de R carre et la grande valeaur de MSE.

4-6 Régression Linéaire Pénalisé : LASSO vs RIDGE with cross_val_score :

Le but de cette comparaison c'est de voir le comportement des deux modeles score de prediction tenant compte desdifferetes valeurs de alpha



A partir de ces résultats, dans ce dataset avec la technique de cross_val_score sur les deux modelés Ridge et Lasso, on voit que plus alpha tend vers zéro, la précision des deux modèles aussi converge vers zéro, mais en montant de zéro vers 0.1, la précision du modèle Lasso monte en exponentiel tant que la précision de modèle Ridge monte d'une façon moins exponentielle par rapport à Lasso. Et vers la valeur alpha = 0.1 converge ensemble, et Lasso garde toujours le dessus sur Ridge.

5. Avantages et les limites :

L'utilisation de la pénalisation de RidgeCV ou LassoCV est donne presque le meme résultat qu'une régression linéaire multivalue ordinaire et même le taux d'erreur est presque identique, bien que RidgeCV et LassoCV présentent de nombreux avantages, ils prennent trop de temps et d'efforts à régler parfaitement le modèle et à trouver le meilleur paramètre alpha pour maximiser le rendement du chaque modèle. Une meilleure utilisation du temps peut être d'étudier plus avant les caractéristiques du Linear Régression multivaluate modèle qu'on fait avec les paramètres nécessaires.

L'ingénierie de fonctionnalités et la sélection de sous-ensembles de fonctionnalités peuvent augmenter (ou diminuer) considérablement les performances de ce modèle. Cela demandera beaucoup plus d'efforts que de brancher des nombres dans une grille de paramètres, mais, en retour, on développera également davantage une compréhension de l'ensemble de données et on découvrira éventuellement de nouvelles relations entre les entités.

6. Conclusion

A partir des deux études effectuer avec RidgeCV et LassoCV, on peut conclure qu'il n y'a pas vraiment de meilleur modèle sur cette étude puisque tous les modèles de régression linéaire avec ou sans RidgeCV ou LassoCV, donnent presque un même coefficient de détermination de 44%(pour le modèle de régression linéaire sans aucune pénalisation qui est le meilleur en tous les cas) et une erreur quadratique qui ne dépasse pas 150 pour les trois modèles.

Contrairement aux algorithmes qu'on a déjà vu comme SVM et KNN, les modèles de régression ont une sortie qui est continue mais pas de sorties discrètes. Donc, on ne peut pas réellement dire qui est meilleur que l'autre car ils sont utilisés dans différents contextes, et ils sont complémentaires.