Université Mohammed premier

ENSAO

IDSCC-4

Modélisation d'indicateur PSA

Réalisé par :
Yahya SGHIOURI

Encadré par:
Pr.Abdelmounaim
Kerkri

May 20, 2023

Contents

1 Exploratory data analysis

- Les données de cette manipulation proviennent d'une étude menée par Stamey et al. (1989) qui examinait la corrélation entre le niveau d'antigène spécifique de la prostate (PSA) et plusieurs mesures cliniques chez 97 hommes sur le point de subir une prostatectomie radicale. Le PSA est une protéine produite par la glande de la prostate. Plus le niveau de PSA d'un homme est élevé, plus il est probable qu'il ait un cancer de la prostate.

L'objectif est de prédire le logarithme du PSA (lpsa) à partir de plusieurs mesures, notamment le logarithme du volume du cancer (lcavol), le logarithme du poids de la prostate (lweight), l'âge, le logarithme de la quantité d'hyperplasie prostatique bénigne (lbph), l'invasion des vésicules séminales (svi), le logarithme de la pénétration capsulaire (lcp), le score de Gleason (gleason) et le pourcentage de scores de Gleason 4 ou 5 (pgg45).

Les données sont modélisées en profondeur dans le livre intitulé "The Elements of Statistical Learning", qui est disponible gratuitement sur le site : " http://statweb.stanford.edu/ tibs/ElemStatLearn/ " .

1.1 Présentation des variables

On importe une base des données concernant la concentration de l'antigène spécifique de la prostate (PSA) .

1. *lcavol*: logarithme du volume du cancer, mesuré en millilitres (cc). La zone de cancer a été mesurée à partir d'images numérisées et multipliée par une épaisseur pour produire un volume.

2. *lweight*: logarithme du poids de la prostate, mesuré en grammes.

3. *age*: l'âge du patient, en années.

4. *lbph*: logarithme de la quantité d'hyperplasie prostatique bénigne (BPH), une hypertrophie non cancéreuse de la glande de la prostate, exprimée en cm^2 dans une image numérisée.

5. *svi*: invasion des vésicules séminales, un indicateur binaire (0/1) indiquant si les cellules cancéreuses de la prostate ont envahi les vésicules séminales.

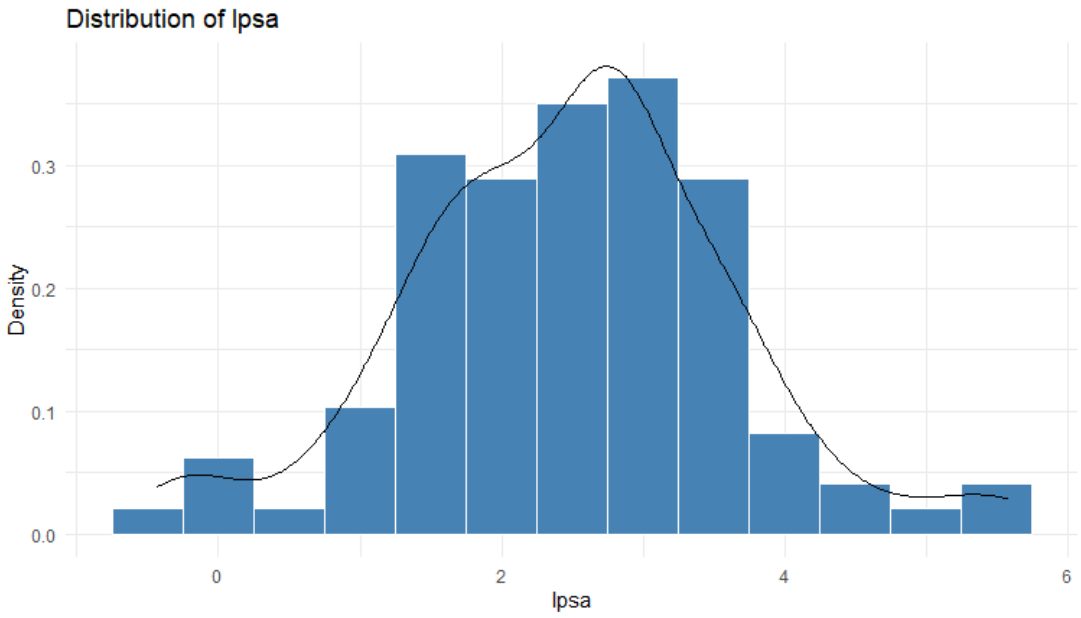
6. *lcp*: logarithme de la pénétration capsulaire, qui représente le niveau d'extension du cancer dans la capsule (le tissu fibreux qui agit comme revêtement externe de la glande de la prostate), mesuré comme l'étendue linéaire de la pénétration, en cm.

7. *gleason*: score de Gleason, une mesure du degré d'agressivité de la tumeur. Le système de classification de Gleason attribue une note (de 1 à 9) à chacune des deux plus grandes zones de cancer dans les échantillons de tissu, 1 étant le moins agressif et 9 le plus agressif ; les deux notes sont ensuite ajoutées pour obtenir le score de Gleason.

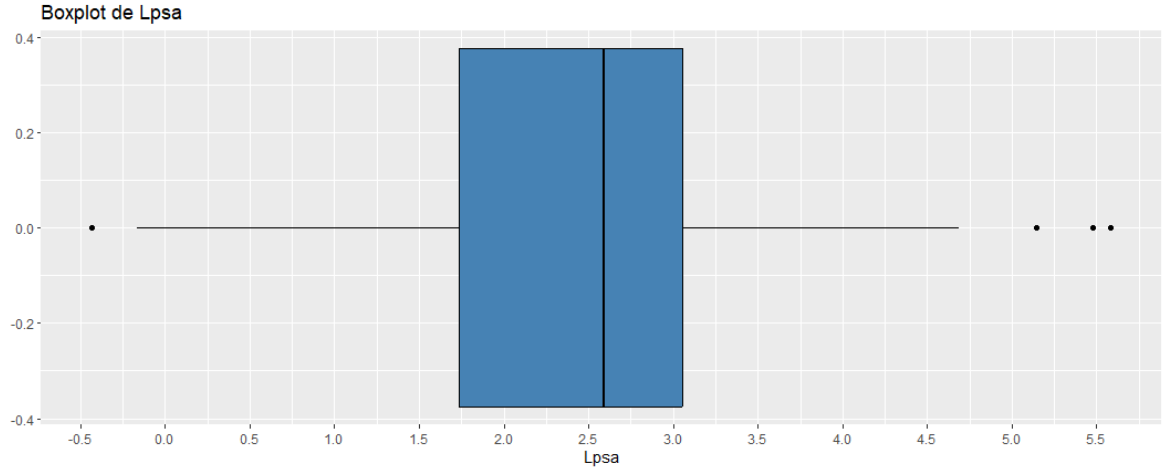
8. *pgg*45: pourcentage des scores de Gleason qui sont de 4 ou 5.

9. *lpsa*:(target variable) logarithme de l'antigène spécifique de la prostate (PSA), une concentration mesurée en ng/mL.

1.2 Distribution de variable de sortie



- la variable dépendante est proche d'une distribution normal , les résidus du modèle seront également susceptibles de suivre une distribution normale. Cela va nous aider à obtenir des estimations de paramètres plus précises et des intervalles de confiance plus fiables , et on a pas besoin d'effectuer des transformations sur cette dérnière .



- La variable Lpsa , contient des valeurs négatives par ce qu'on a pris le logarithme de concentration de PSA .
En plus cette variable contient des valeurs abérrantes , par exemple supérieures à 4.5 , ce sont des niveaux qu'on rencontrer , et ils sont des indicateurs sur la présence du cancer,les enlever ça va influencer la performance de notre modèle à modéliser des situations réelles .

-Plus une personne a un niveau important en lpsa plus il est plus suscéptible d'avoir la maladie , et il va devoir faire d'autre tests pour le confirmer .

variable	Min	Max	Mean	Median	$1^{er} quantile$	$3^{eme} quantile$
lpsa	-0.4308	5.5829	2.5915	2.4784	1.7317	3.0564

1.3 Variable Age

- Le risque de développer des problèmes de prostate, y compris le cancer de la prostate, augmente avec l'âge .

On remarque fréquemment des problèmes de prostate , chez les gens agés de 40 ans et plus , c'est pour cela que la plupart d'entre eux font des tests à cet période .

les niveaux de PSA considérés comme acceptables peuvent varier selon les différents groupes d'âge.les intervalles de PSA spécifiques à l'âge approximatives couramment utilisées comme lignes directrices :

Pour les hommes dans la quarantaine : Des niveaux de PSA jusqu'à 2,5 ng/mL peuvent être considérés comme normaux.
Pour les hommes dans la cinquantaine : Des niveaux de PSA jusqu'à 3,5 ng/mL peuvent être considérés comme normaux.

Pour les hommes dans la soixantaine : Des niveaux de PSA jusqu'à 4,5 ng/mL peuvent être considérés comme normaux.
Pour les hommes de plus de 70 ans : Des niveaux de PSA jusqu'à 6,5 ng/mL peuvent être considérés comme normaux.

- Le clustering de cette variable pourrait etre utile si nous étions dans un problème de classification par exemple du cancer de prostate .ce qui n'est pas le cas ici , puisque i lest utilisé pour diagnostiquer les patients conjointement avec le niveau lpsa .

- Distribution:

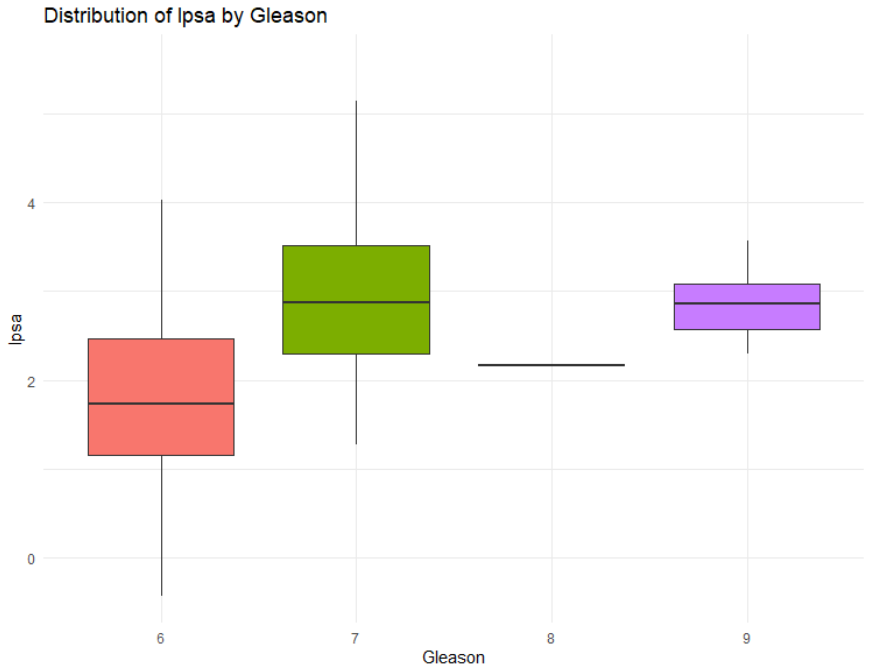
variable	Min	Max	Mean	Median	$1^{er} quantile$	$3^{eme} quantile$
Age	41.00	79.00	63.87	65.00	60.00	68.00

1.4 Variable gleason

- gleason s'agit du degré d'aggressivité de tumeur .

- c'est une variable catégorielle , voyons sa relation avec la variable de sortie .

Correlation entre "gleason" et "'lpsa" : 0.3689868



- On remarque une difference significative entre le niveau 7 et 6 .

- On va le confirmer avec "on way anova test" ,est ce que gleason a un effet significatif sur la variable lpsa .

- Puisque la $Pvalue = 1.21e-05 < 0.05$, on rejette l'hypothése qu'on a pas de différenes significatives dans entres les moyennes des groupes .au seuil de 5% ,ici on a utilisé une statistique de Fisher.

- En effectuant un test de tukey sur les differents catégories de gleason , nous remarquons une différence significative entre la catégorie 6 et 7 ,$pvalue = 0.0000042 < 0.05$

- Donc pour simplifier le modèle on va coder 6 par 1 , 7 par 2 et les autres par 3 , cela va nous permettre de réduire la variance du paramètre lié à la variable gleason .

- On remarque qu'aprés la corrélation entre les deux variables a augmenté pour atteindre : 0.42846269 , c'est un bon signe .

1.5 Correlation entre les variables

- il y a un problème de corrélation entre les régresseurs .

- On remarque qu'on a des correlations fortes , par exemple la variable "lcp" avec "lcavol" , la correlation entre les deux nous donne 0.675 . Ceci peut avoir un impact sur la performance du modèle et l'interprétabilité des coefficients .

- la variable la plus importante dans notre dataset est "lcavol" a la plus forte corrélation avec la variable de sortie .

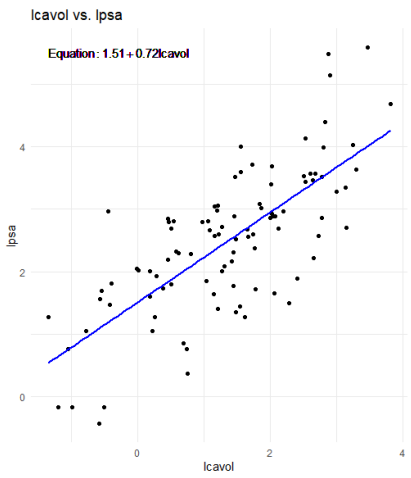


Figure 1: Corrélation entre lcavol et lpsa

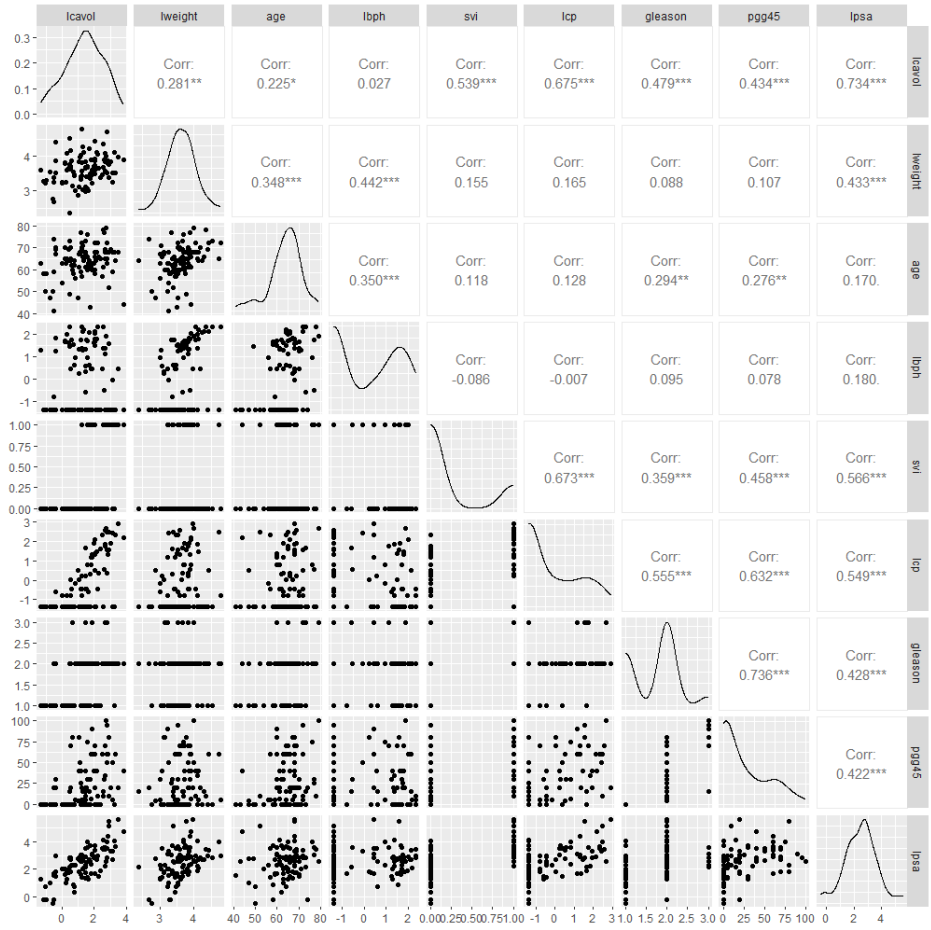


Figure 2: Corrélation entre les régresseurs

2 Multiple linear regression

-On va maintenant créer un modèle , de régression avec une les variables les plus significatives pour faire la comparaison entre ce modèle et les modèles multivariate (ridge ,Lasso,Elsatic-net,ACP)qu'on va établir aprés .

- La formule pour estimer les paramétres :$\hat{\beta} = (X^t X)^{-1} X^t y$

- nous allons utiliser la fonction lm en R pour faciliter la recherche des variables significatives pour la prédiction

Coefficients:	Estimate	Std.Error	t_value	P_value
Intercept	-0.007742	0.078892	-0.098	0.922136 ***
lcavol	0.530931	0.096403	5.507	7.16e-07 ***
lweight	0.273450	0.074802	3.656	0.000525 ***
svi	0.192917	0.097168	1.985	0.051458

- Mais avant tout on va faire un scaling aux données ,et puis nous allons diviser notre dataset , en des données de train et de test , nous avons 97 observations , 30 pour le test , et 67 pour le train .

- Aprés une recherche manuelle des meilleurs régresseurs , on trouve que la meilleure combinaison : "lcavol" "lweight" "svi" la combinaison de ces trois explique mieux la variable de sortie .

- On va utiliser le Rmse pour faire la comparaison entre les modèles .

- Rmse pour ce modèle OLS nous a donné : 0.5482622

3 Régression ridge

-Pour la régression Ridge nous avons un hyperparametre λ , qui est un terme de pénalité permettant de réduire la variance des paramétre en augmentant le biais ,(bias,variance tradeoff) .

- Sa formule $\hat{\beta} = (X^t X + \lambda I_{mm})^{-1} X^t y$,elle va nous permettre de réduire les effets de multicolinéarité entre les variables .

- Pour ceci nous allons construire une pile de valeurs pour chercher le lambda qui donne le Rmse le plus bas .

-On a utilisé une approche dichotomique , on part avec des valeurs dispersés , puis on retraicit notre espace de recherche sur des intervalles qui sont petits.

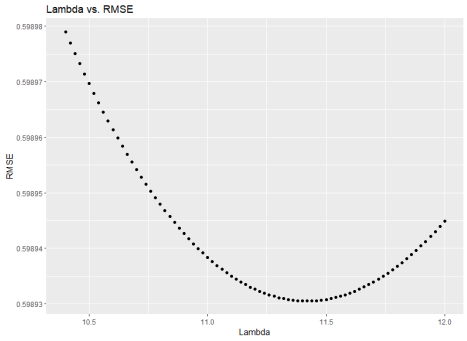


Figure 3: Rmse vs lambda

- RMSE ridge : 0.598931 - meilleur lambda : 11.42

4 Régression Lasso

- Nous allons adopter la meme méthode qu'on a utilisé dans ridge ,
- le lambda le plus optimal est dans le voisinage de la valeur 5 .

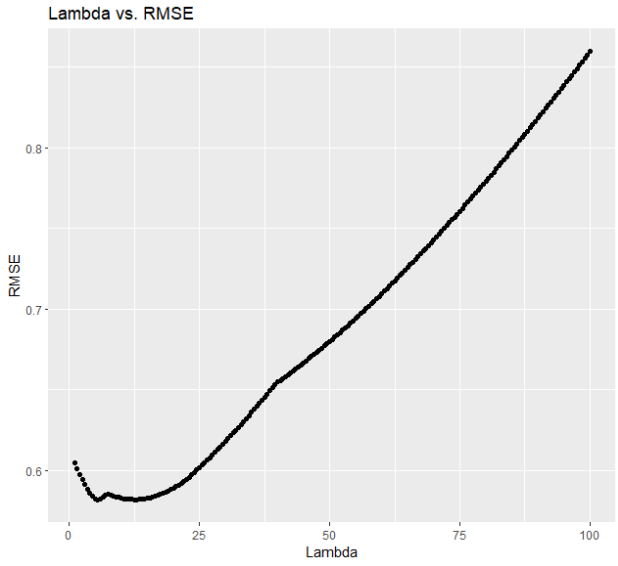


Figure 4: Rmse vs lambda lasso avant sélection

- La régression Lasso est trés utile dans la sélection des variables .

- Best lambda avant sélection $\lambda = 5.30$, $Rmse = 0.581267$
- Best lambda aprés selection $\lambda = 1$, $Rmse = 0.549158$
- La meilleure combinaison trouvé : ("lcavol","lweight","svi") .

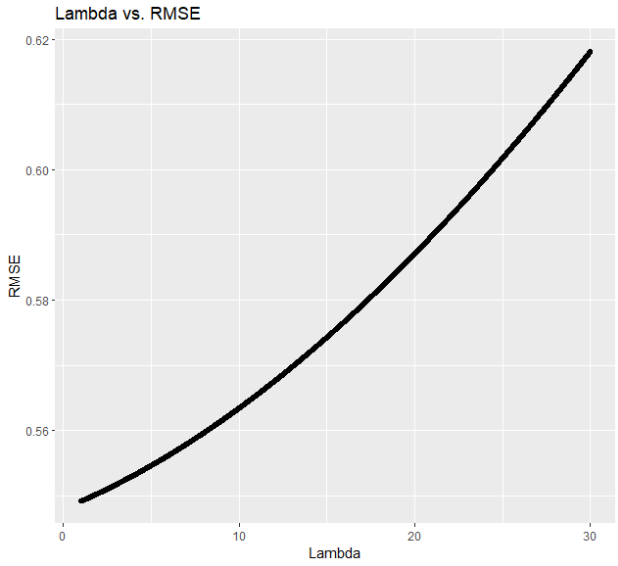


Figure 5: Rmse vs lambda lasso apres sélection

5 Régression Acp

La régression avec l'ACP peut être utile dans notre cas puisqu'on a de nombreuses variables indépendantes corrélées.ça va nous permettre de réduire la dimensionnalité tout en préservant les informations les plus importantes pour la prédiction de la variable dépendante.

- le meilleure RMSE obtenu , concerne 7 premieres composantes .
- Rmse obtenu : $Rmse_{pca} = 0.582156$

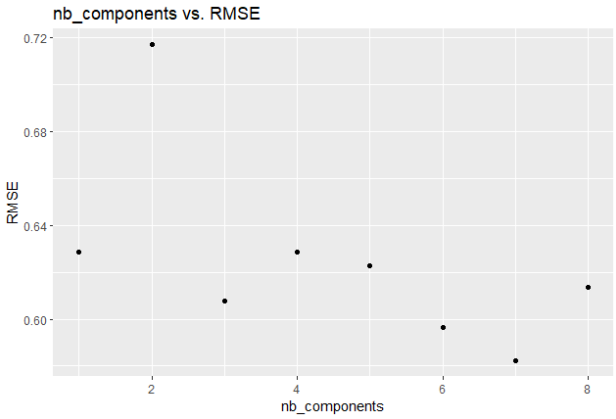


Figure 6: Rmse vs nombre de composantes

6 Régression Elastic net

- la régression Elastic Net peut gérer des situations où les variables sont fortement corrélées et effectuer une sélection de variables automatique. Il combine à la fois une pénalité de L1 (Lasso) et une pénalité de L2 (Ridge), ce qui permet de mettre à zéro les coefficients de régression pour certaines variables moins importantes.

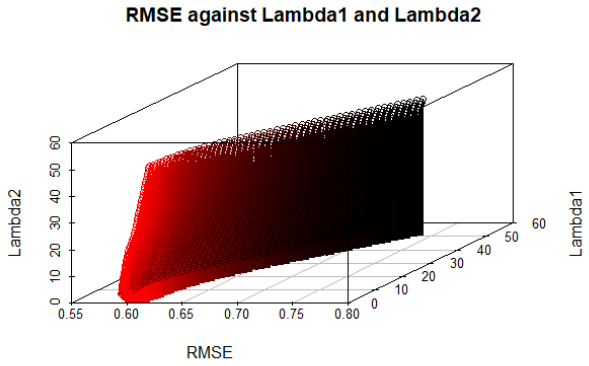


Figure 7: Rmse vs lambda1 et lambda 2

- Rmse obtenu : $Rmse_{net} = 0.582108$ - avec $\lambda 1 = 0.0040$ et $\lambda 2 = 0$

7 Comparaison des modèles

- Maintenant la partie comparaison des modèles , en ce qui concerne la performance prédictive .

Modèle:	Rmse sur les données de test
Ols	0.5482622
$Ridge$	0.598931
PCA	0.582156
$Elastic\_net$	0.582108
$Lasso$	0.549158

- On remarque bien que les modèle OLS et Lasso sont les modèles les plus performants en terme d'Rmse .

- Avoir une connaissance approfondie dans un domaine particulier est extrêmement bénéfique lors de la modélisation, car cela permet d'économiser du temps et de l'effort dans la création de multiples modèles. vous êtes en mesure de déterminer quelles variables sont pertinentes pour votre modèle. Vous pouvez identifier les facteurs clés qui influencent la variable cible et sélectionner les variables appropriées à inclure dans votre modèle.