UNIVERSITÉ DE LORRAINE

FST FACULTÉ DES SCIENCES ET TECHNOLOGIES Nancy

Projet de Statistique pour données de grande dimension

M2 IMSD 2023-2024

Sujet : Attrition des employés

MARNISSI Yassine

SOUMAHORO Moriba

Encadrante :

Mme Anne Gegout-Petit

Table des matières

INTRODUCTION :

Dans le cadre de ce projet, notre objectif est d'explorer les facteurs qui influent sur le départ des employés au sein d'une entreprise. En comprenant ces facteurs, nous ambitionnons de créer un modèle prédictif robuste qui sera en mesure d'anticiper les départs potentiels. La première étape de notre démarche consiste à effectuer un traitement des données. Dans la deuxième phase, nous mettrons en œuvre l'Analyse en Composantes Principales (ACP) afin de réduire la dimensionnalité des variables. Cette technique nous permettra de saisir les tendances fondamentales et d'identifier les variables qui contribuent le plus à la variabilité observée dans le départ des employés. À la suite de l'ACP, nous procéderons à une sélection des variables par différentes méthodes, ensuite nous évaluerons la performance des modèles créés. Enfin, après avoir évalué plusieurs modèles, nous choisirons le modèle le plus adapté pour prédire le départ des employés.

PARTIE 1 : TRAITEMENT DE DONNEES

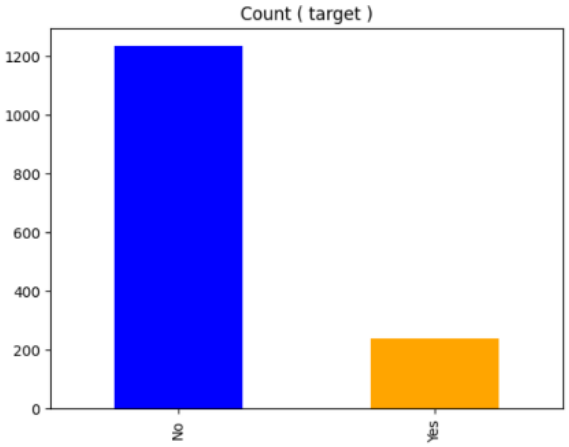
1.1-Visualisation du jeu de données

Notre jeu de données comporte 1470 lignes et 35 variables dont 15 sont quantitatives et 20 qualitatives. La variable à prédire ‘ Attrition ‘ est binaire.

Le pourcentage de chaque classe :

Classe 'Yes’ : 16.12%

Classe 'No’ : 83.88%



On remarque que notre jeu de données est déséquilibré.

Le déséquilibre dans les classes d’un jeu de données peut avoir plusieurs conséquences lors de la construction et de l'évaluation des modèles prédictifs. Par exemple, les modèles peuvent développer un biais en faveur de la classe majoritaire. Cela signifie qu’ils peuvent être moins sensibles à la classe minoritaire et avoir des performances inférieures pour prédire les instances de cette classe.

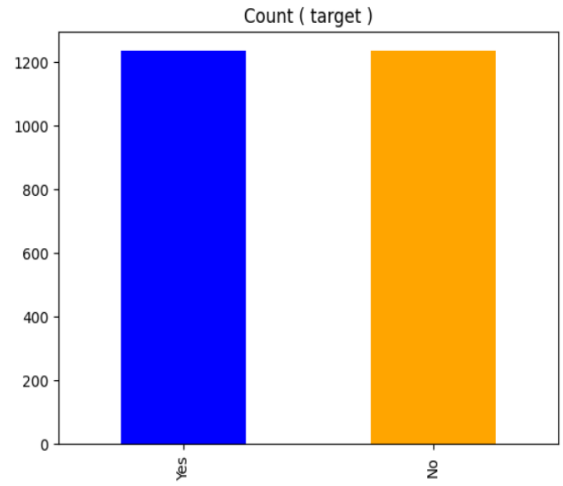
Pour éviter ce problème, nous effectuons un rééchantillonnage de notre jeu de données.

1.2-Rééchantillonnage du jeu de données

Les techniques pour équilibrer les classes sont le sous-échantillonnage et le suréchantillonnage. Nous utiliserons dans notre projet le suréchantillonnage qui consiste à équilibrer les classes en augmentant le nombre d'instances de la classe minoritaire. Cela permet au modèle de mieux apprendre les caractéristiques de la classe minoritaire et d'améliorer ses performances lors de la prédiction de cette classe.

Ci-dessous une illustration de la distribution des classes après rééchantillonnage.

```
After Random OverSampling :  Counter({'Yes': 1233, 'No': 1233})
<Axes: title={'center': 'Count ( target )'}>
```



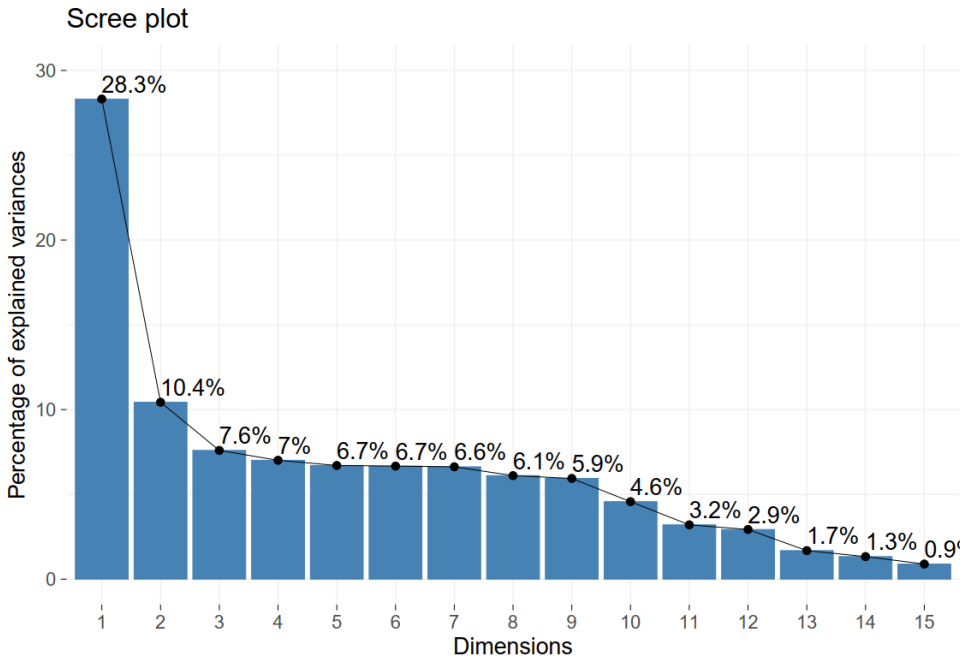
Tout au long du projet, nous utiliserons le jeu de données obtenu avec la méthode de suréchantillonnage.

PARTIE II : ANALYSE EN COMPOSANTE PRINCIPALE

Dans cette partie, nous réaliserons l'analyse en composante principale qui permet de décrire un jeu de données, de le résumer, d'en réduire la dimensionnalité.

2.1 Visualisation de la distribution de l'inertie des axes

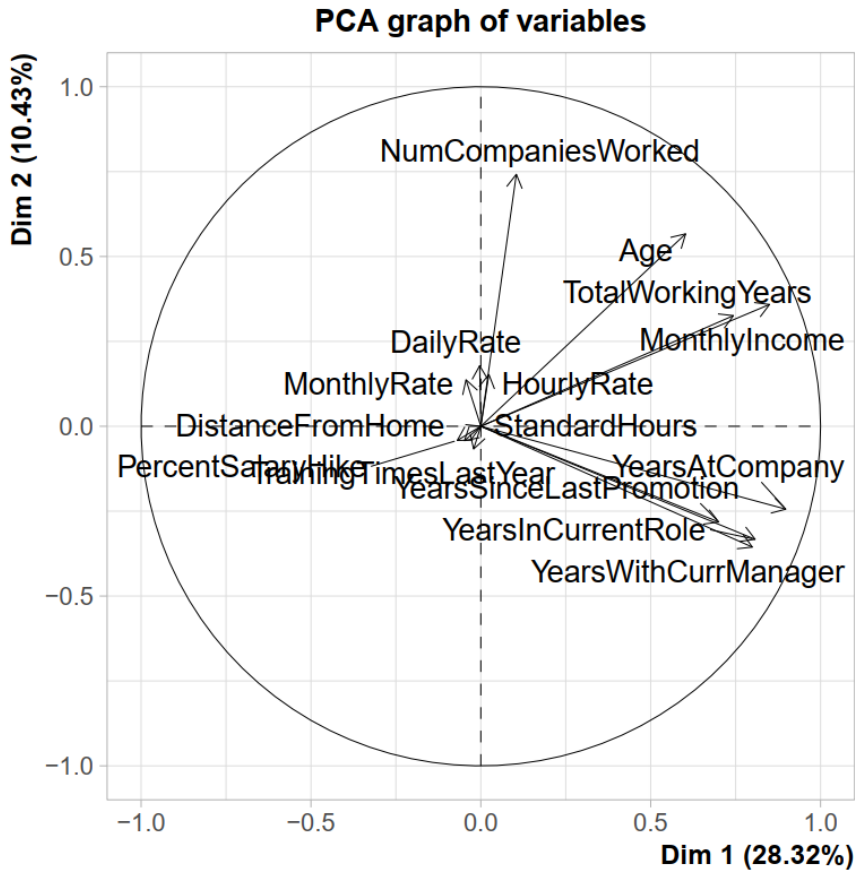
Après avoir effectué l'ACP, nous avons obtenu le graphe du pourcentage des variances expliquée pour chaque composante principale :



Les deux premières composantes contiennent 38 ,7 % de l'inertie totale. Nous nous limitons aux deux premières composantes principales car les autres ne captent pas beaucoup d'inertie totale.

2.2-Visualisation et description des composantes principales

Représentons maintenant le graphique des variables dans le premier plan factoriel :



"TotalWorkingYears","YearsAtCompany","YearsInCurrentRole","YearsSinceLastPromotion","YearsWithCurrManager" sont bien représentées positivement sur la première composante principale.

Elle semble capturer le parcours professionnel de l'employé dans une entreprise spécifique (expérience totale, la stabilité au sein de l'entreprise, la durée dans le rôle actuel, le temps depuis la dernière promotion, et la durée de la relation de travail avec le manager actuel) .

La variable ''NumCompaniesWorked '' est bien représentée sur la deuxième composante principale. Elle semble capturer le nombre total d'entreprises pour lesquelles l'employé a travaillé.

Nous aborderons maintenant la sélection de variables par différentes méthodes.

PARTIE 3 : SELECTION DE VARIABLES

Avant d'entamer l'ajustement de nos modèles de régression logistique, nous commencerons d'abord par tester la covariabilité entre notre variable explicative "Attrition" et les variables qualitatives d'une part, et d'autre part, la covariabilité entre "Attrition" et les variables quantitatives. Ensuite, nous ajusterons les p-values en utilisant la méthode de Benjamin-Hochberg. Enfin, nous

sélectionnerons les variables significatives ayant une p-value inférieure à 0.05 et réviserons notre jeu de données en conséquence.

Les variables pertinentes obtenue par la méthode de Benjamin Hochberg sont :

'Department', 'StockOptionLevel', 'YearsAtCompany', 'DistanceFromHome', 'YearsWithCurrManager', 'RelationshipSatisfaction', 'OverTime', 'BusinessTravel', 'TrainingTimesLastYear', 'DailyRate', 'EnvironmentSatisfaction', 'MonthlyIncome', 'JobSatisfaction', 'Age', 'JobRole', 'Attrition', 'TotalWorkingYears', 'NumCompaniesWorked', 'MaritalStatus', 'Education', 'EducationField', 'YearsSinceLastPromotion', 'JobInvolvement', 'JobLevel', 'YearsInCurrentRole', 'WorkLifeBalance'

Dans la suite, nous utiliserons les variables ci-dessus.

3.1-Sélection de variables par la méthode Forward et Backward

Dans cette partie, nous ajusterons les modèles de régression logistique en utilisant la méthode Forward. À chaque itération, nous calculerons la valeur de l' AIC, d'abord pour le modèle avec l'intercept, puis nous incrémenterons progressivement une variable de manière à minimiser la valeur de l'AIC. Ce processus se poursuivra jusqu'à ce que nous atteignions le modèle le plus performant.

Ci-dessous le début de l'itération.

```
## Start:  AIC=3426.41
## Attrition ~ 1
##
##                            Df Deviance    AIC
## + OverTime                  1   3150.4 3166.0
## + JobRole                   8   3096.0 3166.3
## + TotalWorkingYears         1   3238.3 3253.9
## + JobLevel                  1   3248.6 3264.2
## + MonthlyIncome             1   3259.1 3274.7
## + YearsInCurrentRole        1   3261.8 3277.4
## + Age                       1   3273.5 3289.1
## + YearsWithCurrManager      1   3277.4 3293.1
## + MaritalStatus             2   3270.0 3293.4
## + YearsAtCompany            1   3307.4 3323.0
## + StockOptionLevel          1   3318.2 3333.8
## + BusinessTravel            2   3326.2 3349.7
```

Après les itérations, le modèle obtenu est le suivant :

```
## Step:  AIC=2431.16
## Attrition ~ OverTime + JobRole + MaritalStatus + EnvironmentSatisfaction +
##     BusinessTravel + JobSatisfaction + JobInvolvement + YearsInCurrentRole +
##     YearsSinceLastPromotion + DistanceFromHome + NumCompaniesWorked +
##     Age + RelationshipSatisfaction + WorkLifeBalance + YearsWithCurrManager +
##     YearsAtCompany + TrainingTimesLastYear + TotalWorkingYears +
##     EducationField
##
##                     Df Deviance    AIC
## <none>                   2173.4 2431.2
## + DailyRate          1   2166.0 2431.6
## + Department         2   2163.0 2436.4
## + StockOptionLevel   1   2171.7 2437.2
## + MonthlyIncome      1   2171.9 2437.4
## + JobLevel           1   2172.6 2438.2
## + Education          1   2173.4 2439.0
```

Maintenant, nous ajusterons le modèle par régression logistique en utilisant la méthode Backward qui calcule à chaque fois l' AIC, d'abord pour le modèle complet, après elle exclut à chaque fois une

variable de telle sorte que la valeur de l'AIC soit minimale jusqu'à ce qu'elle arrive au modèle le plus performant.

Ci-dessous le début des itérations

```
## Start:  AIC=2464.71
## Attrition ~ Department + StockOptionLevel + YearsAtCompany +
##     DistanceFromHome + YearsWithCurrManager + RelationshipSatisfaction +
##     OverTime + BusinessTravel + TrainingTimesLastYear + DailyRate +
##     EnvironmentSatisfaction + MonthlyIncome + JobSatisfaction +
##     Age + JobRole + TotalWorkingYears + NumCompaniesWorked +
##     MaritalStatus + Education + EducationField + YearsSinceLastPromotion +
##     JobInvolvement + JobLevel + YearsInCurrentRole + WorkLifeBalance
##
##                            Df Deviance    AIC
## - Education                 1   2152.3 2456.9
## - JobLevel                  1   2152.3 2456.9
## - MonthlyIncome             1   2153.1 2457.7
## - StockOptionLevel          1   2154.1 2458.7
## - Department                2   2162.2 2458.9
## - DailyRate                 1   2160.0 2464.6
```

Après les itérations, nous avons obtenu le modèle suivant :

```
## Step:  AIC=2431.16
## Attrition ~ YearsAtCompany + DistanceFromHome + YearsWithCurrManager +
##     RelationshipSatisfaction + OverTime + BusinessTravel + TrainingTimesLastYear +
##     EnvironmentSatisfaction + JobSatisfaction + Age + JobRole +
##     TotalWorkingYears + NumCompaniesWorked + MaritalStatus +
##     EducationField + YearsSinceLastPromotion + JobInvolvement +
##     YearsInCurrentRole + WorkLifeBalance
##
##                           Df Deviance    AIC
## <none>                         2173.4 2431.2
## - EducationField           5   2213.0 2431.7
## - TotalWorkingYears        1   2184.0 2433.9
## - TrainingTimesLastYear    1   2184.8 2434.8
## - RelationshipSatisfaction 1   2189.0 2439.0
## - WorkLifeBalance          1   2189.2 2439.1
## - Age                      1   2190.6 2440.5
## - YearsWithCurrManager     1   2204.1 2454.0
## - JobInvolvement           1   2208.0 2457.9
## - YearsSinceLastPromotion  1   2208.7 2458.7
## - YearsAtCompany           1   2208.9 2458.8
## - YearsInCurrentRole       1   2209.2 2459.1
```

Nous constatons que les modèles sélectionnés par les deux méthodes sont équivalents .

3.2 : Sélection de variables avec régularisation Ridge, Lasso et Elastic net

Dans cette partie, nous effectuerons la sélection de variables en appliquant les régularisations Ridge, Lasso et Elastic net.

Après l' application de ces régularisations à nos modèles de régression logistique nous avons obtenu les coefficients suivants :

- Pour Ridge

	s1		
(Intercept)	3.405811e+00	JobRoleManager	-1.792735e-01
(Intercept)	.	JobRoleManufacturing Director	-2.289917e-01
DepartmentResearch & Development	-1.653038e-01	JobRoleResearch Director	-1.338924e+00
DepartmentSales	2.256828e-01	JobRoleResearch Scientist	-3.137731e-02
StockOptionLevel	-1.885001e-01	JobRoleSales Executive	4.779260e-02
YearsAtCompany	3.447334e-02	JobRoleSales Representative	8.024920e-01
DistanceFromHome	2.756382e-02	TotalWorkingYears	-2.088163e-02
YearsWithCurrManager	-7.281625e-02	NumCompaniesWorked	1.215076e-01
RelationshipSatisfaction	-1.362707e-01	MaritalStatusMarried	8.851531e-02
OverTimeYes	1.382271e+00	MaritalStatusSingle	6.755982e-01
BusinessTravelTravel_Frequently	9.437773e-01	Education	2.877451e-03
BusinessTravelTravel_Rarely	2.594300e-01	EducationFieldLife Sciences	-2.117974e-01
TrainingTimesLastYear	-9.557372e-02	EducationFieldMarketing	-1.130401e-02
DailyRate	-3.391970e-04	EducationFieldMedical	-1.861822e-01
EnvironmentSatisfaction	-2.720328e-01	EducationFieldOther	-6.221167e-02
MonthlyIncome	-8.912171e-06	EducationFieldTechnical Degree	5.595778e-01
JobSatisfaction	-2.523278e-01	YearsSinceLastPromotion	8.676684e-02
Age	-2.555284e-02	JobInvolvement	-3.516984e-01
JobRoleHuman Resources	3.924634e-01	JobLevel	-7.731213e-02
JobRoleLaboratory Technician	6.448210e-01	YearsInCurrentRole	-8.396999e-02
		WorkLifeBalance	-2.059391e-01

- Pour Lasso

	s1		
(Intercept)	3.359330357	TotalWorkingYears	-0.027203620
(Intercept)	.	NumCompaniesWorked	0.135547191
DepartmentResearch & Development	-0.421623869	MaritalStatusMarried	0.060389314
DepartmentSales	.	MaritalStatusSingle	0.785380294
StockOptionLevel	-0.148110989	Education	.
YearsAtCompany	0.038824167	EducationFieldLife Sciences	-0.084948080
DistanceFromHome	0.029629798	EducationFieldMarketing	.
YearsWithCurrManager	-0.077522122	EducationFieldMedical	-0.046084823
RelationshipSatisfaction	-0.137052864	EducationFieldOther	.
OverTimeYes	1.604945810	EducationFieldTechnical Degree	0.704207076
BusinessTravelTravel_Frequently	1.175543389	YearsSinceLastPromotion	0.097553368
BusinessTravelTravel_Rarely	0.397750133	JobInvolvement	-0.373029105
TrainingTimesLastYear	-0.085421773	JobLevel	-0.017726458
DailyRate	-0.000309856	YearsInCurrentRole	-0.096397569
EnvironmentSatisfaction	-0.299923489	WorkLifeBalance	-0.205269406
MonthlyIncome	.		
JobSatisfaction	-0.283859902		
Age	-0.028994623		
JobRoleHuman Resources	0.192533256		
JobRoleLaboratory Technician	0.741018610		
JobRoleManager	-0.161625583		
JobRoleManufacturing Director	-0.124814786		
JobRoleResearch Director	-1.610813704		
JobRoleResearch Scientist	.		
JobRoleSales Executive	.		
JobRoleSales Representative	0.897253335		

- Pour Elastic net :

	s1		
(Intercept)	3.303128655	TotalWorkingYears	-0.021376750
(Intercept)	.	NumCompaniesWorked	0.123346633
DepartmentResearch & Development	-0.299371817	MaritalStatusMarried	0.012899976
DepartmentSales	0.114897837	MaritalStatusSingle	0.691007527
StockOptionLevel	-0.160523561	Education	.
YearsAtCompany	0.025968404	EducationFieldLife Sciences	-0.092107534
DistanceFromHome	0.027592557	EducationFieldMarketing	.
YearsWithCurrManager	-0.067464897	EducationFieldMedical	-0.060565922
RelationshipSatisfaction	-0.127519252	EducationFieldOther	.
OverTimeYes	1.514288135	EducationFieldTechnical Degree	0.637043254
BusinessTravelTravel_Frequently	1.004106935	YearsSinceLastPromotion	0.089264849
BusinessTravelTravel_Rarely	0.264608920	JobInvolvement	-0.359720775
TrainingTimesLastYear	-0.079303906	JobLevel	-0.046123583
DailyRate	-0.000301757	YearsInCurrentRole	-0.084130579
EnvironmentSatisfaction	-0.281077926	WorkLifeBalance	-0.193829760
MonthlyIncome	.		
JobSatisfaction	-0.263867613		
Age	-0.027522150		
JobRoleHuman Resources	0.265361965		
JobRoleLaboratory Technician	0.683083633		
JobRoleManager	-0.124823536		
JobRoleManufacturing Director	-0.142882883		
JobRoleResearch Director	-1.432401535		
JobRoleResearch Scientist	.		
JobRoleSales Executive	.		
JobRoleSales Representative	0.830477573		

Les coefficients représentés par '.' dans les modèles de Lasso et Elastic net signifient que les variables associées ont été exclues, car ces régularisations ont fortement réduit leur impact sur la prédiction. Cela suggère que ces variables ne contribuent pas de manière significative à la prédiction des modèles.

PARTIE 4 : EVALUATION DES MODELES

Suite à la sélection des variables à l'aide des méthodes ci-dessus, notre objectif est maintenant d'évaluer ces modèles. Nous commencerons par une évaluation en divisant l'ensemble de données en ensembles d'entraînement et de test, puis nous effectuerons l'évaluation sur l'ensemble de test. Ensuite, nous procéderons à une deuxième évaluation en utilisant la validation croisée.

Nous divisons notre ensemble de données en ensemble d'entraînement qui contient 80% d'observations et 20% pour l'ensemble de test pour effectuer l'évaluation.

Après l'entrainement des modèles, nous avons obtenu les performances ci-dessous sur l'ensemble de test :

- Pour le modèle de régression logistique avec la méthode Backward :

```
## Taux de bonne classification : 0.7687627
```

Matrice de confusion :

```
          Predicted
Actual  No Yes
     0 193  55
     1  59 186
```

- Et pour les modèles avec régularisation Lasso ,Ridge et Elastic net :

```
## Matrice de confusion pour Elastic_net_model
##       Predicted
## Actual  No Yes
##      0 191  57
##      1  62 183
## Taux de bonne classification pour Elastic_net_model : 0.7586207
## Matrice de confusion pour Ridge_model
##       Predicted
## Actual  No Yes
##      0 193  55
##      1  60 185
## Taux de bonne classification pour Ridge_model : 0.7667343
## Matrice de confusion pour Lasso_model
##       Predicted
## Actual  No Yes
##      0 193  55
##      1  62 183
## Taux de bonne classification pour Lasso_model : 0.7626775
```

Nous remarquons que le modèle obtenu avec la méthode Backward est le plus performant.

Procédons maintenant à l'évaluation par la méthode de la validation croisée .

Les résultats obtenus après l'évaluation par la validation croisée :

- Pour le modèle de régression logistique avec la méthode Backward :

```
## Taux de bonne classification moyen pour le modèle avec la méthode Backward: 0.7919028
```

- Et pour les modèles avec régularisation Ridge, Lasso et Elastic net:

```
## Taux de bonne classification moyen pour Ridge avec validation croisée : 0.720442
## Taux de bonne classification moyen pour Lasso avec validation croisée : 0.7667605
## Taux de bonne classification moyen pour elasticnet avec validation croisée : 0.7686223
```

Apres l'évaluation par la méthode de validation croisée, nous constatons que le modèle obtenu par la méthode Backward est plus performant que les autres.

PARTIE 5 : CONCLUSION

- Pour le modèle de régression logistique avec la méthode Backward **:**
  - Performance sur l'ensemble de test (0.7687627).
  - Performance après la validation croisée (0.7919028).
- Elastic Net Model :
  - Performance sur l'ensemble de test (0.7586207).
  - Performance après la validation croisée (0.7694208).
- Lasso Model **:**
  - Performance sur l'ensemble de test (0.7626775).
  - Performance après la validation croisée (0.7667605).
- Ridge Model :
  - Performance sur l'ensemble de test (0.7667343).
  - Performance après la validation croisée (0.720442).

Au regard de ce qui précède, le modèle Backward est le choix privilégié en raison de sa robustesse. Cependant si nous devons choisi un modèle avec régularisation, nous choisirons le modèle Elastic Net car il est le deuxième modèle ayant le bon taux de classification.