Projet de Python et Machine Learning

Douleurs Lombaires

Prédire le risque de développer des douleurs lombaires.

Master 2 économie appliquée parcours économie de la santé

Université Paris-Est Créteil

Aleenzahra HAIDER RAZA, équipe

Table des matières

I. Introduction

Les douleurs lombaires, aussi appelées lombalgie ou mal de dos, sont des douleurs localisées dans la région inférieure du dos, au niveau des vertèbres lombaires de la colonne vertébrale. Cette partie fragile et très sollicitée du corps (permet une grande partie des mouvements) subit souvent des tensions et blessures car c'est elle qui supporte le poids du tronc. Les principales causes de douleurs lombaires sont les tensions musculaires, les entorses, les hernies discales, l'arthrose, la sténose spinale, les troubles musculosquelettiques, les traumatismes, les mauvaises postures, le surpoids, le manque d'exercice… En outre, selon l'Assurance Maladie, les lombalgies représentent un enjeu de santé publique. De fait, c'est un symptôme assez répandu, en France 4 personnes sur 5 souffriront de lombalgie commune au cours de leur vie, et plus de la moitié de la population française a eu au moins un épisode de mal de dos sur une période de douze mois.[1]

Un ensemble de données comprenant 310 observations sur des détails physiques collectés sur la colonne vertébrale est mis à notre disposition. L'objectif principal de cette étude est de prédire le risque de développer des douleurs lombaires à partir de notre échantillon. Dans ce cadre, nous aurons recours dans un premier temps au Machine Learning. Cette technique permettant une réalisation de modèles prédictifs d'événements ou de comportements nous donnera la possibilité de mettre en place un modèle de prédiction des douleurs lombaires. Par la suite nous mobiliserons deux algorithmes qui sont une forêt aléatoire (en anglais, Random Forest) et un XGBoost (eXtreme Gradient Boosting) afin d'apporter des résultats plus concrets à notre étude.

II. Base de données

Dans le cadre de cette étude, nous utiliserons la base de données à notre disposition qui contient 310 individus et 13 variables. La variable que nous expliquerons est celle dénommée « Class_att » qui représente le risque de développer des douleurs lombaires. Celle-ci est binaire, c'est-à-dire qu'elle prend la valeur 1 lorsqu'il existe un risque de développer des douleurs lombaires et 0 dans le cas inverse. Les autres variables utilisées dans nos modèles sont référencées en annexe 1. Ces variables représentent des outils d'analyses radiologiques (angles rachidiens ou pelviens) permettant de savoir si une personne est dans les normes morphologiques. De fait, les types de morphologies ont des niveaux de risques plus ou moins importants de créer des douleurs lombaires.

[1] Assurance Maladie

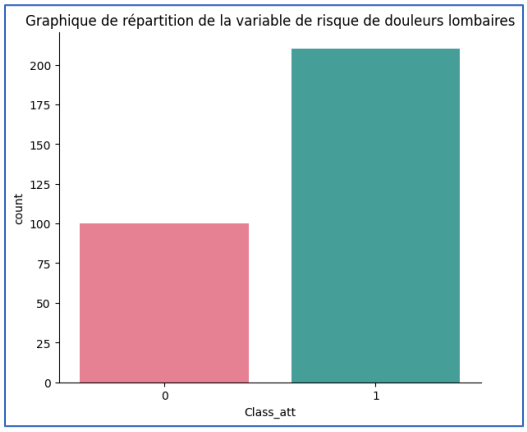
III. Statistiques descriptives :

1) Nettoyage de la base de données :

a) Valeurs manquantes

Avant de nous lancer dans nos analyses, nous évaluons notre base de données. Nous cherchons d'abord la présence de valeurs manquantes et nous remarquons que la base de données n'en comporte aucune. Nous avons donc une table exempte de valeurs manquantes.

b) Répartition des variables



Concernant notre variable expliquée, nous évaluons la fréquence de risque de développer des douleurs lombaires dans notre échantillon.

Nous pouvons voir qu'il y a 67.7% d'individus qui ont un risque de douleurs lombaires contre seulement 32.3% qui en n'ont pas. De fait, notre échantillon présente deux fois plus d'individus pouvant souffrir de douleurs lombaires.

Concernant les variables explicatives que nous utiliserons dans nos analyses, nous procéderons à l'évaluation de leur répartition afin de détecter la présence de valeurs extrêmes, nous traiterons ces valeurs en fonction de la pertinence. Par la suite nous procéderons à la description de nos variables à l'aide de statistiques descriptives (moyenne, écart-type…). ***(Cf. Annexe 1 et 2)***.

Après analyse, nous avons décelé la présence de valeurs extrêmes dans quatre variables. Nous en avons dans la variable « pelvic_radius » qui caractérise la ligne entre l'axe de la hanche et le coin postérieur du plateau vertébral. Par ailleurs, nous remarquons que la distribution est en forme de cloche (normale), dans ce cas, par précaution nous ne procédons pas à la suppression des valeurs extrêmes de cette variable.

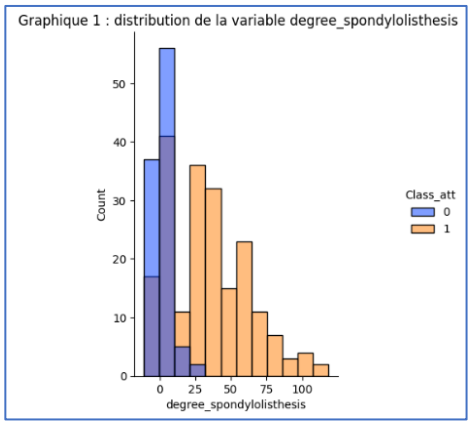
Pour la plupart des variables, on peut voir que la moyenne et la médiane sont assez proches, à l'exception de la variable « degree_spondylolisthesis » qui présente le degré de la quantité de glissement d'une vertèbre sur une autre.
Sur cette variable, nous décelons la présence de valeurs extrêmes. Étant donné que la médiane est faiblement sensible aux valeurs extrêmes, la moyenne, très sensible, n'a pas l'air d'être biaisée. Cependant, nous pouvons voir que l'écart type montre que la plupart des données sont éloignées de la moyenne.
Nous remarquons également la présence de valeurs extrêmes dans la variable « sacral_slope » qui présente l'angle d'inclinaison du plateau sacré de profil par rapport à l'horizontale ainsi que dans la variable « lumbar_lordosis_angle ».

Pour ces trois variables, nous supprimerons les valeurs extrêmes car leur distribution est comprise entre 99% et 100%.

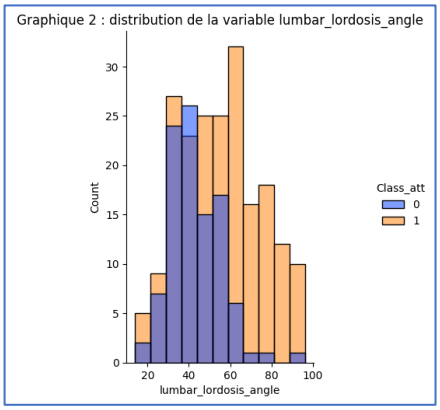
2) Analyse des interactions entre notre variable d'intérêt et les variables explicatives du modèle :

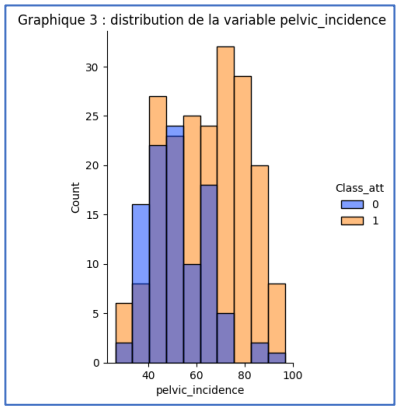
Nous cherchons dans cette partie à analyser l'interaction entre les variables explicatives (qui traitent de la santé des individus) du modèle et la variable expliquée qui nous indique le risque de développer des douleurs lombaires.

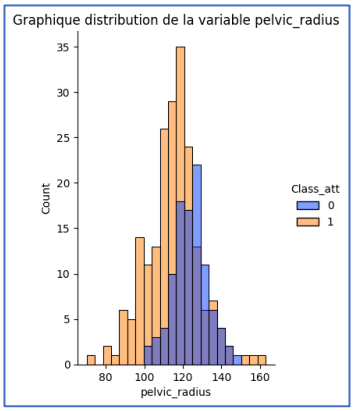
La manière la plus simple de visualiser ces corrélations est de les mettre sous forme de graphique. Nous avons alors la répartition des individus ayant développés ou non des douleurs lombaires en fonction de différentes variables explicatives.



Nous nous intéresserons tout d'abord à la répartition de la variable degré spondylolisthésis définie comme un trouble de la moelle épinière dans lequel une vertèbre glisse sur celle en dessous, provoquant ainsi des douleurs dans le bas du dos. Nous nous attendons à ce que celle-ci soit corrélée positivement avec le fait que l'individu déclare avoir des douleurs lombaires, ce qui est effectivement le cas. En effet, les individus n'ayant pas déclaré de douleurs au dos ont un très faible degré spondylolisthésis, seuls les individus ayant des douleurs au dos déclarent un degré supérieur à 30.







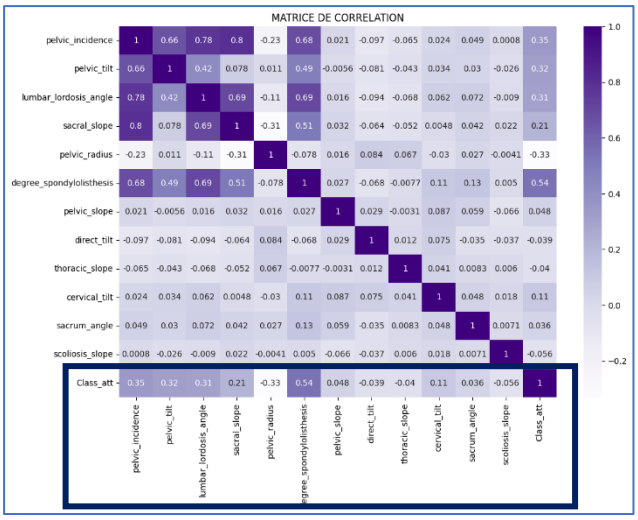
Pour le graphique n°2 sur l'angle du lordose lombaire, nous voyons que la plupart de l'échantillon se situe dans la norme (entre 39° et 53°). Nous avons également beaucoup d'individus allant au-delà du 60°, majoritairement ceux ayant déclaré avoir des douleurs au dos. Nous tenons cependant à rappeler que nous avons une surreprésentation des individus avec des douleurs lombaires, ce qui explique en partie cette distribution.

Même chose pour l'incidence pelvienne (graphique n°3), une grande partie de l'échantillon se situe dans la moyenne, donc entre 30° et 77°, ne vont au-delà de cette fourchette que ceux déclarant avoir des douleurs lombaires (nous avons tout de même quelques-uns ayant répondu ne pas avoir de douleurs). Dans l'ensemble, une distribution assez équilibrée des individus avec

des douleurs lombaires dont une grande partie se situant dans la moyenne. Nous aurions pu penser que ces individus auraient plutôt présenté une incidence plus élevée que ceux ayant répondu négativement, mais ce n'est pas ce que nous observons sur nos graphiques.

Pour le dernier graphique sur le rayon pelvien, nous remarquons que la distribution est en forme de cloche et est assez similaire pour les deux groupes, même si celle du groupe n'ayant pas déclaré des douleurs lombaires se décalent un peu plus vers la droite. Donc, plus le rayon est grand, moins nous avons d'individus ayant déclarés avoir des douleurs au dos.

3) Matrice de corrélation :



Nous pouvons à présent nous intéresser à la corrélation entre nos variables explicatives et notre variable d'intérêt « Class_att ». Pour cela nous utilisons ce qui s'appelle une matrice de corrélation (ou *heatmap*), celle-ci nous donne un coefficient de corrélation compris entre 1 et -1. Plus on est proche de 1, plus la variable est corrélée positivement (respectivement -1, négativement). Si le coefficient est proche de 0, alors nous avons une absence de corrélation entre les deux variables.

Sur notre matrice, nous voyons que les variables corrélées positivement à notre Y sont les variables suivantes : « degree_spondylolisthesis » à 54% (la plus corrélée), « pelvic_incidence » à 35%, « pelvic_tilt » à 32%, et d'autres plus faiblement corrélées comme « lumbar_lordosis_angle » et « sacral_slope ». Nous avons également la variable « pelvic_radius » qui est corrélée à notre variable à expliquer, mais négativement à -33%.

Nous avons jusque là résumé la base de données et l'interaction entre nos différentes variables. Nous pouvons à présent passer à la partie principale qui est la modélisation de la prédiction de notre variable d'intérêt.

IV. Modélisation

Notre variable Y est binaire, nous allons donc faire un modèle supervisée pour résoudre ce problème de classification. Cela consiste à classer les données dans la catégorie « déclarer des douleurs lombaire » ou dans la catégorie « ne pas déclarer de douleurs lombaires » en fonction des caractéristiques médicales.

La méthode consiste à diviser aléatoirement la base en deux, en données d'entraînement et données de test, 70% et 30% respectivement. Les données d'entrainement sont utilisées pour entraîner le modèle, et comprendre le lien qui existe entre les variables indépendantes et dépendante.

Les données de test consistent à évaluer la performance du modèle. En effet, l'algorithme va prédire le risque de déclarer des douleurs lombaires (Y_pred) en prenant en compte les variables

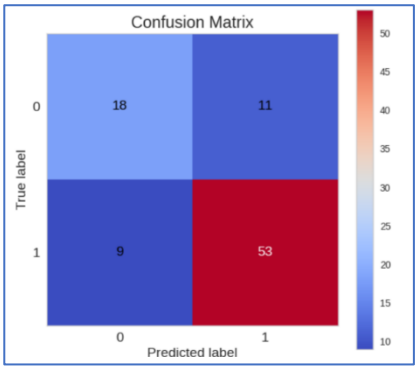
indépendantes des données de test. Cette prédiction est ensuite comparée à la variable dépendante Class_att des données de test.

Les modèles qui seront réalisés sont les suivants : forêt aléatoire et XGBoost. Le choix de ces modèles s'explique par leur robustesse aux valeurs extrêmes et le peu de sensibilité à la multicolinéarité des variables comparément à d'autres modèles.

1) Forêt aléatoire

La forêt aléatoire (Random Forest, en anglais), est une technique d'apprentissage automatique utilisée pour la classification. Elle utilise un ensemble d'arbres de décision construits de manière aléatoire pour prédire la classe (déclarer des douleurs lombaires ou ne pas en déclarer) d'une nouvelle donnée. Les prédictions de chaque arbre de décision sont combinées pour produire une prédiction finale de classe.

a) La matrice de confusion :



La matrice de confusion est un outil qui permet d'évaluer la performance d'un modèle de classification en comparant les prédictions du modèle avec les vraies classes des données. Elle est représentée sous forme d'une matrice carrée avec les lignes correspondant à la réalité et les colonnes à la prédiction. Quatre éléments se présentent: les vrais positifs (VP), les vrais négatifs (VN), les faux positifs (FP) et les faux négatifs (FN).

```
La base TRAIN à les dimensions suivantes : (211, 12)
La base TEST à les dimensions suivantes : (91, 12)
-------------------------------------------
Résultats du Random Forest

Accuracy: 0.78
Recall: 0.85
Precision: 0.83
AUC : 0.74
```

Les vrais positifs (VP) correspondent aux individus qui ont réellement des douleurs lombaires et qui ont été correctement identifiés comme tels par le modèle de classification (en bas à droite). Les vrais négatifs (VN) sont les individus correctement classés comme n'ayant pas de douleurs lombaires, c'est également une prédiction réaliste (en haut à gauche). Les faux positifs (FP) représentent le nombre d'individus qui ont été identifiés à tort comme ayant des douleurs lombaires par le modèle de classification (en haut à droite). Les faux négatifs (FN) sont les individus qui ont réellement des douleurs lombaires mais qui ont été incorrectement identifiés comme ne les ayant pas (en bas à gauche).

La matrice de confusion permet également de calculer des métriques importantes telles que l'exactitude, le rappel, la précision et l'AUC. Ces métriques fournissent des informations sur la qualité des prédictions du modèle et permettent de mesurer sa capacité à prédire (avec précision) les différentes catégories de données.

L'exactitude est la proportion d'individus correctement classés parmi tous les individus.

Exactitude=(VN+VP)/(VN+FP+FN+VP)

=> (18+53)/(18+11+9+53) => (71/91)*100➔78%

Cela signifie que le modèle de classification a correctement classé 78% des individus.

Le rappel (recall) mesure la proportion d'individus ayant réellement des douleurs lombaires qui ont été correctement identifiés par le modèle de classification.

Rappel = VP / (VP + FN)=>53/ (53+9)= (53/62)*100 ➔ 85,4%

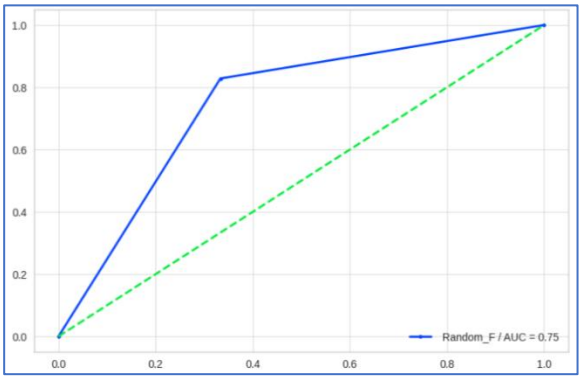
Cela signifie que le modèle a correctement identifié environ 85,4% des individus ayant réellement des douleurs lombaires, tandis que 14,6% de ces individus ont été mal classés (faux négatifs).

La précision mesure la proportion d'individus identifié comme ayant des douleurs lombaires qui ont effectivement des douleurs lombaires.

Précision = VP / (VP + FP) => 53/(53+11)*100 => 53/(64)*100 ➔82,8%

Parmi les individus prédits comme ayant des douleurs lombaires, 82,8% ont effectivement des douleurs lombaires, tandis que 17,2% ont été mal classés.

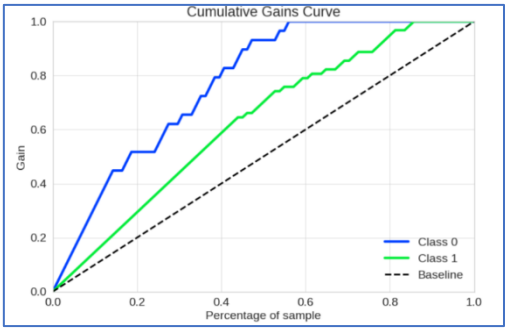
b) La courbe ROC & AUC



L'aire sous la courbe ROC (AUC) est une mesure qui évalue la performance d'un modèle de classification en mesurant sa capacité à différencier les individus qui ont des douleurs lombaires de ceux qui n'en ont pas. Cela se fait en examinant les probabilités de prédiction du modèle pour différents seuils de décision et en traçant une courbe représentant le taux de vrais positifs en fonction du taux de faux positifs.

Dans notre cas, l'AUC vaut 0,75 et la courbe est au-dessus de la courbe linéaire aléatoire ((la courbe pointillée en vert) donc le modèle est meilleur qu'un modèle aléatoire pour prédire les individus atteints de douleurs lombaires. Il indique une performance de prédiction plutôt raisonnable dans sa capacité à différencier les individus atteints de douleurs lombaires de ceux qui n'en ont pas, mais il existe encore une certaine marge d'amélioration. Plus l'AUC est élevée, meilleur est le modèle de classification dans sa capacité à différencier les individus atteints de douleurs lombaires de ceux qui n'en ont pas.

c) La courbe de gains cumulés



La courbe de gains cumulés permet de comparer la performance du modèle à celle d'une prédiction aléatoire (la courbe noire en pointillée). la phrase manque de clarté, il serait plus clair de dire Nous nous concentrons ici sur la courbe en vert, qui représente la capacité du modèle à prédire les individus atteints de douleurs lombaires. Lorsqu'on cible 20% des individus à risque prédit par notre modèle, on identifie 30% d'individus ayant des

douleurs lombaires. Si on cible 40% des personnes, on obtient plus de la moitié des individus ayant des douleurs lombaires (60%).

2) XGBoost

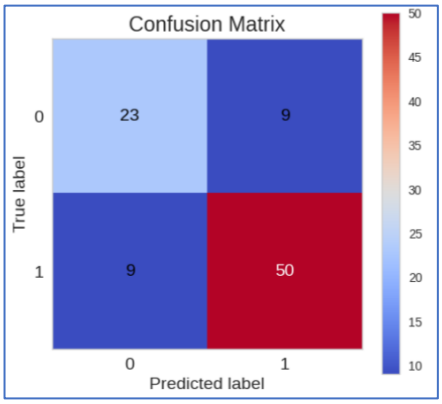
a) La matrice de confusion

XGBoost est également un algorithme de machine learning qui se base sur des arbres de décision pour résoudre des problèmes de classification. Il utilise une méthode de renforcement itérative pour entraîner des arbres de décision faibles. Cette méthode s'adapte en apprenant de ses erreurs et en cherchant à les corriger à chaque étape pour améliorer ses prédictions.

```
La base TRAIN à les dimensions suivantes : (211, 12)
La base TEST à les dimensions suivantes : (91, 12)
-------------------------------------------
Résultats du XGBoost

Accuracy: 0.80
Recall: 0.85
Precision: 0.85
AUC : 0.78
```

Notre modèle est exacte à 80%, il a correctement identifié 80% des individus soit légèrement supérieur à celui du modèle de Random Forest (78%). Exactitude => (23+50)/(23+9+9+50) => (73/91)*100 ➔80,2%



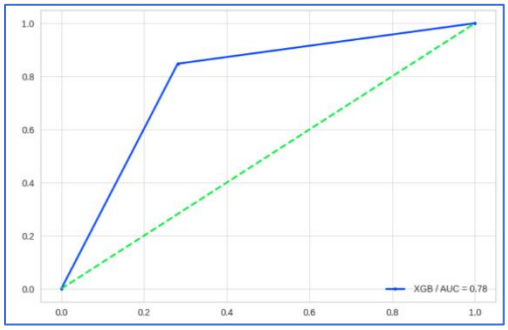
Notre modèle de XGBoost a identifié 85% des individus qui ont déclaré avoir des douleurs lombaires dans notre modèle soit un peu moins que le dernier modèle (85,4%).

Rappel = VP / (VP + FN)=>50/(50+9)= (50/59)*100 ➔ 85%

Parmi la prédiction des individus ayant des douleurs lombaires, 85% ont effectivement des douleurs lombaires. Ce qui est plus précis par rapport au dernier modèle (83%).

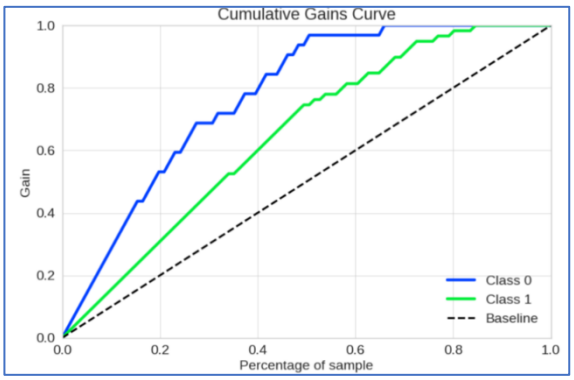
Précision = VP / (VP + FP) => 50/(50+9)*100 => 50/(59)*100 ➔85%

b) La courbe ROC & AUC



La courbe de ROC est supérieure à la droite (en pointillé) linéaire aléatoire et l'AUC de notre modèle vaut 0,78, proche de 1, ce qui se situe dans le coin supérieur gauche du graphique. Le modèle a une performance de prédiction raisonnable pour différencier les individus atteints de douleurs lombaires de ceux qui n'en ont pas, avec une amélioration nette par rapport au dernier modèle

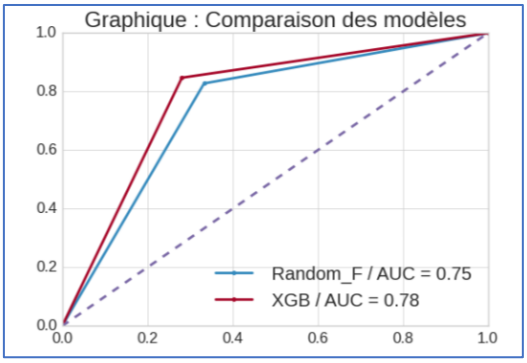
c) La courbe de gains cumulés



La courbe de gains cumulés indique qu'en sélectionnant 20% de notre échantillon, on identifie 35% des individus ayant des douleurs lombaires. Si on cible 40% des personnes, on obtient 60% des individus ayant des douleurs lombaires. Lorsqu'on cible 40% des individus de l'échantillon, on obtient plus de 80% d'individus ayant des douleurs lombaires.

3) Comparaison des courbes :

L'étude consiste ici à comparer les courbes ROC pour identifier le modèle le plus performant. Le sommet de la courbe du modèle XGBoost est plus proche de 1 (coin en haut à gauche) que le sommet de la courbe Random Forest. De plus, le modèle XGBoost est plus précis (0,80>0,78) que le modèle de forêt aléatoire et il est plus apte à identifier les classes des individus (0,78>0,75).

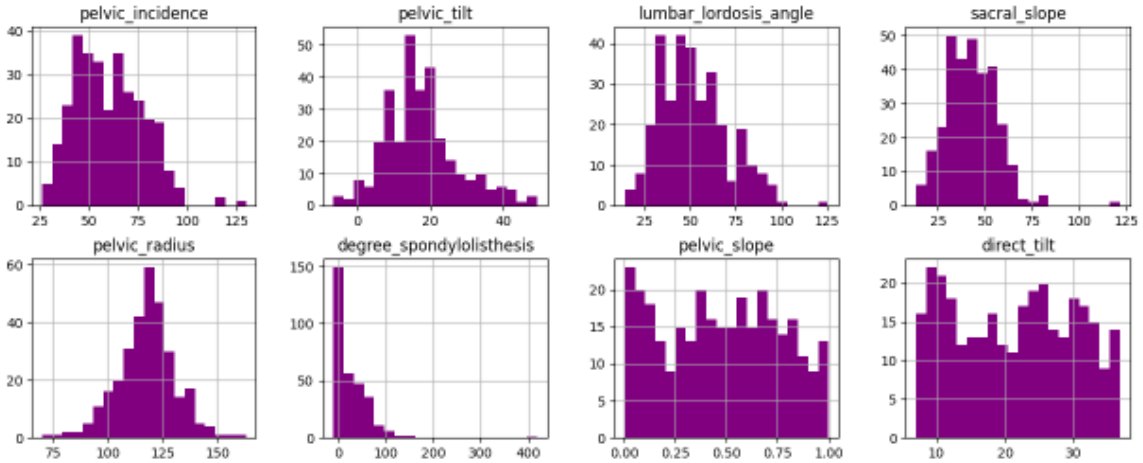


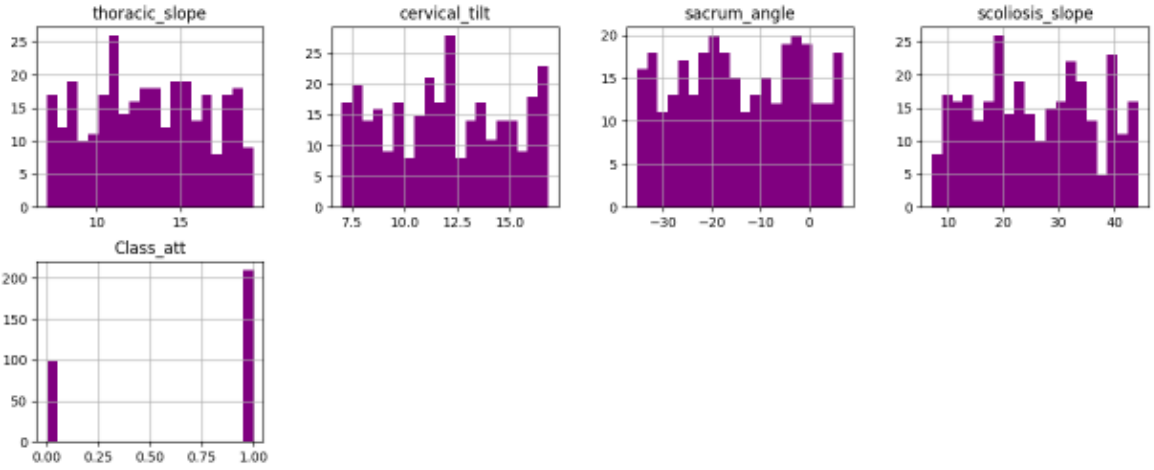
V. Annexe :

a) Annexe 1 : tableau de statistiques descriptives :

	count	mean	std	min	5%	25%	50%	75%	95%	99%	max
pelvic_incidence	310.00	60.50	17.24	26.15	35.99	46.43	58.69	72.88	87.87	96.55	129.83
pelvic_tilt	310.00	17.54	10.01	-6.55	3.38	10.67	16.36	22.12	37.55	46.20	49.43
lumbar_lordosis_angle	310.00	51.93	18.55	14.00	26.85	37.00	49.56	63.00	85.60	95.04	125.74
sacral_slope	310.00	42.95	13.42	13.37	23.49	33.35	42.40	52.70	63.43	78.30	121.43
pelvic_radius	310.00	117.92	13.32	70.08	95.34	110.71	118.27	125.47	139.14	148.47	163.07
degree_spondylolisthesis	310.00	26.30	37.56	-11.06	-4.08	1.60	11.77	41.29	81.69	124.39	418.54
pelvic_slope	310.00	0.47	0.29	0.00	0.04	0.22	0.48	0.70	0.93	0.99	1.00
direct_tilt	310.00	21.32	8.64	7.03	8.55	13.05	21.91	28.95	34.98	36.34	36.74
thoracic_slope	310.00	13.06	3.40	7.04	7.59	10.42	12.94	15.89	18.44	19.18	19.32
cervical_tilt	310.00	11.93	2.89	7.03	7.43	9.54	11.95	14.37	16.46	16.74	16.82
sacrum_angle	310.00	-14.05	12.23	-35.29	-33.20	-24.29	-14.62	-3.50	5.14	6.53	6.97
scoliosis_slope	310.00	25.65	10.45	7.01	9.80	17.19	24.93	33.98	42.68	43.87	44.34
Class_att	310.00	0.68	0.47	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00

b) Annexe 2 : répartition des variables de la base de données :




thoracic_slope
cervical_tilt
sacrum_angle
scoliosis_slope
Class_att