Analyse de données

Biomechanical-features-of-orthopedic-patients

Margaux Peschiera, Tom Bourg, Yunfei Zhao

19 juin 2020

Sommaire

Description des variables

Les variables

- pelvic incidence (**PI**) : Incidence pelvienne.
- pelvic tilt numeric(**PT**) : Inclinaison pelvienne.
- sacral slope (**SS**) : La pente sacrée.
- pelvic radius (**PR**) : Rayon pelvien.
- lumbar lordosis angle (**LLA**) : Angle de la lordose lombaire.
- degree spondylolisthesis(**DS**) : Indicateur pour mesurer le niveau de spondylolisthésis.

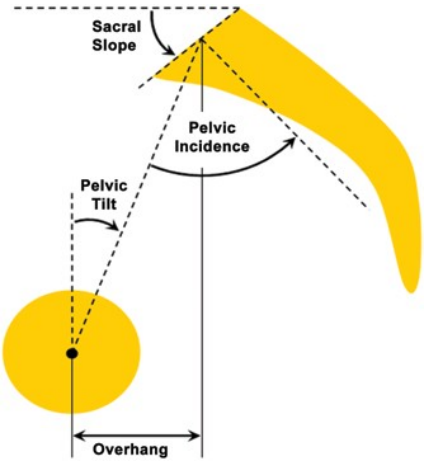

Figure: Angles Pelviens

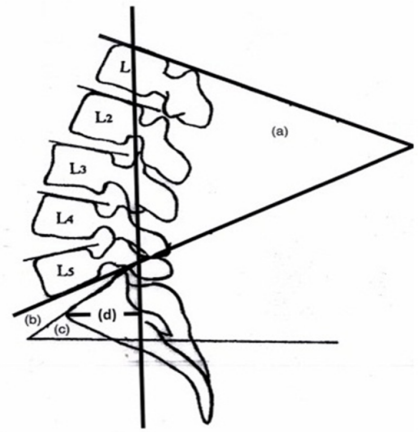

Figure: Représentation des angle LLA

Analyse exploratoire

Jeu de données

- Tableau individus-variables (310 l x 7 col)
- Une ligne correspond aux données d'un patients concernant 6 caractéristiques (des angles) et une colonne étiquette pour savoir si le patient est malade ou non.
- Maladies de l'appareil locomoteur
- $PI = PT + SS$

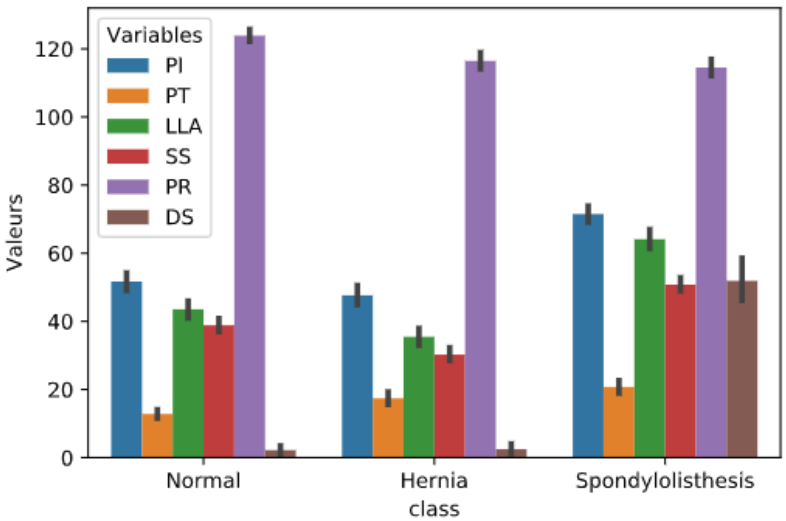

Figure: Barplot

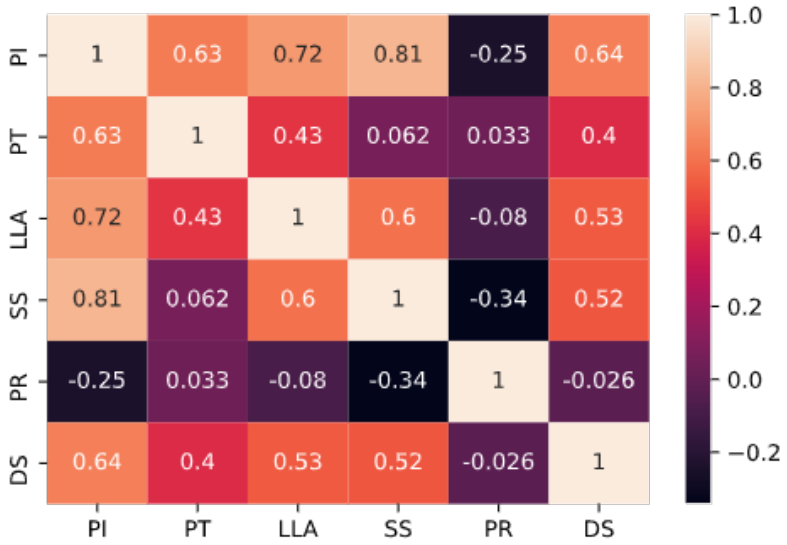

Figure: Tableau de covariance

Deux types d'apprentissages

Apprentissage non supervisé

- Apprentissage autonome, données non étiquetées.

Apprentissage supervisé

- L'utilisateur "aide" l'algorithme en lui fournissant des données étiquetés. Le modèle apprend alors de chaque exemple en ajustant. Le but étant d'être capable de généraliser l'apprentissage à de nouveaux cas.

Évaluation des modèles

- Tailles des classes ne sont pas équilibrées
- Jeu de données concerne des maladies donc il faut éviter les faux négatifs
- Nested Cross Validation pour determiner les hyper-paramètre
- Utilisation du F-1 score pour un équilibre entre précision et rappel

$$F1\_score = 2 * \frac{Precision * Rappel}{Precision + Rappel}$$

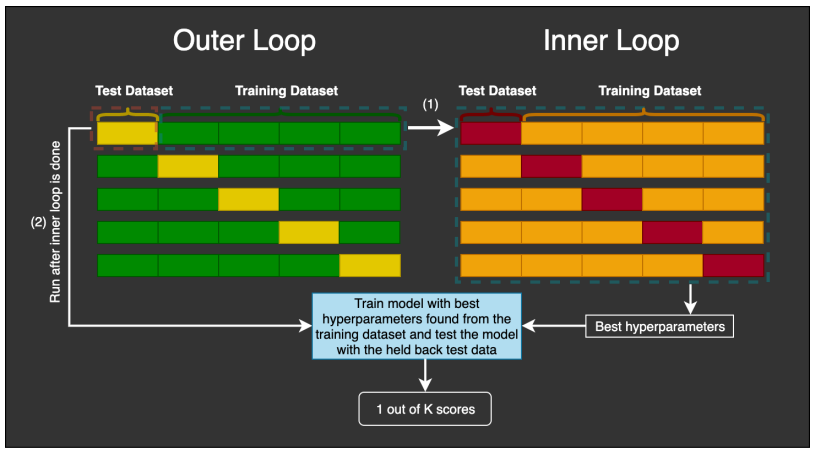

Figure: Nested Cross Validation

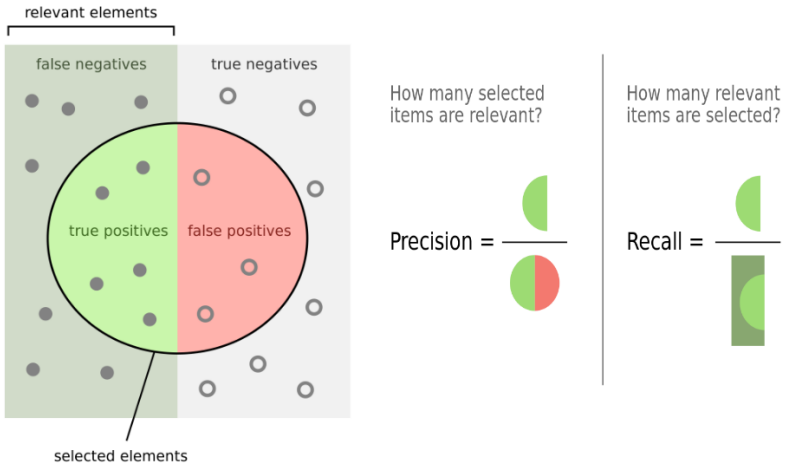

Figure: Précision - Rappel

Apprentissage non supervisé - ACP

Analyse en composantes principales

- L'objectif est de réduire nos données multidimensionnelles afin de pouvoir les visualiser dans un plan.

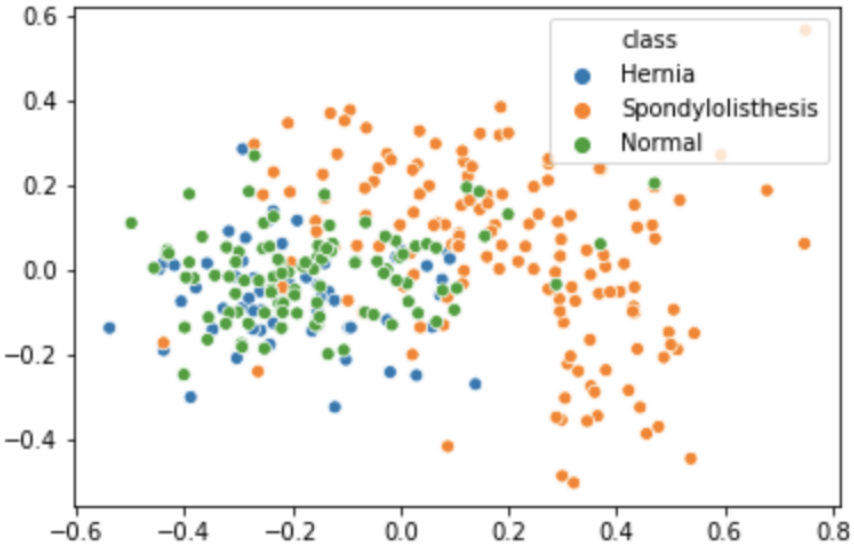

Figure: ACP 3 classes

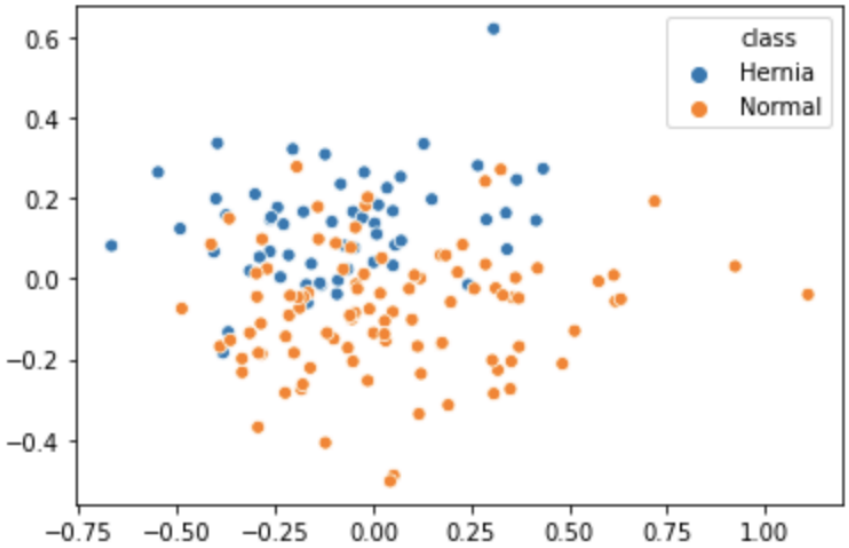

Figure: ACP sans Spondylolisthesis

Apprentissage non supervisé - Kmeans et CAH

Kmeans et CAH

- Classification automatique par Kmeans et CAH, K = 3

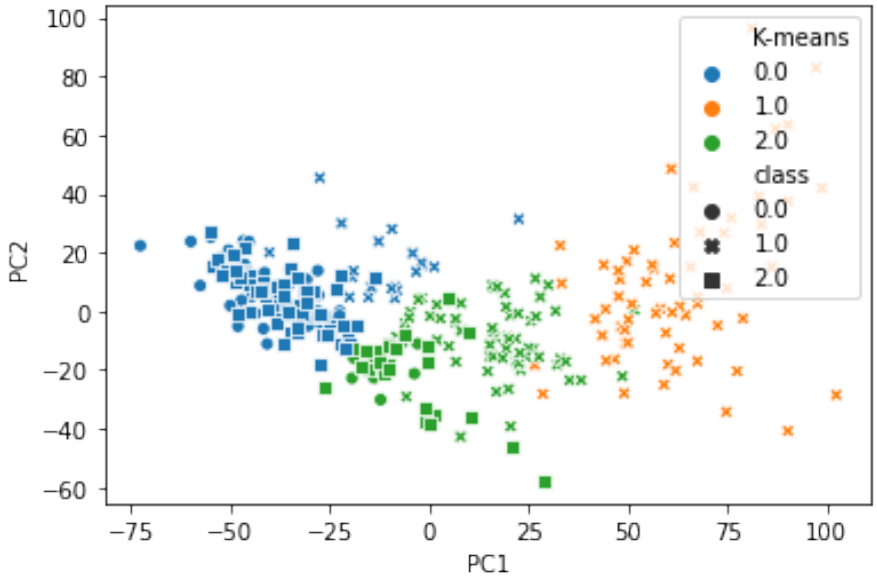

Figure: Résultats de l'algorithme des KMeans, F1-Score : 47,02%

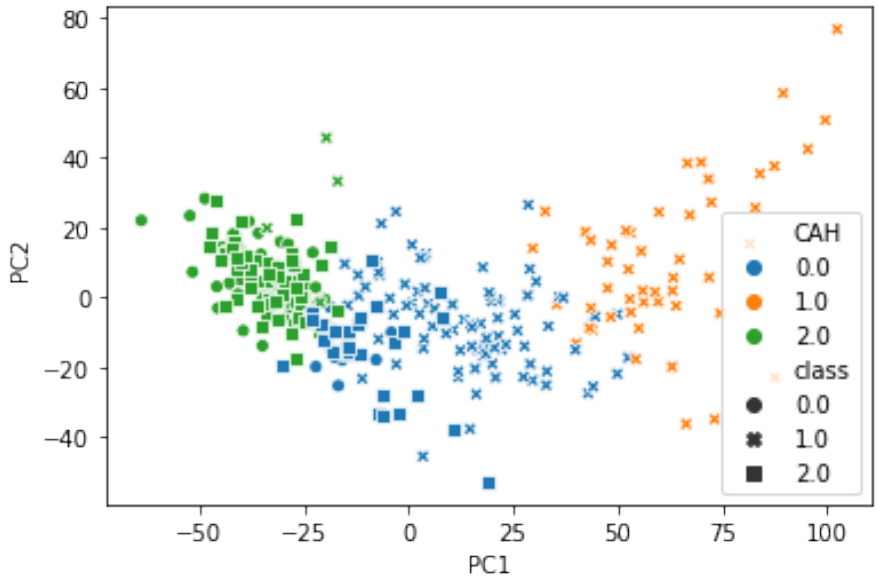

Figure: Résultats de la CAH, F1-Score : 47,68%

Apprentissage supervisé - KNN et ACP

KNN et ACP

- Les variables étant corrélées, nous avons appliqué les KNN sur les dimensions obtenues avec l'ACP.
- F1-Score moyen supérieur à celui obtenu avec KNN sur dimensions initiales.
- F1-score : 83,14%

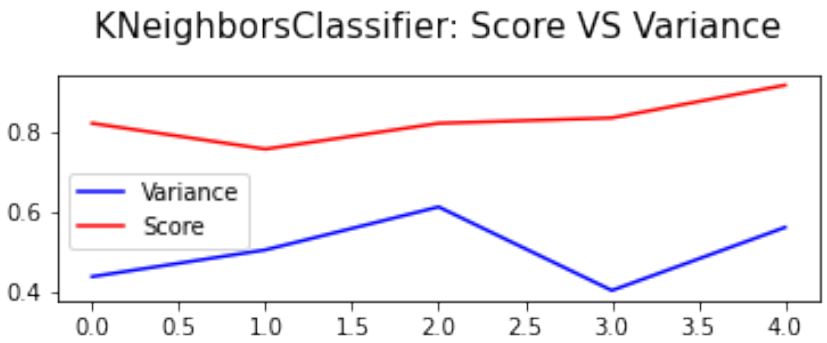

Figure: F1-Scores et Variances du KNN (dimensions ACP) pour 5 itérations de la boucle extérieur

Apprentissage supervisé - Régression Logistique

Régression Logistique

- Estimer directement les probabilités d'appartenance aux classes
- F1-score : 85,46%

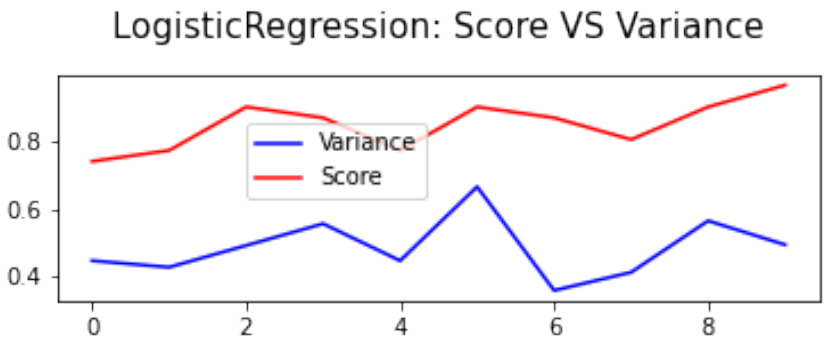

Figure: Affichage des F1-Scores et Variances de la régression logistique pour 10 itérations de la boucle extérieur

Apprentissage supervisé - Arbre de décision

Arbre de Décision

- Tracer le processus de décision et comprendre l'impact de chaque variable
- F1-Score : 79,03%

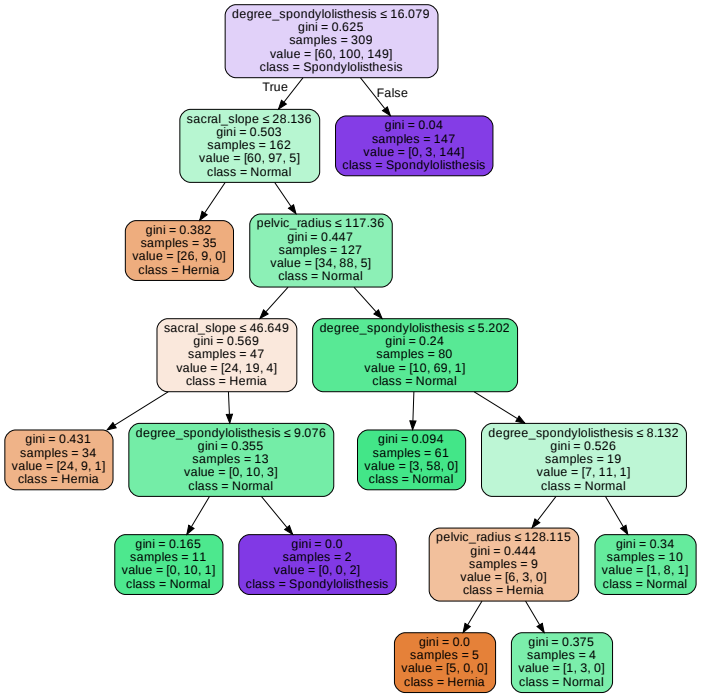

Figure: arbre de décision avec $ccp - alpha$: 0.008

Apprentissage supervisé - Forêt aléatoire

Forêt Aléatoire

- Création de 200 arbres de décision à partir de tirage avec remise dans les données
- F1-score : 82,86%

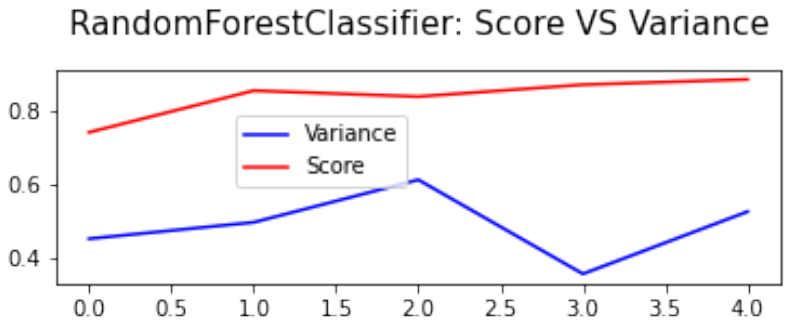

Figure: Forêt aléatoire scores

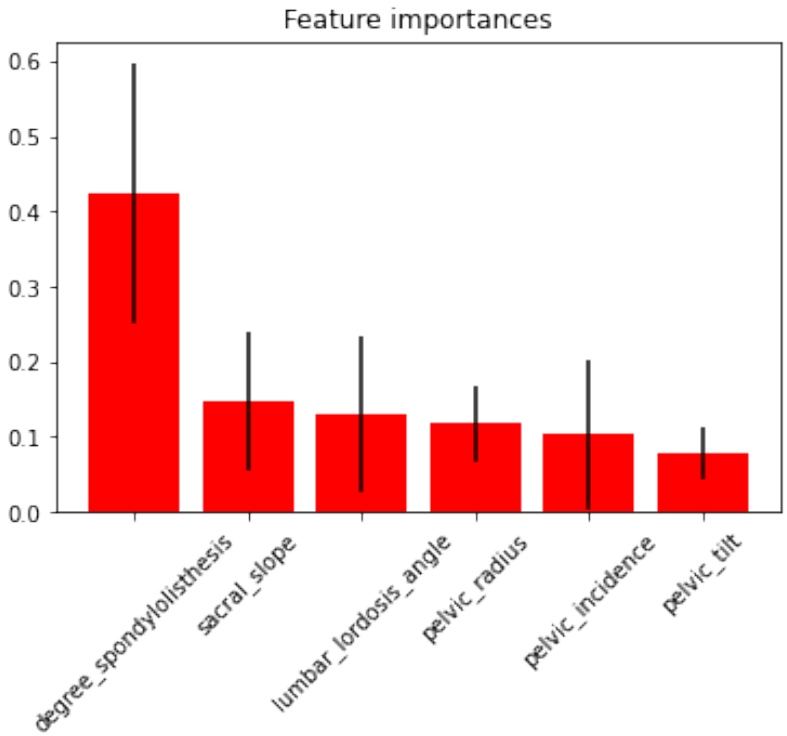

Figure: Axes importants de la forêt aléatoire

Apprentissage supervisé - NCA

Analyse des composantes du voisin

- Méthode de réduction de dimension
- Utilise la distance Mahalanobis.
- F1-score : 86,44%

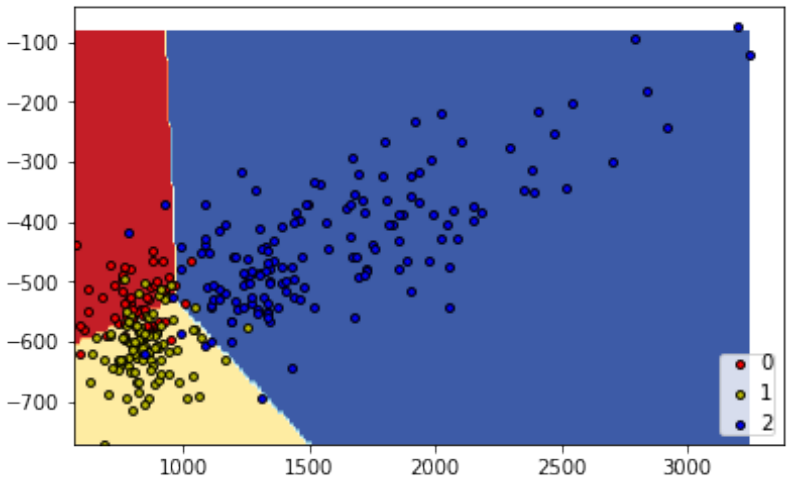

Figure: NCA + LR

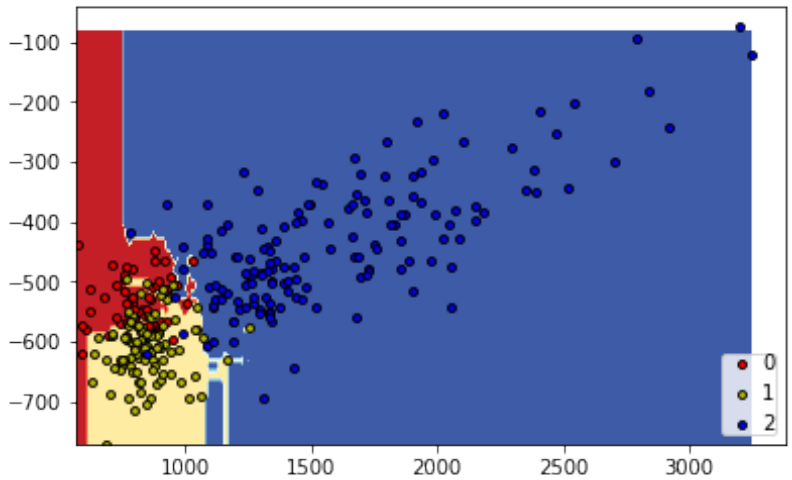

Figure: NCA + RF

Conclusion

- Comparaison des scores pour chaque méthode
- Meilleur score obtenu avec l'apprentissage supervisé (50% vs 85%)
- Meilleur F1-score : 86,44%

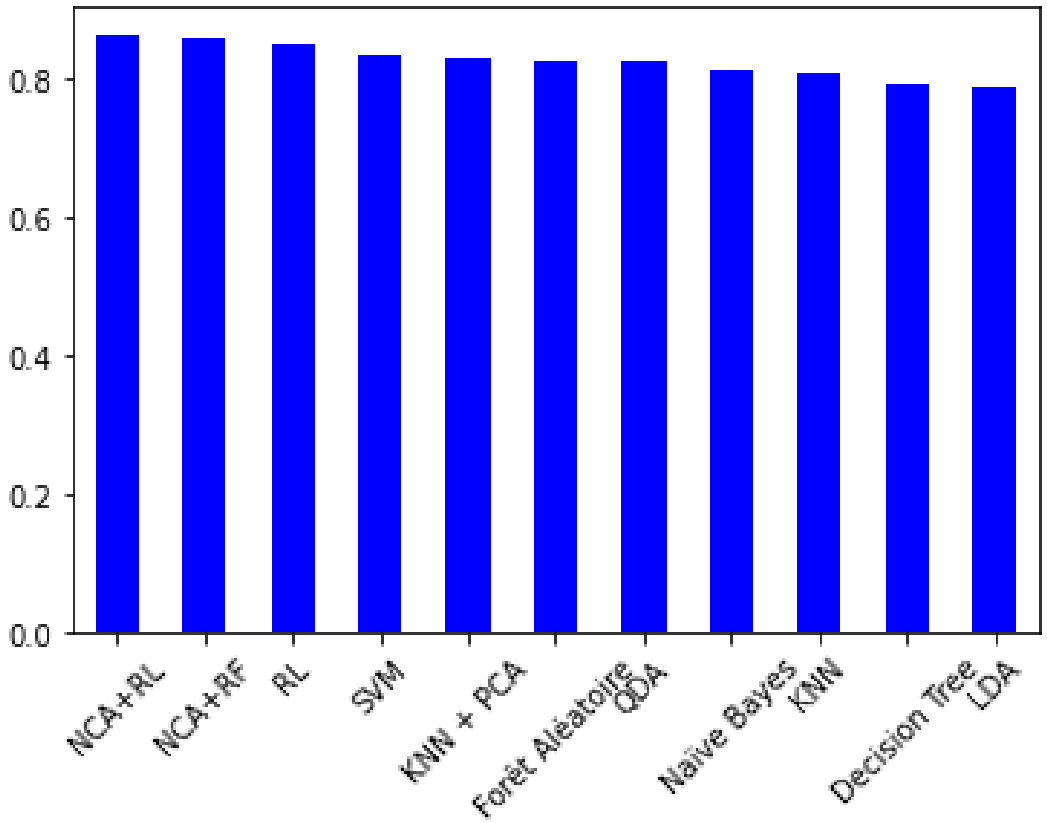

Figure: Barplot des F1-Scores obtenus avec les méthodes supervisées