Projet SY09 - Analyse de données orthopédiques de la colonne lombaire

Reyhan AMAROUCHE; Alexandre BERTOLOTTO; Pauline GUILLET

Printemps 2019

1 Introduction

Nous étudions dans ce document un jeu de données orthopédiques. L'orthopédie est la discipline qui traite des affections de l'appareil locomoteur et de la colonne vertébrale (os, articulations, ligaments, tendons et muscles) [1].

D'après [2], les données ont été collectées par le Dr. Henrique DA MOTA au sein du Groupe de recherche appliquée en orthopédie du Centre Médico-Chirurgical de Réadaptation des Massues de Lyon (France). Celles-ci sont mises à la disposition du grand public par l'Université de Californie à Irvine (USA) et par la plateforme Kaggle.

Dans le cadre de l'unité de valeur SY09 (Analyse de Données et Data Mining), nous avons réalisé un projet en R dont l'objectif est de mettre en oeuvre les méthodes d'analyse de données étudiées en cours sur des données réelles. L'objectif de ce document est de présenter le travail qui a été effectué et les réflexions qui ont été menées lors de l'analyse des données relatives à la colonne lombaire.

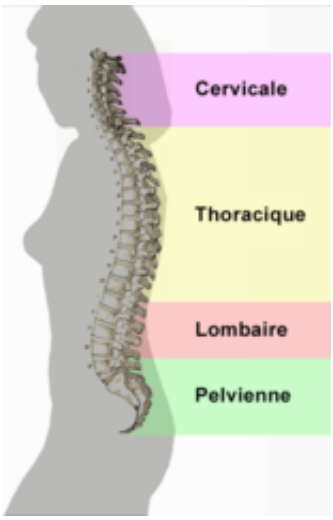

FIGURE 1 – Aperçu des différentes régions de la colonne vertébrale

2 Aperçu de l'orthopédie lombaire

2.1 Description anatomique

Chez l'être humain, la colonne vertébrale (également nommée rachis) est un empilement d'os articulés appelés vertèbres. Les vertèbres sont séparées par des disques intervertébraux qui jouent le rôle d'amortisseurs grâce à un noyau liquide. La colonne vertébrale comporte quatre régions différentes que l'on peut observer en 1, nous nous concentrerons ici sur les deux zones du bas du dos : la zone lombaire et la zone pelvienne (également appelée sacrum). Les régions lombaires et pelviennes sont soumises au quotidien à des pressions importantes. Cela peut conduire à des douleurs et pathologies au niveau de ces zones.

2.2 Pathologies de la zone lombaire

La hernie discale et la spondylolisthésis sont des affections répandues de la zone lombaire. Le vieillissement et la pratique de sports extrêmes (surf par exemple) sont des facteurs aggravants pour le développement de ces pathologies.

2.2.1 Hernie discale

Cette affection de la colonne vertébrale est caractérisée par la saillie anormale du noyau liquide des disques intervertébraux comme on peut le voir en annexe A. Elle peut provoquer de fortes douleurs pour le patient en cas d'irruption en dehors de leur position classique d'après [3].

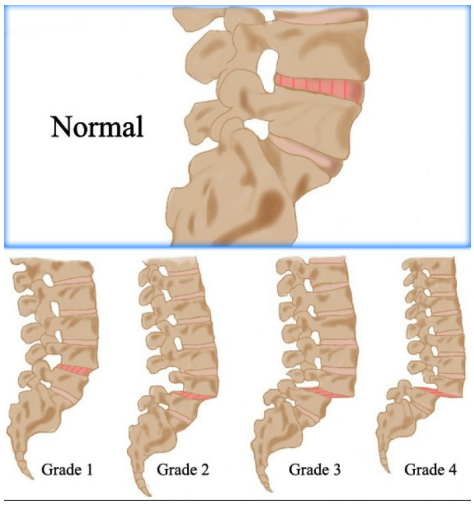

FIGURE 2 – Comparaison d'individus présentant divers degrés de spondylolisthésis par rapport à la normale

2.2.2 Spondylolisthésis

Cette affection de la colonne vertébrale est caractérisée par un glissement en avant ou en arrière d'une vertèbre par rapport à à celle en dessous d'elle, entraînant avec elle le reste de la colonne vertébrale. Différents degrés de cette affection sont présentés en 2.Tout comme la hernie discale, cette pathologie peut engendrer de fortes douleurs.

3 Présentation du jeu de données

3.1 Description générale

Le jeu de données comporte deux fichiers contenant chacun un tableau individus-variables. Les deux tableaux individus-variables représentent les mêmes individus anonymes et les mêmes caractéristiques orthopédiques. Les variables correspondent à des caractéristiques de la colonne lombaire et du sacrum des individus.

3.2 Un jeu de données propre et de faible volume

Les jeux de données en statistiques peuvent parfois être très lacunaires et nécessiter des traitements préalables. Ce jeu de données est particulièrement propre. Il n'y a aucune valeurs manquantes et les noms de variables, bien que techniques, sont assez explicites.

Le jeu de données est de petite taille, il y a en effet seulement 310 individus. Cela peut surprendre à l'ère du *Big Data*, c'est en fait lié à la nature des données. Les données médicales sont complexes à recueillir dans un cadre aussi spécialisé que l'orthopédie lombaire. Le recours à l'imagerie médicale (radiographie entre autres) est nécessaire pour cette collecte. Il est important de signaler que les données ont été recueillies par un seul chercheur au sein d'un seul établissement chirurgical. Par ailleurs, l'espace des variables n'est pas de très grande dimension (7 variables). Comparativement au nombre de variables, le nombre d'individus est donc satisfaisant pour conduire des analyses.

3.3 Deux modes de classement différents

Nous disposons donc de deux jeux de données qui diffèrent uniquement par les étiquettes associées aux individus. Dans le premier jeu de données, les individus sont répartis en trois classes tandis qu'ils sont classés en deux classes dans le second.

Les trois classes distinctes du premier jeu de données sont :

- Les patients ne présentant pas de pathologie de la colonne lombaire. (étiquetés NO pour *Normal*)
- Les patients atteints d'une hernie discale. (étiquetés DH pour *Disk Hernia*)
- Les patients atteints de la spondylolisthésis.(étiquetés SL pour *Spondylolisthesis*)

Dans le second jeu de données, les individus sont identiques mais ceux présentant une affection de la colonne lombaire sont regroupés en une seule classe (étiquetés AB, pour *Abnormal*).

Nous adopterons occasionnellement le raccourci individus normaux/individus anormaux afin de désigner respectivement les individus atteint d'une pathologie de la colonne lombaire (spondylolisthésis ou hernie discale) et ceux ne présentant aucune des deux pathologies.

3.4 Explication des variables

Hormis la variable qualitative nominale qui représente l'étiquette de classe, toutes les variables sont quantitatives continues. Nous disposons des variables suivantes :

- incidence pelvienne (*pelvic_incidence*) : mesure de l'embase sur laquelle repose la colonne vertébrale.
- inclinaison pelvienne (*pelvic_tilt*) : mesure de l'inclinaison du sacrum par rapport à la verticale.
- angle de la lordose lombaire (*lumbar_lordosis_angle*) : mesure de l'angle de

la courbure vers l'avant du corps au niveau des vertèbres lombaires (lordose lombaire).
— inclinaison du sacrum (*sacral_ slope*) : mesure l'inclinaison de l'os sacrum (pelvis) par rapport à l'horizontale.
— angle de la lordose pelvienne (*pelvic_ radius*) : mesure de l'angle de la courbure vers l'arrière du corps au niveau de l'os sacrum (lordose pelvienne).
— degré de spondylolysthésis (*degree_ spondylolisthesis*) : évaluation du degré auquel l'individu est atteint de la spondylolisthésis.

En dehors du degré de spondylolisthésis et de l'étiquette de classe, toutes les variables des jeux de données sont des angles mesurés en degrés. Cela se remarque par une analyse anatomique de la façon dont sont prises les mesures notamment sur la figure 3.

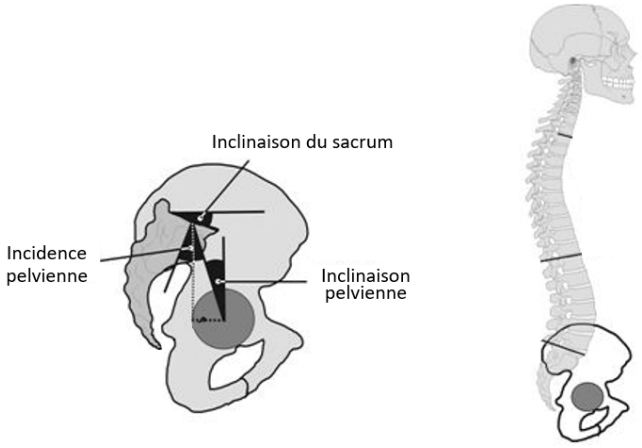

FIGURE 3 – Visualisation des emplacements de mesure de certaines variables du jeu de données

3.5 Problématique

L'objectif ici est de prédire pour un individu donné s'il présente ou non une pathologie de la colonne lombaire, et si c'est le cas, de diagnostiquer de quelle pathologie il est atteint entre hernie discale et spondylolisthésis. Une telle prédiction peut être un outil supplémentaire et une aide à la décision pour le professionnel de santé.

4 Analyse exploratoire des données

4.1 Analyse descriptive

Une première analyse du jeu de données nous montre qu'il est constitué de 310 individus répartis en deux ou trois classes comme suit 1.

TABLE 1 – Effectifs des différentes classes du jeu de données

	Anormaux		Normaux
	Spondylolisthésis	Hernie discale	
Effectif	150	60	100

Les individus sont classés par ordre de classe dans le jeu de données, il faut donc être vigilant et bien mélanger les données afin de les exploiter dans certaines situations.

En dehors de l'étiquette de classe, toutes les variables sont quantitatives continues. Ce sont toutes des angles (exprimés en degrés), mis à part le degré de spondylolisthésis. L'espace angulaire est particulier car deux valeurs peuvent faire référence au même angle (-10°et 350°par exemple). Les angles sont compris entre -6,555°et 163,07°, cette contrainte peut donc être écartée. Les diagrammes en boîte B permettent d'avoir un aperçu de la distribution des données.

Les variables sont quantitatives et leur nombre est raisonnable, on peut donc réaliser un graphe de dispersion, non présenté ici pour des raisons de lisibilité. Celui-ci semble montrer quelques légères corrélations linéaires entre les variables. On visualise plus facilement les corrélations linéaires remarquables sur le graphique associé à la matrice des corrélations de Pearson 4.

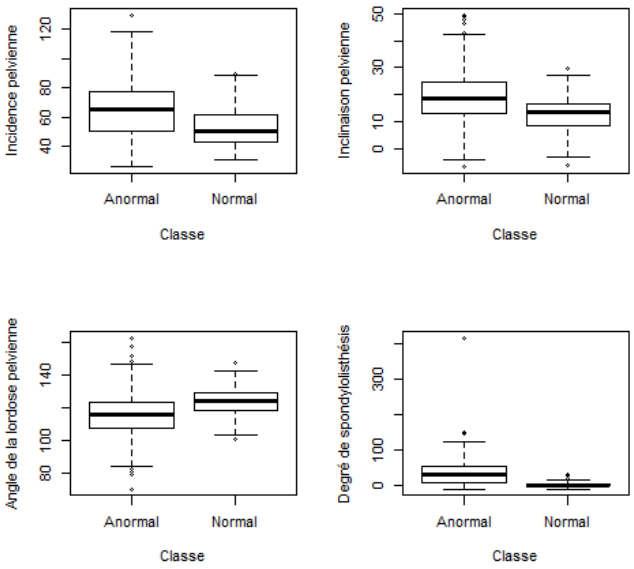

FIGURE 5 – Diagrammes en boîtes de certaines variables quantitatives en fonction de la classe

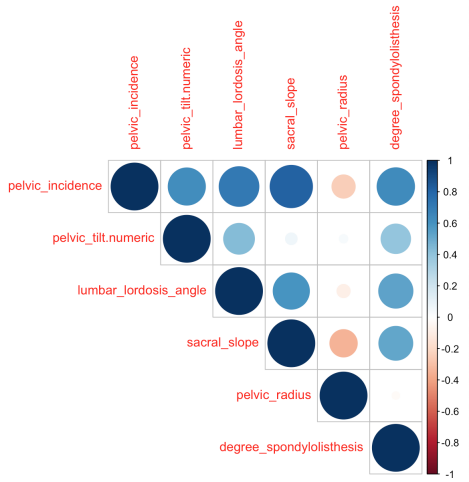

FIGURE 4 – Matrice des corrélations de Pearson

4.2 Analyse du jeu de données à deux classes

On peut également tracer la distribution des valeurs des variables quantitatives en fonction de la variable qualitative de classe sur 5. Les diagrammes en boîte pour les autres variables sont présentés en annexe C.

On remarque que l'étendue interquartile est plus importante pour les individus atteints d'une pathologie par rapport à ceux n'en présentant pas. Il y a des différences en termes de distribution, notamment au niveau de la médiane. Le degré de spondylolisthésis est très proche de 0 pour les individus normaux, c'est cohérent avec le sens de la variable. Il y a quelques valeurs aberrantes, en particulier une valeur très élevée de degré de spondylolisthésis.

Ce jeu de données ne permet pas de répondre totalement à la problématique car, en cas de pathologie, il faut diagnostiquer entre hernie discale et spondylolisthésis.

Pour cela, nous réalisons une analyse similaire en présence des trois classes.

4.3 Analyse du jeu de données à trois classes

Les diagrammes en boîte de la figure 6 et de l'annexe D montrent des différences qui semblent significatives au niveau de la distribution des données. Les individus qui ont une hernie discale et les individus normaux sont plus proches que ceux atteint de spondylolisthésis. Cela se remarque en particulier pour le degré de spondylolisthésis et l'incidence pelvienne.

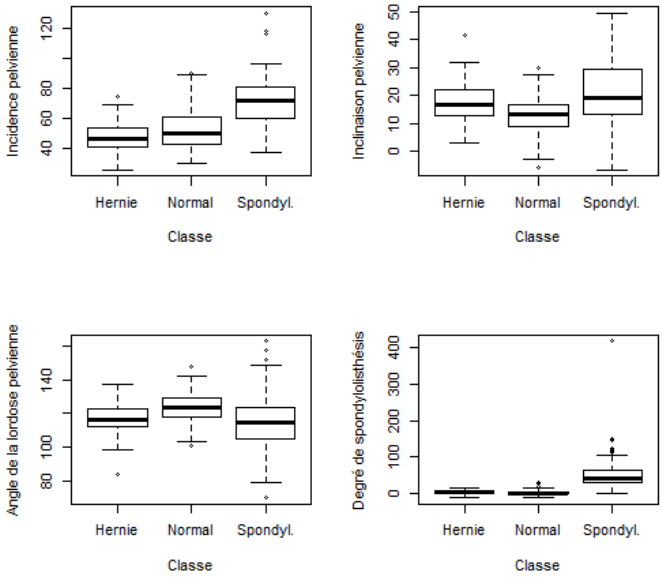

FIGURE 6 – Diagrammes en boîtes de certaines variables quantitatives en fonction de la classe

D'après les représentations graphiques, il semble y avoir un lien entre la classe et les valeurs des autres variables quantitatives. Un test ANOVA (analyse de la variance) est intéressant afin d'étudier l'appartenance aux classes en fonction des variables quantitatives. Néanmoins les hypothèse de ce test sont assez fortes (distribution normale des données dans les classes et égalité des variances dans les différents échantillons).

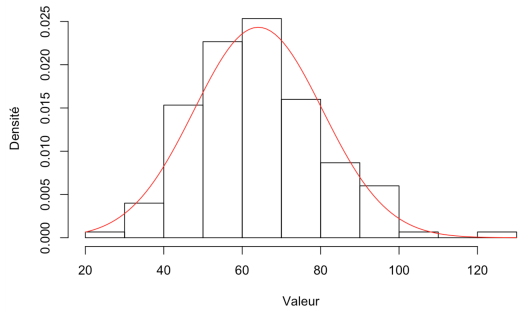

FIGURE 7 – Distribution d'un échantillon sur le jeu de données

Afin de réaliser un test ANOVA, nous vérifions si l'ensemble des hypothèses du test sont satisfaites (normalité et égalité des variances). Par un test de Shapiro-Wilk sur les différents échantillons, nous constatons que certains vérifient l'hypothèse de normalité comme le montre la figure 7 et d'autres non. Par mesure de précaution, nous réalisons alors un test de Kruskal-Wallis aux hypothèses moins fortes afin de limiter au mieux les risques de faux positif. Le test nous permet alors de conclure que les distributions des différents échantillons sont différentes en fonction du groupe (Normal, Hernie, Spondylolisthésis). La classe influence donc significativement les autres variables.

Le jeu de données à deux classes ne présentent pas d'intérêt majeur par rapport à celui à trois classes. En effet, il ne répond que partiellement à la problématique, il ne présente aucune information supplémentaire par rapport au jeu à trois classes. Il faut juste noter que la population des individus atteints de spondylolisthésis et celle des individus atteints d'une hernie discale constituent la population des individus atteints d'une pathologie.

5 Analyse en composantes principales

5.1 Motivations

L'objectif de l'analyse en composantes principales (ACP) est ici de représenter les différents individus dans un espace de plus faible dimension que l'espace initial des variables (à 6 dimensions, en excluant la variable de classe). On peut utiliser l'ACP quand toutes les variables sont quantitatives. Cette méthode est donc particulièrement intéressante dans notre cas. Nous allons en effet réaliser une ACP afin de chercher à obtenir une représentation simplifié du nuage d'individus : par exemple dans un plan.

5.2 ACP sur le jeu de données à deux classes

A partir de l'analyse des données et de leur sémantique, nous avons déterminé que nous ne sommes pas en présence d'un effet taille. En effet, toutes les mesures anatomiques sont dans l'espace angulaire et l'influence de la taille de l'individu est négligeable, beaucoup plus que si nous avions des mesures linéaires. Il n'y a donc pas de traitement préalable à effectuer.

On étudie l'inertie expliquée par les différents axes factoriels issus de l'ACP. La méthode du coude (cf figure 8) nous permet de déterminer le nombre d'axes à retenir.

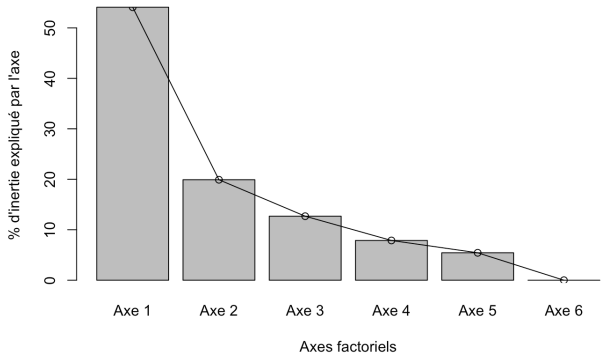

FIGURE 8 – Pourcentages d'inertie expliquée par les différents axes factoriels

Par cette méthode, nous retenons les deux premiers axes qui forment le premier plan factoriel. L'étude de l'inertie expliquée par les axes nous indique que le premier plan factoriel explique 74 % de l'inertie du nuage de point. Cela nous permet de restituer 74 % de l'information dans le plan 14, beaucoup plus facile à visualiser qu'un espace à six dimensions.

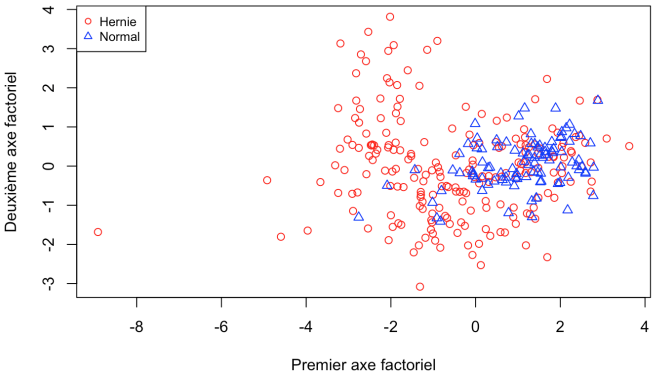

FIGURE 9 – Premier plan factoriel de l'ACP

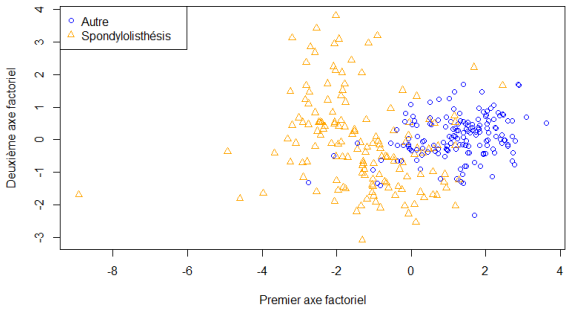

FIGURE 11 – ACP sur les données à trois classes en distinguant la classe Spondylolisthésis des autres classes

5.3 ACP sur le jeu de données à trois classes

5.3.1 Distinction entre les trois classes sur le premier plan factoriel

Après cette première ACP sur les données à deux classes, nous nous sommes demandés si les données à trois classes ne seraient pas plus simples à exploiter.

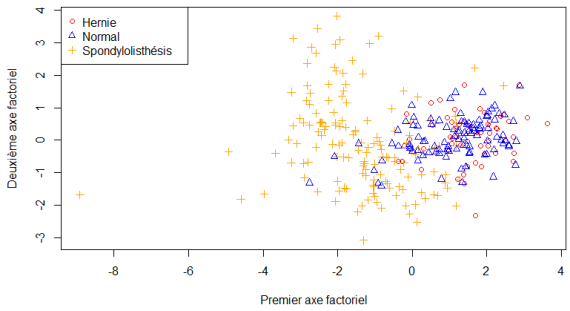

FIGURE 10 – ACP sur les données à trois classes

Nous avons alors remarqué que la classe Spondylolisthésis se distingue des deux autres classes. En revanche, comme pressenti lors de l'analyse exploratoire, la classe Hernie ne se distingue presque pas de la classe Normal ici. Nous avons alors essayé de garder la même ACP mais en distinguant le Spondylolisthésis du reste des données (classes Hernie et classe Normal confondues).

Sur la Figure 11, nous pouvons constater que les individus atteints de la spondylolisthésis se distinguent globalement assez facilement des autres individus.

Pour la Hernie, cela semble plus compliqué. Une autre question qui nous est alors venue est à propos des variables explicatives : peut-être que certaines variables ne permettent pas de distinguer la Hernie de Normal tandis que d'autres oui. Nous avons alors fait des tests d'homogénéité sur les variables entre ces deux classes. Il en ressort que seules 4 des 6 variables sont réellement différentes entre les classes Normal et Hernie. En effet, le degré de spondylolisthésis et l'incidence pelvienne ne sont pas explicatives pour distinguer entre la classe Normal et Hernie (tests de Student). C'est tout à fait cohérent avec le sens de la variable "degré de spondylolisthésis". Pour l'incidence pelvienne, on peut intuitivement le comprendre par le peu d'impact que l'hernie discale A a a priori sur cette mesure 3.

5.3.2 Sélection de variables pour la distinction entre les classes Hernie et Normal

Nous avons alors repris l'ACP en ignorant les individus atteint de spondylolisthésis, et en ne prenant en compte que les quatre variables significatives.

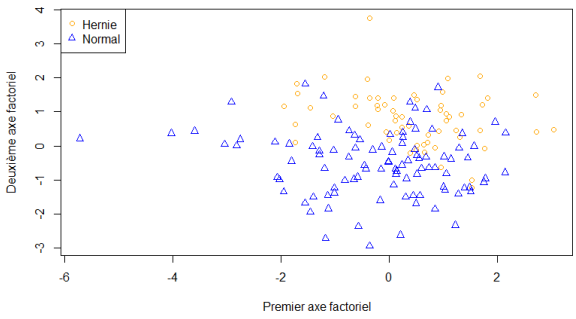

FIGURE 12 – ACP sur les données à trois classes en sélectionnant quatre variables sur six et en ignorant les individus atteint de spondylolisthésis

Sur la figure 12, nous voyons qu'il est encore compliqué de distinguer clairement les deux zones (Hernie et Normal) mais la frontière est plus simple qu'auparavant.

Nous avons ainsi décidé de continuer nos analyses et notamment la construction de nos classifieurs sur le fichier avec trois classes et non celui avec deux classes.

Ces représentations des données guident nos choix de modèles pour la suite de l'étude.

Les données présentes des étiquettes de classe, nous nous situons donc dans un problème de discrimination faisant appel à de l'apprentissage statistique supervisé. Nous devons expliquer la variable de classe. Pour cela, l'algorithme des K plus proches voisins nous a paru particulièrement adapté.

6 Méthode des K plus proches voisins

6.1 Principe de l'algorithme

L'algorithme des K plus proches voisins (ou K-NN pour *K-nearest neighbors*) est l'un des algorithmes d'apprentissage statistique les plus connus et les plus répandus. C'est un algorithme d'apprentissage supervisé, il permet de discriminer des données, c'est-à-dire de déterminer leurs classes. Il consiste à affecter le vecteur forme $\boldsymbol{x}$ à la classe w_k la plus représentée parmi celles des K plus proches voisins de $\boldsymbol{x}$. K est un paramètre donné en entrée de l'algorithme.

Nous partitionnons aléatoirement le jeu de données en ensemble d'apprentissage, de validation et de test. L'ensemble d'apprentissage (ou d'entraînement) permet, comme son nom l'indique, d'entraîner un modèle en apprenant sur les données. L'ensemble de validation, quant a lui, est ici utilisé afin d'optimiser l'hyper-paramètre K. Enfin, lorsqu'on a déterminé le K optimal (notons le K*), nous estimons les performances en calculant le taux de mauvais classement (taux d'erreur) sur l'ensemble de test en faisant tourner l'algorithme. Le fait d'utiliser trois ensembles distincts nous permet d'obtenir une estimation non biaisée du taux d'erreur sur l'ensemble de test.

6.2 Motivations

Cet algorithme s'applique particulièrement bien à des variables quantitatives continues, ce qui est le cas ici.

6.2.1 Une distribution de classes équilibrée

Les classes sont du même ordre de grandeur en termes d'effectifs. Cet équilibre permet d'appliquer l'algorithme des K plus proches voisins. Le cas contraire, la classe d'effectif majoritaire aurait tendance à l'emporter très souvent comme expliqué ci-dessous.

> *Une faiblesse de la classification basique par "vote majoritaire" apparaît quand la distribution de classe est asymétrique. C'est-à-dire, des exemples d'une classe plus fréquente tendent à dominer la prédiction de classification du nouvel entrant, car elle serait statistiquement plus fréquente parmi les k plus proches voisins par définition.* Wikipédia [4]

6.2.2 Un faible volume de données

Cet algorithme est coûteux car il requiert à la fois, le calcul de toutes les distances entre les individus de l'ensemble de test et de l'ensemble d'apprentissage, ainsi que leur stockage en mémoire. Néanmoins, nos données sont de faible volume, nous pouvons donc utiliser cet algorithme avec des ressources modérées et en un temps raisonnable. De par la complexité de la collecte des données, il n'y a pas de risques d'explosion du volume du jeu de données en cas de nouveaux individus.

6.3 Sélection de l'hyper-paramètre K

Afin de déterminer l'hyper-paramètre K minimisant le taux d'erreur de classification, nous effectuons la méthode des K-NN sur un ensemble de validation. Différents K sont testés (de 1 à 20 en évitant les K pairs afin d'éviter les situations d'ex-aequo), le K* retenu est celui ayant généré le taux d'erreur minimal sur l'ensemble de validation.

Le taux d'erreur de prédiction du modèle est estimé avec l'ensemble de test, il varie selon le partitionnement effectué. Afin de mesurer la qualité de notre modèle, nous l'appliquons donc avec plusieurs partitionnements différents (en trois ensembles). Ainsi on apprend sur l'ensemble d'apprentissage, on détermine K* avec celui de validation et on estime le taux d'erreur des K*-NN avec celui de test. La qualité globale du modèle est caractérisée par la moyenne des taux d'erreurs associés aux différents K* (pour chaque partitionnement). Un exemple du taux d'erreur moyen pour chaque K dans ce processus peut être visualisé sur la figure 13. La mesure du taux d'erreur effectuée sur l'ensemble d'apprentissage souligne le fait que l'on a plus tendance à surapprendre le modèle lorsque K est petit (notamment K=1), mais ce n'est pas toujours le cas.

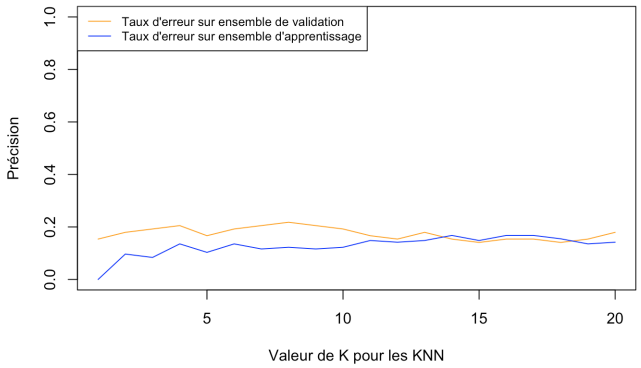

FIGURE 13 – Mesure de la qualité du modèle pour différentes valeurs de K (ici K* = 15) pour un ensemble d'apprentissage et un ensemble de validation donné

Étant donné la dimension de l'espace des individus par rapport à celui des variables, l'utilisation d'un ensemble de validation pour déterminer le K optimal est convenable. Cependant, il existe des méthodes moins sensibles au manque de données. La validation croisée, que nous n'avons pas eu le temps de mettre en oeuvre, est une méthode intéressante. Le principe est le suivant (*"k-fold cross-validation"*) :

> *La méthode consiste à diviser l'échantillon original en k échantillons, puis à sélectionner un des k échantillons comme ensemble de validation et les $k-1$ autres échantillons constitueront l'ensemble d'apprentissage. On calcule comme dans la première méthode le score de performance, puis on répète l'opération en sélectionnant un autre échantillon de validation parmi les $k-1$ échantillons qui n'ont pas encore été utilisés pour la validation du modèle. L'opération se répète ainsi k fois pour qu'en fin de compte chaque sous-échantillon ait été utilisé exactement une fois comme ensemble de validation. La moyenne des k erreurs quadratiques moyennes est enfin calculée pour estimer l'erreur de prédiction.*

6.4 Discrimination directe entre trois classes

Une première idée est d'appliquer le classifieur des k plus proches voisins de manière naïve sur le jeu de données à trois classes. La faible dimension de l'espace ($\mathbb{R}^6$) ne rend pas nécessaire une réduction de dimension (par ACP notamment). Ainsi, nous effectuons un partitionnement du jeu de données en trois comme indiqué précédemment puis nous déterminons K* avant de mesurer le taux d'erreur sur l'ensemble de test. En agissant de la sorte, les résultats sont assez satisfaisants, lors d'un premier test nous avons obtenu un taux d'erreur de 15,6 %, puis en mesurant le taux d'erreur moyen sur plusieurs partitionnements, nous obtenons un taux d'erreur moyen d'environ 17,5 %. A chaque itération (et donc partitionnement), un nouveau K optimal est déterminé comme souligné en 6.3. Étant donné le peu de données à notre disposition (310 individus) les résultats semblent acceptables. Néanmoins, dans la lignée de l'étude menée lors de l'ACP en 5.3.2, une amélioration semble possible. En effet le groupe Spondylolisthésis se détache de manière assez distincte des deux autres groupes. Ainsi une classification en deux temps semble être intéressante.

6.5 Amélioration par une méthode en deux étapes

6.5.1 Idée

Suite à nos analyses lors de l'ACP, nous avons poussé la réflexion un peu plus loin que l'algorithme K-NN naïf, et avons cherché à distinguer dans un premier temps le Spondylolisthésis des autres classes. Une fois cette distinction faite, pour toutes les données détectées comme n'étant pas un Spondylolisthésis, nous avons refait une classification entre les classes Normal et Hernie. Le principe de la méthode est présenté sur la figure suivante :

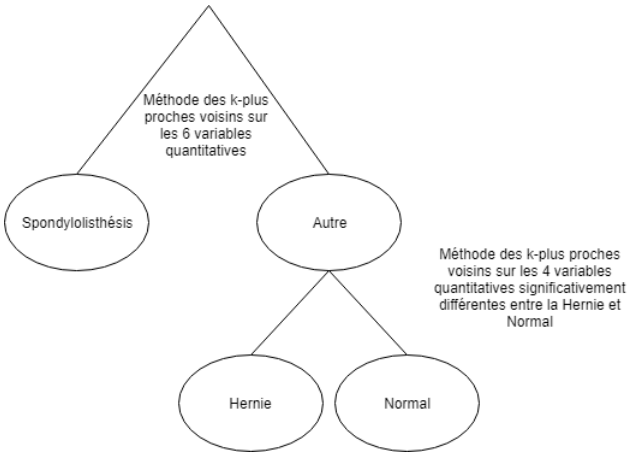

FIGURE 14 – Représentation graphique de la méthode des k-plus proches voisins amélioré pour notre jeu de données

L'algorithme prendra alors en compte les six variables quantitatives pour la distinction du Spondylolisthésis et seulement les quatre variables explicatives pour la Hernie (exclusion de degré de spondylolisthésis et de l'incidence pelvienne). Les individus classés comme atteint de spondylolisthésis sont exclus de la seconde phase.

Cette distinction va permettre d'avoir une méthode des k-plus proches voisins plus performante. En effet, plus l'espace est petit et plus le nombre faible de données que nous avons rendra le classifieur performant.

6.5.2 Résultats

En exécutant cet algorithme, et en utilisant l'optimisation de l'hyper-paramètre faite précédemment, nous obtenons un taux d'erreur de 14,6 % en moyenne. Ce classifieur semble plus performant par rapport au K-NN naïf, il faut cependant se méfier et savoir quel est le type d'erreur car dans le domaine médical, les conséquences d'une erreur de diagnostic peuvent être graves.

Il se trouve qu'en regardant les données plus en détail, il y a davantage de faux positifs que de faux négatifs, ce qui signifie que l'on a plus tendance à diagnostiquer des pathologies sur des individus à la colonne lombaire normale que l'erreur inverse. Il est préférable que ce soit dans ce sens là car une personne détectée avec une pathologie recevra des examens plus avancés qui permettrons d'infirmer ou confirmer ce résultat.

Nous avons souhaité compléter cette étude par l'utilisation d'autres classifieurs, assez répandus dans le domaine médical.

7 Analyse discriminante

7.1 Classifieur bayésien naïf

7.1.1 Principe

Pour mettre en place un classifieur bayésien naïf :
- On détermine un ensemble d'apprentissage
- On détermine des probabilités à priori de chaque classe (en observant les effectifs par exemple)
- On applique le théorème de Bayes :

$$\mathbb{P}(Z \mid X) = \frac{\mathbb{P}(X \mid Z).\mathbb{P}(Z)}{\mathbb{P}(X)} \tag{1}$$

- On choisit la classe la plus probable. Pour ce faire, nous devons estimer $\mathbb{P}(X|Z)$. On note Z la variable aléatoire représentant la classe et X celle représentant les observations.

7.1.2 Propriétés intéressantes

Le fait de faire l'hypothèse :

$$\mathbb{P}(X \mid w_k) = \mathbb{P}(w_k).\frac{\prod_{j=1}^{p} \mathbb{P}(X_j \mid w_k)}{\mathbb{P}(X)} \tag{2}$$

permet de limiter les paramètres à estimer. En effet, l'hypothèse d'indépendance des variables X_j conditionnellement à Z permet de se contenter de la variance de chacune d'entre elles dans chaque classe sans calculer à proprement parler de matrice de covariance $\sum$.

Le classifieur de Bayes possède la propriété intéressante d'avoir le plus faible taux d'erreur théorique. L'intérêt d'utiliser un tel classifieur malgré les hypothèses d'indépendance assez fortes est qu'il requiert relativement peu de données d'entraînement pour estimer les paramètres nécessaires à la classification (moyennes et variances des différentes variables).

7.1.3 Résultats

En appliquant le modèle à nos données, nous obtenons sur plusieurs partitionnements un taux d'erreur moyen de 18 %. Ce résultat ne semble pas satisfaisant par rapport aux modèles précédents, essayons de l'expliquer. La classification naïve bayésienne présuppose l'indépendance conditionnelle des variables. En y regardant de plus près, on constate qu'au sein d'une même classe, la matrice de covariance n'est pas diagonale, ainsi l'hypothèse n'est pas respectée. De même, ce classifieur suppose également que la distribution est normale au sein des classes, ce qui n'est pas le cas pour toutes les

variables comme vu précédemment lors de l'analyse exploratoire. Ceci peut alors expliquer un tel résultat. On peut également expliquer ce résultat par la faible quantité de données disponible, ne permettant donc pas une estimation précise des probabilités à priori. En retirant les variables les plus corrélées entre elles, le classifieur passe à un taux d'erreur de 16,9 %.

Il peut ainsi être intéressant de faire appel à un modèle avec des hypothèses moins fortes.

7.2 Analyse discriminante linéaire

7.2.1 Principe

Ici, contrairement au classifieur bayésien naïf, nous ne supposons pas l'indépendance conditionnelle. On pose néanmoins une hypothèse d'homoscédasticité, c'est-à-dire qu'on suppose la matrice de covariance égale d'une classe à l'autre.

7.2.2 Résultats

Le modèle d'analyse discriminante linéaire (ADL) donne un taux d'erreur moyen de 19 %. Ce résultat assez mauvais peut s'expliquer par le fait que les matrices de covariance sont évidemment différentes entre les classes. L'hypothèse d'homoscédasticité est largement violée. Un avantage de cette méthode est qu'elle reste robuste au non respect de ses hypothèses comme cela à pu être montré dans la publication scientifique *Robust Linear Discriminant Analysis* [5]. Par curiosité, nous explorons également un classifieur encore plus général.

7.3 Analyse discriminante quadratique

7.3.1 Principe

Avec l'analyse discriminante quadratique (ADQ), on ne fait pas d'hypothèse sur les matrices de covariance ou sur l'indépendance conditionnelle. Ainsi, les matrices de covariance $\sum_k$ peuvent différer entre les classes. On suppose simplement que le vecteur de caractéristique $\mathbf{X}$ suit, conditionnellement à chaque classe w_k une loi normale multidimensionnelle.

7.3.2 Résultats

Ainsi, le cas général nous donne un taux d'erreur moyen de 15,6 %. Ce résultat est meilleur que celui obtenu avec le classifieur bayésien naïf ou avec l'analyse discriminante linéaire. Cela semble logique dans la mesure où ce modèle est plus général et fait moins d'hypothèses. En contrepartie, il faut estimer davantage de paramètres, ce qui peut conduire à des estimations imprécises quand le volume de données n'est pas suffisant. Ainsi, le modèle est plus flexible car moins rigide en termes d'hypothèses, mais les erreurs d'estimation peuvent au final s'avérer plus contraignantes que ces hypothèses plus ou moins fortes.

8 Comparaison entre les modèles

Nous avons pu étudier ici différents modèles afin d'effectuer la discrimination entre les individus atteints d'une hernie discale, ceux atteints de spondylolisthésis et ceux ne présentant aucune pathologie.

D'après la table 2, le classifieur des K-NN amélioré par un méthode à 2 étapes en 6.5 est celui qui donne le taux d'erreur moyen le plus faible.

TABLE 2 – Taux d'erreur moyen mesurés pour chacun des classifieurs étudiés dans le document

K-NN	K-NN amélioré	Bay. naïf	ADL	ADQ
17,5 %	14,6 %	18 %	19 %	15,6 %

Le classifieur des K-NN amélioré propose à la fois le meilleur résultat et une cohérence par rapport aux données. L'algorithme se base en effet sur notre observation des données et se prête bien à un volume plutôt faible de données ainsi qu'à un équilibre dans les effectifs des classes.

Au niveau des modèles d'analyse discriminante, l'analyse discriminante quadratique propose les meilleurs résultats. Cependant, il ne faut pas juger la qualité d'un modèle uniquement à la lumière de ses résultats. Malgré sa performance, ce modèle exige l'estimation de nombreux paramètres. Or, nous ne disposons pas d'un gros volume de données, les estimations de ces paramètres ne sont donc pas très précises. On lui préfère donc le classifieur bayésien naïf ou l'analyse discriminante linéaire qui, malgré des taux d'erreur plus élevés, comblent la faiblesse de l'ADQ par des hypothèses simplifiant le modèle. Ces hypothèses ne sont pas totalement respectées mais les modèles restent robustes en fournissant des résultats du même ordre de grandeur. On pourrait encore améliorer les résultats en effectuant un travail supplémentaire pour limiter au maximum les corrélations entre les variables et ainsi être davantage en adéquation avec les hypothèses du modèle naïf de Bayes (comme montré en 7.1.3) ou de l'analyse discriminante linéaire.

Par ailleurs, les modèles doivent être considérés au regard de leur contexte médical. Il est donc important de contrôler les aspects de faux-positifs et faux-négatifs. Il parait en effet inconcevable d'indiquer à un individu atteint d'une pathologie qu'il est totalement sain. Parmi les modèles satisfaisants évoqués ci-dessus, un prolongement de l'étude tendrait à privilégier les modèles limitant en priorité le nombre de faux négatifs.

9 Conclusion

Au cours de l'analyse de ce jeu de données, nous avons pu développer des connaissances dans un domaine que nous ne connaissions absolument pas auparavant : l'orthopédie lombaire.

Les recherches effectuées à ce sujet nous ont permis de comprendre le sens des données et de prendre plaisir à les analyser dans un objectif de diagnostic médical.

Nous avons du pousser nos réflexions pour obtenir les meilleurs diagnostics possibles, et surtout pour proposer des modèles compréhensibles et adaptés aux données. Nous avons découvert ce deuxième aspect grâce au projet.

Nous avons ainsi été confrontés à des problématiques des métiers de la donnée, et nous sommes rendu compte de l'importance des hypothèses de chaque modèle. Quel que soit le modèle et son taux d'erreur, celui-ci ne remplace pas le diagnostic d'un expert mais peut lui être complémentaire.

Les modèles proposés ont encore un taux d'erreur assez élevé face aux enjeux du diagnostic médical. En prolongement de l'étude, nous pourrions essayer de diminuer le taux d'erreur en couplant des modèles tels que K-NN et un modèle d'analyse discriminante. Enfin, cette étude pourrait être approfondie en prenant en compte les coûts (physiques, moraux, sociaux, financiers) induits par le type d'erreur.

A Visualisation de l'hernie discale

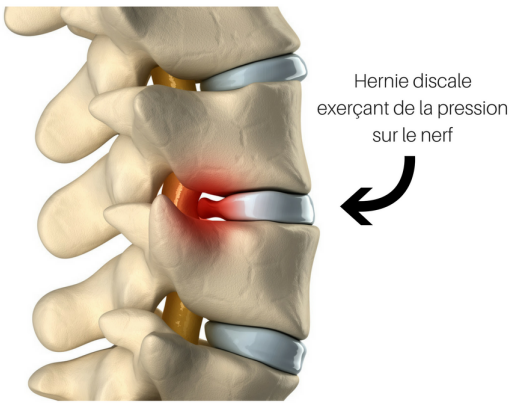

FIGURE 15 – Illustration de l'hernie discale : saillie anormale du noyau des disques intervertébraux

B Distribution de variables angulaires du jeu de données

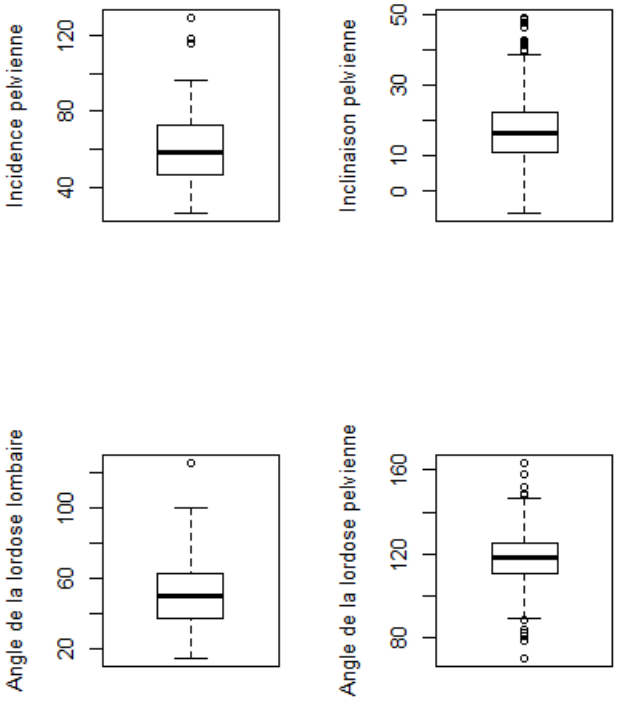

FIGURE 16 – Diagrammes en boîtes du jeu de données global

C Distribution de variables angulaires selon la classe (Anormal/Normal)

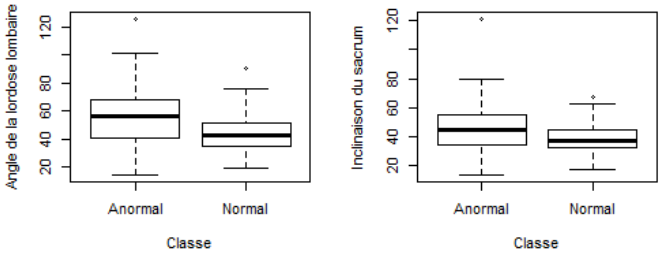

FIGURE 17 – Diagrammes en boîtes de certaines variables quantitatives en fonction de la classe

D Distribution de variables angulaires selon la classe

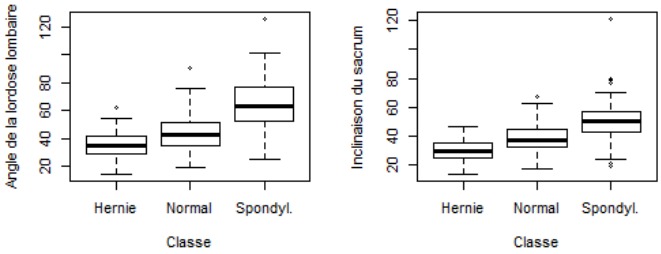

FIGURE 18 – Diagrammes en boîtes de certaines variables quantitatives en fonction de la classe

Références

[1] Dictionnaire Larousse, *Définition de l'orthopédie*. https://www.larousse.fr/dictionnaires/francais/orthop%C3%A9die/56613/

[2] UCI Machine Learning, *Vertebral Column Data Set*. http://archive.ics.uci.edu/ml/datasets/vertebral+column

[3] Chiropracticien.com, *Souffrir d'une hernie discale*. https://www.chiropraticien.com/douleur-hernie-discale/

[4] Wikipédia, *Méthode des k plus proches voisins*. https://fr.wikipedia.org/wiki/M%C3%A9thode_des_k_plus_proches_voisins/

[5] Journal of Mathematics and Statistics, *Robust Linear Discriminant Analysis*. https://thescipub.com/pdf/10.3844/jmssp.2016.312.316?fbclid=IwAR2brn4Mk51Q1NTgsaS-VhjzZaphEHpNnqNkSfhBEVr65TR2VuhjM6dvzhA/