Analyse et apprentissage sur le jeu de données "Palmer Penguins"

Destephen Bastien, Doan Nhat-Minh, Garnier Kilian

17 juin 2021

Résumé

Dans le cadre du projet de l'UV SY09, nous avons eu pour mission d'analyser le jeu de données intitulé *Palmer Penguins*, et d'y appliquer les méthodes d'apprentissages vues en cours. Nous pourrons ainsi gagner de l'expérience en opérant sur des données réelles.

Ce rapport détaille notre démarche d'analyse exploratoire, de sélection de variables et d'application de plusieurs modèles d'apprentissages supervisés. Les limites du jeu de données et des approches utilisées seront également décortiquées.

1 Introduction

Cette section introduira au lecteur le jeu de données que nous allons analyser, ainsi qu'une explication des diverses parties contenues dans la suite de ce rapport.

Les données de *Palmer Penguins* ont été collectées de 2007 à 2009 par le Dr. Kristen Gorman, dans le cadre d'une étude menée au *Palmer Long-Term Ecological Research*. Deux versions de ce jeu de données sont disponibles, nous analyserons celle décrite comme "brute", qui contient toutes les données d'origine, afin d'avoir le plus d'informations possibles.

En menant des recherches préliminaires sommaires, une information revient de manière récurrente : les *Palmer Penguins* constituent une excellente alternative au fameux jeu de données *Iris*, que nous avons étudié dans le cadre des TD. La problématique d'*Iris* est de classifier les individus, en l'occurrence les fleurs, en trois espèces en fonction de quatre mesures concernant leur anatomie. La problématique des *Palmer Penguins* sera assez similaire : établir un modèle de classification des pingouins au sein de différentes espèces.

Ce rapport sera divisé en plusieurs parties :

1. La présente introduction qui se terminera par la formulation de la problématique ;
2. La partie principale, articulée selon trois axes : l'analyse exploratoire des données, la sélection des variables pour l'apprentissage, puis l'apprentissage des modèles supervisés vus en cours ;
3. La conclusion qui résumera nos résultats, leurs limites et présentera des perspectives de continuation ;

Voici donc, pour clore cette introduction, une formulation de la problématique associée à notre étude : *Est-il possible de classifier, selon leur espèces, les pingouins de l'étude de Palmer en utilisant les mesures effectuées par les experts ?*

2 Analyse du jeu de données et apprentissage supervisé

Cette partie principale présente un résumé du travail effectué sur les données, ainsi que les modèles d'apprentissage testés et leurs résultats respectifs, qui seront critiqués.

2.1 Pré-traitement

Un jeu de données vient toujours avec ses défauts, redondances et valeurs nulles. Il s'agit ici de traiter ces imperfections avant de commencer tout travail d'analyse.

Palmer Penguins Raw est le jeu de données qui contient toutes les variables de l'étude :

- StudyName : Nom de l'étude ;
- Sample Number : Numéro d'échantillon ;
- Species : Espèce du pingouin ;
- Region : Région de vie ;
- Island : Île de vie ;
- Stage : État de l'oeuf s'il existe ;
- Individual ID : Identifiant individuel ;
- Clutch Completion : Indique si un oeuf du nid a déjà éclot ;
- Date Egg : Date de l'éclosion de l'oeuf ;
- Culmen Length : Longueur du bec en mm ;
- Culmen Depth : Profondeur du bec en mm ;
- Flipper Length : Longueur des nageoires en mm ;

— Body Mass : Masse du pingouin en gramme ;
— Sex : Male ou Femelle ;
— Delta 15 N (o/oo) : Ratio isotopique d'hydrogène dans le sang ;
— Delta 13 N (o/oo) : Ratio isotopique de carbone dans le sang ;
— Comments : Commentaires de l'expert concernant l'individu.

Avant d'explorer ces données, il faut les rendre exploitable et s'assurer d'éliminer le superflu ou les redondances. Les paragraphes suivants détaillent les traitements effectués.

Suppression de variables :

1. *StudyName* est supprimée : il existe trois îles, et trois études, dont chacune porte sur une île spécifique. Redondance. ;
2. *Region* et *Stage* sont supprimées, car elles ne prenaient qu'une unique valeur pour tous les individus. ;
3. *IndividualID* et *Sample Number* sont des identifiants n'apportant aucune information. ;
4. *Comments*, inutile une fois exploitée (voir le paragraphe suivant). ;

Traitement des valeurs nulles :

1. *Longueurs* : il existe deux individus pour lesquels aucune longueur n'est renseignée. En vérifiant les commentaires de l'expert dans *Comments*, on s'aperçoit qu'il a lui même notifié que les données n'ont pas pu être recueillies, ces individus ne représentent donc rien de concret. Ils sont supprimés du jeu de données. ;
2. *Sex* : introduction de la valeur *UNKNOWN* pour ne pas introduire de biais.
3. *Delta 15N, Delta 13N* : remplacement par la moyenne de la colonne, calculée à partir des pingouins de même espèce. ;

Pour mener à bien l'analyse exploratoire, nous avons fait le choix de convertir les variables qualitatives *Island* et *Species* en variables quantitatives, en utilisant la procédure *Label Encoding*. Elle consiste à remplacer chaque valeur de la variable par un nombre entier, dans notre cas parmi 0, 1, 2, car elles ne prennent chacune que trois valeurs possibles. On précisera qu'ici il n'y a aucune notion d'ordre entre les valeurs, cette manipulation permettant uniquement une analyse plus complète de la variance des variables.

2.2 Analyse exploratoire des données

L'analyse exploratoire des données est une phase cruciale : il s'agit de décortiquer le jeu de données pour en avoir une meilleure compréhension et relever des indices permettant d'aiguiller les choix effectués dans les parties qui suivront. Cette partie peut-être très longue, ainsi nous ne présenterons que les informations essentielles que nous en avons extrait.

Il y a 342 individus dans le jeu de données, ce qui est assez peu. La répartition mâles / femelles au sein des sexes est homogènes, là ou la distribution des pingouins au sein des espèce ne l'est pas : les *Chinstrap* sont moins présents que les *Adelie, Gentoo*. Ce fait pourrait introduire un biais lors de l'apprentissage.

Seuls les pingouins de l'espèce *Adelie* vivent sur les trois îles de l'étude, les *Chinstrap* étant présents uniquement sur *Dream* et les *Gentoo* sur *Biscoe*. Cette information est précieuse pour la phase d'apprentissage.

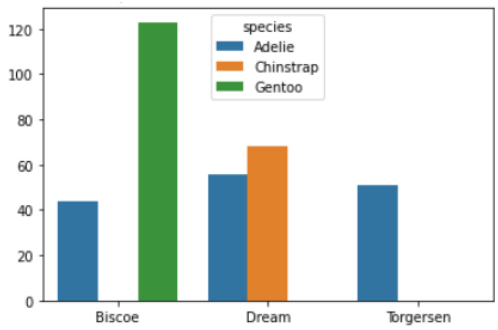

FIGURE 1 – Diagramme en barres des distribution des espèces par île.

Pour les variables quantitatives de longueurs et de masse, les mâles ont souvent des valeurs plus élevées que les femelles, et ce peu importe l'espèce.

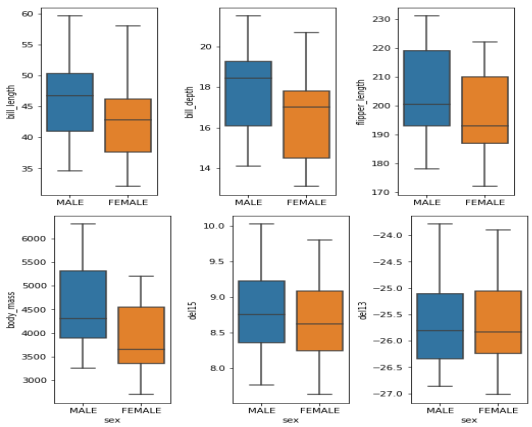

FIGURE 2 – Diagramme en boîte des données quantitatives en fonction du sexe.

Les pingouins *Chinstrap* et *Gentoo* ont le bec plus long, les *Adelie* et *Gentoo* ont un bec plus profond. On remarquera qu'en général, les *Gentoo* ont des valeurs plus élevées partout, ce qui devrait bien les séparer des deux autres.

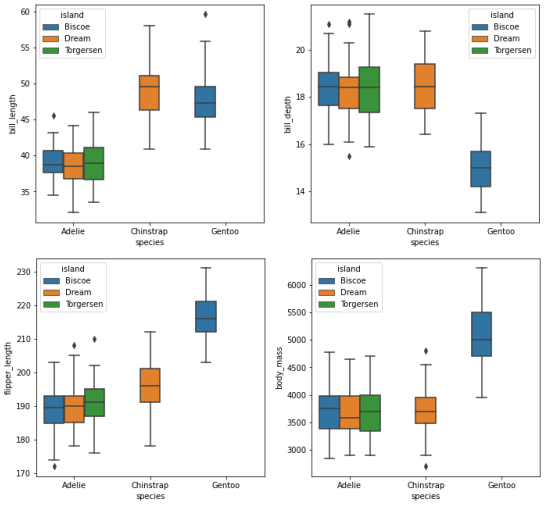

FIGURE 3 – Comparaison des différentes caractéristiques par île et par espèce

Pour analyser les variables deux-à-deux, nous avons fait le choix de commenter le graphique matriciel ci-dessous.

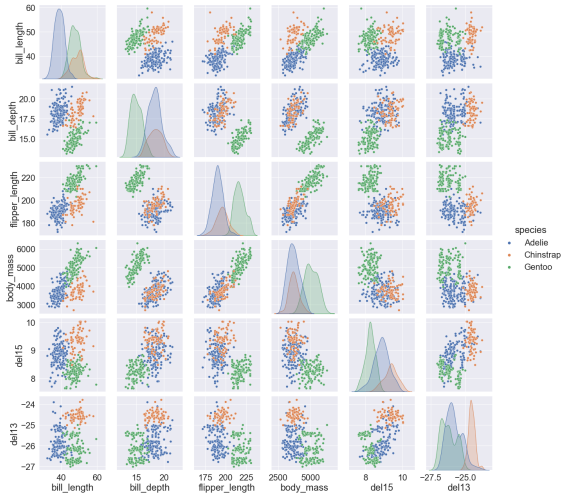

FIGURE 4 – Graphique matriciel des données quantitatives, en fonction de l'espèce.

L'information principale émergent de cette figure est donnée par la teinte des individus en fonction de leur espèce : selon le nuage de point considéré, on peut observer une séparation très nette des individus en trois groupes, qui coïncident avec la couleur des espèces. Les pingouins *Gentoo* semblent se séparer de ceux des deux autres espèces, prenant des valeurs plus élevées, par exemple sur le graphique présentant la masse en fonction de la taille des nageoires, confirmant les indices relevés précédemment.

Ces constats sont très positifs car on parvient à identifier des classes nettes qui correspondent aux trois espèces existantes.

Poursuivons avec l'analyse des variances et des corrélations entre les variables. On s'intéresse aux variables fortement corrélées avec *Species*, car c'est celle qu'on cherchera à prédire.

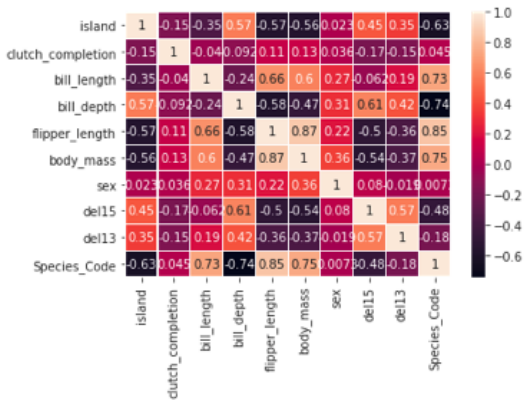

FIGURE 5 – Table des corrélations entre les variables.

Sex et *Clutch Completion* ont une corrélation quasi-nulle à *Species*. Intuitivement, on se doutait déjà que le sexe d'un pingouin ou le fait qu'un des oeufs de sa ponte ait éclot aurait assez peu de signification pour nous aider à prédire l'espèce. Il a été mentionné plus haut que le ratio mâles / femelles était très équilibré, apportant donc peu d'informations pour notre problème.

On remarque certaines corrélations assez fortes, par exemple entre *Species* et *Flipper lentgh*, ou encore entre *Flipper lentgh* et *Body Mass*. Ces constats sont assez intuitifs : un pingouin plus "grand" sera plus lourd.

Procédons maintenant à une analyse de la variance pour sélectionner les meilleures variables pour prédire *Species*. Pour ce faire, on utilise une procédure d'analyse de la variance et le test F.

L'analyse de la variance (ANOVA) a pour objectif de déterminer si les moyennes de chaque variable est différente au sein des classes d'individus, classes formées par les individus prenant tous la même modalité pour la variable à prédire, ici *Species*. Nous avons utilisé l'ANOVA

couplée aux *tests F* ; le résultat obtenu est un *F-Score* dont la signification est simple : plus il est élevé, plus la variable concernée sera utile pour prédire la variable à expliquer. Les paragraphes suivants expliquent la procédure mathématique.

F = Variation entre les moyennes d'échantillonnage / Variation à l'intérieur des échantillons

Numérateur : Variation entre les moyennes d'échantillonnage

Pour mesurer la variance entre les moyennes de l'échantillon, nous utiliserons le carré moyen ajusté, qui est la somme des écarts à la moyenne au carré, divisée par le nombre de degrés de liberté ; plus les moyennes des groupes sont écartées les unes des autres, plus cette variance est importante.

Dénominateur : Variation à l'intérieur des échantillons

Pour calculer cette variance, nous devons calculer à quelle distance chaque observation est de sa moyenne de groupe, c'est la somme des écarts au carré de chaque observation de la moyenne de son groupe divisé par le degré de liberté de l'erreur. Si les observations pour chaque groupe sont proches de la moyenne du groupe, la variance à l'intérieur des échantillons est faible. Toutefois, si les observations pour chaque groupe sont plus éloignées de la moyenne du groupe, la variance à l'intérieur des échantillons est plus élevée. [1]

Dans Python, nous avons utilisé la fonction **SelectKBest**, avec le paramètre **f_classif** car nous traitons un problème de classification. Voici les résultats obtenus :

Variable	*Score*
flipper_length	594.801627
bill_length	410.600255
bill_depth	359.789149
body_mass	343.626275
del13	234.897119
del15	229.625963
island	150.350633
clutch_completion	5.554315
date_egg	4.368453
sex	0.283857

TABLE 1 – F-Score pour chaque variable explicative de notre jeu de données

1. Source :https ://blog.minitab.com/fr/comprendre-lanalyse-de-la-variance-anova-et-le-test-f

Les résultats concordent avec ceux trouvés précédemment : *Sex, Date Egg, Clutch Completion* obtiennent un score très faible, signe qu'elles n'apportent pas beaucoup d'informations lorsque l'on veut classifier nos pingouins en fonction de l'espèce. On remarque qu'il y a un ratio d'environ 30 entre *Clutch Completion* et *Island*, illustrant la différence dans la qualité de l'information apportée.

Conclusion de la phase exploratoire :

1. Les variables apportent dans leur globalité de bonnes informations pour prédire l'espèce d'un individu. On espère donc de bons résultats lors des phases d'apprentissage. ;
2. Le jeu de données ne contient que 342 individus, ce qui est très peu. Pour valider les modèles testés, il faudra donc utiliser une validation croisée imbriquée avec partitionnement aléatoire pour réduire au maximum le biais et le sur-apprentissage.
3. Des classes correspondant aux espèces étaient déjà facilement visibles sur certains nuages de points.

2.3 Sélection des variables pour l'apprentissage

Cette étape est la conclusion logique de l'analyse exploratoire : on sélectionne les meilleurs variables pour prédire l'espèce lors de la phase d'apprentissage. Voici les changements finaux apportés aux données :

1. *Date Egg, Clutc Completion, Sex* sont supprimées, apportant peu d'informations pour la prédiction, et ajoutant trois dimensions supplémentaires au problème. Nous avons peu de données, et une réduction de dimension sera forcément bénéfique, surtout si elle est si bien justifiée.
2. *Island* est convertie au format OHE (One Hot Encoding).
3. *Species* est rétablie au format d'origine, c'est-à-dire que ses valeurs sont le nom des îles. Cette variable est séparée du reste du jeu de donnée sous forme de vecteur colonne car c'est la variable à prédire.

Remarques :

1. *One Hot Encoding* consiste à introduire autant de colonnes qu'il existe de valeurs prises par la variable qualitative qu'on remplace. Ici, il existe trois îles, donc la procédure crée trois nouvelles colonnes, dont les valeurs sont soit 1 si l'individu appartenait à l'île représentée par la colonne, soit 0 sinon. Cette manipulation nous permet de nous

défaire de la notion d'ordre introduite avec le *Label Encoding* effectué plus tôt. ;

2. Dans la section *Apprentissage supervisé*, il sera fait mention d'un jeu de données *centré et réduit*. Il s'agit du jeu de données traité comme décrit précédemment, que nous aurons centré et réduit en suivant la procédure standard (soustraction de la moyenne et division par l'écart type). Cette manipulation a pour but de s'affranchir de l'effet "taille" des données, afin de ne pas introduire de biais dans les modèles testés.

2.4 Analyse en composantes principales

Principe et objectifs de la méthode

Pour conclure cette phase , il nous semblait obligatoire d'opérer une analyse en composantes principales, afin d'observer nos pingouins sur des plans simplifiés. Le résultat de l'ACP sera aussi utilisé comme base pour certains modèles d'apprentissages, car nous avons peu de données, ainsi une réduction de dimension ne pourra qu'être appréciée.

La procédure consiste à chercher successivement les axes (donc les vecteurs les générant) qui maximiseront l'inertie de la projection du nuage des individus sur ces axes. L'inertie d'un nuage de points par rapport à un axe représente la distance des individus à cet axe : si l'on veut représenter nos pingouins dans un espace de dimension plus faible, disons 2 ou 3, on va chercher à minimiser la perte d'information due à la projection du nuage sur les axes formant ces dimensions. Ceci revient à maximiser l'inertie du nuage projetée sur l'espace complémentaire à l'axe, du fait que nous sommes dans un sous-espace vectoriel.

En terme de calculs, nous utiliserons les fonctions Python vues en TD, qui diagonalisent la matrice $B = VM$. Il suffit ensuite de trouver les composantes principales, soit la matrice $C = XMU$, U étant la matrice des vecteurs propres de B ordonnés par ordre de valeurs propres décroissantes. En analysant la proportion d'inertie cumulée expliquée par les composantes principales, on choisira combien en conserver pour les phases d'apprentissage.

Résultats

On trouve qu'en gardant trois composantes principales, on sauvegarde 90% de l'information, ce qui est parfait pour représenter nos pingouins dans les premiers plans factoriels, et même en trois dimensions.

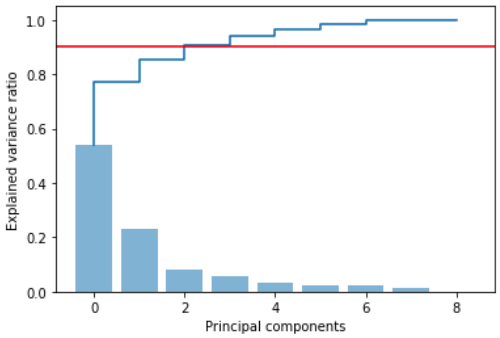

FIGURE 6 – Variance cumulée expliquée par les axes issus de l'ACP ; niveau vertical à 90%.

Nous pouvons maintenant faire des nuages de points dans les plans factoriels afin de visualiser la forme des classes dans un espace de dimension réduite.

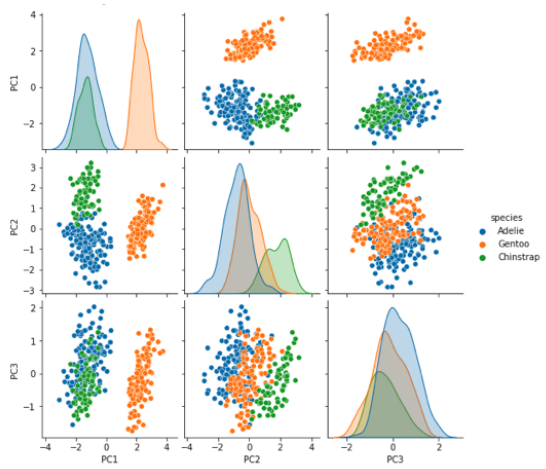

FIGURE 7 – Graphique matriciel des individus projetés sur les axes de l'ACP.

On observe des classes très nettes, avec encore une fois les pingouins *Gentoo* qui se séparent plus distinctement des autres. Selon le plan considéré, on peut aussi observer une séparation nette entre les *Adelie* et les *Chinstrap*. Cette réduction de dimensions nous sera précieuse lors de la phase d'apprentissage.

3 Apprentissage supervisé

3.1 K plus proches voisins

La méthode des K plus proches voisins est l'un des classifieurs les plus simples qui existent : le principe consiste à ranger l'individu au sein de la classe la plus représentée parmi les K plus proches voisins de son entourage, au sens par exemple de la distance euclidienne sur l'espace du nuage.

Pour apprendre le modèle, il faut déterminer son paramètre : le nombre K de voisins à considérer. La performance du classifieur sera mesurée en fonction de la fréquence des individus mal classés par le modèle.

Le problème étant que nous n'avons pas le luxe, avec 342 individus, de pouvoir avoir des ensembles d'apprentissage, de test et de validation assez grand pour garantir la fiabilité des résultats. Nous nous sommes donc tournés vers la procédure de *validation croisée imbriquée*.

Le principe consiste à faire une boucle qui redéfinira à chaque itération l'ensemble d'apprentissage et celui de test. A l'intérieur de la boucle, on estime les paramètres du modèles avec les individus d'apprentissage puis on calcule le taux d'erreur sur l'ensemble de test. En sortie, on pourra établir un intervalle de confiance sur le taux d'erreur et conclure sur le modèle.

Résultats

Nous avons appliqué les K plus proches voisins sur les variables issues de l'ACP, en gardant trois composantes principales : une réduction de dimension est appréciable lorsqu'on doit calculer des distances, et le nombre faible d'individus est une nouvelle fois une motivation clé.

Après avoir expérimenté sur plusieurs valeurs de K, dans l'intervalle [0, 100], on trouve que la meilleure valeur sera 6 voisins, avec une précision qui chute à partir d'un K trop grand. Ce qui est assez évident car comme il n'y a que peu de données, elles vont alors toutes se mélanger.

En appliquant le modèle sur l'ensemble de test, on obtient une précision de 100%. Nous avons d'abord pensé au phénomène de sur-apprentissage au vu de ce résultat. Néanmoins, comme vu sur les figures de l'ACP, la séparation entre individus est assez nette, et comme il y a assez peu d'individus à prédire (un tiers des individus), le résultat est plausible.

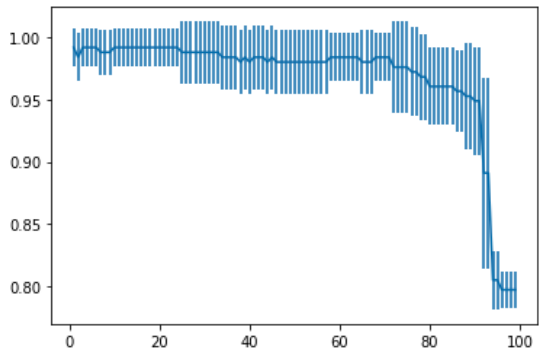

FIGURE 8 – Précision du classifieur pour K appartenant à [1, 100].

Matrice de confusion

```
[38  0  0]
[ 0 16  0]
[ 0  0 32]
```

Rapport de classification

	precision	recall	f1-score	support
Adelie	1.00	1.00	1.00	38
Chinstrap	1.00	1.00	1.00	16
Gentoo	1.00	1.00	1.00	32
accuracy			1.00	86
macro avg	1.00	1.00	1.00	86
weighted avg	1.00	1.00	1.00	86

FIGURE 9 – Rapport de classification des KNN sur les données de test.

En conclusion ici, les K plus proches voisins, modèle simple, semble pertinent pour notre problème de classification, et sa performance peut-être attribuée aux données nous facilitant la tâche, et au pré-traitement qui leur a été appliqué.

3.2 Régression logistique

La régression logistique est une méthode qui va estimer en classification les probabilités d'appartenance aux classes. Dans notre cas, nous allons appliquer des méthodes de régression logistique dites "multinomiales" car notre problème possède 3 résultats discrets possibles. L'idée va donc être de modéliser les probabilités à posteriori d'appartenance à chaque classe, sachant un vecteur de variables explicatives. Le modèle de régression logistique que nous allons utiliser par la suite est le modèle *logit*.

Pour pouvoir bien interpréter les poids associés par la régression logistique à chaque variable explicative nous allons utiliser ici un jeu de données centré-réduit. Nous avons utilisé la validation croisée avec partitionnement aléatoire à 50 plis pour tester l'efficacité des différentes méthodes de régression logistique sur notre problème de classification.

Sur les 50 résultats de classification, pour chacune des différentes méthodes de régression logistique utilisées, au maximum deux résultats de précision ne sont pas égaux à 100%. Voici par exemple l'histogramme (10) obtenu en utilisant l'algorithme *Limited-memory Broyden–Fletcher–Goldfarb–Shanno* qui nous montre bien la présence d'une seule valeur différente d'une précision de 100%.

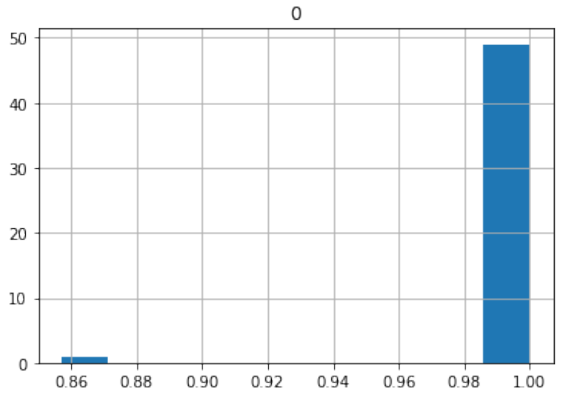

FIGURE 10 – Histogramme de la précision de prédiction sur 50 plis de validation croisée pour la régression logistique "lbfgs"

Cet algorithme est relativement similaire à celui de *Newton-Raphson* vu en cours à la différence qu'il utilise une approximation de l'inverse de la matrice hessienne dans le calcul des estimations des vecteurs de poids. Les autres méthodes utilisées ayant montrées des résultats presque identiques il est compliqué d'interpréter ces prédictions autrement qu'en disant que notre jeu de données possédant très peu de données, il est difficile d'éviter le sur-apprentissage et une très bonne performance de ces modèles.

On peut cependant facilement interpréter le vecteur de poids des variables correspondant à notre méthode de régression logistique utilisant l'algorithme *lbfgs* qui nous donnera des indications sur l'importance de nos variables explicatives dans la classification des espèces.

Ce tableau nous montre clairement que les variables quantitatives *bill length*, *bill depth*, *del13* et *flipper length* sont les plus importante dans la détermination de la probabilité d'appartenance d'un individu à une classe. Un résultat assez cohérent lorsque l'on regarde de plus près le graphique 4, où l'on observe que ce sont ces variables qui vont le plus différer en fonction des espèces. On peut aussi comparer ces résultats aux *tests F* effectués durant la phase d'analyse exploratoire (1). Les résultats correspondent assez bien à la différence que la variable *body mass* a finalement beaucoup moins d'importance lors de la régression logistique qu'annoncé lors des *tests F* et que *flipper length* n'est pas du tout la variable possédant le poids le plus élevé.

Nom de la variable	Valeur du coefficient
bill length	-1.9722
bill depth	1.3164
flipper length	-0.6852
body mass	0.1332
del15	0.2571
del13	-0.8159
island Biscoe	-0.1469
island Dream	-0.3578
island Torgersen	0.5047

TABLE 2 – Coefficient associé à chaque variable dans la régression logistique "lbfgs"

Comme nous avons déjà obtenu des résultats de précision de 100%, nous n'avons pas jugés nécessaires d'appliquer une ACP avant la régression logistique qui aurait seulement rendu l'interprétation des coefficients de poids beaucoup plus compliquée.

3.3 Analyses discriminantes

Le but d'une analyse discriminante (AD) est de prédire l'appartenance à une classe en faisant diverses hypothèses sur les distributions des variables ainsi que sur les propriétés de leurs variances, afin de diminuer le nombre de paramètres à estimer. Trois types d'analyses discriminantes ont été étudiées en cours / TD : l'analyse discriminante linéaire (ADL), l'analyse discriminante quadratique (ADQ) et le classifieur Bayésien naïf, et seront comparées dans la suite de ce rapport. Les hypothèses sous-jacentes à chacun des modèles seront également discutées, à la lumière des résultats obtenus.

L'ADL nécessite spécifiquement l'égalité des matrices de covariance, la normalité multivariée et l'indépendance des observations. De plus, elle ne modélise que des frontières de décision linéaires. D'autre part, l'ADQ est un type étendu de l'ADL, qui permet à modéliser des frontières quadratiques. L'ADQ nécessite également la normalité multivariée et l'indépendance des observations, mais n'a pas de limitation concernant l'homosedasticité. Enfin, le classifieur bayésien naïf considère les mêmes hypothèses que l'ADL, en ajoutant toutefois l'indépendance conditionnelle des variables.

Hypothèses liées aux modèles

Pour estimer les probabilités, l'AD modélise la distribution des prédicteurs X séparément dans chacune des classes (en sachant les valeurs de Y la variable à prédire), puis utilise le théorème de **Bayes** pour les transformer en estimations de

$$\mathbb{P}(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{c=1}^{K} \pi_c f_c(x)}$$

avec π_k la probabilité a priori et f_k la fonction de densité de X pour une observation venant de la k-ième classe. À partir des graphiques figurant sur la diagonale de la figure 4, nous supposons que les variables X suivent une distribution Gaussienne multivariée, et ensuite :

- Avec la même matrice de covariance dans chaque classe, cela conduit à l'ADL. ;
- Avec différentes matrices de covariance dans chaque classe, nous obtenons l'ADQ. ;
- Avec l'indépendance conditionnelle dans chaque classe (matrices de covariance diagonales), nous obtenons le classifieur bayésien naïf.

Présentation des résultats

Les trois modèles d'AD ont été appliqués sur le jeu de données préparé pour l'apprentissage, après centrage et réduction des données. Ayant peu de données, l'analyse en composantes principales a été appliquée afin de réduire le nombre de dimensions des données, car l'ADQ reste une méthode complexe et que si nous réduisons les dimensions sur un modèle d'AD, il faut le faire pour les trois.

Méthode	Matrice de confusion	Rapport de classification
ADL	[36 0 0] [0 15 0] [0 0 35]	precision recall f1-score support Adelie 1.000 1.000 1.000 36 Chinstrap 1.000 1.000 1.000 15 Gentoo 1.000 1.000 1.000 35 accuracy 1.000 86 macro avg 1.000 1.000 1.000 86 weighted avg 1.000 1.000 1.000 86
ADQ	[36 1 0] [0 14 0] [0 0 35]	precision recall f1-score support Adelie 0.973 1.000 0.986 36 Chinstrap 1.000 0.933 0.966 15 Gentoo 1.000 1.000 1.000 35 accuracy 0.988 86 macro avg 0.991 0.978 0.984 86 weighted avg 0.989 0.988 0.988 86
NB	[36 1 0] [0 14 0] [0 0 35]	precision recall f1-score support Adelie 0.973 1.000 0.986 36 Chinstrap 1.000 0.933 0.966 15 Gentoo 1.000 1.000 1.000 35 accuracy 0.988 86 macro avg 0.991 0.978 0.984 86 weighted avg 0.989 0.988 0.988 86

FIGURE 11 – Résultats de classification des méthodes d'analyse discriminante.

Voici une figure qui représente les frontières de décision des modèles sur le premier plan factoriel (12).

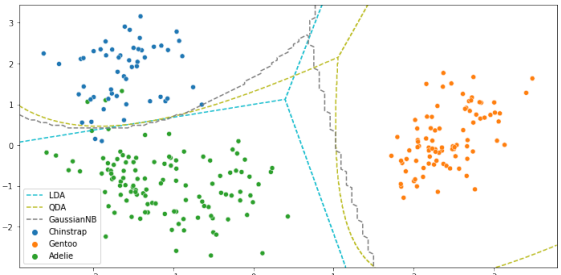

FIGURE 12 – Frontières de décisions tracées sur le premier plan factoriel.

Interprétation des résultats

L'ADL est un classificateur beaucoup moins flexible que l'ADQ, sa variance est donc nettement inférieure. Cela peut potentiellement conduire à une meilleure performance de prédiction (dans notre cas, c'est 100% contre 98.8% - figure 11). Ici, même si nous avons un jeu de données déséquilibré (figure 1) en terme de probabilités a priori, l'ADL tend à être un meilleur choix que l'ADQ en raison du nombre faible d'observations. En revanche, l'ADQ serait recommandée si l'ensemble des individus était très grand, de sorte que la variance du classificateur ne soit pas une préoccupation majeure, ou si l'hypothèse d'une matrice de covariance commune est clairement intenable.

3.4 Arbres de décision

Les arbres de décision binaires constituent une méthode d'apprentissage permettant de résoudre à la fois des problèmes de régression et de classification. Le principe de ces méthodes est de partitionner de manière récursive l'espace des caractéristiques en régions homogènes au sens des valeurs de la variable à expliquer.

Cette méthode est très intéressante de par sa facilité à être interprétée, mais elle possède plusieurs limites comme par exemple sa tendance à facilement surapprendre si on ne contrôle pas suffisamment la profondeur maximale de l'arbre. Étant donné que nous n'avons que peu de données et de variables explicatives, le phénomène de sur-apprentissage s'est vite manifesté.

Nous avons utilisé la validation croisée avec partitionnement aléatoire à 50 plis pour tester l'efficacité des arbres décisionnels sur notre problème de classification. Avec une profondeur maximale de l'arbre à 2 on arrive déjà à une grande majorité de résultats de prédiction à 100%, ce qui se traduit par une précision moyenne de 0.9357 et une médiane à 1 (voir le box-plot 13). Ce

résultat prouve une fois encore que le problème peut facilement se résoudre et que seules quelques frontières de partitionnement suffisent afin de bien classifier les pingouins.

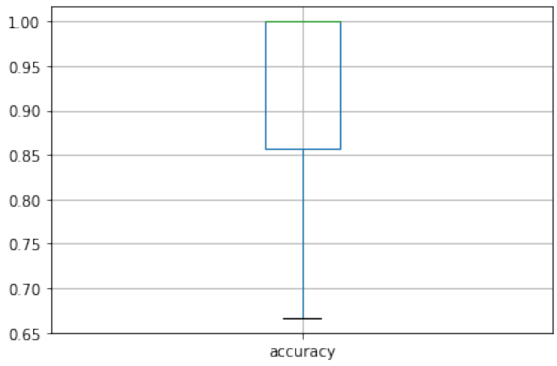

FIGURE 13 – Box-plot de la précision de prédiction sur 50 arbres de décision avec une profondeur maximale de 2

Si on essaye avec une profondeur maximale de 3 on obtient une moyenne de 0.9619 et une médiane toujours à 1. Ce qui signifie qu'il y a encore plus de résultats de précision égaux à 100% sur les 50 tests. Comme notre problème contient peu de variables explicatives et seulement 3 classes, augmenter significativement la profondeur maximale ne servirait à rien, à part sur-apprendre car les arbres n'ont pas besoin d'aller très profond pour obtenir des feuilles possédant un critère d'impureté le plus faible possible. De plus il faut deux fois plus de données pour occuper l'arbre à chaque niveau de profondeur supplémentaire ajouté, or comme notre jeu de données ne contient que très peu d'individus, il est impossible d'avoir des arbres de décision très profonds.

La grande majorité de précision à 100% qu'on observe peut nous signifier que notre jeu de données possède trop peu de données pour éviter le sur-apprentissage.

Observons désormais la représentation graphique d'un arbre pour essayer de comprendre quelles variables sont les plus importantes dans la caractérisation d'une espèce.

Ce premier noeud (14) permet à lui seul de classer presque parfaitement tous les pingouins de l'espèce *Gentoo*, ce qui n'est pas surprenant compte tenu de la séparation de cette espèce avec les autres pingouins mentionnés durant l'analyse exploratoire. En effet, ces pingouins possédant des caractéristiques physiques bien différentes de leurs contemporains, un test sur la variable *flipper length* permet de les classer correctement avec un indice de Gini de seulement 0.06. On peut également observer que seul un individu de l'espèce *Gentoo*

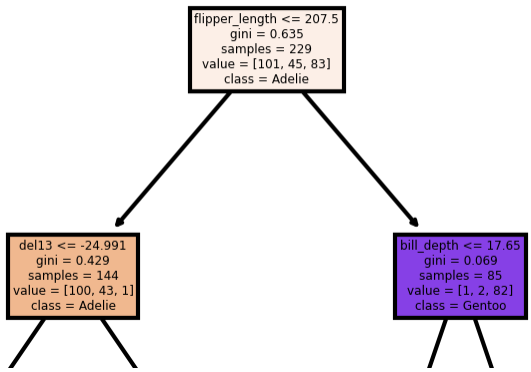

FIGURE 14 – Premier noeud d'un arbre de décision avec une profondeur maximale de 3 et une précision de 0.9469

n'a pas passé le test sur la variable *flipper length*, ce qui signifie que la longueur des nageoires des pingouins est extrêmement importante pour l'identification au sein de l'espèce *Gentoo* comme elle est plus élevée chez eux.

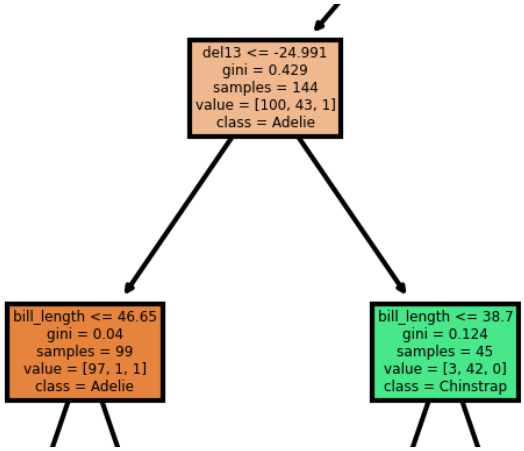

FIGURE 15 – Deuxième noeud d'un arbre de décision avec une profondeur maximale de 3 et une précision de 0.9469

Ce deuxième noeud (15) nous montre que la variable *del13* sépare très bien les pingouins *Chinstrap* et *Adelie* une fois que les *Gentoo* ont été écartés. En effet, lorsqu'on observe le graphique 4, on voit bien que les pingouins de l'espèce *Chinstrap* ont un ratio isotopique de carbone dans le sang nettement plus fort que les autres espèces. Les pingouins *Adelie* et *Gentoo* présentant une distribution de cette variable assez similaire, cette variable est très utile pour identifier les pingouins de l'espèce *Chinstrap*.

Appliquer une méthode de réduction de variables telle que l'ACP avant la méthode des arbres binaires serait contre-productif car on ne pourrait plus interpréter si facilement les arbres.

4 Conclusion

L'étude du jeu de données *Palmer Penguins* nous a permis d'appliquer une vaste palette de méthodes et d'expérimenter divers modèles d'apprentissage. La phase exploratoire, menant au pré-traitement des données pour les phases suivantes, a été plutôt aisée. Les difficultés sont venues de nos résultats trop "parfaits" : toutes nos méthodes présentent entre 96 et 100% de précision, il est alors difficile de les optimiser. Nous pouvons néanmoins décortiquer les résultats et poser des hypothèses sur leurs limites.

Notre jeu de données préparé et la réduction de dimensions ont contribué à ces résultats, néanmoins, nous pensons qu'en raison du faible nombre de données, nous sommes soit tombé dans le cas du sur-apprentissage, soit qu'il n'y avait pas assez de nuances au sein des données pour exprimer la complexité réelle du terrain, donnant ainsi une estimation très optimiste du taux d'erreur réel des classifieurs.

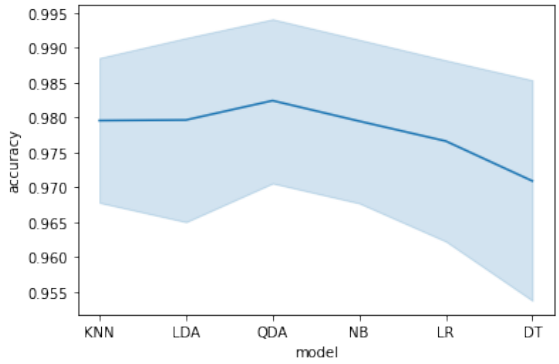

FIGURE 16 – Comparaison des intervalles de précision des différents classifieurs utilisés

Lorsque l'on compare les performances des différents classifieurs utilisés (16), il est difficile de désigner un modèle dominant comme les intervalles de confiance de précision coïncident tous. De plus, notre jeu de données ayant une taille très réduite, chacun des modèles va classer nos individus très rapidement, on ne peut donc pas vraiment sélectionner de modèle plus rapide en temps de calcul pour notre problème.

En réalité, nous avons travaillé sur un jeu de données "modèle", tout comme *Iris* qui poursuit une portée éducative : on obtient des résultats nets, qui illustrent bien les concepts théoriques. Ainsi, malgré un nombre d'individus faible, nous sommes satisfaits des résultats obtenus.

Références

[1] Yonatan-Carlos Carranza-Alarcon (2021) Yonatan Web, Teachings. *Available online at this URL*.

[2] Patrick Fore (2011). Categorical encoding using Label-Encoding and One-Hot-Encoder. *Disponible en ligne à cette URL*.

[3] Christopher W. Smith (2021). Classify That Penguin ! 100% Accuracy. *Disponible en ligne à cette URL*.

[4] Avinash Navlani (2018). Decision Tree Classification in Python. *Disponible en ligne à cette URL*.