Compte rendu de projet de SY09 - Jeu de données "Palmer Penguins"

Maria Al Bejjani - Paul Grimal - Mathias Vast

Juin 2021

1 Introduction

Ce rapport est rédigé dans le cadre du projet de l'UV SY09. Ce projet consiste en l'analyse et l'étude d'un jeu de données. Le jeu de données qui sera traité est intitulé *Palmer Penguins*[1], construit par le Dr Kristen Gorman et qui regroupe différentes mesures réalisées au cours de 3 études sur 3 espèces de manchots, dans un but scientifique. La page GitHub présente une version nettoyée du jeu de données (où les lignes vides et certaines colonnes ont déjà été enlevées lors d'un pré-traitement) ainsi qu'un jeu de données brut où aucun traitement n'a été appliqué en amont. Dans le cadre du projet, c'est ce dernier qui servira de base.

Parmi les mesures et informations qu'il contient, nous avons accès à des variables quantitatives, qui sont un ensemble de mesures de caractéristiques phénotypiques décrivant notamment le bec du manchot, ses nageoires ainsi que son poids, mais également des informations nutritionnelles. Le reste des variables sont quant à elles qualitatives, avec par exemple des variables décrivant l'espèce ou le sexe du manchot, l'île sur laquelle il était lorsqu'il a été enregistré ou encore concernant la ponte des oeufs.

Le but de notre étude est d'appliquer les méthodes vues en TD et en cours de SY09 afin de construire un classifieur capable de déterminer l'espèce ainsi que le sexe d'un manchot. Pour cela, nous nous appliquerons tout d'abord à analyser le jeu de données afin de formuler des hypothèses qui nous servirons par la suite à déterminer les traitements qu'il sera nécessaire d'appliquer avant de passer à l'application de méthodes d'apprentissage non-supervisé. Les résultats obtenus avec ces méthodes nous permettrons d'évaluer la qualité de notre représentation avant de pouvoir appliquer les méthodes d'apprentissage supervisé qui nous permettront de répondre à la problématique. Nous pourrons comparer les résultats obtenus par nos méthodes avec la vraie classification par sexe et par espèce présente dans le jeu de données.

1. https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-07-28/readme.md

2 Présentation des données

Table 1 – Description des variables

Variable	**Type**	**Signification**
studyName	character	Nom de l'étude où l'enregistrement a été réalisé
Sample Number	double	Numéro de l'enregistrement
Species	character	Espèce du manchot parmi Adelie, Chinstrap, Gentoo
Region	character	Région du globe où l'enregistrement a été réalisé
Island	character	Ile où l'enregistrement a été réalisé
Stage	character	Développement de l'oeuf
Individual ID	character	ID identifiant le manchot
Clutch Completion	character	Si l'oeuf du manchot a éclos ou non
Date Egg	double	Date de ponte
Culmen Length (mm)	double	Longueur du bec en mm
Culmen Depth (mm)	double	Profondeur du bec en mm
Flipper Length (mm)	double	Longueur des nageoires en mm
Body Mass (g)	double	Poids du manchot en g
Sex	character	Sexe du manchot
Delta 15 N (o/oo)	double	Quantité d'isotopes du nitrogène dans le sang
Delta 13 C (o/oo)	double	Quantité d'isotopes du carbone dans le sang
Comments	character	Commentaire lié à l'enregistrement

Les mesures des quantités d'isotope du carbone et du nitrogène dans le sang des manchots permettent d'estimer, en se basant sur leur demi-vie, à quand

remonte le dernier repas d'un individu. On a ainsi accès à un indicateur sur la qualité de la nutrition des manchots.

A l'origine, le jeu de données *Palmer Penguins* comporte 344 enregistrements, comprenants à la fois des caractéristiques phénotypiques sur les manchots et des informations liées aux études scientifiques au cours desquelles ces relevés ont été faits. Il y a 17 variables au total dans le jeu de données originel, parmi lesquelles 9 variables qualitatives et 8 variables quantitatives. La liste de ces variables est détaillée ci-dessus dans la table 1.
Dans le cadre de notre étude, toutes ces variables ne sont pas nécessaires, d'autant plus que toutes ne sont pas valuées pour chaque enregistrement. Ainsi, nous avons décidé de laisser de côté la liste de variables suivante :

- *studyName*, car cette variable n'est pas porteuse d'information sur la base de laquelle nous souhaitons pouvoir distinguer les espèces ;
- *Sample Number*, car cette variable est redondante avec l'index et qu'en plus elle est incomplète une fois que l'on retire les enregistrements contenants des valeurs non renseignées ;
- *Region*, car cette information est la même pour chaque manchot ;
- *Stage*, car cette variable ne présente pas d'intérêt pour notre étude ;
- *Individual ID*, car cette variable ne présente pas d'intérêt pour notre étude ;
- *Clutch Completion*, car cette variable ne présente pas d'intérêt pour notre étude ;
- *Date Egg*, car cette variable ne présente pas d'intérêt pour notre étude ;
- *Delta 15 N (o/oo)*, car cette variable n'est pas porteuse d'information sur la base de laquelle nous souhaitons pouvoir distinguer les espèces ;
- *Delta 13 C (o/oo)*, car cette variable n'est pas porteuse d'information sur la base de laquelle nous souhaitons pouvoir distinguer les espèces ;
- *Comments*, car cette variable ne présente pas d'intérêt pour notre étude.

Avant de pouvoir analyser ce jeu de données, il reste encore à le nettoyer des enregistrements incomplets et à renommer les colonnes afin d'obtenir un jeu de données plus facilement exploitable. Au final, le jeu de données retravaillé comporte 333 enregistrements, selon 7 variables : *Species*, *Island*, *Bill_ length* (anciennement *Culmen Length (mm)*), *Bill_ depth* (anciennement *Culmen Depth (mm)*), *Flipper_ length*, *Body_ mass* et *Sex*.

3 Analyse exploratoire des données

Maintenant que ce dernier est nettoyé, nous pouvons commencer à l'analyser afin de mieux comprendre les interactions des variables les unes avec les autres. Pour cela, nous commençons par regarder la répartition du nombre d'enregistrements selon l'espèce et le sexe disponible dans la table 2.

Table 2 – Répartition du nombre de mâles et de femelles au sein de chaque espèce

Espèce	Nombre femelle	Nombre mâle	Total
Adelie	73	73	146
Gentoo	58	61	119
Chinstrap	34	34	68

Nous remarquons que le nombre d'enregistrements est équilibré pour le sexe au sein de chaque espèce. Cependant, au niveau des espèces, on constate un déséquilibre entre le nombre d'Adelie, le nombre de Gentoo et de Chinstrap.

Nous nous sommes également intéressés à la provenance des manchots pour observer les liens entre cette information et les autres. Il y a trois îles, prénommées Biscoe, Dream et Torgersen. Au niveau de la répartition de chaque espèce, Biscoe semble accueillir les espèces Adelie et Gentoo tandis que Dream semble accueillir les espèces Adelie et Chinstrap. Torgensen semble accueillir uniquement l'espèce Adelie. Le problème avec cette variable, c'est que nous n'avons aucune garantie que cette répartition décrit la répartition réelle des espèces. En effet, l'île correspond seulement au lieu où le manchot a été enregistré mais nous ne savons pas si c'est également son site de nichage ou même si cette information est immuable. On ne considèrera donc pas l'île comme une donnée discriminante dans la suite de cette étude et nous ne l'intégrerons pas lors des prochaines étapes.

Maintenant que nous commençons à avoir un meilleur aperçu du jeu de données, nous allons nous intéresser aux corrélations entre nos différentes variables quantitatives. En observant le jeu de données de manière globale (et non par espèce), on remarque que la *Flipper_ length* et la *Body_ mass* sont significativement corrélées. Cette information est à conserver car nous pourrions être en présence d'un effet de taille qu'il faudrait alors annuler en amont de l'analyse en composantes principales.

En observant les corrélations par espèce, nous remarquons que la largeur des becs est davantage corrélée à la taille de celui-ci. Le poids du corps également, en plus d'être toujours corrélé à la taille des nageoires.

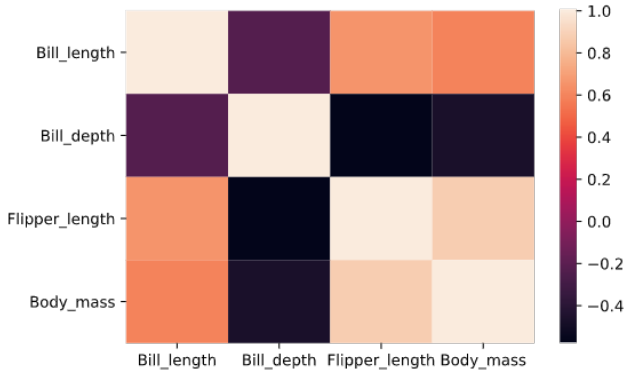

FIGURE 1 – Matrice de corrélation des variables quantitatives

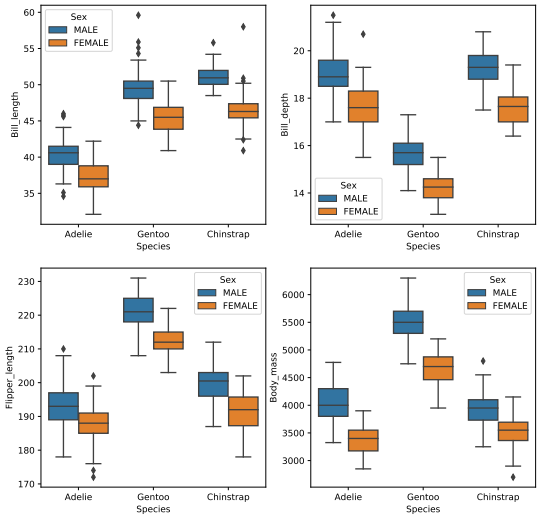

FIGURE 2 – Boxplot des caractères phénotypiques

En observant les attributs phénotypiques par espèce et par sexe, nous remarquons que l'espèce Adelie est facilement différenciable des deux autres grâce à la *Bill_ length*. De plus, on observe sur les boxplots de la figure 2 que les membres de l'espèce Gentoo ont une *Bill_ Depth* plus petite, une *Flipper_ Length* plus importante et une *Body_ Mass* plus importante que les membres des deux autres espèces. On peut également voir qu'au sein de chaque espèce, les caractères phénotypiques des femelles sont en moyenne moins élevés que ceux des mâles. Il semble alors possible de différencier les espèces et le sexe des manchots, à partir des différents caractères phénotypiques retenus.

Maintenant que nous avons un meilleur aperçu du jeu de données et des relations entre les différentes variables, nous allons pouvoir entamer le traitement de celui-ci.

4 Analyse en composantes principales

Afin de visualiser plus facilement ces données multidimensionnelles, nous utiliserons l'analyse en composantes principales (ACP). Cette méthode factorielle à pour but de decorréler les variables phénotypiques en créant de nouveaux axes représentant des combinaisons linéaires des variables initiales.

Dans un premier temps, nous avons appliqué l'ACP sans aucun traitement préalable sur les données. Malheureusement cette représentation ne nous permettait pas de bien séparer les espèce Adelie et Chintsrap, comme on peut le constater sur la figure 3.

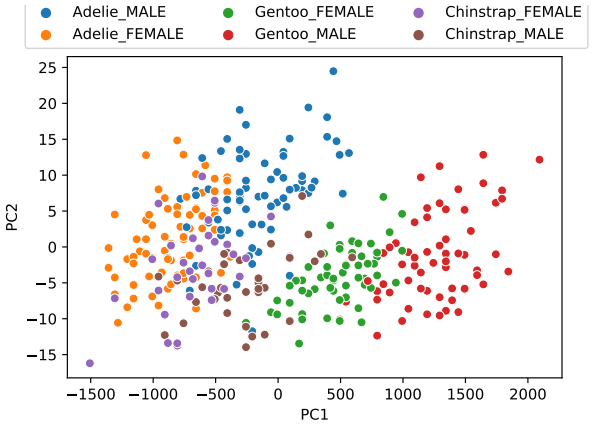

FIGURE 3 – Représentation des manchots dans le premier plan factoriel (ACP sans pré-traitement des variables) *Flipper_ Length* et *Body_ Mass*

La troisième espèce est plus facilement identifiable comme nous l'avions prédit aux vues des analyses menées précédemment. De plus, le premier axe porte une très grande partie de l'inertie expliquée (environ 99.9%). Il s'agit de l'effet taille que nous avons brièvement évoqué dans la partie précédente. En effet, les variables *Flipper_ Length* et *Body_ Mass* sont fortement corrélées positivement, comme nous pouvons le voir dans la figure 1 et qui est confirmé par un graphique de dispersion par paires (cf. figure 10 disponible dans la section Annexe 8). Nous observons encore cet effet de taille avec la figure 4, où un manchot avec une *Body_ mass* importante aura tendance à avoir une *Flipper_ length* également plus im-

portante.

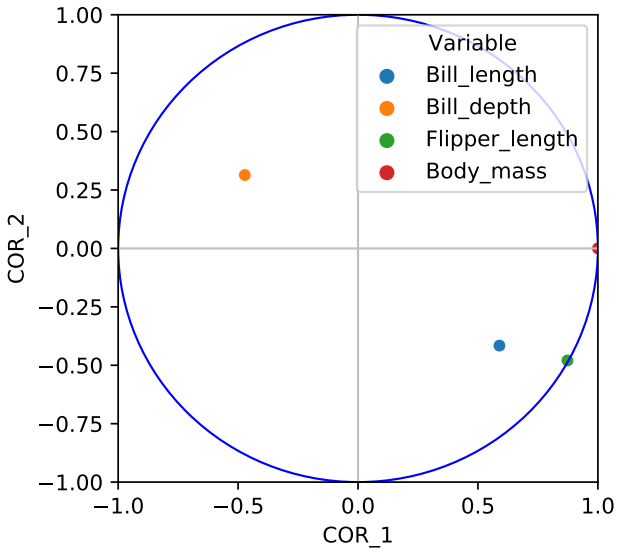

FIGURE 4 – Composantes principales normées, associées aux variables sur les deux premiers axes

Afin d'éviter de regrouper les manchots par leur taille, nous avons traités les variables *Flipper_Length* et *Body_Mass*, de sorte à avoir des variables comparables, en divisant les observations de ces deux variables par la somme de leurs valeurs. Puis nous avons réalisé une ACP. Cette méthode nous permet de distinguer les trois espèces de manchots, comme nous pouvons le voir dans la figure 5, en réduisant l'effet de taille constaté précédemment. Dans ce cas, les deux premiers axes factoriels portent la totalité de l'inertie expliquée (table 3).

TABLE 3 – Pourcentage d'inertie expliqués par les 4 axes factoriels après traitement des données

	Axe 1	Axe 2	Axe 3	Axe 4
% d'inertie expliquée	0.89	0.11	≈ 0	≈ 0

Cette fois, les deux espèces Adelie et Gentoo sont bien mieux séparées. Cette représentation est donc plus intéressante dans l'objectif de prédire par la suite à quelle espèce un individu appartient. Pour le sexe, on peut voir également sur cette représentation que les femelles ont tendance à être séparées des mâles selon le deuxième axe factoriel. Cela confirme les différences observées entre mâle et femelle au sein d'une même espèce vue dans la section 3 et observable sur la figure 2.

Ces résultats étant encourageants, nous avons décidé de poursuivre ce travail d'analyse en continuant

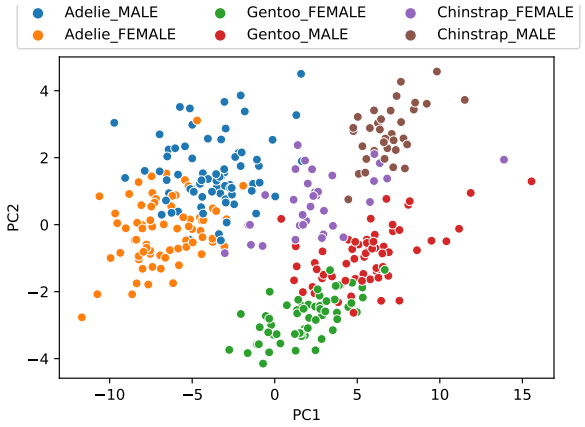

FIGURE 5 – Représentation des manchots dans le premier plan factoriel (ACP avec pré-traitement des variables *Flipper_Length* et *Body_Mass*)

le pré-traitement sur les données. On peut remarquer sur la figure 5 que la dispersion de chaque espèce sur le graphique est encore étendue. Afin d'y remédier, nous avons décidé de centrer et réduire les deux vecteurs correspondants aux variables *Flipper_length* et *Body_mass*. Le résultat de l'ACP avec tout ces traitements sur les données est visible sur la figure 6.

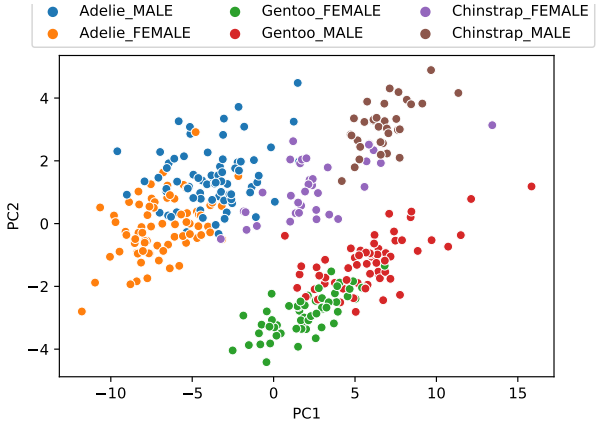

FIGURE 6 – Représentation des manchots dans le premier plan factoriel (ACP avec centrage-réduction des variables *Flipper_Length* et *Body_Mass*)

Les pourcentages d'inertie expliqués selon les axes factoriels sont disponibles dans la table 4.

On peut voir sur la figure 6 que les espèces sont plus facilement identifiables que précédemment tandis que la séparation selon le sexe reste similaire. Les deux premiers axes expliquent cette fois environ 98% de l'inertie totale. Le troisième axe explique quant à lui environ 2%.

Centrer et réduire les variables nous permet de les

TABLE 4 – Pourcentage d'inertie expliqués par les 4 axes factoriels sur le jeu de données partiellement centré et réduit

	Axe 1	Axe 2	Axe 3	Axe 4
% d'inertie expliquée	0.866	0.111	0.02	0.003

"ramener" à des tailles comparables. Nous avons appliqué ce traitement uniquement sur les variables qui provoquaient le plus l'effet de taille. On peut voir sur la figure 14, disponible dans la section Annexe 8, que nos trois premières composantes sont très interessantes pour séparer les espèces et le sexe. On peut également voir sur cette figure que le quatrième axe factoriel ne permet pas de séparer distinctement les différentes classes. C'est pourquoi nous avons décidé de ne pas prendre cet axe en compte pour la suite de notre étude.

Nous avons aussi essayé de réaliser les opérations de centrage et de réduction sur l'ensemble des variables mais les résultats étaient moins bons par rapport à notre objectif. Les individus Chinstrap et Adelie se retrouvaient mélangés en projetant sur les deux premiers axes, comme il est possible de le constater sur la figure 15 disponible en Annexe 8. Le problème de cette représentation était également que les deux premiers axes n'expliquaient pas totalement l'inertie, comme on peut le voir sur la table 5, alors qu'ils étaient les deux seuls à séparer correctement les différentes classes.

TABLE 5 – Pourcentage d'inertie expliqués par les 4 axes factoriels pour l'ACP sur le jeu de données entièrement centré et réduit

	Axe 1	Axe 2	Axe 3	Axe 4
% d'inertie expliquée	0.69	0.19	0.09	0.03

Ajouter les deux autres axes permettaient alors d'expliquer une plus grande part de l'inertie totale, sans pour autant mieux séparer les classes. C'est pour cette raison que nous avons décidé de nous en tenir à la version partiellement centrée et réduite plutôt qu'à cette dernière version.

Ainsi nous nous baserons pour la suite de notre étude sur l'ACP avec centrage et réduction sur *Flipper_Length* et *Body_Mass*. Nous avons de plus, décidé d'utiliser les trois premières composantes, réduisant ainsi notre nombre de variables initiales.

Afin d'estimer la qualité de notre séparation des espèces et ainsi vérifier que les méthodes d'apprentissage supervisé que nous appliquerons par la suite fonctionneront bien, nous allons poursuivre notre étude avec de l'apprentissage non-supervisé.

5 Apprentissage non-supervisé

Nous cherchons ici à retrouver des clusters cohérents qui séparent par espèces et si possible, également par sexe. Ainsi, nous allons procéder dans un premier temps à une classification hiérarchique : la méthode de Ward. Dans un second temps, nous appliquerons sur le jeu de données des méthodes de classification non hiérarchique : les K-means et K-means adaptatifs.

En appliquant la méthode de Ward, avec les distances euclidiennes et pour 3 clusters, sur ce jeu de données, on obtient des résultats mitigés. Les classes obtenues sont en fait à la perpendiculaire des "vraies" classes dessinées par les espèces. Avec 6 clusters en revanche, on obtient une classification très convaincante, qui reprend les classes dessinées par la décomposition des individus selon leur sexe et espèce. Le clustering ainsi obtenu est disponbile sur la figure 7.

Nous avons choisi la méthode de Ward car nous sommes en présence de variables quantitatives. L'ensemble à classifier correspond donc à un nuage de points que l'on peut munir d'une pondération $\frac{1}{n}$ dans $\mathbb{R}^n$ et de la distance euclidienne. Cette méthode va construire les classes en minimisant le critère d'inertie intra-classe.

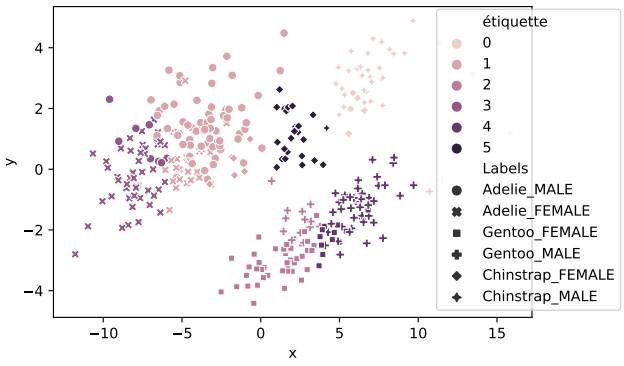

FIGURE 7 – Clustering avec la méthode de Ward, distance euclidienne et 6 clusters

Avec les méthodes des K-means et des K-means adaptatifs, nous obtenons des résultats similaires. L'étude de l'inertie en fonction du nombre de clusters (figure 8) confirme également que 6 est un nombre de clusters satisfaisant (méthode du coude).

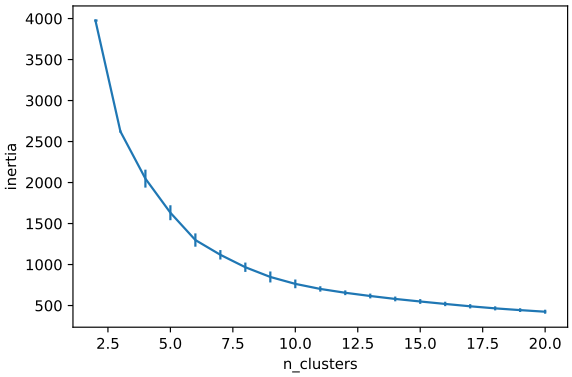

FIGURE 8 – Inerties obtenues avec la méthode des K-means en fonction du nombre de clusters

Sur la figure 9, disponible en Annexe 8, se trouve le clustering obtenu avec respectivement 3 et 6 clusters en appliquant la méthode des K-means adaptatifs.

TABLE 6 – Indice de Rand ajusté (ARI) des partitions obtenues avec les différentes méthodes de clustering et de la partition initiale sur les espèces

Méthodes de partionnement	ARI
K-means (3 clusters)	0.56
K-means (6 clusters)	0.55
K-means Adaptatifs (3 clusters)	0.69
K-means Adaptatifs (6 clusters)	0.49

Les résultats des K-means et K-means adaptatifs pour 3 et 6 clusters sont disponibles dans la table 6. Pour 3 clusters, le clustering a été fait en comparant les résultats avec la classification par espèce alors que pour 6 clusters, le score a été calculé en comparaison avec la classification par espèce et sexe. Les valeurs des indices de Rand ont été calculées en réalisant une moyenne sur 1000 essais. On peut voir que les résultats ne sont pas parfaits mais restent néanmoins convenables. Il est intéressant de relever que si la méthode des K-means adaptatifs est plus performante dans le cas de la labélisation par espèce que la méthode des K-means, elle l'est en revanche moins dans le cas de la labélisation par espèce et sexe. Comme on peut le voir sur le graphique de gauche de la figure 9, cette meilleure performance est dûe à la distribution des points au sein des clusters, qui est mieux capturée par cet algorithme qui tient compte des distances. En revanche, sur la figure de droite, on peut voir que les clusters sont bien moins étendus mais également plus imbriqués les uns dans les autres. C'est une situation où l'algorithme des K-means conviendra mieux.

Si les méthodes de clustering ne nous permettent pas de répondre à notre problématique, elles nous permettent néanmoins d'estimer la qualité de notre représentation. En effet, on peut voir que les clusters obtenus sont de bonne qualité et semblent correspondre aux répartitions réelles par sexe et espèce que nous avions constaté. On peut donc supposer que si ces méthodes ont bien performé, les méthodes d'apprentissage supervisé que nous allons voir par la suite devraient également donner de bons résultats.

6 Apprentissage supervisé

6.1 K plus proches voisins

Afin de répondre à la problématique, nous avons premièrement décidé de mettre en place la méthode des K plus proches voisins. Pour évaluer nos modèles, nous avons utilisé la validation croisée en séparant notre jeu de données en 10 échantillons. Cela permet d'avoir des jeux d'entraînement composés de 90% du jeu de données initial. Nous réalisons cela pour s'assurer que nous aurons assez de points afin d'avoir un classifieur pertinent. En effet, le jeu de données n'est pas équilibré avec l'espèce Chinstrap qui est moins représentée. En cherchant le sexe, il est important de s'assurer que nous aurons assez de données de Chinstrap femelle et de Chinstrap mâle dans le jeu d'entraînement.

Dans un premier temps, nous nous sommes interessés à prédire seulement les espèces. Pour déterminer le K optimal, nous avons mis en place un processus de validation croisée. Le nombre de voisins optimal obtenu sur 100 essais est 3. La comparaison des prédictions du classifieur sur le jeu de test par rapport aux vrais labels de ce jeu nous donne une précision de 0.99, en moyenne, sur ces 100 essais. Ce classifieur est donc très performant.

Nous pouvons noter que le nombre de voisins ne doit pas être trop élevé. En effet, le nombre de Chinstrap étant sous représenté, si on prend un trop grand nombre de voisins, il y a des risques pour que les Chinstraps soient de moins en moins identifiables. C'est un constat que nous avons pu confirmer avec des tests où nous avons vérifié que la précision baissait beaucoup lorsque le degré de liberté devenait faible.

Dans un second temps, nous nous sommes concentrés sur un bon classifieur capable de prédire à la fois l'espèce et le sexe d'un individu. Nous avons à nouveau réalisé une validation croisée en itérant sur le nombre de voisins et en calculant la précision de notre classifieur pour 100 essais. Le K optimal trouvé est 5, avec une précision

moyenne de 0.90. Le degré de liberté [2] correspondant étant de 59.8 en moyenne.

6.2 Analyse discriminante

Nous voulions comparer plusieurs méthodes afin de pouvoir mettre en perspective leurs résultats et nous avons donc décidé de tester ensuite les méthodes d'analyse discriminante. Nous avons donc appliqué l'analyse discriminante linéaire, quadratique ainsi que le classifieur naïf de Bayes, sur le jeu de données labélisé à la fois espèce et sexe. Nous avons évalué la performance de nos modèles avec la validation croisée. Les résultats sont récapitulés dans la Table 7.

TABLE 7 – Scores obtenus avec les différentes méthodes d'analyse discriminante

Méthodes d'analyse	Score (validation croisée)
LDA	0.892
QDA	0.898
Naive Bayes	0.893

Nous pouvons voir que les performances des 3 classifieurs sont assez comparables entre elles.

Si la LDA performe de manière comparable à la QDA sur ce jeu de données, c'est parce que le nombre de données disponibles est relativement faible. Ainsi, l'analyse discriminante linéaire souffre moins de son manque de flexibilité, contrairement à l'analyse discriminante quadratique. En revanche, l'analyse discriminante quadratique fonctionne également bien puisque les classes ne partagent pas toutes la même matrice de variance. En effet, si nous prenons par exemple les matrices de variance de la classe *Adelie_Male* et de la classe *Adelie_Femelle*, Nous remarquons que les deux sont différentes (Table 8 et 9).

TABLE 8 – Matrice de variance de la classe *Adelie_Male*

	Bill_len	Bill_dep	Flipper_len	Mass
Bill_len	1636.50			
Bill_dep	770.26	364.79		
Flipper_len	-24.36	-11.57	0.59	
Mass	-8.00	-3.81	0.20	0.22

2. $\frac{N}{Nb_{voisin}}$: Quantité d'information dont on a besoin pour prendre la décision. Plus on a besoin de points par rapport à la taille du jeu de données pour prendre une décision, plus le degré de liberté est faible. Réciproquement si le degré de liberté est élevé, on a besoin de moins de points. Un degré de liberté de 59.8 est plutôt élevé.

TABLE 9 – Matrice de variance de la classe *Adelie_Femelle*

	Bill_len	Bill_dep	Flipper_len	Mass
Bill_len	1392.18			
Bill_dep	656.85	311.41		
Flipper_len	-35.10	-16.56	1.04	
Mass	-38.73	-18.25	1.02	1.20

Les conditions pour que chaque méthode puisse déployer leur plein potentiel ne sont donc pas réunies dans les deux cas. Cela peut expliquer que leur performance soient au final comparables. Si nous regardons la variance de leur performance, l'analyse discriminante quadratique est celle dont la variance est la plus faible, comme on peut le voir sur la figure 11 disponible en Annexe 8. C'est donc la méthode d'analyse discriminante la plus fiable pour cette représentation.

Si nous nous intéressons uniquement à la labélisation par espèce, les résultats obtenus avec ces méthodes sont bien différents comme nous pouvons le voir sur la table 10.

TABLE 10 – Scores obtenus avec les différentes méthodes d'analyse discriminante, avec la labélisation uniquement sur l'espèce

Méthodes d'analyse	Score (validation croisée)
LDA	0.988
QDA	0.988
Naive Bayes	0.97

6.3 Régression logistique

Dans cette partie, nous allons estimer directement les probabilités d'appartenance aux classes. Pour cela, nous allons utiliser la régression logistique. Comme nous sommes en présence de plus de deux classes nous avons appliqué une regression logistique multinomiale. Nous avons à nouveau utilisé la validation croisée pour évaluer les performances de nos modèles.

Nous obtenons une perfomance de 0.99 pour classifier les espèces et 0.89 pour classifier l'espèce et le sexe.

6.4 Méthodes arborescentes

Pour terminer notre recherche de classifieur, nous avons tenté de discriminer les espèces et les sexes à l'aide des méthodes arborescentes. Nous avons évalué les performances à l'aide de la validation croisée.

Nous avons commencé par réaliser des classifieurs uniquement sur les espèces de manchots. Nous avons

construit l'arbre de décision et obtenu une précision de 0.94 en limitant le nombre de feuille terminale à 3 pour les 3 espèces. Si l'on répète cette méthode mais en spécifiant cette fois la valeur de lambda pour laquelle le compromis erreur-complexité est le meilleur, on obtient une précision de 0.963. L'arbre obtenu 12 est disponible en Annexe 8. Par la suite, nous avons essayé de diminuer la variance de nos estimateurs pour avoir un modèle plus flexible en utilisant la méthode de Bagging et en choisissant 10 jeux de données différents aléatoires. Nous obtenons une performance par validation croisée de 0.976. Afin de diversifier encore les méthodes et avoir un classifieur plus robuste, nous avons ensuite utilisé la méthode de Random Forest avec laquelle nous obtenons un score de 0.99.

En répétant le même processus pour retrouver l'espèce mais aussi le sexe des individus, nous obtenons cette fois-ci les résultats de la table 11.

TABLE 11 – Scores obtenus avec les différentes méthodes arborescentes, avec la labélisation sur les espèces et le sexe

Méthodes d'analyse	Score (validation croisée)
DecisionTree	0.81
DecisionTree (lambda optimal)	0.85
BaggingClassifier	0.844
RandomForest	0.877

L'arbre de décision correspondant à l'arbre avec la valeur de lambda optimal 13 est également en Annexe 8.

7 Conclusion

Ainsi nous trouvons en générale des bonnes performances pour les classifieurs sur les espèces, et un peu en deçà pour les classifieurs sur les espèces et le sexe. D'ailleurs, en regardant de plus près les erreurs de nos modèles sur cette dernière classification, nous nous sommes rendus compte que les erreurs étaient uniquement des erreurs sur le sexe du manchot. Nous allons maintenant expliquer les performances de nos modèles et tenter de conclure sur notre étude.

Pour les méthodes d'analyse discriminante et la régression linéaire, nous cherchons à déterminer une frontière de décision. Plus les classes sont distinctes et plus les résultats seront performants. Nos classes étant plutôt bien séparées dans l'espace, nous obtenons logiquement des précisions assez élevées avec ces méthodes. Concernant la classification par espèce et sexe, les points sont plus mélangés, ce qui explique des perfomances plus faibles.

Le classifieur des K plus proches voisins se base aussi sur la représentation des variables dans l'espace, ce qui explique également les précisions élevées obtenues en classifiant selon les espèces (proche de 100%). On obtient également de meilleurs résultats sur la classification selon le sexe et l'espèce. Nous expliquons cela par le fonctionnement de l'algorithme. En effet, il n'y a pas cette notion de frontière de décision, assez rigide, puisque le classifieur regarde les K plus proches voisins pour prendre une décision. On obtient ainsi des fontières plus adaptées aux contours des différentes classes.

Pour les méthodes arborescentes, on constate encore une fois des résultats très élevés (proche même de 100%) lorsqu'il s'agit de classifier seulement selon les espèces. Cela est dû au choix de la représentation qui a permis de clairement séparer les manchots selon les différents axes factoriels. Les performances un peu en deçà, obtenues avec la labélisation espèce + sexe est également dûe à cette représentation qui distingue moins clairement les mâles des femelles au sein des espèces.

En conclusion, nous pouvons dire qu'au terme de cette étude, nous obtenons plusieurs classifieurs pertinents, qui permettent de répondre de manière satisfaisante à la problématique. Ces classifieurs présentent tous des performances comparables mais dans l'objectif de prédire à la fois le sexe et l'espèce d'un individu, nous préférerons utiliser les K plus proche voisins pour les raisons évoquées ci-dessus.

Si ces classifieurs performent aussi bien, c'est notamment grâce aux choix de représentation qui ont été faits, ainsi qu'au pré-traitement des données. Cela nous a permis de séparer de manière convaincante les différentes classes, comme nous avons pu le vérifier avec les méthodes de clustering. Ce faisant, nous avons également pu vérifier que la distinction des espèces entre elles était plus évidente que celle des espèces et des sexes.

8 Annexes

8.1 Clusters avec la méthode des K-Means adaptatifs

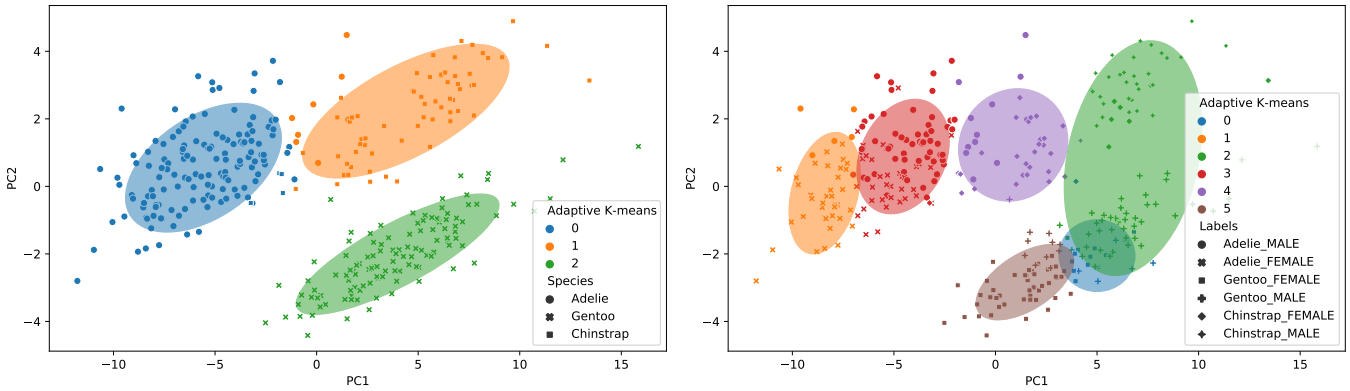

FIGURE 9 – Clusters obtenus avec la méthode des K-means adaptatifs pour 3 et 6 clusters

8.2 Distribution des manchots selon leur *Flipper_Length* et leur *Body_Mass*

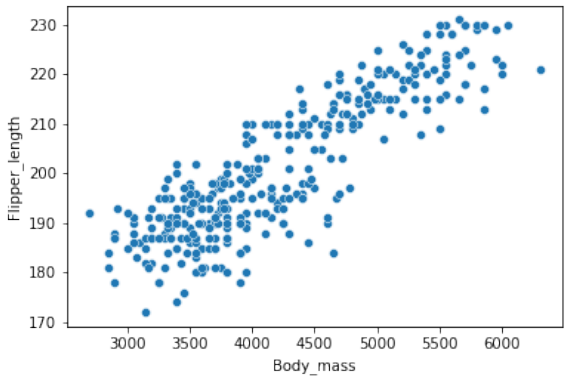

FIGURE 10 – Distribution des manchots selon leur *Flipper_Length* et leur *Body_Mass*

8.3 Comparaison des différentes méthodes d'analyse discriminante

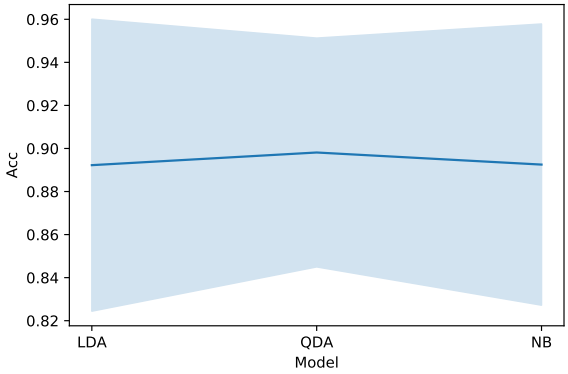

FIGURE 11 – Comparaison des précisions obtenues avec les différentes méthodes d'analyse discriminante

8.4 Arbres de décision

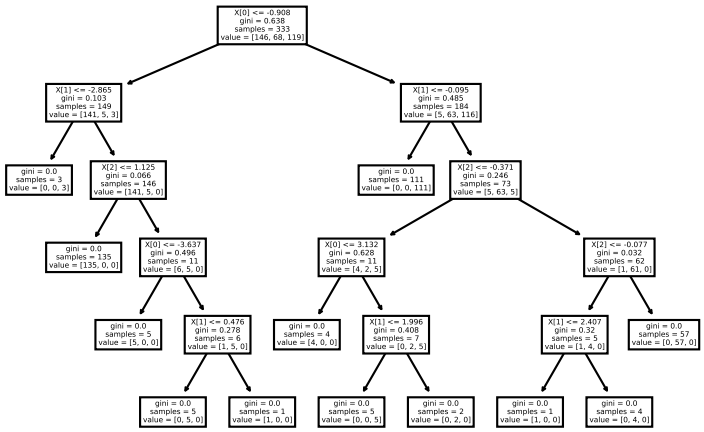

FIGURE 12 – Arbre de décision pour classer le jeu de données selon les espèces

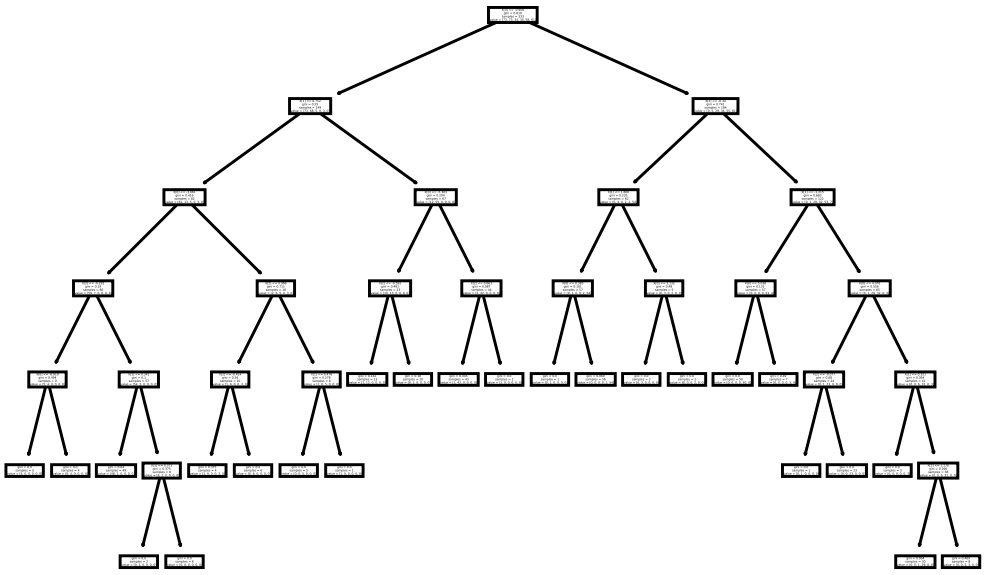

FIGURE 13 – Arbre de décision pour classer le jeu de données selon les espèces et le sexe

8.5 Visualisation de l'ACP obtenue après centrage-réduction des deux colonnes corrélées

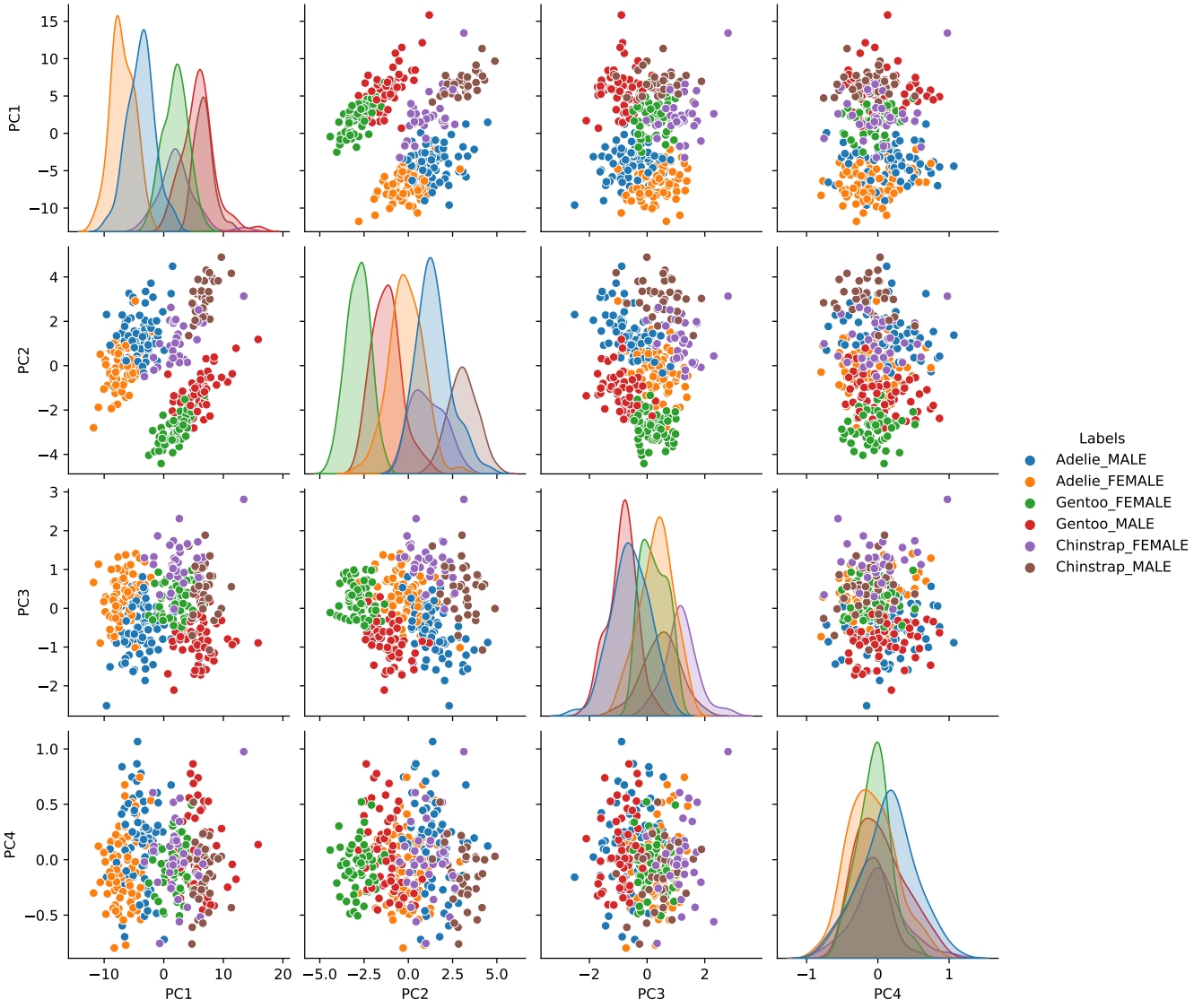

FIGURE 14 – Distribution des données selon les différents axes de l'ACP obtenue après centrage-réduction des deux colonnes corrélées

8.6 Visualisation de l'ACP obtenue après centrage-réduction de toutes les colonnes

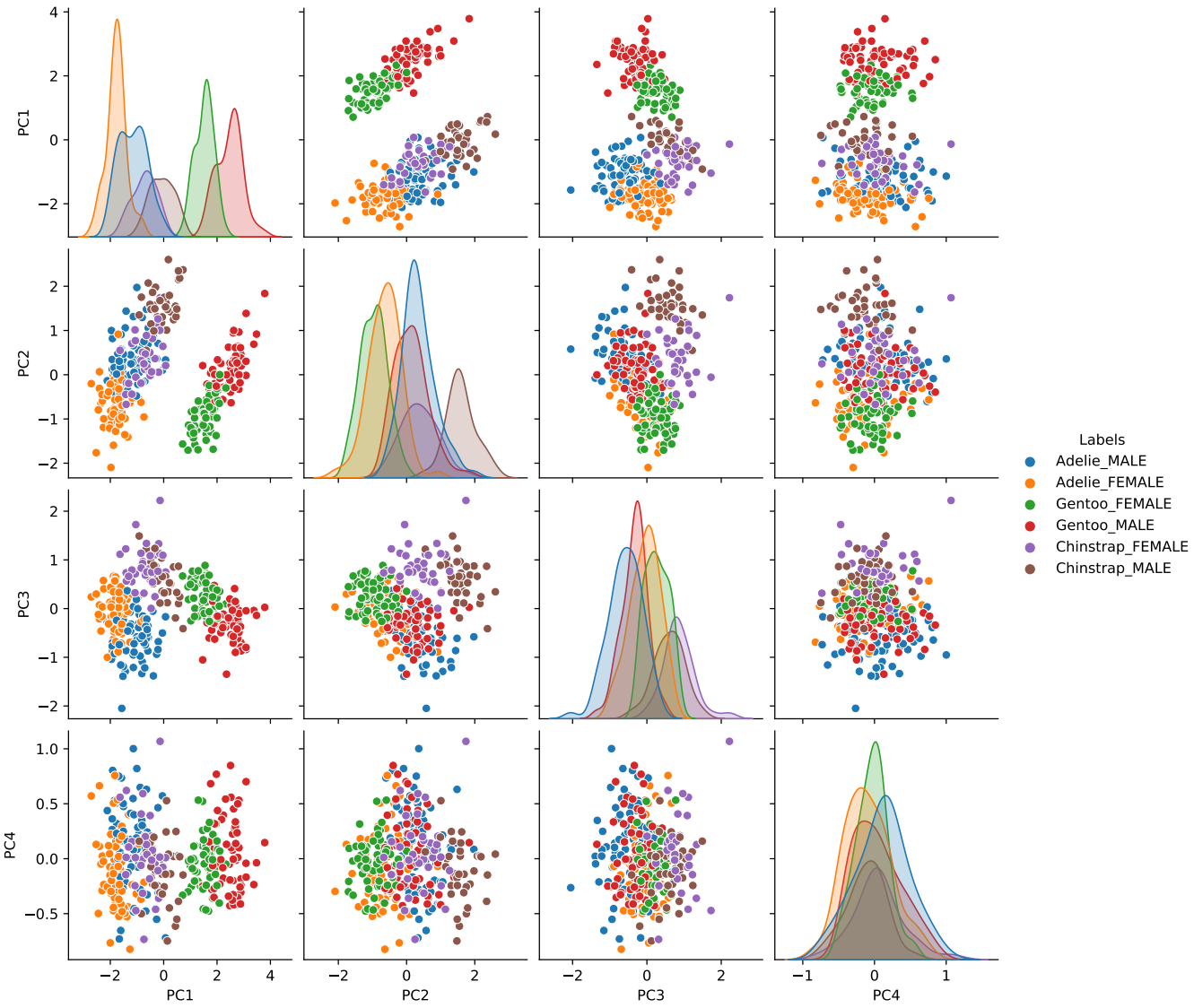

FIGURE 15 – Distribution des données selon les différents axes de l'ACP obtenue après centrage-réduction de toutes les colonnes