Les Lémuriens

Le Boudec Nathan, Louvion Léa

21 mai 2022

Résumé

Dans le cadre de l'UV **SY09** *Analyse des données et Data Mining*, il nous est demandé d'analyser un jeu de données. Initialement, nous avions choisi le jeu de données *Extinct Plants* extrait du *Tidy Up Tuesday de 2020*. Nous avons décidé de changer pour finalement porter notre choix sur le jeu de données *Lemurs* extrait du *Tidy Up Tuesday de 2021*. Le rapport présentera cette analyse.

1 Introduction

Le jeu de données *Lemurs*, provenant de la réserve de lémuriens *Duke Lemur Center*, présente différentes données sur 2270 individus de lémuriens, une espèce de primates en voie d'extinction. Les données les plus anciennes datent de 1966, et les plus récentes datent de 2019. 27 groupes taxonomiques différents sont représentés parmi les individus. Le jeu de données est accompagné d'un article présentant un travail de recherche publié en 2014 dans le journal scientifique *Nature* [1]. Plusieurs données sont disponibles : le sexe, l'espèce ainsi que des données sur la reproduction et sur la masse corporelle des lémuriens. Toutes les données proviennent d'individus captifs comme spécifié dans l'article de Zehr et al. (2014)[1]. L'objectif de cette étude consiste à mieux comprendre les différences entre les espèces de lémuriens et le sexe des lémuriens. Quelles populations de lémuriens sont représentées dans ce jeu de données ? Est-ce que certaines variables sont déterminantes pour l'espérance de vie du lémurien ? Peut-on élaborer un modèle pour prédire l'appartenance à l'espèce, ou encore l'appartenance à une classe d'âge ?

Nous nous intéresserons d'abord à l'analyse exploratoire sur les données afin de comprendre les relations entre les différentes variables. Cette première analyse nous permettra d'effectuer différentes classifications sur le jeu de données, dans l'optique de faire de l'apprentissage non-supervisé. Nous terminerons par la construction de différents modèles pour prédire l'appartenance des lémuriens selon différents groupes.

2 Présentation des données et analyse

2.1 Analyse exploratoire des données, construction des hypothèses sur les lémuriens

Pour l'analyse exploratoire, il a été nécessaire de regrouper les lignes de données correspondant à un individu particulier afin de ne pas compter les individus plusieurs fois. L'objectif de l'analyse exploratoire était de faire une première analyse des individus, d'observer s'il y avait des tendances en fonction de l'appartenance à l'espèce, de l'âge ou encore du sexe des lémuriens.

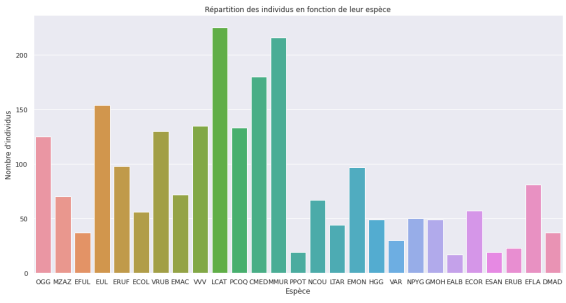

FIGURE 1 – Répartition des lémuriens en fonction de leur espèce

Les espèces sont représentées par un code à 3 ou 4 lettres (se reporter à la Table 1 en annexe). Il y a légèrement plus d'individus mâles que d'individus femelles, et certains sont de sexe indéterminé.

Les individus nés en milieu captif sont largement majoritaires par rapport à ceux nés en milieu sauvage.

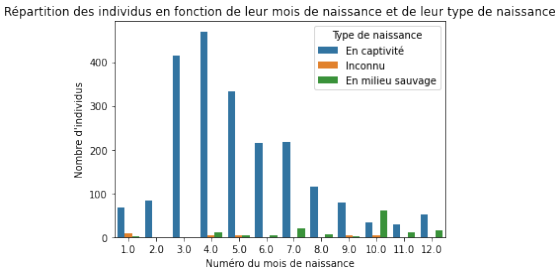

FIGURE 2 – Comptage des individus selon leur mois de naissance

On remarque que la plupart des lémuriens nés en captivité sont nés autour de mars, avril et mai. Quant aux individus nés en milieu sauvage, leurs mois naissance se situent plutôt autour d'octobre, novembre, décembre, et dans une moindre mesure en juillet. En effet, les individus nés en milieu sauvage (en grande majorité à Madagascar) ont le mois de naissance opposé par rapport aux États-Unis, Madagascar étant dans l'Hémisphère Sud et les États-Unis dans l'Hémisphère Nord.

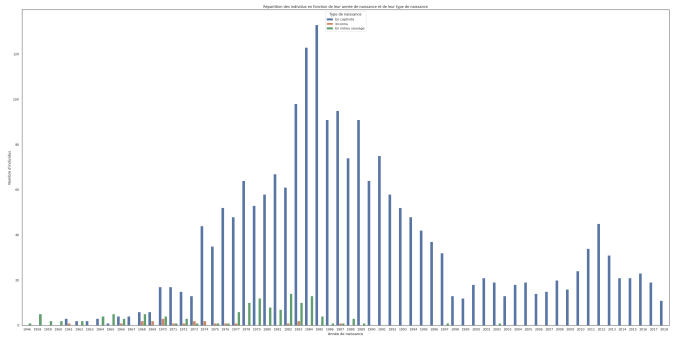

FIGURE 3 – Comptage des individus selon leur année de naissance

On peut remarquer une grande concentration d'individus nés dans les années 1980, et une croissance assez importante au début des années 1970. Cette période correspond à la création du *Duke Lemur Center*. Nous avons aussi essayé de visualiser le lieu de naissance des lémuriens. La carte interactive est disponible au format *HTML*[1]. Sans surprise, la majorité des lémuriens est née aux États-Unis, dans le *Duke Lemur Center*.

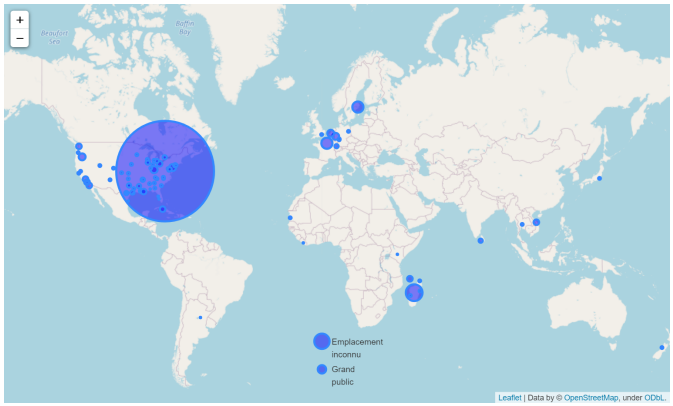

FIGURE 4 – Carte représentant le lieu de naissance des lémuriens

On a pu aussi remarquer, comme l'ont souligné d'autres articles [3], que pour certaines espèces, et particulièrement pour le sous-groupe des *Eulemur*, les femelles sont dominantes par rapport aux mâles, et sont donc plus lourdes. Les sous-groupe des *Eulemur* englobe tous les individus avec un taxon commençant par la lettre E.

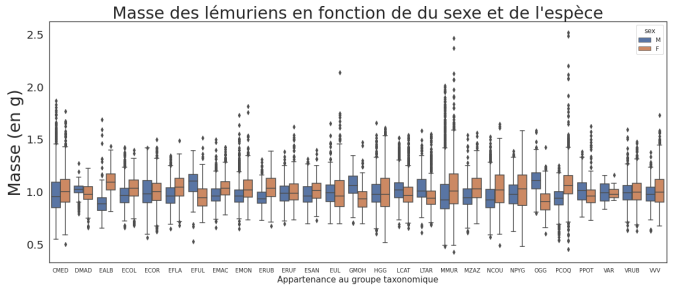

FIGURE 5 – Poids des lémuriens en fonction de leur sexe et de l'espèce

Un autre objectif de l'analyse exploratoire était de voir si certaines espèces étaient plus ou moins proches, si l'on pouvait constituer des groupes d'espèces. On peut par exemple remarquer que des espèces sont plus ou moins proches que d'autres par rapport à l'espérance de vie.

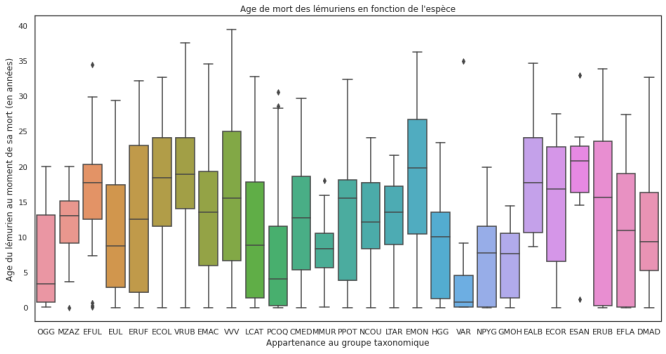

FIGURE 6 – Diagrammes en boîte des individus de l'espérance de vie des individus en fonction du taxon

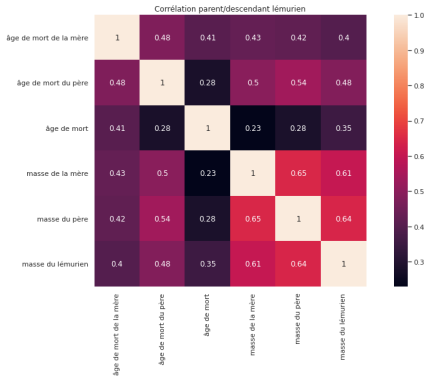

FIGURE 8 – Corrélation entre les caractéristiques des lémuriens parents et enfants

L'enjeu était ensuite de distinguer les individus par leur masse en fonction de leur âge et de l'espèce, en séparant les catégories d'âges. Cela a permis de confirmer par exemple que durant la phase juvénile, l'augmentation de la masse est assez importante durant les premiers jours. A l'âge adulte on peut constater une stagnation de la masse du lémurien, voire une diminution en fin de vie.

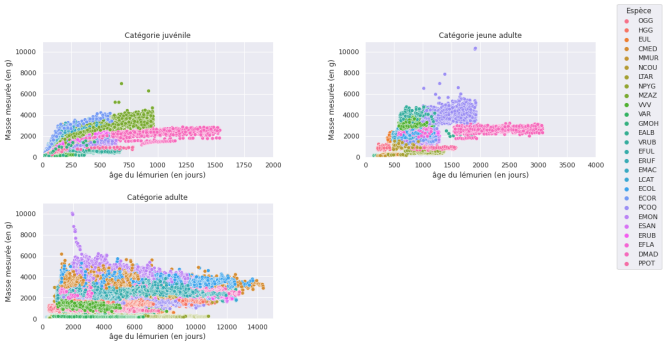

FIGURE 7 – Masse du lémurien en fonction de l'âge

Nous avons constitué un nouveau *Data Frame* avec de nouvelles colonnes pour faire correspondre, quand elles étaient connues, les durées de vie et le poids des parents à celles de leurs descendants. Cela nous a permis de savoir si certains traits de lémuriens étaient héréditaires. Nous avons pu observer une corrélation entre la masse corporelle du lémurien et la masse corporelle de son parent (de 0,6 à 0,7). Cette corrélation est en grande partie expliquée par les différences physiologiques intrinsèques aux espèces. Cependant, il est difficile de dire que l'espérance de vie du lémurien et de ses parents sont corrélées dans ce jeu de données.

2.2 Classification des individus

Nous avons commencé par réaliser une analyse en composantes principales (ACP) sur tous les taxons puis sur les groupes d'espèces les plus présentes au sein de l'échantillon comme indiqué en Figure 1, l'ACP avec les 27 espèces n'étant pas lisible. Cela nous a permis de mieux visualiser notre jeu de données et de réduire le nombre de dimensions. Lorsque l'on représente les inerties expliquées par chacun des axes, on remarque bien que les deux premiers axes permettent de bien représenter le jeu de données. En effet, la variance est en grande partie expliquée par l'âge et le poids de l'individu.

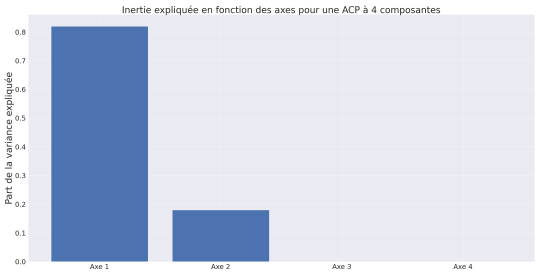

FIGURE 9 – Inertie expliquée en fonction des axes pour l'ACP

Nous avons ensuite voulu réaliser une analyse phylogénétique entre les espèces, *via* une classification hiérarchique. Cela a permis de montrer les similarités et les différences entre espèces, de pouvoir les regrouper. Nous avons pu le comparer à l'arbre phylogénétique donné par l'article *Nature* (Zehr et al., 2014) et,malgré quelques nuances, les résultats sont similaires. On retrouve bien le clade composé des *Eulemur* ainsi que celui composé des *Varecia*. Bien sûr, les distances entre les espèces sont indicatives et ne représentent pas réellement les

distances phylogénétiques entre les espèces, notre analyse ne se portant que sur quelques paramètres physiologiques que nous avons normalisés (poids moyen à l'âge adulte, âge moyen de mort, mois de naissance et taille de portée moyen) et non sur les gènes.

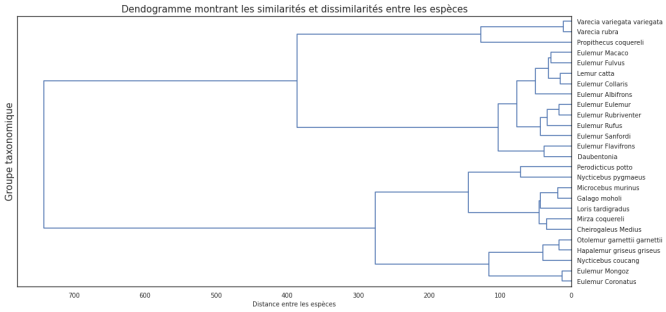

FIGURE 10 – Dendogramme montrant les relations phylogénétiques entre espèces

2.3 Prédiction de l'appartenance des individus à une espèce ou à un groupe d'âge

L'objectif était d'être capable de prédire les catégories d'âge d'individus. Cela permettrait, dans le cas d'un nouvel arrivant d'origine inconnue, de connaître sa catégorie d'âge, et ainsi d'estimer son âge, ce qui est très intéressant pour un centre comme le *Duke Lemur Center*. Nous nous sommes également fixés l'objectif de bien différencier les individus adultes de deux des groupes taxonomiques les plus présents au *Duke Lemur Center*, les *Daubentonia madagascariensis* (DMAD) et les *Propithecus coquereli* (PCOQ).

Pour cela, nous avons appliqué différents modèles. Le modèle des *k plus proches voisins*, l'*analyse discriminante*, la *régression logistique* et enfin des *arbres de décision*. Enfin, nous avons essayé d'être capable de bien différencier les individus des deux espèces précédemment citées quelque soit leurs catégories d'âge.

2.3.1 K plus proches voisins

En premier lieu, nous allons utiliser le modèle des *k plus proches voisins*. Nous nous restreignons aux 8 taxons qui représentent la grande majorité des individus présents au *Duke Lemur Center*. Nous trouvons que le meilleur paramètre K est 13. A l'aide d'un algortihme générant des données *train* aléatoirement pour chaque espèce, nous obtenons une précision moyenne de 0.88.

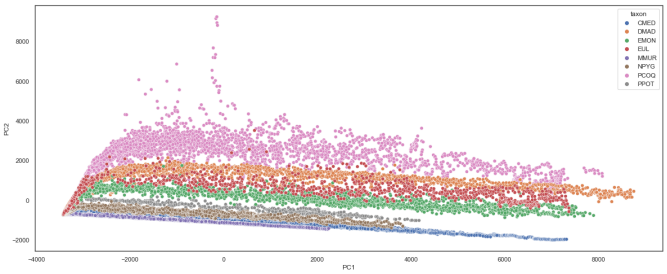

FIGURE 11 – Représentation des axes 1 et 2 d'une ACP réalisée sur les 8 taxons les plus représentatifs

2.3.2 Analyse discriminante

Nous avons ensuite essayé de différencier les individus *PCOQ* et *DMAD* à l'aide d'une analyse discriminante. Pour cela, nous avons utilisé trois modèles : les modèles d'analyse linéaire discriminante, quadratique et bayésienne naïve. Respectivement, nous obtenons des précision de 0.93, 0.98 et 0.95. L'analyse quadratique est donc un modèle particulièrement adapté. Cependant, lorsque les individus adultes et juvéniles sont inclus, la précision baisse drastiquement (0.68, 0.72, 0.72) : les modèles ne sont pas adaptés.

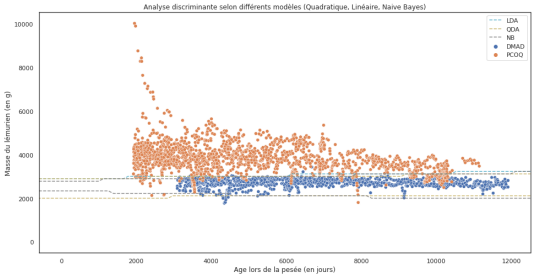

FIGURE 12 – Analyse discriminante pour prédire une espèce

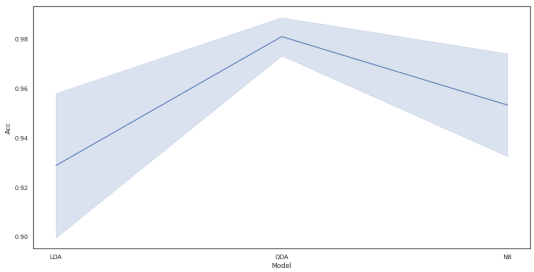

FIGURE 13 – Précision des différents modèles d'analyse discriminante

Pour différencier les catégories d'âge, les modèles quadratique et bayésien naïf accuse d'une bonne précision.

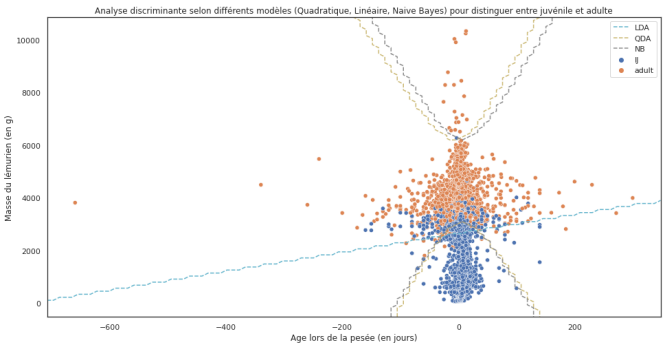

FIGURE 14 – Analyse discriminante pour prédire une catégorie d'âge

2.3.3 Régression logistique

Nous avons construit un modèle de régression logistique afin de prédire l'appartenance à un groupe d'âge. Cette régression avait pour paramètres le sexe, le poids, la variation moyenne de poids journalière, ainsi que l'espèce. Nous avons choisi de coder l'espèce en *One Hot* : chaque espèce disposait d'une colonne, un 1 indiquait l'appartenance à cette espèce et 0 le cas contraire. La matrice de confusion nous permet d'évaluer le nombre de réponses justes par rapport à des données test et leur prédiction[2]. Dans notre test, 1689 individus juvéniles sur 3065 individus juvéniles ont été effectivement classés dans la catégorie juvéniles. 11244 individus adultes sur 11504 individus adultes ont été classés dans la catégorie adultes. 28 individus jeunes adultes sur 1459 individus adultes ont été classés dans la catégorie jeunes adultes. La régression logisitique est donc plus performante pour classer les individus adultes, mais les performances sont assez mauvaises pour classer les jeunes adultes. L'usage de la fonction "classification report" permet de confirmer ce constat. Le modèle de régression logistique a un score de précision de 0,81 de manière générale.

2.3.4 Arbre de décision

Nous avons ensuite appliqué le modèle d'arbre de décision. Nous avons commencé par l'appliquer à la détermination de la catégorie d'âge, et le résultant est probant, le score de précision pour ce modèle est de 0,87.

Nous l'avons ensuite appliqué à la prédiction des espèces *PCOQ* et *DMAD*. Nous obtenons une précision de 0.96.

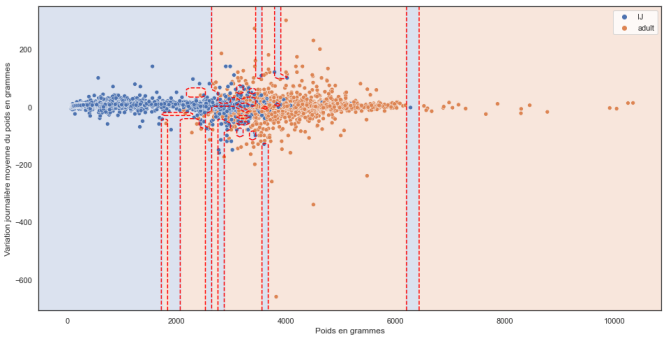

FIGURE 15 – Visualisation des frontières de décision après application du modèle d'arbre de décision sur les classes d'âges

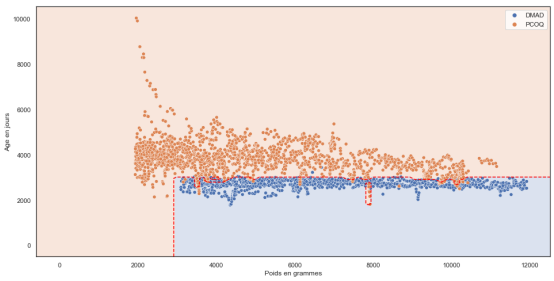

FIGURE 16 – Visualisation des frontières de décision après application du modèle d'arbre de décision sur les classes d'âges

La frontière étant constante, les coupes que réalise l'arbre de décision sont particulièrement adaptées.

2.3.5 Support Vector Classifier

Enfin, nous avons essayé de bien différencier les individus *DMAD* et *PCOQ* quelque soit leur âge. Cependant, nous voyons bien que l'évolution du poids des individus en fonction du temps suit une loi non-linéaire. C'est en réalité ce qui pourrait s'apparenter à une fonction sigmoïde. Nous avons réessayé une analyse discriminante, mais cette fois en utilisant un *Support Vector Classifier*. Un *Support Vector Classifier* est en fait une généralisation des classifieurs linéaires. En utilisant la fonction *SVC* du la librairie *sklearn*, on obtient une précision de 0.91 (contre 0.68, 0.72 et 0.72 pour respectivement l'analyse linéaire discriminante, quadratique et bayésienne naïve). C'est bien plus important qu'avec les outils d'analyse discriminante utilisés précédemment.

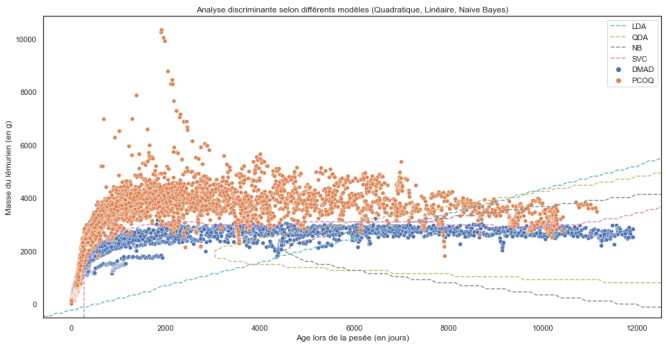

FIGURE 17 – Représentation du Support Vector Classifier pour prédire l'espèce

2.3.6 Conclusion sur les différents modèles de prédictions

Le modèle le plus précis est celui des *k plus proches voisins*. Cependant, il est bien plus coûteux en temps du fait qu'il faille comparer chaque point à tous les autres. Finalement, un modèle au compromis coût-précision intéressant serait celui du *SVC*.

3 Conclusion

L'analyse des informations sur les lémuriens nous a permis de comprendre les relations entre les différentes variables, de comprendre leur importance. Savoir, par exemple, à quel mois de l'année les lémuriens sont en gestation, ou encore quels sont les liens phylogénétiques entre les différents taxons, peut permettre de mieux agir sur la protection de l'espèce. Nous avons pu faire de la prédiction sur l'appartenance des lémuriens à des groupes d'espèces, ou encore à des groupes d'âges, de manière satisfaisante. Cependant, le nombre de variables quantitatives significatives est particulièrement limité dans ce jeu de données. Nous n'avions par exemple pas de données de taille sur les lémuriens et leurs attributs, uniquement le poids. Certaines variables, comme mentionné dans l'article de Zehr et al. (2014) ne sont pas utilisables pour l'analyse. Pour avoir une meilleure précision, dans les modèles, il faudrait ainsi rajouter des données. On pourrait également utiliser des réseaux de neurones en déterminant des fonctions d'activation. On obtiendrait peut-être de meilleurs résultats.

4 Annexe

TABLE 1 – Tableau de correspondances entre le code de lettres et le nom des lémuriens en latin

Code lettres	Nom de l'espèce en latin
CMED	Cheirogaleus medius
DMAD	Daubentonia madagascariensis
EALB	Eulemur albifrons
ECOL	Eulemur collaris
ECOR	Eulemur coronatus
EFUL	Eulemur fulvus
EMF	Eulemur flavifrons
EMM	Eulemur macaco
EMON	Eulemur mongoz
ERUB	Eulemur rubriventer
ERUF	Eulemur rufus
ESAN	Eulemur sanfordi
EUL	Eulemur
GMOH	Galago moholi
HGG	Hapalemur griseus griseus
LCAT	Lemur catta
LTAR	Loris tardigradus
MMUR	Mircocebus murinus
MZAZ	Mirza coquereli
NCOU	Nycticebus coucang
NPYG	Nycticebus pygmaeus
OGG	Otolemur garnettii garnettii
PCOQ	Propithecus coquereli
PPOT	Perodicticus potto
VAR	Varecia
VRUB	Varecia rubra
VVV	Varecia variegata variegata

A noter que les taxons désignés sous le nom d'*Eulemur* et de *Varecia* sont des espèces hybrides.

Références

[1] Zehr, SM, Roach RG, Haring D, Taylor J, Cameron FH, Yoder AD.(2014) Life history profiles for 27 strepsirrhine primate taxa generated using captive data from the Duke Lemur Center. Sci. Data 1 :140019 doi : 10.1038/sdata.2014.19 *Disponible en ligne à cette URL.*

[2] Bharathi (2021) Confusion Matrix for Multi-Class Classification *Disponible en ligne à cette URL.*

[3] Virginie Montmartin (2015) Les femelles lémuriens sont dominantes grâce... à la testostérone Sciences et Avenir *Disponible en ligne à cette URL.*

[4] Corrigés des TDs en SY09

[5] Notre dépôt GitHub, stockant différents fichiers utilisés pour l'analyse *Disponible en ligne à cette URL.*