

**THESE
POUR LE DIPLOME D'ETAT
DE DOCTEUR EN PHARMACIE**

**Soutenue publiquement le 21 septembre 2023
Par M. RIHANI Emir Kaïs**

**APPLICATION DE MODELES D'APPRENTISSAGE MACHINE
A LA CLASSIFICATION DES MACROMYCETES**

Membres du jury :

Président :

Pr LEMDANI Mohamed, PU en Biomathématiques, Faculté de Pharmacie de Lille

Directeur, conseiller de thèse :

Dr HAMONIER Julien, MCU en Biomathématiques, Faculté de Pharmacie de Lille

Assesseur :

Dr WELTI Stéphane, MCU en Sciences Végétales et Fongiques, Faculté de Pharmacie de Lille

Membre extérieur :

Dr MOUSSET Caroline, Pharmacien-Ingénieur, Responsable AQ Clients, Delpharm Lille

Faculté de Pharmacie de Lille
3 Rue du Professeur Laguesse – 59000 Lille
03 20 96 40 40
<https://pharmacie.univ-lille.fr>

Université de Lille

Président
Premier Vice-président
Vice-présidente Formation
Vice-président Recherche
Vice-présidente Réseaux internationaux et européens
Vice-président Ressources humaines
Directrice Générale des Services

Régis BORDET
Etienne PEYRAT
Christel BEAUCOURT
Olivier COLOT
Kathleen O'CONNOR
Jérôme FONCEL
Marie-Dominique SAVINA

UFR3S

Doyen
Premier Vice-Doyen
Vice-Doyen Recherche
Vice-Doyen Finances et Patrimoine
Vice-Doyen Coordination pluriprofessionnelle et Formations sanitaires
Vice-Doyen RH, SI et Qualité
Vice-Doyenne Formation tout au long de la vie
Vice-Doyen Territoires-Partenariats
Vice-Doyenne Vie de Campus
Vice-Doyen International et Communication
Vice-Doyen étudiant

Dominique LACROIX
Guillaume PENEL
Éric BOULANGER
Damien CUNY
Sébastien D'HARANCY
Hervé HUBERT
Caroline LANIER
Thomas MORGENTHOTH
Claire PINÇON
Vincent SOBANSKI
Dorian QUINZAIN

Faculté de Pharmacie

Doyen
Premier Assesseur et Assesseur en charge des études
Assesseur aux Ressources et Personnels
Assesseur à la Santé et à l'Accompagnement
Assesseur à la Vie de la Faculté
Responsable des Services
Représentant étudiant

Delphine ALLORGE
Benjamin BERTIN
Stéphanie DELBAERE
Anne GARAT
Emmanuelle LIPKA
Cyrille PORTA
Honoré GUISE

Professeurs des Universités - Praticiens Hospitaliers (PU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	ALLORGE	Delphine	Toxicologie et Santé publique	81
M.	BROUSSEAU	Thierry	Biochimie	82
M.	DÉCAUDIN	Bertrand	Biopharmacie, Pharmacie galénique et hospitalière	81
M.	DINE	Thierry	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
Mme	DUPONT-PRADO	Annabelle	Hématologie	82
Mme	GOFFARD	Anne	Bactériologie - Virologie	82
M.	GRESSIER	Bernard	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	ODOU	Pascal	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	POULAIN	Stéphanie	Hématologie	82
M.	SIMON	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	STAELS	Bart	Biologie cellulaire	82

Professeurs des Universités (PU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	ALIOUAT	El Moukhtar	Parasitologie - Biologie animale	87
Mme	AZAROUAL	Nathalie	Biophysique - RMN	85
M.	BLANCHEMAIN	Nicolas	Pharmacotechnie industrielle	85
M.	CARNOY	Christophe	Immunologie	87
M.	CAZIN	Jean-Louis	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	CHAVATTE	Philippe	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	COURTECUISSE	Régis	Sciences végétales et fongiques	87
M.	CUNY	Damien	Sciences végétales et fongiques	87
Mme	DELBAERE	Stéphanie	Biophysique - RMN	85
Mme	DEPREZ	Rebecca	Chimie thérapeutique	86
M.	DEPREZ	Benoît	Chimie bioinorganique	85
M.	DUPONT	Frédéric	Sciences végétales et fongiques	87

M.	DURIEZ	Patrick	Physiologie	86
M.	ELATI	Mohamed	Biomathématiques	27
M.	FOLIGNÉ	Benoît	Bactériologie - Virologie	87
Mme	FOULON	Catherine	Chimie analytique	85
M.	GARÇON	Guillaume	Toxicologie et Santé publique	86
M.	GOOSSENS	Jean-François	Chimie analytique	85
M.	HENNEBELLE	Thierry	Pharmacognosie	86
M.	LEBEGUE	Nicolas	Chimie thérapeutique	86
M.	LEMDANI	Mohamed	Biomathématiques	26
Mme	LESTAVEL	Sophie	Biologie cellulaire	87
Mme	LESTRELIN	Réjane	Biologie cellulaire	87
Mme	MELNYK	Patricia	Chimie physique	85
M.	MILLET	Régis	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	MUHR-TAILLEUX	Anne	Biochimie	87
Mme	PERROY	Anne-Catherine	Droit et Economie pharmaceutique	86
Mme	ROMOND	Marie-Bénédicte	Bactériologie - Virologie	87
Mme	SAHPAZ	Sevser	Pharmacognosie	86
M.	SERGHERAERT	Éric	Droit et Economie pharmaceutique	86
M.	SIEPMANN	Juergen	Pharmacotechnie industrielle	85
Mme	SIEPMANN	Florence	Pharmacotechnie industrielle	85
M.	WILLAND	Nicolas	Chimie organique	86

Maîtres de Conférences - Praticiens Hospitaliers (MCU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	BLONDIAUX	Nicolas	Bactériologie - Virologie	82
Mme	DEMARET	Julie	Immunologie	82
Mme	GARAT	Anne	Toxicologie et Santé publique	81
Mme	GENAY	Stéphanie	Biopharmacie, Pharmacie galénique et hospitalière	81
M.	LANNOY	Damien	Biopharmacie, Pharmacie galénique et hospitalière	80

Mme	ODOU	Marie-Françoise	Bactériologie - Virologie	82
-----	------	-----------------	---------------------------	----

Maîtres de Conférences des Universités (MCU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	AGOURIDAS	Laurence	Chimie thérapeutique	85
Mme	ALIOUAT	Cécile-Marie	Parasitologie - Biologie animale	87
M.	ANTHÉRIEU	Sébastien	Toxicologie et Santé publique	86
Mme	AUMERCIER	Pierrette	Biochimie	87
M.	BANTUBUNGI-BLUM	Kadiombo	Biologie cellulaire	87
Mme	BARTHELEMY	Christine	Biopharmacie, Pharmacie galénique et hospitalière	85
Mme	BEHRA	Josette	Bactériologie - Virologie	87
M.	BELARBI	Karim-Ali	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	BERTHET	Jérôme	Biophysique - RMN	85
M.	BERTIN	Benjamin	Immunologie	87
M.	BOCHU	Christophe	Biophysique - RMN	85
M.	BORDAGE	Simon	Pharmacognosie	86
M.	BOSC	Damien	Chimie thérapeutique	86
M.	BRIAND	Olivier	Biochimie	87
Mme	CARON-HOUDE	Sandrine	Biologie cellulaire	87
Mme	CARRIÉ	Hélène	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
Mme	CHABÉ	Magali	Parasitologie - Biologie animale	87
Mme	CHARTON	Julie	Chimie organique	86
M.	CHEVALIER	Dany	Toxicologie et Santé publique	86
Mme	DANEL	Cécile	Chimie analytique	85
Mme	DEMANCHE	Christine	Parasitologie - Biologie animale	87
Mme	DEMARQUILLY	Catherine	Biomathématiques	85
M.	DHIFI	Wajdi	Biomathématiques	27
Mme	DUMONT	Julie	Biologie cellulaire	87
M.	EL BAKALI	Jamal	Chimie thérapeutique	86
M.	FARCE	Amaury	Institut de Chimie Pharmaceutique Albert Lespagnol	86

M.	FLIPO	Marion	Chimie organique	86
M.	FURMAN	Christophe	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	GERVOIS	Philippe	Biochimie	87
Mme	GOOSSENS	Laurence	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	GRAVE	Béatrice	Toxicologie et Santé publique	86
Mme	GROSS	Barbara	Biochimie	87
M.	HAMONIER	Julien	Biomathématiques	26
Mme	HAMOUDI-BEN YELLES	Chérifa-Mounira	Pharmacotechnie industrielle	85
Mme	HANNOThIAUX	Marie-Hélène	Toxicologie et Santé publique	86
Mme	HELLEBOID	Audrey	Physiologie	86
M.	HERMANN	Emmanuel	Immunologie	87
M.	KAMBIA KPAKPAGA	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	KARROUT	Younes	Pharmacotechnie industrielle	85
Mme	LALLOYER	Fanny	Biochimie	87
Mme	LECOEUR	Marie	Chimie analytique	85
Mme	LEHMANN	Hélène	Droit et Economie pharmaceutique	86
Mme	LELEU	Natascha	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	LIPKA	Emmanuelle	Chimie analytique	85
Mme	LOINGEVILLE	Florence	Biomathématiques	26
Mme	MARTIN	Françoise	Physiologie	86
M.	MOREAU	Pierre-Arthur	Sciences végétales et fongiques	87
M.	MORGENROTH	Thomas	Droit et Economie pharmaceutique	86
Mme	MUSCHERT	Susanne	Pharmacotechnie industrielle	85
Mme	NIKASINOVIC	Lydia	Toxicologie et Santé publique	86
Mme	PINÇON	Claire	Biomathématiques	85
M.	PIVA	Frank	Biochimie	85
Mme	PLATEL	Anne	Toxicologie et Santé publique	86
M.	POURCET	Benoît	Biochimie	87
M.	RAVAUX	Pierre	Biomathématiques / Innovations pédagogiques	85

Mme	RAVEZ	Séverine	Chimie thérapeutique	86
Mme	RIVIÈRE	Céline	Pharmacognosie	86
M.	ROUMY	Vincent	Pharmacognosie	86
Mme	SEBTI	Yasmine	Biochimie	87
Mme	SINGER	Elisabeth	Bactériologie - Virologie	87
Mme	STANDAERT	Annie	Parasitologie - Biologie animale	87
M.	TAGZIRT	Madjid	Hématologie	87
M.	VILLEMAGNE	Baptiste	Chimie organique	86
M.	WELTI	Stéphane	Sciences végétales et fongiques	87
M.	YOUS	Saïd	Chimie thérapeutique	86
M.	ZITOUNI	Djamel	Biomathématiques	85

Professeurs certifiés

Civ.	Nom	Prénom	Service d'enseignement
Mme	FAUQUANT	Soline	Anglais
M.	HUGES	Dominique	Anglais
M.	OSTYN	Gaël	Anglais

Professeurs Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	DAO PHAN	Haï Pascal	Chimie thérapeutique	86
M.	DHANANI	Alban	Droit et Economie pharmaceutique	86

Maîtres de Conférences Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	CUCCHI	Malgorzata	Biomathématiques	85
M.	DUFOSSEZ	François	Biomathématiques	85
M.	FRIMAT	Bruno	Pharmacologie, Pharmacocinétique et Pharmacie clinique	85
M.	GILLOT	François	Droit et Economie pharmaceutique	86
M.	MASCAUT	Daniel	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86

M.	MITOUMBA	Fabrice	Biopharmacie, Pharmacie galénique et hospitalière	86
M.	PELLETIER	Franck	Droit et Economie pharmaceutique	86
M.	ZANETTI	Sébastien	Biomathématiques	85

Assistants Hospitalo-Universitaire (AHU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	CUVELIER	Élodie	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	GRZYCH	Guillaume	Biochimie	82
Mme	LENSKI	Marie	Toxicologie et Santé publique	81
Mme	HENRY	Héloïse	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	MASSE	Morgane	Biopharmacie, Pharmacie galénique et hospitalière	81

Attachés Temporaires d'Enseignement et de Recherche (ATER)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	GEORGE	Fanny	Bactériologie - Virologie / Immunologie	87
Mme	N'GUESSAN	Cécilia	Parasitologie - Biologie animale	87
M.	RUEZ	Richard	Hématologie	87
M.	SAIED	Tarak	Biophysique - RMN	85
M.	SIEROCKI	Pierre	Chimie bioinorganique	85

Enseignant contractuel

Civ.	Nom	Prénom	Service d'enseignement
M.	MARTIN MENA	Anthony	Biopharmacie, Pharmacie galénique et hospitalière

Faculté de Pharmacie de Lille

3 Rue du Professeur Laguesse – 59000 Lille
03 20 96 40 40
<https://pharmacie.univ-lille.fr>

**L'Université n'entend donner aucune approbation aux opinions
émises dans les thèses ; celles-ci sont propres à leurs auteurs.**

J'adresse mes sincères remerciements :

A la dream-team de tous ceux que je vais remercier.

Table des matières

1 Abréviations et conventions	17
1.1 Liste des abréviations	17
1.2 Conventions utilisées	17
2 Introduction	19
2.1 Propos liminaire	19
2.2 But de l'étude	19
2.3 État de l'art des lots de données mycologiques	20
3 Création du lot de données	23
3.1 Configuration matérielle et logicielle	23
3.2 Description de l'objet de l'étude	23
3.3 Principes de conception d'un lot de données synthétiques	24
3.3.1 Principes généraux	24
3.3.2 Principes de génération des paramètres quantitatifs	25
3.3.3 Principes de génération des paramètres qualitatifs	32
4 Principes de l'apprentissage machine	33
4.1 Types d'apprentissage machine	33
4.2 Jeux de données	33
4.3 Méthodes de construction des jeux de données	36
4.4 Modèles utilisés	37
4.4.1 Analyses discriminantes	37
4.4.2 Arbres de décision	41
4.4.3 Forêts aléatoires	46
4.5 Optimisation par plans d'expérience	48
4.6 Évaluation des performances des modèles	49
5 Apprentissage machine et classification binaire	53
5.1 Définition des critères de classification binaire	53
5.2 Analyse exploratoire des données (EDA)	54
5.3 Optimisation et sélection des modèles	58
5.3.1 Stratégie d'optimisation	58
5.3.2 Modèle naïf	59
5.3.3 Modèles d'analyse discriminante	60
5.3.4 Modèles d'arbres de décision	61
5.3.5 Forêts aléatoires	67
5.4 Résultats	73
5.4.1 Protocole d'évaluation	73
5.4.2 Performances des modèles de forêts aléatoires	73

6 Apprentissage machine et classification multiclasse	75
6.1 Classification par familles	75
6.1.1 Modèles d'arbres de décision	75
6.1.2 Forêts aléatoires	75
6.1.3 Résultats	77
6.2 Classification par espèce	80
6.2.1 Modèles d'arbres de décision	80
6.2.2 Forêts aléatoires	80
6.2.3 Résultats	82
7 Conclusion et perspectives	85
8 Références bibliographiques	86
A Annexe 1 : Algorithme de génération de lot synthétique	91
A.1 Initialisation	91
A.2 Préparation des données	91
A.3 Génération du lot de données	94
B Annexe 2 : Analyse exploratoire des données (EDA)	97
C Annexe 3 : Algorithmes d'apprentissage machine	99
C.1 Initialisation	99
C.2 Création des jeux d'entraînement, optimisation et évaluation	100
C.3 Entraînement et optimisation des modèles	100
C.3.1 Arbre de classification et régression	100
C.3.2 Rborist	103
C.4 Évaluation des performances des modèles	107
D Annexe 4 : Langage de balisage Rmarkdown	109
D.1 Introduction et préparation	109
D.2 Initialisation	111
D.3 Rédaction du corps de texte	112

1 Abréviations et conventions

1.1 Liste des abréviations

AUC : *Area Under Curve* (aire sous la courbe)
CART : *Classification And Regression Tree* (arbre de classification et de régression)
CSV : *Comma Separated Values* ([fichier de] valeurs séparées par des virgules)
DOE : *Design of Experiments* (plan d'expériences)
EDA : *Exploratory Data Analysis* (analyse exploratoire des données)
ESE : *Enhanced Stochastic Evolutionnary* ([algorithme] évolutionnaire stochastique amélioré)
LDA : *Linear Discriminant Analysis* (analyse linéaire discriminante)
LHS : *Latin Hypercube Sample* (échantillonnage par hypercube latin)
ML : *Machine Learning* (apprentissage machine)
NOLH : *Nearly Orthogonal Latin Hypercube* (hypercube latin quasi orthogonal)
PDA : *Penalized Discriminant Analysis* (analyse discriminante pénalisée)
RF : *Random Forest* (forêt aléatoire)
RF-RI : *Random Forest Random Inputs* (forêt aléatoire à [variables d']entrées aléatoires)
ROC : *Receiver Operating Characteristic* (fonction d'efficacité du récepteur)
VCS : *Version Control System* (logiciel de contrôle des versions)
YAML : *YAML Ain't Markup Language* (YAML n'est pas un langage de balisage)

1.2 Conventions utilisées

- Les références bibliographiques seront indiquées par des exposants numériques : ¹, ², ³...
- Les références de pied de page seront indiquées par des exposants alphabétiques : ^a, ^b, ^c...
- Les notations scientifiques utiliseront la *notation E* : $3e -5 = 3 \times 10^{-5}$.

2 Introduction

2.1 Propos liminaire

L'identification des macromycètes, c'est à dire des champignons visibles à l'œil nu est un sujet difficile, ne devant évidemment pas être pris à la légère. Les espèces rencontrées varient considérablement d'un écosystème à un autre, d'un continent à un autre, et aucun lot de données ni ouvrage sur les champignons ne saurait couvrir toute la diversité du monde fongique.

Le lot de données mycologiques constitué dans cette étude, bien que constituant l'un des lots en libre accès les plus complets du domaine de la *data science*, n'est bien entendu pas exhaustif.

Ce lot se concentre exclusivement sur les champignons habituellement rencontrés au Nord de la France. Nombre de variétés, parfois très connues, ne sont donc pas présentes, parmi lesquelles nous pouvons par exemple citer les représentants du genre *psilocybe*, connus pour leurs propriétés psychédéliques. Certains critères pourront également varier de manière considérable selon le stade de maturité du champignon : alors que les chapeaux vert-olive de l'*Amanita phalloides* mature sont faciles à reconnaître, les spécimens jeunes sont blancs et pourraient facilement être confondus avec des espèces comestibles (par exemple du genre *Agaricus*).¹

L'ingestion de certains de ces champignons est *mortelle*, même en faible quantité. Le diagnostic de l'intoxication fongique peut être difficile, et parfois trop tardif pour un traitement efficace. Des composés toxiques tels que les amanitines ne sont pas altérés ou détruits par cuisson ou congélation, et seront absorbés par l'intestin, avant de passer dans la circulation sanguine afin d'être filtrés par le foie, détruisant les cellules hépatiques, puis excrétées dans l'intestin, réabsorbées, refiltrées... chaque passe détruisant les cellules hépatiques ayant survécu à la précédente, dans un cycle connu sous le nom de réabsorption hépato-entérique.²

Il ne faut jamais, *sous aucune circonstance*, utiliser les lots de données générés par des méthodes similaires à celles de notre étude dans le but de déterminer si un champignon est comestible ou non.

2.2 But de l'étude

L'identification des plantes et champignons est un problème de classification classique, qui est usuellement effectué de façon manuelle, à l'aide de clés d'identification. La plupart de ces clés sont basées sur un processus utilisant des arbres décisionnels, ce qui pourrait sembler logique car rappelant la logique en arbre de l'évolution. Quoique séduisant, cet argument rencontre certaines limites :

La première limite est le nombre de chaînons manquants. Certaines espèces sont évidemment éteintes, ce qui signifie que certaines branches et nœuds de l'arbre phylogénétique seront manquants,

ce qui peut compliquer l'analyse quand deux espèces apparentées ont un nombre élevé de chaînons et nœuds communs manquants. Certaines similarités entre espèces peuvent également ne pas être identifiables de façon macroscopique.

La seconde limite, plus profonde, est la logique inhérente au processus évolutionnaire. Deux phénomènes antagonistes sont en jeu : convergence et divergence évolutives. Ces deux phénomènes sont liés à la nécessaire adaptation des espèces à leurs environnements. La divergence évolutionne explique par exemple la diversité des mammifères : les chauves-souris, baleines et chevaux sont apparentés, mais ont des aspects très dissemblables en raison de leur adaptation à des environnements très différents. D'un autre côté, la convergence évolutionne explique la similarité entre l'aile de la chauve-souris et celle de l'abeille. Toutefois, malgré leur apparence dissimilaire, l'aile de la chauve-souris est plus proche de la main humaine ou de la nageoire de la baleine que de l'aile de l'abeille. La façon la plus fiable pour évaluer le processus évolutionnaire et trouver les liens phylogénétiques de la manière la plus précise possible est l'analyse des génomes : les caractéristiques visibles peuvent être trompeuses. Malheureusement, ces caractéristiques sont souvent les seules aisément identifiables.

Le troisième problème est le critère principal de la classification. Ce critère peut être lié ou non au processus évolutionnaire ou aux critères visibles, surtout si ce critère principal est vague. Le critère de comestibilité ou de non-comestibilité retenu pour les lots de données mycologiques usuellement utilisés en *data science* souffre de ce problème : en effet, il est essentiellement centré sur la toxicité contre les humains, or de nombreux mécanismes de toxicité peuvent exister, et une toxicité ou non-toxicité d'un métabolite fongique ou végétal peut être liée à des variations métaboliques très ténues entre une espèce et une autre.

Pour ces raisons parmi d'autres, la logique arborescente, bien qu'utilisée habituellement dans l'identification des champignons et des plantes, et souvent justifiée par la nature arborescente du processus évolutionnaire, pourrait ne pas nécessairement être l'approche optimale à la classification des espèces basée sur des critères macroscopiques.

Le but principal de cette étude sera de déployer des algorithmes d'apprentissage machine afin d'effectuer cette tâche de classification basée sur des indices visuels limités, et d'évaluer les performances relatives de différentes stratégies et méthodes d'apprentissages machine dédiées à la classification.

2.3 État de l'art des lots de données mycologiques

Le tout premier lot de données mycologiques en libre accès mentionné en *data science* est probablement le *Mushroom Dataset* créé par Jeff Schlimmer en 1987.³.

Ce lot comprend 8124 spécimens, caractérisés par 23 variables qualitatives telles que la co-mestibilité, la forme et la surface du chapeau, la couleur et l'espacement des lames, l'odeur du champignon... .

Un lot de données plus conséquent a été publié par Dennis Wagner en 2021⁴ et mis en libre accès sous le nom de *Secondary Mushroom Dataset*.

Ce lot de données inclut 61069 spécimens et apporte une rationalisation des variables du lot précédent, se limitant ainsi à 18 variables qualitatives, et y ajoute 3 variables quantitatives apportant des caractéristiques dimensionnelles.

Des bases de données dédiées à la mycologie sont également disponibles,^{5,6} mais leur caractère généraliste empêche leur utilisation en data science sans un processus complexe de moissonnage (*data scraping*) et de nettoyage des données (*data cleaning*).

3 Création du lot de données

3.1 Configuration matérielle et logicielle

Le code d'apprentissage machine, les méthodes d'évaluation, ainsi que cette thèse ont été rédigés sur l'équipement suivant, présenté ici à fins de reproductibilité :

- CPU : AMD Ryzen 5 5600G
- RAM : 2x16 Go DDR4-3200
- SSD : Crucial P5 M2 NVMe
- OS : Xubuntu Linux 22.04.2 LTS
- R : version 4.2.2 (2022)
- IDE : RStudio version 2022.7.2.576, “Spotted Wakerobin”
- VCS : git version 2.34.1
- Librairies : tidyverse⁷ (v1.3.2), microbenchmark⁸ (v1.4.9), MASS⁹ (v7.3.58.2), caret¹⁰ (v6.0.93), GGally¹¹ (v2.1.2), twinning¹² (v1.0), rpart¹³ (v4.1.19), rpart.plot¹⁴ (v3.1.1), party¹⁵ (v1.3.11), ranger¹⁶ (v0.14.1), rFerns¹⁷ (v5.0.0), Rborist¹⁸ (v0.3.2), rmarkdown¹⁹ (v2.20), knitr²⁰ (v1.41), ggpubr²¹ (v0.6.0), DiceDesign²² (v1.9), DiceEval²³ (v1.5.1), bookdown²⁴ (v0.32).

3.2 Description de l'objet de l'étude

Notre étude portant sur la classification des champignons macroscopiques (macromycètes) par leur morphologie, il convient d'en apporter une définition succincte.

Le règne fongique est l'un des grands règnes du vivant, se distinguant à la fois du règne animal et du règne végétal par des caractères propres.^{2,25}

L'élément le plus atypique du règne fongique est probablement sa morphologie, caractérisée par son aspect essentiellement diffus, indéterminable et intimement lié à son substrat (sol, bois...), le champignon constituant ainsi un feutrage arachnéen, qualifié de *mycélium*.² Cette particularité singulière permet à certains champignons de constituer des organismes extrêmement massifs, pouvant s'étendre sur des surfaces considérables : le plus grand être vivant de notre planète est ainsi un spécimen d'*Armillaria ostoyae*, mesurant près de quatre kilomètres de long sur deux de large et vieux de plusieurs millénaires.²⁶

Les champignons se reproduisent à l'aide de spores, dont l'ensemble constitue la *sporée*, portée par un organe reproducteur spécialisé appelé le *sporophore*. Le terme de *champignon* du langage usuel et de *sporophore* du langage mycologique font tous deux référence à cet organe reproducteur susceptible d'être identifié, récolté et mangé.

Le sporophore “typique” est constitué d’un *stipe* (pied), surmonté d’un *chapeau*, ce dernier portant en sa face inférieure un *hyménophore*, souvent constitué de *lames* ou de *tubes*, portant les spores. Certains sporophores peuvent également porter un *voile général* ou un *voile partiel*, qui les enrobaient avant leur “éclosion” et dont les restes peuvent être présents sur le sporophore adulte. Ces différents éléments anatomiques, ainsi que l’environnement dans lequel prospère le champignon, représentent autant de caractéristiques permettant l’identification d’une espèce, et qui seront exploitées dans cette étude (voir figure 1).

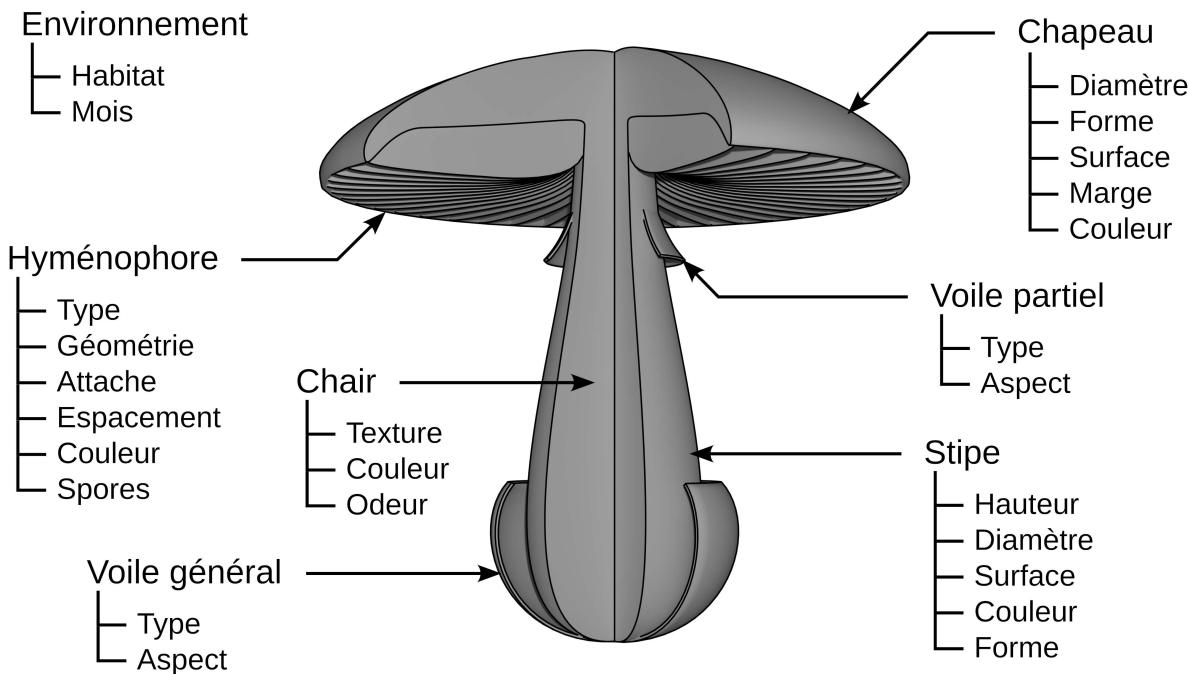


Figure 1: Éléments anatomiques du sporophore et caractéristiques exploitées pour son identification.

3.3 Principes de conception d'un lot de données synthétiques

3.3.1 Principes généraux

Un lot de données synthétiques est un lot de données généré par un algorithme, par opposition aux lots de données issus d'une collecte effectuée en “vie réelle”.

Trois stratégies sont usuellement utilisées :

- Données factices (*dummy data*) : l’ensemble des données est généré aléatoirement.
- Données générées à partir de règles (*rule-based data*) : l’ensemble des données est généré suivant des lois définies au préalable (distribution, valeurs moyennes, minimales, maximales...)
- Données générées par intelligence artificielle (*AI generated*) : l’ensemble des données est généré suivant des lois extraites par l’IA suite à l’analyse d’un échantillon de données obtenues en “vie réelle”.

Les données générées par ces stratégies peuvent être de types variés, que nous pouvons grossièrement regrouper en données alphanumériques (quantitatives et qualitatives), en séries temporelles, et en données d'imagerie.

Pour des raisons pratiques et de maturité des technologies disponibles à l'heure actuelle, la méthode retenue pour créer le lot de données exploité dans notre étude sera la génération de données alphanumériques à partir de règles, extraites d'ouvrages mycologiques de référence.^{1,2,25,27}

3.3.2 Principes de génération des paramètres quantitatifs

Dans le cadre de cette étude, les variables quantitatives générées aléatoirement sont :

- La longueur du stipe L_S ,
- Le diamètre du stipe D_S ,
- Le diamètre du chapeau D_C .

En première approximation, nous pouvons considérer que toutes ces valeurs sont intrinsèquement liées à la croissance du champignon. Ces trois variables peuvent, dans l'absolu, être susceptibles de varier indépendamment des autres au cours de la croissance du champignon, les variables L_S , D_S et D_C obéissant alors aux lois suivantes :

$$\begin{cases} L_S = L_{S_{max}} \cdot F_{Ls} \\ D_S = D_{S_{max}} \cdot F_{Ds} \\ D_C = D_{C_{max}} \cdot F_{Dc} \end{cases}$$

Avec :

- $L_{S_{max}}$, $D_{S_{max}}$ et $D_{C_{max}}$ les valeurs maximales de longueur de stipe, diamètre du stipe et diamètre de chapeau de chaque variété fongique, extraites de la littérature,
- F_{Ls} , F_{Ds} , F_{Dc} des variables générées aléatoirement dans l'intervalle $]0; 1]$, et représentatives de la croissance du spécimen.

Toutefois, nos recherches bibliographiques n'ont pas permis de distinguer de différences de la cinétique de croissance de chacune de ces trois caractéristiques dimensionnelles du sporophore. Nous supposerons donc, en première approximation, que la croissance du stipe en longueur et en largeur, ainsi que la croissance du chapeau s'effectuent à des vitesses identiques. Nous obtenons par conséquent :

$$F_{Ls} = F_{Ds} = F_{Dc} = F_T$$

Avec F_T un facteur représentatif de la taille globale de chaque spécimen, généré aléatoirement.

Ainsi, le problème de génération de nos trois variables aléatoires se simplifie en un problème de génération d'une seule variable aléatoire : le facteur de taille de chaque spécimen. Un certain nombre de distributions d'intérêt sont susceptibles d'être utilisées afin de générer des facteurs de taille F_T aléatoires. Il convient donc de définir le cahier des charges de la distribution la plus adaptée au sujet de cette étude.

Les critères de sélection retenus afin de choisir la loi la plus appropriée sont :

- Efficience calculatoire,
- Distribution continue,
- Distribution bornée, ou aisément normalisable sur un intervalle $[0; 1]$,
- Distribution asymétrique.

L'efficience calculatoire n'est, en pratique, pas un facteur limitant dans notre cas, les temps de calcul pour la génération d'un nombre de facteurs de taille F_T suffisant s'avérant typiquement inférieurs à 200 ms (pour 10^6 facteurs générés) avec la plupart des distributions d'intérêt (voir figure 2).

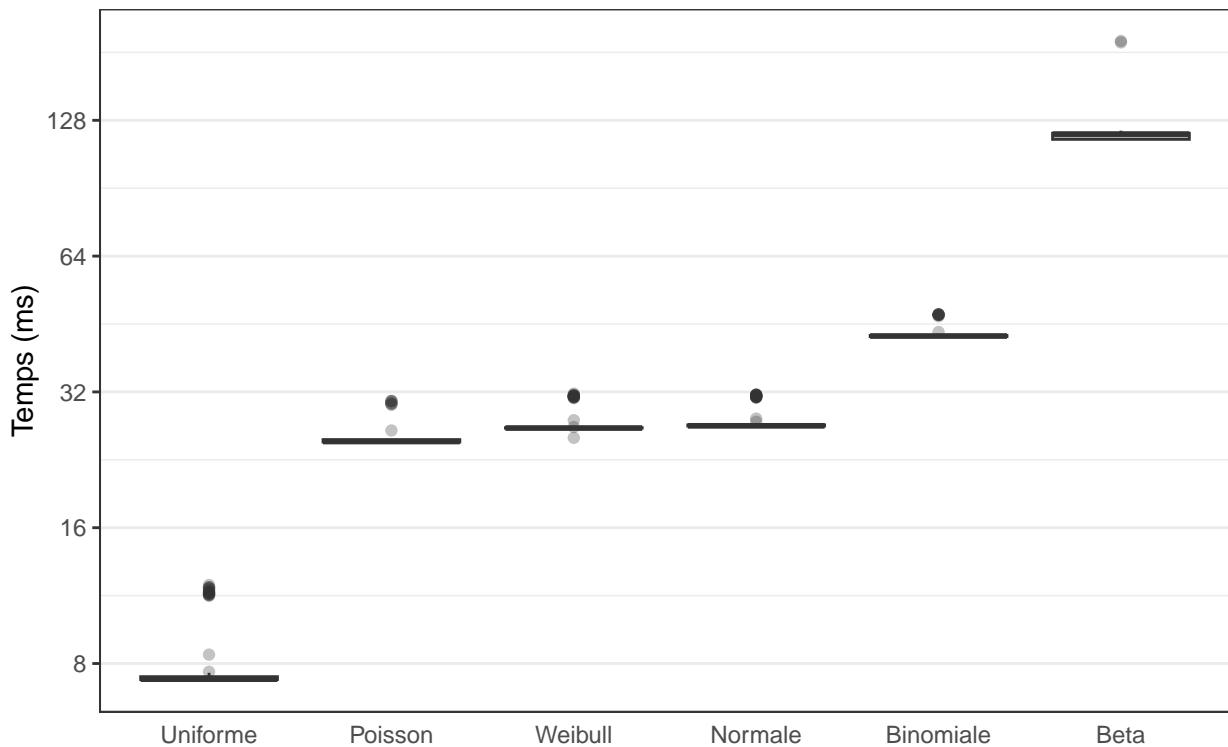


Figure 2: Temps de calcul des principales distributions d'intérêt pour $1e+06$ facteurs, (100 iter.)

Les critères de continuité et de normalité n'appellent que peu de commentaires. Ces critères permettent simplement de garantir la possibilité d'une infinité de valeurs dimensionnelles, dans l'intervalle considéré. Le critère de continuité proscrit toutefois l'utilisation de lois de distributions discrètes telles que la loi binomiale ou la loi de Poisson, et celui de normalité écarte des distributions telles que la loi de Weibull, dont la normalisation est parfois délicate.

Le critère d'asymétrie est un critère permettant de tenir compte des différents paramètres pouvant impacter la distribution de taille des spécimens prélevés, parmi lesquels :

- Différences de cinétique de croissance d'une famille à une autre,
- Particularités de la croissance fongique, notamment par la croissance hyphale,^{28,29}
- Probabilité de prélèvement variable selon la taille du spécimen (par difficulté de détection, considérations éthiques, intérêt mycologique ou gastronomique...).

Le premier paramètre évoqué précédemment n'a pu être exploité dans le cadre de cette étude en raison du manque de données concernant les cinétiques relatives de croissance des sporophores des différentes familles de macromycètes. Le modèle que nous proposons permet toutefois des développements ultérieurs dans ce domaine.

Les deux derniers paramètres permettent de supposer que la distribution de taille des spécimens d'une même espèce issus d'une récolte en vie réelle ne sera pas symétrique, d'une part en raison de la rapidité de la croissance fongique, et d'autre part parce que le prélèvement se fera préférentiellement en épargnant les spécimens de petite taille.

Ainsi, la génération de la variable aléatoire F_T obéira idéalement à une loi de distribution asymétrique vers la droite ($G_1 < 0$). Ce critère d'asymétrie écarte par conséquent les lois de distribution symétriques telles que la loi normale ou la loi uniforme.

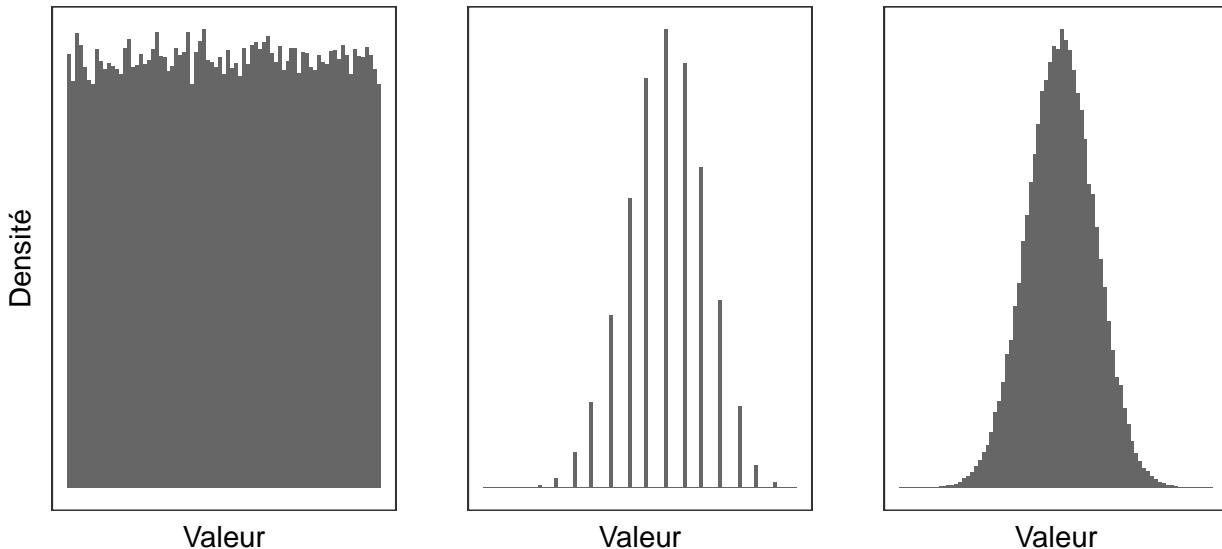


Figure 3: Exemples de distributions de la loi uniforme (à gauche), binomiale (au centre) et normale (à droite)

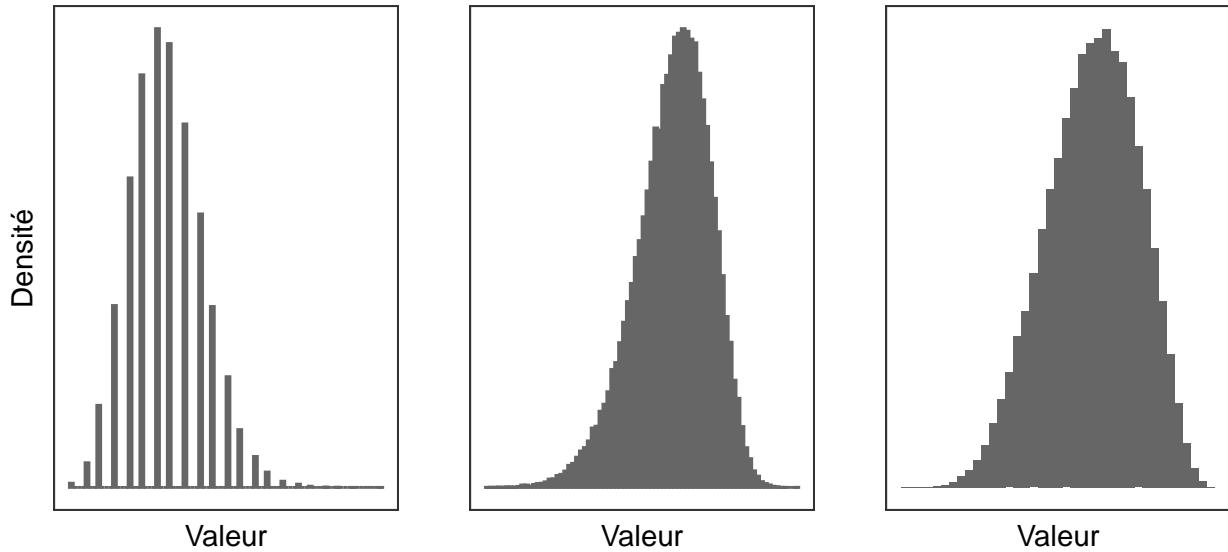


Figure 4: Exemples de distributions de la loi de Poisson (à gauche), de Weibull (au centre) et bêta (à droite)

En raison des contraintes imposées précédemment ainsi que de par sa grande polyvalence,³⁰ la loi retenue dans le cadre de cette étude pour la génération des facteurs de taille aléatoires (F_T) est une loi bêta non-centrale, définie comme la fonction de distribution de :^{30,31}

$$X = \frac{\chi^2_{2\alpha}(\lambda)}{\chi^2_{2\alpha}(\lambda) + \chi^2_{2\beta}}$$

Avec χ^2 la loi du khi-deux, α et β les paramètres de forme de la loi bêta, et λ son paramètre de non-centralité. Nous définirons empiriquement pour cette étude les relations suivantes entre ces trois paramètres :

$$\begin{cases} \alpha = 6 F_c & (\text{shape1}) \\ \beta = 4 & (\text{shape2}) \\ \lambda = F_c/2 & (\text{ncp}) \end{cases}$$

F_c est ici défini comme un facteur de croissance permettant de rendre compte de la cinétique de croissance de chaque variété d'une part, et du prélèvement préférentiel des spécimens de plus grande taille d'autre part, comme l'illustre la figure 5.

Le modèle défini à ce stade impose une stricte proportionnalité entre diamètre du chapeau D_c , diamètre du stipe D_s et longueur du stipe L_s . Dans un souci de réalisme, il apparaît souhaitable d'améliorer ce modèle mathématique en y ajoutant un facteur de dispersion, chargé de générer de légères variations des rapports entre ces trois caractéristiques dimensionnelles. Il nous semble ici souhaitable de proposer une distribution en cloche pour ce facteur de dispersion, afin de favoriser la génération de champignons “harmonieux”, c'est-à-dire de spécimens dont les proportions ne s'éloignent pas de celles typiques de leur espèce.

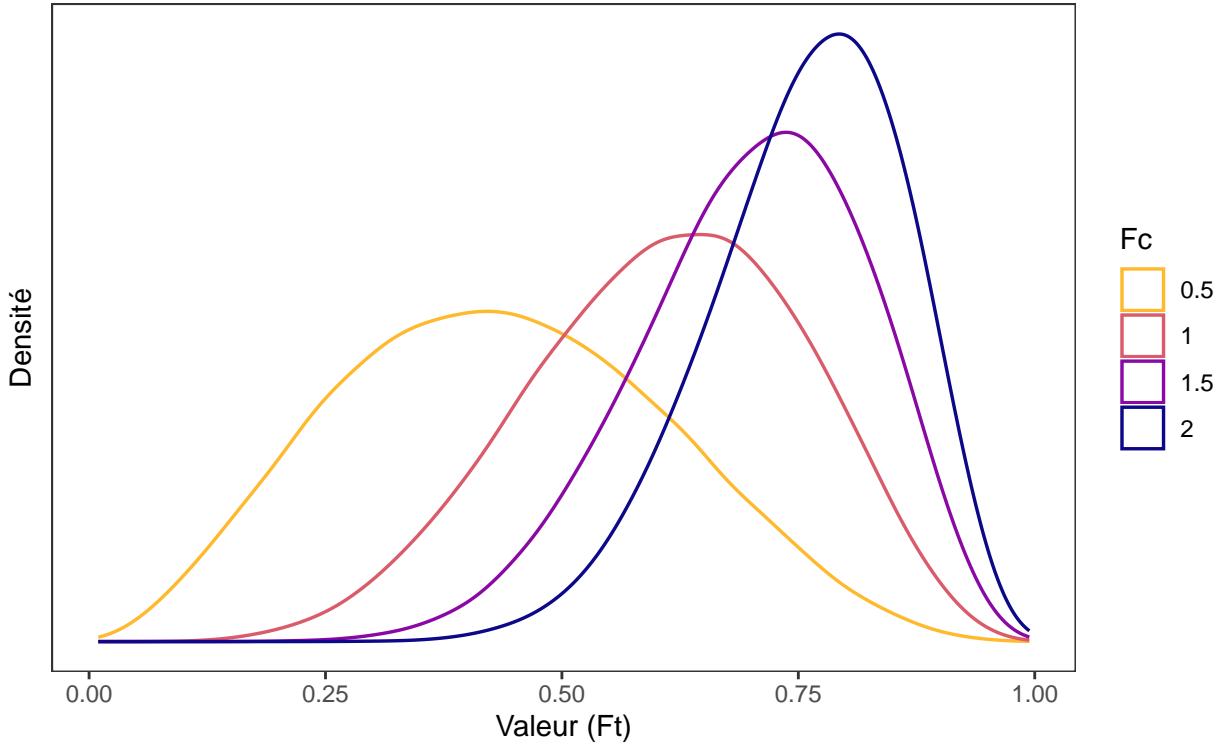


Figure 5: Distribution de différentes lois bêta, en fonction du facteur de croissance F_c

Nous proposons ainsi le modèle suivant :

$$\left\{ \begin{array}{ll} L_S = L_{Smax}.F_T.\delta_{Ls} & \text{avec } \delta_{Ls} \sim \mathcal{N}(\mu = 1; \sigma = 0.05) \\ D_S = D_{Smax}.F_T.\delta_{Ds} & \text{avec } \delta_{Ds} \sim \mathcal{N}(\mu = 1; \sigma = 0.05) \\ D_C = D_{Cmax}.F_T.\delta_{Dc} & \text{avec } \delta_{Dc} \sim \mathcal{N}(\mu = 1; \sigma = 0.05) \end{array} \right.$$

L'impact de cette dispersion sur la distribution des paramètres de taille L_S , D_S et D_C est illustré par les figures 6 et 7.

La dispersion ainsi créée permet de générer de légères variations des rapports entre les différents paramètres de taille, tout en se situant à proximité de la première bissectrice et majoritairement dans la zone 50-90% de la taille maximale (voir figure 7). Cette dispersion autorise par ailleurs l'existence d'une faible proportion de spécimens dépassant les valeurs dimensionnelles maximales généralement admises par la littérature.

Une simulation de Monte Carlo unidimensionnelle effectuée sur 10^5 spécimens nous permet ainsi d'évaluer la proportion de spécimens "hors normes" dépassant la valeur dimensionnelle maximale à environ 0.4 % (voir figure 8). La même simulation nous permet d'évaluer que la proportion de spécimens "exceptionnels", dépassant de plus de 10% cette valeur maximale, sera quant à elle inférieure à 0.01 %.

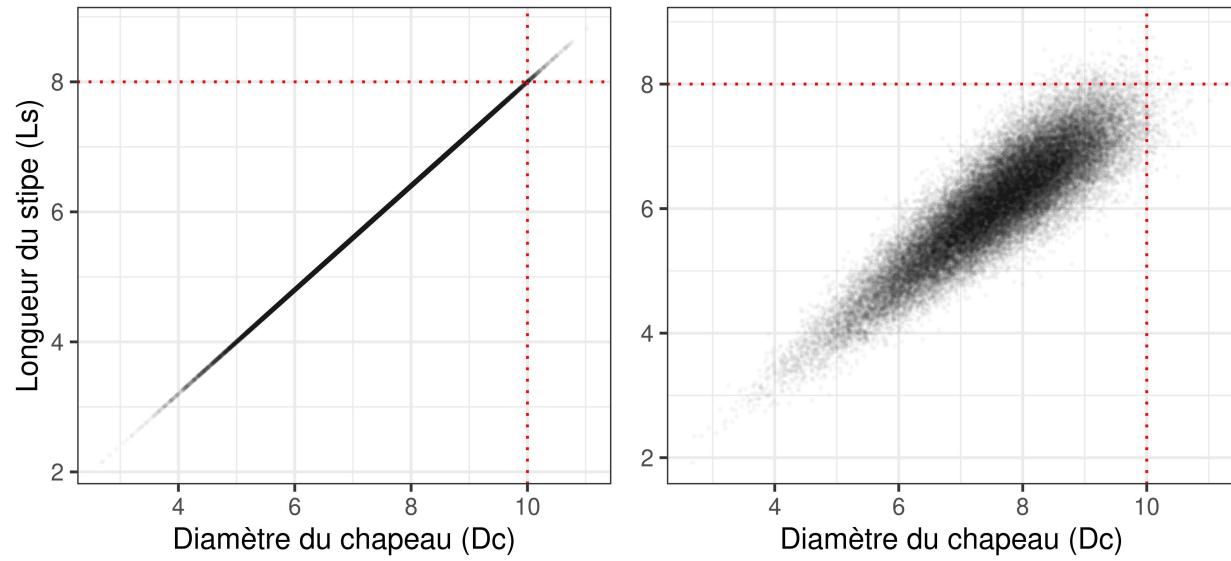


Figure 6: Nuages de points de 2 paramètres de taille (L_s et D_c), sans dispersion (à gauche) et avec dispersion (à droite), pour 50000 champignons

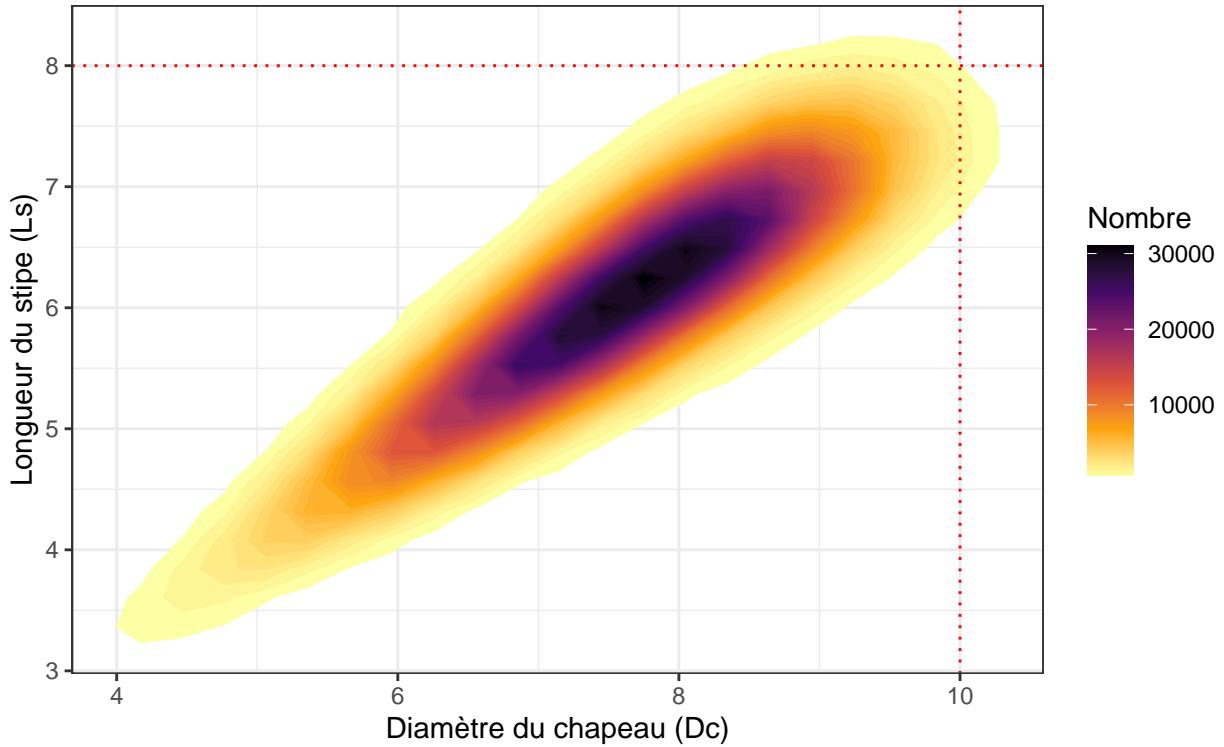


Figure 7: Diagramme de densité de 2 paramètres de taille, avec dispersion

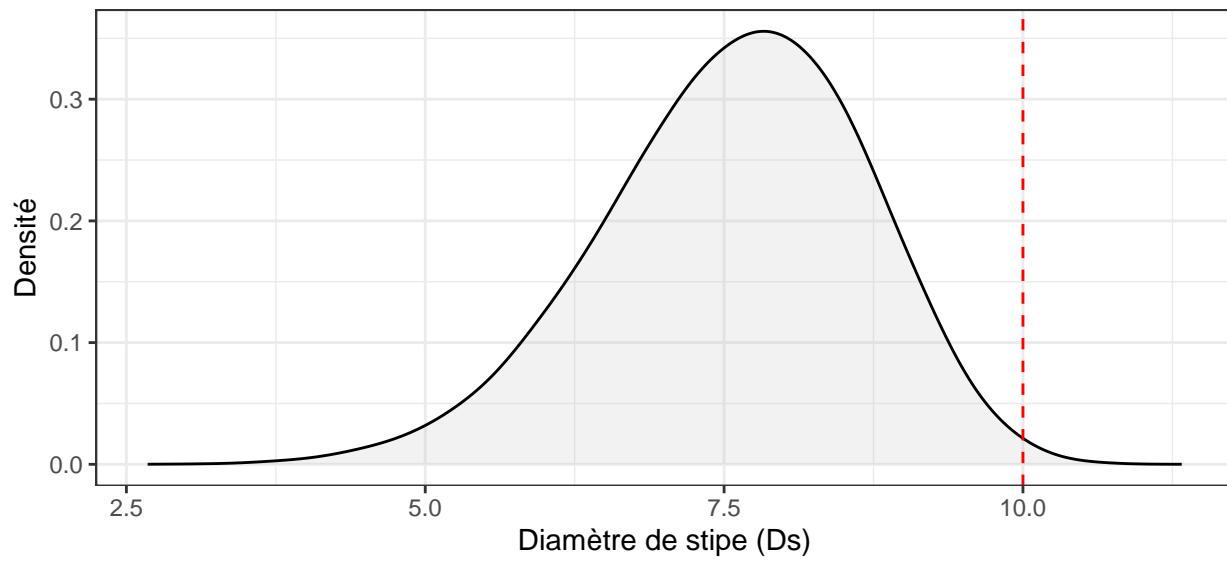


Figure 8: Distribution du diamètre de stipe D_s , pour $D_{s\max} = 10$

3.3.3 Principes de génération des paramètres qualitatifs

La génération des paramètres qualitatifs, tels que la couleur des spores ou le type d'hyménophore, est nettement moins complexe que celle des paramètres quantitatifs.

L'ensemble des valeurs quantitatives possibles pour un critère et pour une variété donnée est insérée dans un vecteur de valeurs, et une valeur sera tirée aléatoirement, et de manière équitable, parmi celles contenues dans ce vecteur afin de caractériser le paramètre en question pour chaque spécimen.

Toutefois, certaines caractéristiques ont pu, lors de notre recherche bibliographique, être identifiées comme rares. Afin d'en tenir compte, nous avons, lorsqu'une telle valeur "rare" est présente dans notre vecteur contenant les valeurs possibles, choisi de dupliquer à 10 les autres valeurs dans le vecteur de valeurs parmi lesquels aura lieu le tirage équitable. (cf. annexes, p. 92)

Notre générateur permettra donc aux caractéristiques "communes" d'avoir une probabilité d'être sélectionnées 10 fois plus élevée que celle de ces caractéristiques "rares".

4 Principes de l'apprentissage machine

L'apprentissage machine est un domaine scientifique se situant à l'interface entre les statistiques et l'informatique, et constitue un domaine de l'intelligence artificielle. L'objet de l'apprentissage machine est conceptuellement de permettre aux machines d' "apprendre" de manière autonome, c'est-à-dire d'évaluer et d'améliorer leurs performances sans l'intervention d'un programmeur, par l'entraînement.³²

En pratique, ce domaine se consacre à la génération d'algorithmes autonomes capables d'améliorer leurs performances prédictives par exposition à un lot de données et de prédire par inférence la meilleure décision ou réponse à une question.³³

4.1 Types d'apprentissage machine

L'apprentissage machine peut se dérouler suivant un certain nombre de paradigmes, concernant en particulier la stratégie d'apprentissage – définie à partir du type de données disponibles – parmi lesquels nous pouvons notamment citer :³²⁻³⁴

- Apprentissage machine supervisé : les données du problème sont disponibles sous forme de couples exemple-annotation. Chaque point de données possède des caractéristiques (covariables) et une annotation. Le principe de cet apprentissage est donc de générer une application liant les vecteurs des entrées (covariables) et une variable de sortie (annotation). L'apprentissage machine supervisé permet également de mesurer les performances du modèle obtenu.
- Apprentissage machine semi-supervisé : les données du problème contiennent une petite quantité de données annotées, associée avec une grande quantité de données non annotées. La présence de cette grande quantité de données non-annotées peut apporter une amélioration considérable de la précision des prédictions par rapport à l'utilisation d'un seul lot de données annotées de plus petite taille.
- Apprentissage machine non supervisé : les données ne sont pas annotées. Ce paradigme permet d'acquérir une connaissance concise de la structure du lot de données, notamment par des méthodes de partitionnement des données (*clustering*), de réduction de dimensionnalité ou de réseaux neuronaux. Il peut notamment être utilisé pour l'analyse exploratoire de données ou pour diminuer la quantité de données fournie à des algorithmes supervisés afin d'augmenter l'efficience calculatoire du système.

4.2 Jeux de données

Les jeux de données dédiés à l'apprentissage machine supervisé ou semi-supervisé sont tous construits sur la base de couples données-résultats. Selon l'étape de ces types d'apprentissage machine, le résultat peut être fourni à la machine ou lui être caché, le but étant dans le premier cas de permettre à la machine d'effectuer son apprentissage, et dans le second cas d'évaluer les performances de la prédition par rapport au résultat réel.

Le déroulement de l'apprentissage machine supervisé se décompose conceptuellement en trois étapes principales, mettant en jeu trois lots de données distincts :

1. Entraînement : le modèle d'apprentissage machine est exposé à un *jeu de données d'entraînement* (*training data set*), censé être représentatif^a des données auquel le modèle sera exposé en utilisation réelle. Cette phase est la phase d'apprentissage du modèle.
2. Validation : le modèle d'apprentissage machine élaboré à l'étape précédente, est ici soumis à un *jeu de données de validation* (*validation data set*) et tentera d'apporter des prédictions quant à une variable d'intérêt considérée comme le résultat (ex: comestibilité, espèce...), sur la base des informations contenues dans le lot de données de validation (ex : dimensions, couleurs, morphologie du champignon...). Ces prédictions sont comparées avec les valeurs réelles (ex : comestibilité, espèce...), ce qui permet d'évaluer les performances du modèle proposé en fonction des indicateurs retenus (spécificité, sensibilité, indice de Rand, temps de calcul...). Les étapes d'apprentissage et de validation sont répétées de manière itérative en explorant l'ensemble des paramètres de configuration du modèle (hyperparamètres) à fins d'optimisation.
3. Test : les performances du meilleur modèle (avec hyperparamètres optimaux), sélectionné à l'issue de l'étape de validation, sont évaluées vis-à-vis d'un *jeu de données test* (*test ou holdout data set*).

La séparation entre étapes d'optimisation et de test peut sembler artificielle. Le problème est en partie lié à un flou sémantique : si l'étape initiale d'entraînement ou d'apprentissage ne pose que peu de problèmes conceptuels, l'étape intermédiaire, dite de *validation* correspond en réalité à une étape d'*optimisation* du modèle et de ses hyperparamètres. Par ailleurs, l'étape finale de *test* sera parfois qualifiée d'étape de *validation* dans la littérature, ce qui peut entretenir la confusion entre ces étapes.³⁵

Une distinction sémantique plus nette entre phases d'*apprentissage*, d'*optimisation* et de *test* permet de comprendre plus aisément le fondement épistémologique de cette dernière phase pouvant parfois sembler superflue : l'optimisation effectuée lors de l'étape de validation aboutit à un modèle potentiellement biaisé par surajustement (problème dit d'*overfitting*) vis-à-vis du jeu de données utilisé comme référence lors de cette étape. Seule une exposition du modèle à des données n'ayant jamais servi à son entraînement ou son optimisation permettra réellement d'évaluer avec précision son caractère prédictif, donc sa validité.^b

Les phases d'entraînement, d'optimisation et d'évaluation utilisent chacune un lot de données spécifique. Chacun de ces lots de données est habituellement obtenu suite à dichotomies successives (voir figure 9) du lot de données initial, avec des proportions pouvant être variables d'une scission à l'autre :

^aVoir section 3.3.1

^bDans un souci de clarifier le propos, nous utiliserons les termes de lots et de phases d'entraînement, d'optimisation et d'évaluation dans la suite de cette étude.

1. Découpage du jeu de données initial, en un jeu d'évaluation d'une part, et un jeu d'entraînement et optimisation d'autre part,
2. Découpage du jeu de données entraînement et optimisation, en un jeu d'entraînement et un jeu d'optimisation.

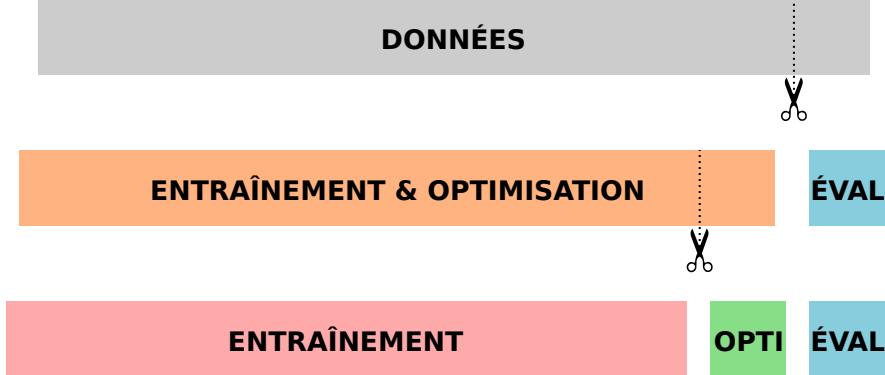


Figure 9: Principe de scissions successives d'un jeu de données initial en jeux d'apprentissage, d'optimisation et d'évaluation.

Les rapports de taille entre jeux de données d'entraînement, d'optimisation et d'évaluation de cette étude suivront la loi $p : \sqrt{p} : \sqrt{p} + 1$, avec p le nombre de coefficients du modèle.³⁶

Ce nombre p de coefficients peut être approché par l'expression $p \approx \sqrt{N_u}$, avec N_u le nombre de lignes uniques de notre jeu de données, c'est-à-dire, dans notre cas, le nombre de champignons contenus dans le lot de données.³⁶

Il est possible d'obtenir ce rapport $p : \sqrt{p} : \sqrt{p} + 1$ par une première scission entre jeu d'entraînement et optimisation d'une part (de taille relative $p + \sqrt{p}$) et jeu d'évaluation d'autre part (de taille relative $\sqrt{p} + 1$), avec, pour ce dernier, une taille représentant la fraction du lot total :

$$f_{test} = \frac{\sqrt{p} + 1}{p + 2\sqrt{p} + 1} = \frac{1}{\sqrt{p} + 1}$$

Cette première dichotomie peut être suivie par une seconde dichotomie entre jeu d'entraînement (de taille relative p) et jeu d'optimisation (de taille relative \sqrt{p}), de fraction :

$$f_{opti} = \frac{\sqrt{p}}{p + \sqrt{p}} = \frac{1}{\sqrt{p} + 1} \Rightarrow f_{test} = f_{opti} = \frac{1}{\sqrt{p} + 1}$$

En pratique, pour notre lot de données contenant $N_u = 98700$ spécimens, nous pouvons calculer $p \approx 314.2$, soit deux dichotomies successives de ratio 18:1.

4.3 Méthodes de construction des jeux de données

Les méthodes de division mises en œuvre dans cette étude appellent quelques précisions, car elles apportent certaines améliorations par rapport à l'utilisation de deux scissions successives effectuées de manière purement aléatoire.

La première division, entre jeu d'entraînement/optimisation et jeu d'évaluation, mettra en œuvre une méthode de découpage basée sur les points-supports³⁷ (*support-points based splitting*) exploitant un algorithme du plus proche voisin (*nearest neighbour*), afin d'optimiser la représentativité des jeux de données par rapport à ceux pouvant être obtenus par un découpage aléatoire.^{38,39}

Notre seconde division, entre jeu d'entraînement et d'optimisation, exploitera quant à elle la méthode de validation croisée à k blocs (*k-folds cross-validation*). Le principe de la validation croisée repose sur une rotation de la séparation créée entre jeux d'entraînement et d'optimisation (voir figure 10).

Le jeu d'entraînement/optimisation est découpé, de façon aléatoire, en k blocs de données de taille égale, dont k-1 sont utilisés pour l'entraînement du modèle prédictif et 1 pour son optimisation. Cette opération est répétée k fois, en utilisant un jeu d'optimisation différent à chaque itération. L'évaluation de la performance globale s'effectue en évaluant la performance moyenne des k itérations. Cette méthode permet de limiter les biais potentiels générés par une simple dichotomie des données d'entraînement et d'optimisation en exploitant la totalité des données du lot afin d'effectuer ces deux tâches.

Comme démontré précédemment, une validation croisée *k-folds* avec $k = 19$ permettrait d'optimiser l'apprentissage et l'optimisation des modèles de cette étude.³⁶

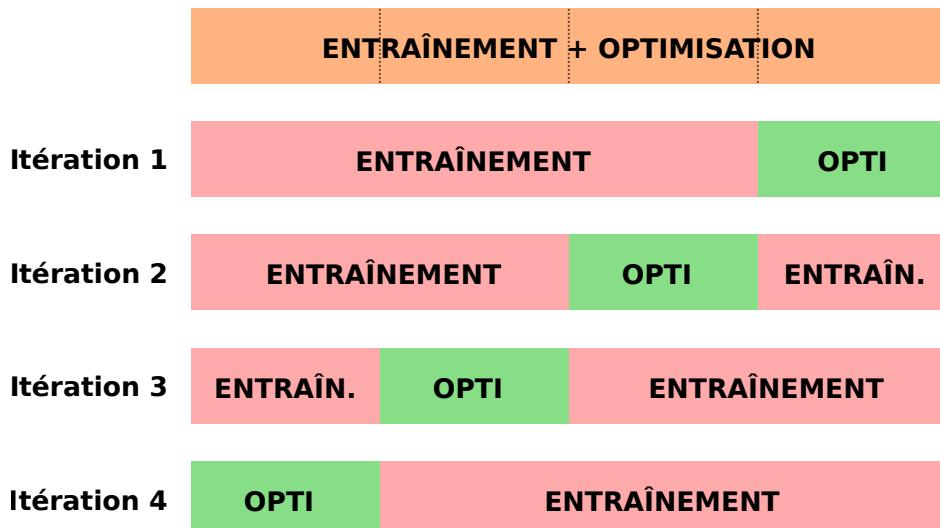


Figure 10: Principe de la validation croisée à k blocs (*k-fold cross-validation*), pour $k = 4$.

4.4 Modèles utilisés

4.4.1 Analyses discriminantes

Cette étude proposera plusieurs classificateurs s'appuyant sur des méthodes d'analyse discriminante, en particulier l'analyse discriminante linéaire (LDA : *Linear Discriminant Analysis*).

L'analyse discriminante linéaire est une méthode ayant été proposée par Ronald Fisher en 1936^{40,41} pour résoudre des problèmes de classification taxonomique dans le domaine de la botanique.^c La LDA est basée sur la construction de l'hyperplan de projection permettant de maximiser la distance entre les moyennes projetées des différentes classes et de minimiser la variance intraclasse (voir figure 11).⁴² La LDA peut être utilisée à fins de classification, mais aussi pour effectuer des réductions de dimensionnalité ou encore afin de faciliter l'interprétation de l'importance de certaines caractéristiques.

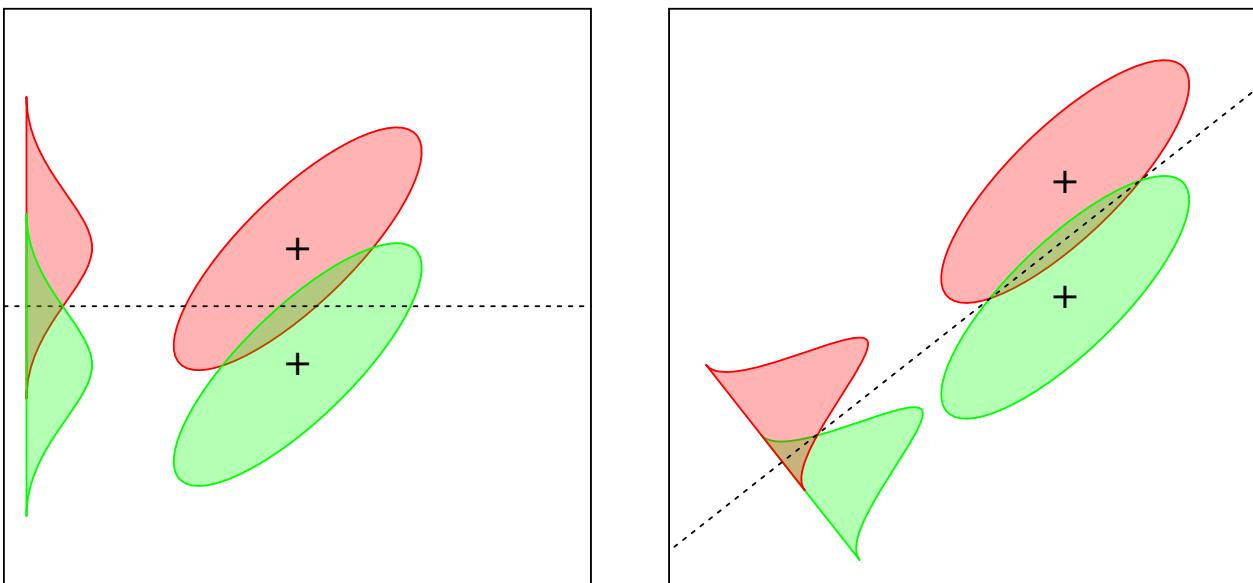


Figure 11: Séparation par distance maximale des moyennes interclasses (à gauche), et par projection sur l'hyperplan optimal tenant compte des variances intraclasse (LDA, à droite)

En pratique, la LDA consiste à construire un indice synthétique, combinaison linéaire des caractéristiques des classes, dont les coefficients permettent de rendre les points du problème originel le plus aisément "séparables". La LDA étant utilisée dans cette étude pour construire un classifieur binaire, c'est ce type de classificateur qui sera présenté dans cette section, et illustré un

^cCette étude, proposant une méthode de classification des variétés *Iris setosa*, *Iris virginica* et *Iris versicolor* est par ailleurs à l'origine du célèbre jeu de données *Iris*.

exemple extrait du jeu de données *Iris*, dans laquelle nous séparerons les espèces *Iris versicolor* et *Iris setosa*.

Dans ce cadre, la LDA vise ainsi à définir la fonction linéaire :

$$X = \sum_{i=1}^n \lambda_i \cdot x_i$$

avec n le nombre de paramètres caractérisant les individus, x_i les caractéristiques mesurées pour chaque individu et chaque paramètre i , et λ_i des coefficients à optimiser, de sorte que la fonction X maximise le rapport entre les différences des moyennes de chaque classe D et la somme des produits des caractéristiques intraclasses S (proportionnelle à la variance intraclasse), définis par :

$$D = \sum_{i=1}^n \lambda_i \cdot d_i$$

avec d_i la différence des caractéristiques moyennes pour chaque paramètre i , et

$$S = \sum_{p=1}^n \sum_{q=1}^n \lambda_p \cdot \lambda_q \cdot S_{pq}$$

avec S_{pq} la somme des produits des caractéristiques intraclasses pour chaque combinaison de paramètres p et q .

L'application sur les espèces *Iris versicolor* et *Iris setosa* nous donne les résultats présentés dans les tables 1 et 2 :

Table 1: Moyennes et différences de moyennes des 4 paramètres d'Iris setosa et versicolor

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
difference	-0.930	0.658	-2.798	-1.080

Table 2: Produits des différences à la moyenne des 4 paramètres d'Iris setosa et versicolor (S_{pq})

	S.I	S.w	P.I	P.w
S.I	19.1434	9.0356	9.7634	3.2394
S.w	9.0356	11.8658	4.6232	2.4746
P.I	9.7634	4.6232	12.2978	3.8794
P.w	3.2394	2.4746	3.8794	2.4604

La maximisation du rapport entre les carrés des distances des moyennes interclasses et les variances intraclasses revient à maximiser D^2/S pour chaque coefficient λ_i soit, par dérivation pour chacun des λ_i :

$$\frac{\partial}{\partial \lambda_i} \frac{D^2}{S} = 0 \Leftrightarrow \frac{1}{S} \frac{\partial}{\partial \lambda_i} D^2 + D^2 \frac{\partial}{\partial \lambda_i} \frac{1}{S} = 0 \Leftrightarrow \frac{D}{S^2} \left(2S \frac{\partial D}{\partial \lambda_i} - D \frac{\partial S}{\partial \lambda_i} \right) = 0 \Leftrightarrow \frac{1}{2} \frac{\partial S}{\partial \lambda_i} = \frac{S}{D} \frac{\partial D}{\partial \lambda_i}$$

En supposant que les distributions des classes soient unimodales, cette équation admet une solution unique. Le rapport S/D étant un facteur supposé constant pour tous les coefficients λ_i inconnus, ces coefficients sont donc les solutions du système :

$$\begin{cases} d_1 = S_{11}\lambda_1 + S_{12}\lambda_2 + S_{13}\lambda_3 + S_{14}\lambda_4 \\ d_2 = S_{21}\lambda_1 + S_{22}\lambda_2 + S_{23}\lambda_3 + S_{24}\lambda_4 \\ d_3 = S_{31}\lambda_1 + S_{32}\lambda_2 + S_{33}\lambda_3 + S_{34}\lambda_4 \\ d_4 = S_{41}\lambda_1 + S_{42}\lambda_2 + S_{43}\lambda_3 + S_{44}\lambda_4 \end{cases} \Rightarrow \mathbf{S} \cdot \boldsymbol{\lambda} = \mathbf{D} \Leftrightarrow \boldsymbol{\lambda} = \mathbf{S}^{-1} \cdot \mathbf{D}$$

avec \mathbf{S} la matrice des produits S_{pq} , \mathbf{D} le vecteur des différences des moyennes d_i et $\boldsymbol{\lambda}$ celui des coefficients λ_i .

En indiquant les facteurs :

- $i = 1$ pour la longueur de sépale L_s ,
- $i = 2$ pour la largeur de sépale ℓ_s ,
- $i = 3$ pour la longueur de pétale L_p ,
- $i = 4$ pour la largeur de pétale ℓ_p .

Nous pouvons calculer les coefficients :

$$\begin{cases} \lambda_1 = 0.0311507 \\ \lambda_2 = 0.1839077 \\ \lambda_3 = -0.222104 \\ \lambda_4 = -0.3147364 \end{cases}$$

Soit, après normalisation sur le facteur λ_1 :

$$\begin{cases} \lambda_1 = 1 \\ \lambda_2 = 5.904 \\ \lambda_3 = -7.13 \\ \lambda_4 = -10.104 \end{cases}$$

$$X = L_s + 5.904 \cdot \ell_s - 7.13 \cdot L_p - 10.104 \cdot \ell_p$$

Le seuil de séparation est alors défini par :

$$X_{sep.} = \frac{\overline{X_{ver.}} + \overline{X_{set.}}}{2}$$

Avec $\overline{X_{ver.}}$ et $\overline{X_{set.}}$ les moyennes respectives des X pour *Iris setosa* et *Iris versicolor*.

La valeur absolue des coefficients λ_i calculés précédemment nous indique la pondération de chaque caractère dimensionnel dans l'indice synthétique X permettant d'obtenir une séparation optimale, ainsi que l'illustrent les figures 12 et 13.

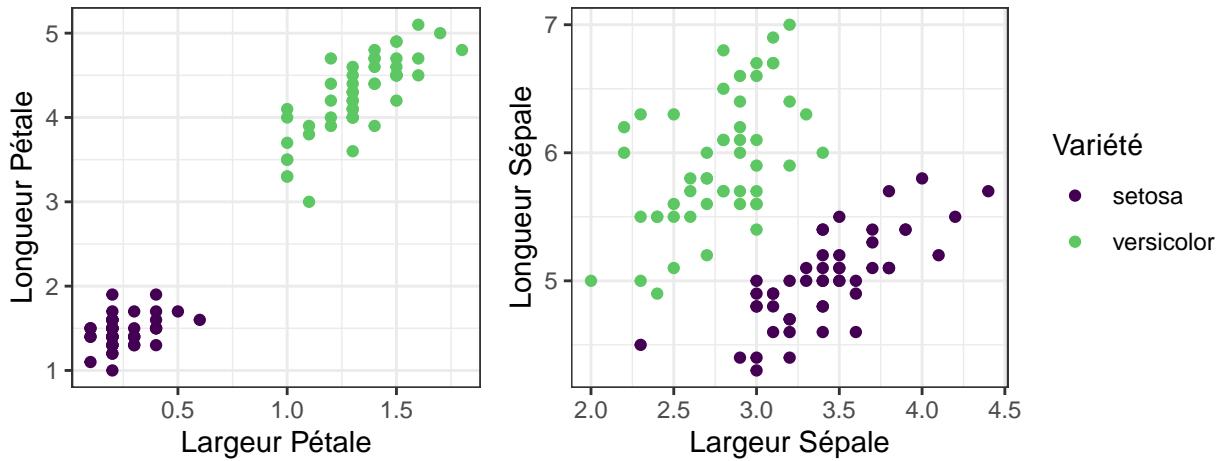


Figure 12: Distribution des variétés setosa et versicolor en fonction de leurs caractéristiques (paramètres fortement pondérés à gauche, faiblement pondérés à droite)

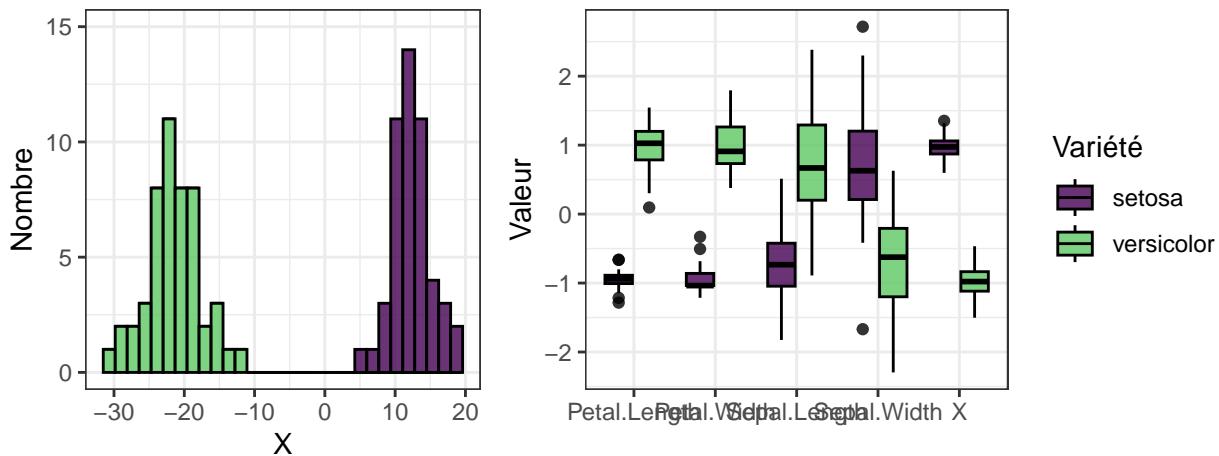


Figure 13: Distribution de X (à gauche) et des paramètres dimensionnels normalisés (à droite) en fonction des espèces

Nous pouvons conclure cette présentation de la LDA en évoquant ses limites, soulignées par certains développements de cette section :

- Les modèles d'analyse discriminante ne sont adaptés qu'à des données quantitatives ou qualitatives ordinaires,
- Dans la LDA, la séparation entre les différentes classes est basée sur la construction d'un indice synthétique purement linéaire : sur l'hyperplan de projection, la séparation entre "territoires" des différentes classes sera toujours une droite,

- La plupart des modèles d'analyse discriminante se basent sur des hypothèses tacites concernant les distributions des spécimens, notamment normalité et homoscédasticité.

4.4.2 Arbres de décision

Les arbres de décision sont l'une des formes les plus populaires de représentation des connaissances utilisées dans le *data mining*.⁴³ Une des raisons de cette popularité est que les modèles obtenus sont généralement considérés comme performants, et que l'apprentissage comme les prédictions sont aisément compréhensibles, contrairement à certaines approches dont la complexité peut les rapprocher de modèles assimilables à des boîtes noires.⁴⁴

Les arbres de décision peuvent aussi bien être utilisés pour des problèmes de classification (prédictions de classes) que de régression (prédiction de valeurs numériques). Les arbres de décision sont des structures hiérarchiques, séquentielles, composées de nœuds reliés par des branches. Ces nœuds sont conceptuellement de deux types :

- Les nœuds internes, qui possèdent des descendants ; le premier d'entre eux, dépourvu de prédécesseur, est qualifié de racine. A chacun de ces nœuds est associé un test, à deux ou plusieurs issues, dont chacune constitue une branche qui aboutira au nœud suivant.
- Les nœuds terminaux, dépourvus de descendants, qualifiés de feuilles, responsables de la prédiction : classe (arbre de classification) ou valeur, voire modèle numérique (arbre de régression).

Le fonctionnement de l'arbre de décision s'articule donc sur une succession de tests – les plus importants se situant près de la racine – dont les résultats permettent d'avancer dans la hiérarchie de l'arbre, étape après étape, afin d'aboutir à une décision finale, dans une démarche pouvant mimer certains aspects du raisonnement humain.⁴⁴

Les tests des arbres de décision peuvent être :

- Univariés, usuellement sous forme de test d'inégalité de type $x_i < S$ avec x_i la caractéristique à mesurer et S un seuil permettant la décision. Ce test d'inégalité aboutit typiquement à une bifurcation, mais certains modèles peuvent regrouper plusieurs tests sur une caractéristique unique pour aboutir à une multifurcation.
- Multivariés, qui exploitent plus d'une caractéristique. Le cas le plus habituel est la scission oblique (*oblique split*), basée sur un hyperplan représentant une combinaison linéaire de caractéristiques. Ce type de test multivarié se rapproche ainsi de la LDA.^d

Les différents modèles d'arbres de classification pourront mettre en œuvre des tests purement univariés (ID3, C4.5, CHAID...), purement obliques (arbres dits *perceptron*) ou autoriser les deux types de tests (arbres CART : *Classification And Regression Trees*).⁴⁵

^dcf. section 4.4.1, page 37

L'algorithme le plus utilisé pour la construction de la structure des arbres de classification est l'induction du haut vers le bas, basée sur l'algorithme classique dit "diviser pour régner" (*divide and conquer*), qui consiste à découper le problème de base en sous-problèmes, puis à résoudre ces sous-problèmes avant de les combiner en une solution globale.

En pratique, un nœud racine est tout d'abord créé, auquel sera associé la totalité du jeu d'apprentissage. En partant de ce nœud initial sera appliqué un algorithme récursif qui essaie de diviser ces données à chaque nœud. A chaque étape de cet algorithme, l'opportunité d'une scission est évaluée et chaque nœud créé pourra soit être marqué en tant que feuille (auquel cas l'induction est terminée pour cette branche), soit en tant que nœud interne (auquel cas l'algorithme se poursuit et l'arbre continuera à s'étoffer). Notons qu'à chaque étape de cet algorithme, la taille du jeu de données disponible pour les nœuds descendants se réduira en raison de la scission, et par conséquent, la probabilité que ces descendants deviennent des feuilles augmentera.

L'optimalité du critère de scission appliqué par chaque nœud est usuellement obtenue en minimisant un critère dit d'*impureté*, représentatif :

- De l'inhomogénéité des classes peuplant chaque nœud postérieur à la scission pour les arbres de classification,
- De la somme de la valeur absolue des résidus ou de leurs carrés pour les arbres de régression.

Pour l'arbre CART qui sera utilisé dans la classification binaire de cette étude, le critère d'impureté mise en œuvre au sein du modèle est le coefficient de Gini :⁴³

$$G = 1 - p^2 - (1 - p)^2$$

avec p la fréquence relative de l'une des deux classes, au sein du nœud considéré.

Les arbres obtenus peuvent enfin être rationalisés à l'aide d'un algorithme d'élagage (*pruning*) visant à réduire la complexité de l'arbre. Cet algorithme effectue l'analyse de l'arbre des feuilles vers la racine – donc dans le sens inverse de l'induction qui l'a construit – et en évaluant si chaque nœud intermédiaire pourrait être avantageusement remplacé par une feuille ou par une branche. Différents critères d'optimisation existent, parmi lesquels nous pouvons citer l'élagage coût-complexité,⁴⁶ exploité par le modèle CART, et qui vise à minimiser le facteur coût-complexité $CC(T)$ de notre arbre T :

$$CC(T) = Erreur(T) + \alpha \times Taille(T)$$

Avec α un hyperparamètre de complexité^e strictement positif évalué expérimentalement.^f

Les résultats de la mise en œuvre d'une classification par arbre de décision sur le jeu de données Iris sont représentés par la figure 14.

^ecf. section C.3.1, p. 101

^fcf. section 4.5, p. 48

Des exemples d'application sur les macromycètes sont également illustrés par les figures 15 et 16, qui permettent d'appréhender la potentielle complexité de la classification des champignons, même en limitant la prédiction à un critère binaire (comestible ou toxique).

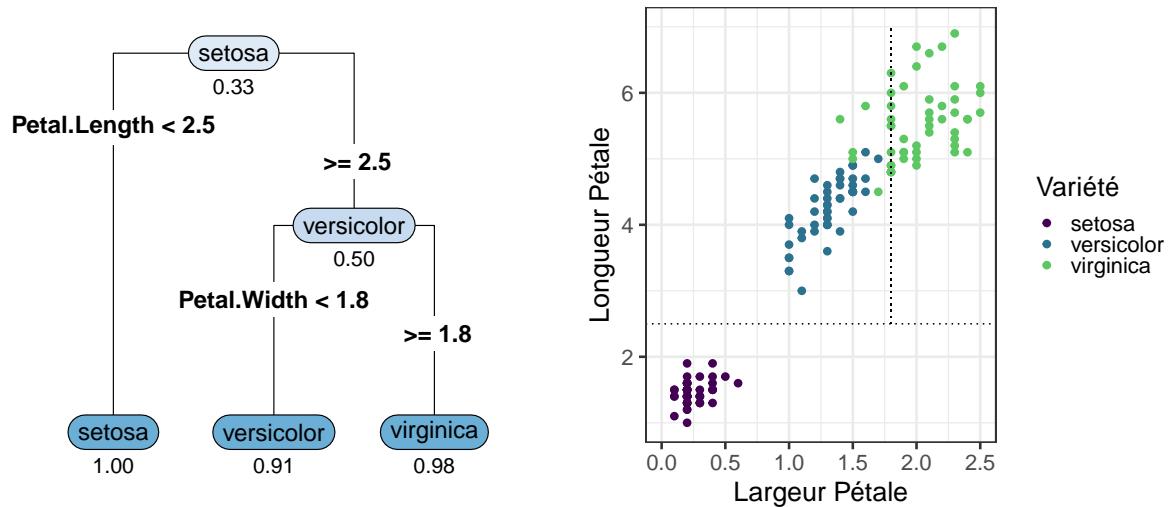


Figure 14: Arbre et critères de classification des trois espèces du lot de données Iris

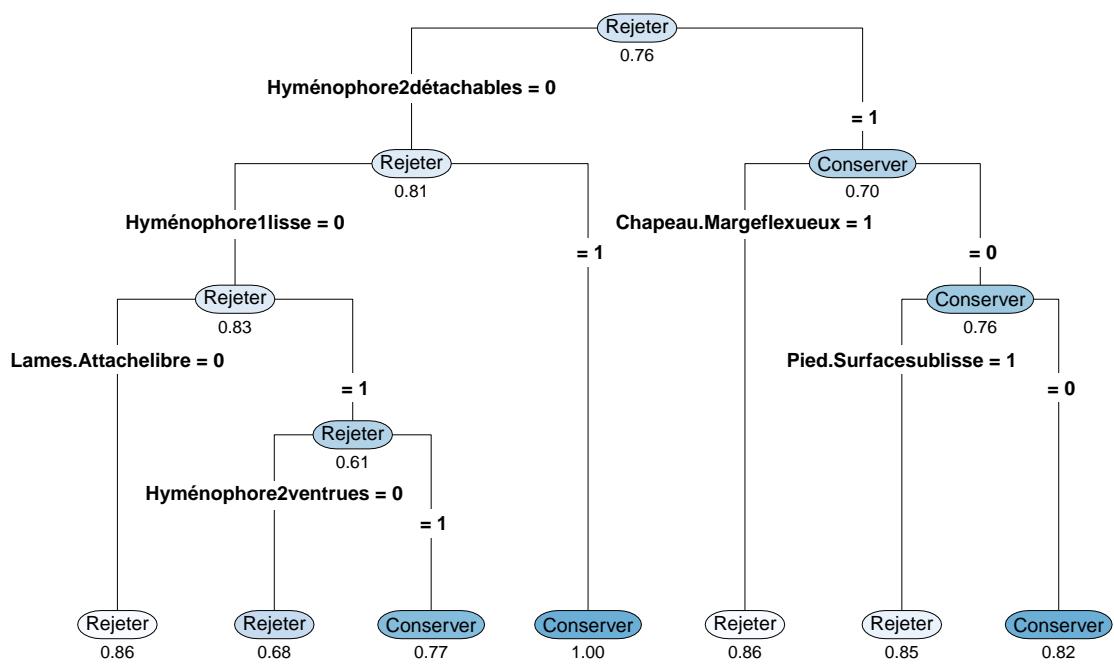


Figure 15: Arbre simplifié proposé par CART pour la classification des champignons

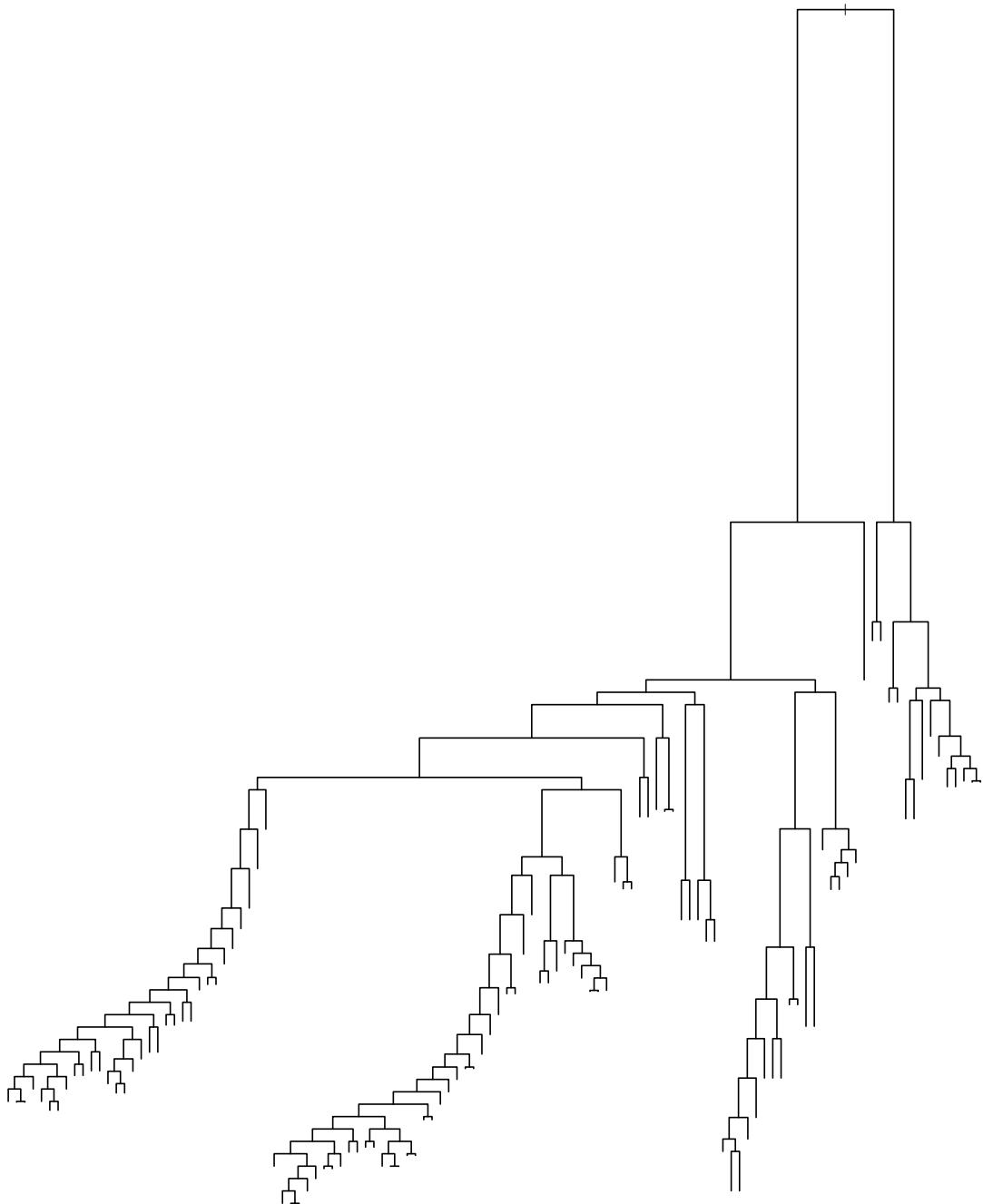


Figure 16: Structure arborescente complète proposée par CART pour la classification des champignons

4.4.3 Forêts aléatoires

Les arbres décisionnels conventionnels ont montré, au cours de leur utilisation dans le traitement des données, quelques limites récurrentes :⁴⁸

- Tendance à l'instabilité des structures arborescentes, même avec des perturbations mineures : une petite modification du lot d'entraînement peut entraîner des modifications majeures de l'arborescence d'un arbre décisionnel,
- Apparition, grâce notamment aux avancées de la génomique, de problèmes dits de type *grand p, petit n*, avec un grand nombre de prédicteurs p pour un petit nombre d'observations n ,^g qui mettent en échec la vision classique de parcimonie d'un modèle unique pour un problème unique,
- Performances limitées face à des combinaisons linéaires de facteurs,
- Impossibilité usuelle d'inférence théorique, liée à la nature adaptative de la construction des arbres décisionnels.

La génération de forêts s'est imposée comme une solution élégante à ces limitations. Une forêt est un ensemble d'arbres qui, individuellement, sont suboptimaux, mais dont la combinaison en un comité améliore considérablement les performances.⁵⁰ L'exploitation de nombreux arbres offre la possibilité d'utiliser plus d'informations, et d'avoir ainsi une meilleure compréhension des données : des arbres différents peuvent proposer des cheminements alternatifs vers une solution.

Une forêt aléatoire est un type particulier de forêt, dont la définition canonique correspond, d'après Breiman,⁵¹ à une collection $(\hat{h}(\cdot, \Theta_1), \dots, \hat{h}(\cdot, \Theta_q))$ d'arbres de prédiction avec $\Theta_1, \dots, \Theta_q$ des variables indépendantes, identiquement distribuées, aléatoirement sélectionnées dans le lot de données \mathcal{L}_n comptant n observations et q prédicteurs.

L'agrégation des données s'obtient par un vote majoritaire :

$$\hat{h}_{RF}(x) = \underset{1 \leq c \leq C}{\operatorname{argmax}} \sum_{\ell=1}^q \mathbf{1}_{\hat{h}(x, \Theta_\ell)=c}$$

Le type le plus commun de forêt aléatoire est la forêt aléatoire à entrées aléatoires (RF-RI : *Random Forests Random Inputs*).^h Le terme *entrées aléatoires* est ici à interpréter comme *variables d'entrées aléatoires* : le principe de la forêt RF-RI est de construire une forêt aléatoire avec des prédicteurs arborescents dont les variables d'entrées sont aléatoires, chacun construit à partir d'un échantillon bootstrapé.ⁱ

^gEn l'espèce, les problèmes posés par la génomique imposent souvent d'analyser des dizaines de milliers de gènes (p), sur quelques dizaines ou centaines de patients (n).

^hLes forêts RF-RI sont d'ailleurs le type de forêt aléatoire auquel il est souvent tacitement référence lorsque le terme *forêt aléatoire* est utilisé dans la littérature.

ⁱL'échantillonnage bootstrap correspond à la construction d'un échantillon par prélèvement d'individus avec remplacement.

En pratique, l'algorithme de génération d'une forêt aléatoire RF-RI, pour un lot de données structuré autour de n observations et p prédicteurs, fonctionne de la façon suivante, également résumée par la figure 17 :⁵¹

1. Prélever un échantillon bootstrap de taille n ,
2. Appliquer un partitionnement récursif sur cet échantillon bootstrap. A chaque nœud, sélectionner aléatoirement q prédicteurs parmi les p (avec $q \ll p$),
3. Poursuivre le partitionnement récursif jusqu'à la création d'un arbre^j : $\hat{h}(\cdot, \Theta_\ell)$, $1 \leq \ell \leq q$,
4. Répéter les étapes précédentes jusqu'à l'obtention d'une forêt, puis agréger les résultats, étape souvent présentée en classification comme le fruit du *vote des arbres*.

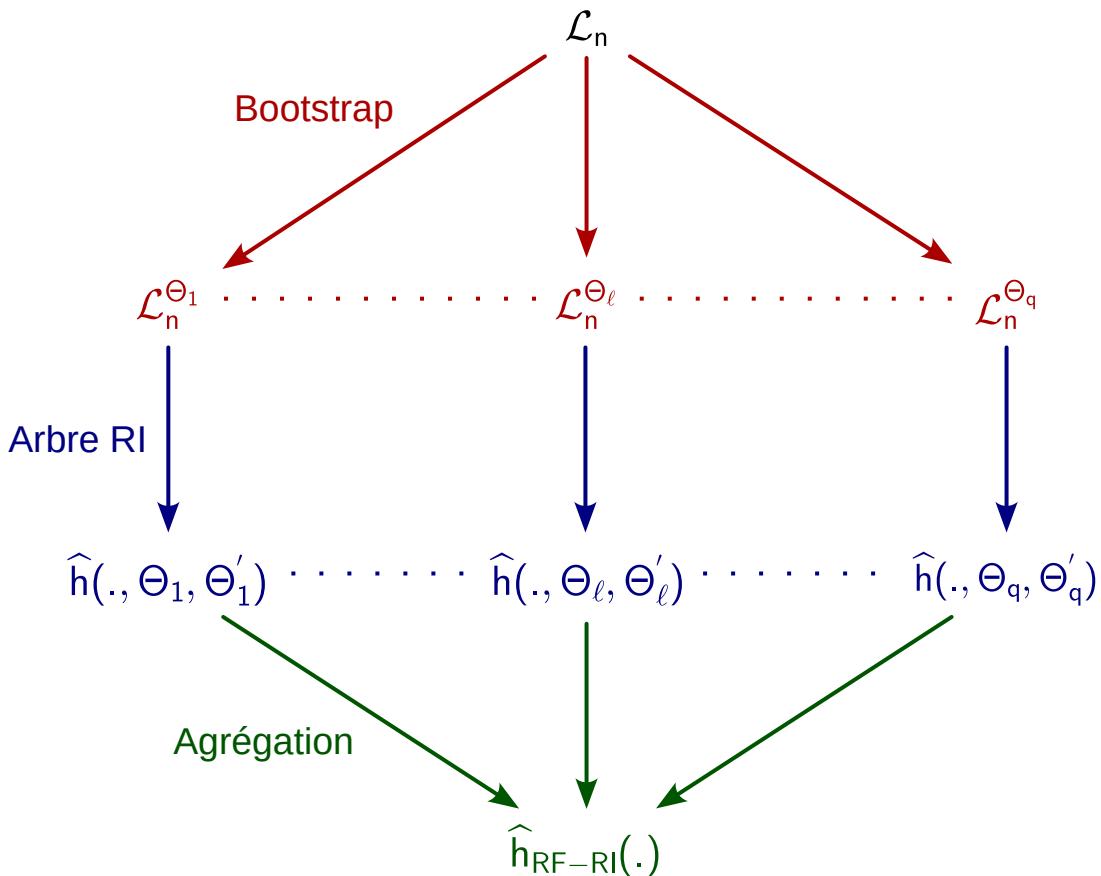


Figure 17: Algorithme de génération des forêts aléatoires de type RF-RI

Bien que les algorithmes de type forêts aléatoires soient des outils particulièrement performants, ils ont toutefois quelques limites, la principale étant la relative opacité du modèle obtenu, assimilable à une véritable “boîte noire”, et donc inadapté en tant qu’outil descriptif permettant d’acquérir une meilleure compréhension des données.

^jcf. section 4.4.2, page 41

4.5 Optimisation par plans d'expérience

Certains modèles nécessiteront une optimisation de leurs hyperparamètres, afin d'obtenir des performances maximales. Cette optimisation relève du domaine des plans d'expérience (*DOE : Design Of Experiments*). De nombreux plans et stratégies sont envisageables, le choix dépendant en partie des caractéristiques du processus à optimiser.

En effet, l'optimisation des paramètres d'un modèle informatique présente quelques particularités notables ayant un impact sur l'utilisation des plans d'expérience :

- La réalisation d'une expérience supplémentaire a un coût faible,
- Plusieurs métriques relatives aux performances peuvent coexister,^k
- La fonction de réponse peut s'avérer relativement complexe.

Ces particularités imposent d'explorer de manière méthodique la totalité de l'espace expérimental. Il existe une multitude de méthodes permettant de générer des plans expérimentaux dits SFD (*Space Filling Design*), afin d'optimiser l'occupation de l'espace expérimental. La méthode retenue pour cette étude sera celle des hypercubes latins, en raison de son utilisation répandue⁵² et de sa simplicité conceptuelle.

La méthode des hypercubes latins est une extension du principe des carrés latins. Un carré latin est une grille $n \times n$, remplie de n éléments distincts arrangés de sorte que chaque ligne et chaque colonne ne contienne qu'un seul exemplaire de chacun des n éléments. Dans le domaine des plans d'expériences, l'application des carrés latins revient à diviser un domaine expérimental bidimensionnel en une grille $n \times n$, et à placer une expérience et une seule sur chaque ligne et chaque colonne.

L'application du concept de carré latin dans un domaine expérimental à trois dimensions aboutit au cube latin. La généralisation dans un espace n -dimensionnel mène au concept d'hypercube latin.

De nombreux plans expérimentaux basés sur les hypercubes latins peuvent être générés. Nous pouvons citer principalement trois types d'hypercubes latins :

- Aléatoires,
- Optimisés, afin d'améliorer l'occupation spatiale,
- Orthogonaux, visant à minimiser la corrélation entre estimateurs des effets principaux.

Dans le cadre de cette étude, nous utiliserons des hypercubes latins quasi-orthogonaux, dont les propriétés nous permettront de modéliser de façon plus précise les performances de nos modèles en fonction de leurs paramètres de configuration (*hyperparamètres*).

^kcf. section 4.6, page 49

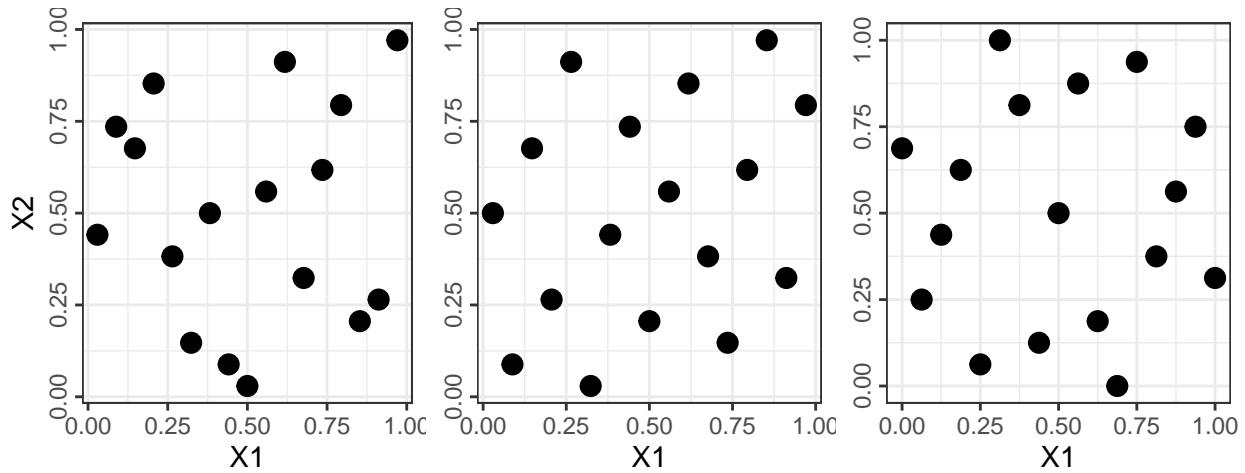


Figure 18: Carré latin aléatoire (à gauche), carré latin avec optimisation évolutive ESE maximin (au milieu), carré latin quasi-orthogonal (à droite)

Le but des plans expérimentaux de cette étude ne sera pas l'obtention d'une prédition exacte de la valeur de la réponse, mais plus modestement la recherche des facteurs permettant d'obtenir cette réponse optimale. A cet effet, la modélisation de la performance pourra s'effectuer à l'aide d'un modèle quadratique avec interactions, de formule générale :

$$Y = \beta_0 + \sum_{i=1}^k \beta_i \cdot X_i + \sum_{i < j} \sum_{j > 1}^k \beta_{ij} \cdot X_i \cdot X_j + \sum_{i=1}^k \beta_{ii} \cdot X_i^2 + \varepsilon$$

Avec β_n les coefficients des effets principaux, X_n leurs facteurs réduits et ε le terme d'erreur.

4.6 Évaluation des performances des modèles

L'optimisation des modèles ainsi que la comparaison de leurs performances relatives implique nécessairement de définir quel sera le critère vis-à-vis duquel cette performance sera évaluée.

De nombreux critères sont utilisables, en fonction du cahier des charges défini pour la résolution du problème, mais également du type de tâche effectuée : régression, classification binaire, classification multiclasse.

Dans une tâche de classification binaire, les critères usuels sont la spécificité, la sensibilité, et l'aire sous la courbe de fonction d'efficacité du récepteur (*AUC ROC*, parfois abrégé en *ROC*). Il conviendra bien évidemment, avant d'utiliser des indicateurs tels que la spécificité et la sensibilité, de définir la notion de test positif et test négatif.

D'autres indicateurs d'intérêt existent, nous retiendrons ici l'indice de Youden⁵³ pondéré, qui permet d'ajuster l'importance relative accordée à la spécificité et à la sensibilité, au sein d'un index synthétique.⁵⁴ Cet indicateur présente un intérêt particulier lorsqu'il apparaît souhaitable de tenir compte de la différence d'impact entre un faux positif et un faux négatif, sans pour autant autoriser des sensibilités ou spécificités trop faibles.

En l'espèce, l'index J de Youden pondéré nous permet d'élaborer un indice synthétique tenant compte du fait qu'il est plus grave de classer à tort comme comestible un champignon toxique que d'éjecter à tort un champignon parfaitement comestible – sans pour autant autoriser le modèle à éjecter un nombre inconsidéré de champignons comestibles.

L'indice de Youden pondéré est donné par :⁵⁴

$$J_w = 2 \cdot (w \cdot Sen + (1 - w) \cdot Spe) - 1 \quad \text{avec} \quad w \in [0; 1]$$

Avec *Sen* la sensibilité et *Spe* la spécificité du modèle.

Dans la classification binaire de cette étude, problème qui revient classiquement en mycologie à classer les espèces en fonction de leur comestibilité, la valeur positive sera ici arbitrairement attribuée à la valeur "champignon non-comestible". Nous cherchons donc à maximiser la sensibilité de la détection, afin d'éjecter les espèces toxiques, la spécificité – c'est-à-dire la capacité à ne pas éjecter trop d'espèces comestibles – apparaissant alors comme un critère relativement secondaire. En établissant arbitrairement un index J de Youden pondéré accordant 10 fois plus d'importance à la sensibilité qu'à la spécificité, nous pouvons établir $w = 10/11$.

Le problème de classification binaire étant relativement simple (*comestible* ou *non-comestible*), nous fixerons arbitrairement le critère de performance minimum à atteindre à $J_w \geq 0.999$, soit :

$$\begin{cases} Sen_{max} = 1 \Leftrightarrow Spe_{min} = 0.9945 \\ Spe_{max} = 1 \Leftrightarrow Sen_{min} = 0.99945 \end{cases}$$

Dans les tâches de classification multiclasse, d'autres indicateurs d'intérêt pourront être utilisés, tels que le kappa de Cohen, l'indice de Rand (*accuracy*, ou précision), mais aussi la sensibilité et la spécificité moyennes.

Nous retiendrons dans les classifications multiclassées effectuées au cours de notre étude le kappa de Cohen,⁵⁵ calculé à partir de la matrice de confusion (illustrée en figure 19), et donné par :

$$\kappa = \frac{\pi_0 - \pi_e}{1 - \pi_e}$$

Avec π_0 la probabilité d'accord entre notre modèle et la classe réelle du champignon, et π_e la probabilité d'un même accord résultant du pur hasard.

Landis et Koch ont élaboré une échelle de validité du kappa de Cohen, avec un accord qualifié de *quasi-parfait* pour $\kappa > 0.80$.⁵⁶ Nous considérerons donc que ce critère sera le minimum requis pour qu'un modèle de classifieur multiclasse élaboré au cours de cette étude puisse être considéré comme ayant des performances acceptables.

		Valeur réelle			
		A. arvensis	A. bitorquis	A. campestris	
Prédiction	A. arvensis	90	2	1	3
	A. bitorquis	1	89	1	2
	A. bitorquis	2	2	91	3
	A. campestris	1	1	1	85

Figure 19: Extrait d'une matrice de confusion, pour une classification multiclasse

L'interprétation du kappa pouvant parfois être assez contre-intuitive, cette étude la complétera parfois par l'indice de Rand R , métrique moins robuste en présence de données non-équilibrées¹, mais présentant l'avantage d'être de compréhension plus aisée, car représentant le pourcentage de prédictions exactes.

¹Ce cas se présente habituellement lors d'une surreprésentation de certaines classes dans les jeux de données.

5 Apprentissage machine et classification binaire

5.1 Définition des critères de classification binaire

Le critère de notre classification binaire est un critère très simple, qui répond à l'éternelle question “*Ce champignon est-il comestible ou non ?*”

Malgré la simplicité apparente du critère binaire, la comestibilité ou la non-comestibilité d'un champignon se situe en réalité sur un continuum :

- Champignons particulièrement réputés pour leurs qualités gustatives,
- Champignons réputés comestibles,
- Champignons comestibles après cuisson,
- Champignons non-toxiques, mais à la comestibilité médiocre en raison de leur amertume (bolet de fiel, *Tylopilus Felleus*), de leur piquant (lactaire poivré, *Lactarius piperatus*)...
- Champignons dont la toxicité se manifeste à partir d'une certaine quantité consommée,
- Champignons toxiques, avec des syndromes de sévérité variée,
- Champignons mortels.

Si le classement des deux premières et des deux dernières catégories au sein d'un critère binaire est trivial, il n'en est pas de même pour les trois autres. L'acceptation ou le rejet des champignons relevant de ces catégories intermédiaires constituera par définition un critère arbitraire.

Nous pouvons toutefois esquisser des contours – forcément subjectifs – de notre critère de comestibilité en supposant que les champignons seront soumis à une cuisson, que tout champignon à la comestibilité médiocre sera à rejeter car susceptible de rendre la totalité du plat immangeable, et que tout champignon présentant une possible toxicité, même minime, sera à écarter par mesure de prudence : à titre d'exemple, le tricholome doré (*Tricholoma Auratum*) a pu être successivement qualifié de *comestible très réputé*, puis de champignon ayant provoqué des empoisonnements mortels par rhabdomolyse.^{2,25}

Nous choisirons donc de fixer les catégories binaires suivantes :

- À accepter : excellents comestibles, comestibles, comestibles cuits.
- À rejeter : comestibles médiocres, toxiques au delà d'une certaine quantité, toxiques, mortels.

5.2 Analyse exploratoire des données (EDA)

BROUILLON, à finir maintenant que j'ai mes données...

Le lot de données d'origine contient 98700 spécimens de champignons, caractérisés par 26 propriétés morphologiques ou environnementales. La structure de ce lot de données est résumée dans le tableau 3.

Ce lot de données original a été découpé en un jeu d'apprentissage/optimisation et un jeu de données d'évaluation, avec un rapport 18:1, conformément aux principes mentionnés dans nos développements précédents.^m

Toutes les distributions des variables du lot d'entraînement ont ensuite été tracées par histogrammes pour les variables numériques, et diagrammes en barres pour les variables alphabétiques et catégorielles.

Les diagrammes en barres n'ont rien illustré de particulièrement remarquable et n'ont pas été inclus dans le rapport. Toutefois, les distributions dimensionnelles sont plus intéressantes : à première vue, elles semblent suivre une courbe en cloche (figure 20), avec une longue queue à droite. Une transformation logarithmique (figure 21) montre plus nettement la forme de cette queue.

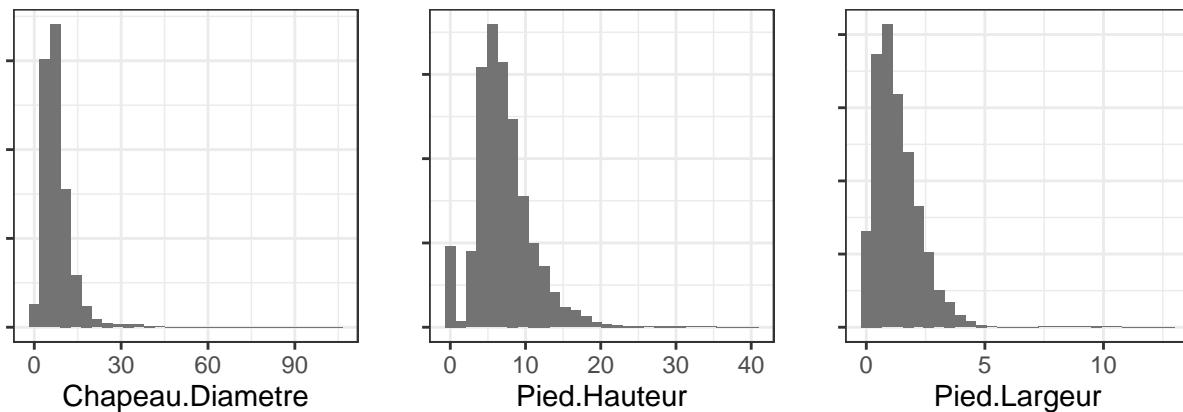


Figure 20: Distribution des diamètres de chapeau, longueur de stipe, diamètre de stipe

La distribution du diamètre du chapeau D_C a l'apparence d'une courbe en cloche, avec une longue queue à droite, mais est en réalité bimodale, avec un mode principal à 5 cm, et un mode secondaire beaucoup plus petit pour $D_C \approx 75$ cm. Cette taille exceptionnelle est attribuable à des variétés telles que *Polyporus squamosus*.^{2,25}

^mcf. section 4.2, page 35.

Table 3: Structure du lot de données initial

	Type	Niveaux
Type	factor	2
Hyménophore1	factor	3
Hyménophore2	factor	27
Chair.Type	factor	5
Chapeau.Forme	factor	35
Chapeau.Surface	factor	53
Chapeau.Couleur	factor	24
Chapeau.Marge	factor	42
Lames.Attache	factor	8
Lames.Espace	factor	14
Lames.Couleur	factor	25
Spore.Couleur	factor	11
Pied.Forme	factor	26
Pied.Surface	factor	49
Pied.Couleur	factor	30
Chair.Couleur	factor	30
Odeur	factor	76
VG.Type	factor	3
VG.Type2	factor	9
VP.Type	factor	8
VP.Type2	factor	42
Habitat	factor	65
Mois	integer	12
Chapeau.Diametre	numeric	3793
Pied.Hauteur	numeric	2511
Pied.Largeur	numeric	891

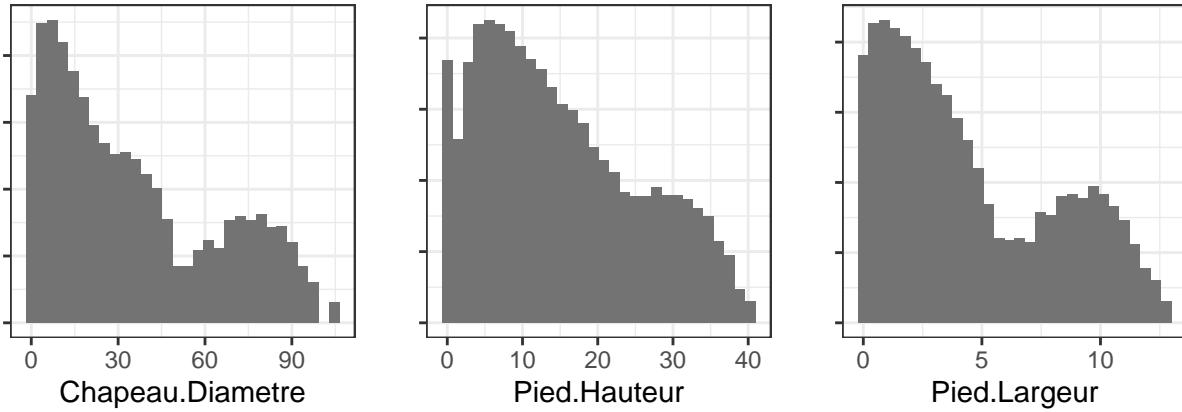


Figure 21: Distribution des diamètres de chapeau, longueur de stipe, diamètre de stipe (échelle logarithmique en ordonnée)

La distribution de la longueur de stipe L_S a également une forme de courbe en cloche avec une longue queue à droite, un mode principal à 5 cm et un mode secondaire à 0 cm. Cette valeur peut également sembler surprenante, mais certains champignons du lot n'ont pas de pied, ce qui explique cette valeur.

La distribution du diamètre de stipe D_S a aussi l'apparence d'une courbe en cloche avec une longue queue à droite, et un pic à $D_S \approx 10\text{-}15$ mm. Dans toutes ces distributions, la longue queue à droite peut probablement s'expliquer par l'utilisation, dans le *Secondary Mushroom Dataset*, d'une distribution normale pour chaque variété, associée à l'impossibilité d'avoir des valeurs dimensionnelles négatives.

L'étude des distributions dans un espace bidimensionnel (figure ??, p. ??) permet déjà de dégager des associations de caractéristiques d'intérêt, ici représentées sous formes de zones unicolores où un type de champignons (comestible ou non) dominera, ce qui permet déjà d'élaborer des stratégies de classification basiques.ⁿ

Ainsi, il est déjà possible de constater, d'après le nuage de points en haut à droite de la figure ?? que dans le lot de données, *tous* les champignons pour lesquels $\text{cap.diameter} \geq 40$ seront comestibles, et de même pour $\text{stem.height} \geq 20$. Toutefois, ainsi que l'illustre le graphique de densité situé dans le coin opposé, cette possible séparation ne concernera qu'une minorité de champignons, l'immense majorité des spécimens ayant des caractéristiques $\text{stem.height} \leq 10$ et $\text{cap.diameter} \leq 15$, où les courbes de densité se superposent, laissant entendre que cette association de critères ne sera que peu d'intérêt pour classer un grand nombre de champignons.

La distribution du diamètre de stipe D_S a aussi l'apparence d'une courbe en cloche avec une longue queue à droite, et un pic à $D_S \approx 10\text{-}15$ mm. Dans toutes ces distributions, la longue queue à droite peut probablement s'expliquer par l'utilisation, dans le *Secondary Mushroom*

ⁿCe type de classifieur sera développé ultérieurement, en section 5.3.2, page 59

Dataset, d'une distribution normale pour chaque variété, associée à l'impossibilité d'avoir des valeur dimensionnelles négatives.

De même, les graphiques de distribution des variables *cap.diameter* et *stem.height* en fonction des variables *veil.type* et *has.ring* montrent des distributions multimodales, qui mettront probablement en difficulté des modèles d'analyse discriminante linéaire, basées sur l'hypothèse de variables distribuées normalement – ou a minima dotées de distributions monomodales.[°]

Cette brève analyse exploratoire des données illustre donc l'intérêt de l'EDA, qui constitue un outil puissant permettant de mieux appréhender un lot de données, voire d'orienter le choix des stratégies d'apprentissage machine qui pourraient être exploitées à fins de classification.

[°]cf. section 4.4.1, page 40

5.3 Optimisation et sélection des modèles

Il existe une grande variété de modèles exploitables pour bâtir un système d'apprentissage machine. Cette section expliquera la stratégie utilisée pour l'évaluation de certains de ces modèles, ainsi que pour l'exploration de l'espace de leurs hyperparamètres à fins d'optimisation et la mesure de leurs performances.

Les modèles sélectionnés pour cette étude sont de types variés :^p

- Analyse naïve
- Analyse discriminante linéaire : *Linear Discriminant Analysis* (lda2), *Penalized Discriminant Analysis* (pda)
- Modèle arborescent : *Classification And Regression Tree* (rpart, rpartCost), C5.0 tree
- Forêt aléatoire : *Random Ferns* (rferns), *Random Forest* (ranger, Rborist)

5.3.1 Stratégie d'optimisation

Les algorithmes d'apprentissage machine développés au cours de cette étude mettent en œuvre les méthodes présentées dans les sections précédentes afin d'effectuer automatiquement les tâches suivantes :

1. Découpage du lot de données en un jeu d'entraînement/optimisation et en un jeu de validation, avec adaptation des rapports de taille en fonction du volume de données du lot initial,^q
2. Apprentissage sur le jeu d'entraînement, exploitant une validation croisée à k blocs, avec adaptation du nombre de blocs à la taille du lot de données,^r
3. Exploration de l'ensemble de l'espace expérimental des hyperparamètres du modèle, via la méthode des hypercubes latins quasi-orthogonaux,^s
4. Mesure des performances en exploitant une métrique adaptée,^t
5. Modélisation des performances en fonction des hyperparamètres, via un modèle quadratique avec interactions,^t
6. Sélection des hyperparamètres permettant d'optimiser les performances du modèle,
7. Mesure des performances de chaque modèle avec les hyperparamètres optimaux,
8. Sélection des modèles les plus performants pour prédiction et mesure finale des performances contre le lot d'évaluation.

^pLes termes entre parenthèses font référence aux noms de modèles qu'utilise la librairie caret.

^qcf. section 4.2, page 35

^rcf. section 4.3, page 36

^scf. section 4.5, page 48

^tcf. section 4.6, page 49

5.3.2 Modèle naïf

Le premier classifieur présenté dans cette étude est un classifieur naïf, dont l'algorithme est extrêmement simple :

1. Considérer par défaut tous les champignons comme toxiques,
2. Tester chaque combinaison de n variables, à la recherche des valeurs qualitatives et quantitatives pour lesquelles tous les champignons sont comestibles.

Trois classificateurs sont ainsi proposés : un classificateur “stupide”, ne tenant compte d'aucun critère ($n = 0$) et considérant tous les champignons comme toxiques, un classificateur monovariable ($n = 1$), et un classificateur exploitant des combinaisons de 2 variables ($n = 2$).

Les performances respectives de ces différents modèles sont synthétisées dans le tableau 4.

Table 4: Performances de différents classificateurs naïfs

	n	Sens	Spec	Jw	Kappa	Temps (s)
Stupide	0	1.000	0.000	0.818	0.000	0.0
MonoCritere	1	1.000	0.399	0.891	0.504	19.7
BiCritere	2	0.995	0.989	0.989	0.982	15466.7

Les performances de ces modèles paraissent relativement bonnes, mais sont à relativiser en raison de la définition de notre indice de Youden pondéré qui accorde une très grande importance à la sensibilité. Ainsi, notre classificateur dit “stupide” – excessivement pusillanime et d'un intérêt pratiquement nul car rejetant tous les champignons – semble montrer une performance honorable d'après ce critère, avec $J_w = 0.818$, alors que, par construction, $\kappa = 0$.

Les performances du modèle monovariable ne sont guère plus élevées, ce qui démontre l'inefficacité des adages populaires prétendant garantir la comestibilité de tous les champignons d'une couleur donnée.^u

En revanche, les performances du modèle bivariable sont honorables, avec une amélioration sensible de la spécificité ($Spec \approx 0.989$), donc de la capacité à réellement distinguer des champignons comestibles. Les performances de ce modèle ($J_w = 0.989$) ne répondent toutefois pas au critère de performance défini précédemment ($J_w \geq 0.999$)^v. Les performances calculatoires sont extrêmement médiocres ($t > 4$ h), et probablement à mettre en relation avec un code peu optimisé.

^uMême si le modèle venait à montrer le contraire, il convient de rappeler que ce lot de données se limite aux espèces les plus courantes, d'une fonge limitée dans l'espace et dans le temps.

^vcf. section 4.6, page 49

5.3.3 Modèles d'analyse discriminante

Les modèles d'analyse discriminante choisis pour cette étude sont Ida2 (*LDA* : *Linear Discriminant Analysis*) et pda (*Penalized Discriminant Analysis*). Le modèle Ida2 dispose d'un seul hyperparamètre (*dimen*, nombre de fonctions discriminantes). Le modèle pda possède également un unique hyperparamètre (*lambda*, pénalité de réduction des coefficients).

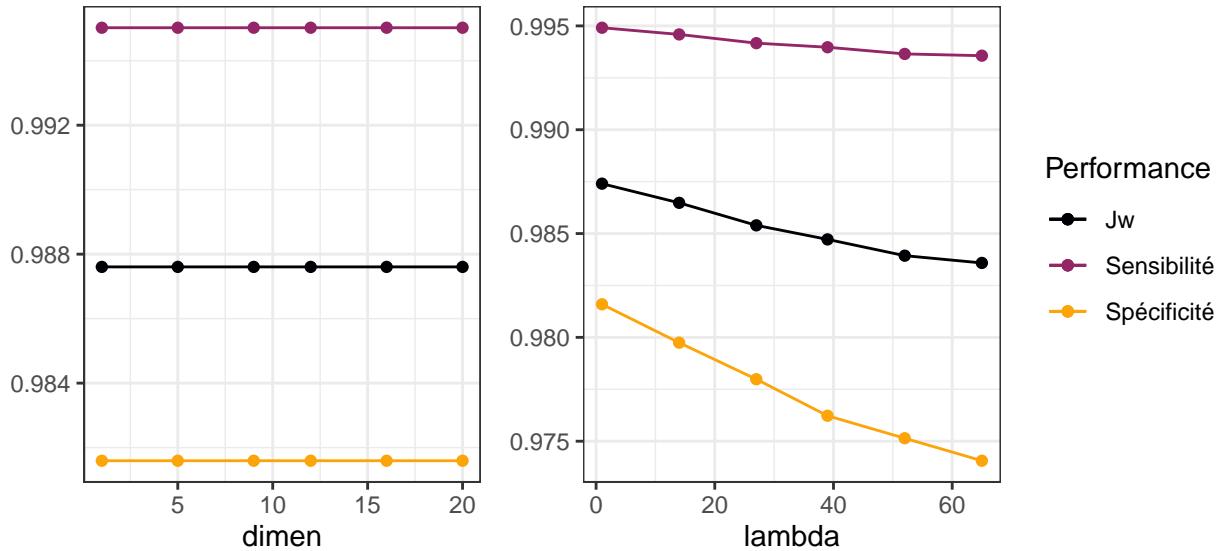


Figure 22: Performances des modèles Ida2 (à gauche) et pda (à droite), dans une tâche de classification binaire

Comme l'illustre la figure 22, les performances des modèles PDA et LDA sont très proches, et relativement constantes sur la totalité de l'espace expérimental de leurs hyperparamètres.

Le paramètre *dimen* du modèle Ida2 n'a absolument aucun effet sur les performances prédictives, avec un index J de Youden pondéré constant ($J_w = 0.988$). Ce caractère constant était en réalité prévisible, car le nombre de fonctions discriminantes n_{fd} dépend du nombre de prédicteurs p et du nombre de degrés de libertés entre les n_g groupes ($ddl = n_g - 1$), de sorte que :

$$n_{fd} = \min(ddl, p) = \min((n_g - 1), p)$$

Dans une classification binaire, ($n_g = 2$) il n'y a qu'un degré de liberté, c'est-à-dire une seule séparation entre les groupes, donc une seule fonction discriminante à définir.

L'efficience calculatoire est correcte, avec $t_{moy} = 4.94$ min ($n = 6$ itérations).

De même, le paramètre *lambda* du modèle pda n'impacte ses performances que de manière très marginale, avec des lambdas faibles donnant une légère amélioration des résultats ($J_{w_{max}} = 0.987$). L'efficience calculatoire est nettement en retrait par rapport au modèle Ida2, avec $t_{moy} = 15.77$ min ($n = 6$ itérations).

Toutefois, les performances de ces deux modèles restent malheureusement insuffisantes pour notre étude, aussi bien sur le J de Youden pondéré qu'en sensibilité ou en spécificité :

$$\left\{ \begin{array}{l} J_w \approx 0.986 \\ Sen \approx 0.995 \\ Spe \approx 0.98 \end{array} \right.$$

Ces performances prédictives perfectibles, car légèrement en retrait par rapport à celles d'un modèle naïf bidimensionnel, s'expliquent par le fonctionnement même des modèles d'analyse discriminante qui, s'ils peuvent analyser des données qualitatives à fins de classifications, ne peuvent le faire que si une quantification sous-jacente est possible, par exemple :

- Données binaires ou booléennes,
- Données qualitatives ordinale.

L'inclusion de ces modèles d'analyse discriminante, qui présentent ici des performances correctes mais néanmoins insuffisantes, a un intérêt essentiellement didactique, permettant de souligner l'intérêt d'une connaissance élémentaire des fondamentaux mathématiques et algorithmiques des modèles d'apprentissage machine mis en œuvre, pour en connaître les limites ou évaluer les besoins de nettoyage préalable des données avant déploiement de l'apprentissage machine, afin d'éviter de confronter certains modèles face à des problèmes de classification pour lesquels ils n'ont pas été conçus.

5.3.4 Modèles d'arbres de décision

Les modèles basés sur des arbres de décision ont un intérêt tout particulier pour cette étude, pour deux raisons majeures :

- La logique en arbre de décision est habituellement usitée pour la classification manuelle des champignons,
- Les arbres de décision obtenus peuvent être tracés, et facilement interprétés par l'humain.

Les premiers modèles présentés dans le cadre de notre étude sont deux modèles de type CART (*Classification And Regression Tree*). Le modèle CART le plus simple proposé dans notre étude (rpart) ne dispose que d'un seul hyperparamètre : *cp* (complexité).

Le modèle CART le plus simple propose des performances excellentes sur une large plage d'hyperparamètres (figure 23, page 62), remplissant notre critère $J_w \geq 0.999$, avec les indicateurs de performance suivants :

$$\left\{ \begin{array}{l} J_{w_{max}} = 0.9994 \\ Spe_{Jw_{max}} = 0.9992 \\ Sen_{Jw_{max}} = 0.9997 \end{array} \right.$$

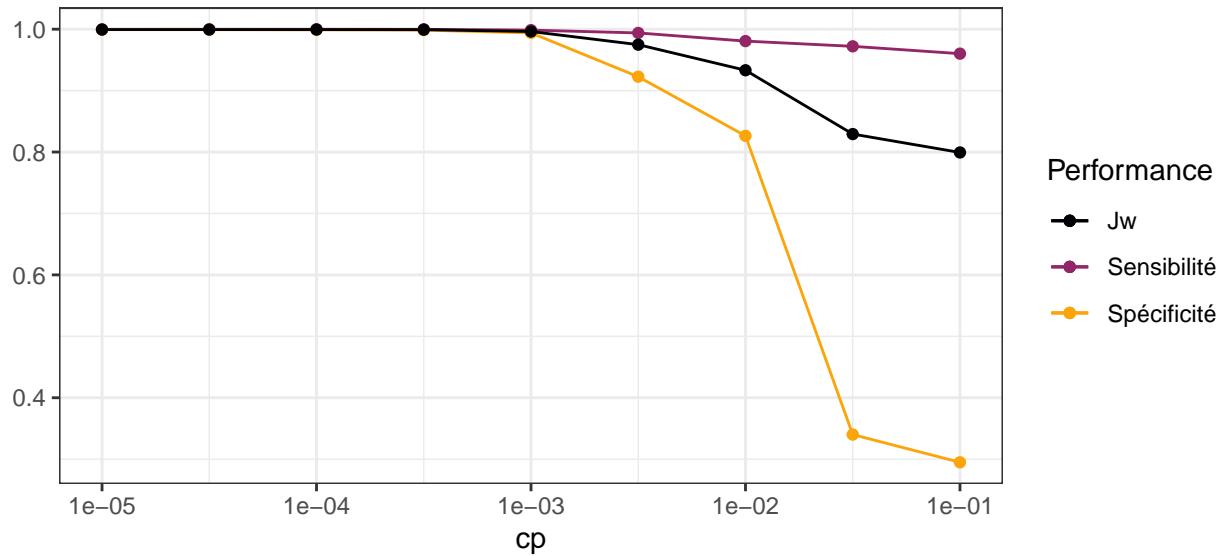


Figure 23: Performances du modèle rpart en classification binaire, en fonction du paramètre de complexité (cp)

L'efficience calculatoire de ce modèle est remarquable, avec $t_{moy} = 0.57$ min ($n = 9$ itérations).

Le second modèle CART utilisé dans cette étude (`rpartCost`) est plus élaboré et associe deux hyperparamètres :

- Complexité (cp), régissant la complexité et donc la taille de l'arbre,
- Coût ($Cost$), qui permet d'appliquer une pénalité variable selon le type d'erreur (ici, faux positif ou faux négatif).

Les graphiques de sensibilité et de spécificité en fonction des hyperparamètres (figure 24) illustrent bien, dans leur partie supérieure ($cp \geq 0.05$) – c'est à dire pour des arbres de complexité limitée – l'impact de cet hyperparamètre de coût, qui se traduit par la notion classique de compromis entre sensibilité et spécificité : dans cette zone, toute amélioration de la sensibilité se fera inévitablement au détriment de la spécificité, et réciproquement.

En pratique, pour $cp \geq 0.05$, notre modèle d'IA basé sur ce type d'arbre de décision se montrera soit excessivement sévère, rejetant un nombre considérable de champignons comestibles (quadrant supérieur gauche, $cost \leq 1.5$), soit au contraire excessivement laxiste, admettant un nombre important de champignons non-comestibles (quadrant supérieur droit, $cost \geq 1.5$).

C'est dans la section inférieure de ces graphiques ($cp \leq 0.025$), pour des arbres bien plus complexes, que le modèle montrera une performance acceptable tant en sensibilité qu'en spécificité.

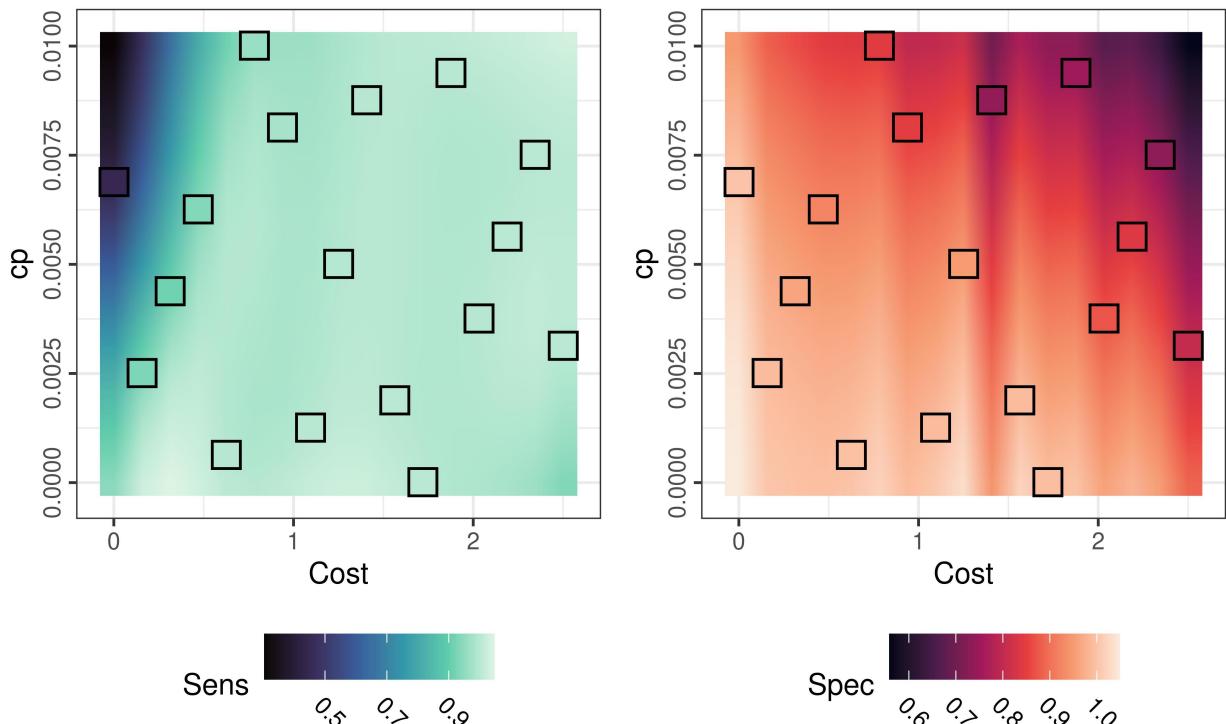


Figure 24: Sensibilité (à gauche) et spécificité (à droite) de `rpartCost` en classification binaire, en fonction de la complexité et du coût (interpolation Kriging quadratique, points expérimentaux encadrés en noir)

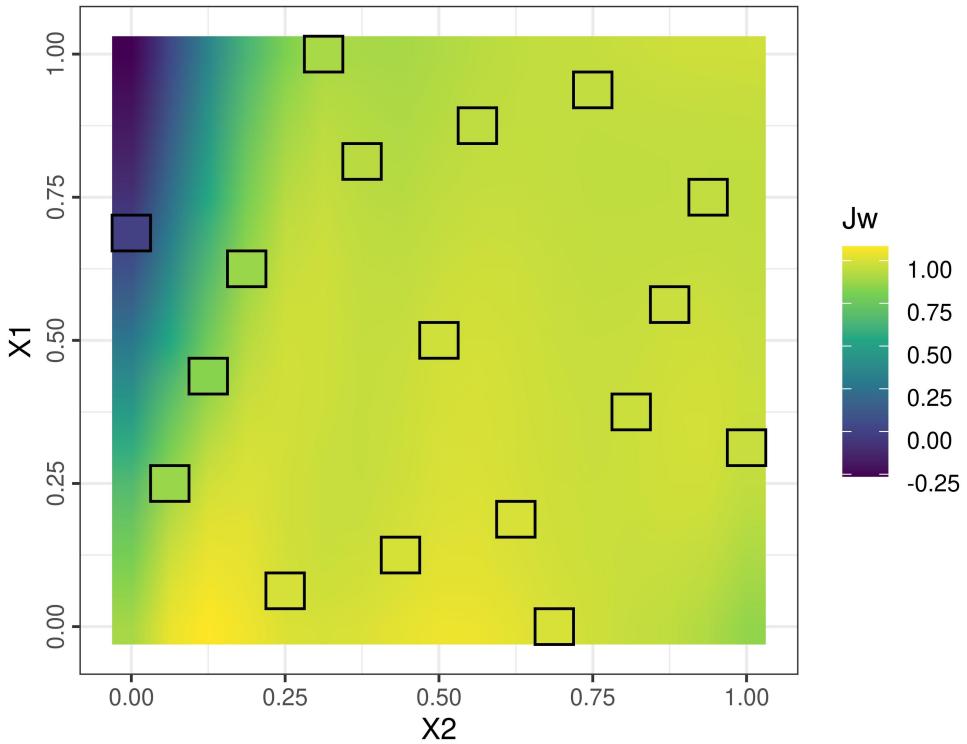


Figure 25: Performances (index J de Youden pondéré 10:1) de rpartCost en classification binaire, en fonction des hyperparamètres réduits de complexité X_1 et de coût X_2 (interpolation Kriging quadratique, points expérimentaux encadrés en noir)

En posant comme facteurs réduits :

- $X_1 \in [0; 1]$ pour le paramètre *minNode*,
- $X_2 \in [0; 1]$ pour le paramètre *predFixed*,

Nous pouvons modéliser la réponse Y (J_w) par un modèle quadratique avec interaction^w :

$$Y = b_0 + b_1.X_1 + b_2.X_2 + b_{12}.X_1.X_2 + b_{11}.X_1^2 + b_{22}.X_2^2$$

Avec Y l'indice J de Youden pondéré, X_1 le facteur réduit dans la plage $[0;1]$ associé à l'hyperparamètre de complexité (*cp*), X_2 le facteur réduit associé à l'hyperparamètre de coût (*Cost*) et b_n les coefficients des effets. La modélisation permet de calculer les effets suivants :

$$\begin{cases} b_0 = 0.7334 \\ b_1 = -0.8011 \\ b_2 = 1.7202 \end{cases} \quad \begin{cases} b_{12} = 1.097 \\ b_{11} = -0.0109 \\ b_{22} = -1.7102 \end{cases}$$

Les performances maximales seront ici atteintes pour :

$$\begin{cases} X_1 = 0 \quad \text{soit} \quad cp = 1e-05 \\ X_2 = 0.5 \quad \text{soit} \quad Cost = 1.25 \end{cases}$$

^wcf. section 4.5, page 48

Ces hyperparamètres optimaux permettent au modèle d'atteindre :

$$\begin{cases} J_{w_{max}} = 0.9998 \\ Spe_{J_{w_{max}}} = 0.9994 \\ Sen_{J_{w_{max}}} = 0.9999 \end{cases}$$

Les performances du modèle rpartCost sont excellentes, dépassant celles de notre modèle précédent, au prix d'une efficience calculatoire amoindrie ($t_{moy} = 4.01$ min, $n = 17$ itérations). Ce modèle remplit pleinement notre critère concernant l'index J de Youden pondéré minimal à atteindre. L'équilibre entre sensibilité et spécificité est à souligner, malgré la pondération 10:1 de ces critères dans notre indice synthétique.

Le dernier modèle d'arbre décisionnel proposé dans notre étude est C5.0 tree (c50tree). Ce modèle ne dispose d'aucun hyperparamètre.

Les performances obtenues sont :

$$\begin{cases} J_w = 0.9996 \\ Spe = 0.9982 \\ Sen = 1 \end{cases}$$

Bien que ne disposant d'aucun hyperparamètre, ce modèle a donné d'excellents résultats, remplissant aisément le critère $J_w \geq 0.999$, sans aucune optimisation nécessaire. En outre, ce modèle s'illustre par une sensibilité – c'est à dire une capacité à rejeter les champignons non comestibles – particulièrement élevée.

Le temps de calcul est relativement long ($t = 22.25$ min), signe d'une efficience calculatoire modeste, mais cette remarque est à modérer en raison des gains que peut apporter un modèle s'affranchissant totalement de la phase d'optimisation des hyperparamètres :

- Pas de séparation entre lots d'entraînement et d'optimisation,
- Pas de détermination du domaine expérimental et du plan d'expériences,
- Pas de balayage du domaine expérimental et de collecte des indicateurs de performance,
- Pas de modélisation des performances à la recherche des hyperparamètres optimaux.

Ces gains potentiels sont à considérer *lato sensu* : en temps de calcul évidemment, mais aussi de développement et d'optimisation des algorithmes correspondants à ces étapes.

Les modèles d'arbres de classification sont réputés pour être particulièrement adaptés aux problèmes de classification avec variables quantitatives et surtout qualitatives, et ont pu s'illustrer dans cette tâche de classification en fournissant des résultats excellents.

En effet, la totalité des modèles d'arbres testés ici dépassent les exigences imposées par le critère de performance que nous avions défini pour les classifieurs binaires de notre étude.

5.3.5 Forêts aléatoires

Le premier modèle de forêt aléatoire évalué dans notre étude est le modèle de fougères aléatoires rFerns (*Random Ferns*). Ce modèle ne possède qu'un seul hyperparamètre, la profondeur (*depth*).

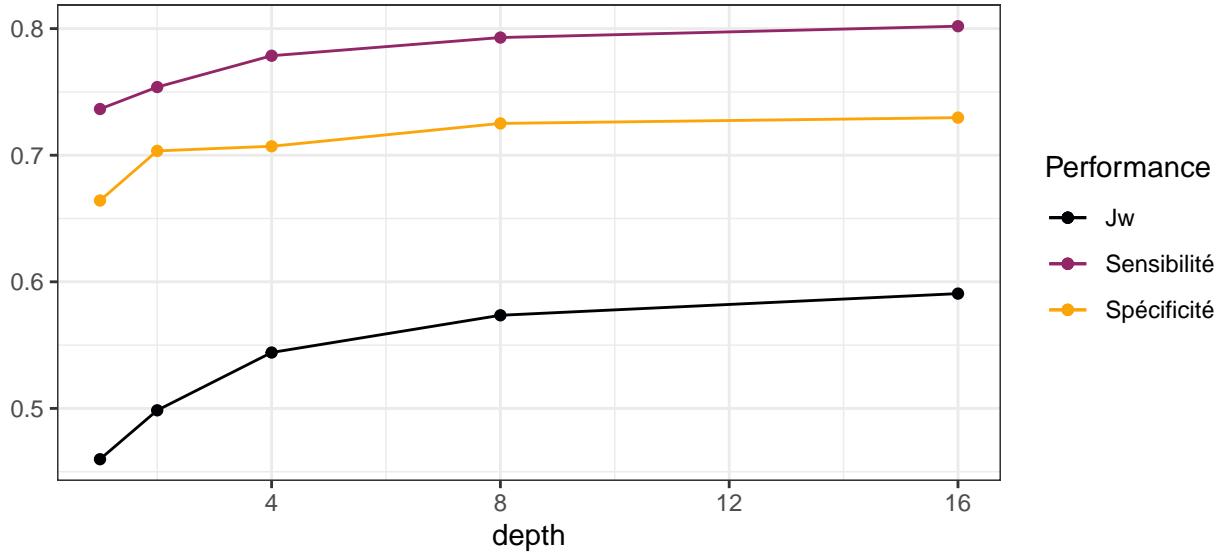


Figure 26: Performances du modèle de fougères aléatoires, en classification binaire

Le modèle de fougères aléatoires a fourni des résultats peu satisfaisants avec :

$$\begin{cases} J_{W_{max}} = 0.591 \\ Spe_{J_{W_{max}}} = 0.73 \\ Sen_{J_{W_{max}}} = 0.802 \end{cases}$$

Le second modèle de forêt aléatoire évalué dans cette étude est Rborist. Deux hyperparamètres régissent ce modèle :

- Le nombre de prédicteurs testés pour une scission (*predFixed*),
- Le nombre minimal de lignes-références distinctes avant de scinder un nœud (*minNode*).

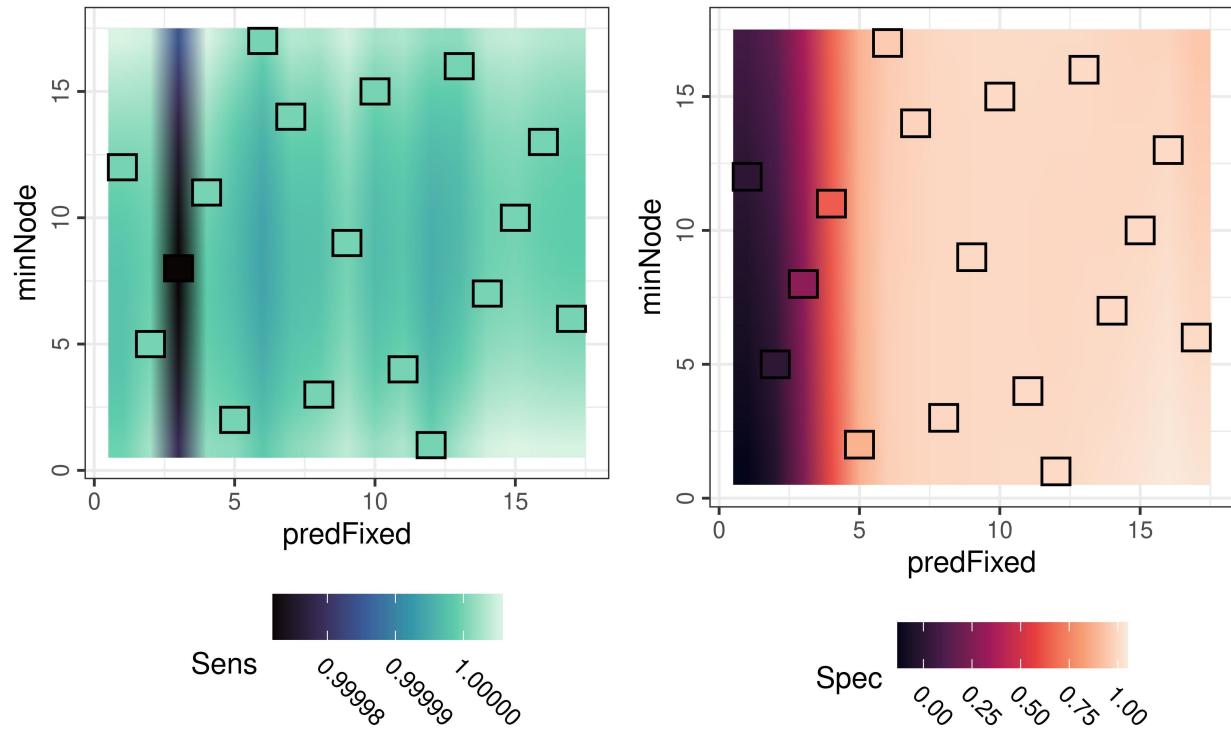


Figure 27: Sensibilité (à g.) et spécificité (à d.) du modèle Rborist en classification binaire, en fonction de ses hyperparamètres (interpolation Kriging quadratique, points expérimentaux en noir)

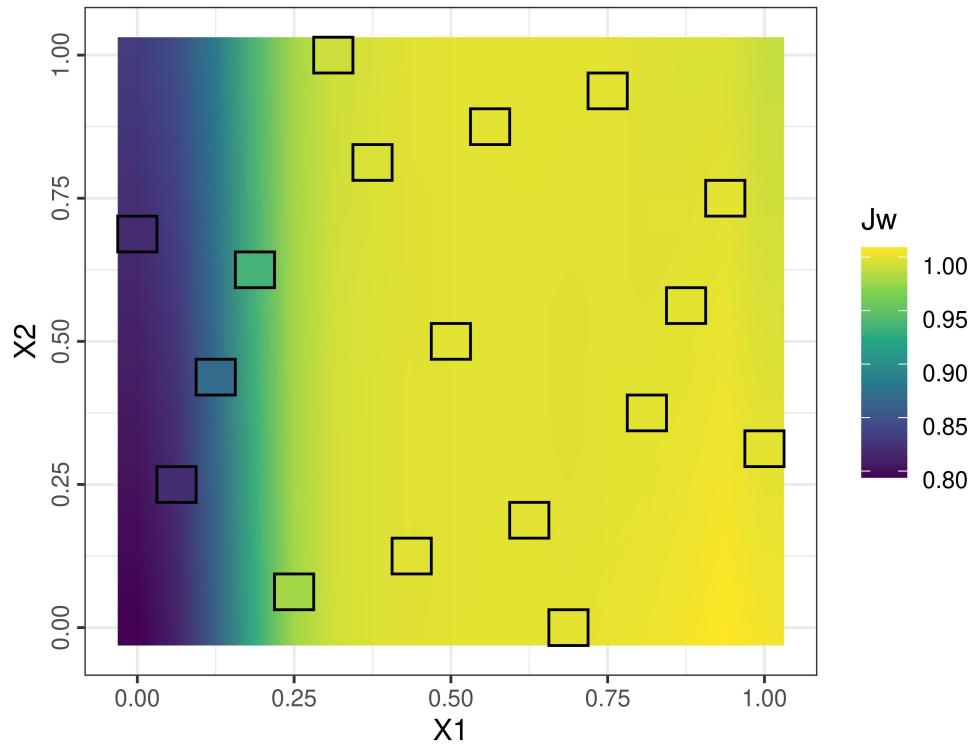


Figure 28: Performance (indice J de Youden pondéré) du modèle Rborist en classification binaire, en fonction de ses hyperparamètres réduits : nombre de prédicteurs X1 et de mbre de références avant scission X2 (interpolation Kriging quadratique, points expérimentaux encadrés en noir)

L'interprétation visuelle des graphiques des indicateurs de performance laisse entrevoir une très haute sensibilité – c'est à dire une excellente capacité à rejeter les champignons non comestibles – sur une très large plage d'hyperparamètres, et une spécificité beaucoup plus variable en fonction des hyperparamètres. L'hyperparamètre semblant avoir le plus d'effet est le nombre de prédicteurs testés pour une scission (*predFixed*). L'efficience calculatoire est très bonne, avec $t_{moy} = 7.44$ min ($n = 17$ itérations).

Nous pouvons modéliser la réponse par un modèle quadratique avec interaction^x :

$$Y = b_0 + b_1.X_1 + b_2.X_2 + b_{12}.X_1.X_2 + b_{11}.X_1^2 + b_{22}.X_2^2$$

Avec Y l'indice J de Youden pondéré J_w , X_1 le facteur réduit dans la plage [0;1] associé à l'hyperparamètre de nombre de prédicteurs testés par scission (*predFixed*), X_2 le facteur réduit associé à l'hyperparamètre de nombre minimal de références distinctes avant scission (*minNode*) et b_n les coefficients des effets.

Le calcul numérique nous permet d'obtenir les estimations des effets :

$$\begin{cases} b_0 = 0.8183 \\ b_1 = 0.5392 \\ b_2 = 0.0242 \end{cases} \quad \begin{cases} b_{12} = -0.0442 \\ b_{11} = -0.3588 \\ b_{22} = 0.0025 \end{cases}$$

Ce modèle quadratique avec interactions semble confirmer l'interprétation graphique, avec $b_1 \gg b_2$, et permet d'évaluer les hyperparamètres optimaux permettant de maximiser l'indice J de Youden pondéré (*minNode* = 1 et *predFixed* = 13) afin de lancer la prédiction sur un modèle optimisé.

Les performances obtenues par le modèle ainsi optimisé sont excellentes :

$$\begin{cases} J_{w_{max}} = 0.99981 \\ Spe_{Jw_{max}} = 0.99896 \\ Sen_{Jw_{max}} = 1 \end{cases}$$

Le dernier modèle de forêt aléatoire que nous évaluons dans cette étude est le modèle ranger. Celui-ci dispose de trois hyperparamètres :

- La taille minimale de nœud (*min.node.size*),
- Le nombre de caractéristiques à séparer à chaque nœud (*mtry*)
- La règle contrôlant cette séparation (*splitrule*).

Les mesures de performance sur l'espace expérimental des hyperparamètres sont illustrées en figures 29, 30 et 31, pages 70 et 71.

^xcf. section 4.5, page 48

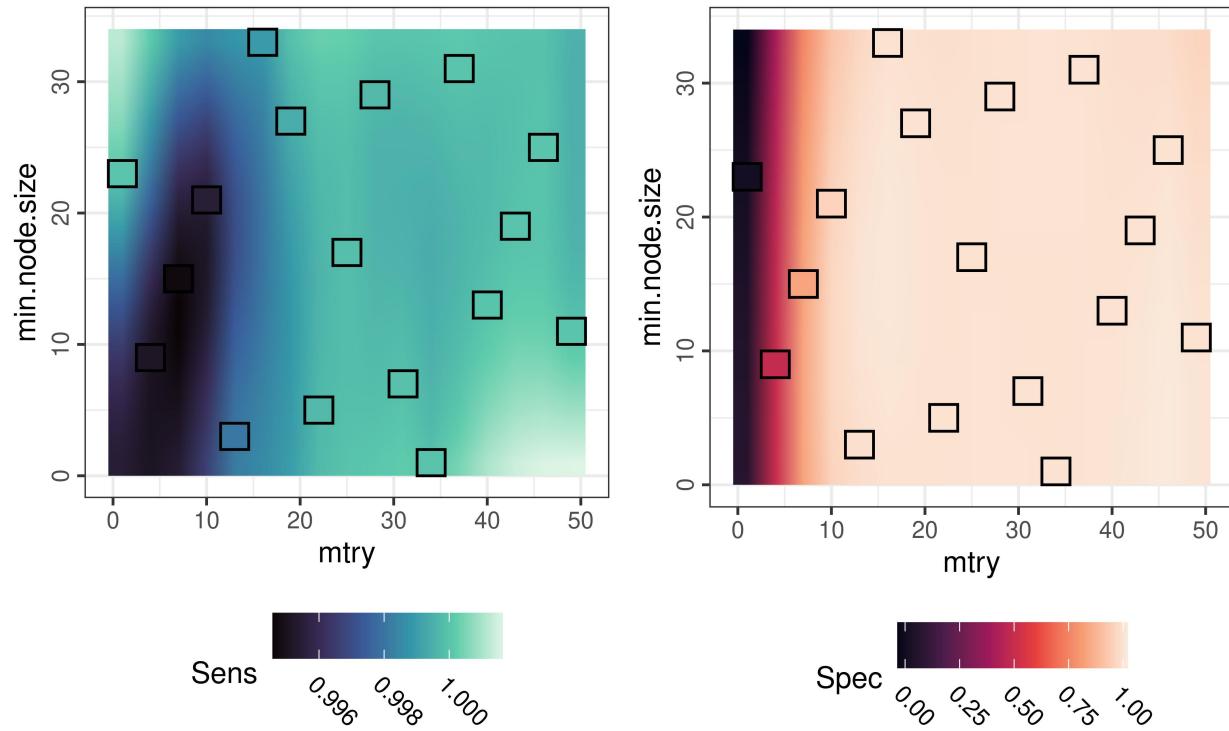


Figure 29: Sensibilité (à gauche) et spécificité (à droite) du modèle Ranger en classification binaire, en fonction des 2 hyperparamètres (algorithme de scission : Gini)

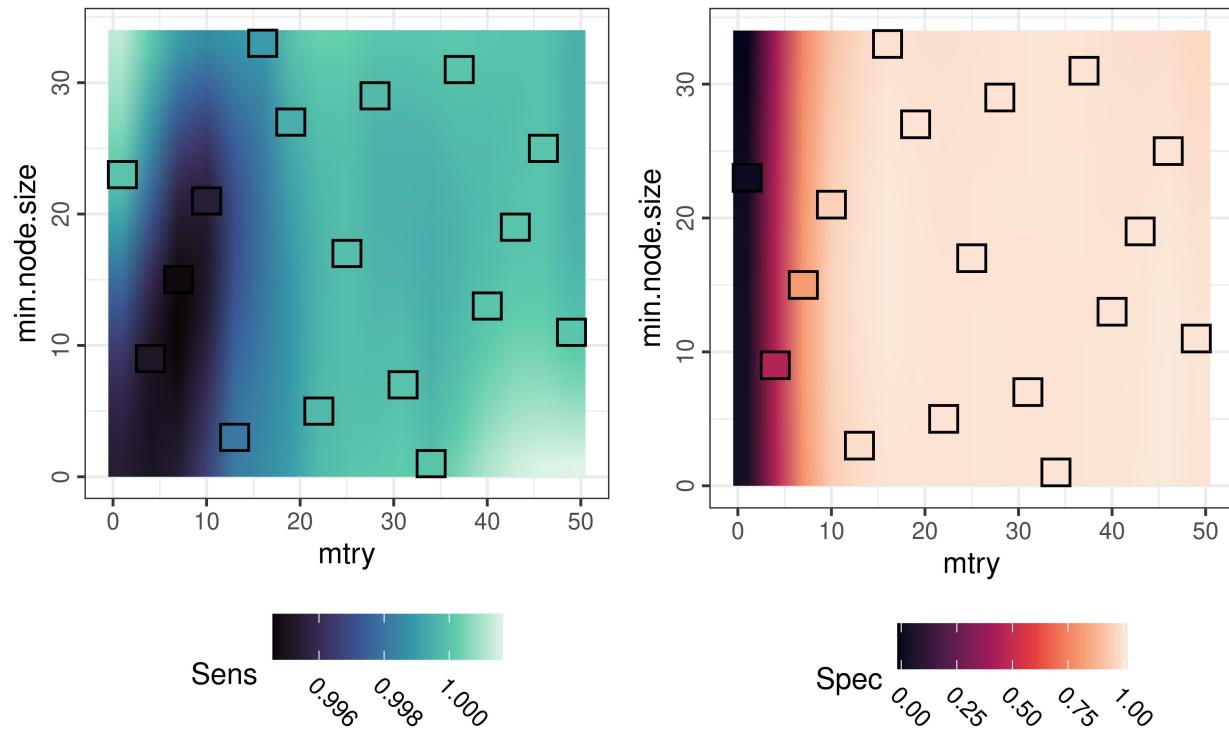


Figure 30: Sensibilité (à gauche) et spécificité (à droite) du modèle Ranger en classification binaire, en fonction des 2 hyperparamètres (algorithme de scission : extratrees)

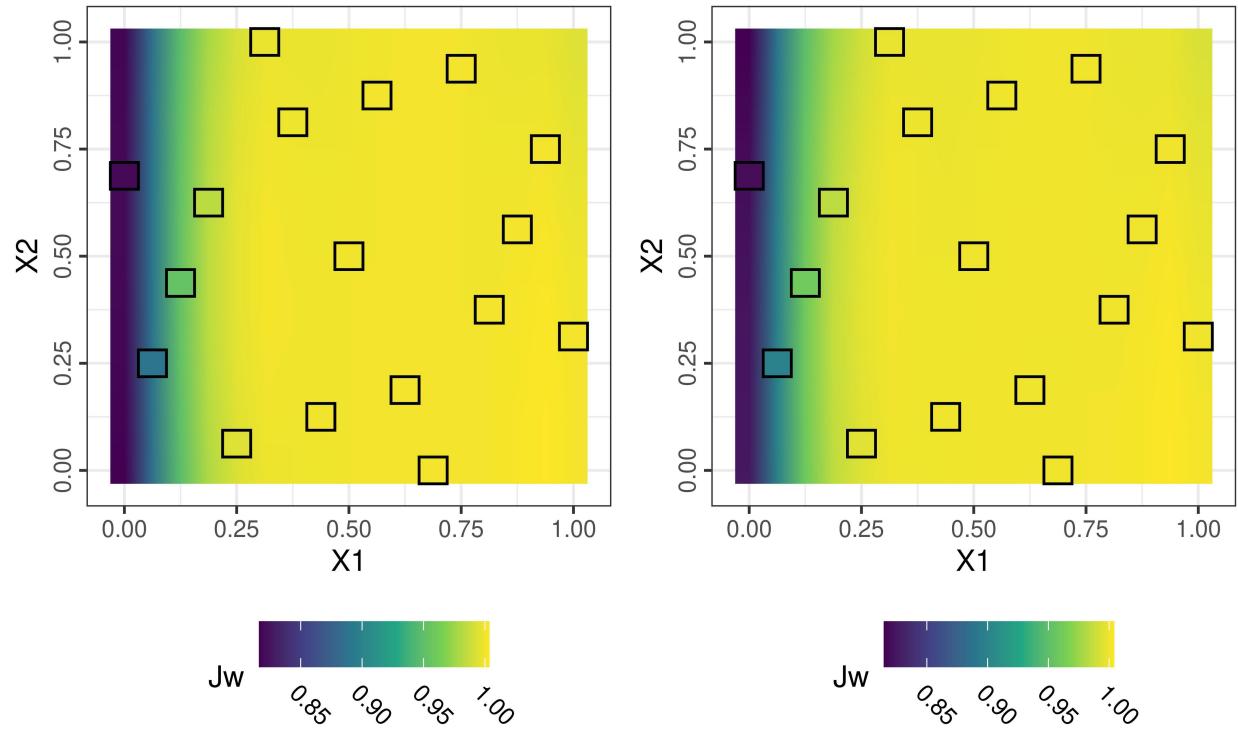


Figure 31: Performances du modèle Ranger en classification binaire, en fonction de l'algorithme de scission (extratrees à gauche, gini à droite) et des hyperparamètres réduits : caractéristiques à séparer X_1 et taille minimale de noeud X_2 (interpolation Kriging quadratique, points expérimentaux encadrés en noir)

L'interprétation graphique laisse entrevoir un comportement relativement similaire à celui du modèle Rborist, avec une sensibilité très élevée sur l'ensemble du domaine expérimental des hyperparamètres, et une spécificité dépendant largement du nombre de caractéristiques à séparer (*mtry*). L'efficience calculatoire est excellente, avec $t_{moy} = 1.12$ min ($n = 34$ itérations).

Nous pouvons proposer pour ce modèle la modélisation quadratique suivante :

$$Y = b_0 + b_1.X_1 + b_2.X_2 + b_3.X_3 + b_{12}.X_1.X_2 + b_{23}.X_2.X_3 + b_{13}.X_1.X_3 + b_{11}.X_1^2 + b_{22}.X_2^2$$

Avec Y l'indice J de Youden pondéré, X_1 le facteur réduit associé au paramètre de nombre de caractéristiques à séparer à chaque nœud (*mtry*), X_2 le facteur réduit associé au paramètre de taille minimale de noeud (*min.node.size*), X_3 le facteur régissant la règle de séparation (*splitlevel*, la valeur 0 étant attribuée à *gini*, 1 à *extratrees*) et b_n les estimations des coefficients des effets. Le facteur X_3 n'ayant que deux niveaux, il est évidemment impossible de lui attribuer une composante quadratique.

La modélisation permet de calculer les effets suivants :

$$\left\{ \begin{array}{l} b_0 = 0.8181 \\ b_1 = 0.6458 \\ b_2 = 0.0031 \\ b_3 = -0.0013 \end{array} \right. \quad \left\{ \begin{array}{l} b_{12} = -0.003 \\ b_{23} = 0.0033 \\ b_{13} = -7 \times 10^{-4} \\ b_{11} = -0.4726 \\ b_{22} = -0.0064 \end{array} \right.$$

La modélisation quadratique confirme les conclusions de l'interprétation graphique, avec $b_1 \gg b_2$. Une optimisation des hyperparamètres grâce à la modélisation quadratique (*min.node.size* = 4, *mtry* = 34 et *splitlevel* = *gini*) a donné d'excellents résultats :

$$\left\{ \begin{array}{l} J_{W_{max}} = 1 \\ Spe_{J_{W_{max}}} = 1 \\ Sen_{J_{W_{max}}} = 1 \end{array} \right.$$

Le modèle Ranger a donné des résultats assez similaires à ceux du modèle Rborist, avec une sensibilité, une spécificité et un indice de Youden excellents.

Ces résultats soulignent un fait intéressant : tous les modèles de forêts aléatoires ne sont pas égaux. Notre étude montre une différence considérable en sensibilité et spécificité entre les forêts aléatoires de type rFerns, Ranger et Rborist. Lors des étapes préliminaires de cette étude, d'autres algorithmes de forêts aléatoires ont également montré de grandes disparités d'efficience sur le plan calculatoire, ce qui nous a conduit à écarter certains modèles pour des raisons pratiques, alors que d'autres se sont avérés sensiblement plus rapides et ont donc pu être retenus pour notre étude.

5.4 Résultats

5.4.1 Protocole d'évaluation

Les modèles ayant atteint les performances requises ($J_w \geq 0.999$) lors de l'étape d'optimisation ont été choisis pour l'évaluation. Les deux modèles retenus sont deux modèles de type forêt aléatoire :

- Forêt aléatoire avec algorithme de type Ranger,
- Forêt aléatoire avec algorithme de type Rborist.

Tous les modèles ont été entraînés sur le jeu de données d'apprentissage, après application des hyperparamètres optimaux obtenus précédemment^y par modélisation des performances via un modèle quadratique avec interactions. Les performances de nos modèles face au jeu de données d'évaluation, auquel ils n'ont encore jamais été exposés^z, seront évaluées avec le même critère que précédemment : $J_w \geq 0.999$

5.4.2 Performances des modèles de forêts aléatoires

La matrice de confusion du modèle ranger (table 5) donne les résultats détaillés de ses prédictions.

Table 5: Matrice de confusion du modèle Ranger (prédictions à gauche, référence en haut)

	Rejeter	Conserver
Rejeter	3961	0
Conserver	0	1234

La forêt aléatoire de type Ranger a donné d'excellents résultats, sa précision finale étant égale à 1, avec un intervalle de confiance à 95% de [0.9993 ; 1], le tout en un temps raisonnable (28.92min), preuve de son efficience calculatoire.

La forêt aléatoire de type Rborist a donné des résultats assez similaires, atteignant une précision finale égale à 1, avec un intervalle de confiance à 95% de [0.9993 ; 1]. Le modèle Rborist s'est de plus avéré extrêmement efficient sur le plan calculatoire (11.15min).

Table 6: Performances des modèles Ranger et Rborist (jeu d'évaluation)

	Sensibilité	Spécificité	J de Youden	Durée (min)
Ranger	1	1	1	1.12
Rborist	1	1	1	7.44

^ycf. section 5.3.5

^zcf. section 4.3, p. 36

6 Apprentissage machine et classification multiclasse

Étant données les performances qu'ont montré les différents modèles lors de la classification binaire, seuls les modèles basés sur les arbres décisionnels et les forêts aléatoires seront évalués dans cette section.

6.1 Classification par familles

6.1.1 Modèles d'arbres de décision

Le modèle d'arbre de décision présenté dans cette partie est rpart, le plus simple des modèles CART.

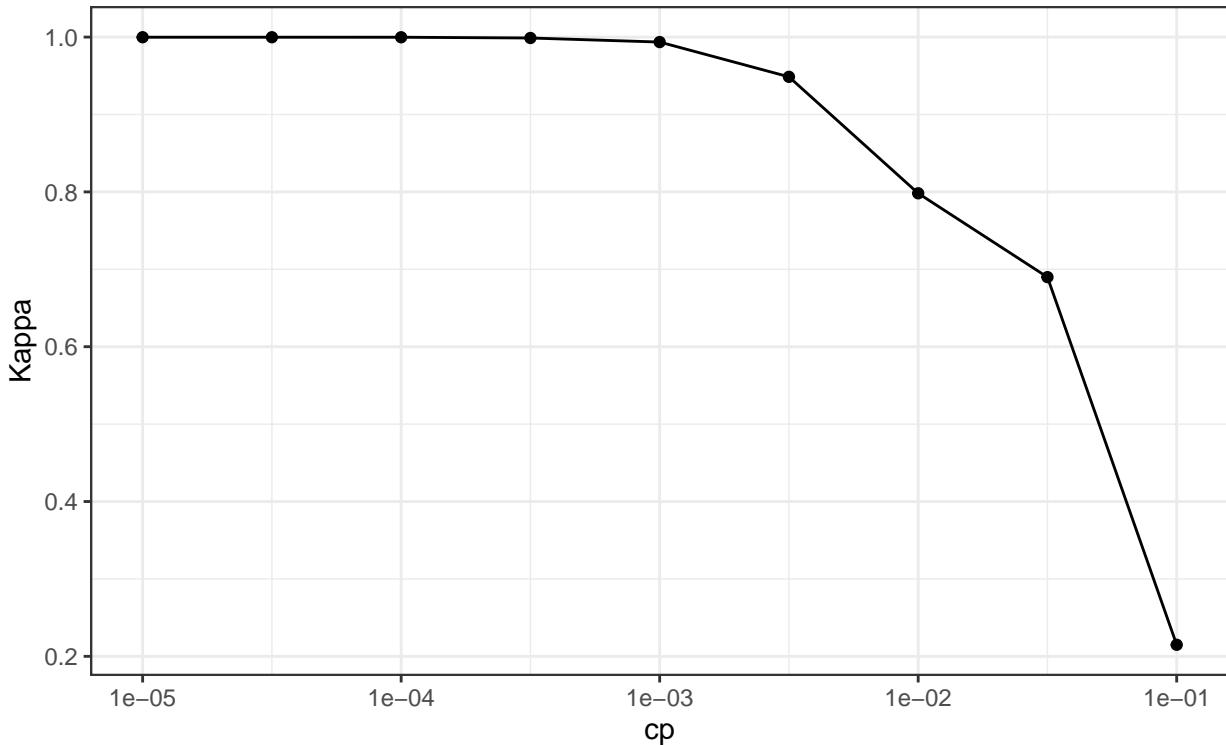


Figure 32: Performances du modèle CART (rpart) dans une classification par familles, en fonction du paramètre de complexité

Ce modèle, pourtant très simple, donne déjà de très bons résultats globaux, avec ($\kappa_{max} = 1$ et une précision $R_{max} = 1$).

6.1.2 Forêts aléatoires

Le premier modèle de forêt aléatoire évalué dans cette partie est ranger, que nous avons déjà présenté précédemment. Les graphiques des performances en fonction des hyperparamètres laissent entrevoir d'excellentes caractéristiques sur une large plage d'hyperparamètres.

Après optimisation des hyperparamètres (`min.node.size = 29`, `mtry = 28` et `splitrule = extratrees`), ce modèle a donné d'excellents résultats.

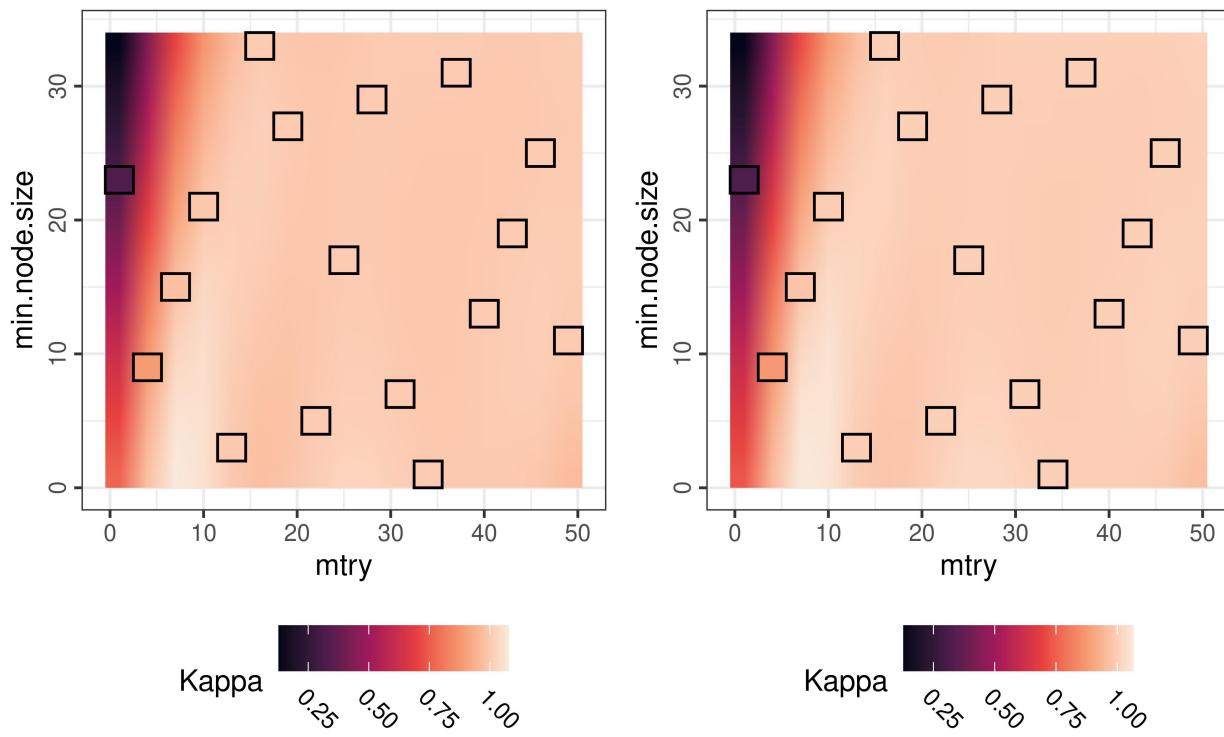


Figure 33: Performances du modèle Ranger dans une classification par familles, en fonction de ses 2 hyperparamètres (algorithme de scission : Gini à gauche, extratrees à droite)

Table 7: Performances du modèle Ranger (hyperparamètres optimaux)

mtry	min.node.size	splitrule	Accuracy	Kappa
28	29	extratrees	1	1

Le dernier modèle de forêt aléatoire est Rborist.

Avec des paramètres optimaux (*predFixed* = 5 et *minNode* = 2), la performance est estimée à :

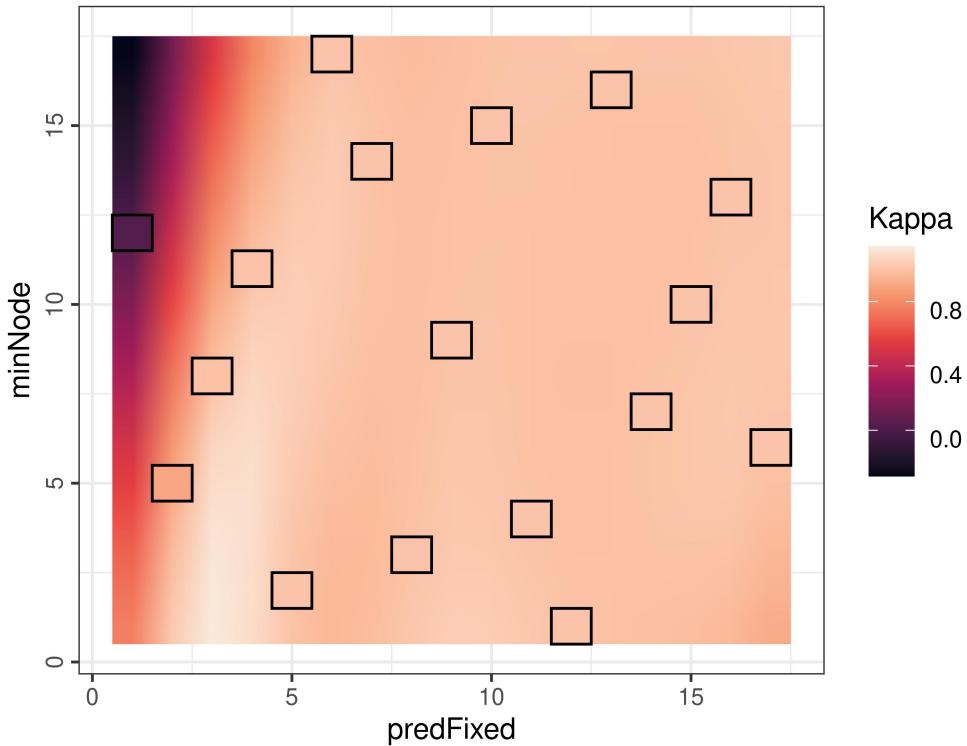


Figure 34: Performances du modèle Rborist dans une classification par familles, en fonction de ses deux hyperparamètres (interpolation Kriging quadratique, points expérimentaux encadrés en noir)

Table 8: Performances du modèle Rborist (hyperparamètres optimaux)

predFixed	minNode	Accuracy	Kappa
5	2	1	1

Le modèle Rborist a donné des résultats similaires à ceux du modèle Ranger, avec d'excellentes performances.

6.1.3 Résultats

Les critères et le protocole de l'évaluation sont les mêmes que ceux évoqués précédemment.

L'évaluation finale du modèle ranger donne la matrice de confusion de la table 9, page 78.

Table 9: Matrice de confusion du modèle Ranger (prédictions à gauche, références en haut)

	Tubariaceae	0	0	0
	Tricholomataceae	0	0	0
	Tricholomataceae	0	0	0
	Strophariaceae	0	0	0
	Russulaceae	0	0	0
	Psathyrellaceae	0	0	0
	Polyporales	0	0	0
	Pluteaceae	0	0	0
	Marasmiaceae	0	0	0
	Lepistiae	0	0	0
	Lepiotaceae	0	0	0
	Inocybaceae	0	0	0
	Hymenogastraceae	0	0	0
	Hygrophoraceae	0	0	0
	Hydnangiaceae	0	0	0
	Gomphidiaceae	0	0	0
	Hydnangiaceae	0	0	0
	Hygrophoraceae	0	0	0
	Hymenogastraceae	0	0	0
	Inocybaceae	0	0	0
	Lepiotaceae	0	0	0
	Lepistiae	0	0	0
	Marasmiaceae	0	0	0
	Pluteaceae	0	0	0
	Polyporales	0	0	0
	Psathyrellaceae	0	0	0
	Russulaceae	0	0	0
	Strophariaceae	0	0	0
	Tricholomataceae	0	0	0
	Tricholomataceae	0	0	0
	Tubariaceae	0	0	0
Agaricaceae	189	0	0	0
Amanitaceae	0	284	0	0
Bolbitiaceae	0	0	64	0
Boletaceae	0	0	0	364
Cantharellales	0	0	0	0
Cortinariaceae	0	0	0	0
Crepidotaceae	0	0	0	0
Entolomataceae	0	0	0	0
Gomphidiaceae	0	0	0	0
Hydnangiaceae	0	0	0	0
Hygrophoraceae	0	0	0	0
Hymenogastraceae	0	0	0	0
Inocybaceae	0	0	0	0
Lepiotaceae	0	0	0	0
Lepistiae	0	0	0	0
Marasmiaceae	0	0	0	0
Pluteaceae	0	0	0	0
Polyporales	0	0	0	0
Psathyrellaceae	0	0	0	0
Russulaceae	0	0	0	0
Strophariaceae	0	0	0	0
Tricholomataceae	0	0	0	0
Tricholomataceae	0	0	0	0
Tubariaceae	0	0	0	0

La précision finale est égale à $R = 0.9996$, avec un intervalle de confiance à 95% de [0.9986 ; 1]. La forêt aléatoire de type Ranger a donné d'excellents résultats, en un temps très raisonnable (2.62 min).

La forêt aléatoire de type Rborist a donné des résultats similaires, avec une précision finale égale à 1, avec un intervalle de confiance à 95% de [0.9993 ; 1]. Le modèle Rborist, donnant des résultats sensiblement identiques à Ranger, s'est avéré plutôt efficient sur le plan calculatoire (11.85 min).

Nous pouvons noter que le modèle ranger s'est ici avéré plus rapide que Rborist.

Table 10: Performances des modèles Ranger et Rborist (évaluation)

	Précision	Kappa	Durée (min)
Ranger	0.99962	0.99958	2.62
Rborist	1.00000	1.00000	11.85

6.2 Classification par espèce

Dans cette partie, la difficulté de la classification augmente sensiblement, les modèles ne sont plus chargés de classifier les champignons par familles, mais de déterminer précisément l'espèce de chaque spécimen du lot de données. Les modèles utilisés dans cette partie sont les mêmes que ceux de la classification par familles.

6.2.1 Modèles d'arbres de décision

Comme précédemment, le modèle d'arbre de décision retenu pour la classification par espèces est rpart.

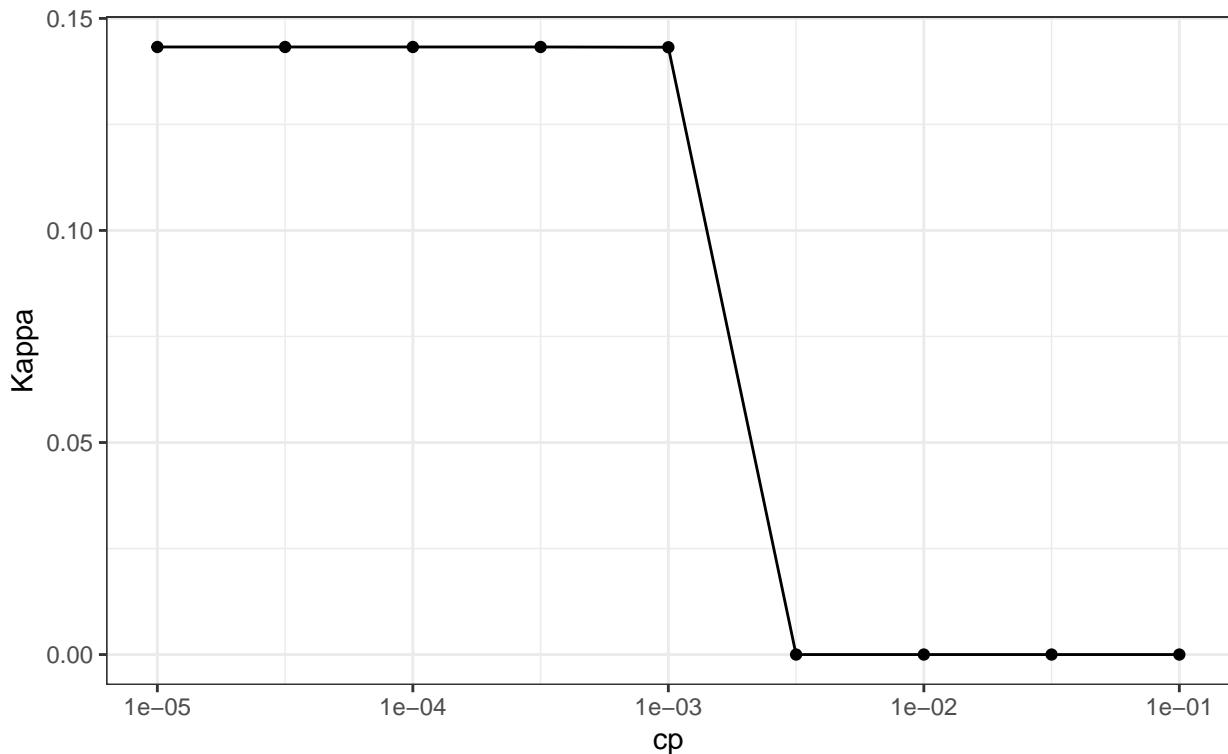


Figure 35: Performances du modèle CART (rpart) dans une classification par espèces, en fonction du paramètre de complexité

Ce modèle, bien que relativement simple, donne encore des résultats honorables, bien que le kappa ($\kappa_{max} = 0.143$) comme la une précision ($R_{max} = 0.146$) n'atteignent pas les objectifs de cette étude.

6.2.2 Forêts aléatoires

Comme lors de la classification par familles, notre premier modèle de forêt aléatoire évalué dans cette partie est ranger. L'exploration de l'espace expérimental des hyperparamètres de ce modèle laisse entrevoir de très bonnes performances sur une large plage d'hyperparamètres.

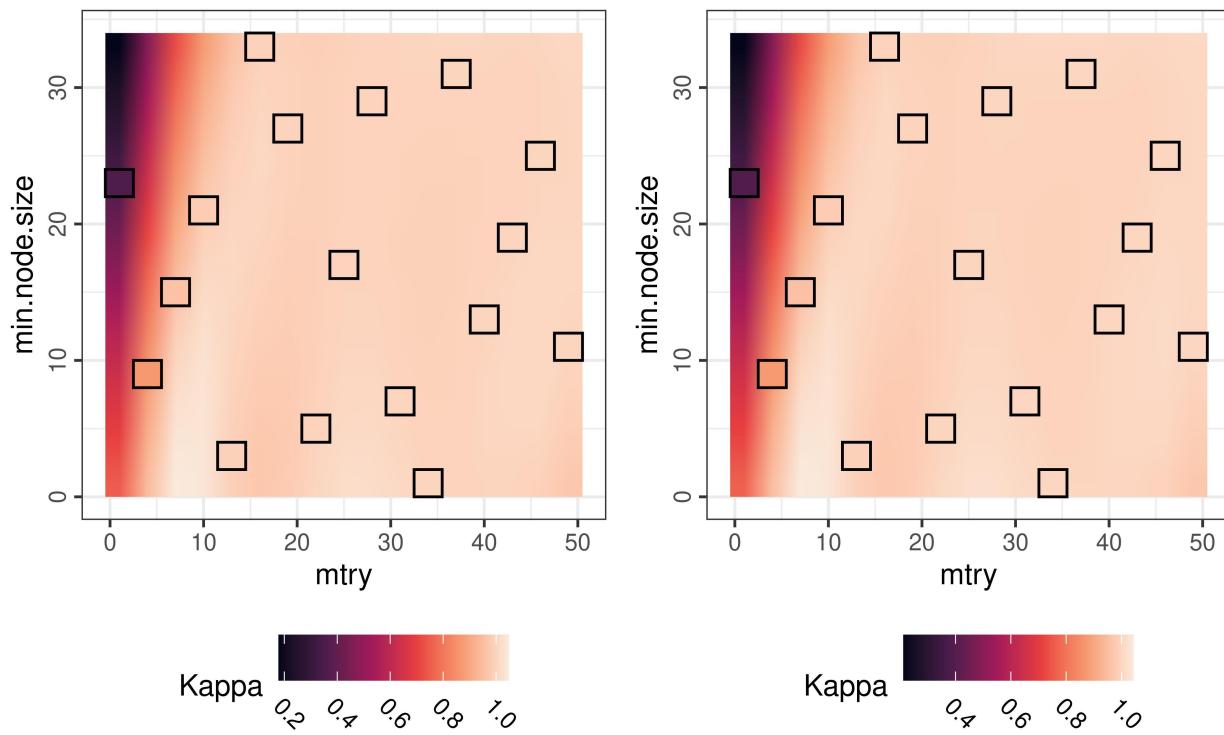


Figure 36: Performances du modèle Ranger dans une classification par espèces, en fonction de ses 2 hyperparamètres (algorithme de scission : Gini à gauche, extratrees à droite)

Après optimisation des hyperparamètres ($\text{min.node.size} = 11$, $\text{mtry} = 49$ et $\text{splitrule} = \text{gini}$), ce modèle a donné, comme lors des tests précédents, d'excellents résultats, malgré la complexité accrue du problème.

Table 11: Performances du modèle Ranger (hyperparamètres optimaux)

mtry	min.node.size	splitrule	Accuracy	Kappa
49	11	gini	0.99948	0.99947

Le dernier modèle de forêt aléatoire exploité dans cette tâche de classification par espèces est Rborist.

Avec des paramètres optimaux ($\text{predFixed} = 7$ et $\text{minNode} = 14$), la performance est estimée à :

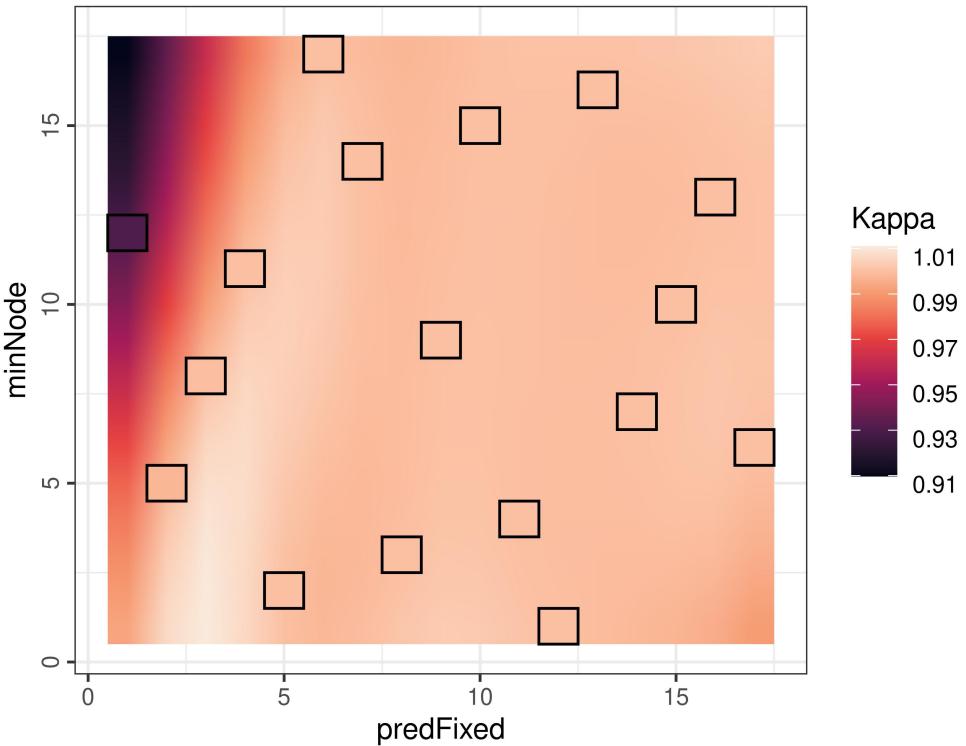


Figure 37: Performances du modèle Rborist dans une classification par espèces, en fonction de ses deux hyperparamètres (interpolation Kriging quadratique, points expérimentaux encadrés en noir)

Table 12: Performances du modèle Rborist (hyperparamètres optimaux)

predFixed	minNode	Accuracy	Kappa
7	14	0.99994	0.99994

Le modèle Rborist a donné des résultats similaires à ceux du modèle Ranger, avec d'excellentes performances en phase d'apprentissage et optimisation.

6.2.3 Résultats

Les critères et le protocole de l'évaluation sont les mêmes que ceux mentionnés pour les autres tâches de classification.

La précision finale du modèle Ranger est égale à $R = 0.994$, avec un intervalle de confiance à 95% de [0.9986 ; 0.9959]. La forêt aléatoire de type Ranger a donné d'excellents résultats, en un temps très raisonnable (2.62 min).

La matrice de confusion, ne reprenant ici que les espèces ayant posé problème à notre modèle (31:5195), est résumée par la table 13.

Table 13: Matrice de confusion des erreurs de Ranger, (prédictions à g., références en h.)

	<i>Xerocomus_pruinatus</i>	<i>Xerocomus_chrysenteron</i>	<i>Trametes_versicolor</i>	<i>Trametes_gibbosa</i>	<i>Piptoporus_betulinus</i>	<i>Lenzites_betulinus</i>	<i>Heterobasidion_annosum</i>	<i>Entoloma_rhodopolium</i>	<i>Entoloma_nidorosum</i>	<i>Daedaleopsis_tricolor</i>	<i>Cortinarius_salor</i>	<i>Cortinarius_delibutus</i>	<i>Amanita_phalloides_var._alba</i>	<i>Amanita_phalloides</i>
<i>Amanita_phalloides_var._alba</i>	1	16	0	0	0	0	0	0	0	0	0	0	0	0
<i>Cortinarius_delibutus</i>	0	0	16	13	0	0	0	0	0	0	0	0	0	0
<i>Entoloma_rhodopolium</i>	0	0	0	0	0	1	15	0	0	0	0	0	0	0
<i>Lenzites_betulinus</i>	0	0	0	0	10	0	0	0	0	16	0	0	0	0
<i>Piptoporus_betulinus</i>	0	0	0	0	0	0	0	2	0	16	0	0	0	0
<i>Trametes_versicolor</i>	0	0	0	0	0	0	0	0	0	0	0	1	16	0
<i>Xerocomus_chrysenteron</i>	0	0	0	0	0	0	0	0	0	0	0	0	0	15
<i>Xerocomus_pruinatus</i>	0	0	0	0	0	0	0	0	0	0	0	0	1	13

La forêt aléatoire de type Rborist a donné des résultats très proches de ceux obtenus par le modèle ranger, avec un taux d'erreur de 0:5195.

La précision finale de Rborist est ainsi égale à 1, avec un intervalle de confiance à 95% de [0.9993 ; 1]. Le modèle Rborist, donnant des résultats proches à Ranger, s'est de plus avéré assez efficient sur le plan calculatoire (16.67 min).

Nous pouvons noter que si Rborist affiche des performances marginalement supérieures, le modèle ranger s'est encore une fois avéré sensiblement plus rapide que Rborist sur des problèmes complexes.

Table 14: Performances des modèles Ranger et Rborist (évaluation)

	Précision	Kappa	Durée (min)
Ranger	0.99403	0.99401	4.14
Rborist	1.00000	1.00000	16.67

7 Conclusion et perspectives

A écrire... .

8 Références bibliographiques

1. Courtecuisse R, Moreau P-A, Welti S. Initiation à la reconnaissance des champignons du Nord de la France - Clé pour la détermination des espèces les plus fréquentes. Département des Sciences Végétales et Fongiques, Faculté de Pharmacie de Lille; 2020.
2. Courtecuisse R, Duhem B. Champignons de France et d'Europe. Delachaux et Niestlé; 2013. (Guides Delachaux).
3. Schlimmer J. Mushroom Data Set. University of California. 1987; Disponible sur: <https://archive.ics.uci.edu/ml/datasets/Mushroom>
4. Wagner D, Heider D, Hattab G. Mushroom data creation, curation, and simulation to support classification tasks. *Scientific Reports* [Internet]. avr 2021 [cité 10 déc 2022];11(1):8134. Disponible sur: <https://www.nature.com/articles/s41598-021-87602-3>
5. FungiFrance – Association pour le Développement d Outils Naturalistes et Informatiques pour la Fonge [Internet]. [cité 26 juin 2023]. Disponible sur: <https://fungi.fongifrance.fr/>
6. MycoDB : Base de données de champignons [Internet]. [cité 26 juin 2023]. Disponible sur: <https://www.mycodb.fr/>
7. Wickham H. tidyverse: Easily Install and Load the Tidyverse [Internet]. 2022. Disponible sur: <https://CRAN.R-project.org/package=tidyverse>
8. Mersmann O. microbenchmark: Accurate Timing Functions [Internet]. 2021. Disponible sur: <https://github.com/joshuaulrich/microbenchmark/>
9. Ripley B. MASS: Support Functions and Datasets for Venables and Ripley's MASS [Internet]. 2023. Disponible sur: <http://www.stats.ox.ac.uk/pub/MASS4/>
10. Kuhn M. caret: Classification and Regression Training [Internet]. 2022. Disponible sur: <https://github.com/topepo/caret/>
11. Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, et al. GGally: Extension to ggplot2 [Internet]. 2021. Disponible sur: <https://CRAN.R-project.org/package=GGally>
12. Vakayil A, Joseph R. twinning: Data Twinning [Internet]. 2022. Disponible sur: <https://CRAN.R-project.org/package=twinning>
13. Therneau T, Atkinson B. rpart: Recursive Partitioning and Regression Trees [Internet]. 2022. Disponible sur: <https://CRAN.R-project.org/package=rpart>
14. Milborrow S. rpart.plot: Plot rpart Models: An Enhanced Version of plot.rpart [Internet]. 2022. Disponible sur: <http://www.milbo.org/rpart-plot/index.html>
15. Hothorn T, Hornik K, Strobl C, Zeileis A. party: A Laboratory for Recursive Partitioning [Internet]. 2022. Disponible sur: <http://party.R-forge.R-project.org>
16. Wright MN, Wager S, Probst P. ranger: A Fast Implementation of Random Forests [Internet]. 2022. Disponible sur: <https://github.com/imbs-hl/ranger>
17. Kursa MB. rFerns: Random Ferns Classifier [Internet]. 2021. Disponible sur: <https://gitlab.com/mbq/rFerns>

18. Seligman M. Rborist: Extensible, Parallelizable Implementation of the Random Forest Algorithm [Internet]. 2022. Disponible sur: <https://CRAN.R-project.org/package=Rborist>
19. Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, et al. rmarkdown: Dynamic Documents for R [Internet]. 2023. Disponible sur: <https://CRAN.R-project.org/package=rmarkdown>
20. Xie Y. knitr: A General-Purpose Package for Dynamic Report Generation in R [Internet]. 2022. Disponible sur: <https://yihui.org/knitr/>
21. Kassambara A. ggpubr: ggplot2 Based Publication Ready Plots [Internet]. 2023. Disponible sur: <https://rpkgs.datanovia.com/ggpubr/>
22. Franco J, Dupuy D, Roustant O, Kiener P, Damblin G, Iooss. B. DiceDesign: Designs of Computer Experiments [Internet]. 2021. Disponible sur: <http://dice.emse.fr>
23. Dupuy D, Helbert C. DiceEval: Construction and Evaluation of Metamodels [Internet]. 2022. Disponible sur: <https://CRAN.R-project.org/package=DiceEval>
24. Xie Y. bookdown: Authoring Books and Technical Documents with R Markdown [Internet]. 2023. Disponible sur: <https://CRAN.R-project.org/package=bookdown>
25. Courtecuisse R. Photo-guide des Champignons d'Europe. Lausanne: Delachaux et Niestlé; 2000.
26. Ferguson BA, Dreisbach TA, Parks CG, Filip GM, Schmitt CL. Coarse-scale population structure of pathogenic *Armillaria* species in a mixed-conifer forest in the Blue Mountains of northeast Oregon. Canadian Journal of Forest Research [Internet]. avr 2003 [cité 17 août 2023];33(4):612-23. Disponible sur: <http://www.nrcresearchpress.com/doi/10.1139/x03-065>
27. Courtecuisse R. Clé de détermination macroscopique des champignons supérieurs des régions du Nord de la France. Société mycologique du Nord de la France; 1986.
28. Money N. Insights on the mechanics of hyphal growth. Fungal Biology Reviews [Internet]. 2008 [cité 11 févr 2023];22(2):71-6. Disponible sur: <https://www.sciencedirect.com/science/article/pii/S1749461308000195>
29. Porter DL, Naleway SE. Hyphal systems and their effect on the mechanical properties of fungal sporocarps. Acta Biomaterialia [Internet]. juin 2022 [cité 11 févr 2023];145:272-82. Disponible sur: <https://www.sciencedirect.com/science/article/pii/S1742706122002161>
30. Johnson NL. Continuous univariate distributions, volume 2. 2nd ed. New York [etc: John Wiley & sons; 1995. (Wiley series in probability et mathematical statistics Applied probability et statistics; vol. 2).
31. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2022. Disponible sur: <https://www.R-project.org/>
32. Journal Officiel de la République Française. Vocabulaire de l'intelligence artificielle [Internet]. 2018 [cité 18 août 2023]. Disponible sur: <https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000037783813>

33. Bironneau M, Coleman T. Machine Learning with Go Quick Start Guide : Hands-on techniques for building supervised and unsupervised machine learning workflows [Internet]. Birmingham: Packt Publishing; 2019. Disponible sur: <https://univ-scholarvox-com.ressources-electroniques.univ-lille.fr/book/88871068>
34. Amr T. Hands-On Machine Learning with scikit-learn and Scientific Python Toolkits: A practical guide to implementing supervised and unsupervised machine learning algorithms in Python. Birmingham: Packt Publishing; 2020.
35. Brownlee J. What is the Difference Between Test and Validation Datasets? [Internet]. MachineLearningMastery.com. 2017 [cité 14 févr 2023]. Disponible sur: <https://machinelearningmastery.com/difference-test-validation-datasets/>
36. Joseph VR. Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal [Internet]. 2022 [cité 15 févr 2023];15(4):531-8. Disponible sur: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11583>
37. Mak S, Joseph VR. Support points. The Annals of Statistics [Internet]. déc 2018 [cité 15 févr 2023];46(6A):2562-92. Disponible sur: <https://projecteuclid.org/journals/annals-of-statistics/volume-46/issue-6A/Support-points/10.1214/17-AOS1629.full>
38. Joseph VR, Vakayil A. SPLIT: An Optimal Method for Data Splitting. Technometrics [Internet]. avr 2022 [cité 15 févr 2023];64(2):166-76. Disponible sur: <https://doi.org/10.1080/00401706.2021.1921037>
39. Vakayil A, Joseph VR. Data Twinning. Statistical Analysis and Data Mining: The ASA Data Science Journal [Internet]. 2022 [cité 12 mars 2023];15(5):598-610. Disponible sur: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sam.11574>
40. Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics [Internet]. 1936 [cité 25 avr 2023];7(2):179-88. Disponible sur: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x>
41. Anderson TW. R. A. Fisher and Multivariate Analysis. Statistical Science [Internet]. 1996 [cité 25 avr 2023];11(1):20-34. Disponible sur: <https://www.jstor.org/stable/2246198>
42. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning - Data Mining, Inference, and Prediction. 2nd éd. Springer; 2016. (Springer Series in Statistics).
43. Wu X, Kumar V. The Top Ten Algorithms in Data Mining. First. CRC Press; 2009.
44. Rokach L, Maimon O. Data mining with decision trees: theory and applications. Second edition. Hackensack, New Jersey: World Scientific; 2015.
45. Loh W-Y. Fifty Years of Classification and Regression Trees: Fifty Years of Classification and Regression Trees. International Statistical Review [Internet]. déc 2014 [cité 20 mai 2023];82(3):329-48. Disponible sur: <https://onlinelibrary.wiley.com/doi/10.1111/insr.12016>
46. Kretowski M. Evolutionary Decision Trees in Large-Scale Data Mining [Internet]. Cham: Springer International Publishing; 2019 [cité 19 mai 2023]. (Studies in Big Data; vol. 59). Disponible sur: <http://link.springer.com/10.1007/978-3-030-21851-5>

47. Zhang H, Singer BH. Recursive Partitioning and Applications [Internet]. New York, NY: Springer; 2010 [cité 19 mai 2023]. (Springer Series in Statistics; vol. 0). Disponible sur: <https://link.springer.com/10.1007/978-1-4419-6824-1>
48. Zhang C, Ma Y, éditeurs. Ensemble Machine Learning: Methods and Applications [Internet]. New York, NY: Springer; 2012 [cité 19 mai 2023]. Disponible sur: <https://link.springer.com/10.1007/978-1-4419-9326-7>
49. Breiman L. Bagging predictors. *Machine Learning* [Internet]. août 1996 [cité 16 juin 2023];24(2):123-40. Disponible sur: <https://doi.org/10.1007/BF00058655>
50. Breiman L. Random Forests. *Machine Learning* [Internet]. oct 2001 [cité 16 juin 2023];45(1):5-32. Disponible sur: <https://doi.org/10.1023/A:1010933404324>
51. Genuer R, Poggi J-M. Random Forests with R [Internet]. Cham: Springer International Publishing; 2020 [cité 19 mai 2023]. (Use R!). Disponible sur: <http://link.springer.com/10.1007/978-3-030-56485-8>
52. Santiago J, Claeys-Bruno M, Sergent M. Construction of space-filling designs using WSP algorithm for high dimensional spaces. *Chemometrics and Intelligent Laboratory Systems* [Internet]. avr 2012 [cité 4 mars 2023];113:26-31. Disponible sur: <https://linkinghub.elsevier.com/retrieve/pii/S0169743911001195>
53. Youden WJ. Index for rating diagnostic tests. *Cancer* [Internet]. 1950 [cité 4 mars 2023];3(1):32-5. Disponible sur: <https://onlinelibrary.wiley.com/doi/abs/10.1002/1097-0142%281950%293%3A1%3C32%3A%3AAID-CNCR2820030106%3E3.0.CO%3B2-3>
54. Rücker G, Schumacher M. Summary ROC curve based on a weighted Youden index for selecting an optimal cutpoint in meta-analysis of diagnostic accuracy. *Statistics in Medicine* [Internet]. 2010 [cité 4 mars 2023];29(30):3069-78. Disponible sur: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3937>
55. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* [Internet]. avr 1960 [cité 12 mars 2023];20(1):37-46. Disponible sur: <https://doi.org/10.1177/001316446002000104>
56. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. *Biometrics* [Internet]. 1977 [cité 12 mars 2023];33(1):159-74. Disponible sur: <http://www.jstor.org/stable/2529310>

A Annexe 1 : Algorithme de génération de lot synthétique

Le principe de cet algorithme est de prendre un tableau au format csv contenant les caractéristiques typiques des macromycètes extraites de la littérature mycologique (type de sporophore, dimensions maximales du stipe, du chapeau, type de lames, couleur de sporée, etc.) et de générer un lot de données exploitable, sous forme d'une liste de spécimens, de chaque espèce données, possédant des caractéristiques individuelles censées être représentatives de leur espèce.

A.1 Initialisation

La seule bibliothèque utilisée lors de la création du lot de données synthétique est le *tidyverse*,⁷ collection de bibliothèques spécialisées dans le domaine de la *data science* et notamment dédiées au traitement, au nettoyage et à la visualisation de données.

```
library(tidyverse)
```

Notre algorithme charge ensuite le fichier csv contenant les caractéristiques typiques des champignons, qui est lu avec les paramètres de décimale et de séparateur utilisés lors de la création et de la sauvegarde du tableau csv (respectivement , et ;), puis attribué à un *dataframe* que nous nommerons `data_champis`. Les lignes commentées correspondent à l'utilisation d'un fichier identique hébergé à distance sur un dépôt GitHub.

```
#URL <- "https://github.com/EKRhani/champis/raw/master/donnees_champis.csv"
#download.file(URL, fichier_data)
fichier_data <- "donnees_champis.csv"

data_champis <- read.csv(fichier_data,
                         header = TRUE,
                         sep = ";",
                         dec = ",",
                         stringsAsFactors = FALSE)
```

A.2 Préparation des données

L'étape suivante consiste à extraire des données d'intérêt relatives à structure de notre objet `data_champis`, qui seront utiles pour les étapes ultérieures du générateur.

La première partie de cette étape s'applique à déterminer le caractère numérique (*i.e.* dimensionnel) ou textuel de chaque variable, à l'aide de la fonction `class`, appliquée sur la totalité de notre objet via une fonction `sapply`. Les données numériques sont ici définies comme des données entières (`integer`) ou numériques (`numeric`). Les données textuelles sont quant à elles définies comme toutes les données n'obéissant pas à ces deux critères.

La seconde partie de cette étape extrait le nombre d'espèces présentes dans notre lot de données et indice chacune d'entre elle dans une variable dédiée N.

```
structure <- sapply(X = data_champis, FUN = class, simplify = TRUE)
numeriques <- which(structure %in% c("integer", "numeric"))
textes <- names(structure[-numeriques])

n_especies <- nrow(data_champis)
data_champis$N <- 1:n_especies
```

Les données relatives à la saisonnalité des champignons ont été créées, dans notre tableau initial, sous la forme *mm-MM*, avec *mm* le mois de début, et *MM* le mois de fin de fructification. Afin de pouvoir générer chacun des mois de cet intervalle, nous pouvons créer un vecteur contenant chacune des valeurs de cet intervalle, à l'aide d'une fonction *ad hoc*.

Cette fonction extrait tout d'abord les deux chaînes numériques *mm* et *MM*, avant d'effectuer une comparaison entre ces deux valeurs. En effet, si *mm* > *MM*, alors la saison de fructification devra inclure le passage de décembre à janvier, et donc générer toutes les valeurs entières de *mm* à 12, puis celles de 1 à *MM*. L'ensemble de ces valeurs est intégrée dans une chaîne de caractères via la fonction `str_flatten\verb`, avec la séparation appropriée. Cette fonction est ensuite appliquée sur l'ensemble de la variable `Mois\verb`.

```
ConversionMois <- function(fcn_mois){
  date_extraction <- str_extract_all(fcn_mois, '[:digit:]+' )[[1]]
  ifelse(
    test = as.numeric(date_extraction[1]) < as.numeric(date_extraction[2]),
    yes = liste_mois <- seq.int(from = date_extraction[1], to = date_extraction[2]),
    no = liste_mois <- c(seq.int(from = date_extraction[1], to = 12),
                          seq.int(from = 1, to = date_extraction[2]))
  )
  str_flatten(liste_mois, collapse = ", ")
}

data_champis$Mois <- lapply(X = data_champis$Mois, FUN = ConversionMois)
```

Une autre particularité de notre lot de données est la prise en compte de caractéristiques rares, parfois rencontrées chez certains individus, mais néanmoins mentionnées dans la littérature mycologique comme étant exceptionnelles. Ces données ont été enregistrées entre parenthèses dans le tableau csv.

La stratégie retenue ici consiste à effectuer un tirage équiprobable, dans un vecteur contenant une occurrence unique de chaque valeur rare, et une répétition de chaque valeur commune. Le nombre de répétition de chaque valeur commune est ici défini par la variable `ratio_cr`.

La fonction créée afin de générer ce vecteur effectue tout d'abord un test visant à identifier les chaînes de caractères de type (xxx), synonymes de présence d'une caractéristique rare. En l'absence d'une caractéristique rare, les données d'entrées seront inchangées par l'algorithme. En la présence d'une telle caractéristique, chaque valeur (séparée par une virgule) est extraite individuellement par la fonction `strsplit`.

Un vecteur `n_repet`, contenant le nombre de répétitions de chaque valeur (ici, 10 pour une valeur commune, 1 pour une rare) est généré, par détection de l'absence de parenthèses autour des valeurs. Ce vecteur servira d'argument à la fonction `rep` pour générer la liste complète. Les parenthèses sont ensuite ôtées, et toutes les valeurs concaténées, avec une virgule en tant que séparateur.

La fonction ainsi créée sera appliquée à l'ensemble des variables précédemment identifiées comme textuelles, pour chaque ligne de l'objet `data_champis`, via la construction du texte constituant la commande, puis son exécution par la combinaison de fonctions `eval` et `parse`.

```
ratio_cr <- 10      # Ratio communs/rares

ConversionRares <- function(fcn_facteur){
  if(str_detect(fcn_facteur, pattern = "\\\\([[:alpha:]]|[:space:]\\)+\\\\)")){
    valeurs <- strsplit(fcn_facteur, split = ",")[[1]]

    n_repet <- valeurs %>%
      str_match(string = ., pattern = "\\\\([[:alpha:]]|[:space:]\\)+\\\\)") %>%
      is.na() %>%
      "*"*(ratio_cr-1)+1

    rep(valeurs, n_repet) %>%
      str_remove(., '\\\\(') %>%
      str_remove(., '\\\\)') %>%
      str_flatten(., collapse = ", ")
  }
  else{
    fcn_facteur
  }
}

for (n in 1:n_especes){
  ordre_rares <- paste0("data_champis[",n,",$", textes,
                        " <- ConversionRares(data_champis[",n,",$", textes, "])")
  eval(parse(text = ordre_rares))
}
```

A.3 Génération du lot de données

La génération du lot de données à proprement parler peut commencer. La stratégie globale est de générer un sous-lot de données par espèce, contenant un nombre défini de spécimens, puis de les regrouper dans le lot final, qui sera par la suite exporté sous forme d'un tableau csv.

La première étape de la génération des sous-lots consiste à créer une liste par espèce, nommée `champN`, avec `N` l'indice attribué précédemment à chaque espèce. La génération implique tout d'abord de créer une liste des noms de listes, ainsi qu'une fonction *ad hoc* permettant de séparer toutes les valeurs séparées par des virgules.

Pour chaque variété sera ensuite créée une liste, simple copie de la ligne correspondante effectuée à partir de notre *dataframe* `data_champis`. Chaque chaîne de caractères de la liste sera ensuite séparée à l'aide de la fonction définie précédemment, afin de définir des vecteurs. La sortie d'une fonction `str_split` étant par définition une chaîne de caractère, toutes valeurs dimensionnelles que nous avions précédemment identifiées comme numériques lors de l'analyse structurelle des données du lot seront quant à elles forcées en tant que numériques via la fonction `as.numeric`.

```
fonc_split <- function(fcn_split){str_split(string = fcn_split,
                                              pattern = ",\\s*",
                                              simplify = TRUE)}

champ_liste <- paste0("champ", data_champis$N)

for (n in 1:n_especes){
  assign(champ_liste[n], NULL)
  assign(champ_liste[n], as.list(data_champis[n,]))
  assign(champ_liste[n], map(.x = eval(parse(text = champ_liste[n])),
                           .f = fonc_split))

  ordre <- paste0(champ_liste[n], "[numeriques] <-",
                  "map(.x = ", champ_liste[n], "[numeriques],",
                  ".f = as.numeric)")

  eval(parse(text = ordre))
}
```

La préparation de la seconde et dernière étape de la génération des données consiste à définir la liste des lots de chaque espèce, appelés `lotN` (avec `N` l'indice attribué à chaque espèce), ainsi que quelques variables : nombre de champignons par espèce (`n_champis`), facteur de croissance appliqué à la loi bêta (`f_crois`), de même qu'une fonction permettant de générer la dispersion suivant la loi normale $\mathcal{N}(\mu = 1; \sigma = 0.05)$ et arrondissant la valeur finale à deux décimales.^{aa}

^{aa}cf. section 3.3.2, page 25 et suivantes.

```

lots_liste <- paste0("lot", data_champis$N)

n_champis <- 3e2      # Nombre de champignons pour chaque espèce
f_crois <- 2          # Facteur de croissance
tailles <- names(structure[numériques])

func_alea <- function(x){round(x * rnorm(n = n_champis,
                                             mean = 1,
                                             sd = .05),
                               digits = 2)}

```

La génération des sous-lots se décompose fondamentalement en trois étapes :

1. La construction d'un objet vide pour chaque lot,
2. La construction des chaînes de caractères de chaque commande,
3. L'exécution de ces commandes par l'enchaînement de fonctions `parse` et `eval`

Les chaînes de caractères correspondent aux commandes suivantes :

- `ordre_texte` insère, pour toutes les valeurs textuelles, une valeur piochée au hasard dans celles de la liste,
- `ordre_fac` définit, en suivant une loi bêta, le facteur de taille,
- `ordre_num1` multiplie ce facteur de taille par la valeur maximale de chaque caractéristique dimensionnelle,
- `ordre_num2` y applique la fonction de dispersion et d'arrondi définie précédemment,
- `ordre_df` crée l'objet de type `dataframe` regroupant toutes ces données,
- `ordre_suppr` et `ordre_rm` nettoient l'environnement en supprimant les données inutiles.

```

for (n in 1:n_especes){
  assign(lots_liste[n], NULL)

  ordre_texte <- paste0(lots_liste[n], "$", textes,
                        "<- sample(x = ", champ_liste[n], "$", textes,
                                   ", size = n_champis, replace = TRUE)")
  ordre_fac <- paste0(lots_liste[n], "$FacteurTaille <- rbeta(n = n_champis,
                                                               shape1 = 6*f_crois,
                                                               shape2 = 4,
                                                               ncp = .5*f_crois)")

  ordre_num1 <- paste0(lots_liste[n], "[tailles] <- lapply(", 
                       champ_liste[n], "[tailles], '*', ",
                       lots_liste[n], "$FacteurTaille)")

  ordre_num2 <- paste0(lots_liste[n], "[tailles] <- map(.x = ",
                       lots_liste[n], "[tailles], .f = func_alea)")

```

```

ordre_df <- paste0("lot",n, " <- data.frame(lot", n, ")")
ordre_suppr <- paste0("lot",n, "$FacteurTaille <- NULL ")
ordre_rm <- paste0("rm(champ",n, ")")

eval(parse(text = ordre_texte))
eval(parse(text = ordre_fac))
eval(parse(text = ordre_num1))
eval(parse(text = ordre_num2))
eval(parse(text = ordre_df))
eval(parse(text = ordre_suppr))
eval(parse(text = ordre_rm))
}

}

```

L'ensemble des sous-lots est ensuite concaténé. Cette concaténation a pour effet de joindre les sous-lots, mais place également la totalité des spécimens de chaque espèce ensemble. Nous effectuons donc un mélange aléatoire des spécimens par échantillonnage de leurs indices.^{ab}

```

lot_la_totale <- do.call(rbind, mget(paste0("lot",1:n_especes)))
lot_final <- lot_la_totale[sample(1:nrow(lot_la_totale)), ]

```

La toute dernière étape dans la préparation du lot est de détecter toutes les valeurs "concolores" ou "subconcolores" caractérisant la couleur du stipe, de la chair ou des lames de certains spécimens et d'en extraire les indices respectifs. La liste de ces indices permettra d'aligner les valeurs concernées sur la valeur de couleur du chapeau.

```

Concol_Pied <- which(lot_final$Pied.Couleur %in% c("concolore", "subconcolore"))
lot_final$Pied.Couleur[Concol_Pied] <- lot_final$Chapeau.Couleur[Concol_Pied]

Concol_Chair <- which(lot_final$Chair.Couleur %in% c("concolore", "subconcolore"))
lot_final$Chair.Couleur[Concol_Chair] <- lot_final$Chapeau.Couleur[Concol_Chair]

Concol_Lames <- which(lot_final$Lames.Couleur %in% c("concolore", "subconcolore"))
lot_final$Lames.Couleur[Concol_Lames] <- lot_final$Chapeau.Couleur[Concol_Lames]

```

Le lot résultant est finalement exporté sous format csv, compressé dans un fichier zip, et prêt à être exploité dans les autres parties de cette étude.

```

write_csv(x = lot_final, file = "lot_champis.csv")
zip(zipfile = "lot_champis.zip", files = "lot_champis.csv")

```

^{ab}Cet échantillonnage s'effectuant *sans* remplacement, il s'agit donc d'un simple mélange, par opposition à un *bootstrap*, qui s'effectuerait *avec* remplacement.

B Annexe 2 : Analyse exploratoire des données (EDA)

Code EDA... A finir.

C Annexe 3 : Algorithmes d'apprentissage machine

Nous détaillerons ici principalement les algorithmes utilisés pour le classifieur binaire. Les particularités d'intérêt des classifieurs multiples seront évoquées brièvement lors des développements de cette section.

C.1 Initialisation

Les bibliothèques utilisées lors des étapes d'apprentissage machine sont :

- tidyverse,⁷ collection de bibliothèques spécialisées dans le domaine de la *data science*,
- DiceDesign,²² bibliothèque spécialisée dans la création de plans d'expériences hypercubiques,
- DiceEval,²³ bibliothèque spécialisée dans la modélisation des résultats de plans d'expériences hypercubiques,
- caret,¹⁰ collection d'outils dédiés à l'apprentissage machine.
- twinning,¹² outils dédiés à la génération de jeux de données d'entraînement, optimisation, validation équilibrés.

```
library(tidyverse)
library(DiceDesign)
library(DiceEval)
library(caret)
library(twinning)
```

Le chargement des données s'effectue de la même façon que lors des sections précédentes. L'argument stringsAsFactors = TRUE revêt une importance particulière, car la classe factor est essentielle au bon fonctionnement des classifieurs. Dans le cadre d'une classification binaire, nous définissons arbitrairement, à l'aide de la fonction relevel, la classe "Rejeter" comme étant la valeur positive. Cette définition n'est pas nécessaire pour les classifieurs multiconnues.

Dans le cadre des classifications binaire et multiconnue, nous ôterons respectivement le nom de chaque spécimen (variable Nom), son groupe (variable Groupe), et/ou sa comestibilité (variable Type), ces variables étant censées être inconnues du modèle prédictif et ne représentant pas la caractéristique à prédire.

```
fichier_data <- tempfile()
fichier_data <- "~/projects/champis/lot_champis.zip"
fichier_data <- unzip(fichier_data, "lot_champis.csv")
dataset <- read.csv(fichier_data,
                    header = TRUE,
                    sep = ",",
                    stringsAsFactors = TRUE)
dataset$Type <- relevel(dataset$Type, ref = "Rejeter")
dataset <- dataset %>% select(!Nom) %>% select(!Groupe)
```

C.2 Création des jeux d'entraînement, optimisation et évaluation

La création du jeu d'évaluation s'effectue en deux étapes. La première est la définition des rapports des dichotomies entre jeux d'entraînement et optimisation d'une part, et d'évaluation d'autre part. Cette définition implique l'évaluation du nombre d'individus, le calcul du nombre de coefficients p , puis du rapport de dichotomie $f = \sqrt{p} + 1$.^{ac}

```
BI_n_champis <- nrow(dataset)
BI_split_p <- sqrt(BI_n_champis)
BI_split_facteur <- round(sqrt(BI_split_p)+1)
```

La seconde partie consiste à effectuer la scission proprement dite. Cette scission implique la définition d'une liste d'index, de fraction $1 : f$ du nombre d'individus, qui servira via inclusion (jeu d'évaluation) ou exclusion (jeu d'apprentissage et d'évaluation) booléennes des lignes correspondantes, à de constituer chaque jeu de données. La fonction `set.seed` assure la reproductibilité.

```
set.seed(7)
index1 <- twin(data = dataset, r = BI_split_facteur)
BI_lot_appr_opti <- dataset[-index1,]
BI_lot_evaluation <- dataset[index1,]
```

Les lots ainsi obtenus seront ensuite utilisés pour l'entraînement, l'optimisation et l'évaluation finale des performances des modèles.

C.3 Entraînement et optimisation des modèles

Cette partie ne prétend pas à l'exhaustivité, elle se limitera à la présentation de l'entraînement, l'optimisation et la génération de graphiques pour deux modèles : un modèle CART (`rpart`) et un modèle RF (`Rborist`). Un certain nombre de tâches telles que l'entraînement du modèle ou la génération de graphiques sont en réalité attribuées à des fonctions créées *ad hoc* dans le but de clarifier l'organisation du code de l'algorithme, car elles sont effectuées à de nombreuses reprises. Nous décrirons ici le code source sans faire appel à ces fonctions.

C.3.1 Arbre de classification et régression

La première étape est de définir l'indice J de Youden,^{ad} et les pondérations respectives de la sensibilité et de la spécificité. Cette définition n'est pas nécessaire pour les classificateurs multiclasse, le kappa (κ) et l'indice de Rand (R) étant des métriques évaluées nativement par la librairie `caret`.

```
BI_w <- 10
BI_RatioSens <- 2*BI_w/(BI_w+1)
BI_RatioSpec <- 2*(1-BI_w/(BI_w+1))
```

^{ac}cf. section 4.2, page 35.

^{ad}cf. section 4.6, page 49

La seconde définition à préciser est celle de l'espace expérimental des hyperparamètres. En l'espèce le seul hyperparamètre du modèle rpart est la variable cp.

```
BI_grid_rpart_cp <- data.frame(cp = 10^seq(from = -5, to = -1, by = .5))
```

L'étape suivante est de définir les paramètres d'entraînement et d'évaluation des performances du modèle en vue de son optimisation. La fonction trainControl permet ici de préciser les principaux paramètres régissant cette étape :

- classProbs, afin de permettre le calcul des indicateurs de performance (ROC, spécificité, sensibilité...)
- summaryFunction indique que les métriques de performance à utiliser sont celles d'un classifieur binaire (l'argument multiClassSummary sera utilisé pour un classifieur multiclasse)
- method, afin de préciser la méthode de construction des jeux d'entraînement et d'optimisation, ici validation croisée (*CV : cross-validation*)
- number, afin d'indiquer le nombre de blocs de la validation croisée, calculé précédemment.

Ici aussi, la fonction set.seed assure la reproductibilité du processus.

```
set.seed(1)
tr_ctrl <- trainControl(classProbs = TRUE,
                         summaryFunction = twoClassSummary,
                         method = "cv",
                         number = BI_split_facteur)
```

L'entraînement du modèle peut avoir lieu. Ici, le modèle mathématique retenu est l'attribution d'une prédiction sur Type en fonction de toutes les autres variables (Type ~ .). Les arguments data, trControl, tuneGrid font appel aux éléments décrits dans les paragraphes qui précèdent.

```
BI_fit_rpart_cp <- train(Type ~.,
                           method = "rpart",
                           data = BI_lot_appr_opti,
                           trControl = tr_ctrl,
                           tuneGrid = BI_grid_rpart_cp)
```

L'objet résultant est d'une structure relativement complexe. Notre algorithme peut notamment en extraire les résultats relatifs aux performances du modèle, et y adjoindre le calcul du J de Youden pondéré J_w . Dans le cadre des classifieurs multiconfondus, ce calcul est inutile, l'objet généré à l'étape précédente contenant déjà les indicateurs de performance que nous utilisons : kappa (κ) et indice de Rand (R , accuracy).

```
BI_fit_rpart_cp_resultats <- BI_fit_rpart_cp$results %>%
  mutate(Jw = Sens*BI_RatioSens + Spec*BI_RatioSpec - 1)
```

L'objet de type *dataframe* ainsi créé peut être appelé afin d'en extraire des résultats d'intérêt ou d'en inclure le tableau dans le rapport (tableau 15).

Table 15: Tableau des résultats de l'entraînement de rpart

cp	ROC	Sens	Spec	ROCS	SensSD	SpecSD	Jw
0.0000100	0.9997665	0.9997196	0.9992782	0.0002782	0.0002873	0.0008694	0.9993590
0.0000316	0.9998027	0.9997477	0.9991880	0.0001954	0.0003007	0.0009246	0.9993936
0.0001000	0.9997206	0.9997757	0.9988722	0.0002194	0.0002844	0.0009918	0.9993871
0.0003162	0.9996668	0.9997337	0.9984212	0.0002476	0.0003438	0.0013171	0.9992287
0.0010000	0.9990151	0.9984861	0.9939553	0.0007993	0.0007378	0.0033364	0.9961483
0.0031623	0.9716711	0.9938743	0.9226287	0.0052619	0.0019674	0.0113167	0.9747948
0.0100000	0.9340307	0.9805856	0.8263100	0.0060302	0.0026859	0.0130453	0.9331210
0.0316228	0.6577760	0.9720770	0.3400703	0.0046521	0.0019142	0.0094595	0.8292437
0.1000000	0.6275154	0.9601201	0.2949107	0.0059971	0.0024566	0.0110368	0.7992930

L'algorithme génère également un graphique synthétisant les performances du modèle (sensibilité, spécificité, J_w , κ , R ou autre indicateur d'intérêt) en fonction de son hyperparamètre :

- `ggplot` est la fonction de génération du graphique, et permet d'appeler l'objet servant à générer le graphique, ainsi que certains paramètres complémentaires via `aes`. Ici, la variable servant d'abscisse.
- `geom_point` permet de tracer le nuage de points. Ici encore, `aes` permet de préciser, pour chaque nuage de points, la variable d'ordonnée (`Sens`, `Spec` ou `Jw`), ainsi que la légende associée à la couleur des points (*Sensibilité*, *Spécificité* ou *Jw*).
- `geom_line` permet de tracer les lignes correspondant au nuage de points.
- `labs` permet de légender correctement l'attribut `color` de notre légende.
- `ylab` permet de définir la légende l'axe des ordonnées. Ici, de la supprimer, car nous avons trois variables différentes en ordonnées.
- `scale_x_log10` nous permet ici de définir un axe logarithmique décimal en abscisse.
- `theme_bw` attribue le thème(couleur de fond, d'axes, grilles) de type `bw` (*black and white*) à notre graphique.

```
BI_fit_rpart_cp_graphe <- ggplot(data = BI_fit_rpart_cp_resultats, aes(x = cp)) +
  geom_point(aes(y = Sens, color = "Sensibilité")) +
  geom_line(aes(y = Sens, color = "Sensibilité")) +
  geom_point(aes(y = Spec, color = "Spécificité")) +
  geom_line(aes(y = Spec, color = "Spécificité")) +
  geom_point(aes(y = Jw, color = "Jw")) +
  geom_line(aes(y = Jw, color = "Jw")) +
  labs(color = "Performance") +
  ylab(NULL) +
  scale_x_log10() +
  theme_bw()
```

Le graphique ainsi généré peut être intégré dans notre rapport :

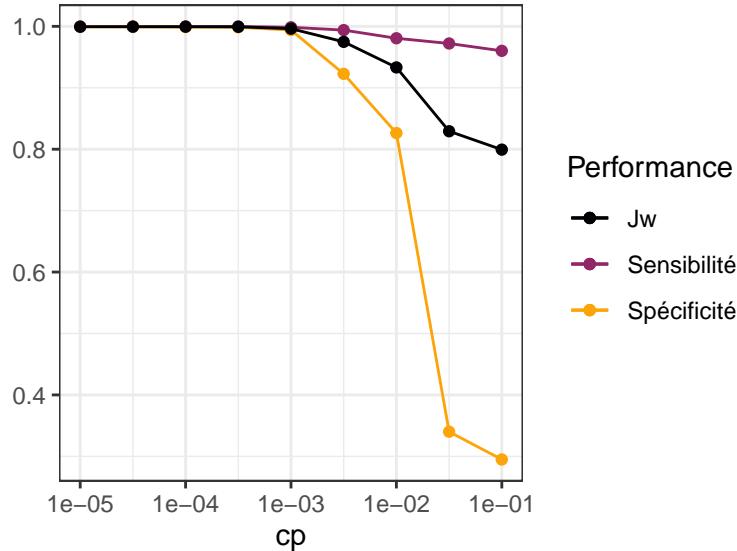


Figure 38: Graphique des performances de rpart

C.3.2 Rborist

La première étape est, comme précédemment, de définir l'espace expérimental. Ici, s'agissant d'un modèle à plusieurs hyperparamètres, l'algorithme utilisera un plan d'expériences basé sur les hypercubes latins. La fonction `nolhDesign` permet de créer un hypercube latin quasi-orthogonal (NOLH : *Near Orthogonal Latin Hypercube*), ici paramétré avec 2 dimensions, dans l'espace $[0; 1]^2$. Le plan d'expérience en est extrait, puis inséré dans un objet de type `dataframe`, avec colonnes nommées d'après nos variables réduites X_1 et X_2 .

```
BI_LHS <- nolhDesign(dimension = 2, range = c(0, 1))$design
BI_LHS <- data.frame(BI_LHS)
colnames(BI_LHS) <- c("X1", "X2")
```

L'hypercube latin des hyperparamètres (`predFixed` et `minNode`) est généré à partir de l'hypercube latin des paramètres réduits. Ces hyperparamètres sont des valeurs entières. L'hypercube latin quasi-orthogonal de dimension 2 possédant 17 expériences équitablement réparties dans l'espace $[0; 1]^2$, il apparaît souhaitable, pour des raisons d'homogénéité dans l'espace expérimental, que $F_n = k \times X_n$ avec k multiple de 16. En pratique, nous n'avons jamais rencontré d'erreurs d'arrondi lors de cette étape mais l'usage de la fonction `round` constitue une précaution supplémentaire garantissant que le produit sera bien un entier.

```
BI_grid_Rborist <- data.frame(BI_LHS) %>%
  mutate(predFixed = round(1+X1*16,0)) %>%
  mutate(minNode = round(1+X2*16,0))
```

Table 16: Plan d'expériences d'entraînement et optimisation du modèle Rborist

X1	X2	predFixed	minNode
0.3125	1.0000	6	17
0.0625	0.2500	2	5
0.1250	0.4375	3	8
0.1875	0.6250	4	11
0.7500	0.9375	13	16
1.0000	0.3125	17	6
0.6250	0.1875	11	4
0.5625	0.8750	10	15
0.5000	0.5000	9	9
0.6875	0.0000	12	1
0.9375	0.7500	16	13
0.8750	0.5625	15	10
0.8125	0.3750	14	7
0.2500	0.0625	5	2
0.0000	0.6875	1	12
0.3750	0.8125	7	14
0.4375	0.1250	8	3

L'entraînement du modèle se déroule de la même façon que pour le modèle rpart.^{ae}

```
set.seed(1)
tr_ctrl <- trainControl(classProbs = TRUE,
                         summaryFunction = twoClassSummary,
                         method = "cv",
                         number = BI_split_facteur)
BI_fit_Rborist <- train(Type ~ .,
                         method = "Rborist",
                         data = BI_lot_appr_opti,
                         trControl = tr_ctrl,
                         tuneGrid = BI_grid_Rborist[c('predFixed', 'minNode')])
```

^{ae}cf. section C.3.1, page 101.

L'algorithme extrait les résultats relatifs aux performances du modèle et y adjoint le calcul de J_w (ou κ pour les classifieurs multiclass), comme précédemment. L'objet obtenu ne contenant que les facteurs expérimentaux (non réduits) des hyperparamètres, il convient d'y adjoindre les facteurs réduits. La table BI_grid_Rborist générée précédemment contient toutes les informations qui permettent, via une jonction, de lier facteurs réduits et facteurs expérimentaux.

```
BI_fit_Rborist_resultats <- BI_fit_Rborist$results %>%
  mutate(Jw = Sens*BI_RatioSens + Spec*BI_RatioSpec - 1) %>%
  left_join(x = .,
            y = BI_grid_Rborist,
            by = c("predFixed", "minNode"))
```

Nous chargeons ensuite notre algorithme de calculer le modèle quadratique avec interactions permettant d'évaluer, à partir des résultats obtenus suite à l'entraînement, la réponse J_w en fonction des X_1 et X_2 , suivant la formule :

$$Y = b_0 + b_1.X_1 + b_2.X_2 + b_{12}.X_1.X_2 + b_{11}.X_1^2 + b_{22}.X_2^2$$

```
BI_mod_Rborist_jw <- modelFit(X = BI_fit_Rborist_resultats[,c("X1", "X2")],
                                  Y = BI_fit_Rborist_resultats$Jw,
                                  type="Kriging",
                                  formula= Y ~ X1 + X2 + X1:X2 + I(X1^2) + I(X2^2))
```

Ces résultats permettent notamment de modéliser les performances sur la totalité de l'espace expérimental des hyperparamètres. A cette fin, l'algorithme est chargé de générer l'ensemble des couples de valeurs (X_1, X_2) possibles, à l'aide de la fonction expand.grid, avant de calculer la valeur J_w correspondante à chaque couple de points.

```
CodBI_pred_Rborist <- expand.grid(CodBI_fit_Rborist_resultats[,c("X1","X2")]) %>%
  mutate(Jw = modelPredict(CodBI_mod_Rborist_jw, .[,c("X1","X2")]))
```

L'ensemble des données expérimentales et modélisées obtenues permettent de générer un graphique bidimensionnel des performances en fonction des hyperparamètres. Pour des raisons didactiques, nous séparerons ici le graphique résultant de l'expérimentation de celui résultant de la modélisation quadratique.

La génération du graphique reprend des principes similaires à ceux présentés précédemment pour le graphique des performances de rpart, notamment au niveau des arguments utilisés dans aes. Les seules fonctions appelant à commentaires sont l'utilisation de geom_raster pour la modélisation, à laquelle nous superposons le graphique généré via geom_tile pour les points expérimentaux. L'utilisation de la gamme de couleurs proposée par viridis permet une visualisation plus aisée des résultats obtenus.

```

BI_pred_Rborist %>% ggplot() +
  geom_raster(data = BI_pred_Rborist,
              aes(x = X1, y = X2, fill = Jw), interpolate = TRUE) +
  geom_tile(data = BI_fit_Rborist_resultats,
            aes(x = X1, y = X2, fill = Jw), color = 'black', linewidth = .5) +
  scale_fill_viridis_c(option = "D", direction = 1) +
  theme(axis.text.y = element_text(angle=90, vjust=.5, hjust=.5)) +
  theme_bw()

```

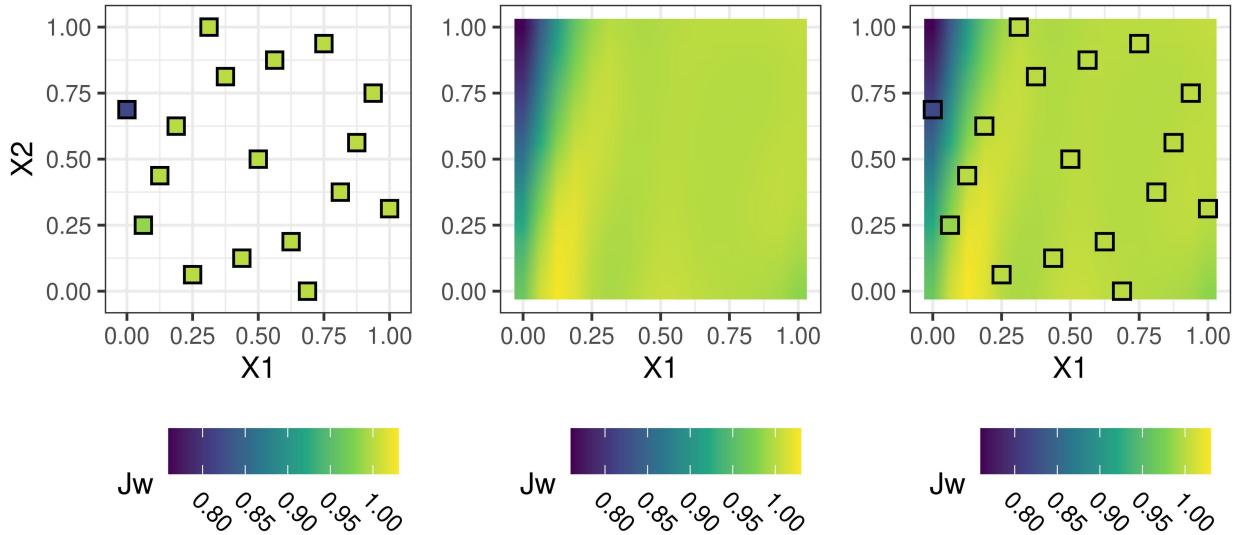


Figure 39: Points expérimentaux (à gauche), modélisation (au centre) et superposition (à droite) des résultats obtenus avec Rborist

L'étape suivante est d'exploiter ce modèle quadratique pour évaluer les hyperparamètres permettant de maximiser J_w . A cet effet, une table de l'ensemble des points de l'espace expérimental des facteurs réduits est générée, puis la fonction quadratique évaluée précédemment est appliquée, afin d'obtenir une prédiction approximative de J_w . Les hyperparamètres sont également calculés, afin de pouvoir être appliqués par la suite.

```

BI_modelquad_Rborist <- expand.grid(X1 = seq(from = 0, to = 1, length.out = 17),
                                      X2 = seq(from = 0, to = 1, length.out = 17)) %>%
  mutate(Jw = BI_mod_Rborist_jw$model@trend.coef[1] +
         BI_mod_Rborist_jw$model@trend.coef[2]*X1 +
         BI_mod_Rborist_jw$model@trend.coef[3]*X2 +
         BI_mod_Rborist_jw$model@trend.coef[4]*X1^2 +
         BI_mod_Rborist_jw$model@trend.coef[5]*X2^2 +
         BI_mod_Rborist_jw$model@trend.coef[6]*X1*X2) %>%
  mutate(predFixed = round(1+X1*16,0)) %>%
  mutate(minNode = round(1+X2*16,0))

```

Le J_w théorique maximal est ensuite évalué, ainsi que les hyperparamètres qui y sont associés.

```
BI_modelquad_Rborist_top <- BI_modelquad_Rborist[which.max(BI_modelquad_Rborist$Jw),
                                                 c("predFixed", "minNode")]
```

Il est ensuite possible de relancer un entraînement, comme précédemment, avec les paramètres optimisés, et en extraire les indicateurs de performances (J_w).

```
set.seed(1)
BI_fit_Rborist_best <- train(Type ~ .,
                                method = "Rborist",
                                data = BI_lot_appr_opti,
                                trControl = tr_ctrl,
                                tuneGrid = BI_modelquad_Rborist_top[c('predFixed', 'minNode')])

BI_fit_Rborist_best_resultats <- BI_fit_Rborist_best$results %>%
  mutate(Jw = Sens*CodBI_RatioSens + Spec*CodBI_RatioSpec - 1)
```

C.4 Évaluation des performances des modèles

L'étape finale est celle de l'évaluation des performances du modèle, ici présentée pour Rborist. Cette évaluation commence par l'extraction des valeurs réelles de comestibilité – c'est à dire des réponses attendues de la part de notre modèle – à partir du jeu de données d'évaluation, ainsi que leur conversion en valeurs booléennes, à fins de comparaison avec les valeurs qui seront prédites. Cette étape n'est évidemment pas nécessaire lors d'une classification multiclasse.

```
BI_evaluation <- BI_lot_evaluation %>%
  mutate(reference = as.factor(case_when(Type == "Rejeter" ~ TRUE,
                                           Type == "Conserver" ~ FALSE)))
```

Les performances du modèle peuvent être évaluées, en termes d'efficience calculatoire, par son temps d'exécution. Un moyen de mesurer le temps d'exécution de n'importe quelle portion de code est de mesurer l'heure de début et de fin d'exécution du code à l'aide de `Sys.time`, puis d'en mesurer la différence via la fonction `difftime`.

Le code exécuté correspond ici à l'entraînement du modèle (cf. *supra*) et à la prédiction sur le lot d'évaluation, enregistré dans un objet dédié.

```
chrono_debut <- Sys.time()
BI_fit_Rborist_final <- train(Type ~ .,
                                 method = 'Rborist',
                                 data = BI_lot_appr_opti,
                                 trControl = tr_ctrl,
                                 tuneGrid = BI_modelquad_Rborist_top[c('predFixed', 'minNode')])
BI_pred_Rborist_final <- predict(object = BI_fit_Rborist_final,
```

```

        newdata = BI_lot_evaluation)

chrono_fin <- Sys.time()
CodBI_temps_Rborist <- difftime(chrono_fin, chrono_debut) %>%
  as.numeric %>%
  round(., 2)

```

L'objet correspondant aux prédictions est ensuite comparé aux valeurs références que notre modèle devait prédire, ce qui permet notamment d'obtenir la matrice de confusion associée aux prédictions.

```

BI_CM_Rborist_final <- confusionMatrix(data = BI_pred_Rborist_final,
                                         reference = BI_lot_evaluation$Type)
BI_CM_Rborist_final$table

```

Nous pouvons également extraire de cette comparaison des indicateurs de performances : sensibilité, spécificité, J_w dans le cas d'une classification binaire, kappa et indice de Rand pour la classification multiclasse, ainsi que le temps de calcul.

```

BI_resultats_Rborist <- BI_CM_Rborist_final$byClass %>%
  t(.) %>%
  as.data.frame(.) %>%
  select(c(Sensitivity, Specificity)) %>%
  mutate(Jw = Sensitivity*BI_RatioSens + Specificity*BI_RatioSpec - 1) %>%
  mutate(temp = BI_temps_Rborist)

```

Enfin, les principaux objets volumineux qui n'ont pas vocation à être exploités par la suite – c'est-à-dire les jeux de données, ainsi que les objets issus des fonctions `train` – sont retirés de l'environnement, qui est ensuite sauvegardé. Cette sauvegarde contient les graphiques, tableaux et valeurs d'intérêt, à fins d'insertion automatisée dans le fichier Rmarkdown qui constitue le corps de texte de cette thèse.

```

rm(dataset, BI_evaluation, BI_lot_appr_opti, BI_lot_evaluation,
   BI_fit_rpart_cp, BI_fit_Rborist, BI_fit_Rborist_best, BI_fit_Rborist_final)

save.image(file = "EKR-Champis-CodeSourceBi.RData")

```

D Annexe 4 : Langage de balisage Rmarkdown

D.1 Introduction et préparation

Rmarkdown est un langage de balisage permettant d'associer les possibilités offertes par R et Latex. Ce langage a été utilisé pour rédiger cette étude, et cette annexe propose d'esquisser les possibilités qu'il offre, en reprenant succinctement la partie décrivant la LDA.

Rmarkdown permet d'utiliser des blocs de code R afin d'effectuer certains calculs, ou d'importer les objets générés par un script R et enregistrés via la fonction `save.image`. C'est cette dernière approche que nous exploiterons.

Le bloc suivant reprend les éléments principaux du script R chargé de calculer la LDA :

```
library(tidyverse)
RMD_Iris_Lot <- iris %>% filter(Species != "virginica") %>% droplevels()
colnames(RMD_Iris_Lot) <- c("Long.Sep", "Larg.Sep", "Long.Pet", "Larg.Pet", "Espece")

# Moyennes intraclasses
RMD_Iris_Moyennes <- RMD_Iris_Lot %>%
  aggregate(. ~ Espece, mean) %>%
  add_row(cbind("Espece"="difference",
    .[1, names(.) != "Espece"] - .[2, names(.) != "Espece"])) %>%
  column_to_rownames("Espece")

# Différences, puis carrés des différences (table III)
RMD_Iris_Deltas <- RMD_Iris_Lot %>%
  group_by(Espece) %>%
  mutate_all(~. - mean(.)) %>%
  ungroup() %>% select(!Espece) %>%
  as.matrix()
RMD_Iris_Produits <- rbind(RMD_Iris_Deltas[,1] %*% RMD_Iris_Deltas,
  RMD_Iris_Deltas[,2] %*% RMD_Iris_Deltas,
  RMD_Iris_Deltas[,3] %*% RMD_Iris_Deltas,
  RMD_Iris_Deltas[,4] %*% RMD_Iris_Deltas)
rownames(RMD_Iris_Produits) <- colnames(RMD_Iris_Produits)

RMD_Iris_InvProduits <- solve(RMD_Iris_Produits) %>% as.matrix() # Matrice inverse (table IV)
RMD_Iris_Coeffs <- as.matrix(RMD_Iris_Moyennes["difference",]) %*% RMD_Iris_InvProduits # Coeffs bruts
RMD_Iris_CoeffsNorm <- RMD_Iris_Coeffs / RMD_Iris_Coeffs[1] # Normalisation

# Coeffs et graphiques LDA
RMD_Iris_Lot <- RMD_Iris_Lot %>%
  mutate(X=rowSums(mapply(`*`, .[,names(.) != "Espece"], RMD_Iris_CoeffsNorm)))

RMD_Iris_GraphMAX <- ggplot(data=RMD_Iris_Lot, aes(x=Larg.Pet, y=Long.Pet, color=Espece)) + geom_point()
RMD_Iris_GraphMin <- ggplot(data=RMD_Iris_Lot, aes(x=Larg.Sep, y=Long.Sep, color=Espece)) + geom_point()
RMD_Iris_GraphX <- ggplot(data=RMD_Iris_Lot, aes(x=X, fill=Espece)) + geom_histogram()

RMD_Iris_Norm <- RMD_Iris_Lot %>%
  mutate_at(., scale, .vars=which(names(.) != "Espece")) %>%
  pivot_longer(data=., cols=which(names(.) != "Espece"))

RMD_Iris_GraphTotale <- ggplot(data=RMD_Iris_Norm, aes(x=name, y=value, fill=Espece)) + geom_boxplot()

save.image(file="CodeSourceIris.RData")
```

Cette partie présente aussi une figure au format vectoriel, enregistrée sous le nom de fichier LDA.pdf, ainsi que des références bibliographiques, enregistrées dans un fichier au format Bibtex, sous le nom IrisBiblio.bib, dont le bloc suivant reprend le contenu :

```
@article{fisher_use_1936,
  title = {The {Use} of {Multiple} {Measurements} in {Taxonomic} {Problems}},
  volume = {7},
  issn = {2050-1439},
  url = {https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x},
  doi = {10.1111/j.1469-1809.1936.tb02137.x},
  number = {2},
  journal = {Annals of Eugenics},
  author = {Fisher, R. A.},
  year = {1936},
  pages = {179--188},
}

@article{anderson_r_1996,
  title = {R. A. Fisher and Multivariate Analysis},
  volume = {11},
  issn = {0883-4237},
  url = {https://www.jstor.org/stable/2246198},
  number = {1},
  journal = {Statistical Science},
  author = {Anderson, T. W.},
  year = {1996},
  note = {Publisher: Institute of Mathematical Statistics},
  pages = {20--34},
}

@book{hastie_elements_2016,
  edition = {2nd},
  series = {Springer Series in Statistics},
  title = {The Elements of Statistical Learning - Data Mining, Inference, and Prediction},
  isbn = {978-0-387-84857-0},
  publisher = {Springer},
  author = {Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome},
  year = {2016},
}
```

D.2 Initialisation

L'initialisation de tout fichier Rmarkdown, commence par un préambule YAML chargé de définir les paramètres à appliquer pour le rendu.

Ce préambule est délimité par les balises ---, et contient, entre autres possibilités :

- La langue dans laquelle est rédigée le fichier,
- La taille de police,
- Le format de sortie,
- Les éventuelles dépendances à ajouter (ici, des librairies Latex),
- La présence ou non d'une table des matières,
- L'indentation des paragraphes,
- Le fichier contenant les bibliographies,
- Le fichier de définition du style de bibliographie.

```
---
```

```
lang: fr
fontsize: 12pt
output:
  bookdown::pdf_document2:
    number_sections: yes
    extra_dependencies: ["float", "placeins", "amssymb", "amsmath"]
    toc: no
  indent: true
  bibliography: [IrisBiblio.bib]
  csl: https://www.zotero.org/styles/vancouver
---
```

La rédaction du texte lui-même peut commencer. Toutefois, dans un souci de clarification du code, la première étape que nous réalisons est la préparation de R, avec le chargement des librairies souhaitées :

- knitr, pour le paramétrage fin des blocs de code sous R,
- ggpubr pour la génération et la mise en page des graphiques,
- kableExtra pour l'amélioration de la mise en page des tableaux.

```
```{r, include = FALSE, warning = FALSE}
load(file = "CodeSourceIris.RData")
library(ggpubr)
library(kableExtra)
library(knitr)
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE, fig.pos = "!H")
...``
```

Les blocs de code sous R commencent tous par ```r, et se terminent par ```. Le préambule de chaque bloc de code indique toujours les paramètres utilisés pour le bloc en question, ici :

- include = FALSE : ni le code, ni les résultats ne doivent apparaître dans le texte,
- warning = FALSE : les messages d'avertissement ne doivent pas apparaître.

Les lignes suivantes n'appellent pas à commentaire particulier, il s'agit du chargement des données générées par le code mentionné précédemment, et des trois librairies d'intérêt.

La toute dernière ligne est plus intéressante, elle permet de définir les paramètres par défaut pour le reste de l'exécution du code :

- echo = TRUE : les résultats du code doivent apparaître dans le texte,
- message = FALSE : les messages relatifs au code ne doivent pas apparaître,
- fig.pos = "!H" : les figures et tableaux "flottants" doivent être insérés au plus près de l'endroit où se situe le code qui les a générés.

### D.3 Rédaction du corps de texte

Après la préparation des sources et l'initialisation via le préambule, la rédaction du texte proprement dit peut commencer :

```
Principes de l'apprentissage machine
Modèles utilisés
Analyses discriminantes
\paragraph*{}
```

Cette étude proposera plusieurs classificateurs s'appuyant sur des méthodes d'analyse discriminante, en particulier l'analyse discriminante linéaire (LDA\ : \*Linear Discriminant Analysis\*).

Ce court passage permet de présenter quelques points typiques de la rédaction d'un texte sous Rmarkdown :

- #, ## et ### servent respectivement à indiquer des sections, sous-sections et sous-sous-sections, ils sont l'équivalent des commandes \section, \subsection et \subsubsection sous Latex,
- \paragraph\*{} est une commande Latex standard, permettant d'insérer un paragraphe non-numéroté,
- \ : permet d'insérer un symbole deux-points précédé d'une espace insécable,
- Les astérisques encadrant un texte permettent de l'écrire en italique (équivalent de \emph sous Latex).

```
\paragraph*{}
```

L'analyse discriminante linéaire est une méthode ayant été proposée par Ronald Fisher en 1936\footnote{Cette étude, proposant une méthode de classification des variétés \emph{Iris setosa}, \emph{Iris virginica} et \emph{Iris versicolor} est par ailleurs à l'origine du célèbre jeu de données \emph{Iris}.} pour résoudre des problèmes de classification taxonomique dans le domaine de la botanique.

[@fisher\_use\_1936; @anderson\_r\_1996] La LDA est basée sur la construction de l'hyperplan de projection permettant de maximiser la distance entre les moyennes projetées des différentes classes et de minimiser la variance intraclasse (voir figure \ref{fig:RMD-Principe-LDA}).[@hastie\_elements\_2016] La LDA peut être utilisée à fins de classification, mais aussi pour effectuer des réductions de dimensionnalité ou encore afin de faciliter l'interprétation de l'importance de certaines caractéristiques.

Le passage qui précède fait usage de certaines commandes Latex typiques :

- \footnote pour l'insertion d'une note de pied de page,
- \emph pour le texte en italique, les astérisques de Rmarkdown ne fonctionnant pas à l'intérieur d'un environnement Latex tel que footnote,
- \ref pour indiquer le numéro de référence d'une figure, ici la figure étiquetée fig:Principe-LDA.

Il y associe une commande Rmarkdown :

- [@xxxx] qui permet d'insérer une citation (équivalent de \cite sous Latex), avec xxxx étant ici la référence de la citation dans le fichier Bibtex.

L'insertion de figures s'effectue en suivant la syntaxe Latex :

```
\paragraph*{}
```

```
\begin{figure}
 \centering
 \includegraphics[width=\linewidth]{LDA}
 \caption{Séparation par distance maximale des moyennes interclasses (à gauche), et par projection sur l'hyperplan optimal tenant compte des variances intraclases (LDA, à droite)}
 \label{fig:RMD-Principe-LDA}
\end{figure}
```

Les commandes utilisées dans ce bloc sont les suivantes :

- \begin{figure} et \end{figure} permettent de délimiter l'environnement figure, qui permet d'insérer une image,
- \centering indique que l'alignement de l'image doit être centré,

- `\includegraphics[width=\linewidth]{LDA}` indique d'insérer une image, dont le nom de fichier est LDA (cf. *supra*), avec une largeur égale à la totalité de la largeur du texte (`width = \linewidth`),
- `\caption{xxx}` indique la légende,
- `\label{xxx}` indique l'étiquette de la figure, qui servira pour les références (fonction `\ref`).

Le bloc suivant n'appelle pas de commentaires particuliers, étant constitué exclusivement de texte :

```
\paragraph*{}
```

En pratique, la LDA consiste à construire un indice synthétique, combinaison linéaire des caractéristiques des classes, dont les coefficients permettent de rendre les points du problème originel le plus aisément "séparables". La LDA étant utilisée dans cette étude pour construire un classifieur binaire, c'est ce type de classifieur qui sera présenté dans cette section, et illustré avec un exemple extrait du jeu de données *\*Iris\**, dans laquelle nous séparerons les espèces *\*Iris versicolor\** et *\*Iris setosa\**.

```
\paragraph*{}
```

Dans ce cadre, la LDA vise ainsi à définir la fonction linéaire\ :

Nous pouvons ensuite introduire des notations mathématiques, toujours au format Latex, ce qui permet d'exploiter un des points forts de ce langage, qui propose un rendu de haute qualité des expression mathématiques :

```
$$X = \sum_{i=1}^n \lambda_i x_i$$
```

avec `$n$` le nombre de paramètres caractérisant les individus, `$x_i$` les caractéristiques mesurées pour chaque individu et chaque paramètre `$i$`, et `$\lambda_i$` des coefficients à optimiser, de sorte que la fonction `$X$` maximise le rapport entre les différences des moyennes de chaque classe `$D$` et la somme des produits des caractéristiques intraclasses `$S$` (proportionnelle à la variance intraclasse), définis par\ :

```
\paragraph*{}
```

```
$$D = \sum_{i=1}^n \lambda_i d_i$$
```

avec `$d_i$` la différence des caractéristiques moyennes pour chaque paramètre `$i$`, et

```
$$S = \sum_{p=1}^n \sum_{q=1}^n \lambda_p \lambda_q S_{pq}$$
```

avec `$S_{pq}$` la somme des produits des caractéristiques intraclasses pour chaque combinaison de paramètres `$p$` et `$q$`.

Les notations mathématiques sous Latex utilisent globalement deux types de balises :

- `$` pour les notations mathématiques intégrées au corps de texte,
- `$$` pour les notations mathématiques et équations à séparer du texte.

Les caractères les plus usuellement utilisés sont

- $\backslash\text{xxx}\{\}$  pour les fonctions et caractères typiques (fonctions trigonométriques, exponentielles, opérateurs de comparaison, flèches, lettres grecques...),
- $\_x\{x\}$  pour l'insertion du caractère x en indice,
- $\^x\{x\}$  pour l'insertion du caractère x en exposant,
- $\backslash\text{frac}\{x\}\{y\}$  pour insérer une fraction de numérateur x et dénominateur  $\^y\{y\}$
- $\backslash\leftX$  et  $\backslash\rightX$  pour insérer des parenthèses, crochets, accolades en remplaçant X par le caractère approprié.

Ainsi, l'expression  $\$X = \backslash\sum_{i=1}^n \backslash\lambda_i.x_i$  sera rendue sous la forme d'une équation intégrée dans le texte,  $X = \sum_{i=1}^n \lambda_i.x_i$ .

L'expression  $\$S=\backslash\sum_{p=1}^n\backslash\sum_{q=1}^n\backslash\lambda_p.\backslash\lambda_q.S_{pq}$  sera quant à elle rendue sous la forme :

$$S = \sum_{p=1}^n \sum_{q=1}^n \lambda_p \cdot \lambda_q \cdot S_{pq}$$

Le bloc de code suivant illustre l'insertion de tableaux issus de *dataframes* générés par R.

```
\paragraph*{}

L'application sur les espèces *Iris versicolor* et *Iris setosa* nous donne les résultats présentés dans les tables \ref{tab:RMD-LDA-TableMoy} et \ref{tab:RMD-LDA-TableProd} :

```{r, RMD-LDA-TableMoy, echo = FALSE}
kable(RMD_Iris_Moyennes, caption = "Moyennes et différences de moyennes des 4 paramètres d'Iris setosa et versicolor") %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

```{r, RMD-LDA-TableProd, echo = FALSE}
kable(RMD_Iris_Produits, caption = "Produits des différences à la moyenne des 4 paramètres d'Iris setosa et versicolor ($S_{pq}$)") %>%
  kable_styling(latex_options = c("striped", "hold_position"))
```

```

L'insertion de tableaux sous Rmarkdown s'effectue grâce à la fonction `kable`, et les paramètres de style peuvent être configurés plus finement à l'aide de la fonction `kable_styling` apportée par la librairie `kableExtra`. Nous avons notamment ajouté des tons bicolores pour le fond des lignes du tableau (`striped`), et bloqué la position des tableaux vis-à-vis du texte (`hold_position`).

Le code se poursuit par des éléments textuels et mathématiques n'appellant pas à commentaires particuliers.

\paragraph\*

La maximisation du rapport entre les carrés des distances des moyennes interclasses et les variances intraclasses revient à maximiser  $D^2/S$  pour chaque coefficient  $\lambda_i$  soit, par dérivation pour chacun des  $\lambda_i$  :

$$\frac{\partial}{\partial \lambda_i} \frac{D^2}{S} = 0$$

\Leftrightarrow

$$\frac{1}{S} \frac{\partial}{\partial \lambda_i} D^2 + D^2 \frac{\partial}{\partial \lambda_i} \frac{1}{S} = 0$$

\Leftrightarrow

$$\frac{D^2 S^2}{S} \left( 2S \frac{\partial}{\partial \lambda_i} D - D \frac{\partial}{\partial \lambda_i} S \right) = 0$$

\Leftrightarrow

$$\frac{1}{2} \frac{\partial}{\partial \lambda_i} S = \frac{1}{D} \frac{\partial}{\partial \lambda_i} D$$

En supposant que les distributions des classes soient unimodales, cette équation admet une solution unique. Le rapport  $S/D$  étant un facteur supposé constant pour tous les coefficients  $\lambda_i$  inconnus, ces coefficients sont donc les solutions du système :

$$\begin{aligned} & \left. \begin{aligned} d_1 &= S_{11}\lambda_1 + S_{12}\lambda_2 + S_{13}\lambda_3 + S_{14}\lambda_4 \\ d_2 &= S_{21}\lambda_1 + S_{22}\lambda_2 + S_{23}\lambda_3 + S_{24}\lambda_4 \\ d_3 &= S_{31}\lambda_1 + S_{32}\lambda_2 + S_{33}\lambda_3 + S_{34}\lambda_4 \\ d_4 &= S_{41}\lambda_1 + S_{42}\lambda_2 + S_{43}\lambda_3 + S_{44}\lambda_4 \end{aligned} \right\} \\ & \Rightarrow S \lambda = D \quad \Leftrightarrow \lambda = S^{-1} \cdot D \end{aligned}$$

avec  $S$  la matrice des produits  $S_{pq}$ ,  $D$  le vecteur des différences des moyennes  $d_i$  et  $\lambda$  celui des coefficients  $\lambda_i$ .

\paragraph\*

En indiquant les facteurs :

- \*  $i = 1$  pour la longueur de sépale  $L_s$ ,
- \*  $i = 2$  pour la largeur de sépale  $\ell_s$ ,
- \*  $i = 3$  pour la longueur de pétales  $L_p$ ,
- \*  $i = 4$  pour la largeur de pétales  $\ell_p$ .

La section de code suivante mêle éléments mathématiques sous Latex et éléments numériques extraits du script R ayant effectué les calculs de la LDA, et permet à Rmarkdown de s'illustrer en exploitant pleinement l'association de Latex pour les formules mathématiques et de R pour les calculs numériques.

```
Nous pouvons calculer les coefficients\ :
\paragraph*{}
$$\left\{ \begin{array}{l} \lambda_1 = `r RMD_Iris_Coeffs[1]` \\ \lambda_2 = `r RMD_Iris_Coeffs[2]` \\ \lambda_3 = `r RMD_Iris_Coeffs[3]` \\ \lambda_4 = `r RMD_Iris_Coeffs[4]` \\ \end{array} \right. .\$\$
```

Soit, après normalisation sur le facteur  $\lambda_1$  :

```
$$\left\{ \begin{array}{l} \lambda_1 = `r round(RMD_Iris_CoeffsNorm[1],3)` \\ \lambda_2 = `r round(RMD_Iris_CoeffsNorm[2],3)` \\ \lambda_3 = `r round(RMD_Iris_CoeffsNorm[3],3)` \\ \lambda_4 = `r round(RMD_Iris_CoeffsNorm[4],3)` \\ \end{array} \right. .\$\$
```

```
\$ X = L_s +
`r round(RMD_Iris_CoeffsNorm[2],3)`.\ell_s
`r round(RMD_Iris_CoeffsNorm[3],3)`.L_p
`r round(RMD_Iris_CoeffsNorm[4],3)`.\ell_p \$\$
```

L'insertion de valeurs – ou de calculs plus complexes – dans le texte s'effectue toujours très simplement, en ouvrant par `r et en fermant par `, qu'il s'agisse :

- D'un environnement mathématique (encadré par \$ ou \$\$),
- D'un tableau créé via Latex (environnement array),
- De texte conventionnel.

Le dernier bloc de code présenté dans cette annexe permet d'insérer des graphiques. De même que pour les tables et les valeurs numériques, Rmarkdown s'illustre ici par ainsi sa grande souplesse qui permet la génération automatique de fichiers textes intégrant graphiques, tableaux et données numériques ou textuelles calculées et mises à jour automatiquement.

```
Le seuil de séparation est alors défini par\ :
$$X_{sep.} = \frac{\overline{X}_{ver.} + \overline{X}_{set.}}{2}$$
\paragraph*{}
```

Avec  $\overline{X_{ver.}}$  et  $\overline{X_{set.}}$  les moyennes respectives des  $X$  pour \*Iris setosa\* et \*Iris versicolor\*.

#### \paragraph\*

La valeur absolue des coefficients  $|\lambda_i|$  calculés précédemment nous indique la pondération de chaque caractère dimensionnel dans l'indice synthétique  $X$  permettant d'obtenir une séparation optimale, ainsi que l'illustrent les figures [\ref{fig:RMD-LDA-MinMax}](#) et [\ref{fig:RMD-LDA-Separation}](#).

```
```{r RMD-LDA-MinMax, echo = FALSE, fig.height = 2.5, fig.cap = "Distribution des variétés setosa et versicolor en fonction de leurs caractéristiques (paramètres fortement pondérés à gauche, faiblement pondérés à droite)"}
plot(ggarrange(widths = c(1, 1.5),
  ncol = 2,
  RMD_Iris_GraphMAX + theme(legend.position = "none"),
  RMD_Iris_GraphMin
)
)
```

```{r RMD-LDA-Separation, echo = FALSE, fig.height = 2.5, fig.cap = "Distribution de X (à gauche) et des paramètres dimensionnels normalisés (à droite) en fonction des espèces"}
plot(ggarrange(widths = c(1, 1.5),
  ncol = 2,
  RMD_Iris_GraphX + theme(legend.position = "none"),
  RMD_Iris_GraphTotale
)
)
```
```

```

Ce bloc de code... [DESCRIPTION DU BLOC PRECEDENT A FINIR]

La fin du code n'appelle pas à commentaire particulier. La table des matières s'insérera automatiquement dans la toute dernière page, conformément au préambule, et avant les annexes éventuelles.

```
### Arbres de décision
etc.

\newpage
# Mini-Bibliographie
```

Le résultat ainsi obtenu est présenté dans les pages suivantes.

1 Principes de l'apprentissage machine

1.1 Modèles utilisés

1.1.1 Analyses discriminantes

Cette étude proposera plusieurs classificateurs s'appuyant sur des méthodes d'analyse discriminante, en particulier l'analyse discriminante linéaire (LDA : *Linear Discriminant Analysis*).

L'analyse discriminante linéaire est une méthode ayant été proposée par Ronald Fisher en 1936¹ pour résoudre des problèmes de classification taxonomique dans le domaine de la botanique.(1,2) La LDA est basée sur la construction de l'hyperplan de projection permettant de maximiser la distance entre les moyennes projetées des différentes classes et de minimiser la variance intraclasse (voir figure 1).(3) La LDA peut être utilisée à fins de classification, mais aussi pour effectuer des réductions de dimensionnalité ou encore afin de faciliter l'interprétation de l'importance de certaines caractéristiques.

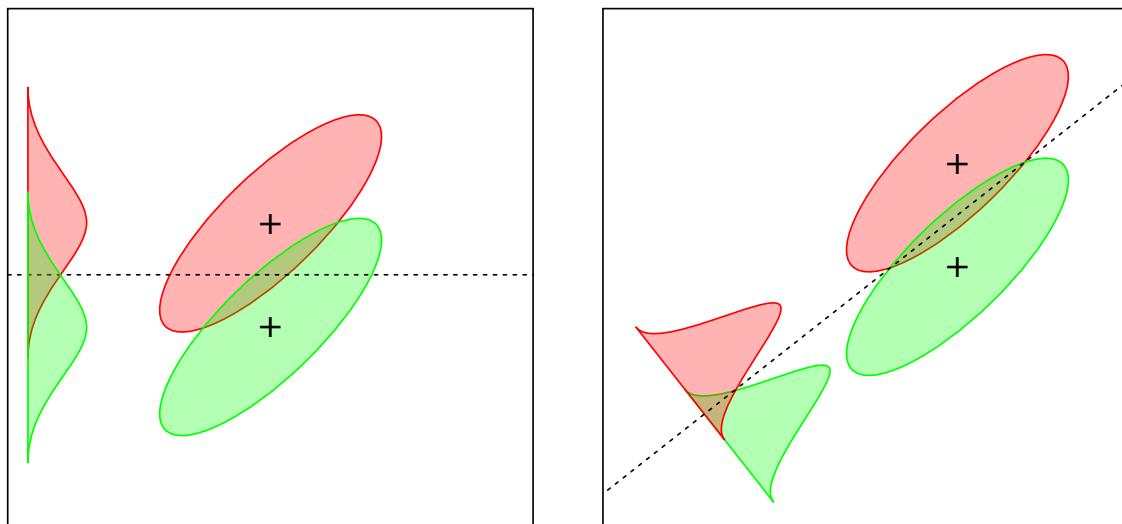


Figure 1: Séparation par distance maximale des moyennes interclasses (à gauche), et par projection sur l'hyperplan optimal tenant compte des variances intraclasses (LDA, à droite)

En pratique, la LDA consiste à construire un indice synthétique, combinaison linéaire des caractéristiques des classes, dont les coefficients permettent de rendre les points du problème originel le plus aisément “séparables”. La LDA étant utilisée dans cette étude pour construire un classificateur binaire, c'est ce type de classifieur qui sera présenté dans cette section, et illustré

¹Cette étude, proposant une méthode de classification des variétés *Iris setosa*, *Iris virginica* et *Iris versicolor* est par ailleurs à l'origine du célèbre jeu de données *Iris*.

avec un exemple extrait du jeu de données *Iris*, dans laquelle nous séparerons les espèces *Iris versicolor* et *Iris setosa*.

Dans ce cadre, la LDA vise ainsi à définir la fonction linéaire :

$$X = \sum_{i=1}^n \lambda_i \cdot x_i$$

avec n le nombre de paramètres caractérisant les individus, x_i les caractéristiques mesurées pour chaque individu et chaque paramètre i , et λ_i des coefficients à optimiser, de sorte que la fonction X maximise le rapport entre les différences des moyennes de chaque classe D et la somme des produits des caractéristiques intraclasses S (proportionnelle à la variance intraclasse), définis par :

$$D = \sum_{i=1}^n \lambda_i \cdot d_i$$

avec d_i la différence des caractéristiques moyennes pour chaque paramètre i , et

$$S = \sum_{p=1}^n \sum_{q=1}^n \lambda_p \cdot \lambda_q \cdot S_{pq}$$

avec S_{pq} la somme des produits des caractéristiques intraclasses pour chaque combinaison de paramètres p et q .

L'application sur les espèces *Iris versicolor* et *Iris setosa* nous donne les résultats présentés dans les tables 1 et 2 :

Table 1: Moyennes et différences de moyennes des 4 paramètres d'Iris setosa et versicolor

	Long.Sep	Larg.Sep	Long.Pet	Larg.Pet
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
difference	-0.930	0.658	-2.798	-1.080

Table 2: Produits des différences à la moyenne des 4 paramètres d'Iris setosa et versicolor (S_{pq})

	Long.Sep	Larg.Sep	Long.Pet	Larg.Pet
Long.Sep	19.1434	9.0356	9.7634	3.2394
Larg.Sep	9.0356	11.8658	4.6232	2.4746
Long.Pet	9.7634	4.6232	12.2978	3.8794
Larg.Pet	3.2394	2.4746	3.8794	2.4604

La maximisation du rapport entre les carrés des distances des moyennes interclasses et les variances intraclasses revient à maximiser D^2/S pour chaque coefficient λ_i soit, par dérivation pour chacun des λ_i :

$$\frac{\partial}{\partial \lambda_i} \frac{D^2}{S} = 0 \Leftrightarrow \frac{1}{S} \frac{\partial}{\partial \lambda_i} D^2 + D^2 \frac{\partial}{\partial \lambda_i} \frac{1}{S} = 0 \Leftrightarrow \frac{D}{S^2} \left(2S \frac{\partial D}{\partial \lambda_i} - D \frac{\partial S}{\partial \lambda_i} \right) = 0 \Leftrightarrow \frac{1}{2} \frac{\partial S}{\partial \lambda_i} = \frac{S}{D} \frac{\partial D}{\partial \lambda_i}$$

En supposant que les distributions des classes soient unimodales, cette équation admet une solution unique. Le rapport S/D étant un facteur supposé constant pour tous les coefficients λ_i inconnus, ces coefficients sont donc les solutions du système :

$$\begin{cases} d_1 = S_{11}\lambda_1 + S_{12}\lambda_2 + S_{13}\lambda_3 + S_{14}\lambda_4 \\ d_2 = S_{21}\lambda_1 + S_{22}\lambda_2 + S_{23}\lambda_3 + S_{24}\lambda_4 \\ d_3 = S_{31}\lambda_1 + S_{32}\lambda_2 + S_{33}\lambda_3 + S_{34}\lambda_4 \\ d_4 = S_{41}\lambda_1 + S_{42}\lambda_2 + S_{43}\lambda_3 + S_{44}\lambda_4 \end{cases} \Rightarrow \mathbf{S} \cdot \boldsymbol{\lambda} = \mathbf{D} \Leftrightarrow \boldsymbol{\lambda} = \mathbf{S}^{-1} \cdot \mathbf{D}$$

avec \mathbf{S} la matrice des produits S_{pq} , \mathbf{D} le vecteur des différences des moyennes d_i et $\boldsymbol{\lambda}$ celui des coefficients λ_i .

En indiquant les facteurs :

- $i = 1$ pour la longueur de sépale L_s ,
- $i = 2$ pour la largeur de sépale ℓ_s ,
- $i = 3$ pour la longueur de pétale L_p ,
- $i = 4$ pour la largeur de pétale ℓ_p .

Nous pouvons calculer les coefficients :

$$\begin{cases} \lambda_1 = 0.0311507 \\ \lambda_2 = 0.1839077 \\ \lambda_3 = -0.222104 \\ \lambda_4 = -0.3147364 \end{cases}$$

Soit, après normalisation sur le facteur λ_1 :

$$\begin{cases} \lambda_1 = 1 \\ \lambda_2 = 5.904 \\ \lambda_3 = -7.13 \\ \lambda_4 = -10.104 \end{cases}$$

$$X = L_s + 5.904 \cdot \ell_s - 7.13 \cdot L_p - 10.104 \cdot \ell_p$$

Le seuil de séparation est alors défini par :

$$X_{sep.} = \frac{\overline{X_{ver.}} + \overline{X_{set.}}}{2}$$

Avec $\bar{X}_{ver.}$ et $\bar{X}_{set.}$ les moyennes respectives des X pour *Iris setosa* et *Iris versicolor*.

La valeur absolue des coefficients λ_i calculés précédemment nous indique la pondération de chaque caractère dimensionnel dans l'indice synthétique X permettant d'obtenir une séparation optimale, ainsi que l'illustrent les figures 2 et 3.

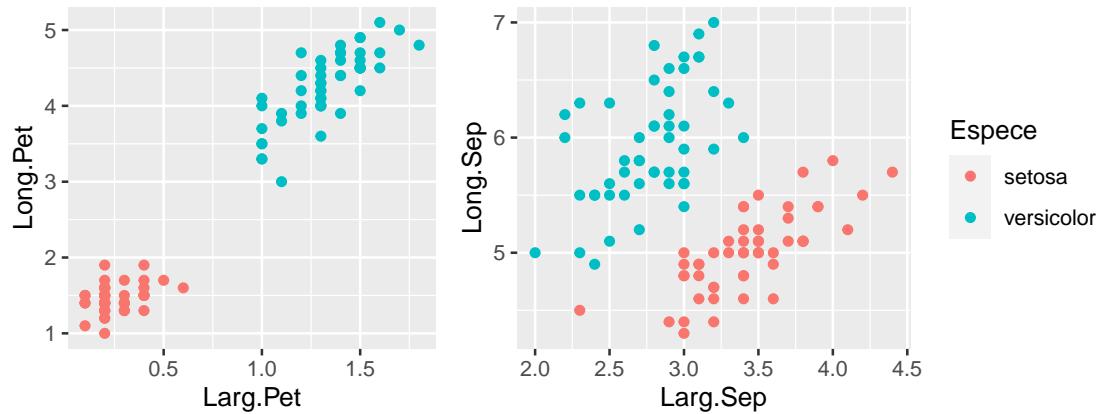


Figure 2: Distribution des variétés setosa et versicolor en fonction de leurs caractéristiques (paramètres fortement pondérés à gauche, faiblement pondérés à droite)

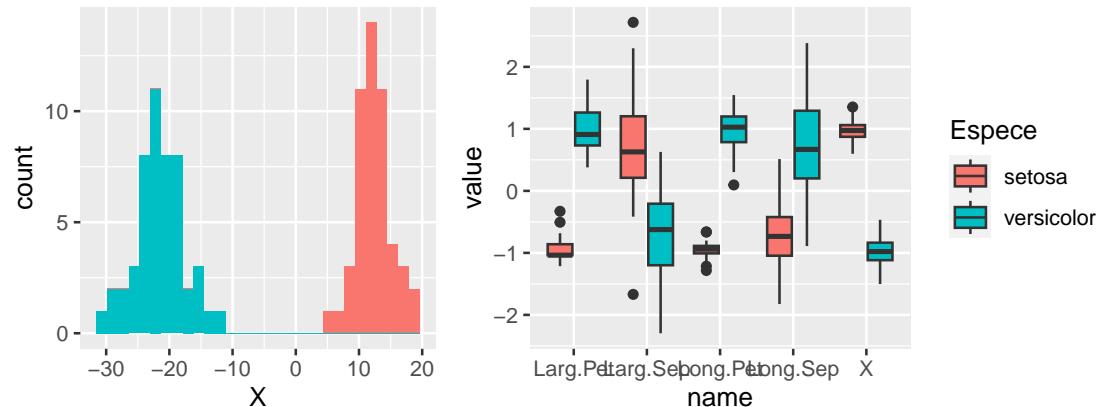


Figure 3: Distribution de X (à gauche) et des paramètres dimensionnels normalisés (à droite) en fonction des espèces

1.1.2 Arbres de décision

etc.

Mini-Bibliographie

1. Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* [Internet]. 1936;7(2):179-88. Disponible sur: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x>
2. Anderson TW. R. A. Fisher and Multivariate Analysis. *Statistical Science* [Internet]. 1996;11(1):20-34. Disponible sur: <https://www.jstor.org/stable/2246198>
3. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. 2nd éd. Springer; 2016. (Springer Series in Statistics).

Université de Lille
FACULTE DE PHARMACIE DE LILLE
DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE
Année Universitaire 2022/2023

Nom : RIHANI

Prénom : Emir Kaïs

Titre de la thèse : Application de modèles d'apprentissage machine à la classification des macromycètes

Mots-clés : intelligence artificielle, apprentissage machine, *machine learning*, classification, R, champignons, mycologie

Résumé : Cette étude propose de soumettre un lot synthétique de macromycètes à différents types de modèles d'apprentissage machine supervisé (analyse discriminante linéaire, arbres décisionnels, forêts aléatoires) et d'évaluer leurs performances respectives dans des tâches de classification par critères de comestibilité, par familles, puis par espèces.

Membres du jury :

Président :

Pr LEMDANI Mohammed, PU en Biomathématiques, Faculté de Pharmacie de Lille

Directeur, conseiller de thèse :

Dr HAMONIER Julien, MCU en Biomathématiques, Faculté de Pharmacie de Lille

Assesseur :

Dr WELTI Stéphane, MCU en Sciences Végétales et Fongiques, Faculté de Pharmacie de Lille

Membre extérieur :

Dr MOUSSET Caroline, Pharmacien-Ingénieur, Responsable AQ Clients, Delpharm Lille