

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

**Soutenue publiquement le 12 septembre 2023
Par M. RIHANI Emir Kaïs**

**APPLICATION DE MODELES D'INTELLIGENCE ARTIFICIELLE
A LA CLASSIFICATION DES MACROMYCETES**

Membres du jury :

Président : Nom, Prenom, titre et lieu de fonction

Directeur, conseiller de thèse : Nom, Prenom, titre et lieu de fonction

Assesseur(s) : Nom, Prenom, titre et lieu de fonction

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 à :

La Dream Team de tous ceux que je vais remercier

Table des matières

1	Introduction	16
1.1	Propos liminaire	16
1.2	But de l'étude	16
1.3	Etat de l'art des lots de données mycologiques	17
2	Création du lot de données	18
2.1	Configuration matérielle et logicielle	18
2.2	Conception d'un lot de données synthétiques	18
2.2.1	Principes généraux	18
2.2.2	Génération des paramètres quantitatifs	18
2.2.3	Génération des paramètres qualitatifs	24
3	Principes de l'apprentissage machine	25
3.1	Jeux de données	25
3.2	Modèles utilisés	26
4	Apprentissage machine et classification binaire	27
5	Apprentissage machine et classification multiclasse	28
6	Robustesse de la classification	29
	Références bibliographiques	30

1 Introduction

1.1 Propos liminaire

L'identification des macromycètes 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. Plusieurs genres, parfois très connus, ne sont pas présents, parmi lesquels nous pouvons par exemple citer le genre *psylocybe*, connu pour ses 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*).

L'ingestion de certains de ces champignons est *mortelle*, même à de petites doses. Le diagnostique 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.

Il ne faut jamais, *sous aucune circonstance*, utiliser ce type de lot de données afin de déterminer si un champignon est comestible ou non.

1.2 But de l'étude

L'identification des plantes et champignons est un problème de classification classique, qui est habituellement effectué manuellement à l'aide de clés d'identification. La plupart de ces clés sont basées sur un processus utilisant des arbres décisionnels, ce qui semble logique car rappelant la logique en arbre de l'évolution. Toutefois, cet argument rencontre quelques limites :

La première limite est le nombre de chaînons manquants. Certaines espèces sont évidemment éteintes, ce qui signifie que certaines branches et noeuds de l'arbre phylogénétique sont manquants, ce qui peut compliquer l'analyse quand deux espèces apparentées ont un nombre élevé de chaînons et noeuds communs manquants. Certaines similarités entre espèces peuvent également ne pas être identifiées.

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 un aspect très différent en raison de leur adaptation à des environnements très différents. D'un autre côté, la convergence évolutive explique la similarité entre l'aile de la chauve-souris et l'aile 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 est l'analyse des génomes : les caractéristiques visibles peuvent être trompeuses. Malheureusement, ces caractéristiques sont souvent les seules aisément observables.

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 : il est essentiellement centré sur la toxicité contre les humains, 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 évolutif, pourrait ne pas nécessairement être l'approche optimale à la classification des espèces basée sur des critères macroscopiques. Le but de cette étude est d'effectuer cette tâche de classification basée sur des indices visuels limités, et d'évaluer les performances relatives de différentes stratégies de classification.

1.3 Etat 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 1987.¹

Un lot de données plus conséquent a été publié par Dennis Wagner en 2021² et mis en libre accès sous le nom de *Secondary Mushroom Dataset*.

2 Création du lot de données

2.1 Configuration matérielle et logicielle

Le code d'apprentissage machine, les méthodes d'évaluation, ainsi que cette thèse ont été rédigés sur l'équipement suivant :

- CPU : AMD Ryzen 5 5600G
- RAM : 2x16 Go DDR4-3200
- SSD : Crucial P5 M2 NVMe
- OS : Xubuntu Linux 20.04 LTS
- R : version 4.1.0 (2021)
- RStudio : version 1.4.1717 "Juliet Rose"
- Librairies : tidyverse³ (v1.3.1), microbenchmark⁴ (v1.4.9), SPLIT⁵ (v1.2), MASS⁶ (v7.3.54), caret⁷ (v6.0.88), ?GGally?⁸ (v2.1.2), ?mda?⁹ (v0.5.2), rpart¹⁰ (v4.1.15), ?plyr?¹¹ (v1.8.6), C50¹² (v0.1.5), party¹³ (v1.3.9), ranger¹⁴ (v0.13.1), e1071¹⁵ (v1.7.8), rFerns¹⁶ (v5.0.0), Rborist¹⁷ (v0.2.3), rmarkdown¹⁸ (v2.11), knitr¹⁹ (v1.34), ggpibr²⁰ (v0.4.0).

2.2 Conception d'un lot de données synthétiques

2.2.1 Principes généraux

Un lot de données synthétiques est un lot de données générée 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 type variés, que nous pouvons grossièrement regrouper en données alphanumériques (quantitatives et qualitatives) et en données d'imagerie.

Pour des raisons pratiques, 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.²¹⁻²³

2.2.2 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 :

$$\begin{cases} L_S = L_{Smax} \cdot F_{Ls} \\ D_S = D_{Smax} \cdot F_{Ds} \\ D_C = D_{Cmax} \cdot F_{Dc} \end{cases}$$

Avec :

- L_{Smax} , D_{Smax} et D_{Cmax} les valeurs maximales de longueur de stipe, diamètre du stipe et diamètre de chapeau de chaque variété de champignon, 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, la recherche bibliographique sur la cinétique de croissance des sporophores n'ayant pas permis de distinguer de différences de la cinétique de croissance de chacun de ces trois paramètres, nous supposerons en première approximation que la croissance du stipe en longueur, en largeur, ainsi que la croissance du chapeau s'effectuent à des vitesses identiques, nous obtenons donc :

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

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.

Le premier critère n'est, en pratique, pas un facteur limitant, les temps de calcul pour la génération d'un nombre de facteurs de taille F_T suffisant étant typiquement inférieurs à 400 ms (pour 10^6 itérations) avec la plupart des distributions d'intérêt (voir figure 1).

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.

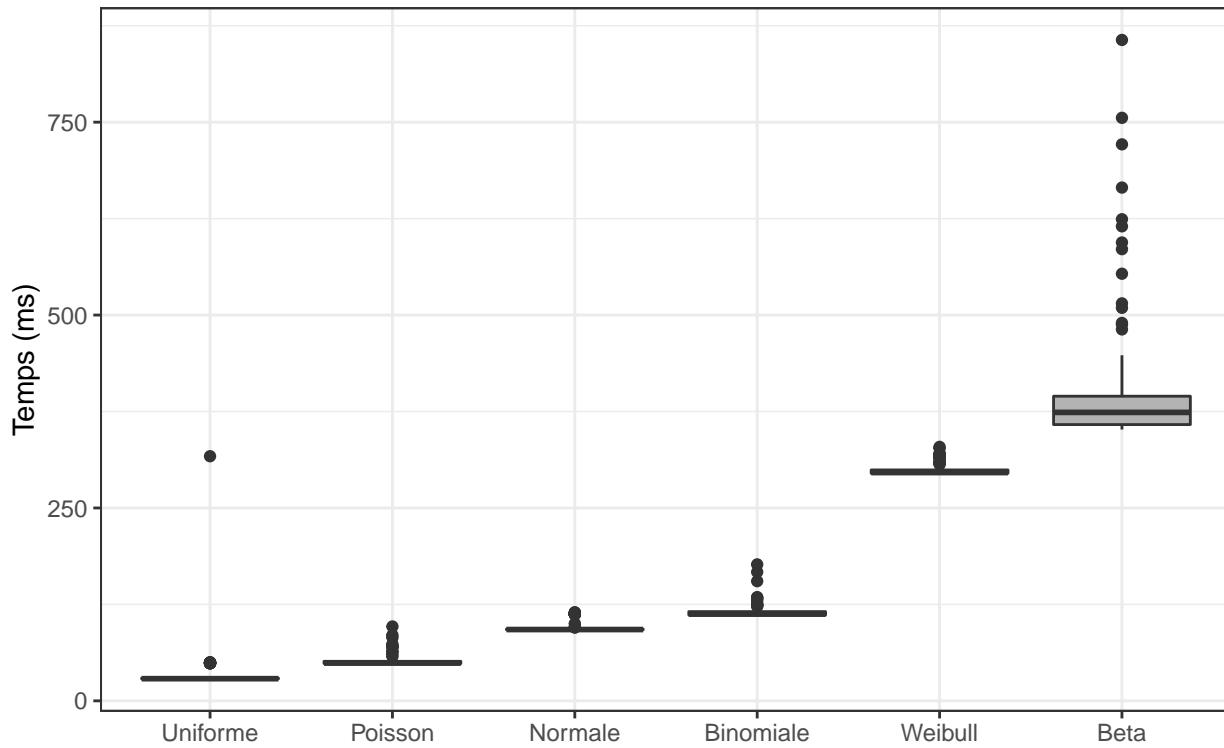


FIGURE 1 – Temps de calcul des principales distributions d'intérêt pour 1e+06 itérations

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,^{24,25}
- 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 à l'issue 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.

En raison des contraintes imposées précédemment ainsi que de sa grande polyvalence²⁶, la loi retenue dans le cadre de cette étude pour la génération des facteurs de taille aléatoires (F_T)

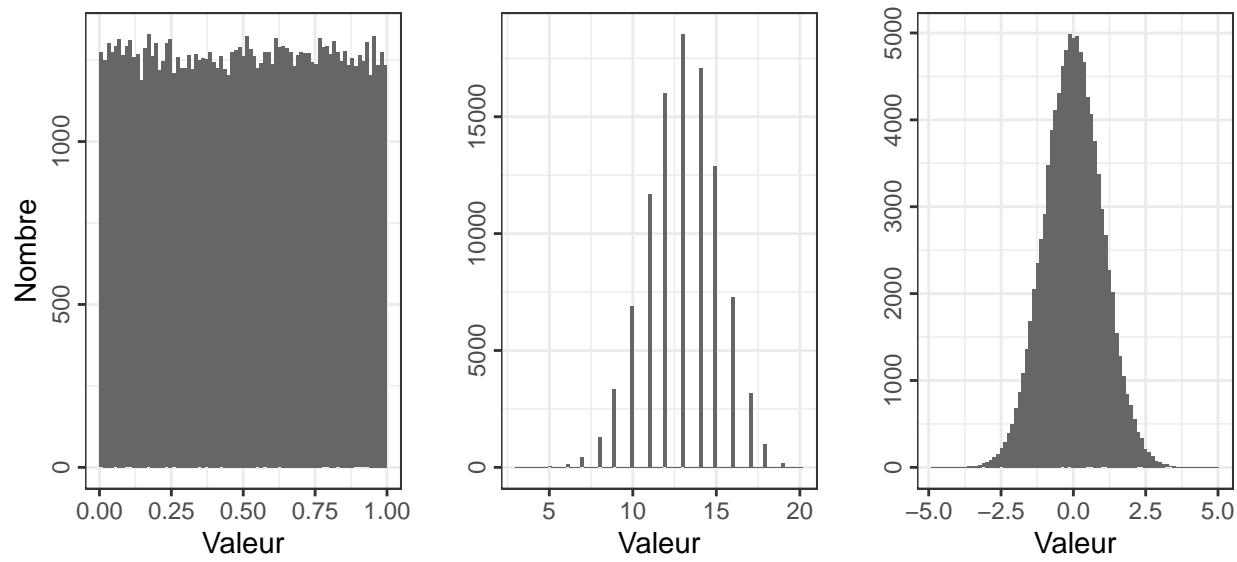


FIGURE 2 – Exemples de distributions de la loi uniforme (à gauche), binomiale (au centre) et normale (à droite)

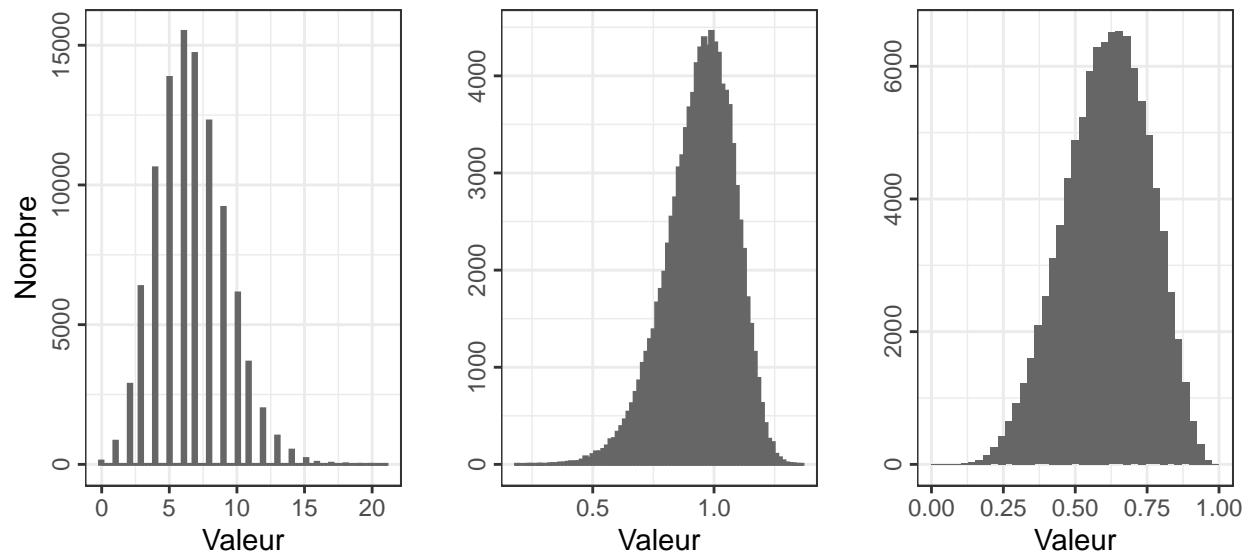


FIGURE 3 – Exemples de distributions de la loi de Poisson (à gauche), de Weibull (au centre) et bêta (à droite)

est une loi bêta non-centrale, définie comme la fonction de distribution de :^{26,27}

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

Avec, comme paramètres définis empiriquement pour cette étude :

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

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 4.

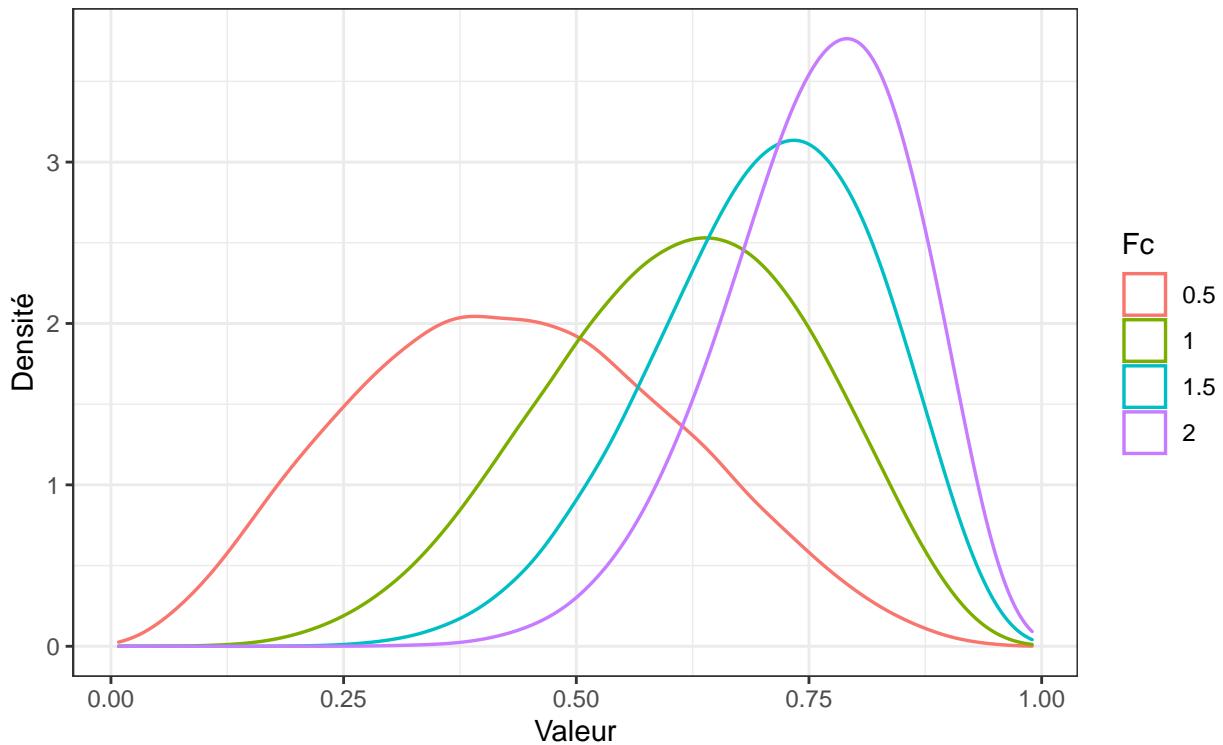


FIGURE 4 – Distribution de différentes lois bêta, en fonction du facteur de croissance F_c

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, afin de proposer le modèle suivant :

$$\begin{cases} L_s = L_{smax}.F_T.\delta_{Ls} & \text{avec } \delta_{Ls} \sim \mathcal{N}(\mu = 1; \sigma = 0.05) \\ D_s = D_{smax}.F_T.\delta_{Ds} & \text{avec } \delta_{Ds} \sim \mathcal{N}(\mu = 1; \sigma = 0.05) \\ D_c = D_{cmax}.F_T.\delta_{Dc} & \text{avec } \delta_{Dc} \sim \mathcal{N}(\mu = 1; \sigma = 0.05) \end{cases}$$

L'impact de la dispersion sur la distribution des paramètres de taille L_s , D_s et D_c est illustré par les figures 5 et 6.

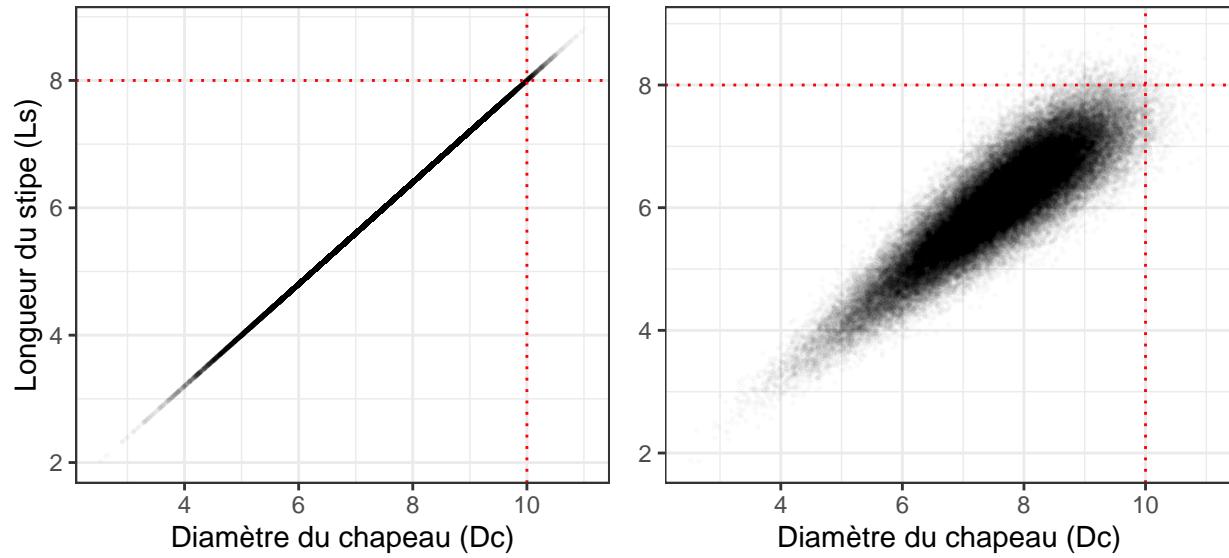


FIGURE 5 – Nuages de points de 2 paramètres de taille, sans dispersion (à gauche) et avec dispersion (à droite)

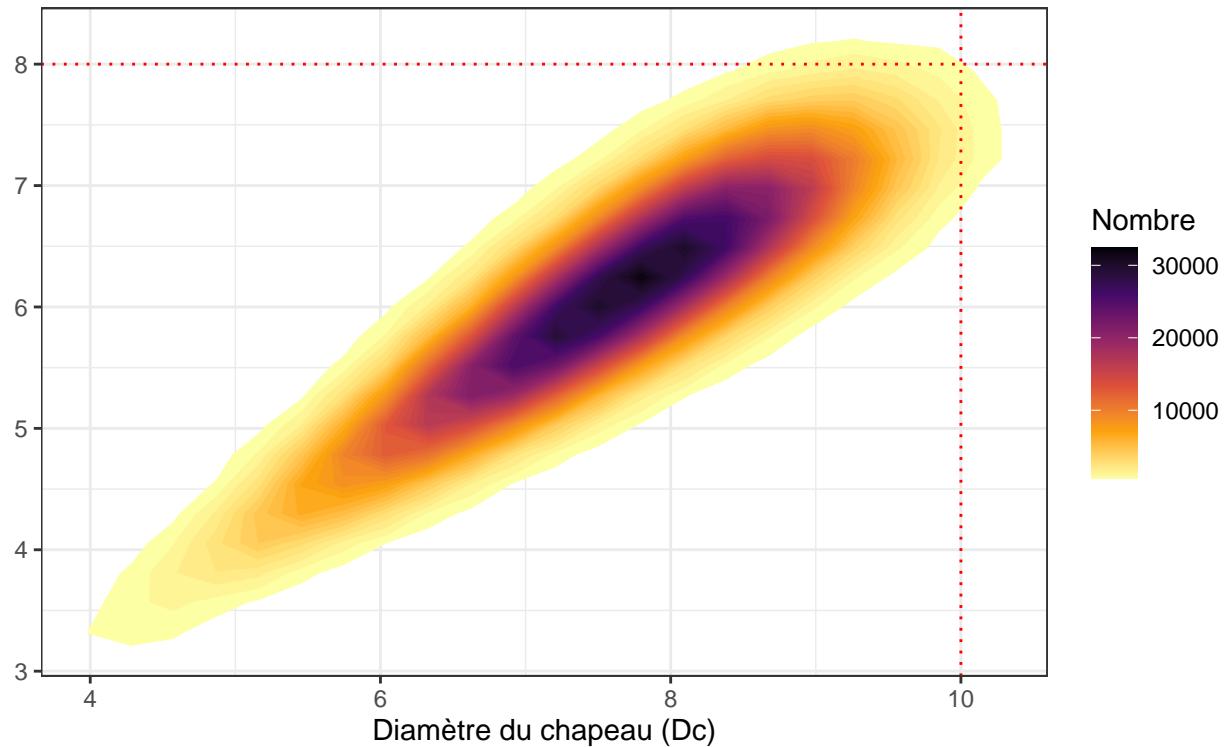


FIGURE 6 – Diagramme de densité de 2 paramètres de taille, avec dispersion

Une simulation de Monte Carlo unidimensionnelle effectuée sur 10^5 spécimens nous permet d'évaluer la proportion de spécimens "hors normes" dépassant la valeur dimensionnelle maximale extraite de la littérature à environ 0.428 % (cf. figure 7), et la proportion de spécimens dépassant de plus de 10% cette valeur maximale est inférieure à 0.1 %.

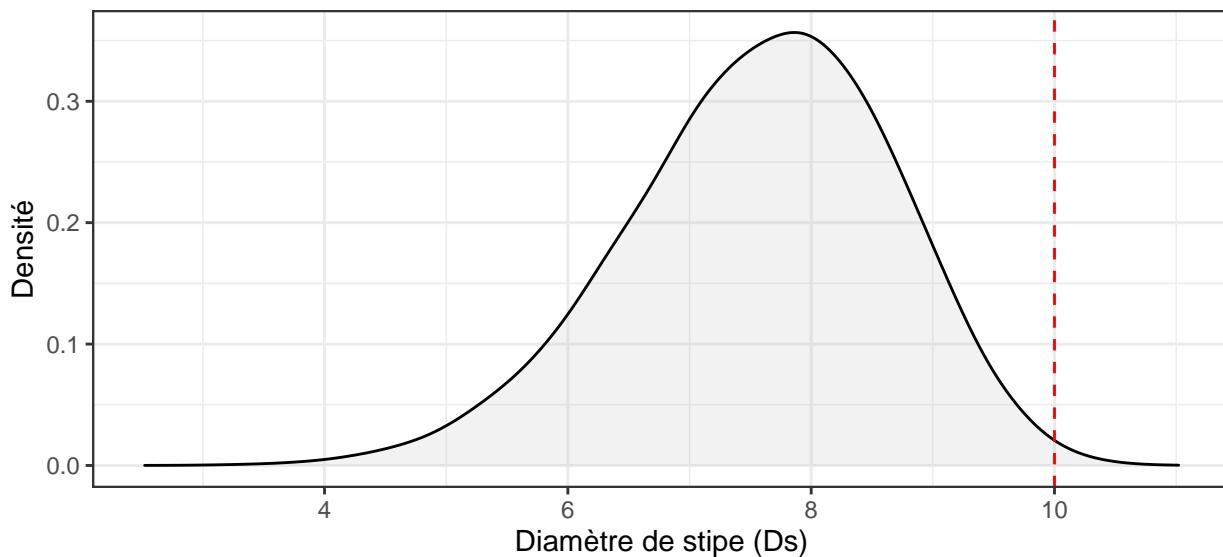


FIGURE 7 – Distribution du diamètre de stipe D_s , pour $D_{s\max} = 10$

2.2.3 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 valeur, et une valeur sera tirée aléatoirement parmi celles de ce vecteur pour caractériser chaque spécimen.

3 Principes de l'apprentissage machine

3.1 Jeux de données

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

1. Entraînement : le modèle d'apprentissage est exposé à un *jeu de données d'entraînement* (*training data set*), censé être représentatif (cf. section 2.2.1) des données auquel le modèle sera exposé en utilisation réelle.
2. Validation : le modèle d'apprentissage développé à l'étape précédente, sera soumis à un *jeu de données de validation* (*validation data set*). Les prédictions (ex: comestibilité, espèce...) proposées par le modèle d'apprentissage sur la base des informations contenues dans le lot de données de validation (ex : dimensions, couleurs, morphologie du champignon...) sont comparées avec les valeurs réelles (ex : comestibilité, espèce...), ce qui permet d'évaluer les performances prédictives du modèle proposé en fonction des indicateurs retenus (spécificité, sensibilité, F1-score, 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) – idéalement en suivant un plan d'expériences – 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* est parfois qualifiée d'étape de *validation* dans la littérature.²⁸

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 : l'optimisation effectuée lors de l'étape de validation aboutit à un modèle potentiellement biaisé (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.

Dans un souci de clarté, nous utiliserons les termes lots et de phases d'entraînement, d'optimisation et d'évaluation dans la suite de cette étude.

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 du lot de données initial, avec des proportions variables :

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

[Schéma Split Apprentissage/Optimisation/Validation]

Le rapport de taille entre jeux de données entraînement, optimisation, évaluation de cette étude suit la loi $p : \sqrt{p} : \sqrt{p} + 1$, avec p le nombre de coefficients du modèle.²⁹

Pour un nombre de coefficients compris entre 100 et 200, cette règle nous conduit à retenir :

$$\begin{cases} R_{entr} = 85 \pm 2\% \\ R_{opti} = 7 \pm 1\% \\ R_{eval} = 8 \pm 1\% \end{cases}$$

Ce rapport 85:7:8 peut en pratique être obtenu par deux découpages successifs avec un rapport 92:8.

Dans cette étude, la division de ces trois jeux de données utilise une méthode de découpage basée sur les points-supports³⁰ (*support-points based splitting*) exploitant un algorithme du plus proche voisin (NN : *Nearest Neighbour*), basé sur un arbre Kd (*k-dimensional tree*), afin d'optimiser la représentativité des jeux de données par rapport à ceux pouvant être obtenus par un découpage aléatoire.³¹

3.2 Modèles utilisés

4 Apprentissage machine et classification binaire

texte

5 Apprentissage machine et classification multiclasse

texte

6 Robustesse de la classification

texte

Références bibliographiques

1. Schlimmer J. Mushroom Data Set. University of California. 1987; Disponible sur: <https://archive.ics.uci.edu/ml/datasets/Mushroom>
2. 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>
3. Wickham H. tidyverse: Easily Install and Load the Tidyverse [Internet]. 2021. Disponible sur: <https://CRAN.R-project.org/package=tidyverse>
4. Mersmann O. microbenchmark: Accurate Timing Functions [Internet]. 2021. Disponible sur: <https://github.com/joshuaulrich/microbenchmark/>
5. Vakayil A, Joseph R, Mak S. SPLIT: Split a Dataset for Training and Testing [Internet]. 2022. Disponible sur: <https://CRAN.R-project.org/package=SPlit>
6. Ripley B. MASS: Support Functions and Datasets for Venables and Ripley's MASS [Internet]. 2021. Disponible sur: <http://www.stats.ox.ac.uk/pub/MASS4/>
7. Kuhn M. caret: Classification and Regression Training [Internet]. 2021. Disponible sur: <https://github.com/topepo/caret/>
8. 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>
9. Trevor Hastie & Robert Tibshirani. Original R port by Friedrich Leisch S original by, Hornik K, code. BDRipleyBN has contributed to the upgrading of the. mda: Mixture and Flexible Discriminant Analysis [Internet]. 2020. Disponible sur: <https://CRAN.R-project.org/package=mda>
10. Therneau T, Atkinson B. rpart: Recursive Partitioning and Regression Trees [Internet]. 2019. Disponible sur: <https://CRAN.R-project.org/package=rpart>
11. Wickham H. plyr: Tools for Splitting, Applying and Combining Data [Internet]. 2020. Disponible sur: <https://CRAN.R-project.org/package=plyr>
12. Kuhn M, Quinlan R. C50: C5.0 Decision Trees and Rule-Based Models [Internet]. 2021. Disponible sur: <https://topepo.github.io/C5.0/>
13. Hothorn T, Hornik K, Strobl C, Zeileis A. party: A Laboratory for Recursive Partytioning [Internet]. 2021. Disponible sur: <http://party.R-forge.R-project.org>
14. Wright MN, Wager S, Probst P. ranger: A Fast Implementation of Random Forests [Internet]. 2021. Disponible sur: <https://github.com/imbs-hl/ranger>
15. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [Internet]. 2021. Disponible sur: <https://CRAN.R-project.org/package=e1071>
16. Kursa MB. rFerns: Random Ferns Classifier [Internet]. 2021. Disponible sur: <https://gitlab.com/mbq/rFerns>
17. Seligman M. Rborist: Extensible, Parallelizable Implementation of the Random Forest Algorithm [Internet]. 2019. Disponible sur: <https://CRAN.R-project.org/package=Rborist>
18. Allaire J, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, et al. rmarkdown: Dynamic Documents for R [Internet]. 2021. Disponible sur: <https://CRAN.R-project.org/package=rmarkdown>

19. Xie Y. knitr: A General-Purpose Package for Dynamic Report Generation in R [Internet]. 2021. Disponible sur: <https://yihui.org/knitr/>
20. Kassambara A. ggpubr: ggplot2 Based Publication Ready Plots [Internet]. 2020. Disponible sur: <https://rpkgs.datanovia.com/ggpubr/>
21. Courtecuisse R. Clé de détermination macroscopique des champignons supérieurs des régions du Nord de la France. Société mycologique du Nord de la France; 1986.
22. Courtecuisse R, Duhem B. Champignons de France et d'Europe. Delachaux et Niestlé; 2013. (Guides Delachaux).
23. Courtecuisse R, Moreau PA, 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.
24. Money NP. 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>
25. 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>
26. 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).
27. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Disponible sur: <https://www.R-project.org/>
28. 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/>
29. 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>
30. 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>
31. 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>

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'intelligence artificielle à la classification des macromycètes

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

Résumé : L'IA c'est génial !

Membres du jury :

Président : Nom, Prenom, titre et lieu de fonction

Assesseur(s) : Nom1, Prenom1, titre et lieu de fonction
Nom2, Prenom2, titre et lieu de fonction
Nom3, Prenom3, titre et lieu de fonction

Membre(s) extérieur(s) : Nom1, Prenom1, titre et lieu de fonction
Nom2, Prenom2, titre et lieu de fonction
Nom3, Prenom3, titre et lieu de fonction