Université de Lille
Année Universitaire 2022/2023

Faculté de Pharmacie de Lille

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	Régis BORDET
Premier Vice-président	Etienne PEYRAT
Vice-présidente Formation	Christel BEAUCOURT
Vice-président Recherche	Olivier COLOT
Vice-présidente Réseaux internationaux et européens	Kathleen O'CONNOR
Vice-président Ressources humaines	Jérôme FONCEL
Directrice Générale des Services	Marie-Dominique SAVINA

UFR3S

Doyen	Dominique LACROIX
Premier Vice-Doyen	Guillaume PENEL
Vice-Doyen Recherche	Éric BOULANGER
Vice-Doyen Finances et Patrimoine	Damien CUNY
Vice-Doyen Coordination pluriprofessionnelle et Formations sanitaires	Sébastien D'HARANCY
Vice-Doyen RH, SI et Qualité	Hervé HUBERT
Vice-Doyenne Formation tout au long de la vie	Caroline LANIER
Vice-Doyen Territoires-Partenariats	Thomas MORGENROTH
Vice-Doyenne Vie de Campus	Claire PINÇON
Vice-Doyen International et Communication	Vincent SOBANSKI
Vice-Doyen étudiant	Dorian QUINZAIN

Faculté de Pharmacie

Doyen	Delphine ALLORGE
Premier Assesseur et Assesseur en charge des études	Benjamin BERTIN
Assesseur aux Ressources et Personnels	Stéphanie DELBAERE
Assesseur à la Santé et à l'Accompagnement	Anne GARAT
Assesseur à la Vie de la Faculté	Emmanuelle LIPKA
Responsable des Services	Cyrille PORTA
Représentant étudiant	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.	DHIFLI	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 Liste des abréviations

AI : *Artificial Intelligence* (intelligence artificielle)
EDA : *Exploratory Data Analysis* (analyse exploratoire des données)
GAM : *Generalized Additive Model* (modèle additif généralisé)
IA : Intelligence Artificielle
LDA : *Linear Discriminant Analysis* (analyse linéaire discriminante)
PDA : *Penalized Discriminant Analysis* (analyse discriminante pénalisée)
RF : *Random Forest* (forêt aléatoire)
XGB : *eXtreme Gradient Boosting*

2 Introduction

2.1 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 foi, détruisant les cellules hépatiques, puis excrétées dans l'intester, 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.

2.2 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 évolutive 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 chave-souris et l'aile de l'abeille. Toutefois, malgré leur apparente dissimilarité, 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 visible, 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.

2.3 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 2021[2] et mis en libre accès sous le nom de *Secondary Mushroom Dataset*.

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 :

- CPU : AMD Ryzen 5 5600G
- RAM : 2x16 Go DDR4-3200
- SSD : Crucial P5 M2 NVMe
- OS : Xubuntu Linux 20.04.1 LTS
- R : version 4.2.2 (2022)
- RStudio : version 2022.7.2.576 "Spotted Wakerobin"
- Librairies : tidyverse[3] (v1.3.2), microbenchmark[4] (v1.4.9), SPlit[5] (v1.2)„ MASS[6] (v7.3.58.2), caret[7] (v6.0.93), ?GGally?[8] (v2.1.2), ?mda?[9] (v0.5.3), rpart[10] (v4.1.19), ?plyr?[11] (v1.8.8), C50[12] (v0.1.8), party[13] (v1.3.11), ranger[14] (v0.14.1), e1071[15] (v1.7.13), rFerns[16] (v5.0.0), Rborist[17] (v0.3.2), rmarkdown[18] (v2.20), knitr[19] (v1.41), ggpubr[20] (v0.6.0).

3.2 Conception d'un lot de données synthétiques

3.2.1 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 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]

3.2.2 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 :

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

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 200 ms (pour 10^6 facteurs générés) avec la plupart des distributions d'intérêt (voir figure 1).

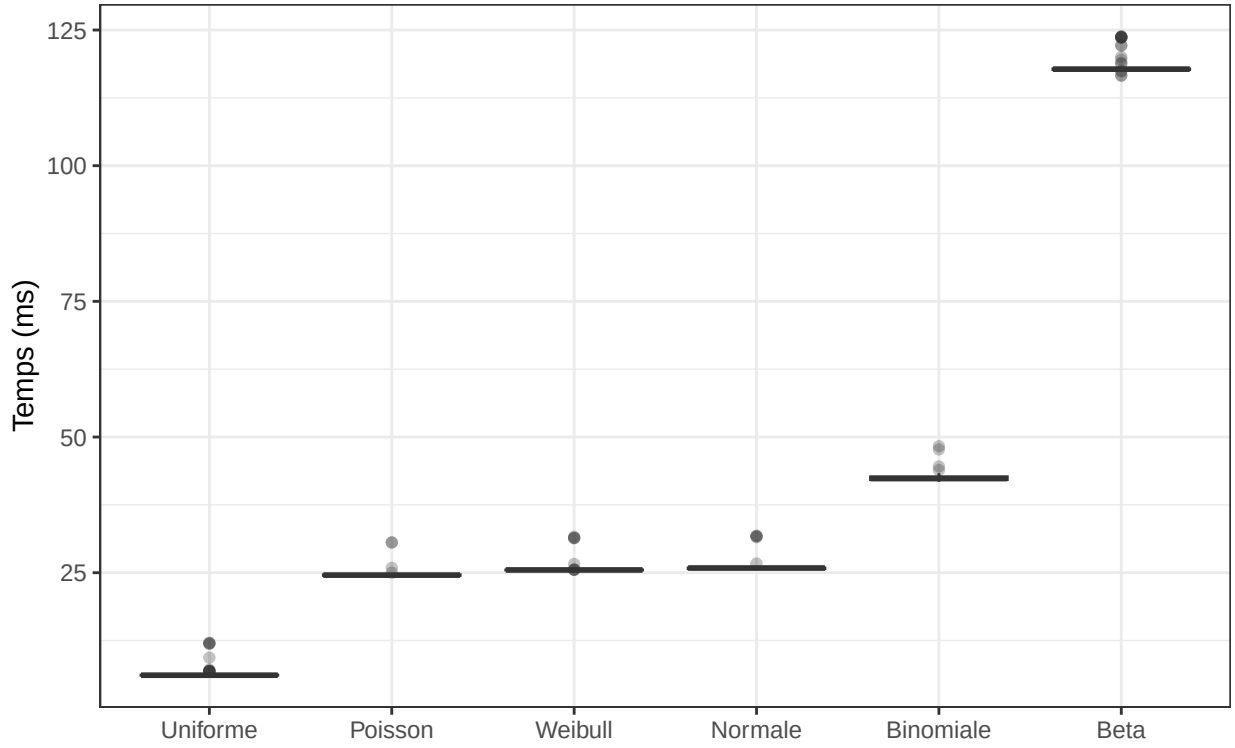

Figure 1: Temps de calcul des principales distributions d'intérêt pour 1e+06 facteurs, (100 iter.)

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,[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.

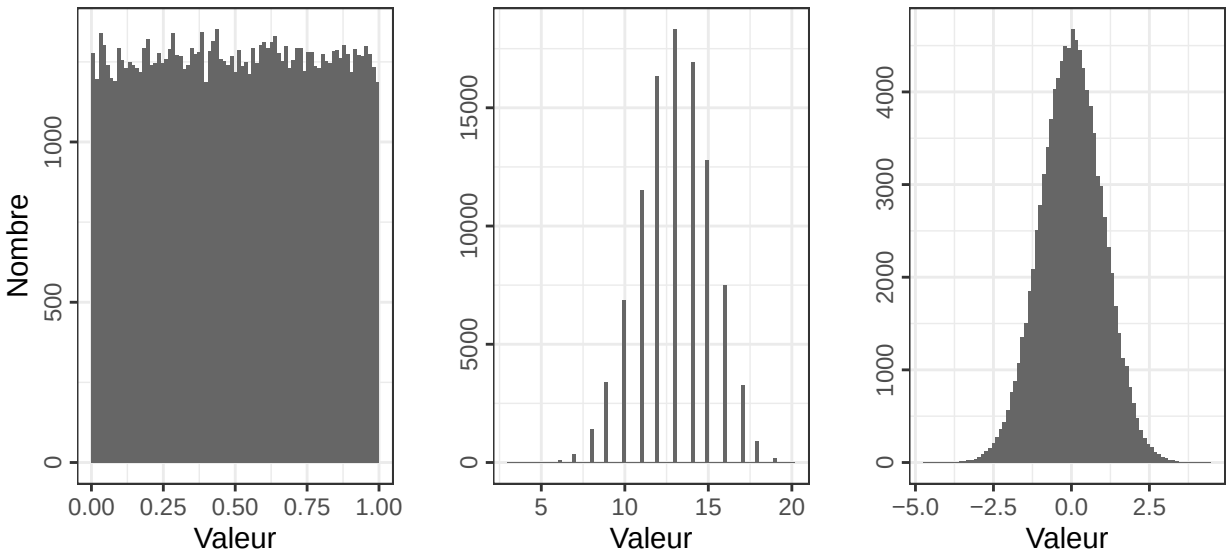

Figure 2: Exemples de distributions de la loi uniforme (à gauche), binomiale (au centre) et normale (à droite)

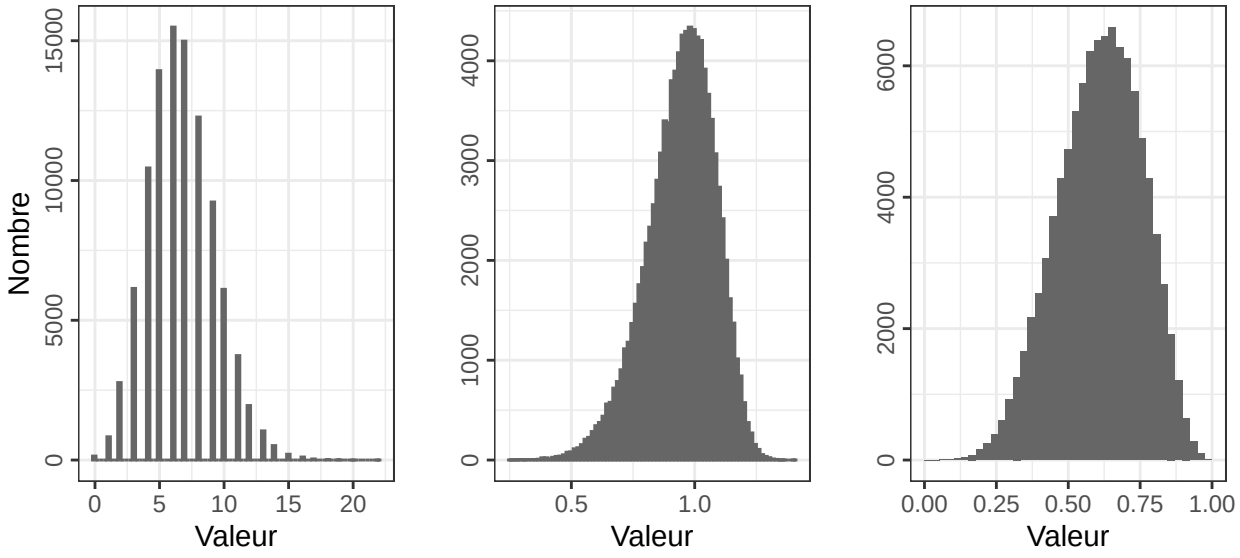

Figure 3: Exemples de distributions de la loi de Poisson (à gauche), de Weibull (au centre) et bêta (à droite)

En raison des contraintes imposées précédemment ainsi que de par sa grande polyvalence[26], 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 :[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 :

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

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.

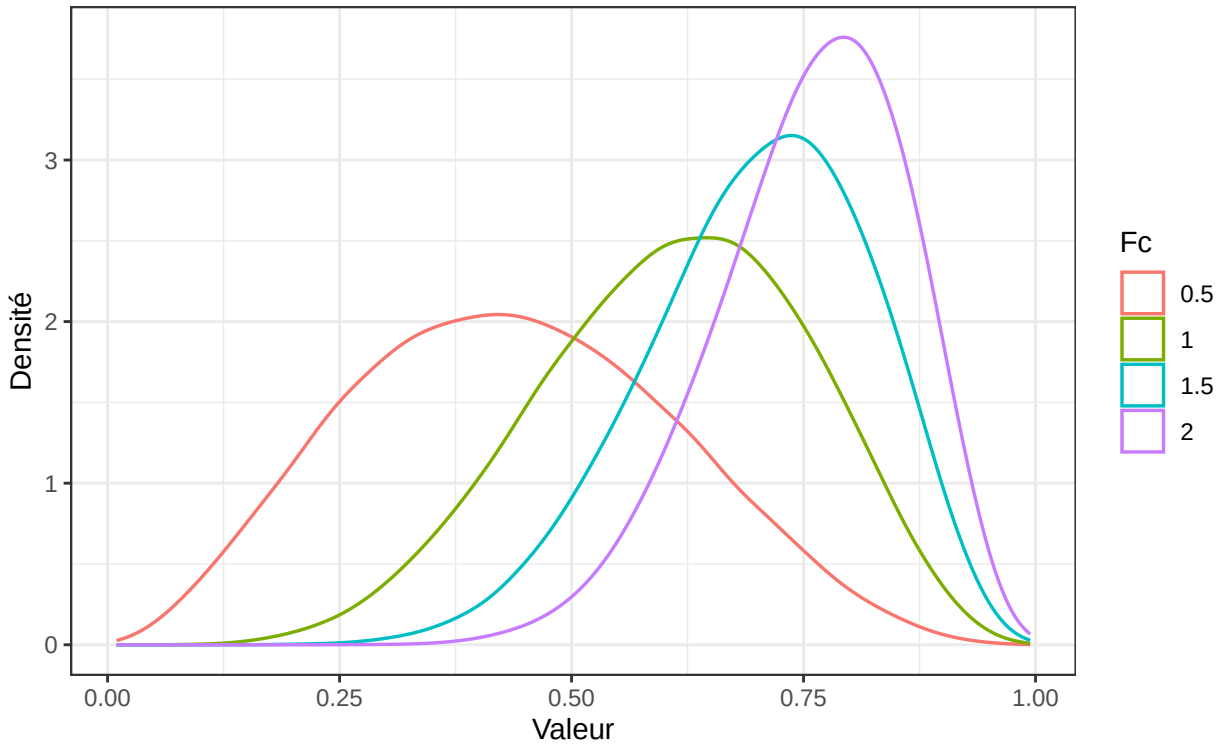

Figure 4: Distribution de différentes lois bêta, en fonction du facteur de croissance Fc

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 :

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

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.

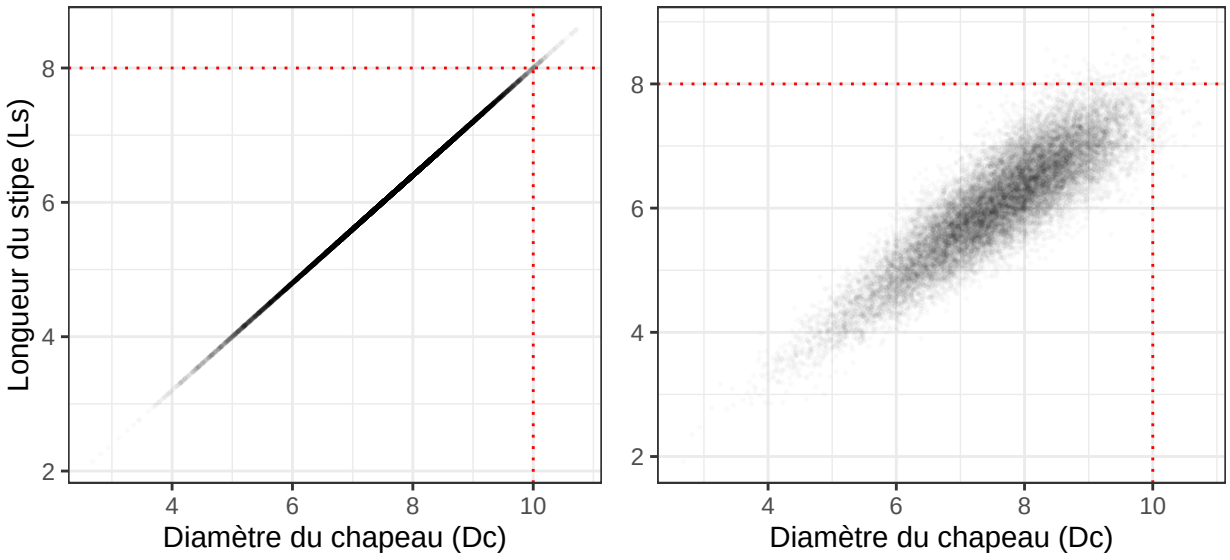

Figure 5: Nuages de points de 2 paramètres de taille, sans dispersion (à gauche) et avec dispersion (à droite)

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 0.444 % (cf. figure 7), et la proportion de spécimens dépassant de plus de 10% cette valeur maximale est inférieure à 50 %.

3.2.3 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.

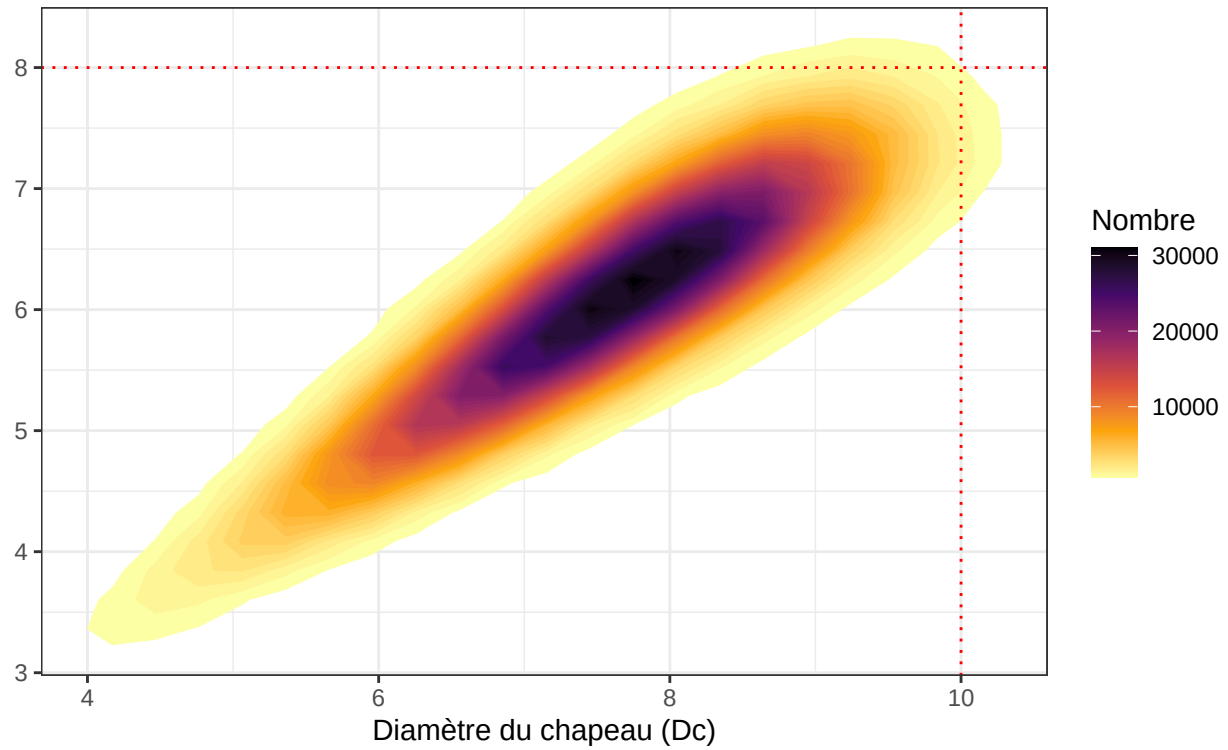

Figure 6: Diagramme de densité de 2 paramètres de taille, avec dispersion

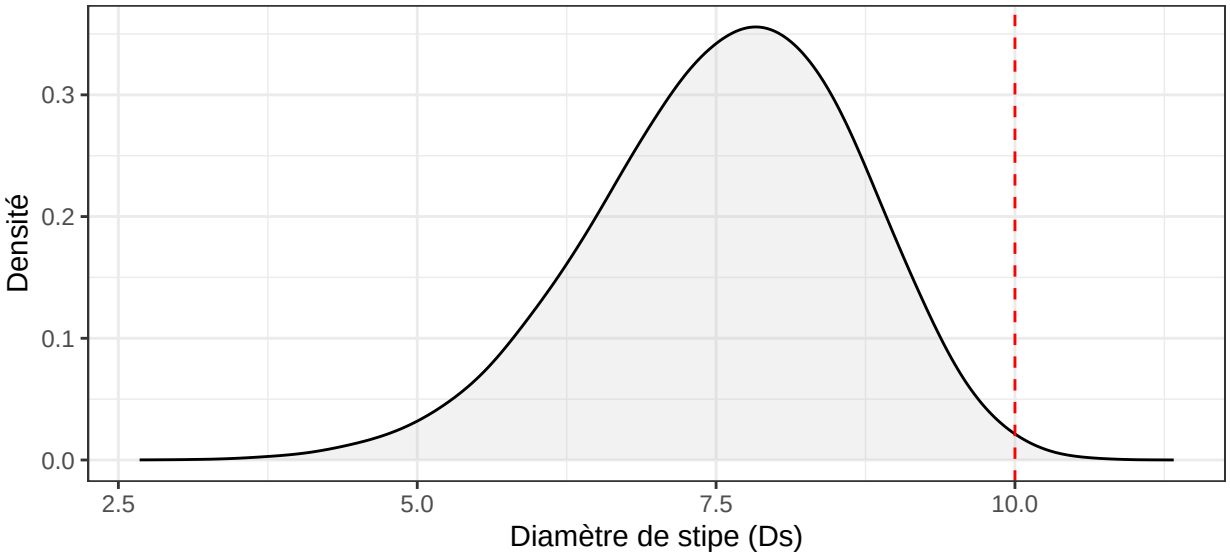

Figure 7: Distribution du diamètre de stipe Ds, pour Dsmax = 10

4 Principes de l'apprentissage machine

4.1 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 3.2.1) 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 intial, 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, dans la pratique, être obtenu par réalisation de deux découpages successifs suivant des rapports 92:8.

Dans cette étude, la division de ces trois jeux de données utilise une méthode de découpage basée sur les points-supports[30] (*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]

4.2 Modèles utilisés

4.2.1 Analyses discriminantes (lda2, pda)

[discriminant correspondence analysis???]
https://rpubs.com/markloessi/505575
https://en.wikipedia.org/wiki/Altman_Z-score
https://towardsdatascience.com/linear-discriminant-analysis-explained-f88be6c1e00b
https://scikit-learn.org/stable/modules/lda_qda.html

Cette étude propose deux classifieurs linéaires s'appuyant sur des méthodes d'analyse discriminante : un modèle basé sur l'analyse discriminante linéaire (*Linear Discriminant Analysis*, LDA) et un modèle basé sur l'analyse discriminante pénalisée (*Penalized Discriminant Analysis*, PDA).

L'analyse linéaire discriminante (LDA) est une méthode utilisée en statistiques et en datascience pour trouver une combinaison linéaire d'éléments qui caractérisent des éléments, afin de créer un classifieur linéaire, ou d'effectuer des réductions de dimensionnalité. Cet algorithme fonctionne en créant des combinaisons linéaires (fonctions discriminantes) de prédicteurs. [A FINIR]

L'analyse discriminante pénalisée [A FINIR]

4.2.2 Modèle additif généralisé

gamLoess

4.2.3 Arbres de décision

rpart, Ctree, c50tree, Rpartcost

4.2.4 Forêts aléatoires

rFerns, Rborist, ranger

5 Apprentissage machine et classification binaire

A TRADUIRE !!!

5.1 Analyse exploratoire des données (EDA)

Table 1: Dataset structure

	Type	Levels
class	factor	2
cap.diameter	numeric	2571
cap.shape	factor	7
cap.surface	factor	12
cap.color	factor	12
does.bruise.or.bleed	factor	2
gill.attachment	factor	8
gill.spacing	factor	4
gill.color	factor	12
stem.height	numeric	2226
stem.width	numeric	4630
stem.root	factor	6
stem.surface	factor	9
stem.color	factor	13
veil.type	factor	2
veil.color	factor	7
has.ring	factor	2
ring.type	factor	9
spore.print.color	factor	8
habitat	factor	8
season	factor	4

La structure du lot de données est la suivante :

5.2 Training set analysis

The original dataset was first randomly split into a 90%-sized training/validation set, and a 10%-sized evaluation set that is not to be used until the final validation. All distributions were then plotted using a simple loop that gets the column names and a conditional branching that uses the structure to plot an histogram or a barplot (simplified code).

```
for (n in 1:ncol(trainingset)){
   plot <- trainingset %>%
      ggplot(aes_string(x = colnames(trainingset)[n]))
   if(structure_dataset$Final[n] %in% c("integer", "numeric"))
      {plot <- plot + geom_histogram(fill = "gray45")}
   else
      {plot <- plot + geom_bar(fill = "gray45")}
```

```
    plotname <- paste0("distrib_", colnames(trainingset)[n])
    assign(plotname, plot)
}
```

Using *aes_string* instead of the traditional *aes* allows to use a string as an argument.

Barplots didn't show anything remarkable, and thus weren't integrated in the report. However, dimensional distributions were more interesting : at first, they seemed to roughly follow a bell curve (fig. 1), with a long tail towards the higher values. A logarithmic transformation (fig. 2) can show more clearly the shape of this tail.

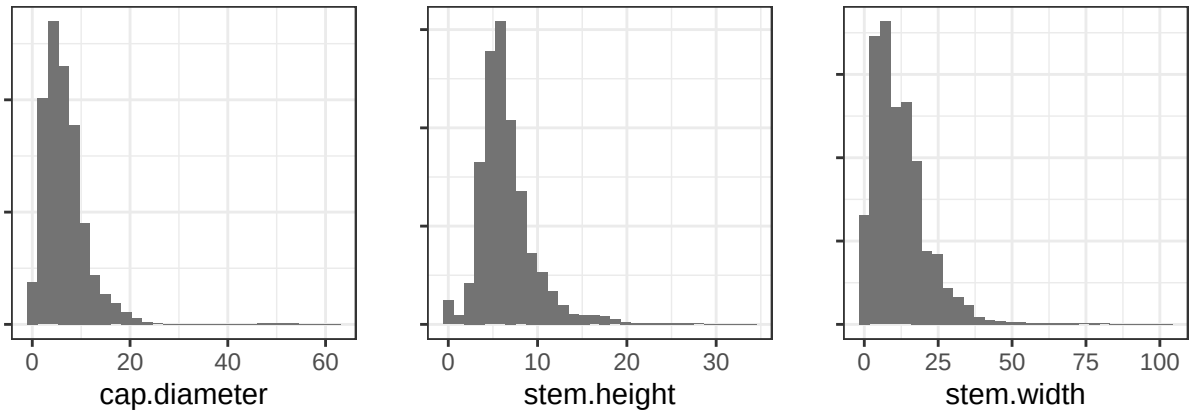

Figure 8: Mushroom cap diameter, stem height and stem width distribution

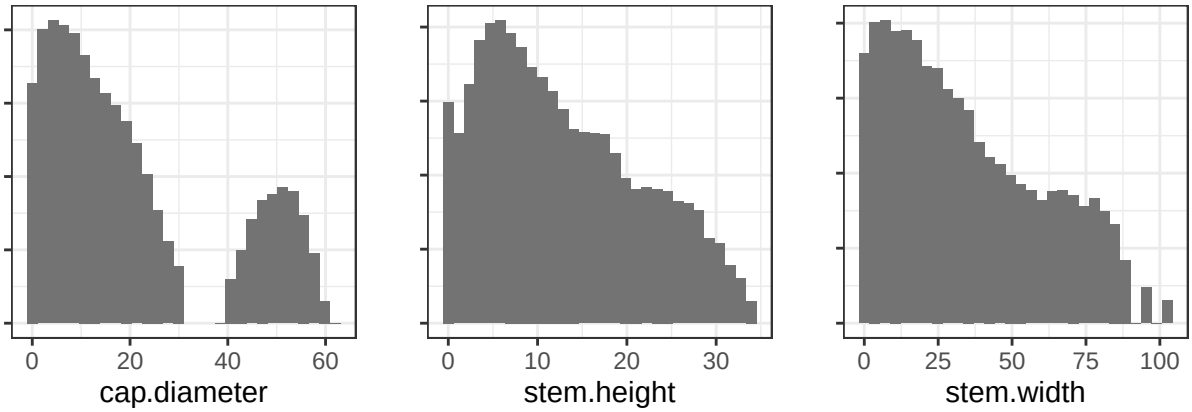

Figure 9: Mushroom cap diameter, stem height and stem width distribution (log Y scale)

The cap diameter distribution looks like a bell curve with a long right tail but actually is bimodal, with a main mode toward 5 cm, and a much smaller secondary mode toward 50 cm. This size can look very surprising, but after further investigation, it appears that some species such as *Polyporus squamosus* (dryad's saddle) can be very large, with some specimens that can weight up to 5 kg.[x]

The stem height distribution also looks like a bell curve with a long right tail, a main mode around 5 cm, and a secondary mode at 0 cm. Again, this can look surprising, but some mushrooms have no stem, which could explain this height value.

The stem width distribution looks like a bell curve with a long right tail, and a peak around 10-15 mm. In all three distributions, the right tail can probably be explained by the impossibility to have negative dimensional values.

5.3 Caret package models

The *caret* packages provides a very convenient and efficient platform for data modelling and inference. This section will explain the strategy used for the evaluation of some of these models. The selected models were of various types :

- Linear Discriminant Analysis : Linear Discriminant Analysis (lda2), Penalized Discriminant Analysis (pda)
- Generalized Additive Model : Generalized Additive Model using LOESS (gamLoess)
- Tree-Based Models : Classification And Regression Tree (CART) (rpart, rpartCost), Single C5.0 Tree (ctree)
- Random Forest : Random Ferns (rferns), Random Forest (ranger, Rborist)

The first step was to build a regression and evaluation function. The caret package allows to define it very simply with :

```
set.seed(1)
tr_ctrl <- trainControl(classProbs = TRUE,
                        summaryFunction = twoClassSummary,
                        method = "cv", number = 10)
train(class ~ .,
      method = [METHOD],
      data = trainvalid_set,
      trControl = tr_ctrl,
      metric = 'Spec',
      tuneGrid = data.frame([PARAMETERS]))
```

The *set.seed* insures reproducibility, the *trainControl* function allows to control various training and evaluation parameters ; in this case, to use a ROC/sensitivity/specificity criterion and a 10-fold cross-validation. The *train* function runs the train and evaluation process, while allowing the use of several parameters, such as the method selection the training/validation set, the metric used for the evaluation and the grid of model parameters.

This block was included in a function to allow easy access and reproducibility. The function returns a list that can be plotted and also includes various data of interest, such as .$results (ROC, Sensibility, Specificity), .$bestTune (best value with the evaluation metric, Specificity in this case), and .$finalModel (miscellanous information that can sometimes be plotted, such as trees).

5.3.1 Linear Discriminant Analysis Models

The two Linear Discriminant Analysis Models selected for this study are lda2 (Linear Discriminant Analysis) and pda (Penalized Discriminant Analysis). The lda2 model has one tuning parameter (*dimen*, number of discriminant functions). The pda model also has a single tuning parameter (*lambda*, shrinkage penalty coefficient).

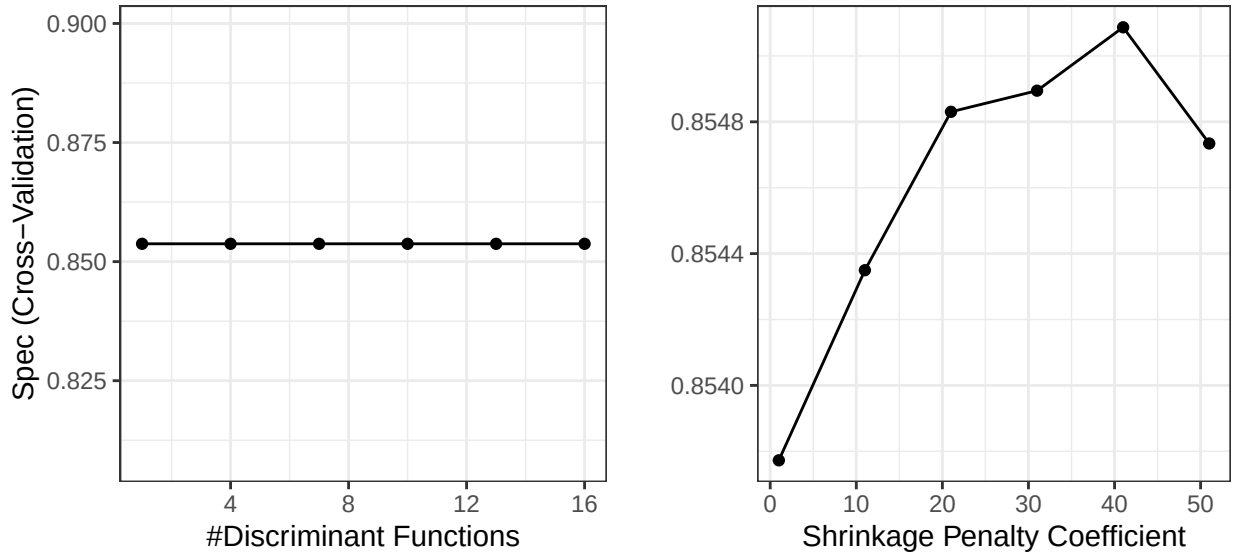

Figure 10: Specificity of lda2 (left) and pda (right) models

The *dimen* parameter of the lda2 model does not seem to have much effect on specificity (*Spec* $=$ 0.854).

The *lambda* parameter of the sda model marginally affects its specificity, with lower lambda values giving slightly better results ($Spec_{max} = 1$).

However, the specificity of both models are far from being sufficient for this study ($Spec \approx$ 0.85), and their sensitivities don't seem to be much higher than the basic 2-criteria model ($Sens \approx$ 0.842 vs xxx).

5.3.2 Generalized Additive Model

The only generalized additive model selected for this study is gamLoess (Generalized Additive Model using Locally Weighted Linear Regression). The caret package documentation indicates that gamLoess has two tuning parameters : *span* (fraction of data points used in the local neighborhood size) and *degree* (degree of linearization).

The *degree* parameter of the gamLoess model does not seem to have much effect on specificity, with $\Delta Spec = 0.018$.

The *span* parameter marginally affects the specificity of the gamLoess model, with an optimal value of *span* $= 0.01$, that gives $Spec_{max} = 0.84$ vs xxx).

The specificity this model does not meet the required specificity level for this study. The GamLoess model actually proved to be inferior to the 2-criteria classifier in both specificity (0.84 vs xxx and sensitivity (0.771 vs xxx).

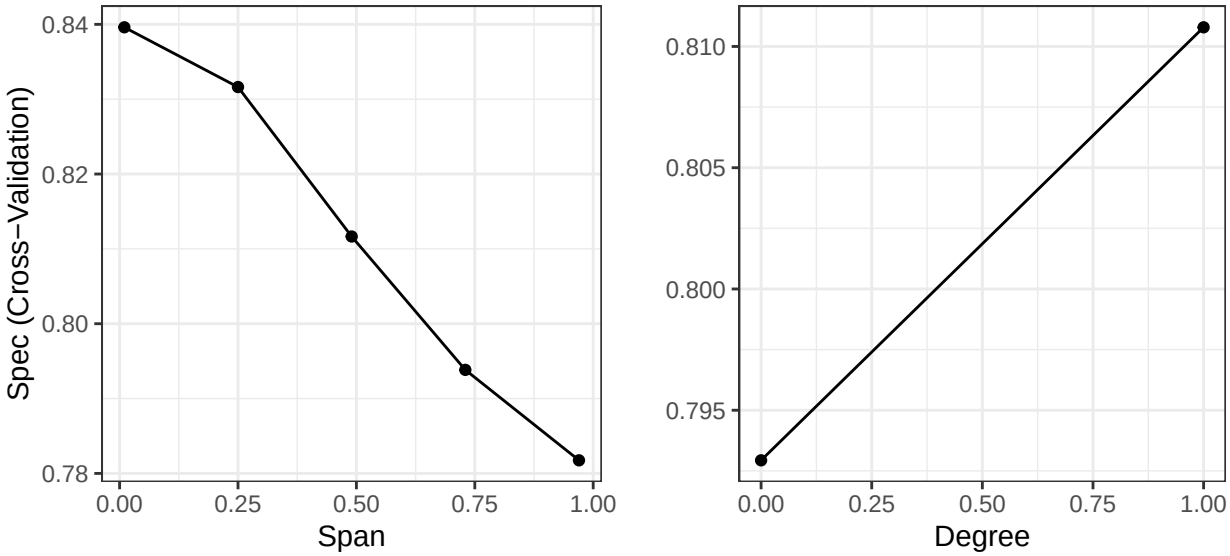

Figure 11: Specificity of lda2 (left) and pda (right) models

Table 2: Performance of the CART (rpart) model

cp	ROC	Sens	Spec
1e-05	0.99820	0.99516	0.99619
1e-04	0.99821	0.99504	0.99606
1e-03	0.99573	0.98555	0.99099
1e-02	0.91057	0.81416	0.91201
5e-02	0.63751	0.75136	0.51526

5.3.3 Tree-Based Models

Tree-based models are of special interest in this study, for two main reasons :

- Tree-based logic is usually used in manual mushrooms classification,
- Tree models can be plotted and easily understood by humans.

The first models presented in this study are two CART (Classification And Regression Tree) models. The basic CART model (rpart) has one complexity parameter (cp).
The basic CART model does never achieve the required specificity. However, this model still comes close and gives excellent results ($Sens_{max} = 0.995$, $Spec_{max} = 0.996$).
The second CART model used in this study (rpartCost) associates a complexity (*cp*) and a cost (*Cost*) parameters.
The best predicted specificity would be achieved with *cp* = 1e-05 and *Cost* = 0.01.
The performance, while very good, did not achieve the specificity requirement.

The last tree-based model was C5.0tree. This model doesn't have any tuning parameter.

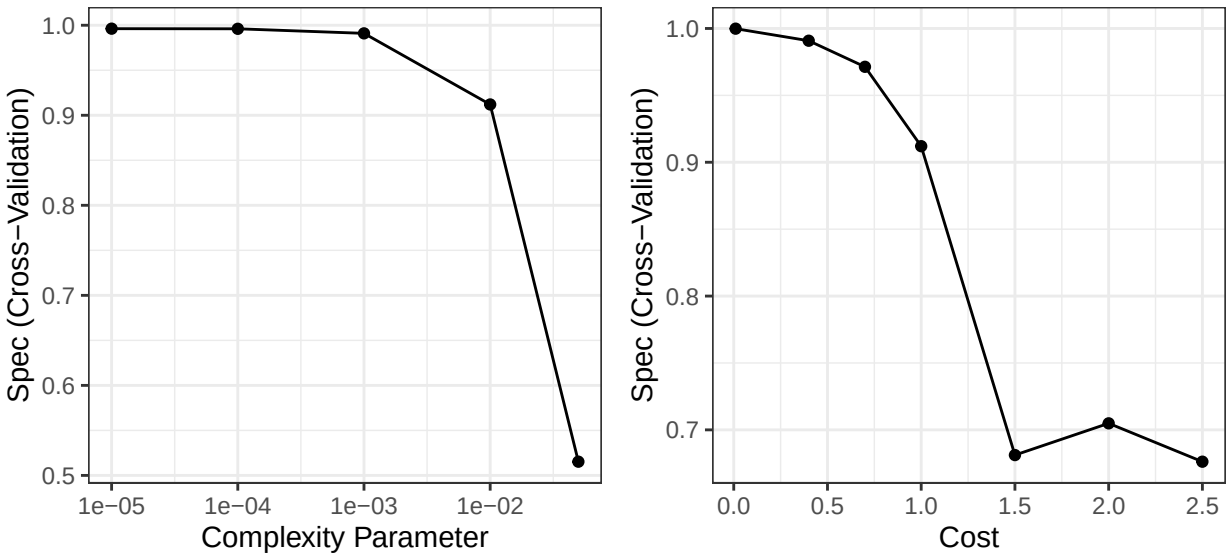

Figure 12: Specificity of rpartCost according to complexity and cost values

Table 3: Performance of the CART (rpartCost) model

cp	Cost	Sens	Spec
1e-05	0.01	0.89569	0.9991

Table 4: Performance of the C5.0 Tree (C5.0Tree) model

ROC	Sens	Spec
0.99858	0.9972	0.99747

Quite interestingly, despite have no tuning parameter, this model gave excellent and very balanced results out of the box, with both very high sensitivity and specificity. Still, this model didn't achieve the required $Spec = 1$ criterion.

5.3.4 Random Forest Models

The Random Ferns (rFerns) model has only one parameter : depth.

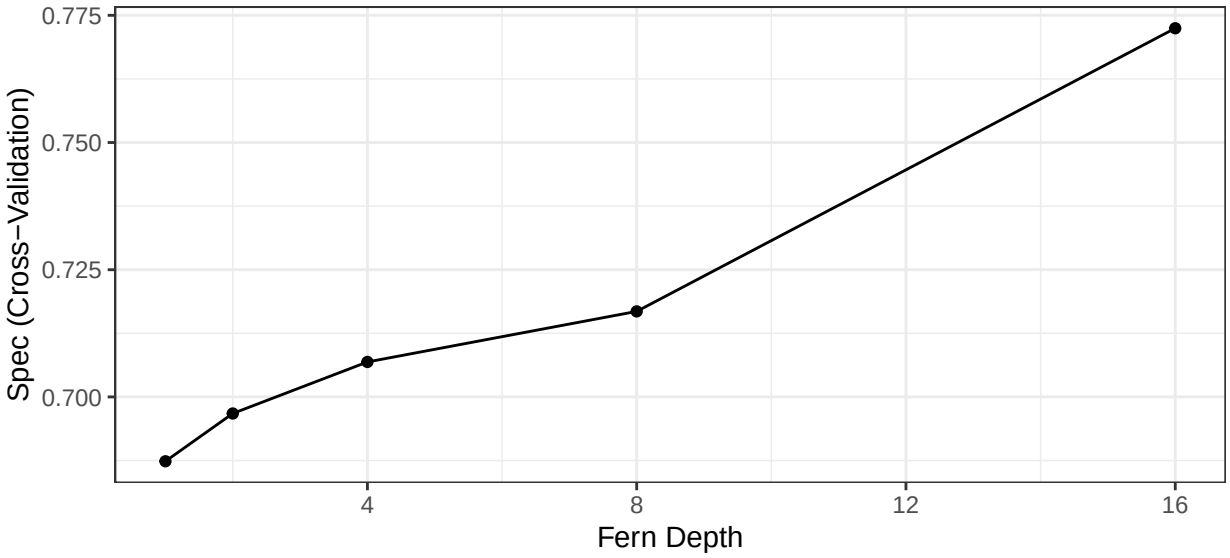

Figure 13: Specificity of Random Ferns model

While pretty fast, the Random Ferns model yielded disappointing results, with a maximum specificity of $Spec_{max} = 0.772$. The maximum sensitivity wasn't very high either ($Sens_{max} = 0.795$). Both metrics were inferior to the basic 2-criteria model ones.

The second random forest model is the ranger model. The caret package documentation mentions three tuning parameters : the minimal node size (*min.node.size*), the number of features to split on each node (*mtry*) and the splitting rule (*splitrule*).

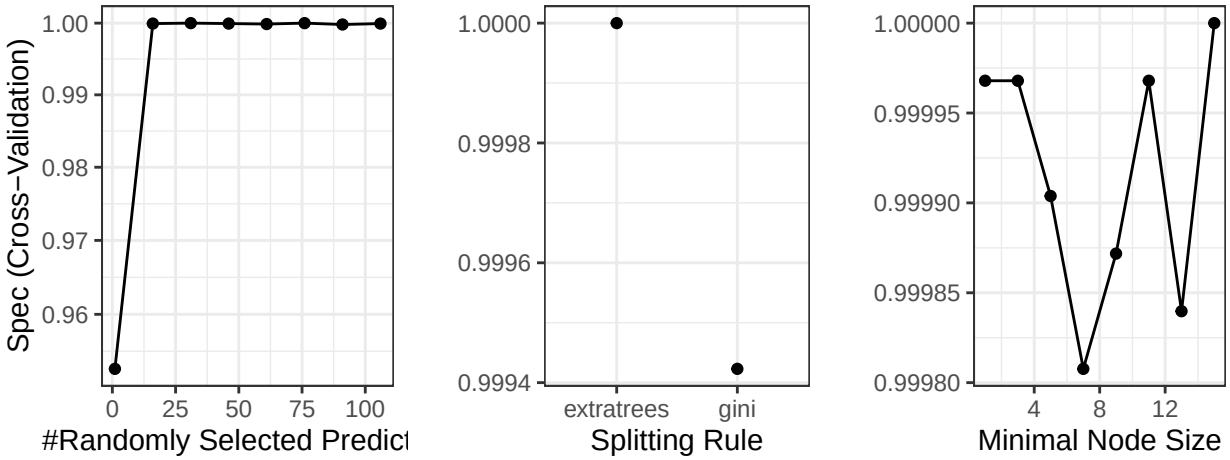

Figure 14: Ranger model specificity

The preliminary tuning step showed very promising results : even with a low number of trees ($n = 6$), the required specificity was already achieved on several occurrences with single-parameter

Table 5: Performance of the Ranger model (optimal settings)

min.node.size	splitrule	mtry	ROC	Sens	Spec
15	extratrees	50	0.9999995	0.9997999	0.9997756

Table 6: Performance of the Rborist model (predFixed tuning)

predFixed	minNode	ROC	Sens	Spec
1	2	0.8856	0.0000	1.0000
11	2	1.0000	1.0000	1.0000
21	2	1.0000	0.9999	0.9999
31	2	1.0000	0.9998	1.0000
41	2	1.0000	0.9997	1.0000

tuning. Tuning this model with optimal parameters (*min.node.size* = 15, *mtry* = 50 and *splitrule* = extratrees) gave interesting results.
This model achieved excellent performance, giving perfect specificity and sensitivity in this evaluation phase.

The last random forest model was provided by Rborist. The caret package mentions two tuning parameters for this model : number of trial predictors for a split (*predFixed*) and minimum number of distinct row references to split a node (*minNode*).

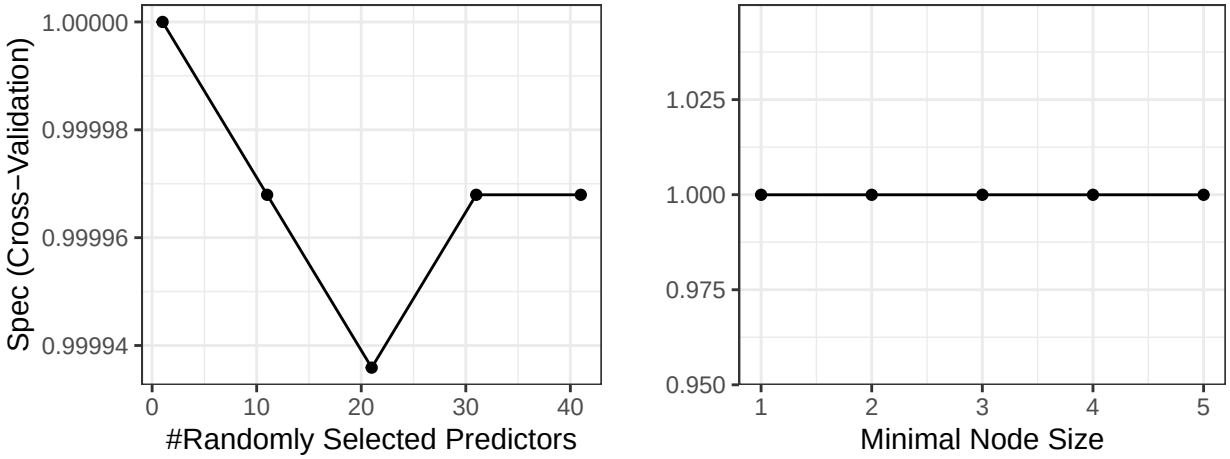

Figure 15: Rborist model specificity

Again, the preliminary tuning step showed promising results : on this model too, the required specificity was achieved on several occurrences using only single-parameter tuning. With optimal parameters (*predFixed* = 6, manually fixed value because of the higher sensitivity), and *minNode* = 1, the performance was estimated to to be :
The Rborist model gave the same excellent results as the Ranger one, with perfect sensitivity and specificity.

Table 7: Performance of the Rborist model (optimal settings)

predFixed	minNode	ROC	Sens	Spec
6	1	1	1	1

These results underline an interesting fact : not all random forest models are created equal. This study shows the considerable difference in sensitivity and specificity between the rFerns and the Ranger or Rborist random forest models. In the first step of the preliminary studies of the caret package models, some random forest models also proved to be extremely slow, while others were considerably faster.

Speed of some algorithms will be explored in the next section. Running time of any code portion can be very easily measured by:

```
start_time <- Sys.time()
[Code to be evaluated]
end_time <- Sys.time()
time <- difftime(end_time, start_time)
```

5.4 Memory Optimization

During the building of this study, an unforeseen and unfortunate event resulted into the addition of a secondary goal : the code had to be able to handle a 61069×21 dataset and run on lower-end computers, with only 4GB RAM + 6GB swap. Memory optimization was an interesting challenge, and implied :

- Image saving on hard drive before environment cleaning,
- Identification and removal of obsolete intermediate values,
- Identification and removal of objects that won't be used in the final report,
- Identification and replacement of large objects,
- Periodic garbage collection.

The identification of large objects can be performed by the following code :

```
object_list <- objects()

obj_size <- function(fcn_object){
   object <- eval(parse(text = fcn_object))
   size <- format(object.size(object), units = "Mb")
   size <- str_remove(size, " Mb")
   size <- as.numeric(size)
   size
```

```
}
size_list <- sapply(X = object_list, FUN = obj_size)
```

The resulting vector can then be converted to a data frame, that returns the size of all objects : xxx
Some of these objects are quite large, but can fortunately be converted into smaller and still useful objects.

For example, the *very* large fit_rFerns_depth train object (xxx Mb) can be split into two useful objects that gather the main metrics of interest : the fitting plot (xxx) and the results (xxx) table.

Some of these objects were much smaller : ggplots are not heavy per se, but while generating plots for all variables proved to be useful at some point in the study, having more than forty 5+ MB plots in memory was not really necessary, especially when having to run training and validation steps on limited hardware.

Periodic data-gathering and object deletion thus permitted to avoid unnecessary memory creep that resulted in major slowdowns or crashes.

6 Results

6.1 Evaluation protocol

The models that attained the specificity requirement during the validation process were selected for the final evaluation. The three selected models are :

- Two-criteria classifier,
- Ranger random forest,
- Rborist random forest.

All models were trained on the training dataset, set with the best hyperparameters values obtained by evaluating the performance against the validation dataset. Their performance against the evaluation dataset will be analyzed, using the same criteria as before :

1. Specificity *must* be equal to one.
2. Sensitivity should be the highest possible.

6.2 Dual criteria classifier performance

Running the dual criteria classifier against the evaluation dataset yields the following results : xxx

The confusion matrix shows more accurately the results : while xxx of the yyy edible specimen were accurately identified, zzz non-edible specimen were incorrectly classified as edible.
xxx
While the model performance is quite honorable, it is not sufficient to completely fulfill the specificity criterion, which was the primary endpoint of this study. A performance decrease between the validation and evaluation stage can typically be attributed to overfitting. It is thus important to find if the performance difference is significant.

xxx The performance difference does not seem to be significant, and the lower specificity just seems to be a result of a trade-off between sensitivity and specificity. The absence of significant overfitting can also be confirmed by the slight increase of the F1 score during the validation stage.

6.3 Random forest performance

The final evaluation of the Ranger gave the following confusion matrix.
The final accuracy was equal to 1, with a 95% confidence interval of [0.9992 ; 1]. The ranger model provided excellent results, in a very reasonable amount of time (26.02 min).

The Rborist model gave similar results with a final accuracy equal to 1, with a 95% confidence interval of [0.9992 ; 1]. The Rborist model, while giving comparable results to the Ranger one, was sensibly slower (3.33 min), despite being run with a lower number of trees ($n = 3$ vs 10).

Table 8: Confusion Matrix of the Ranger model (vs. evaluation)

	e	p
e	2023	0
p	0	2473

Table 9: Ranger and Rborist models performance (vs. evaluation)

	Sensitivity	Specificity	F1 score	Run time (min)
Ranger	1	1	1	26.02
Rborist	1	1	1	3.33

7 Conclusion

Species identification is a classic classification task, traditionally performed by humans using a classification tree strategy.

This study showed that given particular selection conditions imposed by circumstances (toxic mushrooms shouldn't be classified as edible, i.e. $Specificity = 1$), a significant amount of modelling strategies were giving insufficient results, and proved to be inferior to even a quite basic *ad-hoc* bi-criteria classification model.

This bi-criteria model, while already quite efficient (xxx) did not fully achieve the specificity requirement against the evaluation dataset.

The best classification tools were random forest models : ranger and Rborist, which both gave perfect sensitivity and specificity values, but with ranger being about 0 times faster. The rFerns model, while also being based on random forests, was much less accurate.

Apart from the limitations of the starting dataset, this work has some notable limitations :

- Memory optimization could probably be improved, for example with use of local environments,
- The basic classifier code could definitely be optimized and make more use of vectorization,
- Parameter optimization was mostly based on the caret package documentation, which doesn't mention all parameters,
- Parallelization could be an interesting strategy to make some computations much faster,
- Reproductibility could be improved between Linux/Windows systems on R 3.6/4.1, especially on the layout of the report output.

This study will provide a good basis for further personal work and experimentation on these three aspects.

Working on this dataset also gave me nice future project ideas, such as the creation of a future dataset based on the same concept as the two original datasets from Schlimmer[x] and Wagner[y], but with more species and more extensive criteria (such as the smell or the flesh texture) taken from more comprehensive and specialized books.[z]

8 Apprentissage machine et classification multiclasse

texte

9 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]. 2022. 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]. 2023. Disponible sur: http://www.stats.ox.ac.uk/pub/MASS4/
7. Kuhn M. caret: Classification and Regression Training [Internet]. 2022. 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]. 2022. Disponible sur: https://CRAN.R-project.org/package=mda
10. Therneau T, Atkinson B. rpart: Recursive Partitioning and Regression Trees [Internet]. 2022. Disponible sur: https://CRAN.R-project.org/package=rpart
11. Wickham H. plyr: Tools for Splitting, Applying and Combining Data [Internet]. 2022. Disponible sur: https://CRAN.R-project.org/package=plyr
12. Kuhn M, Quinlan R. C50: C5.0 Decision Trees and Rule-Based Models [Internet]. 2023. Disponible sur: https://topepo.github.io/C5.0/
13. Hothorn T, Hornik K, Strobl C, Zeileis A. party: A Laboratory for Recursive Partytioning [Internet]. 2022. Disponible sur: http://party.R-forge.R-project.org
14. Wright MN, Wager S, Probst P. ranger: A Fast Implementation of Random Forests [Internet]. 2022. 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]. 2023. 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]. 2022. 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]. 2023. Disponible sur: https://CRAN.R-project.org/package=rmarkdown

19. Xie Y. knitr: A General-Purpose Package for Dynamic Report Generation in R [Internet]. 2022. Disponible sur: https://yihui.org/knitr/

20. Kassambara A. ggpubr: ggplot2 Based Publication Ready Plots [Internet]. 2023. 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