الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique Et Populaire
وزارة التعليم العالي والبحث العلمي
Ministère De L'enseignement Supérieur Et De La Recherche Scientifique

Université Constantine 1 Frères Mentouri
Faculté des Sciences de la Nature et de la Vie

جامعة قسنطينة 1 الإخوة منتوري
كلية علوم الطبيعة والحياة

Département : microbiologie

قسم : ميكروبيولوجيا

Mémoire présenté en vue de l'obtention du Diplôme de Master

Domaine : Sciences de la Nature et de la Vie

Filière : Sciences Biologiques

Spécialité : microbiologie appliquée

N° d'ordre :
N° de série :

Intitulé :

Développement d'un modèle de machine learning (apprentissage automatique) pour la classification automatique des organismes microscopiques

---

Présenté par : CHIKHI Rania
ZITOUNI Ikram

Le : 12/06/2024

Jury d'évaluation :

Président :	**Abdelaziz Ouidad** (MCB- U Constantine 1 Frères Mentouri).
Encadrant :	**Djama Ouahiba** (MCB - U Constantine 1 Frères Mentouri).
Examinateur(s):	**Chabbi Rabah** (MAA - U Constantine 1 Frères Mentouri).

Année universitaire
2023 - 2024

Remerciement

*Tout d'abord, nous remercions **ALLAH** le tout-puissant de nous avoir guidé tout au long de ce parcours académique et de nous avoir donné la force et la motivation nécessaire pour mener à bien se travail.*

C'est avec un grand plaisir que nous réservons ces lignes en signe de gratitude et de reconnaissance à ceux qui ont contribuée de près ou de loin à l'élaboration de ce travail

Au terme de ce modeste travail, nous tenons a infiniment et avec gratitude notre encadreur Mme Djamaa Ouahiba, qui nous a fourni de précieux conseils et orientations, et pour ses encouragement et son soutien constants a tout moment, a qui nous exprimons notre gratitude et notre appréciation.

A Nour El Houda boubakar merci pour votre accompagnement et votre aides, nos vif remerciements.

Nous remercions tout particulièrement le professeur M. Chabbi Rabah qui a accepté d'examiner et de discuter le présent travail.

Merci au professeur Mme Abdelaziz Ouidad pour avoir accepté de présider le jury de ce modeste travail.

Nous tenons également à remercier tous les étudiants de notre promotion (2023 2024) et particulièrement notre spécialité.

Un grande merci à tous nos enseignants de la faculté des sciences de la nature et de la vie et particulièrement non enseignants de spécialité, Microbiologie Appliquée

Enfin, nos remerciements vont à toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce modeste travail.

Dédicace

Je dédie ce travail

A mon cher père ***Belkacem Djamel*** *pour ses sacrifices et ses encouragements*

A Mes adorables mères : ***Aicha*** *et* ***Sarhouda*** *pour ses soutiens et ses encouragements*

A Mes belles sœurs ***Khawla***, ***Ikhlass*** *et* ***Noussaiba***

A toute ma famille ***ZITOUNI*** *et* ***BOUZIENE***

A tous mes cousines et mes proches

A mes amis et camarades de promotion

A mes belles amies ***INES, Lina, Amira, Ahlem***, ***noussaiba***

A mon binôme Rania

IKRAM

Dédicace

À la mémoire de ma chère grand-mère

Bien que vous ne soyez plus parmi nous, votre amour et votre sagesse continuent de m'accompagner chaque jour. Votre souvenir est une source de force et d'inspiration. Vous resterez à jamais dans mon cœur.

À ma chère maman.

Grâce à votre amour sans conditions et à votre soutien infaillible, j'ai réussi à surmonter tous les défis. Je vous remercie pour toutes vos actions en ma faveur.

À mon cher Père.

J'ai toujours été motivé par votre soutien et vos encouragements pour donner le meilleur de moi-même. Je vous remercie de votre amour et de votre confiance.

À mes chères sœurs, Khouloud et Malak

Votre présence et vos encouragements m'ont toujours motivé à aller de l'avant. Vous êtes mes premières amies et mon soutien infaillible.

À mon fiancé, Redha

Ton amour et ton soutien ont été très précieux pour moi tout au long de cette expérience. Je vous remercie de me croire et de m'encourager quotidiennement.

À tous mes Amis

Merci pour votre amitié sincère et votre soutien tout au long de ce parcours. Vos encouragements et votre compréhension ont été d'une grande aide.

À tous mes professeurs respectés tout au long de mon parcours universitaire

Avec toute ma gratitude.

Chkhi Rania

Résumé

Ce travail consiste à développer des modèles d'apprentissage automatique pour classifier des genres d'organismes microscopiques, en se basant sur les caractéristiques morphologiques extraites à partir des dimensions des formes des microorganismes sur des images. L'objectif principal est de développer un modèle permet de classer les microorganismes avec précision. Les étapes principales pour le développement de ce modèle comprennent la collecte et le traitement des données (dataset), l'exploitation de deux algorithmes d'apprentissage automatique (l'algorithme de forêt aléatoire et l'algorithme de l'arbre de décision), ainsi que l'entraînement et l'évaluation de ce modèle. La performance de notre modèle est évaluée en utilisant plusieurs mesures telles que la précision. Les résultats ont montré que le modèle proposé avec ses algorithmes était capable de classer avec précision les organismes microbiens, avec un taux d'apprentissage supérieur à 98 %.

Mots clé : Microorganisme, Apprentissage Automatique, Foret Aléatoire, Arbre De Décision.

ملخص

يهدف هذا العمل إلى تطوير نماذج التعلم الآلي لتصنيف أنواع من الكائنات الدقيقة بناءً على الخصائص المورفولوجية المستخرجة من أبعاد أشكال الكائنات الحية الدقيقة على الصور (السمات). الهدف الرئيسي هو تطوير نموذج لتصنيف الكائنات الحية الدقيقة بدقة عالية. تشمل الخطوات الرئيسية لتطوير هذا النموذج جمع البيانات ومعالجتها، استغلال خوارزميتين للتعلم الآلي (خوارزمية الغابة العشوائية وخوارزمية شجرة القرار)، بالإضافة إلى تدريب وتقييم هذه النماذج. يتم تقييم أداء هذه الاخيرة باستخدام عدة مقاييس مثل الدقة. اظهرت النتائج ان النموذج المقترح مع خوارزميتيه كان قادرا على تصنيف هذه الكائنات الحية الدقيقة بدقة عالية جاوزت 98 %.

الكلمات المفتاحية : الكائن الحي الدقيق، التعلم الالي، الغابة العشوائية، شجرة القرار.

ABSTRACT

This work aims to develop machine learning models to classify genera of microscopic organisms, based on morphological characteristics extracted from the dimensions of the shapes of microorganisms on images. The main objective is to develop a model to classify microorganisms accurately. The main steps for the development of this model include the collection and processing of data (dataset), the exploitation of two machine learning algorithms (the random forest algorithm and the decision tree algorithm), as well as the training and evaluation of this model. The performance of our model is evaluated using several metrics such as accuracy. The results showed that the proposed model with its algorithms was able to accurately classify microbial organisms, with a learning rate above 98%.

Key words: Microorganism, Machine Learning, Random Forest, Decision Tree.

Listes des figures

Listes de tableaux

Liste des abréviations

MO : Microorganismes

CT : Classification Taxinomique

AI : Intelligence Artificiel

ML : Machine Learning (apprentissage automatique)

DL: Deep Learning (apprentissage profond)

SSL : semi supervised learning

SVR : Support vecteur régression

K-NN: K Nearest Neighbors

K-MEANS: K-moyenne

DT : Decision Tree (arbre de décision)

RF : Random Forest (foret d'arbre décisionnels)

SVM : Support vecteurs machines

RN : Réseau de Neurones

COLAB : Google Colaboratory (service de cloudputing basé sur Jupyter Notebook)

Table des matières

Partie pratique

Chapitre 01 : Matériel et Méthodes

Chapitre 02 : résultats et discussion

INTRODUCTION GENERALE

En microbiologie, la classification des microorganismes est essentielle pour diverses raisons. Elle facilite l'organisation et la structuration de la grande diversité de ces organismes, ce qui facilite leur étude et leur compréhension. En repérant et en classant les micro-organismes en fonction de leurs caractéristiques communes, les chercheurs peuvent approfondir leur compréhension des liens évolutifs, de leurs fonctions écologiques et de leurs conséquences en médecine et en biotechnologie. Les techniques de classification des micro-organismes sont multiples et évoluent en permanence en fonction des progrès technologiques, allant des méthodes morphologiques classiques aux approches moléculaires avancées.

Cependant, il existe plusieurs désavantages aux méthodes classiques traditionnelles de classification des micro-organismes. Le processus est difficile et parfois impraticable en raison de sa complexité et de son coût élevé. Par ailleurs, ces approches peuvent être très subjectives, ce qui peut conduire à des résultats peu fiables. Devant ces difficultés, il est primordial d'explorer de nouvelles méthodes utilisant l'apprentissage automatique qui font partie de l'intelligence artificielle.

Ces méthodes novatrices peuvent rendre le processus de classification plus facile, plus précis, plus rapide et plus accessible, tout en diminuant les dépenses et les erreurs humaines.

L'intelligence artificielle (IA) joue un rôle crucial dans divers domaines, y compris la biologie, en particulier dans la classification des microorganismes. Les techniques d'apprentissage automatique et d'apprentissage profond permettent d'analyser de grandes quantités de données biologiques, ce qui aide les chercheurs à identifier des modèles et des caractéristiques subtiles. Grâce à ces avancées, il devient possible de classifier les micro-organismes de manière plus précise et efficace, facilitant ainsi la recherche scientifique et la compréhension de la diversité biologique.

L'objectif principal de cette étude est de développer et d'évaluer deux algorithmes d'apprentissage automatique (arbre de décision et forêt aléatoire) capables de classifier avec précision dix types de microorganismes (levures, *Spirogyra, Volvox, Pithophora, Rhizopus, Penicillum, Aspergillus, protozoaires, diatomées, Ulothrix*) en appliquant ces algorithmes à des attributs descriptifs de ces microorganismes. Ces attributs sont des dimensions morphologiques qui décrivent et représentent les images des micro-

organismes. Nous mettrons en œuvre une approche en plusieurs étapes comprenant la collecte et le traitement des données, la conception et l'entraînement des modèles d'apprentissage automatique, ainsi que l'évaluation des performances obtenues. Dans notre manuscrit, nous présentons de manière approfondie les différentes étapes de la construction de modèles de classification. Il est structuré en quatre chapitres :

Dans le premier chapitre, nous examinons les notions générales concernant les micro-organismes, leur définition et leur classification en fonction de leurs structures morphologiques et leur taxinomie.

Il est crucial d'avoir une compréhension préliminaire avant de passer au deuxième chapitre, qui se focalise sur l'apprentissage automatique. Ce chapitre présente les différentes étapes de l'apprentissage automatique, les diverses catégories d'apprentissage et les algorithmes employés, en mettant en évidence les modèles de la forêt aléatoire et de l'arbre de décision.

Le troisième chapitre présente l'étude expérimentale, faisant le lien entre les deux premiers chapitres. Ce chapitre décrit les outils utilisés et la méthode utilisée pour explorer les caractéristiques descriptives, ainsi que la création des modèles d'apprentissage automatique.

Enfin, dans le quatrième chapitre, les résultats obtenus sont exposés et discutés afin d'évaluer l'efficacité et l'application des modèles de classification des micro-organismes utilisant l'IA.

Partie bibliographique

CHAPITRE 01:

Classification taxinomique et morphologique des microorganismes

1. Introduction

La microbiologie constitue une branche de la biologie qui se focalise sur l'analyse des micro-organismes et de leurs interactions avec leur environnement (Drouet, 2011).

En 1676, **Antoni Van Leeuwenhoek** découvre les animalcules, étant le premier à décrire l'existence d'organismes vivants invisibles à l'œil nu (microorganismes). Depuis lors, de nombreux micro-organismes ont été décrits et officiellement regroupés au **XIXe siècle** sous le nom de Protistes, qui signifient « les tout premiers » en Grec **(protistos)** (Guedon, 2019).

Le terme **« microorganisme »** englobe un nombre et une variété considérables d'espèces différentes, incluant des bactéries, des archées et des eucaryotes tels que les levures (Ronin, 2019).

2. Définition

Un microorganisme provient du grec micro qui signifie petit et bios qui signifie vie. Il est constitué d'une cellule unique (Unicellulaire) ou d'un groupe de cellules identiques (non différenciées) (Ronin, 2019).

Les microorganismes, également connus sous le nom de microbes, sont effectivement un ensemble d'organismes vivants de taille microscopique. Ils sont généralement invisibles à l'œil nu et nécessitent l'utilisation d'un microscope pour être observés. Les microorganismes sont considérés comme les plus anciens et les plus petits des organismes vivants, mais leur découverte a été retardée en raison de leur taille minuscule et de leur invisibilité à l'œil nu. Ils sont extrêmement divers par leur morphologie, leur physiologie, leur mode de reproduction et leur écologie (Bousseboua, 2005a).

«Les protistes se composent d'organismes procaryotes (bactéries et archaebactéries) et d'organismes eucaryotes (algues, protozoaires et champignons microscopiques).» (Bousseboua, 2005). Les virus sont considérés comme des entités microbiennes non vivantes et acellulaires qui dépendent directement des cellules hôtes infectées pour leur reproduction et leur fonctionnement (Rahmani, 2019).

3. Morphologie et structure

Les microorganismes sont classés selon leur morphologie. Les familles qui les regroupent sont les suivantes (Cruzel, 2020) :

- Les bactéries
- Les protozoaires
- Les algues microscopiques
- Les champignons microscopiques
- Les virus

Dans ce qui suit, nous présentons 10 genres des microorganismes (*Spirogyra, Volvox, Pithophora*, Yeast, *Rhizopus*, *Penicillum*, *Aspergillus sp*, Protozoa, Diatom, *Ulothrix*)

3. 1 Le genre *Spirogyra*

Spirogyra (les Spirogyres) est une algue verte filamenteuse d'eau douce, appartenant à la famille des Zygnemaceae. On recense plus de 400 espèces de *Spirogyra* dans le monde (Wongsawad et Peerapornpisal, 2015).

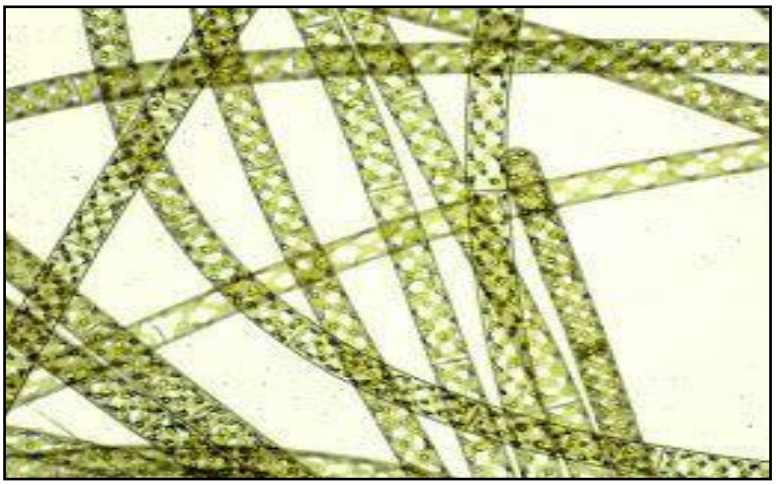

Figure 1 : Filaments de *Spirogyra* (Bouchoukh, 2016)

Les spirogyres sont constituées de filaments coloniaux simples (non ramifiés), principalement transparents et recouverts d'une substance mucilagineuse gluante, disposés de manière désordonnée. Ces filaments, longs de plusieurs décimètres, sont formés d'une succession linéaire de cellules rectangulaires, chacune équipée d'un ou de plusieurs chloroplastes en forme de ruban spiralé (Bouchoukh, 2016).

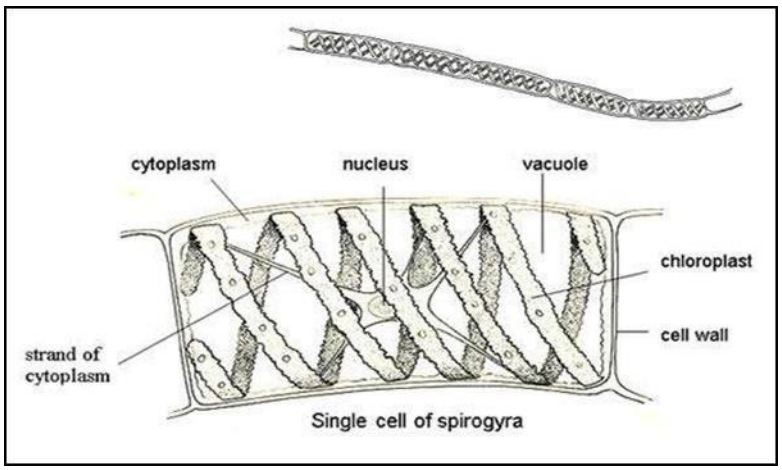

Figure 2 : Cellule de *Spirogyra* (Bouchoukh, 2016)

Les cellules de *Spirogyra* contiennent généralement un ou deux chloroplastes en forme de ruban, disposés en spirale, ce qui explique le nom *Spirogyra*. Ces cellules cylindriques, alignées en rangées, possèdent une paroi cellulosique externe transparente qui leur confère une certaine rigidité. Une fine membrane cytoplasmique adhère à cette paroi du côté interne, également transparente et invisible aux faibles grossissements des microscopes. Près de cette membrane, dans le cytoplasme, se trouvent les chloroplastes, dont le nombre peut varier selon les espèces (Bouchoukh, 2016).

3. 2 Le genre *Volvox*

Le genre *Volvox* rassemble des algues vertes appartenant aux Chlorobiontes, un ensemble qui englobe, en plus des algues vertes, la majorité des plantes terrestres **(**Embryophytes**)**. Ces algues d'eau douce forment des colonies sphériques composées de plusieurs centaines à plusieurs milliers de cellules. Cependant, au sein de ces colonies, on trouve deux types cellulaires haploïdes distincts : les cellules somatiques et les cellules reproductrices, légèrement plus grandes, appelées gonidies. Le thalle, qu'elles forment, est un archéthalle, la catégorie de thalle ayant la structure la plus simple. Chez *Volvox*, cet archéthalle prend la forme d'un cénobe (ou cœnobe), ce qui signifie littéralement «vivre ensemble » (Delarue, 2011).

Figure 3 : Observation générale de *Volvox* (Delarue, 2011)

Toutes les cellules se trouvent sur la surface externe de la sphère. Tandis que le centre est constitué d'un gel mucilagineux produit par les cellules elles-mêmes. Les gonidies, bien qu'elles occupent une position périphérique similaire aux cellules somatiques, sont légèrement en retrait, déplacées vers l'intérieur de la sphère. Les cellules issues des divisions de ces gonidies s'accumulent à l'intérieur de la sphère (Delarue, 2011).

3. 3 Le genre *Pithophora*

Pithophora est un genre d'algues vertes de la famille des Pithophoraceae. La morphologie de *Pithophora* peut être décrite comme celle de filaments ramifiés verts macroscopiques. Ces filaments sont généralement plus gros, moins flexibles et plus étroits. Une caractéristique distinctive de *Pithophora* est la formation facultative d'akinètes, bien que celles-ci ne se forment que pendant la reproduction (Lor *et al.*, 2021).

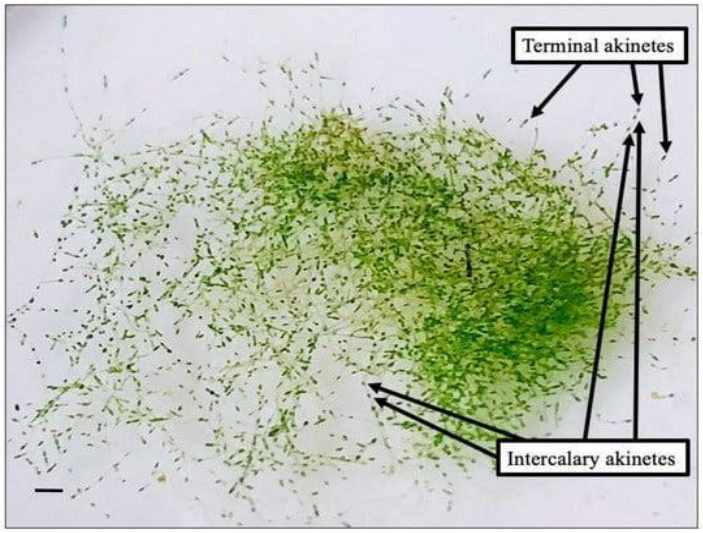

Figure 4 : Aspect macroscopique de *Pithophora roettleri* (Lor *et al.*, 2021)

Pithophora se distingue par ses filaments verts mats, ses cellules cylindriques et ses akinètes foncées agrandies, qui se trouvent à la fois de manière intercalaire et terminale. Chaque cellule renferme plusieurs noyaux et des chloroplastes présentant une structure réticulée. Les parois cellulaires de *Pithophora* sont est composées à la fois de cellulose et de chitine (Baker *et al.*, 2024).

3. 4 Le genre *Rhizopus*

Rhizopus est un genre de moisissures courantes qui se développent sous forme de filaments. Il appartient à l'ordre des Mucorales. L'hyphe est un tube sans cloison, multinucléaire, avec des noyaux haploïdes et une croissance apicale, possédant une paroi composée de chitine et de glucanes. Le protoplasme contient de nombreuses vacuoles qui poussent le cytoplasme et les noyaux vers la périphérie. Les réserves alimentaires sont stockées sous forme de glycogène et de lipides (Bouchoukh, 2016).

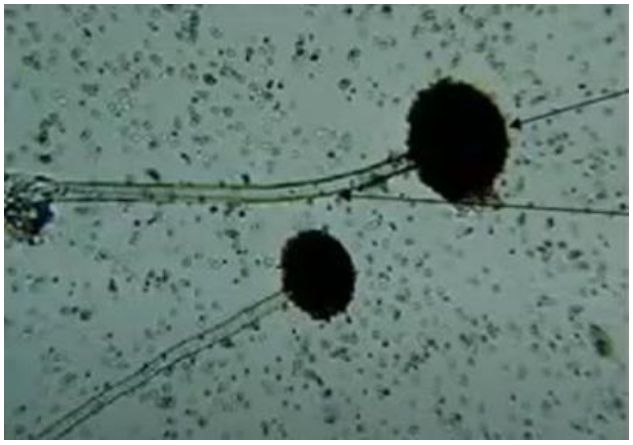

Figure 5 : Observation microscopique du genre *Rhizopus* (Bouchoukh, 2016)

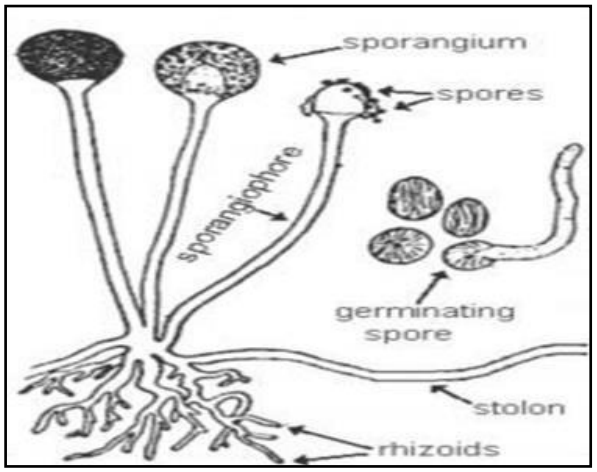

Figure 6 : Caractère morphologique de *Rhizopus* (Bouchoukh, 2016)

Le genre *Rhizopus* se caractérise par la présence de stolons et de rhizoïdes pigmentés, ainsi que par la formation de sporangiophores, émergeant soit individuellement, soit en groupes à partir des nœuds situés au-dessus des rhizoïdes. De plus, il est défini par la présence de sporanges apophyses, columelles et multisports, généralement de forme globuleuse. Après la libération des spores, les apophyses et la columelle s'effondrent souvent, formant une structure similaire à un parapluie. Les sporangiospores, généralement globuleuses à ovoïdes et monocellulaires, peuvent varier en couleur des hyalines à brunes, et présentent souvent des stries dans de nombreuses espèces (Dolatabadi *et al.*, 2014).

3. 5 Le genre *Penicillum*

Ce genre rassemble des champignons filamentaires, faisant partie du phylum des Ascomycètes (Boukhedenna et Merouane, 2013)**.**

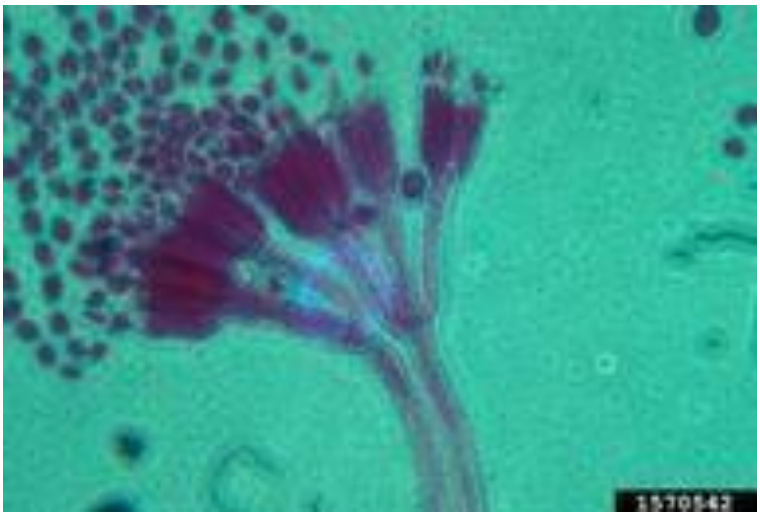

Figure 7 : Observation microscopique du genre *Penicillium* (Bouchoukh, 2016)

Sur le plan morphologique, les *Penicillium* se caractérisent par leur organisation en forme de pinceau. Leur thalle est composé de filaments mycéliens septés et hyalins, sur lesquels poussent des conidiophores lisses ou granuleux, simples ou ramifiés, se terminant par un pénicille. Ces conidiophores peuvent être présents de manière isolée, regroupés en faisceaux dispersés ou agencés en corémies bien définies (Boukhedenna et Merouane, 2013).

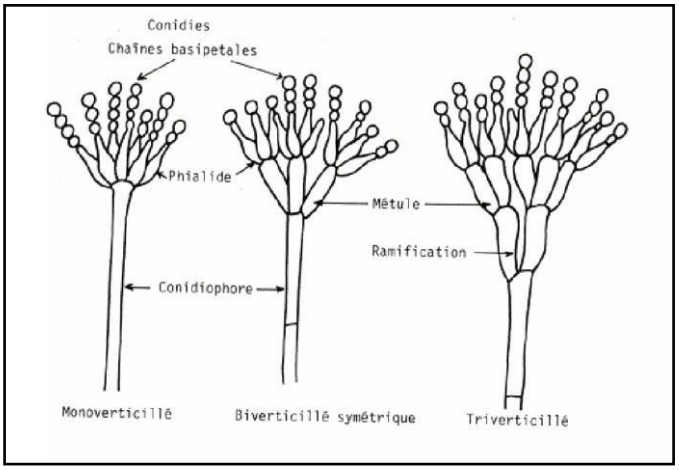

Figure 8 : Caractères morphologiques des *Penicillium* (Boukhedenna and Merouane, 2013)

Les phialides sont disposés en verticilles à l'extrémité des conidiophores. Elles s'insèrent directement (chez les *Penicillium* monoverticillés) ou par l'intermédiaire d'une rangée de métules (chez les *Penicillium* biverticillés) ou de deux rangées successives de métules (chez les *Penicillium* triverticillés) sur les conidiophores. Les conidies, produites en abondance par les phialides, demeurent en chaîne, contribuant ainsi à conférer à la tête conidienne un aspect en pinceau (ou pénicille) (Bouchoukh, 2016).

3. 6 Le genre *Aspergillus*

Le genre *Aspergillus* est affilié à la division des deutéromycètes. Environ 180 espèces, réparties en 18 groupes distincts, constituent la diversité taxinomique du genre Aspergillus (Doghmani *et al.*, 2022).

Les *Aspergillus* se distinguent par la présence d'un appareil végétatif, ou thalle, constitué de filaments mycéliens hyalins (Makhlouf, 2019). Ces filaments sont de diamètre fin et régulier, cloisonnés et ramifiés, portants des conidiophores dressés non ramifiés (Abdelhadi et Boukhroufa, 2011). Les conidiophores se caractérisent par la présence d'une cellule pied basale et se gonflent à leur extrémité distale pour former une vésicule

sphérique ou ovoïde. Sur cette vésicule, une ou deux rangées de stérigmates se développent (selon les espèces) (Makhlouf, 2019).

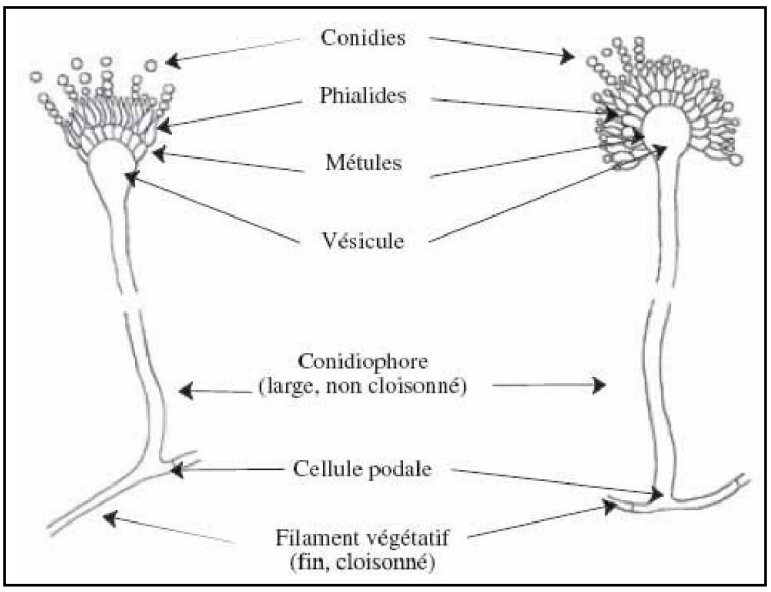

Figure 9: Caractères morphologiques des *Aspergillus* (Makhlouf, 2019)

Les nombreuses conidies, produites par les phialides confèrent à la tête conidienne, donnent un aspect radié lorsque les métules et les phialides recouvrent l'intégralité de la vésicule, ou une apparence en colonne si seule la partie supérieure est fertile. Ces conidies sont toujours disposées en chaînettes. Selon les espèces, elles peuvent être unicellulaires, globuleuses, subglobuleuses ou elliptiques, lisses ou ornées, hyalines ou pigmentées en jaune, brun, noir ou vert. Les cellules à paroi épaissie et les sclérotes peuvent parfois être présentes (Abdelhadi et Boukhroufa, 2011).

3. 7 Les protozoaires

Les protozoaires (protos-premier, zoon-animal) sont des eucaryotes unicellulaires. Comme chez tous les eucaryotes, le noyau est enfermé dans une membrane. Chez les protozoaires autres que les ciliés, le noyau est vésiculaire, avec de la chromatine dispersée donnant un aspect diffus au noyau. Tous les noyaux dans l'organisme individuel apparaissent semblables (Robert et Yaeger, 1996).

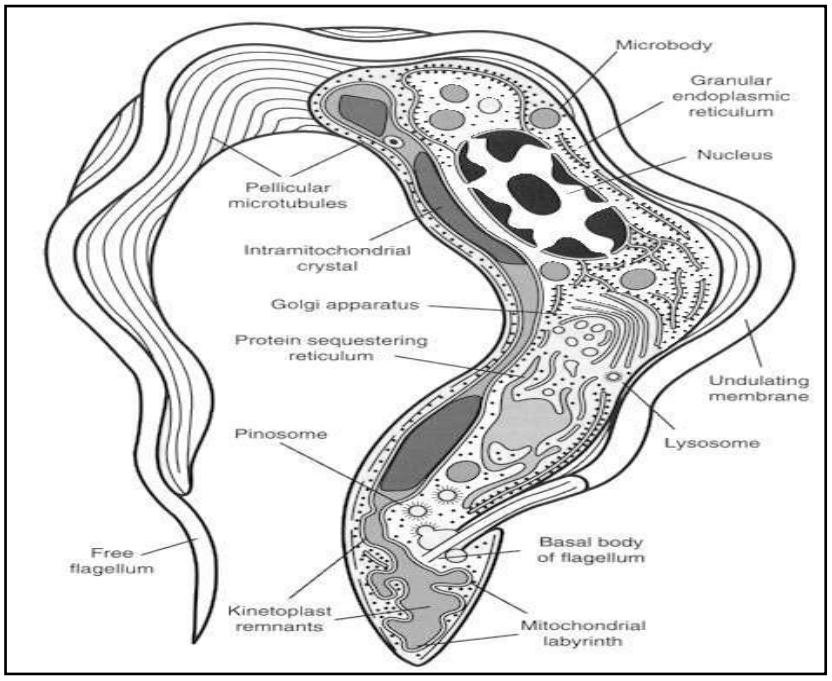

Figure 10: Structure fine d'un parasite protozoaire (Robert et Yaeger, 1996)

La membrane plasmique entourant le cytoplasme recouvre également les structures locomotrices en saillie telles que les pseudopodes, les cils et les flagelles. La couche externe de certains protozoaires, appelée pellicule, est suffisamment rigide pour maintenir une forme distinctive, comme chez les trypanosomes et Giardia (Robert et Yaeger, 1996).

Cependant, ces organismes peuvent facilement se tordre et se plier lorsqu'ils se déplacent dans leur environnement. Chez la plupart des protozoaires, le cytoplasme se différencie en ectoplasme (la couche externe transparente) et en endoplasme (la couche interne contenant les organites); la structure du cytoplasme est plus facilement visible chez les espèces avec des pseudopodes saillants, comme les amibes. Certains protozoaires possèdent un cytosome ou une « bouche » cellulaire pour ingérer des fluides ou des particules solides (Robert et Yaeger, 1996).

3. 8 Les diatomées

Les diatomées, également connues sous le nom de bacillariophycées, sont des organismes unicellulaires microscopiques (eucaryotes) dont la taille varie de quelques micromètres à plus de 500 micromètres (0,5 millimètre). Elles se distinguent par la présence d'un squelette externe en silice (SiO_2) à l'intérieur duquel se trouve le contenu cellulaire comprenant le noyau, les chloroplastes, les mitochondries, les vacuoles, etc. (Lavoie *et al.*, 2008).

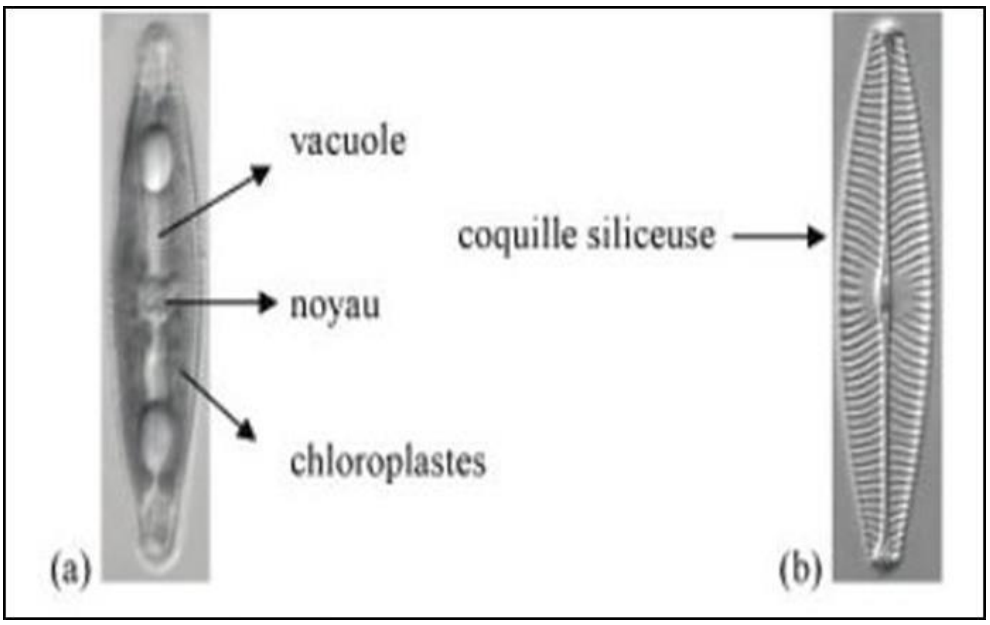

Figure 11: Caractère morphologique de diatomée (Lavoie *et al.*, 2008)

Le frustule (squelette ou coquille siliceuse) est constituée de deux valves s'emboîtant l'une dans l'autre, de manière similaire à une boîte de fromage, où le couvercle est appelé épivalve et le fondhypo-valve. Chaque valve est prolongée par une ou plusieurs ceintures connectives. Les diatomées sont divisées en deux principaux ordres (Rumeau et Coste, 1988) :

- Les centriques, également appelées Centrophycidées, sont généralement des organismes planctoniques présentant une seule symétrie axiale.
- Les pennales ou Pennatophycidées présentent une symétrie par rapport à un plan. Dans ce dernier groupe, les formes comportant un raphé sont prédominantes. Ce raphé est formé par une fente qui s'étend d'un pôle à l'autre le long de l'axe apical et elle est interrompue au centre de la valve.

3. 9 Le genre *Ulothrix*

Le genre *Ulothrix*, regroupe des algues vertes filamenteuses non ramifiées et unisériées (Lokhorst et Vroman, 1972).

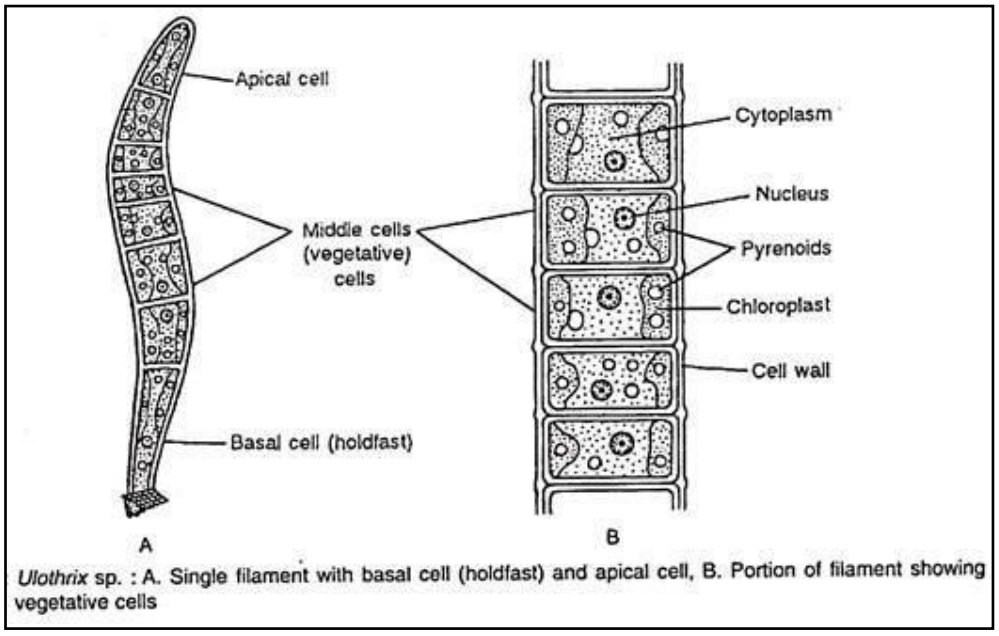

Figure 12: Caractère morphologique du genre *Ulothrix* (Nee l esh, 2016)

Le thalle d'*Ulothrix* est filamenteux, long, non ramifié et multicellulaire, où les cellules sont disposées en une seule rangée (c'est-à-dire unisériées). Les filaments sont d'un vert vif et ils restent attachés à une extrémité au substrat tel que des pierres, des rochers ou d'autres objets solides (Nee l esh, 2016).

Toutes les cellules sauf le crampon basal peuvent se diviser et leur paroi est composée de pectine externe et de cellulose interne. À l'intérieur de la paroi cellulaire se trouve la membrane cellulaire. La membrane cellulaire enferme le protoplaste. Il se compose de cytoplasme, de chloroplaste et de noyau (Nee l esh, 2016).

3. 10 Les levures

Les levures sont des champignons unicellulaires et ils ont évolué plusieurs fois de manière indépendante à travers le règne fongique (Chavez *et al.*, 2024).

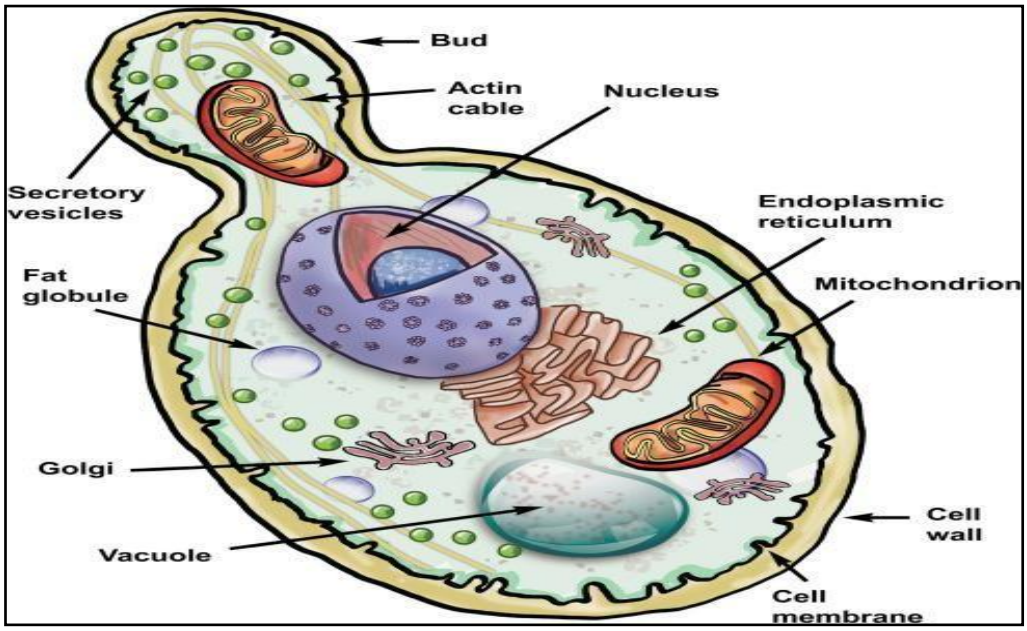

Figure 13: Structure générale d'une cellule de levure (Bensalem et Horchi, 2020)

La morphologie des levures a une importance taxinomique considérable manifestant sous des formes variées telles que sphérique, ovoïde, globuleuse, cylindrique, ellipsoïde, allongée, apiculée, ogivale, triangulaire ou en forme de bouteille. Avec un diamètre cellulaire d'environ 5 à 10 µm, les levures sont nettement plus grandes que les cellules bactériennes dont la taille se situe entre 0,5 et 5 µm (Bensalem et Horchi, 2020).

Les levures sont classées parmi les eucaryotes en raison de la présence d'un noyau et de chromosomes distincts, dont la taille et le nombre varient d'une espèce à l'autre. De plus, la cellule de levure contient divers organites intracellulaires entourés de membranes individuelles telles que l'appareil de Golgi, les mitochondries, le réticulum endoplasmique et les vacuoles (Bensalem et Horchi, 2020).

4. Classification et taxinomie

Le botaniste suisse Augustin-Pyramus de Candolle (1778-1841) a proposé en 1813 le terme « taxonomie » (du grec taxis, qui signifie « ordre », « arrangement » et « nomos », qui signifie « loi ») pour désigner la science des lois de la classification des êtres vivants (Guermazi, 2017).

La taxonomie est l'ensemble des principes et théories qui permettent de classer et de valider le classement des organismes. Elle est divisée en trois domaines distincts. D'une part, la classification implique de regrouper et de classer les microorganismes dans des

ensembles en fonction de leurs similitudes : ces ensembles sont appelés « **taxons** ». Par la suite, la nomenclature implique d'attribuer un nom à chaque taxon. Finalement, le processus d'identification utilise les deux domaines précédents pour reconnaître et attribuer un nom[1] (Bousseboua, 2005a).

Les micro-organismes se divisent en plusieurs familles, chacune avec des milliers d'espèces : bactéries, virus, protozoaires et champignons. Elle est définie hiérarchiquement par domaine, règne, division, classe, ordre, famille, genre et espèce (*Le microbe sous toutes ses formes*, 2017).

Au cours de notre recherche, nous examinerons la classification de dix microorganismes qui diffèrent entre les champignons et les microalgues.

- Les trois microorganismes suivants, à savoir les levures, les diatomées et les protozoaires, sont classés en taxonomie selon le système de classification biologique (tableau 1) :

Tableau 01 : classification taxinomique des microorganismes

MO / **CT**	Levure	Diatomées	Protozoaire
Domaine	Eukariota	Eukariota	Eukariota
Règne	Fungi	Plantae	Protozoa

MO : Microorganisme CT : Classification taxinomique

- Les taxonomies des autres microorganismes sont mentionnées dans le tableau 02.

[1]http://www.microbes-edu.org/etudiant/intro.html

Tableau 02: classification taxinomique des microorganismes

	Aspergillus sp	***Pénicillium***	***Rhizopus***	***Volvox***	***Pithophora***	***Ulothrix***	***S***
ine	Eukaryota	Eukaryota	Eukaryota	Eukariota	Eukariota	Eukariota	E
ne	Fungi	Fungi	Fungi	Plantae	Viridiplantae	Plantae	
ement	Ascomycota	Ascomycota	Zygomycota	Chlorophyta	Chlorophyta	Chlorophyta	Ch
se	Eurotiomycetes	Eurotiomycetes	Zygomycetes	Chlorophyceae	Ulvophyceae	Ulvophyceae	Zygne
re	Eurotiales	Eurotiales	Mucorales	Volvocales	Cladophorales	Ulothrichales	Zy
lle	Trichocomaceae	Trichocomaceae	Rhizopodaceae	Volvocaceae	Pithophoraceae	Ulothrichaceae	Zyg
re	***Aspergillus***	***Pénicillium***	***Rhizopus***	***Volvox***	***Pithophora***	***Ulothrix***	***S***

MO : Microorganisme

CT : Classification taxinomique

CHAPITRE 02:

Apprentissage Automatique

1. Introduction

Le terme "intelligence" trouve son origine dans les termes latins *intellegere* ou *intelligere*, signifiant "choisir entre" (Soudoplatoff, 2018).

L'intelligence Artificielle (IA) représente un champ interdisciplinaire entre l'informatique et les mathématiques, regroupant un ensemble de techniques algorithmiques et de théories visant à développer des machines capables d'imiter l'intelligence humaine. Son objectif fondamental réside dans la reproduction de l'intelligence humaine afin de résoudre des problèmes complexes. Cette démarche implique la modélisation de l'intelligence humaine en tant que phénomène, semblable à celle employée dans les domaines de la physique, de la chimie ou de la biologie (Matteis*et al.*, 2022).

L'intelligence artificielle est un domaine en expansion continu, dont la théorie et les applications s'étendent à de nombreux domaines tels que la théorie des probabilités, les neurosciences, la robotique, la théorie des jeux, la santé et le transport (Matteis*et al.*, 2022).

Le concept d'intelligence artificielle (IA) a été conceptualisé par John McCarthy en 1956, lors d'un séminaire de deux mois qu'il a organisé au Dartmouth College à Hanover, New Hampshire, aux États-Unis. Ce séminaire a rassemblé dix chercheurs américains spécialisés dans la théorie des automates, les réseaux de neurones et l'intelligence. Ce séminaire a abouti à l'officialisation du terme "intelligence artificielle" en tant que désignation officielle d'un nouveau domaine de recherche (Mattei et Villata, 2022).

2. Définition de l’apprentissage automatique

Arthur Samuel a introduit l'apprentissage automatique pour la première fois en 1959, avec la définition suivante : « champ d'études qui permet aux ordinateurs d'apprendre sans être explicitement programmés ». Cette définition se concentre clairement sur le concept d'intelligence artificielle. Ce dernier représente l’environnement dans lequel le machine learning a évolué (Simon, 2016).

L'apprentissage automatique, ou aussi appelé Machine Learning (ML), est un domaine scientifique et plus particulièrement une sous-catégorie de l'intelligence artificielle qui permet aux ordinateurs d'acquérir des capacités d'apprentissage autonomes

sans une programmation explicite. Elle englobe diverses méthodes permettant la création automatique de modèles à partir de données. Ces méthodes se manifestent sous la forme d'algorithmes conçus pour analyser et interpréter les données, facilitant ainsi l'apprentissage autonome des systèmes informatiques (Bastien, 2024). Machine Learning utilise des algorithmes pour générer des modèles statistiques à partir de données structurées (Martin, 2023).

3. Objectifs de l'apprentissage automatique

Selon Tantuğ et Türkmenoğlu (2015), l'apprentissage automatique vise principalement à développer des modèles qui peuvent s'entraîner eux-mêmes pour s'améliorer, à percevoir les modèles complexes et à trouver des solutions aux nouveaux problèmes en utilisant les données des modèles précédents (Tantuğ et Türkmenoğlu, 2015).

Machine Learning est largement employée en science des données et en analyse de données. Il offre la possibilité de concevoir, expérimenter et mettre en œuvre des algorithmes d'analyse prédictive sur diverses données pour anticiper l'avenir. Il s'agit d'offrir aux algorithmes la possibilité de détecter des « patterns », c'est-à-dire des motifs récurrents, dans les ensembles de données (Jérémy, 2020).

L'apprentissage automatique vise principalement à donner aux machines la capacité d'apprendre à partir des données, en repérant des modèles et en les généralisant afin de prendre des décisions ou de prédire de nouvelles données. Il s'agit fréquemment d'algorithmes d'apprentissage supervisé ou non supervisé, qui utilisent des ensembles de données massives pour entraîner des modèles capables de faire des prédictions précises ou de fournir des insights importants (Batta, 2018).

4. Démarche des algorithmes d'apprentissage automatique

D'une manière générale, les algorithmes d'apprentissage automatique se décomposent en plusieurs étapes (Matteis *et al.*, 2022).

4.1. Phase d'entraînement (ou d'apprentissage)

Pendant la phase d'entraînement (ou d'apprentissage), le modèle sélectionné est exposé à un grand nombre d'exemples significatifs. Le système cherche alors à acquérir des règles implicites en se basant sur ces données, appelées données d'entraînement. Cette phase précède généralement l'utilisation du modèle, bien que certains systèmes puissent

continuer à apprendre indéfiniment s'ils reçoivent un retour sur les résultats, phénomène appelé apprentissage en ligne (Matteis *et al.*, 2022).

4.2. Phase de prédiction (inférence)

Durant la phase d'inférence, le modèle entraîné est apte à être utilisé sur de nouvelles entrées. Ces entrées peuvent être traitées même si elles n'ont pas été présentées au modèle lors de la phase d'apprentissage. En effet, grâce à l'extraction de règles implicites, le modèle peut se généraliser à des entrées inconnues (Matteis *et al.*, 2022).

5. Type de système d'apprentissage automatique

Quatre types principaux des algorithmes d'apprentissage automatique peuvent être identifiés (Figure 14) :

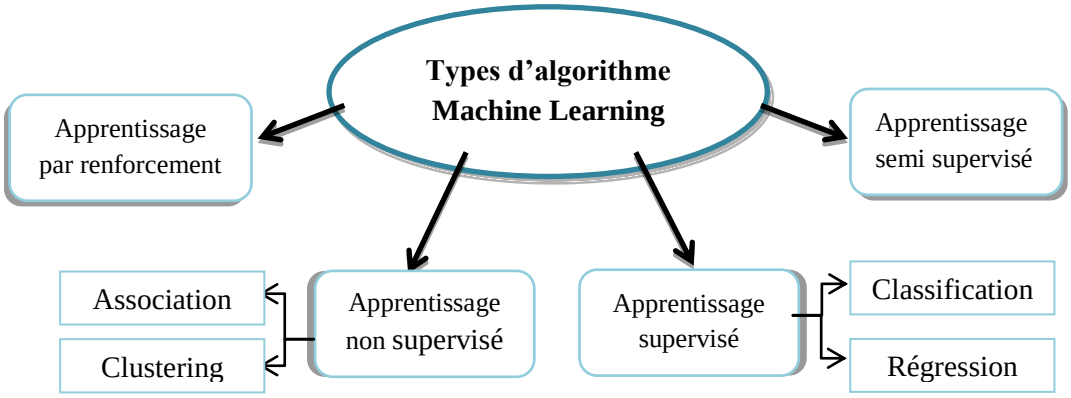

Figure 14: les quatre Types de système d'apprentissage automatique

5.1. Apprentissage par renforcement ou Deep learning (DL)

L'apprentissage profond (ou deep learning en anglais) est une méthode d'apprentissage automatique basée sur le modèle des réseaux neuronaux. Des dizaines, voire des centaines de couches de neurones sont combinées afin d'aggraver la mise en place des règles (Jérémy, 2020).On utilise fréquemment l'apprentissage par renforcement dans des domaines tels que les jeux, la navigation et la robotique (Martin, 2023). Le Q-Learning est l'algorithme d'apprentissage par renforcement le plus couramment employé. Son mécanisme repose sur la détermination de l'action optimale. C'est celle qui permet de maximiser l'espérance des récompenses des prochains états, en tenant compte d'un facteur de mise à jour (Talbi, 2020).leur structure est mentionnée dans la figure 15.

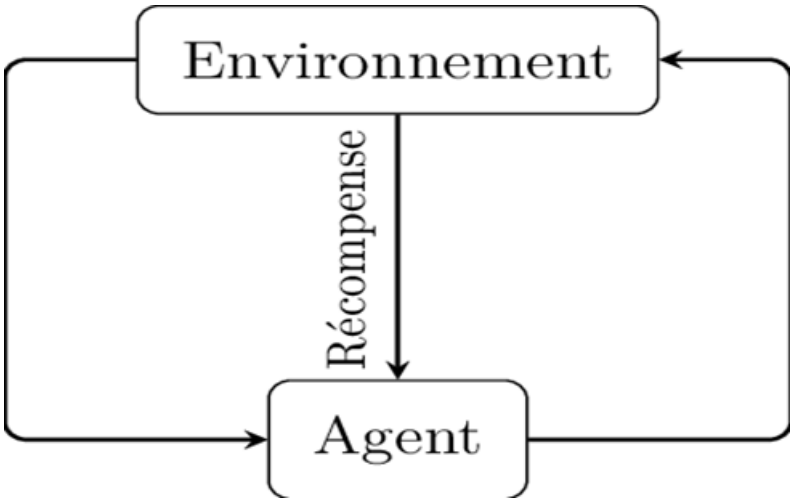

Figure15: Structure de l'apprentissage par renforcement (Lesel, 2016)

5.2. Apprentissage non supervisé

L'apprentissage non supervisé est une technique d'apprentissage automatique où un modèle est élaboré à partir d'un ensemble de données d'entraînement restreint aux entrées. L'apprentissage du modèle consiste à repérer les structures et les modèles présents dans les données sans nécessiter de sorties étiquetées. Les applications de clustering utilisent fréquemment l'apprentissage non supervisé, tel que la segmentation de la clientèle, l'analyse de texte et la détection d'anomalies (Martin, 2023). L'apprentissage non supervisé (unsupervised learning) consiste à faire apprendre l'algorithme de manière indépendante (Talbi, 2019).(figure 16)

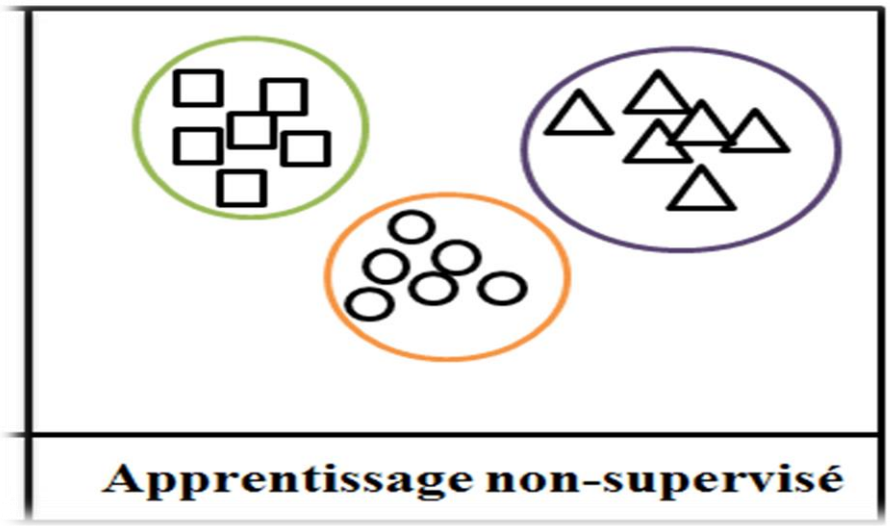

Figure 16: structure de l'apprentissage non supervisé (Raphael, 2022)

5.3. Apprentissage semi supervisé

Le protocole SSL (Semi Supervise Learning) est spécialement conçu pour les secteurs d'application où les données non étiquetées sont fréquentes, comme le traitement des images, la collecte d'informations et la bioinformatique (Chapelle *et al.*, 2009).

La technologie SSL se situe à mi-chemin entre l'apprentissage supervisé et non supervisé, ce qui signifie que l'ensemble des données est séparé en étiquettes et sans étiquettes (Chapelle *et al.*, 2009).

L'apprentissage semi supervisé consiste à utiliser des données non-annotées afin de compléter l'apprentissage supervisé (Chapelle *et al.*, 2009).

5.4. Apprentissage supervisé

L'apprentissage automatique supervisé est la construction des algorithmes qui consiste à créer des modèles et des hypothèses générales en utilisant des instances externes afin de prédire le sort des instances à venir (Singh *et al.*, 2023). On utilise fréquemment l'apprentissage supervisé dans des applications comme la détection de fraude, la reconnaissance d'images et la prédiction de prix (Martin, 2023). (Figure 17)

L'objectif de l'apprentissage supervisé est d'exploiter les informations d'entrée pour anticiper les valeurs de sorties (outputs ou réponses) (Talbi, 2019).

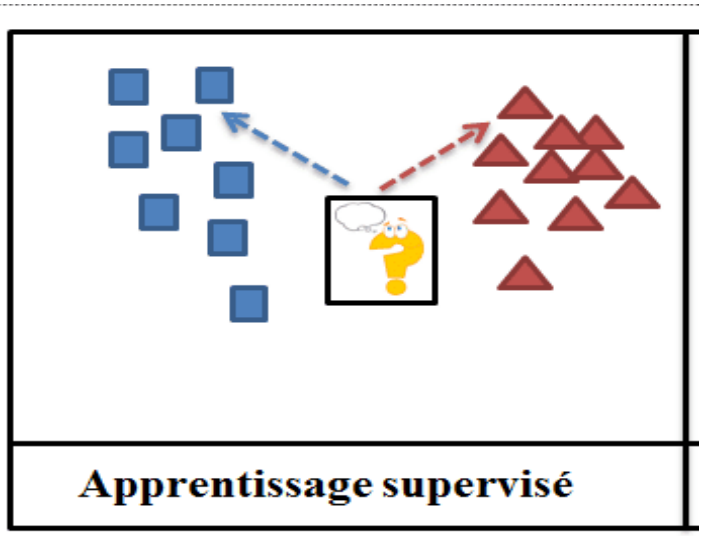

Figure 17: structure de l'apprentissage non supervisé (Raphael, 2022)

Il existe plusieurs algorithmes de l'apprentissage supervisé, mais les plus utilisés sont : la régression et la classification.

5.4.1. Régression

La régression est une méthode d'apprentissage automatique supervisé qui utilise des algorithmes pour prédire des valeurs continues comme les ventes, le salaire, le poids ou la température (Moez, 2022).

La régression est employée lorsqu'il est possible de prédire une sortie qui peut prendre des valeurs continues, ce qui est le cas d'une variable réelle. Par exemple, on peut utiliser un algorithme qui anticipe la consommation électrique d'une installation ou un algorithme qui anticipe le cours des actions en bourse (Matteis *et al.*, 2022).

Les tâches de régression peuvent être effectuées à l'aide de nombreux algorithmes d'apprentissage automatique. Il existe différents types de régression tels que la régression linéaire, le régresseur d'arbre de décision, le k régresseur du voisin le plus proche, le régresseur de forêt aléatoire et les réseaux de neurones (Moez, 2022).

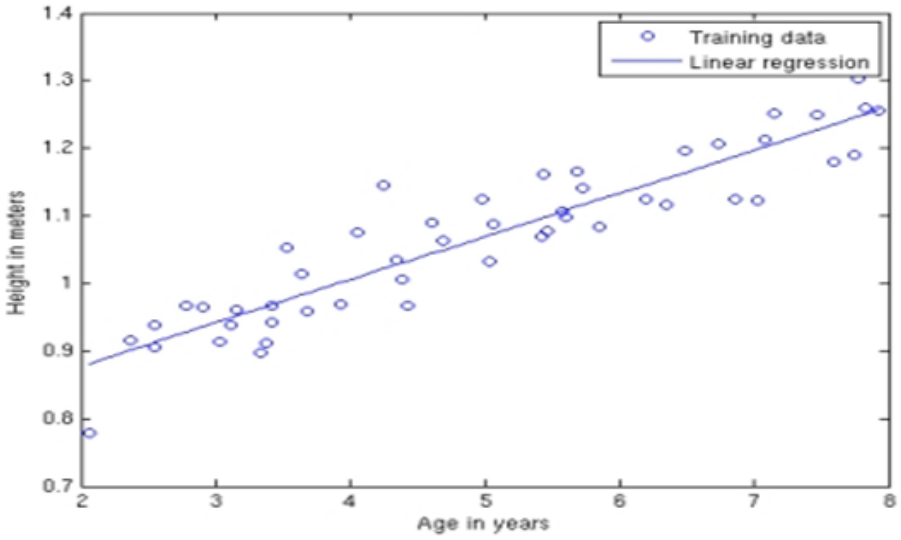

Figure 18 : Modèles de régression linéaire (Padala et *al.*, 2019)

5.4.2. Classification

La classification est un ensemble de tâches réalisées dans un ordre déterminé pour résoudre un problème ou proposer de nouvelles solutions. Comme l'utilisation d'un système d'intelligence artificielle pour apprendre (Raphael, 2022).

Le rôle des algorithmes de classification utilisés dans le machine learning est précisément celui-ci. Grâce à eux, le logiciel peut apprendre de manière autonome à partir de diverses bases de données (Raphael, 2022).

Le concept de classification consiste donc à classer les divers éléments d'un jeu de données en différentes catégories. Ces données sont regroupées en fonction de leur similitude. Étant donné que les données ont des traits communs, il est plus aisé de prédire leur comportement (Raphael, 2022).

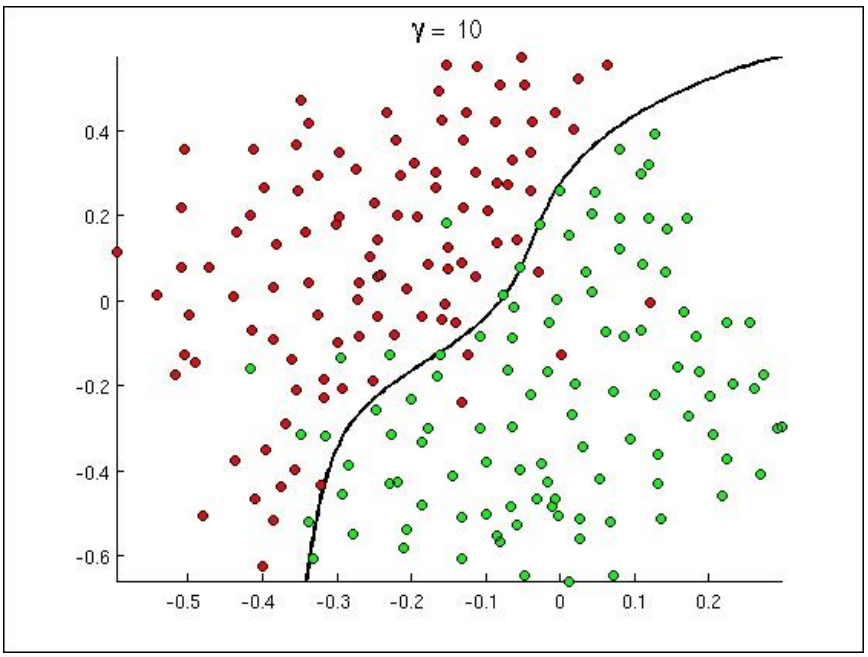

Figure 19 : modèle de classification linéaire (Gullitti et Llc, 2017)

6. Algorithmes d'apprentissage automatique

Selon la complexité du problème, différents algorithmes d'apprentissage automatique sont disponibles (Aiboud et Laskri, 2020).

- k Nearest Neighbors (Le k-NN)
- k -moyenne (K-MEANS)
- Arbres de décision
- Naïve de Bayes
- Support Vecteurs Machines (SVM)
- Réseaux de neurones (RN)

6.1. Choix du modèle

Il existe une diversité considérable d'algorithmes en apprentissage automatique. Il devient donc essentiel de déterminer lequel choisir. Un premier critère de sélection concerne le type d'apprentissage souhaité ou possible. Ainsi, il est nécessaire de décider si l'on opte pour un apprentissage supervisé, non supervisé ou par renforcement, en fonction des exigences de la tâche à accomplir et des données disponibles (Matteis*et al.*, 2022).

Dans ce travail, nous nous concentrerons sur deux algorithmes : la "Forêt d'arbres décisionnels" (Random Forest) et "l'Arbre de Décision" (Decision Tree).

6.1.1. Random Forest

L'algorithme d'apprentissage automatique supervisé Random Forest est un algorithme polyvalent et puissant qui combine différents arbres de décision pour créer une « forêt ».

Le langage de programmation R et Python sont employés pour résoudre des problèmes de classification et de régression. Les Data Scientists ont une grande préférence pour cette méthode de machine learning en raison de ses multiples bénéfices par rapport aux autres algorithmes de données. Son interprétation est simple, sa stabilité est généralement satisfaisante et elle peut être employée pour des tâches de régression ou de classification, ce qui permet de traiter une grande diversité de problèmes en Machine Learning.

Dans Random Forest, le mot « Forest » suggère clairement que cet algorithme utilise des arbres de décision, également connus sous le nom d'arbres décisionnels (Harfi, 2020). (Présentée dans la figure 20)

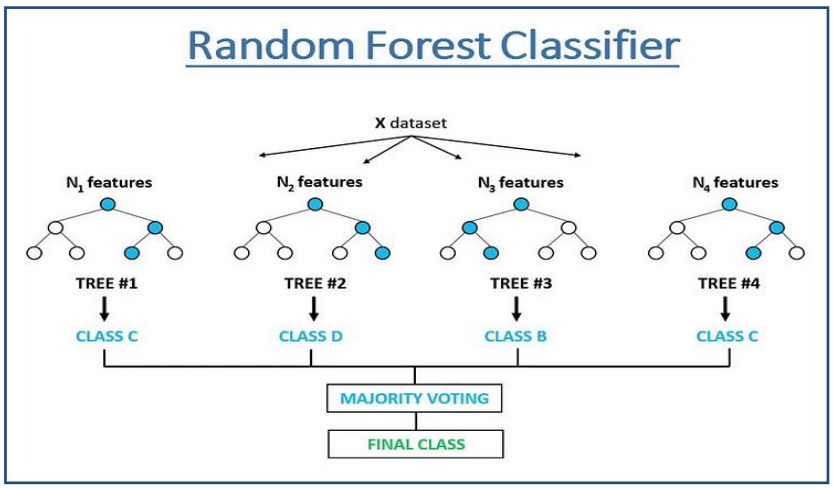

Figure 20 : Exemple d'arbre de Décision (Khushaktov, 2023)

6.1.2. Arbre de Décision

L'arbre de décision est un élément essentiel de l'algorithme de classification dans le domaine de l'apprentissage automatique, permettant également de résoudre les problèmes de régression en utilisant des règles de classification. Dans le domaine contemporain de l'apprentissage automatique, l'arbre de décision est largement employé, et de nombreux algorithmes de ce domaine sont intégrés dans notre quotidien. L'un des principaux algorithmes est l'arbre de décision. L'analyse de l'arbre de décision en tant que modèle prédictif se fait selon une méthode algorithmique, où un ensemble de données est subdivisé en sous-ensembles en fonction des conditions. Le nom de l'algorithme même laisse entendre qu'il s'agit d'un modèle en arbre à base d'instructions si-alors-sinon. La profondeur de l'arbre et la présence de nombreux nœuds améliorent les performances du modèle (Harfi, 2020).

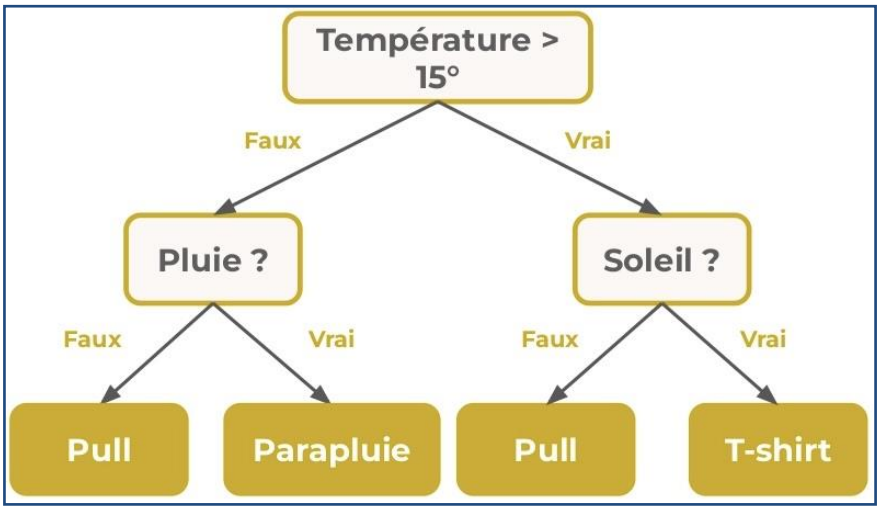

Figure 21 : Exemple d'arbre de Décision (Keldenich, 2022)

Partie pratique

CHAPITRE 01:

Matériel et Méthodes

1. Introduction

Ce chapitre se focalise sur la connaissance des méthodes d'apprentissage automatique et de leurs étapes employées pour classer dix catégories d'organismes microscopiques (algues, protozoaires et champignons). Dans un premier temps, nous examinerons les matériaux employés dans l'étude, y compris les données biologiques et informatiques. Par la suite, nous exposerons en détail la méthode employée, en soulignant ses étapes. Enfin, nous exposerons la manière de concevoir un modèle d'intelligence artificielle utilisant l'algorithme des forêts aléatoires et l'algorithme de l'arbre de décision, en plus de présenter les résultats obtenus.

2. Matériel

2.1 Données biologiques

L'objectif de cette recherche est de classer dix genres de microorganismes à l'aide de modèles d'apprentissage automatique. Afin d'accomplir cela, nous avons recours à une base de données [1] textuelle décrivant les dimensions morphologiques des images de ces micro-organismes, recueillie sur la plateforme en ligne Kaggle[2]. Dans cette collection, on retrouve vingt-six caractéristiques descriptives pour 30 527 individus microscopiques. Le tableau suivant présente les diverses caractéristiques descriptives des microorganismes.

Tableau 3 : attributs descriptives des microorganismes

Attributs	Description
Solidity (Solidité)	La phrase fait référence au concept de « rigidité » en géométrie géométrique, car elle exprime le rapport entre l'aire d'un objet géométrique et l'aire de la coque basale qui l'entoure. Ce rapport est calculé en divisant la surface de l'objet géométrique par la surface de la coque basale qui l'entoure. (solidité).
Eccentricity (Pente)	Le concept de « pente » en géométrie est le rapport entre la longueur du grand axe et la longueur du petit axe d'un objet. Cela signifie qu'il mesure la différence entre la forme d'un objet et la forme d'un cercle parfait.

[1] https://data.mendeley.com/datasets/f9m85ptmvc/3
[2] https://www.kaggle.com/datasets

Attributes	Description
EquivDiameter (diamètre équivalent)	Le « diamètre équivalent » représente : le diamètre qui a la même aire que l'aire spécifiée. En d'autres termes, c'est le diamètre d'un cercle égal à l'aire de la zone donnée**.**
Extrema (points finaux)	Les « extrema » ou « points finaux » d'une région sont les points les plus éloignés de la région ou de la forme donnée. Les points extrêmes sont les quatre coins et les milieux des quatre côtés de la forme. De plus, la déclaration précise que ces points sont représentés dans un vecteur selon un format spécifique, où chaque point est localisé en fonction des directions des points dans la forme.
Fille Area (Zone remplie)	Zone remplie : nombre de pixels illuminés dans l'image remplie, renvoyé sous forme numérique. C'est simplement le nombre de pixels colorés ou actifs dans une image donnée.
Orientation (Direction)	Orientation : « Direction » représente la direction générale du format. La valeur de tendance varie de -90° à 90°, ce qui indique la tendance dans la zone de tendance principale.
EulerNumber (Nombre d'Euler)	Nombre d'Euler : Le nombre d'objets dans la région moins le nombre de trous dans ces objets. Le nombre d'Euler est souvent utilisé pour décrire la topologie des régions dans une image.
BoundingBox (1-4) (Boîte englobant)	Boîte englobant : Position et taille de la plus petite boîte (rectangle) qui englobe l'objet. Cette phrase décrit que la "boîte englobant" représente la position et la taille du plus petit rectangle qui entoure complètement l'objet dans une image. La boîte englobant est souvent utilisée pour délimiter la zone occupée par un objet dans le contexte de la reconnaissance d'objets ou de l'analyse d'images.
Périmètre	Périmètre : Nombre de pixels autour de la bordure de la région. Cette phrase explique que le périmètre d'une région correspond au nombre de pixels situés le long de la bordure de cette région dans une image. Le périmètre est une mesure de la longueur de la frontière externe de la région.

Attributs	Description
ConvexHull (1-4) (Enveloppe convexe)	Enveloppe convexe : Plus petite forme/ polygone convexe qui contient l'objet. Cette phrase explique que l'enveloppe convexe est la plus petite forme ou polygone convexe qui englobe complètement l'objet dans une image. Elle est utilisée pour définir la forme la plus simple qui contient tous les points de l'objet, en éliminant les saillies et les indentations.
MajorAxisLength (Grand axe)	Grand axe : Le grand axe est défini par les extrémités de la plus longue ligne qui peut être tracée à travers l'objet. La longueur (en pixels) du grand axe correspond à la plus grande dimension de l'objet. Cette phrase explique que le grand axe d'un objet est représenté par Grand axe : Le grand axe est défini par les extrémités de la plus longue ligne qui peut être tracée à travers l'objet. La longueur (en pixels) du grand axe correspond à la plus grande dimension de l'objet.
Convex area (région convexe)	Une région convexe est par définition supérieure ou égale à l'aire de la région. Il peut être utilisé pour calculer le facteur de forme appelé «convexité » ou « rigidité », qui est défini comme le rapport entre la surface et la surface convexe.
Extent (Étendue)	Étendue : Rapport de la surface en pixels d'une région par rapport à la surface de la boîte englobant d'un objet. Cette phrase décrit que l'"étendue" est le rapport entre la surface en pixels d'une région et la surface de la boîte englobant de l'objet correspondant. En d'autres termes, cela mesure à quel point la région occupée par l'objet est étendue par rapport à la zone couverte par la boîte englobant.
MinorAxisLength (Axe mineur)	Axe mineur : L'axe perpendiculaire au grand axe est appelé axe mineur. La longueur (en pixels) de l'axe mineur correspond à la plus petite ligne reliant une paire de points sur le contour. Cette phrase explique que l'axe mineur d'un objet est celui qui est perpendiculaire au grand axe. La longueur de l'axe mineur, exprimée en pixels, correspond à la plus petite distance entre deux points du contour de l'objet.

Attributs	Description
Centroid (1-2) (centroïde /point central)	Le centroïde est le point central de l'objet. le point d'intersection des trois médianes du triangle est appelé centre de gravité d'un triangle[3]
Raddi **(**rayon)	Un rayon est une mesure de la distance du centre d'un objet circulaire jusqu'à sa limite la plus extérieure. Un rayon n'est pas seulement une dimension d'un cercle, mais aussi pour une sphère, une demi-sphère, un cône avec une base circulaire, un cylindre ayant des bases circulaire.
Area (Espace découvert)	area : Espace découvert, place publique, cour, vestibule, basse-cour.
Microorganisme	Être vivant microscopique tel que les bactéries, les virus, les champignons unicellulaires (levures), et les protistes.

2.2 Configuration de la machine

Le tableau ci-dessous présente en détail les caractéristiques de l'ordinateur utilisé.

Tableau 4 : les caractéristiques de l'ordinateur utilisé pour l'apprentissage automatique DESKTOP-K7BS3TR

Ordinateur	Caractéristiques
Processeur	Intel(R) Celeron(R) CPU N3350 @ 1.10GHz 1.10 GHz
Memoire installée RAM	4,00 Go (3,83 Go utilisable)
Stockage	57,5 GO
Système d'exploitation	Windows 10 Professionnel
Type de système	Système d'exploitation 64 bits, processeur x64

[3] https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://byjus.com/maths/centroid/&ved=2ahUKEwjLm6vBx8WGAxUmUqQEHQAqApkQFnoECBMQAQ&usg=AOvVaw1FfMOizxkGIV-VAwgfYBjg

2.3 Données informatiques

Dans cette section, nous exposons en détail les outils et les bibliothèques utilisées pour réaliser ce travail.

2.3.1. Outils

On utilise le langage de programmation Python, via Google Colab.

Tableau 5 : principaux outils utilisés

Outil	Description
Python 	Python est un langage facile à utiliser, dynamique, extensible et gratuit, qui offre une possibilité sans l'imposer. La programmation adopte une approche modulaire et orientée objet. Depuis 1989, Guido van Rossum et de nombreux contributeurs bénévoles ont travaillé sur le développement de Python (Benmansour, 2018).
Excel 2010 	Excel est un tableur électronique développé par Microsoft. Un tableur offre la possibilité de réaliser des calculs rapides et précis. La feuille de calcul est une feuille électronique de nombres générée par Excel (Haché, 2003). [4]
Google drive 	Google Drive est une plateforme de partage et de stockage de fichiers en ligne mise en place par Google. Il est aussi compatible avec l'ouverture et la modification de fichiers (doc, ppt, xsl, etc.) et il requiert une adresse Gmail, mais propose un stockage gratuit de 15 Go ainsi qu'un accès aux outils bureautiques.[5]

[4] file:///D:/memoire/PDF/3232.pdf
[5] file:///D:/memoire/PDF/Introduction%20Google%20Drive.pdf

Outil	Description
Google Colab 	« Collab » ou Collaborative est un produit de Google Research. Colab offre la possibilité à tout le monde de créer et de lancer du code Python de manière indépendante via le navigateur, et est spécialement conçu pour l'apprentissage automatique, l'analyse de données et l'éducation (Poornima *et al.*, 2022). [6]
Kaggle 	Kaggle est une plateforme en ligne qui regroupe la communauté de Data Science la plus importante au monde, Elle offre des outils et des ressources puissantes pour soutenir tous les avancées dans le domaine de la science des données. Kaggle propose un environnement Jupyter Notebooks personnalisable et sans nécessité de configuration (Robert, 2021).
Stack Overflow **(débordement de pile)** 	Un débordement de pile est une forme d'erreur de tampon qui survient lorsque le programme informatique essaie d'utiliser plus d'espace mémoire dans la pile d'appels que celui qui lui a été assigné. Le segment de pile ou pile d'appels est un tampon de taille fixe qui conserve les variables de fonction locales et renvoie les données d'adresse lors de l'exécution du programme[7].

[6] https://mail.google.com/mail/u/0/#inbox/FMfcgzGxTPDrSBvnFRzCgPWGkNqNjKTj?projector=1&messagePartId=0.4

[7] https://www.techtarget.com/whatis/definition/stack-overflow

2.3.2. Bibliothèque

La liste des bibliothèques utilisées dans notre travail est présentée dans le tableau ci -dessous, avec leur description.

Tableau 6: différentes bibliothèques python utilisées

Bibliothèque	Description
Pandas	Pandas est un logiciel open source pour l'analyse et la manipulation de données, rapide, puissant, flexible et convivial, basé sur le langage de programmation Python[8].
Numpy	NumPy est une bibliothèque open source Python qui est utilisée dans pratiquement tous les domaines de la science et de l'ingénierie. Il s'agit de l'standard universel pour l'utilisation de données numériques en Python, et il est essentiel dans les écosystèmes scientifiques Python et PyData. Les utilisateurs de NumPy englobent tous les niveaux d'expertise, allant des coders débutants aux chercheurs chevronnés qui mènent des recherches et des développements scientifiques et industriels de pointe. Dans Pandas, SciPy, Matplotlib, scikit-learn, sckit-image et la plupart des autres paquets de Python scientifique et de science des données, l'API NumPy est couramment employée[9].
Matplotlib	Matplotlib est une bibliothèque Python qui permet de concevoir des plans en 2D de matrices. Elle est basée sur l'émulation des commandes graphiques MATLAB®1, mais elle est autonome et peut être utilisée de manière pythonique, orientée objet. Malgré sa conception principalement en Python pur, Matplotlib fait largement appel à NumPy et à d'autres codes d'extension afin de garantir de bonnes performances même pour les ensembles volumineux. (Hunter *et al.*, 2017).[10]

[8] https://pandas.pydata.org/
[9] https://numpy.org/doc/stable/user/absolute_beginners.html
[10] https://drive.google.com/file/d/1aOJlfpHavmfiVMunj5rSPpxMoCx3tAPu/view

Scikit-learn (Sklearn)	La bibliothèque Scikit-learn (Sklearn) est la plus pratique et la plus solide pour l'apprentissage automatique en Python. Il propose une variété d'outils performants pour l'apprentissage automatique et la modélisation statistique, incluant la classification, la régression, le regroupement et la réduction de la dimensionnalité grâce à une interface de cohérence en Python. L'écriture de cette bibliothèque, principalement en Python, repose sur NumPy, SciPy et Matplotlib[11].
Seaborn	Seaborn Python est un outil de visualisation de données en Python qui permet de générer des graphiques statistiques. Son objectif principal est de s'adapter de manière fluide aux structures de données du module Pandas. La production des graphiques Matplotlib est automatisée par Seaborn. Ses modules permettent de générer des graphiques prêts à être exportés pour la rédaction de rapports et de publications scientifiques en suivant quelques instructions[12].

3. Méthode

Après avoir collecté et géré la base de données[13] (dataset) textuelle décrivant les dimensions morphologiques des images des micro-organismes depuis le site Kaggle[14]. Sous forme de fichier Excel **' microbes csv '**, et après avoir téléchargé ce fichier Excel sur Google Drive, qui sert d'espace de stockage dans notre compte Google Colab (fichier thèse. ipynb), on fait une série de manipulations :

3. 1 Connecter Google Colab à Google Drive

Pour connecter Google Colab à Google Drive, nous avons procédé comme suit. Tout d'abord, nous avons importé la bibliothèque nécessaire avec la commande from Google Colab import drive. Ensuite, nous avons monté Google Drive en exécutant drive. Mount ('/content/drive'). Cette commande nous a demandé d'autoriser l'accès à notre

[11] file:///C:/Users/GL%20TECH/Downloads/scikit_learn_tutorial.pdf
[12] https://www.intelligence-artificielle-school.com/ecole/technologies/seaborn-python-tout-savoir-sur-loutil-de-data-visualisation/
[13] https://data.mendeley.com/datasets/f9m85ptmvc/3
[14] https://www.kaggle.com/datasets

Google Drive. Une fois l'autorisation accordée, nous avons pu accéder à nos fichiers dans Google Drive via le chemin /content/drive/My Drive/. En suivant ces étapes, nous avons facilité l'intégration de nos données stockées sur Google Drive avec nos notebooks Colab, nous permettant ainsi de lire et écrire des fichiers facilement.

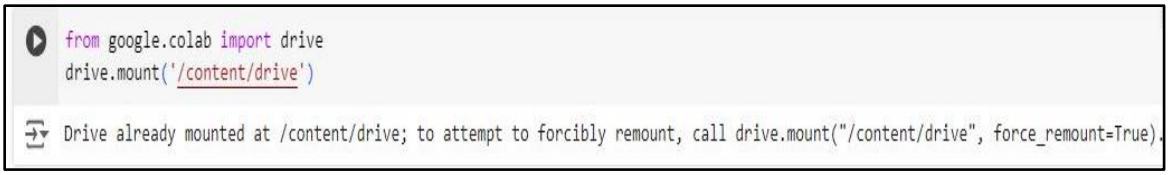

Figure 22 : code python pour connecter Google Colab à Google Drive

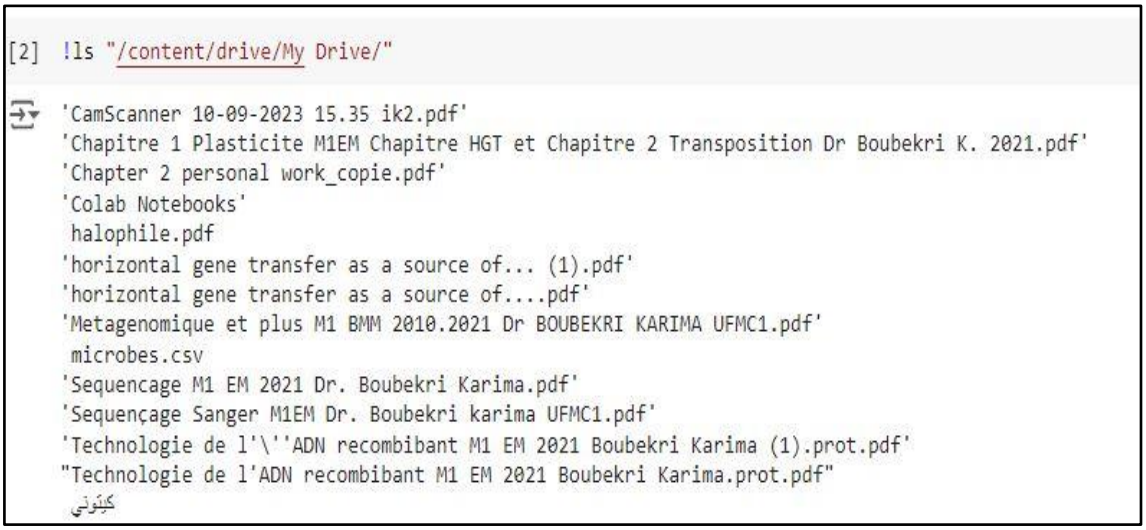

Figure 23 : Code Python permettant de stocker des fichiers de Google Drive dans Google Colab

3. 2 Lecture du Dataset

L'objectif de la phase de « lecture de dataset » dans Google Colab est d'importer et de charger les données nécessaires pour l'analyse ou le traitement. Cette étape est cruciale, car elle permet de préparer les données pour les étapes ultérieures de manipulation, d'exploration, d'entraînement de modèles ou de visualisation.

```
import pandas as pd
data=pd.read_csv('/content/drive/MyDrive/microbes.csv')
print(data.head(14))
```

```
    Unnamed: 0  Solidity  Eccentricity  EquivDiameter  Extrema  FilledArea  \
0            0     10.70          15.8           5.43     3.75       0.785
1            1      5.60          18.3           4.14     6.16       0.364
2            2      8.32          19.8           4.63     6.66       0.415
3            3     10.10          17.9           7.29    11.10       1.470
4            4      6.27          20.2          20.10    10.70      14.700
5            5      9.47          18.4           4.27    14.60       0.400
6            6     13.50          19.8           4.50    17.30       0.493
7            7     15.60          19.6           4.16    19.60       0.352
8            8      8.95          20.0           8.59     3.28       1.520
9            9     11.20          17.4           3.53     4.33       0.332
10          10     11.80          20.9           3.94     5.09       0.354
11          11      6.32          21.7          12.20     6.16       3.040
12          12      5.66          22.2           4.42     7.44       0.368
13          13      5.62          19.5          18.40    11.60       8.510

    Extent  Orientation  EulerNumber  BoundingBox1  ...  ConvexHull4  \
0     8.14       2.1500         22.3          2.97  ...         2.97
1     3.51      18.6000         22.5          5.41  ...         5.47
2     5.85      21.0000         22.4          5.96  ...         5.96
3     6.30       9.9400         21.9          8.81  ...         8.88
4     3.97       2.5800         11.9         10.20  ...        10.20
```

Figure 24 : code python pour le chargement sur mémoire et lecture de dataset

3. 3 Prétraitement des données

Le principal but de la préparation des données est de garantir que les informations relatives sont précises et cohérentes. Effectivement, les données sont fréquemment générées avec des valeurs manquantes, des erreurs ou d'autres erreurs. En outre, les groupes de données sont fréquemment enregistrés dans des fichiers ou des bases de données avec des formats distincts, ce qui nécessite donc une harmonisation. Une grande partie de la préparation des données est consacrée à la correction des erreurs et à la jointure des ensembles.[15]

Différentes fonctionnalités sont employées pour filtrer et améliorer la pratique des données. Les principales étapes que nous effectuons pour nettoyer et prétraiter sont les suivantes :

3.3.1. Suppression des valeurs nulles

Les valeurs nulles correspondent à l'absence de données dans une cellule ou un champ. Différents symboles, comme NA, NaN, ou des espaces, peuvent être utilisés en fonction de la source de données et du format. Il est possible que les valeurs nulles signifient que les données sont inconnues, inappropriées ou non applicables. Il peut aussi s'agir d'erreurs commises lors de la collecte, de la saisie ou du traitement des données. Lors de l'exploration de données, les valeurs nulles peuvent causer des difficultés, car elles

[15] https://www.lemagit.fr/definition/Preparation-des donnees#:~:text=L'un%20des%20principaux%20objectifs,inexactitudes%20ou%20d'autres%20erreurs

peuvent diminuer la taille de l'échantillon, altérer les statistiques et impacter les performances des algorithmes.[16]

L'objectif de la phase de « suppression des valeurs nulles » est de nettoyer toutes les données en supprimant les valeurs manquantes ou les points vides. Cela favorise l'amélioration de la qualité et de la précision des données pour les analyses et les modèles statistiques ultérieurs.

Nous avons utilisé la fonction dropna pour supprimer les lignes contenant des valeurs nulles dans le dataset.

```
data.dropna(inplace=True)   #NN

[ ]  print(data)     #NN

       Unnamed: 0  Solidity  Eccentricity  EquivDiameter  Extrema  FilledArea
0               0     10.70          15.8           5.43     3.75      0.7850
1               1      5.60          18.3           4.14     6.16      0.3640
2               2      8.32          19.8           4.63     6.66      0.4150
3               3     10.10          17.9           7.29    11.10      1.4700
4               4      6.27          20.2          20.10    10.70     14.7000
...           ...       ...           ...            ...      ...         ...
30522       30522      3.01          22.6           4.90    20.00      0.4340
30523       30523      5.19          22.6           2.07    19.40      0.0788
30524       30524      9.21          22.7           2.07    21.00      0.0790
30525       30525      8.21          22.6           1.87    20.50      0.0641
30526       30526      6.57          21.0           2.13    21.20      0.0840

       Extent  Orientation  EulerNumber  BoundingBox1  ...  ConvexHull4  \
0        8.14         2.15         22.3          2.97  ...         2.97
1        3.51        18.60         22.5          5.41  ...         5.47
2        5.85        21.00         22.4          5.96  ...         5.96
3        6.30         9.94         21.9          8.81  ...         8.88
4        3.97         2.58         11.9         10.20  ...        10.20
...       ...          ...          ...           ...  ...          ...
30522    1.40        19.90         22.1         18.10  ...        18.60
```

Figure 25 : code python pour supprimer les valeurs nulles

3.3.2. Suppression des valeurs en double

Une duplication ou une répétition d'une même information dans une liste, une base de données ou tout autre ensemble de données est appelée un « doublon ».

Cela implique qu'un même élément possède deux ou plusieurs entrées identiques ou très similaires. On peut rencontrer des problèmes tels que des incohérences, des erreurs de rapport, une diminution de l'efficacité et des problèmes de gestion des données en cas de doublons dans une base de données.

Le but consiste à garantir la qualité des données en évitant les répétitions et en assur ant que chaque élément est représenté de manière unique dans la base de données.[17]

[16] https://fr.linkedin.com/advice/3/what-best-way-identify-remove-null-values-data-cleaning?lang=fr
[17]https://www.callofsuccess.com/lexique/doublon#:~:text=C'est%20pourquoi%20la%20d%C3%A9tection,dans%20la%20base%20de%20donn%C3%A9es

Nous avons utilisé la fonction drop - duplicates pour supprimer les lignes contenant des valeurs en double dans le dataset

```
data.drop_duplicates() #NN
```

	Unnamed: 0	Solidity	Eccentricity	EquivDiameter	Extrema	FilledArea	Extent	Orientation	EulerNumber	BoundingBox1	...	ConvexHull4	MajorAxisLength
0	0	10.70	15.8	5.43	3.75	0.7850	8.14	2.15	22.3	2.97	...	2.97	1.340
1	1	5.60	18.3	4.14	6.16	0.3640	3.51	18.60	22.5	5.41	...	5.47	1.520
2	2	8.32	19.8	4.63	6.66	0.4150	5.85	21.00	22.4	5.96	...	5.96	1.630
3	3	10.10	17.9	7.29	11.10	1.4700	6.30	9.94	21.9	8.81	...	8.88	2.040
4	4	6.27	20.2	20.10	10.70	14.7000	3.97	2.58	11.9	10.20	...	10.20	7.780
...	...	...	...	...	...	...	...	...	...	...	...	...	...
30522	30522	3.01	22.6	4.90	20.00	0.4340	1.40	19.90	22.1	18.10	...	18.60	4.350
30523	30523	5.19	22.6	2.07	19.40	0.0788	1.67	5.09	22.8	19.20	...	20.00	1.460
30524	30524	9.21	22.7	2.07	21.00	0.0790	5.81	12.70	22.8	20.10	...	20.10	1.060
30525	30525	8.21	22.6	1.87	20.50	0.0641	5.96	11.30	22.8	20.20	...	20.20	1.010
30526	30526	6.57	21.0	2.13	21.20	0.0840	3.77	16.50	22.8	20.70	...	20.80	0.969

30527 rows × 26 columns

```
data.drop_duplicates() #NN
```

ientation	EulerNumber	BoundingBox1	...	ConvexHull4	MajorAxisLength	MinorAxisLength	Perimeter	ConvexArea	Centroid1	Centroid2	Area	raddi	microorganisms
2.15	22.3	2.97	...	2.97	1.340	1.610	0.683	0.1950	3.63	12.10	1.310	7.99	Spirogyra
18.60	22.5	5.41	...	5.47	1.520	1.520	1.010	0.2150	6.01	20.60	0.765	7.99	Spirogyra
21.00	22.4	5.96	...	5.96	1.630	1.380	1.110	0.1820	6.55	11.50	0.953	7.99	Spirogyra
9.94	21.9	8.81	...	8.88	2.040	2.120	0.715	0.3710	10.30	12.00	2.340	7.99	Spirogyra
2.58	11.9	10.20	...	10.20	7.780	6.210	6.800	4.4400	14.00	9.55	17.600	7.99	Spirogyra
...	...	...	...	...	...	...	...	...	...	...	...	...	...
19.90	22.1	18.10	...	18.60	4.350	1.350	1.710	0.5490	19.10	13.20	1.070	7.15	Ulothrix
5.09	22.8	19.20	...	20.00	1.460	0.411	0.458	0.0597	19.80	17.20	0.197	7.15	Ulothrix
12.70	22.8	20.10	...	20.10	1.060	0.257	0.319	0.0341	20.70	11.60	0.198	7.15	Ulothrix
11.30	22.8	20.20	...	20.20	1.010	0.285	0.350	0.0314	20.80	20.40	0.162	7.15	Ulothrix
16.50	22.8	20.70	...	20.80	0.969	0.642	0.385	0.0502	21.00	18.40	0.208	7.15	Ulothrix

Figure 26 : code python pour supprimer les valeurs en doubles

3.3.3. Conversion du texte en nombre avec Label Encoder

Le processus de conversion du texte en nombre dans un ensemble de données, appelé aussi « tokenization » ou « codage », vise à convertir les données textuelles en représentations numériques pour les rendre compréhensibles par les modèles d'apprentissage automatique. Il s'agit habituellement de relier chaque mot ou symbole présent dans le texte à un identifiant numérique spécifique. Cette phase revêt une

importance capitale dans le traitement du langage naturel (NLP) et d'autres opérations liées au texte dans le domaine de l'apprentissage automatique.

Nous avons utilisé la classe label Encoder de la bibliothèque scikit-learn pour convertir les variables textuelles en nombres. Cette classe permet d'attribuer chaque catégorie unique à un nombre entier.

```
from sklearn.preprocessing import LabelEncoder   #NN
# Initialize LabelEncoder
label_encoder = LabelEncoder()

# Fit and transform the 'Color' column using label encoding
data['microorganisms'] = label_encoder.fit_transform(data['microorganisms'])

# Display the transformed DataFrame
print(data)
```

```
       Unnamed: 0  Solidity  Eccentricity  EquivDiameter  Extrema  FilledArea  \
0               0     10.70          15.8           5.43     3.75      0.7850
1               1      5.60          18.3           4.14     6.16      0.3640
2               2      8.32          19.8           4.63     6.66      0.4150
3               3     10.10          17.9           7.29    11.10      1.4700
4               4      6.27          20.2          20.10    10.70     14.7000
...           ...       ...           ...            ...      ...         ...
30522       30522      3.01          22.6           4.90    20.00      0.4340
30523       30523      5.19          22.6           2.07    19.40      0.0788
30524       30524      9.21          22.7           2.07    21.00      0.0790
30525       30525      8.21          22.6           1.87    20.50      0.0641
30526       30526      6.57          21.0           2.13    21.20      0.0840

       Extent  Orientation  EulerNumber  BoundingBox1  ...  ConvexHull4  \
0        8.14         2.15         22.3          2.97  ...         2.97
1        3.51        18.60         22.5          5.41  ...         5.47
2        5.85        21.00         22.4          5.96  ...         5.96
3        6.30         9.94         21.9          8.81  ...         8.88
4        3.97         2.58         11.9         10.20  ...        10.20
...       ...          ...          ...           ...  ...          ...
```

Figure 27 : code python pour Converser du texte en nombre avec Label Encoder

3.3.4. Affichage des Coefficients de Corrélation entre les attributs

L'objectif de la phase d'affichage des coefficients de corrélation entre les attributs dans un ensemble de données est de donner des renseignements sur la relation linéaire entre les différentes variables. Cette étape permet de saisir la relation entre les variables, ce qui est crucial pour l'analyse exploratoire des données et la sélection des caractéristiques pertinentes pour les modèles d'apprentissage automatique. En montrant les coefficients de corrélation, il est possible de repérer les liens positifs, négatifs ou nulles entre les attributs, ce qui facilite la prise de décisions éclairées lors de la modélisation des données.

Nous avons utilisé les bibliothèques Seaborn et Matplotlib pour afficher les coefficients de corrélation entre les attributs de la base de données.

```python
import seaborn as sns            #NN
import matplotlib.pyplot as plt
corr_matrix = data.corr()

# Create heatmap
plt.figure(figsize=(15, 15))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix Heatmap')
plt.show()
```

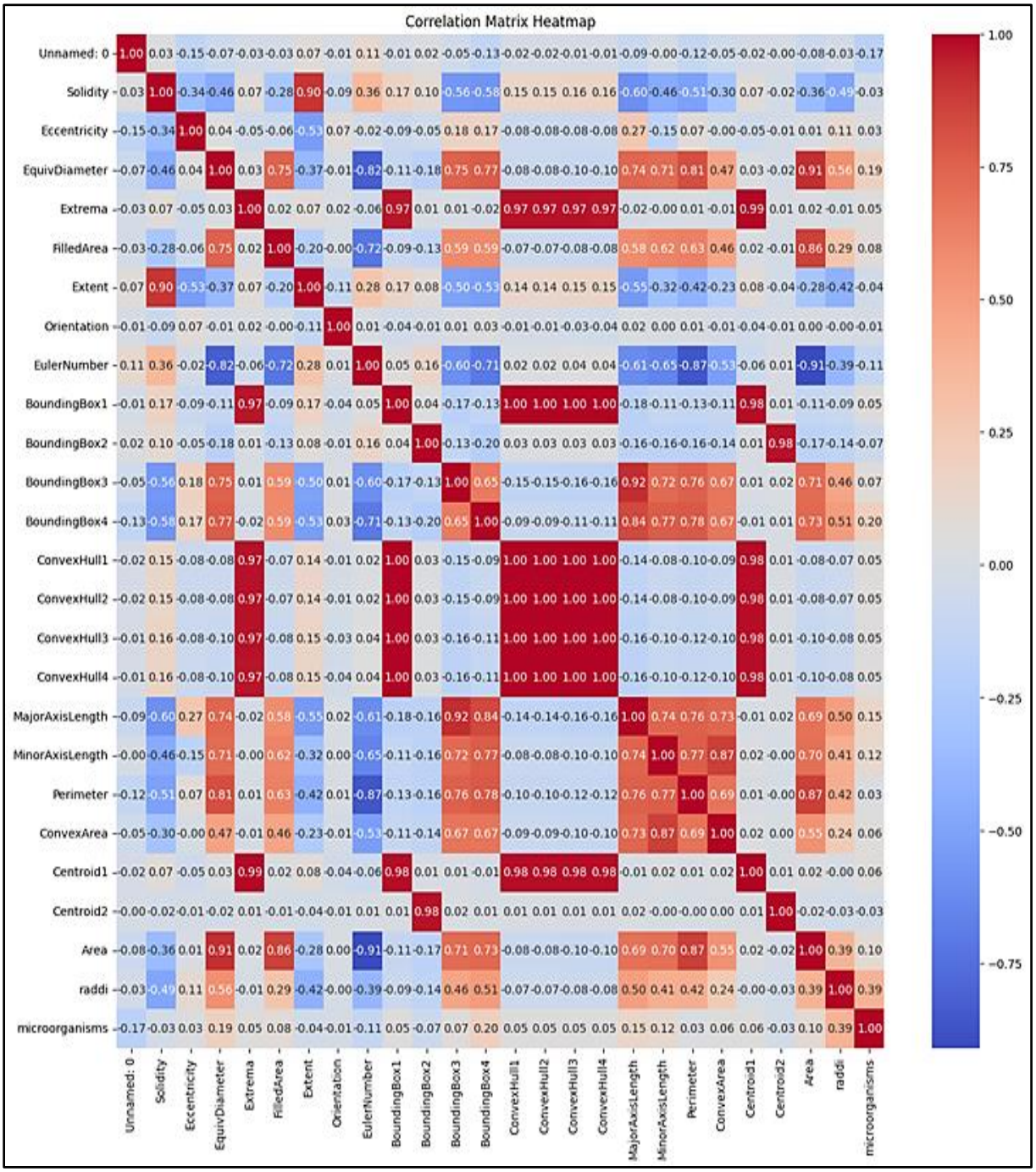

Figure 28 : code python pour afficher les coefficients de corrélation entre les attributs

3.3.5. Suppression des colonnes supplémentaires

Les colonnes supplémentaires (unnamed) sont supprimées dans un ensemble de données lorsqu'elles n'ont pas été nommées ou n'ont pas d'importance pour l'analyse. Cette étape a pour but de purifier et de simplifier le jeu de données, ce qui peut améliorer les performances des modèles d'apprentissage automatique et simplifier l'interprétation des résultats.

Nous avons utilisé la fonction dropna pour supprimer la colonne « unnamed » des attributs dans notre base de données.

```
data = data.drop("Unnamed: 0", axis='columns') #NN

print(data)
```

```
       Solidity  Eccentricity  EquivDiameter  Extrema  FilledArea  Extent  \
0         10.70          15.8           5.43     3.75      0.7850    8.14
1          5.60          18.3           4.14     6.16      0.3640    3.51
2          8.32          19.8           4.63     6.66      0.4150    5.85
3         10.10          17.9           7.29    11.10      1.4700    6.30
4          6.27          20.2          20.10    10.70     14.7000    3.97
...         ...           ...            ...      ...         ...     ...
30522      3.01          22.6           4.90    20.00      0.4340    1.40
30523      5.19          22.6           2.07    19.40      0.0788    1.67
30524      9.21          22.7           2.07    21.00      0.0790    5.81
30525      8.21          22.6           1.87    20.50      0.0641    5.96
30526      6.57          21.0           2.13    21.20      0.0840    3.77

       Orientation  EulerNumber  BoundingBox1  BoundingBox2  ...  ConvexHull4  \
0             2.15         22.3          2.97         10.90  ...         2.97
1            18.60         22.5          5.41         19.20  ...         5.47
2            21.00         22.4          5.96         10.20  ...         5.96
3             9.94         21.9          8.81         10.70  ...         8.88
4             2.58         11.9         10.20          1.22  ...        10.20
...            ...          ...           ...           ...  ...          ...
30522        19.90         22.1         18.10          9.92  ...        18.60
30523         5.09         22.8         19.20         16.20  ...        20.00
30524        12.70         22.8         20.10         11.40  ...        20.10
30525        11.30         22.8         20.20         20.20  ...        20.20
30526        16.50         22.8         20.70         18.00  ...        20.80
```

Figure 29 : code pour supprimer la colonne supplémentaire (unnamed)

3. 4 Apprentissage

3.4.1. Séparation des caractéristiques d'entrée et des étiquettes de sortie

Le processus d'analyse des données consiste à séparer les caractéristiques d'un ensemble de données en deux ensembles distincts, les caractéristiques d'entrée et les caractéristiques de sortie (résultats). Les variables qui seront utilisées pour prédire ou analyser d'autres variables peuvent être séparées par cette division. Les données d'entrée sont celles qui servent à prédire ou à analyser, tandis que les données de sortie sont les

résultats que l'on cherche à prédire ou à comprendre. Il est crucial de faire cette division afin de modéliser les données et de développer des algorithmes d'apprentissage automatique efficaces.

Nous identifions les caractéristiques d'entrée en sélectionnant toutes les colonnes sauf la cible, et nous identifions les étiquettes de sortie en sélectionnant uniquement la colonne cible.

```
# Sample dataset (features and labels)
y = data["microorganisms"]  # Binary classification labels
X = data.drop("microorganisms", axis='columns') #NN
```

Figure 30 : code python pour séparer les caractéristiques d'entrée et les étiquettes de sortie

3.4.2. Division des données en ensembles d'apprentissage

En utilisant la fonction Train_test_split, cette méthode permet de séparer les données de manière aléatoire en ensembles d'apprentissage et de test. 80% (24421 attributs) des données sont des données d'apprentissage, tandis que 20% (6105 attributs) sont des données de test. La taille de l'ensemble de Test est indiquée par l'argument test_size.

```
# Split the data into training and testing sets (80% train, 20% test)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Figure 31 : code python effectuer la diffusion du dataset (ensemble pour l'entrainement et l'ensemble pour le test)

3.4.3. Création du modèle

Un modèle de classification qui associe à la fois la forêt aléatoire et l'arbre de décision est développé en instanciant la classe "Random Forest Classifier" et la classe "Decision Tree Classifier". Par la suite, en raison de ses avantages. Les algorithmes de forêt aléatoire et d'arbre de décision sont largement utilisés dans de nombreux domaines pour leur efficacité et leur polyvalence. La forêt aléatoire combine plusieurs arbres de décision pour fournir des prédictions plus précises et plus robustes, réduisant ainsi le risque de sur apprentissage. D'autre part, l'algorithme de l'arbre de décision offre une interprétabilité élevée, ce qui signifie qu'il est facile de comprendre les décisions prises par le modèle. En outre, ces deux méthodes sont capables de gérer des ensembles de données de grande taille avec une grande efficacité, ce qui en fait des outils précieux pour la modélisation prédictive et l'analyse de données.

Par la suite, le modèle est adapté aux informations d'apprentissage en utilisant la méthode "fit". Le modèle est entraîné en utilisant des ensembles d'apprentissage, ce qui lui permet d'apprendre à partir des données et de trouver les liens entre les attributs non résultats et les attributs résultats du microorganisme.

```
# Initialize the Random Forest Classifier
clf = RandomForestClassifier(random_state=42)

# Fit the classifier to the training data
clf.fit(X_train, y_train)
```

Figure 32 : code et modèle de classification Random Forest

```
# Initialize the Decision Tree Classifier
clf = DecisionTreeClassifier(random_state=42)

# Fit the classifier to the training data
clf.fit(X_train, y_train)
```

Figure 33 : code et modèle de classification Decision Tree

a) Prédiction

La phase de prédiction dans le contexte des données consiste à utiliser des modèles statistiques ou d'apprentissage automatique pour estimer ou anticiper des valeurs cibles ou des classes pour de nouvelles observations. L'objectif principal de cette phase est de fournir des prédictions précises et fiables pour les données non observées, en utilisant des modèles entraînés sur des données historiques. Cela permet de prendre des décisions éclairées, d'identifier des tendances ou des schémas cachés dans les données.

Le modèle sert à prédire les données de test en employant la méthode 'predict'.

```
# Predict using the test set
y_pred = clf.predict(X_test)
```

Figure 34: code pour faire des prédictions sur les données de test Random Forest

```
# Predict using the test set
y_pred = clf.predict(X_test)
```

Figure 35 : code pour faire des prédictions sur les données de test Decision Tree

b) Evaluation

La phase d'évaluation, dans le contexte des données, consiste à évaluer la performance des modèles de prédiction développés sur des données invisibles ou non utilisées lors de l'entraînement. L'objectif principal de cette phase est de mesurer à quel point les modèles sont capables de généraliser et de faire des prédictions précises sur de nouvelles données. Cela permet de déterminer si les modèles sont suffisamment fiables pour être déployés dans des applications réelles et de comparer différentes approches pour choisir la meilleure. En outre, l'évaluation aide à identifier les éventuels problèmes tels que le sur apprentissage ou le sous-apprentissage, ce qui peut conduire à des ajustements ou des améliorations des modèles.

Les mesures d'évaluation du modèle sont calculées en utilisant les étiquettes réelles 'y_test ′et les étiquettes prédites 'y_ pred′.

```
# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

Figure 36 : code de calcule des performances du modèle Random Forest

```
# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
```

Figure 37 : code de calcule des performances du modèle Decision Tree

Dans le dernier chapitre, nous exposerons les résultats obtenus pour démontrer l'efficacité des étapes de ce travail appliqué en mettant en évidence le degré d'apprentissage de chaque modèle et en les comparants. Ainsi, nous évaluerons leur efficacité et leur capacité à être utilisés dans la classification des micro-organismes.

CHAPITRE 02:

Résultats et Discussion

1. Validation et vérification

Nos recherches ont produit des résultats prometteurs, segmentés en deux parties distinctes : dataset et modèles d'apprentissage automatique.

1.1. Validation et vérification de dataset

La première partie a impliqué l'acquisition d'un dataset comprenant globalement dix genres de microorganismes. Ces micro-organismes inclus étaient les levures, les algues et les protozoaires avec différents genres. Cette dataset contient 26 attributs (colonnes) et 30527 observations (lignées ou individus) classés les microorganismes selon les dimensions des formes dans l'image qui représentent la morphologie des microorganismes. Le dataset textuel choisi pour notre étude est considérée comme l'un des types les plus précis.

Tout d'abord, la sélection appropriée de dataset car le prétraitement des données a été efficace, grâce à l'absence de doublons et de valeurs nulles ; parce que ces derniers peuvent entrainer des biais, perturber les algorithmes et fausser les résultats des analyses statistiques, des modelés prédictifs et des visualisations de données. Ainsi qu'une taille de dataset est conséquente. Ce qui nous permet d'éviter les problèmes de sous apprentissage et sur apprentissage.

1.2.Validation et vérification des modèles d'apprentissage automatique

Dans la deuxième partie, après l'application de deux algorithmes ; les forêts aléatoires (Random Forest) et l'arbre de décision. Random Forest a démontré un score d'apprentissage de 98,79 %, tandis que l'arbre de décision a atteint 98,72 %. Ces résultats indiquent une performance remarquable des modèles d'apprentissage automatique exploités.

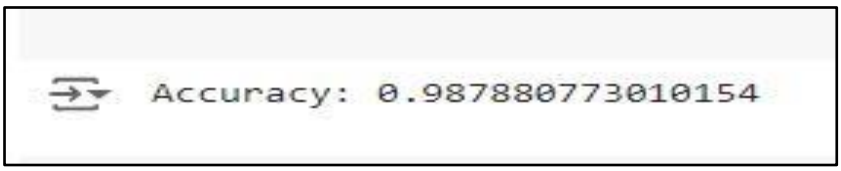

Figure 38 : Le degré d'apprentissage de Random Forest

Figure 39 : Le degré d'apprentissage d'arbre de décision

Grâce aux résultats obtenus ci-dessus et classés dans la figure 1 et 2, nous constatons une simple différence de valeurs entre les deux algorithmes, car le degré d'apprentissage de Random Forrest(les forêts aléatoires) est plus élevé que celui de Decision Tree (l'arbre de décision) en raison de:

En premier lieu, les forêts aléatoires ont la capacité de gérer de manière efficace des ensembles de données complexes avec de multiples caractéristiques, ce qui est fréquent avec des données d'images. Ensuite, ils diminuent le risque d'ajustements incorrects par rapport à un seul arbre de décision, ce qui revêt une importance capitale pour la généralisation des modèles. Ensuite, les forêts aléatoires ont la capacité d'évaluer automatiquement l'importance des différentes caractéristiques, ce qui peut être bénéfique pour repérer les caractéristiques descriptives les plus pertinentes dans la classification des algues et des champignons à partir d'images. Enfin, ils ont une sensibilité réduite aux valeurs aberrantes, ce qui renforce la résistance du modèle face à des données bruitées ou incohérentes, contrairement à l'arbre de décision.

Pour confirmer cela, nous avons évalué la prédiction en comptant le nombre d'individus (30527) avec une prédication juste et le nombre d'individus avec une prédiction erronée pour les deux algorithmes et nous avons obtenu les résultats décrits dans le tableau ci-dessus.

Tableau7 : évaluation de prédiction

	Accuracy (Score d'apprentissage)	**Nombre des individus (lignes) avec prédiction juste**	**Nombre des individus (lignes) avec prédiction erronée**
Random Forest	98,79 %	6031	74
Decision Tree	98,72 %	6027	78

2. Situation de notre travail parmi les travaux connexes

En comparant notre travail avec (Jing., 2022) et (Shallu *et al*., 2022). Nous pouvons prendre en compte déférents critères tels que : Taille et contenu de dataset, les modèles d'apprentissage automatique exploités et la précision (accuracy).

Tableau 8 : comparaison du notre travail avec un autre approche

	Notre travail	**(Shallu et *al.*, 2022)**	**(Jing., 2022)**
Taille de dataset	30527 lignes 26 attributs	100 lignes 48 attributs	30500 lignes 15 attributs
Types des données (images / texte)	texte	Image	texte
Contenu de dataset	attributs descriptifs des dimensions des formes qui remplacent les images des microorganismes (algues, champignonset protozoaires)	la classification des bactéries urinaires. Ils 100 images des bactéries	Classification 12 genres des microorganismes (*Spirogyra*, *Volvox*, *Pithophora*, Yeast, *Rhizopus*, *Penicillum*, *Aspergillus sp*, Protozoa, Diatom, *Ulothrix*)
Modèle utilisée	Arbre de décision Forêt aléatoire (Random Forest)	ANN (réseau de neurones artificiel)	K-nearest neighbors, Logistic regression, Random forests, and Gradient boosted machine.
Accuracy (Score d'apprentissage)	Arbre de décision 98,79 % Forêt aléatoire 98,72 %	76 %	90 %

Dans notre travail, la taille est la plus grande. La taille d'un ensemble de données, tel que celui de 30 527 lignes, est cruciale pour la précision des algorithmes d'apprentissage automatique. Un ensemble de données plus vaste offre une richesse d'informations permettant aux algorithmes d'apprendre plus efficacement et de généraliser mieux. Cela réduit les risques d'erreurs dues au sur apprentissage ou au sous-apprentissage, améliorant ainsi la précision et la fiabilité des modèles prédictifs.

La nature textuelle d'une base de données joue un rôle essentiel dans la précision des algorithmes d'apprentissage automatique. Les données textuelles permettent d'éliminer une grande tâche de traitement des images car la qualité des images influe sur la précision d'entrainement et de la prédiction des systèmes d'apprentissage automatique. Pour ce là, plusieurs techniques doivent être exploitées pour améliorer la qualité des images.

Nous avons développé un modèle d'intelligence artificielle basé sur l'apprentissage automatique, utilisant deux algorithmes : la forêt aléatoire et l'arbre de décision. Ces algorithmes ont démontré une haute performance d'apprentissage avec une précision de 98%≃ 99 %.

Comparativement à d'autres études qui ont utilisé des différents algorithmes de classification, nos résultats sont significativement supérieurs. En effet, ces autres travaux ont obtenu des taux de précision de 76% et 90%.

Le choix judicieux de l’algorithme d'apprentissage automatique, comme la forêt aléatoire et l'arbre de décision, est crucial pour la précision des algorithmes. La forêt aléatoire, avec sa robustesse et sa capacité à réduire le sur apprentissage, et l'arbre de décision, avec sa simplicité et son interprétabilité, sont essentiels pour améliorer les performances et la précision des modèles.

CONCLUSION ET PERSPECTIVE

L'objectif principal de notre travail est de développer un modèle d'apprentissage automatique capable de classer avec précision des microorganismes. Nous avons commencé notre travail par une recherche d'un dataset (base de données) qui étudie la description morphologique et taxinomique des microorganismes basée sur leurs images. Cette description morphologique a été extraite à partir des dimensions des formes des microorganismes sur leurs images. Cette analyse a permis de représenter les principales différences entre ces genres et leur morphologie. Ce dataset présente la description de dix genres de microorganismes (levures, *Spirogyra, Volvox, Pithophora, Rhizopus, Penicillum, Aspergillus, protozoaires, diatomées, Ulothrix*) et il regroupe 30 527 micro-organismes.

Le modèle d'apprentissage automatique développé exploite ce dataset et il se base sur des méthodes comme les arbres de décision Random (Decision Tree) et les forêts aléatoires (Forest Forest), constitue une méthode efficace et prometteuse pour appréhender et caractériser la multi-biodiversité. Les arbres de décision permettent de comprendre facilement les critères de classification, tandis que les forêts aléatoires augmentent la résistance et la précision en combinant les prédictions de plusieurs arbres un dataset.

Les résultats obtenus ont été très promoteur avec une précision de classification très élevée. Ces résultats démontrant clairement l'efficacité et la capacité de notre modèle.

Ce modèle peut classifier efficacement les microorganismes en fonction de leurs propriétés biologiques.

Notre travail représente une avancée significative dans la microbiologie car il permet de traiter de vastes ensembles de descriptions morphologiques et d'identifier des schémas et des relations complexes entre les caractéristiques des microorganismes. Notre travail offre des outils puissants pour la taxonomie microbienne, ouvrant la voie à de nouvelles découvertes dans les domaines de la santé, de l'environnement et de la biotechnologie et la préservation de l'environnement.

Comme perspectives, nous visons à exploiter des algorithmes de machine learning (apprentissage automatique) avancés afin de détecter des modèles, des corrélations et des interactions entre les microorganismes. Ce qui contribue à une meilleure compréhension de leur écologie, de leur évolution et de leurs rôles dans les écosystèmes.

De plus, le Machine Learning joue un rôle crucial dans la lutte contre les maladies infectieuses en permettant la détection rapide et précise de pathogènes, le développement de nouveaux antibiotiques et la prédiction des épidémies. Les applications biotechnologiques

bénéficient également du Machine Learning, notamment dans l'optimisation des processus de fermentation, la découverte de nouvelles enzymes et la conception de biocatalyseurs.

Dans le domaine de la médecine personnalisée, le Machine Learning peut être utilisé pour analyser le microbiote humain et prédire les réponses individuelles aux traitements, ouvrant ainsi la voie à des thérapies plus ciblées et efficaces. Enfin, le Machine Learning contribue à la préservation de l'environnement en permettant la surveillance et la gestion des écosystèmes microbiens, notamment dans le domaine de la dépollution, de la gestion des déchets et de la conservation de la biodiversité.

En somme, le Machine Learning offre un potentiel considérable pour révolutionner notre compréhension et notre utilisation du monde microbien, avec des implications majeures pour la santé, l'environnement et la biotechnologie.

REFERENCES BIBLIOGRAFIQUE

Liste des références

Chapitre 01

Abdelhadi, A.et Boukhroufa, F.Z. (2011) ‘Pathologies humaines causée par 1 les espèces du genre Aspergillus. Mémoire de Fin D'étude Pour L'obtention Du Diplôme Des Etudes Supérieures en biologie. algerie : universite jijle,p 70.

Baker, A.L. *et al.* (2024) Phycokey - Pithophora. [en ligne] (page consultee le 23/05/2024). https://cfb.unh.edu/phycokey/Choices/Chlorophyceae/filaments/branched/PITHOPHORA/Pithophora_key.html

Bensalem, O.et HORCHI, D. (2020) ‘Contribution à l’étude de la production de cellulase levurienne par fermentation en milieu solide à base de déchets d’agrumes. Mémoire présenté en vue de l’obtention du Diplôme de Master en biochimie.algerie : Université des Frères Mentouri Constantine 1 Faculté des Sciences de la Nature et de la Vie, p 68.

Bouchoukh, I. (2016) ‘Cours de Botanique. lieu d' edition constantine, p 130.

Boukhedenna, N. et Merouane, I. (2013) ‘Production de la pénicilline V et G in vitro par*Penicillium chrysogenum’.* Mémoire présenté en vue de l’obtention du Diplôme de Master en microbiologie. algerie : Université des Frères Mentouri Constantine 1 Faculté des Sciences de la Nature et de la Vie,p 75.

Bousseboua, H. (2005) DEFINITIONS, CLASSIFICATION ET NOMENCLATURE DES BACTERIES [en lign].(page consultee le 18/05/2024). http://www.microbes-edu.org/etudiant/intro.html#:~:text=La%20taxonomie%20est%20l'ensemble,relations%20ph%C3%A9n%C3%A9tiques%20et%2Fou%20phylog%C3%A9n%C3%A9tiques.

Bousseboua, H. (2005) *Element de MICROBIOLOGIE* : microorganisme. editeur. lieu d;edition.p : 3

Chavez, C.M.; Groenewald3,M.; Hulfachor A. B.; Kpurubu,G.; Huerta, R.; Hittinger,C.T.et Rokas, A. (2024) ‘The cell morphological diversity of *Saccharomycotina* yeasts’. *FEMS Yeast Research*, 24. p : 1-9. https://doi.org/10.1093/femsyr/foad055.

Cruzel, M.. (2020) ‘Les différentes familles de micro-organismes’. [en ligne] (page consultee le 20/05/2024). https://sa.maxime-cruzel.fr/sa_cap/co/module_.html.

Delarue, M. (2011) ‘Volvox, structure et cycle de vie d’une algue verte originale’. Plante Vie [en ligne], (page consultee le 21/05//2024). https://planet-vie.ens.fr/thematiques/vegetaux/anatomie-vegetale-et-histologie-vegetale/volvox-structure-et-cycle-de-vie-d.

Doghmani, N. , Rezaiguia, D. et Ferdi, K. (2022) ‘Recherche des moisissures dans les aliments des ruminants’. Mémoire de Master en parasitologie. algerie : UNIVERSITE 8 MAI 1945 GUELMA, p 103.

Dolatabadi *et al*. (2014) ‘Rhizopus’.[en ligne] (page consultee le 21/05/2024). https://www.adelaide.edu.au/mycology/rhizopus.

Drouet, E. (2012) ‘Le monde Microbien. Université Joseph Fourier, Grenoble, p 58.

Guedon, E. (2019). Historique et description biologique des microorganismes commensaux et pathogenes traitements conventionnels des pathologies microbiennes et leurs limites . Université de lorraine, nancy, p 25

Guermazi, W. (2017) ‘Diversité des Parazoaires aux protostomiens’. L1, Universite De Gabes, tunisie, p 23.

Lavoie, I., Campeau, S., Grenier, M., Dillon , P. J.et Hamilton, P.B. (2008) Guide d’identification des diatomées des rivières de l’Est du Canada. Morphologie des diatomees. societe de developpement des entreprises culturelles (SODIC). Canada. p : 2-3

https://books.google.fr/books?id=FmfNonZIKT8C&printsec=frontcover&hl=fr#v=onepage&q&f=false.

Le microbe sous toutes ses formes (2017) *rts.ch*. Available at: https://www.rts.ch/decouverte/sante-et-medecine/corps-humain/microbes/9006414-le-microbe-sous-toutes-ses-formes.html (Accessed: 18 May 2024).

Lokhorst, G.M. et Vroman, M. (1972) 'Taxonomic study on three freshwater Ulothrix species'. Acta 801. Neerl. 21(5). p : . 449-480

Lor, B.;Zohn, M.; Meade, M.J.; Cahoon,A.B.; Manoylov, K.M. (2021) 'A Morphological and Molecular Analysis of a Bloom of the Filamentous Green Alga Pithophora', *Water*, 13(6), p. 760. https://doi.org/10.3390/w13060760.

Madigan, M.T. et Martinko, J.M. (2010) *(PDF) Brock Biology of Microorganisms (11th edn).* INTERNATIONAL MICROBIOLOGY (2005) 8:149-152 www.im.microbios.org https://www.researchgate.net/publication/41584156_Brock_Biology_of_Microorganisms_11th_edn_Michael_T_Madigan_John_M_Martinko_eds (Accessed: 22 May 2024).

Makhlouf, J. (2019) Caractérisation de la biodiversité des souches d'Aspergillus de la section Flavi isolées d'aliments commercialisés au Liban*:* approche moléculaire, métabolique et morphologique. these En vue de l'obtention du DOCTORAT en Pathologie, Toxicologie, Génétique et Nutrition.france : L'UNIVERSITÉ DE TOULOUSE, p 142.

Nee l esh (2016) 'Ulothrix: Occurrence, Features and Reproduction', *Biology Discussion*, 24 August. Available at: https://www.biologydiscussion.com/algae/ulothrix-occurrence-features-and-reproduction/46848 (Accessed: 21 May 2024).

Rahmani, S. (2019) Polycopié de cours de Microbiologie Générale. L2, Université Hassiba Ben Bouali, Chlef, p 119. Available at: https://www.studocu.com/row/document/universite-mohammed-premier-oujda/microbiologie/microbiologie-cours-tp/47795760 (Accessed: 22 May 2024).

Robert, G. et Yaeger (1996) 'Protozoaires : structure, classification, croissance et développement', in Microbiologie médicale. Baron, S. Texas. 4e édition.

Ronin, A. (2020) Quelle représentation les enfants de maternelle ont-ils des microbes ? Une action pédagogique peut-elle modifier cette vision ? . Mémoire de fin de cycle en vue de l'obtention du master. Université de Franche-Comté, Besançon en France, p 65

Rumeau, A. et Coste, M. (1988) 'Initiation à la systématique des diatomées d'eau douce. Pour l'utilisation pratique d'un indice diatomique générique', Bulletin Français de la Pêche et de la Pisciculture, (309), p : 1–69. Available at: https://doi.org/10.1051/kmae:1988009.

Wongsawad, P. et Peerapornpisal, Y. (2015) 'Morphological and molecular profiling of Spirogyra from northeastern and northern Thailand using inter simple sequence repeat (ISSR) markers', Saudi Journal of Biological Sciences, 22(4), p : 382–389. Available at: https://doi.org/10.1016/j.sjbs.2014.10.004.

Chapitre 02

Aiboud, L. et Laskri, S. (2020) Appréciation de la qualité des leads dans le marketing numérique à l'aide de l'apprentissage profond. Mémoire de fin d'études En vue de l'obtention du diplôme de Master en Informatique. Algérie : Université Mouloud Mammeri de Tizi-Ouzou, p 82.

Bastien, L. (2024) Machine Learning et Big Data : définition et explications de la combinaison. [En ligne] (Page consultée le 22/05/2024). https://www.lebigdata.fr/machine-learning-et-big-data.

Batta, M. (2018) 'Machine Learning Algorithms -A Review'. [En ligne] (Page consultée le 23/05/2024). https://www.researchgate.net/publication/344717762_Machine_Learning_Algorithms_-A_Review

Chapelle, O., Scholkopf, B. et Zien, Eds., A. (2009) 'Semi-Supervised Learning. [En ligne] (Page consultee le22/05/2024). [Book reviews]', IEEE Transactions on Neural Networks, 20(3), p: 542–542. https://doi.org/10.1109/TNN.2009.2015974.

Gullitti, T. et LLC, R.B. (2017) 'Application Of Machine Learning Algorithms To On-Board Diagnostics (Obd Ii) Threshold Determination'.

Harfi, R. (2020) Amélioration des forêts aléatoires pour la classification des données médicales. : Mémoire Présenté En Vue De L'obtention Du Diplôme De Master Intelligence Artificielle Et Traitement De l'Information. Algérie : université Badji Mokhtar -Annaba. p 118.

Jérémy, R. (2020) Machine Learning : Définition, fonctionnement, utilisations. [En ligne] (Page consultée le 18/05/2024) https://datascientest.com/machine-learning-tout-savoir

Keldenich, T. (2022) 'Arbre de Décision Comment l'Utiliser - Meilleur Tutoriel'. [En ligne] (Page consultée le 22/05/2024) https://inside-machinelearning.com/arbre-decision/.

Khushaktov, M.F. (2023) 'Introduction Random Forest Classification By Example', Medium. [En ligne] (Page consultée le 22/05/2024) https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91.

Lesel (2016) 7-Structure de l'apprentissage par renforcement. | Download Scientific Diagram. [En ligne] (Page consultée le 18/05/2024) https://www.researchgate.net/figure/Structure-de-lapprentissage-par-renforcement_fig26_292906935.

Martin, D. (2023) Machine Learning et Big Data : explication détaillée et utilisation. [En ligne] (Page consultée le 18/05/2024) https://fr.linkedin.com/pulse/machine-learning-et-big-data-explication-d%C3%A9taill%C3%A9e-martin-delattre.

Mattei, P.-A. et Villata, S. (2022) Introduction à l'intelligence artificielle et aux modèles génératifs. Université Côte d'Azur, Inria. . Bruno Martin; Sara Riva. Informatique Mathématique: Une photographie en 2022, CNRS Editions, 2022. ffhal-03849387fp 30.

Matteis, L.D.; JANNY, S.; NATHAN, S.; QUARTIER, W.S. (2022) 'Introduction à l'apprentissage automatique'. paris. p. 18.

Moez, A. (2022) Supervised Machine Learning. [En ligne] (Page consultée le 05/06/2024). https://www.datacamp.com/blog/supervised-machine-learning

Padala, V.S., Gandhi, K. et Dasari, P. (2019) 'Machine Learning: The New Language for Applications', IAES International Journal of Artificial Intelligence (IJ-AI), 8(4), p. 411. Available at: https://doi.org/10.11591/ijai.v8.i4.pp411-421.

Raphael, K. (2022) 'Algorithme de classification : Définition et principaux modèles', Data Science, 22 November. Available at: https://datascientest.com/algorithme-de-classification-definition-et-principaux-

modeles#:~:text=Une%20classification%20supervis%C3%A9e,a%20pas%20de%20classes%20pr%C3%A9d%C3%A9finies.

Simon, K. (2016) L'apprentissage automatique, ou comment les ordinateurs apprennent à partir des données, Decideo - Actualités sur le Big Data, Business Intelligence, Data Science. Available at: https://www.decideo.fr/L-apprentissage-automatique-ou-comment-les-ordinateurs-apprennent-a-partir-des-donnees_a8338.html (Accessed: 27 May 2024).

Singh, A., Thakur, N. et Sharma, A. (2016) 'A review of supervised machine learning algorithms', in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). Available at: https://www.infona.pl//resource/bwmeta1.element.ieee-art-000007724478 (Accessed: 7 June 2024).

Soudoplatoff, S. (2018) L'intelligence Artificielle : L'expertise Partout Accessible À Tous.

Talbi, I. (2019) Panorama de l'apprentissage supervisé. [En ligne] (Page consultée le 18/05/2024).https://larevueia.fr/panorama-de-lapprentissage-supervise/

Talbi, I. (2020) Qu'est-ce que l'apprentissage par renforcement ?, La revue IA. [En ligne] (Page consultée le 05/06/2024). https://larevueia.fr/apprentissage-par-renforcement/

Tantuğ, A. C. ve Türkmenoğlu, C. (2015). Türkçe Metinlerde Duygu Analizi. Yüksek lisans tezi, İstanbul Teknik Üniversitesi, İstanbul.

Chapitre 03

Benmansour, a. (2018) 'Éléments de base du langage Python', in, p. 23. Available at: file:///C:/Users/GL%20TECH/Downloads/Chapitre1_Version_PDF.pdf.

Haché, D. (2003) 'Excel 1 les notions de base (Notes tirées en partie de « Microsoft Office 2000 Premium »)', in. Available at: file:///D:/memoire/PDF/3232.pdf.

Chapitre 04

Jing Lee (2022) « Classifying microorganisms using machine learning » Project 4: Classification proposal [en ligne] https://github.com/lee-jin81/metis_project_4_classification

Shallu Kotwal, Priya Rani, Tasleem Arif, Jatinder Manhas, et Sparsh Sharma (2022) « Automated Bacterial Classifications Using Machine Learning Based Computational Techniques: Architectures, Challenges and Open Research Issues » Arch Comput Methods Eng. ; 29(4): 2469–2490.

Année universitaire : 2023-2024	**Présenté par : CHIKHI Rania** **ZITOUNI Ikram**

Développement d'un modèle de machine learning (apprentissage automatique) pour la classification automatique des organismes microscopiques

Mémoire pour l'obtention du diplôme de Master en :

microbiologie appliquée

Domaine : Sciences de la Nature et de la Vie

Département : microbiologie

Ce travail consiste à développer des modèles d'apprentissage automatique pour classifier des genres d'organismes microscopiques, en se basant sur les caractéristiques morphologiques extraites à partir des dimensions des formes des microorganismes sur des images. L'objectif principal est de développer un modèle permet de classer les microorganismes avec précision. Les étapes principales pour le développement de ce modèle comprennent la collecte et le traitement des données (dataset), l'exploitation de deux algorithmes d'apprentissage automatique (l'algorithme de forêt aléatoire et l'algorithme de l'arbre de décision), ainsi que l'entraînement et l'évaluation de ce modèle. La performance de notre modèle est évaluée en utilisant plusieurs mesures telles que la précision. Les résultats ont montré que le modèle proposé avec ses algorithmes était capable de classer avec précision les organismes microbiens, avec un taux d'apprentissage supérieur à 98 %.

Mots-clefs : Microorganisme, Apprentissage Automatique, Foret Aléatoire, Arbre De Décision.

Président du jury : **Dr Abdelaziz Ouidad** (MCB- U Constantine 1 Frères Mentouri).

Encadrant : **Dr Djama Ouahiba** (MCB - U Constantine 1 Frères Mentouri).

Examinateur(s) : **Dr Chabbi Rabah** (MAA - U Constantine 1 Frères Mentouri).