Datascientest

Projet – Reconnaissance de champignons

Introduction

Le projet porte sur la classification d’espèces de champignons. Issu d’un entrainement où le modèle apprend à reconnaitre des espèces précises de champignons, ce dernier doit être en mesure de classifier de nouvelles observations soumises.

L’algorithme utilise la computer vision pour s’entrainer et pour classifier les images.

Il n’y a pas d’objectif de précision fixé à terme, mais nous visons un taux de précision suffisamment élevé sur une catégorie sélectionnées d’environ dix espèces de champignons. Ces dix espèces ont été sélectionnées en fonction des données chargées pour le modèle ; elles correspondent aux espèces les plus observées du jeu de données.

Expertise des membres du projet

Guillaume Cadet :

Une seule et unique expérience en classification d’images. J’ai testé des modèles générés par Edge Impulse et portés sur des microcontrôleurs embarqués. Aucune connaissance préalable nécessaire pour utiliser le service Edge Impulse. Le travail consistait à entrainer un modèle en fournissant un ensemble d’images, puis récupérer le code généré par Edge Impulse.

Projets personnels :

Projet 1 : Caméra embarquée sur un train miniature (tiny ML).   
Objectifs :

* Localisation par reconnaissance d’objet dans le paysage le long de la voie
* Prise de décision lors de la lecture de la signalisation
* Système connecté au cloud (IoT)

Projet 2 : robotique, exploration et découverte de l’environnement, reconnaissance d’objet et d’animaux domestique.

Guillaume Pot :

Pas d’expérience significative en matière de classification d’images. Le niveau d’expertise dans ce type de projet est un nouveau passage théorique. Concernant le niveau d’expertise en matière de connaissances et compétences à apporter, il est relatif au taux d’avancement dans la formation réalisée actuellement au sein de DataScientest.

En matière de gestion de projet, environ 5 ans d’expérience sur divers projets comprenant l’introduction de la problématique, le chiffrage du projet, la recherche de moyens et la mise en œuvre du dit projet jusqu’à sa mise en production au sein d’un environnement de travail 24h-24h 7j-7j.

Connaissance de projets similaires :

**1/** Développement du train autonome au sein de la direction de l’ingénierie de la S.A SNCF. A travers un ensemble de capteurs et d’informatique embarquée, le projet train autonome a pour but d’aboutir à la mise en service commercial (transport de voyageurs) de trains à conduite sans agent, en premier lieu sur les dessertes fines du territoire (“petites lignes”). Ce projet est en cours de réalisation au sein de la DIR ingénierie de la société mère SNCF.

**2/** Application mobile *PlanNet* destinée à la reconnaissance de différents végétaux. L’application consiste en la soumission de clichés de végétaux (arbres, fleurs, etc.). L’algorithme analyse le cliché et soumet des propositions concernant la spécificité de la plante (identification, origine, lieu). Un pourcentage indique les probabilités de résultat par rapport à l’image soumise. Il y a un aspect collaboratif derrière où il est possible de voter (upvote/downvote) et de permettre à l’algorithme d’apprendre des nouvelles données soumises. Cette application mobile est en service.

Concernant la réalisation de notre projet, l’application *PlanNet* se rapproche très fortement de ce que nous souhaitons réaliser et peut servir de base d’inspiration. Notre projet ne contribuera pas immédiatement à améliorer les deux projets cités ci-haut mais permettent d’étendre les connaissances et la compréhension que nous avons sur ces sujets.

Justine Mialhe :

Au cours de mes études j’ai analysé plusieurs images avec le logiciel *ImageJ*, mais ces analyses ne demandaient pas un travail de programmation pour la reconnaissance d’image.

Projets universitaires :

Analyse d’images à l’aide du logiciel *ImageJ* afin de caractériser les déformations d’un matériau soumis à des contraintes mécaniques directes (traction, flexion etc.), ainsi qu’à des variations de la température et de l’humidité environnementales.

Tâches effectuées : Modification des couleurs (intensité et longueur d’onde) pour mettre en relief certaines zones des images, mesure de distances entre deux points afin d’évaluer certaines dimensions et de les comparer à d’autres.

Partie 1 – Contexte et analyse des données

Choix du Dataset

Pour atteindre les objectifs de notre projet, nous utilisons la base de données de *Mushroom Observer.* Initialement, cette base contient plus de 600 000 images de champignons qui nous permettront d’entraîner le modèle et de le tester. En sus, il est possible d’enrichir la base à l’aide d’images sourcées d’internet (ex. *Google Images*), d’images générées par un prompt IA (Ex. *Midjourney*), ou encore de contributions personnelles (photographies).

Nous possdedons également des datasets associés aux images qui necessitent quelques pré-traitements. Ces derniers proviennent également de la base de données de *Mushroom Observer.*

L’analyse de ces datasets nous permettra d’identifier les variétés de champignons affiliées aux images soumises à l’algorithme, ceci afin de contrôler le bon apprentissage et la réussite de l’algorithme face aux nouvelles données qui seront soumises a posteriori.

Dans le contexte proposé, ces données sont issues du site *Mushroom Observer*. Le site précise que l’ensemble des contributions (images et données) sont sous licence Creative Commons et en libre accès.

Dans le cas où les sources du site seraient utilisées, il est également intéressant de noter que l’intégralité du code source du site est sous licence open source MIT. Ainsi, toutes contributions de notre part peuvent être bienvenues et aider à l’amélioration générale.

Source : <https://mushroomobserver.org/info/intro>

Analyse des données

Première observation

Le jeu de données initial comporte :

* 1 dataset, de taille 647 615 entrées sur 33 colonnes, référençant les variétés de champignons, et diverses observations à propos de ces dernières.
* 647 615 images de champignons, au format JPG, de dimensions variables.

Dans le fichiers de données initial (‘observations\_mushroom.csv’), on remarque rapidement qu’il y a une relation entre la variable cible (‘label’) et les variables suivantes : species, family, kingdom, scientificName, phylum, class, genus, et order.

Par exemple, la variable ‘scientificName’ donne le nom scientifique de l’espèce du champignon. Chacune des autres variables représente un sous-ensemble de classification pour les espèces de champignons.

Toutes ces variables possèdent également un équivalent en valeurs numériques, on peut donc sous-entendre que ces variables sont encodées :

species <-> speciesKey

family <-> familyKey

kingdom <-> kingdomKey

scientificName <-> usageKey

phylum <-> phylumKey

class <-> classKey

genus <-> genusKey

order <-> orderKey

En dépit de ces relations, ces variables n’apportent pas d’éléments caractérisant le champignon en lui-même. Il s’agit d’éléments de classification de l’espèce, et non de caractéristiques propres permettant de l’identifier, telles que la taille, la couleur, etc.

Pour analyser le dataset, il est necessaire d’épurer ce dernier en supprimant les doublons, traitant les valeurs NaN et ejecter les variables non necessaires à l’élaboration du projet.

Sur 11999 espèces représentées, il est fait état que nombre d’entre elles sont clairement sous répresentées comme le montreront les analyses plus bas.

Pour des raisons précisées dans les analyses plus bas, nous avons pris le parti avec notre chef de cohorte de réduire le nombre d’espèces qui seront entrainnées sur le modèle.

Subdivision du dataset initial

Le premier tri selectif s’est portée sur le top 0.1% des observations du jeu de données initial.

Une image contenant texte, Police, reçu, blanc

Description générée automatiquementAinsi, nous réduisons le nombre d’espèces de champignons à classifier à 130 et le nombre d’images à 207 563 entrées. Nous en déduisons que 130 espèces sur les 11999 représentent environ 30% des observations du jeu de données.

Un notebook (‘creation\_datasets’) a été mis en place afin de subdiviser le fichier original (‘observations\_mushroom’) en sous parties :

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

* ‘infos\_images’ est un dataset permettant d’associer la variable cible ‘label’ à l’Url des images. Ce dataset nous permettra d’entrainer nos futurs modèles.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

* ‘especes\_key’ est un dataset reprenant les variables encodées.
* Une image contenant texte, Police, nombre, capture d’écran

  Description générée automatiquement‘especes\_classification est la version non encodée de especes\_key

Analyse de la variable cible

Initialement, la variable cible ‘label’ contient 11999 espèces enregistrées :

Les histogrammes ci-dessous représentent la répartition des espèces observées.

En raison de la distribution initiale de la variable cible, certaines espèces sont très bien représentées, tandis que d’autres sont sous-représentées (par exemple, une des espèces est représentée par 2 images sur un total de 647 615.

A gauche, l’higramme de répartition des 11999 espèces ; A droite, le même histogramme avec uniquement les 130 espèces les plus représentées (top 0.1% du dataset initial)

Une image contenant ligne, Tracé, diagramme, texte

Description générée automatiquementUne image contenant texte, ligne, Tracé, diagramme

Description générée automatiquement

L’histogramme de distribution de la variable cible confirme le déséquilibre de la distribution.

En réduisant à 130 espèces, il demeure toujours un important déséquilibre dans la représentation de ces dernières.

Une image contenant texte, capture d’écran, diagramme, Tracé

Description générée automatiquementCi-contre, une autre représentation de la distribution initiale des 11999 espèces (variable encodée, chaque espèce portant un identifiant unique allant de 0 à 119998).

**Une image contenant texte, capture d’écran, diagramme

Description générée automatiquement**

Distribution issue de DataFrame après tri selectif des 130 espèces les plus représentées :

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

**Une image contenant texte, capture d’écran, Police, algèbre

Description générée automatiquement**

Analyse de la saisonnalité des observations

Les observations renseignées étant datées, un test a été effectué pour déterminer la saisonalité des observations.

Une image contenant capture d’écran, texte, Caractère coloré, Rectangle

Description générée automatiquementUne variable ‘saison’ a été créée en remplacement de la variable ‘date’ : elle représente la saison pendant laquelle l’observation a été réalisée. Cela permet de simplifier le jeu de données en réduisant le nombre de valeurs différentes que peuvent prendre cette variable, tout en conservant sa capacité à contribuer à l’identification de l’espèce de champignon : en effet, certaines espèces ne poussent qu’à certaines saisons.

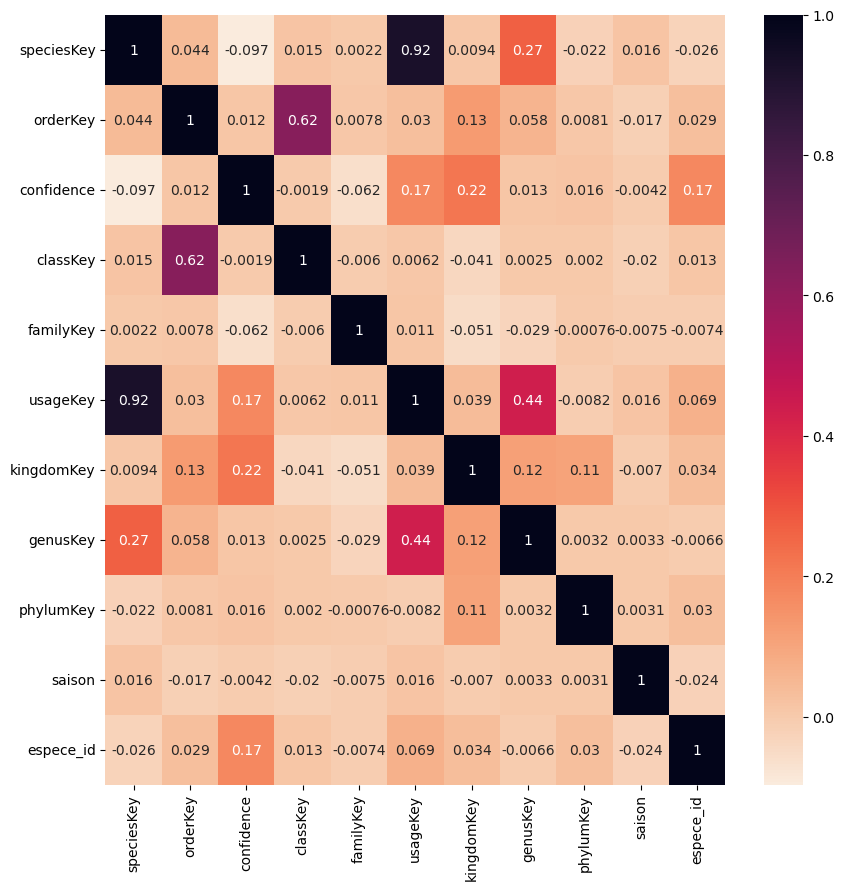
Ci-dessus, les barres empilées représentent la répartition des espèces observées selon les saisons. On remarque qu’il y a plus d’observations en automne et moins au printemps.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquementCertaines espèces ont pu être observées durant plusieurs saisons, tandis que d’autres l’ont été uniquement durant une saison particulière. En conséquence, la saisonnalité n’aura pas d’impact majeur quant à l’apprentissage de notre modèle.

Ce tableau précise cette disparité.

Il y a une corrélation entre le nombre constaté et les saisons : en effet, si une espèce peut être observée durant toutes les saisons de l’année, elle a plus de chance d’être observée.

Recherche de corréltions dans les variables de classification des espèces

La heatmap ci-dessus représente les corrélations entre les variables.

Les variables portant la mention « Key » sont les variables encodées du même nom, sans la suffixe ‘Key’

On ne remarque qu’une seule corrélation, entre les espèces (speciesKey) et le nom scientifique (usageKey).

Un test statistique pourra rejeter ou non cette hypothèse :

Une image contenant texte, capture d’écran, affichage, logiciel

Description générée automatiquementDans le script ci-contre, on teste la corrélation supposée entre les deux variables grâce au test χ².

La p\_value étant de 0.0, l’hypothèse confirmant la corrélation n’est pas rejetée. On peut donc dans notre étude considérer cette hypothèse comme étant valide.

Dimensions des images

Les images constituent la principale source de nos données pour la classification des espèces.

Un tirage aléatoire nous permet de remarquer que les dimensions des images ne sont pas régulières :

Une image contenant capture d’écran

Description générée automatiquementUne image contenant fleur, capture d’écran, Organisme, plante

Description générée automatiquement

Une image contenant texte, diagramme, Tracé, ligne

Description générée automatiquementCi-dessous : histogrammes montrant la distribution de chacune des dimensions en nombre de pixels (largeur, hauteur, et longueur x hauteur) d’un échantillon d’images tiré aléatoirement.

Une analyse de la distribution des dimensions des images nous donne plus de précisions :

* Les images sont généralement de largeur 320 pixels ;
* Les images sont généralement de hauteur 240 px avec également une partie assez importante d’entre elles de hauteurs 220 px et 320 px ;
* Le total de pixels en hauteur x largeur (sans les canaux RGB), correspond en majorité à 75 000 px, une partie importante d’entre elles comporte 70 000 px.

**Une image contenant texte, diagramme, capture d’écran, ligne

Description générée automatiquementUne image contenant texte, diagramme, Police, ligne

Description générée automatiquement**Couleurs des images

**Une image contenant texte, diagramme, Police, ligne

Description générée automatiquement**

**Une image contenant texte, diagramme, ligne, capture d’écran

Description générée automatiquement**L’analyse ci-dessus présente les distributions des moyennes pour chaque canal RGB des images contenues d’un échantillon d’images.

On remarque qu’aucune couleur ne domine les autres en termes de présence et que la répartition des couleurs est assez équilibrée. Cela ne constitue donc pas une variable discriminante pour la classification que nous cherchons à mener.

**Une image contenant diagramme, capture d’écran, pixel

Description générée automatiquement**

Le boxplot permet d’obtenir une représentation graphique de la répartition des valeurs et de remarquer quelques valeurs aberrantes, en particulier au niveau des valeurs des canaux RGB supérieures à ‘200’.

Conclusions d’analyses

Le dataframe ne présente pas de particularités. L’analyse des images nous indique qu’il n’y a pas de tendance particulière au niveau de la colorimétrie des images et que la distribution des observations est relativement déséquilibrées, y compris après le retrait des espèces représentées part moins de 0.1% des observations.

Dans le dataset initial, beaucoup de données ne sont pas exploitables, telle que la variable ‘location’ qui est sur-représentée pour certaines valeurs, non en raison de l’apparition de l’espèce à cet endroit, mais du fait du nombre d’observateurs à cette localisation participant à ces observations (ex : Etats-unis)

Le but de nos travaux étant de développer une méthode de classification d’images, à ce stade il n’y a pas d’autres variables dans le dataset qui puissent être exploitées. Notre analyse se portera essentiellement sur les caractéristiques pouvant être identifiées à partir des images.

A ce stade, il est encore difficile d’entrainer un modèle sur les 130 espèces filtrées. Il demeure encore des espèces avec une sous représsention très pronononcées et les premier jets d’entrainnement de modèle ne nous portait qu’aux alentours de 10% de précision avec un haut taux d’erreurs.

Le choix s’est donc porté sur la classification des 10 espèces les plus réprésentées, nous portant sur une répartition plus équilibrée et permettant de classifier la majorité des espèces observées :

Une image contenant texte, capture d’écran, Caractère coloré, diagramme

Description générée automatiquement

Une image contenant texte, diagramme, ligne, Police

Description générée automatiquementUne image contenant texte, diagramme, ligne, capture d’écran

Description générée automatiquementUne image contenant texte, capture d’écran, diagramme, ligne

Description générée automatiquementUne image contenant texte, diagramme, ligne, Police

Description générée automatiquementUne image contenant texte, capture d’écran, Caractère coloré, diagramme

Description générée automatiquementLes même analyses ont été réalisées sur ces 10 espèces. Cependant, aucune variable discriminante n’a pût être décrite :

Une image contenant diagramme, pixel

Description générée automatiquement

Une image contenant texte, diagramme, capture d’écran, Tracé

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, noir et blanc

Description générée automatiquement

Notre problématique s’apparente à un problème de classification d’images, ici d’espèces de champignons.

A ce stade, nous nous fixons pour objectif de développer un réseau de neuronnes convolutionnel (CNN) capable d’identifier l’espèce de champignon à partir d’une image donnée.

Partie 2 – Traitement des données

Pour les espèces retenues (top 10), nous possédons une banque de 64373 images.

Il demeure des images non exploitables, notamment des images ‘scientifiques’ (analyses microscopiques) ou ne représentant par correctement le champignon :

Une image contenant personne, sourire, plein air, Visage humain

Description générée automatiquement

Une image contenant plante, herbe, plein air, terrain

Description générée automatiquement

Ces images doivent être retirées manuellement du jeu de données.

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Puisque nous devons analyser des images, il est nécessaire de procéder à certaines transformations :

* Normaliser les dimensions des images (réduire les pixels à des valeurs comprises entre 0 et 1) ;
* Redimensionner les images pour l’entraînement avec les algorithmes de deep learning utilisés.

Ces transformations sont appliquées lors de la construction des pipelines Tensorflow.

L’entrainement d’un réseau de neurones ne necessite pas de réaliser de réduction de dimensions. En effet, l’algorithme gèrera lui-même quelles features des images sont importantes et lesquelles mettre de côté.

Nous entrainerons donc un modèle de deep learning pour la classification de 10 espèces de champignons réparties sur **XXX** images.

Partie 3 – Machine learning

Modèles de machine learning

01 sept Quels algorithmes avez-vous essayés ?

01 sept Avez-vous analysé les erreurs de votre modèle ?

01 sept Cela a-t-il contribué à son amélioration ? Si oui, décrivez.

Performances des modèles

01 sept Quelle est la métrique de performance principale utilisée pour comparer vos modèles ? Pourquoi celle-ci ?

01 sept Avez-vous utilisé d’autres métriques de performances qualitative ou quantitative ? Si oui, détaillez-le.

Conclusions sur les modèles essayés

15 sept Avez-vous testé des modèles avancés ? Bagging, Boosting, Deep Learning…  Pourquoi ?

15 sept Décrivez celui/ceux que vous avez retenu et pourquoi ?

15 sept Avez-vous utilisé des techniques d’optimisation de paramètres de type Grid Search et Validation Croisée ?

15 sept Avez-vous utilisé des techniques d’interprétabilité de type SHAP, LIME, Skater… (Grad-CAM pour le Deep Learning...)

15 sept Qu’est ce qui a (ou non) engendré une amélioration significative de vos performances ?

Partie 4 – Conclusion du rapport

29 sept Quel a été le principal verrou scientifique rencontré lors de ce projet ?

29 sept Pour chacun des points suivants, si vous avez rencontré des difficultés, détaillez en quoi elles vous ont ralenti dans la mise en place de votre projet :

* Prévisionnel : tâches qui ont pris plus de temps que prévu, etc.
* Jeux de données : acquisition, volumétrie, traitement, agrégation, etc.
* Compétences techniques / théoriques : timing d’acquisition des compétences, compétence non proposée en formation, etc.
* Pertinence : de l’approche, du modèle, des données, etc.
* IT : puissance de stockage, puissance computationnelle, etc.

Partie 5 – Bilan

Bilan

04 août Êtes-vous entré en contact avec des experts métiers pour affiner la problématique et les modèles sous-jacents ? Si oui, détaillez l’apport de ces interactions.

Au 31/07/2023, nous ne sommes pas encore entrés en contact avec des experts métiers.

29 sept Détaillez quelle a été votre contribution principale dans l’atteinte des objectifs du projet.

29 sept Avez-vous modifié le modèle depuis la dernière itération ? Si oui, détaillez.

29 sept Présentez les résultats obtenus et comparez-les au benchmark

29 sept Pour chacun des objectifs du projet, détaillez en quoi ils ont été atteints ou non.

29 sept S’ils ont été atteints, dans quel(s) process(es) métier(s) votre modèle peut-il s’inscrire ? Détaillez.

Suites du projet

29 sept Quelles pistes d’amélioration suggérez-vous pour augmenter les performances de votre modèle Librairie Optuna

29 sept En quoi votre projet a-t-il contribué à un accroissement de la connaissance scientifique ?

Ressources

Bibliographie

29 sept Sur quels éléments bibliographiques (articles de recherches, blog, livres, etc) vous êtes-vous appuyé pour réaliser votre projet ?

**Deep learning :**

* Chaine Youtube “Science étonnante” du doctorant David Louapre

**Taux de confiance envers les datasets utilisés (contribution communautaire)**

* Théorie de Gustave Lebon « Psychologie des foules » https://fr.wikipedia.org/wiki/Psychologie\_des\_foules\_(livre)
* Chaine youtube “Fouloscopie” du doctorant Mehdi Moussaid

https://www.youtube.com/watch?v=gtXHv95pwyE

https://www.youtube.com/watch?v=cWTn73BZs8c

* Uninformed Individuals Promote Democratic Consensus in Animal Groups

https://www.researchgate.net/publication/51884215\_Uninformed\_Individuals\_Promote\_Democratic\_Consensus\_in\_Animal\_Groups

Annexes

29 sept Diagramme de Gantt.

29 sept Description des fichiers de code.