Fouille de règles numériques pour la prédiction de la dynamique des maladies des plantes

Olivier Gauriau

COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES

ÉCOLE DOCTORALE N° 601
Mathématiques, Télécommunications, Informatique, Signal, Systèmes, Électronique
Spécialité : *INFO*

Par
Olivier GAURIAU

Fouille de règles numériques pour la prédiction de la dynamique des maladies des plantes

Thèse présentée et soutenue à Rennes, le 13 novembre 2024
Unité de recherche : IRISA, UMR 6074

Rapporteurs avant soutenance :

Marie-Odile BANCAL Maîtresse de conférence, AgroParisTech, Palaiseau
Dino IENCO Directeur de recherche, Inrae, Montpellier

Composition du Jury :

Présidente : Céline Robardet
Rapporteurs : Marie-Odile BANCAL, Maîtresse de conférence, AgroParisTech, Palaiseau
Dino IENCO, Directeur de recherche, Inrae, Montpellier
Examinateurs : Céline ROBARDET, Professeur, INSA Lyon, Lyon
Mathilde CHEN, Chargé de recherche, CIRAD, Montpellier
Dir. de thèse : Alexandre TERMIER, Professeur, Université de Rennes, Rennes
David MAKOWSKI, Directeur de recherche, Inrae, Palaiseau
Co-encadrants : Luis GALARRAGA, Chargé de recherche, INRIA, Rennes
François BRUN, Ingénieur de recherche ACTA, Toulouse

Cette thèse n'existerait pas ou n'aurait pas atteint son terme sans le soutien de personnes qui m'ont assisté et accompagnées pendant ces 4 années. Je tiens donc à remercier tous ceux qui m'ont aidé pendant cette période intense de ma vie, sans qui je ne serais peut-être pas parvenu où je suis aujourd'hui.

Premièrement, je souhaite remercier Luis Galarraga et François Brun, mes encadrants et ceux avec qui j'ai le plus collaboré dans le cadre de ma thèse. Ces 4 années n'ont pas été les plus faciles, notamment pendant les périodes où j'éprouvais des difficultés et des blocages. Ils ont toujours fait preuve d'une grande patience et compréhension, m'ont aidé à orienter mon travail, m'ont prodigué de précieux conseils et m'ont permis de prendre du recul lors des moments de doute. Sans leur aide, cette thèse n'aurait pas pu arriver à son terme.

Je souhaite également remercier mes directeurs de thèse, Alexandre Termier et David Makowski. Malgré leur emploi du temps très chargé, ils ont su se rendre disponibles à des moments cruciaux de ma thèse et ont toujours été de bon conseil. Eux aussi ont fait preuve de beaucoup de patience avec moi et m'ont énormément aidé, et je les en remercie.

Je remercie également l'équipe LACODAM dans son ensemble pour m'avoir permis de travailler dans un environnement serein et de me détendre pendant les pauses café. Je souhaite particulièrement remercier Lucie, Julie et Antonin, mes amis et colocataires de bureau, pour les fous rires que nous avons eus ensemble. Merci à Julia, Ambre et Gaëlle, pour m'avoir supporté pendant les pauses et quand j'étais mauvais perdant.

Je remercie aussi tous ceux qui, même s'ils ne font pas partie de l'équipe, m'ont aidé du début à la fin : Lucile Vallet dont les travaux m'ont permis d'avancer rapidement en début de thèse, les experts de l'Institut Technique de la Betterave et de l'Institut Français de la Vigne, dont le travail a permis à ma thèse d'exister et qui ont collaboré avec moi pour valoriser les résultats et modèles obtenus tout en offrant des connaissances agronomiques que seul eux peuvent offrir. Je remercie également les prestataires d'ENEO, qui ont permis de mettre au point la maquette d'outil faisant usage de nos modèles. Enfin, je souhaite remercier particulièrement Gonzalo Méndez, qui a mis à disposition son expertise en data-visualisation et a permis à cet outil de naître.

Je souhaite également remercier l'ACTA qui m'a accueilli et m'a grandement facilité la vie de manière générale, d'un point de vue professionnel et administratif.

Cette thèse a été menée dans le cadre du projet « REGEPI : Apprentissage de règles hybrides pour l'analyse de la dynamique de maladies et ravageurs des plantes en fonction des conditions climatiques » dans le cadre de l'action pilotée par les Ministères de de l'Agriculture, de la Souveraineté Alimentaire et de la Forêt (MASAF), de la Transition écologique, de l'Energie, du Climat et de la Prévention des risques (MTEECPR), de la

Santé et de l'Accès aux soins (MSA) et de l'Enseignement supérieur et de la Recherche (MESR), avec l'appui financier de l'Office Français de la Biodiversité, dans le cadre de l' « appel à projets national Ecophyto 2018 », grâce aux crédits issus de la redevance pour pollutions diffuses attribués au financement du plan Écophyto II+ Elle fait partie de l'institut de convergence DigitAg et a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-16-CONV-0004

Enfin, cette thèse s'inscrit dans le cadre des activités du Réseau Mixte Technologique. Science des Données et Modélisation pour l'Agriculture et l'Agroalimentaire (www.modelia.org), lequel est co-animé pour la période 2020-2025 par plusieurs encadrants de la thèse (François Brun, David Makowski, Luis Galárraga). Cette thèse s'inscrit parfaitement dans le volet 2. Méthodes pour la science des données et la modélisation).

De manière plus personnelle, je remercie mes amis et ma famille, particulièrement mes parents, pour leur soutien constant.

TABLE DES MATIÈRES

INTRODUCTION

Contexte

Subvenir aux besoins alimentaires est une question stratégique à travers le monde. Les manquements à ces besoins peuvent impacter les populations et constituent donc une affaire de santé publique. En amont de la chaîne d'approvisionnement alimentaire, se trouve la production de denrées alimentaires et leurs rendements.

Les rendements des cultures dépendent d'une multitude de facteurs. On distingue souvent les facteurs abiotiques (sécheresse, inondation, fortes températures, gel, salinité, fertilisation, etc.) des facteurs biotiques (maladies et ravageurs des cultures). Parmi les facteurs abiotiques, les conditions climatiques déterminent une grande part de l'élaboration du rendement des cultures. En fonction des différentes phases de développement de la culture, de bonnes conditions météorologiques favorisent la croissance des plantes et augmentent les rendements. Inversement, une mauvaise combinaison de conditions météorologiques peut diminuer ces rendements de manière significative.

Concernant les facteurs biotiques, la propagation de maladies et de ravageurs des plantes peut occasionner des dégâts sur la culture, comme des pertes de capacités photosynthétiques avec des lésions sur les feuilles ou des défoliations, des dégâts sur le système racinaire ou encore des dommages sur les organes de stockage qui constituent la future récolte. Ces différentes maladies et ravageurs ont été une préoccupation majeure de l'Antiquité à nos jours et ont mené à la création de nombreuses méthodes de contrôle [1] , comme la rotation des cultures, la sélection de variétés résistantes ou l'utilisation de produits de biocontrôle. Ces maladies peuvent réduire les rendements, voire même les rendre impropres à la consommation en termes de toxicité (ex. : ergot de seigle, mildiou de la pomme de terre). Ces maladies, combinées aux conditions météorologiques, introduisent une grande variabilité des rendements qui peut avoir des conséquences sévères sur le reste de la chaîne d'approvisionnement [2, 3]. On estime que les pertes de rendement au niveau mondial varient entre 20% et 40% en fonction des années [4]. Ainsi, la majorité des famines historiques sont dues à des conditions climatiques ou épidémiologiques hors normes ayant provoqué des baisses drastiques et durables des rendements agricoles, comme par exemple la crise alimentaire causée par le mildiou de la pomme de terre en Europe dans les années 1840, en partie causée par un climat humide ayant contribué au développement

de la maladie [5].

Les facteurs météorologiques affectent directement les rendements en influençant la croissance des cultures, mais également de manière indirecte à travers leur influence sur le développement des maladies des plantes. Les dégâts occasionnés handicapent les rendements finaux [6]. De la même manière qu'une plante dépend d'une multitude de facteurs pour croître de manière optimale, les dynamiques de propagation des maladies et ravageurs dépendent de certaines conditions climatiques et occasionnent des épidémies différentes d'année en année. Par exemple, les champignons ont évolué pour tirer parti d'environnements humides. Par conséquent, il est logique que les champignons parasites des cultures se développent mieux pendant les années où l'humidité est plus présente. Par exemple, pour le mildiou de la vigne (Plasmopara viticola), un temps humide et de fortes pluies, mêlés à des températures entre 11 et 30 degrés seront propices à son apparition [7].

Pour lutter contre ces agents pathogènes (maladies) et ravageurs, l'utilisation de pesticides s'est développée depuis l'Antiquité par le recours à des substances le plus souvent naturelles, mais dont l'efficacité est limitée. L'utilisation des produits phytosanitaires ou phytopharmaceutiques modernes est apparue au XIXe siècle avec le développement de la chimie minérale, avec les produits à base de sulfate de cuivre et d'autres métaux, puis des produits de la chimie organique à partir du XXe siècle. L'utilisation de produits phytosanitaires en protection des cultures a permis de réduire drastiquement l'impact des maladies et ravageurs des cultures. Ces produits peuvent toutefois avoir un impact négatif et durable sur l'environnement et la santé humaine.

La quantité de produits phytosanitaires utilisés à travers le monde a augmenté de manière constante au cours des dernières décennies [8]. Si cette tendance s'est accompagnée d'une hausse des rendements agricoles, les effets sur l'environnement sont devenus plus visibles. Ces effets peuvent fortement impacter l'environnement de manière négative et durable. Certains produits, comme le sulfate de cuivre, peuvent rester dans les sols sur de longues périodes [9]. D'autres produits peuvent au contraire être plus volatils et se répandre localement. Ils auront ensuite des effets sur la faune et la flore, qui pourront se cumuler au fil des années [10]. Enfin, l'utilisation constante de produits phytosanitaires entraîne la sélection de souches pathogènes résistantes, dites « bio-résistantes » [11]. Ce phénomène risque de faire resurgir la menace que représentaient les maladies des plantes avant l'apparition des produits phytosanitaires. Une solution possible à ce problème est le développement de nouveaux principes actifs ou méthodes agricoles. La mise au point de traitements alternatifs peut toutefois prendre beaucoup de temps.

S'il est impossible d'empêcher totalement ce processus de sélection, il est primordial de ralentir l'apparition de variétés résistantes pour faciliter la transition vers une utilisation plus raisonnée des produits phytosanitaires. La méthode la plus directe est de réduire

la quantité de pesticide utilisée ou le nombre d'applications desdits produits. Jusqu'à maintenant, une application régulière et peu discriminante était utilisée pour s'assurer que les maladies ne pourraient pas impacter les rendements de manière significative. Une stratégie personnalisée peut être mise en place, mais elle nécessite de mobiliser des moyens (souvent humains) pour relever la présence et le développement des pathogènes de visu, ce qui est à la fois imprécis et impraticable sur des cultures de grande taille.

Pour pouvoir réduire la quantité de produits phytosanitaires nécessaires sans entraîner une baisse drastique des rendements, l'agriculture cherche désormais à adopter une logique de protection intégrée des cultures [12] ou encore de protection agroécologique des cultures [13]. Ces approches visent à privilégier les moyens de prévention du développement des maladies par rapport aux moyens de traitement des maladies. Différents leviers agronomiques (rotation des cultures, utilisation de variétés résistantes, suppression des plantes malades, etc.) peuvent être mobilisés à cette fin. Dans les cas où ces leviers ne seraient pas suffisants, l'utilisation de produits phytosanitaires n'est pas exclue, mais reste contrôlée et mesurée.

Il est donc important de savoir avec précision quand l'utilisation d'un pesticide est nécessaire pour être capable de n'en utiliser que lorsque c'est nécessaire. Cela implique de pouvoir prédire dans quels cas une maladie ou un ravageur risque de se développer. Il existe plusieurs manières de répondre à ce problème. Une solution consiste à appliquer des méthodes de modélisation pour prédire la dynamique des maladies des plantes.

Modélisation pour la protection des cultures

Les premiers modèles utilisés pour prédire l'incidence des maladies des plantes sont les modèles mécanistes. Le but de ces modèles est d'établir des relations de cause à effet entre différents composants modélisant des processus biologiques, chimiques ou météorologiques [14] comme l'influence des précipitations au printemps sur le développement d'une maladie. Ces modèles nécessitent de connaître précisément les mécanismes en jeu dans le problème abordé. Dans le cadre de la protection des cultures, ces mécanismes sont le cycle de vie des pathogènes et leurs phases de développement. La base sur laquelle se fondent ces modèles est donc principalement composée de connaissances agronomiques ou biochimiques. Ce genre de modèle est toujours utilisé de nos jours [15] et reste populaire en biologie, car il permet de rassembler et de synthétiser les connaissances actuelles sur le sujet d'étude et présente une fiabilité avérée.

La faiblesse de ce type d'approche réside dans le prérequis d'informations demandé pour obtenir un modèle performant : Mettre au point un modèle mécaniste nécessite de comprendre en profondeur le mode de propagation, le développement des maladies et les

facteurs pouvant participer à leur croissance. Cela demande donc de disposer d'un ensemble exhaustif d'études sur la maladie concernée. Ce besoin en littérature préalable limite donc l'utilisation de ces modèles aux cultures déjà étudiées en profondeur.

Le deuxième type de modèle couramment utilisé est le modèle statistique. La modélisation statistique cherche à utiliser un échantillon de données pour décrire son comportement en se basant sur un ensemble de postulats statistiques. De manière plus simple, un modèle statistique essaie de lier un ensemble de variables aléatoires entre elles pour décrire un comportement réel. Ces modèles ne nécessitent pas, en première instance, de posséder des connaissances précises sur le cycle de développement des maladies. Ils forment automatiquement un ensemble de liens et d'interactions entre les variables disponibles. Les données doivent toutefois respecter un certain nombre de critères pour optimiser les performances du modèle : elles doivent inclure des cas divers et variés, couvrant un large spectre de situations, et respecter des hypothèses (par exemple, une distribution normale des variables prédictives) dont le modèle a besoin pour éviter de former des biais.

Ces modèles sont typiquement utilisés dans une optique de prédiction du risque d'infection ou d'incidence : dans le cas de la prédiction de l'incidence, les modèles vont donner une valeur numérique représentant le stade de développement des symptômes [16]. L'échelle et l'interprétation de ces valeurs varient selon les cas et correspondent souvent au stade de développement de la maladie tel qu'observé par les experts de terrain (par exemple, le pourcentage de feuilles infectées). Plus récemment, ce type de modèle a été utilisé pour prédire l'impact du réchauffement climatique sur le développement des maladies des plantes [17]. Comme indiqué précédemment, les modèles statistiques sont capables de former des relations entre les variables de manière automatique. Ceci les distingue des modèles mécanistes, qui s'appuient sur des connaissances préexistantes. Les relations établies par les modèles statistiques peuvent donc être erronées et doivent être vérifiées. Les approches statistiques requièrent également un volume de données important, à l'inverse des modèles mécanistes qui peuvent se baser sur un volume beaucoup plus réduit.

Le troisième type de modèles utilisés en protection des cultures est celui des modèles de "Machine Learning". Le Machine Learning réunit l'ensemble des méthodes et algorithmes capables "d'apprendre" à effectuer une tâche particulière en se basant sur un ensemble de données, par exemple la prédiction du prix d'une maison. Cet apprentissage se base sur des hypothèses portant sur la structure de ces mêmes données. Le modèle part du principe que ces hypothèses sont valables et cherche à décrire les données le plus précisément possible en suivant ces hypothèses. Les modèles de Machine Learning peuvent être séparés en deux catégories : les modèles supervisés, qui sont utilisés lorsque les données comprennent des labels, c'est-à-dire des informations qui ont déjà été associées à des classes par un être humain ; Les modèles non supervisés se basent sur des données non

labellisées, c'est-à-dire des données ne comprenant pas d'identificateur ou de classes. Ces modèles cherchent à identifier des ensembles de manière empirique en se basant sur des patterns ou des mesures de similitude. Un exemple de problème pouvant être résolu par ces modèles est le regroupement d'images en différentes catégories. Inversement, les modèles supervisés utilisent des données qui ont été classées par un être humain en amont, leur assignant un nom, une classe ou une valeur. Ces identifiants forment une variable cible que le modèle devra suivre et prédire sur de nouveaux cas. Les modèles supervisés cherchent à identifier des relations entre la variable cible et les variables prédictives (les variables présentes dans les données sur lesquelles se baseront les prédictions du modèle). Les modèles supervisés peuvent ensuite être divisés en deux catégories : les modèles de régression et les modèles de classification. La différence entre ces deux types de modèles se situe dans le type de variable que ces modèles doivent prédire : un modèle de classification se base sur une variable discrète (par exemple, un cépage de vigne) et cherchera à la prédire, tandis qu'un modèle de régression se base sur une variable continue (par exemple, la valeur d'incidence d'une maladie). Si les modèles de Machine Learning sont similaires aux modèles statistiques, ils se distinguent toutefois par leurs objectifs. D'une part, les modèles statistiques s'intéressent aux relations entre les variables ; d'autre part, les modèles de Machine Learning sont structurés de manière à obtenir les prédictions les plus précises possibles. Ces modèles sont capables de former des relations plus complexes entre les variables que les modèles statistiques.

Le développement des méthodes de Machine Learning appliquées à la protection des cultures a commencé au début des années 2000, de manière concomitante avec la démocratisation du Machine Learning en général. Il est toujours en cours et nécessite la participation d'experts agronomes ou en data science [18, 19] L'utilisation de modèles de prédiction s'intègre dans l'agriculture de précision, qui tire parti des évolutions technologiques dans le domaine des sciences numériques. Le développement de moyens de collecte de données à distance met à disposition une quantité de plus en plus grande de données pouvant être modélisées. Ces mêmes outils permettent de collecter des données à des échelles de plus en plus fines de manière automatisée, et donc de caractériser plus précisément chaque parcelle observée.

Machine Learning et Interprétabilité

Les modèles de Machine Learning souffrent souvent d'un problème d'interprétabilité qui peut influencer négativement la confiance qu'on leur accorde. L'interprétabilité des modèles est un concept encore débattu au sein de la communauté scientifique. Cette notion se situe à la frontière de plusieurs champs de recherche, de la data science aux sciences cognitives. L'état de l'art comporte plusieurs définitions [20, 21, 22] qui peuvent varier

en fonction du domaine d'application (Rudin2019). De manière générale, on peut décrire l'interprétabilité comme étant le degré auquel il est possible d'expliquer les prédictions d'un modèle de manière compréhensible pour un être humain [21].

Compte tenu de la diversité des modèles de Machine Learning (basés sur des hypothèses et principes variés) et des différentes définitions de l'interprétabilité, il est difficile de trouver une définition définitive et une métrique commune avec laquelle on pourrait comparer les modèles entre eux. Un proxy de l'interprétabilité des modèles peut être la complexité des modèles [23, 21]. On suppose que plus un modèle est complexe, plus il sera difficile d'interpréter ses prédictions. Paradoxalement, si l'on peut informellement classer les différents types de modèles en fonction de leur complexité, aucune métrique unifiée n'existe actuellement.

L'interprétabilité des modèles pose plusieurs questions qui ouvrent des opportunités d'un point de vue de l'application et de l'acceptabilité des modèles. Cette problématique n'est pas nouvelle, car un modèle interprétable bénéficie d'un plus grand degré de confiance de manière générale [20]. En d'autres termes, les utilisateurs finaux ont plus de chances d'accorder leur confiance à un modèle qu'ils comprennent.

Un autre domaine dans lequel l'interprétabilité des modèles peut être utile est la sensibilité des modèles de Machine Learning aux biais. On peut obtenir des modèles qui semblent donner de bons résultats, mais qui en réalité se basent sur des éléments erronés [24]. Il peut être difficile de détecter ces biais dans les modèles les plus complexes. On utilise alors des approches d'interprétation qui permettent d'obtenir des informations sur les prises de décision des modèles. Tous ces facteurs font de l'interprétabilité des modèles un champ d'étude particulièrement intéressant pour la protection des cultures, et de l'agriculture en général [25, 26]. Pour comprendre les approches en termes d'interprétation ou d'explication des résultats, il faut toutefois aborder les différents types de modèles utilisés en protection des cultures.

Objectif de la thèse

Cette thèse s'inscrit dans le contexte de la fourniture d'outils d'aide à la décision (OAD) pour la protection des cultures. Elle a pour objectif de proposer et d'évaluer des méthodes basées sur le machine learning permettant de prédire la dynamique des épidémies de manière précise, tout en proposant des outils pour accompagner les experts dans l'analyse des déterminants agronomiques et météorologiques de ces épidémies. On décompose par conséquent les objectifs en trois grands axes :

- ➔ Évaluer des modèles de machine learning interprétables intégrables dans des outils d'aide à l'orientation des traitements.

- ➔ Examiner le compromis entre l'interprétabilité et les performances des modèles de machine learning pour la protection des cultures.
- ➔ Montrer l'intérêt des modèles interprétables pour la protection des cultures.

Pour atteindre ces objectifs, il est nécessaire d'entraîner des modèles faisant preuve de performances satisfaisantes et stables face aux variations météorologiques annuelles. On devra par la suite les utiliser pour obtenir des informations pertinentes pour les experts, avant de pouvoir leur proposer ces informations de manière efficace.

Les contributions composant cette thèse et répondant à ces questions sont les suivantes :

- ➔ Une analyse du compromis performance/complexité des différentes méthodes de Machine Learning dans le cadre de la protection des cultures, en lien ici avec le développement du mildiou de la vigne et de la cercosporiose de la betterave.
- ➔ Une extraction de connaissances à partir des modèles appris en utilisant des techniques de Machine Learning explicables.
- ➔ Un outil de visualisation ciblant les experts agronomes dont le but est de leur transmettre les connaissances extraites.

Les deux premières contributions ont été valorisées sous forme d'une publication dans le journal Smart Agricultural Technology [27]. La dernière contribution est décrite dans un chapitre de ce manuscrit (le Chapitre 4 intitulé « Représentation visuelle des modèles et explication des prédictions »)).

En premier lieu, un état de l'art des méthodes utilisées en protection des cultures sera effectué dans le Chapitre 1, suivi par un état de l'art des méthodes d'explication des prédictions et de visualisation dans la sous-section 1.2.1. Dans le Chapitre 2, nos cas d'études seront décrits plus en détail. Par la suite, l'approche privilégiée et les résultats obtenus seront présentés dans le Chapitre 3. Le chapitre suivant se concentrera sur la maquette de l'outil d'analyse de nos modèles, développée en collaboration avec des acteurs de terrain (Chapitre 4). Enfin, l'ensemble des perspectives envisageables seront expliquées dans le Chapitre 5 et seront suivies par une conclusion dans le Chapitre 4.

Chapitre 1

Etat de l'art

L'utilisation de modèles de machine learning en protection des cultures présente plusieurs avantages qui expliquent sa popularité grandissante : Ces modèles sont capables de découvrir des relations plus complexes entre les variables que les modèles statistiques. Qui plus est, les modèles de Machine Learning requièrent moins de connaissances préalables et d'expertise humaine que les modèles mécanistes. Ce chapitre présentera donc la méthode générale d'utilisation des modèles de Machine Learning et quelques modèles de *régression* couramment utilisés en protection des cultures [28, 29, 30, 31].

L'application de méthodes de machine learning suit une démarche bien définie : collecte et préparation des données, choix des modèles, entraînement puis évaluation des modèles, puis une phase de déploiement. La préparation des données consiste à les nettoyer et à supprimer les éventuelles anomalies : on cherchera par exemple à détecter des valeurs manquantes ou aberrantes, qui devront donc être retirées ou modifiées. Le choix du modèle peut s'appuyer sur un vaste ensemble de méthodes préexistantes [32]. Ces algorithmes présentent des formalismes et des complexités variées, allant de modèles linéaires classiques (Linear Regression, Lasso, etc.) à des méthodes non linéaires (Support Vector Machine, Random Forest), jusqu'à des méthodes de Deep Learning (Multilayer Perceptron, Long Short Term Memory, Convolution, etc.). Les modèles considérés sont ensuite entraînés sur les données. Une fois cette étape accomplie, on évalue ceux-ci en leur faisant évaluer des cas isolés des données d'entraînement.

La qualité des données peut impacter la précision des prédictions fournies par les modèles. Il est donc généralement nécessaire de travailler en amont sur les données fournies en entrée aux modèles. [33] Le problème le plus souvent observé concerne l'intégrité des données : cela inclut les données manquantes, aberrantes ou dupliquées. Les valeurs manquantes ne peuvent pas être traitées par tous les modèles, et les valeurs aberrantes ou dupliquées peuvent dégrader les performances générales du modèle et introduire un biais dans celui-ci. Pour résoudre ces problèmes, on peut choisir de supprimer tous les cas comprenant des valeurs présentant ces caractéristiques. Pour éviter de réduire la quantité de données disponibles, on peut également choisir de modifier ou compléter ces valeurs en se basant sur la valeur moyenne ou médiane de la variable concernée sur l'ensemble des données.

Un autre problème peut subvenir quand les variables divergent en termes d'échelle : un biais peut alors être induit dans le modèle si celui-ci accorde une importance plus grande aux variables atteignant des valeurs plus élevées. Ce problème peut être contourné en standardisant les données, ce qui consiste à réajuster les valeurs pour que celles-ci soient distribuées dans un intervalle défini (typiquement, $[-1, 1]$ ou $[0, 1]$) ; ou en les normalisant, ce qui consiste à ajuster les valeurs autour de 0 avec un écart-type de 1, revenant donc à remanier les valeurs pour qu'elles suivent une distribution normale centrée réduite. Dans le cas où des variables catégorielles (comme des labels) sont présentes, il peut être nécessaire de les transformer en format numérique. Dans ce cas, on utilisera souvent des méthodes d'encodage comme one-hot qui binarise les valeurs ciblées. Enfin, la présence de variables corrélées entre elles peut poser problème à certains modèles et induire des biais et des redondances. Gérer ce problème peut se faire en effectuant une sélection des variables en se basant sur des mesures d'importance [34], ou en utilisant des techniques de réduction de la dimensionnalité comme l'ACP [35].

Une fois que les données sont préparées de manière adéquate, il s'agit de choisir un modèle adapté à celles-ci : Modèles supervisés ou non supervisés, classification ou régression. Dans notre cas, nous nous limiterons aux modèles supervisés. Par cela, nous entendons les modèles faisant usage de données labellisées pendant la phase d'entraînement. Ce choix est nécessaire dans la mesure où nous abordons le problème sous l'angle de la régression plutôt que de la classification. Chaque instance de données se compose d'entrées (les variables utilisées pour entraîner nos modèles) associées à une sortie. Cette sortie peut être de nature discrète ou continue. Les valeurs discrètes représentent des classes classant les instances dans un nombre limité de catégories distinctes. Ces valeurs peuvent représenter des catégories, comme des espèces d'animaux ou de plantes. Les valeurs continues peuvent prendre des valeurs infinies et peuvent représenter des grandeurs telles qu'un prix ou un poids. En fonction de la nature de ces sorties, on se situera dans un problème de *classification* dans le cas discret ou de *régression* dans le cas continu. Dans le cadre de la protection des cultures, la détection des maladies des plantes à base d'images relève de la classification, alors que la prédiction de l'impact d'une maladie sur les cultures relève plutôt de la régression. Une fois ces facteurs pris en compte, il est nécessaire de garder à l'esprit que les modèles varient en termes de structure et de complexité. En fonction des hypothèses formulées sur les données, ceux-ci seront capables de représenter les relations entre les variables de manière plus ou moins précise. Les modèles linéaires supposent que ces relations sont linéaires. D'autres modèles privilégient des approches non linéaires pour mieux correspondre à la réalité. Cela se fait souvent en échange d'une complexité accrue.

Une fois le modèle choisi, il convient de l'ajuster correctement. La plupart des modèles sont entraînés en suivant certains paramètres délimitant la structure du modèle concerné. Il est

par exemple possible de définir la profondeur maximale d'un arbre de décision ou encore le nombre minimal de cas présents dans une feuille. Celui-ci fournira des prédictions qui permettront d'évaluer la précision du modèle. Cela permettra notamment de déterminer si le modèle est adapté aux données. Dans le cas où le modèle serait trop simple, il serait alors victime de *sous-apprentissage* et échouerait à modéliser correctement les données. Il présenterait alors des performances médiocres, tant pendant la phase d'entraînement que pendant la phase d'évaluation. Dans le cas où le modèle est trop complexe, on observera un effet de sur-apprentissage. Le sur-apprentissage se produit quand un modèle se base de manière excessive sur les spécificités des données. Cela se manifeste par une baisse de performance entre la phase d'entraînement et d'évaluation : le modèle est tellement adapté aux cas observés dans les données utilisées pendant l'entraînement qu'il ne parvient pas à généraliser ses prédictions à des données indépendantes.

Comme indiqué précédemment, cet état de l'art se concentrera sur la régression, et plus particulièrement sur les méthodes de régression les plus pertinentes pour notre domaine d'application, c'est-à-dire des méthodes déjà utilisées en protection des cultures [28, 29, 30, 31] et représentant un panel large en termes de complexité : les modèles de régression linéaire seront nos modèles de complexité basse, les modèles ensemblistes ceux de complexité haute et seront présentés dans cet ordre. On présentera enfin le modèle de complexité intermédiaire utilisé dans notre étude, HiPaR.

1.1 Méthodes de régression

L'objectif de la régression est de prédire une variable réelle, que l'on appelle la variable cible ou variable à prédire, sous la forme d'une valeur numérique continue. Pour ce faire, on utilise un ensemble de variables distinctes que l'on appellera variables prédictives (voir Table 2.1). Ces variables sont, dans notre contexte, la plupart du temps des variables météorologiques, comme des mesures de précipitation ou d'ensoleillement.

Température	Humidité	Pluie	Cépage	...	Incidence
24°	77%	100cm	Pinot	...	10%
17°	36%	50cm	Merlot	...	4%
⋮	⋮	⋮	⋮	⋮	⋮

TABLE 1.1 – Exemple de jeu de données de protection des cultures incluant des variables météorologiques et agronomiques

En présupposant que l'on dispose d'un ensemble de n observations (dans notre contexte, des observations faites dans des parcelles) de la variable cible représentant l'impact d'une maladie, représentées sous la forme d'un vecteur colonne $\boldsymbol{y} \in \mathbb{R}^n$, on associe ces observations à un ensemble d'observations décrivant les variables prédictives, organisées sous

la forme d'une matrice $\boldsymbol{X} \in \mathbb{R}^{n \times d}$. Chaque ligne $\boldsymbol{x}_i^\top \in \mathbb{R}^d$ dans la matrice contient les valeurs observées des variables prédictives associées à chaque observation y_i [1]. Si une variable est catégorielle, c'est-à-dire discrète, comme le cépage du tableau 1.1, on suppose que ses valeurs ont été encodées par des valeurs numériques. Cela peut être accompli en utilisant des méthodes de one-hot encoding ou de réduction de la dimensionnalité. Une fois ces données mises en forme, il est possible d'entraîner des modèles sur celles-ci.

L'objectif de la régression est d'entraîner une fonction $\boldsymbol{y} = f(\boldsymbol{X}) + \boldsymbol{\epsilon}$ de manière à minimiser $\boldsymbol{\epsilon}$. La fonction f est une modélisation des données assignées pour prédire la variable cible pour des instances $\boldsymbol{x}^\top \in \mathbb{R}^d$ des variables prédictives. Le terme $\boldsymbol{\epsilon}$ est l'erreur du modèle de régression, aussi appelé résidu. Ce résidu peut être assimilé à la partie de la valeur cible $\boldsymbol{y}$ que le modèle ne peut pas modéliser par manque d'information. Ce manque d'information peut être dû à l'absence de variables prédictives. Par la suite, on dira que le modèle f est basé sur un ensemble d'observations d'entraînement et de validation. Cela signifie que le modèle formulera des hypothèses à partir des données d'entraînement avant de les tester sur l'ensemble de validation.

Les modèles de régression comprennent un ensemble de méthodes qui se basent chacune sur des principes et des approches différentes, détaillés par la suite.

Modèles de régression classiques

Régression linéaire La régression linéaire pose comme hypothèse l'existence d'une relation linéaire entre les variables prédictives $\boldsymbol{X}$ et la variable cible $\boldsymbol{y}$

$$\boldsymbol{y} = \boldsymbol{\beta}\boldsymbol{X'} + \boldsymbol{\epsilon} \tag{1.1}$$

Où $\boldsymbol{X'} = \mathbb{1} \oplus \boldsymbol{X}$, i.e., $\boldsymbol{X'} \in \mathbb{R}^{n \times (d+1)}$ et $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$ sont les paramètres du modèle (l'opérateur $\oplus$ dénotant la concaténation des colonnes), c'est-à-dire les coefficients associés à chaque variable prédictive auxquels est ajouté l'intercept, ou l'ordonnée à l'origine du modèle.

Les paramètres du modèle sont calculés en minimisant une fonction de perte, par exemple $\mathcal{L}_l(\hat{\boldsymbol{\beta}}) = ||\boldsymbol{y} - \boldsymbol{X'}\hat{\boldsymbol{\beta}}||_2^2$. On effectue cette minimisation en suivant la méthode des moindres carrés. Ce choix se base sur le théorème de Gauss-Markov, selon lequel les estimations des valeurs $\boldsymbol{\beta}$ obtenues en suivant la méthode des moindres carrés sont celles qui minimisent l'erreur d'échantillonnage. On entend par là que ces estimations sont les meilleurs estimateurs non biaisés des paramètres. La méthode des moindres carrés utilisée est la méthode

1. Par la suite, on représentera les vecteurs et matrices en gras pour distinguer les scalaires et les fonctions. Qui plus est, les matrices seront notées en majuscules.

ordinaire, dans laquelle l'importance de chaque mesure de la variable cible targetvar n'est pas pondérée.

La régression linéaire est une méthode fréquemment utilisée en raison de sa simplicité et de sa capacité à décrire efficacement les données disponibles. Elle offre une modélisation à la fois interprétable et relativement peu exigeante en puissance de calcul. Dans certains cas, la régression linéaire peut produire des résultats comparables, voire supérieurs, à des modèles plus complexes et s'adapter de manière adéquate à des jeux de données peu volumineux.

Les coefficients linéaires obtenus par la méthode des moindres carrés nous fournissent une mesure claire et explicite de l'importance d'une variable prédictive sur le résultat final. L'un des points faibles de cette méthode réside dans son hypothèse de départ, qui suppose l'existence d'une relation linéaire entre les variables prédictives et la cible. Cette supposition peut fréquemment entraîner de moins bonnes performances que d'autres modèles non linéaires. Ce cas de figure se manifeste quand la relation entre les variables prédictives et la variable à prédire n'est pas linéaire. Ces propriétés font que les modèles de régression linéaires sont souvent utilisés dans un but d'exploration de jeux de données, ou comme modèle de base auxquels seront comparés d'autres modèles.

Lasso. Un des risques encourus par les modèles de régression linéaire est le surentraînement, ou overfitting. Il se manifeste par une différence significative de performance entre la phase d'entraînement et la phase de validation. Il se produit généralement lorsque le modèle cherche à décrire trop précisément les données d'entraînement. Le modèle présente donc des résultats optimisés sur ces mêmes données, mais échoue à maintenir ces performances sur d'autres cas, et donc à généraliser. Un des facteurs pouvant amener ce problème dépend du volume de données disponibles.

Plus un jeu de données inclut de variables prédictives, plus le modèle qui en résulte sera complexe. S'il est tentant d'inclure autant de variables que possible, avec l'espoir d'y trouver des variables pertinentes, cette approche n'est pas sans risque. Elle augmente la sensibilité à la variance d'échantillonnage et diminue l'explicabilité du modèle. On parle alors de « fléau de la dimension ». Ces problèmes sont prévalents en protection des cultures, car le nombre de variables météorologiques peut varier en fonction de la manière dont les données sont traitées et utilisées. Ces variables sont sujettes à des variations interannuelles significatives qui peuvent affecter les performances des modèles. Contrebalancer ce problème nécessite d'utiliser une plus grande quantité de données. Mais la perte d'explicabilité pouvant être induite par l'augmentation du nombre de variables pose problème vis-à-vis de l'acceptabilité des modèles. Qui plus est, les modèles linéaires sont relativement peu flexibles [32] et ont tendance à surentraîner. Ceci peut être problématique en termes de précision des modèles et donc en termes de confiance accordée à ceux-ci.

Pour éviter de sélectionner un nombre trop grand de variables, l'une des méthodes les plus couramment utilisées est la méthode Lasso, qui se base sur une régularisation L1 de la fonction de perte. Cette pénalisation s'applique au nombre de variables utilisées, ce qui implique que le nombre de coefficients linéaires des modèles différents de zéro est limité. En d'autres termes, la pénalisation permet de sélectionner les variables les plus importantes en punissant l'utilisation de variables d'impact moindre qui pourraient complexifier excessivement le modèle. La pénalisation est accomplie en minimisant la fonction objectif suivante :

$$\boldsymbol{\beta} = argmin_{\hat{\boldsymbol{\beta}}}\ \mathcal{L}_l(\hat{\boldsymbol{\beta}}) + \theta||\hat{\boldsymbol{\beta}}||_1. \tag{1.2}$$

Le coefficient de pénalisation θ est un hyperparamètre qui contrôle le degré de pénalisation appliqué à la fonction de perte pendant la phase d'optimisation. La méthode Lasso sélectionne le jeu de paramètres $\hat{\boldsymbol{\beta}}$ offrant les meilleures performances au terme de la validation croisée, basée sur la fonction de coût pénalisée.

Le Lasso permet d'éviter les problèmes posés par la complexification des modèles : il permet de garder le modèle linéaire interprétable. La sélection des variables les plus importantes permet de limiter le surentraînement et améliore généralement la robustesse des modèles [35, 36].

Si l'hypothèse de linéarité s'applique dans un grand nombre de cas, de nombreux problèmes incluent des paramètres liés à la variable à expliquer de manière non linéaire. Dans ce cas, on peut privilégier des modèles non linéaires, qui seront souvent plus complexes et donc moins interprétables. Une extension naturelle des modèles linéaires appartenant à cette catégorie sont les modèles linéaires par morceaux. Ces modèles forment des partitions de l'espace des variables sur lesquelles se baseront les modèles pour fournir des prédictions. Ces partitions sont caractérisées à l'aide d'un ensemble de conditions portant sur les variables. Ces conditions peuvent être représentées sous la forme d'un arbre de décision.

Decision/Regression Trees. Chaque nœud d'un arbre de décision contient une condition booléenne portant sur une variable prédictive en particulier (par exemple $x_2 < 5.5$ de Fig 1.1). Ces nœuds forment les partitions de l'espace des variables citées précédemment, que l'on peut assimiler à un ensemble de sous-régions découpant ledit espace. Chaque nœud engendre deux enfants, qui sont eux-mêmes des arbres de décision. Chaque feuille est associée à une prédiction de la variable cible par le modèle. Quand cette prédiction est une valeur numérique, on parle alors *d'arbre de régression* [37]. Une approche souvent utilisée pour la construction d'arbres de décision est l'algorithme CART [38], notamment utilisé par défaut dans la librairie scikit-learn. D'autres méthodes, comme ID3 et C4.5

(quinlan1986induction), existent également.

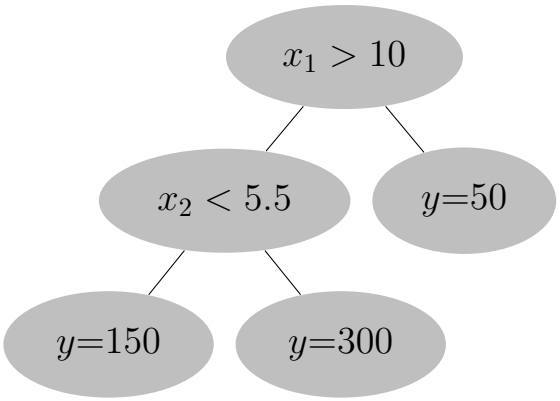


FIGURE 1.1 – Illustration d'un arbre de décision. Les nœuds présents sont associés aux conditions $x_1 > 10$ et $x_2 < 5.5$. Les feuilles sont associées aux valeurs 150, 300 et 50

La construction d'un arbre de décision se fait en se basant sur une approche générale commune.

- ➔ Choix du critère de division
  - ◦ Il est nécessaire de sélectionner un critère qui servira à déterminer la division qui formera les racines par la suite. On utilise généralement des mesures de pureté telles que l'entropie de Shannon [39] ou l'indice de Gini [38].
- ➔ Division du nœud racine
  - ◦ Choix de la variable de division
    - — Une variable est sélectionnée pour générer une division qui formera la racine. La plupart du temps, un algorithme glouton détermine le seuil ou la division la plus avantageuse pour l'ensemble des variables en se basant sur le critère de sélection défini précédemment. La variable et le seuil associé présentant les meilleures performances seront utilisés comme base pour la racine. Dans Fig 1.1, la variable sélectionnée est x_1 et le seuil est 10.
  - ◦ Division
    - — Une fois la condition de la racine choisie et celle-ci formée, des branches divergentes apparaissent : Cela divise le jeu de données en deux parties distinctes en se basant sur le seuil utilisé par la racine. Dans Fig 1.1, les nœuds formés sont le nœud associé à la condition $x_2 < 5.5$ et 50.
- ➔ Itération
  - ◦ Formation des sous-nœuds
    - — Des nœuds sont formés de la même manière que dans le nœud racine, sur les branches formées précédemment. Le nœud associé à la condition $x_2 < 5,5$ est formé de cette manière.

➔ Arrêt

- Si un critère d'arrêt est atteint, le processus récursif prend fin. Le nœud concerné devient donc un nœud terminal, ou feuille. Aucune division ou branche ne sera issue de ce nœud et représentera une des extrémités de l'arbre. Les critères d'arrêt les plus souvent utilisés sont :
  - Si un nœud est pur (toutes les observations contenues dans le nœud sont identiques)
  - Si la profondeur du nœud atteint la limite définie pour limiter la taille de l'arbre.
  - Si le nombre d'observations contenues dans le nœud est inférieur à un seuil défini au préalable.
  - Si le gain de pureté obtenu par la division est trop faible.
- Une fois la feuille formée, une valeur de prédiction est associée à celle-ci. Dans le cas d'une classification, c'est la classe majoritaire parmi les observations de la feuille qui sera choisie. Dans un problème de régression, on pourra choisir la valeur moyenne des observations présentes. Dans Fig 1.1, les feuilles sont associées aux valeurs de prédiction 150, 300 et 50

L'ensemble des conditions associé à chaque observation permet de fournir des explications concernant les prédictions du modèle. Dans notre exemple, on peut supposer qu'une valeur de x_1 supérieure à 10 amène à des valeurs basses, alors qu'une valeur de x_1 inférieure à 10 et une valeur de x_2 inférieure à 5,5 maximisent la valeur prédite. Cela permet aux arbres de décision d'être interprétables pour peu qu'ils ne soient pas trop profonds. On entend par là que le nombre de partitions ne soit pas trop important, au point de le rendre trop complexe pour qu'un être humain puisse le comprendre. Les arbres de décision sont donc classifiés comme des modèles interprétables ou boîte blanche. Par cela, on entend que ces modèles sont intrinsèquement interprétables, dans le sens où ils sont directement compréhensibles. En effet, chaque branche ou chemin final d'un arbre de décision peut être décomposé en un ensemble de patterns de condition délimitant l'espace des données en sous-ensembles disjoints. Ces ensembles forment des patterns. Les méthodes se basant sur l'utilisation de ensembles de condition sont dites *pattern-based* ou *pattern-aided*.

Pattern-aided Regression. Les modèles de régression Pattern-based, ou Rule-based, sont composés de sous-modèles locaux entraînés sur des régions particulières des données. Ces régions sont délimitées à l'aide d'un ensemble de conditions logiques portant sur les variables prédictives. Ces conditions constituent un ensemble de caractéristiques interprétables du sous-ensemble formé par les partitions. Cela permet de contextualiser les modèles locaux et leurs prédictions : Par exemple, on peut imaginer qu'un modèle *pattern-*

aided pourrait former une partition entre les parcelles ayant connu des précipitations plus importantes.

Les modèles locaux sont le plus souvent des modèles considérés comme interprétables comme des modèles linéaires, ce qui permet de modéliser des relations locales entre les variables prédictives et cibles qui ne pourraient pas être décrites dans des modèles continus. Comme montré dans [40, 41], ces méthodes font preuve de performances supérieures aux modèles de régression linéaires classiques en échange d'une augmentation modeste de la complexité des modèles. Parmi les modèles pattern-based, on dénombre les méthodes de régression par morceaux [42], arbres de régression [43], model trees [44] [2], Contrast pattern-aided regression (CPXR) [41], et HiPaR [40]. Les méthodes rule-based sont considérées et utilisées depuis un certain temps en protection des cultures [45]. L'extension qui en est faite dans les méthodes ensemblistes et HiPaR est décrite dans la section suivante. Compte tenu de la popularité des arbres de régression en tant qu'élément de base des méthodes ensemblistes et des gains de performance obtenus par HiPaR comparé à d'autres méthodes de cette catégorie, le reste de cette section se concentrera sur ces méthodes.

Comme dit précédemment, la haute variance des arbres de décision peut être mitigée en utilisant un ensemble d'arbres de décisions plutôt qu'un arbre unique. Les méthodes de *Bagging* et *Boosting* suivent ce principe pour réduire la variance des arbres de décision. Elles seront expliquées dans les sections suivantes, ainsi que les modèles qui les appliquent, les *Random Forests* et *Gradient Boosting*

Méthodes Ensemblistes.

Random Forests. Une forêt aléatoire est un modèle constitué d'un ensemble d'arbres de décision entraînés en suivant le principe du *Bagging* et combinant les prédictions de ces arbres pour obtenir une prédiction agrégée, une technique aussi appelée *Averaging*. Le *Bagging* est une méthode qui développe le principe de *Bootstrap* L bootstrap effectue plusieurs tirages aléatoires avec remise, sur lesquels sera entraîné un modèle. Le bagging entraînera un modèle distinct pour chacun de ces tirages. Pour entraîner chaque modèle, on tire aléatoirement avec remise un ensemble de n sous-ensembles des données. Pour chacun de ces sous-ensembles, on entraîne un arbre de décision ; Chaque arbre est entraîné de manière isolée, , ce qui permet de paralléliser l'entraînement de chacun d'entre eux. Chaque arbre est donc entraîné sur un échantillon d'observations différentes et à l'aide d'un échantillon limité de variables explicatives. Ces modèles sont ensuite utilisés ensemble, où leurs résultats sont moyennés (aussi appelé *Averaging*) pour donner une prédiction finale. Ces tirages permettent d'améliorer les performances d'un modèle en réduisant sa variance,

2. Ces modèles sont des arbres de régression qui incluent un modèle local à chaque nœud terminal, ou feuille, de l'arbre

et donc d'augmenter sa stabilité. Si le Bagging est utilisé typiquement sur des arbres de décision (formant des forêts d'arbres décisionnels), ces méthodes sont théoriquement applicables à tout type de modèle. Les forêts d'arbres décisionnels sont des ensembles de modèles faibles, ici des arbres de décision [43].

En plus du bagging habituel, une des particularités des forêts d'arbres décisionnels par rapport aux arbres de décision classiques est qu'un nombre k de variables est tiré parmi les $\boldsymbol{V}$ variables prédictives à disposition à chaque fois que l'arbre effectue un nœud et forme un nœud. L'abandon de l'algorithme glouton des arbres de décision permet d'accélérer l'entraînement de chaque arbre de la forêt décisionnelle et contribue à séparer chaque arbre, formant ainsi un ensemble d'estimateurs divers, que l'on suppose plus à même de décrire des relations complexes qui pourraient passer inaperçues autrement et indépendants les uns des autres, en leur donnant une vision partielle mais unique des données.

Ces méthodes permettent de réduire significativement la variance et la sensibilité au surentraînement des arbres de décision. Qui plus est, les forêts d'arbres aléatoires, ou random forests, sont des modèles relativement peu gourmands en termes de temps de calcul, notamment en comparaison aux modèles faisant usage du boosting. En conjonction avec leurs performances, cela fait d'eux des modèles très appréciés en protection des cultures [46].

Cela se fait toutefois au prix de la perte d'explicabilité qui résulte de la nature ensembliste des forêts. Celles-ci perdent en effet la capacité d'explication des résultats inhérente aux arbres de décision, une prédiction résultant de l'agrégation de centaines, voire de milliers, d'entre eux. Cette prédiction implique alors de suivre et d'interpréter un grand nombre de chemins par cas. Pour analyser ces modèles, il sera donc nécessaire d'avoir recours à des méthodes d'interprétation.

Gradient Boosting. Une autre méthode ensembliste souvent utilisée en protection des cultures est le *gradient-boosting* [47]. De manière analogue aux forêts d'arbres décisionnels, les modèles de type gradient-boosting utilisent un ensemble de modèles " faibles " dans le but d'obtenir une prédiction plus résiliente que celle que l'on pourrait obtenir d'un seul modèle. Historiquement et encore de nos jours, le gradient-boosting a été utilisé en faisant appel à des arbres de décision.

Le gradient-boosting est une méthode additive itérative. Par cela, on entend que les modèles faibles qui le composent ne sont pas générés indépendamment les uns des autres : chaque modèle h_m est entraîné en fonction de l'erreur du modèle précédent h_{m-1} en se basant sur le gradient de la fonction de perte minimisée. En d'autres termes, chaque itération entraîne un nouveau modèle dont l'objectif est de corriger les erreurs commises

par le modèle précédent.

$$f_m(\boldsymbol{X}) = f_{m-1}(\boldsymbol{X}) + \gamma_m h_m(\boldsymbol{X}) \tag{1.3}$$

$$\gamma_m = \mathcal{L}(\boldsymbol{y}, f_m(\boldsymbol{X})) \tag{1.4}$$

Chaque arbre décisionnel composant le gradient-boosting trees est donc lié aux arbres précédents et suivants dans la chaîne itérative. Contrairement aux arbres d'une forêt d'arbres décisionnels, ces arbres ne sont pas indépendants entre eux, car la structure de l'arbre h_m sera intrinsèquement liée à h_{m-1}.

Les random forests et le gradient-boosting sont souvent comparés en raison de leurs similarités. On notera que le gradient boosting est généralement plus précis que les forêts d'arbres [32] au prix d'un degré supplémentaire de complexité due à leur nature itérative.

Hierarchical Pattern-aided Regression (HiPaR)

HiPaR. Hierarchical Pattern-aided Regression [40] est une méthode de régression pattern-based basée sur l'utilisation de modèles locaux définis par un ensemble compact de règles hybrides portant sur les variables prédictives. Ces règles prennent la forme suivante :

$$p = C_1 \wedge \ldots C_m \Rightarrow \boldsymbol{y} = f_p(\boldsymbol{X}_p). \tag{1.5}$$

Dans cette formulation le pattern p est une conjonction de conditions portant sur les variables prédictives, (Exemple : *vitesse-vent* $> 50 \wedge$ *humidité* > 30). Ces conditions délimitent une sous-région des données $\boldsymbol{X}_p \in \boldsymbol{X}$. Une règle hybride est associée à un modèle linéaire local f_p qui a été entraîné sur $\boldsymbol{X}_p$, ce qui permet de morceler les données de manière à obtenir des prédictions plus précises. Le modèle f que l'on appelle modèle *par défaut* ou modèle général a été entraîné sur l'ensemble des données. Il est conservé et utilisé quand aucune des règles n'est applicable. Cela signifie que si une instance de données ne respecte aucun des ensembles de conditions des règles hybrides, le modèle général sera sollicité.

L'initialisation du modèle se base sur l'entraînement du modèle général. Une fois cette tâche terminée, HiPaR explore les données à la recherche d'un ensemble compact de règles hybrides en deux phases : l'énumération et la sélection.

Algorithm 1: hipar-candidates-enum

Input: a dataensemble : D with attributes A_{num}
parent hybrid rule : $r_p : p \Rightarrow y = f_p(A'_{num})$
patterns of size 1 : C
minimum support threshold : θ
maximum support threshold : γ

Output: a ensemble $\mathcal{R}$ of candidate rules $p \Rightarrow y = f_p(A'_{num})$

$\mathcal{R} := \varnothing$
$\mathcal{C}' := C$
$C_n := \{c \in C | c : a \in I \wedge a \in A_{num}\}$
$\nu :=$ k-th percentile of iv_D in C_n
for $c' \in C'$ **do**
  $\hat{p} := p \wedge c'$
  $C' := C' \setminus \{c'\}$
  if $s_D(\hat{p}) \geq \theta \ \wedge \ s_D(\hat{p}) \leq \gamma \wedge iv_(\hat{p}) > \nu$ **then**
    $p' = \mathbf{cl}(\hat{p})$
    $C' := C' \setminus p'$
    if p *is the left-most parent of* p' **then**
      Learn $r_{p'} : p' \Rightarrow y = f_{p'}(A'_{num})$ on $D_{p'}$
      if $m(r_{p'}) < m(r_{p^*}) \ \forall p^* : p^*$ *is parent of* p' **then**
        $\mathcal{R} = \mathcal{R} \cup \{r_{p'}\}$
        $C'_n := \varnothing$
        for $a \in A'_{num} \setminus attrs(p')$ **do**
          $C'_n := \{c \in discr(a, D_{p'}, y) \mid s_D(c) \geq \theta\} \cup C'_n$
        $\mathcal{R} := \mathcal{R} \cup$ *hipar-candidates-enum*$(D, r_{p'}, (C' \setminus C_n) \cup C'_n, \theta)$

return $\mathcal{R}$

Durant la phase d'énumération présentée dans l'algorithme 1, HiPaR explore l'espace des patterns p en suivant une approche hiérarchique de recherche en profondeur. Quand un pattern p est découvert, HiPaR forme une règle hybride de la forme $p \Rightarrow \boldsymbol{y} = f_p(\boldsymbol{X}_p)$ on $\boldsymbol{X}_p$ l'ensemble des instances correspondant à p puis explore les sous-régions de $\boldsymbol{X}_p$. Ces sous-régions sont définies par des patterns eux-mêmes issus des précédents patterns, mais auxquels sont ajoutées des conditions supplémentaires, c'est-à-dire que les sous-patterns sont plus spécifiques que leurs parents.

Étant donné que l'espace des combinaisons est exponentiel en termes du nombre de variables utilisables, HiPaR utilise un ensemble de stratégies d'élagage qui permettent d'éviter les sous-ensembles les moins prometteurs. Une région est catégorisée comme étant non intéressante si :

— Le modèle local associé à cette région n'améliore pas suffisamment les performances par rapport aux modèles plus étendus déjà étudiés.
— La sous-région se base sur un support trop grand par rapport au jeu de données

entier.

— La sous-région se base sur un support trop restreint.

— La sous-région n'est pas suffisamment distincte du reste des données. Pour évaluer cela, HiPaR utilise la variance inter-classes [40] entre la nouvelle sous-région et le reste des données.

HiPaR base donc sa phase d'énumération sur la RMSE, un seuil de taille minimale et maximale du support θ et ζ, ainsi que sur la valeur de la variance inter-classe.

Malgré ces méthodes d'élagage, il est possible d'obtenir un ensemble de règles hybrides trop étendu. Pour cette raison, HiPaR procède à une phase de sélection des règles après la phase d'énumération. L'objectif de cette sélection est d'obtenir un ensemble limité de règles montrant de bonnes performances prédictives et, dans l'idéal, un recouvrement minime des supports des règles sélectionnées. Contrairement à un arbre de décision qui forme des sous-ensembles strictement disjoints, les règles forment des sous-ensembles qui ne le sont pas. Ceci offre une plus grande flexibilité comparé aux arbres de décision.

La phase de sélection peut se formuler sous la forme d'un problème d'optimisation linéaire en nombre entier :

$$\begin{aligned} \min \quad & \sum_{r_p \in \mathcal{R}} -\alpha_p \cdot z_p + \sum_{r_p, r_q \in \mathcal{R}, p \neq q} (\omega \cdot \mathcal{J}(p,q) \cdot (\alpha_p + \alpha_q)) \cdot z_{pq} \\ \text{s.t.} \quad & \sum_{r_p \in \mathcal{R}} z_p \geq 1 \\ & \forall r_p, r_q \in \mathcal{R}, p \neq q : z_p + z_q - 2z_{pq} \leq 1 \\ & \forall r_p, r_q \in \mathcal{R}, p \neq q : z_p, z_{pq} \in \{0,1\} \end{aligned} \tag{1.6}$$

Chaque règle r_p est associée à une variable binaire. z_p dénote si la règle est utilisée dans le modèle final ou non. La première condition du problème permet de forcer l'algorithme à renvoyer un ensemble de règles non vide.

Les autres termes incluent :

— Le terme α_p associé à cette règle représente le compromis entre le support et l'erreur de la règle. On présente α_p sous la forme $\alpha_p = \bar{s}_{p^\sigma} * \bar{e}_{r_p}$.

— Le terme $\mathcal{J}(p,q)$ est le paramètre de pénalisation de la magnitude des supports des règles sélectionnées, ou Jaccard(p,q) est la valeur de l'indice de Jaccard entre le support des règles p et q $\mathcal{J}(p,q) = \frac{|D_p \cap D_q|}{|D_p \cup D_q|}$.

$\bar{s}(p)^\sigma$ représente la proportion du support de r_p par rapport à la somme de l'ensemble des supports des règles énumérées précédemment, et $\bar{e}_{r_p}$ l'erreur relative de la règle p par

rapport à l'erreur de tous les modèles. On exprime $\bar{e}_{r_p}$ sous la forme $\bar{e}_{r_p} = \dfrac{err(r_p)}{\sum\limits_{r_{p'} \in R} err(r_{p'})}$ et $\bar{s}(p)^{\sigma} = \dfrac{s(p)}{\sum\limits_{r_{p'} \in R} s(p')}$. Le critère α_p pénalise donc les règles au support large et dont les modèles sont peu précis par rapport au reste des règles candidates, et favorise les ensembles de taille plus réduite offrant le plus de gain en termes de performance de prédiction. Le terme $\sigma \in R^+$ représente le paramètre de pénalisation du support contrôlant l'importance accordée à la taille du support de chaque règle : Si $\sigma = 0$, le support des règles n'est pas pris en considération. De manière générale, plus σ est proche de 0, plus l'algorithme favorisera la recherche de modèles performants en ignorant les risques de recouvrement de supports. Le premier terme de la fonction objectif correspond au potentiel de gain de performance par rapport à la taille de leur support des règles sélectionnées.

Contrairement aux modèles basés sur des arbres de décision, les règles hybrides d'HiPaR sont découvertes de manière hiérarchique. Par conséquent, certaines règles peuvent partager des instances de données. Lorsque une nouvelle instance $\boldsymbol{x}^\top$ respecte les conditions de plus d'une règle hybride, la prédiction associée à cette instance sera composée d'une valeur pondérée des prédictions de chacune des règles concernées. La pondération associée à chaque règle est proportionnelle à l'erreur de chaque règle observée sur l'ensemble de validation, exprimée sous la forme $\alpha_{p,\hat{x}} = \frac{\bar{e}(r_p)^{-1}}{\sum_{r_{p'} \in \gamma(\hat{(x)}} err(r_{p'})^{-1}}$. L'erreur est donc normalisée par rapport à l'erreur totale de toutes les règles pertinentes. Cela rend HiPaR plus robuste qu'un modèle de régression linéaire classique ou un arbre de régression, au prix d'une complexité plus importante. Cela étant dit, les règles hybrides sont individuellement des modèles white-box locaux, qui, combinés aux ensembles de conditions associés, permettent d'examiner les variables prédictives les plus importantes et leurs interactions pour une instance $\boldsymbol{x}^\top \in \mathbb{R}^d$.

Une fois que l'on a choisi les modèles qui seront pris en considération, il est nécessaire d'évaluer leur efficacité. Pour cela, on utilisera une approche visant à évaluer leur robustesse.

1.2 Évaluation des modèles de régression

L'évaluation des modèles se base généralement sur leur capacité à fournir des prédictions suffisamment précises. Toutefois, évaluer la précision d'un modèle sur les données sur lesquelles il a été entraîné est insuffisant, car le modèle a été paramétré sur ces mêmes données. Il est donc possible, voire même probable, que le modèle soit surentraîné dans une certaine mesure. Un modèle doit donc être capable de généraliser ses prédictions à des observations qui lui sont inconnues, ce que l'on définit comme étant la *robustesse* (vis-à-

vis du surentraînement) d'un modèle. Cela implique de diviser les données disponibles en deux ensembles : un ensemble d'entraînement (sur lequel on entraînera le modèle) et un ensemble de test (sur lequel on évaluera le modèle). De même, on peut diviser l'ensemble d'entraînement en deux sous-ensembles : un pour l'entraînement et un pour la validation. Le modèle sera entraîné sur ce sous-ensemble. L'ensemble de validation est ensuite utilisé pour orienter l'entraînement de manière à rendre le modèle plus robuste.

Une approche couramment utilisée suivant ce principe est la *validation croisée* [48]. Cette approche consiste à subdiviser l'ensemble des données en n sous-ensembles disjoints les uns des autres. On effectue ensuite n itérations. Pour $i \in [1, n]$, on isole le i-ème sous-ensemble e_i de données du reste. Le modèle est ensuite entraîné sur les sous-ensembles e_j tel que $j \in [1, n]$ et $j \neq i$, puis est évalué sur le sous-ensemble e_i.

De manière générale, que ce soit en régression ou en classification, les métriques se doivent de quantifier l'erreur générale qu'un modèle commet sur l'ensemble de validation, ou d'évaluer la capacité de prédiction du modèle.

Dans le cadre de la régression, en se basant sur y_i les valeurs à prédire et $\hat{y}_i$ les prédictions du modèle, les métriques quantifiant l'erreur d'un modèle de régression les plus utilisées sont :

— la MAE (Mean Absolute Error) $MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$
— la MSE (Mean Square Error), aussi appelée l'erreur quadratique moyenne $MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
— la RMSE (Root Mean Square Error), aussi appelée racine de l'erreur quadratique moyenne $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$

Ces métriques quantifient l'écart entre la valeur de la variable cible pour chaque instance de données et les prédictions du modèle. Il est bon de noter que l'interprétation des valeurs de ces métriques nécessite de les remettre en contexte : L'échelle des valeurs à prédire influe inévitablement sur l'échelle des valeurs de RMSE associée. Pour contourner le problème, on peut choisir de normaliser ou de mettre à l'échelle les valeurs de la variable cible, ou utiliser des métriques indépendantes de cette échelle. Parmi celles-ci, la plus répandue est le coefficient R^2, qui quantifie la qualité de la prédiction du modèle en mesurant la variance expliquée par le modèle. Le R^2 est un rapport entre la variance expliquée et la variance totale du jeu de données. Plus le R^2 s'approche de 1, plus le modèle est performant.

Dans certains cas, évaluer les performances des modèles n'est pas suffisant. Justifier les prédictions d'un modèle est parfois aussi important pour déterminer la confiance qui lui est accordée. Pour résoudre ce problème, on peut utiliser un ensemble de méthodes d'interprétation des modèles.

1.2.1 Méthodes d'interprétation

Les performances des modèles de machine learning ont traditionnellement été le critère le plus important dans le processus de choix de modèle dans la majorité des cas. Toutefois, certains domaines accordent une importance notable à la compréhension du fonctionnement et des prédictions des modèles [23]. On s'intéressera alors particulièrement à l'explicabilité des modèles à disposition. Les notions d'interprétabilité et d'explicabilité sont souvent utilisées de manière interchangeable [3] et représentent le degré auquel un modèle peut être analysé et décrit sans avoir recours à des méthodes supplémentaires [20]. Comprendre le fonctionnement interne d'un modèle est important d'un point de vue éthique et de précision [21]. L'explicabilité prend tout son sens lorsque des individus peuvent être affectés directement par les prédictions d'un modèle. Il est alors important de s'assurer que chaque individu sera traité de manière équitable par le modèle, sans avoir recours à des biais préjudiciables ou injustes à leur égard.

Enfin, l'explicabilité des modèles joue fortement sur leur acceptabilité [49] : on peut prendre dans ce cas l'exemple du domaine de la santé, où les justifications des prédictions sont au moins aussi importantes que la précision du modèle, compte tenu des conséquences possibles sur la santé d'un patient. Enfin, parvenir à expliquer un modèle présente un intérêt en termes d'extraction de connaissances : dans ce genre de cas, un modèle peut être utilisé comme un outil d'analyse. Cela peut permettre de découvrir des facteurs ayant une influence sur le problème étudié et qui auraient pu passer inaperçus jusqu'à maintenant. Cette option dépend toutefois de l'explicabilité des modèles utilisés. L'explicabilité d'un modèle est inversement proportionnelle à sa complexité [20]. Mesurer cette complexité peut donc servir de proxy pour estimer l'explicabilité d'une méthode [23]. De manière générale, on retiendra qu'une bonne explication aura tendance à être concise, ciblée, et aussi le moins abstraite possible [50].

Les méthodes d'interprétation se distinguent en différentes catégories [23]. On dénombre trois grands axes : Nous référerons à ces axes comme le positionnement, l'étendue et la spécificité.

Le positionnement se divise en deux sous-catégories : ante hoc et post hoc. Les méthodes ante-hoc se basent sur l'explicabilité interne des modèles. Les méthodes d'explication ante-hoc sont donc les modèles en eux-mêmes, que l'on définit comme interprétables, ou boîte blanche (white-box). Les composants sont extraits et analysés sans faire usage d'un proxy. Le meilleur exemple de méthode ante-hoc est le modèle linéaire, qui permet d'obtenir une mesure d'importance des variables prédictives en se référant aux coefficients linéaires correspondants. Un autre exemple de modèle boîte blanche couramment utilisé est les modèles pattern-based ou rule-based, comme HiPaR et les arbres de régression.

3. Par la suite, on utilisera exclusivement le terme *explicabilité*

Le Rule-fit [51] est un exemple de modèle rule-based couramment utilisé, qui génère des règles/ensembles de conditions qui seront utilisées comme variables prédictives dans des modèles linéaires pénalisés. Le post-hoc inclut le reste des méthodes d'explication qui forment des approximations du fonctionnement interne du modèle ou de ses prédictions ou évaluent l'impact des variables sur celles-ci. *TREPAN* [52] est une méthode *post-hoc* approximant le fonctionnement interne d'un réseau de neurones à l'aide d'un arbre de décision. *SHAP* est une méthode post-hoc. Celle-ci permet d'évaluer la contribution d'une variable aux prédictions du modèle. Elle considère toutes les permutations possibles de variables et évalue l'impact de l'ajout d'une variable sur la prédiction finale. Cela permet d'estimer sa contribution marginale aux résultats du modèle.

Selon l'étendue des explications ou interprétations que nous pouvons obtenir, on divise les méthodes en deux catégories : les méthodes *globales*, qui expliquent le comportement global du modèle sans réellement prendre en compte les instances individuelles ; ensuite les méthodes locales, qui expliquent les prédictions sur des instances précises des données, ou éventuellement sur un échantillon d'instances. Les méthodes globales les plus couramment utilisées comprennent les Partial Dependence Plots et la RF Feature Importance. Une des méthodes d'explication *locale* les plus connues, *LIME* [53] entraîne des modèles linéaires qui approximent un modèle autour d'une instance cible. Pour ce faire, LIME génère un ensemble de données artificielles basé sur la distribution des cas étudiés. Un modèle linéaire est ensuite entraîné sur ces données artificielles, en pondérant celles-ci en fonction de leur proximité à l'instance cible. Le modèle linéaire est donc un proxy dont les coefficients reflètent l'importance des variables dans la décision du modèle dans un cadre local.

Enfin, le dernier axe décrit l'adaptabilité des méthodes d'explication. Par là, on entend si la méthode est adaptée à un modèle en particulier ou non. On divise donc les méthodes entre les catégories *agnostiques*, applicables à tout type de modèles, et *spécifiques* , structurellement adaptées à un modèle précis. La *RF Feature Importance* est par exemple adaptée uniquement aux modèles basés sur des arbres de décision, et ne peut pas être utilisée sur d'autres types de modèle.

Les méthodes *post-hoc* peuvent être décrites par d'autres caractéristiques. Les méthodes *génératives* sont l'ensemble des méthodes qui génèrent un ensemble de données artificielles sur lesquelles elles se baseront pour fournir des explications. Les méthodes *surrogate* regroupent les méthodes formant des modèles proxy sur le modèle étudié. Les modèles proxy sont ensuite utilisés pour fournir des explications sur le modèle primaire. *SHAP* est un exemple de méthode générative, *Trepan* [52] un autre exemple de méthode surrogate. Ces catégories ne sont pas mutuellement exclusives : *LIME* est un exemple de méthode générative surrogate. Kernel-SHAP [54] est un exemple de méthode non-surrogate générative, et DeepLIFT [55] une méthode non-surrogate non-générative.

Les méthodes que nous utiliserons par la suite sont les *Partial Dependence Plots* et la *Permutation Feature Importance*. Nous avons choisi ces méthodes car elles sont utilisées en protection des cultures, sont adaptées aux modèles black-box et nous fournissent des informations que nous pouvons comparer avec HiPaR.

PDP

Les partial dependence plots (PDP) sont souvent utilisés en protection des cultures. Ils représentent la relation entre une variable prédictive et les prédictions fournies par le modèle en isolant l'action des autres variables. La figure 1.2 est un exemple de partial dependence plots. Les deux premiers PDP représentent la relation entre la variable cible et la température d'une part, et l'humidité d'autre part. Le troisième PDP représente la relation conjointe de la variable cible, de la température et de l'humidité. Les PDP indiquent chaque valeur de la variable prédictive d'intérêt. Les partial dependence plots (PDP) sont souvent utilisés en protection des cultures. Ils représentent la relation entre une variable prédictive et les prédictions fournies par le modèle en isolant l'action des autres variables. La figure 1.2 est un exemple de partial dependence plots. Les deux premiers PDP représentent la relation entre la variable cible et la température d'une part, et l'humidité d'autre part. Le troisième PDP représente la relation conjointe de la variable cible, de la température et de l'humidité. Les PDP indiquent chaque valeur de la variable prédictive d'intérêt et calculent la valeur moyenne de prédiction des observations en fonction de chaque valeur. Les PDP sont particulièrement bien adaptés aux modèles basés sur des arbres de décision. L'étape d'évaluation de chaque valeur d'une variable peut être satisfaite sans se référer aux données, mais à la structure des arbres [32].

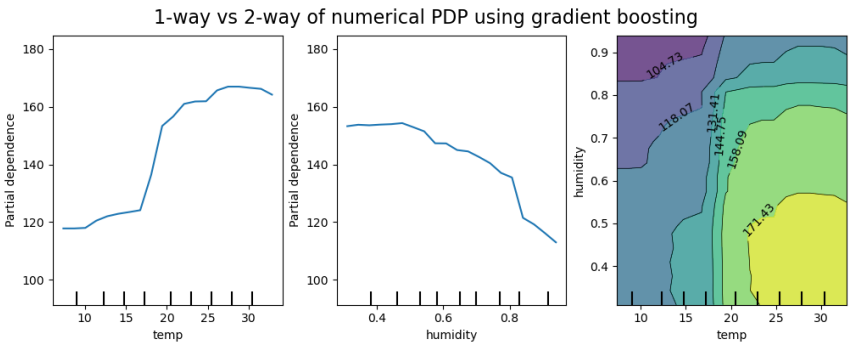


FIGURE 1.2 – Exemple de PDP de la librairie scikit-learn, [56]

Les PDP offrent des opportunités d'explication des modèles [57], qui expliquent leur popularité. Cependant, ils peuvent souffrir de certains problèmes, tels que leur temps de calcul, la difficulté à représenter des relations complexes ou à prendre en compte des corrélations de variables [58, 59]

SaliencyMaps

Les méthodes de saillance (*saliency*) sont des méthodes d'évaluation de l'importance des variables. Elles permettent de quantifier l'importance de chaque variable prédictive sur la prédiction d'un modèle. Dans le cadre de la régression, les méthodes permettant de créer des saliency maps sont les modèles linéaires (et leurs coefficients), LIME, Shapley/SHAP, ou la RF feature importance. Certaines de ces méthodes s'intéressent au modèle en lui-même (MLG), tandis que d'autres perturbent les valeurs de variables pour évaluer leur impact (LIME, RF feature importance) sur le résultat.

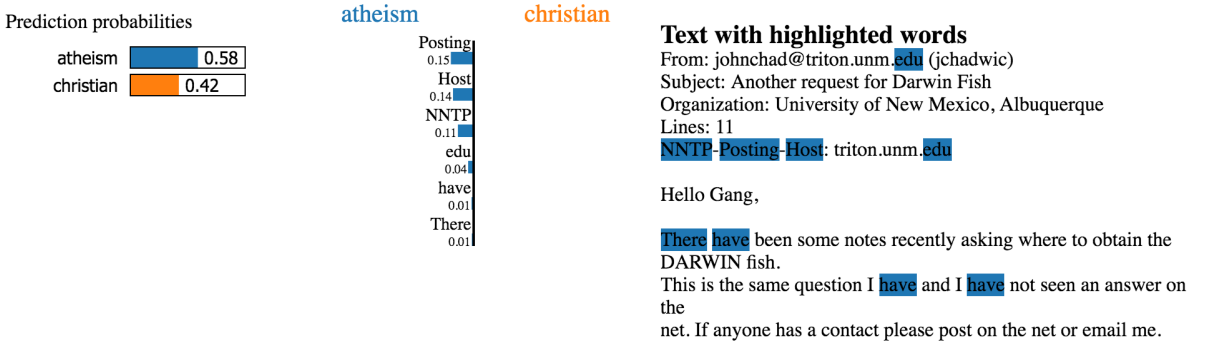


FIGURE 1.3 – Exemple d'explication de Lime [60] appliquée à un classificateur de texte entre les classes "atheism" et "christian".

La figure 1.3 représente le résultat obtenu par LIME à un classificateur de texte. Dans cet exemple, le modèle doit classer le texte dans une catégorie, 'atheism' ou 'christian'. LIME détermine que le classificateur assigne la classe 'atheism' au texte avec une probabilité de 58 %. La partie centrale affiche les facteurs les plus influents sur la prédiction du modèle : dans ce cas, les mots ayant eux le plus d'impact sur le modèle sont 'Posting', 'Host' et 'NNTP'.

Permutation Feature Importance.

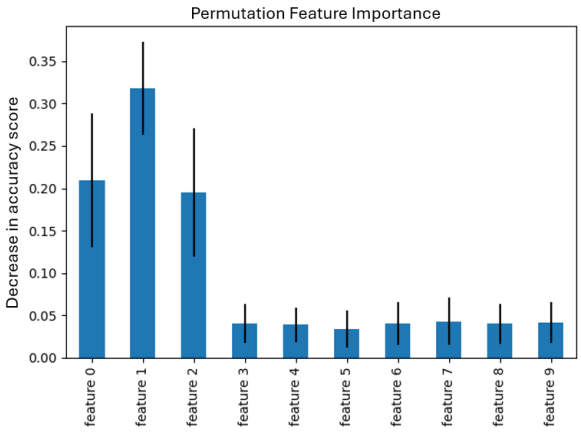


FIGURE 1.4 – Exemple de Permutation Feature Importance de la librairie scikit-learn [61]

La Permutation Feature Importance est une autre méthode d'évaluation de l'importance. La méthode est agnostique et peut être utilisée sur tout type de modèle prédictif. La méthode se base sur une mesure de la baisse de performance des modèles après permutation des valeurs d'une variable : en d'autres termes, on évalue la performance d'un modèle auquel on s'intéresse. Pour chaque variable prédictive, on mélange aléatoirement les

valeurs de celle-ci de manière à rompre le lien entre cette variable et l'objectif. On laisse les valeurs des autres variables inchangées. On évalue ensuite la baisse de performance observée du modèle. Cette baisse représente l'importance qu'à la variable mélangée. Une grande dégradation des performances indique que la variable est influente au sein du modèle.

Une fois que les explications sont générées, il faut être capable de les interpréter. Pour ce faire, il est donc nécessaire de choisir un mode de transmission adapté au cas étudié. Les deux modes de transmission les plus répandus sont les explications textuelles et les explications graphiques. Dans la section suivante, nous nous intéresserons principalement à la représentation graphique de l'information.

1.2.2 Visualisation et machine learning

La visualisation de données (dataviz) est un domaine qui regroupe les méthodes et approches dont le but est de créer des représentations graphiques de grandes quantités de données, quantitatives ou qualitatives. Ces approches ont généralement pour objectif de résumer et de rendre compréhensible des données complexes, idéalement dans le but de mettre en lumière des relations et informations enfouies, qui seraient autrement impossibles à observer, ou de rendre compréhensible des données qui seraient autrement trop complexes. L'objectif primaire de la visualisation des données est donc la transmission de l'information, par le biais de représentations graphiques sélectionnées et adaptées au type de données et d'informations présentes, ainsi qu'aux caractéristiques cognitives et au comportement des utilisateurs. Ces dernières décennies, le traitement et la représentation graphique des données ont été facilités par la démocratisation des ordinateurs, qui a permis de les automatiser. Cela a mené à une formalisation des principes de visualisation des données.

La visualisation suit un certain nombre de principes et de recommandations qui ont pour but de maximiser la compréhension des données. Ces principes comprennent les questions du choix du type de représentation graphique, de la codification des catégories de données (par exemple : le choix de couleurs ou de textures) et de la compréhensibilité/simplicité [62].

Le choix du type de graphique à utiliser dépend du type de données disponibles (quantitatif, qualitatif, ordinal, etc.) ou du type d'information que l'on souhaite mettre en exergue (distribution, quantité, relations, etc.). Le type de représentation choisi doit par conséquent représenter les explications de manière adaptée, c'est-à-dire de façon à ce que l'interface soit informative et intuitive. Par exemple, des méthodes comme LIME et

SHAP, qui évaluent l'importance des variables explicatives, utilisent une représentation sous forme de barres.

Variable	Importance
MedInc	+ 1.83
Longitude	+ 0.64
Latitude	- 0.29
AveRooms	+ 0.14
HouseAge	+ 0.09
Population	- 0.07
AceOccup	0.0
AveBedrms	- 0.0

FIGURE 1.5 – Tableau de valeurs d'importance des variables.

VS

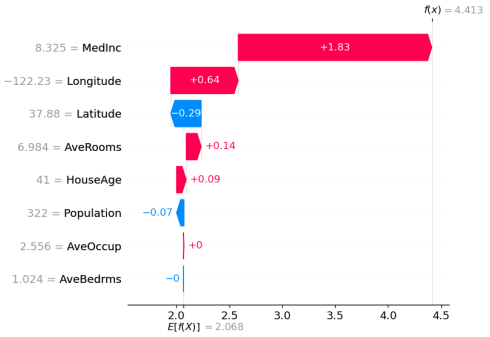


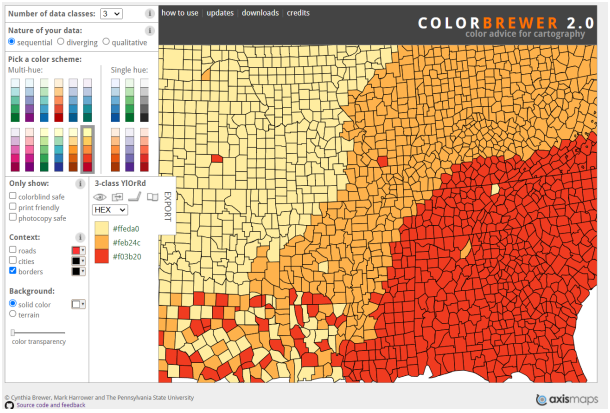
FIGURE 1.6 – Exemple d'explication fournie par SHAP [63].

Dans le cas ci-dessus, le modèle prédit la valeur médiane des logements pour chaque district de Californie en se basant notamment sur le revenu médian des habitants, la taille des logements ou encore la position géographique.

Lorsque l'impact d'une variable sur la prédiction d'un modèle est le plus important, la visualisation sous forme de barres est adaptée, car elle permet de représenter les informations les plus importantes rapidement et sous une forme simple. Dans l'exemple ci-dessus, l'échelle de l'impact des variables est estimable d'un coup d'œil en jugeant de la longueur des barres et en les comparant entre elles (Figure 1.6). De plus, le type d'impact (négatif ou positif) est facilement observable en utilisant des couleurs faciles à distinguer (ici, rouge et bleu). En se fiant à Shap, on peut voir que la variable ayant l'impact le plus important (positivement) sur le prix médian des logements en Californie d'après le modèle est le revenu médian des ménages.

Au-delà de la question du type de représentation privilégiée, il revient ensuite au concepteur de choisir une approche favorisant la compréhensibilité de l'outil de visualisation par les utilisateurs. De manière analogue aux questions portant sur l'explicabilité des modèles d'apprentissage automatique, un compromis entre complexité et explicabilité doit souvent être trouvé. Une explication extrêmement complexe sera plus précise et plus proche de la réalité, mais sera en pratique impossible à interpréter. Ce principe englobe par exemple le nombre de graphiques ou d'axes de visualisation à faire apparaître et la quantité d'information à afficher. On notera par conséquent que toute action permettant de réduire la complexité de la représentation graphique est à privilégier. Plutôt que de représenter toutes les données et informations disponibles, il est donc nécessaire de réduire la dimensionnalité des données sous une forme simplifiée et compréhensible. Cela demandera souvent de procéder à une sélection des informations à représenter en suivant une mesure de priorité.

Le choix de la codification de l'information est un élément central pour la mise au point d'une méthode de visualisation, et est lié à la question de la complexité de la visualisation des données. De manière générale, on essaie de codifier la visualisation de manière à distinguer et contraster les informations. Par exemple, l'utilisation d'un ensemble de couleurs distinctes pour représenter les données est un moyen efficace pour contraster ces dernières. Ce médium est toutefois limité par la capacité du cerveau humain à distinguer les couleurs, particulièrement dans des situations où celles-ci peuvent être mélangées ou difficiles à distinguer. On considère généralement que l'être humain peut distinguer jusqu'à 10 couleurs de manière simultanée. Dans les faits, il est souvent recommandé de limiter ce nombre à 6 pour optimiser l'évaluation visuelle des résultats. Qui plus est, il est bon de noter qu'une proportion non négligeable de la population est victime de daltonisme sous une forme ou une autre. L'objectif de la visualisation des données étant de faciliter la transmission de l'information, on prendra en compte ce facteur en sélectionnant des gradients ou des ensembles de couleurs adaptés.



(a) Exemple d'outil de sélection de gradients de couleur, [64]

Exemple d'outil permettant de sélectionner un gradient de couleur adapté à une approche de dataviz. Cet outil permet de sélectionner au maximum douze couleurs différentes et offre des options annexes, notamment la possibilité de générer des gradients adaptés aux personnes daltoniennes.

1.2.3 Interface utilisateur

Le choix et la complexité de la méthode d'explication, combinés à sa représentation graphique, dépendent également du public ciblé par la méthode. Supposons qu'une méthode d'explication d'un modèle d'apprentissage automatique soit destinée à un domaine d'activité en particulier. On peut diviser grossièrement les publics cibles en fonction de leur degré de connaissance en apprentissage automatique et dans le domaine d'application. Un *néophyte* est un individu ne disposant pas de connaissances dans l'un ou l'autre domaine. Un *expert terrain* aura de solides connaissances dans le domaine d'application, mais généralement peu en apprentissage automatique. Un expert Machine Learning possède des connaissances théoriques sur les modèles utilisés, mais le plus souvent peu sur le domaine d'application [23].

Chaque catégorie a des besoins et des demandes spécifiques en matière de méthode d'ex-

plication : les *néophytes* souhaiteront obtenir des explications simples et concises leur permettant d'interpréter un résultat obtenu par le biais d'un modèle. Cette catégorie d'utilisateurs n'aura pas nécessairement de demande concernant le fonctionnement des modèles ou leurs méthodes d'explication. D'une certaine manière, on peut considérer que les explications fournies doivent être les plus simples et chercher au maximum à ne pas troubler l'utilisateur.

Dans le cas des *experts terrains*, on doit fournir des informations plus précises et plus approfondies. Les *experts terrains* sont aussi particulièrement demandeurs en matière de confiance accordée aux modèles et à leurs explications : cette catégorie souhaitera le plus souvent comprendre *comment* le modèle parvient à un résultat. Cela peut se faire en intégrant des éléments factuels des données en lien avec les explications fournies. De manière générale, on peut dire que l'expert de terrain accordera plus d'intérêt aux méthodes globales que la catégorie précédente. La confiance accordée à la méthode définit si l'expert sera capable de mettre en application ses connaissances en conjonction avec la méthode, et donc in fine l'efficacité de la méthode. Il est donc nécessaire d'adapter la méthode d'explication de manière à permettre aux experts d'extraire de nouvelles informations et connaissances. Il faut également noter que les experts de terrain sont les plus susceptibles de remettre en question les explications fournies, en particulier si celles-ci vont à l'encontre de leurs propres informations.

La dernière catégorie regroupe les individus disposant de connaissances en Machine Learning sans pour autant disposer de connaissances de terrain. Dans ce cas, la méthode de visualisation devra s'attarder sur le fonctionnement du modèle plutôt que sur les interactions avec les données terrain. Cette catégorie se concentre donc principalement sur le fonctionnement interne d'un modèle et sera généralement peu intéressée par des méthodes d'explication destinées aux catégories précédentes.

On cherche désormais à confirmer la pertinence des principes explicités précédemment. Pour ce faire, nous nous baserons sur un ensemble de cas d'application se concentrant sur l'analyse des modèles, en prenant pour cible les experts agronomes.

Chapitre 2

Cas d'études

L'explicabilité des modèles est un élément influant sur l'acceptabilité de l'utilisation de l'approche de machine learning en protection des cultures. On s'intéressera aux cas de maladies des plantes impactant fréquemment les cultures en France, comme le mildiou de la vigne et la cercosporiose de la betterave. Sans traitement efficace, ces maladies peuvent avoir un impact significatif sur les rendements de ces cultures [65]. On utilisera des jeux de données de protection des cultures contenant principalement des informations d'ordre météorologique. Ces données incluent également des données historiques sur l'incidence de maladies sur plusieurs années. Des modèles de machine learning ont été entraînés et testés sur ces données, avec pour objectif de prédire l'incidence ou la date d'apparition des symptômes de maladies. Ces modèles seront ensuite analysés afin d'estimer leur précision et leur explicabilité.

Données épidémiologiques de la cercosporiose de la betterave sucrière

La cercosporiose de la betterave sucrière (SBC) est une maladie fongique se développant dans les feuilles de certains légumes, parmi eux la betterave sucrière. La maladie est polycyclique [66, 67]. Les plantes infectées développent des tâches qui, au cours de l'été, provoquent leur assèchement total. La phase de dissémination initiale a lieu au printemps et dépend de facteurs tels que le vent et la pluie [68]. En dehors de la saison, les inocula survivent sur des résidus de récoltes et dans le sol. Dans le cas des sols, la survie d'un inoculum est inversement proportionnelle à la profondeur d'enfouissement et peut y survivre jusqu'à 3 ans [69]. La perte de rendement engendrée peut être significative si la maladie n'est pas régulée. Les principales méthodes de contrôle de la cercosporiose incluent l'enfouissement profond des résidus de récolte, une rotation des cultures longue, l'utilisation de variétés résistantes et enfin l'utilisation de fongicides. Il est donc particulièrement important de pouvoir prédire et

Figure 2.1 – Feuille de betterave sucrière infectée par la cercosporiose

contrôler le développement de cette maladie pour déterminer quand appliquer les traitements et donc préserver les rendements. Les données expérimentales utilisées dans notre travail ont été collectées sur différentes exploitations. Celles-ci ont pu être collectées grâce à l'aide de l'ITB (Institut Technique de la Betterave) sur une période s'étendant de 2009 à 2020 à travers la France. Chaque parcelle d'expérimentation a été observée sur une année définie pour éviter tout risque de redondance et donc de surentraînement. Les relevés de progression des symptômes ont été effectués de manière hebdomadaire sur une centaine de plants de betterave par des experts agronomes. L'incidence de la maladie est définie comme la proportion (en pourcentage) des feuilles de betterave montrant des tâches symptomatiques de la Cercosporiose. Ces contrôles ont eu lieu de la période d'apparition usuelle des symptômes (généralement au mois de mai) jusqu'à la période de récolte (deuxième quinzaine de septembre). Au total, le jeu de données pré-traité et nettoyé contient des observations de l'incidence de la cercosporiose sur 1 235 parcelles.

On attache une date à chaque parcelle correspondant à la date d'apparition des symptômes de la Cercosporiose. Cette date correspond à la date à laquelle la proportion de feuilles infectées est supérieure à 10 %. L'incidence de la cercosporiose en fin de saison est définie comme l'incidence maximale observée pendant la période s'étendant du 25 août au 15 septembre.

Données épidémiologiques du mildiou de la vigne

Le mildiou est une autre maladie fongique qui peut se développer aux dépens de plusieurs espèces de plantes, comme les pommes de terre ou la vigne. Il se manifeste par des tâches brunes ou des moisissures blanches, typiquement sur les feuilles qui finissent par dépérir. Dans les cas plus extrêmes, des rameaux, voire des pieds entiers peuvent être touchés. À l'inverse de la betterave, la maladie affectant un pied de vigne peut avoir des conséquences beaucoup plus graves à long terme, la betterave étant une culture annuelle, alors qu'un pied de vigne met plusieurs années avant de pouvoir être exploité dans le cadre d'une activité vinicole. Les précipitations et la température sont les facteurs de dissémination et de développement les plus importants du mildiou [70]. Les méthodes de contrôle du mildiou incluent l'utilisation de fongicides, l'entretien des pieds de vigne de manière à limiter le nombre de pousses et de feuilles basses, ainsi que le drainage d'eau stagnante à proximité

FIGURE 2.2 – Feuille de betterave sucrière infectée par la cercosporiose

des pieds.

Les données concernant le développement du mildiou de la vigne ont été collectées sur une période allant de 2010 à 2017. Les parcelles d'expérimentation sont réparties à travers la France et ont été surveillées par l'Institut Français de la Vigne et du Vin (IFV). Pour chaque parcelle considérée, les observations ont été effectuées sur des rangées de vigne non traitées. Chaque rangée utilisée pour les observations est entourée par deux rangées elles-mêmes non traitées. Ceci est fait pour éviter que la rangée soit traitée par inadvertance avec des fongicides. La rangée centrale est surveillée de manière hebdomadaire pour mesurer le développement du mildiou. Le développement du mildiou est mesuré à travers la proportion de feuilles de vigne montrant des signes d'infection par Plasmopara viticola. Ces relevés ont eu lieu entre le débourrement, c'est-à-dire la fin de l'hibernation hivernale des pieds de vigne, et la fermeture des grappes, entre le mois de mars et la deuxième quinzaine de juillet. Le contrôle peut également être arrêté si l'incidence atteint 100 %, c'est-à-dire quand toutes les feuilles sont symptomatiques. Le nombre total d'observations individuelles est de 9 407 observations hebdomadaires, sur un total de 713 parcelles.

Pour chaque parcelle, on attribue une date d'apparition des symptômes. Cette date correspond à la date à partir de laquelle la proportion de feuilles infectées est supérieure à 1 %. On définit l'incidence du mildiou en fin de saison comme la valeur maximale d'incidence sur l'ensemble de la saison.

Données météorologiques

Les variables météorologiques ont été obtenues par l'intermédiaire de la base de données SAFRAN, agrégée et entretenue par Météo-France. SAFRAN subdivise le territoire français à l'aide d'une grille de taille 8×8 Km et conserve les données météorologiques pour chaque cellule de la grille [71]. Des relevés journaliers d'humidité, de température moyenne, de vitesse du vent, de précipitations et d'ensoleillement ont été utilisés pour composer des variables météorologiques pour chacune des maladies.

Dans le cas de la cercosporiose, les informations sont collectées sur une période s'étendant de janvier à juin. Ces données sont ensuite agrégées sur des périodes de 15 jours consécutifs. Chaque période correspond à la première ou à la deuxième moitié de chaque mois concerné. Pour permettre l'utilisation de modèles demandant des données tabulaires plutôt que des séries temporelles, chaque valeur agrégée est stockée sous la forme d'une variable. Par conséquent, chaque variable météorologique est subdivisée en autant de variables que de périodes de 15 jours observées. Ces variables sont nommées selon une convention permettant d'identifier la période de temps et le type de variable concerné. La première partie décrit le mois en utilisant les trois premières lettres de celui-ci, suivies par un 'A' ou un 'B', dénotant la première et la deuxième quinzaine du mois. La deuxième

partie décrit la nature de la variable d'un point de vue météorologique et la manière dont cette variable est calculée. Les suffixes des variables sont décrits dans le tableau 2.1. Par exemple, *JanA-ndRHm60* est la variable correspondant au nombre de jours (**nd**) ou à l'humidité relative (**RH**) pendant la première quinzaine (**A**) de janvier (**A**) où la valeur moyenne a été supérieure à 60 % (**m60**). Les données ont été collectées sur une période s'étendant de 2009 à 2020. Les parcelles se répartissent principalement dans le nord de la France et en région parisienne, même si quelques parcelles sont observables dans le centre du pays.

Dans le cas du mildiou de la vigne, les variables décrivent soit les conditions météorologiques au moment de la collecte des données, soit la somme de celles-ci pendant les quatre semaines précédant cette date. Par exemple, la variable ETP correspond à l'évapotranspiration au moment de l'observation. ETP-4w correspond à la somme de l'évapotranspiration sur les quatre semaines précédentes. Ces jours sont comptabilisés à partir du mois de janvier. Cette longueur de quatre semaines a été choisie en se basant sur des indications d'experts concernant le développement du mildiou. Les seules exceptions sont le nombre de jours considérés comme "pluvieux" et "secs". Les parcelles ont été observées entre 2010 et 2017. Elles se trouvaient majoritairement en Nouvelle-Aquitaine et en Occitanie, ainsi que dans une moindre mesure en Pays de la Loire et dans le Centre-Val de Loire.

Nom	Variable
RHm$\underline{X}$	Humidité relative moyenne inférieure à X (X = {60, 65, 80, 90})
H87	Index d'humidité égal à 87%
H87$\underline{Y}$	Index d'humidité égal à 87 pendant au moins (Y = {6, 10}) hours
Tm$\underline{X}$	Températures moyennes supérieures à (X = {15, 20}) degrés
Tm$\underline{X}$Tinf$\underline{YZ}$	Températures moyennes supérieures à (X = {15}) degrés mais plus basse que (Y = {10}) pour au moins (Z = {3}) heures
TbloX	Nombre de jours ou les températures sont définies comme inhibitrices au développement de la cercosporiose pendant plus de (X = {3,6}) heures.

TABLE 2.1 – Description des variables météorologiques utilisées pour modéliser la dynamique de développement de la cercosporiose de la betterave sucrière (SBC). Les températures sont considérées comme inhibitrices quand comprises entre 10° et 38°

Variables à prédire

Les données météorologiques et agronomiques constituent les jeux de données à notre disposition pour les 4 variables à prédire. Ces variables sont l'incidence en fin de saison et la date d'apparition des symptômes. Ces jeux de données se présentent sous la forme suivante :

➔ Le jeu de données de la cercosporiose en fin de saison contient 1 235 parcelles, décrites par 367 variables qui seront utilisées dans les modèles, parmi lesquelles on

trouve une variable catégorielle et 366 variables numériques. 364 de ces variables numériques correspondent à 26 variables météorologiques agrégées sur une période de 15 jours de janvier à juin. Les deux variables numériques restantes correspondent à la date d'apparition des symptômes et au jour des semis. La variable catégorielle est une variable agronomique nommée zone_à_risque, un indicateur basé sur les connaissances des experts concernant la sensibilité des parcelles à la cercosporiose. Le jeu de données utilisé pour calculer la date d'apparition de la cercosporiose est le même que pour la prédiction de l'incidence en fin de saison. Les variables à prédire décrites dans ce jeu de données, à savoir la date d'apparition des symptômes et l'incidence en fin de saison, sont représentées sous la forme du numéro du jour de l'année durant lequel les symptômes apparaissent et du pourcentage de feuilles symptomatiques.

➔ Le jeu de données du mildiou contient 359 parcelles et 17 variables prédictives numériques dans le cadre de la prédiction des dates d'apparition, ainsi que 700 parcelles pour l'incidence en fin de saison en septembre. Les variables décrites incluent l'évapotranspiration, les précipitations et des mesures de température. Sept de ces variables sont limitées à la date de récolte des données, et quatre représentent la somme de ces variables sur une période de quatre semaines précédant la date de récolte des données. 3 correspondent à des mesures de température agrégées. Les 3 variables restantes représentent les précipitations : une mesure des jours considérés comme "secs", une mesure des jours "humides" et la valeur agrégée des précipitations (en mm).

Chapitre 3

Modélisation et analyse de la dynamique des maladies des plantes

Dans l'objectif de prédire l'incidence de la cercosporiose et la date d'apparition des symptômes du mildiou de la vigne et de la cercosporiose de la betterave sucrière, on entraîne des modèles (Lasso, Random Forest, Gradient Boosting et HiPaR) sur les données présentées dans chap :caset. Ces modèles de régression seront d'abord comparés en fonction de leurs performances et de leur complexité. Puis, on analysera leurs propriétés et les informations que ceux-ci peuvent nous apporter. Plus particulièrement, on examinera HiPaR pour évaluer l'intérêt des modèles de complexité intermédiaire, dans notre cas un modèle pattern-based, du point de vue de l'explicabilité.

3.1 Protocole d'entraînement et de test

3.1.1 Choix et optimisation des modèles

L'évaluation des modèles de notre benchmark suit les protocoles standards d'évaluation, tout en prenant en compte certaines spécificités propres à la protection des cultures. L'approche classique de validation croisée consiste à échantillonner un ensemble d'entraînement depuis l'ensemble des données de manière indiscriminée, et à utiliser les données restantes comme ensemble de test. Un tirage indiscriminé implique que des données provenant d'une même année peuvent se trouver à la fois dans l'ensemble d'entraînement et l'ensemble de test. Cependant, dans le cadre agronomique, la plupart des cultures sont des cultures annuelles, ce qui fait que des sous-ensembles sont de facto présents dans le jeu de données agrégé. Les variations des conditions météorologiques d'une année à l'autre font que ces sous-ensembles peuvent être fortement distincts les uns des autres.

Ceci pose problème pour la validation croisée : le fait d'utiliser des données d'une même année dans les ensembles d'entraînement et de test augmente le risque de sur-apprentissage.

Il existe un risque non négligeable que les modèles semblent performants de prime abord, car ils sont entraînés sur des données d'une même année, mais qu'ils ne puissent pas généraliser ces prédictions sur de nouvelles données.

Qui plus est, la réalité applicative impose de modifier la validation croisée pour l'adapter au cadre de la protection des cultures : les modèles de protection des cultures sont entraînés sur des ensembles de données regroupant des données historiques. En d'autres termes, on entraîne les modèles sur les années précédentes dans l'objectif de prédire les années à venir. Par conséquent, on privilégiera une approche de validation croisée basée non pas sur un échantillonnage aléatoire de l'ensemble des données, mais sur une segmentation année par année des données. À chaque itération, une année sera isolée du reste des données et utilisée comme ensemble de test. Les autres années seront utilisées pour entraîner les modèles. Cette approche permet de simuler le protocole usuel et d'évaluer la robustesse des modèles sur plusieurs années.

Les modèles testés pour chaque année sont le Lasso, le Random Forest, Gradient Boosting et HiPaR. Pour chaque année, le meilleur modèle est sélectionné en optimisant les hyperparamètres de chaque type de modèle. La phase d'énumération d'HiPaR étant trop longue, par exemple quand le seuil de support est trop bas, il n'est pas raisonnable d'envisager d'optimiser les paramètres de cette phase. Par conséquent, on entraînera un seul modèle avec un seuil de support de 30 %, c'est-à-dire que les règles sélectionnées incluront au minimum 30

Le coefficient de détermination linéaire de Pearson (R^2) est utilisé pour mesurer les performances des modèles. Ce choix est fait pour permettre de comparer les modèles de manière uniforme et ne pas dépendre de l'échelle des variables à prédire.

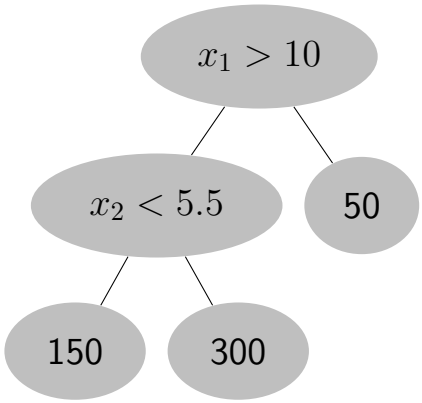
3.1.2 Mesure de complexité

Pour tenter d'estimer l'explicabilité des modèles, on utilisera la complexité des modèles comme proxy de la même manière que décrite dans la section Introduction. Pour évaluer la complexité des modèles, on se basera sur les approches décrites dans [40, 23], qui consistent à dénombrer le nombre d'éléments utilisés dans un modèle. Un élément peut être un coefficient linéaire non nul ou une condition sur une variable prédictive. Cette dernière mesure est applicable aux modèles basés sur des arbres de décision comme les random forests ou le gradient boosting car chaque nœud de chaque arbre représente une condition appliquée sur une variable prédictive. Ce nombre peut être très important quand les modèles sont composés d'un grand nombre d'arbres. Dans le cas d'un modèle mélangeant conditions sur les variables et modèles linéaires, on décompte tous ces éléments pour en évaluer la complexité.

Ces mesures sont toutefois imparfaites. Les modèles peuvent être basés sur des éléments similaires, mais différer dans leur structure. Par exemple, le gradient boosting est entraîné de manière itérative : chaque sous-modèle est lié aux modèles qui le précèdent et le suivent. Ainsi, les éléments de chaque sous-modèle sont liés aux éléments des autres sous-modèles, et sont donc difficilement interprétables en l'état. Ceci entre en conflit avec les Random Forest, où chaque sous-modèle, c'est-à-dire chaque arbre binaire, est formé de manière indépendante par rapport aux autres. Dans ce cas, on compare donc des éléments qui sont techniquement de même nature, mais en réalité divergents.

En suivant cette approche, un modèle Lasso est généralement moins complexe qu'un modèle HiPaR ayant sélectionné plusieurs règles. Ceci est dû au fait que la complexité du Lasso ne dépend que du nombre de coefficients linéaires différents de 0, alors que la complexité d'HiPaR dépend à la fois du nombre de conditions et des coefficients des modèles linéaires locaux.

Si on considère l'arbre de régression T prédisant la valeur de la variable cible y en fonction de x_1 et x_2 :



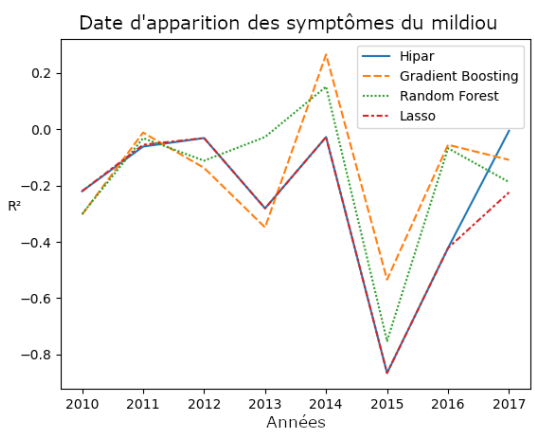
Sa complexité $c(T)$ est de 5 car l'arbre est constitué de 5 nœuds. De la même manière, si on considère la règle R

$$C_1 \wedge C_2 \wedge C_3 \wedge C_4 \Rightarrow \boldsymbol{y} = 3x_1 + 4x_2 - 4x_3 + 8. \tag{3.1}$$

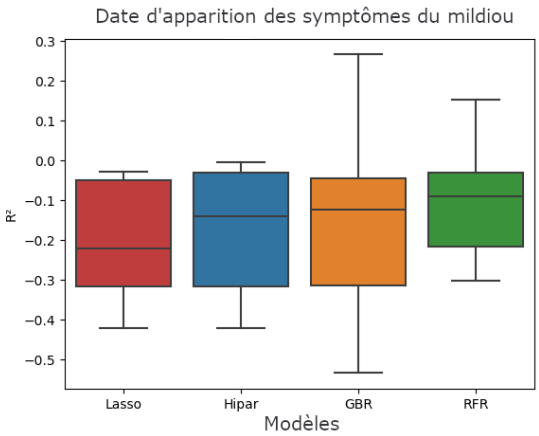
La complexité $c(R)$ de la règle est de 8 car la règle comporte 3 conditions (C_1,C_2,C_3,C_4) et utilise 4 coefficients linéaires (3,4,-4).

3.1.3 Évaluation des modèles année par année

Nous commencerons par évaluer nos modèles en nous intéressant à leurs résultats année par année, présentés dans les graphiques ci-dessous. Cela nous permettra d'évaluer leurs performances prédictives ainsi que leur robustesse, qui sont les principaux critères de sélection des modèles destinés aux OAD.

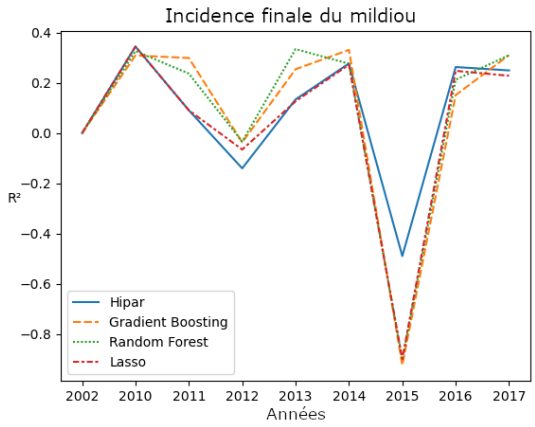


(a) Variation du R^2 année par année

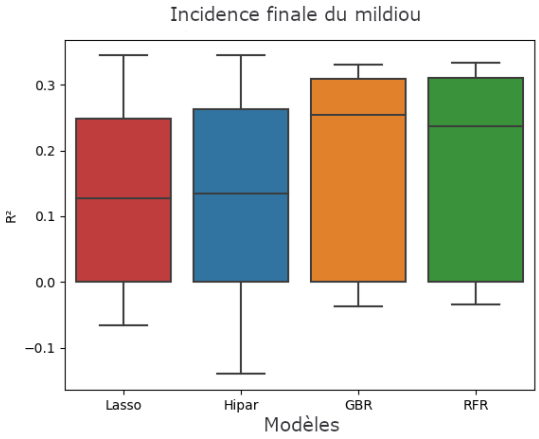


(b) Distribution des valeurs de R^2 année par année

FIGURE 3.1 – Date d'apparition des symptômes du mildiou

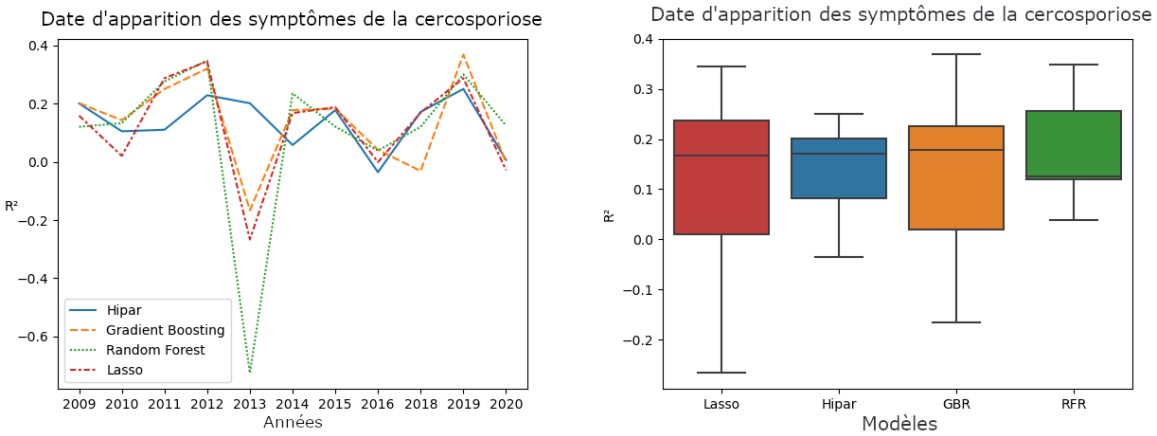


(a) Variation du R^2 année par année



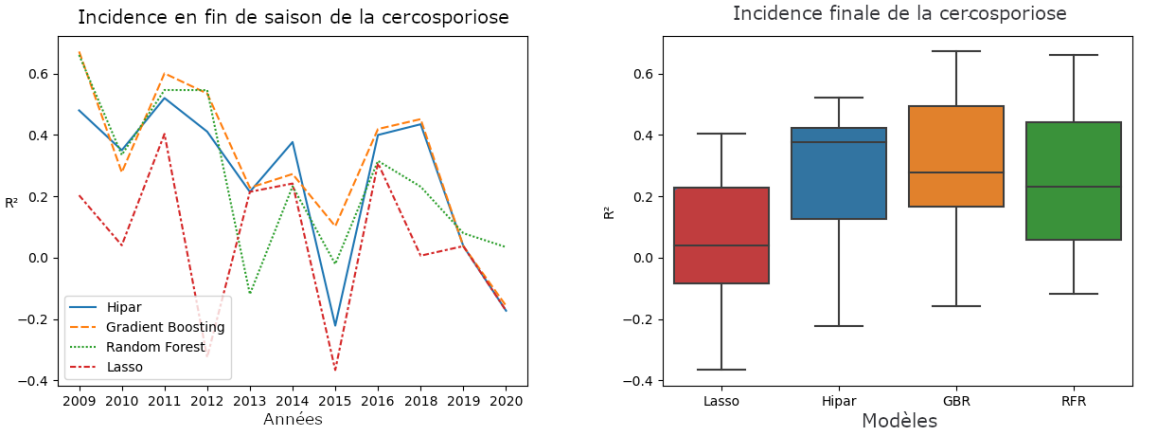
(b) Distribution des valeurs de R^2 année par année

FIGURE 3.2 – Incidence du mildiou en fin de saison



(a) Variation du R^2 année par année (b) Distribution des valeurs de R^2 année par année

FIGURE 3.3 – Date d'apparition des symptômes de la cercosporiose



(a) Variation du R^2 année par année (b) Distribution des valeurs de R^2 année par année

FIGURE 3.4 – Incidence de la cercosporiose en fin de saison

On constate que les performances varient de manière drastique d'une année à l'autre. Dans le cas de la prédiction de la cercosporiose de la betterave, et en dépit des variations notables, les performances de nos modèles sont en moyenne acceptables dans le cadre de la protection des cultures, un R^2 de 0,3 étant considéré comme satisfaisant compte tenu de la complexité des données. On note également que certaines années sont difficiles à modéliser, quelle que soit la variable à prédire (incidence ou date d'apparition) ou la méthode. L'année 2015 est particulièrement problématique et semble mettre en échec la quasi-totalité des modèles. Une explication possible est que cette année était exceptionnellement sèche et chaude [72], avec des températures supérieures de 1 °C par rapport aux normales et une baisse de précipitations de 15 % au niveau national. Ces facteurs météorologiques ont pu déstabiliser nos modèles, en raison de leur nature hors norme, en particulier dans la mesure où les températures et les précipitations sont des facteurs connus de dissémination

et de développement du mildiou et de la cercosporiose [70, 68].

Une tendance descendante notable est observable pour l'incidence en fin de saison de la cercosporiose. Une hypothèse possible permettant d'expliquer cette tendance est l'influence du changement climatique : celui-ci provoque une déviation de chaque nouvelle année par rapport aux précédentes, ce qui entraîne une dégradation des performances des modèles. Aucune tendance particulière ne peut être observée dans les autres cas.

3.1.4 Compromis performance-complexité

Nous allons maintenant évaluer les performances des modèles en fonction de leur complexité. Cela nous permettra de visualiser les gains de performances obtenus par rapport à la perte d'interprétabilité des modèles. La Figure 3.5 représente le compromis entre la complexité et les performances des modèles de machine learning observés. L'axe des abscisses illustre la complexité des modèles en utilisant une échelle logarithmique. Les ordonnées correspondent à la médiane des indices R^2 de chaque modèle obtenus pendant la validation croisée. Plus un modèle se situe dans le coin en haut à gauche, plus il s'approche de l'optimum du compromis entre les performances et la complexité des modèles : faible complexité et bonne performances. Comme suggéré dans [40], les modèles plus complexes, comme les forêts aléatoires et le gradient boosting, font preuve de plus hautes performances en contrepartie d'une complexité accrue. Les modèles Lasso, nos modèles de base, sont souvent les moins performants. HiPaR se positionne entre ces deux groupes de modèles.

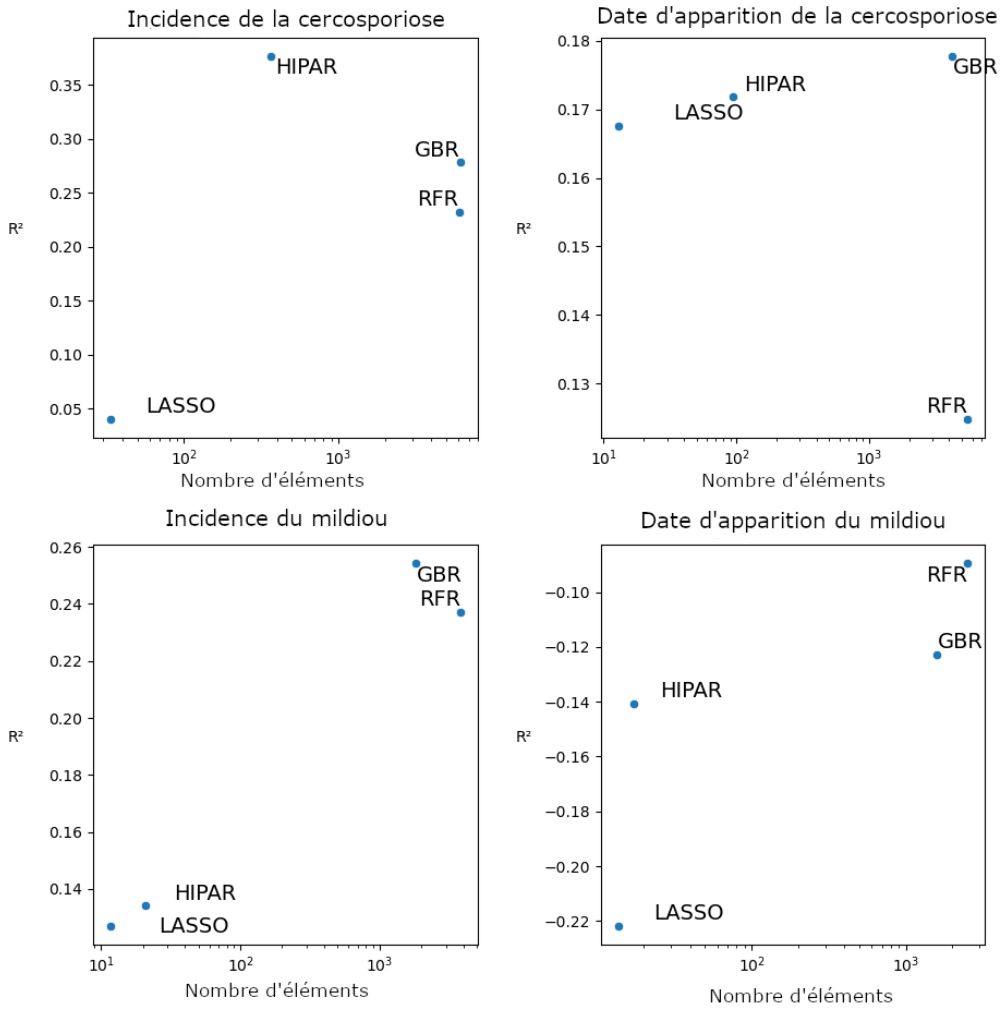


FIGURE 3.5 – Coefficient R^2 des différents modèles de machine learning en fonction de leur complexité. L'axe des abscisses correspond au nombre d'éléments composant chaque modèle (en échelle logarithmique). Les ordonnées correspondent aux valeurs médianes de R^2 obtenues pendant la validation croisée. GBR correspond aux modèles de régression par gradient boosting, et RFR aux forêts d'arbres aléatoires.

On met en exergue que les performances peuvent varier de manière significative entre les différents cas. Tous les modèles semblent éprouver des difficultés à prédire les dates d'apparition des symptômes du mildiou de manière précise, compte tenu des valeurs négatives de R^2 observées (en bas à droite).

Le R^2 varie entre 0,12 et 0,26 pour la prédiction de l'incidence finale de cette même maladie (en bas à gauche), ou le gradient boosting tire son épingle du jeu. HiPaR est proche du Lasso, ce qui signifie qu'aucune ou peu de règles ont été utilisées.

Les différences de performance entre les prédictions de la date d'apparition des symptômes et de l'incidence finale du mildiou dans ce jeu de données peuvent s'expliquer par le faible nombre d'observations utilisées pour la prédiction de la date d'apparition : 359 observations contre 700 pour l'incidence en fin de saison.

La portée des variables agrégées est également limitée en ce qui concerne l'historique. Par cela, on veut dire que ces variables ne couvrent que les 4 semaines précédant une observation spécifique. Même si cela nous aide à confirmer le compromis présent entre complexité et performance, cela a pour conséquence de réduire le nombre de jeux de données intéressants à étudier et d'obtenir des modèles peu efficaces.

Les résultats obtenus sur le jeu de données de la cercosporiose sont plus intéressants. La médiane des scores de R^2 obtenus sur les dates d'apparition varie entre 0,13 et 0,18, le gradient boosting en tête, suivi par HiPaR puis le Lasso. Le Random Forest est moins performant de manière significative. On remarque que si le gradient boosting est le meilleur modèle en termes de R^2, il n'en reste pas moins que le gain de précision est marginal par rapport au Lasso (0,01 R^2), surtout si on prend en compte le gain de complexité induit (de l'ordre de 10^3 éléments).

Dans le cas de l'incidence en fin de saison, le R^2 varie entre 0,05 et 0,35. Dans ce cas, HiPaR est plus performant que le modèle linéaire classique, avec une différence de 0,3 en R^2. HiPaR présente de meilleures performances que les modèles ensemblistes, dépassant le gradient boosting de 0,05 R^2 et le random forest de 0,1. Ce gain est significatif dans la mesure où HiPaR se distingue de ces modèles en termes de complexité par un ordre de grandeur. Les modèles ensemblistes présentent des performances similaires.

Bien que se basant sur des données similaires, les modèles se placent différemment en termes de compromis complexité-performance. On dénombre ici 3 cas de figure :

- ➔ Dans le cas de la modélisation du mildiou, les modèles simples et intermédiaires sont largement dépassés par les modèles ensemblistes. On pourra donc privilégier ces derniers, la complexité des modèles n'apportant rien de significatif dans ce cas.
- ➔ Dans le cas de la prédiction de la date d'apparition des symptômes de la cerco-

sporiose, les gains de performances par rapport à la complexité des modèles sont au mieux marginaux. Il est donc plus intéressant de favoriser les modèles simples comme le Lasso, car les pertes de précision (s'il y en a) sont largement compensées par l'interprétabilité des modèles.

➔ Dans le cas de la prédiction de l'incidence en fin de saison de la cercosporiose, HiPaR dépasse les autres modèles dans une majorité d'années, tout en gardant sa place de modèle de complexité intermédiaire. Il est donc potentiellement intéressant de l'utiliser, tant pour ses performances que pour les informations qu'il pourrait apporter.

Le compromis observé dans nos cas d'usage démontre l'utilité des modèles de complexité intermédiaire, qui sont capables de s'approcher des modèles ensemblistes dans des cas où les modèles linéaires en sont incapables. On peut supposer que ces modèles sont capables de modéliser des interactions plus complexes et, par conséquent, d'encapsuler des informations intéressantes à examiner. Dans la section ci-dessous, nous chercherons à analyser les modèles obtenus pour la prédiction de l'incidence de la cercosporiose de la betterave sucrière en fin de saison en 2009. Ce choix a été fait en raison des performances des modèles sur cette année : le gradient boosting affiche un R^2 de 0,67 et de 0,66 pour les random forests, 0,47 pour HiPaR et 0,3 pour Lasso. Ces valeurs de variance expliquée permettent de supposer que les informations issues de ces modèles sont pertinentes et crédibles. Dans le cas des modèles boîte blanche, il sera possible d'examiner directement les éléments des modèles, sans passer par une méthode d'explication. Les modèles boîte noire demanderont l'utilisation de méthodes d'explication couramment utilisées pour vérifier si les informations obtenues sont concordantes.

3.2 Interprétation : Incidence de la cercosporiose de la betterave sucrière

Les techniques d'interprétation utilisées sont la feature importance ranking (ou importance des variables), les partial dependence plots (ou graphiques de dépendance partielle), et l'analyse de règles hybrides. La première technique nous permet de déterminer quelles sont les variables qui ont le plus d'impact sur les prédictions d'un modèle. Les partial dependence plots et l'analyse des règles hybrides permettent de mettre en exergue des effets de seuil entre les variables prédictives et à prédire. Par effet de seuil, on entend des cas où la relation entre les variables prédictives et la variable prédite varie de manière significative en fonction de la région de l'espace des données dans laquelle on se situe, c'est-à-dire en fonction de seuils portant sur les variables prédictives. Un exemple parlant de fonctions incluant un effet de seuil sont les fonctions non monotones. Les méthodes basées sur des patterns, comme HiPaR, sont particulièrement adaptées pour détecter ce genre d'effets.

De plus, ces méthodes permettent d'affiner les interactions entre les variables prédictives pour chaque sous-région délimitée par les règles.

Notre approche est divisée en trois étapes :

- ➔ On comparera en premier lieu les modèles ensemblistes et linéaires afin d'expliciter les différences de structures entre nos modèles de complexité haute et basse. On utilisera la feature importance ranking sur les modèles complexes pour obtenir les rangs d'importance des variables. Les coefficients linéaires du Lasso nous permettront de classer les variables en fonction de leur importance dans le modèle. Ces rangs d'importance seront utilisés dans un parallel coordinate plot qui nous permettra de visualiser les différences de structure des modèles.
- ➔ Deuxièmement, on cherchera à comparer les modèles ensemblistes et HiPaR. Pour ce faire, on utilisera les Partial Dependence Plots des modèles ensemblistes en conjonction avec les conditions utilisées dans les règles hybrides d'HiPaR pour déterminer s'il est possible d'observer des similarités.
- ➔ Enfin, on comparera les règles hybrides et les modèles locaux d'HiPaR avec le Lasso à l'aide de leurs coefficients linéaires afin d'observer les différences de structure et d'en tirer des informations.

Importance de variables. Une manière simple d'évaluer la manière dont les données sont modélisées par un modèle de machine learning est d'évaluer le rang d'importance des variables prédictives afin d'estimer l'influence de chacune d'entre elles sur les prédictions du modèle. Ce rang peut être basé sur différentes mesures, comme les coefficients d'un modèle linéaire, ou un score d'importance calculé a posteriori pour les modèles boîte noire. Pour les méthodes ensemblistes, qui sont basées sur des arbres de décision, on utilise des mesures adaptées spécialement à celles-ci. On utilisera l'importance de variable par permutation de scikit-learn [61]. Cette approche offre une mesure d'importance en faisant varier les valeurs des variables prédictives associées à chaque observation de $\boldsymbol{X}$. La baisse de performance observée est ensuite utilisée pour déterminer si le modèle dépend de la variable dont on a fait varier la valeur : plus les performances diminuent, plus la variable est importante pour le modèle. Pour les modèles linéaires, comme le Lasso, on utilisera les coefficients linéaires du modèle. Les coefficients du Lasso représentent la contribution réelle des variables prédictives à la prédiction du modèle. Ces coefficients peuvent être de signe positif ou négatif. Un coefficient négatif ne veut pas nécessairement dire que la variable associée est moins impactante, mais que son influence sur la prédiction finale est négative. Il est donc nécessaire d'utiliser les valeurs absolues de ces coefficients pour obtenir un rang d'importance exploitable. On compare les rangs du Lasso, du RFR et du GBR dans le tableau Figure 3.6. On exclut HiPaR en raison de sa nature hybride : l'importance d'une variable par permutation ne permet pas de déterminer si le rang d'une variable est lié à

un effet de seuil (si cette variable est utilisée dans une condition d'une règle hybride) ou à un des modèles linéaires formant HiPaR. Classer en utilisant les coefficients linéaires utilisés par HiPaR est également impossible dans la mesure où celui-ci est composé d'un ensemble de modèles linéaires dont les supports (et donc la prévalence) varient.

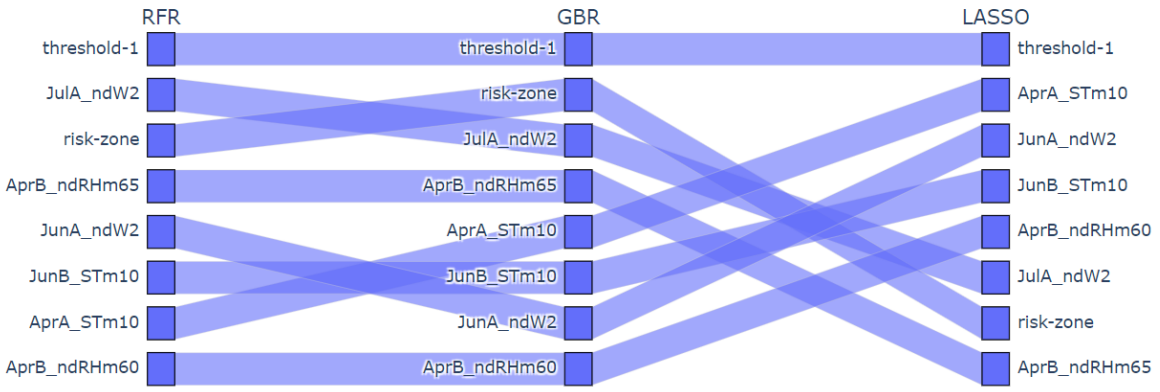


FIGURE 3.6 – Parallel coordinates plot contenant les variables les plus importantes pour les modèles random forest, gradient boosting et Lasso. Pour chaque modèle, on sélectionne les 4 variables les plus importantes de chaque modèle. Pour chaque modèle, les variables situées au delà de la 4e peuvent être de rang inférieur à ce qui est affiché.

Comme on peut l'observer, RFR et GBR se comportent de manière similaire : parmi les quatre variables les plus importantes, la seule différence entre les deux modèles est l'ordre dans lequel les variables sont classées. La variable *threshold-1* est la plus importante quel que soit le modèle. Cette variable représente le jour où les premiers symptômes de cercosporiose apparaissent sur la parcelle. RFR et GBR font usage de la variable *risk-zone*, qui est un indicateur expert. Le modèle linéaire n'inclut pas cette variable parmi les quatre variables les plus importantes.

Si les scores d'importance des modèles basés sur les arbres de décision nous apportent des informations sur les modèles, ceux-ci n'indiquent pas si une variable a une influence négative ou positive sur la prédiction du modèle. Dans le cadre de la protection des cultures, il s'agit par exemple de savoir si un facteur climatique est favorable au développement d'une maladie ou, au contraire, inhibe sa croissance. On peut toutefois obtenir cette information pour le Lasso, car les coefficients sont signés : un coefficient négatif indique une influence inhibitrice sur la croissance de la maladie, tandis qu'un coefficient positif sous-entend que la variable associée est un facteur aggravant l'impact de la maladie.

Variable	Coefficient
Threshold-1	-41
AprA-STm10	27.64
JunA-ndW2	-23.39
JunB-STm10	22.24
AprB-ndRHm60	14.73

TABLE 3.1 – 5 coefficients linéaires les plus importants du Lasso (classés par ordre décroissant)

Le sens des variables présentes est comme suit :

threshold-1 : Date d'apparition des symptômes

AprA-STm10 : Somme des températures moyennes journalières supérieures à 10°C pendant la première quinzaine d'avril

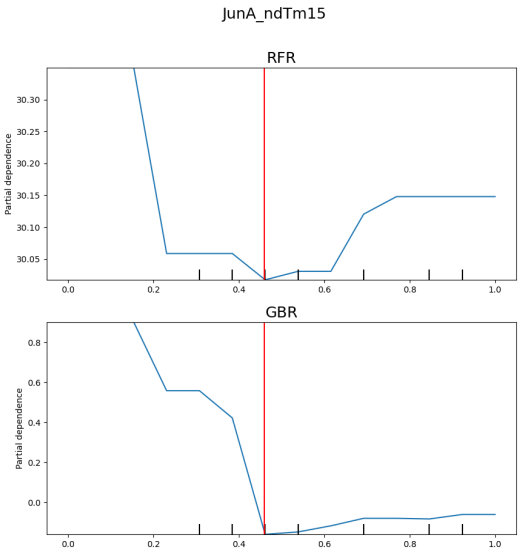
JunB-ndW2 : Nombre de jours de la deuxième quinzaine de juin ou la vitesse moyenne du vent est supérieure ou égale à 2m.s^{-1}

JulB-STm10 : Somme des températures moyennes journalières supérieures à 10°C pendant la deuxième quinzaine de juillet

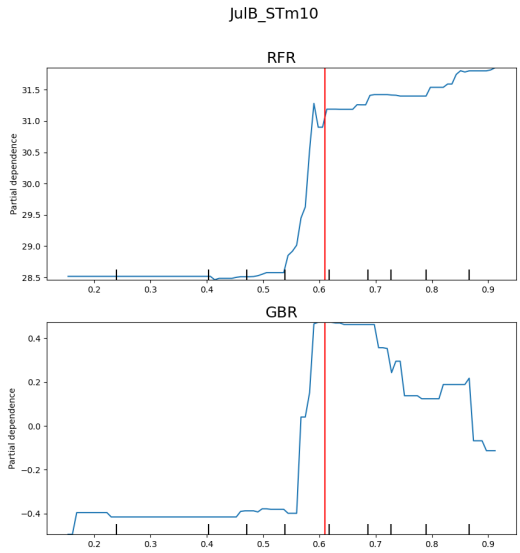
AprB-ndRHm60 : Nombre de jours de la deuxième quinzaine d'avril ou l'humidité relative est supérieure à 60%

La Table 3.1 nous indique logiquement que plus les symptômes apparaissent tardivement, plus l'incidence finale sera faible. L'incidence prédite par le modèle linéaire a tendance à être plus importante lorsque les températures et l'humidité en juin et en avril sont élevées, alors qu'un climat venteux semble inhiber la croissance de la maladie. Ces résultats sont toutefois à relativiser dans la mesure où le R^2 du Lasso est de 0,3. Ceci étant dit, ces variables sont également utilisées par les méthodes ensemblistes, ce qui confirme qu'elles ne sont au minimum pas décorrélées de l'incidence.

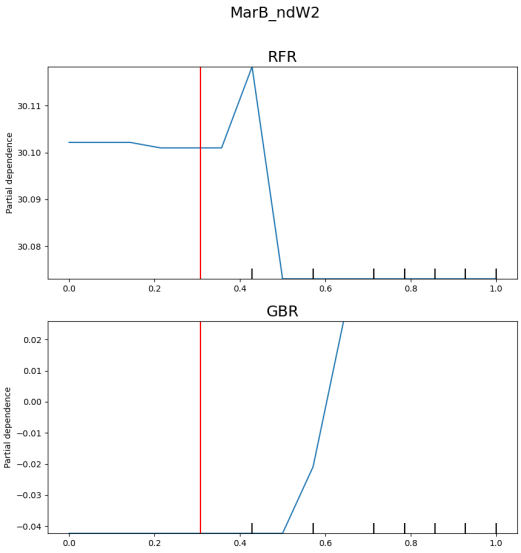
Effets de seuil. Comme énoncé plus tôt, les méthodes de régression pattern-based sont construites de manière à détecter les effets de seuil sur les variables prédictives. Dans le cas d'HiPaR, ces effets sont caractérisés de manière explicite dans l'ensemble de conditions de chaque règle hybride. Pour déterminer si le modèle est parvenu à modéliser ces effets, nous examinerons les règles utilisées par HiPaR sur notre cas d'usage. Les seuils et informations obtenus seront comparés aux modèles complexes de notre benchmark, ici les RFR et GBR. Dans la mesure où ces modèles sont basés sur des estimateurs composés de seuils, on peut observer des effets de seuils dans ces modèles à l'aide de partial dependence plots (PDP) décrits dans la sous-section 1.2.1. Cette méthode est largement utilisée et nous permet d'observer le comportement de la variable cible en relation avec une variable prédictive.



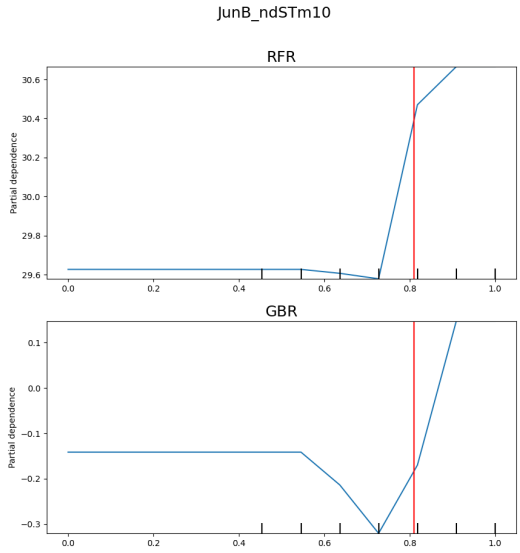
(a) Nombre de jours ou la température moyenne était supérieure à 15°C durant la première quinzaine de juillet (valeurs réelles et normalisées)



(b) Somme des températures journalières ou la température moyenne journalière était supérieure à 10°C pendant la deuxième moitié de juillet. (valeurs réelles et normalisées)



(c) Nombre de jours ou la vitesse moyenne du vent était supérieure à 2km.s^{-1} dans la deuxième moitié du mois de mars (valeurs réelles et normalisées)



(d) Nombre de jours ou la température moyenne était supérieure à 10°C pendant la deuxième quinzaine de juin (valeurs réelles et normalisées)

FIGURE 3.7 – PDP des variables prédictives *MarB-ndW4*, *JulB-STm10*, *JunA-ndTm15* et *JunB-STm10* sur les modèles random forests et gradient boosting. La ligne rouge — représente un seuil d'HiPaR appris par HiPaR dans une de ses règles.

Dans notre cas d'usage, HiPaR sélectionne 3 règles hybrides dont les conditions sont listées dans la référence table :hipar-conditions. Comme on a pu observer précédemment, les seuils (en rouge) utilisés par HiPaR correspondent assez précisément à des changements de comportement des variables prédictives, observables grâce aux PDP. Les variables seuils ne sont pas les plus importantes d'après les scores d'importance, et n'influent donc pas de manière significative sur les prédictions. Toutefois, le fait que les seuils d'HiPaR concordent avec les modèles ensemblistes sur ces points semble indiquer que ces variables agissent possiblement comme des proxys pour des interactions de variables plus complexes : par exemple, les partial dependence plots 3.7a et 3.7c représentent deux conditions provenant d'une même règle hybride. Compte tenu du fait que ces variables décrivent respectivement des facteurs de développement et de dissémination de la cercosporiose, et que ces conditions définissent des parcelles où les températures ont été potentiellement plus basses (au détriment de la maladie) et où le printemps venteux a favorisé la maladie, on peut supposer que la modélisation résultante accordera plus d'importance à des variables intervenant avant la période estivale et privilégiera des variables agissant en conjonction avec le vent au printemps, facteur de dissémination.

Règle 1	*JunA-ndTm15* < 8, *risk-zone=false*
Règle 2	*JulB-STm10* < 324
Règle 3	*MarB-ndW2* ≥ 4, *JunB-ndTm15* < 13

TABLE 3.2 – Conditions des règles hybrides utilisées par HiPaR pour la prédiction de l'incidence en fin de saison de la cercosporiose de la betterave sucrière

Autrement dit, HiPaR subdivise les données en suivant les indices de dangerosité donnés par les experts agronomes (*risk-zone*), les nombres de jours pendant les mois de juin et juillet où les températures sont inférieures à 15 °C. (*JunA-ndTm15*, *JunB-ndTm15*), 10°C *JulB-STm10*, et le nombre de jours de mars considérés comme venteux (*MarB-ndW2*).

Interactions de variables. Chacune des conditions listées dans la Table 3.2 est associée à un modèle linéaire local (dans notre cas, un Lasso). Ces modèles nous révèlent des interactions entre les conditions et les coefficients linéaires de ces modèles. Des variations de coefficients d'un modèle à l'autre peuvent indiquer que certains facteurs n'interviennent que sous certaines conditions. Ces règles sont formées de manière à raffiner le modèle de base, ici le Lasso, entraîné sur l'ensemble du jeu de données.

Des 368 variables prédictives du jeu de données original, l'ensemble des modèles linéaires (locaux ou général) utilisent entre 25 et 55 variables. Cela représente entre 6,7 % et 15 %

des variables utilisées dans tous les modèles. On notera que les modèles locaux sont systématiquement moins complexes que le modèle général, comme l'indique la Table 3.3.

	Règle 1	Règle 2	Règle 3	Modèle par défaut
Règle 1	25	8	3	6
Règle 2		28	6	16
Règle 3			26	12
Modèle par défaut				55

TABLE 3.3 – Nombre de variables utilisées de manière concomitante dont les coefficients sont différents de 0 parmis les sous-modèles utilisés par HiPaR pour la prédiction de la cercosporiose de la betterave sucrière.

Le tableau 3.3 permet d'examiner le nombre de variables communes entre les différents modèles locaux et avec le modèle général. Ce nombre de variables en commun est relativement bas. Cela signifie que chaque modèle se base sur des facteurs différents pour prédire l'incidence finale, et peut indiquer que les conditions de chaque règle permettent d'isoler des facteurs de développement de la cercosporiose différents et conditionnés. La Figure 3.8 illustre l'intensité et le type d'impact de 16 de ces variables pour tous les modèles, locaux ou général.

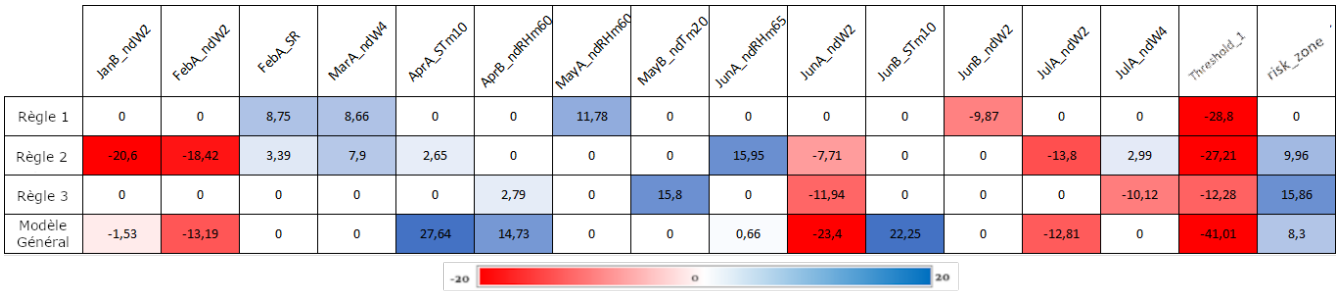


FIGURE 3.8 – Représentation colorisée des coefficients linéaires des règles hybrides utilisées par HiPaR. Les cases blanches ☐ signifient que le coefficient linéaire associé est strictement nul, ce qui signifie que la variable associée n'est pas utilisée par le modèle.

Le premier constat que l'on peut tirer de ce tableau est que, comme on s'y attendait, la date d'apparition (*threshold-1*) est la variable la plus importante, constante entre les modèles, et a un impact négatif sur l'incidence en fin de saison. Les variables *risk-zone* et *JunA-ndW2* (le nombre de jours venteux de la première quinzaine de juin) sont utilisées dans tous les modèles, à l'exception de la première règle. Ceci peut être expliqué par la présence de ces variables dans l'ensemble de condition associé à cette règle (Table 3.2). On peut supposer que l'utilisation de ces conditions permet in fine d'effacer l'influence de celles-ci sur les prédictions du modèle local en prenant en compte leur importance en amont du modèle.

La première règle (ligne "Règle 1" du tableau de la figure 3.8) peut être interprétée de la manière suivante : Les parcelles moins vulnérables à la cercosporiose selon les connaissances des experts (*risk-zone=false*) et soumises une majorité du temps à des températures basses pendant la première quinzaine de juin (*JunA-ndTm15* < 8) sont plus exposées à la cercosporiose lorsque l'humidité en mai (*MayA-ndRHm60*), l'exposition au vent en mars (*MarA-ndW4*)et les précipitations en février (*FebA-SR*) augmentent. Un mois de juin venteux (*JunB-ndW2*) semble avoir un effet inhibant sur la maladie. Le dernier facteur semble confirmer qu'une période estivale sèche contrarie la cercosporiose, le vent étant un facteur d'assèchement. On observe de manière générale que les variables estivales n'ont que peu d'impact dans ces conditions, comparé aux autres périodes. On peut supposer que cela signifie que, dans ce cas, la majorité du développement de la cercosporiose a lieu au printemps.

La deuxième règle suggère que des températures basses en juillet (*JulB-STm10* < 324) affaiblissent la cercosporiose face au vent en janvier, février, juin et juillet (*JanB-ndW2*, *FebA-ndW2*, *JunA-ndW2*, *JulA-ndW2*). De la même manière, un mois de juin humide (*JunA-ndRHm65*) ou un mois de mars venteux (*MarA-ndW4*) semble aggraver la situation. Un mois de juillet venteux (*JulA-ndW4*), un mois de février pluvieux (*FebA-SR*) et un mois d'avril chaud (*AprA-STm10*) diminuent l'incidence finale de la cercosporiose.On peut observer qu'une grande importance est donnée aux variables de vent hivernales, qui correspondent à la période d'hibernation des inocula. Compte tenu de cette information, on peut supposer que les températures élevées en juillet bloquent le développement de la cercosporiose. Les facteurs initiaux d'infection, comme le taux de survie des inocula, deviennent donc d'autant plus importants et le vent a un impact négatif sur ceux-ci.

La 3e règle intervient quand le mois de mars est venteux (*MarB-ndW4* ≥ 4) et un mois de juin ou des journées froides ont été observées (*JunB-ndTm15* < 13). Dans cette situation, les températures élevées en mai (MayB-ndTm20) et l'humidité élevée en avril (*MayB-ndTm20*) sont liées à une augmentation de l'incidence de la cercosporiose. Inversement, le vent en juin (*JunA-ndW2*) et en juillet (*JulA-ndW4*) a un impact négatif sur sa croissance. On notera que cette règle ne fait pas usage des variables hivernales, mais s'intéresse plutôt à la période printanière et estivale.De plus, on observe une baisse drastique de l'importance de la date d'apparition des symptômes et une hausse notable de l'impact de l'évaluation de la vulnérabilité des parcelles (*risk-zone*). On peut donc supposer que dans les cas où le vent en phase de dissémination est suffisamment présent, mais que les températures estivales sont basses, la date d'apparition sera moins impactante, car la période principale de développement sera inhibée. L'incidence finale dépendra alors plus de la vulnérabilité déjà existante de la zone étudiée.

Le modèle général semble agréger l'ensemble de ces facteurs, tout en utilisant d'autres

variables qui ne sont pas présentes dans les modèles locaux. On suppose que cela est dû au fait que le modèle général est entraîné sur l'ensemble des sous-ensembles délimités par les règles hybrides. Cela se traduit par l'utilisation de variables (exemple : *JunB-STm10*) qui semblent avoir une importance notable au niveau global, mais qui peut disparaître sous certaines conditions (*risk-zone=false*).

De manière générale, on observe que les informations obtenues, comme l'influence des températures estivales, sont en accord avec les connaissances actuelles [68]. Certaines règles semblent toutefois sortir en partie de ce cadre et sous-entendre que certains facteurs et périodes (vents hivernaux) jusqu'à maintenant mis de côté pourraient avoir une influence dans certains cas bien précis.

HiPaR, notre modèle de complexité intermédiaire, nous permet donc d'obtenir des informations plus précises que le Lasso, tout en restant plus interprétable et transparent que les méthodes ensemblistes. Les explications fournies jusqu'à présent ont toutefois été mises en forme par une analyse humaine (non automatisée), qui peut être difficile à comprendre pour un non-initié et qui demande de comprendre et de maîtriser l'ensemble des modèles et des méthodes pour prendre sens. Dans l'objectif de fournir des modèles interprétables et d'être capable d'en tirer des explications intelligibles, on cherchera donc à automatiser cette analyse de manière à la rendre plus compréhensible. L'extension naturelle semble donc, dans notre cas, de mettre au point une interface graphique dont le but sera d'exploiter les modèles et résultats obtenus pour mettre en lumière des informations pertinentes pour l'utilisateur.

Chapitre 4

Représentation visuelle des modèles et explication des prédictions

Comme vu précédemment, l'acceptabilité des modèles de Machine Learning dépend en grande partie de leur clarté. On a montré précédemment que même des modèles de complexité intermédiaires demandent de posséder une compréhension de la structure des modèles et des données pour parvenir à interpréter les résultats. Cela peut poser problème quand un individu cherche à comprendre le fonctionnement et le résultat d'un modèle mais ne remplit pas ces prérequis. Celà peut également avoir lieu quand la quantité de données ou de variables est trop grand. Dans le cadre de la protection des cultures, les agronomes chercheront à comparer les résultats à leurs connaissances issues de la bibliographie ou du terrain. Si les résultats et ces connaissances entrent en conflit, être capable de déterminer la source de cette divergence de manière claire permet de détecter des problèmes inhérents aux modèles (dans le cas ou les résultats seraient erronés) ou de fournir des hypothèses sur le fonctionnement du pathosystème jusqu'à maintenant inconnues à condition que le modèle soit suffisamment performant. Cette condition est importante en terme de confiance accordée par les experts au modèle et de fiabilité des informations fournies.

Dans le cas d'un modèle de type boîte noire, il est possible d'associer une méthode d'explication ou d'interprétation des résultats à ce modèle. Cette approche permet de simplifier et de faciliter l'analyse et la compréhension des modèles. Dans l'idéal, on privilégiera une méthode transformant le moins possible les résultats pour conserver l'intérêt des modèles interprétables. Dans le cas de la protection des cultures, la dimension géographique des résultats devra être prise en compte. [73, 74, 75].

Le choix de la catégorie d'utilisateurs finaux doit être pris en compte pendant la mise au point de l'interface. Dans le cadre de la protection des cultures, les publics concernés sont des agronomes. On reprendra les notions de néophytes, experts terrain et experts Machine Learning décrites dans la partie III.

Les agronomes se situent, eux, entre les experts de terrain et machine learning : ils disposent de connaissances agronomiques exhaustives qu'ils souhaiteront comparer aux résultats du modèle, et peuvent posséder des connaissances en machine learning et en statistiques. On exclura donc de proposer des moyens de diagnostic et de compréhension avancée des modèles, ceux-ci étant d'intérêt limité pour nos deux catégories. Les agronomes seront plus attirés par des méthodes d'explication leur permettant d'examiner les modèles avec un horizon plus large à l'aide d'explications plus techniques, mais peuvent parfois être intéressés par des focales plus proches. Le contraste sera d'autant plus important pour les agronomes afin qu'ils puissent confronter leurs connaissances aux résultats des modèles.

L'étude des modèles a permis de mettre en évidence que notre modèle de complexité intermédiaire, HiPaR, semble être en mesure de fournir des informations d'intérêt. En l'état, ces informations sont toutefois très générales et nécessitent de comprendre le fonctionnement de HiPaR pour pouvoir interpréter pleinement les résultats. Pour résoudre ce problème, nous avons conçu une interface graphique permettant de visualiser l'impact de chaque facteur météorologique sur les prédictions des modèles et de les comparer. Afin de faciliter l'interprétation des résultats fournis par HiPaR, nous avons donc cherché à utiliser les informations fournies dans une interface graphique. La maquette de démonstration a été mise au point en collaboration avec la société de prestation informatique ENEO [1] via DigitAg, l'institut technique de la betterave (ITB) et l'institut français de la vigne et du vin (IFV). Pendant la phase de réflexion sur l'ergonomie de la maquette, nous avons également bénéficié des contributions de Gonzalo Mendez, chercheur en visualisation des données à l'université ESPOL [2] en Équateur, que nous avons accueilli en tant que chercheur visiteur à l'INRIA durant ma thèse. Cette maquette a pour but de permettre aux agronomes d'interpréter les résultats obtenus par HiPaR sur les jeux de données de protection des cultures à disposition.

Pour accomplir cette tâche, il était d'abord nécessaire de solliciter les experts en épidémiologie végétale des deux instituts techniques, l'ITB et l'IFV. Cette phase de la mise au point est nécessaire pour s'assurer que les explications fournies soient comprises et acceptées par les utilisateurs. L'approche choisie a été de solliciter les experts et le prestataire à plusieurs reprises pour itérer sur les réflexions et besoins. Des réunions régulières ont donc été organisées en plusieurs étapes : la collecte et la formalisation des besoins, puis la conception et la réalisation informatique.

1. https ://eneo.fr
2. Escuela Superior Politécnica del Litoral

4.1 Collecte des besoins et définition des utilisateurs cibles

Différentes réunions, dont le but était de collecter et de formaliser les besoins, ont eu lieu de novembre 2022 à février 2023. Nous détaillons ici les sujets de discussion de ces réunions.

10-11-2022

Les approches de machine learning et les résultats portant sur les performances et la complexité des modèles discutés dans le Chapitre 3 ont été présentés aux experts de l'Institut Français de la Vigne.
Discussions autour des résultats et de la qualité des données.
Discussions sur les aspects géographiques et temporels d'intérêt.

23-11-2022

Les approches de machine learning et les résultats sur les performances et la complexité des modèles ont été présentés aux experts de l'Institut Technique de la Betterave.
Discussions autour des résultats et des informations/explications des modèles.
Discussions sur l'intérêt de données supplémentaires (en nombre de cas et de facteurs météorologiques) et des modèles parcimonieux.

Ces réunions nous ont permis de déterminer que les points communs les plus importants étaient la notion de contraste, les dimensions spatiales et temporelles des informations fournies, ainsi que la nature des informations et explications fournies.

Par contraste, nous entendons la comparaison d'informations diverses permettant de les contextualiser et de mettre en évidence des points d'intérêt. En termes de contraste, la plupart des experts étaient intéressés par la possibilité de comparer les facteurs importants/influents de plusieurs régions ou modèles entre eux, notamment le modèle général, c'est-à-dire le modèle entraîné sur l'ensemble des données d'entraînement. Concernant l'ergonomie de l'outil, la notion de flexibilité est revenue plusieurs fois. La capacité pour l'utilisateur de regrouper des parcelles choisies et d'analyser le comportement des modèles seulement sur ces parcelles était particulièrement importante.

01-02-2023

Réunion avec les experts de l'ITB, IFV et le prestataire ENEO.
Synthèse des besoins et choix des approches privilégiées.

Cette réunion a permis de faire la synthèse des demandes des experts. Les points abordés concernaient les approches de visualisation privilégiées, les modes de sélection des parcelles

et des variables, ainsi que l'ergonomie de l'interface.

En suivant les demandes des experts, on formalise les différents axes de développement de la maquette. Quatre axes se distinguent et vont orienter le type de représentation graphique privilégié. Ces axes sont :

— L'aspect géographique, ou la capacité de visualisation des parcelles sur une carte de France.
— L'aspect de visualisation des données agronomiques, notamment l'incidence des maladies des cultures agrégée à différents niveaux (Pays, commune, région personnalisée, parcelle).
— La capacité d'observer les variables les plus influentes sur le développement des maladies des plantes d'après un modèle basé sur des règles hybrides, dans notre cas HiPaR.
— La capacité de contraster les informations de différentes régions.

4.2 Données et information à visualiser

4.2.1 Règles : Coefficients, conditions

Les règles hybrides sont définies par le set de conditions et les coefficients du modèle linéaire associé. Les coefficients nous indiquent l'impact des différents facteurs et des tendances générales. Les conditions, en conjonction avec les coefficients, ajoutent un contexte et fournissent des informations sur la signification de ceux-ci et modifient donc leur interprétation. Ces conditions peuvent aussi représenter des effets de bascule ou de seuil.

4.2.2 Support : Appartenance et prévalence

Le support d'une règle permet de déterminer dans quelle mesure celle-ci est prévalente parmi les données. Un grand support implique une tendance, tandis qu'un support réduit implique plutôt des cas marginaux ou exceptionnels.

Pour compléter les explications issues des modèles, il est possible d'ajouter des informations externes. Ces informations permettent d'ajouter du contexte, et donc d'améliorer la compréhension des explications. Par exemple, on peut afficher les performances des modèles, les facteurs météorologiques ou encore l'incidence des maladies, ce qui permet d'évaluer la fiabilité des explications ou de contraster les conditions auxquelles les parcelles sont soumises.

4.2.3 Dimension géographique

Un aspect de la représentation graphique en protection des cultures est le facteur géographique des informations et des cas étudiés. Les parcelles d'expérimentation sont généralement non contiguës. Cette méthode est choisie pour obtenir un éventail large de cas à étudier soumis à des conditions différentes. Sans cela, on pourrait craindre de n'obtenir que des modèles peu robustes et peu généralisables. Qui plus est, les experts agronomes peuvent être familiers avec des régions particulières. Être capable de représenter les explications fournies et de les associer à une zone géographique améliore donc l'acceptabilité des explications, la position géographique faisant partie des facteurs avec lesquels les experts peuvent associer leur propre expertise le plus aisément. La capacité à représenter et utiliser les caractéristiques géographiques des données est un facteur à prendre en compte pour évaluer l'efficacité des méthodes d'explication.

Générale

La première échelle géographique à considérer est l'échelle globale. On s'y intéresse typiquement au modèle général, c'est-à-dire au modèle entraîné sur l'ensemble des données. Cette échelle permet d'observer la dynamique générale du développement d'un pathogène. Toutefois, comme expliqué précédemment, le modèle peut souffrir d'un manque de précision et de performances limitées. Cela peut affecter l'acceptabilité des explications fournies. Cette échelle présente un intérêt comme moyenne ou valeur de référence pour contraster les explications par la suite.

Départementale/communale, locale (groupe)

L'échelle globale peut être considérée comme trop générale et les experts peuvent souhaiter se concentrer sur un nombre limité de parcelles. Il est donc nécessaire de sélectionner des groupes de parcelles pour obtenir des informations plus précises. Plusieurs méthodes de sélection sont possibles : on peut choisir des entités géographiques préexistantes, comme les départements ou les communes. On peut également privilégier une approche plus flexible en permettant à l'utilisateur de délimiter des régions géographiques. Ces sélections peuvent ensuite être utilisées pour examiner et analyser les dynamiques locales.

Individuel

Le dernier échelon géographique est l'échelon individuel, c'est-à-dire une parcelle en particulier. Dans ce cas, on s'intéresse moins aux tendances qu'aux explications au cas par cas. Cela implique que la méthode d'explication doit être capable de fournir une explication individuelle. Dans le cas d'HiPaR et de modèles pattern-based, on portera un intérêt particulier aux patterns ou règles utilisés sur chaque parcelle.

4.2.4 Approches

Les différents degrés de focus décrits précédemment peuvent être résumés en plusieurs approches de visualisation :

L'approche classique consiste à analyser le modèle dans sa globalité. On utilise le modèle et les éléments qui le constituent (conditions, coefficients, support, etc.). D'autres informations peuvent être ajoutées pour contextualiser les explications. Cependant, cette approche est plutôt adaptée aux agronomes qu'aux agriculteurs, compte tenu du manque de spécificité des explications fournies.

L'approche intermédiaire consiste à se concentrer sur le contraste de sous-ensembles de données. On choisira un mode de subdivision ou de sélection qui servira à définir ces sous-ensembles. Les explications fournies seront similaires à l'approche classique. On pourra également s'intéresser à la prévalence des règles (en utilisant notamment le support de chacune). Cela peut servir à contraster des régions entre elles. Dans le même esprit, on peut envisager de comparer des régions au cas général pour observer les divergences qui peuvent apparaître. Par exemple, on peut comparer la moyenne des variables météorologiques aux valeurs locales. Cette approche, encore une fois, semble plus adaptée aux agronomes, mais peut présenter un intérêt pour les agriculteurs.

L'approche par le focus le plus fort se concentre sur les cas individuels. On peut comparer les individus à des sous-régions ou au cas général. Ce degré de focus peut intéresser autant les agronomes que les agriculteurs. Chaque élément pouvant être utilisé comme explication doit être affiché de manière à ce que la compréhension de chacun soit la plus aisée possible. Par ailleurs, il faut faire attention au nombre d'éléments affichés pour éviter de rendre les explications trop complexes. Il est donc nécessaire de sélectionner les éléments utilisés et leur nombre.

4.2.5 Temporalité

Général

Une possibilité est d'autoriser les comparaisons et contrastes entre tous les modèles et toutes les années. Cela permet de comparer les dynamiques d'année en année et de mettre en lumière des différences parmi les variables, notamment météorologiques. Toutefois, cela ajoute un degré de complexité, car cela combine à la fois des différences temporelles et spatiales entre les explications fournies.

Années/saison

L'angle le plus simple pour expliquer les modèles est de se limiter à une année en particulier. Cela a du sens dans la mesure où chaque modèle a été entraîné sur une année en

particulier. La comparaison se fera donc entre des parcelles d'une même année. Dans ce cas, la distinction se fera principalement en termes de position géographique.

4.3 Conception

Une fois les besoins et les axes de développement définis, la phase de conception a pu débuter. Les réunions et les concertations se sont réparties entre les phases de conception et d'implémentation. Les réunions de conception ont été menées en collaboration avec Gonzalo Mendez et ont abouti à un ensemble d'éléments jugés intéressants dans notre cas. Un certain nombre de choix en matière de conception et d'interfaçage ont été faits à ces occasions. Les réunions concernant l'implémentation ont été menées en priorité avec les prestataires ENEO qui nous ont permis de déterminer les choix restants, les fonctionnalités réalisables et les derniers détails de mise au point. Les sujets abordés dans les réunions de conception sont décrits ci-dessous.

04-01-2022

- Choix des options de comparaison et de sélection.
- Discussion sur les contraintes de visualisation liées aux couleurs, à l'ergonomie de l'interface.
- Simplification des options (élimination des options de comparaison inter-années).

25-04-2022

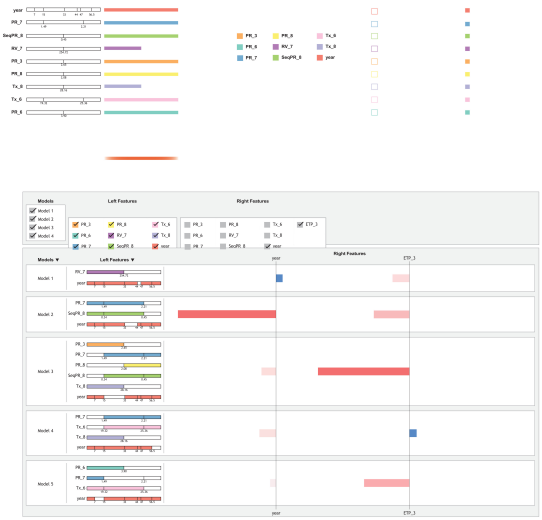


FIGURE 4.1 – Premier prototype

— Discussion sur l'interface.

— Choix :

  ◦ Coefficients ou importance des variables sous forme de barres.

  ◦ Ranking des variables.

  ◦ Représentation des conditions des règles hybrides.

  ◦ questions soulevées sur les variables agrégées.

05-12-2022

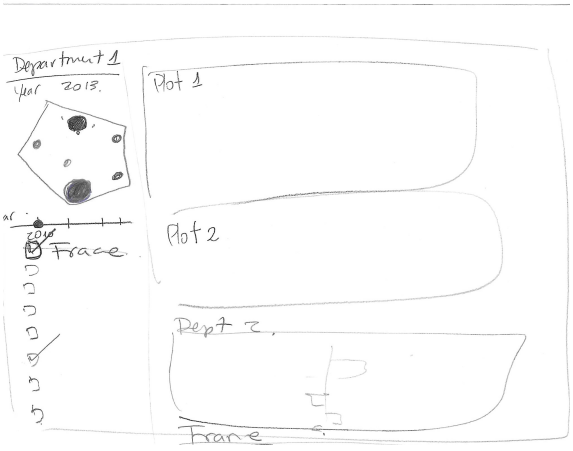


FIGURE 4.2 – Deuxième prototype

— Discussion d'ergonomie.

— Choix des fonctionnalités à privilégier.

— Définition de l'importance des variables.

— Discussion sur les problèmes de couleurs pour la visualisation des données.

25-01-2023

- Discussion des approches de sélection temporelle et géographique.
- Choix de l'approche de visualisation des explications et données aux experts de manière efficace et leur permettre d'évaluer les performances et la viabilité des modèles.
- Choix du mode de sélection des variables d'intérêt et de visualisation des conditions des règles hybrides.

24-02-2023

- Modification du choix du nombre de variable à afficher et de la manière dont sont sélectionnées les variables.
- Choix d'un code couleur pour les différentes règles.
- Simplification de l'interface :
  - Trie des variables en se basant sur des catégories comme les départements.
  - Choix d'afficher les performances des modèles pour chaque sélection.
  - Affichage par défaut de 5 variables.

4.4 Implémentation

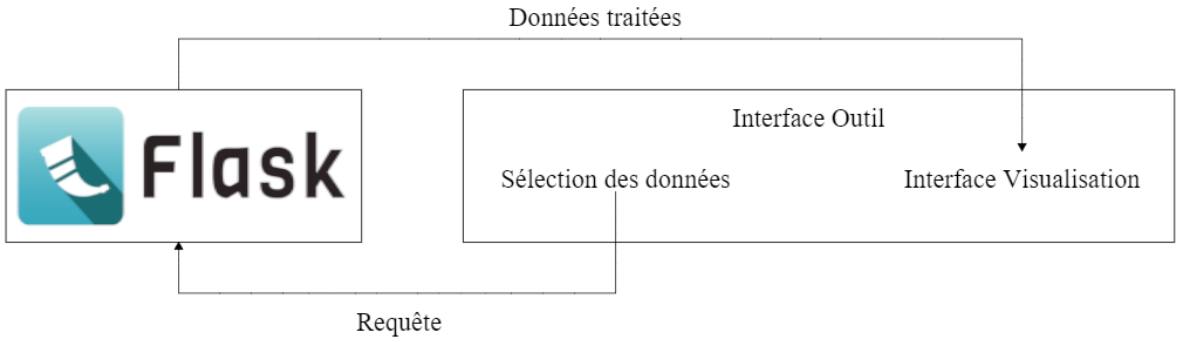


FIGURE 4.3 – Interaction entre l'outil et le serveur Flask

Ces interactions ont mené à la mise au point de l'outil de visualisation, qui repose sur une API permettant d'extraire les informations du modèle et sur une interface utilisateur web.

4.4.1 API d'extraction des informations du modèle

L'API du modèle est basée sur un serveur Flask mis en place pour cette occasion. Flask [76] est un framework léger permettant de créer des applications web en Python. Il a été conçu pour être simple et flexible, permettant ainsi aux développeurs de créer des

applications web de manière rapide et efficace.

L'outil effectue des requêtes sur l'API en fonction des sélections de l'utilisateur et met à jour les informations affichées de manière dynamique. Compte tenu du caractère confidentiel des jeux de données, il a été choisi d'anonymiser autant que possible les informations mises à disposition pour l'outil. De fait, aucune donnée brute n'est fournie à l'outil : les positions géographiques des parcelles sont modifiées en perturbant légèrement les latitudes et longitudes associées à chacune d'entre elles. Cette méthode permet d'anonymiser partiellement les parcelles sans pour autant perdre le potentiel d'explication que l'on peut tirer des modèles. Cela permet d'éviter de divulguer des informations concernant la vulnérabilité des parcelles. Pour les variables prédictives, comme les variables météorologiques, dans le cas où l'incidence est la variable à prédire, aucune valeur brute n'est fournie telle qu'elle : l'impact des variables est précalculé en amont et seules les valeurs calculées sont mises à disposition du serveur Flask. Ces valeurs correspondent au produit du coefficient linéaire du modèle linéaire étudié et de la valeur de la variable.

Étant donné une parcelle x, une règle $p \Rightarrow y = f_p(A'_{num})$ se déclenche si la parcelle évalue "true" pour la condition p de la règle (le modèle général se déclenche dans tous les cas). C'est à partir de cet ensemble de règles déclenchées qu'on calcule les estimations d'importance. Pour trouver le coefficient d'attribution d'une variable $a \in A'_{num}$ pour une règle, l'outil renvoie $|x_a \beta_a|$ où β_a est le coefficient linéaire associé à l'attribut a dans la règle.

Le serveur fournit également une estimation de la performance du modèle sous la forme du score R^2 pour offrir à l'utilisateur une mesure de fiabilité de l'outil.

4.4.2 Interface utilisateur

L'outil Anemone [77] utilise l'architecture vue.js. Vue.js est un framework JavaScript utilisé pour construire des interfaces utilisateur interactives et réactives. Il se concentre principalement sur la création d'applications à page unique (SPA), où une seule page est chargée et où les mises à jour de contenu sont effectuées dynamiquement, sans rechargement complet de la page, dans un souci de réactivité. Vue.js utilise un système de liaison bidirectionnelle qui permet de maintenir automatiquement la synchronisation entre les éléments de la page. Cela signifie que toute modification dans le modèle se reflète automatiquement dans l'interface utilisateur et vice versa. Des librairies supplémentaires permettent d'utiliser une mise en forme spécifique des données. On peut citer parmi elles Openlayers et Chart.js, qui gèrent les aspects géographiques et graphiques respectivement.

Cet outil s'appuie sur ces technologies pour visualiser les données extraites des modèles

sous une forme compréhensible et claire. L'interface a été conçue de manière à ce que son utilisation soit la plus intuitive et facile possible. Elle se présente ainsi sous la forme d'une interface divisée en plusieurs fenêtres. Une carte permet de visualiser et sélectionner les parcelles en les situant à travers la France. Une autre fenêtre permet de sélectionner le nombre de variables à contraster et les modèles à observer. Les informations affichées par l'outil sont l'impact, ou l'importance, d'une variable sur le résultat final du modèle, ainsi que les performances de celui-ci sur la sélection de parcelles étudiées.

Nous allons maintenant examiner l'outil, en détailler les composants et expliquer son utilisation.

Il se présente sous la forme d'une interface divisée en plusieurs parties.

Choix du cas d'usage La partie gauche de l'interface (sections (a), (b) et (c)) correspond à la section de sélection des paramètres géographiques et temporels. La section (a) est un curseur permettant de sélectionner la maladie qui sera étudiée. Elle permet donc de préciser le jeu de données qui sera sollicité.

Choix de l'année La section (b) permet de sélectionner l'année. Ce choix aura une incidence sur la section (c), car les parcelles ne sont pas toutes observées sur toutes les années.

Choix des parcelles Par défaut, les explications fournies concernent l'ensemble des parcelles. La sélection des parcelles à examiner se fait sur la partie inférieure de la carte de France. En utilisant le bouton "Sélectionner des parcelles" de la section (c), l'utilisateur a la possibilité de limiter les explications à un sous-ensemble de parcelles donné. Cette action envoie une nouvelle requête au serveur, qui fournira les informations nécessaires.

Choix des modèles Une fois la requête traitée, les informations sont mises à jour dans les sections (d), (e), (f) et (g). La section (d) nous indique en premier lieu les modèles qui ont été utilisés sur les parcelles sélectionnées pendant l'année demandée. L'utilisateur peut également choisir de s'intéresser non pas au modèle général (comme proposé par défaut), mais aux modèles locaux. La section (e) fournit le coefficient R^2 du modèle sélectionné sur les parcelles concernées. La section (f) affiche le nombre de parcelles sur lesquelles le modèle a été appliqué.

Visualisation La section (g) nous permet de visualiser l'impact des différentes variables sur les prédictions fournies par le modèle. Celles-ci sont représentées sous forme de barres, ce qui permet d'observer graphiquement l'importance accordée à chaque variable.

4.5 Cas d'usage

Admettons qu'un utilisateur souhaite déterminer quels ont été les facteurs les plus impactants sur le développement de la cercosporiose de la betterave sucrière en 2009 selon les modèles. Il doit d'abord sélectionner la culture (ici, la betterave) puis l'année (2009). Les informations fournies par défaut sont celles du modèle général. Après sélection, on peut voir que le R^2 du modèle général sur l'ensemble des parcelles est de 0,2 sur les deux parcelles sur lesquelles il a été utilisé.

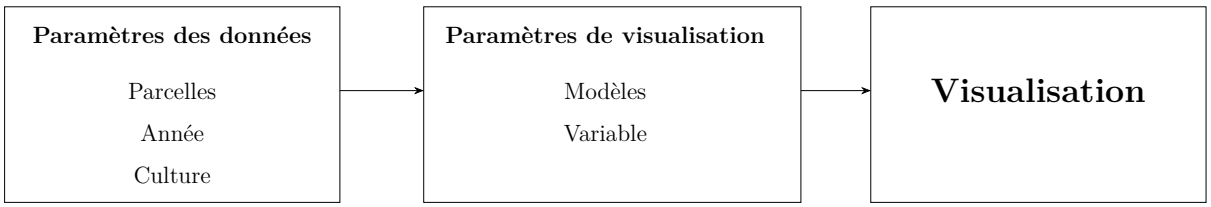
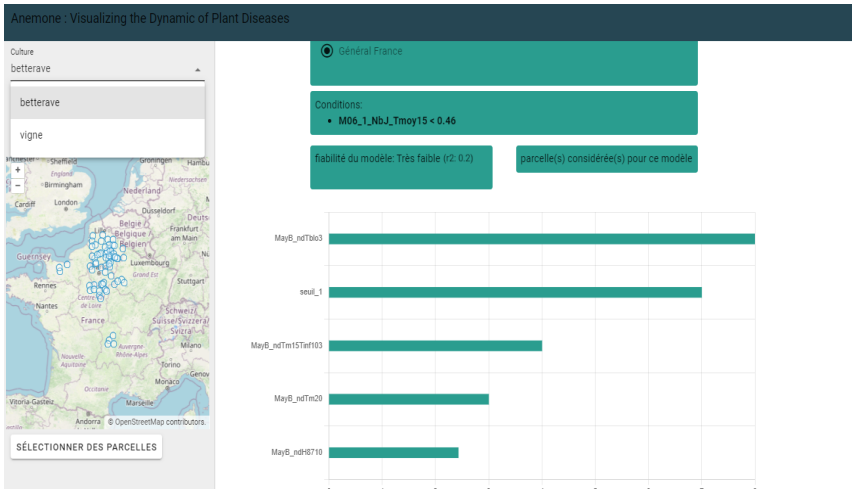


FIGURE 4.4 – Etapes d'utilisation de l'outil



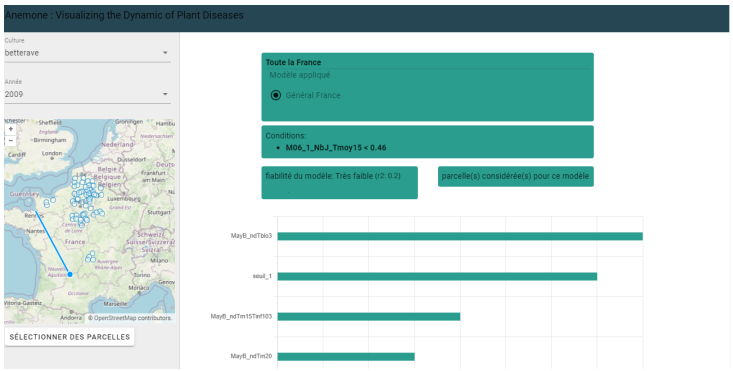
(a) Sélection de la culture

(b) Sélection de l'année

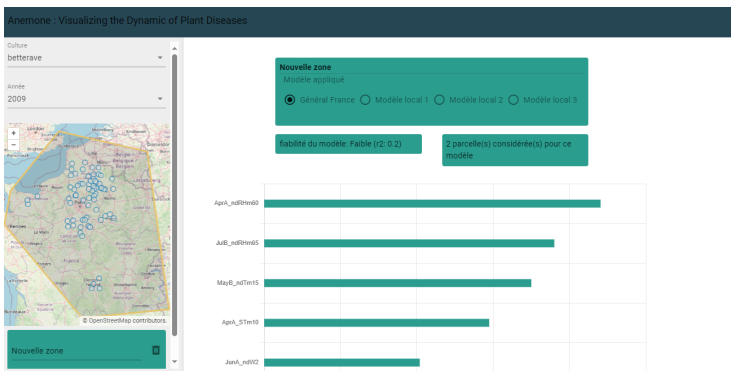
La première étape consiste à déterminer et à définir les données qui nous intéressent : Pour ce faire, on choisira un cas d'étude et une année. Dans notre cas, nous examinerons la cercosporiose de la betterave sucrière. On utilise le curseur situé dans la partie supérieure gauche de l'interface pour sélectionner ce jeu de données.

Le deuxième paramètre à définir est l'année que l'on souhaite examiner spécifiquement. Dans ce cas d'usage, on se concentrera sur l'année 2009.

FIGURE 4.5 – Sélection des paramètres



(a) Sélection des parcelles partie 1

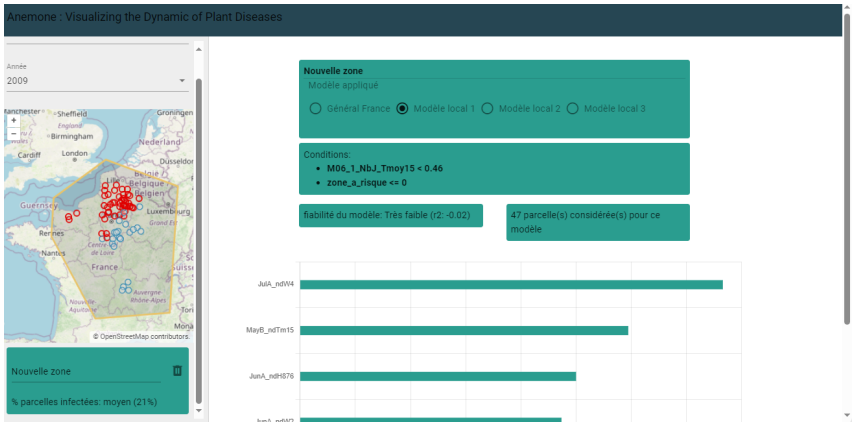


(b) Sélection des parcelles partie 2

Maintenant que ces paramètres ont été définis, il faut choisir les parcelles que l'on souhaite observer. Pour cela, on utilise l'outil de sélection de parcelles situé dans le panneau de gauche de l'interface. La sélection s'effectue à l'aide du bouton « Sélectionner des parcelles », puis en traçant les faces d'un polygone sur la carte.

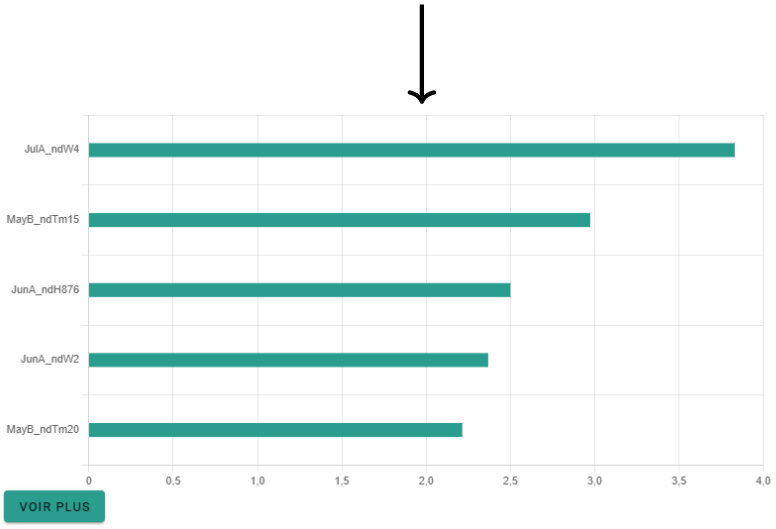
Une fois que les parcelles d'intérêt sont sélectionnées, on ferme le polygone pour terminer de délimiter l'ensemble à étudier. Dans notre cas, on choisit une sélection qui permet d'englober toutes les parcelles : cela nous permettra d'observer l'influence des variables météorologiques selon les règles hybrides apprises. La variable la plus impactante a été AprA-ndRHm60, c'est-à-dire le nombre de jours pendant la première quinzaine d'avril où l'humidité moyenne relative a été supérieure à 60 %, comme on peut le voir sur la Figure 4.5b

FIGURE 4.6 – Sélection des parcelles



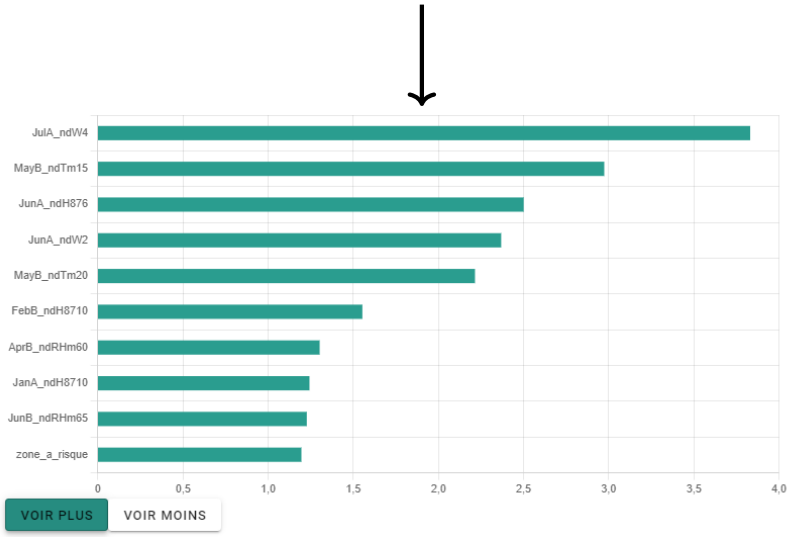
(a) Mise en évidence des parcelles pour la règle sélectionnée

Une fois la sélection effectuée, les modèles présents sur celle-ci sont affichés. On sélectionne l'une de ces règles, la règle 1, pour l'examiner plus en détail.



(b) Affichage par défaut de l'importance des variables

L'impact des variables est affiché dans l'encart inférieur. On souhaite augmenter le nombre de variables examinées. On utilise donc le bouton "Voir plus"



(c) Affichage étendu de l'importance des variables

Par défaut, seules 5 variables sont visibles. Le bouton « Voir plus » permet d'afficher davantage de variables.

FIGURE 4.7 – Sélection des modèles et visualisation

La Figure 4.7a montre à quoi l'interface ressemble une fois que tous les paramètres ont

été définis. La sélection permet à l'outil de récupérer les données spécifiques aux parcelles depuis le serveur Flask. Celui-ci va désormais pouvoir donner accès aux modèles agissant sur les parcelles sélectionnées. Ceux-ci sont visibles et sélectionnables dans l'encart situé dans la partie haute centrale de l'interface. L'affichage par défaut est celui du modèle général. On sélectionne maintenant un modèle local, qui sera le premier disponible.

La Figure 4.7b donne un ensemble d'informations concernant la règle sélectionnée. L'encart situé en dessous de la sélection des modèles locaux contient les conditions de la règle hybride correspondant au modèle local choisi. Celui situé encore plus à gauche nous donne la fiabilité du modèle, évaluée par le coefficient R2 du modèle sur les parcelles concernées. L'encart à sa droite nous informe du nombre de parcelles sur lesquelles le modèle sélectionné agit. Ces parcelles sont également visibles de manière plus explicite dans la partie gauche de l'interface : sur la carte utilisée pour sélectionner les parcelles, les points rouges correspondent aux parcelles qui respectent les conditions associées au modèle étudié. Dans la partie inférieure, on obtient les valeurs d'importance des variables telles que définies précédemment. On peut voir que la variable la plus impactante d'après le modèle local sélectionné est JulA-ndW4, c'est-à-dire le nombre de jours de la première quinzaine de juillet ou la vitesse moyenne du vent était supérieure ou égale à 4 $m.s^{-1}$, ce qui diffère du modèle général. Ceci nous permet de conclure que, d'après les conditions associées à la règle hybride, des températures moyennes et un facteur de risque bas rendent la cercosporiose plus sensible au vent sur les parcelles ayant subi ce genre de conditions météorologiques.

La Figure 4.7c montre que par défaut, 5 variables sont visualisables. Dans notre cas, on peut voir que les variables les plus influentes sont le vent dans la première quinzaine de juillet, les températures dans la deuxième moitié de mai, l'humidité et le vent dans la première quinzaine de juin, ainsi que les températures dans la deuxième moitié de mai (en ordre décroissant d'importance). Le bouton "Voir plus" situé en dessous permet d'afficher davantage de variables, toujours triées par ordre d'importance.

Cela nous permet de visualiser d'autres variables, telles que l'humidité pendant la deuxième quinzaine de février ou l'humidité dans la deuxième moitié d'avril.

4.6 Conclusion

Les échanges avec les experts ont permis de déterminer les moyens les plus aptes à répondre à leurs besoins, comme la possibilité de sélectionner des ensembles de parcelles précises. L'outil répond à ces besoins et permet d'analyser les facteurs d'influence sur le développement de la cercosporiose et du mildiou. Il présente toutefois certaines limites et ne permet pas de comparer différents cas ensemble.

Plusieurs conclusions peuvent être tirées de notre démarche :

— Les explications fournies doivent obligatoirement être contextualisées de manière adaptée à la protection des cultures. Le contexte spatial et géographique, ainsi que le contexte temporel et annuel, est particulièrement parlant pour les experts, qui possèdent des connaissances et une expérience à même d'être comparées, contrastées ou mises en opposition avec les explications fournies.
— L'explicabilité des modèles ayant un impact sur la confiance accordée aux modèles, il est important de prendre en compte ce facteur lorsque le choix du mode de représentation des explications doit être fait. Le public ciblé n'étant pas nécessairement expert en machine learning, il nous revient de faire les choix adaptés pour permettre une compréhension rapide des informations fournies. La fiabilité des modèles est quant à elle à fournir à travers des indicateurs comme le R^2 ou la $RMSE$, car elle participe également à ce que les explications fournies soient acceptées.

L'outil doit encore être validé par les utilisateurs, ce qui demanderait une étude quantitative et/ou qualitative faisant intervenir un nombre significatif d'agronomes. Une telle étude permettrait d'évaluer la pertinence de l'interface actuelle et d'obtenir des retours sur d'éventuelles améliorations ou ajouts pertinents. On peut imaginer que les experts seraient intéressés par des variables particulières et souhaiteraient être en mesure de visualiser la distribution de celles-ci sur une année et une région précise, voire même d'être capables de contraster ces distributions entre plusieurs régions ou années. Il serait également intéressant d'adapter d'autres méthodes d'interprétation au contexte de la protection des cultures, ce qui nous permettrait de comparer la pertinence de chaque méthode et l'intérêt de l'approche privilégiée par notre outil.

Chapitre 5

Discussion

Notre discussion se construira sur trois axes : le compromis performance/complexité abordé précédemment, les possibles implications de la complexité en lien avec l'interprétabilité des modèles et les informations agronomiques extraites des modèles obtenus.

Compromis de complexité. Nos résultats sont en accord avec d'autres travaux sur la complexité des modèles [41, 40], c'est-à-dire que la tendance générale observée est que les modèles complexes surpassent en termes de performance de prédiction les modèles plus simples dans la majorité des cas. Il est toutefois important de noter que cette tendance est observée dans les cas où les modèles ont été paramétrés et entraînés correctement. Un modèle complexe présentera un risque de surentraînement si le jeu de données d'entraînement est de taille limitée, en particulier si le nombre de paramètres du modèle est élevé. Qui plus est, il sera parfois préférable de choisir un modèle de complexité moindre dans les cas où les données sont facilement modélisables par ce type de modèle. En effet, si les données adhèrent aux hypothèses initiales des modèles simples (c'est-à-dire la linéarité), le gain de performance obtenu par des modèles plus complexes sera comparativement moindre et le gain de complexité moins acceptable in fine. Enfin, la stabilité d'un modèle n'est pas assurée et peut faillir dans certains cas, particulièrement lorsque la divergence entre les données d'entraînement et de test est trop grande. Une telle disparité peut être observée dans la prédiction de nos modèles pour les années 2013 et 2015 (Chapitre 2). Ces années ont été caractérisées par des conditions climatiques hors normes, en particulier pour des variables d'importance dont les valeurs sortaient des valeurs observées les années précédentes. Cela s'est traduit par une chute significative de la précision des prédictions, visible dans la perte de R^2 correspondant à ces années (Section 3.1.3) et qui n'est pas corrigée par les modèles plus complexes. Le compromis offert par les modèles de complexité intermédiaire est donc intéressant dans la mesure où leurs performances sont comparables et offrent plus de possibilités d'explication des prédictions. La prise en charge de ces années problématiques est complexe car elle demande de prévoir l'occurrence de conditions météorologiques hors normes qui vont perturber les données. On peut imaginer d'adapter l'approche ensembliste dans laquelle les ensembles sont alignés sur des années particulières, tout en essayant d'inclure des années où ont été observées des conditions

climatiques extrêmes.

Enfin, la prédiction de la date d'apparition des maladies semble être un cas bien plus complexe à modéliser que l'incidence en fin de saison. Les pistes les plus intéressantes pour améliorer les performances des modèles cherchant à prédire ces dates pourraient consister à étudier les différentes manières de définir les dates d'apparition (dans notre cas, la date à laquelle 10 % d'une plante était affectée par la maladie) ou à réunir d'autres indicateurs ou des indicateurs de meilleure qualité pour augmenter les performances des modèles.

Interprétabilité. Comme indiqué précédemment, le lien entre la complexité d'un modèle et son interprétabilité est largement accepté, même s'il reste difficile à quantifier de manière claire. Ce phénomène peut être observé dans nos cas d'usage. Tous les modèles simples utilisés, qu'ils soient linéaires ou pattern-based, nous ont permis d'obtenir des informations sur leur structure. En revanche, nos modèles plus complexes, comme les random forêts et le gradient boosting trees, ont nécessité l'utilisation de méthodes d'explication de modèle telles que les Permutation Feature Importance et les partial dependence plots (PDP). Bien qu'elles soient efficaces, ces techniques ne sont pas parfaites. Les scores d'importance ne donnent pas d'indication sur le sens de l'impact des variables sur la prédiction d'un modèle, mais le quantifient de manière absolue. Les PDP peuvent associer jusqu'à deux variables simultanément, mais ils se basent sur des hypothèses d'indépendance qui se heurtent souvent à la réalité du terrain et à la corrélation des variables, particulièrement des variables climatiques. Comme les PDP représentent les valeurs moyennées de toutes les réponses du modèle en fonction des variables prédictives restantes, il est possible que les PDP considèrent des combinaisons de variables illogiques et impossibles à observer dans la vraie vie. Il est alors nécessaire de garder à l'esprit ce point, en particulier quand on suspecte la présence de corrélations entre les variables prédictives. Ceci étant dit, les PDP obtenus des RFR et GBR nous ont permis de montrer une concordance entre ceux-ci et les conditions obtenues dans HiPaR. On pourra interpréter cela comme une concordance dans la considération des effets de seuils observables dans les jeux de données à disposition. On notera également que les RFR, GBR et HiPaR font tous usage de seuils sous une forme ou une autre, et que le gain de performance observé entre ces modèles et le Lasso semble confirmer l'intérêt de passer outre l'hypothèse de linéarité dans le cas de la prédiction de la dynamique des maladies des plantes se basant sur des variables météorologiques.

Information agronomique. En se basant sur notre étude de la cercosporiose de la betterave sucrière, on observe que l'utilisation de variables météorologiques, associée à une période de temps limitée, permet de modéliser une partie non négligeable de la variation dans l'incidence des maladies des plantes.

L'hiver et les années précédentes semblent constituer un ensemble de conditions initiales : la période hivernale est en effet la période durant laquelle l'inoculum de la cercosporiose repose dans le sol sous la forme de spores. Le printemps constitue la période de développement initial pour les cultures et la cercosporiose. Enfin, l'été englobe à la fois la fin de saison et le moment où les symptômes de la maladie, ainsi que leurs conséquences, sont le plus visibles et les plus répandus.

En règle générale, un été sec semble inhiber l'impact final de la cercosporiose. On peut supposer que cela s'explique par le fait que la plupart des champignons se développent mieux dans des environnements humides. Cette conclusion a été tirée des valeurs d'importance assignées aux facteurs liés au vent et aux températures des mois de juin et juillet. Qui plus est, un hiver sec semble également bloquer le développement de la maladie et sa propagation. On pourra supposer que l'humidité contribue à préserver les spores ou à aider à leur propagation avant la période de sporulation. À l'inverse de ces constatations, un été chaud et humide semble être le facteur le plus aggravant en matière de développement de la cercosporiose. À l'aide des règles hybrides d'HiPaR, on peut définir des relations entre variables plus précises que celles observables avec un Lasso. La deuxième règle, trouvable dans (Table 3.2) , nous indique qu'un mois de juillet relativement doux devrait inciter les agronomes à s'intéresser aux conditions initiales observées en hiver, en particulier le vent et les températures. Qui plus est, un printemps venteux associé à des températures clémentes en juin oriente vers une importance accrue des conditions climatiques observées au printemps, en particulier les facteurs de température et d'humidité, qui semblent être corrélés positivement avec l'incidence finale de la cercosporiose. Dans tous les cas, une date d'apparition rapide est le facteur le plus indicatif de l'incidence en fin de saison, ce qui indique également qu'une surveillance poussée et une détection rapide des premiers symptômes est, en l'état, la meilleure manière de combattre la cercosporiose.

Il nous a été impossible d'obtenir des informations crédibles sur le mildiou de la vigne dans la mesure où les modèles les plus simples n'ont pas réussi à expliquer suffisamment bien la variance de l'incidence (R^2 de 14%). Les résultats de la prédiction de la date d'apparition des symptômes sont encore plus décevants et ne permettent donc pas non plus d'exploiter les modèles. On suppose que le problème principal des modèles réside dans la nature des données disponibles : le jeu de données se base en effet sur des données météorologiques limitées dans le temps et exploitant peu l'historique des parcelles. Cela signifie que la plupart des variables ne comprennent pas d'historique agrégeant leurs valeurs sur une période de quatre semaines. En d'autres termes, ce jeu de données ne semble pas avoir la même précision que le jeu de données de la betterave sucrière en termes de définition temporelle des variables ou de l'historique. Or, les conditions initiales sont un facteur décisif dans la dynamique des maladies des plantes [78] et celles-ci dépendent souvent des saisons passées. Ceci tend encore à confirmer l'importance de la précision des relevés

de données qui seront utilisées pour prédire la dynamique de développement des maladies des plantes. D'un point de vue temporel, l'impact de la précision de ces relevés semble également être un axe de recherche intéressant à explorer. On peut supposer qu'il existe une relation entre les performances d'un modèle et la complexité du jeu d'entraînement, de la même manière que la relation entre la complexité d'un modèle et ses performances. Certains travaux tendent à montrer que des variations extrêmement limitées dans le temps peuvent impacter significativement le développement des maladies des plantes. Il semble donc nécessaire de prendre en compte la granularité des variables météorologiques d'un point de vue temporel dans la phase de prétraitement des données, ou de manière générale en amont de la conception des modèles. On peut également supposer que le compromis entre cette complexité et les performances d'un modèle présente un risque double : un plan trop large ne permettrait pas d'obtenir des données expliquant la variance de la dynamique de développement des maladies des plantes (comme dans le cas du mildiou de la vigne). Inversement, un focus trop précis pourrait mener les modèles vers un sur-entraînement et donc des performances amoindries. Notre hypothèse est qu'encore une fois, un compromis ou une optimisation des jeux de données est préférable. Cette complexité des données impacte directement le problème du compromis entre la complexité d'un modèle et ses performances : de manière générale, un modèle se basant sur un jeu de données plus volumineux (en termes de nombre de variables) sera plus complexe (à paramètres égaux). On suppose donc qu'il existe un compromis désirable entre la complexité des jeux de données, des modèles et de leurs performances. Liée à cette question se pose la problématique de l'ajout ou non de nouveaux indicateurs. Cet ajout devra se faire sur la base à la fois des connaissances pré-existantes des agronomes, puis en se fiant à l'influence que ces indicateurs auront sur les nouveaux modèles. Enfin, on pourra s'intéresser à l'intérêt de dupliquer l'approche utilisée pour la cercosporiose de la betterave et de diviser les indicateurs sous la forme de séries temporelles (ou équivalent) ainsi qu'à l'influence de cette approche sur l'explicabilité et les performances des modèles.

Chapitre 6

Conclusion

Cette thèse s'inscrit dans le cadre du plan Ecophyto, et, de manière générale, de la recherche en apprentissage automatique interprétable appliqué à la protection des cultures. Ce champ de recherche est crucial pour la mise en place d'approches d'agriculture raisonnée permettant de réduire l'utilisation de produits phytosanitaires sans entraîner une baisse de rendement significative. La thèse a permis de montrer l'intérêt de prendre en compte la problématique posée par la complexité des modèles appliqués à la prédiction de la dynamique des maladies des plantes. On admet généralement que les modèles complexes, comme les réseaux de neurones ou le gradient boosting, sont plus efficaces que les modèles de régression linéaire. Cependant, cette augmentation de complexité se fait au prix d'une perte d'interprétabilité. Cette interprétabilité est nécessaire si l'on cherche à extraire des informations des modèles ou si la transparence est importante pour des raisons d'acceptabilité ou légales. Les explications post-hoc peuvent aider à extraire ces informations des modèles boîte noire, mais elles ne sont pas les seules solutions à notre disposition : nos modèles de complexité intermédiaire basés sur de la régression par pattern se montrent capables d'entrer en compétition avec des modèles plus complexes tout en conservant un meilleur degré d'interprétabilité. Nos modèles atteignent des performances satisfaisantes par rapport aux exigences en protection des cultures dans un certain nombre de cas, bien qu'une instabilité puisse être observée. De plus, les résultats obtenus grâce aux méthodes d'explication post-hoc, comme les Partial Dependence Plots, montrent que certains effets de seuils modélisés par les modèles boîte noire sont également pris en compte dans les modèles pattern-based. Ces modèles fournissent des informations confirmées par les connaissances agronomiques préexistantes et de nouvelles informations sur des facteurs jusqu'à présent considérés comme peu impactants. Cela met en lumière l'intérêt des modèles de régression pattern-based pour la protection des cultures. Enfin, notre étude confirme que les variations des conditions météorologiques d'une année à l'autre rendent la prédiction de la dynamique des maladies des plantes difficile, et que cette tâche ne peut être accomplie qu'à l'aide de recherches complémentaires décrites ci-après et de données d'aussi bonne qualité que possible.

À l'avenir, il sera possible d'envisager d'étudier l'influence de la granularité des données

spatiales et temporelles sur les performances des modèles. On pourrait également envisager d'employer des méthodes de détection d'anomalies dont le but serait de stabiliser les performances des modèles en détectant les années dont les conditions météorologiques sortent de la norme. Un autre axe de recherche prometteur pourrait consister à appliquer des techniques d'apprentissage de représentation dans l'objectif de trouver de nouveaux indicateurs météorologiques permettant d'améliorer la précision de nos modèles. On pourra également envisager d'utiliser des modèles non supervisés, compte tenu de la nature exploratoire de l'approche. Ces approches plus souples pourraient également se révéler utiles dans le cadre du changement climatique, celui-ci aggravant les variations des conditions climatiques et météorologiques.

BIBLIOGRAPHIE

[1] Ronald J Howard. Cultural control of plant diseases : a historical perspective. *Canadian Journal of Plant Pathology*, 18(2) :145–150, 1996.

[2] Ian Heap. Global perspective of herbicide-resistant weeds. *Pest Management Science*, 70(9) :1306–1315, 2014. ISSN 15264998. doi : 10.1002/ps.3696.

[3] Thomas van Klompenburg, Ayalew Kassahun, and Cagatay Catal. Crop yield prediction using machine learning : A systematic literature review. *Computers and Electronics in Agriculture*, 177(January) :105709, 2020. ISSN 01681699. doi : 10.1016/j.compag.2020.105709. URL `https://doi.org/10.1016/j.compag.2020.105709`.

[4] FAO. Fao's plant production and protection division, 2022.

[5] James S Donnelly Jr. *The great Irish potato famine*. The History Press, 2002.

[6] JC Zadoks. On the conceptual basis of crop loss assessment : the threshold theory. *Annual Review of Phytopathology*, 23(1) :455–473, 1985.

[7] Biology, epidemiology. `https://ephytia.inra.fr/en/C/6967/Grapevine-Biology-epidemiology`.

[8] FAO. Pesticides use, pesticides trade and pesticides indicators global, regional and country trends, 1990–2020, 2022.

[9] Laurence Denaix, Laetitia Anatole-Monnier, and Denis Thiery. Effet de l'utilisation répétée de bouillie bordelaise sur la contamination des sols, la biodisponibilité du cuivre et son accumulation dans la vigne. In *46. Colloque du Groupe Français des Pesticides*, page np, Bordeaux, France, May 2016. URL `https://hal.science/hal-01869886`.

[10] Isra Mahmood, Sameen Ruqia Imadi, Kanwal Shazadi, Alvina Gul, and Khalid Rehman Hakeem. Effects of pesticides on environment. *Plant, soil and microbes : volume 1 : implications in crop science*, pages 253–269, 2016.

[11] George P Georghiou. *Pest resistance to pesticides*. Springer Science & Business Media, 2012.

[12] Integrated Pest Management. https://food.ec.europa.eu/plants/pesticides/sustainable-use-pesticides/integrated-pest-management-ipm_en.

[13] Jean-Philippe Deguine, Caroline Gloanec, Philippe Laurent, Alain Ratnadass, and Jean-Noël Aubertot. *Protection agroécologique des cultures*. Editions Quae, 2016.

[14] Ruth E. Baker, Jose Maria Peña, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community ? *Biology Letters*, 14(5) :1–4, 2018. ISSN 1744957X. doi : 10.1098/rsbl.2017.0660.

[15] Elisa González-Domínguez, Tito Caffi, Vittorio Rossi, Irene Salotti, and Giorgia Fedele. Plant disease models and forecasting : changes in principles and applications over the last 50 years. *Phytopathology®*, 113(4) :678–693, 2023.

[16] URL https://inoki.ctifl.fr/XhtmlContent/Modele/Bibliographie/Modele_27/bulletin%20semences%20de%20la%20fnams%20n202%20.pdf.

[17] David Gouache, Arnaud Bensadoun, François Brun, Christian Pagé, David Makowski, and Daniel Wallach. Modelling climate change impact on septoria tritici blotch (stb) in france : Accounting for climate model and disease model uncertainty. *Agricultural and forest meteorology*, 170 :242–252, 2013.

[18] F. K. van Evert, S. Fountas, D. Jakovetic, V. Crnojevic, I. Travlos, and C. Kempenaar. Big Data for weed control and crop protection. *Weed Research*, 57(4) :218–233, 2017. ISSN 13653180. doi : 10.1111/wre.12255.

[19] Estimation du risque lié aux charançons du bourgeon terminal. https://www.terresinovia.fr/-/charancon-bourgeon-colza.

[20] Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Interpretable Machine Learning. *Queue*, 19(6) :28–56, 2021. ISSN 15427749. doi : 10.1145/3511299.

[21] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations : An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pages 80–89, 2018. doi : 10.1109/DSAA.2018.00018.

[22] Adrien Bibal and Anthony Cleve. Interpretability and Explainability in Machine Learning with Application to Nonlinear Dimensionality Reduction.

[23] Nadia Burkart and Marco F. Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70 :245–317, 2021. ISSN 10769757. doi : 10.1613/JAIR.1.12228.

[24] Efstathios D. Gennatas, Jerome H. Friedman, Lyle H. Ungar, Romain Pirracchio, Eric Eaton, Lara G. Reichmann, Yannet Interian, José Marcio Luna, Charles B. Simone, Andrew Auerbach, Elier Delgado, Mark J. van der Laan, Timothy D. Solberg, and Gilmer Valdes. Expert-augmented machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 117(9) :4571–4577, 2020. ISSN 10916490. doi : 10.1073/pnas.1906831117.

[25] Masahiro Ryo. Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture*, 6 :257–265, 2022. ISSN 25897217. doi : 10.1016/j.aiia.2022.11.003. URL `https://doi.org/10.1016/j.aiia.2022.11.003`.

[26] Jay Ram Lamichhane. Pesticide use and risk reduction in European farming systems with IPM : An introduction to the special issue. *Crop Protection*, 97 :1–6, 2017. ISSN 02612194. doi : 10.1016/j.cropro.2017.01.017. URL `http://dx.doi.org/10.1016/j.cropro.2017.01.017`.

[27] Olivier Gauriau, Luis Galárraga, François Brun, Alexandre Termier, Loïc Davadan, and François Joudelat. Comparing machine-learning models of different levels of complexity for crop protection : A look into the complexity-accuracy tradeoff. *Smart Agricultural Technology*, 7 :100380, 2024.

[28] Ryan H.L. Ip, Li Minn Ang, Kah Phooi Seng, J. C. Broster, and J. E. Pratley. Big data and machine learning for crop protection. *Computers and Electronics in Agriculture*, 151(November 2017) :376–383, 2018. ISSN 01681699. doi : 10.1016/j.compag.2018.06.008. URL `https://doi.org/10.1016/j.compag.2018.06.008`.

[29] Gianni Fenu and Francesca Maridina Malloci. Review forecasting plant and crop disease : An explorative study on current algorithms. *Big Data and Cognitive Computing*, 5(1) :1–24, 2021. ISSN 25042289. doi : 10.3390/bdcc5010002.

[30] Konstantinos G. Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. Machine learning in agriculture : A review. *Sensors (Switzerland)*, 18(8) :1–29, 2018. ISSN 14248220. doi : 10.3390/s18082674.

[31] Rohit Sharma, Sachin S. Kamble, Angappa Gunasekaran, Vikas Kumar, and Anil Kumar. A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Computers and Operations Research*, 119 :104926, 2020. ISSN 03050548. doi : 10.1016/j.cor.2020.104926. URL `https://doi.org/10.1016/j.cor.2020.104926`.

[32] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The*

elements of statistical learning : data mining, inference, and prediction, volume 2. Springer, 2009.

[33] Sotiris B Kotsiantis, Dimitris Kanellopoulos, and Panagiotis E Pintelas. Data preprocessing for supervised leaning. *International journal of computer science*, 1(2) : 111–117, 2006.

[34] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[35] Pádraig Cunningham. Dimension reduction. In *Machine learning techniques for multimedia : Case studies on organization and retrieval*, pages 91–112. Springer, 2008.

[36] S. Velliangiri, S. Alagumuthukrishnan, and S. Iwin Thankumar Joseph. A Review of Dimensionality Reduction Techniques for Efficient Computation. *Procedia Computer Science*, 165 :104–111, 2019. ISSN 18770509. doi : 10.1016/j.procs.2020.01.079. URL `https://doi.org/10.1016/j.procs.2020.01.079`.

[37] Stefan Kramer. Structural regression trees. In *AAAI/IAAI, Vol. 1*, pages 812–819, 1996.

[38] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. Cart. *Classification and regression trees*, 1984.

[39] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1 :81–106, 1986.

[40] Luis Galárraga, Olivier Pelgrin, and Alexandre Termier. *HiPaR : Hierarchical Pattern-Aided Regression*, volume 12712 LNAI. Association for Computing Machinery, 2021. ISBN 9783030757618. doi : 10.1007/978-3-030-75762-5_26.

[41] Guozhu Dong and Vahid Taslimitehrani. Pattern-aided regression modeling and prediction model analysis. *2016 IEEE 32nd International Conference on Data Engineering, ICDE 2016*, pages 1508–1509, 2016. doi : 10.1109/ICDE.2016.7498398.

[42] Victor E McZgee and Willard T Carleton. Piecewise regression. *Journal of the American Statistical Association*, 65(331) :1109–1124, 1970.

[43] Leo Breiman. Random forests. *Machine learning*, 45 :5–32, 2001.

[44] Yong Wang and Ian H. Witten. Inducing Model Trees for Continuous Classes. In *ECML Poster Papers*, 1997.

[45] Gareth Edwards-Jones. Knowledge-based systems for crop protection : theory and practice. *Crop Protection*, 12(8) :565–578, 1993. ISSN 02612194. doi : 10.1016/0261-2194(93)90119-4.

[46] Dhivya Elavarasan, Durai Raj Vincent, Vishal Sharma, Albert Y. Zomaya, and Kathiravan Srinivasan. Forecasting yield by integrating agrarian factors and machine learning models : A survey. *Computers and Electronics in Agriculture*, 155 (October) :257–282, 2018. ISSN 01681699. doi : 10.1016/j.compag.2018.10.024. URL `https://doi.org/10.1016/j.compag.2018.10.024`.

[47] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting algorithms as gradient descent. *Advances in neural information processing systems*, 12, 1999.

[48] Frederick Mosteller and John W Tukey. Data analysis, including statistics. *Handbook of social psychology*, 2 :80–203, 1968.

[49] Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim. On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, 109(4) :853–898, 2020. ISSN 15730565. doi : 10.1007/s10994-019-05856-5.

[50] Tim Miller. Explanation in artificial intelligence : Insights from the social sciences. *Artificial intelligence*, 267 :1–38, 2019.

[51] Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), September 2008. ISSN 1932-6157. doi : 10.1214/07-aoas148. URL `http://dx.doi.org/10.1214/07-AOAS148`.

[52] Mark William Craven. *Extracting comprehensible models from trained neural networks*. The University of Wisconsin-Madison, 1996.

[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You ?" Explaining the Predictions of Any Classifier. *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Proceedings of the Demonstrations Session*, pages 97–101, 2016. doi : 10.18653/v1/n16-3020.

[54] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[55] Ramprasaath R Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam : Why did you say that ? *arXiv preprint arXiv :1611.07450*, 2016.

[56] Partial dependence and individual conditional expectation plots. `https://scikit-learn.org/stable/modules/partial_dependence.html`.

[57] Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions : Visual inspection of black-box machine learning models. *Conference on Human Factors in Computing Systems - Proceedings*, pages 5686–5697, 2016. doi : 10.1145/2858036.2858529.

[58] Daniel W. Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society. Series B : Statistical Methodology*, 82(4) :1059–1086, 2020. ISSN 14679868. doi : 10.1111/rssb.12377.

[59] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking Inside the Black Box : Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1) :44–65, 2015. ISSN 15372715. doi : 10.1080/10618600.2014.907095.

[60] Lime package screenshots. `https://github.com/marcotcr/lime`.

[61] Scikit permutation feature importance. `https://scikit-learn.org/1.5/modules/permutation_importance.html`.

[62] Stephen R. Midway. Principles of Effective Data Visualization. *Patterns*, 1(9) : 100141, 2020. ISSN 26663899. doi : 10.1016/j.patter.2020.100141. URL `https://doi.org/10.1016/j.patter.2020.100141`.

[63] SHAP. `https://github.com/shap/shap`.

[64] Color Brewer. `https://colorbrewer2.org/#type=sequential&scheme=YlOrRd&n=3`.

[65] ITP. Bilan d'activité, 2020.

[66] V Rossi. Cercospora leaf spot infection and resistance in sugarbeet. 2000.

[67] GD Franc. Ecology and epidemiology of cercospora beticola. *Cercospora leaf spot of sugar beet and related species*, pages 7–19, 2010.

[68] George N Skaracis, Ourania I Pavli, and Enrico Biancardi. Cercospora leaf spot disease of sugar beet. *Sugar Tech*, 12 :220–228, 2010.

[69] J Khan, LE del Río, R Nelson, V Rivera-Varas, GA Secor, and MFR Khan. Survival, dispersal, and primary infection site for cercospora beticola in sugar beet. *Plant Disease*, 92(5) :741–745, 2008.

[70] Annamaria Vercesi, Silvia L Toffolatti, Graziano Zocchi, Raffaella Guglielmann, and Liliana Ironi. A new approach to modelling the dynamics of oospore germination in plasmopara viticola. *European Journal of Plant Pathology*, 128 :113–126, 2010.

[71] Pere Quintana-Segui, Paul Le Moigne, Yves Durand, Eric Martin, Florence Habets, Martine Baillon, Claire Canellas, Laurent Franchisteguy, and Sophie Morel. Analysis of near-surface atmospheric variables : Validation of the safran analysis over france. *Journal of applied meteorology and climatology*, 47(1) :92–107, 2008.

[72] Météo France. Bilan climatique de l'année 2015. `https://meteofrance.fr/sites/meteofrance.fr/files/files/editorial/Bilan_ann%C3%A9e-2015-definitif_defma_0.pdf`, 2015.

[73] Leonel Deleon, Michael J. Brewer, Isaac L. Esquivel, and Jonda Halcomb. Use of a geographic information system to produce pest monitoring maps for south Texas cotton and sorghum land managers. *Crop Protection*, 101 :50–57, 2017. ISSN 02612194. doi : 10.1016/j.cropro.2017.07.016. URL `http://dx.doi.org/10.1016/j.cropro.2017.07.016`.

[74] Petr Kubicek, Jiri Kozel, Radim Stampach, and Vojtech Lukas. Prototyping the visualization of geographic and sensor data for agriculture. *Computers and Electronics in Agriculture*, 97 :83–91, 2013. ISSN 01681699. doi : 10.1016/j.compag.2013.07.007. URL `http://dx.doi.org/10.1016/j.compag.2013.07.007`.

[75] Anabelle Laurent, Peter Kyveryga, David Makowski, and Fernando Miguez. A framework for visualization and analysis of agronomic field trials from on-farm research networks. *Agronomy Journal*, 111(6) :2712–2723, 2019. ISSN 14350645. doi : 10.2134/agronj2019.02.0135.

[76] Flask. `https://flask.palletsprojects.com`.

[77] Maquette de l'outil Anemone : Visualizing the Dynamic of Plant Diseases. `https://anemone.hdigitag.fr/`.

[78] X. M. Xu and M. S. Ridout. Effects of initial epidemic conditions, sporulation rate, and spore dispersal gradient on the spatio-temporal dynamics of plant disease epidemics. *Phytopathology*, 88(10) :1000–1012, 1998. ISSN 0031949X. doi : 10.1094/PHYTO.1998.88.10.1000.

Résumé

Cette thèse se concentre sur la prédiction de la dynamique des maladies des plantes et les informations qui peuvent être tirées des modèles obtenus. Pour trouver un compromis entre la performance et la complexité du modèle, on a utilisé des modèles de complexité intermédiaire dits pattern-based.

L'objectif était de parvenir à obtenir des modèles de ce type suffisamment performants en se basant sur des données météorologiques (comme les précipitations et l'ensoleillement) et agronomiques. Ces modèles ont été comparés à des modèles couramment utilisés en protection des cultures d'un point de vue du tradeoff entre la complexité et les performances des modèles. Compte tenu du caractère hybride des modèles pattern-based, on a cherché à comparer leurs structures et les informations qu'ils fournissent aux autres modèles. Ceci nous a permis de confirmer que les modèles pattern-based s'approchent des modèles plus complexes (RF, Gradient-Boosting...) tout en restant plus simples à comprendre. Ceci nous permet de supposer que les explications fournies par ces modèles sont pertinentes.

Enfin, ces modèles ont été utilisés dans la mise au point d'un outil de visualisation : Cet outil a été mis au point en collaboration avec des experts agronomes d'instituts techniques pour obtenir un résultat adapté à leurs besoins. Cela a permis d'isoler des principes importants pour eux, comme la notion de contraste des informations fournies. L'outil permet de visualiser les facteurs agronomiques et météorologiques les plus impactants sur un ensemble de parcelles défini.

COLLEGE MATHS, TELECOMS
DOCTORAL INFORMATIQUE, SIGNAL
BRETAGNE SYSTEMES, ELECTRONIQUE

Université de Rennes

Titre : Fouille de règles numériques pour le prédiction de la dynamique des maladies des plantes

Mot clés : Protection des cultures, Régression, Dataviz, Explicabilité

Résumé : Avec l'importance croissante de l'agriculture raisonnée, le développement de méthodes d'aide à la décision appliquées à la protection des cultures fait partie d'un ensemble d'approches visant à atteindre l'objectif de réduire la dépendance vis-à-vis des produits phytosanitaires. Ces outils peuvent se baser sur des modèles de machine-learning, ce qui permet d'automatiser l'entraînement des modèles et donc faciliter l'exploitation des données. Cette thèse étudie l'intérêt des modèles de complexité intermédiaire (pattern-based) comparé à des modèles simples comme la régression linéaire ou les modèles complexes comme le Random Forest. On montre que les performances des modèles intermédiaires sont comparables aux modèles plus complexes, et que les informations extractibles de ces modèles sont pertinentes et plus compréhensibles. Ces modèles pattern-based ont été ensuite utilisés comme base pour un outil de visualisation dont l'objectif est de transmettre des explications aux agronomes pour leur offrir des informations d'intérêt d'un point de vue agronomique.

Title: Numerical rule mining for the prediction of plant disease dynamics.

Keywords: Crop protection, Regression, Dataviz, Explicability

Abstract: With the growing importance of integrated agriculture, the development of decision-support methods applied to crop protection is part of a range of approaches aimed at achieving the objective of reducing the dependency of phytosanitary products. These tools can be based on machine-learning models, making it possible to automatize model training and thus facilitate data exploitation. This thesis studies the interest of pattern-based models compared to simple models such as linear regression or complex models such as Random Forest. It is shown that the performance of mid-complexity models is comparable to that of more complex models, and that the information extracted from these models is relevant and more comprehensible. These pattern-based models were then used as the basis for a visualization tool whose aim is to convey explanations to agronomists, offering them information of interest from an agronomic point of view.