NOM Prénom : WANE Mountaga

RAPPORT DE STAGE
2ème ANNEE

THEME DU STAGE : Machine learning pour l'estimation de productions agricoles

STRUCTURE D'ACCUEIL : Agence Nationale de la Statistique et de la Démographie (ANSD)

Adresse du stage : Rocade Fann Bel-air Cerf-volant. BP 116 Dakar RP -Sénégal

Promotion : 2025

Maître de stage : Moussa DIALLO

Référent pédagogique : Marion GOUSSE

Résumé

Ce rapport présente une étude réalisée au sein de l'ANSD visant à développer des modèles de machine learning pour estimer les productions agricoles au Sénégal. En se basant sur une approche hybride combinant des indices calculés à partir de données satellites, tels que l'indice de végétation par différence normalisée (NDVI), et des relevés météorologiques (pluviométrie et température du sol), l'objectif était de créer des modèles prédictifs robustes capables de fournir des estimations précises des rendements agricoles. Les travaux se sont concentrés sur l'utilisation de réseaux de neurones pour capturer les relations complexes et non linéaires entre les variables. Les résultats obtenus montrent que l'utilisation d'un modèle de réseau de neurones adapté, combiné avec une fonction de perte personnalisée, permet de réaliser des prédictions plus précises, notamment pour des cultures spécifiques comme l'arachide, le manioc, le sorgho ou encore le riz.

Abstract

This report presents a study conducted at the ANSD aimed at developing machine learning models to estimate agricultural production in Senegal. Based on a hybrid approach that combines indices calculated from satellite data, such as the Normalized Difference Vegetation Index (NDVI), and meteorological records (rainfall and soil temperature), the objective was to create robust predictive models capable of providing accurate estimates of agricultural yields. The work focused on the use of neural networks to capture the complex and non-linear relationships between variables. The results obtained show that using an adapted neural network model, combined with a customized loss function, allows for more accurate predictions, particularly for specific crops such as peanuts, cassava, sorghum and rice.

Table des matières

1 Introduction

L'agriculture est un secteur clé pour le développement économique et social du Sénégal, représentant une part significative du produit intérieur brut (PIB) et employant une grande partie de la population. Cependant, ce secteur fait face à de nombreux défis, notamment la variabilité climatique, la dégradation des sols et des pratiques agricoles souvent peu optimisées. Dans ce contexte, le recours aux technologies de l'information et à des méthodes avancées comme le machine learning offre une opportunité unique pour améliorer la gestion des ressources agricoles et optimiser les rendements. Ce stage de deuxième année à l'Agence Nationale de la Statistique et de la Démographie (ANSD), au sein du Bureau de la Recherche et de l'Innovation (BRI) de la Direction de la Méthodologie de la Coordination statistique et de l'Innovation (DMCI), s'est inscrit dans cet effort d'innovation.

L'objectif de ce stage était de développer une méthodologie permettant d'estimer les rendements agricoles à partir de données multisources, combinant des images satellites (Sentinel-2, Landsat-8), des données météorologiques (pluviométrie, température) et des statistiques agricoles. Les données satellitaires ont été utilisées pour extraire des indices de végétation tels que le NDVI, un indicateur clé de la santé et de la densité de la végétation. Ces indices ont ensuite été couplés à des informations météorologiques pour alimenter des modèles de machine learning, notamment des réseaux de neurones. Le travail a consisté à définir un cadre méthodologique pour la collecte et le prétraitement des données, la sélection des modèles appropriés, leur entraînement et leur évaluation, tout en prenant en compte les spécificités locales du contexte sénégalais. Ce rapport détaille les différentes étapes de cette démarche, les défis rencontrés et les solutions mises en œuvre pour parvenir à des estimations de rendement fiables et pertinentes.

2 Environnement de travail

« Créée par la loi n°2004-21 du 21 juillet 2004 portant organisation des activités statistiques du Système statistique national (SSN), l'Agence Nationale de la Statistique et la Démographie (ANSD) est une structure administrative dotée de la personnalité juridique et d'une autonomie de gestion. »

Ce passage est tiré de la page d'accueil du site officiel de l'ANSD. Il s'agit de la structure publique en charge de la production et de la diffusion des statistiques au Sénégal. Elle dépend du Ministère de l'Economie. Nous pourrions la considérer comme l'équivalent de l'INSEE (Institut National de la Statistique et des Etudes Economiques) en France.

Mon stage s'est déroulé au sein de la Direction de la Méthodologie, de la Coordination statistique et de l'Innovation de l'Agence, plus précisément dans le Bureau de la Recherche et de l'Innovation (BRI/DMI/DMCI). Il s'agit d'une Direction récente de l'Agence. En effet, elle a été créée il y a moins d'un an et a pour but de créer un écosystème innovant propice au développement de la recherche et de l'Innovation dans l'Agence. Par ailleurs, il s'agit également de disposer d'une entité capable de collaborer de manière indépendante avec d'autres structures publiques de traitement de données telles que le DAPSA (Direction de l'Analyse, de la Prévision et des Statistiques Agricoles) ou encore l'ISRA (Institut Sénégalais de Recherches Agricoles) dans des projets structurants pour le pays.

Mon maître de stage est Moussa Diallo. Il est DataScientist et chef du BRI. Ce stage s'inscrit dans la naissance de la Direction telle que décrite plus haut. Il convient également de préciser que la création d'un DIL (Data Innovation Lab) est en cours de finalisation. Celui-ci sera rattaché à la Direction. C'est en réalité lui qui sera l'intermédiaire entre l'ANSD et les autres structures.

Image 1 : Siège de l'ANSD, Dakar

3 La problématique et les données

3.1 Enjeux et Difficultés

Le projet sur lequel je travaille correspond à l'un des sujets que le DIL (Data Innovation Lab) de l'ANSD souhaite traiter : il s'agit de l'estimation de productions agricoles. Au Sénégal, bien que la part du secteur primaire, très largement dominé par l'agriculture, dans le PIB ne tourne autour que de 15%, c'est un secteur qui emploie de façon directe ou indirecte plus de 70% de la population.

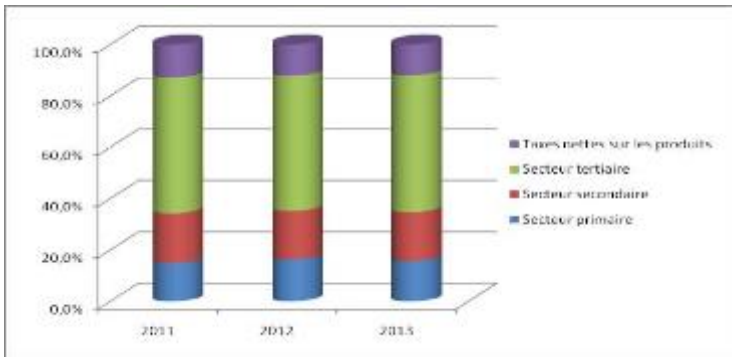

Image 2 : Répartition des différents secteurs de l'économie sénégalaise dans le PIB

L'enjeu est donc assez bien perceptible. Une meilleure maîtrise des dynamiques d'évolution des productions agricoles est primordiale afin de mieux orienter certaines politiques de l'Etat. Il peut par exemple s'agir d'aides pour les agriculteurs en cas de mauvaises récoltes ou encore de fournitures d'équipements agricoles là où des productions à venir sont jugées importantes.

Un premier défi est l'accès aux données. Ainsi, il était initialement prévu que mon stage se déroule en partenariat avec la DAPSA (Direction de l'Analyse, de la Prévision et des Statistiques Agricoles) du Ministère de l'Agriculture. Le sujet concernait la prévision de rendements sur des parcelles agricoles à partir de données satellites. Il a finalement été convenu que je travaille sur des données au niveau départemental, notamment en raisons de lenteurs administratives, étant donné que j'étais en poste pour 2 mois seulement.

Par la suite, nous nous rendrons d'ailleurs compte que les données de la DAPSA ne couvriraient peut-être pas une fenêtre temporelle assez importante pour mener une étude intéressante et de laquelle nous tirerions des résultats fiables sur le long terme.

3.2 Présentation des données

Dans le cadre de cette étude, nous disposons de 3 bases de données différentes. Elles ont toutes été trouvées en OpenSource. (cf bibliographie)

3.2.1 Les fichiers shapefiles des départements

Il y a 14 régions au Sénégal. Chacune d'entre elles comporte 3 à 4 départements pour une meilleure gestion administrative. La région de Dakar fait office d'exception, elle contient 5 départements : Dakar, Pikine, Guédiawaye, Rufisque et Keur Massar. Ce dernier n'existe que depuis 2021 par démembrement de celui de Pikine. Par conséquent, tout le long de l'étude, nous ne considérerons que les 45 départements qui existaient avant 2021, en considérant donc Keur Massar comme intégré à Pikine.

Image 3 : Carte des 45 départements du Sénégal avant 2021

3.2.2 Les productions agricoles

Il s'agit des données concernant les productions agricoles par département, pour plusieurs types de cultures. Les relevés sont faits annuellement de 2017 à 2022, inclus.

Cette base comportait, pour certaines cultures, une quantité non négligeable de données manquantes. Nous verrons par la suite comment elles ont été traitées.

Pour la suite de l'étude, nous allons nous concentrer autour d'une culture de référence sur laquelle nous testerons et mettrons au point tous nos modèles de prédiction avant un éventuel

élargissement : notre choix se porte sur l'arachide. En effet, il s'agit de la culture la plus répandue au Sénégal, avec une zone de production plus accrue vers le centre du pays. De plus, c'est l'une des rares cultures disponibles pour laquelle il n'y avait aucune donnée manquante et il n'y a pas d'énormes disparités de production d'un département à un autre comme cela peut être le cas pour d'autres cultures, nous le verrons plus tard.

Suite à l'obtention de ces données, nous avons pu représenter les cartes de chaleur de la production d'arachides de 2017 à 2022. Ci-après, nous représentons celles pour 2017 et 2022 :

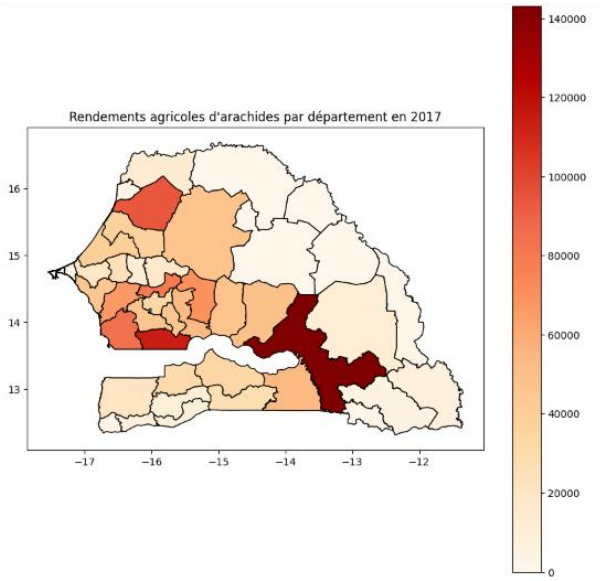

Image 4 : Production d'arachides par département en 2017, en tonnes

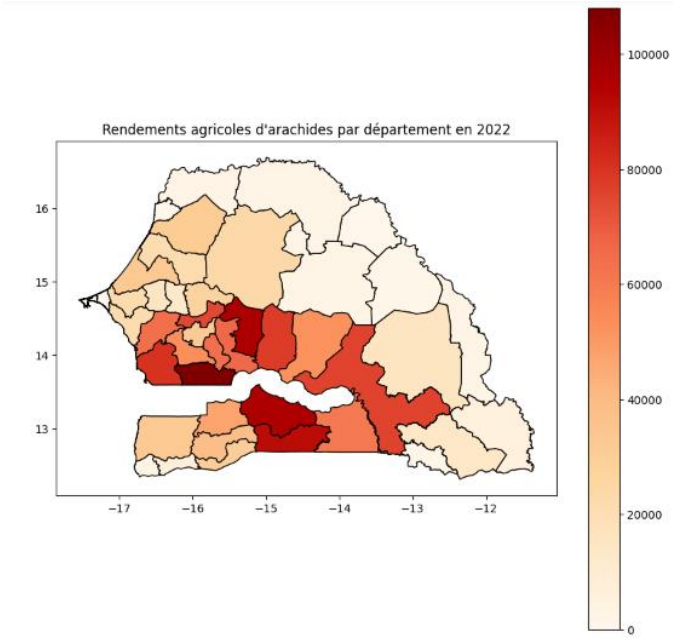

Image 5 : Production d'arachides par département en 2022, en tonnes

3.2.3 Les données d'entrée du modèle

Après avoir enfin trouvé les données pour notre variable cible, c'est-à-dire les productions agricoles, il faut maintenant trouver des éléments qui pourraient permettre d'expliquer ces productions. Plusieurs choix s'offraient à nous.

- Approche uniquement satellitaire ? : Dans l'estimation de productions agricoles, beaucoup de méthodes produites ces dernières années se basent sur l'étude de données géo spatiales. Cela est notamment dû au fait que celles-ci sont de plus en plus accessibles, en bonne résolution et en libre-service surtout. Cela peut permettre de réduire les coûts d'une étude portant sur l'agriculture et les productions agricoles de façon assez significative. De plus, des modèles innovants ont été développés autour de cette problématique. C'est donc un travail en pleine expansion.
  Cependant, il ne s'agit pas de l'approche que nous avons choisie. En effet, les données dont nous disposons sont réparties en départements. Il ne s'agit pas de productions par parcelle. Cette approche aurait été plus pertinente dans ce dernier cas car certains départements sont vastes. Faire la moyenne des valeurs des canaux des images de façon systématique pour chaque département peut s'avérer être préjudiciable. De plus, ces méthodologies innovantes ont été développées dans un but d'économies de prospection de données. Or, dans le cas que nous traitons actuellement, les données sont beaucoup moins volumineuses et nous pouvons donc diversifier un peu nos ressources.
- Approche hybride ? : Nous faisons donc le choix d'une approche hybride dans laquelle nous combinons des indices calculés à partir de données satellites avec des données issues de relevés météorologiques. Nous avons pu trouver des données d'entrée qui correspondent à ce que nous cherchions. Elles sont disponibles sur https://www.aagwa.org/. AAGWA (African Agriculture Watch) est une initiative née de plusieurs DataScientists et chercheurs africains dont le but est de collecter des données et de faire de la documentation de méthodes de machine learning innovantes dans le but de relever les challenges liés à l'agriculture en Afrique.
  Nos données d'entrée seront donc la pluviométrie, la température du sol et le NDVI (Normalized Difference Vegetation Index). Le NDVI (Indice de Végétation par Différence Normalisée) est un indicateur couramment utilisé en télédétection pour mesurer la densité et la santé de la végétation. Il varie entre -1 et +1. Plus le NDVI est élevé, plus la végétation est dense et en bonne santé. Il est calculé à partir des réflectances dans les bandes du proche infrarouge (NIR) et du rouge (Red) des images satellitaires. En agriculture, le NDVI est essentiel pour les modèles de prédiction de production agricole car il fournit des informations sur l'état de la végétation et la vigueur des cultures au cours de la saison de croissance. En intégrant ces données dans des modèles prédictifs, on peut estimer plus précisément les rendements agricoles et optimiser les pratiques de gestion agricole.
  Les 3 features ont chacune des temporalités différentes. Afin d'harmoniser l'étude, nous les mettrons toutes au format mensuel en faisant des moyennes.
  Ci-dessous, nous avons un aperçu de la carte NDVI moyen du Sénégal en Avril 2024 :

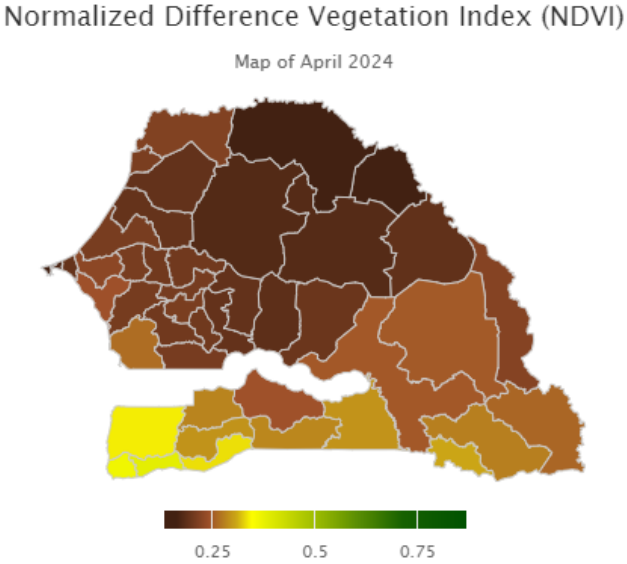

Image 6 : NDVI par département du Sénégal en Avril 2024

Notons que l'initiative AAGWA réalise le même travail pour plusieurs pays en Afrique. Les relevés que nous trouvons commencent en 2020. En recoupant avec les données que nous avons trouvé précédemment sur les productions, nous obtenons 3 années de données exploitables pour un entraînement de machine learning.

4 Inférence statistique

4.1 Choix du modèle

Comme précisé précédemment, depuis quelques temps, les thématiques tournant autour du machine learning pour l'agriculture intéressent de plus en plus d'ingénieurs et de chercheurs, surtout en Afrique. En effet, les défis liés à la sécurité alimentaire restent toujours à relever dans plusieurs pays et l'Intelligence Artificielle, plus précisément le Deep Learning, sont assez souvent mis à contribution afin de produire des modèles robustes et fiables capables de produire des résultats intéressants.

Dans cette étude, nous utiliserons des réseaux de neurones afin de faire nos prédictions. En effet, ces modèles réussissent à capturer des relations complexes, non linéaires notamment, entre plusieurs variables. Cela permet à terme de découvrir des patterns qui seraient plus difficiles à repérer avec les méthodes classiques. Les résultats rendus par les autres méthodes telles que les forêts aléatoires ou les Gradient Boosting se sont avérés mauvais.

4.1.1 Qu'est-ce qu'un réseau de neurones ?

Un réseau de neurones est une architecture mathématique inspirée des réseaux neuronaux du cerveau humain, utilisée pour résoudre des problèmes complexes de traitement de données, comme la classification d'images ou encore les prédictions. Il est composé de plusieurs couches de « neurones » artificiels, organisées en une séquence structurée. Chaque neurone prend en entrée un ensemble de données, applique une fonction d'activation pour transformer cette entrée, et produit une sortie qui sert d'entrée à la couche suivante. Les couches du réseau sont reliées entre elles par des poids, des paramètres qui contrôlent l'importance de chaque connexion. Ces poids sont ajustés automatiquement pendant l'apprentissage pour réduire l'écart entre les prédictions du réseau et les résultats attendus, ce qui permet d'améliorer progressivement la précision du modèle.

Les couches du réseau peuvent être de différents types, chacune ayant un rôle particulier : les couches de convolution identifient et extraient les caractéristiques importantes des données, comme les motifs ou les textures dans une image; les couches de normalisation contrôlent la distribution des données pour stabiliser et accélérer l'apprentissage; les couches de pooling réduisent la dimensionnalité des données tout en préservant l'essentiel de l'information, limitant ainsi le risque de surapprentissage.

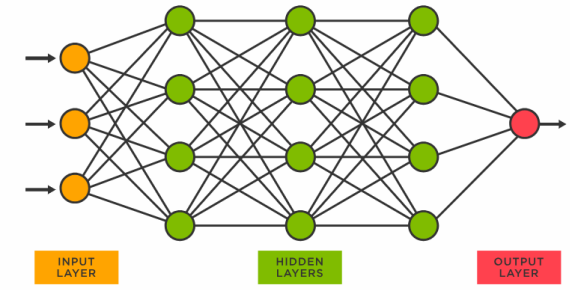

Image 7 : Représentation générale d'un réseau de neurones

L'apprentissage d'un réseau de neurones se fait en deux phases : la propagation et la rétropropagation ou back-propagation. Lors de la propagation, les données d'entrée traversent toutes les couches du réseau pour générer des prédictions, que l'on compare ensuite aux valeurs réelles. Cette comparaison génère une erreur qui mesure la qualité des prédictions. Durant la back-propagation, on ajuste les poids des connexions du réseau afin minimiser cette erreur. Ce processus de propagation et de back-propagation est répété sur de nombreux cycles, appelés « epochs », afin d'optimiser progressivement les performances du réseau sur les données d'apprentissage.

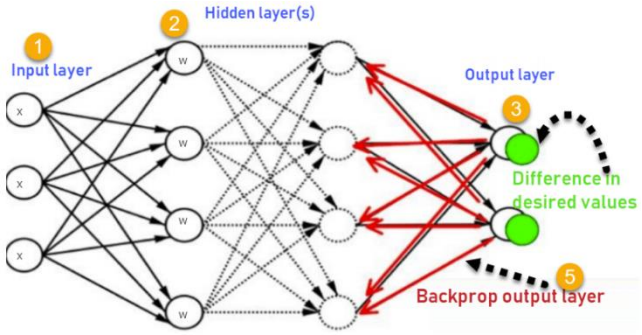

Image 8 : Schématisation de la back-propagation dans un réseau de neurones

Pour implémenter ces modèles en Python, PyTorch est un framework particulièrement adapté, utilisé pour construire et entraîner des réseaux de neurones, notamment convolutifs. La bibliothèque torch.nn permet de définir les différentes couches du réseau, de créer des fonctions d'activation, et de configurer les paramètres d'apprentissage, comme les taux d'apprentissage. Elle offre également des outils pour suivre les performances du modèle et l'améliorer continuellement en ajustant ses poids.

Dans un premier temps, on met en place un modèle simple afin de voir les performances que nous pouvons obtenir en prenant n=3 par exemple. Il va donc là s'agir d'un réseau comprenant 3 couches de convolution (Conv1d) chacune suivie d'une couche de normalisation (BatchNorm1d) et d'une couche ReLU (pour Rectified Linear Unit qui est notre fonction d'activation). Nous ferons varier les différents paramètres de ces couches afin d'améliorer le résultat, notamment la taille du noyau et les tailles d'entrée et de sortie des couches de convolution. Ensuite, après la dernière couche de normalisation, il faut ajouter une couche de pooling pour perdre une dimension sur notre tenseur. Il peut s'agir d'une couche de MaxPooling (maximum) ou d'une couche de AveragePooling (moyenne).

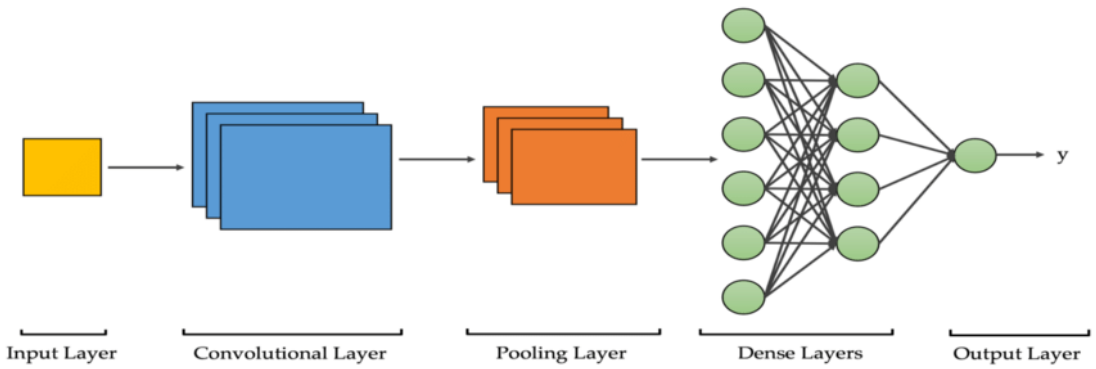

Image 9 : Différents types de couches d'un réseau de neurones

Notons que ce modèle peut avoir plusieurs variantes consistant à augmenter ou à diminuer le nombre de couches de convolution, en mettant les hyperparamètres à jour, en rajoutant des couches fortement connectées (Fully Connected), etc. Il faut trouver un compromis entre complexité du modèle tout en optimisant les performances. Rappelons que, dans l'idéal, nous avons un modèle simple afin de pouvoir mieux interpréter les résultats.

Notons qu'il existe plusieurs modèles « boîte noire » dont les architectures sont déjà mises en place et qui, dans de nombreux cas, produisent des résultats plus intéressants que les modèles classiques. Il s'agit par exemple de LSTM qui est une architecture de réseaux de neurones récurrents destinée à capturer des dépendances à long terme dans les séquences ou encore Inception Time qui lui est spécialement dédié aux séries temporelles. Nous n'aurons toutefois pas le temps nécessaire pour appliquer ces modèles à nos données et en visualiser les résultats ; d'autant plus que les modèles classiques produisent déjà des prédictions globalement correctes.

4.1.2 Détails d'implémentation

Il est temps de tester les modèles. Pour cela, nous devons choisir notre fonction de coût (criterion) et notre optimiseur (optimizer). Pytorch en propose plusieurs et, selon le cas, certains sont plus efficaces et mèneront à de meilleurs résultats. Nous utiliserons dans l'extrême majorité des cas la fonction de coût CrossEntropyLoss() et l'optimizer optim.Adam() dont nous ajusterons le taux d'apprentissage (paramètre lr) au besoin (sinon, il sera établi à 10exp-3 qui rend déjà de bons résultats). C'est un optimiseur populaire utilisé pour ajuster les poids lors de l'apprentissage. Il met en œuvre l'algorithme d'optimisation Adam (Adaptative Moment Estimation).

Il faut bien sûr au préalable désigner les jeux de données d'entraînement et de test. Les années d'entraînement correspondent à 2020 et 2021. 2022 servira d'année test. Nous comparerons les productions prédites pour cette année-là aux productions réelles puis nous en tirerons les métriques d'erreur associées.

A présent, il faut choisir le nombre d'epochs d'apprentissage. C'est un paramètre important de la modélisation, il a une influence directe sur nos résultats. En effet, si on le choisit trop bas, l'entraînement ne dure pas assez longtemps et les performances sont limitées. Si on contraire, on le choisit trop élevé, on risque de faire du sur apprentissage : c'est ce qui arrive lorsqu'un modèle est beaucoup trop adapté à ses données d'entraînement au point de faire de très bons résultats dessus mais qu'il est très difficile voire impossible de le généraliser à d'autres données brutes ou à des données de test car il est trop spécialisé.

Lors de l'étude, nous nous sommes également rendus compte de l'impact important qu'avait le taux d'apprentissage sur nos résultats. En effet, il influence aussi grandement la qualité des prédictions. Il détermine la vitesse à laquelle un modèle ajuste ses poids lors de l'apprentissage. Un taux trop élevé peut conduire à des oscillations et à une convergence difficile, tandis qu'un taux trop faible ralentit l'apprentissage et peut empêcher le modèle d'atteindre une solution optimale.

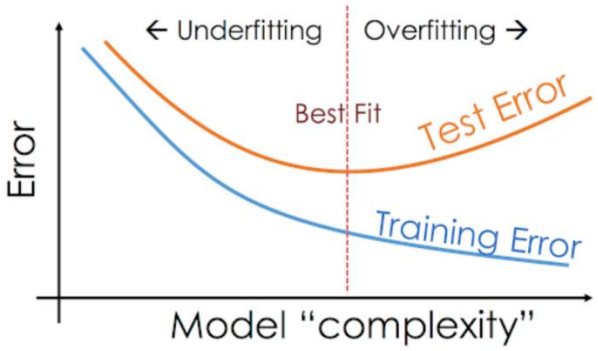

Image 10 : Illustration des cas d'Overfitting et d'Underfitting

4.2 Description de l'étude

4.2.1 Pré-traitement des données

Maintenant que nous avons décrit le modèle que nous allions utiliser, expliqué pourquoi il est intéressant de l'utiliser sur notre problème et exposé les points de vigilance importants dans son utilisation, il est temps de commencer notre étude.

Nous chargeons donc la base des productions départementales annuelles en arachides ainsi que les données météorologiques (pluviométrie, température du sol) et le NDVI pour lesquels nous faisons des moyennes mensuelles. Le seul problème de données manquantes auquel nous avons fait face dans la base des features s'est révélé lors du chargement de la pluviométrie : il n'y avait aucune donnée sur le département de Guédiawaye. Bien heureusement, c'était le seul cas. Nous avons donc convenu, pour ce département, de considérer que sa pluviométrie moyenne était la moyenne des 2 départements les plus proches, autrement dit Dakar et Pikine, comme nous le voyons sur l'illustration ci-après :

Image 11 : Départements de Dakar Région, avant 2021

Nous avions par ailleurs fait remarquer que l'un des éléments qui motivait notre choix de travailler avec l'arachide dans un premier temps était également l'absence de données manquantes, sauf pour les départements de Dakar, Pikine et Guédiawaye dans lesquels la production est en fait nulle en raison de leur fort taux d'urbanisation. Cela est valable quelle

que soit la culture que nous choisissons. Cela veut donc dire qu'en réalité, le problème que nous avons rencontré avec Guédiawaye pour la pluviométrie n'aura pas un grand impact sur nos résultats. Nous concaténons ensuite l'ensemble des données pour les 3 années d'étude.

4.2.2 Construction des premiers réseaux

Comme dit précédemment, nous mettrons nous même au point les différents réseaux que nous allons utiliser grâce au package torch.nn. Nous décidons dans un premier temps d'entraîner notre modèle uniquement sur l'année 2021 et tester les prédictions obtenues pour 2022. Avant de faire passer les features dans le réseau, il convient de les normaliser. En général, cela permet d'accélérer la convergence de l'apprentissage et améliorer les performances du modèle en garantissant que toutes les caractéristiques ont une échelle comparable. Le réseau utilisé est un réseau simple contenant 4 couches fortement connectées, d'une couche de dropout et d'une couche d'activation (ReLU ici). Nous avons ensuite fait un deuxième modèle comportant lui 3 couches de convolution et des couches d'activation ReLU. Nous entraînons les modèles sur une centaine d'epochs. Ci-après les résultats obtenus avec le 2ème modèle qui fait légèrement mieux que le premier:

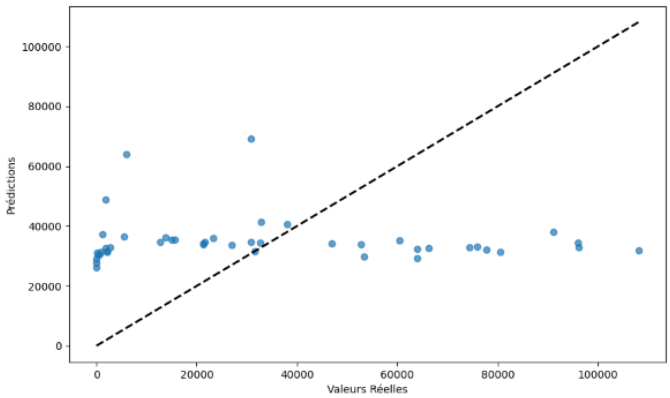

Image 12 : Comparaison des Valeurs réelles et des Prédictions pour 2022

Ce graphique présente en abscisses les vraies valeurs de production par département pour 2022 et en ordonnées les prédictions faites pour cette même année. Chaque point correspond donc à un département. Nous voyons que l'ajustement n'est pas bon du tout. Le R2 score est proche de 0. L'ajout de couches de batch normalization n'améliore que très légèrement les résultats qui restent mauvais. Ces couches permettent de normaliser les sorties des couches en amont permettant ainsi de stabiliser l'entraînement. L'ajustement est encore moins bon lorsqu'on utilise le modèle sans normaliser les données en entrée, comme on pouvait s'y attendre. Le modèle apprend très peu, voire pas du tout. La courbe de prédiction reste quasiment horizontale.

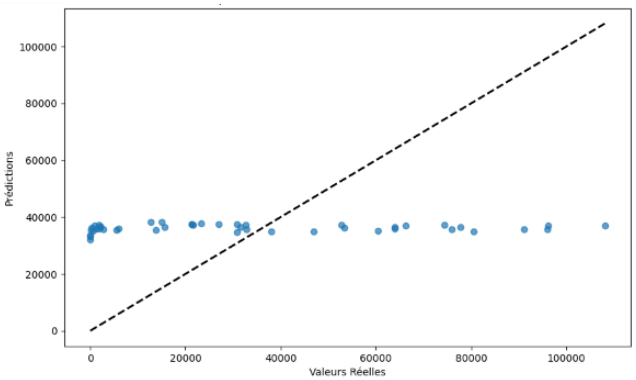

Image 13 : Comparaison des valeurs réelles aux prédictions sans normalisation des données d'entrée

Rappelons que nous travaillions en considérant uniquement l'année 2021 pour année d'entraînement. Nous allons maintenant utiliser les 2 années disponibles pour entraîner notre modèle et faire les tests sur l'année 2022. Nous utilisons cette fois un réseau de neurones similaire au 2ème. Il comporte 3 couches de convolution avec des couches d'activation en sortie de chacune des couches, et une couche de pooling en sortie de réseau. Le « learning rate » ou taux d'apprentissage, fixé arbitrairement à 0.01 lors des précédentes estimations, est augmenté. Nous le faisons passer à 0.1. En effet, le modèle semblait ne pas apprendre des données, ou apprendre beaucoup trop lentement pour le nombre d'epochs fixé. Il s'agit des 2 paramètres influant grandement sur la qualité de l'apprentissage : le nombre d'epochs et le taux d'apprentissage. Ci-après le graphique de comparaison Valeurs prédites/Valeurs réelles :

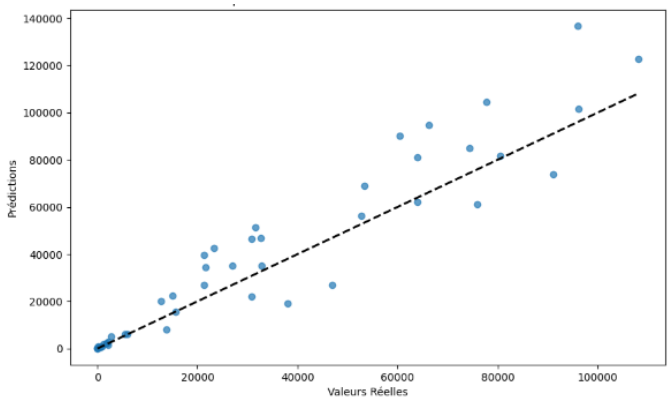

Image 14 : Comparaison des valeurs réelles aux prédictions, learning rate à 0.1

Le R2 score est 0.8206. Nous obtenons un ajustement correct des données. Sur le graphique, nous voyons que tous les départements se situent plus ou moins proches de la première bissectrice. On s'aperçoit également, comme nous pourrions nous y attendre, qu'en général, plus les prédictions sont élevées, plus les erreurs commises, donc les distances des points à la droite, sont grandes.

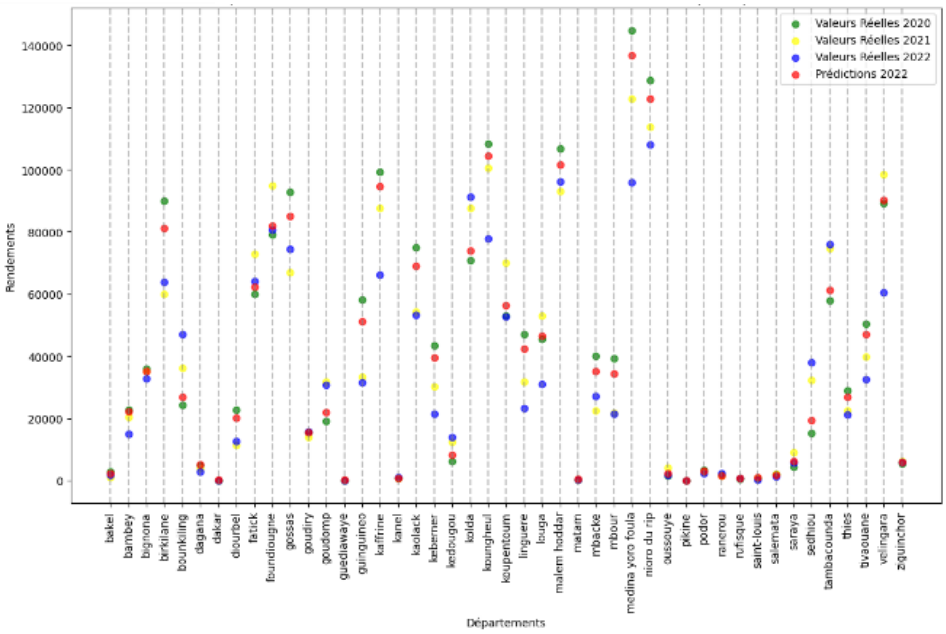

Image 15 : Comparaison Valeurs réelles/Prédictions par département

Le graphique ci-dessus nous permet de comparer les données d'entraînement aux données test et aux prédictions pour chacun des départements. D'abord, sa forme confirme le bon ajustement que nous avons obtenu : les prédictions (en rouge) sont généralement proches des valeurs réelles (en bleu) pour 2022. Dans un second temps, nous comparons ces prédictions avec les productions en 2020 et 2021. Nous constatons que, dans la quasi-totalité des cas, la prédiction pour 2022 se trouve entre les valeurs pour 2020 et 2021. Par exemple, pour le département de Medina Yoro Foula, la production d'arachides en 2022 est largement inférieure à ce qu'elle était les 2 années précédentes. Pourtant, la prédiction pour 2022 se trouve entre les productions 2020 et 2021. Pour remédier à cela nous pensons à 2 solutions.

4.2.3 Amélioration des modèles

La première est de modifier légèrement notre réseau de neurones afin d'inclure des couches qui permettront un meilleur apprentissage à partir des features. Nous créons donc une version modifiée du réseau que nous utilisons ici en lui ajoutant deux couches fortement connectées (fully connected) en fin de réseau. Ces couches relient chacun des neurones de la couche précédente à la couche suivante et permettent ainsi de recueillir toute l'information avant la prise de décision ; dans notre cas, avant la prédiction. Nous appelerons ce réseau « modèle fully connected ». Le modèle est entraîné pour 300 epochs.

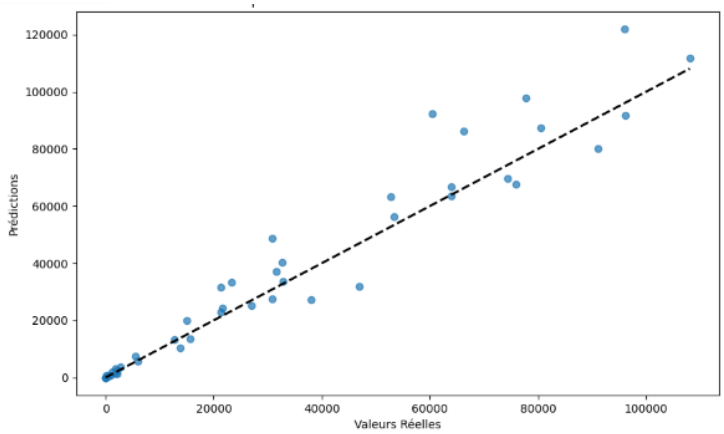

Image 16 : Comparaison Valeurs réelles/Valeurs prédites, modèle fully connected

Nous obtenons un R2 score de 0.91. L'ajustement des valeurs s'est amélioré.

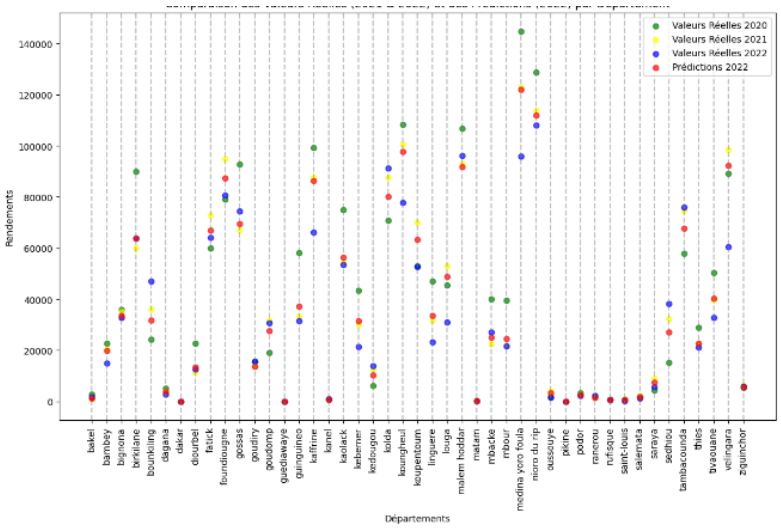

Image 17 : Comparaison Valeurs réelles 2020 2021 2022/Valeurs prédites, modèle fully connected

En regardant de plus près le graphique ci-dessous, nous nous apercevons que le problème mentionné tout à l'heure est toujours présent mais n'est plus la règle. En effet, prenons le département de Nioro du Rip. La prédiction pour 2022 est hors du segment 2020-2021. Elle tend à se rapprocher de la réelle production 2022 pour ce département : notre modèle a mieux appris des données.

Notre deuxième solution consiste à modifier la fonction de perte du modèle. Rappelons que dans un réseau de neurones, à chaque epoch, les sorties du réseau sont comparées aux vraies valeurs via une fonction de perte. A partir de cette comparaison, les poids des branches du réseau sont actualisés (backward) puis les données sont repassées à travers le réseau (forward) et ainsi de suite. Ainsi, le choix de cette fonction de perte a une influence non négligeable sur nos résultats. Pytorch dispose déjà de fonctions de perte implémentées.

Depuis le début de notre étude, nous utilisons la fonction de perte MSELoss() qui calcule l'erreur quadratique moyenne (Mean Squared Error) entre les prédictions du modèle et les valeurs cibles. C'est la fonction généralement utilisée pour les problèmes de régression. Nous décidons de créer notre propre fonction de perte dont le but est de donner plus de poids aux années les plus récentes dans le calcul de l'erreur. Pour être plus précis, nous donnons 70% d'importance à 2021 et le reste, soit 30%, à 2020. S'agissant de séries temporelles, courtes certes, les productions pour une année donnée sont plus en corrélation avec celles des années récentes que celles des années plus anciennes. C'est l'hypothèse que nous faisons. Nous tentons d'optimiser le taux d'apprentissage qui sera à 1.5 mais également le nombre d'epochs que nous fixons à 300. Ci-dessous, le graphique obtenu après avoir apporté ces modifications et entraîné notre modèle :

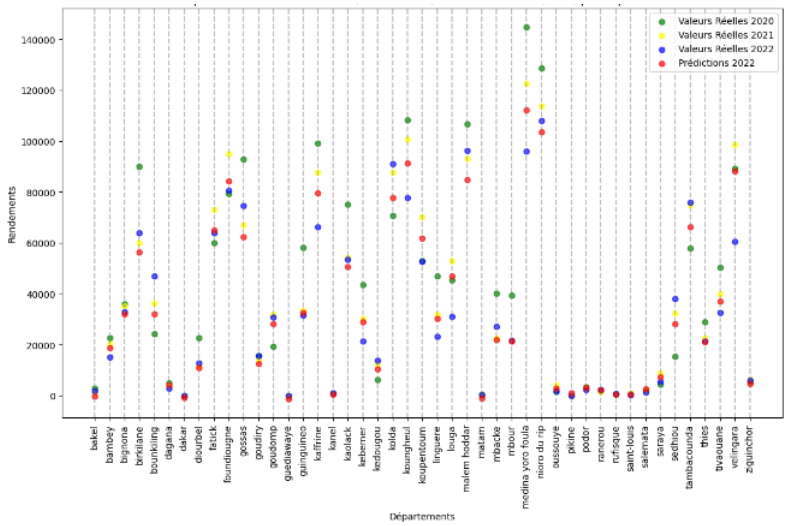

Image 18 : Comparaison Valeurs réelles 2020 2021 2022/Valeurs prédites, modèle fully connected, Fonction de perte implémentée

Nous notons tout d'abord une amélioration du R2 score : celui-ci est passé à 0.9369. Il s'agit du meilleur résultat obtenu depuis le début de notre étude. Regardons le graphique de plus près : nous avons quasiment réussi à traiter le problème que nous cherchions à résoudre en entamant cette démarche. Les prédictions ne sont plus systématiquement limitées entre les productions 2020 et 2021. L'apprentissage est plus poussé et le modèle semble mieux apprendre des données. Il suffit de comparer, pour chacun des départements, l'évolution de la prédiction par rapport à la valeur réelle sur les images 15, 17 et 18. Les prédictions sont de plus en plus exactes.

Il convient de rappeler les modèles que nous construisons, vu la taille de notre fenêtre temporelle d'étude, ne peuvent être utilisés que pour des prédictions à court terme. En effet, le modèle ne sera pas assez robuste pour faire des prédictions à long terme, voire à moyen terme. Les données disponibles ne s'étalent pas assez dans le temps afin qu'il puisse améliorer son apprentissage. Or, nous savons que, sauf exception, plus il y a de données de qualité, plus les réseaux de neurones sont performants.

4.2.4 Fiabilité des modèles

Toutefois, les modèles déjà produits peuvent facilement être adaptés lorsque plus de données seront disponibles. Il faudra certainement revoir quelques paramètres d'implémentation tels que le taux d'apprentissage, les tailles des canaux d'entrée et de sortie des couches du réseau ou encore le nombre d'epochs d'apprentissage entre autres.

Maintenant que nous avons un modèle qui produit des prédictions correctes et dont l'ajustement est convenable, nous voulons avoir une idée de la précision de nos résultats, autrement dit, construire les intervalles de confiance de nos prédictions. Ces intervalles de confiance seront en fait des indicateurs du degré de précision de nos prédictions et seront liés, non pas à la variabilité inhérente aux données elles-mêmes, mais plutôt à la stabilité de ton modèle d'apprentissage automatique vis-à-vis du processus d'entraînement.

En effet, il faut noter que d'une exécution à une autre, le modèle produit en général des prédictions différentes pour un même département. Cela s'explique principalement par le fait que les poids des connexions dans un réseau sont initialisés de manière aléatoire au début de l'apprentissage, impliquant que différentes trajectoires d'apprentissage peuvent être suivies. Le modèle converge donc vers différents minima locaux de la fonction de perte, entraînant par conséquent une différence dans les prédictions finales du modèle.

Ce que nous voulons à présent, c'est pouvoir quantifier cette divergence à travers la construction d'intervalles de confiance adaptés. Pour cela, nous allons fixé un nombre d'itérations, 100 par exemple. Nous allons répéter l'entraînement ce nombre de fois et produire des prédictions à chaque fois que nous stockerons dans une liste « predictions ».

Ensuite, nous allons ajuster des bornes inf et sup pour chaque département en supposant que les résidus suivent une loi normale. Nous pouvons vérifier cela en traçant l'histogramme des erreurs et en y superposant la courbe de la loi normale associée au caractéristiques de nos donénes (moyenne et écart-type). Le résultat est présenté ci-après :

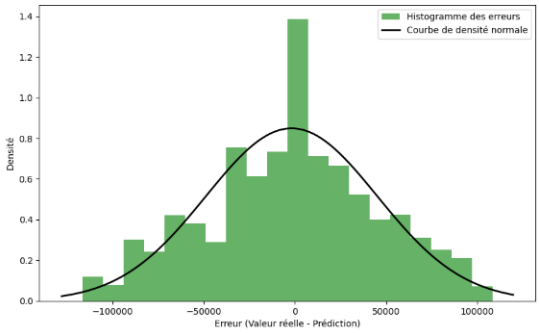

Image 19 : Distribution des erreurs et ajustement à la loi normale

Nous voyons que les erreurs s'ajustent plutôt bien à la loi normale. Nous en profitons vérifier l'indépendance des résidus. Pour cela, nous allons effectuer un test de Durbin-Watson sur le vecteur des erreurs. Le résultat du test est 1.77, soit un résultat proche de 2. Nous pouvons donc considérer que les erreurs sont indépendantes. Enfin, concernant l'hypothèse d'homoscédasticité, nous ne la vérifierons pas car nous ne la considérons pas critique dans notre cadre d'étude, comme elle pourrait l'être pour les modèles linéaires classiques. En effet, les réseaux de neurones ont la capacité de s'adapter à des distributions de données complexes et à des variances non constantes des erreurs.

Une fois les intervalles calculés, nous les agrégeons puis en faisons la représentation graphique ci-dessous, en y superposant les valeurs réelles de productions agricoles pour 2022 ainsi que les prédictions moyennes sur les 100 itérations. Le graphique ci-après :

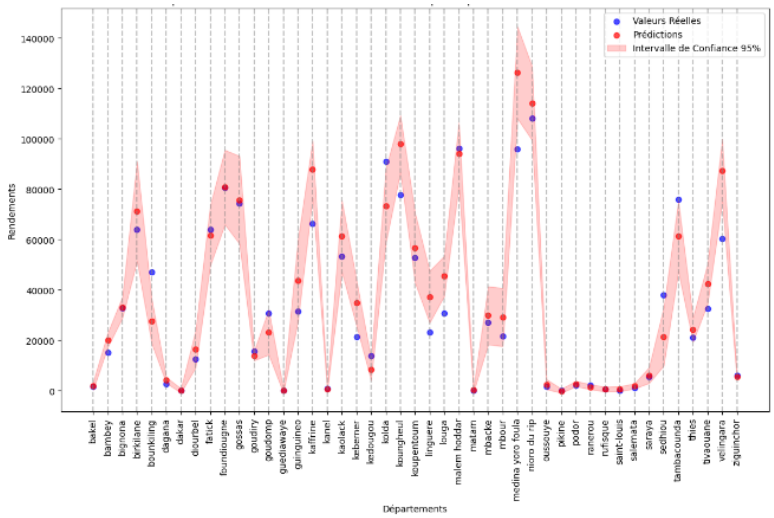

Image 20 : Comparaison Valeurs réelles/Valeurs prédites avec Intervalle de confiance

Dans la majorité des cas, l'intervalle que nous avons construit contient bien la production réelle pour un département donné.

Ce constat d'une différence de prédictions d'une exécution à une autre nous a d'ailleurs conduit à vouloir trouver des méthodes afin de stabiliser l'apprentissage. Pour cela, nous nous inspirons du travail produit pour le calcul des intervalles de valeurs. Nous choisissons un nombre n d'itérations, créons puis entraînons un réseau n fois, faisons des prédictions pour 2022 que nous stockons et calculons à chaque fois les métriques d'erreur du modèle, soit le RMSE (Root Mean Square Error), MSE, r2 score ou encore le MAE (Mean Absolute Error). Nous déduisons de ces calculs les métriques d'erreur moyenne du modèle. Cela permettra de plus facilement comparer différents modèles implémentés entre eux. Notons quand même que, dans le cadre d'une prédiction, il conviendrait d'utiliser celle produite par l'itération la plus performante, soit celle produisant le rmse le plus faible/r2 score le plus élevé.

Après ces travaux sur l'arachide qui nous ont occupé la majeure partie du temps de stage, nous avons voulu étudier les productions pour d'autres cultures. Pour plus de détails, nous pourrons nous référer aux annexes.

5 Conclusion

Au terme de ce stage, plusieurs objectifs ont été atteints pour améliorer l'estimation des rendements agricoles au Sénégal grâce à l'application des techniques de machine learning et à l'utilisation de données multisources. Le travail a débuté par une phase de collecte et de prétraitement des données, incluant l'acquisition d'indices issus d'images satellitaires de la mission Sentinel-2, et de données météorologiques locales. Les indices de végétation calculés à partir des images satellitaires, tels que le NDVI, ont été utilisés comme variables explicatives pour les modèles prédictifs, en combinaison avec les données climatiques.

Au-delà de la création des modèles de deep learning utilisés, une attention particulière a été portée à l'optimisation des paramètres d'apprentissage, tels que le taux d'apprentissage, le nombre d'itérations et la régularisation, afin d'améliorer la précision et la robustesse des prédictions. Une fonction de perte personnalisée a, par exemple, également été développée pour pénaliser davantage les sous-estimations, un choix pertinent dans le contexte où les erreurs de sous-estimation peuvent avoir des conséquences socio-économiques significatives. (cf Annexes)

Les résultats obtenus indiquent que l'utilisation de réseaux de neurones, notamment lorsqu'ils sont alimentés par des données satellites et météorologiques, permet d'améliorer de manière significative la précision des estimations de rendement agricole par rapport aux méthodes traditionnelles. Les modèles développés se sont montrés particulièrement efficaces pour prédire les rendements. Cependant, des défis subsistent, notamment en ce qui concerne la gestion des données manquantes, l'amélioration de la qualité des données et l'intégration de nouvelles sources d'information pour renforcer encore la précision des modèles. Il faudrait aussi encourager l'acquisition de données pertinentes sur des périodes temporelles plus élargies et avec un niveau de précision spatiale plus accru.

En conclusion, ce stage a permis de travailler sur des techniques d'utilisation du machine learning pour l'agriculture au sein de l'Agence. Il reste encore plusieurs pistes d'amélioration à explorer qui, je l'espère, feront l'objet de travaux complémentaires afin de solidifier ce qui a été déjà accompli. Il souligne également l'importance d'une collaboration interdisciplinaire et l'intégration de données variées pour améliorer la prise de décision dans le domaine de la sécurité alimentaire et de la gestion des ressources agricoles. Ces travaux pourraient ainsi contribuer à orienter les politiques publiques et les pratiques agricoles vers une utilisation plus efficace et durable des ressources.

6 Bibliographie

- *Superficie, Rendement et Production agricoles* https://senegal.opendataforafrica.org/ekihmme/superficie-rendement-et-production-agricoles
- Données d'entrée des modèles – Features https://www.aagwa.org/home
- Fichiers shapefiles départements https://www.africageoportal.com/datasets/c741b6c0cb404a42a702ff4999c0bc90/about
- Margaux Masson-Forsythe, *Deep Learning for Crop Yield Prediction in Senegal,* https://www.data4sdgs.org/resources/deep-learning-crop-yield-prediction-senegal
- *How convolutionnal layers work in deep learning neural networks ?*, Jinglescode https://jinglescode.github.io/2020/11/01/how-convolutional-layers-work-deep-learning-neural-networks/
- Page Wikipédia sur les réseaux de neurones : https://fr.wikipedia.org/wiki/R%C3%A9seau_de_neurones_artificiels
- Saint-Cirgue Guillaume, *Machine Learning Français Formation Complete*, https://www.youtube.com/playlist?list=PLO_fdPEVlfKqUF5BPKjGSh7aV9aBshrpY
- Wang, Zhiguang and Yan, Weizhong and Oates, Tim, *Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline* https://arxiv.org/pdf/1611.06455.pdf

7 Annexes : Extension à d'autres cultures

Après avoir élaboré l'ensemble de nos modèles sur les productions d'arachides et au vu de la qualité des résultats obtenus, nous décidons de d'étendre notre étude à d'autres types de cultures pour lesquelles nous disposons de données, principalement 3 : le manioc, le sorgho et le riz. Nous allons brièvement décrire ce qui a été fait pour chacune des cultures.

7.1 Manioc

Il y avait des données manquantes pour le manioc. Après une analyse plus approfondie, nous nous rendons compte qu'il s'agit essentiellement de cas pour lesquels la production est en réalité nulle et le registre n'a pas été rempli. Nous remplaçons donc les Nas par des 0. Nous choisissons d'entraîner un modèle convolutif avec 3 couches de convolution et couches de batch normalization sur nos données. Comme pour l'arachide, 2020 et 2021 sont les années d'entraînement et 2022 est l'année test. Ci-dessous, le graphique de comparaison des valeurs réelles aux prédictions pour 2022 :

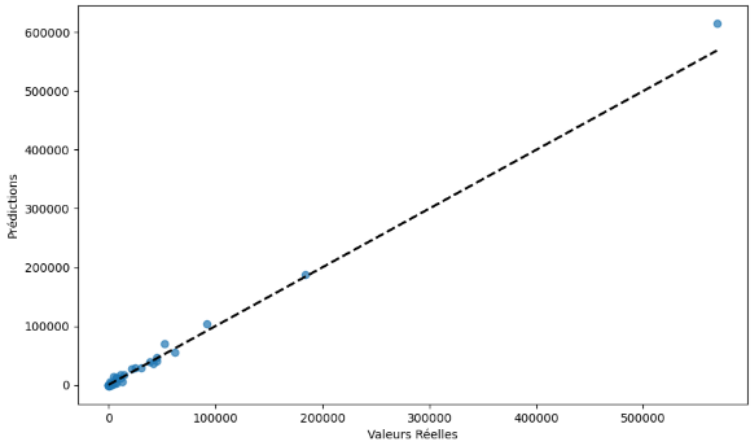

Image 20 : Comparaison Valeurs réelles/Valeurs prédites pour le manioc

L'ajustement des données est très bon. Nous avons un R2 score de 0.9910. Cependant, nous faisons une remarque dans le même temps. Un département se dégage de façon nette dans la quantité de manioc produite et prédite en 2022 : il s'agit de Tivaouane. A priori, sa présence peut sembler ne pas poser problème. En effet, son point représentatif reste quand même assez proche de la droite de régression Valeurs réelles/Valeurs prédites, ce qui signifie que c'est un point influent du modèle, et non pas aberrant. Toutefois, cette grande valeur prise par cet individu peut s'avérer problématique si nous souhaitions généraliser notre modèle. Sa présence aura une grande influence sur les prédictions. Il est donc recommandé de le retirer de notre apprentissage. Nous obtenons un R2 score, certes légèrement plus faible (0.9792), mais unmodèle plus robuste aux changements, et donc plus fiable.

Une fois Tivaouane retiré, un autre département a un comportement similaire vis-à-vis des départements restants : c'est celui de Thiès. Nous le retirons de l'apprentissage pour les

mêmes raisons et obtenons un R2 score de 0.9712, soit en légère baisse une nouvelle fois, mais très correct tout de même.

Nous refaisons la même modélisation en utilisant cette fois nos modèles *fully connected* décrit plus haut, avec des couches fortement connectées donc. Les résultats sont assez similaires. Ils étaient déjà très bons.

7.2 Sorgho

La spécificité de cette 2ème culture est que, comme c'était le cas pour l'arachide, c'est la quasi absence de données manquantes dans la base. Il s'agit essentiellement d'absence de production, que nous allons donc remplacer par des 0. De plus, c'est une culture assez bien répartie sur l'ensemble du territoire national : il n'y a pas de points influents ou aberrants comme nous avons pu le constater avec le manioc dans la partie précédente. Ci-après, la carte de chaleur de la production de sorgho en 2022 au Sénégal :

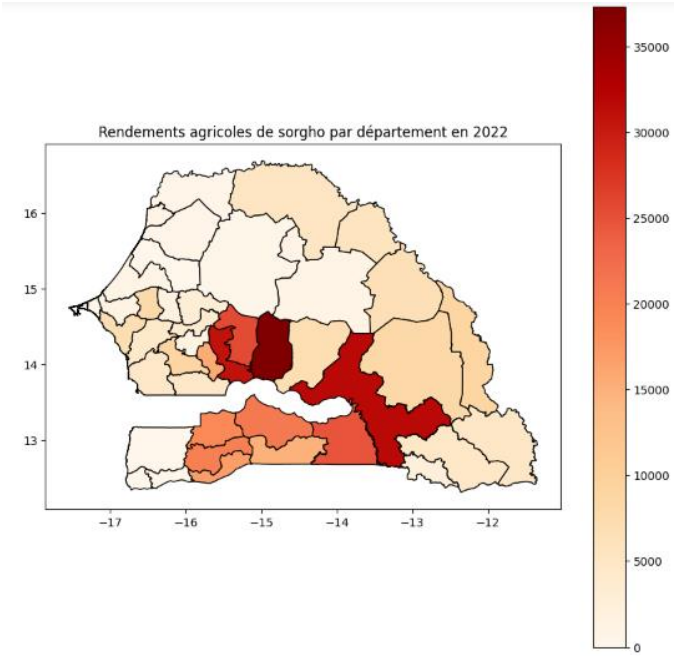

Image 21 : Carte des productions de sorgho par département en 2022

Nous voyons bien sur cette carte qu'un département ne ressort pas de façon flagrante par rapport à tous les autres concernant sa production en sorgho. Koungheul, par exemple, en produit certes beaucoup mais est talonné d'assez près par Tambacounda. Nous considérons donc les données comme équilibrées.

La modélisation utilisée est similaire aux précédentes et produira des R2 score allant globalement de 0.82 à 0.89, sans changements ou innovations majeurs dans les choix des hyper paramètres.

7.3 Riz

Nous avons également appliqué nos modèles aux données de production de riz. Ce travail a été un peu plus laborieux car les chiffres sont très déséquilibrés : certains départements produisent énormément de riz lorsque d'autres n'en produisent pas du tout. De plus, il y avait beaucoup de données manquantes. Nous décidons dans un premier temps de conserver tous les départements, qu'ils aient ou non une production non nulle de riz, et de remplacer les valeurs manquantes par des 0, en supposant qu'il s'agit de productions nulles. Nous appliquons sur ces données les modèles que nous avons conçu auparavant, sans at avec les couches fortement connectées et nous faisons les prédictions pour 2022. Ci-dessous, nous donnons la représentation des résultats avec le modèle sans couche fortement connectée en fin de réseau, en précisant que les 2 modèles ont donné des résultats similaires :

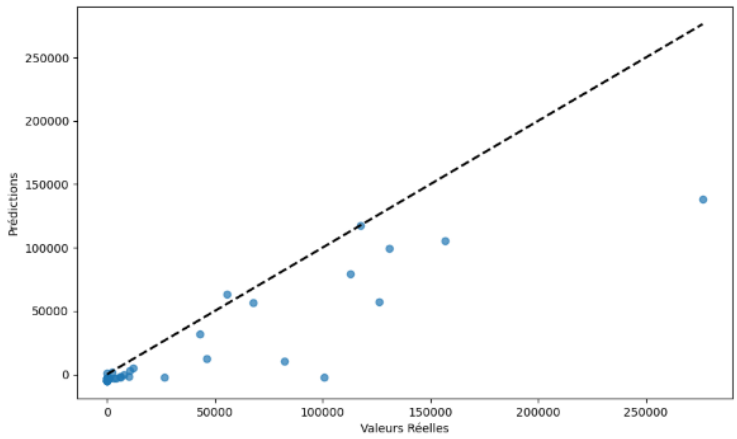

Image 22 : Comparaison Valeurs réelles/Valeurs prédites pour le riz

L'ajustement du modèle n'est pas très bon : nous avons un R2 score de 0.6744. Une première remarque que nous arrivons assez facilement à faire est que, à une exception près, toutes les productions de riz pour 2022 ont été sous-estimées. C'est la première fois que nous faisons face à ce type de comportement. Dans un premier temps, nous avons pensé que cela était dû à la façon dont nous avons traité les données manquantes. Nous décidons donc de retirer de l'entraînement tous les départements pour lesquels il y a au moins 2 valeurs manquantes entre 2020 et 2022 inclus : il en reste 29.

Les résultats des prédictions ne s'améliorent pas, voire se dégradent dans certains cas. Mais surtout, les prévisions continuent d'être largement sous-estimées. Afin de régler ce problème, nous pensons à 2 solutions. Toutes 2 s'appuient sur une modification de la fonction de perte du modèle. Nous avons déjà eu à discuter de l'impact que peut avoir le choix de la fonction de perte ou loss function lorsque nous discutions des résultats de nos modèles sur les productions d'arachide.

Dans un premier temps, nous allons simplement choisir une fonction de perte du module nn autre que MSELoss, qui est celle que nous avons utilisé jusqu'à présent. Nous utilisons la fonction huber qui est recommandée dans les cas où il y a des outliers, ou des valeurs extrêmes. Cette fonction combine la MSE pour les petites erreurs et la MAE pour les grandes erreurs afin d'avoir un entraînement globalement plus robuste, et par conséquent moins sensible aux valeurs aberrantes. Après visualisation des résultats, nous constatons que l'ajustement s'est amélioré mais les prévisions sont toujours sous-évaluées.

La deuxième solution consiste à implémenter nous même une fonction de perte qui pénalisera plus les sous-estimations que les sur-estimations lors de l'apprentissage. Nous appelons cette fonction UnderEstimateLoss(). Après entraînement puis test sur les données 2022, le problème n'est toujours pas réglé.

Nous pensons donc finalement à tout simplement modifier la structure de notre réseau de neurones et de le rendre plus sophistiqué. Nous faisons passer à 4 le nombre de couches de convolution. Elles sont toutes suivies d'une couche de batch normalization et d'une couche d'activation. Nous avons enfin des couches fortement connectées en fin de réseau pour un apprentissage plus fin ainsi que des couches de dropout afin d'éviter le sur-apprentissage. Ci-après, la visualisation de nos résultats de prédiction :

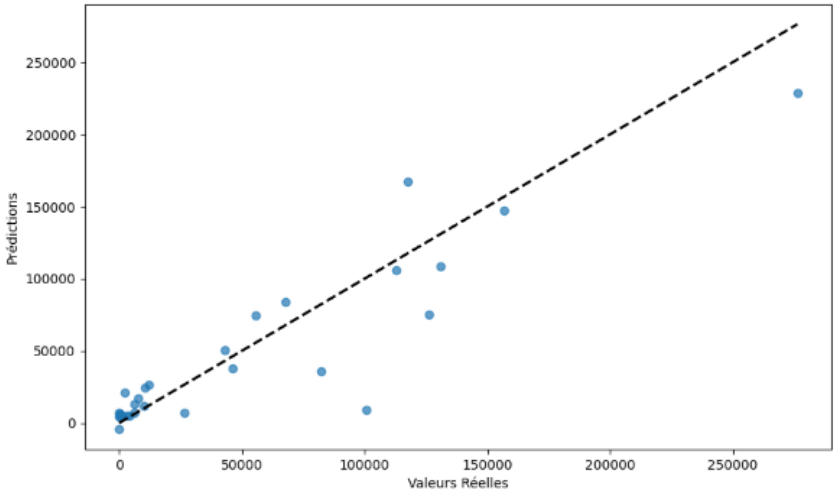

Image 23 : Comparaison Valeurs réelles/Valeurs prédites pour le riz, modèle amélioré

L'ajustement est bon à présent. Les prévisions ne sont plus systématiquement sous évaluées.