Université d'Orléans

Rapport

Rendu de Projet - Big Data Analytics

Élèves :
BARRAUD Lorenzo
JACQUET Léo
MIRZA Simon

Enseignant :
TOKPAVI Sessi

21 janvier 2024

Résumé

Ce rapport explore les dynamiques des communautés en Côte d'Ivoire en utilisant des méthodes de pénalisation et d'agrégation pour prédire la pauvreté des populations locales. En exploitant des données géospatiales telles que la température de l'air et les précipitations, les méthodes d'agrégation, notamment l'arbre de régression et la forêt aléatoire, se démarquent en fournissant les meilleures prédictions des taux de pauvreté. Ces résultats suggèrent que les politiques publiques devraient accorder une attention particulière aux infrastructures et aux initiatives d'adaptation au changement climatique pour soutenir les communautés les plus vulnérables aux phénomènes sociaux et environnementaux. L'intégration de ces approches innovantes dans la planification budgétaire pourrait contribuer significativement à améliorer les conditions de vie dans ces régions nécessiteuses.

Abstract

This report explores the dynamics of communities in Côte d'Ivoire using penalization and aggregation methods to predict poverty among local populations. By leveraging geospatial data such as air temperature and precipitation, aggregation methods, notably regression trees and random forests, stand out in providing the most accurate predictions of poverty rates. These findings suggest that public policies should pay special attention to infrastructure and climate change adaptation initiatives to support communities most vulnerable to social and environmental phenomena. Integrating these innovative approaches into budget planning could significantly contribute to improving living conditions in these needy regions.

Table des matières

1 Introduction **3**

2 Présentation de la base de données **5**
2.1 Le taux de pauvreté comme variable cible 5
2.2 Variables explicatives géospatiales 6

3 Méthodes utilisées **8**
3.1 Régression Linéaire Multiple avec MCO 8
3.2 Méthodes de Pénalisation 8
3.2.1 Ridge 8
3.2.2 LASSO 9
3.2.3 Adaptive-LASSO 9
3.2.4 Elastic-Net 10
3.3 Méthodes d'agrégation 10
3.3.1 Arbre de Régression 10
3.3.2 Forêt Aléatoire 11
3.3.3 Gradient Boosting 11

4 Critères de comparaisons **12**
4.1 Mean Absolute Error (MAE) 12
4.2 Root Mean Squared Error (RMSE) 12
4.3 Concordance Correlation Coefficient (CCC) 13
4.4 Coefficient de Détermination (R^2) sur Out-of-Sample 13

5 Analyse des résultats **14**
5.1 Performances des modèles 14
5.2 Interprétation des Résultats 15
5.3 Importance des variables 16
5.4 Shapley Values 17

6 Conclusion **19**

7 Annexes **21**

1 Introduction

La Côte d'Ivoire, au cœur de sa dynamique économique, se retrouve à l'intersection complexe des promesses de croissance et des défis persistants de la pauvreté. Malgré les avancées notables, la réalité socio-économique du pays est marquée par des disparités et des enjeux spécifiques qui entravent le progrès vers une prospérité partagée. Dans cette équation délicate, la nécessité d'une approche pour comprendre, cibler et réduire la pauvreté au niveau des communautés les plus vulnérables s'impose comme une impérative.

Le contexte mondial, façonné par les Objectifs de Développement Durable (ODD) énoncés par les Nations Unies en 2015, souligne la nécessité d'éradiquer la pauvreté sous toutes ses formes d'ici 2030. La Côte d'Ivoire s'inscrit dans cette vision globale mais doit également naviguer à travers ses réalités locales complexes. Le premier objectif des ODD, "Pas de pauvreté", reflète l'engagement crucial de cibler les populations les plus vulnérables, mettant en lumière la nécessité d'approches adaptées.

La fin des Objectifs du Millénaire pour le Développement en 2015 a été marquée par des constats contrastés. Alors que la croissance économique mondiale était vigoureuse, des millions de personnes étaient toujours piégées dans un cycle persistant de pauvreté. En Côte d'Ivoire, malgré les progrès depuis la fin de la guerre civile en 2007, près de 40% de la population vit toujours dans la pauvreté en 2018 (World Development Indicators, Banque mondiale). Cette situation souligne l'urgence d'innovations dans les stratégies de réduction de la pauvreté.

Les approches conventionnelles, telles que les transferts monétaires directs et les travaux publics, jouent un rôle vital, mais leur efficacité dépend de la précision du ciblage des communautés nécessiteuses. Dans ce contexte, l'utilisation de données géospatiales émerge comme un élément clé pour comprendre les dynamiques locales et améliorer l'efficacité des programmes de réduction de la pauvreté.

Le présent rapport s'inscrit dans cette perspective en proposant une approche innovante qui intègre des données climatiques et géographiques dans des modèles de prévision de la pauvreté à l'échelle communautaire. En se focalisant sur des paramètres tels que la lumière nocturne, la température de l'air, et d'autres indicateurs géospatiaux, notre objectif est d'établir des connexions entre ces données et le bien-être des populations, offrant ainsi une vision plus holistique de la pauvreté au niveau local.

Cette démarche trouve sa justification dans la compréhension que les données géospatiales peuvent agir comme des ponts essentiels entre les conditions environnementales, l'activité économique, l'agriculture et le bien-être des communautés. En allant au-delà des mesures monétaires traditionnelles, nous espérons dévoiler des tendances qui pourraient échapper aux analyses conventionnelles.

L'originalité de ce rapport réside dans son engagement à explorer les dynamiques locales souvent négligées par les approches classiques, jetant ainsi une lumière nouvelle sur les stratégies potentielles pour améliorer le bien-être des communautés les plus défavorisées.

Le rapport s'engage ensuite dans une analyse comparative des méthodologies statistiques, évaluant l'efficacité des méthodes de pénalisation telles que les MCO, Ridge, Lasso et Adaptive Lasso ou Elastic-Net ainsi que des méthodes d'agrégation telles qu'un arbre de régression, Random Forest et Gradient Boosting. Cette évaluation vise à identifier les méthodologies les plus adaptées pour améliorer la précision des prévisions de pauvreté, facilitant ainsi des interventions plus ciblées et éclairées.

Ce rapport se positionne comme une contribution significative à la lutte contre la pauvreté en Côte d'Ivoire. En intégrant des données géospatiales avancées et en adoptant des méthodologies statistiques innovantes, nous aspirons à fournir des outils précieux pour les décideurs, les chercheurs et les acteurs du développement. Nous nous efforçons de dépasser les modèles traditionnels en offrant une vision plus complète, plus contextuelle et plus adaptée à la réalité locale de la pauvreté, ouvrant ainsi la voie à des politiques plus efficaces, équitables et humaines.

2 Présentation de la base de données

Cette section se consacre à détailler les variables utilisées dans le cadre de la prédiction de la pauvreté à une échelle spatiale locale en Côte d'Ivoire. L'ensemble de données fournit des informations concernant pas moins de 11 734 communautés, englobant tout le territoire et pour lesquelles des données géoréférencées uniques, telles que la latitude et la longitude, sont disponibles. Ce groupe constitue notre échantillon de travail pour l'estimation et l'évaluation des modèles alternatifs.

2.1 Le taux de pauvreté comme variable cible

Les indicateurs de pauvreté inclus dans notre ensemble de données découlent d'une démarche méticuleuse, reposant sur la désagrégation des données de l'Étude de mesure des niveaux de vie en 2018. Ces données fournissent des insights précieux sur le statut de pauvreté des ménages, lesquels sont rigoureusement catégorisés comme étant soit pauvres, soit non pauvres, en conformité avec les critères définis pour les besoins fondamentaux. Il convient de noter que cette classification s'applique uniformément à tous les individus au sein de chaque ménage. Il est possible de représenter le taux de pauvreté en fonction des communautés sur une carte.

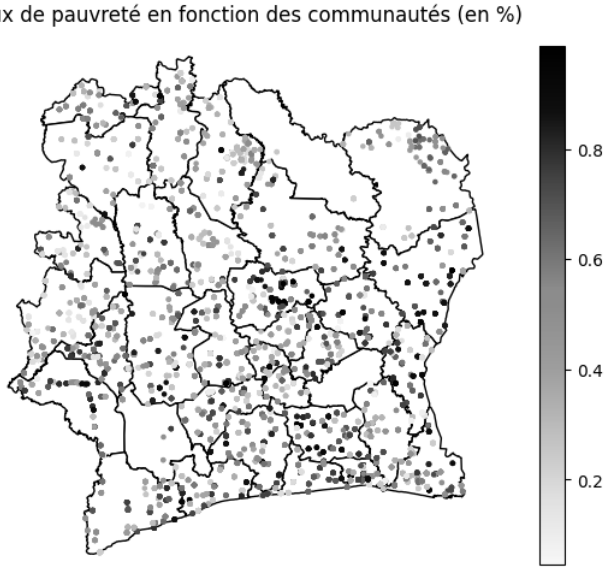

FIGURE 1 – **Réprésentation graphique du taux de pauvreté**

Ainsi, la méthode de calcul des taux de pauvreté au niveau communautaire s'opère en déterminant la proportion d'individus caractérisés comme étant dans un état de pauvreté au sein de chaque communauté. Cette approche, ancrée dans la réalité socio-économique des ménages, permet d'appréhender de manière nuancée et représentative la dynamique de la pauvreté à l'échelle locale.

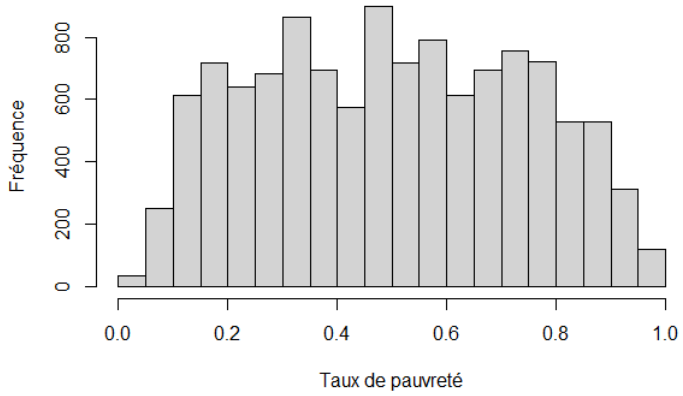

FIGURE 2 – **Histogramme - Taux de pauvreté**

2.2 Variables explicatives géospatiales

Au sein de notre base de données, nous disposons d'un ensemble d'informations essentielles liées à plusieurs variables géospatiales, toutes prises en compte dans le cadre de notre démarche de prédiction des taux de pauvreté à l'échelle communautaire.

Ces données comprennent des éléments tels que l'éclairage nocturne, les niveaux de précipitations, la température de l'air, ainsi que l'indice standardisé d'évapotranspiration des précipitations (IESP). En plus de ces paramètres, nous avons également intégré deux variables supplémentaires, à savoir la densité de population et la densité de la zone bâtie, pour enrichir notre modèle prédictif.

Ces variables géospatiales jouent un rôle crucial dans notre analyse, offrant une perspective contextuelle et environnementale. L'inclusion de l'éclairage nocturne permet d'appréhender les variations d'activité humaine, tandis que les données sur les précipitations, la température de l'air et l'IESP fournissent des indications sur les conditions climatiques locales. Par ailleurs, la prise en compte des variables de densité de population et de densité de la zone bâtie contribue à intégrer des aspects démographiques et d'aménagement du territoire dans notre modèle.

TABLE 1

Statistic	Mean	St. Dev.	Min	Max
Poverty rate	0.500	0.240	0.045	0.987
Nightlights	6.861	11.292	0.185	59.410
Precipitation	112.083	33.432	52.117	240.417
Air temperature	27.027	0.638	24.425	28.692
SPEI	−0.778	0.377	−2.375	0.276
Density	4,003.172	5,247.509	0.000	28,361.180
Built up Area Density	20.510	29.030	0.000	99.235

Afin de compléter notre base de données, et afin de présenter des résultats pertinents lors des méthodes que nous utiliserons dans la prochaine section, nous décidons de créer un certains nombre de variables :

— L'ensemble des variables explicatives au carré.
— L'ensemble des variables explicatives au cube.
— Des interactions entre les variables explicatives deux à deux.
— Des interactions entre les variables explicatives trois à trois.

Cela nous amène à une base de données avec **59 variables**.

3 Méthodes utilisées

Dans cette section, nous présentons les différentes méthodes utilisées, à savoir :

1. Une méthode d'estimation par MCO.
2. Des méthodes de pénalisation telles que Ridge, Lasso, AdaptiveLasso et ElasticNet.
3. Des méthodes d'agrégations telles qu'un arbre de régression, une forêt aléatoire et un GradientBoosting.

3.1 Régression Linéaire Multiple avec MCO

Considérons un modèle de régression linéaire multiple avec k variables indépendantes $X_1, X_2, \ldots, X_k$ et une variable dépendante Y. Le modèle peut être représenté comme suit :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \tag{1}$$

où :
— Y est la variable dépendante,
— $X_1, X_2, \ldots, X_k$ sont les variables indépendantes,
— β_0 est une constante,
— $\beta_1, \beta_2, \ldots, \beta_k$ sont les coefficients de régression,
— ε est le terme d'erreur.

Le but des Moindres Carrés Ordinaires (MCO) est de minimiser la somme des carrés des écarts entre les valeurs observées et les valeurs prédites par le modèle.

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}))^2 \tag{2}$$

où n est le nombre d'observations.

La solution des MCO peut être obtenue en résolvant les équations normales :

$$X^T X \beta = X^T Y \tag{3}$$

où X est la matrice des variables explicatives.

3.2 Méthodes de Pénalisation

En plus des Moindres Carrés Ordinaires (MCO), il existe des méthodes de pénalisation qui peuvent être utilisées dans le cadre de la régression linéaire multiple pour régulariser les coefficients de régression. Les méthodes de pénalisation introduisent une pénalité dans la fonction objectif pour prévenir le surajustement du modèle. Voici quelques-unes des méthodes populaires que nous utilisons :

3.2.1 Ridge

La méthode Ridge, également connue sous le nom de régularisation $L2$, vise à réduire la magnitude des coefficients de régression en ajoutant une pénalité quadratique à la fonction objectif des Moindres Carrés Ordinaires (MCO). La fonction objectif à minimiser devient :

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \left[\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}))^2 + \lambda \sum_{j=1}^{k} \beta_j^2 \right] \tag{4}$$

où λ est le paramètre de pénalité qui contrôle l'intensité de la régularisation. Plus λ est grand, plus la pénalité est forte, ce qui conduit à une réduction plus importante des coefficients.

La pénalité $L2$ favorise des coefficients plus petits en ajoutant une composante quadratique à la somme des carrés des coefficients. Cela a pour effet de "répartir" la contribution des différentes variables plutôt que de favoriser des coefficients extrêmement élevés pour certaines variables.

L'avantage principal de la méthode Ridge est qu'elle peut gérer des situations où il y a une multicollinéarité élevée entre les variables indépendantes, car elle "étale" l'impact des variables corrélées.

3.2.2 LASSO

La méthode LASSO introduit une pénalité absolue dans la fonction objectif :

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \left[\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}))^2 + \lambda \sum_{j=1}^{k} |\beta_j| \right] \tag{5}$$

Contrairement à la méthode Ridge, la pénalité LASSO peut conduire à la sélection automatique de variables, car elle a la propriété de mettre à zéro certains coefficients de régression. En d'autres termes, LASSO favorise la parcimonie en forçant certains coefficients à être exactement égaux à zéro, réduisant ainsi le modèle à une forme plus simple avec moins de variables.

LASSO est particulièrement utile dans des situations où l'on suspecte qu'un petit nombre de variables explicatives est réellement pertinent pour le modèle, tandis que d'autres peuvent avoir une influence négligeable. Elle peut être considérée comme une méthode de sélection de variables intégrée, ce qui est bénéfique dans des contextes où l'interprétabilité du modèle est importante.

3.2.3 Adaptive-LASSO

L'Adaptive-LASSO est une variation de LASSO avec des poids adaptatifs sur les coefficients :

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \left[\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}))^2 + \lambda \sum_{j=1}^{k} w_j |\beta_j| \right] \tag{6}$$

où les poids w_j sont choisis de manière à réduire la pénalité pour les variables importantes.

La particularité de l'Adaptive-LASSO réside dans le choix des poids w_j . Ces poids sont souvent définis de manière à réduire la pénalité pour les variables importantes, ce qui signifie que des poids plus élevés sont attribués aux variables jugées plus cruciales pour la

prédiction. Le choix des poids peut être basé sur des mesures d'importance des variables, telles que des scores de régression partielle ou d'autres critères pertinents au contexte.

L'Adaptive-LASSO combine ainsi les avantages de la parcimonie de LASSO avec une certaine flexibilité permettant de traiter les variables de manière différenciée en fonction de leur importance relative. Cette méthode est particulièrement utile lorsque certaines variables ont un impact significatif tandis que d'autres ont une influence moindre.

3.2.4 Elastic-Net

L'Elastic-Net est une méthode de régression qui combine les pénalités de Ridge (L2) et de LASSO (L1). Cela permet de profiter des avantages de la régularisation L1 en favorisant la parcimonie des coefficients, tout en incorporant la régularisation L2 pour traiter la multicollinéarité. La fonction objectif à minimiser est formulée comme suit :

$$\min_{\beta_0,\beta_1,\ldots,\beta_k} \left[\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}))^2 + \lambda_1 \sum_{j=1}^{k} |\beta_j| + \lambda_2 \sum_{j=1}^{k} \beta_j^2 \right] \quad (7)$$

où λ_1 et λ_2 contrôlent les pénalités LASSO et Ridge respectivement.

L'Elastic-Net offre une solution équilibrée entre les avantages de Ridge et de LASSO. Il est particulièrement utile lorsque nous avons un grand nombre de variables potentiellement corrélées et que nous souhaitons bénéficier de la sélection automatique de variables tout en atténuant les effets de la multicollinéarité.

3.3 Méthodes d'agrégation

3.3.1 Arbre de Régression

Un arbre de régression est un modèle prédictif utilisé dans l'apprentissage automatique et la statistique. Contrairement aux arbres de classification qui visent à prédire une classe discrète, un arbre de régression est conçu pour prédire une valeur continue. Cela en fait un outil puissant pour modéliser des relations non linéaires entre les variables. L'arbre de régression fonctionne en divisant récursivement l'ensemble des données en sous-ensembles homogènes, jusqu'à ce que chaque sous-ensemble représente une valeur de sortie prédictive. Chaque nœud de l'arbre représente une condition sur une variable, et les branches résultantes mènent à d'autres nœuds ou feuilles.

La Figure 3 illustre schématiquement un arbre de régression. Chaque nœud teste une condition sur une variable, et les branches gauche et droite représentent les résultats de la condition. Les feuilles de l'arbre contiennent les valeurs de sortie prédites.

L'avantage des arbres de régression réside dans leur capacité à capturer des relations complexes entre les variables sans avoir besoin de spécifier une forme fonctionnelle a priori. Cependant, ils peuvent être sensibles au surajustement, et des techniques telles que la taille maximale de l'arbre ou la taille minimale de l'échantillon dans une feuille sont souvent utilisées pour contrôler cela.

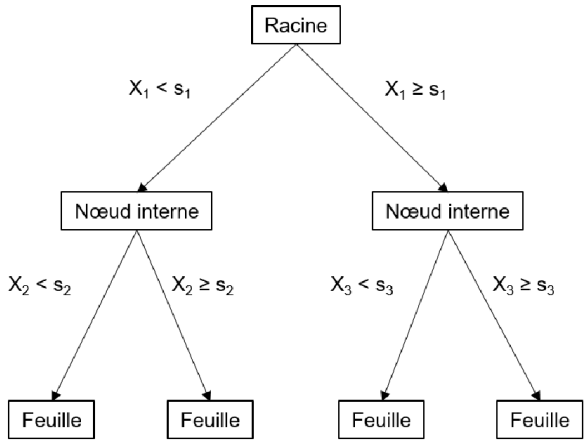

FIGURE 3 – **Exemple schématique d'un arbre de régression.**

3.3.2 Forêt Aléatoire

Une forêt aléatoire est une technique d'ensemble qui combine plusieurs arbres de régression pour améliorer la précision et la robustesse du modèle. Plutôt que de s'appuyer sur la prédiction d'un seul arbre, une forêt aléatoire agrège les prédictions de plusieurs arbres et fournit une prédiction finale basée sur leur moyenne (pour la régression).

L'avantage principal des forêts aléatoires réside dans leur capacité à réduire le surajustement et à améliorer la généralisation du modèle. Chaque arbre est construit à partir d'un sous-ensemble aléatoire des données d'entraînement, et chaque nœud de l'arbre teste une condition sur une variable sélectionnée aléatoirement parmi un sous-ensemble des variables. Cela introduit de la diversité dans les arbres individuels, rendant la forêt plus robuste.

3.3.3 Gradient Boosting

Le Gradient Boosting est une technique d'ensemble similaire à la forêt aléatoire, mais qui construit les arbres séquentiellement plutôt qu'en parallèle. À chaque étape, un nouvel arbre est ajouté pour corriger les erreurs résiduelles des prédictions précédentes. La prédiction finale est obtenue en combinant les prédictions de tous les arbres.

Le Gradient Boosting est reconnu pour sa capacité à bien s'adapter à divers types de données. Cependant, il peut nécessiter davantage de temps de calcul par rapport à une forêt aléatoire.

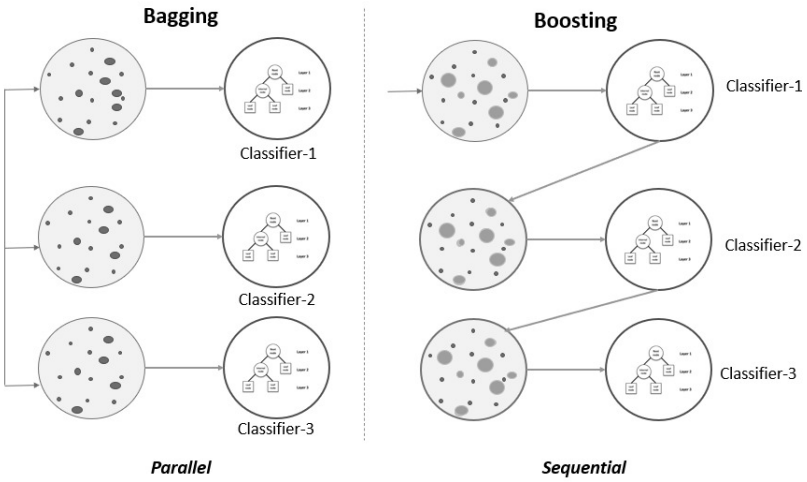

FIGURE 4 – **Différence entre le bagging et le boosting.**

4 Critères de comparaisons

Dans cette section, nous présentons les métriques d'évaluation que nous utiliserons pour comparer les performances de nos modèles de prédiction. Les métriques choisies sont le Mean Absolute Error (MAE), le Root Mean Squared Error (RMSE), le Concordance Correlation Coefficient (CCC) et le coefficient de détermination (R^2) calculé sur l'ensemble de validation (out-of-sample).

4.1 Mean Absolute Error (MAE)

Le Mean Absolute Error (MAE) mesure la moyenne des valeurs absolues des erreurs entre les prédictions de notre modèle (y_{pred}) et les valeurs réelles (y_{true}).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_{\text{true},i} - y_{\text{pred},i}| \tag{8}$$

où n est le nombre total d'observations.

4.2 Root Mean Squared Error (RMSE)

Le Root Mean Squared Error (RMSE) est une mesure qui pénalise davantage les grandes erreurs que le MAE, car il prend la racine carrée de la moyenne des carrés des erreurs.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{\text{true},i} - y_{\text{pred},i})^2} \tag{9}$$

Nous utilisons le RMSE au lieu de la Mean Squared Error (MSE) car la racine carrée permet de revenir à l'échelle d'origine de la variable dépendante, rendant l'interprétation de l'erreur plus intuitive.

4.3 Concordance Correlation Coefficient (CCC)

Le Concordance Correlation Coefficient est une mesure de la concordance entre deux ensembles de données, utilisée pour évaluer à la fois l'accord (concordance) et la précision entre les prédictions de notre modèle (y_{pred}) et les valeurs réelles (y_{true}).

Il est défini comme suit :

$$CCC = \frac{2 \cdot \rho \cdot \sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{10}$$

où :
— ρ est la corrélation entre les deux ensembles de données,
— σ_x et σ_y sont les écarts types des deux ensembles de données,
— μ_x et μ_y sont les moyennes des deux ensembles de données.

L'indice de performances CCC varie de -1 à 1. Une valeur de 1 indique une concordance parfaite, 0 indique aucune concordance, et -1 indique une discordance parfaite.

4.4 Coefficient de Détermination (R^2) sur Out-of-Sample

Le coefficient de détermination (R^2) mesure la proportion de la variance dans la variable dépendante qui est prévisible à partir de la variable indépendante. Lorsqu'il est calculé sur l'ensemble de validation (out-of-sample), il fournit une évaluation de la capacité de généralisation du modèle.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{\text{true},i} - y_{\text{pred},i})^2}{\sum_{i=1}^{n}(y_{\text{true},i} - \bar{y}_{\text{true}})^2} \tag{11}$$

où $\bar{y}_{\text{true}}$ est la moyenne des valeurs réelles.

L'utilisation de ces métriques d'évaluation, à savoir le MAE, le RMSE, et le R^2 sur l'ensemble de validation, nous permettra d'avoir une vision complète des performances de nos modèles. Ces métriques fournissent des informations sur la précision, la robustesse et la capacité de généralisation des modèles par rapport aux données réelles.

5 Analyse des résultats

L'évaluation des performances des modèles de prédiction de la pauvreté à l'échelle spatiale locale en Côte d'Ivoire est cruciale pour comprendre l'efficacité de chaque méthode.

Dans cette section, nous présentons une analyse approfondie des résultats obtenus à partir de différentes techniques, mettant en évidence les mesures d'erreur absolue moyenne (MAE), la racine carrée de l'erreur quadratique moyenne (RMSE), le coefficient de corrélation de Concordance (CCC), et le coefficient de détermination (R^2).

5.1 Performances des modèles

Les performances des différents modèles sont résumées dans le tableau 2. Les méthodes traditionnelles telles que la Régression Linéaire par les Moindres Carrés Ordinaires (MCO) montrent des résultats prometteurs, avec une MAE de 0.19493 et un RMSE de 0.22975. Cependant, les mesures de concordance (CCC) et le coefficient de détermination (R^2) indiquent une corrélation limitée avec les données réelles.

La méthode de régularisation LASSO, en particulier, se distingue par une MAE de 0.19717 et un RMSE de 0.23219, montrant une efficacité similaire à la MCO. Cependant, malgré cette comparabilité, les mesures de concordance (CCC) et le coefficient de détermination (R^2) suggèrent une corrélation relativement modérée avec les données réelles, soulignant les limites de ces approches linéaires. D'autre part, l'Adaptative-LASSO présente des résultats inférieurs avec une MAE de 0.20375 et un RMSE de 0.23845. Cette performance moindre met en lumière l'importance critique de l'ajustement des paramètres dans les méthodes de régularisation. L'Adaptative-LASSO, en ajustant automatiquement les pondérations des coefficients, nécessite une attention particulière pour optimiser ses performances.

Les modèles basés sur des arbres, tels que l'Arbre de Régression, le RandomForest et l'AdaBoost, démontrent des performances plus robustes. Le RandomForest se distingue particulièrement avec une MAE de 0.18971, indiquant une meilleure capacité à capturer la variabilité des données.

Méthode	MAE	RMSE	CCC	R^2
MCO	0.194 93	0.229 75	0.155 98	0.088 69
Ridge	0.199 42	0.234 96	0.068 22	0.041 73
LASSO	0.197 17	0.232 19	0.118 86	0.063 12
Adaptative-LASSO	0.203 75	0.238 45	0.016 05	0.011 91
Elastic-Net	0.197 35	0.232 28	0.113 48	0.062 50
Arbre de régression	0.191 80	0.228 12	0.190 30	0.107 52
RandomForest	0.189 71	0.224 47	0.194 69	0.135 87
AdaBoost	0.197 08	0.232 02	0.093 61	0.076 77

TABLE 2 – **Performances des modèles de prédiction de la pauvreté.**

5.2 Interprétation des Résultats

L'observation de performances relativement similaires entre les méthodes traditionnelles et les méthodes de régularisation souligne l'importance de choisir la méthode en fonction du contexte spécifique de la prédiction de la pauvreté. Alors que les méthodes de régularisation peuvent être efficaces dans la gestion des problèmes de multicollinéarité et de surajustement, leur performance peut être sensible aux paramètres choisis.

L'analyse des résultats suggère que les modèles basés sur des techniques d'ensemble, tels que le RandomForest et l'AdaBoost, surpassent les approches linéaires traditionnelles. Leur capacité à capturer des relations non linéaires dans les données spatiales semble être un atout crucial pour améliorer les prédictions de la pauvreté. Cependant, malgré des performances prometteuses, tous les modèles présentent des limites dans la prédiction de la pauvreté à une échelle spatiale locale. Des recherches ultérieures pourraient explorer des variables supplémentaires ou des techniques plus avancées pour améliorer davantage la précision des modèles.

En outre, afin d'acquérir une compréhension approfondie de la contribution de chaque variable à la prédiction, une analyse détaillée de l'importance des variables sera entreprise, mettant particulièrement l'accent sur le modèle RandomForest. Cette approche permettra de démystifier les facteurs qui jouent un rôle significatif dans la capacité du modèle à prédire la pauvreté à l'échelle spatiale locale en Côte d'Ivoire.

L'analyse de l'importance des variables dans le contexte du RandomForest s'appuiera sur des métriques telles que le Gain d'Information, qui mesure l'effet de chaque variable sur la réduction de l'incertitude dans la prédiction. Cette métrique offre un aperçu de la contribution relative de chaque variable dans la prise de décision du modèle.

Parallèlement, l'utilisation de l'approche Shapley fournira une perspective encore plus approfondie sur l'impact de chaque variable. Cette technique, issue de la théorie des jeux, attribue une valeur à chaque variable en mesurant son influence marginale dans la prédiction du modèle. L'application des valeurs de Shapley permettra ainsi d'identifier les interactions entre les variables et d'évaluer leur contribution globale à la performance du modèle.

En examinant spécifiquement le RandomForest, qui a démontré des performances prometteuses dans la section précédente, cette analyse approfondie de l'importance des variables offrira des insights cruciaux pour interpréter les mécanismes sous-jacents du modèle. Les résultats obtenus aideront à identifier les facteurs les plus pertinents dans la prédiction de la pauvreté et à orienter les futures étapes d'amélioration du modèle.

5.3 Importance des variables

Le graphique d'importance des variables présenté ci-dessous est basé sur un modèle de forêt aléatoire utilisé pour prédire le taux de pauvreté dans une zone donnée. Les variables sont classées par ordre d'importance, de la plus importante à la moins importante.

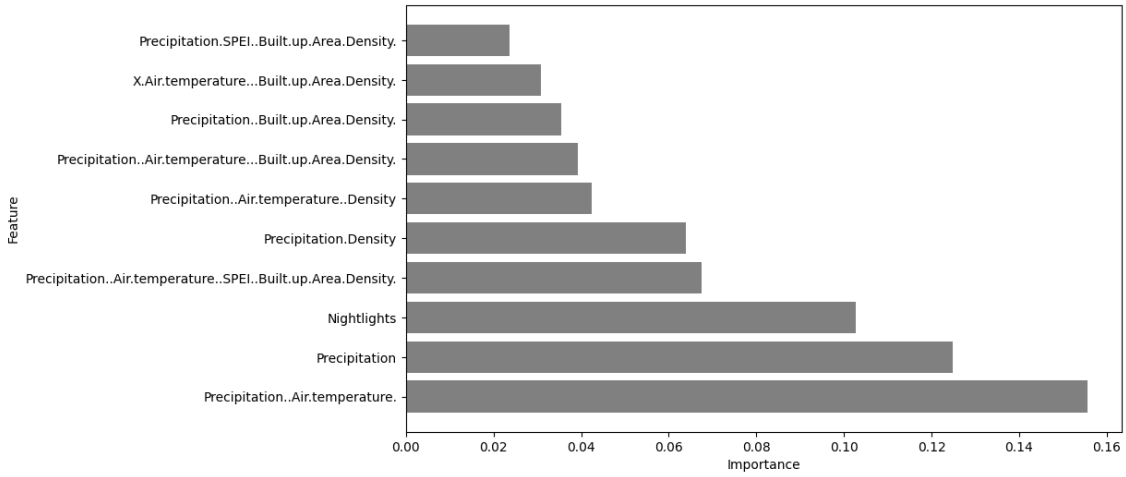

FIGURE 5 – **Importance des variables**

Les résultats de l'étude montrent que les variables les plus importantes pour prédire le taux de pauvreté sont les suivantes :

— Densité de la zone urbanisée (Built-up Area Density)
— Température moyenne (Air Temperature)
— Précipitation (Precipitation)

La densité de la zone urbanisée est la variable la plus importante, ce qui signifie qu'elle est la variable qui explique le plus le taux de pauvreté. Une densité de zone urbanisée plus élevée est associée à un taux de pauvreté plus faible. La température moyenne est la deuxième variable la plus importante. Une température moyenne plus élevée est associée à un taux de pauvreté plus faible. La précipitation est la troisième variable la plus importante. Une précipitation plus élevée est associée à un taux de pauvreté plus élevé.

Les autres variables sont moins importantes, mais elles peuvent également jouer un rôle dans la prédiction du taux de pauvreté.

En particulier, la combinaison de la précipitation et de la température moyenne est également importante. Une précipitation plus élevée et une température moyenne plus élevée sont associées à un taux de pauvreté plus élevé.

La combinaison de la densité de la zone urbanisée et de la température moyenne est également importante. Une densité de zone urbanisée plus élevée et une température moyenne plus élevée sont associées à un taux de pauvreté plus faible.

Les résultats de cette étude suggèrent que les politiques et les programmes de lutte contre la pauvreté devraient se concentrer sur les facteurs suivants :

— Augmentation de la densité de la zone urbanisée
— Réduction de la température moyenne
— Réduction de la précipitation

Ces mesures pourraient contribuer à réduire le taux de pauvreté dans les zones concernées.

5.4 Shapley Values

La figure ci-dessous affiche les 20 valeurs de Shapley les plus importantes en valeur absolue pour l'individu dont le taux de pauvreté prédit est le plus élevé. Ces valeurs mesurent l'impact de chaque variable explicative sur la prédiction de la pauvreté de cet individu.

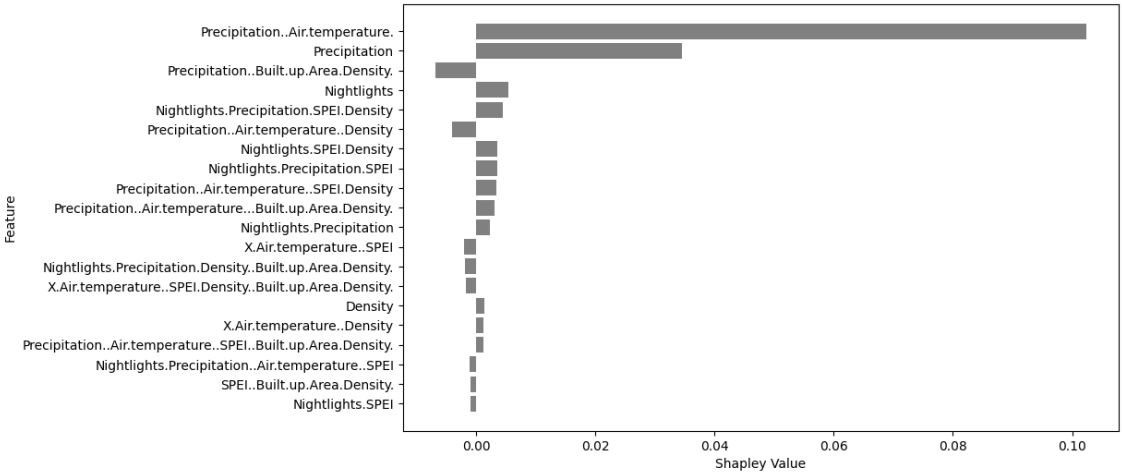

FIGURE 6 – **Valeurs de Shapley**

Ces résultats suggèrent que les facteurs environnementaux, tels que la densité de population construite, le SPEI, la température de l'air et les précipitations, jouent un rôle important dans la détermination du risque de pauvreté. Ces facteurs peuvent avoir un impact sur les revenus, l'accès aux services essentiels et les opportunités d'emploi.

Il est important de noter que ces valeurs de Shapley ne sont que des estimations. Elles peuvent être influencées par la qualité des données et la complexité du modèle utilisé.

Sur la base de ces résultats, les politiques publiques visant à réduire la pauvreté devraient cibler les zones à forte densité de population construite, à faible niveau d'humidité du sol et à températures élevées. Ces politiques pourraient inclure :

— Des investissements dans les infrastructures et les services essentiels, tels que l'éducation, la santé et les transports.
— Des mesures pour atténuer les effets des changements climatiques, tels que l'adaptation aux sécheresses et aux inondations.
— Des programmes de développement économique visant à créer des emplois et à améliorer les revenus.

En ciblant ces facteurs, les politiques publiques peuvent contribuer à réduire le risque de pauvreté et à améliorer le bien-être des populations vulnérables.

En conclusion, cette analyse fournit des informations précieuses sur la performance des modèles de prédiction de la pauvreté, ouvrant la voie à des améliorations futures et à une meilleure compréhension des facteurs influant sur la pauvreté à l'échelle locale en Côte d'Ivoire.

6 Conclusion

Au sein de ce projet, nous avons mis en place un ensemble de méthodes visant à prédire le taux de pauvreté dans différents villages en Côte d'Ivoire.

Dans le cadre de la Régression Linéaire Multiple (RLM) avec la Méthode des Moindres Carrés Ordinaires (MCO), bien que certains coefficients soient significatifs, la grande majorité a un impact très faible sur la variable cible.
Du côté des méthodes de régularisation, ces résultats d'estimation se confirment avec la plupart des coefficients ramenés à 0. D'autre part, le modèle RLM et les modèles de pénalisation montrent des performances assez similaires. En revanche, nous pouvons considérer ces performances comme étant mauvaises étant donné la part de variabilité expliquée par ces derniers.
Le modèle RLM affiche les meilleures performances parmi l'ensemble des modèles linéaires.

Le Machine Learning démontre de meilleures performances sur tous les indicateurs. Cela peut s'expliquer en partie grâce à sa capacité à comprendre les interactions multiples entre les variables et les relations non-linéaires, allant au-delà de la simple création de variables d'interaction comme dans les modèles linéaires.
Le RandomForest présente les meilleures performances parmi les modèles de Machine Learning. En comparant ce dernier avec l'AdaBoost, on pourrait émettre l'hypothèse que cela s'explique par le fait que Scikit-Learn offre plus de possibilités d'hyperparamètres pour le RandomForest, ce qui permet une meilleure optimisation et, par conséquent, de meilleures performances.

D'autre part, les modèles de Machine Learning peuvent également servir de référence en tant que benchmark. Bien que leur interprétabilité puisse constituer une limitation, ils peuvent néanmoins être utilisés comme point de comparaison pour évaluer les performances des modèles classiques et s'assurer de leur validité.

Enfin, il aurait été intéressant d'avoir des données dynamiques, c'est-à-dire sous forme de panel. En effet, avec plus d'observations, nous aurions pu mener une étude plus précise et comparer les différents villages, par exemple avec des modèles de panel ou même de Panel VAR. Cela aurait été particulièrement utile dans le cas où l'on pourrait imaginer que les conditions climatiques passées impactent finalement les conditions de pauvreté actuelles.

Références

[1] World Bank Open Data — donnees.banquemondiale.org. `https://donnees.banquemondiale.org/indicator/SI.POV.NAHC?locations=CI`. [Accessed 21-01-2024].

[2] E. Biernat, M. Lutz, and Y. LeCun. *Data science : fondamentaux et études de cas : Machine learning avec Python et R*. Eyrolles, 2015.

[3] L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Taylor & Francis, 1984.

[4] A. Burkov. *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019.

[5] Adele Cutler, David Cutler, and John Stevens. *Random Forests*, volume 45, pages 157–176. 01 2011.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[7] Christophe Hurlin. Econométrie des variables qualitatives. *Cours de maîtrise d'économétrie, France Université d'Orléans*, 59, 2003.

[8] P. Lemberger, M. Batty, M. Morel, and J.L. Raffaëlli. *Big Data et Machine Learning - 2e éd.* Dunod, 2016.

[9] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

7 Annexes

Variable	Estimation	Pr(>|t|)
Density	0.00302	0.03682
Nightlights..Built.up.Area.Density.	0.02928	0.00728
X.Air.temperature..Density	-0.00011	0.03299
SPEI..Built.up.Area.Density.	0.87763	0.02745
Density..Built.up.Area.Density.	-0.00005	0.02093
Nightlights.Precipitation.Density	-0.00000	0.02271
Nightlights..Air.temperature...Built.up.Area.Density.	-0.00098	0.01296
Nightlights.SPEI.Density	-0.00032	0.00002
Nightlights.SPEI..Built.up.Area.Density.	0.05229	0.00008
X.Air.temperature..SPEI..Built.up.Area.Density.	-0.03137	0.03017
X.Air.temperature..Density..Built.up.Area.Density.	0.00000	0.02450
SPEI.Density..Built.up.Area.Density.	-0.00007	0.00042
Nightlights.Precipitation..Air.temperature..Density	0.00000	0.01613
Nightlights.Precipitation.SPEI.Density	0.00000	0.00000
Nightlights.Precipitation.SPEI..Built.up.Area.Density.	-0.00003	0.00000
Nightlights.Precipitation.Density..Built.up.Area.Density.	0.00000	0.00003
Nightlights..Air.temperature..SPEI.Density	0.00001	0.00003
Nightlights..Air.temperature..SPEI..Built.up.Area.Density.	-0.00180	0.00015
Precipitation.SPEI.Density..Built.up.Area.Density.	0.00000	0.00040
X.Air.temperature..SPEI.Density..Built.up.Area.Density.	0.00000	0.00043

TABLE 3 – **Régression linéaire Coefficients estimés significatifs**

Variable	Estimation
Intercept	0.1558319
Precipitation	0.00003
Density	0.00000
Nightlights	0.00215

TABLE 4 – **Ridge - Extrait des coefficients estimés**

Variable	Estimation
Intercept	2.57715
Precipitation	-0.010467
Density	-0.00004
Nightlights	0.01235

TABLE 5 – **LASSO - Extrait des coefficients estimés**

Variable	Estimation
Intercept	0.04174
Precipitation	0.00000
Density	-0.00002
Nightlights	0.0000

TABLE 6 – **Adaptative-LASSO - Extrait des coefficients estimés**

Variable	Estimation
Intercept	1.54969
Precipitation	-0.00801
Density	-0.00002
Nightlights	0.00007

TABLE 7 – **Elastic-Net - Extrait des coefficients estimés**

L'arbre de régression selectionné a comme paramètres :
— Profondeur : 6
— Nombre d'échantillons pour diviser un nœud : 2
— Nombre d'échantillons dans une feuille : 4

Le modèle Random Forest avec les meilleurs performances, ci-dessous, a comme paramètres :
— Nombre d'arbres : 90
— Profondeur : 6
— Nombre d'échantillons pour diviser un nœud : 2
— Nombre d'échantillons dans une feuille : 5

Pour le modèle AdaBoost, les paramètres sont les suivants :
— Learning rate : 0.01
— Nombre d'arbres : 150

Le graphique ci-dessous montre bien l'intérêt de la validation croisée. En effet, graphiquement, lorsque l'algorithme utilise 150 arbres, nous avons le R^2 moyen le plus faible. Au-delà de ce seuil, le R^2 augmente, ce qui peut être synonyme de sur-apprentissage.

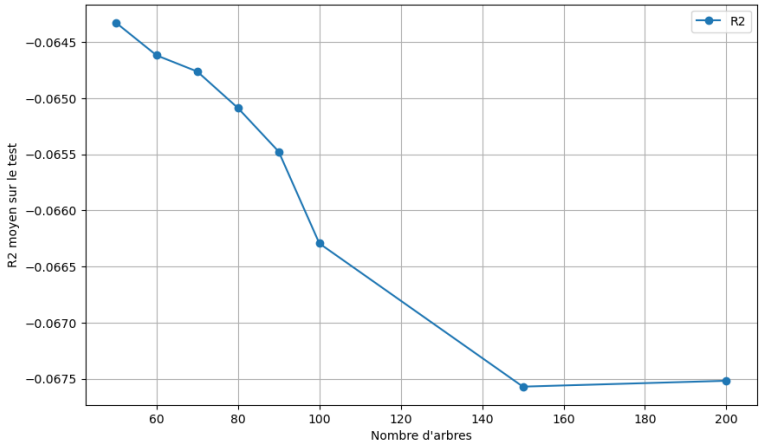

FIGURE 7 –
Evolution du R2 moyen sur l'échantillon test selon le nombre d'abres