Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

École Nationale Supérieure de Statistique et d'Économie Appliquée

ENSSEA (Ex. INPS)

Département économie quantitative et prospective
Option : Économie appliquée et prospective

Rapport du projet : Prédiction de l'indice de richesse à partir de données géospatiales

Module :

Machine Learning

Présenté par:

EL khane amar fakhreddine

2025/2026

Introduction :

L'amélioration de la connaissance du niveau de vie des populations est un enjeu majeur pour les gouvernements, les décideurs publics et les organisations internationales. Cependant, la collecte de données issues des enquêtes auprès des ménages est coûteuse et peu fréquente dans de nombreux pays africains.

Dans ce contexte, l'utilisation de données alternatives telles que les images satellites, les informations sur l'urbanisation, la densité de population et les lumières nocturnes constitue une solution prometteuse pour estimer le niveau de richesse des ménages.

L'objectif principal de ce projet est de prédire un indice de richesse (wealth index) compris entre 0 et 1 à partir de variables géographiques et environnementales disponibles pour plusieurs pays africains. Chaque observation représente un cluster d'enquêtes et la variable cible correspond à la richesse moyenne de ce groupe.

Pour atteindre cet objectif, nous avons appliqué différentes techniques de science des données et d'apprentissage automatique. Le travail commence par une analyse descriptive des données afin de comprendre leur structure et leurs relations avec la variable cible. Ensuite, une analyse non supervisée est réalisée pour identifier des groupes homogènes d'observations. Les informations issues de cette étape sont utilisées pour enrichir le jeu de données par des techniques de feature engineering.

Dans un second temps, des modèles d'apprentissage supervisé sont développés pour prédire l'indice de richesse en tant que variable continue (régression). Une nouvelle variable cible binaire est ensuite construite afin de distinguer les zones riches et pauvres, ce qui permet de mener une seconde analyse supervisée sous forme de classification. Enfin, une procédure d'optimisation des hyperparamètres (tuning) est appliquée afin d'améliorer les performances des modèles et de sélectionner les meilleurs résultats obtenus.

Description des données :

Le jeu de données utilisé dans ce projet est contenu dans le fichier ***wealth.csv***. Il regroupe des informations issues de différentes sources géospatiales et environnementales couvrant l'ensemble du continent africain. Chaque ligne du dataset correspond à un cluster d'enquêtes auprès des ménages, et représente une zone géographique donnée.

La variable cible du projet est l'indice de richesse (*wealth index*), une variable continue comprise entre 0 et 1, où les valeurs proches de 1 indiquent un niveau de richesse élevé et les valeurs proches de 0 un niveau de richesse faible. Cet indice est construit à partir de plusieurs facteurs tels que la possession de biens et les conditions de vie des ménages.

Les variables explicatives sont issues principalement de trois catégories :

- **Données d'urbanisation (GHSL – Global Human Settlement Layer)** : elles décrivent la proportion de surfaces bâties à différentes périodes, la densité de population et la part des surfaces non bâties.

- **Données de couverture des sols (Landcover)** : elles indiquent la proportion de terres agricoles, de zones urbaines et de surfaces d'eau permanentes ou saisonnières autour de chaque cluster.
- **Indicateurs géographiques et économiques indirects** : les lumières nocturnes (*Nighttime_lights*), la distance à la capitale (*Dist_to_capital*) et la distance au littoral (*Dist_to_shoreline*), qui sont des indicateurs reconnus de l'activité économique.

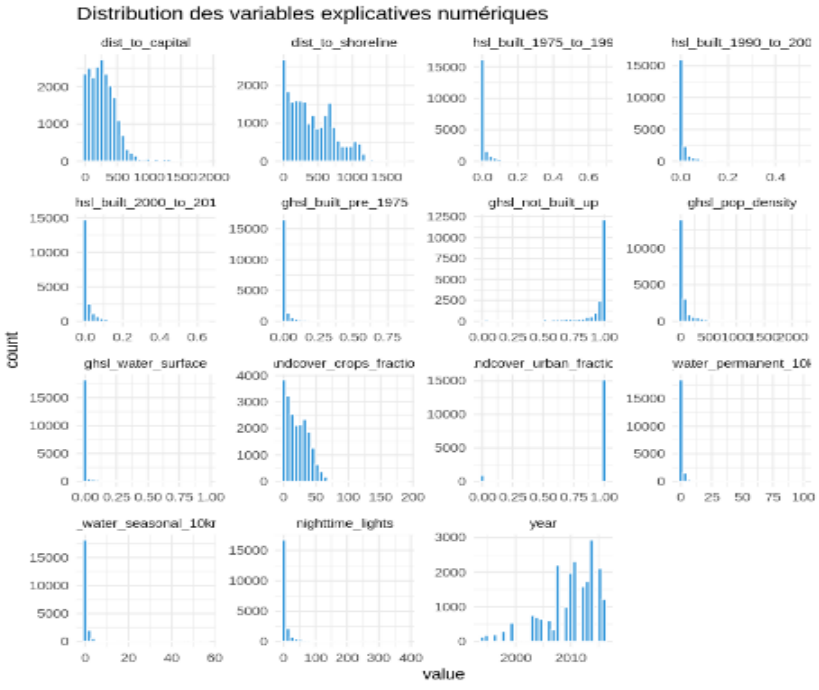
Le jeu de données contient également une variable catégorielle indiquant le type de zone, urbaine ou rurale (*urban_or_rural*).

Avant toute modélisation, une étape de vérification de la qualité des données a été réalisée afin d'identifier la présence de valeurs manquantes, d'anomalies ou de distributions extrêmes et verifie les doublons aussi. Cette étape est essentielle pour garantir la fiabilité des analyses statistiques et des modèles d'apprentissage automatique.

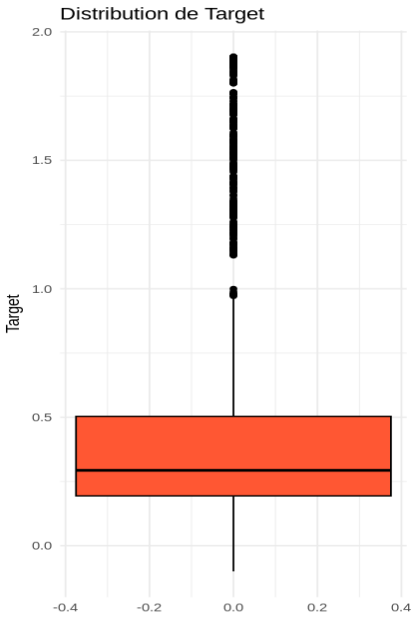
Enfin, les données ont été préparées pour les différentes analyses par des opérations de normalisation et de transformation lorsque cela était nécessaire, notamment pour les méthodes sensibles à l'échelle des variables.

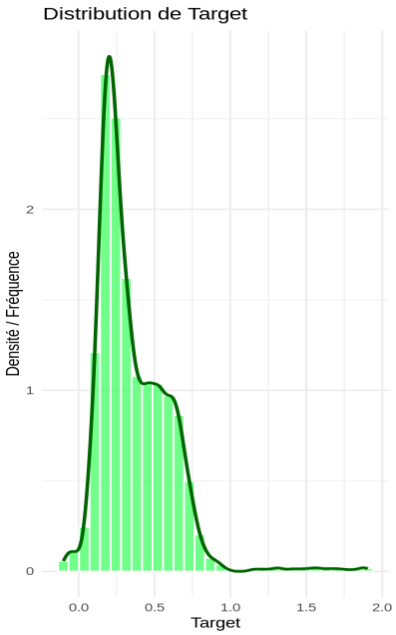
Analyse descriptive (EDA) :

Une analyse exploratoire des données a été réalisée afin de mieux comprendre la distribution des variables et leurs relations avec la variable cible, l'indice de richesse. Cette étape permet d'identifier les tendances générales, les valeurs atypiques et les variables les plus informatives pour la modélisation.



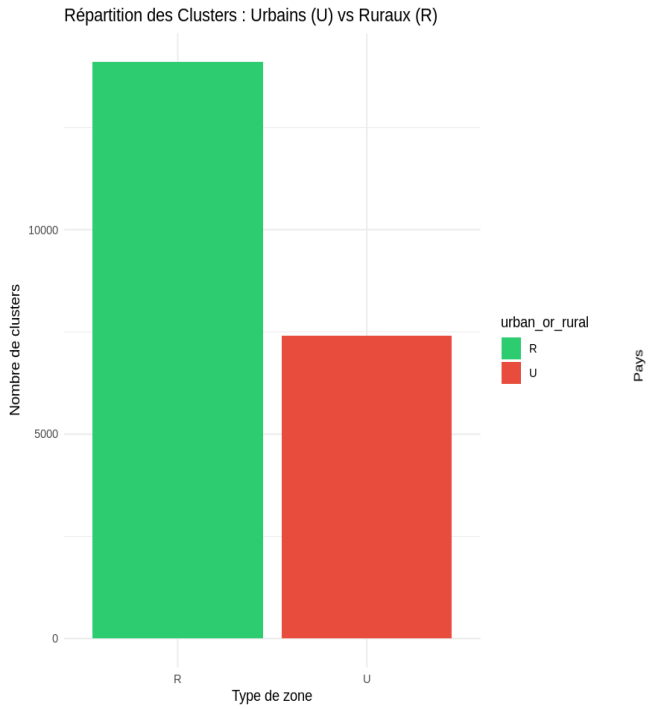
Dans un premier temps, des statistiques descriptives (moyenne, médiane, écart-type, minimum et maximum) ont été calculées pour l'ensemble des variables numériques. Les résultats montrent une forte hétérogénéité entre les différentes zones géographiques, notamment pour des variables telles que la densité de population, les lumières nocturnes et la proportion de zones urbaines.

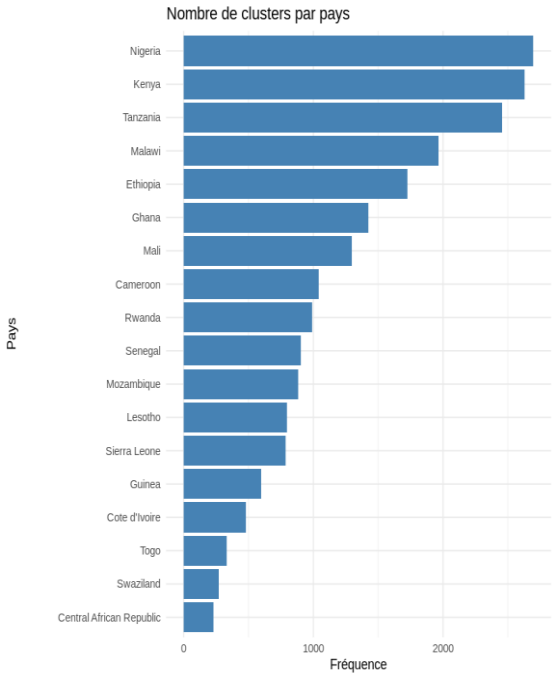




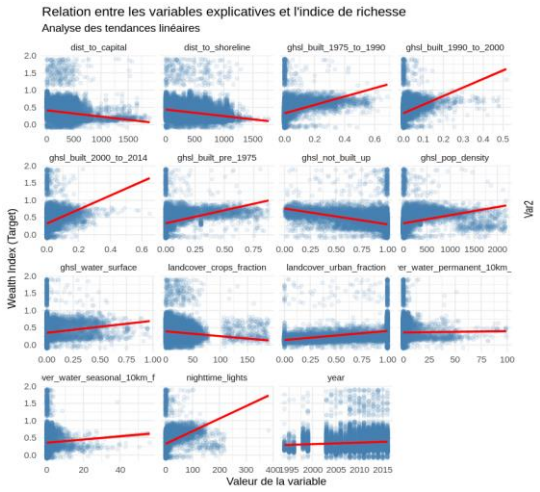
L'étude des distributions met en évidence que l'indice de richesse n'est pas uniformément réparti : la majorité des observations se situe dans des valeurs faibles à intermédiaires, ce qui reflète des disparités économiques importantes entre les régions étudiées.

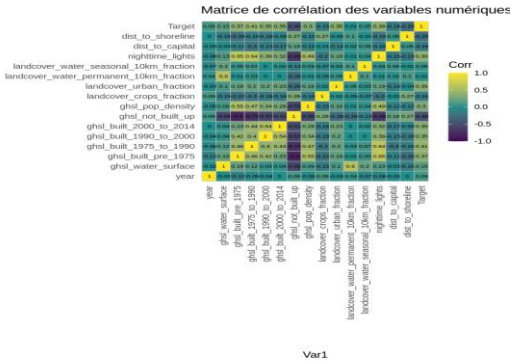
les données décrivent une réalité majoritairement **rurale** et sont portées par des géants démographiques tels que le **Nigeria**. Pour une analyse globale, il sera crucial de prendre en compte ce biais rural ainsi que le poids prépondérant des pays d'Afrique de l'Est et de l'Ouest dans les résultats finaux.





Ensuite, une analyse de corrélation a été effectuée afin d'évaluer les relations entre les variables explicatives et la variable cible. Les résultats indiquent que certaines variables présentent une corrélation positive notable avec l'indice de richesse, en particulier :

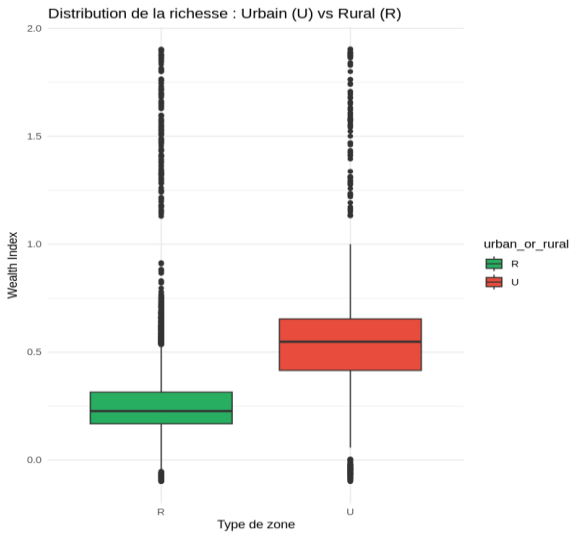




Les variables les plus corrélées positivement avec la richesse sont les indicateurs d'urbanisation (ghsl_built_*, landcover_urban_fraction), la densité de population (ghsl_pop_density) et les lumières nocturnes (nighttime_lights).

À l'inverse, la proportion de zones non construites (ghsl_not_built_up), la distance à la capitale et la distance au littoral sont négativement corrélées avec l'indice de richesse.

Ces résultats confirment que les régions urbaines, proches des centres économiques et plus densément peuplées sont globalement plus riches que les zones rurales et isolées.



Bien que l'étude porte majoritairement sur le milieu rural en termes de volume (plus de clusters), les données révèlent que ces zones sont structurellement plus pauvres que les pôles urbains. On observe un **décalage de richesse systématique** : la classe supérieure rurale atteint à peine le niveau de richesse moyen de la classe urbaine

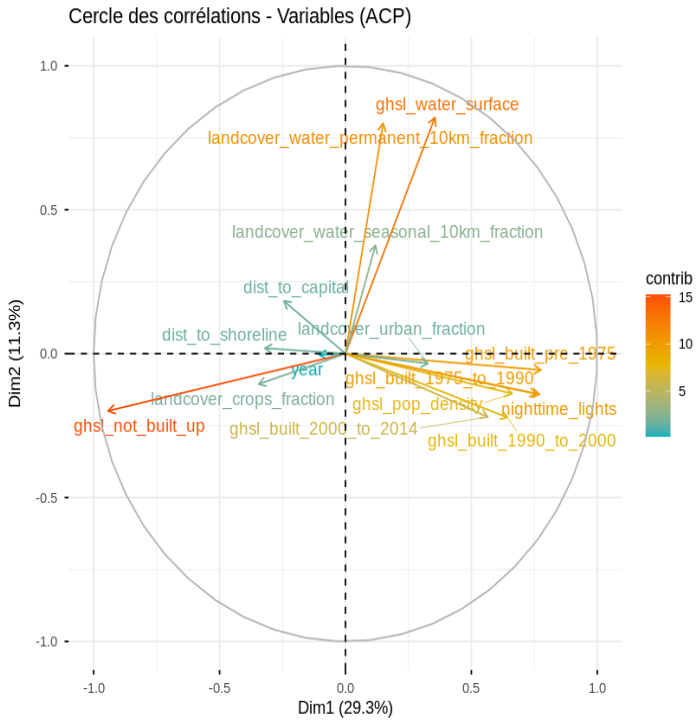
Apprentissage non supervisé :

L'apprentissage non supervisé a été utilisé afin d'explorer la structure intrinsèque des données sans utiliser la variable cible. L'objectif principal est d'identifier des groupes homogènes d'observations partageant des caractéristiques similaires, et de mettre en évidence des profils socio-économiques distincts.

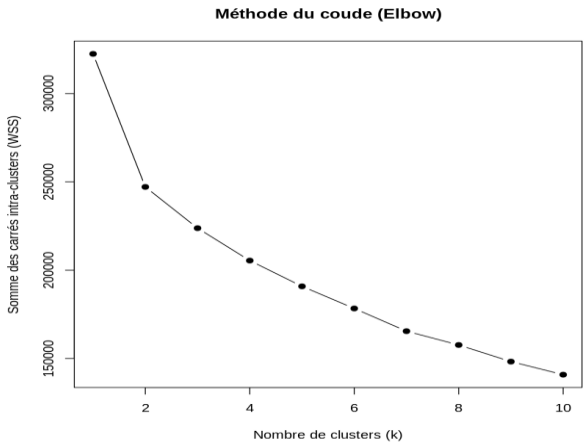
Dans un premier temps, les variables numériques ont été standardisées afin d'éviter que les différences d'échelle n'influencent les résultats des algorithmes de clustering. Une analyse en composantes principales (ACP / PCA) a ensuite été appliquée pour réduire la dimensionnalité du jeu de données et faciliter la visualisation des observations dans un espace à deux dimensions.

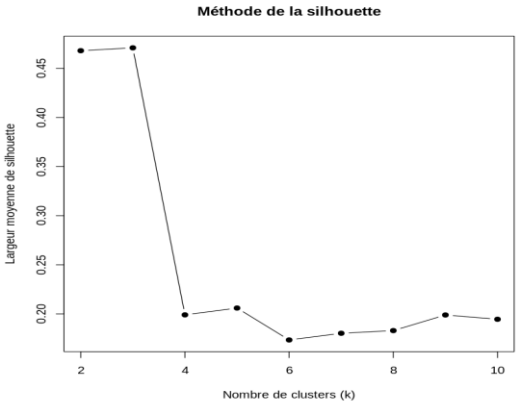
Les valeurs manquantes ont été traitées par imputation à l'aide de la médiane pour les variables numériques, afin de conserver l'ensemble des observations et éviter une perte d'information. Les variables catégorielles ont été imputées par leur modalité la plus fréquente.

Les 2 premières composantes principales expliquent une part importante de la variance totale environ de **40.6%,** ce qui indique que les variables originales peuvent être résumées efficacement par un nombre réduit de dimensions. La projection des données dans cet espace met en évidence une séparation partielle entre différentes catégories d'observations.



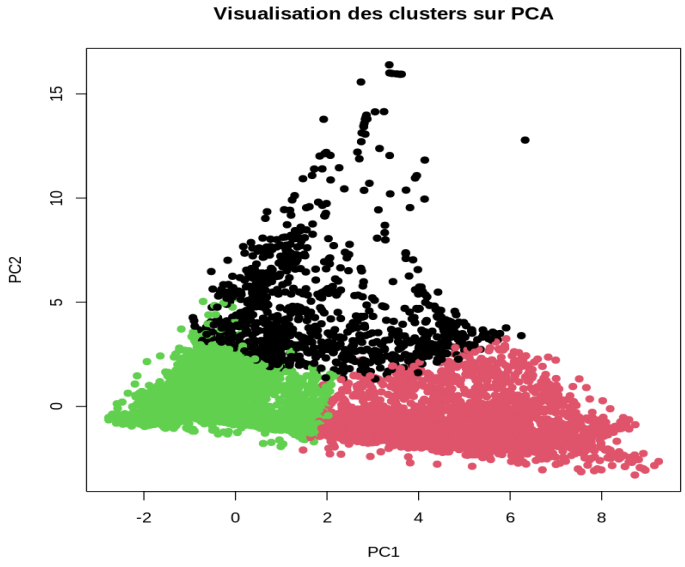
Ensuite, un algorithme de clustering de type **K-means** a été appliqué afin de regrouper les observations en plusieurs clusters. Le nombre optimal de clusters a été déterminé à l'aide de méthodes telles que la courbe du coude (Elbow method) et le coefficient de silhouette, et on obtient k=3.

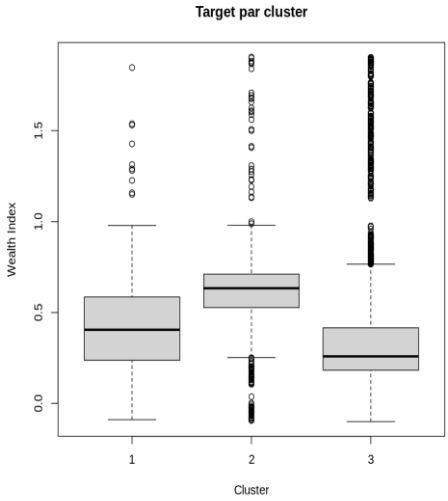




Les résultats montrent l'existence de plusieurs groupes distincts correspondant à différents profils socio-économiques :

1. un cluster caractérisé par de faibles valeurs de lumières nocturnes, une faible urbanisation et une grande distance aux centres économiques, représentant des zones majoritairement pauvres
2. un cluster intermédiaire présentant des valeurs moyennes pour la plupart des variables
3. un cluster regroupant des zones fortement urbanisées, proches des capitales et présentant des niveaux élevés de lumières nocturnes, correspondant à des zones plus riches.





Cette analyse non supervisée permet d'obtenir une première segmentation des données et fournit des informations utiles pour la suite du projet, notamment pour la création de nouvelles variables explicatives dans la phase de feature engineering.

Feature Engineering :

À partir des résultats obtenus lors de l'analyse non supervisée, une étape de *feature engineering* a été réalisée afin d'enrichir le jeu de données et d'améliorer les performances des modèles supervisés.

La principale idée consistait à exploiter les informations issues du clustering pour créer de nouvelles variables explicatives. En particulier, la variable correspondant au cluster d'appartenance de chaque observation a été ajoutée au jeu de données comme une nouvelle caractéristique. Cette variable permet de résumer le profil socio-économique général d'une zone géographique en une seule information.

De plus, les 3 composantes principales obtenues par l'Analyse en Composantes Principales (ACP) ont été utilisées comme variables supplémentaires afin de capturer les relations linéaires entre les variables originales tout en réduisant la dimensionnalité du problème.

Ces nouvelles variables ont été combinées avec les variables initiales afin de constituer un jeu de données enrichi. Cette approche permet aux modèles supervisés de bénéficier à la fois des informations originales et des structures latentes identifiées par l'apprentissage non supervisé.

L'intérêt de cette démarche est double : d'une part, elle améliore la capacité prédictive des modèles en introduisant des variables synthétiques informatives, et d'autre part, elle facilite l'interprétation des résultats en associant chaque observation à un groupe homogène clairement identifiable.

Cette étape constitue un lien essentiel entre l'analyse exploratoire, l'apprentissage non supervisé et l'apprentissage supervisé, et elle permet d'optimiser les performances des modèles construits par la suite.

Apprentissage supervisé : prédiction de l'indice de richesse(regression)

Dans cette partie, l'objectif est de prédire l'indice de richesse en tant que variable continue à l'aide de méthodes d'apprentissage supervisé. Le jeu de données a été divisé en un ensemble d'apprentissage et un ensemble de test, en veillant à ce que les observations provenant des pays Ghana, Kenya et Nigeria soient incluses dans l'échantillon de test, conformément aux consignes du projet.

Plusieurs modèles de régression ont été entraînés afin de comparer leurs performances. Parmi les modèles utilisés figurent notamment la régression linéaire, Random Forest et KNN.

Les performances des modèles ont été évaluées à l'aide de métriques adaptées à un problème de régression, notamment :

- l'erreur quadratique moyenne (RMSE)
- le coefficient de détermination (R^2).

Model	RMSE	R^2
Random Forest	0.18422	43%
KNN	0.20233	33%
régression linéaire	0.18612	39%

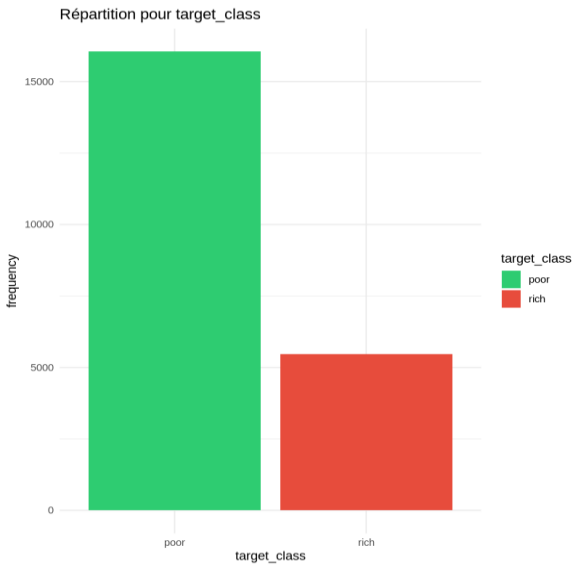
Les résultats montrent que les modèles non linéaires, en particulier la Random Forest, offrent de meilleures performances que les modèles linéaires. Cela s'explique par leur capacité à capturer des relations complexes entre les variables explicatives et la variable cible.
L'intégration des nouvelles variables issues du feature engineering a permis d'améliorer légèrement les résultats, ce qui confirme la pertinence de l'analyse non supervisée réalisée précédemment. Le modèle final de régression sélectionné présente un compromis satisfaisant entre précision et robustesse.

Définition de la variable cible binaire :

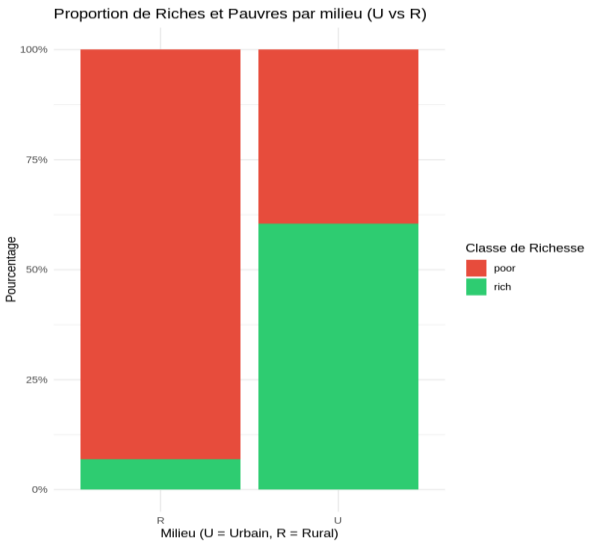
Afin de réaliser une analyse de classification, une nouvelle variable cible binaire a été construite à partir de l'indice de richesse continu. Cette variable permet de distinguer deux catégories socio-économiques : les zones considérées comme riches et les zones considérées comme pauvres.
La règle de classification adoptée est la suivante :

- une observation est classée comme **Rich (riche)** si son indice de richesse est supérieur ou égal à 0.5
- une observation est classée comme **Poor (pauvre)** si son indice de richesse est inférieur à 0.5.

Cette transformation permet de reformuler le problème initial de régression en un problème de classification binaire, plus adapté à certaines applications décisionnelles.



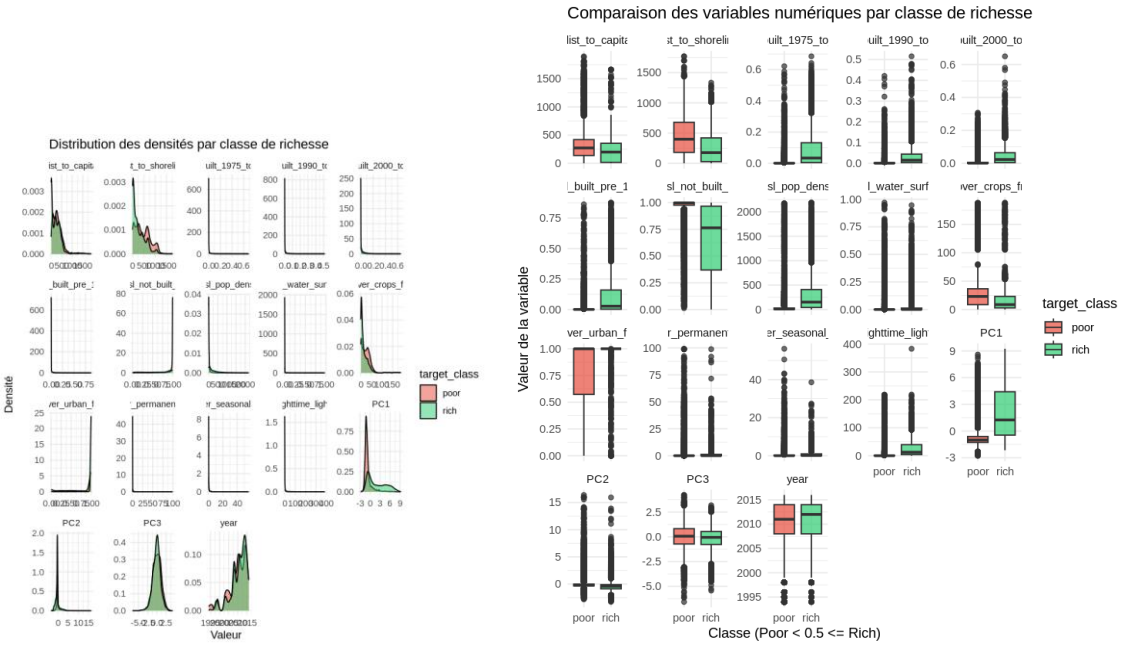
Une analyse de la distribution de cette nouvelle variable cible a été réalisée afin d'évaluer l'équilibre entre les deux classes. Les résultats montrent que les deux catégories ne sont pas parfaitement équilibrées, avec une proportion plus importante de zones classées comme pauvres. Ce déséquilibre doit être pris en compte lors de l'entraînement des modèles afin d'éviter un biais en faveur de la classe majoritaire.



Des visualisations graphiques ont été utilisées pour représenter la répartition des classes selon certaines variables explicatives, notamment le type de zone (urbaine ou rurale). Il apparaît que les zones urbaines sont majoritairement classées comme riches, tandis que les zones rurales sont principalement associées à la classe pauvre.

Cette nouvelle variable cible constitue la base de la seconde phase d'apprentissage supervisé, consacrée à la classification des zones géographiques selon leur niveau de richesse.

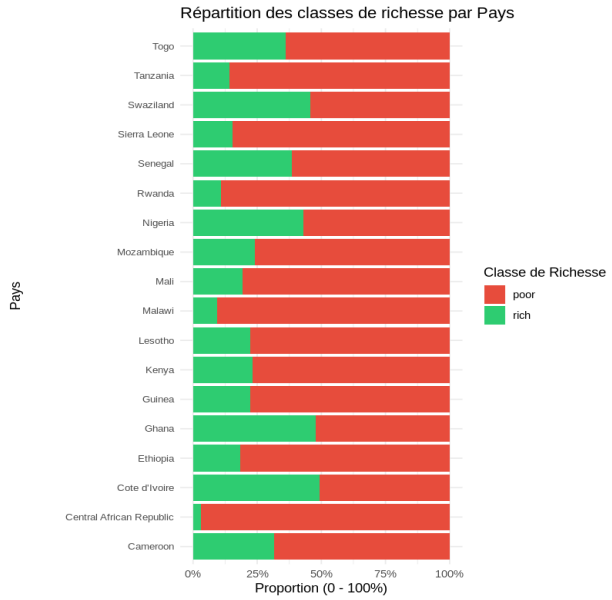
Analyse descriptive (EDA) pour le variable binaire vs autres variables :



L'analyse croisée des données révèle une fracture socio-économique profonde entre les milieux urbains et ruraux, ainsi qu'entre les classes de richesse définies. La base de données est majoritairement composée de clusters ruraux, lesquels affichent un indice de richesse nettement inférieur et plus homogène que leurs homologues urbains. Cette pauvreté rurale est étroitement corrélée à un éloignement géographique des centres décisionnels et des

côtes, ainsi qu'à une dépendance marquée aux activités agricoles. À l'inverse, les clusters classés comme "riches" se concentrent dans les zones urbaines denses, bénéficiant d'infrastructures récentes et d'un accès facilité aux ressources. L'analyse par composantes principales confirme cette séparation, la variable PC1 agissant comme un indicateur discriminant qui isole presque parfaitement les populations aisées des populations précaires selon leurs conditions de vie et leur environnement bâti.

L'examen du graphique révèle des disparités importantes entre les pays étudiés. La République Centrafricaine présente la situation la plus critique avec une population presque entièrement classée comme pauvre. À l'opposé, le Ghana et la Côte d'Ivoire affichent les taux de richesse les plus élevés, avec environ 40 à 45% de leur population classée comme riche.



Le Swaziland et le Nigeria se distinguent également avec des proportions notables de population riche, dépassant les 30%. Le Cameroun, le Sénégal et le Kenya présentent des situations intermédiaires avec environ 20 à 25% de population riche.

En revanche, la majorité des pays tels que la Tanzanie, le Rwanda, le Malawi, le Mali, la Sierra Leone et le Togo montrent des taux de pauvreté dépassant 80%, reflétant des défis économiques considérables.

Apprentissage supervisé : classification des zones riches et pauvres

Dans cette partie, le problème est formulé comme une tâche de classification binaire visant à prédire si une zone est classée comme riche ou pauvre à partir des variables explicatives disponibles. La variable cible binaire présente un déséquilibre entre les deux classes, avec une proportion plus importante d'observations appartenant à la classe *Poor* par rapport à la classe *Rich*.

Afin de corriger ce problème de déséquilibre et d'éviter que les modèles ne soient biaisés en faveur de la classe majoritaire, la technique **SMOTE (Synthetic Minority Over-sampling Technique)** a été appliquée sur l'ensemble d'apprentissage. Cette méthode permet de générer artificiellement de nouvelles observations pour la classe minoritaire en se basant sur les voisins les plus proches, ce qui conduit à un jeu de données plus équilibré.

Plusieurs modèles de classification ont ensuite été entraînés sur les données rééquilibrées, notamment :

- la régression logistique
- Random Forest

Les performances des modèles ont été évaluées à l'aide de métriques adaptées aux problèmes de classification, telles que :

- l'accuracy
- la précision (precision)
- ROC AUC

Les résultats obtenus montrent que l'utilisation de SMOTE améliore significativement la capacité des modèles à détecter correctement la classe minoritaire, en particulier en augmentant le rappel et le score F1. Parmi les différents modèles testés, la Random Forest s'est distinguée par ses meilleures performances globales, ce qui s'explique par sa capacité à modéliser des relations complexes et non linéaires entre les variables explicatives et la variable cible.

L'analyse des matrices de confusion révèle que la majorité des erreurs de classification concerne des observations dont l'indice de richesse est proche du seuil de décision (0.5), ce qui traduit une zone de transition entre les catégories riches et pauvres.

L'intégration de la technique SMOTE combinée aux variables issues du feature engineering a permis d'améliorer la robustesse et la performance des modèles de classification. Cette approche garantit une meilleure prise en compte des deux classes et rend les résultats plus fiables pour une utilisation pratique.

Optimisation des hyperparamètres et sélection du modèle final :

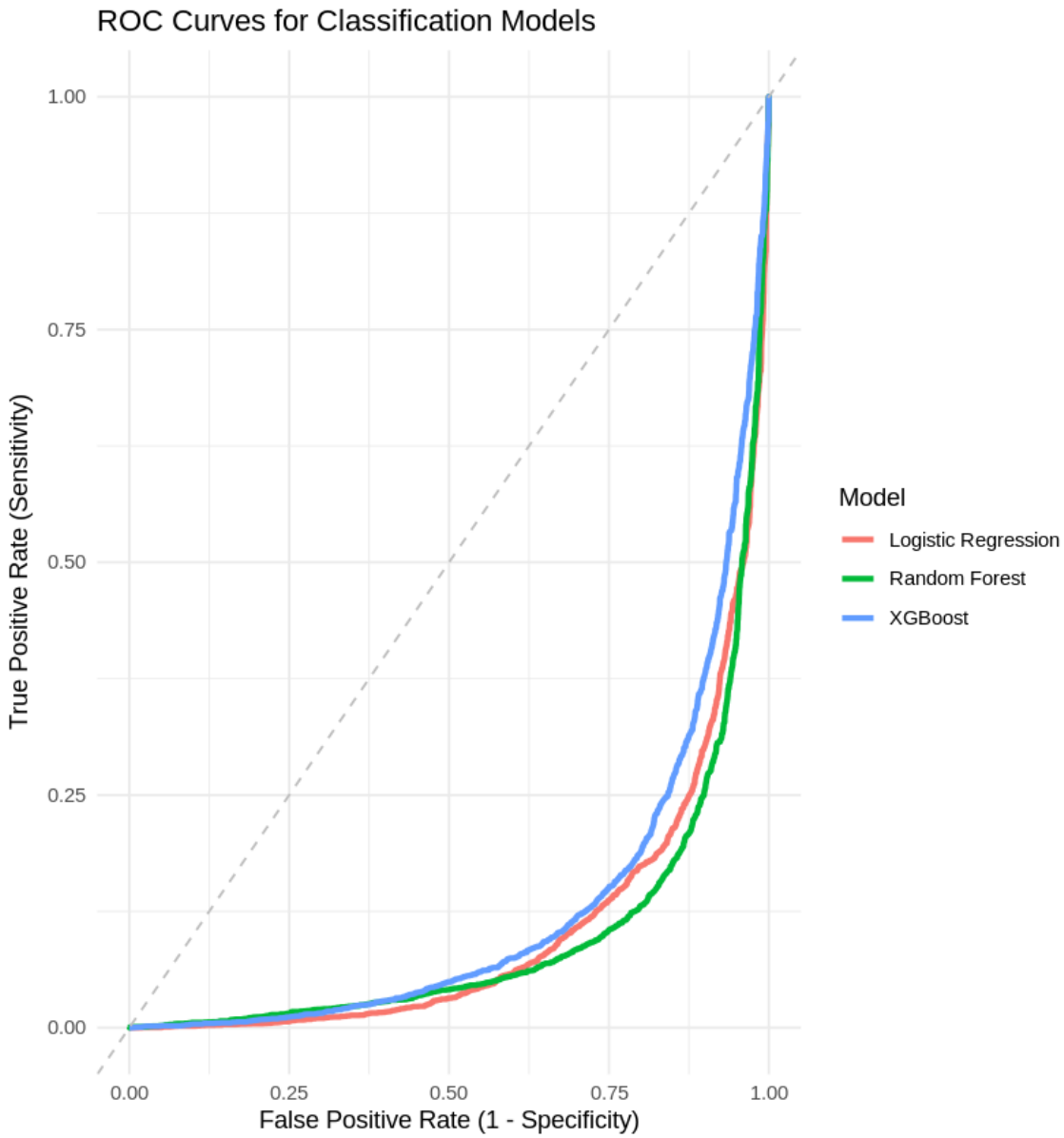
Afin d'améliorer les performances des modèles d'apprentissage supervisé, une étape d'optimisation des hyperparamètres (*hyperparameter tuning*) a été réalisée sur au moins un modèle, conformément aux exigences du projet. Cette optimisation vise à identifier la combinaison de paramètres offrant les meilleurs résultats de prédiction.

Un modèle XGBoost de classification a été entraîné pour prédire la variable binaire *rich/poor*. Le prétraitement des données a été effectué à l'aide du package `recipes` (imputation, normalisation, encodage des variables catégorielles). Afin de corriger le déséquilibre des classes, la méthode SMOTE a été appliquée sur l'ensemble d'apprentissage. Les hyperparamètres du modèle ont été optimisés par validation croisée à 5 plis. Les performances ont été évaluées sur les pays Ghana, Kenya et Nigeria à l'aide des métriques Accuracy, ROC AUC

Les resultat obtenue est :

model	Accuracy	ROC AUC
Logistic Regression	81%	0.10

Random Forest	83%	0.09
XGBoost tuned	81%	0.12



Conclusion :

Ce projet avait pour objectif d'estimer le niveau de richesse des ménages en Afrique à partir de données géospatiales accessibles à grande échelle. À travers une analyse exploratoire approfondie, nous avons mis en évidence des relations significatives entre l'indice de richesse et plusieurs variables explicatives, notamment la luminosité nocturne, le degré d'urbanisation, la densité de population et la distance aux centres économiques.

L'analyse non supervisée a permis d'identifier des structures sous-jacentes dans les données et de mieux comprendre la segmentation des zones géographiques selon leurs caractéristiques socio-économiques. Ces résultats ont également contribué à enrichir le jeu de données via la création de nouvelles variables pertinentes.

Dans la partie supervisée, plusieurs modèles de régression ont été testés afin de prédire l'indice de richesse continu, ainsi que des modèles de classification pour distinguer les zones riches et pauvres. Les modèles basés sur des méthodes d'ensemble, notamment après une phase de tuning des hyperparamètres, ont montré de meilleures performances par rapport aux approches plus simples, soulignant leur capacité à capturer des relations non linéaires complexes.

Les résultats obtenus confirment que les données géospatiales peuvent constituer une alternative crédible aux enquêtes traditionnelles coûteuses pour l'estimation du bien-être économique, en particulier dans des contextes où les données sont rares ou peu fréquentes. Toutefois, certaines limites subsistent, notamment l'agrégation spatiale des données et l'absence d'une dimension temporelle.

En perspective, l'intégration de données supplémentaires, l'analyse temporelle et l'exploration de modèles plus avancés pourraient permettre d'améliorer encore la précision et la robustesse des prédictions. Ce travail ouvre ainsi la voie à des applications concrètes pour l'aide à la décision des gouvernements et des organisations internationales.