Grégoire Sarsat

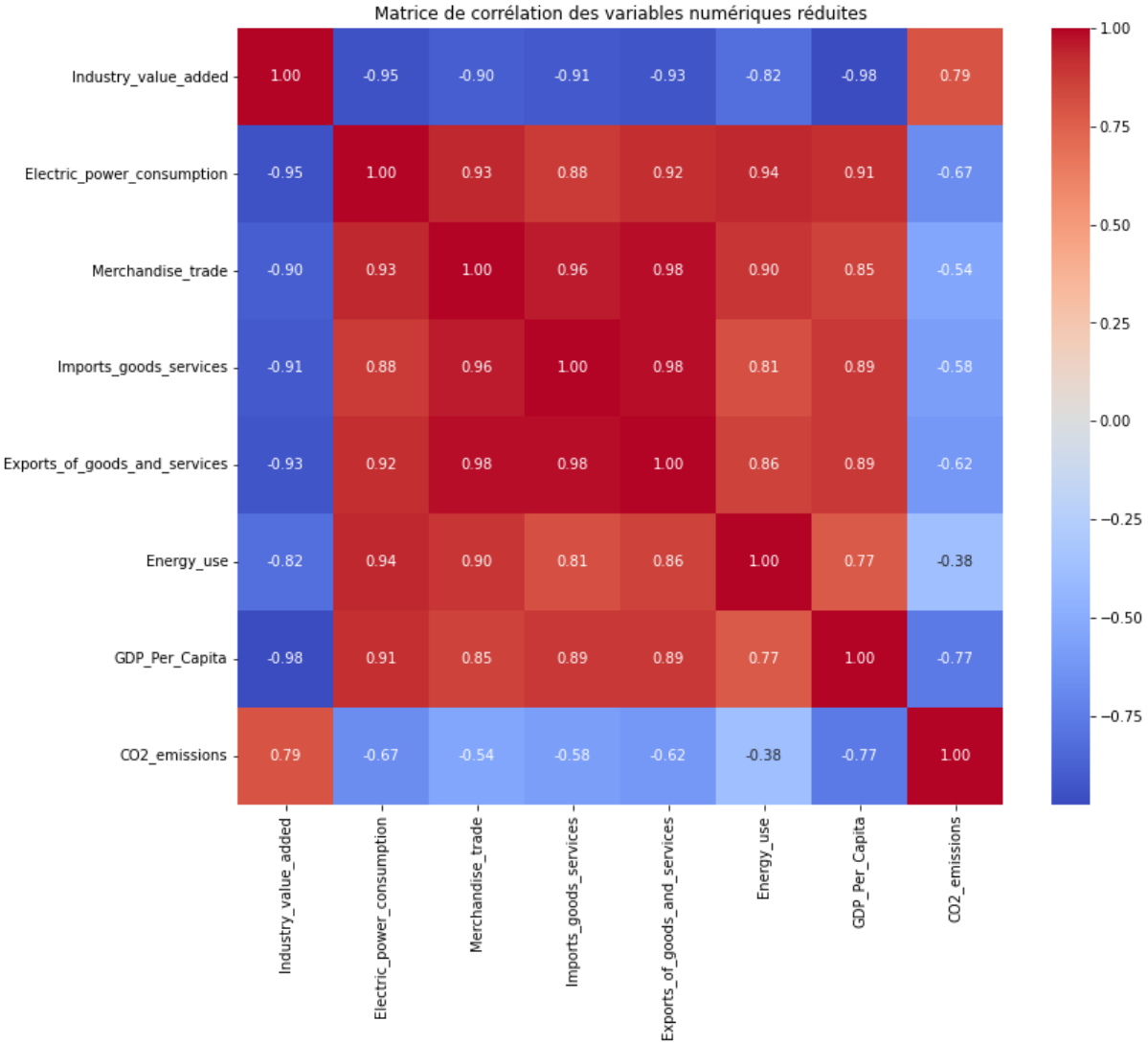
EMISSIONS DE CO2 - UNE ANALYSE MULTIVARIÉE

PROJET DE MACHINE LEARNING

Ce projet vise à analyser les facteurs influençant les émissions de CO2 en France à l'aide de modèles de machine learning. En se basant sur des données économiques et environnementales, l'objectif est de comprendre comment des variables comme l'industrie, la consommation d'énergie, ou le PIB par habitant expliquent ces émissions. Cette analyse permet d'identifier les principaux facteurs influençant les émissions de CO2.

Traitement des données

Les données ont été nettoyées et préparées afin de garantir leur qualité et leur fiabilité pour l'analyse. Les valeurs manquantes ont été interpolées linéairement pour minimiser les biais, et préserver des tendances. Les colonnes non numériques ont été converties en données exploitables. La variable GDP_Per_Capita a été calculée manuellement comme le ratio entre le GDP et la Population. Elle remplace la variable Gross_National_Income_per_capita, car celle-ci contenait des valeurs aberrantes. Enfin, seules les variables pertinentes pour l'étude ont été conservées, réduisant le risque de bruit dans les modèles et améliorant leur performance. Ce traitement assure une base solide pour les analyses multifactorielles et prédictives.



En observant la matrice de corrélation des variables numériques on remarque que Imports_goods_services, Merchandise_trade, et Exports_of_goods_and_services présentent

des corrélations très fortes entre elles (corr > 0,95). On décide de garder uniquement les exportations car elles semblent représenter les deux autres de manière équivalente, cela simplifie l'analyse et réduit le risque de colinéarité. En conservant uniquement Exports_of_goods_and_services, nous représentons efficacement l'effet des échanges commerciaux sans multiplier les variables fortement corrélées.

Analyse des Corrélations

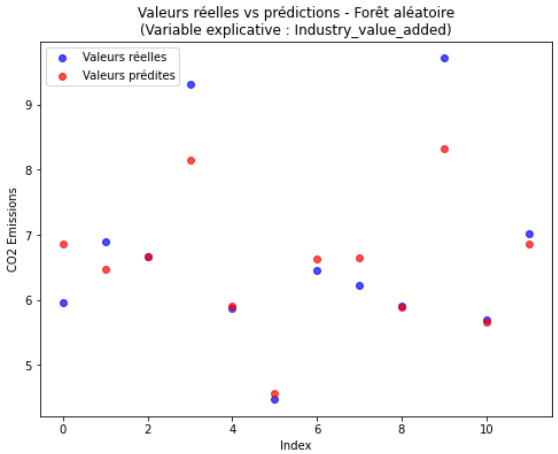
Industry_value_added et CO2_emissions : Corrélation forte et positive (0,79), cela indique que l'augmentation de la valeur ajoutée de l'industrie est associée à une augmentation des émissions de CO2. Hypothèse : L'industrie reste un secteur clé en matière d'émissions.

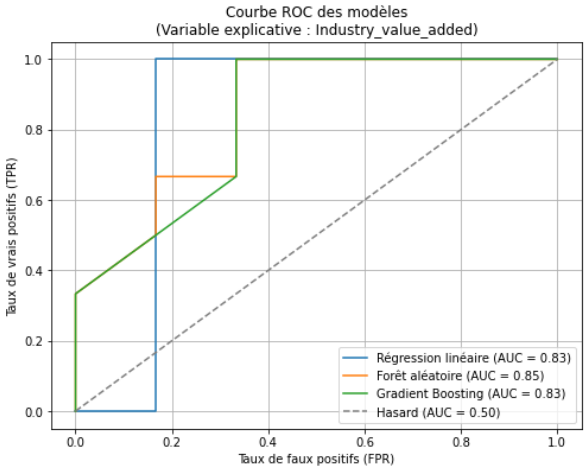
Electric_power_consumption et CO2_emissions : Corrélation négative (-0,67), surprenante car cela peut indiquer que certaines sources d'énergie électrique (comme le nucléaire ou les énergies renouvelables) sont moins émettrices en CO2. Cependant, d'autres facteurs non mesurés pourraient influencer ce lien, comme l'efficacité des infrastructures énergétiques. Une analyse plus approfondie pourrait par exemple révéler si ce lien dépend du mix énergétique utilisé.

Energy_use : Corrélation modérée (-0,38), cela pourrait indiquer que l'augmentation de l'efficacité énergétique a permis de limiter les émissions, même en cas de hausse de la consommation énergétique.

GDP_Per_Capita : Corrélation négative (-0,77), cela pourrait s'expliquer par le fait que les pays avec un PIB par habitant plus élevé investissent davantage dans des technologies propres et des mesures environnementales.

Industry_value_added - Valeur ajoutée du secteur industriel (% du GDP)

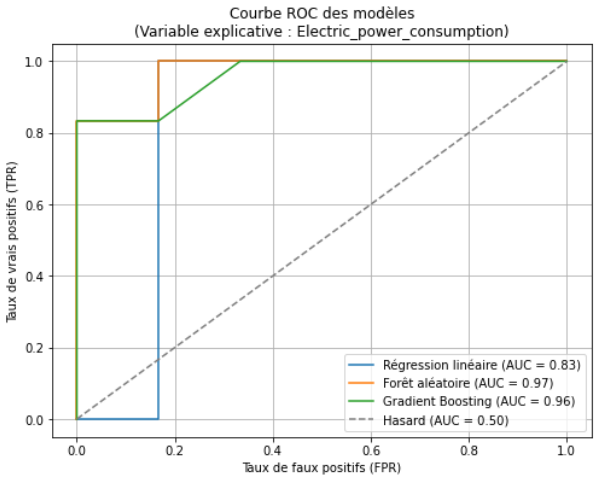




La régression linéaire est moins efficace pour prédire CO2_emissions à partir d'Industry_value_added, elle affiche un MSE de 1.42 et un R^2 de 0,30, La forêt aléatoire obtient un MSE de 0,38 et un R^2 de 0,81 et est le modèle le plus performant, suivie de près par le Gradient Boosting. Ces deux modèles sont adaptés pour capturer des relations non linéaires dans les données.

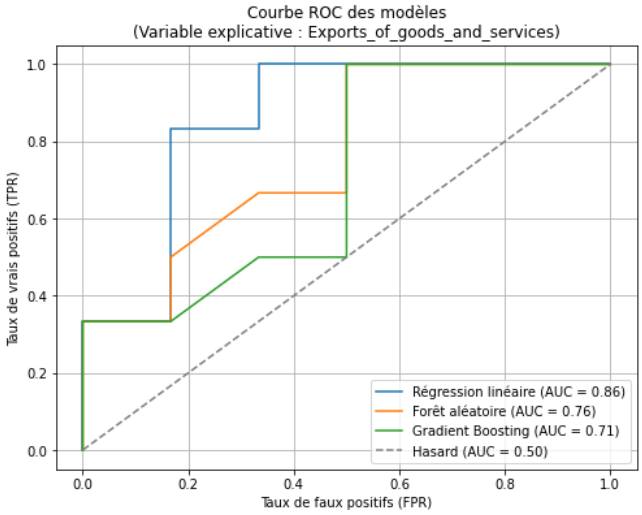
Analyse des résultats pour la variable explicative : Electric_power_consumption

La régression linéaire affiche un MSE de 1.70 et un R^2 de 0,16 montre une performance limitée et n'est pas recommandée pour cette variable explicative. La forêt aléatoire avec un MSE de 0,37 et un R^2 de 0,82 est le meilleur modèle pour prédire CO2_emissions à partir de Electric_power_consumption, suivi par le Gradient Boosting, qui reste une bonne alternative. Ces modèles sont adaptés pour des relations complexes ou non linéaires.



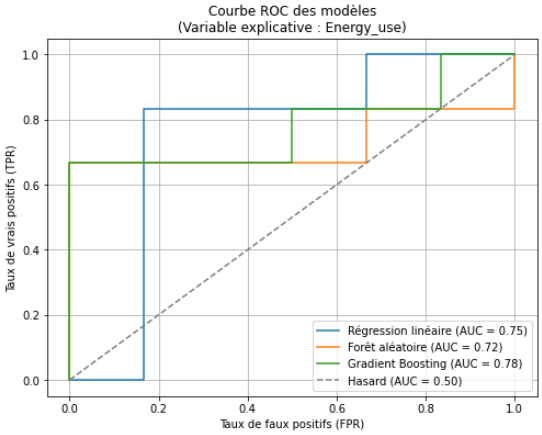
Analyse des résultats pour la variable explicative : Exports_of_goods_and_services

La régression linéaire est la moins performante des trois modèles, avec un faible pouvoir explicatif. La forêt aléatoire est le modèle le plus adapté pour cette variable avec un MSE de 0,86 et un R^2 de 0,57, bien qu'elle n'atteigne pas des performances optimales. Le Gradient Boosting, affiche un MSE de 1.03 et un R^2 de 0,49 légèrement moins performant, reste une alternative valable. Les résultats globaux montrent que cette variable explicative est moins prédictive de CO2_emissions comparée à d'autres variables comme Energy_use ou Electric_power_consumption.



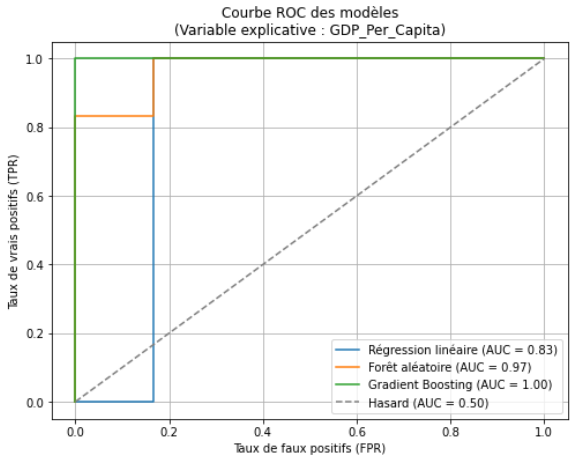
Analyse des résultats pour la variable explicative : Energy_use

La régression linéaire est inadaptée pour prédire CO2_emissions à partir de Energy_use. La forêt aléatoire est le modèle le plus performant, suivi de très près par le Gradient Boosting affiche un MSE de 0,60 et un R^2 de 0,70, qui offre une alternative viable. Ces modèles sont bien adaptés pour des relations complexes et non linéaires.

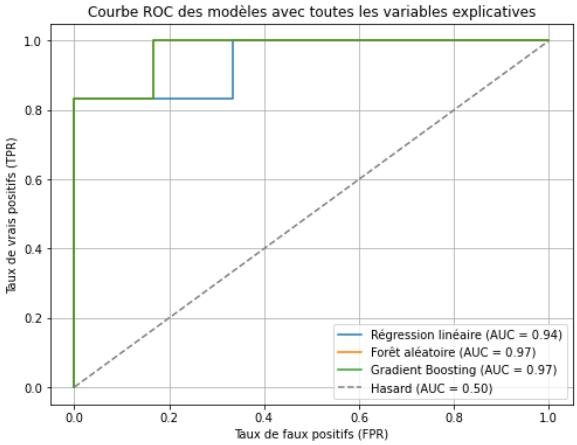


Analyse des résultats pour la variable explicative : GDP_Per_Capita

La régression linéaire est limitée dans sa capacité à modéliser la relation entre GDP_Per_Capita et CO2_emissions. La forêt aléatoire est le modèle le plus performant avec un MSE de 0,29 et un R^2 de 0,86, suivi de près par le Gradient Boosting. Ces deux modèles sont bien adaptés pour capturer des relations complexes et fournir des prédictions précises. Le Gradient Boosting, avec un AUC parfait (1.00), montre une performance exceptionnelle pour distinguer les niveaux d'émissions dans ce contexte.



Enfin l'analyse mutlivariée cdonne un modèle gradient Boosting avec un MSE de 0,05 et un R^2 de 0,98, ce qui explique plus que tous les modèles individuels



Pour conclure, l'industrie est le principal facteur explicatif des émissions de CO2 en France. L'augmentation de la contribution industrielle au PIB s'accompagne d'une augmentation significative des émissions.

Le mix énergétique influence fortement les émissions, comme le montre le rôle de Electric_power_consumption. Une consommation électrique élevée avec un mix énergétique propre (nucléaire, renouvelables) pourrait limiter les émissions, tandis qu'un mix dépendant des combustibles fossiles augmenterait les émissions.

Le PIB par habitant agit comme un facteur de réduction des émissions, probablement en raison de l'adoption de technologies vertes et des politiques environnementales dans les économies avancées.

L'efficacité énergétique joue un rôle modéré mais non négligeable dans la réduction des émissions, notamment en atténuant les effets d'une consommation énergétique croissante.

Les exportations un impact limité sur les émissions, probablement parce que leur effet est indirect, dépendant des secteurs industriels ou énergétiques sous-jacents.