

Machine Learning pour la Prédiction de la Qualité de l'Air

1. Collecte de Données

Les données proviennent de l'article Deciphering Environmental Air Pollution with Large Scale City Data publié dans IJCAI 2022. Ce dataset combine les informations sur les principaux polluants atmosphériques, le trafic urbain, les émissions des industries de production d'électricité, et les données météorologiques pour plus de 50 villes aux États-Unis sur deux ans.

Colonnes Pertinentes :

- **Date** : Correspond à la date de l'échantillon collecté. Cette colonne est essentielle pour l'analyse temporelle des données.
- **Ville** : Indique la ville où l'échantillon a été collecté, permettant des comparaisons géographiques de la qualité de l'air.
- **X_median** : Représente la valeur médiane du jour pour le polluant ou la caractéristique météorologique X. La médiane est utilisée pour minimiser l'impact des valeurs extrêmes et fournir une représentation plus fiable des conditions typiques.
- **mil_miles** : Distance totale parcourue par les véhicules pour l'échantillon donné, utilisée comme indicateur de la pollution due au trafic.
- **pp_feat** : Caractéristique calculée représentant l'influence des centrales électriques voisines sur la pollution de l'air.
- **Population Staying at Home** : Mesure utilisée pour estimer les émissions domestiques, sous-entendant que plus de personnes à la maison pourrait signifier une augmentation de certaines formes de pollution domestique.

Polluants Mesurés :

- **PM2.5** : Particules fines dont le diamètre est inférieur à 2.5 micromètres. Elles peuvent pénétrer profondément dans le système respiratoire.
- **PM10** : Particules dont le diamètre est inférieur à 10 micromètres. Elles sont également nocives mais moins pénétrantes que les PM2.5.
- **NO2 (Dioxyde d'azote)** : Gaz irritant produit par la combustion des carburants. Il est un indicateur clé de la pollution automobile.
- **O3 (Ozone)** : Gaz qui se forme sous l'effet des rayons UV sur certains polluants chimiques. L'ozone au niveau du sol est un polluant majeur qui peut causer divers problèmes respiratoires.
- **CO (Monoxyde de carbone)** : Gaz incolore et inodore produit par la combustion incomplète de carbone. Il est particulièrement dangereux en raison de sa capacité à se lier à l'hémoglobine, réduisant ainsi la capacité du sang à transporter l'oxygène.
- **SO2 (Dioxyde de soufre)** : Gaz produit par la combustion du soufre, souvent issu des centrales électriques et des processus industriels.

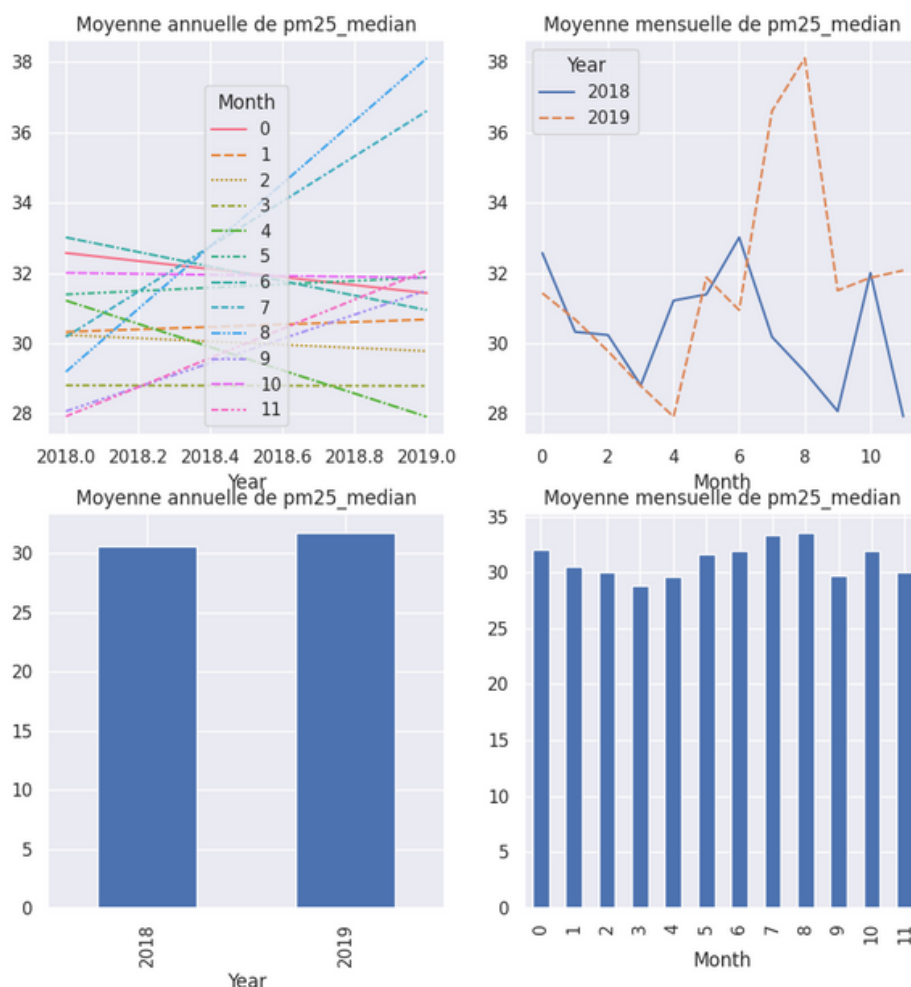
Caractéristiques Météorologiques :

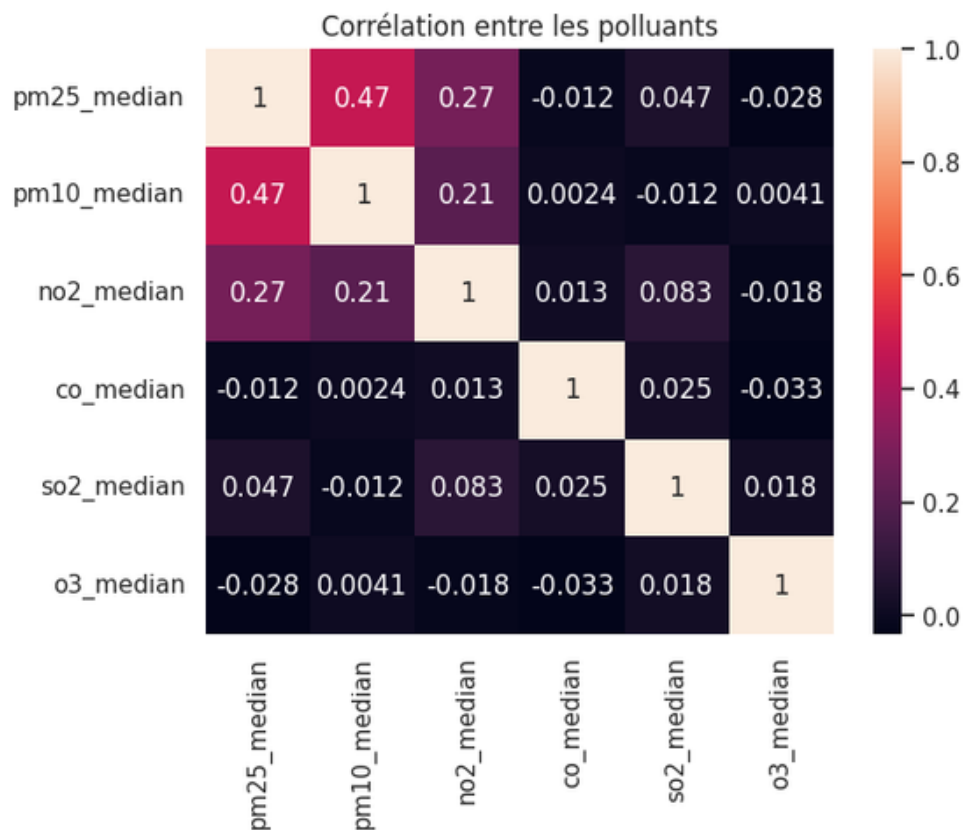
- **Température:** Mesure de la chaleur ou du froid, exprimée en degrés Celsius ou Fahrenheit.
- **Pression :** Pression atmosphérique mesurée en hectopascals (hPa). Elle peut influencer la dispersion des polluants.
- **Humidité :** Pourcentage d'humidité dans l'air, influençant la concentration de certains polluants et le confort humain.
- **Point de rosée:** Température à laquelle l'air doit être refroidi pour atteindre la saturation en vapeur d'eau, reflétant la quantité d'humidité dans l'air.
- **Vitesse du vent :** Vitesse du vent mesurée en kilomètres par heure, qui peut disperser ou accumuler des polluants atmosphériques.

Ces données sont essentielles pour comprendre les facteurs contribuant à la pollution de l'air et pour développer des modèles prédictifs capables d'estimer l'Indice de Qualité de l'Air (AQI) en fonction des variations des concentrations des polluants et des conditions météorologiques. Ces informations servent également à évaluer l'efficacité des politiques environnementales et à sensibiliser le public sur les risques liés à la pollution de l'air.

2. Analyse Exploratoire des Données (AED)

Les graphiques montrent une augmentation notable des niveaux moyens annuels de PM2.5 de 2018 à 2019, avec une variabilité mensuelle élevée en 2019, incluant une diminution significative en fin d'année.





Interpretation des Resultats:

- **PM2.5 et PM10** ont une corrélation de 0.71, ce qui est assez élevé, indiquant que les niveaux de ces particules fines tendent à augmenter ou diminuer ensemble. Cela est logique car les PM2.5 sont souvent une sous-catégorie des PM10, donc les sources d'émission pour ces deux types de particules sont souvent similaires
- **NO2** présente une corrélation modérée avec PM2.5 et PM10 (0.33 et 0.32 respectivement), ce qui pourrait suggérer que lorsque les niveaux de NO2 augmentent, il y a une tendance modérée pour l'augmentation des particules fines également.
- **CO** présente très peu ou pas de corrélation avec PM2.5, PM10 et NO2, ce qui indique que les variations dans les niveaux de CO ne sont pas directement liées à des changements dans les niveaux de ces autres polluants.
- **SO2** a une très faible corrélation positive avec NO2 (0.13), c'est une indication que leurs niveaux ne sont pas fortement liés.
- **O3, ou ozone**, a des valeurs de corrélation faibles ou négatives avec tous les autres polluants, suggérant qu'il n'y a pas de tendance forte pour que les niveaux d'ozone se déplacent en tandem avec les autres polluants mesurés.

3. Prétraitement des Données

Les étapes suivantes ont été prises pour préparer les données à la modélisation :

- Suppression des colonnes non pertinentes contenant des informations redondantes ou spécifiques comme le nom des comtés.
- Imputation des valeurs manquantes en utilisant la moyenne pour chaque polluant.
- Transformation des colonnes de dates pour en extraire le jour de la semaine, le mois et l'année pour une utilisation potentielle comme caractéristiques dans les modèles de prédiction.

4. Sélection et Ingénierie des Caractéristiques :

Le calcul de l'Indice de Qualité de l'Air (AQI) utilise 7 mesures : PM2.5, PM10, SO2, CO et O3.
 Pour les PM2.5, PM10, SO2, NOx et NH3, la valeur moyenne des dernières 24 heures
 Pour le CO et l'O3, la valeur maximale des dernières 8 heures est utilisée.
 L'AQI final est le Sous-Indice maximal, à condition qu'au moins l'un des PM2.5 ou PM10 soit disponible et qu'au moins trois des sept mesures le soient également.

AQI Categories (Index Values)	Ozone (ppm)		Particulate Matter (µg/m³)		Carbon Monoxide (ppm)	Sulfur Dioxide (ppb)	Nitrogen Dioxide (ppb)
	[8-hour]	[1-hour]	PM _{2.5} [24-hour]	PM ₁₀ [24-hour]	[8-hour]	[1-hour]	[1-hour]
Good (Up to 50)	0 - 0.054 None		0 – 12.0 None	0 - 54 None	0 – 4.4 None	0 - 35 None	0 - 53 None
Moderate (51 - 100)	0.055 - 0.070		12.1 – 35.4	55 – 154	4.5 – 9.4 None	36 - 75 None	54 - 100 Unusually sensitive individuals should consider limiting prolonged exertion especially near busy roads.
	Unusually sensitive people should consider reducing prolonged or heavy outdoor exertion.		Unusually sensitive people should consider reducing prolonged or heavy exertion.				
Unhealthy for Sensitive Groups (101 - 150)	0.071 - 0.085	0.125 - 0.164	35.5 – 55.4	155 – 254	9.5 – 12.4 People with heart disease, such as angina, should limit heavy exertion and avoid sources of CO, such as heavy traffic.	76 - 185 People with asthma should consider limiting outdoor exertion.	101 - 360 People with asthma, children and older adults should limit prolonged exertion especially near busy roads.
	People with lung disease (such as asthma), children, older adults, people who are active outdoors (including outdoor workers), people with certain genetic variants, and people with diets limited in certain nutrients should reduce prolonged or heavy outdoor exertion.		People with heart or lung disease, older adults, children, and people of lower socioeconomic status should reduce prolonged or heavy exertion.				
Unhealthy (151 - 200)	0.086 - 0.105	0.165 - 0.204	55.5 – 150.4	255 – 354	12.5 – 15.4 People with heart disease, such as angina, should limit moderate exertion and avoid sources of CO, such as heavy traffic.	186 – 304 Children, people with asthma, or other lung diseases, should limit outdoor exertion.	361 - 649 People with asthma, children and older adults should avoid prolonged exertion near roadways; everyone else should limit prolonged exertion especially near busy roads.
	People with lung disease (such as asthma), children, older adults, people who are active outdoors (including outdoor workers), people with certain genetic variants, and people with diets limited in certain nutrients should avoid prolonged or heavy outdoor exertion; everyone else should reduce prolonged or heavy outdoor exertion.		People with heart or lung disease, older adults, children, and people of lower socioeconomic status should avoid prolonged or heavy exertion; everyone else should reduce prolonged or heavy exertion.				
Very Unhealthy (201 - 300)	0.106 - 0.200	0.205 - 0.404	150.5 – 250.4	355 – 424	15.5 – 30.4 People with heart disease, such as angina, should avoid exertion and sources of CO, such as heavy traffic.	305 – 604 [24-hour] Children, people with asthma, or other lung diseases should avoid outdoor exertion; everyone else should reduce outdoor exertion.	650 - 1249 People with asthma, children and older adults should avoid all outdoor exertion; everyone else should avoid prolonged exertion especially near busy roads.
	People with lung disease (such as asthma), children, older adults, people who are active outdoors (including outdoor workers), people with certain genetic variants, and people with diets limited in certain nutrients should avoid all outdoor exertion; everyone else should reduce outdoor exertion.		People with heart or lung disease, older adults, children, and people of lower socioeconomic status should avoid all physical activity outdoors. Everyone else should avoid prolonged or heavy exertion.				
Hazardous (301 - 500)	-	0.405 - 0.604	250.5 – 500.4	425 – 604	30.5 – 50.4 People with heart disease, such as angina, should avoid exertion and sources of CO, such as heavy traffic; everyone else should limit heavy exertion.	605 – 1004 [24-hour] Children, people with asthma, or other lung diseases, should remain indoors; everyone else should avoid outdoor exertion.	1250 - 2049 People with asthma, children and older adults should remain indoors; everyone else should avoid all outdoor exertion.
	Everyone should avoid all outdoor exertion.		Everyone should avoid all physical activity outdoors; people with heart or lung disease, older adults, children, and people of lower socioeconomic status should remain indoors and keep activity levels low.				

5. Construction de Modèles

A. Régression Linéaire

Choix du Modèle :

Ce modèle a été choisi comme point de référence de base en raison de sa simplicité, de sa rapidité de calcul, et de sa facilité d'interprétation. Il permet d'établir une performance de base pour évaluer l'efficacité des modèles plus complexes.

B. Forêt Aléatoire (Random Forest Regressor)

La forêt aléatoire a été choisie pour sa capacité à gérer les non-linéarités et ses performances généralement élevées pour une large gamme de problèmes. Le modèle est également apprécié pour son utilité dans l'évaluation de l'importance des caractéristiques, permettant de comprendre quelles variables influencent le plus l'AQI.

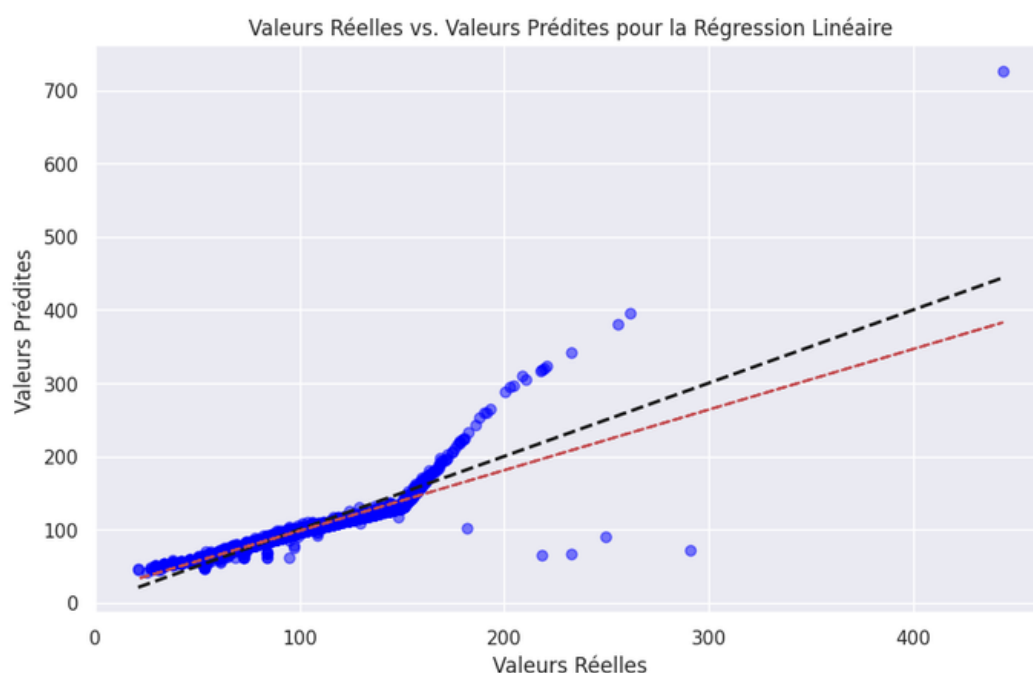
C. Boosting de Gradient (Gradient Boosting Regressor)

Le boosting de gradient a été sélectionné pour sa robustesse et son efficacité reconnues dans de nombreux problèmes de prédiction complexes, y compris ceux avec des données fortement non-linéaires et hétérogènes. C'est un choix courant pour des compétitions de science des données en raison de ses performances souvent supérieures.

6. Paramètres et évaluation

Les modèles ont été évalués principalement sur la base de la Mean Squared Error (MSE) et du coefficient de détermination (R^2). Il y a aussi MSE, et MAE

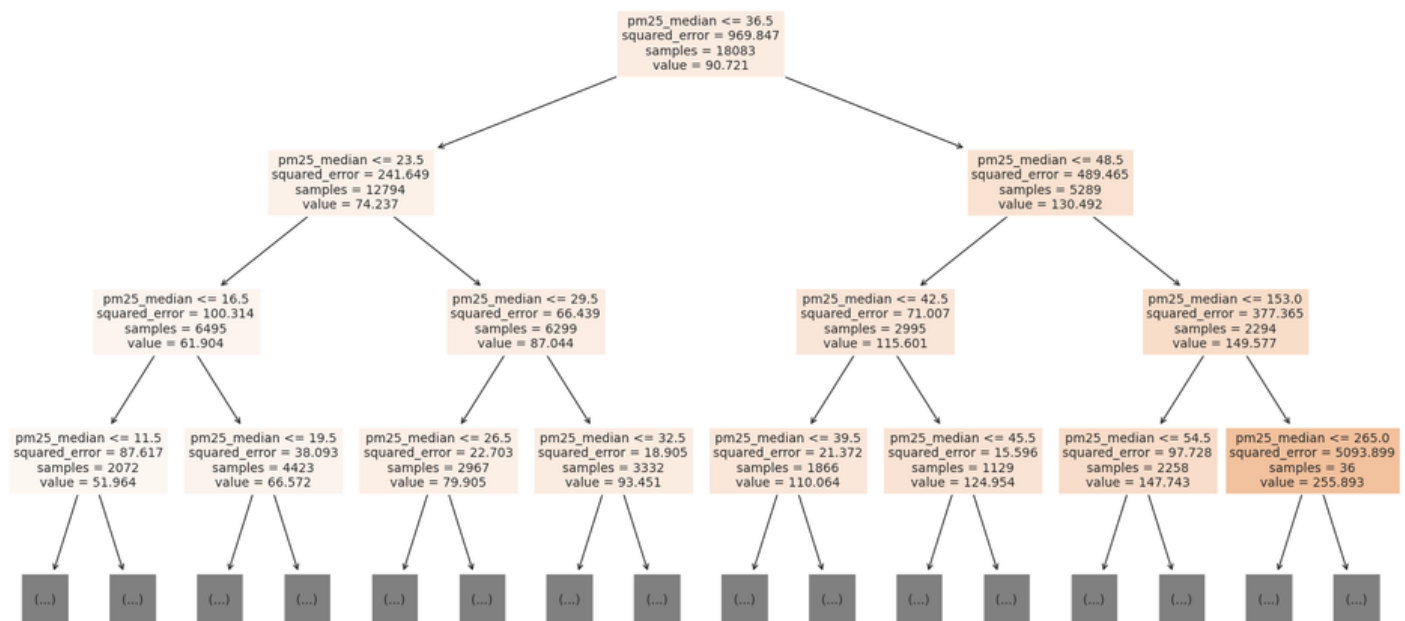
Régression Linéaire : Sans paramètres additionnels significatifs à ajuster, ce qui simplifie son utilisation.



Forêt Aléatoire : Nombre d'estimateurs (arbres) fixé à 150, stratégie de sélection aléatoire des caractéristiques à chaque division pour augmenter la diversité des arbres.

- **MSE : 16.1053, R^2 : 0.9829**

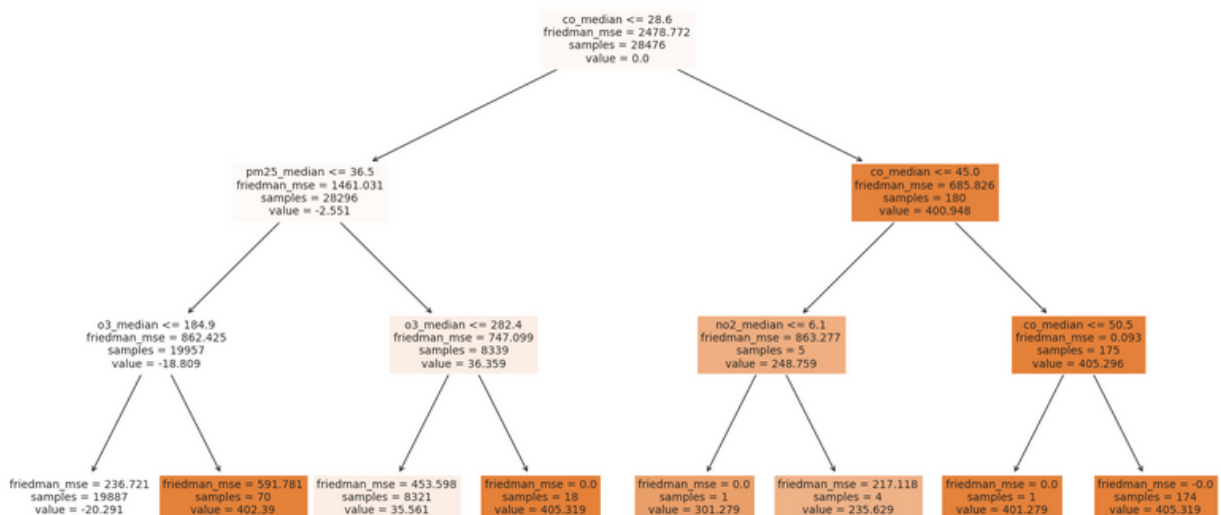
Visualisation d'un Arbre de la Forêt Aléatoire



Boosting de Gradient : j'ai exploré l'utilisation le régresseurs Gradient Boosting et avec nombre d'estimateurs fixé à 150, taux d'apprentissage à 0.1 pour équilibrer la vitesse de convergence et le risque de surapprentissage. Comme resultat:

- **RMSE 3.3577, R^2 : 0.9964**

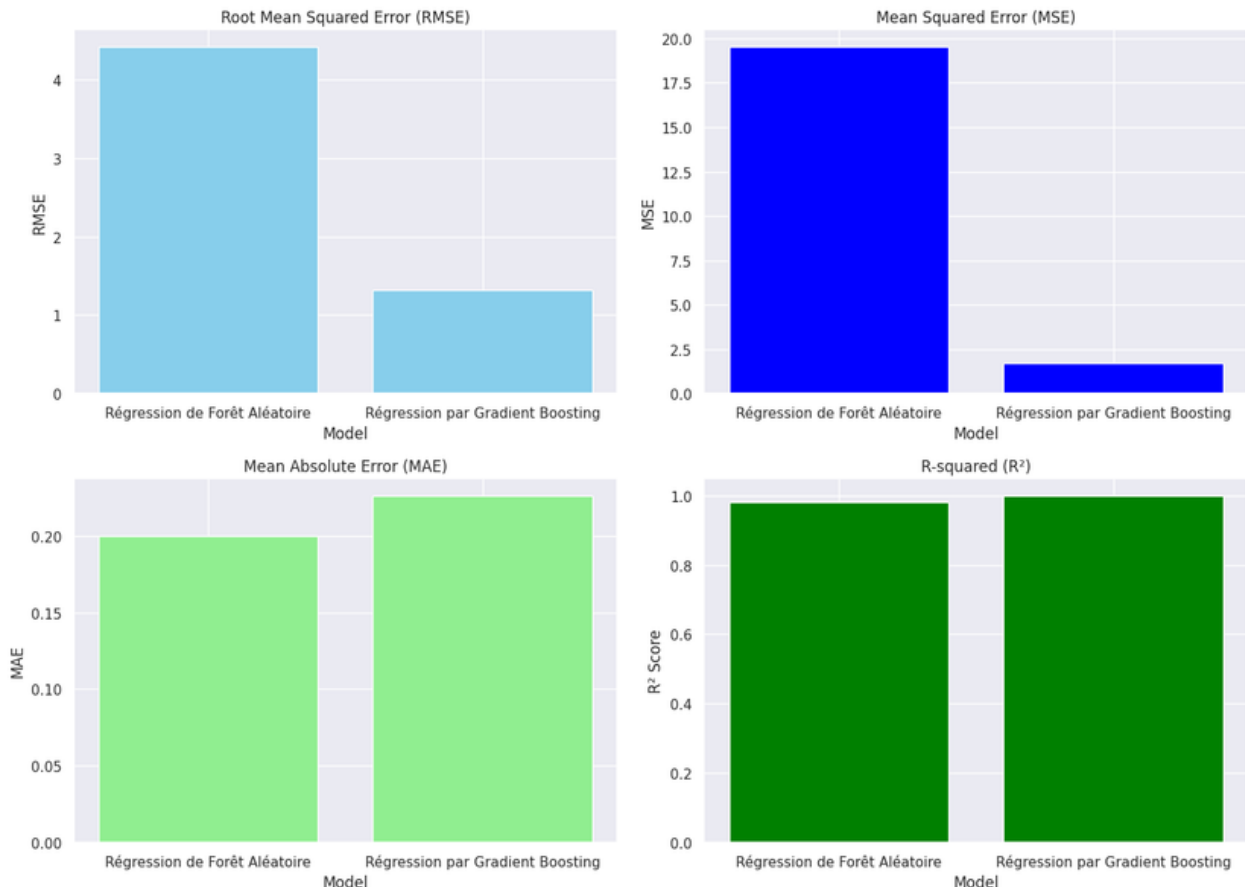
Visualisation of a Tree from Gradient Boosting



Pour améliorer la précision du modèle, notamment en termes d'erreur quadratique moyenne (RMSE), nous avons employé la méthode GridSearchCV afin d'optimiser leurs hyperparamètres.

- **RMSE 1.314 MSE: 1.728 R^2 : 0.9981**

Comparaison

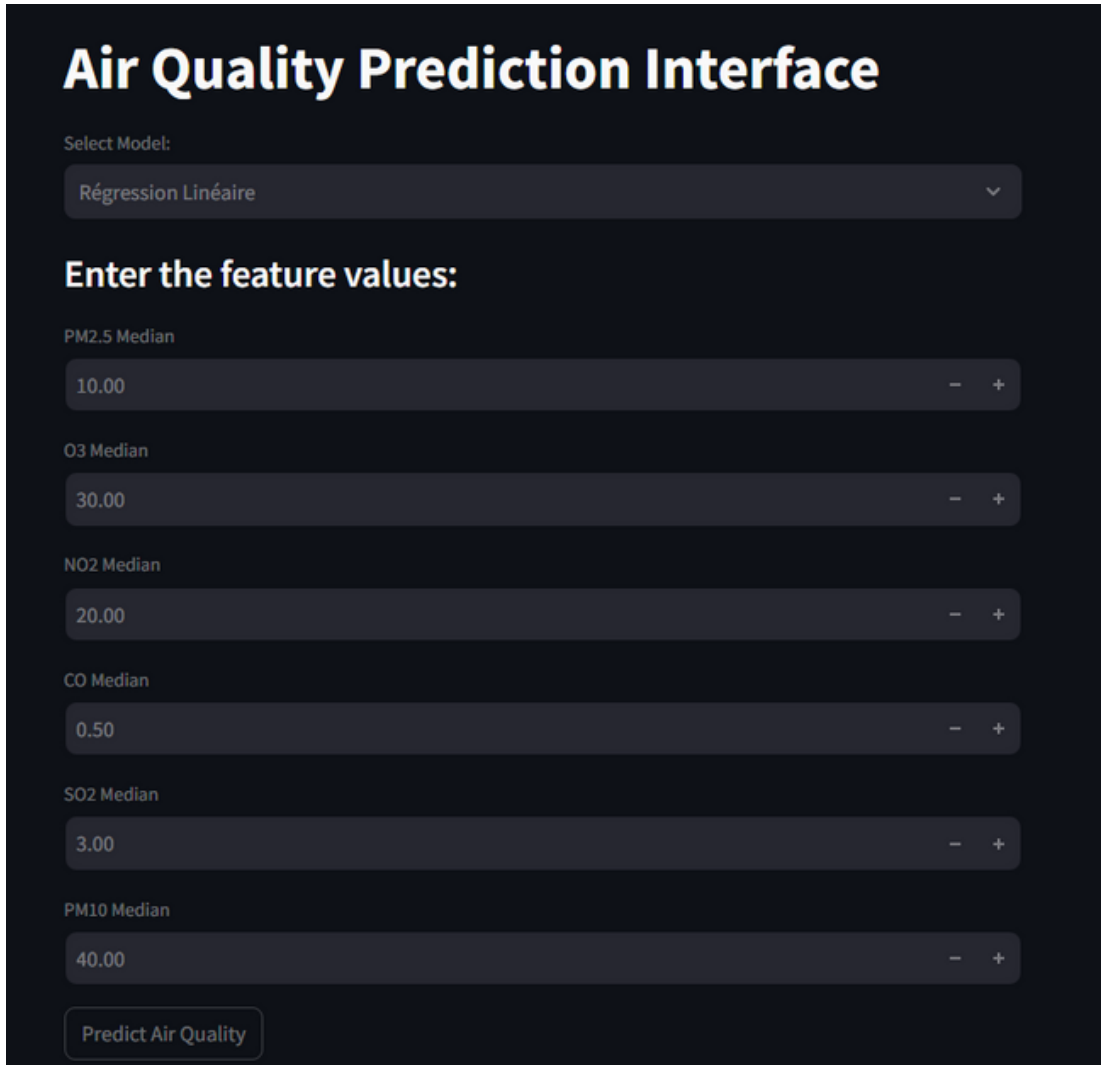


- RMSE: La Régression par Gradient Boosting montre une erreur quadratique moyenne racinée (RMSE) significativement inférieure à celle de la Régression de Forêt Aléatoire, indiquant de meilleures prédictions en moyenne pour le premier modèle.
- MSE: la Régression par Gradient Boosting est beaucoup plus basse que celle de la Régression de Forêt Aléatoire, suggérant une plus grande précision dans les prédictions du modèle Gradient Boosting.
- MAE: la Régression par Gradient Boosting présente également une valeur inférieure, ce qui implique que les prédictions sont en moyenne plus proches des valeurs réelles par rapport à la Régression de Forêt Aléatoire.
- R^2 : les deux modèles affichent des scores élevés, avec un léger avantage pour la Régression par Gradient Boosting.

Cela suggère que, pour ces données et selon ces métriques, **la Régression par Gradient Boosting** semble être plus performante que la Régression de Forêt Aléatoire.

6. Déploiement et Visualisation

SUtilisation du Streamlit pour fair la prediction de la Pollustion

The image shows a web application titled "Air Quality Prediction Interface" with a dark theme. At the top, there is a "Select Model:" dropdown menu currently showing "Régression Linéaire". Below this, a section titled "Enter the feature values:" contains six input fields for different air quality metrics: "PM2.5 Median" (10.00), "O3 Median" (30.00), "NO2 Median" (20.00), "CO Median" (0.50), "SO2 Median" (3.00), and "PM10 Median" (40.00). Each input field has minus and plus buttons on the right for adjustment. At the bottom left, there is a button labeled "Predict Air Quality".

Air Quality Prediction Interface

Select Model:
Régression Linéaire

Enter the feature values:

PM2.5 Median
10.00

O3 Median
30.00

NO2 Median
20.00

CO Median
0.50

SO2 Median
3.00

PM10 Median
40.00

Predict Air Quality

Resources:

Kaggle Notebooks:

- [Calcule du AQI](#)
- [Visualization](#)

Others:

- [GridSearchCV](#)
- [AQI Documentation](#)
- [Research paper](#)