Prédiction de la Qualité de l'Air à la Station Auber RER A avec l'IA

Introduction

Le présent projet s'appuie sur les données de la RATP pour développer un modèle prédictif de la qualité de l'air à la station Auber du RER A. Nous avons suivi un processus rigoureux, défini par un cahier des charges précis, qui inclut le nettoyage des données, une analyse exploratoire, la modélisation par apprentissage automatique et la prédiction des niveaux de qualité de l'air.

I. Nettoyage des Données

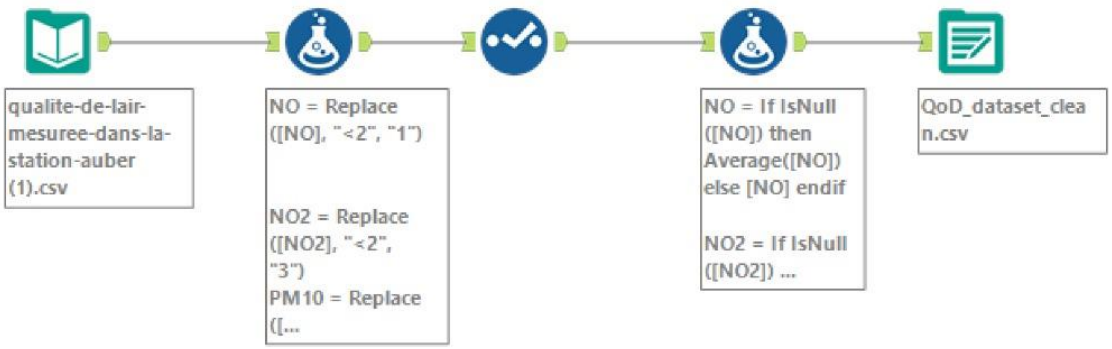
- Outils et Techniques Utilisés:

Nous avons initialement utilisé Alteryx pour nettoyer nos données. Le workflow de nettoyage comprenait plusieurs étapes clés :

Outil Formula : Étant donné que certaines valeurs dans notre dataset étaient de type string, nous avons dû les convertir en valeurs numériques. Par exemple, dans la colonne NO (Monoxyde d'azote), nous avons remplacé les valeurs "<2" par "1" pour pouvoir effectuer des calculs.

Outil Select : Pour uniformiser le type de données, nous avons utilisé l'outil Select pour définir toutes les valeurs comme des entiers.

Outil Formula (Remplacement des valeurs manquantes) : Pour traiter les valeurs manquantes, nous avons utilisé l'outil Formula pour remplacer ces valeurs par la moyenne de la colonne correspondante. Par exemple, si une cellule était vide dans la colonne NO, nous avons calculé la moyenne de la colonne NO et remplacé la valeur manquante par cette moyenne.



- CALCUL DE L'AQI :

Par la suite, nous avons employé un script Python pour procéder au calcul de l'Indice de Qualité de l'Air (AQI), en fonction des niveaux de divers polluants, incluant notamment le dioxyde d'azote (NO), le dioxyde d'azote (NO2), les particules en suspension (PM10 et PM2.5) ainsi que le dioxyde de carbone (CO2). Ce processus a permis de générer une nouvelle colonne nommée "AQI" au sein de notre jeu de données. Par la suite, nous avons archivé ce jeu de données modifié dans un nouveau fichier CSV. Cette étape revêt une importance capitale dans la préparation de nos données en vue de la conception et de l'entraînement de modèles de machine learning supervisés.

II. Analyse Exploratoire des Données (EDA)

Une analyse exhaustive des données a été réalisée pour identifier les tendances, anomalies et corrélations potentielles entre les diverses variables. Cette étape est cruciale pour orienter les choix de modélisation et assurer l'efficacité du modèle de prédiction.

III. Construction du Modèle de Machine Learning

Nous avons opté pour une approche supervisée en utilisant la régression linéaire, pour ses avantages de simplicité et de clarté dans l'interprétation des relations entre les variables. Ce modèle permet de cerner efficacement les liens linéaires entre les différents polluants mesurés et l'indice AQI. La sélection et l'optimisation des hyper-paramètres ont été effectuées pour maximiser la performance du modèle.

IV. Prédictions et Visualisations

- MODELISATION ET EXPLORATION DES DONNEES :

Dans cette phase ultérieure, nous avons élaboré un autre script Python visant à explorer et à modéliser les données. Nous avons initié la séquence en important les librairies fondamentales, notamment NumPy, Pandas et Matplotlib, ainsi que les modules pertinents de scikit-learn pour l'analyse et la construction du modèle. Par la suite, nous avons procédé au chargement des données à partir du fichier "supervisé.csv", contenant les données préalablement préparées, incluant l'Indice de Qualité de l'Air (AQI). Après l'étape de chargement, une préparation des données a été entreprise, comprenant la conversion adéquate de la colonne "DATE/HEURE" ainsi que son exclusion du DataFrame, étant non pertinente pour la modélisation. Ensuite, une partition en ensembles d'entraînement et de test a été réalisée pour l'évaluation du modèle. Utilisant la régression linéaire comme modèle d'apprentissage supervisé, une phase d'entraînement sur l'ensemble d'entraînement a été menée, suivie d'une évaluation des performances sur l'ensemble de test, avec le calcul du MSE et du R^2. Enfin, une visualisation graphique a été produite, permettant une appréciation rapide des prédictions de l'AQI par rapport aux valeurs réelles, renforçant ainsi la compréhension des performances du modèle.

- VISUALISATION DES RESULTATS :

Implémentation: Chargement des données: Les données nettoyées, comprenant la colonne AQI, ont été chargées à partir d'un fichier CSV.

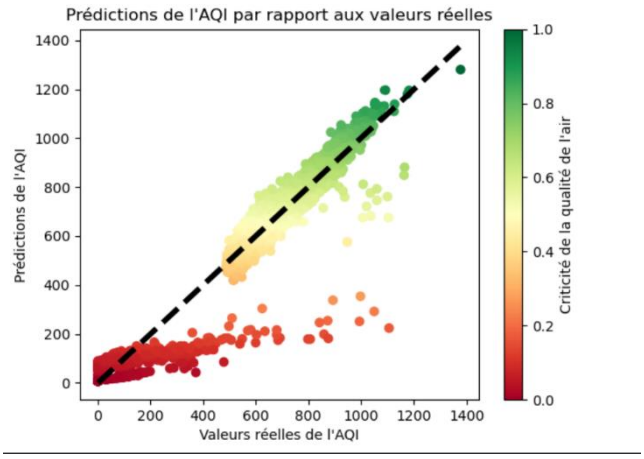
Préparation des données: Conversion et suppression des colonnes non pertinentes, telles que la colonne "DATE/HEURE".

Division des données: Séparation en ensembles d'entraînement et de test.

Entraînement du modèle: Utilisation de la régression linéaire sur l'ensemble d'entraînement.

Visualisation des Résultats:

Présentation des prédictions de l'AQI en comparaison avec les valeurs réelles au moyen de graphiques, facilitant ainsi l'analyse visuelle des performances du modèle.



V. Discussion des Résultats

Les résultats obtenus ont été analysés pour évaluer l'efficacité du modèle de régression linéaire dans le contexte de notre étude. Cette discussion inclut une évaluation de la précision des prédictions, de la pertinence des variables choisies et des perspectives d'amélioration du modèle.

Conclusion

Ce projet démontre la capacité de l'intelligence artificielle à prédire la qualité de l'air en milieu urbain complexe, comme celui de la station Auber du RER A. Les techniques de machine learning, en particulier la régression linéaire, se sont avérées être des outils précieux pour analyser et prédire l'impact des variations des niveaux de pollution sur la qualité de l'air.