Rapport sur l'analyse et la Prédiction de la Qualité de l'Air en Inde (2015 - 2020)

Titre du projet : Classification multiclasse de AQI_Bucket

Indice de Qualité de l'Air (IQA) et données horaires par stations et villes en Inde

Lien: https://www.kaggle.com/datasets/rohanrao/air-quality-data-in-india

La pollution de l'air est une préoccupation croissante à l'échelle mondiale, et avec l'industrialisation et l'urbanisation accrues, il devient crucial de surveiller et de prédire la qualité de l'air en temps réel. L'une des façons les plus fiables de quantifier la pollution de l'air est de calculer l'Indice de Qualité de l'Air (IQA). Dans cet article, nous allons explorer comment prédire l'IQA en utilisant Python, en tirant parti des outils de science des données et des algorithmes d'apprentissage automatique (machine learning).

Qu'est-ce que l'IQA ?

L'Indice de Qualité de l'Air (IQA ou AQI en Anglais) est un indicateur normalisé utilisé pour communiquer le degré de pollution actuel de l'air ou la pollution prévue. L'IQA est calculé à partir des polluants

1. Introduction générale

Ce rapport présente une étude approfondie portant sur la qualité de l'air en Inde, basée sur l'analyse d'un jeu de données réel extrait du fichier city_day.csv. L'objectif principal du projet est d'examiner les tendances de la pollution urbaine, de comprendre les facteurs influençant l'indice de qualité de l'air (AQI) et de développer un modèle de classification capable de prédire la catégorie de qualité de l'air à partir des caractéristiques environnementales mesurées. Cette étude s'inscrit dans une démarche scientifique et technique structurée, reposant sur une

analyse descriptive, un prétraitement rigoureux et la mise en œuvre de plusieurs modèles de machine learning.

2. Objectifs et contexte du projet

L'étude vise à comprendre comment les polluants atmosphériques tels que PM2.5, PM10, NO2, SO2, CO et O3 influencent la qualité globale de l'air dans différentes villes indiennes. Le projet s'appuie sur une approche data-driven permettant non seulement d'évaluer la gravité de la pollution, mais aussi de prédire l'état futur de l'air selon diverses conditions. L'analyse se concentre sur trois axes : la compréhension statistique des données, l'optimisation des modèles prédictifs, et la validation empirique des performances.

3. Chargement et Exploration des Données

Le jeu de données city_day.csv a été chargé et exploré afin de mieux comprendre sa structure. Cette exploration a révélé la présence de valeurs manquantes dans plusieurs colonnes et des types de données variés. Un aperçu statistique et graphique a permis de dégager des tendances générales, notamment une forte variabilité de la pollution selon les villes et les périodes de l'année.

4. Prétraitement des Données

Un prétraitement complet a été réalisé pour garantir la fiabilité des analyses et la robustesse des modèles. Les principales étapes ont été les suivantes :

- Gestion des valeurs manquantes : les colonnes numériques ont été imputées par la médiane pour réduire l'influence des valeurs extrêmes, tandis que la colonne catégorique AQI_Bucket a été complétée avec le mode ('Moderate').
- Encodage des variables catégoriques : AQI_Bucket a été encodée ordinalement selon son ordre de gravité, et City a été encodée par fréquence.
- Ingénierie de caractéristiques : la colonne Date a été transformée en sous-composantes (année, mois, jour), puis supprimée.
- Suppression de caractéristiques : la variable AQI a été retirée pour éviter la fuite d'information, car elle est fortement corrélée avec la cible AQI_Bucket_Encoded.

• Détection des valeurs aberrantes : les méthodes de boîte à moustaches et IQR ont mis en évidence une variabilité naturelle importante, reflétant la diversité géographique et climatique des régions indiennes.

5. Préparation pour la Modélisation

Les données ont ensuite été divisées en deux ensembles : un ensemble d'entraînement (80 %) et un ensemble de test (20 %). Un StandardScaler a été appliqué aux caractéristiques numériques afin de normaliser les données, condition essentielle pour les modèles sensibles aux échelles comme le KNN.

6. Entraînement, Évaluation et Optimisation des Modèles

Trois modèles de classification ont été sélectionnés pour l'expérimentation : K-Nearest Neighbors (KNN), Random Forest (RF) et XGBoost (XGB). Ces modèles ont été évalués selon plusieurs métriques : Accuracy, Précision, Rappel, F1-score, Log-Loss et Matrice de confusion. Les résultats initiaux ont montré que les modèles Random Forest et XGBoost surpassaient KNN, avec un léger avantage pour XGBoost.

Une optimisation par GridSearchCV a ensuite été menée avec validation croisée à 5 plis. Les paramètres ajustés ont permis d'améliorer légèrement les performances des trois modèles :

• KNN : ajustement des paramètres n_neighbors, weights et metric ; amélioration marginale observée.

• Random Forest : optimisation de n_estimators, max_depth et min_samples_leaf ; résultats plus stables.

• XGBoost : ajustement de n_estimators, max_depth, learning_rate et subsample ; meilleures performances globales obtenues.

7. Résultats et Sélection du Modèle Final

Un tableau comparatif des performances des modèles a été élaboré. L'XGBoost optimisé s'est démarqué comme le modèle le plus performant, avec un F1-score de 0.8312 et un Log-Loss de 0.4472. Ces résultats témoignent d'une excellente précision et d'une bonne calibration des probabilités prédites. Sa capacité à capturer des relations non linéaires complexes et sa robustesse face à la variabilité des données justifient son choix comme modèle final.

8. Conclusion et Perspectives

Cette étude a permis de concevoir un pipeline complet d'analyse et de prédiction de la qualité de l'air en Inde, combinant rigueur statistique et performance algorithmique. Le modèle XGBoost optimisé offre une solution efficace pour la classification de la qualité de l'air, pouvant être utilisée par les autorités environnementales pour anticiper les épisodes de pollution et orienter les politiques publiques. Des perspectives futures incluent l'intégration de données météorologiques, le déploiement du modèle dans une application web interactive et la mise à jour continue du modèle par apprentissage en ligne.