Rapport de Projet : Prédiction de la Qualité de l'Air

BENOUADAH Alaeddine Mohamed

December 11, 2024

Introduction

Ce projet vise à développer un modèle de machine learning capable de prédire la qualité de l'air en fonction de facteurs environnementaux et démographiques. L'objectif est d'identifier les zones à risque et de fournir une évaluation précise des niveaux de pollution pour informer les décideurs et la population.

1 Description des Données

1.1 Source des données

Les données utilisées proviennent d'un ensemble contenant 5000 échantillons, chacun décrivant plusieurs caractéristiques environnementales et démographiques influençant la qualité de l'air.

1.2 Caractéristiques principales

- **Température (°C)** : Température moyenne dans la région.
- **Humidité (%)** : Humidité relative enregistrée.
- **PM2.5 (g/m^3)** : Concentration de particules fines.
- **PM10 (g/m^3)** : Concentration de particules grossières.
- **NO2 (ppb)** : Concentration de dioxyde d'azote.
- **SO2 (ppb)** : Concentration de dioxyde de soufre.
- **CO (ppm)** : Concentration de monoxyde de carbone.
- **Proximité des zones industrielles (km)** : Distance par rapport à la zone industrielle la plus proche.
- **Densité de population ($habitants/km^2$)** : Nombre de personnes par kilomètre carré.

Variable cible :

- **Qualité de l'air** : Catégorielle (Bon, Modéré, Médiocre, Dangereux).

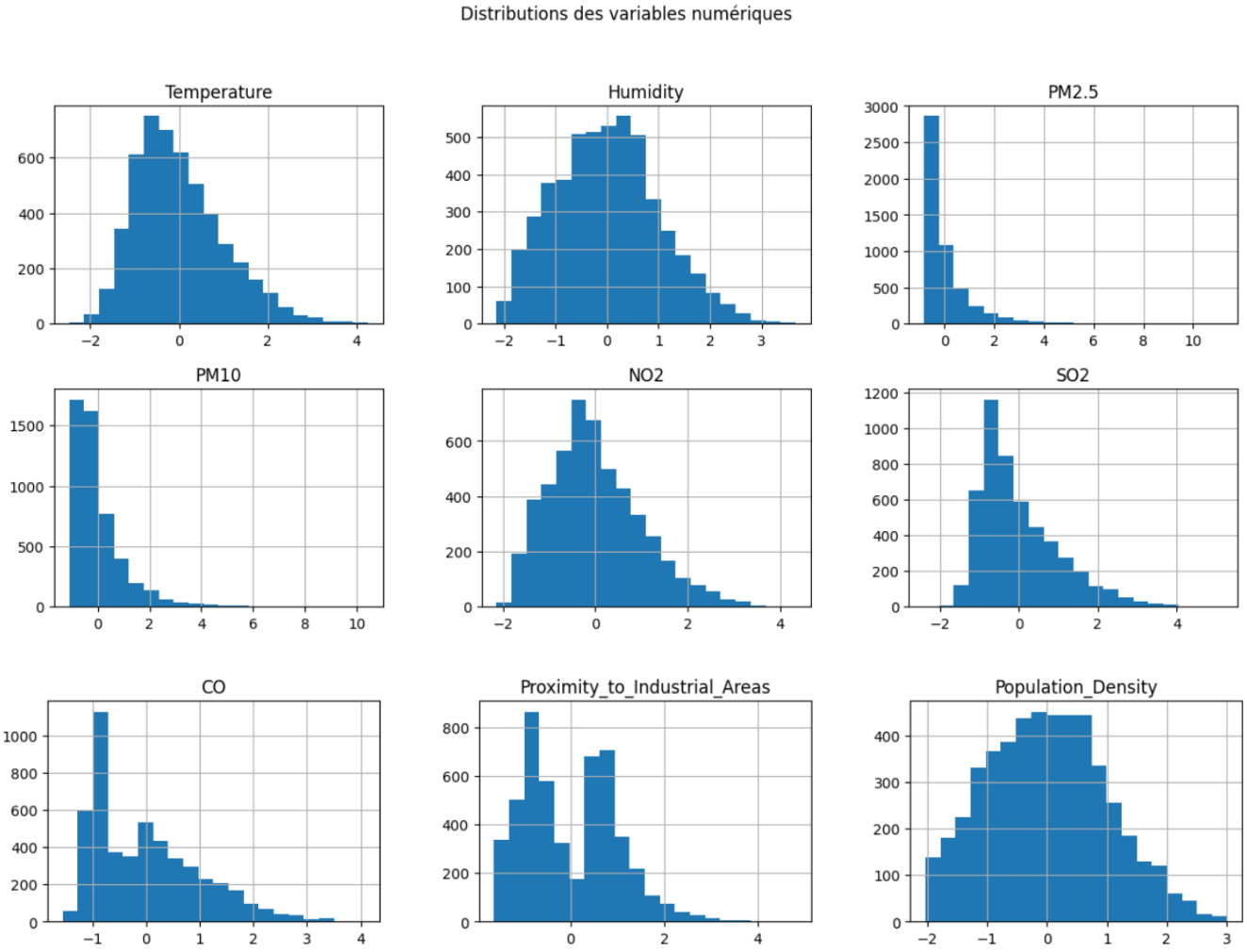
2 Préparation des Données

2.1 Nettoyage et Prétraitement

- **Gestion des valeurs manquantes :** Les valeurs manquantes ont été remplacées par la moyenne pour éviter la perte de données.
- **Normalisation des variables numériques :** Utilisation de `StandardScaler` pour mettre les caractéristiques sur une même échelle et réduire l'effet des unités.
- **Encodage de la variable cible :** La variable `Air Quality` a été encodée en valeurs numériques : Bon (0), Modéré (1), Médiocre (2), Dangereux (3).

2.2 Séparation des données

Les données ont été divisées en un ensemble d'entraînement (80%) et un ensemble de test (20%) pour évaluer les performances.



3 Modèles de Machine Learning

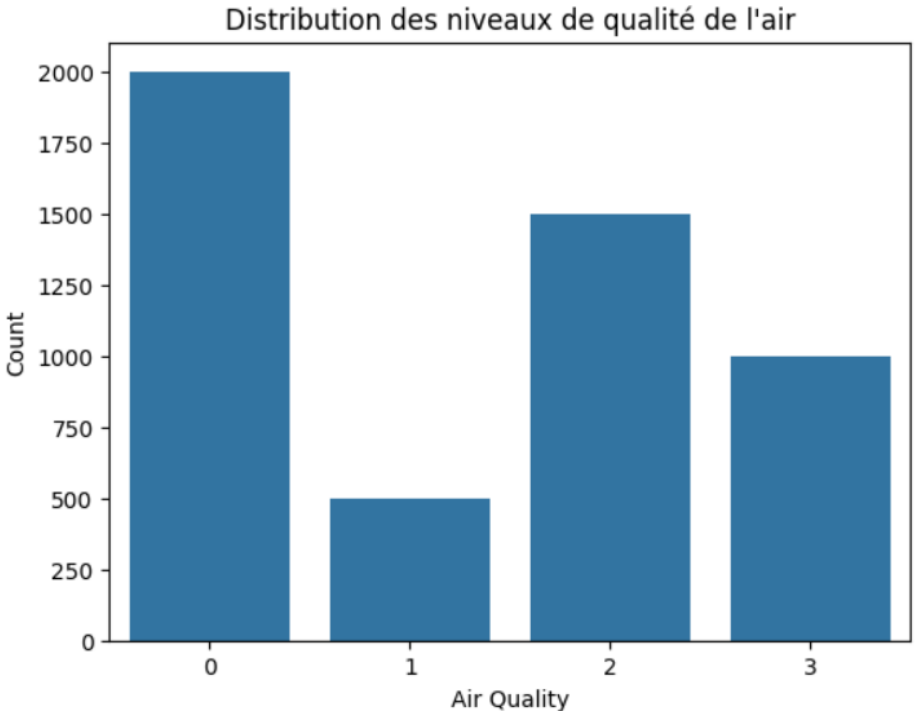
3.1 Choix des modèles

Plusieurs modèles ont été testés pour comparer leurs performances :

- **Régression Logistique** : Modèle simple pour des prédictions rapides.
- **Support Vector Machine (SVM)** : Modèle robuste pour des problèmes avec des classes bien séparées.
- **K-Nearest Neighbors (KNN)** : Méthode basée sur la proximité des données.
- **Arbre de Décision** : Modèle interprétable et rapide. .

3.2 Validation croisée

Chaque modèle a été évalué à l'aide de la validation croisée pour garantir des résultats robustes.



4 Résultats et Comparaison des Modèles

4.1 Performances globales

Les scores moyens de précision (*accuracy*) obtenus pour chaque modèle sont :

- **Régression Logistique** : 0.8
- **SVM** : 0.77
- **KNN** : 0.76
- **Arbre de Décision** : 0.75

4.2 Choix du modèle final

Le modèle **Arbre de Décision** a été retenu en raison de ses meilleures performances globales et de sa robustesse.

5 Évaluation du Modèle Final

5.1 Rapport de classification

Le rapport de classification montre les métriques pour chaque classe (Bon, Modéré, Médiocre, Dangereux) :

Classe	Précision	Rappel	F1-Score	Support
Bon	0.84	0.76	0.80	75
Modéré	0.78	0.83	0.80	76
Médiocre	0.79	0.85	0.82	72
Dangereux	0.91	0.87	0.89	77

Table 1: Résultats de classification par classe.

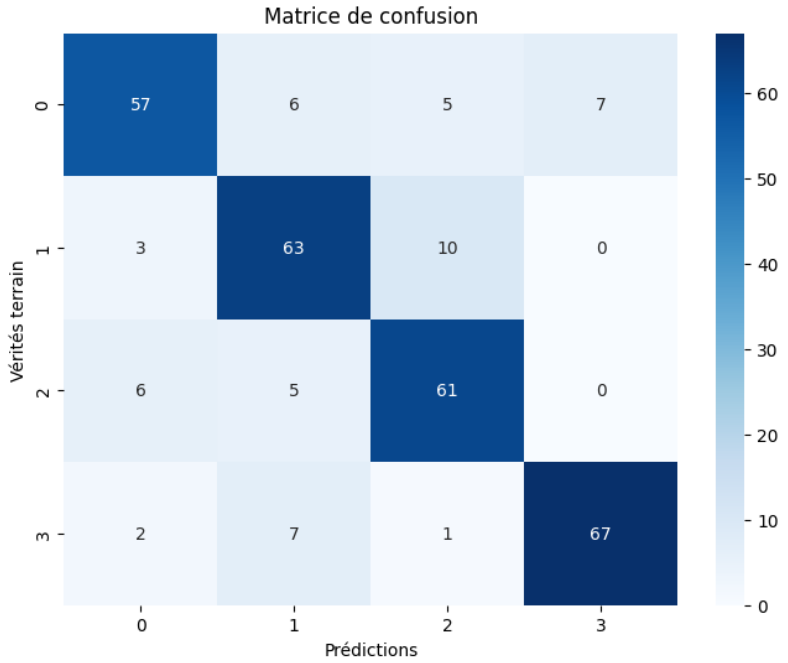
5.2 Précision globale

La précision globale du modèle sur l'ensemble de test est de **85%**.

5.3 Matrice de confusion

La matrice de confusion met en évidence les prédictions correctes et les erreurs :

- Les classes "Bon" et "Dangereux" sont bien prédites.
- Quelques erreurs sont observées entre "Modéré" et "Médiocre".



5.4 Importance des caractéristiques

Les caractéristiques les plus importantes identifiées par le modèle sont :

- **PM2.5** : Principal indicateur de la qualité de l'air.

- **PM10** : Affecte la pollution générale.
- **Proximité des zones industrielles** : Forte corrélation avec des niveaux de pollution élevés.
- **NO2** : Contributeur clé à la pollution.

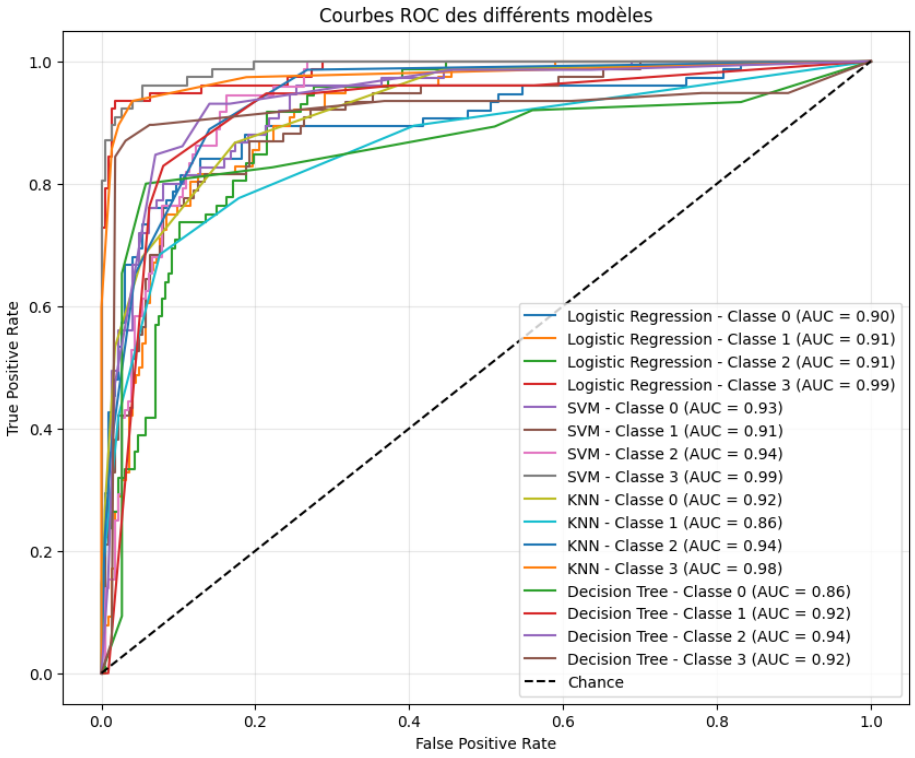
6 Optimisations et Améliorations

Une analyse en composantes principales (PCA) a été effectuée pour réduire la dimensionnalité des données sans perte significative d'information. Avec 5 composantes principales, les performances du modèle sont restées stables à 84%. La sélection des 5 meilleures caractéristiques avec `SelectKBest` a permis de réduire le temps d'entraînement tout en maintenant une précision élevée.

7 Perspectives

Le modèle **Régression logistique** a démontré de bonnes performances pour prédire la qualité de l'air. Il offre une précision globale de **85%** et identifie les principales caractéristiques influençant la pollution, telles que les concentrations de PM2.5 et PM10, ainsi que la proximité des zones industrielles.

- **Amélioration des données** : Intégrer des données météorologiques supplémentaires, comme la vitesse du vent et les précipitations, pour améliorer les prédictions.
- **Déploiement** : Développer une API (par exemple, avec Flask) pour permettre une utilisation en temps réel.
- **Modèles avancés** : Expérimenter avec des modèles comme XGBoost ou LightGBM pour explorer d'autres gains de performances.



8 Analyse Critique des Performances

8.1 1. Performances globales

Le modèle Random Forest a atteint une précision globale de **85%**, ce qui est satisfaisant pour une première itération. Cependant, plusieurs aspects doivent être considérés pour évaluer la pertinence réelle des résultats :

Précision par classe :

- Les classes **"Bon"** et **"Dangereux"** ont des scores élevés (précision et rappel ¿ 85%), indiquant que le modèle distingue bien les extrêmes.
- Les classes intermédiaires **"Modéré"** et **"Médiocre"** sont moins bien prédites, avec des rappels inférieurs (79% pour "Modéré"), ce qui suggère une confusion entre ces classes.

8.2 2. Analyse des caractéristiques

Les caractéristiques identifiées comme les plus importantes sont :

- **PM2.5 et PM10** : Particules fines et grossières sont des indicateurs cruciaux de pollution.
- **Proximité des zones industrielles** : Liée directement à l'exposition aux polluants.
- **NO2** : Polluant courant dans les zones urbaines et industrielles.

8.3 3. Points faibles identifiés

Confusion entre les classes proches :

- Les niveaux intermédiaires (**"Modéré", "Médiocre"**) posent des difficultés, ce qui affecte les rappels et les précisions de ces classes.

Possibilité de surapprentissage :

- Random Forest est un modèle robuste, mais il peut surapprendre si les hyperparamètres ne sont pas bien réglés. La validation croisée a été utilisée, mais des tests supplémentaires sont nécessaires pour garantir que les résultats sont généralisables.

Limitation des données :

- Les données utilisées sont statiques (uniquement des moyennes et mesures ponctuelles). Cela ne prend pas en compte la variabilité temporelle des polluants.

9 Propositions d'Amélioration

Pour améliorer les performances, il serait utile d'élargir le jeu de données avec des informations supplémentaires :

- **Données temporelles :** Inclure des variations horaires ou journalières pour capturer l'évolution des polluants (exemple : concentrations de PM2.5 durant les heures de pointe).

- **Données météorologiques avancées :** Ajouter des variables comme la vitesse du vent, les précipitations et l'ensoleillement, qui influencent la dispersion des polluants.
- **Ensemble Learning :** Combiner plusieurs modèles (*Stacking*) pour capturer différents aspects des données (exemple : un modèle linéaire comme la Régression Logistique pour les classes simples et un modèle non linéaire comme Random Forest pour les classes plus complexes).
- **Réseaux de neurones :** Si les données sont suffisamment volumineuses, les réseaux de neurones (exemple : Multi-Layer Perceptron) peuvent capturer des relations non linéaires plus complexes.

10 Conclusion

Les performances obtenues avec Régression logistique et SVM sont solides, mais il reste des marges d'amélioration, notamment pour mieux différencier les classes intermédiaires et enrichir les données. Des techniques avancées de modélisation, combinées à une meilleure préparation des données, pourraient significativement améliorer les résultats et leur application dans des scénarios réels.