Analyse des résultats et choix des modèles

Introduction :

L'objectif est de prédire la qualité de l'air à partir de différentes données liées à la pollution. On a utilisé deux modèles, random forest et xgboost, pour cette analyse. Ce document explique les résultats obtenus et pourquoi on a choisi ces modèles.
Description des données

- Données utilisées : température, humidité, pm2.5, pm10, no2, so2, co, proximité des zones industrielles et densité de population.

Ce qu'on cherche à prédire : la qualité de l'air, classée en différentes catégories (exemple : bon, mauvais). dans le but de faire une prédiction précise de la qualité de l'air à partir des autres données.

Modèles utilisés

Random forest :

C'est un modèle simple et efficace pour ce genre de données (tableaux avec chiffres et catégories). Il est super utile pour trouver quelles données influencent le plus les résultats.

On a utilisé 100 arbres dans la forêt, avec un paramètre pour assurer que les résultats sont reproductibles.

Résultats :

Précision obtenue : 0.37 % (ce qui ne nous assure une super qualité de prédiction)
Les données les plus importantes sont pm2.5 et pm10, suivies par no2.

Visualisations principales

On utillise la **matrice de corrélation** pour voir comment les différentes données sont liées entre elles.On a remarqué que le pm2.5, pm10 et no2 sont fortement liés, ce qui montre qu'ils jouent un rôle clé dans la qualité de l'air.

Distribution des classes pour vérifier si les catégories de la qualité de l'air sont équilibrées. On a observé que certaines catégories comme "good" et "poor" sont sur-représentées par rapport aux autres, ce qui peut poser un problème pour les prédictions.

On a appris grace au random forest que le pm2.5 et pm10 sont les données les plus importantes, alors que la densité de population ou la proximité des zones industrielles ont un impact plus faible.

Limites et pistes d'amélioration

- Les catégories de la qualité de l'air ne sont pas équilibrées, ce qui peut biaiser les résultats.
- Il manque peut-être des infos importantes comme l'ozone.
- Ce qu'on pourrait faire pour améliorer :
- Utiliser des techniques comme le sur-échantillonnage (smote) pour mieux équilibrer les catégories.

Conclusion

Random forest a été choisi parce qu'il est simple, robuste et facile à interpréter. Les résultats montrent que pm2.5 et pm10 sont les facteurs clés pour déterminer la qualité de l'air. Pour aller plus loin, il faudrait mieux équilibrer les données et tester d'autres approches pour affiner les prédictions.

Le modèle Gradient Boosting Machines

(GBM) parce qu'il est très performant pour les tâches de classification avec des données tabulaires. Voici pourquoi :

Efficacité sur des relations complexes : GBM peut capturer des relations non linéaires entre les variables explicatives et la cible, ce qui le rend adapté à des problèmes où les interactions entre les variables sont importantes.

Précision : Il combine plusieurs arbres de décision en boostant les erreurs à chaque étape, ce qui améliore la précision globale du modèle.

Interprétabilité : Même s'il est plus complexe que des modèles comme la régression logistique, GBM permet de visualiser l'importance des variables pour comprendre les facteurs clés.

Flexibilité : Il fonctionne bien avec des données déséquilibrées (si des ajustements comme la pondération des classes sont faits) et peut être optimisé avec des hyperparamètres comme le taux d'apprentissage ou la profondeur des arbres.

Large adoption : GBM est un modèle robuste et fiable utilisé dans de nombreux cas pratiques, notamment pour la classification et la régression.
En résumé, GBM est un bon choix lorsqu'on cherche un modèle puissant, capable d'offrir à la fois précision et informations sur les variables les plus influentes
Interprétation des résultats obtenus :

Matrice de confusion

- Catégorie “Good” : Bien prédite avec 467 cas corrects, mais il y a des confusions avec “Moderate” et “Hazardous”.
- Catégorie “Moderate” : Beaucoup de confusion avec “Good” (382 erreurs), ce qui montre des similitudes entre ces classes.
- Catégorie “Poor” : Très mal prédite (6 cas corrects), souvent confondue avec “Good”.
- Catégorie “Hazardous” : Peu de bonnes prédictions (22 cas corrects), avec de fortes confusions.

Conclusion :

Le modèle fonctionne bien pour “Good” (classe dominante) mais struggle avec les classes moins fréquentes (“Poor” et “Hazardous”).

Les variables importantes :

- Humidity (humidité) : Facteur principal influençant les prédictions.
- CO (monoxyde de carbone) et SO2 : Indicateurs majeurs de pollution.
- Moins influentes :
- **PM2.5 et PM10 sont moins importantes, probablement parce qu’elles sont fortement corrélées à l’humidité et au CO.**

Interprétation globale :

- Points forts : Bonne performance pour la catégorie “Good”. Les variables clés (humidité, CO, SO2) sont bien identifiées.
- Points faibles : Mauvaise prédiction des classes rares (“Poor”, “Hazardous”) et confusion entre “Good” et “Moderate”.