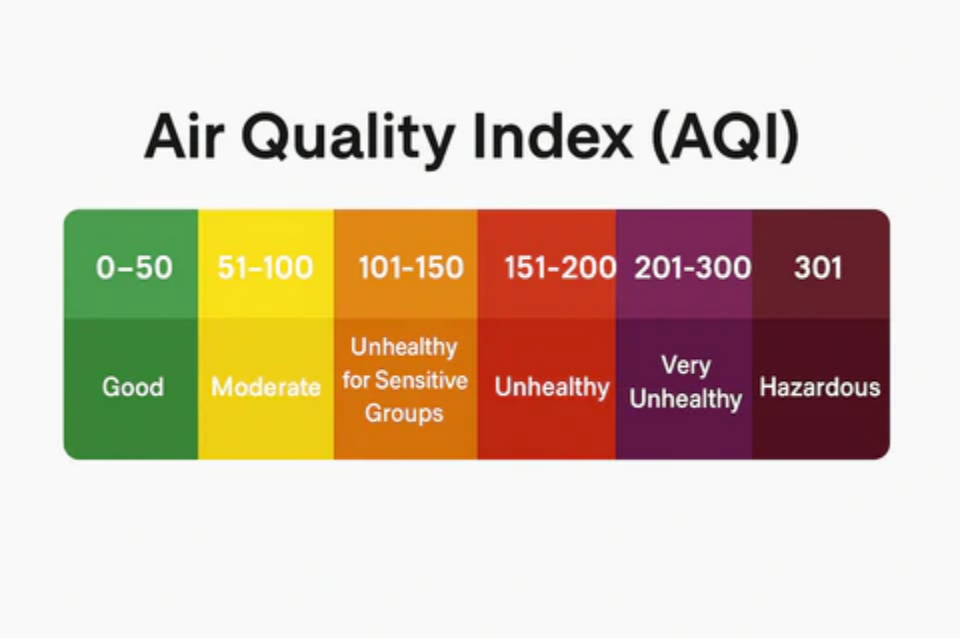
TD Machine Learning - Énoncé Détaillé



Prédiction de la Qualité de l'Air Mondiale

Introduction

Ce TD vous permettra de mettre en pratique les concepts fondamentaux du Machine Learning sur un **enjeu environnemental et de santé publique majeur** : la pollution de l'air.

Vous allez construire un système de prédiction de la qualité de l'air en suivant toutes les étapes d'un projet ML professionnel, en utilisant des **données réelles de plus de 23,000 villes** dans le monde (2017-2022).

Description des Données

Dataset : Global Air Pollution

- **Source** : Kaggle (Hasib Al Muzdadid, 2022)
- **Période** : 2017-2022
- **Couverture** : Plus de 23,000 villes mondiales
- **Taille** : Environ 50 MB

Variables d'entrée (features)

Les mesures de qualité de l'air :

1. **CO AQI Value** : Indice de qualité pour le monoxyde de carbone
2. **Ozone AQI Value** : Indice de qualité pour l'ozone (O3)
3. **NO2 AQI Value** : Indice de qualité pour le dioxyde d'azote
4. **PM2.5 AQI Value** : Indice pour les particules fines (inférieur à 2.5 micromètres)

Informations géographiques

5. **Country** : Pays
6. **City** : Ville
7. **Latitude / Longitude** : Coordonnées GPS

Variable cible (target)

- **AQI Value** : Air Quality Index (0-500)
- **AQI Category** : Catégorie de qualité
    - Good (0-50)
    - Moderate (51-100)
    - Unhealthy for Sensitive Groups (101-150)
    - Unhealthy (151-200)
    - Very Unhealthy (201-300)
    - Hazardous (301+)

Exercices

Partie 1 : Chargement et Exploration des Données

Exercice 1.1 : Chargement des données

1. Téléchargez le dataset depuis Kaggle
    - URL : https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset
    - Ou utilisez le script : python download_kaggle_air_data.py
2. Chargez les données dans un DataFrame pandas
3. Affichez les premières lignes et les informations générales

Questions :

- Combien d'observations avez-vous ?
- Combien de pays et de villes sont représentés ?
- Y a-t-il des valeurs manquantes ?
- Quels sont les types de données de chaque colonne ?

Exercice 1.2 : Analyse statistique descriptive

1. Calculez les statistiques descriptives (moyenne, médiane, écart-type, min, max)
2. Analysez la distribution de la variable cible AQI Value
3. Analysez la distribution des catégories AQI Category
4. Identifiez les valeurs aberrantes (outliers) potentielles

Questions :

- Quelle est la distribution de l'AQI ? Est-elle équilibrée ?
- Quelles sont les villes les plus polluées ?
- Quels pays ont la meilleure/pire qualité d'air en moyenne ?
- Y a-t-il des valeurs qui semblent anormales ?

Exercice 1.3 : Visualisation des données

Créez les visualisations suivantes :

1. **Histogrammes** : Distribution de l'AQI et des polluants
2. **Boxplots** : Détection des outliers par polluant
3. **Barplots** : Distribution des catégories AQI
4. **Heatmap** : Matrice de corrélation entre polluants
5. **Carte géographique** (bonus) : Visualisation mondiale de la pollution

Questions :

- Quels polluants sont fortement corrélés entre eux ?
- Quelles régions du monde sont les plus touchées ?
- Observez-vous des patterns géographiques intéressants ?

Partie 2 : Préparation des Données

Exercice 2.1 : Nettoyage des données

1. Gérez les valeurs manquantes (si présentes)
2. Traitez les outliers identifiés (décision à justifier)
3. Vérifiez la cohérence des données

Exercice 2.2 : Feature Engineering

1. Transformez la variable AQI Category en problème de classification binaire :
    - **Bon air** : Good + Moderate (AQI < 100)
    - **Mauvais air** : Unhealthy for Sensitive Groups et pire (AQI supérieur ou égal à 100)
2. Cette transformation est pertinente car AQI=100 est le seuil d'alerte sanitaire

3. Créez éventuellement de nouvelles features pertinentes :
    - Moyenne des polluants

- Polluant dominant
- Indicateurs géographiques (continent, région)

Questions :

- Pourquoi transformer le problème en classification binaire ?
- Quel est l'impact sur la distribution des classes ?
- Le seuil de 100 AQI est-il pertinent du point de vue santé publique ?

Exercice 2.3 : Séparation des données

1. Séparez les features (X) et la target (y)
2. Divisez le dataset en ensembles d'entraînement (70%) et de test (30%)
3. Utilisez random_state=42 pour la reproductibilité

Important : Vérifiez que la distribution des classes est similaire dans train et test (stratification)

Exercice 2.4 : Normalisation

1. Standardisez les features numériques avec StandardScaler
2. Appliquez la transformation sur train et test (attention au data leakage)

Question :

- Pourquoi est-il important de normaliser les données ?
- Pourquoi fitter le scaler uniquement sur le train set ?

Partie 3 : Modélisation

Exercice 3.1 : Modèle de référence (Baseline)

1. Implémentez un classificateur naïf (prédiction de la classe majoritaire)
2. Calculez l'accuracy sur le test set

Question :

- Pourquoi est-il important d'avoir une baseline ?

Exercice 3.2 : Régression Logistique

1. Entraînez un modèle de Régression Logistique
2. Prédisez sur l'ensemble de test
3. Calculez les métriques suivantes :
    - Accuracy
    - Precision, Recall, F1-Score
    - Matrice de confusion

- Courbe ROC et AUC

Questions :

- Le modèle performe-t-il mieux que le baseline ?
- Quelle métrique est la plus pertinente pour ce problème de santé publique ?
- Que nous apprend la matrice de confusion ?
- Quel type d'erreur est le plus grave (faux positif ou faux négatif) ?

Exercice 3.3 : K-Nearest Neighbors (KNN)

1. Entraînez un modèle KNN avec k=5
2. Évaluez avec les mêmes métriques que précédemment
3. Testez différentes valeurs de k (3, 5, 7, 10, 15)
4. Tracez la courbe de performance en fonction de k

Questions :

- Quelle est la meilleure valeur de k ?
- Comment évolue la performance avec k ?
- KNN performe-t-il mieux que la Régression Logistique ?

Exercice 3.4 : Arbre de Décision

1. Entraînez un arbre de décision
2. Visualisez l'arbre (limité à 3 niveaux de profondeur)
3. Affichez l'importance des features
4. Évaluez les performances

Questions :

- Quels polluants sont les plus importants pour prédire la qualité de l'air ?
- L'arbre est-il interprétable ?
- Observez-vous du surapprentissage ?

Exercice 3.5 : Random Forest

1. Entraînez une Random Forest avec 100 arbres
2. Affichez l'importance des features
3. Évaluez les performances

Questions :

- Random Forest améliore-t-il les performances par rapport à un seul arbre ?
- Les features importantes sont-elles les mêmes qu'avec l'arbre simple ?
- Quels polluants contribuent le plus à la mauvaise qualité de l'air ?

Exercice 3.6 : Support Vector Machine (SVM)

1. Entraînez un SVM avec noyau RBF
2. Évaluez les performances

Question :

- Comment se compare le SVM aux autres modèles ?

Partie 4 : Optimisation et Comparaison

Exercice 4.1 : Optimisation des hyperparamètres

Choisissez votre meilleur modèle et optimisez ses hyperparamètres avec GridSearchCV :

Exemple pour Random Forest :

```
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

Questions :

- Quels sont les meilleurs hyperparamètres trouvés ?
- Quelle amélioration de performance obtenez-vous ?

Exercice 4.2 : Validation croisée

1. Évaluez tous vos modèles avec une validation croisée 5-fold
2. Comparez les scores moyens et écarts-types

Question :

- Quel modèle est le plus stable ?

Exercice 4.3 : Tableau comparatif

Créez un tableau récapitulatif de tous vos modèles avec :

- Accuracy (train et test)
- Precision, Recall, F1-Score (test)
- AUC-ROC (test)
- Temps d'entraînement

Question :

- Quel est le meilleur modèle selon vous ? Justifiez en tenant compte du contexte de santé publique.

Partie 5 : Analyse et Interprétation

Exercice 5.1 : Analyse des erreurs

1. Identifiez les villes mal classifiées par votre meilleur modèle
2. Analysez leurs caractéristiques
3. Proposez des hypothèses sur les causes d'erreur

Questions :

- Quelles villes sont difficiles à classifier ?
- Y a-t-il des patterns géographiques dans les erreurs ?

Exercice 5.2 : Importance des features

1. Analysez l'importance des features de votre meilleur modèle
2. Créez une visualisation claire
3. Interprétez les résultats du point de vue environnemental

Questions :

- Quels polluants sont les plus déterminants pour la qualité de l'air ?
- Ces résultats sont-ils cohérents avec les connaissances scientifiques ?
- Quelles recommandations en tirer pour les politiques publiques ?

Exercice 5.3 : Recommandations environnementales

Rédigez un court rapport (1 page) avec :

1. **Résumé** : Objectif et approche
2. **Résultats** : Performances du meilleur modèle
3. **Insights** : Facteurs clés de la pollution de l'air
4. **Impact santé** : Implications pour la santé publique
5. **Recommandations** : Actions pour améliorer la qualité de l'air

Bonus

Bonus 1 : Analyse géographique

Comparez les performances des modèles par continent ou région.

Bonus 2 : Classification multi-classes

Reprenez le problème avec toutes les catégories AQI (6 classes).

Bonus 3 : Analyse temporelle

Si les données temporelles sont disponibles, analysez l'évolution de la pollution.

Bonus 4 : Prédiction de l'AQI

Transformez le problème en régression pour prédire la valeur exacte de l'AQI.

Bonus 5 : Visualisation géographique

Créez une carte interactive montrant la qualité de l'air mondiale.

Bonus 6 : Deep Learning

Implémentez un réseau de neurones simple avec Keras/TensorFlow.

Livrables Attendus

1. **Notebook Jupyter** complété avec :
    - Code commenté et structuré
    - Visualisations claires et légendées
    - Analyses et interprétations
    - Réponses aux questions
2. **Export PDF** du notebook

3. **Rapport de recommandations environnementales** (1 page)

Critères de Réussite

- Toutes les parties 1 à 4 sont complétées
- Le code est fonctionnel et bien commenté
- Les visualisations sont pertinentes et lisibles
- Les métriques sont correctement calculées et interprétées
- Le meilleur modèle atteint une accuracy supérieure à 75% sur le test set
- Les recommandations environnementales sont argumentées et pertinentes

Conseils Méthodologiques

Bonnes Pratiques

1. **Organisation** : Structurez votre notebook en sections claires
2. **Commentaires** : Expliquez vos choix et vos observations
3. **Visualisations** : Utilisez des graphiques pour illustrer vos propos
4. **Reproductibilité** : Fixez les random_state
5. **Validation** : Vérifiez vos résultats à chaque étape
6. **Contexte** : Pensez toujours à l'impact santé publique

Pièges à Éviter

- Data leakage (normaliser avant de split)
- Overfitting (vérifier train vs test)
- Ignorer le déséquilibre des classes
- Choisir uniquement l'accuracy comme métrique
- Ne pas interpréter les résultats dans le contexte environnemental

Ressources Complémentaires

Contexte Environnemental

- OMS - Qualité de l'air : https://www.who.int/fr/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health
- EPA - Air Quality Index : https://www.airnow.gov/aqi/aqi-basics/
- European Environment Agency : https://www.eea.europa.eu/themes/air

Documentation ML

- Scikit-learn Classification : https://scikit-learn.org/stable/tutorial/statistical_inference/supervised_learning.html
- Guide to Model Evaluation : https://scikit-learn.org/stable/modules/model_evaluation.html

Datasets similaires

- UCI Air Quality Dataset
- OpenAQ - Global Air Quality Data
- WHO Global Air Quality Database

Impact de votre Travail

En réalisant ce TD, vous contribuez à :

- **Protéger la santé publique** en développant des outils de prédiction
- **Aider la prise de décision** politique sur l'environnement
- **Alerter les populations** lors de pics de pollution
- **Sensibiliser** aux enjeux environnementaux

Votre travail a du sens

Bon courage et bon apprentissage