

TD Machine Learning - Master 1

Prédiction de la Qualité de l'Air Mondiale

Air Quality Index (AQI)

0-50	51-100	101-150	151-200	201-300	301
Good	Moderate	Unhealthy for Sensitive Groups	Unhealthy	Very Unhealthy	Hazardous

Informations Générales

Prérequis : Python, pandas, numpy, notions de statistiques

Objectifs Pédagogiques

À l'issue de ce TD, vous serez capable de :

1. **Comprendre et appliquer** le pipeline complet d'un projet Machine Learning
2. **Analyser et préparer** des données réelles pour l'apprentissage automatique
3. **Implémenter et comparer** plusieurs algorithmes de classification
4. **Évaluer** la performance des modèles avec des métriques appropriées
5. **Optimiser** les hyperparamètres d'un modèle
6. **Interpréter** les résultats et formuler des recommandations

Contexte du Projet

Problématique Environnementale

La pollution de l'air est un enjeu majeur de santé publique mondiale. Selon l'OMS, 99% de la population mondiale respire un air dont la qualité dépasse les limites recommandées, causant 7 millions de décès prématurés par an.

Vous travaillez pour une organisation environnementale internationale qui souhaite développer un système de **prédiction et classification de la qualité de l'air** pour alerter les populations et aider à la prise de décision politique.

Dataset

Vous utiliserez le **Global Air Pollution Dataset** de Kaggle :

- **Source** : <https://www.kaggle.com/datasets/hasibalmuzdadid/global-air-pollution-dataset>
- **Période** : 2017-2022 (données récentes)
- **Contenu** : Plus de 23,000 villes dans le monde
- **Variables** : PM2.5, PM10, NO2, CO, O3, AQI (Air Quality Index)
- **Taille** : Environ 50 MB (idéal pour le TD)

Structure du Projet

TD/

```
|── TD3.pdf           (Ce fichier)  
|── sujetdetaille.pdf   (Énoncé détaillé des exercices)  
|── notebook_etudiant.ipynb (Notebook à compléter)  
└── requirements.txt     (Dépendances Python)
```

Installation

1. Créer un environnement virtuel (recommandé)

```
python -m venv venv
```

```
source venv/bin/activate # Sur Windows: venv\Scripts\activate
```

2. Installer les dépendances

```
pip install -r requirements.txt
```

Ressources

Documentation Officielle

- Scikit-learn : <https://scikit-learn.org/stable/>
- Pandas : <https://pandas.pydata.org/docs/>
- Matplotlib : <https://matplotlib.org/stable/contents.html>
- Seaborn : <https://seaborn.pydata.org/>

Consignes

- 1. Travail** : Individuel ou en binôme
- 2. Rendu** : Notebook Jupyter complété avec vos analyses et commentaires
- 3. Format** : .ipynb + export PDF

Conseils

- Prenez le temps de bien comprendre les données avant de modéliser
- Commentez votre code et vos choix méthodologiques
- N'hésitez pas à explorer au-delà des questions posées
- Pensez à la reproductibilité (random_state)
- Documentez vos résultats avec des visualisations claires

Bon travail et bon apprentissage 😊