



***“Analyse de Clustering et Prévission ML pour la  
Caractérisation des Schémas de Pollution  
Atmosphérique et des Conditions Météorologiques”***

**Réaliser par:**

- DAGHMOUMI Marouan

**Encadrer par:**

- Pr. CHRIT Salma

***Année universitaire 2023/2024***

---

# Table des matières

---

<b>Table des matières</b>	<b>2</b>
<b>Tables des Figures</b>	<b>3</b>
<b>CHAPITRE 1 : PRESENTATION DATASET .</b>	<b>4</b>
1. DESCRIPTION DES DONNÉES :	5
2. OBJECTIF DE L'ÉTUDE :	5
3. ANALYSE DU JEU DE DONNÉES :	5
3.1 Description générale du jeu de données:	5
3.2 Statistiques descriptives:	6
3.3 données manquantes :	6
<b>CHAPITRE 2 : PRETRAITEMENT DE DONNEES .</b>	<b>7</b>
1. SUPPRESSION DES COLONNES INUTILES :	8
2. ENCODAGE DES VARIABLE CATEGORIELLES :	8
3. NORMALISATION DES DONNEES :	9
4. GESTION DES MANQUANTES :	9
<b>CHAPITRE 3 : APPRENTISSAGE AUTOMATIQUE .</b>	<b>10</b>
1. INTRODUCTION :	11
2. APPRENTISSAGE NON SUPERVISE : KMeans	11
2. APPRENTISSAGE SUPERVISE :	11
2.1 K plus proches voisins (KNN) :	11
2.2 Support Vector Machine (SVM) :	12
2.3 Arbre de décision (DT):	12
3. COMPARAISON DES PERFORMANCES :	14
<b>Conclusion Générale</b>	<b>15</b>

---

# Tables des Figures

---

Figure 1: Résultat d'exécution df.info() .....	5
Figure 2: Résultat d'exécution df.describe().....	6
Figure 3: Résultat d'exécution df.isnull().sum().....	6
Figure 4: Visualisation les valeurs de colonne Station .....	8
Figure 5: Visualisation les valeurs de colonne wd .....	8
Figure 6: Visualisation les valeurs manquantes .....	9
Figure 7: Résultat de visualisation .....	11
Figure 8: matrice de confusion KNN .....	12
Figure 9: matrice de confusion SVM .....	12
Figure 10: matrice de confusion DT .....	13
Figure 11: DT .....	13

# **CHAPITRE 1 : PRESENTATION DATASET .**

# 1. DESCRIPTION DES DONNÉES :

Le jeu de données utilisé dans cette étude est intitulé "**DataMeteo12Complet.csv**" et comprend des mesures de pollution atmosphérique ainsi que des variables météorologiques collectées à partir d'une station de surveillance spécifique. Les données sont enregistrées à différentes heures pour chaque jour sur une période de plusieurs années.

# 2. OBJECTIF DE L'ÉTUDE :

L'objectif principal de cette analyse est d'explorer les tendances de la pollution atmosphérique et de comprendre les facteurs météorologiques associés. Pour ce faire, nous utiliserons des techniques d'apprentissage automatique telles que le clustering et les méthodes supervisées pour identifier les schémas de pollution et les conditions météorologiques dominantes dans les données.

# 3. ANALYSE DU JEU DE DONNÉES :

Pour mieux comprendre la structure et les caractéristiques du jeu de données "**DataMeteo12Complet.csv**", nous avons effectué une analyse exploratoire. Voici les principales observations :

## 3.1 DESCRIPTION GÉNÉRALE DU JEU DE DONNÉES:

Le jeu de données se compose de **35 064 entrées** et **18 colonnes**. Les colonnes représentent différentes variables telles que la date et l'heure des mesures, les niveaux de pollution atmosphérique (**PM2.5, PM10, SO2, NO2, CO, O3**) ainsi que les variables météorologiques (**température, pression, humidité, pluie, vitesse du vent, direction du vent**)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35064 entries, 0 to 35063
Data columns (total 18 columns):
 #   Column      Non-Null Count  Dtype  
---  --
 0   No          35064 non-null  int64  
 1   year        35064 non-null  int64  
 2   month       35064 non-null  int64  
 3   day         35064 non-null  int64  
 4   hour        35064 non-null  int64  
 5   PM2.5       34436 non-null  float64 
 6   PM10        34624 non-null  float64 
 7   SO2         34618 non-null  float64 
 8   NO2         34372 non-null  float64 
 9   CO          33858 non-null  float64 
10   O3          34558 non-null  float64 
11   TEMP        35044 non-null  float64 
12   PRES        35044 non-null  float64 
13   DEWP        35044 non-null  float64 
14   RAIN        35044 non-null  float64 
15   wd          34986 non-null  object  
16   WSPM        35050 non-null  float64 
17   station     35064 non-null  object  
dtypes: float64(11), int64(5), object(2)
memory usage: 4.8+ MB
```

Figure 1: Résultat d'exécution `df.info()`

### 3.2 STATISTIQUES DESCRIPTIVES:

- Les mesures de pollution atmosphérique présentent des valeurs moyennes et des écarts types variables. Par exemple, la concentration moyenne de PM2.5 est de 84.84  $\mu\text{g}/\text{m}^3$  avec un écart type de 86.23  $\mu\text{g}/\text{m}^3$ .
- Les variables météorologiques comme la température et la pression ont également des moyennes et des écarts types significatifs.

	PM2.5	PM10	SO2	NO2	CO
count	34436.000000	34624.000000	34618.000000	34372.000000	33858.000000
mean	84.838483	108.991096	18.689242	58.097172	1324.350198
std	86.225344	95.341177	24.280665	36.297740	1245.166124
min	2.000000	2.000000	0.571200	2.000000	100.000000
25%	22.000000	38.000000	3.000000	29.000000	500.000000
50%	59.000000	85.000000	9.000000	51.000000	900.000000
75%	116.000000	149.000000	23.000000	80.000000	1600.000000
max	844.000000	995.000000	257.000000	273.000000	10000.000000

Figure 2: Résultat d'exécution `df.describe()`

### 3.3 DONNÉES MANQUANTES :

Nous avons identifié des données manquantes dans plusieurs colonnes, notamment pour les variables de pollution atmosphérique (**PM2.5**, **PM10**, **SO2**, **NO2**, **O3**) et certaines variables météorologiques (**TEMP**, **RAIN**, **WSPM**). Voici le nombre de valeurs manquantes pour chaque variable :

- PM2.5 : 628
- PM10 : 440
- SO2 : 446
- NO2 : 692
- O3 : 506
- TEMP : 20
- RAIN : 20
- WSPM : 14

Nous envisagerons des stratégies de gestion des données manquantes lors de l'étape de prétraitement.

PM2.5	628
PM10	440
SO2	446
NO2	692
O3	506
TEMP	20
RAIN	20
WSPM	14
wd_numeric	78
PRES_scaled	20
CO_scaled	1206
DEWP_scaled	20

Figure 3: Résultat d'exécution `df.isnull().sum()`

# **CHAPITRE 2 : PRETRAITEMENT DE DONNEES .**

## 1. SUPPRESSION DES COLONNES INUTILES :

Nous avons identifié des colonnes qui ne contribuent pas à notre analyse et les avons supprimées du jeu de données. Plus précisément :

- La colonne "station" ne contient qu'une seule valeur pour toutes les entrées, elle a donc été supprimée.

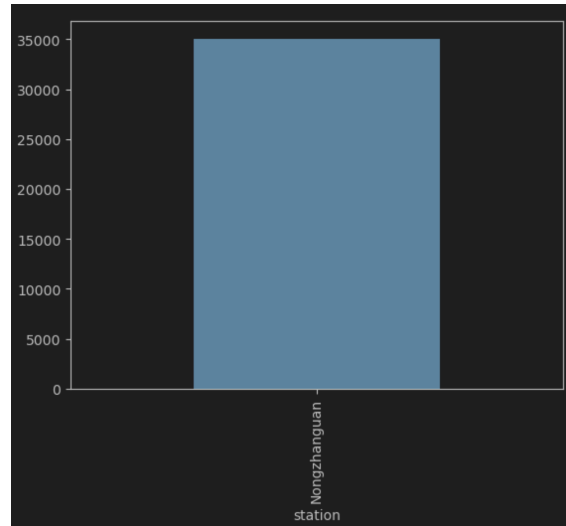


Figure 4: Visualisation les valeurs de colonne Station

- La colonne "No" qui représente un index monotone a également été supprimée.

Après suppression de ces colonnes, nous avons poursuivi avec les étapes suivantes de prétraitement.

## 2. ENCODAGE DES VARIABLE CATEGORIELLES :

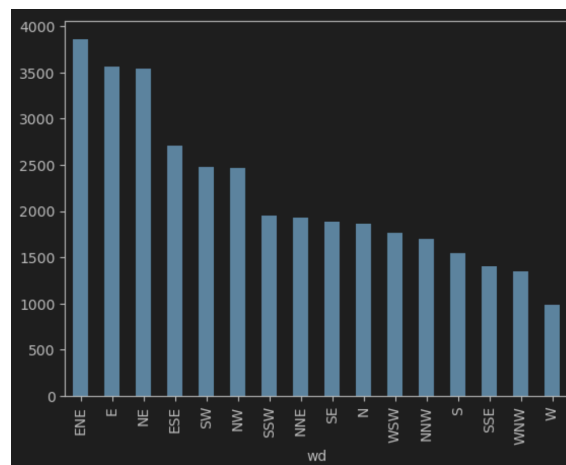


Figure 5: Visualisation les valeurs de colonne wd

Les variables de "wd" a été encodée en variables numériques en utilisant une méthode de mapping. Les valeurs de la direction du vent ont été remplacées par des valeurs numériques correspondant à leurs directions respectives.

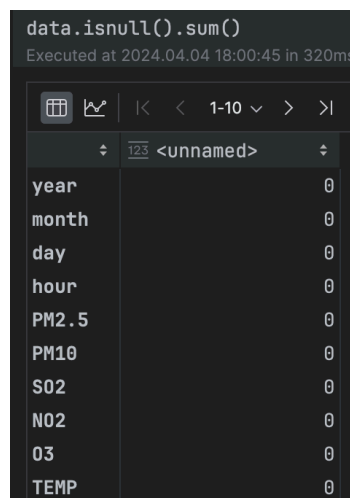


### **3. NORMALISATION DES DONNEES :**

Nous avons normalisé les variables "PRES" (pression atmosphérique) et "CO" (concentration de monoxyde de carbone) en utilisant la méthode de mise à l'échelle MinMaxScaler. Cela permet de ramener ces variables à une plage commune et faciliter leur comparaison.

### **4. GESTION DES MANQUANTES :**

Nous avons rempli les valeurs manquantes dans le jeu de données en utilisant la méthode de l'imputation par la moyenne. Cela garantit que les données manquantes ne compromettent pas l'analyse ultérieure



The screenshot shows a Jupyter Notebook interface with a code cell containing the command `data.isnull().sum()`. Below the code, the execution time is noted as "Executed at 2024.04.04 18:00:45 in 320ms". A table displays the results, showing the count of missing values for each variable. All variables listed have a count of 0, indicating no missing values.

year	0
month	0
day	0
hour	0
PM2.5	0
PM10	0
S02	0
N02	0
O3	0
TEMP	0

Figure 6: Visualisation les valeurs manquantes

# **CHAPITRE 3 : APPRENTISSAGE AUTOMATIQUE .**

## 1. INTRODUCTION :

Dans ce chapitre, nous décrivons en détail les différentes techniques d'apprentissage automatique que nous avons appliquées à nos données de pollution atmosphérique et de variables météorologiques. Chaque méthode est présentée avec ses résultats spécifiques.

## 2. APPRENTISSAGE NON SUPERVISE : KMeans

Nous avons utilisé l'algorithme de clustering Kmeans pour regrouper les observations en 3 clusters distincts. Voici les étapes que nous avons suivies :

- Initialisation de l'algorithme avec 3 clusters et une graine aléatoire de 42.

```
kmeans = KMeans(n_clusters=3, random_state=42)
```

- Entraînement du modèle sur les données.

```
kmeans.fit(data)
```

- Attribution des étiquettes de cluster à chaque observation.

```
data['Target'] = kmeans.labels_
```

- Visualisation de la répartition des clusters à l'aide d'un histogramme.

```
data["Target"].value_counts().plot(kind='bar')  
plt.hist(data["Target"], bins='auto')  
plt.show()
```

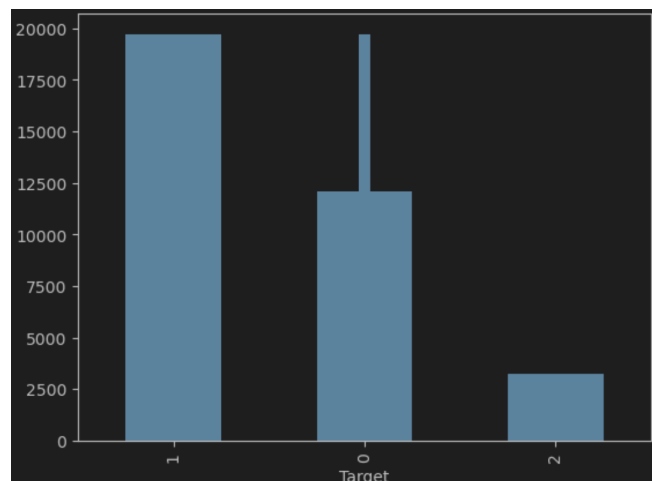


Figure 7: Résultat de visualisation

## 2. APPRENTISSAGE SUPERVISE :

### 2.1 K PLUS PROCHES VOISINS (KNN) :

Nous avons ensuite utilisé l'algorithme des K plus proches voisins (KNN) pour la classification des données. Voici les étapes que nous avons suivies :

- Séparation des données en ensembles d'entraînement et de test.

```
x_train, x_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

- Entraînement du modèle KNN avec 5 voisins sur l'ensemble d'entraînement.

```
knn = KNeighborsClassifier(n_neighbors=5)
```

- Prédiction des étiquettes de classe sur l'ensemble de test.

```
y_pred_KNN = knn.predict(X_test)
```

- Construction et affichage de la matrice de confusion pour évaluer les performances du modèle.

```
Accuracy: 0.9867389134464566
```

↕	123 0	↕	123 1	↕	123 2	↕
0	2359		62		16	
1	58		3911		0	
2	9		0		598	

Figure 8: matrice de confusion KNN

## 2.2 SUPPORT VECTOR MACHINE (SVM) :

Nous avons également utilisé une machine à vecteurs de support (SVM) avec un noyau linéaire pour la classification des données. Voici les étapes que nous avons suivies :

- Entraînement du modèle SVM sur l'ensemble d'entraînement.

```
svm = SVC(kernel='linear')
```

- Prédiction des étiquettes de classe sur l'ensemble de test.

```
Accuracy: 0.9992870383573363
```

- Construction et affichage de la matrice de confusion pour évaluer les performances du modèle.

↕	123 0	↕	123 1	↕	123 2	↕
0	2434		2		1	
1	2		3967		0	
2	0		0		607	

Figure 9: matrice de confusion SVM

## 2.3 ARBRE DE DÉCISION (DT):

Enfin, nous avons utilisé un arbre de décision pour classifier les données. Voici les étapes que nous avons suivies :

- Entraînement du modèle d'arbre de décision sur l'ensemble d'entraînement.

```
DT = DecisionTreeClassifier()
```

- Prédiction des étiquettes de classe sur l'ensemble de test.

```
y_pred_dt = DT.predict(X_test)
```

- Calcul de la précision du modèle.

```
Accuracy: 0.9793241123627549
```

- Construction et affichage de l'arbre de décision pour visualiser les règles de classification.

↕	123 0	↕	123 1	↕	123 2	↕
0	2359		62		16	
1	58		3911		0	
2	9		0		598	

Figure 10: matrice de confusion DT

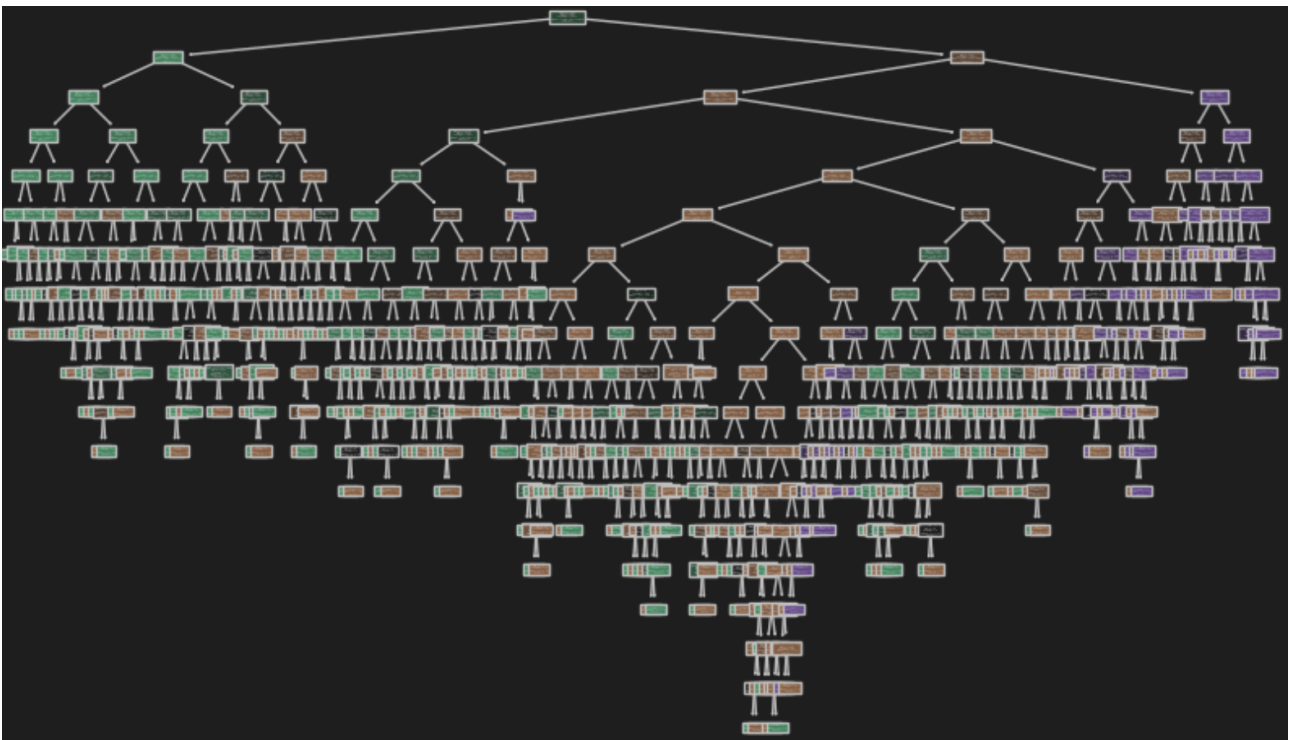


Figure 11: DT

### **3. COMPARAISON DES PERFORMANCES :**

En comparant les performances des différents modèles, nous constatons que le **SVM linéaire** affiche la meilleure précision avec **99.93%**, suivi par **KNN** avec **98.67%**. **L'arbre de décision** se situe légèrement en dessous avec une précision de **97.93%**. Cette comparaison démontre que les modèles d'apprentissage automatique sont capables de fournir des résultats robustes pour la prédiction des schémas de pollution atmosphérique et des conditions météorologiques, avec des performances variables en fonction de l'algorithme utilisé.

---

# Conclusion Générale

---

Ce rapport a examiné les tendances de la pollution atmosphérique et des variables météorologiques associées en utilisant des techniques d'analyse de données et d'apprentissage automatique. Voici les points essentiels :

- **Analyse des Données:** Nous avons commencé par présenter en détail l'ensemble de données, mettant en évidence les mesures de pollution atmosphérique et les variables météorologiques recueillies sur plusieurs années. Cette analyse a permis une meilleure compréhension de la structure et de la distribution des données.
- **Prétraitement des Données:** nous avons effectué un prétraitement des données, comprenant le nettoyage des données, le traitement des valeurs manquantes, la normalisation et la transformation des variables catégorielles en valeurs numériques.
- **L'Apprentissage Automatique:** Nous avons appliqué plusieurs techniques d'apprentissage automatique, dont **Kmeans**, **KNN**, **SVM** et **DT**, pour explorer les schémas de pollution atmosphérique. Chaque modèle a apporté des perspectives uniques sur les relations entre les variables et la prédiction des niveaux de pollution.
- **Résultats et Comparaison:** La comparaison des performances des modèles a révélé des différences significatives. Le **SVM** linéaire a obtenu la meilleure précision, suivi de près par **KNN**, tandis que **l'arbre de décision** a également fourni des résultats solides. Cette comparaison a souligné l'importance de choisir le bon algorithme en fonction des caractéristiques des données.

En conclusion, cette étude a démontré le potentiel des techniques d'analyse de données et d'apprentissage automatique pour mieux comprendre et prédire les schémas de pollution atmosphérique. Ces résultats peuvent être utilisés pour informer les décideurs et contribuer à la protection de la santé publique et de l'environnement.