Comparaison entre méthodes de Machine Learning et Deep Learning pour prédire l'indice de qualité de l'air à partir de données météo

La qualité de l'air est devenue un enjeu majeur pour les grandes villes, avec des impacts significatifs sur la santé publique et l'environnement. Face à l'augmentation des sources de pollution et à l'évolution rapide des conditions météorologiques, prédire l'indice de qualité de l'air est un défi essentiel pour anticiper les pics de pollution et informer les populations. Ce projet s'appuie sur des données météorologiques et environnementales pour améliorer la prévision des indices de qualité de l'air.

Le jeu de données utilisé pour cette étude contient l'évolution de six facteurs météorologiques et six facteurs de pollution mesurés toutes les heures dans 12 stations à Pékin entre 2013 et 2017. Les facteurs météorologiques mesurés sont : la température de l'air (TEMP en °C), la pression atmosphérique (PRES en hPa), la température de rosée (DEWP en °C), les précipitations (RAIN en mm), la direction du vent (wd), la vitesse du vent (WSPM en $m.s^{-1}$). Les polluants atmosphériques mesurés (en concentration) sont : les particules fines dont le diamètre est inférieur à 2.5 microns (PM2.5 en $\mu g.m^{-3}$), les particules fines dont le diamètre est inférieur à 10 microns (PM10 en $\mu g.m^{-3}$), le dioxyde de sodium (SO2 en $\mu g.m^{-3}$), le dioxyde d'azote (NO2 en $\mu g.m^{-3}$), le monoxyde de carbone (CO en $\mu g.m^{-3}$), l'ozone (O3 en $\mu g.m^{-3}$).

Notre sujet s'inscrit dans une approche *One Health* en comparant les performances de méthodes de machine learning et de deep learning pour la prédiction de l'indice de qualité de l'air et ainsi améliorer la gestion des risques liés à la pollution atmosphérique.

Nous comparerons trois modèles prédictifs : deux modèles de machine learning, Random Forest et ARIMA, ainsi qu'un modèle de deep learning, le réseau de neurones récurrent LSTM (Long Short-Term Memory). Chaque modèle sera évalué en termes de précision, robustesse, et capacité à capturer les dynamiques complexes des séries temporelles.

Les résultats attendus permettront d'identifier les points forts et les limites de chaque méthode pour orienter les choix futurs en fonction des besoins spécifiques en termes de précision et de performance de calcul.