Predicting Air Pollutant Concentrations in Beijing Using Regression Models

Ephrem Alemu Mehammed

ephrem.mehammed@tu-dortmunde.de

Md Rounak Jahan Raj

rounak.raj@tu-dortmund.de

TU Dortmund University

International Master of Advanced Methods in Particle Physics (IMAPP)

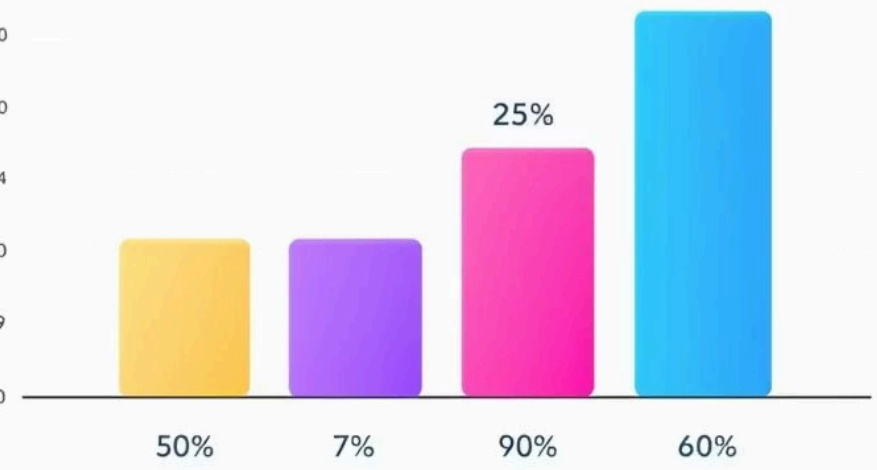
Problem

1 Define the Challenge

Given various meteorological variables, what are the predicted concentrations of air pollutants- PM10, SO_2, and NO_2 in Beijing?

2 Motivation

Air pollution is one of the most serious **environmental challenges** of our time. According to the World Health Organization, it contributes to around **8 million premature deaths** every year. Moreover, **99%** of the global population is exposed to air that does not meet WHO standards. In this project, we aim to **predict the levels of various air pollutants** using machine learning models, helping to better understand and manage air quality risks.



Data Set

Source	UC Irvine Machine Learning Repository
License	Creative Commons Attribution 4.0 International (CC BY 4.0)
Information	Hourly data of 6 air pollutants + 6 meteorological variables from Beijing stations.
Entries	420,768 rows x 17 columns
Important Features	"day", "hour", "TEMP", "PRES", "DEWP", "station", "WSPM"
Targets	"PM10", "SO_2" and "NO_2"
Previous Work	Prior work predicted "PM2.5".

Comparison with Alternative Methods

Dense Neural Network (DNN)

- Captures complex nonlinear relationships.
- Handles multi-target regression effectively.
- Suitable for moderate sized data samples.
- Excels with tabular data structures.

XGBRegressor

- Manages missing data and outliers well.
- Robust to preprocessing and normalization.
- Performs strongly on heterogeneous features.
- Highly effective for tabular datasets.