neue fische
School and Pool for Digital Talent

Machine Learning Project:
Urban Air Pollution

20.10.2025

Who are we?

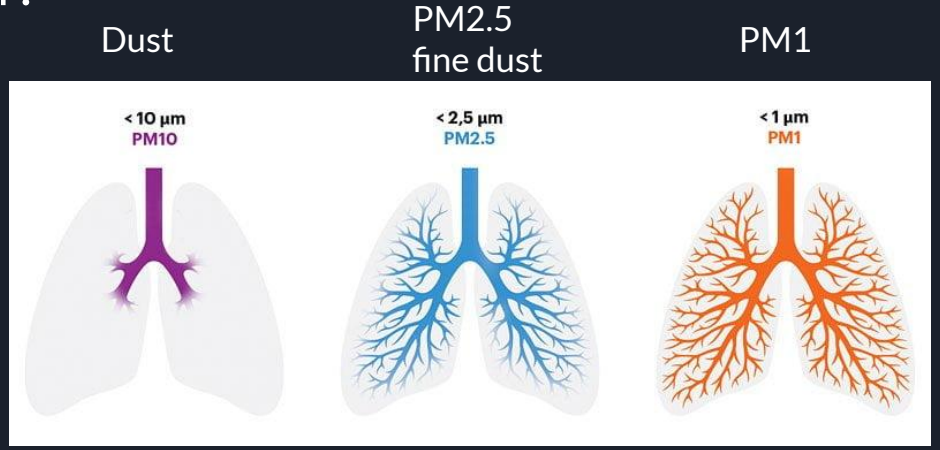
Our Goal:

"Predict air quality even in areas without ground sensors."

What is the problem?

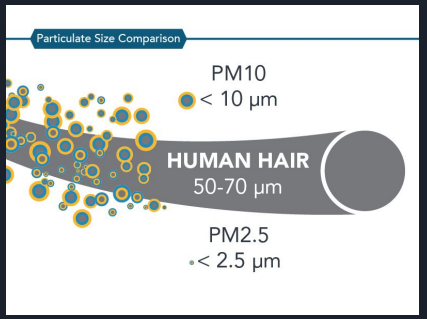
Imagine a clear blue sky - but is the air quality really as good as we think?

Fine dust can be invisible but can penetrate deep into our lungs and damage them.



PM2.5 = particulate matter concentration in **µg/m³**

Normally requires ground-based sensors → only available in cities



What do we want to achieve?

Predict PM2.5 concentration in µg/m3 from satellite data for industry gases (NO_2, O_3, CO) and weather data.

Industry gases

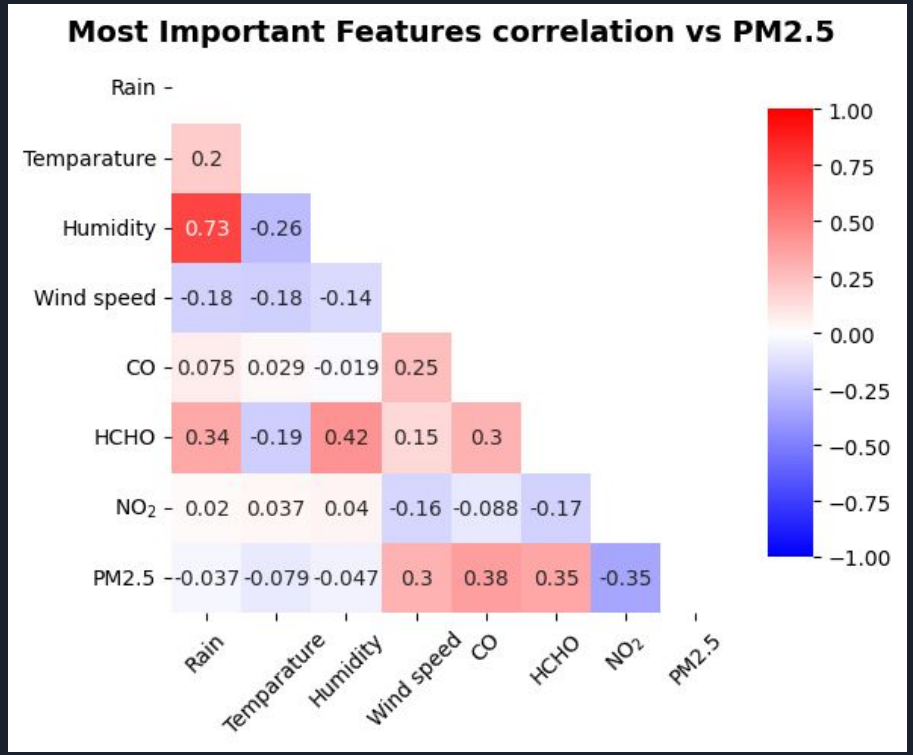
Wind

Humidity & Temperature

Rainfall

What data do we have?

- 30.5k data points
- Jan - April 2020
- humidity, temperature and wind speed from Global Forecast System (GFS)
- various pollutants in the atmosphere from Sentinel 5P satellite (NO_2, O_3, CO, HCHO, SO_2, CH_4)
- hundreds of cities worldwide
- PM2.5 concentration as target



Baseline model

Used features (3 out of 74)

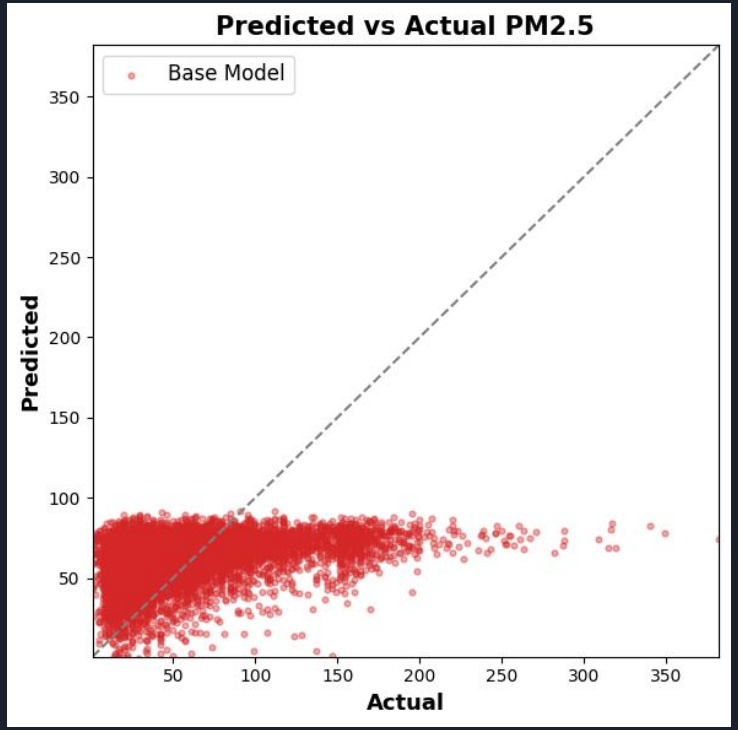
- amount of rainfall
- humidity 2m above ground
- wind speed

Model

- **Linear Regression** -> simple relationship between rain, wind, humidity and the PM2.5 concentration

Result

- RMSE: 42.49 → model has an average deviation of **42.49 µg/m3** from real measured PM2.5 concentration value
- **PM2.5 values** range from **0 to 500**

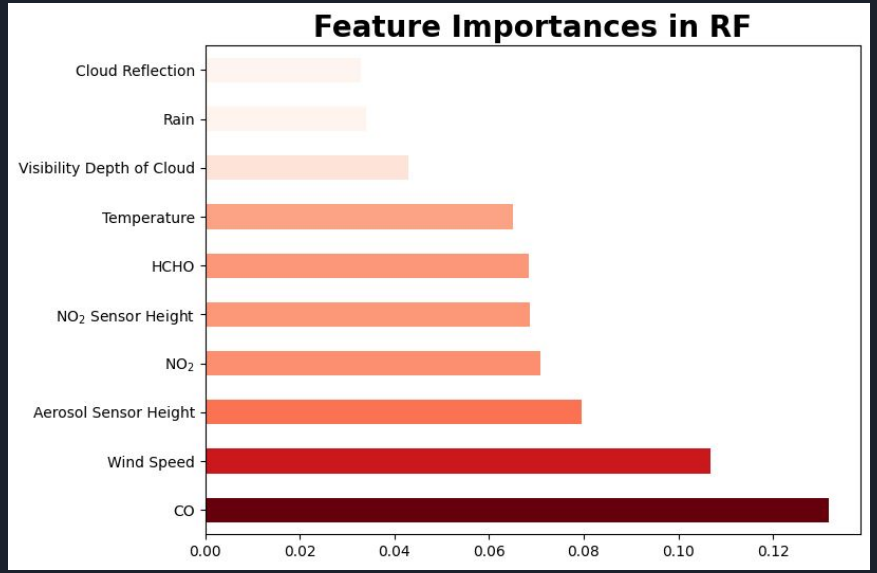


Optimized Air Pollution Model

- Using only weather and gases density data (20 features)
- **RandomForestRegressor** (best model out of 7) → combination of decision trees
- Hyperparameter tuning for best parameter estimation

→ RandomForestRegressor improved RMSE to **28.70**!

→ Prediction of PM2.5 concentration **highly improved!**



Expert Training & Blending Model

- **Categories** got created for:
  - Air Pollution Areas (CO and VOCs)
  - Climate zones
  - Wind speed Bins
  - Urban/Rural Areas

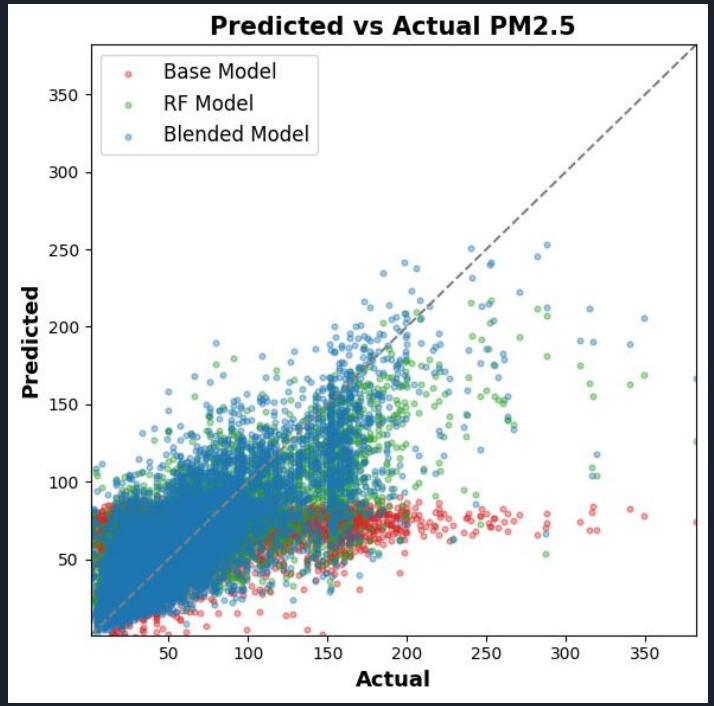
→ Blending LinearRegression and tuning the parameters improved RMSE to **27.6!**

→ Our model works like a team of experts!

Each focused on a different aspect of air quality (climate zoned, CO, NO_2, VOC, wind speed).

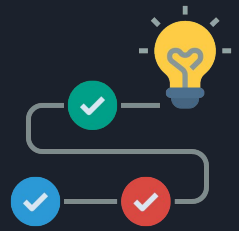
Finally a blending model that learns how to best combine their predictions.

→ Best Prediction of PM2.5 concentration in Zindi is **28.59!**



Conclusion and recommendations

- PM2.5 concentration can be predicted pretty well with weather and satellite data even for abandoned areas
- The CO concentration has the highest impact on PM2.5 concentration



- **Top ares combination of characteristics with best performance:**
  - Rural areas during winter time, with low air pollution and moderate wind speed

- **The model can be used for:**
  - Early warning systems / apps during smog- or heat waves
  - Improvement of urban planning like more green areas, more traffic calmed zones
  - Health improvement by avoiding specific areas on special weather conditions
  - Sustainable investment planning for rural areas

Further Work

- More data points
- Deeper EDA
- more models
- Include geographical data
- Build a Dashboard

Thank you!

…..and remember, if you want to be a “urban air hero” like us, use your bicycle more.

Sources

https://aqicn.org/sources/de/