BÁO CÁO ĐỒ ÁN CUỐI KỲ

Môn : Nhập môn Khoa học Dữ liệu

GV hướng dẫn

Thầy Trần Trung Kiên

Nhóm 10

Vương Thị Ngọc Linh – 18120195

Đặng Đỗ Huỳnh Như - 18120219

fit@hcmus

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

1 ĐỀ TÀI

2 LÝ DO CHỌN ĐỀ TÀI

3 THU THẬP DỮ LIỆU

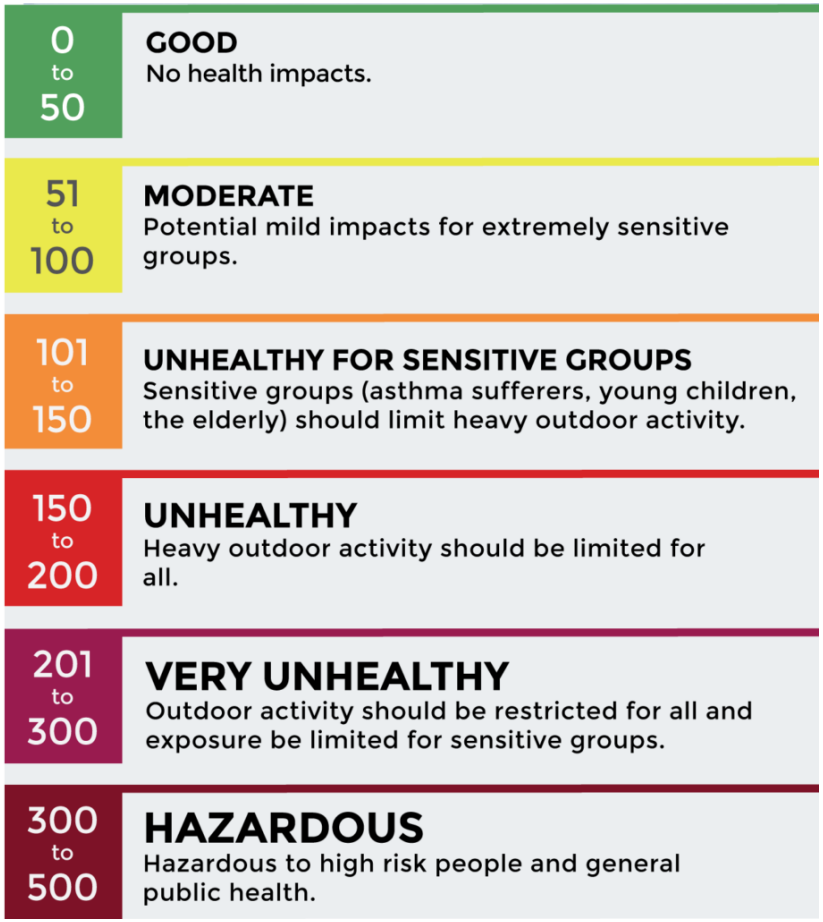
4 TIỀN XỬ LÝ

5 HUẤN LUYỆN MÔ HÌNH

6 TÀI LIỆU THAM KHẢO

1 ĐỀ TÀI

DỰ ĐOÁN CHẤT LƯỢNG KHÔNG KHÍ THEO CẤP ĐỘ



Chất lượng không khí được đánh giá thành 6 cấp độ tương ứng với các khoảng **chỉ số chất lượng không khí** (AQI), AQI được tính dựa trên **chỉ số ảnh hưởng đến chất lượng không khí**:

- *PM2.5* : Các hạt bụi có kích thước đường kính <= 2.5 micromet.
- *PM10* : Các hạt bụi có kích thước đường kính từ 2.5 tới 10 micromet.
- *O3, CO, SO2, NO2*.

Trong đề tài này, nhóm tiến hành dự đoán cấp độ chất lượng không khí dựa vào các **chỉ số ảnh hưởng đến chất lượng không khí**.

2 LÝ DO CHỌN ĐỀ TÀI

- Ngày nay, ô nhiễm không khí đang là một trong những vấn đề được quan tâm nhiều nhất. Ô nhiễm không khí đe dọa sức khỏe của người dân ở khắp mọi nơi trên thế giới.

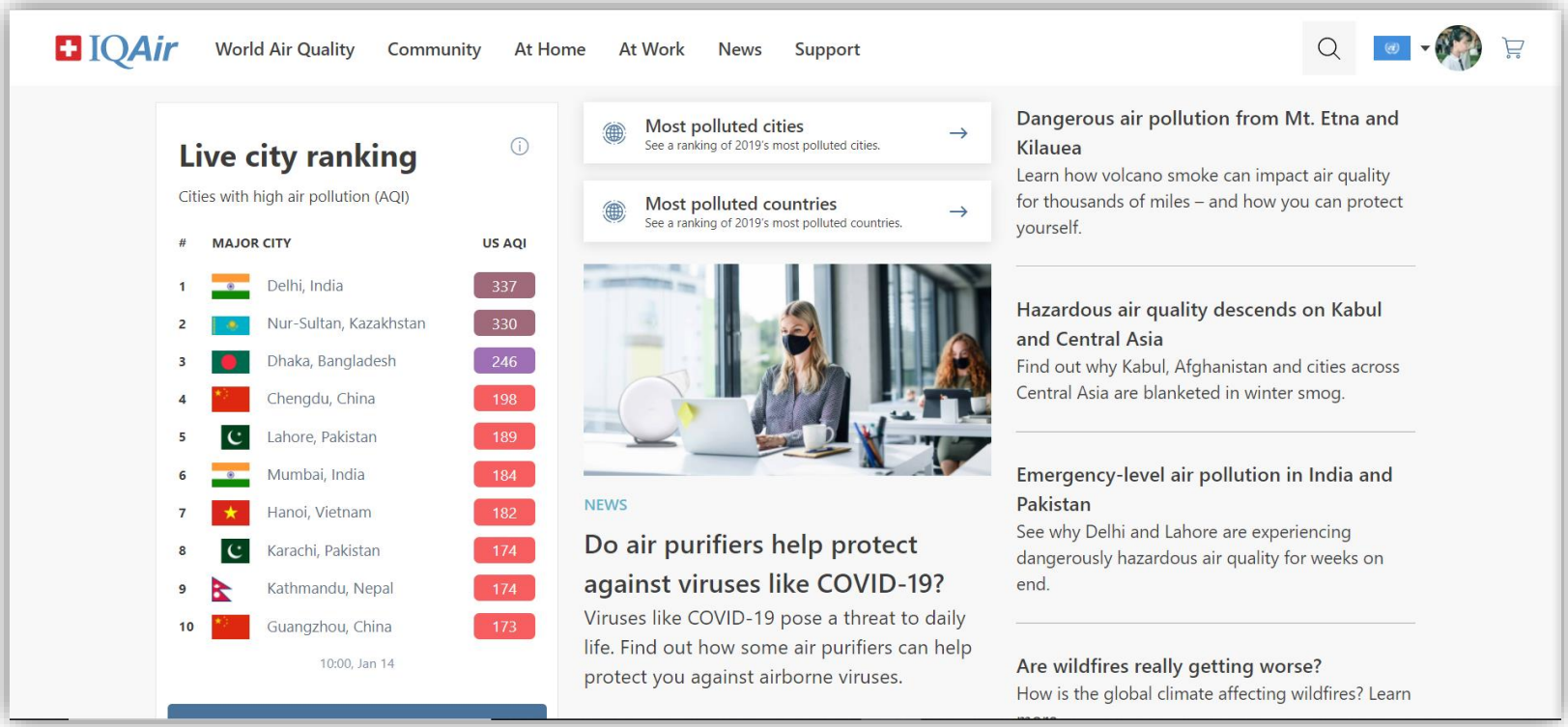
- Ước tính mới đây năm 2018 cho thấy rằng 9/10 người dân phải hít thở không khí chứa hàm lượng các chất gây ô nhiễm cao. Ô nhiễm không khí cả ở bên ngoài và trong nhà gây ra khoảng 7 triệu ca tử vong hàng năm trên toàn cầu.

- Ở Việt Nam, vào năm 2019, ứng dụng Airvisual do iqair phát hành khiến cộng đồng xôn xao về thực trạng chất lượng không khí đáng báo động ở một số thành phố.

- Cách tính AQI khá phức tạp và có sự khác nhau ở nhiều quốc gia, vì vậy nhóm chỉ dựa vào chỉ số PM2.5, PM10, O3, CO, SO2, NO2 để dự đoán mức độ chất lượng không khí.

3 THU THẬP DỮ LIỆU

- **Trang web thu thập dữ liệu:** https://www.iqair.com



- **Phương thức thu thập dữ liệu:** Parse HTML
- **Công cụ sử dụng** : Selenium Webdriver

3 THU THẬP DỮ LIỆU

- **Nội dung thu thập:**

	City	Level	PM2.5	PM10	O3	NO2	SO2	CO
0	Dhaka	Very Unhealthy	227.5	NaN	NaN	NaN	NaN	NaN
1	Delhi	Very Unhealthy	137.7	201.2	9.8	35.0	12.6	1181.7
2	Bishkek	Very Unhealthy	157.4	2.2	NaN	NaN	NaN	NaN
3	Kathmandu	Unhealthy	112.2	198.5	14.8	NaN	NaN	NaN
4	Lahore	Unhealthy	104.5	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...
428	Puertollano, Castilla-La Mancha	Moderate	23.0	15.0	34.0	14.0	5.0	NaN
429	Beasain, Basque Country	Moderate	22.0	NaN	NaN	NaN	2.0	0.0
430	Ciutat Meridiana, Catalunya	Moderate	22.0	32.0	NaN	NaN	NaN	NaN
431	Arwad, Tartus	Moderate	28.7	46.3	NaN	18.4	7.9	NaN
432	Fiq, Quneitra	Moderate	28.8	NaN	NaN	NaN	NaN	NaN

- **Thông tin cột :**
  - *City* : Tên thành phố.
  - *Level* : Cấp độ chất lượng không khí.
  - Các cột *PM2.5, PM10, CO, SO2, NO2, O3* : Chỉ số các chất gây ảnh hưởng chất lượng không khí.

4 TIỀN XỬ LÝ

Các thao tác trong tiền xử lý:

- Xóa dòng có cột Level bị thiếu dữ liệu để không ảnh hưởng đến mô hình.
- Tách tập huấn luyện, validation và tập test.
- Xóa cột City.
- Thay thế dữ liệu bị thiếu (Nan -> 0).
- Số hóa cột Level.

Sử dụng các phương thức trong sklearn:

- FunctionTransformer : Chuyển đổi dữ liệu (không cần tính toán giá trị từ tập huấn luyện).
- StandardScaler : Chuẩn hóa dữ liệu.
- Pipeline.

5 HUẤN LUYỆN MÔ HÌNH

Sử dụng mô hình Neural Network để huấn luyện dữ liệu.

- Thay đổi các tham số alpha và hidden_layer_sizes, chọn ra mô hình tốt nhất để huấn luyện.

alphas = [0.00001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]
li_hidden_layer_sizes = [5, 10, 15, 20, 25, 30]

$\Rightarrow$ **Huấn luyện 6*8 = 48 mô hình.**

Kết quả : Độ lỗi qua các lần chạy < 20%.

6 TÀI LIỆU THAM KHẢO

- Slide bài giảng môn Nhập môn Khoa học Dữ liệu – thầy Trần Trung Kiên
- https://scikit-learn.org/
- https://en.wikipedia.org/wiki/Air_quality_index

THANKS FOR WATCHING