**BÁO CÁO ĐỒ ÁN CUỐI KỲ**

**Nhập Môn Khoa Học Dữ Liệu**

**2020 - 2021**

# 

# *Giáo viên hướng dẫn: Trần Trung Kiên-Phan Thị Phương Uyên*

# Group Information(Nhóm 16)

|  |  |  |
| --- | --- | --- |
| STT | MSSV | Họ và Tên |
| 1. | 18127199 | Lê Thanh Sơn |
| 2.. | 18127190 | Lê Hồng Quang |

# Giới thiệu đồ án:

* Thu thập dữ liệu về các chỉ số và chất lượng không khí tại một số tỉnh/thành phố tại Việt Nam từ 0 giờ ngày 1/6/2021 đến 0 giờ ngày 1/9/2021.
* Các địa điểm : Tiền Giang, Bình Dương, Bình Phước, Lạng Sơn, Quảng Bình, TP.HCM .

# Thu thập dữ liệu:

* Nguồn: <https://openweathermap.org/api/air-pollution>
* Phương pháp thu thập dữ liệu : Dùng API mà trang web cung cấp
* Cú pháp: http://api.openweathermap.org/data/2.5/air\_pollution/history?lat={lat}&lon={long}&start={start}&end={end}&appid={my\_api\_key}
* Ảnh có chứa bàn

  Mô tả được tạo tự động

# Khám phá dữ liệu:

* Dữ liệu có bao nhiêu dòng và bao nhiêu cột? (13206,10)
* Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không?: Mỗi dòng trong DataFrame `air\_df` cho biết dữ liệu về các chỉ số và chất lượng không khí của các tỉnh, thành phố từ 0h ngày 1/8/2021 - 0h ngày 1/9/2021 . Có vẻ không có vấn đề các dòng có ý nghĩa khác nhau.
* Dữ liệu có các dòng bị lặp không? Không
* Ý nghĩa các cột:

|  |  |
| --- | --- |
| Tên cột | Ý nghĩa |
| aqi | Chỉ số chất lượng không khí (1,2,3,4,5) với (1 = Good; 2 = Fair; 3 = Moderate; 4 = Poor ; 5 = Very Poor) |
| co | Nồng độ CO (Carbon monoxide), μg/m³ |
| no | Nồng độ NO (Nitrogen monoxide), μg/m³ |
| no2 | Nồng độ NO2 (Nitrogen dioxide), μg/m³ |
| o3 | Nồng độ O3 (Ozone), μg/m³ |
| so2 | Nồng độ SO2 (Sulphur dioxide), μg/m³ |
| pm2\_5 | Nồng độ PM2\_5 (Fine particles matter), μg/m³ |
| pm10 | Nồng độ PM10 (Coarse particulate matter), μg/m³ |
| nh3 | Nồng độ NH3 (Ammonia), μg/m³ |

* Cột output: Cột output đang ở kiểu dạng số, không có giá trị thiếu và tỉ lệ lớp 3 và 5 trong cột output chênh lệch nhiều so với các lớp còn lại, vì vậy hãy lấy random khoảng 40% lớp 1, 2, 4 để các lớp có tỉ lệ gần nhau nhất.

Ảnh có chứa văn bản

Mô tả được tạo tự độngẢnh có chứa văn bản

Mô tả được tạo tự động

# Câu hỏi:

* Chất lượng không khí được tính như thế nào từ các chỉ số trong không khí đã được đo từ môi trường ?
* Ý nghĩa thực tế của câu hỏi: Theo Tổ chức Y tế thế giới (WHO), hằng năm có 34.000 ca tử vong tại Việt Nam có liên quan đến ô nhiễm không khí. Việc xác định chất lượng không khí giúp người dân có các biện pháp bảo vệ bản thân và cộng đồng kịp thời như đeo khẩu trang, hạn chế sử dụng các phương tiện giao thông gây ô nhiễm, ... . Ngoài ra, chính quyền địa phương có thể hạn chế lượng khí thải từ các nhà máy, xí nghiệp, trồng thêm nhiều cây xanh để làm giảm ô nhiễm không khí.

# Tiền xử lý:

* Sau khi đã khám phá dữ liệu, nhóm sẽ tiến hành tách dữ liệu thành ba tập: tạp train, tập validation, tập test.

Ảnh có chứa văn bản

Mô tả được tạo tự động

* Trong tập huấn luyện:
* Mỗi cột input hiện đang có kiểu dữ liệu gì? Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp không?

Ảnh có chứa văn bản

Mô tả được tạo tự động

* Với mỗi cột input có kiểu dữ liệu dạng số, các giá trị được phân bố như thế nào?

Ảnh có chứa bàn

Mô tả được tạo tự động

* Với các cột dạng số, ta sẽ điền giá trị thiếu bằng giá trị mean của cột (dùng SimpleImputer trong Sklearn).
* Cuối cùng, khi tất cả các cột đã được điền giá trị thiếu và đã có dạng số, ta sẽ tiến hành chuẩn hóa bằng cách trừ đi mean và chia cho độ lệch chuẩn của cột để giúp cho các thuật toán cực tiểu hóa như Gradient Descent, LBFGS, ... hội tụ nhanh hơn (dùng StandardScaler trong Sklearn).

Ảnh có chứa bàn

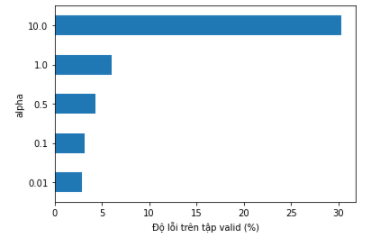
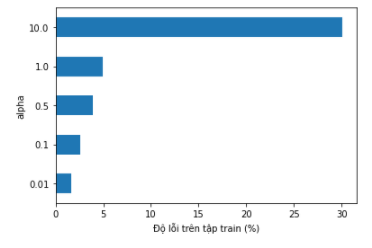
Mô tả được tạo tự động

* Thực hiện transform trên tập validation

# Mô hình hóa:

* Xem xét 3 mô hình :
* Mô hình sử dụng MLP Classifier
* Mô hình sử dụng Softmax Regression
* Mô hình sử dụng KNeighbors Classifier
* Tạo một full pipeline để chứa quá trình tiền xử lý ở trên và một trong 3 mô hình cần xem xét.

1. Mô hình sử dụng MLP Classifier:



Ảnh có chứa văn bản

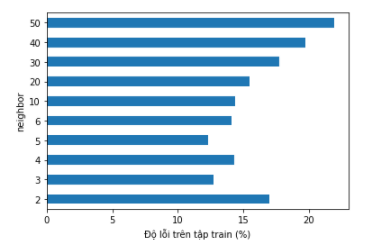
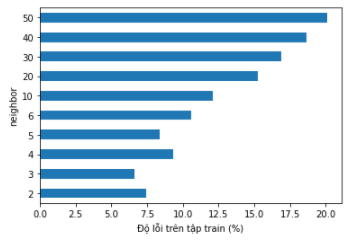
Mô tả được tạo tự động

1. Mô hình sử dụng Softmax Regression

Ảnh có chứa văn bản

Mô tả được tạo tự động

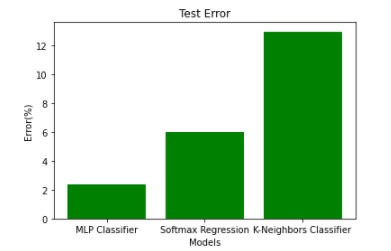
1. Mô hình sử dụng KNeighbors Classifier



Ảnh có chứa văn bản

Mô tả được tạo tự động

* Mô hình tốt nhất:



Dựa vào độ lỗi trên tập test trên ta thấy mô hình MLP Classifier hoạt động hiệu quả hơn hai mô hình còn lại => dự đoán chất lượng không khí tốt nhất.

# Nhìn lại quá trình làm đồ án

1. Khó khăn:

* Quang: lựa chọn mô hình phù hợp với dữ liệu đã dùng, tốn thời gian trong việc tìm kiếm nguồn cung dữ liệu (một số webiste yêu cầu trả phí hoặc gửi email xin xác nhận mục đích sử dụng để truy cập dữ liệu).
* Sơn:

1. Kỹ năng đã học được:

* Quang: qua đồ án này em có thêm kinh nghiệm để chọn lọc và thu thập dữ liệu, hiểu thêm nhiều thuật toán phân lớp hữu ích
* Sơn:

1. Nếu có thêm thời gian: nhóm sẽ tìm và thu thập dữ liệu đa dạng, phức tạp hơn, có thể thử nghiệm thu thập dữ liệu bằng cách parse HTML và mô hình hóa dữ liệu với nhiều thuật toán khác.

# Tham khảo

1. <https://openweathermap.org/api/air-pollution>
2. <https://scikit-learn.org/>
3. <https://stackoverflow.com/>
4. Video bài giảng + demo của môn học+ slide bài giảng
5. Homework:1,2,3