ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

---oOo---

PROJECT 4

AIR QUALITY SENSOR CALIBRATION USING LINEAR MODELS

Nhóm sinh viên thực hiện:

Lê Gia Bảo **MSSV:** 23127325

Vũ Anh **MSSV:** 23127321

Hồ Gia Huy **MSSV:** 23127376

Thành phố Hồ Chí Minh, tháng 4 năm 2025

I. Phân công công việc

Thành Viên	Nhiệm vụ chính
Lê Gia Bảo	Tiền xử lý, Task 1 (Linear models)
Hồ Gia Huy	Feature engineering, Task 2
Vũ Anh	Kiểm định và đánh giá thống kê, Task 3, report

II. Mô tả dữ liệu

- **Train.csv**: có *Time, Ozone, NO2, temp, humidity, no2op1, no2op2, o3op1, o3op2.*
- **Test.csv**: Cùng cấu trúc train.csv, sử dụng để kiểm thử
- **Tiền xử lý**: Loại bỏ giá trị thiếu (nếu có), tách *Time* thành *hour*, mã hóa chu kỳ *hour_sin, hour_cos*, chuẩn hóa (StandardScaler)

III. Phương pháp luận

1. **Task 1: Mô hình tuyến tính cơ bản**
   - Đặc trưng: chỉ 4 điện áp.
   - Mô hình: OLS, Ridge ($\alpha = 1$), Lasso ($\alpha = 1$)
   - Đánh giá: MAE trên tập train
2. **Task 2: Mô hình nâng cao**
   - Mở rộng đặc trưng: voltage + *temp, humidity, hour_sin, hour_cos*
   - Mô hình thử nghiệm: Linear Regression, Random Forest (200 cây, depth = 20), MLP (2 lớp ẩn 128-64, $\alpha = 0.001$).

- Tuning: RandmomizeSearchCV, GridSearchCV.
- Đánh giá thống kê: MSR, R^2, kiểm định Shapiro-Wilk, Breusch-Pagan, BIC cho Linear, đo thời gian train và inference.

3. **Triển khai và kiểm thử**
   - Script predict.py: chọn MLP và chạy test.csv
   - Đầu ra: file predictions.csv gồm *Time, OZONE_PRED, NO2_PRED.*

IV. Kết quả thí nghiệm và phân tích

1. **Task 1: Linear models**

Mô hình	**MAE O_3**	**MAE NO_2**
OLS	5.6259	6.5401
Ridge	5.6259	6.5401
Lasso	5.6804	6.6216

Nhận xét: Cả 3 mô hình tuyến tính cơ bản đều cho MAE > 5.6, vẫn còn sai số lớn (MAE càng nhỏ thì mô hình dự đoán càng chính xác)

2. **Task 2: Advanced Models**

   a. **O_3**

Mô hình	MAE	R^2	Train Time (s)	Inference Times (s)	Shapiro p-val	BP p-val
Linear Regression	5.2010	0.8592	0.00	0.00	0.0000	0.0000
Random Forest	3.4905	0.9191	8.82	0.15	0.0000	0.0000
Neural Network	4.1040	0.9071	6.87	0.00	0.0000	0.0000

Nhận xét:

- Random Forest vượt trội so với Neural Network và Linear Regression.
- Giảm sai số trung bình tuyệt đối (MAE) khoảng 0.61 µg/m³ so với MLP và 1.71 µg/m³ so với linear, đồng thời R^2 cao nhất cho thấy RF có hiệu suất tốt.

b. NO_2

Mô hình	MAE	R^2	Train Time (s)	Inference Times (s)	Shapiro p-val	BP p-val
Linear Regression	6.6295	0.3284	0.00	0.00	0.0000	0.0000
Random Forest	2.2265	0.8911	8.74	0.14	0.0000	0.0000
Neural Network	3.2118	0.8114	10.09	0.01	0.0000	0.0000

Nhận xét:

- Random Forest cũng dẫn đầu so với Neural Network và Linear Regression.
- MAE của RF thấp hơn MLP gần 1.0 µg/m³, và cải thiện mạnh so với linear, R^2 của RF gần 0.89, trong khi linear chỉ ~0.33, tức là RF giải thích tốt hơn gấp 3 lần linear.

c. So sánh hiệu quả tính toán

- Thời gian huấn luyện
  - Linear Regression nhanh nhất, nhưng accurary quá kém.
  - MLP mất ~6.87-10.09 s, lâu hơn RF mà MAE lại cao hơn.

- RF có độ trễ train chấp nhận được (~8.8s), inference nhanh (0.14-0.15s)

- Thời gian suy luận (inference)
  - Cả ba đều thực thi rất nhanh (<0.2s cho toàn bộ tập validation). Không có khác biệt lớn.

d. Đánh giá thống kê (Diagnostics)

- Shapiro-Wilk và Breusch-Pagan
  - Shapiro-Wilk Test: Kiểm định tính phân phối chuẩn của residuals. Cả ba mô hình đều có p-value = 0.0000 (<0.05) => bác bỏ H_0, residuals không phân phối chuẩn
  - Breusch-Pagan Test: Kiểm định phương sai đồng nhất. Cả ba mô hình đều có p-value = 0.0000 (<0.05) => bác bỏ H_0, residuals không phân phối chuẩn
- BIC (chỉ áp dụng cho Linear Regression)
  - BIC O_3 = 15794.06
  - BIC NO_2 = 17838.19

    ⇨ Với BIC này cho thấy mô hình có độ phức tạp vừa phải nhưng likelihood trên dữ liệu chưa tối ưu

3. Task 3: Deployment

- Triển khai mô hình đã huấn luyện dể dự đoán dữ liệu mới và lưu kết quả
- Quy trình:
  - Tải mô hình (joblib.load) và scaler.pkl từ thư mục models/
  - Đọc test.csv, chuyển Time sang datetime, trích xuất hour, tạo hour_sin, hour_cos.
  - Chuẩn hóa test data với scaler.transform()

- Dự đoán O_3 và NO_2, tạo cột OZONE_PRED, NO2_PRED
- Xuất predictions.csv gồm Time, OZONE_PRED và NO2_PRED

- Cú pháp chạy:

V. Kết luận

Mô hình tối ưu (Best Model): Sau khi so sánh các tiêu chí MAE, R^2, thời gian huấn luyện và inference. Random Forest được lựa chọn là mô hình tốt nhất cho cả hai biến mục tiêu O_3 và NO_2 bởi:

1. Độ chính xác cao nhất: MAE giảm đáng kể (3.49 cho O_3, 2.23 cho NO_2) và R^2 trên 0.89
2. Hiệu quả tính toán: Thời gian huấn luyện chấp nhận được (~8.7s) và inference nhanh (~0.15s).
3. Tính ổn định và dễ triển khai: RF ít nhạy cảm với scaling, không cần tuning phức tạp như MLP

VI. Tài liệu tham khảo

1. Shapiro, S. S., & Wilk, M. B. (1965). “An analysis of variance test for normality.” *Biometrika*, 52(3/4), 591–611.
2. Breusch, T. S., & Pagan, A. R. (1979). “A simple test for heteroscedasticity.” *Econometrica*, 47(5), 1287–1294.
3. Breiman, L. (2001). “Random Forests.” *Machine Learning*, 45(1), 5–32.
4. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

5. Pedregosa, F. et al. (2011). “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*, 12, 2825–2830.
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.