|  |
| --- |
| **ĐẠI HỌC QUỐC GIA HÀ NỘI**  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**  A blue circle with white text  Description automatically generated  **Nguyễn Tiến Khôi**  **Tống Duy Tân**  **Nguyễn Quốc Tuấn**  **BÁO CÁO HỌC MÁY DỰ ĐOÁN NHIỆT ĐỘ THỜI TIẾT CAO NHẤT TRONG NGÀY**            **BÁO CÁO DỰ ÁN CUỐI KỲ 2 NĂM HỌC 2023**  Ngành: **Trí tuệ nhân tạo**                              HÀ NỘI - 2023 |

TÓM TẮT

Dự báo thời tiết hiện nay đã trở thành một công cụ không thể thiếu đối với người dân không chỉ với người Hà Nội nói riêng mà cả thế giới nói chung. Nhận thấy rằng công cụ này rất hữu ích và đóng góp rất lớn cho đời sống mọi người nên nhóm chúng em rất có hứng thú với nó. Cùng thời gian đang trong quá trình học Học Máy của cô Diệp, bọn em quyết định làm một báo cáo về việc so sánh các mô hình dự đoán nhiệt độ cao nhất trong ngày tiếp theo, nhận định xem 3 mô hình học máy bọn em làm có điểm gì nổi bật và tối ưu hơn không. Đồng thời cũng rèn luyện kỹ năng phân tích dữ liệu,

thiết kế bài toán học máy hợp lý, ckhả năng diễn giải được kết quả và ưu nhược điểm của mô hình trong bài toán thực tế, ....

Chính vì vậy, trong đề tài này, chúng em đã thực hiện quá trình cào dữ liệu từ trang <https://www.visualcrossing.com/> để lấy dữ liệu, nghiên cứu rồi phân tích thông tin thu được để làm 1 bài báo cáo đề tài này.­­­­­­­­ Do lần đầu được làm báo cáo nên còn nhiều thiếu sót, chúng em mong nhận được sự góp ý, phê bình của thầy để báo cáo của chúng em được hoàn thiện hơn.

Từ khóa: Học Máy, dự báo thời tiết

-

CẤU TRÚC CỦA PHẦN CODE BÁO CÁO (GITHUB)

FINAL-PROJECT-MASTERCHEF

Analyze\_Data\_MasterChef (Folder)

Analyze\_Data\_MasterChef\_Comments.ipynb

Analyze\_Data\_MasterChef\_Images\_Videos.ipynb

Analyze\_Data\_MasterChef\_PostTime.ipynb

Analyze\_Data\_MasterChef\_Reactions.ipynb

Analyze\_Data\_MasterChef\_PostText.ipynb

Data (Folder)

Masterchef.npy

Masterchef.csv

Masterchef.xlsx

Images (Folder)

Crawl\_data\_MasterChef.ipynb

# Phần 1: Thu thập dữ liệu từ Visual Crossing về và xử lý dữ liệu

# (Lưu vào file hanoiweather.csv)

## **I, Thu thập dữ liệu**

### Trang <https://www.visualcrossing.com/> là trang web chuyên cung cấp về dữ liệu thời tiết và API uy tín và có bộ dataset đầy đủ nên chúng em chọn web này để thu thập dữ liệu

### Địa điểm thu thập dữ liệu là Hà Nội, thời gian sẽ thu thập từ năm 1990 đến hiện tại và lấy theo ngày

### Mỗi ngày mỗi tài khoản sẽ được crawl dữ liệu khoảng tầm 1 năm nên chúng em 3 người đã thu thập dữ liệu trong khoảng 10 ngày.

### Phần chia công việc thu thập dữ liệu

### Nguyễn Quốc Tuấn: 1/1/90 - 31/12/99

### Tống Duy Tân: 1/1/2000 - 31/12/2009

### Nguyễn Tiến Khôi: 1/1/2010 – đến nay

A screenshot of a computer

Description automatically generated

### Bộ dữ liệu khá nhiều cột và đầy đủ (37 cột)

## **II, Xử lý dữ liệu**

### Tống Duy Tân:

### Để tận dụng hết các thông tin của cột, em biến đổi thời gian sunrise, sunset trong ngày từ "%Y-%m-%dT%H: %M: %S" sang dạng phút.

### Thêm 1 cột tempmax\_next để dự đoán nhiệt độ cao nhất trong ngày (mục tiêu)

### Vì weather['conditions'] trong dữ liệu được miêu tả như ‘Overcast’, ‘Partially cloudy’ hay ‘Rain, Partially cloudy’ nên sẽ tạo luôn các cột tương ứng với điều kiện thời tiết như vậy và đánh giá true hay false cho tất cả các ngày như là ‘cond\_Overcast', 'cond\_Partially cloudy', hay 'cond\_Rain, Partially cloudy']

### Lấp các giá trị khuyết trong bộ dữ liệu (57/12538) với tỷ lệ chỉ khoảng 0.475% không đáng kể đến chất lượng tổng thể của bộ dữ liệu. Phương pháp "forward fill" lấp đầy các giá trị NaN bằng cách sử dụng giá trị cuối cùng không phải là NaN trước chúng.

# Phần 2: Phân tích dữ liệu và đánh giá

### Tống Duy Tân:

A graph showing a number of lines

Description automatically generated with medium confidence

### Nhiệt độ tối đa trong bộ dữ liệu là: 43.1 °C

### Nhiệt độ tối thiểu trong bộ dữ liệu là: 3.8 °C

A graph with red and blue lines

Description automatically generated

A chart of different colored squares

Description automatically generated with medium confidence

### Phân tích: Từ biểu đồ này, ta có thể phân tích sâu hơn về mối quan hệ giữa các biến, chẳng hạn như khám phá nguyên nhân của các tương quan cao hoặc thấp và cách chúng ảnh hưởng đến các yếu tố thời tiết khác: tempmax, tempmin, feelslikemax, feelslikemin, và feelslike đều có tương quan rất cao với nhau, điều này có ý nghĩa vì chúng đều liên quan trực tiếp đến nhiệt độ. temp (nhiệt độ trung bình) cũng có tương quan rất cao với tempmax, tempmin, và các biến "feels like", thể hiện rõ mối liên hệ giữa nhiệt độ trung bình và các chỉ số nhiệt độ khác.

### Chọn đặc trưng cho mô hình học máy: Tương quan giữa các đặc trưng có thể giúp chúng ta chọn lọc các đặc trưng quan trọng hoặc loại bỏ các đặc trưng thừa cho các mô hình học máy, nhằm cải thiện hiệu quả và độ chính xác của mô hình

# Phần 3: Tạo các bộ dữ liệu dành cho training, validating và testing.

# Phần 4: Thực nghiệm và tối ưu mô hình

## **I, So sánh MSE của data với MSE của mô hình**

### MSE (weather['tempmax\_next']- weather['tempmax']) \*\* 2). mean () = 6.4949137

### Mô hình:

* Validation MSE: 4.769188684210529
* Validation R-squared: 0.8617088947781176
* Test MSE: 4.8152609445773535
* Test R-squared: 0.8572996132849282.

### Mô hình đang hoạt động tốt hơn so với mô hình đơn giản.

## **II, Thực nghiệm**

## **III, Tối ưu**

### Tống Duy Tân: Sử dụng Grid Search để tìm ra bộ tham số tối ưu cho mô hình Random Forest. Sau khi tìm được mô hình tối ưu, em sử dụng mô hình này để dự đoán các giá trị mới và đánh giá chính xác hơn hiệu suất của mô hình trên dữ liệu thực tế. Grid Search giúp đảm bảo rằng mô hình không chỉ phù hợp với tập huấn luyện mà còn có khả năng tổng quát hóa tốt trên dữ liệu mới.

# Phần 5: Đánh giá mô hình

# **Đây sẽ là bài đăng mà có trường hợp khác thường khi xử lý dữ liệu. Những code trên hầu như là không có lỗi nên có vẻ đây sẽ là lỗi dữ liệu khi crawl về. Việc lỗi dữ liệu cũng sẽ là điều không thế tránh khỏi khi crawl 1 dữ liệu khá khá lớn thế này.**