TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN, ĐHQG-HCM

KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN MÔN HỌC

ĐỀ TÀI:

PHÂN TÍCH VÀ DỰ ĐOÁN ĐƯỜNG CHIM BAY (BIRD TRACE)

Môn học: CS313.Q12 - Khai thác dữ liệu và ứng dụng.

Giảng viên hướng dẫn: TS. Võ Nguyễn Lê Duy

Thực hiện bởi nhóm 12, bao gồm:

1.	Nguyễn Thanh Tuấn	23521724	Trưởng nhóm
2.	Trần Minh Vũ	23521819	Thành viên
3.	Đinh Trần Duy Trường	23521688	Thành viên
4.	Nguyễn Đăng Khoa	24520820	Thành viên

MỤC LỤC

Chương I. GIỚI THIỆU ĐỀ TÀI

1.1. Bối cảnh

Hành vi di cư của các loài chim là một trong những hiện tượng tự nhiên kỳ thú và phức tạp nhất trong thế giới động vật. Khả năng định vị và di chuyển qua hàng ngàn km của chúng chịu sự chi phối của nhiều yếu tố nội tại lẫn môi trường. Việc hiểu và dự đoán được đường bay của chim không chỉ mang lại giá trị khoa học to lớn trong việc giải mã tập tính sinh học mà còn mở ra nhiều ứng dụng thực tiễn quan trọng trong đời sống và bảo tồn.

1.2. Ý nghĩa

Nghiên cứu dự báo đường bay mang lại những đóng góp cụ thể cho công tác **Bảo tồn sinh học**, bao gồm:

- **An toàn hàng không:** Cung cấp dữ liệu cảnh báo sớm cho các sân bay để tránh các vụ va chạm giữa chim và máy bay trong mùa di cư.
- **Thực thi pháp luật bảo tồn:** Hỗ trợ cơ quan chức năng khoanh vùng và tối ưu hóa lộ trình tuần tra để ngăn chặn nạn săn bắt chim trái phép.
- **Bảo vệ sinh cảnh:** Xác định các khu vực trọng yếu cần ưu tiên bảo vệ nguồn thức ăn và môi trường sống cho chim.

1.3. Mục tiêu nghiên cứu

Đồ án tập trung vào hai mục tiêu chính:

- **Dự đoán quỹ đạo:** Xây dựng mô hình dự báo vị trí của chim trong tương lai gần (24 giờ tới) với độ chính xác cao.
- **Phân tích yếu tố ảnh hưởng:** Làm rõ mối tương quan giữa hành vi bay và các điều kiện môi trường như nhiệt độ, hướng gió, độ cao.

1.4. Tổng quan Input – Output của bài toán

Mô hình được thiết kế dưới dạng bài toán dự báo chuỗi thời gian (Time-series forecasting):

- **Input (Đầu vào):** Dữ liệu lịch sử trong cửa sổ thời gian **48 giờ** (2 ngày). Các đặc trưng bao gồm: Vị trí địa lý (Kinh độ, Vĩ độ) và các yếu tố môi trường/vận động (Nhiệt độ, Tốc độ di chuyển, Hướng di chuyển).
- **Output (Đầu ra):** Chuỗi tọa độ vị trí dự đoán của đối tượng trong **24 giờ** tiếp theo.

Chương II. BỘ DỮ LIỆU

2.1. Nguồn dữ liệu

Dữ liệu được thu thập thông qua API từ **Movebank.org** - cơ sở dữ liệu trực tuyến toàn cầu về chuyển động của động vật.

- **Đối tượng nghiên cứu:** Loài chim Cắt nhỏ (*Falco naumanni*).
- **Khu vực địa lý:** Dữ liệu tập trung vào quần thể sinh sống và di cư tại khu vực Senegal (Tây Phi).

2.2. Thông tin thuộc tính

Bộ dữ liệu thô bao gồm **30,904** điểm dữ liệu với **16** đặc trưng (features). Qua kiểm tra sơ bộ, dữ liệu đạt chất lượng tốt với không có đặc trưng quan trọng nào bị giá trị NULL.

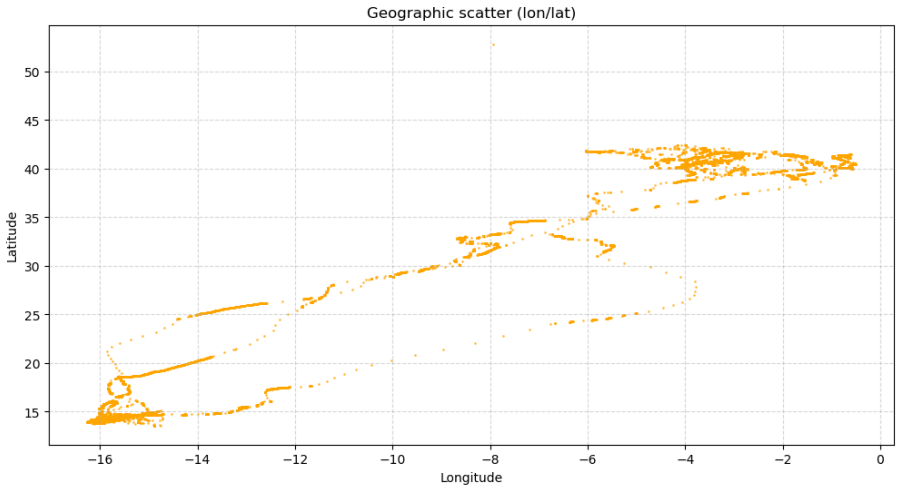
Các đặc trưng chính được sử dụng trong mô hình:

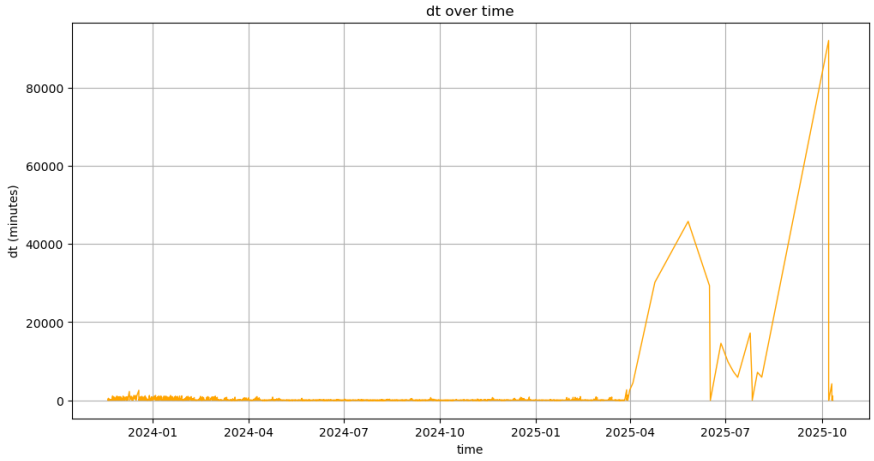
- timestamp: Thời gian ghi nhận.
- location-long: Kinh độ.
- location-lat: Vĩ độ.
- temperature: Nhiệt độ môi trường tại thời điểm ghi nhận.
- speed: Tốc độ bay của chim.
- heading: Hướng bay.
- height: Độ cao so với mực nước biển.

Chương III. PHÂN TÍCH VÀ KHÁM PHÁ DỮ LIỆU

3.1. Phân tích phân bố thời gian

Biểu đồ phân bố tọa độ của loài chim:





Khi trực quan hóa mật độ các điểm dữ liệu theo thời gian, chúng tôi phát hiện một sự bất thường lớn bắt đầu từ tháng **4/2025**. Cụ thể:

- Khoảng cách thời gian (time delta) giữa các bản ghi tăng đột biến.
- Xuất hiện các khoảng trống dữ liệu (gaps) lớn, cho thấy sự ngừng hoặc gián đoạn nghiêm trọng trong quá trình thiết bị thu thập tín hiệu.

3.2. Xử lý và Làm sạch

Để đảm bảo chất lượng dữ liệu đầu vào cho mô hình huấn luyện, quyết định xử lý được đưa ra như sau:

- **Ngưỡng cắt (Cut-off date):** Chỉ giữ lại các dữ liệu được ghi nhận trước ngày **31/03/2025**.
- **Kết quả:**
    - Loại bỏ các bản ghi rời rạc sau tháng 4/2025.
    - Số lượng điểm dữ liệu còn lại: **30,875** mẫu.
    - Dữ liệu sau khi lọc đảm bảo tính liên tục cao, mật độ thu thập ổn định qua các ngày, phù hợp cho việc trích xuất cửa sổ trượt (sliding window) cho bài toán input 48h - output 24h.

Chương IV. TIỀN XỬ LÝ DỮ LIỆU

Để đảm bảo chất lượng dữ liệu đầu vào cho mô hình huấn luyện và nâng cao độ chính xác của dự báo, quá trình tiền xử lý dữ liệu được thực hiện qua 3 bước chính: Lấy mẫu lại (Resampling), Lựa chọn đặc trưng (Feature Selection) và Trích xuất đặc trưng (Feature Extraction).

4.1. Lấy mẫu lại dữ liệu (Data Resampling)

Do đặc thù của việc thu thập dữ liệu di chuyển, khoảng cách thời gian giữa các bản ghi thường không đồng đều. Việc Resampling đóng vai trò cốt lõi nhằm chuẩn hóa tần suất dữ liệu, giảm nhiễu và tăng tính ổn định cho chuỗi thời gian.

Quy trình thực hiện cụ thể như sau:

- **Phân đoạn:** Dữ liệu được chia thành các phiên (sessions) độc lập dựa trên ngưỡng thời gian là **12 giờ**. Nếu khoảng cách giữa hai bản ghi liên tiếp lớn hơn 12 giờ, hệ thống sẽ ngắt và bắt đầu một phiên mới.
- **Tái lập tần suất:** Trong mỗi phiên, dữ liệu được resample về tần suất cố định là **1 giờ**.
- **Nội suy:** Sử dụng phương pháp nội suy tuyến tính (Linear Interpolate) để điền giá trị cho các điểm dữ liệu mới sinh ra, đảm bảo tính liên tục và trơn tru của đường bay.
- **Kết quả:** Sau quá trình resampling, kích thước bộ dữ liệu được chuẩn hóa còn **10,597** điểm dữ liệu chất lượng cao.

4.2. Lựa chọn đặc trưng

Nhằm giảm độ phức tạp tính toán và loại bỏ nhiễu, chúng tôi tiến hành loại bỏ các thuộc tính không mang nhiều ý nghĩa đối với bài toán dự báo đường bay hoặc mang tính chất quản lý hệ thống, ví dụ như: `event-id`, `visible`,...

4.3. Trích xuất và Kỹ thuật đặc trưng

Để giúp mô hình học được các quy luật di chuyển phức tạp, các đặc trưng mới được tạo ra từ dữ liệu gốc:

a. Chuyển đổi hệ tọa độ

- Chuyển đổi tọa độ địa lý (Kinh độ, Vĩ độ) sang hệ tọa độ **UTM (Universal Transverse Mercator)**.
- **Mục đích:** Do Trái đất hình cầu, việc tính toán khoảng cách Euclid trên kinh/vĩ độ sẽ thiếu chính xác. Hệ UTM cho phép tính toán khoảng cách và hướng di chuyển chính xác hơn bằng đại số tuyến tính trên mặt phẳng.

b. Tạo đặc trưng thời gian và ngữ cảnh

Các đặc trưng mới được sinh ra để cung cấp ngữ cảnh cho mô hình:

- `distance`: Khoảng cách giữa hai vị trí liên tiếp.
- `time_of_day`: Phân loại thời điểm trong ngày (Sáng, Chiều, Tối).

- `season`: Mùa trong năm (Xuân, Hạ, Thu, Đông).

c. Mã hóa đặc trưng tuần hoàn (Cyclical Encoding)

Đối với các dữ liệu có tính chất chu kỳ như thời gian (giờ trong ngày, tháng trong năm) hoặc hướng di chuyển (heading - độ), mô hình thông thường sẽ hiểu sai khoảng cách.

- **Giải pháp:** Sử dụng phép biến đổi lượng giác sin và cos để mã hóa các đặc trưng này.
- **Ý nghĩa:** Giúp mô hình nhận biết được tính liên tục giữa điểm đầu và điểm cuối của chu kỳ.

Chương V. XÂY DỰNG MÔ HÌNH

5.1. Thiết lập thực nghiệm

Để đảm bảo tính khách quan và phù hợp với đặc thù dữ liệu chuỗi thời gian, quá trình thực nghiệm được thiết lập như sau:

a. Phân chia dữ liệu (Data Splitting) Dữ liệu được chia theo trình tự thời gian (chronological split) thành 3 tập độc lập để mô phỏng bài toán dự báo thực tế:

- **Tập huấn luyện (Train):** Từ 18/11/2023 đến 30/11/2024. Dùng để huấn luyện mô hình học các quy luật di chuyển dài hạn.
- **Tập kiểm định (Validation):** Từ 01/12/2024 đến 31/01/2025. Dùng để tinh chỉnh siêu tham số và đánh giá sớm nhằm tránh hiện tượng quá khớp (overfitting).
- **Tập kiểm tra (Test):** Từ 01/02/2025 đến 31/03/2025. Dùng để đánh giá hiệu năng cuối cùng của mô hình.

b. Kỹ thuật Cửa sổ trượt (Sliding Window) Để chuyển đổi dữ liệu chuỗi thời gian sang dạng bài toán học có giám sát (Supervised Learning), chúng tôi áp dụng kỹ thuật cửa sổ trượt với các thông số:

- **Input Window (Cửa sổ đầu vào): 48 giờ** (Sử dụng thông tin của 2 ngày quá khứ).
- **Output Horizon (Tầm dự báo): 24 giờ** (Dự đoán vị trí cho 1 ngày tiếp theo).
- **Bước trượt (Step/Stride): 1** (Cửa sổ dịch chuyển từng bước 1 giờ để tạo ra mẫu dữ liệu mới liên tục).

5.2. Các Mô hình

Để tìm ra giải pháp tối ưu cho bài toán dự báo quỹ đạo bay, nhóm nghiên cứu đã tiến hành thực nghiệm trên tập hợp các mô hình học máy và học sâu đa dạng sau:

- **Linear Regression:** Mô hình hồi quy tuyến tính cơ bản, đóng vai trò làm chuẩn (baseline) để xác lập mối quan hệ tuyến tính giữa các biến đầu vào và tọa độ mục tiêu.
- **Random Forest:** Phương pháp học tổ hợp (Ensemble learning) sử dụng đa số phiếu từ nhiều cây quyết định để xử lý dữ liệu phi tuyến và giảm thiểu hiện tượng quá khớp.
- **XGBoost:** Thuật toán Gradient Boosting hiệu năng cao, được tối ưu hóa để xử lý tốt các dữ liệu dạng bảng và nắm bắt các mẫu phức tạp với tốc độ nhanh.
- **KNN (K-Nearest Neighbors):** Thuật toán phi tham số thực hiện dự báo dựa trên sự tương đồng với k điểm dữ liệu lân cận nhất trong không gian đặc trưng.
- **MLP (Multi-Layer Perceptron):** Mạng nơ-ron nhân tạo truyền thẳng (Feedforward Neural Network) với khả năng học các hàm phi tuyến tính thông qua các lớp ẩn.
- **LSTM (Long Short-Term Memory):** Kiến trúc mạng nơ-ron hồi quy (RNN) chuyên dụng cho dữ liệu chuỗi thời gian, có khả năng ghi nhớ các phụ thuộc dài hạn trong hành vi bay của chim.

5.3. Các độ đo đánh giá

Để đánh giá toàn diện hiệu năng mô hình từ khả năng hồi quy tổng quát đến độ chính xác không gian của quỹ đạo bay, nghiên cứu sử dụng 4 độ đo sau:

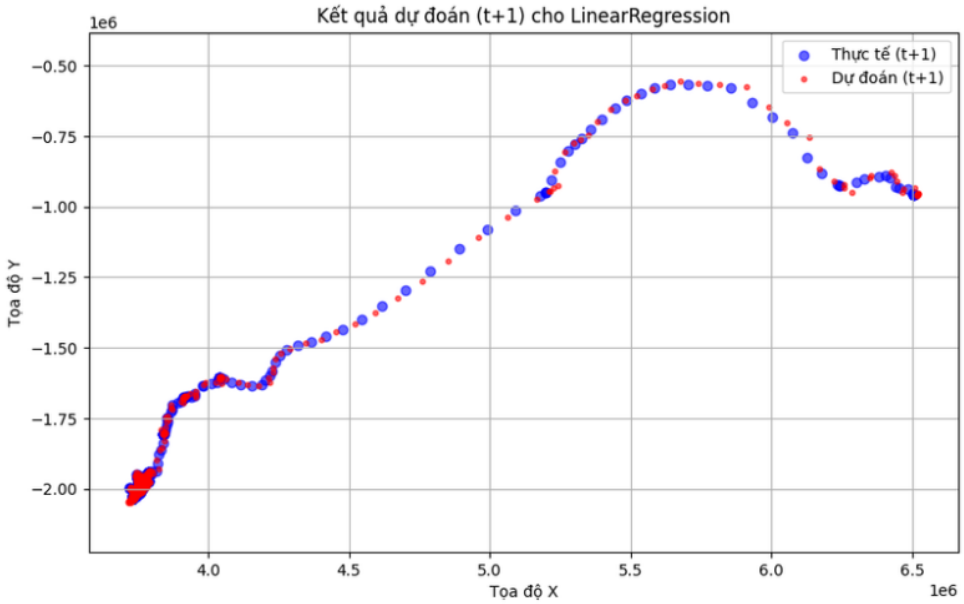
- **MAE (Mean Absolute Error):** Đo mức lệch trung bình giữa vị trí dự báo và thực tế theo khoảng cách địa lý (Haversine), giúp đánh giá chất lượng dự báo tổng thể.
- **R^2(Coefficient of Determination):** Hệ số xác định mức độ mô hình giải thích được sự biến thiên của dữ liệu so với mô hình cơ sở (baseline).
- **ADE (Average Displacement Error):** Đo sai số trung bình trên toàn bộ quỹ đạo 24h, phản ánh độ chính xác liên tục của đường bay theo thời gian.
- **FDE (Final Displacement Error):** Đo sai số tại điểm cuối cùng (giờ thứ 24), đánh giá mức độ chính xác của điểm đến dự kiến.

Chương VI. KẾT QUẢ VÀ ĐÁNH GIÁ

6.1. Kết quả định lượng

	MAE	R^2	ADE	FDE
LinearRegression	**39892.31**	**0.92**	**60555.17**	106329.98
RandomForest	51475.48	0.72	78703.24	101384.48
XGBoost	51544.27	0.67	81189.57	101106.49
KNN	69693.11	0.66	108816.56	117361.30
MLP	63131.78	0.82	98639.18	113682.41
LSTM	52497.29	0.82	82091.74	**97908.78**

6.2. Kết quả định tính



6.3. Đánh giá

Kết quả thực nghiệm cho thấy sự khác biệt rõ rệt giữa các nhóm mô hình:

- **Linear Regression:** Hiệu quả nhất tổng thể (MAE/ADE thấp,R^2 cao) nhờ nắm bắt tốt quán tính (vận tốc + hướng) trong ngắn hạn.
- **LSTM:** Tốt nhất về chỉ số FDE (nhờ thiên kiến chuỗi thời gian) nhưng phải đánh đổi độ khớp trung bình (ADE) do mô hình ưu tiên fit điểm cuối.
- **RandomForest/XGBoost/KNN:** Kém hiệu quả trong bài toán này do lỗi ngoại suy dạng "bậc thang" trên toạ độ liên tục.

Chương VII. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

7.1 Kết luận

Nghiên cứu đã xây dựng và đánh giá hiệu quả các mô hình học máy trong việc dự báo quỹ đạo bay của loài chim. Kết quả thực nghiệm cho thấy Linear Regression là mô hình tối ưu nhất cho bài toán dự báo ngắn hạn (24 giờ) nhờ khả năng nắm bắt tốt xu

hướng và quán tính di chuyển, vượt trội hơn so với các mô hình phức tạp như LSTM hay Random Forest về độ ổn định tổng thể.

Với độ chính xác đạt được, mô hình có khả năng ứng dụng thực tiễn cao trong công tác bảo tồn sinh học, cụ thể là hỗ trợ các cơ quan chức năng khoanh vùng tuần tra để ngăn chặn nạn săn bắt trái phép và giảm thiểu rủi ro va chạm cho ngành hàng không.

7.2 Hướng phát triển

Dựa trên các kết quả thực nghiệm, nhóm đề xuất các hướng nghiên cứu tiếp theo:

1. **Mở rộng phạm vi dữ liệu:** Đánh giá tính tổng quát hóa của mô hình thông qua việc thử nghiệm trên các tập dữ liệu chuyển động của nhiều loài chim khác nhau với đặc tính sinh học và môi trường đa dạng.
2. **Tối ưu hóa dự báo dài hạn:** Cải thiện cấu trúc mô hình để khắc phục hạn chế về sai số tích lũy, từ đó nâng cao độ chính xác khi dự báo quỹ đạo trong khung thời gian dài hơn (long-term prediction).
3. **Kết hợp mô hình (Hybrid Models):** Nghiên cứu phương án kết hợp ưu điểm ngoại suy của mô hình tuyến tính và khả năng học chuỗi của Deep Learning (như LSTM/GRU) để tận dụng sức mạnh của cả hai.

Chương VIII. DEMO

Demo của nhóm được công bố ở đây.

Chương IX. REFERENCES

1. Dolbeer, R. A. (2006). Height distribution of birds recorded by collisions with civil aircraft. The Journal of Wildlife Management, 70(5), 1345–1350.
2. Lees, A. C., & Yuda, P. (2022). The Asian songbird crisis. Current Biology, 32(20), R1063–R1064. https://doi.org/10.1016/j.cub.2022.08.066