TRƯỜNG ĐẠI HỌC MỞ
THÀNH PHỐ HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

THÀNH VIÊN THỰC HIỆN
2251050048 - Hồ Chí Nguyên
2251052011 - Lê Thanh Dân
2251052009 - Trịnh Vĩnh Cường

MÔN HỌC: KHAI PHÁ DỮ LIỆU
ĐỀ TÀI: GLOBAL ANIMA DISEASE

TP. Hồ Chí Minh, Năm 2025

MỤC LỤC

Chương 1

GIỚI THIỆU

Dịch bệnh động vật có thể gây ra những hậu quả nghiêm trọng đối với sức khỏe cộng đồng, an ninh lương thực và nền kinh tế toàn cầu. Việc giám sát dịch bệnh động vật hiệu quả đóng vai trò then chốt trong việc phát hiện sớm, kiểm soát và ngăn chặn sự lây lan của các loại bệnh nguy hiểm. Hệ thống giám sát giúp các nhà chức trách và các tổ chức liên quan đưa ra các biện pháp ứng phó kịp thời, giảm thiểu thiệt hại và bảo vệ sức khỏe cho cả con người và động vật.

Bộ dữ liệu EMPRES Global Animal Disease Surveillance

Link DataSet

Bộ dữ liệu EMPRES Global Animal Disease Surveillance cung cấp thông tin chi tiết về các ổ dịch bệnh động vật trên khắp thế giới, bao gồm vị trí địa lý, thời gian bùng phát, loại bệnh, số lượng động vật bị ảnh hưởng và nhiều yếu tố khác. Đây là một nguồn tài nguyên quý giá để phân tích và hiểu rõ hơn về động thái của dịch bệnh.

Dữ liệu gồm có 24 cột và 17009 dòng.

Mục tiêu Dự đoán và phòng ngừa

Mục tiêu của báo cáo này là áp dụng các kỹ thuật khai phá dữ liệu để phân tích bộ dữ liệu EMPRES, từ đó:

- Phân cụm ổ dịch: Nhận diện các khu vực địa lý và thời gian có xu hướng bùng phát dịch bệnh tương tự nhau
- Tìm kiếm mẫu lây lan: Khám phá các chuỗi sự kiện dịch bệnh phổ biến theo thời gian và không gian.
- Dự đoán số ca bệnh/tử vong: Xây dựng mô hình dự đoán số lượng động vật bị nhiễm bệnh hoặc tử vong dựa trên các yếu tố liên quan.

Thông qua việc khai phá dữ liệu, Bài báo cáo hy vọng sẽ đóng góp vào việc nâng cao hiệu quả giám sát dịch bệnh động vật và hỗ trợ công tác phòng ngừa, kiểm soát dịch bệnh tốt hơn trong tương lai.

Chương 2

TỔNG QUAN VỀ DỮ LIỆU

Phân tích dữ liệu ban đầu

- **Số lượng dòng và cột:** Bộ dữ liệu ban đầu chứa 17 008 dòng và 24 cột.
- **Các loại dữ liệu chính:** Bộ dữ liệu bao gồm sự kết hợp của:
  - Dữ liệu số: bao gồm số nguyên (`int64`) và số thực (`float64`) — liên quan đến số lượng động vật (at risk, cases, deaths, destroyed, slaughtered, humans affected/deaths/age), và thông tin địa lý (latitude, longitude).
  - Dữ liệu đối tượng: (`object`) — bao gồm các thông tin mô tả như nguồn báo cáo, khu vực, quốc gia, loại bệnh, loài vật, v.v.

Chương 3

TIỀN XỬ LÝ DỮ LIỆU (DATA PREPROCESSING)

Trước khi tiến hành phân tích hoặc xây dựng mô hình, dữ liệu cần được làm sạch và chuẩn hóa nhằm đảm bảo chất lượng và tính chính xác của kết quả. Quá trình tiền xử lý dữ liệu bao gồm các bước như đổi tên cột để dễ đọc và nhất quán, xử lý các giá trị thiếu hoặc bất thường, loại bỏ các bản ghi trùng lặp, chuyển đổi định dạng dữ liệu, và tạo ra các đặc trưng mới hỗ trợ phân tích chuyên sâu hơn.

3.1 Chuẩn hóa tên cột (Renaming Columns)

Để đảm bảo tính nhất quán và dễ hiểu trong phân tích, các tên cột được đổi sang dạng có định dạng chuẩn, viết hoa đầu mỗi từ và tách rõ các thành phần.

3.2 Loại bỏ các bản ghi trùng lặp và xử lý giá trị thiếu

- Các dòng dữ liệu bị trùng hoàn toàn được loại bỏ để tránh sai lệch trong thống kê và mô hình.
- Sử dụng .isnull().sum().sum() để thống kê số lượng giá trị bị thiếu trên

toàn bộ DataFrame.Các cột ngày (Observation Date, Reporting Date) được chuyển sang kiểu datetime, xử lý lỗi định dạng bằng errors='coerce'.

- Điền giá trị thiếu bằng trung vị, áp dụng cho các cột định lượng quan trọng:
  - Sum Cases
  - Sum Deaths
  - Sum At Risk
  - Sum Destroyed
  - Sum Slaughtered
- Điền các giá trị là "Unknown" cho các cột có định dạng là "string" khi có giá trị là "NAN" hoặc để trống: Serotypes, Species Description
- Nếu Observation Date bị thiếu, thay thế bằng Reporting Date trừ đi độ trễ trung vị giữa 2 cột

3.3 Phân tích định lượng

Tạo bảng mô tả cho toàn bộ các cột dạng số (numeric) và tính thêm phần trăm giá trị thiếu Dựa trên mô tả, ta sẽ loại bỏ ba cột dữ liệu `HumansDeaths`, `HumansAge`, và `HumansAffected` vì các cột này có hơn 90% giá trị bị thiếu và số lượng dữ liệu còn lại cũng rất nhỏ. Các thông tin này nằm ngoài phạm vi phân tích giám sát bệnh động vật toàn cầu bùng phát trong giai đoạn 2017–2019, bao gồm các dịch bệnh như dịch tả Châu Phi, lở mồm long móng và cúm gia cầm.

3.4 Loại bỏ các cột không cần thiết

Một số cột dữ liệu liên quan đến con người đã bị loại bỏ do chứa quá nhiều giá trị thiếu hoặc không phù hợp với mục tiêu phân tích bệnh động vật. Cụ

thể, ba cột `humansDeaths`, `humansAge` và `humansAffected` bị loại bỏ vì có trên 90% giá trị bị thiếu và số lượng dòng dữ liệu liên quan cũng rất ít. Ngoài ra, các thông tin này không nằm trong phạm vi phân tích chính là giám sát các bệnh động vật toàn cầu bùng phát trong giai đoạn 2017–2019, bao gồm các bệnh: dịch tả lợn Châu Phi (ASF), lở mồm long móng (FMD), và cúm gia cầm (AI).

3.5 Tạo các biến đặc trưng mới

- **Thời gian báo cáo trễ (report late):** Cho biết số ngày trễ giữa ngày quan sát và ngày báo cáo.
- **Vị trí địa lý tổng hợp (location):** Kết hợp giữa quốc gia và khu vực hành chính.
- **Tỷ lệ tử vong (death rate_%):** Tính tỷ lệ tử vong dựa trên số ca tử vong và số ca bệnh.

Chương 4

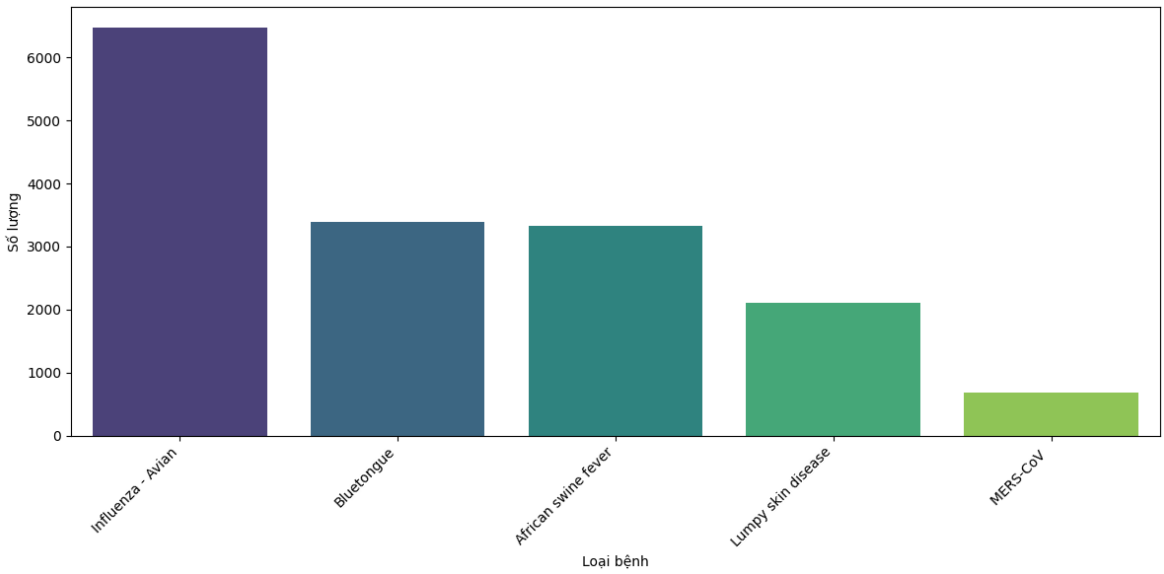
KHÁM PHÁ DỮ LIỆU VÀ TRỰC QUAN HÓA

Trước khi xây dựng mô hình phân tích hay dự đoán, việc khám phá dữ liệu (Exploratory Data Analysis – EDA) là một bước không thể thiếu. Đây là giai đoạn nhằm hiểu rõ bản chất của dữ liệu, phát hiện những đặc điểm nổi bật, mối quan hệ giữa các thuộc tính, cũng như những bất thường có thể ảnh hưởng đến chất lượng mô hình.

4.1 Sự phân bố của các loại bệnh và loài vật chính

- **Các loại bệnh chính:** Các bệnh phổ biến nhất được ghi nhận gồm:
  - *Influenza - Avian* (Cúm gia cầm)
  - *Bluetongue* (Bệnh lưỡi xanh)
  - *African swine fever* (Dịch tả lợn Châu Phi)
  - *Foot and mouth disease* (Bệnh lở mồm long móng)

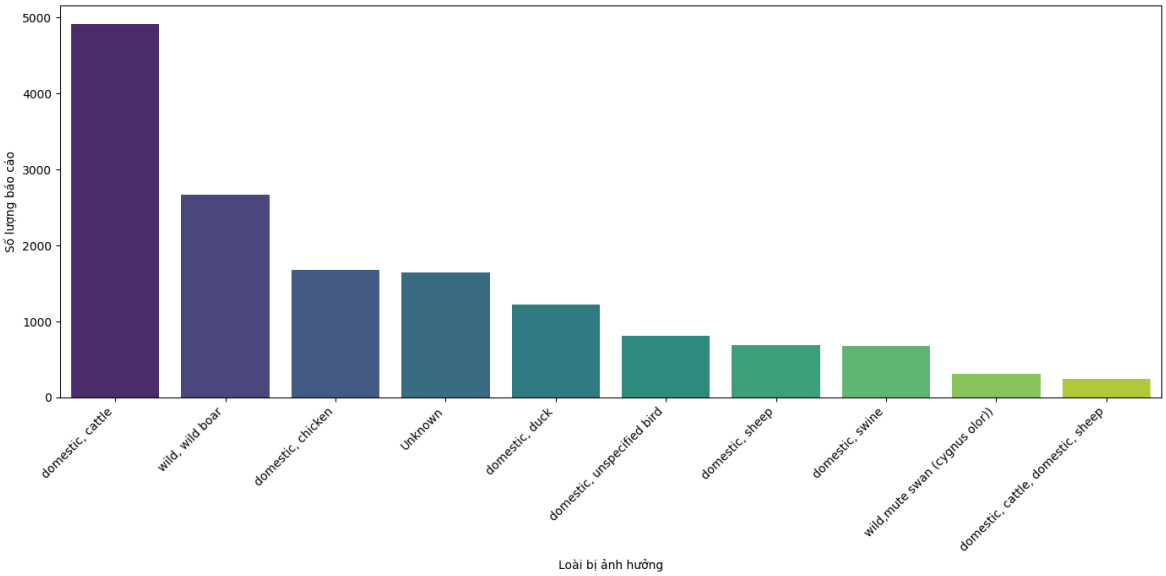
  Sự phổ biến của các bệnh này phản ánh mức độ ảnh hưởng và tần suất bùng phát trong giai đoạn và khu vực được thu thập.



Hình 4.1: Top 5 loại bệnh chính

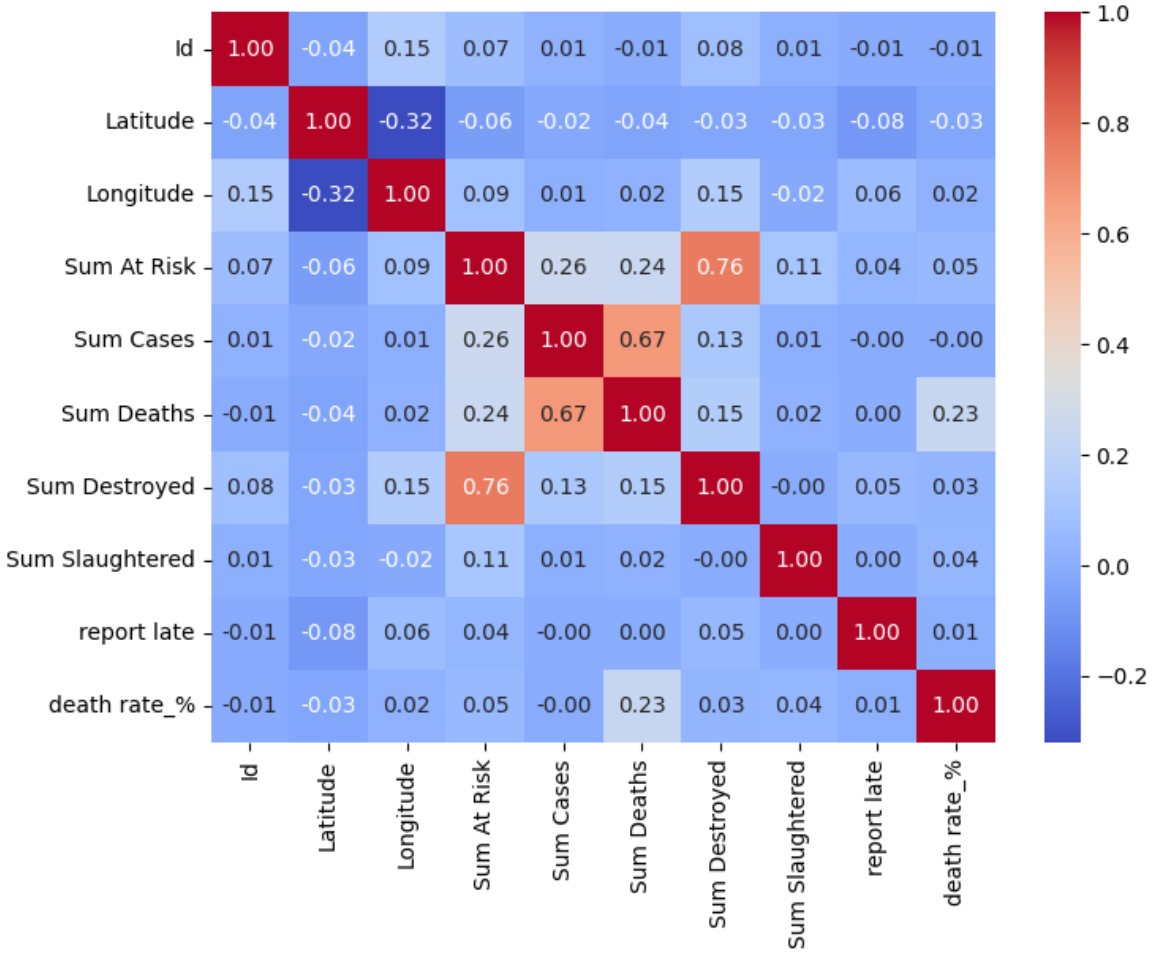
- **Các loài vật chính bị ảnh hưởng:** Dữ liệu ghi nhận ảnh hưởng trên nhiều loài, phổ biến nhất gồm:
  - *Domestic cattle* (gia súc nhà)
  - *Domestic swine* (lợn nhà)
  - *Domestic chicken* (gà nhà)
  - *Unspecified birds* (gia cầm không xác định) thuộc nhóm *domestic* và *wild*

  Điều này cho thấy dịch bệnh có xu hướng tập trung ở các loài vật nuôi phổ biến và cả quần thể chim hoang dã — những yếu tố quan trọng trong việc lây lan mầm bệnh.



Hình 4.2: Top 10 loài bị ảnh hưởng nhiều nhất

4.2 Ma trận tương quan giữa các đặc trưng



Hình 4.3: Ma trận tương quan giữa các đặc trưng

- **Tương quan mạnh giữa các đặc trưng liên quan đến số lượng:** Có tương quan dương mạnh giữa `Sum Cases`, `Sum Deaths`, `Sum Destroyed`, và `Sum Slaughtered`. Điều này là hợp lý vì khi số ca nhiễm (`Sum Cases`) tăng lên, số ca chết (`Sum Deaths`), số lượng vật bị tiêu hủy (`Sum Destroyed`) và số lượng vật bị giết thịt (`Sum Slaughtered`) do dịch bệnh cũng có xu hướng tăng theo.

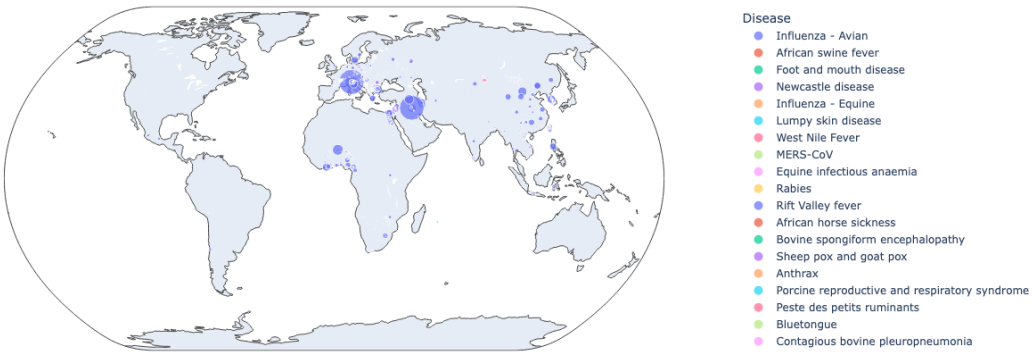
- **Tương quan với `Sum At Risk`:** Cột `Sum At Risk` (tổng số vật nuôi có nguy cơ) có thể có tương quan dương ở mức độ khác nhau với các đặc trưng số lượng khác. Tuy nhiên, mức độ tương quan có thể không quá cao vì số lượng vật nuôi có nguy cơ không phải lúc nào cũng tỷ lệ thuận

trực tiếp với số ca nhiễm hay chết.

- **Tương quan của `Latitude` và `Longitude`:** Biểu đồ sẽ cho thấy chúng có tương quan rất gần 0 với hầu hết các cột khác. Điều này là bình thường vì vị trí địa lý không trực tiếp quyết định số ca bệnh một cách tuyến tính đơn giản.
- **Tương quan của `report late`:** Cột `report late` (thời gian trễ báo cáo) có thể có tương quan (dương hoặc âm) với các đặc trưng khác, nhưng thường không quá mạnh.

4.3 Trực quan hóa Địa lý (Geospatial Visualization)

- **Các khu vực tập trung ổ dịch:** Biểu đồ giúp dễ dàng xác định được những khu vực địa lý có số lượng báo cáo dịch bệnh nhiều hơn, có thể là một số quốc gia hoặc châu lục cụ thể đang chịu ảnh hưởng nặng nề.
- **Sự phân bố của từng loại bệnh:** Màu sắc của các điểm dữ liệu thể hiện từng loại bệnh khác nhau, giúp ta thấy liệu một loại bệnh cụ thể có xu hướng tập trung ở một khu vực nhất định hay không, hoặc có sự phân bố rộng khắp.



Hình 4.4: Geospatial Visualization

Chương 5

PHÂN CỤM (CLUSTERING)

Nhóm các ổ dịch tương tự nhau dựa theo thời gian và không gian. Các phương pháp phân cụm đã thử (DBSCAN, MeanShift, K-Means, HDBSCAN). Giải thích cách chuẩn bị dữ liệu cho từng phương pháp (bao gồm cả việc scale thời gian cho DBSCAN/MeanShift/K-Means/HDBSCAN). Trực quan hóa kết quả phân cụm và nhận xét về các cụm được tạo ra. Từ đó giúp phân tích cụm theo vùng nóng của dịch bệnh (dựa theo thời gian và không gian địa lý). Trình bày kết quả phân tích nhóm theo cụm.

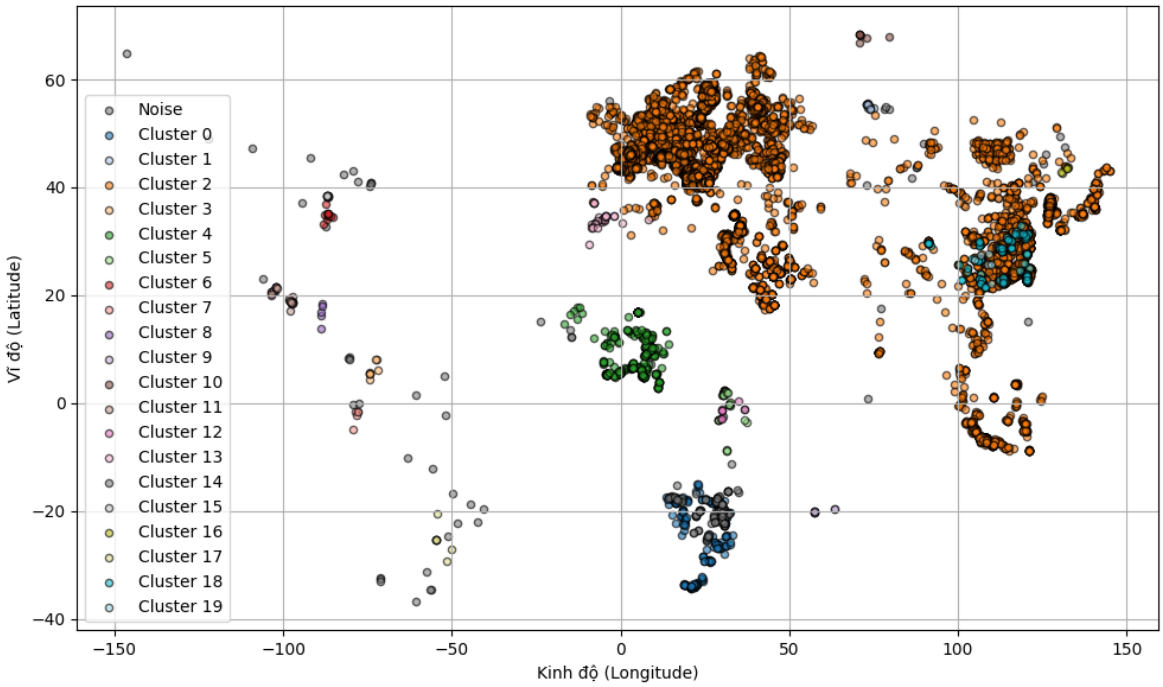
5.1 Phân cụm theo thuật toán DBSCAN

- Khái niệm: (Density-Based Spatial Clustering of Applications with Noise) là một thuật toán phân cụm dựa trên mật độ, hiệu quả trong việc xác định các cụm có hình dạng bất kỳ và xử lý nhiễu trong dữ liệu. Thuật toán này không yêu cầu số lượng cụm ban đầu và có khả năng phát hiện các cụm trong các vùng có mật độ khác nhau.[1]

- Nguyên lý hoạt động của thuật toán này là: Thứ nhất nó lấy một điểm bất kỳ sau đó nó tìm tất cả các điểm nằm trong bán kính mà nó đang xét nếu số lượng điểm lân cận lớn hơn hoặc bằng minPts thì lấy điểm đang xét làm điểm cốt lõi từ đó một cụm mới được tạo ra. Thuật toán mở rộng cụm bằng cách thêm các điểm lân cận trực tiếp vào cụm.Nếu các điểm này đúng là điểm cốt lõi thì sẽ được thêm vào cụm và tiếp tục mở rộng cụm. Các điểm không thuộc cụm nào sẽ được đánh dấu là điểm nhiễu. Quá trình này lặp lại cho đến khi tất cả các điểm đều đã được phân loại.

- Ưu điểm của DBSCAN là: Không cần biết trước số cụm xác định được các điểm nhiễu và đánh dấu nó.

- Nhược điểm của DBSCAN là: Việc lựa chọn tham số epsilon và minPts nếu lựa chọn không phù hợp sẽ đưa ra kết quả không chính xác.



Hình 5.1: Phân cụm theo thuật toán DBSCAN theo không gian và thời gian

- Hình ảnh trên thể hiện kết quả phân cụm ổ dịch theo không gian và thời gian sử dụng thuật toán DBSCAN với trục hoành (X) là kinh độ và trục tung (Y) là vĩ độ của các điểm dữ liệu.
- Như hình thì ta có thể thấy có tất cả 20 cụm và các điểm nhiễu có màu xám (những điểm không thuộc về cụm nào).
- Cụm có màu lớn nhất là cụm màu cam thể hiện một vùng dịch ổ rộng và phức tạp các cụm nhỏ hơn cho thấy các ổ dịch cục bộ hơn, rõ ràng về mặt không gian. Ngoài ra thì các điểm nhiễu cho thấy nó có thể là các ca lẻ tẻ ở một số nơi không thể gom chung thành 1 cụm.

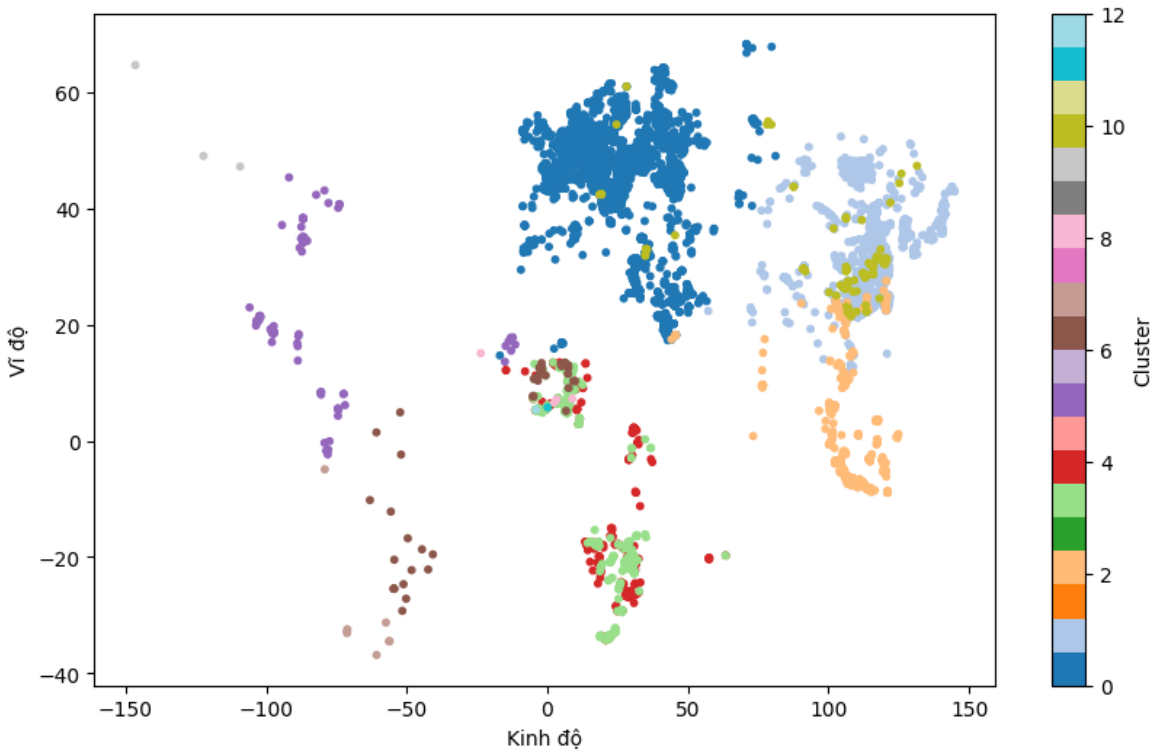
5.2 Phân cụm theo thuật toán Mean Shift

- Khái niệm: Mean Shift là một thuật toán phân cụm bằng cách dịch chuyển các điểm dữ liệu về phía những khu vực có mật độ dữ liệu cao, cho đến khi chúng hội tụ về các trung tâm cụm. [2]

- Nguyên lý hoạt động của thuật toán này là: Thứ nhất chọn một điểm dữ liệu được coi là một tâm cụm tiềm năng. Đối với mỗi điểm dữ liệu thì thuật toán sẽ tính trung bình số của các điểm lân cận (trong một bán kình nhất định) sau đó thì điểm dữ liệu sẽ dịch chuyển về phía trung bình trọng số này. Quá trình dịch chuyển được lặp lại cho đến khi các điểm dữ liệu hội tụ (không còn dịch chuyển đáng kể). Các cụm gần nhau được hợp nhất thành một cụm duy nhất.

- Ưu điểm của Mean Shift là: Không cần biết trước số lượng cụm.

- Nhược điểm của Mean Shift là: Độ phức tạp tính toán cao: Thuật toán có thể chậm khi xử lý các tập dữ liệu lớn.



Hình 5.2: Phân cụm theo thuật toán Mean Shift theo không gian và thời gian

- Hình ảnh trên thể hiện kết quả phân cụm ổ dịch theo không gian và thời gian sử dụng thuật toán Mean Shift với trục hoành (X) là kinh độ và trục tung (Y) là vĩ độ của các điểm dữ liệu.
- Như hình thì ta có thể thấy có tất cả 13 cụm được phân chia với các màu sắc khác nhau.
- Cụm có màu lớn nhất là cụm màu xanh dương đậm thể hiện một vùng dịch ổ rộng và phức tạp các cụm nhỏ hơn cho thấy các ổ dịch cục bộ hơn, rõ ràng về mặt không gian. Điểm khác từ biểu đồ này so với thuật toán DBSCAN là tất cả các điểm dữ liệu đều thuộc một cụm nào đó chứ không phải đánh dấu là điểm nhiễu.

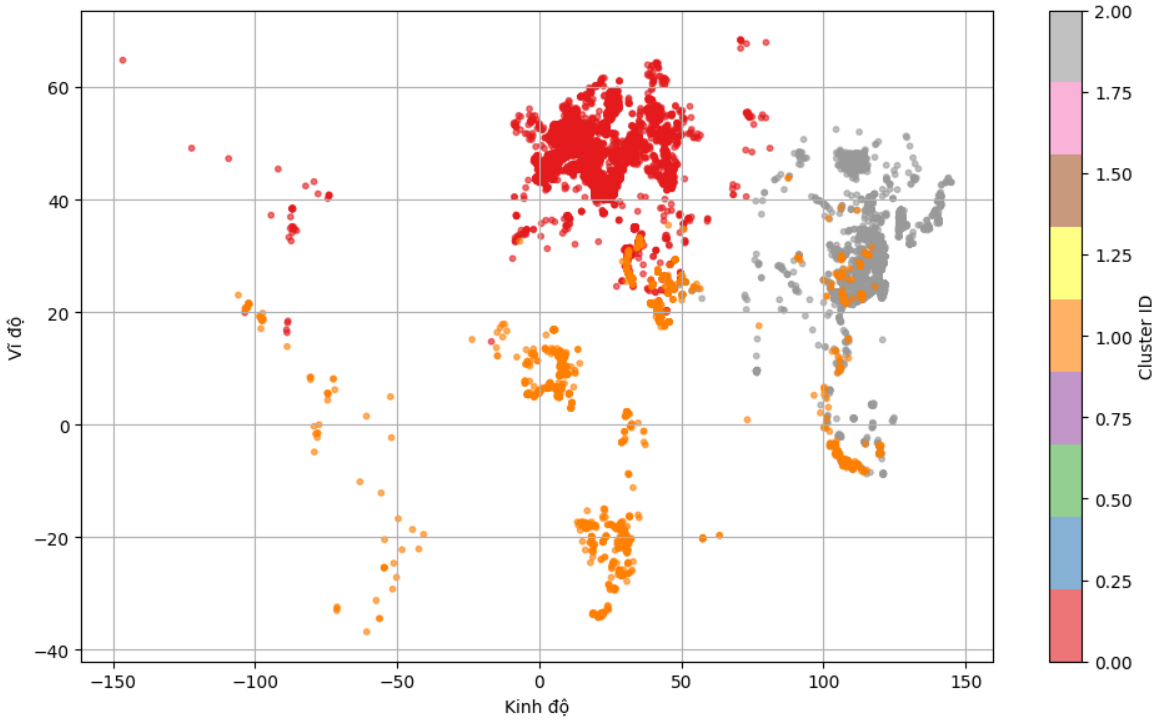
5.3 Phân cụm theo thuật toán K-Means

- Khái niệm: là một phương pháp phân cụm (clustering) trong học máy, được sử dụng để chia một tập dữ liệu thành các nhóm (cụm) dựa trên các đặc điểm tương đồng. Mục tiêu là phân chia dữ liệu sao cho các điểm dữ liệu trong cùng một cụm có độ tương đồng cao, và các cụm khác nhau có độ phân biệt rõ ràng.

- Nguyên lý hoạt động của thuật toán này là: Thứ nhất người dùng cần lựa chọn số cụm sao cho phù hợp. Tiếp theo chọn ngẫu nhiên K điểm dữ liệu để làm tâm của các cụm. Sau đó thì sẽ tính khoản cách các điểm đến tâm các cụm và điểm này sẽ thuộc cụm có khoản cách ngắn nhất. Sau đó tính lại vị trí tâm của cùng bằng cách lấy trung bình cộng của tất cả các điểm dữ liệu trong cụm đó. Lặp lại việc này cho đến khi tâm của các cụm không còn thay đổi thì dừng lại.

- Ưu điểm của K-Means là: Đơn giản, dễ hiểu, làm việc nhanh chóng trên tập dữ liệu lơn.

- Nhược điểm của DBSCAN là: Ta phải xác định được số lượng cụm trước và nhạy cảm với giá trị ngoại lai.



Hình 5.3: Phân cụm theo thuật toán K-Means theo không gian và thời gian

- Hình ảnh trên thể hiện kết quả phân cụm ổ dịch theo không gian và thời gian sử dụng thuật toán K-Means với trục hoành (X) là kinh độ và trục tung (Y) là vĩ độ của các điểm dữ liệu.
- Như hình thì ta có thể thấy có tất cả 3 cụm được phân chia với các màu sắc khác nhau. Điều này tạo nên sự khác biệt với các thuật toán khác là phải xác định trước số cụm cần phân.
- Cụm có màu lớn nhất là cụm màu đỏ thể hiện một vùng dịch ổ rộng và phức tạp các cụm nhỏ hơn cho thấy các ổ dịch cục bộ hơn, rõ ràng về mặt không gian. Điểm khác từ biểu đồ này so với thuật toán DBSCAN là tất cả các điểm dữ liệu đều thuộc một cụm nào đó chứ không phải đánh dấu là điểm nhiễu.

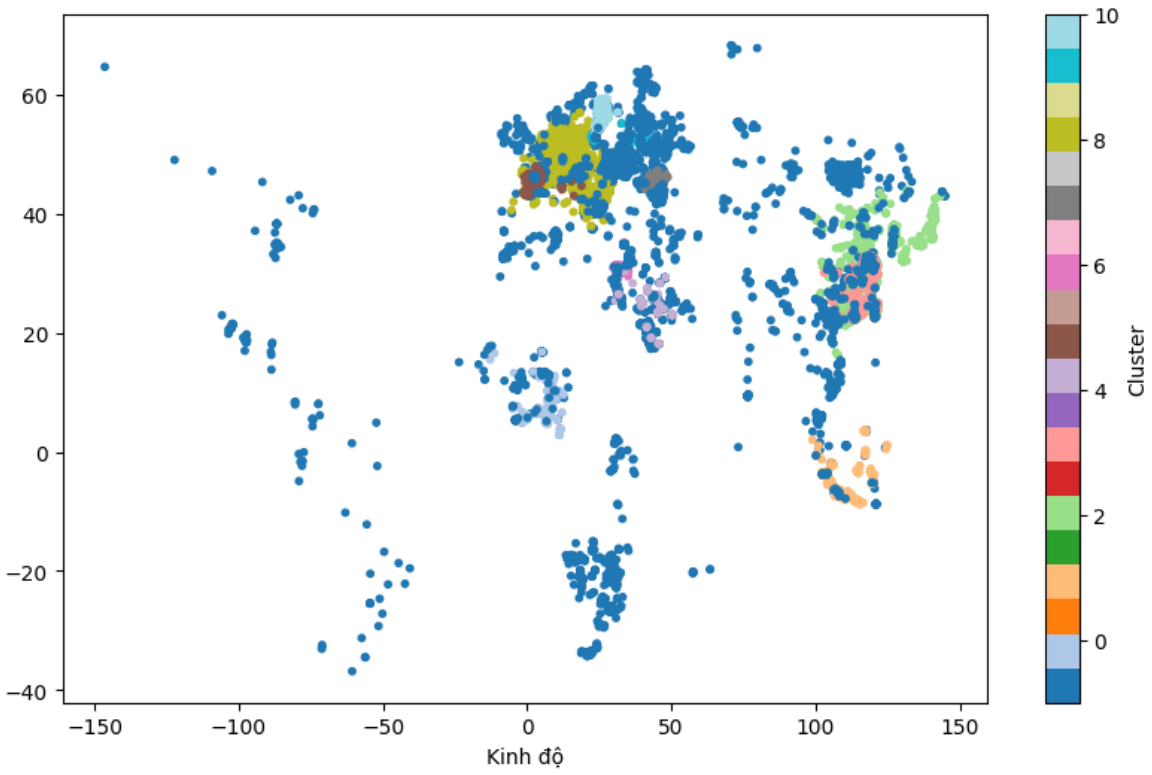
5.4 Gom cụm theo thuật toán HDBSCAN

- Khái niệm: (Hierarchical Density-Based Spatial Clustering of Applications with Noise) là một thuật toán phân cụm nâng cao của DBSCAN. Là một thuật toán phân cụm dựa trên mật độ, có khả năng tìm ra các cụm có hình dạng bất kỳ và xử lý nhiễu trong dữ liệu. [3]

- Nguyên lý hoạt động của thuật toán này là: Thứ nhất nó sẽ chuyển đổi khoản cách sang khoản cách mật độ sau đó nó sẽ Dùng các khoảng cách mật độ để xây dựng cây đồ thị tối tiểu MST (Minimum Spanning Tree). Sau đó cắt dần các cạnh từ MST theo độ dài tạo ra cây phân cấp các cụm (dendrogram) mỗi nhánh tương ứng với một cụm ở các mức mật độ khác nhau tiếp theo thì sẽ chọn cụm tối ưu từ cây phân cấp và gán nhãn cụm cho từng điểm.

- Ưu điểm của HDBSCAN là: Không cần biết trước số cụm xác định được các điểm nhiễu và đánh dấu nó và không đưa vào nó vào cụm nào.

- Nhược điểm của HDBSCAN là: Có độ phức tạp cao hơn thuật toán DBSCAN. Nếu dữ liệu có mật độ tương đương, HDBSCAN có thể không phân cụm hiệu quả.



Hình 5.4: Phân cụm theo thuật toán HDBSCAN theo không gian và thời gian

- Hình ảnh trên thể hiện kết quả phân cụm ổ dịch theo không gian và thời gian sử dụng thuật toán HDBSCAN với trục hoành (X) là kinh độ và trục tung (Y) là vĩ độ của các điểm dữ liệu.
- Theo kết quả phân tích thì ta có tất 11 cụm được phân chia cùng với các điểm nhiễu không thuộc cụm nào.
- Thuật toán đã phát hiện rất nhiều cụm, phản ánh rõ sự phân bố phức tạp và đa dạng của các ổ dịch trong không gian địa lý. Sự xuất hiện của nhiều màu nhỏ, rải rác trong cùng khu vực địa lý cho thấy sự chồng lấn và biến thiên cục bộ, điều mà các thuật toán gom cụm cổ điển khó phát hiện.

5.5 Nhận xét

Cả 4 thuật toán (DBSCAN, Mean Shift, K-Means, HDBSCAN) đều trực quan việc phân các cụm phù hợp tuy nhiên mỗi thuật toán đều có điểm mạnh và yếu khác nhau. Đối với DBSCAN nó có thể không cần biết trước số cụm đầu vào và nó làm tốt việc phân chia các điểm nhiễu tuy nhiên nó phụ thuộc nhiều vào tham số eps (khoảng cách bán kính) nếu chọn sai tham số eps thì việc phân cụm sẽ không hiệu quả. Còn đối với thuật toán Mean Shift nó có thể không cần biết trước số cụm đầu vào tuy nhiên nó không thể đánh dấu các điểm nhiên do đó tát cả các điểm dữ liệu đều nằm trong 1 cụm nào đó và nhược điểm lớn đó chính là độ phức tạp của thuật toán rất cao vì nó phải tính từng điểm 1 đối với một bộ dữ liệu lớn thì đây là một hạn chế lớn. Đối với thuật toán K-Means thì có ưu điểm là tốc độ cao xử lý dữ liệu nhanh chóng tuy nhiên nó phải cho biết được trước số cụm cần phân điều này khó khăn nếu không chọn được số cụm phù hợp thì việc phân cụm sẽ không hiệu quả. Khắc phục hầu như các nhược điểm của các thuật toán trên thì có HDBSCAN vừa không cần biết trước số cụm đầu vào, vừa đánh dấu được các điểm nhiễu, vừa có độ phức tạp tốt và không phụ thuộc vào tham số như DBSCAN. Do đó HDBSCAN là thuật toán tốt nhất để lựa chọn cho việc phân cụm theo thời gian và không gian của các ổ dịch.

Chương 6

KHAI THÁC MẪU (PATTERN MINING)

6.1 Khai phá tập mục phổ biến (Frequent Itemset Mining)

6.1.1 Cơ sở lý thuyết

- **Khái niệm:** Frequent Itemset Mining, hay khai phá tập mục phổ biến, là một kỹ thuật trong khai phá dữ liệu nhằm tìm ra những tập hợp các mục (itemsets) thường xuyên xuất hiện cùng nhau trong một tập dữ liệu giao dịch. [4]

- **Cách hoạt động:** Frequent Itemset Mining hoạt động bằng cách duyệt qua tập dữ liệu giao dịch để tính toán tần suất xuất hiện (support) của từng tập hợp mục (itemset).

  Các itemset có độ hỗ trợ lớn hơn hoặc bằng ngưỡng tối thiểu (min_support) sẽ được giữ lại. Sau đó, các tập mục lớn hơn được sinh ra từ các tập phổ biến trước đó, và quá trình này lặp lại cho đến khi không còn tập nào mới đủ điều kiện.[4]

- **Một vài khái niệm quan trọng:** [4]

  – **Item:** Một mục đơn lẻ trong giao dịch. Ví dụ: *"Sữa"*, *"Trứng"*.

- **Itemset (Tập hợp mục):** Một tập hợp gồm một hoặc nhiều mặt hàng. Ví dụ: {*Sữa*}, {*Sữa, Tã giấy*}.
- **Transaction (Giao dịch):** Một dòng trong dữ liệu giao dịch, chứa nhiều item.
- **Support (Độ hỗ trợ):** Chỉ mức độ phổ biến của một itemset, được tính bằng công thức:

$$\text{support}(X) = \frac{\text{số lượng giao dịch chứa } X}{\text{tổng số giao dịch}}$$

- **Frequent Itemset (Tập phổ biến):** Một itemset có support lớn hơn hoặc bằng một ngưỡng tối thiểu do ta đặt ra, gọi là min_support.
- **Association Rule (Luật kết hợp):** Có dạng $X \to Y$ (Nếu X thì Y). Ví dụ: {Sữa} $\to$ {Tã giấy} (Nếu khách mua Sữa thì họ cũng sẽ mua Tã giấy).
- **Confidence (Độ tin cậy):** Đo lường "độ chính xác" của luật, được tính bằng:

$$\text{Confidence}(X \to Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

- **Lift (Độ nâng):** Đo lường mức độ thú vị, bất ngờ của luật:

$$\text{Lift}(X \to Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X) \times \text{Support}(Y)}$$

- **Các thuật toán phổ biến:**[4]
  - **Thuật toán Apriori:** Sử dụng nguyên lý "cắt tỉa" không gian tìm kiếm: Mọi tập con của một tập phổ biến cũng phải là một tập phổ biến.
  - **Thuật toán FP-Growth:** Thay vì "tạo ứng viên - kiểm tra" lặp đi lặp lại, FP-Growth nén toàn bộ thông tin về các mẫu phổ biến vào một cấu trúc cây thông minh gọi là FP-Tree, sau đó khai thác trực tiếp trên cây này.

6.1.2 Triển khai

Mã hóa giao dịch về dạng Transactions (True/False)

	African horse sickness	African swine fever	Albania	Algeria	Angola	Anthrax	Argentina	Armenia	Austria	Bahrain	...	wild teal	wild tufted duck	wild tundra swan	wild unspecified bird	wild unspecified mammal	wild white stork	wild white tailed eagle	wild white-winged black tern	wild whooper swan	wild wood sandpiper
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
1	False	True	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
4	False	True	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
17003	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
17004	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
17005	False	False	False	False	False	False	False	False	False	False	...	False	False	False	True	False	False	False	False	False	False
17006	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False
17007	False	False	False	False	False	False	False	False	False	False	...	False	False	False	False	False	False	False	False	False	False

17008 rows × 314 columns

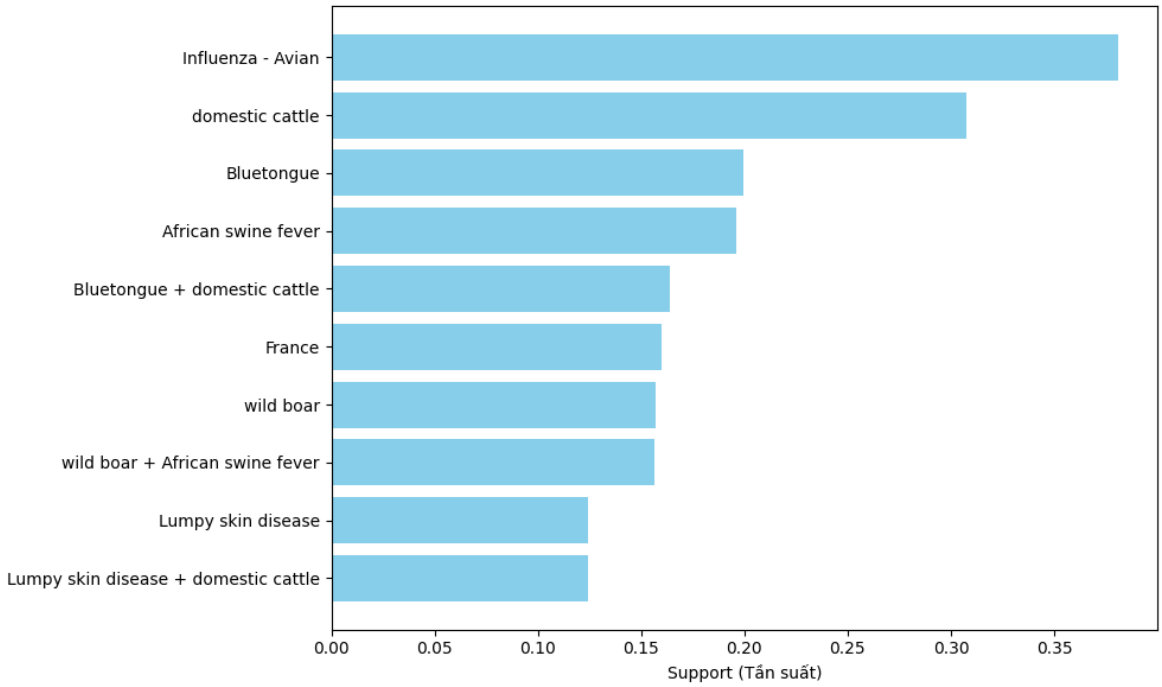
Hình 6.1: Mã hóa về dạng Transactions

Áp dụng thuật toán Apriori

	support	itemsets
13	0.380938	(Influenza - Avian)
32	0.307679	(domestic cattle)
2	0.199377	(Bluetongue)
0	0.195790	(African swine fever)
56	0.163864	(Bluetongue, domestic cattle)
8	0.159984	(France)
41	0.156985	(wild boar)
50	0.156515	(wild boar, African swine fever)
19	0.124177	(Lumpy skin disease)
95	0.124118	(Lumpy skin disease, domestic cattle)

Hình 6.2: Thuật toán Apriori

Sau khi áp dụng thuật toán Apriori để khai thác tập mục thường xuyên, ta thu được một số kết quả đáng chú ý như sau:



Hình 6.3: Top 10 Tổ hợp phổ biến

Sinh các luật kết hợp từ các item phổ biến

	antecedents	consequents	support	confidence	lift
48	Albania	domestic cattle	0.050447	1.0	3.250143
51	Albania	Lumpy skin disease	0.050447	1.0	8.053030
25	Estonia	African swine fever	0.054504	1.0	5.107508
21	domestic duck	Influenza - Avian	0.082020	1.0	2.625096
37	Estonia, wild boar	African swine fever	0.053563	1.0	5.107508
42	Albania, Lumpy skin disease	domestic cattle	0.050447	1.0	3.250143
46	Albania	Lumpy skin disease, domestic cattle	0.050447	1.0	8.056845
44	Albania, domestic cattle	Lumpy skin disease	0.050447	1.0	8.053030
85	Lithuania, wild boar	African swine fever	0.031867	1.0	5.107508
74	Saudi Arabia, unknown	MERS-CoV	0.036571	1.0	25.197037

Hình 6.4: Sinh ra các luật Association Rules

Phân tích kết quả thu được:

- Các bệnh như *Influenza - Avian*, *Bluetongue*, *African swine fever*, cùng với loài *domestic cattle* là những mục xuất hiện với tần suất cao trong dữ liệu

- Các tổ hợp như (*domestic cattle*, *Bluetongue*) hoặc (*African swine fever*, *wild boar*) thể hiện mối liên kết mạnh giữa các bệnh và loài vật, phản ánh xu hướng cùng xuất hiện trong các báo cáo

- Quốc gia *France* xuất hiện thường xuyên trong các giao dịch, cho thấy khả năng cao nước này là điểm nóng về dịch bệnh trong giai đoạn được nghiên cứu

- Các tập hợp mục trong top 10 có độ hỗ trợ dao động từ **12.41%** đến **38.09%**, chứng tỏ rằng chúng xuất hiện trong một tỷ lệ đáng kể các giao dịch, và là những mẫu có giá trị để phân tích sâu hơn

```
Nếu loài A ở vùng B thì bệnh thường là C:
Nếu loài wild boar ở Estonia, thì bệnh thường là African swine fever (confidence=1.00, lift=5.11)
Nếu loài domestic cattle ở Albania, thì bệnh thường là Lumpy skin disease (confidence=1.00, lift=8.05)
Nếu loài wild boar ở Lithuania, thì bệnh thường là African swine fever (confidence=1.00, lift=5.11)
Nếu loài unknown ở Saudi Arabia, thì bệnh thường là MERS-CoV  (confidence=1.00, lift=25.20)
Nếu loài wild boar ở Latvia, thì bệnh thường là African swine fever (confidence=1.00, lift=5.11)
Nếu loài domestic duck ở France, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.63)
Nếu loài domestic sheep ở Serbia, thì bệnh thường là Bluetongue (confidence=1.00, lift=5.02)
Nếu loài domestic duck ở Egypt, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.63)
Nếu loài domestic duck ở China, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.63)
Nếu loài domestic duck ở Republic of Korea, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.63)
Nếu loài domestic chicken ở Republic of Korea, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.63)
Nếu loài domestic chicken ở Indonesia, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.63)
Nếu loài wild boar ở Poland, thì bệnh thường là African swine fever (confidence=1.00, lift=5.11)
Nếu loài domestic cattle ở Serbia, thì bệnh thường là Bluetongue (confidence=1.00, lift=5.02)
Nếu loài domestic chicken ở China, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.63)
Nếu loài domestic chicken ở Nigeria, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.63)
Nếu loài domestic chicken ở Egypt, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.63)
Nếu loài unknown ở China, thì bệnh thường là Influenza - Avian (confidence=1.00, lift=2.62)
Nếu loài domestic cattle ở The former Yugoslav Republic of Macedonia, thì bệnh thường là Lumpy skin disease (confidence=0.99, lift=8.01)
Nếu loài domestic swine ở Ukraine, thì bệnh thường là African swine fever (confidence=0.99, lift=5.08)
Nếu loài domestic cattle ở France, thì bệnh thường là Bluetongue (confidence=0.99, lift=4.98)
Nếu loài domestic cattle ở Greece, thì bệnh thường là Lumpy skin disease (confidence=0.99, lift=7.98)
Nếu loài domestic cattle ở Italy, thì bệnh thường là Bluetongue (confidence=0.99, lift=4.96)
Nếu loài domestic swine ở Russian Federation, thì bệnh thường là African swine fever (confidence=0.99, lift=5.04)
Nếu loài domestic cattle ở Russian Federation, thì bệnh thường là Lumpy skin disease (confidence=0.99, lift=7.94)
Nếu loài domestic cattle ở Bulgaria, thì bệnh thường là Lumpy skin disease (confidence=0.99, lift=7.94)
```

Hình 6.5: Kết quả thu được khi làm Frequent Itemset Mining

6.2 Khai thác sự tuần tự (Sequential Pattern Mining)

6.2.1 Cơ sở lý thuyết

- **Khái niệm:** *Sequential Pattern Mining*, hay khai phá mẫu tuần tự, là một kỹ thuật trong khai phá dữ liệu nhằm tìm ra các chuỗi (*sequences*) gồm các mục (*items*) thường xuất hiện theo một trình tự nhất định trong một tập dữ liệu giao dịch tuần tự (*sequential dataset*) [5].

- **Cách triển khai:**

  – Tạo các chuỗi ứng viên.

  – Đếm số lần xuất hiện của chúng trong tập dữ liệu.

  – Lọc ra các chuỗi có độ hỗ trợ lớn hơn hoặc bằng ngưỡng *min_ support*.

  – Lặp lại với các chuỗi dài hơn cho đến khi không còn chuỗi nào thỏa điều kiện.[5]

- **Một số định nghĩa quan trọng:** [5]

  – **Sequence (Chuỗi tuần tự):** Một chuỗi gồm các *itemset* được sắp xếp theo thứ tự thời gian.
  Ví dụ: $\langle\{A\}, \{B, C\}, \{D\}\rangle$ nghĩa là A xảy ra trước, sau đó B và C đồng thời, cuối cùng là D.

  – **Sequential Pattern (Mẫu tuần tự):** Một chuỗi con xuất hiện trong dữ liệu với độ hỗ trợ lớn hơn hoặc bằng ngưỡng *min_ support*.

- **Một số thuật toán phổ biến:**[5]

  – **Dựa trên Apriori (GSP - Generalized Sequential Pattern):** Áp dụng nguyên lý Apriori mở rộng: "Mọi trình tự con của một trình tự phổ biến cũng phải là trình tự phổ biến." Phương pháp này tạo

và kiểm tra tập ứng viên lặp đi lặp lại, có thể tốn thời gian khi dữ liệu lớn.

– **Dựa trên Pattern-Growth (PrefixSpan):** Không tạo ứng viên một cách tường minh. Thay vào đó, nó khai thác các mẫu theo phương pháp chia nhỏ cơ sở dữ liệu thành các phần gọi là *projected database* dựa trên các tiền tố (prefix), rồi tiếp tục khai phá trên đó. Cách tiếp cận này nhanh và hiệu quả hơn trong nhiều trường hợp.

6.2.2 Triển khai

Tạo mới cột event, gom nhóm theo country, loại bỏ trùng lặp và lọc chuỗi: Tạo danh sách chuỗi sự kiện theo từng quốc gia bằng cách nhóm dữ liệu theo quốc gia, sắp xếp theo thời gian, loại bỏ sự kiện trùng lặp liên tiếp và lọc các chuỗi có độ dài hợp lệ

Áp dụng thuật toán PrefixSpan để khai phá các mẫu chuỗi tuần tự có độ dài từ 2 đến 4 sự kiện và chứa nhiều loại sự kiện khác nhau, sau đó sắp xếp các mẫu theo mức độ phổ biến (*support*) để chọn ra các chuỗi đáng chú ý.

Kết quả của thuật toán PrefixSpan

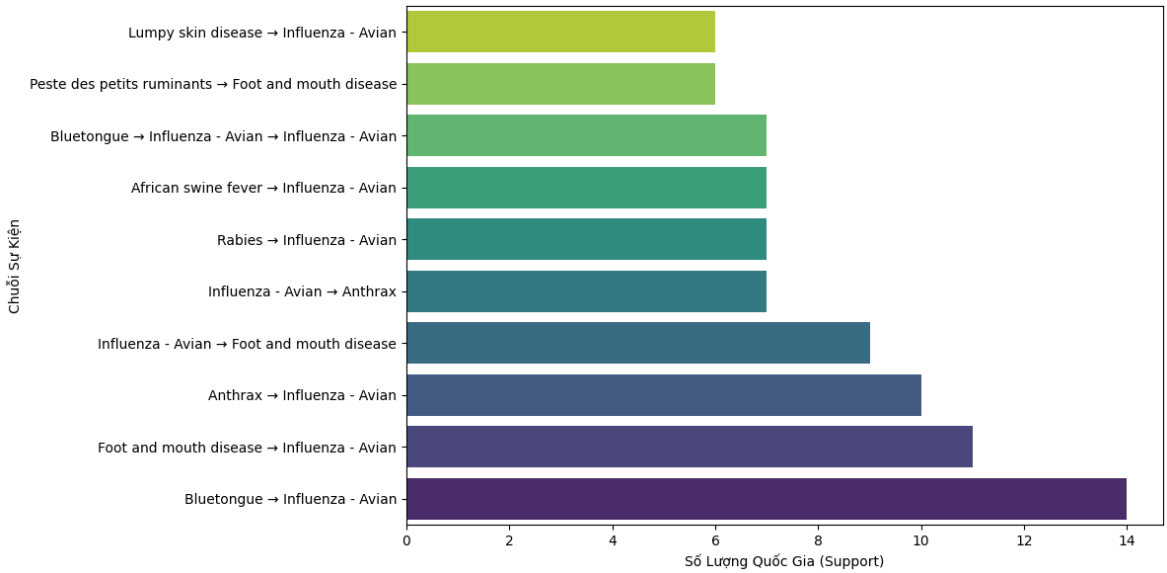
```
Top 10 chuỗi tuần tự phổ biến nhất:
Support: 14 | Sequence: Bluetongue → Influenza – Avian
Support: 11 | Sequence: Foot and mouth disease → Influenza – Avian
Support: 10 | Sequence: Anthrax → Influenza – Avian
Support: 9 | Sequence: Influenza – Avian → Foot and mouth disease
Support: 7 | Sequence: Influenza – Avian → Anthrax
Support: 7 | Sequence: Rabies → Influenza – Avian
Support: 7 | Sequence: African swine fever → Influenza – Avian
Support: 7 | Sequence: Bluetongue → Influenza – Avian → Influenza – Avian
Support: 6 | Sequence: Peste des petits ruminants → Foot and mouth disease
Support: 6 | Sequence: Lumpy skin disease → Influenza – Avian
```

Hình 6.6: Kết quả thuật toán PrefixSpan

Kết quả cho thấy *Influenza - Avian* là bệnh xuất hiện thường xuyên trong các chuỗi sự kiện, đặc biệt là sau các bệnh như *Bluetongue*, *Foot and mouth*

disease, *Anthrax* và *Rabies*. Điều này cho thấy *Influenza - Avian* có xu hướng xảy ra kế tiếp sau nhiều loại bệnh khác, phản ánh khả năng lây lan cao hoặc mối liên hệ giữa các đợt bùng phát bệnh.

Biểu đồ trực quan hóa



Hình 6.7: Top 10 Chuỗi Sự Kiện Bệnh Dịch Phổ Biến Nhất

Chương 7

HỒI QUY (REGRESSION)

Regression (hồi quy) trong khai phá dữ liệu (data mining) là một kỹ thuật dùng để dự đoán giá trị của một biến liên tục (biến số lượng) dựa trên giá trị của một hoặc nhiều biến đầu vào[6]. Với bộ dữ liệu EMPRES, có thể áp dụng hồi quy (regression) để dự đoán số ca bệnh, số con chết hoặc số bị tiêu huỷ dựa vào các yếu tố không gian, thời gian và dịch bệnh.

7.1 Mô hình Decision Tree Regressor dự đoán Sum Cases

Cây quyết định hồi quy (Decision Tree Regressor) là một thuật toán học máy thuộc nhóm cây quyết định được sử dụng cho các bài toán hồi quy, tức là dự đoán một biến mục tiêu liên tục (số lượng). Nó hoạt động bằng cách chia dữ liệu thành các tập con dựa trên giá trị của các đặc trưng, tạo ra một cấu trúc cây.[7]

7.1.1 Đánh giá mô hình Decision Tree Regressor

```
Decision Tree Regressor - Mean Squared Error (MSE): 6170606.474520217
Decision Tree Regressor - R-squared (R2): -0.5067160592947044

Prediction Results (first 5 rows):
```

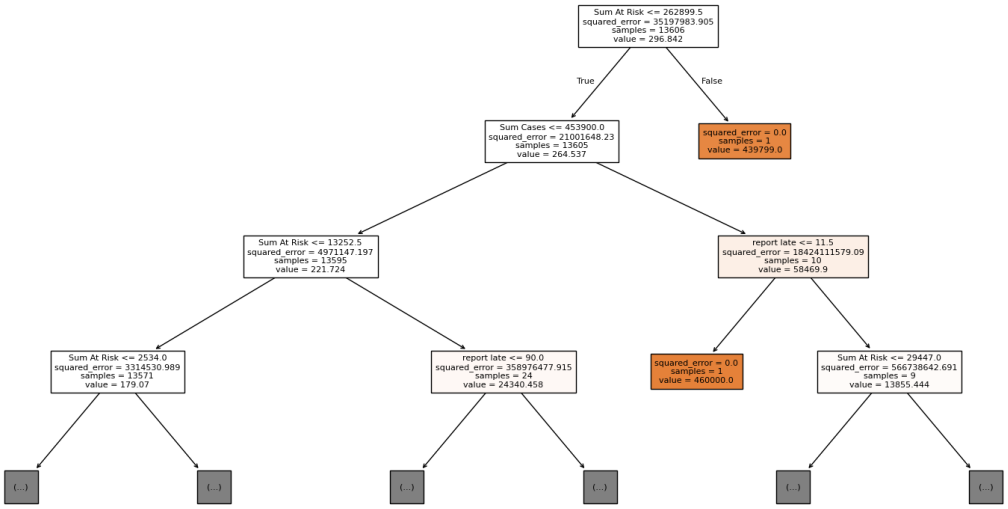
	Actual	Predicted
7665	1.0	2.0
3902	1.0	1.0
744	1.0	1.0
6438	2.0	1.0
9187	11.0	2.0

Hình 7.1: Kết quả dự đoán Sum Cases (Decision Tree Regressor)

Kết quả từ mô hình Decision Tree Regressor dự đoán *Sum Cases* từ các đặc trưng 'Sum At Risk', 'Sum Deaths', 'report late', 'Latitude', 'Longitude', 'Disease', 'Region':

- **Mean Squared Error (MSE)** đạt giá trị **6,170,606.47**, cho thấy sai số bình phương trung bình giữa giá trị dự đoán và thực tế là rất lớn. Điều này phản ánh mô hình dự đoán chưa chính xác.
- **R-squared (R^2)** có giá trị **-0.5067**, là một chỉ số tiêu cực. Điều này cho thấy mô hình hoạt động kém hơn cả việc chỉ đơn giản sử dụng trung bình của toàn bộ tập dữ liệu để dự đoán.
- **Dự đoán cụ thể (5 dòng đầu)** cho thấy mô hình thường dự đoán các giá trị nhỏ (ví dụ: 1.0, 2.0) ngay cả khi giá trị thực tế lớn hơn đáng kể (ví dụ: 11.0). Điều này cho thấy mô hình chưa học được các mẫu có liên quan đến các giá trị cao trong dữ liệu.

7.1.2 Trực quan hoá biểu đồ Decision Tree Regressor



Hình 7.2: Decision Tree Regressor (Max Depth=3)

- **Độ sâu giới hạn:** Việc giới hạn độ sâu của cây (ở đây là 3) giúp biểu đồ dễ hiểu hơn, nhưng cũng có nghĩa là cây không được huấn luyện hết mức và có thể bỏ lỡ các mối quan hệ phức tạp hơn trong dữ liệu.
- **Quy tắc quyết định:** Ta có thể theo dõi các đường đi từ nút gốc đến nút lá để hiểu các quy tắc mà mô hình sử dụng để dự đoán `Sum Cases`. Ví dụ: nếu `Disease_Influenza - Avian > 0.5`, `Sum At Risk > X`, ... thì dự đoán `Sum Cases` là `Y`.

7.2 Mô hình Linear Regression dự đoán Sum Cases

Hồi quy tuyến tính (Linear Regression) là một thuật toán học máy (Machine Learning) thuộc lớp học có giám sát (Supervised Learning). Nó được sử dụng để tìm ra mối quan hệ tuyến tính giữa một biến phụ thuộc và một hoặc nhiều biến độc lập.[8]

7.2.1 Đánh giá mô hình Linear Regression

```
Mean Squared Error (MSE): 2530079.682933488
R-squared (R2): 0.382214421659818 56
```

	Actual	Predicted
7665	1.0	141.507295
3902	1.0	120.227965
744	1.0	-94.057320
6438	2.0	99.478289
9187	11.0	123.438776

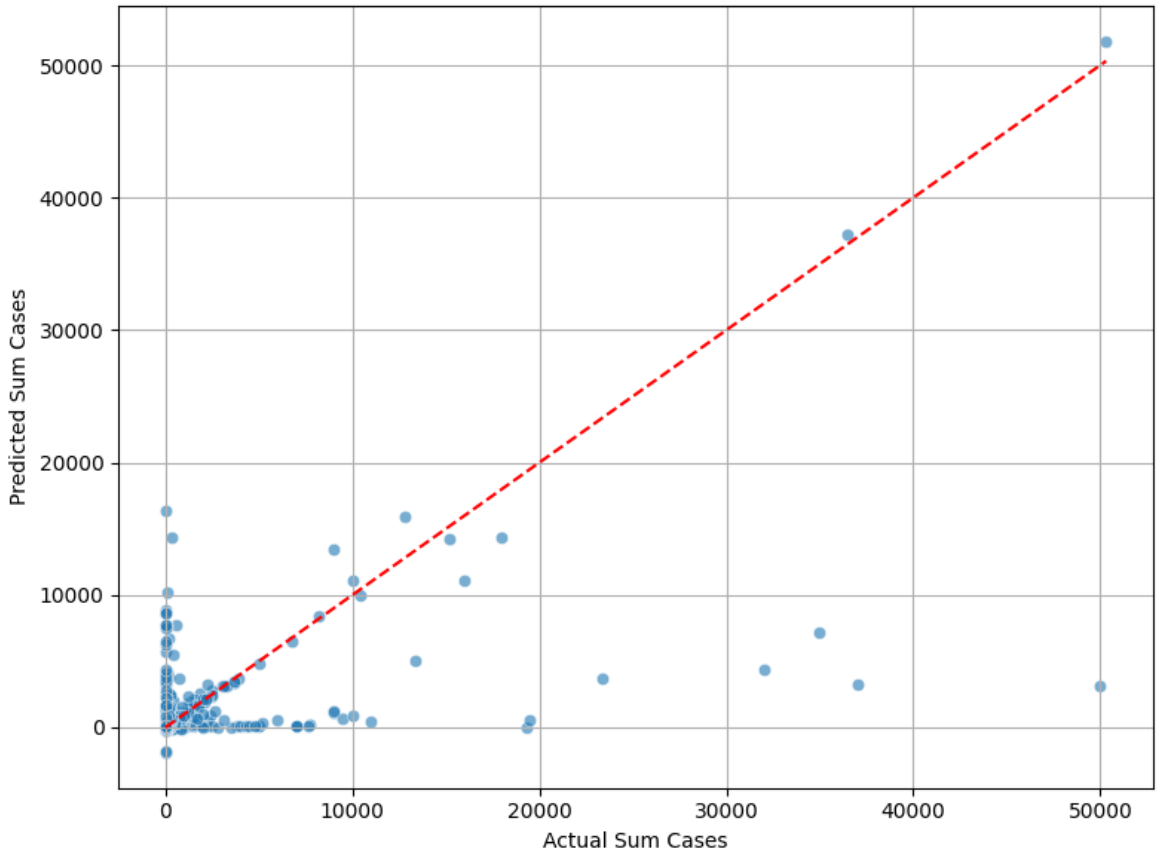
Hình 7.3: Kết quả dự đoán Sum Cases (Linear Regression)

Kết quả từ mô hình Hồi quy Tuyến tính (Linear Regression) dự đoán `Sum Cases` từ các đặc trưng 'Sum At Risk', 'Sum Deaths', 'report late', 'Latitude', 'Longitude', 'Disease', 'Region':

- **Mean Squared Error (MSE)**: Giá trị MSE là **2,530,079.68**. Tương tự như Decision Tree Regressor, đây là một giá trị MSE khá lớn, cho thấy sai số trung bình bình phương giữa giá trị thực tế và dự đoán vẫn cao.
- **R-squared ($\mathbf{R^2}$)**: Giá trị R-squared là **0.3822**. Chỉ số này cho biết mô hình hồi quy tuyến tính hiện tại giải thích được khoảng **38.22%** phương sai của biến mục tiêu `Sum Cases`. Điều này có nghĩa là khoảng **61.78%** biến động của `Sum Cases` vẫn chưa được giải thích bởi các đặc trưng được đưa vào mô hình.
- **Kết quả dự đoán (5 dòng đầu tiên)**: Các giá trị dự đoán là các số thực (có thể âm hoặc dương), trong khi `Sum Cases` là một số nguyên

không âm. Điều này cho thấy mô hình tuyến tính đang gặp khó khăn trong việc mô hình hóa bản chất của biến mục tiêu. Các dự đoán này cũng có sự sai lệch lớn so với giá trị thực tế.

7.2.2 Trực quan hoá biểu đồ Actual vs Predicted



Hình 7.4: Biểu đồ Actual vs Predicted Sum Cases (Linear Regression)

- **Tập trung ở góc dưới bên trái:** Phần lớn các điểm dữ liệu tập trung dày đặc ở góc dưới bên trái của biểu đồ, nơi cả giá trị thực tế và giá trị dự đoán đều thấp.
- **Phân tán khi giá trị lớn:** Khi giá trị `Sum Cases` thực tế tăng lên, các điểm dữ liệu trở nên phân tán hơn và nằm xa đường chéo màu đỏ. Mức độ phân tán tăng lên phản ánh sự không ổn định trong khả năng dự đoán của mô hình tại các giá trị lớn.
- **Điểm nằm dưới đường chéo:** Nhiều điểm nằm dưới đường chéo màu

đỏ, đặc biệt ở vùng giá trị lớn. Điều này cho thấy mô hình có xu hướng dự đoán thấp hơn thực tế khi số ca bệnh cao, dẫn đến sai số nghiêm trọng ở các giá trị cực trị.

- **Có thể có dự đoán âm:** Kết quả dự đoán từ mô hình Linear Regression cho thấy có thể xuất hiện giá trị âm trong đầu ra, không hợp lý trong bối cảnh dự đoán số ca bệnh.

7.3 Mô hình Random Forest Regressor dự đoán Sum At Risk

Hồi quy rừng ngẫu nhiên (Random Forest Regressor) là một thuật toán học máy có giám sát, sử dụng một tập hợp các cây quyết định để đưa ra dự đoán. Thuật toán này kết hợp các kết quả của nhiều cây quyết định để tạo ra một dự đoán duy nhất, thường là trung bình hoặc trung bình có trọng số của các dự đoán từ các cây riêng lẻ. [9]

7.3.1 Đánh giá mô hình Random Forest

```
Random Forest Regressor - Mean Squared Error (MSE) for Sum At Risk: 527855391.6194876
Random Forest Regressor - R-squared (R2) for Sum At Risk: 0.5233499881035808

Prediction Results for Sum At Risk (first 5 rows):
```

	Actual	Predicted
7665	224.0	131.630000
3902	90.0	88.150000
744	90.0	90.000000
6438	2.0	3.778333
9187	30.0	101.860000

Hình 7.5: Kết quả dự đoán Sum At Risk (Random Forest Regressor)

Kết quả từ mô hình Random Forest Regressor dự đoán Sum At Risk từ các đặc trưng 'Sum Cases', 'Sum Deaths', 'report late', 'Latitude', 'Longitude',

'Disease', 'Region':

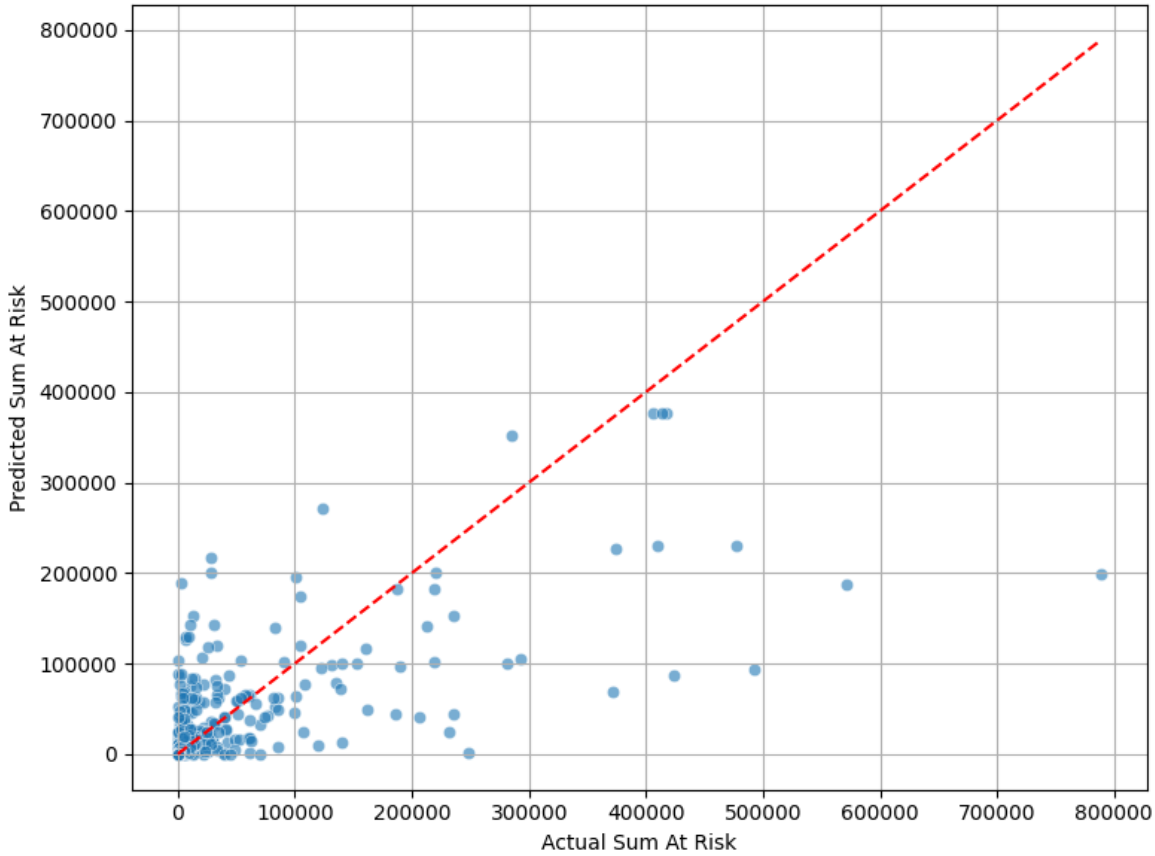
- Về chỉ số R-squared (R^2): Giá trị $R^2 = 0.523$ có nghĩa là mô hình có thể giải thích được khoảng 52.3% sự biến động của biến mục tiêu *Sum At Risk* dựa trên các đặc trưng đầu vào như *Sum Cases*, *Sum Deaths*, *Region*...Đây là kết quả tương đối tốt, đặc biệt khi so sánh với mô hình hồi quy tuyến tính trước đó ($R^2 = 0.38$). Điều này cho thấy mô hình phi tuyến như Random Forest phù hợp hơn với dữ liệu

- Về chỉ số Mean Squared Error (MSE): Giá trị $MSE = 527{,}855{,}391.62$ là một con số rất lớn. MSE phản ánh sai số bình phương trung bình giữa giá trị thực tế và dự đoán. MSE cao cho thấy mô hình bị ảnh hưởng bởi một số giá trị ngoại lệ (outliers), thường là các ổ dịch với số ca bệnh rất lớn.

- Kết quả dự đoán (5 dòng đầu tiên):

  - **Dòng 744 và 3902:** Mô hình dự đoán rất chính xác (*Actual: 90.0*, *Predicted: 90.0 và 88.15*).
  - **Dòng 6438:** Sai số thấp (*Actual: 2.0*, *Predicted: 3.78*).
  - **Dòng 7665 và 9187:** Sai số lớn, đặc biệt dòng 9187 có *Actual: 30* nhưng *Predicted: 101.86*.

7.3.2 Trực quan hoá biểu đồ Actual vs Predicted



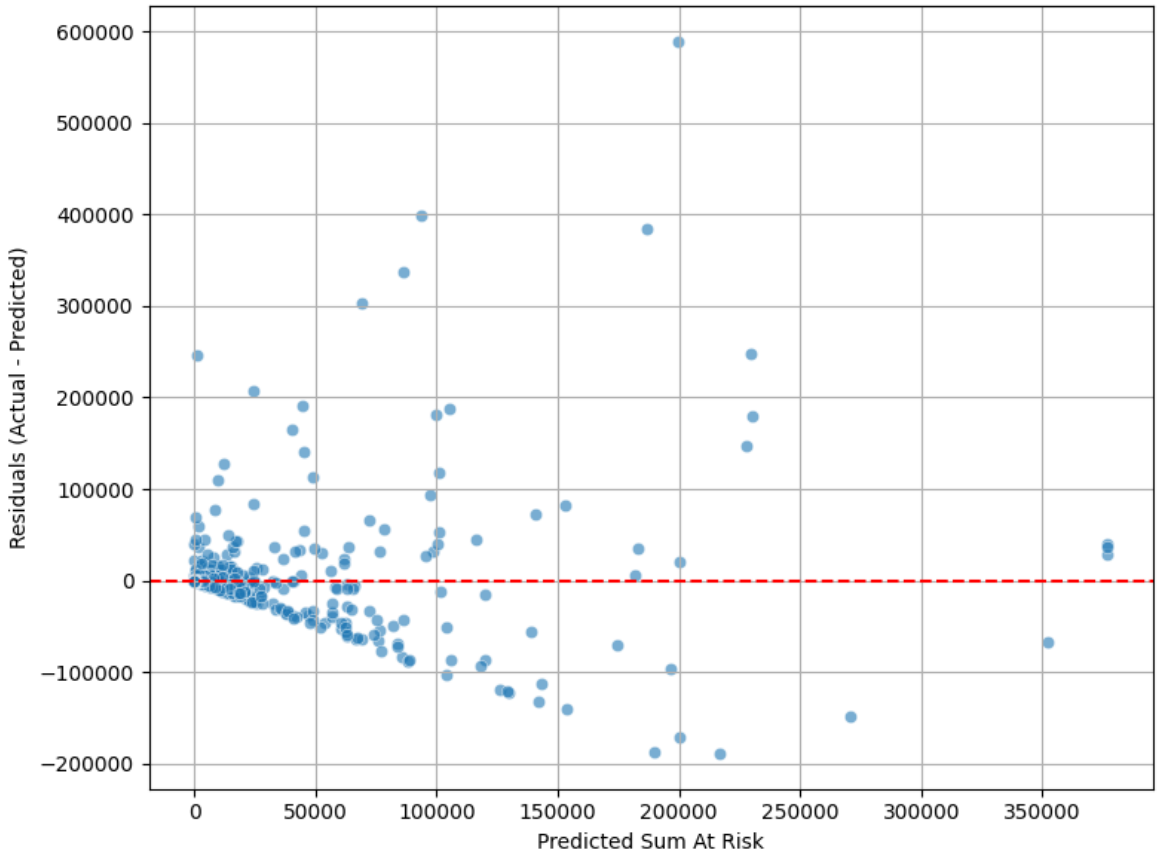
Hình 7.6: Biểu đồ Actual vs Predicted Sum At Risk (Random Forest Regressor)

- **Phân bố dữ liệu bị lệch:**
  - Hầu hết các điểm dữ liệu đều tập trung ở góc dưới bên trái của biểu đồ. Điều này cho thấy phần lớn các đợt dịch có giá trị `Sum At Risk` tương đối nhỏ.
  - Có một vài điểm dữ liệu nằm rất xa về phía bên phải, đây là các giá trị ngoại lệ (outliers) với `Sum At Risk` cực kỳ lớn.
- **Hiệu suất của mô hình:**
  - **Với các giá trị nhỏ:** Nhiều điểm nằm khá gần đường chéo màu đỏ, cho thấy mô hình dự đoán tương đối tốt khi `Sum At Risk` thấp.

– **Với các giá trị lớn:** Các điểm nằm rất xa đường chéo.

- **So sánh với đường chéo lý tưởng:** Sự phân tán rộng của các điểm so với đường chéo màu đỏ khẳng định rằng mô hình còn nhiều sai sót. Nếu mô hình tốt, các điểm sẽ tụ lại thành một dải hẹp xung quanh đường chéo.

7.3.3 Trực quan hoá biểu đồ Residual Plot



Hình 7.7: Biểu đồ Residual Plot Sum At Risk (Random Forest Regressor)

- **Phân bố không đồng đều (Heteroscedasticity):**
  Mô hình dự đoán tốt hơn cho các giá trị Sum At Risk nhỏ, nhưng lại kém chính xác hơn nhiều khi dự đoán các giá trị lớn. Sai số của mô hình trở nên lớn và khó đoán hơn khi giá trị dự đoán tăng.

- **Sự thiên vị (Bias) của mô hình:**
  Phần lớn các điểm có sai số âm (nằm dưới đường $y = 0$), đặc biệt là các

điểm có sai số rất lớn (ví dụ: -2,000,000). Sai số âm (Actual - Predicted < 0) có nghĩa là Predicted $>$ Actual.

Điều này cho thấy mô hình có xu hướng dự đoán cao hơn so với giá trị thực tế (overestimate), đặc biệt là đối với các trường hợp có Sum At Risk lớn.

- **Sự tồn tại của các Outliers:**

  Có một số điểm nằm rất xa so với đường $y = 0$ (cả trên và dưới). Đây chính là những trường hợp mà mô hình đã dự đoán sai lệch rất nhiều, và chính những điểm này đã góp phần làm cho chỉ số MSE (Mean Squared Error) trở nên rất cao như đã thấy ở phần đánh giá trước.

7.4 Mô hình Gradient Boosting Regressor dự đoán Sum Deaths

Gradient Boosting Regressor là một loại thuật toán Machine Learning có giám sát dựa trên học tập tổng hợp . Nó bao gồm một loạt các mô hình tuần tự, mỗi mô hình cố gắng cải thiện lỗi của mô hình trước đó. Nó có thể được sử dụng cho cả nhiệm vụ hồi quy và phân loại. [10]

7.4.1 Đánh giá mô hình Gradient Boosting Regressor

```
Gradient Boosting Regressor - Mean Squared Error (MSE) for Sum Deaths: 680923.9874835924
Gradient Boosting Regressor - R-squared (R2) for Sum Deaths: 0.5907685522556929

Prediction Results for Sum Deaths (first 5 rows):
```

	Actual	Predicted
7665	0.0	-2.209212
3902	1.0	-2.209212
744	1.0	-2.209212
6438	1.0	-2.209212
9187	0.0	-2.209212

Hình 7.8: Kết quả dự đoán Sum Deaths (Gradient Boosting Regressor)

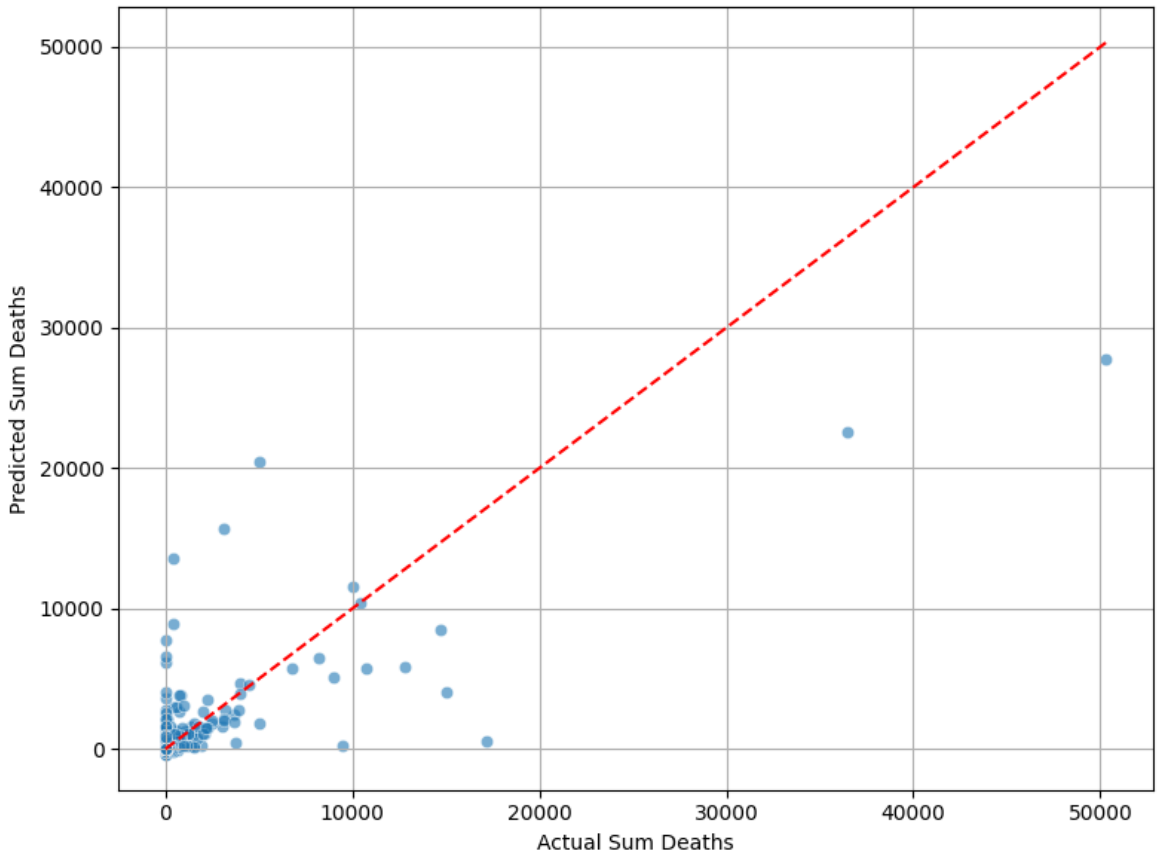
Kết quả từ mô hình Gradient Boosting Regressor dự đoán Sum Deaths dựa vào các đặc trưng 'Sum Cases', 'Sum At Risk', 'report late', 'Latitude', 'Longitude', 'Disease', 'Region':

- **Về chỉ số R-squared (R^2):**
  $R^2 = 0.591$ có nghĩa là mô hình Gradient Boosting có thể giải thích được khoảng 59.1% sự biến động của *Sum Deaths* (số ca tử vong). Đây là kết quả tốt nhất trong số các mô hình hồi quy đã thử nghiệm cho đến nay (cao hơn cả Decision Tree và Random Forest).

- **Về chỉ số Mean Squared Error (MSE):**
  $MSE = 680{,}923.99$ vẫn là một con số rất lớn. Tương tự như các mô hình trước, điều này cho thấy mô hình vẫn gặp khó khăn trong việc dự đoán chính xác các trường hợp có số ca tử vong rất cao (outliers), dẫn đến sai số bình phương lớn.

- **Phân tích các dự đoán cụ thể (5 dòng đầu):**
  Mô hình đang dự đoán ra các giá trị âm (-2.209212) trong khi số ca tử vong (*Sum Deaths*) không thể là số âm.

7.4.2 Trực quan hoá biểu đồ Actual vs Predicted



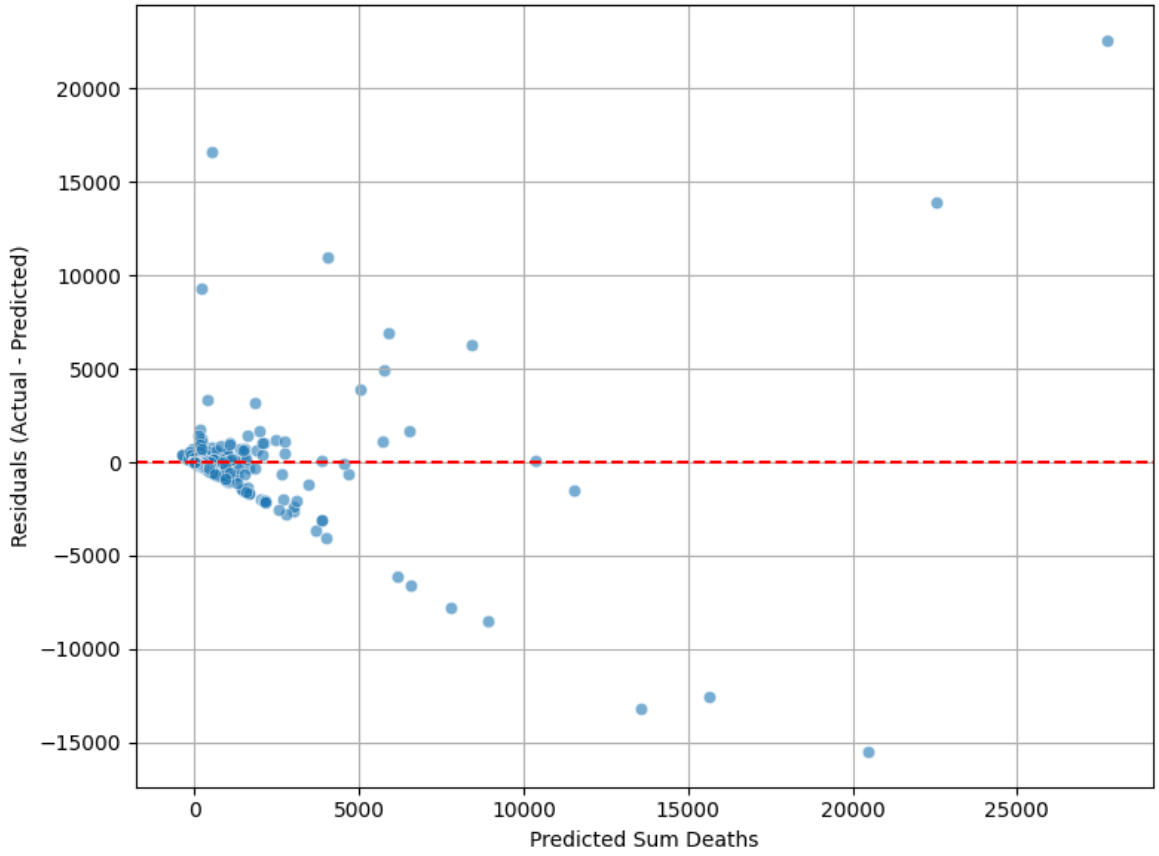
Hình 7.9: Biểu đồ Actual vs Predicted Sum Deaths (Gradient Boosting Regressor)

- **Dự đoán giá trị âm:** Vấn đề lớn nhất là có một dải các điểm dự đoán nằm ở vùng giá trị âm. Mô hình đang đưa ra những dự đoán vô lý trong thực tế, vì số ca tử vong không thể là số âm
- **Phân bố dữ liệu bị lệch (Skewed Distribution):** Hầu hết các điểm dữ liệu đều tập trung ở góc dưới bên trái, cho thấy phần lớn các báo cáo có số ca tử vong thấp.
- **Hiệu suất của mô hình:**
  - Với các giá trị nhỏ (gần 0): Mô hình hoạt động không tốt. Nó không chỉ dự đoán sai mà còn đưa ra các giá trị âm.

– Với các giá trị lớn: Các điểm nằm rất xa đường chéo màu đỏ. Mô hình dường như không thể nắm bắt được các đợt dịch có số ca tử vong cao.

– Nó có xu hướng dự đoán thấp hơn nhiều so với giá trị thực tế (underestimate) cho các điểm này.

7.4.3 Trực quan hoá biểu đồ Residual Plot



Hình 7.10: Biểu đồ Residual Plot Sum Deaths (Gradient Boosting Regressor)

- **Mô hình có cấu trúc rõ ràng (Clear Pattern):**

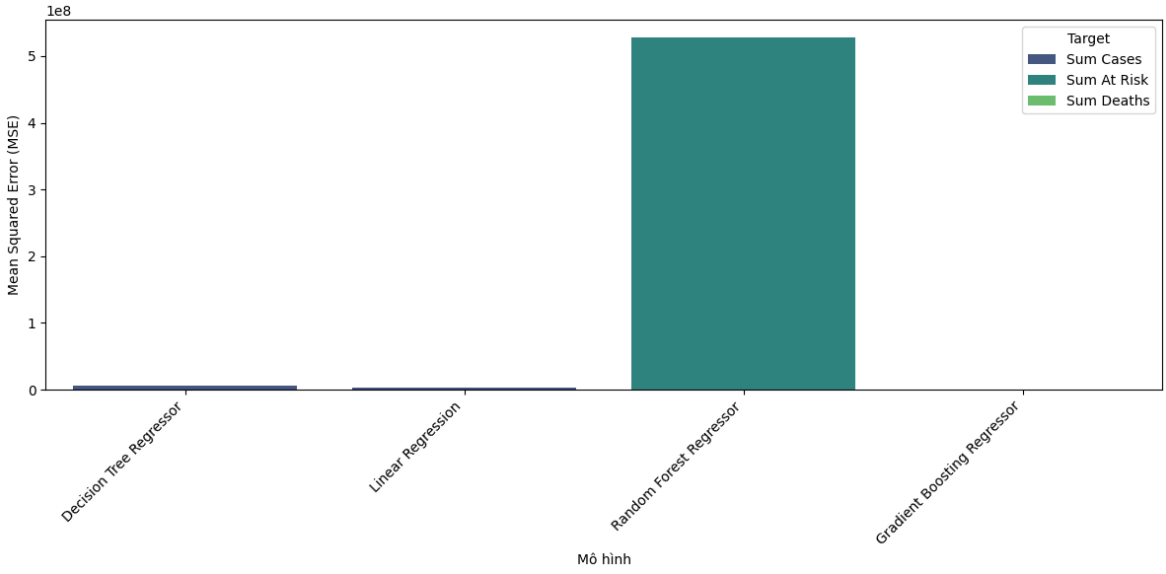
  – Thay vì phân bố ngẫu nhiên xung quanh đường $y = 0$, các điểm sai số tạo thành một cấu trúc rất rõ ràng, giống như một đường cong hoặc một đám mây điểm có xu hướng đi lên.

- Điều này cho thấy mô hình chưa nắm bắt được hết các mối quan hệ giữa các biến đầu vào và đầu ra.

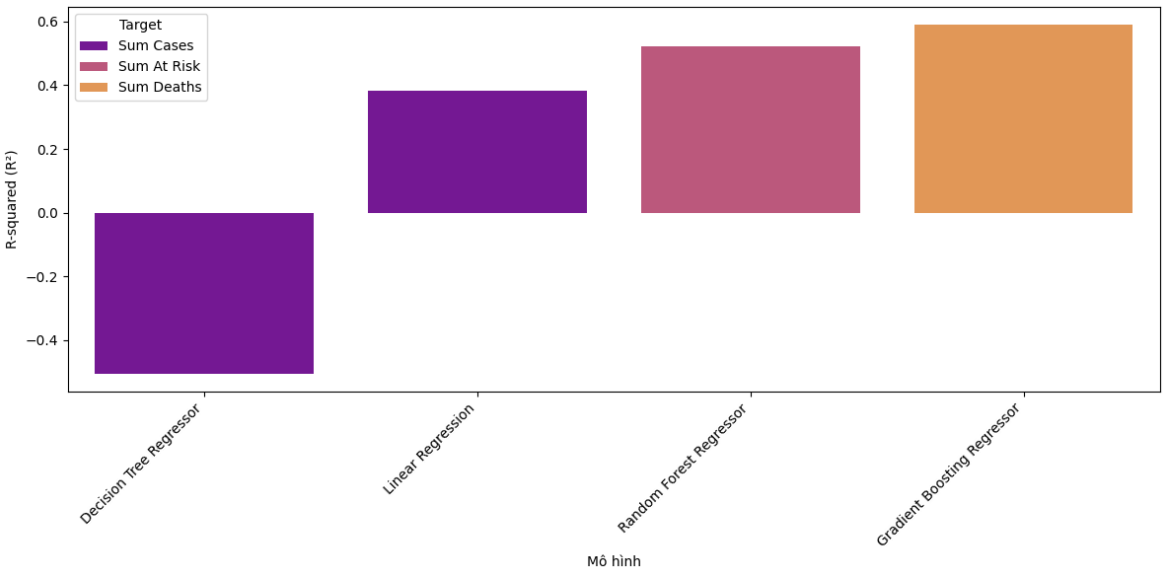
- **Vấn đề với dự đoán âm:** Toàn bộ các điểm dự đoán âm đều có sai số dương.

- **Phương sai không đồng đều (Heteroscedasticity):**
  - Tương tự như mô hình Random Forest, biểu đồ này cũng cho thấy dạng "cái phễu".
  - Khi giá trị dự đoán tăng, sự biến động của sai số cũng tăng lên.
  - Điều này cho thấy mô hình kém tin cậy hơn khi dự đoán các giá trị `Sum Deaths` lớn.

7.5 Kết luận chung REGRESSION



Hình 7.11: So sánh Mean Squared Error (MSE) của các mô hình



Hình 7.12: So sánh R-squared (R^2) của các mô hình

- **Hồi quy tuyến tính (Linear Regression)**: Quá đơn giản, không nắm bắt được mối quan hệ phức tạp, R^2 thấp.
- **Decision Tree Regressor**: Dễ bị quá khớp (overfitting), hiệu suất không cao.
- **Random Forest Regressor**: Tốt hơn đáng kể so với hai mô hình trên ($R^2 \approx 0.52$), nhưng vẫn gặp khó khăn với outliers và dữ liệu lệch.
- **Gradient Boosting Regressor**: Đạt R^2 cao nhất (0.59), cho thấy tiềm năng lớn nhất. Tuy nhiên, nó lại bộc lộ những vấn đề nghiêm trọng nhất khi áp dụng trực tiếp: dự đoán giá trị âm và có các mẫu rõ ràng trong biểu đồ sai số.

Không có mô hình nào trong số này là hoàn hảo cho bộ dữ liệu. Dữ liệu về dịch bệnh có đặc tính rất phức tạp (lệch, nhiều outliers, không âm).

Chương 8

SO SÁNH BỘ DỮ LIỆU COVID-19 DATASET

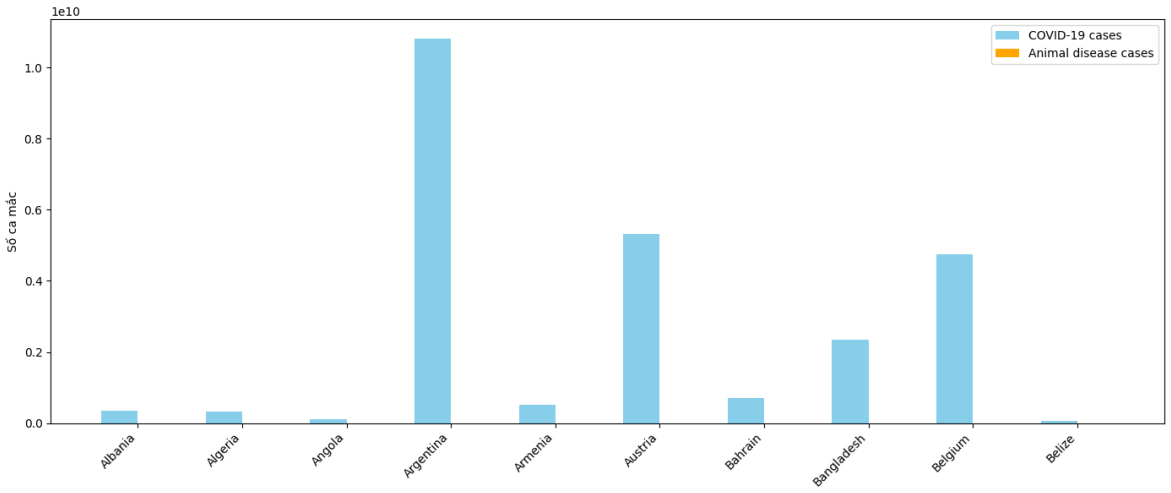
- Mục đích: Việc thực hiện nghiên cứu này nhằm mục đích xem mức độ ảnh hưởng của đại dịch Covid 19 đến con người và các bệnh truyền nhiễm ở động vật. Việc phân tích giúp hiểu hơn về mối tương quan tiềm tàng, mức độ nghiêm trọng và khả năng đe dọa đến sức khỏe cộng đồng, kinh tế - xã hội.

- Nguồn dữ liệu: Đối với Covid-19 sử dụng dữ liệu từ nguồn our world in data còn đối với EMPRES sử dụng dữ liệu từ EMPRES Global Animal Disease Surveillance trên Kaggle.

- Các thuộc sử dụng để thực hiện việc này là Country (Quốc gia), Total cases (Số trường hợp mắc bệnh), Total deaths (Số trường hợp chết)

- Các bước thực hiện: Thứ nhất tiền sử lý dữ liệu đối với hai tập dữ liệu trên. Tiếp thoe là lấy ở một tập dữ liệu gồm 3 cột country, total cases, total deaths để xử lý. Sau đó gộp 2 tập này thành 1 tập duy nhất. Sau đó trực quan hóa lên biểu đồ để thực hiện việc so sánh.

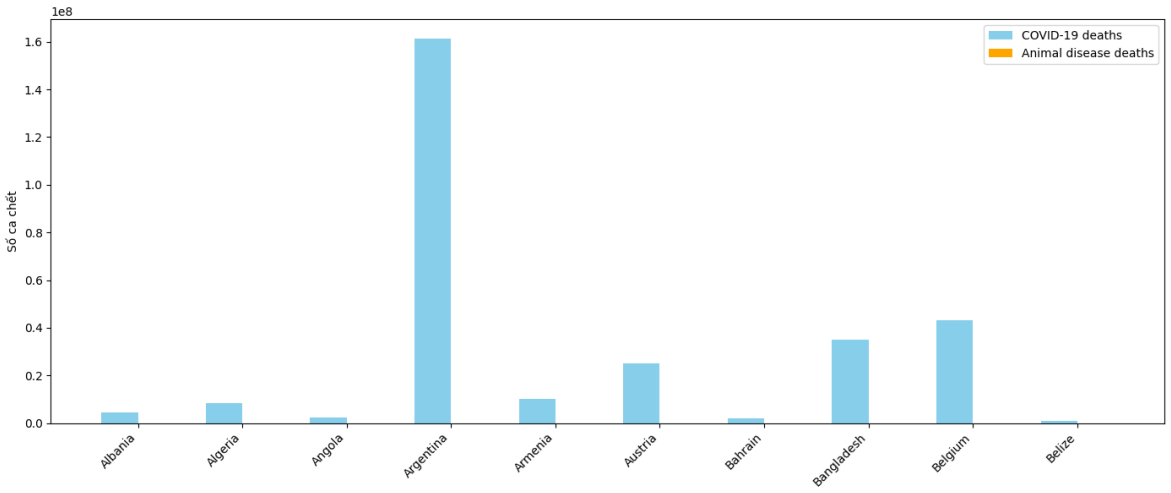
8.1 So sánh số ca nhiễm



Hình 8.1: Biểu đồ so sánh số ca mắc

- Theo kết quả phân tích thì ta có tất 10 nước đầu tiền được đem ra so sánh với nhau để làm dẫn chứng.
- Quan sát biểu đồ ta có thể dễ dàng nhận thấy số ca mắc covid-19 gấp rất nhiều lần so với các loại bệnh động vật trong 4 năm cho thấy mức độ lan nhanh chóng của bệnh covid-19.

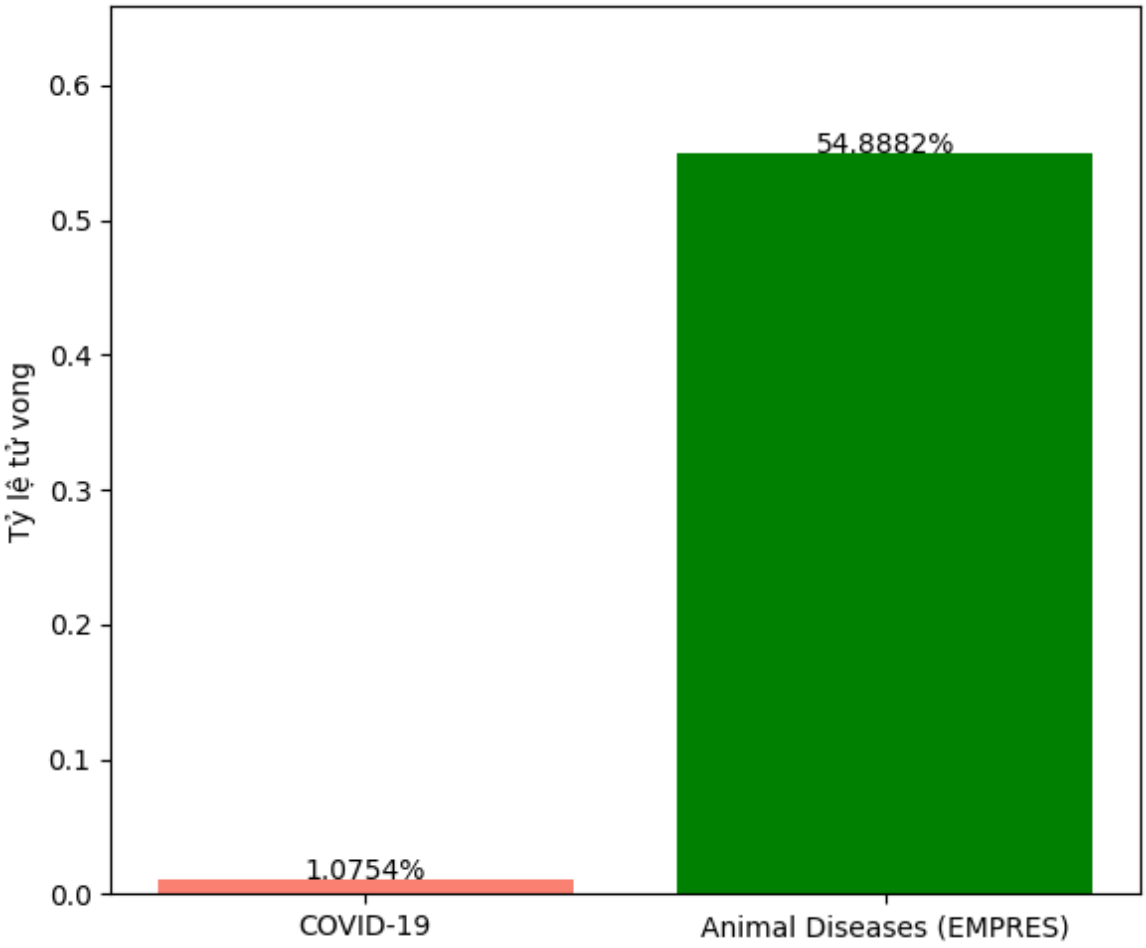
8.2 So sánh số ca chết



Hình 8.2: Biểu đồ so sánh số ca chết

- Theo kết quả phân tích thì ta có tất 10 nước đầu tiền được đem ra so sánh với nhau để làm dẫn chứng.
- Quan sát biểu đồ ta có thể dễ dàng nhận thấy số ca mất do covid-19 nhiều hơn rất nhiều lần so với các ca chết do các bệnh động vật từ đó thấy được mức độ nguy hiểm của đại dịch covid.

8.3 So sánh tỷ lệ tử vong



Hình 8.3: Biểu đồ so sánh tỷ lệ tử vong

- Theo kết quả phân tích thì ta có dễ dàng quan sát tỷ lệ từ vong của covid là 1.0754 và của empres là 54.8882.
- Qua đó cho ta thấy tuy số lượng ca mắc và số lượng tử vong của các dịch bệnh empres thấp hơn rất nhiều so với covid-19 nhưng tỷ lệ từ vong khi mắc phỉa các dịch empres là lớn hơn 50 lần so với covid-19 qua đó ta thấy được sự nguy hiểm của các dịch empres.

Chương 9

KẾT LUẬN

Bài phân tích đã cung cấp một cái nhìn tổng quan về tình hình dịch bệnh động vật dựa trên bộ dữ liệu EMPRES, tập trung vào giai đoạn 2017–2019. Bài báo cáo đã xác định được các loại bệnh và loài bị ảnh hưởng phổ biến nhất, trực quan hóa sự phân bố địa lý của các ổ dịch, và áp dụng các kỹ thuật khai phá dữ liệu để tìm ra các mẫu hình tiềm ẩn và dự đoán số lượng động vật bị nhiễm bệnh hoặc tử vong dựa trên các yếu tố liên quan.

- **Clustering** giúp nhận diện các "điểm nóng" không gian–thời gian của dịch bệnh.
- **Pattern Mining** phát hiện các mối liên hệ thường xuyên giữa bệnh, loài, địa lý và các chuỗi lây lan theo thời gian.
- Các mô hình **Regression** ban đầu (như hồi quy tuyến tính) cho thấy khả năng dự đoán số ca/số chết còn hạn chế. Tuy nhiên, các mô hình phi tuyến như *Random Forest* và *Gradient Boosting* cho kết quả khả quan hơn, đặc biệt trong việc dự đoán `Sum At Risk` và `Sum Deaths`.

Dữ liệu có độ lệch lớn và sự hiện diện của các giá trị ngoại lai (*outliers*) là thách thức không nhỏ đối với các mô hình dự đoán. Dù vậy, phân tích này đã thành công trong việc khám phá cấu trúc và các mẫu hình tiềm ẩn trong dữ liệu dịch bệnh, mang lại những hiểu biết ban đầu về sự lây lan và các yếu tố liên quan.

TÀI LIỆU THAM KHẢO

[1] Ultralytics. Dbscan. https://www.ultralytics.com/vi/glossary/dbscan-density-based-spatial-clustering-of-applications-with-noise, 2025. Accessed July 26, 2025.

[2] Chioka. Meanshift algorithm. https://www.chioka.in/meanshift-algorithm-for-the-rest-of-us-python, 2025. Accessed July 26, 2025.

[3] GeeksforGeeks. Hdbscan in machine learning. https://www.geeksforgeeks.org/machine-learning/hdbscan/, 2025. Accessed July 26, 2025.

[4] geeksforgeeks. Frequent itemset mining. https://www.geeksforgeeks.org/data-science/frequent-item-set-in-data-set-association-rule-mining/, 2023. Truy cập ngày 26 tháng 7 năm 2025.

[5] Yash Sanghvi. Sequence pattern mining. https://hevodata.com/learn/sequence-pattern-mining/, 2024. Truy cập ngày 26 tháng 7 năm 2025.

[6] Grammarly. Regression in machine learning. https://www.grammarly.com/blog/ai/what-is-regression/, 2024. Truy cập ngày 26/07/2025.

[7] Nguyễn Thị Hợp. Decision tree. https://viblo.asia/p/decision-tree-Do754bbBZM6, 2020. Truy cập ngày 26 tháng 7, 2025.

[8] AI Candy. Hồi quy tuyến tính. https://aicandy.vn/hoi-quy-tuyen-tinh-linear-regression-tim-hieu-chi-tiet/, 2023. Truy cập ngày 26/07/2025.

[9] AI Candy. Thuật toán random forest, 2024. Truy cập ngày 26 tháng 7, 2025.

[10] Bùi Tiến Tùng. Gradient boosting, 2021. Truy cập ngày 26 tháng 7, 2025.