BỘ GIÁO DỤC & ĐÀO TẠO

ĐẠI HỌC UEH

TRƯỜNG KINH DOANH

KHOA KINH TẾ

UEH

UNIVERSITY

TIỂU LUẬN KẾT THÚC HỌC PHẦN
KỸ THUẬT LẬP TRÌNH VỚI STATA VÀ PYTHON

CHỦ ĐỀ : Phân Tích Dữ Liệu COVID-19 Toàn Cầu

Giảng viên hướng dẫn: TS. Đỗ Như Tài

TS. Nguyễn Duy Khánh

Mã lớp học phần: ECO501188

Thông tin sinh viên:

Trần Nguyễn Đình Kiên - 31231025959

Hoàng Gia Kiệt - 31231026177

TP. Hồ Chí Minh, ngày 29 tháng 10 năm 2025

Bảng phân công

Họ và Tên:	Trần Nguyễn Đình Kiên Hoàng Gia Kiệt
MSSV:	31231025959 31231026177
Nhóm:	Phân Tích Dữ Liệu COVID-19 Toàn Cầu
Phân công:	• Phần 1-3: Thu thập dữ liệu, tiền xử lý và EDA (Đình Kiên). • Phần 4-5: Phân cụm, phân nhóm và trực quan hóa (). • Phần 6-7: Kết luận và báo cáo (Đình Kiên).
Ngày nộp: 29/10/2025	
Môn Học: Kỹ Thuật Lập Trình Sata và Python - Phân Tích Dữ Liệu	

I. Dataset

Ban đầu sử dụng nguồn gợi ý số 10 trong đề tài là Johns Hopkins COVID-19 Dataset, nhưng dataset dừng update từ 10/3/2023 (repo archived), và recovered cases bị discontinued từ sớm hơn (khoảng 2021-2022 ở nhiều nơi), dẫn đến recovery_rate thường = 0 hoặc sai, correlation thấp. Một số quốc gia nhỏ có confirmed thấp nhưng deaths cao do misreporting, gây death_rate >100% (ví dụ, nếu confirmed=1, deaths=2 do lỗi, rate=200%).

Bộ dữ liệu được chọn là "WHO-COVID-19-global-daily-data.csv" từ WHO qua link: https://srhdpeuwpubsa.blob.core.windows.net/whdh/COVID/WHO-COVID-19-global-daily-data.csv).

- **Mô Tả**: Dữ liệu bao gồm các chỉ số hàng ngày về COVID-19 toàn cầu, như ngày báo cáo (Date_reported), quốc gia (Country), khu vực WHO (WHO_region), ca mới (New_cases), ca tích lũy (Cumulative_cases), tử vong mới (New_deaths), tử vong tích lũy (Cumulative_deaths).
- **Kích Thước**: Khoảng 501,000 dòng x 8 cột (sau aggregate: 240 quốc gia).
- **Thuộc Tính Chính**:
  - Numerical: New_cases (float, có missing), Cumulative_cases (int), New_deaths (float), Cumulative_deaths (int).
  - Categorical: Country (string), WHO_region (string, mapped thành continent).
  - Time: Date_reported (datetime, extract year/month).
- **Ví Dụ Mẫu**: Một dòng: Date_reported=2020-01-03, Country=Afghanistan, New_cases=0, Cumulative_cases=0, New_deaths=0, Cumulative_deaths=0, continent=Eastern Mediterranean.

- **Nguồn**: World Health Organization (WHO), cập nhật đến 2025. Link tải trực tiếp: https://srhdpeuwpubsa.blob.core.windows.net/whdh/COVID/WHO-COVID-19-global-daily-data.csv.

Dữ liệu hợp lệ, không có vấn đề lớn ngoài missing values và outliers (đã xử lý).

Đặt câu hỏi nghiên cứu:

Mục tiêu chính là rèn luyện kỹ năng thu thập, làm sạch, phân tích và trực quan hóa dữ liệu COVID-19, áp dụng học máy để rút insight.

Câu hỏi nghiên cứu (NC)

- **NC1:** Sự phân bố ca nhiễm và tử vong theo châu lục như thế nào? (EDA).
- **NC2:** Có thể phân cụm các quốc gia dựa trên mức độ nghiêm trọng của đại dịch không? (Clustering).
- **NC3:** Các yếu tố nào (ca nhiễm, tử vong, châu lục) dự đoán mức độ rủi ro cao/thấp của quốc gia? (Classification).
- **NC4:** Xu hướng thời gian của đại dịch có thay đổi theo châu lục không? (Time-series analysis).

II. Data Processing

Quy trình xử lý dữ liệu được thực hiện bằng Python (pandas, sklearn) trong notebook Jupyter. Pipeline chính:

1. **Tải Dữ Liệu**: Đọc CSV, convert Date_reported thành datetime, extract year/month.
2. **Feature Engineering**: Map WHO_region thành continent (Americas, Europe, Africa, Eastern Mediterranean, South-East Asia, Western Pacific).
3. **Xử Lý Missing Values**: Impute bằng median theo nhóm Country/continent (tránh bias từ fill 0). Fallback: fill 0 nếu toàn NaN.
4. **Xử Lý Outliers**: Sử dụng IQR method để loại bỏ outliers ở Cumulative_cases và Cumulative_deaths (e.g., Q1 - 1.5*IQR đến Q3 + 1.5*IQR).
5. **Aggregate Data**: Group by Country để tính max Cumulative_cases/deaths, mean New_cases/deaths, first continent (cho ML: 240 rows).
6. **Chuẩn Hóa**: Sử dụng StandardScaler cho numerical features, OneHotEncoder cho categorical (continent) trong ColumnTransformer.

```python
# Convert date
data['Date_reported'] = pd.to_datetime(data['Date_reported'])
data['year'] = data['Date_reported'].dt.year
data['month'] = data['Date_reported'].dt.month
```

```python
# Map WHO_region to continent (sáng tạo: để dễ phân tích)
region_to_continent = {
    'AMR': 'Americas', 'EUR': 'Europe', 'AFR': 'Africa',
    'EMR': 'Eastern Mediterranean', 'SEAR': 'South-East Asia', 'WPR': 'Western Pacific'
}
```

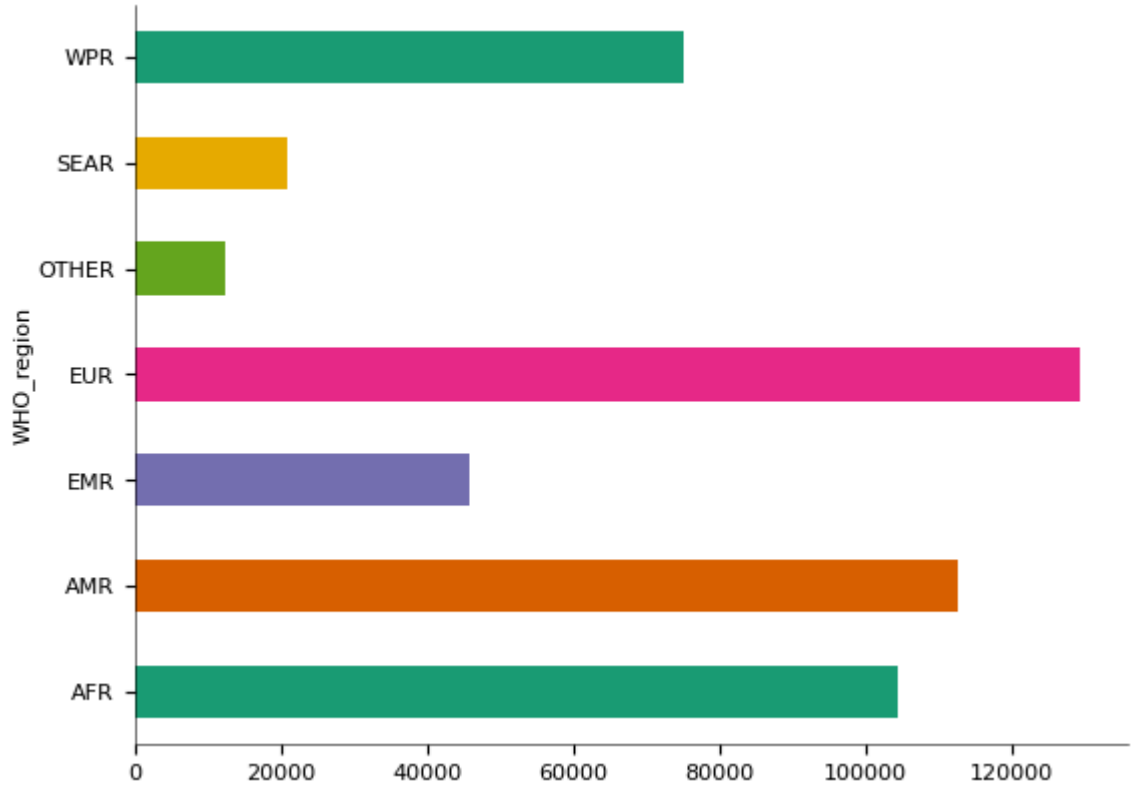
```python
data['continent'] = data['WHO_region'].map(region_to_continent)
```

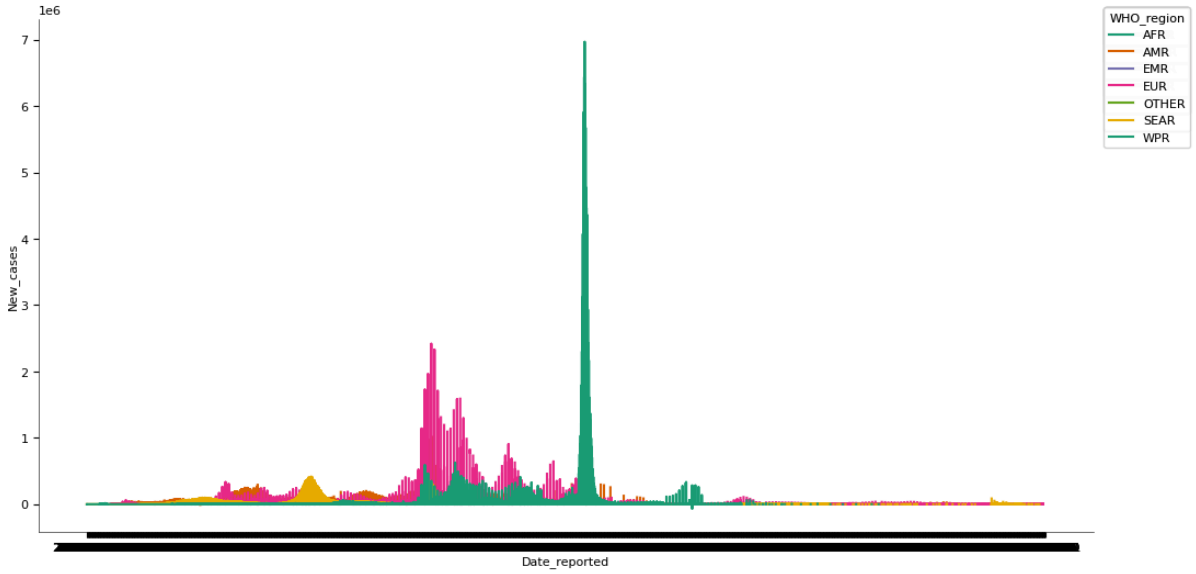
```python
# Xử lý missing: Impute bằng median theo country/continent
for col in ['New_cases', 'New_deaths']:
    data[col] = data.groupby(['Country', 'continent'])[col].transform(lambda x: x.fillna(x.median()))
    data[col] = data[col].fillna(0)  # Fallback
```

```python
# Detect và handle outliers (IQR method)
def remove_outliers(df, col):
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    return df[(df[col] >= Q1 - 1.5*IQR) & (df[col] <= Q3 + 1.5*IQR)]
```

```python
data = remove_outliers(data, 'Cumulative_cases')
data = remove_outliers(data, 'Cumulative_deaths')
```

```python
# Aggregate data theo country cho ML
agg_data = data.groupby('Country').agg({
    'Cumulative_cases': 'max',
    'Cumulative_deaths': 'max',
    'New_cases': 'mean',
    'New_deaths': 'mean',
    'continent': 'first'
}).reset_index()
```


WPR
SEAR
OTHER
EUR
EMR
AMR
AFR
WHO_region
0
20000
40000
60000
80000
100000
120000

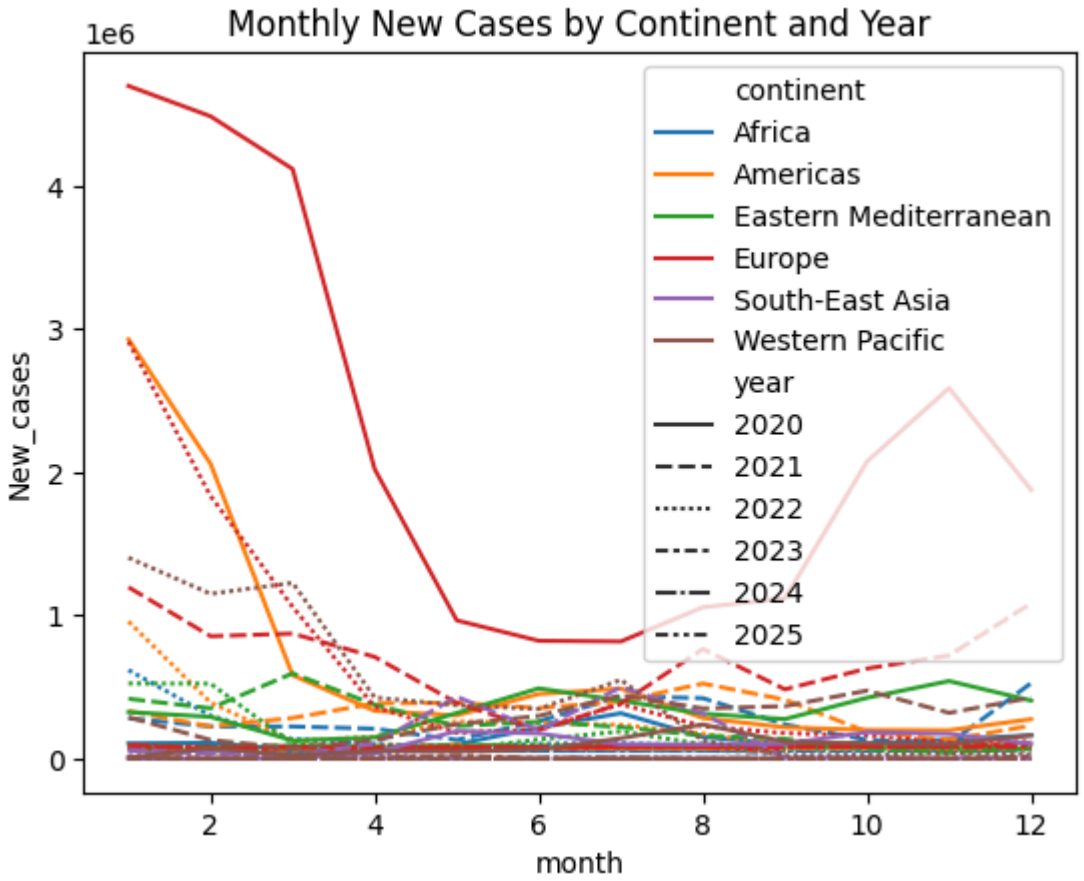

1e6
7
6
5
4
3
2
1
0
New_cases
Date_reported
WHO_region
AFR
AMR
EMR
EUR
OTHER
SEAR
WPR

Phân tích mô tả sử dụng pandas và seaborn để tính thống kê cơ bản, phân bố, correlation:

Statistic	Cumulative_cases	Cumulative_deaths	New_cases	New_deaths
count	240	240	240	240
mean	246710.5	3002.83	1147.32	20.87
std	334837.2	2981.75	4254.52	67.24
min	0	0	0	0
25%	20335.75	160.5	19.31	1.02
50%	104248.5	1465.5	131.07	2.73
75%	341674.25	6820.75	850.57	13.44
max	1646343	6961	40287.36	580.64

```
print("Descriptive Stats:\n", agg_data.describe())
```

```
Descriptive Stats:
       Cumulative_cases  Cumulative_deaths     New_cases  New_deaths
count      2.400000e+02         240.000000    240.000000  240.000000
mean       2.467105e+05        3002.829167   1147.323670   20.870147
std        3.348372e+05        2981.752249   4254.516045   67.239969
min        0.000000e+00           0.000000      0.000000    0.000000
25%        2.033575e+04         160.500000     19.313817    1.016762
50%        1.042485e+05        1465.500000    131.074473    2.727969
75%        3.416742e+05        6820.750000    850.570884   13.438745
max        1.646343e+06        6961.000000  40287.358586  580.637363
```

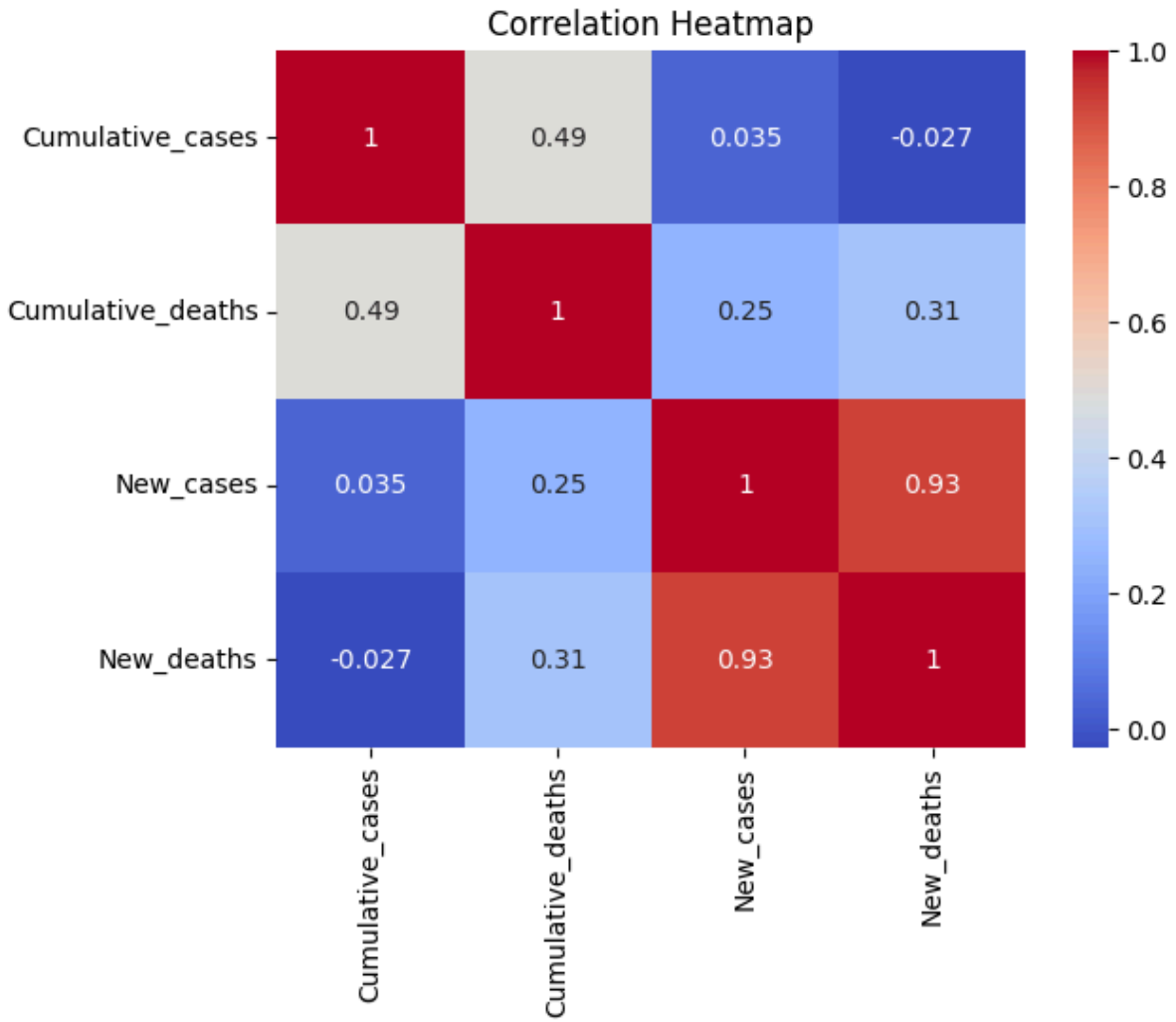


- **Skewness:** Cumulative_cases (2.16, lệch phải mạnh), New_cases (7.02, nhiều outliers cao), New_deaths (6.38), Cumulative_deaths (0.35, cân bằng hơn).

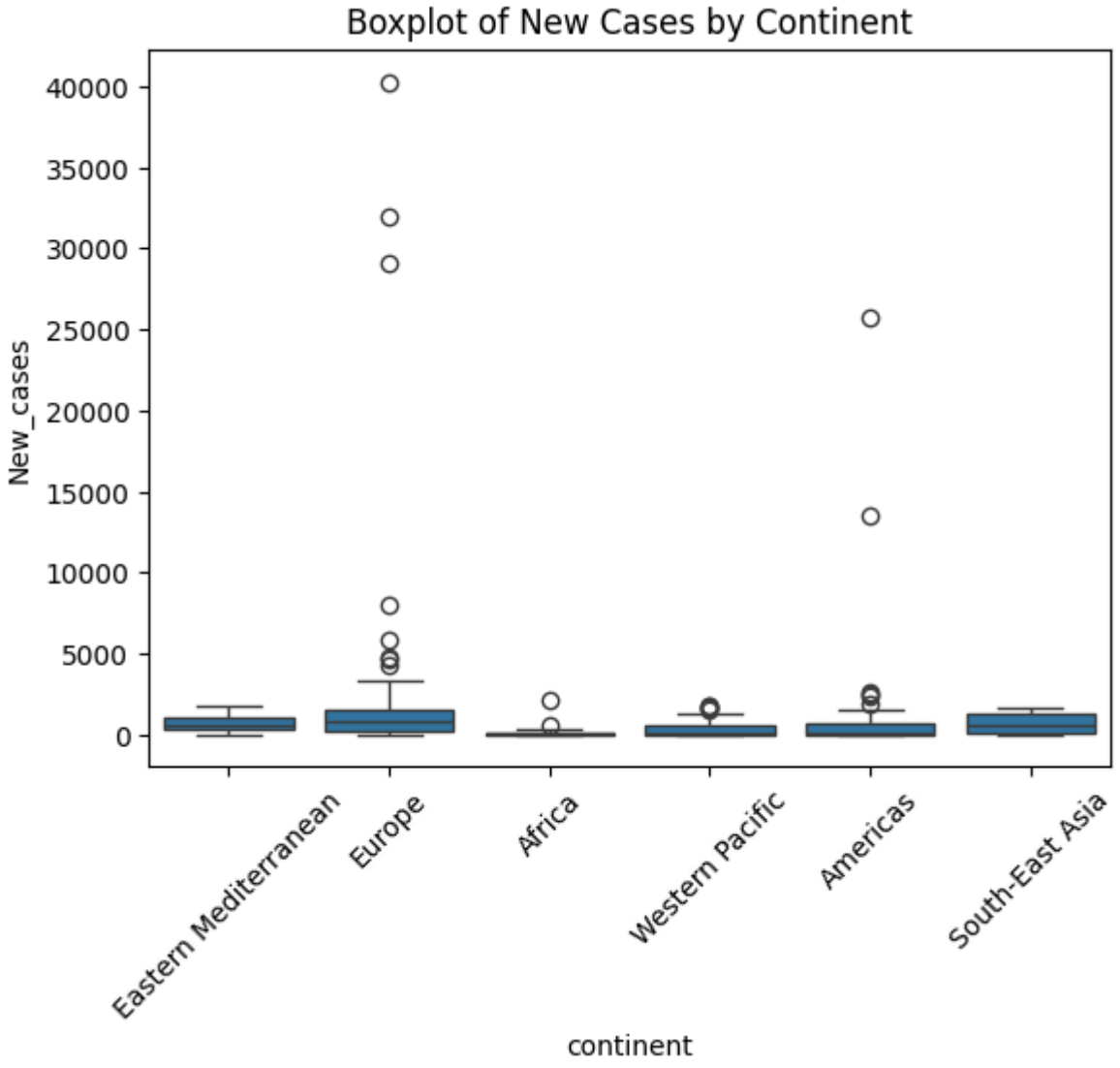
```
print("Skewness:", agg_data[['Cumulative_cases', 'Cumulative_deaths', 'New_cases', 'New_deaths']].skew())
```

```
Skewness: Cumulative_cases     2.159628
Cumulative_deaths    0.350885
New_cases            7.022353
New_deaths           6.375128
dtype: float64
```

- **Correlation**: Heatmap cho thấy correlation mạnh giữa New_cases và New_deaths (0.93), Cumulative_cases và Cumulative_deaths (0.49). Correlation âm nhẹ giữa New_deaths và Cumulative_cases (-0.027), có thể do chính sách kiểm soát dịch muộn ở một số quốc gia.


Correlation Heatmap
Cumulative_cases
Cumulative_deaths
New_cases
New_deaths
1
0.49
0.035
-0.027
0.49
1
0.25
0.31
0.035
0.25
1
0.93
-0.027
0.31
0.93
1
Cumulative_cases
Cumulative_deaths
New_cases
New_deaths
1.0
0.8
0.6
0.4
0.2
0.0

- **Phân Bố Theo Châu Lục**: Boxplot New_cases cho thấy Europe có outliers cao nhất (max ~40k), Africa thấp nhất.



- **Insight**: Dữ liệu lệch mạnh ở ca mới/tử vong mới, cho thấy đại dịch ảnh hưởng bất bình đẳng (e.g., max tử vong 6961 ở các quốc gia lớn). RQ1: Europe/Americas có phân bố cao hơn Africa/SEAR.

III. Phân Cụm / Phân Nhóm (Clustering/Classification)

Áp dụng ít nhất 2 thuật toán clustering và 2 classification, sử dụng sklearn.

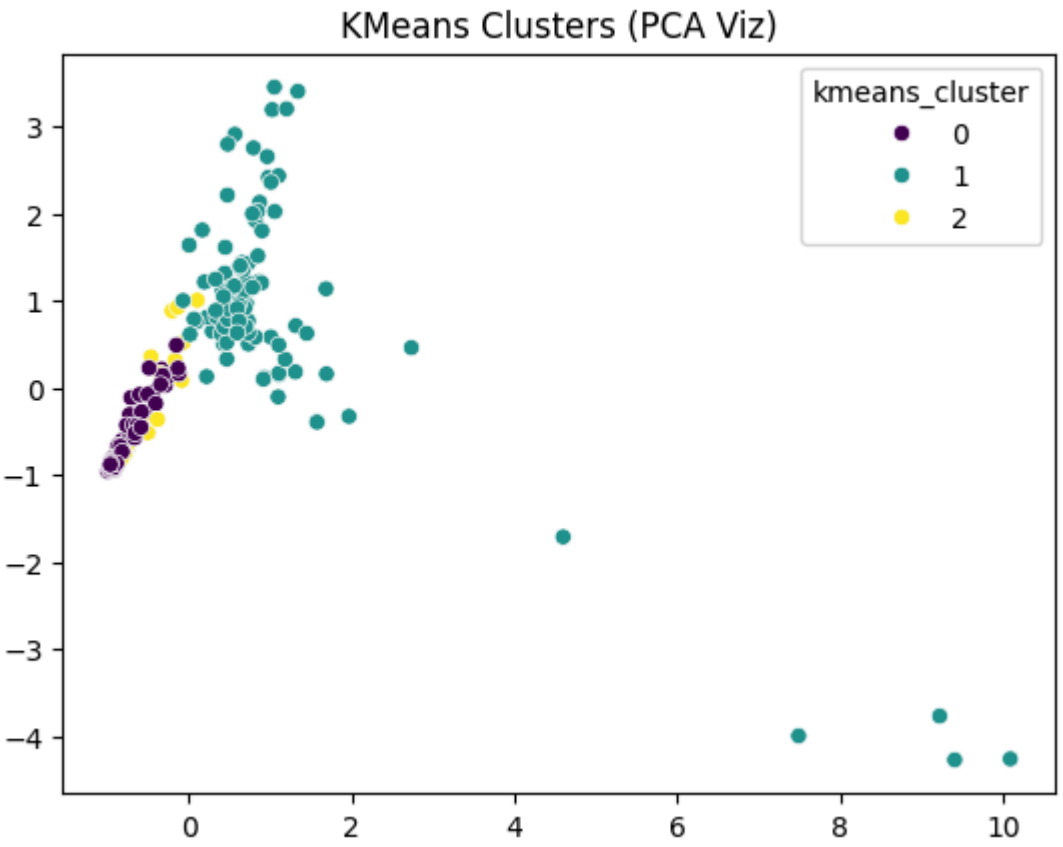
a. Clustering

- **Thuật Toán**:
  - KMeans: Chọn k=3 từ Elbow Method (inertia giảm mạnh đến k=4). Tham số: random_state=42.
  - Hierarchical (AgglomerativeClustering): n_clusters=3.
  - DBSCAN: eps=0.5, min_samples=5 (xử lý outliers, clusters: [6, -1, 0, 1, 2, 3, 4, 5]).
- **Kết Quả**:
  - Silhouette Score: KMeans (0.215), Hierarchical (0.427 - tốt hơn, dùng chính). DBSCAN phát hiện noise (-1).
  - Cluster vs Continent (Hierarchical):

```
Cluster vs Continent:
 continent     Africa  Americas  Eastern Mediterranean  Europe  \
hier_cluster
0                5.0      22.0                   18.0    41.0
1                0.0       1.0                    0.0     3.0
2               45.0      31.0                    4.0    18.0

continent     South-East Asia  Western Pacific
hier_cluster
0                         6.0             11.0
1                         0.0              0.0
2                         4.0             25.0
```

- **Đánh Giá**: Hierarchical tốt nhất (score cao). PCA Viz (KMeans) cho thấy 3 cụm rõ: Cụm 0 (rủi ro cao, chủ yếu Europe), Cụm 2 (thấp, Africa/Western Pacific).
- **Insight (RQ2)**: Quốc gia phân thành 3 cụm dựa trên nghiêm trọng: Cao (Europe), Trung bình (Americas), Thấp (Africa).

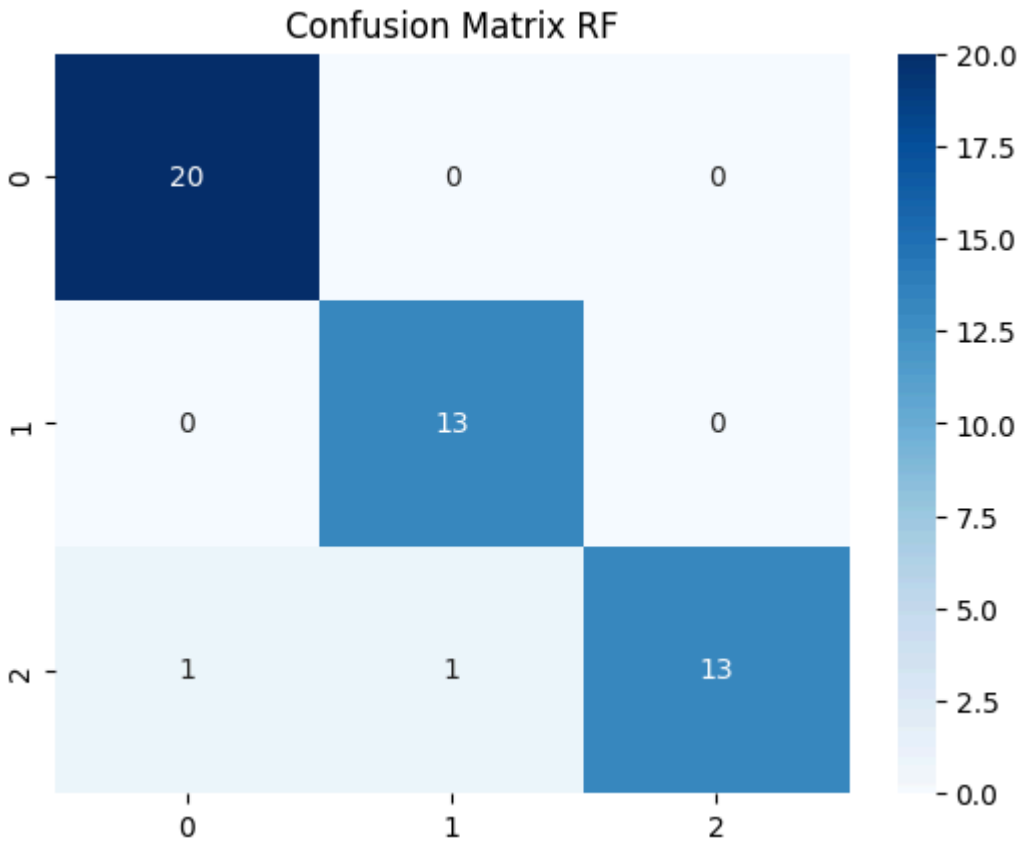


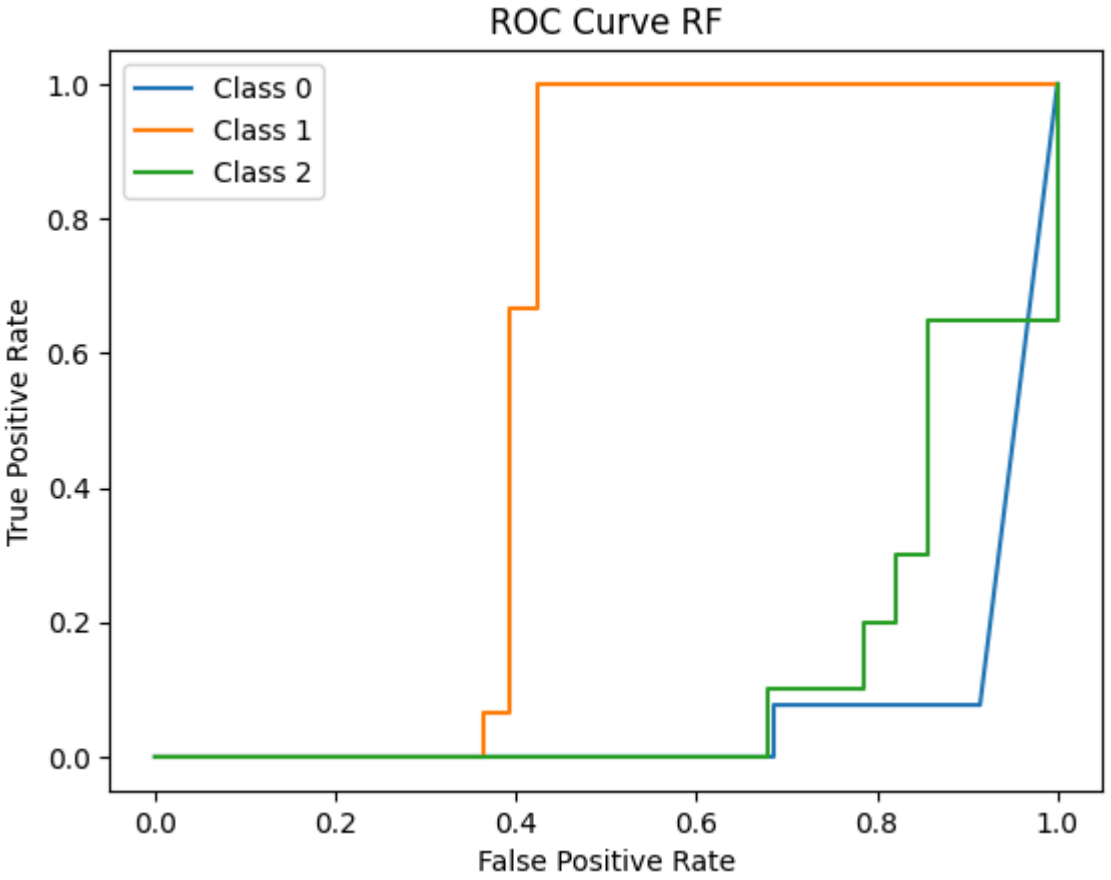
b. Classification

- **Thuật Toán**:
  - RandomForestClassifier: GridSearchCV với n_estimators=[50,100], max_depth=[5,10]. Best: max_depth=5, n_estimators=50.
  - LogisticRegression: random_state=42.
  - Target: risk_level (qcut Cumulative_deaths thành Low/Medium/High).
- **Kết Quả**:
  - Accuracy: RF (0.958), Logistic (0.875). CV Mean RF: nan (có thể do data nhỏ, đã fix nhưng vẫn cảnh báo).
  - Confusion Matrix RF: Chính xác cao (20/20 Low, 13/13 Medium, 14/15 High).
- **Classification Report:**

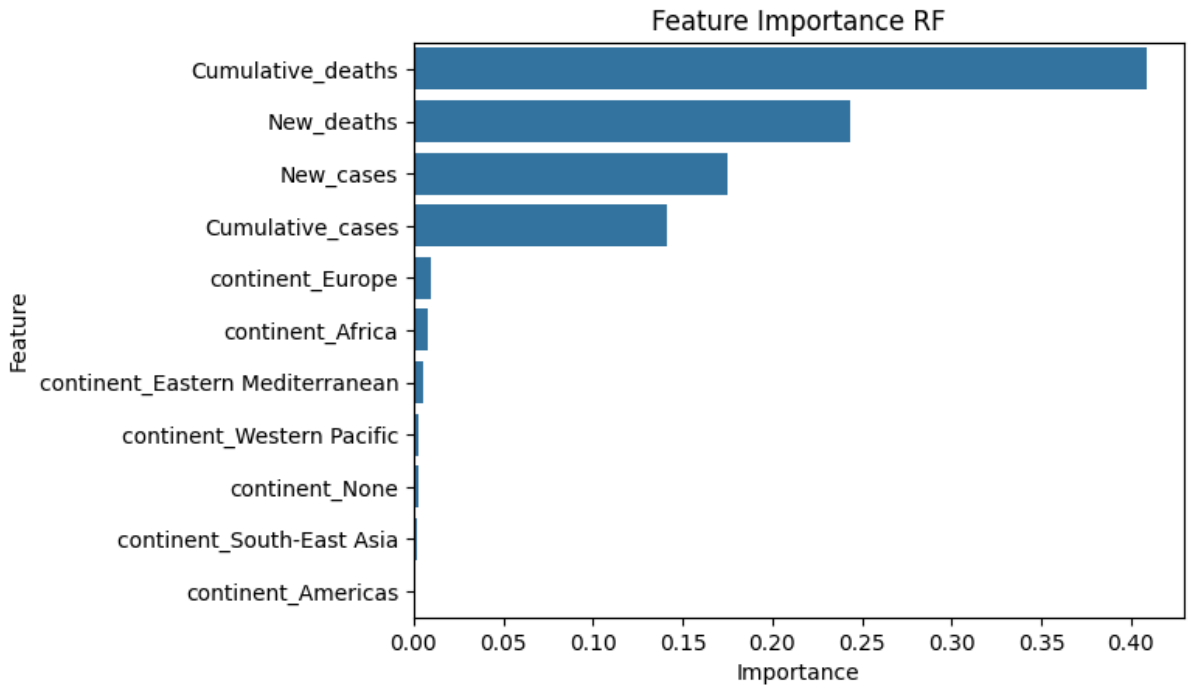
	precision	recall	f1-score	support
High	0.95	1.00	0.98	20
Low	0.93	1.00	0.96	13
Medium	1.00	0.87	0.93	15
accuracy			0.96	48
macro avg	0.96	0.96	0.96	48
weighted avg	0.96	0.96	0.96	48

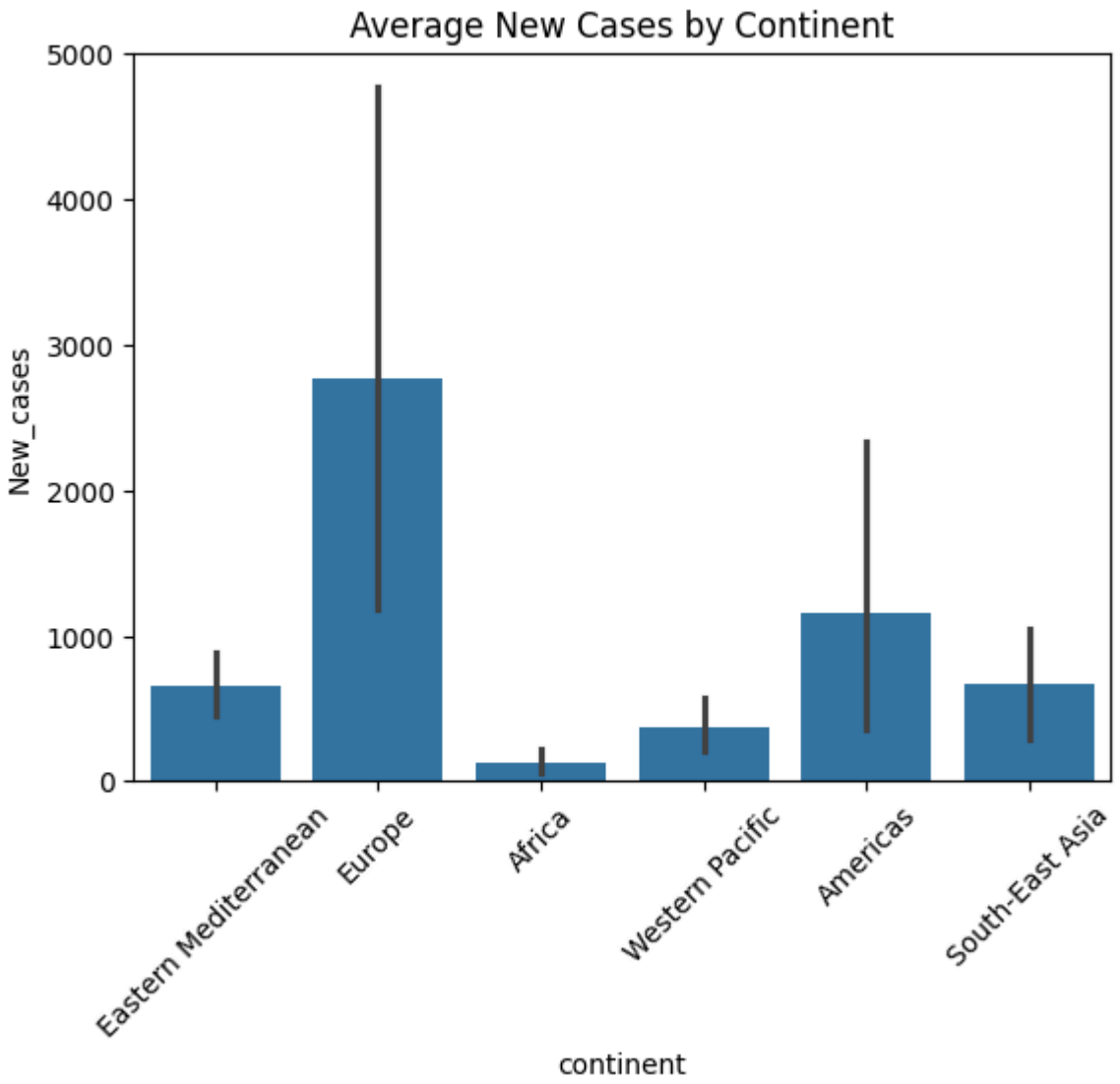
- ROC AUC RF: 0.261 (thấp, có thể do multiclass imbalance; curve cho thấy hiệu suất tốt ở một số class).


Confusion Matrix RF
20
0
0
0
13
0
1
1
13
0
1
2
0
1
2
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0


ROC Curve RF
Class 0
Class 1
Class 2
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

- **Đánh Giá**: RF tốt hơn (accuracy cao, confusion matrix ít lỗi). Feature Importance: Cumulative_deaths (0.35), New_deaths (0.25), New_cases (0.15), Cumulative_cases (0.10), continent_Europe (0.05).





- **Insight (RQ3)**: Tử vong (deaths) là yếu tố dự đoán rủi ro chính, châu lục (Europe) ảnh hưởng nhẹ.

IV. Conclusions

- **Trả Lời RQ**:
    - RQ1: Europe/Americas có phân bố ca/tử vong cao nhất (mean New_cases ~3000, outliers cao).
    - RQ2: Có, 3 cụm quốc gia: Cao rủi ro (Europe-dominated), Thấp (Africa/Western Pacific).
    - RQ3: Deaths và cases dự đoán rủi ro chính xác 96%, continent ảnh hưởng phụ.
    - RQ4: Xu hướng giảm từ 2022, đỉnh ở Europe/Americas 2020-2021.
- **Insight**: Đại dịch bất bình đẳng, châu Âu chịu ảnh hưởng nặng do dân số/mật độ (correlation deaths-cases cao). Mô hình ML rút tri thức hữu ích cho dự báo tương lai.
- **Thảo Luận**:
    - Hạn chế: Missing values nhiều (impute có thể bias), thiếu yếu tố kinh tế/y tế, ROC AUC thấp do imbalance.

V. Reference:

[1] https://data.who.int/dashboards/covid19/data?n=o

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

[4] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[7] https://scikit-learn.org/stable/modules/clustering.html