**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÁO CÁO CUỐI KỲ**Môn học : Dữ liệu lớn

Học kỳ I (2022-2023)

**ĐỀ TÀI:**

**ỨNG DỤNG KỸ THUẬT HỌC MÁY XỬ LÝ BÀI TOÁN TRÊN TẬP DỮ LIỆU NHỮNG CUỘC KHỦNG BỐ Ở MỸ**

Sinh viên thực hiện:

Trần Nhật Tân MSSV: 19522177

Bùi Ngọc Thành MSSV: 19522220

Huỳnh Quốc Khánh MSSV: 19521677

Lê Thế Tiệm MSSV: 19522330

GVHD: ThS. Nguyễn Hồ Duy Tri

Lớp: IS405.N11.HTCL

**Thành phố Hồ Chí Minh, tháng 12 năm 2022**

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÁO CÁO CUỐI KỲ**Môn học : Dữ liệu lớn

Học kỳ I (2022-2023)

**ĐỀ TÀI:**

**ỨNG DỤNG KỸ THUẬT HỌC MÁY XỬ LÝ BÀI TOÁN TRÊN TẬP DỮ LIỆU NHỮNG CUỘC KHỦNG BỐ Ở MỸ**

Sinh viên thực hiện:

Trần Nhật Tân MSSV: 19522177

Bùi Ngọc Thành MSSV: 19522220

Huỳnh Quốc Khánh MSSV: 19521677

Lê Thế Tiệm MSSV: 19522330

GVHD: ThS. Nguyễn Hồ Duy Tri

Lớp: IS405.N11.HTCL

**Thành phố Hồ Chí Minh, tháng 12 năm 2022**

**MỤC LỤC:**

[**LỜI CẢM ƠN** 4](#_Toc122478009)

[**NHẬN XÉT CỦA GIẢNG VIÊN** 5](#_Toc122478010)

[**I.** **TỔNG QUAN ĐỀ TÀI** 6](#_Toc122478011)

[1.1 Lý do chọn đề tài 6](#_Toc122478012)

[1.2 Mô tả dữ liệu 6](#_Toc122478013)

[1.2.1 Nguồn dữ liệu 6](#_Toc122478014)

[1.2.2 Các thuộc tính của tập dữ liệu 7](#_Toc122478015)

[1.3 Mô tả bài toán 9](#_Toc122478016)

[1.4 Kỹ thuật tiền xử lý dữ liệu được lựa chọn 9](#_Toc122478017)

[1.5 Thuật toán khai thác dữ liệu được lựa chọn 9](#_Toc122478018)

[1.5.1 Thuật toán K-Means 9](#_Toc122478019)

[1.6 Khoảng cách Euclidean 11](#_Toc122478020)

[1.7 Khoảng cách Manhattan 13](#_Toc122478021)

[1.8 K-Means Mapreduce 15](#_Toc122478022)

[**II.**  **TIỀN XỬ LÝ DỮ LIỆU** 16](#_Toc122478023)

[2.1 Các thư viện được sử dụng 16](#_Toc122478024)

[2.2 Đọc dữ liệu đầu vào 16](#_Toc122478025)

[2.3 Xóa các cột bị thiếu hoặc dư thừa 17](#_Toc122478026)

[2.4 Chuyển các thuộc tính Object thành Numeric 19](#_Toc122478027)

[2.5 Mô tả số thành phần của dữ liệu 19](#_Toc122478028)

[**III.** **THUẬT TOÁN KHAI THÁC DỮ LIỆU** 21](#_Toc122478029)

[**3.1** **Thuật toán K-Means** 21](#_Toc122478030)

[**IV.Kết quả đạt được** 50](#_Toc122478031)

[4.1 Phát biểu kết quả 50](#_Toc122478032)

[4.2 So sánh và đánh giá 53](#_Toc122478033)

[**V.Kết luận** 54](#_Toc122478034)

[5.1 Ưu điểm 54](#_Toc122478035)

[5.2 Khuyết điểm 54](#_Toc122478036)

[5.3 Hướng phát triển 54](#_Toc122478037)

[**VI.Tài liệu tham khảo** 55](#_Toc122478038)

[**HẾT.** 55](#_Toc122478039)

# **LỜI CẢM ƠN**

Trên thực tế không có sự thành công nào mà không gắn liền với những sự hỗ trợ, giúp đỡ dù ít hay nhiều, dù trực tiếp hay gián tiếp của người khác. Với lòng biết ơn sâu sắc nhất, đầu tiên em xin gửi lời cảm ơn chân thành đến tập thể quý Thầy Cô Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM và quý Thầy Cô khoa Hệ thống thông tin đã giúp cho nhóm có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này.

Đặc biệt nhóm chúng em xin gửi lời cảm ơn chân thành tới thầy Nguyễn Hồ Duy Tri – giảng viên lý thuyết và thực hành môn Dữ liệu lớn đã tận tình giúp đỡ, trực tiếp chỉ bảo, hướng dẫn em trong suốt quá trình làm đề tài. Nhờ đó, chúng em đã tiếp thu được nhiều kiến thức bổ ích trong việc vận dụng cũng như kỹ năng làm đề tài. Nếu không có những lời hướng dẫn, dạy bảo của thầy thì em nghĩ đề tài này của nhóm rất khó có thể hoàn thiện được. Một lần nữa, em xin chân thành cảm ơn thầy.

Cuối cùng, các thành viên trong nhóm đã làm việc hết công suất để hoàn thành tốt đề tài của mình. Xin chân thành cảm ơn!

# **NHẬN XÉT CỦA GIẢNG VIÊN**

# **TỔNG QUAN ĐỀ TÀI**

## Lý do chọn đề tài

Khủng bố là hoạt động phá hoại, đe dọa bằng lời nói, hình ảnh hoặc video giết người do cá nhân hoặc tổ chức thực hiện làm thiệt mạng người, đặc biệt là thường dân, hoặc gây tổn thất cho xã hội và cộng đồng để tác động vào tâm lý đối phương gây hoang mang khiếp sợ, nhằm mục đích chính trị hoặc tôn giáo. Một đặc tính thống nhất của khủng bố là việc sử dụng bừa bãi bạo lực đối với những người không có khả năng chống cự với mục đích là sự nổi tiếng cho một nhóm, một phong trào, một cá nhân hoặc gây áp lực lên đối thủ chính trị buộc họ phải chấp nhận một giải pháp chính trị có lợi cho mình. Các tổ chức khủng bố có thể khai thác nỗi sợ hãi của con người để hỗ trợ đạt được những mục tiêu này. Đối tượng bị khủng bố gây thiệt hại có thể là tính mạng, sức khỏe, danh dự nhân phẩm, tài sản (của cá nhân, tổ chức hay của nhà nước) hoặc sự vững mạnh của một chính quyền nhà nước.

Khủng bố là mối quan tâm lớn đối với các quốc gia trên toàn thế giới, vì nó thường báo hiệu mở đầu cho chiến tranh và dân thường bị mắc kẹt trong cuộc chiến. Chúng ta không thể thay đổi thực tế mà chúng ta không hiểu đầy đủ, vì vậy đây là một số dữ liệu liên quan về chủ đề này Nhận thấy tầm quan trọng ấy, nhóm tập trung nghiên cứu ứng dụng giải thuật K-Means để phân cụm những cuộc tấn công ở các thành phố trong quốc gia United States từ đó đề ra những chiến lược phù hợp giúp chống lại những cuộc khủng bố, góp phần mang lại an toàn cho người dân và xã hội.

## Mô tả dữ liệu

### Nguồn dữ liệu

Tên dữ liệu: Terrorism Data 1970to2017

Nguồn cung cấp: Kaggle.com

Link dữ liệu: [Terrorism Data 1970to2017 | Kaggle](https://www.kaggle.com/datasets/adiagarwalrock/terrorism-data-1970to2017?select=terrorismData.csv)

Mô tả nguồn dữ liệu:

Đây là một cơ sở dữ liệu nguồn mở bao gồm thông tin về các cuộc tấn công khủng bố trên toàn thế giới từ năm 1970 đến năm 2017. Bộ dữ liệu này bao gồm dữ liệu có hệ thống về các vụ khủng bố trong nước cũng như quốc tế đã xảy ra trong khoảng thời gian này.

### Các thuộc tính của tập dữ liệu

**-** Mỗi dòng trong tập dữ liệu là một vụ khủng bố tại Thổ Nhĩ Kỳ.

- Dữ liệu gồm 547 dòng và 17 thuộc tính.

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên cột | Kiểu dữ liệu | Ý nghĩa |
| 1 | Year | Int | Năm xảy ra vụ khủng bố |
| 2 | Month | Int | Tháng xảy ra vụ khủng bố |
| 3 | Day | Int | Ngày xảy ra vụ khủng bố |
| 4 | Country | String | Quốc gia |
| 5 | State | String | Bang của quốc gia |
| 6 | Region | String | Vùng của quốc gia |
| 7 | City | String | Thành phố xảy ra vụ khủng bố |
| 8 | Latitude | Double | Vĩ độ xảy ra cuộc khủng bố |
| 9 | Longtitude | Double | Kinh độ xảy ra cuộc khủng bố |
| 10 | AttackType | String | Kiểu tấn công |
| 11 | Killed | Int | Số người tử vong |
| 12 | Wounded | Int | Số người bị thương |
| 13 | Target | String | Mục tiêu của cuộc tấn công khủng bố |
| 14 | Summary | String | Diễn biến chi tiết của cuộc khủng bố |
| 15 | Group | String | Tổ chức khủng bố |
| 16 | Target\_type | String | Loại mục tiêu của cuộc tấn công khủng bố |
| 17 | Weapon\_type | String | Loại vũ khí dùng để khủng bố |

- Bảng dữ liệu:

Graphical user interface, table

Description automatically generated

## Mô tả bài toán

Mục tiêu của bài toán này là xây dựng một mô hình học máy để phân cụm cuộc khủng bố tại Mỹ để từ đó đưa ra các chiến lược phù hợp nhằm giúp chống lại những cuộc khủng bố, góp phần mang lại an toàn cho người dân và xã hội.

**Đây là project học máy kĩ thuật gom cụm với học không giám sát.**

## Kỹ thuật tiền xử lý dữ liệu được lựa chọn

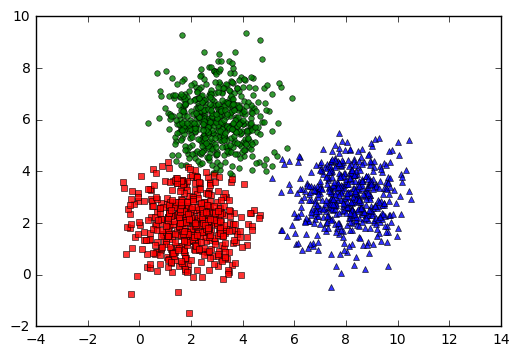
* Tìm cột không cần thiết.
* Tìm cột thiếu dữ liệu (missing values) .
* Tìm những cột chỉ có 1 giá trị.
* Chuyển các thuộc tính Object thành Numeric.
* Mô tả các thành phần của dữ liệu (count,min/max, mean, std)

## Thuật toán khai thác dữ liệu được lựa chọn

### Thuật toán K-Means

Thuật toán phân cụm K-means là một thuật toán phân cụm đơn giản thuộc loại học không giám sát (tức là dữ liệu không có nhãn) và được sử dụng để giải quyết bài toán phân cụm. Và là một phương pháp được sử dụng trong phân tích tính chất cụm của dữ liệu. Nó đặc biệt được sử dụng nhiều trong khai phá dữ liệu và thống kê. K-Means phân vùng dữ liệu thành k cụm khác nhau. Giải thuật này giúp xác định được dữ liệu thuộc về cụm (Cluster) nào.

Trong đó số lượng cụm được cho trước là k. Công việc phân cụm được xác lập dựa trên nguyên lý: Các điểm dữ liệu trong cùng 1 cụm thì phải có cùng 1 số tính chất nhất định. Tức là giữa các điểm trong cùng 1 cụm phải có sự liên quan lẫn nhau. Đối với máy tính thì các điểm trong 1 cụm đó sẽ là các điểm dữ liệu gần nhau.

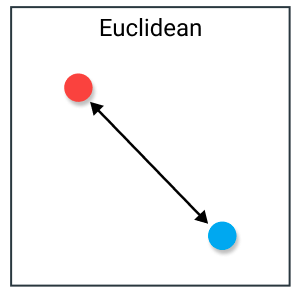


**Hình 1. Kết quả phân cụm bằng thuật toán k-means**

* Ý tưởng của thuật toán :
* Bước 1:   Chọn ngẫu nhiên K tâm (centroid) cho K cụm (cluster). Mỗi cụm được đại diện bằng các tâm của cụm.
* Bước 2:  Tính khoảng cách giữa các đối tượng (objects) đến K tâm (thường dùng khoảng cách Euclidean).
* Bước 3:  Nhóm các đối tượng vào nhóm gần nhất.
* Bước 4:  Xác định lại tâm mới cho các nhóm.
* Bước 5:  Thực hiện lại bước 2 cho đến khi không có sự thay đổi nhóm nào của các đối tượng.
* Ưu điểm :
* Dễ dàng thực hiện.
* Chia dữ liệu thành những tập lớn và đảm bảo sự hội tụ của các phần tử trong cụm.
* Với một số lượng lớn các biến, K-Means có thể nhanh hơn về mặt tính toán so với phân cụm phân cấp (nếu K nhỏ).
* K-means có thể tạo ra nhiều cụm hơn so với phân cụm phân cấp.
* K-Means tạo ra các cụm chặt chẽ hơn so với phân cụm phân cấp, đặc biệt nếu các cụm là hình cầu.
* Nhược điểm :
* Người dùng phải chỉ định k (số lượng cụm) khi thực hiện.
* k-means chỉ có thể xử lý dữ liệu số.
* Bị phụ thuộc vào các giá trị ban đầu.
* Khi xử lý các cụm lớn, nó có thể sẽ không hoạt động tốt.
* Các phân vùng ban đầu khác nhau có thể dẫn đến các cụm cuối cùng khác nhau.
* Không hoạt động tốt với các cụm (trong dữ liệu gốc) có kích thước khác nhau và mật độ khác nhau.

## Khoảng cách Euclidean

Khoảng cách Euclid giữa hai điểm trong không gian Euclid là độ dài của đoạn thẳng nối hai điểm đó. Là thước đo khoảng cách tốt nhất có thể được giải thích là độ dài của một đoạn nối hai điểm.



**Hình 2. Hình ảnh khoảng cách Euclidean**

* Công thức :

Diagram

Description automatically generated with low confidence

* Trường hợp sử dụng :
* Khoảng cách Euclid hoạt động hiệu quả khi bạn có dữ liệu chiều thấp và độ lớn của vector là điều quan trọng cần được đo.
* Các thuật toán như K-nn và HDBSCAN cho kết quả tốt nếu sử dụng khoảng cách Euclid trên dữ liệu chiều thấp.
* Mặc dù nhiều biện pháp khác đã được phát triển để giải quyết những nhược điểm của khoảng cách Euclide, nó vẫn là một trong những thước đo khoảng cách được sử dụng nhiều nhất vì những lý do chính đáng. [4]
* Nó cực kỳ trực quan để sử dụng, đơn giản để thực hiện và cho thấy kết quả tuyệt vời trong nhiều trường hợp sử dụng.
* Ưu điểm :
* Được sử dụng phổ biến, dễ hiểu, dễ thực hiện.
* Cho kết quả tốt trong nhiều usecase.
* Đặc biệt hiệu quả trên các tập dữ liệu ít chiều.
* Nhược điểm :
* Mặc dù nó là một thước đo khoảng cách phổ biến, khoảng cách Euclide không phải là biến thể theo tỷ lệ, có nghĩa là khoảng cách được tính toán có thể bị sai lệch tùy thuộc vào đơn vị của các đối tượng địa lý.
* Cần phải chuẩn hóa dữ liệu trước khi sử dụng thước đo khoảng cách này.
* Euclide distance có thể bị ảnh hưởng bởi đơn vị của feature. Chính vì vậy cần phải thực hiện normalize trước khi tính toán.
* Khi số chiều vector space tăng lên, Euclide Distance trở nên kém hiệu quả. Một phần nguyên nhân do dữ liệu thực tế thường không chỉ nằm trong Euclide Metric Space.

## Khoảng cách Manhattan

* Khoảng cách Manhattan
* Khoảng cách Manhattan tính toán sự khác biệt tuyệt đối giữa các tọa độ của các cặp đối tượng. Nó còn được gọi là khoảng cách L1​ (hay Taxicab/City Block distance).

A picture containing text, clock, device

Description automatically generated

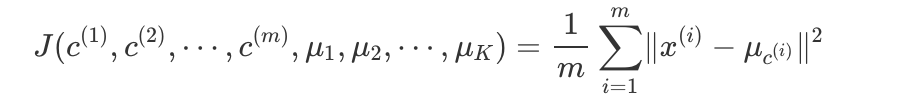
**Hình 3 Khoảng cách Manhattan**

* Công thức:

**Diagram

Description automatically generated**

* Trường hợp sử dụng :
* Khi tập dữ liệu của bạn có các thuộc tính rời rạc và / hoặc nhị phân, Manhattan dường như hoạt động khá tốt vì nó tính đến các đường dẫn mà thực tế có thể được thực hiện trong các giá trị của các thuộc tính đó.
* Ưu điểm :
* Trong một số trường hợp sẽ hiệu quả khi thực hiện trên các tập dữ liệu nhiều chiều.
* Nhược điểm :
* Mặc dù khoảng cách Manhattan có vẻ phù hợp với dữ liệu nhiều chiều , nhưng nó là một phép đo có phần kém trực quan hơn so với khoảng cách euclide, đặc biệt là khi sử dụng trong dữ liệu nhiều chiều.
* Hơn nữa, nó có nhiều khả năng cho một giá trị khoảng cách cao hơn khoảng cách euclide vì nó không phải là đường đi ngắn nhất có thể. Điều này không nhất thiết đưa ra các vấn đề nhưng là điều người dùng nên tính đến.
* Cost :
* Hàm cost dùng để tính tổng bình phương khoảng cách từ các điểm đến tâm tương ứng của chúng. Hàm này được dùng với mục đích tối thiểu hoá chi phí tính toán vì bước này liên quan đến việc gán một điểm dữ liệu cho cụm gần nhất có thể.
* Công thức:



Trong đó:

* c(i) là số thứ tự của cụm.
* là toạ độ tâm cụm K.
  + - * c(i) là toạ độ tâm cụm của mẫu x

## K-Means Mapreduce

* Map:
* Cho một điểm và tập hợp các tâm cụm.
* Tính khoảng cách từ điểm đó đến từng tâm cụm.s
* Phát ra điểm đó và tâm cụm gần nó nhất
* Reduce:
* Cho tâm cụm và các điểm thuộc cụm đó.
* Tính tâm cụm mới cho từng cụm bằng cách lấy trung bình cộng của tất các các điểm dữ liệu đã được gán vào cluster đó.
* Phát ra tâm cụm mới.

# **II. TIỀN XỬ LÝ DỮ LIỆU**

## 2.1 Các thư viện được sử dụng

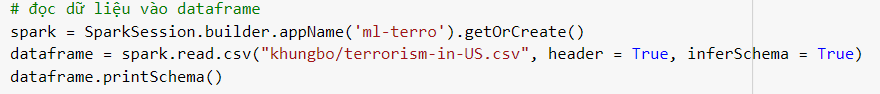
Text

Description automatically generated

**Hình 4. import các thư viện**

## 2.2 Đọc dữ liệu đầu vào

Sử dụng lệnh spark.read.csv để đọc file dữ liệu “terrorism-in-US.csv”.



**Hình 5. đọc file csv**

Sử dụng hàm describe.show để quan sát thống kê của bảng dữ liệu.

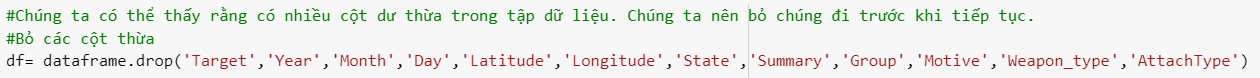
Ảnh có chứa bàn

Mô tả được tạo tự động

**Hình 6. kết quả dataframe của dữ liệu**

## 2.3 Xóa các cột bị thiếu hoặc dư thừa

* Sử dụng hàm drop để loại bỏ các cột dữ liệu dư thừa
* Cột Latitude và Longtitude làm mã định vị của cuộc khủng bố, không mang ý nghĩa trong tập dữ liệu.
* Cột Year, Month, Day mang ý nghĩa ngày, tháng, năm diễn ra cuộc khủng bố, không được chọn để phân tích nên có thể xóa.
* Cột Target, Summary, Group, Motive là những cuộc có kiểu dữl liệu string mô tả, diễn giải cuộc tấn công, không phù hợp để phân tích.



**Hình 7. Loại bỏ dữ liệu trống**

* Tìm **cột có dữ liệu thiếu**
* Thiếu dữ liệu là một sự xuất hiện phổ biến và có ảnh hưởng đáng kể đến các kết luận được rút ra từ dữ liệu. Để có thể kết luận chính xác hơn thì loại bỏ các dòng dữ liệu là cần thiết.

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Hình 8. Kiểm tra giá trị N/a của các cột dữ liệu**

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Hình 9. Kiểm tra giá trị Null của các cột dữ liệu**

* + Có thể thấy rằng cột dữ liệu Killed và Wouded là 2 cột bị thiếu dữ liệu.
* Tìm **cột chỉ có một giá trị**

**Ảnh có chứa bàn

Mô tả được tạo tự động**

**Hình 10. tìm cột chỉ có một giá trị**

* Cột Country và Region đều chỉ có 1 giá trị.

- Tiến hành loại bỏ những dòng giá trị thiếu.

Ảnh có chứa bàn

Mô tả được tạo tự động

**Hình 11. Xóa các côt không chọn**

## 2.4 Chuyển các thuộc tính Object thành Numeric

Vì dữ liệu của cột AttackType và Target\_type là dạng chuỗi nên sẽ không phục vụ được cho việc Mining dữ liệu, nhóm sẽ chuyển đổi dữ liệu sang dạng số

**Ảnh có chứa văn bản

Mô tả được tạo tự động**

**Hình 11. Tiến hành chuyển đổi kiểu dữ liệu**

**Ảnh có chứa bàn

Mô tả được tạo tự động**

**Hình 12 kết quả sau khi chuyển đổi kiểu dữ liệu**

## 2.5 Mô tả số thành phần của dữ liệu

Tiến hành mô tả các thành phần của dữ liệu như:

* Số lượng dòng dữ liệu(count)
* Giá trị cao nhất, thấp nhất(max,min)
* Giá trị trung bình, trung vị(mean)
* Độ lệch chuẩn(stddev)
* Sử dụng hàm describe.show để xem thông tin của tập dữ liệu sau khi xóa các cột.

Ảnh có chứa bàn

Mô tả được tạo tự động

**Hình 13. Kết quả các thành phần của dữ liệu**

* Sử dụng hàm toPandas.info để kiểm tra lại các cột của tập dữ liệu

Ảnh có chứa văn bản

Mô tả được tạo tự động

**Hình 14. kết quả các thành phần của dữ liệu**

# **THUẬT TOÁN KHAI THÁC DỮ LIỆU**

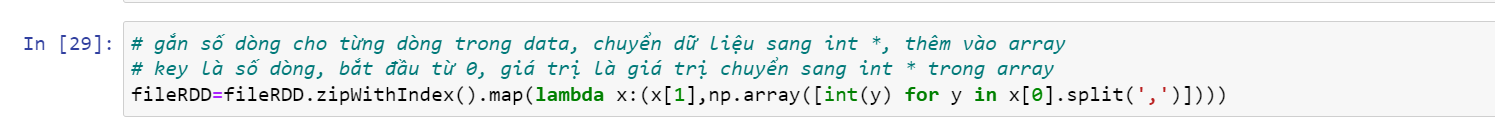
* 1. **Thuật toán K-Means**
* Tiến hành xuất dữ liệu sang file .csv. Sau đó đọc nhiều tệp dữ liệu vào một file RDD với số lượng phân mảnh nhỏ nhất (minPartitions) là 20.

Graphical user interface, text, application

Description automatically generated

**Hình 15. Xuất dữ liệu ra .csv và đọc các tệp vào file RDD**

* Thực hiện gắn số dòng cho từng dòng trong data, chuyển dữ liệu sang int\*, thêm vào array. Với key là số dòng, bắt đầu từ 0, giá trị là giá trị chuyển sang int\* trong array.



**Hình 16. Xử lý trên từng dòng trong data**

* Viết hàm **clust\_assn** có chức năng gắn cụm (cluster) gần nhất cho từng điểm dữ liệu (datapoint) **. Đây là bước tìm tâm cụm (Centroid ).** Hàm nhận tham số đầu vào là từng dòng trong file RDD và trả về giá trị id của cụm, khoảng cách từ datapoint đến cluster gần nhất, số lượng quan sát và tọa độ.

Trong đó :

* X là mỗi dòng của file RDD.
* Trong function return:
  + Temp là id của cụm.
  + dist: khoảng cách từ datapoint tới cụm (cluster) gần nhất.
  + x[0] là số quan sát.
  + x[1] là tọa độ (coordinate).

Graphical user interface, text

Description automatically generated

**Hình 17. Hàm clust\_assn**

* Viết hàm **generate\_initial\_centroids**  để lấy tâm cụm đầu tiên, xác định tọa độ của tâm và in thực hiện Broadcast trên tâm đầu tiên và tìm các điểm dữ liệu (datapoint) ở xa nhất**.** Hàm nhận tham số đầu vào là số cụm k.

Graphical user interface, text, application

Description automatically generated

**Hình 18. Hàm generate\_initial\_centroids**

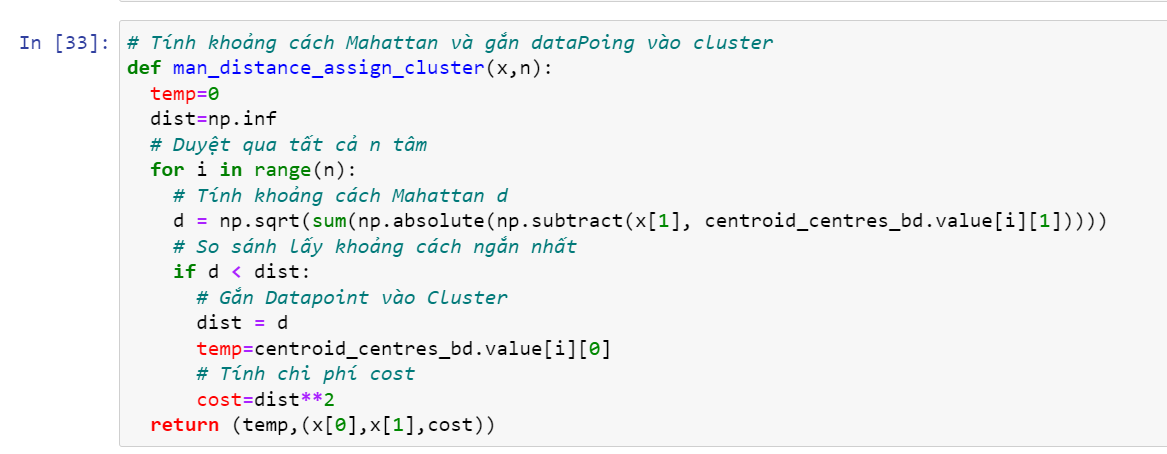
* Viết hàm **euc\_distance\_assign\_cluster** để tính khoảng cách **Euclid** và gắn các điểm dữ liệu (dataPoint) vào cụm (cluster). Hàm nhận tham số đầu vào là mỗi dòng của file RDD và phạm vi n. Kết quả trả về là giá trị id của cụm, số lượng quan sát, tọa độ và chi phí khoảng cách Euclid.

Graphical user interface, text, application, email

Description automatically generated

**Hình 19. Hàm euc\_distance\_assign\_cluster**

* Viết hàm **man\_distance\_assign\_cluster** để tính khoảng cách **Manhattan** và gắn các điểm dữ liệu (dataPoint) vào cụm (cluster). Hàm nhận tham số đầu vào là mỗi dòng của file RDD và phạm vi n. Kết quả trả về là giá trị id của cụm, số lượng quan sát, tọa độ và chi phí khoảng cách Manhattan.



**Hình 20. Hàm man\_distance\_assign\_cluster**

* Viết hàm **kmeans** để tiến hành chạy thuật toán phân cụm k-means. Hàm nhận tham số đầu vào là method, số cụm, số lượng phần tử bên trong và trả về giá trị cost khi chạy trên hai khoảng cách.

**Graphical user interface, text

Description automatically generated**

**Hình 21. Hàm kmeans**

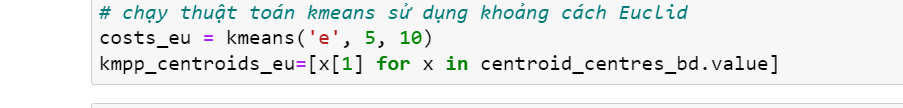
* Tiến hành chạy thuật toán k-means sử dụng khoảng cách **Manhattan.**

A picture containing text

Description automatically generated

**Hình 22. Chạy giải thuật k-means dùng Mahattan Distance**

* + Tiến hành chạy thuật toán k-means sử dụng khoảng cách **Euclid.**



**Hình 23. Chạy giải thuật k-means dùng Euclid Distance**

* + In các nhãn cụm (cluster) của dữ liệu ra và tính toán chi phí thuật toán sử dụng trên hai khoảng cách **Euclid** và **Manhattan**.

Text

Description automatically generated

**Hình 24. In các nhãn cụm của dữ liệu**

* Tính toán chi phí thuật toán sử dụng trên hai khoảng cách **Euclid** và **Manhattan**.

Graphical user interface, text, application, chat or text message

Description automatically generated

**Hình 25. Tính toán chi phí thuật toán sử dụng trên hai khoảng cách**

* + Quan sát được chi phí thuật toán với khoảng cách Euclid cao hơn nhiều so với chi phí trên khoảng cách Mahattan.
* In ra tâm các cụm có k-means sử dụng khoảng cách **Euclid** và khoảng cách **Manhattan**.

Text

Description automatically generated

**Hình 26. In các tâm cụm có kmeans sử dụng 2 khoảng cách**

* Hiệu suất thuật toán phân cụm đạt được khi thực hiện trên khoảng cách **Euclid**.

Chart, line chart

Description automatically generated

**Hình 27. Kết quả hiệu suất giải thuật với khoảng cách Euclid**

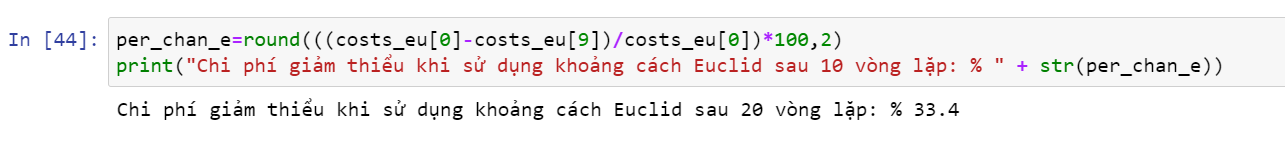
* Hiệu suất thuật toán phân cụm đạt được khi thực hiện trên khoảng cách **Manhattan**.

Chart, line chart

Description automatically generated

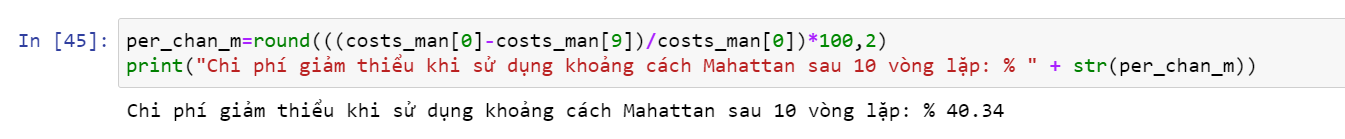
**Hình 28. Kết quả hiệu suất giải thuật với khoảng cách Mahattan**

* Tính toán chi phí giảm thiểu khi sử dụng khoảng cách **Euclid** sau khi thực hiện qua 20 vòng lặp.



**Hình 29. Kết quả chi phí giảm thiểu thực hiện trên Euclid Distance**

* + Chi phí ước tính giảm thiểu khoảng 33.4%.
* Tính toán chi phí giảm thiểu khi sử dụng khoảng cách **Manhattan** sau khi thực hiện qua 20 vòng lặp.



**Hình 30. Kết quả chi phí giảm thiểu thực hiện trên Manhattan Distance**

* + Chi phí ước tính giảm thiểu khoảng 40.34%.
* Import các thư viện để xử lý dữ liệu trước khi thực hiện thuật toán để tìm đặc trưng cụm.

Graphical user interface, text, application

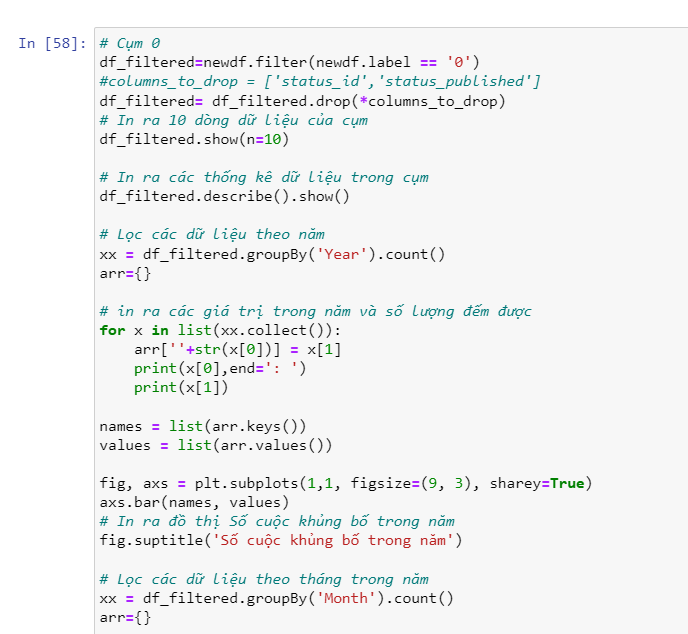
Description automatically generated

Table

Description automatically generated

**Hình 31. Kết quả kiểm tra một vài dòng dữ liệu đầu sau khi xử lý**

* **Cụm 0 :**

****

**Text

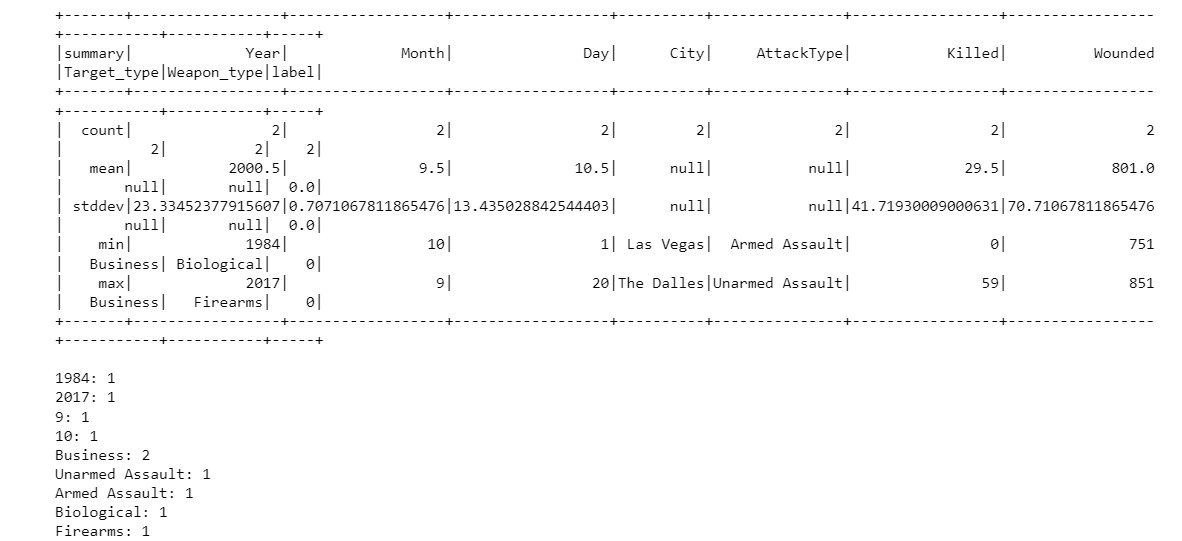
Description automatically generated**

* Hiển thị kết quả

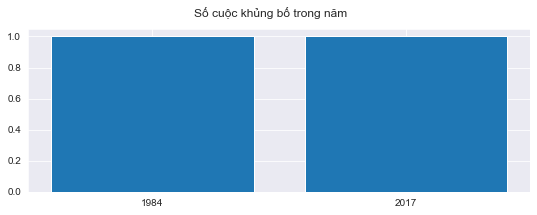
**Table

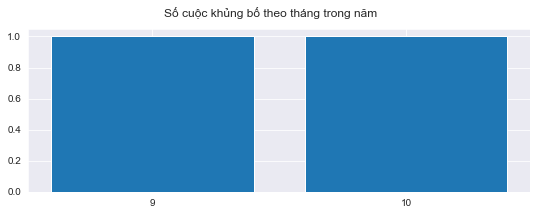
Description automatically generated**

* Hiển thị các thành phần của dữ liệu.



* Kết quả cụm 0





Text

Description automatically generated with low confidence

Chart, bar chart

Description automatically generated

Chart, waterfall chart

Description automatically generated

* **Cụm 1 :**

Text

Description automatically generated

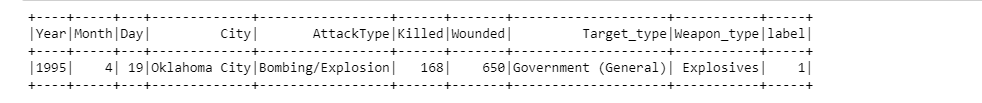
Text

Description automatically generated

Text

Description automatically generated

* Hiển thị kết quả



* Hiển thị các thành phần của dữ liệu

Table

Description automatically generated with low confidence

* Kết quả cụm 1

Chart, treemap chart

Description automatically generated

A picture containing graphical user interface

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Text

Description automatically generated with medium confidence

Text

Description automatically generated with low confidence

* **Cụm 2 :**

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

* Hiển thị kết quả 10 dòng đầu

Table

Description automatically generated

* Hiển thị các thành phần của dữ liệu

Table

Description automatically generated

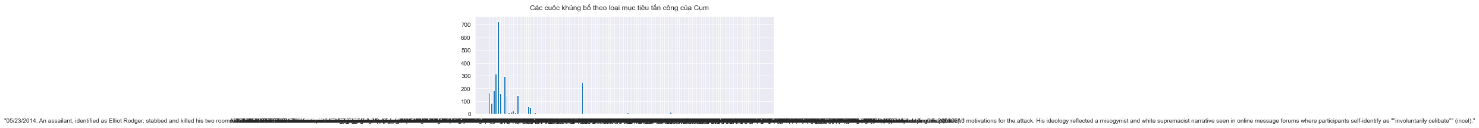
* Kết quả cụm 2

Chart, histogram

Description automatically generated

Chart, bar chart

Description automatically generated



Chart, histogram

Description automatically generated

A picture containing graphical user interface

Description automatically generated

* **Cụm 3** :

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generatedText

Description automatically generated

* Hiển thị kết quả

A picture containing calendar

Description automatically generated

* Hiển thị các thành phần của dữ liệu

Table

Description automatically generatedTable

Description automatically generated

* Kết quả cụm 3

Chart, treemap chart

Description automatically generated

A picture containing text

Description automatically generated

Graphical user interface

Description automatically generated with low confidence

Chart

Description automatically generated with low confidenceA picture containing graphical user interface

Description automatically generated

* **Cụm 4 :**

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

* Hiển thị kết quả

A picture containing diagram

Description automatically generated

* Hiển thị các thành phần của dữ liệu

Table

Description automatically generated

* Kết quả cụm 4

Chart, treemap chart

Description automatically generated

A picture containing text

Description automatically generatedGraphical user interface, text, application

Description automatically generated

Chart

Description automatically generated with medium confidence

A picture containing graphical user interface

Description automatically generated

# **IV.Kết quả đạt được**

## 4.1 Phát biểu kết quả

**Cụm 0:**

* Ý nghĩa phân cụm: những cuộc khủng bố ở cụm 0 chủ yếu là khủng bố theo loại “Assault” với chỉ có hai cuộc khủng bố nhưng có số người chết và bị thương rất cao.

- Phân tích:

Hai cuộc khủng bố xảy ra vào tháng 9 năm 1984 và tháng 10 năm 2017 đều nhắm vào “Business”.

Cuộc khủng bố ở Las Vegas năm 2017 có lượng người chết và tai nạn cao hơn ở The Dalles năm 1984. Mặc dù ở thành phố The Dalles không có thương vong nhưng số người bị thương cũng rất cao

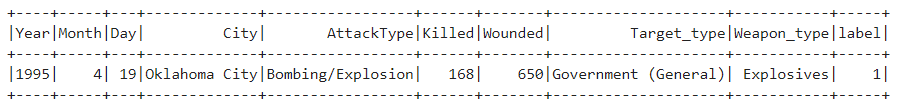
Table

Description automatically generated

**Cụm 1:**

- Ý nghĩa: Mặc dù năm 1995 chỉ xảy ra 1 cuộc khủng bố nhưng hậu quả để lại là cnhiều người chết và bị thương.

- Phân tích: cuộc khủng bố diễn ra vào tháng 4 năm 1995 và mục tiêu bị nhắm tới là “Goverment”. Bọn khủng bố đã sử dụng cách tấn công là “Bombing/Exploision” khiến 168 người thiệt mạng và 650 người bị thương.

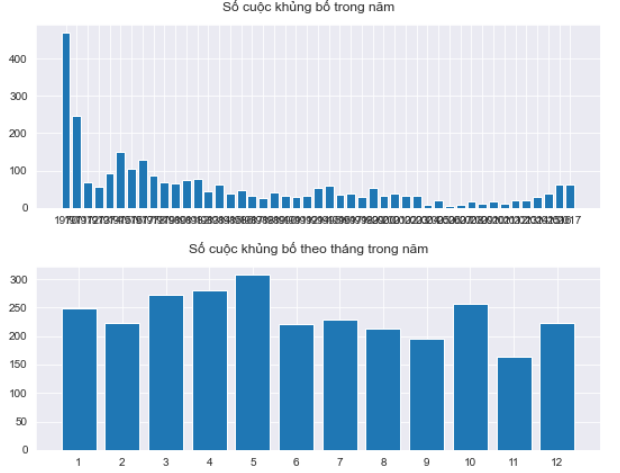


**Cụm 2:**

-Ý nghĩa: Đây là cụm có nhiều cuộc khủng bố nhất kéo dài từ năm 1967 đến năm 2017. Cách tấn công chủ yếu là “Bombing/Exploision” .

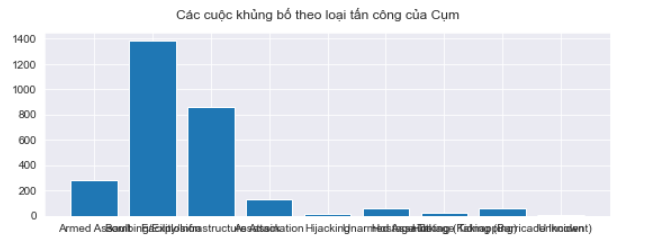
- Phân tích:

Năm 1970 và 1971 có nhiều cuộc khủng nhất trong tất cả các năm(468 và 247) và các cuộc khủng bố xảy ra đều giữa các tháng.



A picture containing table

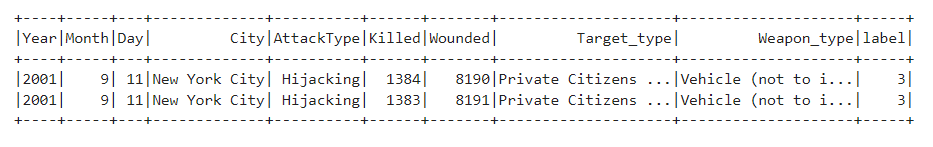
Description automatically generated



**Cụm 3:**

- Ý nghĩa cụm: 2 cuộc khủng bố ở cụm 3 đều có loại tấn công là “Hijacking”. Cả 2 đã gây ra nhiều thương vong và làm rất nhiều người bị thương.

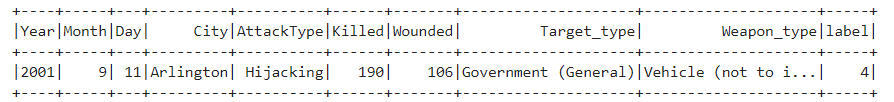
- Phân tích: 2 cuộc khủng bố xảy ra vào tháng 9 năm 2021 và mục tiêu được nhắm đến là các cư dân và tài sản riêng của họ. Cả 2 cuộc khủng bố không có sự chênh lệch về số người tử vong và bị thương.



**Cụm 4:**

- Ý nghĩa: cuộc khủng bố thuộc cụm 4 có loại tấn công là “Hijacking”. Mặc dù chỉ có một cuộc tấn công nhưng cũng khiến nhiều người tử vong và bị thương.

- Phân tích: cuộc khủng bố xảy ra vào tháng 9 năm 2021 và mục tiêu bị nhắm đến là “Goverment” ở thành phố Arlington.



## 4.2 So sánh và đánh giá

Để so sánh hai thuật toán này, tỷ lệ chi phí so với số ô lặp lại được quan sát cho cả khoảng cách euclidean và manhattan.

Đối với kmeans, điểm bắt đầu của chi phí thấp hơn kmeans ++. Tuy nhiên, giá trị điểm cuối của hàm chi phí cao hơn kmeans ++.

Điều này cho thấy đối với kmeans ++ mặc dù nó bắt đầu ở hàm chi phí cao hơn, nhưng sự hội tụ sẽ sớm đạt được với giá trị nhỏ nhất của hàm chi phí so với kmeans.

Điều này chứng minh rằng kmeans ++ có thể là thuật toán phân cụm hiệu quả nhất trong số hai thuật toán này.

|  |  |
| --- | --- |
| **Euclid Distance** | **Manhattan Distance** |
| Giá trị chi phí nhiều hơn | Giá trị chi phí thấp hơn |
| Chi phí giảm thiểu sau sau khi thực hiện ít hơn (33,4%) | Chi phí giảm thiểu sau sau khi thực hiện ít hơn (40.34%) |

**Nhận xét:**

* Thực hiện chạy thuật toán với khoảng cách Manhattan giảm thiểu được nhiều chi phí thực hiện hơn và cho hiệu quả cao hơn.

# **V.Kết luận**

## 5.1 Ưu điểm

* Ứng dụng được bài toán phân cụm và giải quyết được vấn đề đặt ra.
* Áp dụng được các phương pháp đánh giá và so sánh mô hình.

## 5.2 Khuyết điểm

Vì thời gian để nghiên cứu và hiện thực đề tài còn giới hạn do đó còn tồn tại một số hạn chế, như sau:

* Chưa ứng dụng thêm các Distance Measure khác vào xử lý bài toán.
* Chưa nghiên cứu tìm thêm một số thuật toán khác tối ưu hơn.

## 5.3 Hướng phát triển

* Ứng dụng kết quả đề tài vào các dự án theo hướng phân tích mạng xã hội.
* Sử dụng nhiều tập dataset để đánh giá hiệu năng của mô hình.
* Nghiên cứu sử dụng thêmnhiều thuật toán khác để đánh giá và so sánh từ đó tối ưu kết quả đạt được.
* Tiếp tục tối ưu các thuật toán sử dụng để đưa ra kết quả chính xác và thời gian nhanh hơn.
* Kết hợp các thuật toán lại thành một mô hình hoàn chỉnh.

# **VI.Tài liệu tham khảo**

1. Link Dataset “Terrorism Data 1970to2017” - [Terrorism Data 1970to2017 | Kaggle](Terrorism%20Data%201970to2017%20|%20Kaggle)

2. “K-Means Clustering” - <https://machinelearningcoban.com/2017/01/01/kmeans/>

3. “K-means clustering - Wikipedia” - <https://en.wikipedia.org/wiki/K-means_clustering>

4. “9 phép đo khoảng cách trong khoa học dữ liệu” - <https://ichi.pro/vi/9-phep-do-khoang-cach-trong-khoa-hoc-du-lieu-159983401462266>

5. “Distance Measure trong Machine Learning” - [https://viblo.asia/p/distance-measure-trong-machine-learning-ByEZkopYZQ0](https://viblo.asia/p/distance-measure-trong-machine-learning-ByEZkopYZQ0%20)

# **HẾT.**