**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

---🙡 🕮 🙣---



**BÀI TẬP LỚN MÔN HỌC**

**KHAI PHÁ DỮ LIỆU**

ĐỀ TÀI:

Phân tích, khai phá dữ kiệu tập dữ liệu doanh nghiệp

**Sinh viên thực hiện:**

**Đặng Minh Phúc Lê Trung Hậu Hồ Sỹ Quang Trung**

**Lớp:**

**ITEC3417**

**Giảng viên hướng dẫn:**

**Nguyễn Trung Hậu**

**Tháng 11 năm 2023**

Mục lục

[Lời nói đầu 2](#_Toc151257139)

[1. Bảng Phân Công 3](#_Toc151257140)

[2.Giới thiệu về nguồn dữ liệu, phương pháp thu thập 3](#_Toc151257141)

[3. Tiền xử lý dữ liệu: 3](#_Toc151257142)

[3.1 Phân tích tập dữ liệu 3](#_Toc151257143)

[3.2. Xử lý dữ liệu 4](#_Toc151257144)

[3.3 Biến đổi dữ liệu 4](#_Toc151257145)

[4.Mô tả tập dữ liệu sau khi xử lý 4](#_Toc151257146)

[5. Phương pháp khai phá dữ liệu 5](#_Toc151257147)

[5.1 Trực quan hóa, biểu diễn dữ liệu 5](#_Toc151257148)

[5.2 Mô hình hóa 5](#_Toc151257149)

[6. Kết quả thực nghiệm của mô hình và đối sánh với các lập luận đã đưa ra. 5](#_Toc151257150)

[6.1 Trực quan hóa, biểu diễn dữ liệu 5](#_Toc151257151)

[6.2 Mô hình hóa 7](#_Toc151257152)

[7. Kết luận 9](#_Toc151257153)

# Lời nói đầu

Dữ liệu doanh nghiệp là dữ liệu vô cùng lớn với nhiều khịa cạnh khai thác. Nhóm chúng tôi chọn sử dụng tập dữ liệu này với mong muốn tìm hiểu thêm về các ngành, lịch sử phát triển và lý do đằng sau sự phát triển đó

# Bảng Phân Công

|  |  |  |
| --- | --- | --- |
| **MSSV** | **Tên** | **Nhiệm vụ** |
| 2151010290 | Đặng Minh Phúc | - Tiền xử lý dữ liệu thống kê, báo cáo |
| 2151013020 | Lê Trung Hậu | - Chọn mô hình khai phá phù hợp  - Tổng hợp dữ liệu  - Phương pháp đánh giá mô hình |
| 1951052217 | Hồ Sỹ Quang Trung | - Tìm kiếm, gộp, làm sạch data và mô tả tập dữ liệu sau khi xử lý. |

# 2.Giới thiệu về nguồn dữ liệu, phương pháp thu thập

Đầu tiên, nhóm bắt đầu quá trình thu thập dữ liệu từ các nền tảng cung cấp tập dữ liệu mở như Kaggle, Gov, Goggle dataset,... Mục tiêu của nhóm là những thông tin dữ liệu quan đến nhiều khía cạnh khác nhau của doanh nghiệp để tiến hành khai phá và phân tích. Để đạt được điều này, nhóm đã tiến hành quá trình đối chiếu toàn diện trên nhiều nguồn thông tin khác nhau để đúc kết ra một bộ dữ liệu chuyên sâu về doanh nghiệp.

Qua quá trình này, nhóm đã tìm hiểu và tổng hợp được một tập dữ liệu đa dạng và đầy đủ về các khía cạnh của doanh nghiệp.

Các tập dữ liệu mà trong quá trình nhóm tìm hiểu, thu thập được bao gồm:

* Bản ghi ( record ):
* 7+ Million Company ([link](https://www.kaggle.com/datasets/peopledatalabssf/free-7-million-company-dataset))
* Continent List for 2021 Olympics in Tokyo Dataset ([link](https://www.kaggle.com/datasets/arvinthsss/continent-list-for-2021-olympics-in-tokyo-dataset))

# 3. Tiền xử lý dữ liệu:

## 3.1 Phân tích tập dữ liệu

Sau khi hoàn tất quá trình thu thập dữ liệu, cần xác định những thông tin trong tập dữ liệu đã được làm sạch chưa. Nếu dữ liệu chưa được làm sạch sẽ ảnh hưởng đến độ tin cậy của dữ liệu dẫn đến các quyết định không chính xác và gây khó khăn trong quá trình khai phá.

Một số vấn đề nhóm gặp phải khi chưa làm sạch dữ liệu trong tập dữ liệu **7+ Million Company**:

* Dữ liệu không hoàn chỉnh (incomplete): Thiếu các giá trị thuộc tính:
  + Tổng giá trị bị mất: 8,512,762
    - Trong đó:
      * year\_founded: 3510217 (50.11%)
      * industry: 277286 (3.96%)
      * country: 2440634 (34.84%)
      * locality: 2284625 (32.62%)
* Nhiễu/Lỗi (noise/error):
  + Các giá trị year\_founded lớn hơn năm hiện tại: 7 (2029, 2027, 2103, 2025, 2025, 2025, 2025)

Giải pháp:

- Tiến hành lọc từng cụm dữ liệu: xóa các đối tượng thiếu giá trị, lọc các giá trị nhiễu,…

Một số vấn đề nhóm gặp phải khi chưa làm sạch dữ liệu trong tập dữ liệu **Continent List for 2021 Olympics in Tokyo Dataset**:

* Dữ liệu không trùng khớp (mismatched): một số giá trị với cùng tính chất nhưng khác tên gây thiếu sót trong gộp dữ liệu bên cạnh một số thuộc tính thiếu so với tập dữ liệu cần gộp

Giải pháp:

- Tiến hành chuẩn hóa tên các thuộc tính, bổ sung các giá trị còn thiếu,…

## 3.2. Xử lý dữ liệu

- Đối với tập dữ liệu ‘7+ Million Company’:

+ Tiến thành xóa trống các đối tượng chứa thuộc tính nhiễu, thiếu nhưng không làm thay đổi tính chất, sự cân bằng của tập dữ liệu gốc.

- Đối với tập dữ liệu ‘’Continent List for 2021 Olympics in Tokyo Dataset’:

+ Tiến hành đối chiếu các thuộc tính không trùng giữa tập dữ liệu này và ‘7+ Million Company’, tiến hành tinh chỉnh tên giá trị, thêm các giá trị bị thiếu, loại bỏ 1 số giá trị không mang nhiều ý nghĩa để giảm nhẹ chi phí tính toán mô hình mà không làm ảnh hướng dữ liệu chung.

## 3.3 Biến đổi dữ liệu

Biển đổi dữ liệu là việc chuyển toàn bộ tập giá trị của một thuộc tính sang một tập giá trị thay thế sao cho giá trị cũ tương ứng giá trị mới.

Trong bài tập lớn này nhóm chúng tôi sử dụng các phương pháp biến đổi sau:

-Làm trơn (smoothing): Loại bỏ nhiễu/lỗi khỏi tập dữ liệu.

-Chuẩn hóa ( normalization): Đưa các giá trị về một khoảng được chỉ định.

# 4.Mô tả tập dữ liệu sau khi xử lý

**A screenshot of a computer

Description automatically generated**

* Tổng quan tập dữ liệu sau tiền xử lý:
  + 7046000 đối tượng
  + 8 thuộc tính
  + Số giá trị bị mất 0 (0%)
  + Các cột giá trị bị lệch: ‘year\_founded’, ’ current\_employee\_estimate’, ‘ total\_employee\_estimate’
  + Các cột có giá trị 0:
    - current\_employee\_estimate’: 1414862 (20.2%)

# 5. Phương pháp khai phá dữ liệu

## 5.1 Trực quan hóa, biểu diễn dữ liệu

Chúng tôi sử dụng các biểu đồ cũng như các bảng để thể hiện các xu hướng dữ liệu, sự phân bổ để tìm ra các xu hướng cũng như giải thích, chứng minh cho cá xu hướng đó

## 5.2 Mô hình hóa

Chúng tôi sử dụng mô hình gom cụm cùng thuật toán K-means cho tập dữ liệu này vì đây là mô hình khá phổ biển cũng như dễ áp dụng để tìm ra các đặc trưng trong các cụm dữ liệu.

Vấn đề đặt ra cho việc gom cụm dữ liệu là chọn các k clusters phù hợp. Có nhiều cách tiếp cận để tìm k cluster như: Elbow method, AffinityMatrix, SilhoutteScore,… và dựa vào kinh nghiệm cá nhân. Trong tập dữ liệu này chúng tôi tiếp cận với Elbow method là một trong những cách phổ biến nhất để tối ưu k clusters. Chúng tôi lặp lại các giá trị của k từ 1 đến 11 và tính toán độ biến dạng hoặc quán tính cho từng giá trị của k trong phạm vi đã cho. Trong đó Quán tính là tổng bình phương khoảng cách của các mẫu đến tâm cụm gần nhất của chúng. Để xác định số cụm tối ưu, chúng ta phải chọn giá trị k tại “khuỷu tay”, tức là điểm mà sau đó độ méo/quán tính bắt đầu giảm theo kiểu tuyến tính.

# 6. Kết quả thực nghiệm của mô hình và đối sánh với các lập luận đã đưa ra.

## 6.1 Trực quan hóa, biểu diễn dữ liệu

Tổng quan:

* Thuộc tính năm thành lập:

A screenshot of a computer

Description automatically generated

A graph of a number of companies

Description automatically generated

**Biểu đồ phân bố năm thành lập sau 1900**

* Phần lớn các công ty thành lập vào những năm 2000 với giá trị Mean rơi vào giá trị 2001. Trong đó công ty thành lập cũ nhất vào năm 1451 và mới nhất vào 2023
* Các thuộc tính khác:

A close-up of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with orange bars

Description automatically generated

Ngành công nghệ thông tin chiếm đa số, phần lớn công ty thành lập có quy mô nhỏ và Mỹ là quốc gia với nhiều công ty nhất.

* Phần lớn các công ty CNTT ra đời vào nhừng năm 2000 cùng với sự phát triển nhanh của mạng internet.
* Xu hướng phát triển các công ty đang theo hướng ngành CNTT trong thời đại phát triển nhanh của các công nghệ hiện đại

## 6.2 Mô hình hóa

- Sau khi thực thi thuật toán chúng tôi tìm ra số cụm tối ưu cho tập dữ liệu này là 6

A graph with a line

Description automatically generated

* Sau đó chúng tôi tiến hành gom cụm bằng thuật toán K-means

Kết quả sau khi gom cụm như sau:

A black rectangular object with white text

Description automatically generated A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* Các cụm phân chia theo quy mô công ty. Trong đó đặc biệt với cụm 6 gồm các công ty với quy mô lớn và số lượng nhân viên đông đảo nhất tập trung ở các quốc gia thuộc châu Âu, châu Á và châu Mỹ. Trong khi đó các cụm còn lại với quy mô thấp hơn phân bổ đều cho các châu lục

# 7. Kết luận

Sau khi thực hiện khai phá nhóm rút ra các kết quả, hạn chế cũng như các đề xuất phát triển như sau:

* Về kết quả: Nhóm đã dùng các kỹ thuật để tiến hành khai phá dữ liệu, tìm hiểu được các góc nhìn của tập dữ liệu, là bước đầu cho các công việc trong việc khai thác tri thức
* Về mặt hạn chế:
  + Nhóm còn hạn chế trong kinh nghiệm đọc các dữ liệu sau phân tích, các kết quả sau phân tích thường được thể hiện rõ qua trực quan mà chưa có kinh nghiệm tron khai thác sâu dữ liệu
  + Do hạn chế về kiến thức nên nhóm chỉ thực hiện 1 mô hình, vì không đảm bảo tối ưu và kết quả khi áp dụng nhiều mô hình
  + Xử lý dữ liệu chưa thật sự tốt với các kết quả sau gom cụm còn lọt nhiều giá trị nhiễu gây ảnh hướng kết quả chung
* Về mặt đề xuất phát triển:
  + Nhóm có thể dành thêm thời gian đầu tư tham khảo các phương pháp khai phá khác, học hỏi thêm kinh nghiệm về dữ liệu
  + Tối ưu, cải tiến xử lý dữ liệu sâu hơnn để đảm bảo không lọt các giá trị không đáp ứng yêu cầu của mô hình nhằm đạt kết quả cao hơn
  + Áp dụng nhiều loại mô hình nhiều các tối ưu mô hình hơn đế có thêm nhiều tri thức, góc nhìn sâu hơn khi gặp các tập dữ liệu phức tạp khác.

# Phục lục:

Github: [link](https://github.com/Haule9-2/Data_mining_propject.git)