TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

**KHOA CÔNG NGHỆ THÔNG TIN**

---🙡 🕮 🙣---



**BÁO CÁO MÔN HỌC**

**KHAI PHÁ DỮ LIỆU**

**Đề tài**

**ĐÁNH GIÁ THANG ĐO HẠNH PHÚC DỰA THEO DỮ LIỆU THỐNG KÊ**

**Sinh viên thực hiện: Trần Nhật Minh - 2051052081**

**Hồ Quang Văn - 2051050561**

**Lê Đông Anh Kiệt - 2051052068**

**Lớp: IT2002**

**Giảng viên hướng dẫn: TS. Nguyễn Tiến Đạt**

**Thầy Nguyễn Văn Bảy**

**Tháng 08 năm 2023**

**MỤC LỤC**

[Chương 1: Giới thiệu đề tài 3](#_Toc143443043)

[**1.** **Mô tả dữ liệu bài toán** 3](#_Toc143443044)

[**2.** **Mô tả hướng khai thác dữ liệu** 3](#_Toc143443045)

[Chương 2: Công cụ và thuật toán 4](#_Toc143443046)

[**1.** **Công cụ Google colab** 4](#_Toc143443047)

[**2.** **Chuẩn hóa dữ liệu** 4](#_Toc143443048)

[Chương 3: Đặc tả bài toán 7](#_Toc143443049)

[1. Các quốc gia có mức xếp hạng cao trong 8 năm 7](#_Toc143443050)

[2. Phân tích các chỉ số theo khu vực trong năm 2022 10](#_Toc143443051)

[**2.1** **Theo chỉ số hạnh phúc** 10](#_Toc143443052)

[**2.2** **Theo chỉ số GDP** 11](#_Toc143443053)

[**2.3** **Theo tuổi thọ** 11](#_Toc143443054)

[**2.4** **Theo chế độ phúc lợi xã hội** 12](#_Toc143443055)

[**2.5** **Theo mức độ đời sống tinh thần tự do** 13](#_Toc143443056)

[**2.6** **Theo mức độ hào phóng** 14](#_Toc143443057)

[**2.7** **Theo mức độ nhận thức về tham nhũng** 14](#_Toc143443058)

[3. Mối tương quan của các số liệu trong 8 năm 15](#_Toc143443059)

[4. Hồi quy tuyến tính 16](#_Toc143443060)

[5. Phân cụm KMeans 17](#_Toc143443061)

[6. Luật kết hợp 19](#_Toc143443062)

[7. Phân lớp 20](#_Toc143443063)

[8. Cây quyết định 22](#_Toc143443064)

[Chương 4: Kết luận 24](#_Toc143443065)

# Chương 1: Giới thiệu đề tài

1. **Mô tả dữ liệu bài toán**

Phạm vi nghiên cứu: các tệp thống kê dữ liệu World Happiness Report từ năm 2015 đến năm 2022 mà giảng viên cung cấp.

Gồm các cột thuộc tính: 'Country', 'Region', 'Happiness Rank', 'Happiness Score', 'Standard Error', 'Economy (GDP per Capita)', 'Family', 'Health (Life Expectancy)', 'Freedom', 'Trust (Government Corruption)', 'Generosity', 'Dystopia Residual'.

Ý nghĩa thuộc tính ảnh hưởng kết luận:

* Dựa vào các cột thuộc tính ta đánh giá sự tương quan giá trị để xem quốc gia nào có mức độ hạnh phúc cao nhất, thấp nhất và bình ổn trong khoảng thời gian năm 2015 đến năm 2022.
* Dựa vào các cột như số hộ gia đình ‘Family’, phát triển về kinh tế GDP của mỗi nước thể hiện đời sống vật chất, mức độ thang đo sức khỏe cho thấy đời sống tinh thần, số điểm hạnh phúc và sắp xếp thứ hạng dựa vào số điểm của mỗi quốc gia.

1. **Mô tả hướng khai thác dữ liệu**

Trước hết, ta thực hiện xử lý chuẩn hóa dữ liệu để sàng lọc được bảng dữ liệu đủ tiêu chí về giá trị. Sau khi đã lấp đầy bảng dữ liệu và không bỏ sót các ô có giá trị rỗng thì lúc này tệp dữ liệu đã đạt điều kiện.

Ta xét các quốc gia qua từng năm.

# Chương 2: Công cụ và thuật toán

1. **Công cụ Google colab**

Ở đề tài này, ta sử dụng ngôn ngữ lập trình Python và công cụ thực hiện là Google colab.

Đây là một chức năng tiện ích của Google cung cấp miễn phí cho người dùng có tài khoản google. Google colab sử dụng trực tuyến cho phép người dùng download và upload tệp tin cá nhân lên, sử dụng chung dung lượng của Google drive. Không cần cài đặt như các trình IDE hao tốn tài nguyên trên máy nên rất phù hợp cho người dùng là lập trình viên sử dụng, ngoài ra, với quy mô của Google colab cũng có thể chia sẻ quyền truy cập để người dùng khác có thể cùng tham gia dự án mà không cần thao tác phức tạp như github. Ngay cả khi không sử dụng thiết bị cá nhân, chỉ cần đăng nhập tài khoản trên thiết bị bất kỳ kết nối mạng đều có thể thao tác với dự án một cách linh hoạt, tiện lợi.

Ưu điểm:

* Thuận tiện, đơn giản, linh hoạt cho việc lập trình thường xuyên phải di chuyển.
* Bảo mật tốt vì google có cơ chế xác thực người dùng bảo mật hai lớp khi đăng nhập trên bất cứ thiết bị nào ngoài thiết bị gốc.
* Có phân quyền người dùng khác được phép truy cập.
* Miễn phí cho mọi đối tượng tài khoản.
* Môi trường thực thi tài nguyên mạnh mẽ không phụ thuộc vào thiết bị truy cập như chạy các IDE cục bộ khác phải tùy dung lượng RAM mỗi máy.
* Hỗ trợ nhiều ngôn ngữ lập trình và công cụ dùng cho mảng phân tích khái phá dữ liệu.
* Các dữ liệu được mô tả dễ nhìn, dễ tiếp cận.

Khuyết điểm:

* Thiết bị dùng truy cập phải kết nối mạng.
* Chưa hỗ trợ nhiều tính năng đặc sắc như các trình IDE khác.

1. **Chuẩn hóa dữ liệu**

Bộ dữ liệu của năm 2020, 2021 chưa có cột ‘Happiness Rank’ nên cần phải xác định lại dựa theo giá trị index của các hàng dữ liệu cộng thêm 1.

A screenshot of a computer code

Description automatically generated

Để tiện cho việc phân tích nên bộ dữ liệu của các năm cần thêm năm để khi gộp dữ liệu lại có thể xác định được dữ liệu nó thuộc năm nào, và các thuộc tính của các năm được đặt dưới các tên khác nhau cần chuẩn hóa lại để cho tiện việc xử lý dữ liệu tránh nhầm lẫn.

A white background with red and white text

Description automatically generated

Kiểm tra các giá trị rỗng trong bộ dữ liệu phát hiện năm 2018 có 1 trường bị rỗng và năm 2022 có 10 trường bị rỗng nên đã tiến hành xóa.

A screenshot of a computer code

Description automatically generated

Dữ liệu của năm 2022 các cột cần phân tích đều có giá trị string nên cần chuyển qua float để xử lý.

A screen shot of a computer code

Description automatically generated

Tổng hợp dữ liệu các năm để tiện việc xử lý và phân tích.

A computer code with many colorful text

Description automatically generated with medium confidence

# Chương 3: Đặc tả bài toán

## Các quốc gia có mức xếp hạng cao trong 8 năm

Mục đích tìm hiểu ở đây là dựa vào thông số Happiness Rank để tìm ra sự phân bố mức độ hạnh phúc của các quốc gia.

Cột dùng làm đối tượng là ‘Happiness Rank’, ‘Year’.

Lấy ra top 5 các quốc gia mỗi năm có chỉ số hạnh phúc cao nhất.

Code demo thực thi khai thác.



Kết quả nhận được

Từ năm 2015 đến năm 2022, quốc gia Finland nhiều lần giữa vị trí top 1 trong bảng xếp hạng, top 2 là Denmark, top 3 là Iceland, top 4 chiếm nhiều nhất cũng là Iceland cho thấy có vẻ đây là một quốc gia chiếm nhiều vị trí trong bảng xếp hạng, top 5 là Netherland.

Lấy ra độ dài record của mỗi năm để xem số lượng quốc gia được vào danh sách xếp hạng.

Code demo thực thi khai thác

A screenshot of a computer

Description automatically generated

Kết quả nhận được

A close up of numbers

Description automatically generated

Tạo biểu đồ cột top 5 các quốc gia có chỉ số hạnh phúc ‘Happiness Rank’ cao nhất mỗi năm từ năm 2015 đến năm 2022.

Code demo thực thi khai thác

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

Kết quả nhận được

A graph of different colored bars

Description automatically generated with medium confidence

A graph of different colored bars

Description automatically generated with medium confidence

A comparison of different colored bars

Description automatically generated

Sau khi phân tích dựa trên top 5 các quốc gia, ta sẽ xem khái quát rộng hơn là mức độ hạnh phúc ở từng khu vực.

## Phân tích các chỉ số theo khu vực trong năm 2022

### **Theo chỉ số hạnh phúc**

A computer code with text

Description automatically generated

A map of the world

Description automatically generated

Nhận xét: Theo biểu đồ ta thấy được chỉ số hạnh phúc của khu vực Bắc Mỹ, Châu Âu, và Châu Úc là khu vực có chỉ số cao nhất và nó luôn đồng đều trong 1 khu vực. Châu Á, Nam Mỹ chỉ số hạnh phúc nằm ở phần trung tiệm cận với các khu vực top đầu còn Châu Phi là nơi thấp nhất và có những nước không được xếp hạng về điểm hạnh phúc. Tuy nhiên có 1 khu vực ở Châu Á có chỉ số hạnh phúc thấp và đặc biệt có một quốc gia có diện tích thứ 7 trên thế giới nhưng chỉ số hạnh phúc lại rất thấp đó là Ấn Độ có một số yếu tố đã góp phần vào chỉ số hạnh phúc đang diễn ra trong nước. Chúng bao gồm quá trình đô thị hóa nhanh chóng và tình trạng quá tải ở các thành phố, lo ngại về sự an toàn của thực phẩm và nước, chi phí chăm sóc sức khỏe leo thang, sự an toàn của phụ nữ và vấn đề ô nhiễm môi trường, vốn gắn liền với tình trạng sức khỏe tinh thần ngày càng suy giảm.

### **Theo chỉ số GDP**

A computer code with text

Description automatically generated

A blue and white map

Description automatically generated

Nhận xét: Dựa theo việc biểu đồ phân hóa về chỉ số hạnh phúc ta cũng có thể nhận thấy được các nước có chỉ số hạnh phúc cao thì GDP luôn top cao nhất, nhưng có một đặc điểm là khu vực Nam Mỹ có chỉ số hạnh phúc nhỉnh hơn khu vực Châu Á nhưng về chỉ số GDP lại có phần kém hơn một ít. Khu vực Nam Á luôn là khu vực thấp vì các vấn đề về chính trị ảnh hướng tới đời sống kinh tế nơi đây.

### **Theo tuổi thọ**

A computer code with text

Description automatically generated

A map of the world

Description automatically generated

Nhận xét: Khu vực có chỉ số hạnh phúc cao và GDP cao thì tuổi thọ của khu vực đó cũng được cải thiện vì các khu vực đó được đầu tư cho vấn đề y tế và sức khỏe con người hơn, khu vực Châu Phi tuổi thọ của họ rất thấp vì kinh tế nên vấn đề y tế của họ còn chưa phát triển khi các bệnh thông thường có khả năng chưa được ở các nước có GDP cao thì ở các nước khu vực Châu Phi những bệnh này lại có khả năng tử vong rất cao.

Trong đại dịch Covid vừa qua, năm 2020 đến năm 2021 tuổi thọ trung bình bị giảm mạnh từ sấp sỉ 70% trải qua đại dịch tuổi thọ trung bình tuột xuống tới 52%, đến năm 2022 tình hình dịch được cải thiện và ổn định vì thế mà tuổi thọ trung bình đang trên đà tăng.

A close-up of a computer code

Description automatically generated

### **Theo chế độ phúc lợi xã hội**

A computer code with red and blue text

Description automatically generated

A map of the world

Description automatically generated

Nhận xét: Hầu hết các nước có điểm hạnh phúc cao và nền kinh tế ổn định sẽ có chế độ phúc lợi ổn định. Các nước có nền kinh tế kém phát triển thì chế độ phúc lợi xã hội càng ít, đặt biệt là Afghanistan 1 nước có tình hình chính trị bất ổn định nhất từ trước tới nay thì về chế độ phúc lợi dường như nằm ở con số 0.

### **Theo mức độ đời sống tinh thần tự do**

A computer screen shot of a code

Description automatically generated

A map of the world

Description automatically generated

Nhận xét: về mức độ đời sống tinh thần tự do thì bất ngờ Việt Nam ta nằm ở top các nước có đời sống tinh thần tự do cao nhất là 0.7 hơn cả Mỹ và các nước Châu Âu. Châu Phi và Nam Á luôn là các nước có chỉ số thấp, nhưng về chỉ số này Ấn Độ được xếp hạng rất cao tới 0.6.

### **Theo mức độ hào phóng**

A computer code with text

Description automatically generated

A map of the world

Description automatically generated

Nhận xét: đặc biệt thay các nước có chỉ số hạnh phúc cao thì chỉ số về độ hào phóng lại rất thấp như các nước ở khu vực Châu Âu, Châu Úc.

### **Theo mức độ nhận thức về tham nhũng**

A computer code with text

Description automatically generated

A map of the world

Description automatically generated

Nhận xét: Các nước ở khu vực Châu Úc, Châu Âu và Bắc Mỹ có nhận thức về tham nhũng rất cao nhưng các khu vực như Châu Á, Châu Phi, Nam Mỹ thì hoàn toàn ngược lại.

## Mối tương quan của các số liệu trong 8 năm

A screenshot of a computer

Description automatically generated

Nhận xét: Theo biểu đồ và các số liệu được phân tích trong năm 2022 ở trên, ta có thể thấy được các chỉ số ảnh hưởng to lớn đến điểm xếp hạng về hạnh phúc con người của các nước ít phụ thuộc vào sự hào phóng của họ, mà hầu hết sự ảnh hướng lớn đến xếp hạng là về GDP, sức khỏe, mức hỗ trợ xã hội, tự do trong cuộc sống và nhận thức về vấn đề tham nhũng trong xã hội.

## Hồi quy tuyến tính

Hồi quy tuyến tính cho ra mức độ phù hợp với dữ liệu thực tế là gần 0.72 trên 246 gái trị, và gần 0.74 trên 983 giá trị phù hợp với mô hình dữ liệu huấn luyện.

A screenshot of a computer program

Description automatically generated

## Phân cụm KMeans

A screenshot of a computer program

Description automatically generated

Qua đó, ta nhận được kết quả từ phương pháp Elbow

A graph with a line

Description automatically generated

Ta sẽ lấy 4 cụm.

A screenshot of a computer code

Description automatically generated

Đây là kết quả phân tán của 4 cụm.

Theo GDP và Happiness Rank

A screen shot of a graph

Description automatically generated

Theo Happiness Score và Happiness Rank

A screen shot of a computer code

Description automatically generated

A graph showing a line of colors

Description automatically generated with medium confidence

## Luật kết hợp

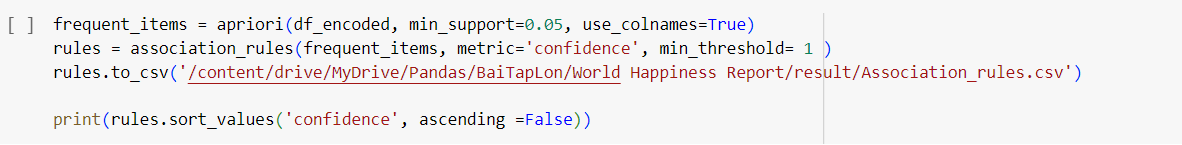
Qua thử nghiệm ta quy định các mức cho các cột của dữ liệu để tiện làm luật.

A screenshot of a computer

Description automatically generated

A close-up of a computer code

Description automatically generated



Kết quả:

A screenshot of a computer

Description automatically generated

Nhận xét: qua các luật ta thấy các mối liên hệ giữa chỉ số hạnh phúc, GDP, mức độ hỗ trợ xã hội và cuộc sống tự do sẽ luôn luôn kéo theo về vấn đề sức khỏe và tuổi thọ.

Những người hạnh phúc có nhiều khả năng chăm sóc sức khỏe của họ hơn và có lối sống lành mạnh hơn. Họ cũng có nhiều khả năng có mạng lưới xã hội mạnh mẽ, điều này có thể giúp họ vượt qua những căng thẳng trong cuộc sống. Những người sống trong các quốc gia có GDP cao tiếp cận được các dịch vụ chăm sóc sức khỏe chất lượng cao và có thể mua thực phẩm, chỗ ở và các nhu yếu phẩm khác cần thiết cho sức khỏe. Trong các quốc gia có mức độ hỗ trợ xã hội cao có được sự hỗ trợ tài chính và xã hội khi họ cần, điều này có thể giúp họ duy trì sức khỏe tốt. Những người sống trong các quốc gia có nhiều tự do quyền kiểm soát cuộc sống của họ hài lòng trong cuộc sống, điều này giúp họ có sức khỏe tinh thần tốt hơn.

## Phân lớp

Dựa theo dữ liệu ta sẽ quy định điểm số hạnh phúc từ 5 trở lên sẽ là đạt là quốc gia hạnh phúc, dưới 5 sẽ là không đạt.

A screenshot of a computer program

Description automatically generated

Chia dữ liệu để phân lớp tập test sẽ chứa 20% dữ liệu còn lại là tập train.

A white screen with black text

Description automatically generated

Kết quả:

A screenshot of a computer program

Description automatically generated

Nhận xét: Trong trường hợp này, f1-score cho các trường hợp đạt là 0,87, trong khi f1-score cho các trường hợp không đạt là 0,77. Điều này có nghĩa là mô hình có thể dự đoán chính xác 80% các trường hợp đạt và 77% các trường hợp không đạt.Mặc dù f1-score cho các trường hợp đạt cao hơn f1-score cho các trường hợp không đạt, nhưng tổng thể mô hình vẫn có thể được coi là tốt. Điều này là do f1-score cho các trường hợp đạt cao hơn 0,7, được coi là ngưỡng chấp nhận được cho hầu hết các nhiệm vụ phân loại.Tuy nhiên, mô hình có thể được cải thiện bằng cách tập trung vào việc cải thiện độ chính xác của các trường hợp không đạt. Điều này có thể được thực hiện bằng cách thu thập thêm dữ liệu cho các trường hợp không đạt hoặc bằng cách sử dụng thuật toán phân loại khác.

Ma trận nhầm lẫn cho thấy có 66 trường hợp của lớp "Không đạt" được dự đoán đúng là "Không đạt" và 19 trường hợp bị dự đoán sai thành "Đạt". Còn đối với lớp "Đạt", có 140 trường hợp dự đoán đúng là "Đạt" và 21 trường hợp bị dự đoán sai thành "Không đạt".

## Cây quyết định

A screenshot of a computer

Description automatically generated

A computer screen shot of a map

Description automatically generated

A screenshot of a computer

Description automatically generated

Nhận xét: F1-score đối với lớp “Không đạt” là 0.86, còn đối với lớp “Đạt” là 0.93. Độ chính xác của lớp “Không đạt” là 0.87 còn “Đạt” là 0.92 cho thấy mức độ phù hợp của cây quyết định ở nhiệm vụ phân loại là khá cao và hoàn toàn có thể dùng mô hình này phân loại các dữ liệu khác trong tương lai.

Ma trận nhầm lẫn cho thấy có 72 trường hợp của lớp "Không đạt" được dự đoán đúng là "Không đạt" và 13 trường hợp bị dự đoán sai thành "Đạt". Còn đối với lớp "Đạt", có 150 trường hợp dự đoán đúng là "Đạt" và 11 trường hợp bị dự đoán sai thành "Không đạt".

# Chương 4: Kết luận

Trong báo cáo này, chúng ta đã thực hiện một phân tích toàn diện về dữ liệu từ Báo cáo Hạnh phúc thế giới từ năm 2015 đến năm 2022 để hiểu sâu hơn về yếu tố ảnh hưởng đến mức độ hạnh phúc của các quốc gia trên khắp thế giới. Qua việc áp dụng các công cụ và phương pháp phân tích, chúng ta đã rút ra những kết luận quan trọng về sự tương quan giữa các yếu tố kinh tế, xã hội và tâm lý trong việc xác định mức độ hạnh phúc của mỗi quốc gia:

1. **Tương quan với GDP:** Mối liên hệ mạnh mẽ giữa GDP và mức độ hạnh phúc đã được thể hiện. Các quốc gia có GDP cao thường có mức độ hạnh phúc tốt hơn, tuy nhiên, điều này cũng cần được kết hợp với các yếu tố xã hội và tâm lý khác để tạo nên một môi trường hạnh phúc toàn diện.
2. **Tầm quan trọng của sức khỏe và tuổi thọ:** Chất lượng dịch vụ y tế và tuổi thọ cao đóng vai trò quan trọng trong việc định hình mức độ hạnh phúc của một quốc gia. Đây là minh chứng cho việc đầu tư vào lĩnh vực sức khỏe có tác động tích cực đến chất lượng cuộc sống và tâm trạng của người dân.
3. **Tự do cá nhân và quyền lựa chọn:** Tự do cá nhân cũng được xác định là một yếu tố quan trọng trong việc định hình hạnh phúc. Các quốc gia tạo điều kiện cho sự tự do và quyền lựa chọn cá nhân thường có mức độ hạnh phúc cao hơn, điều này thể hiện tầm quan trọng của quyền tự do và sự linh hoạt trong cuộc sống.
4. **Tác động của vấn đề tiêu cực:** Tham nhũng và các yếu tố tiêu cực khác có khả năng ảnh hưởng tiêu cực đến mức độ hạnh phúc của một quốc gia. Những vấn đề này không chỉ ảnh hưởng đến khía cạnh kinh tế mà còn tạo ra môi trường không thuận lợi cho sự phát triển và hạnh phúc của người dân.
5. **Tác động của đại dịch Covid-19:** Dịch bệnh đã thể hiện sự ảnh hưởng đáng kể đến mức độ hạnh phúc của các quốc gia, với sự giảm sút về tuổi thọ trung bình và mức độ hạnh phúc trong giai đoạn khủng hoảng. Điều này là minh chứng cho tầm quan trọng của sức khỏe và ổn định xã hội trong việc duy trì hạnh phúc.

Tóm lại, những phân tích chi tiết và những mối quan hệ tương quan trong dữ liệu từ Báo cáo Hạnh phúc thế giới từ năm 2015 đến năm 2022 đã cung cấp một cái nhìn sâu sắc về những yếu tố ảnh hưởng đến mức độ hạnh phúc của các quốc gia. Hiểu rõ hơn về mối quan hệ giữa các yếu tố kinh tế, xã hội và tâm lý là cơ sở quan trọng để đưa ra các chính sách và quyết định có mục tiêu cải thiện chất lượng cuộc sống và hạnh phúc của con người trên toàn cầu.

**HẾT**