**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

An orange circle with white text and a map

Description automatically generated

**BÁO CÁO BÀI TẬP LỚN**

**ĐỀ TÀI: ỨNG DỤNG KHAI PHÁ VÀ PHÂN TÍCH DỮ LIỆU ĐỂ DỰ BÁO CHỈ SỐ HẠNH PHÚC CỦA VIỆT NAM TRONG TƯƠNG LAI**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn** | : Nguyễn Dương Hùng |
| **Học phần** | : Khai phá và phân tích dữ liệu |
| **Lớp** | : 231IS23A01 |
| **Nhóm** | : 9 |

**Hà Nội – 2023**

**HỌC VIỆN NGÂN HÀNG**

**KHOA HỆ THỐNG THÔNG TIN QUẢN LÝ**

An orange circle with white text and a map

Description automatically generated

**BÁO CÁO BÀI TẬP LỚN**

**ĐỀ TÀI: ỨNG DỤNG KHAI PHÁ VÀ PHÂN TÍCH DỮ LIỆU ĐỂ DỰ BÁO CHỈ SỐ HẠNH PHÚC CỦA VIỆT NAM TRONG TƯƠNG LAI**

|  |  |  |  |
| --- | --- | --- | --- |
| **Thành viên** | **Mã sinh viên** | **Tỉ lệ đóng góp** | **Xác nhận** |
| Nguyễn Duy Hưng  ( Nhóm trưởng) | 24A4042596 | 21% |  |
| Hà Trọng Cường |  | 20% |  |
| Trần A Khang | 24A4040141 | 17% |  |
| Đinh Thành Nam | 24A4042599 | 25% |  |
| Trần Quốc Việt |  | 17% |  |

**Hà Nội – 2023**

**BẢNG PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Họ và tên** | **Nhiệm vụ** |
| 1 | Nguyễn Duy Hưng  ( Nhóm trưởng) | Lên ý tưởng đề tài và đóng góp thuật toán |
| 2 | Hà Trọng Cường | Chương III và đóng góp thuật toán |
| 3 | Trần A Khang | Chương IV |
| 4 | Đinh Thành Nam | Word + PowerPoint và thực hiện thuật toán |
| 5 | Trần Quốc Việt | Chương I + chương II |

**LỜI NÓI ĐẦU**

Bài tập lớn "Ứng dụng khai phá và phân tích dữ liệu để dự báo chỉ số hạnh phúc của Việt Nam trong tương lai" là kết quả của sự cống hiến và phối hợp đồng đội của tất cả thành viên trong nhóm 9. Báo cáo không chỉ đánh dấu sự hoàn thiện của dự án mà còn là kỷ niệm đáng trân trọng cho những nỗ lực của chúng em trong suốt quá trình thực hiện. Trong quá trình này, chúng em nhận được sự hướng dẫn và hỗ trợ nhiệt tình từ các giảng viên và các bạn trong khoa Hệ Thống Thông tin - Học Viện Ngân Hàng. Chúng em xin gửi lời cảm ơn chân thành và sâu sắc đến các giảng viên đã dành thời gian và công sức để truyền đạt những kiến thức quan trọng cho chúng em, giúp chúng em tích lũy đủ kiến thức và kỹ năng để hoàn thành bài tập lớn này.

Đặc biệt, chúng em muốn bày tỏ lòng biết ơn đến thầy Nguyễn Dương Hùng, người đã hỗ trợ chúng em trong từng bước hoàn thiện của dự án. Thầy không chỉ là người hướng dẫn tận tâm, mà còn giúp đỡ và hỗ trợ chúng em suốt quá trình làm việc. Trong quá trình học và làm việc cùng thầy, chúng em đã học được thái độ nghiêm túc trong công việc, cũng như sự phối hợp hiệu quả để hoàn thành công việc được tốt nhất.

Mặc dù chúng em đã cố gắng hoàn thiện bài tập trong phạm vi và khả năng của mình, cũng không thể tránh khỏi những thiếu sót. Chúng em mong nhận được sự thông cảm và góp ý từ các giảng viên và các bạn để chúng em có thể hoàn thiện và cải thiện dự án của mình.

Cuối cùng, chúng em xin gửi lời chúc sức khỏe và thành công đến các giảng viên, luôn là người dẫn dắt thế hệ sinh viên của Học Viện Ngân Hàng trưởng thành và trở thành những nhân tố quan trọng trong thị trường lao động. Em xin chân thành cảm ơn!

**LỜI CAM ĐOAN**

Nhóm 9 xin cam đoan báo cáo này là sản phẩm nghiên cứu và thực hành của nhóm. Bài báo cáo đảm bảo tính liêm chính trong học tập, không đạo văn, gian lận, bịa đặt. Các thông tin tham khảo được trích dẫn nguồn đầy đủ và minh bạch.

Nhóm 9 xin chịu toàn bộ trách nhiệm nếu bài báo cáo vi phạm các điều trên.

Hà Nội, ngày 22 tháng 12 năm 2023

Đại diện nhóm 9

Nguyễn Duy Hưng

**MỤC LỤC**

[**CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI 1**](#_Toc154093015)

[**1. Đặt vấn đề 1**](#_Toc154093016)

[**2. Cơ sở hình thành đề tài 1**](#_Toc154093017)

[**3. Mục tiêu đề tài 2**](#_Toc154093018)

[**4. Đối tượng và phương pháp nghiên cứu 2**](#_Toc154093019)

[**5. Ý nghĩa của đề tài 2**](#_Toc154093020)

[**6. Bố cục đề tài 2**](#_Toc154093021)

[**CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 3**](#_Toc154093022)

[**1. Tổng quan kỹ thuật khai phá dữ liệu 3**](#_Toc154093023)

[**1.1. Khái niệm khai phá dữ liệu 3**](#_Toc154093024)

[**1.2. Các giai đoạn của quá trình khai phá dữ liệu 3**](#_Toc154093025)

[**1.3. Các phương pháp khai phá dữ liệu 4**](#_Toc154093026)

[**2. Bài toán hồi quy trong khai phá dữ liệu 5**](#_Toc154093027)

[**2.1. Khái niệm hồi quy 5**](#_Toc154093028)

[**2.2. Ứng dụng của hồi quy 6**](#_Toc154093029)

[**3. Cơ sở lý thuyết về chỉ số hạnh phúc 6**](#_Toc154093030)

[**3.1. Tổng quan về chỉ số hạnh phúc 6**](#_Toc154093031)

[**3.2. Ứng dụng của khai phá dữ liệu trong dự đoán chỉ số hạnh phúc 7**](#_Toc154093032)

[**CHƯƠNG 3: XÂY DỰNG MÔ HÌNH DỰ BÁO 8**](#_Toc154093033)

[**1. Thu thập dữ liệu 8**](#_Toc154093034)

[**2. Xác định kiểu dữ liệu của thuộc tính. 9**](#_Toc154093035)

[**3. Lựa chọn thuật toán sử dụng 10**](#_Toc154093036)

[**3.1. Lý do lựa chọn thuật toán 10**](#_Toc154093037)

[**3.2. Tổng quan về các thuật toán 11**](#_Toc154093038)

[**CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ 15**](#_Toc154093039)

[**4.1. Thực nghiệm 15**](#_Toc154093040)

[**4.1.1. Chuẩn bị dữ liệu 15**](#_Toc154093041)

[**4.1.2. Sử dụng mô hình ARIMA để dự đoán cho từng biến độc lập và mô hình Linear Regression để dự đoán chỉ số hạnh phúc của Việt Nam trong năm 2024 16**](#_Toc154093042)

[**4.1.3. Sử dụng mô hình SARIMAX để dự đoán cho từng biến độc lập và mô hình Linear Regression để dự đoán chỉ số hạnh phúc của Việt Nam trong năm 2024 18**](#_Toc154093043)

[**4.1.4. Sử dụng mô hình Random Forest Regressor để dự đoán cho chỉ số hạnh phúc của Việt Nam. 21**](#_Toc154093044)

[**4.2. Đánh giá 22**](#_Toc154093045)

[**KẾT LUẬN 24**](#_Toc154093046)

[**TÀI LIỆU THAM KHẢO 25**](#_Toc154093047)

**DANH MỤC HÌNH ẢNH**

[Hình 1. Các giai đoạn tiền xử lý dữ liệu 4](#_Toc154100068)

[Hình 2. Bộ dữ liệu ban đầu 8](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100069)

[Hình 3. Đồ thị ACF và PACF để xác định giá trị p, d, q, P, D, Q 16](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100070)

[Hình 4. Bộ dữ liệu chỉ còn lại Việt Nam 19](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100071)

[Hình 5. Kết quả dự đoán các biến độc lập bằng ARIMA 20](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100072)

[Hình 6. Kết quả dự đoán Life Ladder bằng Linear Regression dựa trên ARIMA 21](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100073)

[Hình 7. Kết quả đánh giá mô hình Linear Regression 21](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100074)

[Hình 8. Kết quả dự đoán các biến độc lập bằng SARIMAX 23](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100075)

[Hình 9. Kết quả dự đoán Life Ladder bằng Linear Regression dựa trên SARIMAX 23](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100076)

[Hình 10. Kết quả dự đoán Life Ladder bằng Random Forest Regressor dựa trên SARIMAX 24](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100077)

[Hình 11. Kết quả đánh giá mô hình Random Forest Regressor 25](file:///C:\Users\PV\Downloads\Chương-1-2-3.docx#_Toc154100078)

# CHƯƠNG 1: TỔNG QUAN VỀ ĐỀ TÀI

## Đặt vấn đề

Trong những thập kỷ gần đây, công nghệ thông tin (CNTT) đã có những bước phát triển vượt bậc, tác động sâu sắc đến mọi mặt của đời sống xã hội. Một trong những lĩnh vực phát triển nhanh chóng nhất trong CNTT là khai phá dữ liệu. Sự phát triển của khai phá dữ liệu sẽ tiếp tục được đẩy mạnh trong tương lai. Với sự phát triển của công nghệ, chúng ta sẽ có thể thu thập và lưu trữ một lượng dữ liệu khổng lồ. Khai phá dữ liệu và máy học sẽ là những công cụ quan trọng giúp chúng ta phân tích và hiểu được những dữ liệu này, từ đó đưa ra những quyết định sáng suốt hơn. Kỹ năng khai phá dữ liệu đang được nghiên cứu tại Việt Nam, kỹ thuật này còn tương đối mới mẻ tuy nhiên cũng đã được đưa vào ứng dụng trong nhiều lĩnh vực khác nhau, đem lại nguồn tri thức vô tận và giúp con người nhiều hơn trong quá trình phát triển kinh tế - xã hội.

Chỉ số hạnh phúc là một thước đo tổng quát về mức độ hạnh phúc của một cá nhân, cộng đồng hoặc quốc gia. Chỉ số hạnh phúc có thể được sử dụng để đánh giá hiệu quả của các chính sách, chương trình phát triển kinh tế - xã hội. Khai phá dữ liệu có thể được sử dụng để dự đoán chỉ số hạnh phúc. Thông qua việc phân tích dữ liệu từ các nguồn khác nhau, khai phá dữ liệu có thể giúp phát hiện các yếu tố ảnh hưởng đến hạnh phúc của con người. Từ đó, các chính sách, chương trình phát triển kinh tế - xã hội có thể được xây dựng nhằm nâng cao chỉ số hạnh phúc của người dân.

## Cơ sở hình thành đề tài

Việt Nam là một quốc gia đang phát triển với tốc độ nhanh chóng. Trong những năm gần đây, Việt Nam đã đạt được những thành tựu đáng kể về kinh tế - xã hội, đời sống của người dân được cải thiện rõ rệt. Tuy nhiên, vẫn còn nhiều vấn đề cần được giải quyết để nâng cao chất lượng cuộc sống của người dân. Việc nghiên cứu và dự đoán chỉ số hạnh phúc tại Việt Nam trong tương lai là một vấn đề quan trọng, giúp các nhà hoạch định chính sách có những giải pháp phù hợp để nâng cao chất lượng cuộc sống của người dân.. Từ đó ứng dụng thuật toán dự đoán chỉ số hạnh phúc trong tương lai sẽ giúp chúng ta trong việc biết nên cải thiện những gì và hạn chế những gì. Trong bài tập lớn này, chúng tôi tìm hiểu và trình bày về kĩ thuật trong khai phá dữ liệu để ứng dụng vào đề tài. Dự đoán chỉ số hạnh phúc của Việt Nam trong tương lai là một vấn đề quan trọng, có ý nghĩa thực tiễn và khoa học. Việc ứng dụng khai phá dữ liệu trong dự đoán chỉ số hạnh phúc là một hướng nghiên cứu mới mẻ và có tiềm năng ứng dụng rất cao.

## Mục tiêu đề tài

Dự đoán chỉ số hạnh phúc của Việt Nam trong tương lai là một vấn đề quan trọng, có ý nghĩa thực tiễn và khoa học. Việc ứng dụng khai phá dữ liệu trong dự đoán chỉ số hạnh phúc là một hướng nghiên cứu mới mẻ và có tiềm năng ứng dụng rất cao. Mục tiêu đề tài là đưa ra các dự đoán về chỉ số hạnh phúc, từ đó phân tích và đưa ra các giải pháp giúp cải thiện về chỉ số hạnh phúc của Việt Nam.

## Đối tượng và phương pháp nghiên cứu

* **Đối tượng nghiên cứu của đề tài:** Chỉ số hạnh phúc của Việt Nam với các nước châu Á và trên thế giới từ năm 2005 đến 2023.
* **Phương pháp nghiên cứu:** Sử dụng thuật toán Linear Regression (Hồi quy tuyến tính) và các thuật toán khác có liên quan để đưa ra dự đoán.

## Ý nghĩa của đề tài

Dựa vào tập cơ sở dữ liệu về chỉ số hạnh phúc của Việt Nam với các nước châu Á và trên thế giới từ năm 2005 đến 2023, sử dụng thuật toán Linear Regression và các thuật toán khác có liên quan để đưa ra dự đoán về chỉ số hạnh phúc của Việt Nam trong tương lai.

## Bố cục đề tài

Đề tài gồm có 5 phần:

* Chương 1: Tổng quan về đề tài
* Chương 2: Cơ sở lý thuyết
* Chương 3: Xây dựng mô hình dự đoán
* Chương 4: Thực nghiệm và đánh giá
* Chương 5: Kết luận

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Tổng quan kỹ thuật khai phá dữ liệu

### Khái niệm khai phá dữ liệu

Khai phá dữ liệu là quá trình tính toán để tìm ra các mẫu và thông tin trong các bộ dữ liệu lớn. Mục tiêu tổng thể của quá trình khai thác dữ liệu là trích xuất thông tin từ một bộ dữ liệu và chuyển nó thành một cấu trúc dễ hiểu để sử dụng tiếp. Khai phá dữ liệu là một trong các bước quan trọng nhất trong toàn bộ quá trình khám phá tri thức (KDD), triển khai các thuật toán chuyên dụng để chiết xuất ra các mẫu cũng như các mô hình có ích từ dữ liệu. Trong bối cảnh cạnh tranh khốc liệt hiện nay, các tổ chức/doanh nghiệp luôn tự ý thức được rằng cần phải đạt được một lợi thế cạnh tranh tốt nhất có thể, trong đó việc tìm ra và sử dụng các tri thức thông tin hữu ích ẩn chứa trong dữ liệu có thể trở thành một công cụ cạnh tranh cấp chiến lược. Có nhiều tham số quan trọng khác nhau trong Data Mining, chẳng hạn như quy tắc kết hợp, phân loại, phân cụm và dự báo.

**Một số tính năng chính của Data Mining:**

* Dự đoán các mẫu dựa vào xu hướng trong hàng loạt dữ liệu.
* Tính toán, xử lý kết quả đã được phân tích.
* Đưa ra các thông tin phản hồi để phân tích.
* Tập trung phân tích, xử lý dữ liệu lớn hơn.
* Phân chia, sắp xếp các cụm dữ liệu trực quan.

### Các giai đoạn của quá trình khai phá dữ liệu

Khai phá dữ liệu không hẳn là một hệ thống hoàn toàn tự động mà con người cần phải tương tác thường xuyên với Database thông qua các thuật toán chuyên biệt. Quy trình khai phá dữ liệu là một chuỗi lặp gồm các giai đoạn bắt đầu từ dữ liệu thô (raw data) và kết thúc với tri thức (knowledge of interest):

- Làm sạch dữ liệu: Trước tiên cần phải làm sạch tất cả dữ liệu quy trình để dữ liệu đó phù hợp với tiêu chuẩn ngành. Dữ liệu bẩn hay không đầy đủ dẫn đến hiểu biết kém và hệ thống bị lỗi gây tốn thời gian và tiền bạc.

- Tích hợp dữ liệu: Đây là một trong những kĩ thuật khai thác hàng đầu để hợp lý hóa toàn bộ quá trình trích xuất, chuyển đổi và tải. Nhiều chuyên gia thực hiện dọn dẹp dữ liệu bổ sung trong các CSDL khác nhau trong giai đoạn này.

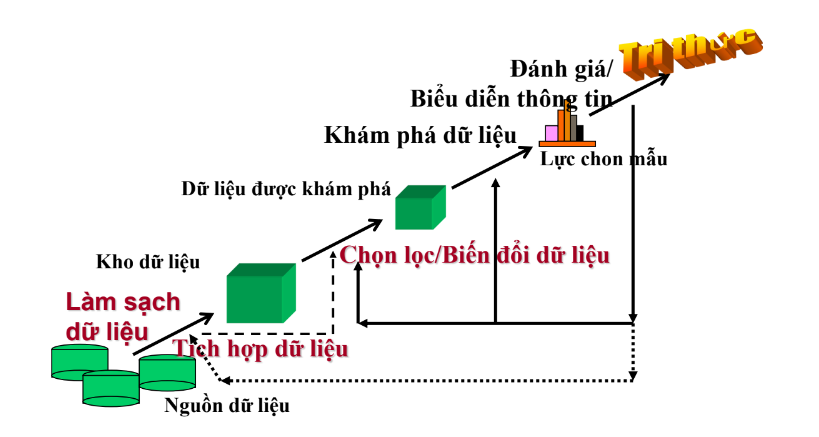
- Chọn lựa dữ liệu: Trong bước này, dữ liệu được trích xuất từ cơ sở dữ liệu.

- Biến đổi dữ liệu: Chuyển đổi dữ liệu sang dạng có thể chấp nhận được để phù hợp với các mục tiêu khai thác. Hợp nhất dữ liệu chuẩn bị để tối ưu hóa quy trình Data mining và giúp dễ dàng phân biệt các mấu trong tập dữ liệu cuối cùng.

- Khai phá dữ liệu: Sử dụng các ứng dụng Data mining để trích xuất các xu hướng hữu ích và tối ưu hóa việc khám phá kiến thức để tạo ra thông tin kinh doanh.

- Đánh giá mẫu: Giai đoạn này sẽ đưa những hiểu biết sâu sắc vào thế giới thực, bên cạnh đó còn xác định bất kì mẫu hữu ích nào có thể tạo ra kiến thức kinh doanh.

- Biểu diễn tri thức: Cuối cùng sẽ sử dụng kết hợp trực quan hóa dữ liệu, báo cáo và các công cụ khai thác để chia sẻ thông tin với những người khác.



Hình 1. Các giai đoạn tiền xử lý dữ liệu

### Các phương pháp khai phá dữ liệu

* Phân loại (Classification): Là phương pháp dự báo, cho phép phân loại một đối tượng vào một hoặc một số lớp cho trước.
* Hồi quy (Regression): Khám phá chức năng học dự đoán, ánh xạ một mục dữ liệu thành biến dự đoán giá trị thực.
* Phân nhóm (Clustering): Một nhiệm vụ mô tả phổ biến trong đó người ta tìm cách xác định một tập hợp hữu hạn các cụm để mô tả dữ liệu.
* Tổng hợp (Summarization): Một nhiệm vụ mô tả bổ sung liên quan đến phương pháp cho việc tìm kiếm một mô tả nhỏ gọn cho một bộ (hoặc tập hợp con) của dữ liệu.
* Mô hình ràng buộc (Dependency modeling): Tìm mô hình cục bộ mô tả các phụ thuộc đáng kể giữa các biến hoặc giữa các giá trị của một tính năng trong tập dữ liệu hoặc trong một phần của tập dữ liệu.
* Dò tìm biến đổi và độ lệch (Change and Deviation Detection): Khám phá những thay đổi quan trọng nhất trong bộ dữ liệu.

## Bài toán hồi quy trong khai phá dữ liệu

### Khái niệm hồi quy

Hồi quy là một loại kỹ thuật học máy có giám sát được sử dụng để dự đoán bất kỳ thuộc tính có giá trị liên tục nào. Kỹ thuật này giúp bất kỳ tổ chức kinh doanh nào phân tích mối quan hệ của biến mục tiêu và biến dự báo. Nó là một công cụ quan trọng nhất để phân tích dữ liệu có thể được sử dụng cho dự báo tài chính và mô hình chuỗi thời gian.

Hồi quy liên quan đến kỹ thuật nối một đường thẳng hoặc một đường cong trên nhiều điểm dữ liệu. Nó xảy ra theo cách mà khoảng cách giữa các điểm dữ liệu và điểm chữa khỏi là thấp nhất. Các loại hồi quy phổ biến nhất là hồi quy tuyến tính (Linear Regression) và logistic. Ngoài ra, nhiều loại hồi quy khác có thể được thực hiện tùy thuộc vào hiệu suất của chúng trên một tập dữ liệu riêng lẻ.

Hồi quy có thể dự đoán tất cả các tập dữ liệu phụ thuộc, được thể hiện bằng biểu thức của các biến độc lập và xu hướng có sẵn trong một khoảng thời gian hữu hạn. Nó cung cấp một cách tốt để dự đoán các biến, nhưng có một số hạn chế và giả định nhất định như tính độc lập của các biến, phân phối chuẩn vốn có của các biến. Trước khi áp dụng phân tích hồi quy, dữ liệu cần được nghiên cứu cẩn thận và thực hiện một số thử nghiệm sơ bộ nhất định để đảm bảo quy tắc có thể áp dụng được. Có những bài kiểm tra không tham số có sẵn trong những trường hợp như vậy.

**Regression được chia thành năm loại khác nhau:**

* Linear Regression
* Logistic Regression
* Lasso Regression
* Ridge Regression
* Polynomial Regression

### Ứng dụng của hồi quy

Hồi quy là một kỹ thuật rất phổ biến, và nó có ứng dụng rộng rãi trong các doanh nghiệp và ngành công nghiệp. Quy trình hồi quy liên quan đến biến dự đoán và biến phản hồi. Các ứng dụng chính của hồi quy bao gồm:

* Mô hình môi trường
* Phân tích hành vi kinh doanh và tiếp thị
* Dự báo tài chính hoặc dự báo
* Phân tích các xu hướng và mô hình mới

## Cơ sở lý thuyết về chỉ số hạnh phúc

### Tổng quan về chỉ số hạnh phúc

Chỉ số hạnh phúc là một thước đo tổng thể về hạnh phúc, phúc lợi và các khía cạnh của tính bền vững và khả năng phục hồi. Chỉ số này được tính toán dựa trên dữ liệu khảo sát, bao gồm các yếu tố như sự hài lòng trong cuộc sống, cảm giác hạnh phúc, sức khỏe tâm lý, sức khỏe, cân bằng thời gian, cộng đồng, hỗ trợ xã hội, giáo dục, nghệ thuật và văn hóa, môi trường, quản trị, phúc lợi vật chất và công việc.

Chỉ số hạnh phúc có thể được sử dụng cho nhiều mục đích khác nhau, bao gồm:

* Đánh giá mức độ hạnh phúc của một cá nhân, nhóm người hoặc quốc gia
* Xác định các yếu tố ảnh hưởng đến hạnh phúc
* Đo lường sự tiến bộ của một quốc gia về mặt hạnh phúc
* So sánh mức độ hạnh phúc của các quốc gia khác nhau

Chỉ số hạnh phúc là một công cụ quan trọng để hiểu rõ hơn về hạnh phúc và phúc lợi của con người. Nó có thể được sử dụng để thúc đẩy các chính sách và chương trình nhằm cải thiện chất lượng cuộc sống của mọi người.

### Ứng dụng của khai phá dữ liệu trong dự đoán chỉ số hạnh phúc

Khai phá dữ liệu là một quá trình tìm kiếm các mẫu, xu hướng và mối quan hệ tiềm ẩn trong dữ liệu. Khai phá dữ liệu có thể được sử dụng trong nhiều lĩnh vực khác nhau, bao gồm cả dự đoán chỉ số hạnh phúc. Khai phá dữ liệu có thể được sử dụng để dự đoán chỉ số hạnh phúc bằng cách phân tích dữ liệu khảo sát về các yếu tố ảnh hưởng đến hạnh phúc. Các kỹ thuật khai phá dữ liệu có thể được sử dụng để tìm kiếm các mối quan hệ giữa các yếu tố này và chỉ số hạnh phúc.

Khai phá dữ liệu có thể được sử dụng để dự đoán chỉ số hạnh phúc ở các cấp độ khác nhau, bao gồm cá nhân, nhóm người và quốc gia. Đặc biệt là ở cấp độ quốc gia, khai phá dữ liệu có thể được sử dụng để dự đoán chỉ số hạnh phúc của một quốc gia dựa trên các yếu tố như:

* Thông tin kinh tế: GDP bình quân đầu người, tỷ lệ thất nghiệp,...
* Thông tin xã hội: tỷ lệ tội phạm, tỷ lệ ly hôn,...
* Thông tin môi trường: chất lượng không khí, chất lượng nước,...

Tóm lại, khai phá dữ liệu là một công cụ mạnh mẽ có thể được sử dụng để dự đoán chỉ số hạnh phúc ở các cấp độ khác nhau. Khai phá dữ liệu có thể giúp chúng ta hiểu rõ hơn về các yếu tố ảnh hưởng đến chỉ số hạnh phúc của Việt Nam.

# CHƯƠNG 3: XÂY DỰNG MÔ HÌNH DỰ BÁO

## Thu thập dữ liệu

Dữ liệu là một yếu tố quan trọng không thể thiếu trong tất cả các bài toán Machine Learning. Với bài toán dự báo chỉ số hạnh phúc dựa trên thông số cho trước, chúng em đã tìm kiếm và tải về bộ dữ liệu tại trang Kaggle để phục vụ cho việc tiến hành thu thập dữ liệu và khởi chạy chương trình.

Bộ dữ liệu của bài toán bao gồm 2220 bản ghi và 10 thuộc tính cùng với rất nhiều các thông tin đã được thu thập trên thực tế từ năm 2005 đến năm 2023. Với bộ dữ liệu này thì chúng em có thể tiến hành việc khai phá dữ liệu dựa trên các thuật toán khác nhau tùy thuộc vào từng tình huống.

A screenshot of a computer

Description automatically generated

Hình 2. Bộ dữ liệu ban đầu

Dữ liệu cần thu thập là những thông tin về tên quốc gia, năm thống kê dữ liệu, các chỉ số liên quan nhằm phục vụ cho bài toán dự báo chỉ số hạnh phúc, cụ thể là:

|  |  |
| --- | --- |
| **Thuộc tính** | **Mô tả** |
| Country Name | Tên quốc gia cụ thể từ dữ liệu được thu thập. Mỗi quốc gia có các đặc điểm văn hóa, kinh tế và xã hội riêng biệt ảnh hưởng đến chỉ số hạnh phúc. |
| Regional Indicator | Phân loại vùng miền địa lý mà quốc gia thuộc về. Nó phản ánh sự tương đồng giữa các khu vực xã hội, kinh tế hoặc văn hóa. |
| Year | Năm thống kê chỉ số hạnh phúc. Theo dõi thay đổi của chỉ số qua các năm có thể giúp phân tích xu hướng và sự thay đổi trong hạnh phúc của một quốc gia theo thời gian. |
| Life Ladder | Chỉ số đo lường mức độ hạnh phúc với cuộc sống của người dân trong một quốc gia. |
| Logged GDP Per Capita | Chỉ số GDP trên đầu người được tính bằng cách lấy logarit tự nhiên của GDP trên đầu người. |
| Social Support | Chỉ số đo lường mức độ hỗ trợ xã hội mà người dân nhận được từ gia đình, bạn bè, cộng đồng trong một quốc gia. |
| Healthy Life Expectancy At Birth | Tuổi thọ trung bình mà một đứa trẻ được dự đoán sẽ sống trong điều kiện khỏe mạnh tại thời điểm sinh trong một quốc gia. Nó cho thấy chất lượng cuộc sống và sức khỏe của người dân. |
| Freedom To Make Life Choices | Chỉ số đo lường mức độ tự do của người dân trong việc lựa chọn cách sống của mình, như là quyền lựa chọn nghề nghiệp, hôn nhân, và các quyết định cá nhân khác. |
| Generosity | Chỉ số đo lường mức độ rộng lượng của người dân trong một quốc gia. Nó phản ánh tầm ảnh hưởng của tình cảm xã hội đến mối quan hệ xã hội. |
| Perceptions Of Corruption | Chỉ số đo lường mức độ tham nhũng trong một quốc gia. Mức độ này có thể ảnh hưởng đến niềm tin và sự tin tưởng vào các cơ quan chính phủ và xã hội, ảnh hưởng đến hạnh phúc chung của người dân. |

## Xác định kiểu dữ liệu của thuộc tính.

|  |  |  |
| --- | --- | --- |
|  | **Thuộc tính** | **Kiểu dữ liệu** |
|  | Tên quốc gia (Country Name) | object |
|  | Phân vùng địa lý (Regional Indicator) | object |
|  | Năm thống kê (Year) | int64 |
| Biến phụ thuộc (Target) | Chỉ số hạnh phúc (Life Ladder) | float64 |
| Biến độc lập (Feature) | Chỉ số GDP trên đầu người | float64 |
| Chỉ số hỗ trợ xã hội (Social Support) | float64 |
| Tuổi thọ trung bình (Healthy Life Expectancy At Birth) | float64 |
| Chỉ số tự do lựa chọn cách sống (Freedom To Make Life Choices) | float64 |
| Chỉ số mức độ rộng lượng (Generosity) | float64 |
| Chỉ độ tham nhũng (Perceptions Of Corruption) | float64 |

## Lựa chọn thuật toán sử dụng

### Lý do lựa chọn thuật toán

Để xây dựng mô hình dự báo có thể đem lại kết quả chính xác, việc áp dụng các phương pháp và thuật toán phù hợp là rất quan trọng. Vì vậy chúng em đã xem xét việc sử dụng các thuật toán Linear Regression, Random Forest, mô hình ARIMA để dự đoán chỉ số hạnh phúc của Việt Nam. Mỗi thuật toán đều có ưu, nhược điểm và đặc trưng riêng, và việc chọn lựa phù hợp giữa chúng sẽ được đánh giá dựa trên các thuộc tính GDP, chỉ số hỗ trợ xã hội, tuổi thọ trung bình và nhiều yếu tố khác.

Về Linear Regression, thuật toán này thường được sử dụng để xác định mối quan hệ tuyến tính giữa các biến đầu vào và đầu ra. Việc sử dụng thuật toán này có thể phản ánh mối quan hệ tuyến tính giữa các yếu tố như GDP, chỉ số hỗ trợ xã hội, tuổi thọ khỏe mạnh với chỉ số hạnh phúc.

Về Random Forest, đối với các biến độc lập phức tạp và có mối quan hệ phi tuyến tính, 2 thuật toán này có thể sẽ rất phù hợp. Các thuộc tính như GDP, chỉ số hỗ trợ xã hội, tuổi thọ trung bình có thể không có mối quan hệ tuyến tính hoàn hảo.

Về mô hình ARIMA, dữ liệu về chỉ số hạnh phúc có xu hướng thay đổi theo thời gian (time series), vì vậy sử dụng mô hình này sẽ rất thích hợp. Nó có thể phân tích và dự đoán xu hướng thay đổi của chỉ số hạnh phúc theo thời gian.

### Tổng quan về các thuật toán

#### **Linear Regression**

Linear Regression là một thuật toán học có giám sát (supervised learning) trong Machine Learning, nó là một phương pháp thống kê dùng để ước lượng mối quan hệ giữa các biến độc lập (input features) và biến phụ thuộc (output target). Linear Regression giả định rằng sự tương quan giữa các biến là tuyến tính, từ đó tìm ra hàm tuyến tính tốt nhất để biểu diễn mối quan hệ này. Thuật toán này dự báo giá trị của biến output từ các giá trị của các biến đầu vào.

Linear Regression có hai loại chính:

- Simple Linear Regression: Mô hình này chỉ có một biến độc lập (input feature) mô tả mối quan hệ tuyến tính giữa biến phụ thuộc (output target) và biến độc lập.

- Multiple Linear Regression: Mô hình này có nhiều hơn một biến độc lập, biểu diễn mối quan hệ tuyến tính giữa các biến độc lập và biến phụ thuộc.

Mục tiêu của Linear Regression là tìm ra hệ số góc và điểm giao với trục tung sao cho hàm dự đoán tuyến tính đạt được sai số nhỏ nhất. Một trong những cách phổ biến để ước lượng các hệ số là sử dụng phương pháp Ordinary Least Squares (OLS), trong đó chúng ta cần tối thiểu hóa tổng bình phương sai số.

Linear Regression được ứng dụng rộng rãi trong nhiều lĩnh vực, như:

- Dự báo giá cả: dự đoán giá nhà, giá cổ phiếu, giá nhiên liệu dựa trên các yếu tố như vị trí, kích thước, chất lượng, lượng cung cầu,…

- Dự báo điểm số: dự đoán điểm số của học sinh dựa trên thời gian học, nỗ lực, kỹ năng, trình độ giáo viên,…

- Dự báo sản phẩm: dự đoán đầu ra sản xuất dựa trên thời gian, công suất, nguyên liệu, lao động,…

- Phân tích chuỗi thời gian: dự đoán xu hướng và chu kỳ của các chuỗi dữ liệu, như bất động sản, thời tiết, xu hướng sản xuất,…

#### **Random Forest**

Random Forest là thuật toán học máy dựa trên kỹ thuật lắp ghép, kết hợp các cây phân lớp. Random Forest xây dựng cây phân lớp bằng cách lựa chọn ngẫu nhiên một nhóm nhỏ các thuộc tính tại mỗi nút của cây để phân chia cho mức tiếp theo của cây phân lớp. Ngoài ra tập mẫu của mỗi cây cũng được lựa chọn ngẫu nhiên bằng phương pháp Bootstrap từ tập mẫu ban đầu. Số lượng các cây phân lớp trong rừng là không hạn chế và thuật toán sử dụng kết quả dự đoán của tất cả cây trong rừng làm kết quả cuối cùng của thuật toán.

Random Forest được xây dựng dựa trên 3 thành phần chính là:

- Thuật toán xây dựng cây quyết định.

- Học máy.

- Tổng hợp theo bootstrap/bagging.

**Ưu và nhược điểm của thuật toán**

**a. Ưu điểm**

- Thuật toán giải quyết tốt các bài toán có nhiều dữ liệu nhiễu, thiếu giá trị. Do cách chọn ngẫu nhiên thuộc tính nên các giá trị nhiễu, thiếu ảnh hưởng không lớn đến kết quả.

- Có những sự ước lượng nội tại như độ chính xác của mô hình phỏng đoán hoặc độ mạnh và liên quan giữa các thuộc tính(Out of bag)

- Dễ dàng thực hiện song song. Thay vì một máy thực hiện cả thuật toán, ta có thể sử dụng nhiều máy để xây dựng các cây sau đó ghép lại thành rừng

- Các sai số được giảm thiểu do kết quả của Random Forest được tổng hợp thông qua nhiều người học (Cây phân lớp).

- Việc chọn ngẫu nhiên tại mỗi bước trong Random Forest sẽ làm giảm mối tương quan giữa các người học trong việc tổng hợp các kết quả.

- Lỗi chung của một rừng các cây phân lớp phụ thuộc vào lỗi riêng của từng cây trong rừng cũng như mối tương quan giữa các cây.

**b. Hạn chế**

- Dữ liệu huấn luyện cần được đa dạng hóa và cân bằng về số nhãn lớp. Việc không cân bằng nhãn lớp khiến kết quả dự đoán của thuật toán có thể lệch về số đông nhãn lớp.

- Thời gian huấn luyện của rừng có thể kéo dài tùy số cây và số thuộc tính phân chia.

**c. Out of bag**

Random Forest sử dụng Bootstrap để tạo tập mẫu cho các cây. Khi tập mẫu được rút ra từ một tập huấn luyện của một cây với sự thay thế thì theo ước tính có khoảng 1/3 các phần tử không có nằm trong mẫu này. Điều này có nghĩa là chỉ có khoảng 2/3 các phần tử trong tập huấn luyện tham gia vào trong các tính toán, và 1/3 các phần tử này được gọi là dữ liệu out-of-bag. Dữ liệu out-of-bag được sử dụng để ước lượng lỗi tạo ra từ việc kết hợp các kết quả từ các cây tổng hợp trong Random Forest cũng như dùng để ước tính độ quan trọng thuộc tính. Hơn nữa có thể sử dụng chính tập huấn luyện để kiểm thử mô hình từ thuật toán trước khi đưa vào ứng dụng.

#### **ARIMA**

Dự báo chuỗi thời gian là một lớp mô hình quan trọng trong Machine Learning. Sở dĩ chúng ta gọi lớp mô hình này là chuỗi thời gian (time series) là vì mô hình được áp dụng trên các chuỗi đặc thù có yếu tố thời gian. Một mô hình chuỗi thời gian thường dự báo dựa trên giả định rằng các qui luật trong quá khứ sẽ lặp lại ở tương lai. Do đó xây dựng mô hình chuỗi thời gian là chúng ta đang mô hình hóa mối quan hệ trong quá khứ giữa biến độc lập (biến đầu vào) và biến phụ thuộc (biến mục tiêu). Dựa vào mối quan hệ này để dự đoán giá trị trong tương lai của biến phụ thuộc.

Mô hình SARIMAX là một biến thể của mô hình ARIMA được mở rộng để xử lý các yếu tố mùa vụ và các biến giải thích bên ngoài.

**a. Các tham số của ARIMA và SARIMAX**

* ARIMA (p, d, q)

p: Bậc của phần tự hồi quy (AR). Chỉ số này xác định số lượng giá trị quá khứ sẽ được sử dụng để dự đoán giá trị hiện tại.

d: Bậc của phần tích hợp (I). Chỉ số này xác định số lần lấy sự chênh lệch để loại bỏ xu hướng từ chuỗi thời gian.

q: Bậc của phần trung bình động (MA). Chỉ số này xác định số lượng lỗi dự đoán trước đó sẽ được sử dụng để dự đoán giá trị hiện tại.

* SARIMAX(p, d, q) x (P, D, Q, s) + X

p: Bậc của phần tự hồi quy (AR).

d: Bậc của phần tích hợp (I).

q: Bậc của phần trung bình động (MA).

P: Bậc của phần tự hồi quy mùa vụ (Seasonal AR).

D: Bậc của phần tích hợp mùa vụ (Seasonal I).

Q: Bậc của phần trung bình động mùa vụ (Seasonal MA).

s: Chu kỳ mùa vụ.

X: Biến giải thích bên ngoài (exogenous variables)

**b. Quá trình tự hồi quy AR**

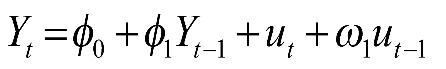
Quá trình tự hồi quy bậc p có dạng như sau:

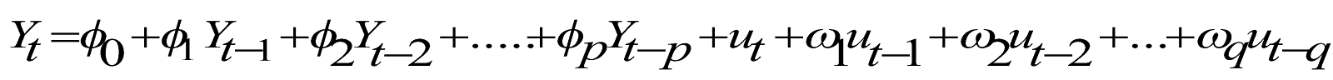
**c. Quá trình trung bình trượt MA**

Quá trình MA(q) là quá trình có dạng:

**d. Quá trình trung bình trượt và tự hồi quy ARMA**

Cơ chế sản sinh ra Y không chỉ có AR hoặc MA mà có thể kết hợp cả hai yếu tố này. Khi kết hợp cả hai yếu tố này ta có quá trình trung bình trượt và tự hồi quy Yt ARMA.

Quá trình ARMA(1,1) nếu y có thể biểu diễn dưới dạng:

Tổng quát ta có quá trình ARMA(p,q) nếu có dạng:

**e. Quá trình trung bình trượt, tích hợp, tự hồi quy ARIMA**

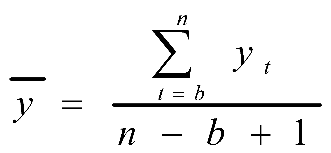
Chuỗi thời gian xuất phát có thể dừng hoặc không dừng. Để làm chuỗi dừng có chúng ta sẽ lấy sai phân. Chuỗi được gọi là đồng liên kết bậc d nếu chuỗi sai phân bậc d là chuỗi dừng. Áp dụng mô hình ARIMA (p,q) cho ta quá trình trung bình trượt, tích hợp, tự hồi quy ARIMA(p,d,q)

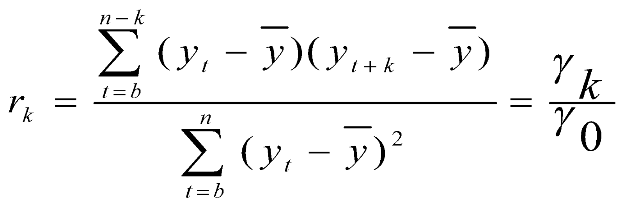
**f. Phương pháp BOX JENKINS**

Phương pháp này là sử dụng chuỗi thời gian trong quá khứ để dự báo cho tương lai. Việc kết hợp mô hình ARIMA trình bày ở trên và phương pháp Box-Jenkins sẽ giúp ta xử lý vấn đề này. Nội dung của phương pháp gồm 5 bước cơ bản:

**Bước 1. Kiểm tra tính dừng**

Có 3 phương pháp:

* Vẽ đồ thị chuỗi thời gian, nhận xét qua hình dạng của đồ thị.
* Thông qua hàm tự tương quan mẫu: SACF

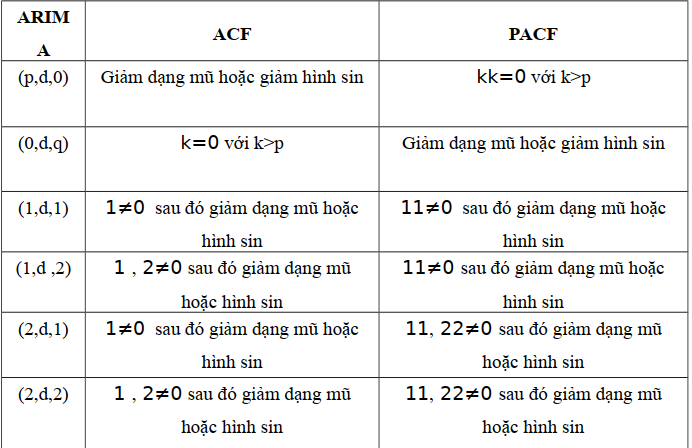


Nếu: SACF giảm nhanh và không có xu thế → chuỗi dừng.

SACF giảm chậm, có xu thế → chuỗi không dừng

* Sử dụng phương pháp kiểm định nghiệm đơn vị

**Bước 2. Nhận dạng mô hình**

Nhận dạng mô hình tức là chúng ta phải tìm ra các giá trị p,q và d. Để tìm được d ta dùng kiểm định nghiệm đơn vị DF hoặc ADF, kiểm định nghiệm đơn vị (làm chuỗi dừng). Từ chuỗi dừng nhận được ta phải tìm ra giá trị p,q. Ta có bảng tổng kết một số trường hợp:

Hình 3. Đồ thị ACF và PACF để xác định giá trị p, d, q, P, D, Q

Các quá trình bậc cao hơn cần phải thử và kết hợp các phương pháp định dạng khác nhau. Từ lược đồ tương quan có thể cho ta kết quả nhiều giá trị p, q khác nhau nên ta có các mô hình ARIMA khác nhau. Các mô hình khác nhau thì cho ta kết quả dự báo khác nhau, thế mô hình nào cho ta kết quả dự báo tốt nhất, giải quyết vấn đề này ta dựa trên các tiêu chuẩn lựa chọn.

**Bước 3: Dự đoán và đánh giá**

Sau khi định dạng mô hình ta biết bậc sai phân d của chuỗi xuất phát để thu được chuỗi dừng. Và ta cũng đã biết p,q. Do đó đã có thể đưa ra dự đoán. Để có cơ sở cho việc ra quyết định cần phải đánh giá dự báo nhằm xem xét kết quả dự báo chính xác đến như thế nào? Để đánh giá dự báo ta dựa vào một số tiêu chí như:

* Mean Absolute Error (MAE): giá trị tuyệt đối trung bình của sự chênh lệch giữa dự đoán và giá trị thực tế.
* Mean Squared Error (MSE): trung bình của bình phương của sự chênh lệch giữa dự đoán và giá trị thực tế.
* Root Mean Squared Error (RMSE): RMSE là căn bậc hai của MSE. Nó cũng đo lường độ lớn của sự chênh lệch, nhưng nó có thể giúp giảm sự nhạy cảm đối với các giá trị lớn bằng cách lấy căn bậc hai.
* Hệ số xác định (R-squared, R2): đo lường phần trăm biến thiên của biến phụ thuộc mà mô hình có thể giải thích, càng gần 1, mô hình càng tốt.

# CHƯƠNG 4. THỰC NGHIỆM VÀ ĐÁNH GIÁ

## 4.1. Thực nghiệm

### 4.1.1. Chuẩn bị dữ liệu

Như đã nói ở trên, bộ dữ liệu đã thu thập chứa tới 2220 bản ghi bao gồm tất cả các nước trên thế giới với dữ liệu từ năm 2005 tới 2023. Nhưng ở đây chỉ quan tâm tới Việt Nam và các nước láng giềng lân cận, nên nhóm sẽ thu hẹp phạm vi ở Đông Nam Á và Đông Á và lưu và file mới Asia.csv.

df = pd.read\_csv('WHR2005-2023.csv')

Asia = df[df['Regional Indicator'].isin(['Southeast Asia', 'East Asia'])]

Asia = Asia.round(4)

Asia.to\_csv('Asia.csv', index = False)

Bởi bộ dữ liệu còn nhiều ô null nên tiến hành xử lý giá trị null là không thể bỏ qua. Xử lý bằng cách lặp qua từng quốc gia và điền giá trị trung bình của cột đó của quốc gia đó cho các ô null rồi lưu vào file Asia1.csv.

# Lặp qua từng quốc gia và điền giá trị trung bình cho các ô null

for country in countries:

    country\_data = Asia[Asia['Country Name'] == country]

    for column in country\_data.columns:

        if country\_data[column].dtype == 'float64':

            Asia.loc[Asia['Country Name'] == country, column] = country\_data[column].fillna(country\_data[column].mean())

Asia.to\_csv('Asia1.csv', index = False)

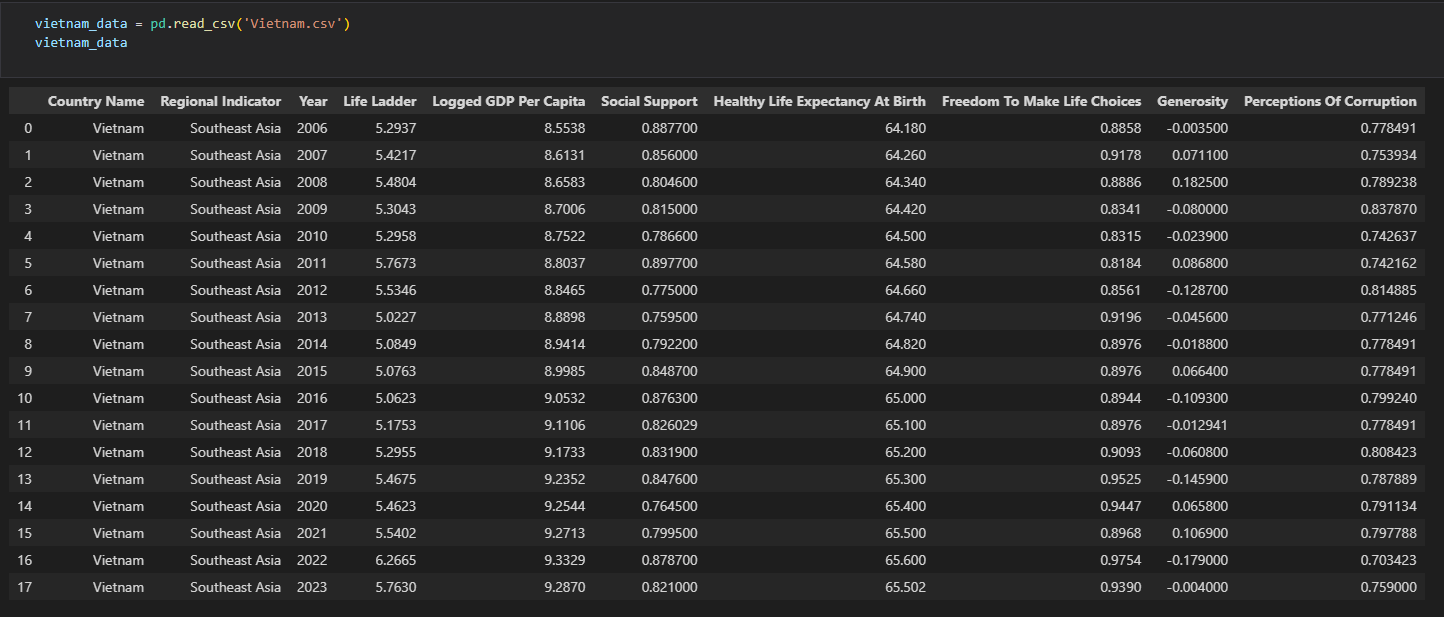
Và cuối cùng là tách dữ liệu của Việt Nam và lưu vào Vietnam.csv.

asia\_data = pd.read\_csv('Asia1.csv')

# Tách dữ liệu của Việt Nam

vietnam\_data = asia\_data[asia\_data['Country Name'] == 'Vietnam']

vietnam\_data.to\_csv('Vietnam.csv', index=False)



Hình 4. Bộ dữ liệu chỉ còn lại Việt Nam

### 4.1.2. Sử dụng mô hình ARIMA để dự đoán cho từng biến độc lập và mô hình Linear Regression để dự đoán chỉ số hạnh phúc của Việt Nam trong năm 2024

Đọc dữ liệu từ Vietnam.csv đã được tạo ở trên. Tạo dataframe chỉ chứa các cột cần thiết rồi chia dữ liệu thành tập huấn luyện và tập kiểm tra với tỉ lệ 80% train và 20% test.

# Đọc dữ liệu từ file CSV

vietnam\_data = pd.read\_csv('Vietnam.csv')

# Chọn các cột cần thiết

selected\_columns = ['Logged GDP Per Capita', 'Social Support', 'Healthy Life Expectancy At Birth', 'Freedom To Make Life Choices', 'Generosity', 'Perceptions Of Corruption', 'Life Ladder', 'Year']

data\_selected = vietnam\_data[selected\_columns]

# Chia dữ liệu thành tập huấn luyện và tập kiểm tra

train\_data, test\_data = train\_test\_split(data\_selected, test\_size=0.2, random\_state=42)

# Tạo mô hình Linear Regression

model = LinearRegression()

# Huấn luyện mô hình

X\_train = train\_data[selected\_columns[:-2]]  # Chọn tất cả các cột ngoại trừ 'Life Ladder' và 'Year'

y\_train = train\_data['Life Ladder']

model.fit(X\_train, y\_train)

# Kiểm tra mô hình trên tập kiểm tra

X\_test = test\_data[selected\_columns[:-2]]  # Chọn tất cả các cột ngoại trừ 'Life Ladder' và 'Year'

y\_test = test\_data['Life Ladder']

predictions = model.predict(X\_test)

Sau khi chuẩn bị dữ liệu xong, tiến hành sử dụng mô hình ARIMA để dự đoán cho các biến độc lập. Sử dụng vòng lặp for lặp qua từng biến độc lập. Với forecasted\_values dùng để lưu trữ giá trị dự đoán.

# Dự đoán cho năm 2024 sử dụng ARIMA cho từng biến độc lập

independent\_variables = ['Logged GDP Per Capita', 'Social Support', 'Healthy Life Expectancy At Birth', 'Freedom To Make Life Choices', 'Generosity', 'Perceptions Of Corruption']

forecasted\_values = []  # Define an empty list to store forecasted values

for variable in independent\_variables:

    # Tạo mô hình ARIMA với chỉ mục là 'Year'

    model\_arima = sm.tsa.ARIMA(train\_data[variable], order=(0, 0, 1))  # Thay p, d, q bằng các giá trị thích hợp

    results = model\_arima.fit()

    # Dự đoán giá trị cho năm 2024

    forecast = results.get\_forecast(steps=1)

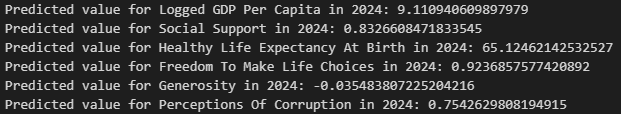
    forecasted\_value = forecast.predicted\_mean.iloc[0]

    forecasted\_values.append(forecasted\_value)

    # In kết quả

    print(f'Predicted value for {variable} in 2024: {forecasted\_value}')

Trong mô hình ARIMA cho các giá trị p,d,q tương ứng với order = (0, 0, 1). Kết quả được lưu trong biến results. Huấn luyện biến results bằng get\_forecast(steps=1) với steps = 1 để dự đoán giá trị cho bước tiếp theo, cụ thể ở đây là năm 2024. Sau đó sử dụng predicted\_mean.iloc[0] để lấy kết quả dự đoán đầu tiên trong dãy dự đoán.



Hình 5. Kết quả dự đoán các biến độc lập bằng ARIMA

Sử dụng kết quả từ ARIMA đưa vào mô hình Linear Regression để dự đoán Life Ladder cho năm 2024.

predicted\_life\_ladder\_2024 = model.predict([forecasted\_values])

print(f'Predicted Life Ladder for 2024: {predicted\_life\_ladder\_2024[0]}')

Sử dụng phương thức predict() của mô hình Linear Regression để dự đoán từ các chỉ số đã dự đoán và cho in ra kết quả dự đoán đầu tiên.



Hình 6. Kết quả dự đoán Life Ladder bằng Linear Regression dựa trên ARIMA

Đánh giá mô hình dựa trên 4 chỉ số:

# Đánh giá mô hình Linear Regression

mae\_lr = mean\_absolute\_error(y\_test\_lr, predictions\_lr)

mse\_lr = mean\_squared\_error(y\_test\_lr, predictions\_lr)

rmse\_lr = np.sqrt(mse\_lr)

r2\_lr = r2\_score(y\_test\_lr, predictions\_lr)

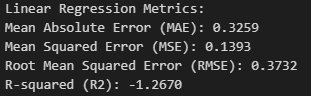
print("Linear Regression Metrics:")

print(f'Mean Absolute Error (MAE): {mae\_lr:.4f}')

print(f'Mean Squared Error (MSE): {mse\_lr:.4f}')

print(f'Root Mean Squared Error (RMSE): {rmse\_lr:.4f}')

print(f'R-squared (R2): {r2\_lr:.4f}')



Hình 7. Kết quả đánh giá mô hình Linear Regression

Kết quả cho ra các chỉ số MAE, MSE, RMSE khá ổn duy chỉ có R2 là hơi thấp.

### 4.1.3. Sử dụng mô hình SARIMAX để dự đoán cho từng biến độc lập và mô hình Linear Regression để dự đoán chỉ số hạnh phúc của Việt Nam trong năm 2024

Đọc dữ liệu rồi chia dữ liệu thành tập huấn luyện và tập kiểm tra với tỉ lệ 80% train và 20% test tương tự như trên.

# Đọc dữ liệu từ file CSV

vietnam\_data = pd.read\_csv('Vietnam.csv')

# Chọn các cột cần thiết

selected\_columns = ['Logged GDP Per Capita', 'Social Support', 'Healthy Life Expectancy At Birth', 'Freedom To Make Life Choices', 'Generosity', 'Perceptions Of Corruption', 'Life Ladder', 'Year']

data\_selected = vietnam\_data[selected\_columns]

# Chia dữ liệu thành tập huấn luyện và tập kiểm tra

train\_data, test\_data = train\_test\_split(data\_selected, test\_size=0.2, random\_state=42)

# Tạo mô hình Linear Regression

model\_lr = LinearRegression()

# Huấn luyện mô hình Linear Regression

X\_train\_lr = train\_data[selected\_columns[:-2]]  # Chọn tất cả các cột ngoại trừ 'Life Ladder' và 'Year'

y\_train\_lr = train\_data['Life Ladder']

model\_lr.fit(X\_train\_lr, y\_train\_lr)

# Kiểm tra mô hình Linear Regression trên tập kiểm tra

X\_test\_lr = test\_data[selected\_columns[:-2]]  # Chọn tất cả các cột ngoại trừ 'Life Ladder' và 'Year'

y\_test\_lr = test\_data['Life Ladder']

predictions\_lr = model\_lr.predict(X\_test\_lr)

Sau khi chuẩn bị dữ liệu xong, tiến hành sử dụng mô hình SARIMAX để dự đoán cho các biến độc lập. Các bước thực hiện giống như mô hình ARIMA chỉ khác ở các chỉ số order = (1, 0, 1), seasonal\_order=(1, 1, 1, 12), enforce\_stationarity = True: áp đặt điều kiện về tính đồng nhất, enforce\_invertibility = False: Không áp đặt điều kiện về tính nghịch đảo

Kết quả được lưu trong biến results\_sarimax.

# Dự đoán cho năm 2024 sử dụng SARIMAX cho từng biến độc lập

independent\_variables = ['Logged GDP Per Capita', 'Social Support', 'Healthy Life Expectancy At Birth', 'Freedom To Make Life Choices', 'Generosity', 'Perceptions Of Corruption']

forecasted\_values\_sarimax = {}

for variable in independent\_variables:

    # Tạo mô hình SARIMAX với chỉ mục là 'Year'

    model\_sarimax = SARIMAX(train\_data[variable], order=(1, 0, 1), seasonal\_order=(1, 1, 1, 12), enforce\_stationarity=True, enforce\_invertibility=False)

    results\_sarimax = model\_sarimax.fit()

    # Dự đoán giá trị cho năm 2024

    forecast\_sarimax = results\_sarimax.get\_forecast(steps=1)

    forecasted\_value\_sarimax = forecast\_sarimax.predicted\_mean.iloc[0]

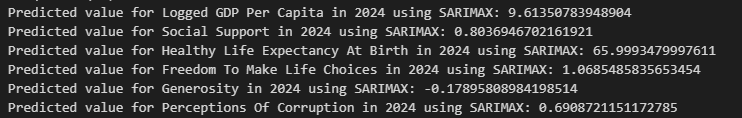
    # Lưu giá trị dự đoán vào từ điển

    forecasted\_values\_sarimax[variable] = forecasted\_value\_sarimax

    # In kết quả

    print(f'Predicted value for {variable} in 2024 using SARIMAX: {forecasted\_value\_sarimax}')

Huấn luện biến results\_sarimax bằng get\_forecast(steps=1) với steps = 1 để dự đoán giá trị cho bước tiếp theo, cụ thể ở đây là năm 2024. Sau đó sử dụng predicted\_mean.iloc[0] để lấy kết quả dự đoán đầu tiên trong dãy dự đoán.



Hình 8. Kết quả dự đoán các biến độc lập bằng SARIMAX

Sử dụng kết quả từ SARIMA đưa vào mô hình Linear Regression để dự đoán Life Ladder cho năm 2024.

# Sử dụng mô hình Linear Regression để dự đoán Life Ladder cho năm 2024 sử dụng giá trị từ SARIMAX

X\_test\_lr\_2024 = np.array(list(forecasted\_values\_sarimax.values())).reshape(1, -1)

predicted\_life\_ladder\_2024\_sarimax = model\_lr.predict(X\_test\_lr\_2024)

print(f'Predicted Life Ladder for 2024: {predicted\_life\_ladder\_2024\_sarimax[0]}')

Chuyển đổi danh sách giá trị thành 1 mảng NumPy và thay đổi hình dạng thành mảng 2D bởi mô hình Linear Regression yêu cầu đầu vào là một mảng 2D sau đó đưa vào mô hình.

Hình 9. Kết quả dự đoán Life Ladder bằng Linear Regression dựa trên SARIMAX

### 4.1.4. Sử dụng mô hình Random Forest Regressor để dự đoán cho chỉ số hạnh phúc của Việt Nam.

Tạo mô hình Random Forest Regressor và tạo tập huấn luyện, tập kiểm tra tương tự Linear Regression.

from sklearn.ensemble import RandomForestRegressor

# Tạo mô hình Random Forest Regressor

forest\_model = RandomForestRegressor(random\_state=42)

# Huấn luyện mô hình trên tập huấn luyện

forest\_model.fit(X\_train, y\_train)

# Dự đoán trên tập kiểm tra

forest\_predictions = forest\_model.predict(X\_test)

Sau khi mô hình đã được huấn luyện, sử dụng phương thức predict() để dự đoán giá trị trên tập kiểm tra (X\_test). Kết quả dự đoán được lưu vào forest\_predictions.

# Dự đoán Life Ladder cho năm 2024

predicted\_life\_ladder\_2024\_sarimax = forest\_model.predict(X\_test\_lr\_2024)

print(f'Predicted Life Ladder for 2024 using Random Forest Regressor: {predicted\_life\_ladder\_2024\_sarimax[0]}')



Hình 10. Kết quả dự đoán Life Ladder bằng Random Forest Regressor dựa trên SARIMAX

Đánh giá mô hình dựa trên 4 chỉ số:

# Đánh giá mô hình Random Forest

mae\_forest = mean\_absolute\_error(y\_test, forest\_predictions)

mse\_forest = mean\_squared\_error(y\_test, forest\_predictions)

rmse\_forest = np.sqrt(mse\_forest)

r2\_forest = r2\_score(y\_test, forest\_predictions)

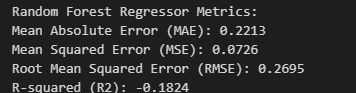
print("Random Forest Regressor Metrics:")

print(f'Mean Absolute Error (MAE): {mae\_forest:.4f}')

print(f'Mean Squared Error (MSE): {mse\_forest:.4f}')

print(f'Root Mean Squared Error (RMSE): {rmse\_forest:.4f}')

print(f'R-squared (R2): {r2\_forest:.4f}')



Hình 11. Kết quả đánh giá mô hình Random Forest Regressor

Kết quả cho ra các chỉ số đẹp hơn mô hình trên rất nhiều, đặc biệt R2 chỉ bằng -0.18. Tốt hơn hẳn so với -1.267 của Linear Regression.

## 4.2. Đánh giá

Kết quả cho thấy tất cả các giá trị MAE, MSE, và RMSE của Random Forest Regressor đều thấp hơn so với Linear Regression, chỉ ra rằng Random Forest Regressor có hiệu suất tốt hơn.

R-squared của cả hai mô hình đều âm, cho thấy mô hình không giải thích được mức độ lớn của biến độc lập. Tuy nhiên, giá trị R2 của Random Forest Regressor là gần 0 hơn, điều này có thể được coi là một cải thiện so với Linear Regressiontrong việc dự đoán giá trị.

Cả hai mô hình đều có thể cần được cải thiện và điều chỉnh để có kết quả tốt hơn, nhưng với các số liệu này, Random Forest Regressor có vẻ hiệu quả hơn trong bối cảnh của công việc dự đoán chỉ số hạnh phúc cho Việt Nam.

# KẾT LUẬN

Trong bối cảnh ngày càng phức tạp của thế giới hiện đại, nghiên cứu và dự đoán chỉ số hạnh phúc của Việt Nam đã đưa ra những thông tin quan trọng về sức khỏe kinh tế và xã hội của đất nước. Các biến đổi trong các yếu tố ảnh hưởng đến chỉ số hạnh phúc như GDP bình quân đầu người, tuổi thọ khỏe mạnh khi mới sinh, và tự do trong việc lựa chọn cuộc sống đều là những yếu tố quan trọng được dự đoán và đánh giá.

Dựa trên mô hình SARIMAX, chúng tôi đã thực hiện dự đoán cho các biến có tính ổn định như 'Logged GDP Per Capita', 'Healthy Life Expectancy At Birth', và 'Freedom To Make Life Choices', cũng như các biến không ổn định như 'Social Support', 'Generosity', và 'Perceptions Of Corruption'. Các kết quả dự đoán cho thấy sự biến động của chỉ số hạnh phúc trong tương lai và sự ảnh hưởng của các yếu tố khác nhau.

Tuy nhiên, cần lưu ý rằng dự đoán chỉ là ước lượng dựa trên dữ liệu hiện có và không thể hoàn toàn đảm bảo tính chính xác trong thực tế. Thêm vào đó, các biến số không dự đoán được như sự biến động trong chính trị, sự kiện quốc tế, hay thậm chí là những yếu tố tự nhiên có thể ảnh hưởng đến kết quả cuối cùng.

Tóm lại, bài nghiên cứu này cung cấp cái nhìn tổng quan về dự đoán chỉ số hạnh phúc của Việt Nam và có thể làm cơ sở cho các quyết định chính trị và xã hội trong tương lai. Tuy nhiên, để đạt được hiệu quả tốt nhất, cần kết hợp thông tin từ nhiều nguồn và cân nhắc đến sự biến động đa chiều của môi trường xã hội và kinh tế.

# TÀI LIỆU THAM KHẢO

Giáo trình Khai phá dữ liệu – HVNH

[Khai phá dữ liệu – Wikipedia tiếng Việt](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u#:~:text=Khai%20ph%C3%A1%20d%E1%BB%AF%20li%E1%BB%87u%20(data,c%E1%BB%A7a%20khoa%20h%E1%BB%8Dc%20m%C3%A1y%20t%C3%ADnh.)

[Data Mining là gì? Các công cụ khai phá dữ liệu phổ biến nhất 2023 | TopOnSeek](https://www.toponseek.com/blogs/data-mining-la-gi/)

[Regression trong Data Mining - w3seo các loại hồi quy trong data mining (websitehcm.com)](https://websitehcm.com/regression-trong-data-mining/)

[Chỉ Số Hạnh Phúc Là Gì? Tổng Quan Về Chỉ Số Hạnh Phúc - Glints Vietnam Blog](https://glints.com/vn/blog/chi-so-hanh-phuc-la-gi/)

<https://trituenhantao.io/machine-learning-co-ban/bai-3-linear-regression-hoi-quy-tuyen-tinh/>

<https://trituenhantao.io/kien-thuc/decision-tree/>

<https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html>

<https://svcuong.github.io/post/ensemble-learning/>