TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐỀ CƯƠNG CHUYÊN ĐỀ HỌC PHẦN**

**TÊN HỌC PHẦN**

**ĐỀ TÀI:**

**Sử dụng thuật toán hồi quy tuyến tính áp dụng cho dự đoán doanh thu phim, trạng thái mới nhất của Covid-19**

**tại Ấn Độ, trạng thái mới nhất của Covid-19 tại USA**

|  |  |
| --- | --- |
|  |  |
| |  |  | | --- | --- | | **Sinh viên thực hiện** | **: Nguyễn Tiến Tài**  **Nguyễn Minh Quang**  **Nguyễn Mạnh Cương** | | **Giảng viên hướng dẫn** | **: (IN HOA)** | | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀN** | | | **Lớp** | **: D13CNPM7** | | | **Khóa** | **: 2018-2023** | | |  |
|  |  |

***Hà Nội, tháng 06 năm 2021***

**PHIẾU CHẤM ĐIỂM**

Sinh viên thực hiện:

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| Nguyễn Tiến Tài |  |  |
| Nguyễn Minh Quang |  |  |
| Nguyễn Mạnh Cương |  |  |

Giảng viên chấm:

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1 : |  |  |
| Giảng viên chấm 2 : |  |  |

MỤC LỤC

[LỜI CẢM ƠN 4](#_Toc75985492)

[**CHƯƠNG I: TỔNG QUAN VỀ NHẬP MÔN HỌC MÁY** 5](#_Toc75985493)

[**I. Khái niệm cơ bản** 5](#_Toc75985494)

[**1. Học máy là gì?** 5](#_Toc75985495)

[**2. Học tăng cường là gì?** 5](#_Toc75985496)

[**3. Học có giám sát là gì?** 6](#_Toc75985497)

[**4. Học không giám sát là gì?** 7](#_Toc75985498)

[**II. Hồi quy tuyến tính** 8](#_Toc75985499)

[**1. Phân tích hồi quy là gì?** 8](#_Toc75985500)

[**2. Điều kiện tiên quyết của việc sử dụng mô hình hồi quy tuyến tính** 8](#_Toc75985501)

[**3.Biểu diễn thống kê** 9](#_Toc75985502)

[**4. Phương trình hồi quy có nghĩa là gì?** 9](#_Toc75985503)

[**5. Hệ số xác định - Đánh giá mức độ phù hợp** 13](#_Toc75985504)

[**CHƯƠNG 2. SỬ DỤNG THUẬT TOÁN HỒI QUY TUYẾN TÍNH ÁP DỤNG CHO BÀI TOÁN CỤ THỂ** 15](#_Toc75985505)

[**I. Sử dụng thuật toán hồi quy tuyến tính áp dụng cho dự đoán doanh thu phim** 15](#_Toc75985506)

[**1.1. Dữ liệu bài toán** 15](#_Toc75985507)

[**1.2. Cài đặt chương trình** 16](#_Toc75985508)

[**1.3. Kết quả** 18](#_Toc75985509)

[**2. Sử dụng thuật toán hồi quy tuyến tính áp dụng cho trạng thái mới nhất của Covid-19 tại Ấn Độ** 20](#_Toc75985510)

[**2.1. Dữ liệu bài toán** 20](#_Toc75985511)

[**2.2. Cài đặt chương trình** 21](#_Toc75985512)

[**2.3. Kết quả** 23](#_Toc75985513)

[**3. Sử dụng thuật toán hồi quy tuyến tính áp dụng cho trạng thái mới nhất của Covid-19 tại USA** 25](#_Toc75985514)

[**3.1. Dữ liệu bài toán** 25](#_Toc75985515)

[**3.2. Cài đặt chương trình** 26](#_Toc75985516)

[**3.3. Kết quả** 28](#_Toc75985517)

[**4. Đánh giá** 31](#_Toc75985518)

[**KẾT LUẬN** 32](#_Toc75985519)

# LỜI CẢM ƠN

Sau quá trình học tập và rèn luyện tại khoa Công nghệ thông tin trường Đại học Điện Lực chúng em đã được trang bị các kiến thức cơ bản, các kỹ năng thực tế để có thể hoàn thành bài báo cáo môn học của mình.

Chúng em xin gửi lời cảm ơn chân thành đến thầy Ngô Hoàng Huy đã quan tâm hướng dẫn truyền đạt học những kiến thức và kinh nghiệm cho chúng em trong suốt thời gian học học tập bộ môn Nhập môn học máy .

Trong quá trình làm báo cáo môn không tránh khỏi được những sai sót, chúng em mong nhận được sự góp ý của quý thầy cô và các bạn để được hoàn thiện hơn.

Chúng em xin chân thành cảm ơn !

**CHƯƠNG I: TỔNG QUAN VỀ NHẬP MÔN HỌC MÁY**

**I. Khái niệm cơ bản**

**1. Học máy là gì?**

Nhưng làm thế nào một cỗ máy có thể được lập trình để xử lý thông tin (tức là dữ liệu) một cách thông minh? Câu trả lời cho điều đó là Học máy (ML). ML là một tập hợp con của AI cụ thể đến từ sự kết hợp giữa thống kê và khoa học máy tính. Nó đề cập đến quá trình máy tính học cách thực hiện một tác vụ thay vì làm theo hướng dẫn từng bước. Điều này thường được thực hiện lặp đi lặp lại bởi các nhà khoa học dữ liệu hướng dẫn một máy tính nếu các quyết định của nó là đúng hay sai. Tùy thuộc vào kết quả, máy tính sẽ điều chỉnh cách nó đưa ra quyết định trong tương lai — nói cách khác, “nó học”.



Mối quan hệ giữa AI, ML và ba loại ML rộng.

Khi nói đến ML về cơ bản có ba loại chính:

1. Học tăng cường,

2. Học tập không giám sát và

3. Học tập có giám sát.

**2. Học tăng cường là gì?**

Reinforcement Learning có lẽ được biết đến nhiều nhất thông qua máy tính Deep Blue của IBM, một "robot" học cách chơi cờ vua và đánh bại nhà vô địch thế giới loài người.

Học tăng cường là một loại kỹ thuật cho phép thuật toán học bằng cách thử và sai, sử dụng phản hồi từ các hành động và kinh nghiệm của chính nó. Giống như Pavlov và chú chó của anh ấy, Reinforcement Learning liên quan đến các quyết định khen thưởng dẫn đến thành công và trừng phạt các quyết định dẫn đến bất cứ điều gì khác ngoài thành công - cuối cùng làm cho thuật toán trở nên thông minh hơn trong quá trình này.

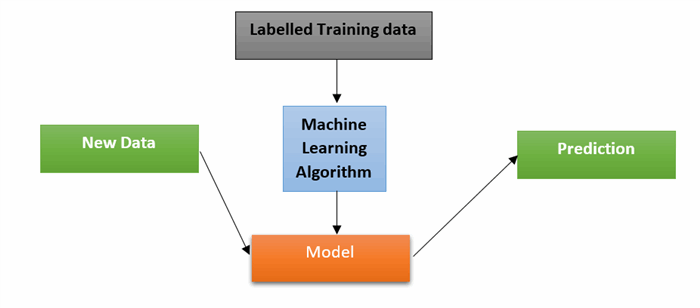
Các ví dụ về học tăng cường được áp dụng trong nhân sự là một chút tinh gọn, mặc dù phổ biến nhất trong các lĩnh vực như giáo dục (tức là áp dụng nội dung dựa trên sự tiến bộ của sinh viên), tài chính và đầu tư (tức là dự báo nâng cao), hoạt động chuỗi cung ứng (tức là , rô bốt thực hiện đơn đặt hàng trong nhà kho), tối ưu hóa luồng giao thông và chăm sóc sức khỏe (tức là phân loại chính xác các hình ảnh sinh thiết).

**3. Học có giám sát là gì?**

Các hình thức ML phổ biến nhất trong các ngành, và cụ thể là lĩnh vực nhân sự, là Học có giám sát, sau đó là Học không giám sát.

Trong Học tập có giám sát, chúng tôi cố gắng dự đoán một kết quả, chẳng hạn như liệu một nhân viên sẽ rời công ty, nguy cơ nhân viên bị thương hay mức lương khởi điểm lý tưởng của một nhân viên mới.

Để đưa ra dự đoán, chúng ta cần các biến đầu vào khác nhau (tức là các biến được các nhà khoa học dữ liệu gọi là “Đặc điểm”). Các tính năng đầu vào của chúng tôi chỉ giới hạn trong trí tưởng tượng của chúng tôi (tức là những gì chúng tôi nghĩ là quan trọng), dữ liệu nào chúng tôi có thể thực hành hoặc dữ liệu nào chúng tôi có thể tạo (ví dụ: bằng cách biết ai đó làm việc và nơi họ sống, chúng tôi có thể tạo một biến tập trung vào khoảng cách đi làm của nhân viên).



Ví dụ: Học tập có giám sát thông báo về Doanh thu của nhân viên

Hãy xem xét một ví dụ chi tiết hơn về Học tập có giám sát — dự đoán ai sẽ rời tổ chức. Hãy tưởng tượng rằng 1/5 tân binh rời tổ chức trong 12 tháng đầu nhiệm kỳ của họ. Để ngăn chặn sự thay đổi đó, chúng tôi có thể xây dựng một mô hình học tập có giám sát dự đoán khả năng những người mới bắt đầu rời đi, để các đồng nghiệp quản lý và nhân sự của chúng tôi có thể can thiệp.

Trong ví dụ này, kết quả của mô hình được dự đoán là rủi ro doanh thu và các tính năng được sử dụng để dự đoán rủi ro doanh thu có thể bao gồm các đặc điểm về nhân khẩu học và việc làm của nhân viên (ví dụ: tuổi, trình độ học vấn, cấp độ vai trò, mức lương so với thị trường, tháng làm việc, sự hiện diện của các kế hoạch phát triển, v.v.).

Giả sử mô hình như vậy có độ chính xác cao, nó sẽ cho phép chúng tôi hiểu được doanh thu của nhóm người mới bắt đầu của chúng tôi từ ba góc độ.

1. Thứ nhất, những yếu tố nào ảnh hưởng nhiều nhất đến việc dự đoán doanh thu trong dân số của chúng ta. Ví dụ về đầu ra mô hình như vậy được trình bày trong hình bên dưới, minh họa liệu một tính năng ngăn cản doanh thu (thanh màu xanh lá cây) hay thúc đẩy doanh thu (đường màu đỏ) và tầm quan trọng tương đối của mỗi tính năng trong việc dự đoán doanh thu (tức là các đường dài hơn biểu thị mức độ quan trọng hơn ).

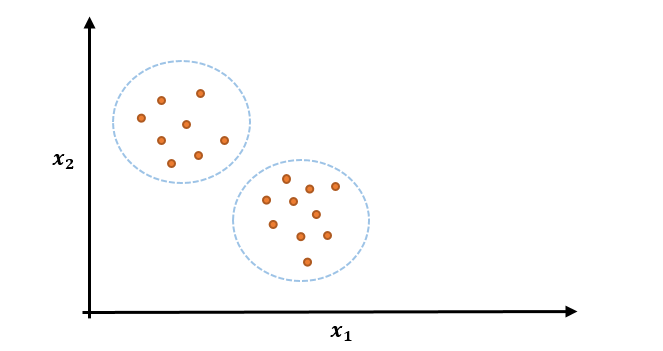
2. Thứ hai, mô hình cũng đánh giá khả năng mỗi người mới bắt đầu rời công ty, cho phép can thiệp tập trung (tức là rủi ro mà Adam sẽ rời đi trong 12 tháng đầu tiên).

3. Thứ ba, mô hình xác định các tính năng ngăn ngừa hoặc thúc đẩy rủi ro doanh thu cho từng nhân viên. Đầu ra được cá nhân hóa này có thể cho phép các chuyên gia nhân sự thực hiện hành động được thông báo và cá nhân hóa, bất kể họ có biết từng nhân viên hay không.

**4. Học không giám sát là gì?**

Không giống như Học có giám sát mà chúng ta đang cố gắng dự đoán một kết quả, Học không giám sát phân tích đồng thời nhiều biến số để xác định các điểm tương đồng, các mẫu hoặc các mối quan hệ trong dữ liệu. Học không giám sát thiên về hiểu những gì có trong dữ liệu. Hai cách sử dụng phổ biến nhất của học tập không giám sát được tập trung vào:

• Phân cụm : tự động chia tập dữ liệu thành các nhóm dựa trên sự tương đồng giữa các tính năng được phân tích. Được áp dụng theo kiểu cổ điển cho người tiêu dùng, nhưng cũng phù hợp với các tổ chức, nhờ đó chúng tôi hiểu các phân khúc nhân viên của mình (tức là các cụm) và xác định xem chính sách nhân sự của chúng tôi có phục vụ các phân khúc hay không.



• Khai thác liên kết : xác định các tập hợp các biến thường xuất hiện cùng nhau trong tập dữ liệu của bạn. Ví dụ, xác định các mô hình thương tích giữa các công nhân tại các địa điểm cụ thể.

Ví dụ: Học tập không giám sát thông báo về Doanh thu của nhân viên

Phân tích cụm, hình thức học tập không giám sát nổi tiếng nhất, cũng có thể giúp chúng ta hiểu rõ hơn về sự tiêu hao của nhân viên. Cách tiếp cận này có thể giúp nhóm các nhân viên dựa trên các đặc điểm tương tự (ví dụ: vị trí, nhiệm kỳ, quốc tịch, trình độ học vấn, tuổi, mức độ hiệu suất, v.v.)

Hình dưới đây mô tả kết quả phân tích các đặc điểm nhân khẩu học của nhân viên. Trước tiên, nhiều đối tượng nhân khẩu học được giảm xuống thành hai thứ nguyên bằng cách sử dụng một phương pháp gọi là Manifold Learning (một phương pháp không được giám sát khác) và hai thứ nguyên mới này sau đó được nhóm lại bằng một phương pháp gọi là T-SNE. Hình dưới đây cho chúng ta thấy cách các nhân viên có thể được nhóm lại với nhau, trong trường hợp này là mười hai nhóm, dựa trên các đặc điểm nhân khẩu học của họ.

Sau khi được nhóm thành các cụm, bước tiếp theo là xác định rủi ro doanh thu cho từng nhóm. Hơn nữa, thật thú vị khi xác định xem có một số yếu tố rủi ro được chia sẻ hay không, thực tế cho thấy rằng các nhân viên trong một nhóm đang trải nghiệm nơi làm việc theo cách tương tự.

Thông tin chi tiết cuối cùng này có ý nghĩa thực tế đáng kể, vì nó có thể giúp chúng tôi điều chỉnh các biện pháp can thiệp nhắm mục tiêu đến các nhóm nhân viên cụ thể, do đó mang lại tác động tối đa (tức là giữ chân nhân viên và giảm chi phí doanh thu) và lợi tức đầu tư của chúng tôi (tức là với mỗi đô la chi tiêu, chúng tôi tạo ra đô la xxx tiết kiệm do giảm doanh thu).

**II. Hồi quy tuyến tính**

**1. Phân tích hồi quy là gì?**

Phân tích hồi quy đơn giản hóa một số tình huống rất phức tạp, gần như một cách kỳ diệu. Nó giúp các nhà nghiên cứu và các chuyên gia tương quan các biến số đan xen nhau. Hồi quy tuyến tính là một trong những mô hình hồi quy đơn giản và được sử dụng phổ biến nhất. Nó dự đoán mối quan hệ nguyên nhân và kết quả giữa hai biến số.

Mô hình sử dụng phương pháp Bình phương nhỏ nhất thông thường (OLS) , phương pháp này xác định giá trị của các tham số chưa biết trong một phương trình hồi quy tuyến tính. Mục tiêu của nó là giảm thiểu sự khác biệt giữa các câu trả lời được quan sát và những câu trả lời được dự đoán bằng cách sử dụng mô hình hồi quy tuyến tính. Có một số yêu cầu nhất định mà bạn cần phải đáp ứng để sử dụng mô hình này. Nếu không, kết quả có thể gây nhầm lẫn và mơ hồ.

**2. Điều kiện tiên quyết của việc sử dụng mô hình hồi quy tuyến tính**

* Số lượng quan sát là hữu hạn.
* Giả định chính là có sai số không đáng kể trong giá trị của biến độc lập (X) hoặc biến hồi quy. Nó tuân theo nguyên tắc ngoại đồng nghiêm ngặt, có nghĩa là không có sai số.
* Các biến hồi quy hoặc biến độc lập phải là hằng số hoặc biến ngẫu nhiên được xác định trước.
* Sự khác biệt giữa điểm dữ liệu tương ứng với đường hồi quy càng nhỏ thì mô hình kinh tế lượng càng phù hợp với dữ liệu.
* Nó cung cấp công cụ ước lượng khả năng xảy ra tối đa, tối đa hóa sự thống nhất của mô hình kinh tế lượng đã chọn với dữ liệu quan sát.:

**Hiện tượng là:** Kinh nghiệm làm việc và lương thưởng là những biến số có liên quan với nhau. Mô hình hồi quy tuyến tính có thể giúp dự đoán mức lương thưởng của một nhân viên dựa trên kinh nghiệm làm việc của họ.

**3.Biểu diễn thống kê**

Bây giờ vấn đề đặt ra là làm thế nào để biểu diễn nó một cách thống kê.

Có hai dòng hồi quy - Y trên X và X trên Y. Y trên X là khi giá trị của Y chưa biết. X trên Y là khi giá trị của X chưa biết. Dưới đây là các đại diện thống kê của chúng: Giả sử, giá trị thù lao là = Y và giá trị kinh nghiệm là = X

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Dòng hồi quy của Y trên X** | **Dòng hồi quy của X trên Y** | | Sau đó, để dự đoán giá trị chưa biết của thù lao (Y), biểu diễn thống kê sẽ là: | Nếu ngược lại hoặc giá trị của kinh nghiệm làm việc là không xác định, thì biểu diễn thống kê sẽ là: | | Y = a + b (X) | X = c + d (Y) | | Trong đó 'b' là hệ số của X và 'a' là hệ số chặn của Y. | Trong đó 'd' là hệ số của Y và 'c' là giao điểm của X. | |

**Lựa chọn dòng hồi quy**

Biểu diễn thống kê ở trên là một ví dụ để chỉ ra cách phát triển các mô hình kinh tế lượng khi giá trị của một trong các biến được biết đến và giá trị của một biến khác chưa biết. Tuy nhiên, ***điều này không có nghĩa là cả hai cách biểu diễn đều đúng*** .

Điều này là do thù lao có thể phụ thuộc vào kinh nghiệm làm việc của một cá nhân nhưng ngược lại thì không đúng. Kinh nghiệm không phụ thuộc vào thù lao. Do đó, bạn sẽ phải cẩn thận chọn biến phụ thuộc và sau đó là dòng hồi quy.

**4. Phương trình hồi quy có nghĩa là gì?**

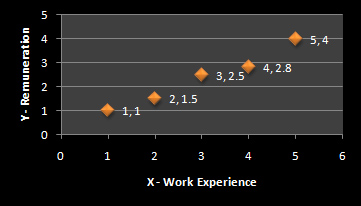
Phương trình hồi quy tuyến tính cho thấy phần trăm tăng hoặc giảm giá trị của biến phụ thuộc (Y) với phần trăm tăng hoặc giảm giá trị của biến độc lập (X).

Giả sử giá trị của X và Y đã biết.

**Bảng 1**

|  |  |
| --- | --- |
| **X** | **Y** |
| 1 | 1 |
| 2 | 1,5 |
| 3 | 2,5 |
| 4 | 2,8 |
| 5 | 4 |

Về mặt đồ họa, nó được biểu thị là:

  
**Đồ thị 1**

Đường màu trắng kết nối tất cả các dấu chấm trong biểu đồ trên biểu thị lỗi hoặc dự đoán. Nhưng bây giờ bạn muốn tìm đường hồi quy phù hợp nhất để giảm thiểu sai số của dự đoán. Mục đích là giúp tìm ra đường hồi quy phù hợp nhất.

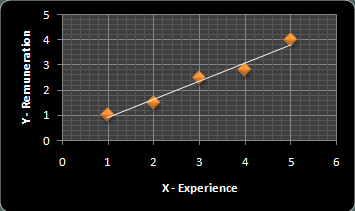
**Làm thế nào để Tìm Đường hồi quy Phù hợp nhất?**

Bằng cách sử dụng phương pháp bình phương nhỏ nhất thông thường!

Hãy để chúng tôi tiếp tục với ví dụ trên:

**bảng 2**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **X** | **Y** | **XY** | **X-X '** | **Y-Y '** | **(X-X ') (Y-Y')** | **(X-X ') 2** | **(Y-Y ') 2** |
|  | 1 | 1 | 1 | -2 | -1,16 | 2,32 | 4 | 1.346 |
|  | 2 | 1,5 | 3 | -1 | -0,66 | 0,66 | 1 | 0,436 |
|  | 3 | 2,5 | 7,5 | 0 | 0,34 | 0,34 | 0 | 0,116 |
|  | 4 | 2,8 | 11,2 | 1 | 0,64 | 0,64 | 1 | 0,410 |
|  | 5 | 4 | 20 | 2 | 1,84 | 3,68 | 4 | 3.386 |
| **Tổng** | 15 | 10,8 | 42,7 |  |  | 7.64 | 10 | 5,69 |
| **Nghĩa là** | X '= 3 (15/5) | Y '= 2,16 (10,8 / 5) |  |  |  |  |  |  |

  
**Đồ thị 2**

Phương trình chính là: Y = a + b (X) + e (thuật ngữ lỗi)

Trong trường hợp này, e bằng 0 vì giả thiết rằng biến độc lập (X) có sai số không đáng kể. Do đó, nó vẫn là Y = a + b (X). Bây giờ chúng ta hãy tìm giá trị của b.

b = [∑ XY - (∑Y) (∑X) / n] / ∑ (X-X ') 2

**Thay thế các giá trị trong công thức trên:**

b = [42,7 - (15 \* 10,8) / 5] / 10 = [42,7 - 162/5] / 10 = [42,7 - 32,5] / 10 = 10,2 / 10 = 1,02

**b = 1,02**

**Do đó** ,

a = Y - b (X)

a = Y - 1,02 (X) hoặc a = ∑Y / n - 1,02 (∑X / n)

a = 2,16 - 1,02 \* 3 = 2,16 - 3,06

**a = -1,06**

**Bằng cách thay giá trị của a, b và X, chúng ta có thể tìm thấy những giá trị tương ứng của Y** .

Y = a + b (X) = -1,06 + 1,02X

Khi X = 1

Y = -1,06 + 1,02 \* 1 = -0,04

Khi X = 2

Y = - 1,06 + 1,02 \* 2 = 0,98

Khi X = 3, Y sẽ là 2

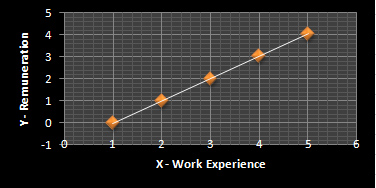
Khi X = 4, Y sẽ là 3,02

Khi X = 5, Y sẽ là 4,04

**bảng 3**

|  |  |
| --- | --- |
| X | Y |
| 1 | -0.04 |
| 2 | 0,98 |
| 3 | 2 |
| 4 | 3.02 |
| 5 | 4.04 |

Đường hồi quy phù hợp nhất sẽ là:

  
**Đồ thị 3**

**Thuộc tính của Đường hồi quy / Công cụ ước tính phù hợp nhất**

* Đường hồi quy đi qua X ', là 3 trong trường hợp này. (Tham khảo Đồ thị 3)
* b, hệ số hồi quy của X, là mức thay đổi trung bình trong Y. Trong trường hợp này, b = 1,02 là mức thay đổi trung bình trong các giá trị của y: -0,04, 0,98, 2, 3,02 và 4,04. (Tham khảo Bảng 3)
* Đường hồi quy đi qua Y ', trong trường hợp này là 2,16. (Tham khảo Đồ thị 3)

**5. Hệ số xác định - Đánh giá mức độ phù hợp**

Bất cứ khi nào bạn sử dụng mô hình hồi quy, điều đầu tiên bạn nên xem xét là - mô hình kinh tế lượng phù hợp với dữ liệu như thế nào hoặc phương trình hồi quy phù hợp với dữ liệu như thế nào.

Đây là nơi xuất hiện khái niệm ***hệ số xác định*** . Các mô hình hồi quy thường được phù hợp bằng cách sử dụng cách tiếp cận này.

Nó xác định mức độ mà biến phụ thuộc có thể được dự đoán từ biến độc lập. Nó đánh giá mức độ phù hợp của Mô hình hồi quy bình phương tối thiểu thông thường.

Ký hiệu là R 2 , giá trị của nó nằm trong khoảng từ 0 đến 1. Khi:

* R 2 = 0: giá trị của biến phụ thuộc không thể dự đoán được từ biến độc lập.
* R 2 = 1, giá trị của biến phụ thuộc có thể dễ dàng dự đoán từ biến độc lập. Không có lỗi trong dữ liệu.

Giá trị R 2 càng cao thì mô hình càng phù hợp với dữ liệu.

Bây giờ chúng ta hãy hiểu cách tính R 2 . Công thức tìm R 2 là:

R 2 = {(1 / n) \* ∑ (X-X ') \* (Y-Y')} / (σx \* σy) 2

Trong đó n = số lần quan sát = 5

∑ (X-X ') \* (Y-Y') = 7,64 (Tham khảo Bảng 2)

σx là độ lệch chuẩn của X và σy là độ lệch chuẩn của Y

σx = căn bậc hai của ∑ (X-X ') 2 / n = √10 / 5 = √ 2 = 1.414

σy = căn bậc hai của ∑ (Y-Y ') 2 / n = √5,69 / 5 = √1.138 = 1,067

Bây giờ hãy xác định giá trị của R 2

R 2 = {1/5 (7,64)} / (1,414 \* 1,067) 2 = 1,528 / (1,509) 2 = 1,528 / 2,277 = 0,67

Do đó **R 2 = 0,67**

Giá trị của hệ số xác định càng cao thì sai số tiêu chuẩn càng thấp. Kết quả chỉ ra rằng khoảng 67% sự thay đổi trong thù lao có thể được giải thích bởi kinh nghiệm làm việc. Nó cho thấy rằng kinh nghiệm làm việc đóng một vai trò quan trọng trong việc xác định mức thù lao.

**Thuộc tính của Công cụ ước tính**

* b là không thiên vị và độc lập.
* σx là không thiên vị vì tính ngoại đồng nghiêm ngặt. Trong trường hợp ngoại sinh không nghiêm ngặt, giá trị sẽ bị sai lệch trong các mẫu hữu hạn.
* (σy) 2 là sai lệch nhưng căn bậc hai giảm thiểu sai số.

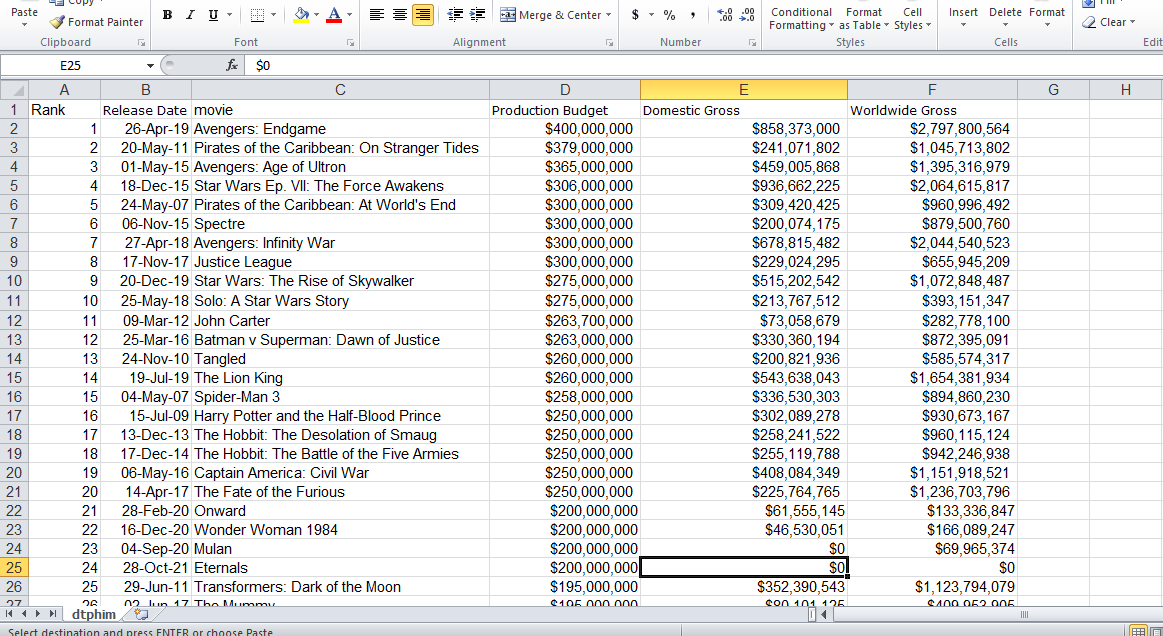
Mô hình hồi quy tuyến tính được sử dụng khi có mối quan hệ tuyến tính giữa các biến phụ thuộc và độc lập. Khi giá trị của một biến phụ thuộc dựa trên nhiều biến (nhiều hơn một), chúng tôi sử dụng phân tích hồi quy nhiều lần.

**CHƯƠNG 2. SỬ DỤNG THUẬT TOÁN HỒI QUY TUYẾN TÍNH ÁP DỤNG CHO BÀI TOÁN CỤ THỂ**

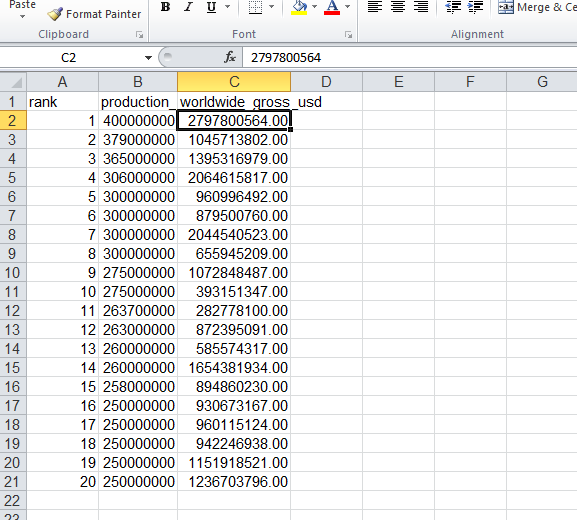
**I. Sử dụng thuật toán hồi quy tuyến tính áp dụng cho dự đoán doanh thu phim**

**1.1. Dữ liệu bài toán**

Bộ dữ liệu của bài toán dự đoán phim bao gồm có ngày sản xuất, tên phim, chi phí phim, doanh thu trong nước, doanh thu toàn cầu của các bộ phim khác nhau và được lấy từ trang https://www.the-numbers.com/ movie/budgets thống kê về Hồ sơ Ngân sách và Hiệu suất Tài chính Phim, bộ dữ liệu ban đầu gồm có như hình sau:



Sau khi chỉnh sửa lại bảng dữ liệu và một số ký tự đặc biệt để tránh phát sinh lỗi trong quá trình dự đoán ta được file dữ liệu sau :



**1.2. Cài đặt chương trình**

- Khai báo thư viện và đọc file từ bộ dữ liệu và in ra màn hình

data = pd.read\_csv("dtp.csv")

data.describe()

print(data)

print(data.keys())

print(data.shape)

print(data.describe())

**-** Trước khi chia dữ liệu thành tập dữ liệu để train - test, chúng ta cần chia dữ liệu thành hai giá trị : giá trị đích và giá trị dự báo. Hãy gọi giá trị đích Y và các giá trị dự báo X. Như vậy:

Y = DataFrame(data, columns=['worldwide\_gross\_usd'])

X = DataFrame(data, columns=['production\_budget\_usd'])

- Khởi tạo bộ dữ liệu train – test và in ra màn hình

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.33, random\_state=5)

print(X\_train.shape)

print(X\_test.shape)

print(Y\_train.shape)

print(Y\_test.shape)

- Nếu chúng ta kiểm tra shape của mỗi biến, chúng ta đã có được bộ dữ liệu với tập dữ liệu thử nghiệm có tỷ lệ 66,66% đối với dữ liệu train và 33,33% đối với dữ liệu test.

-Khởi tạo mô hình hồi quy tuyến tính và tính sai số

regression = LinearRegression()

regression.fit(X, Y)

regression.coef\_

regression.intercept\_

regr = linear\_model.LinearRegression().fit(X\_train, Y\_train)

print('Hệ số W: \n', regr.coef\_)

print('Hệ số Bias: \n', regr.intercept\_)

Y\_pred = regr.predict(X\_test)

print('Sai số - bình phương sai số : %.2f '% mean\_squared\_error(Y\_test, Y\_pred))

print('Sai số - Hệ số xác định: %.2f'% r2\_score(Y\_test, Y\_pred))

- Tiếp theo, chúng ta thực hiện vẽ biểu đồ với dữ liệu và ngân sách và doanh thu phim

plt.scatter(X, Y, alpha=0.3)

plt.scatter(Y\_test, Y\_pred)

plt.title('film cost vs global revenue')

plt.xlabel('Production Budget $')

plt.ylabel('Worldwide Gross $')

plt.figure(figsize=(10,6))

plt.scatter(X, Y, alpha=0.3)

plt.plot(X, regression.predict(X), color = 'red', linewidth = 4)

plt.title('film cost vs global revenue')

plt.xlabel('Production Budget $')

plt.ylabel('Worldwide Gross $')

plt.ylim(0, 3000000000)

plt.xlim(200000000, 450000000)

plt.show()

- Tính lỗi bình phương trung bình (RMSE)

rms = numpy.sqrt(mean\_squared\_error(Y\_test, Y\_pred))

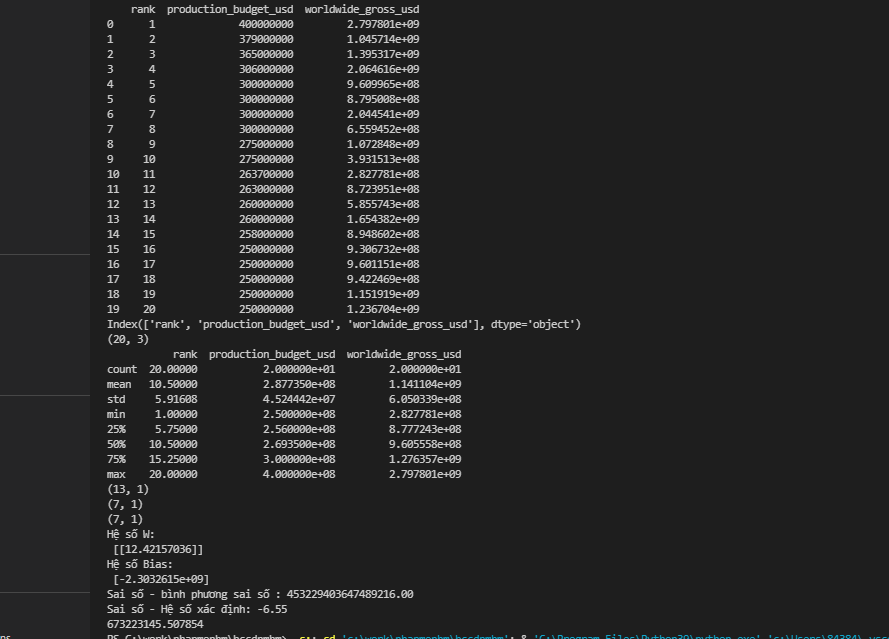
print(rms)

**1.3. Kết quả**

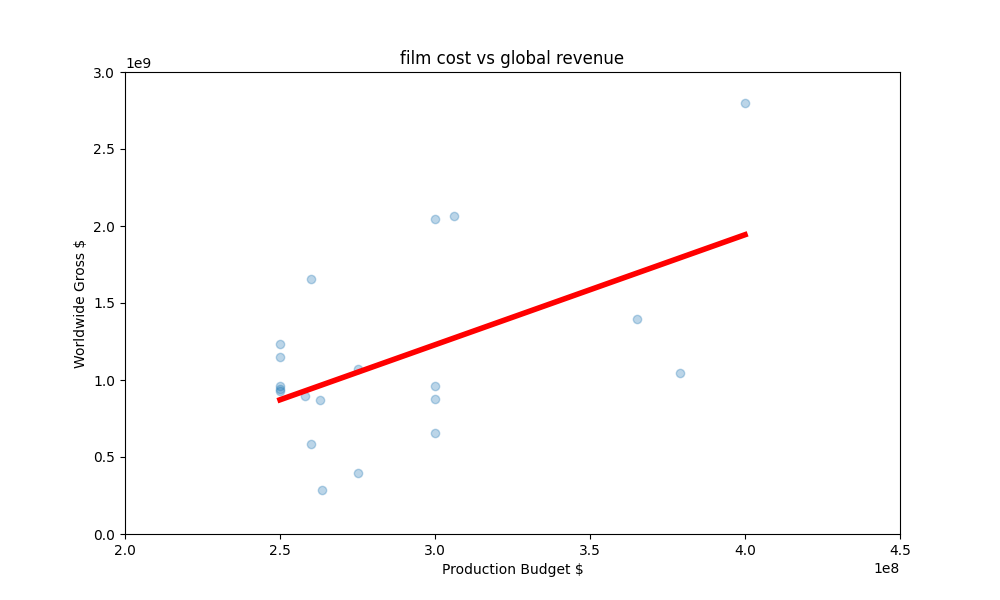
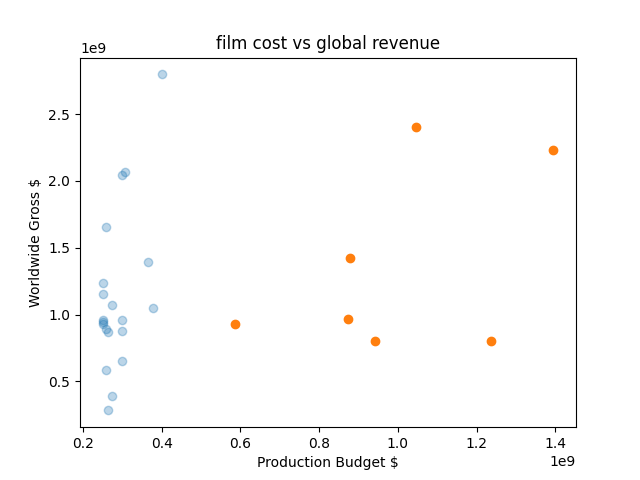
- Hiển thị giá trị của bộ dữ liệu

- Tổng số hàng cột và các dữ liệu của bộ dữ liệu

- Hệ số và sai số của chương trình



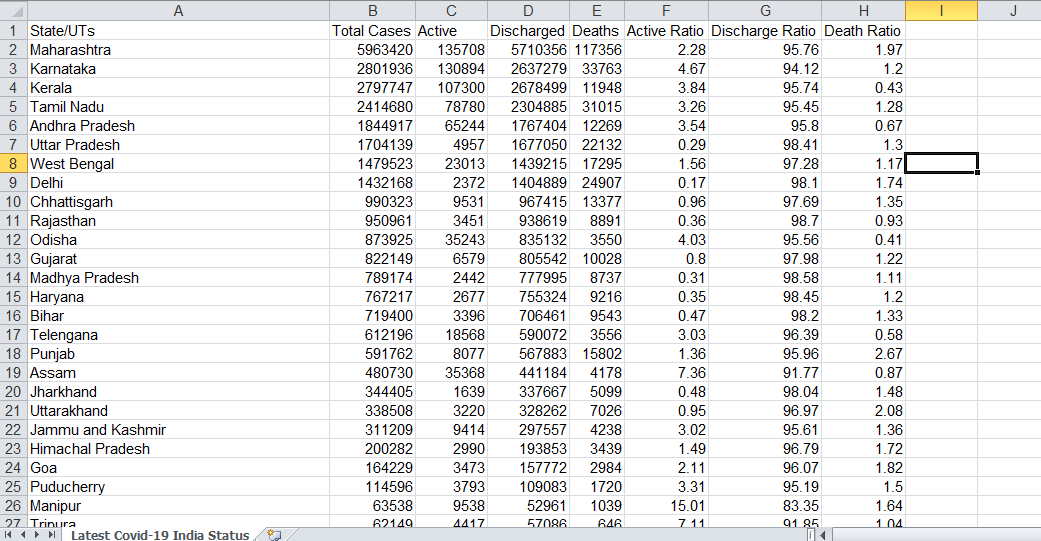
- Sơ đồ biểu diễn



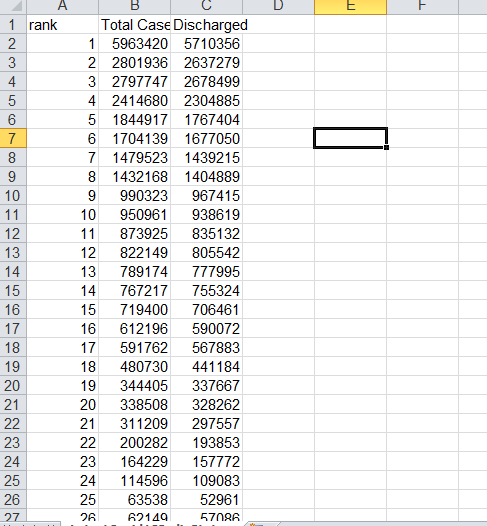
**2. Sử dụng thuật toán hồi quy tuyến tính áp dụng cho trạng thái mới nhất của Covid-19 tại Ấn Độ**

**2.1. Dữ liệu bài toán**

Bộ dữ liệu của bài toán trạng thái mới nhất của Covid-19 Ấn Độ bao gồm có các ca nhiễm, các ca khỏi, số liệu tử vong,…. của các bang khác nhau và được lấy từ trang https://www.kaggle.com/datasets thống kê về tình hình covid-19 tại ấn độ, bộ dữ liệu ban đầu gồm có như hình sau:



Sau khi chỉnh sửa lại bảng dữ liệu và một số ký tự đặc biệt để tránh phát sinh lỗi trong quá trình dự đoán ta được file dữ liệu sau :



**2.2. Cài đặt chương trình**

- Khai báo thư viện và đọc file từ bộ dữ liệu và in ra màn hình

data = pd.read\_csv("LatestCovid19IndiaStatus.csv")

data.describe()

print(data)

print(data.keys())

print(data.shape)

print(data.describe())

**-** Trước khi chia dữ liệu thành tập dữ liệu để train - test, chúng ta cần chia dữ liệu thành hai giá trị : giá trị đích và giá trị dự báo. Hãy gọi giá trị đích Y và các giá trị dự báo X. Như vậy:

Y = DataFrame(data, columns=['Discharged'])

X = DataFrame(data, columns=['Total Cases'])

- Khởi tạo bộ dữ liệu train – test và in ra màn hình

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.30, random\_state=4)

print(X\_train.shape)

print(X\_test.shape)

print(Y\_train.shape)

print(Y\_test.shape)

- Nếu chúng ta kiểm tra shape của mỗi biến, chúng ta đã có được bộ dữ liệu với tập dữ liệu thử nghiệm có tỷ lệ 70% đối với dữ liệu train và 30% đối với dữ liệu test.

-Khởi tạo mô hình hồi quy tuyến tính và tính sai số

regression = LinearRegression()

regression.fit(X, Y)

regression.coef\_

regression.intercept\_

regr = linear\_model.LinearRegression().fit(X\_train, Y\_train)

print('Hệ số W: \n', regr.coef\_)

print('Hệ số Bias: \n', regr.intercept\_)

Y\_pred = regr.predict(X\_test)

print('Sai số - bình phương sai số : %.2f '% mean\_squared\_error(Y\_test, Y\_pred))

print('Sai số - Hệ số xác định: %.2f'% r2\_score(Y\_test, Y\_pred))

- Tiếp theo, chúng ta thực hiện vẽ biểu đồ với số ca khỏi bệnh và số ca nhiễm bệnh

plt.scatter(X, Y, alpha=0.4)

plt.scatter(Y\_test, Y\_pred)

plt.title('Total Cases vs Discharged in Idian')

plt.xlabel('Total Cases')

plt.ylabel('Discharged')

plt.figure(figsize=(10,6))

plt.scatter(X, Y, alpha=0.4)

plt.plot(X, regression.predict(X), color = 'red', linewidth = 3)

plt.title('Total Cases vs Discharged in Idian')

plt.xlabel('Total Cases')

plt.ylabel('Discharged')

plt.ylim(0, 7000000)

plt.xlim(0, 7000000)

plt.show()

- Tính lỗi bình phương trung bình (RMSE)

rms = numpy.sqrt(mean\_squared\_error(Y\_test, Y\_pred))

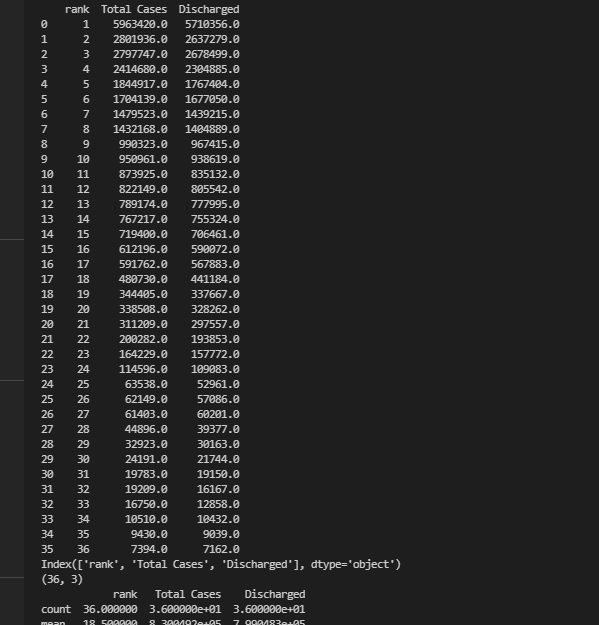
print(rms)

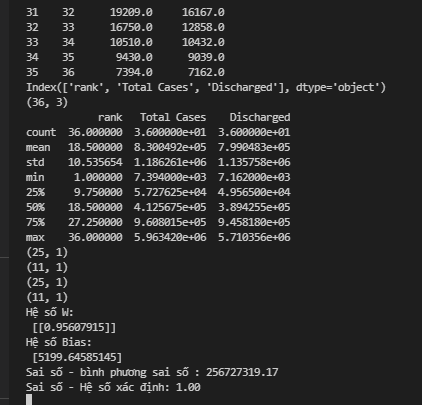
**2.3. Kết quả**

- Hiển thị giá trị của bộ dữ liệu

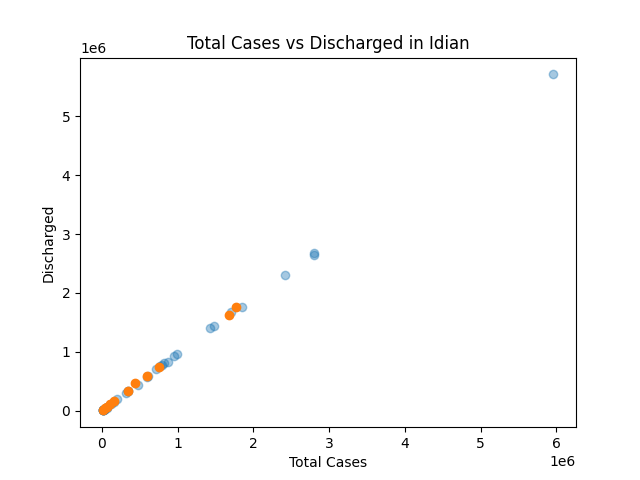
- Tổng số hàng cột và các dữ liệu của bộ dữ liệu

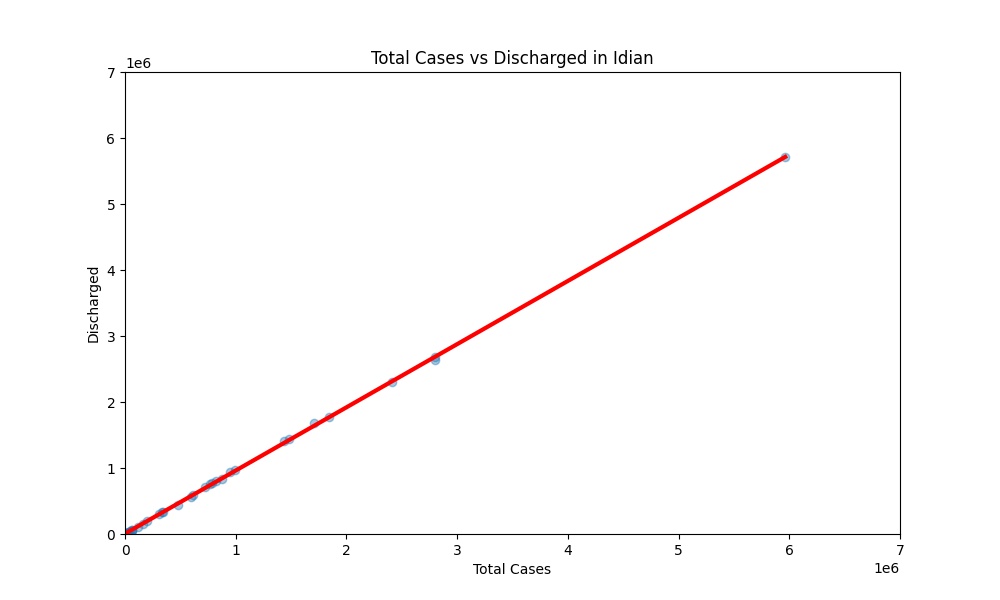
- Hệ số và sai số của chương trình





- Sơ đồ biểu diễn

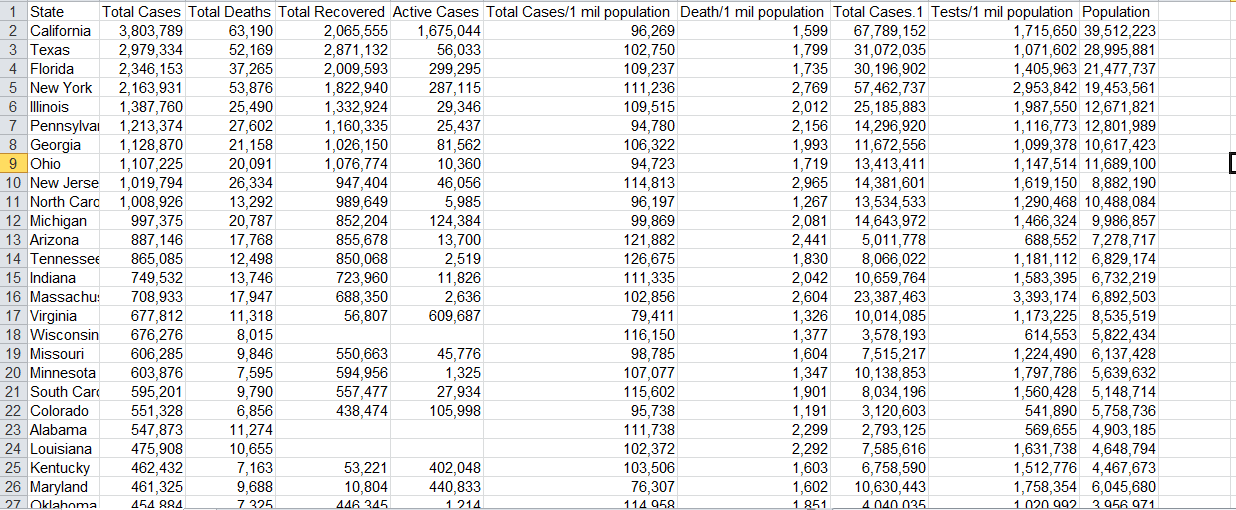




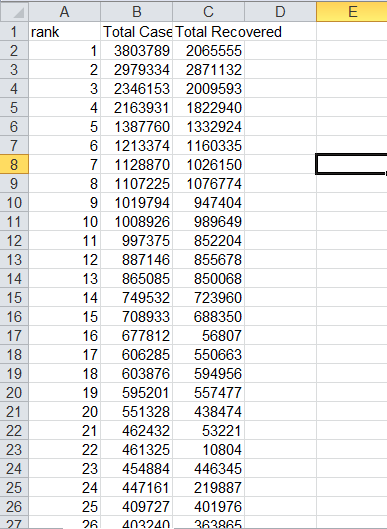
**3. Sử dụng thuật toán hồi quy tuyến tính áp dụng cho trạng thái mới nhất của Covid-19 tại USA**

**3.1. Dữ liệu bài toán**

Bộ dữ liệu của bài toán cho trạng thái mới nhất của Covid-19 tại USA bao gồm tổng số ca nhiễm, số ca tử vong, số ca khỏi bệnh,…. và được lấy từ trang https://www.kaggle.com/datasets thống kê trạng thái mới nhất của Covid-19 tại USA, bộ dữ liệu ban đầu gồm có như hình sau:



Sau khi chỉnh sửa lại bảng dữ liệu và một số ký tự đặc biệt để tránh phát sinh lỗi trong quá trình dự đoán ta được file dữ liệu sau :



**3.2. Cài đặt chương trình**

- Khai báo thư viện và đọc file từ bộ dữ liệu và in ra màn hình

data = pd.read\_csv("USAcoviddata.csv")

data.describe()

print(data)

print(data.keys())

print(data.shape)

print(data.describe())

**-** Trước khi chia dữ liệu thành tập dữ liệu để train - test, chúng ta cần chia dữ liệu thành hai giá trị : giá trị đích và giá trị dự báo. Hãy gọi giá trị đích Y và các giá trị dự báo X. Như vậy:

Y = DataFrame(data, columns=['Total Recovered'])

X = DataFrame(data, columns=['Total Cases'])

- Khởi tạo bộ dữ liệu train – test và in ra màn hình

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.20, random\_state=5)

print(X\_train.shape)

print(X\_test.shape)

print(Y\_train.shape)

print(Y\_test.shape)

- Nếu chúng ta kiểm tra shape của mỗi biến, chúng ta đã có được bộ dữ liệu với tập dữ liệu thử nghiệm có tỷ lệ 80% đối với dữ liệu train và 20% đối với dữ liệu test.

-Khởi tạo mô hình hồi quy tuyến tính và tính sai số

regression = LinearRegression()

regression.fit(X, Y)

regression.coef\_

regression.intercept\_

regr = linear\_model.LinearRegression().fit(X\_train, Y\_train)

print('Hệ số W: \n', regr.coef\_)

print('Hệ số Bias: \n', regr.intercept\_)

Y\_pred = regr.predict(X\_test)

print('Sai số - bình phương sai số : %.2f '% mean\_squared\_error(Y\_test, Y\_pred))

print('Sai số - Hệ số xác định: %.2f'% r2\_score(Y\_test, Y\_pred))

- Tiếp theo, chúng ta thực hiện vẽ biểu đồ với dữ liệu số ca nhiễm và số ca khỏi bệnh

plt.scatter(X, Y, alpha=0.2)

plt.scatter(Y\_test, Y\_pred)

plt.title('Total Cases vs Discharged in USA')

plt.xlabel('Total Cases')

plt.ylabel('Total Recovered')

plt.figure(figsize=(10,6))

plt.scatter(X, Y, alpha=0.2)

plt.plot(X, regression.predict(X), color = 'red', linewidth = 4)

plt.title('Total Cases vs Discharged in USA')

plt.xlabel('Total Cases')

plt.ylabel('Total Recovered')

plt.ylim(0, 4000000)

plt.xlim(0, 4000000)

plt.show()

- Tính lỗi bình phương trung bình (RMSE)

rms = numpy.sqrt(mean\_squared\_error(Y\_test, Y\_pred))

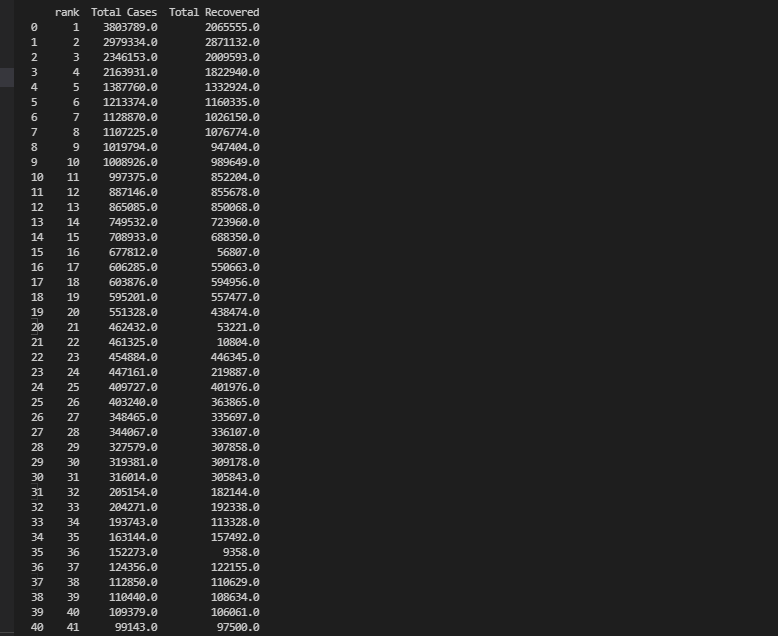
print(rms)

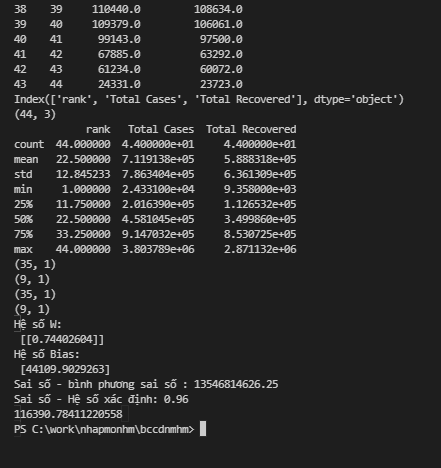
**3.3. Kết quả**

- Hiển thị giá trị của bộ dữ liệu

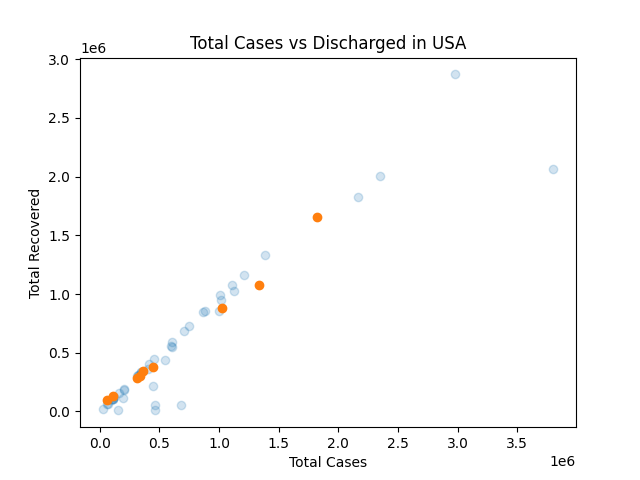
- Tổng số hàng cột và các dữ liệu của bộ dữ liệu

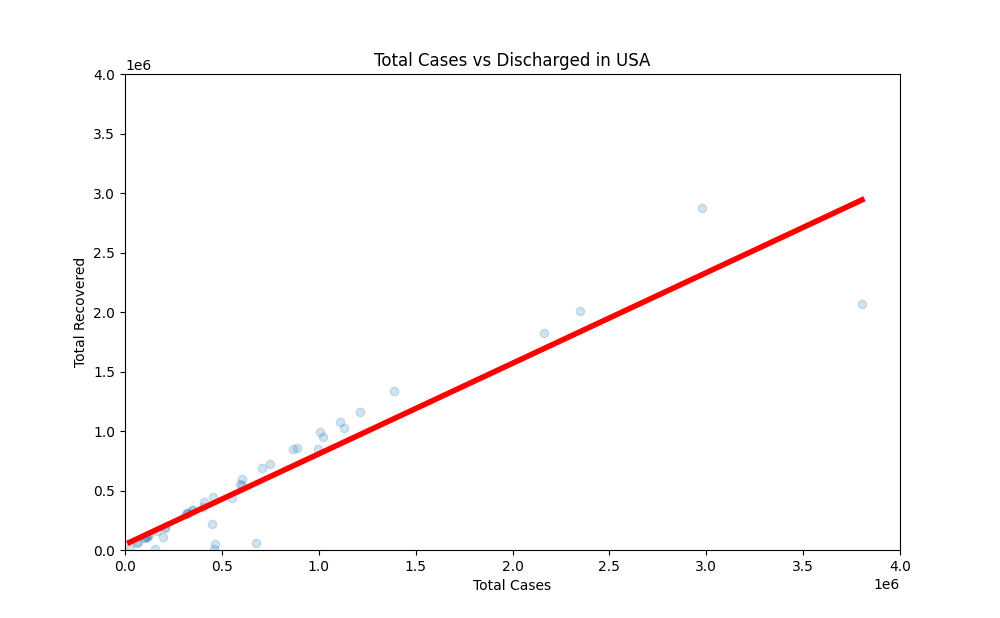
- Hệ số và sai số của chương trình





- Sơ đồ biểu diễn





**4. Đánh giá**

- Ưu điểm:

* Nhanh chóng để mô hình hóa và đặc biệt hữu ích khi mối quan hệ được mô hình hóa không quá phức tạp và nếu bạn không có nhiều dữ liệu.
* Hồi quy tuyến tính là đơn giản để hiểu, nó rất có giá trị cho các quyết định kinh doanh.

- Nhược điểm:

* Đối với dữ liệu phi tuyến tính, hồi quy đa thức có thể khá khó khăn để thiết kế, vì người ta phải có một số thông tin về cấu trúc của dữ liệu và mối quan hệ giữa các biến tính năng.
* Kết quả của những điều trên, các mô hình này không tốt như các mô hình khác khi nói đến dữ liệu rất phức tạp.

**KẾT LUẬN**

Trong thời gian và điều kiện nhất định, mặc dù đã rất cố gắng để hoàn thành đồ án với tất cả sự nỗ lực. Tuy nhiên, do bước đầu đi vào thực tế, tìm hiểu và xây dựng báo cáo trong thời gian có hạn, với lượng kiến thức còn hạn chế, nhiều bỡ ngỡ, nên bài làm của chúng em không thể tránh khỏi những thiếu sót. Chúng em rất mong nhận được sự quan tâm, thông cảm và những đóng góp quý báu của các thầy cô và các bạn để đồ án này ngày càng hoàn thiện hơn.

Chúng em xin gửi lời cảm ơn đến giáo viên giảng dạy bộ môn Nhập môn học máy – thầy Ngô Hoàng Huy. Cô đã tận tình chỉ dạy, quan sát và giúp đỡ em từ những ngày đầu cũng như trong quá trình nghiên cứu và học tập.

Chúng em xin chân thành cảm ơn!