**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**ĐÁNH GIÁ CÁC YẾU TỐ CÁCH HƯỞNG ĐẾN RATING CỦA MỘT ĐẦU SÁCH**

| Sinh viên thực hiện: **(KHÔNG ghi GVHD)** | | |
| --- | --- | --- |
| STT | Họ tên | MSSV |
| 1 | Nguyễn Thành Phúc | 19522040 |
| 2 | Trần Nhật Nam | 19521872 |
| 3 | Nguyễn Võ Thiên Ân | 19521186 |

**TP. HỒ CHÍ MINH – 12/2020**

**MỤC LỤC**

[**1. GIỚI THIỆU**](#_heading=h.gjdgxs) **1**

[**2. NỘI DUNG**](#_heading=h.30j0zll) **1**

[Thu thập & tiền xử lý dữ liệu](#_heading=h.1fob9te) 1

[Thu thập dữ liệu](#_heading=h.3znysh7) 1

[Xử lý bộ dữ liệu](#_heading=h.tyjcwt) 2

[Bộ dữ liệu](#_heading=h.1t3h5sf) 4

[Phân tích thăm dò](#_heading=h.4d34og8) 5

[Các thuộc tính phân loại](#_heading=h.2s8eyo1) 5

[Các thuộc tính liên tục.](#_heading=h.26in1rg) 6

[Mô hình dự đoán](#_heading=h.1ksv4uv) 8

[Xây dựng mô hình](#_heading=h.44sinio) 8

[Đánh giá mô hình](#_heading=h.z337ya) 9

[**3. KẾT LUẬN**](#_heading=h.4i7ojhp) **10**

# 1. GIỚI THIỆU

Đề tài của chúng tôi là Đánh giá các yếu tố ảnh hưởng đến rating của một đầu sách trên bộ dữ liệu tự thu thập từ trang **Goodread.com**[[1]](#footnote-0) và từ đó xây dựng một mô hình dự đoán rating của người dùng với một đầu sách trên trang web này. Bộ dữ liệu có hơn 8000 đầu sách được thu thập từ list ***Best Books of the 21st Century:*** *The best books published during the 21st century (January 1st, 2001 through December 31st, 2100).*[1]

Chúng tôi chọn thu thập các đầu sách từ list này vì 2 nguyên do: thứ nhất, các đầu sách thuộc list này là những sách được xuất bản trong vòng 20 năm trở lại đây, đảm bảo được tính “mới” của dữ liệu để phù hợp với mục tiêu xây dựng mô hình dự đoán; thứ hai là do bản chất một *list* của Goodreads là bao gồm các đầu sách do người dùng bình chọn, điều này có nghĩa là các đầu sách đã có một lượng người đọc nhất định, tránh trường hợp thu thập phải những cuốn sách “quá mới” cũng gây ảnh hưởng đến việc xây dựng mô hình vì có thể thiếu các yếu tố quan trọng.

Bộ dữ liệu sử dụng thư viện *BeautifulSoup*[1]của ngôn ngữ Python để thu thập. Sau khi tiền xử lý xóa bớt các thuộc tính ít quan trọng và chuẩn hóa, bộ dữ liệu có 14 thuộc tính với 8000 dòng dữ liệu. Mục tiêu là phân tích và tìm ra các thuộc tính quan trọng, có độ tương quan so với thuộc tính rating, từ đó phát triển mô hình dự đoán rating cho ra kết quả có độ chính xác cao, sau cùng là sử dụng các độ đo để đánh giá hiệu suất của mô hình.

# 2. NỘI DUNG

## Thu thập & tiền xử lý dữ liệu

Diagram

Description automatically generated

***Hình 2. Quy trình phân tích dữ liệu***

### Thu thập dữ liệu

Bộ dữ liệu được thu thập bằng các thư viện *BeautifulSoup, Numpy* và *Pandas* của ngôn ngữ python, thực hiện trên môi trường Google Colab.

A picture containing diagram

Description automatically generated

***Hình 2.a. Quy trình thu thập dữ liệu.***

Đầu tiên ta truy cập vào list *Best Books of the 21st Century*trên trang Goodreads. List này bao gồm 9,198 đầu sách đáp ứng đủ hai tiêu chí: được bình chọn bởi một lượng người dùng nhất định và được xuất bản trong thế kỷ 21; danh sách này được cập nhật liên tục. Số sách được chia thành 91 trang (pagination) với 100 sách cho mỗi trang, tại đây ta thu thập được 9,198 đường dẫn đến đầu sách. Sau khi đã có các đường dẫn đến từng cuốn sách, ta tiến hành thu thập dữ liệu cụ thể của từng đầu sách, dữ liệu thu được ở dạng dataframe có độ lớn 9,198 x 19 cột như sau:

***Bảng 1. Bộ dữ liệu thô.***

| *id* | Mã id | *review\_count* | Số lượt bình luận | *isbn* | Mã tiêu chuẩn quốc tế |
| --- | --- | --- | --- | --- | --- |
| *title* | Tên sách | *numbers\_of\_pages* | Số trang | *isbn13* | Mã vạch |
| *link* | Link sách | *date\_published* | Ngày xuất bản | *setting* | Bối cảnh trong sách |
| *author* | Tác giả | *publisher* | Nhà xuất bản | *characters* | Nhân vật trong sách |
| *author\_link* | Link tác giả | *original\_title* | Tên tác phẩm gốc | *description* | Mô tả |
| *rating\_count* | Số lượt vote | *genre\_and\_votes* | Thể loại và lượt vote | Kích thước bộ dữ liệu:  9,098 x 19 | |

### Xử lý bộ dữ liệu

Phần tiền xử lý gồm các bước:

* Xóa cột dư thừa & cột có tỷ lệ missing data cao
* Điền các giá trị bị khuyết
* Chuẩn hóa một số cột.

Có một thuộc tính quan trọng và khá phức tạp chúng tôi cần giải thích cụ thể.

*genre\_and\_votes:* đây là cột phức tạp nhất của bộ dữ liệu. Thể loại của một cuốn sách được trang web xác định bằng cách thu thập dữ liệu của cộng đồng từ kệ sách của người dùng, và kệ sách này mỗi người dùng có thể tạo và đặt tên cho chúng. Trang web sẽ sử dụng thuật toán để xếp hạng các thể loại dựa trên các kệ sách của người dùng có trùng tên mà cuốn sách được thêm vào kệ sách đó[3]. Cơ chế này dẫn đến việc một cuốn sách nhưng có thể có từ vài thể loại đến hơn vài trăm ngàn thể loại. Bản thân trang Goodreads cũng tự nhận xét rằng đây là phương pháp chưa tối ưu, song theo cơ chế này, các thể loại có nhiều lượt bình chọn nhất có độ chính xác cao nhất. Xử lý cột này có hai bước, bước đầu ta chọn thể loại có bình chọn cao nhất, ví dụ:

*Fantasy 51882, Young Adult 16499,… 🡪 Fantasy*

Vì chỉ còn chứa 1 thể loại duy nhất cho mỗi cuốn sách, ta đổi tên cột thành *genre.*

Tiếp theo đó, dữ liệu sau khi format lại như trên có 196 thể loại, số lượng giá trị quá nhiều dẫn đến khó phân tích sau này, đầu tiên ta gộp các thể loại con vào thể loại chính của nó.

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

Description automatically generated

***Hình 2.b. Thuộc tính thể loại và thể loại con.***

Sau khi gộp xong, ta tiến hành thu hẹp phạm vi cột này bằng cách xem xét các thể loại hiếm (chỉ có số rất ít sách thuộc thể loại đó) bằng cách đồng kiểm lại trên Goodreads và chọn thể loại khác nằm trong top 3.

Đối với một số thể loại có bản chất tương đồng nhau, ta gộp chung các thể loại đó với nhau.

Các giá trị bị NaN là các sách chưa được một lượng người dùng nhất định xếp vào các kệ sách của họ, dẫn đến việc trang web bị thiếu cơ sở để xếp loại thể loại cho cuốn sách, vì vậy ta gán các giá trị bị thiếu bằng None.

Tổng số các thể loại sau khi xử lý là 63.

## Bộ dữ liệu

Giới thiệu bộ dữ liệu sạch sau khi tiền xử lý và chuẩn hóa:

* **Nguồn dữ liệu:** Bộ dữ liệu được thu thập trực tiếp từ Goodreads.com
* **Ý nghĩa bộ dữ liệu:** Bộ dữ liệu chứa thông tin cơ bản về một đầu sách và các đánh giá, bình luận của độc giả về cuốn sách đó.
* **Số thuộc tính:** 14
* **Số điểm dữ liệu:** 8,133
* **Dữ liệu bị khuyết:** Không có dữ liệu bị khuyết

***Bảng 2. Bảng thống kê thông tin các thuộc tính bộ dữ liệu***

| STT | Tên thuộc tính | Kiểu dữ liệu | Miền giá trị |
| --- | --- | --- | --- |
| 1 | id | int64 | liên tục từ 1 đến 59008505 |
| 2 | title | object | Harry Potter and the Deathly Hallows, The Hunger Games, The Kite Runner, etc. |
| 3 | series | object | Harry Potter, The Hunger Games, Single, etc |
| 4 | author | object | J.K. Rowling, Suzanne Collins, Khaled Hosseini, etc. |
| 5 | rating\_count | float64 | liên tục từ 0.0 đến 7026399.0 |
| 6 | review\_count | float64 | liên tục từ 0.0 đến 179928.0 |
| 7 | number\_of\_pages | float64 | liên tục từ 0.0 đến 4081.0 |
| 8 | date\_published | float64 | liên tục từ 1972.0 đến 2021.0 |
| 9 | publisher | object | Arthur A. Levine Books, Scholastic Press, Riverhead Books, etc. |
| 10 | genre | object | Fantasy, Young Adult, Fiction, etc. |
| 11 | isbn | int64 | 0, 1 |
| 12 | description | int64 | liên tục từ 0 đến 25116 |
| 13 | awards | int64 | liên tục từ 0 đến 40 |
| 14 | rating | float64 | liên tục từ 0.0 đến 5.0 |

## Phân tích thăm dò

### Các thuộc tính phân loại

Trong quá trình phân tích các thuộc tính phân loại chúng tôi nhận thấy tất cả các thuộc tính đều có sự ảnh hưởng đến rating những do giá trị của các thuộc tính quá nhiều cho nên tính phân loại bị giảm. Điều này cũng dễ giải thích do những thuộc tính như: Series, author, publisher thì mỗi sách hầu như được viết bởi những các giả khác nhau thuộc những series khác nhau và xuất bản khác nhau[2]. Trên thế thới có rất nhiều tác giả cho nên những thuộc tính này bị giảm tính phân loại đi. Mặc dù vậy ở trong đối tượng sách thuộc top đánh giá của năm 2021 thì rất có thể sẽ có những tác giả nổi bất cùng với những series đặc biệt, những nhà xuất bản có uy tín sẽ được đánh giá cao.

Với những ý nghĩ từ việc xử lý ở trên chúng tôi dự đoán thuộc tính *genre* là thuộc tính phân loại có số giá trị thấp nhất. Tính phân loại có thể là cao nhất, và được coi là thuộc tính chúng tôi kỳ vọng nhất trong bài toán.

Chart, histogram

Description automatically generated

***Hình 2.c. Biểu đồ phân bố thuộc tính thể loại.***

Thuộc tính thể loại có tính chất khá phức tạp và độ dàn trải các giá trị không đều, lên tới 63 giá trị dù đã qua bước xử lý thủ công. Ta thấy các thể loại có nhiều sách nhất thường là các thể loại hướng giải trí nhiều hơn hướng học thuật.

Chart, box and whisker chart

Description automatically generated

***Hình 2.d. Biểu đồ thể hiện sự phân bố của biến thể loại so với rating***

Biểu đồ hộp cho thấy các thể loại có sự phân bố rating khác nhau: thể loại Young Adult (Tiểu thuyết dành cho người trẻ) và Fiction (Tiểu thuyết hư cấu) có lượng sách lớn, mục tiêu hướng tới các độc giả trẻ, rating của các đầu sách có thể loại này nằm ở mức trung bình; các đầu sách thuộc thể loại Historical (Tiểu thuyết có yếu tố lịch sử) và Fantasy (Phiêu lưu), Science Fiction (Khoa học viễn tưởng) có rating nằm ở mức trung bình khác và cao; các đầu sách bị khuyết thuộc tính thể loại là các đầu sách khá “mới”, chưa có một lượng độc giả xếp vào kệ sách để trang web lấy đó làm cơ sở xếp hạng thể loại, các đầu sách này thường có số lượng rating (rating\_count) thấp, rating cũng nằm ở mức trung bình.

Không nằm ngoài dự đoán thì genre ảnh hưởng lớn đến rating , đồng thời là thuộc tính quan trọng trong mô hình của chúng tôi.

### Các thuộc tính liên tục.

Quan sát độ tương quan giữa các biến liên tục.

Graphical user interface, application

Description automatically generated

***Hình 2.e. Độ tương quan giữa các biến liên tục***

Dựa vào đồ thị tương quan cùng trong quá trình phân tích chúng tôi không nhận thấy sự tương quan giữa các biến liên tục so với rating. Có thể nhận thấy các thuộc tính này không ảnh hưởng đến rating. Chúng ta có thể giải thích như sau. Với các thuộc tính *rating\_count, review\_count* những sách nội dung tệ vẫn có thể thu hút số lượng lớn người dùng vào đánh giá, thậm chí những quyển sách gây ra mâu thuẫn lớn cũng có thể được rất nhiều người tò mò vào để đọc và nhận đánh giá.

Thuộc tính *description*  mức độ ảnh hưởng cũng không có vì chúng tôi dựa trên những phân tích cụ thể với đồ thị. Những với số lượng dữ liệu lớn và số lượng mô tả cũng không nhiều, khi chúng tôi phân tích dưới dạng là thuộc tính phân loại thì vẫn thấy sự ảnh hưởng.

Cùng theo như sự dự đoán thì thuộc tích giải thưởng ít nhiều sẽ ảnh hưởng đến rating. Điều này cũng hợp lý vì những sách có giải thưởng thì phần lớn là những sách thu hút người đọc và nội dung hay.

Chart, box and whisker chart

Description automatically generated

***Hình 2.f. Đồ thị sự phân bố giữa giải thưởng và đánh giá***

Quan sát đồ thị thì những thuộc tính giải thưởng sẽ có ảnh hưởng đến rating ở những đầu sách có nhiều giải thưởng và từ 8 trở lên. Và đây là thuộc tính chúng tôi sử dụng để đưa vào mô hình.

Qua quá trình phân tích và thăm dò chúng tôi nhận thấy thì hầu hết các thuộc tính đều có ảnh hưởng ít đến rating và 2 trong số thuộc tính quan trọng là *genre* và *award* sẽ là thuộc tính để chúng tôi xây dựng các mô hình.

Cuối cùng chúng cũng với 2 thuộc tính quan trọng là các thuộc tính: *series, author, rating\_count, review\_count, number\_of\_pages,  publisher,  description*. Chúng tôi sẽ tiến hành xây dựng mô hình dự đoán rating.

## Mô hình dự đoán

### Xây dựng mô hình

Diagram

Description automatically generated

***Hình 2.g. Sơ đồ các bước xây dựng mô hình***

Do các cột không là số (series, publisher, genre) đều không là các giá trị binary hoặc có thể xác định trước nên không thể tách thành các cột. Chúng tôi xử lý bằng cách gán số cho nó. Mỗi unique sẽ được gán 1 con số riêng khác nhau bằng các sử dụng module *Pickle* của Python.

*Pickle là một module của Python dùng để chuyển đổi một đối tượng Python sang dạng binary và ngược lại. Sử dụng cách thức chuyển đổi riêng (chỉ dành riêng cho Python), không gắn với bất cứ một chuẩn chuyển đổi nào.Không có cơ chế định nghĩa dữ liệu tách biệt với việc mã hóa dữ liệu đó.*

Sau khi đã chuyển các giá trị object (chuỗi) sang các giá trị int (số) thông qua thư viện Pickle để dữ liệu có thể đưa vào mô hình, ta tiến hành chia tập huấn luyện & kiểm thử với tỷ lệ 80-20 để đưa vào mô hình huấn luyện. Chúng tôi sử dụng 3 mô hình từ thư viện *sklearn* của Python:

* Linear Regression.
* Support Vector Regression (C=0.5, gamma=0.1)
* Multi-Layer Perceptron Regressor (Neural Network)

### Đánh giá mô hình

Chart, box and whisker chart

Description automatically generated

***Hình 2.h. Đồ thị so sánh giá trị thực tế với tập train của từng mô hình***

Có thể thấy dữ liệu dự đoán (màu xanh) biến thiên thuận với dữ liệu thực tế. Tuy nhiên sai số khá lớn ở dữ liệu dự đoán thể hiện ở hai đầu của đường phác họa.

Table

Description automatically generated with medium confidence

***Hình 2.i. Đánh giá các mô hình bằng các độ đo MAE, MSE, R2***

Chúng tôi đánh giá mô hình dựa vào các thang đo MAE, MSE, R2. Ở hình trên ta thấy các giá trị ở các thang đo khá thấp, ngay cả ở tập huấn luyện giá trị MAE chỉ xấp xỉ 0.2, R2 xấp xỉ 0.1, cá biệt ở mô hình Polynomial Regressor bậc 6 có R2 lên đến 0.6. Các giá trị này ở cả 2 tập huấn luyện và kiểm thử cho thấy bộ dữ liệu (hay các thuộc tính) không phù hợp với việc đưa vào mô hình dự đoán.

# 3. KẾT LUẬN

Trong đồ án này, nhóm chúng tôi đã sử dụng thư viện *BeautifulSoup* của Python thu thập được bộ dữ liệu thô chứa các thông tin về các đầu sách trên Goodreads.com. Khi đã có bộ dữ liệu thô nhóm chúng tôi tiếp tục tiền xử lý và chuẩn hóa lại dữ liệu, quá trình chuẩn hóa này bao gồm cả chuẩn hóa tự động và thực hiện chuẩn hóa thủ công một thuộc tính có độ phức tạp cao. Sau khi có được bộ dữ liệu hoàn thiện, nhóm đã tiến hành phân tích và tìm được 3 thuộc tính quan trọng là *genre, description* và *awards*.

Nhóm chúng tôi sử dụng module *Pickle* chuyển các thuộc tính phân loại thành kiểu thuộc tính số để xây dựng được 3 mô hình dự đoán rating dựa vào 9 thuộc tính trong đó bao gồm 3 thuộc tính insights đã tìm được trước đó. Tuy nhiên, mô hình xây dựng có độ chính không được cao vì các nguyên nhân: trong thực người đọc có xu hướng đánh giá cuốn sách dựa vào nội dung của nó, vì vậy việc huấn luyện mô hình dự đoán rating trên các biến thông tin là không khả quan; thứ hai là nguồn dữ liệu thu thập từ trang Goodreads có một số hạn chế, như việc thuộc tính *genre* (thể loại) không phải hoàn toàn chính xác vì thể loại của sách được quyết định bằng cách thu thập dữ liệu người dùng, bên cạnh đó các thuộc tính *rating\_count* và *review\_count* cần thời gian để có một lượng người dùng đánh giá và bình luận, điều này khiến việc dự đoán rating các cuốn sách mới, sách bị khuyết 2 thuộc tính này trở nên khó khăn, làm giảm hiệu suất của mô hình.

Tóm lại, từ bộ dữ liệu thu thập được ta có thể tìm được cách thu thập và xử lý dữ liệu một các hiệu quá hơn, hoặc mở rộng độ lớn của bộ dữ liệu để nâng cao hiệu suất của mô ình dự đoán. Mô hình được xây dựng trong đồ án này chỉ mang tính chất tham khảo và còn nhiều thiết sót, có thể sẽ tìm được được mô hình phù hợp hơn ở các nghiên cứu sau.

**TÀI LIỆU THAM KHẢO**

1. Best Books of the 21st Century, https://www.goodreads.com/list/show/7.Best\_Books\_of\_the\_21st\_Century, 12/2021.
2. Fellipe Marcellino, Data X Webscraping, 2020.
3. Christine Alvarado, Briana B. Morrison, Barbara Ericson, Mark Guzdial, Brad Miller, Performance and Use Evaluation of an Electronic, 12/ 2021.
4. How can I set my book's genres?, <https://help.goodreads.com/s/article/How-can-I-set-my-book-s-genres>, 12/2021

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

| **STT** | **Thành viên** | **Nhiệm vụ** |
| --- | --- | --- |
| 1 | Nguyễn Thành Phúc | Thu thập dữ liệu, xử lý dữ liệu, viết code, làm slide, viết báo cáo |
| 2 | Trần Nhật Nam | Thu thập dữ liệu, xử lý dữ liệu, xây dựng mô hình, viết code, làm slide, viết báo cáo |
| 3 | Nguyễn Võ Thiên Ân | Thu thập dữ liệu, xử lý dữ liệu, viết code, làm slide, viết báo cáo |

1. Goodreads là một mạng xã hội về sách lớn nhất thế giới, dành cho các độc giả với tiêu chí giúp mọi người tìm và chia sẻ những cuốn sách mà họ yêu thích. Được thành lập vào năm 2007, đến tháng 10 năm 2012, Goodreads thông báo đã phát triển lên 11 triệu thành viên với 395 triệu đầu sách được thêm vào. Hiện nay trang web đã phát triển lên đến hơn 90 triệu người dùng tính đến năm 2019. [↑](#footnote-ref-0)