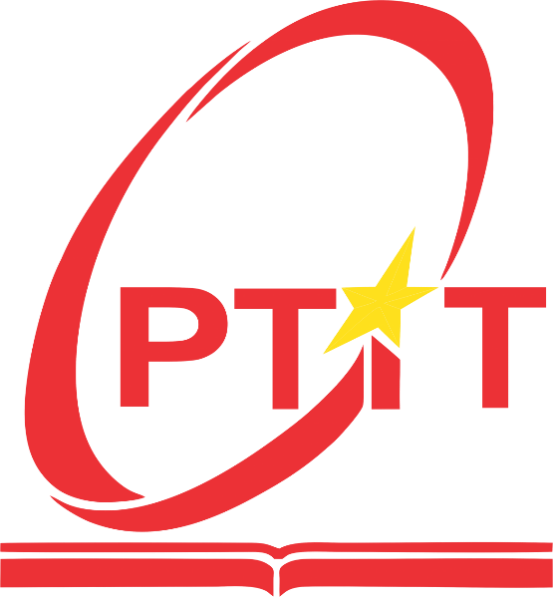
**Học viện Công nghệ Bưu chính Viễn Thông cơ sở tp.HCM**

****

**Đề tài: Dự đoán rating của một cuốn sách**

**Môn: Nhập môn khoa học dữ liệu**

**GVHD: Nguyễn Ngọc Duy**

**Thành viên:**

**Nguyên Thành Long N21DCCN**

**Tô Phan Kiều Thương N21DCCN184**

**Lê Anh Tình N21DCCN**

**Mục lục**

[**Mở đầu** 2](#_Toc180365278)

[**I. Giới thiệu** 3](#_Toc180365279)

[**II. Cơ sở lý thuyết và công nghệ** 6](#_Toc180365280)

[**1.Goodread** 6](#_Toc180365281)

[**2. Python** 6](#_Toc180365282)

[**3. Mô hình** 7](#_Toc180365283)

[**III. Phân tích thiết kế** 15](#_Toc180365284)

[**1. Thu thập dữ liệu** 15](#_Toc180365285)

[**2. Xử lý dữ liệu:** 16](#_Toc180365286)

[**3. Trực quan hóa dữ liệu** 21](#_Toc180365287)

[**4. Huấn luyện dữ liệu:** 22](#_Toc180365288)

[**5. Đánh giá mô hình:** 22](#_Toc180365289)

[**IV. Xây dựng hệ thống** 22](#_Toc180365290)

[**1.Thu thập dữ liệu** 22](#_Toc180365291)

[**2. Xử lý dữ liệu** 28](#_Toc180365292)

[**3. Trực quan hóa dữ liệu** 48](#_Toc180365293)

[**4. Huấn luyện dữ liệu** 49](#_Toc180365294)

[**5. Đánh giá mô hình** 50](#_Toc180365295)

[**V. Chương trình minh họa** 51](#_Toc180365296)

[**VI. Kết luận** 53](#_Toc180365297)

# **Mở đầu**

Khi một doanh nghiệp kinh doanh sách muốn dự đoán rating (đánh giá) của một cuốn sách, họ đang cố gắng xác định mức độ hấp dẫn và tiềm năng thành công của cuốn sách trên thị trường. Rating cao thường liên quan trực tiếp đến doanh số bán hàng, mức độ hài lòng của khách hàng và danh tiếng của thương hiệu. Do đó, việc dự đoán chính xác rating của sách sẽ giúp doanh nghiệp tối ưu hóa chiến lược kinh doanh, từ việc nhập hàng đến các chiến dịch marketing.

Dưới góc nhìn của một doanh nghiệp, dự đoán rating có thể mang lại nhiều lợi ích quan trọng:

1. **Quản lý danh mục sản phẩm**: Dự đoán rating giúp doanh nghiệp quyết định nên tập trung vào những đầu sách nào để tăng tính cạnh tranh. Những sách dự đoán có rating cao có thể được ưu tiên hơn trong việc nhập hàng và trưng bày, trong khi những sách có tiềm năng thấp hơn có thể được bán với mức chiết khấu.
2. **Tối ưu hóa chiến lược marketing**: Rating của sách có thể ảnh hưởng đến chiến lược tiếp thị của doanh nghiệp. Ví dụ, các sách được dự đoán có rating cao có thể được quảng bá mạnh mẽ hơn, với các chiến dịch truyền thông nhắm mục tiêu đến nhóm đối tượng khách hàng cụ thể.
3. **Cải thiện trải nghiệm khách hàng**: Bằng cách hiểu trước được đánh giá của khách hàng tiềm năng, doanh nghiệp có thể cung cấp những gợi ý phù hợp hơn cho từng nhóm độc giả. Điều này không chỉ giúp tăng doanh thu mà còn xây dựng lòng trung thành của khách hàng.
4. **Quyết định giá bán**: Những sách dự đoán có rating cao có thể được định giá cao hơn, trong khi những sách có rating thấp hơn có thể phải giảm giá để thúc đẩy doanh số.

Việc dự đoán rating của sách đòi hỏi doanh nghiệp phải phân tích dữ liệu từ nhiều nguồn khác nhau, chẳng hạn như đánh giá của khách hàng trước đó, đặc điểm của tác giả, thể loại sách, và nội dung mô tả. Các kỹ thuật phân tích dữ liệu và học máy (machine learning) có thể được áp dụng để xây dựng các mô hình dự đoán nhằm đưa ra các đánh giá chính xác và hỗ trợ quyết định kinh doanh.

# **I. Giới thiệu**

**Nội dung:** Trình bày quá trình xây dựng hệ thống dự đoán đánh giá sách từ dữ liệu lấy từ Goodreads (https://www.goodreads.com). Việc này nhằm mục đích áp dụng các kỹ thuật khoa học dữ liệu để tạo ra một mô hình có thể dự đoán độ đánh giá của các cuốn sách dựa trên các đặc điểm như tiêu đề, series, thể loại và các yếu tố khác.

**Cơ sở lý thuyết và công nghệ liên quan:**

Phần này sử dụng nền tảng Extensions Jupyter Notebook của Visual Studio Code để thực hiện các bước xử lý dữ liệu và huấn luyện mô hình. Jupyter Notebook cung cấp môi trường Python trực tuyến thuận tiện, đặc biệt với thư viện như pandas,….Một số thư viện của Python và một số mô hình liên quan hồi quy tuyến tính

**Thư viện:**

BeautifulSoup

* BeautifulSoup là một thư viện dùng để phân tích cú pháp HTML và XML. Nó hỗ trợ việc tìm kiếm, trích xuất và sửa đổi các phần tử trong tài liệu HTML hoặc XML.
* Các ứng dụng phổ biến của BeautifulSoup bao gồm: web scraping (lấy dữ liệu từ các trang web), phân tích cú pháp tài liệu HTML/XML và xử lý dữ liệu văn bản.
* BeautifulSoup thường được kết hợp với thư viện Requests để tự động hóa quá trình thu thập dữ liệu từ web.
* Cách thức hoạt động: BeautifulSoup chuyển đổi nội dung HTML thành một cấu trúc cây (parse tree), sau đó cho phép bạn dễ dàng tìm kiếm và thao tác với các phần tử của tài liệu, như tìm thẻ HTML, trích xuất dữ liệu từ các thuộc tính, hay xử lý các liên kết.

Requests

* Requests là một thư viện Python mạnh mẽ và đơn giản để gửi HTTP requests. Nó giúp việc giao tiếp với các trang web trở nên dễ dàng hơn bằng cách cung cấp các phương thức đơn giản để gửi GET, POST và các loại yêu cầu HTTP khác.
* Requests có thể tải về nội dung của một trang web, gửi dữ liệu lên server, hoặc thực hiện các tác vụ khác như xác thực, quản lý cookies, và xử lý các mã trạng thái HTTP.
* Thư viện này được yêu thích bởi cú pháp thân thiện với người dùng và tính linh hoạt cao. Ví dụ, chỉ với một dòng lệnh, bạn có thể gửi một yêu cầu GET đến một URL và nhận được phản hồi từ máy chủ.

Pandas

* Pandas là một thư viện mạnh mẽ để xử lý và phân tích dữ liệu dạng bảng (dataframes) trong Python. Nó cung cấp các cấu trúc dữ liệu linh hoạt và các công cụ để thao tác, chuyển đổi, và tổng hợp dữ liệu.
* Các thành phần chính của Pandas là Series và DataFrame. Series là một mảng một chiều, còn DataFrame là một bảng hai chiều với các hàng và cột, tương tự như bảng tính trong Excel hoặc bảng dữ liệu trong cơ sở dữ liệu.
* Pandas hỗ trợ việc đọc và ghi dữ liệu từ nhiều định dạng như CSV, Excel, SQL, và thậm chí là HTML. Nó cũng có thể thực hiện các thao tác phức tạp trên dữ liệu như lọc, nhóm, và phân tích thống kê.
* Đây là thư viện cực kỳ phổ biến trong lĩnh vực khoa học dữ liệu và phân tích dữ liệu vì tính tiện lợi và hiệu quả của nó trong việc xử lý dữ liệu lớn.

**Mô hình:**

Hồi quy tuyến tính (Linear Regression)

Hồi quy tuyến tính là một phương pháp thống kê nhằm mô hình hóa mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập thông qua một hàm tuyến tính. Phương pháp này dựa trên giả định rằng mối quan hệ giữa các biến là tuyến tính, có nghĩa là sự thay đổi của biến độc lập sẽ dẫn đến sự thay đổi tỷ lệ tương ứng của biến phụ thuộc. Mục tiêu chính là tìm ra đường thẳng tốt nhất phù hợp với dữ liệu để dự đoán giá trị của biến phụ thuộc.

MLP Regression (Multilayer Perceptron Regression)

MLP Regression là một phương pháp hồi quy sử dụng mạng nơ-ron nhân tạo (Artificial Neural Network - ANN) để dự đoán các giá trị liên tục. Mô hình bao gồm nhiều lớp nơ-ron (input layer, hidden layers, output layer), với các kết nối được tối ưu hóa thông qua quá trình học. MLP Regression có khả năng mô hình hóa các mối quan hệ phi tuyến và phức tạp trong dữ liệu, nhờ vào việc sử dụng các hàm kích hoạt phi tuyến và cấu trúc nhiều lớp.

Support Vector Regression (SVR)

Support Vector Regression là một biến thể của thuật toán Support Vector Machine (SVM), được áp dụng cho các bài toán hồi quy. Mục tiêu của SVR là tìm một siêu phẳng (hyperplane) trong không gian nhiều chiều để dự đoán các giá trị liên tục. Điểm nổi bật của SVR là khả năng tối ưu hóa khoảng cách giữa siêu phẳng và các điểm dữ liệu, giúp mô hình đạt được độ tổng quát tốt hơn và xử lý các dữ liệu phi tuyến bằng cách sử dụng các hàm kernel để ánh xạ dữ liệu sang không gian cao hơn.

Cả ba mô hình này đều được sử dụng rộng rãi trong các bài toán dự đoán giá trị liên tục, nhưng mỗi phương pháp có cách tiếp cận và ưu nhược điểm riêng, phù hợp với các loại dữ liệu và yêu cầu khác nhau.

# **II. Cơ sở lý thuyết và công nghệ**

## **1.Goodread**

Goodreads là một mạng xã hội dành riêng cho người yêu sách, ra mắt vào năm 2006 bởi Otis Chandler. Đây là một trong những trang web lớn nhất cho phép người dùng tìm kiếm, đề xuất, và chia sẻ thông tin về các cuốn sách. Goodreads cung cấp thông tin chi tiết về hàng triệu cuốn sách từ nhiều thể loại khác nhau, từ tiểu thuyết đến phi tiểu thuyết, văn học cổ điển đến các sách khoa học kỹ thuật.

Mỗi cuốn sách trên Goodreads đều có một trang riêng chứa các thông tin cơ bản như tiêu đề, tác giả, nhà xuất bản, ngày xuất bản, thể loại, và mô tả. Ngoài ra, Goodreads cũng cho phép người dùng đánh giá và nhận xét về cuốn sách, từ đó tạo ra một cộng đồng chia sẻ kinh nghiệm và ý kiến. Dữ liệu từ trang Goodreads có thể được truy cập thông qua API chính thức hoặc bằng cách phân tích cú pháp HTML từ trang web của từng cuốn sách.

Trang Goodreads cung cấp thông tin chi tiết về sách qua API hoặc phân tích dữ liệu từ HTML trang sách. Quá trình lấy dữ liệu dựa trên các thẻ HTML như <h1>, <a>, và các phần tử khác.Goodreads là một nguồn tài nguyên phong phú để thu thập dữ liệu liên quan đến sách phục vụ các mục đích nghiên cứu, phát triển hệ thống đề xuất, và các ứng dụng khác trong lĩnh vực học máy và xử lý ngôn ngữ tự nhiên

## **2. Python**

Để thu thập dữ liệu từ các trang web, Python cung cấp các thư viện mạnh mẽ như Requests và BeautifulSoup giúp thực hiện nhiệm vụ này một cách hiệu quả.

**Requests:**

Request là một thư viện Python mạnh mẽ và đơn giản để gửi các yêu cầu HTTP. Nó được sử dụng rộng rãi trong việc thu thập dữ liệu từ web, cho phép người dùng tải nội dung của một trang web một cách dễ dàng bằng cách gửi các yêu cầu GET hoặc POST. Requests hỗ trợ giao thức HTTPS và có khả năng quản lý các phiên, cookies, và các cấu hình khác khi tương tác với một trang web.

**BeautifulSoup:**  
BeautifulSoup là một thư viện Python dùng để phân tích cú pháp các tài liệu HTML và XML. Nó cung cấp một giao diện dễ sử dụng để trích xuất dữ liệu từ các trang web bằng cách xác định và xử lý các phần tử HTML cụ thể như thẻ <h1>, <a>, <div>, và nhiều thẻ khác. BeautifulSoup có thể làm việc với nhiều trình phân tích cú pháp khác nhau, phổ biến nhất là html.parser, để chuyển đổi nội dung HTML thành một cây cấu trúc có thể duyệt qua.

Công cụ này rất hữu ích trong việc xử lý và làm sạch dữ liệu không có cấu trúc, chẳng hạn như dữ liệu từ các trang web không hỗ trợ API chính thức. Bằng cách sử dụng BeautifulSoup, các nhà phát triển có thể dễ dàng trích xuất thông tin từ một trang web, như tiêu đề sách, tác giả, thể loại, và các đánh giá.

Thư viện BeautifulSoup hỗ trợ phân tích cú pháp HTML, trong khi requests được dùng để gửi yêu cầu HTTP và lấy nội dung trang.

## **3. Mô hình**

Mô hình:

**3.1 Hồi quy tuyến tính(Linear Regression)**

1. Phương trình của hồi quy tuyến tính

* Hồi quy tuyến tính đơn giản

Hồi quy tuyến tính đơn giản mô hình hóa mối quan hệ giữa một biến phụ thuộc y và một biến độc lập x bằng phương trình:

Trong đó:

* y: Biến phụ thuộc (biến cần dự đoán).
* x: Biến độc lập (biến đầu vào).
* ​: Hệ số chặn (intercept), là giá trị của y khi x = 0.
* ​: Hệ số hồi quy (slope), biểu thị mức độ thay đổi của y khi x thay đổi một đơn vị.
* : Thành phần nhiễu (error term), thể hiện sai số giữa giá trị thực và giá trị dự đoán.
* Hồi quy tuyến tính đa biến

Hồi quy tuyến tính đa biến mô hình hóa mối quan hệ giữa một biến phụ thuộc y và nhiều biến độc lập x1, x2,…, xn ​ với phương trình:

Trong đó:

* x1, x2,…, xn ​ ​ là các biến độc lập.
* là các hệ số hồi quy tương ứng.

1. Mục tiêu của hồi quy tuyến tính

* Mục tiêu chính của hồi quy tuyến tính là tìm các giá trị tối ưu cho các hệ số sao cho sai số giữa giá trị dự đoán và giá trị thực là nhỏ nhất. Để đạt được điều này, phương pháp phổ biến nhất là phương pháp bình phương nhỏ nhất (Ordinary Least Squares - OLS), trong đó tổng bình phương sai số (Sum of Squared Errors - SSE) được tối thiểu hóa:

Trong đó:

là giá trị thực của điểm dữ liệu thứ i.

​ là giá trị dự đoán của điểm dữ liệu thứ i.

n là số lượng điểm dữ liệu.

1. Đánh giá mô hình hồi quy tuyến tính

* Một số chỉ số phổ biến để đánh giá hiệu quả của mô hình hồi quy tuyến tính bao gồm:
* R-squared (R²): Biểu diễn tỷ lệ phương sai của biến phụ thuộc được giải thích bởi các biến độc lập. Giá trị R² nằm trong khoảng từ 0 đến 1, với giá trị càng cao thì mô hình càng phù hợp.
* Adjusted R-squared: Điều chỉnh giá trị R² để phản ánh số lượng biến độc lập trong mô hình, nhằm tránh tình trạng mô hình phức tạp quá mức.
* Mean Absolute Error (MAE): Đo lường độ chính xác của mô hình bằng cách tính trung bình các sai số tuyệt đối.
* Mean Squared Error (MSE): Đo lường độ chính xác của mô hình bằng cách tính trung bình các bình phương sai số.
* Root Mean Squared Error (RMSE): Là căn bậc hai của MSE, giúp đánh giá sai số dự đoán theo cùng đơn vị với biến phụ thuộc.

1. Ưu và nhược điểm của mô hình hồi quy tuyến tính

Ưu điểm:

* Dễ hiểu và dễ triển khai.
* Cho phép giải thích rõ ràng mối quan hệ giữa các biến.
* Có thể hoạt động tốt với dữ liệu tuyến tính.

Nhược điểm:

* Không hiệu quả với các dữ liệu có mối quan hệ phi tuyến.
* Nhạy cảm với các giá trị ngoại lệ.
* Nếu các giả định của mô hình không được thỏa mãn, kết quả sẽ không đáng tin cậy.

1. Ứng dụng của hồi quy tuyến tính

Hồi quy tuyến tính được sử dụng rộng rãi trong nhiều lĩnh vực như:

* Kinh tế và tài chính: Dự đoán giá cổ phiếu, lạm phát, lợi nhuận.
* Y tế: Phân tích các yếu tố ảnh hưởng đến sức khỏe, dự đoán chi phí y tế.
* Khoa học xã hội: Nghiên cứu mối quan hệ giữa các yếu tố xã hội và hành vi con người.
* Kỹ thuật và sản xuất: Dự đoán tuổi thọ của thiết bị, quản lý chất lượng sản phẩm.

**3.2 Support Vector Regression**

Support Vector Regression (SVR) là một kỹ thuật hồi quy dựa trên nguyên lý của Support Vector Machine (SVM), được sử dụng để dự đoán các giá trị liên tục. SVR mở rộng SVM từ các bài toán phân loại sang các bài toán hồi quy, với mục tiêu tìm ra một hàm dự đoán sao cho sai số giữa giá trị thực và giá trị dự đoán nằm trong một ngưỡng chấp nhận được.

1. Ý tưởng chính của SVR

SVR cố gắng tìm một đường hồi quy sao cho tất cả các điểm dữ liệu nằm trong một dải sai số ϵ\epsilonϵ. Mục tiêu là làm phẳng đường hồi quy càng nhiều càng tốt, trong khi vẫn cho phép một số điểm dữ liệu nằm ngoài dải sai số này. Điều này được thực hiện bằng cách tối thiểu hóa độ phức tạp của mô hình, tức là tối thiểu hóa độ lớn của các hệ số hồi quy.

1. Phương trình của SVR

Trong SVR, hàm hồi quy có dạng:

Trong đó:

* là vector trọng số.
* b là hệ số chặn (intercept).

SVR cố gắng tìm và b sao cho các giá trị thực ​ của các điểm dữ liệu nằm trong phạm vi sai số từ giá trị dự đoán . Điều này có nghĩa là:

Nếu một số điểm dữ liệu nằm ngoài phạm vi sai số , sẽ có các biến bổ sung (slack variables) được sử dụng để xử lý.

1. Hàm mất mát epsilon-insensitive

SVR sử dụng hàm mất mát epsilon-insensitive để tính toán sai số giữa giá trị thực và giá trị dự đoán. Hàm này được định nghĩa như sau:

* Nếu sai số nhỏ hơn , hàm mất mát bằng 0.
* Nếu sai số lớn hơn , hàm mất mát bằng .

1. Hàm mục tiêu trong SVR

Hàm mục tiêu của SVR là tối thiểu hóa độ phức tạp của mô hình (được thể hiện qua ) và đồng thời giảm thiểu sai số dự đoán vượt quá phạm vi . Hàm mục tiêu có dạng:

Trong đó:

* là độ phức tạp của mô hình.
* ​ là các biến bổ sung để xử lý những điểm dữ liệu nằm ngoài phạm vi ..
* C là tham số điều chỉnh mức độ chấp nhận sai số vượt quá phạm vi .

1. Kernel trick trong SVR

SVR có thể sử dụng các hàm kernel để chuyển đổi dữ liệu từ không gian ban đầu sang không gian có chiều cao hơn, giúp xử lý các bài toán hồi quy phi tuyến. Một số kernel phổ biến gồm:

* Linear kernel: Phù hợp với các bài toán tuyến tính.
* Polynomial kernel: Biểu diễn các mối quan hệ đa thức.
* Radial Basis Function (RBF) kernel: Phù hợp với các bài toán phi tuyến phức tạp.

1. Ưu điểm và nhược điểm của SVR

Ưu điểm:

* Khả năng xử lý các bài toán hồi quy phi tuyến nhờ sử dụng kernel.
* Có khả năng kiểm soát độ phức tạp của mô hình thông qua tham số CCC và ϵ\epsilonϵ.

Nhược điểm:

1. Thời gian tính toán có thể lâu hơn khi làm việc với bộ dữ liệu lớn hoặc kernel phi tuyến.
2. Cần điều chỉnh nhiều tham số (C, ϵ\epsilonϵ, kernel parameters) để đạt được hiệu suất tốt nhất.
3. Ứng dụng của SVR

SVR được áp dụng rộng rãi trong nhiều lĩnh vực như dự đoán giá chứng khoán, dự báo nhu cầu, phân tích chuỗi thời gian, và các bài toán khoa học, kỹ thuật đòi hỏi sự dự đoán giá trị liên tục.

**3.3 MLP Regression(Multilayer Perceptron Regression)**

MLP Regression (Multilayer Perceptron Regression) là một kỹ thuật hồi quy dựa trên mạng nơ-ron nhân tạo (Artificial Neural Networks - ANN). Mô hình MLP Regression sử dụng cấu trúc mạng nhiều lớp (multilayer) để học các mối quan hệ phức tạp giữa các biến đầu vào (biến độc lập) và biến đầu ra (biến phụ thuộc).

a) Kiến trúc của MLP Regression

MLP Regression thường bao gồm các thành phần sau:

* Lớp đầu vào (Input Layer): Chứa các nơ-ron (neurons) đại diện cho các đặc trưng của dữ liệu đầu vào.
* Lớp ẩn (Hidden Layers): Một hoặc nhiều lớp nơ-ron ở giữa lớp đầu vào và lớp đầu ra. Các lớp ẩn này giúp học các mối quan hệ phức tạp và phi tuyến tính trong dữ liệu.
* Lớp đầu ra (Output Layer): Có một nơ-ron duy nhất trong trường hợp hồi quy đơn biến, hoặc nhiều nơ-ron nếu là hồi quy đa biến.

Các nơ-ron trong mỗi lớp được kết nối với nhau thông qua các trọng số (weights), và các trọng số này sẽ được tối ưu hóa trong quá trình huấn luyện mô hình.

b) Cách hoạt động của MLP Regression

MLP Regression hoạt động thông qua quá trình lan truyền tiến (feedforward) và lan truyền ngược (backpropagation):

* Lan truyền tiến (Feedforward): Đầu vào được truyền từ lớp này sang lớp khác, từ lớp đầu vào đến lớp ẩn và cuối cùng là lớp đầu ra. Tại mỗi nơ-ron, một hàm kích hoạt (activation function) được áp dụng để tính toán đầu ra của nơ-ron đó.
* Lan truyền ngược (Backpropagation): Sai số giữa giá trị đầu ra thực tế và giá trị đầu ra dự đoán sẽ được tính toán. Sai số này sau đó được sử dụng để điều chỉnh các trọng số nhằm tối ưu hóa mô hình. Quá trình này tiếp tục cho đến khi sai số đạt đến mức nhỏ nhất có thể.

c) Hàm kích hoạt

Hàm kích hoạt được sử dụng để thêm tính phi tuyến vào mô hình, giúp MLP Regression có khả năng học các mối quan hệ phi tuyến tính trong dữ liệu. Một số hàm kích hoạt phổ biến:

* ReLU (Rectified Linear Unit): được sử dụng rộng rãi do khả năng giảm bớt vấn đề gradient biến mất.
* Sigmoid: ​ , được dùng trong các mạng nơ-ron đơn giản hơn nhưng ít phổ biến trong các mô hình phức tạp.
* Tanh (Hyperbolic Tangent: ​ ​, có dải giá trị từ -1 đến 1.

d) Hàm mất mát và tối ưu hóa

Trong bài toán hồi quy, hàm mất mát phổ biến nhất là Mean Squared Error (MSE), được tính bằng:

Trong đó:

* là giá trị thực của mẫu thứ i.
* ​ là giá trị dự đoán của mẫu thứ i.
* n là số lượng mẫu.

Tối ưu hóa trọng số của mô hình thường được thực hiện bằng cách sử dụng các thuật toán như Gradient Descent, Stochastic Gradient Descent (SGD), hoặc các biến thể khác như Adam, RMSprop, giúp tăng tốc độ hội tụ và tối ưu hóa tốt hơn.

e) Ưu điểm và nhược điểm của MLP Regression

Ưu điểm:

* Có khả năng học các mối quan hệ phi tuyến phức tạp trong dữ liệu.
* Linh hoạt trong việc lựa chọn số lớp, số nơ-ron, và các hàm kích hoạt khác nhau.

Nhược điểm:

* Dễ bị quá khớp nếu không có kỹ thuật giảm thiểu thích hợp.
* Cần nhiều dữ liệu để mô hình hóa tốt.
* Thời gian huấn luyện lâu hơn so với các phương pháp hồi quy đơn giản.

f) Ứng dụng của MLP Regression

MLP Regression được sử dụng rộng rãi trong các lĩnh vực yêu cầu dự đoán các giá trị liên tục, chẳng hạn như:

* Dự đoán giá cổ phiếu hoặc các tài sản tài chính.
* Dự báo thời tiết.
* Dự đoán nhu cầu tiêu thụ điện năng.
* Phân tích chuỗi thời gian trong kinh tế và sản xuất.

# **III. Phân tích thiết kế**

## **1. Thu thập dữ liệu**

Quá trình thu thập dữ liệu bao gồm các bước sau:

**Bước 1**: Vào [Browser lists trong](https://www.goodreads.com/list/show/1.Best_Books_Ever) “[Recent updates | Goodreads](https://www.goodreads.com/)” lấy ra đường dẫn của tất cả các danh mục có trong goodread.

**Bước 2**: Từ đó sẽ vào từng link đã thu thập được ở trên để lấy link chi tiết của từng cuốn sách.

**Bước 3**: Ta có danh sách link của từng cuốn sách. Bắt đầu xử lý những đường dẫn bị trùng rồi thực hiện thu thập thông tin của từng cuốn sách.

Sử dụng thư viện requests để lấy dữ liệu từ Goodreads thông qua các URL chứa thông tin chi tiết của từng cuốn sách. Sau đó dùng BeautifulSoup để trích xuất dữ liệu từ các thẻ HTML.

Quá trình thu thập dữ liệu bắt đầu từ việc định danh các thông tin cần lấy từ trang Goodreads, bao gồm tiêu đề sách, tác giả, thể loại, số lượng đánh giá, và các thông tin chi tiết khác. Dữ liệu được thu thập thông qua các URL của từng cuốn sách, từ đó phân tích HTML của các trang để trích xuất thông tin cần thiết.

Mô tả thông tin cần lấy của một cuốn sách:

id : Mã của cuốn sách

title : Tên của cuốn sách

link : Đường dẫn đến cuốn sách đó

series: Tên của tập sách chứa cuốn sách

author : Tác giả cuốn sách

author\_link: Đường dẫn đến thông tin về tác giả

rating\_count: Số lượng người đánh giá

review\_count: Số lượng người để lại bình luận

number\_of\_pages : Số trang của cuốn sách

date\_published: Ngày được xuất bản

publisher : Nhà xuất bản

original\_title : Tên gốc

genre\_and\_votes: Thể loại và lượt bình chọn

isbn : Mã số tiêu chuẩn quốc tế cho sách

isbn13: Mã vạch

settings : Bối cảnh diễn ra chính trong cuốn sách

characters: Các nhân vật trong cuốn sách

description: Tóm tắt cuốn sách

awards : Các giải thưởng mà cuốn sách nhận được

rating : Đánh giá trung bình của người đọc

followers : Số lượng follower của tác giả

## **2. Xử lý dữ liệu:**

+ Sử dụng Jupyter Notebook và Python (thư viện pandas) để làm sạch dữ liệu.

+ Các bước xử lý dữ liệu đã thu thập bao gồm:

* Khai báo thư viện cần thiết cho việc xử lý dữ liệu (pandas).
* Thực hiện đọc các file đã thu thập dữ liệu từ https://www.goodreads.com
* Kiểm tra tập dữ liệu:

Kiểm tra kiểu dữ liệu đầu vào:

A screenshot of a computer code

Description automatically generated

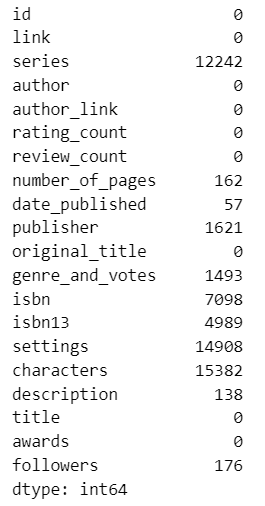
Hình 3.2.1 Kiểm tra kiểu dữ liệu

Kiểm tra dữ liệu bằng cách tính tổng số giá trị NaN trong từng cột:

A black rectangular object with colorful circles and text

Description automatically generated

Hình 3.2.2 Tổng số NaN từng cột



Hình 3.2.3 Tổng số NaN trên từng cột

Tính tỷ lệ phần trăm giá trị NaN trong từng cột so với tổng số hàng của DataFrame ‘df\_X’. ‘df\_X’ là bảng thông tin chi tiết của những cuốn sách không bao gồm thuộc tính ‘rating’.

A black rectangular with white and green text

Description automatically generated

Hình 3.2.4 Tỉ lệ NaN



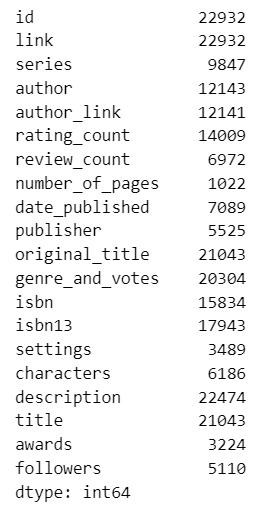
Hình 3.2.5 Tỉ lệ NaN

Kiểm tra xem số lượng giá trị duy nhất trong từng cột:

A black rectangular object with colorful circles and text

Description automatically generated

Hình 3.2.6 Kiểm tra số giá trị duy nhất từng cột



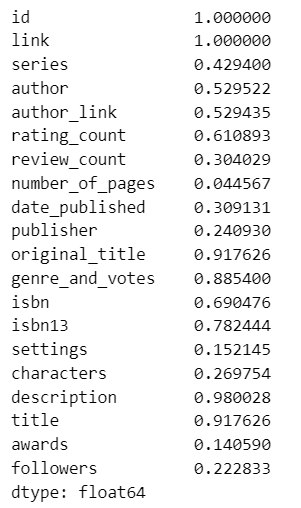
Hình 3.2.7 Số lượng giá trị duy nhất từng cột

Kiểm tra tỉ lệ số lượng các giá trị duy nhất so với số lượng toàn bộ bảng:

A black rectangular with white text and green dot

Description automatically generated

Hình 3.2.8 Tỉ lệ của số giá trị duy nhất từng cột



Hình 3.2.9 Tỉ lệ số giá trị duy nhất từng cột

* Bước làm sạch dữ liệu (data cleaning)
  + Loại bỏ những dòng dữ liệu bị trùng lặp.
  + Xử lý những cột không ảnh hưởng(không liên quan) đến việc đánh giá một cuốn sách.
  + Xử lý những cột có nhiều giá trị NaN.
  + Xử lý những cột có nhiều giá trị độc nhất.
  + Xử lý những dòng dữ liệu có giá trị NaN bằng cách thay vào giá trị 0.
* Chuẩn hóa dữ liệu:
  + Chuẩn hóa dữ liệu của từng cột thành một thể thống nhất, có ý nghĩa liên quan đến việc đánh giá.
  + Chuẩn hóa dữ liệu dạng định tính thành dữ liệu định lượng.
  + Đối với những cột có ảnh hưởng nhưng không có cách nào đưa chúng vào một thể thống nhất thì gán cho chúng một giá trị độc nhất.

## **3. Trực quan hóa dữ liệu**

- Xác định mục tiêu và đối tượng: Hiểu rõ mục đích và đối tượng sử dụng trực quan hóa.

- Chuẩn bị dữ liệu: Làm sạch, chuyển đổi và tóm tắt dữ liệu cần thiết.

- Chọn loại biểu đồ: Lựa chọn biểu đồ phù hợp (cột, đường, scatter, v.v.) với mục tiêu.

- Thiết kế biểu đồ: Đảm bảo nhãn, màu sắc và tỷ lệ hợp lý.

## **4. Huấn luyện dữ liệu:**

- Chia dữ liệu thành các tập huấn luyện và kiểm thử: Thường chia theo tỷ lệ như 70-30 hoặc 80-20. Tập huấn luyện dùng để huấn luyện mô hình, trong khi tập kiểm thử dùng để đánh giá hiệu suất của mô hình.

- Lựa chọn và xây dựng mô hình: Chọn thuật toán học máy phù hợp: Tùy theo bài toán (phân loại, hồi quy, v.v.), chọn các thuật toán phù hợp như hồi quy tuyến tính, mạng nơ-ron, cây quyết định, SVM, v.v.

- Xây dựng mô hình: Định cấu hình mô hình ban đầu với các tham số mặc định.

- Huấn luyện mô hình với tập dữ liệu huấn luyện: Sử dụng dữ liệu huấn luyện để tối ưu hóa các tham số của mô hình.

- Kiểm thử với tập kiểm thử (validation): Trong một số trường hợp, tập kiểm thử được sử dụng để điều chỉnh các tham số siêu (hyperparameter tuning) nhằm tối ưu hóa hiệu suất.

## **5. Đánh giá mô hình:**

- Xác định mục tiêu và chọn các chỉ số đánh giá như độ chính xác, F1-score, MSE; - Thiết kế phương pháp đánh giá, chia tách dữ liệu hoặc dùng cross-validation

- Tiến hành đánh giá trên tập kiểm thử.

- Phân tích lỗi để hiểu nguyên nhân dự đoán sai.

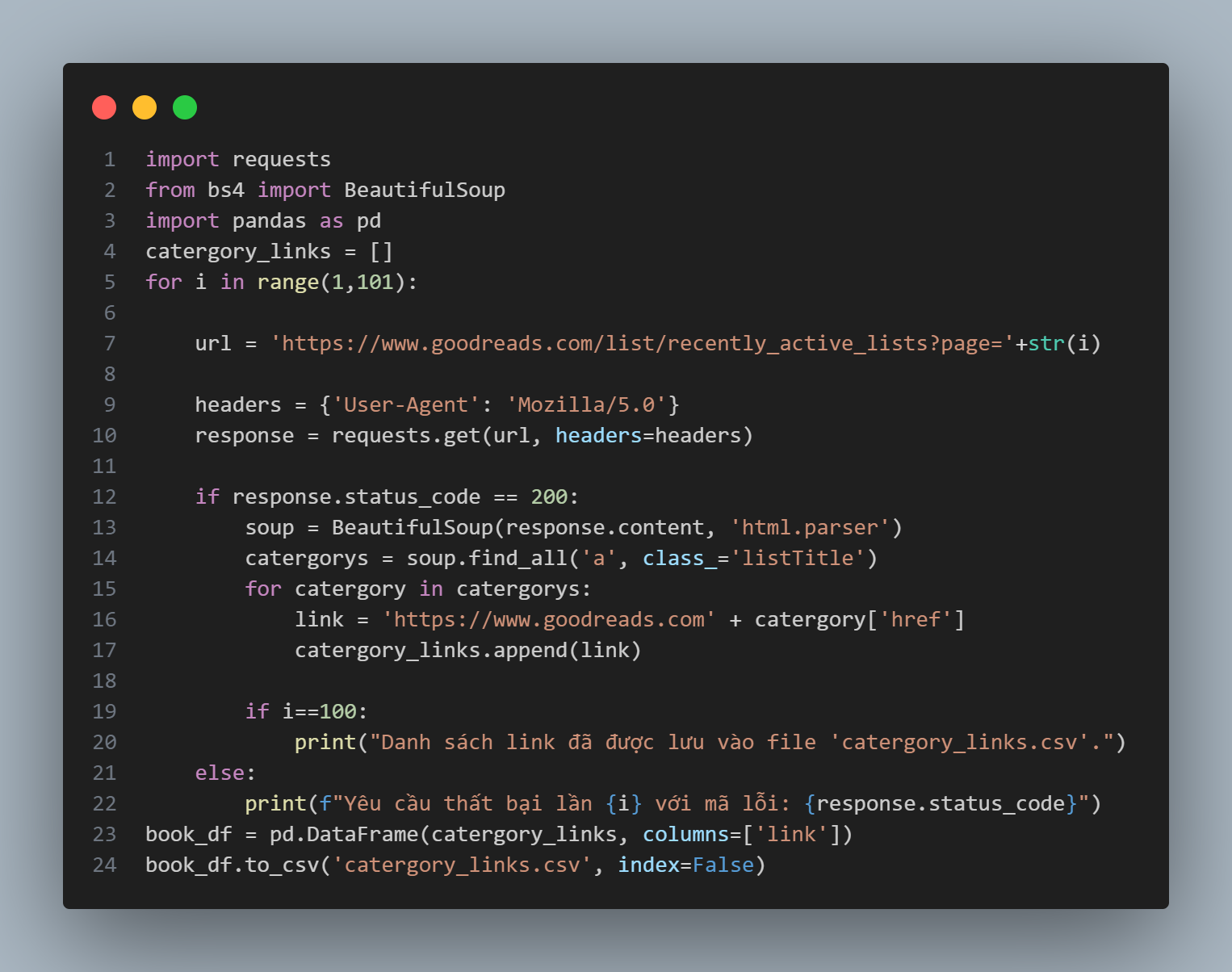
- Tối ưu hóa mô hình qua tuning tham số siêu hoặc thử các thuật toán khác.

- Đánh giá lại mô hình và báo cáo kết quả để cải thiện hiệu suất.

# **IV. Xây dựng hệ thống**

## **1.Thu thập dữ liệu**

- Thu thập các đường dẫn của các danh mục chứa sách ở đường dẫn ‘[listopia | Goodreads](https://www.goodreads.com/list?ref=nav_brws_lists)’



Hình 4.1.1 Thu thập đường dẫn của các danh mục chứa sách

Kết quả trả về ví dụ: “[Best Cozy Mystery Series (2307 books) | Goodreads](https://www.goodreads.com/list/show/3810.Best_Cozy_Mystery_Series)” là đường dẫn của một danh mục chứa sách.

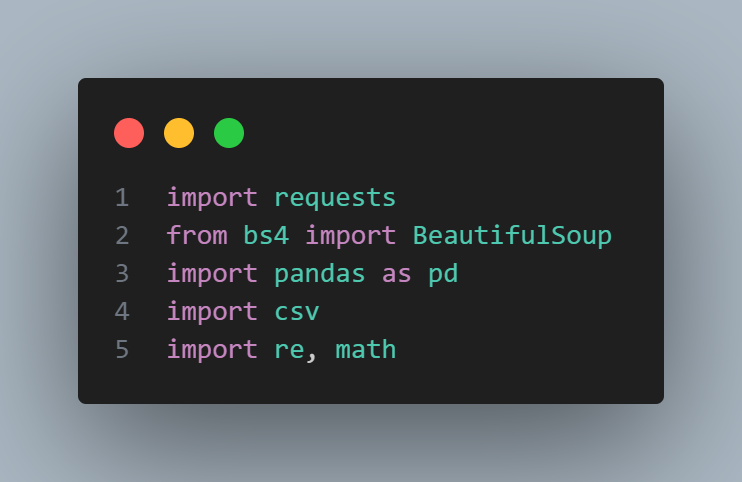
- Thu thập các đường dẫn của các cuốn sách từ các danh mục

Ta sẽ truy cập vào từng đường dẫn danh mục sách ở tệp “catergory\_links

.csv” để lấy tất cả các đường dẫn của các cuốn sách ở tất cả các trang của danh mục đó.

Ta sẽ thu thập tất cả các thông tin đã đề cập ở phần mô tả ở phần III mục 1 thu thập dữ liệu.

Dưới đây là phần xử lí để lấy những thông tin đó:



Hình 4.1.2 Import những thư viện cần thiết

A screen shot of a computer program

Description automatically generated

Hình 4.1.3 Lấy đường dẫn của các cuốn sách(p1)

A screen shot of a computer program

Description automatically generated

Hình 4.1.4 Lấy đường dẫn của các cuốn sách(p2)

Kết quả trả về ví dụ: “[Murder at the Vicarage (Miss Marple, #1) by Agatha Christie | Goodreads](https://www.goodreads.com/book/show/16331.Murder_at_the_Vicarage)” đây là đường dẫn của một cuốn sách được lấy từ danh mục sách “[Best Cozy Mystery Series (2307 books) | Goodreads](https://www.goodreads.com/list/show/3810.Best_Cozy_Mystery_Series)”.

Sau đó, ta sẽ loại bỏ những đường dẫn bị trùng lặp bằng cách dùng hàm “dropduplicates()” của thư viện pandas.

Và bắt đầu trích xuất thông tin của các cuốn sách từ đường dẫn đã thu được ở trên.Dưới đây là một số hàm xử lý để lấy thông tin.



Hình 4.1.5 Hàm lấy số người theo dõi của tác giả.



Hình 4.1.6 Hàm lấy thông tin của cột “genres and votes”(p1).

A screen shot of a computer code

Description automatically generated

Hình 4.1.7 Hàm lấy thông tin của cột “genres and votes”(p2).

A screen shot of a computer program

Description automatically generated

Hình 4.1.8 Lấy thông tin ID của cuốn sách

A screen shot of a computer code

Description automatically generated

Hình 4.1.9 Lấy thông tin tên của cuốn sách

A screen shot of a computer code

Description automatically generated

*Hình 4.1.1*0 *Lấy thông tin series của cuốn sách.*

## **2. Xử lý dữ liệu**

+ Các bước xử lý dữ liệu đã thu thập bao gồm:

* Import thư viện pandas với alias là pd

A black rectangular with colorful dots and text

Description automatically generated

Hình 4.2.1 Import pandas

* Thực hiện đọc các file đã thu thập dữ liệu từ https://www.goodreads.comA screen shot of a computer program

  Description automatically generated

Hình 4.2.2 Đọc dữ liệu

* Bước làm sạch dữ liệu (data cleaning)

1. Loại bỏ những dòng dữ liệu bị trùng lặp

Tập hợp tất cả các thông tin của một cuốn sách về cùng một tệp để loại bỏ dòng dữ liệu bị trùng để tránh việc không đồng bộ.

A screen shot of a computer program

Description automatically generated

Hình 4.2.3 Kết hợp 2 DataFrame

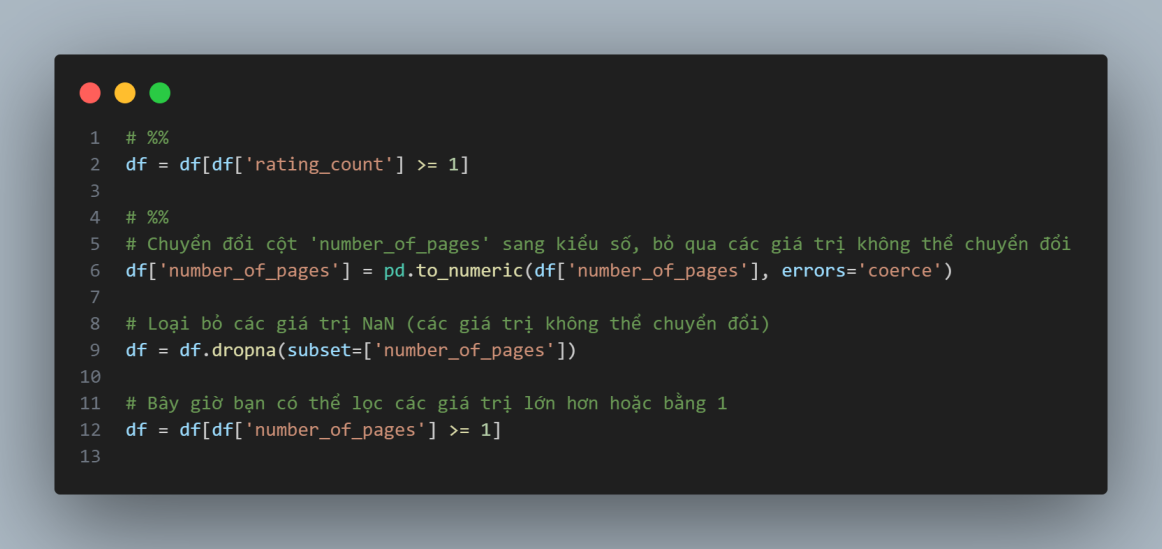
A screen shot of a computer code

Description automatically generated

Hình 4.2.4Loại bỏ trùng lặp và thiết lập lại chỉ mục

1. Cho giá trị của các cột về cùng một kiểu dữ liệu.

Xử lý cột ‘number\_of\_pages’, ‘rating\_count’



Hình 4.2.5 Xử lý cột number\_of\_pages, rating\_count

1. Loại bỏ các cột và hàng trong bảng dữ liệu dựa trên tỷ lệ giá trị thiếu (NaN). Các cột có hơn 50% giá trị NaN sẽ bị loại bỏ, và sau đó các hàng có hơn 50% giá trị NaN cũng sẽ bị loại bỏ. Cuối cùng, chỉ mục của DataFrame mới (df\_dropped\_2) được đặt lại để đảm bảo tính nhất quán trong chỉ số.A computer screen shot of a black screen

   Description automatically generated

Hình 4.2.6 Loại bỏ cột và hàng nếu có số cột/hàng quá 50%

A screenshot of a computer code

Description automatically generated

Hình 4.2.7 Thông tin dữ liệu sau khi xử lý

1. Loại bỏ các cột không liên quan đến việc đánh giá.

Loại bỏ cột ‘link’ và ‘author\_link’, ‘settings’. Các cột này không liên quan đến mục tiêu dự đoán và có nhiều giá trị NaN. Do đó được loại bỏ khỏi DataFrame ‘df\_X’.

A black rectangle with orange and blue text

Description automatically generated

Hình 4.2.8 Loại bỏ link, author\_link

A black rectangular object with text

Description automatically generated

Hình 4.2.9 Loại bỏ settings

1. Xử lý những dòng dữ liệu có giá trị của một cột bất kì bằng NaN

Xử lý giá trị thiếu NaN trong cột ‘followers’ bằng cách thay chúng bằng giá trị ‘0’

A black rectangular with white text

Description automatically generated

Hình 4.2.10 Xử lý followers

1. Loại bỏ các cột có nhiều giá trị độc nhất.

Loại bỏ các cột ‘title’, ‘original\_title’, ‘isbn’, ‘isbn13’, ‘character’’ vì chúng có quá nhiều giá trị độc nhất điều này có thể dẫn đến việc mô hình học được nhiều nhiễu và không hiệu quả.

A black rectangle with orange text

Description automatically generated

Hình 4.2.11 Loại bỏ cột title, original\_title, isbn, íbn13

A black rectangular object with text

Description automatically generated

Hình 4.2.12 Loại bỏ character

Loại bỏ cột ‘author’

A computer screen shot of text

Description automatically generated

Hình 4.2.13 Tính số author duy nhất

A white rectangular object with black text

Description automatically generated

Hình 4.2.14 Số lượng tác giả duy nhất

Do có quá nhiều giá trị khác biệt và đồng thời chúng ta đã có cột ‘followers’ thay thế được độ ảnh hưởng của của tác giả lên đánh giá sách nên ta sẽ loại bỏ cột ‘author’.

A black rectangular with colorful text

Description automatically generated

Hình 4.2.15 Loại bỏ author

* Chuẩn hóa dữ liệu:

Chuẩn hóa dữ liệu của một cột về một thể thống nhất, chuyển đổi dữ liệu định tính sang dữ liệu định lượng.

Chuẩn hóa cột ‘series’: Sẽ xử lý hai bước vì cột này có thể là NaN hoặc một giá trị cụ thể, giá trị này được lưu trữ dưới dạng abc #1. Ở bước đầu của cột này, ta sẽ xử lý khi nó là một giá trị cụ thể bằng cách chuyển đổi abc #1 chỉ còn abc.

A screenshot of a computer program

Description automatically generated

Hình 4.2.16 Xử lý series

A screenshot of a computer

Description automatically generated

Hình 4.2.17Xử lý series

Chuẩn hóa cột ‘date\_published’: dữ liệu của cột này có thể là dd/mm/yyyy có thể là yyyy nên ta sẽ chuyến hết về định đạng yyyy



Hình 4.2.18 Xử lý date\_published

Giải thích logic:

- Kiểm tra nếu giá trị là NaN.

- Sử dụng biểu thức chính quy để tìm tất cả các từ trong chuỗi và lấy từ cuối cùng.

- Nếu từ cuối cùng là số, chuyển đổi thành float và trả về năm.

- Nếu không, trả về NaN.

A screenshot of a computer

Description automatically generated

Hình 4.2.19 Xử lý date\_published

Chuẩn hóa cột ‘genre and votes’: lấy ra genre có số vote cao nhất

A screen shot of a computer program

Description automatically generated

Hình 4.2.20 Xử lý genre and votes

A screenshot of a computer

Description automatically generated

Hình 4.2.21 Xử lý genre và votes

Chuẩn hóa ‘award’:

A screen shot of a computer program

Description automatically generated

Hình 4.2.22 Tính số lượng giải duy nhất

A screenshot of a computer code

Description automatically generated

Hình 4.2.23 Số lượng giải duy nhất

Do có quá nhiều giải khác nhau. Không thể tách thành các cột cho từng giải. Ta xét số lượng giải nhận được. Những dòng NaN sẽ nhận giá trị 0.

A screenshot of a computer program

Description automatically generated

Hình 4.2.24 Xử lý award

Giải thích logic:  
 Hàm formatTextAward(text):

* Kiểm tra nếu giá trị là NaN và trả về 0 nếu đúng.
* Chuyển đổi chuỗi chứa danh sách giải thưởng thành danh sách Python và trả về số lượng giải thưởng. Nếu có lỗi trong quá trình chuyển đổi, nó cũng trả về 0.

Hàm formatAward(df):

* Tạo một bản sao của DataFrame để tránh thay đổi dữ liệu gốc.
* Áp dụng hàm formatTextAward cho cột awards để tính số lượng giải thưởng cho từng hàng.
* Trả về DataFrame đã được cập nhật với số lượng giải thưởng.

A screenshot of a computer program

Description automatically generated

Hình 4.2.25 Xử lý award

Chuẩn hóa cột ‘description’: Vì dữ liệu cột này có quá nhiều sự khác biệt về nội dụng và độ dài giữa các dòng với nhau nên ta sẽ chuyển đổi nó thành số lượng các ký tự.



Hình 4.2.26 Xử lý description

A screenshot of a computer

Description automatically generated

Hình 4.2.27 Xử lý description

Chuẩn hóa các cột mang dữ liệu định tính. Ta sẽ xử lý bằng cách gán số cho nó. Mỗi unique sẽ được gán 1 con số riêng khác nhau.

Xử lý series, publisher:

A screen shot of a computer program

Description automatically generated

Hình 4.2.28 Giải thuật ánh xạ giá trị

Giải thích logic: sử dụng thư viện pickle để lưu trữ và tải lại các đối tượng Python, cho phép lưu trữ ánh xạ (mapping).

# a. Hàm save\_map

- Chức năng: Lưu một đối tượng (trong trường hợp này là map) vào một tệp .pkl để sử dụng sau này.

- Quy trình: Mở tệp trong chế độ ghi nhị phân và sử dụng pickle.dump để lưu đối tượng.

# b. Hàm load\_map

- Chức năng: Tải một đối tượng từ tệp .pkl đã lưu.

- Quy trình: Mở tệp trong chế độ đọc nhị phân và sử dụng pickle.load để lấy lại đối tượng.

# c. Hàm createMappingTable

- Chức năng: Tạo một bảng ánh xạ từ các giá trị văn bản trong một cột của DataFrame thành các số nguyên.

- Quy trình:

- Tạo một bản sao của DataFrame để tránh thay đổi dữ liệu gốc.

- Khởi tạo một từ điển rỗng text\_digit\_vals\_column.

- Lặp qua các giá trị duy nhất trong cột và gán cho mỗi giá trị một số nguyên duy nhất.

- Cập nhật cột trong bản sao DataFrame bằng các số đã ánh xạ.

- Trả về bản sao DataFrame với cột đã được ánh xạ.

# d. Hàm mapData

- Chức năng: Ánh xạ các giá trị văn bản trong cột của DataFrame đã có ánh xạ trước đó thành các số nguyên.

- Quy trình:

- Tương tự như createMappingTable, nhưng sử dụng ánh xạ đã có sẵn để chuyển đổi cột thành các giá trị số.

A screen shot of a computer code

Description automatically generated

Hình 4.2.29 Thực hiện ánh xạ 3 cột: series, publisher, genre

Gọi hàm createMappingTable để ánh xạ các cột 'series', 'publisher', và 'genre' trong DataFrame df\_X.

Đến đây thì đã xử lý xong dữ liệu, tiếp theo ta sẽ chuẩn bị những dữ liệu này cho mô hình dự đoán.

A screen shot of a computer program

Description automatically generated

Hình 4.2.30 Tiền xử lý (p1)

A screen shot of a computer code

Description automatically generated

Hình 4.2.31 Tiền xử lý (p2)

A screenshot of a computer code

Description automatically generated

Hình 4.2.32 Tiền xử lý (p3)

## **3. Trực quan hóa dữ liệu**

Ta sẽ lựa chọn nhữn biểu đồ phù hợp để biểu diễn thông tin của cuốn sách có ảnh hưởng đến việc đánh giá một cuốn sách.

Ta sẽ sử dụng một số thư viện cần thiết như matplotlib, seaborn để có thể vẽ được các biểu đồ

A computer screen shot of a program code

Description automatically generated

Hình 4.3.1 Hàm biểu diễn bằng biểu đồ phân tán

A screen shot of a computer program

Description automatically generated

*Hình* 4.3.2 *Hàm biểu diễn bằng biểu đồ cột*

*A graph of a number of people

Description automatically generated with medium confidence*

*Hình* 4.3.3 *Đồ thị biểu diễn bằng biểu đồ cột*

## **4. Huấn luyện dữ liệu**

- Chia tập dữ liệu huấn luyện và tập dữ liệu kiểm thử theo tỷ lệ 80-20

A black rectangular object with yellow and blue text

Description automatically generated

*Hình* 4.4.1 *Chia tập dữ liệu*

Ta thực hiện huấn luyện dữ liệu trên 3 mô hình:

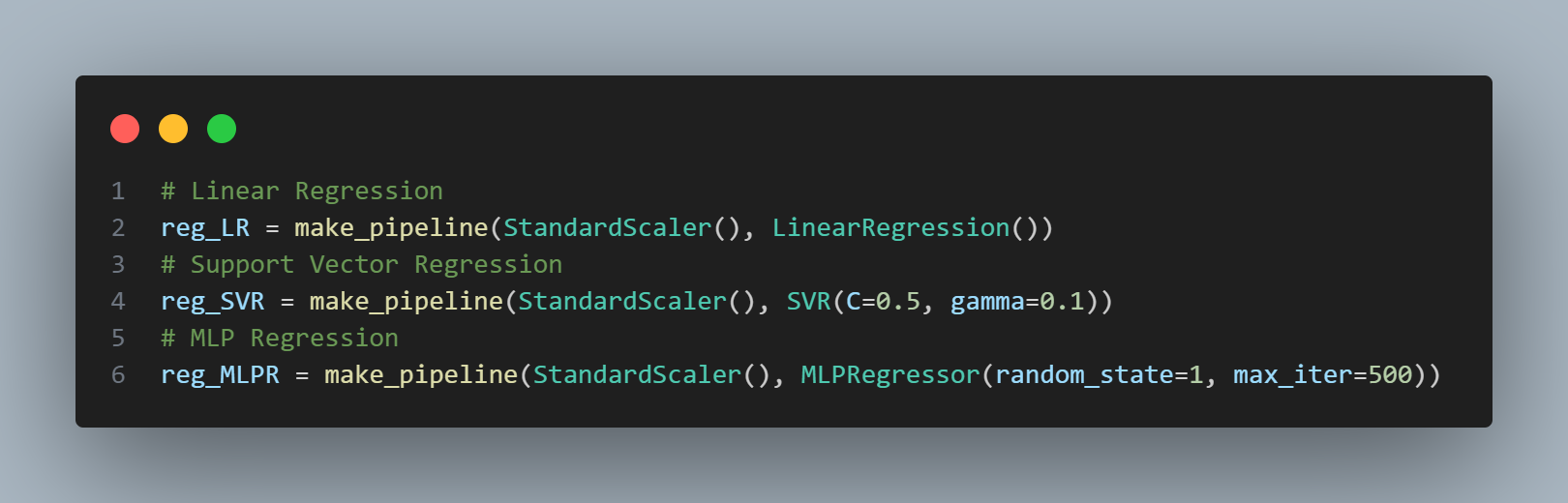
+ Linear Regression

+ Support Vector Regression

+ MLP Regression

Để đưa ra kết luận về việc nên sử dụng mô hình nào sẽ tốt nhất đối với tập dữ liệu đã thu thập được ở trên.

Dưới đây là một số hàm xử lý việc huấn luyện tập dữ liệu



*Hình 4.4.2 Huấn luyện tập dữ liệu trên nhiều mô hình(p1)*

*A black rectangular with white and green text

Description automatically generated*

*Hình 4.4.3 Huấn luyện tập dữ liệu trên nhiều mô hình(p2)*

## **5. Đánh giá mô hình**

Có rất nhiều thang đo để đánh giá một mô hình có tốt hay không như MAE , MSE , R2,..Nhưng ở đây ta sẽ dùng ba thang đo MAE , MSE , R2 để đánh giá kết quả của cả ba mô hình trên. Trước đó ta phải hiểu 3 thang đó có ý nghĩa như thế nào

+ MAE (Mean Absolute Error): Đo lường sai số tuyệt đối trung bình giữa giá trị thực và giá trị dự đoán. Nó thể hiện độ lệch trung bình của dự đoán so với giá trị thực.

+ MSE (Mean Squared Error): Đo lường sai số bình phương trung bình giữa giá trị thực và giá trị dự đoán. Nó nhấn mạnh các sai số lớn hơn.

+ R2 (R-squared): Đo lường mức độ phù hợp của mô hình, thể hiện tỷ lệ biến thiên của dữ liệu đầu ra có thể được giải thích bởi mô hình.

Dưới đây là hàm đưa ra kết quả dựa trên 3 thang đo đối với từng loại mô hình.

A computer screen shot of a program

Description automatically generated

*Hình 4.5.1 Đánh giá mô hình dự đoán bằng cách tính ba thang đo.*

Đây là kết quả của 3 thang đo đối với từng loại mô hình.

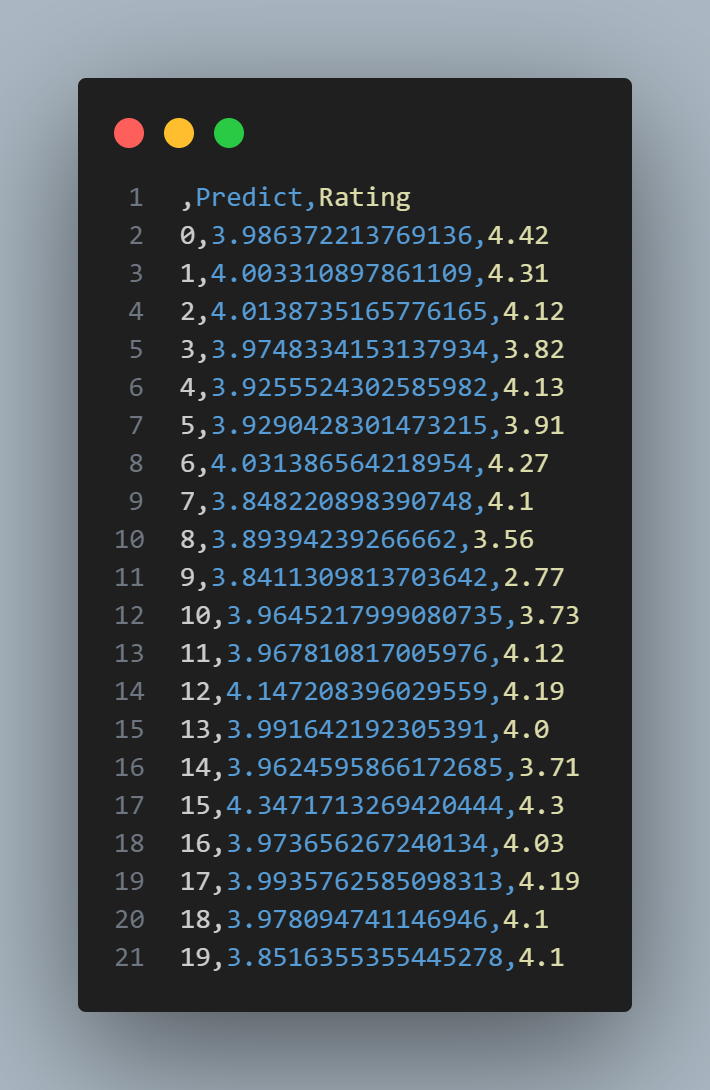
*A screenshot of a computer

Description automatically generated*

*Hình 4.5.2 Kết quả ba thang đo đối với ba loại mô hình.*

# **V. Chương trình minh họa**

Dự đoán rating của một cuốn sách dựa trên tập dữ liệu thử nghiệm



*Hình 5.1.1 Kết quả của mô hình SVR.*

# **VI. Kết luận**

- Tạo ra được một mô hình có thể dự đoán được mức độ đánh giá của một cuốn sách. Nhưng còn cần cải thiện hơn về độ chính xác để mang lại một mô hình tốt hơn