TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

**Logo, company name

Description automatically generated**

**BÁO CÁO CUỐI KÌ MÔN NHẬP MÔN HỌC MÁY – PHẦN LÀM NHÓM**

*Người hướng dẫn:* **GV Lê Anh Cường**

*Người thực hiện:* **Huỳnh Phạm Tố Ngân – 519H0200**

**Trần Tuyết Nhi – 519H0112**

*Lớp:* **19H50201**

*Khoá:* **23**

**HỒ CHÍ MINH, 2023**

**LỜI CẢM ƠN**

Chúng em xin gửi lời cảm ơn chân thành đến thầy Lê Anh Cường – giảng viên bộ môn “Nhập môn học máy” đã trang bị cho em những kiến thức, kỹ năng cơ bản cần có để hoàn thành báo cáo này.

Tuy nhiên trong quá trình hoàn thiện báo cáo, do kiến thức chuyên ngành còn hạn chế nên chúng em vẫn còn nhiều thiếu sót khi tìm hiểu, đánh giá và trình bày. Rất mong nhận được sự quan tâm, góp ý của các thầy/ cô giảng viên bộ môn.

Xin chân thành cảm ơn.

**ĐỒ ÁN ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là sản phẩm đồ án của riêng tôi và được sự hướng dẫn của GV Cao Xuân Phương. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình.** Trường đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 22 tháng 12 năm 2023*

*Tác giả*

*(ký tên và ghi rõ họ tên)*

*Trần Tuyết Nhi*

*Huỳnh Phạm Tố Ngân*

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN

**Phần xác nhận của GV hướng dẫn**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**Phần đánh giá của GV chấm bài**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Tp. Hồ Chí Minh, ngày tháng năm

(kí và ghi họ tên)

**MỤC LỤC**

[PHẦN II: PHẦN LÀM NHÓM 7](#_Toc154265663)

[***(1)*** ***Phân tích thống kê trên dữ liệu, vẽ các đồ thị để hiểu bài toán, hiểu dữ liệu. Tìm hiểu các đặc trưng và đánh giá vai trò của các đặc trưng đối với mục tiêu bài toán:*** 7](#_Toc154265664)

[***(2)*** ***Ứng dụng các mô hình học máy cơ bản để giải quyết bài toán, bao gồm cả các mô hình thuộc Ensemble Learning:*** 11](#_Toc154265665)

[***(3)*** ***Sử dụng Feed Forward Neural Network và Reccurent Neural Network (hoặc mô hình thuộc loại này) để giải quyết bài toán:*** 15](#_Toc154265666)

[***(4)*** ***Áp dụng các kỹ thuật tránh Overfiting trên các mô hình của câu (2) và câu (3) để giải quyết bài toán:*** 18](#_Toc154265667)

[***(5)*** ***Sau khi huấn luyện xong mô hình thì muốn cải thiện độ chính xác, ta sẽ làm gì để giải quyết nó? Phân tích các trường hợp sai, đề ra giải pháp và thực hiện nó, sau đó đánh giá xem có cải tiến so với trước không.*** 21](#_Toc154265668)

# **PHẦN II: PHẦN LÀM NHÓM**

1. ***Phân tích thống kê trên dữ liệu, vẽ các đồ thị để hiểu bài toán, hiểu dữ liệu. Tìm hiểu các đặc trưng và đánh giá vai trò của các đặc trưng đối với mục tiêu bài toán:***
2. *Import thư viện và đọc dữ liệu:*

Bước này nhằm import các thư viện cần thiết cho việc xử lý và trực quan hóa dữ liệu, cũng như đọc dữ liệu từ tệp CSV ('prog\_book.csv'). Việc sử dụng thư viện như Pandas giúp dễ dàng thao tác và phân tích dữ liệu tabular.

A screenshot of a computer program

Description automatically generated

1. *Chuyển đổi dữ liệu và thống kê mô tả:*

Chuyển đổi dữ liệu ở đây giúp đảm bảo rằng cột 'Reviews' được hiểu đúng kiểu dữ liệu (số nguyên) để có thể thực hiện các phép toán số học. Thống kê mô tả cung cấp cái nhìn tổng quan về phân phối và tính chất của dữ liệu, giúp phát hiện giá trị ngoại lai và hiểu rõ hơn về tính chất của các biến.

A screenshot of a computer

Description automatically generated

1. *Trực quan hóa dữ liệu:*
   * Biểu đồ cột:

Hiển thị mối quan hệ giữa loại sách và giá. Giúp nhận diện sự phân bố giá của từng loại sách và so sánh chúng. Có thể phát hiện ra các loại sách có giá cao hoặc thấp hơn so với các loại khác.

A screen shot of a graph

Description automatically generated

Biểu đồ trình bày giá sách theo từng loại. Nhìn vào biểu đồ, chúng ta sẽ thấy sách 'Boxed Set Hardcover Books' là đắt nhất.

* + Biểu đồ hộp:

Hiển thị phân phối của độ đánh giá cho từng loại sách. Biểu đồ hộp giúp định rõ về trung bình, phạm vi và sự phân bố của đánh giá, có thể phát hiện được sự biến động đánh giá giữa các loại sách.

Biểu đồ cột thể hiện mối quan hệ giữa loại sách và giá, và biểu đồ hộp thể hiện phân phối độ đánh giá theo loại sách.

A screenshot of a graph

Description automatically generated

* + Ma trận tương quan:

Hiển thị mối quan hệ tuyến tính giữa các biến. Ma trận tương quan có thể giúp xác định các biến nào có mối quan hệ mạnh mẽ và có thể ảnh hưởng lẫn nhau. Có thể dùng để lựa chọn biến quan trọng cho mô hình hóa.

Ma trận tương quan để xem mối quan hệ tuyến tính giữa các biến.

A screenshot of a computer

Description automatically generated

* + Biểu đồ phân tán:

Trực quan hóa mối quan hệ giữa giá sách và các biến khác như số trang, số lượng đánh giá và độ đánh giá. Giúp xác định xu hướng, điểm ngoại lai và hiểu rõ hơn về phân phối của dữ liệu.

Biểu đồ phân tán để kiểm tra mối quan hệ giữa giá sách và số trang, số lượng đánh giá, và độ đánh giá.

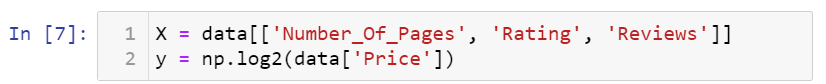
A screenshot of a computer screen

Description automatically generated

1. ***Ứng dụng các mô hình học máy cơ bản để giải quyết bài toán, bao gồm cả các mô hình thuộc Ensemble Learning:***

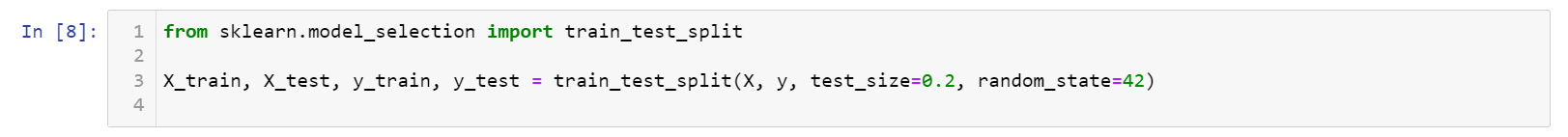
* ***Linear Regression***: Bước này giúp xây dựng một mô hình dự đoán giá sách dựa trên các biến đầu vào được chọn. Đánh giá MSE giúp đánh giá độ chính xác của mô hình, tức là mức độ chệch giữa giá trị dự đoán và giá trị thực tế. Nếu MSE thấp, mô hình có xu hướng dự đoán tốt.
  1. *Chọn biến đầu vào và đầu ra cho mô hình:*

Lựa chọn các biến đầu vào (features) và biến đầu ra (target) để xây dựng mô hình. Trong trường hợp này, biến đầu ra là 'Price', và các biến đầu vào là 'Number\_Of\_Pages', 'Rating', và 'Reviews'. Cột 'Price' được chuyển đổi bằng hàm log để giảm độ biến động và làm cho phân phối dữ liệu đều hơn.



* 1. *Phân chia dữ liệu thành tập huấn luyện và tập kiểm tra:*

Phân chia tập dữ liệu thành tập huấn luyện và tập kiểm tra để đào tạo mô hình trên một phần dữ liệu và đánh giá hiệu suất của nó trên dữ liệu mà mô hình chưa thấy.



* 1. *Xây dựng và huấn luyện mô hình Linear Regression:*

Tạo một đối tượng mô hình Linear Regression và huấn luyện nó trên tập huấn luyện. Mô hình sẽ cố gắng học mối quan hệ tuyến tính giữa các biến đầu vào và biến đầu ra

A close-up of a white box

Description automatically generated

* 1. *Dự đoán và đánh giá mô hình:*

Sử dụng mô hình đã huấn luyện để dự đoán giá trị trên tập kiểm tra và đánh giá hiệu suất của mô hình bằng cách tính Mean Squared Error (MSE), một thước đo đánh giá mức độ chệch giữa giá trị dự đoán và giá trị thực tế.

A close-up of a person

Description automatically generated

* ***Decision Tree:*** Bước này mở rộng phạm vi của phân tích bằng cách sử dụng mô hình Decision Tree, một mô hình dựa trên quyết định. Mô hình này có khả năng học mối quan hệ phi tuyến tính và tạo ra các quy tắc quyết định phức tạp hơn so với mô hình Linear Regression. Đánh giá MSE giúp so sánh hiệu suất giữa mô hình Linear Regression và mô hình Decision Tree trong việc dự đoán giá sách.
  1. *Sử dụng mô hình Decision Tree:*

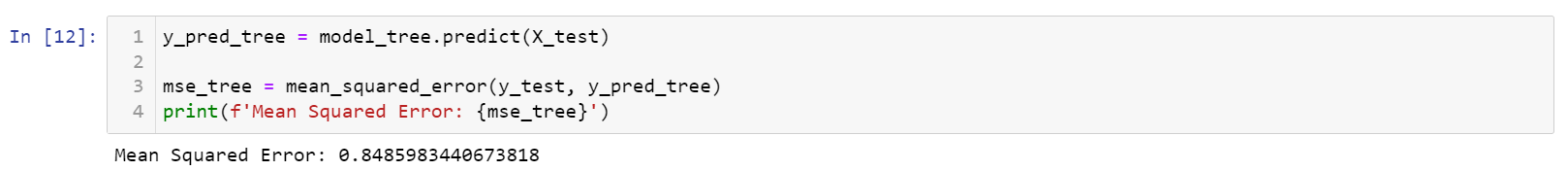
Tạo một đối tượng mô hình Decision Tree và huấn luyện nó trên tập huấn luyện. Mô hình Decision Tree sẽ học cách tạo ra các quy tắc quyết định dựa trên giá trị của các biến đầu vào để dự đoán giá trị đầu ra.

A close-up of a white sign

Description automatically generated

* 1. *Dự đoán và đánh giá mô hình Decision Tree:*

Sử dụng mô hình Decision Tree đã huấn luyện để dự đoán giá trị trên tập kiểm tra và đánh giá hiệu suất của mô hình bằng cách tính Mean Squared Error (MSE). MSE được sử dụng để đánh giá chất lượng của mô hình, trong trường hợp này, của mô hình Decision Tree.



* ***Random Forest:*** Bước này mở rộng phạm vi của phân tích bằng cách sử dụng mô hình Random Forest, một mô hình Ensemble Learning mạnh mẽ. Việc đánh giá MSE giúp so sánh hiệu suất giữa mô hình Random Forest và các mô hình trước đó. Trực quan hóa đặc trưng quan trọng giúp xác định những biến quan trọng nhất đối với mô hình.
  1. *Sử dụng mô hình Random Forest:*

Tạo một đối tượng mô hình Random Forest với 100 cây quyết định và huấn luyện nó trên tập huấn luyện. Random Forest là một mô hình thuộc Ensemble Learning, tức là nó kết hợp nhiều cây quyết định để tạo ra một dự đoán mạnh mẽ và ổn định hơn.

A close-up of a computer screen

Description automatically generated

* 1. *Dự đoán và đánh giá mô hình Random Forest:*

Sử dụng mô hình Random Forest đã huấn luyện để dự đoán giá trị trên tập kiểm tra và đánh giá hiệu suất của mô hình bằng cách tính Mean Squared Error (MSE).



\*n\_estimatiors: số lượng cây quyết định trong rừng. Thường thì càng nhiều cây thì càng tốt, nhưng cũng tăng chi phí tính toán.

\*random\_state: 1 giá trị seed được sử dụng để đảm bảo tính nhất quán khi huấn luyện model.

\*Mỗi cây trong rừng được huấn luyện trên một tập dữ liệu con ngẫu nhiên được rút ra từ tập dữ liệu đào tạo. Điều này tạo ra sự đa dạng giữa các cây và giúp mô hình chống lại overfitting.

* 1. *Trực quan hóa đặc trưng quan trọng của mô hình:*

Trực quan hóa đặc trưng quan trọng của mô hình Random Forest. Điều này giúp hiểu rõ hơn về vai trò của từng biến đầu vào trong quá trình dự đoán giá sách. Biểu đồ cột ngang (barh) hiển thị 10 đặc trưng quan trọng nhất theo mô hình.

A screenshot of a graph

Description automatically generated

1. ***Sử dụng Feed Forward Neural Network và Reccurent Neural Network (hoặc mô hình thuộc loại này) để giải quyết bài toán:***

* ***Neural Network:*** Bước này giúp áp dụng mô hình Neural Network, một mô hình mạnh mẽ trong học máy, để dự đoán giá sách. Đánh giá MSE giúp so sánh hiệu suất của mô hình Neural Network với các mô hình trước đó và đánh giá chất lượng của dự đoán.
  1. *Xây dựng mô hình Neural Network:*

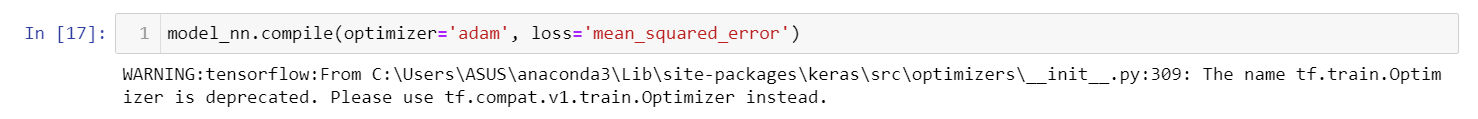
Xây dựng một mô hình Neural Network với kiến trúc đơn giản bao gồm ba lớp (lớp đầu vào có 64 nơ-ron, lớp ẩn có 32 nơ-ron và lớp đầu ra có 1 nơ-ron). Các hàm kích thích ('relu' cho các lớp ẩn và 'linear' cho lớp đầu ra) được chọn để tối ưu hóa hiệu suất.

A screenshot of a computer code

Description automatically generated

* 1. *Compile mô hình:*

Compile mô hình bằng cách chọn optimizer là 'adam' và hàm mất mát là 'mean\_squared\_error'. 'adam' là một thuật toán tối ưu hóa và 'mean\_squared\_error' là hàm mất mát được sử dụng để đánh giá hiệu suất của mô hình trong trường hợp dự đoán giá trị số.



* 1. *Huấn luyện mô hình Neural Network:*

Huấn luyện mô hình trên tập huấn luyện với 50 epochs và batch size là 32. Validation\_split=0.2 chia tập huấn luyện thành tập huấn luyện và tập validation, giúp kiểm tra và theo dõi hiệu suất của mô hình trong quá trình huấn luyện.

A screenshot of a computer

Description automatically generated

* 1. *Dự đoán và đánh giá mô hình Neural Network:*

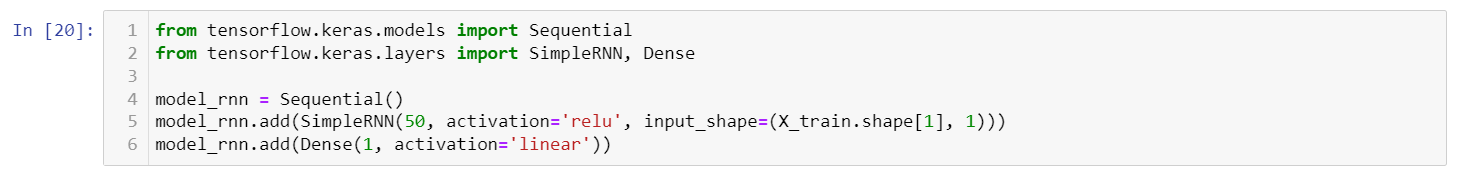
Sử dụng mô hình Neural Network đã huấn luyện để dự đoán giá trị trên tập kiểm tra và đánh giá hiệu suất của mô hình bằng cách tính Mean Squared Error (MSE).

A white rectangular object with a black border

Description automatically generated

* ***Reccurent Neural Network:*** Bước này giúp áp dụng mô hình Recurrent Neural Network, một mô hình mạnh mẽ trong học máy, để dự đoán giá sách. Đánh giá MSE giúp so sánh hiệu suất của mô hình RNN với các mô hình trước đó và đánh giá chất lượng của dự đoán. Mô hình RNN thích hợp cho các tác vụ dự đoán dựa trên chuỗi thời gian hoặc dữ liệu có tính chất chuỗi.
  1. ***Xây dựng mô hình Reccurent Neural Network:***

**Xây dựng một mô hình Recurrent Neural Network (RNN) với một lớp SimpleRNN có 50 nơ-ron và hàm kích thích 'relu'. Lớp đầu vào của mô hình được thiết lập để chấp nhận dữ liệu theo dạng (số lượng quan sát, số lượng biến, chiều dài chuỗi), trong đó chiều dài chuỗi là 1 (1 chiều).**

****

* 1. ***Compile mô hình:***

**Compile mô hình bằng cách chọn optimizer là 'adam' và hàm mất mát là 'mean\_squared\_error'. 'adam' là một thuật toán tối ưu hóa và 'mean\_squared\_error' là hàm mất mát được sử dụng để đánh giá hiệu suất của mô hình trong trường hợp dự đoán giá trị số.**

****

* 1. ***Huấn luyện mô hình Reccurent Neural Network:***

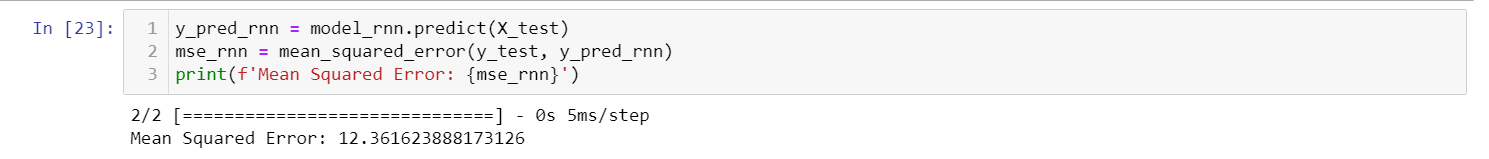
**Huấn luyện mô hình trên tập huấn luyện với 50 epochs và batch size là 32. Validation\_split=0.2 chia tập huấn luyện thành tập huấn luyện và tập validation, giúp kiểm tra và theo dõi hiệu suất của mô hình trong quá trình huấn luyện.**

**A screenshot of a computer

Description automatically generated**

* 1. ***Dự đoán và đánh giá mô hình Recurrent Neural Network:***

**Sử dụng mô hình Recurrent Neural Network đã huấn luyện để dự đoán giá trị trên tập kiểm tra và đánh giá hiệu suất của mô hình bằng cách tính Mean Squared Error (MSE).**



1. ***Áp dụng các kỹ thuật tránh Overfiting trên các mô hình của câu (2) và câu (3) để giải quyết bài toán:***

Kỹ thuật ***Cross-Validation*** giúp đánh giá mô hình Linear Regression trên nhiều tập dữ liệu con khác nhau, giúp đảm bảo rằng độ đo hiệu suất được tính toán không phụ thuộc quá nhiều vào cách chia dữ liệu. So sánh MSE của các mô hình khác nhau giúp lựa chọn mô hình có hiệu suất tốt nhất cho bài toán dự đoán giá sách.

1. *Cross-Validation cho Linear Regression:*

Sử dụng kỹ thuật Cross-Validation để đánh giá hiệu suất của mô hình Linear Regressuin trên tập dữ liệu. Trong trường hợp này, sử dụng 5-fold-Cross-Validation và đánh giá mô hình dựa trên Mean Squared Error (MSE).

**A white screen with black text

Description automatically generated**

1. ***So sánh MSE của các mô hình:***

**So sánh Mean Squared Error (MSE) của mô hình Linear Regression với các mô hình khác, bao gồm Decision Tree, Random Forest, Neural Network và Recurrent Neural Network. Biểu đồ cột hiển thị MSE của từng mô hình để so sánh chất lượng dự đoán.**

A screenshot of a computer

Description automatically generated

* Linear Regression (MSE: 0.848): Mô hình này có MSE tương đối thấp, cho thấy nó dự đoán khá tốt.
* Decision Tree (MSE: 0.863): MSE cao hơn so với Linear Regression, có thể là do mô hình quá phức tạp hoặc overfitting.
* Random Forest (MSE: 0.503): MSE thấp hơn so với Decision Tree, đây là một dấu hiệu tích cực của sức mạnh của Ensemble Learning.
* Neural Network (MSE: 4.675): MSE rất cao, có thể là do mô hình quá phức tạp, cần kiểm tra lại cấu trúc mô hình và thử tuning tham số.
* Recurrent Neural Network (MSE: 1.946): MSE cao hơn so với Random Forest nhưng thấp hơn so với Neural Network, có thể là do RNN không phù hợp cho bài toán cụ thể này.
* So sánh:
  + Random Forest có vẻ là mô hình tốt nhất với MSE thấp nhất.
  + Linear Regression và Decision Tree cũng có hiệu suất tốt, nhưng Linear Regression đôi chút ổn định hơn.
  + Neural Network và Recurrent Neural Network có MSE cao hơn đáng kể, có thể cần kiểm tra lại cấu trúc mô hình và thực hiện tuning tham số.
* ***Cross-Validation MSE Scores:***
* Điểm số MSE trên từng fold của cross-validation: [1.49, 0.23, 0.22, 0.31, 1.17].
* Kết quả này cho thấy mức độ chênh lệch giữa các fold. Một số fold có MSE cao hơn so với các fold khác, đây có thể là dấu hiệu của overfitting hoặc phương sai cao.
* ***Mean MSE:***
* MSE trung bình của tất cả các fold là 0.68.
* Đây là một chỉ số tổng quan về hiệu suất trung bình của mô hình trên toàn bộ cross-validation.
* ***Phân Tích Chi Tiết và Đề Xuất Giải Pháp:***
* Overfitting: Sự chênh lệch cao giữa các fold và MSE trung bình có thể là dấu hiệu của overfitting, tức là mô hình quá tốt trên dữ liệu huấn luyện nhưng không tổng quát hóa tốt cho dữ liệu mới.
* Giải Pháp: Có thể thử giảm độ phức tạp của mô hình, tăng lượng dữ liệu huấn luyện, hoặc sử dụng kỹ thuật chính quy như L1 hoặc L2 regularization để kiểm soát overfitting.
* ***So Sánh Hiệu Suất:***
* Để so sánh hiệu suất giữa các mô hình khác nhau, bạn có thể thực hiện so sánh trực tiếp các MSE hoặc sử dụng các chỉ số đánh giá khác như R-squared, Mean Absolute Error (MAE), hay Root Mean Squared Error (RMSE).
* Quan trọng nhất là, nếu MSE vẫn cao, bạn có thể cần xem xét lại các yếu tố như chất lượng dữ liệu, tính đại diện của các biến đầu vào, và lựa chọn mô hình. Cải thiện mô hình có thể đòi hỏi sự thay đổi đa dạng, từ tinh chỉnh siêu tham số đến thêm biến đầu vào hay sử dụng các kỹ thuật chính quy.

1. ***Sau khi huấn luyện xong mô hình thì muốn cải thiện độ chính xác, ta sẽ làm gì để giải quyết nó? Phân tích các trường hợp sai, đề ra giải pháp và thực hiện nó, sau đó đánh giá xem có cải tiến so với trước không.***
2. *Phân tích các trường hợp sai:*

* Xem xét các điểm dữ liệu mà mô hình dự đoán sai lệch lớn so với giá trị thực tế.
* Kiểm tra xem có xuất hiện các outlier không dự đoán được hay không.
* Xem xét các biến đầu vào và kiểm tra liệu có sự tương quan mạnh giữa chúng không.

1. *Giải pháp cho các trường hợp sai:*

* Xử lý outlier: Nếu có outlier, xem xét xử lý chúng bằng cách giảm ảnh hưởng của chúng hoặc loại bỏ chúng nếu cần thiết.
* Kiểm tra lại dữ liệu: Đảm bảo rằng dữ liệu được làm sạch và chuẩn hóa đúng cách, có thể thực hiện thêm các bước tiền xử lý như loại bỏ giá trị thiếu, chuẩn hóa dữ liệu, hay chuyển đổi biến phù hợp.
* Kiểm tra mô hình: Xem xét kiến trúc của mô hình và xem liệu nó phản ánh đúng mối quan hệ trong dữ liệu hay không. Có thể thử nghiệm với các kiến trúc mô hình khác nhau để tìm ra một kiến trúc phù hợp hơn.

1. *Tăng cường dữ liệu (Data Augmentation):*

* Nếu dữ liệu ít, có thể tăng cường dữ liệu bằng cách tạo thêm các biến mới từ dữ liệu hiện có, thay đổi góc chụp, ánh sáng, hay áp dụng các biến đổi khác để mô hình học được sự đa dạng của dữ liệu.

1. *Thử nghiệm mô hình mới hoặc tinh chỉnh tham số:*

* Thử nghiệm các mô hình máy học khác nhau hoặc tinh chỉnh các tham số của mô hình hiện tại để xem xét liệu có cải thiện độ chính xác không. Điều này có thể bao gồm việc điều chỉnh siêu tham số, tăng số lượng lớp, hoặc sử dụng mô hình có hiệu suất cao hơn.

1. *Kiểm tra đánh giá mô hình:*

* Sau khi thực hiện các biện pháp cải thiện, đánh giá lại mô hình trên tập kiểm tra và tập validation để xem liệu có cải thiện độ chính xác không. Sử dụng các độ đo đánh giá phù hợp như Mean Squared Error, R-squared, hay các độ đo khác tùy thuộc vào bối cảnh bài toán.

1. *Tích hợp các mô hình Ensemble:*

* Kết hợp nhiều mô hình dự đoán để tạo thành một mô hình Ensemble có thể cải thiện độ chính xác. Các mô hình Ensemble, chẳng hạn như Bagging hoặc Boosting, có thể giúp giảm overfitting và tăng tính ổn định của mô hình.

**THAM KHẢO**

[1] [Phân lớp bằng Random Forests trong Python](https://viblo.asia/p/phan-lop-bang-random-forests-trong-python-djeZ1D2QKWz)

[2] [Decision Tree algorithm](https://machinelearningcoban.com/tabml_book/ch_model/decision_tree.html)

[3] [Random Forest algorithm](https://machinelearningcoban.com/tabml_book/ch_model/random_forest.html)

[4] [Top 270 Computer Science / Programing Books](https://www.kaggle.com/datasets/thomaskonstantin/top-270-rated-computer-science-programing-books)