**BỘ GIÁO DỤC VÀ ĐÀO TẠO**



**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT HƯNG YÊN**

BÀI TẬP LỚN

**HỌC MÁY CƠ BẢN**

APPLYING MACHINE LEARING TO PREDICT GDP BASED ON LABOR STRUCTURE DATA AND UNEMPLOYMENT RATE

NGÀNH: CÔNG NGHỆ PHẦN MỀM

SINH VIÊN: **1. NGUYỄN VĂN LINH**

**2. ĐÀO ANH QUÂN**

LỚP: **12522W.2**

NGƯỜI HƯỚNG DẪN: **TS. HOÀNG QUỐC VIỆT**

**HƯNG YÊN – 2025**

# NHẬN XÉT

## Nhận xét của giáo viên hướng dẫn

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

...........................................................................................................................................

## GIÁO VIÊN HƯỚNG DẪN

**Hoàng Quốc Việt**

**MỤC LỤC**

[BỘ GIÁO DỤC VÀ ĐÀO TẠO 1](#_Toc212066575)

[NHẬN XÉT 2](#_Toc212066576)

[Nhận xét của giáo viên hướng dẫn 2](#_Toc212066577)

[GIÁO VIÊN HƯỚNG DẪN 2](#_Toc212066578)

[DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ 6](#_Toc212066579)

[CHƯƠNG 1: GIỚI THIỆU BÀI TOÁN 9](#_Toc212066580)

[1.1 Bài toán 9](#_Toc212066581)

[1.3 Tiền xử lý dữ liệu 11](#_Toc212066582)

[1.4 Làm sạch dữ liệu 12](#_Toc212066583)

[1.5 Trực quan hoá dữ liệu 12](#_Toc212066584)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 13](#_Toc212066585)

[2.1 Random Forest Regressor 13](#_Toc212066586)

[2.1.1 Giới thiệu thuật toán Random Forest Regressor 13](#_Toc212066587)

[2.1.2 Cấu trúc của mô hình RandomForest 13](#_Toc212066588)

[2.1.3 Quá trình xây dựng mô hình RandomForest 13](#_Toc212066589)

[2.1.4 Ưu điểm của RandomForest 13](#_Toc212066590)

[2.1.5 Nhược điểm của RandomForest 14](#_Toc212066591)

[2.1.6 Đánh giá hiệu suất 14](#_Toc212066592)

[2.2 Linear Regressor 14](#_Toc212066593)

[2.2.1 Giới thiệu thuật toán Linear Regressor 14](#_Toc212066594)

[2.2.2 Cấu trúc của mô hình Extra Trees Regressor 15](#_Toc212066595)

[2.2.3 Quá trình xây dựng mô hình Linear Regression 16](#_Toc212066596)

[2.2.4 Ưu điểm của Linear Regression 16](#_Toc212066597)

[2.2.5 Nhược điểm của KNN 16](#_Toc212066598)

[2.2.6 Đánh giá hiệu suất 16](#_Toc212066599)

[2.3 Support Vector Regression (SVR) 17](#_Toc212066600)

[2.3.1 Giới thiệu thuật toán SVR 17](#_Toc212066601)

[2.3.2 Cấu trúc của mô hình SVR 18](#_Toc212066602)

[2.3.3 Quá trình xây dựng mô hình SVR 18](#_Toc212066603)

[2.3.4 Ưu điểm của SVR 18](#_Toc212066604)

[2.3.5 Nhược điểm của SVR 18](#_Toc212066605)

[2.3.6 Đánh giá hiệu suất 18](#_Toc212066606)

[CHƯƠNG 3: GIẢI PHÁP, CÀI ĐẶT THỰC HIỆN 19](#_Toc212066607)

[3.1 Mã nguồn tiền xử lý dữ liệu 19](#_Toc212066608)

[3.2 Mã nguồn chức năng làm sạch dữ liệu 20](#_Toc212066609)

[3.2.1 Xử lý giá trị thiếu 20](#_Toc212066610)

[3.2.2 Xoá outlier 22](#_Toc212066611)

[1) Tính IQR: 23](#_Toc212066612)

[2) Giới hạn: 23](#_Toc212066613)

[3) Lọc dữ liệu: 23](#_Toc212066614)

[4) boxplots(data\_no\_outliers): 23](#_Toc212066615)

[3.2.3 Ma trận tương quan 24](#_Toc212066616)

[1. numeric\_df: 24](#_Toc212066617)

[2. plt.figure(figsize=(10, 6)): 24](#_Toc212066618)

[3. sns.heatmap(): 24](#_Toc212066619)

[4. plt.title(): 24](#_Toc212066620)

[3.3 Mã nguồn chức năng Trực quan hoá dữ liệu 25](#_Toc212066621)

[3.3.1 Biểu đồ thể hiện phân phối đặc trưng các phân loại 25](#_Toc212066622)

[3.3.2 Biểu đồ thể hiện tỷ lệ thất nghiệp trung bình của 15 quốc gia ngẫu nhiên 27](#_Toc212066623)

[3.3.3 Biểu đồ thể hiện các đặc trưng liên tục 28](#_Toc212066624)

[3.3.4 Biểu đồ thể hiện xu hướng tăng GDP theo thời gian của một số quốc gia 29](#_Toc212066625)

[3.3.5 Biểu đồ thể hiện Biểu đồ thể hiện mối quan hệ giữa GDP và các nhóm lao động 30](#_Toc212066626)

[3.3.6 Biểu đồ so sánh GDP thực tế và GDP dự đoán từ mô hình Extra Trees Regressor 31](#_Toc212066627)

[3.4 Chuyển đổi dữ liệu 33](#_Toc212066628)

[3.4.1 Xử lý dữ liệu chuỗi 33](#_Toc212066629)

[3.4.2 Xử lý dữ liệu số 33](#_Toc212066630)

[3.4.3 Chia dữ liệu huấn luyện và kiểm thử 34](#_Toc212066631)

[3.5 Huấn luyện mô hình SKLEARN 34](#_Toc212066632)

[3.5.1 Tìm kiếm mô hình 34](#_Toc212066633)

[3.5.2 Đánh giá hiệu suất mô hình 35](#_Toc212066634)

[3.5.3 Huấn luyện mô hình Extra Trees Regressor 35](#_Toc212066635)

[3.5.4 Huấn luyện mô hình Gradient Boosting Regressor 38](#_Toc212066636)

[3.5.5 Huấn luyện mô hình Linear Regression 41](#_Toc212066637)

[3.5.6 Huấn luyện mô hình Random Forest Regressor 44](#_Toc212066638)

[3.5.7 Huấn luyện mô hình Support Vector Regression (SVR) 47](#_Toc212066639)

[3.5.8 So sánh và tổng kết giai đoạn huấn luyện 50](#_Toc212066640)

[CHƯƠNG 4: XÂY DỰNG ỨNG DỤNG 52](#_Toc212066641)

[KẾT LUẬN 53](#_Toc212066642)

[1. Kết quả đạt được trong kiến thức: 53](#_Toc212066643)

[2. Hạn chế của đề tài: 54](#_Toc212066644)

[3. Hướng phát triển của đề tài: 54](#_Toc212066645)

[TÀI LIỆU THAM KHẢO 56](#_Toc212066646)

# DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

[Hình 1. 1 Hiển thị 5 hàng đầu và 5 hàng cuối của dữ liệu 9](#_Toc212067630)

[Hình 1. 2 Thông tin tổng quan về cấu trúc Dataframe 9](#_Toc212067631)

[Hình 1. 3 Thông tin mô tả về dữ liệu số 10](#_Toc212067632)

[Hình 3. 1 Import thư viện cần thiết, load, hiển thị 5 hàng đầu của DataFrame 18](#_Toc212067633)

[Hình 3. 2 Kết quả hiển thị 5 hàng đầu của DataFrame 18](#_Toc212067634)

[Hình 3. 4 Thông tin về DataFrame 19](#_Toc212067635)

[Hình 3. 5 Mã nguồn xử lý giá trị thiếu 19](#_Toc212067636)

[Hình 3. 6 Hiển thị phần trăm giá trị thiếu của từng cột 20](#_Toc212067637)

[Hình 3. 7 Loại bỏ tất cả các hàng trong DataFrame data có ít nhất một giá trị bị thiếu (NaN) và hiển thị 5 hàng cuối của DataFrame 20](#_Toc212067638)

[Hình 3. 8 Kết quả hiển thị 5 hàng cuối của DataFrame 20](#_Toc212067639)

[Hình 3. 9 Mã nguồn đếm số lượng giá trị thiếu (NaN) trong từng cột của DataFrame data. 21](#_Toc212067640)

[Hình 3. 10 Kết quả liệt kê số giá trị thiếu theo cột. 21](#_Toc212067641)

[Hình 3. 11 Mã nguồn xử lý dữ liệu ngoại lai 21](#_Toc212067642)

[Hình 3. 12 Boxplot minh họa dữ liệu sạch. 22](#_Toc212067643)

[Hình 3. 13 Mã nguồn ma trận tương quan 23](#_Toc212067644)

[Hình 3. 14 Kết quả Heatmap minh họa mức độ tương quan giữa các yếu tố 24](#_Toc212067645)

[Hình 3. 15 Mã nguồn biểu đồ thể hiện phân phối đặc trưng các phân loại 25](#_Toc212067646)

[Hình 3. 16 Kết quả biểu đồ phân phôi GDP 25](#_Toc212067647)

[Hình 3. 17 Kết quả biểu đồ pie thể hiện phân phối đặc trưng các phân loại lao động của Việt Nam 2004 25](#_Toc212067648)

[Hình 3. 18 Mã nguồn biểu đồ thể hiện tỷ lệ thất nghiệp trung bình của 15 quốc gia ngâu nhiên 26](#_Toc212067649)

[Hình 3. 19 Kết quả biểu đồ thể hiện tỷ lệ thất nghiệp 26](#_Toc212067650)

[Hình 3. 20 Mã nguồn biểu đồ thể hiện các đặc trưng liên tục 27](#_Toc212067651)

[Hình 3. 21 Kết quả biểu đồ thể hiện các đặc trưng liên tục 27](#_Toc212067652)

[Hình 3. 22 Mã nguồn biểu đồ thể hiện xu hương tăng GDP theo thời gian của một số quốc gia 28](#_Toc212067653)

[Hình 3. 23 Kết quả biểu đồ thể hiện xu hương tăng GDP theo thời gian của một số quốc gia 28](#_Toc212067654)

[Hình 3. 24 Mã nguồn biểu đồ thể hiện mối tương quan giữa GDP và các phần tử còn lại 29](#_Toc212067655)

[Hình 3. 25 Kết quả biểu đồ thể hiện mối tương quan giữa GDP và Nông nghiệp 29](#_Toc212067656)

[Hình 3. 26 Kết quả biểu đồ thể hiện mối tương quan giữa GDP và Công nghiệp 29](#_Toc212067657)

[Hình 3. 27 Kết quả biểu đồ thể hiện mối tương quan giữa GDP và Dịch vụ 30](#_Toc212067658)

[Hình 3. 28 Mã nguồn biểu đồ thể hiện sự trực quan hoá giá trị thiếu 30](#_Toc212067659)

[Hình 3. 29 Kết quả biểu đồ thể hiện sự trực quan hoá giá trị thiếu 31](#_Toc212067660)

[Hình 3. 30 Import 1 số thư viện cần thiết 32](#_Toc212067661)

[Hình 3. 31 Mã nguồn xử lý dữ liệu chuỗi 32](#_Toc212067662)

[Hình 3. 32 Kết quả xử lý dữ liệu chuỗi 32](#_Toc212067663)

[Hình 3. 33 Mã nguồn xử lý dữ liệu số 32](#_Toc212067664)

[Hình 3. 34 Kết quả xử lý dữ liệu số 32](#_Toc212067665)

[Hình 3. 35 Mã nguồn chia dữ liệu huấn luyện và kiểm thử 33](#_Toc212067666)

[Hình 3. 36 Mã nguồn chia dữ liệu huấn luyện và kiểm thử 33](#_Toc212067667)

[Hình 3. 37 Mã nguồn tìm kiếm mô hình dùng Lazypredict 33](#_Toc212067668)

[Hình 3. 38 Mã nguồn đánh giá hiệu suất mô hình 34](#_Toc212067669)

[Hình 3. 37 Mã nguồn so sánh GDP thực tế và GDP dự đoán 34](#_Toc212067670)

[Hình 3. 38 Kết quả so sánh GDP thực tế và GDP dự đoán 35](#_Toc212067671)

[Hình 3. 40 Mã nguồn sai số GĐP 35](#_Toc212067672)

[Hình 3. 41 Kết quả sai số GĐP 36](#_Toc212067673)

[Hình 3. 42 Mã nguồn tần số phân phối sai số GĐP 36](#_Toc212067674)

[Hình 3. 43 Kết quả tần số phân phối sai số GĐP 37](#_Toc212067675)

[Hình 3. 44 Mã nguồn biểu đồ so sánh GDP thực tế và GDP dự đoán 37](#_Toc212067676)

[Hình 3. 45 Kết quả biểu đồ so sánh GDP thực tế và GDP dự đoán 38](#_Toc212067677)

[Hình 3. 46 Mã nguồn biểu đồ sai số GDP 38](#_Toc212067678)

[Hình 3. 47 Kết quả biểu đồ sai số GDP 39](#_Toc212067679)

[Hình 3. 48 Mã nguồn biểu đồ phân phối sai số GDP 39](#_Toc212067680)

[Hình 3. 49 Kết quả biểu đồ phân phối sai số GDP 40](#_Toc212067681)

[Hình 3. 50 Mã nguồn biểu đồ so sánh GDP thực tế vơi GDP dự đoán 40](#_Toc212067682)

[Hình 3. 51 Kết quả biểu đồ so sánh GDP thực tế vơi GDP dự đoán 41](#_Toc212067683)

[Hình 3. 52 Mã nguồn biểu đồ sai số GDP 41](#_Toc212067684)

[Hình 3. 53 Kết quả biểu đồ sai số GDP 42](#_Toc212067685)

[Hình 3. 54 Mã nguồn phân phối sai số GDP 42](#_Toc212067686)

[Hình 3. 56 Kết quả phân phối sai số GDP 43](#_Toc212067687)

[Hình 3. 57 Mã nguồn biểu đồ so sánh GDP thực tế vơi GDP dự đoán 43](#_Toc212067688)

[Hình 3. 58 Kết quả biểu đồ so sánh GDP thực tế vơi GDP dự đoán 44](#_Toc212067689)

[Hình 3.59 Mã nguồn biểu đồ sai số GDP 44](#_Toc212067690)

[Hình 3. 60 Kết quả biểu đồ sai số GDP 45](#_Toc212067691)

[Hình 3. 61 Mã nguồn phân phối sai số GDP 45](#_Toc212067692)

[Hình 3.62 Kết quả phân phối sai số GDP 46](#_Toc212067693)

[Hình 3.63 Mã nguồn biểu đồ so sánh GDP thực tế vơi GDP dự đoán 46](#_Toc212067694)

[Hình 3.64 Kết quả biểu đồ so sánh GDP thực tế vơi GDP dự đoán 47](#_Toc212067695)

[Hình 3.64 Mã nguồn biểu đồ sai số GDP 47](#_Toc212067696)

[Hình 3.65 Kết quả biểu đồ sai số GDP 48](#_Toc212067697)

[Hình 3.66 Mã nguồn phân phối sai số GDP 48](#_Toc212067698)

[Hình 3.67 Kết quả phân phối sai số GDP 49](#_Toc212067699)

[Hình 3.68 Mã nguồn so sách các mô hình 49](#_Toc212067700)

[Hình 3.69 Kết quả so sách các mô hình 49](#_Toc212067701)

[Hình 3.70 Mã nguồn so sánh R² giữa các mô hình 50](#_Toc212067702)

[Hình 3.71 Kết quả so sánh R² giữa các mô hình 50](#_Toc212067703)

[HÌnh 4. 1 Sử dụng Streamlit để đưa bài toán lên giao diện web 51](#_Toc212067704)

# CHƯƠNG 1: GIỚI THIỆU BÀI TOÁN

## Bài toán

Tổng sản phẩm quốc nội(GDP) là một trong những chỉ số quan trọng phản ánh mức độ phát triển của một quốc gia. Dự đoán chính xác GDP thông qua việc phân tích dữ liệu và áp dụng các thuật toán học máy có thể hỗ trợ các chính phủ, tổ chức và doanh nghiệp xây dựng chính sách kinh tế, kế hoạch đầu tư và các chiến lược phát triển bển vững.

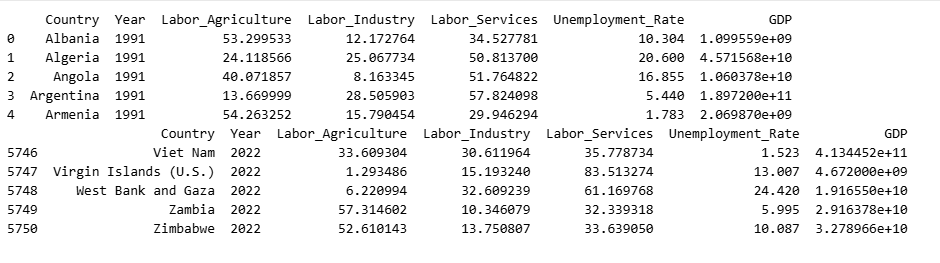
Bài toán "Áp dụng các thuật toán học máy vào dự đoán GDP dựa trên dữ liệu cơ cấu lao động và tỷ lệ thất nghiệp" tập trung vào việc khai thác dữ liệu kinh tế - xã hội và áp dụng các thuật toán học máy để xây dựng mô hình dự báo GDP. Từ đó, chúng ta sẽ xây dựng các mô hình dự báo sử dụng các kỹ thuật học máy như hồi quy logistic, rừng ngẫu nhiên, và máy vector hỗ trợ (SVM) và các bộ dữ liệu được sử dụng bao gồm các thông tin cơ bản như: tên quốc gia, năm, tỷ lệ lao động trong nông nghiệp, tỷ lệ lao động trong công nghiệp, tỷ lệ lao động trong dịch vụ, tỷ lệ thất nghiệp .

Dựa vào tích dữ liệu GDP từng năm của các quốc gia, các mô hình học máy có thể phát hiện các mối quan hệ tiềm ẩn giữa cơ cấu lao động và tốc độ tang trưởng GDP. Từ đó, nghiên cứu hướng tới việc:

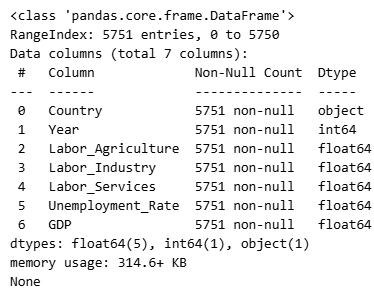
* + - Xác định mức độ ảnh hưởng của từng khu vực lao động đến tăng trưởng GDP.
    - Đánh giá vai trò của tỷ lệ thất nghiệp trong việc phản ánh tiềm năng phát triển của quốc gia.
    - Xây dựng so sách các mô hình học máy có độ chính xác cao trong dự đoán GDP hỗ trợ các chính sách và chiến lược phát triển.

Đề tài không chỉ thể hiện giá trị thực tiễn của học máy trong lĩnh vực kinh tế mà còn mang ý nghĩa khoa học sâu sắc trong việc sử dụng trí tuệ nhân tạo vào phân tích, dự báo và hỗ trợ ra quyết định. Kết quả nghiên cứu có thể góp phần nâng cao hiệu quả điều hành kinh tế, giúp các quốc gia ra chính sách, giải pháp hợp lý nhằm ổn định và phát triển kinh tế một cách bền vững.

* 1. **Trình bày dữ liệu bài toán** Dữ liệu được lấy từ trang web này: [Stroke Prediction Dataset | Kaggle](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)



Hình 1. 1 Hiển thị 5 hàng đầu và 5 hàng cuối của dữ liệu

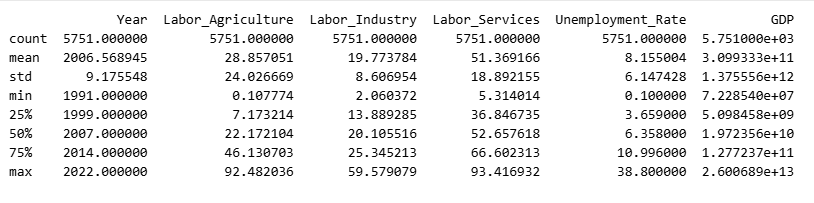
**

Hình 1. 2 Thông tin tổng quan về cấu trúc Dataframe

* Bộ dữ liệu được sử dụng trong bài toán này bao gồm các đặc trưng liên quan đến sức khỏe và nhân khẩu học của bệnh nhân, cụ thể như sau:
  1. Country Name: tên quốc gia .
  2. Year: năm quan sát.
  3. Employment Sector: Agriculture: Tỷ lệ lao động ngành Nông nghiệp.
  4. Employment Sector: Industry: Tỷ lệ lao động ngành Công nghiêp.
  5. Employment Sector: Services: Tỷ lệ lao động ngành dịch vụ
  6. Unemployment Rate: Tỷ lệ thất nghiệp
  7. GDP (in USD): Tổng sản phẩm quốc nội tính theo USD
* Dữ liệu bài toán là 1 file csv gồm 5751 rows × 7 columns

+ Có 7 feature và mỗi feature có 5751 dữ liệu đầu vào

-Sau khi mô tả dữ liệu ta có:



Hình 1. 3 Thông tin mô tả về dữ liệu số

+ Năm thống kê lớn nhất là **2022** và nhỏ nhất là **1991**.

+ Tỷ lệ lao động trong lĩnh vực **nông nghiệp (Agriculture)** lớn nhất là **92.48%** và nhỏ nhất là **0.11%**.

+ Tỷ lệ lao động trong lĩnh vực **công nghiệp (Industry)** lớn nhất là **59.58%** và nhỏ nhất là **2.06%**.

+ Tỷ lệ lao động trong lĩnh vực **dịch vụ (Services)** lớn nhất là **93.42%** và nhỏ nhất là **5.31%**.

+ **Tỷ lệ thất nghiệp (Unemployment Rate)** lớn nhất là **38.80%** và nhỏ nhất là **0.10%.**

+ **GDP (tính theo USD)** lớn nhất là **2.60×10¹³ USD** và nhỏ nhất là **7.23×10⁷ USD.**

## Tiền xử lý dữ liệu

1. Hiển thị 5 hàng đầu của dữ liệu
2. Hiển thị số hàng, cột của dữ liệu
3. Hiển thị thông tin của dữ liệu
4. Mô tả dữ liệu
5. Hiển thị giá trị duy nhất trong mỗi cột

## Làm sạch dữ liệu

1. Xử lý giá trị thiếu
2. Xoá outliner
3. Ma trận tương quan

## Trực quan hoá dữ liệu

1. Biểu đồ thể hiện phân phối đặc trưng các phân loại
2. Biểu đồ phân bố GDP
3. đồ ma trận tương quan
4. Biểu đồ phân tán GDP theo tỷ lệ lao động nông nghiệp
5. Biểu đồ phân tán GDP theo tỷ lệ lao động công nghiệp
6. **Biểu đồ phân tán GDP theo tỷ lệ lao động dịch vụ**
7. Biểu đồ hộp về tỷ lệ thất nghiệp

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Random Forest Regressor

## Giới thiệu thuật toán Random Forest Regressor

Random Forest Regressor là một phương pháp học máy có giám sát (Supervised Learning), thuộc nhóm Ensemble Learning.  
Thuật toán này hoạt động bằng cách xây dựng nhiều cây quyết định (Decision Trees) độc lập và kết hợp kết quả dự đoán của chúng (lấy trung bình trong bài toán hồi quy).  
Mục tiêu của Random Forest là giảm phương sai (variance) của mô hình và tăng khả năng tổng quát hóa, nhờ việc kết hợp nhiều mô hình con yếu thành một mô hình mạnh.

## Cấu trúc của mô hình RandomForest

Mô hình Random Forest bao gồm:

* + - * **Cây quyết định (Decision Trees):** Các mô hình con học các mối quan hệ phi tuyến giữa đặc trưng và biến mục tiêu.
      * **Tập dữ liệu mẫu (Bootstrap Samples):** Mỗi cây được huấn luyện trên một mẫu ngẫu nhiên có thay thế từ tập dữ liệu gốc.
      * **Tập con đặc trưng (Feature Subset):** Ở mỗi nút, chỉ một số đặc trưng được chọn ngẫu nhiên để tìm điểm chia tối ưu.

## Quá trình xây dựng mô hình RandomForest

* + - * **Khởi tạo các cây :** Xác định số lượng cây (tham số n\_estimators)
      * **Lấy mẫu dữ liệu:** Mỗi cây được huấn luyện trên tập con dữ liệu riêng.
      * **Chọn tập con đặc trưng:** Mỗi nút chia dữ liệu dựa trên tập con đặc trưng được chọn ngẫu nhiên.
      * **Xây dựng cây:** Từng cây được huấn luyện độc lập, sử dụng chỉ tiêu như “Mean Squared Error” để tối ưu việc chia.
      * **Dự đoán:** Kết quả cuối cùng là **trung bình dự đoán của toàn bộ cây** trong rừng.

## Ưu điểm của RandomForest

* + - * **Khả năng tổng quát hóa tốt:** Kết hợp nhiều cây giúp mô hình ổn định hơn và tổng quát hóa tốt.
      * **Hiệu suất cao:** Làm việc hiệu quả với dữ liệu lớn, phi tuyến.
      * **Đánh giá quan trọng đặc trưng:** Cho biết yếu tố nào (ví dụ: tỷ lệ lao động nông nghiệp, công nghiệp, dịch vụ, thất nghiệp) ảnh hưởng mạnh nhất đến GDP.
      * **Độ chính xác cao:** Thường cho kết quả chính xác hơn mô hình hồi quy tuyến tính đơn giản.
      * **Xử lý dữ liệu mất mát:** Có khả năng ước lượng giá trị bị thiếu trong dữ liệu.

## Nhược điểm của RandomForest

* + - * **Khó giải thích:** Mặc dù mỗi cây dễ hiểu, nhưng tổng thể rừng lại phức tạp.
      * **Chi phí tính toán cao:** Huấn luyện nhiều cây đòi hỏi tài nguyên lớn.
      * **Dễ bị bias:** Nếu dữ liệu không được chuẩn hóa hoặc phân bố không đồng đều.

## Đánh giá hiệu suất

Hiệu suất của RandomForestClassifier thường được đánh giá thông qua các chỉ số như Accuracy, MAE (Mean Absolute Error, RMSE (Root Mean Squared Error), và **R² (Coefficient of Determination)**. Đối với bài toán dự đoán GDP , mô hình được kiểm tra dựa trên khả năng dự đoán chính xác GDP thực tế của các quốc gia trong giai đoạn quan sát.

## Linear Regressor

## Giới thiệu thuật toán Linear Regressor

Linear Regression (Hồi quy tuyến tính) là thuật toán **cơ bản nhất trong học máy** và **thống kê kinh tế**, được sử dụng để **mô hình hóa mối quan hệ tuyến tính** giữa biến phụ thuộc (trong bài này là **GDP**) và một hoặc nhiều biến độc lập (các yếu tố như cơ cấu lao động và tỷ lệ thất nghiệp).

Ý tưởng chính của hồi quy tuyến tính là **tìm một đường thẳng hoặc mặt phẳng tốt nhất** (đường hồi quy) sao cho **khoảng cách giữa các điểm dữ liệu thực tế và đường hồi quy là nhỏ nhất** (theo tiêu chí bình phương sai số tối thiểu).

Công thức tổng quát của mô hình hồi quy tuyến tính đa biến:

GDP = b0 + b1X1 + b2X2 + ... + bn\*Xn + e

Trong đó:

* GDP: giá trị cần dự đoán (biến phụ thuộc).
* Xi: các biến đầu vào (tỷ lệ lao động nông nghiệp, công nghiệp, dịch vụ, thất nghiệp, năm).
* βi​: hệ số hồi quy thể hiện mức độ ảnh hưởng của từng biến XiX\_iXi​ đến GDP.
* ε: sai số (noise) giữa giá trị thực và giá trị dự đoán.

## Cấu trúc của mô hình Extra Trees Regressor

Mô hình Linear Regression bao gồm:

* Tập dữ liệu huấn luyện: Gồm các quốc gia, năm, tỷ lệ lao động trong từng lĩnh vực và tỷ lệ thất nghiệp tương ứng với GDP thực tế.
* Hệ số hồi quy (βi): Được học thông qua việc tối thiểu hóa sai số bình phương (Least Squares Method).
* Giá trị dự đoán (y^): Tính bằng công thức tuyến tính giữa các đặc trưng đầu vào và hệ số đã học.

Mục tiêu của quá trình huấn luyện là tìm các hệ số β tối ưu sao cho sai số giữa GDP thực tế và GDP dự đoán là nhỏ nhất.

## Quá trình xây dựng mô hình Linear Regression

* **Thu thập và tiền xử lý dữ liệu:**
  + Làm sạch, loại bỏ giá trị thiếu và ngoại lai.
  + Chuẩn hóa dữ liệu bằng StandardScaler để đảm bảo các đặc trưng có cùng thang đo.
* **Chia dữ liệu thành tập huấn luyện và kiểm thử (Train/Test Split):**
  + Dữ liệu được chia với tỉ lệ 80/20 để mô hình có thể học và đánh giá công bằng.
* **Huấn luyện mô hình:**
  + Sử dụng phương pháp Ordinary Least Squares (OLS) để ước lượng các hệ số β.
  + Mô hình học được mối quan hệ tuyến tính giữa GDP và các yếu tố lao động, thất nghiệp, năm.
* **Dự đoán:**
  + Với các giá trị đầu vào mới (năm, cơ cấu lao động, thất nghiệp), mô hình dự đoán GDP tương ứng

## Ưu điểm của Linear Regression

* **Đơn giản và dễ hiểu:** Là mô hình nền tảng, dễ triển khai và diễn giải.
* **Tốc độ huấn luyện nhanh:** Do chỉ cần tính toán ma trận hệ số.
* **Khả năng giải thích rõ ràng:** Mỗi hệ số β cho thấy mức độ ảnh hưởng của biến đầu vào đến GDP.
* **Hiệu quả với dữ liệu tuyến tính:** Khi GDP có quan hệ gần tuyến tính với các yếu tố lao động và thất nghiệp.

## Nhược điểm của KNN

* + - * **Giả định tuyến tính:** Không phù hợp khi quan hệ giữa GDP và các yếu tố đầu vào là phi tuyến.
      * **Nhạy cảm với ngoại lai (Outliers):** Các giá trị bất thường có thể làm sai lệch mô hình.
      * **Không mô hình hóa tương tác phức tạp:** Không thể hiện được các tác động kết hợp giữa các yếu tố (ví dụ: mối quan hệ giữa tỷ lệ thất nghiệp và lao động dịch vụ).
      * **Dễ bị đa cộng tuyến (Multicollinearity):** Khi các biến đầu vào có quan hệ chặt chẽ, mô hình trở nên không ổn định.

## Đánh giá hiệu suất

Hiệu suất của Linear Regression được đánh giá bằng các chỉ số:

* **MAE (Mean Absolute Error):** Trung bình sai số tuyệt đối giữa GDP dự đoán và GDP thực tế.
* **RMSE (Root Mean Squared Error):** Sai số bình phương trung bình, nhấn mạnh các sai lệch lớn.
* **R² (Hệ số xác định):** Đo lường tỷ lệ biến thiên của GDP được giải thích bởi các biến đầu vào.

Trong bài toán dự đoán GDP, mô hình Linear Regression đóng vai trò mô hình cơ sở (baseline) để so sánh hiệu suất với các mô hình phi tuyến mạnh hơn như Random Forest, Extra Trees và Gradient Boosting.

## Support Vector Regression (SVR)

## Giới thiệu thuật toán SVR

**Support Vector Regression (SVR)** là biến thể hồi quy của **Support Vector Machine (SVM),** được thiết kế để giải quyết bài toán **dự đoán biến liên tục.**  
SVR tìm một **hàm hồi quy tối ưu (optimal hyperplane)** sao cho sai số giữa giá trị dự đoán và giá trị thực **nằm trong một giới hạn cho phép (epsilon)**, đồng thời **biên (margin)** giữa các điểm dữ liệu được tối đa hóa.

## Cấu trúc của mô hình SVR

* + - * **Siêu phẳng (Hyperplane):** Là đường hoặc mặt phẳng biểu diễn mối quan hệ giữa đặc trưng và GDP.
      * **Margin:** Khoảng cách cho phép sai số không bị phạt.
      * **Hạt nhân (Kernel):** Ánh xạ dữ liệu sang không gian cao hơn để mô hình hóa mối quan hệ phi tuyến (ví dụ: RBF, Polynomial, Sigmoid).

## Quá trình xây dựng mô hình SVR

* + - * **Lựa Chọn kernel:** Thường sử dụng **RBF kernel** vì phù hợp với dữ liệu kinh tế phi tuyến.
      * **Tìm siêu phẳng tối ưu:** Sử dụng tham số C để điều chỉnh mức phạt và ε để xác định ngưỡng sai số cho phép.
      * **Đánh giá:** Mô hình được huấn luyện để giảm sai số và dự đoán GDP của quốc gia mới.

## Ưu điểm của SVR

* + - * **Hiệu quả cao:** Là một trong những thuật toán tốt nhất cho các bài toán phân loại nhị phân.
      * **Giảm overfitting:** Nhờ cơ chế margin tối ưu..
      * **Khả năng tổng quát tốt:** Làm việc tốt khi dữ liệu có phân bố phức tạp.
      * **Hiệu quả với dữ liệu nhỏ:** Hoạt động tốt ngay cả khi kích thước dữ liệu nhỏ.

## Nhược điểm của SVR

* + - * **Thời gian huấn luyện lâu:** Khi dữ liệu lớn hoặc nhiều đặc trưng.
      * **Cần tinh chỉnh tham số:** Việc chọn C, ε, và kernel ảnh hưởng lớn đến hiệu suất.
      * **Khó diễn giải:** Mô hình hoạt động như “hộp đen”, khó hiểu rõ quan hệ bên trong.

## Đánh giá hiệu suất

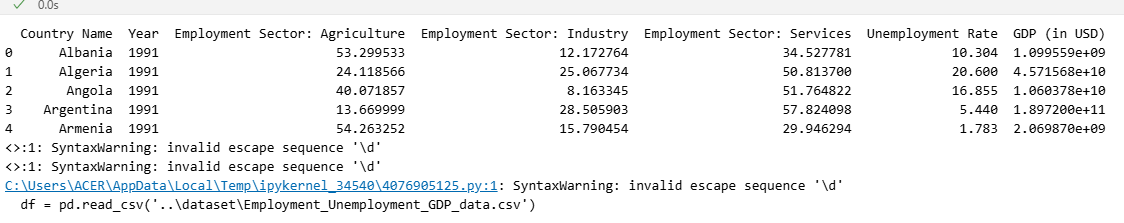
Hiệu suất của mô hình SVR được đánh giá bằng các chỉ số **MAE**, **RMSE**, và **R²**.  
Trong bài toán dự đoán GDP, SVR có khả năng **mô hình hóa mối quan hệ phi tuyến** giữa GDP và các yếu tố như tỷ lệ lao động, thất nghiệp, và năm, cho kết quả ổn định ngay cả khi dữ liệu huấn luyện không lớn.

# CHƯƠNG 3: GIẢI PHÁP, CÀI ĐẶT THỰC HIỆN

## Mã nguồn tiền xử lý dữ liệu

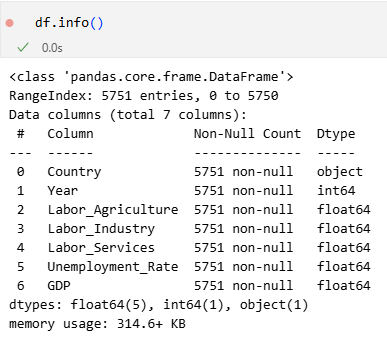
****

Hình 3. 1 Import thư viện cần thiết, load, hiển thị 5 hàng đầu của DataFrame

**

Hình 3. 2 Kết quả hiển thị 5 hàng đầu của DataFrame

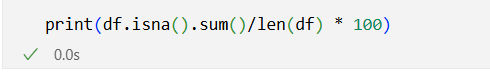
Đầu tiên em sử dụng hàm df.info() để kiểm tra các feature nào bị thiếu và có kiểu dữ liệu thế nào.Em hiển thị lại hàm df và dùng lại hàm info để kiểm tra xem dữ liệu. Và hình ảnh dưới là kết quả sau khi đã điền các dữ liệu:



Hình 3. 4 Thông tin về DataFrame

## Mã nguồn chức năng làm sạch dữ liệu

## Xử lý giá trị thiếu

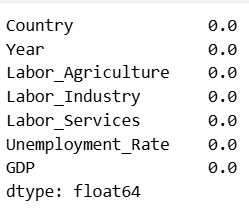
****

Hình 3. 5 Mã nguồn xử lý giá trị thiếu

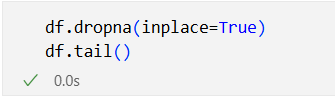
Câu lệnh này tính toán tỷ lệ phần trăm giá trị thiếu trong từng cột của DataFrame data.

* + - * **df.isna()**: Xác định các giá trị thiếu (NaN) trong data và trả về DataFrame với giá trị True/False.
      * **.sum()**: Tính tổng số True (giá trị thiếu) cho mỗi cột.
      * **len(data)**: Số lượng hàng trong data.
      * **\*100**: Chuyển đổi tỷ lệ sang phần trăm.

Kết quả: Hiển thị phần trăm giá trị thiếu của từng cột, hỗ trợ xác định cần xử lý những cột nào.



Hình 3. 6 Hiển thị phần trăm giá trị thiếu của từng cột

**

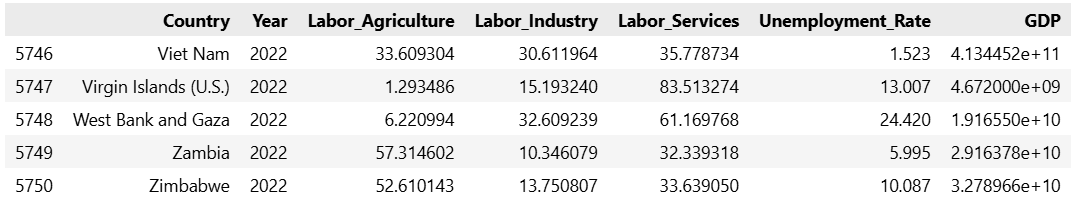
Hình 3. 7 Loại bỏ tất cả các hàng trong DataFrame data có ít nhất một giá trị bị thiếu (NaN) và hiển thị 5 hàng cuối của DataFrame

Câu lệnh này được sử dụng để xử lý giá trị thiếu và kiểm tra dữ liệu sau khi loại bỏ các hàng chứa giá trị thiếu:

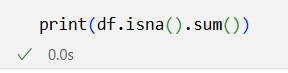
**df.dropna(inplace=True)**: Loại bỏ tất cả các hàng trong DataFrame data có ít nhất một giá trị bị thiếu (NaN). Tham số inplace=True đảm bảo thay đổi được thực hiện trực tiếp trên DataFrame data mà không tạo một bản sao.

**df.tail()**: Hiển thị 5 hàng cuối cùng của DataFrame data sau khi loại bỏ các giá trị thiếu.

Mục đích: Loại bỏ dữ liệu không đầy đủ và xem nhanh các dòng cuối để đảm bảo dữ liệu đã được xử lý đúng cách.



Hình 3. 8 Kết quả hiển thị 5 hàng cuối của DataFrame



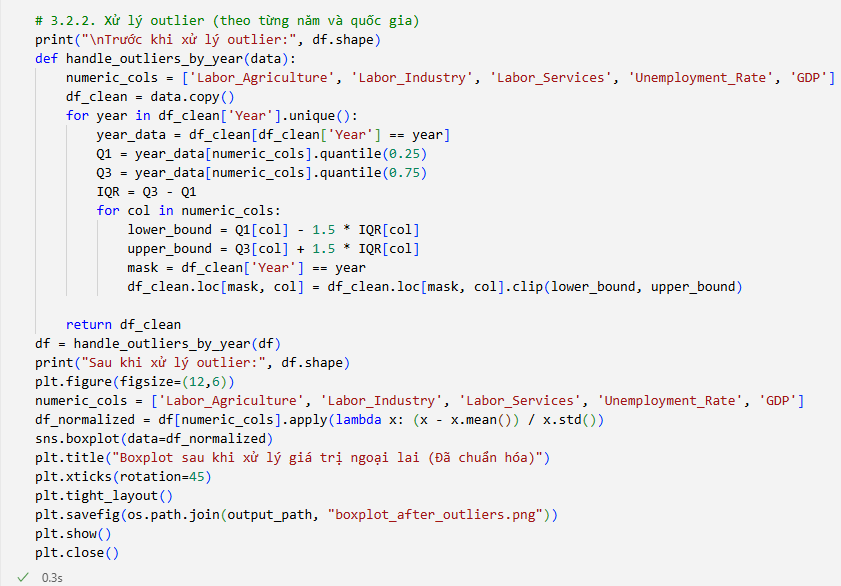
Hình 3. 9 Mã nguồn đếm số lượng giá trị thiếu (NaN) trong từng cột của DataFrame data.

* + - * Đếm số lượng giá trị thiếu (NaN) trong từng cột của DataFrame data.
      * Kết quả: Một bảng liệt kê số giá trị thiếu theo cột.



Hình 3. 10 Kết quả liệt kê số giá trị thiếu theo cột.

## Xoá outlier

****

Hình 3. 11 Mã nguồn xử lý dữ liệu ngoại lai

Hàm **handle\_outliers** loại bỏ giá trị ngoại lai trong các cột bằng cách sử dụng **IQR (Interquartile Range)**:

## Tính IQR:

* + Q1 (25%) và Q3 (75%) là các phân vị của dữ liệu.
  + IQR = Q3 - Q1, thể hiện khoảng giữa 50% dữ liệu trung tâm.

## Giới hạn:

* + **Lower bound** = Q1 - 1.5 \* IQR.
  + **Upper bound** = Q3 + 1.5 \* IQR.

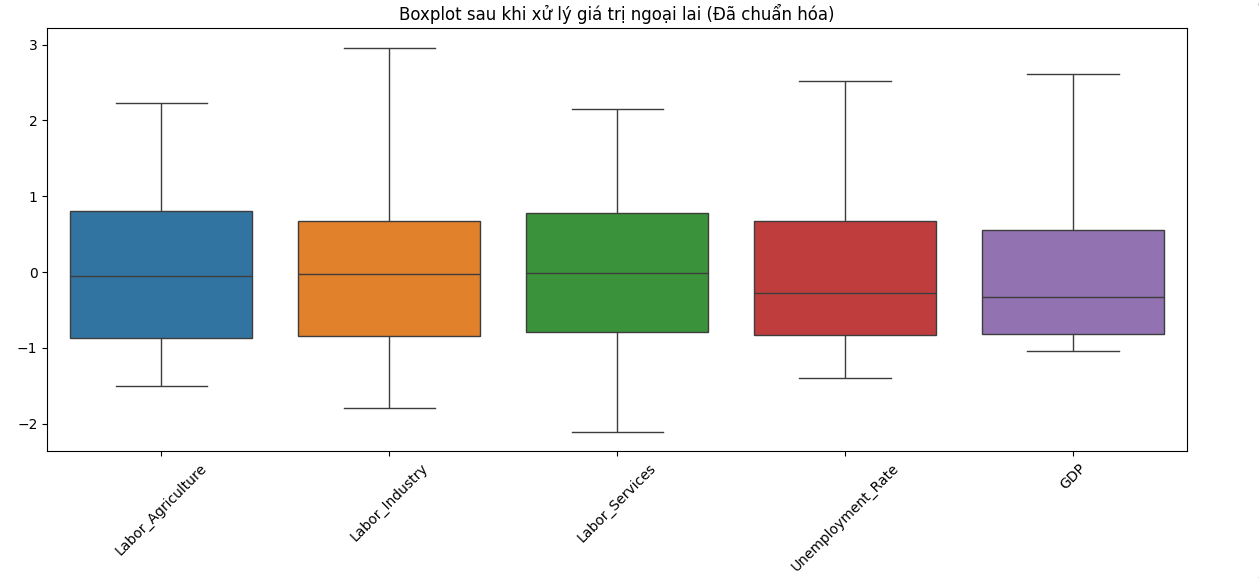
## Lọc dữ liệu:

* + Loại bỏ các dòng có giá trị nằm ngoài khoảng [Lower bound, Upper bound] ở bất kỳ cột số nào
  + Sau khi lọc, dữ liệu chỉ còn lại các giá trị nằm trong phạm vi hợp lý, đảm bảo tính ổn định cho quá trình huấn luyện.

## boxplots(data\_no\_outliers):

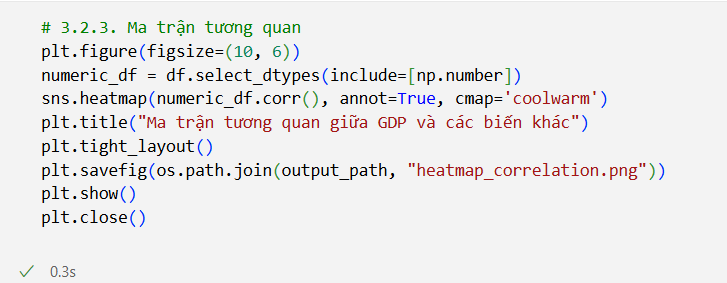
* + Hàm vẽ biểu đồ **Boxplot** thể hiện dữ liệu sau khi loại bỏ ngoại lai.
  + Biểu đồ được **lưu tại thư mục** output/figure với tên *boxplot\_no\_outliers.png*, giúp trực quan hóa kết quả làm sạch dữ liệu.

Kết quả: DataFrame data\_no\_outliers không còn ngoại lai, và boxplot minh họa dữ liệu sạch.



Hình 3. 12 Boxplot minh họa dữ liệu sạch.

## Ma trận tương quan

****

Hình 3. 13 Mã nguồn ma trận tương quan

**Biểu đồ Heatmap** để thể hiện mối quan hệ tương quan giữa các cột số học trong dữ liệu đã xử lý ngoại lai:

## numeric\_df:

* + Lọc ra các cột có kiểu dữ liệu số trong DataFrame (df.select\_dtypes(include= [np.number])).
  + Bao gồm các cột như: ['Year', 'Labor\_Agriculture', 'Labor\_Industry', 'Labor\_Services', 'Unemployment\_Rate', 'GDP'].

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

* + Thiết lập kích thước biểu đồ là 10x6 inch.

## sns.heatmap():

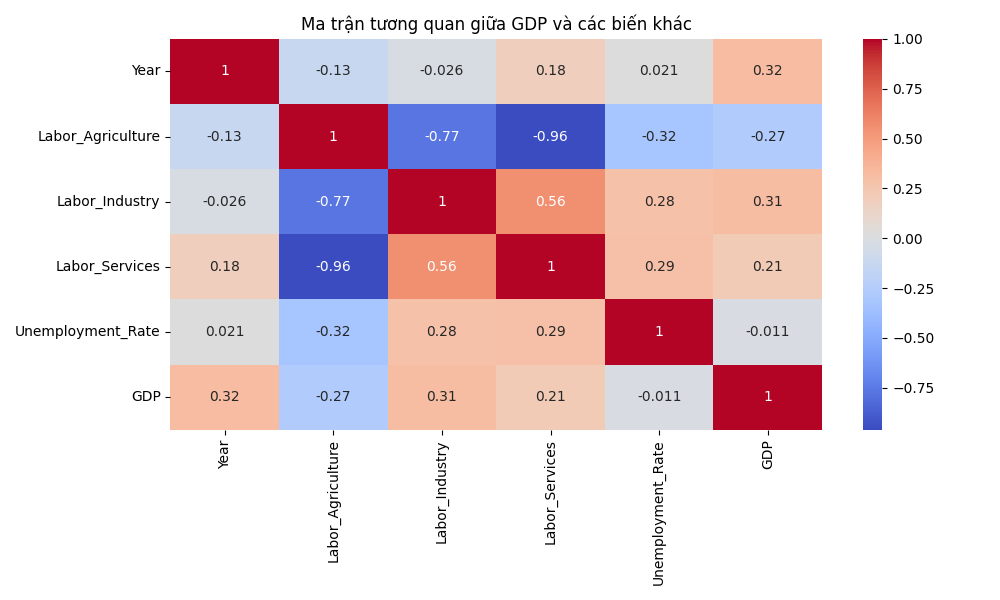
* + Tính toán ma trận tương quan giữa các biến số trong dữ liệu.
  + annot=True: Hiển thị giá trị tương quan trên heatmap.
  + Cmap = ‘coolwarm’:Áp dụng bảng màu giúp dễ nhận biết giá trị tương quan dương(đỏ) và âm (xanh).

## plt.title():

* + Đặt tiêu đề: " Ma trận tương quan giữa GDP và các biến khác”".

1. **plt.savefig():** Lư biểu đồ vào thư mục output/figure với tên heatmap\_correlation.png
2. **plt.show()**: Hiển thị biểu đồ.

**Kết quả**: Heatmap minh họa mức độ tương quan giữa các yếu tố, các vùng màu thể hiện mối quan hệ tuyến tính giữa GDP và các yếu tô kinh tế.



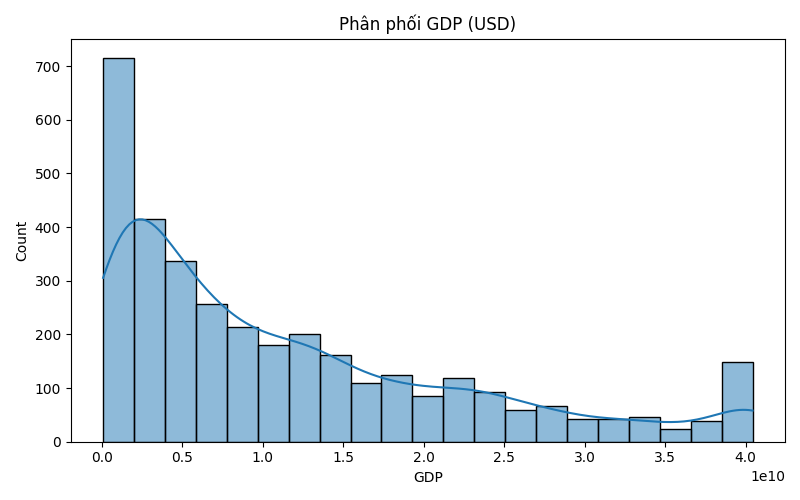
Hình 3. 14 Kết quả Heatmap minh họa mức độ tương quan giữa các yếu tố

## Mã nguồn chức năng Trực quan hoá dữ liệu

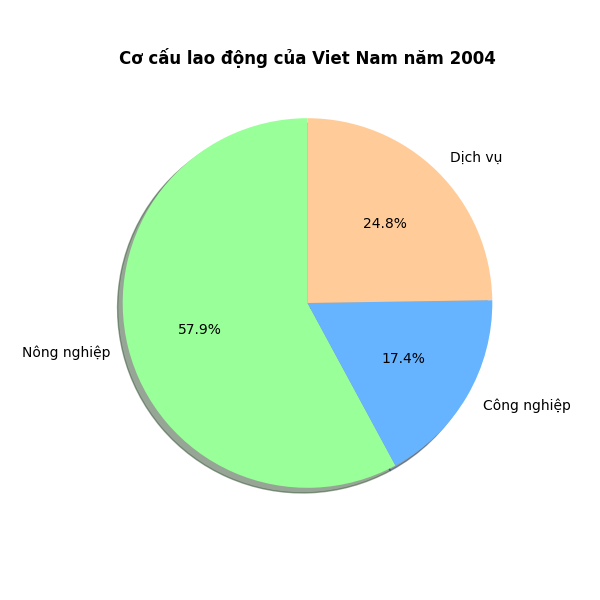
## Biểu đồ thể hiện phân phối đặc trưng các phân loại

****

Hình 3. 15 Mã nguồn biểu đồ thể hiện phân phối đặc trưng các phân loại

**

Hình 3. 16 Kết quả biểu đồ phân phôi GDP

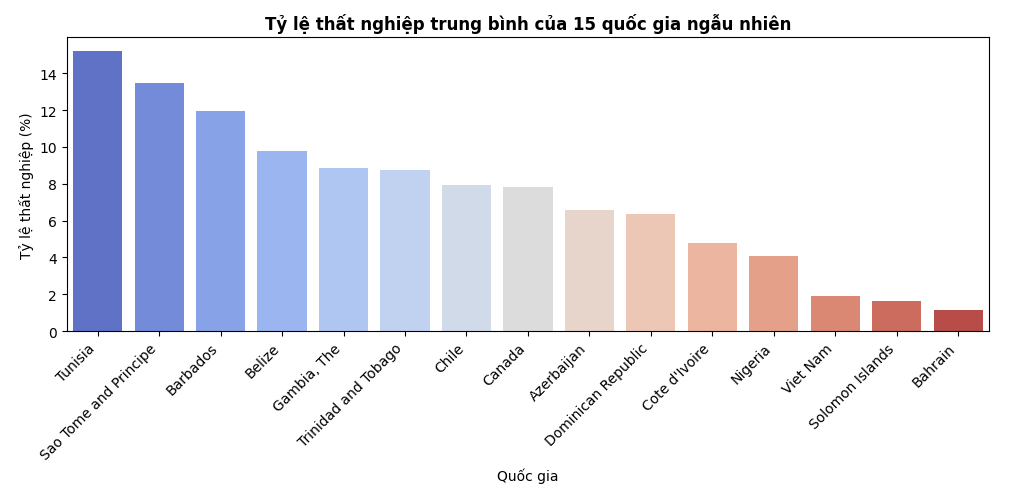


Hình 3. 17 Kết quả biểu đồ pie thể hiện phân phối đặc trưng các phân loại lao động của Việt Nam 2004

## Biểu đồ thể hiện tỷ lệ thất nghiệp trung bình của 15 quốc gia ngẫu nhiên

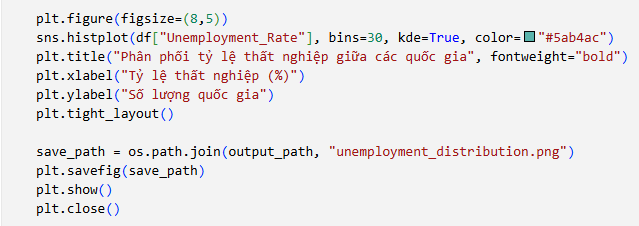
****

Hình 3. 18 Mã nguồn biểu đồ thể hiện tỷ lệ thất nghiệp trung bình của 15 quốc gia ngâu nhiên

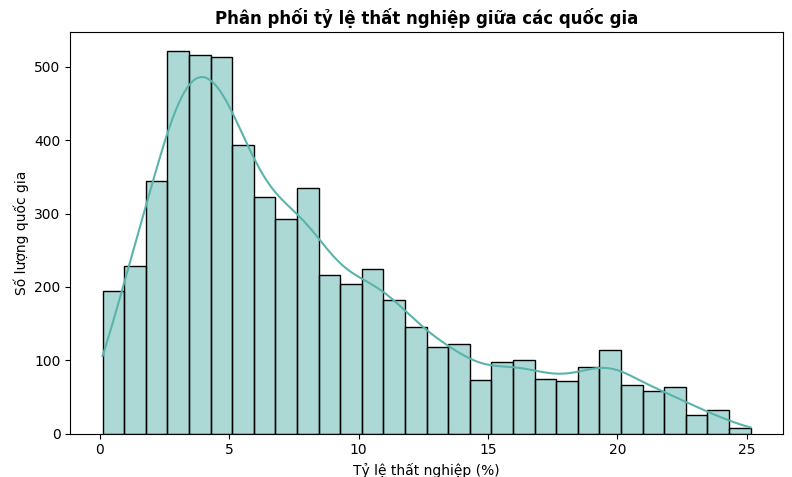


Hình 3. 19 Kết quả biểu đồ thể hiện tỷ lệ thất nghiệp

## Biểu đồ thể hiện các đặc trưng liên tục

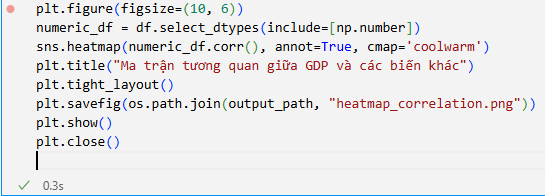
****

Hình 3. 20 Mã nguồn biểu đồ thể hiện các đặc trưng liên tục

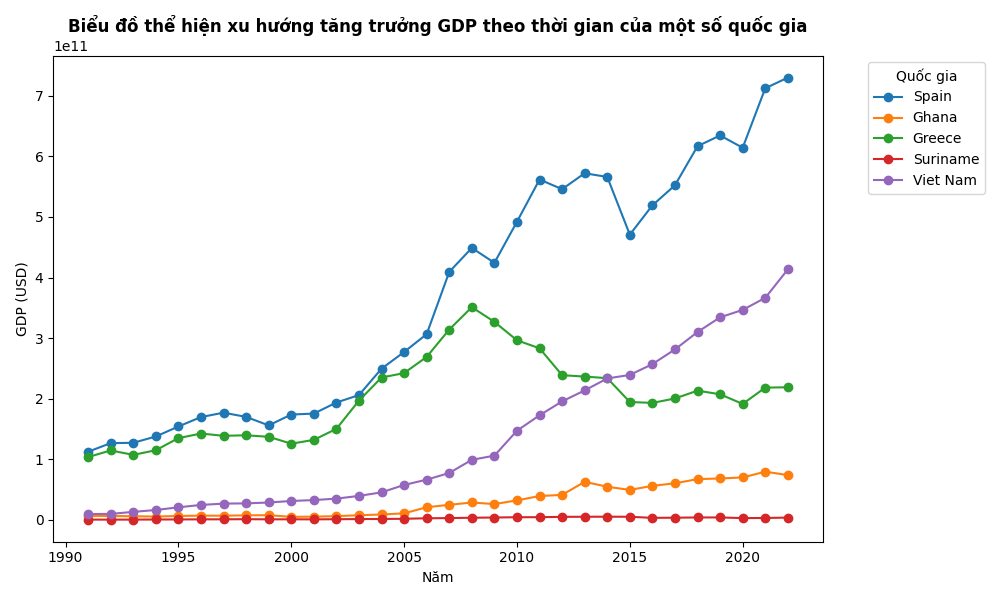
**

Hình 3. 21 Kết quả biểu đồ thể hiện các đặc trưng liên tục

## Biểu đồ thể hiện xu hướng tăng GDP theo thời gian của một số quốc gia

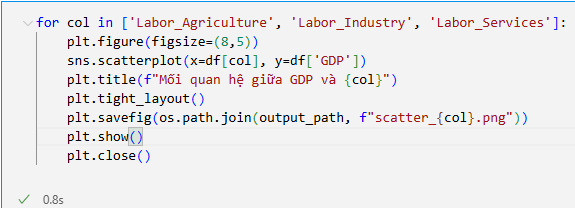
****

Hình 3. 22 Mã nguồn biểu đồ thể hiện xu hương tăng GDP theo thời gian của một số quốc gia

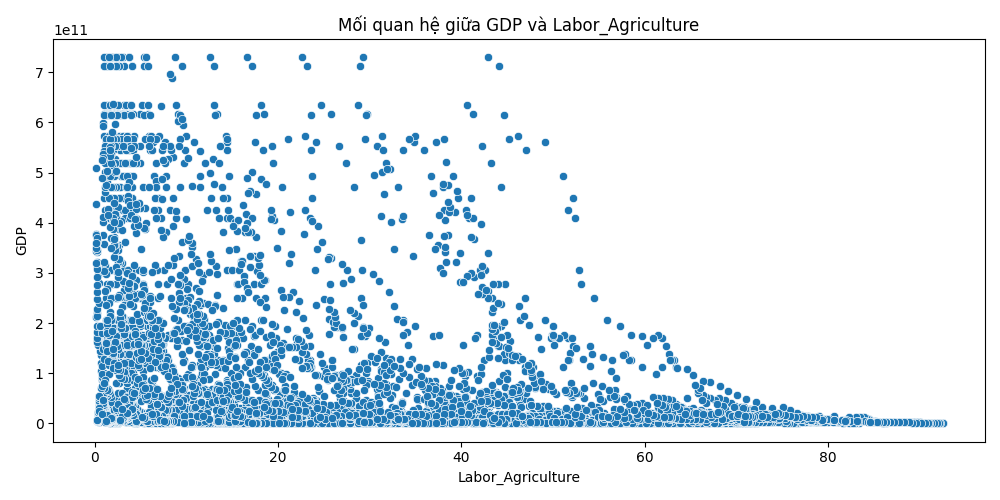
**

Hình 3. 23 Kết quả biểu đồ thể hiện xu hương tăng GDP theo thời gian của một số quốc gia

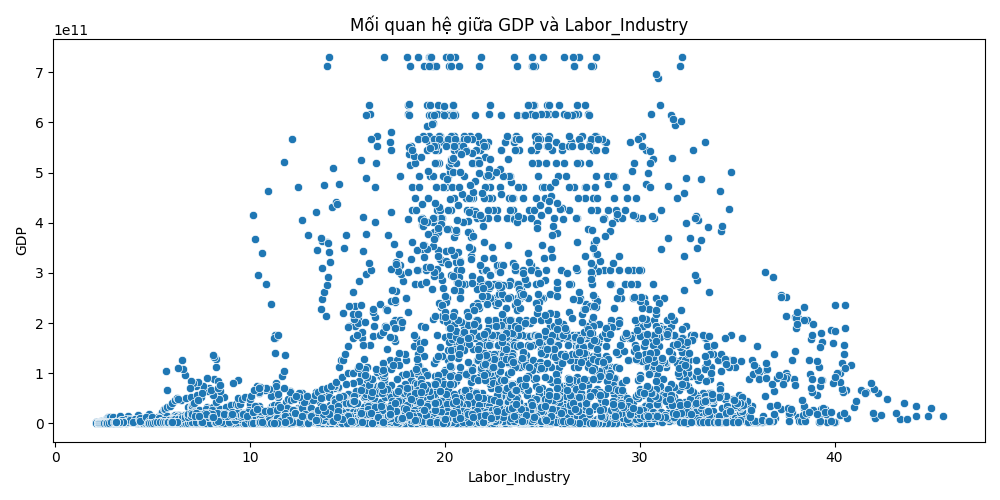
## Biểu đồ thể hiện Biểu đồ thể hiện mối quan hệ giữa GDP và các nhóm lao động

****

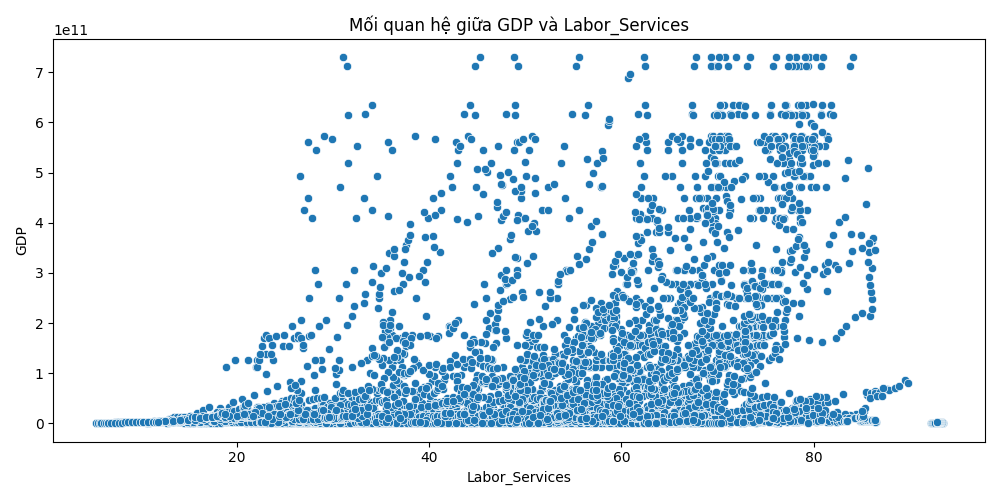
Hình 3. 24 Mã nguồn biểu đồ thể hiện mối tương quan giữa GDP và các phần tử còn lại

**

Hình 3. 25 Kết quả biểu đồ thể hiện mối tương quan giữa GDP và Nông nghiệp

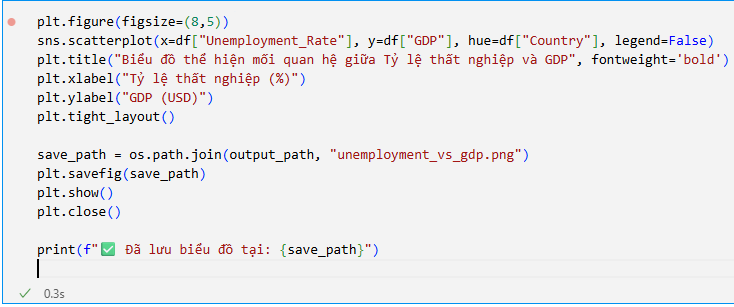
**

Hình 3. 26 Kết quả biểu đồ thể hiện mối tương quan giữa GDP và Công nghiệp

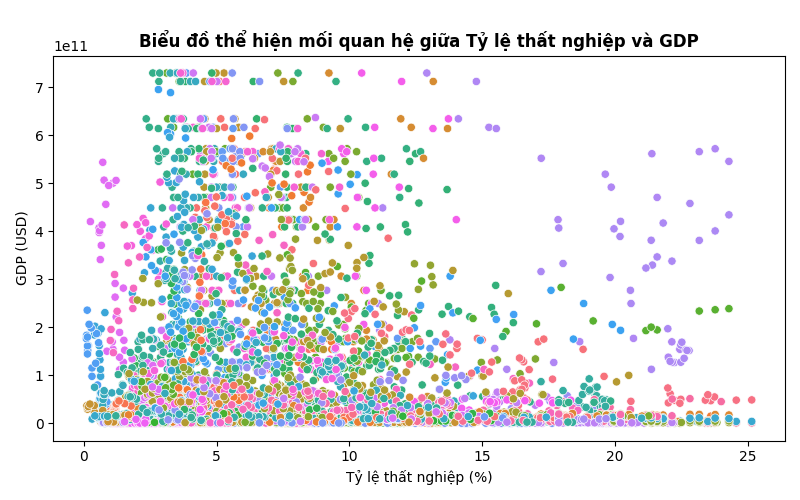
**

Hình 3. 27 Kết quả biểu đồ thể hiện mối tương quan giữa GDP và Dịch vụ

## Biểu đồ so sánh GDP thực tế và GDP dự đoán từ mô hình Extra Trees Regressor

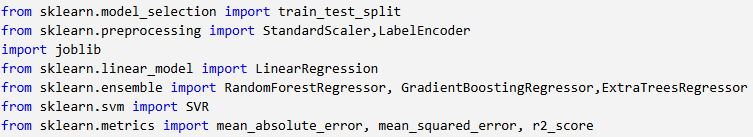
****

Hình 3. 28 Mã nguồn biểu đồ thể hiện sự trực quan hoá giá trị thiếu

**

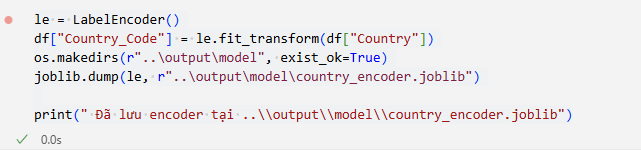
Hình 3. 29 Kết quả biểu đồ thể hiện sự trực quan hoá giá trị thiếu

## Chuyển đổi dữ liệu

****

Hình 3. 30 Import 1 số thư viện cần thiết

## Xử lý dữ liệu chuỗi

****

Hình 3. 31 Mã nguồn xử lý dữ liệu chuỗi

**

Hình 3. 32 Kết quả xử lý dữ liệu chuỗi

## Xử lý dữ liệu số

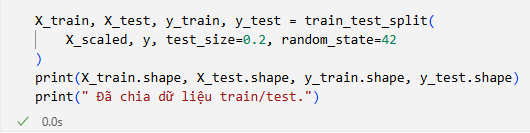


Hình 3. 33 Mã nguồn xử lý dữ liệu số

**

Hình 3. 34 Kết quả xử lý dữ liệu số

## Chia dữ liệu huấn luyện và kiểm thử

**

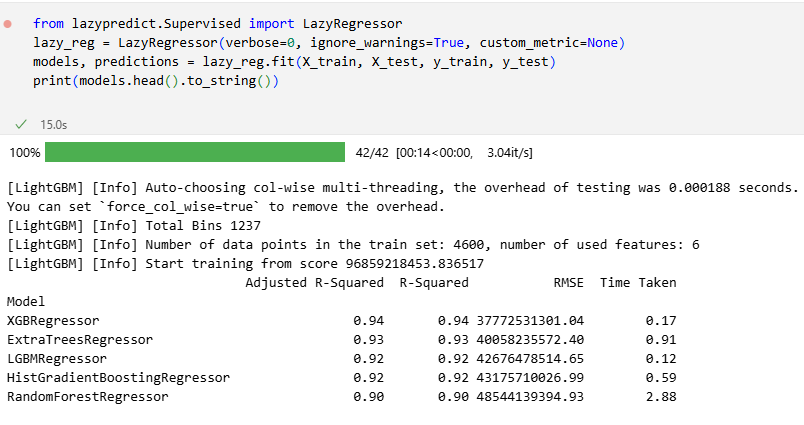
Hình 3. 35 Mã nguồn chia dữ liệu huấn luyện và kiểm thử

**

Hình 3. 36 Mã nguồn chia dữ liệu huấn luyện và kiểm thử

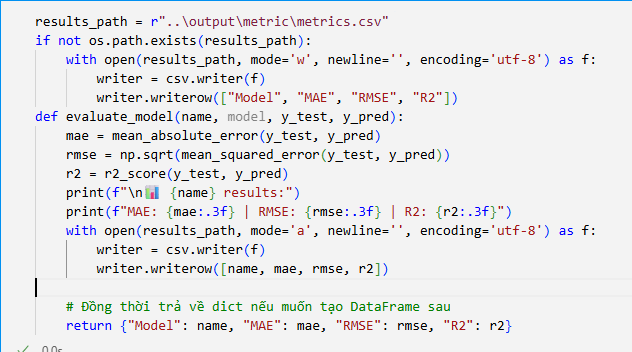
## Huấn luyện mô hình SKLEARN

## Tìm kiếm mô hình

****

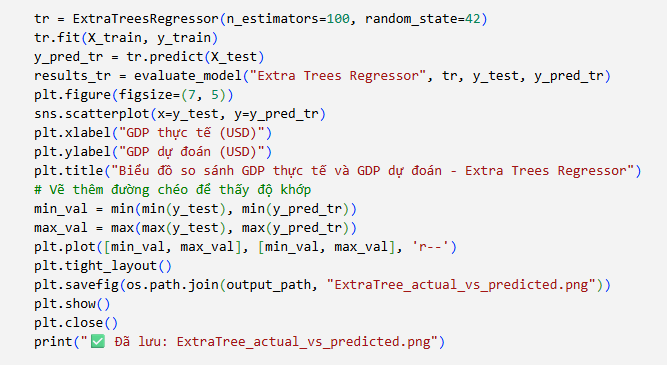
Hình 3. 37 Mã nguồn tìm kiếm mô hình dùng Lazypredict

## Đánh giá hiệu suất mô hình

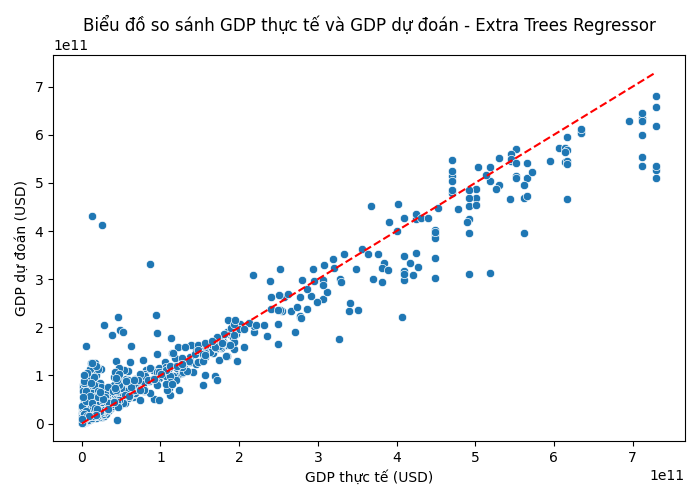


Hình 3. 38 Mã nguồn đánh giá hiệu suất mô hình

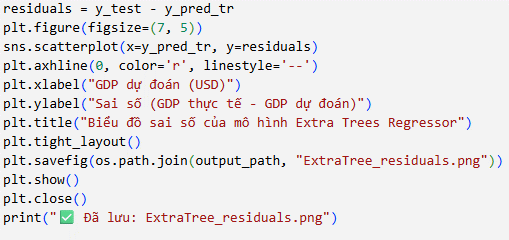
## Huấn luyện mô hình Extra Trees Regressor



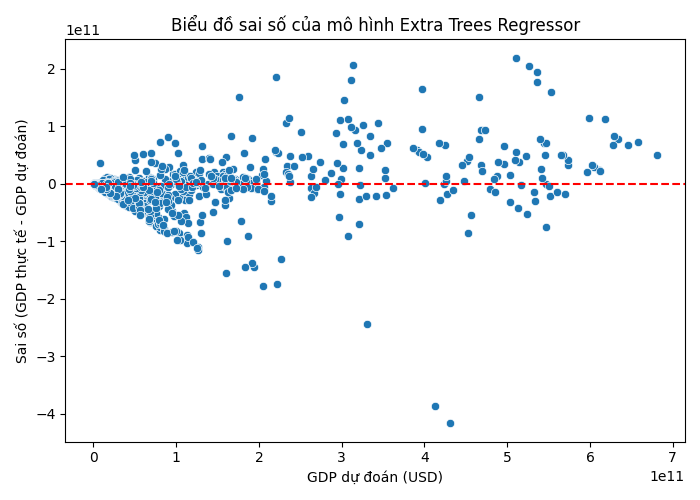
Hình 3. 37 Mã nguồn so sánh GDP thực tế và GDP dự đoán



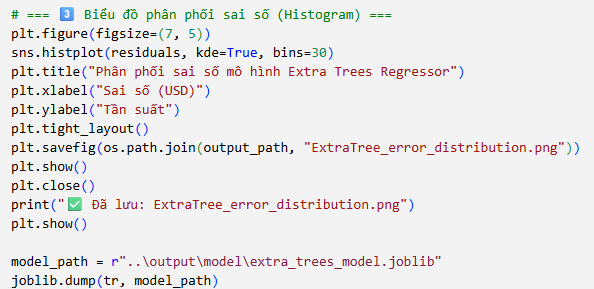
Hình 3. 38 Kết quả so sánh GDP thực tế và GDP dự đoán



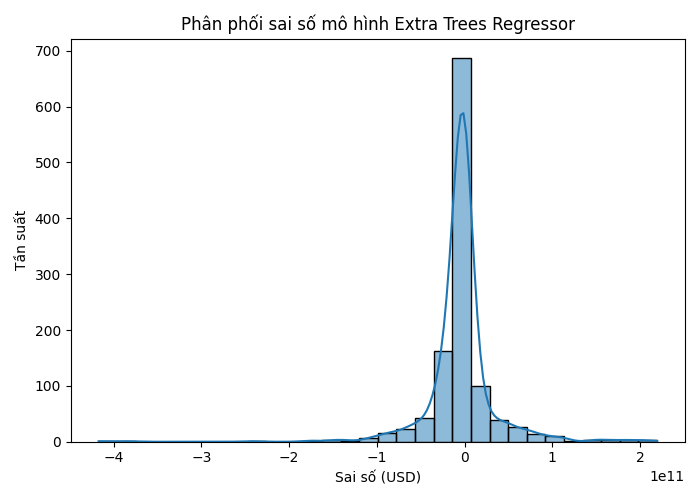
Hình 3. 40 Mã nguồn sai số GĐP



Hình 3. 41 Kết quả sai số GĐP

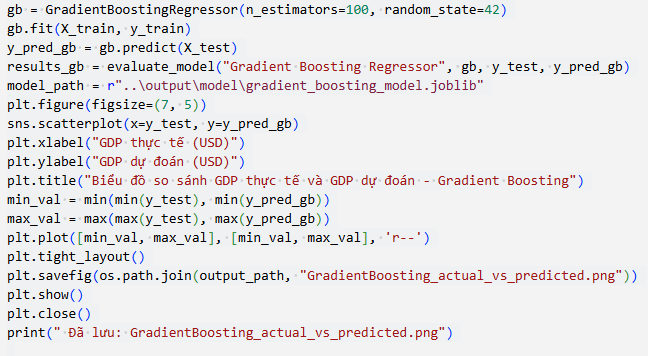


Hình 3. 42 Mã nguồn tần số phân phối sai số GĐP

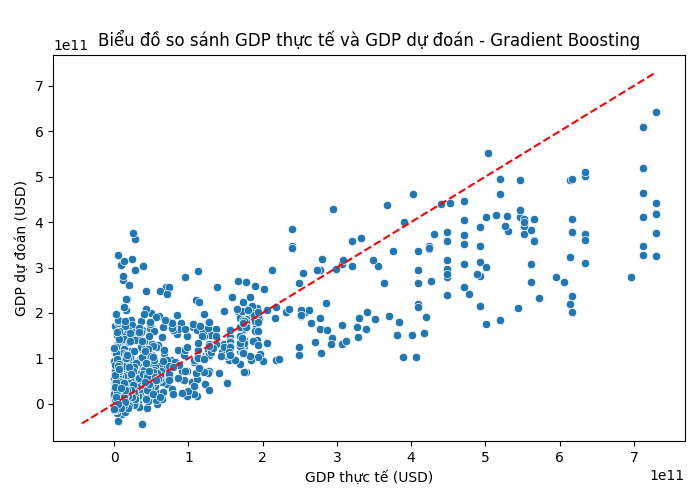


Hình 3. 43 Kết quả tần số phân phối sai số GĐP

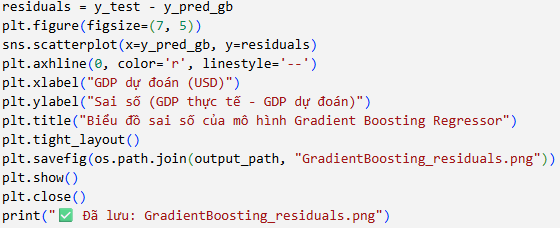
## Huấn luyện mô hình Gradient Boosting Regressor



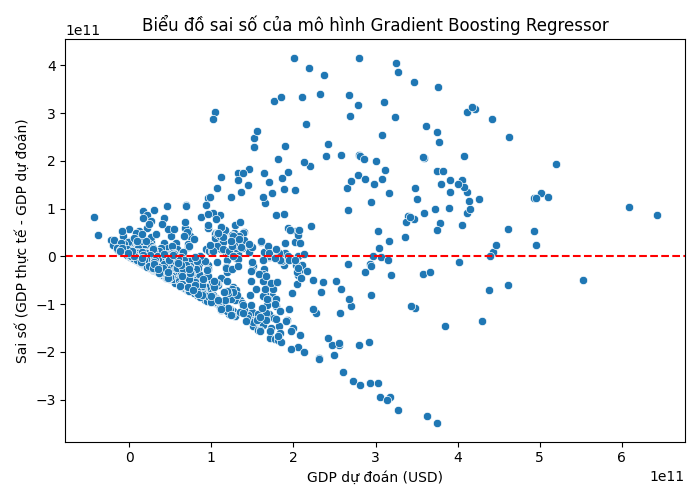
Hình 3. 44 Mã nguồn biểu đồ so sánh GDP thực tế và GDP dự đoán



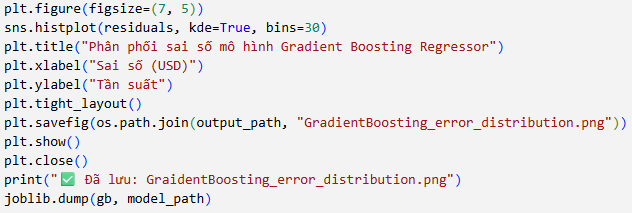
Hình 3. 45 Kết quả biểu đồ so sánh GDP thực tế và GDP dự đoán



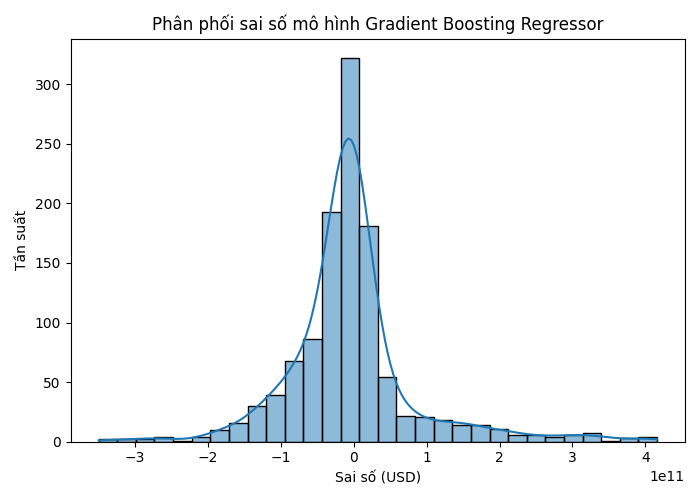
Hình 3. 46 Mã nguồn biểu đồ sai số GDP



Hình 3. 47 Kết quả biểu đồ sai số GDP

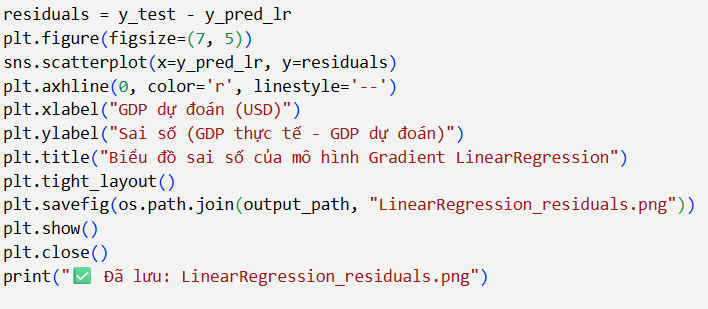


Hình 3. 48 Mã nguồn biểu đồ phân phối sai số GDP

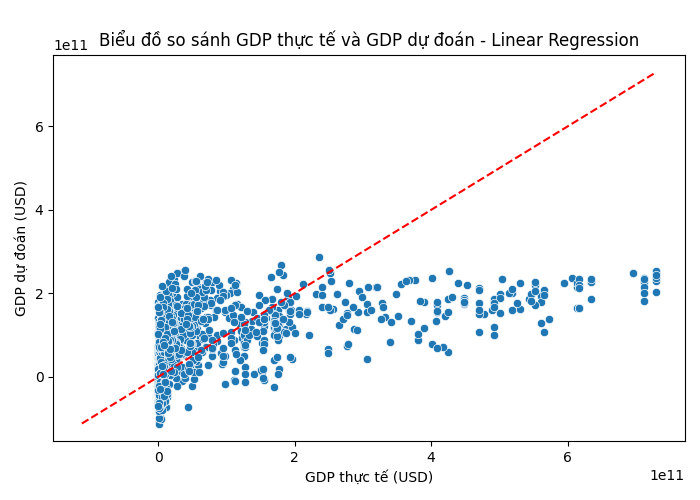


Hình 3. 49 Kết quả biểu đồ phân phối sai số GDP

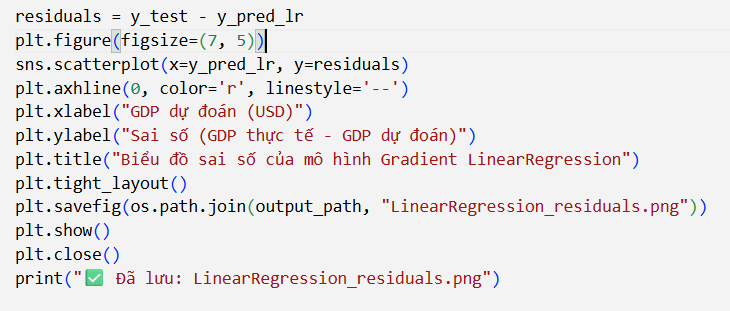
## Huấn luyện mô hình Linear Regression



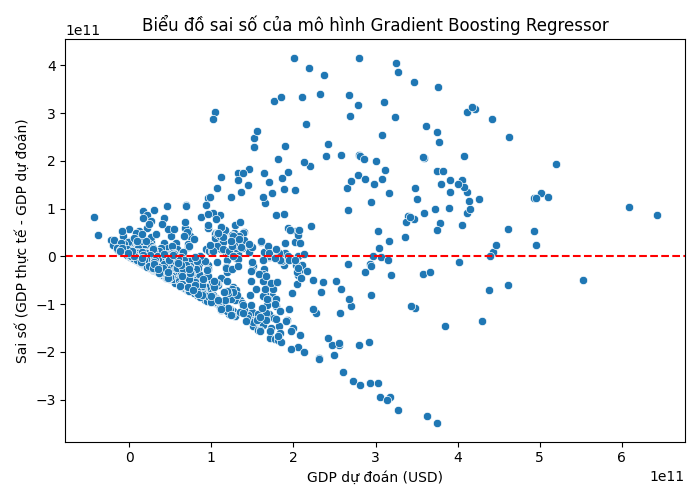
Hình 3. 50 Mã nguồn biểu đồ so sánh GDP thực tế vơi GDP dự đoán



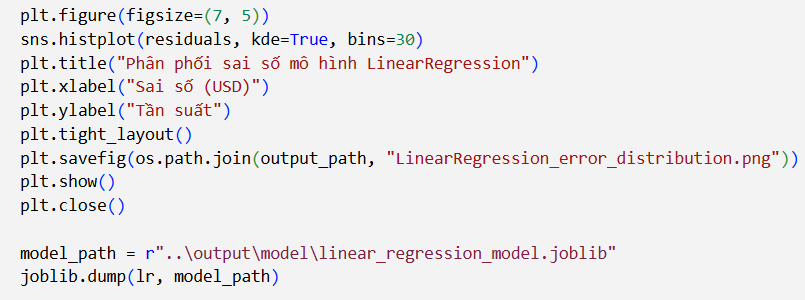
Hình 3. 51 Kết quả biểu đồ so sánh GDP thực tế vơi GDP dự đoán



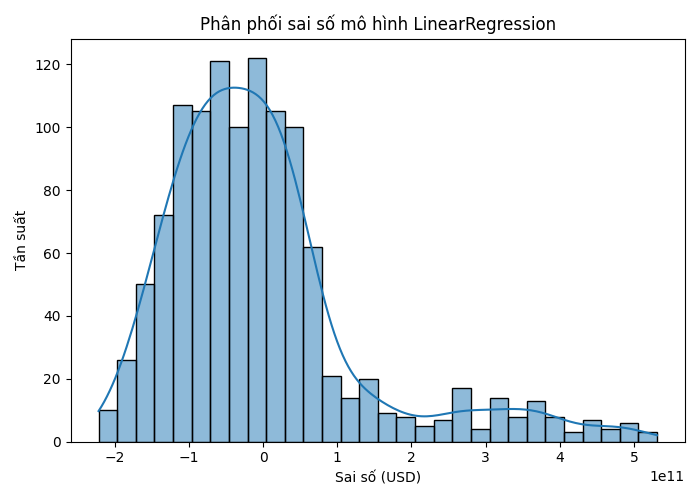
Hình 3. 52 Mã nguồn biểu đồ sai số GDP



Hình 3. 53 Kết quả biểu đồ sai số GDP

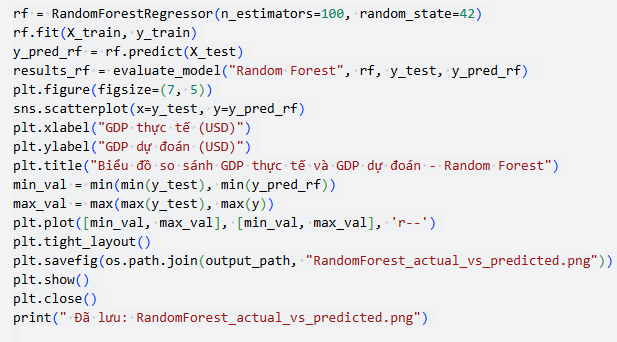


Hình 3. 54 Mã nguồn phân phối sai số GDP

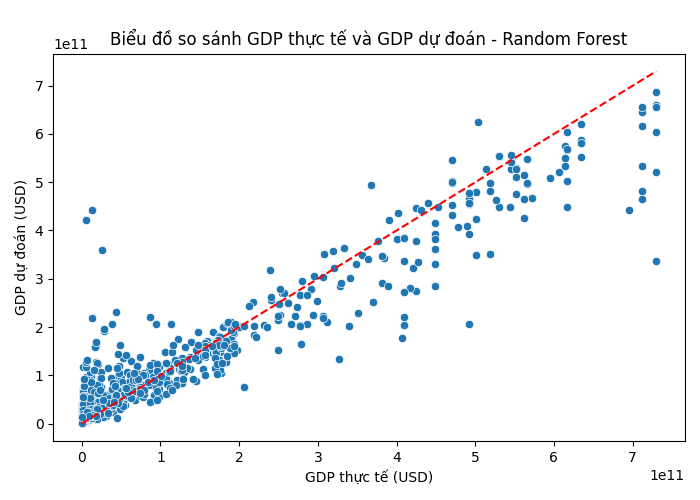


Hình 3. 56 Kết quả phân phối sai số GDP

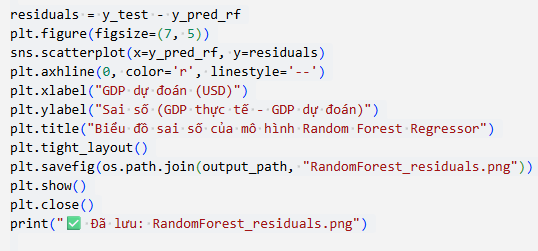
## Huấn luyện mô hình Random Forest Regressor



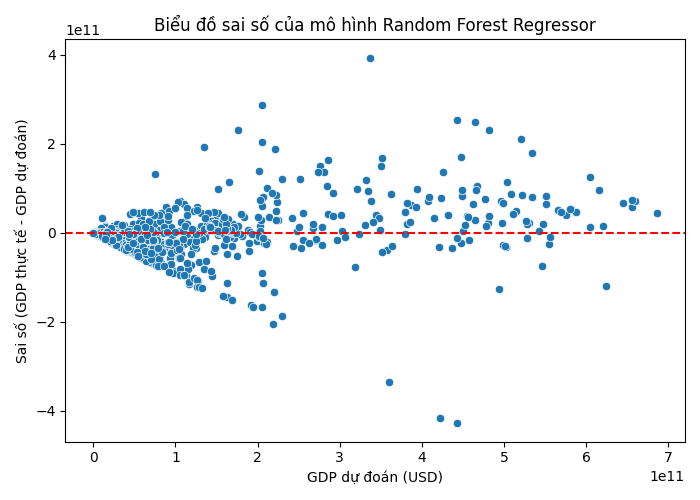
Hình 3. 57 Mã nguồn biểu đồ so sánh GDP thực tế vơi GDP dự đoán



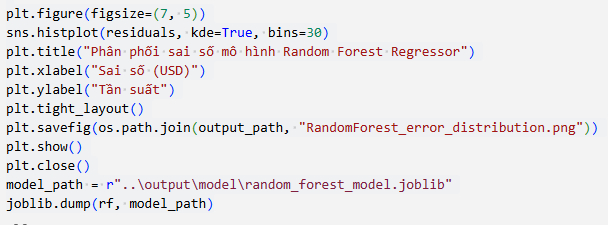
Hình 3. 58 Kết quả biểu đồ so sánh GDP thực tế vơi GDP dự đoán



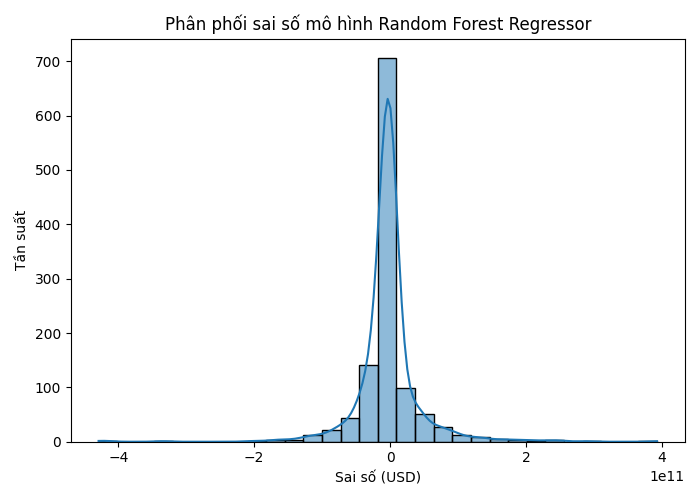
Hình 3.59 Mã nguồn biểu đồ sai số GDP



Hình 3. 60 Kết quả biểu đồ sai số GDP

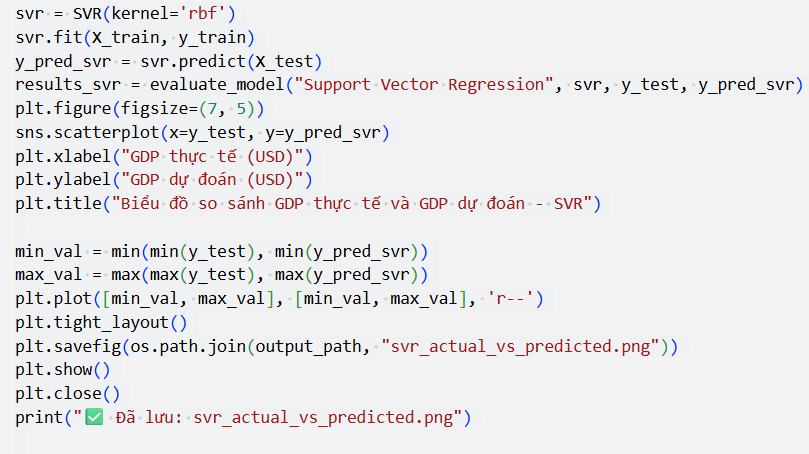


Hình 3. 61 Mã nguồn phân phối sai số GDP

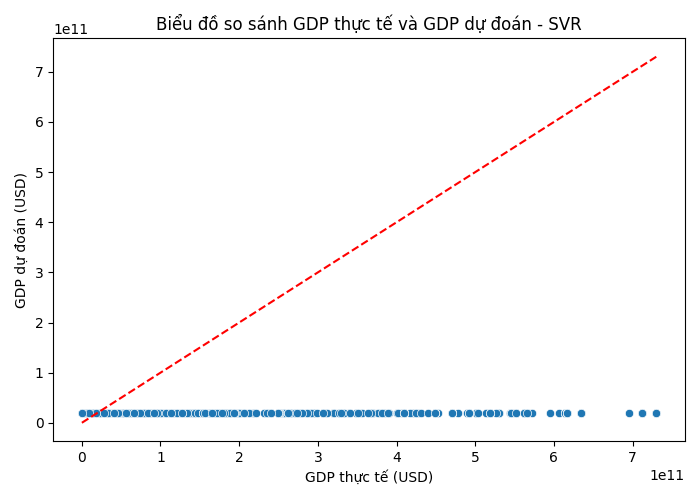


Hình 3.62 Kết quả phân phối sai số GDP

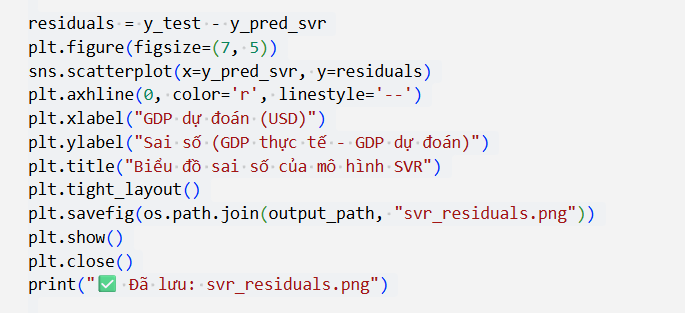
## Huấn luyện mô hình Support Vector Regression (SVR)

x

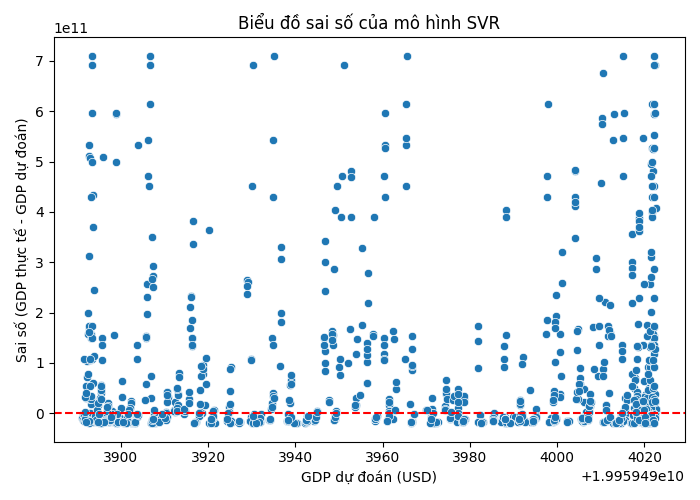
Hình 3.63 Mã nguồn biểu đồ so sánh GDP thực tế vơi GDP dự đoán



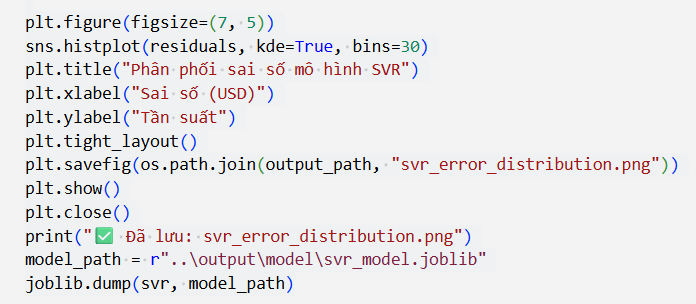
Hình 3.64 Kết quả biểu đồ so sánh GDP thực tế vơi GDP dự đoán



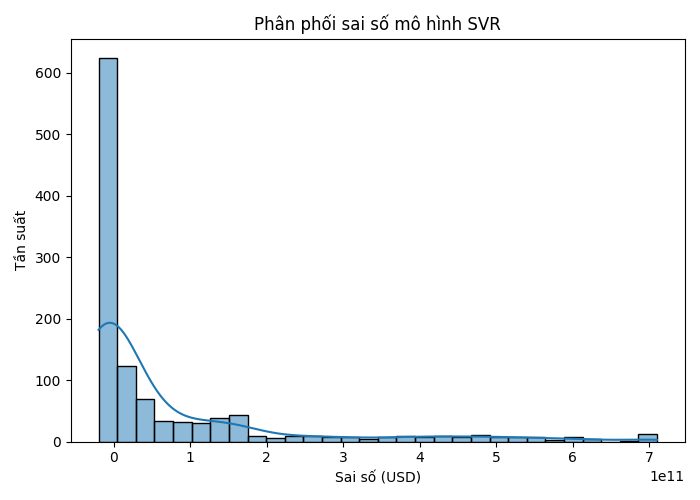
Hình 3.64 Mã nguồn biểu đồ sai số GDP



Hình 3.65 Kết quả biểu đồ sai số GDP

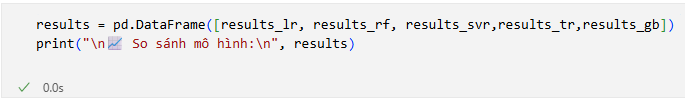


Hình 3.66 Mã nguồn phân phối sai số GDP

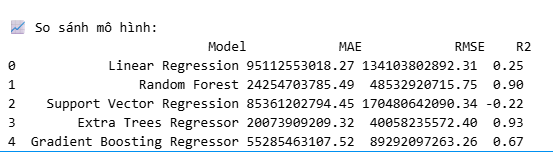


Hình 3.67 Kết quả phân phối sai số GDP

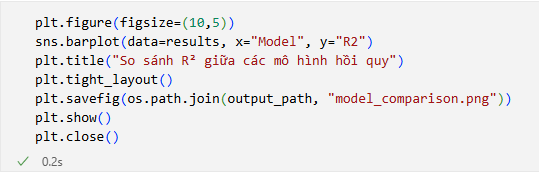
## So sánh và tổng kết giai đoạn huấn luyện



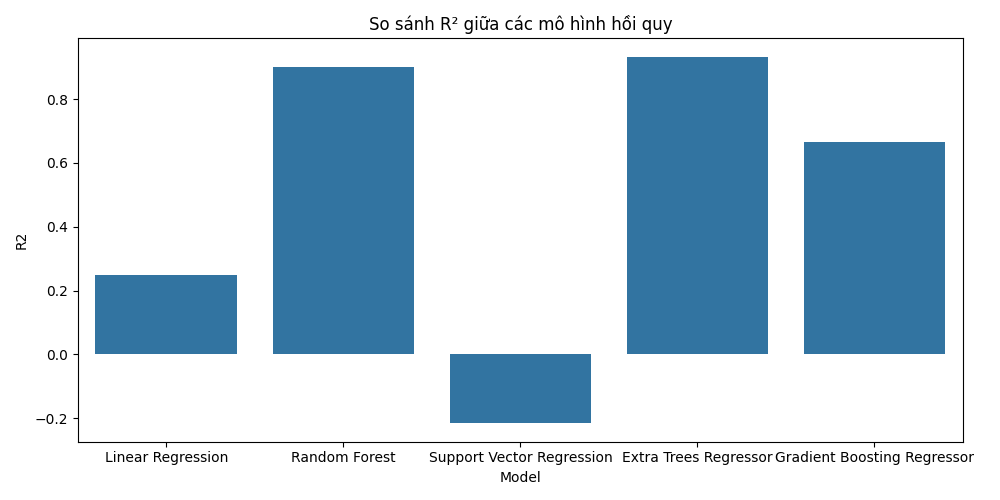
Hình 3.68 Mã nguồn so sách các mô hình

**

Hình 3.69 Kết quả so sách các mô hình

**

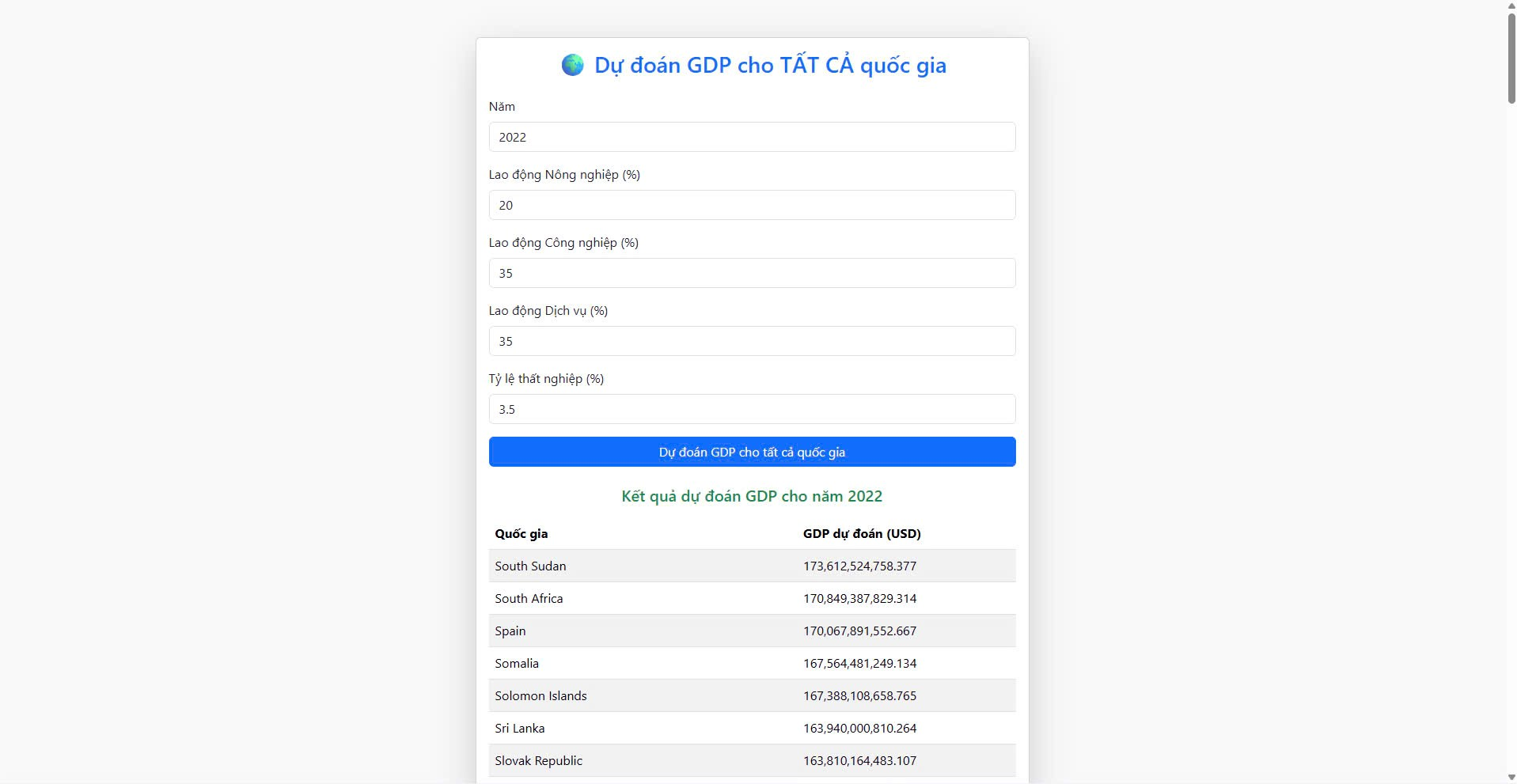
Hình 3.70 Mã nguồn so sánh R² giữa các mô hình

**

Hình 3.71 Kết quả so sánh R² giữa các mô hình

# CHƯƠNG 4: XÂY DỰNG ỨNG DỤNG

Bằng cách sử dụng Streamlit của Python em đã đưa bài toán dự đoán bệnh đột quỵ lên giao diện web.



HÌnh 4. 1 Sử dụng Streamlit để đưa bài toán lên giao diện web

# KẾT LUẬN

## Kết quả đạt được trong kiến thức:

* + **Hiểu rõ cơ chế hoạt động của các mô hình hồi quy trong học máy:**
    - Nắm được nguyên lý của Linear Regression trong việc mô hình hóa mối quan hệ tuyến tính giữa GDP và các yếu tố lao động, thất nghiệp.
    - Hiểu cách Random Forest Regressor và Extra Trees Regressor kết hợp nhiều cây quyết định để tăng độ chính xác và giảm sai số tổng thể.
    - Áp dụng thành công Support Vector Regression (SVR) và Gradient Boosting Regressor để mô hình hóa mối quan hệ phi tuyến giữa các yếu tố kinh tế.
  + **Xây dựng được quy trình xử lý dữ liệu kinh tế hoàn chỉnh:**
    - Thực hiện các bước làm sạch dữ liệu, xử lý giá trị thiếu và loại bỏ giá trị ngoại lai (bằng IQR) để đảm bảo chất lượng dữ liệu đầu vào.
    - Chuẩn hóa dữ liệu bằng StandardScaler và mã hóa các biến phân loại bằng LabelEncoder, giúp mô hình học ổn định và hiệu quả.
    - Phân chia dữ liệu train/test hợp lý để đánh giá khách quan năng lực mô hình.
  + **Thực hiện trực quan hóa dữ liệu kinh tế:**
    - Xây dựng các biểu đồ như Heatmap, Histogram GDP, Boxplot tỷ lệ thất nghiệp, Biểu đồ mối quan hệ GDP – lao động, giúp phân tích xu hướng và mối tương quan giữa các biến.
    - Thực hiện biểu đồ pie thể hiện cơ cấu lao động theo 3 nhóm ngành (Nông nghiệp, Công nghiệp, Dịch vụ) cho từng quốc gia, hỗ trợ trực quan hóa cấu trúc kinh tế.
  + **Huấn luyện, so sánh và lựa chọn mô hình tối ưu**:
    - Sử dụng thư viện LazyPredict để đánh giá nhanh hiệu suất nhiều mô hình hồi quy.
    - Thực nghiệm chi tiết với Extra Trees Regressor, Random Forest Regressor, Gradient Boosting Regressor, Linear Regression, SVR,…
    - Ghi nhận các chỉ số đánh giá như MAE, RMSE, R², và trực quan hóa sai số dự đoán qua biểu đồ scatter và histogram sai số.
    - Kết quả cho thấy Extra Trees Regressor đạt hiệu suất tốt nhất, được chọn làm mô hình chính cho bài toán dự đoán GDP.
  + **Ứng dụng mô hình vào dự đoán thực tế:**
    - Xây dựng hệ thống dự đoán GDP bằng mô hình Extra Trees Regressor, cho phép người dùng nhập dữ liệu đầu vào để nhận kết quả GDP dự đoán.
    - Phát triển giao diện web thân thiện bằng Flask + Bootstrap, giúp mô hình có thể được sử dụng trực tiếp qua trình duyệt mà không cần kỹ năng lập trình.

## Hạn chế của đề tài:

* + **Dữ liệu kinh tế hạn chế về phạm vi và độ cập nhật:**
    - Dữ liệu chủ yếu thu thập từ các năm trước, một số quốc gia thiếu thông tin đầy đủ.
    - Các yếu tố ảnh hưởng khác như lạm phát, thương mại quốc tế, đầu tư, dân số… chưa được tích hợp.
  + **Độ chính xác phụ thuộc vào tiền xử lý:**
    - Mặc dù đã loại bỏ ngoại lai và chuẩn hóa, nhưng sai số vẫn có thể phát sinh do sự biến động tự nhiên của GDP giữa các năm.
    - Việc chọn tham số mô hình (số cây, độ sâu, learning rate, v.v.) chủ yếu dựa trên thực nghiệm, chưa tối ưu toàn diện.
  + **Mô hình chưa tính đến yếu tố thời gian:**
    - GDP là dữ liệu chuỗi thời gian (time-series), tuy nhiên đề tài hiện mới dừng lại ở mô hình hồi quy tĩnh.
    - Chưa áp dụng các mô hình dự báo động như LSTM hoặc ARIMA.

## Hướng phát triển của đề tài:

* + **Mở rộng và cập nhật tập dữ liệu:**
    - Thu thập dữ liệu GDP, cơ cấu lao động và tỷ lệ thất nghiệp theo quý hoặc theo vùng để tăng độ chi tiết và độ chính xác.
    - Bổ sung thêm các yếu tố kinh tế vĩ mô như xuất nhập khẩu, đầu tư, dân số, lạm phát… để mô hình phản ánh thực tế toàn diện hơn.
  + **Nâng cao mô hình dự đoán:**
    - Áp dụng các mô hình hiện đại hơn như XGBoost, CatBoost, hoặc Deep Learning (LSTM, GRU) để dự báo GDP theo thời gian.
    - Sử dụng GridSearchCV hoặc Bayesian Optimization để tự động tìm tham số tối ưu cho từng mô hình.
  + **Phát triển ứng dụng thực tế:**
    - Hoàn thiện API Flask và giao diện web để triển khai thành nền tảng dự báo kinh tế trực tuyến.
    - Mở rộng thành dashboard trực quan hóa với biểu đồ tương tác (Plotly, Dash, hoặc ReactJS).
    - Tích hợp hệ thống này vào các tổ chức nghiên cứu hoặc doanh nghiệp, giúp hỗ trợ ra quyết định về chính sách và đầu tư.

# TÀI LIỆU THAM KHẢO

1. Dữ liệu

[Stroke Prediction Dataset | Kaggle](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)

1. Tài liệu

|  |  |
| --- | --- |
| [1] | Giáo trình Học máy cơ bản ( Nhà xuất bản khoa học và kỹ thuật) |
| [2] | Các slide bài thuyết trình Học máy cơ bản – các nhóm trình bày. |
| [3] | Các slide bài giảng Học máy cơ bản – thầy TS. Hoàng Quốc Việt |
| [4] | Deep Learning - Tác giả: Ian Goodfellow, Yoshua Bengio, Aaron Courville |
| [5] | Python Machine Learning - Tác giả: Sebastian Raschka, Vahid Mirjalili |
| [6] | [https://www.coursera.org/learn/machine-learning](http://www.coursera.org/learn/machine-learning) |
| [7] | https://scikit-learn.org/stable/index.html |
| [8] | [https://www.kaggle.com/](http://www.kaggle.com/) |
| [9] | https://chatgpt.com/ |