ĐẠI HỌC BÁCH KHOA HÀ NỘI

BÁO CÁO BÀI TẬP LỚN

HỆ HỖ TRỢ QUYẾT ĐỊNH

DỰ BÁO GDP CÁC NƯỚC

Giảng viên hướng dẫn: **TS. Trần Ngọc Thăng**

Sinh viên thực hiện: **Nguyễn Nam Thắng**

Mã số sinh viên: **20227183**

HÀ NỘI – 2025

LỜI CẢM ƠN

Em xin chân thành cảm ơn các Thầy, Cô giảng viên tại Khoa Toán – Tin, Trường Đại học Bách khoa Hà Nội. Thầy, Cô đã tận tình giảng dạy và trang bị cho chúng em những kiến thức chuyên môn vững chắc, là nền tảng quan trọng để nhóm có thể thực hiện và hoàn thành đề tài nghiên cứu này.

Thông qua quá trình thực hiện đề tài, em không chỉ được củng cố kiến thức lý thuyết mà còn rèn luyện thêm nhiều kỹ năng cần thiết như tư duy phân tích, xử lý dữ liệu và trình bày học thuật – những kỹ năng quan trọng phục vụ cho học tập và công việc sau này.

Đặc biệt, em xin gửi lời cảm ơn sâu sắc đến **TS. Trần Ngọc Thăng** – người đã trực tiếp hướng dẫn và đồng hành cùng nhóm trong suốt quá trình thực hiện đề tài. Thầy đã tận tình chỉ bảo, định hướng và đưa ra những góp ý quý báu, giúp nhóm hoàn thiện báo cáo một cách tốt nhất.

Tuy đã nỗ lực để hoàn thiện đề tài, nhưng chắc chắn vẫn không tránh khỏi những thiếu sót. Nhóm chúng em rất mong nhận được những ý kiến đóng góp từ Thầy để có thể tiếp tục cải thiện và phát triển tốt hơn trong những nghiên cứu tiếp theo.

Em xin chân thành cảm ơn thầy vì đã luôn đồng hành, định hướng và truyền đạt những kiến thức chuyên môn quý giá, giúp em hoàn thành tốt quá trình nghiên cứu này.

Em xin trân trọng cảm ơn!

LỜI NÓI ĐẦU

Trong bối cảnh nền kinh tế toàn cầu đang trải qua những biến động liên tục do ảnh hưởng của các yếu tố như khủng hoảng tài chính, biến đổi khí hậu, xung đột địa chính trị và tiến bộ công nghệ, việc dự báo **Tổng sản phẩm quốc nội (GDP)** của các quốc gia đóng vai trò ngày càng quan trọng. Dự báo GDP không chỉ giúp các nhà hoạch định chính sách đánh giá xu hướng phát triển kinh tế trong tương lai, mà còn hỗ trợ doanh nghiệp và nhà đầu tư đưa ra các quyết định chiến lược một cách hiệu quả và kịp thời.

Với sự phát triển nhanh chóng của khoa học dữ liệu, đặc biệt là các kỹ thuật **học máy (machine learning)** và **trí tuệ nhân tạo (AI)**, việc ứng dụng các mô hình tiên tiến để phân tích và dự báo GDP đã trở nên khả thi và ngày càng chính xác hơn so với các phương pháp thống kê truyền thống.

Đề tài này được thực hiện nhằm hai mục tiêu chính:

- **Khai thác và xử lý dữ liệu kinh tế vĩ mô** từ các nguồn uy tín như Ngân hàng Thế giới (World Bank) hoặc Quỹ Tiền tệ Quốc tế (IMF), bao gồm các biến số liên quan như lạm phát, tỷ lệ thất nghiệp, đầu tư, tiêu dùng, xuất nhập khẩu, ...
- **Áp dụng và so sánh hiệu quả các mô hình học máy** như Linear Regression, Random Forest ... để dự báo GDP, từ đó đánh giá độ chính xác của từng mô hình và đề xuất mô hình phù hợp nhất cho từng nhóm quốc gia hoặc khu vực cụ thể.

Kết quả của đề tài không chỉ mang ý nghĩa học thuật mà còn có giá trị thực tiễn trong việc hỗ trợ hoạch định chính sách, xây dựng kịch bản phát triển và định hướng đầu tư trong bối cảnh kinh tế ngày càng phức tạp và khó lường.

Mục lục

CHECKLIST VÀ BẢNG ĐÁNH GIÁ MÔ HÌNH

STT	Loại yêu cầu	Yêu cầu	Điểm chữ	Điểm số	Check	Minh chứng
1	Xử lý dữ liệu (2 điểm)	Mô tả bài toán, đầu vào, đầu ra, yêu cầu xử lý	A	1	x	Trang 6-8
2		Đánh nhãn & Tiền xử lý dữ liệu	A	1	x	Trang 9-10
3		Thống kê dữ liệu mẫu	A	1	x	Trang 12-14
4		Xử lý mất cân bằng dữ liệu (cho bài toán phân lớp) hoặc Chuyển đổi dữ liệu (cho bài toán hồi quy)	A	1	x	Trang 10-12
5	Đánh giá mô hình (1 điểm)	Đề xuất và lựa chọn các tiêu chí đánh giá (về độ chính xác, tốc độ, khả năng ứng dụng,...)	A	1	x	Trang 15-17
6		Thống kê và phân tích lỗi	A	1	x	Trang 17-19
7	Cải tiến mô hình (4 điểm)	Kiến trúc mô hình 1	A	1	x	Trang 20-21
8		Kiến trúc mô hình 2	A	1	x	Trang 21-22
9		Kiến trúc mô hình 3	A	1	x	Trang 22-23
10		Kiến trúc mô hình 4	A	1	x	Trang 23-24
11		Kiến trúc mô hình 5	A	1	x	Trang 24-25
12		Kiến trúc mô hình 6	A	1	x	Trang 25-26
13		Kiến trúc mô hình 7	A	1	x	Trang 26-27
14		Kiến trúc mô hình 8	A	1	x	Trang 27-28
15	Đóng gói mô hình (3 điểm)	Có sử dụng mô hình tiên tiến trong 3 năm trở lại đây (chỉ ra paper liên quan)	A	1	x	Trang 28-29
16		Có khả năng ứng dụng vào một ngữ cảnh cụ thể (ứng dụng vào bài toán nghiệp vụ nào, ai là người sử dụng)	A	1	x	Trang 30-31
17		Các chỉ số đánh giá mô hình đủ điều kiện để ứng dụng vào thực tế	A	1	x	Trang 31-33
18		Đóng gói giao diện demo chương trình	A	1	x	Trang 31-33
19		Làm slide báo cáo	A	1	x	
20		Thuyết trình trên lớp	A	1	x	
		Tổng điểm/20		20	0	
		Tổng điểm/10		10		

Bảng mô tả chi tiết mô hình

STT	Tên mô hình	Điều kiện dừng	Phương pháp tối ưu hóa siêu tham số	Siêu tham số của mô hình	Kết quả đánh giá trên dữ liệu test theo các chỉ số				Chú giải
					R^2	MAE	RMSE	MAPE (%)	
1	Linear Regressor	Mặc định của mô hình	Siêu tham số được chọn thủ công	Mô hình theo kiến trúc mới	0.973	11.09	12.73	3.56	Mô hình baseline
2	ARIMA (AutoRegressive Integrated Moving Average)	ARIMA tự dừng khi hội tụ theo tiêu chí AIC	sử dụng thư viện auto_arima	Mô hình theo kiến trúc mới	0.897	21.38	22.99	6.74	Mô hình baseline
3	Random Forest Regressor	Random Forest dừng sau khi xây xong 50 cây	Siêu tham số được chọn thủ công	Mô hình theo kiến trúc mới	0.9639	10.78	13.61	3.11	mô hình nâng cao trên dữ liệu đã loại bỏ xu hướng
4	SVR (Support Vector Regression)	Mặc định của mô hình	Siêu tham số được chọn thủ công	Mô hình theo kiến trúc mới	0.9578	12.42	15.25	3.99	Mô hình baseline
5	KNN (K-Nearest Neighbors Regressor)	Mặc định của mô hình	Siêu tham số được chọn thủ công	Mô hình theo kiến trúc mới	0.8593	17.83	25.16	4.63	Mô hình baseline
6	Elastic Net Regression	max_iter=10000 (ElasticNet hội tụ)	Siêu tham số được chọn thủ công	Mô hình theo kiến trúc mới	0.9483	14.4	16.3	4.62	mô hình có regularization nâng cao
7	Bayesian Ridge Regression	Tối đa 300 vòng lặp (max_iter=300), hoặc khi thay đổi nhỏ hơn tol=1e-3	Tự động ước lượng bằng tối đa hóa xác suất hậu nghiệm	Mô hình theo kiến trúc mới	0.9448	15.04	16.84	4.94	Mô hình baseline
8	Neural Basis Expansion Analysis for Time Series Forecasting	Số epoch cố định (30), Optuna pruning dừng sớm trial kém.	Dùng Optuna (TPE sampler), tối ưu kiến trúc và tham số huấn luyện	Mô hình theo kiến trúc mới	0.8856	23.45	26.2	7.71	Mô hình baseline

Chương 1

PHÁT BIỂU BÀI TOÁN

1.1 Mô tả bài toán

Bối cảnh và Tầm quan trọng của Bài toán

Tổng sản phẩm quốc nội (GDP) là một trong những chỉ số kinh tế vĩ mô quan trọng nhất, phản ánh tổng giá trị thị trường của tất cả các hàng hóa và dịch vụ cuối cùng được sản xuất ra trong phạm vi một quốc gia trong một khoảng thời gian nhất định (thường là một năm). Việc theo dõi và dự báo GDP có ý nghĩa then chốt đối với nhiều đối tượng:

- **Chính phủ và các nhà hoạch định chính sách:** Sử dụng dự báo GDP để xây dựng kế hoạch ngân sách, điều chỉnh chính sách tài khóa và tiền tệ, đánh giá hiệu quả của các chương trình kinh tế, và đặt ra các mục tiêu tăng trưởng.
- **Doanh nghiệp và Nhà đầu tư:** Dựa vào dự báo GDP để đưa ra quyết định đầu tư, mở rộng sản xuất, quản lý rủi ro, và đánh giá tiềm năng thị trường tại các quốc gia khác nhau.
- **Các tổ chức quốc tế (IMF, World Bank, UN):** Sử dụng GDP và dự báo GDP để theo dõi sức khỏe kinh tế toàn cầu, cung cấp hỗ trợ tài chính, và đưa ra các khuyến nghị chính sách.
- **Giới học thuật và Nghiên cứu:** Phân tích xu hướng GDP giúp hiểu rõ hơn về các chu kỳ kinh tế, tác động của các cú sốc, và các yếu tố thúc đẩy tăng trưởng dài hạn.

Trong bối cảnh kinh tế toàn cầu ngày càng biến động và phức tạp, việc xây dựng các mô hình dự báo GDP chính xác và đáng tin cậy trở nên cấp thiết hơn bao giờ hết. Bài toán này tập trung vào việc áp dụng các kỹ thuật phân tích dữ liệu và học máy để dự báo GDP hàng năm của các quốc gia trên thế giới.
Bài toán được đặt ra là phát triển và đánh giá một tập hợp các mô hình dự báo nhằm ước tính giá trị GDP theo giá hiện tại (đơn vị: Tỷ đô la Mỹ) cho các quốc gia trong danh sách dữ liệu cho các năm tiếp theo (ví dụ: 2024–2028). Dữ liệu lịch sử GDP kéo dài từ năm 1980 đến năm 2023 sẽ được sử dụng làm cơ sở để huấn luyện và kiểm định các mô hình.

1.2 Đầu vào (Input) của Bài toán

Dữ liệu chính: Bộ dữ liệu GDP lịch sử (1980-2023) từ kaggle có tên `World GDP Dataset.csv` .

Kích thước bộ dữ liệu :

- Số hàng ban đầu : 230
- Số cột ban đầu : 45
- Tổng số phần tử ban đầu: 10350

Cấu trúc bộ dữ liệu : Bộ dữ liệu có tổng cột 45 cột trong đó

- **Cột 1 (GDP, current prices (Billions of U.S. dollars)):** Tên các quốc gia trên thế giới
- **Các cột còn lại :** Mỗi năm là 1 cột thể hiện cho GDP của các nước qua từng năm

Dưới đây là hình ảnh 5 dòng đầu tiên của bộ dữ liệu gốc :

	GDP, current prices (Billions of U.S. dollars)	1980	1981	1982	1983	1984	1985	1986	1987	1988	...	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
0	Afghanistan	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	20.616	20.057	18.020	18.883	18.401	18.876	20.136	0.000	0.000	0.000
1	Albania	1.946	2.229	2.296	2.319	2.290	2.339	2.587	2.566	2.530	...	13.246	11.389	11.862	13.053	15.157	15.399	15.161	18.310	18.256	18.842
2	Algeria	42.346	44.372	44.780	47.529	51.513	61.132	61.535	63.300	51.664	...	213.810	165.979	160.034	170.207	175.372	171.680	144.922	162.711	187.155	190.254
3	Andorra	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	3.267	2.789	2.895	2.993	3.217	3.155	2.885	3.330	3.302	3.360
4	Angola	6.639	6.214	6.214	6.476	6.864	8.457	7.918	9.050	9.818	...	145.712	116.194	101.124	122.022	101.353	84.516	58.125	75.179	124.794	135.558
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
225	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
226	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
227	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
228	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
229	©IMF, 2022	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

230 rows × 45 columns

Bộ dữ liệu trước khi xử lý

1.3 Đầu ra (Output) của bài toán

Mục tiêu cuối cùng của dự án này là cung cấp một bộ kết quả đầu ra toàn diện, cho phép đánh giá và ứng dụng các mô hình dự báo GDP.

Dự báo GDP Điểm (Point Forecasts) cho Tương lai: Giá trị GDP cụ thể được dự báo cho mỗi quốc gia trong N năm tiếp theo (ví dụ, N=5, từ năm 2024 đến 2028).

Đánh giá Định lượng Hiệu suất Mô hình: Bảng tổng hợp hoặc biểu đồ so sánh hiệu suất của 8 mô hình đã xây dựng, giúp xác định các phương pháp tiếp cận phù hợp nhất cho bài toán và bộ dữ liệu này.

Phân tích Trực quan Kết quả Dự báo và Chẩn đoán Mô hình:

- **Biểu đồ so sánh Thực tế và Dự báo:** Biểu đồ đường thể hiện GDP thực tế, dự đoán trên tập kiểm tra và dự báo cho tương lai từ các mô hình tiêu biểu.
- **Biểu đồ phân tích phần dư**
- **Mục đích:** Chẩn đoán tính phù hợp của mô hình, phát hiện sai lệch và đánh giá tính ổn định.

Báo cáo Phân tích và Diễn giải Kết quả:

- Đánh giá mô hình tốt nhất và lý do.
- So sánh nhóm mô hình sử dụng trong bài.
- Nhận xét về ảnh hưởng của đặc trưng, tham số và hạn chế mô hình.
- Gợi ý cải tiến tương lai.

1.4 Yêu cầu xử lý

Để đạt mục tiêu dự án, quá trình xử lý và phân tích dữ liệu gồm các bước chính sau:

1. **Thu thập và tiền xử lý:**
    - Đọc dữ liệu GDP từ file nguồn, loại bỏ dòng trống, footer.
    - Xử lý giá trị "0" như dữ liệu thiếu, chuẩn hóa tên cột.
2. **Chuyển đổi định dạng:**
    - Chuyển từ dạng *wide* sang *long* (Quốc gia–Năm–GDP).
    - Đảm bảo định dạng số cho cột năm và GDP.
3. **Thống kê khám phá:**
    - Thống kê mô tả và trực quan hóa GDP theo thời gian từng quốc gia.
4. **Sắp xếp Dữ liệu và Hoàn thiện:**
    - Sắp xếp dữ liệu theo thứ tự tăng dần của hai cột: `'Country Name'` và `'Year'` sau khi chuyển đổi và xử lý dữ liệu.

Chương 2

XỬ LÝ DỮ LIỆU

2.1 Thu thập dữ liệu

Như em đã mô tả ở phần input ta có :

Nguồn dữ liệu: Bộ dữ liệu về GDP theo giá hiện tại của các quốc gia trên thế giới, lưu dưới dạng file `World GDP Dataset.csv`.

Mô tả ban đầu:

- Cột đầu tiên: Tên quốc gia.
- Các cột tiếp theo: GDP (Tỷ USD) từ năm 1980 đến 2023.

Phương pháp thu thập : Dữ liệu được đọc bằng Python sử dụng thư viện `pandas`, hàm `pd.read_csv()`.

	GDP, current prices (Billions of U.S. dollars)	1980	1981	1982	1983	1984	1985	1986	1987	1988	...	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
0	Afghanistan	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	20.616	20.057	18.020	18.883	18.401	18.876	20.136	0.000	0.000	0.000
1	Albania	1.946	2.229	2.296	2.319	2.290	2.339	2.587	2.566	2.530	...	13.246	11.389	11.862	13.053	15.157	15.399	15.161	18.310	18.256	18.842
2	Algeria	42.346	44.372	44.780	47.529	51.513	61.132	61.535	63.300	51.664	...	213.810	165.979	160.034	170.207	175.372	171.680	144.922	162.711	187.155	190.254
3	Andorra	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	...	3.267	2.789	2.895	2.993	3.217	3.155	2.885	3.330	3.302	3.360
4	Angola	6.639	6.214	6.214	6.476	6.864	8.457	7.918	9.050	9.818	...	145.712	116.194	101.124	122.022	101.353	84.516	58.125	75.179	124.794	135.558
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
225	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
226	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
227	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
228	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
229	©IMF, 2022	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

230 rows × 45 columns

Bộ dữ liệu trước khi xử lý

2.2 Đánh nhãn dữ liệu

Trong bối cảnh của bài toán dự báo Tổng sản phẩm quốc nội (GDP) này, một bước "đánh nhãn dữ liệu" là không cần thiết

Bộ dữ liệu được cung cấp (World GDP Dataset.csv) đã bao gồm các giá trị GDP lịch sử cho từng quốc gia qua các năm. Sau quá trình tiền xử lý và chuyển đổi cấu trúc dữ liệu, cột chứa các giá trị GDP này tự nó đã đóng vai trò là biến mục tiêu (target variable) hay "nhãn" mà các mô hình sẽ học để dự đoán.

2.3 Tiền xử lý dữ liệu

2.3.1 Làm sạch dữ liệu cơ bản

- **Loại bỏ dòng trống:** Các hàng hoàn toàn không có dữ liệu đã bị xóa khỏi bộ dữ liệu.
- **Xóa dòng thông tin phụ:** trong bộ dữ liệu này có một dòng chứa chú thích/bản quyền (ví dụ: "©IMF") được xác định là không thuộc dữ liệu chính nên em cũng xóa bỏ.
- **Chuẩn hóa tên cột định danh:** Tên cột đầu tiên 'GDP, current prices (Billions of U.S. dollars)' được đổi thành `'Country Name'` để thống nhất và dễ xử lý.
- **Chuyển đổi giá trị :** các giá trị '0' được chuyển thành NaN

2.3.2 Kết quả sau làm sạch cơ bản:

- DataFrame `df_raw` gồm 196 hàng và 45 cột, tổng cộng 8,820 phần tử.
- Các giá trị "0" đã được chuyển thành `NaN`.

	Country Name	1980	1981	1982	1983	1984	1985	1986	1987	1988	...	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
0	Afghanistan	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	20.616	20.057	18.020	18.883	18.401	18.876	20.136	NaN	NaN	NaN
1	Albania	1.946	2.229	2.296	2.319	2.290	2.339	2.587	2.566	2.530	...	13.246	11.389	11.862	13.053	15.157	15.399	15.161	18.310	18.256	18.842
2	Algeria	42.346	44.372	44.780	47.529	51.513	61.132	61.535	63.300	51.664	...	213.810	165.979	160.034	170.207	175.372	171.680	144.922	162.711	187.155	190.254
3	Andorra	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	3.267	2.789	2.895	2.993	3.217	3.155	2.885	3.330	3.302	3.360
4	Angola	6.639	6.214	6.214	6.476	6.864	8.457	7.918	9.050	9.818	...	145.712	116.194	101.124	122.022	101.353	84.516	58.125	75.179	124.794	135.558
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
191	Vietnam	35.357	17.617	23.369	35.204	61.171	19.045	43.009	53.385	29.501	...	232.888	236.795	252.146	277.071	303.091	327.873	342.941	366.201	413.808	469.620
192	West Bank and Gaza	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	13.990	13.972	15.405	16.128	16.277	17.134	15.532	18.037	18.818	19.398
193	Yemen	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	43.229	42.444	31.318	26.842	21.606	21.888	18.850	19.911	27.594	28.099
194	Zambia	4.246	4.385	4.232	3.653	3.003	2.848	1.962	2.431	4.095	...	27.145	21.245	20.965	25.874	26.312	23.309	18.111	21.313	27.025	28.798
195	Zimbabwe	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	19.499	19.969	20.555	21.385	36.945	22.995	23.181	32.868	38.280	37.303

196 rows × 45 columns

Bộ dữ liệu sau khi xử lý cơ bản

2.4 Chuyển đổi dữ liệu

Trong bước này, em tập trung vào việc chuyển đổi dữ liệu về định dạng và kiểu dữ liệu phù hợp nhằm phục vụ tốt hơn cho quá trình phân tích và mô hình hóa. Cụ thể, dữ liệu được tổ chức lại từ định dạng bảng rộng (wide format), vốn không thuận tiện cho xử lý chuỗi thời gian, sang định dạng dài (long format), nơi mỗi bản ghi đại diện cho một quan sát đơn lẻ theo từng quốc gia và năm. Đồng thời, các cột chứa thông tin năm và GDP cũng được chuyển đổi về kiểu số (numeric) để đảm bảo tính toàn vẹn và khả năng tính toán trong các bước phân tích thống kê, trực quan hóa và huấn luyện mô hình dự báo sau này.
Phương thức **pd.melt()** của thư viện Pandas được sử dụng .

2.4.1 Chuyển đổi từ "Wide" sang "Long"

Cột `'Country Name'` giữ lại làm biến định danh.Các cột năm được xoay thành hai cột mới: `'Year'` và `'GDP'`.
Kết quả là bộ dữ liệu sau khi chuyển đổi có kích thước : 8,624 hàng x 3 cột (25,872 phần tử).

	Country Name	Year	GDP
0	Afghanistan	1980	NaN
1	Albania	1980	1.946
2	Algeria	1980	42.346
3	Andorra	1980	NaN
4	Angola	1980	6.639
...	...	...	...
8619	Vietnam	2023	469.620
8620	West Bank and Gaza	2023	19.398
8621	Yemen	2023	28.099
8622	Zambia	2023	28.798
8623	Zimbabwe	2023	37.303

8624 rows × 3 columns

Bộ dữ liệu sau khi chuyển đổi

2.4.2 Xử lý dữ liệu khi sau chuyển đổi

Sau khi chuyển đổi dữ liệu từ định dạng "wide" sang "long", bước tiếp theo là đảm bảo kiểu dữ liệu của các cột phù hợp cho phân tích định lượng. Cụ thể:

- Hai cột **'Year'** và **'GDP'** được chuyển sang kiểu số để phục vụ cho các phép tính toán và mô hình hóa sau này.Trong quá trình chuyển đổi, các giá trị không thể ép kiểu thành số (do chứa ký tự không hợp lệ hoặc bị thiếu) được tự động chuyển thành giá trị thiếu **NaN**.

- Sau đó, các hàng có chứa **NaN** ở bất kỳ một trong hai cột**'Year'** và **'GDP'** đều bị loại bỏ để đảm bảo tính toàn vẹn của dữ liệu.

- Sắp xếp bộ dữ theo **'Country Name'** và sau đó là **'Year'** , Dữ liệu được sắp xếp trước tiên dựa trên tên quốc gia theo thứ tự bảng chữ cái (A-Z). Kết quả là tất cả các dòng dữ liệu thuộc về cùng một quốc gia sẽ được nhóm lại với nhau.Sau khi đã nhóm theo quốc gia, bên trong mỗi nhóm quốc gia, dữ liệu tiếp tục được sắp xếp dựa trên năm theo thứ tự tăng dần (từ năm nhỏ nhất đến năm lớn nhất).

- Chỉ số (index) được đặt lại để đảm bảo thứ tự nhất quán.

Kích thước dữ liệu lúc này: 7852 hàng × 3 cột (23556 phần tử)

	Country Name	Year	GDP
0	Afghanistan	2002	4.367
1	Afghanistan	2003	4.553
2	Afghanistan	2004	5.146
3	Afghanistan	2005	6.167
4	Afghanistan	2006	6.925
...	...	...	...
7847	Zimbabwe	2019	22.995
7848	Zimbabwe	2020	23.181
7849	Zimbabwe	2021	32.868
7850	Zimbabwe	2022	38.280
7851	Zimbabwe	2023	37.303

7852 rows × 3 columns

Kết quả xử lý dữ liệu

2.5 Thống kê dữ liệu mẫu

Sau khi thực hiện các bước xử lý , đánh nhãn , chuyển đổi dữ liệu.... Bây giờ em sẽ thực hiện thống kê lại dữ liệu điều này cho ta biết các thống kê về kích thước, kiểu dữ liệu, sự phân bố giá trị, tình trạng giá trị thiếu, và các đặc điểm trung tâm cũng như độ phân tán của GDP . Những thông tin này sẽ cung cấp rõ ràng về dữ liệu đang được sử dụng, đồng thời là cơ sở quan trọng để nhận diện các đặc điểm nổi bật và các vấn đề cần lưu ý trước khi chuyển sang giai đoạn xây dựng mô hình.

2.5.1 Thống kê dữ liệu chung

– Kích thước : 7852 hàng x 3 cột

– Số lượng giá trị duy nhất cho mỗi cột :

- Country Name: 196
- Year: 44
- GDP: 7052

Dưới đây là 1 số thông số thống kê cho cột **'GDP'** và **'Year'**

	Year	GDP
count	7852.000000	7852.000000
mean	2002.741722	263.121139
std	12.397475	1245.564570
min	1980.000000	0.014000
25%	1993.000000	3.260250
50%	2003.000000	15.974000
75%	2013.000000	99.016500
max	2023.000000	26185.210000

Thống kê tóm tắt cho năm và GDP

2.5.2 Thống kê dữ liệu cho Việt Nam

Với bộ dữ liệu trên chúng ta có thể thực hiện phân tích và dự báo cho 196 quốc gia trên thế giới tuy vậy trong bài báo cáo này em sẽ tập trung vào phân tích và dự báo cho **GDP** của **Vietnam**

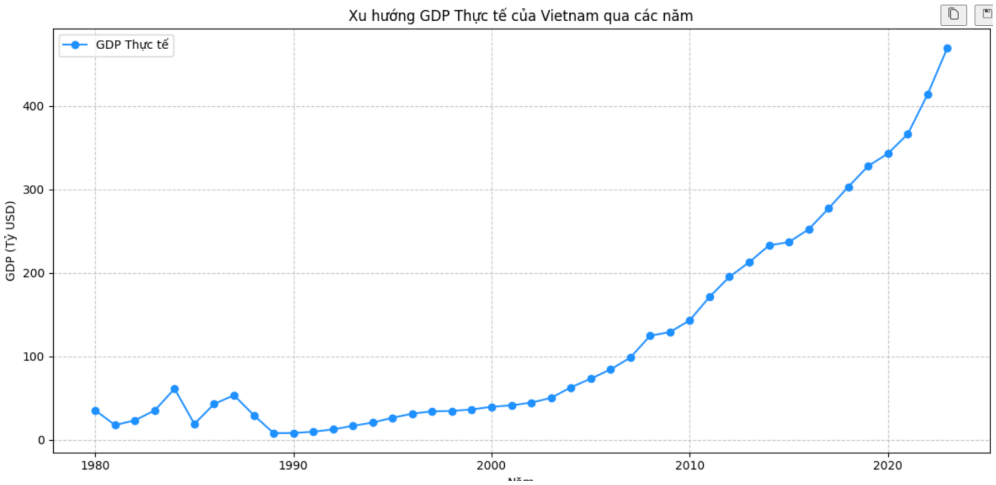
	Country Name	Year	GDP
0	Vietnam	1980	35.357
1	Vietnam	1981	17.617
2	Vietnam	1982	23.369
3	Vietnam	1983	35.204
4	Vietnam	1984	61.171
...	...	...	...
39	Vietnam	2019	327.873
40	Vietnam	2020	342.941
41	Vietnam	2021	366.201
42	Vietnam	2022	413.808
43	Vietnam	2023	469.620

Dữ liệu GDP Việt Nam

	count	mean	std	min	25%	50%	75%	max
GDP	44.0	119.308795	126.304471	7.991	30.88925	51.809	199.55875	469.62

Thống kê mô tả GDP của Việt Nam

Trực quan hóa GDP Lịch sử của Việt Nam



Biểu đồ lịch sử GDP Việt Nam

Biểu đồ đường cho thấy xu hướng tăng trưởng GDP của Vietnam qua các năm.Cho thấy sự tăng trưởng mạnh mẽ từ đầu những năm 2000 tới nay.

Chương 3

ĐÁNH GIÁ MÔ HÌNH

3.1 Các tiêu chí sử dụng đánh giá mô hình

3.1.1 Hệ số xác định R^2

Hệ số xác định, thường được gọi là R^2 là một chỉ số thống kê cho biết mức độ mà một mô hình có thể giải thích được sự biến thiên của biến phụ thuộc dựa trên biến độc lập. Nói cách khác, R^2 đo lường tỷ lệ phần trăm biến động của biến phụ thuộc mà mô hình giải thích được.

Công thức :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Trong đó:

- y_i là giá trị thực tế của biến phụ thuộc.
- $\hat{y}_i$ là giá trị dự đoán của biến phụ thuộc.
- $\bar{y}$ là giá trị trung bình của các giá trị thực tế.

Ý nghĩa:

- R^2 có giá trị từ 0 đến 1.
- $R^2 = 1$: Mô hình giải thích hoàn toàn sự biến thiên của dữ liệu.
- $R^2 = 0$: Mô hình không giải thích được bất kỳ sự biến thiên nào của dữ liệu.

- Giá trị R^2 càng cao, mô hình càng phù hợp.

Hạn chế:

- R^2 không thể đánh giá độ chính xác của các dự đoán, chỉ đo lường độ phù hợp của mô hình.
- R^2 có thể tăng lên khi thêm biến vào mô hình, ngay cả khi các biến thêm vào không có ý nghĩa thống kê.

3.1.2 MAE - Sai số tuyệt đối trung bình

Công thức:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Trong đó:

- y_i: Giá trị thực tế tại quan sát thứ i.
- $\hat{y}_i$: Giá trị dự đoán tại quan sát thứ i.
- n: Tổng số quan sát.

Ý nghĩa:

- MAE đo lường sai số tuyệt đối trung bình giữa giá trị thực tế và giá trị dự đoán.
- Giá trị MAE càng nhỏ, mô hình càng chính xác.

Ưu điểm:

- Dễ hiểu, trực quan.
- Ít nhạy cảm với các giá trị ngoại lai.

Hạn chế:

- Không phản ánh rõ sai số lớn như RMSE.
- Không chuẩn hóa, khó so sánh giữa các tập dữ liệu khác nhau.

3.1.3 RMSE - Căn bậc hai của sai số bình phương trung bình

Công thức:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Trong đó:

- y_i: Giá trị thực tế tại quan sát thứ i.
- $\hat{y}_i$: Giá trị dự đoán tại quan sát thứ i.
- n: Tổng số quan sát.

Ý nghĩa:

- RMSE đo mức độ sai số bình phương trung bình, nhấn mạnh các sai số lớn.
- Giá trị càng nhỏ, mô hình càng chính xác.

Ưu điểm:

- Nhạy với sai số lớn.

- Phổ biến trong các bài toán hồi quy.

Hạn chế:

- Nhạy cảm với giá trị ngoại lai.
- Không chuẩn hóa, khó so sánh giữa các mô hình.

3.1.4 MAPE - Sai số phần trăm tuyệt đối trung bình

Công thức:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Ý nghĩa:

- MAPE biểu thị sai số trung bình theo phần trăm so với giá trị thực tế.
- Dễ hiểu và có thể so sánh giữa các mô hình với đơn vị khác nhau.

Trong đó:

- y_i: Giá trị thực tế tại quan sát thứ i.
- $\hat{y}_i$: Giá trị dự đoán tại quan sát thứ i.
- n: Tổng số quan sát.
- Nếu $y_i = 0$, công thức không xác định và cần xử lý đặc biệt.

Ưu điểm:

- Không phụ thuộc đơn vị đo.
- Phù hợp cho báo cáo trực quan và so sánh mô hình.

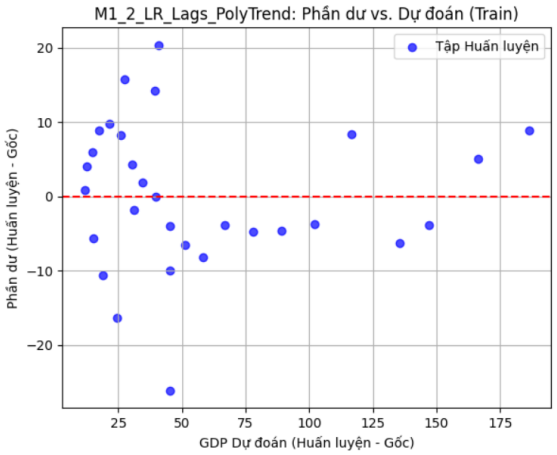
Hạn chế:

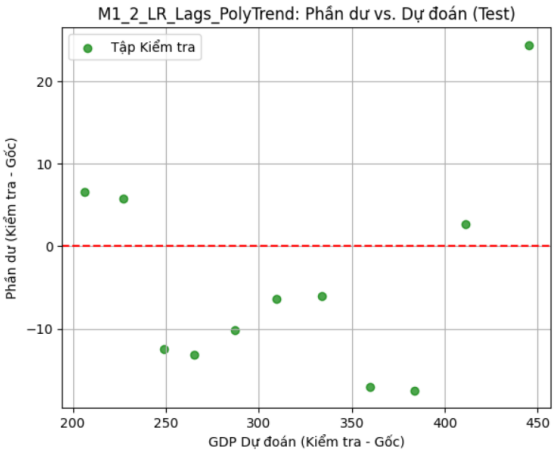
- Không xác định khi $y_i = 0$.
- Nhạy cảm với giá trị y_i nhỏ.

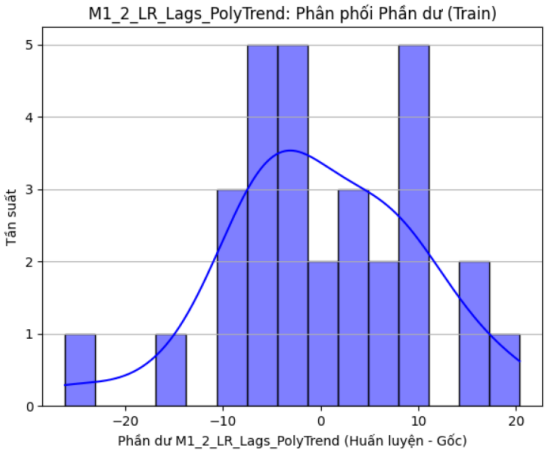
3.2 Thống kê và phân tích lỗi

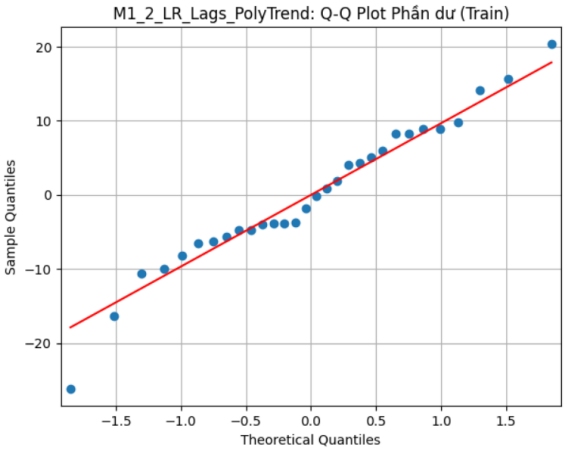
Chúng em sau khi chạy mô hình thì sẽ sẽ tiến hành phân tích phần dư cho một số mô hình tiêu biểu, đặc biệt là những mô hình cho kết quả tốt nhất và có thể là một mô hình cho kết quả kém để so sánh. Ở phần này sẽ phân tích cho mô hình tốt nhất đó là mô hình mô hình mà em đã trình bày ở trên lớp . **Mô hình hồi quy tuyến tính**

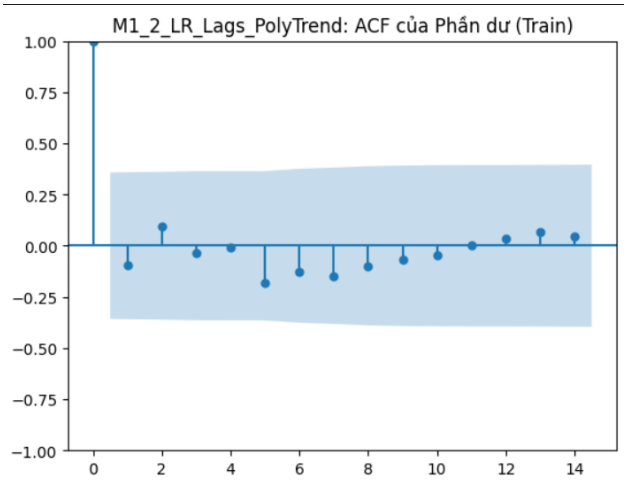
Mô hình hồi quy tuyến tính











Mô hình Hồi quy Tuyến tính này đã thành công trong việc loại bỏ phần lớn tự tương quan trong dữ liệu huấn luyện và không có dấu hiệu rõ ràng về heteroscedasticity hay mối quan hệ phi tuyến bị bỏ sót. Tuy nhiên, phân phối của phần dư không hoàn toàn chuẩn, với khả năng có các giá trị ngoại lệ hoặc đuôi nặng hơn, điều này có thể ảnh hưởng đến độ tin cậy của các kiểm định thống kê và khoảng tin cậy nếu chúng được xây dựng dựa trên giả định chuẩn. Hiệu suất trên tập test cần được xem xét cẩn thận do số lượng điểm ít.

Chương 4

CẢI TIẾN MÔ HÌNH

4.1 Các mô hình sử dụng giải quyết bài toán

Sử dụng nhiều mô hình khác nhau là một chiến lược hiệu quả nhằm nâng cao độ chính xác, giảm thiểu rủi ro thiên vị và tạo ra các dự báo ổn định, đáng tin cậy hơn. Việc kết hợp các mô hình này không chỉ cho phép khai thác tối đa các ưu điểm riêng biệt của từng mô hình, mà còn góp phần tăng khả năng giải thích và tối ưu hóa kết quả dự báo GDP quốc gia. Trong phạm vi đề tài này, nhóm em đã áp dụng 8 mô hình với 4 kiến trúc mô hình khác nhau để giải quyết bài toán dự báo GDP quốc gia.

1. Mô hình hồi quy tuyến tính (Linear Models)

Đây là mô hình *hồi quy tuyến tính* sử dụng các *giá trị GDP trong quá khứ (lag features)* và *xu hướng thời gian theo năm (trend polynomial)* để dự báo GDP của một quốc gia trong tương lai. Các đặc trưng đầu vào bao gồm các giá trị GDP ở các năm trước đó và các biến biểu diễn xu hướng như năm hiện tại (`Year`), bình phương của năm (`Year`^{2}), hoặc lũy thừa bậc cao hơn tùy theo cấu hình. Mô hình có thể áp dụng biến đổi logarit lên GDP nếu cần thiết để xử lý sự phân tán dữ liệu lớn.

$$\text{GDP}_t = \beta_0 + \beta_1 \cdot \text{GDP}_{t-1} + \beta_2 \cdot \text{GDP}_{t-2} + \beta_3 \cdot \text{GDP}_{t-3} + \gamma_1 \cdot \text{Year}_t + \gamma_2 \cdot \text{Year}_t^2 + \gamma_3 \cdot$$

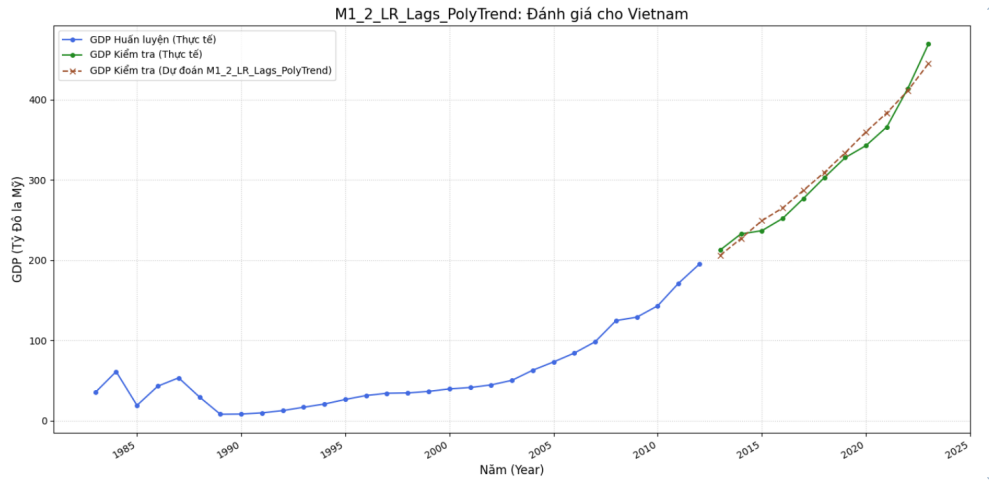
Trong đó:

- GDP_t: GDP tại năm cần dự báo,
- GDP_{t-i}: GDP của các năm trước đó
- $\text{Year}_t, \text{Year}_t^2$: đặc trưng xu hướng theo thời gian
- β_i, γ_j: hệ số mô hình,
- ε_t: sai số ngẫu nhiên.

Kết quả các chỉ số đánh giá:

- **R-squared (Test)**: 0.9730
- **MAE (Test)**: 11.09

- **RMSE (Test)**: 12.73
- **MAPE (Test)**: 3.56%



2.Mô hình ARIMA (AutoRegressive Integrated Moving Average)

Mô hình ARIMA được sử dụng để dự báo giá trị GDP dựa trên chính chuỗi thời gian GDP trong quá khứ. ARIMA tự động nhận diện cấu trúc phụ thuộc theo thời gian trong dữ liệu. Dạng tổng quát của mô hình ARIMA(p, d, q) được viết lại theo ngữ cảnh dữ liệu GDP như sau:

$$\text{GDP}_t = c + \sum_{i=1}^{p} \phi_i \cdot \text{GDP}_{t-i} + \sum_{j=1}^{q} \theta_j \cdot \varepsilon_{t-j} + \varepsilon_t$$

Trong đó:

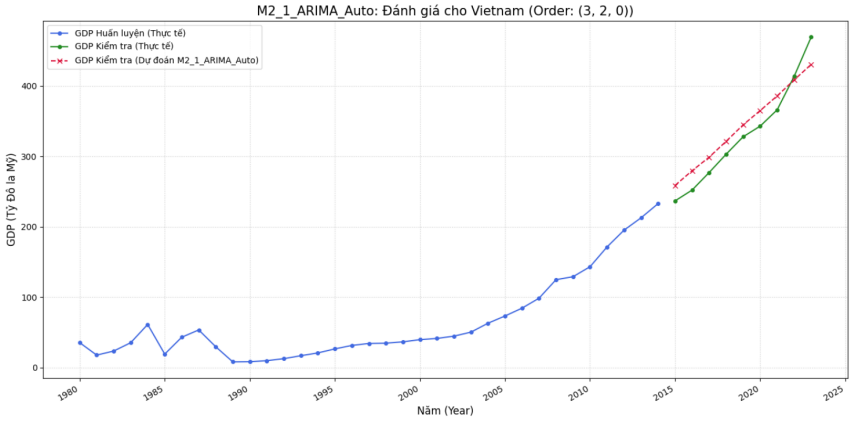
- GDP_t: GDP tại năm t,
- GDP_{t-i}: GDP ở các năm trước đó (thành phần tự hồi quy, AR),
- ε_t: sai số ngẫu nhiên tại năm t,
- ϕ_i: hệ số của thành phần AR (Autoregressive),
- θ_j: hệ số của thành phần MA (Moving Average),
- c: hằng số (intercept) của mô hình.

Trước khi huấn luyện, chuỗi GDP có thể được lấy sai phân d lần để đảm bảo tính dừng. Các tham số tối ưu (p, d, q) có thể nhập thủ công nhưng để cho mô hình tối ưu nhất có thể các tham số được xác định tự động bằng hàm **auto_arima**, dựa trên tiêu chí AIC (Akaike Information Criterion).

Kết quả các chỉ số đánh giá:

- **R-squared (Test)**: 0.8971
- **MAE (Test)**: 21.38

- **RMSE (Test)**: 22.99
- **MAPE (Test)**: 6.74%



3. Mô hình Random Forest Regressor

Mô hình gồm hai bước chính: (1) ước lượng và loại bỏ thành phần xu hướng trong chuỗi GDP, và (2) dự báo phần biến động còn lại bằng mô hình Random Forest sử dụng các giá trị trễ (lags).

1. **Ước lượng xu hướng GDP bằng hồi quy tuyến tính:**

$$\widehat{\text{GDP}}_t^{\text{trend}} = \alpha_0 + \alpha_1 \cdot \text{Year}_t + \alpha_2 \cdot \text{Year}_t^2 + \alpha_3 \cdot \text{TimeIndex}_t$$

2. **Tính phần GDP đã loại bỏ xu hướng (detrended):**

$$\text{GDP}_t^{\text{detrended}} = \text{GDP}_t - \widehat{\text{GDP}}_t^{\text{trend}}$$

3. **Dự báo phần detrended bằng Random Forest:**

$$\widehat{\text{GDP}}_t^{\text{detrended}} = f_{\text{RF}}(\text{GDP}_{t-1}^{\text{detrended}}, \text{GDP}_{t-2}^{\text{detrended}}, \ldots, \text{GDP}_{t-k}^{\text{detrended}})$$

4. **Tái tạo dự báo GDP cuối cùng:**

$$\widehat{\text{GDP}}_t = \widehat{\text{GDP}}_t^{\text{trend}} + \widehat{\text{GDP}}_t^{\text{detrended}}$$

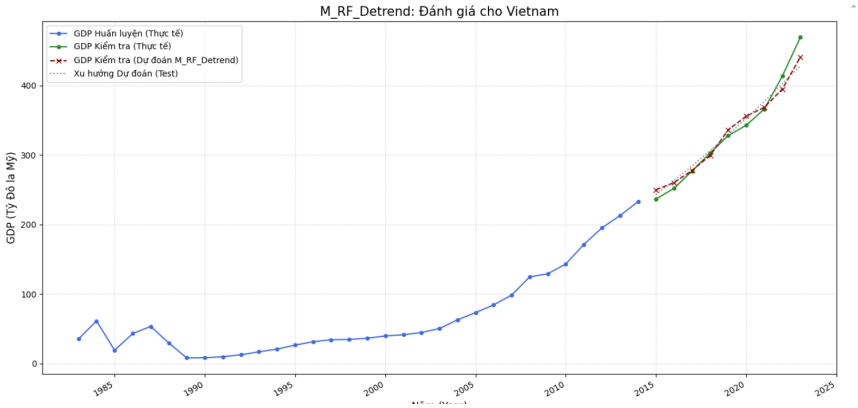
Trong đó:

- Year_t: năm tương ứng với thời điểm t,
- TimeIndex_t: chỉ số thời gian tuần tự (0, 1, 2, ...),
- f_{RF}: hàm dự báo của mô hình Random Forest,
- k: số lượng lags sử dụng để huấn luyện Random Forest.

Kết quả các chỉ số đánh giá:

- **R-squared (Test)**: 0.9639

- **MAE (Test)**: 10.78
- **RMSE (Test)**: 13.61
- **MAPE (Test)**: 3.11%



4. Mô hình SVR (Support Vector Regression)

Đây là mô hình *Support Vector Regression (SVR)* sử dụng **5 giá trị GDP quá khứ (lags)**, **năm hiện tại** và **bình phương của năm** làm đặc trưng đầu vào để dự báo GDP trong tương lai. Dữ liệu đầu vào được chuẩn hóa trước khi huấn luyện để đảm bảo hiệu quả mô hình.

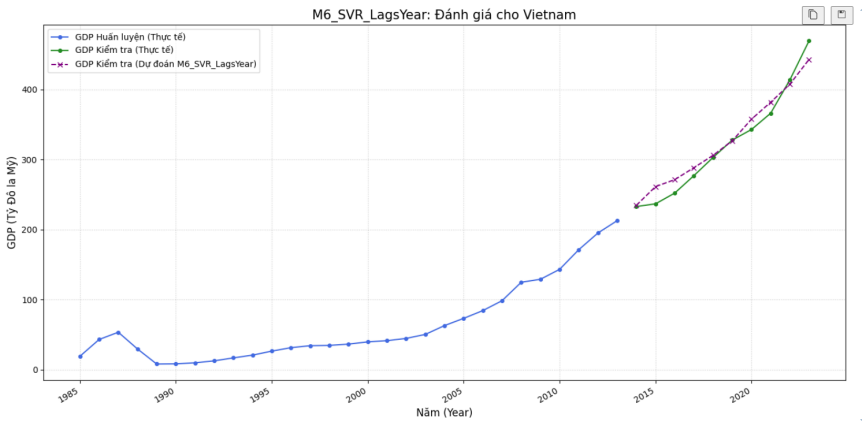
$$\text{GDP}_t = w_0 + \sum_{i=1}^{5} w_i \cdot \text{GDP}_{t-i} + \gamma_1 \cdot \text{Year}_t + \gamma_2 \cdot \text{Year}_t^2 + \varepsilon_t$$

Tham số mô hình:

- **Loại mô hình**: SVR
- **Kernel**: `linear`
- **C**: 1.0 (hệ số phạt sai số lớn)
- **Epsilon**: 0.1 (vùng không phạt sai số)
- **Số lượng lags**: 5
- **Đặc trưng xu hướng**: `Year`, $\texttt{Year}^2$

Kết quả các chỉ số đánh giá:

- **R-squared (Test)**: 0.9578
- **MAE (Test)**: 12.42
- **RMSE (Test)**: 15.25
- **MAPE (Test)**: 3.99%



5. Mô hình KNN (K-Nearest Neighbors Regressor)

Mô hình *K-Nearest Neighbors Regression (KNN)* được áp dụng để dự báo GDP bằng cách học từ chuỗi **sai phân bậc 1 của GDP** và các **lag đặc trưng** của chuỗi này. Sau khi dự báo sai phân $\widehat{\Delta\text{GDP}}_t$, ta tích hợp ngược với giá trị GDP_{t-1} để thu được $\widehat{\text{GDP}}_t$.

Công thức mô hình KNN sai phân:

Đầu tiên, tính sai phân bậc 1 của GDP:

$$\Delta\text{GDP}_t = \text{GDP}_t - \text{GDP}_{t-1}$$

Sau đó, tạo các đặc trưng lag từ chuỗi sai phân:

$$X_t = [\Delta\text{GDP}_{t-1}, \Delta\text{GDP}_{t-2}, \Delta\text{GDP}_{t-3}]$$

Mô hình KNN hồi quy được huấn luyện để dự đoán sai phân hiện tại:

$$\widehat{\Delta\text{GDP}}_t = f_{\text{KNN}}(X_t)$$

Cuối cùng, dự báo GDP được tích hợp ngược lại từ sai phân và GDP của năm trước:

$$\widehat{\text{GDP}}_t = \text{GDP}_{t-1} + \widehat{\Delta\text{GDP}}_t$$

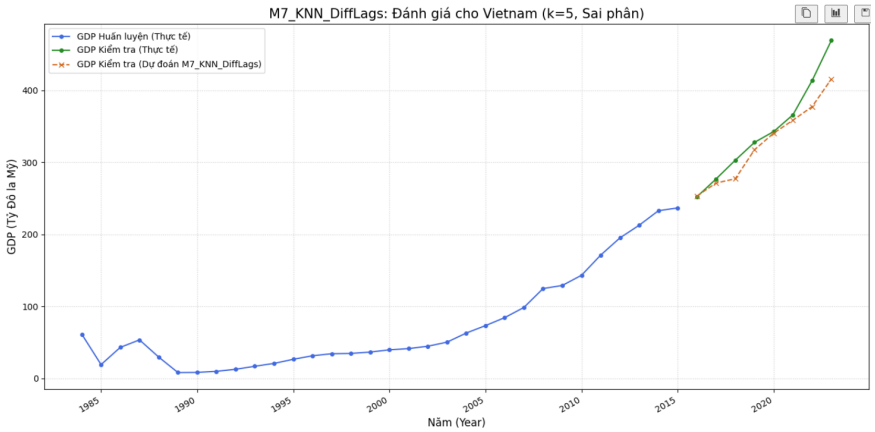
Tham số mô hình:

- **Loại mô hình**: KNN Regression
- **Số lượng láng giềng (K)**: 5
- **Sai phân sử dụng**: bậc 1 ($\Delta\text{GDP}_t = \text{GDP}_t - \text{GDP}_{t-1}$)
- **Số lags đầu vào**: 3 lags của ΔGDP
- **Chuẩn hóa**: Các đặc trưng đầu vào được chuẩn hóa (standardization)
- **Trọng số**: `distance-based` (láng giềng gần hơn ảnh hưởng lớn hơn)

- **Khoảng cách**: Minkowski ($p = 2$, tương đương khoảng cách Euclid)

Kết quả các chỉ số đánh giá:

- **R-squared (Test)**: 0.8593
- **MAE (Test)**: 17.83
- **RMSE (Test)**: 25.16
- **MAPE (Test)**: 4.63%



6. Mô hình Elastic Net Regression Mô hình dự đoán GDP dựa vào GDP các năm trước (3 lags), năm hiện tại (Year) và bình phương năm (Year^2) , mô hình kết hợp L1 (Lasso) và L2 (Ridge) để vừa chọn lọc đặc trưng, vừa giảm quá khớp. Có khả năng tự động làm co nhỏ hoặc loại bỏ các hệ số không quan trọng.

Đặc trưng đầu vào:

- GDP trễ: GDP_{t-1}, GDP_{t-2}, GDP_{t-3}
- Năm hiện tại: Year_t
- Bình phương năm: Year_t^2

Công thức hồi quy:

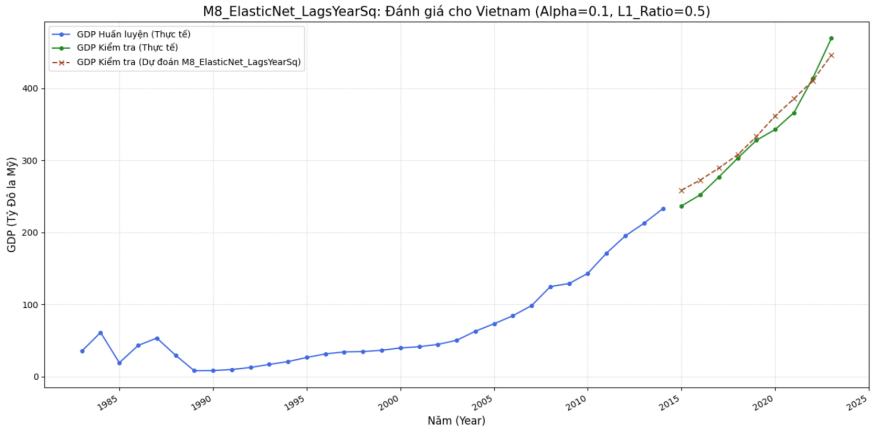
$$\text{GDP}_t = \beta_0 + \beta_1 \cdot \text{GDP}_{t-1} + \beta_2 \cdot \text{GDP}_{t-2} + \beta_3 \cdot \text{GDP}_{t-3} + \beta_4 \cdot \text{Year}_t + \beta_5 \cdot \text{Year}_t^2 + \varepsilon$$

Tham số mô hình ElasticNet:

- $\alpha = 0.1$: Cường độ regularization tổng thể. Giá trị càng lớn, mô hình càng bị phạt nhiều.
- l1_ratio = 0.5: Cân bằng giữa L1 (Lasso) và L2 (Ridge).

  Kết quả các chỉ số đánh giá:

  - **R-squared (Test)**: 0.9483
  - **MAE (Test)**: 14.40
  - **RMSE (Test)**: 16.30
  - **MAPE (Test)**: 4.62%



7. Mô hình Bayesian Ridge Regression

Đây là biến thể của mô hình hồi quy tuyến tính sử dụng các đặc trưng gồm các giá trị GDP trễ, năm hiện tại và bình phương của năm, trong đó các hệ số hồi quy được ước lượng theo phương pháp Bayes. Dữ liệu được chuẩn hóa trước khi huấn luyện. Mô hình giúp kiểm soát overfitting thông qua ước lượng phân phối của các hệ số.

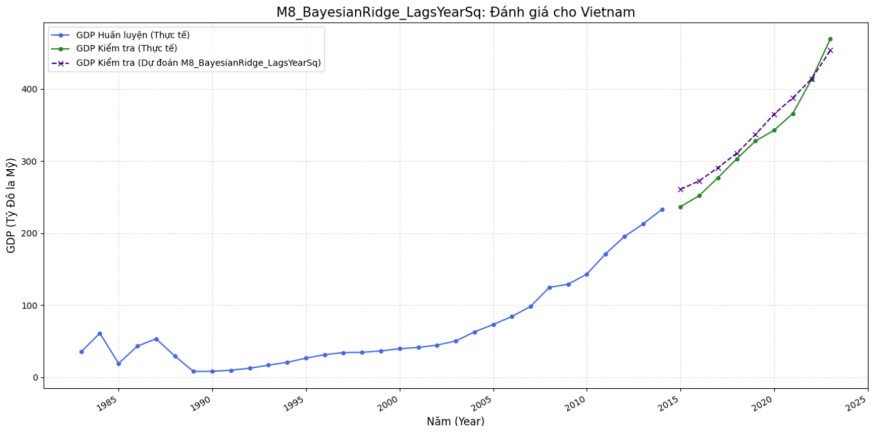
$$\text{GDP}_t = \beta_0 + \sum_{i=1}^{k} \beta_i \cdot \text{GDP}_{t-i} + \gamma_1 \cdot \text{Year}_t + \gamma_2 \cdot \text{Year}_t^2 + \varepsilon_t$$

Trong đó:

- GDP_t: GDP tại năm t,
- GDP_{t-i}: GDP ở các năm trước đó (các đặc trưng trễ),
- Year_t, Year_t^2: đặc trưng xu hướng thời gian,
- β_i, γ_j: hệ số hồi quy với phân phối xác suất theo Bayes,
- ε_t: sai số ngẫu nhiên với phương sai α^{-1},
- α: tham số điều khiển độ nhiễu (noise),
- λ: tham số điều khiển độ phức tạp mô hình (sự phân tán của hệ số).

Kết quả các chỉ số đánh giá:

- **R-squared (Test)**: 0.9448
- **MAE (Test)**: 15.04
- **RMSE (Test)**: 16.84
- **MAPE (Test)**: 4.94%



8. Mô hình N-BEATS (Neural Basis Expansion Analysis for Time Series Forecasting)

Mô hình N-BEATS (M4_1_NBEATS_Small):

N-BEATS (Neural Basis Expansion Analysis for Time Series) là một mô hình deep learning chuyên biệt cho dự báo chuỗi thời gian, sử dụng kiến trúc các khối nối tiếp nhau (stacked blocks). Trong mô hình này, chuỗi GDP theo năm của một quốc gia được dùng làm dữ liệu đầu vào và được chuẩn hóa bằng Darts Scaler.

Cơ chế học của mô hình dựa trên cấu trúc sliding window:

$$[\text{GDP}_{t-l+1}, \ldots, \text{GDP}_t] \longrightarrow \text{GDP}_{t+1}$$

Trong đó:

- l: độ dài chuỗi đầu vào (input chunk length),
- GDP_{t+1}: giá trị cần dự báo tại thời điểm tiếp theo,
- Mỗi khối trong N-BEATS học dự báo phần dư (residual forecast) từ đầu ra của khối trước đó.

Đầu ra cuối cùng là tổng các dự báo từ các khối:

$$\widehat{\text{GDP}}_{t+1} = \sum_{i=1}^{n} f_i([\text{GDP}_{t-l+1}, \ldots, \text{GDP}_t])$$

Trong đó f_i là khối dự báo thứ i, và n là tổng số khối. Mô hình này phù hợp với bài toán dự báo GDP dài hạn do khả năng học phi tuyến và không cần giả định cấu trúc chuỗi trước.

Kết quả các chỉ số đánh giá:

- **R-squared (Test)**: 0.7796
- **MAE (Test)**: 31.17
- **RMSE (Test)**: 36.36
- **MAPE (Test)**: 9.43%

Trong mô hình N-BEATS, việc tối ưu siêu tham số được thực hiện bằng công cụ **Optuna** nhằm cải thiện độ chính xác dự báo GDP. Các siêu tham số như số lượng khối (stacks, blocks), số lớp (layers), độ rộng lớp ẩn

(layer width), tốc độ học (learning rate), kích thước batch, và số epoch được lựa chọn tự động thông qua quá trình thử nghiệm nhiều tổ hợp. Tại mỗi vòng thử (trial), một mô hình N-BEATS được huấn luyện trên tập huấn luyện và đánh giá trên tập kiểm tra thông qua chỉ số sai số RMSE.
Trong mô hình của em đây là bộ tham số tối ưu mà **Optuna** tìm được :

- `num_stacks`: 3
- `num_blocks`: 2
- `num_layers`: 2
- `layer_widths`: 128
- `learning_rate`: 0.0007591104805282694
- `batch_size`: 8
- `n_epochs`: 57

Giá trị RMSE tốt nhất trên tập validation: 26.2026
Kết quả khi chạy bộ dữ liệu với tham số tối ưu :

- **R-squared (Test)**: 0.8856
- **MAE (Test)**: 23.45
- **RMSE (Test)**: 26.20
- **MAPE (Test)**: 7.71%

4.2 Có sử dụng mô hình tiên tiến trong 3 năm trở lại đây

Mô hình tiên tiến mà em sử dụng trong 3 năm trở lại gần đây là mô hình **N-BEATS** (Neural Basis Expansion Analysis for Time Series Forecasting)
N-BEATS (Neural Basis Expansion Analysis for Time Series) là một mô hình học sâu (deep learning) hiện đại được thiết kế đặc biệt để dự báo chuỗi thời gian. Ra đời vào năm 2020 bởi các nhà nghiên cứu tại Element AI và được công bố tại hội nghị ICLR, N-BEATS đã nhanh chóng trở thành một trong những mô hình mạnh mẽ nhất cho bài toán dự báo chuỗi thời gian một biến (univariate forecasting), vượt qua nhiều phương pháp truyền thống và học máy khác trong các bài kiểm thử chuẩn.
Chúng ta có thể kể đến một số thông tin và bài báo liên quan đến các cải tiến và ứng dụng của **N-BEATS** trong 3 năm gần đây:

- Lin, Hao, and Chundong Wang. "DIGWO-N-BEATS: An evolutionary time series prediction method for situation prediction." Information Sciences 664 (2024): 120316.
- Nayak, GH Harish, et al. "N-BEATS deep learning architecture for agricultural commodity price forecasting." Potato Research (2024): 1-21.

- Aiwansedo, Konstandinos, et al. "CNN-N-BEATS: Novel Hybrid Model for Time-Series Forecasting." International Conference on Deep Learning Theory and Applications. Cham: Springer Nature Switzerland, 2024.

Chương 5

ĐÓNG GÓI MÔ HÌNH

5.1 Có khả năng ứng dụng vào một ngữ cảnh cụ thể

Ứng dụng mô hình vào ngữ cảnh thực tế

Sau khi xây dựng và đánh giá các mô hình dự báo GDP, việc ứng dụng chúng vào thực tế là bước quan trọng nhằm phát huy giá trị của mô hình trong các ngữ cảnh nghiệp vụ cụ thể. Một trong những ứng dụng điển hình là trong lĩnh vực **hoạch định chính sách kinh tế và đầu tư tài chính**.

Ngữ cảnh ứng dụng cụ thể

Giả sử bạn là chuyên viên phân tích tại *Văn phòng Chính phủ* hoặc *Bộ Kế hoạch và Đầu tư*, việc có một hệ thống dự báo GDP đáng tin cậy là điều vô cùng quan trọng. Mỗi năm, Chính phủ cần đưa ra các quyết định về ngân sách, chi tiêu công, chính sách thuế và định hướng phát triển kinh tế quốc gia. Các quyết định này phụ thuộc phần lớn vào dự báo tăng trưởng GDP trong trung và dài hạn.

Hệ thống dự báo GDP được xây dựng trong đề tài này có thể được tích hợp vào quy trình phân tích dữ liệu tại các cơ quan quản lý, đóng vai trò như một công cụ hỗ trợ ra quyết định, giúp:

- Ước lượng quy mô nền kinh tế trong 3–5 năm tới.
- Phân tích tác động của các chính sách (ví dụ: kích cầu, cắt giảm thuế) đến tăng trưởng kinh tế.
- So sánh tốc độ phát triển giữa các quốc gia trong bối cảnh hội nhập hoặc đàm phán quốc tế.

Ngoài ra, mô hình còn có thể được sử dụng trong các tổ chức tài chính như *ngân hàng thương mại*, *quỹ đầu tư*, hoặc *tập đoàn đa quốc gia* để:

- Đánh giá rủi ro kinh tế vĩ mô khi ra quyết định đầu tư.
- Xác định thời điểm mở rộng thị trường hoặc thiết lập cơ sở sản xuất tại quốc gia đang phát triển.
- Dự báo lợi nhuận gián tiếp thông qua mức tăng GDP bình quân đầu người.

Người sử dụng mô hình

- **Chuyên gia kinh tế, nhà hoạch định chính sách:** sử dụng đầu ra mô hình để điều chỉnh chính sách phát triển.
- **Nhà đầu tư tài chính, chuyên viên phân tích rủi ro:** đánh giá xu hướng vĩ mô làm cơ sở cho các chiến lược đầu tư.
- **Chuyên viên dữ liệu tại cơ quan nhà nước:** tích hợp mô hình vào hệ thống dữ liệu kinh tế để khai thác tự động.

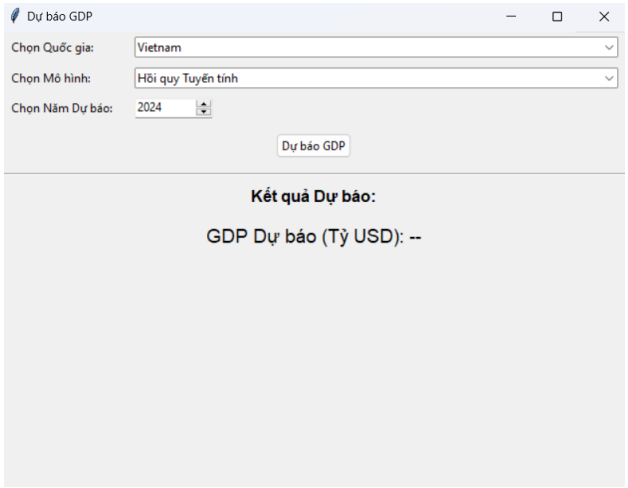
Lợi ích mang lại

- Giảm sự phụ thuộc vào các mô hình dự báo thủ công hoặc mang tính định tính.
- Tăng độ chính xác và khách quan trong dự báo nhờ áp dụng kỹ thuật học sâu.
- Cung cấp công cụ linh hoạt có thể mở rộng cho nhiều quốc gia khác nhau chỉ bằng cách thay đổi dữ liệu đầu vào.

5.2 Đóng gói giao diện demo chương trình

Xây dựng demo chương trình

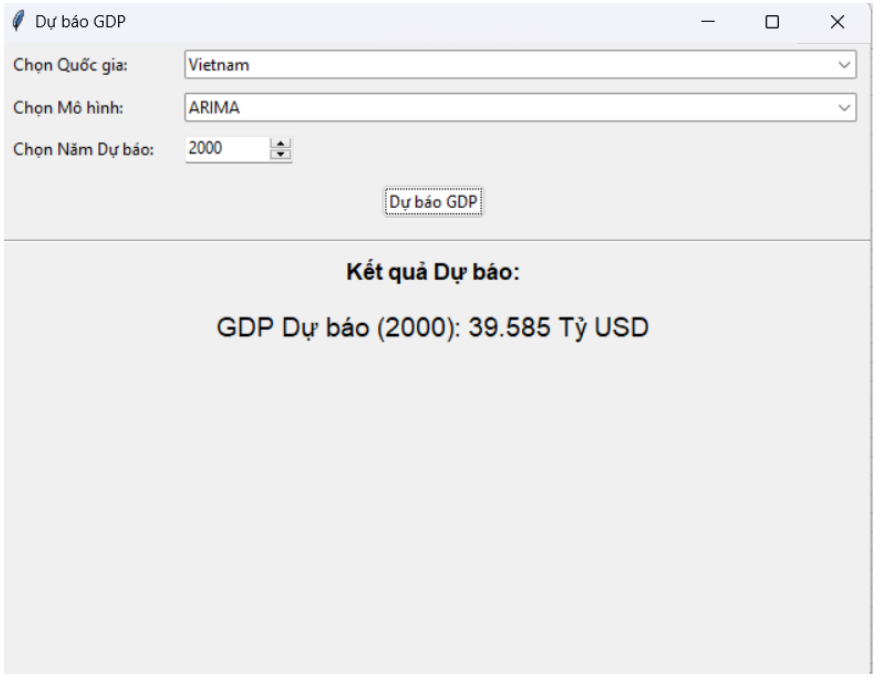
Công cụ sử dụng: ở đây em sử dụng python và các thư viện có sẵn như là sử dụng Tkinter cho giao diện, Pandas và NumPy cho xử lý dữ liệu, cùng một loạt các thư viện mạnh mẽ như Scikit-learn, Statsmodels, pmdarima, Darts ....để triển khai nhiều mô hình dự báo chuỗi thời gian khác nhau.
INPUT: Nhập tên quốc gia , nhập năm muốn dự báo và lựa chọn mô hình dự báo . Ở đây chúng ta sẽ có 8 sự lựa chọn mô hình dự báo chính là các mô hình mà em đã xây dựng bên trên .
OUTPUT: GDP được dự báo của quốc gia và năm chúng ta muốn dự báo.
Giao diện chương trình:



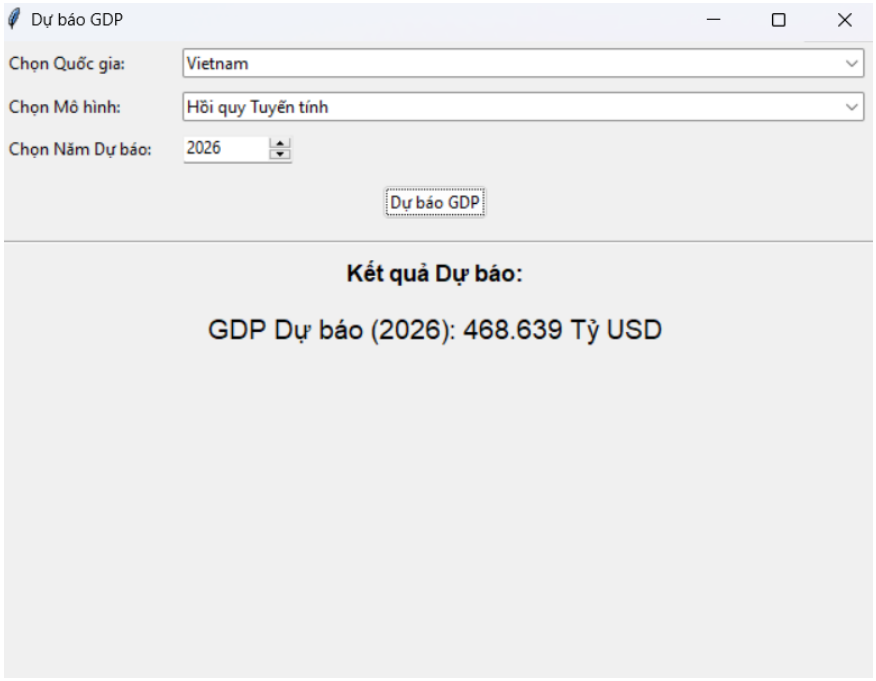
Cách xây dựng:

- Sử dụng **Tkinter** để xây dựng giao diện người dùng, cho phép chọn quốc gia, mô hình dự báo (*Linear Regression, ARIMA, Random Forest, SVR, KNN, ElasticNet, Bayesian Ridge, NBEATS*) và năm cần dự báo.
- Các thư viện như **Pandas** và **NumPy** để đọc dữ liệu từ tệp CSV, xử lý và chuyển đổi dữ liệu GDP về định dạng phù hợp cho việc huấn luyện mô hình.
- Các thư viện **Scikit-learn**, **Statsmodels**, **pmdarima** và **Darts**... để xây dựng và huấn luyện các mô hình dự báo. Mỗi mô hình được xây dựng là một hàm riêng
- Sử dụng cơ chế liên kết sự kiện trong Tkinter để khi người dùng nhấn nút "Dự báo", ứng dụng sẽ:
  - Gọi đúng hàm tương ứng với mô hình đã chọn.
  - Xử lý và đưa dữ liệu GDP của quốc gia tương ứng vào mô hình.
  - Hiển thị kết quả dự báo GDP ngay trên giao diện.

Dưới đây là 1 số ví dụ dự báo với các mô hình :



GDP Việt Nam năm 2000 với mô hình ARIMA



GDP Việt Nam năm 2026 với mô hình Hồi quy tuyến tính

Nhận xét và đánh giá

Ta thấy dữ liệu dự đoán cũng khá tương đối sát so với thực tế.Từ đó có thể ứng dụng trong 1 số ngữ cảnh cụ thể như :

- **Hoạch định chính sách kinh tế:** Dự báo GDP giúp chính phủ đánh giá xu hướng tăng trưởng, từ đó đưa ra các chính sách tài khóa, tiền tệ, đầu tư công phù hợp để ổn định và thúc đẩy nền kinh tế.
- **Quản lý ngân sách nhà nước:** Biết trước quy mô GDP giúp ước tính nguồn thu từ thuế, điều chỉnh chi tiêu công và cân đối ngân sách hiệu quả hơn.
- **Hỗ trợ doanh nghiệp ra quyết định:** Doanh nghiệp có thể dựa vào dự báo GDP để dự tính nhu cầu thị trường, mở rộng sản xuất, đầu tư, hoặc điều chỉnh kế hoạch tài chính.
- **Phân tích rủi ro và đầu tư tài chính:** Các tổ chức tài chính, ngân hàng và nhà đầu tư sử dụng dự báo GDP để đánh giá triển vọng tăng trưởng, từ đó xây dựng chiến lược đầu tư và quản lý rủi ro.

Tuy nhiên với bộ dữ liệu GDP chỉ từ năm 1980 đến 2023 có một số hạn chế nhất định. Thời gian hơn 40 năm là tương đối ngắn đối với các mô hình dự báo phức tạp, đặc biệt là các mô hình học sâu vốn yêu cầu lượng dữ liệu lớn. Dữ liệu không phản ánh đầy đủ các chu kỳ kinh tế dài hạn hay các biến động lớn trước 1980, làm giảm khả năng phát hiện xu hướng bền vững. Khi dự báo cho các năm xa trong tương lai, mô hình dễ gặp sai số cao do phải ngoại suy ngoài phạm vi dữ liệu huấn luyện. Ngoài ra, dữ liệu trong giai đoạn đầu (nhất là với các nước đang phát triển) có thể thiếu ổn định, ảnh hưởng đến độ tin cậy của mô hình.

Chương 6

Kết luận

Thông qua đề tài này, em đã có cơ hội tiếp cận và tìm hiểu sâu về các mô hình dự báo GDP, từ các phương pháp truyền thống như *Linear Regression*, *ARIMA* đến các mô hình học máy hiện đại như *Random Forest*, *SVR* và mô hình học sâu *N-BEATS*. Quá trình thực hiện giúp em hiểu rõ hơn về cách xử lý dữ liệu chuỗi thời gian, tiền xử lý dữ liệu kinh tế, cũng như áp dụng các tiêu chí đánh giá mô hình như R^2, MAE, RMSE, và MAPE để so sánh hiệu quả các phương pháp.

Kết quả cho thấy mô hình **Linear Regression** hoạt động hiệu quả nhất trong bài toán này, với độ chính xác cao và sai số thấp (MAPE chỉ 3.56%). Mô hình *Random Forest* và *SVR* cũng cho kết quả dự báo tốt, trong khi mô hình *N-BEATS* tuy hiện đại nhưng chưa đạt hiệu quả tối ưu do hạn chế về dữ liệu đầu vào và thời gian huấn luyện.

Ngoài ra, việc xây dựng giao diện demo bằng Python giúp em rèn luyện thêm kỹ năng lập trình và tăng tính ứng dụng thực tế cho mô hình. Qua đó, em nhận thấy mô hình có thể được áp dụng để hỗ trợ cho các cơ quan quản lý kinh tế, nhà đầu tư hoặc các tổ chức cần dự báo GDP phục vụ ra quyết định.

Tuy nhiên, do giới hạn về dữ liệu (chỉ từ năm 1980 đến 2023) và nguồn lực cá nhân, em chưa thể triển khai được các mô hình sâu hơn hoặc sử dụng thêm nhiều biến vĩ mô liên quan. Trong các nghiên cứu sau, em mong muốn mở rộng dữ liệu, bổ sung các yếu tố ảnh hưởng khác và thử nghiệm các mô hình tiên tiến như *LSTM* hoặc *Transformer* để cải thiện độ chính xác và khả năng dự báo dài hạn.

Qua đề tài này, em không chỉ củng cố kiến thức lý thuyết mà còn nâng cao được kỹ năng phân tích dữ liệu, mô hình hóa và tư duy phản biện – những năng lực quan trọng trong lĩnh vực khoa học dữ liệu và ra quyết định.

References

[1] Trần Ngọc Thăng, Slide Bài giảng Học Máy, Đại học Bách Khoa Hà Nội.

[2] Nayak, GH Harish, et al. "N-BEATS deep learning architecture for agricultural commodity price forecasting." Potato Research (2024): 1-21.

[3] Aiwansedo, Konstandinos, et al. "CNN-N-BEATS: Novel Hybrid Model for Time-Series Forecasting." International Conference on Deep Learning Theory and Applications. Cham: Springer Nature Switzerland, 2024.

[4] Lin, Hao, and Chundong Wang. "DIGWO-N-BEATS: An evolutionary time series prediction method for situation prediction." Information Sciences 664 (2024): 120316.

[5] Ramesh Sharda, Dursun Delen, Efraim Turban. Business Intelligence and Analytics-Symtems for Decision Support. Tenth Edition. Pearson Education. 2014.

[6] Phạm Văn Toàn, Bài toán dự đoán dựa trên mô hình hồi quy trong machine learning, Viblo, 2016.https://viblo.asia/p/bai-toan-du-doan-prediction-dua-tren-mo-hinh-hoi-quy-trong-machine-learning-YmjeoLgzkqa