ỨNG DỤNG MÔ HÌNH HỌC MÁY TRONG BÀI TOÁN XÁC ĐỊNH MỨC ĐỘ TÁC ĐỘNG CỦA LẠM PHÁT TRÊN NHÓM NGƯỜI TIÊU DÙNG

APPLICATION OF MACHINE LEARNING MODEL IN THE PROBLEM OF DETERMINING THE LEVEL OF IMPACT OF INFLATION ON CONSUMER GROUPS

Đặng Thị Hồng Hà[1], Vũ Việt Thắng[2,*], Nguyễn Như Quỳnh[3], Vương Thị Tuyên[1], Bùi Lê Hiền Mai[4]

DOI: https://doi.org/10.57001/huih5804.2026.037

TÓM TẮT

Bài báo nghiên cứu mức độ tác động của lạm phát đến hành vi tiêu dùng của người dân qua việc ứng dụng phân tích mối quan hệ giữa lạm phát và hành vi tiêu dùng của người dân, từ đó sử dụng mô hình Random Forest để đánh giá tác động của CPI lên chi tiêu trung bình hàng năm và xác định nhóm chịu ảnh hưởng lớn nhất từ CPI. Kết quả cho thấy nhóm thu nhập thấp ở nông thôn (Group 1) là đối tượng dễ tổn thương nhất do phụ thuộc vào hàng hóa thiết yếu, khả năng đa dạng hóa chi tiêu thấp, và thị trường kém cạnh tranh. Trong khi đó, nhóm thu nhập cao ở thành thị (Group 5) chịu tác động tương đối lớn do cơ cấu tiêu dùng đa dạng, chi phí sinh hoạt cao và hiệu ứng lạm phát kỳ vọng. Ngoài ra, nghiên cứu sử dụng mô hình ARIMA để dự báo CPI năm 2025, dựa trên dữ liệu CPI từ 2015 - 2024, kết quả giúp hiểu tác động của lạm phát và đồng thời dự báo xu hướng tăng nhẹ của CPI, phù hợp với bối cảnh kinh tế hiện tại.

Từ khóa: *Mô hình học máy; lạm phát; nhóm người tiêu dùng.*

ABSTRACT

The paper studies the impact of inflation on people's consumption behavior by applying the analysis of the relationship between inflation and people's consumption behavior, thereby using the Random Forest model to assess the impact of CPI on average annual expenditure, and identify the group most affected by CPI. The results show that the low-income group in rural areas (Group 1) is the most vulnerable due to dependence on essential goods, low ability to diversify expenditure, and less competitive market. Meanwhile, the high-income group in urban areas (Group 5) is relatively affected due to diverse consumption structure, high cost of living, and expected inflation effect. In addition, the study uses the ARIMA model to forecast CPI in 2025, based on CPI data from 2015 - 2024, the results help understand the impact of inflation and at the same time forecast a slight upward trend of CPI, in line with the current economic context.

Keywords: *Machine learning model; inflation; consumer groups.*

[1]Trường Kinh tế, Trường Đại học Công nghiệp Hà Nội
[2]Phòng Đào tạo, Trường Đại học Công nghiệp Hà Nội
[3]Sinh viên Trường Kinh tế, Trường Đại học Công nghiệp Hà Nội
[4]Ngân hàng TMCP Quân đội MB
[*]Email: vuvietthang@haui.edu.vn

Ngày nhận bài: 10/5/2025
Ngày nhận bài sửa sau phản biện: 20/8/2025
Ngày chấp nhận đăng: 26/02/2026

CHỮ VIẾT TẮT

CPI	Chỉ số giá tiêu dùng
ARIMA	Autoregressive Integrated Moving Average
ACF	Autocorrelation Function
PACF	Partial Autocorrelation Function
AIC	Akaike Information Criterion
BIC	Bayesian information criterion

1. GIỚI THIỆU

Trong bối cảnh kinh tế toàn cầu đã hứng chịu nhiều biến động bởi đại dịch, xung đột địa chính trị và những thay đổi trong chuỗi cung ứng, lạm phát đã và đang nổi lên như một vấn đề kinh tế vĩ mô quan trọng, tác động trực tiếp đến đời sống người dân và hành vi tiêu dùng của họ. Tại Việt Nam, những năm gần đây chứng kiến những đợt biến động giá cả hàng hóa, thực phẩm, năng lượng gây ảnh hưởng không nhỏ đến thu nhập thực tế và quyết định chi tiêu của người dân, đặc biệt là nhóm thu nhập thấp và trung bình. Vì vậy, chống lạm phát cũng là nhiệm vụ thường trực của các quốc gia.

Lạm phát cao đã có tác động đến nhiều mặt của đời sống kinh tế - xã hội, làm đảo lộn cuộc sống của dân chúng. Theo một nghiên cứu của bà Mai Thị Thanh Xuân, Khoa Kinh tế Chính trị, Trường Đại học Kinh tế, Đại học Quốc gia Hà Nội [1], lạm phát cao sẽ có ảnh hưởng trực tiếp đến nhóm đối tượng có thu nhập thấp nhất là những người sống chủ yếu dựa vào tiền lương như công nhân, viên chức, người về hưu, người hưởng trợ cấp xã hội khác, nông dân và những người kinh doanh nhỏ lẻ. Vì vậy, việc xác định mức độ tác động của lạm phát lên nền kinh tế nói chung và lên từng nhóm người tiêu dùng (xác định dựa vào thu nhập) nói riêng là việc làm rất quan trọng trong quá trình giảm bớt gánh nặng lạm phát cho người có thu nhập thấp và cải thiện mức sống của người dân Việt Nam.

Trong khuôn khổ nghiên cứu này, nhóm tác giả ứng dụng mô hình Random Forest nhằm xác định và phân tích mức độ tác động của lạm phát thông qua chỉ số giá tiêu dùng CPI đến 5 nhóm người tiêu dùng tại 2 khu vực Nông thôn và Thành thị ở Việt Nam, đồng thời xác định nhóm thu nhập nào bị ảnh hưởng nhiều nhất bởi lạm phát và sự khác biệt giữa các khu vực. Ngoài ra, nghiên cứu còn ứng dụng mô hình chuỗi thời gian ARIMA trong dự báo chỉ số giá tiêu dùng CPI trong 12 tháng của năm 2025. Kết quả nghiên cứu hướng tới cung cấp cơ sở dữ liệu và công cụ hỗ trợ quá trình ra quyết định cho các cơ quan quản lý, doanh nghiệp và tổ chức có liên quan trong bối cảnh kinh tế nhiều biến động.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Mô hình nghiên cứu

Để xác định mức độ của lạm phát trên nhóm người tiêu dùng, nghiên cứu sử dụng các mô hình học máy và phương pháp triển khai, tập trung vào hai mô hình nổi bật: ARIMA và Random Forest.

Mô hình ARIMA

Mô hình ARIMA (Autoregressive Integrated Moving Average) là một trong những công cụ phổ biến nhất trong phân tích chuỗi thời gian, được phát triển bởi Box và Jenkins [2]. Mô hình ARIMA được sử dụng để mô hình hóa và dự báo các hiện tượng kinh tế có tính liên tục theo thời gian như lạm phát, GDP, tỷ giá, hay chỉ số giá tiêu dùng (CPI). Mô hình ARIMA kết hợp ba thành phần chính để mô tả và dự báo các chuỗi thời gian. Mô hình ARIMA được ký hiệu là ARIMA(p,d,q) [3], trong đó: p: bậc của phần tự hồi quy (AR), d: bậc của sai phân để làm cho chuỗi dừng (I), q: bậc của phần trung bình trượt (MA).

Mô hình ARIMA là một trong những mô hình phổ biến trong phân tích chuỗi thời gian, được sử dụng để dự báo các chỉ số kinh tế theo xu hướng quá khứ. Trong nghiên cứu này, mô hình ARIMA tự động được áp dụng để dự báo chỉ số giá tiêu dùng (CPI) một chỉ số quan trọng phản ánh mức độ lạm phát. Mô hình ARIMA được lựa chọn vì khả năng tự động điều chỉnh các tham số tối ưu (p, d, q) mà không cần can thiệp thủ công. Cụ thể, mô hình này sử dụng thông tin từ các chuỗi thời gian trước đó để dự báo giá trị của CPI trong tương lai, từ đó cung cấp các dự báo về xu hướng lạm phát. ARIMA là lựa chọn thích hợp khi chuỗi dữ liệu có tính dừng và không có xu hướng thay đổi đột ngột

Chuỗi thời gian ARIMA(p,d,q) là chuỗi thời gian trung bình trượt kết hợp với tự hồi quy, với p biểu thị bậc tự hồi quy, d biểu thị số lần chuỗi thời gian được tính sai phân cho đến khi có tính dừng va q là bậc trung bình trượt. Mô hình ARIMA(p,d,q) tổng quát sẽ có dạng như sau:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t \quad (1)$$

- Y_t: là giá trị chuỗi thời gian tại thời điểm t.

- c: là hằng số.

- ϕ_i: là các hệ số tự hồi quy.

- ϵ_t: là sai số ngẫu nhiên (nhiễu trắng) tại thời điểm t.

Box - Jenkin đề xuất sử dụng mô hình ARIMA để dự báo chuỗi thời gian dừng, gồm 4 bước cơ bản: nhận dạng mô hình thử nghiệm, ước lượng, kiểm định bằng chẩn đoán và dự báo [4].

Mô hình Random Forest

Random Forest là một thuật toán học máy thuộc nhóm học có giám sát (supervised learning), dựa trên kỹ thuật ensemble learning (học tổ hợp), kết hợp nhiều mô hình con để cải thiện độ chính xác và độ ổn định. Random Forest xây dựng nhiều cây quyết định (decision trees) trên các tập con ngẫu nhiên của dữ liệu huấn luyện và tổng hợp kết quả dự đoán từ các cây đó thông qua biểu quyết đa số (với bài toán phân loại) hoặc trung bình (với bài toán hồi quy). Việc sử dụng nhiều cây cùng lúc không chỉ tăng hiệu suất mà còn làm giảm rủi ro overfitting - một vấn đề

phổ biến khi sử dụng mô hình cây đơn lẻ. Thuật toán Random Forest được đề xuất bởi Leo Breiman vào năm 2001 [5] và kể từ đó trở thành một trong những thuật toán mạnh mẽ và phổ biến nhất trong học máy, được áp dụng rộng rãi trong nhiều lĩnh vực như tài chính, y tế, marketing, và khoa học dữ liệu. Mô hình này sử dụng kỹ thuật đóng gói (bagging) [6] cho phép lựa chọn một nhóm nhỏ các thuộc tính tại mỗi nút (node) của cây phân lớp để phân chia thành các mức tiếp theo.

Quy trình xây dựng và dự đoán của Random Forest bao gồm:

- *Tạo các tập con:* Từ tập dữ liệu gốc, tạo ra n tập con ngẫu nhiên bằng phương pháp lấy mẫu có hoàn lại (bootstrap sampling).

- *Huấn luyện từng cây:* Trên mỗi tập con dữ liệu này, một cây quyết định được huấn luyện. Trong quá trình huấn luyện từng cây, tại mỗi nút chia, chỉ một tập con ngẫu nhiên các đặc trưng được lựa chọn để tìm điểm chia tốt nhất.

- *Tổng hợp dự đoán:* Đối với bài toán phân loại, khi dự đoán cho một quan sát mới, tất cả các cây trong rừng sẽ đưa ra một "phiếu bầu" cho một lớp nhất định. Kết quả cuối cùng là lớp nhận được biểu quyết đa số (majority voting). Đối với bài toán hồi quy, mỗi cây sẽ đưa ra một giá trị dự đoán. Random Forest sẽ tính giá trị trung bình của tất cả các dự đoán này để đưa ra kết quả cuối cùng.

Ưu điểm của mô hình Random Forest là khả năng xử lý được các giá trị ngoại lai và các nhiễu [7]. Ngoài phân lớp hay dự báo, rừng ngẫu nhiên có thể được xác định được tầm quan trọng của các biến trong mô hình. Điều này giúp đưa ra các yếu tố quyết định trong việc phân lớp hay dự báo [8].

2.2. Dữ liệu nghiên cứu

Trong nghiên cứu này, để khảo sát mức sống của người dân sống tại nông thôn và thành thị, nhóm tác giả tiến hành lấy dữ liệu từ năm 2015 đến năm 2024 của Tổng cục Thống kê (bảng 1).

Bảng 1. Khảo sát mức sống của người dân tại nông thôn và thành thị giai đoạn 2015 - 2024

Year	Region	CPI	Group	Income	Spending
2015	Countryside	0,63	1	649,6	564,9
2016	Countryside	2,66	1	696	659,8
2017	Countryside	3,53	1	767,4	667,3
2018	Countryside	3,54	1	819,4	770,6
2029	Countryside	2,79	1	915,9	796,4
2020	Countryside	3,23	1	950,8	826,8
2021	Countryside	1,84	1	1075,5	935,2
2022	Countryside	3,15	1	1081,1	940
2023	Countryside	3,25	1	1297,9	1128,6
2024	Countryside	3,63	1	1405,1	1221,8
2015	City	0,63	1	1457,1	1267
2016	City	2,66	1	1544,2	1352,4
2017	City	3,53	1	1669,8	1452
2018	City	3,54	1	1946,5	1651,1
2029	City	2,79	1	2071,8	1801,6
2020	City	3,23	1	2119,2	1842,8
2021	City	1,84	1	2408,9	2094,7
2022	City	3,15	1	2358,9	2051
2023	City	3,25	1	2415,7	2100,6
2024	City	3,63	1	2596,6	2257,9

(Nguồn: Tổng cục Thống kê)

Dữ liệu được thu thập gồm các biến ở bảng 2.

Bảng 2. Mô tả dữ liệu nghiên cứu

Chỉ tiêu	Ý nghĩa	Kiểu dữ liệu
Year	Năm (10 năm tính từ 2015 - 2024)	Số nguyên (int)
Region	Vùng miền - Countryside: Nông thôn - City: Thành thị	Chuỗi (string), dạng phân loại (categorical)
CPI	Tỷ lệ lạm phát (Thể hiện qua chỉ số CPI)	Số thực (float)
Group	5 nhóm thu nhập từ thấp đến cao	Số nguyên (int)
Income	Thu nhập (Đơn vị: Nghìn đồng)	Số thực (float)
Spending	Chi tiêu (Đơn vị: Nghìn đồng)	Số thực (float)

(Nguồn: Tổng cục Thống kê)

Tập dữ liệu có 100 quan sát, mỗi quan sát đại diện cho một năm, khu vực, và nhóm thu nhập. Dữ liệu thu thập ở 2 khu vực: "Nông thôn" (50 quan sát), "Thành thị" (50 quan sát), trong đó có 5 nhóm thu nhập, mỗi nhóm có 20 quan sát. Ngoài ra, nghiên cứu sử dụng dữ liệu chỉ số giá tiêu dùng (CPI) của Việt Nam trong giai đoạn 10 năm kể từ năm 2015 đến hết năm 2024. Dữ liệu được thu thập gồm 2 biến là biến thời gian (Date) trong 10 năm kể từ tháng 01/2015 đến hết tháng 12/2024 (120 quan sát) tạo điều kiện cho việc phân tích xu hướng biến động và tác động theo thời gian. Dữ liệu được phân tích để đánh giá tác động của lạm phát "CPI" đến mức chi tiêu "Spending" của các hộ gia đình tại khu vực nông thôn và thành thị, đồng thời để kiểm tra sự khác biệt trong hành vi chi tiêu giữa các nhóm thu nhập (Group).

Bảng 3. Dữ liệu CPI của Việt Nam giai đoạn 2015 - 2024

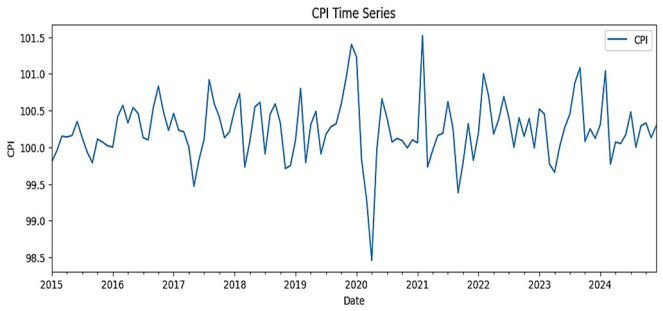
Date	01/2015	02/2015	03/2015	04/2015	05/2015	06/2015	07/2015	08/2015	09/2015	10/2015	11/2015	12/2015
CPI	0.998	0.9995	1.0015	1.0014	1.0016	1.0035	1.0013	0.9993	0.9979	1.0011	1.0007	1.0002
Date	01/2016	02/2016	03/2016	04/2016	05/2016	06/2016	07/2016	08/2016	09/2016	10/2016	11/2016	12/2016
CPI	1	1.0042	1.0057	1.0033	1.0054	1.0046	1.0013	1.001	1.0054	1.0083	1.0048	1.0023
Date	01/2017	02/2017	03/2017	04/2017	05/2017	06/2017	07/2017	08/2017	09/2017	10/2017	11/2017	12/2017
CPI	1.0046	1.0023	1.0021	1	0.9947	0.9983	1.0011	1.0092	1.0059	1.0041	1.0013	1.0021
Date	01/2018	02/2018	03/2018	04/2018	05/2018	06/2018	07/2018	08/2018	09/2018	10/2018	11/2018	12/2018
CPI	1.0051	1.0073	0.9973	1.0008	1.0055	1.0061	0.9991	1.0045	1.0059	1.0033	0.9971	0.9975
Date	01/2019	02/2019	03/2019	04/2019	05/2019	06/2019	07/2019	08/2019	09/2019	10/2019	11/2019	12/2019
CPI	1.001	1.008	0.9979	1.0031	1.0049	0.9991	1.0018	1.0028	1.0032	1.0059	1.0096	1.014
Date	01/2020	02/2020	03/2020	04/2020	05/2020	06/2020	07/2020	08/2020	09/2020	10/2020	11/2020	12/2020
CPI	1.0123	0.9983	0.9928	0.9846	0.9997	1.0066	1.004	1.0007	1.0012	1.0009	0.9999	1.001
Date	01/2021	02/2021	03/2021	04/2021	05/2021	06/2021	07/2021	08/2021	09/2021	10/2021	11/2021	12/2021
CPI	1.0006	1.0152	0.9973	0.9996	1.0016	1.0019	1.0062	1.0025	0.9938	0.998	1.0032	0.9982
Date	01/2022	02/2022	03/2022	04/2022	05/2022	06/2022	07/2022	08/2022	09/2022	10/2022	11/2022	12/2022
CPI	1.0019	1.01	1.007	1.0018	1.0038	1.0069	1.004	1	1.004	1.0015	1.0039	0.9999
Date	01/2023	02/2023	03/2023	04/2023	05/2023	06/2023	07/2023	08/2023	09/2023	10/2023	11/2023	12/2023
CPI	1.0052	1.0045	0.9977	0.9966	1.0001	1.0027	1.0045	1.0088	1.0108	1.0008	1.0025	1.0012
Date	01/2024	02/2024	03/2024	04/2024	05/2024	06/2024	07/2024	08/2024	09/2024	10/2024	11/2024	12/2024
CPI	1.0031	1.0104	0.9977	1.0007	1.0005	1.0017	1.0048	1	1.0029	1.0033	1.0013	1.0029

(Nguồn: Tổng cục Thống kê)

3. KẾT QUẢ NGHIÊN CỨU VÀ THẢO LUẬN

3.1. Kết quả mô hình ARIMA dự báo CPI

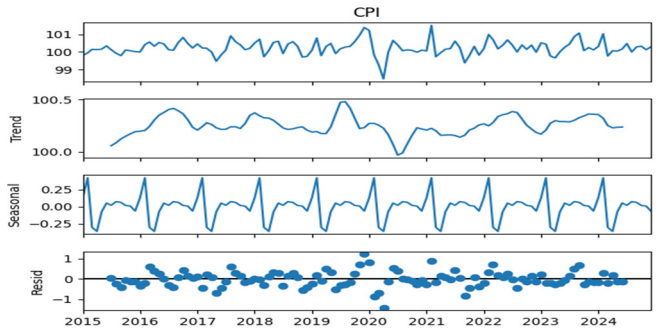
Nhóm tác giả tiến hành trực quan hóa dữ liệu nhằm phân tích xu hướng tổng thể chỉ số CPI và thu được kết quả như hình 1.



Hình 1. Xu hướng tổng thể chỉ số CPI giai đoạn 2015 - 2024

Biểu đồ chuỗi thời gian của chỉ số giá tiêu dùng (CPI) giai đoạn 2015 - 2024 cho thấy, CPI dao động quanh mốc 100, phản ánh mức giá tiêu dùng tương đối ổn định trong dài hạn. Tuy nhiên, biểu đồ không cho thấy xu hướng tăng hoặc giảm rõ rệt trong dài hạn mà chủ yếu thể hiện các dao động ngắn hạn. Đáng chú ý, vào đầu năm 2020, CPI ghi nhận một sự sụt giảm mạnh xuống gần mức 98,5%, giai đoạn này cũng trùng với giai đoạn bùng phát đại dịch COVID-19, khi các hoạt động tiêu dùng, sản xuất và lưu thông hàng hóa bị gián đoạn nghiêm trọng. Ngay sau đó, từ cuối năm 2020 đến đầu năm 2021, CPI bật tăng trở lại nhanh chóng, đạt đỉnh trên 101,5%, phản ánh quá trình phục hồi kinh tế và nhu cầu tiêu dùng tăng mạnh sau các đợt giãn cách. Trong giai đoạn 2022 - 2024, biểu đồ tiếp tục cho thấy các dao động có tính chu kỳ với các đỉnh và đáy xen kẽ, cho thấy sự bất ổn ngắn hạn về giá cả có thể bắt nguồn từ các biến động về nguồn cung (như giá xăng dầu, lương thực), chính sách điều hành giá trong nước hoặc các yếu tố kinh tế quốc tế. Từ những quan sát trên, có thể kết luận rằng CPI trong giai đoạn này nhìn chung duy trì ổn định, song vẫn chịu tác động đáng kể từ các cú sốc kinh tế lớn, đặc biệt là đại dịch COVID-19. Các dao động ngắn hạn và tính chu kỳ của dữ liệu gợi ý rằng các mô hình dự báo có khả năng nắm bắt yếu tố mùa vụ như SARIMA có thể phù hợp hơn trong việc dự báo chỉ số CPI.

- Phân tích xu hướng mùa vụ của CPI



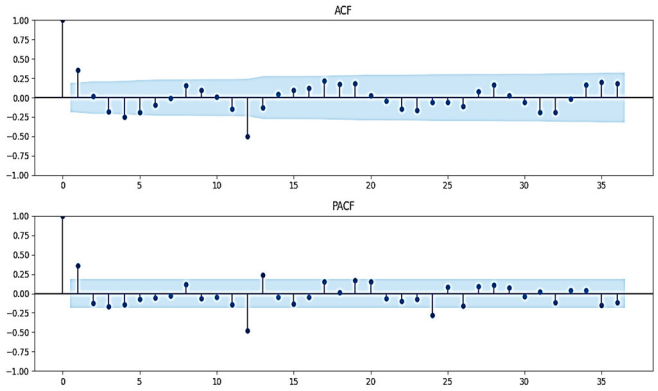
Hình 2. Xu hướng mùa vụ của CPI giai đoạn 2015 - 2024

Kết quả xu hướng mùa vụ của CPI từ năm 2015 đến 2024 cho thấy cấu trúc chuỗi thời gian bao gồm ba thành phần chính: xu hướng (trend), mùa vụ (seasonal), và phần dư (residual). Thành phần xu hướng cho thấy chỉ số CPI nhìn chung duy trì ổn định trong dài hạn, với mức tăng nhẹ giai đoạn 2015 - 2018, sau đó dao động quanh mốc 100,3 đến 100,5 trong suốt thời gian còn lại. Thành phần mùa vụ biểu hiện rõ rệt các chu kỳ lặp lại hàng năm với biên độ khoảng ±0,25, phản ánh đặc điểm tiêu dùng theo mùa, chẳng hạn như dịp Tết và các kỳ cao điểm tiêu dùng trong năm. Cuối cùng, phần dư dao động quanh giá trị trung bình bằng 0 và không cho thấy cấu trúc rõ rệt, ngoại trừ một số điểm ngoại lệ trong thời kỳ khủng hoảng, cho thấy quá trình tách xu hướng và mùa vụ đã được thực hiện hiệu quả. Kết quả này cho thấy dữ liệu CPI có đặc tính mùa vụ mạnh mẽ và xu hướng tương đối ổn định.

- Kết quả kiểm định ACF và PACF

Biểu đồ ACF (Autocorrelation Function) và PACF (Partial Autocorrelation Function) của chuỗi dữ liệu CPI sau khi loại bỏ xu hướng và mùa vụ cung cấp thông tin

quan trọng cho việc xác định cấu trúc mô hình ARIMA. Biểu đồ ACF cho thấy hệ số tự tương quan giảm dần một cách chậm rãi và dao động quanh mức 0, với một số giá trị tại các độ trễ 1, 2, 5 và 13 vượt khỏi khoảng tin cậy 95%, cho thấy chuỗi vẫn còn phần tự tương quan đáng kể tại một vài độ trễ. Trong khi đó, biểu đồ PACF thể hiện sự suy giảm đột ngột sau độ trễ 1 và 2, với các giá trị đáng kể tại các độ trễ 1, 2, 13, 14 và 25, điều này gợi ý chuỗi có thể được mô hình hóa tốt với một phần AR ở bậc thấp. Ngoài ra, do chuỗi CPI ban đầu có yếu tố mùa vụ rõ rệt như đã phân tích trước đó, mô hình SARIMA có thể phù hợp hơn với s = 12 để phản ánh chu kỳ mùa vụ hàng năm.



Hình 3. Kết quả kiểm định ACF và PACF dữ liệu CPI

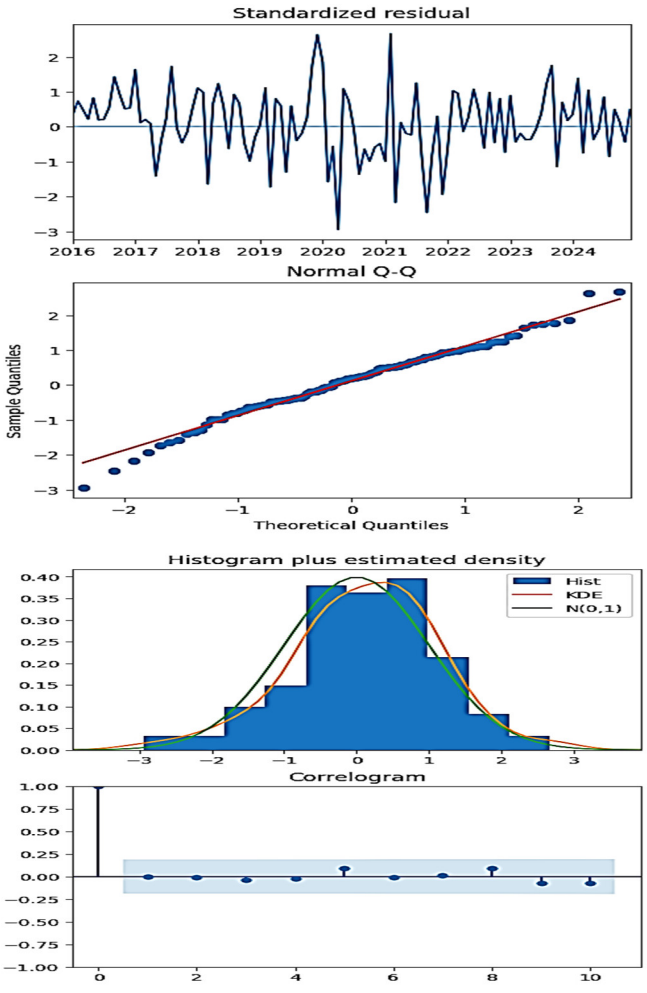
- Kết quả mô hình SARIMA

```
Best model:  ARIMA(2,0,1)(0,1,2)[12]
Total fit time: 73.902 seconds
                                     SARIMAX Results
==========================================================================================
Dep. Variable:                                 y   No. Observations:                  120
Model:             SARIMAX(2, 0, 1)x(0, 1, [1, 2], 12)   Log Likelihood                -56.077
Date:                                 Thu, 22 May 2025   AIC                         124.155
Time:                                         14:02:38   BIC                         140.247
Sample:                                     01-01-2015   HQIC                        130.680
                                          - 12-01-2024
Covariance Type:                                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          1.1968      0.119     10.057      0.000       0.964       1.430
ar.L2         -0.4486      0.071     -6.354      0.000      -0.587      -0.310
ma.L1         -0.8698      0.126     -6.889      0.000      -1.117      -0.622
ma.S.L12      -0.9727      0.173     -5.608      0.000      -1.313      -0.633
ma.S.L24       0.1059      0.117      0.906      0.365      -0.123       0.335
sigma2         0.1393      0.023      6.064      0.000       0.094       0.184
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                 2.85
Prob(Q):                              1.00   Prob(JB):                         0.24
Heteroskedasticity (H):               0.84   Skew:                            -0.30
Prob(H) (two-sided):                  0.62   Kurtosis:                         3.53
===================================================================================
```

Hình 4. Mô hình SARIMA

Kết quả (hình 4) thể hiện đầu ra từ mô hình SARIMA(2,0,1)(0,1,2) [12]. Đây là mô hình được lựa chọn tối ưu với tiêu chí AIC = 124,155, BIC = 140,247 và HQIC = 130,680. Những chỉ số này tương đối thấp, hàm ý mô hình có độ phù hợp tốt so với các mô hình thay thế. Mô hình bao gồm 2 bậc AR (AR.L1 và AR.L2), 1 bậc MA (MA.L1), và phần mùa vụ là (0,1,2)[12], trong đó có 2 hệ số MA mùa vụ tại độ trễ 12 và 24 (ma.S.L12 và ma.S.L24).

Phần kiểm định chẩn đoán mô hình cho thấy không có dấu hiệu vi phạm giả định. Kiểm định Ljung-Box (Q) với p-value = 1,00 chỉ ra rằng phần dư không có tự tương quan, tức là mô hình đã xử lý tốt cấu trúc chuỗi. Kiểm định Heteroskedasticity (H) với p-value = 0,62 cho thấy không có hiện tượng phương sai sai số thay đổi. Ngoài ra, kiểm định Jarque-Bera (p = 0,24) chỉ ra rằng phần dư gần như tuân theo phân phối chuẩn, với độ lệch Skew = -0,30 và độ nhọn Kurtosis = 3,53 nằm gần mức kỳ vọng của phân phối chuẩn (Skew = 0, Kurtosis = 3). Như vậy, mô hình SARIMA(2,0,1)(0,1,2))[12] không những phù hợp về mặt thống kê mà còn thỏa mãn các giả định quan trọng về phần dư, là một công cụ dự báo hiệu quả cho chuỗi dữ liệu CPI có yếu tố mùa vụ.

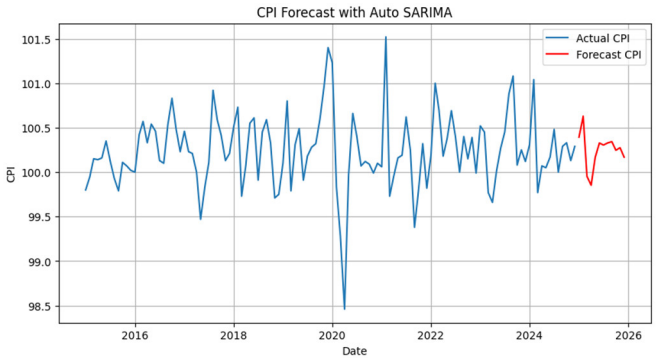


Hình 5. Biểu đồ chẩn đoán phần dư

Các biểu đồ chẩn đoán phần dư cho thấy mô hình SARIMA được xây dựng là phù hợp và đáng tin cậy (hình 5). Biểu đồ phần dư chuẩn hóa (Standardized Residual) cho thấy phần dư dao động ngẫu nhiên quanh giá trị trung bình bằng 0, không xuất hiện xu hướng hay chu kỳ rõ rệt, chứng tỏ không có cấu trúc chưa được mô hình hóa trong dữ liệu. Biểu đồ histogram kết hợp với đường mật độ ước lượng (KDE) và phân phối chuẩn chuẩn hóa (N(0,1)) cho thấy phần dư phân bố gần giống phân phối chuẩn, chỉ có sai lệch nhẹ ở hai đuôi. Q-Q Plot cũng xác

nhận điều này khi hầu hết các điểm nằm gần đường chéo, ngoại trừ một số điểm ở hai đầu. Cuối cùng, biểu đồ tương quan tự động phần dư (Correlogram) cho thấy các hệ số tương quan đều nằm trong khoảng tin cậy 95%, cho thấy phần dư không còn tự tương quan đáng kể. Tổng thể, các kết quả này cho thấy mô hình SARIMA đã loại bỏ được hầu hết cấu trúc phụ thuộc trong chuỗi thời gian gốc, phần dư đạt được các giả định cần thiết, và do đó, mô hình hoàn toàn có thể sử dụng để dự báo trong tương lai.

- Kết quả thực tế và dự báo CPI



Hình 6. CPI thực tế và dự báo

Biểu đồ hình 6 thể hiện kết quả dự báo chỉ số giá tiêu dùng (CPI) bằng mô hình SARIMA, trong đó đường màu xanh biểu diễn giá trị CPI thực tế từ năm 2015 đến năm 2024, còn đường màu đỏ thể hiện giá trị CPI dự báo cho năm 2025. Quan sát cho thấy mô hình SARIMA đã tái hiện khá tốt xu hướng dao động của dữ liệu thực tế với tính chất chu kỳ rõ rệt. Trong giai đoạn dự báo, mô hình cho thấy CPI duy trì ổn định quanh mức 100, không có biến động lớn, phản ánh kỳ vọng lạm phát được kiểm soát trong năm tới. Điều này phù hợp với xu hướng CPI thực tế trong các năm gần đây khi chỉ số này dao động trong biên độ hẹp. Tuy nhiên, do dữ liệu đầu vào có nhiều biến động bất thường (ví dụ như đột biến vào năm 2020), kết quả dự báo nên được sử dụng cẩn trọng và kết hợp với các yếu tố vĩ mô khác để đưa ra quyết định chính sách phù hợp. Tổng thể, mô hình SARIMA cho thấy khả năng dự báo ngắn hạn khá ổn định và có thể là công cụ hữu ích hỗ trợ phân tích xu hướng lạm phát.

3.2. Kết quả mô hình Random Forest

Kết quả hiệu suất mô hình

Mô hình Random Forest đã được huấn luyện thành công với dữ liệu mẫu gồm 100 quan sát và 6 đặc trưng đầu vào. Sau quá trình tìm kiếm tối ưu, mô hình cho thấy khả năng học tốt trên tập huấn luyện nhưng hiệu suất giảm đáng kể trên tập kiểm tra, phản ánh khả năng overfitting, do số lượng dữ liệu còn hạn chế, thể hiện trong hình 7.

```
KẾT QUẢ HIỆU SUẤT MÔ HÌNH
==========================
RMSE - Train:  0.0077
RMSE - Test:   0.0224
R² - Train:    0.9260
R² - Test:     0.5462
MAE - Train:   0.0052
MAE - Test:    0.0133

Hyperparameter tốt nhất:
max_depth: 10
min_samples_leaf: 1
min_samples_split: 5
n_estimators: 100
```

Hình 7. Kết quả hiệu suất mô hình

Phân tích tác động theo nhóm thu nhập

Nghiên cứu sử dụng mô hình Random Forest để phân tích tác động theo nhóm thu nhập, kết quả thu được ở hình 8.

```
PHÂN TÍCH TÁC ĐỘNG THEO NHÓM THU NHẬP
================================================
Bảng phân tích chi tiết:
       Impact_Mean  Impact_Std  Impact_Min  Impact_Max  CPI_Mean  Income_Mean
Group
1           0.8763      0.0238      0.8482      0.9480     2.825     1512.370
2           0.8429      0.0182      0.7836      0.8597     2.825     2788.920
3           0.8294      0.0082      0.8248      0.8555     2.825     4035.610
4           0.8133      0.0202      0.8063      0.8930     2.825     5715.615
5           0.8002      0.0223      0.7873      0.8555     2.825    11538.630

       Spending_Mean  Impact_Rank
Group
1           1319.125          1.0
2           2353.095          2.0
3           3345.315          3.0
4           4647.260          4.0
5           9249.470          5.0
```

Hình 8. Phân tích tác động theo nhóm thu nhập

```
PHÂN TÍCH TÁC ĐỘNG THEO VÙNG MIỀN
================================================================
Tỷ lệ tác động trung bình theo vùng và nhóm:
                  Impact_Ratio     CPI     Income
Region      Group
city        1           0.8680   2.825    2058.87
            2           0.8420   2.825    3635.75
            3           0.8297   2.825    5030.15
            4           0.8171   2.825    7073.60
            5           0.8062   2.825   14360.41
countryside 1           0.8845   2.825     965.87
            2           0.8438   2.825    1942.09
            3           0.8292   2.825    3041.07
            4           0.8095   2.825    4357.63
            5           0.7941   2.825    8716.85
```

Hình 9. Phân tích tác động theo vùng miền

Kết quả cho thấy, nhóm thu nhập thấp hơn (Nhóm 1) có mức tác động trung bình (Impact_Mean) cao nhất và mức tác động này giảm dần khi thu nhập tăng lên (đến Nhóm 5). Điều này được hỗ trợ bởi Impact_Rank nếu hạng 1,0 biểu thị mức độ tác động lớn nhất. Mặc dù chỉ số CPI trung bình là như nhau cho tất cả các nhóm, thu nhập trung bình và chi tiêu trung bình tăng đáng kể từ nhóm thu nhập thấp nhất đến nhóm cao nhất. Độ biến động của tác động (Impact_Std) có vẻ cao nhất ở hai cực của dải thu nhập (Nhóm 1 và Nhóm 5) và thấp nhất ở nhóm giữa (Nhóm 3).

Kết quả phân tích tác động theo vùng miền, tác động cao nhất luôn nằm ở Nhóm 1, thấp nhất ở Nhóm 5: Dù là thành thị hay nông thôn, nhóm thu nhập thấp nhất (Group 1) đều có mức Impact_Ratio cao nhất. Nhóm thu nhập cao nhất (Group 5) có Impact_Ratio thấp nhất, kết quả phù hợp với phân tích trước: nhóm thu nhập cao ít bị ảnh hưởng bởi lạm phát hơn, kết quả thể hiện ở hình 9.

So sánh giữa thành thị và nông thôn:

Bảng 4 cho thấy, nhóm thu nhập thấp ở nông thôn có mức bị tác động bởi lạm phát cao hơn nhóm cùng mức thu nhập ở thành thị. Ngược lại, các nhóm thu nhập trung bình và cao ở thành thị lại có mức tác động cao hơn so với cùng nhóm ở nông thôn.

Bảng 4. So sánh giữa thành thị và nông thôn

So sánh theo nhóm	Impact_Ratio (countryside)	Impact_Ratio (city)	Nhận xét
Group 1	**0,8845**	0,8680	**Nông thôn bị ảnh hưởng cao hơn**
Group 2	0,8438	0,8420	Gần như tương đương
Group 3	0,8228	0,8297	**Thành thị cao hơn**
Group 4	0,8095	0,8171	**Thành thị cao hơn**
Group 5	**0,7941**	0,8062	**Thành thị cao hơn**

(Nguồn: Nhóm tác giả tổng hợp)

Đối với nhóm thu nhập thấp (Group 1): Ở nông thôn, thu nhập rất thấp (chỉ ~965,87), nên chi phí sinh hoạt tăng lên (do lạm phát) dễ làm ảnh hưởng mạnh đến mức chi tiêu. Hạn chế trong tiếp cận các chính sách bảo vệ giá cả hoặc hàng hóa thiết yếu.

Đối với nhóm thu nhập cao (Group 5): Ở thành thị, mặc dù thu nhập cao, nhưng cũng có thể phải đối mặt với mức giá tiêu dùng cao hơn, dẫn đến tỷ lệ bị ảnh hưởng (Impact_Ratio) cao hơn so với nhóm thu nhập cao ở nông thôn.

Sự khác biệt giữa vùng miền và nhóm: Dữ liệu cho thấy không chỉ thu nhập, mà bối cảnh sống (region) cũng ảnh hưởng đến việc hộ gia đình bị tác động bởi lạm phát.

4. KẾT LUẬN VÀ KHUYẾN NGHỊ

Mô hình ARIMA, được triển khai thông qua phương pháp Auto ARIMA, thể hiện khả năng dự báo xu hướng CPI trong dài hạn, bao gồm cả giai đoạn từ 2015 đến 2026. Dự báo CPI cho giai đoạn 2024 - 2026, đặc biệt trong quý 1 năm 2025, cho thấy xu hướng tăng nhẹ từ mức khoảng 100 lên xấp xỉ 101. Điều này phản ánh sự ổn định tương đối của CPI trong bối cảnh kinh tế hiện tại, phù hợp với các biến động lịch sử được mô hình hóa. Tuy nhiên, độ chính xác của dự báo cần được kiểm chứng thêm với dữ liệu thực tế trong tương lai.

Kết quả từ mô hình Random Forest cho thấy tiềm năng lớn trong việc phân tích tác động của lạm phát theo vùng miền và nhóm thu nhập. Phân tích cho thấy lạm phát có tác động không đồng đều đến các nhóm thu nhập và khu vực địa lý. Mô hình giúp làm nổi bật các yếu tố đặc trưng ảnh hưởng đến hành vi tiêu dùng, đồng thời chỉ ra rằng các nhóm có thu nhập thấp, đặc biệt ở khu vực nông thôn, thường chịu ảnh hưởng nặng nề hơn trước biến động giá cả. Trong khi đó, người tiêu dùng ở khu vực thành thị, dù có mức chi tiêu cao hơn, lại có khả năng thích nghi linh hoạt hơn với lạm phát nhờ vào cơ hội tiếp cận thông tin và điều chỉnh tiêu dùng hiệu quả hơn.

Mặc dù nghiên cứu đã đạt được những kết quả nhất định trong việc phân tích mối quan hệ giữa chỉ số giá tiêu dùng (CPI), lạm phát và chi tiêu của người tiêu dùng, cũng như xây dựng mô hình dự báo CPI trong tương lai, tuy nhiên vẫn còn tồn tại một số điểm có thể được xem xét kỹ lưỡng hơn trong những nghiên cứu tiếp theo.

Trước hết, do phạm vi và mục tiêu của nghiên cứu tập trung vào phân tích mối quan hệ giữa CPI, lạm phát, thu nhập và chi tiêu, nên một số yếu tố kinh tế vĩ mô khác có khả năng ảnh hưởng đến hành vi tiêu dùng như giá cả các mặt hàng thiết yếu, niềm tin người tiêu dùng, tỷ lệ thất nghiệp, hay tác động từ chính sách tài khóa và tiền tệ chưa được đưa vào mô hình. Đây là những yếu tố khách quan nằm ngoài phạm vi dữ liệu hiện có và vượt ngoài giới hạn phân tích của nghiên cứu ở thời điểm hiện tại. Do vậy, việc tích hợp thêm các biến kinh tế vĩ mô khác như thu nhập bình quân, tỷ lệ thất nghiệp, lãi suất, hoặc chỉ số niềm tin người tiêu dùng sẽ giúp làm rõ hơn các yếu tố tác động đa chiều đến hành vi chi tiêu. Đồng thời, các mô hình kinh tế lượng nâng cao như mô hình hồi quy phi tuyến hoặc hồi quy có kiểm soát nội sinh cũng có thể được xem xét để phản ánh tốt hơn những mối quan hệ phức tạp trong thực tiễn.

Cuối cùng, một hướng phát triển có giá trị thực tiễn là phân tích tác động của các chính sách điều hành giá, điều chỉnh lãi suất hoặc trợ giá tiêu dùng đến mối quan hệ giữa lạm phát và chi tiêu, qua đó cung cấp thông tin hữu ích cho quá trình hoạch định chính sách kinh tế vĩ mô.

TÀI LIỆU THAM KHẢO

[1]. Mai Thị Thanh Xuân, "Tác động của lạm phát đến đời sống của người thu nhập thấp ở Việt Nam hiện nay," *Tạp chí Khoa học ĐHQGHN, Kinh tế - Luật*, 24(2), 102-113, 2008.

[2]. Box G. E. P., Jenkins G. M., *Time Series Analysis: Forecasting and Control, Holden-Day*. San Francisco: Holden-Day, Inc., 1970.

[3]. Analytics India Magazine, *Quick way to find p, d and q values for ARIMA*, 2022. https://analyticsindiamag.com/quick-way-to-find-p-d-and-q-values-for-arima/.

[4]. Trần Văn Anh, *Mô hình dự báo ARIMA: Phương pháp Box-Jenkins và ứng dụng trong dự báo ngắn hạn*. Tài liệu khóa học, Trường Đại học Kinh tế - Đại học Quốc gia Hà Nội, 2014.

[5]. Breiman L., "Random Forests," *Machine Learning*, 45(1), 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[6]. Salman H. A., Kalakech A., Steiti A., "Random forest algorithm overview," *Babylonian Journal of Machine Learning*, 69-79, 2024.

[7]. Yeh C.C, Chi D. J, Lin Y. R, *"*Going-concern prediction using hydric random forests and rough set a pproach,*" Information Sciences*, 254, 98-110, 2014.

[8]. Maione C., Batista B. L., Campiglia A. D, Barbosa F. Jr., Barbosa R. M, "Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry," *Computers and Electronics in Agriculture*, 121, 101-107, 2016.

[9]. Tổng cục thống kê, *Y tế, mức sống dân cư, văn hóa, thể thao, trật tự an toàn xã hội và môi trường,* 2024, https://www.nso.gov.vn/y-te-muc-song-dan-cu-van-hoa-the-thao-trat-tu-an-toan-xa-hoi-va-moi-truong/.

AUTHORS INFORMATION

Dang Thi Hong Ha[1], Vu Viet Thang[2], Nguyen Nhu Quynh[3], Vuong Thi Tuyen[1], Bui Le Hien Mai[4]

[1]School of Economics, Hanoi University of Industry, Vietnam

[2]Department of Academic Affairs, Hanoi University of Industry, Vietnam

[3]Student of School of Economics, Hanoi University of Industry, Vietnam

[4]Military Commercial Joint Stock Bank, Vietnam