The Economics and Finance Letters
2025 Vol. 12, No. 3, pp. 596-609
ISSN(e): 2312-430X
ISSN(p): 2312-6310
DOI: 10.18488/29.v12i3.4432
© 2025 Conscientia Beam. All Rights Reserved.

Application of artificial intelligence using linear regression and the Naïve Bayes model in forecasting and analyzing the consumer price index in Vietnam

Pham Thi Ha An[1]
Truong Quoc Tri[2+]
Nguyen Thanh Phuc[3]

[1,3]*Faculty of Finance and Banking, Van Lang University, Ho Chi Minh City, Vietnam.*
[1]*Email: an.pth@vlu.edu.vn*
[3]*Email: phuc.nt@vlu.edu.vn*
[2]*Faculty of Mechanical - Electrical and Computer Engineering, School of Technology, Van Lang University, Ho Chi Minh City, Vietnam.*
[2]*Email: tri.truong@vlu.edu.vn*

(+ Corresponding author)

Article History
Received: 10 June 2025
Revised: 8 August 2025
Accepted: 16 September 2025
Published: 30 September 2025

Keywords
Artificial intelligence
CPI
Forecast
Inflation
Linear regression
Naïve Bayes.

JEL Classification:
C11; C45; C53; E31; E37; O53.

ABSTRACT

The investigation of applying Artificial Intelligence (AI) and Naïve Bayes to forecast the Consumer Price Index (CPI) in Vietnam marks a significant contribution to advancing accurate inflation prediction capabilities. The study leverages rigorous methodological standards and reliable data sources by utilizing a comprehensive 2003 to 2023 dataset comprising seven input variables and the CPI as the output variable. A correlation coefficient of 0.99 indicates a robust correlation between the predicted value and the actual value. The model demonstrates efficacy in forecasting the CPI in both the training dataset and the testing dataset. Furthermore, the histogram visually represents the distribution of errors. The errors are primarily clustered at a minimal magnitude, predominantly falling within the range of -0.05 to 0.03. This suggests that the model tends to make predictions that are quite close to the actual value. The achieved Mean Squared Error (MSE) value of 0.03 demonstrates the model's remarkable accuracy, validating the effectiveness of AI in capturing intricate patterns within CPI data. This research paves the way for further exploration of advanced machine-learning techniques tailored to Vietnam's economic landscape, contributing to improved economic forecasting, proactive policy decisions, and sustainable growth.

Contribution/Originality: This study contributes to the existing literature by demonstrating a strong correlation (0.96) between actual and predicted values. Data points closely align with the regression line, and prediction errors range from -0.04 to 0.04, highlighting the model's high accuracy and reliability in predicting outputs on the training set.

1. INTRODUCTION

Forecasting inflation and macroeconomic indicators is one of the crucial tasks of central banks worldwide, as inflation stabilization is a primary goal of monetary policy management (Juarsa, Janwari, Hasanuddin, Ridwan, & Athoillah, 2025; Wahyudin, 2025). Central banks decide to adjust inflation or deflation through the impact of monetary policy transmission channels with significant lags (Ogbonnaya, Maduka, & Okafor, 2025; Wadood, 2025). Inflation rates tend to converge with global inflation rates and move in parallel trends across countries (Alomani, Kayid, & Abd El-Aal, 2025). This is important in developing inflation forecasting models in the context of complex changes in other factors today. Elevated inflation impacts microeconomic activity and associated macroeconomic indices (Azhar, Ilyas, Ali, & Shafiq, 2025). Rising inflation influences real interest rates, affects the economy's capacity to mobilize capital, and intensifies pressure on production and business operations, resulting in repercussions that the economy must endure, such as economic recession and heightened unemployment (Daniel, Israel, Chidubem, & Quansah, 2021; Muhammad, 2023). High inflation also causes speculation, serious imbalances in the supply-demand relationship of goods in the market, causing economic disruptions and large gaps in income and living standards between the rich and the poor (Dabrowski, 2022). Inflation affects the profitability of long-term contracts, which in turn influences decisions regarding capital mobilization, savings, and investment by the private sector (Faust & Wright, 2013). Therefore, the state needs to control inflation close to the inflation target, on the one hand helping to stabilize the macroeconomy, on the other hand helping the economy develop. To control inflation close to the inflation target, policymakers need to understand what factors affect inflation, forecast future inflation, and then make appropriate adjustments to achieve the set target. There are many different inflation forecasting models, each with its advantages and disadvantages. However, using forecasting tools such as machine learning, deep learning, and artificial intelligence helps produce more accurate forecasting results than traditional forecasting models, particularly for limited databases.

The number of studies on forecasting inflation and macroeconomic indicators in Vietnam remains quite modest. The studies all propose monetary policy implications based on forecasting results from different approaches. Research by Thảo (2015), Đông, Trang, and Lam (2022), Nguyễn, Lê, and Đinh (2021), and Ly and Hà (2022) used the ARIMA or VAR approach to forecast Vietnam's inflation. The research results showed some similarities as follows: (1) the results shared the view that past inflation plays an important role in current inflation fluctuations; (2) the shock of economic fluctuations affecting inflation is relatively small. However, the limitations of time series estimation tools are the requirement of a stationary series, poor ability to handle nonlinear relationships, while inflation itself is affected by many nonlinear factors such as economic quality, political shocks, fluctuations in commodity and currency markets, etc. Most of the previously published studies share the common finding of suitable variables and models for forecasting the inflation index. However, with data sets built in different periods and influenced by varying factors, the choice of inflation index forecasting models and input variables will differ. Therefore, the group of authors recognizes the importance of the problem and wants to study and research the application of ANN and Naïve Bayes, a more advanced technique, to inflation forecasting. The research results are expected to provide additional perspectives and comments that are not only theoretically meaningful but also have flexible application value in supporting monetary policy management decisions for the State Bank of Vietnam.

In recent years, the application of machine learning methods to forecast economic indicators has been widely adopted due to their simple processing and high flexibility in practical applications (Elhoseny, Metawa, Sztano, & El-Hasnony, 2025; Kontopoulou, Panagopoulos, Kakkos, & Matsopoulos, 2023; Sonkavde et al., 2023; Yoon, 2021). Studies confirm the advantages of applying advanced forecasting methods such as machine learning and ANN when providing results with high similarity between forecast values and actual values. In addition, these methods also require simple processing as they do not need prior information about the distribution and probability of the data, can learn from past experience, and work with incomplete and/or nonlinear data. Most current studies build ANN models at the national level and often use annual or monthly time series data. Most current studies build ANN models at the national level and often use annual data series (Elhoseny et al., 2025; Kontopoulou et al., 2023; Sonkavde et al., 2023). Research on applying ANN models to forecast some economic indicators for Vietnam, especially quarterly inflation, has not yet received the attention of researchers. The selection of a model with superior accuracy to forecast inflation has very high practical significance, helping the Central Bank assess the ability to achieve the macroeconomic target, thereby enabling timely policy responses to promote economic growth drivers, aiming to achieve the economic growth target as planned.

The structure of the research is divided into five parts: In the first part, the group of authors introduces the research. In the second part, there is a brief overview of previous studies related to forecasting macroeconomic indicators. The third part provides an overview of the model and research methods. Part four presents the results of forecasting inflation in Vietnam using ANN and Naïve Bayes models. The fifth part presents the conclusions of the study.

2. LITERATURE REVIEW

In fact, there are many econometric models applied to inflation forecasting, especially time series models such as AR, VAR, ARDL, ARIMA, etc. In Vietnam, Đông et al. (2022) used the VAR model to analyze inflation based on quarterly income data from the third quarter of 2006 to the fourth quarter of 2021 in Vietnam. The results showed that, in the long term, the impact of past inflation gradually decreased over time but still explained the fluctuations of current inflation. The remaining factors also affected inflation and contributed to forecasting inflation in the second quarter of 2022, which increased by 3.27% compared to the same period last year. Thành (2012) used the ARDL model to analyze the relationship between inflation and budget deficit, thereby providing policy suggestions in macroeconomic management. In the US, Binner, Elger, Nilsson, and Tepper (2006) applied the AR model to forecast inflation, but the authors confirmed that the inflation forecast using the AR model was not as good as the Markov Autoregressive Model (MS-AR). Safitri and Iwari (2025) forecast the inflation rate for Lampung province, Indonesia, using the ARIMA method. The study is based on the criteria of the Akaike Information Criterion (AIC) and Mean Absolute Percentage Error (MAPE) to select the best ARIMA model among the five tested. The ARIMA method has several limitations in its forecasting role because it does not indicate the influence of independent variables (Nurfadila & Aksan, 2020), but this method is effective for short-term forecasting (Safitri & Iwari, 2025).

Carriero, Clark, Marcellino, and Mertens (2024) employ the BVAR model to project financial market indices in the United States during the COVID-19 pandemic, as the pandemic induces significant volatility in the indicators within the study model. The paper advocates for the use of Student-t distributed random variables and outliers in the stochastic volatility VAR (vector autoregressive) model to address random fluctuations. The integration of artificial neural networks and data mining enables the development of robust economic indicator forecasting models. Urrutia, Longhas, and Mingo (2019) provide a projection of Philippine GDP from Q1 2018 to Q4 2022. The authors evaluated the forecasting methodologies ARIMA and Bayesian ANN for predicting Philippine GDP using variables such as MSE, NMSE, MAE, RMSE, and MAPE. The study demonstrated that ARIMA (1,1,1) and Bayesian ANN exhibited substantial concordance between the predicted and actual values; however, the Bayesian ANN method yielded lower error rates. Smalter and Cook (2017) utilize the Deep Neural Networks (DNN) methodology to predict the unemployment rate for the next quarter, relying solely on the monthly lag of the US unemployment rate from 1948 to 2016 as input data. The findings indicate that the research model exhibits superior performance in the short term (during the next quarter), although fails to provide accurate forecasts for the subsequent 2 to 4 quarters. Costa e Silva, Lopes, Correia, and Faria (2020) highly appreciated the predictive ability of the Logistic regression model or Chang, Chang, and Wu (2018) chose XGBoost for their prediction problem. In addition, many studies have also applied deep learning to improve the accuracy of the model. For instance, Ko, Lin, Do, and Huang (2022) and Graves and Graves (2012) demonstrated the effectiveness of ANN, CNN, and LSTM algorithms.

Lidiema (2017) implemented the Holt–Winters model and the SARIMA model to predict CPI in Kenya from November 2011 to October 2016. The two prediction models were compared using MASE, MAE, and MAPE. For the SARIMA model, the results were 0.059, 0.0036, and 0.073, respectively; for the Holt–Winters model, the results were 0.643, 0.595, and 0.400, respectively. This indicates that the SARIMA model is more accurate than the Holt–Winters model based on the selected parameters. Riofrío, Chang, Revelo-Fuelagán, and Peluffo-Ordóñez (2020) use Support Vector Regression (SVR), LSTM, and SARIMA models to forecast Ecuador CPI. The dataset used consists of 174 samples, collected monthly from January 2005 to June 2019. The results show that the SVR model has the best performance with a Mean Absolute Percentage Error (MAPE) of 0.00171, followed by the LSTM model with an MAPE of 0.00173. Puka and Zaçaj (2018) utilized the SARIMA and ANN methodologies to forecast the US CPI from January 1980 to December 2013. The statistical techniques employed to assess the model encompass AIC and BIC.

A reduced AIC value indicates that the SARIMA model's forecast is more suitable in the short term compared to that of the multiple regression model. Peirano, Kristjanpoller, and Minutolo (2021) utilize ANN, Fuzzy Inference System (FIS), ANN-FIS, and SARIMA-ANN as benchmarks to evaluate the efficacy of SARIMA-LSTM in forecasting inflation rates for Brazil, Mexico, Chile, Colombia, and Peru. The suggested model demonstrates superior accuracy compared to all benchmark models (SARIMA, FIS, ANNFIS, SARIMA-ANN, LSTM, SARIMA-LSTM). For all the examined economies, the suggested model has decreased the MSE by 0.91% for Brazil, 5.51% for Mexico, 3.33% for the inflation rate in Chile, 4.15% for Colombia, and 5.61% for Peru.

Research on forecasting utilizing ANN approaches predominantly emphasizes the optimization of models to achieve superior forecasting performance for indices. Harahap, Lipikorn, and Kitamoto (2020) conducted a study utilizing macroeconomic indicators to predict the Nikkei 225 (N225) and Nikkei 400 (N400) indexes in Japan. Polamuri, Srinivas, and Mohan (2019) investigated stock market forecasting utilizing various machine learning models, including Linear Regression, Multivariate Regression, Random Forest, and Extra Tree Regressor. Additionally, Song, Tang, Wang, and Ma (2023) forecasted stock market volatility by integrating macroeconomic variables through GARCH-MIDAS and deep learning models. In Vietnam, Hai (2024) examines the application of machine learning to analyze the Vietnamese stock market concerning macroeconomic issues. While each research route possesses distinct objectives, it is evident that employing findings from the influence research direction to ascertain input variables for forecasting studies, in conjunction with machine learning techniques to develop a forecasting model, will yield novel outcomes. Zhang, Patuwo, and Hu (1998) assert that the ANN model only predicts well for nonlinear cases, but for linear relationships, the ANN model does not predict as accurately as the linear regression model. In another study by Binner et al. (2010) on inflation in the US economy, the authors concluded that the KRLS model predicts better than the ANN model. In addition, the study by Zhang (2003) and Khashei and Bijari (2011) also concluded that the hybrid model between ANN and ARIMA provides better forecasting results than using these models alone. There are different conclusions regarding the forecasting performance of ANN models, depending on each country and the dataset used by the researcher. In the following sections, the authors will present more details about ANN and its forecasting performance compared to the Naïve Bayes regression model, using the macroeconomic dataset of Vietnam.

3. METHODOLOGY

3.1. AI- Linear Regression

Depicting the linear relationship between a dependent variable and one (univariate) or multiple independent variables (multivariate), the linear regression model is a statistical technique. A multivariate linear regression model comprises an output variable (dependent), frequently represented as y, which is subject to the influence of numerous input variables (inputs), frequently denoted as $\boldsymbol{x} = [x_1, x_2, \dots, x_k]^T$.

The CPI index is the dependent variable in this study. A linear regression model can be employed to ascertain the relationship between the CPI and other economic factors (independent variables), including the following: other commodities and services, such as food, beverages, and tobacco; housing and construction materials; household appliances; and education.

The following equation represents the relationship between x and y.

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Where x_i $(i = 1, \dots, k)$ are independent variables; $\beta_1, \beta_2, \dots, \beta_k$ are estimated coefficients and ε is an error term. In order to optimize the coefficients, the equation's output will be designed to predict values that are closer to the actual results.

3.2. Naïve Bayes Model

The Naïve Bayes model is an algorithm for machine learning tasks that classifies data probabilistically. It finds extensive application in classification and prediction-related fields, including but not limited to text classification, object classification in images, and fate classification in finance.

The foundation of the Naïve Bayes algorithm lies in Bayes' theorem, a statistical probability principle. In particular, it calculates the probability of an independent variable by utilizing the Bayesian formula and the probabilities of other independent variables. All independent variables are assumed to be uncorrelated, or wholly independent, by the model.

$$P\ (Y|X) = P(X|Y) \times P(Y)/P(X)$$

Where $P\ (Y|X)$ is the probability of the independent variable Y, based on data on variable X; $P(X|Y)$ is the probability of variable X, based on the data of variable Y; $P(X)$ is the prior probability of the variable X; $P(Y)$ is the prior probability of variable Y.

One notable benefit of employing the Naïve Bayes classification model is its minimal training data requirement, which is sufficient to calculate the classification parameters (variable means and variances). Given the assumption of independent variables, the variance of the variables for each category is the only one that needs to be calculated, rather than the total variance. It applies to multiclass and binary classification problems.

3.3. Data

This research employs time series data spanning from January 2003 to December 2023. Data on seven input variables food, drinks, tobacco, housing and building materials, household appliances, education, and other products and services and one outcome variable the total consumer price index were collected for the study from January 2003 to December 2023. The consumer price index quantifies the average variation in prices of consumer products and services across time, based on consumer spending patterns. The selected input variables food, beverages, tobacco, housing and construction materials, domestic appliances, education, and other goods and services constitute the basket of goods utilized for calculating the Consumer Price Index (CPI) in Vietnam during the research period, as per the General Statistics Office of Vietnam. Data is sourced directly from the General Statistics Office of Vietnam (GSO) in accordance with the Consumer Price Index (CPI). The study designated the research period from 2003 to 2023 due to: (1) the sample size guaranteeing an adequate number of observations for the ANN and Naïve Bayes model forecasts, and (2) the research team's ability to gather sufficient data on input and output variables throughout this period, thereby mitigating errors from missing data and ensuring the accuracy and reliability of the analysis and model.

4. EMPIRICAL RESULTS

4.1. AI- Linear Regression Model Results in Forecasting the Consumer Price Index (CPI)

4.1.1. Evaluating the Model on the Training Dataset

Figure 1 illustrates a strong correlation between the actual value and the predicted value of the model. The vertical axis represents the predicted value, while the horizontal axis represents the actual value. Many data points are closely clustered around or aligned with a straight line, indicating a robust association between the two variables. Data points that are close to the straight line show that the Linear Regression model is effective in accurately estimating and forecasting values based on the input variables. This demonstrates that the model acquired knowledge efficiently from the training data and achieved high accuracy in its predictions.

As displayed in Figure 2, the presence of a high correlation coefficient (0.99) indicates a strong and direct association between the input variable (prediction) and the output variable (actual). A positive correlation coefficient indicates a direct relationship between the projected price and the actual value. This demonstrates the efficacy of the trained model since the projected value closely aligns with the actual value.

Figure 3 depicts a histogram chart that graphically assesses the error of the training set. The vertical axis represents the frequency of events, while the horizontal axis represents the measurement of errors. Upon examining the figure, it is evident that most mistakes fall between -0.05 and 0.03. This indicates that the training set model makes predictions close to the actual values. Specifically, most predictions fall within the range of -0.03 to 0.01, occurring 23 times, and around -0.01 to 0.01, occurring over 35 times. Nevertheless, within the range of -0.05 to -0.01, the model consistently underestimates the true value, whereas within the range of 0.01 to 0.03, the model consistently overestimates. This indicates that the model possesses a relatively strong understanding of the data; however, it still exhibits some deviation, either positive or negative, when compared to the actual values. Furthermore, we observe a decrease in significant deviations, specifically those falling outside the range of -0.05 to 0.03. The model exhibits challenges in accurately forecasting larger error magnitudes.

Regarding Figure 4, upon examining the Mean Squared Error (MSE) of the model, which quantifies the average discrepancy between the predicted and actual values, it is determined that the model's MSE is 0.03. This figure signifies that the overall mean squared error of the model is relatively low, which is typically regarded as a favorable indication. Additionally, numerous data points on the chart exhibit proximity, indicating a strong correlation and alignment between the projected and observed values. The model trained on the training set is performing exceptionally and demonstrates precise predictive capabilities. After analyzing the evaluations of Linear Regression, it can be concluded that the model worked well on the training set and exhibited accuracy, strong correlation, and precise prediction capability. Nevertheless, to guarantee the model's robustness, it is imperative to continue evaluating and verifying its performance on testing sets.

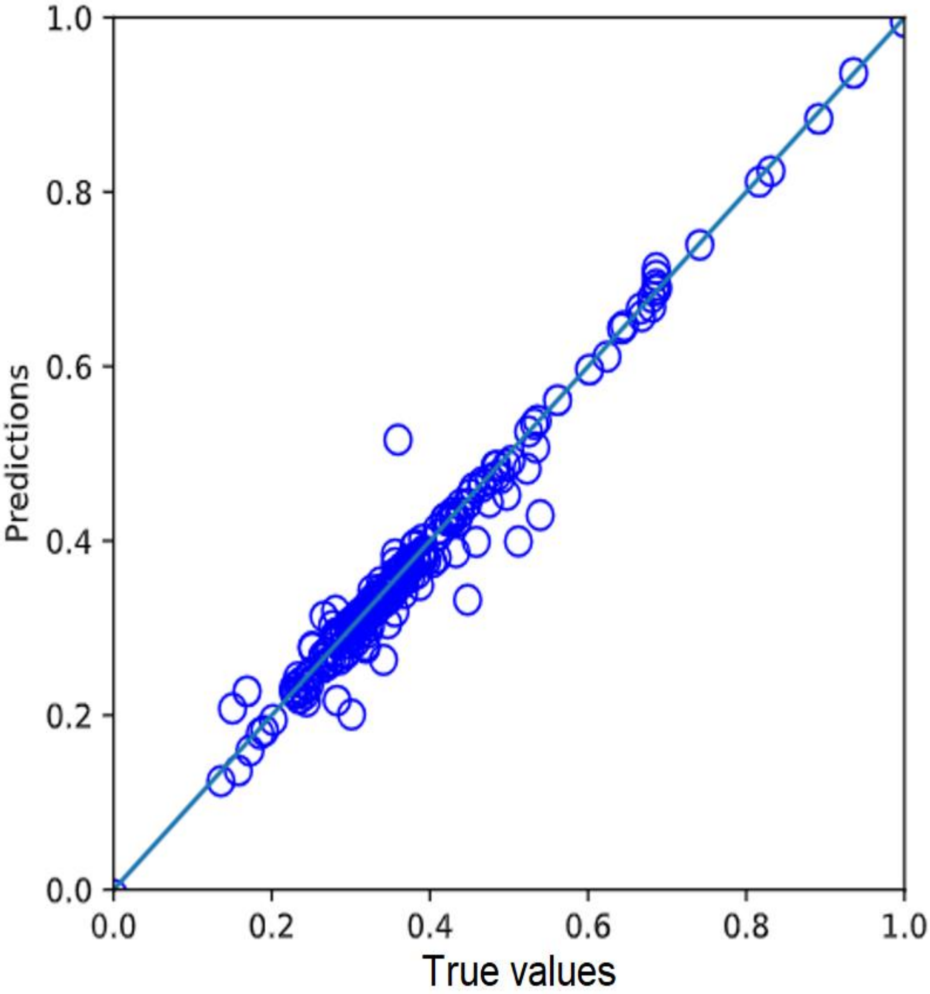

Figure 1. Scatter plot of the relationship between the actual value and the predicted value on the training dataset.

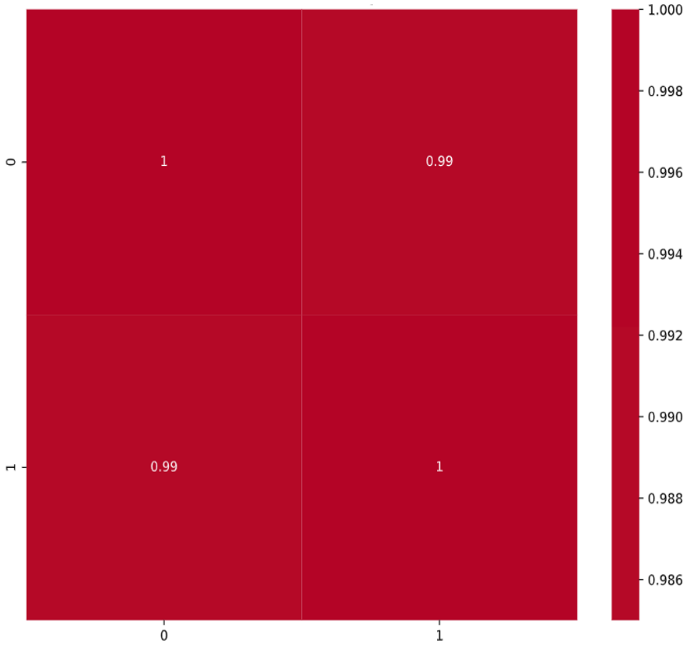

Figure 2. Correlation coefficient between the actual value and the predicted value of the training dataset.

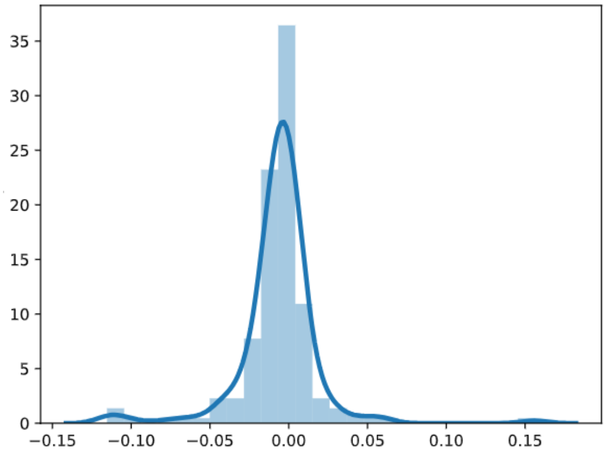

Figure 3. Visualize the Histogram chart to evaluate the error of the training dataset.

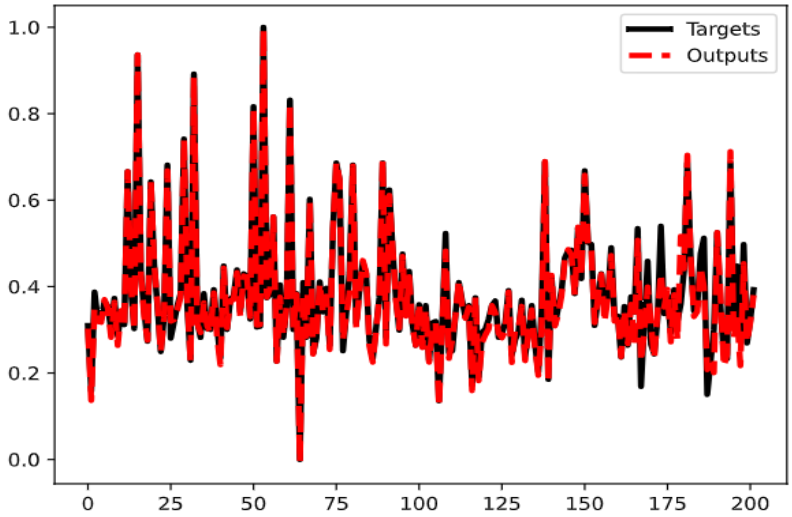

Figure 4. Illustrates the true and predicted values of the training dataset.

Note: The mean squared error (MSE) of the model on the training set is 0.03. The red line represents the predicted value, while the black line represents the true value.

4.1.2. Evaluating the Model on the Test Dataset

For Figure 5, regression analysis on the testing set reveals that numerous data points remain close to or on the regression line, similar to the training set. This illustrates the correlation and precision of the model in calculating the projected value using the input variables. Nevertheless, we also observe a small number of data points that significantly diverge from the regression line. These limited data points may be considered outliers or exhibit significant deviations from the projected model. This implies that the model may have limitations in accurately estimating outliers or in unique circumstances. Despite this, a linear correlation exists between the predicted values and the actual values on the testing set. However, some degree of skewness can occur in the data. To enhance the comprehensiveness of our evaluation, we will proceed with additional analyses to assess the model's performance on the testing set. Figure 6 illustrates the correlation coefficient (0.9) between the predicted and actual values for the test dataset. This score indicates a strong correlation between the predicted and target values. The correlation coefficient in the testing set (0.9) is lower than that in the training set (0.99). This phenomenon is common in machine learning models, particularly linear regression models. The Vietnam CPI dataset used for model training was collected from 2003 to 2023. The dataset is relatively small, resulting in a testing set that exhibits complex patterns and a different data distribution compared to the training set.

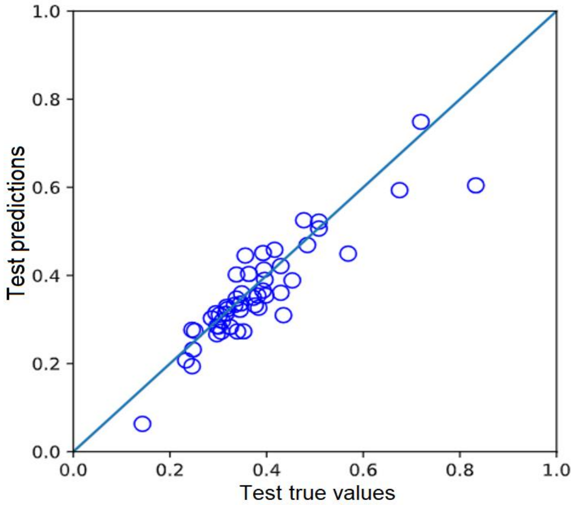

Figure 5. Scatter plot of the relationship between the actual value and the predicted value on the testing dataset.

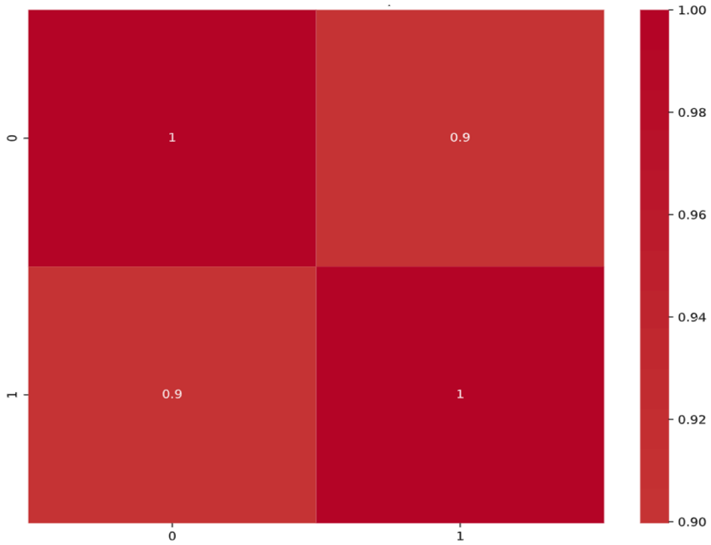

Figure 6. Correlation coefficient between the actual value and the predicted value of the testing dataset.

4.2. Naïve Bayes Model Yields Outcomes While Examining the CPI

The Naïve Bayes model is used to generate results during training. A performance evaluation of the model will be undertaken on both the training and testing sets. This enables us to assess the model's proficiency in categorizing situations as either "Good" or "Bad".

4.2.1. Evaluating the Model on the Training Dataset

The performance evaluation of the classification model on the training set, considering the parameters, is reported in Table 1 with detailed descriptions as follows:

Error Rate:

The model's error rate, computed as 0.405, indicates the extent of inaccuracies in the model's classification of samples within the training set. More precisely, this ratio quantifies the proportion of incorrectly identified samples of the total number of samples in the training set.

The analysis reveals that the model's predictions were inaccurate for around 40.5% of the entire sample set used for training. This might greatly diminish the model's dependability and efficiency in categorizing instances.

Recall:

The model's recall is 0.405. Given this value, the model's ability to correctly recall "Good" cases in the training dataset is approximately 40.5%. This indicates that the model tends to overlook certain positive examples, which can lead to the loss of crucial information or introduce ambiguity in the classification of "Good" cases.

F1-Score:

The F1-score is a metric used to measure the performance of a classification model. It combines precision and recall into a single value, providing a balanced assessment of the model's accuracy. The F1-score is 0.629, which indicates medium performance in categorization. According to this value, the model demonstrates a favorable balance between accuracy (the capacity to accurately predict) and recall (the ability to identify all positive cases).

1-Precision:

The precision number, denoted as 1-Precision, is 0.257. This indicates that the model incorrectly identified approximately 25.7% of the "Good" occurrences in the test dataset. The misprediction rate can be observed as roughly 25.7% of the total number of "Good" events in the testing set.

Confusion of Matrix:

The performance of the model can be summarized based on the findings obtained from the confusion matrix in the following manner:

- True Positives (TP): A total of 27 samples were accurately identified as good.
- False Negatives (FN): 12. Good samples have been erroneously categorized as Bad.
- False Positives (FP): A total of 59 erroneous samples have been incorrectly categorized as Good.
- True Negatives (TN): A total of 77 samples were accurately identified as bad.

Consequently, the training set contains a total of 175 samples. Upon completion of training, the model accurately classified 104 samples, comprising 27 true positives and 77 true negatives. The model incorrectly identified 71 samples, comprising 12 false negatives and 59 false positives.

The existing model must be enhanced to attain superior categorization performance. Possible enhancements can be achieved by optimizing the model's hyperparameters, augmenting the dataset, mitigating noise in the input data, or modifying the network architecture. This will improve the precision, recall, and error reduction in the classification of "Good" and "Bad" situations.

Table 1. Model evaluation on the training dataset.

Training dataset						
Supervised learning (Naïve Bayes)						
Parameters						
Good				1		
Bad				2		
Classifier performances						
Error rate		0.405				
Value prediction				**Confusion matrix**		
Recall	F1-score	1-Precision		Good	Bad	Total
0.405	0.629	0.257	Good	27	12	39
			Bad	59	77	136
						175

4.2.2. Evaluating the Model on the Test Dataset

As shown in Table 2, the evaluation results of the classification model on the testing set, based on the given parameters, are presented as follows:

Error Rate:

The model's error rate is 0.329, indicating that approximately 32.9% of the samples in the testing set are categorized incorrectly. The error rate of the current training dataset is lower than that of the prior dataset (0.405), indicating an improvement in the model's classification performance.

Recall:

The model's recall rate is 0.67, indicating that it successfully recalls approximately 67% of the "Good" positive cases in the test dataset. The current training dataset (0.405) represents an improvement compared to the previous one, indicating that the model has achieved a higher level of recall.

F1-Score:

The F1-score has increased to 0.729, surpassing the previous training dataset's value of 0.629. This score indicates that the model has superior performance in effectively combining precision and recall.

1-Precision:

The 1-Precision is 0.108, which is lower than the precision ratio of the prior training data set (0.257). This indicates that the model incorrectly classified approximately 10.8% of the "Good" instances in the testing set.

Confusion matrix:

The confusion matrix provides a detailed analysis of the model's performance in classifying situations as "Good" or "Bad".

- True Positives (TP): 8 samples were accurately identified as Good.
- False Negatives (FN): 1 acceptable sample is erroneously categorized as unacceptable.
- False Positives (FP): 24 defective samples are erroneously categorized as acceptable.
- True Negatives (TN): A total of 43 samples were accurately identified as bad.

Consequently, the testing set contains a total of 76 samples. The model accurately identified 51 samples, comprising 8 true positives and 43 true negatives. The model incorrectly identified 25 samples, comprising 1 false negative and 24 false positives.

Table 2. Model evaluation on the test dataset.

Test dataset						
Supervised learning (Naïve Bayes)						
Parameters						
Good				1		
Bad				2		
Classifier performances						
Error rate	0.329					
Value prediction				**Confusion matrix**		
Recall	F1-score	1-Precision		Good	Bad	Total
0.67	0.729	0.108	Good	8	1	9
			Bad	24	43	67
						76

5. CONCLUSION

The research employs Artificial Intelligence models, specifically Linear Regression and Naïve Bayes models, to forecast and analyze the Consumer Price Index (CPI) in Vietnam. Both models utilize a dataset comprising 251 samples collected from 2003 to 2023. The Linear Regression model predicts the CPI with a high correlation coefficient of 0.99 on the training dataset, which decreases to 0.9 on the testing dataset. Similarly, the Naïve Bayes model produces favorable results when evaluating the CPI index. However, the outcomes on both datasets indicate that the models have difficulty reliably classifying the "Bad" occurrences, suggesting limitations in their predictive capabilities for certain classifications.

5.1. The Application of a Linear Regression Model Yields Accurate Predictions for the Consumer Price Index (CPI)

The scatterplot illustrates the correlation between the observed and predicted values in the training set. The proximity of numerous data points to the regression line demonstrates the model's accuracy and strong estimation capability. A correlation coefficient of 0.99 indicates a robust correlation between the predicted value and the actual value. The model demonstrates efficacy in forecasting the dependent variable based on the independent variable within the training dataset. Furthermore, the histogram visually represents the distribution of errors. The errors are primarily clustered at a minimal magnitude, predominantly falling within the range of -0.05 to 0.03. This suggests that the model tends to make predictions that are quite close to the actual value. Ultimately, the model assesses the predicted and actual values, and the comparison graph between them demonstrates the correlation and precision of the model. A Mean Squared Error (MSE) value of 0.03 indicates a high level of accuracy in the model.

The scatterplot on the testing set illustrates the correlation between the observed and forecasted values. While the majority of data points adhere to the regression line, there are a few outliers that deviate from it. This is evident from the correlation coefficient on the testing set, which decreases to 0.9. This indicates a moderate association between the predicted values and the actual values. This drop could indicate challenges in extrapolating the model. The error distribution closely resembles that of the training set, with a high concentration of errors occurring near zero. The model exhibits strong predictive capability on the majority of the data. Ultimately, when comparing the predicted and actual values, any additional discrepancies observed on the testing set may indicate a lack of generalization or a probable decline in performance on the data.

5.2. The Naïve Bayes Model Yields Outcomes when Examining the Consumer Price Index (CPI)

The performance of the Naïve Bayes model is assessed by applying a set of performance measures to both the training and testing datasets. The model exhibits the capacity to accurately categorize significant errors in both the training set and the testing set, indicating a need for improvement.

Recall and F1-Score: The testing set exhibits an improvement in memory and F1-Score relative to the training set. While the model demonstrates superior memory and precision in weighing, further enhancements are necessary to guarantee consistent performance. The confusion matrix reveals that the model exhibits high accuracy in identifying "Good" cases, but it still encounters difficulty in accurately classifying "Bad" cases.

This study provides data and results that are more current than those of prior research on inflation predictions in Vietnam. The artificial intelligence model exhibits robust predictive capabilities throughout the majority of the data, showing a significant correlation between forecasted and actual values, while the Naïve Bayes model displays great precision in distinguishing "Good" cases during forecasting. Recent domestic and international economic instabilities, including heightened export taxes on goods, have diminished export market share to the US and European markets. Additionally, the fuel supply issue resulting from conflicts in certain nations escalates oil and gold prices, perhaps leading to a resurgence of inflation. Consequently, precise forecasting of macroeconomic variables is crucial for effective and adept monetary policy administration. The synchronized coordination of state agencies in regulating key input variables of the CPI index, including food, beverages, tobacco, housing, construction materials, household appliances, education, and other goods and services, will also aid in managing inflation in Vietnam in the future. By regulating the CPI index, the government can focus on stabilizing and enhancing other key macroeconomic variables of the economy.

Despite numerous contributions to the application of contemporary methodologies in forecasting inflation for Vietnam, the study is constrained by restrictions, notably that both models indicate the dataset is insufficient to generalize the data features. Future research may concentrate on employing approaches such as augmentation to enhance the dataset and cross-validation to guarantee that the model acquires more generalized patterns.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Data Availability Statement: The corresponding author can provide the supporting data of this study upon a reasonable request.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.

REFERENCES

Alomani, G., Kayid, M., & Abd El-Aal, M. F. (2025). Global inflation forecasting and uncertainty assessment: Comparing ARIMA with advanced machine learning. *Journal of Radiation Research and Applied Sciences*, *18*(2), 101402. https://doi.org/10.1016/j.jrras.2025.101402

Azhar, M., Ilyas, S., Ali, S. Y., & Shafiq, A. (2025). Assessing the macroeconomic consequences of climate change: Impacts on GDP growth, inflation volatility, and agricultural productivity in developing economies. *Review Journal of Social Psychology & Social Works*, *3*(2), 250-268. https://doi.org/10.71145/rjsp.v3i2.191

Binner, J. M., Elger, C. T., Nilsson, B., & Tepper, J. A. (2006). Predictable non-linearities in US inflation. *Economics Letters*, *93*(3), 323-328. https://doi.org/10.1016/j.econlet.2006.06.001

Binner, J. M., Tino, P., Tepper, J., Anderson, R., Jones, B., & Kendall, G. (2010). Does money matter in inflation forecasting? *Physica A: Statistical Mechanics and its Applications*, *389*(21), 4793-4808. https://doi.org/10.1016/j.physa.2010.06.015

Carriero, A., Clark, T. E., Marcellino, M., & Mertens, E. (2024). Addressing COVID-19 outliers in BVARs with stochastic volatility. *Review of Economics and Statistics*, *106*(5), 1403-1417. https://doi.org/10.1162/rest_a_01213

Chang, Y.-C., Chang, K.-H., & Wu, G.-J. (2018). Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing*, *73*, 914-920. https://doi.org/10.1016/j.asoc.2018.09.029

Costa e Silva, E., Lopes, I. C., Correia, A., & Faria, S. (2020). A logistic regression model for consumer default risk. *Journal of Applied Statistics*, *47*(13-15), 2879-2894. https://doi.org/10.1080/02664763.2020.1759030

Dabrowski, M. (2022). *Demand-and supply-side factors behind the higher inflation*. Paper presented at the 18th EUROFRAME Conference on Economic Policy Issues in the European Union on 'Macroeconomic Policy Challenges in Pandemic Times', Helsinki, June.

Daniel, S. U., Israel, V. C., Chidubem, C. B., & Quansah, J. (2021). Relationship between inflation and unemployment: Testing Philips curve hypotheses and investigating the causes of inflation and unemployment in Nigeria. *Traektoriâ Nauki= Path of Science*, *7*(9), 1013-1027. https://doi.org/10.22178/pos.74-13

Đông, P. T., Trang, N. K., & Lam, P. T. (2022). Applying the VAR model to analyze some factors affecting inflation and forecasting inflation in Vietnam. *Journal of Economics and Management*, *168*, 14–23.

Elhoseny, M., Metawa, N., Sztano, G., & El-Hasnony, I. M. (2025). Deep learning-based model for financial distress prediction. *Annals of Operations Research*, *345*(2), 885-907. https://doi.org/10.1007/s10479-022-04766-5

Faust, J., & Wright, J. H. (2013). Forecasting inflation. *Handbook of Economic Forecasting*, *2*, 2-56. https://doi.org/10.1016/B978-0-444-53683-9.00001-3

Graves, A., & Graves, A. (2012). *Supervised sequence labelling with recurrent neural networks. Studies in Computational Intelligence* (Vol. 385). Berlin, Germany: Springer.

Hai, N. V. (2024). Application of machine learning in analyzing the Vietnamese stock market in relation to macroeconomic issues. *Journal of Financial Data Science and Economics*, *5*(1), 45–60.

Harahap, L. A., Lipikorn, R., & Kitamoto, A. (2020). *Nikkei stock market price index prediction using machine learning*. Paper presented at the Journal of Physics: Conference Series.

Juarsa, E., Janwari, Y., Hasanuddin, M., Ridwan, A. H., & Athoillah, M. A. (2025). The role of central banks in inflation and exchange rate stability amidst global economic challenges: Monetary policy approach. *Strata International Journal of Social Issues*, *2*(1), 29-36.

Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. *Applied Soft Computing*, *11*(2), 2664-2675. https://doi.org/10.1016/j.asoc.2010.10.015

Ko, C. H., Lin, P. C., Do, Q. H., & Huang, Y. H. (2022). Application of ANN, CNN, and LSTM in predictive modeling: A comparative study. *Journal of Artificial Intelligence and Data Science*, *4*(2), 100–112.

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., & Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in data driven networks. *Future Internet*, *15*(8), 255. https://doi.org/10.3390/fi15080255

Lidiema, C. (2017). Modelling and forecasting inflation rate in Kenya using SARIMA and Holt-Winters triple exponential smoothing. *American Journal of Theoretical and Applied Statistics*, *6*(3), 161-169. https://doi.org/10.11648/j.ajtas.20170603.15

Ly, N. H., & Hà, H. T. T. (2022). Combining machine learning and statistical models in time series forecasting: The case of inflation in Vietnam in the period 2000-2021. *Journal of Economic Science*, *10*(1), 18-27.

Muhammad, A. A. (2023). Examining the relationship among unemployment, inflation, and economic growth. *Journal of Business and Economic Options*, *6*(2), 23-31.

Nguyễn, Đ. T., Lê, H. A., & Đinh, T. P. A. (2021). Forecasting Vietnam's economic growth and inflation: A comparison between Var, Lasso and MLP models. *Journal of Commercial Science*, *154*(1), 1–13.

Nurfadila, K., & Aksan, I. (2020). Arima Box-Jenkins method application for forecasting daily mobile data usage. *Journal of Mathematics: Theory and Applications*, *2*(1), 5-10. https://doi.org/10.31605/jomta.v2i1.749

Ogbonnaya, K. S., Maduka, O. D., & Okafor, S. O. (2025). Monetary policy and price stability in Nigeria. *NAU Eco Journals*, *22*(1), 125-146.

Peirano, R., Kristjanpoller, W., & Minutolo, M. C. (2021). Forecasting inflation in Latin American countries using a SARIMA–LSTM combination. *Soft Computing*, *25*(16), 10851-10862. https://doi.org/10.1007/s00500-021-06016-5

Polamuri, S. R., Srinivas, K., & Mohan, A. K. (2019). Stock market prices prediction using random forest and extra tree regression. *International Journal of Recent Technology and Engineering*, *8*(3), 1224-1228. https://doi.org/10.35940/ijrte.C4314.098319

Puka, L., & Zaçaj, O. (2018). *Forecasting consumer price index (CPI) using time series models and multi regression models (Albania case study)*. Paper presented at the Proceedings of the 10th International Scientific Conference "Business and Management 2018" Vilnius Gediminas Technical University, Vilnius, Lithuania.

Riofrío, J., Chang, O., Revelo-Fuelagán, E. J., & Peluffo-Ordóñez, D. H. (2020). Forecasting the consumer Price index (CPI) of ecuador: A comparative study of predictive models. *International Journal on Advanced Science, Engineering and Information Technology*, *10*(3), 1078-1084. https://doi.org/10.18517/ijaseit.10.3.10813

Safitri, I., & Iwari, A. R. P. (2025). Forecasting the inflation rate in Lampung Province using the ARIMA method. *Bit-Tech*, *7*(3), 1046-1056. https://doi.org/10.32877/bt.v7i3.2335

Smalter, H. A., & Cook, T. R. (2017). *Macroeconomic indicator forecasting with deep neural networks*. Federal Reserve Bank of Kansas City Working Paper No. 17-11.

Song, Y., Tang, X., Wang, H., & Ma, Z. (2023). Volatility forecasting for stock market incorporating macroeconomic variables based on GARCH-MIDAS and deep learning models. *Journal of Forecasting*, *42*(1), 51-59. https://doi.org/10.1002/for.2899

Sonkavde, G., Dharrao, D. S., Bongale, A. M., Deokate, S. T., Doreswamy, D., & Bhat, S. K. (2023). Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. *International Journal of Financial Studies*, *11*(3), 94. https://doi.org/10.3390/ijfs11030094

Thành, S. Đ. (2012). Budget deficit and inflation: Empirical evidence in Vietnam. *Economic Development Magazine*, *259*(5), 12-18.

Thảo, N. T. P. (2015). Applying autoregressive model combined with moving average to forecast inflation rate in Vietnam in 2015. *Journal of Science – Hue University*, *109*(10), 273-282.

Urrutia, J. D., Longhas, P. R. A., & Mingo, F. L. T. (2019). *Forecasting the Gross Domestic Product of the Philippines using Bayesian artificial neural network and autoregressive integrated moving average*. Paper presented at the AIP Conference Proceedings.

Wadood, M. R. (2025). Monetary policy transmission in Bangladesh: Evaluating the most effective channels. *Journal of Business and Economic Options*, *8*(1), 15-27.

Wahyudin, A. (2025). The role of monetary policy in addressing economic and financial challenges: Effective strategies for managing inflation and growth. *Journal of Multi Currency*, *1*(1), 1-13.

Yoon, J. (2021). Forecasting of real GDP growth using machine learning models: Gradient boosting and random forest approach. *Computational Economics*, *57*(1), 247-265. https://doi.org/10.1007/s10614-020-10054-w

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*(1), 35-62. https://doi.org/10.1016/S0169-2070(97)00044-7

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175. https://doi.org/10.1016/S0925-2312(01)00702-0

Views and opinions expressed in this article are the views and opinions of the author(s), The Economics and Finance Letters shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.