The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 1, pp.159-164, January 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2023.10.1.159

JCCT 2024-1-19

경제지표를 활용한 다중선형회귀 모델 기반 국제 휘발유 가격 예측

A study of Predicting International Gasoline Prices based on Multiple Linear Regression with Economic Indicators

한명은*, 김지연**, 이현희*, 김세인***, 박민서****

Myeongeun Han*, Jiyeon Kim, Hyunhee Lee*, Sein Kim***, Minseo Park******

요 약 국내 석유 시장은 국제 석유 가격의 변동에 매우 민감하기 때문에 그 변동성에 대한 파악과 대처가 중요하다. 특히, 높은 소비량을 보이는 휘발유의 가격이 어떠한 요인에 인해 변화하는지 명확하게 파악하는 것이 필요하다. 국제 휘발유 가격은 휘발유 수급, 지정학적 사건, 미국 달러화 가치 변동 등 글로벌 요인에 영향을 받는다. 그러나 기존의 연구들은 휘발유의 수급에만 초점에 맞추어 진행하였다는 한계가 존재한다. 본 연구에서는 다양한 머신러닝 기반의 회귀 모델을 활용하여 거시적 경제지표와 국제 휘발유 가격 간의 인과관계를 탐색한다. 첫째, 다양한 세계 경제지표 데이터를 수집한다. 둘째, 데이터 전처리를 진행한다. 셋째, 다중선형회귀, Ridge 회귀, Lasso(Least Absolute Shrinkage and Selection Operator) 회귀 모델을 활용하여 모델링한다. 실험 결과, 테스트 데이터 셋에서 다중선형회귀 모델이 가장 높은 정확도(97.3%)를 보였다. 우리는 국제 휘발유 가격의 예측은 국내 경제 안정성과 에너지 정책 결정에 도움이 될 수 있을 것으로 기대한다.

주요어 : 국제 휘발유 가격, 경제지표, 머신러닝, 다중선형회귀

Abstract The domestic petroleum market is highly sensitive to changes in international oil prices. So, it is important to identify and respond to those changes. In particular, it is necessary to clearly understand the factors causing the price fluctuations of gasoline, which exhibits high consumption. International gasoline prices are influenced by global factors such as gasoline supplies, geopolitical events, and fluctuations in the U.S. dollar. However, previous studies have only focused on gasoline supplies. In this study, we explore the causal relationship between economic indicators and international gasoline prices using various machine learning-based regression models. First, we collect data on various global economic indicators. Second, we perform data preprocessing. Third, we model using Multiple linear regression, Ridge regression, and Lasso(Least Absolute Shrinkage and Selection Operator) regression. The multiple linear regression model showed the highest accuracy at 96.73% in test sets. As a result, Our Multiple linear regression model showed the highest accuracy at 96.73% in test sets. We will expect that our proposed model will be helpful for domestic economic stability and energy policy decisions.

Key words : International Gasoline Prices, Economic Indicators, Machine Learning, Multiple Linear Regression

*준회원, 서울여자대학교 데이터과학전공 학부생
**준회원, 서울여자대학교 디지털미디어학과 학부생
***준회원, 서울여자대학교 데이터사이언스학과 학부생
****정회원, 서울여자대학교 데이터사이언스학과 조교수(교신저자)Dept. of Data Science , Seoul Women's Univ, Korea
접수일: 2023년 10월 5일, 수정완료일: 2023년 10월 21일
게재확정일: 2023년 11월 5일

Received: October 5, 2023 / Revised: October 21, 2023
Accepted: November 5, 2023
****Corresponding Author: mpark@swu.ac.kr

Ⅰ. 서 론

국가통계포털에서 발표한 에너지 수급 통계에 따르면, 2022년 12월 현재 국내 석유 제품 소비량은 전체 에너지 소비량의 약 88.5%를 차지한다[1]. 그러나 국내 에너지 수급 구조의 한계로 인해 석유 제품 소비에 필요한 대부분을 수입에 의존하고 있다[2]. 국내 석유 시장은 국제 석유 가격의 변동에 매우 민감하므로, 국제 석유 가격의 예측은 우리나라의 경제 안정성과 에너지 정책 결정에 큰 영향을 미친다[3, 4].

국제 석유 가격은 석유 시장에서의 수요, 공급을 포함하여 미 달러의 환율, 지정학적 사건, 시장 참여자의 심리 등 다양한 요인에 의해 변동된다[5]. 석유 가격의 변동은 즉각적으로 이루어지기 때문에 그 변동성에 대한 파악과 대처가 필요하다. 특히, 도로 기반 수송이 상대적으로 발달한 우리나라의 경우 대표적인 도로 수송용 연료인 휘발유의 소비량이 다른 석유제품에 비해 상대적으로 높기 때문에, 국제 휘발유 가격이 어떠한 요인으로 인해 변화하는지 명확하게 파악하는 것이 필요하다[6].

기존 국제 석유 가격에 관한 연구는 수급 요인에만 중점을 두어 진행되고 있다. 그러나 국제 석유 가격은 수급 요인 외에도 다양한 세계 경제지표에 의해 영향을 받는다[5]. 세계 경제지표는 정치, 경제, 금리 요인을 포함하며 이는 심리적, 기술적, 제도적 요인으로 인해 다양한 시차를 가지고 있다.

따라서 본 연구에서는 다양한 세계 경제지표를 수집하여 시차를 반영하고, 이를 독립 변수로 사용하여 국제 휘발유 가격을 예측하는 머신러닝 기반의 회귀 모델을 제안한다. 본 논문은 5장으로 구성되어 있다. 2장에서는 국제 석유 가격 예측에 관한 선행연구를 언급한다. 3장에서는 모델을 설계하고 검증 과정을 통해 최적의 모델을 탐색한다. 4장에서는 결과를 분석하고 5장에서는 결론을 언급한다.

Ⅱ. 선행연구

국제 석유 가격의 변동은 국내외 경제 흐름에 많은 영향을 미치기 때문에, 다양한 국제 석유 가격 예측에 관한 연구가 진행되고 있다. 과거의 국제유가 예측은 통계적인 시계열 분석으로 진행되었다[7-9]. Morana는 유가 예측을 위해 GARCH(Generalized Auto Regressive Conditional Heteroskedasticity) 모델을 활용하여 유가 단기 예측을 진행하였다[7]. Yousefi은 wavelet-based 알고리즘을 사용하여 유가를 예측하였다[8]. 송경재는 1984년 1월 ~ 2004년 12월 84분기의 WTI (West Texas Intermediate)의 평균 가격을 ARIMA(Auto Regressive Integrated Moving Average) 시계열 모델에 적용하여 WTI 유가 추이를 예측하였다[9]. 그러나 단순한 시계열 모델은 유가 변동에 영향을 미치는 다양한 요인을 파악하고 그 관계를 정의하기 어렵다는 한계가 있다. 이와 같은 한계를 극복하기 위해 최근 연구에서는 변수의 인과관계를 파악할 수 있는 머신러닝 기반의 회귀 모델을 활용하여 국제유가를 예측한다[10, 11]. 박강희는 1992년 1월 ~ 2008년 7월의 월별 OPEC(Organization of the Petroleum Exporting Countries) 석유 생산량, OECD 석유 수요량, 세계 석유 총생산량, 사우디 석유 생산량, 세계 석유 총 수요량 등의 6개 독립변수를 활용해 국제 석유 가격을 예측하는 로지스틱 회귀(Logistic Regression) 모델을 제안하였다[10]. 김선미는 2004년 1월 ~ 2020년 12월의 WTI 유가, 국가별 생산량, 소비량, 수출량 데이터인 정형데이터와 구글의 트랜드 검색 키워드 추세를 지수화한 데이터를 사용해 국제 석유 가격을 예측하는 방법을 제시하고, 다양한 머신러닝 모델을 통해 향상된 모델 성능을 증명하였다[11].

이와 같이 기존 연구는 석유의 생산량, 소비량과 같은 수급 요인에 기반을 두어 분석한다. 그러나 국제유가는 수급 요인 이외에도 국제적 정세에 의해 영향을 받아 변화하기 때문에 다양한 경제지표를 함께 분석할 필요가 있다. 따라서 본 논문에서는 머신러닝을 기반으로 다양한 경제지표를 활용하여 국제 휘발유 가격을 예측하고자 한다.

Ⅲ. 국제 휘발유 가격 예측 모델 설계

본 연구에서는 다양한 세계 경제지표 데이터를 수집하여 시차를 반영한 다음 이를 바탕으로 국제 휘발유를 예측하는 머신러닝 모델을 제안한다. 본 연구에서 제안하는 모델의 전체 프로세스는 그림 1과 같다. 그림 1(a)는 데이터 수집(Data Collection)을, 그림 1(b)는 데이터 전처리(Data Preprocessing)을, 그림 1(c)는 머신러닝

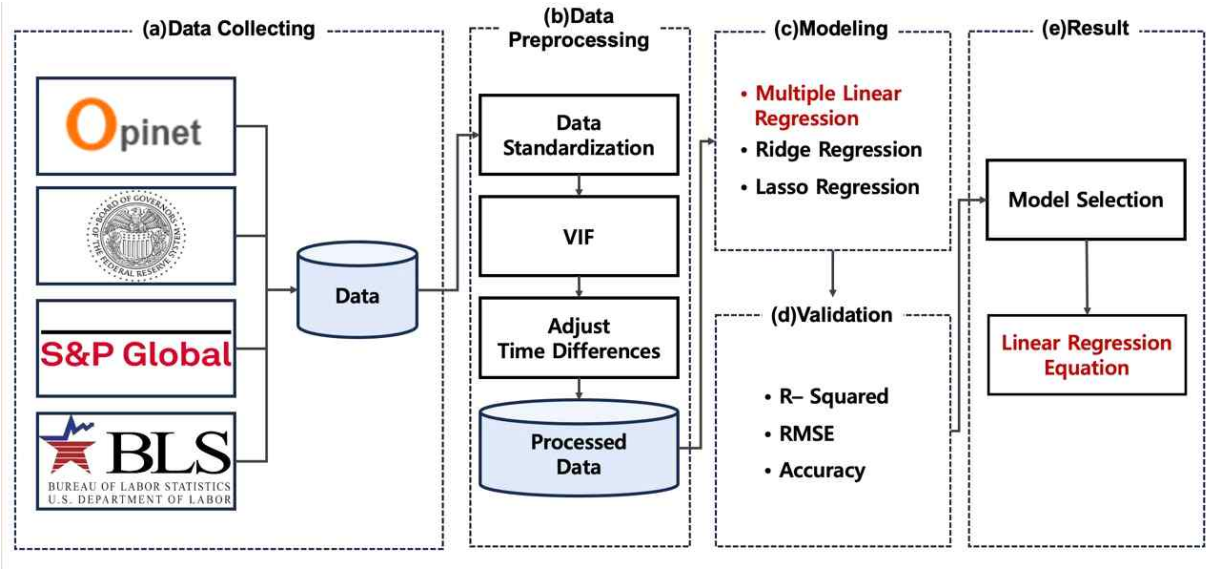

그림 1. 국제 휘발유 가격 예측 모델 : 다양한 전세계적 요인을 수집하고 데이터를 전처리 한 후 모델을 설계 및 검증한다.
Figure 1. Flow diagram of global oil price prediction model: We collect various global factors, preprocess, model and validate the data.

기반의 회귀 모델을 활용한 모델링(Modeling)을, 그림 1(d)는 모델링의 성능 검증(Validation)을, 그림 1(e)는 본 연구에서 제안하는 모델의 결과(Result)를 의미한다.

1. 데이터 수집(Data Collection)

본 연구에서는 2008년 7월 ~ 2022년 12월의 물가, 증권, 통화 정책 등과 관련된 21개의 변수로 구성된 174개의 월간 수치 데이터를 활용한다.

2. 데이터 전처리(Data Preprocessing)

전체 변수 중 유의한 변수를 탐색하기 위해 데이터 표준화(Data Standardization), 시차 반영, 분산팽창지수(Variance Inflation Factor, VIF)를 활용한 유의변수 탐색의 전처리를 수행하였다.

수집한 데이터를 그대로 모델에 활용하는 경우 큰 척도나 범위를 가지는 데이터에 의해 편향된 모델을 도출할 수 있다. 데이터의 편향 문제를 해결하여 정확한 예측 모델을 도출하기 위해 독립변수의 값의 범위를 평균 0, 표준편차 1인 정규 분포 형태로 데이터 표준화를 수행하였다[12].

표준화 한 독립변수 중 유의 변수를 탐색하기 위해 VIF를 활용하여 다중 공선성 기준 유의성 평가를 진행하였다. VIF 기준 10 이하인 변수만 도출하였다. 보통 10 이하인 경우 변수 간의 독립성이 보장되며, 유의한 독립변수라고 해석한다[13]. VIF 값이 가장 높은 변수를 반복적으로 제거하는 과정을 반복해 유의한 독립변수를 획득하였다. 모델에 사용할 유의한 변수는 지정학적 위험지수(Geopolitical Risk Index, GPR), Dollar Index, 미국 수입 석유 물가지수(U.S. Import petroleum and petroleum products), 브렌트유 선물 가격, 미국 기준 금리(The U.S. Benchmark interest rate), 다우존스 산업지수(Dow Jones Industrial Average), S&P 500(Standard & Poor's 500 stock)의 7개 변수이다. 표 1은 본 연구에서 사용한 독립변수의 VIF를 보여준다.

정확한 국제 휘발유 가격 예측을 위해 각 독립변수가 국제 휘발유 가격에 영향을 미치는 적합한 시점을 파악하였다. Dollar Index와 미국 수입 석유 물가지수에 각각 2개월, 1개월의 시차를 반영하였다[14, 15]. 달러의 가치를 나타내는 Dollar Index는 미국의 수입 석유 물가에 대한 예측력을 가진다. Dollar Index가 상승하면 미국 달러

표 1. 각 독립변수의 분산팽창지수
Table 1. Variance Inflation Factor of Independent variables

t is monthly from July 2008 to December 2022

독립변수	VIF
GPR	1.3
Dollar Index	5.8
미국 수입 석유 물가지수	8.0
다우존스 산업지수	2.4
미국 기준금리	1.8
브렌트유 선물 가격	8.3
S&P 500	4.7

의 가치가 상승하며, 이에 따른 영향으로 미국의 수입 석유 물가가 하락한다[14]. 물가지수와 국제유가는 1개월 및 2개월의 시차를 바탕으로 변화한다[15]. 즉, 국제유가 변동의 원인이 되는 물가지수와 국제유가 간의 1개월의 시차를 가진다고 가정하였을 때, Dollar Index는 물가지수를 선행하기 때문에 국제유가와 2개월의 시차를 가진다.

3. 모델링 및 검증(Modeling and Validation)

모델링을 위해 전체 2008년 7월 ~ 2022년 12월의 총 174개 월간 데이터 셋에서 훈련 데이터(Training sets)와 테스트 데이터(Test sets)를 각각 80%, 20%의 비율로 나누어 구성하였다. 다중선형회귀와 Ridge 회귀, Lasso(Least Absolute Shrinkage and Selection Operator) 회귀의 성능 비교를 통해 규제항의 여부에 따른 모델의 성능을 분석하였다. Ridge 회귀, Lasso 회귀의 규제항의 가중치는 각각 0.01, 0.1, 1, 10으로 설정하였다.

최적의 모델을 찾기 위해 결정 계수(R-squared), 평균제곱근 오차(Root Mean Squared Error, RMSE), 정확도(Accuracy)를 측정하였다. R-squared는 종속변수에 대한 독립변수의 설명력을 정량적으로 파악하기 위해 사용하는 평가 지표이다. R-squared 값이 1에 가까울수록 설명력이 좋은 모델이다[16]. RMSE는 모델의 예측력을 평가하는 평가 지표로, 실제값과 예측값 간의 차이를 기준으로 값을 산출한다. RMSE의 값이 작을수록 실제값과 예측값 사이의 차이가 작다는 것으로 해석할 수 있으며, 이는 모델의 예측 정확도가 높다고 판단할 수 있다. 표 2는 다중선형회귀와 Ridge 회귀, Lasso 회귀의 성능 평가 결과이다. Ridge 회귀와 Lasso 회귀의 규제항의 크기와 관계없이, 모든 모델에서 다중선형회귀보다 낮은 성능을 보였다. 이는 독립변수의 개수가 적은 회귀 모델에서는 규제항에 큰 영향을 받아 모델의 성능이 하락할 수 있다는 것을 의미한다. 다중선형회귀 모델의 Accuracy는 훈련 데이터셋 96.93%, 테스트 데이터셋 96.73%로, 상당히 높은 정확도를 보였다. 따라서 국제 휘발유 가격 예측을 위한 최적의 모델로 다중선형회귀 모델을 채택하였다.

Ⅳ. 결과 분석

본 연구에서는 다중선형회귀 모델을 통해 도출한 회귀식을 활용하여 국제 휘발유 가격에 영향을 미치는 변인을 파악하고, 실제 각 변인의 중요도를 파악하였다[3]. 수식 (1)은 다중선형회귀를 통해 도출한 선형 회귀식이다. 이 때, 각 독립변수의 가중치는 종속변수의 중요도를 의미한다. 국제 휘발유 가격 예측에 가장 높은 중요도를 보이는 독립변수는 브렌트유 선물 가격이었다. 브렌트유 선물 가격, 미국 수입 석유 물가지수, Dollar Index, 다우존스 산업 지수, GPR은 국제 휘발유 가격과 양의 상관관계를 가지며, S&P 500과 미국 기준 금리는 국제 휘발유 가격과 음의 상관관계를 가졌다.

그림 2는 테스트 데이터 셋의 예측값과 실제값의 차이를 나타낸 그래프이다. x축은 랜덤하게 추출한 테스트 데이터 셋을 실제값 기준으로 정렬하여 새롭게 부여한 Index 번호이며, y축은 해당 Index에 해당하는 국제 휘발유 가격을 의미한다. 테스트 데이터 셋의 실제값을 회색 O 기호로, 예측값을 빨간색 X 기호로 표현하였다. O와 X

표 2. 국제 휘발유 가격 예측 모델 검증 결과
Table 2. Validation results of the international gasoline price prediction model

Models	Weight	R-Square		RMSE		Accuracy(%)	
		Train Sets	Test Sets	Train Sets	Test Sets	Train Sets	Test Sets
Multiple Linear	-	**0.9395**	**0.9357**	**46.3425**	**48.8712**	**96.93**	**96.73**
Ridge	0.01	0.9395	0.9357	46.3425	48.8784	96.93	96.73
	0.1	0.9395	0.9356	46.3435	48.9434	96.93	96.73
	1	0.9392	0.9337	46.4390	49.6300	96.91	96.63
	10	0.9292	0.9133	50.1057	56.7821	96.40	95.57
Lasso	0.01	0.9395	0.9357	46.3425	48.8790	96.93	96.73
	0.1	0.9394	0.9355	46.3470	48.9468	96.92	96.72
	1	0.9382	0.9328	46.8235	49.9931	96.86	96.58
	10	0.9071	0.8953	57.3798	62.3797	95.24	94.62

$$
\begin{aligned}
International\ Gasoline\ \mathrm{Price}(t) = & +131.5 \times Brent\ crude\ oil\ price\ (t) \\
& +85.3 \times U.S.Import\ petroleum\ \text{and}\ petroleum\ products\ (t-1) \\
& +76.5 \times Dollar\ Index\ (t-2) \\
& -53.8 \times S\&P\ 500\ (t) \\
& +39.0 \times Dow\ Jones\ Industrial\ Average\ (t) \\
& -26.4 \times U.S.\ benchmark\ Interest\ rate(t) \\
& +13.2 \times GPR\ (t)
\end{aligned} \tag{1}
$$

t is monthly from July 2008 to December 2022

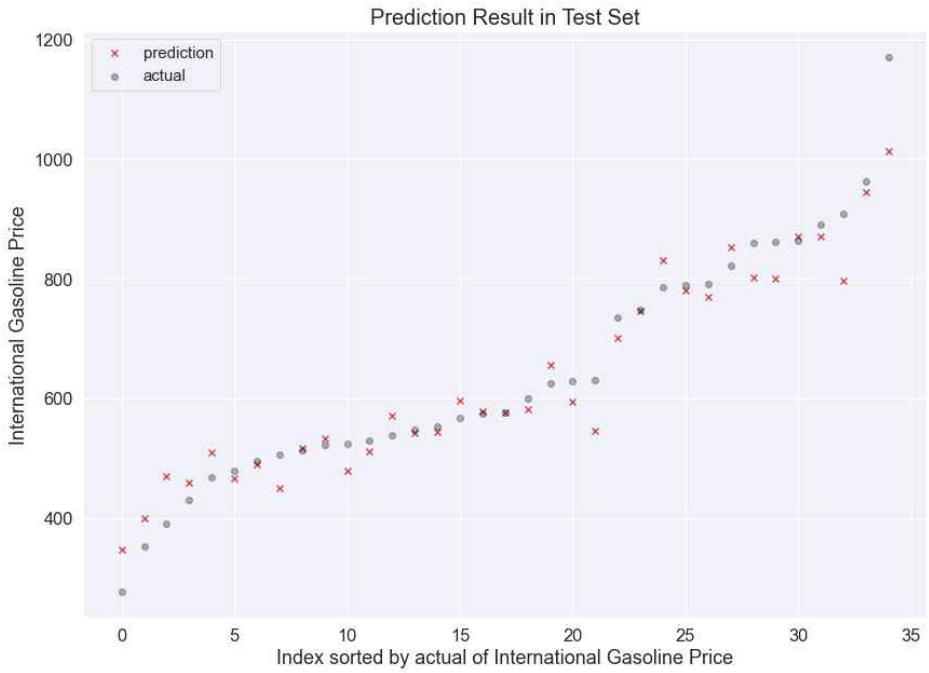

그림 2. 선형회귀 분석의 테스트셋 데이터 결과: 국제 휘발유 가격에 대한 선형회귀 분석 결과. 빨간색 X는 예측값을 나타내고 회색 O는 실제 값을 나타낸다.

Figure 2. The test dataset results of the linear regression: The results of the linear regression for the International gasoline price. Red X represents predictions and gray O represents actual values.

의 차이가 작을수록 해당 값을 제대로 예측한 결과이다. 이를 통해, 회색으로 표현된 실제값과 빨간색으로 표현된 예측값의 추세가 유사한 것을 알 수 있다. 또한, 일부 실제값과 예측값이 겹치는 것을 그림 2를 통해 확인할 수 있다.

V. 결 론

우리는 국제 휘발유 가격을 예측하기 위해 머신러닝 기법 중 다양한 변수를 활용하여 미래의 값을 예측하는 데 효과적인 다중선형회귀 모델을 제안하였다. 이를 위해 물가, 증권, 통화 정책 등과 관련된 다양한 변수를 수집하였다. 변수 간의 척도나 범위의 차이로 인한 데이터 편향 문제를 방지하기 위해 데이터 표준화(Data Standardization)를 수행하였다. 이후, 유의한 독립변수와 국제 휘발유 가격 간의 관계를 더 잘 정의하기 위해 각 독립변수에 시차를 반영하였다. 또한 유의 변수를 탐색하기 위해 분산팽창지수(Variance Inflation Factor, VIF)를 기준으로 유의성을 평가하였다.

시차를 반영한 독립변수를 활용하여 국제 휘발유 가격을 예측하는 제안 모델의 성능을 평가하기 위해 다중선형회귀, Ridge 회귀, Lasso(Least Absolute Shrinkage and Selection Operator) 회귀 모델의 성능을 확인하였다. 모델의 성능은 결정 계수(R-squared), RMSE(Root Mean Squared Error), 정확도(Accuracy)로 검증하고 평가하였다. 성능 평가 결과, 다중선형회귀 모델이 가장 우수한 결과를 보였다. 회귀식을 통해 확인한 국제 휘발유

가격을 예측하는 데에 가장 높은 중요도를 보이는 독립변수는 브렌트유 선물 가격이었다.

그러나 본 연구에서는 월 단위 데이터를 사용하였다는 한계점이 있다. 실시간으로 변화하는 국제 휘발유 가격의 경우 더욱 정확한 예측을 위해서는 월 단위보다 작은 일 단위 혹은 시간 단위의 데이터가 필요하며, 경제지표 또한 세분화 된 데이터가 필요하다. 또 다른 한계점은 전 세계적인 전염병 확산과 같은 이례적인 현상의 영향에 대해 예측하는 것에 한계가 존재하였다. 따라서 향후 실시간으로 변화하는 국제 휘발유 가격을 더 정확하게 분석하고 예측하기 위해서는 세분화된 데이터 수집과 함께 데이터 증강 기법을 통한 모델의 설계가 필요하다.

References

[1] Korean Statistical information Service (https://kosis.kr/index/index.do)

[2] Hyundai Economic Research Institute, *Report on the Domestic Inflationary Impact of International Oil Price Increases*, 2018

[3] S. Yoon and M. Park, “Media-based Analysis of Gasoline Inventory with Korean Text Summarization,” *The Journal of the Convergence on Culture Technology(JCCT)*, Vol. 9, No. 5, pp. 509-515, Sep 2023. DOI:10.17703/JCCT.2023.9.5.509

[4] A. K. Singh, N. Kumar, A. Amardeep, and P. Kumar, “A Comparative Study on the Performance and Emission Analysis of a Dual Fuelled Diesel Engine with Karanja Biodiesel and Natural Gas,” *The International Journal of Advanced Culture Technology(IJACT)*, Vol. 4. No. 1, pp. 10-18, 2016. DOI:https://doi.org/10.17703/IJACT.2016.4.1.10

[5] S. Pyo, M. Jo, G. Lee, and M. Jung, "Interrelationship analysis between crude oil price and the world economic indices," *Korean Institute Of Industrial Engineers*, pp. 412-416, 2012.

[6] H. Kim, “Analysis of Changes in Petroleum Product Price Determination Structure,” *Korea Energy Economics Institute*, Vol. 09, No. 03, 2009.

[7] C. Morana, “A semiparametric approach to short-term oil price forecasting,” *Energy Economics*, Vol. 23, No. 3, pp. 325-338, May 2001. DOI:10.1016/S0140-9883(00)00075-x

[8] S. Yousefi, I. Weinreich, and D. Reinarz, “Wavelet-based pre-diction of oil prices,” *Chaos Solitons and Fractals*, Vol. 25, No. 2, pp. 265-275, Jul 2005. DOI:10.1016/j.chaos.2004.11.015

[9] K. Song and H. Yang, “A Study on the Nymex WTI Prices Forecasting Using Time Series Analysis,” *Journal of Korean official statistics*, Vol. 10, No. 1, pp. 4, 2005.

[10] K. Park, T. Hou, and H. Shin, “Oil Price Forecasting Based on Machine Learning Techniques,” *Journal of Korean Institute of Industrial Engineers*, Vol. 37, No. 1, pp. 64-73, Mar 2011.

[11] S. Kim and D. Cho, “Forecasting Crude Oil Prices with Google Trends Data Based on Machine Learning Methods,” *The Korean Journal of Economics*, Vol. 29, No. 2, pp. 175-193, Dec 2022. DOI:10.46228/kje.29.2.3

[12] M. Gal, D. Rubinfeld, “Data Standardization,” *NYU Law and Economics Research Paper,* Vol. 17, No. 19, Jun 2019. DOI:10.2139/ssrn.3326377

[13] D. E. Farrar, and R. R. Glauber, “Multicollinearity in Regression Analysis: The Problem Revisited,” *The Review of Economics and Statistics*, Vol. 49, No. 1, pp. 92-107, Feb 1967. DOI:10.2307/1937887.

[14] J. Kim and S. Kim, “A Study on the Effect of Won dollar Exchange Rate Changes on Export Price and Import Price,” J*ournal of CEO and Management Studies,* Vol. 20, No. 1, pp. 55-68, Apr 2017.

[15] S. Yoon and H. Jeon, “Consumer Price Outlook and Implications through the Lens of International Oil Prices,” *Insurance Research Institute*, 2022.

[16] S. Lee and K. Cho, “Prediction of Solar Photovoltaic Power Generation by Weather Using LSTM,” *Journal of the Korea Society of Computer and Information*, Vol. 27, No. 8, pp. 23-30, Aug 2022.