ILI 예측모델 개발 프로젝트
: 인플루엔자 B의 예측을 통한 의약품 재고 최적화

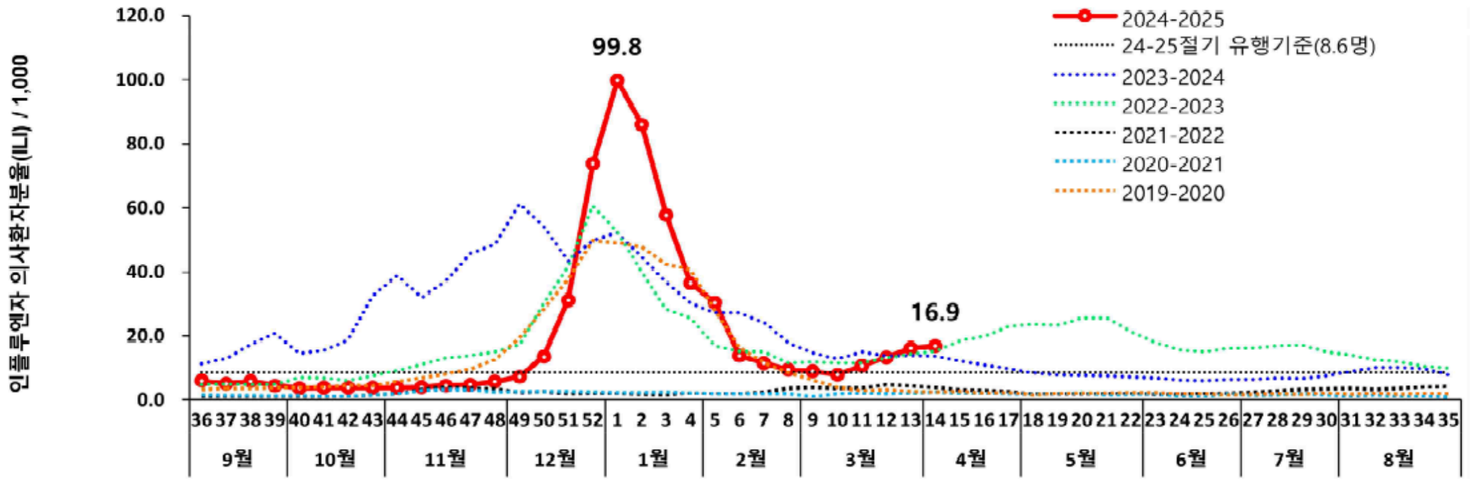
중앙대학교 SCM 분석 학회 씀씀이 B팀
강다훈 김희연 김태형 이정우 유현승 윤설리 한예호 전주현

CONTENTS

01

인플루엔자 B형 유행 예측의 필요성

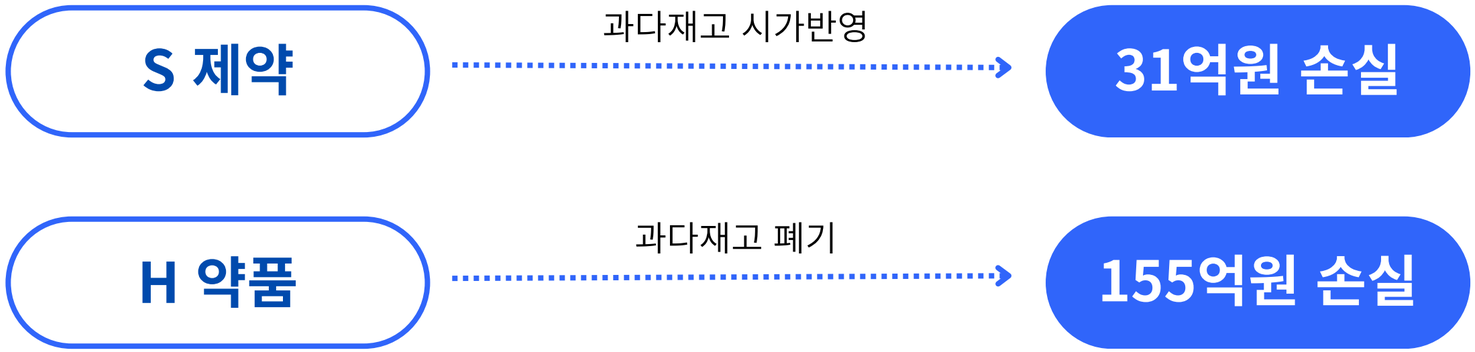
- 전 세계 계절성 인플루엔자: 매년 약 10억 건 이상 발생
- 2025년 국내 ILI(의사환자) 비율: 외래환자 1,000명당 평균 8.6명
- 인플루엔자 A형과 달리 봄철(3~4월) 학령기 아동·청소년 중심 확산



▲ 질병관리청 보도자료_인플루엔자 의사환자 분율 (2016~2025년 14주차)

01

인플루엔자 B형 유행 예측의 필요성



국내기업: 과거 판매 실적과 단순 계절성 패턴을 기반으로 수요 예측

비선형적이고 변동성이 큰 인플루엔자의 특성 반영 한계

02

분석 목표 및 예측모델 구축 전략

분석 목표:

ILI 데이터를 바탕으로 인플루엔자 B형 유행 강도와 시기를 정밀하게 예측하는 AI 모델 구축 및 SCM 적용

예측모델 구축 전략:

1. 데이터 수집 및 시각화

- 시계열 기반 데이터 전처리
- 변수 상관성 분석

2. 예측 모델링

- SARIMAX
- LSTM

3. 모델 성능 평가

- 회귀(R^2, MSE, MAE)
- 분류(Accuracy, F1, Recall)

4. 결과 해석

- 변수 간 상관관계 분석
- 주요 인자 도출
- 정책 적용 가능성 평가

03

선행 연구 고찰 및 적용 가능성 평가

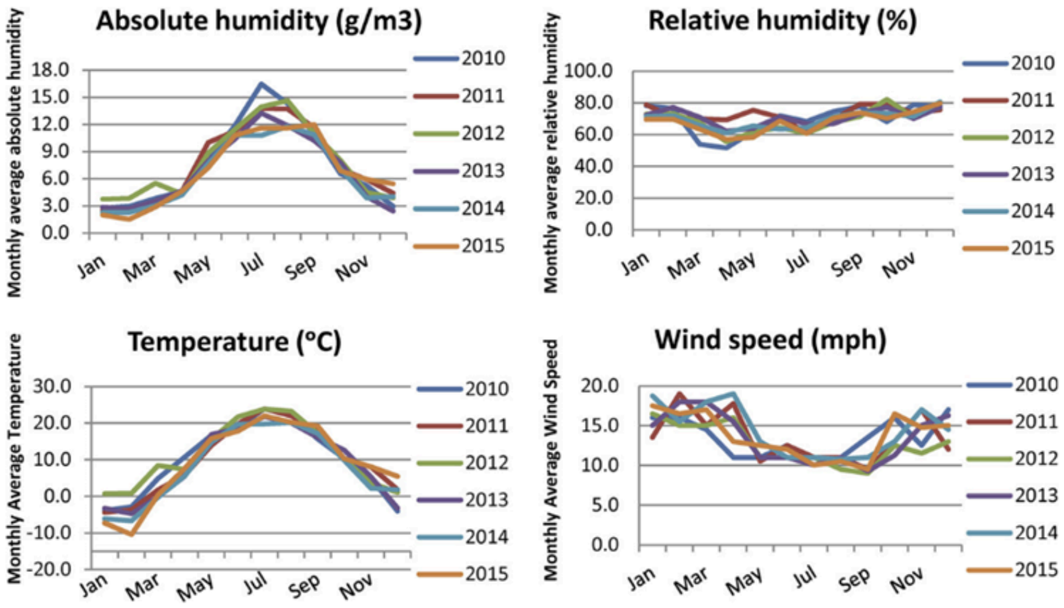
- 연구 1. 캐나다 토론토 지역 기반 인플루엔자 B형 분석

- 연구 2. LightGBM을 사용한 SHAP 기반의 설명 가능한 인플루엔자 발생 예측

- 연구 3. FORECASTING SEASONAL INFLUENZA-LIKE ILLNESS IN SOUTH KOREA AFTER 2- AND 30-WEEKS USING GOOGLE TRENDS AND INFLUENZA DATA FROM ARGENTINA

03 선행 연구 고찰 및 적용 가능성 평가

- **연구 1. 캐나다 토론토 지역 기반 인플루엔자 B형 분석**



▲토론토 절대습도, 상대습도, 기온 및 풍속

연구 개요

- 인플루엔자 B형의 유행과 주요 기상 변수 간의 관계 분석
- 캐나다 토론토, 2010~2015년 데이터 반영
- 기상변수(기후) 데이터 : 기온, 절대습도, 상대습도, 풍속
- 기온, 절대습도, 상대습도, 풍속

03 선행 연구 고찰 및 적용 가능성 평가

- **연구 1. 캐나다 토론토 지역 기반 인플루엔자 B형 분석**

TABLE 6 Adjusted RH nonlinear negative binomial regression models exploring the relationship of environmental factors and nonlinearity of RH with influenza A and B viruses[a]

	All influenza viruses			Influenza A virus			Influenza B virus		
Demographic or climatic factor	IRR (95% CI)	*P* value	Nonlinearity *P* value	IRR (95% CI)	Association *P* value	Nonlinearity *P* value	IRR (95% CI)	*P* value	Nonlinearity *P* value
Age group (yr)									
65+	1.00	NA	NA	1.00	NA	NA	1.00	NA	NA
<1	0.67 (0.53-0.85)	<0.0001*	NA	0.69 (0.50-0.95)	0.0024*	NA	0.73 (0.49-1.10)	<0.0001	NA
1–4	1.32 (1.07–1.63)	<0.0001*	NA	1.18 (0.88–1.59)	<0.0001	NA	1.89 (1.33–2.67)	<0.0001*	NA
5–19	2.42 (1.95–3.01)	<0.0001*	NA	1.82 (1.33–2.48)	0.005*	NA	4.32 (3.07–6.09)	0.0037*	NA
20–64	1.19 (1.01–1.42)	0.0012*	NA	1.22 (0.96–1.54)	0.0247	NA	1.33 (0.99–1.78)	0.1316	NA
Outbreak status									
Yes	1.00	NA	NA	1.00	NA	NA	1.00	NA	NA
No	0.27 (0.23–0.32)	<0.0001*	NA	0.23 (0.19–0.29)	<0.0001*	NA	0.30 (0.22–0.39)	<0.0001*	NA
RH[b]	NA	0.0013*	0.0056*	NA	<0.0001*	0.4923	NA	<0.0001*	<0.0001*
Temp	NA	<0.0001*	<0.0001*	NA	<0.0001*	<0.0001*	NA	<0.0001*	<0.0001*
WS	1.00 (0.98–1.02)	0.8200	NA	1.00 (0.97–1.03)	0.8879	NA	0.99 (0.95–1.02)	0.6061	NA
Temp fluctuation	1.03 (1.01–1.05)	<0.0001*	<0.0001	0.99 (0.97–1.01)	<0.0001	NA	1.09 (1.06–1.11)	<0.0001*	NA

[a]RH, relative humidity; WS, wind speed; IRR, incidence relative risk, CI, confidence interval; NA, the measurement is not applicable for that variable. The RH nonlinear regression model explored the association of environmental factors with influenza activity as well as the nonlinearity of the association for RH and temperature with influenza A and B viruses and all influenza viruses. The left column lists independent/predictable variables for which this model was adjusted. The total weekly numbers of positive influenza A and B virus counts were used as dependent variables. A significant result (*) for association is considered when the 95% confidence interval does not cross 1 and the *P* value is <0.05. A significant result for nonlinearity is considered when the *P* value is <0.05. The incidence relative risk of 1.00 indicates the category used for reference/comparison. AH, temperature, and WS were measured by the use of weekly median measurements. Influenza A virus and influenza B virus represent the total weekly numbers of positive specimens. All influenza viruses represent the sum of influenza A and B virus-positive specimens.
[b]RH was also examined for a nonlinear association with influenza A and B viruses.

▲ 비선형성 회귀 모형 결과

연구 결과

- **온도 변동:**
  일교차가 클수록 인플루엔자 B형 감염자 수가 유의하게 증가
- **절대 및 상대습도:**
  절대습도 10.5g/m^3, 상대습도 60%를 초과하면 확진자 수가 급격히 감소
- **연령:**
  5~19세 학령기 아동 및 청소년의 상대 위험도(IRR: 4.32) 가장 높음
- **기온:**
  15°C를 넘는 구간부터 B형 확진자 수 급감

03 선행 연구 고찰 및 적용 가능성 평가

- **연구 2. LightGBM을 사용한 SHAP 기반의 설명 가능한 인플루엔자 발생 예측**
  - 국내 데이터를 바탕으로 LightGBM 및 SHAP(Shapley Additive Explanations)를 활용해 인플루엔자 발생을 예측
  - 타국 ILI 환자수와 머신러닝 알고리즘 기반
  - SHAP 분석을 통해 각 변수의 중요도를 직관적으로 해석할 수 있도록 설계
  - 국내 실존 데이터를 활용해 로컬라이징 용이
  - 타겟 변수로 A형과 B형을 구분하지 않고 전체 인플루엔자 및 유사 질환을 통합하여 분석

→ B형 고유의 계절성이나 민감한 환경 반응을 반영한 모델 구성 불가능

→ B형 특성에 대한 근본적인 통찰을 얻기에는 한계

03 선행 연구 고찰 및 적용 가능성 평가

- **연구 3. Forecasting seasonal influenza-like illness in South Korea after 2- and 30-weeks using Google Trends and influenza data from Argentina**

  - ARMAX 및 SARIMAX와 같은 통계 기반 시계열 기법 활용
  - 아르헨티나의 인플루엔자 데이터를 외부 변수로 삼아 대한민국 내 인플루엔자 발생을 장기(30주 후) 예측하는 모델을 설계
  - Rolling Window 기법 등 시계열 모델링 측면에서 유용한 아이디어를 제공
  - Google Trends 등 해외 검색 데이터 활용 → 데이터 해석 및 구조 차이로 인해 재구성 필요
  - 타겟 변수로 A형과 B형을 구분하지 않고 전체 인플루엔자를 통합하여 분석

  → B형 고유의 계절성이나 민감한 환경 반응을 반영한 모델 구성 불가능

  → B형 특성에 대한 근본적인 통찰을 얻기에는 한계

04

예측 모델 설계 및 변수 구성 전략

- 예측 목표

INF_B 환자수 예측 → 인플루엔자 B형의 유행 여부 및 시기 판단

↓

의약품 재고 및 생산 계획에 반영

↓

제약기업의 재고 손실과 공급망 리스크 최소화

↓

인플루엔자 B형의 유행 여부 및 시기 판단

- 유행 기준 정의

유행 기준 (주차 t) : 최근 3년간 주별 INF_B 평균 + (1 × 표준편차)

→ 예측된 INF_B가 이 기준을 초과할 경우 “유행”으로 간주

*미국 CDC의 권고안을 준용

04

예측 모델 설계 및 변수 구성 전략

- 분석 단위 및 기간

예측 단위	주 단위 (Weekly)
학습 기간	2014년 1월 ~ 2024년 4월(코로나 기간 제외)
평가 기간	2024년 5월 ~ 2025년 4월

- 모델링 접근 전략

SARIMAX: 계절성과 외생 변수 처리가 가능한 전통 통계 기반 모델

LSTM: 시계열 내 장기 의존성과 비선형적 특성을 반영할 수 있는 딥러닝 기반 모델

04

예측 모델 설계 및 변수 구성 전략

- 최종 사용 변수

한국 ILI	국내 INF_B 주간 환자수
타국 ILI	선행 시그널
평균 기온	국내 서울 기준 주간 기온
일교차	평균 일일 기온 차
상대 습도	주간 평균 상대습도

04

예측 모델 설계 및 변수 구성 전략

- **제외된 변수 및 사유**

강수량	기존 문헌에서 상관성 낮음, 예측 타당성 부족
연령별 독감 데이터	A/B형 미구분 데이터로 예측 정확도 저하 우려
백신 접종률	연 단위만 존재하여 주단위 예측과 시간 불일치
검색량 지표	주 단위 누락 및 불연속성으로 분석 정확도 저하 우려

*Google Trends, Naver

05 모델링 | 사용 데이터

본 프로젝트는 2가지 주요 범주의 데이터 활용하여 예측 모델 구성

INF_B 전처리 데이터

- 대상: 전 세계 국가별 B형 인플루엔자 주간 발생률
- 주요 컬럼
  - COUNTRY: 국가명
  - DATE, YEAR, WEEK, YEAR_WEEK : 날짜정보
  - INF_B: 해당 주차의 INF_B 건수
  - occur_t-0 ~ occur_t-52: 시점 t 기준 직전 0~52주의 lag 값

국내 기상 변수

- 대상: 주간 단위로 집계된 기후 정보
- 주요 컬럼
  - 일자: 주차 날짜 (YYYY-MM-DD)
  - 평균 기온, 일교차, 상대습도

05 모델링 | 데이터 전처리

COVID-19의 구조적 비정상성 제거를 위해 제외 기간 선정

사용 기간 및 제외 기간

- 사용 기간 : 2014.01 ~ 2025.05
- 제외 기간 : 2020~2022년

주요 전처리 항목

- 모든 시계열에 대해 YEAR_WEEK 생성
- INF_B의 결측은 0으로 대체
- 국가별 row 수를 동일하게 맞추기 위해 padding 수행
- occur_t-0부터 occur_t-52까지의 lag 변수 생성
- 모든 시계열 변수는 MinMax Scaling을 통해 정규화하여 비교 가능하도록 조정

05 모델링 | 국가 간 ILI 유사도 분석

분석 방식

외국의 ILI 시계열을 예측 모델의 외생 변수로 활용하기 위해
코사인 유사도(Cosine Similarity)와 **동적 시계열 거리(Dynamic Time Warping, DTW)** 이용해 각 국가별 유사도 정량적 측정

코사인 유사도

한국 INF_B 시계열 기준으로, 각 국가의 occur_t-0부터 occur_t-52까지의 lag 시계열과의 유사도 계산

- 특징 : 시계열 크기보다 변화 방향에 민감, 시점 동일할 경우 패턴 유사성 효과적으로 측정

- 목적 : 한국과 **동일한 시점에서 유사한 추세** 보이는 국가 식별하여, 예측 모델의 외생 변수로 활용

DTW 거리

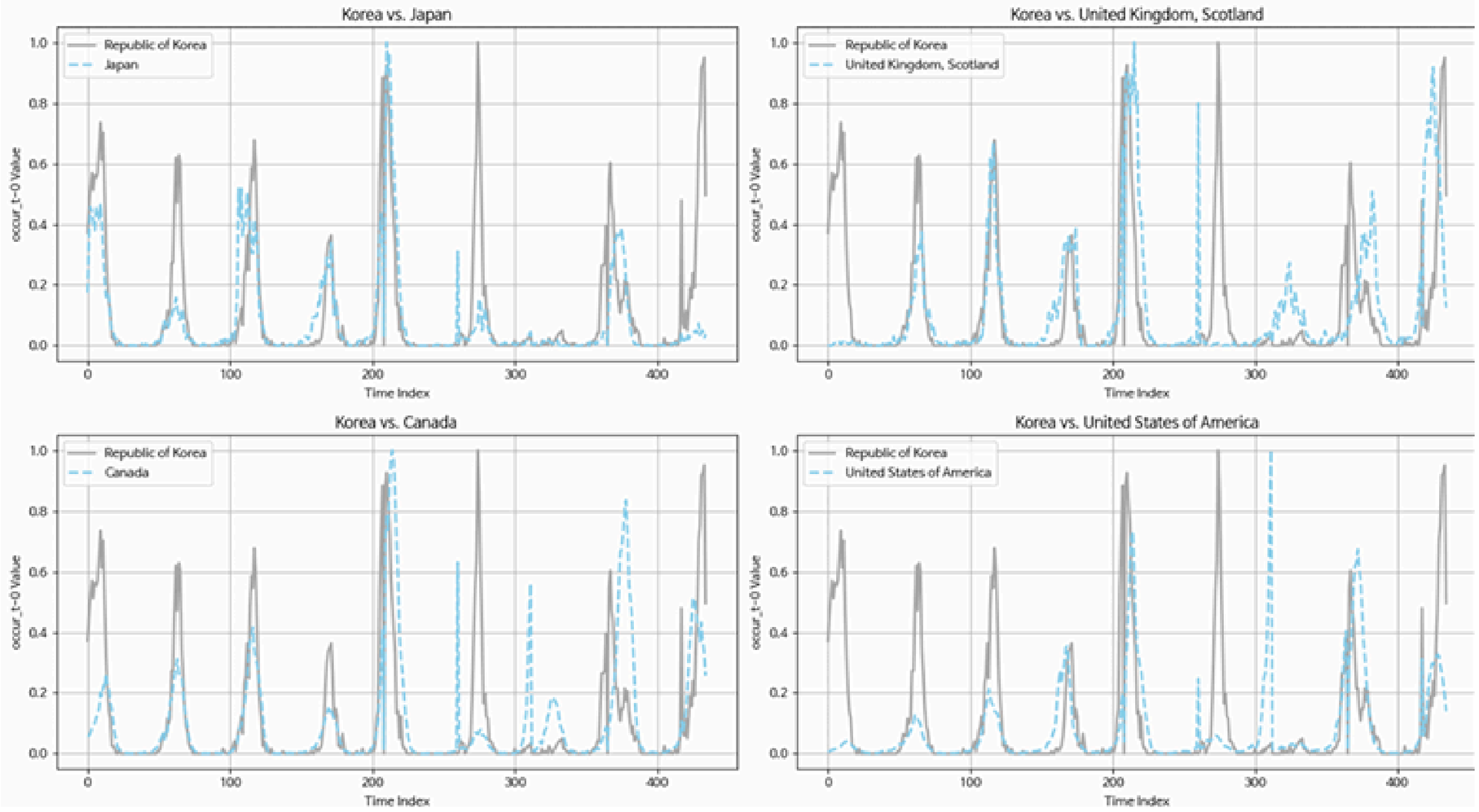
한국 시계열과 각 국가의 시계열 간 시간 왜곡을 허용한 상태에서 전체 패턴 유사성 측정
DTW 거리 값 계산 후 이를 0~1 범위로 정규화

- 특징 : 시점이 어긋난 유사한 패턴 비교 가능, 전체적인 곡선 구조의 유사성 반영

- 목적 : **유사한 패턴 보이되 시점 차이가 존재**하는 국가 선별하여, 선행 예측 신호로 활용 가능한 외생 변수 후보 식별하는데 사용

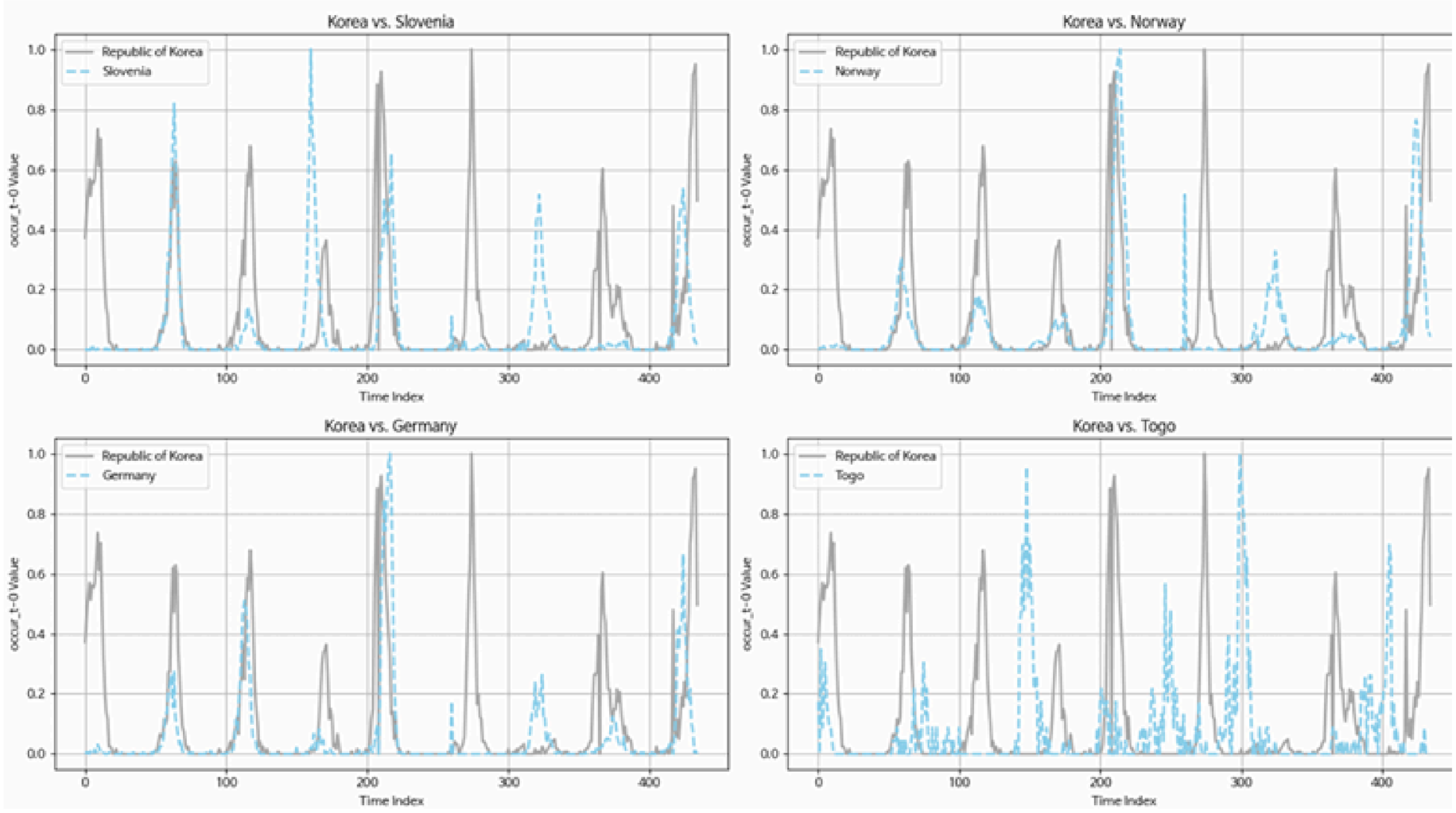
05 모델링 | 국가 간 ILI 유사도 분석

DTW 상위 10개국 vs 한국 ILI 시계열 플롯 : 일본, 영국 스코틀랜드, 캐나다, 미국



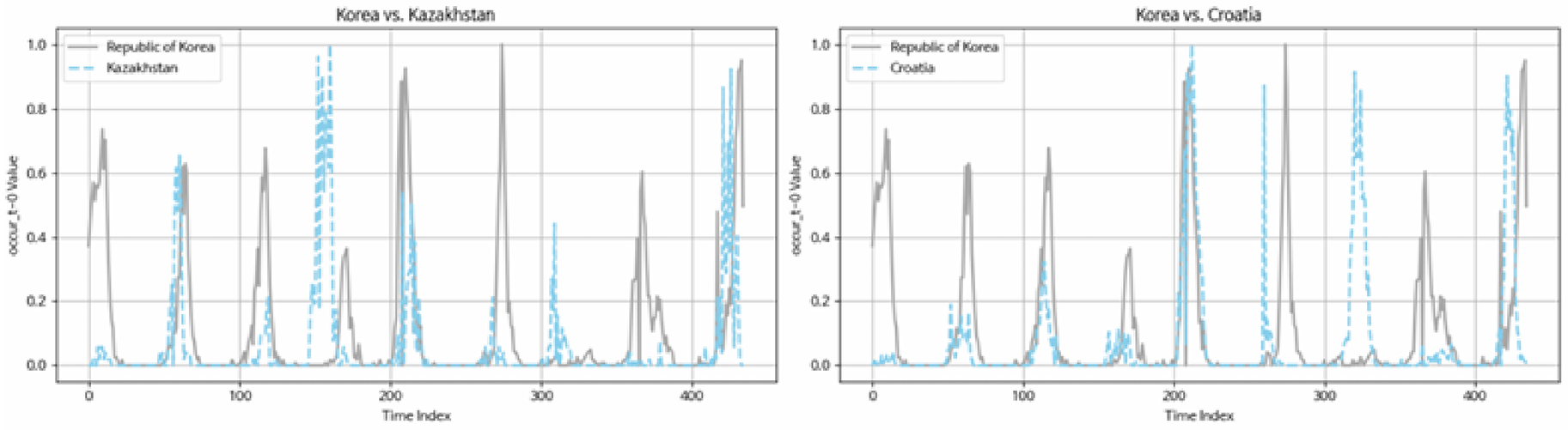
05 모델링 | 국가 간 ILI 유사도 분석

DTW 상위 10개국 vs 한국 ILI 시계열 플롯 : 슬로베니아, 노르웨이, 독일, 토고



05 모델링 | 국가 간 ILI 유사도 분석

DTW 상위 10개국 vs 한국 ILI 시계열 플롯 : 카자흐스탄, 크로아티아



최종 점수 (final_score)

- final_score = cosine_similarity - scaled_dtw
- **유사도가 높고 거리 차가 적은 국가**가 우선 선별

05 모델링 | 국가 간 ILI 유사도 분석

주요 결과 (상위 10개국)

순위	국가명	final score	순위	국가명	final score
1	Japan	0.7367	6	Norway	0.3606
2	Canada	0.6218	7	Turkey	0.3582
3	UK, Scotland	0.5778	8	Slovenia	0.3288
4	USA	0.4252	9	Argentina	0.2996
5	China	0.3819	10	Germany	0.2895

05 모델링 | SARIMAX 예측 모델

ADF 정상성 검정 및 계절성 분석 | 정상성 검정

SARIMAX |

시계열 예측 모델로 정상 시계열을 전제. 평균과 분산, 자기상관 구조가 시간에 따라 일정하게 유지되는 시계열

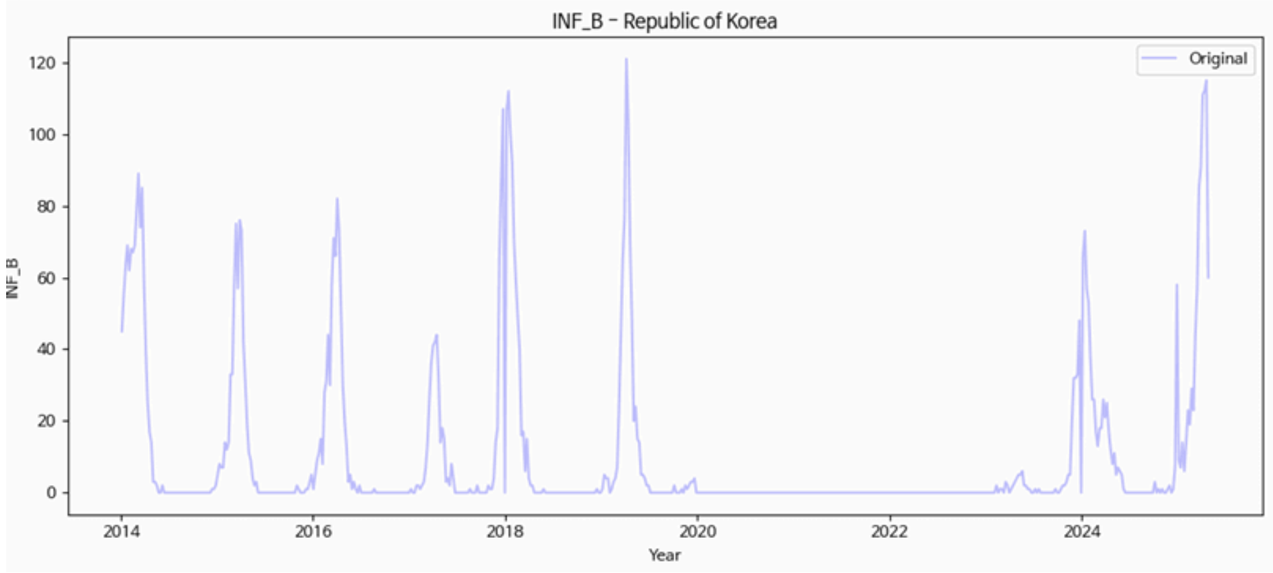
ADF 검정 | Augmented Dickey-Fuller

시계열에 단위근이 존재하는 여부 평가하며 이는 곧 비정상성 의미.

- 귀무가설(H_0) : “단위근이 존재한다(비정상 시계열)”
- 대립가설(H_1) : “단위근이 존재하지 않는다(정상 시계열)”

검정 결과

- ADF TEST STATISTIC: -5.8286
- P-VALUE: 0.0000
- 임계값(CRITICAL VALUES):
  - 1%: -3.4456
  - 5%: -2.8683
  - 10%: -2.5704



05 모델링 | SARIMAX 예측 모델

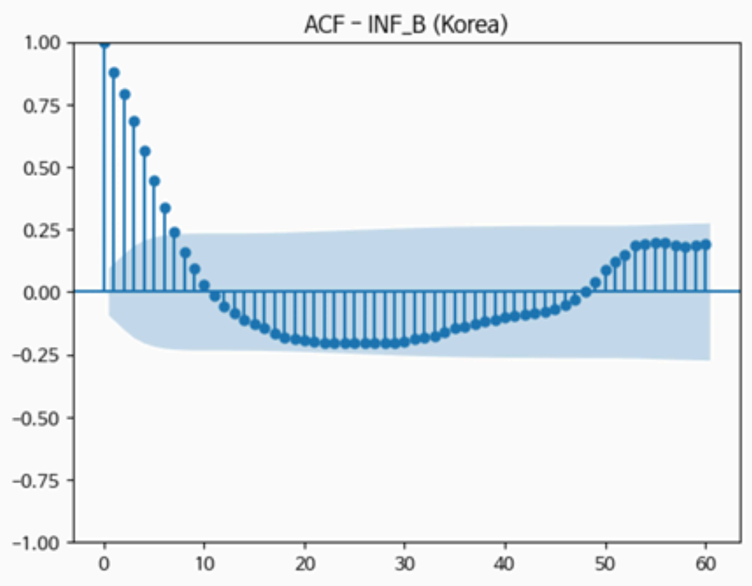
ADF 정상성 검정 및 계절성 분석 | 계절성 분석

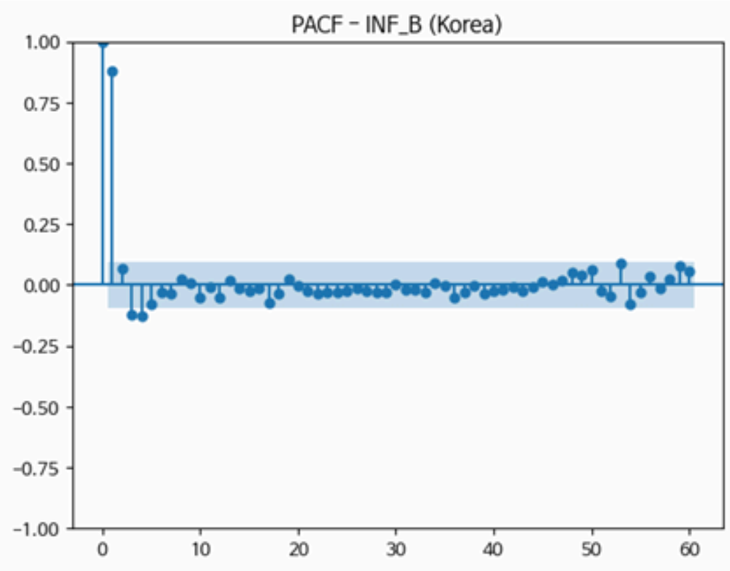
자기상관 함수(ACF) |

ACF : 시차 52주 근처에서 반등하는 형태 → 이는 연 단위의 계절 주기 존재할 가능성 시사

부분 자기상관 함수(PACF) |

PACF : 1차 시점에서 뚜렷한 spike 확인, 이후 급격히 감소하는 패턴 → AR(1) 구조가 적합할 가능성 높음





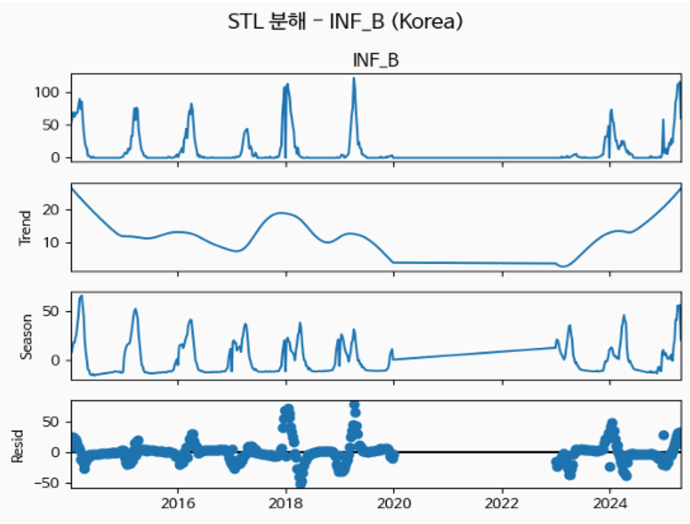
05 모델링 | SARIMAX 예측 모델

ADF 정상성 검정 및 계절성 분석 | 계절성 분석

STL(Seasonal-Trend decomposition using Loess) 분해 분석
시계열이 뚜렷한 계절성. INF_B 원시 시계열은 매년 겨울철 정점 이후 급격히 하강하는 계절적 유행 패턴 반복

OLS 회귀 분석
계절성 정량적 검증 위해 주차 기준으로 52개 더미 변수 설정 → OLS 회귀 분석 실시
전체 주차 중 약 절반 이상 변수에서 유의수준 5% 이하 유의성 → 주차 인플루엔자 발생에 영향 미치는 계절적 요인 입증

종합 : INF_B 시계열은 정상성과 뚜렷한 계절성 보유 → SARIMA 또는 SARIMAX 모델 적용 기반 요건 충족



OLS Regression Results

Dep. Variable:	INF_B	R-squared:	0.404
Model:	OLS	Adj. R-squared:	0.323
Method:	Least Squares	F-statistic:	4.973
Date:	Thu, 10 Jul 2025	Prob (F-statistic):	7.98e-21
Time:	17:16:46	Log-Likelihood:	-1911.6
No. Observations:	435	AIC:	3929.
Df Residuals:	382	BIC:	4145.
Df Model:	52		
Covariance Type:	nonrobust		

05 모델링 | SARIMAX 예측 모델

(p,d,q,S) 설정 및 모델(한국 단독) Fitting | SARIMAX

비계절 모수(p,d,q)

비계절 차수 : ACF 및 PACF 분석 결과 → AR(1) 구조 적절 → (p=1, d=0, q=0)으로 고정

PACF 그래프에서 1차 시점 강한 spike 이후 급격한 감소 → 자기회귀(AR) 성분 1차까지만 유의

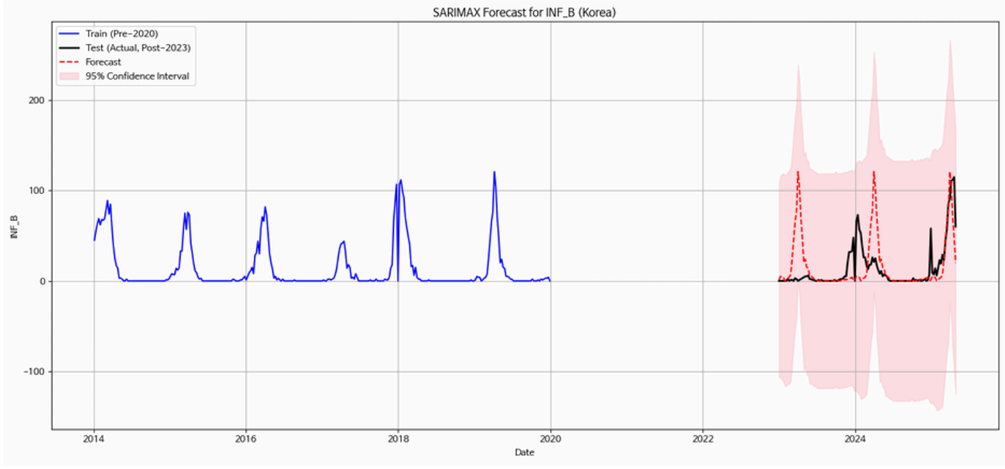
계절 모수(P, D, Q, s)

계절 주기 s=52(주 단위, 1년 기준) 고정, P, D, Q 각각 대해 [0,1] 범위에서 grid search

평가 기준인 Akaike Information Criterion(AIC)가 최소가 되는 조합을 최적값

최종 도출 계절 파라미터 (P=0, D=1, Q=0) → 계절적 변화 대한 1차 차분 필요, 계절성 자체에 AR 또는 MA 요소가 크지 않음

- 모델 : SARIMAX(1,0,0) × (0,1,0,52)
- 학습 기간 : 2014년 ~ 2019년
- 테스트 기간 : 2023년 1월 ~ 2025년 4월
- 평균절대오차(MAE) : 16.13
- 평균제곱오차(MSE) : 약 875.61
- 루트 평균제곱오차(RMSE) : 29.59
- 결정계수(R^2) : -0.4437



05 모델링 | SARIMAX 예측 모델

외생변수(타국ILI) 추가 SARIMAX

- 외생 변수 후보 국가 : 앞서 수행한 국가 간 유사도 분석 결과 바탕으로 선정 (코사인 유사도 및 DTW 거리 기반의 종합 점수 (final_score) 기준으로 유사도 높은 국가 순으로 시계열 차례로 반영)

각 국가별로 최적 시차 고려하여 시계열 이동 → 해당 변수를 exogenous variable로 추가

외생 변수의 선택 방법

1. final score 기반 랭킹 상위 국가를 대상으로 순차적으로 모델 외생변수에 추가
2. 국가별 최적 lag를 적용하여 시계열 shift
3. 외생 변수 계수의 통계적 유의성($p < 0.05$) 확인
4. 예측 성능 지표 RMSE 개선 여부를 기준으로 채택 여부 결정

05 모델링 | SARIMAX 예측 모델

외생변수(타국ILI) 추가 SARIMAX

- 데이터 : 2014년 1월 ~ 2019년 12월까지의 훈련 데이터, 2023년 1월부터 2024년 4월까지 테스트 데이터 (기존과 동일)
  - 팬데믹 기간인 2020~2022년 제외
- 모델 구조 : SARIMAX(1,0,0) × (0,1,0,52)
- 외생 변수로 첫 번째 추가 된 국가는 일본으로 lag 없이 동시 시점에서 INF_B 데이터 적용
  - RMSE : 29.59 → 29.17 (약 1.44%) 감소, 변수 계수 p=0.000
- 두 번째 추가 된 국가 캐나다
  - RMSE : 32.40, 통계적 유의하지 않음 → 최종 모델에 포함하지 않음

단계	추가 국가(lag)	모델 RMSE	R^2 Score	성능 개선 여부	외생변수 유의성(p<0.05)
0	(Baseline)	29.5906	-0.4437	-	-
1	Japan (0)	29.1655	-0.4025	개선 (↓1.44%)	p = 0.000 < 0.05 (유의)
2	Canada (0)	32.4021	-0.7310	성능 저하	추가하지 않음

05 모델링 | SARIMAX 예측 모델

SARIMAX 결론

예측 정확도와 실용적 설명력 확보하기 어려움
유연하게 비선형성과 복합적 요인 반영할 수 있는 LSTM 기반 딥러닝 모델로 분석 전환

시사점

- 일부 국가의 ILI 발생 패턴이 한국의 유행과 일정한 상관성
- 선행 지표로서의 가능성 지님

한계

- 팬데믹에 따른 시계열 연속성 훼손으로 전체 패턴 일관성 약화 → 모델 예측력 크게 감소
- SARIMAX는 선형 추세와 고정된 계절성 가정 → 최근 급격하게 변화한 유행 양상이나 예외적 패턴 반영의 한계
  - 결정계수 : 외생 변수 유무와 관계없이 음수(-0.44 이하) 기록 → 단순 평균값 예측보다 낮은 수준의 설명력

05 모델링 | LSTM 예측 모델

LSTM(Long Short-Term Memory) |
시계열 데이터 내의 장기 의존성과 시점 간 시간 지연 효과를 효과적으로 학습할 수 있는 구조로,
기존의 선형 기반 모델이 다루기 어려운 비정형적 패턴이나 외부 요인의 비선형 결합 등을 학습하는 데 적합한 모델로 평가됨

AS-IS

기존 SARIMAX 모델의 문제점과 한계

비선형성과 구조적 변화는 고정된 계절성과 선형 추세를 전제로 설계된 SARIMAX 모델이 충분히 반영하기 어려운 요인이며, 실제 예측 성능에서도 그 한계가 확인됨

TO-BE

딥러닝 기반의 LSTM 모델 사용

본 프로젝트의 경우, 해외 국가의 ILI 발생 패턴이나 기후 변수 등의 외생 정보를 함께 반영할 필요가 있었기 때문에, LSTM 구조의 적용이 더욱 타당하다고 판단함

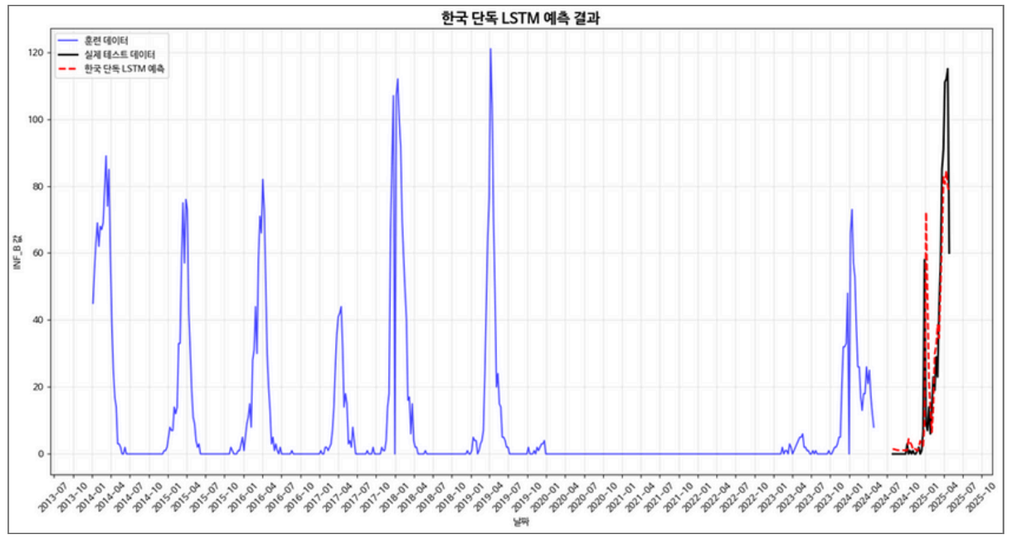
05 모델링 | LSTM 예측 모델

한국 ILI 기반 LSTM 모델 | 예측성능 평가

RMSE: 17.7684 / MAE: 10.2214 / 결정계수 R^2: 0.7406

→ 기존 SARIMAX 모델 대비 크게 향상된 수치로, 모델이 단순히 평균 추세만을 학습하는 것이 아닌 유행 시점과 강도 등 변동성을 일정 수준 이상 반영하고 있음
→ R^2 지표가 양수로 전환되며 0.74를 기록한 점은 모델의 설명력이 통계적으로 유의함을 보여줌

- 입력 시퀀스 길이: 12주
- 예측 대상: INF_B
- 모델 구조: LSTM(64) → DENSE(1)
- 최적화 및 손실 함수: RMSPROP (OPTIMIZER), MSE (LOSS)
- EPOCHS: 80
- BATCH SIZE: 8
- VALIDATION SPLIT: 20%



05 모델링 | LSTM 예측 모델

한국 + 일본, 캐나다, 영국, ... | 북반구 국가 ILI 기반 LSTM 모델의 한계

순위	국가명	final score	순위	국가명	final score
1	Japan	0.7367	6	Norway	0.3606
2	Canada	0.6218	7	Turkey	0.3582
3	UK, Scotland	0.5778	8	Slovenia	0.3288
4	USA	0.4252	9	Argentina	0.2996
5	China	0.3819	10	Germany	0.2895

일본, 캐나다, 영국 등 유사도 상위 국가 기반 모델 시도

❗ 유사도 상위 국가임에도 아르헨티나, 뉴질랜드 등
남반구 국가 기반 모델에 비해 성능 저하 현상

05 모델링 | LSTM 예측 모델

한국 + 일본, 캐나다, 영국, ... | 북반구 국가 ILI 기반 LSTM 모델의 한계

순위	국가명	final score	순위	국가명	final score
1	Japan	0.7367	6	Norway	0.3606
2	Canada	0.6218	7	Turkey	0.3582
3	UK, Scotland	0.5778	8	Slovenia	0.3288
4	USA	0.4252	9	Argentina	0.2996
5	China	0.3819	10	Germany	0.2895

LSTM

- 여러 입력 정보(피처)를 동시에 학습하는 구조
- 입력 변수 증가 → 중요 정보 + 불필요한 정보(노이즈) 섞임

→ 북반구 데이터: 불필요한 분산 유발

→ 남반구 국가 중 한국과 시계열 유사도 상위 국가인 **아르헨티나** 선정

05 모델링 | LSTM 예측 모델

한국 + 아르헨티나(Lag 32) ILI 기반 LSTM 모델 | 예측성능 평가

RMSE: 16.7174 / MAE: 10.6196 / R^2: 0.7926

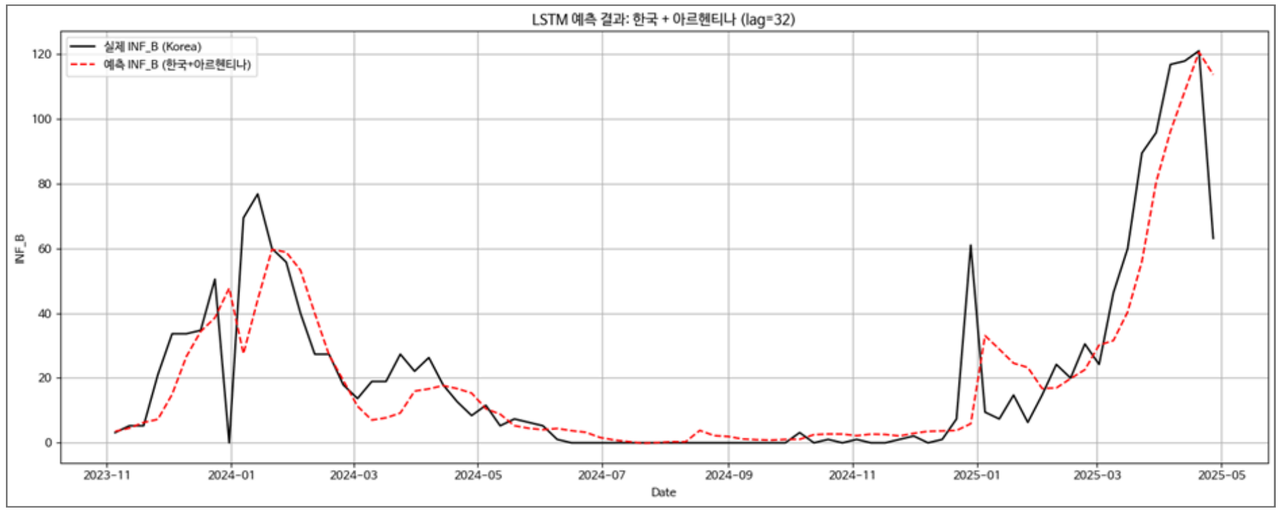
- 다변량 LSTM 모델을 구축 후 실험 결과, 한국 단독 모델보다 낮은 R^2 값을 기록하며 하락

→ 남반구 국가 중 한국과의 시계열 유사도 상위 국가에 주목

- 해당 모델 단일 변수 모델(R^2 = 0.7406) 대비 유의미한 성능 향상

→아르헨티나의 시차 적용된 시계열이 한국 INF_B의 유의미한 선행 지표 역할을 할 수 있음을 시사함

- 입력 시퀀스 길이: 12주
- 예측 대상: INF_B
- 모델 구조: LSTM(64) → DENSE(1)
- 최적화 및 손실 함수: RMSPROP (OPTIMIZER), MSE (LOSS)
- EPOCHS: 80
- BATCH SIZE: 8
- VALIDATION SPLIT: 20%

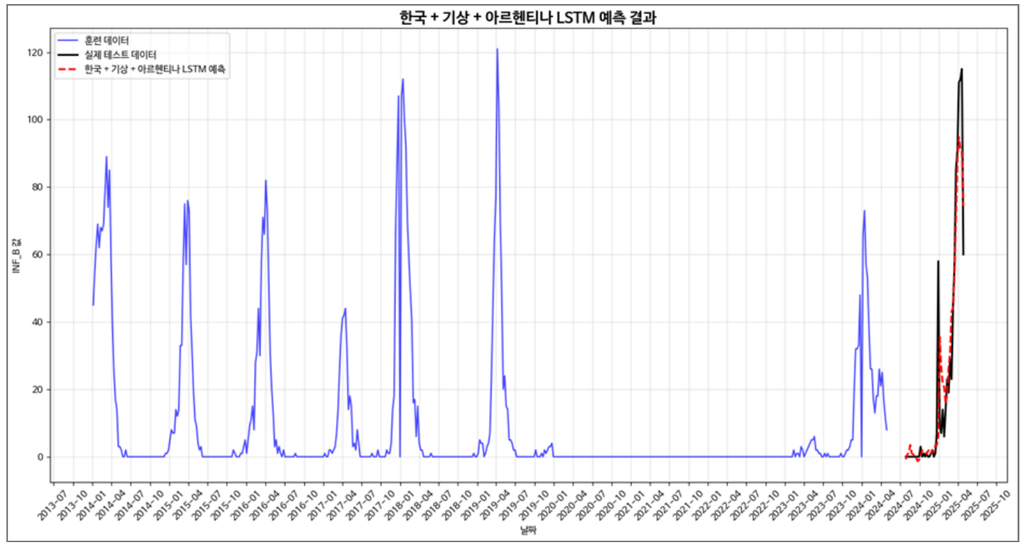


05 모델링 | LSTM 예측 모델

한국 + 아르헨티나(Lag 32) ILI + 국내 기상 변수 기반 LSTM 모델
RMSE: 12.5929 / MAE: 7.0014 / R^2: 0.8697

기존 모든 실험 대비 가장 높은 예측 정확도를 보임.
특히 RMSE 기준으로 한국 단독 모델 대비 약 29%, 아르헨티나 추가 모델 대비 약 25% 이상 감소하는 성과.

- 입력 시퀀스 길이: 12주
- 입력 변수: INF_B, ARG_INF_B, 기온, 일교차, 습도
- 모델 구조: LSTM(64) → DENSE(1)
- 최적화 및 손실 함수: RMSPROP (OPTIMIZER), MSE (LOSS)
- EPOCHS: 80
- BATCH SIZE: 8
- VALIDATION SPLIT: 20%



05 모델링 | LSTM 예측 모델

한국 + 아르헨티나 + 남반구 다국가 기반 LSTM 모델

RMSE: 17.7562 / R^2: 0.7661

*뉴질랜드가 포함된 모델 기준

예상과는 달리 다국가 병합 모델의 성능은 기존 2개국 모델 대비 하락.
남아공, 호주, 칠레, 브라질 등 다른 국가들을 병합한 모델에서도 유사한 경향이 반복됨

문제 원인 분석 |

- 다국가 병합 시 발생하는 변수의 과도한 증가, 데이터 품질의 이질성이 원인으로 분석됨.
- 학습 데이터가 상대적으로 제한된 상황에서 입력 차원이 급격히 늘어날 경우, 모델이 노이즈에 민감해지며 오버피팅될 가능성

예측 모델 해석 결론 |

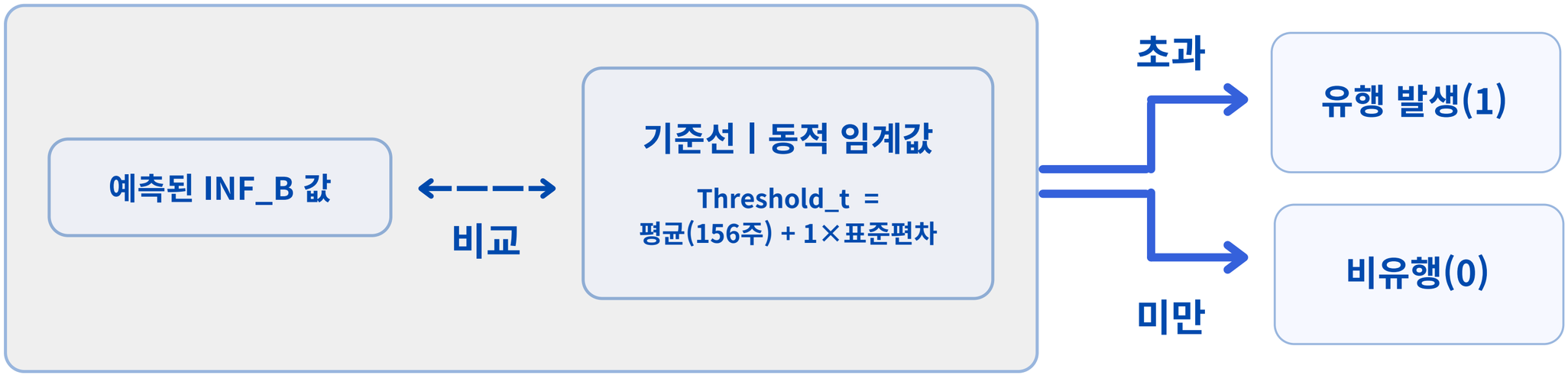
- 아르헨티나 단일국가와 기후 변수를 조합한 모델이 예측 효율성과 실용성 측면에서 가장 우수함
- 다국가 병합 전략은 일반화 성능을 저해할 가능성 높음

국가	lag	cosine 유사도	최종 점수
South Africa	31	0.447921	0.054122
Chile	19	0.443564	-0.393376
Australia	23	0.393162	0.0157
New Zealand	30	0.381004	0.010079
Brazil	43	0.276682	-0.255194

05 모델링 | LSTM 예측 모델

이진 분류 평가 기준 |

정량 예측된 INF_B 값의 일정 기준에 대한 초과 여부

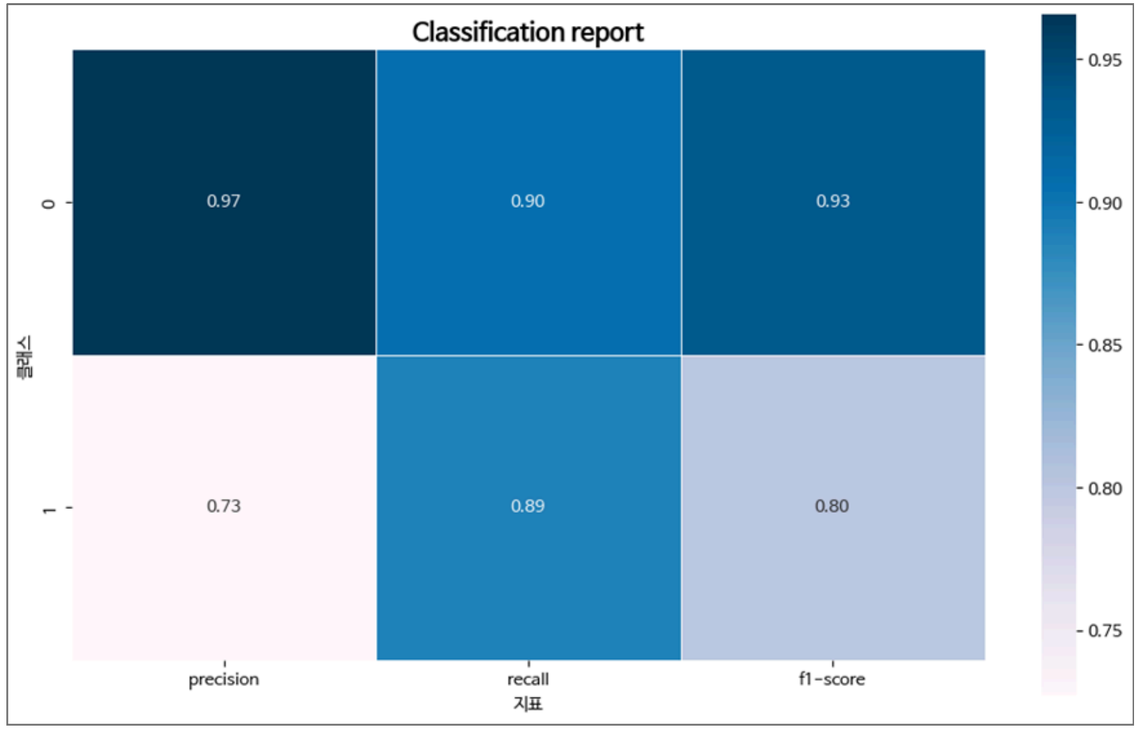


05

모델링 | LSTM 예측 모델

이진 분류 평가 결과 |

- 전체 정확도(ACCURACY): 0.900
- 전체 F1 점수: 0.800



시계열 예측값과 실제 INF_B 값의 시각적 비교 결과에서도 일관된 패턴 대응이 확인

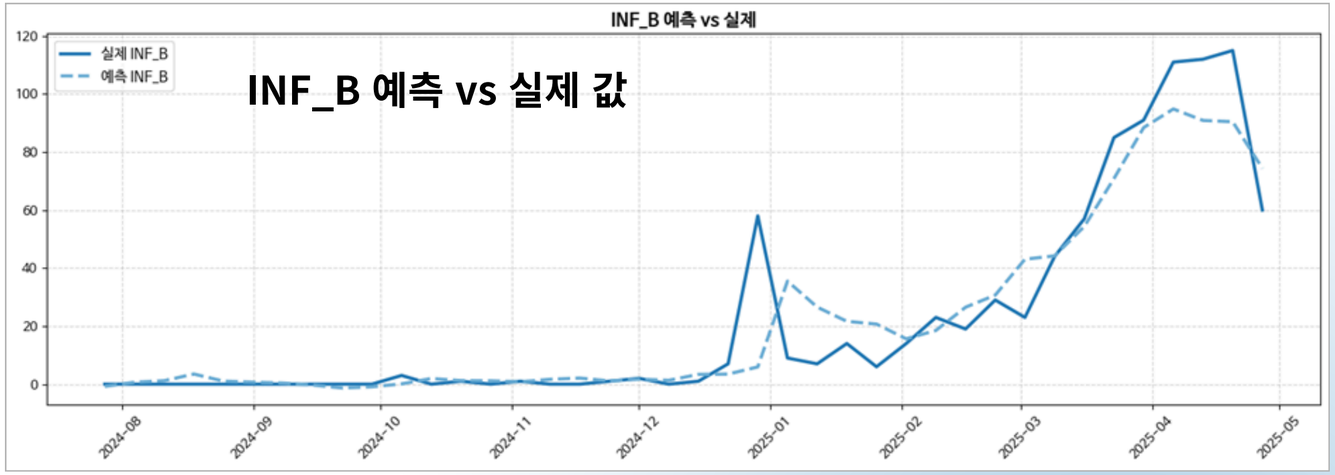
05 모델링 | LSTM 예측 모델

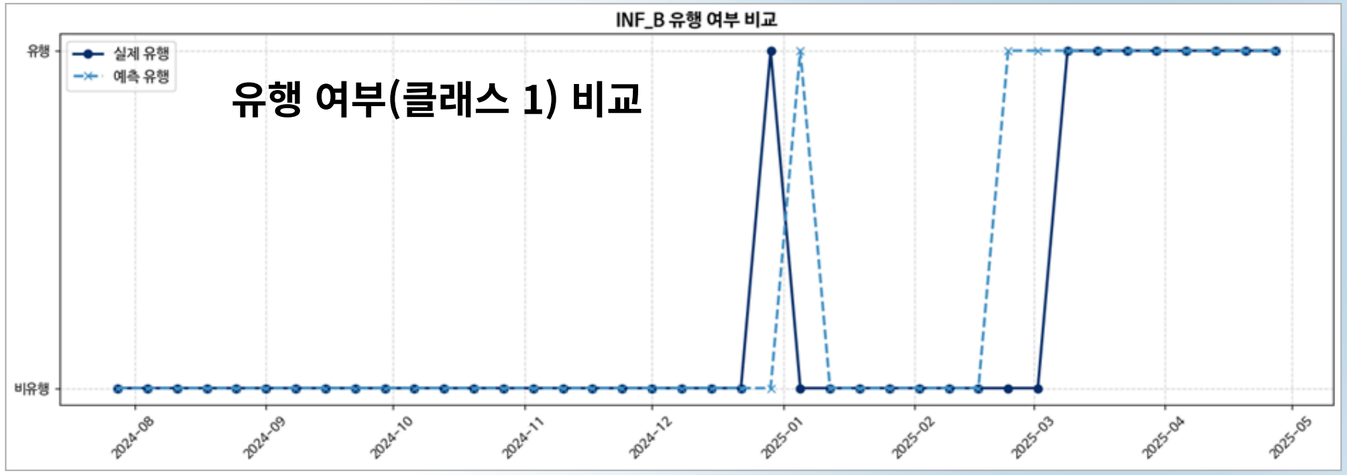
이진 분류 평가 결과 |

- **RMSE: 12.5929**
- **MAE: 7.0014**
- **R^2: 0.8697**

→ 정량 예측 성능 자체도 우수

→ 판단 결과 또한 신뢰도 높은 경보 지표로 기능





05

모델링 | 해석 및 결론

본 예측 모델에서 중요한 역할을 한 주요 변수

아르헨티나 INF_B (lag 32주)

선행 유행 예측에 효과적인 변수로, 한국 유행의 조기 탐지에 기여함

기상 변수 (기온·습도·일교차)

모델의 정밀도를 향상시켜 유행 강도와 지속성 추정에 보완적인 효과를 줌

한국 INF_B (기준 시계열)

예측 모델의 중심축이 되는 데이터로, 외생 변수들과의 상호작용을 통해 정확도 향상에 기여함

- **예측 성능: R^2 0.8697, RMSE 약 12.6**
- **이진 분류: F1-score 0.800, 재현율(Recall) 0.889**

05

모델링 | 해석 및 결론

향후 모델링시 고려 요소 및 조건

1. 시계열 구조적 유사도 확보 (cosine, DTW 기반 사전 검증)

2. 적절한 시차 보정 적용 (최적 lag 분석 후 shift)

3. 실질적 예측 성능 개선이 확인된 경우에만 채택 (RMSE, R^2 기준)

- 일관된 해외 감시 시계열과 기후 요인을 통합한 **비선형 딥러닝 기반** 감염병 예측 전략은 실용성과 예측력을 모두 갖춘 유의미한 접근 방식임을 입증
- **제약 산업** 및 보건의료 분야에서 초기 의사결정 및 자원 배분에 활용 가능한 정량적 예측 인프라를 마련 가능

06

기대 효과 및 향후 보완점

한계 및 향후 보완점 I

도메인 지식 기반 변수 설계의 한계

정성적 인사이트를 모델링에 명시적으로 반영하지 못한 부분은 향후 보완될 필요

선행 연구 및 학술적 근거 기반의 제약

모델 선택이나 변수 조합에 있어 ‘경험적 최적화’에 의존하는 측면이 일부 존재

학습 데이터의 제약 및 불균형

코로나19 기간의 구조적 이상치로 인해 이 구간을 제외함에 따라 학습에 활용 가능한 데이터가 제한

모델 확장성의 한계 및 과적합 위험

외생 변수의 수가 증가할수록 오히려 예측 성능이 저하되는 경우가 발생

06

기대 효과 및 향후 보완점

실질적 기대효과 I

유행 사전 예측을 통한 재고 손실 최소화

반복적으로 발생해온 백신 및 의약품의 초과 재고 문제를 완화, 자산 손실 감소에 기여 가능

공급망(SCM) 민첩성 향상 및 운영 리스크 사전 차단

단기 재고 운영 및 월간 생산계획에 반영해 공급지연, 병목, 과잉 납품 등 공급망 리스크 요인 사전 제어

경영 의사결정의 정량화 및 재무 안정성 확보

기업의 자본 운용 효율성 및 영업이익률 개선에 직결되며, 재무 리스크를 줄이는 방향으로 기능 가능

06 기대 효과 및 향후 보완점

실질적 기대효과 I

AS-IS

비정형화된 백신 수요 예측 방식 사용

- 실무자의 감에 의존
- 참고 국가에 대한 논리적인 근거 미비
- 전년도 판매 실적을 기반으로 한 짧은 예측

TO-BE

정형화된 백신 수요 예측 모델 구축

- LSTM 모델 구축으로 논리적 예측
- 아르헨티나 단일 국가 선정으로 높은 예측 성능 확보

참고문헌

1. Chen, X., Tao, F., Chen, Y., Cheng, J., Zhou, Y., & Wang, X. (2025). Forecasting influenza epidemics in China using transmission dynamic model with absolute humidity. Infectious Disease Modelling, 10, 50–59.
2. Caini, S., & Kusznierz, G. (2019). The epidemiological signature of influenza B virus and its B/Victoria and B/Yamagata lineages in the 21st century. PLOS ONE, 14(9), e0222381.
3. Ashraf, M. A., & Raza, M. A. (2024). A comprehensive review of influenza B virus, its biological and clinical aspects. Frontiers in Microbiology, 15, 1467029.
4. Moon, J., Jung, S., Park, S., & Hwang, E. (2021). Machine learning-based two-stage data selection scheme for long-term influenza forecasting. Computers, Materials & Continua, 68(3), 2945–2959.
5. Kandula, S., & Shaman, J. (2019). Near-term forecasts of influenza-like illness. Epidemics, 27, 41–51.
6. Lee, K. C.-Y., Lin, L. C. Y., Leung, C. T., Yau, D. S.-W., Chan, J. Y. N., Ip, D. K. M., & Lau, E. H. Y. (2024). An adaptive weight ensemble approach to forecast influenza activity in an irregular seasonality context. Nature Communications, 15, Article No. 4040.
7. O'Donnell MJ, Fang J, Mittleman MA, Kapral MK, Wellenius GA; Investigators of the Registry of Canadian Stroke Network. Fine particulate air pollution (PM2.5) and the risk of acute ischemic stroke. Epidemiology. 2011 May;22(3):422-31. doi: 10.1097/EDE.0b013e3182126580. PMID: 21399501; PMCID: PMC3102528.
8. Choi SB, Ahn I (2020) Forecasting seasonal influenza-like illness in South Korea after 2 and 30 weeks using Google Trends and influenza data from Argentina. PLoS ONE 15(7): e0233855. https://doi.org/10.1371/journal.pone.0233855
9. Sungwoo, P. (2020). SHAP-based explainable influenza occurrence forecasting using lightGBM
10. Soo Beom Choi. (2020). Forecasting seasonal influenza-like illness in South Korea after 2 and 30 weeks using Google Trends and influenza data from Argentina

E.O.D

ILI 예측모델 개발 프로젝트
: 인플루엔자 B의 예측을 통한 의약품 재고 최적화

중앙대학교 SCM 분석 학회 씀씀이 B팀
강다훈 김희연 김태형 이정우 유현승 윤설리 한예호 전주현