미세먼지-호흡기질환 전처리 결과 보고서

1. 목적 및 배경

미세먼지(PM10) 및 초미세먼지(PM2.5)가 호흡기 질환(천식·비염) 외래 방문에 미치는 영향을 분석하기 위해,국민건강보험공단 진료정보와 대기환경(PM10/PM2.5) 데이터를 "시도별·월별" 단위로 통합한 **최종 분석용 데이터셋**을 구축한다.

---

2. 데이터 개요

항목	설명
분석 기간	2006 년 1 월 ~ 2024 년 10 월 (226 개월)
공간 단위	전국 17 개 시·도 (사용 파일에 이미 "시도별·월별"로 가공됨)
데이터 구성	진료 건수, 성별, 연령대, 지역, PM10/PM2.5 농도
미세먼지(PM10) 전처리	시도별·월별 PM10 평균 농도(㎍/㎥)
초미세먼지(PM2.5) 전처리	시도별·월별 PM2.5 평균 농도(㎍/㎥)
천식 환자 전처리	시도별·월별 천식 외래 환자 수
비염 환자 전처리	시도별·월별 비염 외래 환자 수, 동일한 컬럼 구조
연령대 매핑 기준	연령대('0–5 세', '6–17 세', ..., '65+세') → 대표 나이 매핑
날짜 매핑(캘린더 차원) 기준	date_id(예: 200601) → year, month, quarter, season
데이터 출처	국민건강보험공단, 공공데이터포털

---

3. 전처리 목표

시도별·월별 공통ㅈㅈ 키 생성

모든 파일은 year_month("YYYY-MM")와 region("서울특별시", "부산광역시" 등 시도명)을 공통 키로 가짐

미세먼지(PM10/PM2.5)와 환자 수 데이터 병합

PM10/PM2.5 평균 농도(시도×월) ↔ 천식·비염 외래 환자 수(시도×월)

연령대·성별 합산(요약)

최종 분석용으로는 "시도×월별 총 환자 수(성별·연령대 합산)"을 사용
연령별·성별 민감도 분석 시에는 원본 구분을 유지

계절 변수 및 추가 칼럼 생성

year_month 에 대응하는 season(봄/여름/가을/겨울) 삽입

변수간 일치성·정규화 확인

파일 간 동일한 지역명(시도명) 사용 여부 확인
필요 시 "서울특별시" 등 미리 표준화된 형태를 유지

---

4. 전처리 상세 과정

4.1 컬럼명 정리

- PM10/PM2.5
  - (pm10_processed_20250529_v1.0.xlsx, pm25_processed_20250529_v1.0.xlsx)
  - year_month (문자열, 예: "2006-01")
  - "서울특별시", "부산광역시", ..., "제주특별자치도"
- 천식/비염
  - (pm10_asthma_processed_20250529_v1.0.xlsx, pm10_rhinitis_processed_20250530_v1.0.xlsx):
  - year_month (문자열), region(시도명), gender("남"/"여"), age_group(연령대 구분), visit_count
  - 연령대 매핑(reference_agegroup_mapping.xlsx):
  - 구분(연령대 문자열), 나이(대표값)
- 날짜 매핑
  - (reference_date_mapping.xlsx):
  - date_id(예: 200601), year(정수), month(정수), quarter, season(string)

4.2 날짜 정규화

- 각 파일 모두 year_month 형식("YYYY-MM")으로 통일되어 있으며, datetime64 타입 변환→ 분기(quarter)/계절(season) 결합 시 아래 과정을 거침.

4.3 지역 정제 및 통일

- 주요 시도명(예: "서울특별시", "부산광역시", "대구광역시" ..., "제주특별자치도")이 일치.

4.4 결측치 및 이상치 처리

- 결측 <10%인 컬럼(자치구·도 단위): 전후일 평균으로 보강
- 결측 발생 원인: 센서 설치 시기 차이
- 일부 소규모 지역은 결측을 0(노출 없음)으로 처리,
- 주요 자치구(서울·경기·대전 등)는 인접 관측소 평균값 보강
- PM 수치의 극단값 → IQR 기반 이상치 확인 (필요 시 제거 예정)

---

5. 향후 활용 계획

1. 상관분석 기반 가설 검정
    - PM10/PM2.5 와 천식·비염 환자 수 간 Pearson/Spearman 상관계수 계산
    - 라그(lag) 효과(1~3 개월 지연 효과) 분석
2. 회귀분석 모델링
    - 단순선형회귀(단일 오염원), 다중회귀(기상변수, 라그 포함)
    - 정규화(Regularization) 기법(Lasso, Ridge)을 통한 변수 중요도 도출
3. 시계열 예측 모델
    - Prophet, XGBoost, LSTM 기반 시계열 예측 비교
    - 평가 지표: R^2, MAE, MAPE
4. 취약 계층(고령자·어린이) 민감도 분석
    - 원본 asthma_df, rhinitis_df 를 활용하여 "연령대별 상관분석" 및 "회귀분석"
    - 연령별 회귀계수 비교를 통해 상대적 민감도 파악

5. 대시보드/서비스 구현

- Streamlit 기반 실시간 입력(미세먼지 농도, 기상 변수) → 환자 수 예측 인터페이스
- 자동 알림(문자/이메일) 기능 연계 검토

6. 정책 제언 및 인사이트 도출

- 계절별 집중 관리 방안(예: 겨울·봄철 고농도 경보)
- 시도별 자원 배치(병상, 의료 인력) 최적화