기상청

일별 혈관질환 발병위험도 예측을 통한 대국민 알림 서비스

참 가 번 호	220168	팀 명	입상

1. 공모 배경

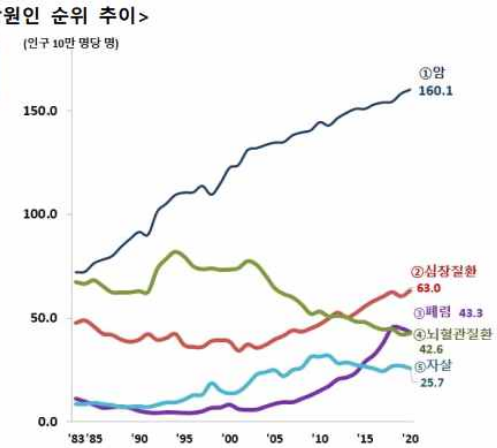
□ 한국인 사망원인 2위, 심혈관질환

○ 통계청에서 2020년에 발표한『한국인 사망원인 통계』에 의하면 심혈관 질환이 암에 이어 한국인 사망 원인 2위를 차지함.[1)]

<사망원인 순위 추이>

(단위: 인구 10만 명당 명)

순위	사망원인	사망률	'19년 순위 대비
1	악성신생물(암)	160.1	-
2	심장 질환	63.0	-
3	폐렴	43.3	-
4	뇌혈관 질환	42.6	-
5	고의적 자해(자살)	25.7	-
6	당뇨병	16.5	-
7	알츠하이머병	14.7	-
8	간 질환	13.6	-
9	고혈압성 질환	11.9	⬆(+1)
10	패혈증	11.9	⬆(+1)



- 뇌혈관질환의 경우 사망률이 감소하는 추세를 보이지만 여전히 심장질환에 의한 사망률은 증가하는 추세임을 알 수 있음.

[표 8] 순환계통 질환의 성별 사망률 추이, 2010-2020

(단위: 인구 10만 명당 명, %)

		순환계통 질환	고혈압성 질환	심장 질환	허혈성[1] 심장 질환	기타[2] 심장 질환	뇌혈관 질환
남녀 전체	2010년	112.5	9.6	46.9	26.7	20.2	53.2
	2019년	117.4	11.0	60.4	26.7	33.8	42.0
	2020년	**121.1**	**11.9**	**63.0**	**27.4**	**35.6**	**42.6**
	'19년 대비 증감	3.8	0.9	2.6	0.7	1.9	0.5
	'19년 대비 증감률	3.2	8.3	4.2	2.6	5.5	1.2

- 인구 10만 명당 사망률의 추이를 보면 순환계통 질환에 의한 사망률이 큰 폭으로 증가하고 있음을 알 수 있음.
- 혈관질환의 경우 사망률이 높은 질환이기 때문에 사전에 혈관질환의 발생 원인을 파악하고 발병위험도가 높을 것으로 예상되는 경우 사전에 안내함으로써 혈관질환의 발생을 방지해야 함.

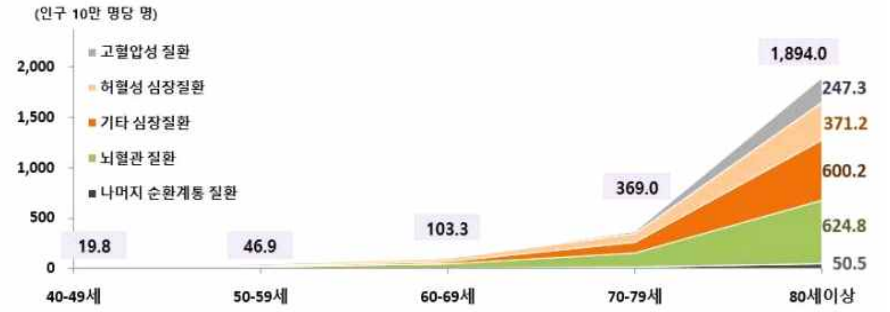
○ 연령이 증가함에 따라 순환계통 질환의 사망률이 높아짐.

- 연령이 증가할수록 사망률이 큰 폭으로 증가하는 추세를 보이며 60세 이상 노인 집단에서 순환계통 질환에 의한 사망률이 매우 큰 폭으로 증가함.
- 따라서 60세 이상 노인을 고위험군으로 분류하고 혈관질환 발생을 방지할 수 있도록 사전

1) 출처: 통계청『한국인 사망원인 통계, 2021』

에 발병위험도 및 행동 수칙을 안내함으로써 혈관질환 발병을 방지해야 함

[그림 7] 순환계통 질환의 연령별 사망률, 2020



2. 분석 데이터 정의

□ 데이터 정의

○ 기상 변수

- 일별 기상 데이터와 다음 날의 기상에 대한 예보 데이터로 구성됨.
  - 예) 2012년 1월 1일의 경우 2012년 1월 1일의 기상 데이터와 2012년 1월 2일의 기상에 대한 예보 데이터로 구성됨.

○ 인구통계학적 변수

- 지역별, 연도별 노인인구 수, 전체인구 수 및 1인당 평균 소득 데이터를 변수로 사용함.

□ 파생변수

○ 기상 변수

- 일교차를 계산한 후 해당 일자의 최저기온이 낮으면 더 큰 값을 갖도록 최저기온 변수에 가중치를 부여한 값을 곱해서 사용함.

  - $(\text{가중치를고려한일교차}) = (\text{최고기온} - \text{최저기온}) \times \frac{1}{(1+e^{\text{최저기온}/30})}$
  - 예) 일교차가 -10℃일 때, 최저기온이 10℃인 경우 5.8257, 최저기온이 10℃인 경우 4.1743의 값을 가짐.

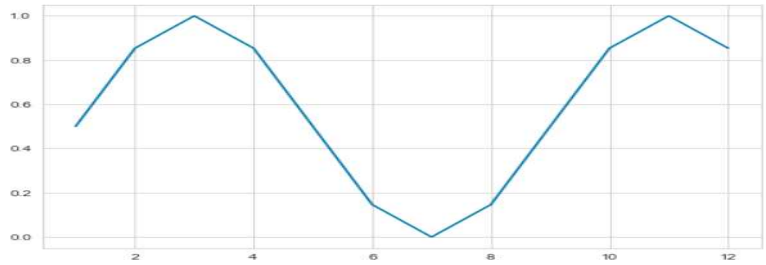
- 체감기온 변수 및 대기 오염도의 지연효과를 확인하기 위해 예측일 기준 1일 전, 2일 전, 3, 7, 10일 전의 PM10, 오존 농도 변수를 생성함.

○ 기타 변수

- 연도별 공휴일 목록을 구성한 후, 해당 날짜의 평균 발병 빈도가 전체 데이터의 발병 빈도의 평균보다 10% 이상 높은 경우만 공휴일 변수로 사용함.
- 연, 월, 일, 요일, 주말 변수 생성
  - 월 변수의 경우 환절기와 겨울에 발병 빈도가 높아지고 여름에 낮아진다는 점을 고려하여 코사인 변환을 수행한 값을 사용함.

- 지역별 혈관질환 발병의 양상이 다를 것으로 판단하고 총인구 대비 발병 빈도의 비율에 따라 4가지의 집단으로 분류함.

- 월 변수의 코사인 변환 결과

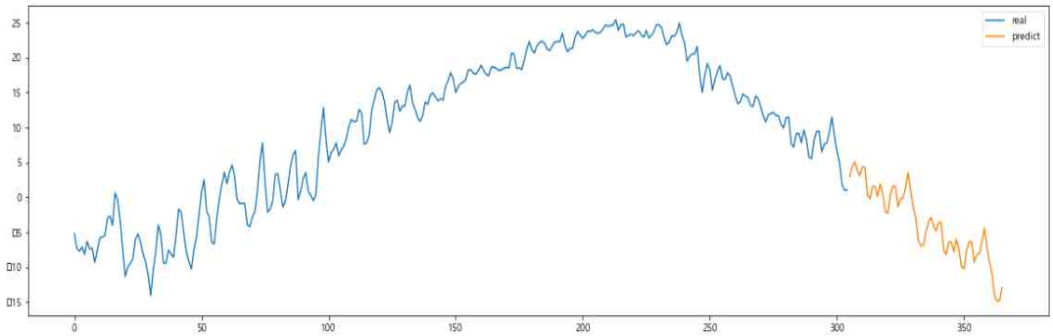


3. 활용 분석기법 및 모델링 결과

□ 데이터 전처리

○ 예보 데이터 결측 처리

- 예측 모형에서 예보 데이터는 일 최고기온, 일 최저기온, 습도를 사용함.
- 2012년 11월, 12월의 경우 모든 지역에서 예보 데이터가 결측 데이터임.
- 시계열 모형인 SARIMAX 모형을 통해 1월부터 10월까지의 추세로부터 11월, 12월의 값을 예측한 후 사용함.
- 예) 충청북도 지역의 최저기온 예측 결과



○ PM10, O3 데이터 결측 처리

- 인접한 지역에서 대기질이 비슷하다는 점을 이용해서 인접한 지역의 값으로 대체함.

□ 반응변수 변환

○ 일자별 혈관질환 발병 빈도는 전날의 기상 변수를 통해 예측

- 전날의 기상 변수와 전날 발표된 기상예보 데이터를 통해 해당 일자의 발병 빈도를 예측함. 이는 서비스 측면에서 일자별 혈관질환 발병위험도를 전날에 예측하여 사전에 안내하기 위함임.
- 예) 2012년 1월 2일의 발병위험도는 2012년 1월 1일 23시에 예측한 후 안내 대상자들에게 알림 서비스로 전송됨.

○ 혈관질환 발병 빈도 대신 총인구 100만 명당 발병 빈도로 변환

- 발병 빈도 자체를 사용하는 경우 총인구 수에 큰 영향을 받게 됨. 이를 방지하기 위해 발병 비율을 예측하는 방식으로 변경함. 이렇게 함으로써 총인구수와 관계없이 발병 비율이라는 동일한 의미를 갖도록 함.
- 최종 예측값 계산 시에 예측된 발병 비율에 총인구수를 곱한 값을 제출했으며, 실제 서비스

에서는 예측된 발병 비율을 이용함.

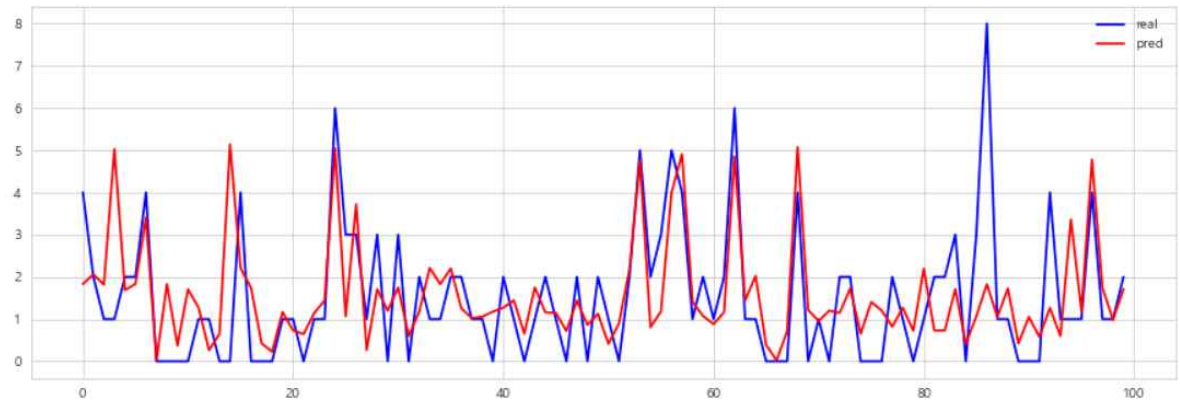
□ 모형 학습 및 모델링 결과

○ 최종 모형

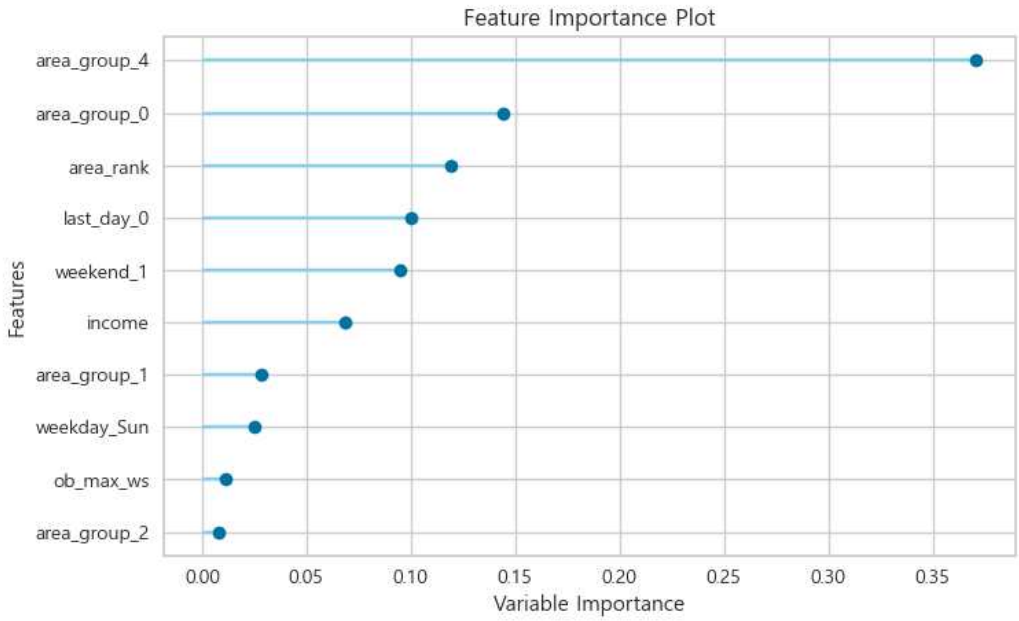
- Gradient Boosting Regressor 알고리즘을 사용함.
- 학습용 데이터를 구축한 후 모형 선정 과정에서 AutoML 라이브러리를 활용함. 여러 회귀 모형 후보 중에서 예측 오차가 가장 낮은 모형을 선정해서 하이퍼 파라미터 튜닝을 수행함.
- 여러 개의 Decision tree를 통해서 강력한 예측 모형을 만드는 앙상블 기법의 한 종류이며, Random Forest와 다르게 독립적인 개별 모형들의 예측을 기반으로 하는 것이 아니라 이전 모형의 오차를 학습해가면서 오차를 줄여나가는 식으로 학습함.

○ 예측 결과

- 훈련 데이터의 20%를 검증 데이터로 활용하였으며, RMSE는 1.2317임.
- 검증 데이터에서 임의로 선택한 100개의 관측치에 대한 예측 결과



□ 중요 변수



- 일별 혈관질환 발병의 예측에서는 주로 기상 요소를 제외한 기타 변수들이 중요도가 높았으며 특히 65세 이상 노인 인구 수 대비 발병 비율에 따라 집단을 분류한 변수의 중요도가 가장 높게 나옴. 이어서 지역별로 노인인구(65세 이상) 수 대비 발병 빈도에 대한 순위가 중

요도가 높게 나왔으며, 매월 마지막 날 여부, 주말 여부, 소득 수준이 뒤를 이었음.

- 기상 변수 중에서는 예측일의 풍속, 최고기온(예보), 예측일의 기압, 예측일 3일 전의 오존 농도, 습도(예보), 예측일의 일교차 순으로 중요도가 높게 나타나지만 앞선 기타 변수에 비해 중요도가 떨어짐.

4. 서비스 활용 방안

□ 대국민 혈관질환 발병위험도 알림 서비스

○ 매일 23시마다 다음 날의 혈관질환 발병위험도 계산

- 일별 혈관질환 발생 빈도 예측 모형에서 예측일을 기준으로 다음 날의 날씨(일 최고기온, 일 최저기온, 평균 습도)에 대한 예보 데이터를 사용했음.
- 모형 학습에 사용된 예보 데이터는 매일 23시에 발표된 수치를 사용했으며, 해당 수치를 통해 당일의 혈관질환 발병위험도를 계산해서 자정부터 조회할 수 있도록 함.
- 계산된 발병위험도를 아침 7시에 일괄적으로 전송함.

○ 혈관질환 발병위험도 등급화

- 혈관질환 발병위험도의 등급은 지역마다 다르게 산정하며, 연도별로 모형을 통해 예측한 전년도의 1년간의 혈관질환 발병위험도를 기준으로 함.
- 지역마다 예측 확률을 5개의 구간으로 나누어 1~5단계의 등급으로 안내함.
- 예측 모형에서 변수 중요도가 높게 나온 기상 요소를 선정해서 해당 요소들의 수치를 위험도 등급과 함께 제시 후 행동 수칙을 안내함.

□ 알림 서비스 활용 매체

○ 스마트폰 앱을 이용한 푸시 알림 서비스

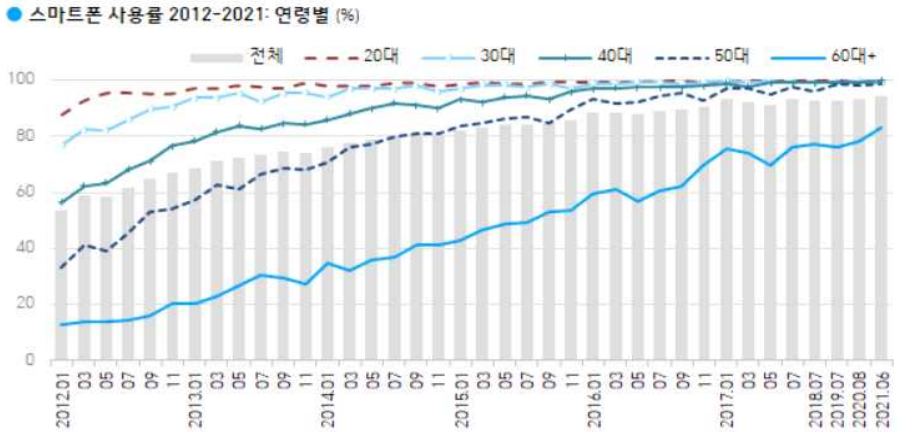
- 이동통신사에서 알림서비스 앱을 기본 앱으로 제공함.
- 매일 23시에 다음 날의 기상예보 데이터가 확보되면 예측 모형을 통해서 다음 날의 혈관질환 발병위험도를 계산함.
- 앱에서 당일의 혈관질환 발병위험도를 제시함.
- 매월 마지막 날이나 공휴일의 경우 발병 빈도가 증가하는 경향이 있음. 이는 과음에 의한 것으로 판단되어 해당 날짜에는 과음에 대한 경고 메시지및 예방 수칙을 추가로 전송함.

○ 문자 메시지

- 60세 이상의 스마트폰 보급률은 80%에 불과함.[1)]
- 60세 이상 인구가 고위험군으로 분류된다는 점을 고려했을 때, 해당 집단 내 스마트폰 미가입자에게도 정보를 제공할 수 있는 방안이 마련되어야 함.
- 스마트폰 미가입자를 대상으로 스마트폰 앱을 통한 푸시 알림과 동일한 방식으로 발병위험도를 문자 메시지로 안내함.
- 60세 이상 인구가 고위험군으로 분류된다는 점을 고려했을 때, 해당 집단 내 스마트폰 미가입자에게도 정보를 제공할 수 있는 방안이 마련되어야 함.

1) 출처: 한국갤럽『스마트폰 사용률 & 브랜드, 스마트워치, 무선이어폰에 대한 조사, 2021』

- 스마트폰 미가입자를 대상으로 스마트폰 앱을 통한 푸시 알림과 동일한 방식으로 발병위험도를 문자 메시지로 안내함.



○ 기타 매체

- 위의 두 가지 방법은 휴대폰이라는 수단에 의존하게 된다는 한계점을 가짐.
- 이러한 한계점을 보완하기 위해 뉴스, 신문, 라디오 등의 매체에서 기상정보와 함께 발병위험도를 안내함.

○ 추가 방안

- 발병위험도가 4단계 이상인 경우 안내 대상자들에게 안전 안내 문자를 발송함.
- 70세 이상 노인이나 저소득층을 대상으로 일별 발병위험도를 알려주는 알림 팔찌를 지급함으로써 사각지대 문제를 방지함.

□ 실제 서비스 적용

○ 성과 측정

- 알림 서비스 전·후의 일별 평균 혈관질환 발병 빈도 비교

○ 서비스 구축

- 기상 데이터 및 기상예보 데이터 실시간 수집 체계 구축
- 앱 내 발병위험도 및 행동 수칙 안내 서비스 구축

□ 기대 효과

○ 혈관질환 발병률 감소 및 대응 체계 구축

- 일별 혈관질환 발병위험도를 사전에 계산할 수 있는 시스템이 구축되면 고위험군에 대한 사전 안내 및 예방 수칙을 제안함으로써 혈관질환 발병률을 낮출 수 있음.
- 혈관질환에 단기적으로 영향을 미치는 요인을 파악할 수 있음.
- 혈관질환의 발병위험도가 높은 날에는 각 병원에서 환자 수를 예상하고 병상 확보, 의료진 확충 등의 실시간 대응 체계를 구축할 수 있음.

○ 환자 데이터 수집

- 알림 팔찌에 건강 상태를 측정하는 기능을 추가하면 혈압, 심박수 등의 데이터를 수집함으로써 혈관질환 예측에 활용할 수 있음.
- 알림 팔찌를 통해 수집된 데이터를 통해 혈관질환 이외에도 실시간으로 개인의 건강 상태를 진단하는 효과를 기대할 수 있음.