건강검진데이터를 활용한 질병 예측 및 건강보조식품 추천 시스템 - 중간 발표

과목 : 헬스케어데이터사이언스

일자 : 2024.10.28.

조 이름 : 헬스프로텍터

조원 : 김다은 김우신 김태현 유승찬 임과림 전지은

한양대학교

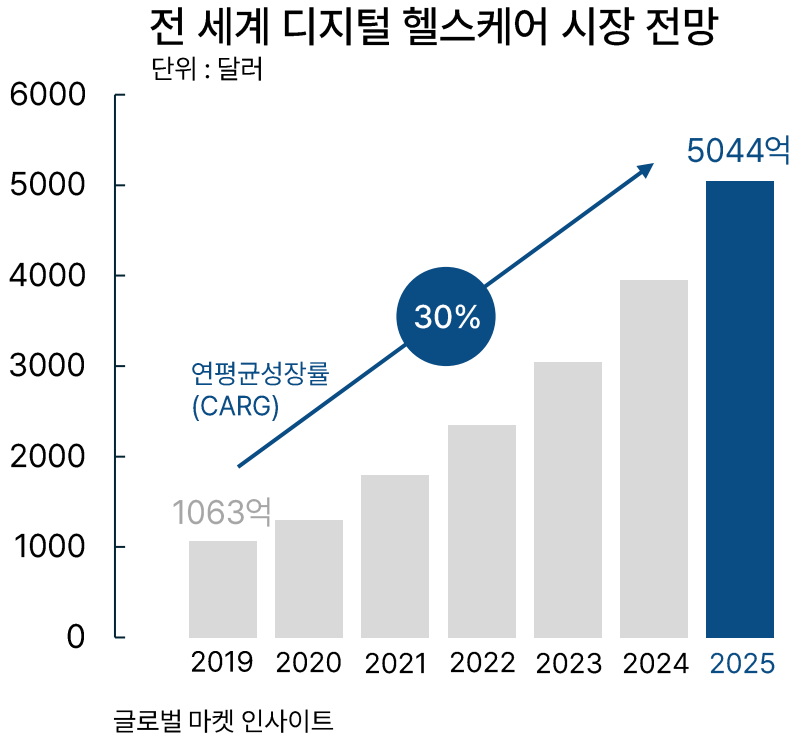
Table of contents

한양대학교

문제해결 시나리오

건강검진 데이터를 통한 특정 질환 예측 및 개인 맞춤형 건강보조식품 추천 알고리즘 개발

최근 디지털 헬스케어 분야는 CES 2024에서도 주요 키워드로 다뤄질 만큼 주목을 받고 있으며, 개인 맞춤형 건강 정보를 제공받고자 하는 수요도 빠르게 증가하고 있다. 건강기능식품 쇼핑몰을 운영하는 ㈜글로잇은 개인의 건강 상태에 맞춰 적합한 제품을 추천하는 방안을 모색하고 있으며, 이를 위해 자사가 보유한 국가건강검진 데이터를 활용해 특정 질환을 예측하고, 이에 기반한 모델 및 알고리즘을 개발하고자 한다.



프로젝트 목표

- 대규모의 건강검진 정보를 통해 데이터를 분석하고 예측하는 모델 개발
- 완성된 모델을 통해 각 개인에게 알맞은 맞춤형 건강보조식품 추천

요구사항

01 국가건강검진 데이터를 통해 도출할 수 있는 질병 및 특정 의학적 상태에 대한 발굴

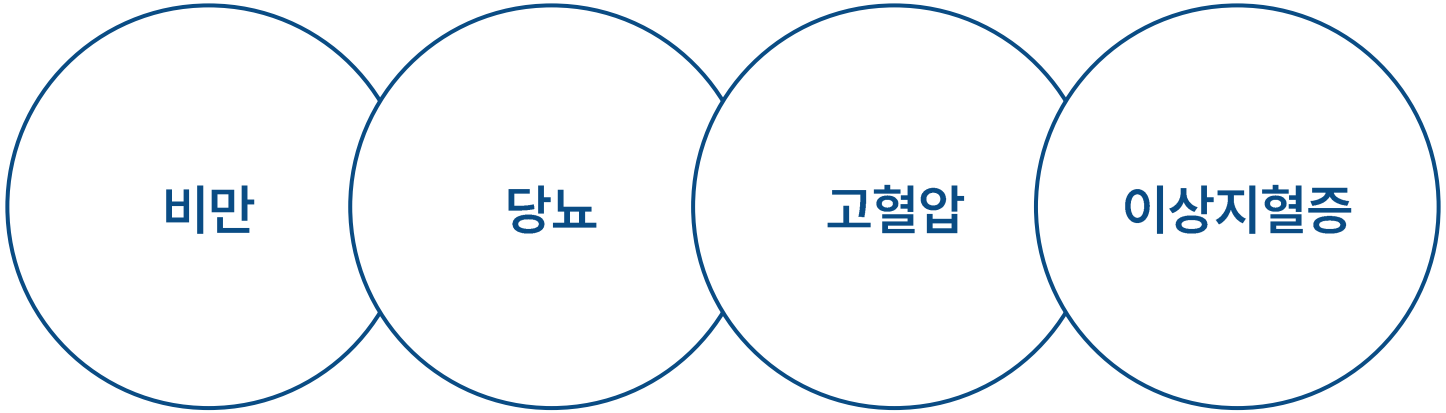
02 예측 모델 알고리즘 개발을 통해 가능성 있는 질병에 대해 위험도 제시

03 건강보조식품과 연계하여 사용자의 상태를 개선해줄 수 있는 상품 연결

요구사항

01 국가건강검진 데이터를 통해 도출할 수 있는 질병 및 특정 의학적 상태에 대한 발굴

기존 건강보험 데이터와 추가 데이터셋에서 확인 가능한 네 가지의 질병을 타겟 질환으로 설정



요구사항

02 예측 모델 알고리즘 개발을 통해 가능성 있는 질병에 대해 위험도 제시

1. 각 질병을 특성 수치 Rule 기반으로 라벨링(e.g. 고혈압 1기 : 수축기 140~159 / 이완기 90~99, 고혈압 도 정상, 고혈압 이전, 1기, 2기 등)

2. 질병에 대한 확률값 존재 (e.g. 고혈압 : 정상 0.6, 고혈압 위험, 0.2, 1기 0.15, 2기 0.05 등)
   → 이를 바탕으로 질병의 발병 위험 확률 또는 단계 예측

요구사항

03 건강보조식품과 연계하여 사용자의 상태를 개선해줄 수 있는 상품 연결

타겟 질환으로 설정한 네 가지의 질병(비만, 당뇨, 고혈압, 이상지혈증)에 대해 질병 위험도를 예측하고, 해당 질병들에 대한 건강보조식품추천 알고리즘 개발

국내 영양제 중 아래의 기준에 따라 영양제 엄선 후, 맞춤형 추천 알고리즘 개발

1. 원료의 안전성 : 안전성을 검증 받은 원료를 사용하는 제조사의 제품
2. 함량의 적절성 : 같은 성분의 영양제라도 함량이 높은 영양제
3. 가격의 합리성 : 합리적인 가격과 좋은 품질의 영양제

기본 데이터셋

: 국민건강보험공단 건강검진 데이터

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	기준년도	가입자일련	성별코드	연령대코드	시도코드	신장(5Cm단	체중(5Kg단	허리둘레	시력(좌)	시력(우)	청력(좌)	청력(우)	수축기혈압	이완기혈압	식전혈당(공	총콜레스테	트리글리세	HDL콜레스	LDL콜레스	혈색소	요단백
2	2017	1	1	8	43	170	75	90	1	1	1	1	120	80	99	193	92	48	126	17.1	1
3	2017	2	1	7	11	180	80	89	0.9	1.2	1	1	130	82	106	228	121	55	148	15.8	1
4	2017	3	1	9	41	165	75	91	1.2	1.5	1	1	120	70	98	136	104	41	74	15.8	1
5	2017	4	1	11	48	175	80	91	1.5	1.2	1	1	145	87	95	201	106	76	104	17.6	1
6	2017	5	1	11	30	165	60	80	1	1.2	1	1	138	82	101	199	104	61	117	13.8	1
7	2017	6	1	11	41	165	55	75	1.2	1.5	1	1	142	92	99	218	232	77	95	13.8	3
8	2017	7	2	10	27	150	55	69	0.5	0.4	1	1	101	58	89	196	75	66	115	12.3	1
9	2017	8	1	8	48	175	65	84.2	1.2	1	1	1	132	80	94	185	101	58	107	14.4	1
10	2017	9	1	12	41	170	75	84	1.2	0.9	1	1	145	85	104	217	100	56	141	15.1	1
11	2017	10	1	9	41	175	75	82	1.5	1.5	1	1	132	105	100	195	83	60	118	13.9	1
12	2017	11	1	10	41	155	55	79.2	1	1	1	1	118	70	90	183	55	42	130	12.9	1
13	2017	12	1	14	27	155	75	98	1.2	9.9	1	1	109	69	137	115	137	31	57	16.5	1
14	2017	13	2	12	41	150	55	72.3	1.2	0.9	1	1	130	80	106	183	214	51	89	13.1	1
15	2017	14	1	7	48	175	75	88	1.2	1.2	1	1	118	72	82	200	77	55	129	15.7	1
16	2017	15	2	7	41	160	50	76	0.9	1	1	1	129	77	79	205	219	53	108	14.5	1
17	2017	16	1	9	47	170	65	80	1	1	1	1	113	72	104	113	35	44	62	16	2
18	2017	17	2	6	42	160	65	73	1.2	0.9	1	1	126	78	96	148	60	54	82	12.3	1
19	2017	18	1	6	28	170	65	78	1.2	1.2	1	1	119	67	100	147	54	51	85	14.8	1
20	2017	19	1	11	41	170	85	99	0.7	0.8	1	1	121	74	99	180	169	43	103	14.4	1
21	2017	20	1	13	44	165	60	85	0.3	0.7	1	1	120	85	105	197	222	42	111	15.2	1
22	2017	21	2	8	11	170	50	67	1	0.8	1	1	111	65	88	174	46	66	98	12.1	1

추가 데이터셋 1

: 국민건강보험공단 건강검진 데이터

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
기준년도	가입자일련번호	시도코드	성별	연령대코드(5세단위)	신장(5cm단위)	체중(5kg단위)	허리둘레	시력(좌)	시력(우)	청력(좌)	청력(우)	수축기혈압	이완기혈압	식전혈당(공복혈당)	총콜레스테롤	트리글리세라이드
2020	5	26	2	6	155	40	62.5	1.5	2	1	1	95	60	81		
2020	8	11	1	13	165	70	84	0.7	1.5	1	1	139	93	80	261	55
2020	11	49	1	9	170	80	89	1.2	1	1	1	118	84	89	206	195
2020	16	46	2	12	155	65	74	1	0.9	1	1	120	80	128		
2020	25	29	1	12	170	65	83	0.8	1.2	1	1	104	72	75	203	38
2020	26	11	1	9	165	45	84	0.7	0.7	1	1	130	70	91	235	85
2020	27	27	2	10	155	55	80	1.5	1.2	1	1	131	80	90		
2020	33	11	2	5	155	50	65.5	1	0.9	1	1	109	69	89		
2020	34	41	2	8	155	55	79	1.2	1.5	1	1	101	51	98		
2020	40	47	2	14	140	50	86	0.5	0.7	1	1	119	68	76		
2020	48	41	1	16	160	55	82.5	0.1	0.2	1	1	133	76	103		
2020	54	42	1	10	170	60	81	0.9	1	1	1	139	105	89	189	121
2020	64	11	1	7	165	75	87	1.2	1.2	1	1	123	66	88		
2020	71	47	1	15	165	60	86	0.7	0.6	1	2	132	65	96	198	84
2020	72	41	2	9	165	50	73	1.2	1.2	1	1	97	67	96	231	62
2020	75	11	2	7	165	45	60.5	1	1.5	1	1	99	64	87		
2020	79	28	1	15	160	70	86	0.9	0.9	1	1					
2020	82	31	1	11	170	70	91	0.7	0.7	1	1	100	67	93		
2020	83	28	2	6	160	50	63	1	0.7	1	1	104	70	93		
2020	85	28	1	8	165	65	84	1	0.9	1	1	120	76	89		
2020	86	29	2	5	160	55	67.2	1	1.2	1	1	103	60	81		
2020	88	11	1	11	170	65	88	1	0.9	1	1	124	80	83	148	153
2020	96	11	1	11	170	70	79	1.2	1	1	1	123	74	87	206	151
2020	97	30	1	14	155	65	91	0.7	0.8	1	1	138	86	102	252	101
2020	101	47	1	12	165	80	94	1.5	1	1	1	135	83	102		
2020	107	49	2	12	150	50	77	1	0.8	1	1	117	72	99	205	49
2020	119	45	1	15	160	65	86	0.8	0.5	1	1	107	77	100	246	181
2020	131	28	1	8	170	65	77.9	1.5	1.5	1	1	110	70	89	157	124
2020	133	48	1	8	180	80	85	1.2	1.2	1	1	135	90	103		
2020	138	46	1	10	175	70	87.4	0.6	0.7	1	1	120	73	117	177	155
2020	144	11	1	12	170	80	96	1.5	1	1	1	115	76	98		
2020	149	41	1	8	170	55	70.5	0.9	0.9	1	1	105	68	91		
2020	154	43	2	16	150	50	80	0.7	0.6	1	2	130	80	123		
2020	159	45	2	13	155	70	86	0.8	0.6	1	1	132	76	101	289	186

국민건강보험공단의 다른 년도
건강검진 데이터 사용 예정

* 국민건강보험공단
https://www.data.go.kr/data/15007122/fileData.do

추가 데이터셋 2

: KOSIS 데이터

연령별(1)	성별(1)	2022							
		위장질환진단유무							대장항문질환진단
		소계	위궤양	위축성 위염	장상피화생	위용종	기타	없음	소계
계	합계	14,645,200	1,235,287	1,902,979	287,714	439,192	1,180,263	9,599,765	14,308,370
	남자	5,810,556	557,446	669,050	125,003	213,322	424,284	3,821,451	5,768,472
	여자	8,834,644	677,841	1,233,929	162,711	225,870	755,979	5,778,314	8,539,898
20 ~ 24세	합계	184,206	2,504	4,756	146	362	6,302	170,136	183,452
	남자	-	-	-	-	-	-	-	-
	여자	184,206	2,504	4,756	146	362	6,302	170,136	183,452
25 ~ 29세	합계	295,979	6,295	11,690	479	1,829	13,257	262,429	294,357
	남자	-	-	-	-	-	-	-	-
	여자	295,979	6,295	11,690	479	1,829	13,257	262,429	294,357
30 ~ 34세	합계	520,770	17,916	28,463	1,394	6,656	26,799	439,542	516,627
	남자	-	-	-	-	-	-	-	-
	여자	520,770	17,916	28,463	1,394	6,656	26,799	439,542	516,627
35 ~ 39세	합계	395,281	20,677	26,738	1,529	6,759	20,514	319,064	391,573
	남자	-	-	-	-	-	-	-	-
	여자	395,281	20,677	26,738	1,529	6,759	20,514	319,064	391,573
40 ~ 44세	합계	1,660,105	139,693	153,268	14,589	42,461	131,390	1,178,704	1,614,949
	남자	699,162	58,836	48,771	5,217	18,567	48,297	519,474	695,617
	여자	960,943	80,857	104,497	9,372	23,894	83,093	659,230	919,332
45 ~ 49세	합계	1,277,152	126,402	157,323	20,049	37,424	111,676	824,278	1,232,599
	남자	538,726	57,163	53,912	7,634	17,632	42,262	360,123	534,279
	여자	738,426	69,239	103,411	12,415	19,792	69,414	464,155	698,320
50 ~ 54세	합계	2,258,442	226,739	330,299	50,296	65,576	187,766	1,397,766	2,188,200

연령대 및 성별에 따른 수치,
질병 분포를 참고하기 위하여 사용

* KOSIS
https://kosis.kr/statHtml/statHtml.do?orgId=350&tblId=DT_35007_N099&conn_path=I2

추가 데이터셋 3

: KNHANES 영양조사데이터

원시자료

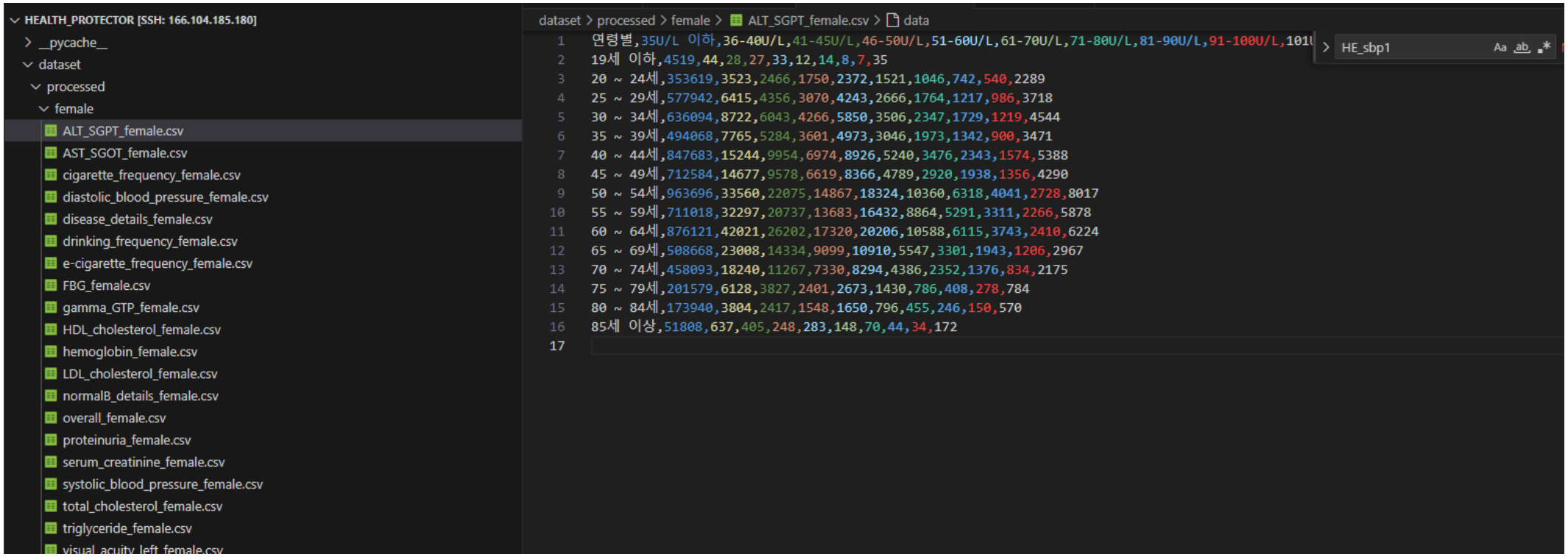
연도	영역		내용	DB	SAS	SPSS	최종수정일
	분과별 상세DB	검진조사	가정 실내공기질 및 환경유해 물질	HN_IAQ	다운로드	다운로드	2024-03-22
	분과별 상세DB	검진조사	가정 실내공기질 및 환경유해물질 생체지표 조사	HN_IAQ	다운로드	다운로드	2023-07-31
2022	기본 DB		검진조사, 건강설문조사, 영양조사	기본DB	다운로드	다운로드	2024-05-21
2022	분과별 상세DB	영양조사	식품섭취조사(개인별 24시회상조사)	HN22_24RC	다운로드	다운로드	2024-01-11
2021	기본 DB		검진조사, 건강설문조사, 영양조사	기본DB	다운로드	다운로드	2023-12-07
2021	분과별 상세DB	영양조사	식이보충제조사	HN21_SUP	다운로드	다운로드	2023-12-07
2021	분과별 상세DB	영양조사	식품섭취조사(개인별 24시회상조사)	HN21_24RC	다운로드	다운로드	2023-01-18
2020	기본 DB		검진조사, 건강설문조사, 영양조사	기본DB	다운로드	다운로드	2023-12-07
2020	분과별 상세DB	영양조사	식이보충제조사	HN20_SUP	다운로드	다운로드	2023-12-07
2020	분과별 상세DB	영양조사	식품섭취조사(개인별 24시회상조사)	HN20_24RC	다운로드	다운로드	2023-01-18
2019	기본 DB		검진조사, 건강설문조사, 영양조사	기본DB	다운로드	다운로드	2023-12-07
2019	분과별 상세DB	검진조사	구강검사	HN19_OE	다운로드	다운로드	2023-01-18
2019	분과별 상세DB	영양조사	식이보충제조사	HN19_SUP	다운로드	다운로드	2023-12-07
2019	분과별 상세DB	영양조사	식품섭취조사(개인별 24시회상조사)	HN19_24RC	다운로드	다운로드	2023-01-18
2018	기본 DB		검진조사, 건강설문조사, 영양조사	기본DB	다운로드	다운로드	2023-12-07
2018	분과별 상세DB	검진조사	구강검사_제7기(2016-2018) 통합	HNY7_OE	다운로드	다운로드	2022-04-05
2018	분과별 상세DB	영양조사	식이보충제조사	HN18_SUP	다운로드	다운로드	2023-12-07
2018	분과별 상세DB	영양조사	식품섭취조사(개인별 24시회상조사)	HN18_24RC	다운로드	다운로드	2023-01-18

- 학습 및 테스트에 사용
- 기존 건강검진 데이터와 유사하지만, 추가 설문 결과 및 건강검진에서 진단 가능한 질병의 유병 여부 또한 나타나있음 (질병 유병 여부는 설문 및 규칙 기반으로 구성)

* KNHANES
https://knhanes.kdca.go.kr/knhanes/sub03/sub03_02_05.do

데이터 전처리 - KOSIS 데이터셋 정제

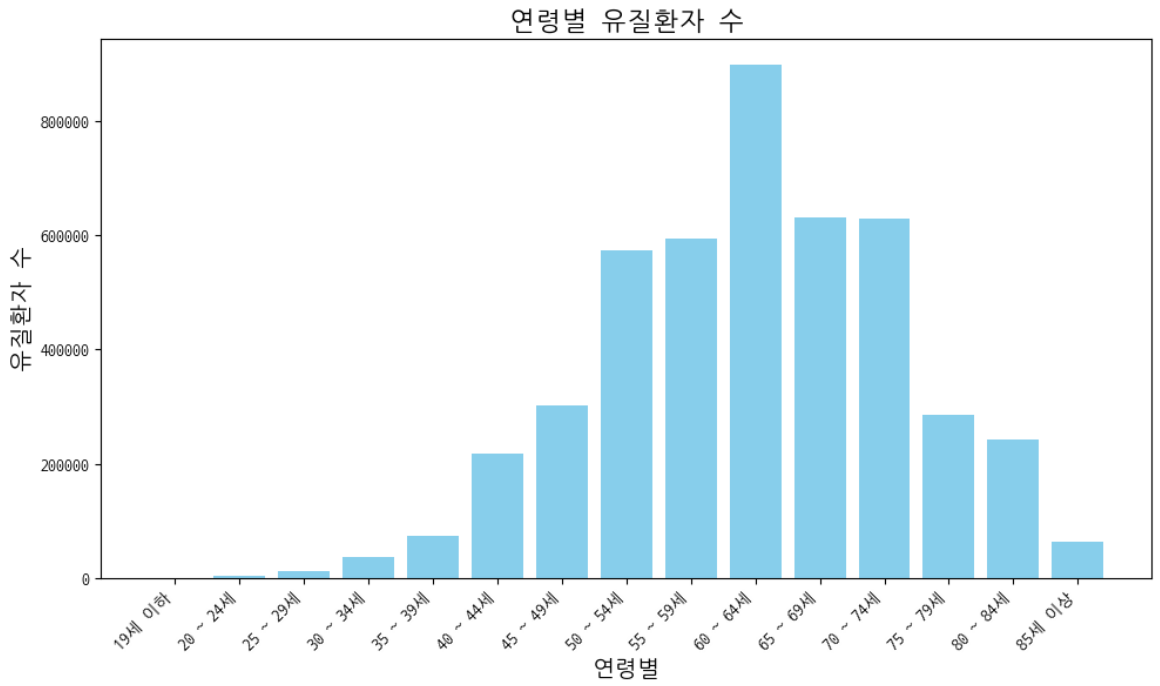
: KOSIS 데이터셋을 각 검진 수치 또는 유병 여부 별로 연령대에 따른 분포 정리



데이터 분석

: KOSIS 데이터 분포 분석

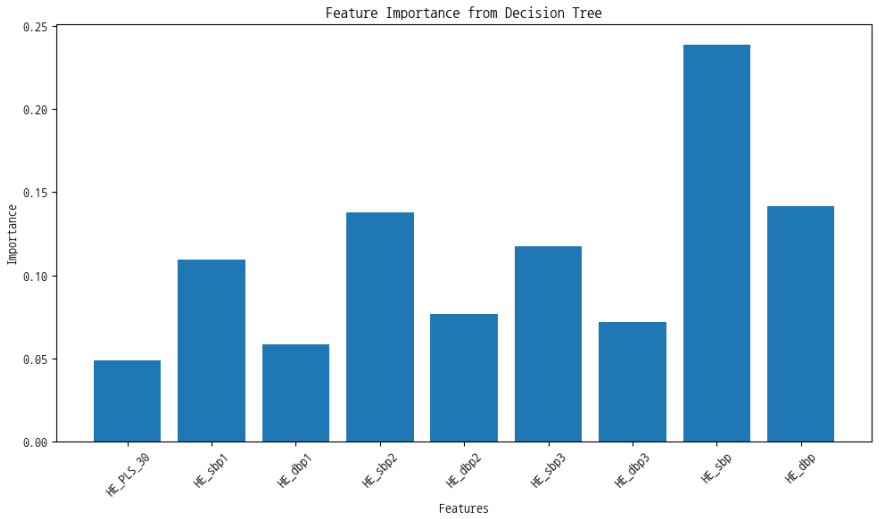
- KOSIS 데이터셋을 시각화한 결과, 유질환자가 고령에서 많이 나타난다는 것을 확인 할 수 있음
- 이를 통해 성별이나 연령에 대한 가중치 고려 가능



데이터 분석

: KNHANES Feature importance 분석

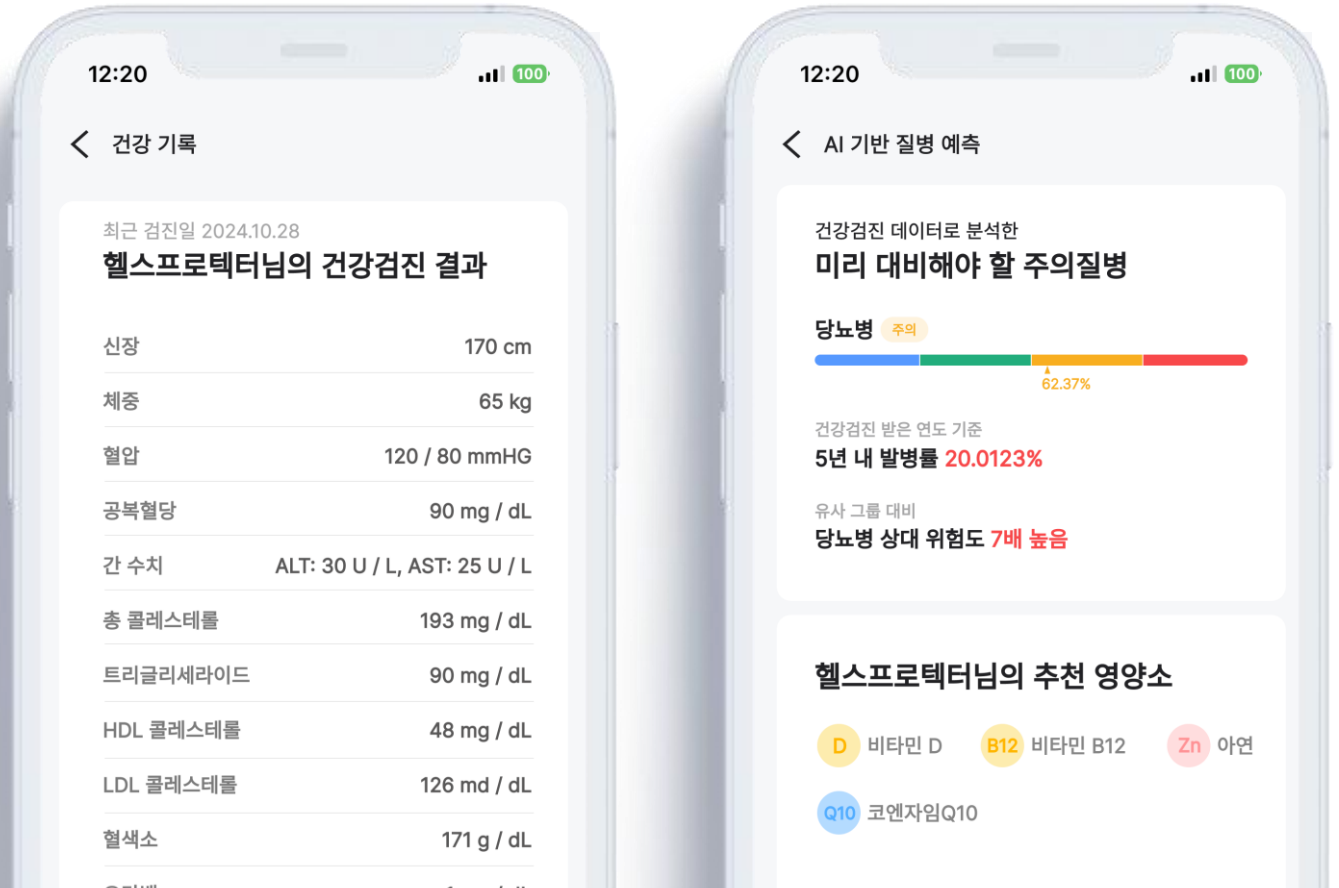
- KNHANES에서 비만, 고혈압, 당뇨병, 이상지혈증에 대한 여러 검사 수치의 상관관계 분석 진행
- Scikit-learn의 RandomForestClassifier에서 제공하는 feature_importance를 사용하여 분석
- 해당 분석을 통해 과도하게 상관관계가 높은 피쳐들을 제거하고 학습을 진행할 수 있음



- 건강검진 데이터셋에서 규칙 기반 라벨링 후 모든 피쳐(검사 수치)를 이용하여 모델을 학습시키면, 결국 규칙 기반 방법을 학습시키는 것이기에 의미가 없음
- 즉, 규칙 기반에 사용된 핵심 피쳐를 제외한 나머지를 이용하여 예측 성능을 유지하는 것이 필요함
- 다만, 현재 건강검진 데이터셋에서 규칙기반에 사용된 핵심 피쳐를 제거하면 성능이 매우 크게 떨어짐 (정확도 0.1 ~ 0.3 정도로 확인)
- 이는 건강검진 데이터셋에 존재하는 (핵심 피쳐를 제외한) 피쳐만으로는 질병 유병 여부를 설명하기 어려움을 의미
- 이때 KNHANES에서 조사한 선택형 설문 결과(ex 식사 빈도 또는 식단 조사 등) 및 추가적인 측정 피쳐들을 사용하면 성능을 올릴 수 있을 것으로 예측됨

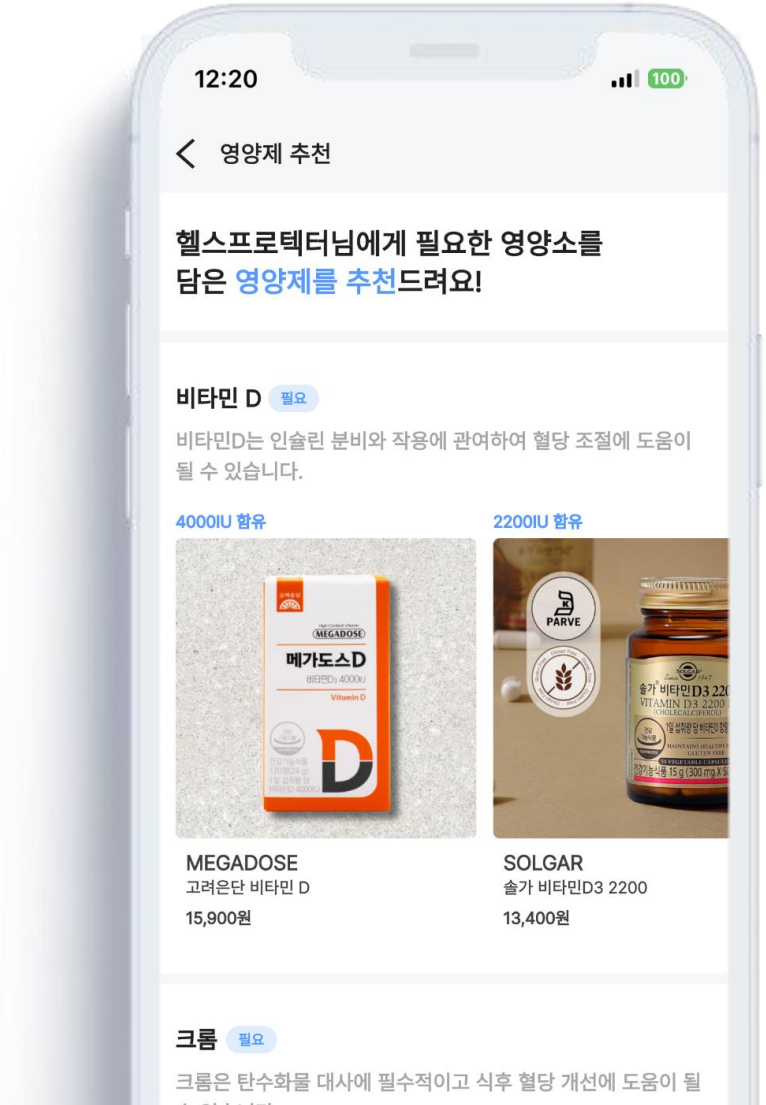
건강검진 결과를 토대로,
발병 가능성 있는 질병에 대해 위험도 제시

- KNHANES에서 조사한 선택형 설문 결과 및 추가적인 측정 피쳐들을 사용하면 성능을 향상시킬 수 있을 것으로 예상
- 사용자가 BMI 수치나 공복혈당 수치를 정확히 몰라도 서술형 설문조사를 통해 질병을 예측할 수 있는 사용자 친화적인 시스템 구현 가능
- 따라서 서술형 설문을 추가하여 자연어 모델을 이용한 분석 방법 또한 고려 중



질병 위험도에 따른
개인 맞춤형 건강보조식품 추천

- 안전성을 검증 받은 원료를 사용하는 제품인가?
- 같은 성분일지라도, 영양소 함량이 높은가?
- 품질 대비 가격이 합리적인가?



Q & A

한양대학교