헬스케어데이터사이언스

IC-PBL 과제 설명서

건강검진 데이터 분석을 통한
질병 예측 모델 개발

문제상황 시나리오 (1)

- 디지털 헬스케어는 최근 CES(Consumer Technology Association) 2024에서 중요한 키워드로 포함될 만큼 관심이 높아지고 있음
- 또한 개인 맞춤형 건강 정보 및 상품을 추천받고 싶어하는 수요가 증가하고 있음
- 건강기능식품 쇼핑몰을 운영하고 있는 ㈜글로잇은 개인의 고유한 건강 상태 등에 따라 자사의 상품을 개인에게 추천하고자 함
- 현재 기업에서 보유하고 있는 데이터는 국가건강검진 데이터로 해당 건강검진 정보를 통해 예측할 수 있는 질병을 발굴하고, 예측 모델 및 알고리즘을 개발하고자 함

문제상황 시나리오 (2)

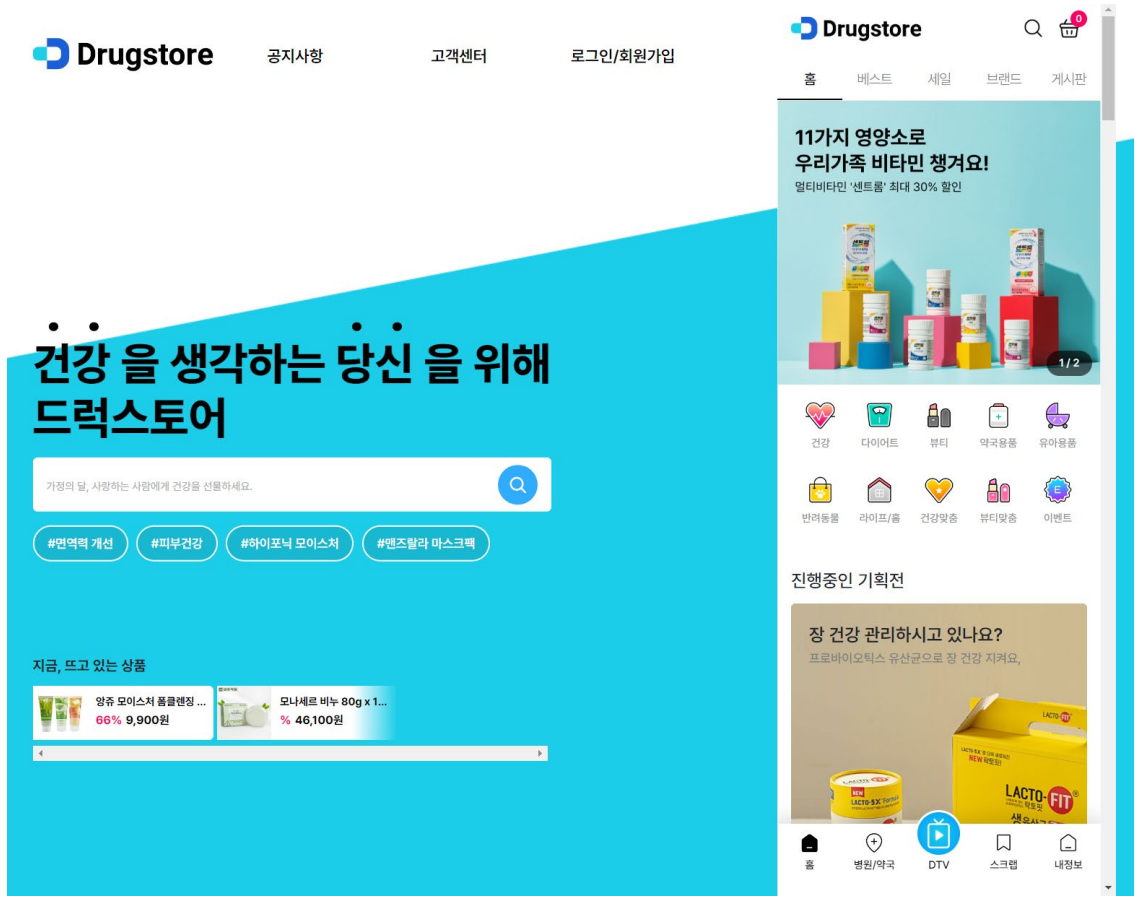
- 요구사항
  - 국가 건강검진 데이터를 통해 도출 할 수 있는 질병 및 특정 의학적 상태에 대한 발굴
    예) 공복혈당 수치를 통한 당뇨병 단계별 진단
  - 예측 모델 및 알고리즘 개발을 통해 가능성 있는 질병에 대해 위험도 제시
  - 시중에 유통되고 있는 건강보조식품과 연계하여 사용자의 의학적 상태를 개선해 줄 수 있는 상품 연결

- 참여자의 역할:
  - 주식회사 글로잇의 데이터사이언티스트로 이번 프로젝트를 수행함

문제상황 시나리오 (3)

- 기업 정보

  https://www.drugstore.kr/



데이터 소개

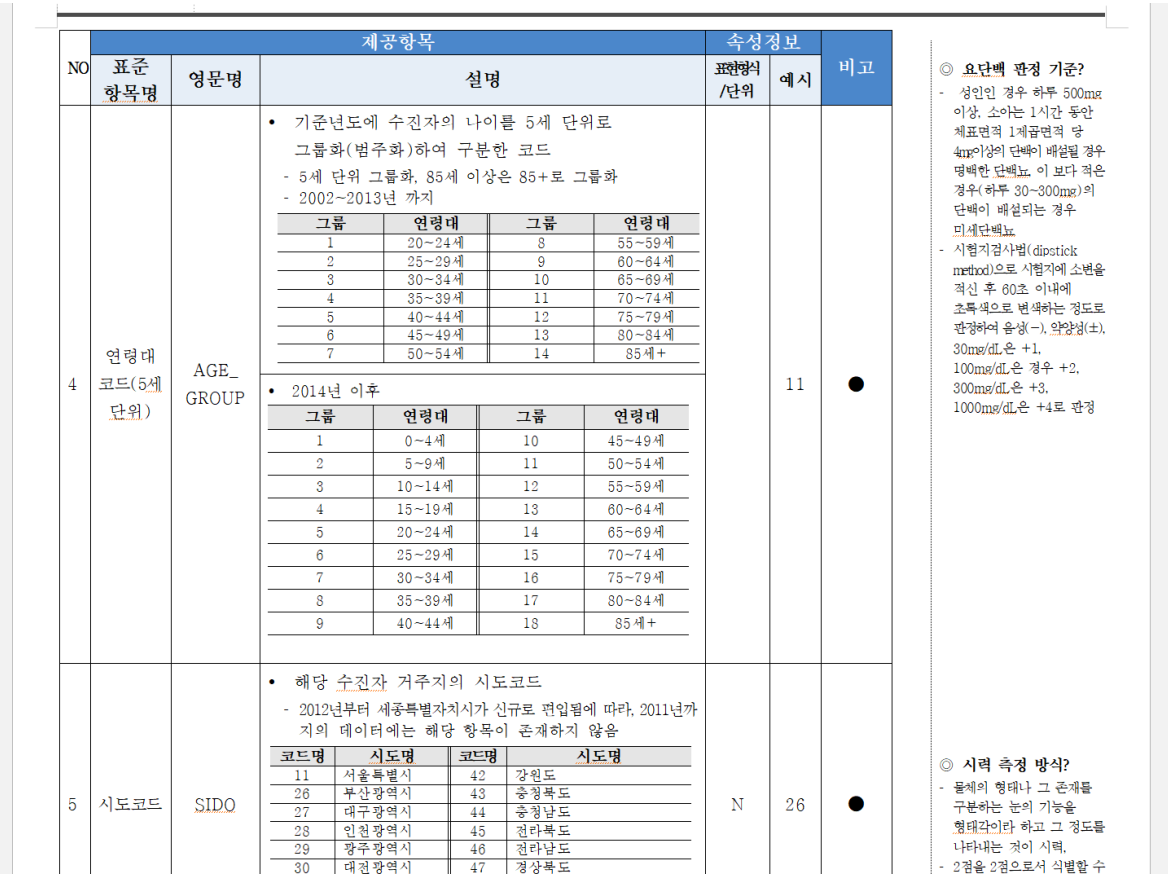
- 국가 건강검진 데이터
  - 1,000,000건

	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	트리글리서	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

건강검진데이터 샘플

데이터 소개

- 국가 건강검진 데이터 사용 가이드



NO	제공항목			속성정보		비고
	표준 항목명	영문명	설명	표현형식 /단위	예시	
4	연령대 코드(5세 단위)	AGE_ GROUP	• 기준년도에 수진자의 나이를 5세 단위로 그룹화(범주화)하여 구분한 코드 - 5세 단위 그룹화, 85세 이상은 85+로 그룹화 - 2002~2013년 까지		11	●
5	시도코드	SIDO	• 해당 수진자 거주지의 시도코드 - 2012년부터 세종특별자치시가 신규로 편입됨에 따라, 2011년까지의 데이터에는 해당 항목이 존재하지 않음	N	26	●

그룹	연령대	그룹	연령대
1	20~24세	8	55~59세
2	25~29세	9	60~64세
3	30~34세	10	65~69세
4	35~39세	11	70~74세
5	40~44세	12	75~79세
6	45~49세	13	80~84세
7	50~54세	14	85세+

• 2014년 이후

그룹	연령대	그룹	연령대
1	0~4세	10	45~49세
2	5~9세	11	50~54세
3	10~14세	12	55~59세
4	15~19세	13	60~64세
5	20~24세	14	65~69세
6	25~29세	15	70~74세
7	30~34세	16	75~79세
8	35~39세	17	80~84세
9	40~44세	18	85세+

코드명	시도명	코드명	시도명
11	서울특별시	42	강원도
26	부산광역시	43	충청북도
27	대구광역시	44	충청남도
28	인천광역시	45	전라북도
29	광주광역시	46	전라남도
30	대전광역시	47	경상북도

◎ 요단백 판정 기준?
- 성인인 경우 하루 500mg 이상, 소아는 1시간 동안 체표면적 1제곱면적 당 4mg이상의 단백이 배설될 경우 명백한 단백뇨. 이 보다 적은 경우(하루 30~300mg)의 단백이 배설되는 경우 미세단백뇨
- 시험지검사법(dipstick method)으로 시험지에 소변을 적신 후 60초 이내에 초록색으로 변색하는 정도로 판정하여 음성(-), 약양성(±), 30mg/dL은 +1, 100mg/dL은 경우 +2, 300mg/dL은 +3, 1000mg/dL은 +4로 판정

◎ 시력 측정 방식?
- 물체의 형태나 그 존재를 구분하는 눈의 기능을 형태각이라 하고 그 정도를 나타내는 것이 시력.
- 2점을 2점으로서 식별할 수

건강검진데이터 사용 가이드 샘플

학습내용

- 대규모의 건강검진 정보를 통해 데이터를 분석하고 예측 모델을 개발하는 방법에 대해 학습한다
- 완성된 모델을 실제 건강기능식품 추천까지 연결하여 인공지능 프로젝트의 전 과정에 대해 학습한다

핵심 학습 목표

- 의료 데이터 분석 및 모델링 능력 향상
- 의료 인공지능 모델의 활용 방법 학습