헬스케어 데이터 분석을 통한 입원 여부 예측 서비스 개발

A분반 2조 데이터뿌시기
김용국, 김진영, 조정범, 손학영

프로젝트 개요

1

프로젝트 주제

상용 헬스케어 데이터 분석을 활용한 입원 여부 예측 서비스 개발

2

프로젝트 진행 배경

- 의료정보에 대한 접근성 ↑
- 대중적인 상용 헬스케어 서비스의 부재
- 쉽게 이용하는 헬스케어 서비스를 개발하여 누구나 쉽게 의료데이터에 접근할 수 있게 함

3

프로젝트 진행 순서 및 도구

- 순서: 주제선정, 데이터 전처리, 시각화, 모델링, 서비스 개발, 발표
- 도구 : Colab, VS Code/ Numpy, Pandas/ Scikit Learn/ Matplotlib, Seaborn, Plotly/FastAPI, WIX

4

프로젝트 진행 내용

- 데이터 탐색 및 Target과 Feature간의 관계 시각화
- 종속변수(입원여부)에 영향을 미치는 적합한 독립변수 판단
- Scikit-Learn을 활용한 예측모델 구현
- Fast-API를 활용하여 Web 구현

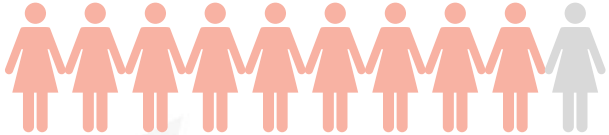
팀 구성 및 역할

A반 2조 데이터뿌시기

100%

손학영

데이터 전처리, 데이터 시각화, 코드병합, 데이터 모델링, FastAPI, HTML,CSS 발표



90%

조정범

데이터 전처리, 데이터 시각화, 데이터 모델링, FastAPI(HTML), WIX, CSS, 프로젝트 관리

80%

김진영

데이터 전처리, 데이터 시각화, 데이터 모델링, 발표자료 준비

70%

김용국

데이터 전처리, 데이터 시각화, 데이터 모델링

프로젝트 Work-flow & 데이터 상세

Work-flow

구분	기간	활동	도구
사전기획	9월 27일 ~ 28일	주제 선정, 자료분석 및 탐색, 데이터 전처리, 시각화	Colab, VS Code, GitHub
	9월 29일	주제 선정, 기획서 제출, 일정 수립, 자료조사 및 데이터 전처리, 시각화	Colab, VS Code, GitHub, MS Word
분석	9월 30일 ~ 10월 1일	데이터 전처리, 시각화, 모델링	Colab, VS Code, GitHub
	10월 4일	모델링, Fastapi, html	Colab, VS Code, GitHub
수정 / 보완 / 발표준비	10월 5일	데이터 합치기, Fastapi, html, CSS, WIX, 발표자료 준비	Colab, VS Code, GitHub, Fastapi, WIX, PowerPoint
발표준비	10월 6일	발표 준비 및 연습	PowerPoint
프로젝트 발표	10월 6일	발표	PowerPoint

데이터 상세

출처	데이터 이름	제공형태	요약
공공데이터포털	국민건강보험공단_진료내역정보	CSV	국민건강보험공단 진료내역정보 (13178345, 19) 기간: 2019.01.01 – 2019.12.31

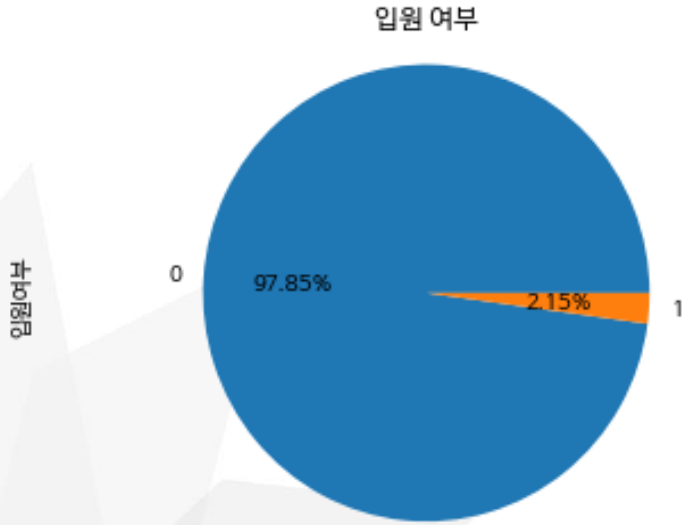
시각화를 위한 Pre-processing

- Target : **입원 여부 (0,1)**
- 입원 정의 : 서식코드 02, (04, 06, 07, 10, 12)
- 현재 데이터는 02(의과입원), 03(의과외래), 08(보건기관외래) 존재
- 범주 재설정, 숫자형 코드 → 한글 범주, 새로운 Feature 생성(입원여부, 요일) 등

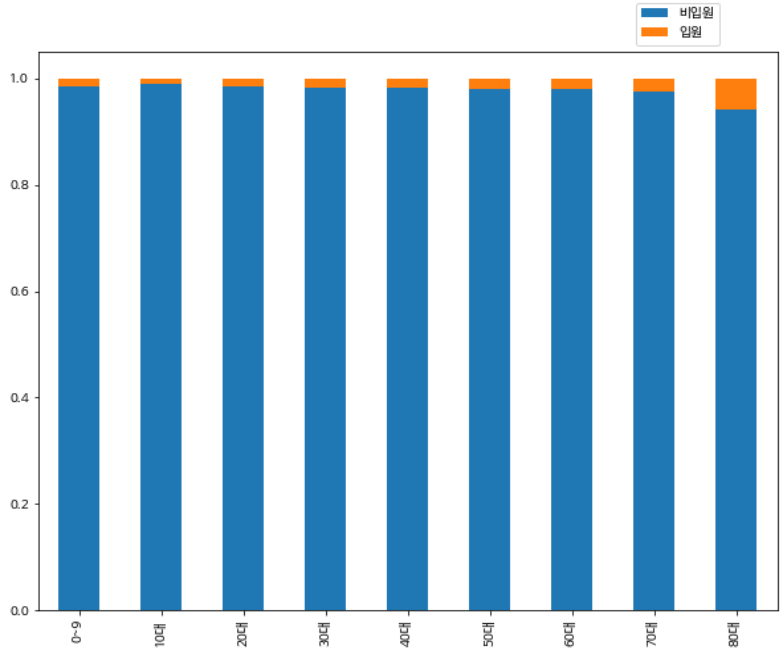
년도	성별 코드	연령대코드	시도코드	요양개시일자	서식코드	진료과목 코드	주상병 코드	부상병 코드	요양일수	입내원일수	심결가산율	심결요양급여 비용총액	심결본인 부담금	심결보험 자부담금	총처방일 수	데이터 기준 일자
2020	2	15	41	20200309	3	1	I109	H814	1	1	0.15	11540	1500	10040	0	20210929
2020	2	16	41	20200123	3	1	I109	I209	1	1	0.15	11540	1500	10040	0	20210929
2020	2	15	41	20200108	3	1	I109	E785	1	1	0.15	11540	1500	10040	0	20210929
2020	2	15	41	20200106	3	1	I109	M170	1	1	0.15	11540	1500	10040	0	20210929
2020	2	15	41	20200121	3	1	I109	E782	1	1	0.15	11540	1500	10040	0	20210929

EDA 탐색적 데이터 분석

변수 특수성 시각화



입원한 경우의 비율

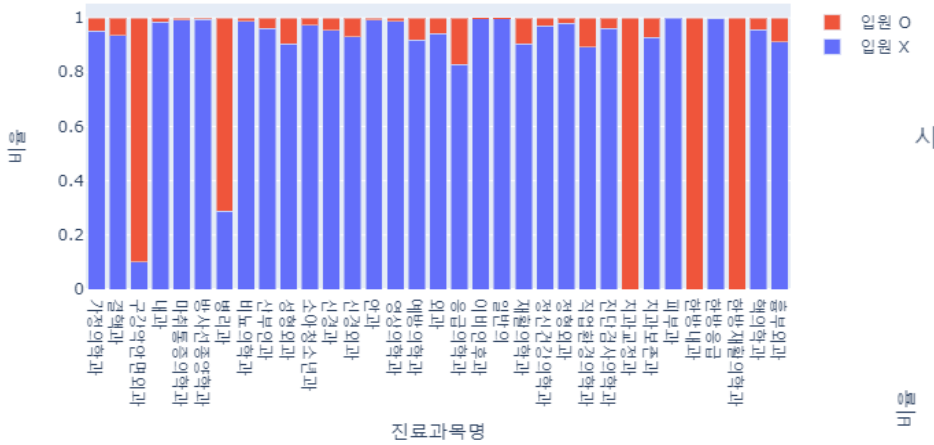


연령대 - 입원여부

EDA 탐색적 데이터 분석

변수 특수성 시각화

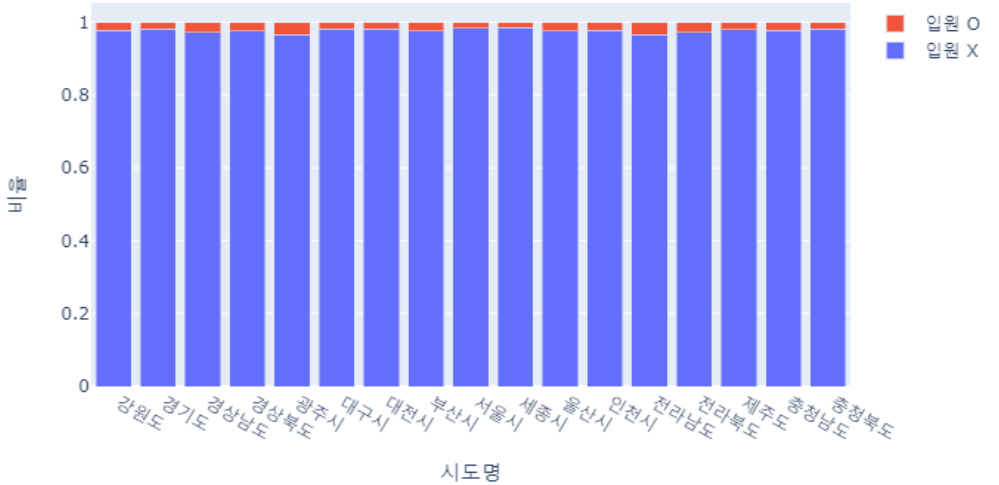
진료과목별 입원 여부 비율



진료과목 - 입원여부

진료받은 시도 - 입원여부

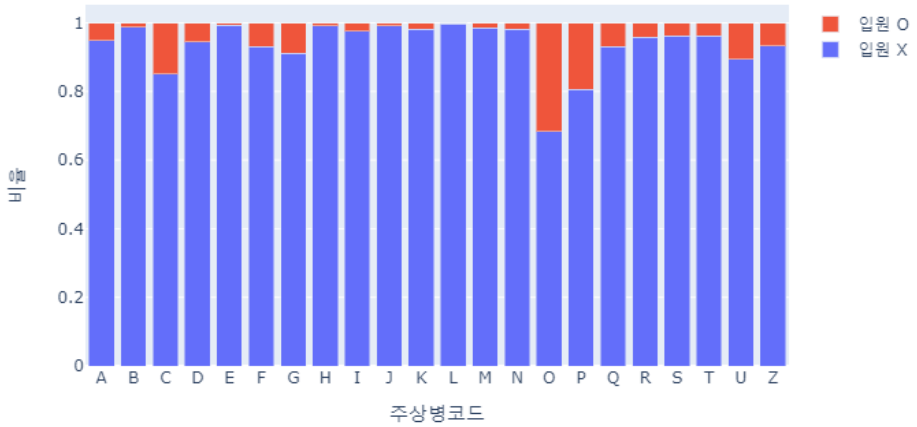
시도별 입원 여부 비율



EDA 탐색적 데이터 분석

변수 특수성 시각화

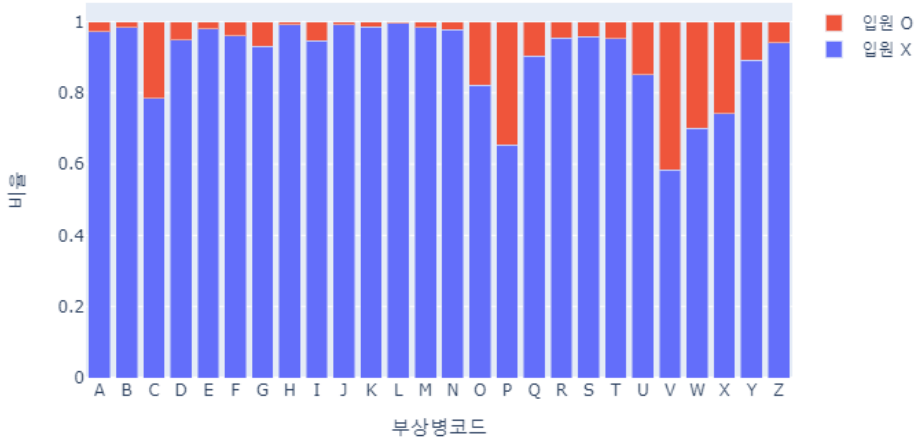
주상병코드 별 입원 여부 비율



주상병코드 - 입원여부

부상병코드 - 입원여부

부상병코드 별 입원 여부 비율



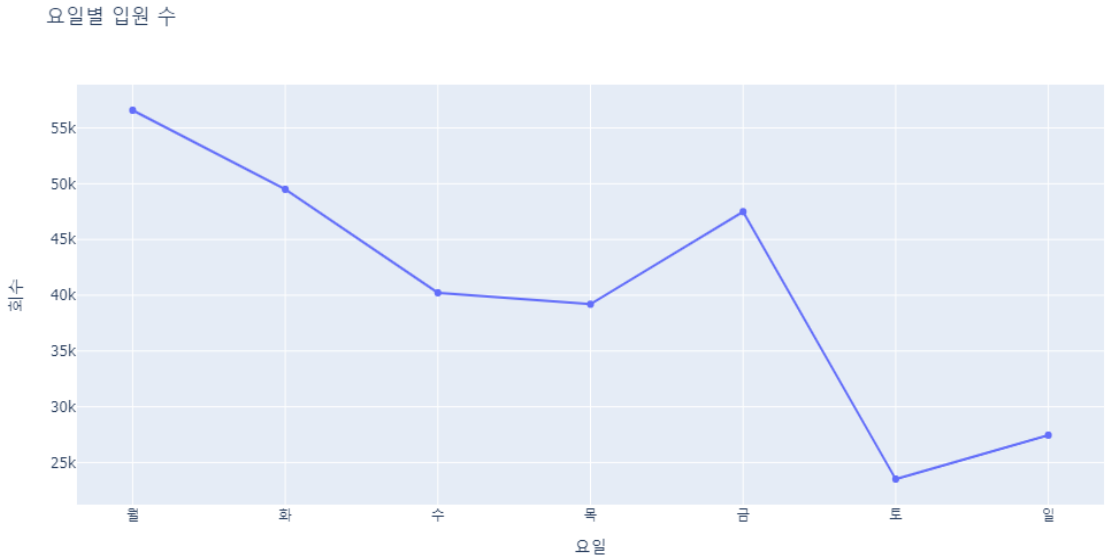
EDA 탐색적 데이터 분석

변수 특수성 시각화

요양개시일자

일자
계절

요일



요일 - 입원여부

탐색 결과

Target Class의 불균형

97.85%
비입원

2.15%
입원

"UnderSampling"

50%
비입원

50%
입원

Feature와 Target 변수 간의 상관성

카이제곱검정 – 관찰된 빈도가 기대되는 빈도와 의미 있게 다른지 검정하기 위해 범주형 자료 검정에 사용

Before Undersampling

```
성별코드 : The P-Value of the ChiSq Test is: 1.4459695826147777e-15
연령대코드 : The P-Value of the ChiSq Test is: 0.0
시도코드 : The P-Value of the ChiSq Test is: 0.0
진료과목코드 : The P-Value of the ChiSq Test is: 0.0
주상병코드 : The P-Value of the ChiSq Test is: 0.0
부상병코드 : The P-Value of the ChiSq Test is: 0.0
심결가산율 : The P-Value of the ChiSq Test is: 0.0
요일 : The P-Value of the ChiSq Test is: 0.0
```

After Undersampling

```
성별코드 : The P-Value of the ChiSq Test is: 5.695149210988395e-09
연령대코드 : The P-Value of the ChiSq Test is: 0.0
시도코드 : The P-Value of the ChiSq Test is: 0.0
진료과목코드 : The P-Value of the ChiSq Test is: 0.0
주상병코드 : The P-Value of the ChiSq Test is: 0.0
부상병코드 : The P-Value of the ChiSq Test is: 0.0
심결가산율 : The P-Value of the ChiSq Test is: 0.0
요일 : The P-Value of the ChiSq Test is: 0.0
```

카이제곱검정 결론:
모든 Feature에서 P-value가 0.05이하 → Feature와 Target의 상관성 확인

Modeling

분류 모델

1 K-NN

2 Naive Bayes

3 Logistic Regression

4 Decision Tree

모델링 전 데이터 전 처리 - Undersampling 적용 및 Feature를 숫자-카테고리 변수로 변환

```
Before undersampling:  Counter({0: 12894383, 1: 283962})
After undersampling:  Counter({0: 283962, 1: 283962})
```

KNN

Parameter : K = 10, weights = 'distance'

모델생성 → FIT → SCORE → K-fold cross Validation

k	Accuracy score	Cross validation score
k = 3	0.8227580	NA
k = 5	0.8266881	NA
k = 7	0.8277305	NA
k = 10	0.8286672	0.82900881 0.82980117 0.82707545 0.83063579 0.8283221 0.83028715

Naive Bayes

Multinomial

모델생성 → FIT → SCORE → K-fold cross Validation

Parameter estimation and event models	Accuracy score	Cross validation score
Multinomial naïve Bayes	0.60743338897	0.60666216 0.60478163 0.6079405 0.60972595 0.6086906 0.60576415

Logistic Regression

모델생성 → FIT → SCORE → K-fold cross Validation

Model	Accuracy score	Cross validation score
Logistic Regression	0.79730386460	0.79727217 0.79686014 0.79631077 0.79712426 0.79484227 0.79482114

Decision Tree

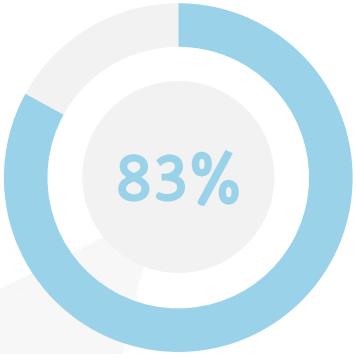
Parameter : depth=11, 불순도="gini계수"

모델생성 → FIT → SCORE → K-fold cross Validation

max_depth	Accuracy score	Cross validation score
3	0.84711334615	NA
6	0.84955029194	NA
11	0.85623428487	0.8574915 0.85784013 0.85677309 0.856921 0.8546707 0.85862193
14	0.85939667983	NA

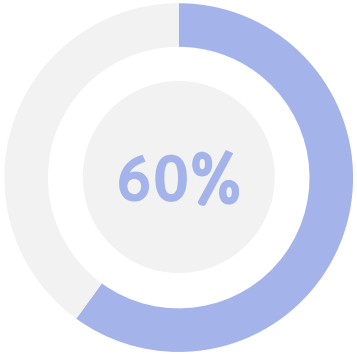
Modeling Accuracy

분류 모델 별 모델링 정확도

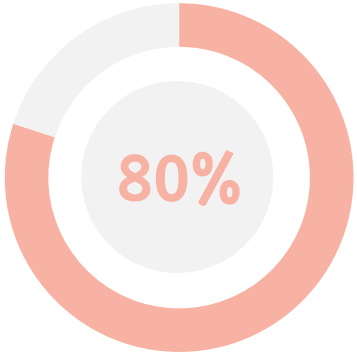


KNN

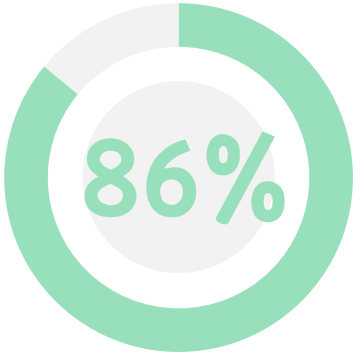
k = 10



Naive Bayes



Logistic Regression



Decision Tree

max_depth = 11

입원 여부 예측 서비스

FastAPI – Web View – WIX

결론

프로젝트 결과 및 향후 발전 방향

결과

나이, 연령대, 성별과 같은 환자 정보데이터와
진료과목, 질병코드와 같은 의료 데이터 간의 **연관성 분석**

사전 정보를 입력한 환자의
입원여부를 미리 확인할 수 있는 **예측 서비스**를 제작

향후 발전 방향

입원예측모델을 기반으로 의료정보와 치료비용과의 연관성을 밝혀
비용산출 모델 등을 개발하여

서비스 이용자(일반환자 등) 모두가
의료관련 정보에 **쉽게 접근 및 활용**할 수 있도록 **서비스 개선**

소감

이름	잘한 점	아쉬운 점	갈등	해결방안
김용국	새로운 것을 시도하면서 모르는 부분이 많았지만 **배우면서 포기하지 않고** 프로젝트를 마무리하기 위해 **노력**했다.	평소 수업 시간에 배울 때는 적용하기 쉬울 것 같던 내용도 **실제로 적용**하려고 하니 생각하지 못했던 **문제점들**이 많이 있었다. 더 많은 실습을 통해 익숙해지는 시간이 필요하다.	이론을 배우는 것과 **실제 활용하는 것의 난이도** 차이를 실감할 수 있었다.	모르는 것은 **다른 분들에게 물어보거나 검색을 통해 배워나갔다**. 배운 이론들을 익숙하게 활용하기 위해 연습을 생각보다 더 많이 해야 한다는 것을 느꼈다.
김진영	서로의 실력차이를 인지하고 역할분담을 하였고, 참여를 독려하여 **의사소통이 활발**하게 되었다.	다른 팀원들에 비해 **실력이 뒤떨어져** 코딩에 어려움이 있었고 팀에 도움이 되지 못했다. 기회가 된다면 연속형 데이터도 다뤄보고 싶다.	실력차이로 인해 일부 팀원들의 **참여도가 떨어졌다**.	다른 팀원들이 꾸준히 참여를 독려하고, **어려운 부분을 함께 이야기**하며 참여율을 높였다.
조정범	**부족한 점을 인정**하고, 열정을 가지고 끝까지 팀프로젝트를 완수하기 위해 노력하였으며, 프로젝트를 진행하는 내내 팀원들과 소통하였다.	데이터 분석프로젝트를 진행하는 것이 처음이다보니 주제 선정부터 서비스 구현까지 낭비되는 시간들이 많아 **보다 나은 결과물을 도출하지 못한** 것이 아쉽다. 하지만 그동안 배운 파이썬, 머신러닝들을 실제로 코딩해보고, 간단하게나마 서비스 구현을 위해 웍스, 깃허브를 활용한 웹호스팅 등과 같은 새로운 것들을 직접 찾아보고 실행해보면서 많은 것을 배울 수 있어서 나에게는 의미가 깊은 경험이 되었다고 생각한다.	실제로 데이터분석 프로젝트를 진행해본 것이 처음이라 보다 체계적으로 계획을 수립하고 단계적으로 진행하지 못하여 **시간을 보다 효율적**이고 효과적으로 쓰지 못하였다.	지식과 경험이 부족한 현재상태를 인정하고 다른 사람들과 최대한 소통을 하고 **의견을 교류**하며 프로젝트를 진행하였으며, 그 과정에서 팀원들한테서 코딩과 프로젝트를 진행만 하느라 놓쳤던 포인트들을 다시 보고, 나와 **다른 관점들을 공유**하면서 많은 것을 배울 수 있었다.
손학영	모델링을 처음 해봄에도 불구하고, 팀원 모두 구글링, 수업자료 등을 통해 **포기하지 않고 끝까지 완성**하였다.	Feature의 데이터 타입이 범주형만 있는 데이터여서 아쉬웠다. **연속형 변수**를 활용하여 스케일링 등도 해보면 좋았을 것 같다.	머신러닝을 이론과 기초 데이터로만 수행했었기에 분석주제부터 서비스 구현까지는 처음이라 **어떤 흐름으로 진행**해야 하는지 알아가는게 어려웠다.	**구글링**과 기존에 **학습했던 이론 자료**들을 바탕으로 단계 별 배운 이론과 코드를 적용시키고 **강사님께 질문**드려 해결했다. (구글링은 위대하다.)

감사합니다