

PBL 콘텐츠 설계서

PROJECT 1: No-show 분석



2021 March 30

매직에코



**X**

# 목차

[목차 1](#_Toc71560405)

[백그라운드 스토리 2](#_Toc71560406)

[학습해야 하는 키워드에 대한 설명 2](#_Toc71560407)

[환자 데이터 분석을 통한 No-Show 예측 2](#_Toc71560408)

[프로젝트의 목적 2](#_Toc71560409)

[텍스트 데이터 분석 및 학습 2](#_Toc71560410)

[수준 구분 3](#_Toc71560411)

[관련 기술 3](#_Toc71560412)

[Require Skill 3](#_Toc71560413)

[Acquire Skill 3](#_Toc71560414)

[미션 별 목표 및 평가기준 3](#_Toc71560415)

[Mission 1 – 환자 데이터 분석 3](#_Toc71560416)

[Mission 2 - No Show 예측 모델 개발 5](#_Toc71560417)

[모범 답안 6](#_Toc71560418)

[데이터 파일 설명 6](#_Toc71560419)

[Mission 1: Medical No Show 6](#_Toc71560420)

[Mission 2: Head CT-Hemorrhage 6](#_Toc71560421)

[학습 지원 정보 7](#_Toc71560422)

[학습 리소스 7](#_Toc71560423)

[포럼 7](#_Toc71560424)

# 백그라운드 스토리

예약 부도(No Show): 진료·수술 일시를 잡고 미리 병·의원에게 알리지 않고 환자가 본인 마음대로 예약을 취소하는 행위, 다른 환자가 진료받을 기회를 빼앗아 간다. 검사나 수술이 시급한 환자들의 빠른 치료 기회가 줄어든다는 점이 가장 큰 문제이다.

오늘도 바쁜 XX병원 데스크. 환자 응대하랴, 걸려오는 예약 전화 받으랴 정신이 없다. 오늘따라 방문을 희망하는 환자들이 많은데 이를 어쩌나, 오전과 오후 시간대 모두 수술 환자가 꽉 잡혀 있어서 다른 환자의 진료가 어렵다. 수술 예약 환자 방문을 10분 앞두고 전화가 또 걸려온다. "너무 아파서 그런데... 지금 바로 가서 진료 좀 볼 수 있을까요?" 하지만 도리없다. "조금 뒤 원장님이 수술을 들어가야 해서, 오늘은 진료가 힘드실 것 같습니다." 응대하는 데스크도 찝찝하다. 너무 아프다고 하는데 진료를 못 봐주니, 방법을 찾아주고 싶지만 불가능하다. 그리고 10분 뒤, 수술 예약한 환자는 오지도 않고 연락도 되지 않는다.

이러한 상황이 반복되니 XX병원 혁신팀은 노쇼를 방지하고자 환자들에게 '예약 선입금'을 시도하려 했지만, 신규 환자 유치를 막는다는 이유로 병원장이 퇴짜를 놓았다. 그리하여 혁신팀은 차선책으로 '예약 대기 환자 시스템'을 제안했다. 예약 대기 환자 시스템이란 부득이 예약이 취소되는 경우를 대비하여 대기 예약 환자를 배정하는 정책이다. 물론 먼저 예약했던 환자가 제시간에 도착한다면 예약 대기 환자는 진료를 받을 수 없지만, 이전의 예약이 부도가 난다면 (노쇼가 발생한다면) 우선 순위는 대기환자에게 돌아가 진료를 받을 수 있게 한다.

하지만 이 시스템을 적용하기에도 아직 어려움이 남아있었다. 모든 시간에 예약 대기환자를 받는다면 예약 대기환자에게 우선순위가 돌아가는 확률은 매우 낮을 것으로 예상되어 예약 대기 환자 시스템을 운영하는 실효성이 떨어진다. 따라서 혁신팀은 환자들의 데이터를 활용하여 특정 환자가 예약을 부도낼 확률, 즉 노쇼 확률(No - Show Probability)를 계산하여 그 확률이 높은 환자들의 예약에만 대기 환자를 받자고 제안했다.

# 학습해야 하는 키워드에 대한 설명

## 환자 데이터 분석을 통한 No-Show 예측

환자에게 적절하고 정확한 의료 서비스를 제공하고 병원 자원을 효율적으로 사용하기 위 하여 대부분의 병원은 예약 진료 시스템을 사용하고 있다. 그러나 사회가 고령화 되어감에 따라 의료 서비스 요구가 증가하고, 이는 곧 예약의 증가, 예약 대기의 증가, 그리고 예약 부도 (no-show)의 증가로 이어지고 있다. 이러한 상황에서 환자들은 불필요하게 길어진 예약 대기 때문에 건강 상태의 불확실성이 증가하게 되며, 불필요한 응급 진료를 받는 경우가 발생하기도 한다. 또한, 병원에서는 의료진들의 진료 효율성이 저하되고 병원 자원을 비효율적으로 운영하게 되어 병원의 수익성이 악화될 수 있다.

no-show를 설명하고 예측하는 연구는 실제 데이터를 이용하여 no-show에 영향을 줄 수 있는 요인들을 나열하고 통계적 분석 등의 방 법을 통해 핵심 요소들을 찾는다. 그리고 이를 이용하여 병원 예약 관리에 시사점을 제시하거나, 대안을 제시한 뒤 적용 사례연구를 통해 대 안의 효과에 대해 분석하고, 환자 별 no-show 확률에 대해서 예측하기도 한다.

프로젝트의 목적

텍스트 데이터 분석 및 학습

|  |  |
| --- | --- |
| 구분 | 설명 |
| **1** | 데이터 저장 및 준비 등 데이터 전처리 과정을 경험할 수 있다. |
| **2** | 데이터를 자세하게 분석하고 해당 내용에 대해 설명할 수 있다. |
| **3** | 모델링을 경험해볼 수 있다. |

# 수준 구분

* AI/Data분석/Edu4
* AI/모델개발/Edu4
* AI/AIOps/Edu4

# 관련 기술

### Require Skill

* Fundamentals > Python Programming > Python Basics
* Fundamentals > Python Programming > Numpy
* Fundamentals > Python Programming > Pandas
* Data Scientist > Visualization > Matplotlib
* Data Scientist > Visualization > seaborn
* Machine Learning > General
* Machine Learning > Tools > scikit – learn

### Acquire Skill

* Data Engineer > Summary of Data Formats
* Data Engineer > Data Discovery
* Data Engineer > Data Integration
* Data Engineer > Data Fusion
* Deep Learning > Tools > Tensorflow
* Deep Learning > Tools > PyTorch

# 미션 별 목표 및 평가기준

## Mission 1 – 환자 데이터 분석

**키워드**

NumPy, Pandas, matplotlib, seaborn, data analysis

**설명**

* 환자 데이터를 읽어오고 데이터를 전처리한다.
* 이 때, 데이터에 존재하는 오타를 수정하고 데이터를 학습을 위한 데이터 타입으로 변환한다.
* 필요하다면 학습을 위해 새로운 데이터를 생성해도 된다.
* 이 후 데이터들을 분석 분석한 내용을 토대로 보고서를 작성하고 데이터 분석에 관한 발표 동영상을 만들어라

**목표**

|  |  |  |
| --- | --- | --- |
| 분류 | 경험 | 난이도 |
| Data분석(데이터 추출 및 가공 역량) | 수집된 데이터의 속성을 결정하고 관련 내용을 정리할 수 있다. | 3 |
| Data분석(데이터 추출 및 가공 역량) | 수집된 데이터에서 잘못된 데이터를 찾아 수정 삭제할 수 있다 | 3 |
| Data분석(데이터 추출 및 가공 역량) | 시각화 도구를 사용해 수집된 데이터를 가공할 수 있다. | 3 |
| Data분석(데이터 추론 역량) | 모집단으로부터 추출된 표본을 바탕으로 통계적 추론이 가능하다 | 4 |
| Data분석(데이터 추론 역량) | 대용량의 자료에서 데이터의 관계, 패턴, 규칙을 탐색하고 활용할 수 있다 | 4 |

**결과물**

* 데이터 분석 보고서 (형태: 자유)
* 데이터 분석 발표 동영상

**평가 기준**

|  |  |
| --- | --- |
| 멘토 평가 항목 | 구분 |
| 1. 구글드라이브에 저장했는가? | Data Store |
| 2. try except의 사용으로 데이터 엑세스 에러에 대처했는가? | Data Store |
| 1. 오타를 고쳤는가? | Data Explore & Preprocess |
| 2. 학습에 필요없는 데이터를 삭제했는가? | Data Explore & Preprocess |
| 3. 데이터 타입을 변환했는가? | Data Explore & Preprocess |
| 4. 데이터 에러를 확인하려는 노력을 했는가? (검출하지 못했어도 다양하게 노력한 것으로만 점수 부여) | Data Explore & Preprocess |
| 5. 새로운 데이터를 생성했는가? (높은 점수 부여) | Data Explore & Preprocess |
| 6. 많은 데이터를 분석할수록 높은 점수 부여 | Data Analysis |
| 7. 보고서를 잘 만들수록 높은 점수 부여 | Data Analysis |
| 8. 발표를 잘할수록 높은 점수 부여 | Data Analysis |

**확인 질문**

|  |  |  |
| --- | --- | --- |
| 분류 | 문제 | 정답 |
| Data Explore & Preprocess | 판다스 데이터 프레임을 간단하게 요약해서 보여주며, 주로 데이터 타입을 확인할 때 사용하는 메소드는? | info() (sample(), describe()을 적었을 경우 점수 반만 부여) |
| Data Explore & Preprocess | 열에 어떤 데이터들이 있는지 확인해 보고 싶을 때 사용하는 메소드는? | set(), unique() (둘중 하나만 적어도 정답) |
| Data Analysis | 판다스 데이터의 개수를 세기 위해서 사용하는 메소드. 예를 들어 no show의 yes와 no의 개수를 세기 위해서 어떤 메소드를 사용해야 하는가? | size() |
| Data Analysis | 병원은 쉬는 요일이 있다. 언제인가? | 토요일, 일요일 |

## Mission 2 - No Show 예측 모델 개발

**키워드**

Decision Tree Classifier, Random Forest Classfier, Grid Search CV, Bayes Classifier, XGBoost, Logistic Regression

**설명**

* 현재까지 가공 및 분석한 환자 데이터를 토대로 CNN 모델을 개발한다.
* 이 때, 데이터를 2 개 이상의 알고리즘으로 모델링하고 score를 확인하라
* 또한 팀원들 앞에서 사용한 알고리즘을 간단히 설명하고 모델링 결과를 설명하라

**목표**

|  |  |  |
| --- | --- | --- |
| 분류 | 경험 | 난이도 |
| Data분석(데이터 추론 역량) | 모델링을 위한 성능 평가 지표를 정의하고 구현할 수 있다 | 4 |
| AI모델개발(ml/dl/기초 지식 이해력) | 머신러닝과 관련된 공개된 학습모델(사이킷런 라이브러리의 문서 등)을 따라서 모델을 돌려볼 수 있다. | 2 |

**결과물**

* 코드
* 모델링 결과 발표 동영상

**평가 기준**

|  |  |
| --- | --- |
| 멘토 평가 항목 | 구분 |
| 1. 많은 알고리즘으로 모델링할 수록 높은 점수 부여 | Data Modeling |
| 2. score가 높은 순서대로 높은 점수 부여 | Data Modeling |
| 3. 발표를 잘할수록 높은 점수 부여 | Data Modeling |
| 4. 알고리즘에 대해서 잘 설명한 순서대로 높은 점수 부여 | Data Modeling |

**확인 질문**

|  |  |  |
| --- | --- | --- |
| 분류 | 문제 | 정답 |
| Data Modeling | 모델의 학습을 위해 data type이 object인 것을 숫자 값으로 변환해 주어야 한다. 어떤 메소드를 어떻게 사용해야 하는지 서술하라. | sklearn의 labelEncoder()와 fit\_transform()을 사용하여 범주형 데이터를 정수형 유효값으로 치환한다. |
| Data Modeling | 모델의 정확도를 측정하기 위해 사용하는 메소드는? | score() |

# 모범 답안

<https://github.com/edu10make/Project-X_/blob/main/Project1-Medical%20Analysis/001_Medical%20No%20Show/001_Medical%20No%20Show.ipynb>

# 데이터 파일 설명

## Medical no show.csv

브라질 환자들의 병원 예약 정보와 예약 취소 여부에 대한 데이터

* PatientId: 환자 ID
* AppointmentID: 예약 ID
* Gender: 성별
* DataMarcacaoConsulta: 예약 날짜
* DataAgendamento: 진료 날짜
* Age: 나이
* Neighbourhood: 지역
* Scholarship: 장학금
* Hipertension: 고혈압
* Diabetes: 당뇨
* Alcoholism: 알콜중독
* Handcap: 장애
* SMS\_received: 환자에게 보낸 문자
* No-show: 예약취소

# 학습 지원 정보

## 학습 리소스

* <https://en.wikipedia.org/wiki/Convolution_neural_network>  
  : Convolution Neural Network(합성곱 신경망)에 대한 일반적인 이해 (위키피디아)
* <https://en.wikipedia.org/wiki/Scikit-learn>  
  : scikit-learn에 대한 일반적인 이해 (위키피디아)
* <https://en.wikipedia.org/wiki/CT_scan>  
  : CT scan에 대한 일반적인 이해 (위키피디아)
* [TensorFlow KR 페이스북 그룹](<https://www.facebook.com/groups/TensorFlowKR>) : TensorFlow를 학습하거나, TensorFlow, 다른 Machine Learning 도구에 관한 연구, 응용 제작을 하는 분들의 모임. 질의 응답이 매우 활발함
* •[Machine Learning Yearning](<https://www.deeplearning.ai/machine-learning-yearning/>) : 앤드유 응교수의 ML 프로젝트를 진행할 때 고려해야 할 우선순위, 성능에 대한 고려 사항 (딥러닝 기술 자체를 다룬 책은 아니라 필독). PDF 버전을 인터넷에서 쉽게 구할 수 있음
* •[모두를 위한 딥러닝]([https://www.inflearn.com/course/기본적인-머신러닝-딥러닝-강좌#](https://www.inflearn.com/course/기본적인-머신러닝-딥러닝-강좌%23)) : 머신 러닝과 딥러닝에 대해 더 이해하고 본인들의 문제를 TensorFlow를 이용하여 풀 수 있게 도와주는 동영상 강의. 이 강좌는 수학이나 컴퓨터 공학적인 지식이 없어도 어렵지 않게 볼 수 있음.
* <https://www.youtube.com/playlist?list=PLS8gIc2q83OjStGjdTF2LZtc0vefCAbnX&ab_channel=NeoWizard>  
  : 머신 러닝, 딥러닝
* [https://www.youtube.com/watch?v=AHdGc4nR5f4&list=PLS8gIc2q83OhM0RTktKDitgZGX5dHo7Vs&ab\_channel=NeoWizard](https://www.youtube.com/watch?v=AHdGc4nR5f4&list=PLS8gIc2q83OhM0RTktKDitgZGX5dHo7Vs&ab_channel=NeoWizard%20)   
  : TensorFlow

# 포럼

* 학습 리소스: <https://github.com/edu10make/Project-X_>
* 토론과 Q&A: https://prj.innoaca.kr/forum/project/2/forum
* 코드리뷰: https://prj.innoaca.kr/forum/project/2/codereview
* 동료평가: https://prj.innoaca.kr/forum/project/2/evaluation
* 멘토: https://prj.innoaca.kr/forum/project/2/mentor
* 동료목록: https://prj.innoaca.kr/forum/project/2/peer