AMP®-Parkinson's Disease Progression Prediction

미니 프로젝트 : Python 을 활용한 통계 분석 및 웹 서비스 구현

고용노동부
KOREATECH 직업능력심사평가원

AMP®-Parkinson's Disease Progression Prediction

MDS-UPDRS를 이용한 파킨슨 질병 진행 예측

프로젝트 팀 구성

팀 구성 및 역할
개발 환경
프로젝트 기간

프로젝트 개요

대회 목적
연구배경
평가 지표
순서도

프로젝트 수행 절차

테이블 정의서
데이터 전처리
피처 엔지니어링

수행 결과 I

ERD
기초통계분석
상관 분석
평가 지표

머신 러닝
데이터 분리
모델 생성 및 학습
모델 평가 및 성능 검증

수행 결과 II

대시 보드

Home / Description / Data

EDA / STAT / ML

자체 평가 의견

분석 결과 요약
한계점 및 개선사항
출처

Index

01

프로젝트 팀 구성

팀 구성 및 역할
개발 환경
프로젝트 기간

01 프로젝트 개요

03 수행 절차 및 방법

04 수행 결과 Ⅰ

05 수행 결과 Ⅱ

06 평가 의견

팀 구성 및 역할

성열민

통계 검증

머신러닝

이한영

데이터 전처리

머신러닝

이태선

대시보드 제작

PPT제작

김동욱

시각화

PPT제작

임광현

시각화

머신러닝

프로젝트 개발환경

개발환경

Google Colaboratory

kaggle

배포

Streamlit

GitHub

언어

python

라이브러리

- matplotlib (3.7.1)
- numpy (1.24.3)
- pandas (2.0.1)
- pingouin (0.5.3)
- plotly (5.14.1)
- plotly-express (0.4.1)
- scikit-learn (1.2.2)
- seaborn (0.12.2)
- statsmodels (0.13.5)
- streamlit (1.22.0)
- streamlit-option-menu (0.3.2)
- utils (1.0.1)
- xgboost (1.7.5)

프로젝트 기간

기간 : 4/24 ~ 5/17

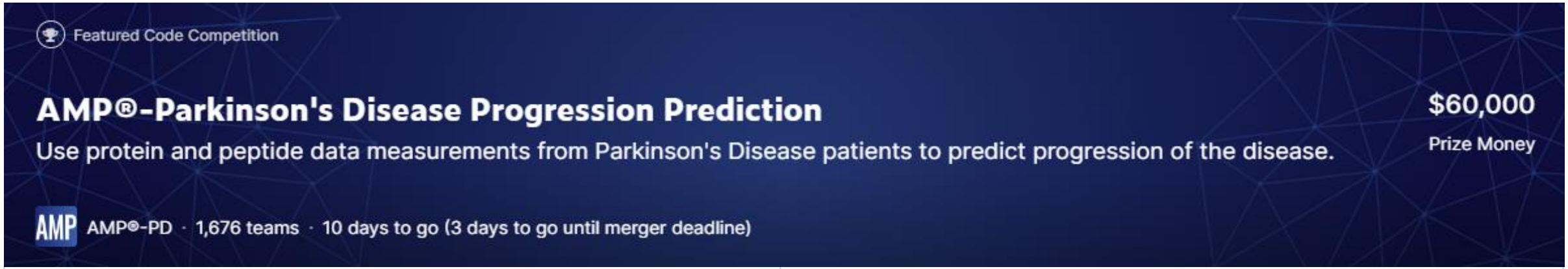
SUN	MON	TUE	WED	THU	FRI	SAT
23	24 프로젝트 시작	25	26	27	28	29
30	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17 프로젝트 마감	18	19	20

02

프로젝트 개요

대회 목적
연구배경
평가 지표
순서도

프로젝트 개요



대회 목표

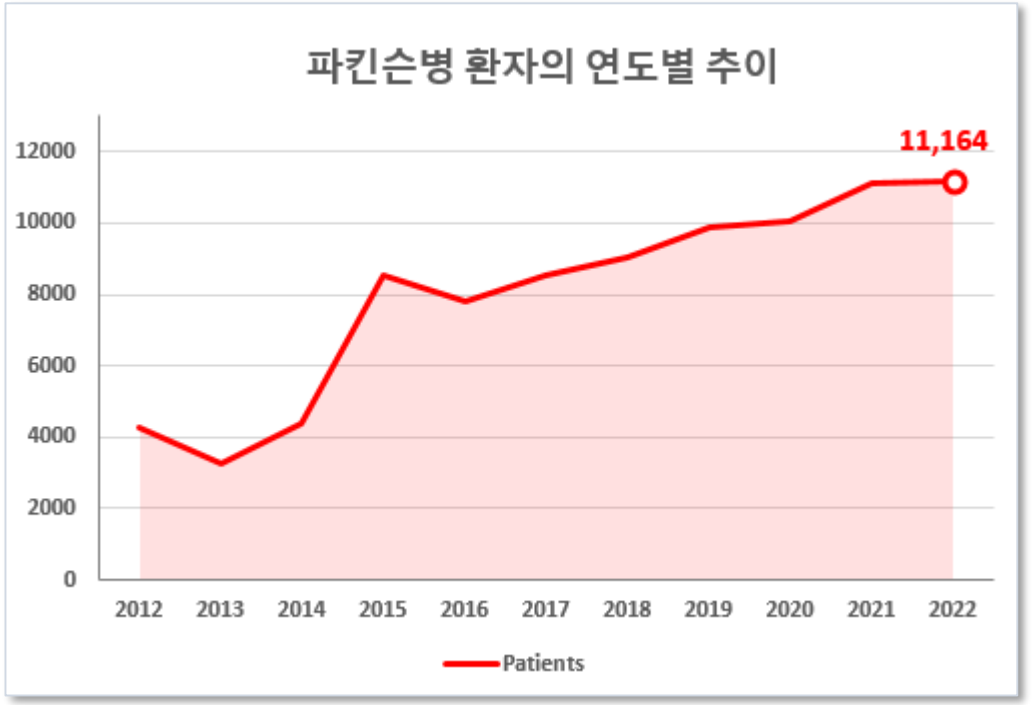
이 대회의 목적은 파킨슨병 환자의
임상 데이터를 사용하여,
파킨슨병 환자의 진행을 측정하는
MDS-UPDRS (통합 파킨슨병 평가척도) 예측

MDS-UPDRS란?

MDS-UPDRS		점수 범위
Part 1	일상생활에서 비 운동성 상태	0 ~ 52
Part 2	일상생활에서 **운동성 증상** 상태	0 ~ 52
Part 3	클리닉에서 검사하는 운동성 증상 상태	0 ~ 132
Part 4	운동성 합병증	0 ~ 24

▶ 파킨슨병의 진행 정도를 평가할 수 있는 객관적인 지표

프로젝트 개요



데이터 출처 : 서울대학교 병원 파킨슨 센터

파킨슨병은 흑질, 도파민 신경세포의 소실 로
운동, 인지, 수면 및 기타 정상인 기능에 영향을 끼치는
치료법이 없는 뇌 질환

파킨슨병 환자는 매년 증가 하는 추세
(**2022** 년 기준 **11,164** 명)

※ **2037**년까지 미국에서는 **160**만 명 이 파킨슨병에 걸릴 것으로 추정,
이로 인한 경제적 비용 은 **800**억 달러 로 예상

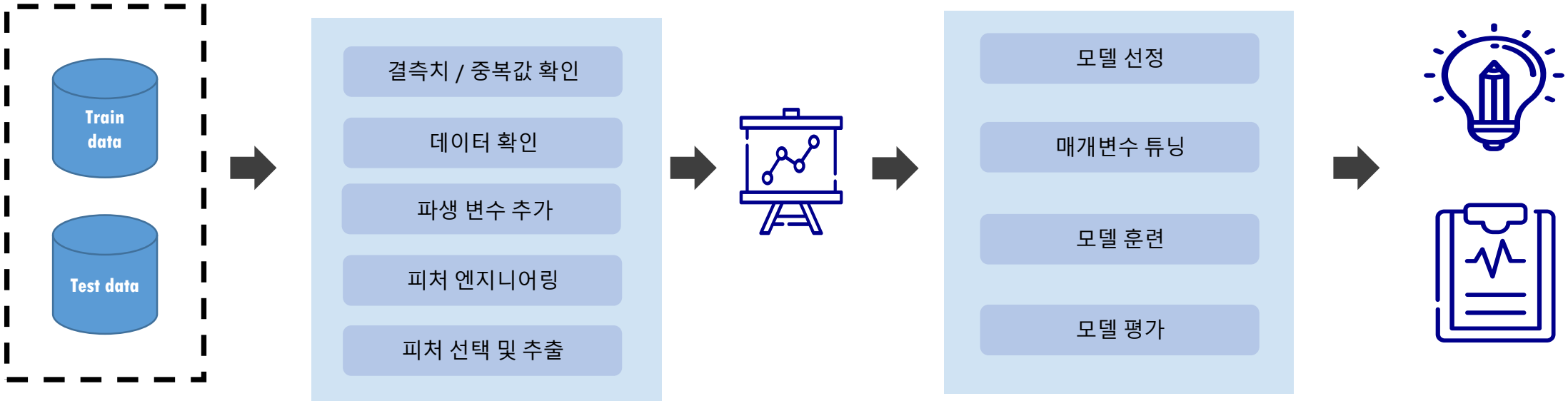
프로젝트 개요

평가 지표	수식	설명	값 해석
MAE **(Mean Absolute Error)**	$MAE = \frac{\sum \lvert y - \hat{y} \rvert}{n}$	평균 절대 오차, 실제 값과 예측 값의 절대값 평균	값이 작을 수록 모델의 성능이 좋음
MSE **(Mean Squared Error)**	$MSE = \frac{\sum_{i=1}^{n}(y - \hat{y})^2}{n}$	상관 계수의 제곱 한 값 변수간 인과 관계를 설명	값이 작을 수록 모델의 성능이 좋음
R2 score **(R-squared)**	$R^2 = 1 - \frac{SSE}{SST}$	결정계수, 예측값의 분산 / 실제 값의 분산	**0 ~ 1** 의 범위, **1**에 가까울 수록 선형 모델이 해당 변수에 대한 높은 연관성 가짐
SMAPE **(Symmetric mean absolute percentage error)**	$SMAPE = \frac{100}{n} \times \sum_{i=0}^{n} \frac{\lvert Y_i - \hat{Y}_i \rvert}{(\lvert Y_i \rvert + \lvert \hat{Y}_i \rvert)/2}$	평균 절대 백분율 오차, 실제값과 예측값의 비율 차이의 절대값 평균	**0~100** 또는 **0~200**의 범위, 값이 작을 수록 모델의 성능이 좋음

SMAPE 사용 하여 실제 값과 모델이 예측한 값의 오차**(Error)** 를 확인 ▶ 모델의 성능 을 검증함

프로젝트 개요

데이터 → 데이터 전처리 → **EDA** 수행 데이터 시각화 → 모델 구축 → 예측 결과 도출 / 대시보드 구현



Index

03

프로젝트 수행 절차

- ERD
- 테이블 정의서
- 데이터 전처리
- 피처엔지니어링

01 프로젝트 개요

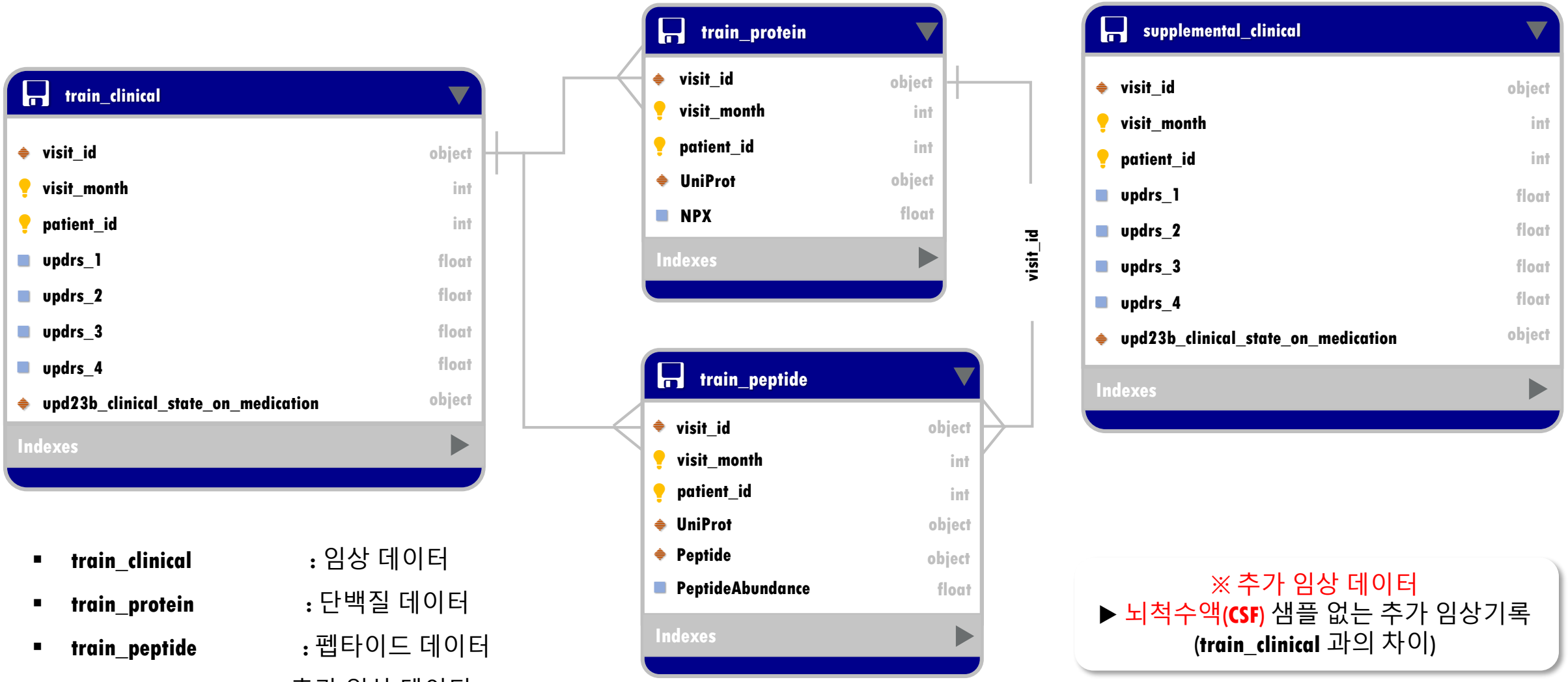
02 팀원 구성 및 역할

04 수행 결과 Ⅰ

05 수행 결과 Ⅱ

06 평가 의견

프로젝트 수행 절차



- **train_clinical** : 임상 데이터
- **train_protein** : 단백질 데이터
- **train_peptide** : 펩타이드 데이터
- **supplemental_clinical** : 추가 임상 데이터

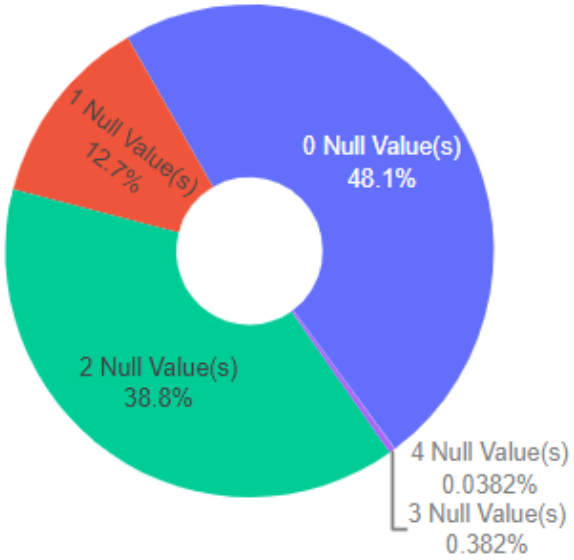
※ 추가 임상 데이터
▶ 뇌척수액(**CSF**) 샘플 없는 추가 임상기록
(**train_clinical** 과의 차이)

테이블 명	컬럼명	자료형	NULL	KEY	비고	
train_clinical / supplemental_ clinical	visit_id	object	비허용	X	방문자 ID 코드	
	visit_month	int	비허용	X	방문 월 (환자 첫 방문 기준)	
	patient_id	int	비허용	X	환자 ID 코드	
	updrs_1	float	비허용	X	통합 파킨슨병 평가 척도	기분 및 행동
	updrs_2	float	비허용	X		일상 생활의 운동 경험
	updrs_3	float	비허용	X		운동 평가
	updrs_4	float	비허용	X		운동 합병증
	upd23b_clinical_state_ on_medication	object	비허용	X	환자의 약물 복용 여부 (점수 평가기간 동안)	

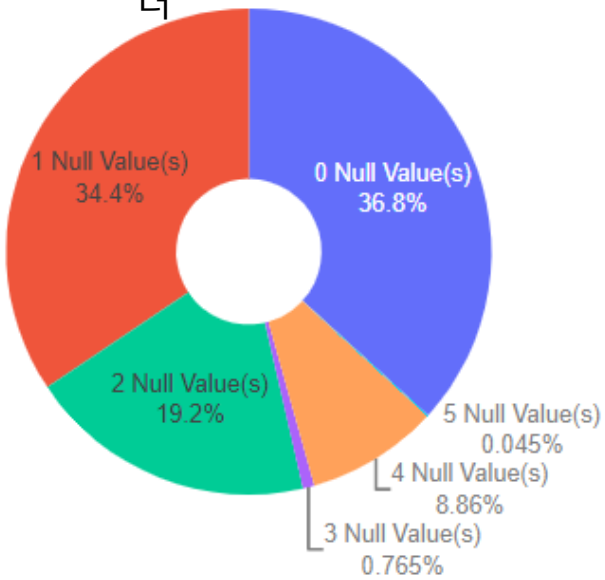
테이블 명	컬럼명	자료형	NULL	KEY	비고
train_proteins	visit_id	object	비허용	X	방문자 ID 코드
	visit_month	int	비허용	X	방문 월 (환자 첫 방문 기준)
	patient_id	int	비허용	X	환자 ID 코드
	UniProt	object	비허용	X	단백질 UniProt ID 코드
	NPX	float	비허용	X	정규화된 단백질 발생 빈도
train_peptides	visit_id	object	비허용	X	방문자 ID 코드
	visit_month	int	비허용	X	방문 월 (환자 첫 방문 기준)
	patient_id	int	비허용	X	환자 ID 코드
	UniProt	object	비허용	X	단백질 UniProt ID 코드
	peptide	object	비허용	X	peptide에 포함된 아미노산 서열
	PeptideAbundance	float	비허용	X	아미노산 빈도

결측치 확인

임상 데이터



추가 임상 데이터



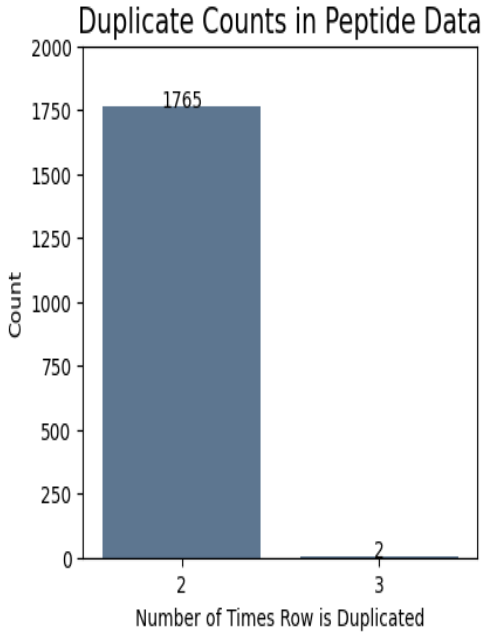
단백질 데이터

펩타이드 데이터

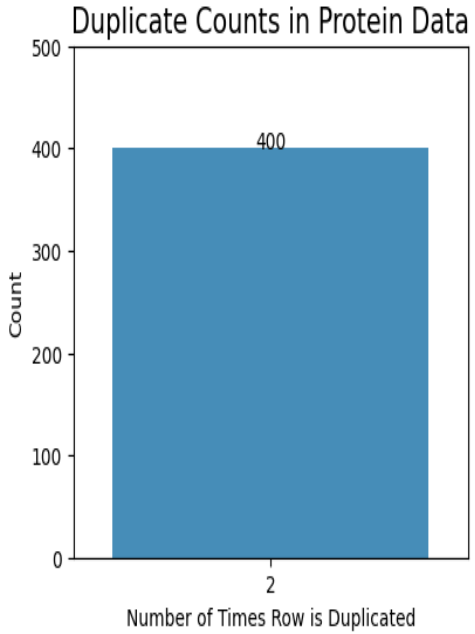
임상데이터 , 추가 임상 데이터의 결측치 확인

단백질 데이터, 펩타이드 데이터에는 결측치 없음

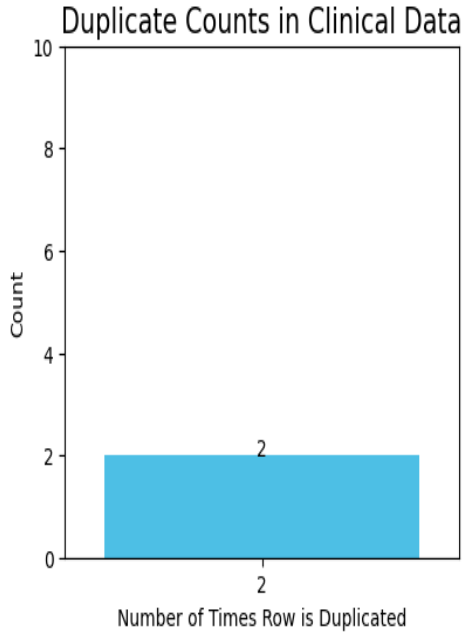
중복값 확인



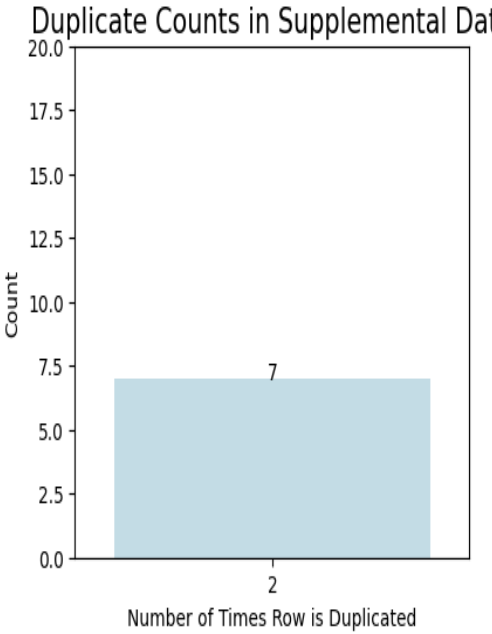
0.15%



0.63%



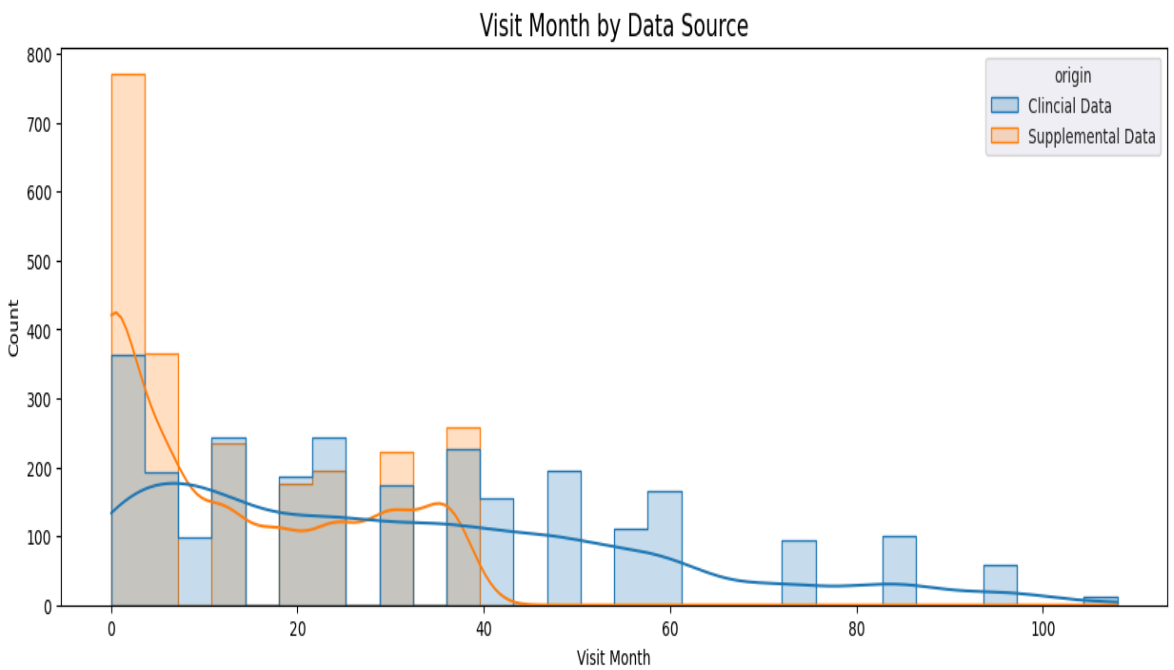
0.35%

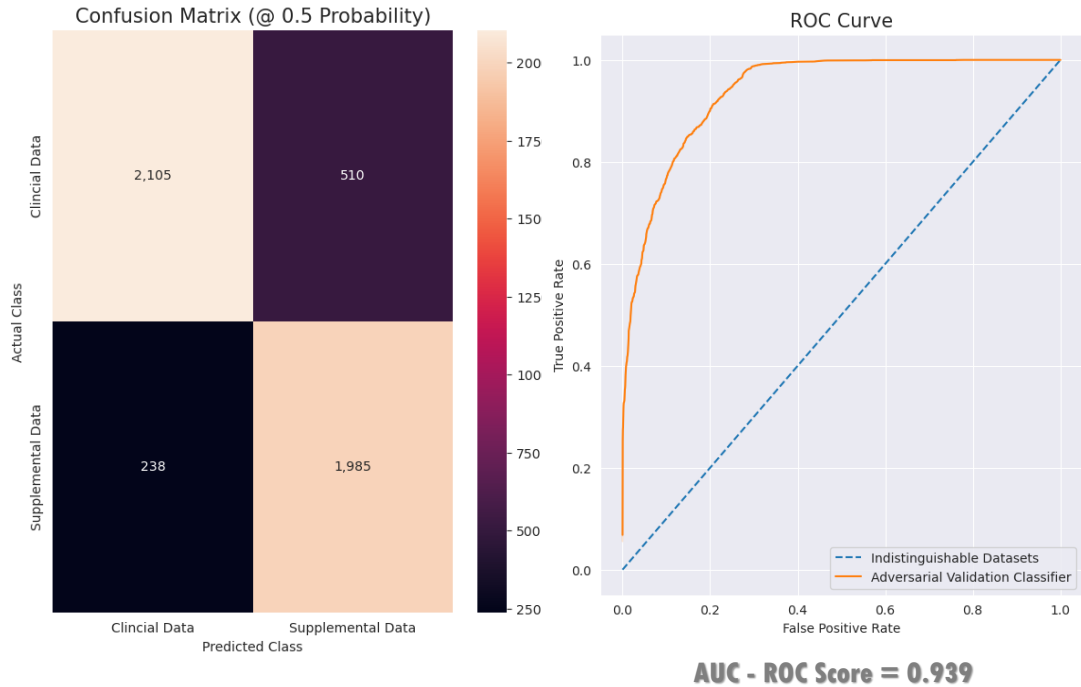


0.36%

각 데이터의 중복값이 거의 영향을 미치지 않을 것으로 판단됨

데이터 확인



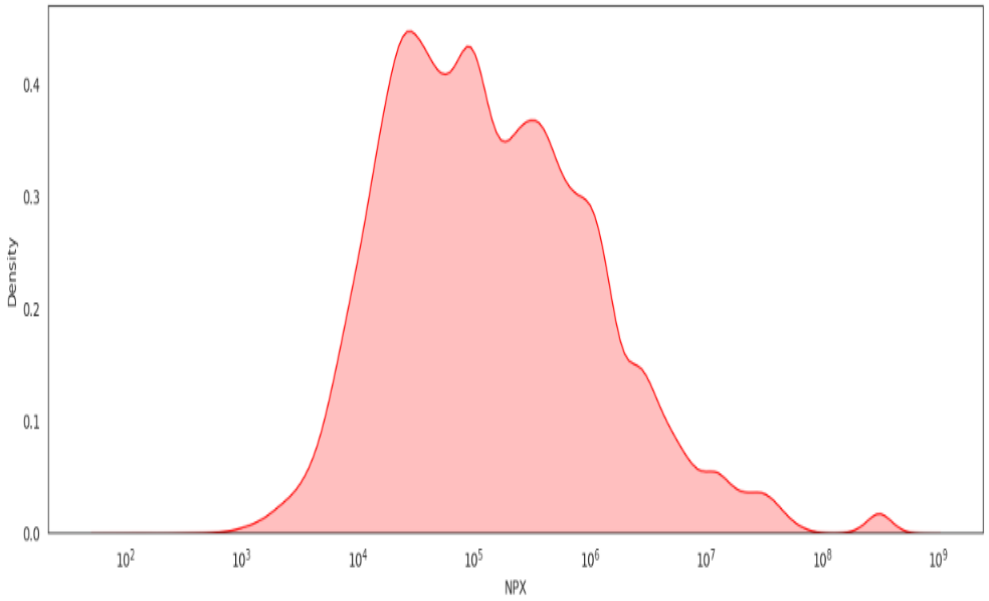


임상 데이터에서는 **0~108**개월 사이 방문
추가 임상 데이터에서는 **0~36**개월 사이 방문
추가 임상 데이터에서 **0**개월 데이터가 가장 많음

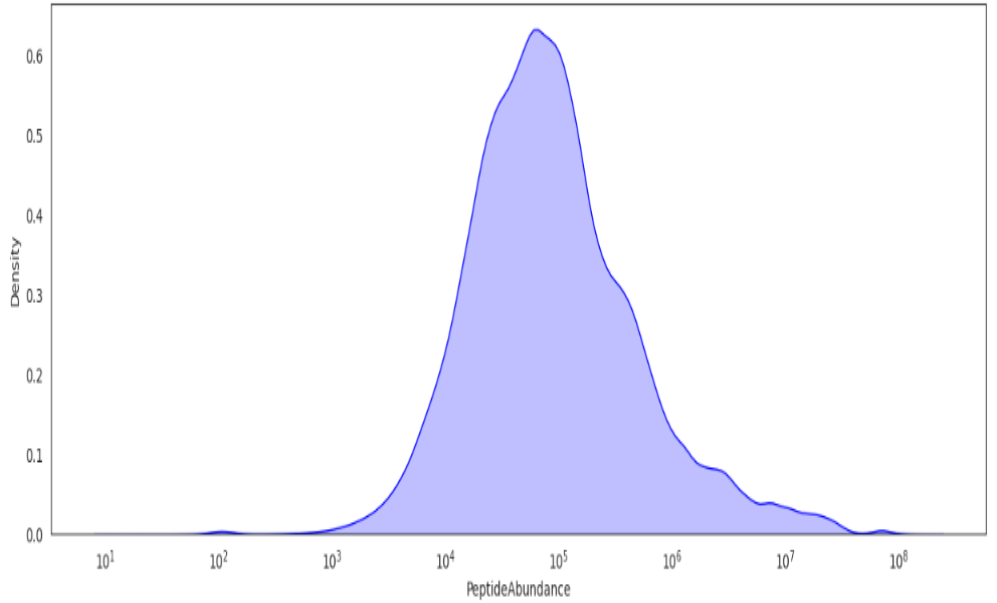
임상 데이터와 추가 임상 데이터를 쉽게 구분 가능
⇨ 두 데이터 세트의 특징이 다르므로
합쳐서 사용할 때 주의해야함

데이터 확인

protein_data

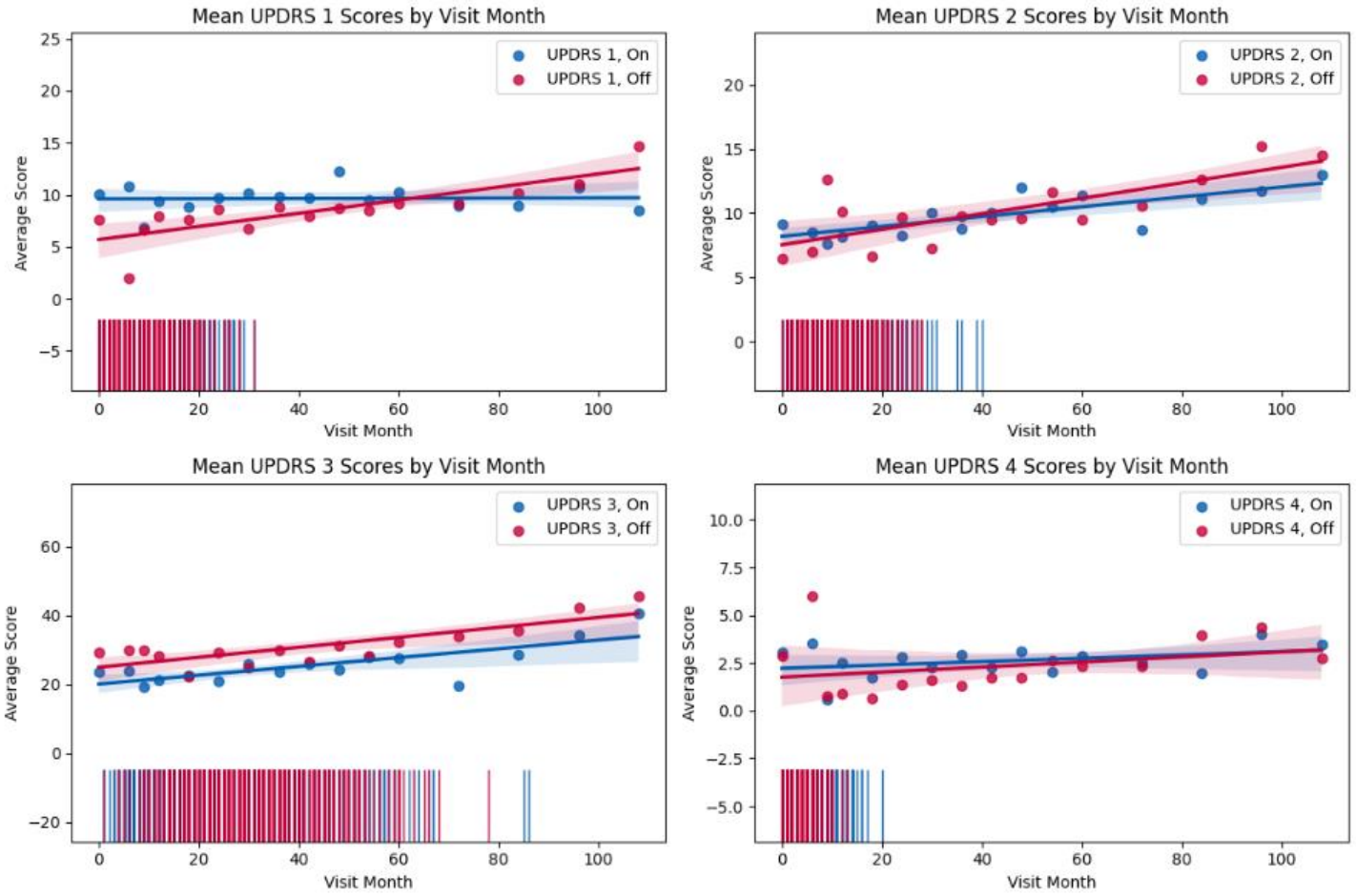


peptide_data



단백질 데이터의 **NPX,** 펩타이드 데이터의 **PeptideAbundance log** 변환 값
⇨ **NPX, PeptideAbundance** 값의 분포도가 광범위

약물 복용 여부에 따른 UPDRS 점수

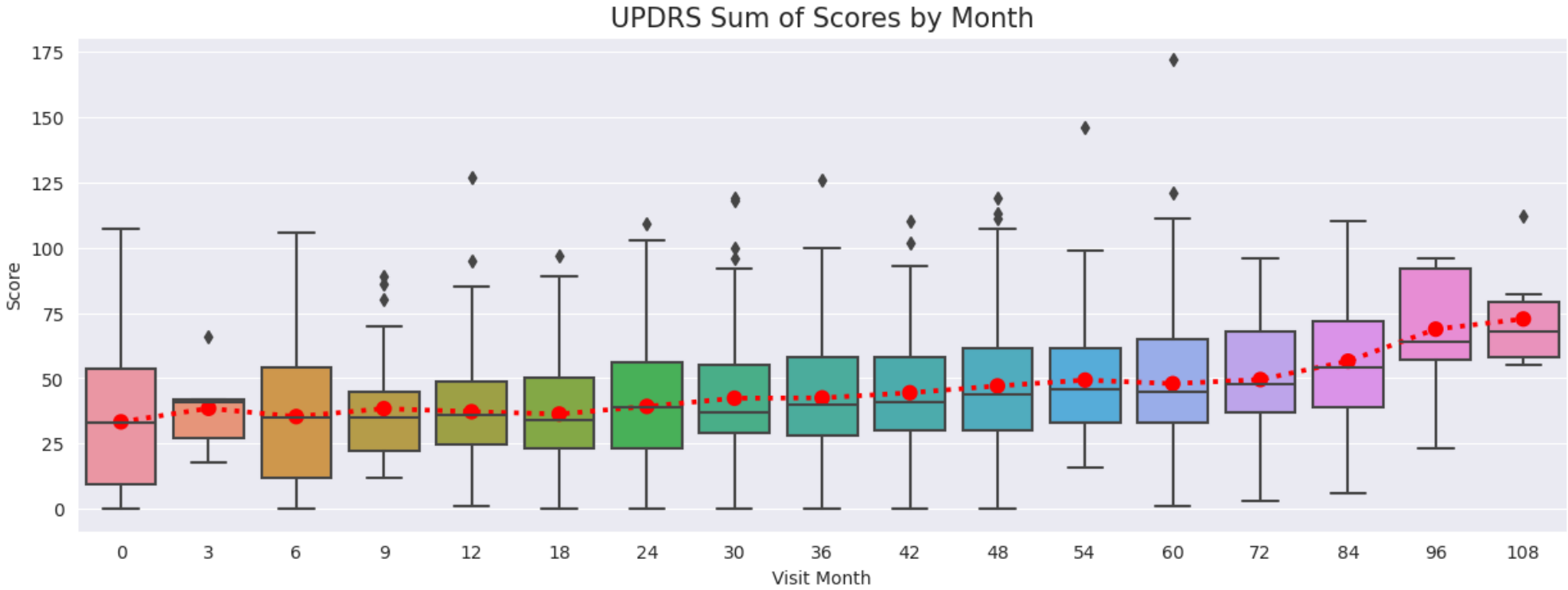


약물 복용 OFF일 경우
UPDRS_1,2,3에서 점차적으로 증가

약물 복용 ON일 경우
OFF 에 비해 비교적 평탄함 유지

※ 기초 통계 분석 페이지에서 다룰 예정

약물 복용 여부에 따른 UPDRS 점수



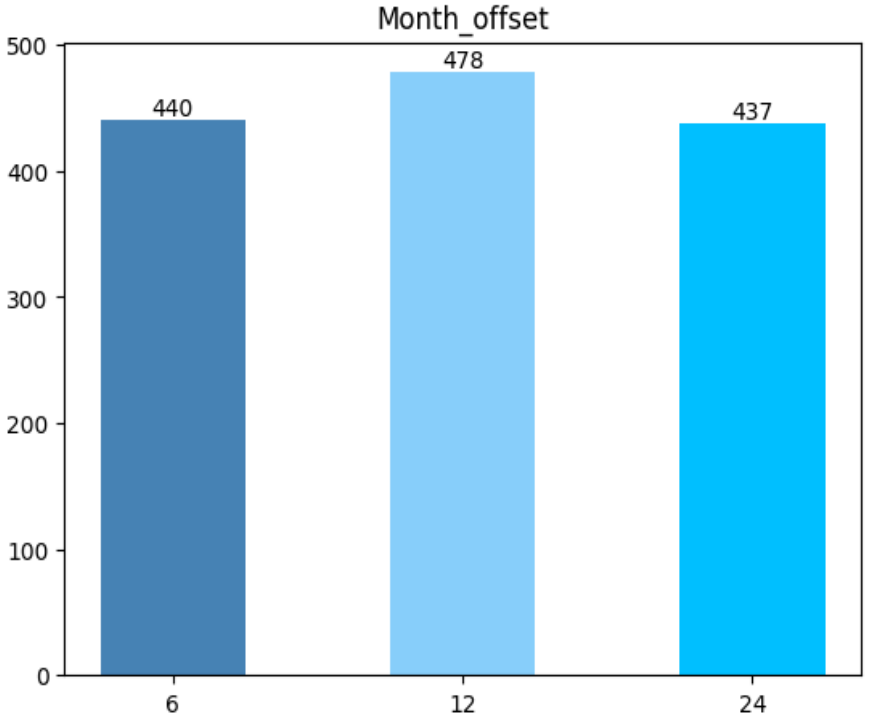
약물복용 **ON** + 약물복용 **OFF** -> 시간에 따라 **updrs** 점수 점진적으로 증가

결측치 처리

updrs_4, medication의 결측치 많음.

⇨ **updrs_4**의 값을 **0**으로 설정

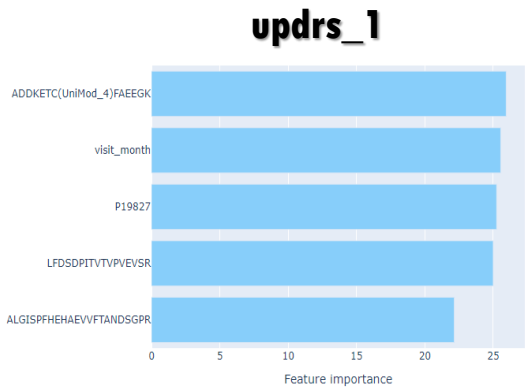
파생 피처 생성

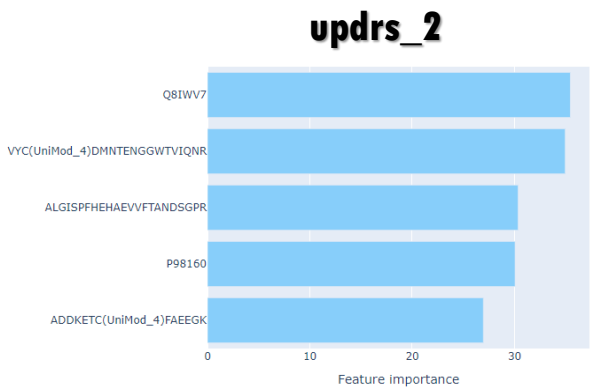


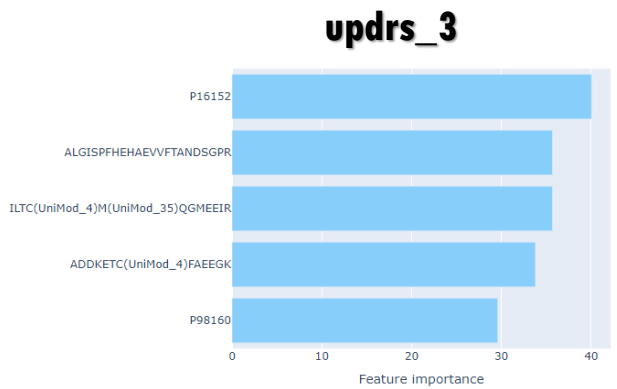
visit_month를 이용한 파생 변수 추가

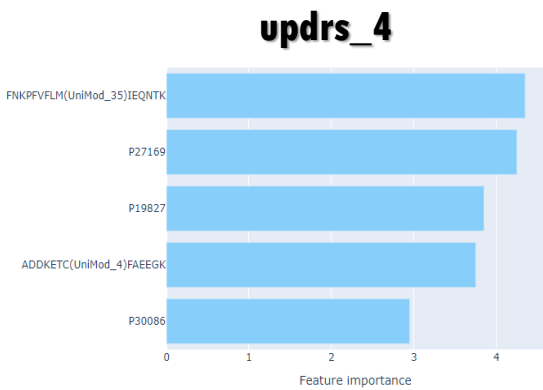
⇨ **test data**에서 **6, 12, 24**개월만 사용

피처 중요도 확인









	updrs_1	updrs_2	updrs_3	updrs_4
visit_month	25.550000	39.950000	42.750000	9.350000
SC(UniMod_4)SPELQQK	33.750000	41.150000	45.850000	5.550000
LPPTSAHGNVAEGETKPDPDVTER	33.250000	42.650000	50.400000	7.250000
ADDKETC(UniMod_4)FAEEGK	25.950000	26.950000	33.800000	3.750000
TTPPVLDSDGSFFLYSK	32.150000	38.850000	45.600000	5.250000

04

수행 결과 Ⅰ

- 기초통계분석
- 상관 분석
- 평가 지표

- 머신 러닝
- 데이터 분리
- 모델 생성 및 학습
- 모델 평가 및 검증

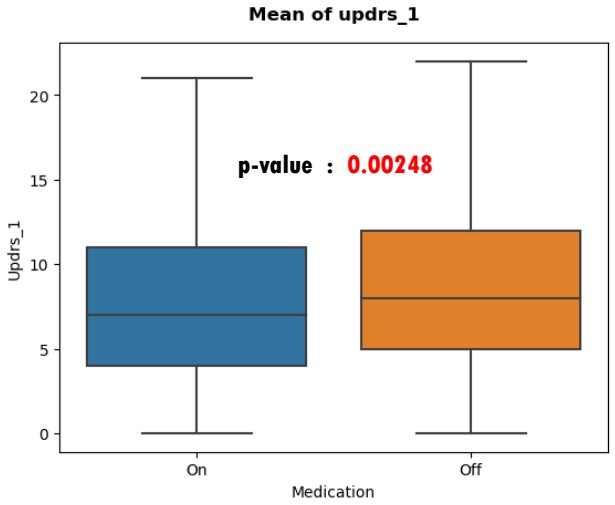
Medication On / Off - updrs 점수 분석

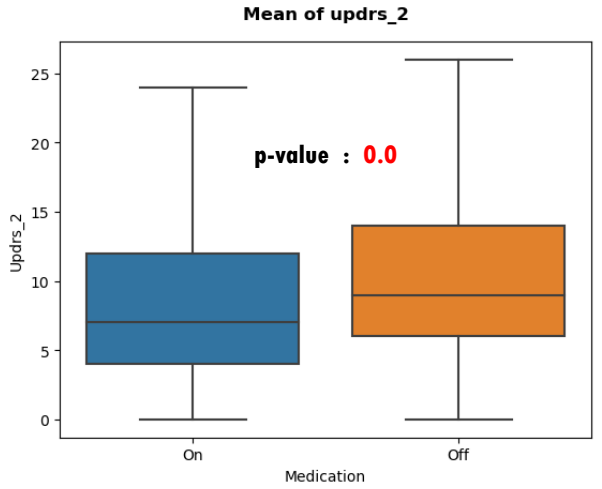
두 집단간 평균 차이 검정 (Two-sample t-test)

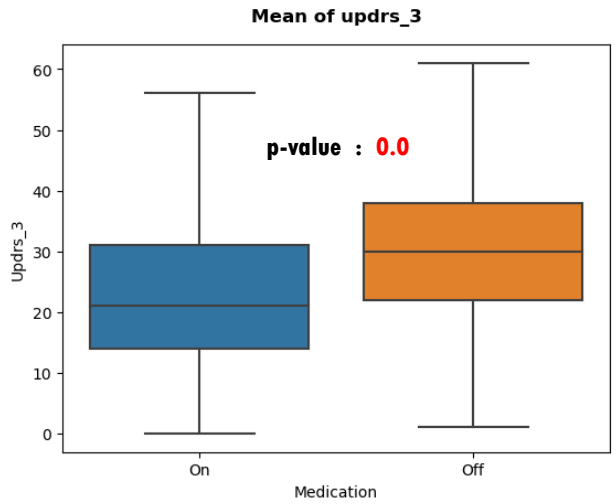
$$t = (\overline{X_1} - \overline{X_2}) \,/\, s_p(\sqrt{1/n_1 + 1/n_2})$$

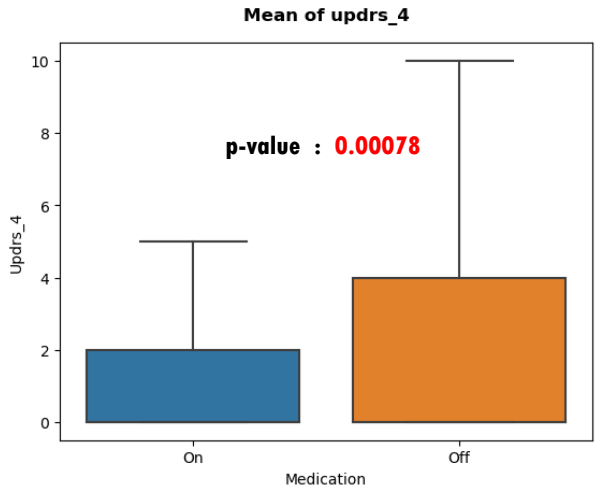
귀무가설(H0) : 두 모집단 (On, Off) 평균이 같다

대립가설(H1) : 두 모집단 (On, Off) 평균이 같지 않다







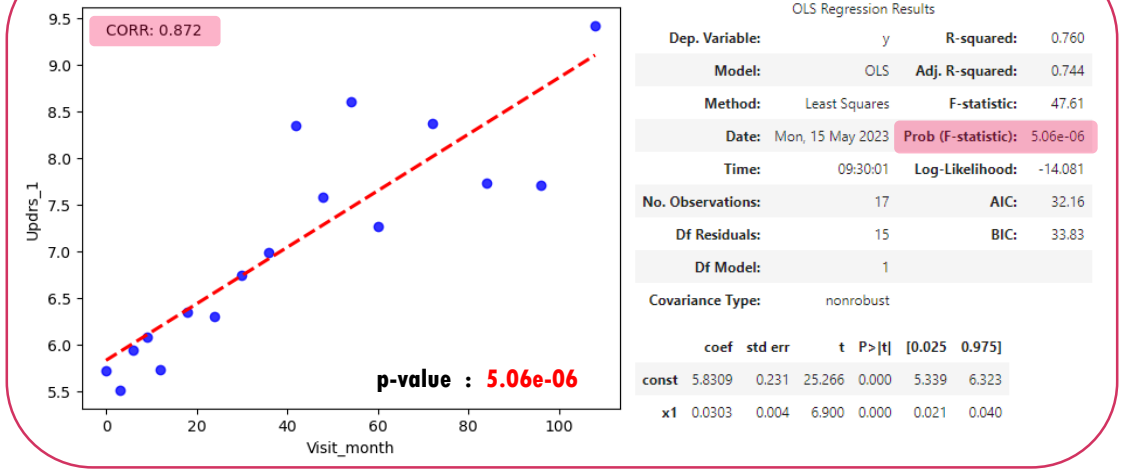


▣ t-test 를 통해 나온 **p-value < 0.05**

∴ 귀무가설(H0) 기각 / 대립가설(H1) 채택,

On, Off 의 평균이 다르다고 판단 할 수 있음

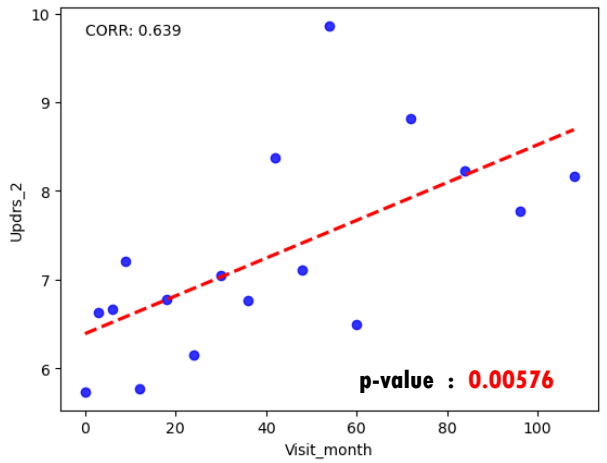
visit_month - updrs 점수 분석



OLS Regression Results

Dep. Variable:	y	R-squared:	0.760
Model:	OLS	Adj. R-squared:	0.744
Method:	Least Squares	F-statistic:	47.61
Date:	Mon, 15 May 2023	Prob (F-statistic):	5.06e-06
Time:	09:30:01	Log-Likelihood:	-14.081
No. Observations:	17	AIC:	32.16
Df Residuals:	15	BIC:	33.83
Df Model:	1		
Covariance Type:	nonrobust		

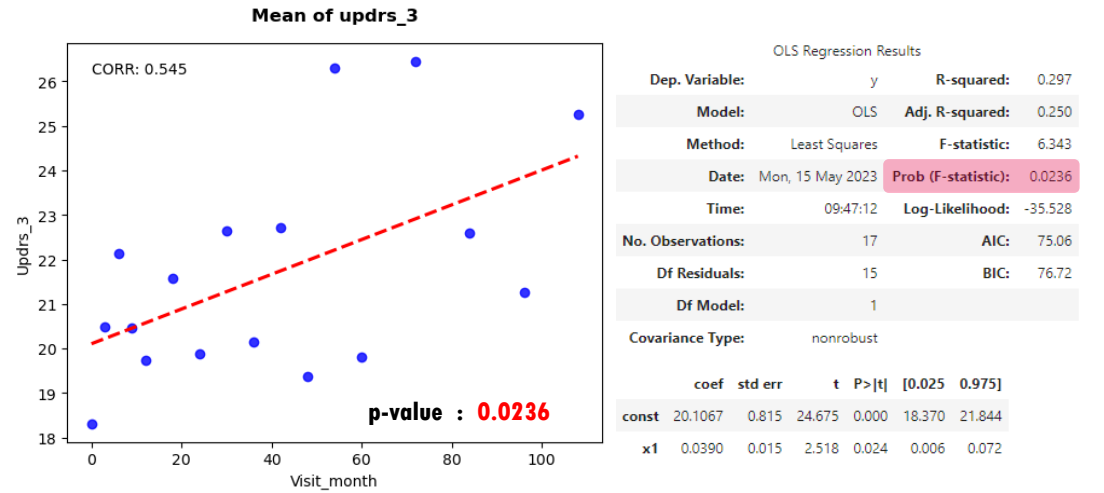
	coef	std err	t	P>|t|	[0.025	0.975]
const	5.8309	0.231	25.266	0.000	5.339	6.323
x1	0.0303	0.004	6.900	0.000	0.021	0.040



OLS Regression Results

Dep. Variable:	y	R-squared:	0.408
Model:	OLS	Adj. R-squared:	0.369
Method:	Least Squares	F-statistic:	10.35
Date:	Mon, 15 May 2023	Prob (F-statistic):	0.00576
Time:	09:46:15	Log-Likelihood:	-21.095
No. Observations:	17	AIC:	46.19
Df Residuals:	15	BIC:	47.86
Df Model:	1		
Covariance Type:	nonrobust		

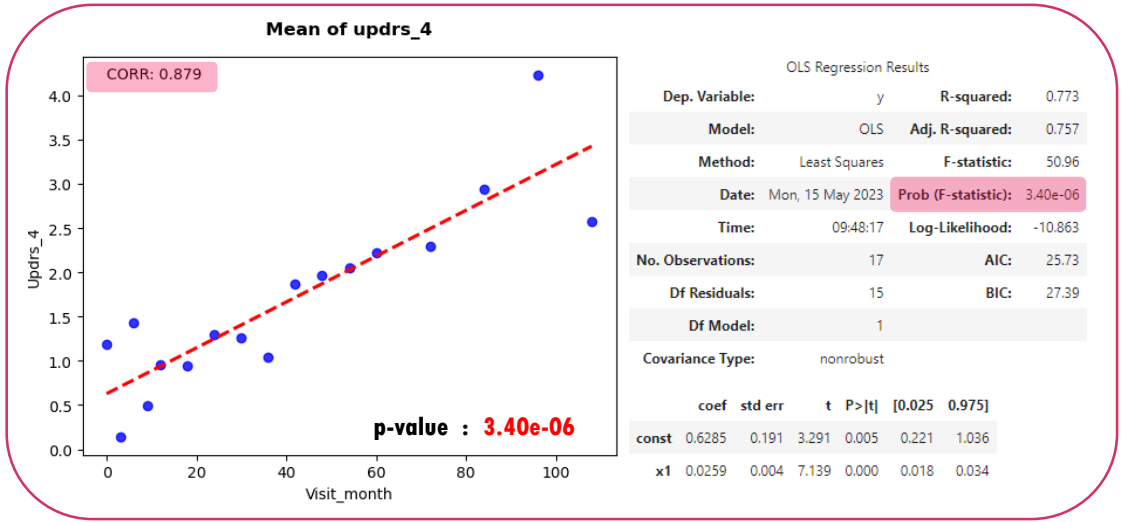
	coef	std err	t	P>|t|	[0.025	0.975]
const	6.3896	0.349	18.328	0.000	5.647	7.133
x1	0.0213	0.007	3.217	0.006	0.007	0.035



OLS Regression Results

Dep. Variable:	y	R-squared:	0.297
Model:	OLS	Adj. R-squared:	0.250
Method:	Least Squares	F-statistic:	6.343
Date:	Mon, 15 May 2023	Prob (F-statistic):	0.0236
Time:	09:47:12	Log-Likelihood:	-35.528
No. Observations:	17	AIC:	75.06
Df Residuals:	15	BIC:	76.72
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	20.1067	0.815	24.675	0.000	18.370	21.844
x1	0.0390	0.015	2.518	0.024	0.006	0.072



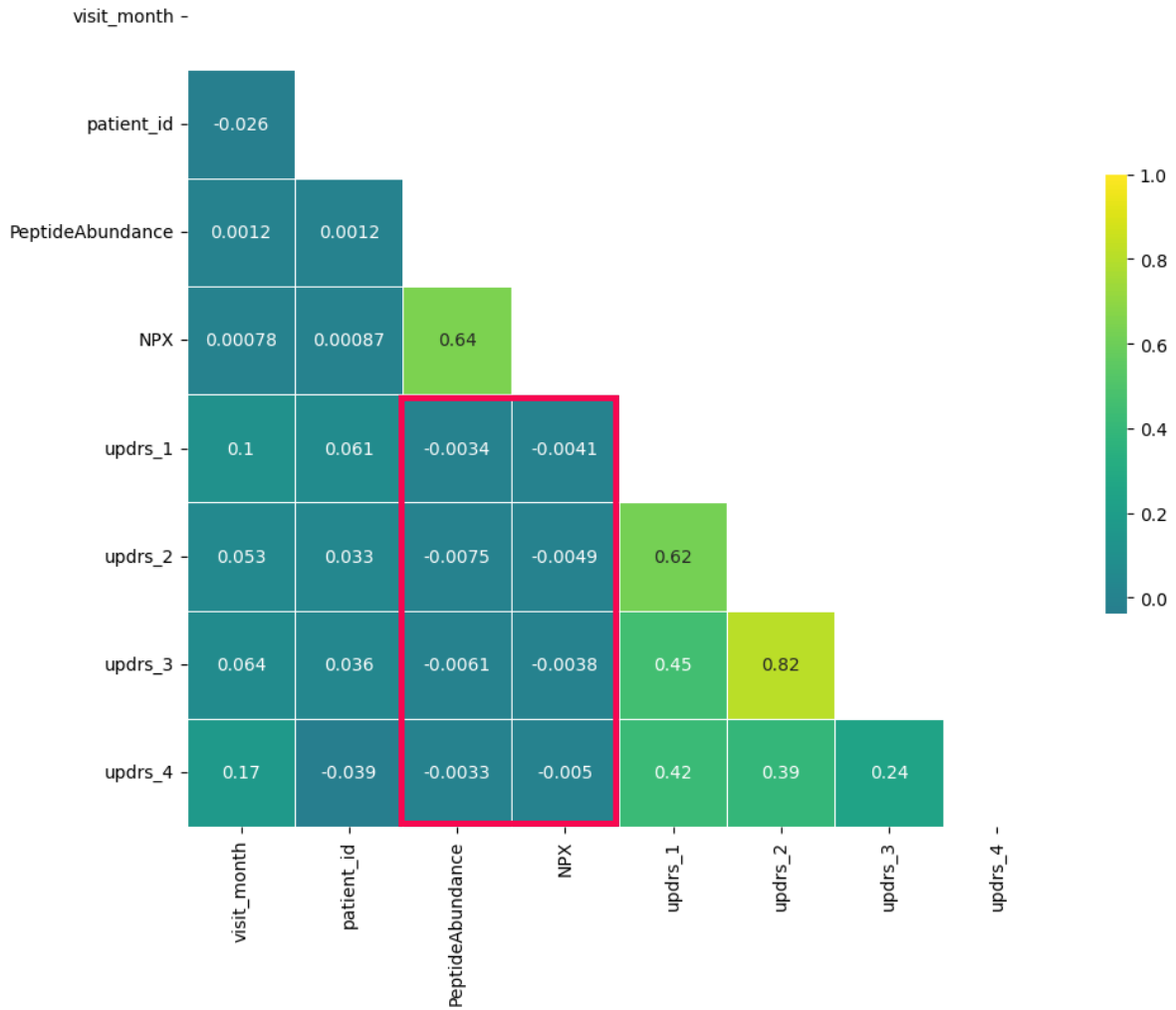
OLS Regression Results

Dep. Variable:	y	R-squared:	0.773
Model:	OLS	Adj. R-squared:	0.757
Method:	Least Squares	F-statistic:	50.96
Date:	Mon, 15 May 2023	Prob (F-statistic):	3.40e-06
Time:	09:48:17	Log-Likelihood:	-10.863
No. Observations:	17	AIC:	25.73
Df Residuals:	15	BIC:	27.39
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6285	0.191	3.291	0.005	0.221	1.036
x1	0.0259	0.004	7.139	0.000	0.018	0.034

수행 결과 I

전체 피처 별 상관관계



① **visit_month – updrs**

▶ **updrs_1, 4 :** 약한 양의 상관관계 (**0.1, 0.17**)

② **updrs** 간 상관관계

▶ **updrs_1 - 2 :** 양의 상관관계 (**0.62**)

▶ **updrs_2 - 3 :** 양의 상관관계 (**0.82**)

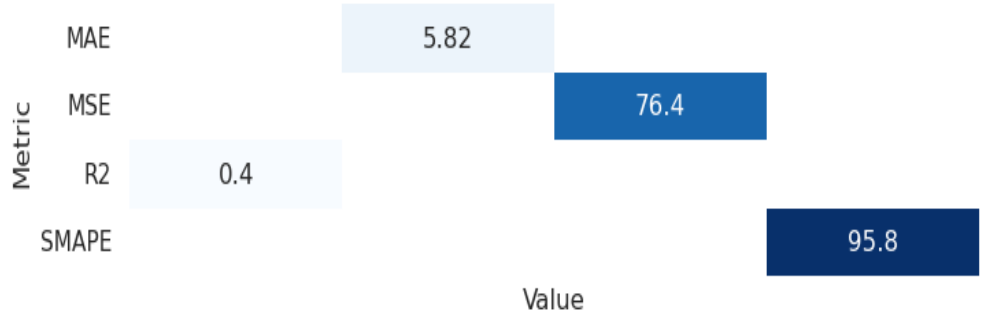
③ **NPX – PeptideAbundance**

▶ 양의 상관관계 (**0.64**)

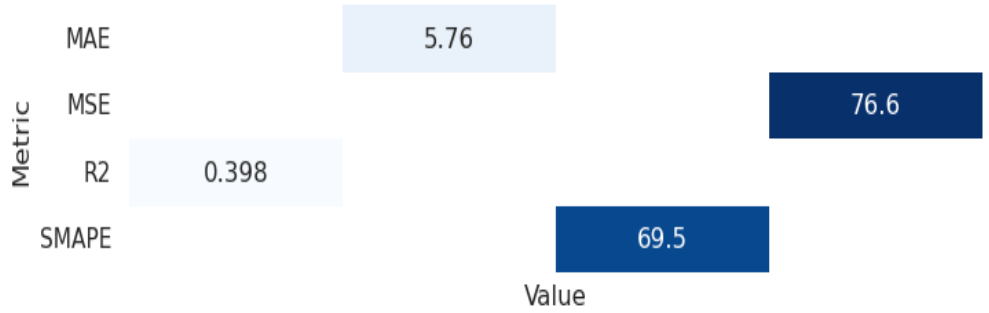
④ **updrs – NPX & PeptideAbundance**

▶ 상관관계가 거의 없다고 판단

평가지표 별 점수 비교



updrs_4 != 0으로 설정

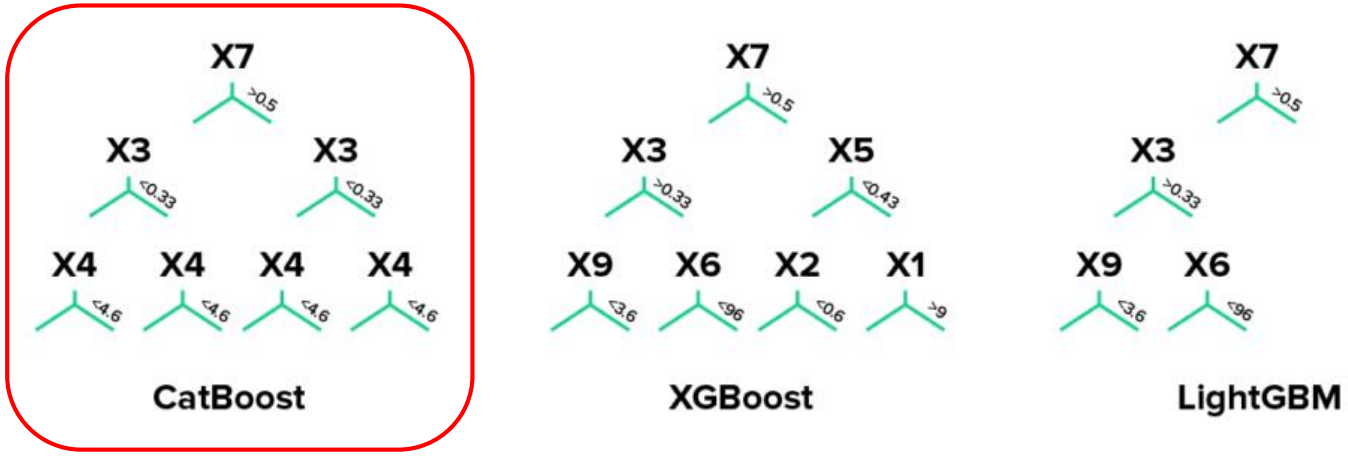


updrs_4 == 0으로 설정

평가지표 **MAE, MSE, R2**는
UPDRS_4의 값에 크게 영향을 받지 않음.

평가지표 **SMAPE**는
UPDRS_4의 값에 크게 영향을 받음.

모델별 트리 성장 전략 비교



📌 XGBoost

레벨 별 나무 성장 전략을 사용하여 나무를 한 번에 한 레벨 씩 성장시키는 기법

▶ 균형 잡히고 대칭적인 트리를 만드는 것이 목표

📌 LightGBM

잎 단위 트리 성장 전략을 사용하여 가장 높은 이득을 가진 잎 노드를 확장하는 기법

▶ 더 복잡하고 비대칭적인 트리를 생성 (데이터가 작을 때 효과적)

CatBoost

1. 균형 잡힌 트리 구조여서 오버피팅을 방지

▶ **Light GBM** : **leaf-wise**

▶ **XGboost** : 대칭 트리

2. Ordered Boosting

▶ 기존의 부스팅 모델

: 모든 훈련 데이터를 대상으로 잔차 계산

▶ **Catboost** 는 일부만 가지고 잔차를 계산

이 값으로 모델을 만들고,

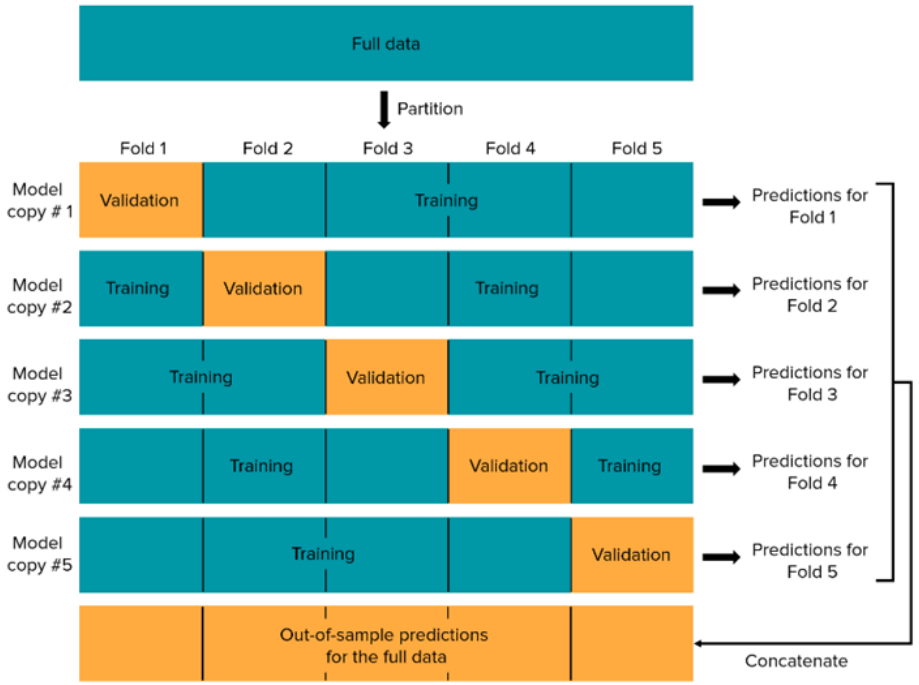
데이터의 잔차는 이 모델로 예측한 값을 사용

3. Hyper Parameter

▶ 파라미터가 최적화가 잘 되어있음

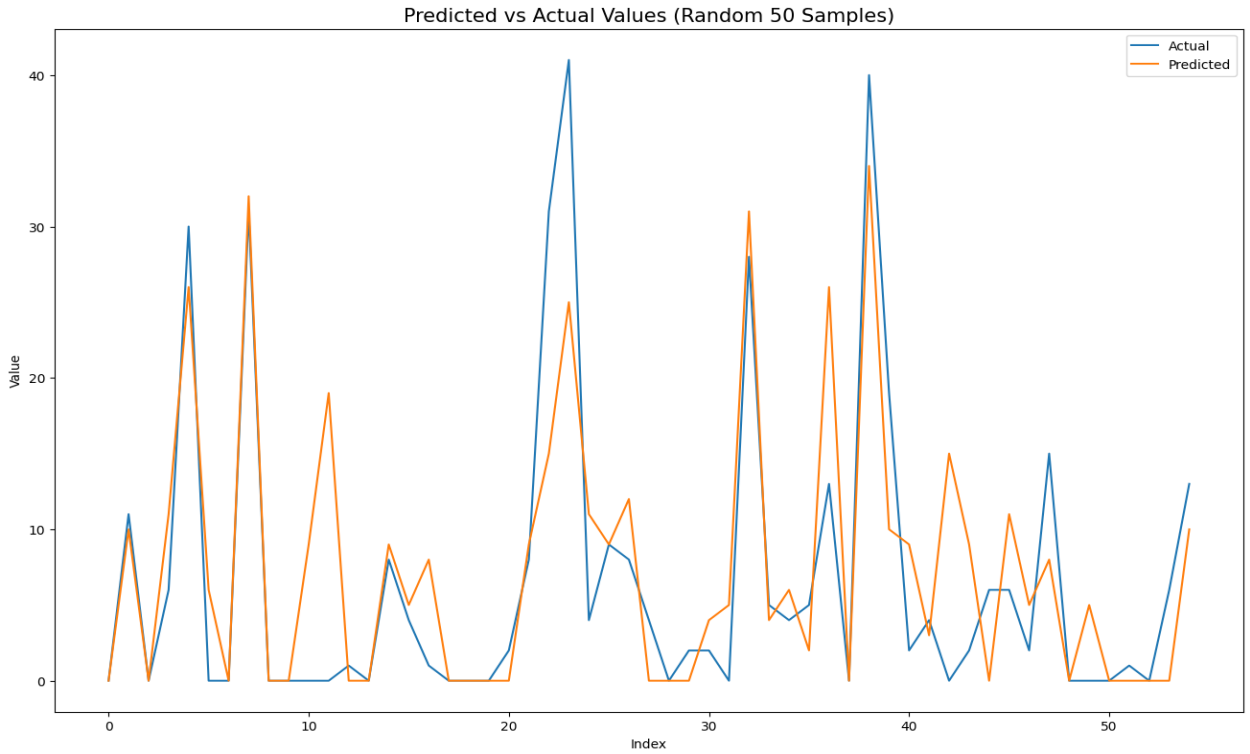
▶ 파라미터 튜닝에 크게 신경쓰지 않아도 됨

훈련 / 검증 데이터 분리



random_state : 시드값 고정
k-fold (5) : 교차검증

실제 값 / 예측 값 확인

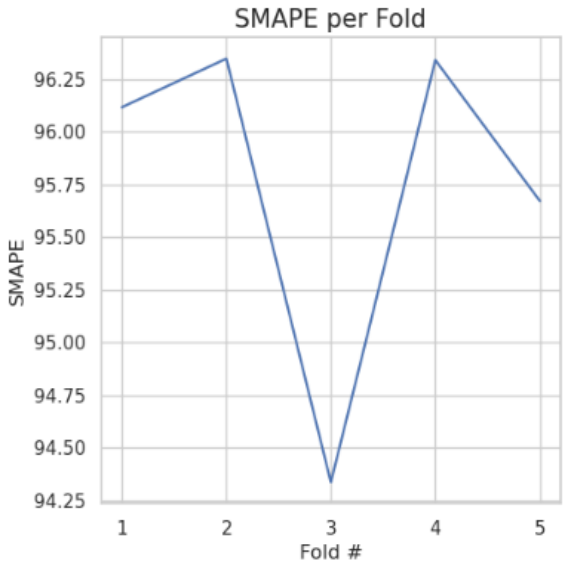


예측 값 과 실제 값을 **50**개씩 랜덤 추출
전반적으로 실제 값 과 근접한 수준으로 나타남

CatBoost 모델 생성 및 학습

모델 ①

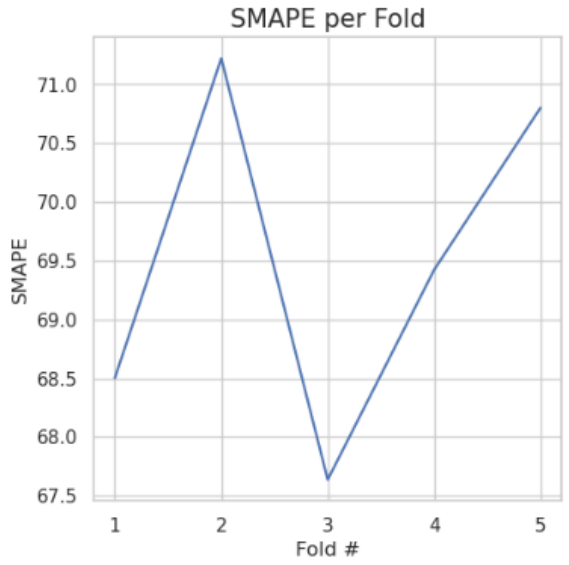
임상 데이터
피처 : **visit_month, month_offset**
조건: **UPDRS_4 != 0**



SMAPE = 95.76

모델 ②

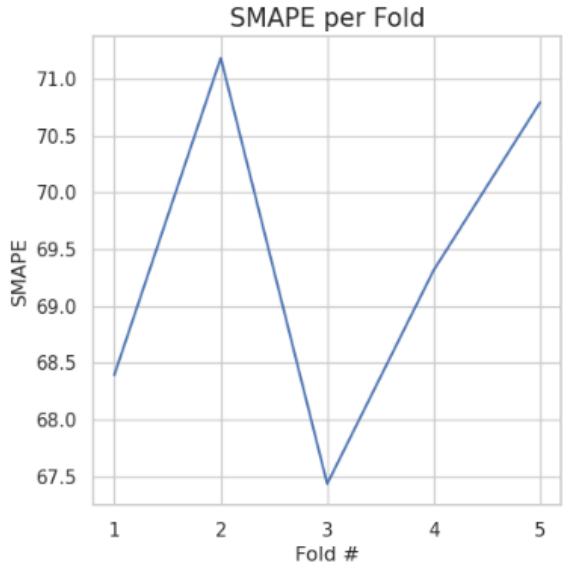
임상 데이터
피처 : **visit_month, month_offset**
조건: **UPDRS_4 == 0**



SMAPE = 69.52

모델 ③

임상 데이터 + 추가 임상 데이터
피처 : **visit_month, month_offset**
조건: **UPDRS_4 == 0**



SMAPE = 69.42

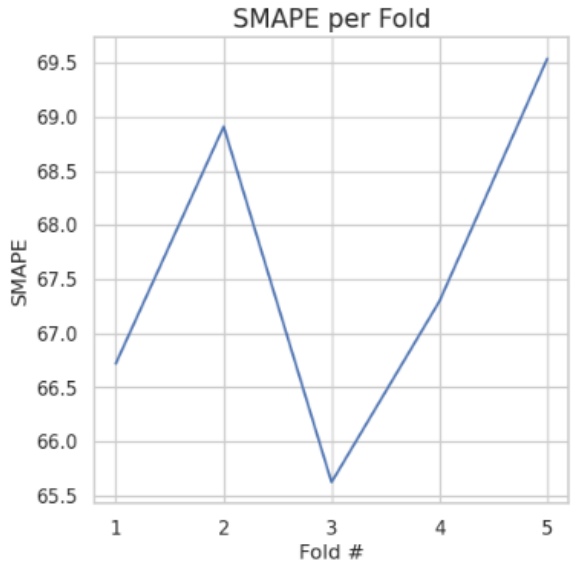
▶ 공통 사항

임상 데이터 / 피처 : **visit_month, month_offset** / 조건 : **UPDRS_4 == 0**

CatBoost 모델 생성 및 학습

모델 ④

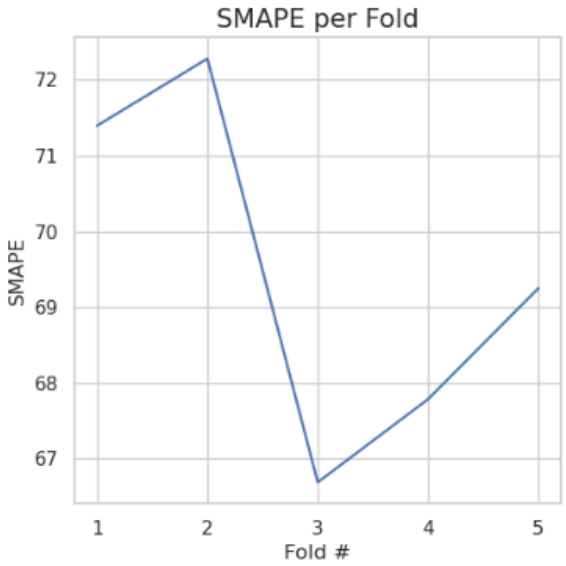
추가 피처 : **medication on/off**



SMAPE = 67.62

모델 ⑤

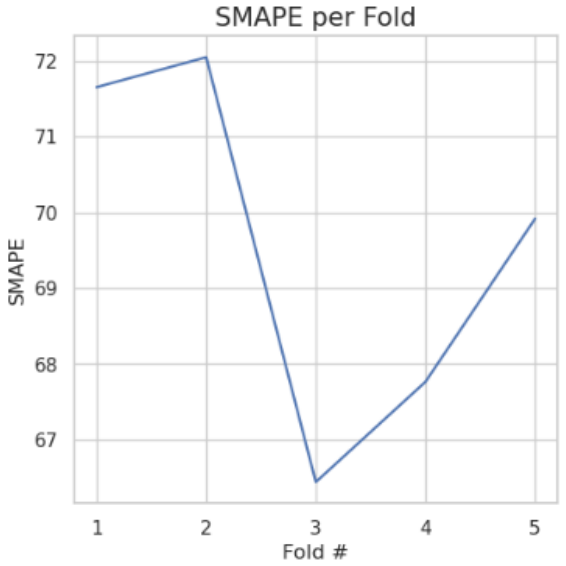
추가 피처 : **protein CV TOP 10**



SMAPE = 69.48

모델 ⑥

추가 피처 : **peptide CV TOP 10**



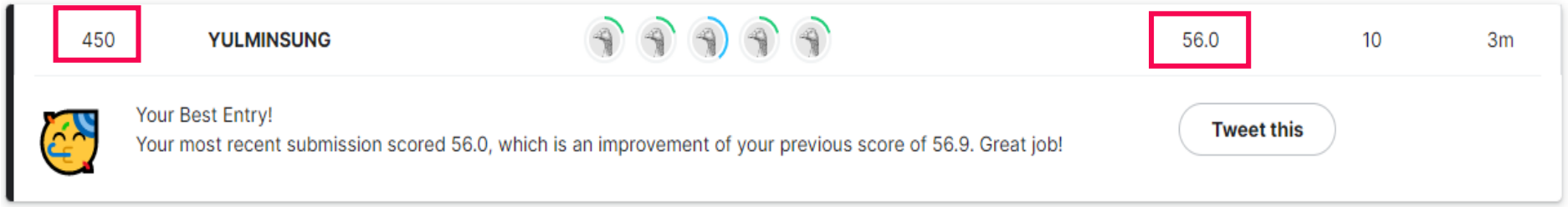
SMAPE = 69.56

CatBoost 모델 별 SMAPE 점수 예측

- **Constant UPDRS 4 : 69.5178**
- **Supplemental Data : 69.4233**
- **Medication State : 67.6184**
- **Protein Data : 69.5638**
- **Peptide Data : 69.5638**

SMAPE 평가지표를 확인했을 때
약물복용여부가 모델의 성능에
가장 많은 영향을 미침

제출 결과



최종 **SCORE** = **56.0**

1,788팀 중 **450** 등

상위 약 **25 %**

Index

05

수행 결과 II

대시 보드

Home / Description / Data

EDA / STAT / ML

01 프로젝트 개요

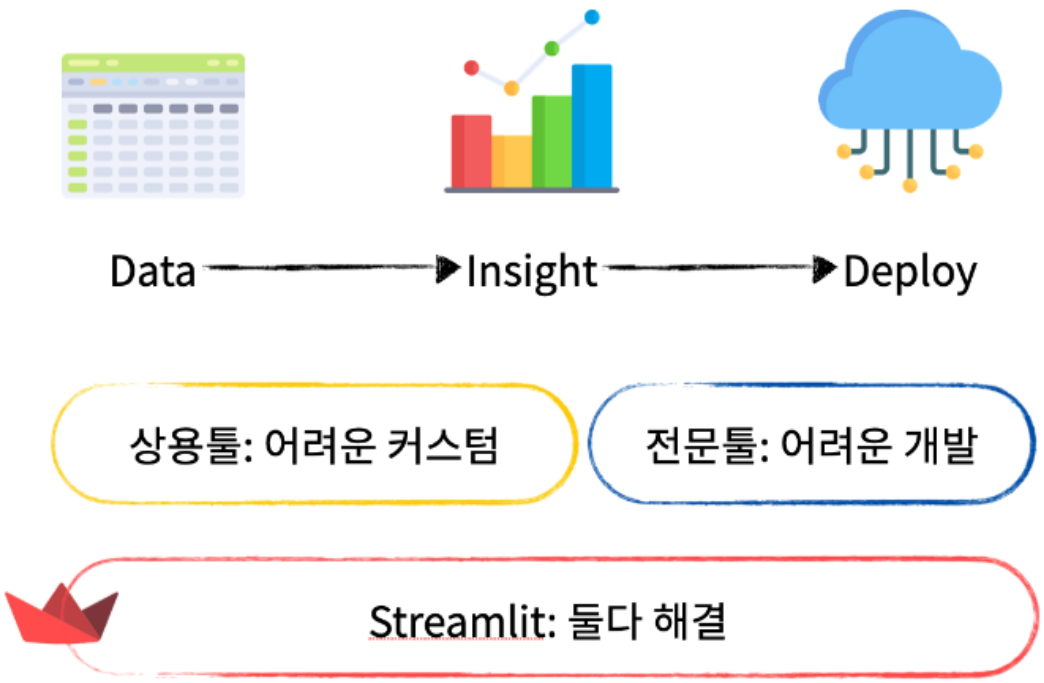
02 팀원 구성 및 역할

03 프로젝트 수행 절차

04 수행 결과 I

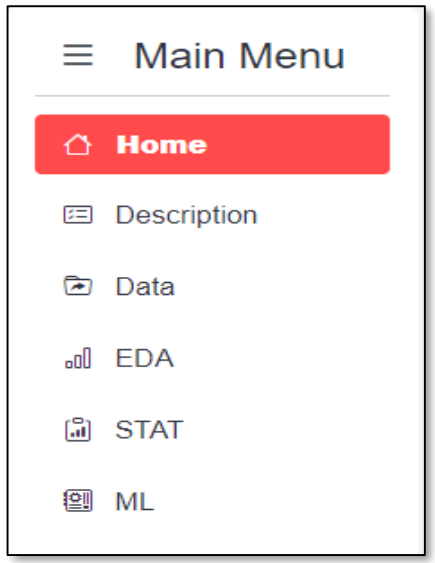
06 평가 의견

Streamlit 소개



Streamlit은 다른 대시보드 **Tool** 에 비해 쉽게 개발이 가능하며, 클라우드 를 활용하여 결과물을 배포 할 수 있는 대시보드 서비스이다.

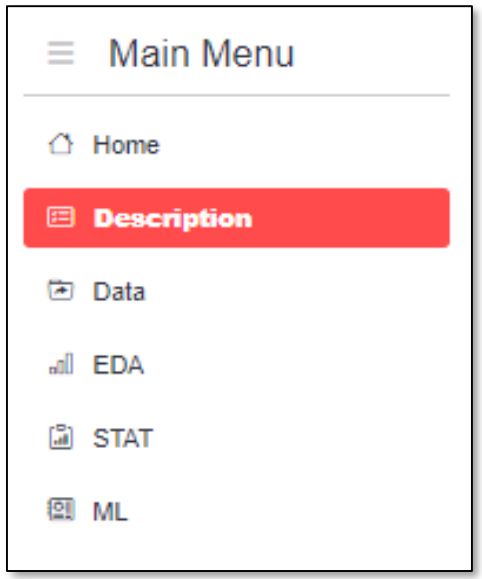
Main Menu 구성 & Home

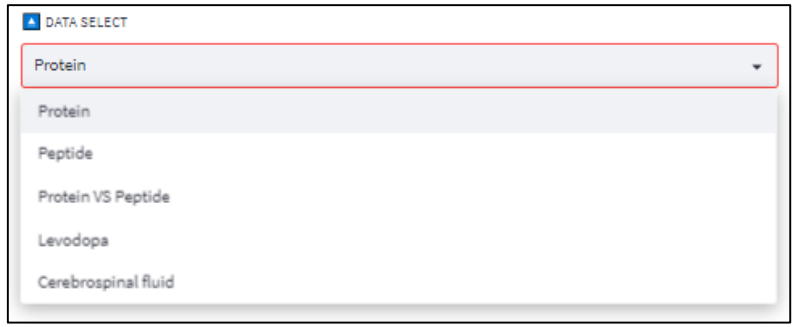


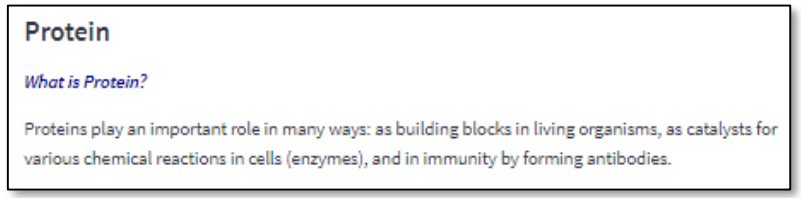
- ✓ **Home** : 대회 목표 및 대회 개요
- ✓ **Description** : 대회에 사용되는 용어 및 평가지표 설명
- ✓ **Data : Dataset** 설명
- ✓ **EDA :** 데이터 시각화로 데이터 탐색과 이해
- ✓ **STAT :** 기초통계량 및 상관관계 설명
- ✓ **ML :** 모델에 학습에 사용한 **Catboost** 및 모델 성능 평가

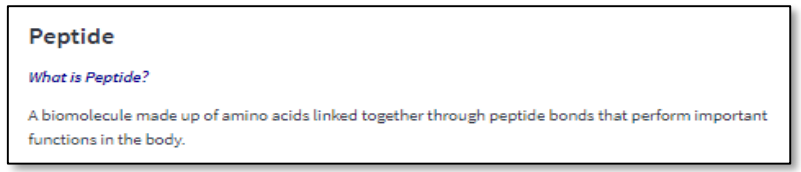
개발환경 / 대회 목표 / 개요 설명

Description





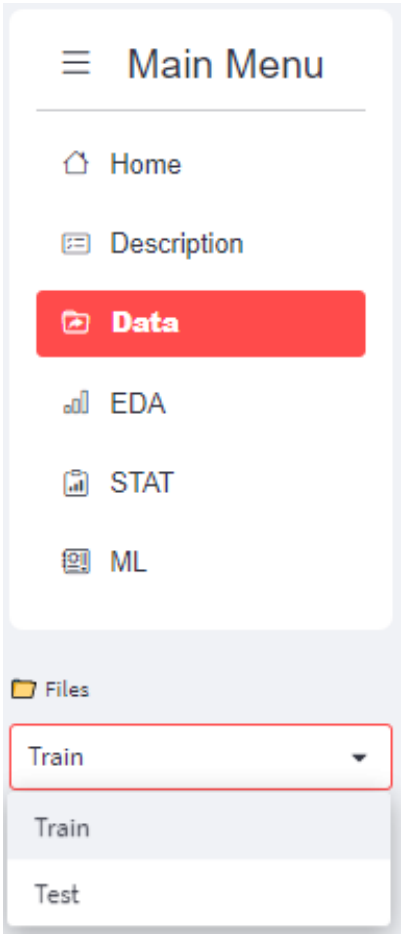




파킨슨 질환과 **MDS-UPDRS**에 대한 설명

Protein, Peptide 등 기본 용어 설명

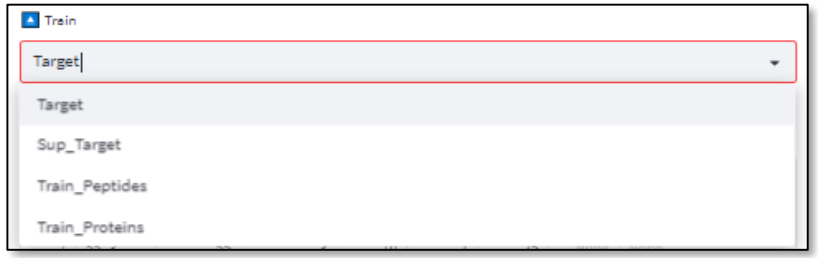
DATA

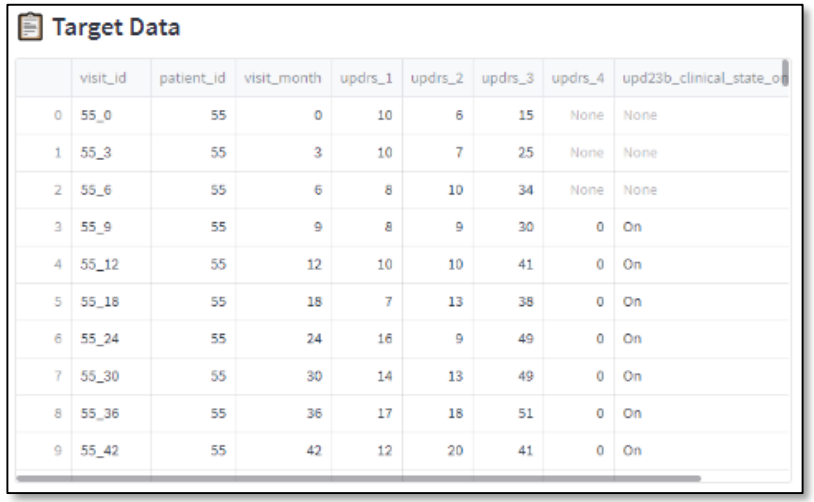


Parkinson's Data

Dataset Column Description

- visit_id - ID code for the visit.
- visit_month - The month of the visit, relative to the first visit by the patient.
- patient_id - An ID code for the patient.
- UniProt - The UniProt ID code for the associated protein. There are often several peptides per protein.
- Peptide - The sequence of amino acids included in the peptide.
- PeptideAbundance - The frequency of the amino acid in the sample.
- NPX - Normalized protein expression. The frequency of the protein's occurrence in the sample.
- updrs_[1-4] - The patient's score for part N of the Unified Parkinson's Disease Rating Scale. Higher numbers indicate more severe symptoms.
- upd23b_clinical_state_on_medication - Whether or not the patient was taking medication such as Levodopa during the UPDRS assessment. Expected to mainly affect the scores for Part 3 (motor function).

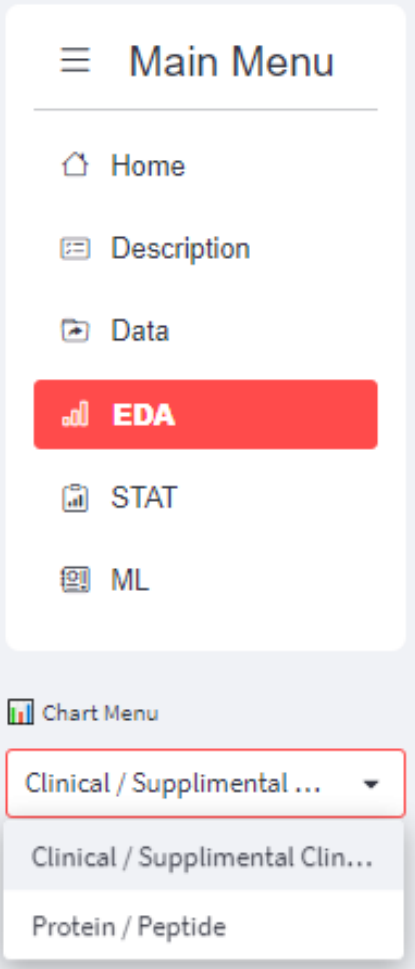


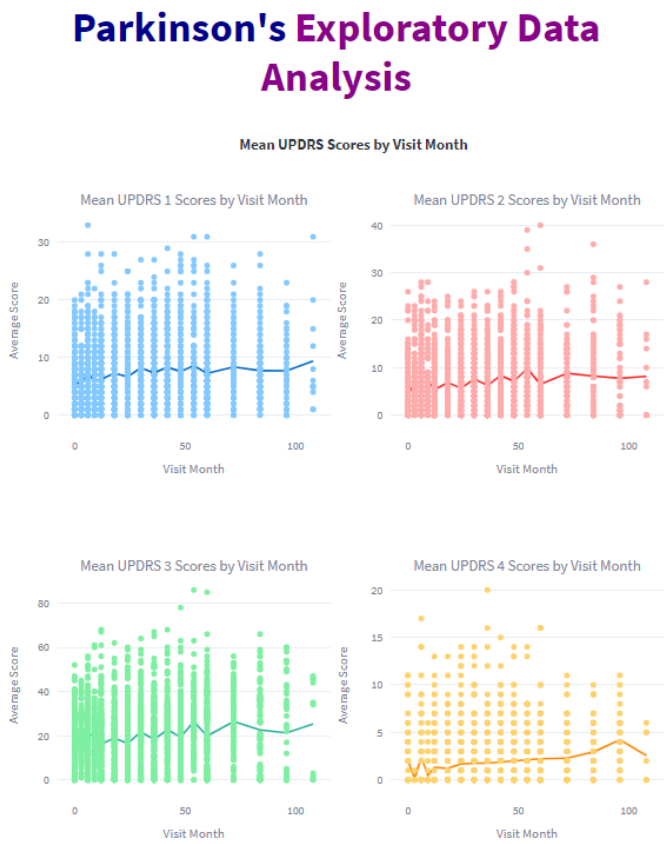


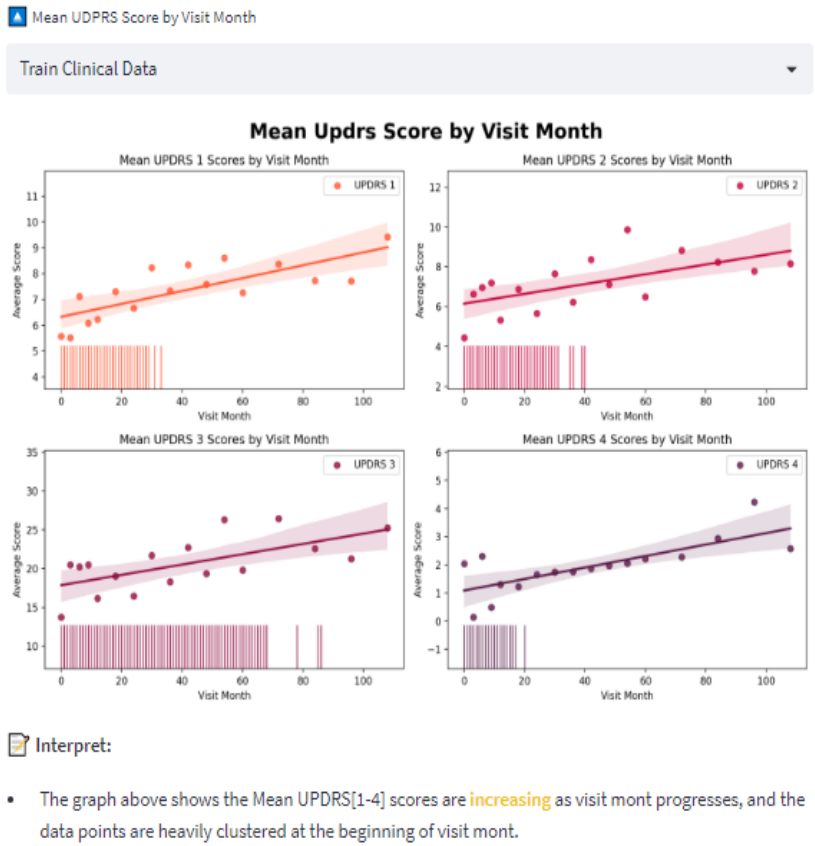
데이터 변수에 대한 설명

데이터 셋 확인 가능

EDA

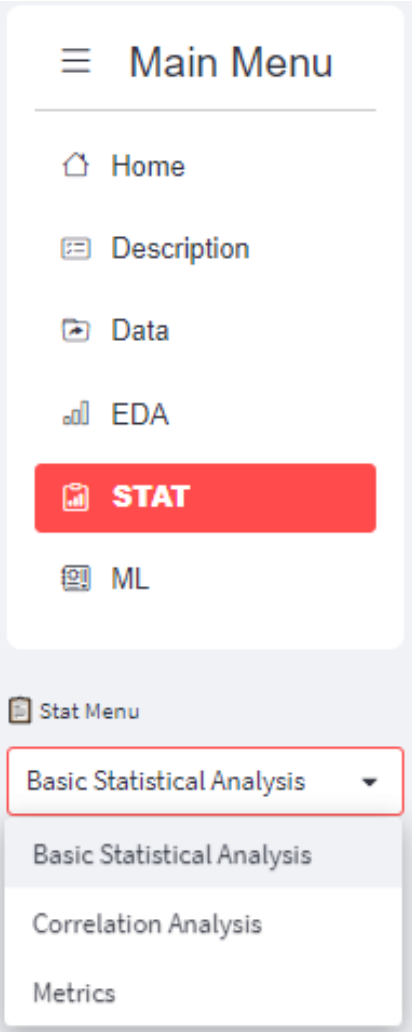


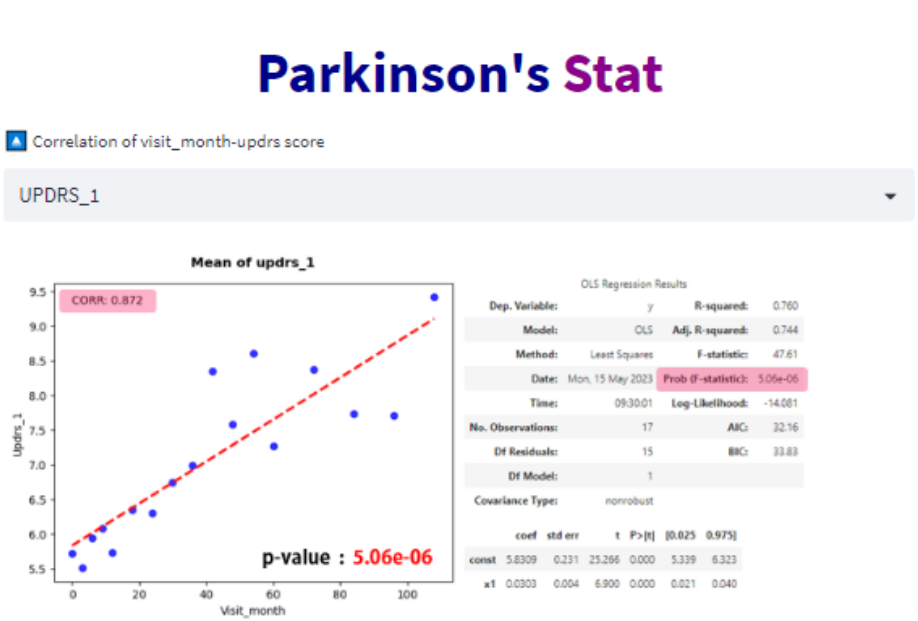


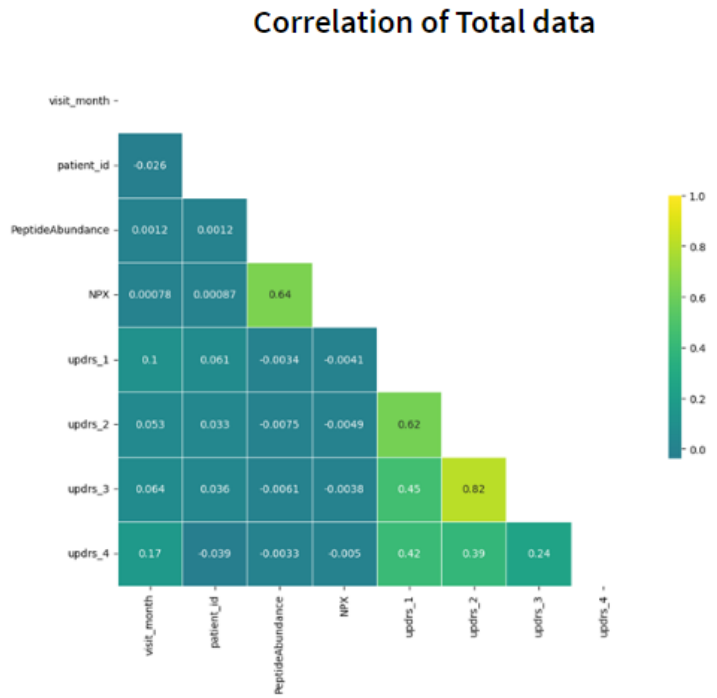


데이터 탐색과 이해를 위한 데이터 시각화

STAT

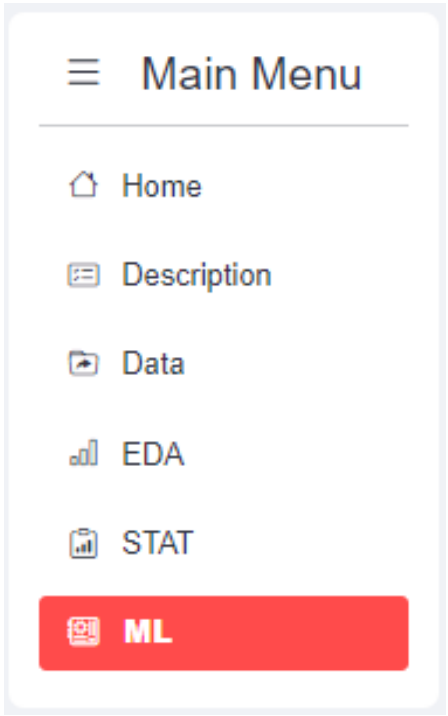


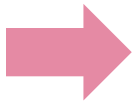


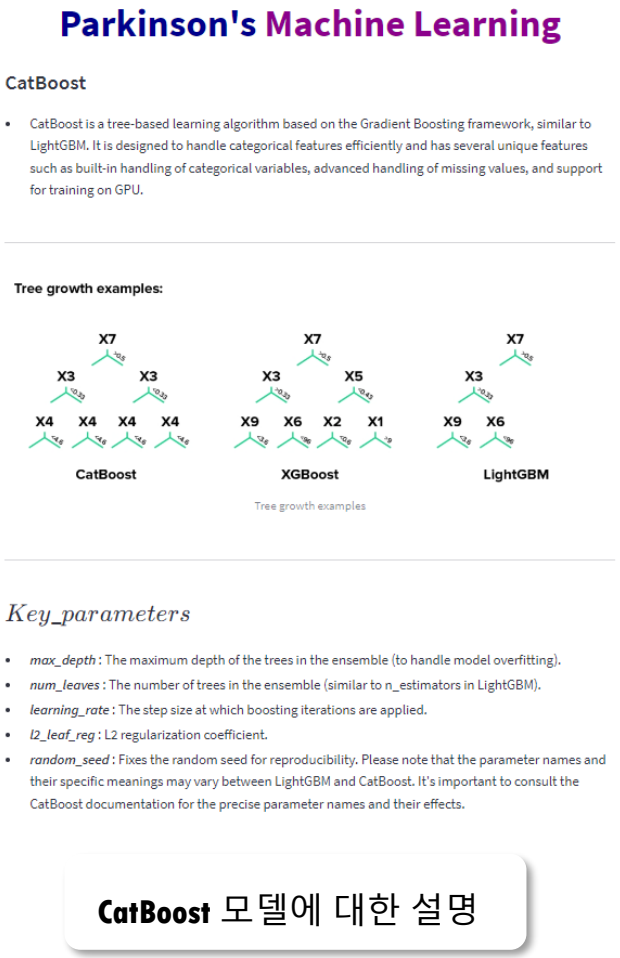


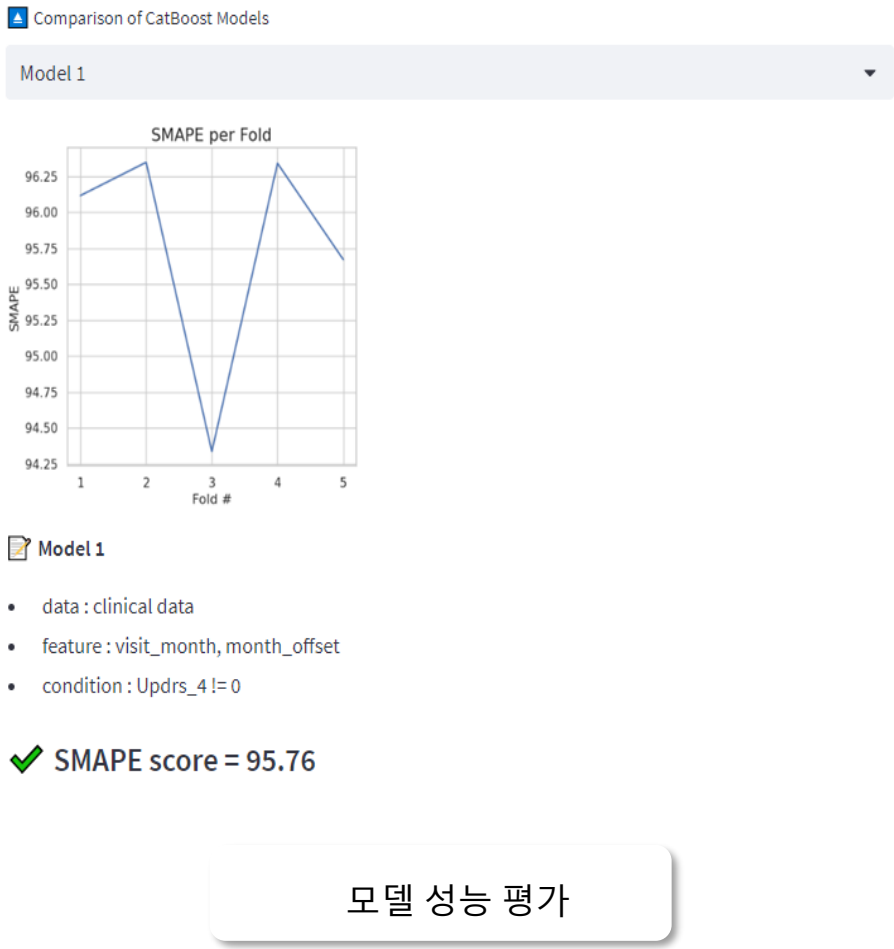
기초통계량 및 상관관계에 대한 설명

ML









CatBoost 모델에 대한 설명

모델 성능 평가

06

자체 평가 의견

분석 결과 요약
한계점 및 개선 사항
출처

01 프로젝트 개요

02 팀원 구성 및 역할

03 프로젝트 수행 절차

04 수행 결과 Ⅰ

05 수행 결과 Ⅱ

자체 평가 의견

분석 결과

Ⅰ. 단일 펩타이드와 단일 단백질로는 **UPDRS**의 점수와 명확한 관계가 있다고 보기 어려움

Ⅱ. 제공 된 데이터에서는 **peptide**, **protein**의 정보보다 피험자의 수(**patient_id**)가 더 중요

Ⅲ. **UPDRS_4**는 다른 **UPDRS**보다 데이터가 더 부족하기 때문에, **UPDRS_4**의 값을 **0**으로 설정했을 때 예측 값이 더 높게 나타남

Ⅳ. 약물을 복용한 사람들이 복용하지 않은 사람들 보다 비교적 느리게 질병이 진행 됨
그러나 시간이 지날수록 둘 다 **UPDRS**의 점수는 증가

자체 평가 의견

한계점 및 개선 사항

Ⅰ. **protein** 데이터가 **visit_month**의 최대 **40%**만 존재하여, 정확한 데이터를 확보했다고 보기 어려움

Ⅱ. **protein**, **pepetide**의 데이터보다 약물복용여부가 더 중요하다고 판단이 되지만,

test_data에서는 약물 복용 여부를 알 수 없음

Ⅲ. 여러가지 모델을 구현 하는 데에 있어 한계가 있음

Ⅳ. 평가지표(**SMAPE**)의 결과값이 대부분 높은 수치를 보이며

SMAPE가 모델 평가에 가장 적합한 평가지표가 맞는지 비교해볼 필요가 있음.

참고 자료

자료 유형	출처
논문	[1] Holden(2018). Progression of MDS-UPDRS Scores Over Five Years in De Novo Parkinson Disease from the Parkinson's Progression Markers Initiative Cohort. Movement Disorders Clinical Practice, 5, 47-53. DOI: 10.1002/mdc3.12553
	[2] Shi(2015). Cerebrospinal Fluid Peptides as Potential Parkinson Disease Biomarkers: A Staged Pipeline for Discovery and Validation*. Molecular & Cellular Proteomics, 14(3), 544–555. DOI: 10.1074/mcp.m114.040576
	[3] Goetz(2008). Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results. Movement Disorders, 23(15), 2129–2170. DOI: 10.1002/mds.22340
	[4] Martinez-Martin(1994). Unified Parkinson's Disease Rating Scale characteristics and structure. Movement Disorders, 9(1), 76-83. DOI: 10.1002/mds.870090112.
웹 사이트	[5] Craig Thomas.(2023)."AMP - EDA + Models" https://www.kaggle.com/code/craigmthomas/amp-eda-models#1.4---Statistical-Breakdown
	[6] 파킨슨센터(2022)."실적 및 통계_외래 진료 환자수" https://www.snumdc.org/mdc/performance/outpatient-practice/
	[7]애옹킴(2022)."파이썬 스트림릿으로 데이터 대시보드 만들기" https://yozm.wishket.com/magazine/detail/1827/
	[8] 김병희 기자(2020). "한 번 처치로 뇌세포 생성해 파킨슨병 치료." https://www.ibric.org/myboard/read.php?Board=news&id=318691