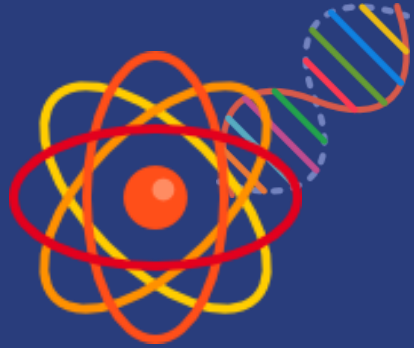


KUMED 14기 박상준 14기 제갈예빈 14기 채윤병

# MoA와 당뇨병- classification

KUBIG 2021-2 의료데이터 프로젝트



# CONTENTS



## MoA Prediction

- 전처리
- 모델링



## AIDD 당뇨병 예측

- 전처리
- 모델링



# MoA Prediction

 Research Code Competition

## Mechanisms of Action (MoA) Prediction

Can you improve the algorithm that classifies drugs based on their biological activity?

**\$30,000**  
Prize Money

 Laboratory for Innovation Science at Harvard · 4,373 teams · a year ago

**MoA** : 특정 약물에 대한 세포 내의 분자들의 생물학적 기능을 알려주는 지표

**대회 목표** : 각 샘플에 대해 반응하는 하나 혹은 그 이상의 MoA를 예측하는 다중 분류 문제

**평가 지표** : Log Loss





# MoA Prediction - DATA

Train\_Target : 206가지 종류의 약물에 대해 반응했는지 여부에 따라 이진 분류

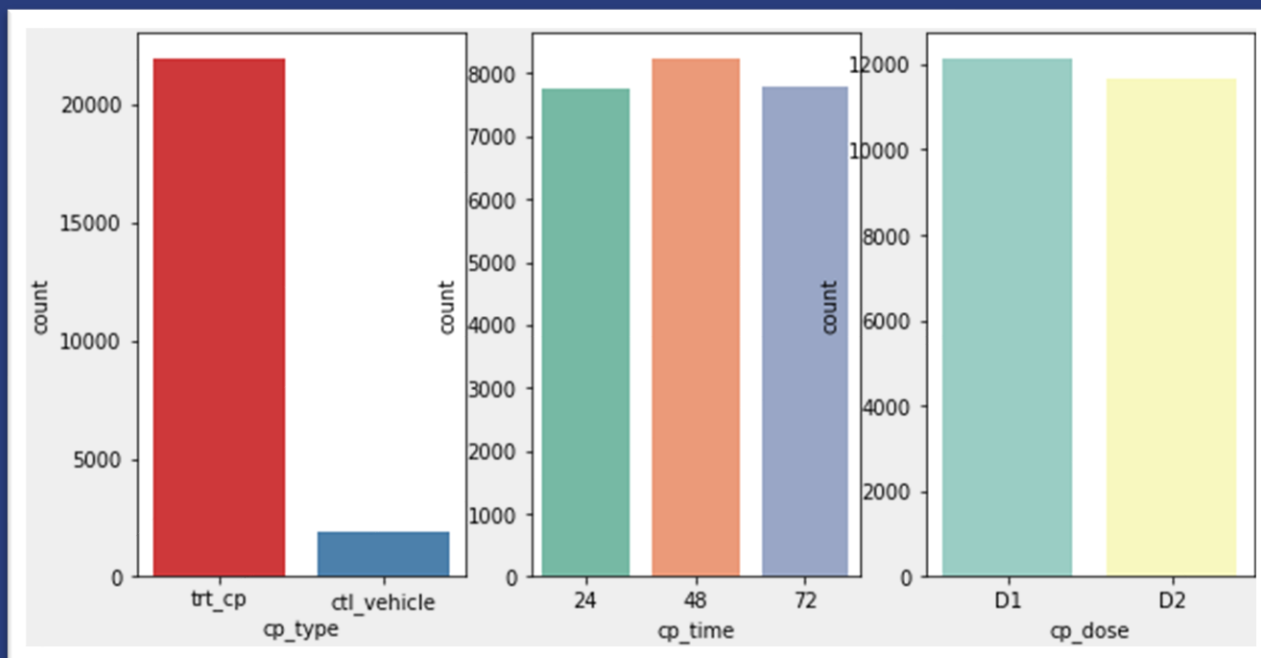
Train\_Feature : 연속형 : g - 유전자 표현 -> 772개의 컬럼

c - 세포 생존 가능성 -> 100개의 컬럼

범주형 : cp type - 샘플 처리 방식 -> trt\_cp(93%), ctl\_vehicle(7%)

cp time - 치료 기간 (24h, 48h, 72h)

cp dose - 복용량 (high or low)



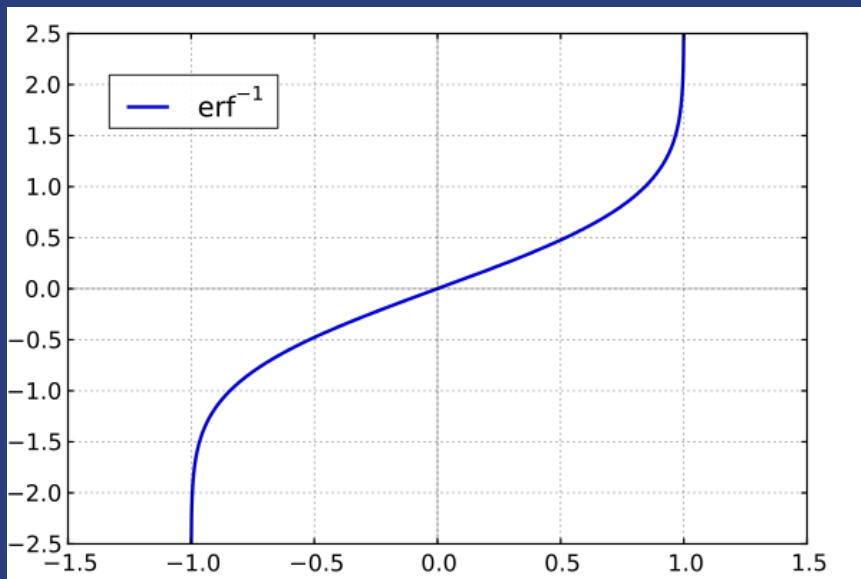


# MoA Prediction - 전처리

## 1. Scaler (Rank Gauss, Standard)

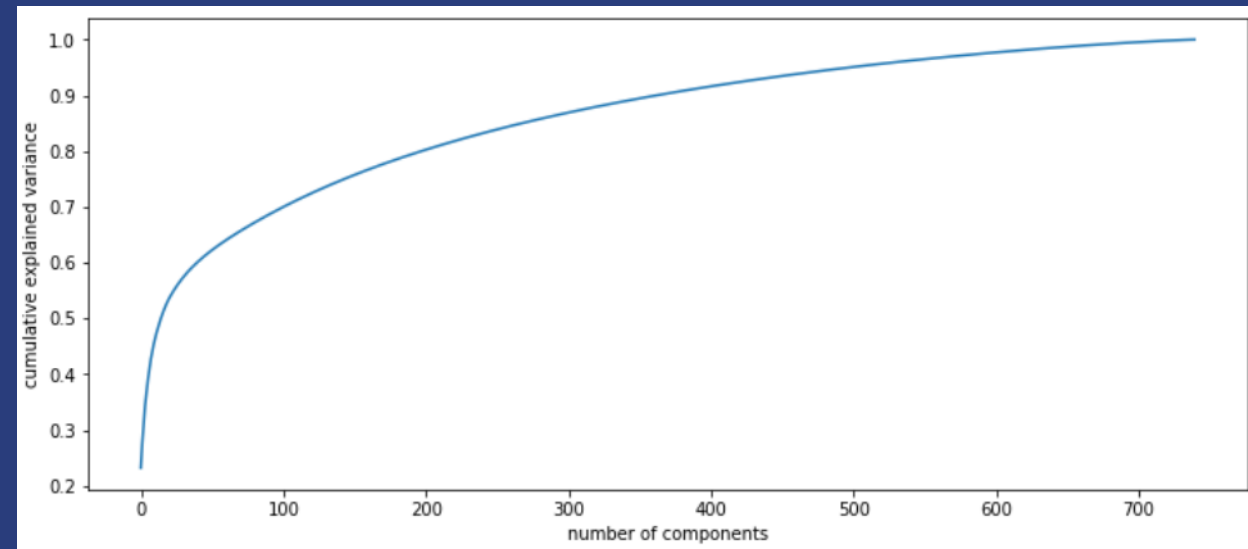
Rank Gauss :

딥러닝에서 정규화를 빠르게 시켜줄 수 있는 방법  
연속형 변수의 각 rank를 측정 한 후에 (-1.1) 범  
위로 scaling해준 후 오차역함수  $\text{erf}^{-1}$ 를 적용  
해 가우시안 분포가 되도록 변형해주는 방식



## 2. PCA (차원 축소)

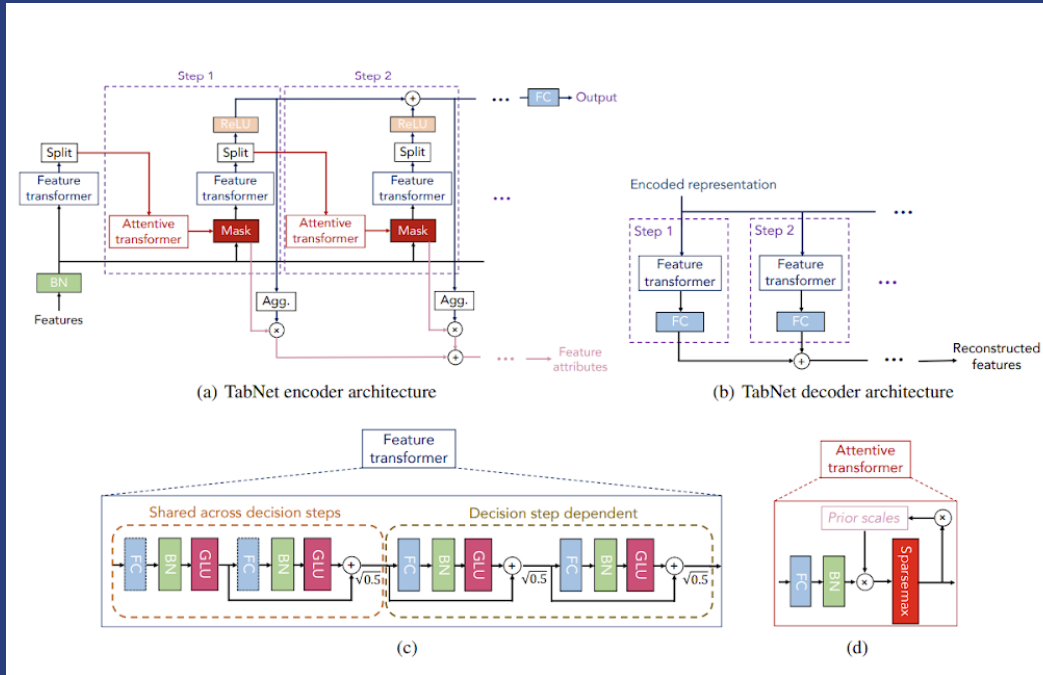
총 컬럼이 876개로 매우 많기 편이므로 차원 축소 필요  
g - 컬럼 중 80%의 설명력을 보장하는 175개의 컬럼  
선정 / 급히 꺾이는 지점인 70개 선정  
c - 1개 만으로도 충분히 설명 가능 / 20개 선정  
총 180개의 컬럼으로 축소 / 총 94개의 컬럼으로 축소





# MoA Prediction - 모델링

## Tabnet



- Tree 기반 gradient boosting을 활용하여 DNN에 반영

- Encoder – feature transformer, attentive transformer, feature masking 으로 구성

→ Attention mechanism을 활용한 feature engineering & selection

- Decoder – feature transformer 블록으로 구성





# MoA Prediction - 모델링

## Tabnet 결과

-----  
FOLDS: 9

\*\*\*\*\*

Device used : cuda

epoch 0		loss: 0.29069		val_logits_ll: 0.02899		0:00:02s
epoch 10		loss: 0.01881		val_logits_ll: 0.01868		0:00:20s
epoch 20		loss: 0.01675		val_logits_ll: 0.01665		0:00:39s
epoch 30		loss: 0.01629		val_logits_ll: 0.01632		0:00:57s
epoch 40		loss: 0.01601		val_logits_ll: 0.01613		0:01:16s
epoch 50		loss: 0.01574		val_logits_ll: 0.016		0:01:35s
epoch 60		loss: 0.01563		val_logits_ll: 0.01592		0:01:54s
epoch 70		loss: 0.0155		val_logits_ll: 0.01591		0:02:13s
epoch 80		loss: 0.01532		val_logits_ll: 0.01598		0:02:32s
epoch 90		loss: 0.01521		val_logits_ll: 0.01587		0:02:51s
epoch 100		loss: 0.01517		val_logits_ll: 0.01597		0:03:10s
epoch 110		loss: 0.01499		val_logits_ll: 0.01578		0:03:29s

Early stopping occurred at epoch 118 with best\_epoch = 98 and best\_val\_logits  
Best weights from best epoch are automatically used!

AutoML 中 optuna 사용  
Best score : 0.01578





# AIDD 대회 개요

- 경희대학교 의료원에서 주최한 당뇨병 예측 모델 개발 대회
- 당뇨병 유무를 확인하기 위한 분류 모델 생성
- 11/19~ 11/22 4일간 진행
- Baseline code를 기반으로 NSML에 제작한 모델을 올리고, 리더보드에서 각 팀의 순위를 매김 (평가지표: AUC score)
- 제출은 1시간에 1번만 가능

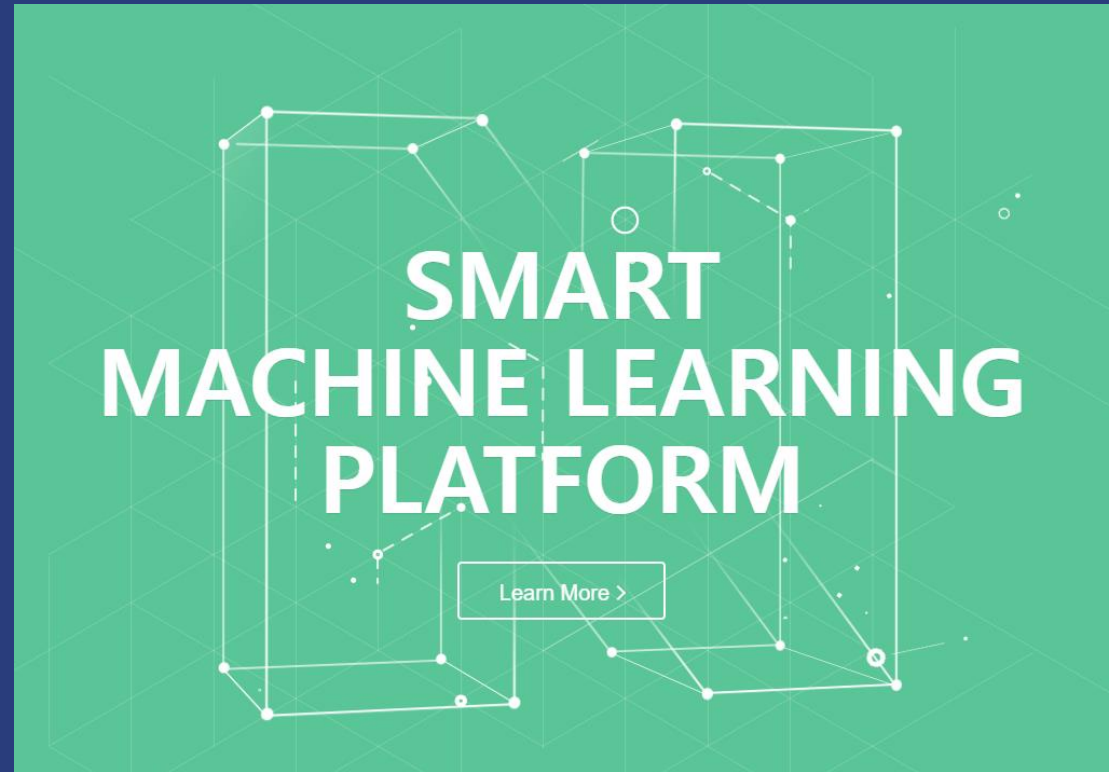






# NSML 설명

- NSML: 네이버에서 개발한 머신러닝 클라우드 플랫폼
- GPU 할당, 학습 과정 시각화, 제작 모델 간 성능 비교를 가능케 함
- NSML와 친숙해지고 코드를 익히느라 하루 정도의 시간이 소요되었음





# 데이터 설명

- Data 개수: 2,438개의 training set을 기반으로 431개의 test set로 모델 평가
- 변수: 22개의 임상변수, baseline & endpoint 건강검진 일자, target 변수로 구성
  - Target 변수: 당뇨병 발병일 때 1, 아닐 때 0 부여
- Training set의 csv파일을 제공하지 않아 nsml 서버에서 raw data를 정제해야 했음

	CDMID	gender	age	date	Ht	Wt	BMI	SBP	DBP	PR	...	Alb	BUN	Cr	CrCl	AST	ALT	GGT	ALP	date_E	target
0	11782172	F	59.0	2015.05.15	159.0	56.0	22.2	99.0	59.0	60.0	...	4.2	16.0	0.6	108.75	22.0	22.0	15.0	82.0	2016.04.11	0
1	11864586	F	51.0	2016.08.26	160.0	56.0	21.9	140.0	85.0	73.0	...	4.4	11.0	0.6	112.02	22.0	18.0	45.0	98.0	2019.10.25	0
2	11498294	M	36.0	2013.06.28	162.0	56.0	21.2	100.0	55.0	60.0	...	4.5	13.0	0.9	76.40	18.0	20.0	19.0	50.0	2019.04.26	0
3	41710350	F	53.0	2012.03.20	150.0	47.2	21.0	105.0	62.0	78.0	...	4.1	9.0	0.7	93.03	30.0	20.0	12.0	157.0	2017.12.18	0
4	41712205	M	61.0	2008.11.10	168.1	66.1	23.4	104.0	67.0	88.0	...	4.1	12.0	0.9	91.18	15.0	29.0	69.0	348.0	2014.12.02	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2433	42158163	F	55.0	2014.05.16	165.3	61.9	22.7	116.0	77.0	63.0	...	4.4	12.0	0.5	136.15	30.0	26.0	77.0	71.0	2019.05.31	0
2434	10716568	M	55.0	2011.04.29	166.2	66.4	24.0	112.0	57.0	50.0	...	4.2	12.0	0.9	74.03	22.0	14.0	13.0	62.0	2018.05.03	0
2435	11380684	M	55.0	2016.11.28	173.0	76.0	25.6	183.0	109.0	58.0	...	4.7	14.0	1.0	76.26	46.0	50.0	215.0	83.0	2019.09.25	0
2436	10905760	M	79.0	2016.05.10	169.0	69.0	24.2	125.0	64.0	56.0	...	4.2	24.0	1.0	76.61	37.0	32.0	39.0	116.0	2018.02.09	0
2437	10844734	M	54.0	2016.03.31	164.0	49.0	18.4	115.0	71.0	69.0	...	4.3	21.0	0.8	62.19	23.0	13.0	20.0	73.0	2019.04.15	0

2438 rows × 26 columns

<Training set>

번호	변수명	설명	번호	변수명	설명
01	gender	성별	13	TG	중성지방
02	age	나이	14	LDL	LDL 콜레스테롤
03	date	Baseline 건강검진 일자	15	HDL	HDL 콜레스테롤
04	Ht	신장	16	Alb	알부민
05	Wt	체중	17	BUN	혈중요소질소
06	BMI	체질량지수(BMI)	18	Cr	크레아티닌
07	SBP	수축기혈압	19	CrCl	크레아티닌 청소율
08	DBP	이완기혈압	20	AST	아스파테이트아미노전이효소
09	PR	맥박	21	ALT	알라닌아미노전이효소
10	HbA1c	당화혈색소	22	GGT	감마글루타밀전이효소
11	FBG	공복혈당	23	ALP	알칼리인산분해효소
12	TC	총콜레스테롤	24	date_E	Endpoint 건강검진 일자

당뇨병 진단 기준: (1) 공복혈당(FBG) 126 mg/dL 이상, 또는 (2) 당화혈색소(HbA1c) 6.5% 이상

<변수 종류>



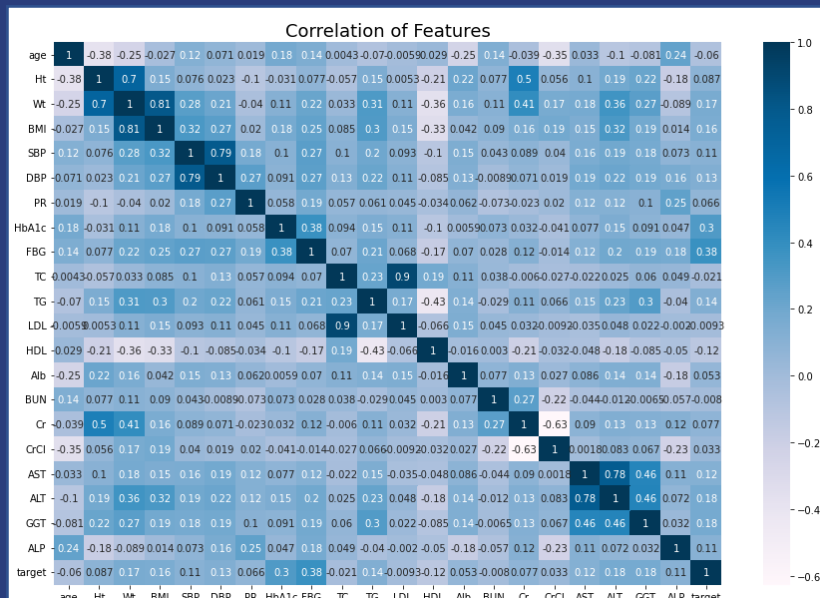


# AIDD - 전처리

- 1. Dataprep으로 변수 별 특징 파악
- 2. Correlation plot을 활용하여 Target과 변수들 간의 관계 파악
- 3. Correlation plot, 로지스틱 회귀분석의 stepwise selection 등의 방법을 통해 여러 조합의 변수를 선택
- 4. 보건정책관리학부 '현대인과 만성질환' 수업에서 제시한 당뇨병 판정 기준을 활용하여 새로운 범주형 변수도 생성해 보는 등 다양한 시도를 함

## 최종 선택 변수

성별, 나이, BMI, 당화혈색소, 공복혈당, 총콜레스테롤, LDL 콜레스테롤, 3가지 전해요소, 알칼리인산분해요소



<Correlation Plot>

## Diagnostic criteria for pre-diabetes and diabetes

FPG and OGTT show your blood glucose level **at the time** of the test

	Fasting plasma glucose test (FPG)	2-hour oral glucose tolerance test (OGTT)
Normal	<100 mg/dL	<140 mg/dL
Pre-diabetes	100-125 mg/dL	140-199 mg/dL
Diabetes	≥126 mg/dL	≥200 mg/dL

### A1C test

- Reflects **average** blood glucose over the **past 3 months**
- Pre-diabetes: HbA1c 5.7-6.4%
- Diabetes: HbA1c ≥ 6.5%

OGTT identify how your body handles glucose after a meal

<당뇨병 판정 기준>



# AIDD - 전처리

- 5. 연속형 변수에 대한 이상치 제거 및 정규화
  - 이상치 기준:  $Q3 + 1.5 * IQR$ 을 초과하거나  $Q1 - 1.5 * IQR$  미만
- 6. 범주형 변수에 대한 더미 변수 부여

```
[ ] def outliers(df, columns):
    outlier_indices = []
    for col in columns:
        Q1 = np.percentile(df[col], 25)
        Q3 = np.percentile(df[col], 75)
        IQR = (Q3 - Q1) * 1.5
        lowest = Q1 - IQR
        highest = Q3 + IQR

        outlier_index = df[col][(df[col] < lowest) | (df[col] > highest)].index

        for index in list(outlier_index):
            if index not in outlier_indices:
                outlier_indices.append(index)

        #outlier_indices.append(outlier_index)
        #outlier_indices.extend(outlier_index)
        #outlier_indices = Counter(outlier_indices)
        #multiple_outliers = list(k for k, v in outlier_indices.items() if v > n)

    return outlier_indices

[ ] outliers_to_drop = outliers(df_all, ['Ht', 'PR', 'TC', 'LDL', 'A1b', 'BUN', 'Cr'])
len(outliers_to_drop)
```

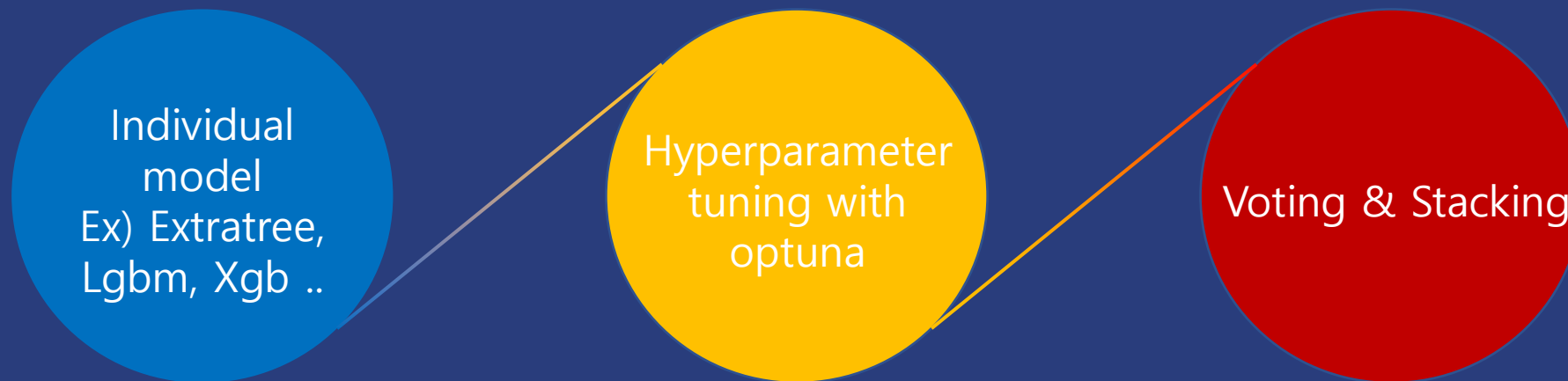
```
[ ] num_pipeline = Pipeline([
    ('imputer', IterativeImputer()),
    ('scaler', StandardScaler()),
])

[ ] from sklearn.compose import ColumnTransformer
    from sklearn.preprocessing import OneHotEncoder

    num_attribs = list(train_x_num)
    cat_attribs = ['gender']
    full_pipeline = ColumnTransformer([
        ("num", num_pipeline, num_attribs),
        ("cat", OneHotEncoder(), cat_attribs),
    ])
```



# AIDD - 모델링



앙상블, 부스팅 기법을 사용하는 머신러닝 모델 채택

- 개별 모델에서 optuna 모듈을 이용해 하이퍼파라미터 최적화
- 모델 Voting & Stacking





# AIDD - 결과

결과 제출의 제한(1시간당 한 번) & Test set 공개 x > 다양한 모델의 조합으로 Voting과 Stacking을 진행하여 결과 제출을 통해 성능 확인

```
XGB = XGBClassifier(alpha=_0.3050062504473983, subsample=_0.6, learning_rate=_0.01, n_estimators=_600,
                    max_depth=_17, random_state=_42, min_child_weight=_1, use_label_encoder=_False)
ET = ExtraTreesClassifier(n_estimators=_810, criterion=_'entropy', max_depth=_14, min_samples_split=_4,
                          min_samples_leaf=_3, random_state=_42)
LGB = LGBMClassifier(lambda_l1=_1.8037439202178834e-05, lambda_l2=_0.06372653492091832, num_leaves=_7,
                     feature_fraction=_0.9819459112971965, bagging_fraction=_0.899465584480253, bagging_freq=_2,
                     min_child_samples=_22, path_smooth=_8.260808399079588e-08, random_state=_42)
```








XGB, ExtraTree, LGBM에 각각 optuna를 사용해 최적화를 진행 > Stacking > 최고의 성능 (0.781)

\* Voting 방법을 사용했을 때 : 0.758





# AIDD - 결과

17		피니톨	0.7840014064697608
18		학부생과 석사 과정생	0.7827355836849508
19		DSS17	0.7814697609001406
20		KUMED	0.7814697609001406
21		김천낙오단	0.7802390998593529
22		PDXen	0.7801687763713079
23		DAI Team	0.7776371308016877

최종 성능 : 0.781  
예선 참가 40팀 중 22위 기록  
본선 참가(20팀)에는 실패...!

아쉬웠던 점

- NSML의 미숙
- 한정된 시간과 제출 빈도 제한

대회를 통한 경험





# Summary

KUBIG 2021-2 의료데이터 프로젝트 활동정리

## 1. MoA Prediction (Kaggle)

- 주어진 데이터(유전자 발현, 세포의 생존도)를 이용한 약물의 생물학적 기능 예측
- 스케일링, PCA -> Tabnet

## 2. AIDD 당뇨병 예측 해커톤

- 환자의 건강검진 데이터를 이용한 당뇨병 발병 여부 예측
- 해커톤 -> 한정된 시간과 제출 빈도 제한
- 다양한 전처리와 모델링 -> 예선 40팀 중 22위 기록