

난임 환자 대상 임신 성공 여부 예측 AI 해커톤

3KS

김태은, 신운서, 권형준

목차

메인타이틀에 대한 세부 설명을 입력해 주세요.

01

도메인지식 및 가설

도메인지식 활용을 통한 가설설정

02

EDA

타이틀에 대한 설명을 입력해 주세요.

03

특성공학

타이틀에 대한 설명을 입력해 주세요.

04

모델

타이틀에 대한 설명을 입력해 주세요.

05

인사이트

타이틀에 대한 설명을 입력해 주세요.

도메인지식 및 가설

도메인지식 활용을
통한 가설설정

도메인지식 및 가설

Intervention

3 groups; Group I: women who had a live birth in the first cycle, Group II those who had a miscarriage, Group III, women who had a negative pregnancy test in their first cycle.

Outcome measures

Pregnancy (PR), Live birth (LBR) & miscarriage rates in the second cycle.

Results

For women < than 40: PR was 46.4% (368/793), miscarriage rate was 29.9% and the LBR was 32.5% (258/793). Women in groups I & II had a statistically higher PR than those in group III 63.3% v 55.2% v 41.9% respectively. LBR was higher 45% v 37.8 v 29.6% respectively. Miscarriage rate was similar.

For women 40 years and older: The PR was 21.0% (73/348), miscarriage rate was 52.1% (38/73) and the LBR was 10.1% (35/348). There was no significant difference in PR among women in groups I, II & III. The LBR and miscarriage rates were similar in all groups.

✓ 이전 성공률이 높은 환자는 다시 성공 확률이 높다.

과거 IVF 성공은 자궁 수용성, 배아 질, 생식 환경을 간접적으로 반영

도메인지식 및 가설

In conclusion, the current study provides the first set of relevant data on CLBRs over multiple IVF cycles in AMA women categorized by age and ovarian reserve. Our findings support the efficacy of extending the number of cycles up to three or four until the age of 43 and recommendations should be given individually considering the age and ovarian reserve. Women above the age of 43 is not cost-effective to continue repeated IVF treatment using their own oocytes. Further work is required to move towards tailored protocols to maximize the IVF success rate of each age-specific POSEIDON group without compromising safety.

✓ 고령 + 반복 시술은 예후 나쁨

고령에서 IVF를 반복해도 짧은 연령만큼 보상이 되지 않고, 40세 이상에서는 반복해도 예후가 급격히 나빠진다

도메인지식 및 가설

Result(s)

Of the 102 sites, considerable numbers were noncompliant with ASRM's guidelines that prohibit varying compensation based on a donor's traits (34%), and recommend an age of 21 years or older (41%), and presentation of risks alongside compensation (56%). Trait-based payment variation was associated with being an agency rather than a clinic, location in the West, not being endorsed by ASRM or Society of Assisted Reproductive Technology (SART), and referring to ASRM's guidelines about compensation. Of sites mentioning traits, prior donation success was the most commonly paid for trait (64%).

Conclusion(s)

Our data, the first to systematically analyze agency and clinic websites reveal that many do not follow ASRM's guidelines. These data have critical implications for policy, practice, and research, suggesting needs for consideration of possible changes in guidelines, and/or improvements in compliance and monitoring by ASRM or others.

✓ ICSI 의존도가 높은 경우 → 남성요인 가능성

ICSI 생성비율 = 미세주입 생성 배아 수 / 총 생성 배아 수

도메인지식 및 가설

- ✓ 배아이식 비율 낮으면 배아 질이 낮을 가능성이 있다.
-

이식 비율 = 이식된 배아 수 / 총 생성 배아 수

도메인지식 및 가설

- ✓ 수정률 낮으면 배아 질 낮다.
-

수정률 = 수정된 난자 수 / 혼합된 난자 수

도메인지식 및 가설

✓ 5일 배양 배아는 일반적으로 성공률 높다.

- 5일까지 *in vitro*에서 생존해 포배기까지 도달한 배아는, 3일 시점에 비해 세포 수가 훨씬 많고 분열·발달이 안정적으로 진행된 선택된 고품질 배아일 가능성이 높다.
- 임상 배양 전략에서 “난자와 배아 수가 충분한 경우 5일까지 배양해 포배기까지 도달한 배아만 선별 이식하면 착상률·임신률이 더 높다.

특성공학(파생변수)

이전 성공률 = 총 임신 횟수/총 시술 횟수

IVF 성공률 = IVF 임신 횟수/ IVF 시술 회수

DI 성공률 = DI 임신 횟수/DI 시술 횟수

특성공학(파생변수)

고령_다회시술 = (나이 >= 35) & (총 시술 횟수>=3)

특성공학(파생변수)

ICSI 생성비율 = 미세주입 생성 배아 수 / 총 생성 배아 수

특성공학(파생변수)

이식비율 = 이식된 배아 수 / 총 생성 배아 수

특성공학(파생변수)

수정률 = 수정된 난자 수 / 혼합된 난자 수

특성공학(파생변수)

PGT_사용 = PGS or PGD or 착상유전자검사

1차 EDA



✓ 특정 시술 유형 칼럼 전처리

!!-> 문제발견: 데이터 중복값 다수 있는거 확인
text 들을 split 해서 AH / BLASTOCYST 가 들어가 있으면 각각 1
이 채워지도록 has_AH, has_Blastocyst 칼럼 생성

ICSI + IVF 가 같이 있으면 mixed 로 변환

ICSI가 있으면 ICSI 로 변환

IVF가 있으면 IVF 로 변환

IUI가 있으면 IUI 로 변환

DI가 있으면 DI 로 변환

Unknown가 있으면 Unknown 로 변환

나머지는 Other 로 변환

→ 결론 Main_procedure / has Blastocyst / has AH 칼럼 생성

2차 EDA

미세주입에서 생성된 배아 수	6291
이식된 배아 수	6291
미세주입 배아 이식 수	6291
저장된 배아 수	6291
미세주입 후 저장된 배아 수	6291
해동된 배아 수	6291
해동 난자 수	6291
수집된 신선 난자 수	6291
저장된 신선 난자 수	6291
혼합된 난자 수	6291
파트너 정자와 혼합된 난자 수	6291
기증자 정자와 혼합된 난자 수	6291
동결 배아 사용 여부	6291

✓ 공통 결측치 6291열들 전부 삭제

!!> 문제발견: 단일 배아 이식 여부 데이터 칼럼들 포함 여러 칼럼에서 공통 결측치 다수 보임 -> dropna(subset) 를 통해 전부 삭제

3차 EDA



✓ 배아 생성 주요 이유 칼럼 결측치 전처리

배아 생성 주요 이유 칼럼에 들어있는 데이터 목록

!!-> 문제발견: 중복되는 값이 많아보여, 정리 필수!
전처리

1. 배아 생성 주요 이유 전처리 함수 생성
2. 5개 섹터(현재시술용, 배아저장용, 난자저장용, 기증용, 연구용)으로 나누어서 정리
3. 입력되는 text에서 빈칸 삭제 + split 단어 분할
4. 해당 내용이 있으면 해당 칼럼 데이터 1 아니면 0

4차 EDA



✓ 나이 칼럼 숫자형으로 변환

!!> 문제발견: 문자열로 되어 있으면, 나이 분포에 따른 데이터 이점을 가져가지 못함

총 6개 데이터

만18~34세 -> 1

만35~37세 -> 2

만38~39세 -> 3

만40~42세 -> 4

만43~44세 -> 5

만45~50세 -> 6

5차 EDA



✓ 불임 원인 칼럼 전처리

1. 불임 원인 - 여성 요인 칼럼 모든 데이터 값이 0이므로 제거
2. 불임 원인 - 자궁경부 문제 데이터값은 10만개 중 한개만 1 나머지 모두 0 이므로 매우 희소한 데이터 값 -> 학습 노이즈 될 수 있음. → 삭제진행

!!-> 문제발견: 데이터 값이 0인 칼럼 존재 -> unique값이 1인 칼럼들 삭제 진행

!!-> 문제발견: 실제 데이터가 26만개 중 10개 미만인 칼럼 데이터 존재 -> 자궁경부 문제 칼럼 삭제

6차 EDA



✓ 횟수 관련 칼럼들 데이터 전처리

!!> 문제발견: 문자열 데이터인 것 확인하고 수치형으로 변형

'0회': 0,

'1회': 1,

'2회': 2,

'3회': 3,

'4회': 4,

'5회': 5,

'6회 이상': 6

7차 EDA



✓ 수치형 데이터 중 nan 값 처리

난자 채취 경과일 [0, nan] -> 그냥 칼럼 삭제
난자 혼합 경과일, 배아 이식 경과일, 배아 해동 경과일 nan 값들
기록되지 않음. -> 모두 0일 처리

8차 EDA

[임신 시도 또는 마지막 임신 경과 연수]

결측치 수: 240993

[nan 10. 6. 7. 8. 12. 9. 16. 11. 15. 5. 13. 17. 4. 14. 3. 19. 20.
18. 2. 0. 1.]

[난자 해동 경과일]

결측치 수: 248624

[nan 0. 1.]

[난자 혼합 경과일]

결측치 수: 47444

[0. nan 3. 2. 1. 5. 6. 4. 7.]

[배아 이식 경과일]

결측치 수: 37275

[3. nan 2. 5. 1. 0. 4. 6. 7.]

[배아 해동 경과일]

결측치 수: 209691

[nan 0. 5. 3. 2. 6. 1. 4. 7.]

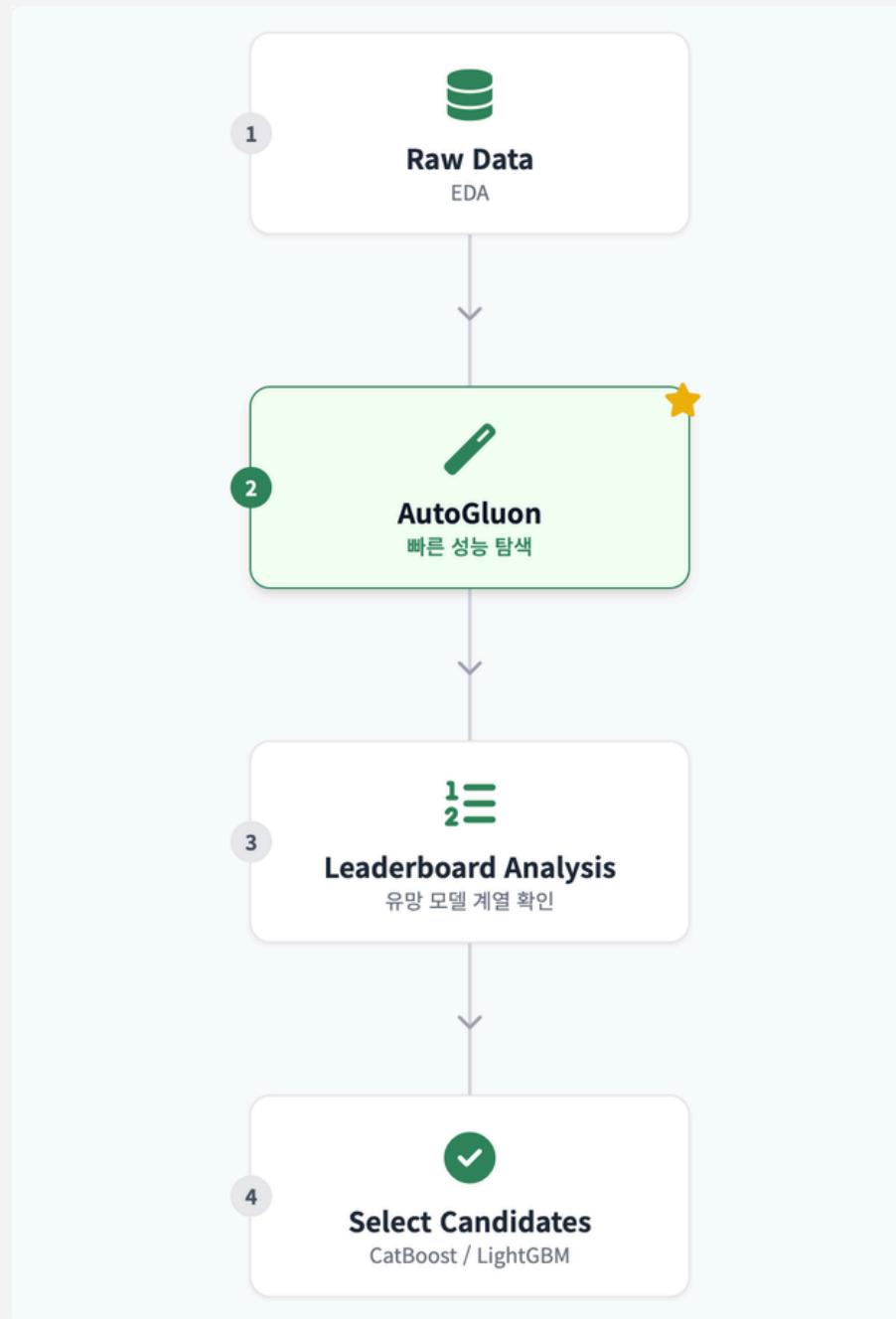
✓ 임신시도, 난자해동 경과일 칼럼 전처리

!!> 문제발견: 임신 시도 또는 마지막 임신 경과 연수, 난자 해동
경과일 칼럼 등 약 20만 개 nan 값 존재확인

unique로 데이터값 존재 유무 확인해보니 → 의미있는 숫자열
데이터 존재 → nan 값들은 대부분 기록되지 않은 값으로 판단 ->
0처리

모델 접근 전략

AutoML을 활용한 초기 탐색부터 최종 모델 선정까지



✓ 핵심 목표

단기간 내 성능 상한선(Upper Bound) 확인
데이터가 가진 최대 잠재력을 빠르게 파악

✓ 사용 도구

AutoGluon (best_quality)
복잡한 튜닝 없이 SOTA급 성능 기준점 확보

✓ 전환 전략

탐색은 AutoML로,, 최종 제출은 단일 모델로
복잡한 튜닝 없이 SOTA급 성능 기준점 확보
해석 가능성과 재현성을 위해 구조 단순화

AutoGluon 실험 설정

재현성과 성능의 균형을 맞춘 파이프라인 구성



presets = "best_quality"

가장 강력한 모델 조합 및 양상을 탐색



dynamic_stacking = False

데이터셋에 따른 가변적 구조 방지 → [안정성 확보](#)



time_limit = 12000

약 3.3시간 동안 충분한 모델 학습 시간 부여



Bagging 10 / Stacking 1

10-Fold Bagging으로 과적합 방지, 1-Level Stacking

experiment_pipeline.py

```
from autogluon.tabular import TabularPredictor

# 1. 평가 지표 및 문제 유형 정의
predictor = TabularPredictor(label='target', problem_type='binary', eval_metric='roc_auc')

# 2. 안정적인 학습을 위한 고정 파라미터 설정
predictor.fit(
    train_data=train_df,
    presets='best_quality',          # 최고 성능 모드
    time_limit=12000,                # 12,000초 (3시간 20분)
    dynamic_stacking=False,           # 구조 가변성 제한 (재현성 핵심)
    num_bag_folds=10, num_stack_levels=1, num_bag_sets=1,
    ag_args_fit={'num_cpus': 10, 'random_seed': 42, 'num_gpus': 0}
)
```



AutoGluon

리더보드 인사이트

Top 6 모델 성능 비교 및 단일 모델 가능성 확인



	model	score_val	eval_metric	pred_time_val	fit_time	pred_time_val_marginal	fit_time_marginal	stack_level	can_infer	fit_order
0	WeightedEnsemble_L3	0.740053	roc_auc	66.239602	8199.148921	0.030196	22.688951	3	True	33
1	WeightedEnsemble_L2	0.738983	roc_auc	19.128392	7435.501222	0.034054	13.437738	2	True	18
2	CatBoost_r9_BAG_L2	0.738595	roc_auc	41.213669	8088.372936	0.199237	143.855068	2	True	32
3	CatBoost_BAG_L2	0.738532	roc_auc	41.230330	8539.060117	0.215899	594.542248	2	True	23
4	CatBoost_r177_BAG_L2	0.738493	roc_auc	41.128280	8042.535508	0.113848	98.017640	2	True	29
5	LightGBM_r131_BAG_L2	0.738429	roc_auc	42.111127	8065.656666	1.096695	121.138798	2	True	31

왜 AutoGluon이 아닌 단일 모델을 선택했는가?

성능을 넘어서, 재현성과 해석 가능성을 고려한 전략적 선택



복잡한 구조

모델의 앙상블/스태킹으로 얹혀 있어

**개별 변수의 영향도 추적이
어려움.**



재현/통제 난이도

블랙박스 형태의 자동 튜닝으로 인해

**특정 파라미터 변경 시
동일 결과 재현이 어려움.**



설명 가능성 (XAI)

"왜 점수가 좋아졌는가?"에 대한
명확한 비즈니스 설명이 부족하여
설득력이 떨어짐.



전략적 결론

AutoML로 탐색하고, 최종 제출은 단일 모델로 학습

Single Model

단일 모델 학습: CatBoost / LightGBM

AutoML의 통찰을 바탕으로 직접 통제 가능한 파이프라인 구축

공통 전처리 파이프라인
ID 제거 · 타깃 분리 · 범주형 결측을 'Unknown'으로 통일

검증 전략 (Validation)
5-fold Stratified CV + OOF(Out-Of-Fold) AUC로 “일관된 성능” 확인

CatBoost

- AutoGluon 상위권 계열 → 파라미터 고정 후 재학습
- OOF AUC: **0.7381** (Best fold: 0.7396)

LightGBM

- AutoGluon 상위권 계열 → 규제/샘플링 포함해 안정화
- OOF AUC: **0.7381** (Best fold: 0.7402)

```
single_model_params.py

CatBoost params AUC
from catboost import CatBoostClassifier

cat_params = dict(
    loss_function="Logloss",
    eval_metric="AUC",

    iterations=6000,           # early stopping 기반
    learning_rate=0.03,
    depth=6,
    l2_leaf_reg=6,
    random_strength=1.0,
    bagging_temperature=0.7,

    # 불균형 보정(가볍게)
    class_weights=[1.0, 2.0], 

    # 학습 편의
    verbose=300,
    random_seed=42,
    allow_writing_files=False
)

LightGBM params AUC
import lightgbm as lgb

lgb_params = {
    # 학습 목표
    "objective": "binary",
    "metric": "auc",
    "boosting_type": "gbdt",

    # 학습 조절
    "learning_rate": 0.03,
    "n_estimators": 10000,      # early stopping
    "num_leaves": 64,
    "min_child_samples": 50,

    # 샘플링(과적합 완화)
    "subsample": 0.8,
    "subsample_freq": 1,
    "colsample_bytree": 0.8,

    # 규제
    "reg_lambda": 1.0,
    "reg_alpha": 0.0,
}
```

단일 모델 결과

CatBoost vs LightGBM 비교 후, LightGBM을 최종 제출 모델로 선택

CatBoost

```
# OOF / 테스트 예측 저장

oof_cb = np.zeros(len(X))
test_pred_cb = np.zeros(len(X_test))

for fold, (tr_idx, va_idx) in enumerate(skf.split(X, y), 1):
    X_tr, X_va = X.iloc[tr_idx], X.iloc[va_idx]
    y_tr, y_va = y.iloc[tr_idx], y.iloc[va_idx]

    model = CatBoostClassifier(**cat_params)

    model.fit(
        X_tr, y_tr,
        cat_features=cat_idx,
        eval_set=(X_va, y_va),
        use_best_model=True,
        early_stopping_rounds=300
    )

    oof_cb[va_idx] = model.predict_proba(X_va)[:, 1]
    auc = roc_auc_score(y_va, oof_cb[va_idx])
    cb_fold_scores.append(auc)

    print(f"[CatBoost] Fold {fold} AUC: {auc:.5f}")

    test_pred_cb += model.predict_proba(X_test)[:, 1] / skf.n_splits

print("CatBoost OOF AUC:", roc_auc_score(y, oof_cb))
print("CatBoost Folds:", cb_fold_scores)
```

CatBoost OOF AUC: 0.738139
CatBoost Best Fold AUC: 0.739594

3KS

LightGBM

```
import lightgbm as lgb

# 컬럼명 공백만 처리 (경고 방지)
X_lgb = X.copy()
X_test_lgb = X_test.copy()
X_lgb.columns = X_lgb.columns.str.replace(" ", "_")
X_test_lgb.columns = X_test_lgb.columns.str.replace(" ", "_")

# 범주형 dtype 맞추기
cat_cols_lgb = X_lgb.select_dtypes(include=["object", "category"]).columns
for c in cat_cols_lgb:
    X_lgb[c] = X_lgb[c].astype("category")
    X_test_lgb[c] = X_test_lgb[c].astype("category")

oof_lgb = np.zeros(len(X_lgb))
test_pred_lgb = np.zeros(len(X_test_lgb))

for fold, (tr_idx, va_idx) in enumerate(skf.split(X_lgb, y), 1):
    X_tr, X_va = X_lgb.iloc[tr_idx], X_lgb.iloc[va_idx]
    y_tr, y_va = y.iloc[tr_idx], y.iloc[va_idx]

    model = lgb.LGBMClassifier(**lgb_params)

    model.fit(
        X_tr, y_tr,
        eval_set=[(X_va, y_va)],
        eval_metric="auc",
        callbacks=[lgb.early_stopping(300)]
    )

    oof_lgb[va_idx] = model.predict_proba(X_va)[:, 1]
    auc = roc_auc_score(y_va, oof_lgb[va_idx])
    lgb_fold_scores.append(auc)

    print(f"[LGBM] Fold {fold} AUC: {auc:.5f}")

    test_pred_lgb += model.predict_proba(X_test_lgb)[:, 1] / skf.n_splits

print("LGBM OOF AUC:", roc_auc_score(y, oof_lgb))
print("LGBM Folds:", lgb_fold_scores)
```

LGBM OOF AUC: 0.7381228334281063
LGBM Best Fold AUC: 0.7401632766402618

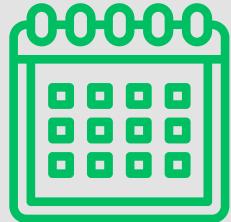
핵심 인사이트

문제발견 → 전처리 진행



단순 결측치 대체가 아닌, 변수 유형(수치형·범주형·명목형)에 맞는 전처리 전략을 설계하고 적용함. 의료 Tabular 데이터에서 전처리가 모델 성능에 큰 영향을 미친다는 점을 확인함.

하이퍼파라미터 최적화 자동화 적용

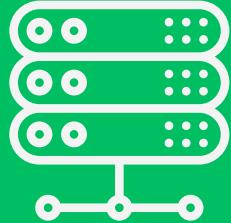


Optuna 또는 AutoML 기반 하이퍼파라미터 튜닝을 적용하여 모델 성능을 개선함. 수동 파라미터 조정보다 자동 탐색 방식이 효율적이며 재현성이 높음을 확인함.

협업 기반 모델 개발 프로세스 경험



팀 단위 프로젝트에서 실험 결과, 전처리 방법, 모델 성능을 지속적으로 공유하며 협업 개발을 진행함. 데이터 분석 프로젝트에서 커뮤니케이션이 개발 효율에 중요한 요소임을 체감함.



도메인 지식 기반 프로젝트

의료 데이터 특성을 반영한 파생변수를 설계하여 모델 성능을 향상시킴. 단순 모델링보다 도메인 이해가 예측 정확도 개선에 핵심적인 역할을 함을 확인함.