PNU UGRP 결과보고

[질환 예측 AI 모델링]

다낭성 난소 증후군 (PCOS)

다낭성 난소 증후군 (PCOS) 관련
질병 위험 예측을 위한 AI 모델링 및 생애 주기별 분석

부산대학교 의료인공지능전공

정보컴퓨터공학부 | 이진솔

목차

CONTENTS

1. 연구 주제 및 목적

TOPIC

[연구 주제]

다낭성 난소 증후군(PCOS) 관련 질병 위험 예측을 위한
AI 모델링 및 생애 주기별 분석

연구 목적

- ☑ 생애 주기별 PCOS 관련 질병의 주요 위험 요인을 분석
- ☑ AI 기반 모델을 사용하여 주요 질병(Obesity, Type2 Diabetes 등)의 위험도를 예측
- ☑ SHAP을 사용하여 모델의 해석 가능성을 확보

[연구 배경 및 목적]

다낭성 난소 증후군
Polycystic ovary syndrome

호르몬 불균형에 의해
불규칙한 월경, 다모증, 여드름
같은 증상이 나타나고
가임력에 영향을 미치는 질환

PCOS 관련 질환 예측

PCOS는 가임기 여성에게 흔히 발생하는 내분비 장애로,
비만, 당뇨병, 심혈관계 질환 및 비타민 D 결핍과 같은 여러 동반 질환과 밀접하게 관련되어 있습니다.
특히, 주변 20대 초중반 친구들에게도 생각보다 흔히 발견되는 질환이라는 점에서, 본 연구의 필요성이 더욱 강조되었습니다.

AI 모델링 및 생애 주기별 분석

본 연구는 AI 모델을 활용하여 생애 주기별
(초기 가임기, 중기 가임기, 후기 가임기, 폐경기)로 PCOS와 관련된 주요 질환의 위험도를 예측하고,
이를 통해 개인화된 예방 및 관리 방안을 제공하고자 합니다.

[연구 목표 및 내용]

목표
AIM

- 생애 주기(LIFE STAGE)를 기반으로 PCOS 관련 질병 위험 요인을 도출
- 비만, 제2형 당뇨병, 심혈관 질환, 비타민 D 결핍 등 주요 질환의 위험도 예측

내용
CONTENT

- 데이터 수집 및 전처리 (KAGGLE 데이터 활용)
- 주요 변수 정의 및 타겟 변수 생성
- AI 모델(RANDOM FOREST, XGBOOST, LIGHTGBM) 학습 및 평가
- SHAP을 활용한 변수 중요도 분석
- 생애 주기별 질병 위험도 시각화

참고 논문

- Askar, S. S., & Kumar, S. S. (2023). *Optimized Polycystic Ovarian Disease Prognosis and Classification Using AI-Based Models*
- Gupta, A., & Sharma, R. (2024). *Unveiling the Role of Artificial Intelligence (AI) in Polycystic Ovary Syndrome (PCOS) Management: A Comprehensive Review*
- Khan, M. E., & Rahman, S. A. (2023). *AI-Driven Detection of PCOS for Personalized Health Improvement*

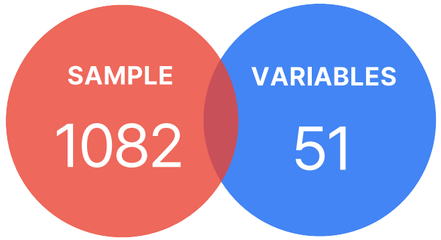
METHOD

[연구 방법론]

사용한 데이터

https://www.kaggle.com/code/jagatheeswari/pcos-dataset

- PCOS_data_without_infertility.xlsx
- PCOS_infertility.csv



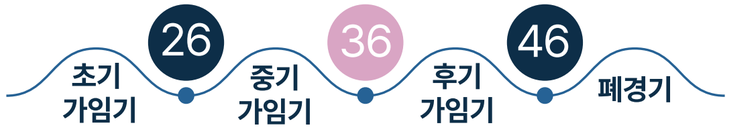
데이터 전처리

결측치
결측치가 있는 변수들을 제거하지 않고 대체 (mean, most_frequent) 방식으로 처리하여 데이터 손실 최소화.

생애 주기 변수 생성
나이(Age (yrs))를 기반으로 생애 주기를 정의.

범주형 데이터 인코딩
One-Hot Encoding을 사용하여 생애 주기 (Life Stage)를 인코딩.

Life Stage by Age

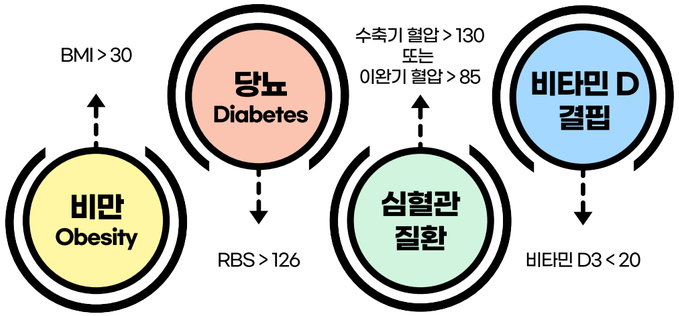


METHOD

[연구 방법론]

데이터 전처리

질환 타겟 변수 추가



데이터 모델링

사용된 모델

- Random Forest
- LightGBM
- XGBoost

평가 지표

- Precision, Recall, F1-Score, AUC-ROC

성능 평가 방법

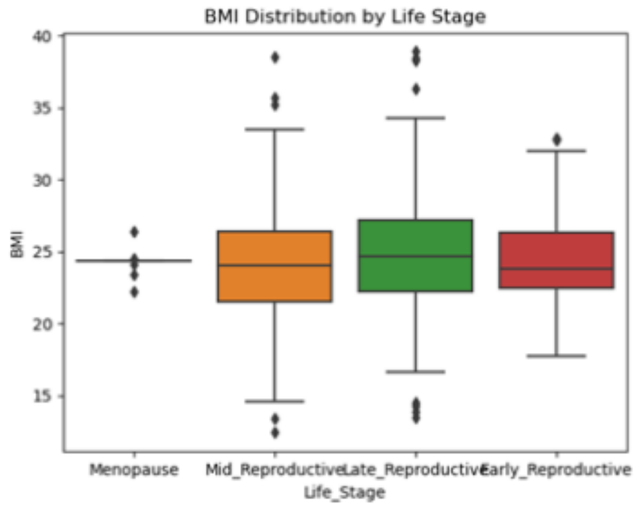
- 학습 데이터와 테스트 데이터를 80:20으로 분할
- Precision-Recall Curve를 통해 Decision Threshold 최적화
- SHAP 해석으로 변수 중요도를 도출

[모델 성능 비교]

MODEL	Obesity AUC	Type2 Diabetes AUC	Cardiovascular Risk AUC	Vitamin D Deficiency AUC
Random Forest	1.0	0.80	0.76	0.81
XGBoost	1.0	0.63	0.72	0.78
LightGBM	1.0	0.66	0.74	0.78

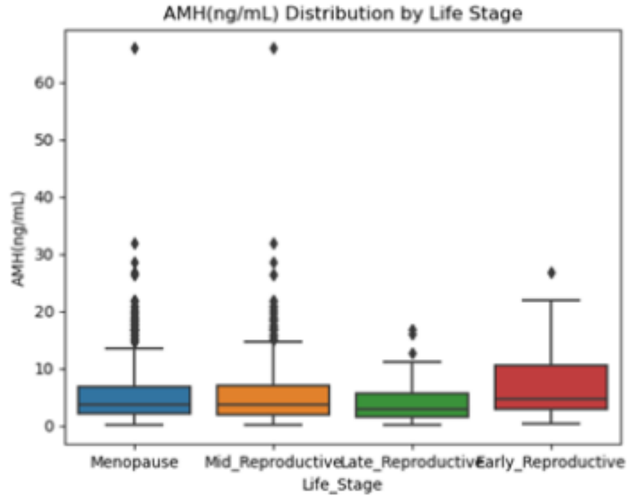
[생애 주기별 데이터 분석]

BMI 분포



BMI는 생애 주기별로 변동을 보였으며, 후기 가임기(LATE REPRODUCTIVE) 단계에서 높은 BMI 값을 나타냈습니다. 이는 이 단계에서 비만과 같은 대사 질환 위험이 증가할 가능성을 시사합니다.

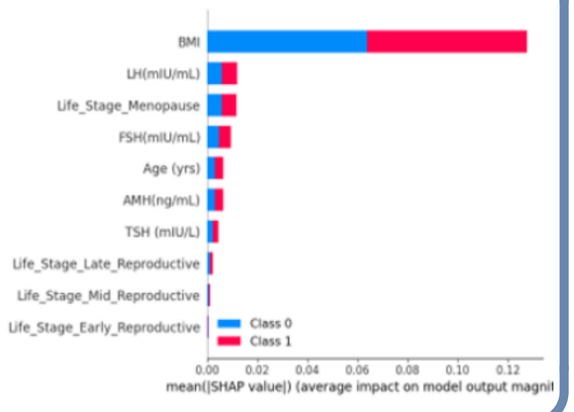
AMH 분포



AMH(NG/ML) 수치는 생애 주기 전반에 걸쳐 가임기 단계에서 높은 값을 보였습니다. 이는 생식 건강 상태를 나타내는 주요 지표로, 가임기 단계의 여성에서 중요한 의미를 가집니다.

[SHAP 변수 중요도]

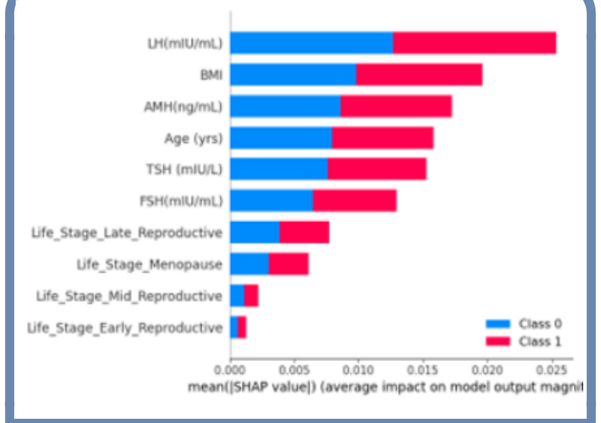
Obesity



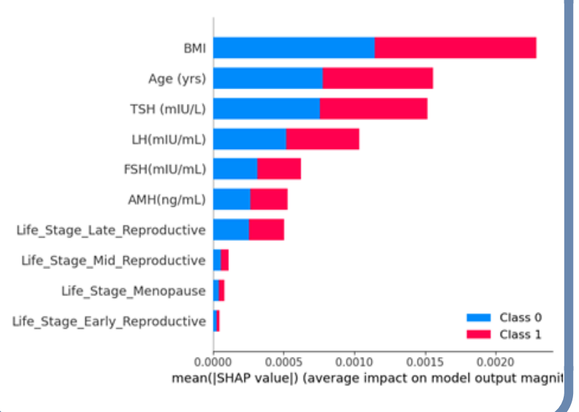
BMI는 비만 모델에서
가장 중요한 특징으로,
예측에 가장 큰 영향을 미칩니다.

Type2 Diabetes

LH와 BMI가 당뇨병 예측에서
중요한 특징으로 나타났습니다.



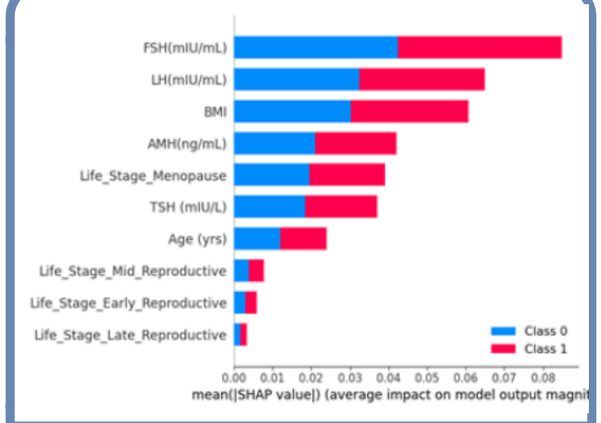
Cardiovascular Risk



BMI가 심혈관 위험 예측에
주요한 변수로 작용하며,
AGE와 TSH가 뒤를 잇습니다.

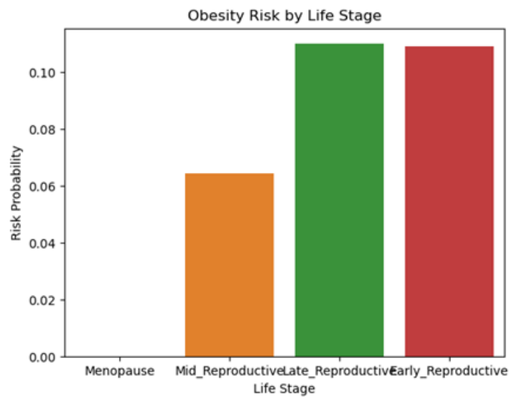
Vitamin D Deficiency

FSH와 LH가 비타민 D 결핍 예측
에서 중요한 변수로 나타났습니다.



[생애 주기별 위험도 시각화]

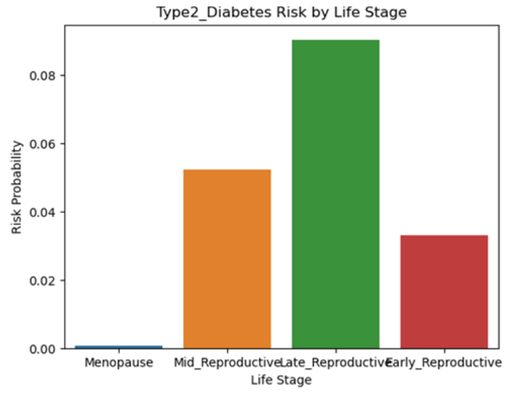
비만 위험도



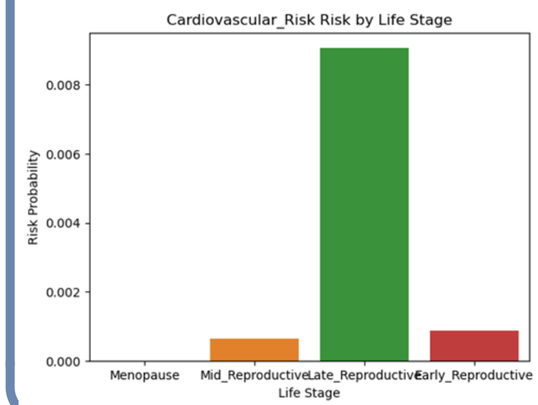
초기 가임기와 후기 가임기
단계에서 비만 위험도가
높게 나타났습니다.

제2형 당뇨병 위험도

후기 가임기 단계에서
제2형 당뇨병 위험도가
가장 높았습니다.



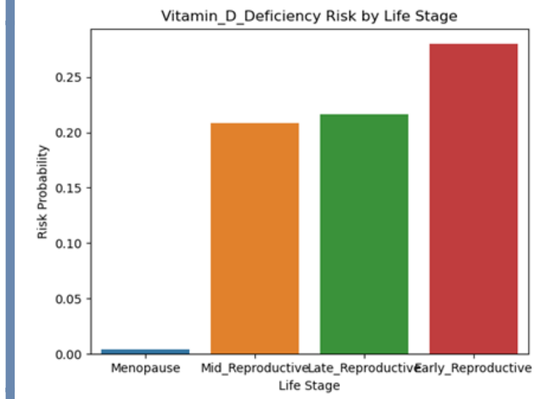
심혈관 질환 위험도



후기 가임기 단계에서
심혈관 질환 위험도가 두드러졌습니다.

비타민 D 결핍 위험도

초기 가임기 단계에서
비타민 D 결핍 위험도가
가장 높게 나타났습니다.



[기대효과 및 활용분야]

기대효과

- **생애 주기별 분석을 통해** 각 연령대에 적합한 **맞춤형 건강 관리 방안을 제공**할 수 있습니다. 이를 통해 개인별 건강 상태를 개선하고, 삶의 질을 높일 수 있습니다.
- **주요 질병의 발병 위험을 사전에 예측하고 적절한 관리 전략을 수립**함으로써 합병증 예방에 기여할 수 있습니다. 이를 통해 효율적인 예방과 치료가 가능해집니다.
- **의료 및 건강 IT 솔루션 개발의 기초 자료로 활용**할 수 있습니다 이를 통해 보다 개인화되고 효율적인 의료 서비스를 지원할 수 있습니다.

활용분야

- PCOS 관련 질병의 위험도를 예측하여 **연령대별 맞춤형 건강 관리 방안**을 제공하는 서비스 개발에 활용될 수 있습니다.
- 연구에서 개발된 AI 모델은 병원 및 연구 기관에서 **데이터 분석 및 환자 맞춤형 치료 설계**에 활용될 수 있습니다.
- 질병 위험이 높은 집단을 대상으로 한 **건강 캠페인**과 **교육 프로그램 설계**에 기여할 수 있습니다. 이를 통해 **공공 건강 정책 수립**이 보다 효과적으로 이루어질 수 있습니다.

[향후계획 (지속연구)]

01 연구확장
임상 데이터 기반 추가 모델 검증
대규모 데이터 확보로 예측 정확도 개선

02 데이터 증강
SMOTE를 활용하여 소수 클래스 성능 보완

03 새로운 연구방향
비타민 D 결핍 및 심혈관 위험에 대한 추가 연구
생애 주기별 맞춤형 건강 관리 프로그램 개발

CONCLUSION

[결론]

본 연구를 통해 생애 주기별로 PCOS가 미치는 영향을 분석하고
AI를 활용한 개인 맞춤형 질환 예측 모델을 개발하였습니다. 주요 결과는 다음과 같습니다.

01 초기 가임기 단계에서 비만 및 비타민 D 결핍 위험이 증가한다.

02 후기 가임기 단계에서 대사 질환 및 심혈관 질환 위험이 높다.

03 BMI, LH 등의 변수는 주요 예측 원인으로 확인되었다.

**이 결과는
개인 맞춤형 건강 관리와
질병 예방 전략 수립에
중요한 기초 자료로
활용될 수 있습니다.**

[연구 요약]

본 연구는 다낭성 난소 증후군(PCOS)과 관련된 주요 질병의 생애 주기별 위험 요인을 분석하고, AI 모델을 활용해 질병 예측과 개인 맞춤형 건강 관리 방안을 제시했습니다.
Kaggle 데이터셋을 활용하여 데이터 전처리, Random Forest, XGBoost, LightGBM 모델 학습, SHAP 분석을 통해 주요 변수와 성능을 도출했습니다.
연구 결과, 후기 가임기 (Late Reproductive) 단계에서 대사 및 심혈관 질환 위험이 높고, 초기 가임기 (Early Reproductive) 단계에서 비만과 비타민 D 결핍 위험이 증가함을 확인했습니다.
이 결과는 개인 맞춤형 건강 관리, 질병 예방 전략, 의료 데이터 분석, 공공 건강 정책 등 다양한 분야에 활용 가능하며, PCOS 관련 건강 문제 해결에 기여할 것으로 기대됩니다.

CONCLUSION

[한줄요약]

> PCOS로 힘들어하는 모든 이들이
> 나이에 따라 겪을 수 있는 추가 질환의 고통을 막고,
> 건강한 삶을 누릴 수 있도록.

[일부 코드]

데이터 전처리

```python
def define_life_stage(age):  1개의 사용 위치  mogId *
    if age < 26:
        return 'Early_Reproductive'
    elif age < 36:
        return 'Mid_Reproductive'
    elif age < 46:
        return 'Late_Reproductive'
    else:
        return 'Menopause'

def preprocess_data(data):  2개의 사용 위치  mogId *

    # 생애 주기 정의
    data['Life_Stage'] = data['Age (yrs)'].apply(define_life_stage)

    # 결측치 처리
    numeric_cols = ['BMI', 'FSH(mIU/mL)', 'AMH(ng/mL)', 'LH(mIU/mL)', 'RBS(mg/dl)', 'TSH (mIU/L)']
    numeric_imputer = SimpleImputer(strategy='mean')
    data[numeric_cols] = numeric_imputer.fit_transform(data[numeric_cols])

    # 범주형 데이터 처리
    encoder = OneHotEncoder(handle_unknown='ignore')
    life_stage_encoded = encoder.fit_transform(data[['Life_Stage']]).toarray()
    data = pd.concat( objs: [data, pd.DataFrame(life_stage_encoded)], axis=1)

    print("전처리 완료: 데이터 크기 ->", data.shape)
    return data
```

모델링

```python
X_train, X_test, y_train, y_test = train_test_split( *arrays: X, y, test_size=0.2, random_state=42)

models = {
    'RandomForest': MultiOutputClassifier(RandomForestClassifier(random_state=42)),
    'XGBoost': MultiOutputClassifier(XGBClassifier(use_label_encoder=False, eval_metric='logloss', random_state=42)),
    'LightGBM': MultiOutputClassifier(LGBMClassifier(random_state=42))
}

model_results = []

# 각 모델 학습 및 평가
for model_name, model in models.items():
    print(f"\n{model_name} 모델 학습 중...")
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)

    report = classification_report(y_test, y_pred, target_names=target_columns, zero_division=0, output_dict=True)
    auc_scores = []
    for i, col in enumerate(target_columns):
        try:
            auc = roc_auc_score(y_test[col], model.predict_proba(X_test)[i][:, 1])
            auc_scores.append(auc)
        except Exception as e:
            auc_scores.append(None)
            print(f"{model_name}의 {col} AUC-ROC 계산 중 오류 발생: {e}")

    model_results.append({
        'Model': model_name,
        'Classification Report': report,
        'AUC-ROC Scores': dict(zip(target_columns, auc_scores))
    })

    print(f"{model_name} 모델 학습 완료")
```

감사합니다


정보컴퓨터공학부 이진솔


purnsol1001@naver.com


GITHUB
https://github.com/mogld/ugrp_2024