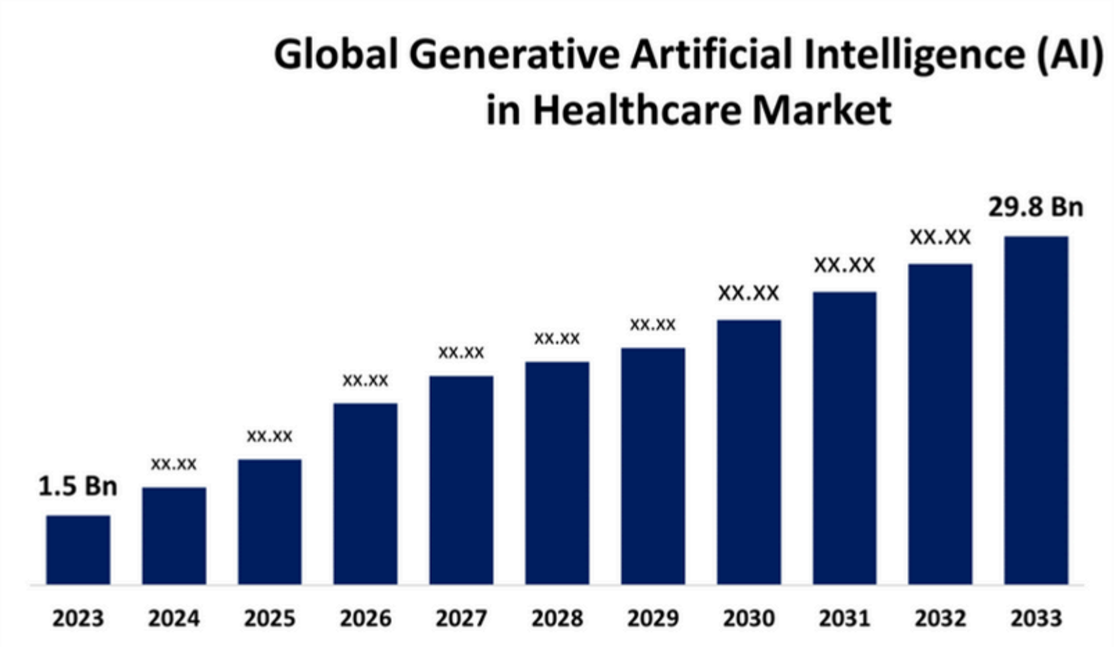
갑상선 암 진단
이진 분류 모델

Team : Mango (발표자 : 김종민, 김석민, 김수민)

Overview

프로젝트 배경 및 목표

- 최근 인공지능 의학의 관심이 크게 증가하고 머신 러닝을 이용한 예측 기술은 암 진단 및 치료의 혁신적인 미래 의료 기술로 기대되고 있다.
- 우리나라에서 새로 발생한 282,047건의 암 중 갑상선암은 33,914건으로 전체의 12.0%를 차지하며 가장 높은 발생률을 기록했다.



출처: sphericalinsights

순위	암종	발생자수	분율
	모든 악성암	282,047	1.000
1	갑상선	33,914	120
2	대장	33,158	118
3	폐	32,313	115
4	유방(5)	29,528	105
5	위(4)	29,487	105
6	전립선	20,754	74
7	간	14,913	53
8	췌장	9,78	35
9	담낭 및 기타담도	7,848	28
10	신장	6,963	25

출처: https://www.cancer.go.kr/index.do

프로젝트 배경 및 목표

- 갑상선 암은 조기에 발견해 치료할 경우 5년 생존율이 100%에 이를 만큼 예후가 좋은 암이지만, 기존 초음파 검사와 조직검사는 침습적이고 시간이 많이 소요되는 한계가 있어 보다 신속하고 비침습적인 조기 선별 방법에 대한 필요성이 커지고 있다.
- 본 연구는 제공된 데이터를 활용해 갑상선 암 악성·양성 분류 모델의 예측 정확도를 혁신적으로 개선하는 것을 목표로 한다.

Q&A로 풀어본 갑상선암

조기에 발견해 수술·치료하면 생존율 99%!

고대안암병원 갑상선센터 김경진 교수(내분비내과)

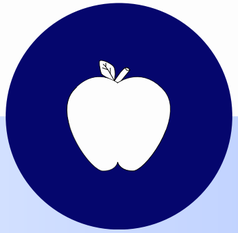
국내 암 중에서 증가율 1위 갑상선암. 7년 사이 2배 이상 환자가 늘었다. 다행히 다른 암에 비해 비교적 예후가 좋고, 조기에 발견해 치료하면 5년 생존율이 100%에 가깝다. 그러나 언제 나쁜 암으로 돌변할지 모르기 때문에 적절한 치료가 중요하다. 고대안암병원 갑상선센터 김경진 교수(내분비내과)와 함께 갑상선암에 대한 궁금증을 풀어봤다.

출처 : https://www.kumc.or.kr/seasonPress/KUMM_07/kumm34.jsp

기대 효과

서비스

접근성 향상
균등한 의료 서비스 제공
비용 절감



진단

정확도 향상
효율성 증대
오진 감소
시간 단축

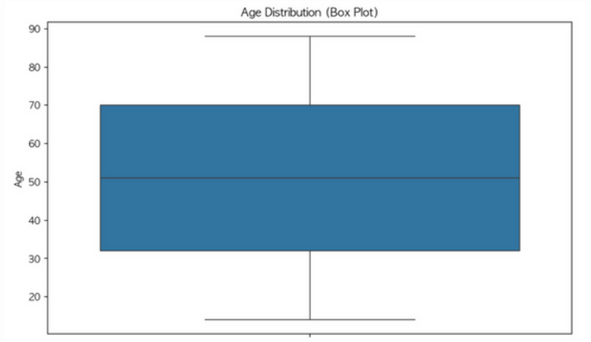
치료

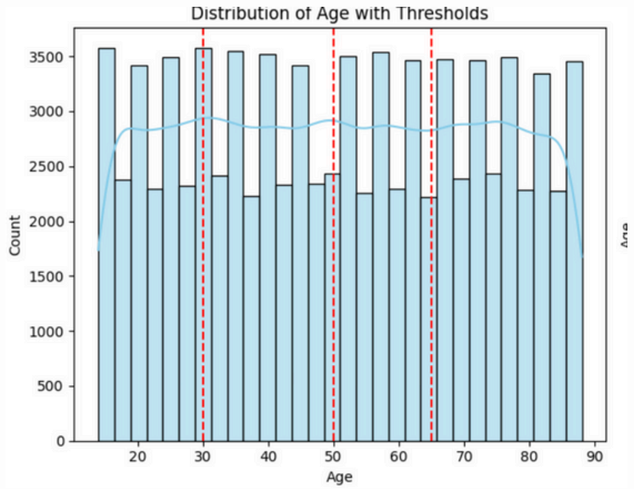
빠른 치료 방향성
완치율 향상

데이터 수집 및 전처리

- 수치형 데이터를 범주형 데이터로 변환

변수명	설명	데이터 타입
ID	샘플별 고유 ID	object
Age	환자의 나이	int64
Gender	성별	object
Country	국적	object
Race	인종	object
Family_Background	가족력 여부	object
Radiation_History	방사선 노출 이력	object
Iodine_Deficiency	요오드 결핍 여부	object
Smoke	흡연 여부	object
Weight_Risk	체중 관련 위험도	object
Diabetes	당뇨병 여부	object
Nodule_Size	갑상선 결절 크기	float64
TSH_Result	TSH 호르몬 검사 결과	float64
T4_Result	T4 호르몬 검사 결과	float64
T3_Result	T3 호르몬 검사 결과	float64
Cancer	갑상선암 여부 (0 : 양성, 1 : 악성)	int64





수치형 데이터 분포 확인 & 이상치 확인

```python
class add_new_feature(BaseEstimator, TransformerMixin):
    def t4_category(self, x):
        if x < 6:
            return 'T4_Low'
        elif x > 11.98:
            return 'T4_High'
        else:
            return 'T4_Normal'
    def t3_category(self, x):
        if x < 1.4:
            return 'T3_Low'
        elif x > 3:
            return 'T3_High'
        else:
            return 'T3_Normal'
    def tsh_category(self, x):
        if x < 0.27:
            return 'TSH_Low'
        elif x > 4.2:
            return 'TSH_High'
        else:
            return 'TSH_Normal'

    def age_category(self, x):
        if x < 30:
            return 'Young'
        elif x < 50:
            return 'Middle'
        elif x < 65:
            return 'Senior'
        else:
            return 'Elderly'

    def nodule_size_category(self, x):
        if x < 1.0:
            return 'Small'
        elif x < 2.0:
            return 'Medium'
        elif x < 4.0:
            return 'Large'
        else:
            return 'VeryLarge'
```

데이터 분석

- 트리 기반 모델을 활용해 각 특성의 중요도를 산출.
- 중요도가 낮은 하위 특성들은 모델 성능 향상과 과적합 방지를 위해 제거.

```
# 결과 출력
print(grouped_importance)
```

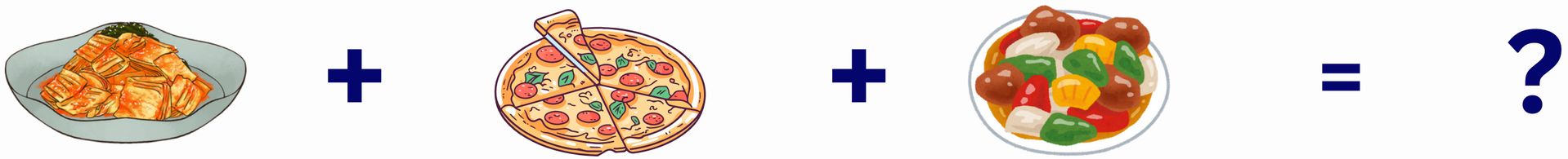
	base_feature	importance
0	Country	0.314868
1	Race	0.145911
2	Family_Background	0.112343
3	Weight_Risk	0.095491
4	Smoke	0.091877
5	Diabetes	0.088775
6	Iodine_Deficiency	0.072213
7	Radiation_History	0.038487
8	Gender	0.026066
9	Age	0.003983
10	T4_Result	0.002553
11	TSH_Result	0.002543
12	T3_Result	0.002472
13	Nodule_Size	0.002420

```
98  # 파이프라인 구성
99  useless_feature = ['T3_Result', 'T4_Result', 'TSH_Result',
100                    'Age', 'Nodule_Size',
101                    'T3_Cat', 'T4_Cat', 'T3_Result_Cat']
102 preprocessor = Pipeline([
103     ('add_new_feature', add_new_feature()),
104     ('dropper', dropper(useless_feature)),
105     ('encoder', custom_encoder())
106 ])
```

데이터 분석

- **특성 공학**
  - 모델의 성능 향상을 위해 데이터를 가공&변환하는 과정
  - 보통 기존의 변수들을 변형, 선택, 조합하는 과정

예시:



데이터 분석

- 3가지 파생 변수로 만들어진 3가지 가설

1. T3 및 T4 호르몬 수치는 갑상선 암 발병과 관련이 있다.

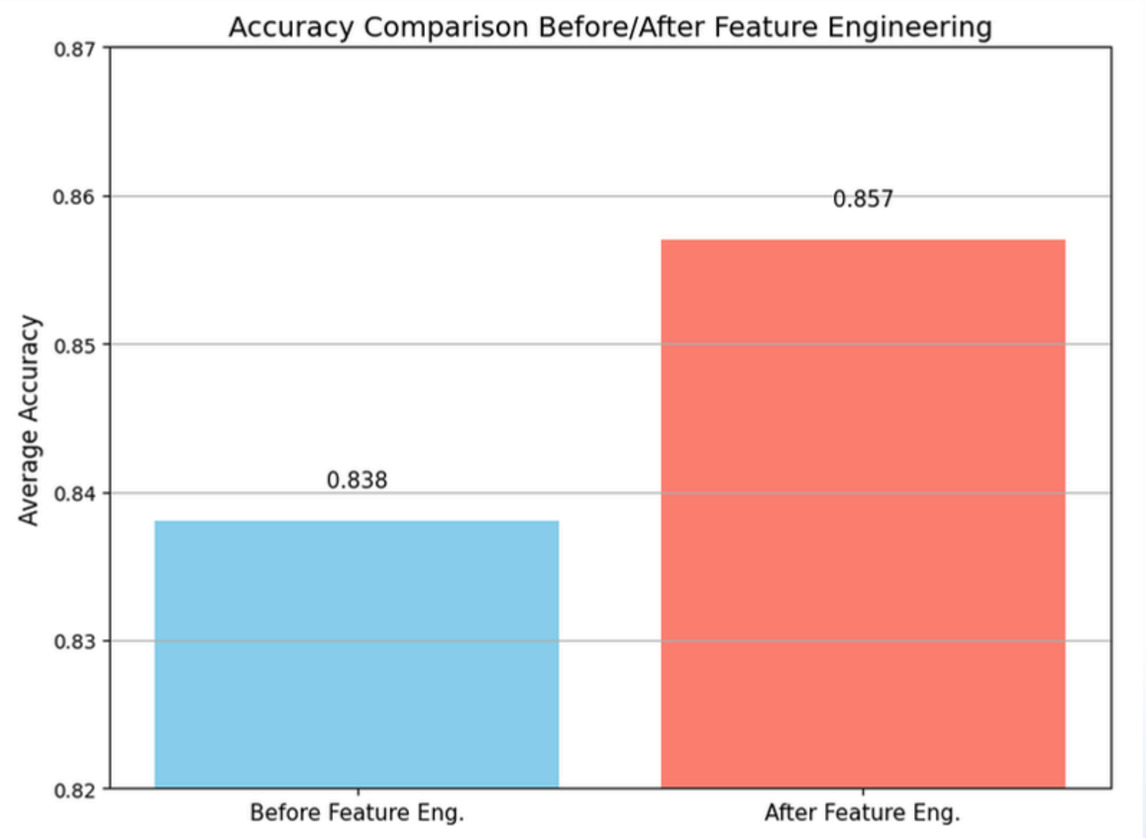
2. 인종과 T3호르몬 수치는 갑상선 암 발병과 관련이 있다.

3. 가족력 여부와 요오드 결핍 여부는 갑상선 암 발병과 관련이 있다.

결과와 해석

모델 학습 및 교차 검증 (10-fold)

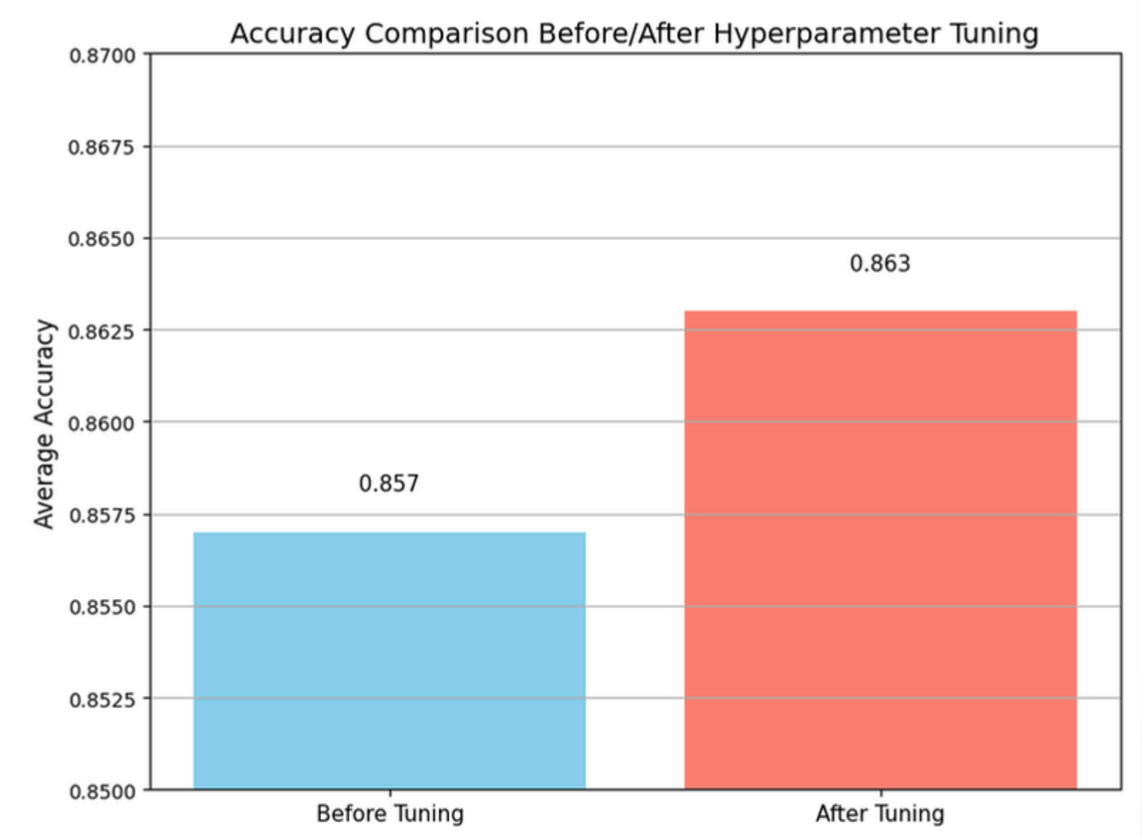
- StratifiedKFold를 이용해 클래스 비율을 일정하게 유지하면서 교차검증을 시행
- 각 fold를 검증셋으로 사용하여 파생 특성 추가 전/후의 성능을 기록
- 파생 특성 추가 전/후의 성능에 대해 대응 표본 t-검정을 시행
- 1.9%p 의 성능 향상 확인



결과와 해석

튜닝 전과 튜닝 후의 모델 성능 비교

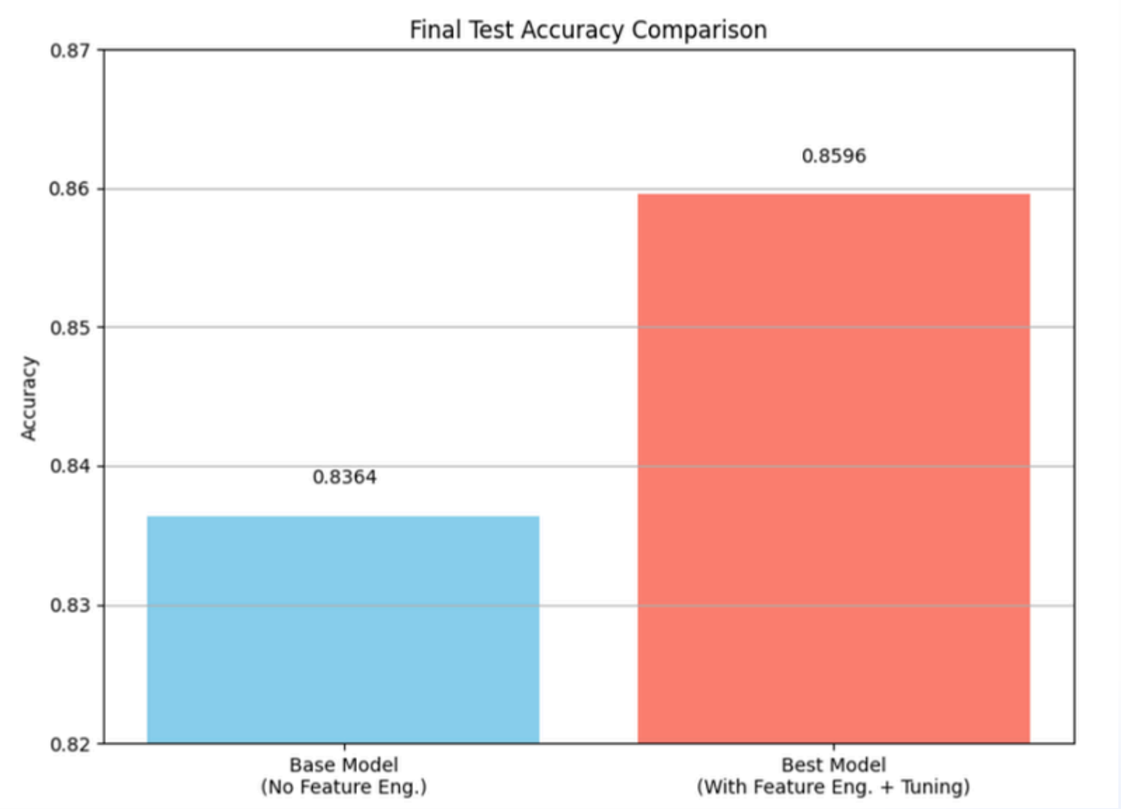
- GridSearchCV를 통해 최적의 파라미터를 탐색
- 튜닝 전과 비교하여 튜닝 후에 0.6%p 성능 향상
- 대응표본 t-검정을 시행하고 통계적으로 유의미함을 확인함



결과와 해석

베이스 모델과 최적화된 모델의 성능 차이

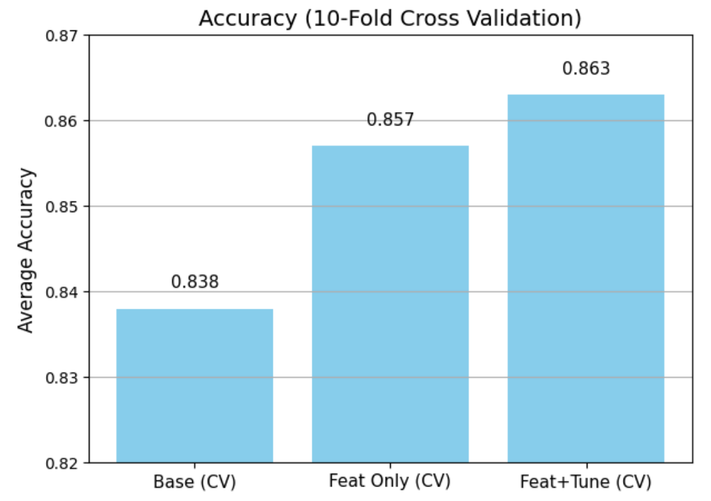
- 최종적으로 전체 train셋으로 학습하고 test 셋으로 성능 확인
- 2.32%p의 정확도 향상을 확인



결과와 해석

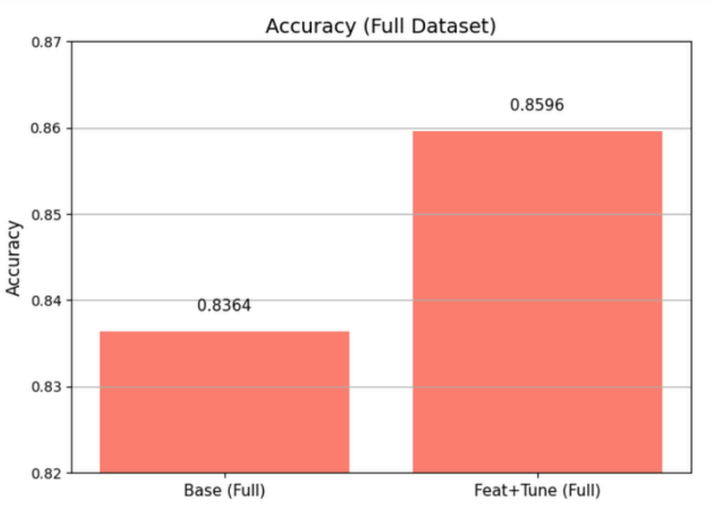
Ten-Fold Cross Validation

- 베이스 모델 정확도: 0.838 ± 0.006
- 파생변수 추가 후 튜닝 전 모델 정확도: 0.857 ± 0.004
- 파생변수 추가 후 튜닝 후 모델 정확도: 0.863 ± 0.004



전체 데이터셋을 이용한 성능 평가

- 베이스 모델 정확도 : 0.8364
- 파생변수 추가 후 튜닝 후 모델 정확도: 0.8596



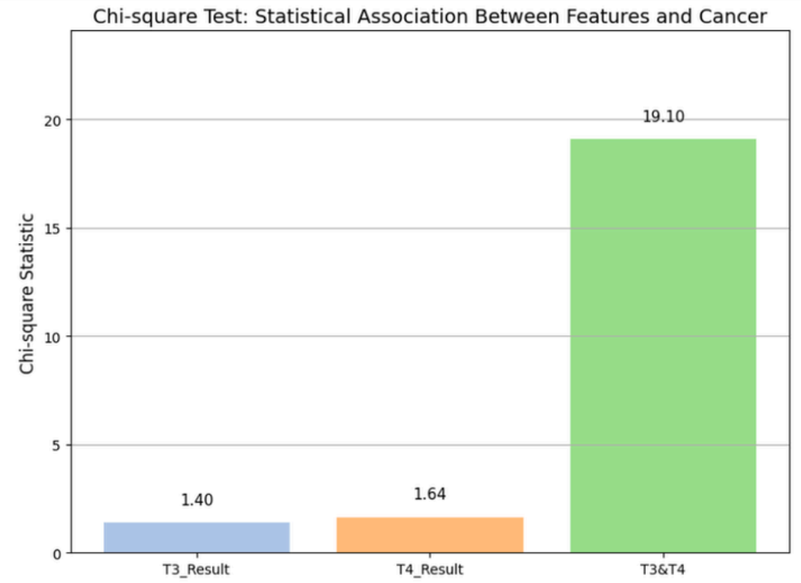
결과와 해석

1. T3 및 T4 호르몬 수치는 갑상선 암 발병에는 관련이 없다.

[가설 1-1] T3호르몬 수치와 갑상선 암 발병에는 관련이 없다
카이제곱 통계량: 1.3974 | p-value: 0.4972

[가설 1-2] T4호르몬 수치와 갑상선 암 발병에는 관련이 없다
카이제곱 통계량: 1.6389 | p-value: 0.4407

[가설 1-3] T3 및 T4 호르몬 수치는 갑상선 암 발병에는 관련이 없다
카이제곱 통계량: 19.1033 | p-value: 0.0143

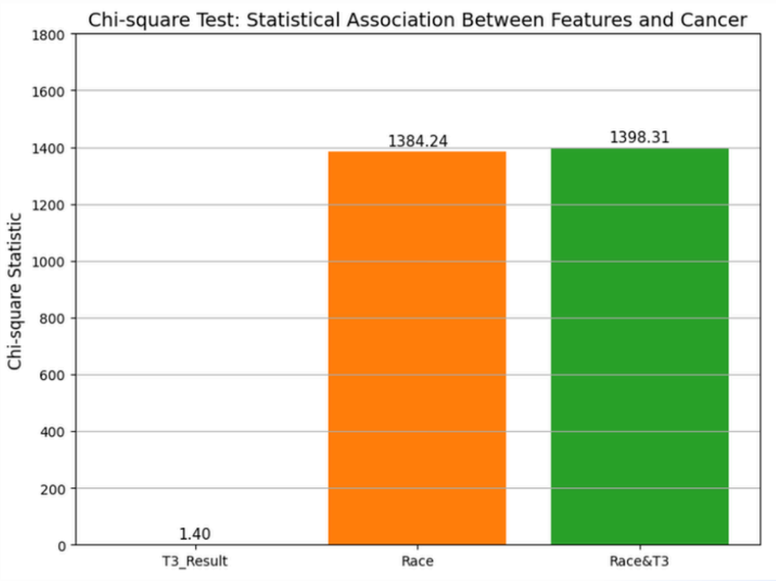


2. 인종과 T3호르몬 수치는 갑상선 암 발병에 관련이 없다

[가설 2- 1] T3호르몬 수치와 갑상선 암 발병에는 관련이 없다
카이제곱 통계량: 1.3974 | p-value: 0.4972

[가설 2-2] 인종과 갑상선 암 발병에는 관련이 없다
카이제곱 통계량: 1384.2394 | p-value: 0.000

[가설 2-3] 인종과 T3호르몬 수치는 갑상선 암 발병에 관련이 없다
카이제곱 통계량: 1398.3145 | p-value: 0.000



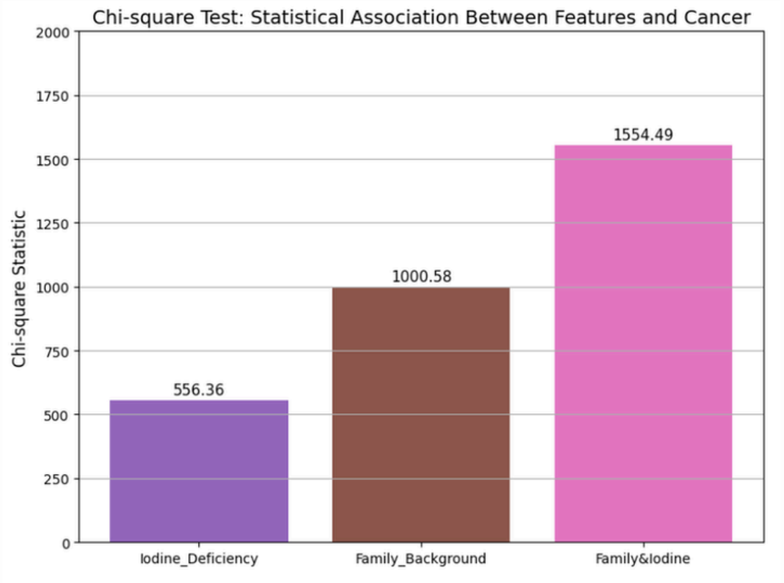
결과와 해석

3. 가족력과 요오드 결핍은 갑상선 암 발병에는 관련이 없다

[가설 3-1] 요오드 결핍과 갑상선 암 발병에는 관련이 없다
카이제곱 통계량: 556.3569 | p-value: 0.000

[가설 3-2] 가족력과 갑상선 암 발병에는 관련이 없다
카이제곱 통계량: 1000.5772 | p-value: 0.000

[가설 3-3] 가족력과 요오드 결핍은 갑상선 암 발병에는 관련이 없다
카이제곱 통계량: 1554.4941 | p-value: 0.000



시사점

- 독립적으로 특성을 고려했을 때보다 특성 간 상호작용을 반영한 모델이 더 높은 예측 성능을 보였다는 점에서, 변수 간 관계를 고려한 접근이 모델 성능 향상에 중요하다.

- 초음파나 조직검사 없이 간단한 임상 정보와 혈액검사 결과만으로 갑상선암 위험을 빠르고 부담 없이 예측할 수 있다.

- 이번 프로젝트에서는 두 개의 특성을 조합하여 하나의 파생변수를 생성하고 예측 성능을 높였다. 이를 바탕으로, 세 개 이상의 더욱 다양한 특성들을 조합하여 고차원 파생변수를 생성하고 모델의 예측 성능을 높일 가능성이 있으므로 특성 공학에 대한 연구 가치는 매우 높다.

역할 분담

LEADER

김종민

가설 검증
데이터 분석
회의록 작성
최종 PPT 제작 및 발표

홍종효

가설 검증
데이터 시각화
보고서 작성
제안 PPT 제작 및 발표

김석민

가설 검증
데이터 분석
회의록 및 보고서 검토
최종 PPT 제작 및 발표

역할 분담

김수민

가설 검증
데이터 분석
PPT 검토
최종 PPT 제작 및 발표

윤세혁

가설 검증
데이터 분석 및 시각화, 전처리
자료조사
제안 PPT 제작 및 발표

프로젝트 회고

의학 도메인에 대한 배경 지식이 부족하여 결과를 해석하는데 한계가 있었지만 통계적 검정을 바탕으로 유의미한 파생변수를 도출하고, 이를 통해 모델의 성능을 향상시킬 수 있다는 점에서 의미 있는 경험이었습니다.

모델 학습 과정에서 클래스 불균형 문제가 존재했고, 이를 해결하기 위한 다양한 방법을 조사한 결과로 SMOTE 기법을 적용하고 앙상블 학습 기반의 랜덤 포레스트 모델을 사용하여 데이터의 균형을 맞추고 모델의 일반화 성능을 향상시키는 노력을 하였습니다.

제안 발표를 마치고 “이미 다양한 통계들이 존재하는데 이 주제가 무슨 의미를 가지는가”라는 질문을 받고 프로젝트에서 정말 중요한건 단순히 데이터를 분석하는 것보다, 문제 정의와 방향 설정이라는 것을 깨닫는 계기가 되었습니다.

Thank's For Watching