위스콘신 유방암 예측

**위스콘신 유방암 예측**

**프로젝트 개요**

본 프로젝트는 위스콘신 유방암 진단 데이터셋(Wisconsin Breast Cancer Dataset)을 활용하여 다양한 머신러닝 분류 모델을 비교 분석하고, 전처리 및 하이퍼파라미터 튜닝을 통해 최적의 모델을 선정하는 것을 목표로 합니다. 모델의 성능은 정확도(Accuracy), F1 Score, ROC AUC 등을 기준으로 평가하며, 최종적으로 가장 우수한 성능을 보이는 모델을 도출합니다.

**데이터셋 개요**

* 출처 : UCI 머신러닝 저장소
* 샘플 수 : 569건
* 피처 수 : 30개 (모두 수치형)
* 타겟 변수 : Diagnosis (M=Malignant, B=Benign)
* 클래스 분포 : M (악성) 212건, B (양성) 357건 → 약간의 불균형 존재

**탐색적 데이터 분석 (EDA)**

* 기초 통계 분석
* describe() 메서드로 평균, 표준편차, 최대/최소값 확인
* 일부 피처는 큰 분산과 왜도를 가짐
* 클래스 분포 확인
* B: 63%, M: 37%로 클래스 불균형 존재
* 상관계수 분석
* heatmap으로 변수 간 상관관계 시각화
* 상관계수 높은 변수들 다수 확인됨(공선성 원인)

**데이터 전처리**

* 타겟, 피처 분리
* 타겟 변수는 Diagnosis
* 피처는 30개 수치형 변수(ID 및 기타 불필요한 변수 제외)
* 다중공선성 제거
* 상관계수가 0.9이상인 변수들 중 일부 제거
* 로그 변환
* 왜도가 높은 피처 변환 적용
* 표준화(StandardScaler)
* 전체 피처를 평균 0, 표준편차 1로 스케일링
* 학습, 테스트 데이터 분리
* train\_test\_split(8:2) 비율 사용, random\_state=42고정

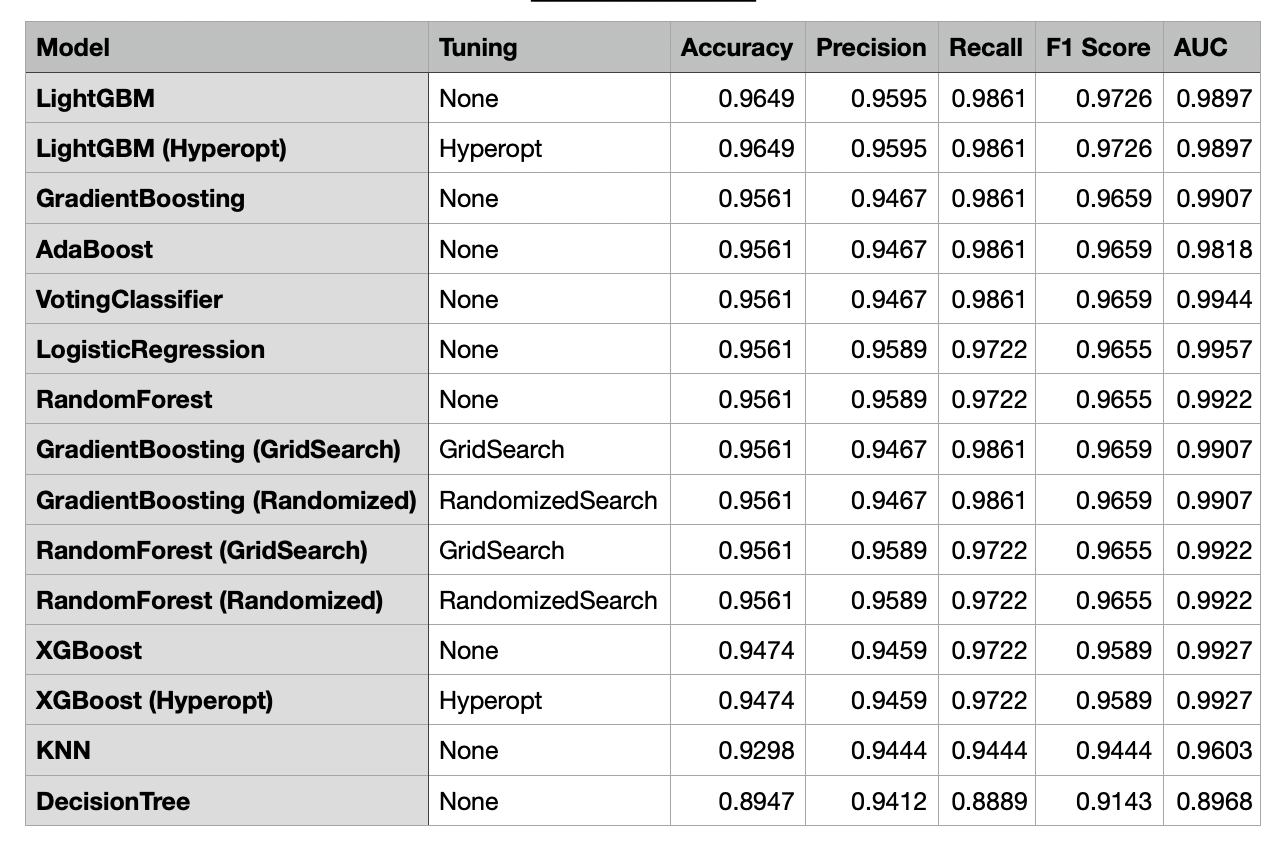
**모델 학습 및 평가**

* 사용한 모델 목록
* 로지스틱 회귀(Logistic Regression)
* KNN
* 랜덤포레스트(Random Forest)
* 그래디언트 부스팅(Gradient Boosting)
* 의사결정트리(Decision Tree)
* 에이다부스트(AdaBoost)
* XGBoost
* LightGBM
* 앙상블(Voting)
* 소프트 보팅 방식 적용
* 구성 모델 : 로지스틱 회귀, 랜덤포레스트, XGBoost, LightGBM
* 성능 비교
* 가장 높은 정확도 : LightGBM(Accuracy : 0.9649, F1 Score : 0.9726)
* 높은 AUC : XGBoost(0.9927)
* 낮은 성능 : KNN, Decision Tree

**하이퍼파라미터 튜닝**

* 적용 모델 및 방식
* LightGBM : Hyperopt
* RandomForest : GridSearchCV & RandomizedSearchCV
* GradientBoosting : GridSearchCV & RandomizedSearchCV
* XGBoost: Hyperopt
* 튜닝 후 성능 변화 요약
* LightGBM : 성능 변화 없음 (기본 모델이 이미 우수함)
* RandomForest, GradientBoosting : 튜닝 효과 미미
* XGBoost: 정확도와 F1 Score 향상 (튜닝 전 대비 개선)

**전체 모델 성능 비교**



**결론**

* LightGBM은 튜닝 전후 모두 가장 뛰어난 성능을 보였으며 실무에서도 바로 적용 가능한 강력한 분류기임을 확인했습니다.
* XGBoost는 정확도는 다소 낮지만 AUC가 높아 민감도 기반 판단이 필요한 분야에 적합합니다.
* GradientBoosting, RandomForest는 성능이 안정적이지만 튜닝에 따른 성능 향상은 미미했습니다.
* Voting 모델은 균형 잡힌 성능을 제공하며 여러 모델의 장점을 결합한 점에서 실무 활용성이 높습니다.
* 다양한 모델과 튜닝 전략을 통해 실전 문제 해결에 유용한 경험을 얻었으며, 성능 평가는 Accuracy 외에도 F1 Score, AUC 등 다양한 지표를 함께 고려해야 함을 체감했습니다.

# **프로젝트 소개**

## 1.1 과제 배경 및 목적

본 프로젝트는 위스콘신 유방암 데이터셋을 활용해 양성(Benign)과 악성(Malignant) 종양을 머신러닝 모델을 통해 분류하고 다양한 모델 및 전처리 기법, 하이퍼파라미터 튜닝을 적용해 모델별 성능을 비교 분석하는 것을 목적으로 합니다.

단순히 하나의 모델을 최적화하는 것이 아닌 다양한 모델과 튜닝 전략을 적용해 어떤 조합이 가장 높은 정확도를 달성하는지 파악하고자 합니다.

## 1.2 데이터 셋 소개 (Wisconsin Breast Cancer Dataset)

사용된 데이터는 UCI 머신러닝 저장소에서 제공하는 Wisconsin Breast Cancer Diagnostic 데이터셋입니다.

총 569개의 샘플로 구성되어 있으며 각각은 환자의 유방 종양에 대한 30개의 수치형 특성(피처)을 포함하고 있습니다. 예를 들어 종양의 크기, 경계의 부드러움, 밀도, 대칭성 등의 정보가 포함됩니다.

각 샘플은 양성(B) 또는 악성(M)으로 라벨링 되어있으며 이진 분류 문제는 종양이 악성인지 아닌지를 정확하게 예측하는 것이 목표입니다.

* 피처 수 : 30개 (모두 연속형 수치)
* 타겟 변수 : Diagnosis(M=Malignant, B=Benign)
* 샘플 수 : 569건
* 클래스 비율 : 악성(B) 약 37%, 양성(M) 약 63% → 조금 불균형한 데이터셋

## 1.3 문제 정의 및 성능 평가 기준

이 프로젝트는 이진 분류 문제로 정의됩니다. 즉, 주어진 종양 특성 데이터를 기반으로 해당 샘플이 양성인지 악성인지 예측하는 것이 목표입니다.

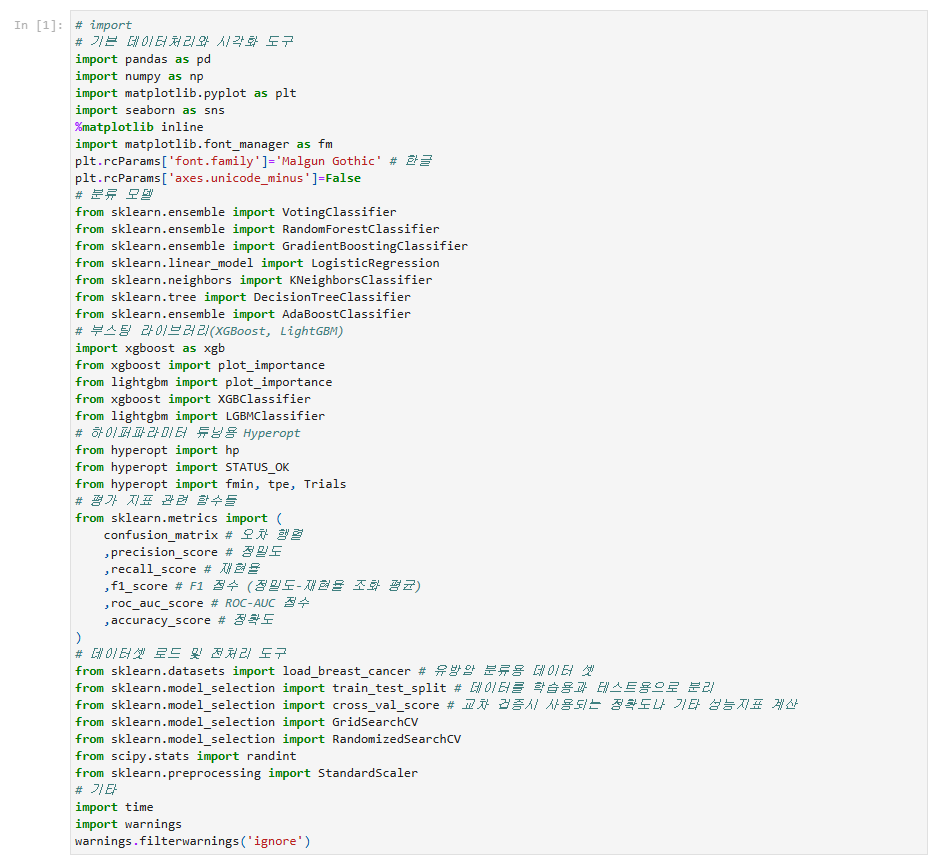
모델의 성능을 평가하기 위해 다음과 같은 지표를 사용했습니다 :

* Accuracy(정확도) : 전체 예측 중 정답 비율. 직관적이고 가장 기본적인 지표
* Precision(정밀도) : 양성으로 예측한 것 중 실제 양성 비율 (False Positive 방지 중요할 때 사용)
* Recall(재현율) : 실제 양성 중 모델이 양성으로 맞춘 비율 (False Negative 방지 중요할 때 사용)
* F1 Score : 정밀도와 재현율의 조화 평균. 불균형 데이터셋에서 균형 잡힌 평가 제공
* ROC AUC : 다양한 임계값에 대한 분류 성능을 종합적으로 평가

본 프로젝트의 주요 비교 기준은 **정확도(Accuracy)**이며 부가적으로 F1 Score와 AUC도 함께 비교하였습니다.

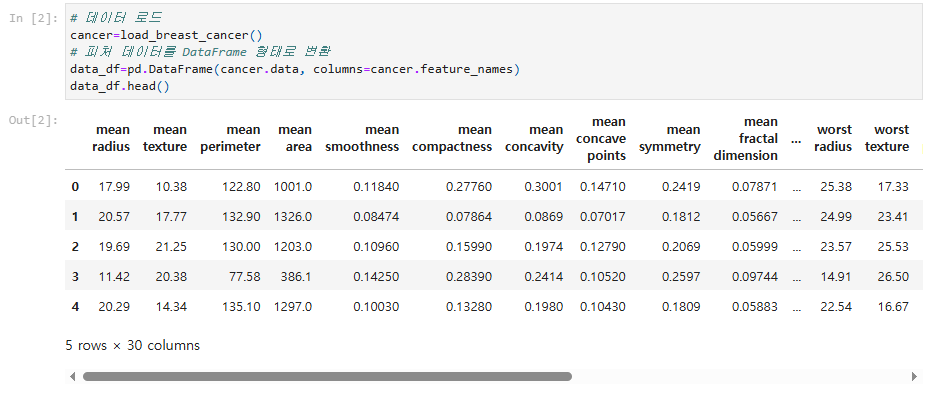
이는 모델의 일반적인 성능 수준을 파악하고 다양한 튜닝과 전처리 전략이 얼마나 정확도 개선에 기여했는지를 중심으로 평가하기 위함입니다.

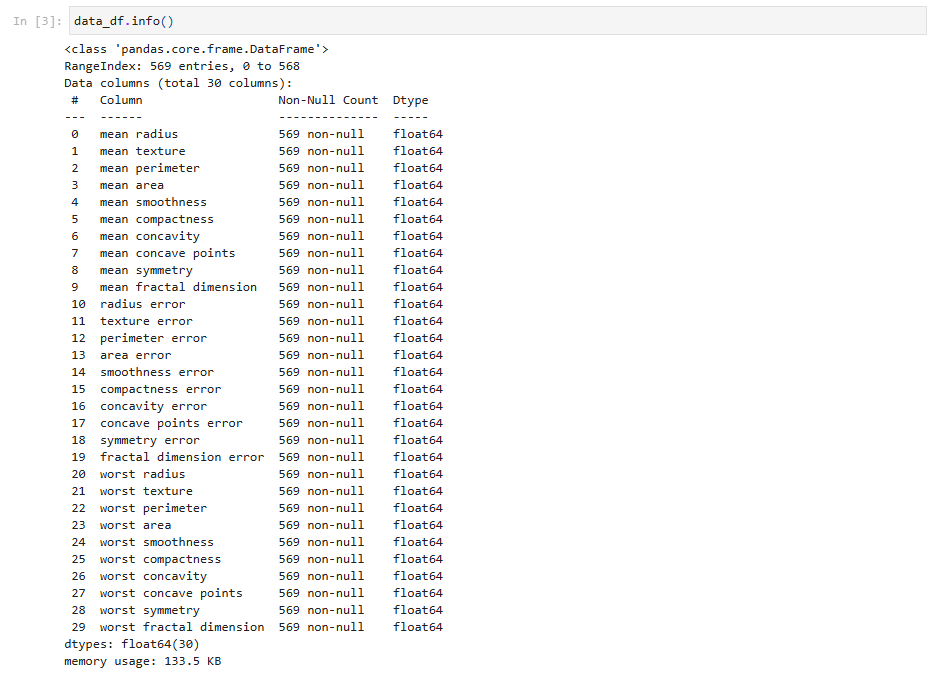
# **데이터 탐색 및 분석(EDA)**



## 2.1 데이터 로딩 및 구조 확인

load\_breast\_cancer() 함수를 통해 위스콘신 유방암 데이터를 로드하고 이를 pandas DataFrame 형태로 변환하여 분석에 활용하였습니다. 각 컬럼의 이름, 데이터 타입, 결측치 유무, 기본적인 구조(shape) 등을 확인하고 데이터가 분석 가능한 형태로 잘 구성되어 있음을 확인했습니다.





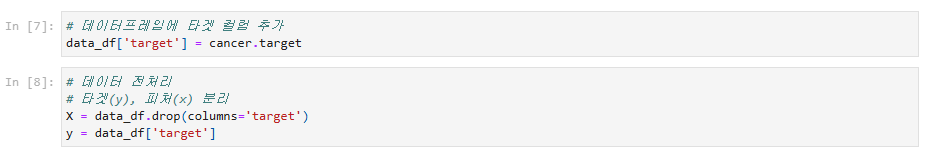
*# RangeIndex: 569 entries, 0 to 568 = 569개 샘플(행)*

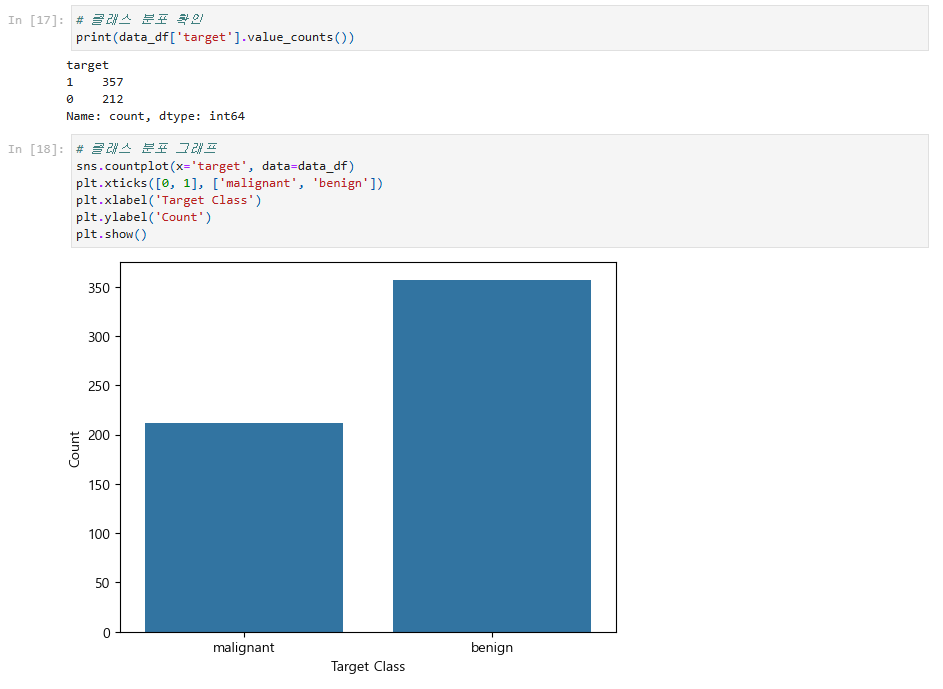
*# Data columns (total 30 columns): = 30개 특성(피처)*

*# non-null = 결측값 없음 (전부 569개로 동일)*

## 2.2 타겟(label) 클래스 분포 확인

타겟 변수인 'diagnosis' 컬럼은 M(Malignant, 악성), B(Benign, 양성) 두 개의 클래스로 구성되어 있으며 value\_counts() 및 시각화를 통해 클래스 불균형이 존재함을 확인했습니다.





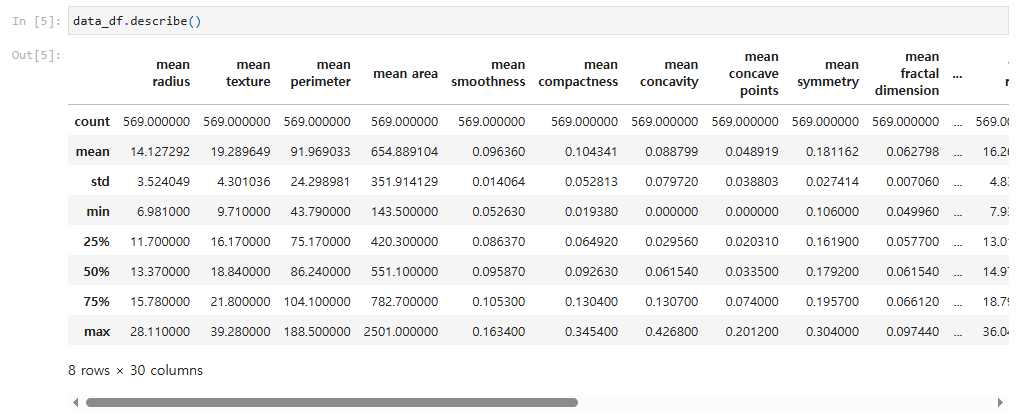
*# 0 : malignant(악성종양) = 212개*

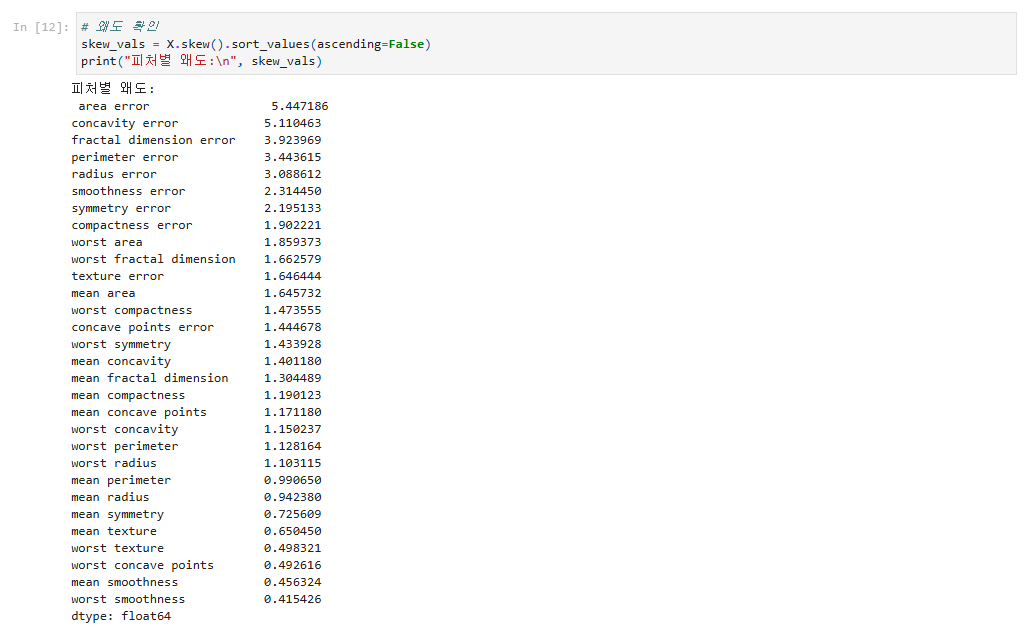
*# 1 : benign(양성종양) = 357개*

*# 양성 샘플이 더 많음 -> 클래스 살짝 불균형*

## 2.3 기초 통계 분석

describe() 함수를 통해 각 피처의 평균, 표준편차, 최소값, 최대값 등의 기초 통계량을 확인하였습니다. 대부분의 피처는 연속형 변수이며 값의 스케일이 서로 다르기 때문에 향후 정규화(StandardScaler)가 필요함을 인식했습니다. 또한 일부 피처의 왜도가 크다는 것도 함께 확인되었습니다.

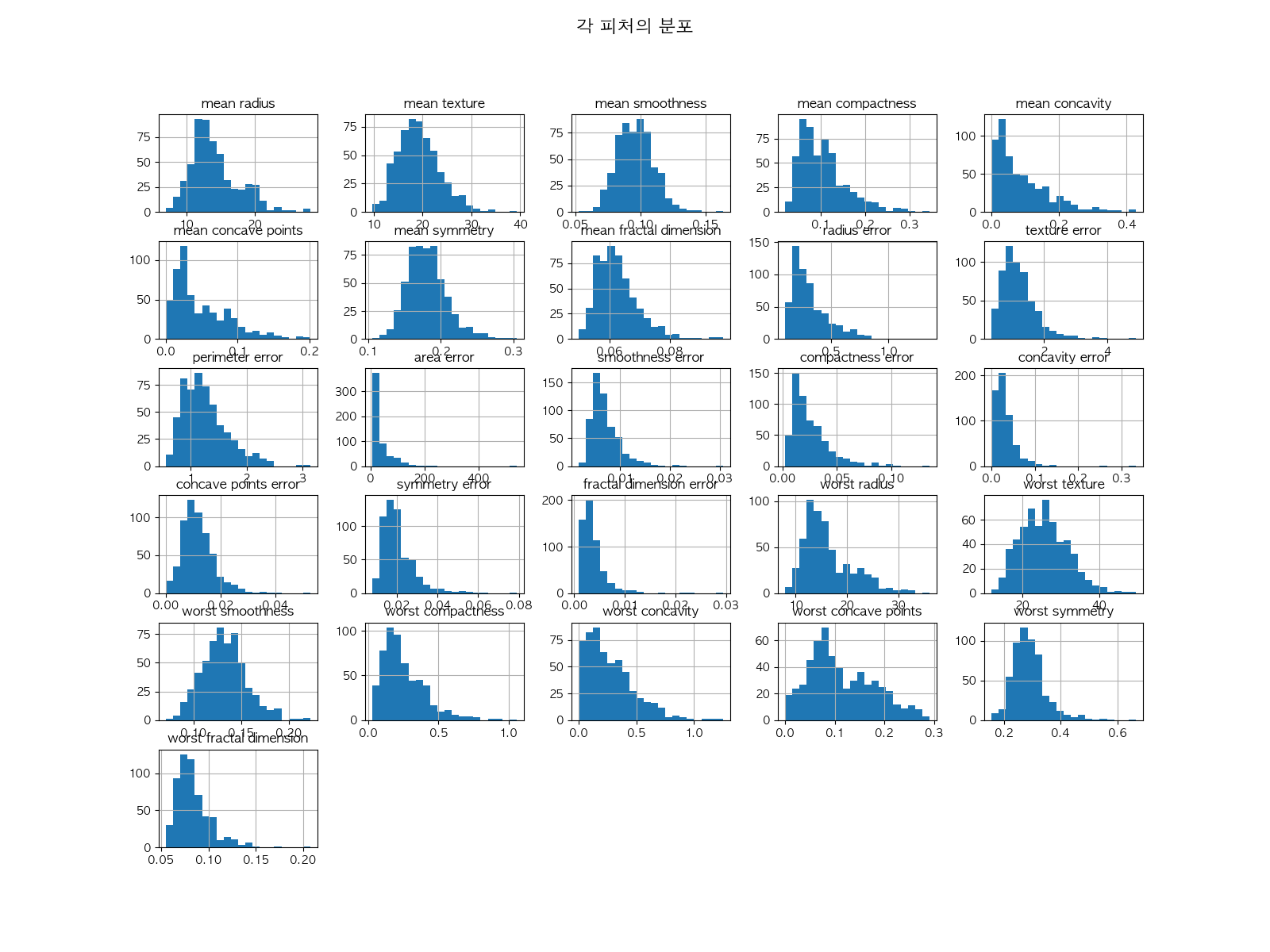


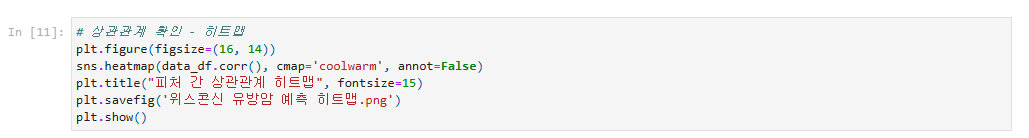


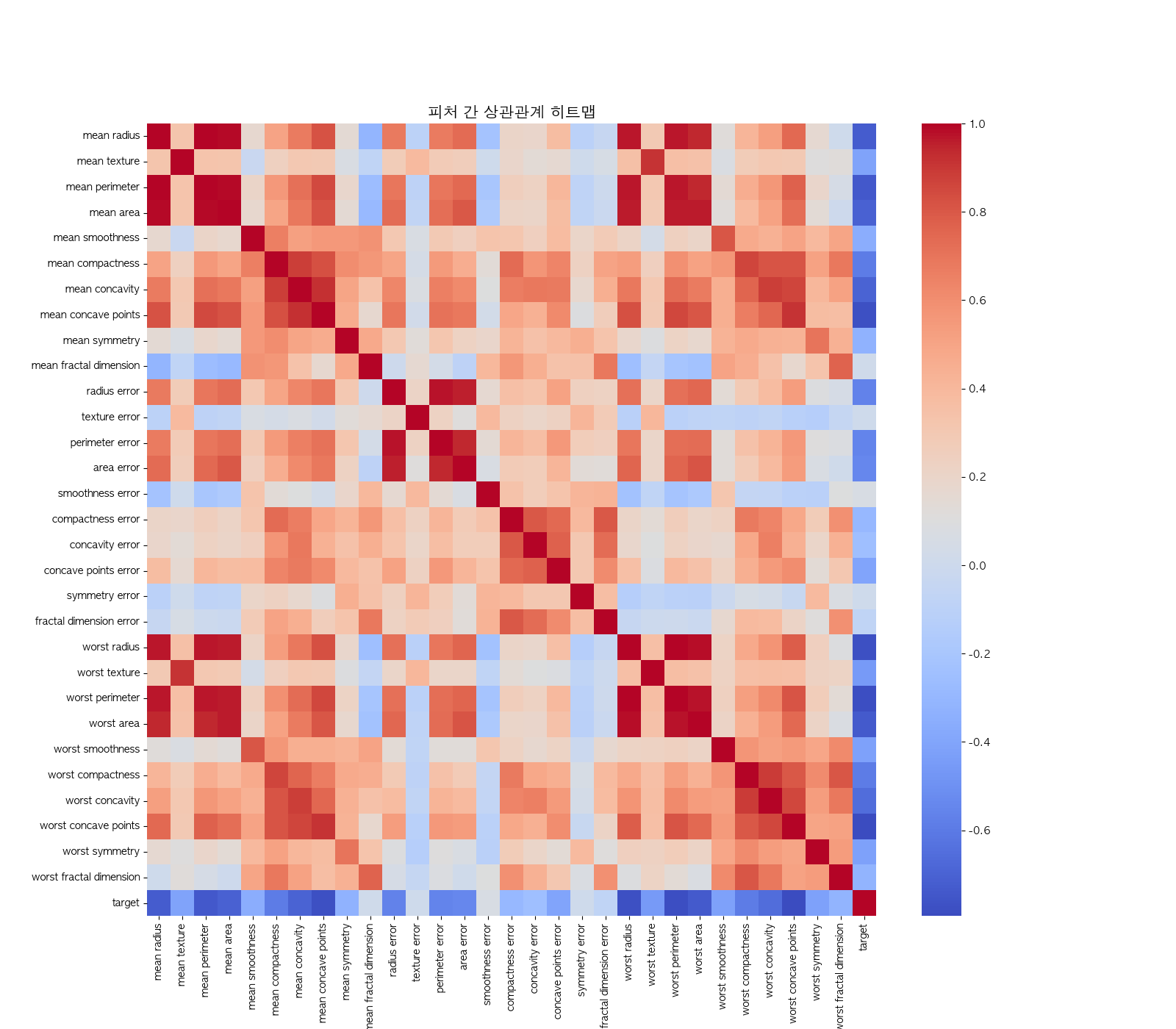
## 2.4 상관계수 분석 및 시각화

corr() 메서드와 seaborn의 heatmap()을 활용하여 피처 간의 상관관계를 시각화하였습니다. 그 결과 일부 변수들 간에 매우 높은 상관관계가 존재함을 발견하였고,이는 다중공선성 문제를 유발할 수 있음을 알게되었습니다.





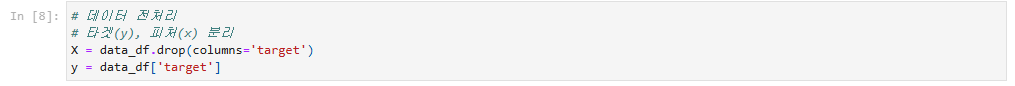




# **데이터 전처리**

## 3.1 타겟, 피처 분리

'X'에는 30개의 수치형 피처를, 'y'에는 진단 결과(M 또는 B)를 각각 할당하여 모델 학습에 사용할 독립 변수와 종속 변수를 분리하였습니다. 이 과정은 데이터 분석 및 모델 학습의 첫 단계로 이후 전처리 및 모델링 과정에서 일관된 데이터 사용을 가능하게 합니다.



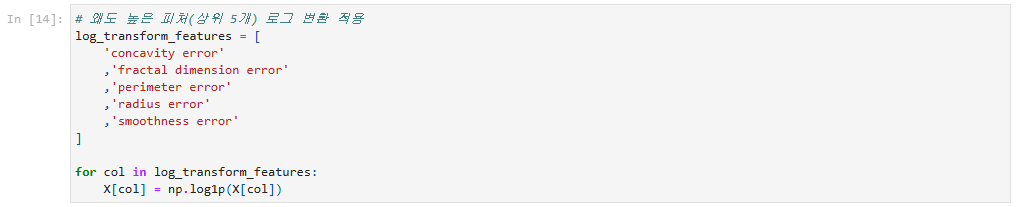
## 3.2 다중공선성 제거

앞서 상관계수 분석에서 확인된 높은 상관관계를 가진 피처 중 일부를 제거함으로써 다중공선성 문제를 완화하였습니다. 이는 특히 로지스틱 회귀와 같은 선형 모델에서 성능 저하를 방지하고 모델의 안정성과 해석 가능성을 높이기 위해 중요한 단계입니다.



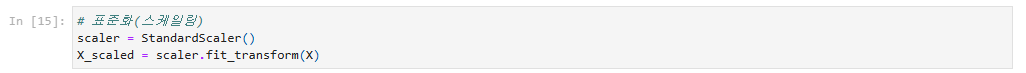
## 3.3 왜도 높은 피처 로그 변환

기초 통계 분석에서 확인된 왜도가 큰 피처들에 대해 로그 변환(np.log1p())을 적용하여 정규성에 가까운 분포로 보정하였습니다. 이를 통해 모델이 피처의 극단값에 덜 민감하도록 만들고, 성능 향상을 기대할 수 있었습니다.



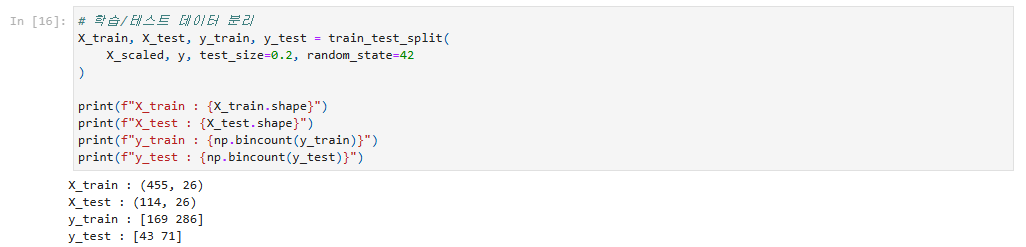
## 3.4 표준화(StandardScaler)

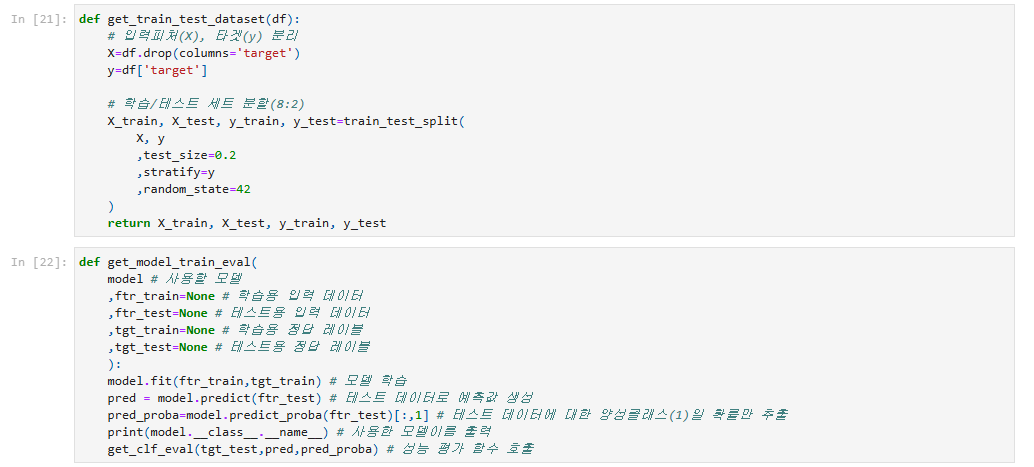
모든 피처에 대해 StandardScaler를 적용하여 평균이 0이고 표준편차가 1인 정규분포 형태로 스케일링을 수행하였습니다. 이는 다양한 머신러닝 알고리즘에서 학습의 수렴 속도를 높이고 특정 피처가 과도한 영향을 미치지 않도록 하기 위한 일반적인 전처리 단계입니다.



## 3.5 학습/테스트 데이터 분리

전처리가 완료된 데이터를 학습용과 테스트용으로 분리하기 위해 train\_test\_split()을 사용하였습니다. 일반적으로 8:2의 비율로 분리하였으며, 랜덤 시드를 고정하여 실험의 재현성을 확보하였습니다. 학습 데이터는 모델 학습과 검증에 사용되며, 테스트 데이터는 최종 평가에 사용됩니다.





1. **모델 개발 및 학습**

4.1 기본 모델 학습 및 평가

다양한 분류 모델을 적용하여 학습 및 성능 평가를 수행하였습니다. 기본적으로 fit()과 predict() 메서드를 사용하였고, 테스트 데이터셋을 통해 각 모델의 정확도, F1 Score, AUC 등을 산출하였습니다.

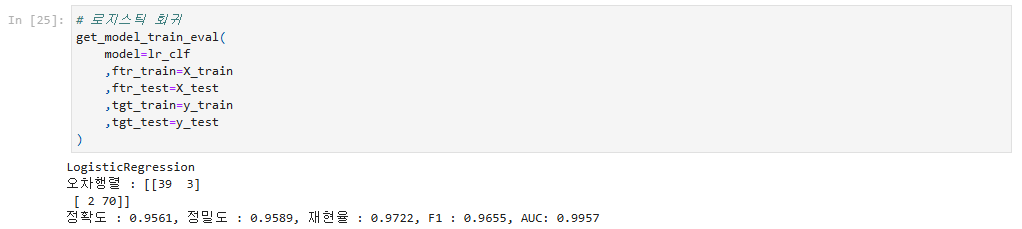


*# Voting은 RandomForest, GradientBoosting, XGBoost, LightGBM 트리 기반 강력한 모델들만을 조합하여 구성함.*

*# 각 모델은 서로 다른 방식의 앙상블/부스팅 전략을 사용하므로 보팅 시 개별 약점을 보완하고 성능을 극대화할 수 있음.*

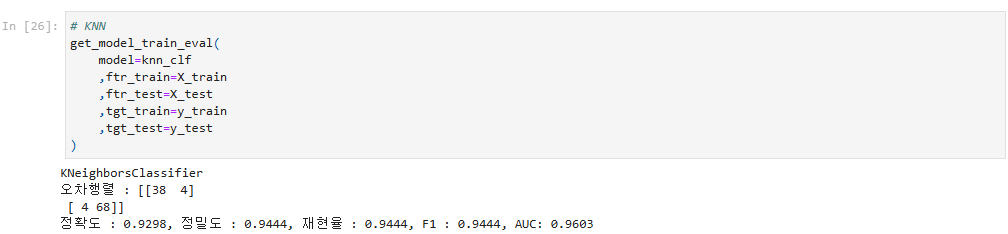
### 로지스틱 회귀(Logistic Regression) :

선형 모델로 해석력이 높으며 베이스라인 모델로 활용하기에 적합합니다.



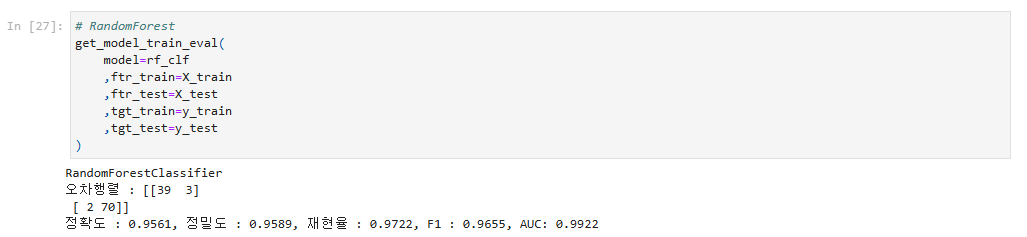
### KNN :

거리 기반 비모수 모델로 스케일링에 민감하며 비교적 단순한 구조입니다.



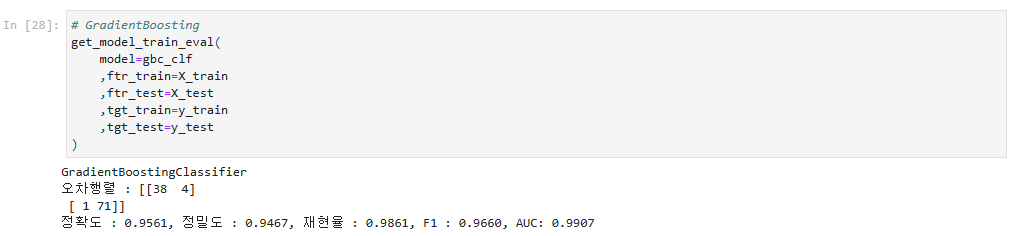
### 랜덤포레스트(Random Forest) :

다수의 결정 트리를 앙상블하는 모델로 과적합에 강하고 변수 중요도 해석이 가능합니다.



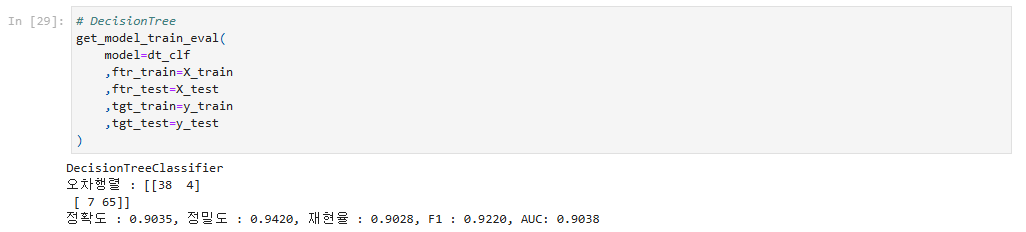
### 그래디언트 부스팅(Gradient Boosting) :

약한 학습기를 순차적으로 결합하여 성능을 개선하는 부스팅 모델입니다.



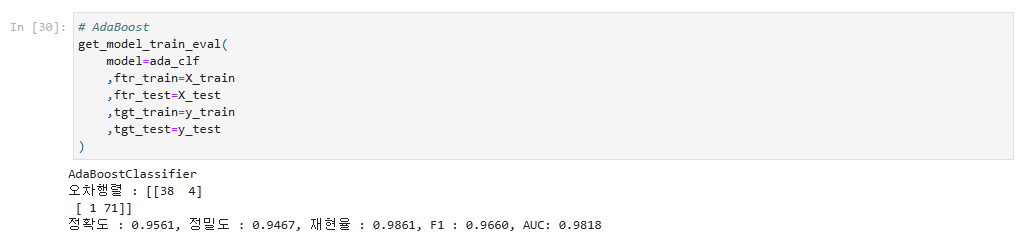
### 의사결정트리(Decision Tree) :

단일 트리 모델로 이해가 쉬우며 시각화에 유리하나 과적합에 민감합니다.



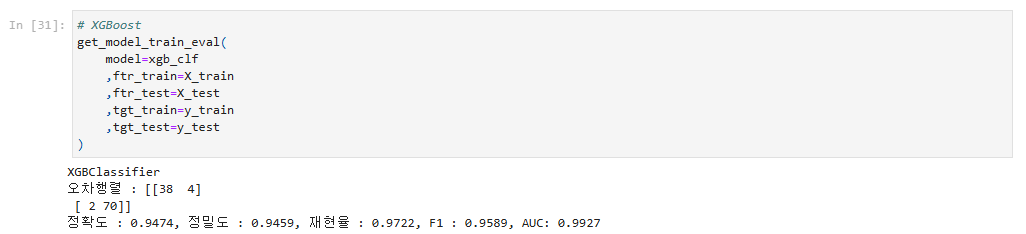
### 에이다부스트(AdaBoost) :

예측이 어려운 샘플에 가중치를 두고 학습하는 부스팅 계열 모델입니다.



### XGBoost :

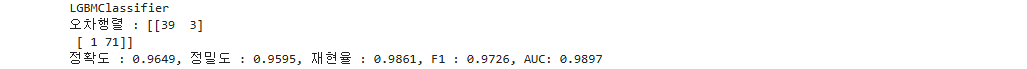
성능과 속도에서 강력한 부스팅 모델로 정규화와 병렬 처리 기능이 강화되어 있습니다.



* LightGBM :

대용량 데이터 처리에 최적화된 부스팅 모델로 빠른 학습 속도와 높은 정확도를 자랑합니다.

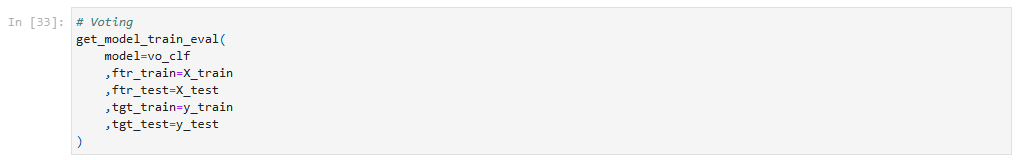


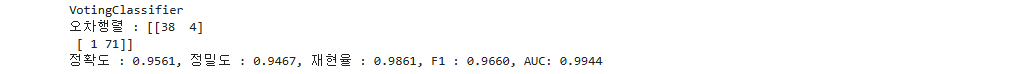


## 4.2 앙상블(Voting) 모델 학습 및 평가

성능이 우수한 여러 모델을 결합하여 VotingClassifier를 구성하였습니다. 소프트 보팅 방식으로 예측 확률의 평균을 기반으로 최종 클래스를 예측하였으며 보팅에 참여한 모델로는 로지스틱 회귀, 랜덤 포레스트, XGBoost, LightGBM 등을 사용하였습니다.

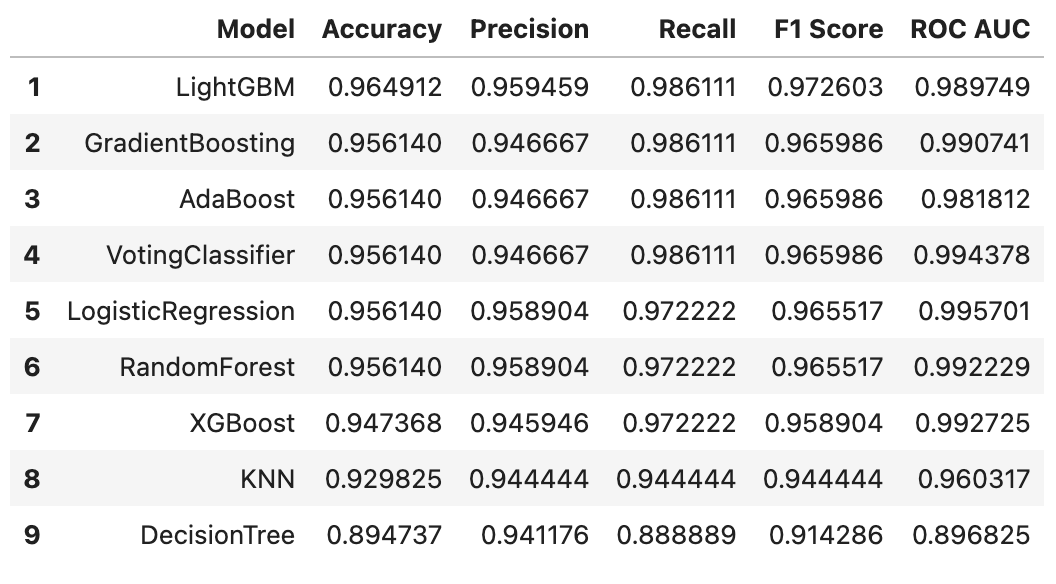
Voting 모델은 개별 모델 대비 더욱 안정적이고 균형 잡힌 성능을 보여주는 장점이 있습니다.





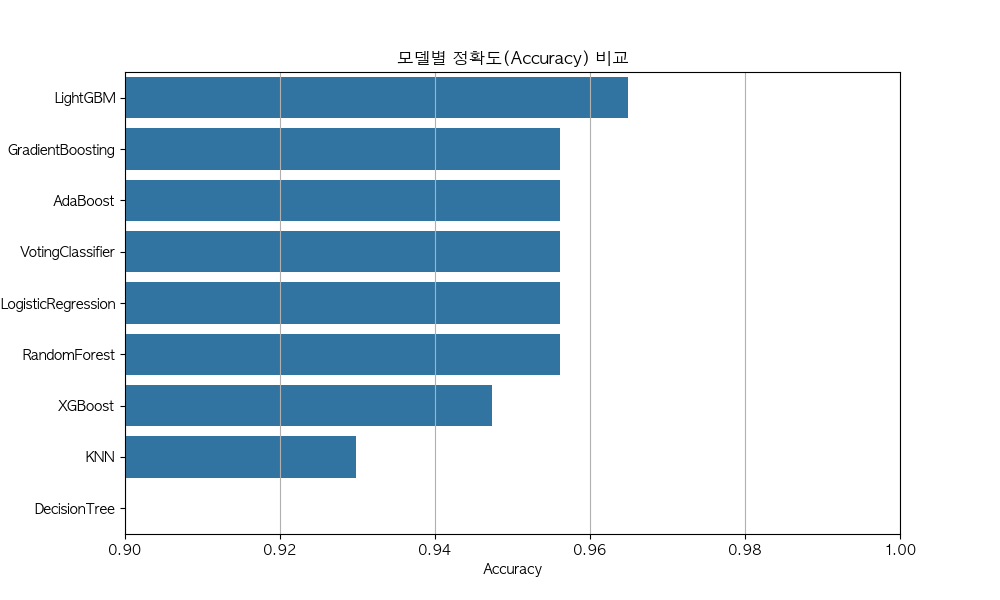
## 4.3 모델 성능 비교

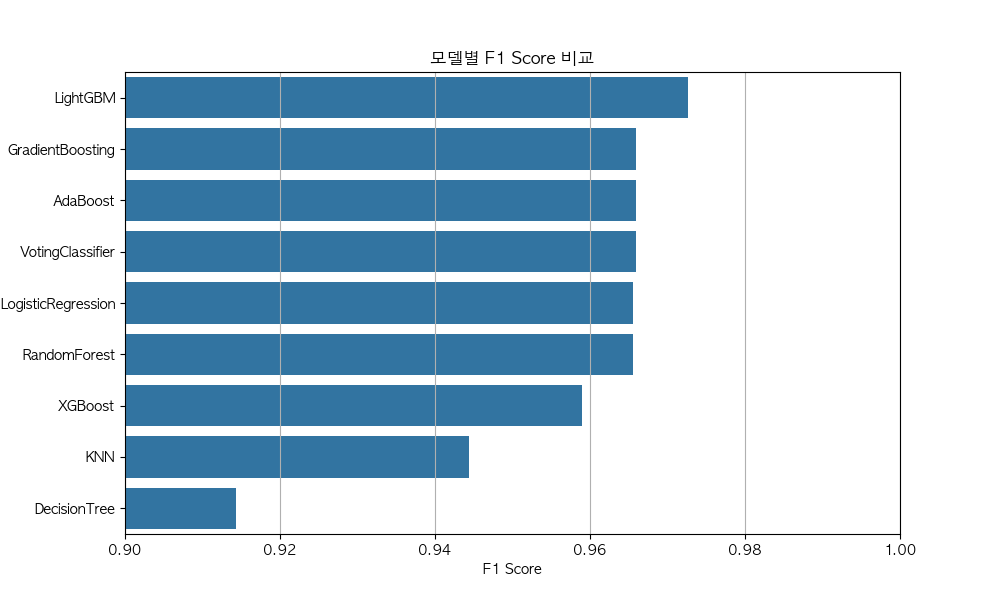
기본 모델 및 보팅 모델의 성능을 정확도, F1 Score, AUC 등 다양한 지표 기준으로 비교하였습니다. 이를 통해 어떤 모델이 가장 우수한 성능을 보였는지 확인하였고, 이후 하이퍼파라미터 튜닝이 성능에 미치는 영향을 분석하는 기초 자료로 활용되었습니다. 시각화 자료(막대 그래프 등)를 통해 성능 차이를 직관적으로 표현하였습니다.

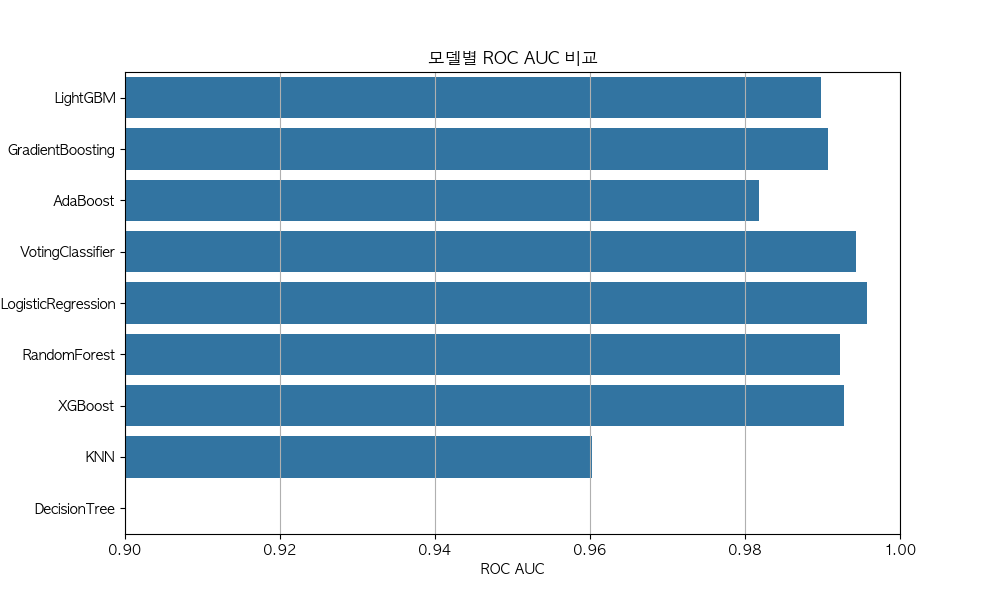












*# 유방암 분류*

*# - 악성(malignant=0) vs 양성(benign=1) : 클래스 수가 비슷해보이지만 조금 불균형*

*# - 정확도 기준으로 성능 파악*

*# - 암 진단은 놓치는게 더 위험하기 때문에 재현율과 F1 Score도 중요*

# **하이퍼파라미터 튜닝**

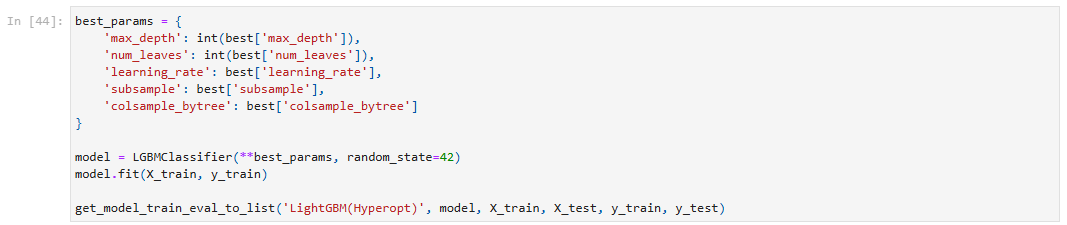
## 5.1 모델 튜닝

LightGBM, RandomForest, Gradient Boosting 모델은 사전 실험에서 비교적 정확도가 높았던 모델 순서에 따라 튜닝을 우선 진행하였습니다. 이후 XGBoost 모델은 성능과 범용성을 고려하여 추가로 포함하고 튜닝하였습니다.

### LightGBM : Hyperopt

베이지안 최적화 알고리즘인 Hyperopt를 사용하여 max\_depth, num\_leaves, learning\_rate 등을 튜닝.

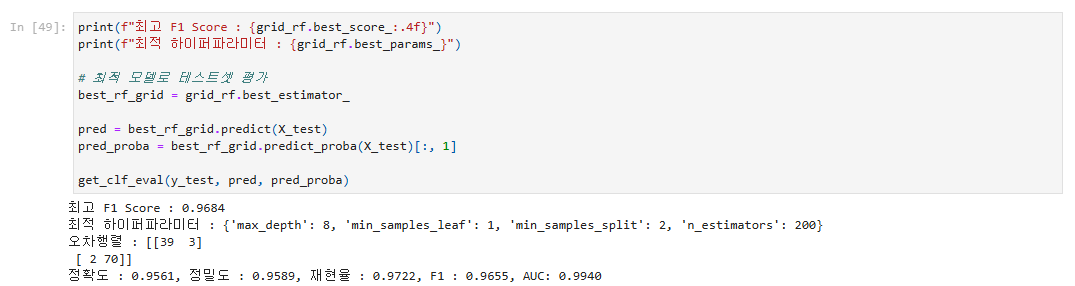




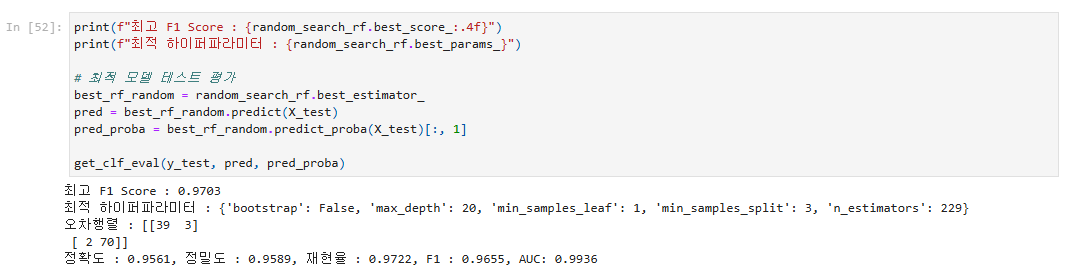
### RandomForest : Grid/Random Search

n\_estimators, max\_depth, min\_samples\_split 등 파라미터를 GridSearchCV 및 RandomizedSearchCV를 통해 탐색.



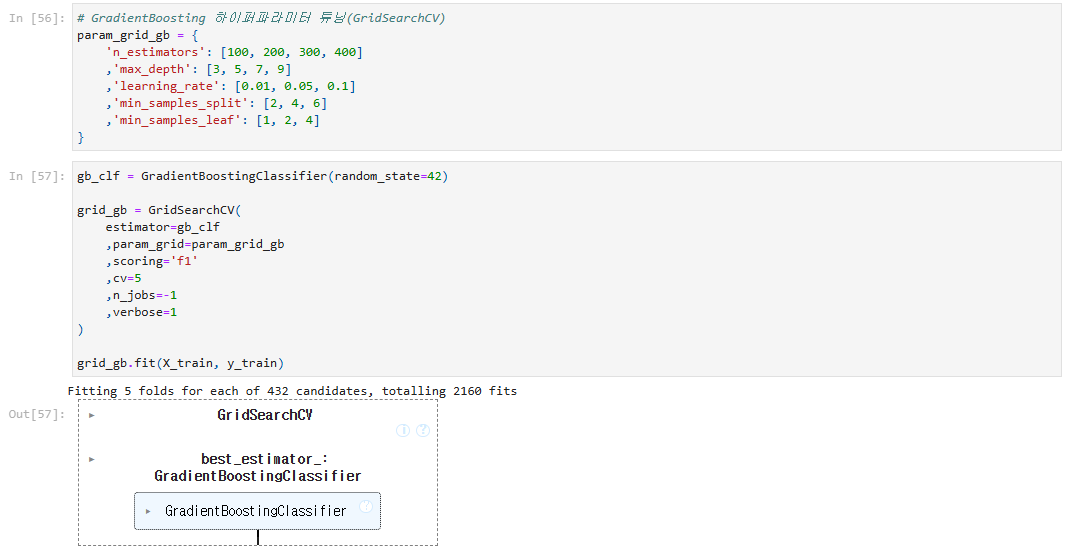


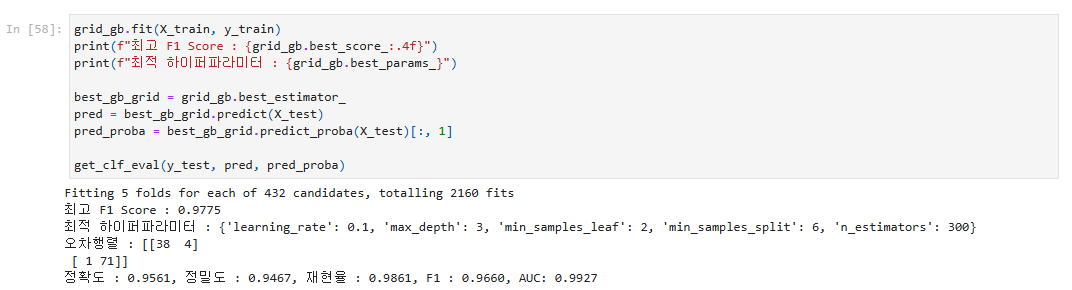


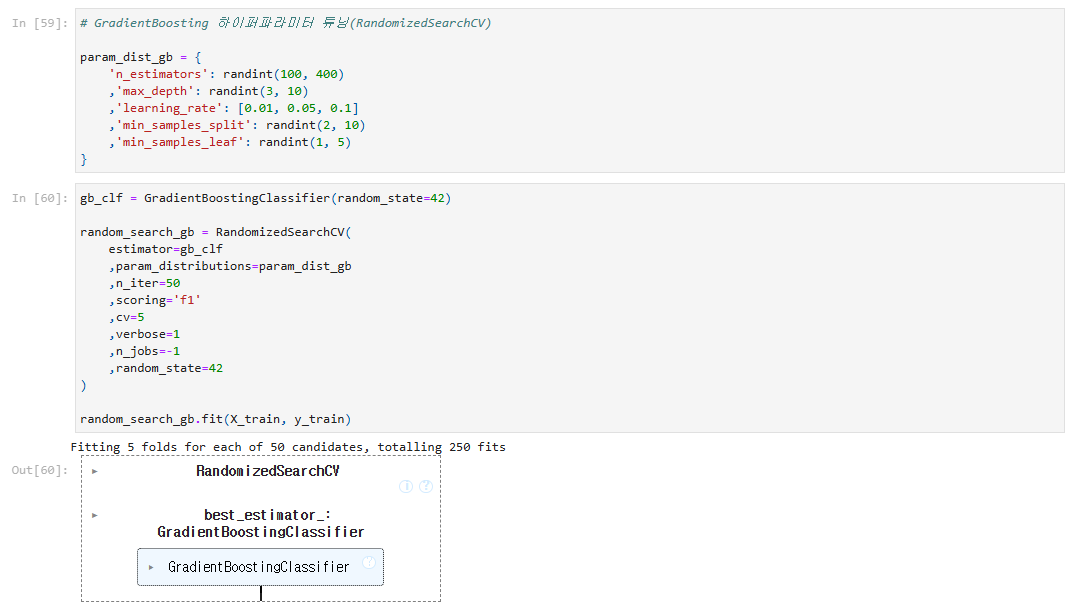


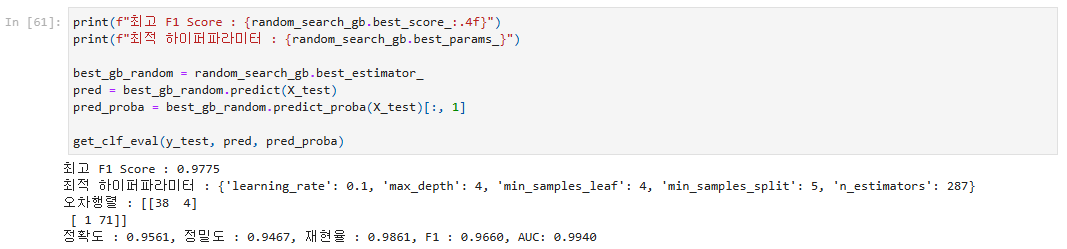
### GradientBoosting : Grid/Random Search

GridSearchCV와 RandomizedSearchCV 모두 적용하여 튜닝된 파라미터 조합을 확보.







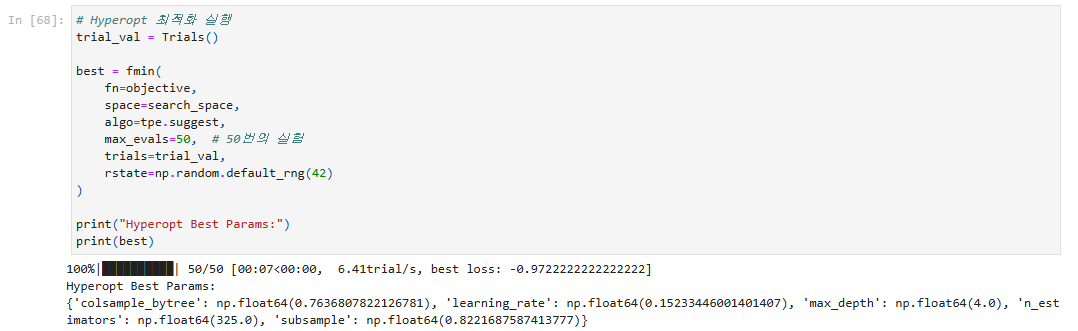


### XGBoost : Hyperopt

Hyperopt를 사용하여 max\_depth, learning\_rate, subsample, gamma 등을 최적화함.







## 5.2 튜닝 결과 비교

### LightGBM : Hyperopt



*# 정확도 떨어짐*

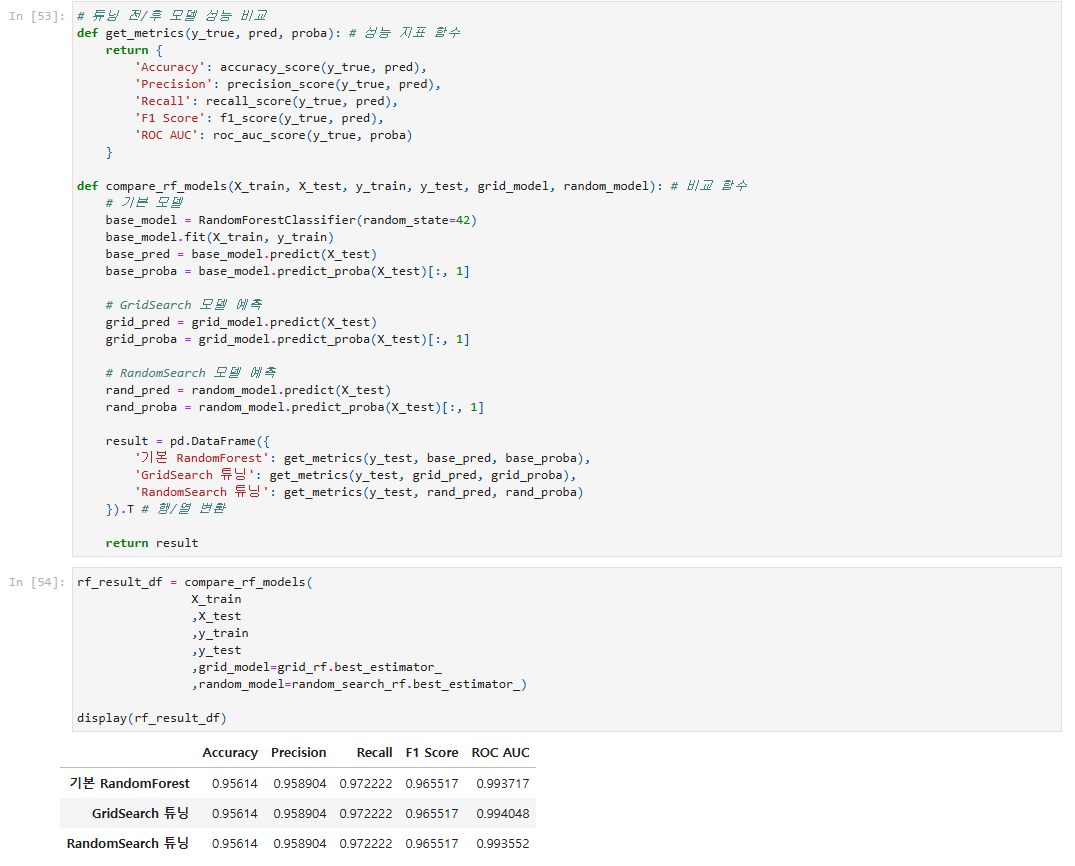
*# 정밀도 떨어짐*

*# 재현율 동일*

*# F1 Score 떨어짐*

*# AUC 증가*

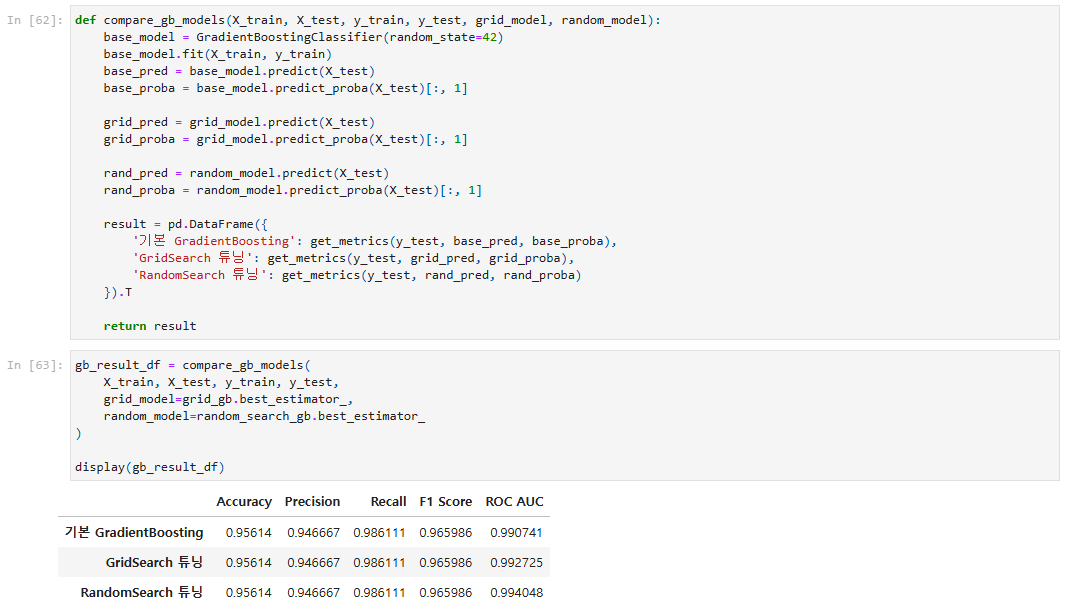
### RandomForest : Grid/Random Search



*# 튜닝 효과가 거의 없음*

*# GridSearch, RandomSearch 둘 다 기본 모델과 큰 차이 없음*

### GradientBoosting : Grid/Random Search



*# 튜닝 효과 거의 없음*

*# AUC에서만 소폭 증가*

### XGBoost : Hyperopt





*# 정확도 증가*

*# 정밀도 증가*

*# 재현율 증가*

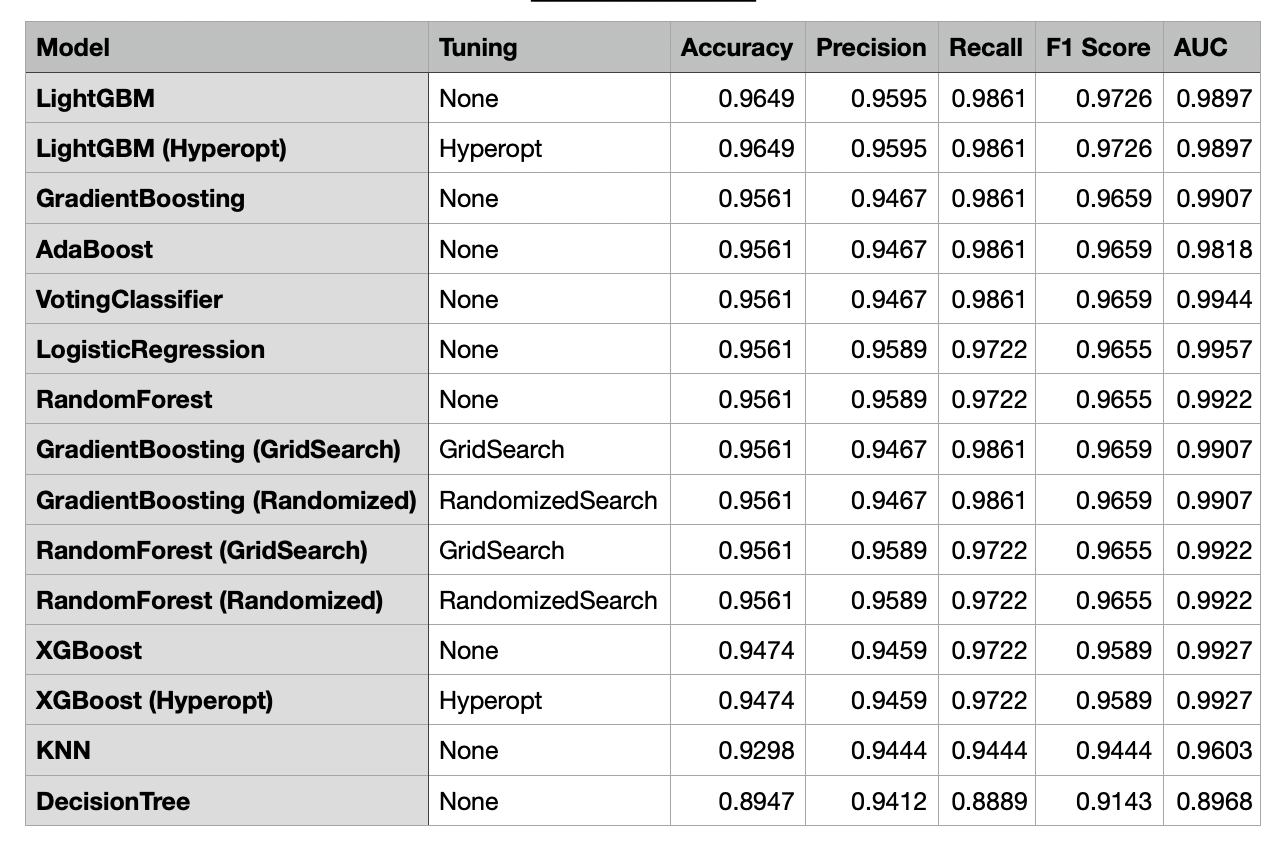
*# F1 Score 증가*

*# AUC 동일*

# **전체 모델 평가 및 성능 비교**

## 6.1 정확도 기준 전체 모델 성능 비교

아래는 기본 모델 및 하이퍼파라미터 튜닝이 적용된 모델들을 포함한 최종 성능 비교 표입니다. 정확도(Accuracy)를 기준으로 내림차순 정렬하였습니다.



## 6.2 성능 해석 요약

* 정확도 기준 최고 모델은 LightGBM (기본/튜닝 모두 동일)으로 정확도 0.9649, F1 Score 0.9726의 매우 우수한 성능을 기록했습니다.
* 정확도 0.9561 그룹에는 GradientBoosting, RandomForest, AdaBoost, Voting, LogisticRegression 등 다양한 모델이 포함되며 튜닝 유무에 관계없이 전체적으로 비슷한 성능을 나타냈습니다.
* XGBoost는 정확도는 상대적으로 낮은 0.9474였지만 AUC는 0.9927로 매우 높아 임계값 조정이 중요한 경우에 적합한 모델입니다.
* KNN과 DecisionTree는 상대적으로 정확도 및 F1 Score가 낮아, 본 문제에서는 비효율적인 모델로 평가됩니다.

## 6.3 결론

* 정확도, F1 Score 기준으로는 LightGBM이 전반적으로 가장 안정적이고 우수한 성능을 보였으며, 튜닝을 하지 않더라도 강력한 베이스라인 모델임을 확인할 수 있었습니다.
* GradientBoosting과 RandomForest 역시 튜닝을 통해 일정 수준 이상의 성능을 안정적으로 확보할 수 있는 모델이었습니다.
* Voting과 AdaBoost는 복잡도를 많이 늘리지 않고도 준수한 성능을 보여 실무에서 빠르게 성능 확보가 필요한 경우 좋은 선택이 될 수 있습니다.
* XGBoost는 정확도보다는 민감도(AUC) 기반 의사결정에 유리한 모델로 분류됩니다.

# **결론 및 인사이트**

이번 프로젝트에서는 위스콘신 유방암 데이터셋을 활용하여 다양한 머신러닝 모델을 적용하고 전처리 및 하이퍼파라미터 튜닝을 통해 모델 성능을 향상시키는 과정을 수행하였습니다.

결론 :

전처리 단계에서 다중공선성 제거, 로그변환, 표준화 등의 기법을 통해 데이터 품질을 향상시켰으며 이는 전반적인 모델 성능 개선에 기여했습니다.

기본 모델 중에서는 LightGBM, GradientBoosting, RandomForest가 우수한 정확도와 F1 점수를 보여주었고 하이퍼파라미터 튜닝을 적용한 LightGBM은 튜닝 전과 정확도가 동일하여 성능 향상에는 영향을 주지 않았음을 확인했습니다.

튜닝 여부와 상관없이 성능이 크게 달라지지 않은 모델도 있었으나 모델별 최적화 가능성 확인 측면에서는 의미 있는 실험이었습니다.

인사이트 및 적용 방안 :

다양한 모델과 전처리, 튜닝 전략을 비교하는 것은 실무에서 최적 모델을 선정하는 데 있어 중요한 사전 작업이 될 수 있습니다.

단순히 하나의 지표가 아닌 정확도, F1 Score, AUC 등 여러 지표를 함께 고려하여 상황에 맞는 모델을 선택하는 것이 중요합니다.

LightGBM은 높은 성능과 빠른 학습 속도 측면에서 실무 적용 가능성이 매우 높은 모델로 판단됩니다.

이번 분석을 통해 머신러닝 워크플로우의 전반적인 흐름(EDA → 전처리 → 모델링 → 튜닝 → 평가)을 체계적으로 경험할 수 있었으며 이는 향후 데이터 분석 및 AI 프로젝트 수행에 기반이 될 수 있습니다.