**[데이터 설명]**

ID: 각 샘플을 구분하는 고유 식별자.

SUBCLASS: 각 샘플에 대한 암종을 나타내는 레이블 (타겟 변수).

A2M ~ ZYX: 유전자 정보(열)가 담긴 열로, 각 열은 유전자 변이를 나타냅니다.

WT(Wild Type)는 해당 유전자에 변이가 없음을, 그 외 값은 특정 변이가 있음을 나타냅니다.

**[기술 통계 확인]**

SUBCLASS는 26개의 고유 값이 있고, 이는 26개의 서로 다른 암종이 존재

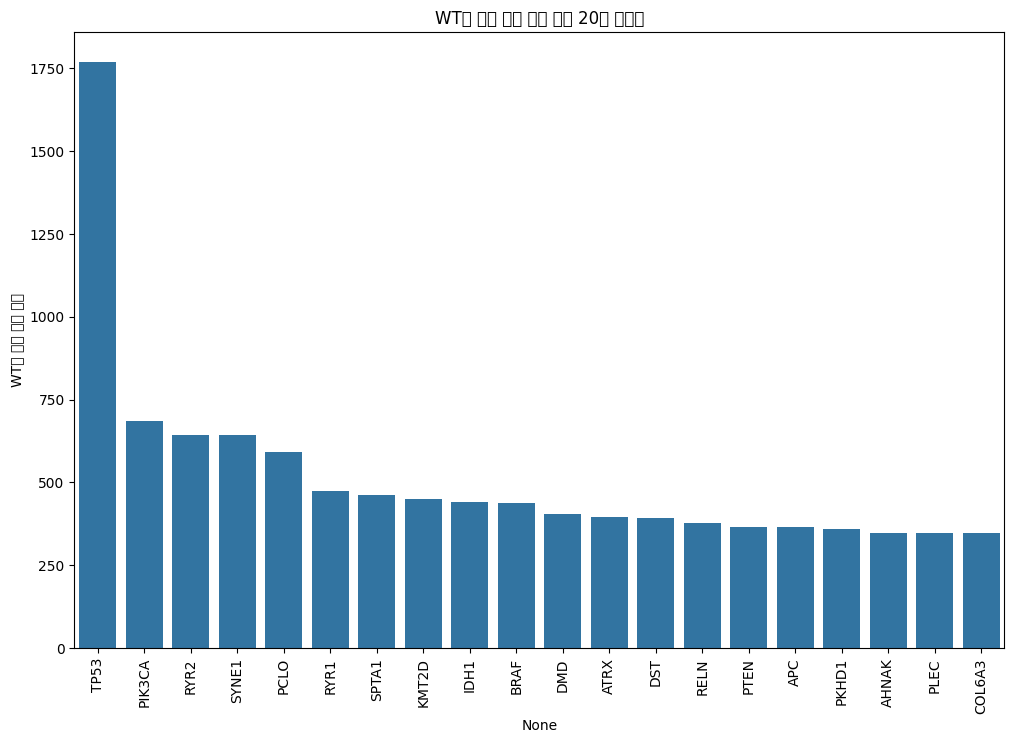
SUBCLASS에서 가장 많이 나타나는 값은 BRCA로, 786번 등장

A2M과 같은 다른 변수는 가장 많이 나타나는 값이 WT로, 이는 변이가 없음

데이터 결측치 없음

**[데이터 분석]**

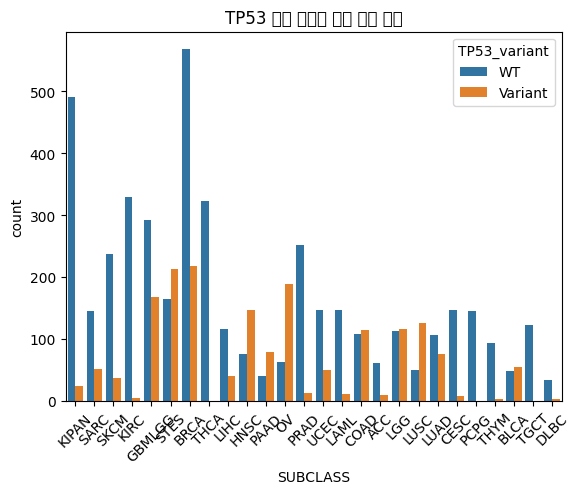
1. **각 유전자 특정 변이(WT가 아닌 값)가 자주 발생하는 상위 20개 유전자를 시각화**
   1. 이슈 : TP53열이 WT가 아닌 값의 개수가 1750 값으로 가장 많은 값으로 측정



1. **TP53 변이 여부에 따른 암종 분포 시각화**
   1. TP53유전자 변이 암종 분류에 얼마나 중요한지 파악.
   2. TP53변이가 있는 데이터와 없는 데이터 나눠 암종 분포 차이 확인
   3. 이슈 : "BRCA"와 GBM 암종에서 WT와 변이 모두 많이 관찰

SARC, SKCM, LUSC 등 몇몇 암종에서 변이가 상대적으로 많이 관찰

ESCA, HNSC, LUAD 등은 변이와 WT 모두 고르게 분포



1. **TP53 변이가 암종 분류에 미치는 영향 분석**
   1. 카이제곱 검정을 사용하여 TP53 변이와 특정 암종 간의 유의미한 관계를 확인
   2. 이를 통해 TP53 변이 여부가 암종 분류에 중요한지 평가하고, 모델에 변이 여부를 특징으로 추가할 수 있음.
   3. p-value가 유의수준(예: 0.05)보다 작다면, TP53 변이가 암종 분류에 유의미한 영향을 미친다고 할 수 있음
   4. 이슈
      1. p-value는 0.0으로 두 변수 간 통계적으로 매우 유의미한 상관 관계가 있음
      2. TP53 변이 여부는 암종 분류하는데 유의미한 영향 미침
      3. TP53 변이가 각 암종에서 중요한 분류 기준이 될 수 있음

1. **히트맵 분석 - WT가 아닌 값이 있는 상위 20개 유전자 간의 상관관계**
   1. 현재 제공된 히트맵은 WT가 아닌 값이 있는 상위 20개 유전자 간의 상관관계를 보여줌
   2. 이 상관관계 히트맵을 통해 각 유전자 간 변이 패턴이 얼마나 비슷하게 나타나는지 분석할 수 있음
   3. 이슈

(1) TP53 유전자

TP53은 그래프 상에서 다른 유전자들과 거의 상관관계가 없습니다.

대부분의 유전자와 상관계수가 0에 가까운 것으로 나타나며, 특히 PIK3CA 및 "BRAF" 등과는 음의 상관관계를 보이고 있습니다.

이는 TP53 변이가 독립적으로 발생하거나, 다른 유전자의 변이와는 연관성이 낮다는 것을 나타냅니다.

TP53 변이는 암에서 매우 중요한 유전자 중 하나로 알려져 있기 때문에, 이 유전자의 독립적인 특성을 고려하여 암종 분류에서 독립 변수로 사용하는 것이 좋습니다.

(2) RYR1, SYNE1, RYR2의 상관관계

RYR1, RYR2, SYNE1은 상관관계가 상대적으로 높은 편입니다. 이들 간의 상관계수는 약 0.25~0.28로, 다른 유전자들에 비해 상호간의 변이 패턴이 비슷하게 나타날 가능성이 있습니다.

이 상관관계는 특정 암종에서 공통적으로 변이가 일어날 가능성을 시사합니다. 이들 유전자 그룹은 암종 분류에서 함께 고려해볼 만합니다.

(3) IDH1과 ATRX의 높은 상관관계(<- 이부분 분석함)

IDH1과 ATRX는 상관계수가 0.40으로 상대적으로 높은 상관관계를 보여줍니다.

이 두 유전자는 실제로 뇌종양 등 특정 암종에서 자주 함께 변이가 나타나는 것으로 알려져 있으며, 이 결과는 이를 반영한 것으로 보입니다.

이 상관관계를 바탕으로 IDH1과 ATRX가 동시에 변이하는 환자군을 따로 분석하거나 이 두 유전자를 암종 분류에 중요한 특징으로 활용할 수 있습니다.

(4) 다른 유전자 간의 미미한 상관관계

그 외 유전자 간 상관계수는 대부분 0에 가깝거나, -0.2~0.2 범위에 분포하고 있습니다.

이는 대다수의 유전자들이 독립적으로 변이가 일어나고 있음을 나타냅니다. 다시 말해, 이 유전자들은 특정 유전자가 변이를 일으킬 때 다른 유전자도 같이 변이하는 패턴이 적다는 것을 보여줍니다.

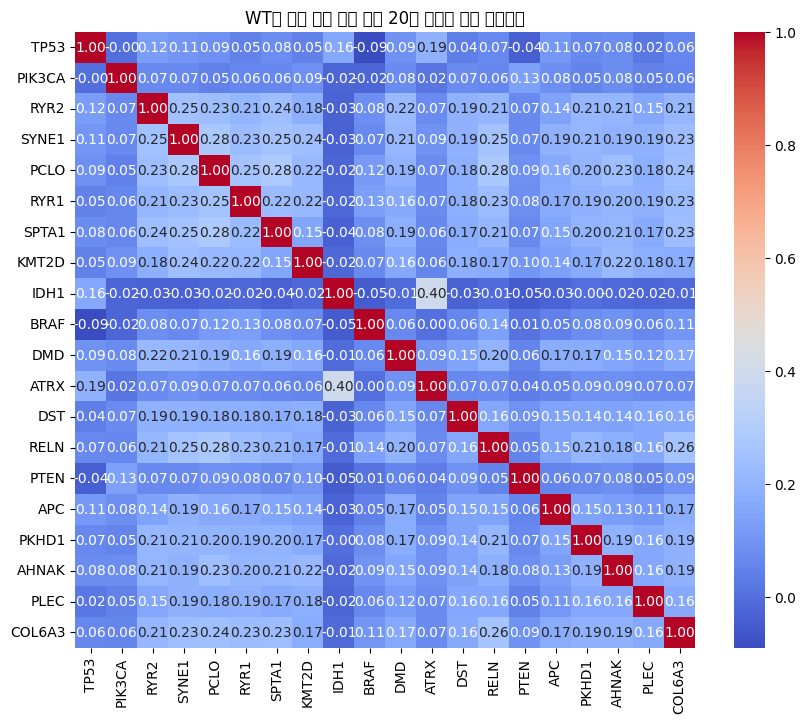
(5) 특정 음의 상관관계

몇몇 유전자 쌍 간에는 음의 상관관계가 존재합니다. 예를 들어, TP53과 BRAF의 상관계수는 -0.09로 나타났습니다. 이는 한 유전자에서 변이가 발생할 때 다른 유전자에서는 변이가 덜 발생하는 경향이 있음을 시사합니다.

이런 음의 상관관계는 변이 패턴이 서로 상반되는 유전자들이 있음을 나타내며, 이를 모델링할 때 함께 고려하면 예측 성능을 향상시킬 수 있습니다.

(6) 분석의 의미

유전자 간의 변이 패턴 파악: 상관관계가 높은 유전자들은 함께 변이하는 경향이 있기 때문에 암종 분류에 있어 중요한 역할을 할 수 있습니다.

이러한 유전자들을 중심으로 특징 선택(feature selection)을 하면 더 나은 모델을 만들 수 있습니다.

1. **WT가 아닌 값이 있는 상위 20개 유전자 간의 상관관계 중, IDH1과 ATRX 유전자 데이터가 높은 상관관계로 나타남 .**

두 유전자는 실제로 뇌종양 등 특정 암종에서 자주 함께 변이가 나타나는 것으로 알려져 있어, 이 상관관계를 바탕으로 IDH1과 ATRX가 동시에 변이하는 환자군을 분석해보기

1. WT를 0으로, 다른 변이 값을 1로 변환
2. IDH1\_ATRX 동시 변이 여부에 따른 암종 분포
   1. 이슈

GBMLGG, LGG 암종에서 IDH1과 ATRX의 동시 변이가 집중적으로 발생하고 있는 것을 볼 수 있습니다.

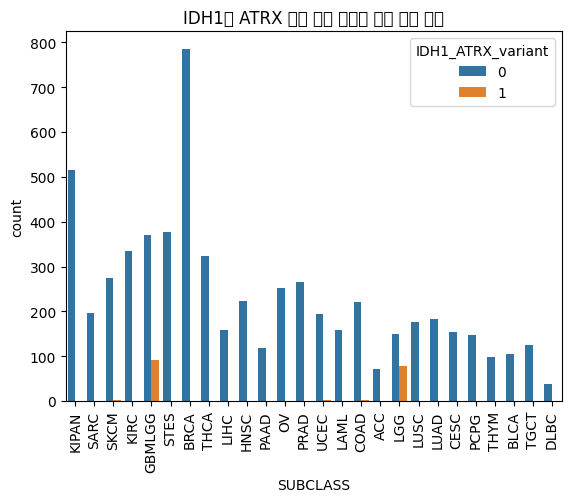
이는 GBMLGG에서 이 두 유전자의 변이가 중요한 역할을 할 수 있음을 나타냅니다.

나머지 암종에서는 \*\*동시 변이(1)\*\*가 거의 발생하지 않으며,

대부분의 샘플에서 IDH1과 ATRX가 독립적으로 변이하거나 변이가 없는(0) 상태입니다.

이러한 결과는 IDH1과 ATRX의 동시 변이가 특정 암종(예: GBMLGG)에만 주로 나타난다는 것을 시사합니다.

이는 실제로도 알려진 사실로, IDH1과 ATRX는 주로 뇌종양에서 변이를 일으키는 유전자입니다.

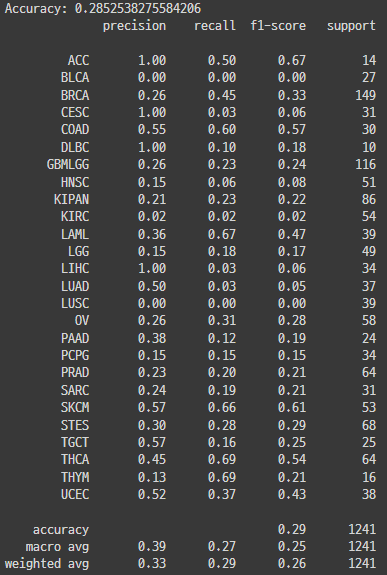
* 1. 

**6. 다음 분석 제안**

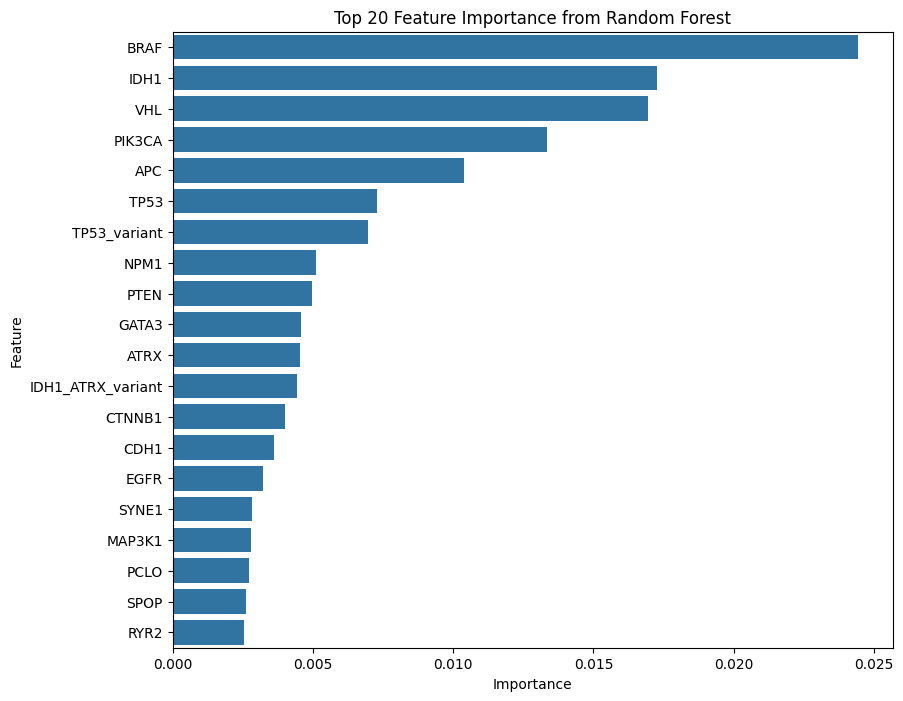
1. **Random Forest** 모델을 사용하여 중요한 유전자 변이를 파악
   1. 설명 : IDH1\_ATRX\_variant 열을 포함한 전체 유전자 데이터를 사용해 Random Forest를 훈련하고, 그 결과로 중요한 유전자를 찾는 방법

**방법1) 랜덤포레스트 모델 성능 분석**

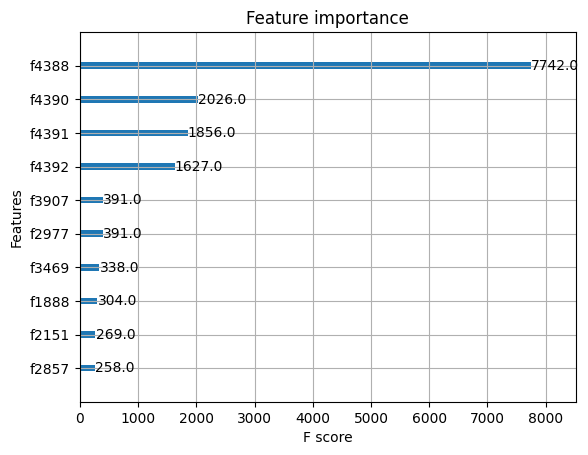
* + 1. 전체 정확도 : 0.29로 낮은편, 이 모델이 암종 분류하는데 어려움있음
    2. 정밀도, 재현율, f1-score
    3. BLCA(유방암), LUSC은 정밀도와 재현율이 매우 낮음 -> 잘못된 예측을 많이 하고 있음을 의미
    4. 잘 예측된 암종 : **THCA(갑상선암), ACC(부신피질암)**의 경우 상대적으로 높은 정확도와 f1-socre를 보임.



**방법2) Feature Importance 분석**

* + 1. 상위 중요도 유전자 : **BRAF, IDH1** 유전자가 가장 중요한 특징. 이 유전자들이 다양한 암종에서 중요한 역할을함
    2. 두번째 중요도 유전자 : VHL, PIK3CA, APC, TP53도 중요한 유전자 변이.
    3. IDH1는 상위 중요도 특징이고 IDH1\_ATRX\_variant는 중간 정도 중요도. -> **IDH1과 ATRX의 동시 변이**가 특정 암종에서 중요한 역할을 시사함
  1. 방법1,2에 따른 결론
     1. **Feature Importance** 분석에서는 **BRAF**, **IDH1** 등의 유전자가 중요한 역할을 하는 것으로 나타났습니다. 이러한 유전자들은 모델에서 암종을 분류하는 데 중요한 정보로 사용되고 있습니다.
     2. **모델 성능**은 전반적으로 개선이 필요합니다. 특히 일부 암종에 대해 예측 성능이 매우 낮으며, 특정 암종에서는 전혀 예측하지 못하는 문제가 발생하고 있습니다. 이를 해결하기 위해서는 데이터 불균형을 해결하거나 모델을 튜닝할 필요가 있습니다.

**7. baseline\_4.ipynb의 피처 중요도 시각화**



* **f4388 실제 피처 이름 : mutation\_freq**
* **f4390 실제 피처 이름 : mutation\_freq\_ratio**
* f4391 실제 피처 이름 : mutation\_freq\_TP53\_interaction
* f4392 실제 피처 이름 : mutation\_freq\_ratio\_TP53\_interaction

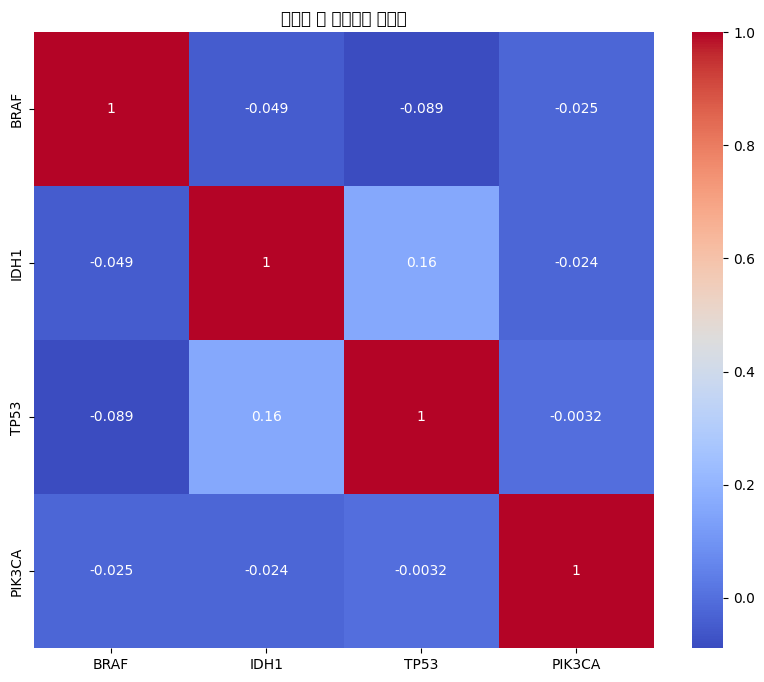
**mutation\_freq ?**

* **mutation\_freq**는 모든 유전자에 대해 전체적으로 변이가 얼마나 자주 발생하는지를 보여주는 지표입니다.
* **mutation\_freq**는 환자 전체의 유전적 변이의 수준을 나타냅니다. 이는 특정 암종이 여러 유전자에서 동시에 변이를 보일 수 있음을 의미합니다.
* 즉, 특정 유전자 변이(IDH1, BRAF)만을 고려하는 것보다 **mutation\_freq**는 더 많은 변이 정보를 반영하고 있어 전체적인 변이 경향을 포괄적으로 나타낼 수 있습니다.

**mutation\_freq\_ration ?**

* mutation\_freq\_ratio는 변이 빈도와 총 변이의 비율을 통해 "상대적인 변이 중요성"을 반영합니다. 이는 유전자 변이가 상대적으로 얼마나 중요한지를 나타낼 수 있기 때문에, 중요한 피처로 작용할 수 있습니다.

**8. 특정 변이 정보 시각화**

1. BRAF, IDH1, TP53, PIK3CA 간의 상관관계

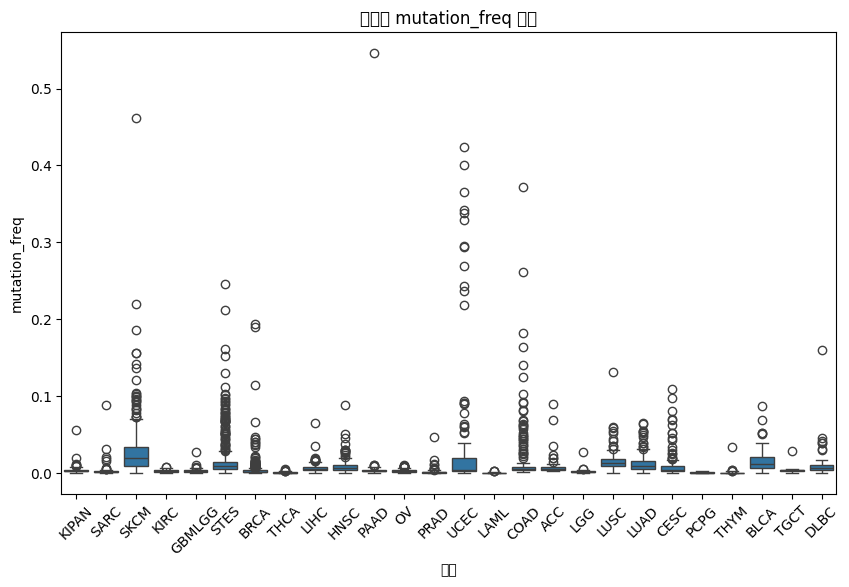
**그래프 해석**

1. **상관 계수 값**:
   * 상관 계수는 두 변수 간의 선형 관계를 나타내며, 값의 범위는 -1에서 1입니다.
     + **1**: 완벽한 양의 상관관계 (한 변수가 증가할 때 다른 변수도 증가).
     + **0**: 선형 관계 없음.
     + **-1**: 완벽한 음의 상관관계 (한 변수가 증가할 때 다른 변수는 감소).
2. **유전자 간 상관관계**:
   * **IDH1과 TP53**: 상관 계수가 **0.16**으로, 약간의 양의 상관관계를 보입니다. 이는 두 유전자가 일부 암종에서 함께 변이될 가능성이 있다는 것을 시사합니다.
   * **다른 유전자 조합**: 대부분의 상관 계수가 0에 가까운 값을 가지고 있어, 다른 유전자들 간에는 특별한 상관관계가 없거나 매우 약한 선형 관계를 보입니다. 예를 들어, BRAF와 PIK3CA는 상관 계수가 **-0.025**로 거의 상관이 없습니다.
3. **색상 해석**:
   * **빨간색**: 강한 양의 상관관계 (값이 1에 가까움). 예를 들어, BRAF 유전자는 스스로와 완벽한 상관관계를 가집니다 (대각선에 있는 값들).
   * **파란색**: 상관관계가 거의 없거나 음의 상관관계 (값이 0 또는 음수). 예를 들어, TP53와 PIK3CA 사이의 상관관계는 **-0.0032**로 거의 없습니다.

### **결론:**

이 그래프는 BRAF, IDH1, TP53, PIK3CA 유전자 변이 간의 상관관계가 대부분 약하거나 거의 없다는 것을 보여줍니다. 다만, **IDH1과 TP53**의 상관관계는 약간 긍정적이며, 이 두 유전자는 특정 암종에서 함께 변이가 발생할 가능성을 시사합니다. 이러한 상호작용은 모델에서 중요한 예측 변수로 작용할 수 있습니다.

1. 암종별로 mutation\_freq(변이 빈도)의 분포를 상자 수염 그림으로 나타내 변이 빈도가 어떻게 분포하는지 보기

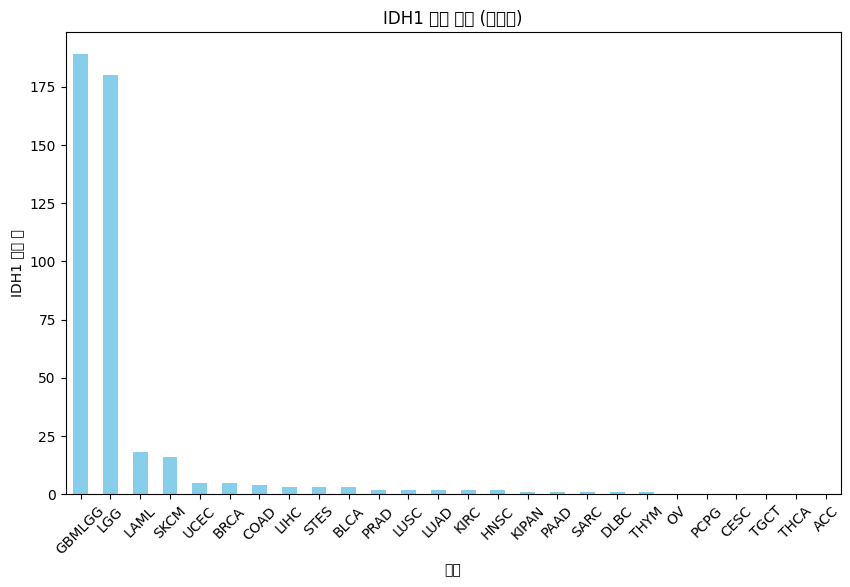


### **그래프 해석:**

1. **세로축 (Y축):**
   * mutation\_freq: 유전자 변이 빈도를 나타냅니다. 값이 높을수록 해당 유전자에서 변이가 자주 발생했다는 의미입니다.
   * 값의 범위는 0에서 0.5 이상으로, 변이가 거의 발생하지 않은 암종부터 변이가 상대적으로 많이 발생한 암종까지 다양하게 나타나 있습니다.
2. **가로축 (X축):**
   * 암종의 이름이 표시되어 있습니다. 예를 들어, 'KIPAN', 'SARC', 'SKCM', 'GBMLGG' 등 다양한 암종들이 나열되어 있습니다.
   * 각각의 암종에 대해 상자 수염 그래프가 그려져 있으며, 이는 해당 암종에서의 mutation\_freq 분포를 보여줍니다.
3. **각 암종에서의 상자 수염 해석:**
   * **중앙값 (상자의 중간 선):** 각 암종에서 mutation\_freq의 중앙값을 나타냅니다. 대부분의 암종에서 중앙값이 0에 가까운 값으로 나타나며, 이는 대부분의 샘플이 낮은 mutation\_freq 값을 가진다는 것을 의미합니다.
   * **사분위 범위 (상자의 크기):** 각 상자의 상단과 하단은 1사분위수(Q1)와 3사분위수(Q3)를 나타내며, 상자 크기가 클수록 암종 내 변이 빈도 값의 변동성이 큽니다. 예를 들어, **GBMLGG**, **SARC**, **STES**와 같은 암종들은 사분위 범위가 넓어 변이 빈도에 더 큰 변동성이 있음을 나타냅니다.
   * **이상치 (상자의 위아래에 위치한 동그라미):** 상자 밖의 점들은 **이상치**를 나타냅니다. 이는 다른 샘플들과 비교해 변이 빈도가 매우 높거나 낮은 샘플을 의미합니다. 예를 들어, **GBMLGG**, **SARC**, **UCEC** 등에서 이상치가 다수 발견됩니다. 이는 일부 샘플에서 매우 높은 변이 빈도가 나타났다는 뜻입니다.
4. **암종 간 비교:**
   * **GBMLGG**와 **SARC**와 같은 암종은 다른 암종에 비해 변이 빈도가 더 높은 샘플들이 존재함을 보여줍니다. 특히 **GBMLGG**의 경우, 변이 빈도가 0.5에 이르는 샘플이 있습니다.
   * **KIPAN**, **BLCA**, **TGCT**, **DLBC** 등의 암종은 변이 빈도가 매우 낮은 편으로, 대부분의 샘플이 0에 가까운 mutation\_freq를 가집니다.

### **결론:**

이 그래프는 각 암종에서 유전자 변이 빈도의 분포를 시각적으로 보여줍니다. **GBMLGG**, **SARC**, **UCEC** 등의 암종에서는 변이 빈도의 변동성이 크고, 특정 샘플에서는 변이 빈도가 매우 높은 이상치가 존재합니다. 반면, **KIPAN**, **BLCA**, **TGCT** 등은 변이 빈도가 매우 낮은 경향을 보입니다. 이를 통해 특정 암종에서 변이 빈도가 중요한 역할을 할 수 있음을 시사하며, 이는 암종 분류 모델에서 유용한 피처가 될 수 있습니다.

(3) 암종별로 IDH1 유전자 변이 빈도 나타내기

### 

### 

### **그래프 해석:**

1. **가장 변이가 많이 발생한 암종:**
   * **GBMLGG**와 **LGG**에서 IDH1 변이가 가장 많이 발생했습니다. **GBMLGG**는 약 180건 이상의 변이, **LGG**는 약 160건 이상의 변이가 발생했습니다.
   * **GBMLGG**와 **LGG**는 뇌암 또는 신경계 관련 암종으로, IDH1 변이가 이들 암종에서 매우 중요한 역할을 한다는 것을 알 수 있습니다. 이 변이는 이러한 암종의 특이적 바이오마커일 가능성이 높습니다.
2. **다른 암종과의 차이:**
   * **LAML** (급성 골수성 백혈병)에서도 약 25건 이상의 IDH1 변이가 발생했으며, **SKCM**, **UCEC**, **BRCA** (유방암)와 같은 암종에서 소수의 IDH1 변이가 발생했습니다.
   * 그러나 **GBMLGG**와 **LGG**에 비해 대부분의 암종에서는 IDH1 변이가 거의 발생하지 않거나 아주 적게 발생했습니다. 예를 들어, **ACC**, **TGCT**, **THCA**와 같은 암종에서는 IDH1 변이가 거의 없는 것으로 나타납니다.
3. **특이한 변이 패턴:**
   * 이 그래프는 IDH1 변이가 특정 암종에서만 주로 나타난다는 점을 시사합니다. 특히 **GBMLGG**, **LGG**, **LAML** 등의 암종에서 IDH1 변이가 매우 빈번하게 발생하며, 이는 이 변이가 이들 암종에서의 발생과 밀접한 관계가 있음을 의미합니다.
   * 반면에, 다른 암종에서는 IDH1 변이가 거의 나타나지 않으므로, IDH1 변이는 특정 암종에 대한 강력한 예측 신호로 사용될 수 있습니다.
4. **임상적 의미:**
   * IDH1 변이는 주로 **뇌암** 및 **혈액암**에서 자주 발생하는 것으로 보입니다. 따라서, **GBMLGG**나 **LGG**와 같은 암종에서 IDH1 변이 여부를 활용한 분류가 매우 중요한 역할을 할 수 있으며, IDH1 변이가 있는지 여부가 진단에 중요한 정보를 제공할 수 있습니다.
   * **GBMLGG**와 **LGG**에서 IDH1 변이 수가 압도적으로 많기 때문에, 이 변이는 특정 암종의 바이오마커로 활용될 가능성이 큽니다.

### **결론:**

이 그래프는 **IDH1 변이**가 특정 암종에서 특히 자주 발생하며, 특히 **GBMLGG**와 **LGG**에서 매우 중요한 바이오마커일 가능성을 보여줍니다. 이 변이 정보는 암종 분류 모델에서 중요한 피처로 사용될 수 있으며, **뇌암**과 같은 특정 암종의 분류 성능을 향상시킬 수 있습니다.

**8. 다음 전략?**

특정 변이 정보를 바탕으로 유전자 암종을 정확하게 분류하고 성능을 높이기 위해 다음과 같은 전략을 사용할 수 있습니다:

### **1. 주요 유전자 변이 정보 활용**

* **mutation\_freq**와 **mutation\_freq\_ratio**는 전체 변이의 빈도를 반영하는 중요한 피처로 확인되었습니다. 이 피처들은 각 유전자의 변이 빈도를 종합하여 암종을 분류하는 데 중요한 역할을 할 수 있습니다. 이를 통해 각 암종별로 특정 변이가 얼마나 자주 발생하는지를 반영할 수 있습니다.
* 또한, **TP53**와 **PIK3CA** 등의 주요 유전자는 암 발생에 중요한 역할을 하므로, 이러한 유전자의 변이 정보 및 상호작용 정보를 사용하면 성능을 높일 수 있습니다.

### **2. 유전자 상호작용 피처 강화**

* **mutation\_freq\_TP53\_interaction**과 같은 유전자 간 상호작용 피처는 모델이 단일 유전자의 변이보다 더 복잡한 패턴을 학습하도록 도와줍니다. 유전자 간 상호작용은 암종의 발생 및 진행에 중요한 역할을 할 수 있으므로, 더 많은 유전자 상호작용 피처를 추가하여 성능을 향상시킬 수 있습니다.

### **3. 성능 개선을 위한 추가 전략**

* **암종별 클러스터링**: 특정 암종은 유사한 유전자 변이 패턴을 보일 수 있습니다. 예를 들어, GBMLGG와 LGG는 IDH1과 ATRX의 동시 변이가 자주 나타납니다. 이러한 클러스터링 정보를 반영하여 모델이 암종을 더 정확히 분류할 수 있도록 도와줄 수 있습니다.
* **앙상블 모델 적용**: XGBoost뿐만 아니라 LightGBM, CatBoost, RandomForest 등 다양한 모델을 결합한 앙상블 모델을 활용해 성능을 더욱 향상시킬 수 있습니다.
* **하이퍼파라미터 튜닝**: 각 모델의 하이퍼파라미터를 최적화하여 모델 성능을 더욱 극대화할 수 있습니다. GridSearchCV 또는 RandomSearchCV와 같은 방법을 사용할 수 있습니다.

**9. 모델 추천**

-> XGBoost, LightGBM, CatBoost