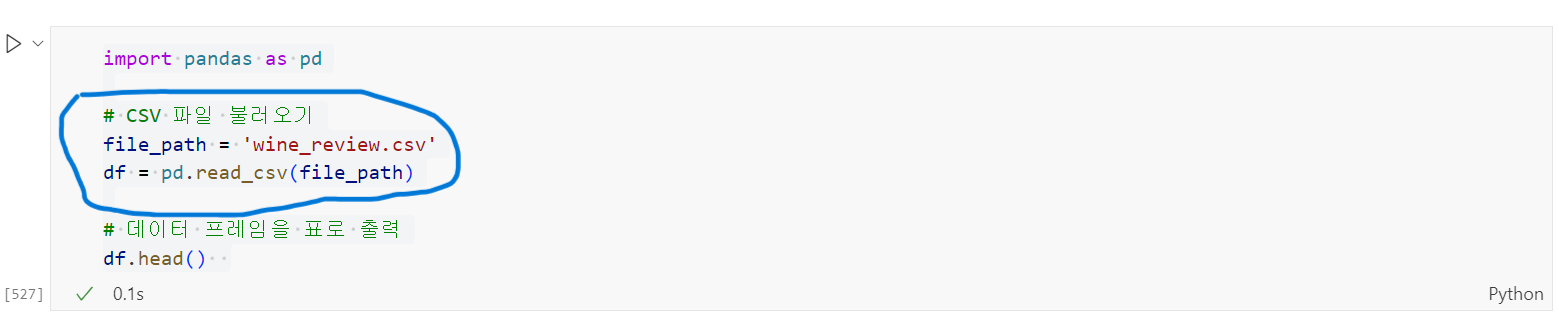
1. **감성 분석의 기본 이해**: 감성 분석이 무엇이며 그 응용에 대해 학습
2. **데이터 전처리**: 머신 러닝 작업을 위해 데이터 전 처리하는 경험을 쌓음
3. **모델 선택 및 학습**: 수업에서 활용한 모델을 활용해서 구현하고 그 강점과 약점을 이해
4. **평가**: 적절한 지표를 사용하여 모델의 성능을 평가
5. **시각화 및 해석**: 결과를 시각화하고 해석

202021018 이해든

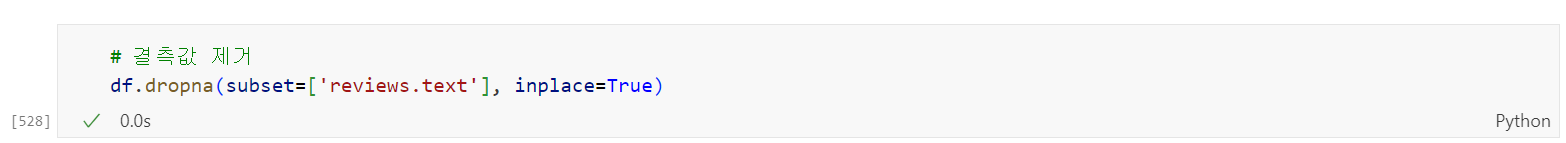
**감성 분석이란?**

감성 분석은 디지털 텍스트를 분석 후 감정적 어조가 긍정적인지, 부정적인지 또는 중립적인지를 확인하는 것이다. 이메일이나 리뷰 등의 텍스트 데이터를 감성 분석으로 글을 쓴 사람의 감정을 분석할 수 있다.

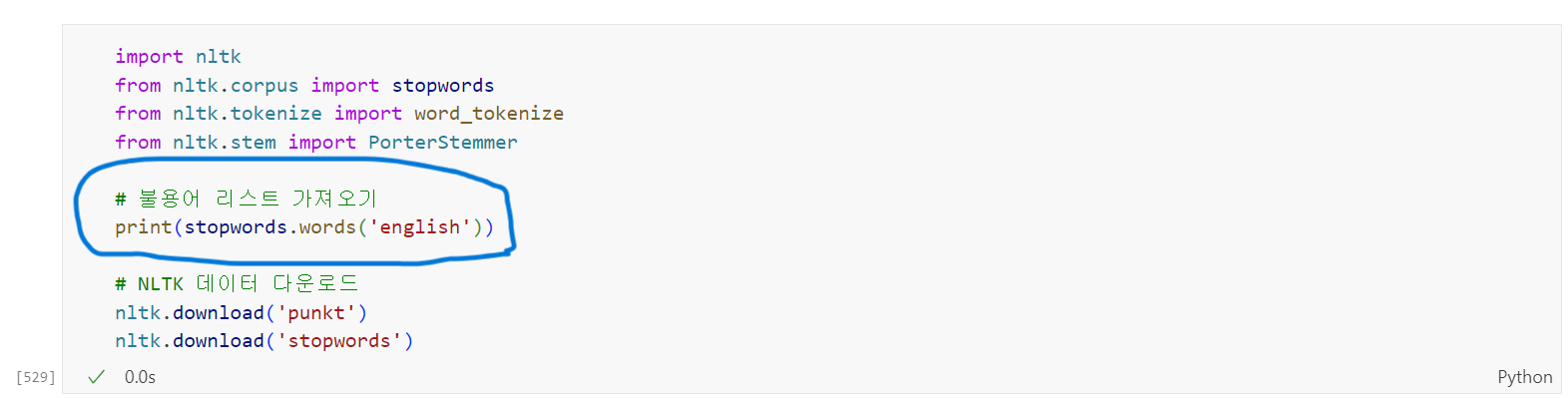
**데이터 전처리**

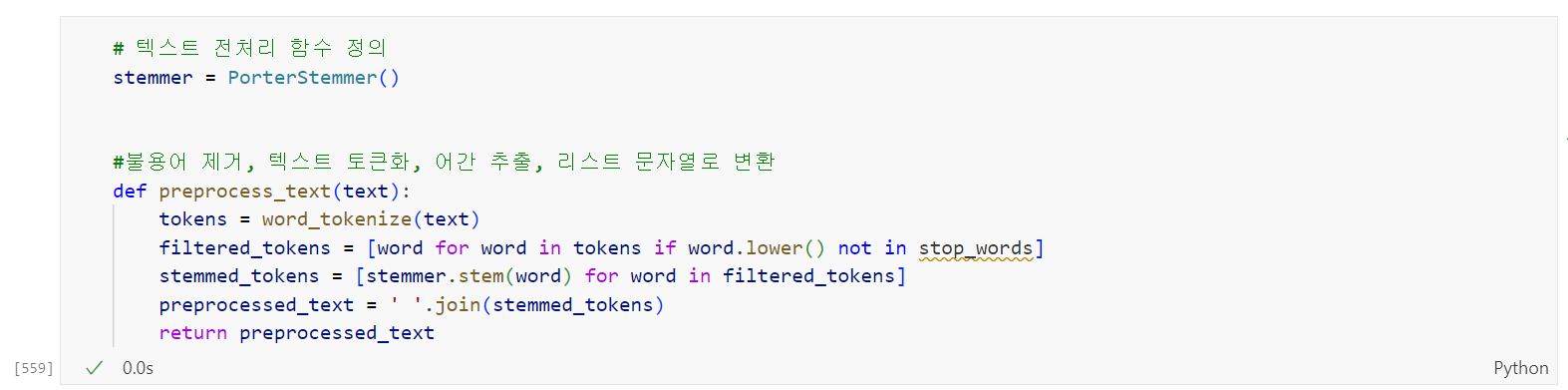


Wine\_review.csv를 데이터 프레임으로 불러온다.



reviews.text 에 결측값을 제거한다. 이후에는 결측값이 없는 데이터로 출력이된다.



불용어 리스트를 가져온다. 

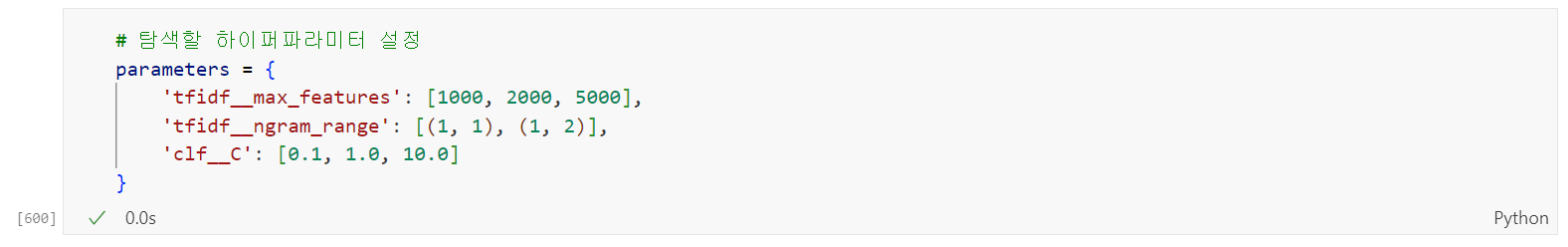
입력 받은 텍스트를 토큰화, 불용어 제거, 어간을 추출하여 전처리된 텍스를 반환한다.

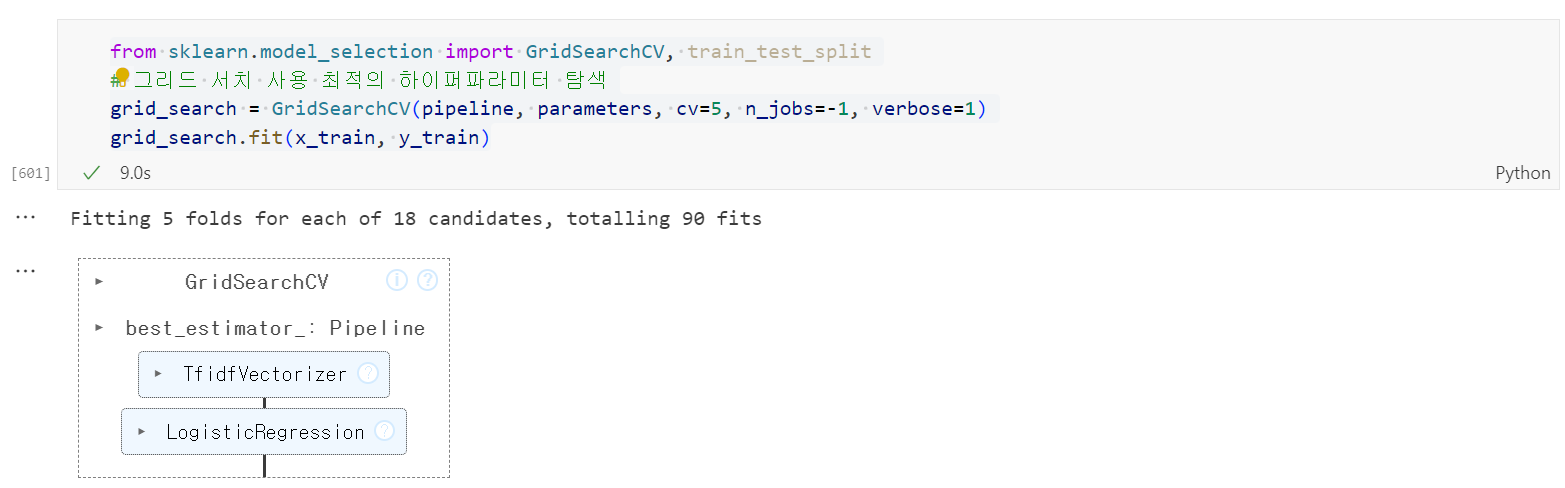


reviews.text 각 행에 preprocess\_text 함수 적용, 전처리 된 텍스트를 cleaned\_text에 저장

* 리뷰 텍스트가 전처리되어 불용어 제거, 어간 추출된 상태로 저장된다.

**모델 구현**





TF-IDF 벡터화, 로지스틱 회귀 모델을 사용하여 감성 분석 모델을 구축하고 학습하였다.

TF-IDF 강점

* 텍스트를 수치화할 수 있다.
* 직관적 해석이 가능하다. (단어 중요도)
* 단어의 상대적 중요성을 고려하여 추전이 가능하다.
* 상대적 단어 빈도를 고려한다.

단점

* 문맥을 완전히 이해하기 어렵다.
* 동일한 단어가 다른 의미를 가지고 있어도 문맥에서 동일한 단어로 처리할 수 있다.

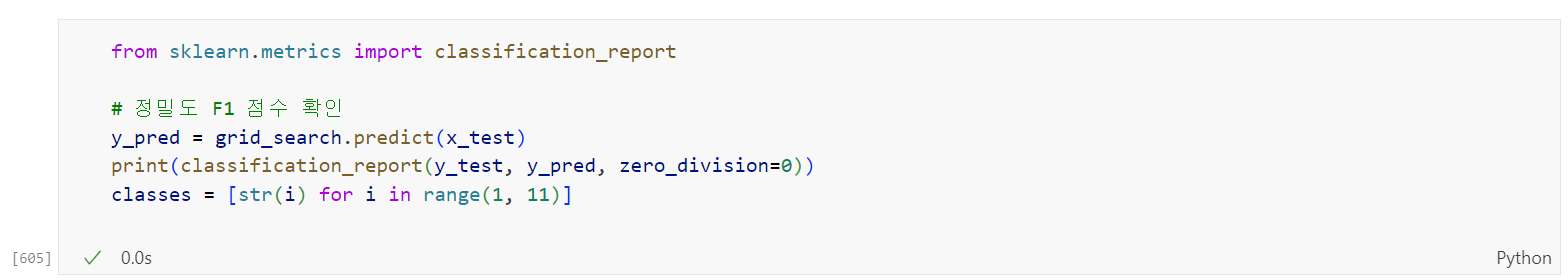
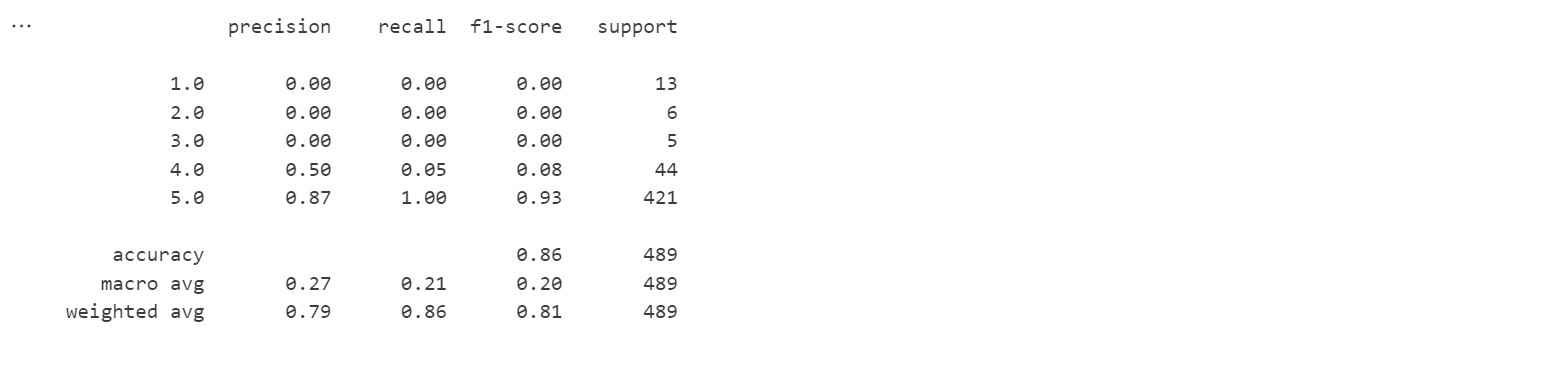
로지스틱 회귀 모델 강점

* 분류만 할 수 있는 모델들과 달리 확률값을 얻을 수 있다.
* 오버피팅을 방지할 수 있다.
* 간단하다.

단점

* 복잡한 비선형 관계 모델링을 하는데 어려움이 있다.
* 상호작용이나 비선형 관계를 직접적으로 처리하지 않는다.

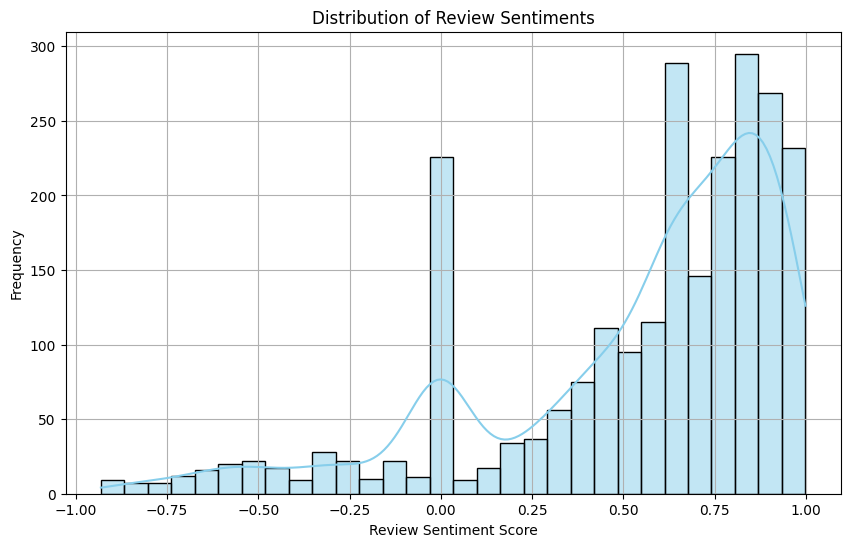
**지표를 사용하여 모델 성능 평가**

Class5.0을 제외하고는 낮은 성능을 보인다. Class4.0에서는 그래도 값을 보여주긴 하지만 1.0, 2.0, 3.0 에서는 정확도, 정밀도, F1\_score 모두 0.00으로 클래스를 예측하는데 어려움이 있는 것으로 보인다.

**시각화 및 해석**

****



리뷰 감정 점수 분포를 시각화한 것이다. 매우 부정부터 매우 긍정적 감정을 나타낸다.

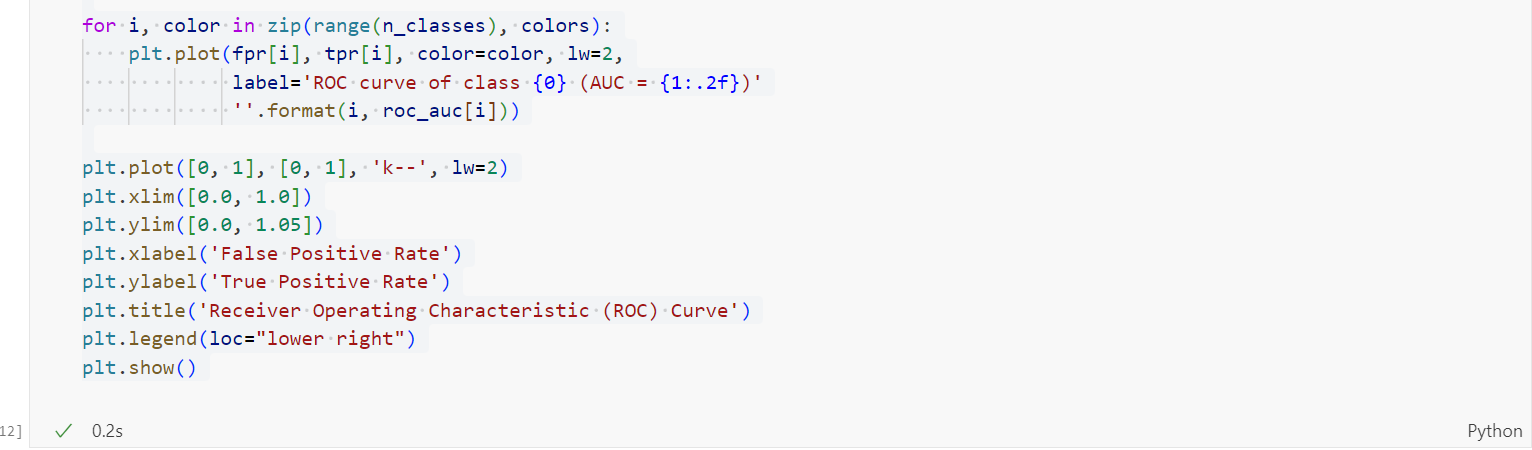
X축: 0.0이 꽤 높게 치솟아 있는 걸 보면 중립적인 감정이 있음을 알 수 있다.

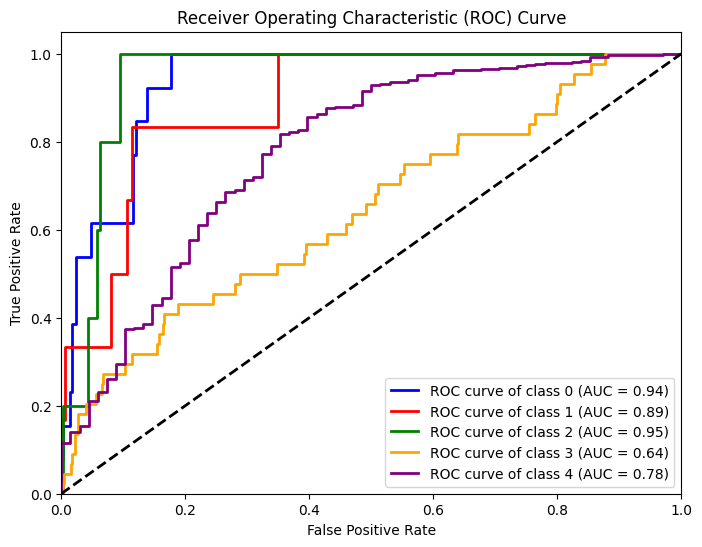
0.5 ~ 1.0 사이에 많은 긍정적인 리뷰가 분포되어 있음을 알 수 있다.

Y축: 특정 감정 점수를 가진 리뷰의 수를 보여준다.

막대의 높이로 해당 구간에 얼마나 많은 리뷰가 있는 지 알 수 있다.

그래프를 보면 중립적인 리뷰도 많지만 긍정적인 점수가 훨씬 많음을 알 수 있고 부정적인 리뷰는 상대적으로 적은 편인 것을 볼 수 있다.



다중 클래스 분류 문제에서 각 클래스에 대한 ROC (Receiver Operating Characteristic) 곡선이다.

AUC 값이 높을 클래스일수록 모델의 예측력이 뛰어남을 의미한다.

클래스 0 (파란색): AUC = 0.94, 높은 성능

클래스 1 (빨간색): AUC = 0.89, 좋은 성능

클래스 2 (녹색): AUC = 0.95, 매우 높은 성능

클래스 3 (주황색): AUC = 0.64, 낮은 성능

클래스 4 (보라색): AUC = 0.78, 중간 정도의 성능

따라서 클래스 2(녹색)의 AUC가 가장 높아 매우 높은 성능을 보여줌을 알 수 있다.