**코로나 보고서**

**학번: 2218087**

**이름: 최 한 비**

**Github address:**

1. **안전 관련 머신러닝 모델 개발의 목적** 
   1. Tab 키를 사용하여 들여쓰기 방식으로 글을 작성할 것
   2. 학습 모델 활용 대상: **세계국가**
   3. 데이터의 어떠한 독립 변수를 사용하여 어떠한 종속 변수를 예측하는 지

**: 종속 변수는 data 데이터 프레임에서 ' Deathrate'가 예측하고자 하는 값을 나타내며, 나머지 열들은 이 값을 예측하는데 사용되는 변수들이다. 'Deathrate'가 종속 변수이며, 나머지 열들은 독립 변수로 설정된다.**

* 1. 개발의 의의

: **국가별 건강 상태를 평가하거나 사망률 추세를 분석하는데 활용될 수 있다. 이를 통해 정부, 의료기관 또는 보건 담당자들이 적절한 조치를 취하고 의사결정을 내릴 때 유용한 정보를 얻을 수 있다. 또한, 머신러닝 모델을 통해 예측 및 분석하는 프로세스 자체도 더 나은 이해와 통찰을 제공하여 향후 비슷한 문제에 대한 접근 방식을 개선하는데 도움이 된다.**

1. **안전 관련 머신러닝 모델의 네이밍의 의미** 
   1. **전 국가의 GDP, 사망율, 출생율 등 모든 나라들의 수치를 나타내며 코로나와 관련된 현황들을 보여준다.**
2. **개발 계획**
   1. 데이터에 대한 요약 정리 및 시각화

: **데이터를 불러와 결측치를 제거하고, 문자열 값을 가진 열을 삭제한 후 데이터를 용약하여 설명 통계량을 계산한다. 시각화를 통해 'Population' 열의 분포를 히스토그램으로 확인한다.**

* 1. 데이터 전처리 계획

**① 결측치 처리: 'dropna()'를 사용하여 결측치가 있는 행들을 삭제한다.**

**② 열 삭제:'Country','Region'열과 같이 문자열 값을 가진 열을 삭제한다.**

* 1. 어떠한 머신러닝 모델을 사용할 것인지 (해당 머신러닝 모델의 이론 추가)

**: RandomForestClassifier를 사용하여 머신러닝 모델을 학습한다. 이 모델은 여러 결정 트리들을 앙상블한 모델이다. 각 결정 트리는 부트스트랩 샘플을 이용하고 무작위로 선택된 특성들에 대해 분할을 수행하여 개별적인 예측을 합친다.**

* 1. 머신러닝 모델 예측 결과가 어떠할 지

**: 학습된 모델을 사용하여 테스트 데이터의 'Dreathrate'를 예측하고, 이를 평가하기 위해 정확도(accuracy)를 계산한다. 성능 평가를 위해 혼동 행렬(confusion matrix)을 생성하고 시각화하여 모델의 성능을 시각적으로 확인한다.**

* 1. 사용할 성능 지표

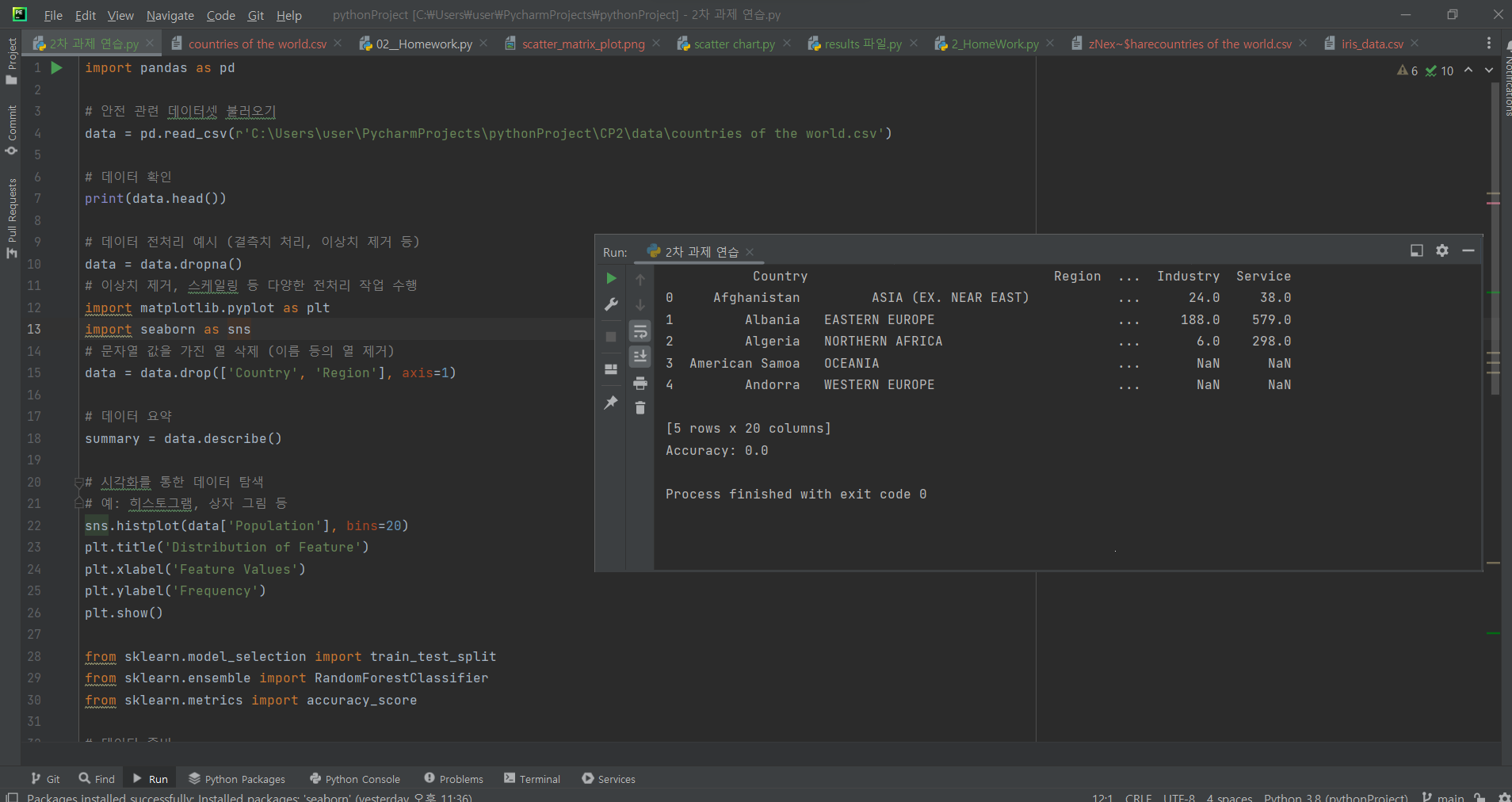
**: 코드에서는 정확도(accuracy)를 사용하여 모델의 성능을 평가하고 있다. 또한, 행렬을 사용하여 모델의 예측 결과를 더 자세히 살펴보고 있다.**

* 1. 성능 검증 방법 계획 등

**: 성능 검증은 학습 데이터와 테스트 데이터를 분리하여 모델을 학습 및 평가하고 있으며, 이때 테스트 데이터의 크기는 전체 데이터의 20%로**

**설정되어 있다.**

1. **개발 과정** 
   1. 계획 후 실제 학습 모델 개발 과정을 기록 (\*개발 과정 캡쳐 필수)



* 1. 각 함수는 어떻게 동작하는 지 구체적으로 설명

**1. 라이브러리 불러오기:**

**\* pandas: 데이터를 다루고 분석**

**\* matplotlib.pyplot과 seaborn: 데이터 시각화**

**\* sklearn.model\_selection의 RandomForestClassfier: 분류를 위한 앙상블 학습 방법**

**\* sklearn.metrics에서 accuracy\_score와 confusion\_matrix: 모델의 성능 평가**

**2. 데이터 불러오기”**

**\* 지정된 파일 경로에서 csv파일을 읽음**

**3. 데이터 검토 및 전처리:**

**\* data.head(): 데이터셋의 처음 몇 줄을 출력**

**\* data = data.dropna(): 결측치가 있는 행 제거**

**4. 데이터 시각화:**

**\* sns.histplot(): Population열의 분포를 시각화**

**5. 데이터 준비:**

**\* 독립변수(x)와 종속변수(y)로 데이터를 분리하고, country, region, population열을 x에서 제거합니다.**

* 1. 에러 발생 지점 및 해결 과정

**: data파일 열에 있는 문자열을 읽지 못해서 오류가 계속 발생하였는데, data.dropna() 결측치 제거를 통해 그 열을 삭제하였다.**

* 1. 학습 모델의 성능 평가

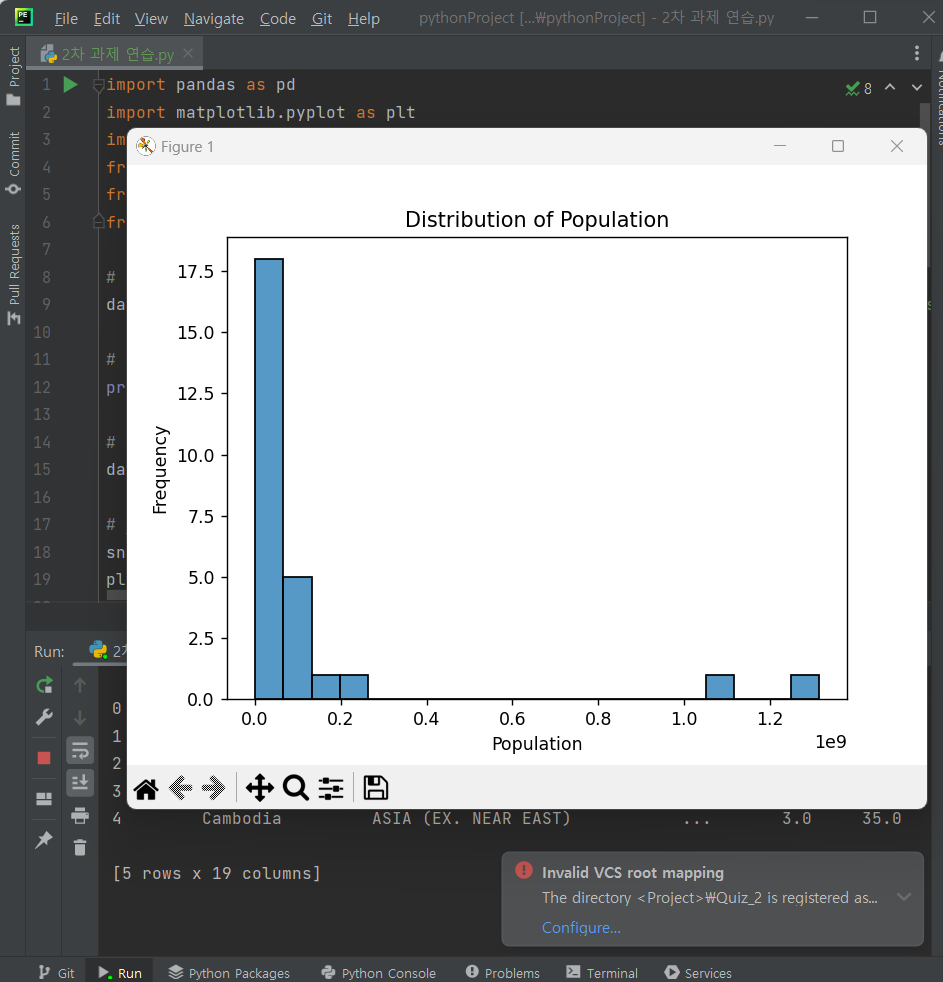
**\* Accuracy(정확도)**

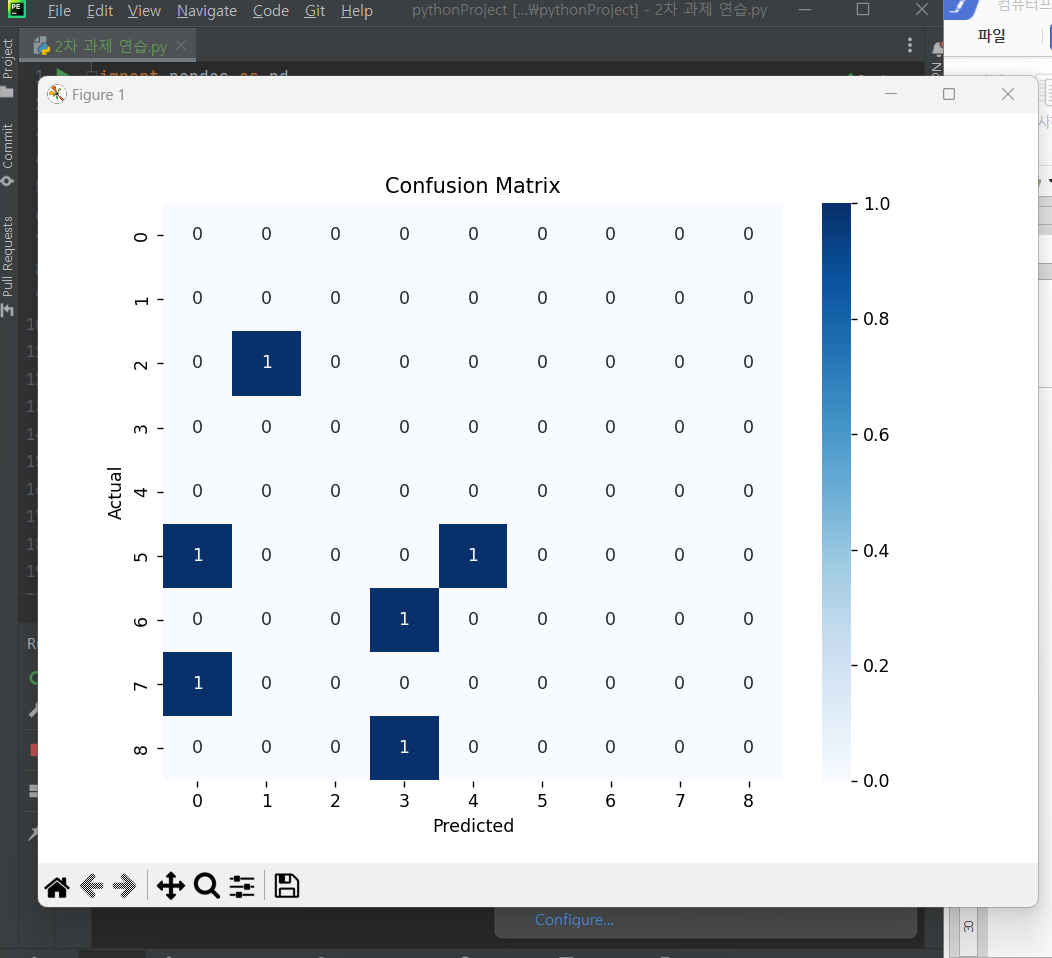
**- 코드에서는 accuracy\_score(y\_test, predictions)'를 사용하여 테스트 세트인 'y\_test'와 모델의 예측값 'predictions'를 비교하여 정확도를 계산한다.**

**\* Confusion Matrix(혼동행렬)**

**- 코드에서는 confusion\_matrix(y\_test, predictions)를 사용하여 실제값 'y\_test'와 예측값 'predictions'사이의 혼동 행렬을 생성한다.**

* 1. 결과 시각화



1. **개발 후기**
   1. 개발 후 느낀 점 설명

**: 데이터를 다뤄야 할 때 데이터의 형태와 내용을 이해하는 것이 중요하다는 것을 알게 되었다. 데이터를 이해하지 못한채 코딩을 했을때 예기치 못한 오류가 많았었다. 그래서 데이터의 특성을 파악하고 전처리 과정에서 결측치나 이상치를 처리하는 것이 중요하다는 것을 알게되었다.**

**모델을 평가하고 성능을 이해하는 것이 중요하다. 정확도와 혼동 행렬을 통해 모델의 예측 능력과 각 클래스에 대한 성능을 이해할 수 있다.**