제 기말 프로젝트의 주제는 잠수함에서 쓰이는 소나파를 분석하여 기뢰인지 암석인지를 구분하는 머신러닝 모델을 만드는 것입니다.

데이터로는 UC Irvine Machine Running Repository의 Sonar, Mines versus Rocks 데이터셋을 사용했고, 알고리즘은 의사결정트리의 한 종류인 랜덤 포레스트를 사용하였습니다.

먼저 필요한 라이브러리를 설치한 후, 필요한 라이브러리를 import하여 데이터의 모양을 살펴보면

데이터는 208개, 각 데이터에는 특징이 각각 60개씩 있다는 것을 알 수 있습니다.

데이터의 값에는 기뢰 즉 물속에 있는 지뢰를 뜻하는 Mines인 M과 암석을 뜻하는 Rocks인 R이 있다는 것을 알 수 있습니다.

데이터의 형태를 살펴보기 위하여 레포지토리에서 다운받은 파일 중 sonar.all-data 파일을 살펴보면, 데이터의 각 특성은 소나 주파수 대역에서 측정된 에너지 값인 0이상 1이하의 실수로 나타내어져 있고, 마지막에는 데이터의 결과값인 M 또는 R이 있다는 것을 알 수 있습니다.

데이터의 결과값인 M과 R을 숫자형으로 변환하고,

데이터를 학습 데이터와 테스트 데이터로 분할합니다.

저는 테스트 데이터의 크기를 10%로 지정했습니다.

랜덤 포레스트 모델을 불러온 후

트리 개수, 트리 깊이 제한, 분할 최소 샘플 개수, 리프 노드 최소 샘플 개수, 각 트리가 사용할 특성 개수, 랜덤 스테이트를 설정하여 학습을 시킵니다.

학습을 시킨 후, 훈련 정확도와 검증 정확도를 출력하면 훈련은 정확히 되어 100%가 된 것을 확인할 수 있고, 검증 정확도는 ?????%입니다.

마지막으로 가시적으로 학습을 확인하기 위하여 그래프를 출력하였습니다.

아무래도 군용이다 보니까 결측값이 없는 데이터셋을 제공했다고 써있음

기뢰 데이터 111개

암석 데이터 97개

다른 사람들 정확도: XGBoost 80% 초반대, Extra Tree 모델 90% 후반대

XGBoost를 사용하였을 때는 검증 정확도가 80%대