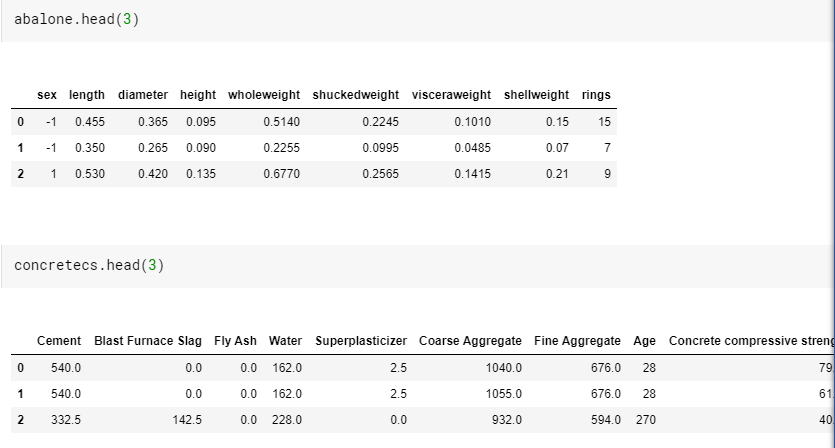
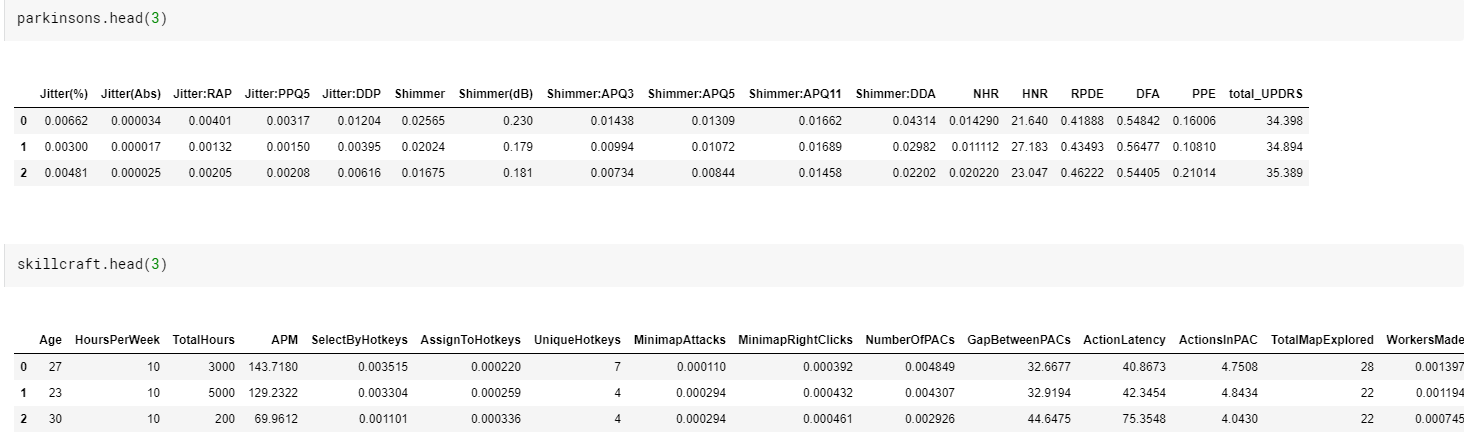
**Assignment 6: AutoML**

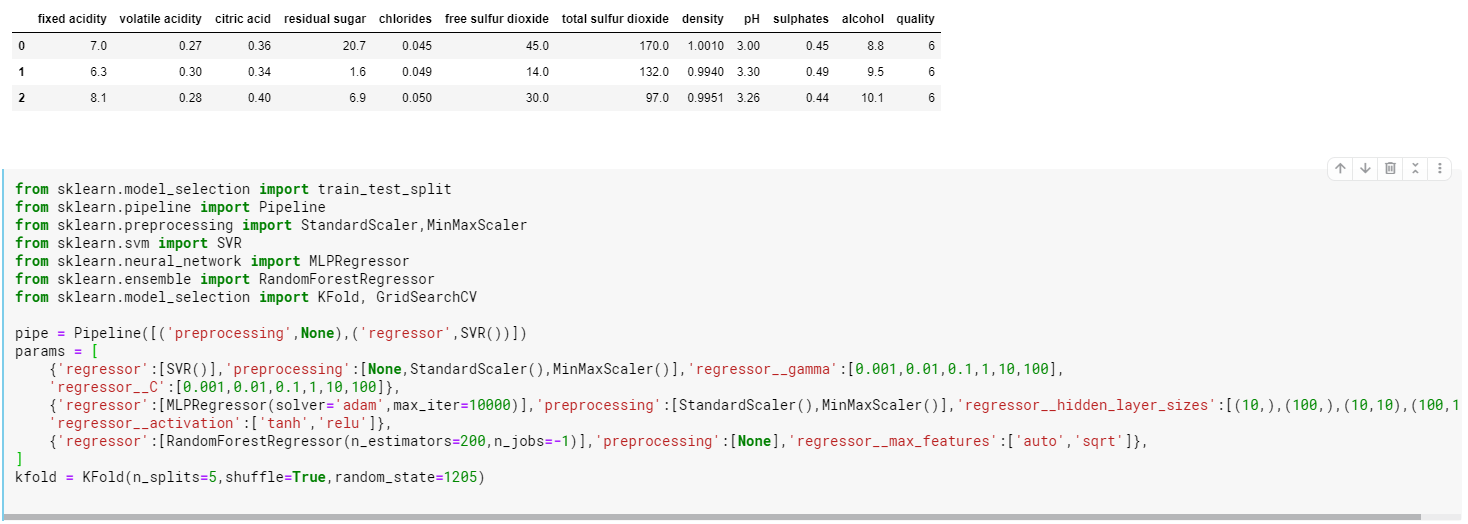
**2016314726 정영준**

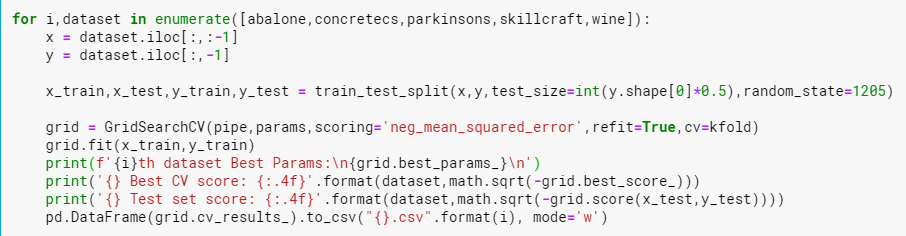
**Jupyter notebook 환경에서 프로그래밍을 하였습니다. 사용한 데이터셋은 abalone, concretecs, skillcraft, winequality-white, parkinsons 데이터셋입니다. 5개의 데이터셋 모두 연속형 label을 가진 회귀를 적용하여야 하는 데이터였습니다. 마지막 열이 label이고 나머지 열은 feature를 나타내고 있어 slicing을 통해 feature인 x와 label인 y로 나누었습니다. 그리고 훈련세트와 테스트세트로 데이터를 나누기 위해 sklearn의 train\_test\_split을 사용하여 test\_size를 절반으로 설정하였습니다. Pipeline과 GridSearchCV를 사용하여 여러 모델과 hyperparameters에 따라 성능을 비교하고 가장 성능이 좋은 모델을 채택하는 코드를 작성하였습니다. 먼저 데이터셋을 모두 확인하여 Categorical feature가 있는지 확인하였습니다. 0과 1로 작성된 성별을 제외하고 numeric feature로 모두 이루어져 있었습니다. Pipeline의 기초 형태를 preprocessing -> regressor로 설정하였습니다. SVR, MLPRegressor, RandomForestRegressor을 사용할 Regressor로 정하였습니다. 세 모델 모두 성능이 크게 떨어지지 않으며 내부 매커니즘이 다르기 때문에 데이터셋에 따라 적절한 모델이 선정될 것이라고 생각하였습니다. 먼저 Preprocessing는 None, Standard Scaler, MinMax Scaler를 사용하였습니다. RandomForestRegressor가 regressor로 사용될 때는 Preprocessing은 None으로 지정됩니다. 이는 scaling에 영향을 받지 않는 모델이기 때문에 Computation cost를 줄이기 위해서 입니다. 다음으로 SVR은 gamma와 C값의 범위는 [0.001, 0.01 ,0.1, 1, 10, 100]로 설정하였습니다. gamma와 C값이 모두 모델의 performance에 영향을 중요하게 끼치기 때문입니다. MLP 회귀 모델에서는 solver는 adam으로 고정하였습니다. 여러 데이터셋에서 모두 gradient descent가 충분히 루어지도록 max\_iter를 10000으로 설정하였습니다. Scale에 따라 영향을 크게 받기 때문에 Preprocessing 후보에서 None을 제거하였습니다. 활성화 함수는 [relu, tanh]로 정하였고 은닉층의 size는 [(10,), (100,), (10,10), (100,10)]으로 다양한 형태의 신경망에서 모델의 성능을 비교할 수 있도록 하였습니다. 마지막으로 RandomForesetRegressor의 경우 n\_estimators를 200으로 고정하고 max\_feature를 [‘auto’, ‘sqrt’]로 설정하였습니다. 5개의 fold에서 cross validation이 이루어지도록 하였습니다. Abalone 데이터셋에서 standard scaler를 사용한 activation=tanh, hidden\_layer-size=(100,10)으로 설정된 MLP 회귀모델이 최고의 성능을 보였으며 test set에서 2.0593의 RMSE를 기록하였습니다. Concretecs 데이터셋에서는 scaling을 하지 않은 max\_feature=auto로 설정된 RandomForestRegressor가 가장 좋은 모델이었고 5.6110의 RMSE를 test set에서 기록하였습니다. Parkinsons 데이터셋에서는 MinMax scaler를 사용하고 C=100, gamma=10인 SVR이 최적 모델이며 8.7620의 test set RMSE를 기록하였습니다. Skillcraft 데이터셋에선 max\_features=sqrt인 RandomForestRegressor가 0.0002의 test set RMSE를 보였습니다. 마지막으로 winequality데이터셋에서도 마찬가지로 max\_features=sqrt인 RandomForestRegressor가 0.6414의 RMSE를 기록하였습니다. SVR과 MLP에서 hyperparameters의 범위를 다시 조정하고 feature와 label의 분포를 확인하고 log transform을 적용하는 개선이 이루어지면 더 성능 좋은 모델을 얻을 수 있습니다.**

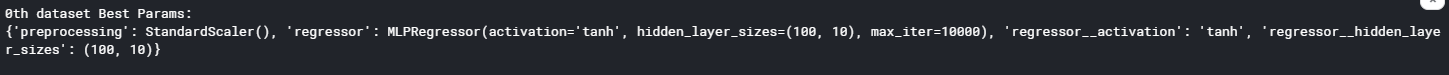










Abalone output





Concretecs output



Parkinsons output







Skillcraft output







Winequality-white output





