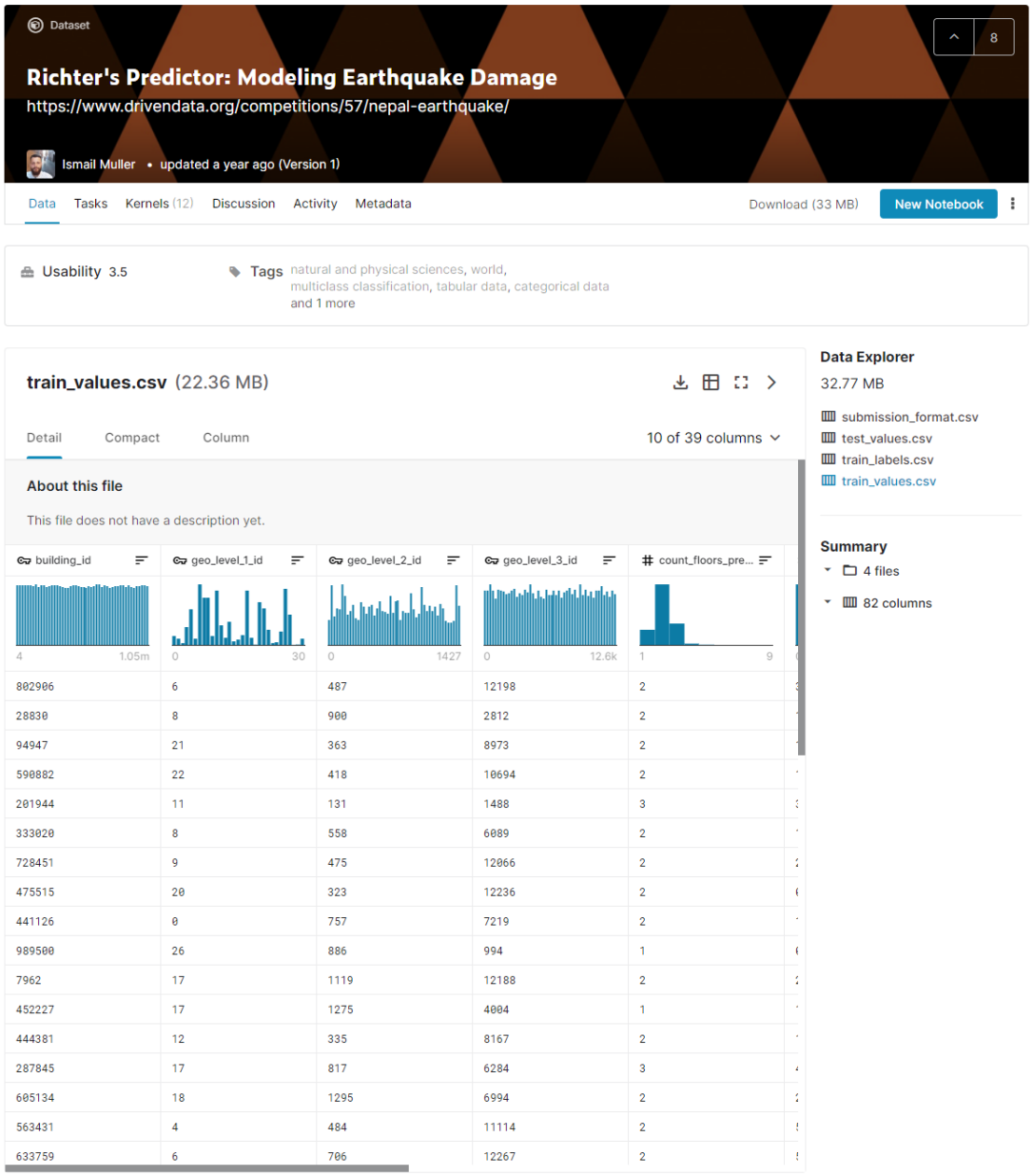
Multivariate Data Analysis Assignment #5

Artificial Neural Network for Classification

Dataset: Richter's Predictor: Modeling Earthquake Damage

(https://www.kaggle.com/mullerismail/richters-predictor-modeling-earthquake-damage?select=train_values.csv)

해당 데이터셋은 총 260,602개의 건물에 대한 지진 피해 정도를 기록한 데이터셋이다. 첫 번째 컬럼인 building_id는 각 건물의 일련번호이며, 마지막 컬럼인 damage_grade는 해당 건물의 지진 피해 정도로써 분류모형의 종속변수로 사용된다. 이 외 모든 컬럼은 입력변수이다.



전체 데이터 셋에 대해 다음 물음에 답하시오.

[Q1] 입력 변수의 속성이 numeric 이 아닌 변수들을 모두 확인하고, 각 변수들의 요인 값들에 대한 Bar chart 를 도시하시오.

[Q2] [Q1]에서 선택된 변수들에 대해 1-of-C coding (1-hot encoding) 방식을 통해 명목형(요인형) 변수를 범주의 개수만큼의 이진형(binary) 변수들로 구성되는 dummy variable 을 생성하시오.

전체 데이터셋을 임의로 150,000 개의 빌딩이 포함된 Training dataset 과 50,000 개의 Validation dataset, 그리고 60,602 개의 Test dataset 으로 구분한 뒤 다음 각 물음에 답하시오. 분류 성능을 평가/비교할 때는 3-class classification 의 Accuracy 와 Balanced Correction Rate (BCR)을 이용하시오.

[Q3] Training 및 Validation 데이터셋을 바탕으로 실습 시간에 사용한 "nnet" package 를 사용하여 다음 hyper-parameter 조합들에 대한 Grid search 를 수행하여 최적의 조합을 찾아보시오(합계 최소 21 가지 이상 탐색). 아래 두 가지 이외의 hyper-parameter 는 nnet package 의 default value 를 사용하시오. 성능이 가장 우수한 조합과 가장 열등한 조합과의 Accuracy 와 BCR 의 차이는 얼마인가?

- Number of hidden node: 최소 7 가지 이상 탐색
- maxit: 최소 3 가지 이상 탐색

[Q4] [Q3]에서 선택된 최적의 모델 구조에 nnet의 rang 옵션(default value = 0.7)을 세 가지 이상 변경해 가면서 성능 변화를 살펴보시오. [Q3]의 방식과 마찬가지로 Training/Validation 데이터셋을 사용하여 탐색하시오.

[Q5] [Q3]에서 확인된 최적의 hidden node의 수와 최적의 maxit, 그리고 [Q4]에서 확인된 가장 우수한 rang option으로 ANN 모델 구조를 확정하고 Training dataset과 Validation dataset을 결합한 데이터셋에 대해 학습을 수행하시오. 학습이 완료된 ANN 모델을 Test dataset에 적용하여 Accuracy와 BCR을 살펴보시오. 동일한 과정을 10회 반복 수행하여 수행 회차에 따라 정확도의 변동성은 어느 정도 되는지 확인하시오.

[Q6] [Q5]에서 사용된 모델 하이퍼파라미터를 고정시킨 상태에서 Genetic Algorithm을 이용한 변수 선택 모듈을 추가로 작성하여 최적으로 선택된 변수 조합을 확인해 보시오. Genetic Algorithm에서 변경 가능한 옵션의 경우 본인의 Assignment 3에서 설정했던 최적 값들을 그대로 사용하시오. Chromosome의 초기값을 다르게 설정하고 변수 선택을 3회 수행한 뒤, 동일한 변수 조합들이 선택되는지, 아니면 반복 회차마다 다른 변수들이 선택되는지 확인해보시오.

[Q7] [Q6]에서 GA를 통해 한번 이상 주요 변수로 선택된 모든 변수들을 사용하여 나머지 설정은 [Q5]와

동일하게 하여 Test dataset에 대한 Accuracy와 BCR을 확인하고 모든 변수를 사용한 모델인 [Q5]의 결과와 비교해 보시오.

[Q8] 동일한 Training/Validation 데이터셋을 이용하여 최적의 의사결정나무 하이퍼파라미터 조합을 탐색해 보시오. 이후 최적의 하이퍼파라미터 조합을 이용하여 Training/Validation 데이터셋을 결합한 데이터셋으로 의사결정나무를 학습한 뒤, Test Dataset에 대한 Accuracy와 BCR을 산출해 보시오.

[Q9] Training dataset과 Validation dataset을 결합하여 Multinomial logistic regression을 학습하고, 이를 Test dataset에 적용하여 Accuracy와 BCR을 산출하고 이를 [Q5], [Q7], [Q8]의 결과와 비교해보시오. (Hint: 아래와 같은 표를 작성해보시오)

Test Performance	ANN (all variables)	ANN (Selected variables by GA)	Decision Tree	Multinomial Logistic Regression
Accuracy				
BCR				