Multivariate Data Analysis Assignment #4

Decision Tree for Classification

(준비사항) 본인의 과제 2에서 사용했던 데이터셋

[Q1] 본인이 생각하기에 “예측 정확도”도 중요하지만 “예측 결과물에 대한 해석”이 매우 중요할 것으로 생각되는 분류 문제를 다루고 있는 데이터셋을 1개 선정하고 선정 이유를 설명하시오. 데이터셋 탐색은 아래에서 제시된 Data Repository를 포함하여 여러 Repository를 검색해서 결정하시오. 보고서에는 데이터를 다운로드할 수 있는 링크를 반드시 제공하시오.

- Kaggle Datasets: https://www.kaggle.com/datasets
- UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/index.php
- 공공 데이터 포털: https://www.data.go.kr/

(가이드라인) 두 가지 데이터셋에 대해서 학습:검증:테스트 용도로 적절히 분배하시오(예: 60:20:20). 본인이 분배한 비율에 대해서 간략히 그 근거를 설명하시오. 분류 성능을 평가/비교할 때는 TPR, TNR, Precision, Accuracy, BCR, F1-Measure, AUROC 를 복합적으로 고려하여 의견을 서술하시오.

(두 가지 데이터셋에 대해서 아래 질문들을 각각 수행하시오)

[Q2] 실습 시간에 사용한 "tree” package 를 사용하여 Classification Tree 를 학습한 결과물을 Plotting 하고 이에 대한 해석을 수행하시오. 또한 해당 Tree 를 pruning 을 수행하지 않은 상태에서 Test dataset 에 대한 분류 성능을 평가하시오.

[Q3] 앞에서 생성한 Tree에 대해서 적절한 Post-Pruning을 수행한 뒤 결과물을 Plotting하고 이에 대한 해석을 수행하시오. Pruning 전과 후에 Split에 사용된 변수는 어떤 변화가 있는가? Test dataset에 대한 분류 성능을 평가하고 [Q2]의 결과와 비교해보시오.

[Q3] "party” package를 사용하여 다음 조건에 맞는 Pre-pruning을 수행하여 가장 최적의 min_criterion, min_split, max_depth 값을 찾아보시오.

(조건 1) 각 하이퍼파라미터들은 최소 5 가지 이상의 후보 값을 가질 것(총 125가지 이상)

(조건 2) 평가 지표는 Validation Dataset에 대한 AUROC 사용

[Q4] 최적의 결정나무의 Plot을 그리고, 대표적인 세 가지 규칙에 대해서 설명해보시오.

[Q5] [Q4]에서 선택한 하이퍼라라미터 조합을 이용하여 Training Dataset과 Validation Dataset을 결합한 데이터셋을 학습한 뒤, Test Dataset에 적용해보고 분류 성능을 평가하시오.

[Q6] 과제 2를 수행하기 위해 사용된 데이터셋(Dataset 1)과 이번 과제 수행을 위해 선택된 데이터셋(Dataset 2)에 대해서 각각 로지스틱 회귀분석을 수행하여 Test Dataset에 대한 다음 Confusion Matrix를 채우고 이에 대한 결과를 해석해 보시오.

Dataset	Model	TPR	TNR	Accuracy	BCR	F1-Measure
Dataset 1	Logistic Regression					
	Decision Tree					
Dataset 2	Logistic Regression					
	Decision Tree					

[Q7] 각 데이터셋마다 Logistic Regression에 의해 중요하다고 판별된 변수들과 의사결정나무에 의해 중요하다고 판별된 변수들을 확인해보고 차이가 있는지의 여부와, 차이가 존재할 경우 그 이유에 대한 본인의 생각을 서술해 보시오.