Multivariate Data Analysis Assignment #3

Decision Tree & Neural Network

[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) 해당 데이터셋에 대해서 학습:검증:테스트 용도로 적절히 분배하시오(예: 60:20:20). 본인이 분배한 비율에 대해서 간략히 근거를 설명하시오. 분류 성능을 평가/비교할 때는 TPR, TNR, Precision, Accuracy, BCR, F1-Measure, AUROC를 복합적으로 고려하여 서술하시오.

[Q2] (Decision Tree) 아래 두 가지의 경우에 대한 테스트 데이터셋에 대한 분류 성능을 평가하고 그 결과를 비교해보시오.

1) 학습 데이터만을 이용해서 학습한 Full Tree
2) 학습 데이터를 사용하여 학습한 후 검증 데이터를 사용하여 Post-pruning 을 수행한 Tree

[Q3] (Decision Tree) 학습 데이터와 검증 데이터를 이용하여 Pre-pruning 을 수행해보시오. Pre-pruning 을 수행하기 위해 사용된 하이퍼파라미터를 설명하고, 각 하이퍼파라미터마다 탐색 범위를 어떻게 설정했는지 서술하시오. 검증 데이터에 대한 AUROC 를 기준으로 최적의 하이퍼파라미터 조합을 찾아보시오.

[Q4] (Decision Tree) [Q2]와 [Q3]에서 생성한 Post-pruning 모델과 Pre-pruning 모델의 결과물을 각각 Plotting하고 이에 대한 해석을 수행하시오. 각 Pruning 방식에 따라 Split에 사용된 변수는 어떤 변화가 있는가?

[Q5] (Decision Tree) 최적의 결정나무의 Plot을 그리고, 대표적인 세 가지 규칙에 대해서 설명해보시오.

[Q6] (Neural Network) 동일한 데이터셋에 대하여 Neural Network 학습을 위해 필요한 최소 3가지 이상의 하이퍼파라미터를 선정하고, 각 하이퍼파라미터마다 최소 3개 이상의 후보 값(최소 9가지 조합)을 사용하여 grid search를 수행한 뒤, 검증데이터에 대한 AUROC 기준으로 최적의 하이퍼파라미터 조합을 찾아보시오.

[Q7] (Decision Tree/Neural Network 공통) [Q3]에서 선택한 최적의 Pre-pruning Decision Tree 모델과 [Q6]에서 선택한 최적의 Neural Network, 그리고 로지스틱 회귀분석을 사용하여 학습 데이터를 학습한 뒤, 테스트 데이터에 적용한 결과를 아래의 Confusion Matrix와 같이 작성하고 이에 대한 결과를 해석해 보시오.

Dataset	Model	TPR	TNR	Accuracy	BCR	F1-Measure
Dataset Name	Logistic Regression					
	Decision Tree					
	Neural Network					

[Q8] 이번에는 본인이 생각하기에 "예측 정확도"가 "예측 결과물에 대한 해석"보다 훨씬 더 중요할 것으로 생각되는 분류 문제를 다루고 있는 데이터셋을 1개 선정하고 선정 이유를 설명하시오. 이 외 가이드라인은 [Q1]의 가이드라인과 동일합니다.

[Q9] (Decision Tree/Neural Network 공통) [Q8]에서 선택한 데이터셋을 사용하여 [Q3]에서 수행한 최적의 pre-pruning Decision Tree 모델 찾기, [Q6]에서 수행한 최적의 Neural Network 모델 찾기를 동일하게 수행하시오.

[Q10] (Decision Tree/Neural Network 공통) [Q8]에서 선택된 최적의 Pre-pruning Decision Tree 모델과 최적의 Neural Network, 그리고 로지스틱 회귀분석을 사용하여 학습 데이터를 학습한 뒤, 테스트 데이터에 적용한 결과를 아래의 Confusion Matrix와 같이 작성하고 이에 대한 결과를 해석해 보시오. 데이터셋 선정 당시 본인의 예상과 [Q7]의 결과표 및 아래 결과표가 일치하는지 확인해보시오. 일치하지 않는다면 왜 일치하지 않는지 그 이유를 서술해보시오(일치 여부가 평가 점수에 영향을 미치지 않음).

Dataset	Model	TPR	TNR	Accuracy	BCR	F1-Measure
Dataset Name	Logistic Regression					
	Decision Tree					
	Neural Network					