왜 이 데이터 셋을 골랐는지

1. missing data가 많다 -> 이를 처리해주는 preprocessing에 대해 공부 가능

2. 생소한 수의학 분야

3. 칼럼의 개수가 많고,,,

Baseline

Horse Colic 데이터 셋은 missing data가 많은 데이터 셋이므로, NaN을 먼저 처리해 주어야합니다. NaN의 data type이 String인 경우에는 mode를, Numeric인 경우에는 median을 이용하여 NaN을 채웠습니다. 보다 정확한 학습 결과를 위해 reindex를 해준 이후, outcome을 제외한 String type의 칼럼들에 대해 one\_hot 인코딩을 해주었습니다. 종속변수인 outcome의 경우에는 live, died, euthanized에 대해 각각 1, 0, -1의 값을 가지도록 Label 인코딩을 해주었습니다.

<Classification Models>

사용한 모델과 Hyper-parameter은 다음과 같습니다.

1) OneR (based on Decision tree with depth 1)

2) K-Nearest Neighbors

K:3, k:5

3) Linear SVM (Kernel: Linear)

C (Penalty parameter of the error term): 0.025

4) RBF SVM (Kernel: RBF)

C (Penalty parameter of the error term): 0.025

5) Decision Tree

Max-Depth: 10

6) Random Forest

Max-Depth: 10, number of estimators: 300

7) Neural Net

Alpha (L2 Regularization term): 1, hidden layer: (100,50)

8) AdaBoost (Ensemble based on the Decision Tree)

Number of estimators: 300

9) Naive Bayes

Default

표

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1 Score |
| Linear SVM |  |  |  |  |
| AdaBoost |  |  |  |  |
| Neural Network |  |  |  |  |
| Decision Tree |  |  |  |  |
| Random Forest |  |  |  |  |
| Nearest Neighbors |  |  |  |  |
| OneR |  |  |  |  |
| RBF SVM |  |  |  |  |
| Zero R |  |  |  |  |
| Naïve Bayes |  |  |  |  |

<Evaluation>

Cross Validation을 이용하여 각 모델의 성능을 평가하였습니다.

// cv를 선택한 이유도 써야할까