<예측 애널리틱스 과제#6>

2014170852 산업경영공학부 조영관

<data set>

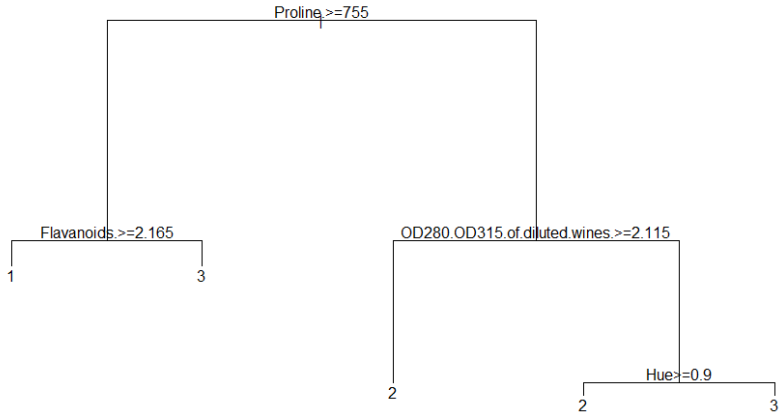
3 종류의 wine을 분류한 data

13개의 입력변수와 1개의 출력변수(Y)가 존재한다.

<전체 데이터 이용 -> 의사결정나무 모델 구축>

데이터 셋 전체를 이용해서 의사결정나무 모델을 구축하였다.

그 결과 다음과 같다.




```
Classification tree:
rpart(formula = Class ~ ., data = data, method = "class")

Variables actually used in tree construction:
[1] Flavanoids.                   Hue
[3] OD280.OD315.of.diluted.wines. Proline.

Root node error: 107/178 = 0.60112

n= 178

        CP nsplit rel error  xerror     xstd
1 0.495327      0   1.00000 1.00000 0.061056
2 0.317757      1   0.50467 0.46729 0.056040
3 0.056075      2   0.18692 0.27103 0.046047
4 0.028037      3   0.13084 0.19626 0.040222
5 0.010000      4   0.10280 0.20561 0.041037
```

위에는 시각화 한 것이고, 아래는 결과값에 대한 설명이다.

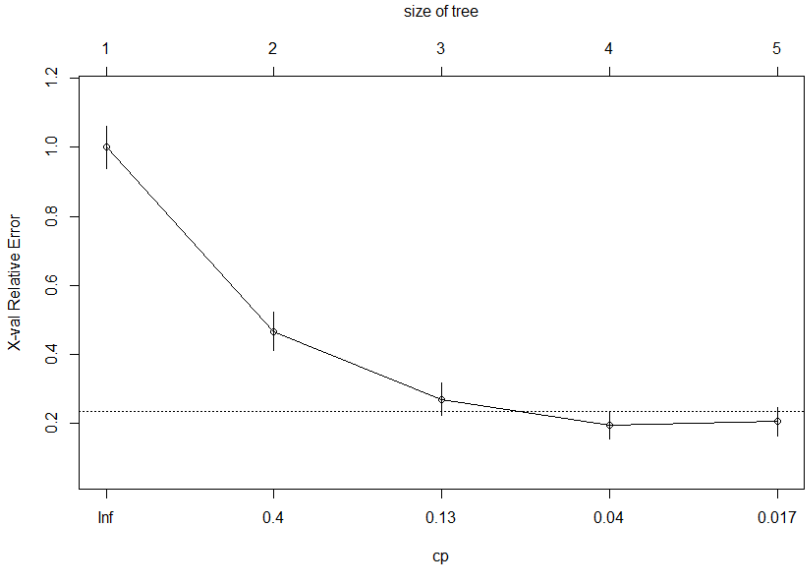
분류를 함에 있어 실질적으로 사용한 변수는 위 4개다. Flavanoids, Hue, OD280.OD315.of.diluted.wines, Proline 이렇게 4가지다.

root node의 분류 error는 0.60112, 즉 root node에서 107개의 observation의 분류가 정확하지 않다.

결과 값 아래의 1,2,3,4,5는 split 개수에 다른 xerror값의 변화를 보여준다.

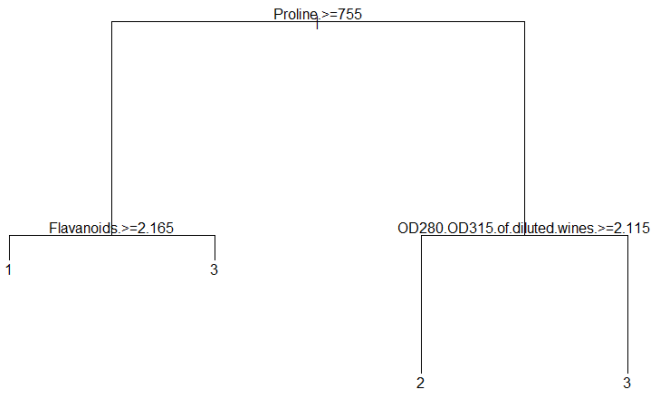
Xerror 값이 가장 작은 경우가 최적의 분류 모델이다.

따라서 4개의 split이 제일 최적의 분류이다. 시각화 하면 다음과 같다.



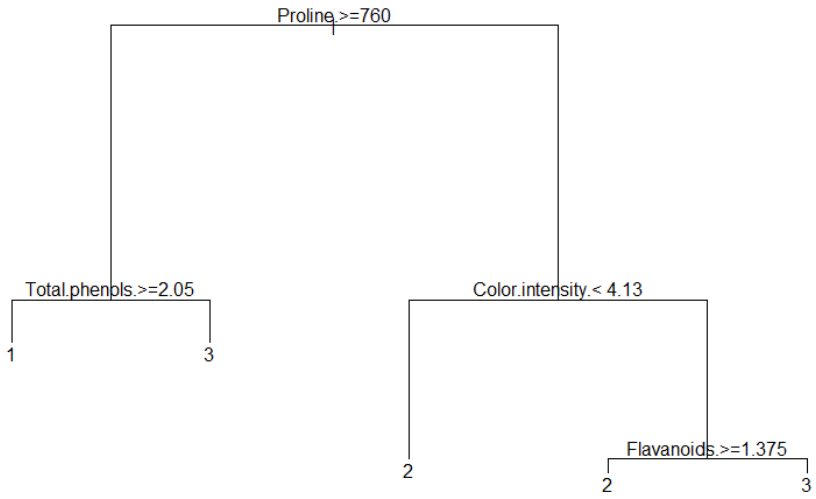
위 결과에 따라 가지치기를 진행한다. (Pruning)

그 결과 최종 구축 모델은 다음과 같다.



<Training data를 이용해 의사결정나무 모델 구축, test data를 이용해 분류 정확도 구하기>

7대 3으로 training set과 test set을 분류하였다. 그리고 의사결정나무 모델을 구축하였다.



Training set을 통해 학습한 모델 결과는 위와 같았다.

Test set을 이용해 모델의 성능을 예측해보았다.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 12  0  0
         2  5 21  0
         3  0  0 14

Overall Statistics

               Accuracy : 0.9038
                 95% CI : (0.7897, 0.968)
    No Information Rate : 0.4038
    P-Value [Acc > NIR] : 6.531e-14

                  Kappa : 0.8521
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.7059   1.0000   1.0000
Specificity            1.0000   0.8387   1.0000
Pos Pred Value         1.0000   0.8077   1.0000
Neg Pred Value         0.8750   1.0000   1.0000
Prevalence             0.3269   0.4038   0.2692
Detection Rate         0.2308   0.4038   0.2692
Detection Prevalence   0.2308   0.5000   0.2692
Balanced Accuracy      0.8529   0.9194   1.0000
```

위 matrix를 보면 예측 값이 잘 분류되었음을 확인할 수 있다.

정확도 accuracy도 0.9038로 높은 편이다.

p-value도 유의수준 0.05보다 작아서 유의미한 결과이다.

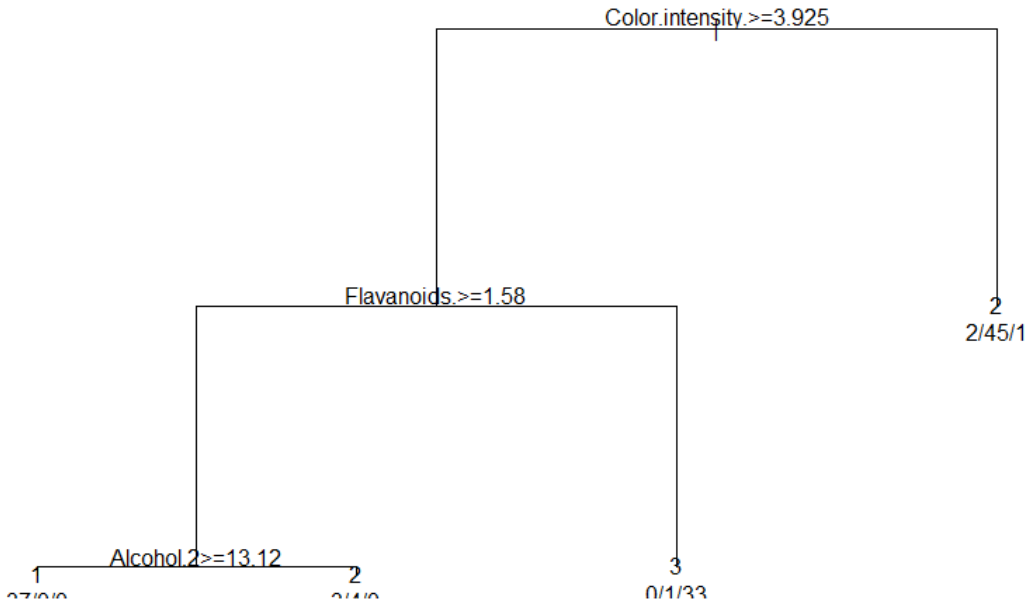
각각의 class에 관한 balanced accuracy도 0.8529, 0.9194, 1 로 높은 편이다.

즉, 예측 성능이 좋은 모델임을 확인할 수 있다.

<비용함수가 변화함에 따른 결과의 차이>

Gini를 사용했을 때와 information gain 비용 함수를 사용했을 때를 비교해보겠다.

먼저 gini를 사용했을 때의 의사결정나무 결과를 확인하면 다음과 같다.



왼쪽부터 순서대로 type A, B, C, B로 분류된다. (다수가 속한 곳으로)

CONFUSION MATRIX를 통해 성능을 확인해보자.

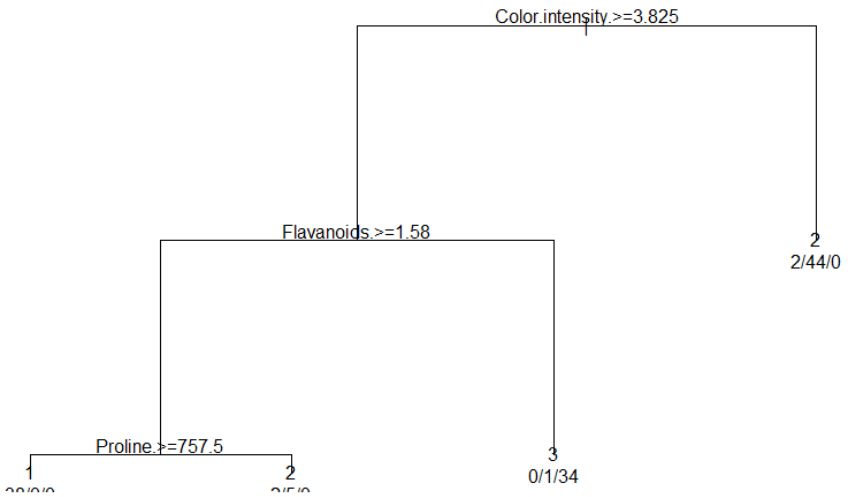
```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 12  0  0
         2  5 21  0
         3  0  0 14

Overall Statistics

               Accuracy : 0.9038
                 95% CI : (0.7897, 0.968)
    No Information Rate : 0.4038
    P-Value [Acc > NIR] : 6.531e-14

                  Kappa : 0.8521
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.7059   1.0000   1.0000
Specificity            1.0000   0.8387   1.0000
Pos Pred Value         1.0000   0.8077   1.0000
Neg Pred Value         0.8750   1.0000   1.0000
Prevalence             0.3269   0.4038   0.2692
Detection Rate         0.2308   0.4038   0.2692
Detection Prevalence   0.2308   0.5000   0.2692
Balanced Accuracy      0.8529   0.9194   1.0000
```

Accuracy는 0.9038 값이 나왔다.

p-value도 유의수준 0.05 보다 작으므로 유의미한 결과다.

마지막으로 balanced accuracy 값도 각각의 class에서 충분히 큰 값이 나온다.

다음으로 information gain을 이용하여 의사결정나무 모델을 추출한 결과를 확인해보자.



분할 값에서 약간의 차이가 있지만 gini와 거의 유사함을 알 수 있다.

왼쪽부터 순서대로 type이 A, B, C, B로 분류된다. (GINI 와 동일)

CONFUSION MATRIX를 통해 성능을 확인해보자.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 15  0  0
         2  2 21  0
         3  0  0 14

Overall Statistics

               Accuracy : 0.9615
                 95% CI : (0.8679, 0.9953)
    No Information Rate : 0.4038
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9412
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.8824   1.0000   1.0000
Specificity            1.0000   0.9355   1.0000
Pos Pred Value         1.0000   0.9130   1.0000
Neg Pred Value         0.9459   1.0000   1.0000
Prevalence             0.3269   0.4038   0.2692
Detection Rate         0.2885   0.4038   0.2692
Detection Prevalence   0.2885   0.4423   0.2692
Balanced Accuracy      0.9412   0.9677   1.0000
```

ACCURACY 값이 0.9615로 GINI보다 높게 나온다.

따라서 위 데이터 셋은 ENTROPY 비용함수 최소화 방법을 통해 의사결정나무의 가지를 뻗어 나가는 것이 더 효과적이다.

P-VALUE도 유의수준 0.05에서 충분히 작으며 BALANCED ACCURACY도 각각의 CLASS에서 값이 충분히 크다.