Predictive Analytics Team Project

Poisonous Mushroom Data set

팀 명 : 박유임 조
팀원 : 임재인 박종범 조영관 유현준

INDEX

1. 문제 정의

식용 가능 버섯

항암 식품, 건강 보조 식품

독버섯

죽음으로 이르게 되는 치명적인 식품

화학적 성분을 분석하지 않고도, 외관의 특징만으로
독버섯인지 유무를 판단하기 위한 분석 모델을 만들기로 함

2. 데이터 준비 및 탐구

Data set은 'Kaggle'에 있는 Mushroom Classification 데이터를 이용하여 프로젝트를 진행함

입력 변수

명목형 21개

▶

출력 변수

독성 여부
Binary variable

총 8,124개의 데이터

2. 데이터 준비 및 탐구

입력 변수

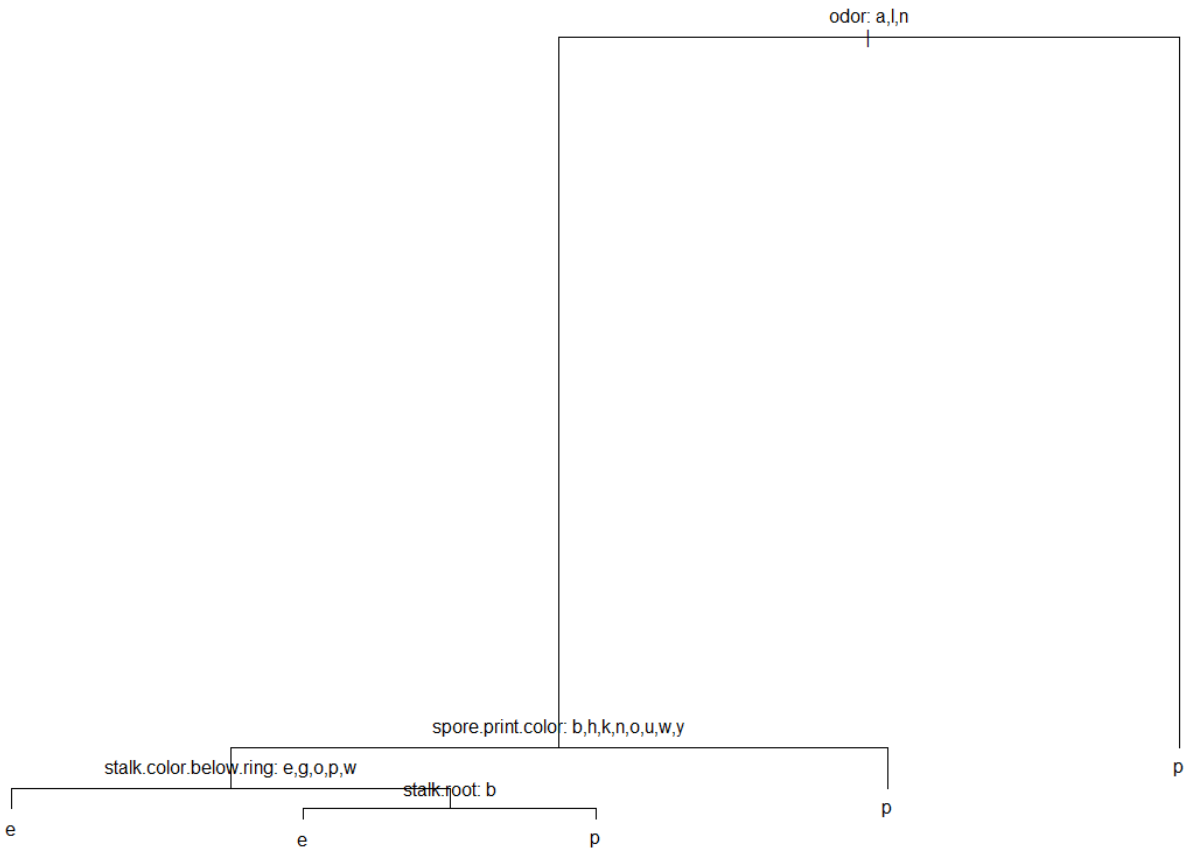
habitat	cap.color	gill. attachment	gill.color	stalk. surface. above.ring	stalk.color. below.ring	ring.type
cap.shape	bruises	gill. spacing	stalk. shape	stalk. surface. below.ring	veil.color	spore.print. color
cap. surface	odor	gill.size	stalk.root	stalk.color. above.ring	ring. number	population

출력 변수

classes : edible = e, poisonous = p

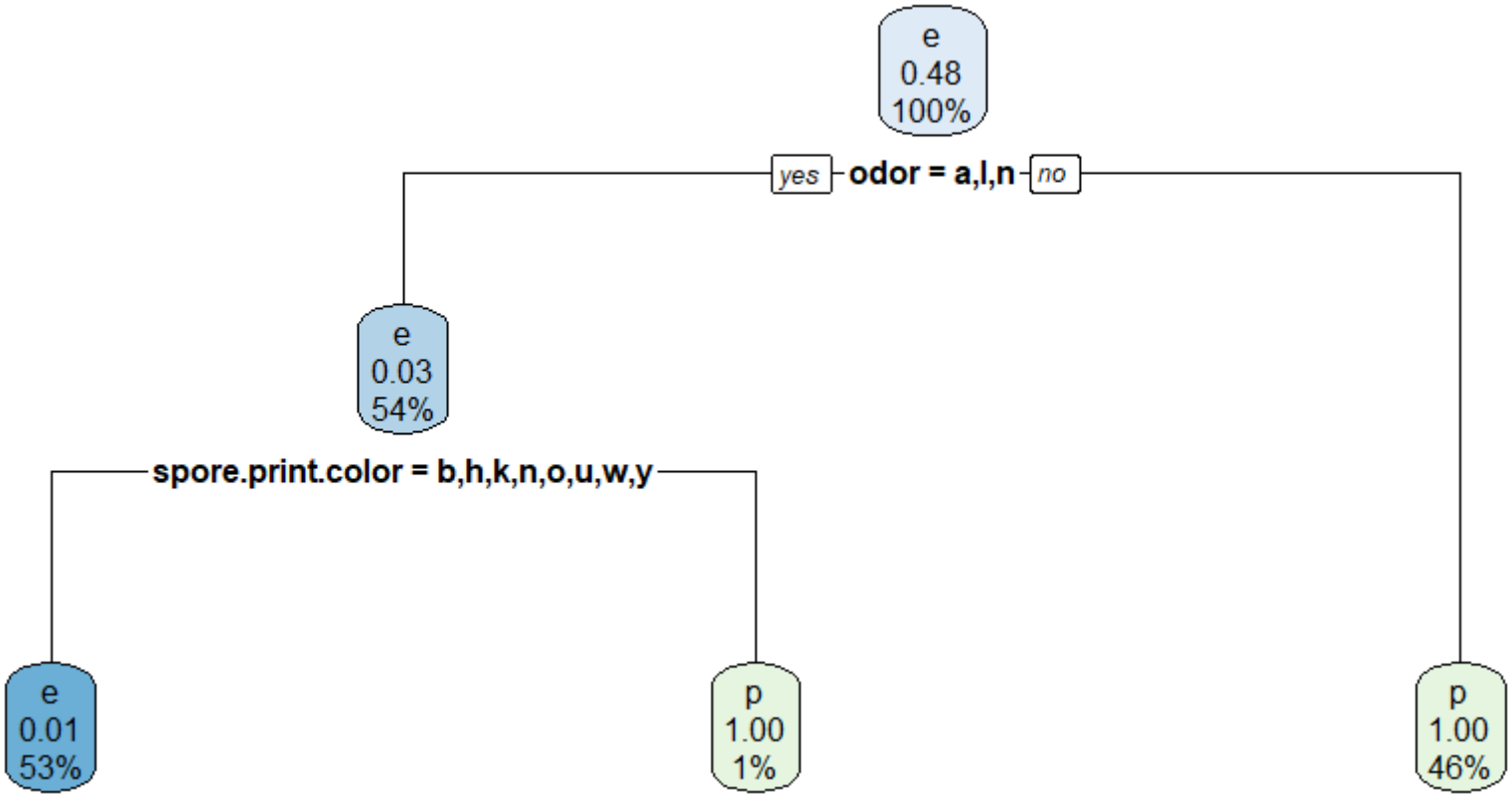
3. 데이터 분석 및 결과 도출 – Decision Tree

tree 패키지



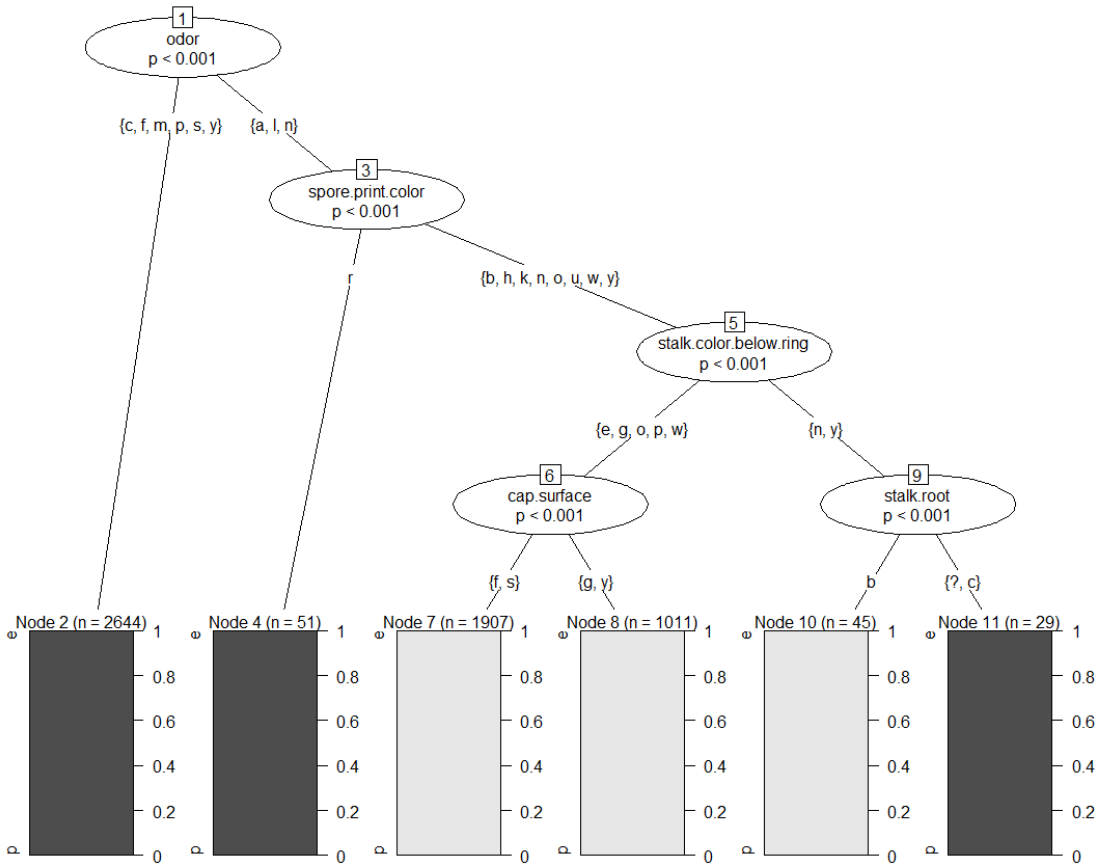
3. 데이터 분석 및 결과 도출 – Decision Tree

rpart 패키지



3. 데이터 분석 및 결과 도출 – Decision Tree

party 패키지



3. 데이터 분석 및 결과 도출 – Decision Tree

tree 패키지

```
> tree_cfm
   tree_prey
       e    p
  e 1249    0
  p    4 1184
```

rpart 패키지

```
> rpart_cfm
   rpart_prey
       e    p
  e 1249    0
  p   15 1173
```

party 패키지

```
> party_cfm4
   party_prey4
       e    p
  e 1249    0
  p    4 1184
```

```
> Perf_Table
                 TPR Precision TNR  Accuracy       BCR F1-Measure
tree(tree)     0.9966330         1   1 0.9983586 0.9983151  0.9983137
tree(rpart)    0.9873737         1   1 0.9938449 0.9936668  0.9936468
tree(party)    0.9966330         1   1 0.9983586 0.9983151  0.9983137
```

3. 데이터 분석 및 결과 도출 – Logistic Regression

Full

```
Call:
glm(formula = class ~ ., family = "binomial", data = m_scaled_trn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9824  -0.1492   0.0000   0.1292   1.9150

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.90691   19.24733  -0.203 0.839147
cap.shape                  0.01002    0.06874   0.146 0.884155
cap.surface                0.38055    0.09301   4.092 4.28e-05 ***
cap.color                 -0.35834    0.08473  -4.229 2.34e-05 ***
bruises                    1.30411    0.18545   7.032 2.04e-12 ***
odor                      -2.71702    0.17632 -15.409  < 2e-16 ***
gill.attachment           -5.28449  146.39445  -0.036 0.971205
gill.spacing              -8.59871    0.42686 -20.144  < 2e-16 ***
gill.size                 10.12748    0.46554  21.754  < 2e-16 ***
gill.color                -0.71021    0.10385  -6.839 7.98e-12 ***
stalk.shape               -1.17892    0.22780  -5.175 2.28e-07 ***
stalk.root                -9.88557    0.56431 -17.518  < 2e-16 ***
stalk.surface.above.ring  -8.40464    0.42526 -19.764  < 2e-16 ***
stalk.surface.below.ring   0.38389    0.11755   3.266 0.001091 **
stalk.color.above.ring    -0.40240    0.11508  -3.497 0.000471 ***
stalk.color.below.ring    -0.22790    0.11388  -2.001 0.045363 *
veil.color                14.82305  140.19004   0.106 0.915792
ring.number                0.34887    0.15312   2.278 0.022701 *
ring.type                  8.92338    0.54312  16.430  < 2e-16 ***
spore.print.color         -0.37082    0.16616  -2.232 0.025638 *
population                -1.51947    0.15611  -9.733  < 2e-16 ***
habitat                    0.29556    0.08980   3.291 0.000997 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7879.4  on 5686  degrees of freedom
Residual deviance: 1485.5  on 5665  degrees of freedom
AIC: 1529.5

Number of Fisher Scoring iterations: 18
```

3. 데이터 분석 및 결과 도출 – Logistic Regression

Forward Selection

```
Call:
glm(formula = class ~ gill.size + stalk.surface.above.ring +
    gill.spacing + bruises + odor + veil.color + stalk.surface.below.ring +
    stalk.shape + spore.print.color + stalk.root + cap.surface +
    habitat + gill.color + stalk.color.above.ring + ring.type +
    stalk.color.below.ring + population, data = m_scaled_trn)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.32551  -0.12676  -0.02047   0.10748   0.97239

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.483624   0.003391 142.603  < 2e-16 ***
gill.size                 0.244528   0.005609  43.596  < 2e-16 ***
stalk.surface.above.ring -0.099837   0.004332 -23.048  < 2e-16 ***
gill.spacing             -0.162515   0.005919 -27.456  < 2e-16 ***
bruises                  -0.159668   0.007877 -20.269  < 2e-16 ***
odor                     -0.045244   0.004313 -10.489  < 2e-16 ***
veil.color                0.088484   0.004108  21.539  < 2e-16 ***
stalk.surface.below.ring -0.040340   0.004547  -8.871  < 2e-16 ***
stalk.shape              -0.033514   0.005337  -6.280 3.64e-10 ***
spore.print.color        -0.120626   0.007069 -17.065  < 2e-16 ***
stalk.root               -0.091918   0.005784 -15.893  < 2e-16 ***
cap.surface               0.030937   0.003796   8.151 4.42e-16 ***
habitat                   0.031750   0.004127   7.693 1.68e-14 ***
gill.color               -0.034469   0.005054  -6.820 1.01e-11 ***
stalk.color.above.ring   -0.016069   0.004366  -3.681 0.000235 ***
ring.type                 0.024165   0.007780   3.106 0.001906 **
stalk.color.below.ring   -0.010799   0.004315  -2.502 0.012361 *
population               -0.007809   0.005037  -1.550 0.121163
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.06536891)

    Null deviance: 1420.64  on 5686  degrees of freedom
Residual deviance:  370.58  on 5669  degrees of freedom
AIC: 646.5

Number of Fisher Scoring iterations: 2
```

3. 데이터 분석 및 결과 도출 – Logistic Regression

Backward Elimination

```
Call:
glm(formula = class ~ cap.surface + bruises + odor + gill.spacing +
    gill.size + gill.color + stalk.shape + stalk.root + stalk.surface.above.ring +
    stalk.surface.below.ring + stalk.color.above.ring + stalk.color.below.ring +
    veil.color + ring.type + spore.print.color + population +
    habitat, data = m_scaled_trn)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.32551  -0.12676  -0.02047   0.10748   0.97239

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.483624   0.003391 142.603  < 2e-16 ***
cap.surface               0.030937   0.003796   8.151 4.42e-16 ***
bruises                  -0.159668   0.007877 -20.269  < 2e-16 ***
odor                     -0.045244   0.004313 -10.489  < 2e-16 ***
gill.spacing             -0.162515   0.005919 -27.456  < 2e-16 ***
gill.size                 0.244528   0.005609  43.596  < 2e-16 ***
gill.color               -0.034469   0.005054  -6.820 1.01e-11 ***
stalk.shape              -0.033514   0.005337  -6.280 3.64e-10 ***
stalk.root               -0.091918   0.005784 -15.893  < 2e-16 ***
stalk.surface.above.ring -0.099837   0.004332 -23.048  < 2e-16 ***
stalk.surface.below.ring -0.040340   0.004547  -8.871  < 2e-16 ***
stalk.color.above.ring   -0.016069   0.004366  -3.681 0.000235 ***
stalk.color.below.ring   -0.010799   0.004315  -2.502 0.012361 *
veil.color                0.088484   0.004108  21.539  < 2e-16 ***
ring.type                 0.024165   0.007780   3.106 0.001906 **
spore.print.color        -0.120626   0.007069 -17.065  < 2e-16 ***
population               -0.007809   0.005037  -1.550 0.121163
habitat                   0.031750   0.004127   7.693 1.68e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.06536891)

    Null deviance: 1420.64  on 5686  degrees of freedom
Residual deviance:  370.58  on 5669  degrees of freedom
AIC: 646.5

Number of Fisher Scoring iterations: 2
```

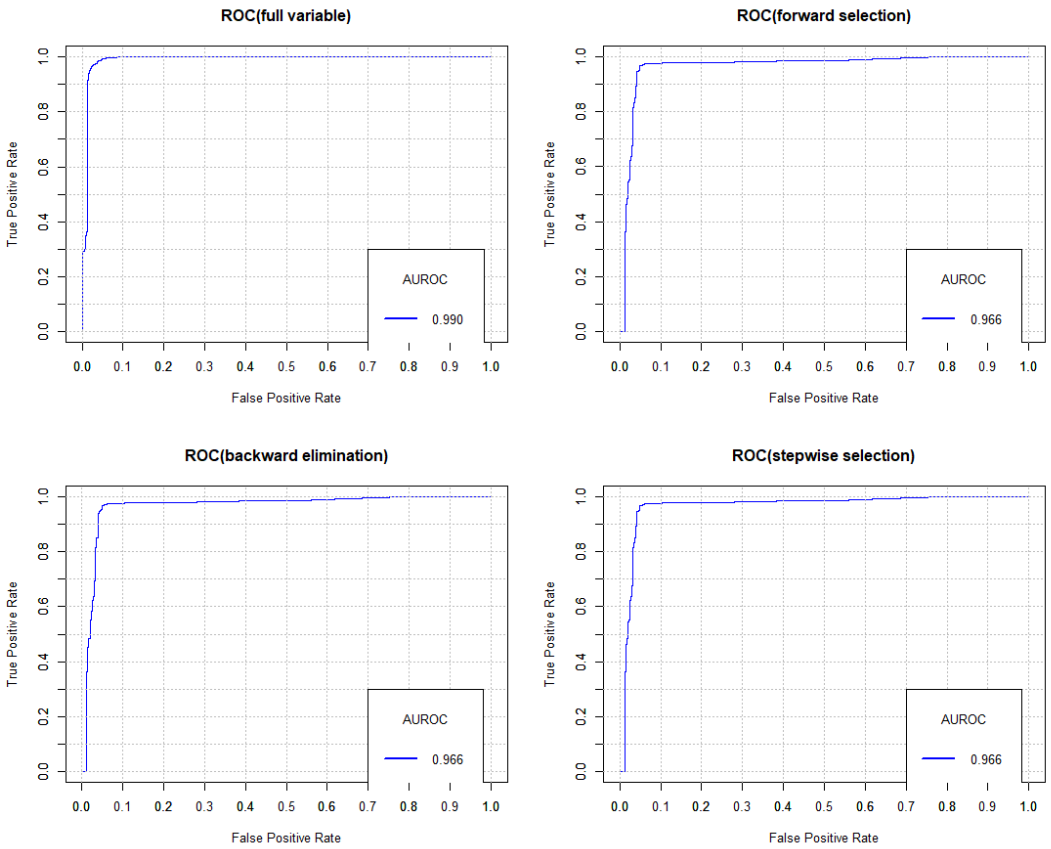
3. 데이터 분석 및 결과 도출 – Logistic Regression

Stepwise Selection

```
Call:
glm(formula = class ~ gill.size + stalk.surface.above.ring +
    gill.spacing + bruises + odor + veil.color + stalk.surface.below.ring +
    stalk.shape + spore.print.color + stalk.root + cap.surface +
    habitat + gill.color + stalk.color.above.ring + ring.type +
    stalk.color.below.ring + population, data = m_scaled_trn)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.32551  -0.12676  -0.02047   0.10748   0.97239

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               0.483624   0.003391 142.603  < 2e-16 ***
gill.size                 0.244528   0.005609  43.596  < 2e-16 ***
stalk.surface.above.ring -0.099837   0.004332 -23.048  < 2e-16 ***
gill.spacing             -0.162515   0.005919 -27.456  < 2e-16 ***
bruises                  -0.159668   0.007877 -20.269  < 2e-16 ***
odor                     -0.045244   0.004313 -10.489  < 2e-16 ***
veil.color                0.088484   0.004108  21.539  < 2e-16 ***
stalk.surface.below.ring -0.040340   0.004547  -8.871  < 2e-16 ***
stalk.shape              -0.033514   0.005337  -6.280 3.64e-10 ***
spore.print.color        -0.120626   0.007069 -17.065  < 2e-16 ***
stalk.root               -0.091918   0.005784 -15.893  < 2e-16 ***
cap.surface               0.030937   0.003796   8.151 4.42e-16 ***
habitat                   0.031750   0.004127   7.693 1.68e-14 ***
gill.color               -0.034469   0.005054  -6.820 1.01e-11 ***
stalk.color.above.ring   -0.016069   0.004366  -3.681 0.000235 ***
ring.type                 0.024165   0.007780   3.106 0.001906 **
stalk.color.below.ring   -0.010799   0.004315  -2.502 0.012361 *
population               -0.007809   0.005037  -1.550 0.121163
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.06536891)

    Null deviance: 1420.64  on 5686  degrees of freedom
Residual deviance:  370.58  on 5669  degrees of freedom
AIC: 646.5

Number of Fisher Scoring iterations: 2
```

3. 데이터 분석 및 결과 도출 – Logistic Regression

ROC Curve



3. 데이터 분석 및 결과 도출 – Logistic Regression

Full

```
> lg_cm
         lg_predicted
lg_target    0    1
        0 1244   41
        1   28 1124
```

Forward Selection

```
> forward_model_cm
   forward_model_prey
       0    1
  0 1233   52
  1   76 1076
```

Backward Elimination

```
> backward_model_cm
   backward_model_prey
       0    1
  0 1233   52
  1   76 1076
```

Stepwise Selection

```
> stepwise_model_cm
   stepwise_model_prey
       0    1
  0 1233   52
  1   76 1076
```

```
> Perf_Table_lg
                       TPR Precision       TNR  Accuracy       BCR F1-Measure
logistic(full)     0.9722222 0.9697733 0.9711769 0.9716865 0.9716994  0.9709962
logistic(forward)  0.9318182 0.9576125 0.9607686 0.9466557 0.9461827  0.9445392
logistic(backward) 0.9318182 0.9576125 0.9607686 0.9466557 0.9461827  0.9445392
logistic(stepwise) 0.9318182 0.9576125 0.9607686 0.9466557 0.9461827  0.9445392
```

3. 데이터 분석 및 결과 도출 – KNN

K = 5

```
> knn_model5_cm
     pred
true    0    1
   0 1285    0
   1    0 1152
```

K = 10

```
> knn_model10_cm
     pred
true    0    1
   0 1285    2
   1    0 1150
```

K = 20

```
> knn_model20_cm
     pred
true    0    1
   0 1282    3
   1    3 1149
```

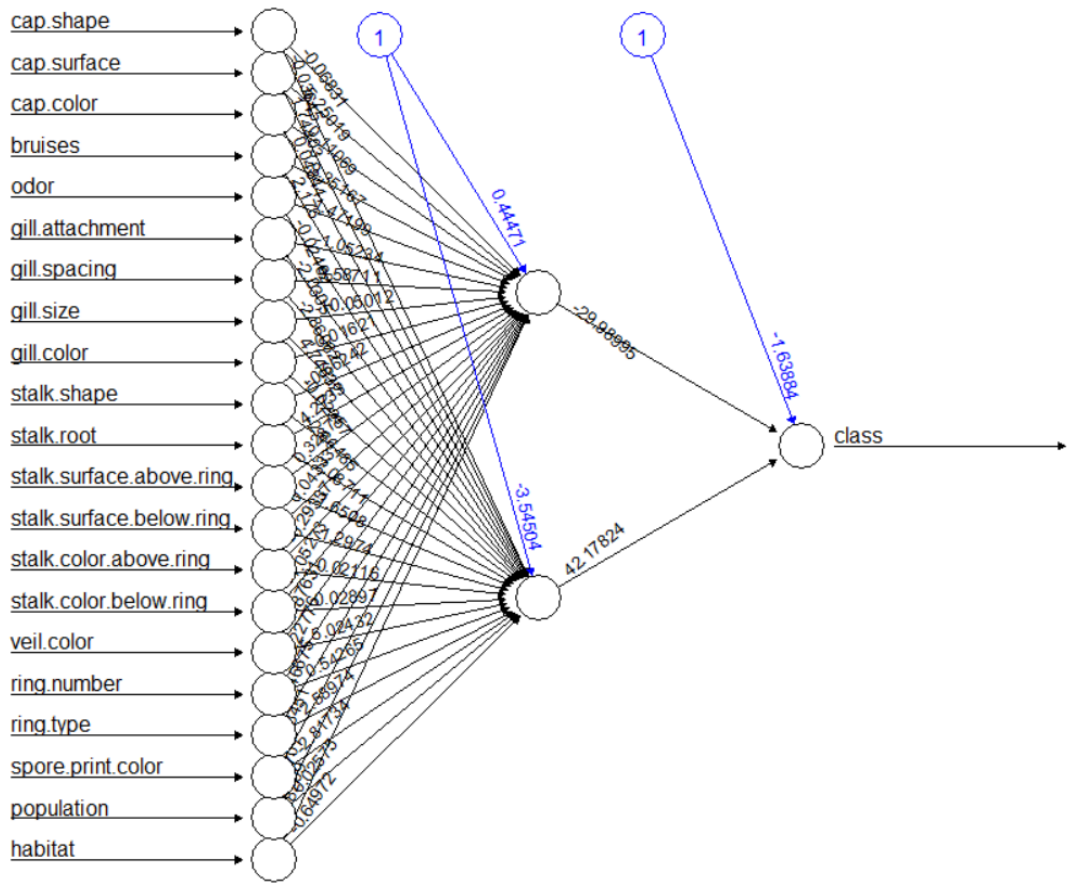
K = 50

```
> knn_model50_cm
     pred
true    0    1
   0 1273   21
   1   12 1131
```

```
> Perf_Table_knn
                TPR Precision       TNR  Accuracy       BCR F1-Measure
knn(k=5)  1.0000000  1.000000 1.0000000 1.0000000 1.0000000  1.0000000
knn(k=10) 1.0000000  0.996633 0.9968077 0.9983586 0.9984026  0.9983137
knn(k=20) 1.0000000  0.989899 0.9904837 0.9950759 0.9952305  0.9949239
knn(k=50) 0.9923274  0.979798 0.9810127 0.9864588 0.9866538  0.9860229
```

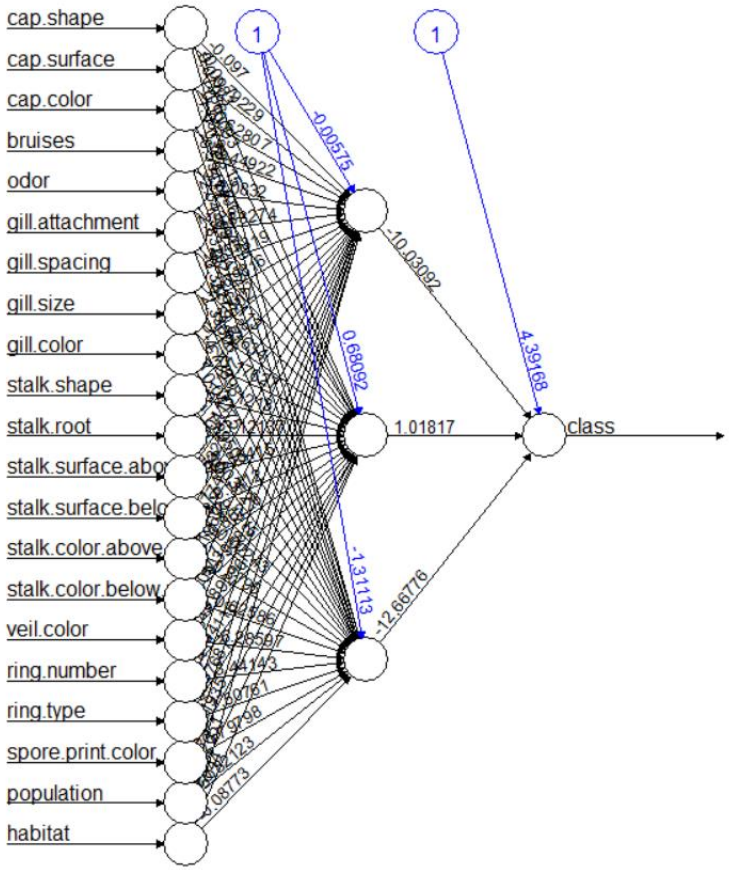
3. 데이터 분석 및 결과 도출 – ANN

Hidden Node = 2, Sigmoid



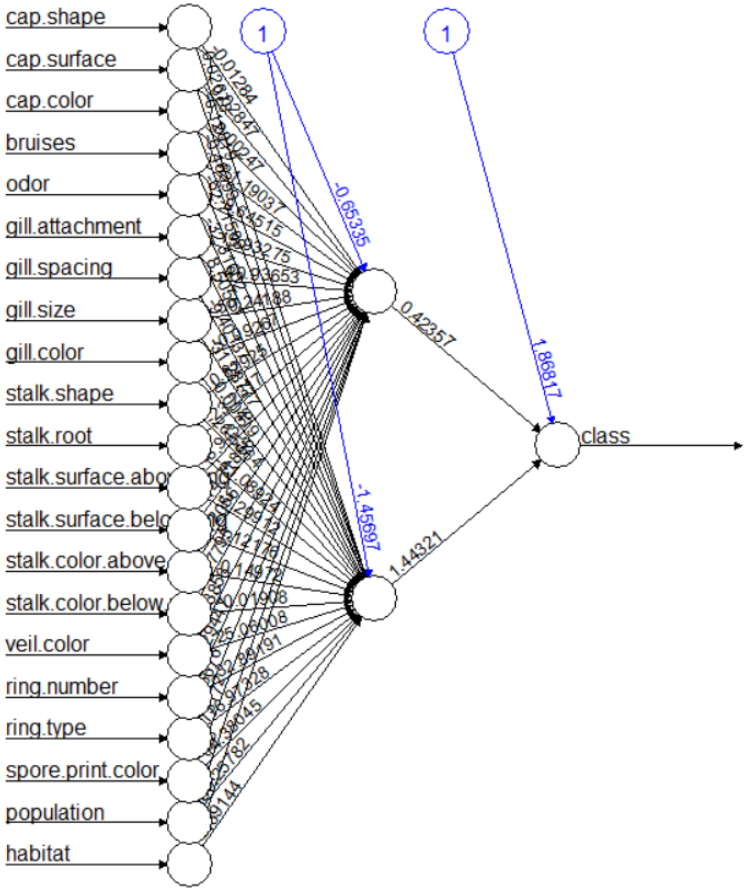
3. 데이터 분석 및 결과 도출 – ANN

Hidden Node = 3, Sigmoid



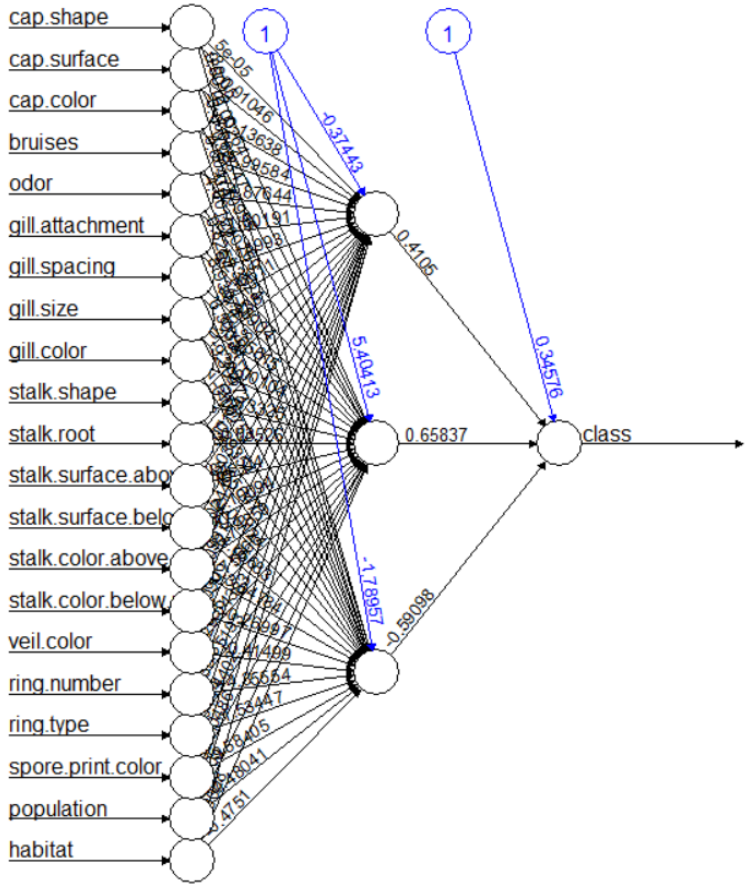
3. 데이터 분석 및 결과 도출 – ANN

Hidden Node = 2, Tanh



3. 데이터 분석 및 결과 도출 – ANN

Hidden Node = 3, Tanh



3. 데이터 분석 및 결과 도출 – ANN

Hidden Node = 2, Sigmoid

```
> ANN_model2_cm
   ANN_model2_prey
        0    1
  0 1210   39
  1   25 1163
```

Hidden Node = 3, Sigmoid

```
> ANN_model3_cm
   ANN_model3_prey
        0    1
  0 1234   15
  1   25 1163
```

Hidden Node = 2, Tanh

```
> ANN_model2_tanh_cm
   ANN_model2_tanh_prey
        0    1
  0 1210   39
  1    4 1184
```

Hidden Node = 3, Tanh

```
> ANN_model3_tanh_cm
   ANN_model3_tanh_prey
        0    1
  0 1210   39
  1    4 1184
```

```
> Perf_Table_ANN
                          TPR Precision       TNR  Accuracy       BCR F1-Measure
ANN(node=2, logistic) 0.9789562 0.9675541 0.9687750 0.9737382 0.9738523  0.9732218
ANN(node=3, logistic) 0.9789562 0.9872666 0.9879904 0.9835864 0.9834629  0.9830938
ANN(node=2, tanh)     0.9966330 0.9681112 0.9687750 0.9823554 0.9826053  0.9821651
ANN(node=3, tanh)     0.9966330 0.9681112 0.9687750 0.9823554 0.9826053  0.9821651
```

4. 최종 해석

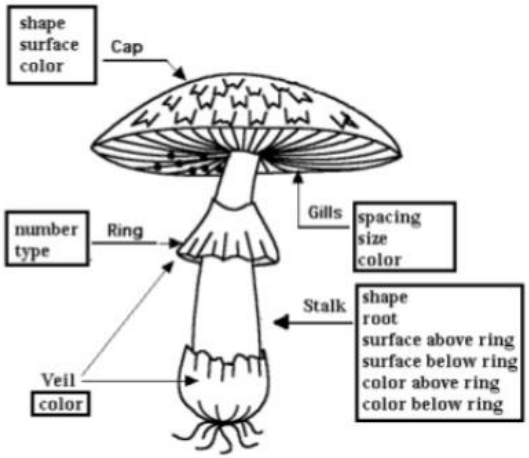
```
> Perf_Table
                          TPR Precision       TNR  Accuracy       BCR F1-Measu
tree(tree)          0.9966330 1.0000000 1.0000000 0.9983586 0.9983151  0.9983137
tree(rpart)         0.9873737 1.0000000 1.0000000 0.9938449 0.9936668  0.9936468
tree(party)         0.9966330 1.0000000 1.0000000 0.9983586 0.9983151  0.9983137
tree(party, d=4)    0.9966330 1.0000000 1.0000000 0.9983586 0.9983151  0.9983137
tree(party, d=3)    0.9873737 1.0000000 1.0000000 0.9938449 0.9936668  0.9936468
logistic(full)      0.9722222 0.9697733 0.9711769 0.9716865 0.9716994  0.9709962
logistic(forward)   0.9318182 0.9576125 0.9607686 0.9466557 0.9461827  0.9445392
logistic(backward)  0.9318182 0.9576125 0.9607686 0.9466557 0.9461827  0.9445392
logistic(stepwise)  0.9318182 0.9576125 0.9607686 0.9466557 0.9461827  0.9445392
knn(k=5)            1.0000000 1.0000000 1.0000000 1.0000000 1.0000000  1.0000000
knn(k=10)           1.0000000 0.9966330 0.9968077 0.9983586 0.9984026  0.9983137
knn(k=20)           1.0000000 0.9898990 0.9904837 0.9950759 0.9952305  0.9949239
knn(k=50)           0.9923274 0.9797980 0.9810127 0.9864588 0.9866538  0.9860229
ANN(node=2, logistic) 0.9789562 0.9675541 0.9687750 0.9737382 0.9738523  0.9732218
ANN(node=3, logistic) 0.9789562 0.9872666 0.9879904 0.9835864 0.9834629  0.9830938
ANN(node=2, tanh)   0.9966330 0.9681112 0.9687750 0.9823554 0.9826053  0.9821651
ANN(node=3, tanh)   0.9966330 0.9681112 0.9687750 0.9823554 0.9826053  0.9821651
```

Best Model !

4. 최종 해석

15 Attributes

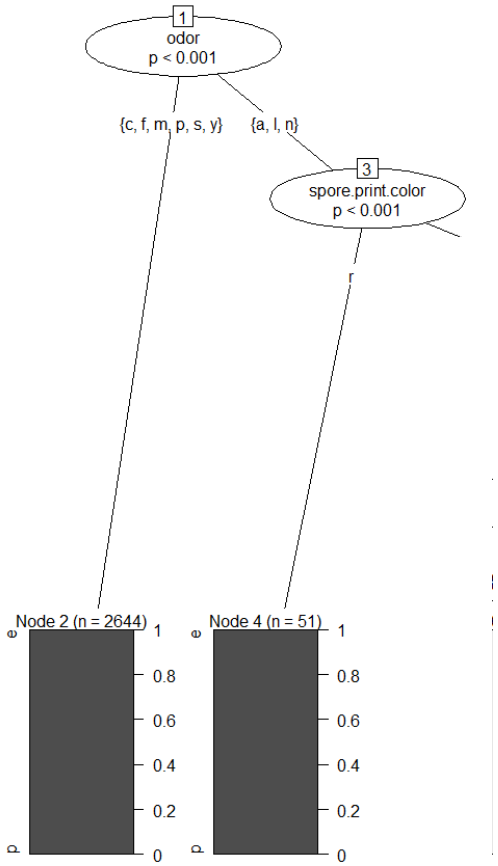
5 Sections:
1. Cap
2. Ring
3. Veil
4. Gills
5. Stalk

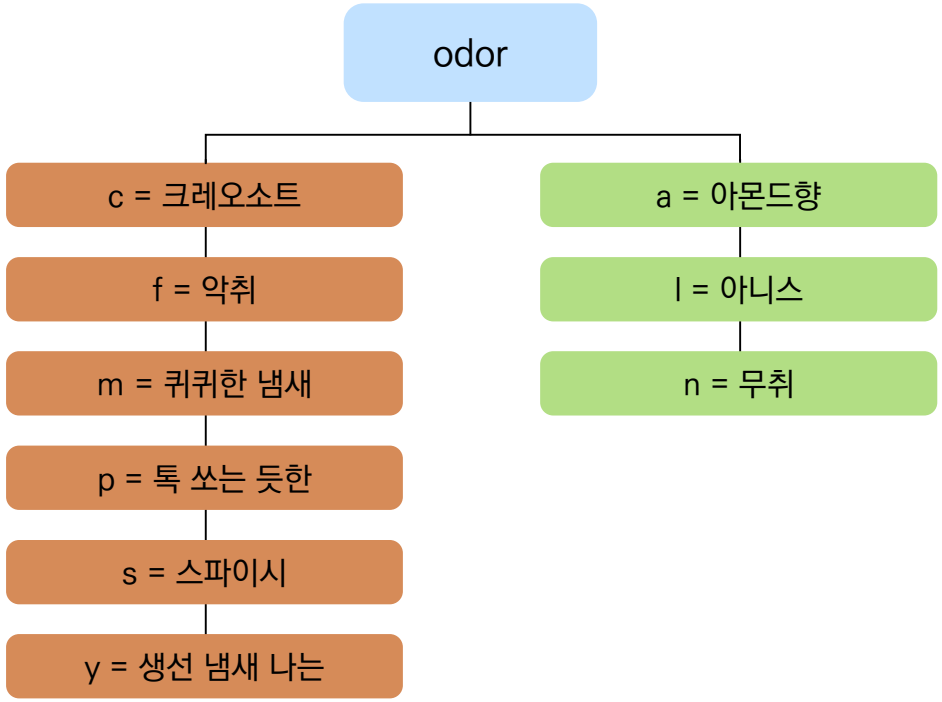


odor	향기	stalk.color.below.ring	고리 밑 자루 색
spore.print.color	포자문 색	stalk.root	자루 뿌리

4. 최종 해석

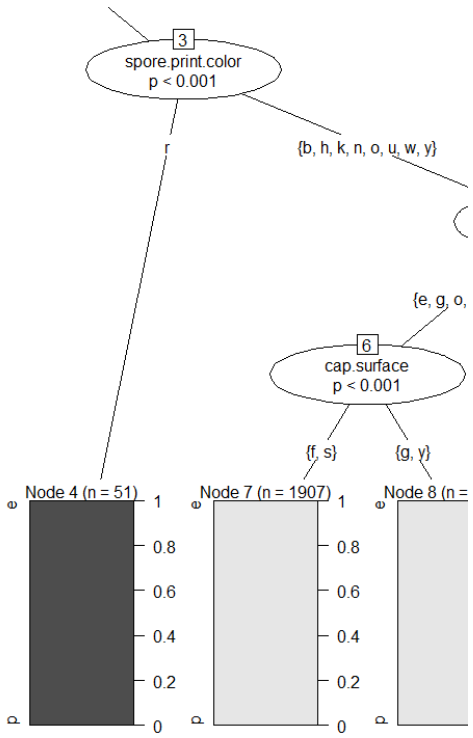
입력 변수 1. odor





4. 최종 해석

입력 변수 2. spore.print.color



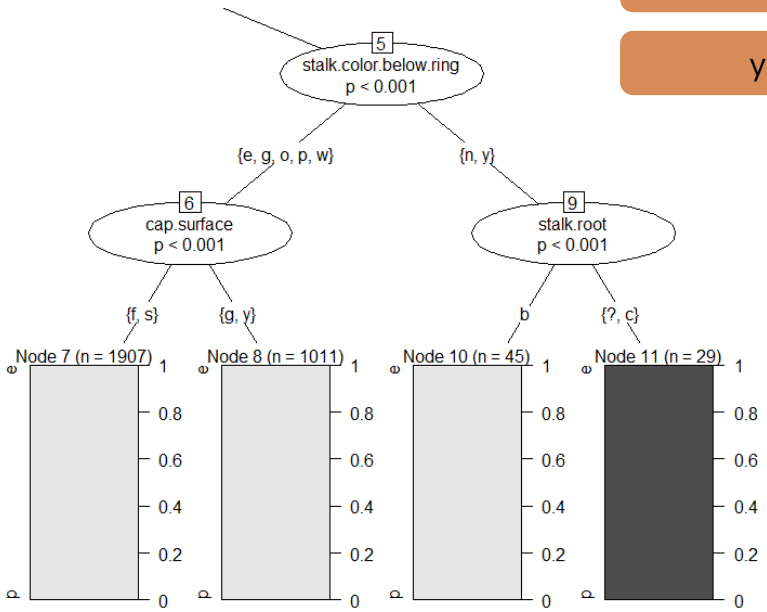
spore.print.color

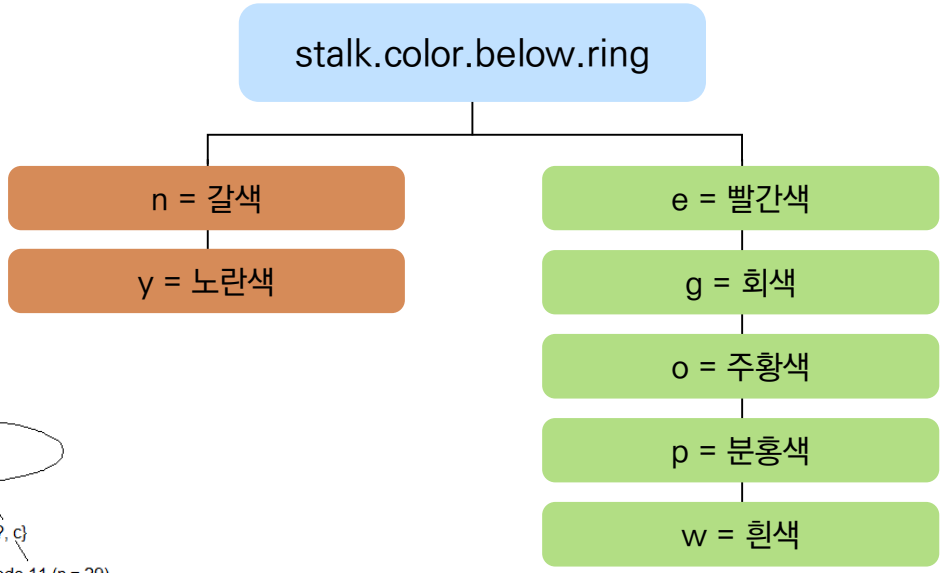
r = 초록색

- b = 황갈색
- h = 초콜릿색
- k = 검정색
- n = 갈색
- o = 주황색
- u = 보라색
- w = 흰색
- y = 노란색

4. 최종 해석

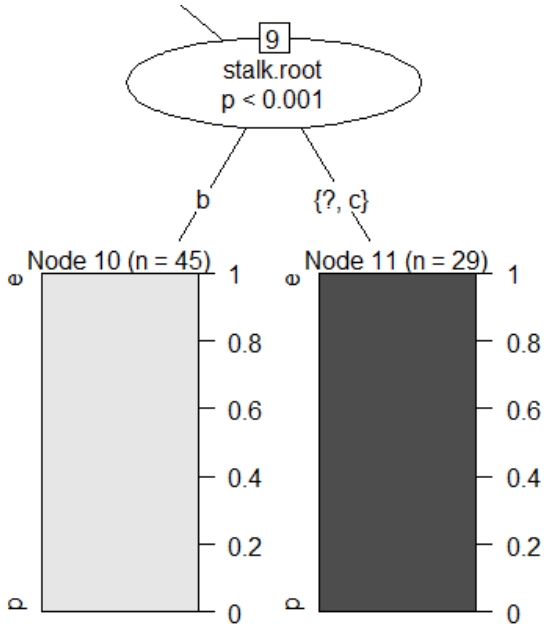
입력 변수 3. stalk.color.below.ring

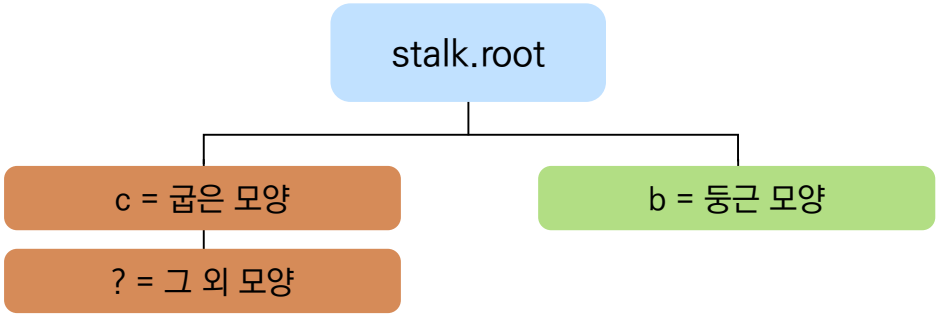


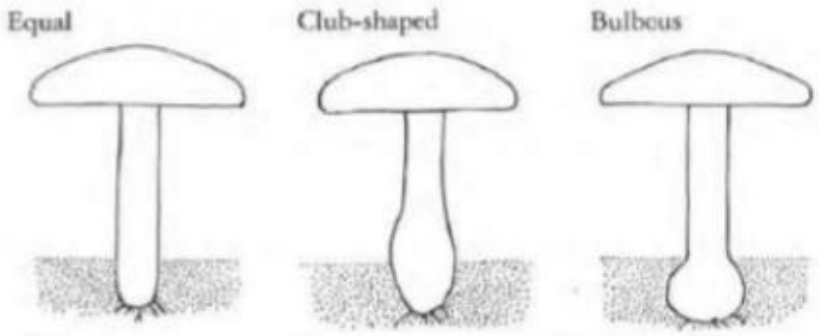


4. 최종 해석

입력 변수 4. stalk.root








End of Document