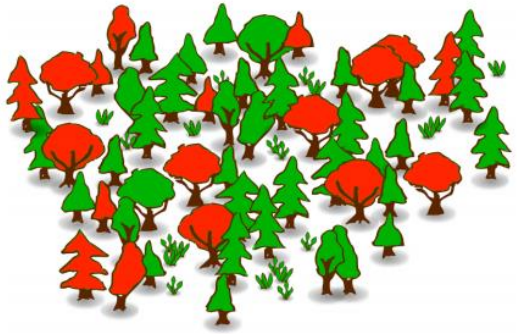
DECISION TREE

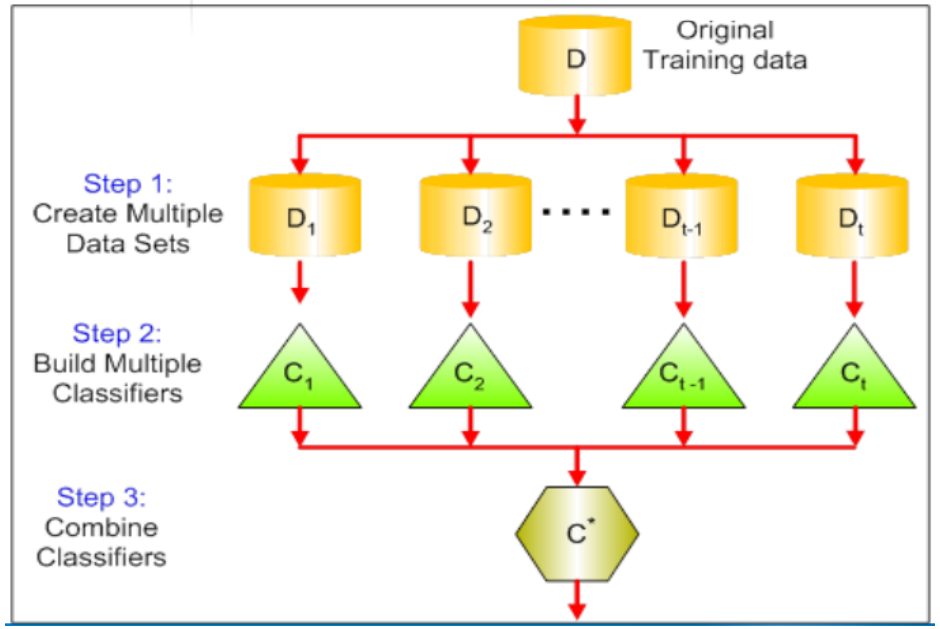
DEFINITION

- A schematic, tree-shaped diagram used to determine a course of action or show a statistical probability.
- **Each branch** of the decision tree represents a possible decision, occurrence or reaction.
- The tree is structured to show how and why one choice may lead to the next, with the use of the branches indicating each option is mutually exclusive.

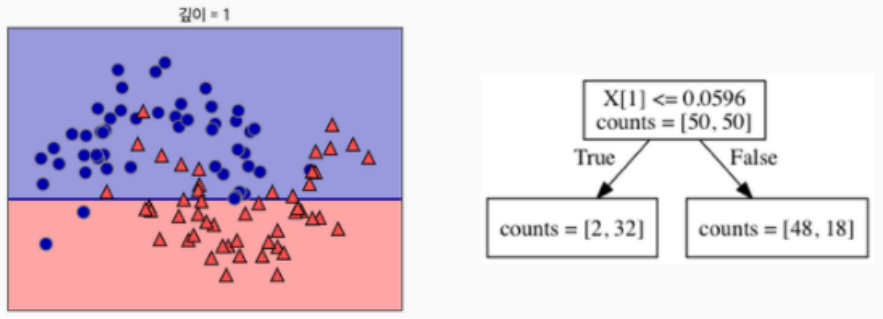
DECISION TREE VS RANDOM FOREST



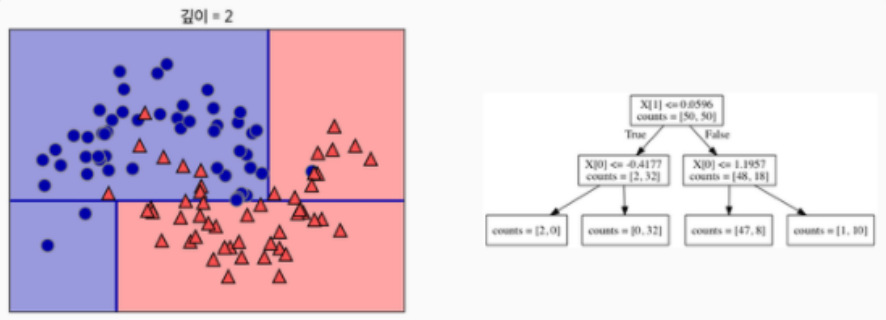
- "**Random forests** are a **combination of tree predictors** such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest."



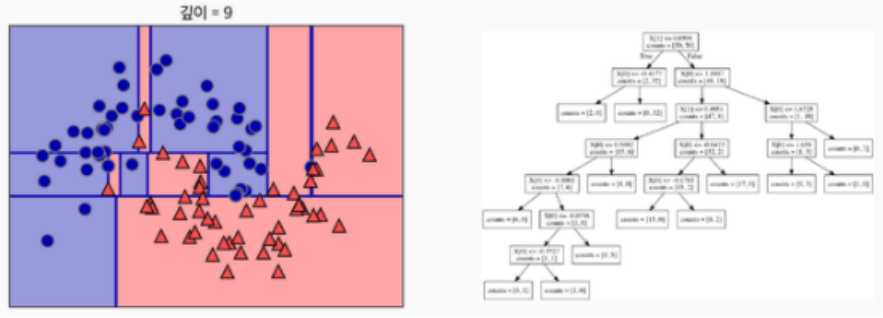
DECISION TREE PROCESS



먼저 위와 같이 데이터를 가장 잘 구분할 수 있는 질문을 기준으로 나눕니다.

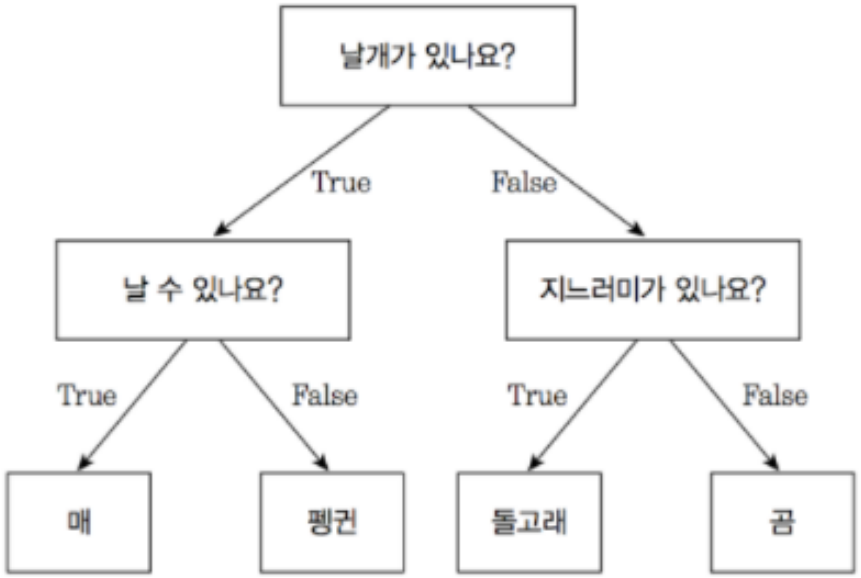


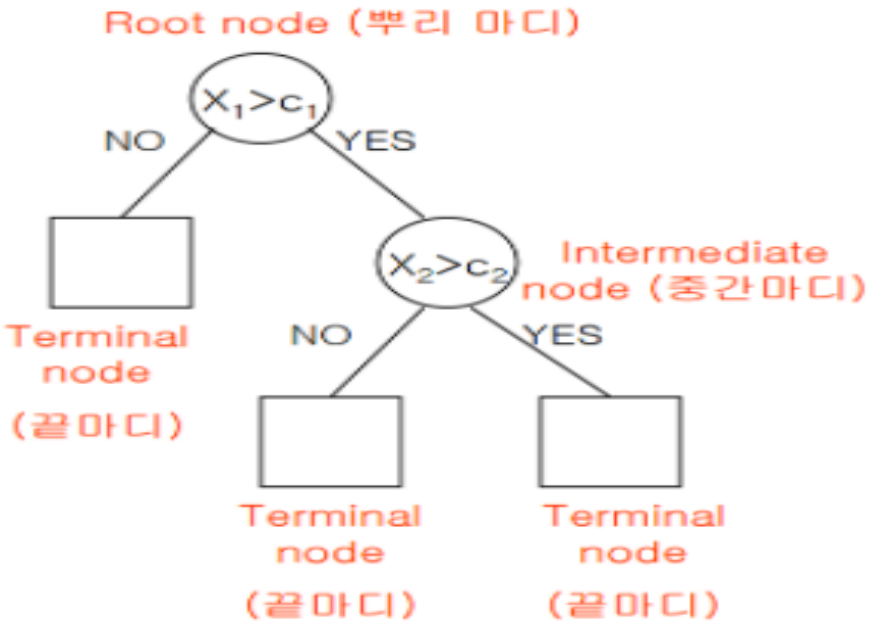
나뉜 각 범주에서 또 다시 데이터를 가장 잘 구분할 수 있는 질문을 기준으로 나눕니다.



지나치게 데이터를 많이 나누면 위 그림과 같이 데이터의 분할이 필요이상으로 진행된 것을 알 수 있습니다. 이를 **Overfitting**이라고 합니다. (**Overfitting**을 막기 위한 방법은 **Pruning**이 있습니다.)

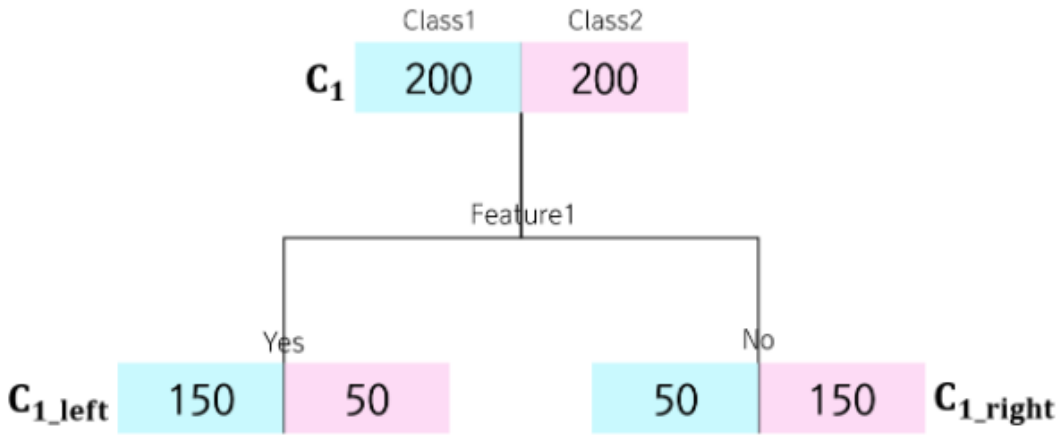
WHAT IS NODE?





- **Node:** 질문이나 정답을 담은 네모 상자
- **Root Node :** 맨 처음 분류 기준 (첫 질문)
- **Leaf Node(Terminal Node) :** 맨 마지막 노드

FEATURE IMPORTANCE



- Gini importance

$$G(N_j) = \sum_{i=1}^{K} p_i(1-p_i) = 1 - \sum_{i=1}^{K} p_i^2$$

$$G(C_1) = 1 - \{(\frac{200}{400})^2 + (\frac{200}{400})^2\} = 0.5$$

$$G(C_{1\_left}) = 1 - \{(\frac{150}{200})^2 + (\frac{50}{200})^2\} = 0.25$$

$$G(C_{1\_right}) = 1 - \{(\frac{50}{200})^2 + (\frac{50}{200})^2\} = 0.25$$

- Information gain

$$I(C_j) = w_j \cdot G(C_j) - w_{j\_left} \cdot G(C_{j\_left}) - w_{j\_right} \cdot G(C_{j\_right})$$

$$\begin{aligned} I(C_j) &= 1 \cdot G(C_j) - \frac{200}{400} \cdot G(C_{j\_left}) - \frac{200}{400} \cdot G(C_{j\_right}) \\ &= 0.5 - 0.5 \cdot 0.25 - 0.5 \cdot 0.25 \\ &= 0.25 \end{aligned}$$

- Feature importance

$$I(f_i) = \frac{\sum_{j:f_i\text{에 의해 split된 모든노드 } C_j} I(C_j)}{\sum_{k \in \text{all node } C_k} I(C_k)} \qquad I(f_i)^{norm} = \frac{I(f_i)}{\sum_{i \in \text{ all feature } f_i} I(f_i)}$$

TABLE OF DATA SET

See first five observations

```
proc print data=sampsio.LAQ(obs=5);
 var LobaOreg MinMinTemp Aconif PrecipAve Elevation ReserveStatus; run;
```

Output 15.1.1: Partial Listing of LAQ

Obs	LobaOreg	MinMinTemp	Aconif	PrecipAve	Elevation	ReserveStatus
1	0	-5.970	44.897	89.623	1567	Matrix
2	0	-6.430	81.585	91.231	1673	Reserve
3	1	-0.893	229.330	154.610	685	Reserve
4	0	-7.476	45.875	110.330	1971	Reserve
5	0	-5.992	81.679	98.739	1597	Reserve

The LAQ data set consists of 30 measurements of environmental conditions, such as **temperature**, **elevation**, and **moisture**, at 840 sites.

These variables are treated as **predictors** for the response variable LobaOreg (our main object), which is coded as 1 if the lichen species Lobaria oregana was present at the site and 0 otherwise.

RE TO EE FOR LOBAOREG

GROW statement specifies the entropy **criterion for splitting the observations** during the process of recursive partitioning that results in a large initial tree

PRUNE statement requests **cost-complexity pruning** to select a smaller subtree that **avoids overfitting the data**.

```
ods graphics on;
proc hpsplit data=sampsio.LAQ;
  class LobaOreg ReserveStatus;
  model LobaOreg (event='1') =
    Aconif DegreeDays TransAspect Slope Elevatio
    PctConifCov PctVegCov TreeBiomass EvapoTra
    MoistIndexAve MoistIndexDiff PrecipAve Preci
    RelHumidDiff PotGlobRadAve PotGlobRadDiff A
    DayTempAve DayTempDiff MinMinTemp MaxM
    AmbVapPressDiff SatVapPressAve SatVapPressDif ReserveStatus;
  grow entropy;
  partition fraction(VALIDATE = 0.3, SEED =123)
  prune costcomplexity;
  output out = scored;
run;
```

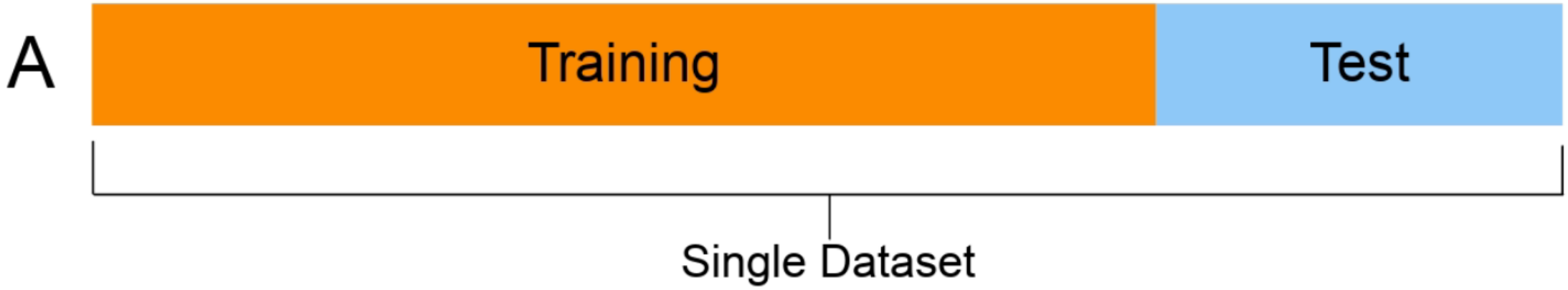
Partition fraction statement decided to divide data into trainset and test set. 'VALIDATE =0.3' means that train set is 70% and test set is 30%

Seed is a parameter for random selection.

In the case of **binary outcomes**, the **EVENT= option** is used to explicitly control the level of the response variable that represents the **event of interest** for computing the area under the curve (**AUC**), **sensitivity**, **specificity**, and values of the receiver operating characteristic (**ROC**) curves.

Note: These fit statistics do not apply to **categorical response** variables that have more than two levels, so the EVENT= option does not apply in that situation. Likewise, this option does not apply to **continuous response** variables.

PARTITION FRACTION(VALIDATE = 0.3, SEED =123) 추가설명

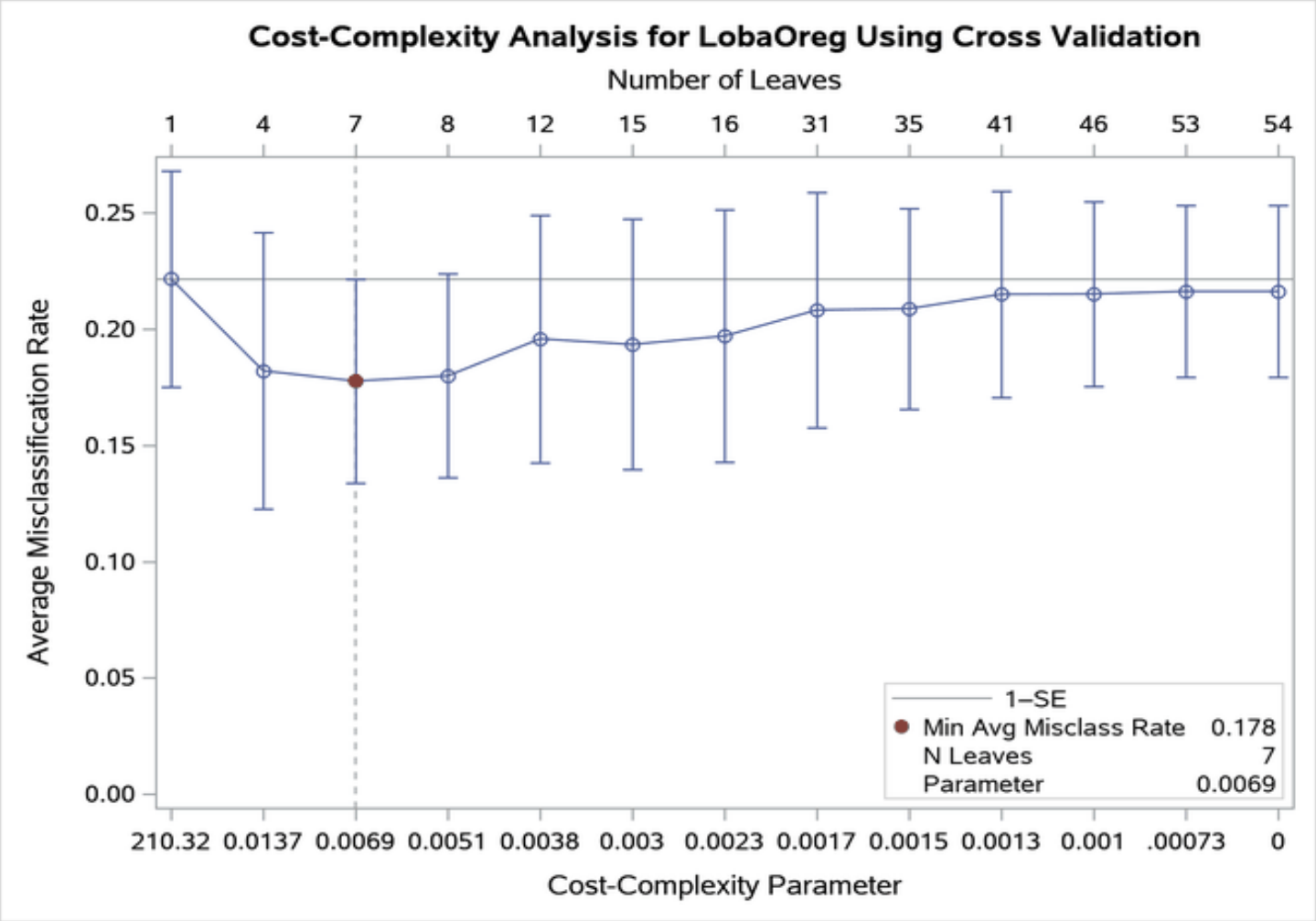


- **Train data set**(훈련/학습 데이터 셋)
모델을 학습시킬 때 사용할 데이터 셋

- **Validation data set** (검증 데이터 셋)
Train set으로 학습한 모델의 성능을 측정하기 위한 데이터 셋

- **SEED**
임의의 데이터 선정을 위한 파라미터이다.
같은 번호를 입력하면 같은 결과가 나오지만 다른 번호를 입력하면 데이터 셋이 달라졌기 때문에 다른 결과가 나온다.

MISCLASSIFICATION RATE (ERROR RATE) AS A FUNCTION OF COST-COMPLEXITY PARAMETER



Definition

- 전체 값에서 오차의 값이 발생한 비율
- 모형이 제대로 예측하지 못한 관측치를 평가하는 지표
- **Misclassification rate**는 전체 관측치 중 실제 값과 예측 치가 다른 정도를 나타내며 **1-accuracy** 또는 다음과 같은 식으로 나타낸다.

$$\text{error rate} = \frac{FP + FN}{TP + TN + FP + FN} = \frac{FP + FN}{P + N} = 1 - accuracy$$

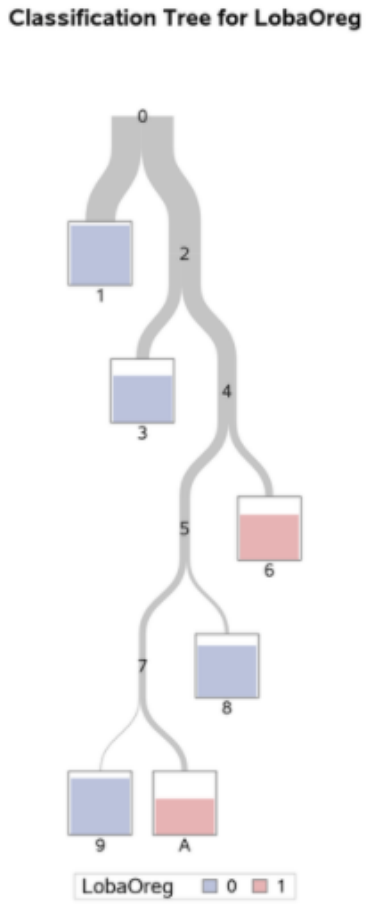
- **Selects the smallest subtree for which the misclassification rate is less than the minimum rate plus one standard error. >> Minimum error rate is at 7 leaves, so select subtree with six leaves.**

OVERVIEW OF FITTED TREE

- The **color** of the bar in each leaf node indicates the most frequent level of LobaOreg and represents the **classification level** assigned to all observations in that node.

- The **height** of the bar indicates the **proportion** of observations (sites) in the node that have the most frequent level.

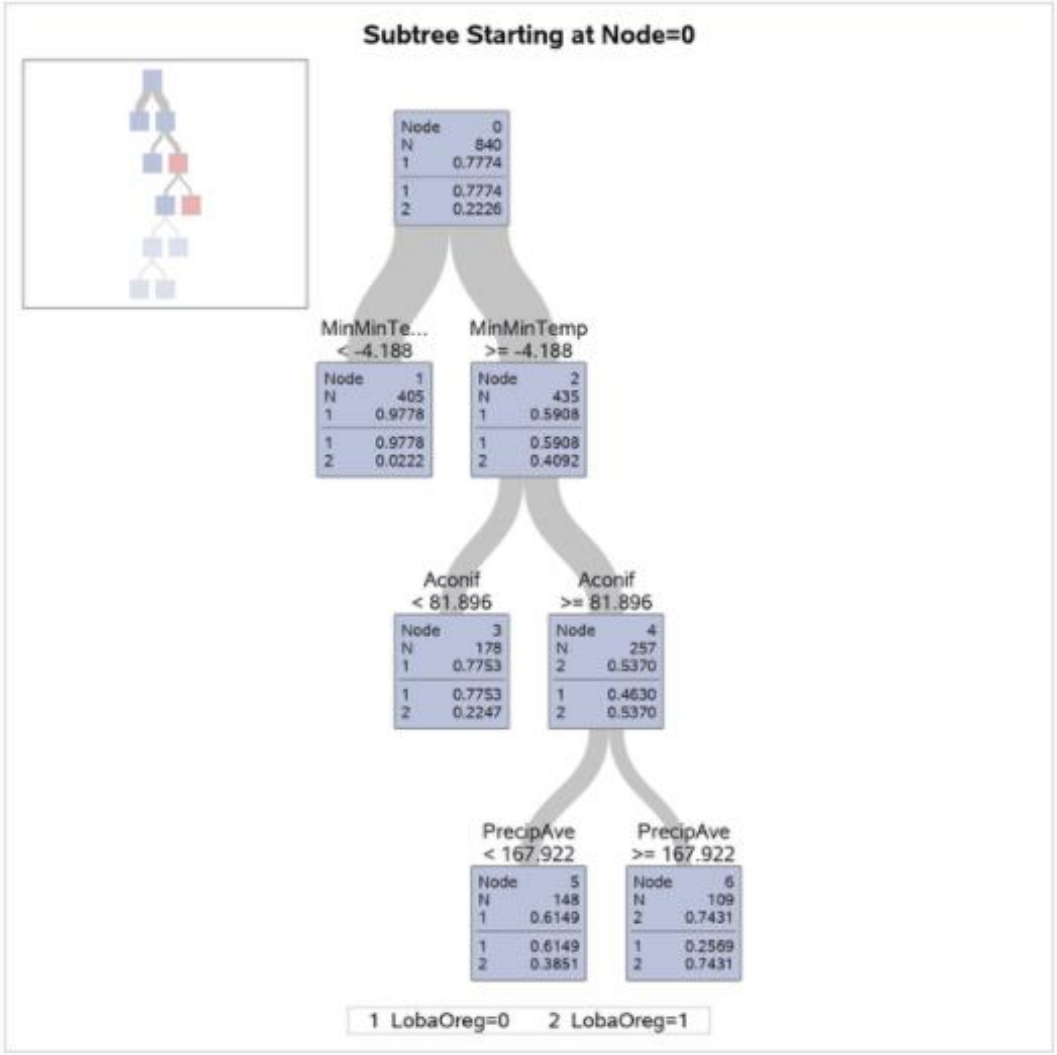
- Note: there is impurity in each leaf node. In other words, there can be LobaOreg observation values 1 and 0 in each node.



FIRST FOUR LEVELS OF FITTED TREE

- The diagram provides more detail about the nodes and splits in the first four levels of the tree. It reveals a model that is highly interpretable.

- The 435 sites for which **MinMinTemp** – 4.188 (node 2) are further subdivided based on the variable **Aconif**, which is the average age of the dominant conifer at the site. Lobaria oregana is present at **53.7%** of the 257 sites for which MinMinTemp – 4.188 and Aconif 81.896 years.

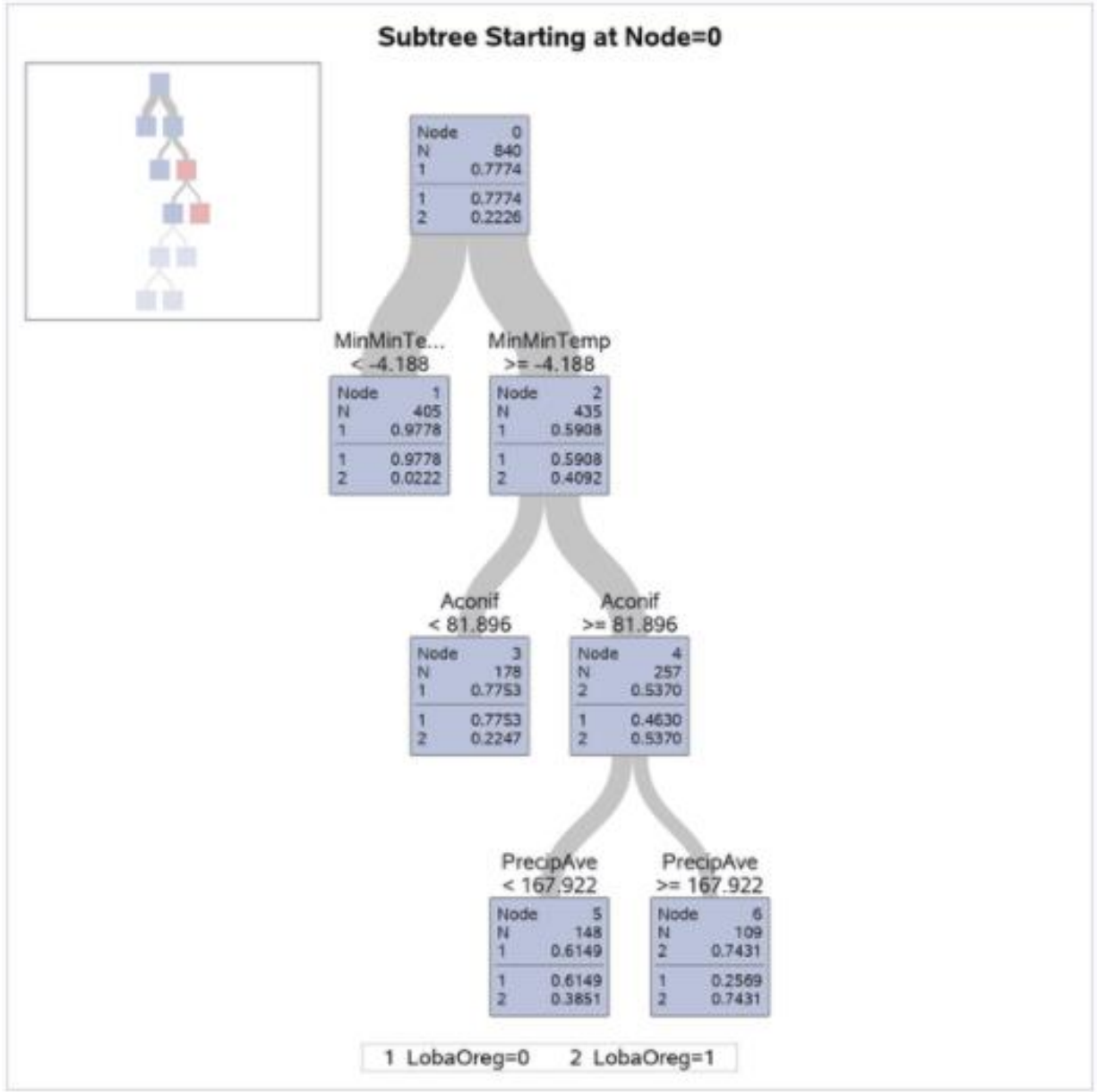


- The 257 sites for which Aconif 81.896 are further subdivided on the basis of **PrecipAve** (average monthly precipitation) with a cutoff value 167.922 mm.

- Lobaria oregana was present at **74.31%** of the 109 sites for which MinMinTemp –4.188, Aconif 81.896 years, and PrecipAve 167.922 mm.

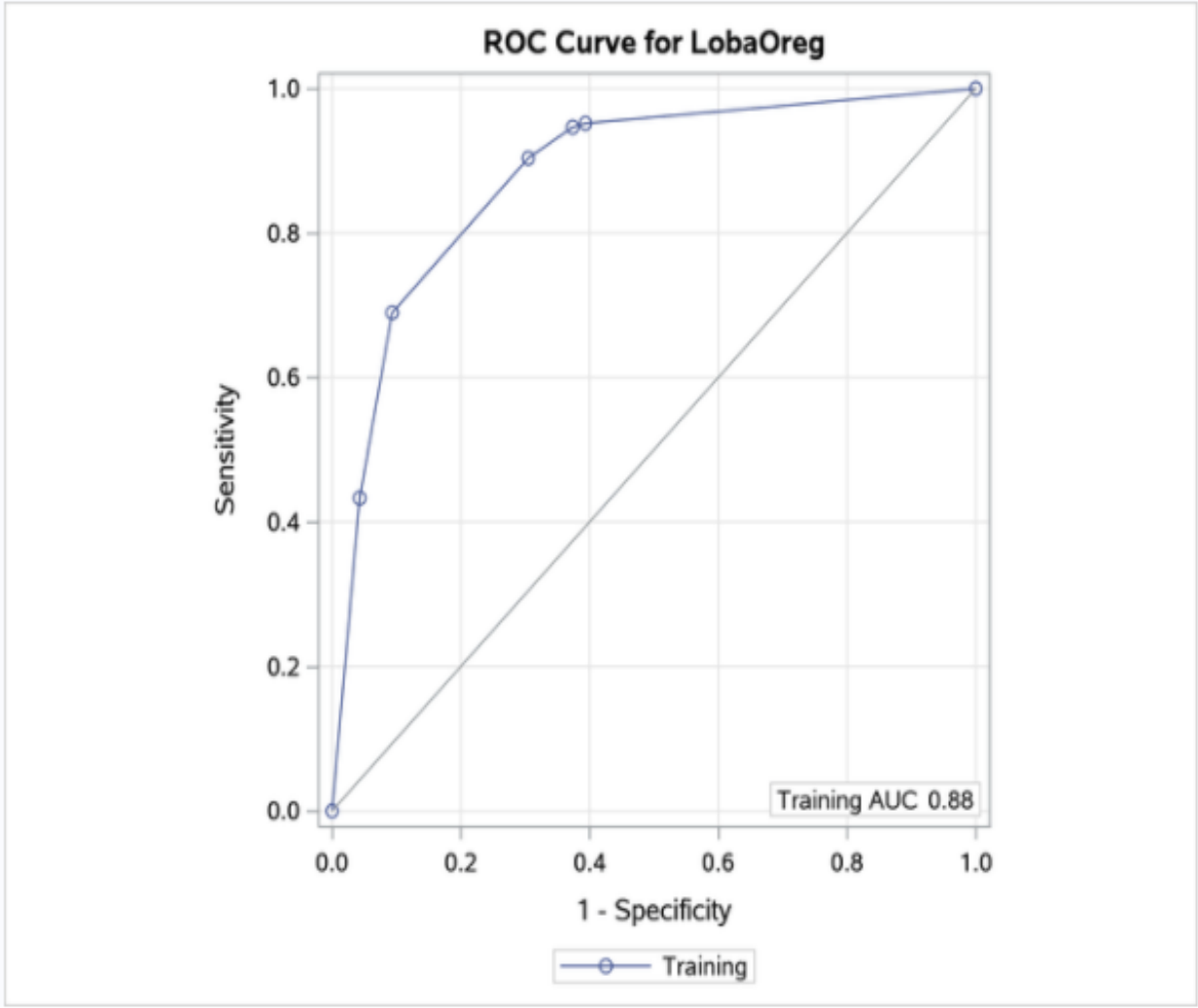
- Contrast this occupancy percentage with the 2.22% for the sites for which MinMinTemp –4.188.

- In summary, based on the first three splits, Lobaria oregana is most likely to be found at sites for which MinMinTemp –4.188, Aconif 81.896, and PrecipAve 167.922.



ROC CURVE FOR CLASSIFICATION

- The AUC statistic and the values of the ROC curve are computed from the training data. When you specify a validation data set by using the **PARTITION statement**, the plot displays an **additional ROC curve and AUC statistic**, whose values are computed from the validation data.
- Note: In this example, the computations of the sensitivity, specificity, AUC, and values of the ROC curve depend on defining LobaOreg=1 as the event of interest by using the EVENT= option in the MODEL statement.
- 진단의 관점에서 **민감도(sensitivity)**는 질병이 있는 사람을 얼마나 잘 찾아 내는가에 대한 값이고 특이도(**specificity**)는 정상을 얼마나 잘 찾아 내는가에 대한 값이다.



ROC CURVE

		실제	
		양성	음성
예측	양성	True Positive	False Positive Type I Error
	음성	False Negative Type II Error	True Negative

- **민감도(Sensitivity, True positive rate(TPR), Recall)**: 실제 병에 걸린 사람이 양성(Positive) 판정을 받는 비율입니다.
- **특이도(Specificity, True Negative rate(TNR))**: 정상인이 음성(Negative) 판정을 받는 비율입니다.
- **False positive rate(FPR)** = 1-specificity
- **정확도(Accuracy)**: 전체 데이터 중 제대로 분류된 데이터 비율
- **에러율(Error Rate)**: 전체 데이터 중 제대로 분류되지 않은 데이터 비율
- **정밀도(Precision)**: Positive로 예측했을 때, 실제로 Positive인 비율

$$\text{민감도(Sensitivity, Recall, True Positive Rate)} = \frac{(1)}{(1)+(3)} = \frac{TP}{TP+FN}$$

$$\text{특이도(Specificity, True Negative rate)} = \frac{(4)}{(2)+(4)} = \frac{TN}{FP+TN}$$

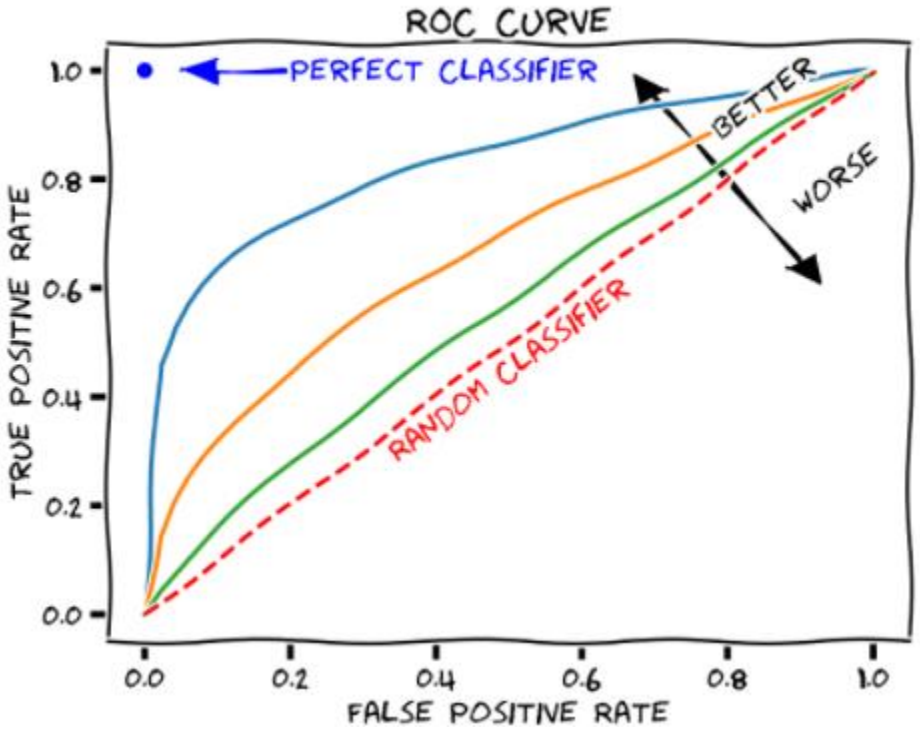
$$\text{False positive rate(FPR)} = \frac{(2)}{(2)+(4)} = \frac{FP}{FP+TN}$$

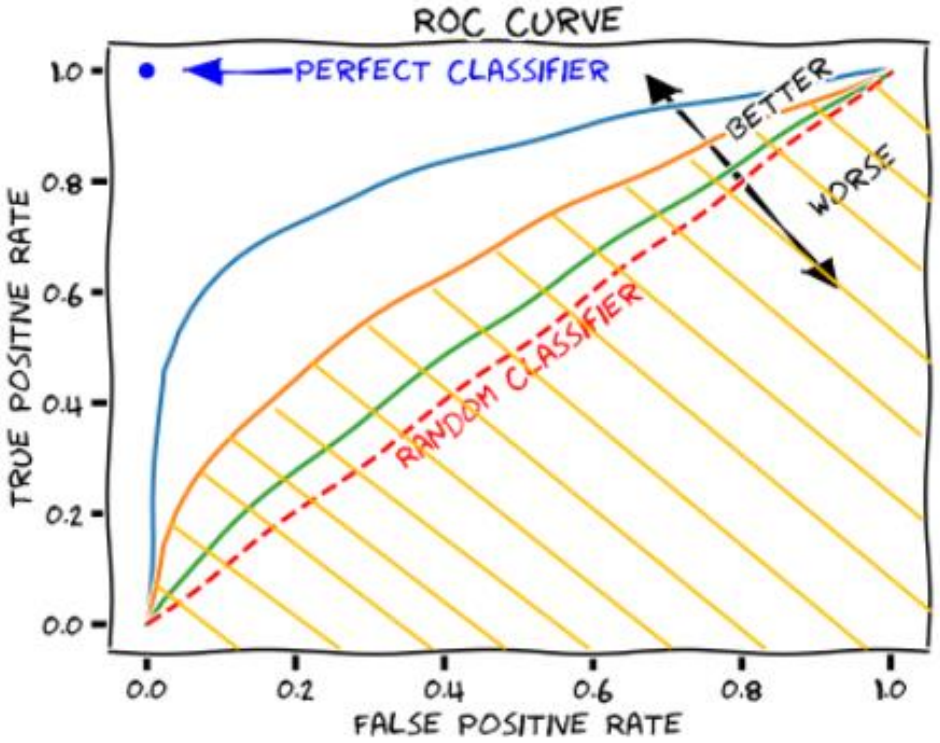
$$\text{정확도(Accuracy)} = \frac{(1)+(4)}{(1)+(2)+(3)+(4)} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{에러율(Error Rate)} = \frac{(2)+(3)}{(1)+(2)+(3)+(4)} = \frac{FP+FN}{TP+FP+FN+TN}$$

$$\text{정밀도(Precision)} = \frac{(1)}{(1)+(2)} = \frac{TP}{TP+FP}$$

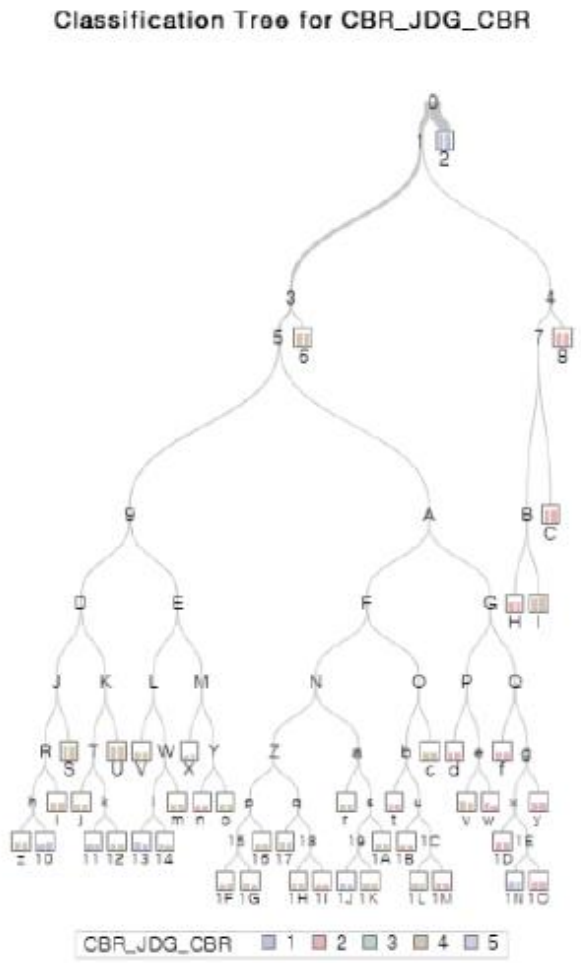
ROC CURVE INTERPRETATION

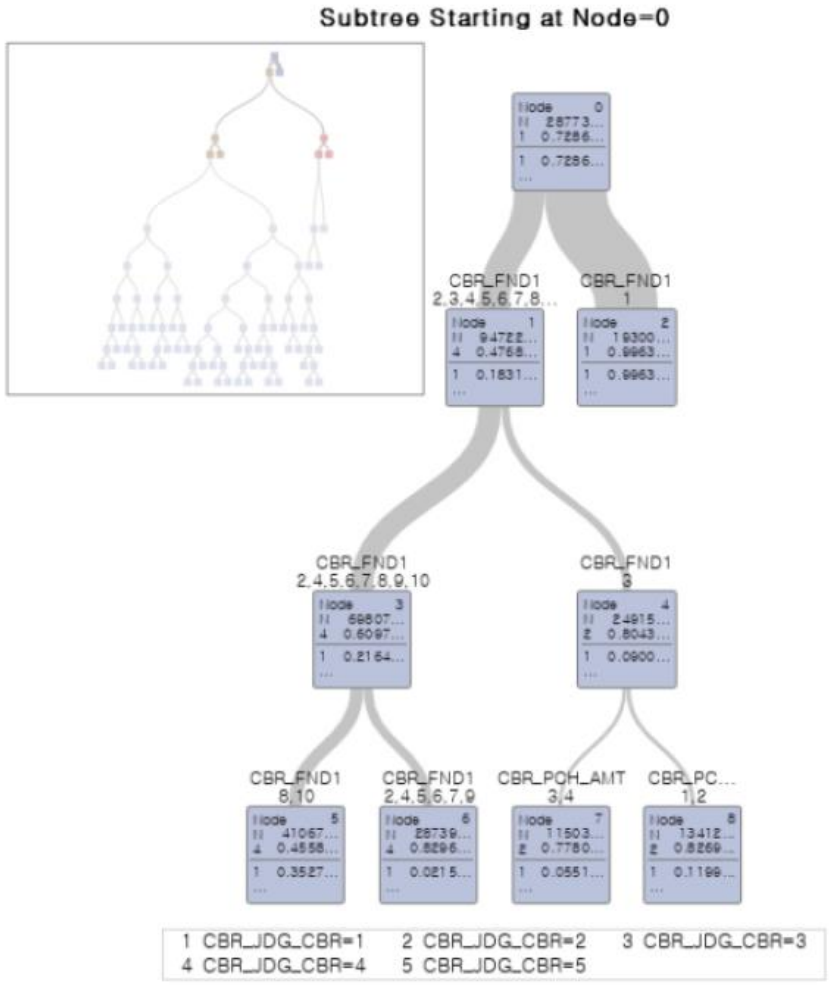




The colored area is AUC (정확성 지표)

RESULT: CBR_JDG_CBR

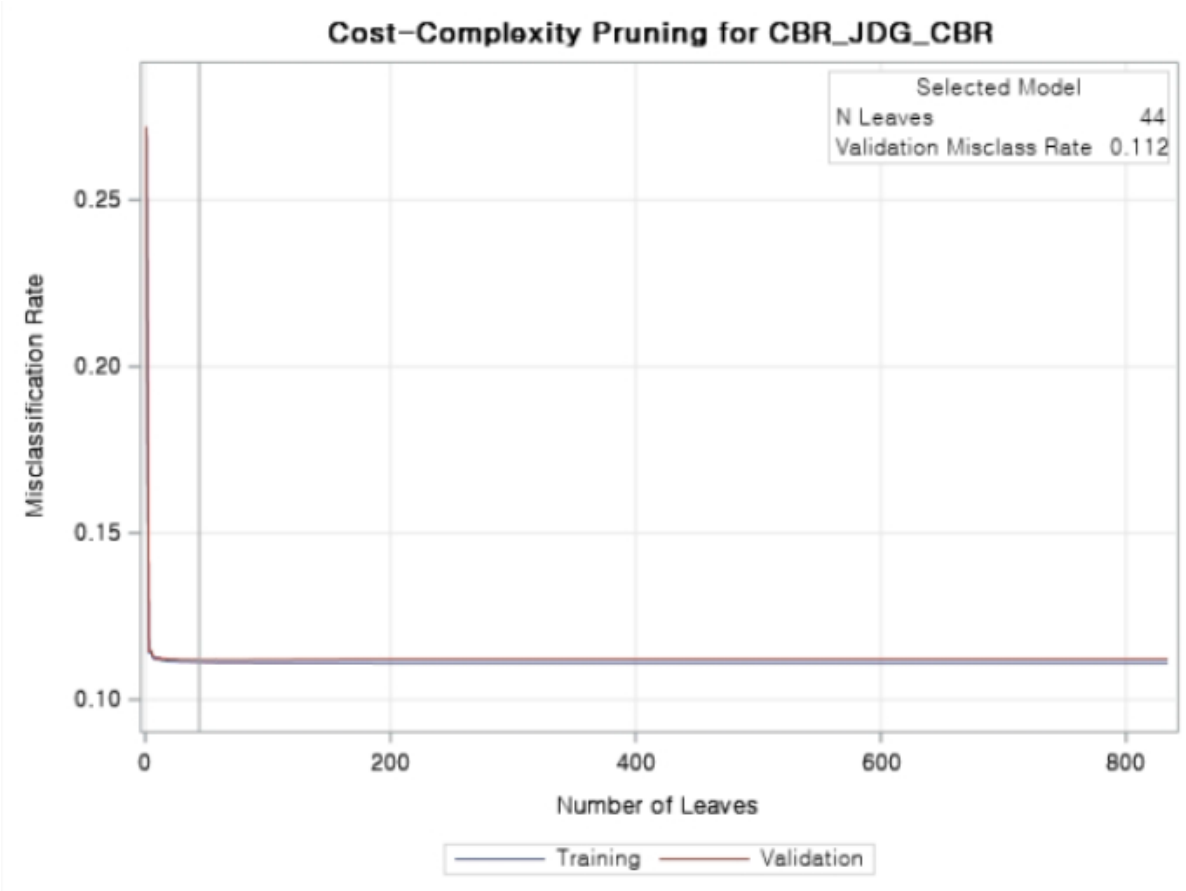




VARIABLE IMPORTANCE

Variable Importance						
		Training		Validation		
Variable	Variable Label	Relative	Importance	Relative	Importance	Relative Ratio
CBR_FND1	CBR_FND1	1.0000	895.8	1.0000	586.5	1.0000
CBR_PCH_AMT	CBR_PCH_AMT	0.0999	89.4555	0.0996	58.4017	0.9971
G1E_SGOT	G1E_SGOT	0.0151	13.5325	0.0160	9.3669	1.0572
MNS_YN_1		0.0144	12.8977	0.0150	8.8062	1.0428
G1E_HDL	G1E_HDL	0.0167	14.9757	0.0144	8.4418	0.8609
G1E_LDL	G1E_LDL	0.0184	16.4825	0.0125	7.3225	0.6785
G1E_FBS	G1E_FBS	0.0124	11.1426	0.0122	7.1286	0.9771
QC_BRFD_DRT	QC_BRFD_DRT	0.0132	11.8578	0.0106	6.2352	0.8031
EXER		0.0099	8.8503	0.0103	6.0589	1.0456
G1E_BP_DIA	G1E_BP_DIA	0.0115	10.3153	0.0098	5.7400	0.8499
QC_PHX_BBR_YN	QC_PHX_BBR_YN	0.0109	9.7226	0.0096	5.6578	0.8888
G1E_SGPT	G1E_SGPT	0.0075	6.7533	0.0079	4.6398	1.0493
G1E_BP_SYS	G1E_BP_SYS	0.0065	5.7978	0.0062	3.6416	0.9593
G1E_GGT	G1E_GGT	0.0051	4.5338	0.0055	3.2186	1.0842
QC_OPLL_YN	QC_OPLL_YN	0.0051	4.5896	0.0053	3.1055	1.0334
G1E_TG	G1E_TG	0.0057	5.1036	0.0039	2.2766	0.6813
QC_MNC_AGE	QC_MNC_AGE	0.0064	5.6988	0.0038	2.2271	0.5969

RESULT: MISCLASSIFICATION RATE



REFERENCE

- https://documentation.sas.com/doc/ko/pgmsascdc/9.4_3.4/stathpug/stathpug_hpsplit_examples01.htm
- https://m.blog.naver.com/sharp_kiss/221826800044
- https://codedragon.tistory.com/9618
- https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Toronto-Data-Mining-Forum/dzieciolowski-randomforests.pdf
- https://towardsdatascience.com/how-to-find-decision-tree-depth-via-cross-validation-2bf143f0f3d6