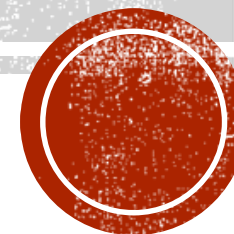




DECISION TREE



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.”

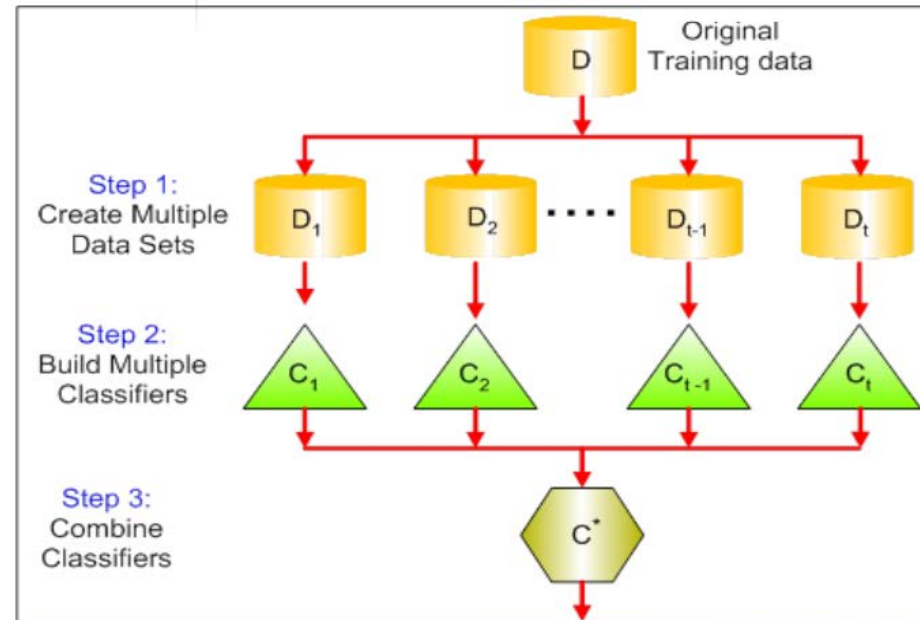


TABLE OF DATA SET

```
proc print data=sampsio.LAQ(obs=5);
```

See first five observations

```
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.



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

ods graphics on;

proc hpsplit data=sampsio.LAQ;

class LobaOreg ReserveStatus;

model LobaOreg (event='1') =

Aconif DegreeDays TransAspect Slope Elevation

PctConifCov PctVegCov TreeBiomass EvapoTra

MoistIndexAve MoistIndexDiff PrecipAve Precip

RelHumidDiff PotGlobRadAve PotGlobRadDiff A

DayTempAve DayTempDiff MinMinTemp MaxM

AmbVapPressDiff SatVapPressAve SatVapPressDiff reservestatus,

grow entropy;

partition fraction(VALIDATE = 0.3, SEED =123)

prune costcomplexity;

output out = scored;

run;

E TO

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

EE FOR LOBAOREG

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%

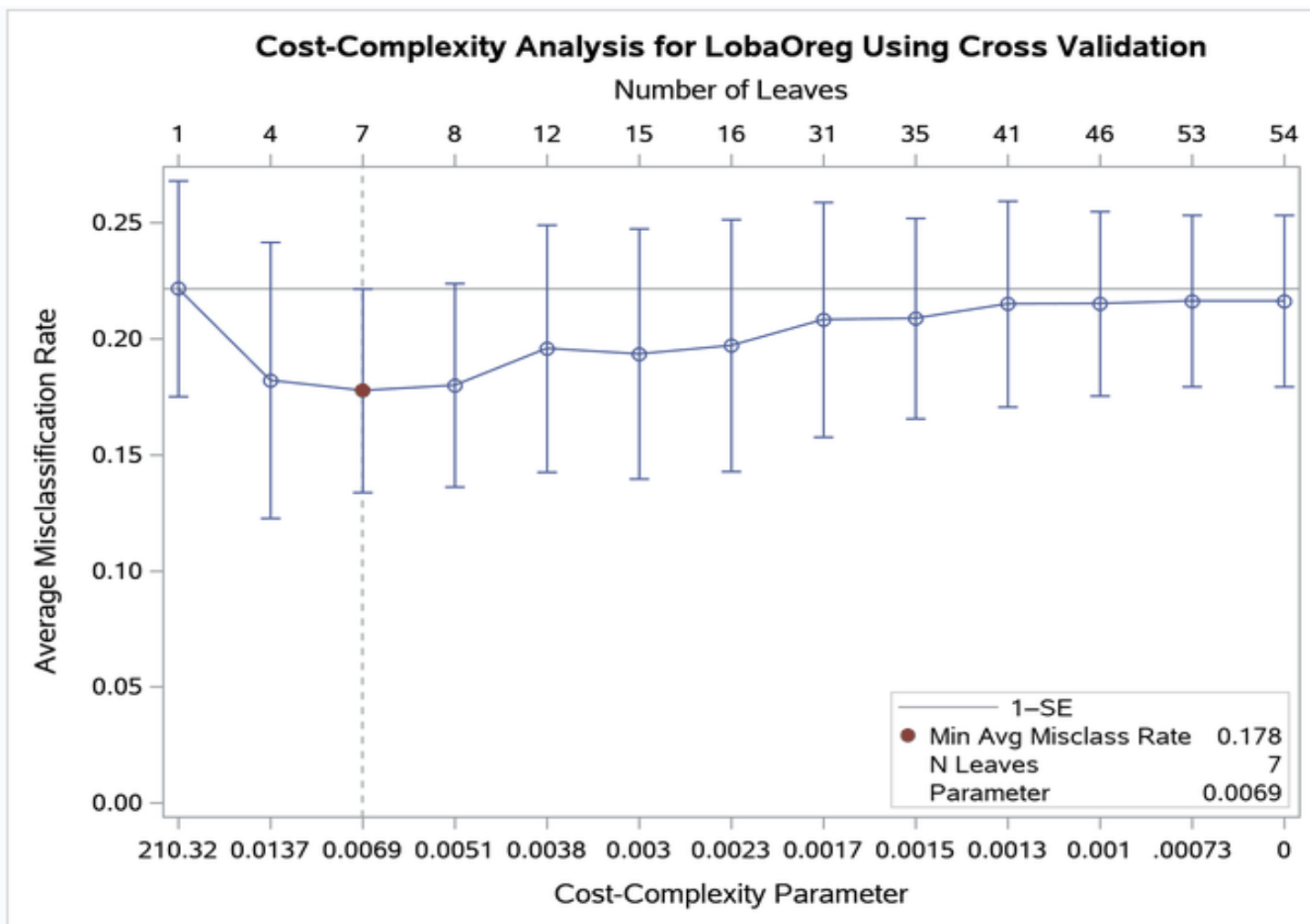
Seed is a parameter for random selection.

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.



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.

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

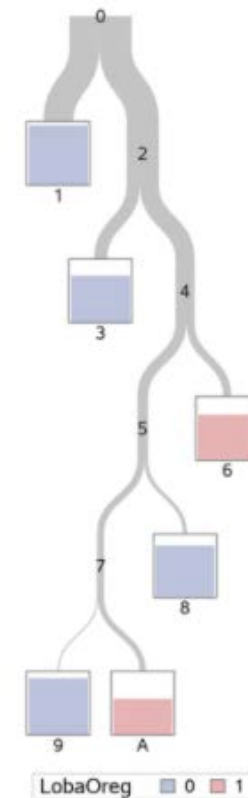
- Selects the smallest subtree for which the misclassification rate is less than the minimum rate plus one standard error. >> Minimum error rate is at 7 leaves, so **select subtree with six leaves.**



OVERVIEW OF FITTED TREE

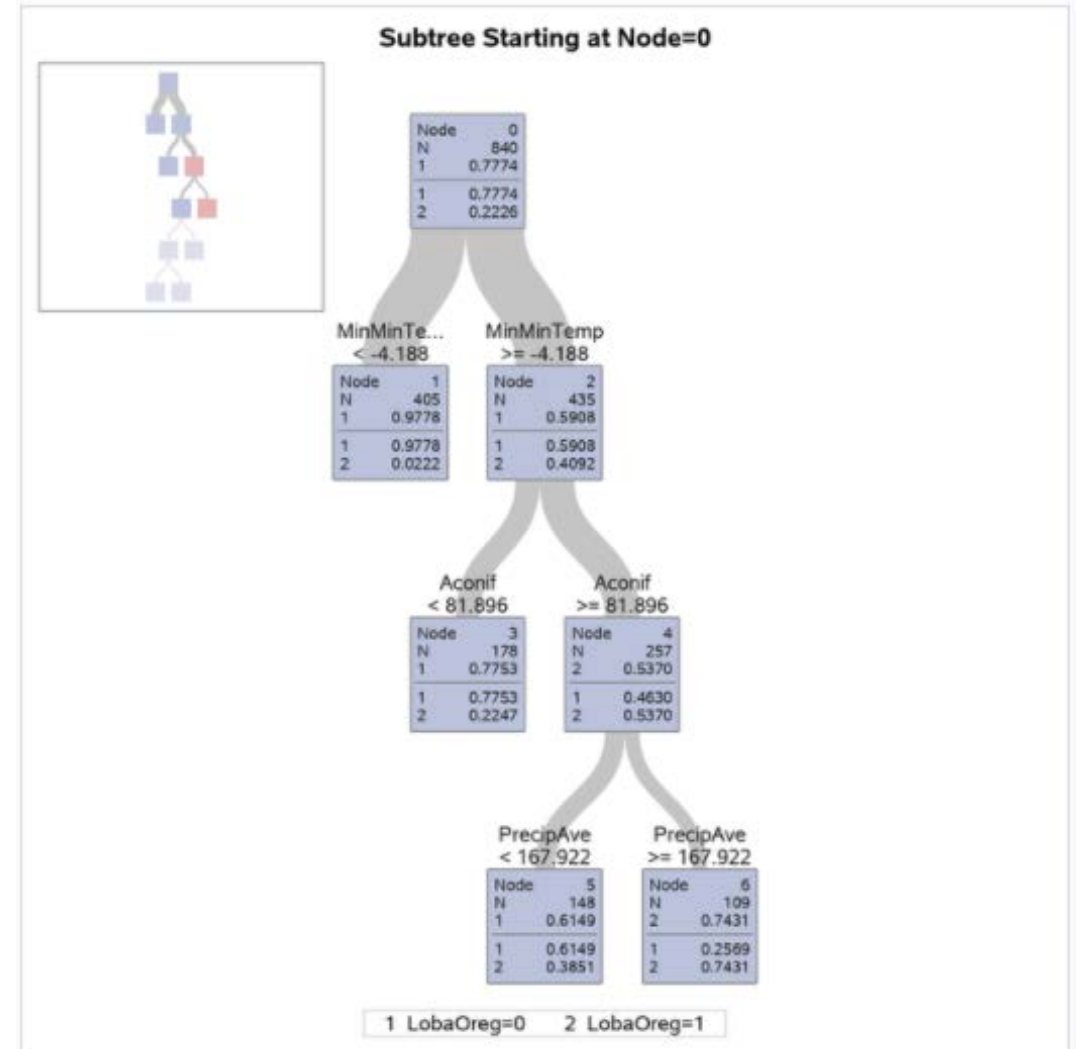
- The **color** of the bar in each leaf node indicates the most frequent level of LobaOreg and represents the **classification level** assigned to all observations in that node.
- The **height** of the bar indicates the **proportion** of observations (sites) in the node that have the most frequent level.
- Note: there is impurity in each leaf node. In other words, there can be LobaOreg observation values 1 and 0 in each node.

Classification Tree for LobaOreg

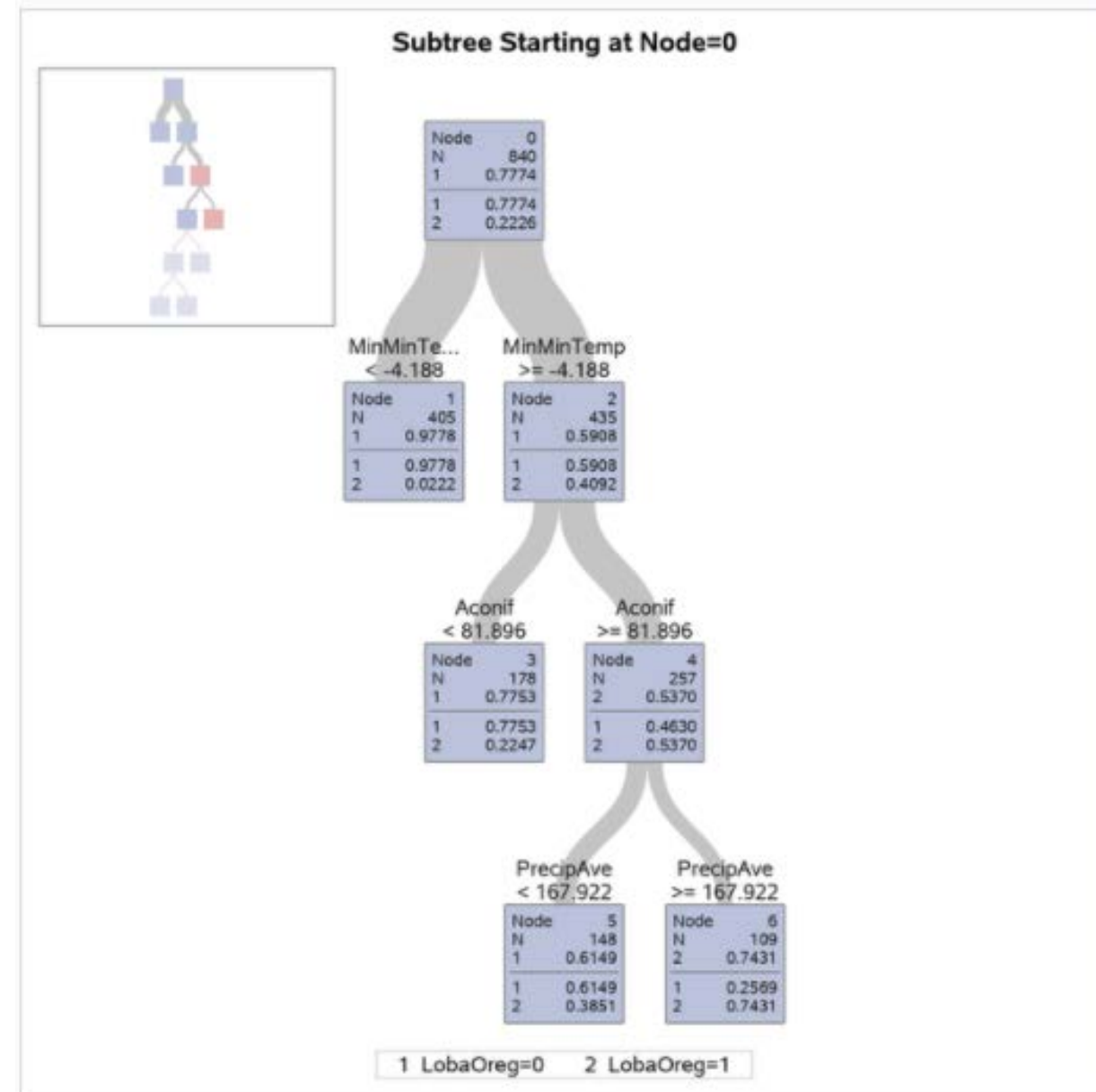


FIRST FOUR LEVELS OF FITTED TREE

- 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.

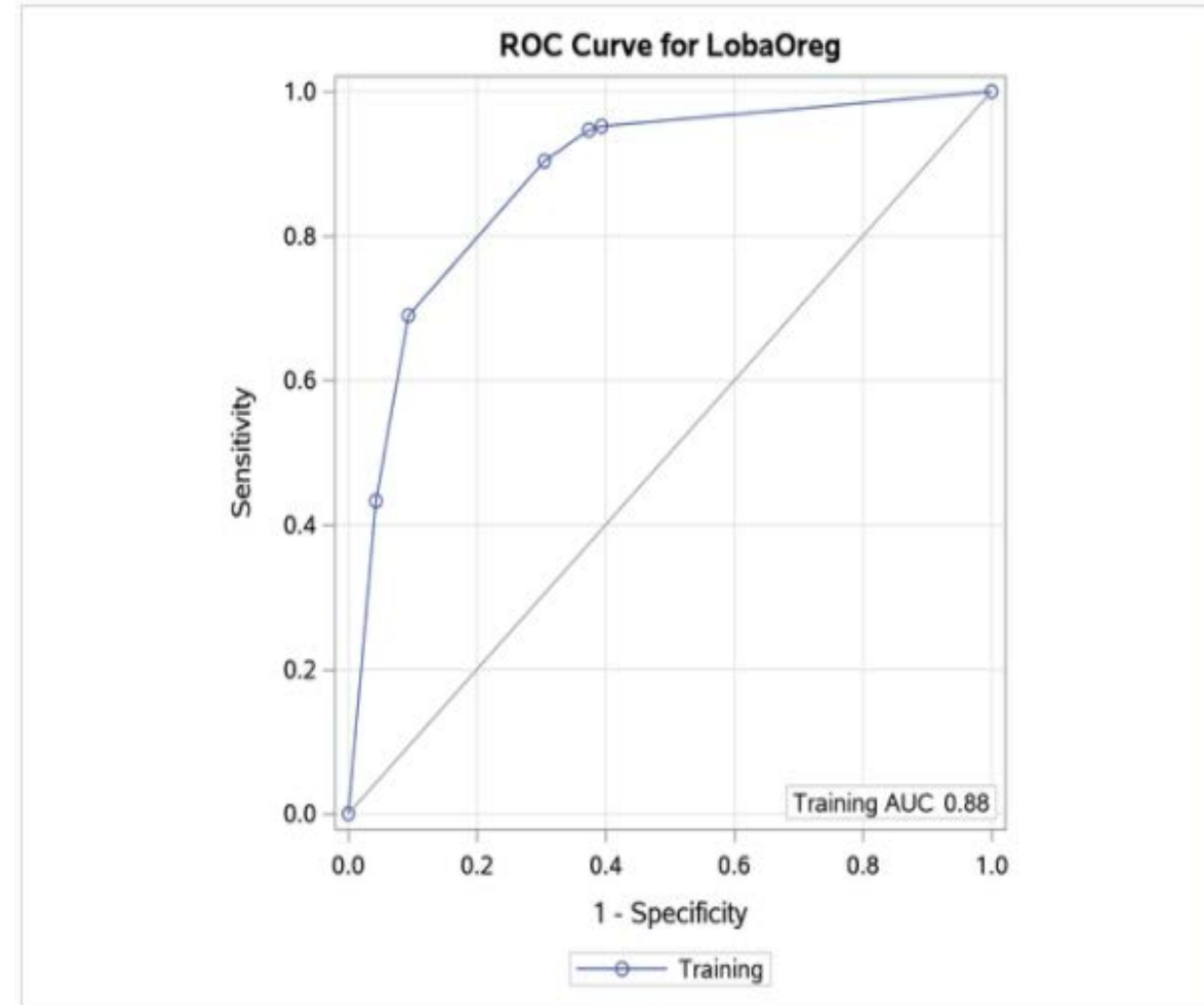


- The 257 sites for which Aconif 81.896 are further subdivided on the basis of **PrecipAve** (average monthly precipitation) with a cutoff value 167.922 mm.
- *Lobaria oregana* was present at **74.31%** of the 109 sites for which MinMinTemp -4.188 , Aconif 81.896 years, and PrecipAve 167.922 mm.
- Contrast 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.

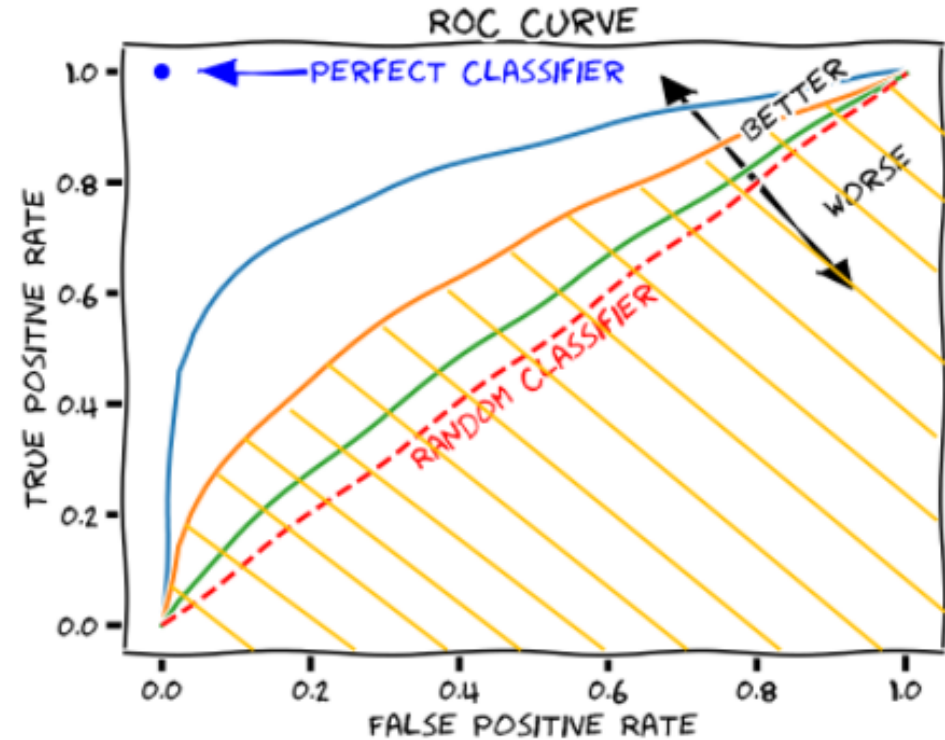
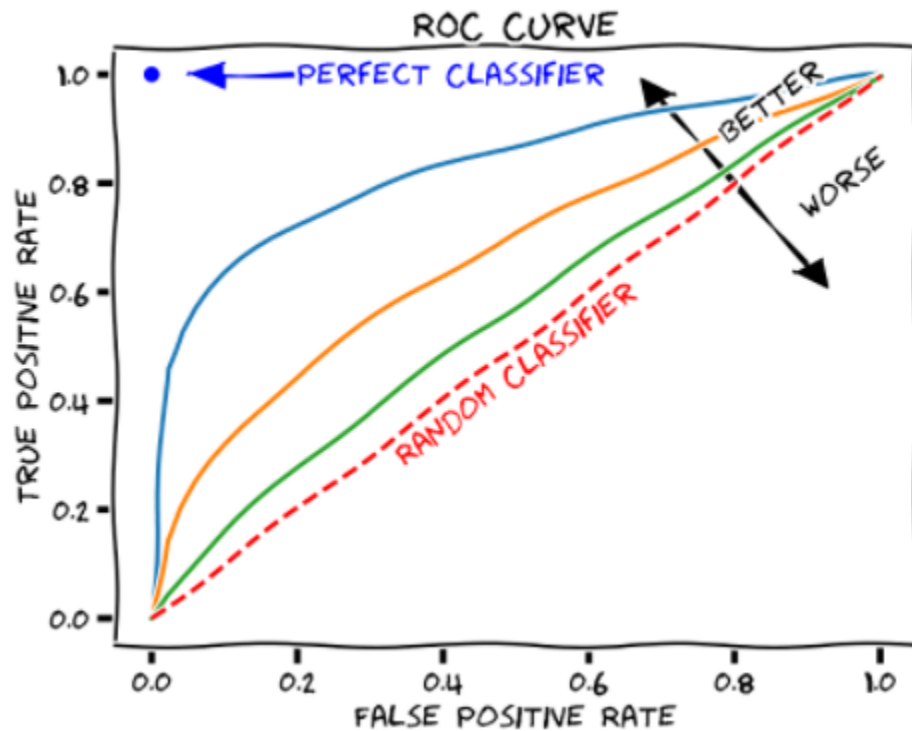


ROC CURVE FOR CLASSIFICATION

- 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 INTERPRETATION



The colored area is AUC (정확성 지표)



REFERENCE

- https://documentation.sas.com/doc/ko/pgmsascdc/9.4_3.4/stathpug/stathpug_hp_split_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>

