**9/4 과제: 서포트 벡터 머신 & 나무 모형**

**데이터 설명: Heart Disease UCI (heart.csv)**

age : age in years

sex : (1 = male; 0 = female)

cp : chest pain type

trestbps : resting blood pressure (in mm Hg on admission to the hospital)

chol : serum cholestoral in mg/dl

fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

restecg : resting electrocardiographic results

thalach : maximum heart rate achieved

exang : exercise induced angina (1 = yes; 0 = no)

oldpeak : ST depression induced by exercise relative to rest

slope : the slope of the peak exercise ST segment

ca : number of major vessels (0-3) colored by flourosopy

thal : 3 = normal; 6 = fixed defect; 7 = reversable defect

target : have disease or not (1=yes, 0=no)

1. **SVM(서포트 벡터 머신)을 이용하여 Classification문제를 해결해보자 (10점)**

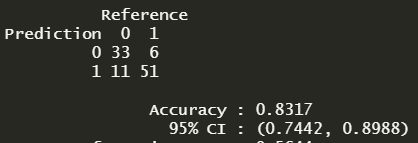
( √ target변수를 as.factor() 함수를 사용해서 범주형 변수로 변환 시켜줘야 함)

**1.1. heart데이터의 2/3을 학습 데이터셋, 1/3을 테스트 데이터 셋으로 분할 하라. (seed값은 1000으로)**

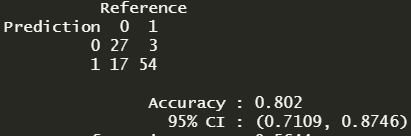
**1.2. 다양한 커널을 사용해서 SVM모델을 만들어라. (커널: radial base, polynomial, sigmoid, linear 등등). 어떤 모델이 가장 분류를 잘 하는가? 그 판단 근거는 무엇인가?**

**- 답: 정확도만을 비교하면 radial base를 사용한 모델이 가장 높으나, 비슷한 정확도의 polynomial (0.80) 모델보다 FN 비율이 높다. 이 데이터는 건강관련 데이터이므로 FN의 cost가 높다고 생각되며, 따라서 비슷한 정확도에 FN의 비율이 더 낮은 polynomial을 사용한 모델이 더 좋다고 할 수 있다.**

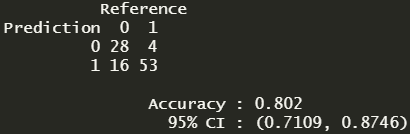
**- radial: Accuracy: 0.83, FP: 11, FN: 6**

****

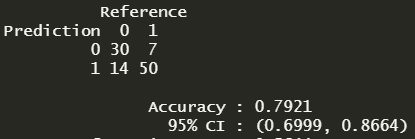
**- polynomial: Accuracy: 0.80, FP: 17, FN: 3**

****

**- sigmoid: Accuracy: 0.80, FP: 16, FN: 4**

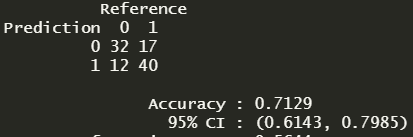
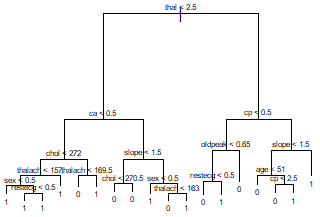
****

**- linear: Accuracy: 0.79, FP: 14, FN: 7**

****

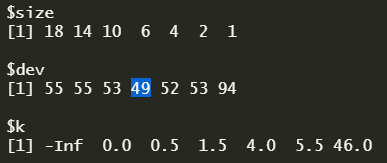
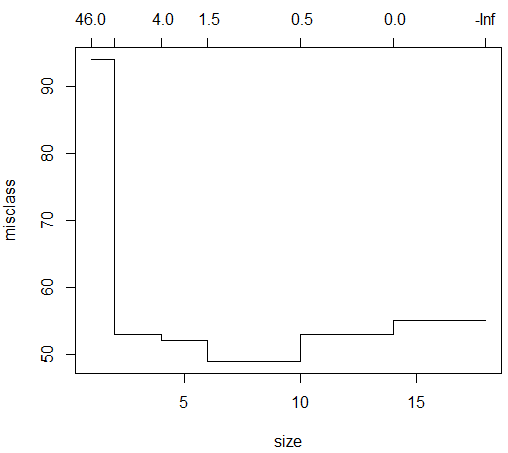
* 1. **Tree모델(model-a)을 만들어라. 정확도는?**

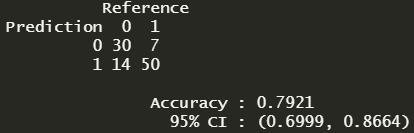
**답: 정확도: 0.7129**

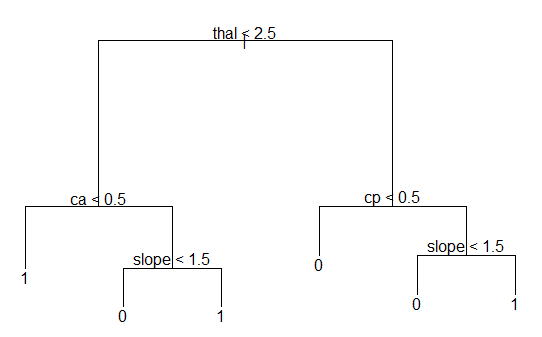
 ****

* 1. **가지치기를 해서 Tree 모델(model-b)을 만들어라. (seed값은 1000) 이때 최적 터미널 노드 수와 정확도는?**

**답: 최적 사이즈 : 6 (deviance: 49로 최저), 정확도 : 0.7921**

****

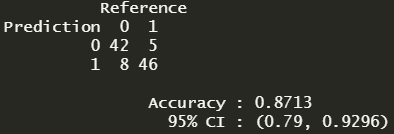
****

****

* 1. **model-a와 model-b 중 정확도가 더 높은 모델은? 그 이유는 무엇인가?**

**답: Model-b 가 accuracy도 더 높고 FN의 비율도 더 낮기 때문에 더 정확도가 높은 모델이라고 볼 수 있다.**

**추가. randomForest로 가장 좋은 모델 만들기**

****

1. **MeanDecreaseGini가 낮은 fbs, restecg 변수 제외**

**Accuracy도 0.87로 가장 높고, FN도 5로 낮은 모델!**