**[마지막 과제]**

**IT응용시스템공학과**

**1494053**

**김희택**

**9.2 FlightDelays.csv는 2004년 1월 동안 워싱턴 D.C에서 출발하여 뉴욕에 도착하는 모든 민간 항공기들에 대한 정보를 담고 있다. 데이터에는 각 항공편에 대한 출발지 및 도착지 공항, 운항 경로 거리, 항공편 예정 날짜 및 시간 등에 관한 정보가 있다. 여기서 예측하고자 하는 변수는 항공편이 연착되는지 아닌지에 대한 여부이다. 연착은 예정된 시간보다 최소 15분 이상 늦게 도착하는 것으로 정의한다.**

**데이터 전처리: DAY\_WEEK 변수를 범주형 변수로 바꾸시오. 항공편 예정 출발 시간을 8구간으로 구간화하시오. (R에서 함수 cut()을 사용하시오) 이 8개의 변수들과 다른 모든 열들(columns)을 예측변수로 사용하시오. (DAY\_OF\_MONTH 변수 제외) 데이터를 학습세트와 검증세트로 나누시오.**

1. **모든 적절한 예측변수들을 사용하여 항공편 연착 변수에 대한 분류나무모델을 만드시오. 예측시점에서는 DEP\_TIME이 알려지지 않으므로 (항공기 이륙 후 연착을 예측하는 것이 아니므로) 모델에는 출발 시간을 포함하지 마시오. 최대 깊이 8수준으로 가지치기된 나무모델을 사용하시오. 나무모델의 결과를 규칙으로 표현하시오.**
2. **월요일 오전 7시에 DCA에서 EWR을 비행해야 된다면, 이 나무모델을 사용할 수 있겠는가? 필요한 다른 정보는 무엇인가? 이 모델은 실제로 사용할 수 있는 모델인가? 중복된 정보는 무엇인가?**