## 머신러닝 데이터 분석 실습

■ 비행기 연착 추측 분류

- 미국 교통부 교통통계국에서 제공된 2011년 10월 한달 간의 비행기 출발/도착 정보를 이용하여 비행기 연착 추측 모델 구축하기
- 데이터 구조: Airport Codes Dataset (공항정보)

| Feature | 의미 |
|---|---|
| Airport_id | 공항코드 |
| City | 도시이름 (공항이 속한) |
| State | 주 이름 (도시가 속한) |
| Name | 공항이름 |

# 머신러닝 데이터 분석 실습

## ■ 비행기 연착 추측 분류

- 데이터 구조: Flight-on-time performance (비행정보)

| feature | 의미 |
|---|---|
| Year | 년도 |
| Quarter | 분기 |
| Month | 월 |
| DayofMonth | 일 |
| DayofWeek | 요일 |
| Carrier | 항공사 코드 |
| OriginAirportID | 출발 공항 코드 |
| DestAirportID | 도착 공항 코드 |
| CRSDepTime | 출발 예정 시각 |

| feature | 의미 |
|---|---|
| DepTimeBlk | 출발 시각 블럭 |
| DepDelay | 출발 지연 (분) |
| DepDel15 | 15분 이상 출발 지연 |
| CRSArrTime | 도착 예정 시각 |
| ArrTimeBlk | 도착 시각 블럭 |
| ArrDelay | 도착 지연 (분) |
| ArrDel15 | 15분 이상 도착 지연 |
| Cancelled | 비행 취소 |
| Diverted | 비행 우회 |

- Target Feature: ArrDel15 (binary classification)

# 머신러닝 데이터 분석 실습

## ■ 비행기 연착 추측 분류

- 데이터 전처리: 비행정보 + 출발 공항정보 병합
  - 공항정보 데이터의 city, state, name 컬럼명을 Ori-city, Ori-state, Ori-airport로 변경(원본 데이터 유지)
  - 비행정보 데이터의 “출발공항ID” 컬럼과 공항정보 데이터의 “공항ID”컬럼을 키 값으로 두 데이터 조인

| 비행정보 <18 개 컬럼> | | |
|---|---|---|
| 출발공항ID | 도착공항ID | … |
| 숫자? | 숫자? | … |
| 숫자? | 숫자? | … |

| 공항정보 < 4개 컬럼 > | | | |
|---|---|---|---|
| 공항ID | (출발) 도시 | (출발) 지역 | (출발) 항공사명 |
| 숫자 | … | … | … |
| 숫자 | … | … | … |

| 비행 + (출발) 공항 정보 <21(22)개 컬럼> | | | | | | |
|---|---|---|---|---|---|---|
| 출발공항ID | 도착공항ID | …… | (출발공항정보) | | | |
| | | | 공항ID | 도시 | 지역 | 항공사명 |
| 숫자? | 숫자? | … | 숫자? | … | … | … |
| 숫자? | 숫자? | … | 숫자? | … | … | … |

# 머신러닝 데이터 분석 실습

## ■ 비행기 연착 추측 분류

- 데이터 전처리: 비행정보 + 출발 공항정보 + 도착 공항정보 병합
  - 공항정보 데이터의 city, state, name 컬럼명을 Dest-city, Dest-state, Dest-airport로 변경(원본 데이터 유지)
  - 비행정보 데이터의 “도착공항ID” 컬럼과 공항정보 데이터의 “공항ID”컬럼을 키 값으로 두 데이터 조인

| 비행 + 출발공항 정보 <21(22)개 컬럼> | | | | | | |
|---|---|---|---|---|---|---|
| 출발 공항ID | 도착 공항ID | … … | (출발공항정보) | | | |
| | | | 공항ID | 도시 | 지역 | 항공사명 |
| 숫자? | 숫자? | … | 숫자 | … | … | … |
| 숫자? | 숫자? | … | 숫자 | … | … | … |

| 공항정보 < 4개 컬럼 > | | | |
|---|---|---|---|
| 공항ID | (도착) 도시 | (도착) 지역 | (도착) 항공사명 |
| 숫자 | … | … | … |
| 숫자 | … | … | … |

| 비행 + 출발공항 + 도착공항 정보 <24(~26)개 컬럼> | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 출발 공항ID | 도착 공항ID | …… | (출발공항정보) | | | | (도착공항정보) | | | |
| | | | 공항ID | 도시 | 지역 | 항공사명 | 공항ID | 도시 | 지역 | 항공사명 |
| 숫자? | 숫자? | … | 숫자 | … | … | … | 숫자 | … | … | … |
| 숫자? | 숫자? | … | 숫자 | … | … | … | 숫자 | … | … | … |

# 머신러닝 데이터 분석 실습

## ■ 비행기 연착 추측 분류

- 데이터 전처리
  - 24개 컬럼 중 학습에 사용할 14개 컬럼만 선택

DayOfWeek, Carrier, DepTimeBlk, DepDelay, DepDel15, ArrTimeBlk, ArrDel15
Ori-city, Ori-state, Ori-airport, Dest-city, Dest-state, Dest-airport, ArrDelay

  - 결측 데이터 처리: 결측치를 포함한 행 삭제
  - 명목형 변수 인코딩: 라벨 인코딩