# Test 1

## Test Details/Instructions

Worth: 33.3%

Due Date: May 26th, 2022 @ 11:59pm

### Submissions:

1. Your python notebook with all the codes to all the questions
2. Explanation of your results written as text in your python notebook
3. Submit your python notebook as: ***firstname\_lastname\_Test1***

### Questions:

1. The file FlightDelays.csv contains information on all commercial flights departing the Washington, DC area and arriving at New York during January 2004. For each flight, there is information on the departure and arrival airports, the distance of the route, the scheduled time and date of the flight, and so on. The variable that we are trying to predict is whether or not a flight is delayed. A delay is defined as an arrival that is at least 15 minutes later than scheduled. Data Preprocessing.
2. Drop Columns: Carrier, dep\_time, fl\_date, fl\_num, day\_of\_month, tail\_num
3. Transform variable day of week info a categorical variable and any other variables necessary
4. Bin the scheduled departure time into eight bins (in Python use function pd.cut() from the pandas package).
5. Partition the data into training (60%) and validation (40%).
6. Run a boosted classification tree for delay (flight status). Except for setting n\_estimators=500 and random\_state=1, use default setting for the DecisionTreeClassifier and the AdaBoostClassifier.
7. Compared with the single tree, how does the boosted tree behave in terms of overall accuracy?
8. Compared with the single tree, how does the boosted tree behave in terms of accuracy in identifying delayed flights?
9. Explain why the boosted model may have outperformed the non-boosted model.
10. The Institute for Statistics Education at Statistics.com offers online courses in statistics and analytics, and is seeking information that will help in packaging and sequencing courses. Consider the data in the file CourseTopics.csv, the first few rows of which are shown in Table 14.14. These data are for purchases of online statistics courses at Statistics.com. Each row represents the courses attended by a single customer. The firm wishes to assess alternative sequencings and bundling of courses. Use association rules to analyze these data, and interpret several of the resulting rules.

*HINT: Please refer to all the metrics used in association rules and explain the use of those metrics in the comment sections. You decide the parameters in your rules along with a simple explanation.*

1. The file EastWestAirlinesCluster.csv contains information on 3999 passengers who belong to an airline’s frequent flier program. For each passenger, the data include information on their mileage history and on different ways they accrued or spent miles in the last year. The goal is to try to identify clusters of passengers that have similar characteristics for the purpose of targeting different segments for different types of mileage offers.
   1. Remove unnecessary columns from your analysis
   2. Apply kmeans clustering to the dataset
   3. Showcase the optimal # of clusters for your analysis and how you came up with it?
   4. Showcase the difference between the optimal clusters
   5. Which clusters would you target for offers and why?