**CS673 – Group Project:**

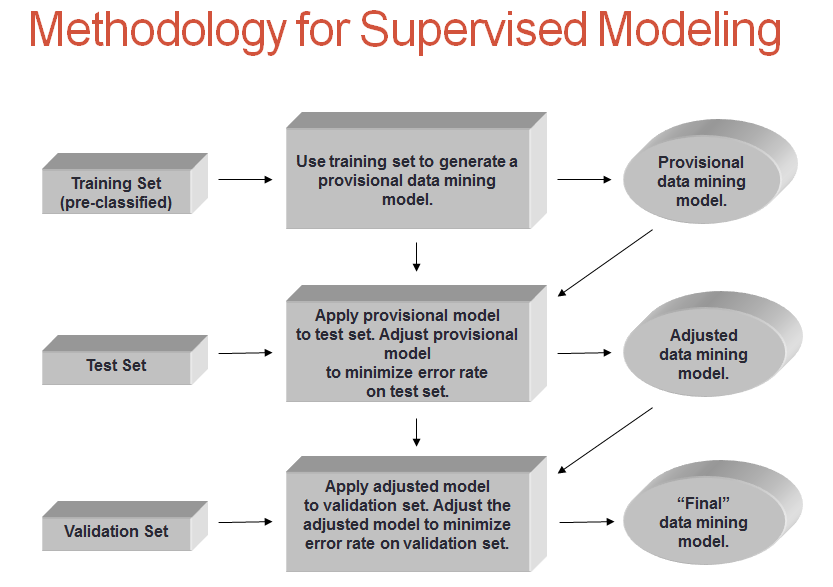
**Part 2:**  For a maximum of 17.5 points

Each team will do the following steps

1. Create a new table in Hive with the format of your year table. Name it “sample” with your name (***yourname*\_sample**)
2. Add a new column to the ***yourname*\_sample** table and label it **Delayed**
3. Select a **random** sample of 30,000 records from all across your year table. See **https://hadoopsters.com/how-random-sampling-in-hive-works-and-how-to-use-it-7cdb975aa8e2**
4. Load your 30,000 sample records into this table and check that all the records are loaded.
5. Populate the Delayed column as follows
   1. if the record has a value of zero or less in the ArrDelay and DepDelay columns, put a “N” in the Delayed column
   2. otherwise, put a “Y” in the Delayed column
6. Extract your 30,000 sample records from Hive and download it to your local device.
7. Combine your file with the other sample year files from your other team members. (180,000 records in total)
8. Each member will choose a different machine learning model that can determine (score) whether new unknown flights have delays or not.
9. Train, validate and test the analytics model with the team’s combined sample data (180,000 records).
10. Team members will choose the team model that best determines whether a record is delayed or not, with at least 90% accuracy and 90% precision. You are not restricted to using RapidMiner for the training, testing and validation work. You can use Python, R, or any other tool you are more comfortable with. **MAKE SURE YOU CAN EXPLAIN THIS WORK IN DETAIL.**
11. Using the model chosen by the team, each member will work with ten records provided by the instructor to determine whether a flight is delayed or not and score it with a Y or N accordingly. (See example below)



**What you need to Upload:**

1. Each individual member will upload the team’s power-point presentation to Classes.
2. The presentation will contain slides legible screenshots of the successful completion of the steps listed above for each individual’s work, as well as a final summary slide with any conclusions.
3. A spreadsheet of your 10 predicted records.
4. 
5. **Include a matrix in the slides describing findings. See example below:**

