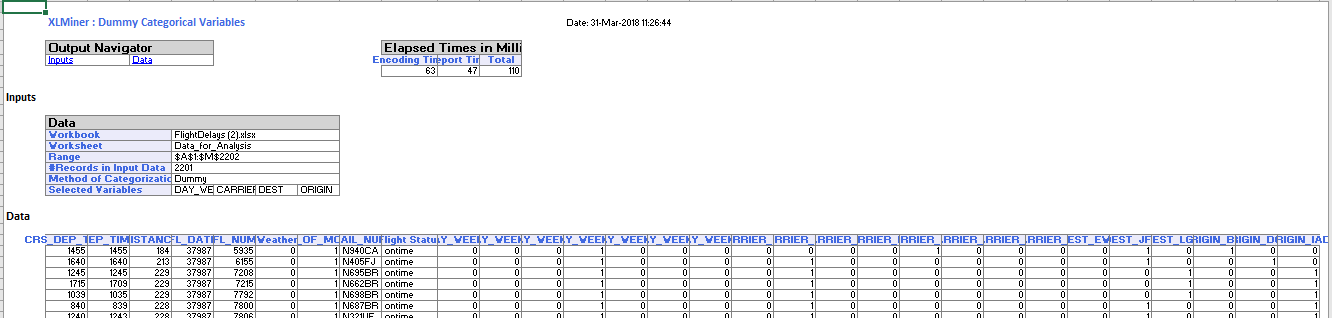
**Name: Surabhi Maheshwari Ninernet: 801028703**

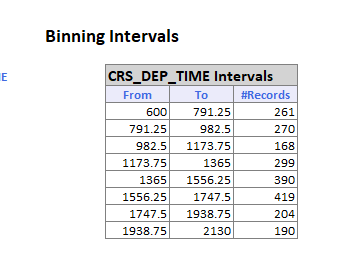
**HOMEWORK 5**

**BIG DATA ANALYTICS FOR COMPETITIVE ADVANTAGE**

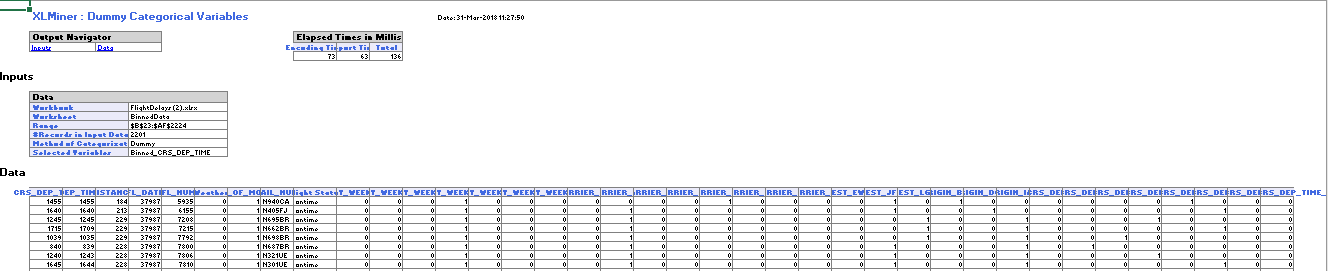
**Data Preprocessing.** Create dummies for day of week, carrier, departure airport, and arrival airport. This will give you 17 dummies.



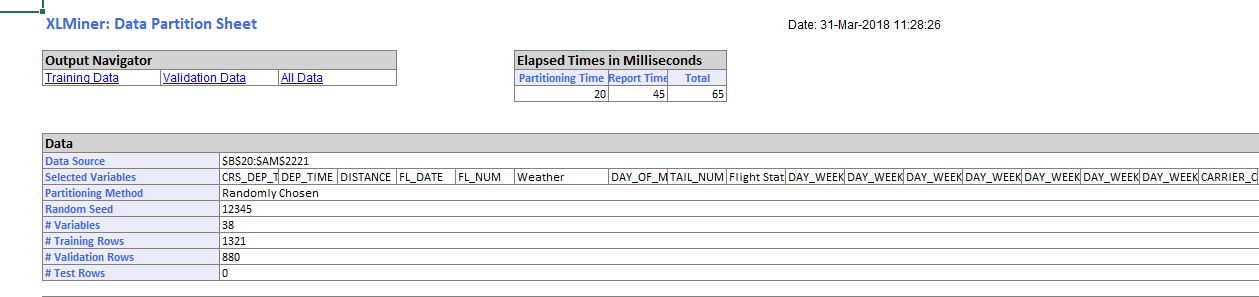
Bin the scheduled departure time into eight bins (in XLMiner use Transform 🡪 Bin Continuous Data and select equal width).



After binning CRS DEP TIME into the 8 bins, this new variable should be broken down into dummies (because the effect will not be linear, due to the morning and afternoon rush hours). This will avoid treating the departure time as a continuous predictor, since it is reasonable that delays are related to rush-hour times.



Partition the data into training and validation sets.

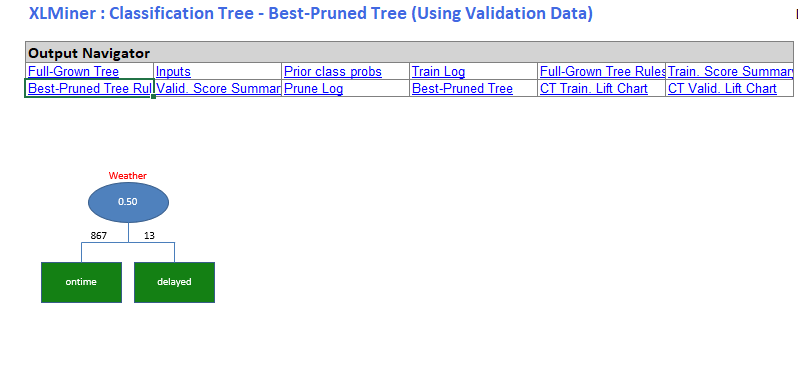


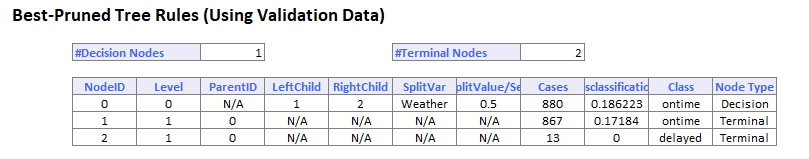
1. Fit a classiﬁcation tree to the ﬂight delay variable using all the relevant predictors. Do not include DEP\_TIME (actual departure time) in the model because it is unknown at the time of prediction (unless we are generating our predictions of delays after the plane takes off, which is unlikely). In the third step of the Classiﬁcation Tree menu, choose “Maximum # levels to be displayed = 6.” Use the best-pruned tree, setting the minimum number of observations in the ﬁnal nodes to 1. Express the resulting tree as a set of rules.

The following variables are included:

1. Weather
2. Day of week
3. Carrier
4. Origin
5. Destination
6. Binned\_CRS\_DEP\_TIME

The other predictors are irrelevant and thus they can be removed.





1. If you needed to ﬂy between DCA and EWR on a Monday at 7 AM, would you be able to use this tree?

No, we cannot predict the delay with the help of the tree because we can just predict if the flight is ontime or delayed based on weather. The details of destination, origin, day of week and departure time are not included in the tree.

What other information would you need?

We need to have the origin airport(DCA), destination airport(EWR), day of week(Monday) and CRS\_DEP\_TIME(7am) in the tree which is not there. We are just getting a tree with single node which is Weather.

Is it available in practice? What information is redundant?

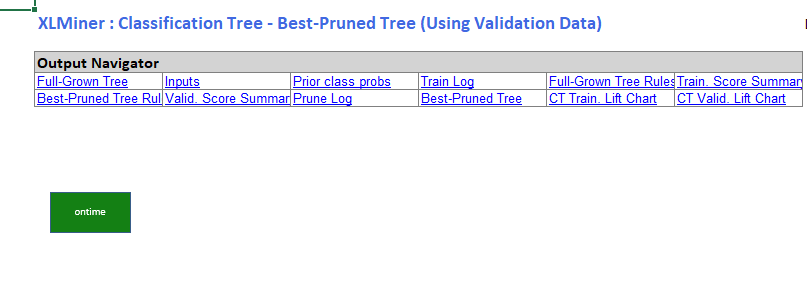
Yes, it is available in practice i.e. practically this information is needed for predicting the flight status. No information is redundant as there is only one node which is weather.

1. Fit another tree, this time excluding the Weather predictor.

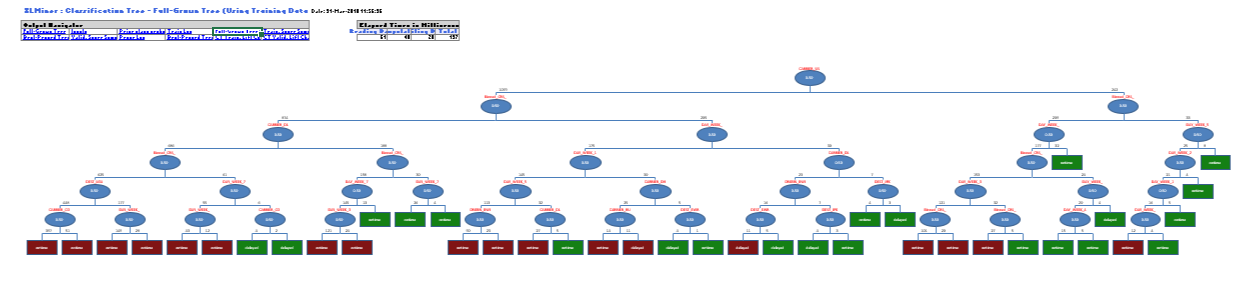
(Why?) The weather cannot be used as a predictor as weather can change and the correct result cannot be derived based on weather. At times, weather prediction cannot be made, or if made the prediction can be wrong, which can hamper the final result.

Select the option of seeing both the full tree and the best-pruned tree. You will ﬁnd that the best-pruned tree contains a single terminal node.

Best Prune Tree:



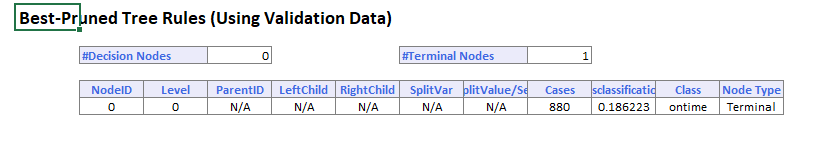
Full Tree:



* 1. How is this tree used for classiﬁcation? (What is the rule for classifying?)

Tree can be used to derive the rules for classification of the dataset and conclude an answer for given conditions.

For classifying, the right branch of the tree is selected which means true. The branch is traversed until we come to a leaf node and the final rule generated is the classifying rule.



* 1. To what is this rule equivalent?

The rule is equivalent to a single node which cannot be used to predict the final output. The rules are equivalent to the tree which contains only single node which is generated by classifying the dataset using a classification algorithm.

* 1. Examine the full tree. What are the top three predictors according to this tree?

The top three predictors in this tree are:

1. Carrier\_US
2. Binned\_CRS\_DEP\_TIME\_5
3. Carrier\_DL
   1. Why, technically, does the pruned tree result in a tree with a single node?

The pruned tree results in a single node as all the other redundant nodes are pruned out of the tree. The data contains biased target classes; thus the tree cannot grow.

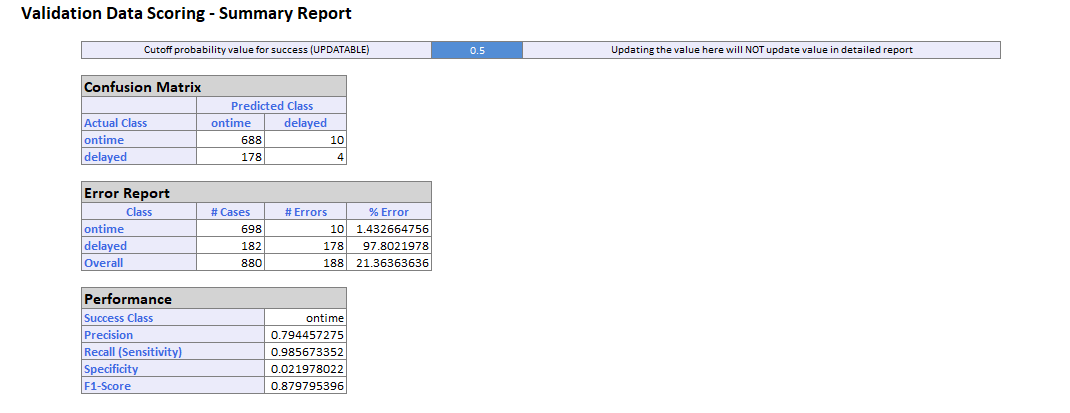
* 1. What is the disadvantage of using the top levels of the full tree as opposed to the best-pruned tree?

The full tree has the problem of overfitting which can be removed by best pruned tree. The top levels of full tree contain many redundant nodes and for traversing the full tree the user must go through all the redundant nodes. Whereas the best pruned tree does not have redundant nodes as pruning is the method of cutting off the unnecessary branches. Thus, the performance is improved by best pruned tree.

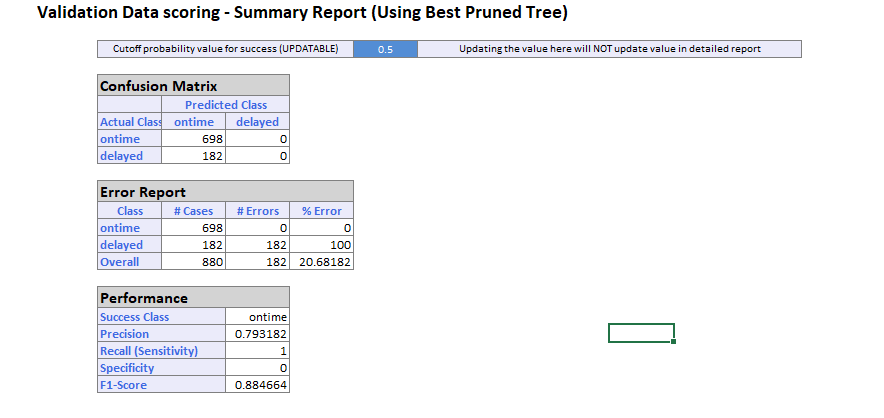
* 1. Compare this general result to that from logistic regression in the example in Chapter 10. What are possible reasons for the classiﬁcation tree’s failure to ﬁnd a good predictive model?

The classification tree has a problem of overfitting which can be eliminated by using logistic regression. The predictive model using logistic regression is more interpretable as compared to classification tree.

The validation report for logistic regression is as follows:



The validation report for classification tree is as follows:



The precision is more in case of logistic regression as compared to that of classification tree.

**\*Detailed report can be seen in the attached excel file.**