# AI Academy: Introduction to Data Mining
# Week 4 Workshop

Workshop 4 contains 2 questions. Please read and follow the instructions.

# 1 Decision Tree Construction

Create decision trees **by hand** for the `hw q1.csv` Titanic survival dataset, as explained below. Note the following:

In the given dataset, all of the input attributes are binary except for the "Pclass" which is categorical. "Pclass" should create a 3-way split if used in the tree.

The output label has two class values: T or F, which represent Survival or Not Survival.

In the case of ties when selecting an attribute, break ties in favor of the leftmost attribute.

When considering a split for the continuous attribute, identify the best value to split on (e.g. $\leq 15$ and $> 15$) by testing all possible split values.

You must show your work when calculating Gini Index for *the split at the root node* (but not for later splits). You can do so by either 1) writing out substeps (e.g. conditional entropy for each child node), or including a code for a program you used to make your calculations.

You should draw a separate tree, like the example in Figure 1, after each attribute split. You can use a program (e.g. tikz with LaTeX, Lucidchart, etc.) to draw your trees, or draw them by hand on paper and scan your results into the final pdf.

The instructor will go though the first split as an example, then you will complete the rest.

1. Construct the decision tree manually, using Gini index to select the best attribute to split on. The maximum depth of your tree should be 3 (count the root node as depth 0), meaning that any node at depth 3 will automatically be a leaf node, even if it has objects with different classes.

## 2 Evaluation Measures & Pruning

This analysis pertains to the *IBM Attrition* dataset, which includes attributes about employess and whether they left the company (Yes/No). The main goal of the analysis is to study the indicators of attrition in order to identify ways that the company can improve employee retention to save money and time spent in hiring and training. To predict the attrition, consider using the decision tree shown in Figure 1 which involves Business Travel Frequency (BTF), Gender, Marital Status (MS) and Engineer Level (EL). Complete the following tasks:

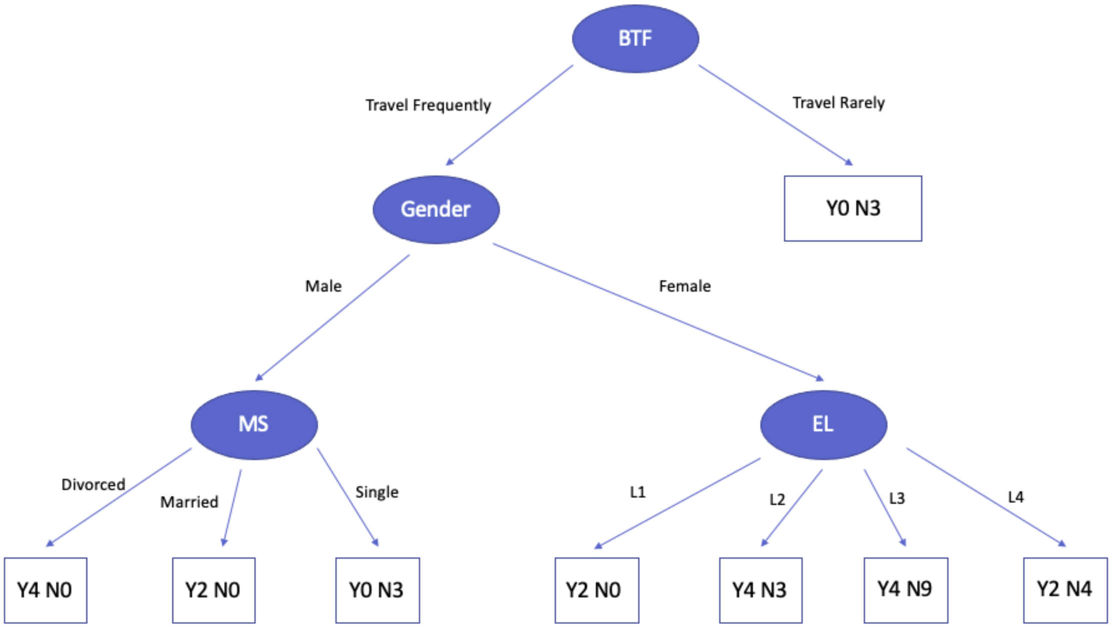

Figure 1: Decision Tree

1. Use the decision tree above to classify the provided dataset. `hw q2.csv`. Construct a confusion matrix and report the test Error Rate. Use "Yes" as the positive class in the confusion matrix.
2. Calculate the optimistic training classification error before splitting and after splitting using **EL**, respectively. **Consider only the subtree starting with the EL node.** If we want to minimize the optimistic error rate, should the node's children be pruned?
3. Calculate the pessimistic training errors before splitting and after splitting using **EL** respectively. Consider only the subtree starting with the EL node. When calculating pessimistic error, use a leaf node error penalty of 0.8. If we want to minimize the pessimistic error rate, should the node's children be pruned?
4. Assuming that the "EL" node is pruned, recalculate the test Error Rate using `hw2q2.csv`. Based on your evaluation using the test dataset in `hw2q2.csv`, was the original tree (with the EL node) over-fitting? Why or why not?