Our implementation successfully implements the ID3 classification algorithm on multivalued class.. it was assumed the training and test data were well formed.

Training process

The implementation is a recursive top down implementation of the greedy-search ID3 classification method.. We first strip down the header row from training data then kick off computation by assigning the head node of the tree to **ID3Rec(DataClean, attrsList)**.. we carry around in our recursive calls the examples that are relevant for computing that branch (filtered on the attribute selected for each value of the children) of the tree as well as a list of currently still considered attributes..
The principle then is simple, **the base case is when all the data elements are homogenous or if there are no more attributes to consider classification on** then a leaf node is created the most common value of examples.
The non trivial case is considering each attribute once at a time and calculating the gain if that attribute was to be picked, this is done with the help of the **example filtering method**, to consider each attribute in turn computing the **infoGain** for each and greedily **picking the highest gain attribute, then computing the TreeNode children recursively for each value of the attribute**.
InfoGain takes care to consider any number of values for classes and this can be seen in the **entropy** function and specifically the loop on **line 283-290,** for computing this method we iterate over possible class values (indexed using idx search method that does a reverse lookup on strings[attributes-1]) we make use of the supplied xlogx function. This works for **any number of values for the class attribute.**

One difficulty was encountered with the potentially null values in strings[attributes-1][x] due to way the index was constructed those, are automatically skipped in relevant parts (while computing gain and entropy).

Classification details

Classification iterates over each test set example, then starting with the top level node value (the first attribute to consider) recursively traverses the decision process down the tree, doing down the child corresponding to the value for the attribute on that example.. when a leaf node is encountered then that is the class of the example.. With the datasets given this yield identical results.