Classification using Decision Trees in Excel

As you now know, decision trees usually involved drawing out a tree, and at each branch you make a decision. We will use Excel to decide which branching is "best," using information gain.[1]

Big idea: compare branches using information gain

Information gain (IG) tells us how much information we gain by creating a certain branch point in our decision tree.

- Look at the outcomes that are attached to the branch (call this set a) and figure out the entropy of all those outcomes (calculate $E(a)$).
- Look at the separate branches you get and for each branch b, calculate the entropy of that branch alone. (That's $E(b)$.)
- The information gain is $E(a) - average(E(b))$, where the average is over each branch.

Example Excel spreadsheet

1. Find, download, and open Decision Tree Spreadsheet.xlsx.
2. We are interested in figuring out who defaults on loans. Banks want to know who fails to pay back loans so that they can stop giving loans to people who won't pay them back! Many groups also want to educate people who take out loans so that they can increase repayment rates. Imagine if you could target small-business training to groups who take out small-business loans but sometimes run into trouble....
3. In cells A21 through E31, you'll find some records (more fake data!) about people who may have defaulted on loans. The set of observations has an ID for each person, whether they're a homeowner or not, their marital status, income information, and whether or not they have defaulted on a loan. I colored this area yellow.
4. In the upper right, I have tables that look at every variable (homeowner, marital status, and income) and split the original data set along those variables.
   (a) First check the "rules" I set up – does it make sense? The formula in M3, for instance, counts all the records that match the conditions in G2 through J3. In Excel, we write `=DCOUNT($A$21:$E$31,"ID",G2:J3)`.
      i. DCOUNT is a command that counts in a "database" or collection of cells.
      ii. `$A$21:$E$31` tells you the cells in the "database." They're the ones in yellow in the spreadsheet.
      iii. "ID" tells the `DCOUNT` command to count all the separate IDs that match the condition ...
      iv. And the condition is given by `G2:J3`.
   (b) By hand, count the number of people who are not homeowners and did not default. Does that match the output you get in cell M3?
   (c) To get the formula for N3, I just copied the formula from M3. When you copy a formula in Excel, the inputs change too. The formula in N3 now says `=DCOUNT($A$21:$E$31,"ID",H2:K3)`

[1] These instructions are for a spreadsheet made by Kaisa Taipale based on the paper, "Teaching Decision Tree Classification Using Microsoft Excel" by Kaan Ataman, George Kulick, Thaddeus Sim.

(d) If you write $ by numbers and letters (if you write `$A$21`, for instance), the row and column don't change when you copy the formula to a different place on the spreadsheet. You'll always get the contents of the cell in column $A$ in row 21.

(e) If you don't write $, then copying the formula will change the inputs, shifting the rows and columns according to where you copy the formula. Why does N3 have the condition H2:K3 now?

(f) N3 should be counting the number of people who are not homeowners and did default.

(g) *If you are confused, stop and talk to a TA! This is hard to write out – it makes a lot more sense when you just LOOK at it.*

5. The cell O3 just adds up the previous two counts. That means it should give the total of people who are not homeowners.
6. Cell Q3 does the entropy calculation for non-homeowners who did not default on a loan; cell R3 does the entropy calculation for non-homeowners who did default on a loan.
7. Cell S3 adds up those entropies.
8. Now instead of splitting on homeowner/not homeowner, we can look at married/divorced/single, and then low/average/high income. For each of these we'll do the same kinds of calculations: count up the people in each category and do the entropy calculations.
9. In the end, we want the lowest entropy (most certainty). Just pick the final entropy that's lowest!

Now you try! Experiment!
This process shows you how to pick the variable that gives you the lowest entropy – the most information gained by splitting on that variable. What is the next variable you should split on? Can you repeat the process I described above to pick the next variable to split on? Remember you'll have to do this separately for each branch.