1. Given the following data, create a decision tree as ID3 would.

x_1	x_2	x_3	x_4	y
T	T	T	T	+
T	T	F	F	+
T	F	T	F	-
T	F	F	T	-
F	T	T	T	+
F	T	F	F	-
F	F	T	F	+
F	F	F	T	-

First, start out by computing the information gain by splitting on each individual feature (x_1, x_2, x_3, x_4):

$$
\begin{aligned}
H(Y) &= -\left[\frac{4}{8}\log_2\left(\frac{4}{8}\right) + \frac{4}{8}\log_2\left(\frac{4}{8}\right)\right] = -\left[-1-1\right] = 1 \\
I(Y|x_1) &= H(Y) - H(Y|x_1) = 1 - \left[\frac{4}{8}H(Y|x_1 = \mathrm{T}) + \frac{4}{8}H(Y|x_1 = \mathrm{F})\right] \\
&= 1 - \left[\frac{1}{2}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) + \frac{1}{2}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\right] \\
&= 1 - \left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] = 1 - \left[0.5 + 0.5\right] = 0 \\
I(Y|x_2) &= H(Y) - H(Y|x_2) = 1 - \left[\frac{4}{8}H(Y|x_2 = \mathrm{T}) + \frac{4}{8}H(Y|x_2 = \mathrm{F})\right] \\
&= 1 - \left[\frac{1}{2}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \frac{1}{2}\left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right)\right] \\
&= 1 - \left[\frac{1}{2}\left(0.311 + 0.5\right) + \frac{1}{2}\left(0.5 + 0.311\right)\right] = 1 - \left[0.311 + 0.5\right] = 0.189 \\
I(Y|x_3) &= H(Y) - H(Y|x_3) = 1 - \left[\frac{4}{8}H(Y|x_3 = \mathrm{T}) + \frac{4}{8}H(Y|x_3 = \mathrm{F})\right] \\
&= 1 - \left[\frac{1}{2}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \frac{1}{2}\left(-\frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right)\right] \\
&= 1 - \left[\frac{1}{2}\left(0.311 + 0.5\right) + \frac{1}{2}\left(0.5 + 0.311\right)\right] = 1 - \left[0.311 + 0.5\right] = 0.189 \\
I(Y|x_4) &= H(Y) - H(Y|x_4) = 1 - \left[\frac{4}{8}H(Y|x_4 = \mathrm{T}) + \frac{4}{8}H(Y|x_4 = \mathrm{F})\right] \\
&= 1 - \left[\frac{1}{2}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) + \frac{1}{2}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right)\right] \\
&= 1 - \left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right] = 1 - \left[0.5 + 0.5\right] = 0
\end{aligned}
$$

There are only two features that give non-zero information gain. Let's arbitrarily pick x_2 as the first feature to split on. The entropy of this state is shown in the intermediate calculations above ($0.311 + 0.5 = 0.811$). Now,

solving for the true branch of x_2:

$$
\begin{aligned}
H(Y) &= 0.811 \\
I(Y|x_1) = H(Y) - H(Y|x_1) &= 0.811 - \left[\frac{2}{4}H(Y|x_1 = \mathrm{T}) + \frac{2}{4}H(Y|x_1 = \mathrm{F})\right] \\
&= 0.811 - \left[\frac{1}{2}\left(-\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) + \frac{1}{2}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right] \\
&= 0.811 - \left[\frac{1}{2}(0) + \frac{1}{2}(0.5)\right] = 0.811 - [0.25] = 0.561 \\
I(Y|x_3) = H(Y) - H(Y|x_1) &= 0.811 - \left[\frac{2}{4}H(Y|x_3 = \mathrm{T}) + \frac{2}{4}H(Y|x_3 = \mathrm{F})\right] \\
&= 0.811 - \left[\frac{1}{2}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \frac{1}{2}\left(-\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right)\right] \\
&= 0.811 - \left[\frac{1}{2}(0) + \frac{1}{2}(0.5)\right] = 0.811 - [0.25] = 0.561 \\
I(Y|x_4) = H(Y) - H(Y|x_4) &= 0.811 - \left[\frac{2}{4}H(Y|x_4 = \mathrm{T}) + \frac{2}{4}H(Y|x_1 = \mathrm{F})\right] \\
&= 0.811 - \left[\frac{1}{2}\left(-\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) + \frac{1}{2}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right] \\
&= 0.811 - \left[\frac{1}{2}(0) + \frac{1}{2}(0.5)\right] = 0.811 - [0.25] = 0.561
\end{aligned}
$$

All the features are equal, so we can arbitrarily pick one to split on, and construct the rest of the tree from there. Looking at the boolean logic, it looks as though $x_2 \wedge (x_1 \vee x_3 \vee x_4)$ would explain this branch and the underlying logic, so the rest of the subtree can be computed using this. Solving for information gain for 100

$$
\begin{aligned}
H(Y) &= 0.811 \\
I(Y|x_1) = H(Y) - H(Y|x_1) &= 0.811 - \left[\frac{2}{4}H(Y|x_1 = \mathrm{T}) + \frac{2}{4}H(Y|x_1 = \mathrm{F})\right] \\
&= 0.811 - \left[\frac{1}{2}\left(\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) + \frac{1}{2}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right] \\
&= 0.811 - \left[\frac{1}{2}(0) + \frac{1}{2}(0.5)\right] = 0.811 - [0.25] = 0.561 \\
I(Y|x_3) = H(Y) - H(Y|x_3) &= 0.811 - \left[\frac{2}{4}H(Y|x_3 = \mathrm{T}) + \frac{2}{4}H(Y|x_3 = \mathrm{F})\right] \\
&= 0.811 - \left[\frac{1}{2}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \frac{1}{2}\left(\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right)\right] \\
&= 0.811 - \left[\frac{1}{2}(0) + \frac{1}{2}(0.5)\right] = 0.811 - [0.25] = 0.561 \\
I(Y|x_4) = H(Y) - H(Y|x_4) &= 0.811 - \left[\frac{2}{4}H(Y|x_4 = \mathrm{T}) + \frac{2}{4}H(Y|x_4 = \mathrm{F})\right] \\
&= 0.811 - \left[\frac{1}{2}\left(\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) + \frac{1}{2}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right)\right] \\
&= 0.811 - \left[\frac{1}{2}(0) + \frac{1}{2}(0.5)\right] = 0.811 - [0.25] = 0.561
\end{aligned}
$$

This seems to have the same properties as the previous subtree where all the splits yield the same information gain. Boolean logic would explain this tree as $\neg x_2 \wedge (\neg x_1 \wedge x_3 \wedge \neg x_4)$.

The completed decision tree is shown below:

2. The decision tree for the following dataset is shown below:

x_1	x_2	y
2	9	+
7	8	+
2	6	-
7	6	-
2	2	-
7	2	+

3. The class boundaries and labels plotted to a two-dimensional space of the dataset from problem 2 is shown below:

4. With the following dataset, show how lookahead logic would create a better decision tree than if splits were chosen in a greedy manner.

x_1	x_2	y
1	1	-
1	3	-
1	5	+
1	7	+
3	1	+
3	3	+
3	7	-
5	5	-

For splitting the tree, there are several options: $x_1 > 2, x_1 > 4, x_2 > 2, x_2 > 4$, and $x_2 > 6$. A greedy operation would take $x_1 > 4$ as the first split (the only split that gives non-zero information gain), then split on either $x_1 > 2$ or $x_2 > 4$ as the second. A sample greedy tree is shown above.

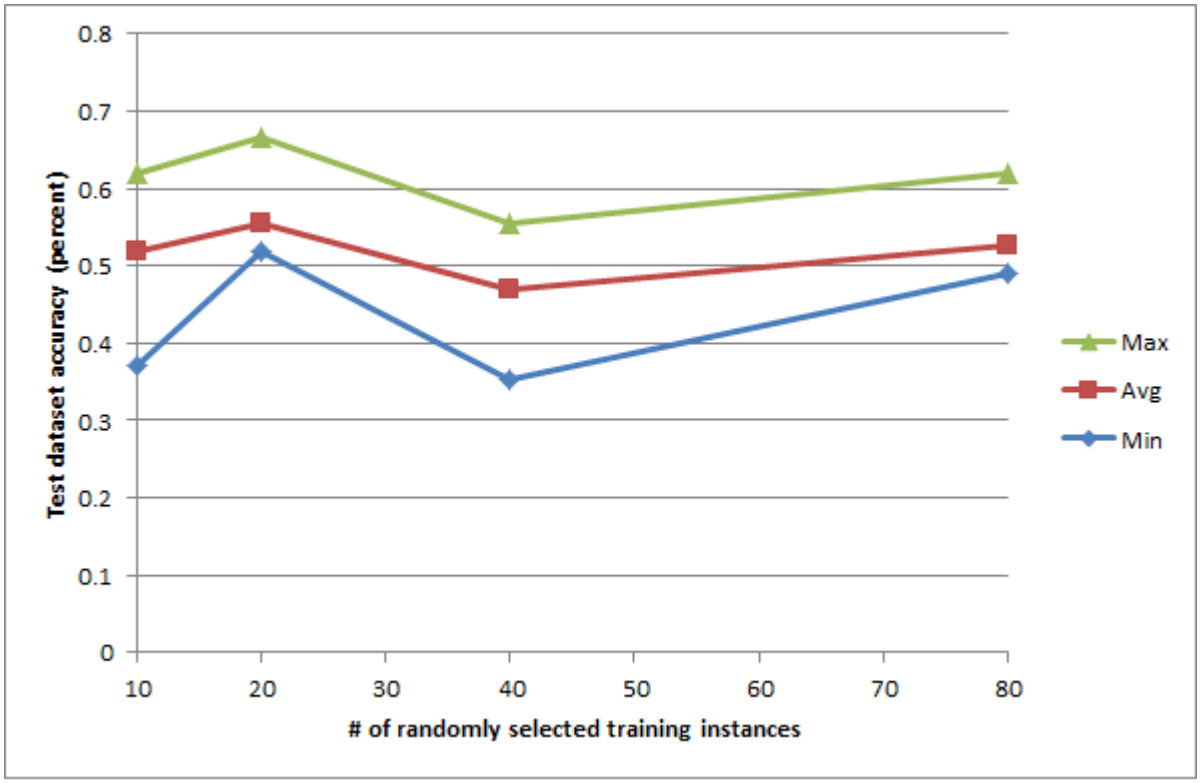
A lookahead algorithm is a bit more intelligent and will be able to find a tree where it can completely split in two levels. The lookahead algorithm will consider all splits, but only the winning solution is shown below ($x_1 > 2$, then true branch $x_2 < 4$ and false branch $x_2 > 4$).

$$
\begin{aligned}
I(Y|x_1 > 2) &= H(Y) - H(Y|x_1 > 2) \\
H(Y|x_1 > 2) &= H(Y|x_1 > 2, x_2 < 4, x_2 > 4) \\
&= \frac{1}{4}\left[H(Y|x_1 > 2, x_2 < 4) + H(Y|x_1 > 2, x_2 \not< 4) + H(Y|x_1 \not> 2, x_2 > 4) + H(Y|x_1 \not> 2, x_2 \not> 4)\right] \\
H(Y|x_1 > 2, x_2 < 4) &= -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - 0 = 0.5 \\
H(Y|x_1 > 2, x_2 \not< 4) &= 0 - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0.5 \\
H(Y|x_1 \not> 2, x_2 > 4) &= -\frac{2}{2}\log_2\left(\frac{2}{2}\right) - 0 = 0.5 \\
H(Y|x_1 \not> 2, x_2 \not> 4) &= 0 - \frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0.5
\end{aligned}
$$

$$
\begin{aligned}
H(Y|x_1 > 2) &= \frac{1}{4}(0.5 + 0.5 + 0.5 + 0.5) = 0.5 \\
I(Y|x_1 > 2) &= H(Y) - H(Y|x_1 > 2) = 1 - 0.5 = 0.5
\end{aligned}
$$

The greedy algorithm gets hooked on the $x_1 > 5$ split, but it turns out not to be such a wise decision in this case as a more direct, complete decision tree exists, as shown below.

5. I ran 1-kNN for 10, 20, 40, and 80 instances with 10 iterations each. From one trial, I got the following results:



6. I ran my model for 1-, 3-, 5-, and 7-kNN. The accuracy results of the test set is shown below in the graph. Overall, the results seemed to hover around 50%, although 1-kNN was surprisingly the most accurate. I guess sonar readings are pretty finicky.