Problem 12.1:

I want to create a decision tree to tell me whether I am likely to enjoy a particular book. Below is the data from 8 books I have read, whether I enjoyed them and the attributes:

1 Fiction/Non Fiction

2 Whether the book has > 500 pages (1) or not (0)

3 Whether the book is about travel (1) or not (0)

Book number	Fiction/NF	> 500 pages?	Travel?	Enjoyed?
1	N	0	0	0
2	F	1	0	0
3	N	0	0	0
4	F	1	1	0
5	N	0	1	1
6	F	0	0	1
7	N	0	1	1
8	F	0	1	1

Problem 12.1.1: Use the decision tree learning algorithm with the information gain heuristic to create a decision tree for these data.

We use the Information Gain (Entropy Loss) function as a heuristic along which to split. The formula for this is:

$$Gain(A) = B\left(\frac{p}{p+n}\right) - Remainder(A)$$

$$Remainder(A) = \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k + n_k}\right)$$

Where A is an attribute taking d values; p and n are the number of positive and negative examples respectively (and p_k and n_k are the number positive and negative examples of the kth value of the attribute). The function B is entropy and is defined as:

$$B(q) = -(q\, log_2(q) + (1-q)\, log_2(1-q)).$$

First, we calculate the information gain for each of the attributes "Fiction/NF", "> 500 pages?", and "Travel":

$$Gain(Fiction/NF) = B\left(\frac{4}{8}\right) - \left(\frac{4}{8}B\left(\frac{2}{4}\right) + \frac{4}{8}B\left(\frac{2}{4}\right)\right) = 0$$

$$Gain(> 500Pages?) = B\left(\frac{4}{8}\right) - \left(\frac{6}{8}B\left(\frac{4}{6}\right) + \frac{2}{8}B\left(\frac{0}{2}\right)\right) = 0.3113$$

$$Gain(Travel?) = B\left(\frac{4}{8}\right) - \left(\frac{4}{8}B\left(\frac{1}{4}\right) + \frac{4}{8}B\left(\frac{3}{4}\right)\right) = 0.1887$$

Book number	Fiction/NF	> 500 pages?	Travel?	Enjoyed?
1	N	0	0	0
2	F	1	0	0
3	N	0	0	0
4	F	1	1	0
5	N	0	1	1
6	F	0	0	1
7	N	0	1	1
8	F	0	1	1
Entropy Loss	**0**	**0.3113**	**0.1887**	

Clearly, "> 500 pages" is the attribute along which to split. Having split, we get two child data sets. For "> 500 pages" = 1, we have:

Book number	Fiction/NF	> 500 pages?	Travel?	Enjoyed?
2	F	1	0	0
4	F	1	1	0

These both have the same value, which is 0. Now we look at "> 500 pages" = 0. These are not all the same value, so we have to apply the B heuristic to the table below. The remaining attributes along which we can split are "Fiction/NF" and "Travel?". There information gain is computed as follows:

$$Gain(Fiction/NF) = B\left(\frac{4}{6}\right) - \left(\frac{4}{6}B\left(\frac{2}{4}\right) + \frac{2}{6}B\left(\frac{2}{2}\right)\right) = 0.2516$$

$$Gain(Travel?) = B\left(\frac{4}{6}\right) - \left(\frac{3}{6}B\left(\frac{1}{3}\right) + \frac{3}{6}B\left(\frac{3}{3}\right)\right) = 0.4591$$

Book number	Fiction/NF	> 500 pages?	Travel?	Enjoyed?
1	N	0	0	0
3	N	0	0	0
5	N	0	1	1
6	F	0	0	1
7	N	0	1	1
8	F	0	1	1
Entropy Loss	**0.2516**	–	**0.4591**	

We see that we need to split along "Travel?". At the end of this round, our decision tree looks like this:
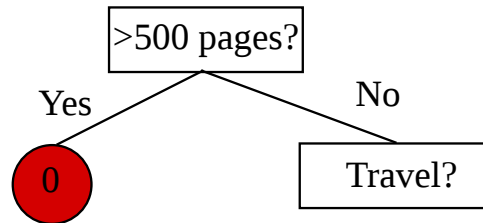
Figure 1: Intermediate Decision Tree

For "Travel?" = 1 we have:

Book number	Fiction/NF	> 500 pages?	Travel?	Enjoyed?
5	N	0	1	1
7	N	0	1	1
8	F	0	1	1

And again these have all the same value, which is 1. For "Travel?" = 0 we have:

Book number	Fiction/NF	> 500 pages?	Travel?	Enjoyed?
1	N	0	0	0
3	N	0	0	0
6	F	0	0	1
Entropy Loss	**0.9183**	−	−	

Just from looking at the data, we can see that we need to split along "Fiction/NF". But the information gain heuristic confirms our decision, as we gain 0.9183 bits of entropy by splitting:

$$Gain(Fiction/NF) = B\left(\frac{1}{3}\right) - \left(\frac{2}{3}B\left(\frac{0}{2}\right) + \frac{1}{3}B\left(\frac{1}{1}\right)\right) = 0.9183$$

Book number	Fiction/NF	> 500 pages?	Travel?	Enjoyed?
1	N	0	0	0
3	N	0	0	0

Book number	Fiction/NF	> 500 pages?	Travel?	Enjoyed?
6	F	0	0	1

As these values for Fiction and Non-Fiction are the all 0 and 1 respectively, we take their values. Our final decision tree looks like this:
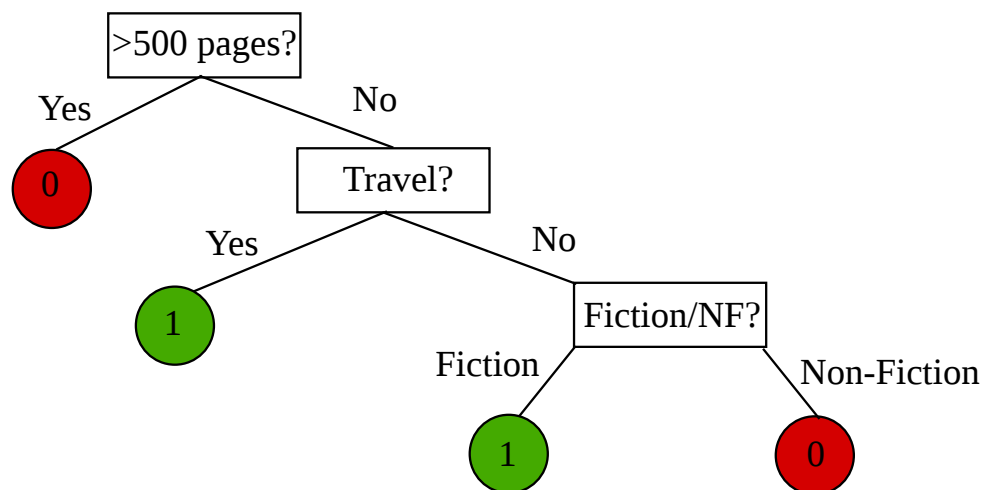

Figure 2: Final Decision Tree

Problem 12.1.2: Philip K. Dick's *Do Androids Dream of Electric Sheep* is fiction, 283 pages and not about travel. Am I likely to enjoy it?

Yes. And it is an excellent book and highly relevant to ethics in Artificial Intelligence