Show your working in all calculations.

2. There are two closely related cultivars of the cassava plant *Manihot esculenta* that grow in South America: one that is edible without any special processing, and another that presents a high content of hydrogen cyanide (HCN), which can be lethal to humans if not properly treated before consumption. Laboratory testing is expensive, and you are trying to investigate whether it is possible to classify plants based on their physical characteristics.

   In the samples available to you, some characteristics of the plants have been measured, and each specimen has been identified as poisonous or not. This data is shown in the table below. Based on this information and your knowledge of data mining, answer the following questions.

Sample	HasRedBranches	IsSmooth	HasPinkSkin	IsBitter	IsPoisonous
A	1	0	0	1	1
B	0	0	0	1	0
C	1	0	1	1	0
D	0	1	0	0	0
E	0	1	1	1	1
F	1	1	0	0	0
G	0	0	1	0	0
H	1	1	1	1	1

   a) What is the equation for calculating **entropy**? What is the entropy of "**IsPoisonous**"? (6 marks)

   b) What is the equation to calculate the **Information Gain** (IG)? Calculate the IG of each of the two attributes "**HasRedBranches**" and "**IsBitter**" for splitting the "**IsPoisonous**" class attribute. Which of these two attributes should you choose as the root of a decision tree? Why? (19 marks)

(Total: 25 Marks)