Show your working in all calculations.

1. This question is related to **Classification**.

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	RedBranches	Smooth	PinkSkin	Bitter	Poisonous
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) Estimate the following **conditional probabilities**:

- $P(RedBranches = 1 \mid Poisonous = 1)$
- $P(RedBranches = 1 \mid Poisonous = 0)$
- $P(Smooth = 1 \mid Poisonous = 1)$
- $P(Smooth = 1 \mid Poisonous = 0)$
- $P(PinkSkin = 0 \mid Poisonous = 1)$
- $P(PinkSkin = 0 \mid Poisonous = 0)$
- $P(Bitter = 1 \mid Poisonous = 1)$
- $P(Bitter = 1 \mid Poisonous = 0)$

b) Write down the **Bayes' theorem** and explain briefly what it means.

c) Use the conditional probabilities estimated in item (a) to predict the value of $(Poisonous)$ for a test sample $(RedBranches = 1; Smooth = 1; PinkSkin = 0; Bitter = 1)$ using the naïve Bayes approach.

d) In the context of Bayesian classification, what is the problem that could arise from having an estimated (prior) conditional probability of zero? What could be done to prevent this problem?