Show your working in all calculations.

1. a) Explain the concepts of **nominal** and **ratio** attributes. Give TWO examples of each type of attribute.

(6 marks)

Model Answer:

- *A nominal attribute has a finite (and usually small) set of possible values that serve only as labels or names. No ordering is implied among the values in a nominal variable. Examples include eye colour, types of drinks in a restaurant menu, zip codes, manufacturers of equipment, names of university staff.*
- *Ratio quantities are numeric variables for which the measurement scheme defines a natural zero, i.e., a value that represents the absence of any quantity. Ratio variables are treated as real numbers. Examples include distance between points, body weight, height, age, and annual income.*

(2.0 marks) *for each correct definition*
(0.5 marks) *for each correct example (for up to 2 examples per variable).*
An incorrect example cancels a correct one if more than 2 are provided for a given type of attribute.

b) Explain why it is important to evaluate the performance of classifiers on data that was not used for training the classification model. What would happen if the performance of a 1-nearest neighbour classifier was evaluated using the same data from the training set?

(6 marks)

Model Answer:

- *Training data error rate is inherently biased above the true error rate; thus testing is essential if we want to use the classifier on new data and need an estimate of its likely performance. The quality of that estimate depends on the amount of test data available.* (4 marks)
- *1-nearest neighbour always gets 100% on the test data, because the nearest neighbour of a point is the point itself, which always has the correct class. This however is not a realistic estimate of true error rate.* (2 marks)

c) Explain what is meant by a **confusion matrix**. Compute the values of the true positive rate and false positive rate for the following confusion matrix:

		Predicted	
		P	N
Actual	P	45	5
	N	10	20

where P stands for positive and N for negative cases. (5 marks)

Model Answer:

- *A confusion matrix tabulates the errors of a classifier on a given set where each row represents the true class frequencies and each column the predicted class.* (3 marks)
- *For the given matrix:*
  - *the true positive rate is* $TPR = TP/(TP + FN) = 45/(45 + 5) = 9/10 = 0.9$ (1 mark)
  - *the false positive rate is* $FPR = FP/(FP + TN) = 10/(10 + 20) = 1/3 = 0.33$ (1 mark)

d) Describe briefly how **Regression** and **Clustering** differ from **classification**. Give TWO example appli-

cations of regression and TWO of clustering. (8 marks)

Model Answer:

- *Regression: The target variable is numeric instead of discrete; this means that the error function cannot be misclassification rate, but instead must be a function such as the sum of squared errors (used e.g. for least squares estimation).* (2 marks) *Example applications include predicting financial markets, predicting missing numeric data, weather forecasting etc.* (2 marks)
- *Clustering: This is an unsupervised learning problem rather than a supervised one (like classification). In general one is interesting in detecting patterns in data, without having a labeled training set as a guide.* (2 marks) *Example applications include investigation of typical customer behaviours (e.g., clustering retailer data), and biological taxonomies.* (2 marks)

(Total: 25 Marks)

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)

Model Answer:

- *The equation for calculating entropy is* $Entropy(t) = -\sum_{j=1}^{n} p_j \log_2(p_j)$. (3 marks)
- *Entropy(IsPoisonous) =* $-3/8 \times \log_2(3/8) - 5/8 \times \log_2(5/8) = 0.9544$ (3 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)

Model Answer:

- *The quality of splitting at node* p *into* k *partitions (children) is given by the IG:* (3 marks)

$$IG(split) = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)\right)$$

- *If split by "**HasRedBranches**", we will have:*

	IsPoisonous	
HasRedBranches	*0*	*1*
0	*3*	*1*
1	*2*	*2*

(1 mark)

  - *For the branch "**HasRedBranches** = 0",* $Entropy = -(3/4)\log_2(3/4) - (1/4)\log_2(1/4) = 0.81$ (2 marks)
  - *For the branch "**HasRedBranches** = 1",* $Entropy = -(2/4)\log_2(2/4) - (2/4)\log_2(2/4) = 1$ (2 marks)
  - $IG(split) = Entropy(IsPoisonous) - (4/8 \times 0.8113 + 4/8 \times 1) = 0.9544 - 0.9512 = 0.0488$ (2 marks)

- *If split by "**IsBitter**", we will have:*

	IsPoisonous	
IsBitter	*0*	*1*
0	*3*	*0*
1	*2*	*3*

(1 mark)

  - *For the branch "**IsBitter** = 0", Entropy =* $-(0/3)\log_2(0/3) - (3/3)\log_2(3/3) = 0$ (2 marks)
  - *For the branch "**IsBitter** = 1", Entropy =* $-(2/5)\log_2(2/5) - (3/5)\log_2(3/5) = 0.9710$ (2 marks)
  - $IG(split) = Entropy(IsPoisonous) - (3/8 \times 0 + 5/8 \times 0.9710) = 0.9544 - 0.9057 = 0.3475$ (2 marks)

- *Since the information gain using "**IsBitter**" is larger than using "**HasRedBranches**", the attribute "**IsBitter**" should be chosen for the split.* (2 marks)

(Total: 25 Marks)

Show your working in all calculations.

3. This question is concerned with the **Naïve Bayes** classifier.

   a) Write down the **Bayes' theorem** and explain briefly what it means. (7 marks)

Model Answer:

- *Let E denote evidence and H denote hypothesis. Then Bayes' theorem can be stated as:* (3 marks)

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)}.$$

- *Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence. $P(H)$, the prior, is the initial degree of belief in H. $P(H|E)$, the posterior, is the degree of belief having accounted for E. The quotient $P(E|H)/P(E)$ represents the support E provides for H.* (4 marks)

   b) A doctor can run a test for the horrible disease *Examophobia*, which is known to affect one out of every 2000 students. The test for this disease has two possible outcomes: positive and negative. It is known that:

   - if a student has *Examophobia*, the test comes out positive 99% of the times.
   - if a student does not have *Examophobia*, the test comes out negative 99% of the times.

   If a student tests **positive** for the disease, what is the probability that he of she really has *Examophobia*? How can you interpret this result?

   (8 marks)

Model Answer:

We have the following information:

- *if Examophobia is present, the test comes out positive 99% of the time:* $p(P|E) = 0.99,\ \ p(N|E) = 0.01$
  (1 mark)
- *If Examophobia is not present, the test comes out negative 99% of the time:* $p(P|\neg E) = 0.01,\ \ p(N|\neg E) = 0.99$
  (1 mark)
- *Examophobia is known to occur in* $1/2000 = 0.05\%$ *of all students:* $p(E) = 0.0005,\ \ p(\neg E) = 0.9995$
  (1 mark)

We need to calculate the probability of a student really has Examophobia if he or she was tested positive for the disease, i.e., $p(E|P)$*:* (3 marks)

$$\begin{aligned} p(E|P) &= \frac{p(P|E) \times p(E)}{p(P)} \\ &= \frac{p(P|E) \times p(E)}{p(P|E) \times p(E) + p(P|\neg E) \times p(\neg E)} \\ &= \frac{0.99 \times 0.0005}{0.99 \times 0.0005 + 0.01 \times 0.9995} = 0.0472 \end{aligned}$$

Even though the exam tested positive, and even though the exam has low marginal error rates (1% *in both cases), the base rate of Examophobia in the student population is so low that the actual probability that a student testing positive actually has the disease is less than* 5%*.* (2 marks)

c) Consider the following dataset which represents the voting record of some U.S. members of Congress in 1984 on three key issues; the class is their party.

Handicapped Infants	Physician Fee Freeze	Budget	Party
yes	yes	no	Republican
no	yes	no	Republican
no	no	yes	Democrat
yes	no	yes	Democrat
no	no	yes	Democrat
no	yes	no	Democrat
no	yes	no	Republican
no	yes	no	Republican
yes	no	yes	Democrat
no	no	no	Republican
no	no	yes	Republican
no	no	yes	Democrat
yes	no	yes	Democrat
no	yes	no	Republican
no	yes	no	Republican
yes	no	yes	Democrat
yes	no	yes	Democrat

For a Congressman with the voting record **(yes, yes, no)**, compute the most probable value for **Party** using the standard Naïve Bayes model.

(10 marks)

(Total: 25 Marks)

Tip: To answer this question you may find it useful to copy the table below to your answer book, and fill in the counts and probabilities in the table. (this is **not** mandatory, just use it if you think it's convenient.)

	Handicapped Infants		Physician Fee Freeze		Budget	
	Yes	No	Yes	No	Yes	No
Party	Counts					
Republican						
Democrat						
Party	Probabilities					
Republican						
Democrat						

Model Answer:

- *Wee need to compare* $P(Party = Rep|HI = Y, PFF = Y, B = N)$ *against* $P(Party = Dem|HI = Y, PFF = Y, B = N)$*, where:*

$$P(Party = X|HI = Y, PFF = Y, B = N) = \frac{P(HI = Y, PFF = Y, B = N|Party = X) \times P(Party = X)}{P(HI = Y, PFF = Y, B = N)}$$

Since the denominator $P(HI = Y, PFF = Y, B = N)$ *is the same for both* rep *and* dem*, it can be ignored in the comparisons.* (3 marks)

- *For* rep*:* (3 marks)

$$\begin{aligned}&P(HI = Y, PFF = Y, B = N|Party = Rep) \times P(Party = Rep)\\ &= P(HI = Y|Party = Rep) \times P(PFF = Y|Party = Rep) \times P(B = N|Party = Rep) \times P(Party = Rep)\\ &= 1/8 \times 3/4 \times 7/8 \times 8/17 \approx 0.0386\end{aligned}$$

- *For* dem*:* (3 marks)

$$\begin{aligned}&P(HI = Y, PFF = Y, B = N|Party = Dem) \times P(Party = Dem)\\ &= P(HI = Y|Party = Dem) \times P(PFF = Y|Party = Dem) \times P(B = N|Party = Dem) \times P(Party = Dem)\\ &= 5/9 \times 1/9 \times 1/9 \times 9/17 \approx 0.00363\end{aligned}$$

- *Since the result for* rep *is larger than that for* dem*, the most probable value for the PArty class in this new observation is* $Republican$*.* (1 mark)