

Total marks available is 100.

I have read and clearly understand the Examination Rules of Beijing University of Technology and University College Dublin and am aware of the Punishment for Violating the Rules of Beijing University of Technology and University College Dublin. I hereby promise to abide by the relevant rules and regulations by not giving or receiving any help during the exam. If caught violating the rules, I would accept the punishment thereof.

Pledger: _____ **Class No:** _____

BJUT Student ID: _____ **UCD Student ID** _____

[illegible]

The exam paper has 4 questions on 5 pages, with a full score of 100 points. Answer Question 1 (worth maximum of 40 points), and any other two questions (worth maximum of 30 points each). You are required to use the given Examination Book only.

Item	Part 1	Part 2	Part 3	Total Score
Full Score	40	30	30	
Obtained Score				

Question 1 (40 marks):

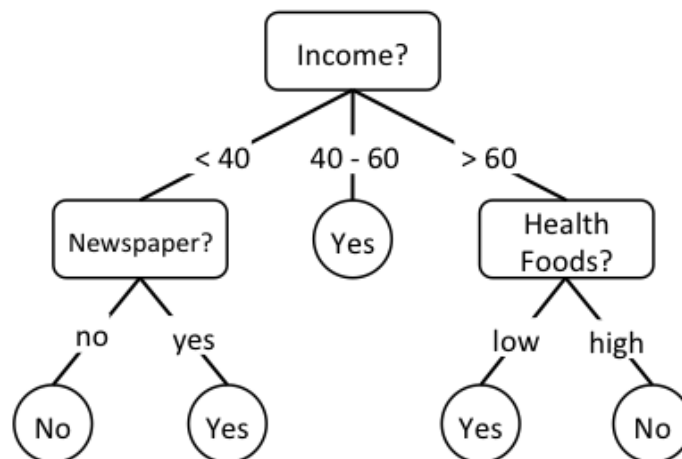
- a) Explain briefly what is meant by **Unsupervised Learning** and **Supervised Learning**. Give an example of each. (5 marks)
- b) Explain briefly **Feature Selection**. Why would you use feature selection? (5 marks)
- c) Briefly explain **Underfitting** and **Overfitting**. Draw a simple plot demonstrating each. (5 marks)
- d) What does it mean when a prediction model is said to **Generalise** well? (5 marks)
- e) List the three types of **data quality issues** and briefly explain each. (5 marks)
- f) What is the **Inductive Bias** of a machine learning algorithm. Give examples. (5 marks)
- g) What is the difference between a **Continuous feature** and a **Categorical feature**? Give examples. (5 marks)
- h) What is the difference between a **Raw feature** and a **Derived feature**? Give examples. (5 marks)

Question 2 (30 marks):

The following table presents a dataset collected by a retail company capturing historical details of which of their customers have responded to promotions the company has run. The information captured covers customer income bracket, customer age, whether or not the customer regularly buys a newspaper, the proportion of health foods typically included in the customer's shopping, and, finally, whether or not they responded to previous promotional mailings.

ID	Income	Age	Newspaper	Health Foods	Respond
C-01	<40	81	no	low	No
C-02	<40	76	no	high	No
C-03	40-60	86	no	low	Yes
C-04	>60	84	no	low	Yes
C-05	>60	45	yes	low	Yes
C-06	>60	66	yes	high	No
C-07	40-60	41	yes	high	Yes
C-08	<40	68	no	low	No
C-09	<40	32	yes	high	Yes
C-10	>60	56	yes	low	Yes
C-11	<40	58	yes	high	Yes
C-12	40-60	52	no	high	Yes
C-13	40-60	90	yes	low	Yes
C-14	>60	69	no	high	No

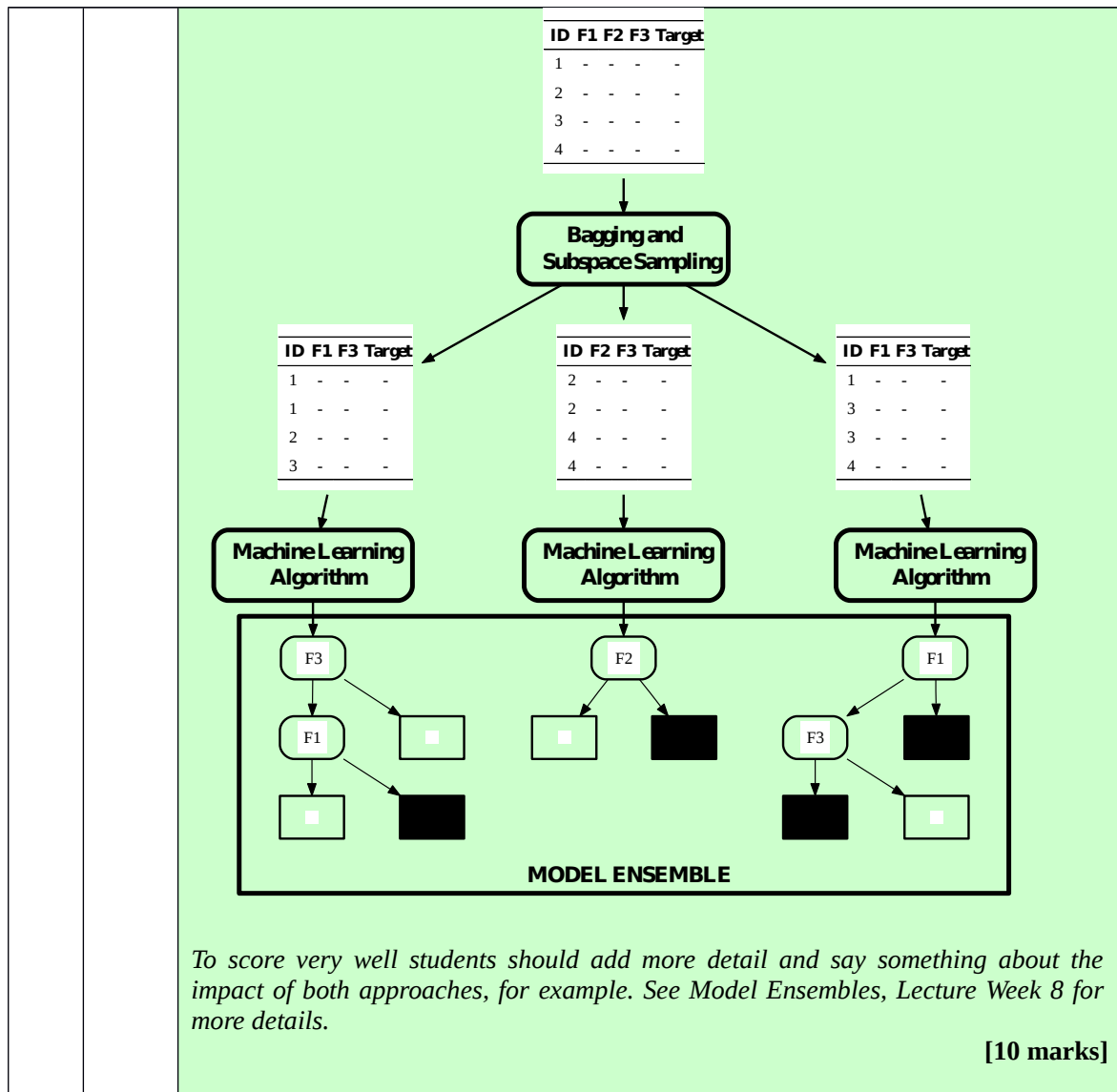
This dataset has been used to induce a **decision tree** that can predict whether or not new customers will respond to promotional mailings. This decision tree is shown below.



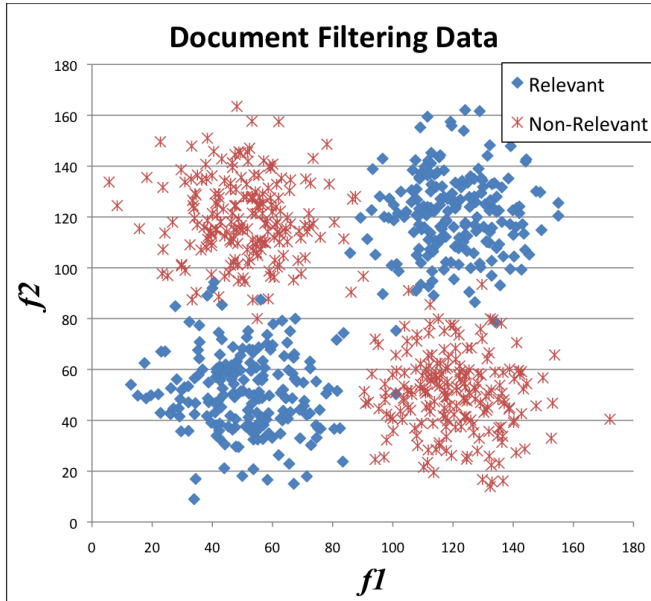
- (i) What is **information gain**? Describe the three step process for calculating information gain using the following equations:

		$H(t, \mathcal{D}) = - \sum_{l \in \text{levels}(t)} (P(t = l) \times \log_2(P(t = l)))$ $\text{rem}(d, \mathcal{D}) = \sum_{l \in \text{levels}(d)} \underbrace{\frac{ \mathcal{D}_{d=l} }{ \mathcal{D} }}_{\text{weighting}} \times \underbrace{H(t, \mathcal{D}_{d=l})}_{\text{entropy of partition } \mathcal{D}_{d=l}}$ $IG(d, \mathcal{D}) = H(t, \mathcal{D}) - \text{rem}(d, \mathcal{D})$
		(10 marks)
		<p>The information gain of a descriptive feature can be understood as a measure of the reduction in the overall entropy of a prediction task by testing on that feature.</p> <p>Computing information gain is a three-step process:</p> <p>1 Compute the entropy of the original dataset with respect to the target feature. This gives us an measure of how much information is required in order to organize the dataset into pure sets.</p> <p>2 For each descriptive feature, create the sets that result by partitioning the instances in the dataset using their feature values, and then sum the entropy scores of each of these sets. This gives a measure of the information that remains required to organize the instances into pure sets after we have split them using the descriptive feature.</p> <p>3 Subtract the remaining entropy value (computed in step 2) from the original entropy value (computed in step 1) to give the information gain.</p> <p>We need to define three equations to formally specify information gain (one for each step).</p> <p>See “Information-based Learning – Decision Trees”, Lectures Weeks 6-8 for full details.</p>
	(ii)	<p>The information gain of the feature <i>Income</i> at the root node of the tree is 0.247. A colleague has suggested that <i>Newspaper</i> would be the best feature to query at the root node of the tree. Demonstrate whether or not this is the case. Show all workings.</p>
		(12 marks)
		<p>In order to answer this question the student needs to calculate the information gain of the <i>Newspaper</i> attribute at the root node and compare it to the information gain for <i>Income</i>.</p> $H = -(5/14 * \log_2(5/14) + 9/14 * \log_2(9/14))$ $= -(0.357 * (-1.49) + 0.643 * (-0.64))$

		$= -(-0.532 + -0.411)$ $= 0.9434$ <p style="text-align: right;">[2 marks]</p> $H(\text{Newspaper}) = 7/14*(-4/7*\log_2(4/7) + -3/7*\log_2(3/7))$ $+ 7/14*(-1/7*\log_2(1/7) + -6/7*\log_2(6/7))$ $= 0.5*(-0.571*(-0.81) + -0.429*(-1.22))$ $+ 0.5*(-0.143*(-2.81) + -0.857*(-0.22))$ $= 0.5*(0.4625 + 0.5233) + 0.5*(0.4014 + 0.1885)$ $= 0.5*0.9858 + 0.5*0.5899$ $= 0.4929 + 0.2949$ $= 0.7879$ <p style="text-align: right;">[3 marks]</p> <p>The information gain for Newspaper is 0.15</p> <p style="text-align: right;">[2 marks]</p> <p>This is not greater than the information gain for Income and so Newspaper would not be a better variable for the root node.</p> <p style="text-align: right;">[3 marks]</p> <p>See “A Worked Example: Predicting Vegetation Distributions”, Lecture Week 7 for more details.</p>
	(iii)	<p>Decision trees are often used as the basis for ensemble models. The key to training effective ensemble models is to introduce diversity into the models in the ensemble. Compare the ways that <i>diversity</i> is introduced into ensembles in the bagging and random forest ensemble methods.</p> <p style="text-align: right;">(10 marks)</p>
		<p>Boosting works by iteratively creating models and adding them to the ensemble. The iteration stops when a predefined number of models have been added. Each new model added to the ensemble is biased to pay more attention to instances that were miss-classified by previous models.</p> <p>Students should explain that in bagging diversity is introduced by performing a random sub-sampling on the rows in the training data so that each model in the training data will be trained using a slightly different sub-sample of the data.</p> <p>Random forest go further in that not only are the instances sub-sampled but the descriptive features in a dataset are also sub-sampled.</p> <p>A diagram such as the following would be useful.</p>

**Question 3 (30 marks):**

(a)	<p>A confused client has come to you with three different customer marketing response prediction applications, each of which uses a particular classification algorithm to perform the response prediction (except for the classification algorithm used, all other aspects of the applications are identical).</p> <p>Describe the evaluation criteria and experiments you would recommend so as these applications could be ranked from best to worst.</p>
	(15 marks)
	<p><i>This is a discursive question so giving a precise marking scheme is not appropriate. See Lec-11_week-12.pdf “The Art of Machine Learning” and “Choosing a Machine Learning Approach”. Also see Lec-5_week-5.pdf Evaluation and Performance Measures.</i></p> <p><i>However, key points that the student should touch on include:</i></p> <ul style="list-style-type: none"> <i>.Predictive accuracy</i> <i>.Evaluation/ Performance measure</i> <i>.Evaluation measure comparisons</i> <i>.Experiments/training and test sets/k-Fold Cross Validation</i>

	<p>.Prediction speed .Capacity for retraining .Interpretability (understanding and insight provided by the model)</p> <p><i>It should be noted also, that these evaluation criteria are application dependent.</i></p> <p style="text-align: right;">[15 marks]</p>
(b)	<p>The image below shows a scatter plot of a dataset from a simple document filtering problem. There are just two continuous features in this dataset, $f1$ and $f2$, and two classes, <i>Relevant</i> and <i>Non-Relevant</i>. In the scatter plot $f1$ is shown on the horizontal axis, $f2$ is shown on the vertical axis and the shapes of the points represent class.</p> <div style="text-align: center;">  </div> <p>Discuss the difficulties associated with building classification models from datasets with characteristics similar to those shown in the scatter plot. In your answer comment on the suitability of specific classification approaches.</p> <p style="text-align: right;">(15 marks)</p> <p><i>The key insight students need to make here is that this dataset is not linearly separable. This means that many classification approaches will not be able to build good models using this data. This is one of the core distinctions that can be made between classification algorithms.</i></p> <p style="text-align: right;">[7 marks]</p> <p>Students should then go on to discuss how specific algorithms cope with non-linearly separable data. Interesting examples would include:</p> <ul style="list-style-type: none"> . Simple linear logistic regression models could not handle this problem. . . k-NN are good at dealing with data in this form, and this is one of their major advantages, as they do not attempt to create a global model. <p style="text-align: right;">[8 marks]</p> <p>[It is possible that students could come up with a completely different answer and these will be marked appropriately. See Lec-11_week-12.pdf “The Art of Machine Learning” and “Choosing a Machine Learning Approach” for other suggestions, and more advantages/disadvantages of different algorithms.]</p>