Student Name | Nome do Aluno:	Number | Número:

GROUP I *(15 VALUES)*

1. Assume the *dataset* with attributes X (nominal), Y (numeric) and a class C (boolean).
   Admita o *dataset* com atributos X (nominal), Y (numérico) e uma classe C (booleana).

X	Y	C
A	10.5	0
	8.3	0
A	11.4	0
A	11.6	1
B	12.1	1
B	20.2	1

(a) Apply 1-R method to attribute X and generate its rule and the corresponding error.
Aplique o método 1-R ao atributo X e gere a sua regra e o respectivo erro.

Attribute **X: rule** and **error** (show all calculations):

(b) Use 1-R method to discretize attribute Y and show the corresponding breakpoints.
Use o método 1-R para discretizar o atributo Y e apresente os respectivos pontos-de-corte (*breakpoints*).

Attribute **Y: breakpoints** for discretization (show all calculations):

(c) Apply 1-R method to attribute Y and generate its rule and the corresponding error.
Aplique o método 1-R ao atributo Y e gere a sua regra e o respectivo erro.

Attribute **Y: rule** and **error** (show all calculations):

(d) According to 1-R method what is rule and the corresponding error that best predicts the class?
De acordo com o método 1-R qual é a regra e correspondente erro que melhor faz a previsão da classe?

1-R: rule and **error** (justify):

Student Name | Nome do Aluno:	Number | Número:

2. For the ID3 decision tree to deal with a numeric attribute it must previously discretized. But, now consider the suggestion (from Mr. T) that "we can deal with it by adding a branch for each value of the attribute".

   Para a árvore de decisão ID3 lidar com atributo numérico ele precisa de ser discretizado. Mas, considere agora a sugestão (do Sr. T) diz "ser possível lidar com isso adicionando um ramo por cada valor do atributo".

(a) Choose, from list below, the most accurate observation about the technique suggested by Mr. T, and justify.

Escolha, da lista abaixo, a observação mais correta sobre a técnica sugerida pelo Sr. T e justifique.

#	Observation
A	Would likely result in a decision tree where each complete-branch (sequence of nodes from the root to a leaf) would be computationally too heavy to cross. Provavelmente resultaria numa árvore de decisão onde cada ramo-completo (sequência de nós da raiz a uma folha) seria computacionalmente demasiado pesado de atravessar.
B	Would likely result in a decision tree with poor results in the training dataset and also with poor results in the testing dataset. Provavelmente resultaria numa árvore de decisão com maus resultados no conjunto de treino e também com maus resultados no conjunto de teste.
C	Would likely result in a decision tree with good results in the training set but poor results in the test set. Provavelmente resultaria numa árvore de decisão com bons resultados no conjunto de treino mas maus resultados no conjunto de teste.
D	Would likely result in a decision tree with poor results in the training set and good results in the test set. Provavelmente resultaria numa árvore de decisão com maus resultados no conjunto de treino e bons resultados no conjunto de teste.

The **most accurate** observation is: | **A** | **B** | **C** | **D** | (circle the correct option).

Justification:

(b) Now, forget Mr. T! Assume a numeric attribute with minimum observed value "-10"; after applying "**equal-width binning**" you get 60 bins each of width 5. What is the maximum observed value for this attribute?

Esqueça Sr. T! Assuma atributo numérico com valor mínimo observado de "-10"; depois de aplicar "***equal-width binning***" obtém 60 *bins* cada um com amplitude 5. Qual o valor máximo observado deste atributo?

Justification (show all calculations):

(c) We have 1000 observed values of an attribute. We apply "**equal-frequency binning**" with 20 bins. Without looking at the dataset is it possible to know how many values are there in each bin? If yes, how many?

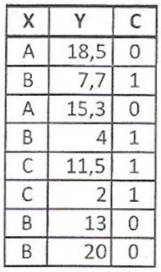
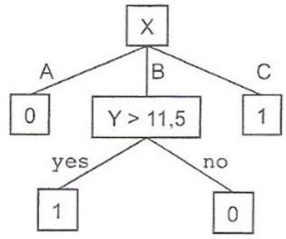
Temos 1000 valores observados de um atributo. Aplicamos "***equal-frequency binning***" com 20 *bins*. Sem olhar para o *dataset* é possível saber quantos valores estarão em cada *bin*? Se sim, quantos?

Justification (show all calculations):

Student Name | Nome do Aluno:	Number | Número:

3. Consider the following dataset and decision-tree.

   Considere o *dataset* e árvore-de-decisão (*decision-tree*) que se seguem.

dataset	*decision-tree*
	

X	Y	C
A	18,5	0
B	7,7	1
A	15,3	0
B	4	1
C	11,5	1
C	2	1
B	13	0
B	20	0

(a) Consider the entire datatset and calculate the entropy of attribute X.

Considere todo o *dataset* e calcule a entropia do atributo X.

For the entire dataset, the entropy of attribute X (show all calculations):

(b) Consider the entire datatset and calculate the entropy of the "virtual binary attribute" defined by "Y > 11.5".

Considere todo o *dataset* e calcule a entropia do "atributo virtual binário" definido por "Y > 11.5".

For the entire dataset, the entropy of "virtual binary attribute" "Y > 11.5" (show all calculations):

(c) Restrict *dataset* to instances that satisfy "X = B" and (for those instances) calculate entropy of "Y > 11.5".

Restrinja *dataset* às instâncias que satisfazem "X = B" e (para essas instâncias) calcule entropia "Y > 11.5".

For dataset where "X = B", the entropy of "virtual binary attribute" "Y > 11.5" (show all calculations):

(d) The decision-tree was built with an entropy-based method. Is it plausible that dataset was the training one?

A árvore-decisão é construída com base na entropia. É plausível que o *dataset* tenha sido o de treino?

It is plausible that **the dataset was used for training the decision-tree: YES | NO** (circle the correct option):
Justification:

(e) Now we use the dataset for testing (evaluating) the decision-tree. What is the associated error-rate?

Agora usamos o *dataset* para testar (avaliar) a árvore-decisão. Qual é a taxa-de-erro associada?

The error-rate of decision-tree using dataset for testing (show all calculations) is: _____

Student Name | Nome do Aluno:	Number | Número:

GROUP II *(5 VALUES)*

1. Consider the attributes X and Y Boolean, an attribute Z with three possible values (0, 1, 2), a binary class C and a training dataset with 10 instances, 5 positive (C=1) and 5 negatives (C=0). The following table shows the result of counting over the (10) instances of the training dataset. Consider the naïve Bayes method.

   Considere os atributos X e Y Booleanos, o atributo Z com três valores (0, 1, 2), a classe binária C e um *dataset* de treino com 10 instâncias, 5 positivas (C=1) e 5 negativas (C=0). A tabela seguinte representa o resultado de uma contagem sobre as (10) instâncias do *dataset* de treino. Considere o método naïve Bayes.

	C = 0	C = 1
X = 1	0	0
Y = 0	3	2
Z = 1	1	4
Z = 2	2	0

(a) The value of **P(X=0 | C=0) x P(C=0)** is: _____

Justification (show all the calculations):

(b) The **likelihood** of **C=0** given the observation **(X=0, Y=0, Z=0)** is: _____

Justification (show all the calculations):

(c) Given this dataset does it make sense to use the Laplace estimator? Why?

Dado este *dataset* faz sentido usar-se o estimador de Laplace? Porquê?

Justification:

2. Mark, with X, for true (T) or false (F). Assinale, com X, verdadeiro (T) ou falso (F).

T	F	Statement | Afirmação
		The market-basket analysis assumes that dataset does not contain a class attribute. O *maket-basket analysis* assume que o *dataset* não contém um atributo classe.
		The APRIORI algorithm does not need the user to provide the minimum support value. O algoritmo APRIORI não precisa que o utilizador forneça o valor de suporte mínimo.
		If {A,B}, {A,C}, {B,D} are frequent 2-itemset, then {A,B,D} is a 3-itemset candidate. Se {A,B}, {A,C}, {B,D} são 2-itemset frequentes, então {A,B,D} é 3-itemset candidato.
		The PRIORI assumes a total order relation between the LHS and RHS of each rule. O APRIORI assume uma relação de ordem total entre o LHS e o RHS de cada regra.
		The confidence anti-monotone property is used for the frequent itemset generation. A propriedade anti-monótona da confiança é usada para gerar os itemset frequentes.