UNIVERSITY OF LONDON
GOLDSMITHS COLLEGE
Department of Computing
B. Sc. Examination 2015

IS53023B
Data Mining

Duration: 2 hours 15 minutes

Date and time:

There are five questions in this paper. You should answer no more than THREE questions. Full marks will be awarded for complete answers to a total of THREE questions. Each question carries 25 marks. The marks for each part of a question are indicated at the end of the part in [.] brackets.

There are 75 marks available on this paper.

Electronic calculators must not be programmed prior to the examination. Calculators which display graphics, text or algebraic equations are not allowed.

THIS PAPER MUST NOT BE REMOVED
FROM THE EXAMINATION ROOM

Question 1

a. Name 3 decision tree algorithms. [3]

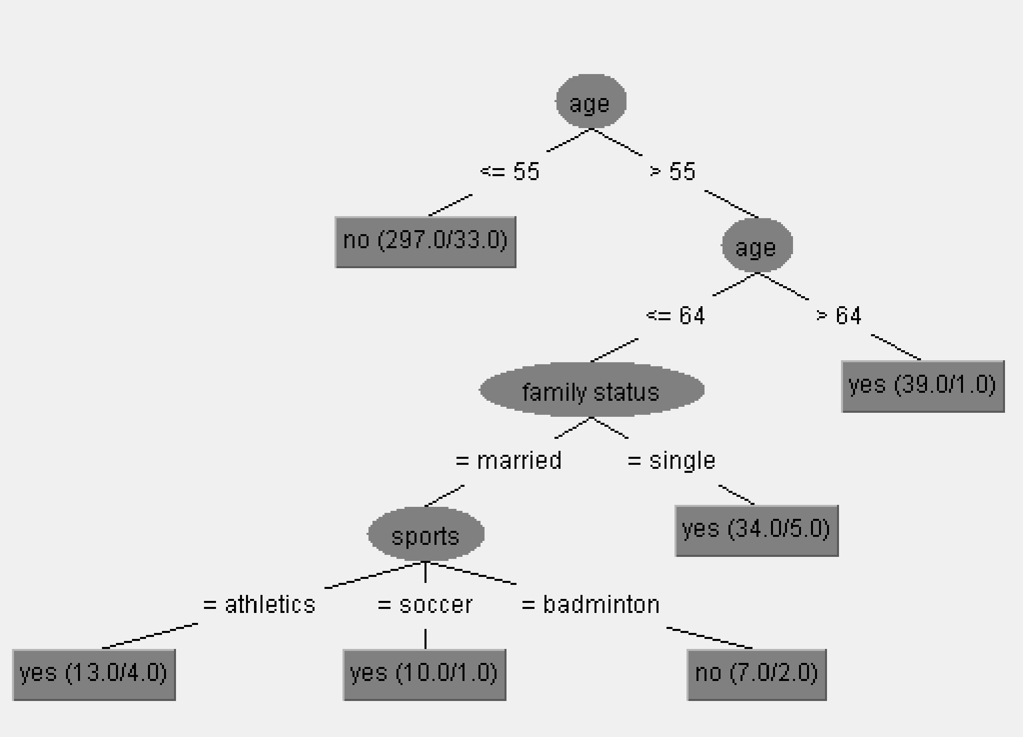
b. The decision tree below represents a supervised learning model for a direct mailing campaign, with the output attribute **respond** whose values are **yes** and **no**. Using the information provided in the decision tree you are required to:

i. Compute the size of the classes **yes** and **no** with respect to the dataset originally used to train the decision tree. [4]

ii. Compute the accuracy of this decision tree, with respect to the same dataset used for training it. [2]

iii. Generate four production rules only from the decision tree, and compute their accuracy and coverage. [12]

c. Prune the decision tree below such that the new decision tree has 3 levels only. You are required to completely draw the new decision tree. [4]



Question 2

a. Briefly explain, in one sentence, one assumption that the Naive Bayes algorithms makes regarding the data. [3]

b. Apply the Naive Bayes algorithm on the dataset below in order to predict the value for LifeInsurance attribute for an instance defined by HomeOwner=No, Retired=Yes, Car=No, and CreditCardInsurance=No. You are required to illustrate how the algorithm works in this case.   [22]

HomeOwner	Retired	Car	CreditCardInsurance	LifeInsurance
No	No	No	No	Yes
No	Yes	Yes	Yes	No
Yes	No	No	No	Yes
No	Yes	Yes	Yes	Yes
No	No	Yes	No	No
Yes	No	No	No	No
No	Yes	Yes	Yes	Yes
Yes	No	No	No	Yes
No	No	No	No	Yes
No	Yes	Yes	No	No

Question 3

In a customer attrition / churning application which involves a dataset with details about customers, the output attribute called **churn** has two values, **yes** and **no**. One builds a predictive model that, before being used in practice, is evaluated on a test dataset. Below one provides two columns from the scored dataset, namely the column **churn**, and the column **predicted_churn,** which provides the predictions by applying the model on this test dataset.

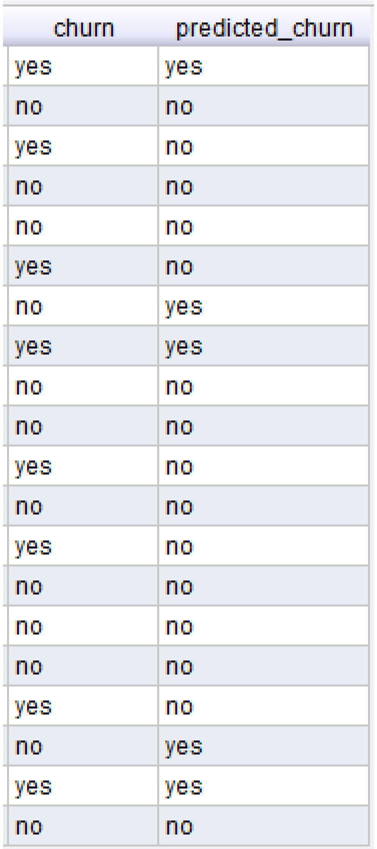
You are required to:

a) Provide the confusion matrix. [4]

b) Calculate the accuracy, precision, sensitivity, specificity, and lift with respect to the class **yes** by showing your work. [10]

c) Briefly explain what the accuracy, precision, sensitivity and lift represent in the context of this mentioned application of predicting churners? Do not use more than one statement for each performance measure. [8]

d) What is the most important performance measure among the ones mentioned in (c) above, as a criterion for the selection of a predictive model in the customer churn application? Justify your answer. [3]



churn	predicted_churn
yes	yes
no	no
yes	no
no	no
no	no
yes	no
no	yes
yes	yes
no	no
no	no
yes	no
no	no
yes	no
no	no
no	no
no	no
yes	no
no	yes
yes	yes
no	no

Question 4

a. Does k-Means work with missing values? Justify your answer. [3]

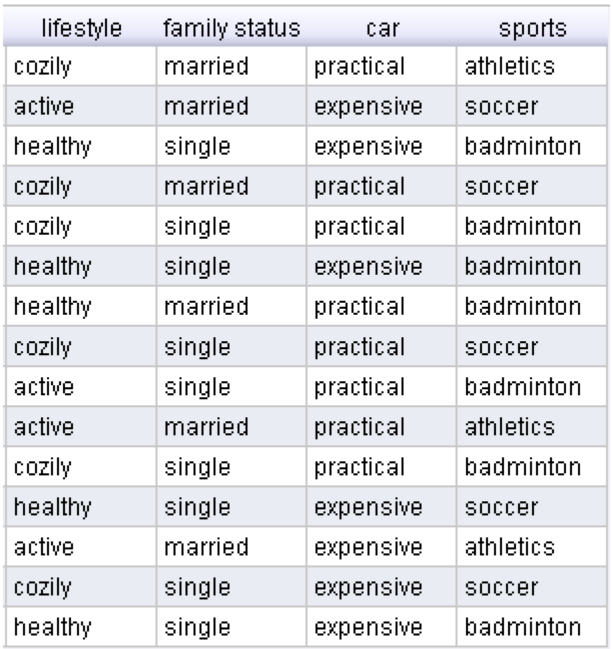
b. Name two other clustering algorithms, different from k-Means. [2]

c. Illustrate the application of the k-Means algorithm for k=2, using the following dataset. The cluster centres to start with are the rows (40, 50) and (60, 50). Only the first two iterations are required to be illustrated. Would more iterations be needed for the completion of the algorithm? Justify your answer. [20]

Col1	**Col2**
40	50
40	110
60	50
60	90
80	70
120	140

Question 5

The following dataset having the attributes **lifestyle**, **family status, car** and **sports** is to be used for the extraction of association rules. You are required to apply the Apriori algorithm with the minimum support of 0.2 and the minimum confidence of 0.6. In particular display all the frequent itemsets. Display all the association rules you get and decide which ones are strong rules. Finally you are required to compute also the support for the strong rules only. [25]



lifestyle	family status	car	sports
cozily	married	practical	athletics
active	married	expensive	soccer
healthy	single	expensive	badminton
cozily	married	practical	soccer
cozily	single	practical	badminton
healthy	single	expensive	badminton
healthy	married	practical	badminton
cozily	single	practical	soccer
active	single	practical	badminton
active	married	practical	athletics
cozily	single	practical	badminton
healthy	single	expensive	soccer
active	married	expensive	athletics
cozily	single	expensive	soccer
healthy	single	expensive	badminton