**UNIVERSITY OF LONDON**
**GOLDSMITHS COLLEGE**

**Department of Computing**

**B. Sc. Examination 2013**

# 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.*

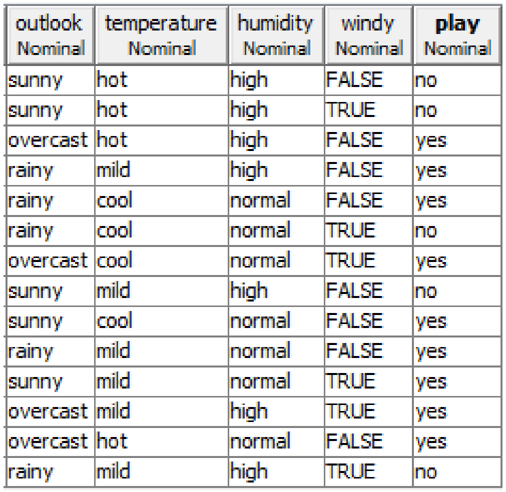
*There are 75 marks available on this paper.*

*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) Demonstrate the application of the decision tree algorithm based on goodness scores, on the dataset given below. You must not build the whole tree, but you are required to generate the root of the decision tree by illustrating in detail the work done by the algorithm. All the attributes of the dataset are input attributes, except *play* which is the output attribute. [15]



| outlook | temperature | humidity | windy | **play** |
|---|---|---|---|---|
| Nominal | Nominal | Nominal | Nominal | Nominal |
| sunny | hot | high | FALSE | no |
| sunny | hot | high | TRUE | no |
| overcast | hot | high | FALSE | yes |
| rainy | mild | high | FALSE | yes |
| rainy | cool | normal | FALSE | yes |
| rainy | cool | normal | TRUE | no |
| overcast | cool | normal | TRUE | yes |
| sunny | mild | high | FALSE | no |
| sunny | cool | normal | FALSE | yes |
| rainy | mild | normal | FALSE | yes |
| sunny | mild | normal | TRUE | yes |
| overcast | mild | high | TRUE | yes |
| overcast | hot | normal | FALSE | yes |
| rainy | mild | high | TRUE | no |

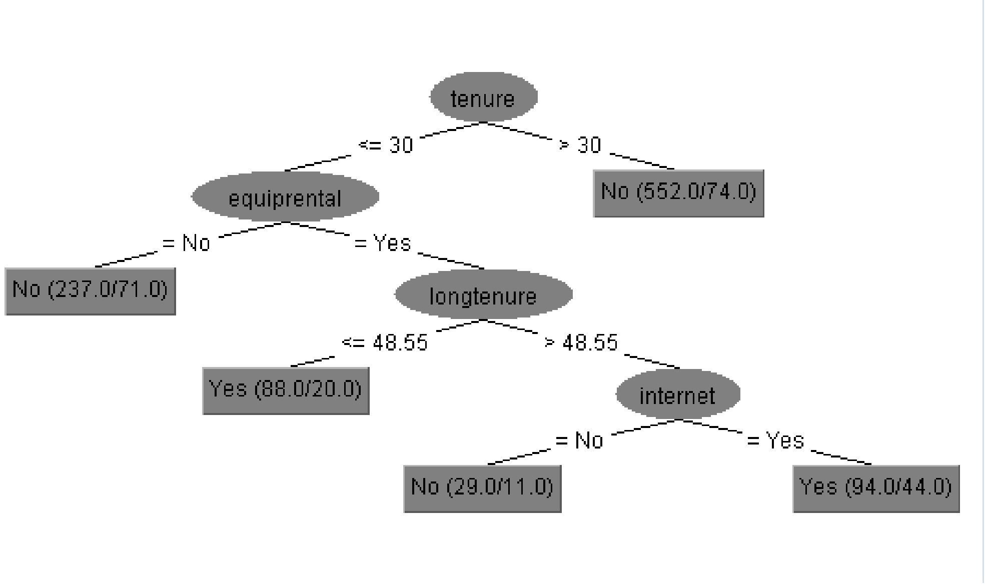
b) In the context of data warehouses, name and briefly describe the OLAP operations. Do not use more than two statements per operation. [10]

**Question 2**

a) Consider the decision tree below built for a classification problem using a dataset whose output attribute, called *churn*, has the possible values *Yes* and *No*. A pair of numbers (x/y) in a terminal node T means that x instances satisfy all the conditions on the path from the root to the node T, and that out of these x instances: (x-y) instances satisfy the decision in the node T, and y instances are exceptions to this decision. For instance the pair (237.0/71.0) in the leftmost terminal node (having the decision *No*) means that 237 instances satisfy the conditions *tenure<=30* and *equiprental=No* and that (237-71) instances satisfy the decision in this terminal node (that is, *churn* is *No* for these instances), and 71 instances do not satisfy this decision. You are required to:

i. Calculate how many instances each of the classes *Yes* and *No* contains (these classes are defined by the respective values of the output attribute *churn*). Calculate also how many instances the dataset contains. [5]

ii. Write all the production rules that can be extracted from the decision tree, and calculate the accuracy and the coverage of each production rule when providing it. [15]



b) Draw a fully connected feed-forward neural network with one hidden layer, which can be learnt in a classification problem with three input numeric attributes. Indicate the input, hidden and output layers in the drawing. [5]

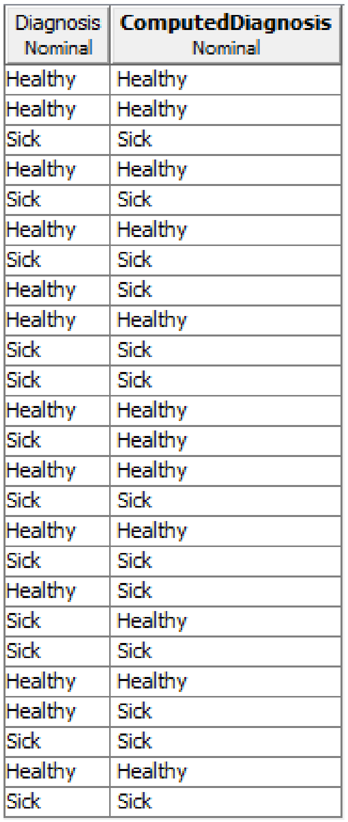
**Question 3**

Assume a data mining session was performed for a classification problem, on a given training dataset containing details and results of medical tests of patients that were diagnosed by a cardiologist regarding heart conditions. The output attribute, called *Diagnosis*, has two values: *Healthy* and *Sick*. A model was built using a data mining algorithm on the training dataset. The model was then applied on a test dataset in order to be evaluated. Following the application of the model on the training dataset, you are provided the table below showing in the first column the actual diagnoses (see the *Diagnosis* attribute), and in the second column the computed diagnoses which are generated by the application of the model (see the *ComputedDiagnosis* generated attribute). Note that each row corresponds to a diagnosed patient. You are required to:

a) Build the confusion matrix. [4]

b) i. Calculate the accuracy, error rate, precision, sensitivity, specificity, and lift with respect to the class *Sick* by showing your work and the formulae you use. [12]
ii. Choose and name one measure from (i) above that you consider to be the most important in evaluating the performance of the model built for this classification problem, and briefly explain your choice. [3]

c) List the names of six data mining algorithms that are appropriate to be applied in the classification problem mentioned above. [6]



| Diagnosis | **ComputedDiagnosis** |
|---|---|
| Nominal | Nominal |
| Healthy | Healthy |
| Healthy | Healthy |
| Sick | Sick |
| Healthy | Healthy |
| Sick | Sick |
| Healthy | Healthy |
| Sick | Sick |
| Healthy | Sick |
| Healthy | Healthy |
| Sick | Sick |
| Sick | Sick |
| Healthy | Healthy |
| Sick | Healthy |
| Healthy | Healthy |
| Sick | Sick |
| Healthy | Healthy |
| Sick | Sick |
| Healthy | Sick |
| Sick | Healthy |
| Sick | Sick |
| Healthy | Healthy |
| Healthy | Sick |
| Sick | Sick |
| Healthy | Healthy |
| Sick | Sick |

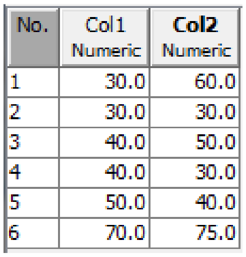
**Question 4**

a) In the context of the k-Means algorithm, state whether the following statements are true or false:

   i. k represents the number of attributes in the dataset. [1]
   ii. k represents the number of clusters. [1]
   iii. The number of clusters is determined by the algorithm. [1]

b) Name and define the measure that is used to evaluate the quality of the clusters that result from the application of the k-Means algorithm. Briefly justify why this is a useful measure, and explain how it is used. [5]

c) Illustrate in detail the first iteration of the k-Means algorithm on the dataset below for k=2, starting with the centres C1 and C2 given by instance No. 2 and instance No. 4, respectively. More precisely, using these two centres, you are required to generate the clusters, then to compute the new centres, and finally to decide if further iterations are needed. Note that only the values of the attributes named *Col1* and *Col2* are to be used in the calculations. [15]

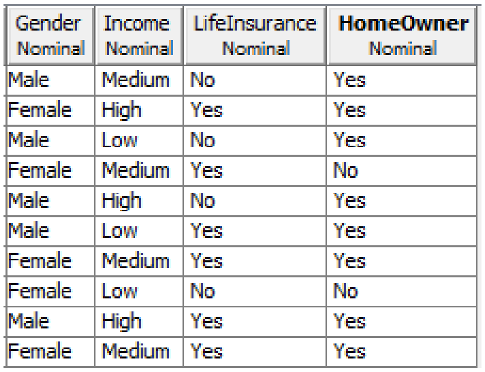


| No. | Col1 | Col2 |
|---|---|---|
| | Numeric | Numeric |
| 1 | 30.0 | 60.0 |
| 2 | 30.0 | 30.0 |
| 3 | 40.0 | 50.0 |
| 4 | 40.0 | 30.0 |
| 5 | 50.0 | 40.0 |
| 6 | 70.0 | 75.0 |

d) Briefly explain how the k-Means algorithm can be extended such that it can handle datasets containing not only numeric attributes but also nominal attributes. Do not use more than two statements. [2]

**Question 5**

a) i. You are required to apply the Apriori algorithm on the dataset below. More precisely generate all the frequent itemsets knowing that the minimum support is 0.4. [10]
ii. Based on the frequent itemsets from (i) above, generate all the association rules and compute their confidences. Finally select the interesting/strong rules with confidence at least 0.8 . [7]



| Gender<br>Nominal | Income<br>Nominal | LifeInsurance<br>Nominal | **HomeOwner**<br>Nominal |
|---|---|---|---|
| Male | Medium | No | Yes |
| Female | High | Yes | Yes |
| Male | Low | No | Yes |
| Female | Medium | Yes | No |
| Male | High | No | Yes |
| Male | Low | Yes | Yes |
| Female | Medium | Yes | Yes |
| Female | Low | No | No |
| Male | High | Yes | Yes |
| Female | Medium | Yes | Yes |

b) Briefly describe four methods used to handle missing data in the preprocessing phase of the KDD process. Do not use more than two statements per method. [8]