Candidates are admitted to the examination room ten minutes before the start of the examination. On admission to the examination room, you are permitted to acquaint yourself with the instructions below and to read the question paper.

Do not write anything until the invigilator informs you that you may start the examination. You will be given five minutes at the end of the examination to complete the front of any answer books used.

May/June 2013 **SE3DM11 2012/13 A 001**

1 Answer Book
Any calculator (including programmable calculator) permitted

UNIVERSITY OF READING

DATA MINING (SE3DM11)

One and a half hours

Answer any **TWO** out of THREE questions.

EACH Question is worth 20 marks

1. Given a classification problem and a dataset, where each record has several attributes and a class label, a learning algorithm can be applied to the data in order to determine a classification model. The model is then used to classify previously unseen data (data without a class label) to predict the class label.

   (a) What is the confusion matrix in the context of a classification model? Provide the definition of accuracy and error rate derived from the confusion matrix. Consider a classification model which is applied to a set of 1000 records (800 belonging to class A and 200 to class B): it correctly predicts 750 records to belong to class A and 190 records to class B. Provide the confusion matrix and compute the accuracy and the error rate. (4 marks)

   (b) Discuss the limitations of accuracy as a performance metric to evaluate a classification model under class imbalance. How can these limitations be overcome with a cost function? (5 marks)

   (c) Provide the definition of precision, recall and F1-measure. Compute them for the example in (a). What is the definition of the general F-measure? (5 marks)

   (d) Describe the holdout method and the cross-validation method for evaluating the performance of a classifier. Compare and contrast them. (6 marks)

2. Given a set of objects each defined by a set of features, clustering is the assignment of the objects into subsets (clusters) so that objects in the same cluster are similar and objects in different clusters are not, according to a given proximity measure defined over the set of features.

   (a) What is the difference between partitional and hierarchical clustering? Compare and contrast them. (5 marks)

   (b) K-means is a partitional clustering algorithm. What is its hierarchical variant? Provide pseudocode for BOTH. (10 marks)

   (c) How can we validate the quality of the set of clusters resulting from a clustering algorithm? (5 marks)

3. A travel agency wants to leverage its vast warehouse of customer data to identify customers most likely to be interested in new products. Targeted mailing campaigns can achieve a large decrease in costs over conventional approaches. The company hires you to identify the customers with affinity for new products. You are required to propose a knowledge discovery and data mining project that will try to identify customer profiles to be used in a targeted mailing campaign.

   (a) Give an overview of the knowledge discovery process you intend to apply. List the different stages and provide a brief description. (10 marks)

   (b) Is a descriptive or a predictive data mining approach appropriate to this task? Identify and justify a specific data mining algorithm that could be applied and discuss its advantages and disadvantages. (5 marks)

   (c) Provide the pseudocode of the specific data mining algorithm you have selected. (5 marks)

(End of Question Paper)