COIS 4400H
Assignment 1

Question 1 (2 marks)
Using the techniques discussed in Week 2, map the following categorical values to numerical representations. Describe which mapping you used and why.

a) Colors of the rainbow
b) Names of Employees
c) Which continent a country belongs to
d) Letters of the Alphabet

Question 2 (4 marks)
Find a recent (within the last five years) journal paper discussing an application of data mining that is not using a decision tree to an application area of your choice. Describe the dataset that was used and the preprocessing techniques that were applied.

Question 3 (4 marks)
When normalizing a dataset, the resulting data will have a minimum value of 0 and a maximum value of 1. However, the dataset we work with in data mining is typically a sample of a population. Therefore, the minimum and maximum for each of the attributes in the population are unknown.
Samples from the population may be added to the dataset over time, and the attribute values for these new objects may then lie outside those you have seen so far. One possibility to handle new minimum and maximum values is to periodically renormalize the data after including the new values. Your task is to think of a normalization scheme that does not require you to renormalize all of the data. Your normalization approach has to fulfill all of the following requirements:

- all values (old and new) have to lie in the range between 0 and 1
- no transformation or renormalization of the old values is allowed

Describe your normalization approach.