Assignment 1

1) Select a real-world company and do your research to answer the following questions.

   a) Describe a big data application that the company is using and use it to illustrate the characteristics of big data analytics. (4)
   b) Suggest and describe one new big data application that would help the company improve their business performance. (4)
   c) Explain why your suggested application is innovative and useful. Discuss the challenges of implementing the application that you proposed. (4)

2) Use data set Breast Cancer Wisconsin (Original)
   https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

   a) Remove the ID attribute. Consider all attributes except the class attribute as numeric attributes. Build a Naïve Bayes model and classify benign and malignant. Show the screenshot of the model when you train the model with all records. (3)

   b) Use one record to explain how the model makes classification. (3)

   c) Show the accuracy of the model using 10-fold cross validation and the confusion matrix. Show and explain the meaning of the precision and recall for malignant. (3)

   d) Discretise the data set using three bins (equal-frequency). (3)

   e) Build a Naïve Bayes model and classify benign and malignant using the discretised dataset. Show the accuracy of the model using 5-fold cross validation and the confusion matrix. Explain the meaning of the numbers in the confusion matrix. (3)

   f) Use one record to explain how the model makes classification using the discretised dataset. (3)