**THE UNIVERSITY OF HONG KONG**
**Department of Computer Science**
**COMP3270B Artificial Intelligence**
**Assignment 3**

Due Date: Sunday, April 9, 2023.

1. Write a python program using `sci-kit learn` (aka `sklearn`) package to build a decision tree for the mushroom dataset from the UCI Machine Learning Repository. (The dataset has been downloaded as `mushroom.csv` and available from Moodle, you may download the same file from the UCI official site). The first attribute is class label (`e` for edible and `p` for poisonous). There are 22 other attributes (features) which are categorical. You will need to convert them to numerical data as the packages expect numerical input. Split the dataset into 75% training and 25% test set, with shuffle. Write down the accuracy for the test set.

2. Repeat the experiment using Naive Bayes Classifier.

3. Use the Naive Bayes and K-NN classifiers on the Breast Cancer (Wisconson) dataset. Again the data has been downloaded as `breast.csv` in moodle. Note that the Breast Cancer dataset contains 11 attributes. The first one is the patient id, which can be ignored. The last one is the class label, 2 for negative (benign) and 4 for positivbe (malignant). The rest features are features valued from 1 to 10. There is a description of all features in the official site. Note that some features are imported as string values, (in particular, feature 6), and has to be converted to numerical data first.

   Repeat the K-NN classification with $k = 1, 3, 5, 7$.

   For each of the classification methods, since the data file is shuffled before splitting between the training and test set, repeat the experiments 10 times and find the average accuray. Write a short report and discuss your result.