INFO411: Data Mining and Knowledge Discovery

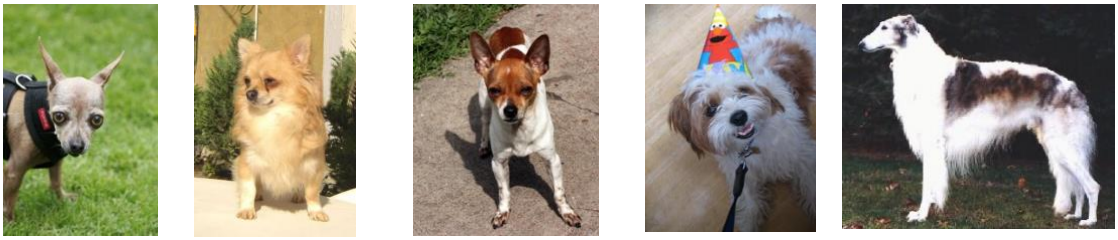
Instructions:
This task is a real-world data mining problem. You are required to prepare a set of presentation slides that must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) your proposed data mining approach and methodology; (3) the strengths and weaknesses of your proposed approach; (4) the performance measures that can evaluate your data mining results; (5) the results and a brief discussion. Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Dogs Breeds Recognition

Background:
Dogs breeds recognition with data mining is a challenging issue. First, there is only slight difference between some dog breeds. Second, a dog of a same breed may appear in different ages, poses, occlusion/self-occlusion and even colors. Furthermore, a large proportion of the images contain humans and are taken in manmade environments leading to greater background variation.

Recently, there has been an increasing interest to develop deep learning models for dogs breed recognition due to their powerful feature representation capability. Briefly, deep models automatically learn feature descriptors from dog images and use them to train classifiers that can distinguish between different dog breeds. Stanford Dogs is a public image dataset and widely used for the development of dog breeds recognition models. More details of the dataset can be accessed from http://vision.stanford.edu/aditya86/ImageNetDogs. Some example images of this dataset are shown below.



The feature descriptors of Stanford Dogs dataset produced with a deep learning model (ResNet-18 trained on ImageNet) have been provided to you with this instruction as the "dog-breeds-recognition.zip" file. By unzipping this file, you shall find the following two files:

1) "training.csv" with 12000 feature descriptors extracted using images from training split of Stanford Dogs dataset. You should use these descriptors for training.
2) "testing.csv" with 8580 feature descriptors extracted using images from testing split of Stanford Dogs dataset. You should use these descriptors for training purpose.
3) Both files has the following data format: image_name<>class_name<>feature_descriptor. There are 120 image classes.

The goal of this task is to train a classification model for dogs breed recognition using provided feature descriptors from Stanford Dogs dataset.

Requirements:
1. Get yourself familiar with the Stanford Dogs dataset and the provided training and test sets. Present a general description of the dataset and present the general properties of the dataset.

2. You are required to implement three classification methods to predict the dog classes. You shall correctly use the provided training and test sets. Also, you need to tune the hyperparameters of your classification models in a principled way.
3. Discuss any data preprocessing or post processing and selection of attributes which have been applied.
4. You need to provide the performance measures of your classification results.
5. Compare the classification models you have implemented and discuss their advantages and disadvantages.