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: Bird Species Recognition

Background:
Bird species recognition with data mining algorithms is a challenging issue. Naturally, birds present in various scenarios appear in different sizes, shapes, colors, and angles from human perspective. Besides, the images present strong variations to identify the bird species more than audio classification. Also, human ability to recognize the birds through the images is limited.

Recently, there has been an increasing interest to develop deep learning based prediction models for bird species recognition due to their powerful feature representation capability. Briefly, deep learning models automatically learn feature descriptors from bird images and use them to train classifiers that can distinguish between different bird species. Caltech-UCSD Birds-200-2011 is a public bird species recognition dataset and widely used for the development of bird species recognition models. More details of the dataset can be accessed from http://www.vision.caltech.edu/visipedia/CUB-200-2011.html. Some example images of this dataset are shown below.

The feature descriptors of Caltech-UCSD Birds-200-2011 dataset produced with a deep learning model (ResNet-18 trained on ImageNet dataset) has been provided to you with this instruction as the "birds-species-recognition.zip' file. By unzipping this file, you shall find the following two files:

1) "training.csv" with 5994 feature descriptors extracted using images from training split of Caltech-UCSD Birds-200-2011 dataset. You should use these descriptors for training.
2) "testing.csv" with 5794 feature descriptors extracted using images from testing split of Caltech-UCSD Birds-200-2011 dataset. You should use these descriptors for test.
3) Both files has the following data format: image_name<>class_name<>feature_descriptor. There are 200 image classes.

The goal of this task is to train classification models for bird species recognition using provided feature descriptors from Caltech-UCSD Birds-200-2011 dataset.

Requirements:
1. Get yourself familiar with the Caltech-UCSD Birds-200-2011 dataset and the provided training and test sets. Present a general description of the dataset and present the general properties of the dataset.
3. You are required to implement three classification methods to predict the bird 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.
4. Discuss any data preprocessing or post processing and selection of attributes which have been applied.
5. You need to provide the performance measures of your classification results.
6. Compare the classification models you have implemented and discuss their advantages and disadvantages.